ANÁLISE DE REGRESSÃO

TÉCNICAS DE MODELAGEM FLORESTAL

João L. F. Batista

Departamento de Ciências Florestais

UNIVERSIDADE DE SÃO PAULO

Escola Superior de Agricultura “Luiz de Queiroz”

Piracicaba

Análise de Regressão: Técnicas de Modelagem Florestal

Departamento de Ciências Florestais
Escola Superior de Agricultura “Luiz de Queiroz”
Universidade de São Paulo
Av. Pádua Dias, 11
Caixa Postal 9
13418-900, Piracicaba - SP

Email: parsival@usp.br

“Twice two equals four: ’tis true,
But too empty, too trite.
What I look for is a clue
To some matters not so light.”

W. Busch, 1909

1 MODELOS E REGRESSÃO LINEAR

Modelos são as unidades básicas do desenvolvimento científico e tecnológico. Qualquer teoria científica pode ser vista como um *modelo conceitual* onde a realizadade é apresentada de forma simplificada através de conceitos abstratos. *Modelos quantitativos* são modelos que utilizam grandezas numéricas e funções matemáticas para representar os conceitos e suas inter-relações.

As atividades práticas da Engenharia Florestal são povoadas por modelos quantitativos. Tanto na pesquisa florestal quanto no manejo de recursos florestais, os modelos biométricos florestais constituem uma ferramenta básica e essencial. A técnica mais utilizada para se construir os modelos biométricos florestais é a Regressão Linear. A Regressão Linear é uma técnica estatística que permite construir um modelo onde uma *variável resposta*, geralmente denotada pela letra Y, é "explicada" em termos de uma ou mais variáveis preditoras que em geral são representadas pela letra X (denotadas por X_1, X_2, etc.). O termo "explicada" tem uma conotação específica no jargão estatístico e veremos o seu significado mais adiante.

1.1 O que são Modelos?

Modelos são representações simplificadas da realidade. Tais representações estão presentes no dia-a-dia de qualquer ser humano, na maioria das vêzes de forma inconsciente. Com efeito, a própria idéia que cada um de nós tem de seu próprio corpo é um modelo, pois nenhum ser humano possui conhecimento perfeito de seu organismo. Alguém conhece todas as células de seu corpo? Ou é capaz de saber as causas de qualquer doença que o aflige sem auxílio da medicina? Em geral, pessoas adultas têm uma razoável noção de como seu corpo reage em situações particulares, mas esta noção é limitada e frequentemente distorcida. O conhecimento imperfeito que temos de nosso próprio corpo pode ser chamado de modelo, pois se trata antes de tudo de uma representação mental do nosso corpo.

O organismo de qualquer pessoa é muito mais complexo do que a imagem que a

própria pessoa tem dele. Se o conhecimento que temos de nosso próprio corpo é tão limitado, o que podemos pensar sobre o conhecimento do mundo que nos circunda? Na verdade criamos representações mentais (modelos) não só do nosso organismo, mas de toda a realidade que nos envolve. A atividade de modelar, isto é, de construir representações mentais, é própria do ser humano, acontecendo tanto no plano consciente quanto nos planos subconsicentes ou inconscientes da mente.

Mas o que difere tais modelos que todas as pessoas constroem dos modelos biométricos florestais? Quais são as características desejáveis de um modelo a ser utilizado na prática florestal? Espera-se que um modelo, em sendo uma simplificação da realidade, mantenha as características fundamentais do fenômeno ou realidade que representa. Nesta visão, um modelo seria um representação imperfeita mas relativamente fiel da verdade. Na Engenharia Florestal, modelos são utilizados para auxiliar a compreensão dos fenômenos estudados e para auxiliar na tomada de decisões. Desta forma, espera-se que os modelos sejam ferramentas úteis à prática florestal.

Modelos biométricos florestais são modelos quantitativos, que representam as grandezas medidas em árvores e florestas e as suas inter-relações com o ambiente físico, biótico e humano. As grandezas utilizadas nos modelos florestais são informações quantitativas ou qualitativas obtidas através de mensuração da floresta, como por exemplo o diâmetro e a altura de árvores, ou a área basal e diversidade de espécies de uma floresta. Os modelos biométricos florestais são, portanto, alimentados por informações obtidas em campo ou em laboratório.

As inter-relações entre as grandezas são representadas por expressões matemáticas cuja a forma funcional implica num modo específico e quantitativo de relacionamento. Por exemplo, ao dizer que o volume de madeira numa floresta *varia* com a área basal estamos fazendo uma afirmação genérica não-quantitativa. Por outro lado, se dissermos que o volume de madeira numa floresta *aumenta linearmente* com a área basal estamos construindo um modelo biométrico. A diferença está no fato que existem inúmeras maneiras de expressar matematicamente a afirmação "*variar*", mas somente uma única expressão matemática pode representar o termo "*aumentar linearmente*".

Exercícios

1.1.1 Construa esquemas gráficos onde a grandeza Y é função da grandeza X, sendo que a relação entre elas é:

a) Y aumenta linearmente com X;
b) Y descresce linearmente com X;
c) Y é diretamente proporcional a X;
d) Y é inversamente proporcional a X;
e) Y tem uma relação parabólica com X.

1.1.2 Procure listar as características fundamentais que um modelo biométrico deveria conter nos seguintes casos:

a) Manejo para produção de madeira de uma floresta nativa.
b) Manejo para produção de madeira de uma floresta plantada de *Pinus sp*.
c) Manejo para conservação de uma floresta nativa.
d) Manejo de florestas nativas ou plantadas para a proteção de mananciais.

1.1.3 Tente relacionar os conceitos abaixo em termos de uma expressão matemática que represente a relação entre eles, onde a primeira grandeza é função da segunda.

a)	Altura de árvores individuais	Diâmetro das árvores (DAP)
b)	Altura média das árvores do povoamento	Fertilidade do solo
c)	Diversidade de espécies arbóreas na floresta	Precipitação anual e temperatura (clima)
d)	Volume de madeira de árvores individuais	Idade das árvores
e)	Taxa de crescimento em biomassa	Idade do povoamento

1.2 Modelos Estatísticos

De forma genérica, um modelo estatístico pode ser definido pelo seguinte esquema:

DADOS = MODELO + ERRO

Os DADOS são as informações obtidas de levantamentos de campo que representam as grandezas medidas, as quais desejamos relacionar quantitativamente. Os DADOS são sempre complexos e de difícil interpretação e manipulação. Eles podem ser constituidos por uma única variável medida em cada observação, como por exemplo altura das árvores, ou por um conjunto com diversas variáveis , por exemplo quando se mede para cada árvore a sua altura, diâmetro, biomassa de tronco, biomassa de folhas, forma do tronco, etc.

O termo MODELO na expressão acima representa uma função matemática que descreve o comportamento dos DADOS. A função matemática estabelece uma *relação funcional* entre as grandezas que se pretende modelar e deve ser fruto de um

conhecimento científico sobre o comportamento destas grandezas, sendo uma explicação teórica para o uso do modelo. Como todo MODELO é uma representação simplificada da realidade, sempre existe uma discrepância entre o MODELO e os DADOS. Esta discrepância é chamada de ERRO.

Note que o ERRO não significa que alguém cometeu algum engano durante o processo de mensuração ou na análise dos dados. O ERRO a que nos referimos é unicamente a diferença que sempre existirá entre os DADOS e o MODELO. Construir um modelo estatístico significa obter um MODELO que seja uma representação adequada dos DADOS isto é, que tenha um pequeno ERRO. No jargão estatístico, construir um modelo é "ajustar" o MODELO aos DADOS.

1.2.1 População versus Amostra

Um modelo estatístico, como simplificação da realidade, pretende representar um objeto de estudo que frequentemente não pode ser observado em seu todo. O objeto de estudo é chamado de *população* e dever ser precisamente definido antes do início da coleta dos dados e modelagem.

Os dados obtidos em campo são, em geral, uma *amostra* da população de interesse e, portanto, são apenas uma fração dos dados passíveis de coleta na população. Para ajustar o modelo estatístico, se utiliza os dados da amostra, mas pretende-se que o modelo construido seja uma boa representação da população. Para deixar mais claro estes fundamentos, vejamos um exemplo.

Exemplo:
Altura de Árvores de *Eucalyptus grandis*

População versus Amostra

A área de estudo é uma fazenda florestal com 1500 ha, com povoamentos de *Eucalyptus grandis* em 1ª rotação e idade variando de 2.1 a 14.4 anos na região de Bofete (Estado de São Paulo). O plantio foi realizado num espaçamento de plantio de 3×2 m com taxa de sobrevivência de 95%.

População: é o conjunto das alturas de todas as árvores da fazenda, isto é, aproximadamente 2,5 milhões de árvores.

Amostra: foram medidas as altura de 213 árvores da fazenda:

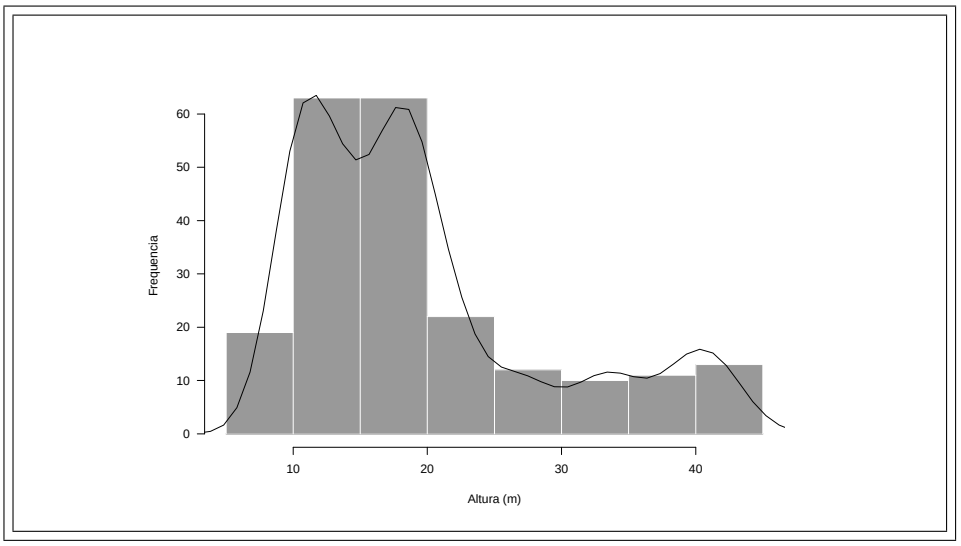
10.96	9.38	10.44	10.20	11.08	10.51	14.24	9.81	13.07	12.48	14.19
12.53	15.59	15.79	29.37	32.23	10.10	9.57	10.37	8.65	10.23	9.91
10.99	13.44	12.96	13.44	12.17	11.53	12.71	14.56	15.41	19.21	9.95
13.64	11.88	16.87	16.81	18.42	22.44	21.40	22.46	18.16	20.93	24.24
27.78	26.48	29.59	26.92	29.72	9.53	14.23	17.01	17.34	15.37	18.28
21.49	21.27	17.96	18.83	19.33	21.62	21.21	25.51	23.49	26.32	23.24
21.74	25.68	26.20	27.56	21.21	18.57	23.97	22.87	32.50	35.27	34.80
28.23	33.83	36.94	40.87	40.14	42.58	33.78	32.62	35.47	38.03	40.49
42.31	34.85	39.72	41.48	39.40	42.42	41.16	43.42	44.91	31.54	32.57
36.46	32.91	39.07	41.85	38.96	38.82	40.02	38.20	41.80	9.76	13.08
13.00	13.97	15.90	16.72	15.32	16.40	17.58	15.54	16.92	16.73	16.85
16.28	17.06	17.35	17.38	19.41	18.52	19.75	17.52	18.14	18.02	19.16
19.48	19.32	19.62	20.45	19.48	19.35	18.95	20.03	19.74	20.87	21.68
22.59	16.35	17.73	17.22	16.70	16.98	15.05	14.39	10.15	12.53	17.22
18.76	18.66	19.03	17.45	18.23	18.66	19.46	19.63	21.05	18.73	18.94
18.78	18.60	18.01	21.43	9.63	9.32	9.01	8.65	10.15	11.69	10.63
12.41	12.53	11.87	10.75	12.43	11.12	10.56	11.70	9.99	11.32	13.02
9.54	7.11	8.31	7.43	9.91	8.98	10.43	10.40	11.63	10.81	11.28
13.12	10.60	11.81	10.87	11.56	10.97	12.70	12.93	10.70	13.53	13.83
14.37	14.37	14.40	14.88							

Dado o tamanho da amostra, fica difícil visualizar o comportamento da altura das árvores, mas o gráfico abaixo mostra que existe uma grande variação e a distribuição não é simétrica.

Exemplo:
Altura de Árvores de *Eucalyptus grandis*

População versus Amostra

(cont.)



1.2.2 Construindo um Modelo Univariado Simples

Para ficar mais claro a estrutura dos modelos estatísticos, construiremos um modelo simples para o exemplo acima. Os dados disponíveis apresentam uma única variável: altura (dados univariados). No modelo mais simples possível, os dados de altura serão representados por uma constante. No caso da *população*, o modelo estatístico pode ser apresentado na seguinte forma:

$$Y_i \quad = \quad \beta_0 + \varepsilon_i \tag{1.1}$$

onde:

Y_i representa a altura da árvore i da fazenda (DADOS).

$i\,(= 1, 2, \ldots, N)$ é um índice que representa cada uma das árvores na fazenda. No exemplo acima $N \approx 2500000$ árvores.

β_0 é uma constante que é o modelo matemático para a altura de todas as árvores da fazenda (MODELO). β_0 é chamado de *parâmetro* pois é uma constante (desconhecida) que se refere à população.

ε_i é o ERRO, isto é, a diferença entre a constante β_0 (MODELO) e a altura observada Y_i (DADOS) para árvore i. Note que ε_i também se refere às árvores da fazenda (população).

β_0 e ε_i são relacionados no sentido que um só é conhecido se o outro for conhecido também. Como ambos se referem às alturas das árvores da fazenda e, não só da amostra, ambos serão sempre desconhecidos. No entanto, este é o modelo hipotético para toda a população.

Para ajustar este modelo aos dados, precisamos apresentá-lo quando somente os dados da *amostra* forem utilizados. Neste caso ele se torna:

$$Y_i = b_0 + e_i$$

onde:

Y_i $(i = 1, 2, \ldots, n)$ é a altura da árvore i da amostra. No exemplo acima, o tamanho da amostra (n) é 213 árvores.

b_0 é um candidato a tomar o lugar de β_0, isto é, a ser a nossa "melhor" estimativa do parâmetro do modelo. Como o nosso modelo é composto de apenas um parâmetro, b_0 é também a nossa "melhor" estimativa para altura das árvores.

e_i é chamado de *resíduo* pois é o que sobra ou falta quando a nossa estimativa é comparada com a altura das árvores da amostra.

Em estatística é comum utilizar uma notação especial para representar a *estimativa* de uma variável observada. Nesta notação, coloca-se o acento circunflexo (ˆ) sobre a letra que representa a variável. No nosso caso temos:

Y_i altura *observada* da árvore i;

$\widehat{Y}_i$ altura *estimada* da árvore i.

O modelo simples que estamos construindo implica que:

$$\widehat{Y}_i = b_0$$

ou seja, a nossa estimativa da altura será a mesma para todas as árvores da fazenda. A constante b_0 será encontrada com base nas alturas das árvores da *amostra* (213 árovres), mas será aplicada a todas as árvores da *população* (todas 2,5 milhões de árvores da fazenda). Como na amostra, o resíduo é a diferença entre a altura observada e a altura estimada pelo modelo, temos que:

$$\begin{aligned} e_i &= Y_i - \widehat{Y}_i \\ e_i &= Y_i - b_0 \end{aligned}$$

Note que o fato de subtrairmos *sempre* o observado do estimado, nesta ordem, implica que:

- resíduo positivo indica *subestimativa*, e
- resíduo negativo indica *superestimativa.*

1.2.3 Critérios para Ajuste de Modelos Estatísticos

Ao encontrarmos um valor numérico para b_0, estaremos ajustado o MODELO (β_0) aos DADOS (Y_i). Um bom ajuste deverá produzir um ERRO pequeno quando aplicado à população. Para encontrarmos b_0 devemos ser mais explícitos sobre o que consideramos como "a nossa melhor estimativa" e o que é "produzir um ERRO pequeno". Há vários critérios que podemos utilizar para medir a discrepância entre os DADOS e o MODELO. Vejamos alguns:

Contagem dos Resíduos (CR): neste critério contaríamos os resíduos (e_i) que fossem diferentes de zero. Formalmente, este critério pode ser representado pela função:

$$CR = \sum_{i=1}^{n} I(e_i \neq 0) = \sum_{i=1}^{n} I(Y_i - \widehat{Y}_i \neq 0) = \sum_{i=1}^{n} I(Y_i - b_0 \neq 0)$$

onde $I(\cdot)$ é uma *função indicadora* que assume o valor 1 se a condição dentro de parênteses for verdadeira e o valor 0 (zero) se for falsa. Na verdade, $I(e_i \neq 0)$ é uma maneira sofisticada de dizer que estamos contando os resíduos cujos valores diferem de zero. Este critério tem o problema de ignorar a magnitude de cada resíduo, assim, resíduos grandes e pequenos teriam a mesma importância ao definir o valor de b_0.

Soma dos Resíduos (SR): este critério consiste simplesmente em somar os resíduos, isto é:

$$SR = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i) = \sum_{i=1}^{n} (Y_i - b_0)$$

A soma dos resíduos tem o problema de que os resíduos positivos e negativos se anularem. Se b_0 for obtido com base neste critério, é possível que ele gere grandes resíduos positivos e grandes resíduos negativos, o que gostaríamos de evitar.

Soma dos Resíduos Absolutos (SRA): a alternativa natural para a soma dos resíduos é ignorarmos o sinal do resíduo:

$$SRA = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |Y_i - \widehat{Y}_i| = \sum_{i=1}^{n} |Y_i - b_0|$$

Este critério tem a vantagem de evitar que resíduos positivos cancelem resíduos negativos. Por outro lado, há o problema de um grande resíduo ser considerado de mesmo peso que uma série de pequenos resíduos. Por exemplo, um modelo que superestime a altura de uma única árvore em 10 m seria equivalente a um modelo que superestima a altura de 10 árvores em apenas 1 m. Em termos práticos, o segundo modelo é muito superior ao primeiro.

Soma do Quadrado dos Resíduos (SQR): esta é uma outra alternativa de remover o sinal dos resíduos:

$$SQR = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0)^2$$

Este critério, além de evitar o cancelamento de resíduos devido ao sinal, dá maior importância aos resíduos maiores, evitando que vários resíduos pequenos tenham a mesma importância que um grande resíduo.

Todos os critério acima, foram apresentados na forma de uma função. Estas funções são chamadas de *funções de perda*, pois quanto maior os seus valores pior o ajuste do MODELO aos DADOS. Se encontrarmos o valor de b_0 que *minimiza* uma função de perda, isto é, que a torne o menor possível para os dados da amostra que possuímos, teremos encontrado o "melhor" valor de b_0 de acordo com o respectivo critério.

Tomemos como exemplo o critério da Soma dos Resíduos (SR). Neste caso, o menor valor desejável para a SR é zero, pois valores negativos indicariam uma tendência a superestimar (lembre-se que $e_i = Y_i - \widehat{Y}_i$). Qual o valor de b_0 que faria $SR = 0$?

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \widehat{Y}_i) = \sum_{i=1}^{n}(Y_i - b_0) &= 0 \\
\sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}b_0 &= 0 \\
\sum_{i=1}^{n}Y_i - nb_0 &= 0 \\
nb_0 &= \sum_{i=1}^{n}Y_i \\
b_0 &= \frac{\sum_{i=1}^{n}Y_i}{n} = \overline{Y}
\end{aligned}$$

Portanto, a média amostral ($\overline{Y}$) é o melhor valor de b_0 segundo o critério da Soma dos Resíduos. Assim, dizemos que a média amostral é o melhor *estimador* segundo a Soma dos Resíduos.

Cada critério apresentado acima terá o seu *melhor estimador* caso a função de perda seja minimizada:

Função de Perda	**Estimador de β_0**
Contagem dos Resíduos	MODA: valor mais frequente de Y_i na amostra
Soma dos Resíduos	MÉDIA: = média amostral de Y_i
Soma dos Resíduos Absolutos	MEDIANA: = valor acima de 50% das observações de Y_i na amostra
Soma de Quadrado dos Resíduos	MÉDIA: = média amostral de Y_i

Vejamos como cada um destes critérios se comportam com os dados do exemplo da altura de árvores de *Eucalyptus grandis*.

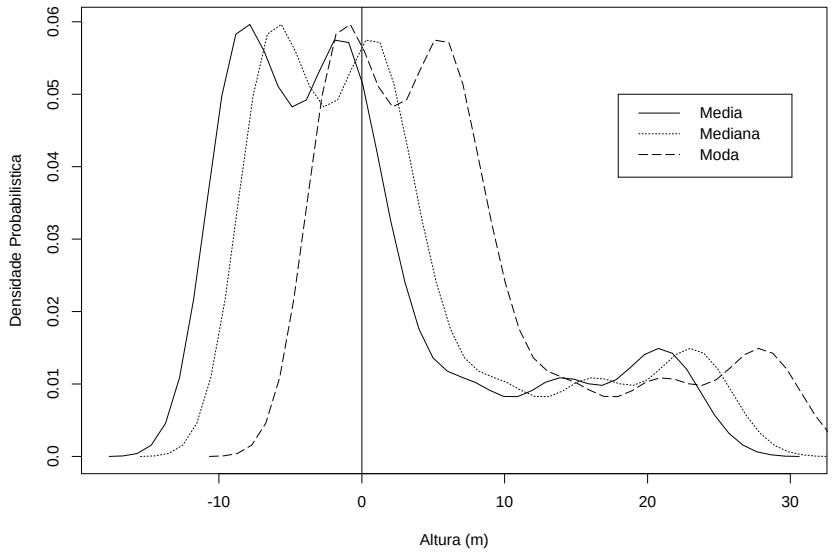
Exemplo:
Altura de Árvores de
Eucalyptus grandis

Critérios de Ajuste

Ajustando-se o modelo (1.1, pág. 6) à amostra da altura de 213 árvores de *E. grandis*, obtem-se o seguinte resultado:

Estimadores	Estimativas na Amostra	Funções de Perda CR	SR	SRA	SQR
Moda	*12.53*	**210**	1494.08	1725.94	29379.48
Mediana	*17.35*	212	467.42	**1481.00**	19925.05
Média	*19.54*	213	**0.00**	1548.00	**18899.32**

Cada critério mostrou que minimiza a sua respectiva função de perda, somente a média amostral foi capaz de minizar dois critérios. A soma de resíduos (SR) indica que a moda e a mediana tendem a gerar resíduos positivos com mais frequência, sendo que o gráfico de distribuição dos resíduos abaixo mostra claramente esta tendência.



1.2.4 O Método dos Quadrados Mínimos

O método de minimizar a Soma dos Quadrados dos Resíduos é chamado de ***Métodos dos Quadrados Mínimos*** e as estimativas obtidas por esse método são ditas ***estimativas de quadrados mínimos***. Este é o critério utilizado em regressão linear para ajustar os modelos pois é o único que satisfaz duas condições muito importantes:

Erro Médio Nulo: os estimadores de quadrados mínimos, além de minimizar a Soma dos Quadrados dos Resíduos, também tornam nula a Soma dos Resíduos. Isto implica que o "*erro médio*" destes estimadores é zero, o que significa que não há tendências de superestimar ou subestimar.

Maior Penalização de Grandes Resíduos: como neste critério os resíduos são elevados ao quadrado, grandes resíduos são fortemente penalizados. No exemplo da altura das árvores, seriam necessários 100 resíduos de 1 m para se alcançar a mesma soma de um único resíduo de 10 m. Grandes resíduos serão evitados pelo Método dos Quadrados Mínimos.

Uma vez que se tenha em mãos uma amostra, a Soma dos Quadrados dos Resíduos será sempre função dos parâmetros a serem estimados. As estimativas de quadrados mínimos serão obtidas minimizando esta função em relação aos parâmetros. A teoria do cálculo diferencial nos garante que para obtermos os pontos extremos de uma função devemos encontrar a sua primeira derivada, igualá-la a zero e solucionar a expressão resultante. A solução nos fornece o ponto extremo, se a segunda derivada da função neste ponto for positiva, este ponto extremo é um ponto de mínimo, isto é, o valor obtido igualando a primeira derivada a zero minimiza a função.

Vejamos como isto pode ser feito no caso do modelo (1.1). A Soma dos Quadrados dos Resíduos é função do estimador b_0:

$$Q(b_0) = \sum_{i=1}^{n}(Y_i - b_0)^2$$

Desenvolvendo o quadrado desta expressão obtemos:

$$\begin{aligned} Q(b_0) &= \sum_{i=1}^{n}(Y_i^2 - 2Y_i b_0 + b_0^2) \\ &= \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} 2Y_i b_0 + \sum_{i=1}^{n} b_0^2 \\ &= \sum_{i=1}^{n} Y_i^2 - 2b_0 \sum_{i=1}^{n} Y_i + nb_0^2 \end{aligned}$$

Tomando a primeira derivada em relação à b_0 e igualando-a a zero obtemos:

$$\begin{aligned} \frac{dQ}{db_0} &= -2\sum_{i=1}^{n} Y_i + 2nb_0 = 0 \\ &= -\sum_{i=1}^{n} Y_i + nb_0 = 0 \Rightarrow b_0 = \frac{\sum_{i=1}^{n} Y_i}{n} \end{aligned}$$

A fórmula obtida para b_0 é a fórmula da média amostral. Logo a função da Soma dos Quadrados dos Resíduos atinge um ponto extremo (máximo ou mínimo) quando o valor de b_0 é substituido pela média amostral.

Para termos certeza de que este ponto extremo é um ponto de mínimo, é necessário mostrar que a segunda derivada da função Q (em relação a b_0) é positiva:

$$\frac{d^2Q}{db_0^2} \quad = \quad 2n > 0$$

Portanto, podemos ter a certeza de que a média amostral minimiza a Soma dos Quadrados dos Resíduos para o modelo (1.1).

No caso do nosso modelo univariado simples, o modelo (1.1), a média amostral é o estimador de quadrados mínimos. Esta exposição justifica o porquê da média arimética ser tão frequentemente utilizada como estatística descritiva de uma amostra. Mas a média amostral não é uma panacéia e, ao adotarmos outros critérios de representação dos dados, outras estatísticas descritivas devem ser utilizadas.

Exemplo:
Altura de Árvores de *Eucalyptus grandis*

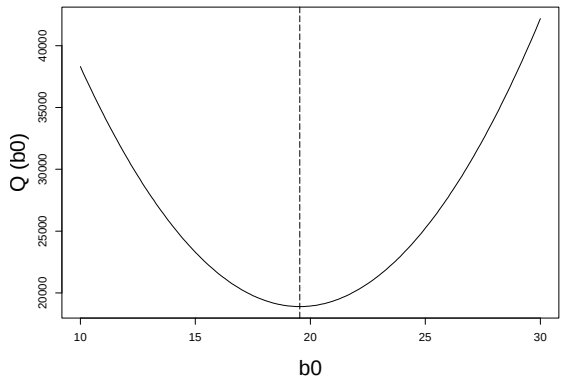
Estimador de Quadrados Mínimos

Uma forma visual de verificar que o estimador de quadrados mínimos obtido pelo método acima de fato minimiza a Soma dos Quadrados dos Resíduos (SQR) é calculá-la para valores arbitrários de b_0 construindo um gráfico.

Para amostra de árvores de *Eucalyptus grandis*, a SQR em função de b_0 fica:

$$\begin{aligned} Q(b_0) &= \sum_{i=1}^{n} Y_i^2 - 2b_0 \sum_{i=1}^{n} Y_i + nb_0^2 \\ &= (100262.3) - 2b_0(4162.97) + 213\, b_0^2 \\ &= 100262.3 - 8325.94\, b_0 + 213\, b_0^2 \end{aligned}$$

Fazendo os valores de b_0 variar entre 10 a 30, obtemos o seguinte gráfico para esta função:



Note que $Q(b_0)$ é uma função quadrática de b_0, isto é, seu gráfico é uma parábola. O ponto de mínimo está exatamente no ponto em que $b_0 = 19.54$, isto é, no ponto em que b_0 é igual à média amostral.

Exercícios

1.2.1 Os dados abaixo são os CAP de 32 árvores de palmiteiro juçara (*Euterpe edulis*) medidas numa propriedade rural no Município de Eldorado, Estado de São Paulo.

18.5	48.0	33.0	16.0	25.0	46.0	21.0	51.5
17.5	32.0	30.0	18.5	43.5	25.0	17.5	17.5
18.5	43.0	20.0	33.5	19.5	19.5	38.0	30.0
20.0	38.0	23.0	16.0	33.5	16.0	19.0	17.5

Tomando como base o modelo (1.1):

a) caracterize a *população* e a *amostra* referentes a esses dados;
b) encontre as estimativas que minimizam a Contagem dos Resíduos, a Soma dos Resíduos, a Soma dos Resíduos Absolutos e a Soma dos Quadrados dos Resíduos;
c) mostre, através de um gráfico, que a média amostral minimiza a Soma dos Quadrados dos Resíduos.

1.2.2 Os dados abaixo são as áreas (ha) de fragmentos de mata degradada na região do Vale do Ribeira, Estado de São Paulo.

4.86	4.54	0.49	3.46	0.01	5.87	0.08	2.97
1.18	2.02	3.16	78.00	4.51	8.29	4.38	2.34

Com base no modelo (1.1):

a) encontre os estimadores que minimizam a Contagem dos Resíduos, a Soma dos Resíduos Absolutos e a Soma dos Quadrados dos Resíduos;
b) calcule os resíduos produzidos por cada estimador;
c) analisando os resíduos responda as seguintes questões:
   - Quais as limitações de cada um dos estimadores?
   - Qual estimador representa melhor os dados?

1.2.3 Num levantamento da regeneração de guarantã (*Esenbekia leiostachia*) na Reserva de Ibicatu, Muncípio de Piracicaba, São Paulo, utilizou-se 40 parcelas e foram encontrados os seguintes números de plantas com altura entre 1 e 2.5 m por parcela:

1	0	0	3	0	3	0	4	2	3
3	0	0	0	2	12	7	1	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Se o modelo (1.1) fosse ajustado a esses dados, qual critério de ajuste deveria ser escolhido? Por que?

1.2.4 A altura comercial (m), isto é a altura até a 1ª bifurcação, foi medida em 30 árvores de jatobá (*Hymenea courbaril*) numa floresta no Município de Bom Jardim, Estado do Maranhão.

4	5	10	8	8	7	8	11	7	6	7	4	6	4	6
5	10	9	4	6	14	14	12	13	10	11	7	11	10	9

Qual estatística descritiva (média, mediana, moda) deveria ser utilizada para respresentar estes dados? Por que?

1.3 Regressão Linear Simples

No modelo univariado simples, construiu-se um modelo estatístico com base em uma única variável que no exemplo das árvores de *Eucalyptus grandis* foi a variável altura. Na regressão linear, no entanto, estaremos interessados em construir modelos com duas ou mais variáveis, sendo que o modelo mais simples envolve apenas duas variáveis.

1.3.1 O Modelo Linear Simples

Na estrutura geral dos modelos estatísticos:

$$\text{DADOS} = \text{MODELO} + \text{ERRO}$$

dois componentes mudam no caso do modelo linear simples quando este é comparado ao modelo univariado apresentado acima (modelo 1.1). Os DADOS não são mais observações de uma única variável, mas *observações pareadas* de duas variáveis:

variável resposta: que é a variável cujo comportamento desejamos modelar, e

variável preditora: qué é a variável que nos auxiliará a representar o comportamento da variável resposta.

O termo "*observações pareadas*" siginifica que ambas as variáveis são medidas conjuntamente nas observações uma-a-uma.

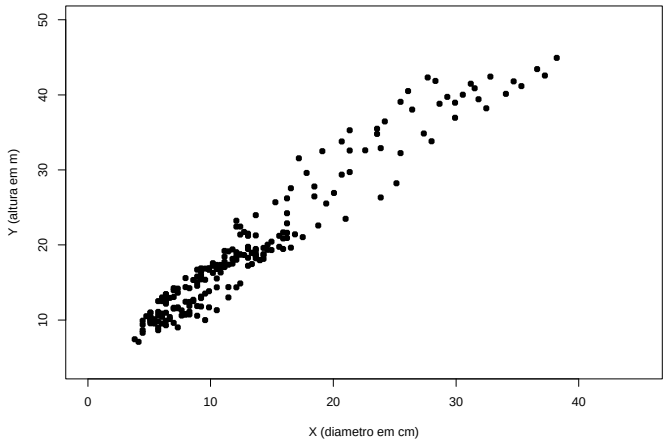
Exemplo:
Relação
Altura-Diâmetro
em Árvores de
Eucalyptus grandis

DADOS

Este exemplo ainda se refere às árvores de *Eucalyptus grandis* do exemplo anterior. Entretanto, interessa-nos agora a relação entre a altura total (m) e o diâmetro (DAP - cm) das árvores. Os DADOS, portanto, consistem de observações pareadas destas duas variáveis árvore-a-árvore:

Árvore	Diâmetro	Altura
1	5.09	10.96
2	4.46	9.38
3	5.09	10.44
4	5.09	10.20
5	5.73	11.08
6	4.77	10.51
7	7.00	14.24
8	5.73	9.81
9	7.00	13.07
10	6.37	12.48
11	7.32	14.19
12	6.05	12.53
⋮	⋮	⋮
211	12.10	14.37
212	11.46	14.40
213	12.41	14.88

A variável que desejamos modelar é a altura total das árvores (variável resposta) enquanto que o diâmetro é a variável preditora. A melhor maneira de visualizar a relação entre altura e diâmetro para construir o modelo de regressão é por meio de um gráfico de dispersão.



Por convensão, a variável resposta é sempre colocada no eixo das ordenadas (eixo-y) e a variável preditora no eixo das abcissas (eixo-x).

No modelo linear simples, a relação funcional entre variável resposta e variável preditora segue um polinômio de 1º grau, que graficamente é representado por uma reta. A expressão matemática da função linear simples é

$$y \quad = \quad \beta_0 + \beta_1 x$$

Note que utilizamos y e x (*letras minúsculas*) na expressão acima para denotar variáveis matemáticas arbitrárias.

Neste modelo matemático, o parâmetro β_0 indica o ponto em que a reta intercepta o eixo das ordenadas, ou valor de y quando $x = 0$. Já o parâmetro β_1, é a inclinação da reta, ou a alteração que ocorre em y, quando x varia em uma unidade. Este parâmetro também pode ser entendido como a razão da taxa de variação de y pela taxa de variação em x:

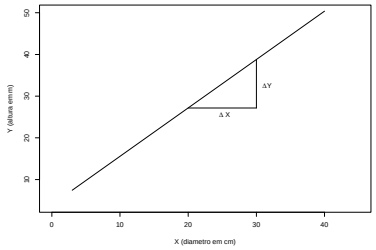
$$\left.\begin{array}{l} y_1 = \beta_0 + \beta_1 x_1 \\ y_2 = \beta_0 + \beta_1 x_2 \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{ll} y_2 - y_1 & = \beta_0 + \beta_1 x_2 - \beta_0 - \beta_1 x_1 \\ y_2 - y_1 & = \beta_1 (x_2 - x_1) \end{array}\right.$$

$$\beta_1 = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}$$

Exemplo:
Relação Altura-Diâmetro em Árvores de *Eucalyptus grandis*

MODELO

No caso da relação altura-diâmetro, o modelo linear simples sugere que a altura das árvores é diretamente proporcional ao diâmetro.



O parâmetro β_1 é a constante de proporcionalidade. Se $\beta_1 = 2$, então a altura (em metros) será o dobro do diâmetro (em centímetros).

Outra forma de entender β_1 é que uma variação de 1 cm no diâmetro resulta numa variação de β_1 m na altura. Portanto, o parâmetro β_1 *possui unidade de medida*, e esta unidade é sempre a razão da unidade da variável resposta pela unidade da variável preditora. Neste exemplo, a unidade de medida de β_1 é m/cm.

O parâmetro β_0 seria a altura de uma árvore cujo diâmetro é zero. Portanto, β_0 tem unidade de medida igual a unidade de medida da variável resposta, que neste exemplo é metro.

Não é muito realista falarmos da altura de árvores com diâmetro zero, mas é importante lembrar que a função matemática do MODELO é uma representação simplificada da realidade e, consequentemente, sempre terá limitações em explicá-la.

Combinando DADOS e o MODELO obtemos o modelo estatístico para regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1.2}$$

onde

Y_i é o valor da *variável resposta* para observação i ($i = 1, 2, \ldots, N$);

X_i é o valor da *variável preditora* para observação i;

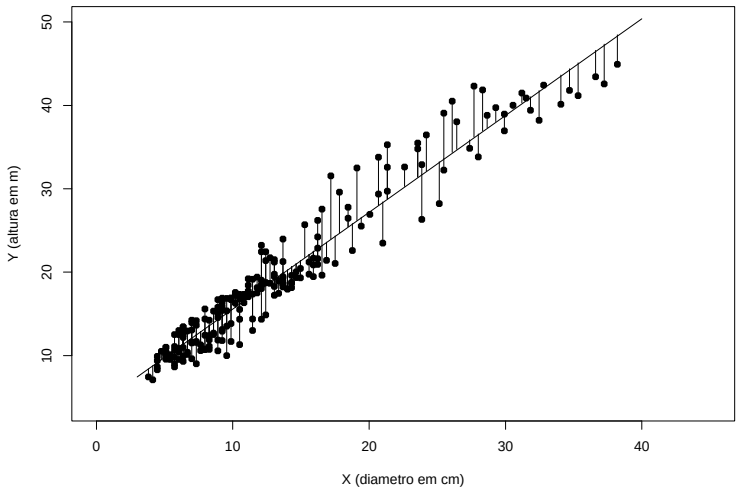
β_0 e β_1 são os *parâmetros*; e

ε_i é o erro na observação i.

Exemplo:
Relação
Altura-Diâmetro
em Árvores de
Eucalyptus grandis

Regressão
Linear Simples

No modelo de regressão haverá sempre discrepância entre a altura observada para as árvores de *Eucalyptus grandis* (Y_i) e a altura estimada pelo modelo linear simples ($\beta_0 + \beta_1 X_i$). Essa discrepância é o ERRO estatístico, que no gráfico de dispersão da altura pelo diâmetro é representado pela *distância vertical* entre a cada observação e a reta que representa a relação funcional altura-diâmetro.



1.3.2 A Função da Soma de Quadrado dos Resíduos

Para encontrarmos as estimativas dos parâmetros do modelo (β_0 e β_1) utilizaremos o método dos Quadrados Mínimos. Numa dada amostra, os resíduos do modelo linear simples são:

$$\begin{aligned} e_i &= Y_i - \widehat{Y}_i \\ &= Y_i - (b_0 + b_1 X_i) \\ &= Y_i - b_0 - b_1 X_i \end{aligned}$$

onde b_0 é a estimativa de β_0 e b_1 é a estimativa de β_1. A soma dos quadrado dos resíduos (SQR) é definida pela função:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

A função da SQR depende agora de duas variáveis: b_0 e b_1, sendo uma função quadrática de ambas. Isto é mais facilmente visualizado se desenvolvermos a

expressão:

$$\begin{aligned} Q(b_0, b_1) &= \sum_{i=1}^{n} \left[Y_i^2 - 2Y_i b_0 - 2b_1 X_i Y_i + b_0^2 + 2b_0 b_1 X_i + b_1^2 X_i^2 \right] \\ &= \sum_{i=1}^{n} Y_i^2 - 2b_0 \sum_{i=1}^{n} Y_i + n b_0^2 - 2b_1 \sum_{i=1}^{n} X_i Y_i + b_1^2 \sum_{i=1}^{n} X_i^2 + 2b_0 b_1 \sum_{i=1}^{n} X_i \end{aligned}$$

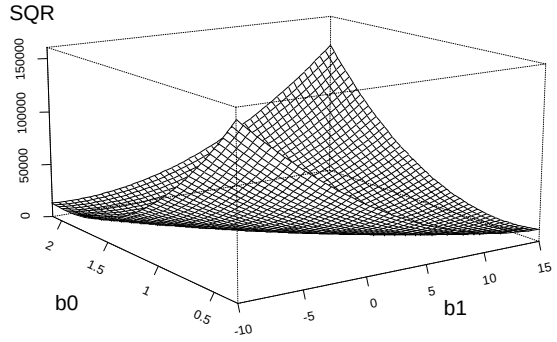
Exemplo:
Relação
Altura-Diâmetro
em Árvores de
Eucalyptus grandis

SQR

Encontrando a função da SQR para relação altura-diâmetro podemos investigar graficamente a sua forma. No caso das árvores de *Eucalyptus grandis* a função da SQR fica:

$$\begin{aligned} Q(b_0, b_1) &= 100262.3 - 8325.94\, b_0 + 213\, b_0^2 - 141736.06\, b_1 \\ &\quad + 51156.04\, b_1^2 + 5699.12\, b_0 b_1 \end{aligned}$$

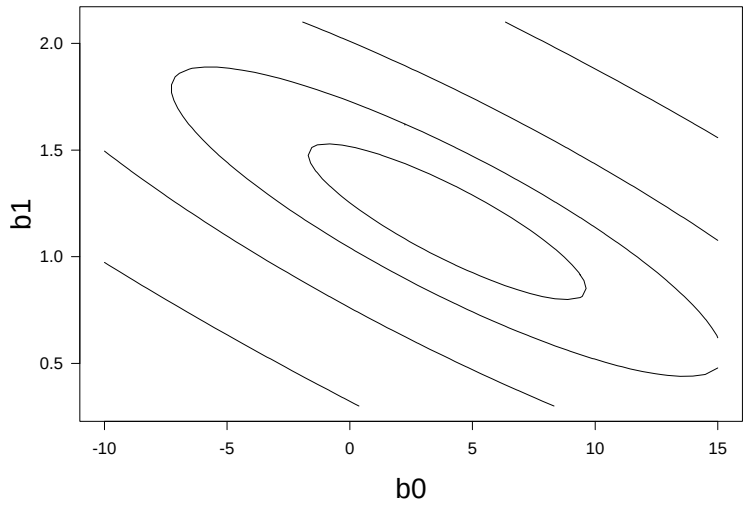
Construindo um gráfico tridimensional para esta função observamos que $Q(b_0, b_1)$ é de fato uma função quadrática, mas com curvatura que difere em relação a b_0 e b_1.



Exemplo:
Relação
Altura-Diâmetro
em Árvores de
Eucalyptus grandis

SQR (cont.)

Outra forma de visualizarmos a função da SQR é através de um gráfico de contornos. O gráfico de contornos é um gráfico bidimensional onde as linhas representam "curvas de nível" (*isolinhas*) em relação à terceira variável. No gráfico abaixo, cada linha é uma isolinha para a SQR, isto é, representa um mesmo valor de SQR.



Relembrando o gráfico tridimensional anterior, conclui-se que o ponto de mínimo da função da SQR está no centro do gráfico.

1.3.3 Estimativas de Quadrados Mínimos

Para encontrarmos o ponto de mínimo desta função devemos encontrar as *derivadas parciais* em relação a b_0 e b_1, igualando-as a zero:

$$\begin{aligned} \frac{\partial Q}{\partial b_0} &= nb_0 + b_1 \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i = 0 \\ \frac{\partial Q}{\partial b_1} &= b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i Y_i = 0 \end{aligned}$$

Note que o sistema obtido é composto de duas equações e duas incógnitas (b_0 e b_1). É importante lembrar que para uma dada amostra todos os termos que envolvem somatórias são constantes, portanto o sistema obtido consiste num sistema **linear** que é facilmente solucionado.

Re-escrevemos aqui o sistema de equações na forma que ele é mais comumente

apresentado:

$$\begin{aligned}\sum_{i=1}^{n} Y_i &= nb_0 + b_1 \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i Y_i &= b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2\end{aligned}$$

Este sistema é a chave para a regressão linear sendo chamado de sistema de *Equações Normais*. Ajustar o modelo aos dados significa encontrar a solução para este sistema. Felizmente, podemos obter uma solução geral para as estimativas dos parâmetros independentemente do conjunto de dados que estejamos analisado.

Para solucionarmos este sistema, primeiramente isolamos b_0 na primeira equação do sistema, obtendo b_0 em função de b_1:

$$\begin{aligned}b_0 &= \frac{1}{n}\left[\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i\right] \\ &= \left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) - b_1 \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\end{aligned}$$

$$\boxed{b_0 = \overline{Y} - b_1 \overline{X}}$$

A estimativa de quadrados mínimos para β_0 pode, portanto, ser interpretada como a diferença entre a média amostral da variável resposta *observada* ($\overline{Y}$) e a média amostral *predita* com base na relação de proporcionalidade com a variável preditora ($b_1\overline{X}$).

Para obtermos b_1, devemos substituir a expressão de b_0 na segunda equação do sistema de equações normais:

$$\begin{aligned}\sum_{i=1}^{n} X_i Y_i &= \left[\frac{\sum_{i=1}^{n} Y_i}{n} - b_1 \frac{\sum_{i=1}^{n} X_i}{n}\right] \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \\ \sum_{i=1}^{n} X_i Y_i &= \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} - b_1 \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} + b_1 \sum_{i=1}^{n} X_i^2 \\ b_1 \left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right] &= \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}\end{aligned}$$

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \left[\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)\right]/n}{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2/n}$$

Exemplo:
Relação
Altura-Diâmetro
em Árvores de
Eucalyptus grandis

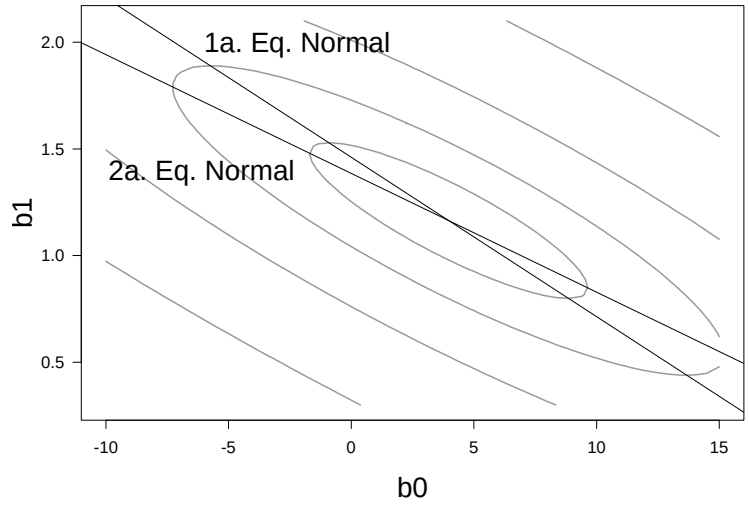
Sistema de
Eq. Normais

Vejamos inicialmente como o Sistema de Equações Normais aparece nos dados de *Eucalyptus grandis*:

$$\begin{array}{rclrcl} \sum Y_i & = & 4162.97 & \sum X_i & = & 2849.56 \\ \sum X_i Y_i & = & 70868.03 & \sum X_i^2 & = & 51156.04 \end{array}$$

$$\begin{array}{rrcl} 1^{\text{a}}\text{ Eq.Normal:} & 4162.97 = & 213\, b_0 + 2849.56\, b_1 \\ 2^{\text{a}}\text{ Eq.Normal:} & 70868.03 = & 2849.56\, b_0 + 51156.04\, b_1 \end{array}$$

O sistema de Equações Normais aparece no gráfico da superfície da SQR como duas linhas, sendo que o cruzamento das linhas indicam o ponto de mínimo da SQR:



Como b_1 é uma razão entre duas grandezas, devemos entender os termos desta razão para podermos intrepretar b_1 adequadamente e compreender como o Método de Quadrados Mínimos estima β_1.

Numerador: é chamado de *Soma de Produtos* de X por Y e pode ser apresentado da

seguinte forma:

$$S_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{[(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)]}{n} = \sum_{i=1}^{n} \left[(X_i - \overline{X})(Y_i - \overline{Y})\right].$$

S_{XY} é na verdade uma soma dos produtos dos *desvios* X e Y em relação às suas médias amostrais. Grandes valores desta soma (em termos absolutos) indicam que grandes desvios de X em relação à sua média são acompanhados de grandes desvios de Y. Por outro lado, pequenos valores (em termos absolutos) da soma indicaram um "descompasso" entre os desvios de X e Y. Portanto, S_{XY} é uma medida de como X e Y *variam conjuntamente*, isto é, da sua co-variância.

Denominador: é chamado de *Soma de Quadrados* de X, podendo ser apresentado na forma:

$$S_{XX} = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

S_{XX} é a soma dos desvios ao quadrado de X em relação à sua média, sendo uma medida da variância de X.

A fórmula de b_1, portanto, pode ser escrita como:

$$b_1 = \frac{S_{XY}}{S_{XX}}$$

isto é, a razão entre a variabilidade conjunta da variável preditora (X) e da variável resposta (Y) pela variabilidade da variável preditora (X). Esta razão pode ser interpretada como a *proporção* da variabilidade conjunta em relação a variabilidade da variável preditora.

Exemplo:
Relação Altura-Diâmetro em Árvores de *Eucalyptus grandis*

Estimativas de Quadrados Mínimos

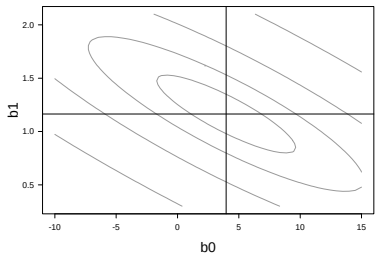
Com os dados das árvores de *Eucalyptus grandis*, podemos obter as grandezas (média e das somas de quadrados e soma de produtos) necessárias para se aplicar as fórmulas deduzidas acima:

$$\overline{Y} = 19.54446 \quad \overline{X} = 13.37822 \quad S_{XX} = 13034.01 \quad S_{XY} = 15174.91$$

Aplicando-se as fórmulas, obtemos as estimativas de quadrados mínimos:

$$\begin{aligned} b_1 &= \frac{15174.91}{13034.01} = 1.164255 \\ b_0 &= 19.54446 - 1.164255(13.37822) = 3.968804 \end{aligned}$$

Tais valores minimizam de fato a SQR, o que podemos verificar plotando-os no gráfico da função da SQR:



Exemplo:
Relação
Altura-Diâmetro
em Árvores de
Eucalyptus grandis

Estimativas de
Quadrados Mínimos
(cont.)

Vejamos o que acontece com as unidades de medida das variáveis originais no exemplo das árvores de *Eucalyptus grandis* (altura e diâmetro) quando encontramos as estimativas de quadrados mínimos. Primeiramente, devemos identificar as unidades das médias e somas de quadrados e produtos utilizadas nas fórmulas:

$$\begin{array}{ll} \overline{Y} = 19.54446\ [m] & \overline{X} = 13.37822\ [cm] \\ S_{XX} = 13034.01\ [cm^2] & S_{XY} = 15174.91\ [cm \cdot m] \end{array}$$

Aplicando as fórmulas e considerando as unidades de medida obtemos:

$$\begin{aligned} b_1 &= \frac{15174.91\ [cm \cdot m]}{13034.01\ [cm^2]} \\ &= 1.164255\ [m/cm] \end{aligned}$$

$$\begin{aligned} b_0 &= 19.54446\ [m] - 1.164255\ [m/cm](13.37822\ [cm]) \\ &= 3.968804\ [m] \end{aligned}$$

Portanto, podemos de fato interpretar b_1 como uma medida da variação na altura das árvores que ocorre com uma variação no diâmetro. O valor encontrado sugere que duas árvores que tenha uma diferença de 1 cm no diâmetro, terão *em média* uma diferença de 1.16 m na altura.

Por outro lado, o valor de b_0 sugere que quando o diâmetro é zero a altura da árvore é 3.97 m. Esta sugestão, no entanto, é inapropriada pois sabemos que o diâmetro é medido a 1.30 m de altura (DAP) e, consequentemente, este deveria ser o valor apropriado.

1.3.4 Aplicação do Modelo

Uma das funções dos modelos quantitativos em geral, e dos modelos florestais em particular, é a sua aplicação em situações práticas onde desejamos conhecer o comportamento da variável resposta, mas possuimos informação apenas da variável preditora. Nesta circunstância, o modelo é utilizado para *estimar* o valor da variável resposta sendo aplicado da seguinte maneira:

$$\widehat{Y}_h = b_0 + b_1 X_h$$

onde:

$\widehat{Y}_h$ é o valor estimado da variável resposta;

X_h é o valor da variável preditora, para o qual desejamos estimar a variável resposta;

b_0, b_1 são as estimativas de quadrados mínimos;

h é o subscrito utilizado para denotar que estamos nos referindo a uma observação h que *não* fazia parte da amostra utilizada para encontar b_0 e b_1.

No caso das observações utilizadas para ajustar o modelo utilizamos sempre o subscrito i ($Y_i; X_i; i = 1, 2, \ldots, n$).

Ao utilizarmos um modelo ajustado por regressão linear para estimar a variável resposta podem aconter duas situações:

Interpolação: o valor da variável preditora (X_h) embora não faça parte da amostra original utilizada para ajustar o modelo, está *dentro da amplitude* dos dados utilizados no ajuste.

Esta é a situação para a qual os modelos de regressão são contruídos. A confiabilidade das estimativas obtidas por interpolação se fundamenta na teoria estatística que desenvolveu os modelos de regressão linear.

Extrapolação: o valor da variável preditora (X_h) está *fora da amplitude* dos dados utilizados no ajuste.

Esta é a situação indesejável que deveria ser evitada, pois não podemos utilizar a teoria estatística para garantir a qualidade de estimativas obtidas por extrapolação. O comportamento estatístico de todo modelo de regressão linear só pode ser analisado *dentro da amplitude* dos dados originais utilizados no ajuste do modelo.

Exemplo:
Relação Altura-Diâmetro em Árvores de *Eucalyptus grandis*

Aplicação do Modelo

Uma vez ajustado os dados da altura e diâmetro de árvores de *Eucalyptus grandis* ao modelo linear simples obtivemos o seguinte modelo para estimar a altura em função do diâmetro:

$$\widehat{h}_h = 3.968804 + 1.164255\ (d_h)$$

onde $\widehat{h}_h$ é a altura a ser estimada e d_h é o diâmetro medido.

Desejamos agora estimar a altura de árvores com os seguintes diâmetros (cm):

2, 10, 20, 30, 60, 80

Utilizando o modelo ajustado obtemos as seguintes estimativas:

d_h (cm)	2	10	20	30	60	80
$\widehat{h}_h$ (m)	6.3	15.6	27.3	38.9	73.8	97.1

Exemplo: Relação Altura-Diâmetro em Árvores de *Eucalyptus grandis*

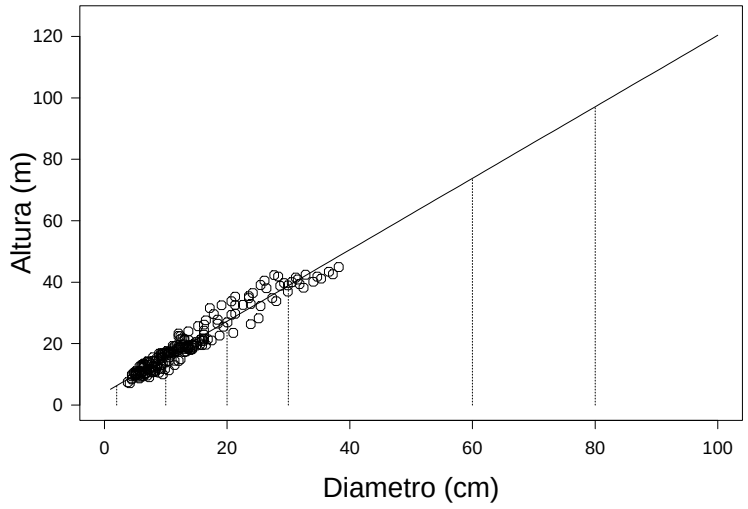
Aplicação do Modelo (cont.)

Analisemos agora estes resultados. As estimativas de altura parecem razoáveis para as árvores com diâmetro até 30 cm, mas para as árvores com os maiores diâmetros (60 e 80 cm) elas parecem desproporcionais. Quantas árvores de 60 cm com 73.8 m de altura você já viu? Seria possível uma árvore ter 97.1 m de altura ?

As árvores mais altas do mundo chegam no máximo a 100 m de altura. Mas estas árvores gigantes não são *Eucalyptus grandis*, tem muito mais que 14 anos e não estão localizadas no Estado de São Paulo.

As alturas estimadas para diâmetros de 60 e 80 cm são *extrapolações*, que neste caso resultaram em estimativas de altura totaltmente inapropriadas. A estimativa da altura para o diâmetro de 2 cm também é uma extrapolação que, embora difícil julgar se é apropriada ou não, pode ser tão irreal quanto as outras.

Para visualizarmos o que é a intrapolação e a extrapolação, bem como dos possíveis enganos resultantes da extrapolação, devemos contruir um gráfico de dispersão onde colocamos os dados originais e o modelo ajustado.



Exercícios

Os exercícios que se seguem utilizarão os dados da tabela abaixo referente a árvores de *Eucalyptus grandis* com idade inferior a 4 anos. Em todos eles serão utilizado o modelo linear simples (modelo 1.2).

Arv.	DAP (cm)	Volume (dm^3)	Arv.	DAP (cm)	Volume (dm^3)	Arv.	DAP (cm)	Volume (dm^3)
1	10.82	74.3	24	14.01	126.4	47	3.82	2.6
2	11.14	77.4	25	16.87	208.9	48	6.37	14.8
3	10.19	63.9	26	7.00	17.4	49	5.73	10.8
4	9.87	59.0	27	6.37	13.9	50	6.05	12.7
5	10.50	68.9	28	7.32	15.8	51	6.68	17.1
6	8.91	43.6	29	5.73	10.0	52	7.00	18.3
7	7.96	32.0	30	6.68	14.8	53	7.96	23.8
8	5.09	6.4	31	7.32	21.0	54	7.64	22.9
9	5.73	14.7	32	5.73	11.1	55	9.23	38.0
10	13.05	106.0	33	8.28	29.3	56	7.64	24.1
11	12.41	107.4	34	8.59	30.5	57	9.23	34.7
12	12.73	106.2	35	8.91	31.7	58	6.05	12.5
13	12.10	96.3	36	8.28	26.2	59	7.32	22.1
14	13.37	109.5	37	7.96	28.4	60	6.37	16.1
15	13.69	115.6	38	8.28	21.9	61	8.59	33.3
16	14.32	125.8	39	8.91	25.0	62	9.23	35.4
17	15.92	182.1	40	9.87	37.0	63	7.96	23.3
18	16.55	197.5	41	9.55	29.6	64	9.55	41.4
19	17.51	227.8	42	10.50	45.0	65	9.87	50.1
20	12.41	102.1	43	11.46	59.0	66	10.50	57.2
21	13.37	119.7	44	5.41	9.3	67	12.10	66.7
22	14.32	132.5	45	4.14	3.9	68	11.46	63.3
23	13.69	123.8	46	4.46	4.7	69	12.41	73.8

1.3.1 Ajuste o modelo linear simples (modelo 1.2) aos dados acima utilizando:

- variável resposta: $Y_i = \text{Volume}_i$;
- variável preditora: $X_i = DAP_i$;

e responda as seguintes questões:

a) Qual os valores de b_0 e b_1 encontrados ?
b) Quais as unidades de medida de b_0 e b_1 ?
c) Qual a interpretação prática para os valores de b_0 e b_1 encontrados ?
d) Qual a estimativa do volume de árvores com DAP igual a: 5, 10, 15, 20, 25 e 30 cm ?
e) Quais das estimativas acima são razoáveis?

1.3.2 Ajuste o modelo linear simples (modelo 1.2) aos dados acima da mesma forma que o exercício anterior, mas utilize agora as seguintes variáveis:

- variável resposta: $Y_i = \text{Volume}_i$;

- variável preditora: $X_i = DAP_i^2$;

Responda as seguintes questões:

a) Qual os valores de b_0 e b_1 encontrados ?
b) Quais as unidades de medida de b_0 e b_1 ?
c) Qual a interpretação prática para os valores de b_0 e b_1 encontrados ?
d) Qual a estimativa do volume de árvores com DAP igual a: 5, 10, 15, 20, 25 e 30 cm ?
e) Quais das estimativas acima são razoáveis?

1.3.3 Ajuste o modelo linear simples (modelo 1.2) aos dados acima da mesma forma que os dois exercício anteriores, mas altere as variáveis do modelo para:

- variável resposta: $Y_i = \log(\text{Volume}_i)$;

- variável preditora: $X_i = \log(DAP_i)$;

- onde $\log$ é o logaritmo neperiano (base e $= 2.718282$).

Responda as seguintes questões:

a) Qual os valores de b_0 e b_1 encontrados ?
b) Quais as unidades de medida de b_0 e b_1 ?
c) Qual a interpretação prática para os valores de b_0 e b_1 encontrados ?
d) Qual a estimativa do volume de árvores com DAP igual a: 5, 10, 15, 20, 25 e 30 cm ?
e) Quais das estimativas acima são razoáveis?

O exercícios que se seguem se baseiam nos dados abaixo e na modelo linear simples (modelo 1.2).

País	POP75 População com + 75 anos (%)	RENDA Renda Per Capta (US$)	País	POP75 População com + 75 anos (%)	RENDA Renda Per Capta (US$)
Australia	2.87	2329.68	Malta	2.47	601.05
Austria	4.41	1507.99	Norway	3.67	2231.03
Belgium	4.43	2108.47	Netherlands	3.25	1740.70
Bolivia	1.67	189.13	New.Zealand	3.17	1487.52
Brazil	0.83	728.47	Nicaragua	1.21	325.54
Canada	2.85	2982.88	Panama	1.20	568.56
Chile	1.34	662.86	Paraguay	1.05	220.56
Taiwan	0.67	289.52	Peru	1.28	400.06
Colombia	1.06	276.65	Philippines	1.12	152.01
Costa.Rica	1.14	471.24	Portugal	2.85	579.91
Denmark	3.93	2496.53	South.Africa	2.28	651.11
Ecuador	1.19	287.77	Rhodesia	1.52	250.96
Finland	2.37	1681.25	Spain	2.87	768.79
France	4.70	2213.82	Sweden	4.54	3299.49
Germany	3.35	2457.12	Switzerland	3.73	2630.96
Greece	3.10	870.85	Turkey	1.08	389.66
Guatemala	0.87	289.71	Tunisia	1.21	249.87
Honduras	0.58	232.44	United.Kingdom	4.46	1813.93
Iceland	3.08	1900.10	United.States	3.43	4001.89
India	0.96	88.94	Venezuela	0.90	813.39
Ireland	4.19	1139.95	Zambia	0.56	138.33
Italy	3.48	1390.99	Jamaica	1.73	380.47
Japan	1.91	1257.28	Uruguay	2.72	766.54
Korea	0.91	207.68	Libya	2.07	123.58
Luxembourg	3.73	2449.39	Malaysia	0.66	242.69

1.3.4 Construa um modelo de regressão linear simples com as seguintes variáveis:

- variável resposta: $Y_i = \text{RENDA}_i$;
- variável preditora: $X_i = \text{POP}75_i$;

Responda as seguintes questões:

a) Qual os valores de b_0 e b_1 encontrados ?
b) Quais as unidades de medida de b_0 e b_1 ?
c) Qual a interpretação prática para os valores de b_0 e b_1 encontrados ?
d) Qual a estimativa da renda per capta para países com população com mais de 75 anos de: 0.1, 0.5, 2.0, 3.0, 4.5, 5.0, 10.0, 15.0 % ?
e) Quais das estimativas acima são razoáveis?

1.3.5 Construa um modelo de regressão linear simples semelhante ao exercício anterior, mas utilize as variáveis:

- variável resposta: $Y_i = \log(\text{RENDA}_i)$;
- variável preditora: $X_i = \text{POP}75_i$;
- onde $\log$ é o logaritmo neperiano (base e $= 2.718282$).

Responda as seguintes questões:

a) Qual os valores de b_0 e b_1 encontrados ?
b) Quais as unidades de medida de b_0 e b_1 ?

c) Qual a interpretação prática para os valores de b_0 e b_1 encontrados ?
d) Qual a estimativa da renda per capta para países com população com mais de 75 anos de: 0.1, 0.5, 2.0, 3.0, 4.5, 5.0, 10.0, 15.0 % ?
e) Quais das estimativas acima são razoáveis?

2 INFERÊNCIA EM REGRESSÃO LINEAR

2.1 Componente Probabístico

Os estimadores de mínimos quadrados garantem a minimização do quadrado dos desvios. Para que possamos utilizar o modelo ajustado dentro de um contexto estatístico é necessário incorporar ao nosso modelo estatístico geral:

DADOS = MODELO + ERRO

um componente probabilístico. Com base nos aspectos probabilísticos do modelo, podemos verificar a qualidade do modelo ajustado em relação aos dados originais e fazer comparações estatísticas utilizando o MODELO.

No caso do modelo linear simples

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

os seus elementos são definidos como:

Y_i é o valor da variável resposta para a i$^{\text{ésima}}$ observação;

X_i é o valor da variável preditora para a i$^{\text{ésima}}$ observação;

ε_i é o erro aleatório (não explicado) associado à i$^{\text{ésima}}$ observação;

β_0 e β_1 são os parâmetros a serem estimados (pelo método dos quadrados mínimos).

Em termos de componente probabilística dos elementos teremos:

X_i é uma variável matemática, isto é, conhecida *sem erro de medição* e *sem efeito aleatório.* Assim o componente $\beta_0 + \beta_1 X_i$ é determinístico, isto é, sem efeito aleatório.

ε_i é uma variável aleatória com as seguintes características:

os ε_i são multuamente independententes;

possuem média zero ($\mu_\varepsilon = 0$);

possuem variância constante (σ^2);

têm distribuição Normal.

Esse modelo estatístico implica que para cada valor da variável preditora X_i, a variável resposta Y_i tem

- média igual a $\beta_0 + \beta_1 X_i$;
- variância constante igual a σ^2;
- distribuição Normal.

A figura 2.1 apresenta uma representação gráfica do modelo linear simples que incorpora os aspectos probabilísticos. Note que para cada valor de X_i, o valor de Y_i esperado segundo o modelo ($\widehat{Y}_i = \beta_0 + \beta_1 X_i$) é a média de uma distribuição normal que possue variância σ^2. Note ainda que a variância σ^2 é constante para todos os valores de X_i.

O modelo de *Regressão Linear Simples* é composto não só pela fórmula

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

como também pelas *pressuposições* probabilísticas que definem o comportamento de Y_i e ε_i.

2.2 Inferência sobre os Parâmetros do Modelo

2.2.1 Propriedades das Estimativas de Quadrados Mínimos

Incluindo o componente probabilístico o modelo de regressão linear simples fica:

$$Y_i \quad = \quad \beta_0 + \beta_1 X_i + \varepsilon_i$$

onde $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$, isto é, os erros ($\varepsilon_i$) são independentes e têm distribuição Normal com média 0 (zero) e variância constante σ^2.

A importância das pressuposições sobre o comportamento dos erros no modelo linear é permitir a dedução de propriedades estatísticas da estimativas de quadrados mínimos. No modelo com erros normais as estimativas de quadrados mínimos b_0 e b_1 terão ambas distribuição Normal. De fato, pode ser provado que:

$$b_0 \quad \sim \quad N\left(\beta_0 \quad , \quad \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right]\right)$$

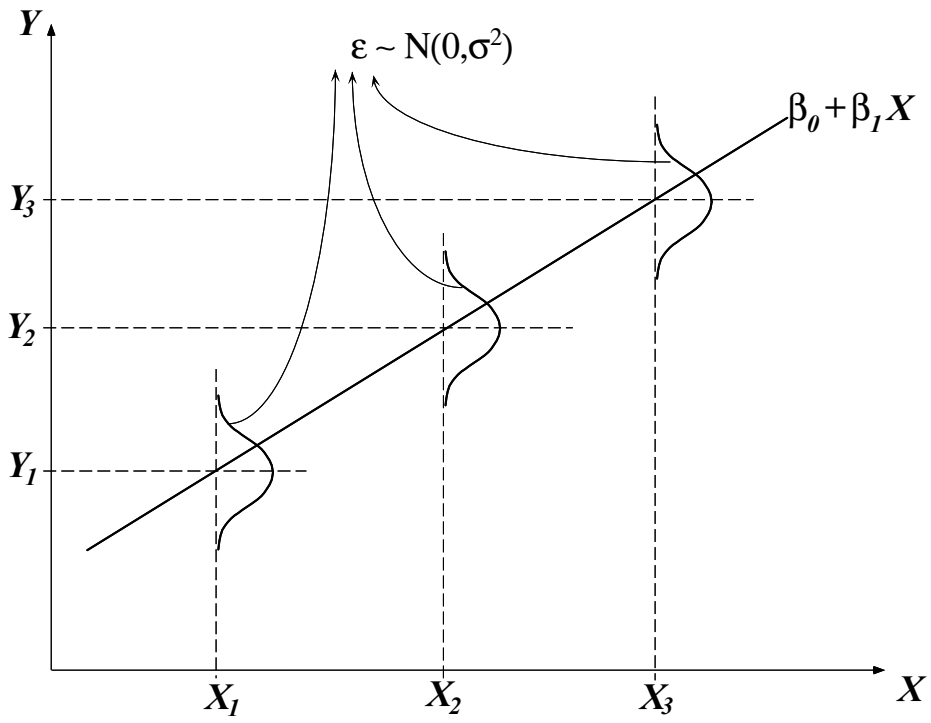


Figura 2.1: Representação gráfica do modelo estatístico linear simples.

$$b_1 \sim N\left(\beta_1 \quad , \quad \frac{\sigma^2}{\sum(X_i - \overline{X})^2}\right)$$

Note que σ^2 se refer a variância dos erros e para encontrarmos as variâncias de b_0 e b_1 precisamos estimar σ^2. O melhor forma de estimar a variância do erro é utilizando a variância dos resíduos, portanto, a estimativa de σ^2 é:

$$\frac{\sum e_i^2}{n-2} = \frac{SQR}{n-2} = QMR$$

onde n é o número de observações e QMR é chamado de "Quadrado Médio dos Resíduos". A SQR é divida pelos graus de liberdade $n-2$, onde o número de observações n é reduzido em 2, pois dois parâmetros foram estimados (β_0 e β_1).

As variâncias das estimativas dos parâmetros são encontradas, portanto, pelas fórmulas:

$$\begin{aligned} s^2\{b_0\} &= QMR\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right] = QMR\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right] \\ s^2\{b_1\} &= \frac{QMR}{\sum(X_i - \overline{X})^2} = \frac{QMR}{S_{XX}} \end{aligned}$$

2.2.2 Testes de Hipóteses

Para testarmos hipóteses sobre estes parâmetros do modelo de regressão podemos utilizar o teste t de Student. Uma hipótese frequentemente testada é se o valor do parâmetro é igual a zero. A notação estatística para testar tal hipótese no caso dos parâmetros do modelo de regressão linear simples é:

Hipótese Nula	$H_0 : \beta_0 = 0$	$H_0 : \beta_1 = 0$
Hipótese Alternativa	$H_\alpha : \beta_0 \neq 0$	$H_\alpha : \beta_1 \neq 0$

No caso de β_0 (intercepto), a hipótese nula implica que o modelo de regressão é de fato

$$Y_i = \beta_1 X_i + \varepsilon_i$$

isto é, a linha de regressão passa pela origem $(X = 0, Y = 0)$. Tal hipótese tem poucas implicações práticas.

Já no caso do parâmetro da inclinação (β_1), a hipótese nula implica no modelo

$$Y_i = \beta_0 + \varepsilon_i$$

o que significa que não existe relação linear entre X e Y, pois o modelo mais adequado é uma constante (β_0). Testar esta hipótese é uma das maneiras de verificar se o modelo ajustado é confiável.

Para utilizar o teste t de Student, basta utilizar a estatística:

$$t_0^* = (b_0 - 0)/s\{b_0\} \qquad t_1^* = (b_1 - 0)/s\{b_1\}$$

Os valores desta estatística devem ser comparados com os valores tabelados de t. Para o nível de significância α o valor tabelado é $t(1 - \alpha/2; n - 2)$, onde n é o número de observações. A regra de decisão fica:

- se $|t^*| \geq t(1 - \alpha/2; n - 2) \Rightarrow$ rejeita-se H_0 e aceita-se H_α;
- se $|t^*| < t(1 - \alpha/2; n - 2) \Rightarrow$ rejeita-se H_α e aceita-se H_0.

2.2.3 Intervalo de Confiança

De modo análogo ao teste de hipóteses, Intervalos de Confiança podem ser construidos para as estimativas dos parâmetros. Os Intervalos de Confiança de $(1 - \alpha)100\%$ para β_0 e β_1 são:

$$b_0 \pm t(1 - \alpha/2; n - 2)s\{b_0\}$$
$$b_1 \pm t(1 - \alpha/2; n - 2)s\{b_1\}$$

Exemplo: Relação Altura-Diâmetro em Árvores de *Eucalyptus grandis*

Inferência sobre os Parâmetros

Para realizarmos a inferência sobre os parâmetros necessitamos do modelo:

$$\widehat{h}_h \quad = \quad 3.9688 + 1.1643\ (d_h)$$

onde $\widehat{h}_h$ é a altura a ser estimada e d_h é o diâmetro medido, e de algumas grandesas relativas aos dados:

$$\begin{array}{ll} n = 213 & \overline{X} = 13.37822 \\ \sum (X_i - \overline{X})^2 = 13034.01 & QMR = 5.84 \end{array}$$

Assim temos os erros padrões das estimativas dos parâmetros ficam:

$$\begin{aligned} s\{b_0\} &= \sqrt{5.84 \left[\frac{1}{213} + \frac{(13.37822)^2}{13034.01} \right]} = \boxed{0.3280} \\ s\{b_1\} &= \sqrt{\frac{5.84}{13034.01}} = \boxed{0.0212} \end{aligned}$$

Teste de hipóteses em relação a b_0 ($\alpha = 0.05$):

$$\left\{ \begin{array}{l} H_0 : \beta_0 = 0 \\ H_\alpha : \beta_0 \neq 0 \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} t^* = 3.9688/0.3280 = 121.000 \\ t(1-\alpha/2; n-2) = t(0.975; 211) = 1.971 \end{array} \right.$$

DECISÃO: como $|t^*| \geq t(1-\alpha/2; n-2)$ rejeita-se H_0.

Teste de hipóteses em relação a b_1 ($\alpha = 0.05$):

$$\left\{ \begin{array}{l} H_0 : \beta_1 = 0 \\ H_\alpha : \beta_1 \neq 0 \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} t^* = 1.1643/0.0212 = 54.920 \\ t(1-\alpha/2; n-2) = t(0.975; 211) = 1.971 \end{array} \right.$$

DECISÃO: como $|t^*| \geq t(1-\alpha/2; n-2)$ rejeita-se H_0.

Intervalo de Confiança de 95%:

$$\begin{array}{ll} b_0 \pm t(1-\alpha/2; n-2) s\{b_0\} & \Rightarrow 3.9688 \pm (1.971)(0.3280) \\ & \Rightarrow 3.9688 \pm 0.6465 \\ \\ b_1 \pm t(1-\alpha/2; n-2) s\{b_1\} & \Rightarrow 1.1643 \pm (1.971)(0.0212) \\ & \Rightarrow 1.1643 \pm 0.0418 \end{array}$$

2.3 Verificando a Adequação do Modelo Linear

Como o modelo linear simples é mais do que uma simples fórmula e incorpora *pressuposições probabilísticas*, é necessário saber se tais pressuposições são razoáveis para os DADOS que dispomos para ajustar o modelo. Pelo método de quadrados mínimos, obtemos estimativas dos parâmetros β_0 e β_1 do modelo. Sabemos que tais estimativas minimizam a Soma de Quadrado dos Resíduos:

$$SQR = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$$

Como os resíduos e_i são os nossos melhores representantes dos erros ε_i, devemos agora verificar se eles têm o comportamento que o modelo linear afirma que os erros devem ter. Podemos enumerar as pressuposições do modelo linear simples como:

Pressuposições do Modelo Linear Simples

1. A relação entre X e Y é linear e o termos dos erros (ε_i) é aditivo.
2. O número de observações (n) é maior que o número de parâmetros a serem estimados (p).
3. A variável preditora (X_i) é não-estocásticas.
4. Os erros ε_i são aleatórios e independentes (não correlacionados).
5. Os erros ε_i têm variância constante (σ^2) em relação ao modelo.
6. Os erros ε_i têm distribuição Normal com com média zero.

As pressuposições (2) a (4) são assumidas como verdadeiras na maioria dos modelos biométricos florestais e, em geral, são verificadas somente em situações especiais. Para a maioria dos dados obtidos em mensuração florestal, estas pressuposições são razoáveis. Na prática, mais atenção é dada às pressuposições (1), (5) e (6), pois elas acarretam implicações sérias sobre o modelo linear caso seja violadas.

2.3.1 Relação Linear e Variância Constante

Para se verificar a pressuposição de que a relação entre X e Y é linear e de que a variância do erro é constante (pressuposições 1 e 5), utiliza-se um gráfico de dispersão

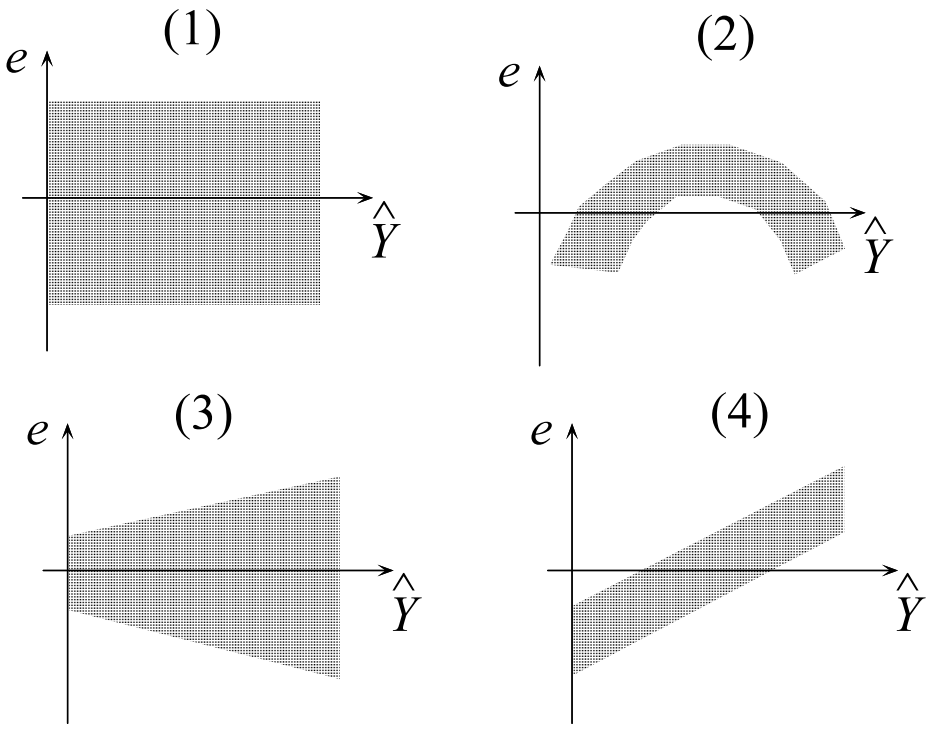


Figura 2.2: Gráficos de dispersão dos resíduos: (1) padrão apropriado, (2) relação não-linear entre X e Y, (3) variância crescente com X, (4) relação não-linear entre X e Y.

do resíduo ($e_i = Y_i - \widehat{Y}_i$) contra os valores estimados pelo modelo ($\widehat{Y}_i$). A figura 2.2 apresenta vários gráfico de dispersão onde os resíduos tem diferentes comportamentos. O comportamento ideal (figura 2.2) se resume em:

a) os resíduos se distribuem ao longo de todo o eixo x;
b) a distribuição tem a forma de uma "faixa" centrada na linha de resíduo igual a zero, com igual amplitude para valores positivos e valores negativos;
c) a largura desta "faixa" é constante (variância constante).

Qualquer padrão de dispersão diferente pode implicar em que a pressuposição de variância constante não seja válida.

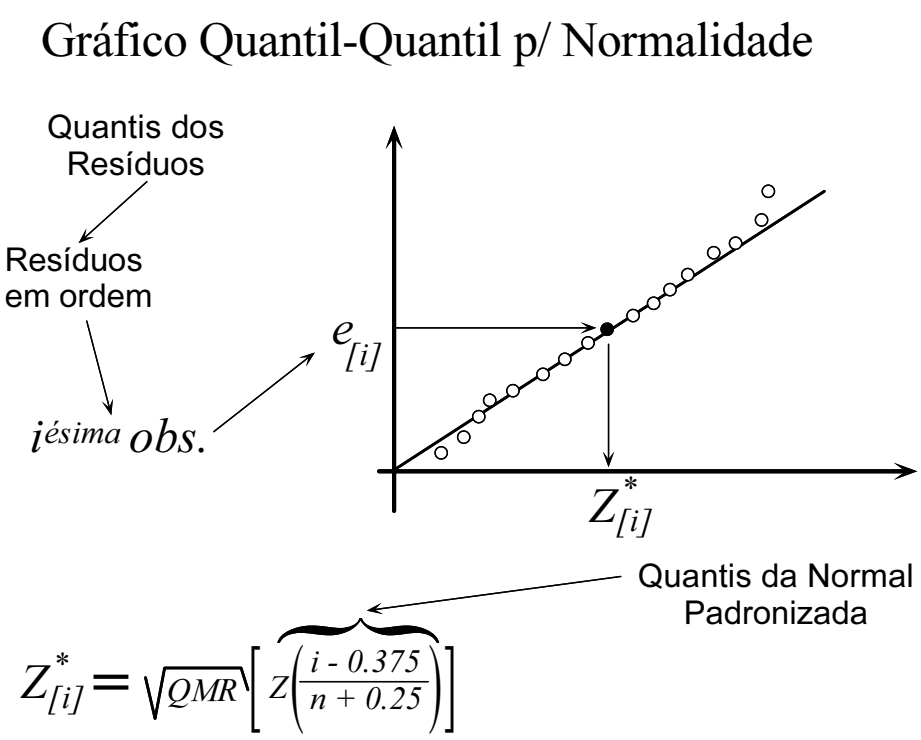


Figura 2.3: Gráfico Quantil-Quantil dos resíduos para verificar a normalidade dos dados.

2.3.2 Normalidade dos Erros

A pressuposição de normalidade dos erros (pressuposição 6) pode ser verificada por teste de ajustamento de distribuições (como o teste de Qui-Quadrado ou Komolgorov-Smirnov). Para se efetuar estes testes os dados são em geral agrupados em classes o que pode gerar perda de informação. Uma análise mais visual dos dados é muitas vêzes mais informativa e neste caso se constroe um gráfico Quantil-Quantil (gráfico QQ). Num gráfico QQ, os quantis *empíricos* da variável sendo estudada são comparados com os quantis de uma distribuição estatística qualquer, no nosso caso a distribuição normal. A figura 2.3 mostra como se constore um gráfico QQ no caso da distribuição Normal. Note que os pontos do gráfico estão posicionados ao longo de uma reta. Este é comportamento esperado para uma resíduos com distribuição Normal quando os quantis dos resíduos é plotado contra os quantis da distribuição Normal padronizada.

A figura 2.4 mostra como a distribuição dos resíduos pode desviar-se da distribuição

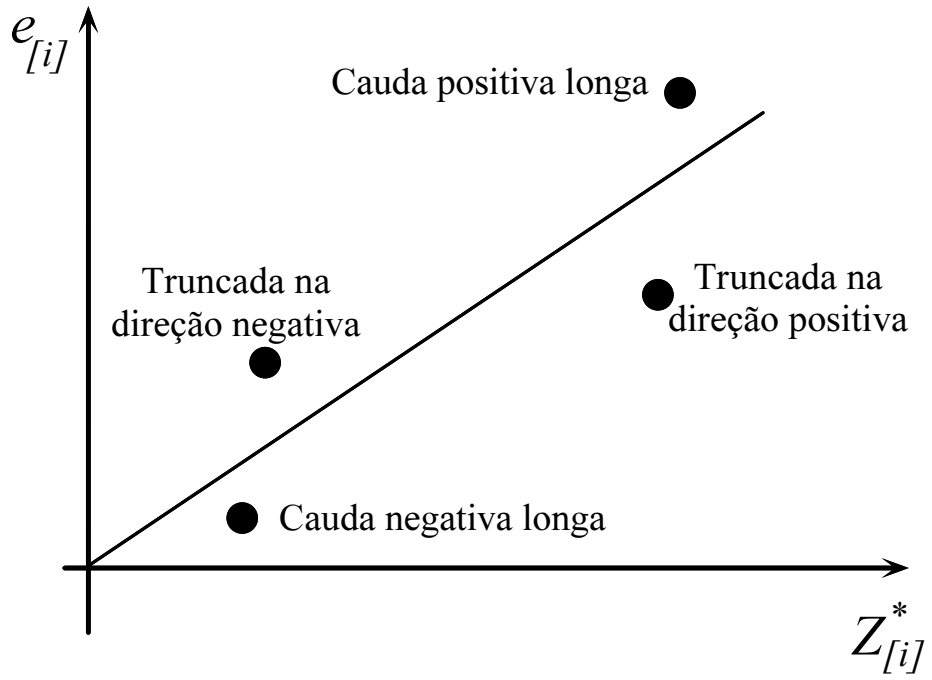


Figura 2.4: Desvios da Normalidade mostrados no gráfico Quantil-Quantil dos resíduos.

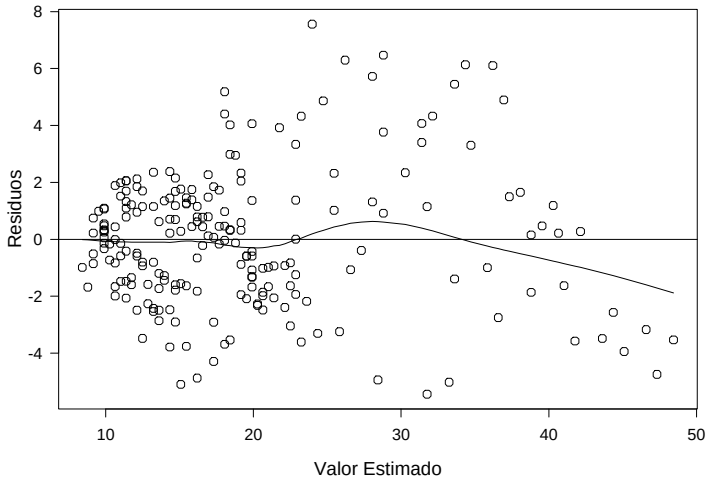
Normal. De modo geral, pequenos desvios da reta na cauda da distribuição são aceitáveis. Já desvios no centro dos dados indicam forte desvio da normalidade. É importante lembrar que o tamanho da amostra (número de pontos no gráfico) influencia o julgamento. Para grandes amostras, pequenos desvios da reta podem ser considerados importantes.

Exemplo:
Relação Altura-Diâmetro em Árvores de *Eucalyptus grandis*

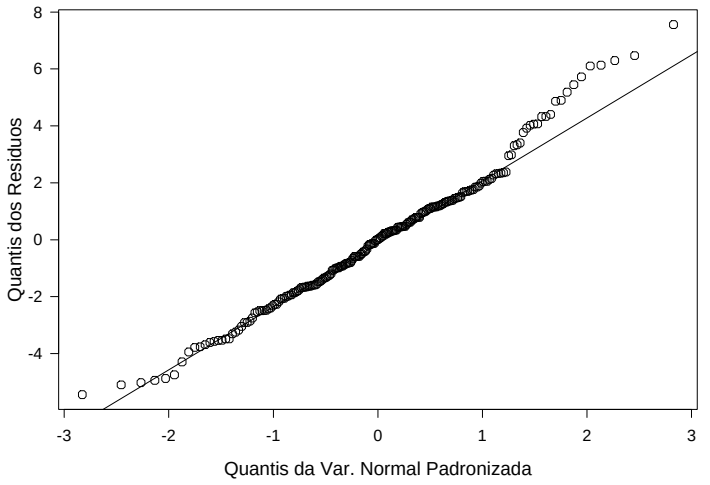
Adequação do Ajuste

Para verificarmos se o modelo é adequado ao dados devemos verficar se as pressuposições da regressão linear podem ser aceitas.

O gráfico de dispersão dos resíduos, mostra que a relação linear é uma pressuposição aceitável, mas provavelmente a variância não é constante.



Já o gráfico QQ aponta para normalidade dos resíduos, embora com uma certa assimetria à direita. O único problema que o modelo parecem apresentar é em relação à variância não ser constante.



2.4 Exercícios

2.4.1 Utilizando os dados de DAP e volume de árvores de *E. grandis*, nos exercícios do capítulo anterior (pag. 32). Ajuste o modelo linear simples tomando considerando dois modelos dendrométricos:

Modelo Dendrométrico 1: Y_i = VOLUME e X_i = DAP.

Modelo Dendrométrico 2: Y_i = ln(VOLUME) e X_i = ln(DAP).

Para cada modelo, realize as seguintes análises:

a) Utilize gráficos para verificar as como cada modelo se comporta em relação às pressuposições do modelo de regressão linear simples. Estabeleça suas conclusões de modo claro e conciso.
b) Teste a hipótese de que o valor dos parâmetros de cada modelo é igual a zero. Interprete os seus resultados.
c) Construa Intervalos de Confiança de 95% relativos aos parâmetros de todos os modelos ajustados. Interprete os seus resultados.

2.4.2 Utilizando os dados demográficos de diversos paises, apresentados nos exercícios do capítulo anterior (pag. 33), ajuste os modelos abaixo por regressão linear:

Modelo 1: Y_i = RENDA e X_i = POP75.

Modelo 2: Y_i = ln(RENDA) e X_i = ln(POP75).

Para cada modelo, realize as seguintes análises:

a) Utilize gráficos para verificar as como cada modelo se comporta em relação às pressuposições do modelo de regressão linear simples. Estabeleça suas conclusões de modo claro e conciso.
b) Teste a hipótese de que o valor dos parâmetros de cada modelo é igual a zero. Interprete os seus resultados.
c) Construa Intervalos de Confiança de 95% relativos aos parâmetros de todos os modelos ajustados. Interprete os seus resultados.

2.5 Verificando o Ajuste do Modelo

Uma vez que temos certeza que as pressuposições do modelo linear foram adequadamente alcançadas podemos então verificar se o modelo construido possui a

qualidade necessária para ser utilizado. “Qualidade” nesse caso significa que os valores observados são razoavelmente estimados pelo modelo. Ao contrário da verificação das pressuposições, nesse caso costuma-se se utilizar índices e testes estatísticos para definir se o modelo representa bem os dados.

2.5.1 Coeficiente de Determinação

O primeiro índice utilizado é o *Coeficiente de Determinação*:

$$R^2 = \frac{(S_{XY})^2/S_{XX}}{S_{YY}} = \frac{SQM}{SQT} = 1 - \frac{SQR}{SQT}$$

onde:

$SQT = S_{YY} = \sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2/n$ é a Soma de Quadrados Total, ou a variabilidade total da variável resposta (Y);

$SQM = (S_{XY})^2/S_{XX}$ é a Soma de Quadrados do Modelo, isto é, a variabilidade da variável resposta que o modelo linear consegue explicar.

A SQT representa a variabilidade total dos dados, enquanto a SQM é a variabilidade explicada pelo modelo linear. O R^2, portanto, representa a proporção da variabilidade total que é explicada pelo modelo, consequentemente: $0 \leq R^2 \leq 1$. Quanto mais próximo de 1 estiver R^2, melhor a qualidade do ajuste.

Exemplo: Relação Altura-Diâmetro em Árvores de *Eucalyptus grandis*

Coeficiente de Determinação

As grandezas necessárias ao cálculo do Coeficiente de Determinação são:

$$\sum(Y_i - \overline{Y})^2 = 18899.32 \qquad \sum(X_i - \overline{x})^2 = 13034.01$$
$$\sum[(Y_i - \overline{Y})(X_i - \overline{x})] = 15174.91$$

As somas de quadrados e produtos e o coeficiente de determinação ficam:

$$SQT = 18899.32$$
$$SQM = \frac{(15174.91)^2}{13034.01} = 17667.46$$
$$R^2 = 1 - \frac{17667.46}{18899.32} = 0.9348$$

Este valor indica que apesar de existir uma forte relação entre a altura total e o DAP das árvores *E. grandis*, e o modelo ajustado explica apenas 93% da variação observada nas alturas das árvores. Trata-se, portanto, de um bom modelo para se estimar a altura das árvores.

Sabemos que quanto mais próximo de 1, melhor o R^2 do modelo. No entanto, o que é estar próximo de 1? Para relações hipsométricas em florestas plantadas é comum trabalharmos com R^2 maiores do que 0.90, assim valores abaixo disto não são considerados bons. Mas em outras relações dendrométricas e florestais modelos com R^2 menores que 0.90 podem ser considerados bons dada a complexidade das variáveis envolvidas. Em quase toadas as situações florestais evitamos utilizar modelos cujo coeficiente de determinação seja inferior a 0.50, pois a qualidade das estimativas se torna seriamente questionável.

2.5.2 Análise de Variância do Modelo

Outra forma de se testar um modelo linear ajustado é através do teste F, o qual é obtido na forma de uma tabela de análise de variância. Nesse caso a variância total é sub-dividida em duas partes uma explicada pelo modelo e a outra não explicada (resíduo). O teste F é uma comparação dessas duas variâncias. A tabela de análise de variância é construida da seguinte maneira:

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Teste F
Modelo	$p-1$	SQM	$QMM = SQM/(p-1)$	QMM/QMR
Resíduo	$n-p$	$SQR = SQT - SQM$	$QMR = SQR/(n-p)$	
Total	$n-1$	SQT		

A hipótese nula formal sendo testada na análise de variância é a seguinte:

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0$$

Ela é testada contra a seguinte hipótese alternativa:

$$H_\alpha : \beta_i \neq 0, \quad \text{para pelo menos dos parâmetros do modelo}$$

Sob H_0, isto é, caso a hipótese nula seja verdadeira, a estatística:

$$F \quad = \quad \frac{QMM}{QMR}$$

tem distribuição F com graus de liberdade $p-1$ para o numerador (ν_1) e $n-p$ para o denominador (ν_2).

Para considerarmos o modelo como tendo um bom ajuste devemos rejeitar a hipótese nula. A hipótese nula é rejeitada ao nível α de probabilidade (em geral $\alpha = 0.05$ ou

5% de probabilidade) quando a estatística calculada é maior ou igual ao valor

$$F_{[1-\alpha;\mu_1=p-1;\mu_2=n-p]}$$

da distribuição de F encontrado em tabelas estatísticas.

O modelo ajustado também deve ser testado em termos das estimativas dos parâmetros do modelo. Caso o modelo proposto seja de fato apropriado para os dados, as estimativas dos parâmetros devem ser estatísticamente diferentes de zero. Isso é testado verificando se os Intervalos de Confiança construidos para as estimativas dos parâmetros incluem o valor zero. Se o intervalo de confiança de uma das estimativas abranger o zero, a estimativa não pode ser considerada estatisticamente diferente de zero, sugerindo que o modelo apropriado deve ser diferente do modelo ajustado.

Exemplo: Relação Altura-Diâmetro em Árvores de *Eucalyptus grandis*

Análise de Variância

Para construirmos a tabela de análise de variância partimos praticamente das mesmas somas de quadrados que utilizamos calcular o R^2:

$$\begin{aligned} SQT &= 18899.32 \\ SQM &= \frac{(15174.91)^2}{13034.01} = 17667.46 \\ SQR &= SQT - SQM = 18899.32 - 17667.46 = 1231.86 \end{aligned}$$

Com estes valores construímos a tabela de análise de variância:

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Teste F
Modelo	$2-1=1$	17667.46	17667.46	$\frac{17667.46}{5.8382} = 3026.18$
Resíduo	$213-2=211$	1231.86	$\frac{1231.86}{211} = 5.8382$	
Total	$213-1=212$	18899.32		

O valor de F encontrado é de 3026.18, que se mostra muito superior ao valor crítico para o nível de probabilidade de 5% ($\alpha = 0.05$):

$$F_{[1-\alpha;\mu_1=p-1;\mu_2=n-p]} = F_{[0.95;\mu_1=1;\mu_2=211]} = 3.885908$$

e, portanto, rejeitamos a hipótese nula. Concluimos que pelo teste F, existe uma forte relação entre a altura e o DAP e o modelo linear simples é capaz de representar esta relação.

2.6 Exercícios

2.6.1 Utilizando os dados de DAP e volume de árvores de *E. grandis*, nos exercícios do capítulo anterior (pag. 32). Ajuste o modelo linear simples tomando considerando dois modelos dendrométricos:

Modelo Dendrométrico 1: Y_i = VOLUME e X_i = textscdap.

Modelo Dendrométrico 2: Y_i = ln(VOLUME) e X_i = ln(textscdap).

Para cada modelo, verifique a qualidade do ajuste através do coeficiente de determinação e a análise de variância.

2.6.2 Utilizando os dados demográficos de diversos paises, apresentados nos exercícios do capítulo anterior (pag. 33), ajuste os modelos abaixo por regressão linear:

Modelo 1: Y_i = RENDA e X_i = POP75.

Modelo 2: Y_i = ln(RENDA) e X_i = ln(POP75).

Para cada modelo, verifique a qualidade do ajuste através do coeficiente de determinação e a análise de variância.

3 REGRESSÃO LINEAR PONDERADA

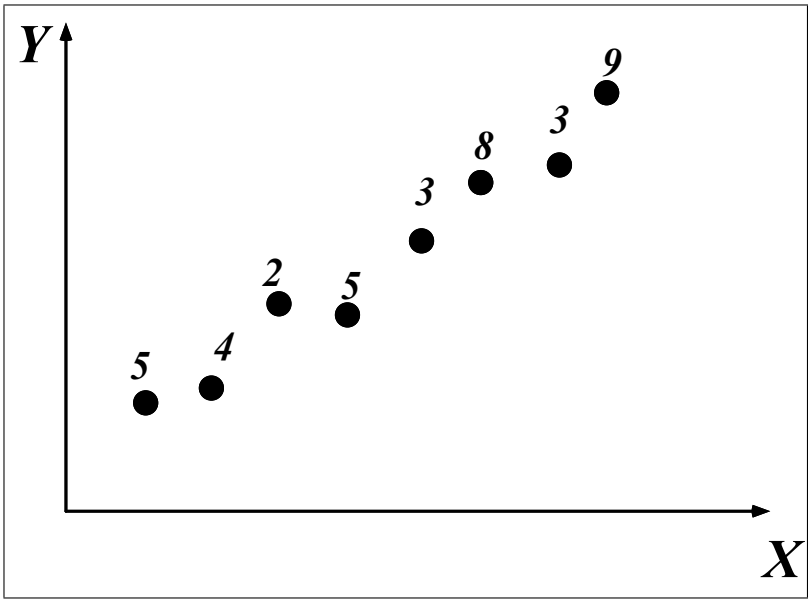
3.1 Quadrados Mínimos Ponderados

Os estimadores de Quadrados mínimos são encontrados, minimizando a **função de perda**:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Mas frequentemente não desejamos dar o mesmo peso a todas as observações.

A título de ilustração, considere o exemplo onde os dados são formados por um conjuntos de médias de Y para cada nível de X, mas o número de observações para cada média são diferentes:



Neste caso, é mais apropriado minimizar a função de perda:

$$Q_w = \sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 X_i)^2$$

onde w_i é o peso de cada observações. No exemplo acima temos:
$w_1 = 5, w_2 = 4, w_3 = 2, w_4 = 5, w_5 = 3, w_6 = 8, w_7 = 3, w_8 = 9$.

A minimização de Q_w com respeito a β_0 e β_1 produz as seguintes Equações Normais:

$$\begin{aligned} \sum w_i Y_i &= b_0 \sum w_i + b_1 \sum w_i X_i \\ \\ \sum w_i X_i Y_i &= b_0 \sum w_i X_i + b_1 \sum w_i X_i^2 \end{aligned}$$

cuja a solução é:

$$\begin{aligned} b_1 &= \frac{\sum w_i X_i Y_i - [(\sum w_i X_i)(\sum w_i Y_i)/n]}{\sum w_i X_i^2 - [(\sum w_i X_i)^2/n]} \\ \\ b_0 &= \frac{\sum w_i Y_i}{\sum w_i} - b_1 \frac{\sum w_i X_i}{\sum w_i} \end{aligned}$$

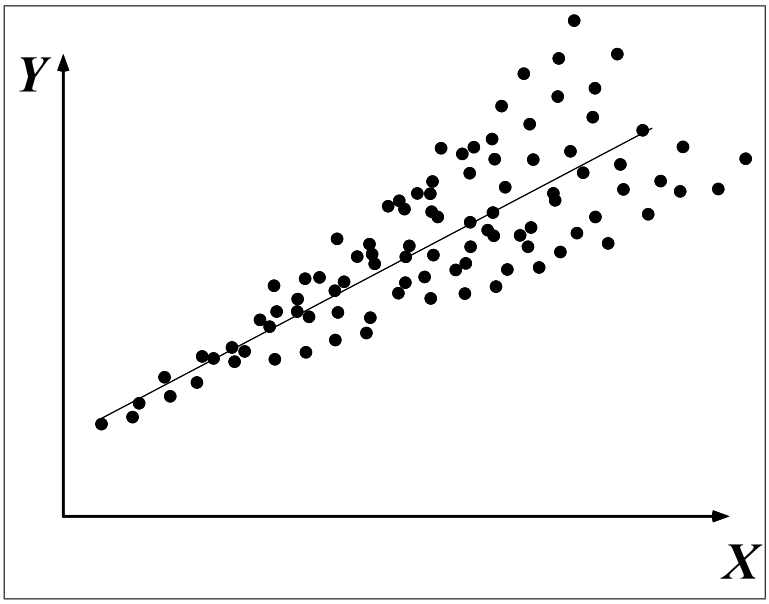
Note que se $w_i = 1 (i = 1, \ldots, n)$, estes estimadores se tornam identicos aos estimadores sem ponderação.

3.2 Contexto de Aplicação

Mas em que contexto é interessante ponderar ? Quando a variância dos erros não é constante. Um caso muito comum na área florestal é o do volume ou biomassa de uma árvore individualmente. É natural que o volume ou biomassa de árvores com grande diâmetro e altura seja mais variável que o volume ou biomassa de árvores pequenas. Uma mesma variação percentual no fator de forma ou na densidade resultará numa vaior variação em metros cúbicos ou kilogramas nas árvores grandes. O resultado é que o gráfico do volume ou biomassa como variável resposta (Y) em função do diâmetro ou altura (X) tende a ter o seguinte aspecto:

O gráfico acima sugere que podemos ter maior confiança nos valores de Y_i para pequenos valores de X_i, pois a variabilidade é menor. Como a variância de Y_i cresce de acordo com X_i, podemos supor que a cada nível i de X teremos uma variância σ_i^2. Para dar maior importância às observações que têm menor variância, podemos utilizar como peso o inverso das variâncias σ_i^2:

$$w_i = \frac{1}{\sigma_i^2}.$$



Em geral, as variâncias σ_i^2 não são conhecidas, mas, como o gráfico sugere, elas são frequentemente proporcionais ao valor de X_i. Se isto ocorrer, podemos utilizar os valores de X_i como peso:

$$\sigma_i^2 \propto X_i^2 \Rightarrow \sigma_i^2 = kX_i^2 \Rightarrow w_i = \frac{1}{X_i^2}$$

pois a constante k será eliminada das Equações Normais. Num contexto mais genérico podemos assumir que:

$$\sigma_i^2 \propto X_i^m \Rightarrow \sigma_i^2 = kX_i^m \Rightarrow w_i = X_i^{-m}$$

onde $m = -5, \ldots, 0, \ldots, +5$.

3.3 Quadrados Mínimos Ponderados através de Transformação

Utilizar o Método dos Quadrados Mínimos Ponderados para ajustar um dado modelo é equivalente a usar o Método dos Quadrados Mínimos não ponderados para ajustar um modelo transformado do modelo original. Suponhamos que o nosso modelo seja

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad\qquad \varepsilon \sim N(0, \sigma^2 X_i^m); m \neq 0$$

o que implica que a variância não é constante, mas é proporciona a X_i. Utilizando como pesos:

$$\sigma_i^2 = kX_i^m \Rightarrow w_i = \frac{1}{X_i^m},$$

a função de perda fica:

$$\begin{aligned} Q_w &= \sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 X_i)^2 \\ Q_w &= \sum_{i=1}^{n} \frac{1}{X_i^m}(Y_i - \beta_0 - \beta_1 X_i)^2 \\ Q_w &= \sum_{i=1}^{n} \left(\frac{Y_i}{X_i^m} - \beta_0 \frac{1}{X_i^{m/2}} - \beta_1 \frac{X_i}{X_i^{m/2}} \right)^2 \end{aligned}$$

Portanto, a regressão ponderada é equivalente a ajustar o modelo

$$\begin{aligned} \frac{Y_i}{X_i^{m/2}} &= \beta_0 \frac{1}{X_i^{m/2}} + \beta_1 \frac{X_i}{X_i^{m/2}} + \frac{\varepsilon_i}{X_i^{m/2}} \\ Y_i^{/} &= \beta_0^* + \beta_1^* X_i^{/} + \varepsilon_i^{/} \end{aligned}$$

que não possui o problema de variância não homogênea, pois

$$\varepsilon_i \sim N(0, \sigma^2 X_i^m) \implies \varepsilon_i^{/} = \frac{\varepsilon_i}{X_i^{m/2}} \sim N(0, \sigma^2).$$

Importante:

- Quadrados mínimos ponderados implica numa transformação da escala da variável resposta.
- Para se corrigir a não homogeneidade da variância é frequentemente necessário testar diversos valores de m ($w_i = X_i^{-m}$), para se encontrar o peso que de fato homogeniza as variâncias.

3.4 Índice de Furnival

Sempre que realizamos a transformação da variável resposta (através de regressão ponderada ou não), modificamos a escala dos resíduos e, portanto, o QMR de modelos alternativos não são diretamente comparáveis.

Por exemplo: os seguintes modelos são comparados:

(1) $Y = \quad b_0 + b_1 X$

(2) $\ln(Y) = \quad b_0 + b_1 \ln(X)$ Transformação logarítimica

(3) $(Y/X) = \quad b_0(1/X) + b_1$ Peso $= (1/X^2)$

(4) $(Y/\sqrt{X}) = \quad b_0(1/\sqrt{X}) + b_1\sqrt{X}$ Peso $= (1/X)$

O Índice de Furnival é:

$$I = \left[f^{/}(Y)\right]^{-1} \sqrt{QMR}$$

- $[Z]$ é a média geométrica de Z:

$$[Z] = \exp\left(\frac{\sum \ln Z_i}{n}\right)$$

- $f^{/}(Y)$ é a primeira derivada da transformação com respeito a Y.
- Como o Índice de Furnival é uma correção da escala do QMR, quanto **menor** o seu valor, **"melhor"** o ajuste.

No exemplo acima temos:

(1) $f(Y) = Y \quad \Rightarrow f^{/}(Y) = 1 \quad \Rightarrow I = \sqrt{QMR}$

(2) $f(Y) = \ln(Y) \quad \Rightarrow f^{/}(Y) = 1/Y \quad \Rightarrow I = \exp\left(\frac{\sum \ln Y_i}{n}\right)\sqrt{QMR}$

(3) $f(Y) = (Y/X) \quad \Rightarrow f^{/}(Y) = 1/X \quad \Rightarrow I = \exp\left(\frac{\sum \ln X_i}{n}\right)\sqrt{QMR}$

(4) $f(Y) = (Y/\sqrt{X}) \quad \Rightarrow f^{/}(Y) = 1/X^{1/2} \quad \Rightarrow I = \exp\left(\frac{\frac{1}{2}\sum \ln X_i}{n}\right)\sqrt{QMR}$

Note que

$$\begin{aligned} \left[1/Z^k\right] &= \exp\left(\frac{\sum \ln(1/Z^k)}{n}\right) = \exp\left(\frac{-k\sum \ln Z}{n}\right) \\ \left[1/Z^k\right]^{-1} &= \exp\left(\frac{k\sum \ln Z}{n}\right) \end{aligned}$$

Importante: o índice de Furnival é uma correção do QMR para as situações onde a variável resposta foi transformada. Portanto, quanto menor o valor do índice, menor o QMR e, consequentemente, melhor o ajuste.

3.5 Exercícios

3.5.1 Utilizando os dados do arquivo `g:\geral\lcf410\exemplos\biomassa.dat` construa uma equação para **biomassa do tronco** das árvores de *E. saligna* em função da variável combinada DAP^2H, segundo o modelo:

$$Y_i \quad = \quad \beta_0 + \beta_1 X_i + \varepsilon_i$$

Encontre o melhor peso para a regressão ponderada utilizando o gráfico de dispersão dos resíduos e o índice de Furnival.

3.5.2 Referia-se ao exercício **1.1**. Encontre o índice de Furnival para cada um dos modelos. Qual dos modelos apresente o melhor ajuste de acordo com este índice ?

3.5.3 Referia-se ao exercício **1.3**. Encontre o índice de Furnival para cada um dos modelos. Qual dos modelos apresente o melhor ajuste de acordo com este índice ?

4 MATRIZES E REGRESSÃO LINEAR

4.1 Regressão Linear Simples por Matrizes

Embora o modelo linear simples possa ser ajustado pelas fórmulas vistas anteriormente, quando utilizamos duas ou mais variáveis preditoras (modelos lineares múltiplos) as fórmulas se tornam muito complicadas. Nestes casos, a abordagem mais prática é utilizar a algebra de matrizes. Iniciamos apresentando como as matrizes são utilizadas nos modelos lineares simples para depois apresentarmos a sua utilização nos modelos lineares múltiplos.

4.1.1 Representação do Modelo Linear Simples em Matrizes

Como foi visto, o modelo linear simples é:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

onde $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Note que o subscrito i indica que a equação acima se repete para $i = 1, 2, \ldots, n$. O modelo, portanto, pode ser escrito como um sistema de equações da forma:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ &\ldots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned}$$

A algebra de matrizes é particularmente indicada para expressar sistemas de equações lineares, pois é mais compacta. O sistema acima pode ser representado pelas seguintes matrizes:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & Y_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Em notação matricial, este sistema é expresso simplismente como

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times 2)}{\boldsymbol{X}} \underset{(2\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

onde

$\boldsymbol{Y}$ é o vetor das observações da variável resposta.

$\boldsymbol{X}$ é chamada de *matrix de delinearmento* e tem na primeira coluna some o número 1 e na segunda os valores da variável preditora X.

$\boldsymbol{\beta}$ é o vetor dos parâmetros (β_0 e β_1).

$\boldsymbol{\varepsilon}$ é dos erros.

4.1.2 Exemplo: Relação DAP-Altura em E. grandis

Utilizando o nosso exemplo da relação DAP-altura em *E.grandis*, esta fórmula para cada árvore formaria o seguinte sistema:

$$h_i = \beta_0 + \beta_1\, d_i + \varepsilon_i$$

$$\begin{aligned}
27 &= \beta_0 + \beta_1\, 18.1 + \varepsilon_1 \\
26 &= \beta_0 + \beta_1\, 13.7 + \varepsilon_2 \\
30 &= \beta_0 + \beta_1\, 15.6 + \varepsilon_3 \\
13 &= \beta_0 + \beta_1\, 5.7 + \varepsilon_4 \\
28 &= \beta_0 + \beta_1\, 15.0 + \varepsilon_5 \\
31 &= \beta_0 + \beta_1\, 21.0 + \varepsilon_6 \\
23 &= \beta_0 + \beta_1\, 12.1 + \varepsilon_7 \\
29 &= \beta_0 + \beta_1\, 16.6 + \varepsilon_8 \\
28 &= \beta_0 + \beta_1\, 14.3 + \varepsilon_9 \\
32 &= \beta_0 + \beta_1\, 18.8 + \varepsilon_{10} \\
24 &= \beta_0 + \beta_1\, 13.7 + \varepsilon_{11} \\
26 &= \beta_0 + \beta_1\, 15.6 + \varepsilon_{12} \\
28 &= \beta_0 + \beta_1\, 18.1 + \varepsilon_{13} \\
16 &= \beta_0 + \beta_1\, 8.6 + \varepsilon_{14} \\
27 &= \beta_0 + \beta_1\, 12.7 + \varepsilon_{15} \\
28 &= \beta_0 + \beta_1\, 20.7 + \varepsilon_{16}
\end{aligned}$$

$$\begin{aligned} 21 &= \beta_0 + \beta_1\, 20.7 + \varepsilon_{17} \\ 27 &= \beta_0 + \beta_1\, 12.7 + \varepsilon_{18} \end{aligned}$$

Este sistema de 18 equações, cada uma representando uma árvore pode ser representado matricialmente da seguinte maneira:

$$\begin{bmatrix} 27 \\ 26 \\ 30 \\ 13 \\ 28 \\ 31 \\ 23 \\ 29 \\ 28 \\ 32 \\ 24 \\ 26 \\ 28 \\ 16 \\ 27 \\ 28 \\ 21 \\ 27 \end{bmatrix} = \begin{bmatrix} 1 & 18.1 \\ 1 & 13.7 \\ 1 & 15.6 \\ 1 & 5.7 \\ 1 & 15.0 \\ 1 & 21.0 \\ 1 & 12.1 \\ 1 & 16.6 \\ 1 & 14.3 \\ 1 & 18.8 \\ 1 & 13.7 \\ 1 & 15.6 \\ 1 & 18.1 \\ 1 & 8.6 \\ 1 & 12.7 \\ 1 & 20.7 \\ 1 & 20.7 \\ 1 & 12.7 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{17} \\ \varepsilon_{18} \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

onde

$\boldsymbol{Y}$ é o *vetor coluna* com as alturas;

$\boldsymbol{X}$ é a matrix com a primeira coluna preenchida com o valor 1, e a segunda com os valores dos DAPs;

$\boldsymbol{\beta}$ é o vetor coluna com os parâmetros do modelo; e

ε é o vetor coluna com os erros.

4.1.3 Método dos Quadrados Mínimos

Vimos que as estimativas dos parâmetros do modelo são encontradas minimizando a Soma do Quadrado dos Resíduos (SQR). Esta solução corresponde a resolver o sistema de Equações Normais que é expresso por:

$$b_0 n + b_1 \sum X_i = \sum Y_i$$

$$b_0 \sum X_i + b_1 \sum X_i^2 \quad = \quad \sum Y_i X_i$$

O sistema de Equações Normais também pode ser organizado nas matrizes

$$\left[\begin{array}{cc} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{array} \right] \left[\begin{array}{c} b_0 \\ b_1 \end{array} \right] \quad = \quad \left[\begin{array}{c} \sum Y_i \\ \sum Y_i X_i \end{array} \right].$$

Na linguagem matricial, o sistema de Equações Normais é compactamente representado por:

$$\boldsymbol{X'Xb} \quad = \quad \boldsymbol{X'Y}.$$

Demonstremos que as matrizes $\boldsymbol{X'X}$ e $\boldsymbol{X'Y}$, de fato representam as somatórias presentes nas Equações Normais:

$$\boldsymbol{X'X} \quad = \quad \left[\begin{array}{ccccc} 1 & 1 & 1 & \ldots & 1 \\ X_1 & X_2 & X_3 & \ldots & X_n \end{array} \right] \times \left[\begin{array}{cc} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_n \end{array} \right] = \left[\begin{array}{cc} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{array} \right]$$

$$\boldsymbol{X'Y} \quad = \quad \left[\begin{array}{ccccc} 1 & 1 & 1 & \ldots & 1 \\ X_1 & X_2 & X_3 & \ldots & X_n \end{array} \right] \times \left[\begin{array}{c} Y_1 \\ Y_2 \\ Y_3 \\ \ldots \\ Y_n \end{array} \right] = \left[\begin{array}{c} \sum Y_i \\ \sum X_i Y_i \end{array} \right]$$

As estimativas de quadrados mínimos dos coeficientes de regressão são obtidas solucionando o sistema de Equações Normais.

$$\begin{aligned} \boldsymbol{[X'X]b} &= \boldsymbol{X'Y} \\ \boldsymbol{[X'X]^{-1}[X'X]b} &= \boldsymbol{[X'X]^{-1}X'Y} \\ \boldsymbol{Ib} &= \boldsymbol{[X'X]^{-1}X'Y} \\ \boldsymbol{b} &= \boldsymbol{[X'X]^{-1}X'Y} \end{aligned}$$

Demonstremos que esta solução matricial é a mesma já obtida para os valores de b_0 e b_1:

$$\boldsymbol{X'X} \left[\begin{array}{cc} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{array} \right] \quad \Longrightarrow \quad \boldsymbol{[X'X]^{-1}} = \frac{1}{n \sum X_i^2 - (\sum X_i)^2} \times \left[\begin{array}{cc} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{array} \right]$$

Note que

$$n\sum X_i^2 - (\sum X_i)^2 \quad = \quad n\left[\sum X_i^2 - (\sum X_i)^2/n\right] = nS_{XX}$$

O produto das matrizes é

$$\begin{aligned} \boldsymbol{X'Y} &= \begin{bmatrix} \sum Y_i \\ \\ \sum X_i Y_i \end{bmatrix} \\ [\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y} &= \begin{bmatrix} \sum X_i^2/nS_{XX} & -\sum X_i/nS_{XX} \\ \\ -\sum X_i/nS_{XX} & n/nS_{XX} \end{bmatrix} \times \begin{bmatrix} \sum Y_i \\ \\ \sum X_i Y_i \end{bmatrix} \end{aligned}$$

o que resulta em

$$[\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y} \quad = \quad \begin{bmatrix} \left[\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i\right]/nS_{XX} \\ \\ \left[n\sum X_i Y_i - \sum X_i \sum Y_i\right]/nS_{XX} \end{bmatrix} = \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Desenvolvendo as expressões para cada estimativa temos:

$$b_1 \quad = \quad \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{nS_{XX}} = \frac{n\left[\sum X_i Y_i - (\sum X_i \sum Y_i)/n\right]}{nS_{XX}} = \frac{nS_{XY}}{nS_{XX}} = \frac{S_{XY}}{S_{XX}}$$

$$\begin{aligned} b_0 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{nS_{XX}} \\ &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i + (\sum X_i)^2 \sum Y_i/n - (\sum X_i)^2 \sum Y_i/n}{nS_{XX}} \\ &= \frac{\sum Y_i \left[\sum X_i^2 - (\sum X_i)^2/n\right] - \sum X_i \left[\sum X_i Y_i - \sum X_i \sum Y_i/n\right]}{nS_{XX}} \\ &= \frac{\sum Y_i \left[S_{XX}\right] - \sum X_i \left[S_{XY}\right]}{nS_{XX}} \\ &= \frac{S_{XX}}{S_{XX}}\frac{\sum Y_i}{n} - \frac{S_{XY}}{S_{XX}}\frac{\sum X_i}{n} = \frac{\sum Y_i}{n} - b_1\frac{\sum X_i}{n} = \overline{Y} - b_1\overline{X} \end{aligned}$$

4.1.4 Exemplo: Relação DAP-Altura em E. grandis

No exemplo da relação hipsométrica de *E. grandis*, temos as seguintes matrizes:

$$\begin{aligned} [\boldsymbol{X'X}] &= \begin{bmatrix} 18 & 273.70 \\ 273.70 & 4449.23 \end{bmatrix} \\ [\boldsymbol{X'X}]^{-1} &= \begin{bmatrix} 4449.23/5174.45 & -273.70/5174.45 \\ \\ -273.70/5174.45 & 18/5174.45 \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
\boldsymbol{X'Y} &= \begin{bmatrix} 464 \\ \\ 7298.6 \end{bmatrix} \\
[\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y} &= \begin{bmatrix} 4449.23/5174.45 & -273.70/5174.45 \\ \\ -273.70/5174.45 & 18/5174.45 \end{bmatrix} \times \begin{bmatrix} 464 \\ \\ 7298.6 \end{bmatrix} \\
\boldsymbol{b} &= [\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y} = \begin{bmatrix} 12.9115 \\ \\ 0.8461 \end{bmatrix}
\end{aligned}
$$

Assim, vemos que por fórmula e por matrizes obtemos as mesmas estimativas de quadrados mínimos para os parâmetros do modelo (as diferenças são devido aos problemas de arredondamento). A algebra matricial, no entanto, é bem tem notação bem mais compacta e conveniente. As operações trabalhosas de inversão e multiplicação de matrizes podem ser programadas para serem realizadas por computadores.

4.2 Um Modelo de Regressão Linear Múltipla

Vejamos agora um modelo linear múltipla com duas variáveis preditoras:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Novamente este modelo representa um sistema de equações

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \varepsilon_2 \\
&\dots \\
Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \varepsilon_n
\end{aligned}
$$

o qual pode ser organizado nas matrizes:

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}
$$

Este sistema pode convenientemente ser representado pela mesma notação matricial anterior, alterando-se apenas a dimensão da matrix $\boldsymbol{X}$ e do vetor $\boldsymbol{\beta}$:

$$
\begin{array}{ccccccc}
\boldsymbol{Y} & = & \boldsymbol{X} & \boldsymbol{\beta} & + & \boldsymbol{\varepsilon} \\
(n \times 1) & & (n \times 3) & (3 \times 1) & & (n \times 1)
\end{array}
$$

As estimativas de quadrados mínimos para os parâmetros β_0, β_1 e β_2 são obtidas solucionando o sistema de Equações Normais

$$\underset{(3\times 3)}{\boldsymbol{X'X}} \quad \underset{(3\times 1)}{\boldsymbol{b}} \quad = \quad \underset{(3\times 1)}{\boldsymbol{X'Y}}$$

o qual difere do caso da regressão linear simples apenas pela dimensão das matrizes envolvidas. A solução que gera as estimativas de quadrados mínimos, no entanto, permanece a mesma

$$\boldsymbol{b} \;=\; [\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y}$$

4.2.1 *Exemplo: Relação DAP-Altura em* E. grandis

No exemplo de *E. grandis* esse modelo poderia representar a seguinte relação hipsométrica, por exemplo:

$$h_i \;=\; \beta_0 + \beta_1 d_i + \beta_2 d_i^2 + \varepsilon_i$$

A diferença está na forma da matrix $\boldsymbol{X}$ (matrix de delineamento) e do vetor $\boldsymbol{\beta}$:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 18.1 & 327.61 \\ 1 & 13.7 & 187.69 \\ 1 & 15.6 & 243.36 \\ 1 & 5.7 & 32.49 \\ 1 & 15.0 & 225.00 \\ 1 & 21.0 & 441.00 \\ 1 & 12.1 & 146.41 \\ 1 & 16.6 & 275.56 \\ 1 & 14.3 & 204.49 \\ 1 & 18.8 & 353.44 \\ 1 & 13.7 & 187.69 \\ 1 & 15.6 & 243.36 \\ 1 & 18.1 & 327.61 \\ 1 & 8.6 & 73.96 \\ 1 & 12.7 & 161.29 \\ 1 & 20.7 & 428.49 \\ 1 & 20.7 & 428.49 \\ 1 & 12.7 & 161.29 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

As operações matriciais resultam nas seguintes matrizes:

$$\boldsymbol{X'X} \;=\; \begin{bmatrix} 18.00 & 273.70 & 4449.23 \\ 273.70 & 4449.23 & 75803.26 \\ 4449.23 & 75803.26 & 1338533.04 \end{bmatrix}$$

$$\left[\boldsymbol{X'X}\right]^{-1} \;=\; \begin{bmatrix} 5.25210087 & -0.729957035 & 0.0238808569 \\ -0.72995703 & 0.107847269 & -0.0036812147 \\ 0.02388086 & -0.003681215 & 0.0001298411 \end{bmatrix}$$

$$\boldsymbol{X'Y} \;=\; \begin{bmatrix} 464.0 \\ 7298.6 \\ 120708.1 \end{bmatrix}$$

As estimativas de quadrados mínimos para os parâmetros são:

$$\boldsymbol{b} = \left[\boldsymbol{X}'\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 5.25210087 & -0.729957035 & 0.0238808569 \\ -0.72995703 & 0.107847269 & -0.0036812147 \\ 0.02388086 & -0.003681215 & 0.0001298411 \end{bmatrix} \begin{bmatrix} 464.0 \\ 7298.6 \\ 120708.1 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} -8.0772303 \\ 4.0816544 \\ -0.1141228 \end{bmatrix}$$

e a relação hipsométrica ajustada fica:

$$\widehat{h}_i = -8.0772303 + 4.0816544\, d_i - 0.1141228\, d_i^2$$

4.3 Modelo Geral de Regressão Linear Múltipla

Note que utilizando a algebra matricial **o mesmo procedimento** para encontrar as estimativas de quadrados mínimos foi utilizado no caso de uma variável preditoras (regressão linear simples) e no caso de duas variáveis preditoras (regressão linear múltipla). Este procedimento é válido para qualquer número de variáveis preditoras. Assim podemos definir o modelo de regressão linear múltipla como:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i(p-1)} + \varepsilon_i$$

onde

Y_i é a variável resposta;

$\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$ são os p parâmetros do modelo;

$X_1, X_2, \ldots, X_{p-1}$ são as variáveis preditoras $(p-1)$;

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ são os erros.

Este modelo representa um sistema de equações que pode ser organizado nas matrizes:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1(p-1)} \\ 1 & X_{21} & X_{22} & \ldots & X_{2(p-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Utilizando a algebra de matrizes, a notação permanece compacta e os resultados obtidos permanecem válidos:

$$\text{Modelo:} \quad \Rightarrow \quad \underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}} \underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

Equações Normais: $\Rightarrow \quad \boldsymbol{X'Xb = X'Y}$

Estimativas de Quad. Mínimos: $\Rightarrow \quad \boldsymbol{b = [X'X]^{-1}X'Y}$

4.4 Exerícios

4.4.1 Utilizando os dados apresentados no exercício **1.1**, ajuste os modelos abaixo utilizando a algebra de matrizes.

Modelo A: $h_i = \beta_0 + \beta_1 d_i + \varepsilon_i$

Modelo B: $\log(h_i) = \beta_0 + \beta_1 \log(d_i) + \varepsilon_i$

Modelo C: $\log(h_i) = \beta_0 + \beta_1 \dfrac{1}{d_i} + \varepsilon_i$

4.4.2 Utilizando os dados apresentados no exercício **1.1**, represente o sistema de Equações Normais (apresentando as matrizes numéricas sem solucioná-lo) para os seguintes modelos:

Modelo A: $\log(h_i) = \beta_0 + \beta_1 d_i + \beta_2 \log(d_i) + \varepsilon_i$

Modelo B: $\dfrac{1}{h_i} = \beta_0 + \beta_1 d_i + \beta_2 d_i^2 + \varepsilon_i$

5 REGRESSÃO LINEAR MÚLTIPLA

5.1 Algumas Matrizes Especiais

Algumas matrizes utilizadas nos cálculos de quantidades associadas à regressão linear são matrizes sem ligação direta com os dados. São elas:

Matriz Identidade: é uma matrix quadrada denotada por $\boldsymbol{I}$ onde os elementos da diagonal principal são todos **1**, e os demais elementos são **0**. Exemplos:

$$\boldsymbol{I}_{3\times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{I}_{5\times 5} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Matriz $\boldsymbol{J}$: é uma matrix $n \times n$ (quadrada) onde todos os elementos são **1**. Exemplos

$$\boldsymbol{J}_{3\times 3} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad \boldsymbol{J}_{5\times 5} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Matriz $\boldsymbol{H}$: outra matrix especial tem ligação direta com os dados, trata-se da matrix $\boldsymbol{H}$. A partir delas muitas quantias são na regressão definidas, pois ela combina todas as variáveis preditoras:

$$\boldsymbol{H} \quad = \quad \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'$$

A matrix $\boldsymbol{H}$ nos permite mostrar que os valores estimados por qualquer modelo de regressão são na verdade combinações da variável resposta (y) e das variáveis de predição. Vejamos: a partir das equações normais podemos representar os valores esperados pelo modelo de regressão.

$$\boldsymbol{X}'\boldsymbol{X}\,\boldsymbol{\beta} \quad = \quad \boldsymbol{X}'\boldsymbol{Y}$$

$$\begin{aligned} \boldsymbol{b} &= [\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y} \\ \\ \widehat{\boldsymbol{Y}} &= \boldsymbol{X}\,\boldsymbol{b} \\ &= \boldsymbol{X}\,[\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y} \\ &= \boldsymbol{H}\,\boldsymbol{Y} \end{aligned}$$

No exemplo da relação DAP-altura em *E. grandis* a matrix $\boldsymbol{H}$ para o modelo linear simples fica:

$$\boldsymbol{H} = \begin{bmatrix} 1 & 18.1 \\ 1 & 13.7 \\ 1 & 15.6 \\ 1 & 5.7 \\ 1 & 15.0 \\ 1 & 21.0 \\ 1 & 12.1 \\ 1 & 16.6 \\ 1 & 14.3 \\ 1 & 18.8 \\ 1 & 13.7 \\ 1 & 15.6 \\ 1 & 18.1 \\ 1 & 8.6 \\ 1 & 12.7 \\ 1 & 20.7 \\ 1 & 20.7 \\ 1 & 12.7 \end{bmatrix} \times \begin{bmatrix} \frac{4449.23}{(18)(287.4694)} & \frac{-15.2056}{287.4694} \\ \\ \frac{-15.2056}{287.4694} & \frac{1}{287.4694} \end{bmatrix} \times \begin{bmatrix} 1 & 18.1 \\ 1 & 13.7 \\ 1 & 15.6 \\ 1 & 5.7 \\ 1 & 15.0 \\ 1 & 21.0 \\ 1 & 12.1 \\ 1 & 16.6 \\ 1 & 14.3 \\ 1 & 18.8 \\ 1 & 13.7 \\ 1 & 15.6 \\ 1 & 18.1 \\ 1 & 8.6 \\ 1 & 12.7 \\ 1 & 20.7 \\ 1 & 20.7 \\ 1 & 12.7 \end{bmatrix}'$$

5.2 Análise de Variância

Na regressão linear múltipla, a análise de variância representa um teste geral do ajuste do modelo aos dados. Se o modelo ajustado é

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i;p-1} + \varepsilon_i,$$

as hipóteses testadas na análise de variância são:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_\alpha &: \text{nem todos } \beta_k = 0 \quad (k = 1, 2, \ldots, p-1) \end{aligned}$$

A tabela de análise de variância da regressão, como foi visto, tem a seguinte forma:

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Teste F
Modelo	$p-1$	SQM	$QMM = SQM/(p-1)$	QMM/QMR
Resíduo	$n-p$	$SQR = SQT - SQM$	$QMR = SQR/(n-p)$	
Total	$n-1$	SQT		

A partir da soma de quadrados, todos os demais valores podem ser calculados utilizando as demais informações da tabela. As fórmulas matriciais para as somas de quadrado são:

- Soma de Quadrados do Resíduo:

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \\ SQR &= \boldsymbol{e'e} = \boldsymbol{Y'Y} - \boldsymbol{bX'Y} \\ &= \boldsymbol{Y'}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \end{aligned}$$

- Soma de Quadrados do Modelo:

$$\begin{aligned} SQM &= \boldsymbol{b'X'Y} - \left(\frac{1}{n}\right)\boldsymbol{Y'JY} \\ &= \boldsymbol{Y'}\left[\boldsymbol{H} - \left(\frac{1}{n}\right)\boldsymbol{J}\right]\boldsymbol{Y} \end{aligned}$$

- Soma de Quadrados Total:

$$\begin{aligned} SQT &= \boldsymbol{Y'Y} - \left(\frac{1}{n}\right)\boldsymbol{Y'JY} \\ &= \boldsymbol{Y'}\left[\boldsymbol{I} - \left(\frac{1}{n}\right)\boldsymbol{J}\right]\boldsymbol{Y} \end{aligned}$$

O Coeficiente de Determinação é calculado por:

$$R^2 = 1 - \frac{SQR}{SQT}$$

5.3 Propriedades das Estimativas dos Parâmetros

5.3.1 Variância das Estimativas dos Parâmetros

Pelo método de matrizes, obtem-se inicialmente a matrix de Variância-Covariância das Estimativas de Quadrados Mínimos dos parâmetros do modelo:

$$\begin{aligned} \boldsymbol{s^2\{b\}} &= \begin{bmatrix} s^2\{b_0\} & s\{b_0,b_1\} & \dots & s\{b_0,b_{p-1}\} \\ s^2\{b_1,b_0\} & s^2\{b_1\} & \dots & s\{b_1,b_{p-1}\} \\ \vdots & \vdots & & \vdots \\ s^2\{b_{p-1},b_0\} & s\{b_{p-1},b_1\} & \dots & s^2\{b_{p-1}\} \end{bmatrix} \\ &= QMR[\boldsymbol{X'X}]^{-1} \end{aligned}$$

Esta matriz apresenta as variâncias da estimativas dos parâmetros na diagonal principal:

$$s^2\{b_k\} = \left[QMR[\boldsymbol{X'X}]^{-1}\right]_{kk}$$

sendo que os demais elementos representam a co-variância entre as estimativas de diferentes parâmetros.

5.3.2 Exemplo: Relação DAP-Altura em E. grandis

No exemplo da relação DAP-altura em *E. grandis* a matrix de co-variância das estimativas dos parâmetros do modelo linear simples fica:

$$\boldsymbol{s^2\{b\}} = (12.8328) \begin{bmatrix} \frac{4449.23}{(18)(287.4694)} & \frac{-15.2056}{287.4694} \\ \frac{-15.2056}{287.4694} & \frac{1}{287.4694} \end{bmatrix} = \begin{bmatrix} 0.8598 & -0.0529 \\ -0.0529 & 0.0035 \end{bmatrix}$$

Assim as variâncias das estimativas dos parâmetros são:

$$\begin{aligned} s^2\{b_0\} &= 0.8598 \\ s^2\{b_1\} &= 0.0035 \end{aligned}$$

enquanto que a co-variância entre b_0 e b_1 é $s\{b_0,b_1\} = -0.0529$.

5.3.3 Testes de Hipótese Envolvendo os Parâmetros

Assim como na regressão linear simples, as estimativas de quadrados mínimos na regressão linear múltipla têm a seguinte propriedade:

$$b_k \sim N(\beta_k, \sigma^2\{b_k\}),$$

isto é, as estimativas de cada estimativa têm distribuição normal centrada no parâmetro sendo estimado (β_k).

Desta forma, no modelo

$$Y_i \quad = \quad \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i;p-1} + \varepsilon_i$$

que possui $p-1$ variáveis preditoras, é possível se testar as hipótes:

$$\begin{aligned} H_0 \quad &: \quad \beta_k = 0 \\ H_\alpha \quad &: \quad \beta_k \neq 0 \end{aligned}$$

onde $k = 1, 2, \ldots, p$, utilizando o teste t de Student:

$$t^* \quad = \quad \frac{b_k}{\sqrt{s^2\{b_k\}}}$$

com a regra de decisão (ao nível α de significância):

- se $t^* \geq t(1 - \frac{\alpha}{2}; n-p)$ rejeitar H_0;
- se $t^* < t(1 - \frac{\alpha}{2}; n-p)$ **não** rejeitar H_0.

5.4 Interpretação da Regressão Linear Mútipla

O modelo de regressão linear múltipla com duas variáveis preditoras tem a forma

$$Y_i \quad = \quad \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

e a sua interpretação envolve os seguintes aspectos:

- O modelo representa um **plano** no espaço tridimencional definido pelos eixos (Y, X_1, X_2).
- Este plano é geralmente definido como superfície de resposta.
- β_0 é o ponto em que o plano intercepta o eixo-Y ($X_1 = 0$ e $X_2 = 0$).
- $\beta_1 =$ alteração na resposta média que resulta da alteração **em uma unidade** na variável X_1, quando X_2 permanece **constante**.
- $\beta_2 =$ alteração na resposta média que resulta da alteração **em uma unidade** na variável X_2, quando X_1 permanece **constante**.

- **MAS** em geral X_1 e X_2 são **correlacionadas** ($s\{X_1, X_2\} \neq 0$), portanto, se X_1 varia, X_2 também vaira.

  Logo, a interpretação dos parâmetros é "artificial", pois não possível X_1 variar e X_2 permanecer constante (e vice-versa).

A interpretação para um modelo com $p-1$ variáveis preditoras é análoga. Sendo o modelo

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i;p-1} + \varepsilon_i,$$

- a superfície de resposta será um **hiperplano**, isto é, um "plano" no hiper-espaço com p dimensões.
- $\beta_0 =$ ponto onde o hiperplano intercepta o eixo-Y ($X_1 = 0, X_2 = 0, \ldots, X_{p-1} = 0$).
- $\beta_k =$ alteração na resposta média resultante da alteração em **uma unidade** em X_k, quando todas as demais variáveis preditoras permanecem constantes.
- Novamente, esta interpretação é "artificial" pois se as variáveis resposta estiverem correlacionadas será impossível uma delas variar e todas as demais permanecerem constantes.

5.5 Exercícios

5.5.1 Utilizando os dados do arquivo `ESA2-PRD.TXT`, compare os modelos abaixo, escolhendo o mais apropriado para representar a altura das árvores dominantes:

$$\begin{aligned} H_{dom;i} &= \beta_0 + \beta_1 (I_i) + \varepsilon_i \\ H_{dom;i} &= \beta_0 + \beta_1 (I_i) + \beta_2 (I_i)^2 + \varepsilon_i \\ H_{dom;i} &= \beta_0 + \beta_1 (I_i) + \beta_2 (I_i)^2 + \beta_3 (I_i)^3 + \varepsilon_i \end{aligned}$$

Em cada modelo, interprete o significado e a significância estatística das estimativas dos coeficientes de regressão.

Observações:

$H_{dom;i} =$ altura média das ávores dominantes;

$I_i =$ idade;

$\overline{D}_i =$ DAP médio;

$G_i =$ área basal.

5.5.2 Utilizando os dados do arquivo ESA2-PRD.TXT, compare os modelos abaixo, escolhendo o mais apropriado para representar a área basal:

$$\begin{aligned} G_i &= \beta_0 + \beta_1(I_i) + \beta_2(I_i)^2 + \varepsilon_i \\ G_i &= \beta_0 + \beta_1(I_i) + \beta_2 H_{dom;i} + \varepsilon_i \\ G_i &= \beta_0 + \beta_1(I_i) + \beta_2 H_{dom;i} + \beta_3 \overline{D}_i + \varepsilon_i \end{aligned}$$

Em cada modelo, interprete o significado e a significância estatística das estimativas dos coeficientes de regressão.

5.5.3 Utilizando os dados do arquivo ESA2-PRD.TXT, construa um modelo para estimar a produção da floresta de *E. saligna*. Inclua no modelo as variáveis que você julgar mais apropriadas para explicar a produção da floresta.

Após escolher o modelo mais apropriado, interprete o significado e a significância estatística das estimativas dos coeficientes de regressão.