Insper Instituto de Ensino e Pesquisa

Análise de Regressão Linear Simples III

Aula 03

Gujarati e Porter – Capítulos 4 e 5

Wooldridge – Seção 2.5

Suposições, Propriedades e Teste t

Suposições e Propriedades

RLS.1 – O modelo de regressão é linear nos parâmetros

No modelo populacional, a variável resposta *y* está relacionada ao regressor *x* e ao erro ε como

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

em que

β_0 **– parâmetro de intercepto populacional (constante);**

β_1 **– parâmetro de inclinação populacional (constante);**

ε **– erro aleatório não observável.**

Suposições e Propriedades

RLS.2 – Amostragem Aleatória

Temos uma amostra aleatória de n observações

$$(x_i, y_i), \quad i = 1, 2, ..., n,$$

proveniente do modelo populacional descrito em RLS.1.

RLS.3 – Variação amostral no regressor

Os resultados amostrais em x, ou seja, $\{x_i, i = 1, 2, ..., n\}$ não são todos de mesmo valor.

Suposições e Propriedades

RLS.4 – Média Condicional Zero

O termo de erro aleatório, ε, tem valor esperado igual a zero, dado qualquer valor do regressor. Ou seja,

$$E(\varepsilon \mid x) = 0.$$

Teorema 1. Sob as suposições RLS.1 a RLS.4, condicional aos valores amostrais do regressor, os estimadores de MQO dos parâmetros do modelo de regressão linear simples são não-viesados, ou seja, $E(\hat{\beta}_j) = \beta_j$, j = 0, 1. (Exercício:PROVE!)

Observação

SUPOSIÇÃO FUNDAMENTAL:

$$E(\varepsilon \mid x) = 0$$

Ou seja, todos os fatores contidos em ε devem ser não correlacionados com o regressor. Além disso, deve ter sido usada a forma funcional correta.

Observação (cont.)

SUPOSIÇÃO FUNDAMENTAL: (cont)

Como pode falhar?

- ***Omissão de regressor relevante, correlacionado com x;***
- ***Forma funcional especificada incorretamente;***
- ***Erro de medida em x;***
- ***Simultaneidade entre y e x;***

Suposições e Propriedades

RLS.5 – Homocedasticidade

O termo de erro aleatório ε tem a mesma variância dado qualquer valor do regressor. Ou seja,

$$\mathrm{Var}(\varepsilon \mid x) = \sigma^2.$$

Observação

De RLS.4 e RLS.5 temos que $E(\varepsilon^2 \mid x) = \sigma^2$, o que significa que σ^2 também é a expectativa incondicional de ε^2. Dessa forma, $\mathrm{Var}(\varepsilon) = \sigma^2$ (variância do erro).

Variância dos Estimadores de MQO

Teorema 2. **Sob as suposições RLS.1 a RLS.5, condicionadas aos valores amostrais da variável explicativa, prova-se que**

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

e

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}.$$

Covariância entre os Estimadores de MQO

Exercício

Mostre que a covariância entre os estimadores para os parâmetros β_0 e β_1 é dada por

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\bar{x}Var\left(\hat{\beta}_1\right)$$

Melhores estimadores lineares não-viesados (BLUE)

- **Sob as suposições RLS.1 a RLS.5, os estimadores de MQO para os parâmetros β_0 e β_1 são os melhores dentre todos os estimadores da classe dos lineares não-viesados.**
- **Isto quer dizer que além de serem não-viesados, tais estimadores apresentam a menor variância dentre os demais estimadores não-viesados, gerando estimadores com menor erro quadrático médio dentre os lineares.**

Vide demonstração em Gujarati e Porter (2011, p. 115)

Distribuição amostral de $\hat{\beta}_j$

RLS.6 – Normalidade

O erro populacional ε é *independente* do regressor x e é normalmente distribuído, com média zero e variância σ^2. Ou seja,

$$\varepsilon \sim N(0;\, \sigma^2)$$

Distribuição amostral de $\hat{\beta}_j$

Teorema 3 – Sob as suposições RLS.1 a RLS.6, condicionado aos valores amostrais do regressor,

$$\hat{\beta}_j \sim N\left(\beta_j, Var\left(\hat{\beta}_j\right)\right)$$

Observação

Tais estimadores são normalmente distribuídos, pois, são combinações lineares dos y´s, que são independentes e normalmente distribuídos.

Distribuição amostral de $\hat{\beta}_j$

Do teorema anterior vem que,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim N(0,1)$$

Do *slide* 9, vimos que as expressões das variâncias dos estimadores dos parâmetros envolvem σ^2, que é um parâmetro desconhecido. Dessa forma, deveremos procurar um estimador para tal parâmetro. Ainda, um estudo da distribuição de probabilidades da nova v.a. gerada deverá ser feito.

Estimação de σ^2

MSR (Quadrado Médio devido aos Resíduos)

$$\hat{\sigma}^2 = MSR = \frac{SSR}{n\text{-}2}$$

em que

$$SSR = \sum_{i=1}^{n} \left(y_i\text{-}\hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_o - \hat{\beta}_1\, x_i\right)^2$$

SSR perde *2* graus de liberdade, pelas 2 restrições impostas pelas condições de primeira ordem de MQO.

Estimação de σ^2

Teorema 4. Sob as suposições RLS.1 a RLS.5

$$E(\hat{\sigma}^2) = E(MSR) = \sigma^2$$

Observação

$\hat{\sigma} = \sqrt{MSR}$ **: é chamado de erro padrão da regressão.**

Erro Padrão dos Estimadores de MQO

Substituindo MSR (que é um estimador não viesado para σ^2) nas expressões provenientes do Teorema 2, teremos que

$$\widehat{Var(\hat{\beta}_0)} = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SST_x}\right),$$

e

$$\widehat{Var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{SST_x}.$$

em que

$$SST_x = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Erro Padrão dos Estimadores de MQO

À raiz quadrada das duas quantidades anteriores damos o nome de erro-padrão associado ao estimador de mínimos quadrados do *i-ésimo* parâmetro do modelo de regressão. A notação comumente utilizada é a seguinte:

$$\hat{\sigma}_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{SST_x}\right)}$$

e

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{SST_x}}.$$

Distribuição amostral de $\hat{\beta}_j$

Teorema 5. Sob as suposições RLS.1 a RLS.6,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{(n\text{-}2)}$$

Vale ressaltar que a v.a. anteriormente obtida não pode ser considerada como uma estatística de teste, uma vez que depende do parâmetro β_j.

INFERÊNCIA

Teste de Hipóteses

Para testar as hipóteses

H_0: $\beta_j = b$ (em particular b = 0)

H_A: $\beta_j \neq b$ (H_A: $\beta_j < b$ ou H_A: $\beta_j > b$),

utilizaremos o fato que, sob H_0 e sob o Teorema 5,

$$\boxed{\frac{\hat{\beta}_j - b}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{(n-2)}}$$

Intervalo de Confiança para β_j

Também, do Teorema 5, não é difícil provar que:

$$IC\left[\beta_j;\gamma\right]=\left(\hat{\beta}_j \pm t_{(n\text{-}2)}^{\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_j}\right)$$

(I.C. para β_j com coeficiente γ = 1-α de confiança)

Retornando ao Exemplo do RH

a) **Os formuladores do exame acreditam que a nota média seja uma variável relevante para explicar o desempenho. Assim sendo, conduza um teste de hipóteses que seja capaz de verificar a veracidade da afirmação feita pelos formuladores do exame. Para tanto, adote α = 5%.**

b) **A partir da construção de um intervalo de confiança para o parâmetro intercepto, o que podemos afirmar sobre a significância estatística do mesmo? Adote α = 5%.**

Resultados - Excel

Coeficiente de correlação linear de Pearson

Estatística de regressão	
R múltiplo	0,7621
R-Quadrado	0,5808
R-quadrado ajustado	0,5721
Erro padrão	6,3332
Observações	50

R^2

n

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	1	2667,85	2667,85	66,51	1,28284E-10
Resíduo	48	1925,27	40,11		
Total	49	4593,12			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*Inferior 95,0%*	*Superior 95,0%*
Interseção	68,51	3,75	18,25	3,10574E-23	60,96	76,06
Variável X 1	1,81	0,22	8,16	1,28284E-10	1,36	2,26

$\hat{\beta}_0$

$\hat{\beta}_1$

Modelo Estimado:

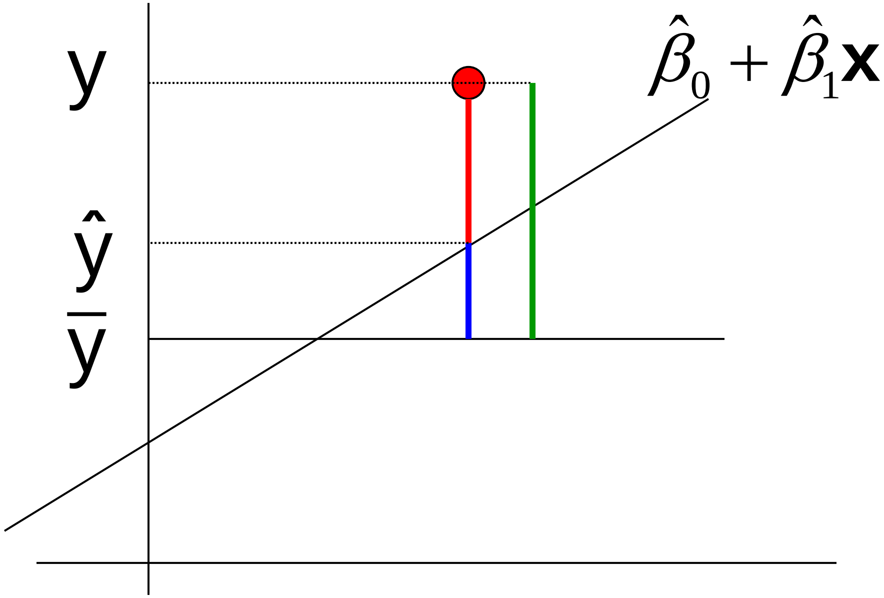
$$\widehat{desempenho} = 68{,}51 + 1{,}81\overline{nota}$$

Leitura Complementar

Retornando ao Exemplo do RH

A partir de um teste de hipóteses adequado, verifique se o modelo de regressão proposto é significante. Adote um nível de significância de 5% (α = 0,05).

Tabela de Análise de Variâncias (ANOVA)



$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SST: soma de quadrados total

SSR: soma de quadrados devido aos resíduos

SSE: soma de quadrados devido à explicação (modelo de regressão)

Tabela de Análise de Variâncias (ANOVA)

SSR, SSE e SST são v.a. e, sob certas condições, é possível provar que:

1. $\dfrac{SSR}{\sigma^2} \sim \chi^2_{(n-2)};$

2. Se $\beta_1 = 0$, $\dfrac{SSE}{\sigma^2} \sim \chi^2_{(1)};$

3. SSR e SSE são independentes.

Tabela de Análise de Variâncias (ANOVA)

Consequências:

$$\text{(a)}\quad E\left(\frac{SSR}{\sigma^2}\right) = n - 2 \Rightarrow E\left(\frac{SSR}{n-2}\right) = E(MSR) = \sigma^2$$

Logo, MSR é um estimador não-viesado de σ^2;

$$\text{(b)}\quad E(SSE) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$$

Se β_1 = 0, então SSE / 1 = MSE é um estimador não-viesado de σ^2;

Tabela de Análise de Variâncias (ANOVA)

Consequências (cont.):

(c) Se $\beta_1 = 0$,

$$\mathrm{E(SST)} = \mathrm{E(SSR)} + \mathrm{E(SSE)} =$$

$$= (\mathrm{n}-2)\sigma^2 + \sigma^2 = (\mathrm{n}-1)\sigma^2$$

Logo, SST/(n-1) é estimador não-viesado de σ^2;

Tabela de Análise de Variâncias (ANOVA)

Consequências (cont.):

(d) Se $\beta_1 = 0$,

$$F_{obs} = \frac{\dfrac{SSE/\sigma^2}{1}}{\dfrac{SSR/\sigma^2}{(n-2)}} = \frac{MSE}{MSR} \sim F_{[1,n\text{-}2]}$$

Tabela de Análise de Variâncias (ANOVA)

Os resultados descritos em (d) podem ser colocados numa tabela conhecida como ANOVA (Análise de Variâncias)

Regressão	**1**	**SSE**	**MSE**	**MSE/MSR**
Resíduo	**n-2**	**SSR**	**MSR**	
Total	**n - 1**	**SST**		

em que

$$\hat{\sigma}^2 = MSR = \frac{\mathrm{SSR}}{\mathrm{n-2}} \qquad \mathrm{M}SE = \frac{\mathrm{SSE}}{1}$$

Teste F

Devido ao item (d), para testarmos

$$\mathbf{H_o: \beta_1 = 0 \quad vs \quad H_A: \beta_1 \neq 0}$$

rejeitamos H_o a um nível α de significância se

$$F_{obs} \geq F_{[1;n-2]}^{(\alpha)}$$

Retornando ao Exemplo do RH

A partir de um teste de hipóteses adequado, verifique se o modelo de regressão proposto é significante. Adote um nível de significância de 5% (α = 0,05).

Resultados - Excel

Modelo Estimado:

$$\widehat{desempenho} = 68{,}51 + 1{,}81\overline{nota}$$

Estatística de regressão	
R múltiplo	0,7621
R-Quadrado	0,5808
R-quadrado ajustado	0,5721
Erro padrão	6,3332
Observações	50

F_{obs}

Valor-p

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1	2667,85	2667,85	66,51	1,28284E-10
Resíduo	48	1925,27	40,11		
Total	49	4593,12			

	Coeficientes	Erro padrão	Stat t	valor-P	Inferior 95,0%	Superior 95,0%
Interseção	68,51	3,75	18,25	3,10574E-23	60,96	76,06
Variável X 1	1,81	0,22	8,16	1,28284E-10	1,36	2,26

Observações Finais

i. Se não rejeitarmos a hipótese nula de que $\beta_1 = 0$, estaremos admitindo que $E(Y|X=x) = \beta_0$, cuja representação gráfica é dada por:

neste caso, o modelo de regressão estimado será dado por $\hat{y} = \bar{y}$

Observações Finais

De (i), temos que a não rejeição de H_0 fará com que adotemos o modelo

$$\hat{y} = \bar{y}$$

Por outro lado, a rejeição de H_0 fará com que adotemos o modelo

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$$

Observações Finais

ii. É importante salientar que a rejeição de H_0 não garante a adequabilidade do modelo $\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}$. Por outro lado, a rejeição de H_0 pode ser interpretada como um indicativo de superioridade do modelo $\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}$ em relação ao modelo $\hat{\mathbf{y}} = \bar{\mathbf{y}}$.

iii. Para verificarmos se o modelo adotado é adequado devemos fazer uma análise de resíduos. Desta forma, a adoção definitiva do modelo $\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}$ deve estar sujeita ainda a análises posteriores.