4

**COMPARAÇÃO DE DOIS INDICADORES DA
DESNUTRIÇÃO MATERNA USANDO REGRESSÃO
E CLASSIFICAÇÃO POR ÁRVORE E LOGÍSTICA
MULTINOMIAL**

LAEL ALMEIDA DE OLIVEIRA

Tese de Doutorado apresentada à
Faculdade de Saúde Pública da
Universidade de São Paulo, Departamento
de Saúde Materno-Infantil, para obtenção do
Título de Doutor em Saúde Pública.

Orientador

**Prof. Dr. ARNALDO AUGUSTO
FRANCO DE SIQUEIRA**

SÃO PAULO
1996

À coluna Tânia, que tem se mostrado um evento raro, consistente ao longo do tempo, de máxima eficiência como fonte de nutrição de ânimo.

Aos heróis Lucas, Jônatas e Mateus, que entenderam que contribuíam, abdicando do tempo de diálogo e lazer, para que o trabalho em gestação viesse a termo.

AGRADECIMENTOS

- Ao Prof. Arnaldo Augusto Franco de Siqueira, pela orientação, apoio e disposição.
- À professora Ana Cristina d'Andretta Tanaka, pela valiosa colaboração desde a coleta dos dados às sugestões e discussão.
- Aos professores colegas do Departamento de Estatística da UFSCAR, em especial à Maria Cecília, ao Milan e ao Benê, pelo tempo e disposição em discutir e opinar.
- A HINDI-Cia. de Investimentos e Participações, pela cessão do programa da técnica CART.
- Ao Lucas, pela colaboração na digitação e confecção do banco de dados.
- Ao “pessoal da ADHONEP”, de São Carlos pela torcida e incentivo.
- À Mara, pela contribuição no resumo em Inglês.
- Ao Auster, Francisco e a IGNIS-Comunicações, pela contribuição na impressão.
- A CAPES/CNPq, pelo suporte financeiro.
- Aos que colaboraram e não foram citados.
- A Deus, pela força e por todas oportunidades de agradecimentos acima.

*"Tudo quanto te vier à mão para fazer,
faze-o conforme as tuas forças, porque no
além para onde tu vais não há obra, nem projetos,
nem conhecimento, nem sabedoria humana."*

Eclesiastes 9:10

ÍNDICE

	página
RESUMO	i
SUMARY	ii
1. INTRODUÇÃO	1
1.1 - Considerações sobre a desnutrição materna.....	1
1.2 - Indicadores do estado nutricional da gestante.....	6
2. CURVAS PARA AVALIAÇÃO DA NUTRIÇÃO MATERNA.....	9
2.1 - A curva de Siqueira.....	9
2.2 - A curva de Rosso.....	10
3. OBJETIVOS.....	13
4. MATERIAL E MÉTODO	14
4.1 - População e Amostra.....	14
4.2 - Abordagem Metodológica.....	19
4.2.1 - Análise da Eficiência, Consistência, Sensibilidade e Especificidade das curvas	19
4.2.2 - Descrição dos dados e Comparações	21
4.2.3. - Classificação: O uso de Árvores e de Regressão Logística.....	22
4.2.3.1 - O modelo logístico para variável dependente multinomial.....	30
4.2.3.1.1 - O modelo logístico simples.....	31
4.2.3.1.2 - O modelo logístico generalizado.....	31

4.2.3.1.3 - A função logito acumulada.....	33
4.2.3.2 - A técnica AID.....	36
A amostra informada.....	38
Taxa de má-classificação	39
Estimador por Resubstituição	40
Estimador por Validação Cruzada	41
4.2.3.3 - A técnica CHAID.....	44
4.2.3.3.1 - O algoritmo da técnica CHAID.....	46
4.2.3.4 - A técnica CART.....	53
Medida de Impureza	53
4.2.3.5 - A técnica do pacote de programas SPLUS	58
4.2.3.6 - A técnica KS.....	63
RESULTADOS E DISCUSSÃO:.....	65
5.1 - Descrição univariada de algumas medidas	66
5.2 - Comparações entre as curvas de Rosso e de Siqueira.....	85
5.2.1 - Construção e Estrutura das Curvas	86
5.3 - Comparação estatística entre as curvas de Rosso e de Siqueira.....	90
5.3.1 - Utilização de Análise exploratória.....	90
Comparação entre as curvas com relação ao prognóstico da adequação do peso do RN.....	102
5.3.2 - Utilização de Regressão Linear por “Stepwise”	116

5.3.3 - Utilização de Regressão Logística Acumulada	118
5.3.4 - Aplicação de Regressão e Classificação por Árvore.....	119
6. COMPARAÇÕES ENTRE AS TÉCNICAS DE CLASSIFICAÇÃO POR ÁRVORE E A LOGÍSTICA MULTINOMIAL	132
6.2 - Validação Cruzada.....	134
6.2.1 - Validação Cruzada em 10 partes.....	134
Estudo da classificação em função de parâmetros.....	135
Estudo da classificação em função do tamanho da árvore.....	146
6.2.2 - Validação cruzada em 5 partes	157
Estudo da Composição das Árvores/Modelos	161
7. CONCLUSÕES.....	166
7.1 - Conclusões relativas às técnicas de classificação.....	166
7.2 - Conclusões relativas às curvas de avaliação do estado nutricional da gestante.....	168
REFERÊNCIAS BIBLIOGRÁFICAS	172
APÊNDICES:.....	180

RESUMO

As curvas de Rosso e de Siqueira para avaliação do estado nutricional de gestantes foram comparadas, em uma amostra de 565 gestantes. Para que essa comparação fosse feita a contento, utilizaram-se técnicas estatísticas tradicionais como: regressão múltipla por stepwise, coeficientes de associação, de concordância, de especificidade e de sensibilidade, métodos gráficos e técnicas mais atuais de classificação, como logística multinomial, regressão e classificação por árvores do tipo AID.

Os resultados indicaram vantagem da curva de Siqueira, em termos de consistência, eficiência e associação, com relação ao prognóstico para a adequação do peso do RN. A análise das interrelações mostraram que a curva de Rosso tem uma tendência em sub-classificar a gestante, com relação ao seu peso e ao peso do seu RN.

Para validar e confirmar tais resultados, procurou-se comparar as técnicas de classificação, uma vez que existiam cinco técnicas, cada qual, devido às suas características, com alguns resultados distintos. A comparação entre as técnicas de classificação foi feita por intermédio da taxa de má-classificação, usando validação cruzada em 10 e 5 partes, obtendo-se vantagem das técnicas CHAID e Logística Multinomial, seguidas de CART, SPLUS, KS.

A validação cruzada permitiu também confirmar a vantagem da curva de Siqueira, com o destaque para o fato que o melhor diagnóstico, com relação ao prognóstico da adequação do peso do RN, é na primeira consulta do pré-natal, abrindo o horizonte de diagnóstico precoce, prevenção e intervenção logo no início da gravidez, o que indica o seu uso com maiores chances de bons resultados.

ABSTRACT

The curves of Rosso and Siqueira, used to evaluate the nutritional state of pregnant women, were compared in a sample of 565 mothers. In order to make this comparison successful, traditional statistical techniques were used such as: multiple regression by stepwise, coefficients of association, concordance, specificity and sensibility, and graphical methods, and the updated techniques of classification, such as: multinomial logistic, regression and classification through trees of AID type.

The results showed the advantage of the Siqueira curve, in terms of consistency, efficiency and association, related to prognosis of the adequacy of the weight of the newborn. The analysis of the interrelations showed that the curve of Rosso has a tendency to under classify the pregnant women concerning their weight and the weight of the newborn.

In order to validate and confirm these results, comparisons of classification of techniques were searched, since there were five techniques and, each one, due to its own characteristics, presented some distinct results. The comparison between the techniques of classification were made by means of misclassification rate, using cross-validation in 10 and 5 parts, where CHAID techniques and Multinomial Logistic showed advantages, followed by CART, SPLUS and KS.

The cross-validation also allowed us confirm the advantage of the Siqueira curve, with emphasis on the fact that the best diagnose, concerning the prognosis of the adequacy of the weight of the newborn, is obtained in the first antenatal care appointment. This expands horizons for precocious diagnose, prevention and intervention at the very beginning of pregnancy, which indicates its use with better chances of good results.

1 - INTRODUÇÃO

1.1 Considerações sobre a desnutrição materna

A relação entre a massa corporal da mãe e o peso do recém-nascido já era conhecida há cerca de 150 anos, embora com interpretações ambíguas e mesmo errôneas (HYTTEN & LEITHC(1964); SIQUEIRA(1981)), o interesse pelo estudo dessa relação já aparecia em meados do século XX (BEILLY & KURLAND(1945); BESSER e col.(1947)); mas somente nas décadas de 60 e 70 a sua importância começou a ser reconhecida(SIQUEIRA(1975); JACOBSON (1975) EASTMAN & JACKSON(1968); NISWANDER e col.(1969)), à medida que se identificava uma relação entre o peso do recém-nascido(RN) e a mortalidade neonatal, como o trabalho de PUFFER & SERRANO(1973) que identifica dentre 1632 óbitos neo-natais em São Paulo, 60% em crianças chamadas Recém Nascidos de Baixo Peso, com 2500 gr ou menos (RNBp). Outros autores também notaram a associação do peso ao nascer com a mortalidade neo-natal (CHASE (1967)).

Se for considerado que em 1994 nasceram 209.372 crianças na cidade de São Paulo (SEADE (1995)), que a porcentagem de

nascimentos de RNB (Recem Nascido de Baixo Peso) está entre 9% e 10% (SIQUEIRA (1979); KRAMER (1987)), e que o número de mortes em idade inferior a 28 dias foi de 3089, tem-se que a força de mortalidade associada ao baixo peso é 13.5 vezes a força de mortalidade associada ao peso acima de 2500 gr, ou seja o risco relativo de morte neo-natal para peso inferior a 2500 gr é de 13.5. No apêndice A1, mostra-se que o mesmo ocorre para o Estado de São Paulo. Problemas desse tipo têm levado a uma maior atenção de pesquisadores, órgãos, entidades, e autoridades envolvidas com a Saúde Pública, para o peso ao nascer.

Diversas pesquisas com animais mostraram que a restrição alimentar durante a gestação resultou em aumento da mortalidade, baixo peso ao nascer, menor tamanho adulto, menor tamanho da ninhada (HSUEH(1974); CHOW(1968)), e também em déficits neurológicos, comportamentais, no aprendizado e desenvolvimento (HSUEH(1974); SIMONSON e col.(1975); SIMONSON&CHOW(1970)).

Em estudos com humanos, segundo Informe de um Grupo de Estudos da OMS(1981), a resistência a enfermidades infecciosas em crianças e adultos depende do seu estado nutricional, o qual deve ter um profundo efeito na imunidade, o que leva a associar o melhoramento das condições de saúde com a disponibilidade de alimentos. CHOW e col(1968) verifica em crianças na idade aproximada de 11 anos, conseqüências depreciadoras da desnutrição materna, tais como incapacidade de recuperação normal do peso depois de um revés nutricional, mesmo tendo à disposição o alimento normal rico em calorias e proteínas, concluindo que os indivíduos filhos de mães que passaram por restrição alimentar na gestação, ficam mal preparados por toda a vida em absorver e aproveitar todo o alimento necessário para recuperar o peso normal, fato este que pode ser atribuído a uma falha no metabolismo devido à dieta materna. Tal experimento pode levantar dúvidas quanto à forma de suplementação alimentar de

crianças desnutridas, pois considerando que a pesquisa de CHOW e col.(1968) pudesse ser generalizada, se a oferta de suplementação alimentar for a mesma para filhos de mães desnutridas e nutridas na gestação, a criança da mãe desnutrida terá maiores chances de não alcançar o peso ideal. Portanto não bastaria socializar o alimento para as crianças, em programa de suplementação alimentar, pois segundo essa tendência a distribuição nesse caso, possivelmente teria maior sucesso se fosse feita em função do estado nutricional da mãe no período de gestação.

Por outro lado KRAMER(1993) verificou modesta alteração em função de suplementação alimentar na gestação, com pouco benefício para as crianças em termos de desenvolvimento físico ou neuro-cognitivo.

NOBREGA(1981) mostra diversos trabalhos em que a desnutrição intra-uterina resulta em problemas ao conceito entre os quais enzimáticos, metabólicos, alterações morfológicas(peso, pele, olhos, músculo esquelético, ossos e dentes, baço, pulmões, coração, artérias, rins, pâncreas, fígado, cérebro, hipófise), imunológicas, aparelho digestivo, pulmonar, sistema nervoso central, e de desenvolvimento. SIQUEIRA⁶⁴, depois de analisar uma gama de trabalhos na área conclui que a desnutrição pode provocar comprometimento do sistema nervoso central, “prejuízos esses que poderão afetar sua “performance” ao longo de sua vida”, além da relação já citada do tamanho do recém-nascido com a mortalidade neo-natal.

Comentando diversos trabalhos, SIQUEIRA(1979) refere que as deficiências nutricionais da mãe provocam prejuízos diferenciados ao conceito dependendo do período de gestação; como é o caso de danos no cérebro em função da desnutrição do sexto mês de gestação até 2 anos de idade(OMS(1990)). Outras pesquisas feitas em diversas regiões da Europa, concluem que suas descobertas estão de acordo com as hipóteses derivadas do conhecimento sobre as fases de crescimento humano. Ou seja, uma restrição

alimentar na última metade da gestação afeta o crescimento fetal, enquanto que uma restrição alimentar nos primeiros meses afeta a formação, o desenvolvimento do embrião e a sua capacidade de sobrevivência (WORTHINGTON e col.(1980) .

Um problema característico de países em crise de desenvolvimento, como o Brasil, em que a população sofre restrições no acesso da aquisição de alimento, embora haja produção suficiente de alimento, não há equidade na sua distribuição, é a desnutrição, que afetando todas as idades, tem na criança e na mãe desnutrida o foco da atenção e preocupação.

Do que foi citado, e de outros estudos, tem-se uma alta probabilidade de que a desnutrição da mãe esteja fortemente correlacionada com a desnutrição do filho. Assim, identificando-se a gestante desnutrida, aumenta-se a chance de identificar o conceito em risco, para que sejam tomadas as providências no sentido de evitar os possíveis danos encontrados nas diversas pesquisas. Já em 1966 JELLIFFE alertava para que em todo exame de pré-natal feito em países em desenvolvimento, deveria ser dada ênfase na detecção precoce da desnutrição e na sua prevenção.

Como concluíram diversos autores (NOBREGA(1981); JURADO GARCIA(1971); CIARI(1974); FIKREE & BERENDES(1994); GEBRE MEDHIN & GOBENZIE (1975); BERGNER & SUSSER (1970); MONTEIRO(1995); JELLIFFE(1966)), deve-se lembrar que o ambiente tem forte influência sobre o indivíduo, e conseqüentemente a desnutrição não deve ser sozinha um fator único causal. FIKREE & BERENDES(1994) encontraram uma influência da educação(sem primário completo, com primário completo) sobre o peso do RN, em um grupo de mães pesquisadas em Karachi, Paquistão; da mesma forma, notaram que o uso de água não encanada residencial é um fator de risco considerável para o baixo peso ao nascer. Mas a identificação da desnutrição é o primeiro passo da

prevenção, que poderia ser aprimorada se houvesse uma política facilitadora de acesso ao alimento, adicionada a uma educação da população quanto à melhor forma de utilização desse alimento.

Não é por acaso que o Comitê de Saúde Materno-Infantil da American Medical Association, publicou o trabalho intitulado Reducing Infant Mortality (1965), comentado por SIQUEIRA(1979), afirmando que “o processo de socialização produz uma tendência para certas características sócio-biológicas se associarem. Grupos sócio-econômicos são caracterizados por similaridades em renda, educação, **hábitos alimentares**, Pessoas dentro de um grupo social são também semelhantes no **físico, estatura, saúde básica**, ...”. (grifo nosso). Motivo pelo qual pode-se usar o peso do RN ou a porcentagem de baixo peso ao nascer para comparar os níveis da qualidade de vida entre países ou regiões, como é feito com a mortalidade neo-natal. MURRAY & LOPEZ(1994) mostraram em um estudo dos padrões de causas de morte em diversas regiões do mundo, no ano de 1990, que existe uma associação entre o desenvolvimento econômico do país ou região e a taxa de mortalidade perinatal. KRAMER(1987) mostra a influência do meio, ao comparar o peso médio do RN e a prevalência de baixo peso ao nascer em função da localização geográfica.

Qualquer que seja o modo de prevenir a desnutrição, só será obtido sucesso se apoiado por uma motivação extensa da população alvo. Inspirados na importância da prevenção da desnutrição, e outros riscos, tem-se intensificado o alerta e a educação de várias populações, como por exemplo o trabalho de KAFATOS e col.(1989), em que a educação continuada em uma população de gestantes na Grécia, com relação ao melhor aproveitamento do alimento disponível, mostrou forte influência sobre o ganho de peso gestacional e a taxa de mortalidade neo-natal. Nesse sentido WORTHINGTON e col(1980)

mostram diversos cuidados e princípios de aprendizado, como motivação e envolvimento da pessoa em aprendizado, combinar o novo com o familiar e medir o aprendizado pela modificação de comportamento.

1.2 - Indicadores do estado nutricional da gestante

Cientes da importância da desnutrição intra-uterina, muitos autores (BIRCH(1972); CIARI(1974); SIQUEIRA(1979); WORTHINGTON e col.(1980); ROSSO(1985)) ressaltam a necessidade de indicadores do estado nutricional das mães, antes e durante a gestação. O diagnóstico nutricional da gestante não tem merecido tanta atenção dos pesquisadores como tem merecido a busca dos fatores associados à desnutrição; embora na presente década tenha aumentado o interesse pelo diagnóstico do estado nutricional da gestante (WHO(1995), KRASOVEC & ANDERSON(1991)).

NOBREGA(1981) enfoca o problema em termos populacionais, no sentido de indicarem-se medidas de saúde pública para a resolução da desnutrição da população como um todo, dependendo para isso de estatísticas oficiais, as quais por seu lado são muito escassas, inconfiáveis ou oscilantes.

Outros pesquisadores procuram analisar o indivíduo, utilizando principalmente os seguintes métodos de avaliação nutricional:

a) Observações clínicas: Segundo WORTHINGTON e col.(1981), tais observações não são muito eficientes devido a uma “considerável variação em relação ao observado na interpretação dos sinais físicos, parcialmente devido a diferenças de conhecimento e experiência, falta de padronização da definição de sinais particulares e a pouca incidência de sinais de desnutrição”. Ainda com relação às dificuldades nas observações clínicas, WORTHINGTON alega que “existe muita divergência na interpretação específica de sinais durante a gravidez, por exemplo, hipertrofia gengival, que muitas vezes pode ocorrer normalmente na gravidez e pode não ser necessariamente um sinal de carência de ácido ascórbico.”

b) Avaliação de Peso e Altura: As medidas antropométricas têm sido usadas tanto pela facilidade de avaliação, como pela alta associação encontrada entre algumas delas e a desnutrição materna. Devido a essas associações já verificadas, a maioria das medidas antropométricas usadas para diagnóstico do estado nutricional têm sido o peso, a altura, perímetro braquial, perímetro cefálico, algumas relações entre peso e altura, medidas como pregas cutâneas, e relações entre essas medidas e a idade(SIQUEIRA (1979); KRASOVEC & ANDERSON(1991)). O aumento ou perda de peso durante a gravidez (peso na consulta - peso pré-gestacional) também tem sido muito usado, por ter sido encontrada alta correlação entre essa medida e o peso do recém-nascido(SIQUEIRA(1981); SIQUEIRA(1979); HSUEH(1974); KRASOVEC & ANDERSON(1991); WHO(1995)). Outras vantagens das medidas antropométricas são o baixo custo, fácil aplicabilidade e não exigência de instalações especiais, portanto especialmente indicadas para a situação brasileira.

c) Testes Laboratoriais: Alguns autores, embora vejam problemas na aplicação de testes laboratoriais para detectar a nutrição materna durante a gestação, indicam os testes laboratoriais para o período do início da gravidez, pela necessidade vital de

avaliação e controle no decorrer da gestação(WORTHINGTON(1980)). Outros autores indicam os testes laboratoriais com ressalvas; SIQUEIRA(1979) e CIARI(1974) mostram que indicadores bioquímicos não devem ser utilizados para tal finalidade, uma vez que têm grande chance de serem registrados com ruído, mascarados ou alterados, como por exemplo devido às alterações hematológicas próprias da gravidez, aumento da volemia, modificações metabólicas, etc...

d) Curvas Indicadoras do Estado Nutricional: Em 1974 CIARI mostrava a necessidade de estabelecer uma metodologia que incluísse uma ficha de registro de Pré-Natal, com anotações que permitissem a “avaliação mais correta possível do grau de nutrição da grávida”. Nessa época, CIARI alertava para o fato que “não se tem feito a comparação do ganho de peso entre gestantes que no início da gravidez apresentam peso mais baixo, mais alto ou normal. Tem-se analisado a curva ponderal da gravidez em grupo de mulheres sem contudo distingui-las”. Acrescente-se a isso o fato de que a utilização de tabelas ou curvas de crescimento fetal, peso ganho, etc., criadas em países de populações de estrutura física diferente da realidade brasileira não são adequadas para o uso no Brasil. Discutindo formas de construir indicadores do estado nutricional, SIQUEIRA e col. em 1975 e ROSSO em 1985 entre outros (TANNER & THOMSON(1970); JELLIFFE(1968); GUERI e col.(1982); KRASOVEC & ANDERSON(1991); WHO(1995)), propõem métodos distintos para diagnosticar o estado nutricional da gestante. Embora ambos sejam diferenciados, são refinamentos e elaborações baseados em medidas antropométricas que, como foi visto são as mais usadas e indicadas no âmbito da saúde pública.

2 - CURVAS PARA AVALIAÇÃO DA NUTRIÇÃO MATERNA:

2.1 - A curva de Siqueira:

Procurando sanar essas lacunas, SIQUEIRA e col.(1975 e 1979) construíram curvas de ganho de peso para a idade gestacional, utilizando medidas antropométricas feitas em 1354 pacientes na cidade de São Paulo, tais como :

Altura da mãe, Peso pré-gestacional, Peso da gestante mês a mês e a respectiva Idade gestacional, Peso no parto, Peso do RN no parto.

Reafirmando que o diagnóstico da desnutrição da mulher não é o fim, SIQUEIRA(1979) lembra que o objetivo de uma técnica, ao diagnosticar uma gestante como desnutrida, é o de “poder afirmar que o risco de que o seu feto sofra um retardo do crescimento, seja maior do que o de uma mulher normalmente nutrida”.

A curva de Siqueira foi construída baseada no fato de que há um comportamento diferente da curva ponderal, em função da situação do peso da mulher no início da gestação(SIQUEIRA(1975); CIARI(1974)). Ela consta de 5 gráficos que foram construídos com base na distribuição dos pesos ganhos, a partir do segundo mês de gestação. Cada um desses gráficos é para um conjunto de mulheres dentro de uma das 5 faixas de estatura: 1,49 m e menos, 1,50 m a 1,54 m, 1,55 m a 1,59 m, 1,60 m a 1,64 m e 1,65m e mais. Assim, para cada uma dessas faixas de estatura, temos em cada mês uma distribuição dos pesos próxima da Gaussiana, com um intervalo em torno da média(média $\pm$ DP) que identifica a faixa de normalidade em aproximadamente 70% dos casos. As mulheres com peso acima do limite superior dessa faixa, aproximadamente 15% dos casos, foram

consideradas obesas, e as mulheres com peso abaixo do limite inferior dessa faixa, aproximadamente 15% dos casos, foram consideradas desnutridas. Assim, de posse da estatura da gestante, a partir do segundo mês é possível diagnosticar o seu estado nutricional com base no seu peso. Logicamente, tais curvas se adequam à realidade brasileira e mais especificamente à realidade do sudeste brasileiro. CIARI(1974) em um ensaio à procura do diagnóstico da desnutrição, construiu uma curva semelhante, mas considerando apenas a média e não um intervalo, e apenas 4 grupos de estaturas.

Para validar seu método, SIQUEIRA(1979) usou o seu conjunto de curvas em diferentes populações apresentando resultados semelhantes; mulheres classificadas como desnutridas por suas curvas também o foram quando outros indicadores já conhecidos eram aplicados.

2.2 - A curva de Rosso:

O método de ROSSO(1985) consta da utilização de um gráfico que foi construído a partir da adequação do peso para a altura de 262 mulheres normais, sem vícios de tabaco e álcool, registradas em um serviço de pré-natal de uma população de baixa renda, racialmente mista e residente em Nova York. As informações registradas no seu estudo foram:

Idade, Altura, Peso pré-gestacional, Peso no parto, Número de gestações, Duração da Gestação, Peso do Recém Nascido.

ROSSO(1985) notou que o peso ganho durante a gestação tem uma clara influência sobre o peso do RN e, além disso, que se trata de uma correlação positiva apenas para mulheres desnutridas e normais, e não para

mulheres obesas, como haviam notado diversos outros pesquisadores (SIQUEIRA(1975); SIQUEIRA(1979); KRAMER(1987)). Para tal verificação, ROSSO(1985) tomou uma tabela de altura versus peso da Sociedade Atuarial(JELLIFFE(1982)), como padrão, e calculou um peso para cada gestante em termos de porcentagem do “peso padrão”. Tal correção foi efetuada nas várias idades gestacionais. Semelhantemente CIARI(1974) em 1974 considerou a média dos pesos de um grupo de gestantes de estaturas próximas como sendo o padrão. Para ROSSO(1985), o gráfico do peso do RN versus o peso da gestante corrigido em termos de porcentagem do padrão, e o estudo da correlação, mostraram uma relação entre o peso do RN e o peso corrigido da gestante sempre crescente, até que a mãe chegasse a 120% do seu padrão. Para qualquer valor do peso da mãe acima de 120% do padrão, o peso do RN passou a se manter constante, indicando a não existência de influência do ganho de peso sobre o peso do RN. Desses resultados, como o crescimento corporal foi de 20% para mulheres normais, e como em média o peso pré-gestacional no seu trabalho foi de aproximadamente 60 kg, tem-se um ganho de peso indicado de 12 kg.

O fato de que o crescimento fetal tem maior chance de alcançar um ponto ótimo, quando a mãe obtém uma massa corporal definida, isto é, 20% a mais do padrão, sugeriu a ROSSO(1985) a construção de um gráfico onde pudesse acompanhar o ganho de peso gestacional, conjuntamente com o peso desejável segundo o padrão.

Para que se pudesse fazer a conversão do peso da gestante no momento da consulta, em termos de porcentagem do peso padrão, para uma dada altura, ROSSO(1985) desenvolveu um nomograma, como da figura 2.2.1, com dois eixos paralelos, referentes respectivamente à altura e ao peso da gestante, e um eixo inclinado que define o valor do peso corrigido em termos de porcentagem do “peso padrão”. ROSSO(1985) não informa como construiu a relação entre esses

eixos, e o ângulo de inclinação do terceiro eixo. A porcentagem do peso padrão é a projeção no terceiro eixo de uma reta que passa pelos valores da altura e do peso da gestante.

A idéia da classificação das gestantes foi a mesma da usada por CIARI(1974) e por SIQUEIRA(1975), ou seja, com base no peso pré-gestacional corrigido, as mães foram classificadas como normais quando seu peso corrigido estava entre 90 e 110 por cento do peso padrão da tabela da Sociedade Atuaria. As mães eram classificadas como desnutridas, quando seu peso corrigido estivesse abaixo de 90 por cento do peso padrão, e eram classificadas como obesas, quando seu peso corrigido estivesse acima de 110 por cento do peso padrão. A partir desses valores, seguindo a classificação inicial, ROSSO(1985) construiu uma curva ponderal registrando a evolução dos pesos, em termos de porcentagem do padrão, constituída de uma faixa de normalidade, outra de obesidade e outra de desnutrição, para um seguimento pré-natal no período de 12 semanas até 42 semanas.

Para validar seu método, ROSSO(1985) verificou que em duas populações distintas, de baixa renda, as mães classificadas como tendo peso maior que 120% do seu padrão, tinham em média peso de RN bem superiores aos demais.

3 - OBJETIVOS:

Como existem duas técnicas para o diagnóstico da desnutrição materna ainda não comparadas, e diversas técnicas estatísticas disponíveis para uso em comparações desse tipo, mas também ainda não comparadas para situações como a apresentada aqui, os objetivos do presente trabalho podem ser divididos em dois:

- 1 . Comparar o método proposto por Rosso com o proposto por Siqueira, quanto à sensibilidade, especificidade, eficiência e consistência para o diagnóstico da desnutrição materna.
- 2 . Fazer uma comparação das técnicas estatísticas disponíveis e usadas na análise dos dados para alcançar o objetivo anterior, procurando encontrar a mais indicada quanto à eficiência e consistência.

4 - MATERIAL E MÉTODO:

4.1 - População e Amostra:

Para que os objetivos fossem alcançados, considerou-se como amostra os dados coletados na Grande São Paulo, por uma equipe de pesquisadores* do Departamento de Saúde Materno-Infantil da Faculdade de Saúde Pública da USP, que consiste em dois grupos de gestantes, de duas localidades distintas. O primeiro grupo, denominado aqui de FSP, consta de todas as gestantes inscritas e atendidas no serviço de pré-natal do Centro de Saúde Geraldo de Paula Souza, da Faculdade de Saúde Pública da USP, no período de janeiro de 1977 a dezembro de 1986, totalizando 530 gestantes. O segundo grupo, denominado aqui de COT, consta de todas as gestantes que eram internadas para dar à luz na Associação Hospital de Cotia, no período de agosto de 1983 a junho de 1985, totalizando 35 gestantes.

Como critério de escolha, descartaram-se os casos com patologias a não ser a desnutrição, de modo que a população estudada foi de apenas gestantes consideradas sadias.

Com o intuito de trabalhar com os dois grupos de dados, foram feitos alguns testes de hipóteses para comparar as clientela da FSP e COT, com relação as médias de algumas variáveis, os quais são apresentados no apêndice A4.2.

A cidade de Cotia faz parte da área metropolitana da grande São Paulo, tendo como atividades principais a agricultura e a indústria. A partir da comparação apresentada no apêndice A4.1, da oferta e acesso dos serviços de assistência à saúde da cidade de São Paulo e da grande São Paulo, onde se encontra Cotia, da análise gráfica e intervalar das médias dos pesos das gestantes no início da gravidez, das estaturas das gestantes, dos pesos dos RN, em função dos tamanhos desses grupos, considerou-se ora a utilização da amostra da FSP ora a junção dos dois grupos em uma única amostra para o estudo proposto.

Os dados do primeiro grupo foram obtidos das fichas de pré-natal do arquivo do citado Centro de Saúde, enquanto os dados do segundo grupo vieram dos registros da carteirinha da gestante .

As medidas registradas para o estudo foram:

- **Local**: Para identificar a origem dos dados, usou-se a codificação COT para o Hospital de Cotia, e FSP para Centro de Saúde da Faculdade de Saúde Pública. Para facilitar, em alguns registros e análises estatísticas, a variável LOCAL será chamada de V2
- **Estatura da gestante**: Registrada em centímetros. Para facilitar, em alguns registros e análises estatísticas, a variável ESTATURA DA GESTANTE será chamada de V3.
- **Peso Pré-Gestacional**: Registrado em hectogramas. Para facilitar, em alguns registros e análises estatísticas, a variável PESO PRÉ-GESTACIONAL será chamada de V4.
- **Peso na Primeira Consulta**: Registrado em hectogramas Para facilitar, em alguns registros e análises estatísticas, a variável PESO NA PRIMEIRA CONSULTA será chamada de V5.

- **Tempo de Gestação na Primeira Consulta:** Registrado em semanas Para facilitar, em alguns registros e análises estatísticas, a variável TEMPO DE GESTAÇÃO NA PRIMEIRA CONSULTA será chamada de V6.
- **Estado Nutricional segundo Siqueira na Primeira Consulta:** Registrado segundo as três possíveis classificações de estado nutricional: DES para desnutrida, NOR para normal e OBS para obesa. Para facilitar, em alguns registros e análises estatísticas, a variável ESTADO NUTRICIONAL SEGUNDO SIQUEIRA NA PRIMEIRA CONSULTA será chamada de V7.
- **Estado Nutricional segundo Rosso na Primeira Consulta:** Registrado segundo as três possíveis classificações de estado nutricional: DES para desnutrida, NOR para normal e OBS para obesa. Para facilitar, em alguns registros e análises estatísticas, a variável ESTADO NUTRICIONAL SEGUNDO ROSSO NA PRIMEIRA CONSULTA será chamada de V8.
- **Valor da Escala de Rosso na Primeira Consulta:** Trata-se do valor numérico da escala do nomograma de Rosso para avaliação do estado nutricional materno. Para facilitar, em alguns registros e análises estatísticas, a variável VALOR DA ESCALA DE ROSSO NA PRIMEIRA CONSULTA será chamada de V9.
- **Peso da Gestante na Última Consulta:** Registrado em hectogramas. Para facilitar, em alguns registros e análises estatísticas, a variável PESO DA GESTANTE NA ÚLTIMA CONSULTA será chamada de V10.
- **Tempo de Gestação na Última Consulta:** Registrado em semanas. Para facilitar, em alguns registros e análises estatísticas, a variável TEMPO DE GESTAÇÃO NA ÚLTIMA CONSULTA será chamada de V11.
- **Estado Nutricional Segundo Siqueira na Última Consulta:** Registrado segundo as três possíveis classificações de estado nutricional: DES para desnutrida, NOR para normal e OBS para obesa. Para facilitar, em alguns registros e análises

estatísticas, a variável ESTADO NUTRICIONAL SEGUNDO SIQUEIRA NA ÚLTIMA CONSULTA será chamada de V12.

- **Estado Nutricional Segundo Rosso na Última Consulta:** Registrado segundo as três possíveis classificações de estado nutricional: **DES** para desnutrida, **NOR** para normal e **OBS** para obesa. Para facilitar, em alguns registros e análises estatísticas, a variável ESTADO NUTRICIONAL SEGUNDO ROSSO NA ÚLTIMA CONSULTA será chamada de V13.
- **Valor da Escala de Rosso na Última Consulta:** Trata-se do valor numérico da escala do nomograma de Rosso para avaliação do estado nutricional materno. Para facilitar, em alguns registros e análises estatísticas, a variável VALOR DA ESCALA DE ROSSO NA ÚLTIMA CONSULTA será chamada de V14.
- **Duração da Gestação 1:** Trata-se do tempo de gestação até o dia do parto, medido em dias. Para facilitar, em alguns registros e análises estatísticas, a variável DURAÇÃO DA GESTAÇÃO 1 será chamada de V15.
- **Duração da Gestação 2:** Trata-se do tempo de gestação até o dia do parto, medido em semanas. Para facilitar, em alguns registros e análises estatísticas, a variável DURAÇÃO DA GESTAÇÃO 2 será chamada de V16.
- **Peso do Recém-Nascido:** Registrado em gramas. Para facilitar, em alguns registros e análises estatísticas, a variável PESO DO RECÉM-NASCIDO será chamada de V17.
- **Sexo do Recém-nascido:** Com o registro F para feminino e M para masculino. Para facilitar, em alguns registros e análises estatísticas, a variável SEXO DO RECÉM-NASCIDO será chamada de V18.
- **Adequação do Peso do Recém-Nascido:** Medida do peso em função da idade gestacional, segundo a curva de Williams(1982). Registrada segundo as três possíveis classificações de adequação nas categorias: **PIG**(pequeno para a idade gestacional), **AIG**(adequado para a idade gestacional) e **GIG**(grande para a idade

gestacional). Para facilitar, em alguns registros e análises estatísticas, a variável ADEQUAÇÃO DO PESO DO RECÉM-NASCIDO será chamada de V19.

- **Tipo de Parto:** Registrado segundo as formas de parto. Cesariana com o código C, Forceps com o código F e Normal com o código N. Para facilitar, em alguns registros e análises estatísticas, a variável TIPO DE PARTO será chamada de V20.

Algumas variáveis calculadas:

A partir das variáveis medidas na amostra, calcularam-se algumas variáveis que têm sido encontradas na literatura, que de alguma forma estão relacionadas com o presente trabalho:

Peso Relativo:

Os pesos da gestante durante a gravidez foram também considerados com relação à estatura usando o Índice de Massa Corporal, também chamado índice de Quetelet(1870) da seguinte forma:

$$\text{pré-gestacional: } V4PV3Q = \frac{V4}{(V3)^2}$$

$$\text{na primeira consulta: } V5PV3Q = \frac{V5}{(V3)^2}$$

$$\text{na última consulta : } V10PV3Q = \frac{V10}{(V3)^2}$$

Peso Ganho na Gestação:

O peso ganho foi calculado como sendo:

PESOG1=V10-V5 com relação ao peso na primeira consulta,
e **PESOG2=V10-V4** com relação ao peso pré-gestacional.

Além disso, o peso ganho na gestação foi também calculado em termos da estatura como o índice de Quetelet, de duas formas;

com relação ao peso da primeira consulta:
$$\text{PESOGRI} = \frac{(V10 - V5)}{(V3)^2}$$

e

com relação ao peso pré-gestacional:
$$\text{PESOGRI2} = \frac{(V10 - V4)}{(V3)^2}$$

Além dessas variáveis calculadas, em alguns momentos da análise, por exigência da técnica empregada, algumas variáveis contínuas foram transformadas em categóricas ao partir o campo de variação em intervalos, recebendo novo nome como:

V3 foi transformada em **V33=trunc(V3/2)** que resultou em 20 intervalos

V4 foi transformada em **V34=trunc(V4/200)** que resultou em 26 intervalos

V5 foi transformada em **V35=trunc(V5/200)** que resultou em 30 intervalos

V10 foi transformada em **V40=trunc(V10/200)** que resultou em 26 intervalos

V15 foi transformada em **V45=trunc(V15/5)** que resultou em 20 intervalos

V17 foi transformada em **V47=trunc(V17/100)** que resultou em 28 intervalos

PESOG1 foi transformado em **tpesog1=trunc(PESOG1/100)** que resultou em 19 intervalos

PESOG2 foi transformado em **tpesog2=trunc(PESOG2/150)** que resultou em 22 intervalos

4.2 - ABORDAGEM METODOLÓGICA

4.2.1 - Análise da Eficiência, Consistência, Sensibilidade e Especificidade das curvas:

A eficiência em física, é a diferença entre a energia aplicada e o trabalho produzido(SMITH & VAN NESS), e em estatística é a propriedade de uma técnica ter menos variação do que outras técnicas, que objetivam medir o mesmo fenômeno(MOOD & GRAYBILL & BOES (1974)).

No contexto da estatística, a consistência está mais voltada para o número de vezes que a técnica é aplicada. A técnica é consistente, se à medida que é aumentado o tamanho da amostra para sua aplicação, seus resultados tenderem a uma constante esperada(MOOD & GRAYBILL & BOES(1974), WILKS(1971)).

Assim, dada a natureza, a variabilidade e multiplicidade de fatores que podem ser observados no pré-natal, e possam estar relacionados com o desenvolvimento dos fetos, deve-se buscar a metodologia de análise que seja eficiente, tanto no sentido de se obter o resultado, da identificação dos riscos aos quais os futuros bebês estão sujeitos, como no sentido de discriminar a mãe desnutrida com recém-nascido de baixo peso. Além disso, o ideal é que além de eficiente, a técnica a ser procurada deve ter resultados com pouca oscilação se aplicada mais que uma vez, e que seja consistente à medida que se aumenta o tamanho amostral.

O cálculo da sensibilidade(FORATTINI(1976)), também conhecida como sensibilidade (FLEISS(1981); SOARES & BARTMAN(1983)), e da especificidade de uma técnica é feito quando a mesma tem como resultado uma

variável dicotômica, como por exemplo uma técnica que procura identificar se o bebê irá nascer com peso abaixo de 2500g(RNBP) ou acima de 2500g(não RNBP), comparando-se posteriormente o diagnóstico da técnica com os valores reais do recém-nascido. Isso implica na existência de uma tabela de dupla entrada com duas linhas e duas colunas no formato da tabela 4.2.1.1, onde a proporção em que a técnica indica RNBP dado que o RN tem peso até 2500g estima a sensibilidade da técnica, a proporção em que a técnica indica não RNBP dado que o RN tem peso acima de 2500g estima a especificidade, e o valor preditivo positivo é estimado pela proporção em que o RN tem peso até 2500g dado que a técnica indica RNBP (FLEISS(1981)).

TABELA 4.2.1.1:

peso real do RN	Resultado da técnica em teste		Total
	RNBP	não RNBP	
até 2500g	A	B	C
acima de 2500g	D	E	F
Total	G	H	N

Assim a sensibilidade para prever peso até 2500g na tabela 4.2.1.1 é dada por A/C , a especificidade é E/F e o valor preditivo é A/G .

Portanto o cálculo da sensibilidade e da especificidade neste trabalho estará restrito a uma alteração nos diagnósticos das curvas, de três categorias (DES, NOR, OBS) para duas categorias(DES e Não DES, NOR e Não NOR,), e a uma alteração na adequação do peso do RN, de tres categorias (PIG, AIG, GIG) para duas categorias(PIG e não PIG, AIG e não AIG,...). Lógicamente ao se fazer essas junções de categorias de diagnóstico e de categorias de adequação

do peso do RN, perdem-se informações. O ideal é que a técnica seja eficiente e consistente, respaldada pela sensibilidade e pela especificidade quando possível.

4.2.2 - Descrições dos dados e Comparações:

Inicialmente foram usados métodos de descrição e comparação gráfica, e por intervalo de 95% de confiança, utilizando um micro-computador 486-DX de 66 Mhz com 8 Mb de memória RAM e os pacotes estatísticos SPSS(Statistical Package for Social Scientists(1995)), SAS(BARRY e col.(1986)) e SPLUS(CLARK & PREGIBON(1992)). Segundo HOAGLIN, MOSTELLER E TUKEY(1983) a caixa do desenho box-plot, também conhecido como desenho esquemático, cobre 50% da distribuição e a distância entre os valores extremos do desenho box-plot cobre 99.3% da distribuição. Além disso, os pontos aberrantes, que caem fora dessa região são destacados, com suas respectivas identificações. Por esse motivo, e porque em muitos casos será de interesse a forma da distribuição e não a estimação de parâmetros, para que se possa ter uma idéia mais abrangente e detalhada da distribuição, muitas vezes neste trabalho o desenho box-plot será mais indicado que o intervalo de confiança ao nível de 95%.

Posteriormente para o estudo da associação entre variáveis categóricas, ou categorizadas, utilizaram-se o Risco Relativo, o coeficiente de contingência de Pearson, o Qui-quadrado de Pearson, os coeficientes Phi e V de Cramér e os coeficientes de concordância Gama e Somer's. Para o estudo da interrelação entre as variáveis contínuas utilizaram-se os coeficientes de correlação de Pearson, e o coeficiente de correlação parcial fixando a variável peso pré-gestacional. Para comparações entre medidas de posição central, foram feitos testes F ou entalhes nos box-plots.

Para o estudo da forma de relação entre as variáveis foram feitos diagramas de dispersão, ajustadas regressões linear simples e múltipla, e regressão polinomial.

Para a comparação entre as curvas além do método gráfico e da análise das correlações, foram usadas regressão linear múltipla por “stepwise”, regressão logística, e regressão por árvore.

Todas as tabelas apresentadas serão para os dados da amostra em estudo.

4.2.3 - Classificação: O uso de Árvores e da Logística:

Uma característica do banco de dados proveniente de pesquisas desse tipo é a diversidade do tipo de variáveis, sendo muito constante a existência de variáveis categóricas. Considerando a variável como sendo categórica, quando seus resultados assumem categorias ou classes, como por exemplo sexo, com as categorias: M=masculino e F=feminino; estado nutricional com as categorias: desnutrida, normal, obesa; Adequação do Peso ao nascer com as categorias: PIG(pequeno para a idade gestacional), AIG(adequado para a idade gestacional) e GIG(grande para a idade gestacional), etc..., várias técnicas estatísticas para trabalhar com esse tipo de dados têm sido desenvolvidas, indo desde a tabulação de dupla entrada, à razão de estatísticas tais como “odds ratio” e “risco relativo”, e técnicas mais complexas tais como modelos de regressão logística e modelos log-lineares. Esses modelos podem ser usados para analisar dados em um contexto similar ao de análise de modelos lineares, tais como regressão, análise de variância, etc... Com a crescente necessidade do tratamento computacional, começaram a aparecer no mercado diversos “softwares” e módulos de “softwares”,

tais como: GLIM(Generalized Linear Interactive Modeling), SAS(CATMOD, LOGISTIC), BMDP, SPSS especialmente construídos para tais análises.

Examinando as listas de disciplinas oferecidas por cursos de graduação e pós-graduação em estatística, das universidades norte-americanas, européias e brasileiras, e mesmo nas publicações, tanto periódicos como livros, veremos que, apesar do aumento do uso de dados categóricos em pesquisas científicas, os métodos para sua análise não têm crescido juntamente, como ocorreu com modelos lineares. Um dos desenvolvimentos de modelos lineares, que tem buscado trabalhar com dados categóricos, é o Modelos Lineares Generalizados (McCULLAGH & NELDER(1989)).

Portanto, muitos pesquisadores que possam estar necessitados em usar técnicas adequadas para esse tipo de dados, ficam simplesmente restritos ao uso do que existe, e mais restritos ainda devido à não popularização de muitas técnicas já existentes.

Para o presente estudo será adotado um modelo em que o peso do RN, ou mais especificamente, a adequação do peso do RN, seja a variável dependente de um conjunto de outras variáveis, medidas direta ou indiretamente na gestante, estando entre elas os diagnósticos fornecidos pelas curvas de Siqueira e de Rosso. O ajuste de um modelo desse tipo deverá encontrar as variáveis que tiverem maior influência sobre o peso do RN e, ao mesmo tempo, facilitar a comparação entre as curvas de Rosso e de Siqueira. O mecanismo esperado para esse modelo deverá ser o de classificar cada indivíduo, a partir da análise de todas suas variáveis independentes, em um dos três grupos de adequação do peso do RN. Em outras palavras, esperar-se-á que usando os registros relativos às gestantes, o modelo as discrimine entre aquelas que provavelmente pertençam a um ou outro dos grupos de adequação do peso do RN, contribuindo no prognóstico do grupo de adequação do peso, ao qual o seu conceito provavelmente deverá pertencer.

A terminologia usada neste trabalho seguirá a proposta por R.A.FISHER(1938) que, segundo JOHNSON & WICHERN(1982), diferencia discriminação de classificação. A Discriminação é um procedimento de separação ou agrupamento, com base em antigas investigações e parâmetros. Por exemplo: em uma população de gestantes, as gestantes que tiverem com peso/estatura no extremo inferior da distribuição dessa estatística são consideradas desnutridas. O procedimento Classificação trata da alocação de objetos ou indivíduos para grupos previamente definidos, por exemplo, considerando que toda gestante com peso/estatura menor ou igual a 33.75(kg/m) deve muito provavelmente ser desnutrida, então, toda gestante que for ao consultório nessa faixa de peso/estatura será considerada desnutrida. A discriminação é mais exploratória, pois busca descrever a diferença nas características dos objetos ou indivíduos, a partir da definição de diversos grupos ou populações conhecidas. A classificação objetiva enquadrar objetos ou indivíduos em um dentre 2 ou mais grupos ou classes, com ênfase na procura de uma regra que possa ser usada para designar otimamente um novo indivíduo ou objeto para as classes ou grupos definidos.

É interessante notar que os novos indivíduos que são alocados pela classificação passam por uma análise segundo uma regra ou modelo, que foi criado a partir de uma discriminação e, à medida que novos indivíduos aparecem para serem classificados, aparecem juntamente novas informações que em média podem alterar o padrão, se for feita nova análise discriminante. Tal fato nos leva a crer que estamos lidando com um modelo dinâmico que, com o passar do tempo, utilizando-se da discriminação em um conjunto aumentado de informações, deve-se corrigir a tendência anterior para uma nova tendência, que passa a ser a nova regra de classificação.

Uma regra usada para discriminação pode em muitos casos servir como regra de classificação, veja figura 4.2.3.1, e inversamente, uma

regra de alocação pode sugerir discriminação. Na prática essas duas funções quase sempre se confundem, e a distinção entre classificação e discriminação não é visível. Neste trabalho ambas serão muitas vezes confundidas. A regra de classificação, ou classificador, pode ser vista como uma função com domínio no espaço das medidas X , e imagem no espaço de categorias da variável dependente, de modo que, com base em suas características, o classificador designa cada indivíduo para uma dessas categorias.

Tanto a classificação como a discriminação não são técnicas incomuns. Desde a tenra idade, quando a criança separa seus brinquedos em tipos e formas, está fazendo uma discriminação; o médico quando, a partir do conhecimento do valor de algumas medidas efetuadas no paciente e de um conjunto de sintomas, identifica um padrão de doenças, obtendo o diagnóstico, está fazendo uma classificação, assim por diante. Uma regra de classificação é criada a partir de um banco de dados históricos, de modo que os dados apresentem similaridades

dentro de uma classe e distinção ou dissimilaridade entre os indivíduos das distintas classes. Assim, com base na analogia entre os indivíduos do banco de dados históricos, pode-se determinar a classe a que um novo indivíduo pertence.

Estudos em que são usadas técnicas de agrupamento a partir de um grupo de objetos, indivíduos ou variáveis, obtendo-se sub-grupos homogêneos internamente, são conhecidos como “Cluster Analysis” ou Análise de Agrupamento. Esta técnica ficou conhecida na literatura a partir do trabalho de SOKAL & SNEATH(1963), sobre princípios de taxonomia numérica. Mas, para se fazer classificações semelhantes a essas com os sub-grupos associados a um nível de uma outra variável, por exemplo a variável adequação do peso do RN, não se estará mais usando apenas a técnica “Cluster Analysis”, pois tem-se uma classificação associada ou dependente da variável adequação do peso do RN.

Pretende-se estudar as técnicas estatísticas que possam se adequar ao problema da discriminação e classificação das gestantes em grupos de adequação do peso do RN, em função de diversas variáveis observadas em acompanhamento pré-natal, de modo que seja possível comparar a eficiência das curvas de Rosso e Siqueira. No interior desse estudo, está a identificação de fatores registrados no pré-natal que possam ser identificados como de risco ao peso do RN, podendo ou não confirmar a importância de alguns como mais discriminantes, que no fundo é o estudo da associação entre um vetor de k variáveis gestacionais e a variável adequação do peso do RN.

Para problemas semelhantes a esses poder-se-ia pensar em técnicas comuns de análise de interrelações, e a literatura tem indicado mais comumente três formas de investigar a associação entre uma variável múltipla e a variável resposta, usando um modelo que indique essa associação:

- no contexto de modelos lineares (regressão linear ou análise de variância) quando a variável resposta for contínua^{24,60},
- no contexto de modelos para dados categóricos como: modelo log-linear para estudo da associação entre variáveis categóricas de qualquer tipo sem o uso de uma das variáveis em estudo como dependente, logística e probito para variáveis respostas binárias^{11,21}, logística para variável resposta multinomial¹,
- em um contexto mais abrangente, juntando as duas formas anteriores em um modelo mais geral, usando modelos lineares generalizados⁴³.

Como pode ser notado pelas citações feitas anteriormente, para pesquisas que buscam discernir a estrutura das ligações entre os fatores envolvidos com o desenvolvimento intra-uterino, o número de variáveis consideradas independentes ou preditoras a serem medidas antes ou durante a gestação pode chegar a ser bem grande. Mas a dimensão do espaço de medidas não necessariamente implica no discernimento dessa estrutura. O conjunto de dados provavelmente será mais rico em informações se tiver dimensão alta, vários tipos de variáveis e número de observações grande.

O modelo log-linear poderia ser usado nesse trabalho, se a variável dependente fosse colocada junto com as variáveis independentes (AGRESTI (1990)), e teria desempenho satisfatório se a dimensão do espaço de medidas fosse pequena, ou existissem informações a priori sobre o conjunto de dados, entre elas a existência de correlação entre variáveis independentes e/ou a existência de variáveis independentes não correlacionadas com a variável dependente. Em caso contrário, o número de parâmetros necessários para uma modelagem é muito grande, crescendo com a dimensão do espaço de medidas.

Atualmente o uso de dados complexos, com alta dimensionalidade e com vários tipos de variáveis, tem aumentado com o uso de computadores, e levado ao desenvolvimento de novas técnicas estatísticas, tais como os procedimentos multivariados de redução da dimensão do espaço de medidas, como por exemplo:

- Regressão por “stepwise”
- Análise Fatorial
- Análise Discriminante

Embora existam obstáculos na aplicação dessas ferramentas multivariadas de redução, elas são respostas à necessidade vigente.

A regressão por “stepwise” deve ser aplicada quando pode-se supor um modelo aditivo, e a variável dependente for quantitativa, o que não ocorre no presente trabalho, a não ser que seja considerada como a variável dependente o peso do RN. A regressão logística é uma opção que pode ser usada de forma “stepwise” considerando a variável dependente categórica.

A análise fatorial busca agrupar as variáveis em estudo, em subgrupos de variáveis mais homogêneas entre si, cada grupo definindo um fator mais geral, não existindo a possibilidade de considerar uma das variáveis como sendo dependente, o que inviabiliza seu uso nesse trabalho.

A alternativa de utilizar a “Análise discriminante” seria adequada, uma vez que essa técnica pode discriminar os indivíduos segundo suas medidas, e criar uma regra de classificação. Mas sua aplicação só é viável se as k variáveis independentes forem numéricas, ou categóricas de modo que possam ser transformadas em numéricas, e que tenham uma distribuição diferente para cada sub-população definida por uma das classes da variável dependente, cujo vetor média seja assintoticamente normal (JOHNSON & WICHERN(1982); CHATFIELD & COLLINS(1980)). Assim, como algumas variáveis independentes em pesquisas

relacionadas com a desnutrição materna, são categóricas, como por exemplo sexo, local do pré-natal, etc., seria necessário criar métodos de modo que se adotasse valores numéricos para as categorias dessas variáveis, de modo que a distribuição do vetor das médias resultantes fosse confiável e, pelo teorema do limite central, fosse considerado assintoticamente normal, o que na prática torna-se muito difícil. Além disso, uma desvantagem adicional é a complexidade de sua interpretação por um não especialista.

Outras opções encontradas na literatura para o presente problema seriam:

- **A técnica AID - “Automatic Interaction Detection”** (MORGAN & SONQUIST(1963); HO(1987))
- **A técnica da “Regressão por Árvore” (CART - Classification and Regression Trees)** (BREIMAN e col.(1984);

Ambas as técnicas fazem divisões do conjunto total de dados em subgrupos como a técnica de discriminação ou agrupamentos, e são chamadas hierárquicas, devido à maneira como os sub-grupos são formados. A classificação escolhe os objetos ou elementos em etapas, produzindo uma sequência de partições, segundo a ordem de influência das suas medidas que serão chamadas de variáveis independentes, digamos $X_1, X_2, \dots, X_k$, que serão representadas por X , sobre a variável dependente Y , resultando em uma árvore de classificação.

4.2.3.1 - O modelo logístico para variável dependente multinomial:

O modelo logístico (AGRESTI(1990)) tem sido usado comumente para o caso em que se procura o ajuste de um modelo linear tipo regressão em que as variáveis independentes podem ser quantitativas ou qualitativas, e a variável dependente é binária.

4.2.3.1.1 - O modelo logístico simples:

Para que seja possível o estudo da função logística com resposta multinomial, considere-se primeiramente a existência de uma variável categórica Y binária, assumindo apenas dois possíveis valores. A proporção de elementos da população, registrados na primeira categoria, considerada de interesse, será denotada por π_1 , que leva à chamada função logito (AGRESTI(1990)):

$$\log \left(\frac{\pi_1}{1 - \pi_1} \right) \quad 4.2.3.1.1$$

Para que seja feito o ajuste de um modelo logístico, considera-se que a proporção de registros referentes à categoria de interesse da variável Y , é dependente das variáveis independentes $\mathbf{X}=(X_1, X_2, \dots, X_k)'$, e é dada pela função logística:

$$\pi(\mathbf{X}) = \frac{\exp(\alpha + \beta\mathbf{X})}{[1 + \exp(\alpha + \beta\mathbf{X})]} \quad 4.2.3.1.2$$

onde α e β são constantes.

O “odds” para a categoria em estudo da variável Y é dado por:

$$\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})} = e^{\alpha + \beta \mathbf{X}} \quad 4.2.3.1.3$$

que leva à regressão logística:

$$\begin{aligned} \log \left[\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})} \right] &= \alpha + \beta \mathbf{X} = \alpha + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \dots + \beta_k \mathbf{X}_k \\ &= \alpha + \sum_{i=1}^k \beta_i \mathbf{X}_i \end{aligned} \quad 4.2.3.1.4$$

4.2.3.1.2 - O modelo logístico generalizado:

Se Y é uma variável dependente categórica com d categorias, existirão $\binom{d}{2}$ pares de categorias, para as quais pode-se construir funções logísticas $\pi_{jh}(\mathbf{X})$, onde j e h seriam as categorias em comparação. Se for fixada h como a última categoria d , de modo que todas outras categorias anteriores a d sejam comparadas apenas com ela, d é chamada de categoria “baseline”, e existiriam $d-1$ funções logito na forma:

$$\log \left[\frac{\pi_j}{\pi_d} \right] \text{ com } j=1,2,\dots, d-1 \quad 4.2.3.1.5$$

Para o presente trabalho, a variável dependente *Adequação do peso do RN* não é dicotômica, pois tem três níveis de categorias, do que se deduz que existem 2 regressões logísticas se uma das categorias for considerada “baseline”.

Usando a notação afeta ao modelo log-linear, considerando duas variáveis independentes, tem-se:

- m_{jit} = a frequência esperada em uma tabela de múltipla entrada, como no apêndice A4.2.3, referente à casela em que Y tem a categoria j , X_1 e X_2 tem as categorias i e t respectivamente.
- π_{jit} = a proporção de um elemento que estão na casela jit
- $\pi_{jit} = \frac{m_{jit}}{N}$

que transforma o modelo 4.2.3.1.5 em:

$$\log \left[\frac{\pi_{jit}}{\pi_{dit}} \right] = \log \left[\frac{m_{jit}}{m_{dit}} \right] = \alpha_j + \beta_{ji}^{x_1} + \beta_{jt}^{x_2} \quad 4.2.3.1.6$$

onde $j = 1, 2, \dots, d-1$; $i =$ um nível de X_1 ; $t =$ um nível de X_2 ; $\alpha_j = (\lambda_j^Y - \lambda_d^Y)$; $\beta_{ji}^{x_1} = (\lambda_{ji}^{X_1 Y} - \lambda_{di}^{X_1 Y})$ e $\beta_{jt}^{x_2} = (\lambda_{jt}^{X_2 Y} - \lambda_{dt}^{X_2 Y})$ da notação de modelos log-lineares (AGRESTI(1990)).

O modelo logito generalizado pode conter não só variáveis independentes categóricas, mas também contínuas e discretas. Considerando que as k variáveis independentes $\mathbf{X} = (X_1, X_2, \dots, X_k)'$ assumem em conjunto o r -ésimo valor, digamos $\mathbf{X}_r = (1, X_{r1}, X_{r2}, \dots, X_{rk})'$, a probabilidade que esse conjunto de valores $\mathbf{X}_r$ tenha o nível j da variável dependente é dada por $\pi_j(\mathbf{X}_r)$.

Em termos de modelos lineares, a função logística generalizada é dada por:

$$\pi_j(\mathbf{X}_r) = \frac{\exp(\beta_j' \mathbf{X}_r)}{\sum_{h=1}^d \exp(\beta_h' \mathbf{X}_r)} \quad 4.2.3.1.7$$

que leva ao modelo

$$\log \left[\frac{\pi_j(\mathbf{X}_r)}{\pi_d(\mathbf{X}_r)} \right] = \log \left[\frac{m_{jr}}{m_{dr}} \right] = \beta' X_r, \quad 4.2.3.1.8$$

onde $j=1,2,\dots, d-1$, $\beta' = \beta'_j - \beta'_d$, que se refere à comparação do j -ésimo nível de $\mathbf{Y}$, com o seu nível d , sendo $\beta'_j = (\alpha_j, \beta_1, \beta_2, \dots, \beta_k)$

4.2.3.1.3 - A função logito acumulada:

Desde que neste trabalho a variável dependente está sendo considerada ordinal, uma forma indicada por Agresti, de construir o modelo 4.2.3.1.8 é partindo o espaço $\mathbf{Y}$ em $d-1$ formas, de modo que se leve em conta a ordem das categorias, e sejam feitos agrupamentos dos níveis das categorias adjacentes na escala ordinal, resultando então na construção de $d-1$ logitos. Por exemplo, supondo que d seja igual a 4, com os níveis y_1, y_2, y_3, y_4 , os $4-1=3$ logitos virão das partições: $(y_1 | y_2, y_3, y_4), (y_1, y_2 | y_3, y_4), (y_1, y_2, y_3 | y_4)$. Em cada uma dessas 3 partições existem sempre duas regiões formadas por grupos de níveis, à esquerda do ponto de corte e à direita do ponto de corte.

Cada uma dessas $d-1$ partições geram um logito que deve ser construído considerando a região da esquerda como se fosse um nível do logito binário, e a região da direita como se fosse o outro nível do logito binário. Para isso, acumulam-se as probabilidades dos níveis de categorias de cada região, da seguinte forma:

- Para a região da esquerda, considerando c como sendo o ponto de corte da sequência de níveis de categorias de $\mathbf{Y}$, a probabilidade acumulada dessa região é dada por: $F_c(\mathbf{X}) = \pi_1(X) + \pi_2(X) + \dots + \pi_c(X)$ onde $c=1,2,\dots,d-1$

- Para a região da direita a probabilidade acumulada dessa região é dada por:
 $1 - F_c(X)$.

Considerando essas duas regiões, a função logito acumulada é definida como uma função logito simples de variável resposta binária, onde as categorias de 1 até c representam uma categoria simples, e as categorias $c+1$ até d representam a segunda categoria simples, como segue:

$$\begin{aligned}
 L_c &= \log \left(\frac{F_c(X)}{1 - F_c(X)} \right) = \\
 &= \log \left(\frac{\sum_{j=1}^c \pi_j(X)}{1 - \sum_{j=1}^c \pi_j(X)} \right) = \\
 &= \log \left(\frac{\sum_{j=1}^c \pi_j(X)}{\sum_{j=c+1}^d \pi_j(X)} \right) \quad \text{com } j=1,2,\dots, d-1 \quad 4.2.3.1.9
 \end{aligned}$$

Nota-se que para $d=2$, temos $L_c=L_1$, que é o logito simples 4.2.3.1.1.

Como em 4.2.3.1.4, para cada logito em 4.2.3.1.9 tem-se o modelo de regressão logística acumulado dado por:

$$L_c(X) = \alpha_c + \beta' X \quad \text{com } c=1,2,\dots,d-1 \quad 4.2.3.1.10$$

Como no presente caso a variável dependente possui $d=3$, ter-se-á dois modelos do tipo 4.2.3.1.10, isto é, $L_1(X)$ e $L_2(X)$, onde o

coeficiente angular β é o mesmo nos dois modelos, do que se conclui que esses modelos são paralelos, diferenciando apenas na intersecção da ordenada Y , e que a diferença entre $L_1(X)$ e $L_2(X)$ é devida à contribuição da mudança de categoria na variável dependente.

Esses dois modelos ficarão na forma:

$$\bullet L_1 = \log\left(\frac{\pi_1}{\pi_2 + \pi_3}\right) \text{ ou } \log\left(\frac{m_{i1}}{m_{i2} + m_{i3}}\right) = \alpha_1 + \beta'X \quad \text{para } c = 1$$

4.2.3.1.11

$$\bullet L_2 = \log\left(\frac{\pi_1 + \pi_2}{\pi_3}\right) \text{ ou } \log\left(\frac{m_{i1} + m_{i2}}{m_{i3}}\right) = \alpha_2 + \beta'X \quad \text{para } c = 2$$

Poder-se-ia pensar na função logística 4.2.3.1.4 genericamente, considerando-se $L_c(X) = \mathbf{X}\beta_c$ de onde tem-se $L_1(X) - L_2(X) = \mathbf{X}\beta$, que depende da mudança de nível de Y e de X . A matriz $\mathbf{X}$ pode indicar as variáveis independentes X como também as interações entre as variáveis independentes, como X_j e X_k , pelo parâmetro β_{jk} de β .

O SAS ajusta o modelo de regressão logística 4.2.3.1.11 pelo método de máxima verossimilhança, usando um algoritmo iterativo de mínimos quadrados ponderados, e testa o paralelismo usando a estatística $-2\log(\text{Razão de Verossimilhança})$. Com pequenas mudanças nos comandos do PROC LOGIST do SAS é possível fazer o teste do paralelismo, medindo a contribuição da mudança de nível da variável dependente. No apêndice A4.3 é apresentado o programa construído, usando o SAS, para ajustar o modelo logístico acumulado com variável dependente múltipla.

4.2.3.2 - A técnica AID:

A técnica AID procura classificar os indivíduos, usando as variáveis independentes segundo a ordem de influência sobre a variável dependente Y , com uma idéia de encontrar um modelo do tipo linear que explique essa influência. Tal classificação começa na partição do conjunto de dados originais em dois subgrupos, chamada partição binária, analisando cada uma das k variáveis independentes de tal forma que sejam formados grupos de combinação de níveis ou categorias, preservando a ordem natural dessas categorias, de modo que sejam o mais homogêneos internamente e heterogêneos entre si. Na literatura relativa à técnica AID, a regra de partição preservando a ordem natural das categorias é chamada **regra monotônica**, enquanto a regra desconsiderando quaisquer ordens como as escalas puramente nominais é chamada **livre**, e a regra de partição em que originalmente as categorias estão em uma escala ordinal, mas existe pelos menos uma exceção devida à forma de registro em que uma ou mais categorias não podem ser colocadas em uma posição juntamente com as demais na escala ordinal, como é o caso de medidas “missing”, perdidas ou não registradas, ou ainda as categorias “outros” ou “não sabe” é chamada **flutuante**.

A variável independente que apresentar grupos de categorias com maior diferença na variável dependente, é escolhida para explicar a variação dessa variável.

O estudo dos níveis ou categorias de cada uma das variáveis independentes, objetivando o agrupamento, é feito usando medidas de similaridades ou dissimilaridades entre grupos, com idéias semelhantes à análise de variância, quando se estuda a variação entre grupos e dentro dos grupos. Busca-se

então inicialmente a variável independente X_i que internamente tem combinações ou junções de níveis de categorias, que indiquem maior mudança na variável dependente Y , fazendo então uma divisão dicotômica dos dados segundo esses grupos de categorias da variável X_i , digamos C_1, C_2 . Para procurar o agrupamento de níveis mais discriminantes, poder-se-ia pensar por exemplo como medida de dissimilaridade a soma de quadrados da variável resposta entre os grupos. Assim, para cada variável independente X_i ($i=1,2,3,\dots,k$, onde k =número de variáveis independentes) procura-se fazer todos possíveis agrupamentos de níveis, buscando aquele que produza a maior “distância” entre as respostas desses grupos.

Por exemplo, supondo a variável Y como sendo quantitativa, procura-se o agrupamento que produza a maior Soma dos Quadrados Entre,

$$SQ_E = N_1(\bar{y}_1 - \bar{y})^2 + N_g(\bar{y}_g - \bar{y})^2 \quad 4.2.3.2.1$$

A variável independente que tiver a maior $SQ_E(x_j)$ para um dado agrupamento seria a discriminante.

Feito isso, dentro de cada um dos sub-grupos G_1 e G_2 , criados a partir das junções de categorias da primeira variável mais significativa, o método estuda as restantes $k-1$ variáveis individualmente, voltando ao processo anterior de analisar os possíveis agrupamentos de níveis, procurando aquele que indique maior mudança na resposta, assim sucessivamente. Esquemáticamente o resultado desses agrupamentos e discriminação é uma árvore ou dendrograma, iniciando o tronco na primeira partição, ficando um ramo em cada agrupamento de categorias, como é visto na figura 4.2.3.2.1.

Essa árvore tem dois propósitos, descobrir uma estrutura preditiva entre a variável dependente e as variáveis independentes, ou classificar um indivíduo com base nas suas medidas em uma categoria da variável dependente.

Para classificar um novo indivíduo, recorre-se a um classificador, que trata de um conjunto de perguntas, definidas em cada nó de uma árvore já construída por dados históricos; sendo tais perguntas criadas pela partição feita nesse nó, em função da variável escolhida. A primeira pergunta é feita no nó inicial, chamado raiz. Se a resposta é afirmativa, o processo continua no ramo da esquerda. Se for negativa, continua no ramo da direita. O processo todo de perguntas e respostas termina quando se alcança um dos nós terminais. Como cada nó terminal define um grupo de indivíduos homogêneos em função da variável dependente, o processo de classificação de um novo indivíduo termina quando ele é alocado em função da variável dependente referente ao nó terminal encontrado. Por outro lado, se a variável dependente for qualitativa, o processo de classificação ocorreria em um nó, quando o indivíduo fosse alocado como pertencendo ao nível da variável dependente mais provável naquele nó.

A literatura define o conjunto de dados “históricos”, usado para construir a árvore, como *Amostra Informada*(ROJAS(1987)) ou *Amostra de Aprendizagem*(BREIMAN(1984)) que receberá a notação $\mathcal{L}$.

Se x é uma variável categórica, com digamos c níveis, então tem-se um total de $2^{c-1} - 1$ possíveis partições binárias, como por exemplo no caso da variável estado nutricional da mãe pelos métodos de Rosso e de Siqueira, que tem três níveis de registros($c=3$), a saber: DES(desnutrida), NOR(normal) e OBS(obesa), e as possíveis partições são($2^{3-1}-1=3$): (DES | NOR,OBS), (DES,NOR | OBS) e (DES,OBS | NOR). Por outro lado, se X_i for categórica ordinal com c níveis ou numérica com c valores distintos, então tem-se $c-1$ partições, de forma que não sejam sobrepostas e nem fora de ordem. Tal tipo de partição permite transformações invariantes e monotônicas de X_i .

Como em modelos lineares, no contexto de classificação por árvore, é necessário medir a bondade do ajuste, no sentido de procurar a árvore que melhor represente o fenômeno em questão, que no caso se refere à capacidade de alocar os indivíduos para suas verdadeiras categorias.

Definição 4.2.3.2.1:

Considera-se que, na realidade, a amostra é dividida em duas partes. A primeira parte é a já mencionada amostra informada $\mathcal{L}$, com a qual se constroi a árvore, e a segunda parte é a amostra em teste, que receberá a notação $\mathcal{X}$, cujos elementos serão classificados segundo a árvore criada por $\mathcal{L}$. A probabilidade de classificar erradamente um elemento de $\mathcal{X}$ é chamada de *Taxa de Má-classificação*, e é usada para quantificar a bondade da árvore, quanto à capacidade de classificação. Assim para uma dada árvore $\mathcal{A}$, a *taxa de má-classificação* é dada por:

$$R(\mathcal{A}) = P(\mathcal{A}(\mathcal{X}) \neq \mathcal{Y}), \quad 4.2.3.2.2$$

onde $\mathcal{A}(\mathcal{X})$ representa a classificação de $\mathcal{X}$ através da árvore $\mathcal{A}$, e $\mathcal{Y}$ representa valores da variável dependente Y .

Definição 4.2.3.2.2:

Para o cálculo de $R(\mathcal{A})$ exatamente da forma que foi definida em 4.2.3.2.2, são necessárias duas amostras, o que muitas vezes não é possível. Na prática apenas $\mathcal{L}$ é obtida, e usada tanto para construir $\mathcal{A}$, como para fazer a classificação $\mathcal{A}(\mathcal{X})$, e para estimar $R(\mathcal{A})$, motivo pelo qual são propostas três formas de estimação dessa taxa, como veremos a seguir:

1. Estimador Por Ressubstituição:

Trata-se do estimador mais comum e menos acurado (BREIMAN e col(1984)). Neste caso a árvore é construída usando as observações de $\mathcal{L}$, e as mesmas observações são classificadas através dessa árvore. Para essa forma de classificação, $\mathcal{L}$ faz o papel da amostra $\mathcal{X}$. A proporção de casos mal classificados é o *estimador por ressubstituição* da taxa de má-classificação, como segue:

$$R^R(\mathcal{A}) = \frac{\text{número de elementos de } \mathcal{L} \text{ mal classificados}}{\text{número de elementos de } \mathcal{L}} \quad 4.2.3.2.3$$

Essa medida tem a desvantagem de se classificar justamente os mesmos elementos que construíram a árvore, que no caso do uso de todos os ramos da árvore logicamente leva a uma taxa de má-classificação zero. Algumas vezes a árvore é muito grande como modelo, então podem-se os ramos de modo que seja facilitada sua interpretação. Nesse caso $R^R(\mathcal{A})$ não deve alcançar o valor zero, porque alguns dos ramos podados deixam a classificação obscura.

NETTER e WASSERMAN (1974) mostram que a regressão clássica utiliza essa medida.

2. Estimador por Amostra Teste:

Neste caso considera-se que a amostra informada $\mathcal{L}$ é dividida em duas amostras, digamos $\mathcal{L}_1$ para construir a árvore de classificação, e $\mathcal{L}_2$ para estimar $R(\mathcal{A})$. O estimador por amostra teste é dado por:

$$R^{AT}(\mathcal{A}) = \frac{\text{número de elementos de } \mathcal{L}_2 \text{ mal classificados}}{\text{número de elementos de } \mathcal{L}_2} \quad 4.2.3.2.4$$

Para a construção do estimador 4.2.3.2.4, $\mathcal{L}_2$ deve ser independente de $\mathcal{L}_1$, mas da mesma distribuição. Por isso o procedimento mais comum é tirar $\mathcal{L}_2$ de $\mathcal{L}$ aleatoriamente. Geralmente $\mathcal{L}_2$ tem 1/3 dos casos de $\mathcal{L}$. Essa medida tem a vantagem sobre a anterior(4.2.3.2.3), com relação ao fato de se ter outro conjunto de dados para serem classificados. Uma restrição a esse estimador é quando o tamanho de $\mathcal{L}$ for pequeno, pois teríamos por exemplo 2/3 de $\mathcal{L}$ para construir a árvore e apenas 1/3 para estimar $R^{AT}(\mathcal{A})$.

3. Estimação por Validação Cruzada:

Neste caso a amostra informada, $\mathcal{L}$, é dividida aleatoriamente em v sub-amostras, digamos $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \dots, \mathcal{L}_v$ de modo que sejam construídas v árvores, da seguinte maneira: constrói-se a árvore $\mathcal{A}_1$ usando a sub-amostra $\mathcal{L} - \mathcal{L}_1$

constrói-se a árvore $\mathcal{A}_2$ usando a sub-amostra $\mathcal{L}-\mathcal{L}_2$

assim por diante, até construir a árvore $\mathcal{A}_v$ usando a sub-amostra $\mathcal{L}-\mathcal{L}_v$.

Então, para cada árvore $\mathcal{A}_j$ construída, estima-se a taxa de má-classificação, como se fosse por amostra teste, da seguinte forma:

$$R^{AT}(\mathcal{A})_j = \frac{\text{número de elementos de } \mathcal{L}_j \text{ mal classificados}}{\text{número de elementos de } \mathcal{L}_j} \quad 4.2.3.2.5$$

onde $j=1,2,\dots,v$.

Finalmente o estimador da taxa de má-classificação por Validação cruzada é dado pela média entre as taxas das sub-amostras:

$$R^{CV}(\mathcal{A}) = \sum_j \frac{R^{AT}(\mathcal{A})_j}{v} = \sum_j \left(\frac{\text{n}^\circ \text{ erros em } \mathcal{L}_j}{v \times n(\mathcal{L}_j)} \right) \quad j=1,2,\dots,v \quad 4.2.3.2.6$$

Nem sempre é possível dividir $\mathcal{L}$ em sub-amostras de tamanhos iguais, mas procura-se ter grupos de tamanhos bem semelhantes ou usam-se artifícios como no presente trabalho em que o tamanho amostral é 565, e as sub-amostras serão de tamanho 56, colocando-se os 5 elementos restantes no complementar de $\mathcal{L}_j$ para a construção da árvore.

A maior dificuldade no uso do estimador por validação cruzada 4.2.3.2.6, é computacional uma vez que, para sua utilização, são necessárias várias partições aleatórias de $\mathcal{L}$, várias construções das respectivas árvores e vários cálculos das respectivas taxas de má-classificação por amostra teste. BREIMAN e col.(1984) afirmam que uma vantagem desse procedimento é a

estabilidade, no sentido que, as reais taxas de má-classificação para cada sub-amostra $R(\mathcal{A})_j$, estão bem próximas da taxa real $R(\mathcal{A})$.

A validação cruzada usa todos os dados uniformemente, pois todos os casos em $\mathcal{L}$ são usados para construir a árvore, e todos os casos são usados exatamente uma vez para construir o estimador por amostra teste. A validação cruzada de tamanho $v = 10$ tem sido usada com sucesso e, em estudos com simulação de dados, tem-se obtido resultados satisfatoriamente próximos de $R(\mathcal{A})$ (BREIMAN(1984)).

4.2.3.3 - A técnica CHAID:

Como a técnica AID é indicada quando a variável dependente é quantitativa, quando a variável dependente for categórica como no caso da adequação do peso do RN, essa técnica não pode mais ser usada. Uma extensão da técnica AID é a técnica CHAID (Chi-Square Automatic Iterative Detection) proposta por Kass³⁹, que no lugar de procurar a SQE máxima, como critério ótimo de seleção do agrupamento de categorias, utiliza a estatística χ^2 de Pearson $\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ ou a Razão de Verossimilhança χ^2 : $2 \sum_{ij} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}} \right)$. Para cada variável independente estudam-se todas as partições e agrupamentos do grupo de categorias em sub-grupos, procurando aquela que tem o maior valor da estatística χ^2 em uma tabela de dupla entrada com a variável dependente.

Se a variável dependente for nominal, a CHAID estará testando a independência entre **X** e **Y** contra um modelo log-linear saturado do tipo: $\ln(E_{ij}) = \lambda + \lambda_i^x + \lambda_j^y + \lambda_{ij}^{xy}$, onde λ é o parâmetro referente a um valor geral, λ_i^x se refere ao efeito apenas da variável **X** sobre a casela ij , e λ_j^y se refere ao efeito apenas da variável **Y** sobre a casela ij , e λ_{ij}^{xy} se refere ao efeito da interação entre as variáveis **X** e **Y** naquela casela. Por outro lado no caso em que a variável dependente é ordinal numérica, a CHAID estará testando a independência entre **X** e **Y** contra um modelo log-linear não saturado do tipo: $\ln(E_{ij}) = \lambda + \lambda_i^x + \lambda_j^y + \theta_i(y_j - \bar{y})$, onde y_j é uma categoria de **Y**, e θ_i um coeficiente desconhecido de y_j , para uma dada categoria **i** de **X**.

Diferentemente da técnica AID, a CHAID permite uma partição múltipla das categorias. A variável que for escolhida é a discriminante, pois o conjunto de dados é particionado em sub-conjuntos de acordo com o agrupamento de categorias dessa variável. Uma nova análise é feita dentro de cada sub-grupo de categoria, estudando-se a influência das restantes $k-1$ variáveis independentes, utilizando-se novamente a estatística χ^2 .

Além disso a técnica CHAID pode trabalhar preservando a ordem natural das categorias das variáveis, como faz a AID, ou sem ordem nas categorias das variáveis, como é o caso da partição livre, ou pode trabalhar com o caso misto da regra de partição flutuante.

Como a técnica AID, a CHAID trabalha em passos, como se vê a seguir:

- 1°.- Para cada variável independente é achada a melhor partição.
- 2°.- Todas as variáveis independentes são comparadas segundo sua melhor partição e a melhor é escolhida.
- 3°.- Cada sub-grupo resultante da divisão da variável independente escolhida é reanalisado independentemente, buscando produzir outra análise e subdivisão.

O tipo de variável independente determina a forma do agrupamento a ser feito nas categorias, para a construção da tabela de contingência, de modo que o valor da estatística χ^2 seja o mais alto possível. Isso implica que o número de observações deve ser grande o bastante, de modo que o valor da estatística não seja contestado.

4.2.3.3.1 - O algoritmo da técnica CHAID:

Um problema nessa análise é o de reduzir a tabela de contingência $c \times d$, onde c =número de categorias da variável independente e d =número de categorias da variável dependente, $d \geq 2$ e $c \geq 2$, para uma tabela mais significativa $j \times d$, com o maior χ^2 , sendo j uma combinação das c categorias.

A proposta de Kass foi a de procurar a melhor tabela de uma maneira passo-a-passo ("stepwise"), como é feito em regressão (DRAPER & SMITH(1981)).

O algoritmo consta de:

1º passo: Para cada variável independente, construir uma tabela de dupla entrada das categorias dessa variável com as categorias da variável dependente. Passando então para os passos 2 e 3.

2º passo: Achar o par de categorias da variável independente, cuja sub-tabela $2 \times d$ é a menos significativa, calculando a estatística χ^2 para esse par de categoria e a variável dependente. Para cada estatística χ^2 , calcular o respectivo nível descritivo p . Para aqueles pares com p maior que o valor crítico determinado α , tomar o par que tiver o menor valor χ^2 e agrupar as duas categorias em uma, considerando como uma única categoria composta. Repetir esse passo para todas categorias existentes.

3º passo: Para cada categoria composta pelo passo anterior, que seja resultado da junção de três ou mais categorias originais, ver

se a junção pode ser considerada significativa pelo valor da estatística χ^2 e fazer partições binárias, procurando a partição dentro da categoria composta que seja mais significativa. Se a estatística estiver além de um valor crítico, oficializar a partição, e retornar para o passo 2. Se mais que uma partição é possível, tomar a partição que tiver maior χ^2 .

4º passo: Para cada variável independente, com as categorias resultantes dos passos anteriores, calcular o nível descritivo **p**, separando a variável mais significativa. Se o nível descritivo for menor que um valor crítico definido, subdividir o conjunto de dados de acordo com as categorias finais dessa variável, se o nível descritivo for maior juntar as categorias. Para as categorias que tiverem poucas observações (menor que o mínimo especificado) juntar com a categoria mais similar. Calcular o ajuste de Bonferroni para a categoria composta.

5º passo: Para cada agrupamento de dados resultante do passo 4, voltar ao passo 1.

A parada ocorre quando todas as variáveis independentes forem analisadas e reanalisadas ou o número de indivíduos é pequeno dentro do sub-grupo.

Os passos 2 e 3 contam com a partição da estatística χ^2 total, referente à tabela inicial, em dois termos, sendo X_1 referente à sub-tabela $2 \times v$, que está sendo analisada para ver se permanece ou não na tabela inicial, e outro termo X_2 referente à tabela resultante da junção de categorias, tal que χ^2 , X_1 , e X_2 ,

têm distribuição assintótica qui-quadrado com $(d-1) \times (c-1)$, $d-1$, $(j-1) \times (d-1)$ graus de liberdade respectivamente. A estatística X_1 é tratada como tendo a distribuição qui-quadrado com $d-1$ graus de liberdade, da mesma forma que a estatística F é usada em regressão por “stepwise” para ajudar a decidir se uma variável deve ou não ser incluída no modelo. O passo 4 requer vários testes χ^2 para as tabelas reduzidas, condicional ao número de categorias resultantes. KASS(1980) apresenta os valores do multiplicador de Bonferroni para os limites de significância dos testes.

No apêndice A4.4 é apresentada a forma de aplicação do algoritmo da CHAID. O pacote estatístico SPSS, versão 6.0 (1995), possui um módulo chamado CHAID, que faz essa análise e apresenta o dendrograma em forma de árvore como do exemplo da figura 4.2.3.3.1, onde o tipo de parto está sendo prognosticado por peso na última consulta e altura da gestante, para uma amostra de 477 pacientes. Trata-se de uma relação entre a variável dependente e as variáveis independentes, como uma regressão por “stepwise”. Os ramos da árvore terminam em nós chamados terminais, que definem um sub-grupo de indivíduos classificados em um dos níveis da variável dependente. Essa classificação tem sido usada na literatura da seguinte forma: se o nó terminal tem as proporções de seus indivíduos nas categorias da variável dependente de tal forma que uma das categorias é maior que as outras, digamos a categoria j , então os indivíduos desse nó são classificados como pertencendo à categoria j .

O valor do nível de significância a ser fixado para comparação de todos os testes das respectivas tabelas de contingência, é para o CHAID por “default” $p=0.05$, e pode ser alterado, dependendo da exatidão requerida na escolha das variáveis que deverão definir os nós e ramos. No fundo o valor p , representa a probabilidade da relação encontrada na amostra estudada entre a variável dependente e a variável independente em estudo, ser ao acaso, quando na realidade essas duas variáveis são independentes. Assim, a variável independente

que tiver o menor p é aquela que é menos provável que seja não associada com a variável dependente.

FIGURA 4.2.3.3.1: Árvore da CHAID com PARTO como variável dependente

OBS: V20=tipo de parto, V10=peso na última consulta, V6=tempo de gestação na primeira consulta
Parâmetros: Merge Level=0.05, Eligibility Level=0.05, before=after=5

Além disso o pacote CHAID permite fixar o tamanho mínimo dos grupos, fornece as tabelas antes e depois de aglutinar as categorias, com os respectivos χ^2 e p , permite dar pesos, “scores”, às categorias da variável dependente, trabalha com variáveis dependentes ordinais ou nominais, permite podar ramos indesejáveis, fornece uma tabela de “força” de cada ramificação chamada “Gains Chart” como da tabela 4.2.3.3.1.

TABELA 4.2.3.3.1: Quadro dos detalhes dos ramos da árvore ajustada na

figura 4.2.3.3.1

Id	size	% of all	score	index	Cum: size	% of all	score	index
7	30	6.3	76.67	193	30	6.3	76.67	193
6	73	15.3	53.42	135	103	21.6	60.19	152
3	46	9.6	39.13	99	149	31.2	53.69	136
2	168	35.2	36.90	93	317	66.5	44.79	113
4	84	17.6	34.52	87	401	84.1	42.64	108
5	28	5.9	25.00	63	429	89.9	41.49	105
1	48	10.1	22.92	58	477	100	39.62	100

Cada linha da tabela 4.2.3.3.1 representa um ramo da árvore da figura 4.2.3.3.1, a coluna “**size**” informa o número de indivíduos que pertencem ao respectivo ramo, a coluna “**% of all**” registra quanto significa cada ramo em termos de porcentagem do total da amostra, a coluna “**score**” informa a “força” do ramo em prognosticar cesariana, estando em ordem decrescente, pois nesse exemplo o ramo 7, ($V10 \leq 72$ e $V6 \geq 27$), é o mais representativo, com 76.67% de cesariana. A coluna “**index**” representa a proporção entre a porcentagem de cesarianas no nó terminal daquele ramo e a porcentagem de cesarianas no nó inicial, chamado raiz. As colunas “**Cum:size**” em diante constituem medidas acumuladas referentes à junção passo-a-passo das ramificações mais importantes. Além dessas o CHAID permite muitas outras formas de análise via dendrograma em forma de árvore.

O CHAID admite ainda o uso de variáveis quantitativas, tanto para a variável dependente como para as variáveis independentes. A única restrição é que, além de se considerar a variável dependente como ordinal, cada variável deve ser transformada com seus valores agrupados em intervalos como se fossem categorias, em um total máximo de 31 categorias.

Por exemplo, procurou-se identificar as variáveis independentes que estivessem mais relacionadas com o Peso do RN. Para que o

CHAID pudesse ser usado para esse mesmo fim, as variáveis independentes contínuas, com mais que 31 valores distintos, foram transformadas como apresentados na seção 4.1

A partir dessas transformações, foi permitido o uso do método CHAID, obtendo-se a árvore da figura 4.2.3.3.2:

FIGURA 4.2.3.3.2: Árvore da CHAID com PESO DO RN como variável

OBS: V47=Peso do RN/100, V16=tempo de gestação em semanas e v10=peso na última consulta
Parâmetros: Merge Level=0.05, Eligibility Level=0.05, before=after=5

Para a figura 4.2.3.3.2, tem-se 7 ramos identificados por duas partições do conjunto de dados segundo duas variáveis: V16 e V10. A ramificação nº.7, com tempo de gestação acima de 35 semanas e peso medido na última consulta acima de 68 kg, representa 60 gestantes nessa situação, e é considerada pela tabela 4.2.3.3.2 a seguir, como a de maior média de V47, ou seja a que tem a maior média do peso do recém-nascido, V47=37.98 ou V17=3798gramas. A ramificação nº.6 com tempo de gestação acima de 35 semanas e peso medido na

última consulta até 68 kg, representa 70 gestantes nessa situação, e tem a segunda média de peso do RN, com $V47=36.61$ ou $V17=3661$ gramas. A ramificação nº.1 é a que tem a menor média do Peso do RN, ou seja 2418 gramas.

TABELA 4.2.3.3.2: Quadro dos detalhes dos ramos da árvore ajustada na figura

4.2.3.3.2

Id	size	% of all	score	index	Cum: size	% of all	score	index
7	60	12.6	37.98	120	60	12.6	37.98	120
6	70	14.7	36.61	116	130	27.3	37.25	118
5	66	13.8	33.48	106	196	41.1	35.98	114
4	88	18.4	31.48	100	284	59.5	34.58	109
3	72	15.1	29.56	93	356	74.6	33.57	106
2	60	12.6	27.58	87	416	87.2	32.70	103
1	61	12.8	24.18	76	477	100.0	31.61	100

As informações da tabela 4.2.3.3.2 são equivalentes as informações da tabela 4.2.3.3.1. A coluna Id, indica o número do ramo, da árvore da figura 4.2.3.3.2. A coluna size, indica o número de indivíduos do ramo; a coluna % of all, indica o número de indivíduos do ramo com relação ao total; a coluna score, indica a média da variável resposta, peso do RN, naquele ramo. A coluna index, é a razão entre os scores do ramo e o nó raiz. A partir da sexta coluna tem-se as medidas anteriores acumuladas.

4.2.3.4 - A técnica CART:

A técnica CART, Classification and Regression Trees (BREIMAN(1984)), consiste em construir a árvore de classificação dividindo o espaço de medidas sucessivamente por partições binárias em sub-conjuntos cada vez menores, de modo que cada nó terminal represente uma categoria de y . Nesse sentido procura-se a melhor forma de fazer a partição do conjunto de dados para cada variável independente x . O “melhor” pode ser pensado como a partição que provoca a maior redução da variância da variável dependente, no nó t . Depois que cada x é usado para encontrar a sua melhor partição em t , a melhor partição dentre essas melhores partições é usada para dividir o conjunto de dados que estão em t , em dois grupos. A idéia principal é seleccionar a forma de partir cada conjunto de dados de tal forma que os sub-conjuntos descendentes sejam mais “puros”, ou homogêneos, que o sub-conjunto anterior. Para isso é necessária uma medida de quantificação da homogeneidade dentro de cada nó, com relação às categorias da variável dependente. Para a classificação por árvores essa quantificação é feita por intermédio de uma **Medida de Impureza**, que é definida por:

Definição 4.2.3.4.1

Considere-se um nó t , onde os dados estão distribuídos segundo as v categorias da variável dependente, e seja $p(j|t)$ a proporção de elementos do nó t que pertencem à categoria j da variável dependente, com $j=1,2,...,v$. A **Medida de Impureza** para o nó t é uma função de todas as proporções das v classes da variável dependente, do tipo $\varphi(t) = \varphi[p(1|t), p(2|t), \dots, p(v|t)]$, de tal forma que:

a) $\varphi[p(1|t), p(2|t), \dots, p(v|t)] \geq 0$

b) se todas as categorias dentro do nó t forem igualmente misturadas, ou seja, $p(j|t) = 1/v$ para todo $j=1,2,\dots,v$, a medida de impureza $\varphi[1/v, 1/v, \dots, 1/v]$ alcança o valor máximo. Nesse caso considera-se que o nó é totalmente impuro, pois para um indivíduo que estiver nesse nó, não existe possibilidade de classificá-lo para uma das categorias da variável dependente.

c) se apenas uma categoria for representada no nó t , a medida de impureza será mínima, ou seja: $\varphi[1,0,0,\dots,0] = \varphi[0,1,0,0,\dots,0] = \varphi[0,0,\dots,1] = 0$. Nesse caso, todos os indivíduos são classificados para apenas uma determinada categoria da variável dependente, então não será mais necessária uma nova partição, e temos um nó terminal.

Quanto maior for a impureza dentro do nó, maior é a variabilidade interna, e aumenta a necessidade de se fazer uma partição do respectivo conjunto de indivíduos.

Para se procurar a partição que seja a melhor, no sentido de gerar nós reduzindo a impureza com relação ao nó original, seleciona-se a partição que tenha a “bondade” máxima, que pode ser definida como segue:

Definição 4.2.3.4.2

Considere um nó t e uma partição S a ser feita em t , de modo que os nós resultantes são: o nó direito t_D e o nó esquerdo t_E . As proporções de elementos que saem de t e vão para cada um desses nós, são respectivamente p_D e p_E , como vemos na figura 4.2.3.4.1.

Então a bondade da partição $\mathcal{S}$ é avaliada pelo decréscimo ocorrido na impureza devido à partição do nó t nos nós t_D e t_E , dada por:

$$\Delta_i(\mathcal{S}, t) = \varphi(t) - [p_E \cdot \varphi(t_E) + p_D \cdot \varphi(t_D)] \quad 4.2.3.4.1$$

onde $\varphi(t)$ é a impureza do nó t , $\varphi(t_E)$ e $\varphi(t_D)$ são as impurezas dos nós descendentes do nó t .

Além do citado no item c) da definição 4.2.3.4.1, um nó é considerado terminal quando o decréscimo na impureza $\Delta_i(\mathcal{S}, t)$ for suficientemente pequeno. Um valor pequeno para $\Delta_i(\mathcal{S}, t)$ o suficiente para não continuar o crescimento da árvore deve ser do tipo:

$$\Delta_i(\mathcal{S}, t) < \beta, \quad 4.2.3.4.2$$

onde $\beta > 0$ é uma constante a ser determinada.

Desde que o tamanho da árvore está associado à capacidade de classificação, dependendo do valor de β temos uma árvore grande ou pequena, pois se β for pequeno a árvore tenderá a ter um tamanho maior do que se β for grande, e respectivamente alta ou baixa capacidade de classificação. Breiman e col⁹ consideraram além do valor de β como parâmetro definidor do tamanho dos ramos, o custo-complexidade em que são atribuídos pesos ou custos para cada tipo de má-classificação. Por exemplo: o diagnóstico de uma gestante como normal sendo que é desnutrida tem um custo biológico e social maior que o da mãe que é diagnosticada como normal sendo obesa, simbolicamente $C(\text{normal/desnutrida}) > C(\text{normal/obesa})$.

BREIMAN e col.(1984) indicam alguns índices de impureza $\varphi(t)$, mostrando exemplos em que o índice de Gini:

$$\varphi(t) = \sum_{j \neq i}^d p(j|t)p(i|t) \quad 4.2.3.4.3$$

apresenta vantagens, principalmente pela simplicidade, facilidade computacional e escolha de melhores partições. Por esse motivo, neste trabalho consideraremos a aplicação de um pacote computacional com o CART, aplicando o índice de Gini. Diversos autores(BREIMAN e col.(1984), HO,L.L(1987), LIGHT & MARGOLIN(1971)) comentam o fato que o índice de Gini tem a interpretação interessante de ser uma medida de variabilidade dentro do nó, proporcional à Soma dos Quadrados Totais como em análise de variância, pois pode ser escrito como :

$$\varphi(t) = 1 - \sum_j p^2(j|t)$$

No caso de se considerar o custo-complexidade, digamos o custo de diagnóstico na classe i considerando que a verdadeira classe é a j , $C(i/j)$, temos o índice de Gini dado por:

$$\varphi(t) = \sum_{j \neq i} c(i/j) p(j|t) p(i|t) \quad 4.2.3.4.4$$

A técnica CART mede a bondade da árvore, usando a taxa de má-classificação $R(\mathcal{A})$, estimando pelos 3 métodos comentados na secção 4.2.4.1. Neste trabalho, será utilizado um programa construído por Iglesias^{58,30}, em um estudo inicial de tendências e variações de bolsas de valores.

Nesse programa, o usuário escolhe o tipo de estimação, se deve ser por Ressubstituição, por Amostra Teste ou por validação cruzada; pode escolher uma dentre quatro medidas de impureza, entre as quais a de Gini, pode escolher o valor de β que é o parâmetro limitador da impureza dos nós, permite fixar uma pergunta fora do “default” que cria o ponto de corte em função de $\Delta_i(\mathcal{S}, t)$, para um ou mais nós; e permite considerar o custo de má-classificação. No apêndice A4.7 é indicada uma matriz do custo do tipo proposto por BREIMAN e col(1984) aplicável no CART.. Também estão nesse apêndice a forma de entrada dos dados, opções e “output” .

4.2.3.5 - A técnica do pacote SPLUS:

A técnica do pacote estatístico SPLUS foi desenvolvida por CLARK e PREGIBON (1992), e tem um algoritmo que como os anteriores busca partir o espaço de medidas X sucessivamente usando partições binárias em sub-conjuntos cada vez menores e mais homogêneos de modo que a distribuição condicional de y dado x , $p(y/x)$, dentro de cada sub-conjunto, não depende de x . O diagrama resultante da aplicação desse algoritmo é do tipo da figura 4.2.3.5.1, que é complementado por mais informações como se vê no apêndice A4.6. Os passos desde a inicialização até o teste para verificar se a partição segundo x é a mais indicada, são os mesmos da técnica CART.

FIGURA 5.2.3.5.1: Árvore da SPLUS com PARTO como variável dependente

OBS: V20=tipo de parto, V40=(peso na última consulta)/200, V6=tempo de gestação na primeira consulta, V11=tempo de gestação na última consulta em semanas, V33=estatura/2, V34=(peso pré-gestacional)/200. Parâmetro de parada $k=9$.

Essa técnica foge ligeiramente das anteriores, estando mais próxima dos modelos lineares clássicos e dos métodos de regressão e classificação de dados. A idéia é a de estimar uma função escada $\tau(\mathbf{x})$ que é relacionada a um parâmetro da distribuição condicional de $\mathbf{y}|\mathbf{x}$, para cada parte de $\mathbf{x}$ resultante da partição no nó anterior. Para que a partição seja feita usa-se como medida de dissimilaridade a estatística chamada “*deviance*”, que é baseada na estatística log da razão de verossimilhança, de modo que se encontre a partição do nó que seja “*mais provável*”.

A árvore será para classificação se a variável dependente for categórica, ou será para regressão se a variável dependente for numérica(contínua ou discreta). Uma característica discutível do ajuste de uma árvore pelo SPLUS, no caso em que a variável dependente for categórica ordinal, é a desconsideração da ordem entre os seus níveis.

O modelo usado para a classificação é baseado na distribuição multinomial. Algumas notações são mais específicas para esse caso, por exemplo para a variável dependente **Adequação do Peso do RN**, tendo três níveis, se o valor de um dado bebê for GIG, a notação para seu registro seria: $\mathbf{y} = (0, 0, 1)$. Para o nó raiz o vetor média representando as probabilidades que $\mathbf{Y}$ caia em cada um dos níveis é dado por $\mu = (p_1, p_2, p_3)$ de modo que $\sum_{j=1}^{d=3} p_j = 1$, e para o nó $\mathbf{t}$ o vetor média é dado por $\mu_{\mathbf{t}} = (p_{1|\mathbf{t}}, p_{2|\mathbf{t}}, p_{3|\mathbf{t}})$ com $\sum_{j=1}^{d=3} p_{j|\mathbf{t}} = 1$.

Considerando que cada modelo tem uma parte estocástica e uma estrutural, isto é, a forma como os dados estão sendo modelados. O modelo de classificação consiste do componente estocástico:

$$y_i \sim \text{Multinomial}(\mu_i) \quad \text{para } i=\text{número do indivíduo} \quad 4.2.3.5.1$$

e componente estrutural: $\mu_i = \tau(x_i)$ (a função escada é do tipo massa (MOOD, GRAYBILL, BOES(1974))) 4.2.3.5.2

A função deviance para uma observação y_i nesse modelo, é definida como:

$$\mathcal{D}(\mu_i; y_i) = -2 \sum_{j=1}^d y_{ij} \log(p_{ij}) \quad 4.2.3.5.3$$

onde j =número da categoria

No caso em que o modelo é de regressão, μ_i é um único valor numérico para y_i , a qual é considerada a média de uma distribuição normal, que consiste no componente estocástico do modelo:

$$y_i \sim \text{Normal}(\mu_i, \sigma^2) \quad \text{com } i=1, 2, \dots, N \quad 4.2.3.5.4$$

e o componente estrutural:

$$\mu_i = \tau(x_i) \quad 4.2.3.5.5$$

A função deviance do tipo 4.2.3.5.3 para uma observação y_i nesse modelo, é definida como:

$$\mathcal{D}(\mu_i; y_i) = (y_i - \mu_i)^2, \quad 4.2.3.5.6$$

que é menos duas vezes o log da verossimilhança ponderado pela variância, (assumida constante para todo i).

Para um dado nó t , o parâmetro μ_t é considerado o mesmo para todas as n_t observações dentro desse nó, e o estimador de μ_t por

máxima-verossimilhança ou equivalentemente por mínima deviance, é dado pelo vetor de proporções $\hat{\mu}_t = (\hat{p}_{1t}, \hat{p}_{2t}, \hat{p}_{3t})$ no caso do modelo de classificação 4.2.3.5.1, ou a média simples dos elementos internos ao nó $\mu_t = \sum_{i=1}^{n_t} \frac{y_i}{n_t}$ no caso do modelo de regressão 4.2.3.5.4.

A deviance de um nó é a soma das deviances de todas observações daquele nó:

$$\mathcal{D}(\hat{\mu}; y) = \sum_i \mathcal{D}(\hat{\mu}_i; y_i) \quad 4.2.3.5.7$$

que tem a mesma função da medida de impureza $\varphi(t)$ da CART, ou seja, a deviance é zero se dentro do nó todos os valores da variável y são os mesmos, isto é, o nó é puro; e aumenta à medida que os valores de y se espalham uniformemente entre as categorias.

Para estudar a bondade da partição $\mathcal{S}$, o SPLUS usa a mesma idéia da CART, comparando a deviance do nó origem com as deviances dos nós gerados pela partição. A partição que maximizar a diferença:

$$\Delta_D = D(\hat{\mu}, y) - D(\hat{\mu}_E, y) - D(\hat{\mu}_D, y) \quad 4.2.3.5.8$$

onde $\hat{\mu}_E$ e $\hat{\mu}_D$ são os estimadores das médias respectivamente do nó t_E descendente esquerdo e do nó t_D descendente direito, é a escolhida como melhor.

O critério usado pelo SPLUS para determinar se um nó é terminal é:

- a) se a deviance do nó é menor que uma pequena fração da deviance do nó raiz. O “default” do programa é 1%.
- b) se o tamanho do nó for pequeno o bastante. (o “default” do programa é $n_t = 10$).

O “default” resulta em uma árvore bem grande com aproximadamente $N/10$ nós terminais. O SPLUS permite alterar tais limites.

Diferenciando dos outros métodos, o critério usado pelo SPLUS na construção da árvore está mais associado com a seleção de variáveis “stepwise backward” em modelos lineares. Em vez de expandir a árvore a partir do nó raiz de uma forma “stepwise forward”, testando nó a nó a bondade do mesmo para então limitar o seu crescimento, o SPLUS constrói a árvore de tamanho máximo, a partir da qual passa a encolhê-la em um estudo de cada ramo, podando aqueles que se apresentarem menos representativos em termos de deviance ou tamanho.

Esse programa permite muitas opções de manuseio na poda, encolhimento, uso de custo-complexidade, uso de validação cruzada, entre outros. No apêndice A4.6 é apresentada a forma de entrada dos dados, opções e “output” desse programa.

4.2.3.6 - A técnica KS:

O KS-KnowledgeSEEKER é um software desenvolvido por FirstMark Technologies Ltd.-Canadá, sendo que o utilizado neste trabalho é a versão 2.06 de 1990.

O algoritmo desse programa faz as partições múltiplas, em uma forma “stepwise forward”, resultando em um diagrama parecido com a árvore da técnica CHAID, conforme figura 4.2.3.6.1. Em cada nó o programa faz um teste não indicado pelo manual, tendo como referência o trabalho de KASS(1980).

FIGURA 5.2.3.6.1: Árvore da KS com PARTO como variável dependente

filtro 0,01 V10=peso na última consulta, V5 peso na 1a.consulta

Para determinar o rigor da análise, que no fundo trata de selecionar a forma de partir cada conjunto de dados de tal forma que os sub-conjuntos descendentes sejam mais “puros”, ou homogêneos, do que o sub-conjunto anterior, o programa KS permite ao usuário a escolha do “*filtro*” que é o nível de significância para testar a validade da relação entre a variável dependente y e cada variável independente x , e por “default” tem o valor 0.05. Quanto maior for o filtro, maior é a árvore, e aumenta o número de detalhes das associações e do conjunto de dados em uso, sempre com os primeiros nós indicando as principais variáveis, as mais confiáveis e estáveis.

Além disso é possível definir o tipo de variável, definir os campos de leitura, a significância dos testes para junção de níveis de categorias como na CHAID, escolha das variáveis independentes a serem incluídas no modelo, edição dos dados, o tipo de regra de partição a ser considerado para cada variável (livre, monotônica, flutuante), podar ramos que possam ter interferências indesejáveis ou instáveis, encolher a árvore alterando o valor do filtro, modificar os detalhes em cada nó, fazer as partições manualmente, nó por nó.

Depois de feita a partição e criado o nó, a KS indica outras 8 possíveis partições que podem ser consideradas para aquele nó. Como as técnicas CART e SPLUS, a KS tem uma opção de calcular as taxas de má-classificação por ressubstituição e por amostra-teste. Uma restrição nos dados, para o uso da KS, é que o número de níveis de categorias sejam 20 ou menos.

No apêndice A4.7 é apresentada a forma de introdução ao KS, leitura dos dados e suas opções.

5 - RESULTADOS E DISCUSSÃO:

5.1 - Descrição de algumas variáveis medidas na amostra e associações:

Para que se possa familiarizar com as informações pertinentes à amostra em estudo e suas nuances, em 5.1.1 serão descritas as variáveis medidas que estão mais diretamente ligadas as curvas em comparação, as distribuições dos pesos dos recém-nascidos, e as distribuições das adequações dos pesos dos recém-nascidos. Posteriormente em 5.1.2 serão descritas as associações e correlações existentes entre as variáveis medidas

Em alguns casos, buscando eliminar possíveis efeitos de Local, serão usados apenas os dados da sub-amostra FSP para verificação e confirmação de comparações feitas.

5.1.1- Descrição univariada de algumas medidas:

Tabela 5.1.1 - Distribuição das gestantes segundo o Local:

Local	frequência	porcentagem
COTIA	35	6.2%
FSP	530	93.8%

Nota-se que o tamanho amostral de Cotia é praticamente insignificante perto do tamanho amostral da Faculdade de Saúde Pública. No apêndice A5.2 verifica-se que a junção dos dados de ambos os locais não interfere na distribuição de várias variáveis.

No apêndice A5.1 são apresentadas as distribuições das variáveis: estatura da gestante(v3), peso pré-gestacional(v4), peso na primeira consulta(v5), tempo de gestação na primeira consulta(v6), peso da gestante na última consulta(v10), tempo de gestação na última consulta(v11), tempo de gestação em dias por ocasião do parto em dias(v15), tempo de gestação por ocasião do parto em semanas(v16), sexo do recém-nascido(v18), tipo de parto(v20), pesos ganhos(PESOG1 e PESOG2), e pesos ganhos relativos(PESOGR1 e PESOGR2).

As figuras 5.1.1 e 5.1.2 terão boa contribuição na comparação entre as curvas de Siqueira e de Rosso.

V7 = estado nutricional segundo Siqueira na primeira consulta

V8 = Estado nutricional segundo Rosso na primeira consulta

Nota-se que a distribuição do estado nutricional segundo Siqueira é ligeiramente assimétrica, com 6% a mais (1.42 vezes) de Obesas do que Desnutridas. Como era de se esperar pelo comentado na seção 2.1 a porcentagem de normais está próximo de 70%.

Comparando-se a figura 5.1.1 com a figura 5.1.2 nota-se que a distribuição do estado nutricional segundo Rosso é diferente do estado nutricional segundo Siqueira. A distribuição de Rosso apresenta maior concentração à esquerda, sendo a proporção de desnutridas 1.48 vezes maior que a

de obesas. A porcentagem de normais segundo a curva de Rosso, 48%, é bem menor que a dada pela curva de Siqueira(70%).

Por esses valores pode-se tirar duas hipóteses para a curva de Rosso: ou ela considera a faixa de normalidade bem menor que a de Siqueira, nesse caso as gestantes consideradas por Rosso como NOR na primeira consulta teriam o peso variando em menor espaço do que os pesos das gestantes consideradas por Siqueira como NOR , ou a curva de Rosso apresenta dificuldades na delimitação entre gestantes desnutridas e normais. Fazendo essa mesma análise para apenas os dados da sub-amostra FSP, encontraram-se as mesmas distribuições e mesmas comparações.

V9 = valor da escala de Rosso na primeira consulta

O menor valor encontrado na amostra para a escala de Rosso, foi 74 e o maior valor foi 136. Nota-se que os valores da escala de Rosso para desnutrição na primeira consulta, comparativamente à distribuição normal, tem comportamento não simétrico, com valores diferentes à esquerda da média e à direita da média, que sozinha não necessariamente dá a informação conclusiva de que se trata de uma distribuição com mais desnutridas que obesas, embora a figura 5.1.2

confirme tal suposição, pois os dados que geraram esse gráfico são para os diversos tempos de gestação como pode ser visto na figura figura A5.1.4 do apêndice A5.1, e a classificação no estado nutricional de Rosso depende do tempo de gestação, como foi visto na seção 2.2. Apesar da diferença de comportamento para as caudas da distribuição dos valores da escala de Rosso, a média igual a 107.28, a mediana, igual a 107.0, o ajuste da normal, e o coeficiente de assimetria igual a 0.2345, permitem afirmar que essa distribuição pode ser considerada simétrica.

V12 = estado nutricional segundo Siqueira na última consulta

V13 = Estado nutricional segundo Rosso na última consulta

Como na primeira consulta, a distribuição do estado nutricional segundo Siqueira continua assimétrica, agora um pouco mais, com 3 vezes a mais de Obesas do que Desnutridas. Nesse caso a porcentagem de normais está próxima do esperado 70%, segundo a seção 2.1.

Nota-se que a distribuição do estado nutricional segundo Rosso na última consulta é bem mais assimétrica do que o estado nutricional de Siqueira nessa consulta, e também bem mais assimétrica do que do próprio Rosso na primeira consulta(figura 5.1.2).

Na primeira consulta a assimetria indicava uma proporção de desnutridas 1.48 vezes à de obesas, e agora na última consulta tem-se um aumento na proporção de desnutridas, para 1.71 vezes à proporção de obesas, distorcendo ainda mais a distribuição pois a porcentagem de normais segundo a curva de Rosso é menor que a de desnutridas, quase a metade da apresentada pela curva de Siqueira. Nessa consulta, fica claro a disparidade entre as curvas, com a de Siqueira mais próxima de uma curva esperada em termos de simetria e proporções, e a de Rosso considerando um número bem maior de gestantes como sendo desnutridas. Novamente pode-se ter as hipóteses levantadas no comentário sobre o comportamento da curva de Rosso na primeira consulta (figura 5.1.2), e no presente caso, o ponto de corte entre desnutrida e normal subiria mais, acarretando o campo de variação para a normalidade menor ainda. Fazendo a mesma comparação para apenas os dados da sub-amostra FSP, encontraram-se idênticos resultados.

V14 = valor da escala de Rosso (porcentagem do padrão) na última consulta

Comparando os valores da escala de Rosso da primeira consulta(fig.5.1.3), com os da última consulta(fig.5.1.6), a distribuição ficou com uma assimetria mais forte à esquerda, concentrando-se mais nos valores maiores. O menor valor encontrado na amostra para a escala de Rosso na última consulta, foi 87 e o maior valor foi 136. Como já foi afirmado com relação à primeira consulta, essa assimetria sozinha não dá a informação conclusiva de que se trata de uma distribuição com mais obesas que desnutridas, como pode ser visto na figura 5.1.5.

A amostra que Rosso usou para construir sua curva apresentou 51 gestantes, aproximadamente 20%, com porcentagens do padrão acima de 150%, mostrando uma diferença entre as composições da sua amostra e a da amostra aqui estudada.

V17 = pesos dos Recém-nascidos

O peso do RN apresentou o valor mínimo de 1750g e valor máximo de 4600g. SIQUEIRA encontrou em 1979 uma média de 3314g para uma amostra de 131 recém-nascidos de gestantes atendidas na FSP. No apêndice A5.3 estudaram-se as distribuições dos pesos dos RN para cada local, a diferença entre as mesmas e o efeito da junção entre os locais.

V19 = Adequação do Peso do Recém-Nascido

As porcentagens de bebês considerados pequenos ou grandes para a idade gestacional, apenas 5,5%(PIG) contra 14%(GIG), indicam uma assimetria na distribuição da adequação, embora a distribuição de peso dos recém-

nascidos da figura A5.1.7 tenha mostrado simetria. Nota-se que a grande maioria dos recém-nascidos foram considerados como tendo peso adequado para a idade gestacional(80%). Essas mesmas tendências foram encontradas por TANAKA(1986).

A partir da seção 5.1.2 será feita uma análise da associação entre as variáveis em estudo buscando principalmente aquelas que tenham forte associação com o peso do RN.

5.1.2: Descrição das associações e correlações existentes entre as variáveis em estudo:

Tabela 5.1.3: Coeficientes de correlações entre as variáveis medidas / nível descritivo do teste $H_0: \rho = 0$

	V3	V4	V5	V6	V10	V11	V15	V16	V17
V3	1	<u>0,4531/</u> 0,000	<u>0,4564/</u> 0,000	0,0129/ 0,760	<u>0,4994/</u> 0,000	0,0673/ 0,110	0,0198/ 0,638	0,0240/ 0,569	<u>0,1687/</u> 0,000
V4		1	<u>0,9269/</u> 0,000	0,0549/ 0,222	<u>0,8798/</u> 0,000	-0,006/ 0,894	-0,030/ 0,507	-0,021/ 0,642	<u>0,3375/</u> 0,000
V5			1	<u>0,2112/</u> 0,000	<u>0,9274/</u> 0,000	0,0386/ 0,359	-0,026/ 0,540	-0,019/ 0,647	<u>0,3572/</u> 0,000
V6				1	0,0401/ 0,341	<u>0,1625/</u> 0,000	<u>0,0950/</u> 0,024	<u>0,1082/</u> 0,010	-0,044/ 0,296
V10					1	<u>0,1154/</u> 0,006	-0,002/ 0,963	-0,001/ 0,984	<u>0,3985/</u> 0,000
V11						1	<u>0,5253/</u> 0,000	<u>0,5173/</u> 0,000	<u>0,1457/</u> 0,001
V15							1	<u>0,9803/</u> 0,000	<u>0,2119/</u> 0,000
V16								1	<u>0,2040/</u> 0,000
V17									1

V3=estatura, V4=peso pré-gestacional, V5=peso na primeira consulta, V6=tempo de gestação na primeira consulta, V10=peso na última consulta, V11=tempo de gestação na última consulta, V15=duração da gestação em dias, V16=duração da gestação em semanas, V17=peso do RN.

As correlações assinaladas com negrito e sublinhado são altamente significantes, e maiores que 0.20. As correlações assinaladas com

apenas negrito são significantes mas abaixo de 0.2. As correlações assinaladas na coluna de V17 confirmam as encontradas na literatura. Nota-se que existe uma maior correlação para a duração da gestação em dias, (V15), com o peso do RN (V17), do que em semanas(V16); e que a correlação entre o peso do RN e o peso da gestante cresce à medida que se aproxima o parto (V4, V5, V10).

TABELA 5.1.4: Correlações entre as variáveis calculadas e algumas variáveis medidas diretamente e nível descritivo do teste $H_0: \rho = 0$:

	V5	V6	V10	V16	V4pV3q	V5pV3q	V10pV3q	pcsog1	pcsog2	pcsogr1	pcsogr2	V17
V5	1	<u>0,211</u> 0,000	<u>0,927</u> 0,000	-0,019 0,647	<u>0,792</u> 0,000	<u>0,876</u> 0,000	<u>0,781</u> 0,000	-0,173 0,000	0,093 0,038	<u>-0,261</u> 0,000	0,019 0,661	<u>0,357</u> 0,000
V6		1	0,041 0,341	0,108 0,010	0,063 0,164	<u>0,228</u> 0,000	0,036 0,391	<u>-0,446</u> 0,000	-0,003 0,939	<u>-0,430</u> 0,000	-0,006 0,896	-0,044 0,296
V10			1	-0,001 0,984	<u>0,715</u> 0,000	<u>0,769</u> 0,000	<u>0,838</u> 0,000	<u>0,208</u> 0,000	<u>0,336</u> 0,000	0,108 0,010	<u>0,255</u> 0,000	<u>0,398</u> 0,000
V16				1	-0,028 0,530	-0,044 0,299	-0,018 0,661	0,048 0,252	0,044 0,333	0,037 0,375	0,038 0,402	<u>0,204</u> 0,000
V4pV3q					1	<u>0,904</u> 0,000	<u>0,845</u> 0,000	-0,187 0,000	<u>-0,246</u> 0,000	<u>-0,191</u> 0,000	<u>-0,235</u> 0,000	<u>0,279</u> 0,000
V5pV3q						1	<u>0,897</u> 0,000	<u>-0,266</u> 0,000	0,018 0,680	<u>-0,271</u> 0,000	0,027 0,546	<u>0,302</u> 0,000
V10pV3q							1	<u>0,163</u> 0,000	<u>0,289</u> 0,000	<u>0,162</u> 0,000	<u>0,304</u> 0,000	<u>0,351</u> 0,000
PESOG1								1	<u>0,646</u> 0,000	<u>0,966</u> 0,000	<u>0,627</u> 0,000	0,115 0,006
PESOG2									1	<u>0,617</u> 0,000	<u>0,973</u> 0,000	<u>0,159</u> 0,000
PESOG1										1	<u>0,629</u> 0,000	0,075 0,076
PESOG2											1	<u>0,128</u> 0,004
V17												1

V5=peso na primeira consulta, V6=tempo de gestação na primeira consulta, V10=peso na última consulta, V16=duração da gestação em semanas, V4pV3q=índice de massa corporal pré-gestacional, V5pV3q=índice de massa corporal na primeira consulta, V10pV3q=índice de massa corporal na última consulta, PESOG1=peso ganho na gestação desde a primeira consulta, PESOG2=peso ganho na gestação desde o início da gravidez, PESOG1=PESOG1/(estatura)², PESOG2=PESOG2/(estatura)², V17=peso do RN.

Na tabela 5.1.4 as correlações das variáveis peso da gestante na primeira consulta(V5), na última consulta(V10), tempo de gestação no parto em semanas(V16), com as variáveis calculadas apresentou-se menor do que

usando as variáveis sem correções da tabela 5.1.3. Por exemplo a correlação entre V4(peso pré-gestacional) e V5(peso na primeira consulta) é maior que a correlação entre V4PV3Q(peso pré-gestacional corrigido) e V5(peso na primeira consulta). A variável tempo de gestação na primeira consulta(V6) mostrou-se associada com o peso medido na primeira consulta(V5 e V5PV3Q) e com o peso ganho com relação a primeira consulta(PESOG1 e PESOGR1). A duração da gestação(V16) apresentou correlação significativa com apenas as outras durações da gestação(V6, V11, V15) e com V17(peso do RN).

Comparando a variável PESOG1 com PESOGR1, e a variável PESOG2 com PESOGR2, nota-se que ambas tem o mesmo comportamento em termos de correlação com as demais variáveis.

As variáveis PESOG2(peso ganho com relação ao peso pré-gestacional) e PESOGR2(peso ganho com relação ao peso pré-gestacional corrigido) apresentaram-se correlacionadas negativamente com a variável V4PV3Q(peso pré-gestacional corrigido). Essas correlações negativas indicam, como alguns autores já haviam encontrado((CIARI(1974); (ROSSO(1985); SIQUEIRA(1981)), que quanto mais magra a mãe no início da gravidez, maior é o peso ganho. Como notou SIQUEIRA(1981), o peso na última consulta é a medida mais correlacionada com o peso do RN, com maior correlação para a medida sem correção V10.

Analisando as tabelas 5.1.3 e 5.1.4, pode-se fazer uma comparação entre a variável peso pré-gestacional e as formas de cálculos dos pesos ganhos, em termos de correlação com o peso do RN. O peso pré-gestacional apresentou-se como uma medida mais correlacionada com o peso do RN($r=0,1687$) do que os pesos ganhos($r=0,115$ para PESOG1, $0,159$ para PESOG2 e valores menores para os pesos ganhos corrigidos). Ainda com relação ao cálculo do peso ganho, pode-se notar que existe uma vantagem do uso da variável v3(peso pré-

gestacional) sobre v5(peso na primeira consulta) ($r=0,159$ versus $0,115$ e $0,128$ versus $0,075$), provavelmente devido ao fato que o peso da gestante na primeira consulta foi medido em tempos distintos de gestação, como pode ser verificado no apêndice A5.1.

A tabela 5.1.5 apresenta as correlações parciais entre as diversas variáveis em estudo, buscando verificar possíveis mudanças da influência dessas variáveis sobre o peso do RN quando o peso pré-gestacional é controlado.

Tabela 5.1.5: Correlações parciais, controlando o Peso Pré- Gestacional(V4):

	V3	V5	V6	V10	V16	V5PV3Q	V10PV3Q	PESOGRI	PESOGRI2	V17
V3	1	0,124/ 0,006	-0,023/ 0,612	<u>0,256/</u> 0,000	0,009/ 0,841	<u>-0,690/</u> 0,000	<u>-0,602/</u> 0,000	0,055/ 0,224	0,098/ 0,030	0,0359/ 0,425
V5		1	<u>0,442/</u> 0,000	<u>0,633/</u> 0,000	-0,018/ 0,683	<u>0,608/</u> 0,000	<u>0,414/</u> 0,000	<u>-0,203/</u> 0,000,,	<u>0,619/</u> 0,000	<u>0,120/</u> 0,007
V6			1	0,005/ 0,912	0,099/ 0,027	<u>0,328/</u> 0,000	0,019/ 0,665	<u>-0,421/</u> 0,000	0,006/ 0,885	-0,067/ 0,135
V10				1	0,041/ 0,364	<u>0,235/</u> 0,000	<u>0,601/</u> 0,000	<u>0,607/</u> 0,000	<u>0,975/</u> 0,000	<u>0,227/</u> 0,000
V16					1	-0,039/ 0,390	0,018/ 0,693	0,064/ 0,154	0,034/ 0,452	<u>0,225/</u> 0,000
V5PV3Q						1	<u>0,766/</u> 0,000	<u>-0,201/</u> 0,000	<u>0,349/</u> 0,000	0,043/ 0,338
V10PV3Q							1	<u>0,448/</u> 0,000	<u>0,710/</u> 0,000	0,149/ 0,001
PESOGRI								1	<u>0,613/</u> 0,000	0,152/ 0,000
PESOGRI2									1	<u>0,222/</u> 0,000
V17										1

V3=estatura, V5=peso na primeira consulta, V6=tempo de gestação na primeira consulta, V10=peso na última consulta, V16=duração da gestação em semanas, V5pV3q=índice de massa corporal na primeira consulta, V10pV3q=índice de massa corporal na última consulta, PESOGRI=PESOG1/(estatura)², PESOGRI2=PESOG2/(estatura)², V17=peso do RN.

Na tabela 5.1.5 tem-se os coeficientes de correlações parciais, controlando o Peso Pré-Gestacional. Comparando-se as tabelas 5.1.3, 5.1.4 e 5.1.5, vê-se que a variável V6(tempo de gestação na primeira consulta)

mostrou-se sempre mais associada com V5(peso na primeira consulta), PESOG1(peso ganho com relação a primeira consulta) e V5PV3Q(peso da primeira consulta corrigido pela estatura), sendo que a associação entre V6 e V5 cresceu com o controle de V4(peso pré-gestacional) (de $\rho=0,211$ para $\rho=0,442$).

Como SIQUEIRA(1981) já havia notado, controlando o peso pré-gestacional(V4), a estatura(V3) não se mostra mais associada com o Peso do RN(V17)($\rho=0,0359$), embora a correlação entre V3(estatura) e V17(peso do RN) sem controlar V4(peso pré-gestacional) seja baixa($\rho=0,1687$). Mesmo controlando V4, a variável V10(peso na última consulta) continua sendo a mais correlacionada com o peso do RN($\rho=0,227$). A variável V10PV3Q(peso na última consulta corrigido) era antes de se controlar V4(peso pré-gestacional), uma das mais altamente correlacionadas com o peso do RN(V17)($\rho=0,351$), e agora sofrendo uma forte influência de V4, apresenta uma correlação de 0.149. O mesmo ocorreu com as variáveis V5(peso na primeira consulta) e V5V3Q(peso na primeira consulta corrigido), pois as correlações do peso do RN(V17) com V5 e com V5PV3Q caíram bastante($\rho=0,357$ para $\rho=0,120$ e $\rho=0,302$ para $\rho=0,043$ respectivamente). Isso significa que o impacto dessas variáveis maternas sobre o peso do RN é mínimo ou nenhum quando se controla o peso pré-gestacional. A correlação da variável V16(duração da gestação) com o peso do RN, aumentou com o controle do peso pré-gestacional(V4). O interessante, é que a variável peso pré-gestacional(V4) continua sendo mais importante que a variável peso na primeira consulta(V5), no cálculo do peso ganho durante a gestação, como pode ser visto ao comparar as correlações dessas variáveis com o peso do RN(V17)($\rho=0,222$ versus $\rho=0,152$).

A forma como ocorre a associação entre o peso do RN(V17) e o peso na última consulta(V10), pode ser visualizada nas figuras 5.1.9, e

5.1.10. Nessas figuras pode-se notar a força de discriminação das variáveis V7(estado nutricional segundo Siqueira) e V8(estado nutricional segundo Rosso).

As figuras 5.1.9 e 5.1.10 mostram que independentemente do método de diagnóstico, a relação entre o peso do RN(V17) e

peso na última consulta(V10) é do mesmo tipo, embora haja alguma diferença no agrupamento de pontos em algumas regiões devido à forma como as curvas de Siqueira e de Rosso interpretam os pesos das gestantes.

Os agrupamentos de pontos nas regiões referentes aos estados nutricionais determinados pelas curvas mostram que ambas as curvas estão cumprindo o seu papel de diagnosticar como desnutrida quando a gestante tem baixo peso na última consulta, como obesa quando tem peso alto e normal quando o peso é intermediário.

A diferença na densidade da nuvem de pontos azuis indica uma defasagem entre as curvas, com a curva de Siqueira tendo a tendência de diagnosticar mais vezes como estado nutricional normal, e a maior densidade da nuvem de pontos vermelhas para Rosso indica a tendência dessa curva em diagnosticar mais vezes como estado nutricional DES.

Passando uma linha vertical próxima de 69 kg, que é o peso limite em que a curva de Rosso tem sua nuvem de pontos referentes ao diagnóstico DES, pode-se notar que essa curva considera várias gestantes DES enquanto a curva de Siqueira considera NOR, pois a nuvem de pontos referente ao diagnóstico NOR de Siqueira tem limite inferior próximo de 52 kg.

A utilização apenas do peso do RN como indicador do crescimento fetal, tem sido rejeitada, devido principalmente à idade gestacional, pois apenas o peso do RN não traz toda informação sobre o tamanho e o desenvolvimento do recém-nascido, motivo pelo qual indica-se o uso da Adequação do peso do RN que envolve a idade gestacional, a qual neste trabalho varia de 228 a 308 dias ou 33 a 44 semanas. Como pode ser notado na figura 5.1.11, a seguir, mais da metade dos bebês PIG tem peso ao nascer dentro do campo de variação dos bebês AIG,

como por exemplo bebês que são considerados PIG com 2650g e bebês considerados AIG com 2050, existindo também intersecção entre as distribuições de AIG e GIG.

Peso do RN(V17) vs Adequação do seu peso(V19):

Calculando as correlações parciais entre todas as variáveis e o peso do RN, fixando o tempo de gestação(V15), o peso da gestante na última consulta (V10) continuou sendo a mais correlacionada, seguida do peso da gestante na primeira consulta(V5).

Os quatro gráficos a seguir mostram como se comporta a distribuição do peso do RN em função do peso pré-gestacional, identificando-se o diagnóstico do estado nutricional de cada curva na primeira e última consultas.

DISTRIBUIÇÃO DO PESO DO RN EM FUNÇÃO DO PESO PRÉ-GESTACIONAL PARA OS DIAGNÓSTICOS DE SIQUEIRA(V7)

DISTRIBUIÇÃO DO DO PESO DO RN EM FUNÇÃO DO PESO PRÉ-GESTACIONAL PARA OS DIAGNÓSTICOS DE ROSSO(V8)

As figuras 5.1.12 e 5.1.13 confirmam a correlação positiva entre o peso do RN(V17) e o peso pré-gestacional(V4) ($\rho=0.3375$). O ajuste de outras curvas diferentes da regressão linear simples não foi melhor. Nota-se que a nuvem de pontos verdes(diagnóstico NOR) é bem maior para a curva de Siqueira do que para a curva de Rosso, e a nuvem de pontos vermelhos(diagnóstico DES) é bem maior para a curva de Rosso, avançando para a direita, confirmando a hipótese que essa curva tem um limiar para o diagnóstico DES bem mais alto do que

a de Siqueira. Além disso nota-se também que para a curva de Rosso, o número de pontos vermelhos misturados com pontos verdes, é em maior quantidade.

DISTRIBUIÇÃO DO PESO DO RN(V17) EM FUNÇÃO DO PESO PRÉ-GESTACIONAL(V4) PARA OS DIAGNÓSTICOS DE SIQUEIRA(V12)

DISTRIBUIÇÃO DO DO PESO DO RN(V17) EM FUNÇÃO DO PESO PRÉ-GESTACIONAL(V4) PARA OS DIAGNÓSTICOS DE ROSSO(V13)

Nota-se nas figuras 5.1.14 e 5.1.15, que a nuvem de pontos verdes(diagnóstico NOR) para a curva de Siqueira, se destaca mais ainda do que na primeira consulta, indicando que o seu limiar inferior para NOR diminuiu; enquanto a nuvem de pontos vermelhos(diagnóstico DES) para a curva de Rosso

aumentou mais ainda, indicando que essa curva aumenta seu limiar superior para o diagnóstico DES. Além disso nota-se também que o número de pontos verdes que estão misturados com pontos vermelhos, para a curva de Rosso, aumentou.

Como foi citado no capítulo 2, Rosso notou que a relação existente entre peso do RN e porcentagem do peso padrão era crescente e linear até 120%, acima do qual o peso do RN não sofria mais influência da porcentagem do peso padrão, passando a ficar constante. No gráfico em que Rosso mostrou esse fenômeno, foi registrada uma porcentagem observada máxima de 160%. A tendência que Rosso encontrou, de uma regressão linear crescente entre $v14=80\%$ e 120% , parando de crescer depois de $v14=120\%$, ficando constante, e o limite máximo de 160% são fatos não encontrados na presente amostra, como já foi citado no comentário da figura 5.1.6 e pode ser visto na figura 5.1.16.

PESO DO RN vs ESCALA DE ROSSO:

Para descrever a relação existente entre o peso do RN e a porcentagem do peso padrão da gestante na última consulta, a regressão encontrada

para a presente amostra com o melhor ajuste foi sempre crescente, e não como Rosso encontrou em sua amostra. A tendência crescente depois de 130% do peso padrão notada na figura 5.1.16 poderia ser modificada, e ficar constante como encontrado por Rosso, se houvesse um grupo de gestantes com porcentagem do padrão entre 140% e 160% sem aumento do peso do RN, o que exigiria uma clientela com maior frequência de obesidade, e peso do RN sem variação para quaisquer valores de v14 nesse intervalo.

5.2 - Comparações entre as curvas de Rosso e de Siqueira:

Embora ambas as curvas tenham sido usadas constantemente em saúde pública, para avaliar o estado nutricional da gestante, e para prever o estado nutricional do futuro bebê, não existem registros na literatura de comparações, indicações ou confirmações sobre a preferência de uma delas.

No entanto, pela sua importância, a curva de Rosso tem merecido diversos estudos, e sua avaliação tem sido uma das prioridades da área de nutrição materno-infantil(WHO(1995); KRASOVEC & ANDERSON(1990)), estando em uso na atenção pré-natal no Chile e no Uruguai(TADDEI e col(1991)), o que remete a indicação dessa curva como parâmetro de comparação para qualquer curva que esteja sendo construída.

Para que sejam feitas as comparações, serão usadas análises da estrutura e construção de ambas curvas, e análises estatísticas visando identificar a eficiência, a consistência, a sensibilidade e especificidade das curvas.

Nessa comparação, deve-se lembrar que no cerne da questão está a relevância da necessidade do estudo das condições nutricionais da gestante, e a compreensão das variáveis que possam estar afetando o desenvolvimento do feto. Para tal, deve-se procurar formas de análise dos dados provenientes de pesquisas pertinentes, que possam identificar a eficiência e a consistência das curvas. Para que se possa ter uma idéia do comportamento dos dados usados por ambos autores, na construção de suas técnicas, a tabela 5.2.1 mostra algumas médias calculadas nesses trabalhos.

Como pode ser notado nessa tabela, as gestantes estudadas por Siqueira eram em média mais velhas, mais baixas, mais magras no início da gestação, e tiveram bebês mais magros do que as estudadas por Rosso.

TABELA 5.2.1: Quadro das médias de algumas medidas calculadas nos trabalhos de Siqueira e de Rosso.

VARIÁVEL	SIQUEIRA	ROSSO
Idade da mãe	25.53	23.82
Altura da mãe	155.58	161.16
Peso Pré-gestacional	53.76	61.30
Duração da gestação	39.09 semanas	39.27 semanas
Peso do RN	3187 gramas	3339 gramas
Peso no final da gravidez	63.48	não publicado

5.2.1 - Construção e Estrutura das Curvas:

Para verificar vantagens de uma curva sobre a outra, vamos levar em consideração a facilidade operacional e praticidade de aplicação, a abrangência de aplicação com relação ao tipo de clientela, tamanho amostral usado para construir as curvas, atualização dos dados usados para a construção das mesmas, tipo de padrão de peso utilizado como referência.

Pela sua construção, algumas **vantagens do método de Siqueira**, são:

- é um método de rápido acesso à adequação do peso para uma dada altura, em qualquer momento da gestação.
- em cada controle do pré-natal, para uma dada altura, é possível verificar a adequação ou defasagem do peso da gestante, possibilitando a indicação de dieta ou suplementação,

- trata-se de uma forma visual de monitorar o progresso gestacional para cada visita do pré-natal,
- a curva de Siqueira considera um intervalo de normalidade, incorporando as variações biológicas e particulares, em cada consulta, para cada faixa de estatura.
- a amostra utilizada por Siqueira é de tamanho 1354, que pode ser considerado um valor muito bom.
- Siqueira considerou todas as mães independentemente do número de filhos, desde que pudessem ser consideradas gestantes normais e sem patologia importante; além disso não houve restrição quanto ao tempo de gestação da paciente.

Pela sua construção, algumas **desvantagens do método de Siqueira**, são:

- operacional, pois a sua curva trata na realidade de 5 curvas, o que exige a escolha de uma das cinco folhas, uma de cada faixa de estatura, para cada diagnóstico. Em um estudo amostral ou populacional, o peso referente a uma gestante deve ser ajustado na curva referente à sua faixa de estatura.
- Outra desvantagem da curva de Siqueira, é a sua desatualização, pois a mesma foi construída para representar as tendências da década de 70, e como afirmam MONTEIRO e col.(1995), existe uma variação na constituição física dos adultos, com redução da prevalência de adultos com baixo Índice de Massa Corporal no Brasil; redução essa de uma média de 50% entre 1974 e 1989. Dessa forma, Monteiro e colaboradores mostram que a estatura no Brasil teve entre 1951 e 1968, um acréscimo médio de 1.48cm para masculino e 1,18cm para feminino.

Pela sua construção, algumas **vantagens do método de**

Rosso, são:

- é uma forma rápida de acesso à adequação do peso para altura em qualquer tempo acima de 12 semanas de gestação,
- para cada estatura materna, há a sugestão do incremento desejável para o peso materno,
- trata-se de uma forma visual para monitorar o progresso gestacional para cada visita do pré-natal.

Pela sua construção, algumas **desvantagens do método**

de Rosso, são:

- Para a construção de sua curva, Rosso considerou para cada semana de gestação, apenas a média dos pesos das gestantes(ponderados pelo peso padrão), e não um intervalo de normalidade. É possível que uma mulher de estatura corporal pequena, ou grande, no pré-parto, saia da tendência da média durante a gravidez e tenha no parto uma massa corporal diferente da prevista pela média da amostra estudada,
- O tamanho amostral pode ser considerado pequeno, comparativamente ao de Siqueira, $n=262$, para detectar todas as variações biológicas, para cada faixa de estatura,
- Existe uma restrição na aplicação dessas curvas, pois as mesmas foram construídas com as informações de mulheres em idade entre 20 e 25 anos, o que implica em inadequação para idades diferentes.
- A construção da curva de Rosso incluiu somente mães primigestas, com a restrição adicional de que o tempo de gestação deveria estar entre 38 e 42 semanas, um grupo bem mais restrito do que o de Siqueira, Aqui existe

uma possível dificuldade de inferir utilizando essa curva, que é o de ter que esperar terminar a gravidez para ver se a gestante pode ser diagnosticada pela curva.

- O método de Rosso foi construído para o diagnóstico de gestantes com idade gestacional de 12 semanas ou mais, enquanto o método de Siqueira é para 8 semanas ou mais.
- Toda construção dessa curva, está baseada em um padrão de peso pré-gestacional construído por JELLIFFE(1966), derivado matematicamente a partir dos níveis “desejados” calculados por companhias de seguro dos Estados Unidos e pela Sociedade Atuaria, estando implícito que o desenvolvimento gestacional para duas mulheres com potencial genético diferente, que tenham um dado peso padrão, segue a mesma função, sem nenhuma variabilidade. Nesses casos o padrão é o peso antes de ficar grávida, e não um padrão para aquele período gestacional. CIARI(1974) comenta que o fato do padrão ser mulheres adultas não grávidas pode distorcer os resultados.
- Além disso, o padrão foi construído segundo as tendências das condições sociais e educacionais da década de 50, que tinha a mulher em contexto social diferente da época em que foi construída, 1985. Com base nas pesquisas citadas anteriormente, principalmente o trabalho de Monteiro e col.(1995) em que mostra a influência do tempo, e no trabalho de EVELETH & TANNER(1976) sobre a interferência do meio, e no trabalho de KRAMER(1987) em que verifica que a definição de “magra”, ou ponto de corte, está associado com a região ou país, pode-se imaginar prováveis alterações nos padrões de peso da população, e necessidade constante de ajustes desse padrão.

5.3 - Comparação estatística entre as curvas de Rosso e de Siqueira:

5.3.1 - Utilização de Análise exploratória:

Já foram feitas algumas comparações gráficas na seção 5.1.1, usando as figuras 5.1.1, 5.1.2, 5.1.4, 5.1.5, que indicam para um vício no método de Rosso, ao diagnosticar as mães na primeira consulta com uma tendência mais para “desnutrida” do que para “obesa”. Tal fato pode ser visto como uma desvantagem para a curva de Rosso, se for considerado que o ideal seja haver simetria entre os extremos dos estados nutricionais. Essa tendência pode ser atribuída ao fato que na primeira consulta muitas gestantes tinham menos de 12 semanas de gestação, como pode ser visto na figura A5.1.4, do apêndice A1, período para o qual essa curva não foi construída, ou devido a definição de “magra” estar associada com a região ou país, como citou KRAMER(1987). Essa tendência aumenta com o passar do tempo gestacional, chegando a um diagnóstico muito defasado, no final da gravidez, como pode ser visto comparando os gráficos das figuras 5.1.4 e 5.1.5 com o da figura 5.1.8.

A partir dessas análises e comparações foram levantadas hipóteses, e com o intuito de verificá-las serão analisados os intervalos de confiança para a média dos pesos das gestantes na primeira e última consultas (V5 e V10) e os desenhos “box-plot” para as distribuições desses pesos. Como a caixa do desenho

box-plot cobre 50% da distribuição e a distância entre os valores extremos do desenho box-plot cobre 99.3% da distribuição, para que se possa ter uma idéia mais abrangente e detalhada da distribuição, e fazer comparações não só pontuais, mas também intervalares e de formatos, algumas comparações serão feitas prioritariamente utilizando o desenho box-plot, algumas vezes reforçadas com intervalo de confiança ao nível de 95%.

Como a variável peso do RN(V17), ou a variável adequação do peso do RN(V19), é o valor real que se procura prever a partir das demais variáveis independentes, entre elas o diagnóstico nutricional da gestante, será feita também uma comparação gráfica entre os valores dos diagnósticos das curvas de Rosso e de Siqueira com os valores do peso do RN e os valores da adequação do peso do RN. Tal comparação deverá dar uma idéia de qual curva está cumprindo de melhor forma o seu papel de diagnosticar uma gestante como desnutrida cujo bebê nasce com peso baixo, assim por diante.

Posteriormente uma análise de tabelas de distribuições conjuntas de V7 com V8, V12 com V13, e ambas as curvas com V19, culminará com uma idéia mais aprimorada da sensibilidade e especificidade dessas curvas.

Como foi visto na figura 5.1.11 existe uma interação do do Peso do RN com a idade gestacional, motivo pelo qual a análise das tabelas e gráficos envolvendo o peso do RN, buscando comparar ambas as curvas, deve ser vista como complemento e não como palavra final.

COMPARAÇÃO ENTRE AS DISTRIBUIÇÕES DE PESOS DAS GESTANTES NOS EXTREMOS DO PERÍODO DE GESTAÇÃO, SEGUNDO OS DIAGNÓSTICOS DO ESTADO NUTRICIONAL DE CADA CURVA

Tabela 5.3.1 - Intervalos de Confiança para as médias dos pesos das gestantes nos extremos do período de gestação, segundo o diagnóstico do estado nutricional para cada curva.

período	estatística	curva	DIAGNÓSTICO DO ESTADO NUTRICIONAL		
			DES	NOR	OBS
primeira consulta	Média	Siqueira(V7)	4710	5653	7257
		Rosso(V8)	5008	5770	7206
	Interv. de Confiança	Siqueira(V7)	(4617,4803) D=186	(5594,5712) D=118	(7101,7413) D=312
		Rosso(V8)	(4934,5083) D=149	(5709,5844) D=135	(7051,7362) D=311
última consulta	Mínimo	Siqueira(V7)	3900	4400	5700
		Rosso(V8)	3900	4300	5200
	Máximo	Siqueira(V7)	6180 Δ =2280	7500 Δ =3100	9750 Δ =4050
		Rosso(V8)	6500 Δ =2600	7500 Δ =3200	9750 Δ =4550
	Média	Siqueira(V12)	5410	6372	8008
		Rosso(V13)	5938	6724	7896
	Interv. de Confiança	Siqueira(V12)	(5275,5544) D=269	(6313,6431) D=118	(7877,8139) D=262
		Rosso(V13)	(5868,6008) D=140	(6640,6809) D=169	(7752,8040) D=288
	Mínimo	Siqueira(V12)	4560	5020	6450
		Rosso(V13)	4560	5180	5480
	Máximo	Siqueira(V12)	6300 Δ =1740	7940 Δ =2920	9999 Δ =3549
		Rosso(V13)	7650 Δ =3090	8480 Δ =3300	9999 Δ =4519

A tabela 5.3.1 mostra que o peso médio aumenta à medida que a gestante é mais nutrida, e comparando as médias dos pesos das gestantes em cada consulta, nota-se que quando o diagnóstico é DES ou NOR, a

média dos pesos das gestantes é maior para o diagnóstico da curva de Rosso, e quando o diagnóstico é OBS a relação se inverte e a curva de Siqueira apresenta maior média. Em outras palavras, em média Rosso considera como DES pessoas mais pesadas que a curva de Siqueira, e para considerar a gestante OBS a curva de Rosso tem um limiar mais baixo do que a de Siqueira.

Os valores $\mathcal{D}$ referem-se ao tamanho do intervalo de confiança, dependem do tamanho da variância e quanto menor forem, maior é a precisão da estimação da média. Os intervalos de confiança para as médias dos pesos das gestantes nas duas consultas apresentam tamanho $\mathcal{D}$, menor para Rosso no diagnóstico DES e maior para Rosso no diagnóstico NOR em ambas as consultas. Para OBS não existe diferença de tamanho dos intervalos na primeira consulta, mas para a última consulta a curva de Rosso mostra maior dispersão.

Os intervalos de variação, Δ , apresentam-se maiores para Rosso em ambas as consultas e todos os diagnósticos.

Como não existe interesse em comparar apenas as médias, e sim em analisar a forma como se comportam as distribuições, e a existência de possíveis intersecções, esses fenômenos são descritos nas figuras 5.3.1 e 5.3.2 a seguir.

Distribuição dos pesos das gestantes segundo as curvas de Rosso e de Siqueira na primeira consulta:

Distribuição dos pesos das gestantes segundo as curvas de Rosso e de Siqueira na última consulta:

Confirmando o encontrado na tabela 5.3.1, as figuras 5.3.1 e 5.3.2 mostram que a curva de Siqueira apresenta pesos mais baixos do que a curva de Rosso para o estado nutricional DES e NOR, e essa relação tende a se inverter no estado nutricional OBS, com a curva de Rosso passando a apresentar pesos ligeiramente mais baixos, embora com diferença não significativa. Nota-se

que as distribuições de pesos para os diagnósticos de Siqueira são mais concentradas em todos diagnósticos e em ambas as consultas. Isso significa que a delimitação das faixas de peso para as gestantes em função dos diagnósticos é em geral mais precisa e definida pela curva de Siqueira do que pela curva de Rosso, o que não necessariamente representa uma vantagem para a curva de Siqueira. Para o diagnóstico DES, o fato da curva de Rosso ter um box-plot maior significa que essa curva engloba mais gestantes em uma faixa maior de peso, principalmente na última consulta, o que poderia ser vantagem se o interesse fosse detectar apenas DES, sem a preocupação da especificidade.

O fato de que os desenhos “box-plot” se deslocam para cima mostra que ambas as curvas estão cumprindo o seu papel de diagnosticar como desnutrida quando a gestante tem baixo peso, e normal quando o peso é um pouco maior, assim por diante. Mas ao existirem intersecções entre os campos de variação de ambas distribuições, implica no diagnóstico ambíguo de gestantes que ao mesmo tempo, na mesma consulta, são consideradas por Siqueira como normais e por Rosso como desnutridas. Por exemplo 92 gestantes foram consideradas ao mesmo tempo, na primeira consulta, por Siqueira como normais e por Rosso como desnutridas.

Nota-se que a medida que o diagnóstico passa de menos para mais nutrida, o campo de variação e a caixa do box-plot aumentam, indicando diminuição de precisão nesse sentido.

Resultados idênticos aos encontrados na tabela 5.3.1 e figuras 5.3.1 e 5.3.2, foram encontrados considerando apenas a sub-amostra da FSP.

Com relação à hipótese de que a curva de Rosso considera a faixa de normalidade bem menor que a de Siqueira, ou seja, que as gestantes consideradas por Rosso como NOR em qualquer consulta teriam o peso variando em menor espaço do que os pesos das gestantes consideradas por Siqueira

como NOR, vê-se pela análise da tabela 5.3.1 e das figuras 5.3.1 e 5.3.1 que a faixa de normalidade para ambas as curvas é praticamente de mesmo tamanho.

Assim, se ambas as curvas levassem em consideração apenas o peso naquela consulta, não existiria uma diferença tão grande como a vista nas distribuições das figuras 5.1.1, 5.1.2, 5.1.4 e 5.1.5. Então resta a hipótese da curva de Rosso ser pouco sensível à normalidade e considerar o ponto de corte para DES muito alto.

Outros comentários com relação a esses desvios serão feitos posteriormente, de modo que se possa alcançar uma conclusão a respeito.

COMPARAÇÃO ENTRE AS DISTRIBUIÇÕES DE PESOS DOS RN PARA OBSERVAÇÕES NOS EXTREMOS DO PERÍODO DE GESTAÇÃO SEGUNDO OS DIAGNÓSTICOS DO ESTADO NUTRICIONAL PARA CADA CURVA:

Além de comparar as curvas quanto ao peso das gestantes, como o objetivo principal do diagnóstico do estado nutricional da gestante é prever o tamanho do bebê para que se possa corrigir o peso com suplementação alimentar, uma forma de comparar as duas curvas é verificar o “poder” das duas em estimar o peso do RN. Será dado início a um estudo com essa finalidade imaginando o tamanho do bebê dado pelo seu peso ao nascer a partir da tabela 5.3.2, onde estão as médias dos pesos dos RN, os intervalos de confiança encontrados para as médias dos pesos dos RN segundo as curvas de Siqueira e de Rosso, em cada diagnóstico feito na primeira e na última consultas, e os respectivos valores de pesos extremos. Reforçando a comparação novamente serão usados os desenhos box-plots para cada distribuição.

O interessante nas distribuições dos pesos dos bebês segundo o estado nutricional durante a gestação de suas mães, é que seguem a

mesma tendência das distribuições dos pesos de suas mães, pois como pode ser notado na tabela 5.3.2, os pesos médios dos bebês das gestantes diagnosticadas em ambas as consultas por Rosso como DES ou como NOR, tendem ser maiores que os pesos dos bebês das gestantes diagnosticadas como DES ou NOR por Siqueira, invertendo-se a relação para o diagnóstico OBS, com a curva de Siqueira tendo maior média dos pesos dos RN. Confirmando a existência da tendência da curva de Rosso em “puxar” para baixo o seu diagnóstico.

TABELA 5.3.2: Algumas estatísticas das distribuições de pesos dos RN segundo diagnósticos nas duas consultas:

período	estatística	curva	DIAGNÓSTICO DO ESTADO NUTRICIONAL		
			DES	NOR	OBS
primeira consulta	Média	Siqueira(V7) Rosso(V8)	2962 3052	3185 3217	3498 3457
	Interv. de Confiança	Siqueira(V7) Rosso(V8)	(2855,3069)Đ=214 (2983,3121)Đ=138	(3144,3227) Đ =83 (3168,3266) Đ =98	(3411,3585)Đ=174 (3372,3542)Đ=170
	Mínimo	Siqueira(V7) Rosso(V8)	1750 1750	2010 2020	2000 2000
	Máximo	Siqueira(V7) Rosso(V8)	3900 Δ=2150 4280 Δ=2530	4460 Δ=2450 4460 Δ=2440	4600 Δ=2600 4600 Δ=2600
última consulta	Média	Siqueira(V12) Rosso(V13)	2818 3083	3182 3237	3445 3410
	Interv. de Confiança	Siqueira(V12) Rosso(V13)	(2676,2960)Đ=284 (3023,3143)Đ=120	(3141,3224) Đ =83 (3182,3293)Đ=111	(3366,3525)Đ=159 (3331,3491)Đ=160
	Mínimo	Siqueira(V12) Rosso(V13)	1750 1750	2010 2320	2000 2000
	Máximo	Siqueira(V12) Rosso(V13)	3900 Δ=2150 4280 Δ=2530	4280 Δ=2270 4280 Δ=1960	4600 Δ=2600 4600 Δ=2600

Novamente, como para o peso da mãe, os intervalos de confiança para as médias dos pesos recém-nascidos nas duas consultas apresentam tamanho Đ, menor para a curva de Siqueira no diagnóstico NOR, e

maior nos diagnósticos extremos. Na última consulta o peso do RN apresentou-se com $\bar{D}$ menor para Rosso no diagnóstico DES e maior para Rosso no diagnóstico NOR em ambas as consultas. Para OBS praticamente não existe diferença de tamanho dos intervalos na primeira consulta e última consulta. Os box-plots das figuras 5.3.3 e 5.3.4 podem elucidar de melhor maneira a comparação entre os campos de variação de ambas as curvas na primeira consulta.

Distribuição dos pesos dos RN segundo as curvas de Rosso e de Siqueira na primeira consulta:

Distribuição dos pesos dos RN segundo as curvas de Rosso e de Siqueira na última consulta:

Como era de se esperar, pelo visto na tabela 5.3.2, embora não haja uma diferença significativa entre as médias dos grupos, os box-plots dos peso do RN referentes as duas curvas mostrados nas figuras 5.3.3 e 5.3.4, indicam uma tendencia de aumentar o peso, a medida que a mãe é diagnosticada como mais gorda, mesmo havendo uma intersecção entre as distribuições vizinhas. Indicando com isso, que existem mães diagnosticadas como NOR tendo filho com baixo peso ou alto peso, existem mães diagnosticadas como OBS tendo filho com peso mediano ou baixo, assim por diante, mas no geral as mães diagnosticadas como DES tem os filhos mais magros e as mães diagnosticadas como OBS tem os filhos mais gordos.

Nota-se que os box-plots do estado nutricional NOR têm as caixas menores, confirmando os valores $\bar{D}$ da tabela 5.3.2, com o significado de que os pesos dos bebês das mães diagnosticadas na gestação como normal variam menos, pois possuem distribuições mais concentradas para esse estado nutricional e maior precisão na previsão do peso do RN a partir desse diagnóstico.

A tendência crescente é mais nítida para a curva de Siqueira do que para a curva de Rosso, mesmo porque de DES para NOR essa curva não apresenta um crescimento muito evidente nos valores dos pesos dos RN como a curva de Siqueira, indicando maior sensibilidade da curva de Siqueira.

Analizando a curva de Rosso, vê-se que a distribuição dos pesos dos RN quando o diagnóstico para a gestante é NOR, está praticamente inclusa na distribuição do diagnóstico DES, o que pode levar a um prognóstico confuso com relação ao peso do RN. Nota-se claramente nessas figuras, que o fato

dos box-plot de ambas as curvas estarem inclusos no box-plot da curva de Rosso para diagnóstico DES, significa que quando a mãe é diagnosticada tanto como DES como NOR pela curva de Rosso o peso do recém-nascido não se altera. O que indica que também com relação ao peso do RN, o ponto de corte da curva de Rosso continua tendo valores altos do peso do RN para gestantes DES, incluindo também gestantes NOR como DES.

A grande diferença entre as duas curvas quanto ao prognóstico do peso do RN é aparentemente o “desvio ótico” associado com o diagnóstico DES, onde a curva de Siqueira se mostra com pesos mais definidos para esse diagnóstico e conseqüentemente com maior precisão. Acrescente-se a isso o fato que a frequência de indivíduos diagnosticados na última consulta por Rosso como DES é maior que diagnosticados como NOR, levando a supor que haja um “vício” que tende em sub-classificar a desnutrição da gestante, como já foi notado nos gráficos 5.1.6, 5.1.7, 5.1.10, 5.1.11, 5.3.1, 5.3.2, 5.3.3, 5.3.4, tabelas 5.3.1 e 5.3.2.

Portanto, dessa análise, nota-se que o problema da curva de Rosso não é apenas a definição de normal, mas também a de desnutrida, pois as esperar-se-ia que as distribuições correspondessem, com um crescimento, à medida que a nutrição aumenta.

Resultados idênticos aos encontrados na tabela 5.3.2 e figuras 5.3.3 e 5.3.4, foram encontrados considerando apenas a sub-amostra da FSP.

Uma outra análise comparativa entre as curvas de Rosso e de Siqueira pode ser feita verificando a frequência conjunta dos diagnósticos como nas tabelas 5.3.3 e 5.3.4.

TABELA 5.3.3: Frequência dos diagnósticos das curvas de Siqueira(V7) e Rosso(V8) na primeira consulta:

		V8			total
		DES	NOR	OBS	
V7	DES	77	4	0	81
	NOR	99	261	12	372
	OBS	0	6	106	112
	total	176	271	118	565

Na tabela 5.3.3 os valores fora da diagonal se referem à incompatibilidade de diagnóstico, chamando a atenção os 99 casos em que V7 diagnostica como NOR e V8 diagnostica como DES. Analisando os totais da tabela 5.3.3, nota-se que esses 99 casos contribuem para a assimetria da distribuição da curva de Rosso na primeira consulta(V8), vista nas figuras 5.1.1 e 5.1.2, e principalmente para a baixa frequência de diagnósticos NOR pela curva de Rosso.

TABELA 5.3.4: Frequência conjunta dos diagnósticos das curvas de Siqueira(V12) e Rosso(V13) na última consulta:

		V13			total
		DES	NOR	OBS	
V12	DES	45	0	0	45
	NOR	187	184	15	386
	OBS	0	13	121	134
	total	232	197	136	565

Nota-se na tabela 5.3.4 que por ocasião da última consulta a discordância é maior do que qualquer concordância, sendo que a maior

incompatibilidade no diagnóstico continua ocorrendo quando a curva de Rosso diagnostica gestantes como DES as que a curva de Siqueira diagnostica como NOR, com maior intensidade na última consulta do que na primeira consulta. Ficando agora claro que a imprecisão vista na curva de Rosso quando diagnostica a gestante como DES levando aos box-plots maiores incluindo os box-plots do diagnóstico NOR, se refere a boa parte dessas 187 gestantes que Siqueira considera NOR. A assimetria que ocorre principalmente no total de V13 já foi vista nos gráficos das figuras 5.1.4 e 5.1.5 e analisada no comentário das tabelas 5.3.1 e 5.3.2

Assim, à medida que passa o tempo de gestação a curva de Rosso tende a aumentar o limiar superior de corte para DES, sub-classificando a nutrição da gestante. Os mesmos resultados ocorreram quando considerou-se apenas a sub-amostra FSP.

COMPARAÇÃO ENTRE AS CURVAS COM RELAÇÃO AO PROGNÓSTICO DA ADEQUAÇÃO DO PESO DO RN:

Outra forma de comparar as duas curvas quanto ao “poder” das mesmas em estimar o tamanho do RN, é usando a Adequação do peso do RN(V19) como medida a ser prevista. A partir da figura 5.3.5 se fará um estudo comparativo dessas duas curvas com essa finalidade.

COMPARAÇÃO ENTRE OS DIAGNÓSTICOS DAS CURVAS E V19:

V7=Curva de Siqueira na 1a.consulta, V8=Curva de Rosso na 1a.consulta, V12=Curva de Siqueira na última consulta, V13=Curva de Siqueira na última consulta, V19=Adequação do peso do RN

Para comparação, os valores referentes à adequação do peso do RN, foram colocados segundo a previsão lógica dos diagnósticos, a saber: PIG com DES, AIG com NOR e GIG com OBS. O caso em que a gestante é classificada como OBS e o seu recém-nascido é classificado como PIG, será considerado erro de classificação, embora existam relatos de associação entre obesidade da gestante e retardo do crescimento fetal, como por exemplo em casos de problemas placentários conforme os trabalhos de NAEYE(1079) e de SIMPSON e col(1975) citados por SIQUEIRA(1981) e de AUBRY e col(1975) citado por WORTHINGTON e col(1980).

Nota-se pela figura 5.3.5 que:

- a) - a curva com a frequência de diagnóstico DES mais próxima da adequação PIG é a curva de Siqueira na última consulta, V12, proporcionalmente uma diferença de $0,025(45/565-31/565)$, seguida da curva de Siqueira na primeira consulta, V7. A tendência da curva de Rosso em diagnosticar mais como desnutrida do que deveria, pode ser confirmada ao se notar que a adequação

do peso do RN está com uma frequência bem abaixo de todas as curvas, e que a frequência das curvas de Rosso para DES é a mais discrepante de V19.

b) - com relação à adequação do peso do RN AIG, a curva de Siqueira na última consulta apresenta uma frequência mais próxima, proporcionalmente 0,120 de diferença, seguida da curva de Siqueira na primeira consulta. Novamente a curva de Rosso mostra frequências mais distantes da adequação do peso do RN.

c) - na estimação da adequação do peso do RN igual a GIG, a curva de Siqueira na primeira consulta apresenta uma frequência mais próxima, com uma diferença proporcional de 0,056, seguida da curva de Rosso na primeira consulta. Neste caso o diagnóstico da curva de Siqueira, na última consulta, não apresentou a mesma performance dos casos anteriores.

Esses resultados foram confirmados ao calcular a medida de aderência $\sum_1^3 \frac{(v_i - V19)^2}{V19}$ onde v_i é a frequência observada da curva, no estado nutricional $i=DES, NOR$ e OBS , e $V19$ é a frequência observada na adequação do peso do RN; como pode ser verificado pelos valores a seguir:

curva	v7(Siq. 1a.cons.)	v8(Ross.1a.cons)	v12(Siq.ult.cons)	v13(Ros.ult.cons)
aderência	108,25	769,00	52,95	1487,68

Considerando apenas a sub-amostra de FSP notou-se que proporcionalmente a relação encontrada na figura 5.3.5 é a mesma para ambas as amostras.

Como o objetivo principal no diagnóstico do estado nutricional da gestante é prever o tamanho do bebê para que se possa corrigir o peso com suplementação alimentar, o estudo do cruzamento da Adequação do Peso do

RN com o diagnóstico das curvas a seguir, pode contribuir com informações sobre a capacidade de previsão de cada curva.

TABELA 5.3.5: Distribuição conjunta dos diagnósticos da curva de Siqueira na primeira consulta(V7) e Adequação(V19):

		V7			total	Proporção de acertos
		DES	NOR	OBS		
V19	PIG	15	13	<u>3</u>	31	15/31=0,4839
	AIG	61	324	69	454	324/454=0,7136
	GIG	<u>5</u>	35	40	80	40/80=0,5
	total	81	372	112	565	
Proporção de acertos		0,0617	0,8709	0,3571		

Os valores assinalados com negrito e sublinhado se referem a um alto grau de má classificação ao usar a curva de Siqueira para o diagnóstico na primeira consulta, perfazendo um total de 8 ou 1.4% dos casos. Quando a gestante foi classificada como OBS e seu bebê foi classificado como PIG, considerou-se que existia má-classificação, embora, como foi citado, existam relatos na literatura de associação entre obesidade da gestante e retardo do crescimento fetal, como no exemplo de problemas placentários. A coluna e a linha proporção de acertos, consideram a categoria da variável V19 como o alvo para prognosticar. A coluna proporção de acertos fornece para os RN classificados naquela categoria de adequação do peso, a chance de ter sido previsto pelo diagnóstico da curva, enquanto a linha proporção de acertos, fornece para as gestantes diagnosticadas naquela categoria de estado nutricional, a chance do RN vir a ter a adequação prevista. Nota-se maior acerto em prever AIG a partir do diagnóstico do estado

nutricional NOR(0,8709), e maior acerto para os bebês com adequação do peso AIG(0,7136) seguida de PIG(0,4839). O coeficiente Gamma de concordância para a tabela 5.3.5 foi de 0.599, e o de contingência foi de 0.358.

TABELA 5.3.6: Distribuição dos diagnósticos da curva de Rosso na primeira consulta(V8) e Adequação(V19):

		V8			total	Proporção de acertos
		DES	NOR	OBS		
V19	PIG	17	11	3	31	17/31=0.5484
	AIG	147	230	77	454	230/454=0.5066
	GIG	12	30	38	80	38/80=0.475
total		176	271	118	565	
Proporção de acertos		0,0965	0,8487	0,3220		

Assinalando da mesma forma os valores com alto grau de má classificação, no diagnóstico usando a curva de Rosso na primeira consulta, obtém-se um total de 15 ou 2.65% dos casos, que é aproximadamente o dobro de erro da curva de Siqueira. Continuando a usar a proporção de acertos, considerando a categoria da variável V19 como o alvo para prognosticar, nota-se que para o diagnóstico do estado nutricional da gestante o maior acerto está em prever AIG com o estado NOR(0,8487), seguindo de GIG com OBS(0,3220); enquanto para as adequações dos RN, o maior acerto está em PIG de ter sido previsto por DES(0,5484) e menor acerto em GIG em ter sido previsto por OBS(0,475). O coeficiente Gamma de concordância para a tabela 5.3.6 foi de 0.491, e o de contingência foi de 0.281.

Comparando-se as tabelas 5.3.5 e 5.3.6 com relação a provável atenção especial devido ao diagnóstico de DES comparativamente ao número de recém-nascidos PIG, temos:

Curva de Siqueira na primeira consulta

com 81 gestantes diagnosticadas como DES sendo
 15 eram realmente PIG e receberam atenção especial através da mãe DES
 66 não eram PIG e a mãe recebeu atenção especial como DES
 e 16 nasceram PIG cujas mães não receberam atenção especial (NOR ou OBS)

Curva de Rosso na primeira consulta

com 176 gestantes diagnosticadas como DES sendo
 17 eram realmente PIG e receberam atenção especial através da mãe DES
 159 não eram PIG e a mãe recebeu atenção especial como DES
 e 14 nasceram PIG cujas mães não receberam atenção especial (NOR ou OBS)

TABELA 5.3.7: Distribuição dos diagnósticos da curva de Siqueira na última consulta(V12) e Adequação(V19):

		V12				
		DES	NOR	OBS	total	Proporção de acertos
V19	PIG	12	15	<u>4</u>	31	12/31=0,3871
	AIG	32	333	89	454	333/454=0,7335
	GIG	<u>1</u>	38	41	80	41/80=0,5125
	total	45	386	134	565	
proporção de acertos		0,2666	0,8627	0,3059		

Os valores assinalados com negrito e sublinhado se referem a um alto grau de má classificação ao usar a curva de Siqueira para o diagnóstico na última consulta, perfazendo um total de 5 ou 0.9% dos casos. Nota-se que para o diagnóstico do estado nutricional da gestante o maior acerto está em prever AIG com o estado NOR(0,8627), seguindo de GIG com OBS(0,3059); enquanto para as adequações dos RN, o maior acerto é para os bebês AIG de ter sido

previsto por NOR(0,7335) e menor acerto nos PIG em ter sido previsto por DES(0,3871).. O coeficiente Gamma de concordância para a tabela 5.3.7 foi de 0,606, e o de contingência foi de 0,353.

TABELA 5.3.8: Distribuição dos diagnósticos da curva de Rosso na última consulta(V13) e Adequação:

		V13			total	Proporção de acertos
		DES	NOR	OBS		
V19	PIG	20	6	<u>5</u>	31	20/31=0,645
	AIG	196	167	91	454	167/454=0,3678
	GIG	<u>16</u>	24	40	80	40/80=0,50
	total	232	197	136	565	
proporção de acertos		0,0862	0,8477	0,2941		

Novamente foram assinalados com negrito e sublinhado os casos com um alto grau de má classificação ao usar a curva de Rosso para o diagnóstico na última consulta, perfazendo um total de 21 ou 3.72% dos casos. Nota-se que para o diagnóstico do estado nutricional da gestante o maior acerto está em prever AIG com o estado NOR(0,8477), seguindo de GIG com OBS(0,2941); enquanto para as adequações dos RN, maior acerto em PIG de ter sido previsto por DES(0,645) e menor acerto em AIG em ter sido previsto por NOR(3678).. O coeficiente Gamma para a tabela 5.3.8 foi de 0,467, e o de contingência foi de 0,265.

Comparando-se as tabelas 5.3.7 e 5.3.8 com relação a provável atenção especial devido ao diagnóstico de DES comparativamente ao número de recém-nascidos PIG, temos:

Curva de Siqueira na última consulta

com 45 gestantes diagnosticadas como DES sendo
12 eram realmente PIG e receberam atenção especial através da mãe DES
33 não eram PIG e a mãe recebeu atenção especial como DES
e 19 nasceram PIG cujas mães não receberam atenção especial (NOR ou OBS)

Curva de Rosso na última consulta

com 232 gestantes diagnosticadas como DES sendo
20 eram realmente PIG e receberam atenção especial através da mãe DES
212 não eram PIG e a mãe recebeu atenção especial como DES
e 11 nasceram PIG cujas mães não receberam atenção especial (NOR ou OBS)

Comparando-se as porcentagens de má-classificação das tabelas 5.3.5, 5.3.6, 5.3.7 e 5.3.8 nota-se claramente uma vantagem da curva de Siqueira no sentido de classificar com menos erro, no sentido de, a partir do diagnóstico do estado nutricional da gestante, prever o estado físico do futuro recém-nascido. Comparando-se a “força de previsão”, dada pelas proporções de acertos, nota-se que dado o estado nutricional pela curva,

- para prever PIG a curva de Siqueira na última consulta(V12) é a melhor,
- para prever AIG, a curva de Siqueira na primeira consulta(v7) é a melhor,
- para prever GIG, a curva de Siqueira na primeira consulta(v7) é a melhor.

Por outro lado, comparando-se as proporções de acertos associadas a sensibilidade da curva, nota-se que dada a adequação do peso do RN,

- entre os PIG, a curva de Rosso na última consulta(v13) é a melhor,
- entre os AIG, a curva de Siqueira na última consulta(v12) é a melhor,
- entre os GIG, a curva de Siqueira na última consulta(v12) é a melhor.

Usando as estatísticas Phi e V de Cramér, Coeficiente de Contingência e Qui-quadrado de Pearson, Tau de Kendall, Gamma, Somer's d, para

as tabelas 5.3.5, 5.3.6, 5.3.7 e 5.3.8, todas estatísticas indicaram maior associação da variável V7 com V19 do que V8 com V19, e da variável V12 com V19 do que V13 com V19, o que aponta novamente para a curva de Siqueira como tendo maior poder de previsão. Para apenas a sub-amostra FSP também tais estatísticas mostraram a mesma vantagem da curva de Siqueira.

Seguem-se duas tabelas que mostram a distribuição completa do cruzamento entre as variáveis V19, V7 e V8, e entre as variáveis V19, V12 e V13.

TABELA 5.3.9: Distribuição conjunta das variáveis V19, V7 e V8:

	Adequação do Peso do Recém-Nascido (V19)						Sub-total	
	A I G		G I G		P I G		n	%
	n	%	n	%	n	%		
V7=DES								
V8= DES	60	10.6%	4	0.7%	13	2.3%	77	13.6%
= NOR	1	0.2%	1	0.2%	2	0.4%	4	0.7%
Sub-total	61	10.8%	5	0.9%	15	2.7%	81	14.3%
V7=NOR								
V8= DES	87	15.4%	8	1.4%	4	0.7%	99	17.5%
= NOR	226	40.0%	26	4.6%	9	1.6%	261	45.2%
= OBS	11	1.9%	1	0.2%	0	0	12	2.1%
Sub-total	324	57.3%	35	5.2%	13	2.3%	372	65.8%
V7=OBS								
V8= NOR	3	0.5%	3	0.5%	0	0	6	1.1%
OBS	66	11.7%	37	6.5%	3	0.5%	106	18.8%
sub-total	69	12.2%	40	7.1%	3	0.5%	112	19.8%

Os valores assinalados com um círculo representam um alto grau de má-classificação devido a V7 e a V8 ao mesmo tempo, enquanto os assinalados com um quadrado representam alto grau de má-classificação apenas por V7 ou por V8. O valor 87 da casela V19=AIG, V7=NOR, V8=DES, assinalado com representa a má classificação que induz à assimetria da distribuição de V8. Nota-se que a coluna onde V19 é GIG, é a que apresenta mais erros de classificação, indicando a mesma tendência verificada nas tabelas 5.3.3 e 5.3.6, de V8 sub-classificar, “puxando” para DES o que é GIG e o que é AIG.

TABELA 5.3.10 - Distribuição conjunta das variáveis V19, V12 e V13:

	Adequação do Peso do Recém-Nascido (V19)						Sub-total	
	A I G		G I G		P I G		n	%
	n	%	n	%	n	%		
V12=DES								
V13=DES	32	5.7%	1	0.2%	12	2.1%	45	8.0%
Sub-total	32	5.7%	1	0.2%	12	2.1%	45	8.0%
V12=NOR
V13=DES	164	29.0%	15	2.7%	8	1.4%	187	33.1%
= NOR	158	28.0%	20	3.5%	6	1.1%	184	32.6%
= OBS	11	1.9%	3	0.5%	1	0.2%	15	2.7%
Sub-total	333	58.9%	38	6.7%	15	2.7%	386	68.3%
V7=OBS								
V8= NOR	9	1.6%	4	0.7%	0	0	13	2.3%
OBS	80	14.2%	37	6.5%	4	0.7%	121	21.4%
sub-total	89	15.8%	41	7.3%	4	0.7%	134	23.7%

As assinalações são as mesmas usadas da tabela 5.3.9, e a distribuição dos valores nas caselas das tabelas 5.3.9 e 5.3.10 é bem parecida. Pode-se concluir pelas má-classificações de médio e alto grau, que as curvas de Rosso têm o vício de “puxar” para baixo não só o que a curva de Siqueira considera NOR e OBS, mas também quando o bebê tem a adequação de peso AIG e GIG, tanto com relação ao diagnóstico feito na primeira como na última consulta.

Algumas suposições foram levantadas como causa desse comportamento da curva de Rosso, tal como o uso do peso pré-gestacional como padrão, a inclinação da reta usada por Rosso para indicar o padrão, a população usada por Rosso como padrão ser de uma região diferente; mas o que tudo indica pela análise feita até aqui, é que o padrão dado pelo nomograma da curva de Rosso tem um ponto de corte alto ou a população usada para construir a sua curva era mais bem nutrida do que a população aqui estudada. Analisando outra população como a do presente estudo sob esse prisma, a curva de Rosso tende a considerar as gestantes normais como desnutridas e as obesas como normais.

As porcentagens de má classificação das tabelas 5.3.5, 5.3.6, 5.3.7 e 5.3.8, são proporções de erros máximos de classificação com relação ao total de possibilidade (565), independentemente do número de bebês de cada categoria da variável Adequação do Peso(V19).

Se for considerada a proporção de erros máximos de classificação com relação ao total de cada categoria da variável V19, a taxa de erros máximos ficaria:

Para Siqueira na primeira consulta (tabela 5.3.5): $5/80+3/31 = 0,159$

Para Rosso na primeira consulta	(tabela 5.3.6):	$12/80+3/31 = 0,247$
Para Siqueira na última consulta	(tabela 5.3.7):	$1/80+4/31 = 0,141$
Para Rosso na última consulta	(tabela 5.3.8):	$16/80+5/31 = 0,361$

que no fundo “tentam” estimar a probabilidade do tipo:

$$P[\text{classificar como não X dado que a Adequação é X}]$$

a qual é o complementar de $P[\text{classificar como X dado que a Adequação é X}]$, que por seu lado teria as características da sensibilidade descrita na seção 4.2.1, se X fosse uma de duas possíveis categorias de uma variável dicotômica.

Assim, das taxas de erros máximos pode-se obter uma idéia aproximada da sensibilidade para não cometer o erro máximo por:

Para Siqueira na primeira consulta	(tabela 5.3.5):	$1 - 0,159 = 0,841$
Para Rosso na primeira consulta	(tabela 5.3.6):	$1 - 0,247 = 0,753$
Para Siqueira na última consulta	(tabela 5.3.7):	$1 - 0,141 = 0,859$
Para Rosso na última consulta	(tabela 5.3.8):	$1 - 0,361 = 0,639$

o que dá uma vantagem para a curva de Siqueira na última consulta e em segundo lugar para a curva de Siqueira na primeira consulta.

Usando as tabelas 5.3.3 a 5.3.10, e dicotomizando as variáveis Adequação do Peso do RN e Diagnóstico de cada curva, calculou-se a sensibilidade e a especificidade de cada curva, considerando as definições da seção 4.2.1, obtendo-se os resultados da tabela 5.3.11

TABELA 5.3.11: Sensibilidade e Especificidade das curvas de Siqueira e de

Rosso:

		SENSIBILIDADE		ESPECIFICIDADE		VALOR PREDITIVO	
		SIQUEIRA	ROSSO	SIQUEIRA	ROSSO	SIQUEIRA	ROSSO
primeira consulta	NORp/ AIG	0,714	0,506	0,568	<u>0,631</u>	0,871	0,848
	DES p/ PIG	0,484	<u>0,566</u>	0,876	0,702	0,185	0,096
	OBS p/ GIG	0,500	0,476	0,876	0,837	0,357	0,322
última consulta	NORp/ AIG	0,733	0,368	0,522	<u>0,730</u>	0,862	0,848
	DES p/ PIG	0,387	<u>0,645</u>	0,938	0,603	0,273	0,086
	OBS p/ GIG	0,512	0,500	0,809	0,802	0,306	0,294

Para as duas consultas, nos 18 possíveis cálculos de sensibilidade, especificidade e valor preditivo, houve uma vantagem da curva de Rosso na razão de 4 para 18. Onde ocorreu vantagem da curva de Rosso, DESp/PIG para sensibilidade, e NORp/AIG para a especificidade, o valor está assinalado com itálico, negrito e sublinhado. Com relação a previsão de PIG, a curva de Rosso se mostrou ligeiramente mais sensível, a de Siqueira se mostrou mais específica e com maior valor preditivo. Tais resultados indicam uma vantagem da curva de Siqueira sobre a de Rosso, principalmente quanto a economia de recursos de sua aplicação em saúde pública.

Os gráficos 5.3.5 e 5.3.6 mostram um fato interessante quando considera-se o peso do RN em apenas duas categorias, até 2500g e 2500g ou mais. Para o primeiro gráfico em um total de 36 recém-nascidos com pesos até 2500g vê-se que Rosso na última consulta(V13) é mais sensível, sensibilidade $27/36=0,75$, com o diagnóstico DES.

GRÁFICOS INDICANDO A FORÇA DE PREVISÃO DAS CURVAS, COM 2500g COMO PONTO DE CORTE

V7=Siqueira na 1a.cons., V8=Rosso na 1a.cons., V12=Siqueira na últ.cons., V13=Rosso na últ.cons.

Para o gráfico 5.3.6 tem-se um total de 529 recém-nascidos com peso acima de 2500g e vê-se que Siqueira na última consulta é mais específica, com especificidade $(365+131)/529=0,938$, estando ainda próximo à curva de Siqueira na primeira consulta.

A defasagem dos resultados em termos de especificidade e sensibilidade com relação aos resultados anteriores se deve à variável a ser prognosticada, à forma como ela é dicotomizada, e ao fato que devem existir outras variáveis interagindo sendo desprezadas em uma tabela de dupla entrada com fatores

dicotômicos como nesse caso. Motivo pelo qual essas medidas não serão usadas para darem a palavra final, mas apenas para contribuírem para reforçar ou não um resultado de uma análise já feita como, por exemplo, as taxas de má classificação encontradas nas tabelas 5.3.5 a 5.3.10.

5.3.2 - Utilização de Regressão Linear por “Stepwise”:

Na procura da curva que melhor prognostique o tamanho do RN, considerando o peso do RN(v17) como variável dependente, e as demais variáveis, excluindo os métodos de Rosso e de Siqueira, como independentes, procuraram-se as variáveis associadas mais significativamente ao peso do RN, primeiramente aplicando uma regressão linear múltipla por “stepwise”, usando o nível de significância do teste F para entrada da variável no modelo de 0.05, e de remoção da variável do modelo de 0.10. Os resultados foram os seguintes:

passo	variável incluída	b	valor de T	valor p
1	V10	0,186	10.05	0,0000
2	V15	7,797	5,488	0,0000
3	V6	-6.528	-2,063	0,0397
	(Constante)	-53.333	-0,131	0,8962

$$\text{PESO RN} = -53,333 + 0,186.V10 + 7,797.V15 - 6,528.V6$$

Novamente, como esperado, a variável peso na última consulta(V10) foi considerada como a mais influente sobre o peso do RN(V17). Em segundo lugar a variável mais significante foi V15(duração da gestação em dias) e em terceiro lugar v6(tempo de gestação na primeira consulta).

Em um segundo ajuste por “stepwise”, foi considerando o peso do RN(V17) como variável dependente, e as demais variáveis, incluindo os métodos de Rosso e de Siqueira, como independentes, usando uma codificação onde DES tinha peso 1, NOR tinha peso 2 e OBS tinha peso 3, com o nível de significância do teste F para entrada da variável no modelo de 0,05, e de remoção da variável do modelo de 0,10. Os resultados foram os seguintes:

passo	variável incluída	b	valor de T	valor p
1	V10	0,1511	5,039	0,0000
2	V15	7,8454	5,572	0,0000
3	V5	-19,617	-2,127	0,0339
4	V7	211,232	3,880	0,0001
	(Constante)	412,675	,237	0,8131

$$\text{PESO RN} = 130,567 + 0,15V10 + 7,98.V15 + 211,69V7 - 4,989V9$$

Portanto, o modelo mantém as variáveis V10(peso na última consulta) e V15(duração da gestação em dias), entrando agora V5(peso na primeira consulta) e V7(estado nutricional segundo Siqueira na primeira consulta). A única variável referente às curvas em comparação a ser incluída no modelo linear foi a a curva de Siqueira na primeira consulta(V7). Embora a prioridade não seja o prognóstico apenas do peso do RN, também para esse caso a curva de Siqueira se destaca, indicando possível vantagem dessa curva, o que já foi encontrado principalmente nas tabelas 5.3.1 e 5.3.2 e figuras 5.3.1 a 5.3.4.

No entanto se a entrada das variáveis V10(peso na última consulta), V15(duração da gestação) e V5(peso na primeira consulta) no modelo

antes de V7(diagnóstico de Siqueira na primeira consulta) for confirmada, significa que essas variáveis são importantes para o prognóstico do peso do RN e que o diagnóstico de Siqueira na primeira consulta(V7) sozinho não traz toda informação sobre o peso do RN.

Considerando apenas a sub-amostra FSP, a regressão linear ajustada por “stepwise” foi:

$$\text{PESO RN} = 61,150 + 0.123.V10 + 7.40.V15 + 144,20V7$$

que não incluiu a variável V5(peso na primeira consulta) como na amostra conjunta, mas confirmando a superioridade da curva de Siqueira na primeira consulta(V7).

5.3.3 - Utilização de Regressão Logística Acumulada:

Na procura da curva que melhor prognostique o tamanho do RN, considerou-se como variável dependente a adequação do peso do RN(V19), e os estados nutricionais segundo Siqueira V7, V8, e segundo Rosso V12 e V13 juntamente com as demais variáveis, como candidatas a variáveis independentes do modelo logístico multinomial; em um processo “stepwise”. O ajuste do modelo logístico multinomial, usando o programa do SAS conforme apêndice A4.3, incluiu as variáveis no modelo na seguinte ordem:

passo	variável incluída	b	Qui-quadrado	valor p
1	V10	-0,109	10,742	0,0010
2	V16	0,249	15,166	0,0001
3	V7	-0,815	7,716	0,0006
	(constante 1)	-7,777	9,112	0,0025
	(constante 2)	-2,484	0,967	0,3254

Portanto a única variável referente às curvas em comparação a ser incluída no modelo logístico foi a V7(estado nutricional segundo Siqueira na primeira consulta), reafirmando o encontrado nas seções 5.3.1, 5.3.2 e 5.3.3, não apenas a vantagem da curva de Siqueira, mas também começando a se destacar a variável V7 que é um diagnóstico precoce, já na primeira consulta.

No pêndice A5.3.3 verifica-se que considerando apenas a sub-amostra FSP, a curva de Siqueira na primeira consulta volta a ser a variável mais significativa para prever a adequação do peso do RN.

5.3.4 - Aplicação de Regressão e Classificação por Árvore:

Considerando a variável dependente ora como sendo o peso do RN, ora a adequação do peso do RN, foram utilizados os programas apresentados no capítulo 4, para construir as árvores como segue.

É bom lembrar que, como foi visto no comentário da figura 5.1.25, a variável peso do RN sozinha não informa sobre o crescimento e desenvolvimento do conceito, pelo fato de existir uma interação com a variável idade gestacional. Por esse motivo as árvores com o peso do RN como variável dependente servirão apenas para descrição de fatos, uma vez que o prognóstico, por exemplo, um peso do RN(V17) de 2600g apenas, não significa que esse bebê é pequeno ou normal, enquanto as árvores com adequação do peso do RN(V19) como variável dependente servirão como descrição, modelo, e para comparações.

Foram ajustadas árvores de classificação quando os dados eram só da FSP, e os resultados, quanto à posição das curvas em comparação, na ordem de hierarquia das variáveis dessas árvores, são os mesmos, conforme se vê no apêndice A5.3.4:

I) - Utilizando o CHAID:

Árvore com V17(peso do RN) como a variável dependente

FIGURA 5.3.8

V47=peso do RN/100, V40=peso na última consulta/200, V18=Sexo do RN, V16=duração da gestação em semanas, V20=tipo do parto, V35=peso na primeira consulta/200, V7=C.Siqueira na 1a.consulta, V18=Sexo do RN. (parâmetros: depth=5, Before=After=5, Merge Level=0.05, Eligibility Level=0.05)

Nota-se que a variável V40, equivalente a peso na última consulta, aparece em primeiro lugar, seguida de V18(sexo), V16(duração da gestação) e V20(tipo de parto) em segundo lugar, V35, equivalente a V5, (duração da gestação) e V7(Siqueira na primeira consulta) em quarto lugar, terminando com V18(sexo) em quinto lugar. Por essa figura vê-se que o diagnóstico precoce de

Siqueira é também o mais significativo das curvas, na previsão do peso do RN, confirmando o encontrado na regressão por “stepwise”.

Os ramos da árvore da figura 5.3.8 são representados na seguinte ordem:

Tabela 5.3.12 Quadro dos detalhes dos ramos da árvore ajustada na figura 5.3.8

Ramo	Variáveis e Partições	Média do Peso do RN em gramas	ordem do ramo
1	V10<53kg	2715	1
2	(53<=V10<61)&(V18=f)	3009	4
3	(53<=V10<61)&(V18=m)&(V5<51kg)	3294	8
4	(53<=V10<61)&(V18=m)&(V5>=51kg)	2993	3
5	(61<=V10<73)&(V16<37)	2989	2
6	(61<=V10<73)&(V16>=37)&(V7=DESouNOR)&(V18=f)	3168	5
7	(61<=V10<73)&(V16>=37)&(V7=DESouNOR)&(V18=m)	3278	7
8	(61<=V10<73)&(V16>=37)&(V7=OBS)	3495	10
9	(73<=V10<91)&(V20=C ou F)	3467	9
10	(73<=V10<91)&(V20=N)	3262	6
11	(V10>=91)	3900	11

O comando “Gains-chart” ordena os ramos informando aquele que leva a um recém-nascido mais pesado ou mais leve, essa informação está na coluna ordem do ramo acima. Assim a ramificação que obtém maior peso do RN é a de número 11, e a que obtém o menor peso do RN é a de número 1.

Como pode ser visto na figura 5.3.9, considerando como variável dependente a Adequação do peso do RN(V19) o CHAID escolheu em primeiro lugar a variável V7(Siqueira na primeira consulta), seguida de V40(ou V10=peso na última consulta), V16(duração da gestação em semanas) e V20(tipo de parto) e posteriormente V12(Siqueira na última consulta), indicando uma vantagem

da curva de Siqueira sobre a curva de Rosso, agora não só na primeira consulta(V7) como também na última consulta(V12).

Utilizando o CHAID: Árvore com V19(adequação do peso do RN) como a variável dependente

FIGURA 5.3.9

V19=Adequação do peso do RN, V7=C.Siqueira na 1a.consulta V40=peso na última consulta/200, V16=duração da gestação em semanas, V20=tipo do parto, V12=C. Siqueira na última consulta. (depth=5, Before=After=5, Merge Level=0.05, Eligibility Level=0.05)

Na tabela 5.3.13 tem-se as ramificações, as respectivas classificações e o número de ordem das mesmas, agora em função das ponderações dadas para a adequação com ponderação de PIG=1, ponderação de AIG=2, e

ponderação de GIG=3. Assim a ramificação que leva a crianças mais pesadas é a de número 3 e a ramificação que leva a crianças mais leves é a de número 1.

Tabela 5.3.13 : Quadro dos detalhes dos ramos da árvore ajustada na figura 5.3.9

Ramo	Variáveis e Partições	Classificação	ordem do ramo
1	(V7=DES)&(V10<53)	AIG	1
2	(V7=DES)&(V10>=53)	AIG	4
3	(V7=NOR)&(V16<=33)	GIG	8
4	(V7=NOR)&(34<V16<38)	AIG	5
5	(V7=NOR)&(V16>=39)&(V12=DES)	AIG	2
6	(V7=NOR)&(V16>=39)&(V12=NOR ou OBS)	AIG	3
7	(V7=OBS)&(V20=C)	AIG	7
8	(V7=OBS)&(V20=F ou N)	AIG	6

II) - Utilizando o CART:

Árvore com V17(peso do RN) como a variável dependente

V14=C.Rosso na última consulta, V5=peso na 1a. consulta, V11=tempo de gestação na última consulta, V6=tempo de gestação na 1a.consulta, V18=sexo do RN, V9=escala de Rosso na 1a.consulta, V15=duração da gestação em dias, V4=peso pré-gestacional, V2=local, V3=estatura ($\beta=0,015$ e índice de Gini)

Para o CART a primeira variável é a V14(escala de Rosso na última consulta), diferenciando bastante do CHAID, com relação à classificação das curvas em comparação. Além disso a variável V9(escala de Rosso na primeira consulta) aparece no degrau 2 partindo o nó número 6, e no degrau 3 partindo o nó número 10. As ramificações assinaladas com c10 são as que o CART prevê maiores pesos do RN, isto é, acima de 4500 gramas, e as ramificações com c1 o CART prevê menores pesos, isto é, abaixo de 2100 gramas.

Esses resultados dão vantagem para a curva de Rosso e, não coincidem com os resultados anteriores, o que deixa uma ressalva quanto à sua validade, precisão e confiabilidade. Os próximos algoritmos deverão contribuir para que se possa reforçar ou não esse resultado.

Arvore com V19(adequação do peso do RN) como a variável dependente

FIGURA 5.3.11

V19=Adequação do peso do RN, V7=C.Siqueira na 1a.consulta V5=peso na 1a.consulta, V16=duração da gestação em semanas, V11=tempo de gestação na última consulta em semanas, V9=escala de Rosso na primeira consulta, V3=estatura, V5=peso na primeira consulta. ($\beta=0,015$ e índice de Gini)

Considerando como variável dependente a variável adequação do peso do RN(V19), a técnica CART escolheu em primeiro lugar a curva de Siqueira na primeira consulta(V7) seguida do peso na primeira

consulta(V5), duração da gestação em semanas(V16), tempo de gestação na última consulta(V11), estatura da gestante(V3) e peso na primeira consulta(V5), sem escolher a curva de Rosso em qualquer consulta. Os nós assinalados com c1 significam que o seu ramo classifica a gestante como futura mãe de um bebê com adequação do peso PIG, os ramos com c2 classificam a gestante como futura mãe de um bebê com AIG. Tal resultado é diferente do encontrado usando a mesma técnica mas com a variável dependente peso do RN, indicando uma certa inconsistência, ou da técnica ou devido ao fato da mudança de variável dependente. Se houvesse sempre mudança devido à variável dependente, os resultados seriam inconsistentes e inconfiáveis.

III) - Utilizando o SPLUS :

As figuras 5.3.12 e 5.3.13 apresentam os resultados da aplicação da técnica SPLUS utilizando respectivamente como variável dependente o peso do RN e a adequação do peso do RN.

Arvore com V17(peso do RN) como a variável dependente

FIGURA 5.3.12

V7=C.Siqueira na 1a.consulta V10=peso na última consulta, V15=duração da gestação em dias, V16=duração da gestação em semanas, V18=sexo do RN,, PESOG=V10-V4. Parâmetro de parada k=9.

Arvore com V19(adequação do peso do RN) como a variável dependente

V19=Adequação do peso do RN, V40=peso na última consulta/200, V6=tempo de gestação na 1a. consulta, V7=C.Siqueira na 1a.consulta, V4=peso pré-gestacional, V20=tipo do parto, V45=(duração da gestação em dias)/5, V16=duração da gestação em semanas. Parâmetro de parada k=9.

Usando a variável peso do RN como a variável dependente, na figura 5.3.12 o SPLUS escolheu em primeiro lugar a curva de Siqueira na primeira consulta(V7), seguida de peso da gestante na última consulta(V10) e duração da gestação em dias(V15). Nos patamares seguintes, não apareceu o diagnóstico de Rosso, indicando vantagem da curva de Siqueira.

Para o caso em que a variável dependente é a adequação do peso do RN(V19), na figura 5.3.13, novamente a curva de Siqueira na primeira consulta(V7) volta a se destacar, não aparecendo o diagnóstico de Rosso, o que confirma os resultados do CHAID e das seções 5.3.1 e 5.3.2.

Nessa técnica não houve mudança grande como no caso da CART.

A SPLUS tem algumas vantagens em termos de opções gráficas como a da figura 5.3.14, onde para cada ramificação tem-se um histograma do peso do RN(V17) indicando o comportamento ambíguo dessa variável para as diferentes classificações de adequação do peso do RN; embora haja coerência quanto a forma da distribuição dos pesos dos RN; pois para os ramos da direita (peso na última consulta > 55kg (27.5x200), V7=OBS), a distribuição dos pesos dos RN, tende a estar mais concentrada em valores altos, enquanto nos ramos da esquerda, (V40 < 27.5 e V7=DES ou NOR), a distribuição de pesos dos RN tende a estar mais concentrada em valores baixos.

V19=Adequação do peso do RN, V40=peso na última consulta/200, V7=C.Siqueira na 1a.consulta, V20=tipo do parto, V45=(duração da gestação em dias)/5, V6=tempo de gestação na 1a.consulta

IV) - Utilizando o KS:

Resultados semelhantes voltam a aparecer na figura 5.3.15, onde a variável peso na última consulta(V10) está relacionada em primeiro lugar com o Peso do RN(V17), a curva de Siqueira aparece no terceiro degrau da árvore, com o diagnóstico da última consulta(V12), e no quarto degrau com o diagnóstico da primeira consulta(V7), enquanto a curva de Rosso não apareceu em nenhum degrau.

Quando a variável dependente é a adequação do peso do RN(V19), a figura 5.3.16 mostra novamente a curva de Siqueira na primeira consulta(V7) em primeiro lugar, seguida de peso na primeira consulta(V5) e da curva de Siqueira na última consulta(V12) no segundo degrau. A curva de Rosso não apareceu em nenhum degrau.

A menos do caso da técnica CART com o peso do RN(V17) como variável dependente a curva de Siqueira apareceu com vantagem sobre a de Rosso, nos outros casos algumas vezes aparecendo ambas as curvas e em outras apenas a de Siqueira.

Mas ainda poderia existir a possibilidade de que as variáveis que entram no modelo podem estar substituindo o efeito de V8 e V13(a curva de Rosso na primeira e última consulta), embora isso não tenha ocorrido com a curva de Siqueira. Por outro lado, pode ainda existir uma pequena possibilidade que a ambigüidade da técnica CART se deve a uma alteração normal em função da estrutura da amostra, talvez devido a uma alteração circunstancial ou sazonal na mesma. Uma primeira alteração seria imaginar que a amostra fosse apenas das gestantes atendidas na FSP. Nesse caso para as 530 gestantes dessa sub-amostra obteve-se as árvores que estão no apêndice A4.5, continuando a vantagem da curva de Rosso na última consulta(V14). O que aumenta a possibilidade de que a presença da curva de Rosso em

primeiro lugar se deva exclusivamente ao fato da variável dependente ser o peso do RN e não a adequação do peso.

A curva de Siqueira na primeira consulta foi a mais consistente das curvas na estimação do peso do RN(V17) e na estimação da adequação do peso do RN(V19), tendo uma presença de 7 em 8 das árvores estudadas. Embora isso tenha ocorrido, pode ser que em um estudo mais detalhado a melhor técnica de regressão ou classificação por árvore indique a curva de Rosso como ocorreu com a CART usando a variável peso do RN(V17) como dependente.

Assim, existem algumas dúvidas a serem resolvidas, quais sejam:

- 1) - Se houvesse alterações na amostra, por exemplo uma amostra que fosse apenas uma parte da atual, digamos as primeiras 200 gestantes atendidas, ou quaisquer 300 delas, a curva de Siqueira continuaria se destacando ou ocorreria o inverso?
- 2) Qual é a árvore de classificação a ser considerada para o prognóstico do tamanho do bebê?
- 3) Qual a técnica de construção de árvore de regressão e classificação mais confiável?

Por esse motivo há ainda a necessidade de comparar tais técnicas quanto à sensibilidade no sentido de classificar adequadamente as mães segundo suas características para as categorias do tamanho do bebê, de modo que se possa confiar com mais segurança nos resultados.

Uma forma de fazer esse estudo exaurindo as possibilidades de alterações na amostra e análise da consistência e eficiência das curvas e das técnicas é usar a técnica de validação cruzada citada na seção 4.2.3.2, o que será feito no capítulo 6.

Utilizando o KS: Arvore com V17(peso do RN) como a variável dependente

Global Legend
Average V17
of records

FIGURA 5.3.15

V40=peso na última consulta/200, V16=duração da gestação em semanas, V3=estatura, V2=local, V12=Siqueira na última consulta, V18=sexo do RN, V4=peso pré-gestacional, V15=duração da gestação em dias, V7=Siqueira na 1a.consulta, V6=tempo de gestação na 1a. consulta.(filtro=0,01)

Utilizando o KS: Arvore com V19(adequação do peso do RN) como a variável dependente

FIGURA 5.3.16

V12=Siqueira na última consulta, V5pV3q=(peso na 1a.consulta)/(estatura)², 0=peso na última consulta/200, V16=duração da gestação em semanas, V3=estatura, V2=local. (filtro 0,01)

6 - COMPARAÇÕES ENTRE AS TÉCNICAS DE CLASSIFICAÇÃO POR ÁRVORE E LOGÍSTICA MULTINOMIAL

Como ficaram algumas dúvidas no capítulo cinco, com relação à concretização da vantagem da curva de Siqueira em possíveis variações da amostra, e sobre: qual técnica de classificação deve ser usada para o prognóstico do tamanho do bebê, a confiança e validade das técnicas de regressão e classificação por árvore e a escolha da melhor quanto à consistência em classificar corretamente, e como não existe na literatura uma comparação ou indicação de preferência por uma delas, neste capítulo será feito um estudo conjunto da validade das técnicas e eficiência das curvas em função da variação amostral, buscando a técnica que apresente a menor taxa de má-classificação, conforme os estimadores definidos no capítulo quatro, sempre considerando um modelo em que a adequação do peso do RN é a variável dependente.

6.1 - COMPARAÇÃO QUANTO À FACILIDADE DE USO:

Uma forma inicial de comparação entre as cinco técnicas, quatro de classificação por árvores e uma regressão logística multinomial, é quanto à facilidade de uso, “manuseio” e interpretação. Nos apêndices A4.3 a A4.7 se encontram os comandos e forma de utilização de cada programa, e algumas “pistas

operacionais”. Quanto ao uso e manuseio, as técnicas SPLUS, KS e CHAID são as mais simples, nessa ordem. Os programas CART e CHAID são os mais demorados, tendo usado em média neste trabalho entre 2 e 20 minutos de CPU, de um PC 486-DX de 66 Mhz com 8 Mb de memória RAM, dependendo do número de variáveis independentes, do tamanho amostral e do parâmetro usado para “poda” da árvore. Quanto à facilidade de interpretação dos resultados do programa, as técnicas de classificação por árvores podem ser classificadas na seguinte ordem:

- 1a.) KS
- 2a.) SPLUS
- 3a.) CHAID
- 4a.) CART

Pois as duas primeiras já fornecem a árvore com as indicações das partições feitas em cada nó, enquanto a CHAID exige uma informação à parte pelos comandos “Settings” e “Variable Categories” para se conhecer como foi feita a partição de cada nó. A CART que está sendo usada neste trabalho não constrói o diagrama, apenas fornecendo uma lista com cada nó e sua partição, exigindo do usuário que desenhe a sua árvore.

Com relação à regressão logística multinomial, o seu uso depende apenas dos comandos do SAS indicados no apêndice A4.3, enquanto que para interpretar o modelo basta lembrar que se trata de um conjunto de duas funções cuja diferença se deve à mudança de categoria das variáveis dependentes, e para estudo da classificação deve-se usar a função inversa da logística como mostrado no apêndice A6.1.

6.2 - VALIDAÇÃO CRUZADA:

Como dentre os estimadores da taxa de má-classificação apresentados na seção 4.2.3, o estimador por validação cruzada apresenta a vantagem de ser bem próximo da taxa real $R(\mathcal{A})$, ele será utilizado para a presente amostra, a qual será a amostra informada $\mathcal{L}$.

Uma outra vantagem da taxa de má-classificação por validação cruzada é que a maioria das variações internas da amostra são cobertas, e suas particularidades são mais facilmente identificadas.

Inicialmente a amostra $\mathcal{L}$ com os 565 dados deve ser dividida em ν sub-amostras, o que remete à necessidade da definição do número de sub-amostras ν . BREIMAN et al(1984) notaram que quanto maior for ν mais acurado fica o estimador da taxa de má-classificação $R^{CV}(T)$, mas por outro lado quanto maior for ν , maior é o tempo de computação.

Como BREIMAN et al(1984) concluíram que $\nu = 10$ fornece bons resultados, em uma análise inicial será usado $\nu = 10$ e posteriormente, procurando estudar possíveis diferenças de resultados, será usado $\nu = 5$.

6.2.1 - VALIDAÇÃO CRUZADA EM 10 PARTES:

Para $\nu = 10$, a amostra informada forneceu 10 sub-amostras de tamanho 56, e foram aplicadas as quatro técnicas de classificação por árvore e a regressão logística multinomial em cada um dos dez conjunto de dados, com V19(adequação do peso do RN) como variável dependente.

Para que fosse efetuada a divisão da amostra informada de tamanho 565 em 10 sub-amostras, de modo que essa divisão em sub-amostras não priorizasse os dados segundo suas ordens de registro, primeiramente foi efetuada a aleatorização dos mesmos criando uma variável chamada RAN, gerada por sorteio aleatório uniforme entre 1 e 565, usando o pacote SPSS, de modo que cada indivíduo tivesse um número entre 1 e 565. Posteriormente, os registros foram ordenados de acordo com a variável RAN em um banco de dados definitivo. A partir desse banco de dados foram tomados os primeiros 56 registros como a primeira sub-amostra que ficaria em um primeiro arquivo computacional, e o seu complementar, os 509 registros restantes ficariam em um segundo arquivo computacional, os segundos 56 registros seriam a segunda sub-amostra e ficariam em um terceiro arquivo computacional, e o seu complementar de 509 registros em um quarto arquivo computacional, assim por diante. Deve-se notar que o terceiro arquivo contém os 56 registros que estão no segundo arquivo mais 8x56 registros que estão em outros oito arquivos de tamanho 56.

Cada um dos 10 arquivos de tamanho 509 é uma sub-amostra do tipo $\mathcal{L}^* - \mathcal{L}_j$ da seção 4.2.3.2, e foi usada para ajustar uma árvore para cada técnica de classificação por árvore e uma logística, e cada um dos arquivos de tamanho 56 é uma sub-amostra $\mathcal{L}_j$ da seção 4.2.3.2, a qual foi usada para testar a classificação indicada pelo modelo ou árvore construída usando a amostra complementar de tamanho 509, calculando então a taxa de má-classificação segundo o estimador 4.2.3.2.6.

ESTUDO DA CLASSIFICAÇÃO EM FUNÇÃO DE PARÂMETROS:

Para que fosse efetuada a comparação entre as técnicas, inicialmente procurou-se estudar a taxa de má-classificação em função da variação de parâmetros próprios de cada técnica, esperando-se que alterações no tamanho da

árvore ou modelo resultassem em alterações nas respectivas taxas de má-classificação. Os parâmetros usados para esse fim estão na tabela 6.2.1.1. Esse estudo levaria à identificação dos parâmetros mais indicados para cada técnica, e a comparação desses resultados permitiriam encontrar aquela técnica que alcançasse o melhor rendimento.

As alterações dos parâmetros que levam a diferentes tamanhos das árvores e modelos podem ser feitas das seguintes formas:

- alterando os parâmetros que definem o tamanho mínimo possível para os subgrupos a serem criados por junção de categorias *before* e *after* do CHAID.
- alterando os parâmetros *merge level* e *eligibility level* do CHAID, que definem respectivamente o nível de dificuldade em combinar categorias de uma variável independente e o erro tipo 1 para uma variável ser considerada estatisticamente significativa. Quanto maior for o nível *merge* maior é a dificuldade em combinar categorias e quanto menor for o nível *eligibility*, mais significativa é a relação.
- podando a árvore pelo comando *prune.tree(arvore,k)* do SPLUS, onde *k* é uma função da deviance e define o tamanho da poda. Quanto menor é o valor de *k*, melhor é o ajuste aos dados e maior é a árvore, o que leva ao encolhimento da árvore à medida que *k* cresce.
- alterando o nível de significância dos testes para CART e KS.
- alterando o nível de significância do teste que inclui variáveis no modelo logístico pelo sub-comando SLE do procedimento LOGISTIC do SAS.

Tabela 6.2.1.1 : Quadro dos parâmetros usados nos programas computacionais de cada técnica de classificação por árvore

TÉCNICA	PARÂMETROS USADOS
SPLUS	prune.tree(arvore,k) com k=7.5 a 12
CHAID	before/after(b/a)=de 1/1 a 20/20 merge level/elegibility level(L)= de 0.05 a 0.15 scores=1 p/PIG 2 p/AIG 3 p/GIG
CART	β o limite de impureza(4.2.3.4.2) , variando de 0.012 a 0.026 medida de impureza de Gini
KS	nível de significância do teste p/ incluir variáveis: filtro de 0.01 a 0.15
LOGÍSTICA	nível de significância p/ incluir variáveis: SLE=0.05 a 0.20

A variação dos parâmetros das técnicas de classificação por árvore e os consequentes resultados em termos de má-classificação são apresentados nas tabelas 6.2.1.2 a 6.2.1.6.

A TÉCNICA SPLUS:

A tabela 6.2.1.2 apresenta os resultados da classificação pela técnica SPLUS nas 10 sub-amostras de tamanho 56, para dez valores do parâmetro K dessa técnica, totalizando 100 árvores. A primeira linha dessa tabela mostra o número de árvores cuja classificação era apenas AIG, sem classificar as crianças como PIG ou GIG, a segunda linha mostra o número de erros de classificação, a terceira linha mostra as taxas de má-classificação, e a última linha mostra o número médio de ramos para os diversos valores do parâmetro K dessa técnica. Quando a classificação era apenas AIG só ocorria erro quando o

verdadeiro valor de V19 era PIG ou GIG. Ocorre que a amostra tem uma quantidade grande de AIG, e se não fosse aplicada qualquer técnica, apenas “chutando” que todos os bebês seriam AIG, a taxa de má-classificação seria a dessas árvores que classificam apenas AIG.

Assim, observando a tabela 6.2.1.2, não seria aconselhável usar, para essa amostra, valores de k maiores ou iguais a 10, pois das 10 árvores da validação cruzada 7 ou mais se apresentaram classificando apenas V19=AIG. Tais resultados indicam que a SPLUS alcança o melhor desempenho quando K está próximo de 9.5 com um total de 112 erros e uma taxa de má-classificação de 0.200.

Tabela 6.2.1.2 - Quadro dos resultados da técnica SPLUS:

K	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12
classif.AIG	1	2	1	2	3	7	8	9	9	9
Nº.de erros	130	128	122	121	112	110	110	110	110	110
RCV	0.2331	0.2286	0.2178	0.2161	0.2000	0.1964	0.1964	0.1964	0.1964	0.1964
Nº. ramos	4.7	4.8	4.9	5.3	6.0	6.7	9.2	10.4	12.8	15.6

A TÉCNICA CHAID:

A tabela 6.2.1.3 apresenta os resultados da classificação pela técnica CHAID nas 10 sub-amostras de tamanho 56, para cinco valores dos parâmetros before/after e três valores dos parâmetros níveis merge/elegibility dessa técnica, totalizando 130 árvores. Para cada valor de *before/after* a primeira linha apresenta o número de árvores cuja classificação era apenas AIG, a segunda linha o número de erros de classificação, a terceira linha as taxas de má-classificação, e a última linha o número médio de ramos para os diversos valores dos parâmetros.

Tabela 6.2.1.3 - Quadro dos resultados da técnica CHAID :

before/after		merge level = eligibility level		
		0.05	0.1	0.15
before = 1 after = 1	classif.só AIG	1	0	0
	Nº.de erros	113	123	135
	RCV	0.2017	0.2196	0.2411
	Nº.de ramos	13.9	19.5	25.2
before = 2 after = 2	classif.só AIG	1	1	1
	Nº.de erros	109	116	120
	RCV	0.1946	0.2071	0.2142
	Nº.de ramos	11.4	14.9	17.9
before = 5 after = 5	classif.só AIG	3	3	1
	Nº.de erros	114	115	123
	RCV	0.2036	0.2053	0.2196
	Nº.de ramos	9.2	10	13.1
before = 10 after = 10	classif.só AIG	8	6	5
	Nº.de erros	115	113	118
	RCV	0.2053	0.2017	0.2107
	Nº.de ramos	7	8.5	10.3
before = 20 after = 20	classif.só AIG		9	
	Nº.de erros		111	
	RCV		0.1982	
	Nº.de ramos		6.8	

Nota-se na tabela 6.2.1.3 que a maioria dos casos de **before** e **after** maiores ou iguais a 10 e **level** menores ou iguais a 0.1 forneceu árvores tão curtas que a classificação era apenas V19=AIG, então para comparação tais casos devem ser desconsiderados. Os resultados dessa tabela indicam que a técnica CHAID alcança o seu melhor desempenho, nesta amostra, quando before=2, after=2 e level=0.05, com um total de 109 erros de classificação e taxa de má-classificação igual a 0.1946. Comparando as taxas de má-classificação dessa técnica com os da técnica SPLUS, nota-se que a CHAID alcançou melhores desempenhos em termos de taxa de má-classificação.

Ao que tudo indica, existe uma tendência de melhores resultados para essa técnica quando os sub-grupos por junção de categoria(before/after) são pequenos e os níveis dos testes para combinar as categorias e escolher as variáveis(merge/elegibility) são menores. Nota-se que o número médio de ramos está associado com os valores desses parâmetros, e que o melhor desempenho está onde o número de ramos é aproximadamente 11.

A TÉCNICA CART:

A tabela 6.2.1.4 apresenta os resultados da classificação pela técnica CART nas 10 sub-amostras de tamanho 56, para quatorze valores do limite máximo da impureza β , totalizando 140 árvores, tendo na primeira linha o número de árvores que classificam apenas AIG, na segunda linha o número de erros de má-classificação, na terceira linha as respectivas taxas de má-classificação, e na última linha o número médio de ramos.

Tabela 6.2.1.4 - Quadro dos resultados da técnica CART:

	VALORES DE β													
	0.012	0.013	0.014	0.015	0.016	0.017	0.018	0.019	0.020	0.021	0.022	0.023	0.024	0.025
classif.AIG	0	0	2	2	3	4	7	8	8	8	8	8	9	9
Nº.de erros	129	130	120	120	122	121	112	112	112	112	112	112	111	111
RCV	0.230	0.232	0.214	0.214	0.218	0.216	0.200	0.200	0.200	0.200	0.200	0.200	0.198	0.198
Nº. ramos	>32	~32	27	27	21	18	7	6.2	6	5.4	5	3.8	3	3

Essa técnica apresentou um problema na oscilação do tamanho das árvores em função da variação do parâmetro β , não ocorrendo uma variação que indicasse uma tendência suave de sub-amostra para sub-amostra da validação cruzada, mas pelo contrário, a oscilação do tamanho da árvore com a variação do parâmetro β foi sempre brusca como pode ser notado na tabela 6.2.1.5.

Tabela 6.2.1.5: Quadro da oscilação do tamanho das árvores em função de β :

Sub-amostra	β	0.015	0.0152	0.01525	0.0153	0.0155	0.016
01	nós	53	53	1	1	1	1
Sub-amostra	β	0.017	0.0171	0.0172	0.0174	0.0175	0.018
02	nós	69	69	5	5	5	5
Sub-amostra	β	0.017	0.02	0.03	0.04	0.05	0.15
03	nós	61	61	57	27	27	17
Sub-amostra	β	0.017	0.02	0.022	0.023	0.024	0.025
04	nós	33	33	33	1	1	1
Sub-amostra	β	0.016	0.0161	0.0162	0.0163	0.0154	0.0165
05	nós	57	57	57	1	1	1
Sub-amostra	β	0.0125	0.0128	0.0129	0.013	0.0135	0.014
06	nós	81	81	25	25	21	7
Sub-amostra	β	0.017	0.0171	0.0176	0.0177	0.0178	0.0180
07	nós	61	61	61	7	7	7
Sub-amostra	β	0.0120	0.0130	0.0140	0.0150	0.0160	0.021
08	nós	83	83	11	11	7	7
Sub-amostra	β	0.016	0.017	0.0171	0.0172	0.0180	0.0190
09	nós	63	63	63	5	5	5
Sub-amostra	β	0.0170	0.0171	0.0172	0.0173	0.0175	0.0180
10	nós	59	59	59	1	1	1

Assim, o número de nós da árvore passa de um valor muito grande, como por exemplo 53 nós na primeira sub-amostra com $\beta=0.0152$

para 1 nó com $\beta=0.01525$, em uma pequena variação de 0.00005 para β . Por outro lado os valores de β que alteram o tamanho dessa árvore, quando aplicados na segunda sub-amostra não alteram o tamanho da árvore, ocorrendo uma alteração brusca na mudança de valor de β de 0.0171 para 0.0172, cujas árvores tiveram 69 nós e 5 nós respectivamente. O mesmo problema da inconstância da alteração do tamanho da árvore em função do valor de β pode ser visto nas demais sub-amostras.

Portanto embora a tabela 6.2.1.4 indique $\beta=0.017$ com a menor ocorrência de erros, a tabela 6.2.1.5 alerta para a possibilidade de existir outro ideal para β em outra amostra motivo pelo qual essa técnica não pode ser indicada como tendo um bom desempenho para um ou alguns valores de β . Mesmo no caso em que isso fosse possível, a técnica CART não teve melhor desempenho do que a CHAID e a SPLUS quando tinha um número baixo de árvores com classificações apenas de AIG.

A TÉCNICA KS:

A tabela 6.2.1.6 apresenta os resultados da classificação pela técnica KS nas 10 sub-amostras de tamanho 56, para seis valores do filtro dessa técnica, totalizando 60 árvores, tendo na primeira linha o número de árvores que classificaram apenas AIG, na segunda linha o número de erros de classificação, na terceira linha as respectivas taxas de má-classificação, e na última linha o número de ramos.

Tabela 6.2.1.6 - Quadro dos resultados da técnica KS:

	F I L T R O					
	0.01	0.02	0.05	0.08	0.1	0.15
classif.AIG	8	8	6	5	3	1
Nº.de erros	117	117	122	133	134	135
RCV	0.2089	0.2089	0.2178	0.2375	0.2393	0.2411
Nº. ramos	5.1	6.3	10.3	13.8	15.6	18.4

Nota-se pela tabela 6.2.1.6 que a técnica KS não alcança bom desempenho como as demais, mesmo se fossem considerados os casos com 5 ou 6 árvores classificando apenas AIG.

Assim considerando os seus parâmetros de poda e encolhimento de árvores, a técnica CHAID se mostrou mais vantajosa. As técnicas CHAID e KS, pela suas propriedades de partição múltipla, resultam em árvores mais “frondosas” com profundidade ou degrau menor. A técnica CART é a mais arriscada ou incerta, e a técnica KS a que teve o pior desempenho.

A LOGÍSTICA MULTINOMIAL:

A tabela 6.2.1.7 mostra os resultados do ajuste do modelo logístico multinomial nas 10 sub-amostras de tamanho 56, para sete valores do nível de significância do teste baseado na estatística $-2\text{LOG(Razão de Verossimilhança)}$ que inclui variáveis no modelo logístico pelo sub-comando SLE do procedimento LOGISTIC do SAS, totalizando 70 ajustes, apresentando na primeira linha o número de modelos que classificam apenas AIG, na segunda linha o número de erros de má-classificação, na terceira linha as respectivas taxas de má-classificação, e na última linha o número médio de variáveis envolvidas.

Tabela 6.2.1.7 - Quadro dos resultados dos ajustes da Logística Multinomial:

	Nível de Significância do teste de inclusão de variáveis						
	0.05	0.08	0.10	0.12	0.15	0.17	0.20
classif.AIG	1	2	2	2	2	2	2
Nº.de erros	111	107	107	107	114	114	114
RCV	0.1982	0.1911	0.1911	0.1911	0.2036	0.2036	0.2036
Nº. variáveis	3.7	4.2	4.2	4.2	4.4	4.4	4.4

Como pode ser visto na tabela 6.2.1.7 a logística alcançou o seu melhor desempenho quando o nível de significância do teste de inclusão de variáveis esteve entre 0.08 e 0.12, obtendo a taxa de má-classificação de valor 0.1911. Um problema encontrado no desempenho da logística multinomial foi que o modelo variou muito pouco dentro da maioria das sub-amostras, em função da mudança do nível de significância do teste de inclusão de variáveis, como pode ser verificado na tabela 6.2.1.8, o que indica que a logística tem uma sensibilidade ao parâmetro de corte do modelo bem diferente das técnicas de árvores.

Como pode ser visto nessa tabela, em 5 sub-amostras da validação cruzada de 10 partes, assinaladas com **, não houve variação alguma em quaisquer valores do nível de significância, e em 4 das sub-amostras, assinaladas com *, houve uma única mudança, apenas no momento em que o nível de significância passava de 0.05 para 0.08. A primeira sub-amostra foi a única em que a logística se apresentou mais sensível ao nível de significância, a qual é a responsável pela tendência crescente na soma total de erros de classificação da tabela 6.2.1.7 e, pelos seus valores, pode-se imaginar que os resultados possam ser diferentes se for alterada essa ou outra sub-amostra. As sub-amostras assinaladas com * são as responsáveis pela tendência decrescente na soma total de erros de classificação.

No entanto, tal fato não desabona essa técnica se for considerado que a logística tenha esse comportamento para esse tipo específico de amostra, e que para outra amostra deve-se ter o cuidado de analisar a taxa por validação cruzada, pois outros resultados podem ser alcançados.

Tabela 6.2.1.8: Quadro da oscilação do tamanho do modelo em função do nível de significância do teste de inclusão de variáveis:

		Nível de Significância do teste de inclusão						
		0.05	0.08	0.010	0.012	0.015	0.017	0.20
Sub-amostra 01	Nº.var.	5	5	5	5	7	7	7
	erros	13	13	13	13	20	20	20
Sub-amostra 02	Nº.var.	3	4	4	4	4	4	4
	erros	9	7	7	7	7	7	7
Sub-amostra 03	Nº.var.	4	4	4	4	4	4	4
	erros	12	12	12	12	12	12	12
Sub-amostra 04	Nº.var.	3	4	4	4	4	4	4
	erros	13	11	11	11	11	11	11
Sub-amostra 05	Nº.var.	4	4	4	4	4	4	4
	erros	11	11	11	11	11	11	11
Sub-amostra 06	Nº.var.	3	3	3	3	3	3	3
	erros	12	12	12	12	12	12	12
Sub-amostra 07	Nº.var.	5	5	5	5	5	5	5
	erros	8	8	8	8	8	8	8
Sub-amostra 08	Nº.var.	4	5	5	5	5	5	5
	erros	8	7	7	7	7	7	7
Sub-amostra 09	Nº.var.	4	4	4	4	4	4	4
	erros	12	12	12	12	12	12	12
Sub-amostra 10	Nº.var.	2	4	4	4	4	4	4
	erros	13	14	14	14	14	14	14

Comparando as taxas de má-classificação da logística multinomial com as obtidas pelas técnicas de classificação por árvore, nota-se praticamente nenhuma vantagem da logística sobre as demais, classificando acertadamente dois indivíduos a mais do que a CHAID, o que é insignificante para um total de 560 classificações, ou proporcionalmente 0.00357 classificações corretas de diferença.

ESTUDO DA CLASSIFICAÇÃO EM FUNÇÃO DO TAMANHO DA ÁRVORE:

Desde que o tamanho das árvores está associado com o número de variáveis usadas no prognóstico, que por seu lado estão associadas com a classificação, uma outra forma de análise seria comparar as técnicas segundo as taxas de má-classificação para árvores com número fixo de ramos, mas como as técnicas CART e S-PLUS fazem partições binárias e as técnicas CHAID e KS fazem partições múltiplas, notou-se em uma análise dos dados das tabelas 6.2.1.2 a 6.2.1.6 que as duas primeiras têm associação entre o número de ramos e o número de erros por má-classificação sendo possível para a SPLUS encontrar a região ou número de ramos em que a árvore consegue alcançar o valor mínimo para a taxa de má-classificação, enquanto a técnica CART tem a dificuldade da oscilação brusca já citada. Nesse estudo a SPLUS teve seu melhor desempenho quando o número de ramos era cinco. Por outro lado as duas últimas técnicas constroem árvores diferentes com o mesmo número de ramos, levando a erros por má-classificação diversificados, impedindo assim que se fizessem comparações segundo o número de ramos.

Uma outra forma de fixar um parâmetro de comparação, seria o uso do tamanho da árvore em profundidade(“depth”) ou degraus. Para isso

foi fixado que as árvores tivessem profundidade 2, 4 ou 6. O uso do tamanho máximo da árvore profundidade 6 se deve ao fato que, à medida que a árvore tem mais ramos e aumenta a profundidade, envolve mais variáveis, maior detalhe na classificação, e maior dificuldade de interpretação, como ocorre em modelos lineares. Além do que a medida que os nós se distanciam do nó raiz a associação entre a variável dependente e a variável independente daquele nó fica mais fraca. Por esse motivo muitas vezes houve necessidade de podar ou de aumentar a copa da árvore como, por exemplo, a árvore resultante do pacote SPLUS que chega a ter 14.335 nós e 61 nós terminais para a amostra de 565 deste trabalho, como pode ser visto no apêndice A.6.2. Em contrapartida, o fato de existir uma grande maioria de categoria AIG na variável resposta V19 remete à necessidade de detectar as categorias menos frequentes de V19; e isso é possível à medida que se aumenta a árvore. Esse estudo foi feito também para a validação cruzada em 5 partes, considerando árvores com profundidade até 4. O problema encontrado quando se fixava o número de ramos não foi encontrado fixando-se a profundidade.

Foram aplicadas as quatro técnicas de classificação por árvore e a regressão logística multinomial, com os melhores parâmetros encontrados no estudo anterior, em cada um dos dez conjuntos de dados, com V19 como variável dependente, resultando em 10 árvores com profundidade 2, 10 árvores com profundidade 4, e 10 árvores com profundidade 6 para cada técnica de classificação por árvore, totalizando 120 árvores, e 10 modelos logísticos multinomiais. Os ajustes de árvores e logitos multinomiais da validação cruzada de partição dez estão na tabela 6.2.1.9.

A coluna DEGRAU representa a ordem em que a variável foi incluída na árvore/modelo, ou seja é profundidade da árvore, contada a partir da raiz. Cada coluna $\mathcal{L}_j$ representa o ajuste de uma árvore/modelo com suas variáveis independentes segundo a ordem de entrada para a sub-amostra $\mathcal{L}_j$.

Posteriormente na linha marcada na primeira coluna com nº.nós term. se refere ao número de nós terminais de cada árvore. Como a regressão logística não tem profundidade, para efeito de comparação, considerou-se o número de variáveis no modelo na direção do degrau.

Posteriormente, na tabela 6.2.1.10, estão os resultados da validação cruzada de dez partes, registrando os números de erros de má-classificação conforme o apêndice A6.3, para ressubstituição $\sum z_R$ e validação cruzada $\sum z_C$, primeiramente para árvores com profundidade 2, depois para árvores com profundidade 4 e finalmente para árvores com profundidade 6. Abaixo dos números de erros estão as respectivas taxas de má-classificação conforme as definições 4.2.3.2.3 e 4.2.3.2.6.

A forma de aplicação da logística na validação cruzada e fórmulas utilizadas se encontram no apêndice A6.4.

TABELA 6.2.1.9 - Resultados dos ajustes de árvores e logísticas multinomiais por validação cruzada em 10 sub-grupos, com as variáveis que entraram no modelo segundo a ordem de entrada

DEGRAU	PROGRAMAS USADOS NO AJUSTE																													
	S - P L U S										C H A I D										C A R T									
	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
1	v10	v7	v10	v7	v10	v10	v10	v7	v7	v7	v10	v7	v14	v7	v7		v7	v7	v7	v7	v7	v7	v10	v7	v7	v7	v7	v7	v7	v7
2	v7	v10		v10	v7		v7	v10	v10	v10		v10	v3	v15	v10	v10	v10	v10	v15	v16	v5	v5	v3	v5	v10	v5	v16	v5	v5	
		v5	v7	v20		v7	v6	v20		v20	v16	v16	v3	v16	v16	v16	v16	v16	v16	v16	v5		v11		v5		v5			
											v20	v20	v16	v20	v20	v7	v20	v20	v20	v16			v3							
3	v15	v15	v4	v11	v15	v4	v15	v11	v16	v4	v20	v16	v3	v18	v10	v4p3	pesg2	v2	pesg2	v18	v5	v9	v10p	v15	pesg2	v9	v11	v5	pesg2	v10p
	v6		v20	v15	v20	v3	v20	v15		v5	v10		pesg	v16	v16	v3		v16		v10		v10p	pesg2	v3	v15	v3	v10		v3	
			v5							pesg2	v7	v16	v4p3	v10	v20	v13			v16	v18	v3				v4		v5p3			
				v10p							v10		v14			v20				v10p					v3		v3			
4	v4	v16	v6	v4p3	v4p3	v15	v4	v4p3	v4p3	v5p3	v8	v18	v7	v8	v18		v2		v2		v5	v6		v6	v5	v3	v9	v11	v16	v11
	v5p3		v16	v9	v5	v6	v10			v11	v6		v18		v5				pesg2				v20	v9	v14	v10p	v11	v4		
	v15		v5			v6				v6			v20		v16				v18						pesg2		v6			
										v15					v5p3											v6				
															v18															
6	v4	v3	v4p3	v9	v5p3	v4p3	v15			v6			v3								v6		v5	v5	v10	v5	v5	v14		v16
	v9	v5	v5	v6	pesg2	v18	v20			v5											v4p3		v5	v16	v15	v10	v6	v2		v5
	v6	v4	v3	v5	v10	v6	v10			v10											v4		v3	v11	v4p3	v14	v3		v6	
		v10	v15		v3	v3	v4			v4p3													v3		v9	v16	v10p		v18	
			v18																								v3	v16		
nº nós term	8	6	9	10	6	9	8	8	5	10	14	10	25	14	15	12	10	11	15	18	8	5	7	7	7	12	7	11	6	6

(continuação)

TABELA 6.2.1.9 - Resultados dos ajustes de árvores e logísticas multinomiais por validação cruzada em 10 sub-grupos, com as variáveis que entraram no modelo segundo a ordem de entrada

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L0
1		V7		V7	V7	V12	V7	V7	V7	V7	V7	V10	V10	V10	V10	V10	V10	V10	V7	V10
2	V3	V10	V3	V5p3	V14	V3	V10	V10	V14	V5p3	V16	V16	V16	V16	V16	V16	V16	V16	V16	V16
	V3	V12	V15		V6	V15		V16												
	V3		V3	V20		V7														
3	V16	V11	V9	V11	V5p3	V4	pesg1	V16	V9	V4p3	V3	V7	V7	V7	V7	V7	V7	V7	V3	V7
	V15	V16	V4	V3	V15	V20	V4	V9	V6	V6										
	V7		V7	V13	pesg1	V16	V15													
4	V11	V4p3	V3	V4p3	V15	V3	V16	V10	V2	V11	V14	V3	V5p3	V3	V5p3	V5p3	V3	V3	V14	V5p3
	V7	V9	V3	V15	V10p	V8	V3		pesg2	V15										
		V3	V8		V11	V11	V9													
					pesg2															
6	V4p3	pesg2	V16	V11	pesg2	V14	V2		V4p3	V9						V14	V14	V14		
	V7	V18	V7	pesg2	V15	V16	V18		V5p3	V10					V3					
		V15	V4p3	V13	V9				V16											
			V5p3	V5p3	V6				V20											
			pesg1																	
nº.nós term	14	14	19	15	17	14	17	10	15	11	5	4	4	4	4	6	5	5	4	4

Sendo:

- V3= estatura da gestante
- V4= peso pré-gestacional
- V5= peso na primeira consulta
- V6= tempo de gestação na primeira consulta
- V7= curva de Siqueira na primeira consulta
- V8= curva de Rosso na primeira consulta
- V9= escala de Rosso na primeira consulta
- V10= peso na última consulta
- V11= tempo de gestação na última consulta
- V12= curva de Siqueira na última consulta
- V13= curva de Rosso na última consulta
- V14= escala de Rosso na última consulta
- V15= Duração da gestação em dias
- V16= Duração da gestação em semanas
- V18= Sexo do RN
- V20= tipo de parto

Tabela 6.2.1.10 - Resultados dos ajustes de árvores e logísticas multinomiais por validação cruzada em 10 sub-grupos com os números de erros de má-classificação para os ajustes da tabela 6.2.1.9 e respectivas taxas de má-classificação

	S - P L U S										C H A I D										C A R T									
	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_6$	$\mathcal{L}_7$	$\mathcal{L}_8$	$\mathcal{L}_9$	$\mathcal{L}_{10}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_6$	$\mathcal{L}_7$	$\mathcal{L}_8$	$\mathcal{L}_9$	$\mathcal{L}_{10}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_6$	$\mathcal{L}_7$	$\mathcal{L}_8$	$\mathcal{L}_9$	$\mathcal{L}_{10}$
Arvp $\sum z_R^2$	97	98	99	102	99	98	103	103	99	96	87	99	97	97	90	93	98	98	95	88	89	98	94	93	94	94	98	121	96	92
taxa Ressub.	$R^R(T) = 0.1953$										$R^R(T) = 0.1851$										$R^R(T) = 0.1904$									
$\sum z_C^2$	14	8	12	9	12	13	8	8	12	15	14	8	12	9	12	15	8	8	11	7	13	8	11	9	12	13	8	8	12	14
taxa V.Cruz.	$R^C(T) = 0.1982$										$R^C(T) = 0.1857$										$R^C(T) = 0.1929$									
Arv.m $\sum z_R^4$	81	94	94	92	89	87	89	90	96	83	84	94	93	88	82	91	97	96	91	84	127	96	93	88	92	81	93	85	95	91
taxa Ressub.	$R^R(T) = 0.1758$										$R^R(T) = 0.1768$										$R^R(T) = 0.1849$									
$\sum z_C^4$	15	8	11	9	12	14	10	12	12	15	14	8	8	13	12	15	8	8	13	7	14	9	11	13	12	13	9	12	12	14
taxa V.Cruz.	$R^C(T) = 0.2107$										$R^C(T) = 0.1893$										$R^C(T) = 0.2125$									
Arv.g $\sum z_R^6$	81	91	84	84	84	87	81	90	92	74	84	94	91	88	82	91	97	96	91	84	74	84	85	81	77	78	85	75	90	77
taxa Ressub.	$R^R(T) = 0.1666$										$R^R(T) = 0.1764$										$R^R(T) = 0.1583$									
$\sum z_C^6$	15	8	12	13	11	14	12	12	12	16	17	8	8	13	12	15	8	8	13	7	14	12	12	10	12	12	11	11	12	18
taxa V.Cruz.	$R^C(T) = 0.2232$										$R^C(T) = 0.1946$										$R^C(T) = 0.2214$									

OBS: Arv.p = árvore com depth=2, Arv.m= árvore com depth=4 e Arv.g = árvore com depth=6
 $\sum z_R^j$ = soma dos erros de má-classificação por ressubstituição para árvore com depth=j

(continuação)

Tabela 6.2.1.10 - Resultados dos ajustes de árvores e logísticas multinomiais por validação cruzada em 10 sub-grupos com os números de erros de má-classificação para os ajustes da tabela 6.2.1.9 e respectivas taxas de má-classificação

	K S										LOGÍSTICA									
	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_6$	$\mathcal{L}_7$	$\mathcal{L}_8$	$\mathcal{L}_9$	$\mathcal{L}_{10}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_6$	$\mathcal{L}_7$	$\mathcal{L}_8$	$\mathcal{L}_9$	$\mathcal{L}_{10}$
nº.nós term	11	11	10	11	11	12	11	10	8	11	-	-	-	-	-	-	-	-	-	-
Arvp $\sum z_R^2$	97	104	99	101	98	96	102	102	98	95	96	98	95	105	97	98	99	99	99	95
taxa Ressub.	$R^R(T) = 0.1949$										$R^R(T) = 0.1927$									
$\sum z_C^2$	14	7	12	9	12	17	8	8	12	15	15	8	11	13	11	12	7	8	12	13
taxa V.Cruz.	$R^C(T) = 0.2037$										$R^C(T) = 0.1964$									
Arv.m $\sum z_R^4$	91	103	98	98	96	94	100	103	98	91	95	97	93	93	94	94	98	98	96	95
taxa Ressub	$R^R(T) = 0.1909$										$R^R(T) = 0.1872$									
$\sum z_C^4$	14	7	12	17	12	17	8	11	12	15	13	7	12	11	11	11	7	8	12	13
taxa V.Cruz.	$R^C(T) = 0.2232$										$R^C(T) = 0.1875$									
Arv.g $\sum z_R^6$	97	102	93	97	94	94	100	100	91	91	99	97	93	93	94	97	102	103	96	95
taxa Ressub	$R^R(T) = 0.1884$										$R^R(T) = 0.1904$									
$\sum z_C^6$	15	7	13	17	12	17	8	12	12	15	12	7	12	11	11	12	8	7	12	14
taxa V.Cruz.	$R^C(T) = 0.2286$										$R^C(T) = 0.1893$									

OBS: Arvp = árvore com depth=2, Arv.m= árvore com depth=4 e Arv.g = árvore com depth=6

$\sum z_R^j$ = soma dos erros de má-classificação por ressubstituição para árvore com depth=j

A linha número de ramos, da tabela 6.2.10, indica que a LOGÍSTICA teve apenas um modelo com seis variáveis, três modelos com cinco variáveis e seis modelos com quatro variáveis enquanto solicitava-se ao programa computacional que ajustasse um modelo com no máximo seis variáveis. Por esse motivo pode-se pensar em comparar os modelos ajustados da LOGÍSTICA com as árvores que tenham profundidade 4.

Analisando os números e as taxas de má-classificações da tabela 6.2.10, vê-se que quando a classificação é testada na própria amostra que construiu a árvore, considerando as taxas de má-classificação por Ressubstituição como medida de avaliação do ajuste, essas taxas indicam as técnicas com maior capacidade de classificar corretamente na seguinte ordem:

- **para degrau=2**: 1º. CHAID, 2º. CART, 3º. LOGÍSTICA, 4º. KS, e em último lugar a SPLUS.
- **para degrau= 4**: 1º. SPLUS, 2º. CHAID, 3º. CART, 4º. LOGÍSTICA, e em último lugar a KS.
- **para degrau= 6**: 1º. CART, 2º. SPLUS, 3º. CHAID, 4º. KS, e em último lugar a LOGÍSTICA.

A KS e a LOGÍSTICA não apresentaram em nenhum caso uma baixa taxa de má-classificação por ressubstituição.

Com relação à capacidade de classificar na amostra complementar da usada para construir a árvore/modelo, considerando as taxas de má-classificação por validação cruzada como medida de avaliação do ajuste, essas taxas indicam as técnicas com maior capacidade de classificar corretamente na seguinte ordem:

- **para degrau= 2:** 1º. CHAID, 2º. CART, 3º. LOGÍSTICA, 4º. SPLUS, e em último lugar a KS.
- **para degrau= 4:** 1º. LOGÍSTICA, 2º. CHAID, 3º. SPLUS, 4º. CART, e em último lugar a KS.
- **para degrau=6:** 1º. LOGÍSTICA(se for considerado que seus modelos têm profundidade 6), 2º. CHAID(ou junto com a logística), 3º. CART, 4º. SPLUS, e em último lugar a KS.

Para a profundidade 4, a diferença entre as técnicas LOGÍSTICA E CHAID se deve à classificação de apenas uma gestação com adequação do peso do RN erradamente, a mais para a técnica CHAID, o que é insignificante para um total de 560 classificações. Para a profundidade 6 a técnica CHAID continua sendo a melhor dentre as técnicas de classificação por árvore, mas a LOGÍSTICA com no máximo 6 variáveis apresentou o melhor desempenho, com a taxa de má-classificação por validação cruzada de 0.1875. Como pode ser notado, praticamente o desempenho da LOGÍSTICA não se alterou quando o modelo passou de 4 para no máximo 6 variáveis, isso se deve principalmente ao fato de apenas um modelo dos dez ter realmente seis variáveis, o que deve estar ligado ao problema da não oscilação do tamanho do modelo visto na tabela 6.2.8. Pode-se notar que o desempenho da LOGÍSTICA no degrau 6 é o mesmo da CHAID no degrau 4, com a taxa de 0,1893.

Em um estudo à parte procurando os casos em que o número de variáveis era igual a 6 para a técnica CHAID, encontrou-se a taxa de má-classificação igual a 0.2089, ou seja fixar o número de variáveis não é interessante para essa técnica, embora mesmo assim seja um valor melhor do que das demais técnicas de classificação por árvore.

ESTUDO DA CLASSIFICAÇÃO EM FUNÇÃO DAS CATEGORIAS DA VARIÁVEL DEPENDENTE:

Uma outra forma de analisar essas técnicas é estudar o desempenho das mesmas em termos de porcentagem de erro de má-classificação para cada uma das categorias da variável dependente V19(adequação do peso do RN). Para isso, foi construída para cada técnica uma tabela cruzando a classificação feita, correta ou errada, com o valor real da variável V19 em cada árvore, a partir da qual calculou-se, para cada categoria, a porcentagem de classificações erradas com relação a todas árvores ajustadas, resultando finalmente na tabela 6.2.11. Devido à diferença de número de parâmetros usados em cada técnica, o número de árvores estudadas foi bem maior na técnica CHAID e menor na LOGÍSTICA.

Tabela 6.2.1.11 - Porcentagem de erros de má-classificação por técnica e categoria da variável V19(adequação do peso do RN)

CATEGORIA	TÉCNICA	Nº.ÁRVORES	% MÍNIMA	% MÁXIMA	% MÉDIA
V19 = PIG (31 casos)	SPLUS	61	75	100	98.36
	CHAID	137	50	100	97.52
	CART	70	100	100	100
	KS	51	51	75	99.51
	LOGÍSTICA	30	100	100	100
V19 = AIG (454 casos)	SPLUS	61	0	18.75	2.96
	CHAID	137	0	19.51	3.11
	CART	70	0	12.50	3.71
	KS	51	0	16.67	4.20
	LOGÍSTICA	30	0	15.69	1.61
V19 = GIG (80 casos)	SPLUS	61	50	100	94.35
	CHAID	137	28	100	93.93
	CART	70	57	100	92.22
	KS	51	71	100	99.44
	LOGÍSTICA	30	57	100	89.38

Os valores entre parênteses na primeira coluna são os números de casos realmente existentes na amostra como tendo essa adequação de peso, e as colunas referentes a porcentagens se referem à média das porcentagens de erros por má-classificação para todos os casos contados nas árvores/modelos ajustadas .

Como pode ser notado na tabela 6.2.11 a porcentagem de erro é maior quando a adequação do peso do RN é PIG, mas isso pode ser também devido ao baixo número de casos realmente existentes na amostra. O que pode ser confirmado ao se notar que para V19 igual a GIG a porcentagem de erros caiu ligeiramente em todas técnicas. Assim os casos PIG e GIG podem ser considerados mais raros e conseqüentemente mais difíceis de serem classificados corretamente. A LOGÍSTICA e a CART não classificaram corretamente nenhum caso em que o RN era PIG. A CHAID foi a técnica que foi mais sensível à categoria PIG, tendo menos casos de má-classificação nessa categoria. No outro extremo, na classificação dos recém-nascidos que eram GIG, a LOGÍSTICA foi a técnica que apresentou menor número de erros. Para a classificação dos recém-nascidos com adequação do peso igual a AIG, a LOGÍSTICA foi a técnica que melhor classificou. Esses resultados mostram que a LOGÍSTICA obteve a menor taxa de má-classificação, indicada na tabela 6.2.2.10, devido ao seu desempenho nas categorias AIG e GIG. Se a adequação do peso do RN fosse distribuída mais uniformemente entre as categorias os resultados em termos de classificação e erro de classificação poderiam ser outros.

O fato das técnicas não classificarem corretamente a categoria PIG, não significa que as curvas tenham falhado no prognóstico da adequação do peso do RN, como foi visto nas tabelas 5.3.5 a 5.3.10.

Um fato interessante que pode explicar a falta de classificação como PIG, foi que nas classificações por árvore, ao fazer a partição dos nós iniciais, a categoria OBS de V7 e de V12 tem estado quase sempre separada de DES+NOR, e as árvores tem ramificações mais longas a partir de DES+NOR e mais curtas em OBS. Isso quer dizer que inicialmente as gestantes OBS são mais diferenciadas de DES e NOR, com o sucesso de classificar mais GIG do que PIG, pois em muitos casos DES fica confundida com NOR, talvez devido principalmente ao pouco número de DES e à diferença de sensibilidade das curvas com relação a DES e OBS.

6.2.2 - VALIDAÇÃO CRUZADA EM 5 PARTES:

Com o intuito de verificar se existe possível interferência do tamanho das sub-amostras da validação cruzada no desempenho das técnicas e na escolha das variáveis a serem incluídas nas árvores/modelos, estudou-se o ajuste das quatro técnicas de classificação por árvores e o modelo logístico multinomial usando a validação cruzada em 5 partes. Para isso, usando a notação da seção 4.2.3 tomou-se $v = 5$, ou seja 5 sub-amostras de tamanho 113, tomadas aleatoriamente da mesma forma como foram obtidas as amostras da validação cruzada em 10 partes. Agora as sub-amostras $\mathcal{L}_j$ para testar a árvore/modelo por classificação são as sub-amostras de tamanho 113, e as sub-amostras complementares $\mathcal{L} - \mathcal{L}_j$ usadas para construir as árvores/modelos são de tamanho 452.

Como o número de casos usados na validação cruzada em cinco partes para construir a árvore/modelo é menor do que o usado na validação

cruzada em 10 partes, como a LOGÍSTICA e a CHAID tendo sido as melhores técnicas na validação cruzada em 10 partes muitas vezes com profundidade significativa até quatro, e como a validação cruzada em 5 partes visa apenas verificar se existiam incongruências com relação à validação cruzada de 10 partes, foi considerado o estudo em que se fixasse o tamanho das árvores com profundidades 2 e 4. Como no caso da validação cruzada em 10 partes, para o modelo logístico multinomial, considerou-se o número de variáveis como profundidade. Os resultados encontram-se na tabela 6.2.2.1.

TABELA 6.2.2.1 Resultados dos ajustes de árvores e logísticas multinomiais por validação cruzada em 5 sub-grupos, com as variáveis que entraram no modelo segundo a ordem de entrada, número de erros e respectivas taxas de má-classificação.

	PACOTES USADOS NO AJUSTE																									
DEGRAU	S - PLUS					CHAID					CART					K S					LOGÍSTICA					
	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	
1	V10	V10	V10	V7	V7	V7	V7	V7	V7	V7	V7	V4	V7	V7	V7					V7	V7	V7	V10	V10	V10	V10
2	V7	V11	V7	V20	V10	V10	V20	V5	V10		V5	V18	V5	V10	V5		V15		V15							
		V7		v10				v10		V10							V7	V3	V3	V20	V14					
																	V3			V10						
3	V15	V20	V4	V15	V6				V11	V3	V9	V11	V4	V5			V3	V9	V9	V16	V6	V3	V7	V7	V7	V7
	V10	V4	V3		V4				V18	V18		V5	V3	V18			V20	V4	V4		V9					
																		V7	V6							
																		v1								
4	V4	V6	v6	V3	V6		V11			V11		V5	V9				V11		V18		V2		V5p3	V5p3	V3	V5p3
		V5											V3													
nº.nós term	5	7	5	7	7	6	6	10	10	14	4	6	5	5	3		10	14	11	8	7	-	-	-	-	-
Arv.p $\sum z_R^1$	90	90	85	91	84	90	90	85	88	80	85	90	85	88	82		90	90	84	94	83	88	84	83	90	85
$\sum z_C^1$	$R^R(T)_1 = 0.1947$					$R^R(T)_1 = 0.1915$					$R^R(T)_1 = 0.1903$					$R^R(T)_1 = 0.1951$					$R^R(T)_1 = 0.1903$					
	21	21	26	28	27	21	21	26	20	29	21	21	26	20	26		21	21	29	28	27	22	22	21	16	26
	$R^{CV}(T)_1 = 0.2177$					$R^{CV}(T)_1 = 0.2071$					$R^{CV}(T)_1 = 0.2018$					$R^{CV}(T)_1 = 0.2230$					$R^{CV}(T)_1 = 0.1893$					
Arv.g $\sum z_R^2$	81	85	78	90	82	85	86	92	82	79	81	86	78	78	79		89	90	74	98	84	87	83	84	89	79
$\sum z_C^2$	$R^R(T)_2 = 0.1841$					$R^R(T)_2 = 0.1876$					$R^R(T)_2 = 0.1779$					$R^R(T)_2 = 0.1925$					$R^R(T)_2 = 0.1867$					
	22	20	25	25	30	20	22	25	24	26	22	24	25	21	27		21	21	28	28	27	21	22	22	15	23
	$R^{CV}(T)_2 = 0.2159$					$R^{CV}(T)_2 = 0.2071$					$R^{CV}(T)_2 = 0.2106$					$R^{CV}(T)_2 = 0.2212$					$R^{CV}(T)_2 = 0.1823$					

OBS: soma dos erros de má-classificação: por ressubstituição = $\sum z_R^j$, por validação cruzada = $\sum z_C^j$; para árvore com depth=j

Analisando os números e as taxas de má-classificações da tabela 6.2.2.1, vê-se que quando a classificação é testada na própria amostra que construiu a árvore, as taxas de má-classificação por Ressubstituição indicam as técnicas com maior capacidade de classificar corretamente na seguinte ordem:

- **para degrau=2**: 1º. CART e LOGÍSTICA, 2º. CHAID, 3º. SPLUS, e em último lugar a KS.
- **para degrau= 4**: 1º. CART, 2º. SPLUS, 3º. LOGÍSTICA, 4º. CHAID, e em último lugar a KS.

A KS não apresentou em nenhum caso uma baixa taxa de má-classificação por ressubstituição, enquanto a LOGÍSTICA, contrariamente ao caso da validação cruzada em 10 partes, alcançou uma boa taxa juntamente com a CART para a profundidade 2. Para a profundidade 4 a vantagem da LOGÍSTICA sobre a CHAID é de 2 classificações corretas, o que é insignificante para o total de 2260 classificações, ou 0,0009 classificações corretas.

Com relação à capacidade de classificar na amostra complementar da usada para construir a árvore/modelo, as taxas de má-classificação por validação cruzada indicam as técnicas com maior capacidade de classificar corretamente na seguinte ordem:

- **para degrau= 2**: 1º. LOGÍSTICA, 2º. CART, 3º. CHAID, 4º. SPLUS, e em último lugar a KS.
- **para degrau= 4**: 1º. LOGÍSTICA, 2º. CHAID, 3º. CART, 4º. SPLUS, e em último lugar a KS.

Assim, como pode-se notar, novamente a LOGÍSTICA se apresentou com melhores valores do que as demais técnicas, enquanto entre as técnicas de classificação por árvore a CHAID e a CART estiveram com desempenhos praticamente iguais. Em ambas as validações cruzadas a LOGÍSTICA se mostrou com melhor desempenho, seguida bem proximamente da técnica de

classificação por árvore CHAID. Da mesma forma, a KS manteve o pior desempenho em termos de taxa de má-classificação por validação cruzada e também por ress substituição.

Nota-se que a taxa de má-classificação por ress substituição é sempre menor do que a taxa de má-classificação por validação cruzada, devido ao fato que no caso da ress substituição a amostra usada para teste é a mesma que foi usada para construir a árvore. BREIMAN e col.(1984) notaram que isso ocorria sempre para árvores com mais de dois nós terminais.

Comparando as tabelas das taxas de má-classificação das duas validações cruzadas, nota-se que à medida que foi aumentando o número de partes da validação cruzada, de 5 para 10, a taxa de má-classificação por validação cruzada ficou menor para S-PLUS e CHAID, ora maior, ora menor para CART e KS e maior para a LOGÍSTICA. BREIMAN e col.(1984) indicam que as taxas de validação cruzada em 10 partes fornecem estimativas mais confiáveis e próximas da verdadeira taxa de má-classificação, motivo pelo qual os resultados da tabela 6.2.1.10 devem ser considerados com maior confiança.

ESTUDO DA COMPOSIÇÃO DAS ÁRVORES/MODELOS:

Para contribuir com a análise das tabelas 6.2.1.9 e 6.2.2.1 e comparar as duas validações cruzadas quanto a estrutura das árvores e utilização das variáveis nas árvores e no modelo logístico, embora haja elementos repetidos na constituição das sub-amostras usadas para construir as árvores, foi construída a tabela 6.2.2.2 com as frequências de vezes que as variáveis apareceram nas árvores/modelos, procurando obter-se uma idéia da importância das variáveis para a amostra inteira.

As variáveis foram ordenadas segundo as frequências de vezes que eram usadas nas árvores/modelos da validação cruzada em 10 partes, para cada programa computacional, e colocadas nessa ordem na tabela 6.2.2.2. As colunas indicadas por cv10 e cv5 referem-se à indicação da validação cruzada.

Tabela 6.2.2.2: Frequências que as variáveis aparecem nos ajustes das árvores / modelo para as validações cruzadas de 10 e 5 partes segundo a técnica usada.

Variável	cor nas tabelas 6.2.1.9 e 6.2.2.1	SPLUS		CHAID		CART		K S		LOGIST		frequência total	
		cv10	cv5	cv10	cv5	cv10	cv5	cv10	cv5	cv10	cv5	cv10	cv5
V7	amarela	10	5	10	5	9	4	11	5	10	5	50	24
V16	azul claro	3	0	18	1	7	0	6	2	6	5	44	8
V10	cinza	15	6	11	4	5	1	4	1	8	4	43	16
V3	contorno azul	5	2	4	1	13	1	12	2	6	2	40	8
V5	sem cor	9	0	1	1	19	6	0	0	0	0	29	7
V15	sem cor	11	2	2	0	3	0	8	2	0	0	24	4
V6	sem cor	10	4	1	0	7	0	3	2	0	0	21	6
V20	sem cor	7	2	10	1	1	0	2	2	0	0	20	5
V12	verde	2	2	7	2	0	0	8	4	0	1	17	9
V11	sem cor	3	1	0	3	6	1	6	1	0	0	15	6
V4pV3	sem cor	7	0	2	0	2	0	3	0	0	0	14	0
V4	sem cor	8	4	0	0	3	2	2	2	0	0	13	8
V9	contorno vermelho	3	0	0	0	5	2	4	3	0	0	12	5
V5pV3q	sem cor	3	0	1	0	1	0	3	0	4	3	12	3
V18	sem cor	2	0	8	0	1	2	0	3	0	0	11	5
pesogr2	sem cor	2	0	3	0	4	0	1	0	0	0	10	0
V14	contorno lilás	0	0	2	0	3	0	3	1	2	0	10	1
V10pV3q	sem cor	1	0	1	0	6	0	1	0	0	0	9	0
V2	sem cor	0	0	3	0	1	0	1	1	0	0	5	1
V8	vermelho	0	0	2	0	0	0	2	0	0	0	4	0
pesogr1	sem cor	0	0	1	0	0	0	2	0	0	0	3	0
V13	lilás	0	0	0	0	0	0	1	0	0	0	1	0
TOTAL		101	28	87	18	96	19	83	31	40	20		

Assim sendo, a tabela 6.2.2.2 mostra que a variável mais freqüente nas árvores/modelo por ambas validações cruzadas foi a curva de Siqueira na primeira consulta; aparecendo em 50 vezes das 50 árvores/modelos da validação cruzada em 10 partes, em 24 das 25 árvores/modelos da validação cruzada em 5 partes, sendo sempre usada pelos programas SPLUS, CHAID e LOGÍSTICA em ambas validações. Tal resultado indica que mesmo nas variações internas das subamostras próprias da validação cruzada a curva de Siqueira foi identificada como a mais significativa para prever a adequação do peso do recém-nascido.

Além disso a vantagem da curva de Siqueira sobre a de Rosso é notada pela ordem das demais variáveis em termos de freqüência de uso nas árvores/modelos, pois a curva de Siqueira na segunda consulta foi mais freqüente do que a de Rosso em ambas as consultas, reforçando os resultados encontrados no capítulo 5 a esse respeito. Essa superioridade é mais evidente ao se notar na tabela 6.2.1.9 e 6.2.2.1 que a variável V7(curva de Siqueira na primeira consulta) aparece mais freqüentemente nos primeiros degraus.

Como pode ser notado na tabela 6.2.2.2, as freqüências indicam que as variáveis mais importantes são as mesmas para ambas as validações cruzadas, embora a variável duração da gestação em semanas(V16) tenha caído para a validação cruzada em 5 partes e a variável diagnóstico da curva de Siqueira na última consulta(V12) tenha aumentado em termos de freqüências de vezes de uso pelas árvores/modelos.

As variáveis peso da gestante, V5(na primeira consulta) e V10(na última consulta), e a estatura(V3), são as medidas antropométricas mais comuns. As variáveis peso da gestante, V4(pré-gestacional) e V10(na última consulta) e estatura(V3) são as medidas antropométricas mais constantemente usadas

na validação cruzada de 5 partes, sendo que a variável peso da gestante na última consulta(V10) se destacou das demais em ambas as validações cruzadas.

A segunda variável em frequência para a validação cruzada em 10 partes é o tempo de gestação em semanas(V16) e para a validação cruzada em 5 partes é o peso da gestante na última consulta(V10), sendo que V10 também se destaca na validação cruzada em 10 partes. A técnica SPLUS usa mais as variáveis V10(peso na última consulta), a CHAID usa mais as variáveis V16(tempo de gestação em semanas) na primeira validação e V7(Siqueira na primeira consulta) na segunda validação, a CART usa mais a variável V5(peso na primeira consulta) em ambas validações, a KS usa mais as variáveis V3(estatura) na primeira e V7(Siqueira na primeira consulta) na segunda validações, e a LOGÍSTICA usa mais a variável V7 em ambas as validações. As técnicas SPLUS e CHAID são bem parecidas em termos de composição da informação necessária para prognosticar V19(adequação do peso do RN), com as variáveis em posições bem semelhantes.

Algumas características interessantes das técnicas quanto ao uso das curvas em teste na construção de suas árvores/modelos são que a CART e a KS foram as técnicas que mais usaram a curva de Rosso, embora relativamente pouco comparando com a de Siqueira, e as variáveis mais usadas, V9 e V14 se referem à escala numérica de Rosso e não ao estado nutricional. Com relação às variáveis calculadas em função de outras existentes na amostra, a CART foi a técnica que mais usou o peso na última consulta relativo à estatura ao quadrado, $V10pV3q$, e a LOGÍSTICA foi a técnica que mais usou o peso na primeira consulta relativa à estatura ao quadrado, $V5pV3q$. O peso ganho desde o peso pré-gestacional relativamente à estatura, PESOGR2, foi usado nas técnicas CART, CHAID e SPLUS, e o peso ganho desde o peso na primeira consulta relativamente à

estatura PESOGR1 foi usado nas técnicas KS e CHAID, apenas na validação cruzada de 10 partes.

A grande frequência de V7(Siqueira na primeira consulta) e V12(Siqueira na última consulta), a baixa frequência de V9(escala de Rosso na primeira consulta) e V14(escala de Rosso na última consulta) e o não aparecimento de V8(Rosso na primeira consulta) e V13(Rosso na última consulta) nas árvores/modelos ajustados, indicam uma grande diferença entre as curvas de Siqueira e de Rosso, com uma vantagem indiscutível da curva de Siqueira, com destaque para sua aplicação na primeira consulta. Além disso, essa curva quando aplicada na última consulta, também apresenta melhor desempenho com relação à de Rosso, quando aplicada em qualquer consulta.

Com a variação da amostra característica da validação cruzada, a presença da curva de Siqueira com maior frequência nos primeiros nós das árvores de ambas as validações cruzadas é um indício de maior eficiência da mesma. A constância dos resultados usando as amostras de tamanho 509 na validação cruzada de 10 partes, as amostras de tamanho 452 da validação cruzada em 5 partes e da amostra total no capítulo 5, indicam a curva de Siqueira como a mais consistente.

7 - CONCLUSÕES:

Para concluir, deve-se lembrar que na determinação da adequação do peso do RN podem existir viéses principalmente devido a efeitos de variáveis possivelmente influentes não estudadas, e que tais efeitos estejam diluídos e distribuídos uniformemente e não concentrados em pontos.

Assim, pelos resultados dos estudos efetuados nos capítulos 5 e 6 pode-se obter as seguintes conclusões:

7.1 - Conclusões relativas às técnicas de classificação:

1. Comparando as técnicas de classificação por árvore, a técnica CHAID mostrou-se mais eficiente, no sentido de ter menor variação em termos de classificação e uso das mesmas variáveis em todas as sub-amostras de ambas as validações cruzadas, e número de má-classificações.
2. Na maioria das sub-amostras a técnica CHAID foi a que obteve o melhor ajuste e menor erro, conseguindo a menor taxa de má-classificação dentre as técnicas de classificação por árvore.
3. Ao alterar o tamanho da validação cruzada, a técnica CHAID apresentou-se como sendo a mais consistente, com a menor taxa de má-classificação para a variação de menor para maior sub-amostra.

4. Comparando as técnicas de classificação por árvores com a logística multinomial, a última se mostrou ligeiramente mais poderosa na classificação, embora com a ressalva de seus resultados dependerem quase exclusivamente de apenas uma das sub-amostras. No entanto tais resultados foram consistentes e com pouca oscilação de uma para outra validação. A LOGÍSTICA e a CHAID foram as técnicas mais sensíveis respectivamente aos eventos raros GIG e PIG.
5. Embora a logística tenha se destacado ligeiramente, a CHAID tem a preferência, por ser mais constante de sub-amostra para sub-amostra, e para os dados aqui estudados, uma vez que foi a técnica mais sensível em classificar corretamente a categoria PIG da variável V19, que pode ser considerada a mais interessante, em termos de saúde pública.
6. Uma outra vantagem da CHAID sobre a LOGÍSTICA é na interpretação dos resultados em termos de visualização das variáveis, e forma de suas utilizações segundo as partições usadas.
7. A técnica KS é a menos indicada, e a SPLUS tem vantagem sobre a CART em termos de facilidade de interpretação dos resultados, por ter menor porcentagem de erro de classificação nas categorias PIG e AIG da adequação do peso do RN, enquanto a CART tem ligeira vantagem sobre a SPLUS em termos de taxa de má-classificação.
8. As vantagens das técnicas LOGÍSTICA e CHAID podem não ser conclusivas, e sim momentaneamente devido à amostra em estudo, pois se houver outra constituição de estrutura das categorias da variável dependente ou mesmo outras variáveis independentes, os resultados poderão ser modificados, motivo pelo qual as conclusões relativas à comparação entre as técnicas devem ser consideradas, no momento, válidas para o conjunto aqui analisado.

7.2 - Conclusões relativas às curvas de avaliação do estado nutricional da gestante:

1. A primeira diferença entre as curvas é com relação às amostras usadas nos trabalhos que originaram tais curvas. Essas amostras diferiram no tamanho, nas médias das medidas registradas nas gestantes, na idade gestacional mínima. Embora a curva de Rosso tenha sido construída sob a limitação de mães primigestas e com tempo de gestação entre 38 e 42 semanas, foi aplicada a todas as gestantes, como tem sido feito nos serviços de saúde.
2. A curva de Rosso apresenta pesos das gestantes mais altos do que a de Siqueira nos seus estados nutricionais DES e NOR, e inversamente mais baixos no seu estado nutricional OBS.
3. A curva de Siqueira mostrou mais precisão do que a de Rosso, ao apresentar a distribuição dos pesos das gestantes mais concentradas em todos diagnósticos e consultas.
4. Considerando a amostra global, Cotia mais FSP, ou apenas a amostra FSP, obteve-se o mesmo resultado quanto à comparação entre as curvas, com vantagem para a de Siqueira na primeira consulta.
5. O maior problema da curva de Rosso é o de usar um ponto de corte alto, entre DES e NOR, o que resultou na apresentação da distribuição assimétrica para as categorias do estado nutricional, com maior frequência em DES, sub-classificando a nutrição da gestante. Assim, a curva de Rosso tende a considerar as gestantes normais como sendo desnutridas e as obesas como sendo normais, embora isso possa ser considerado uma vantagem, pois se não dá para ser preciso é melhor errar chamando de DES uma gestante NOR, o que leva à maior atenção do serviço de saúde, o que não é maléfico. Assim a curva de Rosso tem maior capacidade de detectar mais PIG. Essa tendência pode criar um grupo de risco

“aumentado” com reflexos no custo da intervenção. Por outro lado, no estado nutricional DES, a curva de Siqueira se apresentou com a distribuição do RN mais definida; enquanto para os estados nutricionais NOR e OBS a curva de Rosso foi a que apresentou essa distribuição mais definida.

6. A curva de Rosso teve as distribuições dos pesos dos RN se interceptando quando os diagnósticos do estado nutricional das gestantes eram NOR e DES, o que leva a um prognóstico confuso com relação ao peso do RN, a partir do estado nutricional da sua curva, principalmente se for DES.
7. Com relação ao prognóstico da adequação do peso do RN, a curva de Siqueira mostrou-se mais adequada, principalmente se o RN tinha adequação do peso igual a AIG e GIG em ambas as consultas, tendo melhor desempenho na primeira consulta. Quando o seu diagnóstico do estado nutricional era DES, a curva de Siqueira tinha menor proporção de acertos do que a de Rosso, com pior desempenho na última consulta. Apesar disso a curva de Siqueira foi a mais associada com a adequação do peso do RN.
8. Devido à assimetria da distribuição de frequências de seus diagnósticos dos estados nutricionais, a curva de Rosso tem o pior prognóstico para a adequação do peso do RN, quando o diagnóstico da sua curva é NOR ou OBS, com a tendência em sub-classificar quando o bebê tinha adequação de peso AIG ou GIG. Assim, a curva de Rosso não só tende a sub-classificar o estado nutricional da gestante, como também a sub-classificar a gestante com relação à adequação do peso do RN. Assim, a curva de Rosso tende a considerar as mães de bebês AIG como DES e de bebês GIG como NOR, que leva à mesma vantagem comentada na conclusão 7.2.5. Desse modo a curva de Siqueira mostrou-se mais adequada, principalmente se o RN tinha adequação do peso igual a AIG e GIG.
9. Com relação a sensibilidade, especificidade e valor preditivo das curvas, fazendo uma aproximação por partições dicotômicas, notou-se uma desvantagem da curva

de Rosso. Com relação a previsão de PIG, a curva de Rosso na última consulta mostrou-se ligeiramente mais sensível, 0,645 versus 0,566 de Rosso na primeira consulta; enquanto a curva de Siqueira na última consulta mostrou-se com maior especificidade, 0,938 versus 0,876 para Siqueira na primeira consulta; e a curva de Siqueira na última consulta mostrou-se com maior valor preditivo, 0,273 versus 0,185 para a curva de Siqueira na primeira consulta.

- 10.A regressão linear por “stepwise” com o peso do RN como variável dependente, incluiu a curva de Siqueira da primeira consulta, após a inclusão das variáveis Peso da gestante na última consulta, Duração da gestação em dias e Peso da gestante na primeira consulta. A curva de Rosso não foi sequer incluída na regressão por “stepwise”.
- 11.A regressão logística multinomial por “stepwise” com a adequação do peso do RN como variável dependente incluiu apenas a curva de Siqueira na primeira consulta depois das variáveis Peso na última consulta e Tempo de gestação em semanas, sem incluir também a curva de Rosso.
- 12.A aplicação das técnicas de classificação por árvore, reforçada pela validação cruzada, indicou uma vantagem indiscutível da curva de Siqueira na primeira consulta sobre a de Rosso em qualquer consulta. Essa vantagem é evidenciada tanto pelo fato de ser a variável mais usada pelas árvores/modelos como pelo fato de ser a primeira variável a ser escolhida para as árvores/modelos.
- 13.A frequência de vezes com que a curva de Siqueira foi preferida nos primeiros lugares pelas árvores/modelos, nas diversas sub-amostras das validações cruzadas, indica essa curva como a mais eficiente, e a constância desses resultados de amostras menores até amostras maiores a caracteriza como a mais consistente, além de ter a maior associação com a adequação do peso do RN.

O interessante do fato da curva de Siqueira na primeira consulta aparecer em primeiro lugar na classificação das árvore/modelos, é que isso leva a um diagnóstico no início da gravidez, o que deve facilitar a indicação de intervenção logo no início da gravidez, com melhores chances de bons resultados.

Confirmando outros trabalhos, as variáveis Peso na última consulta e Tempo de gestação foram as variáveis medidas diretamente com maiores associações.

Esses resultados indicam a necessidade de se fazer futuros trabalhos:

1. Em outras amostras, em outros lugares, buscando verificar ou confirmar a vantagem da curva de Siqueira.
2. Incluir outras variáveis em futuros estudos como peso do último filho, número de gestações, nível sócio-econômico, etc...
3. Estudar as técnicas de classificação em nova amostra cuja variável dependente tenha categorias menos discrepantes.
4. Verificar se o modelo logístico multinomial mantém sua vantagem sobre as demais técnicas em outros tamanhos de validações cruzadas, e maior variação do parâmetro SLE de inclusão de variável no modelo.
5. Estudar a inclusão de interação no modelo logístico multinomial.
6. Dar maior importância para a partição $PIG|AIG,GIG$ do que para a partição $PIG,AIG|GIG$, em todas as técnicas de classificação por árvore e logística multinomial.
7. Incluir custo de classificação nas técnicas de classificação por árvore e logística multinomial.
8. Estudar a possibilidade de aperfeiçoamento das curvas de avaliação do estado nutricional.

REFERÊNCIAS BIBLIOGRÁFICAS

1. AGRESTI, A. *Categorical Data Analysis*. John Wiley & Sons. 1990.
2. AUBRY, R.H.; ROBERTS, A.; CUENCA, V.G. *The assessment of maternal nutrition. In Barnes, L.A., and Pitkin, R.M., editors: Clinics in perinatology, symposium on nutrition, Philadelphia, W.B.Saunders Co, 1975.*
3. BARRY, A.J.; GOODNIGHT, J.H.; SALL, J.P.; HELWING, J.; SAS. *A User's Guide to SAS: SAS - Institute Inc. San Francisco: W.F.Freeman. 1986.*
4. BEILLY, J.S. & KURLAND, II. *The relationship of maternal weight gain and weight of newborn infants. Am.J.Obstet Gynecol; 50:202-6. 1945*
5. BERGNER, L. & SUSSER, M.W. *Low birth weight and prenatal nutrition: an interpretative review. Pediatrics 46: 946-66, 1970.*
6. BESSER, O.A; LOWRY, O.H; BROCH, M.J. *The quantitative determination of ascorbic acid in small amounts of white blood cells and platelets. J Biol Chem . 168: 197-205. 1947.*
7. BIRCH, H.G. Summary of session I. In: SEMINAR OF MALNUTRITION IN EARLY LIFE AND SUBSEQUENT MENTAL DEVELOPMENT, Mona, Jamaica, 1972. Nutrition, the nervous system, and behavior: proceedings. Washington, D.C., Pan American Health Organization, p.38-40. (PAHO - Scient.publ., 251), 1972.
8. BMDP: Statistical Software - University of California Press - Berkeley. 198
9. BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. *Classification and Regression Trees*. Wadworth & Brooks/Cole Advanced Books - Pacific Grove, California. 1984.

10. CHASE, H.C. *International comparisons of perinatal and infant mortality* . Vital Hlth Statist. Sur., 3(6) 1967.
11. CHATFIELD, C.; COLLINS, A.J. *Introduction to Multivariate Analysis*. Science Paperbach, 1980.
12. CHOW, B.F. et al. *Maternal nutrition and metabolism of the offspring: studies in rats and men*. Amer.J.Publ.Hlth., 58: 668-77, 1968.
13. CIARI, Jr.,C. *Curva Ponderal de Gestantes normais*. Folha Médica, 68: 141-6, 1974.
14. CLARK, L.A.; PREGIBON, D. *Statistical Models in S*. John M.Chambers and Trevor J.Hastie - AT&T Bell Laboratories. Wadsworth & Books/Cole Advanced Books & Software. Pacific Grove, California. 1992.
15. COMMITTEE ON MATERNAL AND CHILD CARE. *Reducing infant mortality*. J. Amer.Med.Ass., 193: 130-9, 1965.
16. DRAPER, N.R.; SMITH, H. *Applied regression Analysis*. John Wiley & Sons,Inc. 1981.
17. EASTMAN, N.J; JACKSON, E. *Weight relationships in pregnancy. The bearing of maternal weight gain and pre-pregnancy weight on birth weight in full term pregnancies*. Obstet Gynecol; 23: 1003-25.1968.
18. EVELETH, P.B & TANNER, J.M. *Worldwide variation in human growth*. Cambridge University Press, 1976.
19. FIKREE, F.F.; BERENDES, H.W. *Risk factors for term intrauterine growth retardation: a community-based study in Karachi*. Bullet.of the Word Health Organizat., vol 72, n.4, Geneva. 1994.

20. FISHER, R.A. *The statistical utilization of multiple measurements.* Annals of Eugenics, 8: 376-386. 1938.
21. FLEISS, J.L. *Statistical Methods for Rates and Proportions* - John Wiley & Sons. 1981
22. FORATTINI, O.P. *Epidemiologia Geral.* Editora da Universidade de São Paulo. São Paulo. 1976.
23. GEBRE MEDHIN, M.; GOBEZIE, A. *Dietary intake in the third trimester of pregnancy and birth weight of offspring among nonprivileged and privileged women.* Amer.Clin.Nutr., 28:1322-9, 1975.
24. GUERI, M.; JUTSUM, P.; SORHAINDO, B. *Anthropometric assessment of nutritional status in pregnant women: a reference table of weight-for-height by week of pregnancy.* Am.J.Clin.Nutr.;35:609-16. 1982.
25. GUTTMAN, I. *Linear Models: An Introduction,* John Wiley & Sons, N.Y. 1982.
26. HO, L.L. *Alguns aspectos da técnica AID para variável resposta categorizada.* Tese de Mestrado - IME - USP. 1987
27. HOAGLIN, D.C.; MOSTELLER, F.; TUKEY, J.W. *Understanding Robust and Exploratory Data Analysis.* John Wiley & Sons. 1983.
28. HSUEH, A.M. et al. *The importance of the period of dietary restriction of the dam on behavior and growth in the rat.* J.Nutr., 104:37-46, 1974.
29. HYTTEN, F.E; LEITCH, I. *The physiology of human pregnancy.* London, Oxford University Press, 1964.
30. IGLESIAS, P. *Una medida de impureza para la construccion de un arbol binario de predicción multivariada.* Revista brasileira de Probabilidade e Estatística. Outubro de 1987

31. INFORME DE UN GRUPO DE ESTUDIOS EM LA OMS: Dieta, Nutrición y Prevención de Enfermedades Crónicas. OMS - Ginebra . 1990 .
32. JACOBSON, N.H. *Weight and weight gain in pregnancy*. Clin Perinat; 2:233-42. 1975.
33. JELLIFE, D.B. *The assessment of nutritional status of the community*. World Health Organization, Ginebra. 240-1. 1982.
34. JELLIFFE, D.B. *The assessment of the nutritional status of the community*. World Health Organization, Ginebra, 210-241. 1966.
35. JELLIFFE, D.B. et al. *Evaluación del estado de nutrición de la comunidad*. O.M.S. Ginebra. 1968.
36. JOHNSON, R.A; WICHERN, D.W. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Inc, Englewood Cliffs, New Jersey. 1982.
37. JURADO GARCIA, E. *El crecimiento intrauterino: correlación peso/ longitud corporal al nacimiento en función de la edad de la gestación*. Gac. Med. Mex., 102: 227-55, 1971.
38. KAFATOS, A.G.; VLACHONIKOLIS,I.G.; CONDRINGTON,C.A. *Nutrition during pregnancy: the effects of educational intervention program in Greece*. Am.J.Clin.Nutr.; 50:970-9, 1989.
39. KASS, G.V. *An exploratory technique for investigating large quantities for categorical data*. Applied Statistical, 29, 2: 119-127. 1980.
40. KnowledgeSEEKER User's Guide version 2.0. FirstMark Technologies Ltd. Ottawa - Canada. 1990.

41. KRAMER, M.S. *Determinants of low birth weight: methodological assessment and meta-analysis.* Bulletin of the World Health Organ., 65(5):663-737. 1987.
42. KRAMER, M.S. *Effects of energy and protein intakes on pregnancy outcome: an overview of the research evidence from controlled clinical trials.* J.Clin.Nutr. 58:627-35, 1993.
43. KRASOVEC, K; ANDERSON, M.A.; *Maternal nutrition and pregnancy outcomes - Anthropometric Assessment.* Pan American Health Organization. Washington, DC. 1991.
44. LIGHT, R.J.; MARGOLIN, B.H. *An analysis of Variance for Categorical Data.* J. Am.Stat.Assoc. 66:534-44. 1971.
45. McCULLAGH, P. & NELDER, J.A. *Generalized Linear Models.* 2ª ed., Londres, Chapman and Hall, 1989.
46. MONTEIRO, C.A. *Velhos e Novos males da Saúde do Brasil - A evolução do país e de suas doenças .* Editora Hucitec-Nupens/USP. 1995.
47. MONTGOMERY, D.C.; PECK, E.A. *Introduction to Linear Regression Analysis.* John Wiley & Sons, Inc.
48. MOOD, A.M.; GRAYBILL, F.A.; BOES, D.C. *Introduction to theory statistics.* McGRAW-HILL. Kogakusha - Tokio. 1974.
49. MORGAN, J.N.; SONQUIST, J.A. *Problems in the analysis of survey data.* Journal of the American Statistical Association, Washington, 68: 415-435, 1963.
50. MURRAY, C.J.L. & LOPEZ, A.D. *Global and regional cause-of-death patterns in 1990 - Bullet.of the World Health Organiz.: vol 72, 3. Geneva. 1994.*

51. NAEYE, R.L. *Weight gain and the outcome of pregnancy.* Amer. J. Obstet. Gynec., 135:3-9, 1979.
52. NELDER, J. & WASSERMAN, W. *Applied Linear Statistical Models.* Richard D.Irwin, Inc, Illinois. 1974.
53. NISWANDER, K.R.; SINGER, J; WESTPHAL, M; WEISS, M. *Weight gain during pregnancy and pre-pregnancy weight.* Obstet Gynecol; 482-91.1969.
54. NOBREGA, F.J. *Desnutrição Intra-Uterina e Pós-Natal.* São Paulo. Panamed Editorial. 1981.
55. PAYNE, C.D. (ed) *The GLIM System - release.* Oxford. Numerical Algorithms Groups Ltda. 3-77. 1986.
56. PERFIL MUNICIPAL DE SAÚDE Fundação Sistema estadual de Análise de Dados - SEADE - 1995.
57. PHYLLIS, B.; EVELETH, S.; TANNER, J.M. *Wordwide variation in human growth.* Cambridge University Press - 1976.
58. PUFFER, R.R & SERRANO, C.V. *Patterns of mortality in childhood.* Washington, D.C., Pan American Health Organization (PAHO - Scient.publ.,262). 1973.
59. QUÊTELET, A. *Anthropométric ou mesure des differentes facultês de l'homme.* Bruxeles, C.Muquardt, 1870.
60. RAMALINGASWAMI, V. *Nutrition, cell biology and human development.* WHO Chron., 29: 306-12, 1975.
61. ROJAS, M.C.R. *Reconocimiento de patrones a traves de arbores binarios de classificacion. Sistemas expertos.* Memoria para optar al titulo de Ingeniero Civil en Informatica. Universidad Tecnica Federico Santa Maria. Chile. 1987.

62. ROSSO, P. *A new chart to monitor weight gain during pregnancy.* Am.J.Clin. Nut. 41: 644-52, 1985.
63. SEARLE, R.S. *Linear Models.* John Wiley & Sons, 1980.
64. SIMONSON, M.; CHOW, B.F. *Maze studies on progeny of underfed mother rats.* J.Nutr., 100: 685-90, 1970.
65. SIMONSON, M. et al. *Responsability of the obstetrician to the fetus. II Influence of prepregnancy weight and pregnancy weight on birthweight.* Obstet.and Gynec., 45: 481-7, 1975.
66. SIMPSON, J.W. et al *Responsability of the obstetrician to the fetus. II. Influence of prepregnancy weight and pregnancy weight gain on birth weight.* Obstet. and Gynec., 45: 481-7, 1975.
67. SIQUEIRA, A.A.F.DE *Estudo de um conjunto de curvas antropométricas no diagnóstico de estado nutricional de gestantes e sua relação com o tamanho do recém-nascido.* São Paulo, Tese de Doutorado - Faculdade de Saúde Pública - USP. 1979.
68. SIQUEIRA, A.A.F.DE *Indicadores de risco de peso inadequado ao nascer.* (Tese de Livre-Docência - Faculdade de Saúde Pública - USP). 1981.
69. SIQUEIRA, A.A.F.DE et al. *Influência da altura e ganho de peso maternos e da idade gestacional sobre o peso do recém-nascido: estudo de 3 grupos de gestantes normais.* Rev.Saúde Públ., S.Paulo, 9:331-42, 1975.
70. SMITH, J.M.; VAN NESS, H.C. *Introdução à Termodinâmica da Engenharia Química.* Ed. Guanabara Dois. 3a. edição, pg.139.
71. SOARES, J.F.; BARTMAN, F.C. *Métodos Estatísticos em Medicina e Biologia.* 14º.Colóquio Brasileiro de Matemática. Instituto de Matemática Pura e Aplicada do CNPq. Poços de Caldas. 1983.

72. SOKAL, R.; SNEATH, P. *Principle of numeric taxonomie*. San Francisco: W.H.Freeman. 1963
73. SPSS/PC+Advanced Statistical Package For Social Scientists, versão 6.0. SPSS Inc., 1993.
74. TADDEI e col. *O percentual do peso para altura da gestante no prognóstico do baixo peso ao nascer: Estudo multicêntrico*. Rev.Bras.Ginecol.Obstet. 1991
75. TANAKA, A.C.d'A. *Saúde Materna e Saúde Perinatal - Relações entre variáveis orgânicas sócio-econômicas e institucionais*. Tese de Doutorado - FSP - USP. 1986.
76. WHO EXPERT COMMITEE. *Physical status: The use and interpretation of anthropometry*. World Health Organization. Genebra. 1995.
77. WILKS, S.S. *Mathematical Statistics*. John Wiley & Sons, Inc., 1971
78. WILLIAMS, R.L. et al *Fetal growth and perinatal viability in California*. Obstet y Gynecol, 59, pg. 624-632. 1982.
79. WORTHINGTON, B.S.; VERMEERSCH, J.; WILLIAMS, S.R. *Nutrição na gravidez e na lactação*. R.Janeiro.Ed.Interamericana; 1980.

APÊNDICES:

A1 - O RISCO RELATIVO DE ÓBITOS NEO-NATAIS PARA BAIXO PESO:

Segundo informações do SEADE em 1994 foram registrados:

Na cidade de São Paulo: 209.372 nascidos vivos
3.089 óbitos de crianças com idades inferiores a 28 dias
taxa de mortalidade infantil neo-natal = 14.75

No estado de São Paulo: 684.828 nascidos vivos
10.943 óbitos de crianças com idades inferiores a 28 dias
taxa de mortalidade infantil neo-natal = 15.98

Segundo diversos autores(SIQUEIRA(1979), KRAMER(1987)) a porcentagem de nascimentos com baixo peso é de aproximadamente 10%, e segundo PUFFER e SERRANO(1973), dentre os óbitos neo-natais 60% são de crianças de baixo peso, então:

PARA A CIDADE DE SÃO PAULO

peso	óbitos neo-natais	não óbitos neo-natais	totais
inferior a 2500 g	1.853	19.084	20.937
superior a 2500 g	1.236	187.199	188.435
totais	3.089	206.283	209.372

do que tem-se as taxas:

Taxa de óbitos neo-natais em crianças de peso abaixo de 2500 gr :

$$(1853/20937).1000=88,504 \text{ por mil}$$

Taxa de óbitos neo-natais em crianças de peso acima de 2500 gr:

$$(1236/188435).1000=6,56 \text{ por mil}$$

de onde se obtém o Risco Relativo de óbitos neo-natais para peso inferior a 2500gr:

$$RR=88.504/6.56=13,49$$

PARA O ESTADO DE SÃO PAULO

peso	óbitos neo-natais	não óbitos neo-natais	totais
inferior a 2500 g	6.566	61.917	68.483
superior a 2500 g	4.377	611.968	616.345
totais	10.943	673.885	684.828

do que tem-se as taxas:

Taxa de óbitos neo-natais em crianças de peso abaixo de 2500 gr :

$$(6566/68483).1000=95.878 \text{ por mil}$$

Taxa de óbitos neo-natais em crianças de peso acima de 2500 gr:

$$(4377/616345).1000=7.101 \text{ por mil}$$

de onde se obtém o Risco Relativo de óbitos neo-natais para peso inferior a 2500gr:

$$RR=95.88/7.10=13.50$$

A4.1 - Análise buscando verificar a possibilidade de usar os dados da FSP e COT juntamente:

Com o objetivo de comparar as amostras de FSP e de COT, foram feitos alguns testes, como será visto a seguir. No entanto, a diferença dos tamanhos amostrais, e principalmente o tamanho da amostra de Cotia podem servir como impecílio para inferências em termos de testes de hipóteses ou intervalo de confiança para a diferença de médias, porque estimativas calculadas na amostra de FSP são mais estáveis e mais confiáveis, enquanto a amostra de 35 dados de COT fornece estimativas mais instáveis, pois a inclusão de alguns novos registros poderia ter um grande peso, alterando estatísticas como a média, mediana e variância.

Oferta e acesso dos serviços de assistência à saúde da Grande São Paulo e da cidade de São Paulo:

Para que se tenha uma idéia comparativa da oferta e acesso aos serviços públicos de assistência à saúde na grande São Paulo onde se encontra a cidade de Cotia e na cidade de São Paulo onde está a Faculdade de Saúde Pública, será mostrada a seguir a distribuição das gestantes inscritas no programa de assistência à gestante dos Centros de Saúde da Secretaria de Estado da Saúde, e consultas médicas de rotina nesse programa, segundo Anuário Estatísticos do Estado de São Paulo:

	Ano	Grande São Paulo	São Paulo
Inscrições	1983	59.490 (0.00887)	35.678 (0.00772) *
	1984	38.639 (0.00576)	27.181 (0.00588) *
	1985	51.596 (0.00077)	33.694 (0.00729) *
Consultas de Rotina	1983	189.007 (0.0282)	112.533 (0.02436) *
	1984	123.080 (0.0183)	86.206 (0.01866) *
	1985	226.206 (0.0337)	149.854 (0.03243) *
População de Eleitores inscritos	1983	6.701.004	4.620.138

* = proporcionalmente com relação à população de eleitores inscritos, a grande São Paulo tem um pouco mais de Inscrições e consultas, mas não existem diferenças significantes.

Para verificar a possibilidade e adequabilidade da junção das duas amostras referentes a FSP e COT, serão feitos alguns testes de hipóteses.

Como os objetivos 1 e 2 do presente trabalho são comparações entre técnicas e, quanto maior for o conjunto de dados melhor é empregada a metodologia indicada, optou-se por aproveitar todos os dados colhidos pela equipe de pesquisadores do projeto: Nutrição Materna do Departamento de

Comparando com relação à estatura da gestante:

com estaturas médias:

Localidade	média	desvio padrão	valor p	intervalo de confiança para a diferença
Cotia	152,57	6,43	0,008*	(-5,38 ; -0,84)*
FSP	155,68	6,23		
Junção em uma população	155,49	6,28		

* = o teste de Levene indicou desigualdade das variâncias.

O teste de igualdade de médias indica diferença entre os grupos quanto à média das estaturas, diferença essa confirmada pelo desenho box-plot, embora em termos de distribuição não exista muita diferença entre as mesmas.

Comparando com relação ao peso do RN:

com pesos médios dos RN:

Localidade	média	desvio padrão	valor p	intervalo de confiança para a diferença
Cotia	3374,57	502,27	0,035*	(-133,13 ; 129,13)*
FSP	3205,28	455,22		
Junção em uma população	3215,77	459,61		

* = o teste de Levene indicou igualdade das variâncias.

Assim os dois grupos tiveram pesos médios dos recém-nascidos iguais, embora o desenho box-plot indique uma ligeira vantagem do grupo de Cotia.

A42.3 - TABELA DE MULTIPLA ENTRADA E O MODELO LOG-LINEAR:

variável resposta Y	variáveis independentes							
	X ₁							
	X ₁ = x ₁				X ₁ = x _i		X ₁ = x _n	
	variável X ₂ X ₂ = x ₁ ... X ₂ = x _t		variável X ₂ X ₂ = x ₁ ... X ₂ = x _t		variável X ₂ X ₂ = x ₁ ... X ₂ = x _t		variável X ₂ X ₂ = x ₁ ... X ₂ = x _t	
Y=y ₁	m ₁₁₁	m _{1t1}						
.....								
Y=y _j						m _{jit}		
Y=y _d								m _{ntd}

modelo log-linear: $\log[m_{jit}] = \alpha + \lambda_j^Y + \lambda_i^{X_1} + \lambda_t^{X_2} + \lambda_{ij}^{X_1Y} + \lambda_{it}^{X_2Y} + \lambda_{it}^{X_1X_2} + \lambda_{ijt}^{X_1X_2Y}$, onde:

$\lambda_i^{X_1}$ se refere ao efeito da variável X₁ sobre a frequência da casela quando assume o valor X₁ = x_i; $\lambda_{ij}^{X_1Y}$ se refere ao efeito da interação entre a variável X₁ quando assume o valor x_i e a variável dependente Y quando assume o valor Y=y_j; $\lambda_{it}^{X_1X_2}$ se refere ao efeito da interação entre as variáveis independentes X₁ e X₂ sobre a frequência da casela. Sob a hipótese de independência o modelo se torna: $\alpha + \lambda_j^Y + \lambda_i^{X_1} + \lambda_t^{X_2}$.

A4.3 - Comandos do SAS usados para ajustar o modelo

logístico acumulado com variável dependente multinomial :

```
DATA;  
  INPUT NUM V3 V4 V5 V6 V9 V10 V11 V14 V15 V16 PESOG1 PESOG2 V5PV3Q  
  V4PV3Q V10PV3Q V7 V8 V12 V13 V18 V20 V2 V19;  
  CARDS;  
  306 79 24 25 27 45 27 39 51 54 39 2 3 20 20 22 1 1 1 1 1 1 1 2  
  . . . . .  
PROC LOGIST K=2;  
  MODEL V19=V3 V4 V5 V6 V9 V10 V11 V14 V15 V16 PESOG1 PESOG2 V5PV3Q  
  V4PV3Q V10PV3Q V7 V8 V12 V13 V18 V20 V2/  
  SELECTION=FORWARD  
  SLE=0.12;  
RUN;
```

onde K=2 porque a variável dependente V19 tinha 3 níveis.

A4.4 - Aplicação da técnica CHAID:

Para a aplicação da técnica CHAID, o conjunto de dados deve estar em um arquivo tipo ASCII ou em um arquivo criado pelo SPSS com a extensão SAV. A última forma tem as vantagens de facilidade de manuseio do programa SPSS na definição das variáveis, formato, tamanho etc...

Depois que os dados são lidos pelos comandos “File/Open”, o programa pergunta ao usuário qual é a variável dependente, qual é o tipo de preditor que cada variável independente pode se enquadrar(monotônica, livre ou flutuante). Além desses, os próximos comandos para que o CHAID prepare os dados para a análise, são os comandos “Scan dat” e “ordinal” para analisar os dados segundo o número de categorias e possíveis tabelas de contingência, considerando a variável dependente como sendo ordinal.

A4.5 - Aplicação da técnica CART:

Para a utilização da técnica CART usada neste trabalho, deve-se aplicar o programa “arvclass” ao arquivo construído em ASCII, que deve conter os dados em valores numéricos, em colunas para cada variável. Antes dos dados, no mesmo arquivo são colocados os comandos necessários para indicar: o tamanho do conjunto de dados, se deve ser usada uma amostra teste, o número de variáveis, quais são as variáveis categóricas, o valor de beta, o número de classes da variável dependente, qual a medida de impureza a ser usada (a número 3 é a de Gini), se houver duplicação de resultados em termos de partição de um dado nó, imprimir o nó e as variáveis que levaram a tais resultados, se o usuário deseja fazer perguntas fixas em cada partição para cada variável, se o usuário deseja usar custo de má-classificação, e em caso afirmativo deve-se fornecer a matriz de custo. Em cada comando deve-se colocar s para sim ou n para não. Com relação ao comando

desenha a arvore o programa desta versão só tem a opção *n*:

```
NUMERO DE OBSERVACOES= 530
USA AMOSTRA DE PROVA: n
NUMERO DE VARIAVEIS=17
USA VARIAVEIS CATEGORICAS: s 1 6 7 11 12 16 17 0
USA BETA: s 0.0175 0
NUMERO DE CLASSES=3
MEDIDA DE IMPUREZA=3
CALCULA SUBSTITUTAS: n
IMPRIME DUPLICACOES: n
INDUZ PERGUNTAS: n
USA CUSTO DE CLASSIFICACAO: n
SALVA STATUS A CADA NO': n
REINICIO DE EXECUCAO INTERROMPIDA: n
CALCULA ERRO POR CROSS VALIDATION: n
DESENHA A ARVORE: n
353 1 78 24 26 20 1 3 48 31 39 1 3 58 55 40 1 1 1
. . . . .
```

Esse arquivo deve ter extensão .DAT, ser colocado no mesmo diretório onde está o CART, e para sua leitura e análise pelo CART, usa-se

o comando **arvclass** **NOMEDOARQUIVO** sem a extensão **.DAT**. O “output” com a árvore resultante, é colocado no arquivo em formato ASCII com o mesmo nome do ARQUIVO que têm os dados, agora com extensão **LST**, o qual além de descrever os parâmetros usados no arquivo com extensão **DAT**, informa como é constituído cada nó, qual a variável usada para fazer a partição do nó, quando não é terminal, como é feita a partição do nó indicando o ramo da esquerda, quais são os nomes dos nós resultantes da partição, medida de erro com base na medida de impureza escolhida no arquivo extensão **DAT**.

A seguir tem-se um exemplo da matriz de custo de má-classificação do tipo proposto por BREIMAN e col(1984) aplicável no CART, onde os valores reais da variável dependente estão nas colunas e as classificações nas linhas:

		valores reais da variável Y		
		PIG	AIG	GIG
c l a s s i f i c a ç ã o	PIG	0	1	1
	AIG	3	0	2
	GIG	4	1	0

A4.6 - Aplicação da técnica SPLUS:

Para a aplicação da técnica SPLUS, os dados foram colocados em um arquivo à parte, com os nomes das variáveis segundo a ordem em que as colunas de dados das mesmas aparecem. Para a aplicação da técnica utilizou-se do programa:

```

NUTRICA0<-scan("b:nutric10.dat",skip=1,what=list(id="",V2="",V3=0,V4=0,V5=0,V6=0,
V7="",V8="",
v9=0,V10=0,V11=0,V12="",V13="",V14=0,V15=0,V16=0,V18="",V19="",V20="",
pgr1=0,pgr2=0,V5pV3=0, V4pV3=0,V10pV3=0))
nutric <- data.frame(NUTRICA0)
attach(nutric)
win.graph()
arvore<-
tree(V19~V2+V3+V4+V5+V6+V7+V8+v9+V10+V11+V12+V13+V14+V15+V16+V18+
V20+pgr1 +pgr2+V5pV3+V4pV3+V10pV3)
plot.tree(arvore)
text(arvore)

```

Onde "b:nutric10.dat" é o arquivo com os dados a serem analisados, e o comando skip=1 desconsidera a primeira linha que contém os nomes das variáveis, para a leitura dos dados. Esse programa fornece a árvore inteira, sem poda. Para efetuar a poda usou-se os comandos:

```

poda<-prune.tree(arvore,k)
plot.tree(poda)
text(poda)

```

Onde k indica o valor da deviance do modelo, e quanto maior for o seu valor, maior é a árvore.

Além do diagrama da árvore o SPLUS fornece informações sobre o "status" de cada nó e ramificação, como por exemplo: node), split, n, deviance, yval, (yprob)

** denotes terminal node*

```

1) root 565 702.100 AIG ( 0.8035 0.001770 0.13980 0.05487 )
2) V40<27.5 71 88.120 AIG ( 0.7465 0.000000 0.01408 0.23940 ) *
3) V40>27.5 494 567.400 AIG ( 0.8117 0.002024 0.15790 0.02834 )
6) V7:DES,NOR 382 358.500 AIG ( 0.8691 0.002618 0.09948 0.02880 )
12) V45<47.5 6 5.407 GIG ( 0.1667 0.000000 0.83330 0.00000 ) *
13) V45>47.5 376 334.500 AIG ( 0.8803 0.002660 0.08777 0.02926 )
26) V34<22.5 75 18.440 AIG ( 0.9733 0.000000 0.00000 0.02667 ) *
27) V34>22.5 301 300.000 AIG ( 0.8571 0.003322 0.10960 0.02990 )
54) V6<26.5 276 279.100 AIG ( 0.8478 0.000000 0.11960 0.03261 ) *
55) V6>26.5 25 8.397 AIG ( 0.9600 0.040000 0.00000 0.00000 ) *
7) V7:OBS 112 170.900 AIG ( 0.6161 0.000000 0.35710 0.02679 ) *

```

onde a primeira linha indica o nó raiz com $N=565$, medida de heterogeneidade deviance=702,100, a categoria classificada naquele nó (AIG), e entre parenteses as proporções nesse nó de quatro classes de categorias da variável Y.

A4.7 - Aplicação da técnica KS:

Para que os dados sejam lidos pelo programa KS, e se os mesmos não foram salvos pelo próprio KS, a forma mais simples de leitura é importar os dados que estejam em um arquivo tipo dBASE ou Lotus 123 pelo comando **"Import"**. Após a importação dos dados, define-se os dados a serem trabalhados selecionando as variáveis a serem incluídas na análise pelo comando **"Select Fields"**, alterando ou corrigindo alguns dados pelo comando **"Edit data"**, transformando certos valores de uma dada variável pelo comando **"Map Data"**. Além disso deve-se verificar se o tipo de variável registrado em **"Edit view"**, contínua/categórica, e o tipo de regra de partição, livre/flutuante/monotônica, precisa ser alterada, e indicar se existem "missings". Se os dados foram salvos pelo KS, pode-se lê-los diretamente pelo comando **"Open"**.

Definido o conjunto de dados a ser analisado, pode-se iniciar a análise indicando o nível de significância do teste pelo **filtro** (prediction=0,05 ou especificando o valor), e usando o comando **"Grow"** para a construção da árvore.

A5.1 - DISTRIBUIÇÕES DAS VARIÁVEIS MEDIDAS:

Na figura A5.1.1 se encontra a distribuição das estaturas das gestantes. A altura média das gestantes foi de 1,55 m, coincidindo com a média encontrada por Siqueira(1979) nessa população. A menor estatura da amostra teve o valor de 1,36 m e a maior estatura teve o valor de 1,75m. Nota-se

que a distribuição das estaturas é simétrica e próxima de uma normal. No apêndice A5.2 estuda-se a distribuição das estaturas para cada Local, a diferença entre as mesmas e o efeito da junção entre locais.

V3 = Estatura da gestante

A figura A5.1.2 descreve a distribuição dos pesos pré-gestacionais das gestantes.

V4=Peso pré-gestacional

Existem 496 dados na distribuição apresentada na figura A5.1.2 porque foram perdidos 69 registros dessa variável, por falta de informação, ou falta de lembrança, por parte da gestante, do seu peso antes de ficar grávida. A distribuição desses pesos apresentou valor mínimo igual a 37,00 kg e valor máximo

igual a 94,00 kg. Nota-se que a distribuição do peso pré-gestacional é mais concentrada à esquerda do que uma distribuição normal, e apresenta valores altos na faixa de 90kg na cauda direita da distribuição.

V5=Peso da gestante na primeira consulta

A figura A5.1.3 apresenta valor mínimo de 39.00 kg e máximo de 97.50 kg, a distribuição do peso na primeira consulta mantém a mesma tendência do peso pré-gestacional de se concentrar mais à esquerda. No apêndice A5.2 são estudadas as distribuições dos pesos pré-gestacionais das gestantes para cada Local, a diferença entre as mesmas e o efeito da junção entre Locais.

V6=tempo de gestação na primeira consulta

Nota-se na figura A5.1.4 o valor mínimo do tempo de gestação de 1 semana e o valor máximo de 36 semanas. Essa distribuição tem uma assimetria à direita, com a frequência crescendo mais rapidamente antes de $V6=12$ decrescendo a partir do mesmo. Mas isso não indica que esse é o ponto central da distribuição, pois a média de $V6$ é 16.1 e a mediana é 15 semanas.

Isso indica que em média, a demora para vir ao serviço de pré-natal é maior do que 12 semanas, e que existem mais primeiras consultas de pré-natal além de 12 semanas do que antes. Além disso, tais fatos podem indicar provavelmente que a motivação para ir ao pré-natal aumenta à medida que a gestação se aproxima de 12 semanas e diminui à medida que a gestação se afasta além de 13 semanas, talvez devido a orientações tendenciosas nesse sentido.

V10 = peso da gestante na última consulta

A distribuição dos pesos das gestantes na última consulta mantém o mesmo comportamento das distribuições dos pesos pré-gestacionais e dos pesos na primeira consulta, ou seja, com assimetria à direita. A mediana apresentou o valor 65,2 kg, o valor mínimo foi de 45,6 kg e o valor

máximo foi de 99,9 kg. SIQUEIRA(1981) notou que o peso no final da gravidez é a variável de maior influência sobre o peso do RN.

V11 = Tempo de gestação na última consulta

A distribuição dos tempos de gestação na última consulta se mostrou assimétrica à esquerda, talvez devido ao fato que algumas gestantes não retornaram para nova consulta quando estavam com o tempo de gestação mais próximo do parto. O tempo mínimo encontrado foi de 22 semanas e o máximo foi de 43 semanas.

V15 = tempo de gestação em dias por ocasião do parto

V16 = tempo de gestação em semanas por ocasião do parto

Com média de 273,7 dias ou 39,1 semanas e mediana de 275 dias ou 39 semanas, as duas distribuições de tempo de gestação se apresentam ligeiramente assimétricas à esquerda, com mesmo formato. Pela análise visual não se pode afirmar que há diferença no uso de uma das duas escalas, mas isso só poderá ser confirmado com estudos adicionais.

V18 = Sexo do recém-nascido

Não há diferença nessa amostra com relação ao sexo do recém-nascido.

V20 = tipo de parto

Essa distribuição confirma a tendência contemporânea de grande preferência pelo parto com cesariana. TANAKA(1986) encontrou 25% de cesária, 17% de fórcepe e 55% de normal na população de Cotia.

Descrição das principais variáveis calculadas:

PESOG1 = Peso Ganho 1:

A média de peso ganho medido em relação à primeira consulta foi de 8469 gramas, tendo valor mínimo de -170 gramas(3 casos com PESOG negativo) e máximo de 24500 gramas.

PESOG2= Peso Ganho 2

A média de peso ganho medido em relação ao peso pré-gestacional foi de 11383 gramas, tendo um valor mínimo de -1150 gramas e máximo de 28300 gramas.

PESOG1 = Peso ganho relativo 1

A média de ganho de peso relativo à estatura, medido em relação ao peso na primeira consulta foi de 3,49 kg/m², tendo um valor mínimo de -1,0 kg/m² e máximo de 10,0 kg/m². A correção processada no peso ganho concentra mais a distribuição.

PESOG2 = Peso Ganho Relativo 2

FIGURA A5.1.14

A média de ganho de peso relativo à estatura, medido em relação ao peso pré-gestacional foi de 4,70 kg/m², tendo um valor mínimo de -4,0 kg/m² e máximo de 11,0 kg/m². Houve 69 casos sem registro dessa variável, devido ao fato que PESOG2 envolve no seu calculo a variável V4(peso pré-gestacional).

A5.2.1 - Verificação se a inclusão dos dados de COT altera, a distribuição dos Pesos pré-gestacionais:

FIGURA A5.1.1

com distribuições praticamente coincidentes.

peso pré-gestacional	só FSP	FSP + COT
média	5536,659	5528,851
intervalo de confiança(95%)	[5448,43 ; 5624,88]	[5443,64 ; 5614,06] **
mínimo	3700	3700
máximo	9400	9400
tamanho do campo de variação	5700	5700

** = indicando que os intervalos de confiança para a média ao nível de 95% são praticamente coincidentes, levando à não alteração do peso pré-gestacional médio quando da inclusão dos dados de COT. Tal resultado era esperado devido ao teste de hipóteses para diferença de médias em A4.2, embora haja a ressalva da diferença de tamanhos das sub-amostras.

A5.2 - Verificação se a inclusão dos dados de COT altera a distribuição das estaturas:

com distribuições praticamente coincidentes

estatura	só FSP	FSP + COT
média	155,68	155,49
intervalo de confiança(95%)	[155,15 ; 156,21]	[154,97 ; 156,00] **
mínimo	136cm	136cm
máximo	175cm	175cm
tamanho do campo de variação	39cm	39cm

** = indicando que os intervalos de confiança para a média ao nível de 95% são praticamente coincidentes, levando à não alteração da estatura média quando da inclusão dos dados de COT, embora o teste de hipótese para diferença de médias em A4.2 tenha identificado diferença entre os locais.

Tais resultados provavelmente seriam alterados se o tamanho da sub-amostra de COT fosse maior. Tanaka⁶⁹ encontrou em seu trabalho uma média de altura para COT de 1,54 m e para FSP de 1,56 m.

A5.3 - Verificação se a inclusão dos dados de COT altera a distribuição dos Pesos dos RN:

com distribuições praticamente coincidentes

peso do RN	só FSP	FSP + COT
média	3250,28	3215,77
intervalo de confiança	[3166,44 ; 3244,13]	[3177,79 ; 3253,75] **
mínimo	1750g	1750g
máximo	4600g	4600g
tamanho do campo de variação	2850g	2850g

** = indicando que os intervalos de confiança para a média ao nível de 95% são praticamente coincidentes, levando à não alteração do peso do RN médio quando da inclusão dos dados de COT, o que era de se esperar devido ao teste de igualdade de médias visto em A4.2.

A5.3.3 - APLICAÇÃO DA REGRESSÃO LOGÍSTICA ACUMULADA PARA A SUB-AMOSTRA FSP:

Considerando apenas a sub-amostra FSP, o ajuste do modelo, incluiu as variáveis no modelo na seguinte ordem:

passo	variável incluída	b	Qui-quadrado	valor p
1	V10	-0.0005	9.190	0.0024
2	V16	0.249	15.579	0.0001
3	V7	-0.726	5.856	0.0155
	(constante 1)	-8.023	10.024	0.0015
	(constante 2)	-2.777	1.249	0.2636

confirmando a vantagem encontrada nas análises anteriores da curva de Siqueira na primeira consulta.

A5.4 - Aplicação das técnicas de classificação e regressão por árvore na amostra da FSP:

APLICAÇÃO DA TÉCNICA SPLUS:

Variável dependente=V17:

V19=variável dependente

Comparando-se a figura A5.4.1 com a figura 5.3.12 nota-se que, considerando a variável dependente como sendo a V17(peso do RN), houve ausência da variável V7(Siqueira na primeira consulta) quando considerava-se apenas FSP. Comparando-se a figura A5.4.2 com a figura 5.3.13, nota-se que, considerando a variável dependente como sendo a V19(adequação do peso do RN), as variáveis usadas nos primeiros degraus são as mesmas em ambas as figuras, estando em destaque as variáveis V10, V7, com V4 e V15 trocando de posição na hierarquia das variáveis.

APLICAÇÃO DA CHAID :

Variável dependente=V17:

FIGURA A5.4.3

Comparando a figura A5.4.3 com a figura 5.3.8, nota-se que as variáveis usadas nos primeiros degraus são as mesmas em ambas as figuras, estando presentes as variáveis V40(peso na primeira consulta), V16(duração da gestação), V20(parto) e V18(sexo), com destaque para a curva de Siqueira na primeira consulta(V7), como havia ocorrido com as duas sub-amostras conjuntamente.

Variável dependente=V19

Comparando as figura A5.4.4 e 5.3.9, nota-se que, para apenas a FSP, as variáveis referentes à curva de Siqueira, V12 e V7, trocaram de posição, aparecendo a curva de Rosso no terceiro degrau, em sua forma de escala numérica, V13. As variáveis V40, V16 e V20 foram constantes em suas posições em ambos os ajustes. Assim, pode-se dizer que a estrutura de classificação é a mesma nos primeiros degraus das árvores construídas pela técnica CHAID.

APLICAÇÃO DA CART

Variável dependente=V17:

Variável dependente=V19:

Comparando as figuras A5.4.5, A5.4.6 com as suas respectivas figuras, 5.3.10 e 5.3.11, nota-se novamente que no caso em que a variável dependente é o peso do RN, a variável referente à curva de Rosso em forma de escala numérica aparece em primeiro lugar, e para o caso em que a variável dependente é a adequação do peso do RN, a curva de Siqueira na primeira consulta, V7, continua na primeira posição, não aparecendo a curva de Rosso. As variáveis V5, V16 foram constantes em suas posições mantendo primazia em ambos os ajustes. Assim, a estrutura de classificação é praticamente a mesma nos primeiros degraus das árvores construídas pela técnica CART.

A5.4 - APLICAÇÃO DA TÉCNICA KS

variável dependente=V17

variável dependente=V19

FIGURA A5.4.8

Comparando-se as figuras A5.4.7 com 5.3.14 e A5.4.8 com 5.3.15, nota-se que considerando a variável V17 como dependente, a variável V10 é a primeira associada, enquanto com relação às curvas em comparação aparecem V12 e V7 em ambas as árvores, sem aparecer quaisquer das variáveis referentes à curva de Rosso. Quando a variável dependente é a V19, as variáveis V7 e V12 aparecem nos primeiros degraus, e a variável referente à curva de Rosso na primeira consulta aparece no quarto degrau depois das variáveis referentes à curva de Siqueira na primeira e última consultas.

APÊNDICE 6

A.6.1 - INVERSA DA FUNÇÃO DE REGRESSÃO LOGÍSTICA NA VALIDAÇÃO CRUZADA:

Conforme o modelo 4.2.3.11 o modelo logístico é composto de dois modelos de regressão paralelos, onde

$\log\left(\frac{\pi_1}{\pi_2 + \pi_3}\right) = \alpha_1 + \beta' X = L_1$ compara o nível 1 de V19 com a junção dos níveis 2 e 3 de V19.

$\log\left(\frac{\pi_1 + \pi_2}{\pi_3}\right) = \alpha_2 + \beta' X = L_2$ compara o nível 3 de V19 com a junção dos níveis 1 e 2 de V19.

então $\exp(L_1) = \frac{\pi_1}{\pi_2 + \pi_3} = k_1$

$$\exp(L_2) = \frac{\pi_1 + \pi_2}{\pi_3} = k_2 \quad \text{A6.1.1}$$

portanto, fixando a observação X_i tem-se :

$$\begin{aligned} \alpha_1 + \beta' x_i &= L_{1i} \\ \alpha_2 + \beta' x_i &= L_{2i} \end{aligned}$$

que leva para cada X_i aos valores: $k_{1i} = \exp(L_{1i}) = k_1$ e

$$k_{2i} = \exp(L_{2i}) = k_2$$

mas de A6.1.1 como $k_1 = \frac{\pi_1}{\pi_2 + \pi_3}$ então $k_1(\pi_1 + \pi_3) = \pi_1 = P(\text{PIG})$

$$k_2 = \frac{\pi_1 + \pi_2}{\pi_3} \text{ então } \frac{\pi_1 + \pi_2}{k_2} = \pi_3 = P(\text{GIG}) \quad \text{A6.1.2}$$

de onde obtém-se: $\frac{\pi_3(k_2 - k_1)}{k_1 + 1} = \pi_2 = P(\text{AIG})$

A6.1.2 é um sistema de três equações lineares com as três probabilidades como incógnitas, cuja solução é:

$$\pi_1 = \frac{k_1 + 1}{k_1 k_2 + k_1 + k_2 + 1}$$

$$\pi_2 = \frac{k_2 - k_1}{k_1 k_2 + k_1 + k_2 + 1}$$

A6.1.3

$$\pi_3 = \frac{k_2 + 1}{k_1 k_2 + k_1 + k_2 + 1}$$

A6.2 - A CONFUSÃO NAS RAMIFICAÇÕES PARA UMA ÁRVORE COM MUITOS DEGRAUS.(SPLUS COM 14.335 NÓS E 61 NÓS TERMINAIS):

A.6.3 - Contagem do número de má-classificações: Cálculo da variável Z.

Para cada uma das árvores construídas por uma das técnicas utilizando um arquivo de 509(452) observações, tomou-se o arquivo complementar de 56(113) observações e inicialmente foi construída a variável auxiliar X com os valores:

- 1 quando o nó terminal do ramo classificava o indivíduo como PIG,
- 2 quando o nó terminal do ramo classificava o indivíduo como AIG,
- 3 quando o nó terminal do ramo classificava o indivíduo como GIG.

Assim, para cada gestante calculou-se o valor da variável X usando-se expressões lógicas com operadores relacionais como por exemplo:

“if ((V7=“NOR” or V7=“OBS”) and (V10<=26)) then x=3”.

Como no fundo a variável X é indicadora da classificação do ramo, a má-classificação é dada pela variável Z:

$$Z = \begin{cases} 0 & \text{if } (v19 = x) \\ 1 & \text{if } (v19 \neq x) \end{cases}$$

de onde se tem o número total de má-classificações pela soma:

$$\sum_{i=1}^k Z_i \quad \text{para } i=\text{número do ramo e } k \text{ o total de ramos da árvore.}$$

A6.4 - APLICAÇÃO DA INVERSA DA LOGÍSTICA NA VALIDAÇÃO

CRUZADA:

Conforme A6.1, foi preciso calcular π_1, π_2 e π_3 para cada observação x_i de cada arquivo que tinha a amostra $\mathcal{L}_i$ de tamanho 113 quando a validação cruzada era de 5 partes, e de tamanho 56 quando a validação cruzada era de 10 partes. Isso foi feito calculando-se inicialmente duas novas variáveis $k_1 = \exp(L_1)$ e $k_2 = \exp(L_2)$ de A.6.1.1, e com esses valores calculou-se as probabilidades π_1, π_2 e π_3 .

Para o cálculo do número de erros de má-classificação foi considerada a variável indicadora de classificação:

$$Y = \begin{cases} 1 & \text{se } \max(\pi_1, \pi_2, \pi_3) = \pi_1 \\ 2 & \text{se } \max(\pi_1, \pi_2, \pi_3) = \pi_2 \\ 3 & \text{se } \max(\pi_1, \pi_2, \pi_3) = \pi_3 \end{cases}$$

e a variável

$$Z = \begin{cases} 0 & \text{se } Y = V19 \\ 1 & \text{se } Y \neq V19 \end{cases}$$

de modo que $\sum Z_i$ fornece o número de erros de má-classificação, e $\sum Z_i / v \times m$ fornece a taxa de má-classificação por substituição quando m é o tamanho da sub-

amostra $\mathcal{L}_i$, e a taxa de má-classificação por validação cruzada quando m é o tamanho da sub-amostra $\mathcal{L}_i$. O valor v é o número de sub-amostras da validação cruzada.

PROGRAMA USADO NO SPSS PARA CONTAGEM DE MÁ-CLASSIFICAÇÕES PARA O AJUSTE DA LOGÍSTICA:

```
GET
FILE='C:\Vael\nutricao\cross\nordem10.SAV'.
IF (V7=" DES") V7ind = 1 .
IF (V7=" NOR") V7ind = 2 .
IF (V7=" OBS") V7ind = 3 .
IF (V8=" DES") V8ind = 1 .
IF (V8=" NOR") V8ind = 2 .
IF (V8=" OBS") V8ind = 3 .
IF (V12=" DES") V12ind = 1 .
IF (V12=" NOR") V12ind = 2 .
IF (V12=" OBS") V12ind = 3 .
IF (V19=" PIG") V19ind = 1 .
IF (V19=" AIG") V19ind = 2 .
IF (V19=" GIG") V19ind = 3 .
EXECUTE .
COMPUTE k1 = exp(3.4934-0.0375*V4-0.8033*V7ind+0.4862*V8ind-0.4526*V12ind) .
COMPUTE k2 = exp(5.0929-0.0375*V4-0.8033*V7ind+0.4862*V8ind-0.4526*V12ind) .
COMPUTE p3 = (k1+1)/(k1*k2+k1+k2+1) .
COMPUTE p2 = (k2-k1)/(k1*k2+k1+k2+1) .
COMPUTE p1 = k1*(k2+1)/(k1*k2+k1+k2+1) .
EXECUTE .
COMPUTE y = 1 .
IF (max(p1,p2,p3)=p2) y = 2 .
IF (max(p1,p2,p3)=p3) y = 3 .
COMPUTE z = 0 .
EXECUTE .
IF (y ~= V19ind) z = 1 .
FREQUENCIES
VARIABLES=z .
CROSSTABS
/TABLES=V19ind BY y z
/FORMAT=AVALUE NOINDEX BOX LABELS TABLES
/CELLS= COUNT .
EXECUTE .
```