Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Monitoria PMG
Artur de Souza Plácido Teixeira
Pedro Henrique Chaves Cardoso
Pedro Henrique de Matos Araujo
Rafael Dias de Brito Oliveira
e Adrian Luna (orientador)

Sumário

1 Introdução e Contexto Histórico **3**
1.1 Ciência social de dados . . . 3
1.1.1 Etnografia na Ciência de Dados: a Inteligência Artificial como Ator Social 4
1.2 História da análise de dados categóricos . . . 5
1.2.1 A regressão linear . . . 5
1.2.2 Grandes nomes da análise de dados categóricos . . . 5
1.2.3 A regressão logística: um modelo linear com fatores . . . 7
1.3 Software RCommander . . . 8

2 Modelos de Regressão Linear **9**
2.1 Introdução aos Modelos de Regressão Linear . . . 9
2.2 Modelos de Regressão Linear no Rcmdr: Bases de Dados e Janelas de Comandos 9
2.3 Resultados e Interpretação . . . 11

3 Modelos Lineares com Fatores **13**
3.1 Fórmulas de Modelos Lineares . . . 13
3.2 Simbologia do RCommander . . . 13
3.3 Princípio da Marginalidade . . . 14
3.4 Interpretação dos coeficientes . . . 15
3.5 Variáveis Dummy . . . 17
3.6 Exemplos no RCommander . . . 17
3.6.1 Exemplos usando o Canadian Occupational Prestige Data . . . 17
3.6.2 Exemplos usando o Banco de Dados *stress* . . . 21
3.6.3 Exemplos usando o Banco de Dados hsb2 . . . 23
3.7 Outros Contrastes para Fatores . . . 27

4 Modelo Linear Generalizado **30**
4.1 Interpretação dos coeficientes . . . 33
4.2 Exemplo usando o banco de dados binary . . . 34
4.3 Outros Modelos de Regressão . . . 37
4.3.1 Modelos de Regressão Binária: Logit . . . 38

5 Visualizando Modelos Lineares e Lineares Generalizados **41**
5.1 Resíduos Parciais em Gráficos de Efeitos . . . 43

A Ajustando Splines de Regressão e Polinômios **46**
A.1 Termos Polinomiais . . . 46
A.2 Regressão Spline . . . 48

B Intervalos de Confiança e Teste de Hipótese para Coeficientes **50**
B.1 Intervalos de Confiança 50
B.2 Tabelas de Análise de Variância e Análise de Desvio 51
B.3 Testes Comparando Modelos 52
B.4 Testando Hipótese Linear 54

C Escolha de Modelo **58**

Capítulo 1

Introdução e Contexto Histórico

1.1 Ciência social de dados

De uns tempos pra cá, a Ciência de Dados vem ganhando importância e relevância no contexto disciplinar. Como dito por Wil van der Aalst em seu capítulo *"Data Science in Action"*, a ciência de dados pode ser vista como uma amálgama de áreas do conhecimento como a estatística, mineração de dados, banco de dados e distribuição de sistemas [11]. Sua área de abrangência é multidisciplinar, quando levamos em conta que ela pode ser utilizada em benefício de qualquer campo do saber, seja ele matemático, biológico ou social. Mas o que é, de fato, ciência de dados?

Figura 1.1: Wil van der Aalst

Pode-se defini-la como estudo disciplinado de informações e dados que dizem respeito a um determinado assunto e as visões possíveis relacionadas com essa temática. Um cientista de dados, portanto, deve ser capaz de: transformar dados em informações, formular problemas e hipóteses para guiar a resolução dos mesmos, e aplicar os modelos estatísticos adequados para cada análise que precisa ser feita.

Entendido o conceito de ciência de dados, resta saber de que forma as ciências sociais podem usar dessa abordagem para otimizar seus estudos, que envolvem múltiplos constructos como *Big Data*, interação homem-máquina, trabalho corporativo apoiado por computador, software R, entre outros. Para isso, é importante relembrar conceitos como o de epistemologia, etnografia e etnomologia, que são muito utilizados na chamada "ciência social de dados".

1.1.1 Etnografia na Ciência de Dados: a Inteligência Artificial como Ator Social

O exemplo de como a etnografia pode ser abordada como ciência de dados, usando para tanto, a relação íntima entre etnografia e a famosa inteligência artificial.

Pode-se conceituar etnografia como **estudo descritivo das diversas etnias, de suas características antropológicas, sociais etc.** E também buscar a definição da fenomenologia sociológica como "**A fenomenologia sociológica de Alfred Schutz (...) busca o significado subjetivo da conduta social (...) o que implica pensar como os fenômenos se apresentam à consciência**" [3].

Vamos conceituar também o termo "inteligência artificial (IA)", que pode ser entendido, no campo da ciência de dados, como uma ferramenta capaz de predizer soluções para possíveis problemas, utilizando, para esse fim, os dados disponíveis e aprendendo com eles. Atualmente, somos capazes de identificar a IA em diferentes contextos e situações no nosso cotidiano; mas a verdade é que, ainda hoje, o exemplo que mais chama a atenção é o dos carros autônomos, que são capazes de "dirigir" por conta própria [6].

No entanto, com certeza você, leitor, já deve ter ouvido falar no chamado dilema dos carros autônomos - basicamente uma versão atualizada do problema sugerido por Philippa Foot ainda em 1967[5], que envolvia um bonde desgovernado, pessoas prestes a serem atingidas e você, com o poder de salvar/condenar um grupo de trabalhadores ou um inocente desavisado. Tendo isso em mente fica fácil perceber como a etnografia e a sociologia fenomenológica estão relacionadas com discussões de cunho ético [6] e em tomadas de decisões que, de maneira indireta ou direta, implicam na construção da IA como componente interdisciplinar no meio acadêmico e também responsável pelo aumento da qualidade de vida da população como um todo.

Figura 1.2: Philippa Foot

A IA também tem muitas implicações quando o assunto é a estatística. Um bom exemplo é a utilização de métodos de inteligência artificial, como o de redes neurais ou algoritmo de *random forest*, para identificar um fator preditor - que envolve de certa forma a interação entre variáveis dependentes e independentes.

Tais algoritmos precisam que seus comportamentos sejam bem documentados e entendidos para que a vigilância de suas atividades possa ser feita de forma adequada. Contudo, devido ao rápido desenvolvimento com que esses sistemas são capazes de atingir, os próprios desenvolvedores apontam a necessidade de várias frentes investigativas trabalharem juntas para que essa monitorização seja feita [2].

Toda essa problemática culmina em estudos quase como de caráter etnográfico puro. Em

muitas discussões recentes, especialistas sugerem que o estudo de máquinas inteligentes precisa ser conduzido adotando estas como "atores que apresentam padrões ecológicos e comportamentais particulares", ao invés de apenas serem tratados como artefatos de engenharia [2].

Um fato curioso é que outros pesquisadores, por outro lado, argumentam que o estudo comportamental de máquinas inteligentes deve ser feito pautado na etologia, tratando dele como uma espécie de comportamento animal [2]. Diante disso cabe o questionamento: Estamos realmente prontos para as inovações da Ciência de Dados, ou ainda subestimando seus avanços? Esse é outro motivo pelo qual aprender sobre essa nova temática vem sendo cada vez mais necessário.

Além disso, a forma com que os diferentes algoritmos inteligentes interagem também é um outro exemplo de onde a etnografia e a etnometodologia podem ser aplicadas juntamente ao conceito de *Machine Learning* [2]. No entanto, voltando à questão dos modelos de IA úteis para descrever preditores, modelos estatísticos tradicionais também são úteis para ilustrar, de forma simples e menos controversa, como variáveis dependentes e independentes se relacionam. Uma dessas ferramentas mais tradicionais e extremamente poderesa são as regressões lineares.

1.2 História da análise de dados categóricos

1.2.1 A regressão linear

Sir Francis Galton (antropólogo, meteorologista, matemático e estatístico inglês) foi quem usou o termo "regressão" pela primeira vez em 1885. Utilizando essa ferramenta juntamente com os pressupostos da seleção natural.[4]

Atualmente os diversos modelos que levam o nome de "análise de regressão", são utilizados para prever comportamentos com base na associação entre duas ou mais variáveis [1]. Na regressão linear simples, temos uma variável dependente que se relaciona com uma variável independente; na regressão linear múltipla, uma variável dependente que se relaciona com múltiplas variáveis independentes; e também temos a regressão logística, que, diferentemente das regressões que foi dito anteriormente, tem por característica sua variável dependente ser categórica. O fato da variável dependente ser categórica ou qualitativa nos interessa muito, e iremos focar nesse ponto no trabalho a seguir.

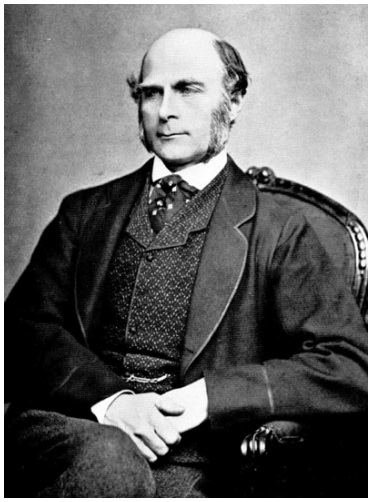

Figura 1.3: Sir Francis Galton

1.2.2 Grandes nomes da análise de dados categóricos

Karl Pearson [1] foi um estatístico, criador do método qui-quadrado de distribuição e com notáveis trabalhos em tabelas de contingência. Em 1900, muito do que se produzia sobre o tema

de “análise de dados categóricos” (ADC) estava restrito à discussões acerca de como resumir associações (associação é um método para descobrir o quanto duas variáveis se relacionam), foi então que em 1904, Pearson chamou de *contingência* a “medida do desvio total da classificação da probabilidade independente”, apresentando também medidas para descrever e estimar correlação [1] (uma forma de descrever de que maneira uma associação ocorre).

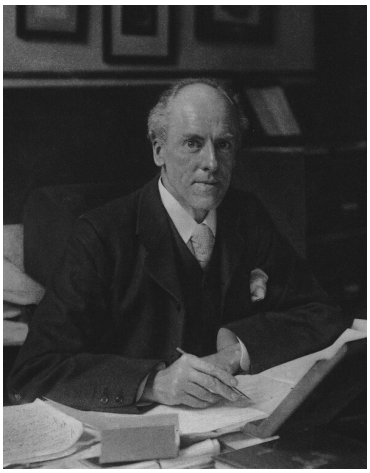

Figura 1.4: Karl Pearson

Depois de Karl, diversas outras contribuições importantes apareceram e discussões riquíssimas foram levantadas, contribuindo e muito para o desenvolvimento da ADC. Exemplo disso foram as discussões entre Pearson e Gorge Yule e mais posteriormente entre o mesmo Pearson e Ronald Fisher [1] - este último, o responsável pela introdução dos graus de liberdade para caracterizar as distribuições qui-quadrado de Karl. Fisher no entanto, é um excelente exemplo de como colocar os métodos de ACD em prática e, através de uma pesquisa na área da genética (onde também tinha domínio), demonstrou que o método qui-quadrado de Pearson em certas tabelas de contingência não era adequado.

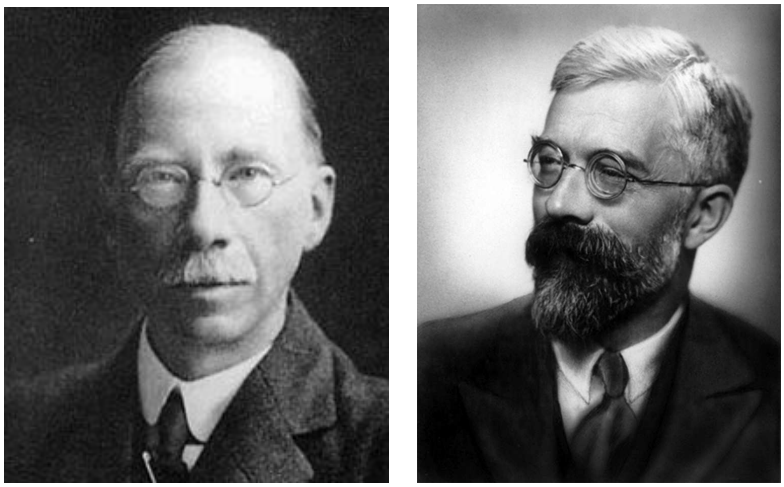

Figura 1.5: Gorge Udny Yule (esquerda) & Ronald Fisher (direita).

Ainda nesse meio tempo, dois estudiosos: William Kruskal e Leo Goodman, se destacaram no trabalho sobre medidas de associação para tabelas de contingência, cada um trabalhando por seus próprios métodos [1].

Figura 1.6: William Kruskal (esquerda) & Leo Goodman (direita).

1.2.3 A regressão logística: um modelo linear com fatores

Mas foi em 1938 que Fisher e seu amigo Frank Yates propuseram uma expressão que, possivelmente, seria capaz de transformar parâmetros binomiais (respostas apenas como 'sucesso' ou 'fracasso') em dados binários para facilitar suas análises. Em 1944, o estatístico e médico Joseph Berkson gostou tanto da expressão que deu o nome de "logit" para essa transformação, contribuindo muito para a popularização da então chamada "regressão logística" [1].

Figura 1.7: Frank Yates

Jerome Cornfield, Danish, Georg Rasch, Daniel McFadden, Peter McCullagh, os bioestatísticos NormanBreslow e Ross Prentice (que aplicaram a regressão logística em estudos de caso-controle), e diversos outros estudiosos e estudiosas também foram importantes nomes na construção da história da regressão logística durante os anos. No entanto, foi R. W. M. Wedderburn e John Nelder que, em 1972, introduziram o conceito de "modelos lineares generalizados" [1].

Esta última contribuição unificou os modelos de regressão logística com diversos outros modelos, permitindo uma compreensão tanto individual quanto coletiva deles.

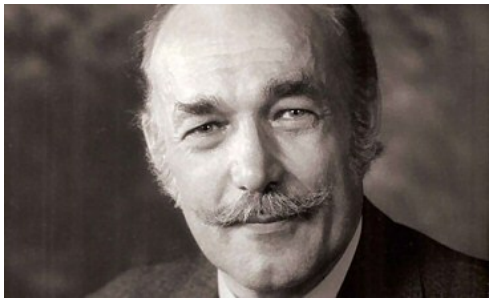

Figura 1.8: John Ashworth Nelder

É importante ressaltar que desde a ciência de dados, sua ligação com as ciências sociais, a estatística e os modelos de regressão será mais bem abordada no restante do trabalho, em que utilzaremos a interface do RCommander.

1.3 Software RCommander

Além da estatística básica, os modelos de regressão de vários tipos estão no cerne dos métodos estatísticos aplicados. Os modelos de regressão rastreiam como a distribuição de uma variável de resposta (ou "dependente") - ou algumas características-chave dessa distribuição, como sua média - está relacionada aos valores de uma ou mais variáveis explicativas ("independentes"). A regressão linear de mínimos quadrados é normalmente introduzida em um curso de estatística básica, enquanto uma consideração mais geral de modelos estatísticos lineares para respostas normalmente distribuídas e modelos lineares generalizados para respostas não normalmente distribuídas é o assunto de um segundo curso em estatísticas aplicadas (ver, por exemplo, Fox, 2016 ou Weisberg, 2014).

O software R é de fato muito utilizado por cientistas de dados, especialmente quando o assunto é a analise estatística de dados. O RCommander por sua vez, é uma interface interativa e de fácil manipulação, que permite ao usuário realizar os complexos modelos estatisticos do software, através de menus simples e caixas de texto explicativas.

Nessa apostila fizemos uso do do capítulo de regressões lineares pertencente ao material criado e idealizado por Fox [8], intitulado *"Using the R Commander. A Point-and-Click Interface for R"* você irá encontrar, no primeiro capítulo, as coordenadas para como ajustar modelos de regressão linear e linear generalizada no RCommander e como realizar cálculos adicionais em modelos de regressão, uma vez que tenham sido ajustados aos dados. Ao tratar os modelos estatísticos como objetos sujeitos a computação posterior, Fox rencoraja um estilo de modelagem estatística em que o analista de dados se envolve em uma conversa de ida e volta com os dados.

Capítulo 2

Modelos de Regressão Linear

2.1 Introdução aos Modelos de Regressão Linear

Conforme mencionado, a regressão linear de mínimos quadrados é normalmente abordada em um curso de estatística básica. O modelo de regressão linear normal é descrito da forma que se segue:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i \\ &= E(yi) + \varepsilon_i \end{aligned}$$

onde y_i é o valor da variável de resposta para as n observações amostradas independentemente; x_i, x_2, ... , x_k os valores de k variáveis explicativas; e os erros ϵ_i são normalmente e independentemente distribuídos com médias 0 e variância constante, $\epsilon_i \sim$ N(0, σ^2). Ambos y e x são variáveis numéricas, e o modelo assume que o valor médio de y é um função linear - isto é, uma soma ponderada simples - de x. Se houver apenas um x (isto é, k = 1), então a equação demonstrada acima, é chamada de modelo de regressão linear simples; se houver mais de um x (k > 1), então é chamado de modelo de regressão linear múltipla. Os modelos lineares são estimados de maneira ótima pelo método dos mínimos quadrados, produzindo o modelo ajustado.

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_k x_k$$

Onde ŷ é o valor ajustado e ϵ_i o resíduo da observação i. O critério dos mínimos quadrados seleciona os valores de β que minimizam a soma dos resíduos quadrados, $\sum \epsilon_i^2$. Os coeficientes de regressão de mínimos quadrados são facilmente calculados e, além de ser estatisticamente relevantes. Além disso, os teste de hipóteses baseados nas estimativas de mínimos quadrados são muito simples, vide, por exemplo, a estatística F, que usamos para avaliar a significância estatística dos β's estimados.

$$F = \frac{MSm}{MSr}$$

Sendo MSm os Quadrados Médios para o Modelo e MSr Quadrados Médios Residuais.

2.2 Modelos de Regressão Linear no Rcmdr: Bases de Dados e Janelas de Comandos

Continuando, a maneira mais simples de ajustar um modelo de regressão linear no RCommander é pela caixa de diálogo Regressão Linear. A mérito de exemplo, será usado os dados de

Prestígio Ocupacional de Duncan. O conjunto de dados de Duncan reside no pacote do **car** e, portanto, é possível ler os dados no RCommander através da sequência:

Dados* → *Conjunto de dados em pacotes* → *Ler dados de pacote "attachado"

Em seguida, selecione:

Estatísticas* → *Ajustes de Modelos* → *Regressão Linear

Isso irá produzir a caixa de diálogo abaixo. É importante ressaltar que esse banco de dados mede o alcance social que cada carreira proporciona ao indivíduo, de onde temos o conceito sociológio de "Prestígio Ocupacional".

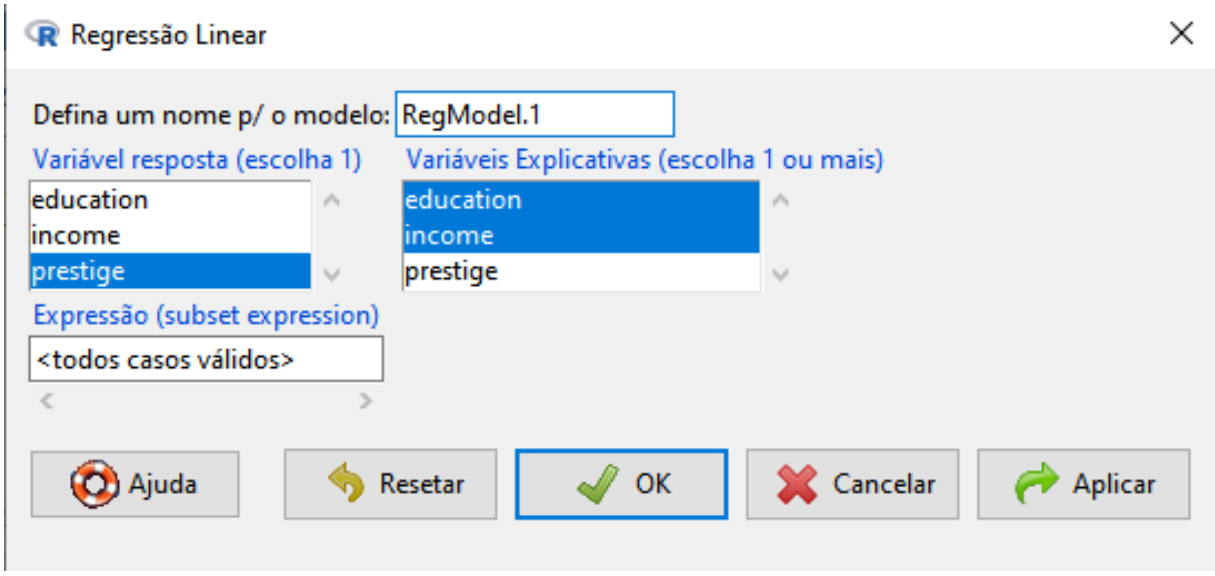

Figura 2.1: Ajuste do modelo linear *prestige* ∼ *education* + *income* ajustada para o banco de dados *Duncan*

Para completar a caixa de diálogo, clique em *prestige* na lista Variável resposta e, com a tecla Ctrl pressionada, clique em *education* e *income* na lista Variáveis explicativas. Finalmente, ao pressionar o botão OK produz a saída mostrada na Figura 2.2.

2.3 Resultados e Interpretação

```
Rcmdr>  RegModel.1 <- lm(prestige~education+income, data=Duncan)

Rcmdr>  summary(RegModel.1)

Call:
lm(formula = prestige ~ education + income, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-29.538  -6.417   0.655   6.605  34.641

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) -6.06466    4.27194  -1.420      0.163
education    0.54583    0.09825   5.555 0.00000173 ***
income       0.59873    0.11967   5.003 0.00001053 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared:  0.8282,        Adjusted R-squared:   0.82
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

Figura 2.2: Saída para o modelo linear *prestige* $\sim$ *education* + *income* ajustado para o banco de dados *Duncan*

Os comandos gerados pela caixa de diálogo Regressão linear usam a função lm (modelo linear) no R para ajustar o modelo, criando RegModel.1 e, em seguida, resume o modelo para produzir a saída impressa. A saída de resumo (ou *summary*) inclui informações sobre a distribuição dos resíduos, estimativas de coeficientes, seus erros-padrão, estatísticas t para testar a hipótese nula de que cada coeficiente de regressão é 0, os valores-p de dois lados para esses testes, o desvio padrão dos resíduos ("erro padrão residual"), graus residuais de liberdade; a correlação múltipla quadrada, R2, para o modelo e R2 ajustada para graus de liberdade e o teste omnibus para a hipótese de que todos os coeficientes da população (aqui os coeficientes de educação e renda) são 0 (H_0: $\beta_1 = \beta_2 = 0$, para o exemplo).

Esta é saída de regressão de mínimos quadrados padrão, semelhante à saída impressa produzida por quase todos os pacotes estatísticos. O que é incomum é que, além da impressão na Figura 2.2, o RCommander cria e retém um objeto de modelo linear no qual é possível realizar cálculos adicionais.

Assim como é possível escolher entre conjuntos de dados que residem na memória (se houver mais de um) pressionando o botão *Conjunto de Dados* na barra de ferramentas, também pode-se escolher entre modelos estatísticos (se houver mais de um) pressionando o botão *Modelo* no RCommander. É possível também escolher:

Modelos → Selecionar modelo ativo

Além disso, o RCommander se encarrega de coordenar conjuntos de dados e modelos, associando cada modelo estatístico ao conjunto de dados ao qual ele se encaixa. Consequentemente, a seleção de um modelo estatístico faz com que o conjunto de dados ao qual foi ajustado seja o conjunto de dados ativos, se ainda não for o caso.

As listas de variáveis no diálogo *Regressão Linear* na Figura 2.1 incluem apenas variáveis numéricas. Por exemplo, o tipo de fato (tipo de ocupação) no conjunto de dados de Duncan, com níveis "bc" (colarinho azul), "wc" (colarinho branco) e "prof" (profissional, técnico ou gerencial), não aparecem em qualquer lista de variáveis. Além disso, as variáveis explicativas selecionadas entram no modelo de forma linear e aditiva. A caixa de diálogo *Modelo Linear*, descrita posteriormente, é capaz de ajustar uma variedade muito maior de modelos de regressão.

Ao completar a caixa de diálogo Regressão linear na Figura 2.1, o nome do modelo é padrão, **RegModel.1**. O RCommander gera nomes de modelo exclusivos automaticamente durante uma sessão, cada vez incrementando o número do modelo (aqui 1).

A *Expressão* (ou subset expression) foi mantida como padrão ou "<todos os casos válidos>", como pode ser visto na Figura 2.1. Se, em vez disso, tivesse inserido o tipo == "bc", por exemplo, o modelo de regressão teria sido adequado apenas para ocupações de colarinho azul. Como neste exemplo, a expressão de subconjunto pode ser uma expressão lógica, retornando o valor VERDADEIRO ou FALSO para cada caso, ou seja, podemos informar um vetor de índices paraa incluir (VERDADEIRO), ou um vetor negativo de índices para excluir (FALSO). Por exemplo, digitar "1:25" na caixa de diálogo da *subset expression* incluiria as primeiras 25 ocupações, enquanto "-c (6, 16)" excluiria as ocupações 6 e 16 para gerar o modelo. Além disso, todos os diálogos de modelagem estatística no RCommander permitem que subconjuntos de casos sejam especificados desta maneira.

Capítulo 3

Modelos Lineares com Fatores

Como a discussão sobre Regressão Linear descrita na próxima seção, o Modelo Linear pode englobar Modelos de Regressão Linear, mas é muito mais flexível: O Modelo de Linear acomoda transformações das variáveis dependentes e independentes, fatores assim como variáveis de explicação numéricas no lado direito do modelo de regressão, funções não-lineares de variáveis dependentes expressas em polinômios e regressão spline, além de interações entre as variáveis de explicação. Tudo isso é alcançado permitindo ao usuário especificar o modelo desejado. Fórmulas de modelos lineares no R são derivadas da linguagem de programação S (Chambers and Hastie, 1992)[7], e são versões da notação para expressar modelos lineares originalmente introduzida por Wikinson and Rogers (1973)[12].

3.1 Fórmulas de Modelos Lineares

Uma fórmula de Modelo Linear R (R linear-model formula), possui como forma geral:

variável.explicada ∼ preditor.linear

O til () em uma fórmula de modelo linear pode ser lido como "é regredido em". Desse modo, nessa forma geral, a variável explicada é elucidada no preditor linear composto pelos termos do lado direito do modelo.

O lado esquerdo do modelo, variável-explicada, é uma expressão R que avalia a variável explicada no modelo, é usualmente simplificado pelo nome da variável explicada. É possível, por outro lado, transformar a variável explicativa diretamente na fórmula do modelo (e.g. log10(*income*)) ou computar a explicativa como uma expressão aritmética mais complexa (e.g. log(*income* + *education***income*)).

3.2 Simbologia do RCommander

A formulação do preditor linear no lado direito de uma fórmula de modelo é mais complexa. O que normalmente são operadores aritméticos (+, -, *, / e ∧) nas expressões do R têm significados especiais em uma fórmula de modelo, assim como os operadores (dois pontos) e %. O numeral 1 (um) pode ser usado para representar a constante de regressão (isto é, o intercepto) em uma fórmula de modelo; isso geralmente é desnecessário, no entanto, porque o intercepto é incluído por padrão. Um ponto (.) representa todas as variáveis no conjunto de dados, com exceção da resposta. Parênteses podem ser usados para agrupamento, assim como em uma expressão aritmética.

Na grande maioria dos casos, é possível formular um modelo usando apenas os operadores + (interpretado como "e") e * (interpretado como "cruzado com") e, portanto, é bom enfatizá-

los aqui. O significado desses e de outros operadores de fórmula de modelo são resumidos e ilustrados na tabela a seguir.

Operador	**Significado**	**Exemplo**	**Interpretação**
+	e	x1 + x2	x1 e x2
:	interação	x1:x2	interação de x1 e x2
*	multiplicar	x1*x2	x1 multiplicando com x2 (i.e., x1 + x2 +x1:x2)
-	remover	x1 - x2	regressão através da origem (para numérico x1)
ˆ k	cruzar para a ordem k	(x1 + x2 + x3)ˆ 2	igual a x1*x2 + x1*x3 + x2*x3
%in%	aninhar	cidade %in% país	cidade aninhada no país
/	aninhar	cidade/país	igual a país + cidade %in% país
Símbolo	**Significado**	**Exemplo**	**Interpretação**
1	interceptar	x1 - 1	suprime a interceptação
.	tudo menos a resposta	y ∼ .	regressa y em todo o resto
( )	agrupar	x1*(x2 + x3)	igual a x1*x2 + x1*x3

Uma sutileza da fórmula final: como explicado, os operadores aritméticos assumem significados especiais no lado direito de uma fórmula de modelo linear. Uma consequência é que não se pode usar esses operadores diretamente para aritmética. Por exemplo, o ajuste do modelo de *prestige* ∼ *type*education + type*income*, que pode ser visto na sessão 3.6.1 estima um coeficiente de regressão separado para cada um. Suponha, no entanto, que deseja-se estimar um único coeficiente para a soma dessas variáveis - na verdade, definindo os três coeficientes iguais uns aos outros. A solução é "proteger" o operador + dentro de uma chamada para a função I (identidade ou inibição), que simplesmente retorna seu argumento inalterado: *savings*∼ I (*wages* + *interest* + *dividends*). Esta fórmula funciona como desejado porque os operadores aritméticos como + têm seu significado usual dentro de uma chamada de função no lado direito da fórmula - implicando em, *savings*∼ log10 (*wages* + *interest* + *dividends*) também funciona conforme pretendido, estimando um único coeficiente para a base logarítmica 10 da soma de salários, juros e dividendos.

3.3 Princípio da Marginalidade

Introduzido por Nelder (1977)[9], o princípio da marginalidade é uma regra para a formulação e interpretação de modelos estatísticos lineares (e semelhantes). De acordo com o princípio da marginalidade, um choque em uma variável explicativa, capta o efeito médio de todas as demais variáveis explicativas. Dessa forma, digamos $x_1 : x_2$, está incluída em um modelo linear, então o mesmo deve acontecer com os principais efeitos, x_1 e x_2, que são marginais a da interação. Da mesma forma, as interações de ordem inferior $x_1 : x_2$, $x_1 : x_3$ e $x_2 : x_3$ são marginais à interação de três variáveis $x_1 : x_2 : x_3$. A constante de regressão (1 em uma fórmula de modelo R) é marginal para todos os outros termos do modelo.

Em resumo, o princípio da marginalidade propõe que, em geral, não é certo interpretar, ou nem mesmo estimar os efeitos principais de uma variável explicativa, enquanto se ignora os efeitos principais marginais a essa interação alvo.

É na maioria das circunstâncias difícil formular modelos que violem o princípio da marginalidade, e tentar fazer isso pode produzir resultados indesejados. Por exemplo, embora possa parecer que o modelo $y \sim f * x - x - 1$, em que f é um fator e x é uma variável explicativa numérica, viola o princípio da marginalidade removendo a constante de regressão e a inclinação. O modelo que realmente se ajusta inclui uma interceptação e inclinação separadas para cada nível do fator f. Assim, o modelo $y \sim f * x - x - 1$ é equivalente a $y \sim f * x$. Quase sempre é melhor evitar essas fórmulas de modelo incomuns.

3.4 Interpretação dos coeficientes

Dessa forma, ao realizarmos um modelagem no RCommander com variáveis de fatores obtemos como saída um valor associado aquele fator. Normalmente, o R trabalha com fatores de forma binária (ver Seção 3.5), portanto, teremos o valor 0 atribuído à um nível do fator, enquanto o valor 1 é atribuído ao outro nível. O caso mais comum é o de gênero, sendo o feminino comumente associado ao 0 e o masculino ao 1.

Portanto, quando temos o valor associado à essa variável categórica, vemos o impacto dela na variável explicada. Utilizando o exemplo citado acima, vamos criar um banco de dados no R com 20 nomes diferentes, apenas para servir como referência. Nela devemos incluir 10 nomes masculinos e 10 femininos (n = 20), com valores salariais arbitrários apenas para fins ilustrativos e os códigos 0 para sexo feminino e 1 para masculino. Dessa forma, com os comandos abaixo, obtemos o banco de dados "Salarios":

```
NOMES <- c(‘‘Mário’’, ‘‘Rodrigo’’, ‘‘Júlia’’, ‘‘Fernanda’’, ‘‘Roberto’’, ‘‘João’’,
           ‘‘Vitória’’, ‘‘Vitor’’, ‘‘Gabriel’’, ‘‘Rafaela’’, ‘‘Thayná’’, ‘‘Natália’’,
           ‘‘Gustavo’’, ‘‘Mayra’’,‘‘Pedro’’, ‘‘Rosa’’, ‘‘Vicente’’, ‘‘Cristina’’,
           ‘‘Artur’’, ‘‘Luiza’’)
GÊNERO<- c(1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0)
SALÁRIO <- c(5300, 6530, 5000, 2500, 6720, 5500, 3200, 6500, 6000, 5400,
             3500, 5500, 7000, 5500, 6545, 2500, 5500, 2500, 5000, 3000)

Salarios <- data.frame(NOMES, GÊNERO, SALÁRIO)
```

Figura 3.1: Código para gerar o banco de dados Salarios

Com esse banco de dados obtemos a Tabela 3.1 vista abaixo.

NOMES	GÊNERO	SALÁRIO
Mário	1	5300
Rodrigo	1	6530
Júlia	0	5000
Fernanda	0	2500
Roberto	1	6720
João	1	5500
Vitória	0	3200
Vitor	1	6500
Gabriel	1	6000
Rafaela	0	5400
Thayná	0	3500
Natália	0	5500
Gustavo	1	7000
Mayra	0	5500
Pedro	1	6545
Rosa	0	2500
Vicente	1	5500
Cristina	0	2500
Artur	1	5000
Luiza	0	3000

Tabela 3.1: Dados arbitrariamente escolhidos para o exemplo.

Agora, para que possamos ver o gráfico de dispersão da Figura 3.2, vamos até:

Gráficos → Diagrama de dispersão...

A partir daí selecionamos para a variável-x, "SALÁRIO", e para a variável-y, "GÊNERO", e finalizamos com "OK".

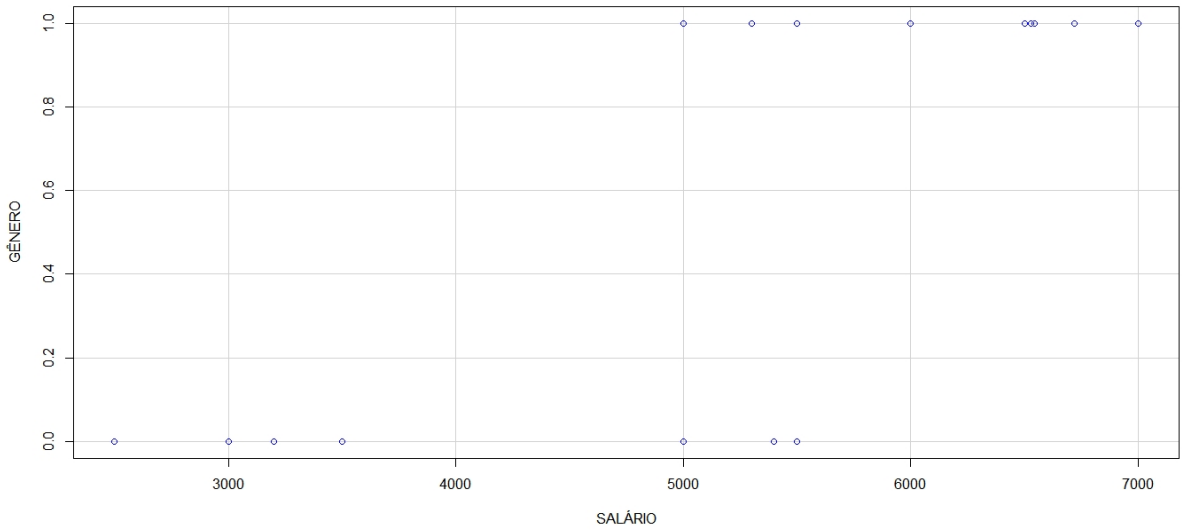

Figura 3.2: No eixo Y vemos os valores marcados em 0 e 1 relacionados à variável "GÊNERO". No eixo X, encontramos os valores da variável "SALÁRIO" em reais.

Segundo a dispersão dos nossos dados, o gênero feminino apresenta maior frequência de salários menores, quando comparado ao sexo masculino. Nesse gráfico de dispersão, os dados estão todos concentrados nos valores de 0 e 1, pois essa é a codificação que estabelecemos para a variável categórica "GÊNERO". Isso justifica o fato de que no eixo Y do gráfico, não existem dados no valor de 0.2, 0.4 ou mesmo 0.8, pois estes são valores impossíveis de serem observados para essa variável.

Se fizermos uma regressão $SALARIO \sim GENERO$, o valor de saída do R será a diferença entre as médias salariais desses dois grupos. Um valor positivo para o coeficiente de GÊNERO mostraria que os homens, na média, ganham mais do que as mulheres. Para fazer o modelo, apenas leia o comando `lm(formula = SALÁRIO  GÊNERO, data = Salarios)` no seu ambiente após carregados os dados no R.

```
> lm(formula = SALÁRIO ~ GÊNERO, data = Salarios)

Call:
lm(formula = SALÁRIO ~ GÊNERO, data = Salarios)

Coefficients:
(Intercept)       GÊNERO
       3860         2200
```

Figura 3.3: Saída para o modelo linear $SALÁRIO \sim GÊNERO$ ajustado para o banco de dados apresentado na Tabela 3.1

Essa foi uma breve explicação que se tornará mais palpável com próximos exemplos.

3.5 Variáveis Dummy

Antes de entrarmos de vez nos exemplos, para que possamos consolidar de fato todos esses conceitos, precisamos apresentar mais um: variáveis dummy ou variáveis fictícias.

Por padrão no RCommander, fatores em fórmulas de modelo linear são representadas por regressões de variáveis Dummy 0/1 geradas pela função *contr.Treatment* no pacote *car*, escolhendo o primeiro nível de um fator como o nível de linha de base. Esta codificação de contraste, junto com algumas outras opções, é mostrada na Tabela 3.2, usando o fator *type* no conjunto de dados *Prestige* como um exemplo. É preciso explicar que variável dummy é aquela que leva apenas o valor 0 ou 1 para indicar a ausência ou presença de algum efeito categórico que possa ser esperado para mudar o resultado. Eles podem ser considerados como stand-ins numéricos para fatos qualitativos em um modelo de regressão, classificando dados em categorias mutuamente exclusivas. Vejamos um exemplo, podemos dizer que uma pessoa é fumante (pertence a esse grupo) ou não:

$$\textit{Variável Dummy}_f = \left\{ \begin{array}{l} 1, se \;\; \text{'f'} \; \textit{for fumante} \\ 0, se \;\; \text{'f'} \; \textbf{não} \; \textit{for fumante} \end{array} \right| \tag{3.1}$$

Dessa forma, se a variável assume o valor de 0, isso significa que o indivíduo em questão não fuma e portanto, não pode pertencer ao grupo de fumantes (categorias mutuamente exclusivas).

3.6 Exemplos no RCommander

3.6.1 Exemplos usando o Canadian Occupational Prestige Data

Para concretizar, formularemos alguns modelos lineares para o Canadian Occupational Prestige Data, regredindo prestige em income , education, women (composição de gênero), e type (tipo de ocupação). A última variável é um fator (variável categórica) e, portanto, não pode entrar no modelo linear diretamente. Quando um fator é incluído em uma fórmula de modelo linear, R gera contrastes (uma combinação linear de variáveis cujos coeficientes somam zero, permitindo a comparação de diferentes tratamentos) para representar o fator - um a menos que o número de níveis do fator. Como explicado anteriormente, o R usa regressores de variável fictícia 0/1, também chamados de variáveis indicadoras, para interpretar fatores.

Uma versão do Canadian Occupational Prestige Data reside no banco de dados Prestige no pacote **car**, e é conveniente ler os dados no RCommander a partir desta fonte, segue os passos:

Dados → Dados em pacotes → Ler dados de um pacote anexado.

Prestige se tornará o conjunto de dados ativo.É importante ressaltar que 4 das 102 ocupações no conjunto Prestige têm valores ausentes (NA) para o tipo ocupacional. Para ajustar vários modelos de regressão ao banco de dados Prestige, nem todos incluindo o tipo, é interessante começar filtrando o conjunto de dados para valores ausentes, selecionando:

Dados → Conjunto de dados ativo → Remover observações com dados faltantes.

Além disso, a ordem alfabética padrão dos níveis do tipo - "bc", "prof", "wc" - não é a ordem natural e, portanto, é possível reordenar os níveis desse fator através da sequêncoa:

Dados → Variáveis de gerenciamento no conjunto de dados ativo →

Reordene os níveis do fator para "bc" , "wc", "prof".

Esta última etapa não é estritamente necessária, mas torna a análise de dados mais fácil de seguir.

Primeiro, é bom ajustar uma regressão dummy aditiva aos dados de prestígio canadense, empregando a fórmula do modelo de $prestige \sim income + education + women + type$. Para fazer isso, selecione Estatísticas → Ajuste de Modelos → Modelo linear nos menus do RCommander, produzindo a caixa de diálogo na Figura 3.4. O nome do modelo fornecido automaticamente é LinearModel.3, refletindo o fato de que já ajustamos um modelo estatístico na sessão, RegModel.2.

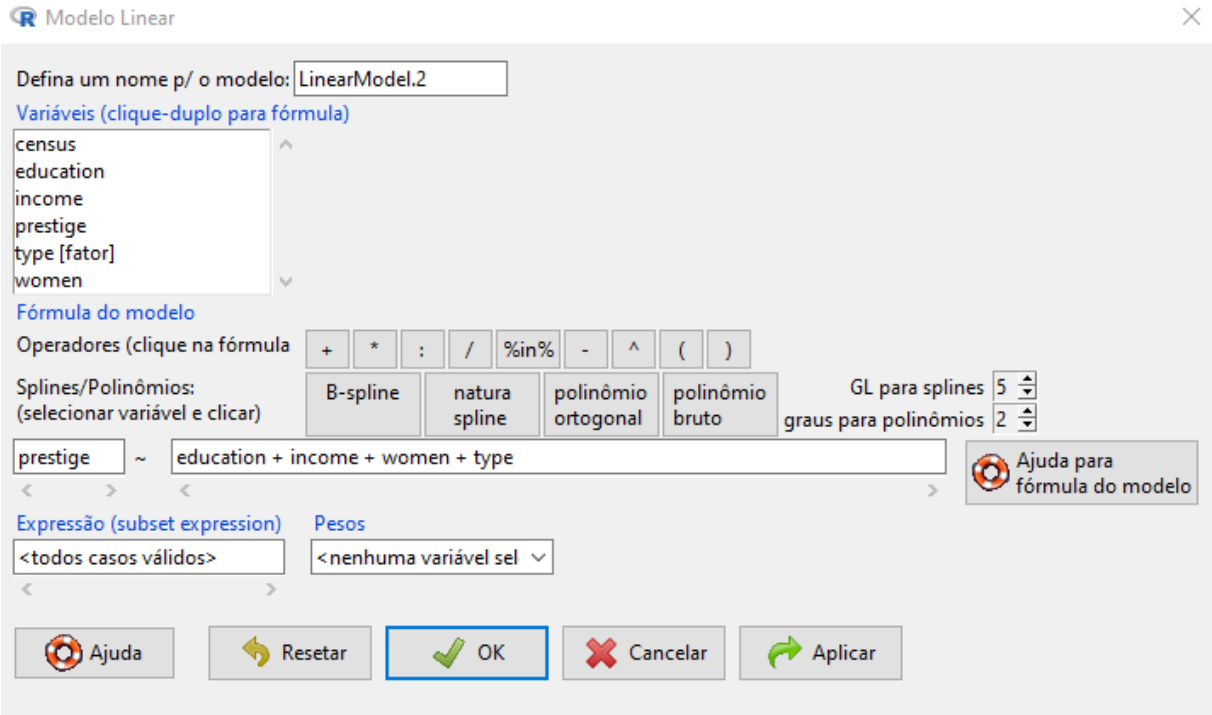

Figura 3.4: Ajuste do modelo linear $prestige \sim education + income + woman + type$ ajustada para o banco de dados *Prestige*

A maior parte da estrutura do diálogo Modelo Linear é comum aos diálogos de modelagem estatística no RCommander. Se a caixa de texto de resposta à esquerda da fórmula do modelo estiver vazia, clicar duas vezes no nome de uma variável na caixa de listagem de variáveis inserirá o nome na caixa de resposta; depois disso, clicar duas vezes nos nomes das variáveis insere os nomes no lado direito da fórmula do modelo, separados por +. Pode-se inserir parênteses e operadores como + e * na fórmula usando a barra de ferramentas na caixa de diálogo. Também é possível digitar diretamente nas caixas de texto da fórmula do modelo. Na Figura 3.4, simplesmente clique duas vezes sucessivamente em prestige, education, income, women e type. Clicar em OK produz a saída mostrada na Figura 3.5.

```
Rcmdr>  LinearModel.2 <- lm(prestige ~ education + income + women + type,
Rcmdr+    data=Prestige)

Rcmdr>  summary(LinearModel.2)

Call:
lm(formula = prestige ~ education + income + women + type, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7485  -4.4817   0.3119   5.2478  18.4978

Coefficients:
               Estimate Std. Error t value    Pr(>|t|)
(Intercept)  -0.8139032  5.3311558  -0.153    0.878994
education     3.6623557  0.6458300   5.671 0.000000163 ***
income        0.0010428  0.0002623   3.976    0.000139 ***
women         0.0064434  0.0303781   0.212    0.832494
type[T.prof]  5.9051970  3.9377001   1.500    0.137127
type[T.wc]   -2.9170720  2.6653961  -1.094    0.276626
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.132 on 92 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.8349,        Adjusted R-squared:  0.826
F-statistic: 93.07 on 5 and 92 DF,  p-value: < 2.2e-16
```

Figura 3.5: Saída para o modelo linear *prestige* $\sim$ *education + income + woman + type* ajustado para o banco de dados *Prestige data*

Já foi explicado o formato geral da saída de resumo do modelo linear em R. O que há de novo na Figura 3.5 é a maneira como o tipo de fator é tratado no modelo linear: regressores de variável dupla são criados automaticamente para o tipo de fator de três níveis. O primeiro regressor dummy, labelled type [T.wc] na saída, é codificado 1 quando tipo é "wc" e 0 caso contrário; o segundo regressor dummy, tipo [T.prof], é codificado 1 quando tipo é "prof" e 0 caso contrário. O primeiro nível do tipo - "bc" - é, portanto, selecionado como o nível de referência ou da linha de base, codificado 0 para ambos os regressores dummy.

Consequentemente, o intercepto na saída do modelo linear é o intercepto para o nível de referência "bc" do tipo e os coeficientes para os outros níveis de diferenças nos interceptos entre cada um desses níveis e o nível de referência. Como os coeficientes de inclinação para as variáveis explicativas numéricas *education*, *income* e *women* neste modelo aditivo não variam por níveis de tipo, os coeficientes das variáveis dummy também são interpretáveis como a diferença média entre cada um dos outros níveis e "bc" para quaisquer valores fixos, como explicado na Seção B.

Para ilustrar um modelo estruturalmente mais complexo, não aditivo, especifique o modelo Canadian Occupational Prestige Data para incluir interações entre *type* e *education* e entre *type* e *income*, no processo de remoção da variável *women* do modelo - na regressão inicial, o coeficiente de mulheres é pequeno com um grande valor p. O Modelo Linear reabre em seu estado anterior, com o nome do modelo incrementado para LinearModel.3. Para ajustar o novo modelo, deve se modificar a fórmula para *prestige*$\sim$*type* $*$ *education* $+$ *type* $*$ *income*. Clicar em OK produz a saída na Figura 3.6.

Com as interações no modelo, existem diferentes interceptos e inclinações para cada nível de

tipo. O intercepto na saída - junto com os coeficientes para education e income - pertence ao nível de linha de base "bc" do tipo. Outros coeficientes representam diferenças entre cada um dos outros níveis e o nível da linha de base. Por exemplo, tipo [T.wc] = - 33,54 é a diferença nos interceptos entre os níveis "wc" e "bc" do tipo; da mesma forma, o tipo de coeficiente de interação [T.wc]: educação = 4,291 é a diferença nas inclinações de educação entre "wc" e níveis "bc". A complexidade dos coeficientes pode torna um pouco difícil de entender o que o modelo diz sobre os dados, por isso é importante avaliar com atenção.

```
Rcmdr>  LinearModel.3 <- lm(prestige ~ type*education +type*income, data=Prestige)

Rcmdr>  summary(LinearModel.3)

Call:
lm(formula = prestige ~ type * education + type * income, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-13.462  -4.225   1.346   3.826  19.631

Coefficients:
                          Estimate  Std. Error t value      Pr(>|t|)
(Intercept)              2.2757530   7.0565809   0.323        0.7478
type[T.prof]            15.3518963  13.7152577   1.119        0.2660
type[T.wc]             -33.5366519  17.6536726  -1.900        0.0607 .
education                1.7132747   0.9572405   1.790        0.0769 .
income                   0.0035224   0.0005563   6.332 0.00000000962 ***
type[T.prof]:education   1.3878090   1.2886282   1.077        0.2844
type[T.wc]:education     4.2908748   1.7572512   2.442        0.0166 *
type[T.prof]:income     -0.0029026   0.0005989  -4.847 0.00000528253 ***
type[T.wc]:income       -0.0020719   0.0008940  -2.318        0.0228 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.318 on 89 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.8747,        Adjusted R-squared:  0.8634
F-statistic: 77.64 on 8 and 89 DF,  p-value: < 2.2e-16
```

Figura 3.6: Saída para o modelo linear *prestige* $\sim$ *type * income + type * education* ajustada para o banco da dados*Prestige*

Função	Nomes de contraste	Níveis de type "bc"	"wc"	"prof"
contr.Treatment	type[T.wc]	0	1	0
	type[T.prof]	0	0	1
contr.Sum	type[S.wc]	1	0	1
	type[S.prof]	0	1	-1
contr.poly	type.L	$-1/\sqrt{2}$	0	$1/\sqrt{2}$
	type.Q	$1/\sqrt{6}$	$-2/\sqrt{6}$	$1/\sqrt{6}$
contr.Helmert	type[H.1]	-1	1	0
	type[H.2]	-1	-1	2

Tabela 3.2: Contrastes para fatores do banco de dados *Prestige*

3.6.2 Exemplos usando o Banco de Dados *stress*

Um outro banco de dados no qual é possível trazer exemplos é o *stress* do pacote **datarium**, regredindo *score* (medida de escore de estresse), *treatment*, *exercise* e *age*. As variáveis *treatment* e *exercise* são categóricas. Os níveis de *treatment* são estar em tratamento psicológico (*yes*) e não estar em tratamento psicológico (*no*), sendo que *yes* é a base do fator. Os níveis de *exercise* são baixa, moderada e alta quantidade de prática de exercícios (*low*, *moderate* e *high*, respectivamente) e estão construídos de forma natural, ou seja, em ordem crescente.

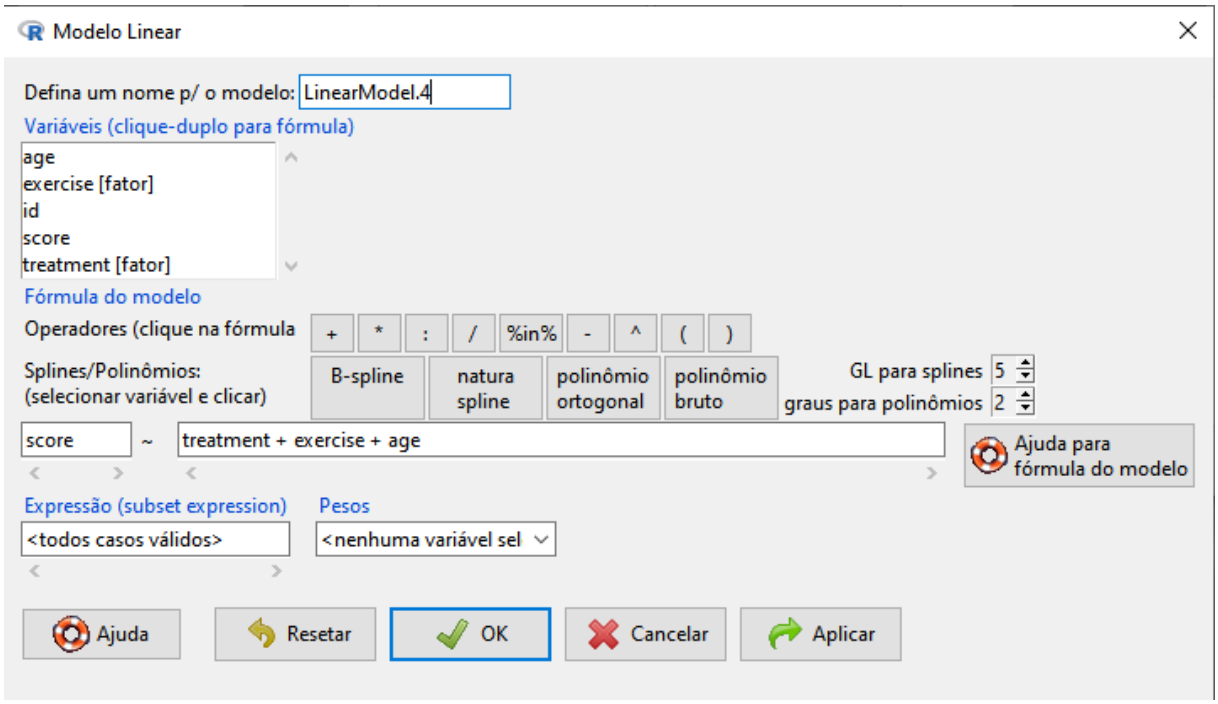

Figura 3.7: Ajuste do modelo linear *score* $\sim$ *treatment* + *exercise* + *age* ajustada para o banco de dados *stress*

De forma bem semelhante ao exemplo anterior. Primeiro, ajuste uma regressão dummy aditiva aos dados de stress, empregando a fórmula do modelo de $score \sim treatment + exercise + age$. Para fazer isso, selecionamos:

Estatísticas → Ajuste de Modelos → Modelo linear

Isso nos menus do RCommander, produzindo a caixa de diálogo na Figura 3.7. O nome

do modelo fornecido automaticamente é LinearModel.4, refletindo o fato de que já ajustamos outros modelos nas sessões anteriores. Clicar em OK produz a saída na Figura 3.8.

Novamente há fatores na saída do modelo gerado no RCommander e sua interpretação ocorre conforme foi explicado no exemplo anterior. Para *tretament* os regressores de variável dupla são criados automaticamente para o tipo de fator de dois níveis. O regressor dummy, *labelled type* [T.no] na saída, é codificado 1 quando tipo é *no* e 0 caso contrário. O nível do tipo - *yes* - é, portanto, selecionado como o nível de referência ou da linha de base, codificado 0 para o regressor dummy. Já para *exercise*, os regressores de variável dupla são criados automaticamente para o tipo de fator de três níveis. O primeiro regressor dummy, *labelled type* [T.moderate] na saída, é codificado 1 quando tipo é *moderate* e 0 caso contrário; o segundo regressor dummy, tipo [T.high], é codificado 1 quando tipo é *high* e 0 caso contrário. O primeiro nível do tipo - *low* - é, portanto, selecionado como o nível de referência ou da linha de base, codificado 0 para ambos os regressores dummy.

```
Rcmdr>  LinearModel.4 <- lm(score ~ treatment +exercise + age, data=stress)

Rcmdr>  summary(LinearModel.4)

Call:
lm(formula = score ~ treatment + exercise + age, data = stress)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0261 -3.7497 -0.4285  3.0943 13.3696

Coefficients:
                     Estimate Std. Error t value   Pr(>|t|)
(Intercept)          55.72934   10.91888   5.104 0.00000427 ***
treatment[T.no]       4.32529    1.37744   3.140    0.00272 **
exercise[T.moderate]  0.08735    1.69032   0.052    0.95897
exercise[T.high]     -9.61841    1.84741  -5.206 0.00000296 ***
age                   0.49811    0.17648   2.822    0.00662 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.288 on 55 degrees of freedom
Multiple R-squared:  0.6045,        Adjusted R-squared:  0.5757
F-statistic: 21.01 on 4 and 55 DF,  p-value: 1.473e-10
```

Figura 3.8: Saída para o modelo linear *score* $\sim$ *treatment* + *exercise* + *age* ajustado para o banco de dados *stress*

Para ilustrar um modelo estruturalmente um pouco mais complexo, especifique o modelo stress para incluir interações entre *age* e *treatment* e entre *age* e *exercise*. O Modelo Linear reabre em seu estado anterior, com o nome do modelo incrementado para LinearModel.5. Para ajustar o novo modelo, deve se modificar a fórmula para $stress \sim treatment * age + exercise * age$ ou para $stress \sim (treatment + exercise) * age$. Clicar em OK produz a saída na Figura 3.9.

Com as interações no modelo, existem diferentes interceptos e inclinações para cada nível de tipo. O intercepto na saída - junto com o coeficiente para *age* - pertence aos níveis de linha de base *low* de *exercise* e *yes* de *treatment*. Outros coeficientes representam diferenças entre cada um dos outros níveis e o nível da linha de base. Por exemplo, *treatment* [T.no] = 32,04 é a diferença nos interceptos entre os níveis *no* e *yes* do tipo; da mesma forma, o tipo de coeficiente

de interação *treatment*[T.no]:*age* = -0,47 é a diferença nas inclinações de *age* entre *no* e *yes*.

```
Rcmdr>  LinearModel.5 <- lm(score ~ (treatment + exercise)*age, data=stress)

Rcmdr>  summary(LinearModel.5)

Call:
lm(formula = score ~ (treatment + exercise) * age, data = stress)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6537 -4.0828 -0.1275  3.2049 12.5670

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               48.8452    19.9401   2.450   0.0177 *
treatment[T.no]           32.0424    21.2464   1.508   0.1376
exercise[T.moderate]     -21.1787    25.0691  -0.845   0.4021
exercise[T.high]         -22.0869    30.6118  -0.722   0.4738
age                        0.6187     0.3251   1.903   0.0626 .
treatment[T.no]:age       -0.4650     0.3551  -1.310   0.1961
exercise[T.moderate]:age   0.3516     0.4096   0.858   0.3946
exercise[T.high]:age       0.2094     0.5202   0.402   0.6890
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.34 on 52 degrees of freedom
Multiple R-squared:  0.6187,        Adjusted R-squared:  0.5673
F-statistic: 12.05 on 7 and 52 DF,  p-value: 0.000000005052
```

Figura 3.9: Saída para o modelo linear *score* ∼ *(treatment + exercise) * age* ajustado para o banco de dados *stress*

3.6.3 Exemplos usando o Banco de Dados hsb2

Usando o comando:

```
hsb2 <- read.csv(‘‘https://stats.idre.ucla.edu/stat/data/hsb2.csv’’)
```

Acessa-se a base de dados Hsb2. Este banco de dados contem 200 observações de uma amostra aleatória de alunos do ensino médio e mediu 11 variáveis que descrevem o perfil do aluno e que medem escores de habilidades de diferentes competências. Neste exemplo usaremos apenas as variáveis *race*, *schtyp*, *ses* e *write*, que para cada indivíduo informa sobre a raça (1 se afro-americano, 2 se asiático, 3 se hispânico e 4 se branco), o tipo da escola (1 se pública e 2 se privada), a renda familiar (1 se baixa, 2 se média e 3 se alta) e o escore em escrita no teste aplicado, respectivamente.

Como as variáveis categóricas *female*, *race*, *ses*, *schtyp* e *prog* estão em tipo numérico, é necessário usar o comando as.factor() para converte-las para o tipo correto, conforme feito abaixo.

```
#corrigindo tipo das variáveis
hsb2$female = as.factor(hsb2$female)
hsb2$race = as.factor(hsb2$race)
hsb2$ses = as.factor(hsb2$ses)
hsb2$schtyp = as.factor(hsb2$schtyp)
hsb2$prog = as.factor(hsb2$prog)
```

Figura 3.10: Conversão das variáveis categóricas do banco de dados *hsb2* para o seu tipo correto (fator)

Primeiro pretende-se fazer um modelo simples em que apenas com o valor indicado para a raça tentamos predizer a nota dos alunos. Desta forma selecionamos a opção de Modelo Linear no RCommander, *write* como a variável resposta do modelo e *race* como a variável explicativa, para chegarmos na mesma tela que a mostrada na Figura 3.11.

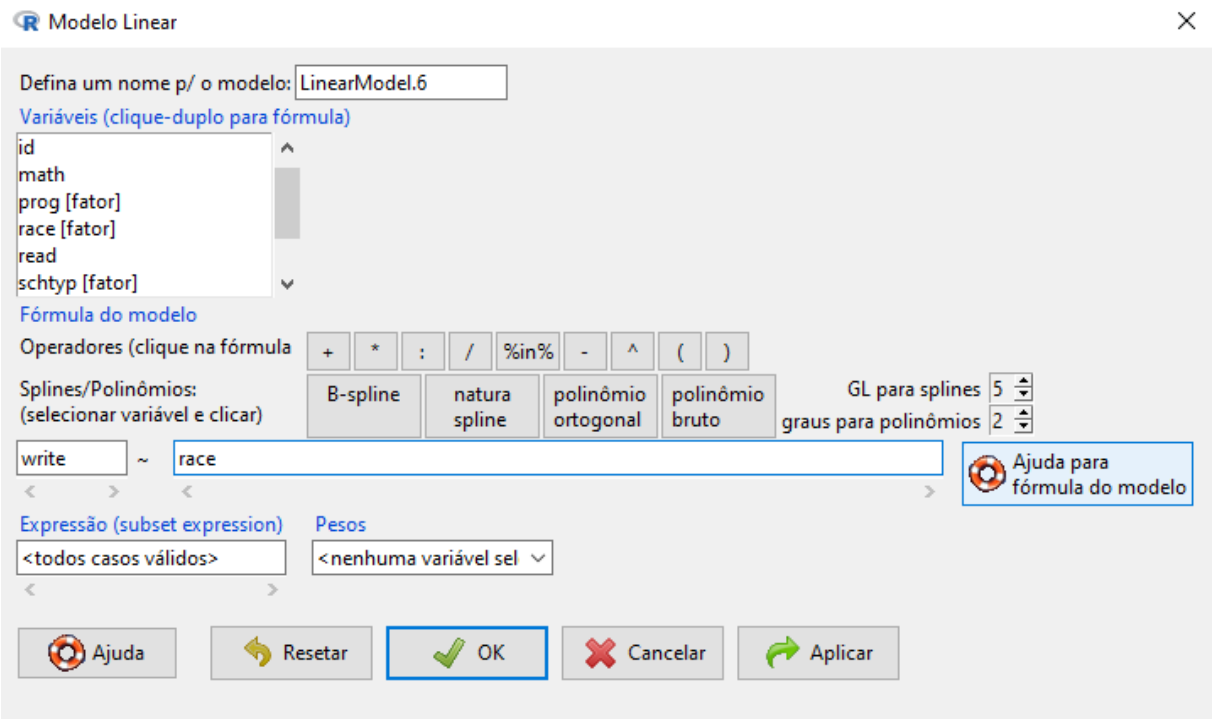

Figura 3.11: Ajuste do modelo linear *write* $\sim$ *race* ajustada para o banco de dados *hsb2*

Com este modelo podemos avaliar melhor as diferenças dos valores do escore por raça. como a variável base (o intercepto) é da raça afro-americana, temos que todos os outros coeficiente serão adições ao valor desta base. Ou seja, já exemplificando, percebemos na figura 3.12 que alunos asiaticos têm em média 11.542 pontos a mais no escore do que alunos negros. Além disso, hispânicos e brancos tiveram um aumento médio de 1.742 e 7.597, respectivamente. Chegamos desta forma à conclusão há diferenças, portanto desigualdes, entre os desempenhos dos alunos, já que os alunos negros da escola têm em média escores menores para escrita.

Lembrete: os motivos da diferenças devem ser explicados por uma pesquisa/análise maior com, por exemplo, indicadores socio-econômicos e de desigualdade. Não interprete estes valores de forma causal, ou seja, raça não implica em escores diferentes.

```
Rcmdr>  LinearModel.6 <- lm(write ~ race, data=hsb2)

Rcmdr>  summary(LinearModel.6)

Call:
lm(formula = write ~ race, data = hsb2)

Residuals:
     Min       1Q   Median       3Q      Max
-23.0552  -5.4583   0.9724   7.0000  18.8000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.458      1.842  25.218  < 2e-16 ***
race[T.2]     11.542      3.286   3.512 0.000552 ***
race[T.3]      1.742      2.732   0.637 0.524613
race[T.4]      7.597      1.989   3.820 0.000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.025 on 196 degrees of freedom
Multiple R-squared:  0.1071,        Adjusted R-squared:  0.0934
F-statistic: 7.833 on 3 and 196 DF,  p-value: 0.00005785
```

Figura 3.12: Saída para o modelo linear *write* $\sim$ *race* ajustado para o banco de dados *hsb2*

O modelo anterior nos trouxe informações interessantes sobre a diferença dos escores dos alunos. mas ele pode ser considerado muito simples e com um $R2$ de 0.0934, que é considerado baixo. A raça certamente não é a única variável para explica o escore em escrita. Desta forma selecionaremos agora a opção de Modelo Linear no RCommander e ainda com *write* como a variável resposta do modelo, mas selecionaremos além de *race* para a variável explicativa as variáveis *ses* e *schtyp*. Assim chegaremos na mesma tela que a mostrada abaixo:

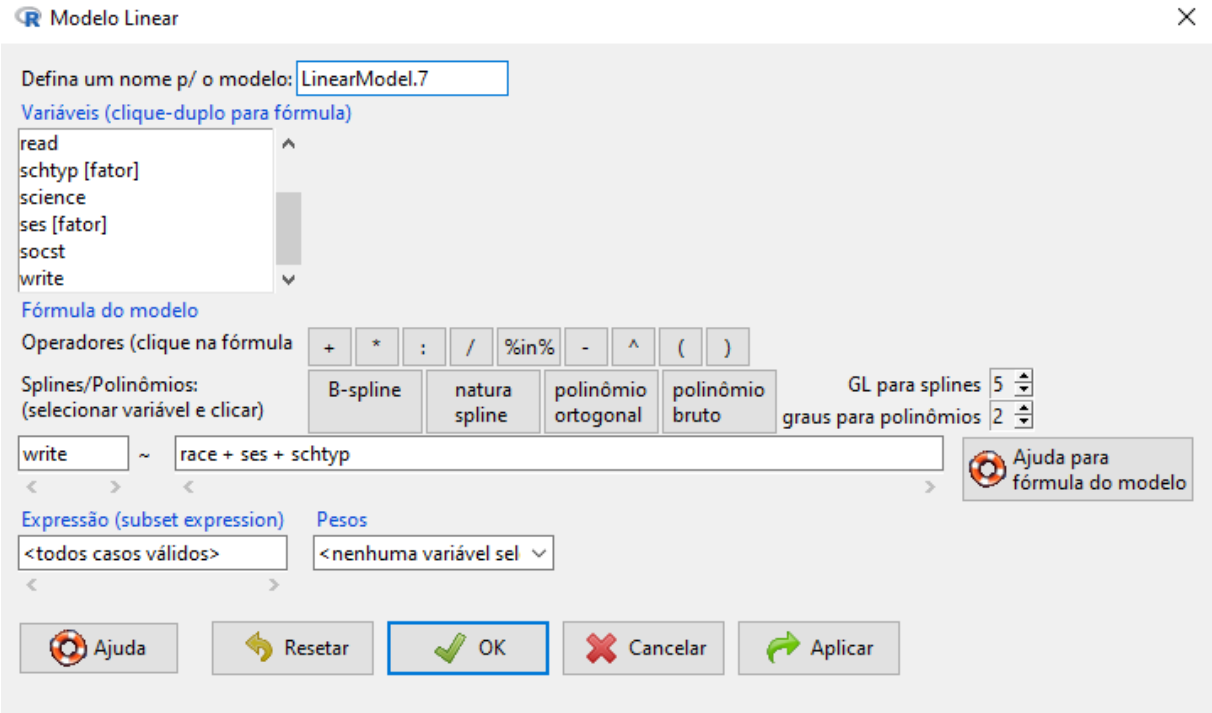

Figura 3.13: Ajuste do modelo linear *write* $\sim$ *race* + *ses* + *schtyp* ajustada para o banco de dados *hsb2*

Agora no novo modelo (visto na imagem 3.14) temos com base (intercepto) as categorias de raça afro-americana, de renda baixa e de escola pública. Para achar o valor estimado de um aluno qualquer, deve-se adicionar o coeficiente da categorias que diferem do conjunto base. Por exemplo, o calculo da estimação do valor do escore de um aluno asiático de renda familiar média e de escola pública deve ser a soma do intercepto com race[T.2] e ses[T.2], já que estes são os coeficientes para as categorias deste aluno que diferem das categorias base. Assim chegamos na estimação da nota de para este aluno hipotético.

Além de podermos estimar valores para novas observações, podemos visualizar as diferenças médias entre as categorias de cada variável. Isso quer dizer que há uma diferença média de -0.4419 e 3.1976 para as notas de alunos com renda familiar média e alta, respectivamente, com relação a alunos de renda baixa e que as notas de alunos de escola privada é em média 2.5652 maior que os de escola pública. Assim, concluímos que ser de renda alta e de escola privada tende a aumentar a nota de escrita.

Para finalizar, nem para os coeficientes sobre raça nem para o valor de $R2$ na saída não diferem muito do valor do modelo anterior. Isso indica, respectivamente, que as variáveis adicionadas não sobre a relação entre *write* e *race* nem adicionam muito o sobre a predição da variável *write*.

```
Rcmdr>  LinearModel.7 <- lm(write ~ race + ses + schtyp, data=hsb2)

Rcmdr>  summary(LinearModel.7)

Call:
lm(formula = write ~ race + ses + schtyp, data = hsb2)

Residuals:
   Min     1Q Median     3Q    Max
-22.94  -6.30   1.70   6.70  19.09

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.9142     2.0615  22.272  < 2e-16 ***
race[T.2]    11.1814     3.2452   3.446 0.000699 ***
race[T.3]     1.6822     2.7094   0.621 0.535406
race[T.4]     6.8273     1.9962   3.420 0.000764 ***
ses[T.2]     -0.4419     1.6646  -0.265 0.790948
ses[T.3]      3.1976     1.8336   1.744 0.082779 .
schtyp[T.2]   2.5652     1.7495   1.466 0.144212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.897 on 193 degrees of freedom
Multiple R-squared:  0.1454,        Adjusted R-squared:  0.1189
F-statistic: 5.474 on 6 and 193 DF,  p-value: 0.0000297
```

Figura 3.14: Saída para o modelo linear *write* $\sim$ *race + ses + schtyp* ajustado para o banco de dados *hsb2*

3.7 Outros Contrastes para Fatores

Assim como as variáveis dummy, existem outros tipos de contrastes para fatores, sendo alguns deles apresentados anteriormente na Tabela 3.2.

A função *contr.Sum* do *car* gera os chamados regressores de contraste "restritos por sigma" ou "soma a zero", como são usadas nos tratamentos tradicionais de análise de variância. A função padrão *contr.poly* do R contrastes ortogonais-polinomiais - um grupo de coeficientes de contraste ortogonal que digitam ou codificam os padrões lineares, quadráticos e de ordem superior nas informações. Neste caso, termos lineares e quadráticos para os três níveis de *type*; no RCommander, *contr.poly* é a escolha padrão para fatores ordenados. Finalmente, *contr.Helmert* gera contrastes Helmert, que comparam cada nível com a média daqueles que o precedem. Selecionando:

Dados → Modificação de variáveis no conjunto de dados.. → Definir contrastes p/ um fator

A caixa de diálogo à esquerda da Figura 3.7 é produzida. O fator *type* está pré-selecionado nesta caixa de diálogo porque é o único fator no conjunto de dados. Você pode usar os botões de opção para escolher entre *treatment*, *sum-to-zero*, *Helmert* e *polynomial contrasts*, ou definir contrastes personalizados selecionando *Other*, como mostrado.

Clicar em *OK* leva ao sub-diálogo mostrado à direita da Figura 3.7. Mude os nomes dos contrastes padrão, .1 e .2, para *[bc.v.others]* e *[wc.v.prof]*, e então preencha os coeficientes dos contrastes (i.e., os valores dos regressores dos contrastes). Essa escolha produz regressores

de contraste chamados *type [bc.v.others]* e *type [wc.v.prof]*, para serem usados quando o fator *type* no conjunto de dados *Prestige* aparecer em uma fórmula de modelo linear. Os contrastes definidos diretamente desta maneira devem ser linearmente independentes e são mais simples de interpretar se obedecerem a duas regras adicionais: (1) Os coeficientes para cada contraste deve somar 0, e (2) cada par de contrastes deve ser ortogonal (i.e., o os produtos dos coeficientes correspondentes para cada par de contrastes somam 0).

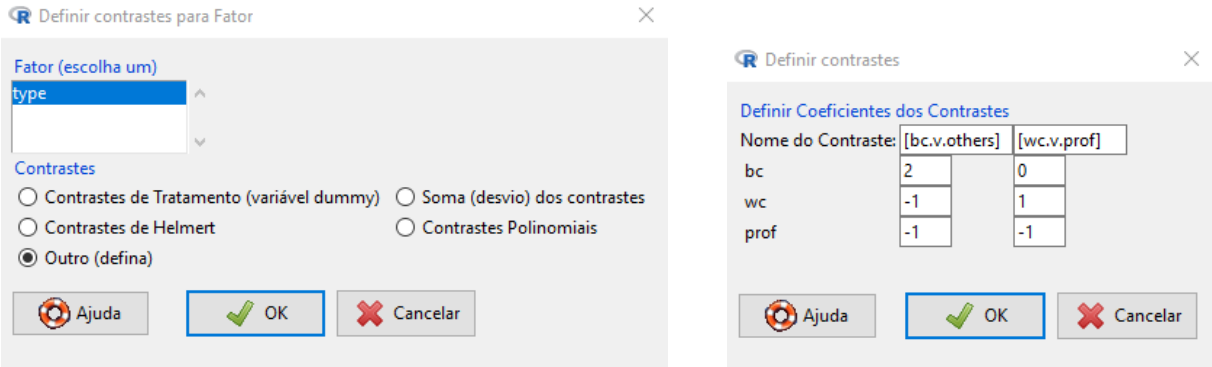

Figura 3.15: A caixa de diálogo para definir Contrastes para fator (à esquerda) e a sub-caixa de diálogo para especificar Contrastes (à direita), criando contrastes para *type* no banco de dados*Prestige*os.

Para ver como esses contrastes são refletidos nos coeficientes do modelo, reajuste a regressão aditiva de *prestige* em *education*, *income*, *women* e *type*, produzindo o resultado da Figura 3.16. O primeiro contraste para *type* estima a diferença entre "*bc*" e a média dos outros dois níveis de *type*, mantendo as outras variáveis explicativas constantes, enquanto o segundo contraste estima a diferença entre "*wc*" e "*prof*". Esta codificação de contraste alternativa para *type* produz estimativas diferentes para a interceptação e coeficientes *type* da codificação do regressor Dummy, mas os dois modelos têm o mesmo ajuste para os dados (por exemplo, R2 = 0,8349)[1].

[1]Para entender como é feito o ajuste de "Splines de Regressão" e de "Polinômios", procure o apêndice A

```
Rcmdr>  LinearModel.8 <- lm(prestige ~ education + income + women + type,
Rcmdr+    data=Prestige)

Rcmdr>  summary(LinearModel.8)

Call:
lm(formula = prestige ~ education + income + women + type, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7485  -4.4817   0.3119   5.2478  18.4978

Coefficients:
                    Estimate Std. Error t value    Pr(>|t|)
(Intercept)        0.1821385  7.0466879   0.026    0.979435
education          3.6623557  0.6458300   5.671 0.000000163 ***
income             0.0010428  0.0002623   3.976    0.000139 ***
women              0.0064434  0.0303781   0.212    0.832494
type[bc.v.others] -0.4980208  1.0194568  -0.489    0.626347
type[wc.v.prof]    4.4111345  1.3968819   3.158    0.002150 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.132 on 92 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.8349,        Adjusted R-squared:  0.826
F-statistic: 93.07 on 5 and 92 DF,  p-value: < 2.2e-16
```

Figura 3.16: Saída para o modelo linear *prestige* $\sim$ *education* + *income* + *women* + *type* ajustado para o banco de dados *Prestige*, usando constrastes customizados para o *type*

Capítulo 4

Modelo Linear Generalizado

Esse conjunto de modelos é uma espécie de prolongamento dos modelos de regressão linear que já abordamos anteriormente, o simples e múltiplo. Os Modelos Lineares Generalizados (ou MLGs) no entanto, permitem uma maior gama de distribuições, que resulta na capacidade de investigar não somente interações lineares entre variáveis!

Resumidamente, os MLGs, introduzidos em um artigo seminal de Nelder e Wedderburn (1972)[10], consistem em três componentes:

1. Componentes aleatórios que especificam a distribuição das variáveis explicativas de resposta. Tradicionalmente, o componente aleatório é membro de uma família exponencial - as famílias gaussiana (normal), binomial, Poisson, gama ou gaussiana inversa - mas tanto a teoria dos modelos lineares generalizados quanto sua implementação em Raras agora mais gerais: além das famílias exponenciais tradicionais, R fornece para famílias quase binomiais e quase Poisson que acomodam binomiais "superdispersos" e dados de contagem.

2. Preditor linear
$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$
em que a expectativa da variável de resposta $\mu_i = E(y_i)$ para as observações independentes, onde os regressores x_{ji} são funções pré-especificadas das variáveis explicativas - variáveis explicativas numéricas, regressores dummy representando fatores, interação de regressores, e assim por diante, exatamente como no modelo linear.

3. Uma função de vínculo invertível pré-especificada que transforma a expectativa da resposta ao preditor linear, $g(\mu_i) = \eta_i$, e assim $\mu_i = g - 1(\eta_i)$. O R implementa identidade, inverso, log, logit, probit log-log complementar, raiz quadrada e links inversos quadrados, com os links aplicáveis variando por família de distribuição

O MLG mais comum além do modelo linear normal (ou seja, a família Gaussiana emparelhada com vínculo de identidade) é o modelo logit binomial, adequado para variáveis de resposta dicotômicas (duas categorias). Para ilustração, usaremos os dados coletados por Cowles e Davis (1987) sobre o voluntariado para um experimento psicológico, em que os sujeitos do estudo eram alunos em uma aula de introdução à psicologia na universidade[1].

Os dados para este exemplo estão contidos no conjunto de dados Cowles no pacote de **car**, que inclui as seguintes variáveis: *neuroticism*, uma dimensão da personalidade com escores inteiros variando de 0 a 24; *extraversion*, outra dimensão da personalidade, também com escores de 0 a 24; *sex*, um fator com níveis "feminino" e "masculino"; e *volunteer*, um fator com níveis "não" e "sim".

Ao analisar os dados, Cowles e Davis realizaram uma regressão logística do voluntariado no sexo e a interação linear por linear entre neurotiscimo e extroversão. Para ajustar o modelo de

[1]Para aprender sobre a visualização de modelos lineares e lineares generalizados, acesse o capítulo 5

Cowles e Davis, primeiro é necessário ler pacote de **car** da maneira usual para poder acessar o banco de dados *Cowles*, tornando este o conjunto de dados ativo no Commander R. Então , no RCommander, selecionaremos:

Estatísticas → Ajuste de Modelos → Modelo Linear Generalizado (GLM)

Isso produz a caixa de diálogo abaixo. A caixa de diálogo do Modelo Linear Generalizado é muito semelhante ao diálogo Modelo Linear da seção anterior: O nome do modelo no topo (GLM.9) é gerado automaticamente, e você pode alterá-lo se desejar. Clicar duas vezes em uma variável na caixa de listagem insere a variável na fórmula do modelo. Existem barras de ferramentas para inserir operadores, linhas de regressão e polinômios na fórmula do modelo, e existem caixas para definir o subconjunto do conjunto de dados e especificar pesos anteriores.

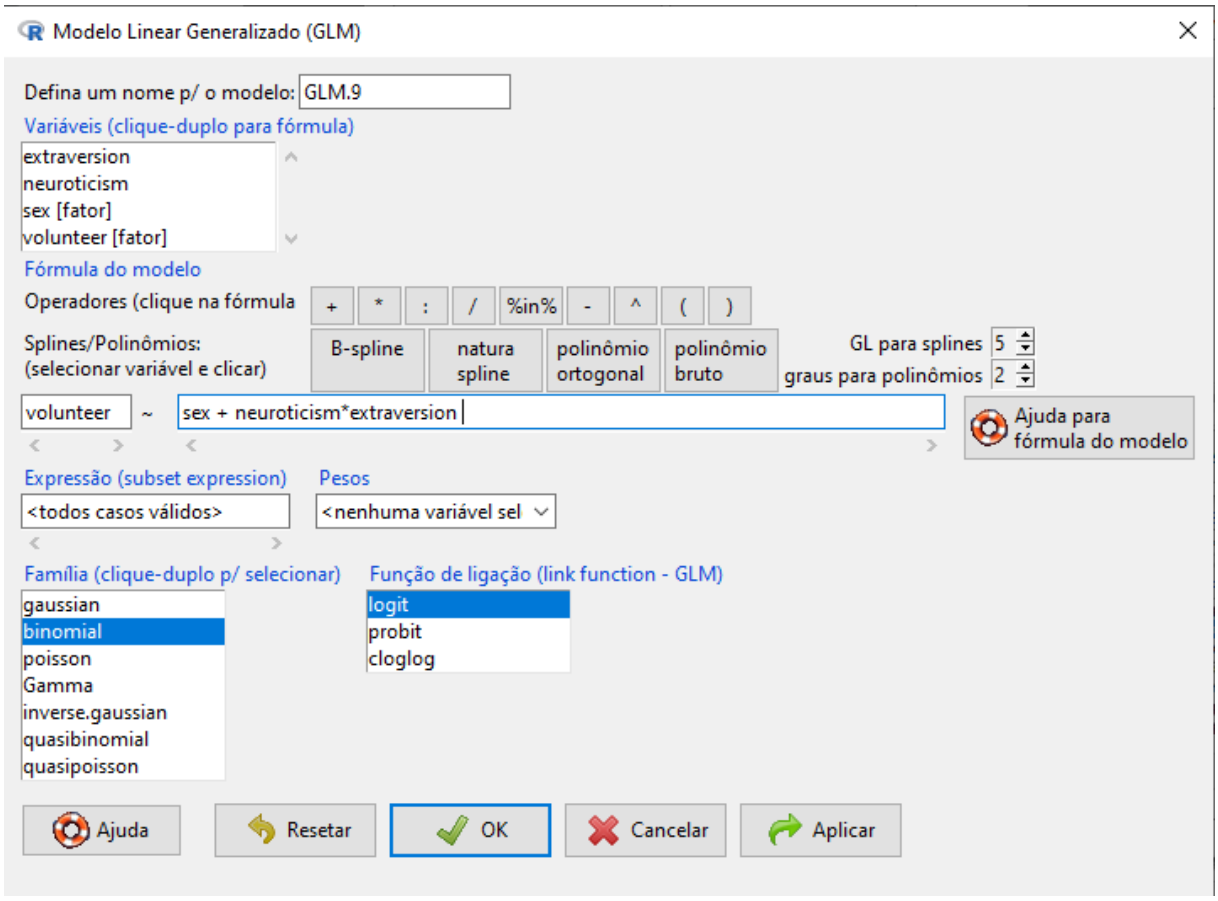

Figura 4.1: Ajuste do modelo de regressão logística *volunteer* $\sim$ *sex + neurotism*extraversion* ajustada para o banco de dados *Cowles*

O que há de novo na caixa de diálogo Modelo Linear Generalizado são as caixas da lista de funções Família e Link, conforme apropriado para um MLG. Famílias e links são coordenados: clicar duas vezes em uma família de distribuição altera os links disponíveis. Em cada caso, o link canônico para uma família específica é selecionado por padrão. As seleções iniciais são família binomial e link de registro canônico correspondente, que coincidentemente são o que desejo para o exemplo.

Prosseguimos para completar o diálogo clicando duas vezes em *volunteer* na lista de variáveis, tornando-a a variável de resposta; em seguida, clicar duas vezes em *sex* e em *neuroticism*; clicando no botão * na barra de ferramentas; e, finalmente, clicando duas vezes em *extraversion* - produzindo o modelo *volunteer* $\sim$ *sex + neuroticism * extravesion.* Como no diálogo Modelo

Linear, uma alternativa é digitar a fórmula diretamente.

As respostas apropriadas para um modelo logit binomial incluem fatores de dois níveis (como voluntário no exemplo atual), variáveis lógicas (ou seja, com valores FALSO e TRUE) e variáveis numéricas com dois únicos valores (mais comumente 0 e 1). Em cada caso, o modelo logit é para a probabilidade do segundo dos dois valores - a probabilidade de que o voluntário seja "sim" no exemplo. Clicar no botão OK produz a saída mostrada abaixo. A caixa do Modelo Linear Generalizado usa a função *glm* do R para ajustar o modelo. A saída de resumo para um Modelo Linear Generalizado é muito semelhante ao de um Modelo Linear, incluindo uma tabela de coeficientes estimados junto com seus erros padrão, valores z (estatísticas de Wald) para testar se os coeficientes são 0 e os valores de p bilateral para esses testes. Para uma regressão logística, o R também imprime os coeficientes exponenciados, interpretáveis como efeitos multiplicativos na escala de probabilidades - aqui as probabilidades de voluntariado, P("sim") / P("não").

Como pode ser visto na saída na figura 4.2, os testes z de Wald sugerem uma interação estatisticamente significativa entre *neuroticism* e *extravesion*, como Cowles e Davis esperava, e um efeito axial significativo, com menos homens propensos a se voluntariar do que mulheres que têm pontuações equivalentes nas dimensões de personalidade.

Embora tenha desenvolvido apenas um exemplo de um MLG nesta seção - um modelo logit para dados binários - a caixa de diálogo do RCommander de Modelos Lineares Generalizados é mais flexível:

- As funções de link probit e log-log complementar (cloglog) também podem ser usadas com dados binários, como alternativas ao link logit canônico.

- A família binomial também pode ser usada quando o valor da variável de resposta para cada caso (ou observação) representa a proporção de "sucessos" em um determinado número de ensaios binomiais, que também pode variar conforme o caso. Neste cenário, o lado esquerdo da fórmula do modelo deve dar a proporção de sucessos, que podem ser acessos / tentativas computados (imaginando que existem variáveis com esses nomes no conjunto de dados ativo) diretamente na caixa à esquerda do modelo fórmula, e a variável que representa o número de tentativas para cada observação (por exemplo, tentativas) deve ser fornecida na caixa de Pesos.

- Alternativamente, para dados binomiais, o lado esquerdo do modelo pode ser uma matriz de duas colunas especificando, respectivamente, os números de sucessos e fracassos para cada observação, digitando, por exemplo, cbind (sucessos,fracassos) (novamente, imaginando que essas variáveis estão no conjunto de dados ativo) na caixa do lado esquerdo da fórmula do modelo.

- Outros modelos lineares generalizados são especificados escolhendo uma família diferente e o vínculo correspondente. Por exemplo, um modelo de regressão de Poisson, comumente empregado para dados de contagem, pode ser ajustado selecionando a família Poisson e o link de registro canônico (ou, para obter erros padrão de coeficiente tipicamente mais realistas, selecionando a família de quasipoisson com o link de registro).

```
Rcmdr>  GLM.9 <- glm(volunteer ~ sex + neuroticism*extraversion,
Rcmdr+    family=binomial(logit), data=Cowles)

Rcmdr>  summary(GLM.9)

Call:
glm(formula = volunteer ~ sex + neuroticism * extraversion, family = binomial(logit),
    data = Cowles)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4749  -1.0602  -0.8934   1.2609   1.9978

Coefficients:
                          Estimate Std. Error z value   Pr(>|z|)
(Intercept)              -2.358207   0.501320  -4.704 0.00000255 ***
sex[T.male]              -0.247152   0.111631  -2.214    0.02683 *
neuroticism               0.110777   0.037648   2.942    0.00326 **
extraversion              0.166816   0.037719   4.423 0.00000975 ***
neuroticism:extraversion -0.008552   0.002934  -2.915    0.00355 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1933.5  on 1420  degrees of freedom
Residual deviance: 1897.4  on 1416  degrees of freedom
AIC: 1907.4

Number of Fisher Scoring iterations: 4


Rcmdr>  exp(coef(GLM.9))  # Exponentiated coefficients (``odds ratios'')
             (Intercept)              sex[T.male]              neuroticism
              0.09458964               0.78102195               1.11714535
            extraversion neuroticism:extraversion
              1.18153740               0.99148400
```

Figura 4.2: Saída para o modelo de regressão logística *volunteer* $\sim$ *sex + neurotism*extraversion* ajustada para o banco de dados *Cowles*

4.1 Interpretação dos coeficientes

Para entendermos como interpretar os coeficiente de variáveis contínuas e dummy, precisamos antes entender o que é a função de ligação e o conceito de razão de chances (ou odds ratio). Modelos Lineares Generalizados incluem uma função de ligação que associa os valores esperados da resposta aos preditores lineares no modelo. Uma função de ligação transforma as probabilidades dos níveis de uma variável de resposta categórica em uma escala contínua que é ilimitada. Depois de concluída a transformação, a relação entre os preditores e a resposta pode ser modelada com regressão linear. Esta é exatemente a função citada no início do capítulo que faz o vínculo entre μ_i e η_i, ou seja, é a função $\mathbf{g}(\mu_i) = \eta_i$, tal que $\mu_i = E(y_i)$. No nosso caso, trabalharemos

apenas com a função de ligação canônica ou função de ligação logit, pois ela é frequentemente utilizada, a interpretação é mais fácil.

Um conceito essencial para entender a função logit é o de razão de chances, também conhecido como razão de possibilidades ou odds ratio. A razão de chances indica o valor da divisão da probabilidade de ocorrer um evento (chamaremos de $P(A)$ ou θ) pela probabilidade de não ocorrer o evento (chamaremos de $P(\bar{A})$ ou $1 - P(A)$ ou $1 - \theta$), ou seja $oddsratio = \frac{P(A)}{(1-P(A))} = \frac{\theta}{(1-\theta)}$. Para entendermos o conceito melhor, vamos supor valores para a razão e o que isso implica no valor da probabilidade do evento em estudo:

- Se $odds = 1$, a probabilidade de ocorrer o evento é igual a de não ocorrer. Também podemos afirmar que $P(A)$ é 0,5;
- Se $odds < 1$, a probabilidade de ocorrer o evento é menor que a de não ocorrer. Também podemos afirmar que $P(A)$ é menor que 0,5;
- Se $odds > 1$, a probabilidade de ocorrer o evento é maior que a de não ocorrer. Também podemos afirmar que $P(A)$ é maior que 0,5.

∗ *Observe que em qualquer caso temos como calcular o valor de $P(A)$, já que $P(A) = \frac{odds}{(1+odds)}$.*

A importância do conceito da razão de chances é que a função de ligação logit é, basicamente, o log da odds. Desta forma, com a função de ligação escolhida, podemos analizar a influência de cada variável no modelo. Considere um MLG logístico simples:

$$ln(\frac{\theta_i}{(1-\theta_i)}) = \beta_0 + \beta_1 \cdot X_{1i}$$

Ou seja

$$\frac{\theta_i}{(1-\theta_i)} = e^{(\beta_0+\beta_1 \cdot X_{1i})}$$

Se X_{1i} aumentar em 1 unidade, temos:

$$e^{(\beta_0+\beta_1 \cdot (X_{1i}+1))} = e^{(\beta_0+\beta_1+\beta_1 X_{1i})} = e^{(\beta_1)} \cdot e^{(\beta_0+\beta_1 \cdot X_{1i})}$$

Ou seja, podemos concluir que aumentar em uma unidade X1i causa um efeito em que a razão de chances é multiplicada por exp(beta1). Isso implica que:

- Se $\beta_1 = 0$, o efeito é nulo;
- Se $\beta_1 > 0$, temos $e^{(\beta_1)} > 1$, aumentando a odds ratio e por consequência também o θ_i;
- Se $\beta_1 < 0$, temos $e^{(\beta_1)} < 1$, diminuindo a odds ratio e por consequência também o θ_i;

*O*bserve que essa interpretação vale para variáveis dummy dicotômicas, já que o valor 1 é exatamente o aumento que indica em um modelo que a observação i pertece à categoria referente à variável X_1. Por exemplo, imagine que X_1 é uma variável categórica que indique 1 caso a observação tenha sido diagnósticada para ansiedade e 0 caso contrário. Caso a observação i perteça à categoria que a variável X_1, seu valor será 1 (um aumento de 1 na variável). Isso implica na multiplicação da razão de chances para esta observação em e^{β_1}.

Em valores porcentuais, podemos dizer que o aumento em uma unidade altera a odds ratio em $100 \cdot (e^{\beta_1} - 1)\%$. No exemplo a seguir poderemos trabalhar de forma mais prática estes conceitos.

4.2 Exemplo usando o banco de dados binary

Um segundo exemplo pode ser obtido pelo banco de dados **binary.csv** disponível em `https://www.kaggle.com/aakashmisra/binary-dataset`.

```
> binary<-read.csv(‘‘binary.csv’’)
> str(binary)
’data.frame’:        400 obs. of  4 variables:
 $ admit: int  0 1 1 1 0 1 1 0 1 0 ...
 $ gre  : int  380 660 800 640 520 760 560 400 540 700 ...
 $ gpa  : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
 $ rank : int  3 3 1 4 4 2 1 2 3 2 ...
```

Figura 4.3: Carregando o banco de dados *binary*

Observando a estrutura dos conjuntos de dados, podemos observar que ela possui 4 variáveis. A variável admit informa se um candidato será admitido ou não (1 se admitido e 0 se não admitido) em um processo para a entrada em certa faculdade. Já as outra variáveis gre, gpa e rank atribuem aos candidatos uma pontuação no exame Graduate Record Examination (GRE), sua nota média no ensino médio (Grade Point Average ou GPA) e na classificação da escola ao qual o aluno provém (sendo 1 o melhor rank e 4 o pior), respectivamente.

Como as variáveis categóricas admit e rank estão em tipo numérico, é necessário usar o comando as.factor() para converte-las para o tipo correto, conforme feito abaixo.

```
> #corrigindo tipo das variáveis
> binary$admit = as.factor(binary$admit)
> binary$rank = as.factor(binary$rank)
```

Figura 4.4: Conversão das variáveis categóricas do banco de dados *binary* para o seu tipo correto (fator)

Utilizando o RCommander, em clica-se em:

Estatísticas → Ajuste de Modelos → Modelo Linear Generalizado (GLM)

Isso gera um modelo que tem como intensão predizer a variável binárias admit, selecionamos ela como resposta e as restantes como explicativas com uma família binomial e função de ligação logit, conforme feito na imagem abaixo. Para facilitar a interpretação, não utilizaremos um intercepto, informando “+ 0” na fórmula para indicar esta escolha.

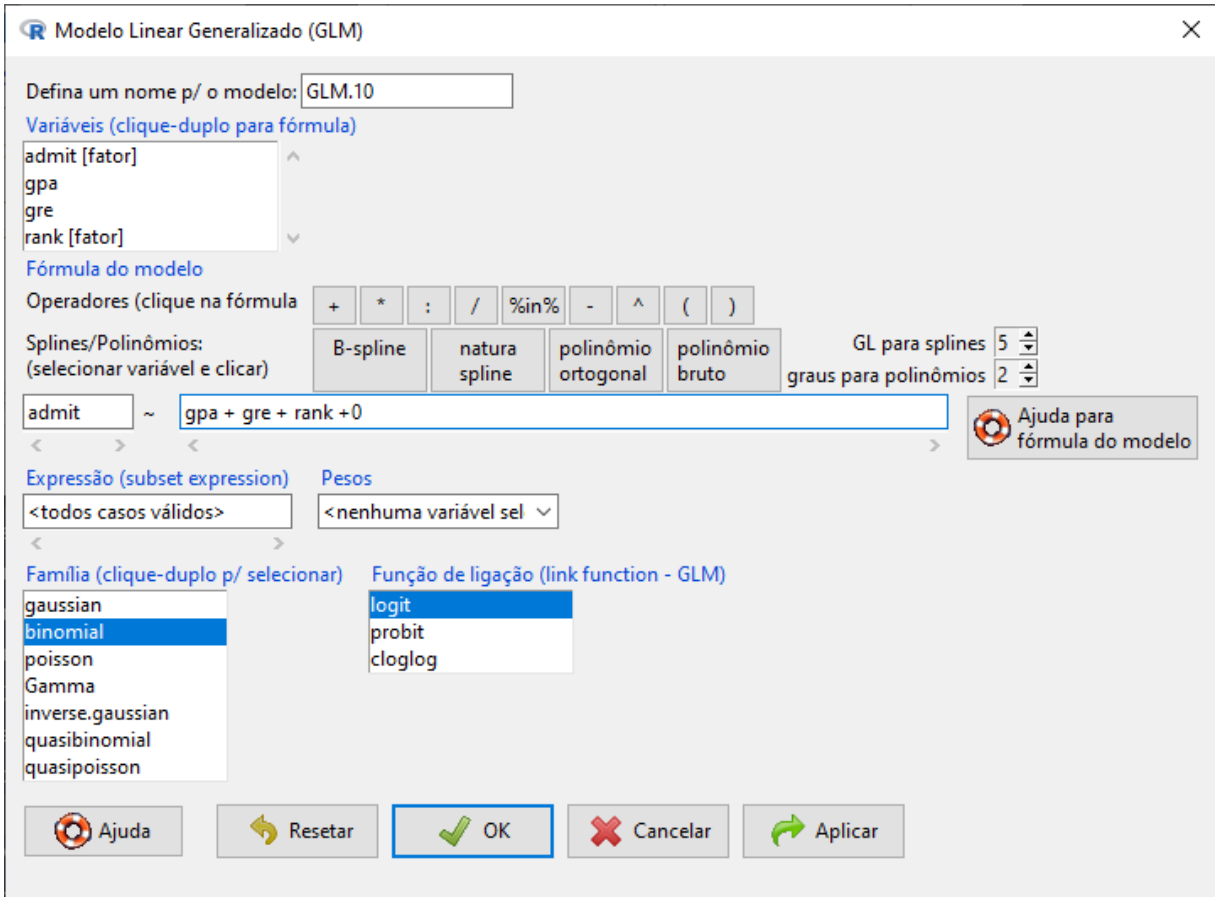

Figura 4.5: Ajuste do modelo de regressão logística *admit* $\sim$ *gpa* + *gre* + *rank* + *0* ajustada para o banco de dados *binary*

Nossa saída está na Figura 4.6. Nesta figura podemos ver as nossas razões de chances ("odds ratio") indicam que o GPA tem uma odds favorável à admissão, GRE nos informa uma influência favorável à odds e o rank da escola indica que, já que a odds entre os fatores de rank é decrescente, um rank pior da escola diminui as chances de o aluno ser admitido na faculdade. Ainda sobre o rank, podemos dizer que a odds muda de 0.0185 para 0.0094 só de o aluno ser de uma escola de rank 2 ao invés de ser de uma de rank 1, ou seja um aluno de escola de rank 1 tem quase 2 vezes mais chances de ser aprovado do que um de escola de rank 2, antes mesmo de considerar o GPA e o GRE. Isso indica que o ranking da escola já traz uma diferença grande na chance de aprovação, mostrando que o sistema não traz chance iguais entre alunos de escolas diferentes (lembrando, sem nem considerar o GPA e o GRE).

```
Rcmdr>  GLM.10 <- glm(admit ~ gre + gpa + rank +0, family=binomial(logit),
Rcmdr+    data=binary)

Rcmdr>  summary(GLM.10)

Call:
glm(formula = admit ~ gpa + gre + rank + 0, family = binomial(logit),
    data = binary)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6268  -0.8662  -0.6388   1.1490   2.0790

Coefficients:
       Estimate Std. Error z value   Pr(>|z|)
gpa    0.804038   0.331819   2.423   0.015388 *
gre    0.002264   0.001094   2.070   0.038465 *
rank1 -3.989979   1.139951  -3.500   0.000465 ***
rank2 -4.665422   1.109370  -4.205 0.00002605 ***
rank3 -5.330183   1.149538  -4.637 0.00000354 ***
rank4 -5.541443   1.138072  -4.869 0.00000112 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 554.52  on 400  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4


Rcmdr>  exp(coef(GLM.10))  # Exponentiated coefficients (``odds ratios'')
        gpa         gre       rank1       rank2       rank3       rank4
2.234544824 1.002266992 0.018500101 0.009415274 0.004843184 0.003920866
```

Figura 4.6: Saída para o modelo de regressão logística *admit* $\sim$ *gpa + gre + rank + 0* ajustada para o banco de dados *binary*

4.3 Outros Modelos de Regressão

Além de Regressão Linear, Modelos Lineares e Modelos Lineares Generalizados, o RCommander pode ajustar modelos de logit multinomial para variáveis de resposta categóricas com mais de duas categorias da seguinte maneira:

Estatísticas → Ajuste de Modelos → Modelo Logito Multinomial

e modelos de regressão ordinal para respostas ordenadas de várias categorias, incluindo o modelo de logit de odds proporcionais e o modelo probit ordenado como segue:

Estatísticas → Ajuste de Modelos → Modelo de regressão ordinal

Embora não vamos ilustrar todos esses modelos aqui, muitos dos itens de menu no menu Modelos se aplicam a essas classes de modelos[2].

4.3.1 Modelos de Regressão Binária: Logit

Como já dito anteriormente (Sessão 3.5), alguns modelos requerem o uso de variáveis categóricas binárias, as chamadas *dummy* ou *dummies*. Portanto, quando queremos modelar esse tipo de variável como explicada, podemos fazer uso do recurso de Modelo logit multinominal do RCommander. Destarte, quando modelamos uma Regressão Logística supomos uma relação de sucesso/fracasso para a variável que desejamos explicar, sendo "1" um sucesso e "0" um fracasso. Dessa forma podemos entender o impacto de outras variáveis na probabilidade de obter um sucesso ou um fracasso.

A utilização desse modelo se dá, em grande maioria, para avaliar resposta da variável categórica à um impulso em uma de suas variáveis exógenas. Isso se deve, pois como o nome já sugere, o modelo logit, é um modelo logarítimo da razão entre a probabilidade de um sucesso e a probabilidade de um fracasso (*log-odds*). Para entender melhor o que foi explicitado aqui, utilizaremos um modelo logit sobre um estudo de painel de eleições.

Exemplo de Regressão logit Usando o Banco de Dados *BEPS*

Utlizaremos o banco de dados *BEPS* que está incluso no pacote **car**. Para importá-lo no RCommander basta:

Dados → Conjunto de dados em pacote → Ler dados de pacote "attachado"

Então selecionar o pacote **carData** e selecionar o banco de dados *BEPS*.

Esse banco de dados traz informações acerca de um estudo de painel das eleições britânicas. Dessa forma, afim de exemplificar o uso de uma regressão logit, basta seguir a indicação dada no início do capítulo. É importante ressaltar que um Modelo de Regressão Logística trabalha com a predição de uma variável categórica, e nesse caso, as variáveis categóricas são regredidas com variáveis categóricas (idade e gênero) e contínuas (idade).

Diante disso, podemos realizar um modelo para probabilidade de votos dado alguns fatores, são eles: idade (*age*), gênero (*gender*), conhecimento político (*political knowledge*). O modelo está descrito na figura 4.7. Dessa forma, podemos avaliar o impacto de um choque exógeno em cada uma dessas variáveis na intenção de voto (sucesso = "1") em cada um dos partidos.

[2]Para entender mais sobre como a escolha do modelo ideal ocorre, acesse o apêndice C

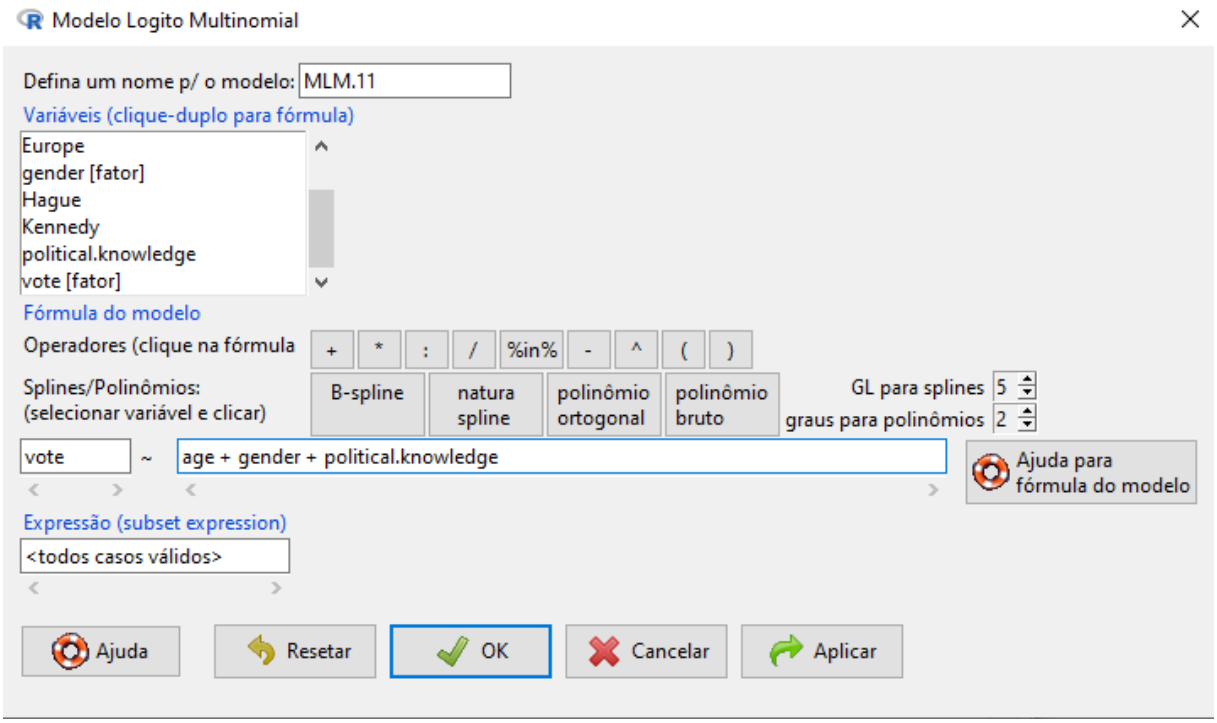

Figura 4.7: Ajuste do modelo de regressão logística múltipla *vote* $\sim$ *age* + *gender* + *political.knowledge* ajustada para o banco de dados *BEPS*

Dito isso, os resultados do modelo (ver Figura 4.8) nos sugerem certas informações interessantes. Dentre elas, podemos ver que o conhecimento sobre política tem um impacto muito mais negativo nas intenções de votos do Partido dos Trabalhadores do que do Partido Liberal Democrata. Além disso, vemos que os homens tem um tendência maior a votar no Partido dos Trabalhadores. Outro ponto interessante, é que existe uma maior probabilidade do Partido dos Trabalhadores vencerem as eleições subsequentes ao estudo, o que veio a se concretizar no ano de 2001.

```
Rcmdr>  MLM.11 <- multinom(vote ~ age + gender + political.knowledge, data=BEPS,
Rcmdr+    trace=FALSE)

Rcmdr>  summary(MLM.11, cor=FALSE, Wald=TRUE)

Call:
multinom(formula = vote ~ age + gender + political.knowledge,
    data = BEPS, trace = FALSE)

Coefficients:
                 (Intercept)         age gender[T.male] political.knowledge
Labour              1.783757 -0.01707121      0.2897226         -0.34596764
Liberal Democrat    0.634807 -0.01587298      0.1477722         -0.07540355

Std. Errors:
                 (Intercept)         age gender[T.male] political.knowledge
Labour             0.2519853 0.003892352      0.1237596          0.05842057
Liberal Democrat   0.2976980 0.004600146      0.1458591          0.07029038

Value/SE (Wald statistics):
                 (Intercept)       age gender[T.male] political.knowledge
Labour              7.078814 -4.385834       2.341011           -5.922017
Liberal Democrat    2.132386 -3.450538       1.013117           -1.072744

Residual Deviance: 3143.818
AIC: 3159.818
```

Figura 4.8: Saída para o modelo de regressão logística múltipla *vote* $\sim$ *age + gender + political.knowledge* ajustada para o banco de dados *BEPS*

Capítulo 5

Visualizando Modelos Lineares e Lineares Generalizados

Introduzido por Fox (1987), *effect plots* são gráficos para visualizar modelos de regressão complexos focalizando variáveis explicativas particulares ou combinação de variáveis explicativas, mantendo outras variáveis explicativas em valores típicos. Uma estratégia é focar sucessivamente nas variáveis explicativas no *high-order terms* do modelo – isto é, termos que não são marginais a outros (ver Seção 3.3).

No RCommander, exibições de efeito podem ser desenhadas para linear, linear generalizado e alguns outros modelos estatísticos por meio de

Modelos* → *Gráficos* → *Effects plots...

A Figura 5.1 mostra a caixa de diálogo resultante para a regressão logística de Cowles e Davis da seção anterior, GLM.9, que é o modelo estatístico atual no RCommander. Por padrão, o diálogo oferece plotar todos os termos de ordem superior no modelo - neste caso, o efeito principal *sex* e a interação *neuroticism* por *extraversion* . Você pode, alternativamente, escolher um subconjunto de *Preditores* (variáveis explicativas) para representar graficamente. Para um modelo linear ou linear generalizado, há também uma caixa de seleção para traçar residuais parciais, desmarcada por padrão, junto com um controle deslizante para a extensão de um ajuste mais suave aos residuais. Resíduos parciais e o mais suave que os acompanha podem ser úteis para julgar desvios da forma funcional do modelo especificado, como ilustraremos posteriormente nesta seção.

Clicar em *OK* produz o gráfico da Figura 5.2. O painel esquerdo mostra o efeito principal *sex*, com *neuroticism* e *extraversion* configurados para níveis médios. O painel direito mostra a interação neuroticismo por extroversão, para um grupo composto por homens e mulheres na proporção de sua representação no conjunto de dados. Em ambos os gráficos, o eixo voluntário vertical é desenhado na escala *logit*, mas os rótulos das marcas de verificação estão na escala de probabilidade estimada - isto é, eles representam a probabilidade estimada de voluntariado.

No gráfico da interação, o eixo horizontal de cada painel é para *neuroticism*, enquanto *extraversion* assume valores sucessivamente maiores em seus intervalos, do painel inferior esquerdo ao painel superior direito. O valor de *extraversion* para cada painel é representado pela pequena linha vertical na faixa denominada *extraversion* na parte superior do painel.

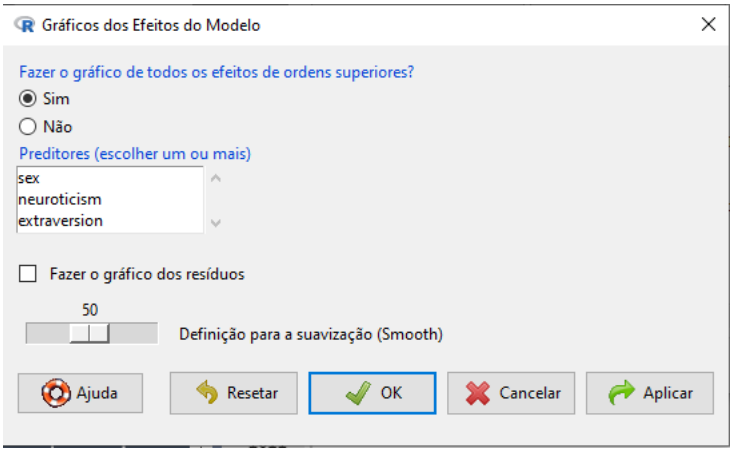

Figura 5.1: Ajuste dos Gráficos dos Efeitos do Modelo para o modelo GML.9

As linhas nos painéis representam o efeito combinado de *neuroticism* e extroversão e são calculadas usando os coeficientes estimados para interação *neuroticism:extraversion* junto com os coeficientes para *neuroticism* e "efeitos principais" de *extraversion*, que são marginais à interação. É claro que há uma relação positiva entre *volunteer* e *neuroticism* no nível mais baixo de *extraversion*, mas que essa relação se torna negativa no nível mais alto de *extraversion*.

As barras de erros no gráfico de efeito para sex e as bandas cinzas no gráfico da interação neuroticismo por extroversão representam intervalos de confiança de 95% pontuais em torno dos efeitos estimados. O "*rug-plot*" (gráfico de tapete) na parte inferior de cada painel na exibição da *interação neuroticismo por extroversão* mostra a distribuição marginal do *neuroticism*, com as linhas ligeiramente deslocadas para diminuir o *over-plotting*. O *rug-plot* não é terrivelmente útil aqui, porque o *neuroticism* assume apenas pontuações inteiras.

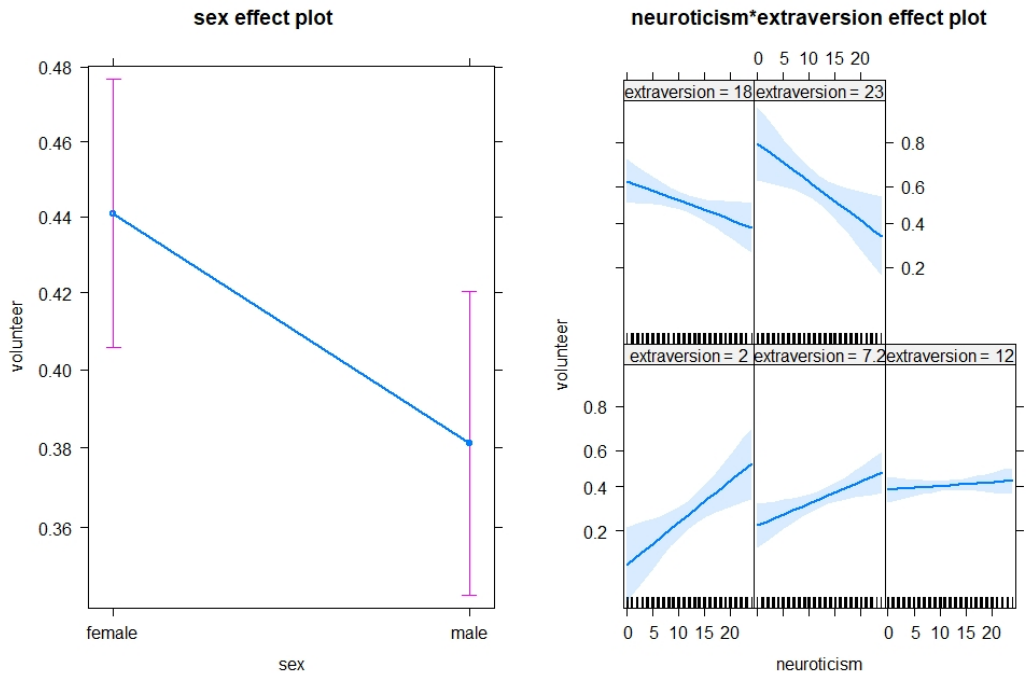

Figura 5.2: Gráficos dos Efeitos do Modelo para o modelo GML.9

5.1 Resíduos Parciais em Gráficos de Efeitos

Adicionar *resíduos parciais* para efetuar gráficos de variáveis explicativas numéricas em modelos lineares e lineares generalizados pode ser uma ferramenta eficaz para julgar desvios da forma funcional (linear ou não) especificada no modelo. Vamos ilustrar usando os dados de prestígio ocupacional canadense. Nas Sessão 3.6 , ajustamos vários modelos aos dados *Prestige*, incluindo um modelo aditivo de regressão simulada (*LinearModel.2* na Figura 3.5),

$$prestige \sim education + income + women + type$$

e um modelo com interações (*LinearModel.3* na Figura 3.6),

$$prestige \sim type * education + type * income$$

entre outros.

O RCommander permite alternar entre modelos e conjuntos de dados e se encarrega de sincronizar os modelos com os conjuntos de dados aos quais eles se ajustam. Depois de tornar *LinearModel.2* o modelo ativo, abre-se novamente a caixa de diálogo *Gráficos dos Efeitos do Modelo*, exibida na Figura 5.3, selecionaremos a caixa *Fazer o gráfico dos resíduos* e clico em *OK*, produzindo o que se apresenta na Figura 5.4. Resíduos parciais são plotados para os preditores numéricos, mas não para o fator *type*; isso se reflete em um aviso impresso nos painéis *Messages*, que simplesmente ignorarei.

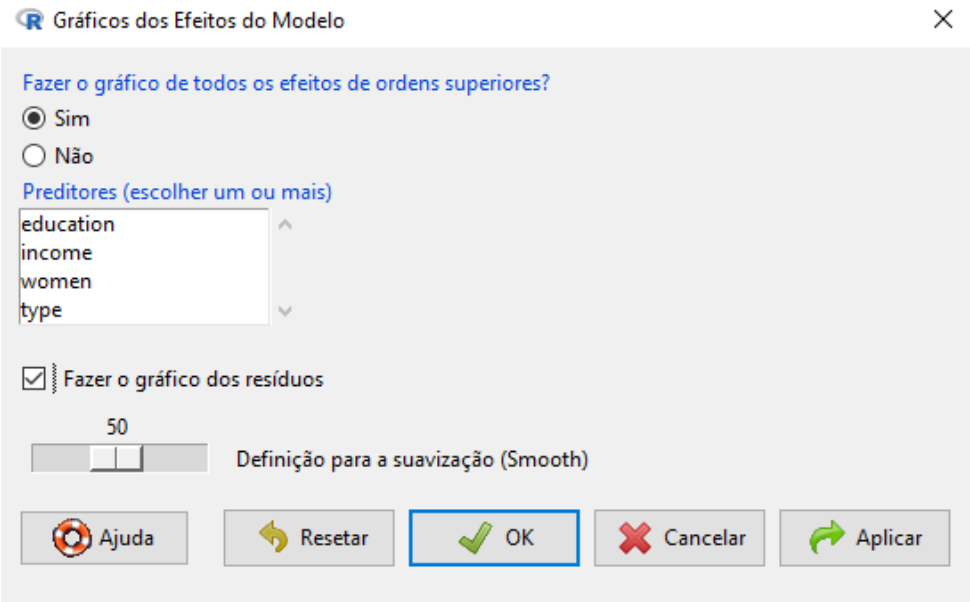

Figura 5.3: Ajuste dos Gráficos dos Efeitos do Modelo para o modelo LinearModel.2

As linhas sólidas nos gráficos de efeito representam o ajuste do modelo aos dados, enquanto as linhas quebradas e suaves dos resíduos parciais. Se as linhas para o modelo ajustado e os resíduos parciais suavizados forem semelhantes, isso dará suporte à forma funcional especificada do modelo. Os residuais parciais são calculados adicionando o residual de cada observação à linha que representa o efeito ajustado. Parece que o efeito *education* foi modelado razoavelmente, mas os efeitos *income* e *women* parecem ser não lineares.

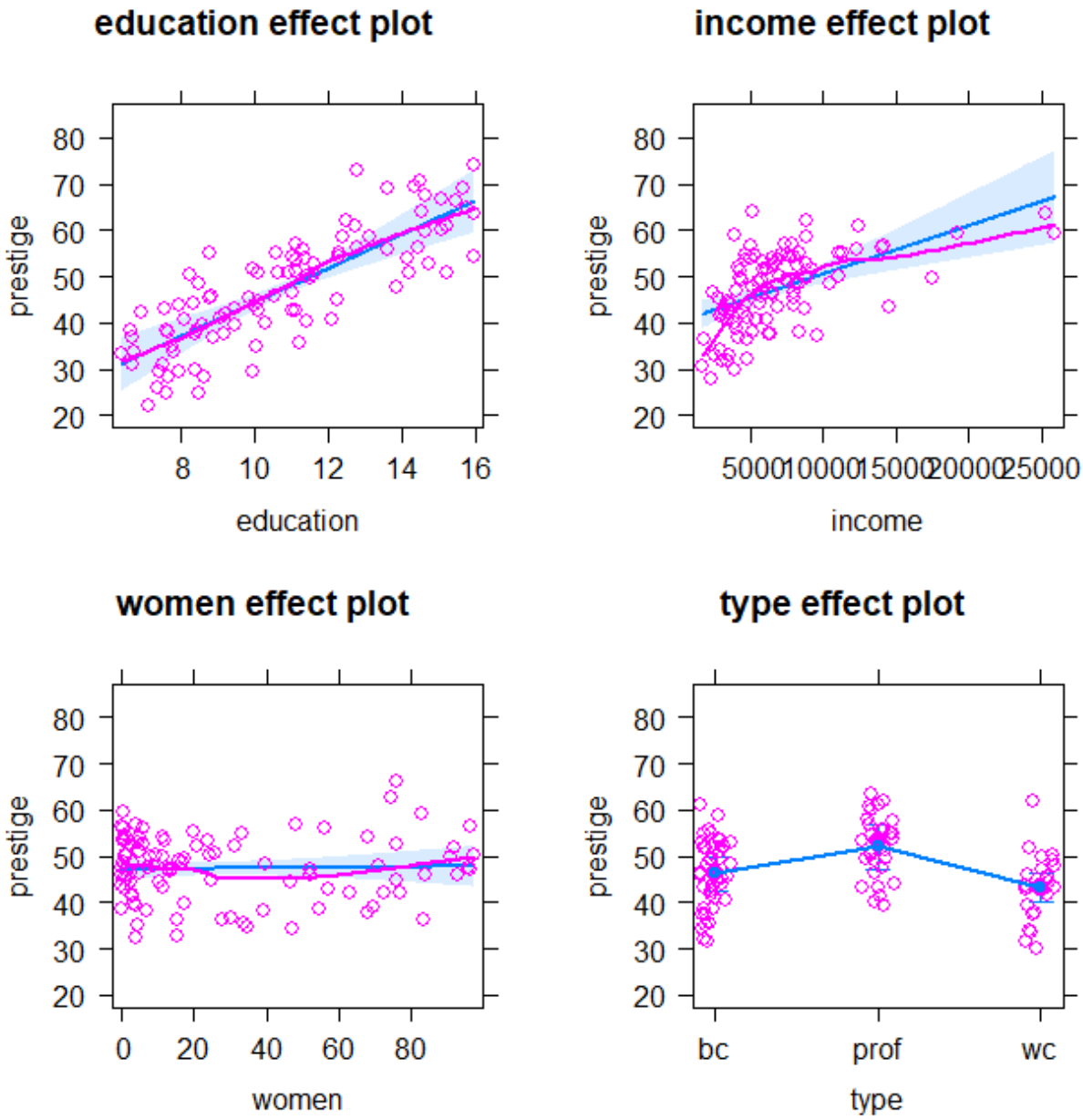

Figura 5.4: Gráficos dos Efeitos do Modelo para o modelo LinearModel.2

LinearModel.3 inclui interações entre *type* e cada um de *education* e *income*. A Figura 5.5 mostra gráficos de efeito com resíduos parciais para os termos de ordem superior neste modelo. Como dividir os dados por *type* deixa relativamente poucos pontos em cada painel dos gráficos, foi definido no exemplo a amplitude do mais suave para um valor alto, 0,9.

A aparente não linearidade na relação entre *prestige* e *income* é explicada pela interação entre *income* e *type*. A exibição do lado direito na Figura 5.5 mostra que a inclinação income é menor para as ocupações profissionais e gerenciais (ou seja, *type* = "prof") do que para as ocupações de *blue-collar* ("bc", colarinho azul) ou de *white-colllar* ("wc", colarinho branco), e as ocupações profissionais tendem a ter rendimentos mais elevados. A exibição à esquerda, para a *interação educação por tipo*, sugere que a inclinação *education* é mais acentuada para ocupações de colarinho branco do que para os outros tipos de ocupações. As suaves dos resíduos parciais indicam que essas relações são lineares dentro dos níveis de *type*.

Os envelopes de confiança nas exibições de efeito com residual parcial na Figura 5.5 também são um ponto pedagógico útil sobre a precisão da estimativa da superfície de regressão:

Onde os dados são esparsos – ou, no extremo, ausentes – a superfície de regressão é estimada de forma imprecisa.

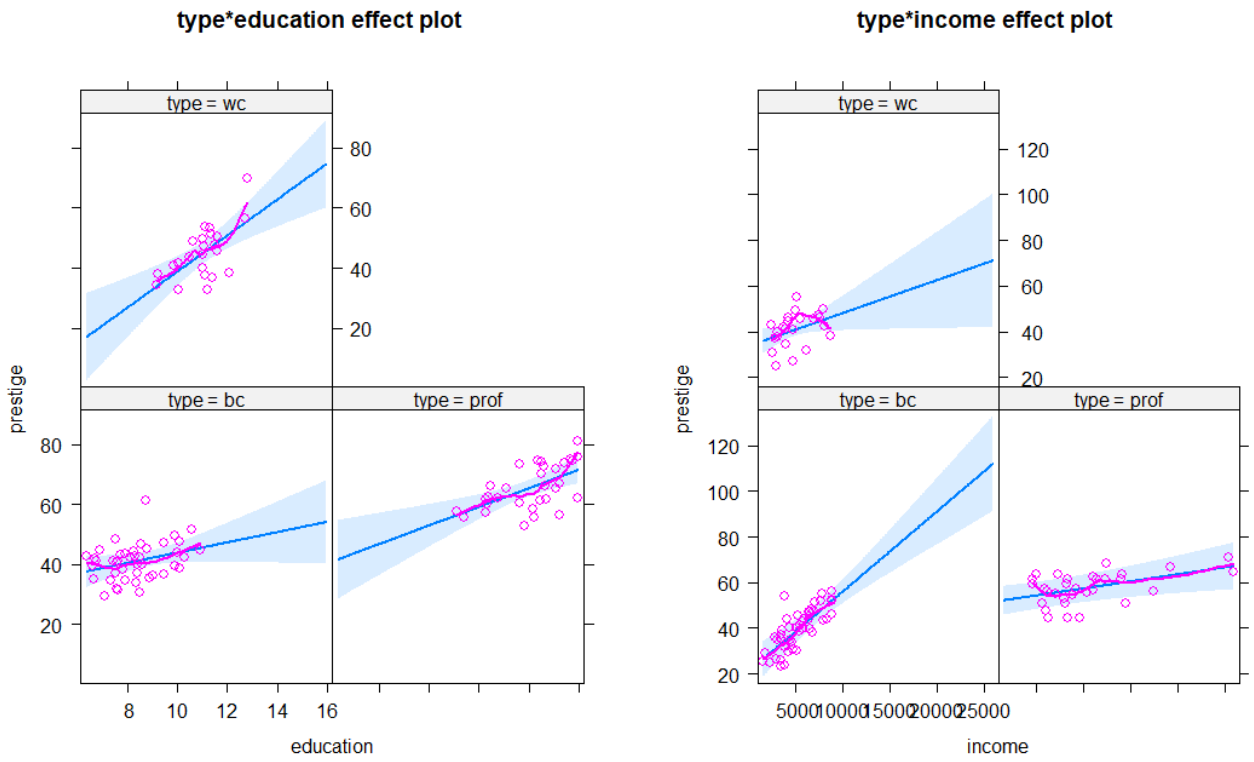

Figura 5.5: Ajuste dos Gráficos dos Efeitos do Modelo para o modelo LinearModel.3

Apêndice A

Ajustando Splines de Regressão e Polinômios

A segunda barra de ferramentas na caixa de diálogo em "Modelo Linear" torna mais fácil modelar regressões polinomiais não lineares e de regressões spline juntamente à um modelo linear.

A.1 Termos Polinomiais

Algumas relações não lineares simples podem ser representadas como polinômios de ordem inferior, como termo quadrático, usando regressores x e x^2 para uma variável x explicativa numérica ou um termo cúbico, usando x, x^2 e x^3. O modelo resultante é não linear na variável explicativa x, mas linear nos parâmetros (os β s). O R e o RCommanders suportam polinômios ortogonais e "brutos" nos modelos lineares.

Para adicionar um termo polinomial ao lado direito do modelo, clique em uma variável numérica na lista de variáveis e, em seguida, pressione o botão da barra de ferramentas apropriado (polinomial ortogonal ou polinômio spline, conforme desejado). Há um spinner na caixa de diálogo em "Modelo Linear" para o grau de um termo polinomial, e o padrão é 2 (ou seja, um quadrático).

Por exemplo, a inspeção dos dados sugere que pode haver uma relação quadrática parcial entre prestígio e mulheres na regressão de prestígio em educação, renda e mulheres para o Dados de Prestígio Ocupacional Canadense, base de dados que já tratamos anteriormente. Uma vez especificada essa relação quadrática no modelo linear, usando um polinômio de segundo grau bruto (ver Figura A.1), produz a saída. O coeficiente quadrático no modelo acaba por não ser estatisticamente significativo ($p = 0,15$), como vemos na Figura A.2.

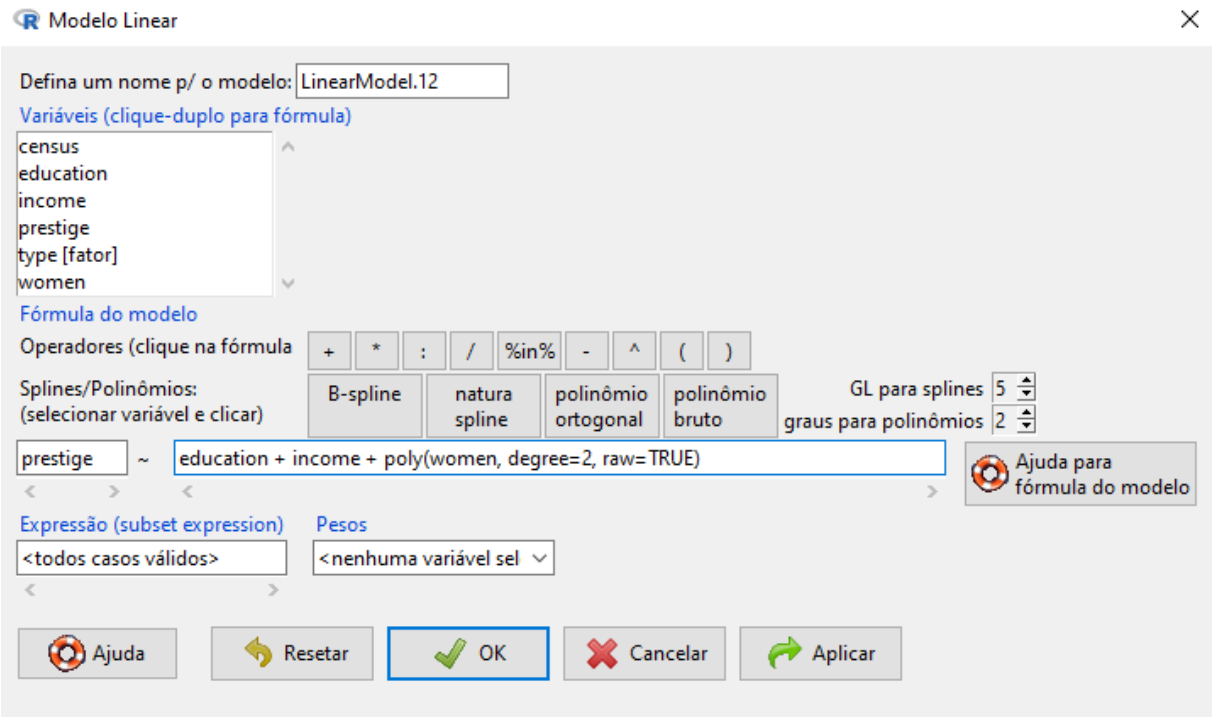

Figura A.1: Ajuste do modelo linear com um termo polinomial (quadrático) *prestige* $\sim$ *education + income + poly(women, degree=2, raw=TRUE)* ajustada para o banco de dados *Prestige*

```
Rcmdr>  LinearModel.12 <- lm(prestige ~ education + income + poly(women, degree=2,
Rcmdr+    raw=TRUE), data=Prestige)

Rcmdr>  summary(LinearModel.12)

Call:
lm(formula = prestige ~ education + income + poly(women, degree = 2,
    raw = TRUE), data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max
-20.2331  -5.3217   0.0987   4.9248  17.0059

Coefficients:
                                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)                          -6.1766164  3.2496931  -1.901    0.0603
education                             4.2601722  0.3899199  10.926   < 2e-16
income                                0.0012720  0.0002778   4.579 0.0000139
poly(women, degree = 2, raw = TRUE)1 -0.1451676  0.0991716  -1.464    0.1465
poly(women, degree = 2, raw = TRUE)2  0.0015379  0.0010660   1.443    0.1523

(Intercept)                          .
education                            ***
income                               ***
poly(women, degree = 2, raw = TRUE)1
poly(women, degree = 2, raw = TRUE)2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.804 on 97 degrees of freedom
Multiple R-squared:  0.8024,        Adjusted R-squared:  0.7943
F-statistic: 98.48 on 4 and 97 DF,  p-value: < 2.2e-16
```

Figura A.2: Saída para o modelo linear com um termo polinomial (quadrático) *prestige* $\sim$ *education + income + poly(women, degree=2, raw=TRUE)* ajustada para o banco de dados *Prestige*

A.2 Regressão Spline

As splines de regressão são funções flexíveis capazes de representar uma ampla variedade de padrões não lineares em um modelo que, como um polinômio de regressão, é linear nos parâmetros. Ambas as splines B e as splines naturais são suportadas pelo RCommander. Adicionar um termo spline ao lado direito de um modelo linear é semelhante a adicionar um termo polinomial: os graus de liberdade de um termo spline são controlados por um botão giratório, com o valor padrão 5; clique uma vez em uma variável na lista e, em seguida, pressione o botão da barra de ferramentas para B-spline ou spline natural.

Por exemplo, mantido a especificação quadrática para mulheres, selecione educação e pressione o botão de spline natural. Em seguida, selecione renda e pressione o botão de spline natural novamente. Em ambos os casos, o spinner manteve o seu valor padrão. Essas escolhas produzem a fórmula do modelo $prestige \sim poly(women, degree = 2, raw = TRUE) + ns(education, df =$

5) + $ns(income, df = 5)$, regredindo o prestígio em mulher ao quadrado e 5-df natural splines na educação e renda (ver Figura A.3). A saída para o modelo de regressão resultante não é mostrada porque o modelo requer interpretação gráfica. As estimativas de coeficientes para as curvas de regressão não são diretamente interpretáveis.

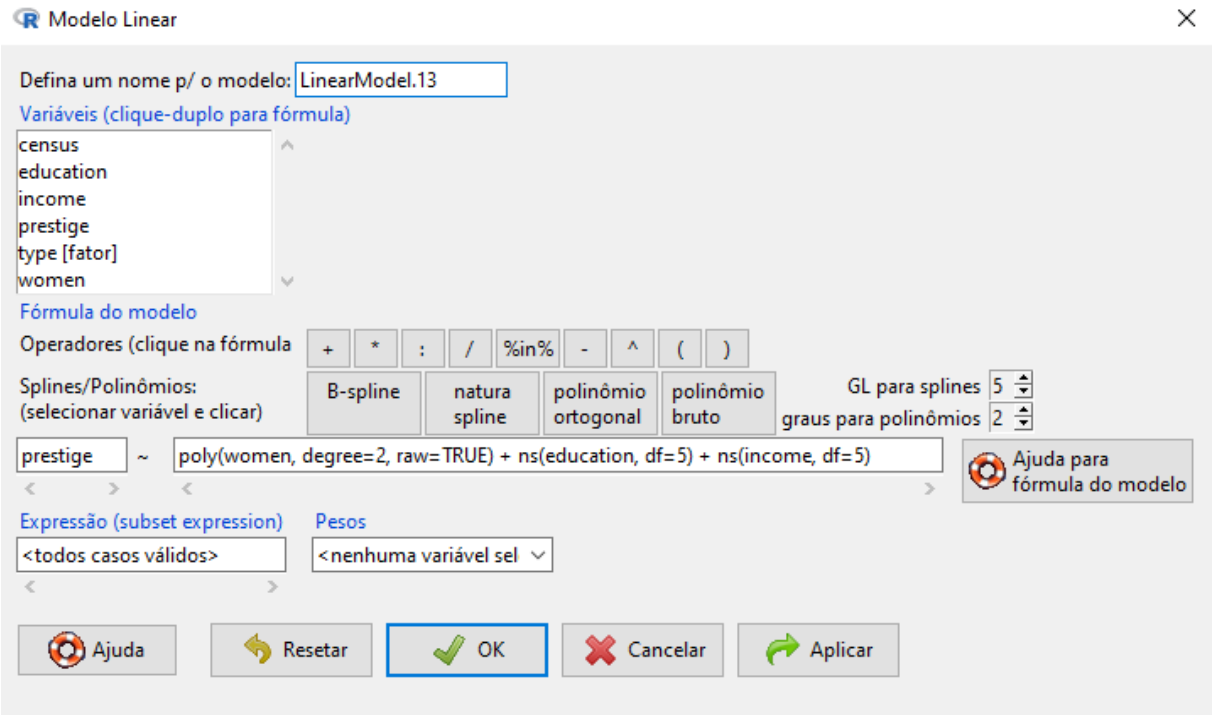

Figura A.3: Ajuste do modelo de regressão-spline com um termo polinomial (quadrático) *prestige* ~ *poly(women, degree=2, raw=TRUE) + ns(education, df=5) + ns(income, df=5)* ajustada para o banco de dados *Prestige*

Apêndice B

Intervalos de Confiança e Teste de Hipótese para Coeficientes

O menu *Modelos* inclui vários itens de menu para construir intervalos de confiança e realizar testes de hipótese para coeficientes de regressão. Como explicado, testes para coeficientes individuais em modelos lineares e lineares generalizados aparecem nos resumos do modelo. Esta seção descreve como realizar testes mais elaborados, por exemplo, para um subconjunto de coeficientes relacionado.

B.1 Intervalos de Confiança

De forma ilustrativa, a regressão de prestígio ocupacional de Duncan (RegModel.1 na Figura 2.2), selecione o modelo estatístico ativo no RCommander, o que automaticamente torna Duncan o conjunto de dados ativo.

Selecione Modelos → Intervalo de Confiança nos menus RCommander leva à caixa de diálogo de amostra a Figura B.1. Mantenha o nível de confiança padrão de 0,95 e clique em *OK* produz a saída descrita na Figura B.1.

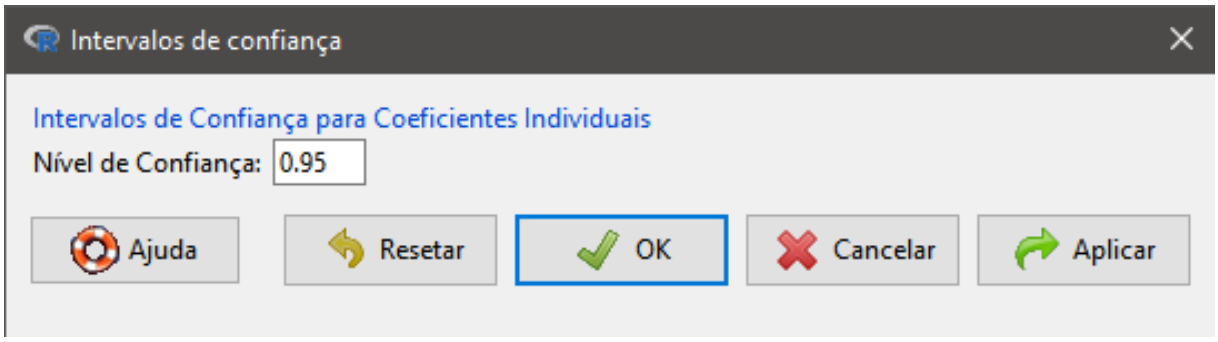

Figura B.1: Seleção do Intervalo de Confiança

```
> Confint(RegModel.1, level=0.95)
              Estimate       2.5 %    97.5 %
(Intercept) -6.0646629 -14.6857892 2.5564634
education    0.5458339   0.3475521 0.7441158
income       0.5987328   0.3572343 0.8402313
```

Figura B.2: Saída do Teste de Intervalo de Confiança para o modelo *Reg.Model1*

B.2 Tabelas de Análise de Variância e Análise de Desvio

Você pode calcular uma tabela de análise de variância (ANOVA) para um modelo linear ou uma tabela análoga de análise de desvio para um modelo linear generalizado via:

Modelos → Testes de Hipóteses → Tabela de ANOVA

A mérito de exemplo, a regressão logística de Cowles e Davis (GLM.9 na Figura 4.1).

A caixa de diálogo da tabela ANOVA na Figura B.3 oferece três "tipos" de testes, convencionalmente chamados de Tipos I, II e III. Selecionando o Tipo II, obtemos a saída da Figura B.4.

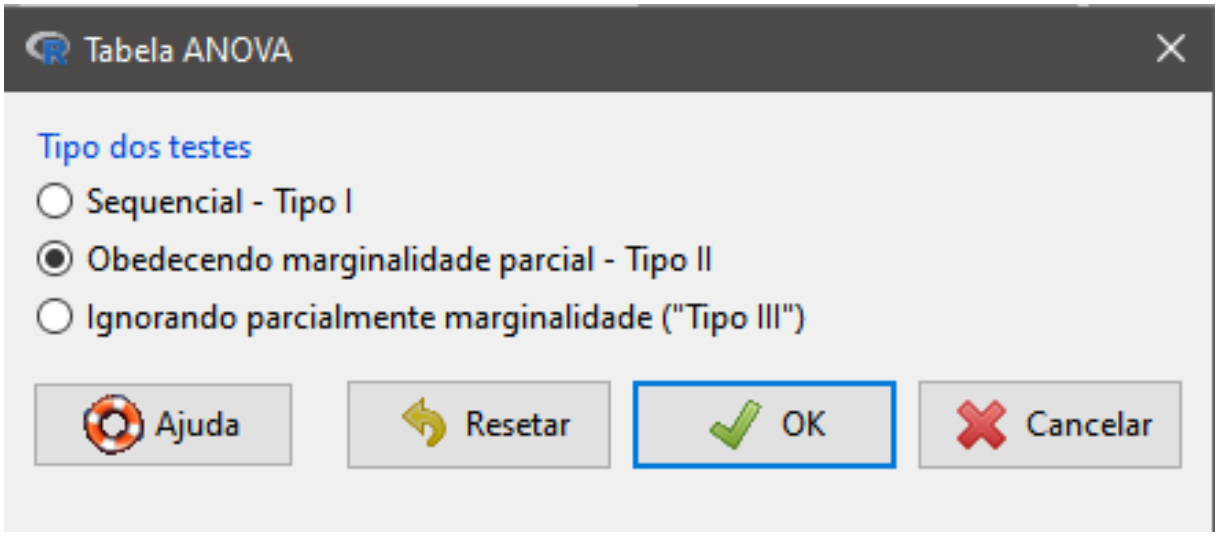

Figura B.3: Seleção do Método

```
> Anova(GLM.9, type=``II'', test=``LR'')
Analysis of Deviance Table (Type II tests)

Response: volunteer
                        LR Chisq Df  Pr(>Chisq)
sex                       4.9184  1    0.026572 *
neuroticism               0.3139  1    0.575316
extraversion             22.1372  1 0.000002538 ***
neuroticism:extraversion  8.6213  1    0.003323 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura B.4: Saída da Tabela ANOVA do Tipo II para o modelo *GLM.9*

- Além do intercepto, há quatro termos no modelo de Cowles e Davis: sexo (*sex*), neuroticismo (*neurotism*), extroversão (*extraversion*) e a interação entre neuroticismo e extroversão (*neurotism:extraversion*). Os testes Tipo 1 são sequenciais e, portanto, testam (em uma terminologia abreviada) o sexo ignorando todo o resto; neuroticismo após o sexo, mas ignorando a extroversão e a interação entre neuroticismo e extroversão; extroversão após o sexo e neuroticismo ignorando a interação; e a interação entre neuroticismo e extroversão após todos os outros termos. Os testes de sequência raramente são sensatos.

- Os testes Tipo II são formulados em conformidade com o princípio da marginalidade (Seção 3.3): sexo após todos os outros termos, incluindo a interação neuroticismo:extroversão; neuroticismo após o sexo e extroversão, mas ignorando a interação neuroticismo:extroversão à qual o "efeito principal" do neuroticismo é marginal; da mesma forma, extroversão após sexo e neuroticismo, mas ignorando a interação neuroticismo:extroversão; e a interação neuroticismo:extroversão após todos os outros termos. Mais geralmente, cada termo é testado ignorando termos aos quais é marginal (ou seja, ignorando seus parentes de ordem superior). Esta é geralmente uma abordagem sensata e é o padrão na caixa de diálogo.

  Os testes Tipo II para a regressão logística de Cowles e Davis são mostrados na Figura B.4. Para um modelo linear generalizado, como a regressão logística de Cowles e Davis, o RCommander calcula uma tabela de análise de desvio com testes de razão de probabilidade, o que implica refazer o modelo (em alguns casos, duas vezes) para cada teste. Neste exemplo, cada teste qui-quadrado de razão de probabilidade tem um grau de liberdade porque cada termo no modelo é representado por um único coeficiente. Como a interação neuroticismo:extroversão é altamente significativa estatisticamente, não se interpreta os testes para os efeitos principais de neuroticismo e extroversão, que assumem que a interação é nula. O efeito do sexo também é estatisticamente significativo.

- Os testes Tipo III são para cada termo, depois de todos os outros. Isso não é sensato para o modelo de Cowles e Davis, embora possa se tornar sensato, por exemplo, centralizando cada neuroticismo e extroversão em seus meios antes de se ajustar ao modelo. Mesmo quando os testes Tipo III podem corresponder a hipóteses sensatas, como uma análise de modelos de variância, eles exigem uma formulação cuidadosa no R. Sugere-se que se evite os testes Tipo III, a menos que se saiba o que está fazendo.

B.3 Testes Comparando Modelos

O RCommander permite que você calcular o teste F de razão de probabilidade ou teste qui-quadrado para dois modelos de regressão, um dos quais está aninhado dentro do outro. Para ilustrar, retornemos ao conjunto de dados de Duncan e ajusta-se uma versão da regressão de Duncan que define os coeficientes de *education* e *income* iguais uns aos outros, especificando a fórmula do modelo linear $prestige \sim I(income + education)$. Isso faz sentido parcialmente porque ambas as variáveis explicativas são percentuais – respectivamente de concluintes do ensino médio e de pessoas de alta renda. Usando oa caixa de diálogo para Modelos Lineares resultando no modelo *LinearModel.14* que está resumido na Figura B.5.

```
Rcmdr>  LinearModel.13 <- lm(prestige ~ I(education + income), data=Duncan)

Rcmdr>  summary(LinearModel.13)

Call:
lm(formula = prestige ~ I(education + income), data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-29.605  -6.708   0.124   6.979  33.289

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -6.06319    4.22540  -1.435    0.159
I(education + income)  0.56927    0.03958  14.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.22 on 43 degrees of freedom
Multiple R-squared:  0.8279,        Adjusted R-squared:  0.8239
F-statistic: 206.8 on 1 and 43 DF,  p-value: < 2.2e-16
```

Figura B.5: Saída para o modelo linear *prestige* $\sim$ *I(education + income)* ajustado para o banco de dados *Duncan*

Selecione:

Modelos → Testes de Hipóteses → Comparar Dois Modelos

Isso leva ao diálogo na Figura B.6. Selecione o RegModel.1 mais geral como o primeiro modelo e o mais específico, restrito, LinearModel.8 como o segundo modelo, mas a ordem das seleções não é importante - o mesmo teste F é produzido em ambos os casos. Clique em *OK* na caixa de diálogo resulta na saída na Figura B.7. A hipótese de coeficientes iguais para *education* e *income* é plausível, p = 0,79 – afinal, os dois coeficientes são bastante semelhantes na regressão original (Figura 2.2), $\beta_{education} = 0,55$ e $\beta_{income} = 0,60$.

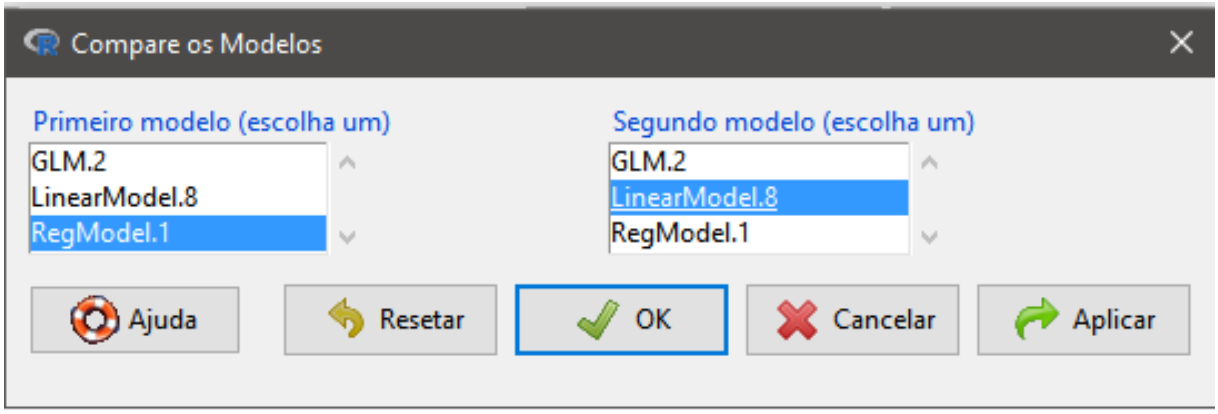

Figura B.6: Seleção dos dois modelos a serem comparados

```
> anova(RegModel.1, LinearModel.13)
Analysis of Variance Table

Model 1: prestige ~ education + income
Model 2: prestige ~ I(education + income)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     42 7506.7
2     43 7518.9 -1   -12.195 0.0682 0.7952
```

Figura B.7: Saída da comparação dos modelos *Reg.Model1* e *LinearModel.13*

B.4 Testando Hipótese Linear

Ao selecionar o menu:

Modelos → Testes de Hipóteses → Hipóteses Lineares

permite-se formular e testar hipóteses lineares gerais sobre os coeficientes em um modelo de regressão. Para ilustrar, será usada novamente a regressão de prestígio ocupacional de Duncan, RegModel.1. As Figuras B.8 e B.10 mostram a caixa de diálogo Teste de Hipótese Linear configurada para testar duas hipóteses diferentes, e as junto c Figuras B.9 e B.11, com as suas saídas respectivamente

1. Nas Figuras B.8 e B.9 , $H_0 = 1 * \beta_{education} - 1 * \beta_{income} = 0$ (i.e., $H_0 : \beta_{education} = \beta_{income}$). Esta é a mesma hipótese testada pela abordagem de comparação de modelos imediatamente acima e, é claro, produz o mesmo teste F.

2. Nas Figuras B.10 e B.11, $H_0 = 1 * \beta_{education} = 0$, $1 * \beta_{income} = 0$, que é equivalente e, portanto, produz o mesmo teste F que a hipótese nula de *omnibus* na saída de resumo do modelo linear, $H_0 : \beta_{education} = \beta_{income} = 0$ (ver Figura 2.2). Como a hipótese linear consiste em duas equações, a estatística F para a hipótese tem dois df no numerador.

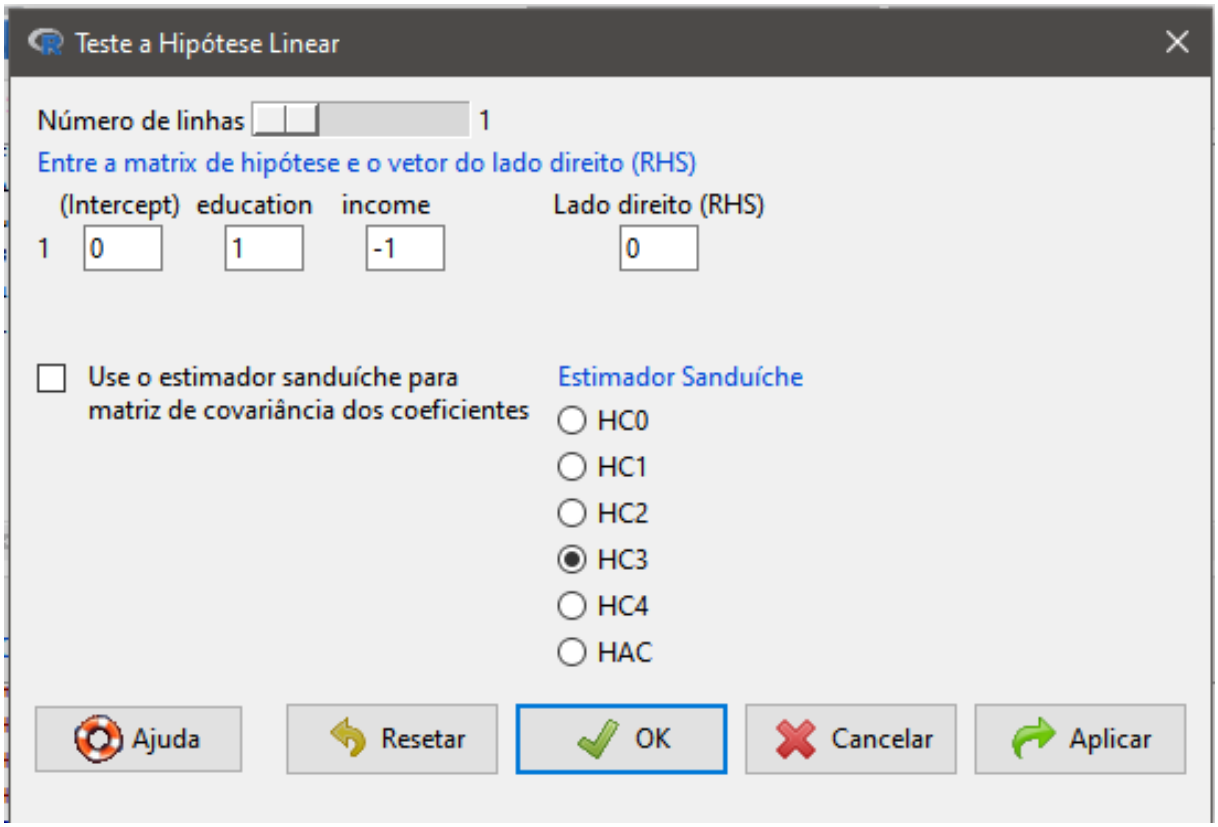

Figura B.8: Testando a hipótese nula de $H_0 : \beta_{education} = \beta_{income}$

```
    > local({
+   .Hypothesis <- matrix(c(0,1,-1), 1, 3, byrow=TRUE)
+   .RHS <- c(0)
+   linearHypothesis(RegModel.1, .Hypothesis, rhs=.RHS)
+ })
Linear hypothesis test

Hypothesis:
education - income = 0

Model 1: restricted model
Model 2: prestige ~ education + income

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     43 7518.9
2     42 7506.7  1    12.195 0.0682 0.7952
```

Figura B.9: Saída do Teste de Hipótese da Figura B.8

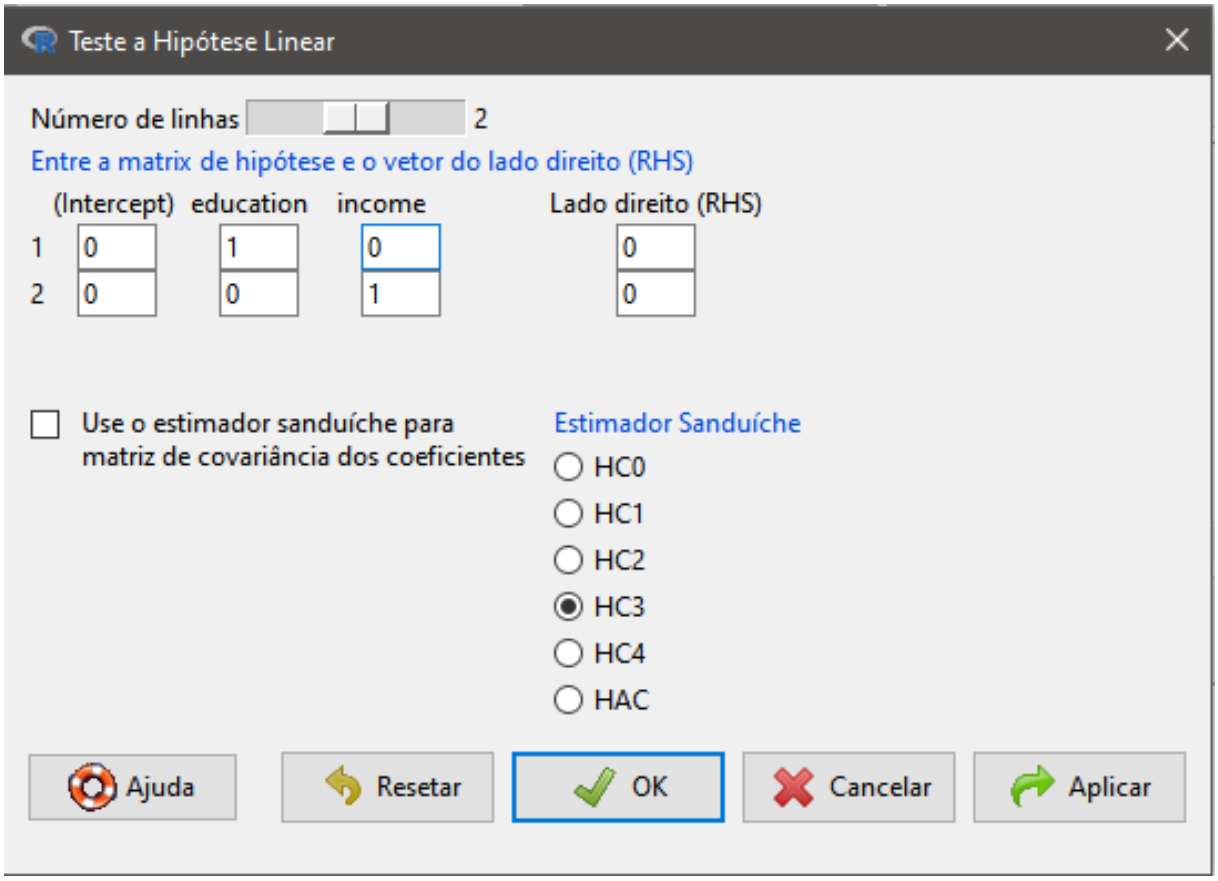

Figura B.10: Testando a hipótese nula de $H_0 : \beta_{education} = \beta_{income} = 0$

```
> local({
+   .Hypothesis <- matrix(c(0,1,0,0,0,1), 2, 3, byrow=TRUE)
+   .RHS <- c(0,0)
+   linearHypothesis(RegModel.1, .Hypothesis, rhs=.RHS)
+ })
Linear hypothesis test

Hypothesis:
education = 0
income = 0

Model 1: restricted model
Model 2: prestige ~ education + income

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     44 43688
2     42  7507  2     36181 101.22 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura B.11: Saída do Teste de Hipótese da Figura B.10

Pode haver tantas equações em uma hipótese linear quanto o número de coeficientes no modelo, com o número de equações controlado pelo controle deslizante na parte superior da

caixa de diálogo. As equações devem ser linearmente independentes umas das outras – ou seja, não podem ser redundantes. Inicialmente, todas as células em cada linha são 0, incluindo a célula que representa o lado direito da hipótese, que geralmente fica em 0. A caixa de diálogo Teste de Hipótese Linear para um modelo linear fornece um estimador "sanduíche" opcional da matriz de covariância de coeficiente, que pode ser usada para ajustar a inferência estatística para erros autocorrelacionados ou heterocedásticos (variância de erro não constante).

Apêndice C

Escolha de Modelo

O menu Modelos do RCommander inclui recursos modestos para comparar modelos de regressão e para seleção automática de modelo. Seguindo os itens:

Modelos → Critério de Informação Akaike (AIC)

e

Modelos → Critério de Informação Bayesiano (BIC)

As estatísticas de seleção de modelo AIC ou BIC para o modelo estatístico atual se apresentam. Além disso:

Modelos → Seleção de Modelo Stepwise

Essa sequência acima, por outro lado, executa a regressão passo a passo para um modelo linear ou linear generalizado, enquanto:

Modelos → Seleção do Modelo Subset

executa a regressão de todos os subconjuntos para um modelo linear.

De forma ilustrativa à seleção do modelo com o conjunto de dados *Ericksen* no pacote *car*. Os dados, descritos por Ericksen et. al (1980), refere-se à subcontagem do Censo dos EUA de 1980 e refere-se a 16 grandes cidades nos Estados Unidos, as partes restantes dos estados aos quais essas cidades pertencem (por exemplo, Estado de Nova York fora da cidade de Nova York) e os outros estados, portanto, 66 casos ao todo. Além do percentual estimado de subcontagem em cada área, o conjunto de dados contém uma variedade de características das áreas. Uma regressão linear de mínimos quadrados de *uncercount* nas outras variáveis no conjunto de dados revela que alguns dos preditores são substancialmente colineares. Ajuste o modelo linear inicial com a fórmula *undercount* $\sim$. e obtive fatores de inflação de variância para os coeficientes de regressão via:

Modelos → Diagnóstico Numérico → Fatores de inflação da variância

O resultado relevante está na Figura: C.2.

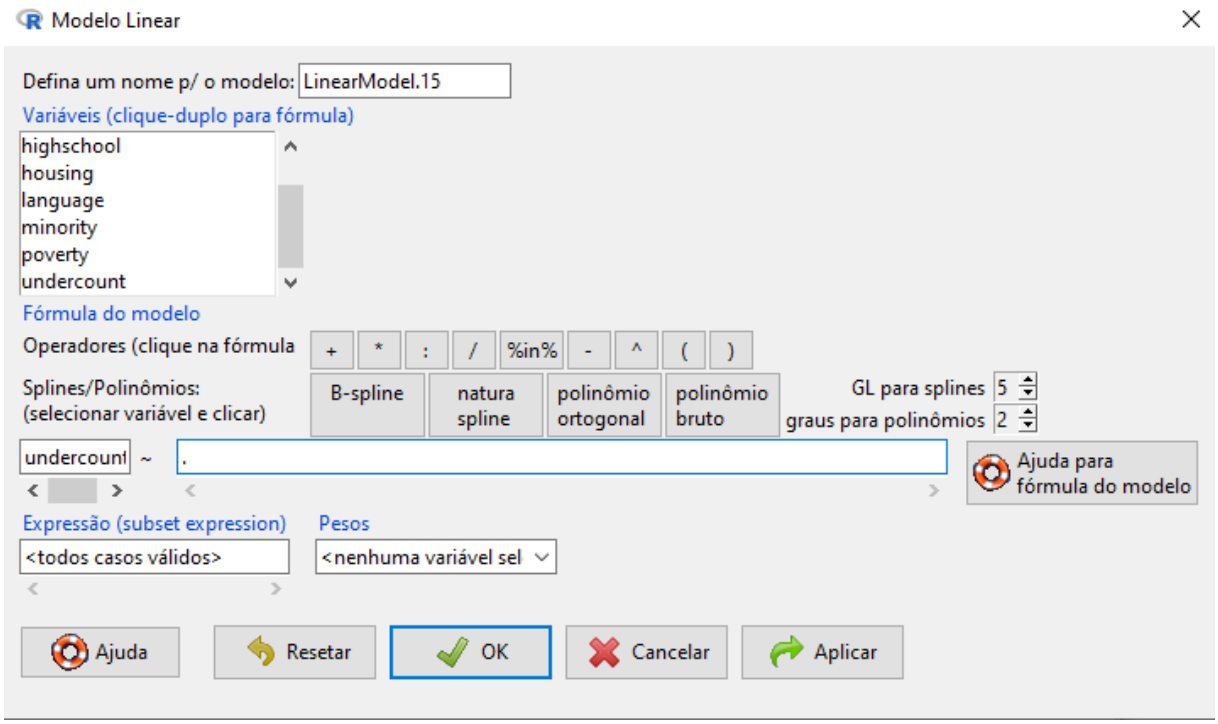

Figura C.1: Ajuste do modelo linear *undercut* $\sim$. ajustada para o banco de dados *Ericksen*

```
Rcmdr>  LinearModel.15 <- lm(undercount ~ ., data=Ericksen)

Rcmdr>  summary(LinearModel.15)

Call:
lm(formula = undercount ~ ., data = Ericksen)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8356 -0.8033 -0.0553  0.7050  4.2467

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.611411   1.720843  -0.355 0.723678
minority       0.079834   0.022609   3.531 0.000827 ***
crime          0.030117   0.012998   2.317 0.024115 *
poverty       -0.178369   0.084916  -2.101 0.040117 *
language       0.215125   0.092209   2.333 0.023200 *
highschool     0.061290   0.044775   1.369 0.176415
housing       -0.034957   0.024630  -1.419 0.161259
city[T.state] -1.159982   0.770644  -1.505 0.137791
conventional   0.036989   0.009253   3.997 0.000186 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.426 on 57 degrees of freedom
Multiple R-squared:  0.7077,        Adjusted R-squared:  0.6667
F-statistic: 17.25 on 8 and 57 DF,  p-value: 1.044e-12
```

Figura C.2: Saída para o modelo linear *undercut* $\sim$. ajustada para o banco de dados *Ericksen*

Ao escolher:

Modelos → Modelo de Seleção Subset

produz-se a caixa de diálogo da Figura C.3. Todas as opções nesta caixa de diálogo permanecem em seus padrões, incluindo o uso do BIC para a seleção do modelo. Pressionar o botão *OK* produz o gráfico na parte inferior da Figura C.4, traçando o "melhor" modelo de cada tamanho $k = 1, \ldots, 9$ de acordo com o BIC. Os preditores incluídos em cada modelo são representados por quadrados preenchidos e valores menores (ou seja, negativos de maior magnitude) do BIC representam modelos "melhores". Observe que a interceptação da regressão está incluída em todos os modelos. O melhor modelo geral, de acordo com o BIC, inclui os quatro preditores *minority*, *crime*, *language* e *conventional*.

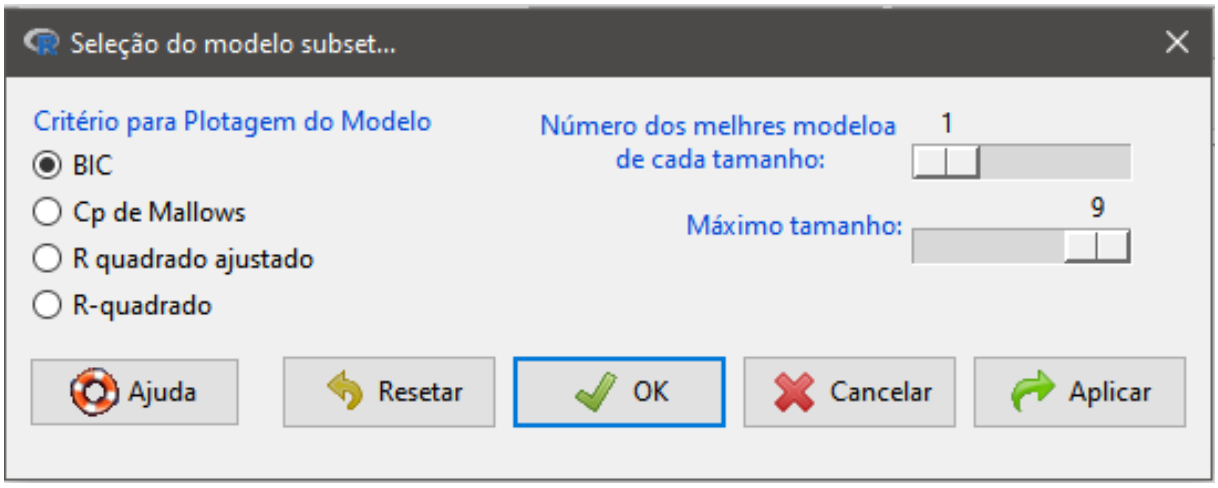

Figura C.3: Caixa de Diálogo para Teste BIC de Escolha de Modelo

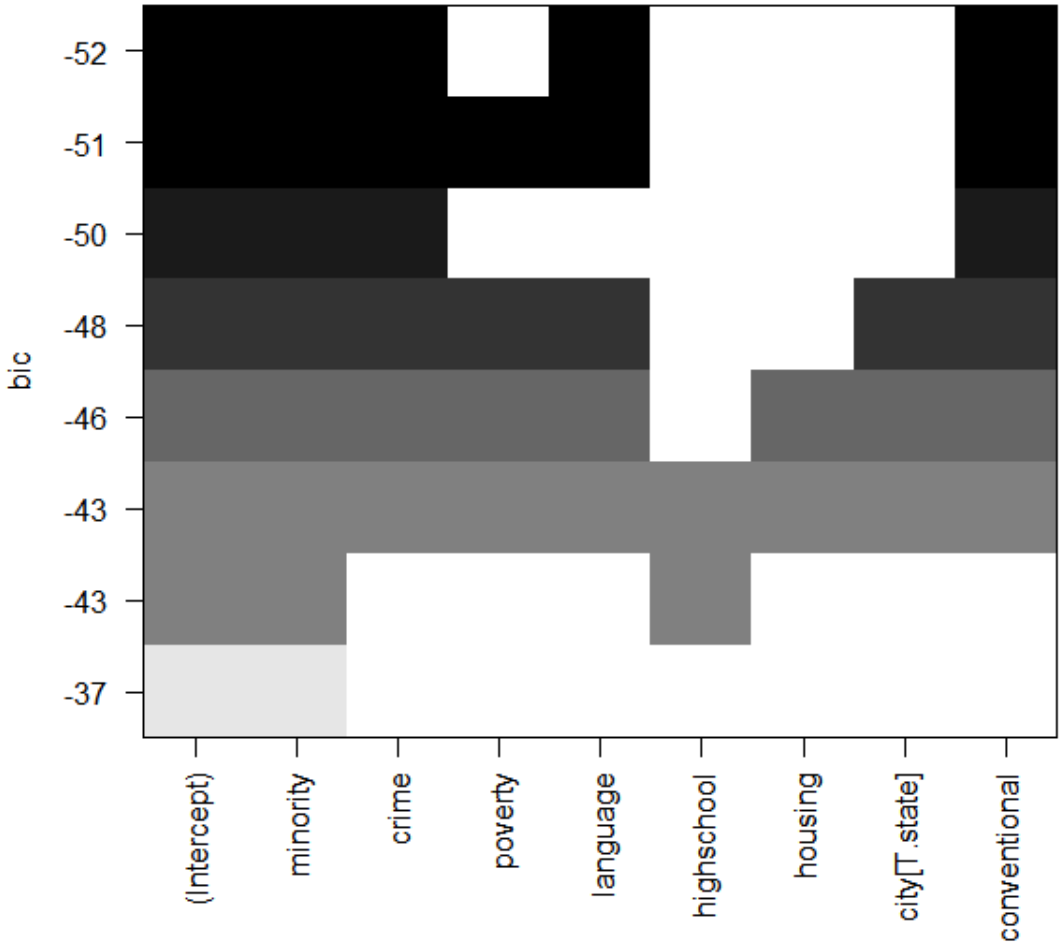

Figura C.4: Saída da Caixa de Diálogo da Figura C.3

Bibliografia

[1] Alan Agresti. *An Introduction to Categorical Data Analysis*. 2°, p. 372.

[2] Gian Marco. Campagnolo. *Social Data Science Xennials. Between Analogue and Digital Social Research*, p. 106.

[3] Nilma Margarida de Castro Crusoé e Edmilson Menezes. Santos. "ALFRED SCHUTZ'S SOCIOLOGICAL PHENOMENOLOGY: CONTRIBUTIONS TO QUALITATIVE RESEARCH IN EDUCATIONAL PRACTICE". Em: *Tempos e Espaços em Educação* (2020), p. 15.

[4] Valdeir Del Cont. "Francis Galton: eugenia e hereditariedade". Em: *scientiæ zudia* (2008), p.201–218.

[5] Philippa. Foot. "The Problem of Abortion and the Doctrine of the Double Effect". Em: *Oxford review* (1967), p. 5.

[6] Noah J. Goodall. "Machine Ethics and Automated Vehicles". Em: *Virginia Center for Transportation Innovation and Research* (2014), p. 93–102.

[7] T.J. Hastie J. M. Chambers. *Statistical Models in S*. No. 1. 1992, p. 624.

[8] Fox John. *Using the R Commander. A Point-and-Click Interface for R*. 2017, p. 223.

[9] J. A. Nelder. *A Reformulation of Linear Models*. No. 1. Vol. 140. 1977, pp. 48–77.

[10] J. A. Nelder e R. W. M. Wedderburn. *Generalized Linear Models*. No. 3. Vol. 135. 1972, pp. 370–384.

[11] van der Aalst W. *Data Science in Action*. 2016, p.3–23.

[12] G. N. Wilkinson e C. E. Rogers. *Symbolic Description of Factorial Models for Analysis of Variance*. No. 3. Vol. 22. 1973, pp. 392–399.