Correlação e Regressão

Notas preparadas por L.A. Bertolo

Índice

Termos básicos e conceitos 1
Regressão simples 5
Regressão Múltipla 13
Terminologia de Regressão 20
Fórmulas de Regressão 21

Termos Básicos e conceitos

1. Um **gráfico de espalhamento** (*scatter plot*) é uma representação gráfica da relação entre duas ou mais variáveis. Num gráfico de espalhamento de duas variáveis x e y, cada ponto no gráfico é um par x-y.

2. Nós usamos regressão e correlação para descrever a variação em uma ou mais variáveis.

A. A **variação** é a soma dos desvios quadrados de uma variável de sua média.

$$\text{Variação} = \sum_{i=1}^{N}(x - \bar{x})^2$$

B. A variação é o numerador da **variância** de uma amostra:

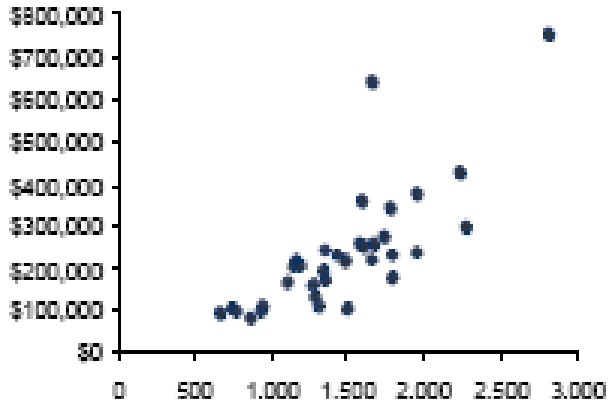
$$\text{Variância} = \frac{\sum_{i=1}^{N}(x - \bar{x})^2}{N - 1}$$

C. Ambas, a variação e a variância, são **medidas da dispersão** de uma amostra.

3. A **covariância** entre duas variáveis aleatórias é uma medida estatística do grau para o qual as duas variáveis se movem juntas.

Exemplo1: Preços de vendas de casas e pés quadrados

Preços de venda de casas (eixo vertical) v. pés quadrados para uma amostra de 34 casas em Setembro de 2005 em St. Lucie County.



A. A covariância captura quanto uma variável é diferente da sua média quando a outra variável for diferente da sua média.

B. Uma covariância positiva indica que as variáveis tendem a se moverem juntas; uma covariância negativa indica que as variáveis tendem a se moverem em direções opostas.

C. A covariância é calculada como a razão da **co-variação** pelo tamanho da amostra menos um:

$$\text{Covariância} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

onde N é o tamanho da amostra
x_i é a i-ésima observação da variável x,
$\bar{x}$ é a média das observações da variável x,
y_i é a i-ésima observação da variável y, e
$\bar{y}$ é a média das observações da variável y.

D. O valor real da covariância não é significante porque ele não é afetado pela a escala das duas variáveis. Isto é o porquê de se calcular o coeficiente de correlação – para tornar algo interpretável da informação da covariância.

E. O coeficiente de correlação, r, é uma medida da intensidade da relação entre ou dentre as variáveis.

Cálculo:

$$r = \frac{\text{covariância entre x e y}}{\left(\begin{matrix}\text{Desvio padrão}\\ \text{de x}\end{matrix}\right)\left(\begin{matrix}\text{Desvio padrão}\\ \text{de y}\end{matrix}\right)}$$

Nota: A correlação não implica que um causa o outro. Podemos dizer que duas variáveis X e Y estão correlacionadas, mas não que X causa Y ou que Y causa X, na média – eles simplesmente estão relacionados ou associados um com o outro.

$$r = \frac{\frac{\left(\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})\right)}{N-1}}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}\sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}}}$$

Observação	x	y	Desvio de x $x - x_{Médio}$	Desvio Quadrado de x $(x - x_{Médio})^2$	Desvio de y $y - y_{Médio}$	Desvio Quadrado de y $(y - y_{Médio})^2$	Produto dos desvios $(x - x_{Médio})(y - y_{Médio})$
1	12	50	-1,50	2,25	8,40	70,56	-12,60
2	13	54	-0,50	0,25	12,40	153,76	-6,20
3	10	48	-3,50	12,25	6,40	40,96	-22,40
4	9	47	-4,50	20,25	5,40	29,16	-24,30
5	20	70	6,50	42,25	28,40	806,56	184,60
6	7	20	-6,50	42,25	-21,60	466,56	140,40
7	4	15	-9,50	90,25	-26,60	707,56	252,70
8	22	40	8,50	72,25	-1,60	2,56	-13,60
9	15	35	1,50	2,25	-6,60	43,56	-9,90
10	23	37	9,50	90,25	-4,60	21,16	-43,70
Soma	135	416	0,00	374,50	0,00	2342,40	445,00

Cálculos

$x_{Médio}$= 135/10 = 13,5

$y_{Médio}$= 416/10 = 41,6

s^2_x= 374,5/9 = 41,611

s^2_y= 2.342,4/9 = 260,267

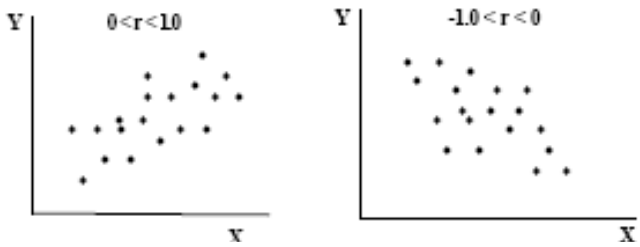
r = $(445/9)/((41,611)^{1/2}(260,267)^{1/2})$ = 49,444/(6,451*16,133) = 0,475

i. O tipo de relação está representada pelo coeficiente de correlação:

r =+1 correlação perfeitamente positiva
+1 >r > 0 relação positiva
r = 0 nenhuma relação
0 > r > –1 relação negativa
r = –1 correlação perfeitamente negativa

ii. Você pode determinar o grau de correlação observando o gráfico de espalhamento.

- Se a relação é para cima existe **correlação positiva**.
- Se a relação é para baixo existe **correlação negativa**.



iii. O coeficiente de correlação está limitado por –1 e +1. Quanto mais próximo o coeficiente estiver de –1 ou +1, mais forte é a correlação.

iv. Com a exceção dos extremos (isto é, r = 1,0 ou r = -1), nós não podemos realmente falar acerca da intensidade de uma relação indicada pelo coeficiente de correlação sem um teste estatístico de significância.

v. As hipóteses de interesse a respeito da correlação da população, ρ, são:

Hipóteses Nulas H0: ρ = 0

Em outras palavras, não existe correlação entre as duas variáveis

Hipóteses Alternativas Ha: ρ =/ 0

Em outras palavras, há uma correlação entre as duas variáveis

Exemplo 2, continuação
No exemplo anterior,
r = 0,475
N = 10

$$t = \frac{\mathbf{0,475\sqrt{8}}}{\sqrt{\mathbf{1-0,475^2}}} = \frac{\mathbf{1,3435}}{\mathbf{0,88}} = \mathbf{1,5267}$$

vi. O teste estatístico está t-distribuído com n-2 graus de liberdade:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

vii. Para tomar uma decisão, compare a estatística-t calculada com a estatística-t crítica para os graus de liberdade apropriados e nível de significância.

Problema
Suponha o coeficiente de correlação como 0,2 e o número de observações como 32. Qual é o teste estatístico calculado? Isto é uma correlação significante usando um nível de significância de 5%?

Solução
Hipóteses:
H0: ρ = 0
Ha: ρ ≠ 0

Estatística-t calculada: $t = \frac{o,2\sqrt{32-2}}{\sqrt{1-0,04}} = \frac{0,2\sqrt{30}}{\sqrt{0,96}} = 1,11803$

Graus de liberdade = 32-1 = 31
O valor-t crítico para um nível de significância de 5% e 31 graus de liberdade é 2,042. Então, não existe correlação significante (1,11803 cai entre os dois valores críticos de –2.042 e +2.042).

Problema
Suponha o coeficiente de correlação como 0,80 e o número de observações como 62. Qual é o teste estatístico calculado? Isto é uma correlação significante usando um nível de significância de 1%?

Solução
Hipóteses:
H0: ρ = 0
Ha: ρ ≠ 0

Estatística-t calculada: $t = \frac{o,80\sqrt{62-2}}{\sqrt{1-0,64}} = \frac{0,80\sqrt{50}}{\sqrt{0,36}} = \frac{5,65685}{0,6} = 9,42809$

O valor-t crítico para um nível de significância de 1% e 11 observações é 3,169. Então, a hipótese nula é rejeitada e concluímos que existe correlação significante.

F. Um **valor afastado** (outlier[1]) é um valor extremo de uma variável. O valor afastado deve ser bem grande ou bem pequeno (onde grande e pequeno são definidos relativamente ao restante da amostra).

   i. Um valor afastado deve afetar a estatística da amostra, tanto quanto um coeficiente de correlação. É possível para um valor afastado afetar o resultado, por exemplo, tal que concluímos que existe uma relação significante quando de fato não existe nenhuma ou concluir que não existe relação quando de fato há uma relação.

   ii. O pesquisador deve exercitar o julgamento (e cuidado) quando decidir se inclui ou exclui uma observação.

G. **Correlação espúria** é uma aparência de uma relação quando de fato não existe relação. Valores afastados podem resultar numa correlação espúria .

   i. O coeficiente de correlação não indica uma relação causal. Certos itens dados podem estar altamente correlacionados, mas não necessariamente um resultado de uma relação causual.

   ii. Um bom exemplo de uma correlação espúria é a caída de neve e os preços de ações em Janeiro. Se fizermos uma regressão histórica dos preços de ações versus o total de caída de neve em Minnesota, obteremos uma relação estatística significante – especialmente para os meses de Janeiro. Desde que não existe uma razão econômica para esta relação, este seria um exemplo de correlação espúria.

Regressão Simples

1. **Regressão** é a análise da relação entre uma variável e alguma outra variável(s), assumindo uma relação linear. Também referida como **regressão dos mínimos quadrados** e **mínimos quadrados ordinários** (*ordinary least squares - OLS*).

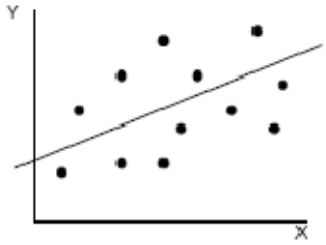
   A. O propósito é explicar a variação numa variável (isto é, como uma variável difere do seu valor médio) usando a variação em uma ou outras mais variáveis.

   B. Suponha que queremos descrever, explicar, ou predizer porque uma variável difere de sua média. Seja a i-ésima observação desta variável representada como Y_i, e seja n indicando o número de observações.

      A variação nos Y_i's (os quais queremos explicar) é:

$$\underset{\text{do Y}}{\text{Variação}} = \sum_{i=1}^{N} (y_i - \bar{y})^2 = SS_{Total}$$

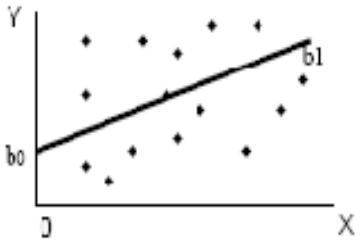
   C. O princípio dos mínimos quadrados é que a linha de regressão é determinada minimizando a soma dos quadrados das distâncias verticais entre os valores reais de Y e os valores previstos de Y.



[1] Uma observação extrema que está bem separada do restante dos dados. Em análise de regressão, nem todos os valores outlying terão uma influência na função de ajuste. Estes outlying com respeito a seus valores X (**alavancagem** alta), e aqueles com valores Y que não são consistentes com a relação de regressão para outros valores (**resíduos** altos) espera-se que sejam influentes. Para testar a influência de tais valores é usada a **estatística Cook**

Uma linha é um ajuste através dos pontos XY tal que a soma dos resíduos quadráticos (isto é, a soma dos quadrados da distância vertical entre as observações e a linha) seja minimizada.

2. As variáveis numa relação de regressão consistem de variáveis dependentes e variáveis independentes.
   A. A **variável dependente** é a variável cuja variação está sendo explicada pela(s) outra(s) variável(s). Também referida como variável explicada, a variável endógena, ou a variável prevista.



   B. A **variável independente** é a variável cuja variação é usada para explicar aquelas da variável dependente. Também referida como a variável explicativa, a variável exógena, ou a variável previsível.

   C. Os parâmetros numa equação de regressão simples são a inclinação (b_1) e o intercepto (b_0):

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

onde y_i é a i-ésima observação da variável dependente,
x_i é a i-ésima observação da variável independente,
b_0 é um intercepto.
b_1 é o coeficiente de inclinação,
ε_i é um resíduo para a i-ésima observação.

   D. A inclinação, b_1, é a variação em Y para uma variação de uma unidade em X. A inclinação pode ser positiva, negativa, ou zero, calculados como:

$$b_1 = \frac{\text{cov}(X,Y)}{\text{var}(x)} = \frac{\frac{\sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})}{N-1}}{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}}$$

Sugestão: Pense na linha de regressão como a média da relação entre a variável independente e a variável dependente . O resíduo representa a distância de quanto um valor observado da variável dependente (i.e., Y) está longe da relação média como descrito pela linha de regressão.

Suponha que:

$$\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x}) = 1.000$$

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = 450$$

N = 30

Então

$$b_1 = \frac{\frac{1.000}{29}}{\frac{450}{29}} = \frac{34,48276}{15,51724}$$

Uma fórmula atalho para o coeficiente de correlação:

$$b_1 = \frac{\frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{N-1}}{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} = \frac{\sum_{i=1}^{N} x_i y_i - \left[\frac{(\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i)}{N}\right]}{\sum_{i=1}^{N} x_i^2 - \left[\frac{(\sum_{i=1}^{N} x_i)^2}{N}\right]}$$

Se isto é realmente um atalho ou não depende do método de realizar os cálculos: manualmente, usando o *Microsoft Excel*, ou usando uma calculadora.

   E. O intercepto, b_0, é a intersecção da linha com o Y- em X=0. O intercepto pode ser positivo, negativo ou zero. O intercepto é calculado como:

$$\hat{b}_0 = \bar{y} - b_1 \bar{x}$$

3. A regressão linear assume o seguinte:

A. Uma relação linear existe entre as variáveis, dependente e independente.

   Nota: se a relação não é linear, pode ser possível transformar uma ou ambas variáveis de modo que exista uma relação linear.

B. A variável independente não está correlacionada com os resíduos; isto é, a variável independente não é aleatória.

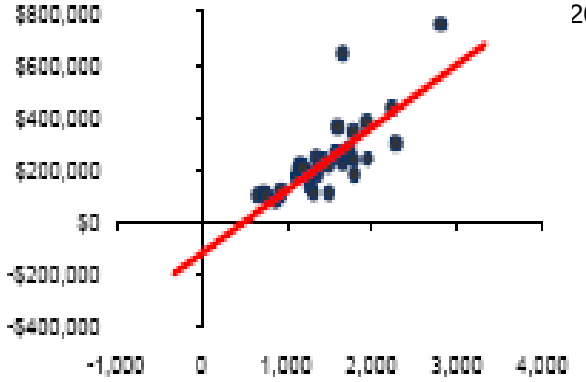
C. O valor esperado do termo distúrbio é zero; isto é, $E(\varepsilon_i)=0$

D. Há uma variância constante do termo distúrbio; isto é, os temos distúrbio ou resíduo são todos extraídos de uma distribuição com uma variância idêntica. Em outras palavras, os termos distúrbios são homoscedásticos. [Uma violação disto é referida como heteroscedasticidade.]

Exemplo 1, continuaçao:

Exemplo 1, continuação:

Preços de vendas (eixo vertical) versus pés quadrados para uma am 2007 em St. L



E. Os resíduos são distribuídos independentemente; isto é, o resíduo ou distúrbio para uma observação não está correlacionado com aquele de outra observação. [Uma violação disto é referida como auto-correlação.]

F. O termo distúrbio (a.k.a. resíduo, a.k.a. error term) é normalmente distribuído.

4. O **erro padrão da estimativa**, SEE, (também referido como o **erro padrão do resíduo** ou **erro padrão da regressão**, e freqüentemente indicado como s_e) é o desvio padrão dos valores previstos da variável dependente ao redor da linha de regressão estimada.

5. Erro padrão da estimativa (SEE) = $\sqrt{s_e^2} = \sqrt{\frac{SS_{Residual}}{N-2}}$

$$SEE = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{b}_0-\hat{b}_i x_i)^2}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{e}_i)^2}{N-2}}$$

Onde $SS_{Residual}$ é a soma dos erros quadráticos;
^ indica o valor predito ou estimado da variável ou parâmetro; e
$\hat{y}_i = \hat{b}_0 - \hat{b}_1 x_i$ é o ponto na linha de regressão correspondente a um valor da variável independente , o x_i; o valor esperado de y, dado a relação média estimada entre x e y.

A. O erro padrão da estimativa ajuda-nos calibrar o "ajuste" da linha de regressão; isto é, quão bem temos descrito a variação na variável dependente.

   i. Quanto menor o erro padrão, melhor o ajuste.

   ii. O erro padrão da estimativa é uma medida da proximidade dos valores estimados (usando a regressão estimada), os $\hat{y}$'s, estão dos valores reais, os Y's.

   iii. Os ε_i's (a.k.a. os termos distúrbios; a.k.a. os resíduos) são as distâncias verticais entre o valor observado de Y e aquele previsto pela equação, os $\hat{y}$ '

iv. Os ε_i's estão nos mesmos termos (unidades de medidas) que os Y's (p.ex, dollars, pounds, billions)

6. O **coeficiente de determinação**, R^2, é a porcentagem da variação da variável dependente (variação dos Y_i's ou a soma dos quadrados total, SST) explicada pela variável independente(s).

A. O coeficiente de determinação é calculado como:

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\text{Variação total} - \text{Variação explicada}}{\text{Variação total}} = \frac{SS_{Total} - SS_{Residual}}{SS_{Total}} = \frac{SS_{Regressão}}{SS_{Total}}$$

Exemplo 2, continuação:

Considere as seguintes observações sobre X e Y:

Observação	x	y
1	12	50
2	13	54
3	10	48
4	9	47
5	20	70
6	7	20
7	4	15
8	22	40
9	15	35
10	23	37
Soma	135	416

A linha de regressão estimada é:

$Y_i = 25{,}559 + 1{,}188\ x_i$

E os resíduos são calculados como:

Observação	x	y	^y	y-^y	e^2
1	12	50	39,82	10,18	103,63
2	13	54	41,01	12,99	168,74
3	10	48	37,44	10,56	111,51
4	9	47	36,25	10,75	115,56
5	20	70	49,32	20,68	427,66
6	7	20	33,88	-13,88	192,65
7	4	15	30,31	-15,31	234,40
8	22	40	51,70	-11,70	136,89
9	15	35	43,38	-8,38	70,22
10	23	37	52,89	-15,89	252,49
				0,00	1.813,77

Portanto,

$SS_{Residual} = 1.813{,}63/8 = 226{,}70$

$SEE = (226{,}70)^{1/2} = 15{,}06$

B. Um R^2 de 0,49 indica que as variáveis independentes explicam 49% da variação da variável dependente.

Exemplo 2, continuação

Continuando o exemplo de regressão anterior, podemos calcular o R^2.

x	y	$(y - y_{Médio})^2$	^y	y-^y	$(^y - y_{Médio})^2$	ε^2
12	50	70,56	39,82	10,18	3,17	103,63
13	54	153,76	41,01	12,99	0,35	168,74
10	48	40,96	37,44	10,56	17,31	111,51
9	47	29,16	36,25	10,75	28,62	115,56
20	70	806,56	49,32	20,68	59,60	427,66
7	20	466,56	33,88	-13,88	59,60	192,65
4	15	707,56	30,31	-15,31	127,46	234,40
22	40	2,56	51,70	-11,70	102,01	136,89
15	35	43,56	43,38	-8,38	3,17	70,22
23	37	21,16	52,89	-15,89	127,46	252,49
	416	2.342,40	416,00	0,00	528,75	1.813,77

R^2 = 528,77/ 2.342,40 = **22,57%** ou

R^2 = 1 – (1.813,63/2.342,40) = 1 – 0,7743 = **22,57%.**

7. Um **intervalo de confiança** é um intervalo de valores de coeficientes de regressão para um dado valor estimado do coeficiente e um dado nível de probabilidade.

A. O intervalo de confiança para um regressão coeficiente $\hat{b}_1$ é calculado como:

$$\hat{b}_1 \pm t_c s_{\hat{b}_1}$$

Ou

$$\hat{b}_1 - t_c s_{\hat{b}_1} < b_1 < \hat{b}_1 + t_c s_{\hat{b}_1}$$

onde t_c é um valor-t crítico para o nível de confiança selecionado. Se existirem 30 graus de liberdades e um nível de confiança 95%, o t_c é 2,042 [tomado de uma tabela-t].

B. A interpretação do intervalo de confiança é que ele é um intervalo que acreditamos que incluirá o parâmetro verdadeiro (1 bˆ S no caso acima) com nível de confiança especificado.

8. Quando o **erro padrão da estimativa** (a variabilidade dos dados ao redor da linha de regressão) subir, a confiança se alarga. Em outras palavras, quanto mais variáveis forem os dados, menos confiante você ficará quando estiver usando o modelo de regressão para estimar o coeficiente.

9. O **erro padrão do coeficiente** é uma raiz quadrada da razão da variância da regressão pela variação da variável independente:

$$S_{\hat{b}_1} = \sqrt{\frac{S_e^{\,2}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

A. Teste de hipóteses: uma variável explicativa individual

i. Para testar hipótese do coeficiente de inclinação (isto é, para ver se a inclinação estimada é igual a um valor hipotético, b_0, H_o: b = b_1, calculamos a estatística t-distribuída:

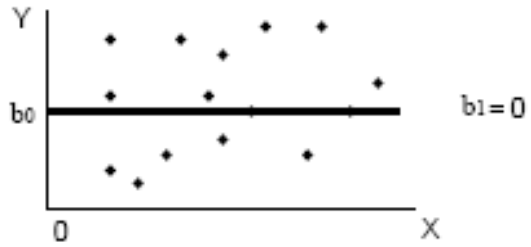
$$t_b = \frac{\hat{b}_1 - b_1}{S_{\hat{b}_1}}$$

ii. O teste estatístico é t–distribuído com N–k–1 graus de liberdade (número de observações (N), menos o número de variáveis independentes (k), menos um).

B. Se a estatística-t é maior que o valor-t crítico para o apropriado grau de liberdade, (ou menor que o valor-t crítico uma inclinação negativa) podemos dizer que o coeficiente de inclinação é diferente do valor hipotético, b_1.

C. Se não existir relação entre a variável dependente e uma variável independente, o coeficiente de inclinação, b_1, será zero.

Nota: A fórmula para o erro padrão do coeficiente tem a variação da variável independente no denominador, não a variância. A variância = variação / n-1.



- Uma inclinação zero indica que não existe variação em Y para uma dada variação em X
- Uma inclinação zero indica que não existe relação entre Y e X.

D. Para testar se uma variável independente explica a variação na variável dependente, a hipótese que é testada é se a inclinação é zero:

Ho: $b_1 = 0$

versus a alternativa (que você conclui se você rejeitar a nula, Ho):

Ha: $b_1 \neq 0$

Esta hipótese alternativa é referida como uma hipótese bilateral. Isto significa que rejeitamos a nula se a inclinação observada é diferente de zero em uma das duas direções (positiva ou negativa).

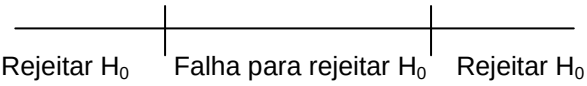
E. Existem hipóteses na economia que se referem ao sinal da relação entre as variáveis dependente e as independentes. Neste caso, a alternativa é direcional (> ou <) e o teste-t é unilateral (usa somente uma cauda da distribuição-t). No caso de uma alternativa unilateral, existe somente um valor-t crítico.

Exemplo 3: Testando a significância de um coeficiente de inclinação

Suponha que o coeficiente de inclinação estimado seja 0,78, o tamanho da amostra seja 26, o erro padrão da coeficiente seja 0.32, e o nível de significância seja 5%. A inclinação é diferente de zero?

O teste estatístico calculado é : tb = $\frac{\hat{b}_{1-}b_1}{S_{\hat{b}_1}} = \frac{0,78-0}{0,32} = 2,4375$

Os valores-t críticos são = 2.060



Portanto, rejeitamos a hipótese nula, concluindo que a inclinação é diferente de zero.

10. Interpretação dos coeficientes.

A. O intercepto estimado é interpretado como o valor da variável dependente (o Y) se a variável independente (o X) tomar um valor zero.

B. O coeficiente estimado de inclinação é interpretado como a variação na variável dependente para uma dada variação de uma unidade na variável independente.

C. Quaisquer conclusões à respeito da importância de uma variável independente na explicação de uma variável dependente exige determinar a significância estatística se o coeficiente inclinar. Simplesmente olhando para a magnitude do coeficiente de inclinação não indica esta matéria de importância da variável.

11. **Previsão** é usar regressão envolve fazer predições acerca da variável dependente baseado nas relações médias observadas na regressão estimada.

A. **Valores preditos** são valores da variável dependente baseado nos coeficientes de regressão estimados e uma predição acerca dos valores das variáveis independentes.

B. Para uma regressão simples, o valor de Y é predito como:

Exemplo 4
Suponha que você estimou um modelo de regressão com as seguintes estimativas:
$\hat{y} = 1,50 + 2,5\ X_1$
Além disso, você tem valores projetados para a variável independente, X_1=20. O valor projetado para y é 51,5:
$\hat{y} = 1,50 + 2,50\ (20) = 1,50 + 50 = 51,5$

$$\hat{y} = \hat{b}_0 + \hat{b}_i x_p$$

onde $\hat{y}$ é um valor previsto da variável dependente, e x_p é um valor previsto da variável independente (input).

12. Uma **análise de tabela de variância** (tabela ANOVA) é um resumo das explicação da variação da variável dependente. A forma básica da tabela ANOVA é como segue:

Fonte de variação	Graus de Liberdade	Soma dos quadrados	Média Quadrática
Regressão (Explicada)	1	Soma das regressões ao quadrado ($SS_{Regressão}$)	Regressão Quadrática Média = $SS_{Regressão+}/1$
Erro (não explicado)	N – 2	Soma dos resíduos ao quadrado ($SS_{Residual}$)	Erro quadrático médio = $SS_{Residuo+}/N-2$
Total	N – 1	Soma dos quadrados total (SS_{Total})	

Exemplo 5 Fonte de variação	Graus de Liberdade	Soma dos quadrados	Média Quadrática
Regressão (Explicada)	1	5.050	5050
Erro (não explicado)	28	600	21.429
Total	29	5.650	
R^2 = 5.050/5.650 = 0,8938 ou 89,38%			
SEE = $(600/28)^{1/2}$ = $(21.429)^{1/2}$ = 4,629			

Regressão Múltipla

1. **Regressão múltipla** é a análise de regressão com mais do que uma variável independente.

   A. O conceito de regressão múltipla é idêntico daquele da análise de regressão simples exceto que duas ou mais variáveis independentes são usadas simultaneamente para explicarem as variações da variável dependente.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$$

   B. Numa regressão múltipla, a meta é minimizar a soma dos erros quadráticos. Cada coeficiente de inclinação é estimado enquanto se mantém as outras variáveis constantes.

Nós não representamos a regressão múltipla graficamente porque ela exigiria gráficos que estão em mais do que duas dimensões.

2. O **intercepto** na equação de regressão tem a mesma interpretação que ela tinha sob o caso linear simples – o intercepto é um valor da variável dependente quando todas as variáveis independentes são iguais a zero.

3. O coeficiente de inclinação é um parâmetro que reflete a variação na variável dependente para uma unidade de variação na variável independente.

   A. Os coeficientes de inclinações (os betas) são descritos como o movimento na variável dependente para uma variação de uma unidade de variação na variável independente – *mantendo todas as outras variáveis independentes constantes*.

   B. Por esta razão, os coeficientes betas numa regressão linear múltipla, são algumas vezes chamados de *betas parciais* ou *coeficientes parciais de regressão*.

Uma inclinação com qualquer outro nome ...

- O coeficiente de inclinação é a elasticidade da variável dependente com respeito à variável independente.
- Em outras palavras, é a *derivada primeira* da variável dependente com respeito à variável independente.

4. Modelo de Regressão:

$$Y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i$$

   onde:

   b_j é a coeficiente de inclinação da j-ésima variável dependente; e
   x_{ji} é a i-ésima observação da j-ésima variável.

   A. Os graus de liberdade para o teste de um coeficiente de inclinação são N-k-1, onde n é um número de observações da amostra e k é um número de variáveis independentes.
   B. Na regressão múltipla, as variáveis independentes podem estar correlacionadas umas com as outras, resultando em estimativas menos confiáveis. Este problema é referido como multi-colinearidade.

5. Um **intervalo de confiança** para uma inclinação da regressão de população numa regressão múltipla é um intervalo centrado na inclinação estimada:

$$\hat{b}_1 \pm t_c s_{\hat{b}_1}$$

   ou

$$\hat{b}_1 - t_c s_{\hat{b}_1} < b_1 < \hat{b}_1 + t_c s_{\hat{b}_1}$$

   A. Este é o mesmo intervalo usado na regressão simples para o intervalo de um coeficiente de inclinação.
   B. Se este intervalo contém zero, concluímos que a inclinação não é estatisticamente diferente de zero.

6. As hipóteses do modelo da regressão múltipla são como segue:

   A. Uma relação linear existe entre as variáveis, dependente e independente.

B. As variáveis independentes não estão correlacionadas com os resíduos; isto é, a variável independente não é aleatória. Além disso, não existe relação linear entre duas ou mais variáveis independentes. [Nota: isto é ligeiramente modificado das hipóteses do modelo de regressão simples.]

C. O valor esperado do termo distúrbio é zero; isto é, $E(\varepsilon_i)=0$

D. Há uma variância constante do termo distúrbio; isto é, os termos distúrbio ou resíduo são todos extraídos de uma distribuição com uma variância idêntica. Em outras palavras, os termos distúrbios são ***homoscedásticos***. [Uma violação disto é referida como ***heteroscedasticidade***[2].]

E. Os resíduos são distribuídos independentemente; isto é, o resíduo ou distúrbio para uma observação não está correlacionado com aquele de outra observação. [Uma violação disto is referida como auto-correlação.]

F. O termo distúrbio (a.k.a. resíduo, a.k.a. error term) é normalmente distribuído.

G. O resíduo (a.k.a. termo distúrbio, a.k.a. error term) éo que não é explicado pelas variáveis independentes.

7. Numa regressão com duas variáveis independentes, o resíduo para a i-ésima observação é:

$$\varepsilon_i = Y_i - (\hat{b}_0 + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i})$$

8. O **erro padrão da estimativa** (SEE) é o erro padrão do resíduo:

$$s_e = SEE = \frac{\sum_{i=1}^{N}(\hat{e}_i)^2}{N-k-1} = \frac{SSE}{N-k-1}$$

9. Os **graus de liberdade**, df, são calculados como:

$$df = \begin{matrix}\text{número de}\\ \text{observações}\end{matrix} - \begin{matrix}\text{número de}\\ \text{variáveis independentes}\end{matrix} - 1 = N - k - 1 = N - (k+1)$$

A. Os graus de liberdade são o número de pedaços de informações independentes que são usadas para estimar os parâmetros de regressão. No cálculo dos parâmetros de regressão, usamos os seguintes pedaços de informações:

- A média da variável dependente.
- A média de cada uma das variáveis independentes.

B. Então,

- se a regressão é uma regressão simples, usamos os dois graus de liberdade na estimação da linha de regressão.
- se a regressão é uma regressão múltipla com quatro variáveis independentes, usamos cinco graus de liberdade na estimação da linha de regressão.

10. **Previsão** (Forecasting) usando regressão envolve fazer predições acerca da variável dependente baseadas nas relações médias observadas na regressão estimada.

---

[2] Em estatística, uma seqüência ou um vetor de variáveis aleatórias é heteroscedástico (heteroskedastistic) se as variaveis aleatórias tiverem variâncias diferentes. O conceito complementar é chamado homocedasticidade (homoscedasticity). (Nota: A ortografia alternativa homo- ou heteroskedasticity é igualmente correta e também é usada freqüentemente). O termo significa "variância diferindo" e vem do Grego "hetero" ('diferente') e "skedastios" ('dispersão').
Quando usar algumas técnicas estatísticas, tais como mínimos quadrados ordinários (ordinary least squares - OLS), várias hipóteses são geralmente feitas. Uma delas é que o termo erro tenha uma variância constante. Isto será verdadeiro se as observações do termo erro forem assumidas serem extraídas de distribuições idênticas. Heteroscedasticidade é uma violação desta hipótese.
Por exemplo, o termo erro poderá variar ou aumentar com cada observação, de certa forma este é o caso freqüente com medidas de seção cruzada ou séries temporais. Heteroscedasticidade é freqüentemente estudada como parte da econometria, que freqüentemente lida com dados exibindo ela.
Com o advento de erros padrões robustos permitindo-nos fazer inferência sem especificar o segundo momento condicional do termo erro, testar a homoscedasticidade condicional não é tão importante quanto no passado.
O econometricista Robert Engle ganhou o 2003 Nobel Memorial Prize for Economics pelos seus estudos sobre análise de regressão na presença de heteroscedasticidade, que conduziu à sua formulação da técnica de modelagem ARCH (Auto Regressive Conditional Heteroscedasticity).

A. Valores Preditos são valores da variável dependente baseados na regressão estimada dos coeficientes e uma predição acerca dos valores das variáveis independentes.

B. Para uma regressão simples, o valor de y é previsto como:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 \hat{x}_1 + \hat{b}_2 \hat{x}_2$$

onde
$\hat{y}$ é o valor previsto da variável dependente,
$\hat{b}_i$ é o parâmetro estimado, e
$\hat{x}_i$ é o valor previsto da variável independente

C. Quanto melhor for o ajuste da regressão (isto é, quanto menor for o SEE), mais confiantes estamos nas nossas predições.

Exemplo 6: Usando informação da análise de variância

Suponha que estamos estimando com o modelo de regressão múltipla que tem cinco variáveis independentes usando uma de 65 observações. Se a soma dos resíduos quadráticos é 789, qual é o erro padrão da estimativa?
Solução
Dado:
$SS_{Residual}$ = 789
N = 65
k = 5
SEE = $\frac{789}{65-5-1} = \frac{789}{59} = 13{,}373$

Cuidado: O intercepto estimado e todas as inclinações estimadas são usadas na predição do valor da variável dependente, mesmo se uma inclinação não for estatisticamente significantemente diferente de zero.

Exemplo 7: Calculando um valor projetado (*forecasted*)
Suponha que você está estimando um modelo de regressão com as seguinte estimativas:
$\hat{Y} = 1{,}50 + 2{,}5\ X_1 - 0{,}2\ X_2 + 1{,}25\ X_3$
Além disso, você tem os valores previstos para as variáveis independentes:
X_1=20 X_2=120 X_3=50
Qual é o valor previsto de y?

Solução

O valor previsto para Y é 90:

$\hat{Y} = 1{,}50 + 2{,}50\ (20) - 0{,}20\ (120) + 1{,}25\ (50)$

$= 1{,}50 + 50 - 24 + 62{,}50 =$ **90**

11. A **estatística-F** é uma medida de quão bem um conjunto de variáveis independentes, como um grupo, explica a variação na variável dependente.

A. A estatística-F é calculada como:

$$F = \frac{\text{Regressão quadrática média}}{\text{Erro médio quadrático}} = \frac{MSR}{MSE} = \frac{\frac{SS_{Regressão}}{k}}{\frac{SS_{Residual}}{N-k-1}} \frac{\sum_{i=1}^{N} \frac{(\hat{y}_i - \overline{y})}{k}}{\sum_{i=1}^{N} \frac{(y_i - \hat{y})}{N-k-1}}$$

B. A estatística-F pode ser formulada para testar *todas as variáveis independentes como um grupo* (a aplicação mais comum). Por exemplo, se existirem quatro variáveis independentes no modelo, as hipóteses são:

H_0: $b_1 = b_2 = b_3 = b_4 = 0$

H_a: no mínimo um $b_i \neq 0$

C. A Estatística-F pode ser formulada para testar subconjuntos de variáveis independentes (para ver se elas tem poder de explicação incremental (incremental explicative power). Por exemplo se existirem quatro variáveis independentes no modelo, um subconjunto poderia ser examinado:

H_0: $b_1 = b_4 = 0$

H_a: b_1 ou $b_4 \neq 0$

12. O **coeficiente de determinação**, R^2, é a porcentagem da variação da variável dependente explicada pelas variáveis independentes.

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação Total}} = \frac{\text{Variação Total} - \text{Variação Inexplicada}}{\text{Variação Total}}$$

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{N}(y - \bar{y})^2} \qquad 0<R^2<1$$

A. Por construção, R^2 varia de 0 a 1,0

B. O **R^2-ajustado** é uma alternativa a R^2:

$$R^2 = 1 - \left(\frac{N-1}{N-k}\right)(1 - R^2)$$

i. O R^2-ajustado é menor que ou igual a R^2 ('igual a' somente quando k=1).

ii. Adicionando variáveis independentes ao modelo o R^2 aumentará. Adicionar variáveis independentes ao modelo pode aumentar ou diminuir o R^2-ajustado (Nota: R^2-ajustado pode ser até negativo).

iii. O R^2-ajustado não tem a explicação "clara " do poder explicativo que o R^2 tem.

13. O propósito da tabela da Análise da Variância (ANOVA) é atribuir a total variação da variável dependente ao modelo de regressão (a fonte de regressão na coluna 1) e os resíduos (a fonte de erro da coluna 1).

A. **SS_{Total}** é a total variação de Y ao redor de sua média ou valor médio (a.k.a. soma dos quadrados total) e é calculada como

$$SS_{Total} = \sum_{i=1}^{N}(y_i - \bar{y})^2$$

onde $\bar{y}$ é a média de Y.

B. **$SS_{Residual}$** (a.k.a. SSE) é a variabilidade isto é não é explicada pela regressão e é calculada como:

$$SS_{Residual} = SSE = \sum_{i=1}^{N}(y_i - \hat{y})^2 = \sum \hat{e}_i$$

onde $\hat{Y}$ é o valor da variável dependente usando a equação de regressão.

C. **$SS_{Regression}$** (a.k.a. $SS_{Explicada}$) é a variabilidade que é explicada pela equação de regressão e é calculada como $SS_{Total} - SS_{Residual}$.

$$SS_{Regressão} = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$$

D. MSE é o erro quadrático médio, ou MSE = $SS_{Residual}$ / (N − k - 1) onde k é o número de variáveis independentes na regressão.

E. MSR é a regressão quadrática média, MSR =$SS_{Regressão}$ / k

Tabela de Análise da Variância (ANOVA)

Fonte	df Graus de Liberdade	SS Soma dos quadrados	SS/df Média Quadrática
Regressão	k	$SS_{Regressão}$	MSR
Erro (não explicado)	N – k - 1	$SS_{Residual}$	MSE
Total	N – 1	SS_{Total}	

$$R^2 = \frac{SS_{Regressão}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

$$F = \frac{MSR}{MSE}$$

14. **Variáveis *Dummy*** são variáveis qualitativas que tomam os valores zero ou um.

   A. A maioria das variáveis independentes representa um fluxo contínuo de valores. Entretanto, Alguma vezes a variável independente é de natureza binária (ela é ou ON ou OFF).

   B. Estes tipos de variáveis são chamadas variáveis *dummy* e aos dados é atribuído um valor de "0" ou "1". Em muitos casos, você aplica o conceito de variável dummy para quantificar o impacto de uma variável qualitativa. Uma variável *dummy* é uma variável dicotômica; isto é, ela toma um valor de um ou zero.

   C. Use uma variável *dummy* a menos que o número de classes (p.ex., se tem três classes, use duas variáveis *dummy*), caso contrário você cairá numa variável *dummy* "emboscada" (multicolinearidade perfeita – hipótese da violação [2]).

   D. Uma variável *dummy* interativa é uma variável *dummy* (0,1) multiplicada por uma variável para criar uma nova variável. A inclinação desta nova variável diz-nos a inclinação incremental.

15. **Heteroscedasticidade** é uma situação em que a variância dos resíduos não é constante em todas as observações.

   A. Uma hipótese da metodologia da regressão é que a amostra é extraída da mesma população, e que a variância dos resíduos é constante nas observações; em outras palavras, os resíduos são homoscedásticos.

   B. Heteroscedasticidade é um problema porque os estimadores não tem a menor variância possível, e portanto o erro padrão dos coeficientes não serão corretos.

16. **Auto-correlação** é uma situação em que os termos de resíduos estão correlacionadas unscom os outros. Isto ocorre freqüentemente em análises de séries temporais.

   A. Auto-correlação aparece geralmente em dados de séries temporais. Se o lucro do ano passado foi maior, isto significa que o lucro deste ano pode ter uma probabilidade maior de ser alto do que ser baixo. Isto é um exemplo de auto-correlação positiva**.** Quando um ano bom for sempre seguido por uma ano ruim, isto é um exemplo de auto-correlação negativa.

   B. Auto-correlação é um problema porque os estimadores não tem a menor variância possível e portanto oerro padrão dos coeficientes não seriam corretos.

17. Multicolinearidade é um problema de alta correlação entre ou dentre duas ou mais variáveis independentes.

   A. Multicolinearidade é uma problema porque

      i. A presença da multicolinearidade pode causar distorções no erro padrão e pode conduzir a problemas com teste significância dos coeficientes individuais, e

      ii. Estimativas são sensíveis às variações nas observações da amostra ou da especificação do modelo.

   B. Se existir multicolinearidade, estamos mais aptos a concluir que uma variável não é importante.

   C. Multicolinearidade está provavelmente presente em certo grau na maioria dos modelos econômicos. **Multicolinearidade perfeita** nos proibirá de estimar os parâmetros de regressão. O caso é então realmente a um dos graus.

18. O significado econômico dos resultados de uma estimação de regressão focaliza principalmente nos coeficientes de inclinação.

A. Os coeficientes de inclinação indicam a variação da variável dependente para uma variação de uma unidade na variável independente. Esta inclinação pode ser então interpretada como uma medida da elasticidade; isto é, a variação em uma variável corresponde a uma variação em outra variável.

B. É possível ter significância estatística, apesar de que não tenha significância econômica (p.ex., retornos anormais significantes associados com um anúncio, mas estes retornos não são suficientes para cobrirem custos de transações).

Para...	**use...**
Testar o papel de uma única variável na explicação da variação da variável dependente	a estatística-t.
Testar o papel de todas as variáveis na explicação da variação da variável dependente	a estatística-F.
Estimar a variação na variável dependente para uma variação de uma unidade na variável independente	o coeficiente de inclinação.
Estimar a variável dependente se todas as variáveis independentes tomarem um valor zero	o intercepto.
Estimar a porcentagem das variações explicadas das variáveis dependentes pelas variáveis independentes	o R^2.
Prever o valor da variável dependente dados os valores estimados da variável independente(s)	A equação de regressão, substituindo os valores estimados da variável independente(s) na equação.

Regressão terminologia

Analysis of variância
ANOVA
Autocorrelação
Coefficient of determination
Confidence interval
Correlation coeficiente
Covariance
Covariation
Cross-sectional
Degrees of freedom
Dependent variável
Explained variável
Explanatory variável
Forecast
Estatística-F
Heteroskedasticity
Homoskedasticity
Invariável dependente
Intercept
Least squares regressão
Mean square error
Mean square regressão
Multicollinearity
Regressão múltipla
Negative correlação
Ordinary least squares
Perfect negative correlação
Perfect positive correlação
Positive correlação
Predicted valor
R^2
Regressão
Residual
Scatterplot
Se
SEE
Simple regressão
Slope
Slope coeficiente
Spurious correlação
$SS_{Residual}$
$SS_{Regression}$
SS_{Total}
Standard error da estimate
Sum of squares error
Sum of squares regressão
Sum of squares total
Time-series
t-statistic
Variância
Variação

Fórmulas de Regressão

$$\text{Variação} = \sum_{i=1}^{N}(x-\bar{x})^2 \qquad \text{Variância} = \frac{\sum_{i=1}^{N}(x-\bar{x})^2}{N-1} \qquad \text{Covariância} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N-1}$$

$$\text{Correlação} \quad r = \frac{\left(\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N-1}\right)}{\sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}}\sqrt{\frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{N-1}}} \qquad t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Regressão

$y_i = b_0 + b_1 x_i + \varepsilon_i$ $\qquad$ $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + \varepsilon_i$

$$b_1 = \frac{\text{cov}(X,Y)}{\text{var}(x)} = \frac{\frac{\sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})}{N-1}}{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}} \qquad \hat{b}_0 = \bar{y} - b_1\bar{x}$$

Testes e intervalos de confiança

$$\text{SEE} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{b}_0 - \hat{b}_i x_i\right)^2}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{e}_i)^2}{N-2}}$$

$$S_{\hat{b}_1} = \sqrt{\frac{S_e{}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

$$t_b = \frac{\hat{b}_1 - b_1}{S_{\hat{b}_1}}$$

$$F = \frac{\text{Regressão quadrática média}}{\text{Erro médio quadrático}} = \frac{\text{MSR}}{\text{MSE}} = \frac{\frac{SS_{\text{Regressão}}}{k}}{\frac{SS_{\text{Residual}}}{N-k-1}} \frac{\sum_{i=1}^{N}\frac{(\hat{y}_i - \bar{y})}{k}}{\sum_{i=1}^{N}\frac{(y_i - \hat{\bar{y}})}{N-k-1}}$$

Previsão

$\hat{b}_1 - t_c s_{\hat{b}_1} < b_1 < \hat{b}_1 + t_c s_{\hat{b}_1}$

$\hat{y} = \hat{b}_0 + \hat{b}_1 \hat{x}_1 + \hat{b}_2 \hat{x}_2$

Análise de Variância

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = SS_{\text{Total}}$$

$$SS_{\text{Residual}} = \text{SSE} = \sum_{i=1}^{N}(y_i - \hat{y})^2 = \sum \hat{e}_i$$

$$SS_{\text{Regressão}} = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$$

$$F = \frac{\text{Regressão quadrática média}}{\text{Erro médio quadrático}} = \frac{\text{MSR}}{\text{MSE}} = \frac{\frac{SS_{\text{Regressão}}}{k}}{\frac{SS_{\text{Residual}}}{N-k-1}} \frac{\sum_{i=1}^{N}\frac{(\hat{y}_i - \bar{y})}{k}}{\sum_{i=1}^{N}\frac{(y_i - \hat{\bar{y}})}{N-k-1}}$$

Regressão

$yi = b_0 + b_1 xi + \varepsilon_i$ $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + \varepsilon_i$