Regressão múltipla

tradução do análise de dados multivariados (Lattin,Carrol & Green)

índice

Introdução

para cada observação i, o valor esperado da variável dependente, condicionado pela informação contida nas variáveis $X_1, ..., X_p$ é

$$E(y_i|x_{i1}, x_{i2}, ..., x_{ip}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

Introduzimis o erro aleatório para refletir a diferença entre o valor real da variável dependente e nossas expectativas:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$$

em termos matriciais:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

onde $\boldsymbol{y}$ e $\boldsymbol{\epsilon}$ $(n \times 1)$ e $\boldsymbol{\beta}$ é um vetor de ordem $[(p+1) \times 1]$

A matrix $\boldsymbol{X}$ é :

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Presupostos do modelo

- 1. A matriz $\boldsymbol{X}$ é fixada de posto completo
- 2. O termo ϵ_i é iid com média 0 e variância σ^2
- 3. $\boldsymbol{X}$ é não correlacionado com ϵ_i.

Propriedades do modelo

Se escolhermos os valores dos parâmetros da regressão para nos dar os valores ajustados mais precisos (R^2 máximo), então a estimativa do parâmetro será dada por:

$$\hat{\beta} = (\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\boldsymbol{y}$$

Esta estimativa, conhecida como de MQ, tem as seguintes propriedades:

- 1. a estimativa de $\hat{\beta}$ é não viesada. $(\boldsymbol{E}[\beta - \hat{\beta}] = 0)$
- 2. A estimativa $\hat{\beta}$ é eficiente. Isto é, $var(\hat{\beta}) = \sigma^2(\boldsymbol{X}`\boldsymbol{X})^{-1}$

A matriz de covariância de $\hat{\boldsymbol{\beta}}$ será útil na inferência sobre os valores dos parâmetros do modelo de regressão.
Os valores ajustados do modelo de regressão, calculados a estimativa de parâmetros $\hat{\boldsymbol{\beta}}$, são dados por

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

As diferenças entre os valores reais $\boldsymbol{y}$ e os valores ajustados $\hat{\boldsymbol{y}}$ são os resíduosm representados por

$$e_i = y_i - \hat{y}_i$$

observe que e_i é uma estimativa do termo erro ϵ_i não observado, baseado em um cálculo envolvendo os coeficientes estimados $\hat{\boldsymbol{\beta}}$ do modelo (em lugar do coeficiente real $\boldsymbol{\beta}$

Estimativa de MQ ordinários

A função objetivo é:

$$\text{escolher } \hat{\boldsymbol{\beta}} \text{ para minimizar } (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})`(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

derivando esta expressão em relação a cada elemento de $\boldsymbol{\beta}$ e igualamos a zero:

$$2\boldsymbol{X}`(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0$$

que resulta em

$$2\boldsymbol{X}`\boldsymbol{y} - 2\boldsymbol{X}`\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

temos então

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\boldsymbol{y}$$

pois sendo X de posto completo, $\boldsymbol{X}`\boldsymbol{X}$ será não singular

Matriz de covariâncias

encontramos a matriz de convariância para o estimador de MQ $\hat{\beta}$:

$$var(\hat{\beta}) = E((\beta - \hat{\beta})(\beta - \hat{\beta})`] \tag{2}$$

Observe que

$$\begin{aligned} \beta - \hat{\beta} &= \beta - (\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\boldsymbol{y} \\ &= \beta - (\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`(\boldsymbol{X}\beta + \boldsymbol{\epsilon}) \\ &= \beta - \beta + (\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\boldsymbol{\epsilon} \end{aligned}$$

ou

$$\beta - \hat{\beta} = (\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\boldsymbol{\epsilon} \tag{3}$$

substituindo 3 em 2 temos:

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}) &= E[(\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\boldsymbol{\epsilon}\boldsymbol{\epsilon}`\boldsymbol{X}(\boldsymbol{X}`\boldsymbol{X})^{-1}] \\ &= (\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}`]\boldsymbol{X}(\boldsymbol{X}`\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}`\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\boldsymbol{X}(\boldsymbol{X}`\boldsymbol{X})^{-1} \end{aligned}$$

de onde temos:

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}`\boldsymbol{X})^{-1} \tag{4}$$

Exemplo

O terreno de Leslie SALt deveria ser adquirida pela cidade de Mountain View, (246,8 acres). Para determinar o valor de mercado justo para a propriedade, avaliadores decidiram construir um modelo de regresão para entender melhor os fatores que poderiam influenciar o valor de mercado. Coletaram dados de 31 propriedades próximas que tinham sido vendidas nos anos anteriores. Além do preço, eles coletaram dados como tamanho, época da venda, elevação, localização e acesso à rede de esgoto.
O objetivo dos especialistas era aproximar-se o mais possível, usando as informações disponíveis das variáveis independentes, para ajustar os preços reais de marcado das 30 propriedades.

Tabela 3.1 Dados da Leslie Salt

	PREÇO	*CONDADO*	*TAMANHO*	*ELEVAÇÃO*	*ESGOTO*	*DATA*	*INUNDAÇÃO*	*DISTÂNCIA*
1	4,5	1	138,4	10	3000	−103	0	0,3
2	10,6	1	52,0	4	0	−103	0	2,5
3	1,7	0	16,1	0	2640	−98	1	10,3
4	5,0	0	1695,2	1	3500	−93	0	14,0
5	5,0	0	845,0	1	1000	−92	1	14,0
6	3,3	1	6,9	2	10000	−86	0	0,0
7	5,7	1	105,9	4	0	−68	0	0,0
8	6,2	1	56,6	4	0	−64	0	0,0
9	19,4	1	51,4	20	1300	−63	0	1,2
10	3,2	1	22,1	0	6000	−62	0	0,0
11	4,7	1	22,1	0	6000	−61	0	0,0
12	6,9	1	27,7	3	4500	−60	0	0,0
13	8,1	1	18,6	5	5000	−59	0	0,5
14	11,6	1	69,9	8	0	−59	0	4,4
15	19,3	1	145,7	10	0	−59	0	4,2
16	11,7	1	77,2	9	0	−59	0	4,5
17	13,3	1	26,2	8	0	−59	0	4,7
18	15,1	1	102,3	6	0	−59	0	4,9
19	12,4	1	49,5	11	0	−59	0	4,6
20	15,3	1	12,2	8	0	−59	0	5,0
21	12,2	0	320,6	0	4000	−54	0	16,5
22	18,1	1	9,9	5	0	−54	0	5,2
23	16,8	1	15,3	2	0	−53	0	5,5
24	5,9	0	55,2	0	1320	−49	1	11,9
25	4,0	0	116,2	2	900	−45	1	5,5
26	37,2	0	15,0	5	0	−39	0	7,2
27	18,2	0	23,4	5	4420	−39	0	5,5
28	15,1	0	132,8	2	2640	−35	0	10,2
29	22,9	0	12,0	5	3400	−16	0	5,5
30	15,2	0	67,0	2	900	−5	1	5,5
31	21,9	0	30,8	2	900	−4	0	5,5

Tabela 3.2 Descrição das variáveis no banco de dados da Leslie Salt

Nome da variável	Descrição
PREÇO	Preço de venda em US$ 000 por acre
CONDADO	San Mateo = 0, Santa Clara = 1
TAMANHO	Tamanho da propriedade em acres
ELEVAÇÃO	Elevação média em pés sobre o nível do mar
ESGOTO	Distância (em pés) da conexão de esgoto mais próxima
DATA	Data da venda, contada retroativamente (em meses) a partir da presente data
INUNDAÇÃO	Sujeita à inundação pela ação da maré = 1, de outro modo = 0
DISTÂNCIA	Distância em milhas da propriedade Leslie (em quase todos os casos em direção a São Francisco)

- Qual deveria ser a variável dependente?
- os tamanhos das propriedades variam amplamente e tende a aumentar com o tamanho da propriedade.
- Podemos modelar o preço por acre (em milhões de dólares) para cada propriedade, no entanto, no histograma abaixo, essa variável é distorcida.

Aplicando Logaritmo?

- A distorção na variável dependente pode ser problemática (pode resultar em problemas com as suposições do modelo).
- podemos lidar com a distorção por meio de uma transformação como $log(y)$ por exemplo.
- O logaritmo tem a vantagem de oferecer uma interpretação clara para a variável transformada $log(Y)$. Mudanças constantes no $log(Y)$ resultam em mudanças constantes de porcentagem em Y.

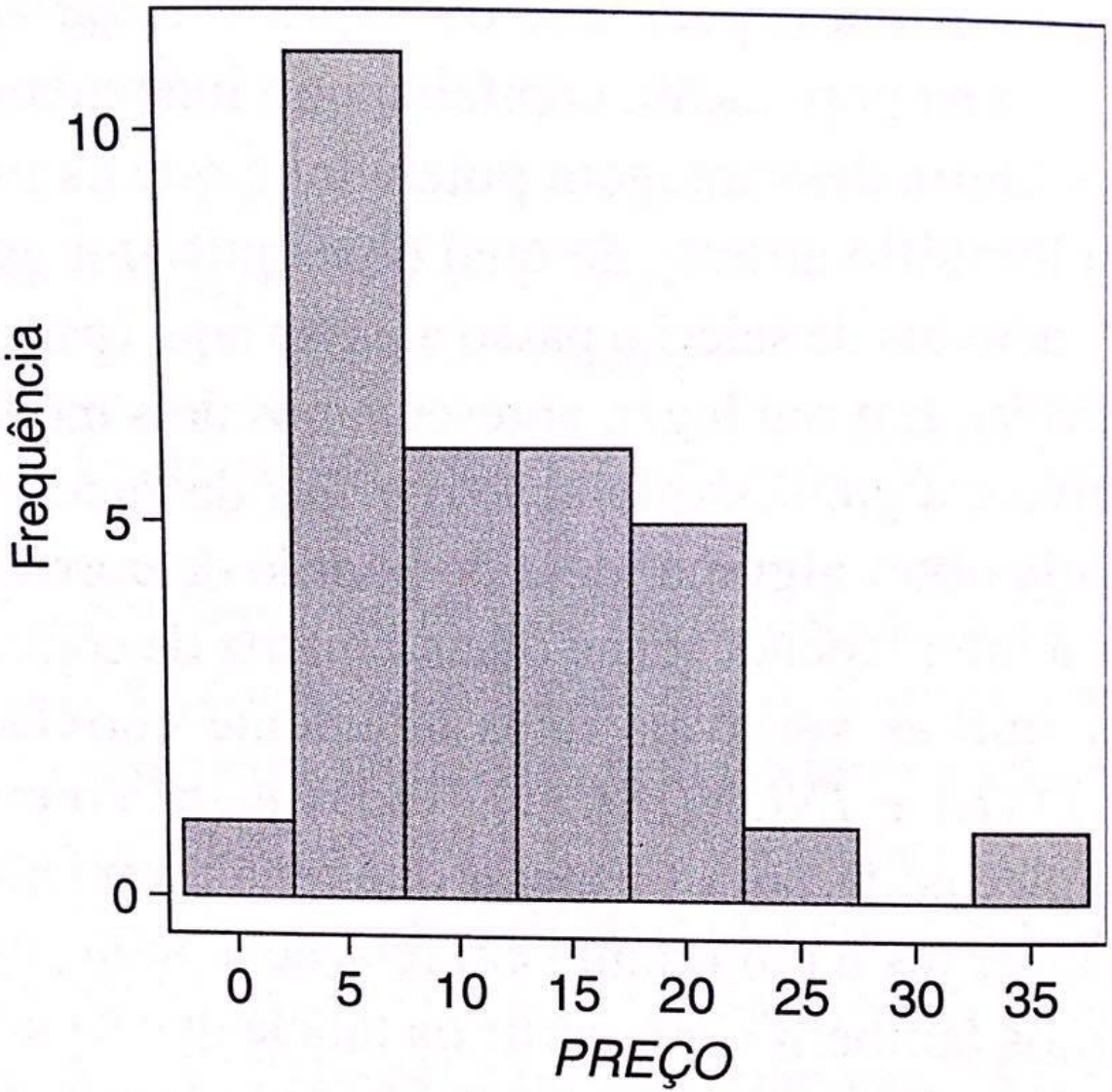


Figura 3.6 Histograma de *PREÇO*.

A distribuição da transformação de log (base 10) do preço por acre está apresentada na figura 3.8. Para fins de ilustração usaremos $Y = log(preco)$ como a variável dependente nas análises.

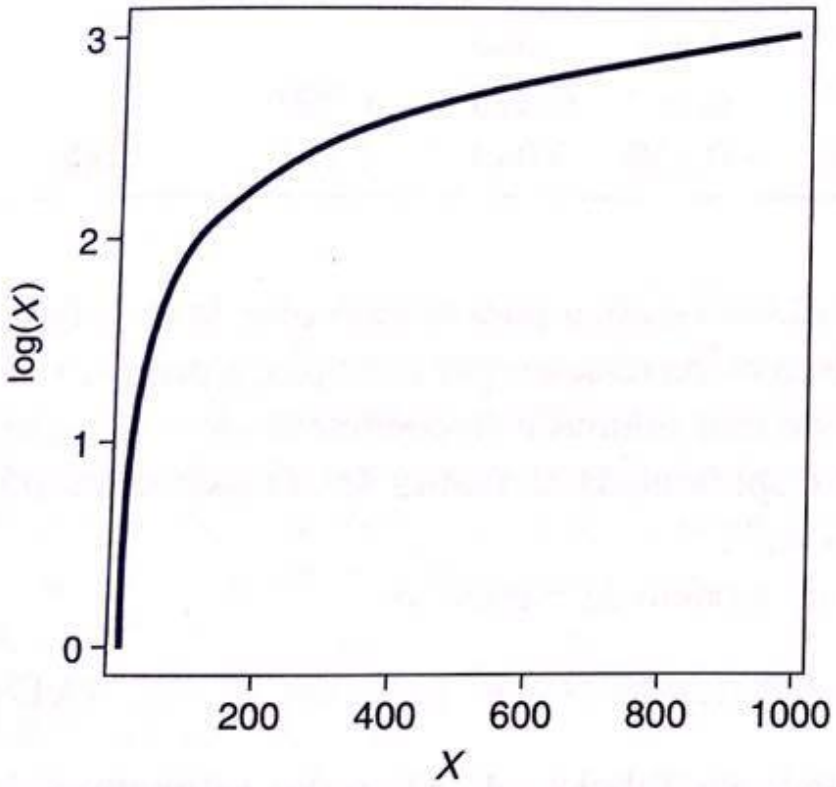


Figura 3.7 Gráfico mostrando o formato da transformação logarítmica.

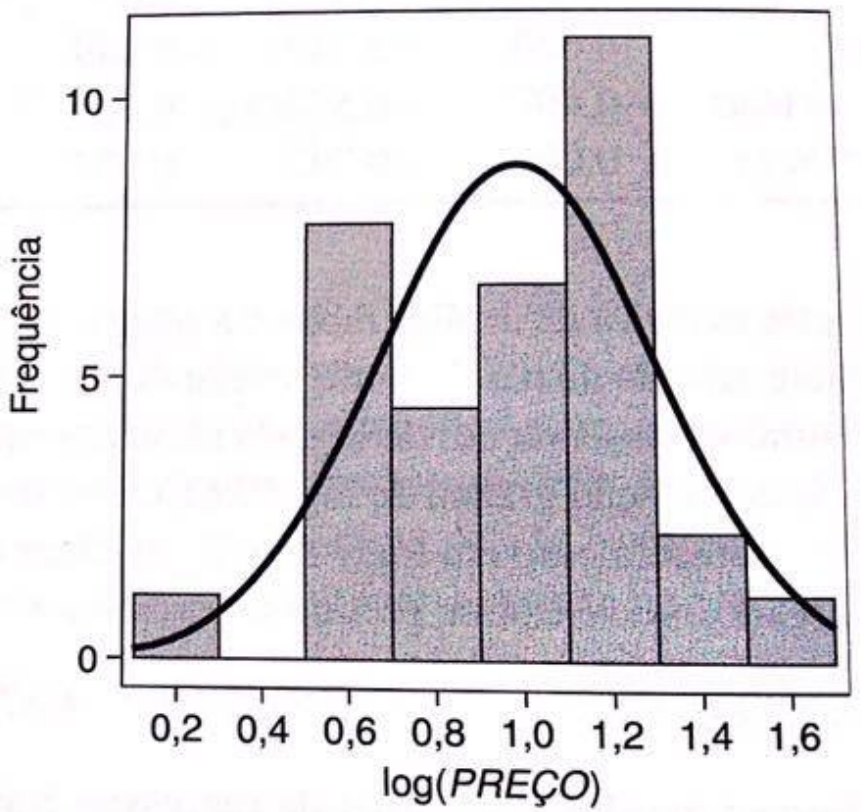


Figura 3.8 Histograma do log(*PREÇO*).

Seleção das variáveis de regressão

Existem dois modelos automatizados passo a passo:

- 1. Seleção para frente (forward): inicia do nada e acrescenta variáveis
- 2. Seleção para trás (backward): começa com todas as variáveis e apaga algumas de acordo com alguns critérios pre-especificados em relação à melhoria no ajuste do modelo.

embora úteis para grande conjunto de dados, a desvantagem é que há uma tendência de superajustar os dados (escolher variáveis com base em sua capacidade de explicar a variação na amostra, que pode ou não ser característica de variação na população).
Outra desvantagem é que o pesquisador pode perder o sentido intuitivo pelos dados.

Uma maneira de observar a interdependência, é olhar a matriz de correlação:

Tabela 3.3 Matriz de correlação para os dados de Leslie Salt

	log(*PREÇO*)	*CONDADO*	*TAMANHO*	*ELEVAÇÃO*	*ESGOTO*	*DATA*	*INUNDAÇÃO*	*DISTÂNCIA*
log(*PREÇO*)	1,000							
CONDADO	−0,044	1,000						
TAMANHO	−0,220	−0,339	1,000					
ELEVAÇÃO	0,433	0,475	−0,209	1,000				
ESGOTO	−0,468	−0,050	0,053	−0,359	1,000			
DATA	0,620	−0,370	−0,349	−0,057	−0,151	1,000		
INUNDAÇÃO	−0,407	−0,552	0,109	−0,373	−0,113	0,015	1,000	
DISTÂNCIA	0,066	−0,742	0,557	−0,362	−0,159	0,044	0,423	1,000

Observe Log(preço) altamente correlacionada com ELEVAÇÃO, ESGOTO, DATA e INUNDAÇÃO. Também podemos ver alta correlação entre as variáveis independentes (ELEVAÇÃO e CONDADO, por exemplo).

outra maneira é avaliar olhando para o diagrama de dispersão (por exemplo (Log(PREÇO) vs DATA):

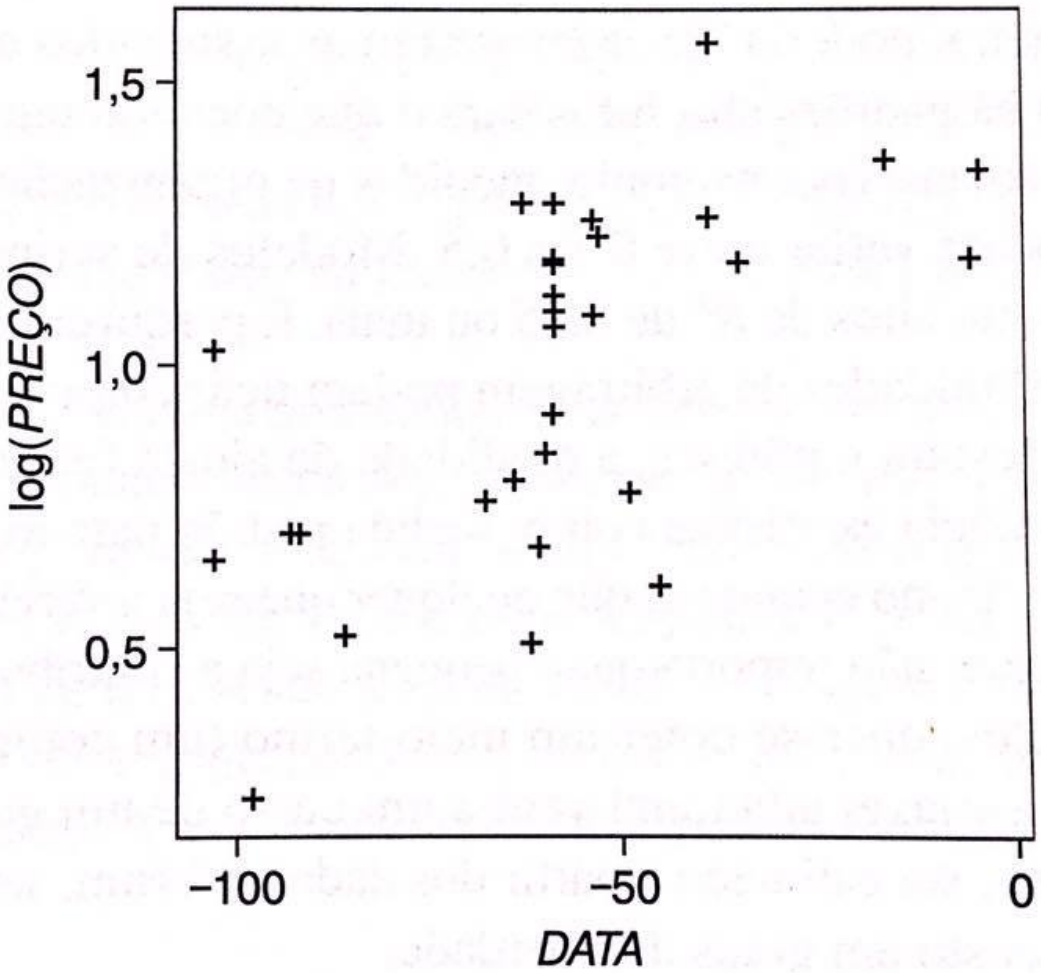


Figura 3.9 Diagrama de dispersão de *DATA versus* log(*PREÇO*).

propomos o seguinte modelo:

$$log(PREÇO) = \beta_0 + \beta_1 ELEVAÇAO + \beta_2 DATA + \beta_3 INUNDAÇÃO + \beta_4 DISTANCIA + \epsilon \quad (5)$$

um resumo dos resultados é:

Tabela 3.4 Resultados do modelo de regressão na Equação (3.17) para os dados de Leslie Salt

	Coeficiente	Erro padrão
Intercepto	1,2266	0,0928
ELEVAÇÃO	0,0324	0,0073
DATA	0,0081	0,0012
INUNDAÇÃO	−0,3383	0,0874
DISTÂNCIA	0,0257	0,0072
$R^2 = 78{,}1\%$		

pode-se observar da tabela acima, que a pesar da variável DISTÂNCIA não ter forte correlação com log(PREÇO), ela é negativamente correlacionada com ELEVAÇÃO e positivamente correlacionada com INUNDAÇÃO. Desta maneria, é possível por exemplo interpretar o coeficiente da variável DISTÂNCIA mantendo constante o efeito das variáveis ELEVAÇÃO, DATA e INUNDAÇÃO. Uma milha mais próxima de São Francisco poderia aumentar o log(PREÇO) em até 0,0257 (em porcentagem: $10^{0.0257} = 1,061$ ou $6,1\%$ por acre).

R^2

R^2 foi descrito como uma medida de variância explicada (ou incerteza reduzida). No caso do exemplo $R^2 = 0,781$ indica um ajuste excelente.

- Não há padrões absolutos sobre que constitui um ajuste aceitável
- dados em ciências sociais: R^2 entre 0,1 e 0,5
- dados de séries históricas $R^2 > 0,95$
- dados de bolsas de valores $R^2 \approx 0,01$, podem-se buscar oportunidades de arbitragem.
- desvantagen: qualquer que seja a variável independente adicionada, o valor de R^2 aumenta. O problema é que cada variável vem a um custo adicional de um grau de liberdade.

R^2_{adj}

$$\bar{R}^2 = R^2_{adj} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2/(n-p-1)}{\sum_i (y_i - \bar{y}_i)^2/(n-1)} \tag{6}$$

em lugar de olhar para o quociente das somas dos desvios ao quadrado, usamos as estimativas da variância da amostra que são responsáveis pelos graus de liberdade usados em cada medida.
Se houver uma melhora insignificante no ajuste, então o aumento em p irá causar aumeno no numerador, reduzindo portanto $\bar{R}^2$ (podendo até ser negativo).
No exemplo $\bar{R}^2 = R^2_{adj} = 0,738$.

Suposições para o ajuste do modelo

As únicas suposições feitas para ϵ eram que os ϵ_i fossem iid. Para fazer inferências sobre a qualidade do ajuste do modelo ou o valor dos parâmetros do modelo suporemos:

$$\boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \tag{7}$$

Teste de significância do modelo

Para testar a significância do modelo geral:

$$H_0 : \boldsymbol{\beta} = \mathbf{0}, \quad vs \quad H_a : \text{ pelo menos um } \beta_i \neq 0$$

utilizamos

$$F = \frac{\sum_i (\hat{y}_i - \bar{y})^2 / p}{\sum_i (y_i - \hat{y}_i)^2 / (n - p - 1)} \sim F_{p, n-p-1} \tag{8}$$

que é o quociente de duas estimativas diferentes de σ^2 (variância do ϵ no modelo de regressão). No denominador usamos a variância dos termos residuias (e): $\sum_i (y_i - \hat{y}_i)^2/(n-p-1)$, no numerador, o erro da média ao quadrado dos valores ajustados da regressão em torno da média: $\sum_i (\hat{y}_i - \bar{y})^2/p$
Quanto maior o valor de F (reflete não somente a variação em e senão também a variação atribuível a Y), maior a probabilidade de rejeitarmos H_0.

ANOVA

Dividimos a variação total (Soma de quadrados) em dois componentes: a soma de quadrados das observações reais ao redor de valores ajustados ($\hat{y}_i$ ao redor de y_i) e a variação dos valores ajustados ao redor da média ($\hat{y}_i$ ao redor de $\bar{y}$).

Tabela 3.5 Análise da tabela de variância para modelo de regressão na Equação (3.21)

	SS	*df*	MS = SS/*dF*	*F*	*p*
Regressão	2,2693	4	0,5673	23,12	0,000
Erro	0,6380	26	0,0245		
Total	2,9072	30			

a estatística F é dada por:

$$F(4, 26) = \frac{0,567}{0,0245} = 23,12$$

o valor crítico na tabela F é $F(4, 26, 0.01) = 4,14$ rejeitamos assim H_0

Teste de parâmetros individuais

Se rejeitarmos H_0, queremos ainda saber quais dos parâmetros são diferentes de zero. O erro padrão do parâmetro é dado pela raíz quadrada da variância na matriz

$$var(\boldsymbol{\beta}) = \sigma^2(\boldsymbol{X}`\boldsymbol{X})^{-1}$$

seja v_{kk} o elemento k da diagonal da $var(\boldsymbol{\beta})$ correspondendo a β_k, então

$$H_0 : \beta_k = 0$$

pode ser testada usando

$$t = \frac{(\hat{\beta}_k - 0)}{\sqrt{v_{kk}}} \sim t_{n-p-1} \tag{9}$$

Tabela 3.6 Estimativas de parâmetro, erros padrão e estatística-t para modelo de regressão na Equação (3.21)

	Coeficiente	**Erro padrão**	t	p
Intercepto	1,2266	0,0928	13,22	0,000
ELEVAÇÃO	0,0324	0,0073	4,43	0,000
DATA	0,0081	0,0012	6,90	0,000
INUNDAÇÃO	−0,3383	0,0874	−3,87	0,001
DISTÂNCIA	0,0257	0,0072	3,58	0,001

Multicolinearidade

Aparece quando existe alta correlação entre as variáveis independentes. O problema aparece na estimativa MQ de $\hat{\beta}$ e de sua matriz de covariância, que são funções de $(\boldsymbol{X}`\boldsymbol{X})^{-1}$. Na presença de alta colinearidade, a matriz inversa $(\boldsymbol{X}`\boldsymbol{X})^{-1}$ fica instável.
Define-se o **indice da condição** (condition index) como:

$$CI = \frac{d_{max}}{d_{min}} \tag{10}$$

que são resultado da decomposição da matriz $\boldsymbol{X}$ nas três componentes matriciais: um conjunto de vetores não correlacionados, uma matriz diagonal e uma matriz de rotação. Os elementos da matriz diagonal são os desvios padrões dos vetores não correlacionados. O IC é a razão entre o maior e o menor desses números.

Um alto valor de CI ($CI > 30$) indica multicolinearidade severa. As soluções passam por: coletar mais dados (pouca ajuda), excluir variáveis do modelo (pode-se perder informação valiosa) ou componentes principais ou análise fatorial (para reduzir a dimensionalidade preservando a informação).

Heterocedasticidade

Acontece heterocedasticidade quando a suposição de variâncias iguais é violada. Acontece isto normalmente quando a magnitude do erro é diretamente relacionada ou à propria variável dependente ou a uma das variáveis independentes.
Uma maneira de lidar com o problema é transformar a variável dependente, dividindo pela variável independente.

MQ ponderados (WLS-weighted least squares)

é outra abordagem para lidar com heterocedasticidade. Suponha que o modelo heterocedástico é

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \tag{11}$$

$$\boldsymbol{\epsilon} \sim \boldsymbol{N}(\boldsymbol{0},\sigma^2\,\boldsymbol{W}) \tag{12}$$

onde $\boldsymbol{W}$ é uma matriz diagonal:

$$\boldsymbol{X} = \begin{pmatrix} w_1 & 0 & 0 & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ 0 & 0 & w_3 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & w_n \end{pmatrix} \tag{13}$$

se multiplicarmos ambos os lados da equação 11, resulta:

$$\boldsymbol{W^{-1/2}y} = \boldsymbol{W^{-1/2}X\beta} + \boldsymbol{W^{-1/2}\epsilon} \tag{14}$$

onde $\boldsymbol{W^{-1/2}}$ é:

$$\boldsymbol{W^{-1/2}} = \begin{pmatrix} 1/\sqrt{w_1} & 0 & 0 & \dots & 0 \\ 0 & 1/\sqrt{w_2} & 0 & \dots & 0 \\ 0 & 0 & 1/\sqrt{w_3} & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1/\sqrt{w_n} \end{pmatrix} \tag{15}$$

sejam

$$\begin{aligned} \boldsymbol{y}^* &= \boldsymbol{W}^{-1/2}\boldsymbol{y} \\ \boldsymbol{X}^* &= \boldsymbol{W}^{-1/2}\boldsymbol{X} \\ \boldsymbol{\epsilon}^* &= \boldsymbol{W}^{-1/2}\boldsymbol{\epsilon} \end{aligned}$$

reescrevemos a equação 14 como

$$\boldsymbol{y}^* = \boldsymbol{X}^*\beta + \boldsymbol{\epsilon}^* \tag{16}$$

Agora a matriz de covariância de $\boldsymbol{\epsilon}^*$ é homocedástica, e utilizamos OLS para estimar $\hat{\beta}$:

$$\hat{\beta} = (\boldsymbol{X}^{*\text{`}}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\text{`}}\boldsymbol{y} \tag{17}$$

reescrevendo $\boldsymbol{X}^*$ e $\boldsymbol{y}^*$ em termos dos dados originais $\boldsymbol{X}$ e $\boldsymbol{y}$ temos:

$$\hat{\beta} = (\boldsymbol{X}\text{`}\,\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\text{`}\,\boldsymbol{W}^{-1}\boldsymbol{y} \tag{18}$$

que é o estimador de mínimos quadrados ponderados.

O efeito é colocar peso baixo ($\sqrt{1/w}$) em observações com grandes variações de erro (w). Os pesos w_i podem ser determinados através de teoría prévia ou pela iteração (por exemplo, começando com OLS e estimando os pesos dos resíduos).

Autocorrelação

Em dados de séries de tempo, os fatores não observados e não mensuráveis que influemciam a variável dependente em algum momento do tempo, podem tender a persistir em alguma extensão para o período seguinte. Como consequência, poderia haver uma correlação positiva entre os erros em períodos consecutivos; i.e, $corr(\epsilon_t, \epsilon_{t+1}) > 0$
A preocupação é que nossa estimativa de $var(\hat{\beta})$ tenha viês para baixo na presença da autocorrelação positiva, isto significa que poderiamos concluir que certos parâmetros do modelo são significativos quando não são.

Maneiras de detetar autocorrelação:

- Desenhar os resíduos versus tempo e buscar um padrão sistemático: **Autocorrelação positiva:** (vários valores positivos, depois vários negativos e novamente positivos). **Autocorrelação negativa**: (os resíduos se alternam para trás e para frente: positivo, negativo, positivo,negativo).
- A estatística de Durbin-Watson (DW):

$$DW = \frac{\sum_t (e_t - e_{t-1})^2}{\sum_t e_t^2} \tag{19}$$

uma boa regra sobre DW é que ela é igual a $2(1-r)$ onde r é a autocorrelação presente nos dados. Portanto, se DW é próximo de 2, não há evidência de autocorrelação. Embora não haja um teste estatístico exato, há tabelas que fornecem limites para determinar se a evidência é conclusiva ou não.

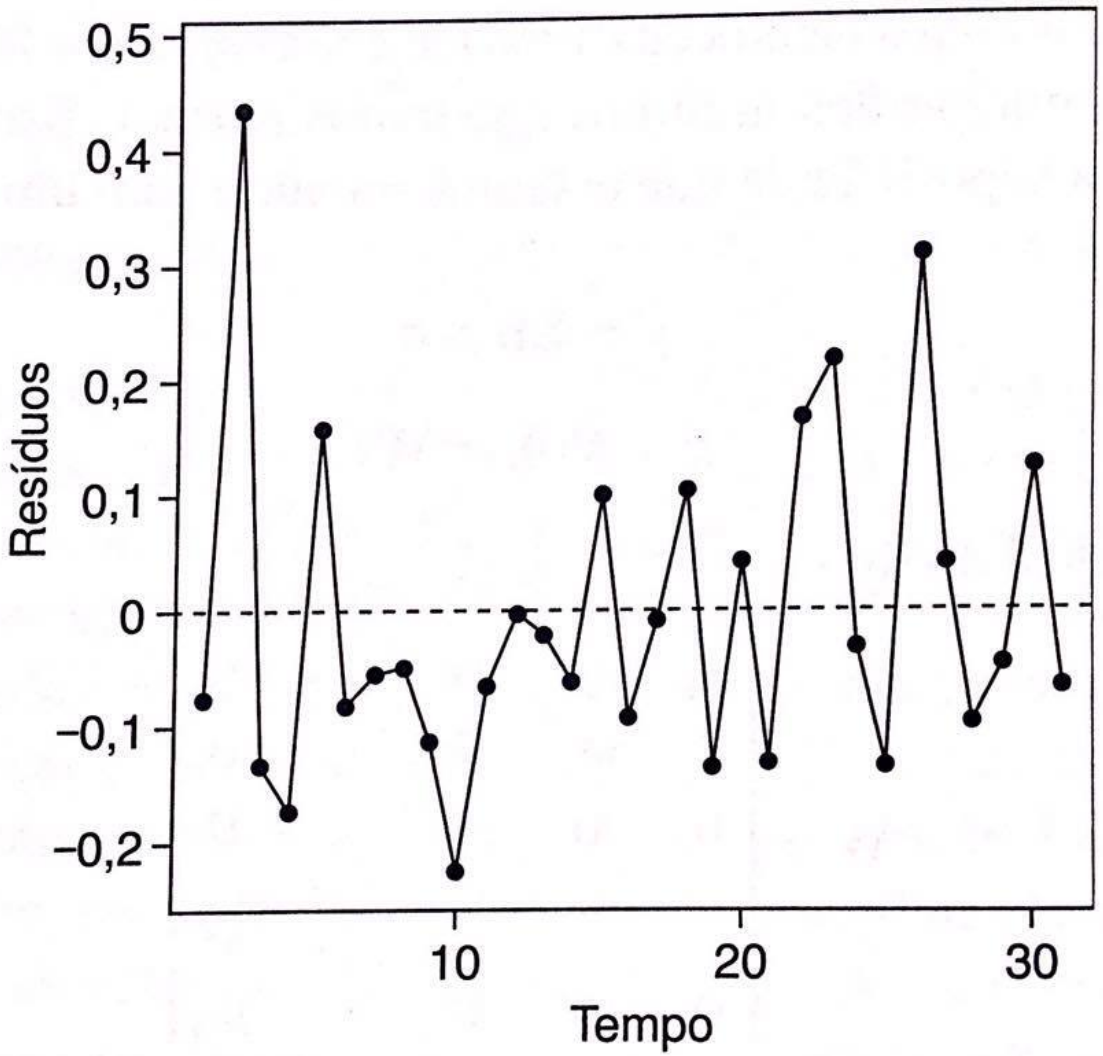
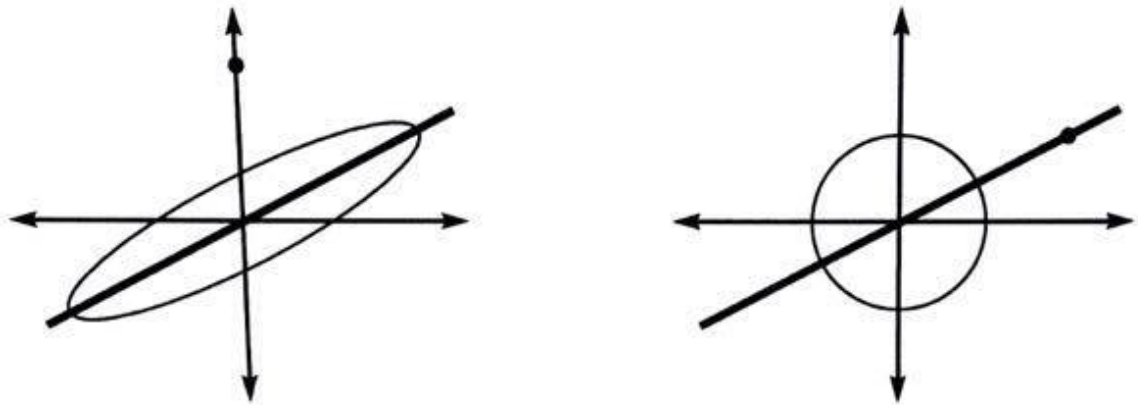


Figura 3.12 Diagrama de dispersão dos resíduos no decorrer do tempo para os dados de Leslie Salt.

Observações influentes

também chamadas de observações discrepantes, em geral assumem uma valor extremo (em X, em Y ou em ambos). Na figura abaixo, podemos observar discrepâncias na variável dependente (não influencia a inclinação estimada da linha de regressão, mas afetará adversamente o ajuste do modelo e a variância estimada do erro) e na variável independente (ela pode determinar a inclinação da reta).



(a) Discrepante em *Y* (b) Observação influente sobre *X*

Figura 3.13 Gráfico representando (a) observação discrepante e (b) observação influente.

Os autores Belsley, Kuh e Welsch (1980) desenvolveram um procedimento baseado no impacto sobre o ajuste, sobre os $\beta`s$ se é removida a observação i. Os diagósticos são baseados na noção de alavancagem: impacto da medida de dependência y_i sobre $\hat{y}_i$:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`\boldsymbol{y} \tag{20}$$

A matriz $\boldsymbol{X}(\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{X}`$ é representada por $\boldsymbol{H}$. A influência de y_i sobre $\hat{y}_i$ reflete-se no i-ésimo elemento da diagonal de $\boldsymbol{H}$ dada por

$$h_{ii} = \boldsymbol{x_i}`(\boldsymbol{X}`\boldsymbol{X})^{-1}\boldsymbol{x_i} \tag{21}$$

onde $\boldsymbol{x_i}`$ é o vetor linha correspondente à observação i e possuem as seguintes propriedades:

- 1. $0 \leq h_{ii} \leq 1$
- 2. $\sum_i h_{ii} = p + 1$

onde p é o número de variáveis independentes.

o valor médio de h_{ii} é $(p+1)/n$. Quando h_{ii} é grande em relação a esse valor médio, os autores sugerem um limite de $2(p+1)/n$ para $p > 10$ e $n - p > 50$, chamamos a observação i como ponto de discrepância. Estudamos também os resíduos estudentizados (úteis para identificar pbservações distantes) e DFBETAs (uma medida da influência da observação nas estimativas dos parâmetros de regressão).

Resíduos estudentizados

Belsley, Kuh e Welsch recomendam:

$$e_i^* = \frac{e_i}{s(i)\sqrt{1 - h_{ii}}} \sim t_{n-p-2} \tag{22}$$

onde h_{ii} é a discrepância do item i e $s(i)$ é o desvio padrão dos resíduos da amostra omitindo a observação i.
Quando i é uma discrepância influente, h_{ii} é grande e $s(i)$ muito menor que s (desvio padrão dos resíduos quando i está presente). O efeito combinado aumenta o valor de e_i^* e a observação se destaca nas representações do resíduo. Podemos também utilizar a tabela t para identificar as discrepâncias significativs (t com $n - p - 2$ g.l.)

DFBETAs

olhamos o impacto de i sobre $\hat{\beta}$. Seja $\boldsymbol{V} = (\boldsymbol{X}`\boldsymbol{X})^{-1}$, a diferença entre a estimativa para $\hat{\beta}$ na presença da observação i e a estimativa para $\hat{\beta}$ na ausência de i é:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = \frac{\boldsymbol{V}\boldsymbol{x_i}\epsilon_i}{(1 - h_{ii})} \tag{23}$$

uma medida padronizada da diferença (DFBETAs) é

$$DFBETAs = \frac{\hat{\beta}_k - \hat{\beta}_k(i)}{s(i)\sqrt{v_{kk}}} \tag{24}$$

onde v_{kk} é o k-ésimo elemento da matriz $\boldsymbol{V}$.
Se i não tem efeito sobre a determinação de $\hat{\beta}_k$ (o coeficiente de inclinação $\hat{\beta}_k$ é o mesmo na presença ou ausência da observação i) então DFBETAs=0. Qualquer observação com um valor de $DFBETAs$ maior que 2, pode ser considerada influente na determinação da estimativa do parâmetro.

No caso do exemplo, observe na tabela a observação 2, o resíduo estudentizado elevado ($e_2^* = 3,7$) é muito maior que qualquer outra observação, e embora não tenha DFBETAs que excedam 2,0, seu impacto sobre a estimativa do parâmetro ELEVAÇÃO é maior.
Dadas estas observações, é necessária uma análise aprofundada por parte dos modeladores para decidir se retirá-la ou não da amostra.

Tabela 3.9 Diagnóstico da regressão para os dados de Leslie Salt

Obs.	h_{ii}	e_i^*	DFBETAs				
			Intercepto	*ELEVAÇÃO*	*DATA*	*INUNDAÇÃO*	*DISTÂNCIA*
1	0,20	−0,52	0,11	0,18	0,07	−0,09	−0,03
2	**0,16**	**3,70**	**−0,44**	**−1,33**	**−0,36**	**−0,40**	**−0,25**
3	0,29	−1,00	0,26	0,36	−0,04	0,06	−0,39
4	0,31	−1,32	0,36	0,44	−0,64	0,17	0,43
5	0,31	1,25	−0,43	−0,39	0,33	0,06	0,43
6	0,16	−0,54	−0,05	0,12	0,14	0,13	0,03
7	0,09	−0,34	−0,05	0,02	0,08	0,04	0,01
8	0,09	−0,31	−0,05	0,01	0,07	0,04	0,01
9	0,48	−0,96	0,28	−0,01	−0,03	−0,86	−0,21
10	0,16	−1,59	−0,44	0,05	0,45	0,52	0,12
11	0,16	−0,43	−0,12	0,01	0,12	0,14	0,03
12	0,10	−0,01	−0,00	0,00	0,00	0,00	0,00
13	0,07	−0,11	−0,02	−0,00	0,02	0,01	0,00
14	0,05	−0,38	−0,00	−0,00	−0,01	−0,05	0,01
15	0,09	0,68	−0,03	0,01	0,03	0,15	−0,01
16	0,07	−0,59	0,01	−0,00	−0,03	−0,10	0,02
17	0,06	−0,04	0,00	−0,00	−0,00	−0,01	0,00
18	0,04	0,70	0,03	−0,00	0,02	0,03	−0,05
19	0,11	−0,89	0,06	−0,01	−0,07	−0,25	0,00
20	0,06	0,30	−0,00	0,00	0,02	0,04	−0,01
21	0,34	−1,01	0,06	−0,01	−0,62	0,14	0,40
22	0,04	1,12	0,09	0,04	0,04	−0,01	−0,10
23	0,06	1,49	0,22	0,05	0,02	−0,22	−0,20
24	0,22	−0,20	−0,00	−0,01	−0,02	0,00	−0,07
25	0,24	−0,95	−0,14	−0,11	0,19	−0,01	−0,47
26	0,08	2,22	0,28	0,33	0,29	0,05	−0,27
27	0,06	0,29	0,05	0,04	0,01	0,00	−0,03
28	0,14	−0,63	−0,10	−0,11	−0,15	0,06	0,13
29	0,14	−0,28	−0,08	−0,09	−0,01	−0,00	0,03
30	0,39	1,05	0,47	0,54	−0,24	0,03	0,59
31	0,22	−0,43	−0,20	−0,19	−0,00	0,06	0,06

Comparação de modelos

Podemos comparar o modelo descrito com um outro modelo onde apareça o efeito interação, onde o impacto de uma variável independente é afetado pelo nível de outra variável independente no modelo.
Quando dois modelos de regressão são aninhados (quando um modelo contém um subconjunto adequado de parâmetros de outro modelo), podemos usar a estatística F para testar a significância do apereiçoamento do ajuste.
Neste caso, testamos o ajuste de um modelo mais geral (f de full) contra um modelo restrito (r de restrito) no qual somente alguns parâmetros são estabelecidos como iguais a zero.

Digamos que $SSE_f = \sum_i (y_i - \hat{y}_i^f)^2$ representa a soma dos erros ao quadrado do modelo completo f, e que $SSE_r = \sum_i (y_i - \hat{y}_i^r)^2$ a soma dos erros ao quadrado do modelo restrito. O teste F para essa comparação é :

$$F = \frac{(SSE - SSE_f)/(df_r - df_f)}{SSE_f / df_f} \tag{25}$$

onde df_f e df_r são o número de g.l. associados ao modelo completo e modelo restrito respectivamente.

Como $1 - R^2$ é diretamente proporcional a SSE, escrevemos o teste em termos de R^2 (não os R^2_{adj}):

$$F = \frac{(R^2_f - R^2_r)/(df_r - df_f)}{(1 - R^2)/df_f} \tag{26}$$

Quando o modelo restrito contém somente um intercepto, então $R^2_r = 0$ e a equação 26 é equivalente ao teste para a significância do modelo geral. Para os dados do exemplo, o R^2 do modelo geral é 0,857 (24 g.l.) e o valor do $R^2_r = 0,781$ (26 g.l.) e o valor de F é

$$F = \frac{(0,857 - 0,781)/2}{(1 - 0,857)/24} = 6,38$$

o valor de F com $F_{2,24} = 5,61$ e concluimos que o efeito de CONDADO ao modelo é significativo.

Previsão

Voltamos ao exemplo. Lembre que a propriedade de Leslie é de aproximadamente 247 acres, localizada em Santa Clara (CONDADO=1), no nível do mar (ELEVAÇÃO=0), não sujeita a inundação pela maré (INUNDAÇÃO=0), afastada de San francisco (DISTANCIA=0), e a ser vendida atualmente (DATA=0):

$$\begin{aligned} Y = & \quad 0,646 + 0,139\ ELEVAÇÃO + 0,008\ DATA - \\ - & \quad 0,132\ INUNDAÇÃO + 0,053\ DISTANCIA + \\ + & \quad 0,561\ CONDADO - 0,115\ CONDADO \times ELEVAÇÃO \\ = & \quad 1,207 \end{aligned}$$

Na verdade estamos interessados em fornecer também um intervalo que diga algo sobre a precisão de nossa previsão e nossa confiança de que o valor real caia dentro do intervalo.
Há duas fontes de erro de variância em nossa previsão. A primeira vem do fato de que nosso modelo reflete uma descrição imperfeita dos dados de nossa amostra.
Quando se trata de previsão fora da amostra, há outra fonte de erro de variância: o erro da amostra.
O erro padrão da previsão considera ambas as fontes de variação. Para a observação $\boldsymbol{x}_0$:

$$se\ f(\boldsymbol{x}_0^{`}) = s\sqrt{\boldsymbol{x}_{\mathbf{0}}^{`}(\boldsymbol{X}^{`}\boldsymbol{X})^{-1}\boldsymbol{x}_{\mathbf{0}} + \mathbf{1}} \tag{27}$$

onde s é o desvio padrão dos resíduos do modelo.
Para os dados de Leslie, o erro padrão da previsão é 0,16, o que significa que um IC de 95% poderia ser grosseiramente mais ou menos dois erros padrão em torno de nosso valor de previsão log(PREÇO) ou $(0,87, 1,54)$

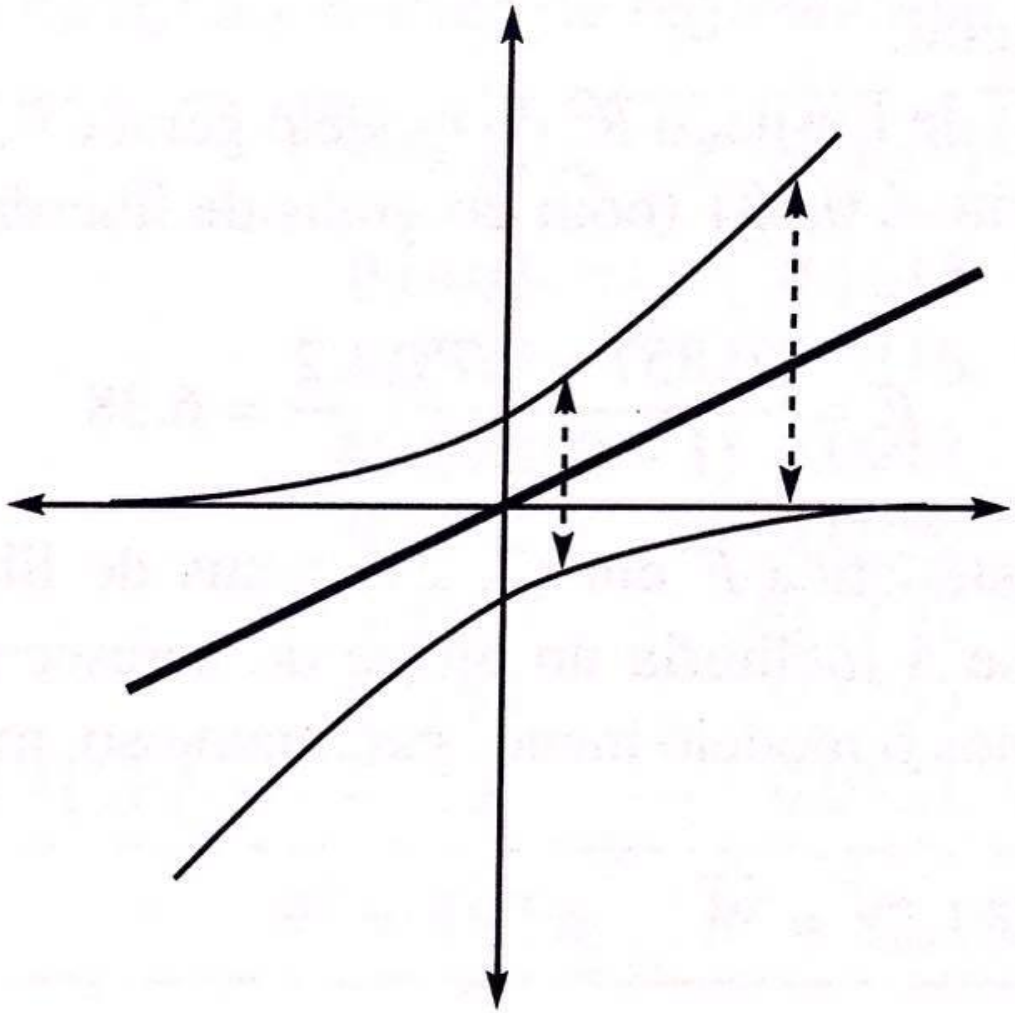
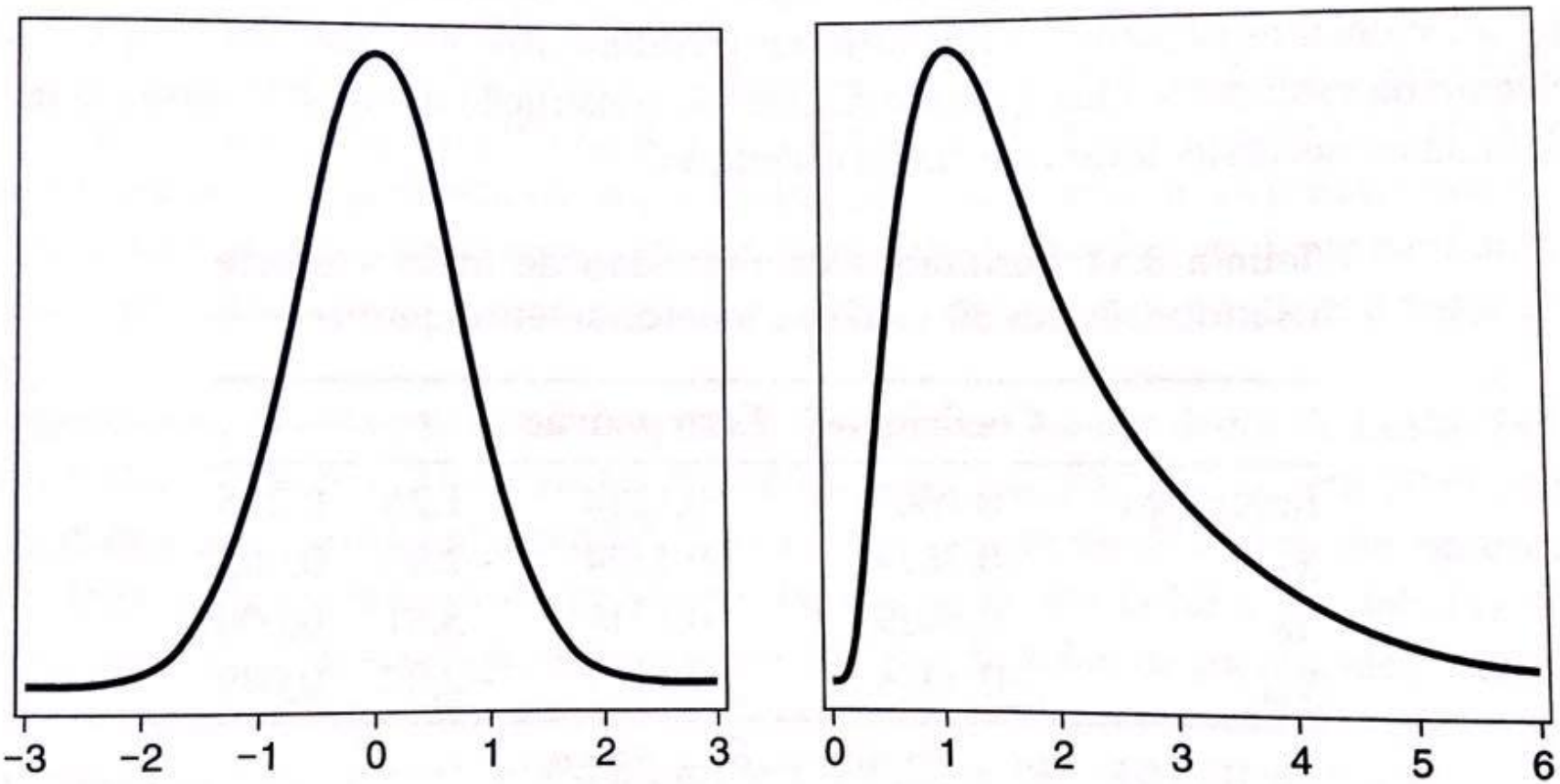


Figura 3.15 Erro padrão da previsão aumenta quando a observação é mais afastada da média.

foi usada a função $log_{10}(Y)$ para transformar a variável dependente. Transformar de volta $10^Y = 10^{0,127} = 16,106$ ou aproximadamente $16,100$ por acre não é rigorosamente correta.
Quando modelamos $log(PREÇO)$ presupomos que $\epsilon \sim N(0, \sigma^2)$. Ao transformar $log(PREÇO)$ de volta a $PREÇO$, $\epsilon \sim Log - Normal$.
A diferença entre a normal e a log-normal é apresentada na figura abaixo.
A fórmula para o valor esperado de uma v.a. log-normal é

$$E(\epsilon) = 10^{(\mu+\sigma^2)/2}$$

onde μ é a média e σ^2 a variável normalmente distribuida. Para a previsão de Leslie, $\mu = 1,207$ e $\sigma^2/2 = (0,13)^2/2 = 0,0087$. Portanto, o valor de nossa previsão é $10^{1,2157} = 16,430$



(a) Distribuição normal

(b) Distribuição log-normal

Figura 3.16 Distribuição normal *versus* distribuição log-normal.

Validação do modelo

Ajuste exagerado: Significa escolher os valores dos parâmetros do modelo para explicar não somente a variação que podemos observar na população como um todo, mas também a variação devida ao ruído. O ajuste do modelo melhora, mas sua capacidade de ajustar as observações fora da amostra torna-se pior porque os padrões captados na amostra não se generalizam para a população.
Isto é o que referimos como capitalizar sobre o acaso.

Validação do modelo

Validação cruzada: Quando o número de observações é suficiente, deve ser possível dividir a amostra (aleatoriamente) em dois conjuntos: Um deles usado para calibrar o modelo (i.e, para estimar os valores dos parâmetros) e outro para validar o modelo (i.e, para testar o desempenho preditivo). Comparamos a precisão preditiva no modelo na amostra de validação à qualidade de ajuste na amostra de calibração e avaliar a extensão da capitalização sobre o acaso

Validação do modelo

Validação Jacknife:
Quando nosso conjunto de dados é pequeno para dividir a amostra (validação cruzada), usamos o método-U (validação jacknife). A idéia é retermos uma observação de cada vez, usar as n-1 observações restantes para estimar os parâmetros do modelo e então usar essas estimativas para calcular o valor predito para a observação retida. O benefício é que nenhum ponto dos dados é usado para ajustar os parâmetros que são também usados para predizeer o valor da variável dependente.
Para o exemplo, os dados são mostrados na tabela abaixo:

Tabela 3.12 Coeficientes estimados da validação jacknife

Cada linha mostra os coeficientes estimados dos dados com a observação particular removida

Obs	Intercepto	*CONDADO*	*ELEVAÇÃO*	*DATA*	*INUNDAÇÃO*	*DISTÂNCIA*	*CONDADO* × *ELEVAÇÃO*
1	0,646	0,563	0,139	0,00786	−0,131	0,053	−0,116
2	0,717	0,543	0,130	0,00932	−0,123	0,054	−0,104
3	0,653	0,540	0,136	0,00761	−0,124	0,052	−0,113
4	0,659	0,519	0,134	0,99741	−0,169	0,053	−0,111
5	0,672	0,545	0,136	0,00796	−0,146	0,051	−0,112
6	0,654	0,555	0,138	0,00771	−0,136	0,051	−0,115
7	0,649	0,560	0,139	0,00784	−0,132	0,052	−0,115
8	0,648	0,561	0,139	0,00785	−0,132	0,053	−0,115
9	0,618	0,605	0,143	0,00788	−0,125	0,055	−0,123
10	0,687	0,566	0,134	0,00788	−0,140	0,050	−0,114
11	0,657	0,563	0,137	0,00786	−0,134	0,052	−0,115
12	0,639	0,564	0,140	0,00784	−0,130	0,053	−0,116
13	0,642	0,563	0,139	0,00784	−0,131	0,053	−0,116
14	0,633	0,581	0,141	0,00793	−0,127	0,054	−0,117
15	0,654	0,551	0,138	0,00778	−0,135	0,052	−0,115
16	0,630	0,584	0,141	0,00795	−0,126	0,054	−0,117
17	0,637	0,574	0,140	0,00790	−0,129	0,053	−0,117
18	0,654	0,550	0,138	0,00781	−0,134	0,052	−0,114
19	0,628	0,582	0,141	0,00798	−0,125	0,054	−0,116
20	0,644	0,564	0,139	0,00786	−0,131	0,053	−0,116
21	0,649	0,559	0,138	0,00786	−0,134	0,053	−0,115
22	0,667	0,529	0,136	0,00771	−0,138	0,051	−0,112
23	0,680	0,506	0,135	0,00765	−0,142	0,050	−0,109
24	0,644	0,562	0,140	0,00782	−0,135	0,053	−0,116
25	0,663	0,549	0,140	0,00786	−0,094	0,050	−0,116
26	0,715	0,514	0,115	0,00810	−0,137	0,050	−0,092
27	0,633	0,566	0,147	0,00772	−0,133	0,053	−0,123
28	0,649	0,559	0,139	0,00786	−0,134	0,053	−0,115
29	0,632	0,582	0,154	0,00796	−0,132	0,053	−0,130
30	0,594	0,559	0,141	0,00706	−0,188	0,054	−0,118
31	0,412	0,729	0,169	0,00705	−0,051	0,061	−0,147

remover algumas observações tem pouco impacto sobre os coeficiente, enquanto outras como a observação 31 tem forte impacto sobre a estimativa do parâmetro para as variáveis CONDADO, ELEVAÇÃO e INUNDAÇÃO e CONDADO x ELEVAÇÃO.
Usamos os coeficientes estimados em cada linha com os dados retidos para essa linha para calcular o valor previsto de $log(PREÇO)$ daquela observação. Calculamos a soma dos erros quadrados entre os valores reais e previstos, e então comparamos as medidas do R^2 (a proporção da variação explicada) para os valores ajustados e previstos.

A soma dos erros quadrados entre os valores reais e previstos é 0,73. Assim, a proporção da variação pelos valores previstos é $(2,91 - 0,73)/2,91 = 75\%$. Este número é mais baixo que o $\bar{R}^2$ (mesmo depois dos ajustes para os graus de liberdade) de 82%, refletindo a capitalização sobre o acaso que ocorreu no ajuste do modelo.
A deteriorização de 82 para 75 é um problema sério?, provavelmente não.