Intro. Econometria Usando R
- Aula 2-

Prof. Mestre. Omar Barroso Khodr

Instituto Brasileiro de Educação, Pesquisa e Desenvolvimento

Tópicos

- Medidas de Ajuste R^2
- Regressão Linear Múltipla
- Inferência

Avaliando a qualidade do ajuste (R^2)

- Existe uma medida conveniente para avaliar o quão bem um determinado modelo estatístico se ajusta aos dados. Ela é chamada de R^2, também chamado de **coeficiente de determinação**. Utilizamos a decomposição de variância recém-introduzida e escrevemos a fórmula como:

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2,$$

$$TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2,$$

$$R^2 = \frac{ESS}{TSS}.$$

Avaliando a qualidade do ajuste (R^2)

- $\hat{y}_i$: Valores estimados de Y
- $\bar{y}$: A média de Y.
- ESS: Soma dos Quadrados Explicada (Explained Sum of Squares). ↔SQE
- TSS: Soma Total dos Quadrados (Total Sum of Squares). ↔ STQ
- $R^2 = \frac{SQE}{STQ}$

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2,$$

$$TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2,$$

$$R^2 = \frac{ESS}{TSS}.$$

Avaliando a qualidade do ajuste (R^2)

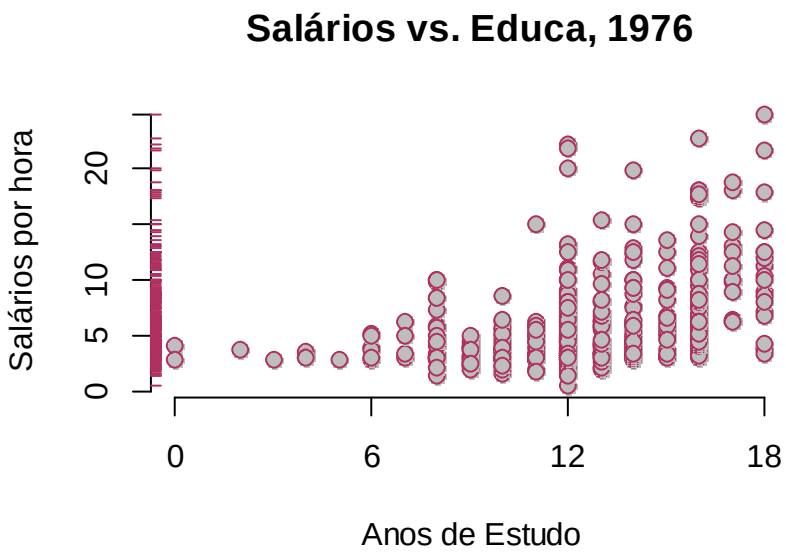
- Nesse contexto,
- TSS = ESS + SSR
- Ou Seja, a Soma Total dos Quadrados é igual a Soma dos Quadrados Explicadas mais a Soma dos Resíduos ao Quadrado (SRQ ou SSR – Sum of Squared Residuals). Em outras palavras...
- STQ = SQE + SQR
- Assim, podemos também Determinar que,
- $1 - \frac{SRQ}{STQ}$

Avaliando a qualidade do ajuste (R^2)

- $R^2 = \frac{SQE}{STQ} = 1 - \frac{SRQ}{STQ} \; \epsilon \; [0,1]$
- Desta maneira, R^2 fica entre o intervalo de 0 e 1.
- É fácil ver que um ajuste perfeito, ou seja, nenhum erro cometido ao ajustar a linha de regressão, implica $R^2 = 1$.
- Caso contrário, $R^2 = 0$.
- Nesse contexto, quanto mais perto de 1 melhor o ajuste do modelo. Caso contrário, um ajuste menor implica um ajuste menor.

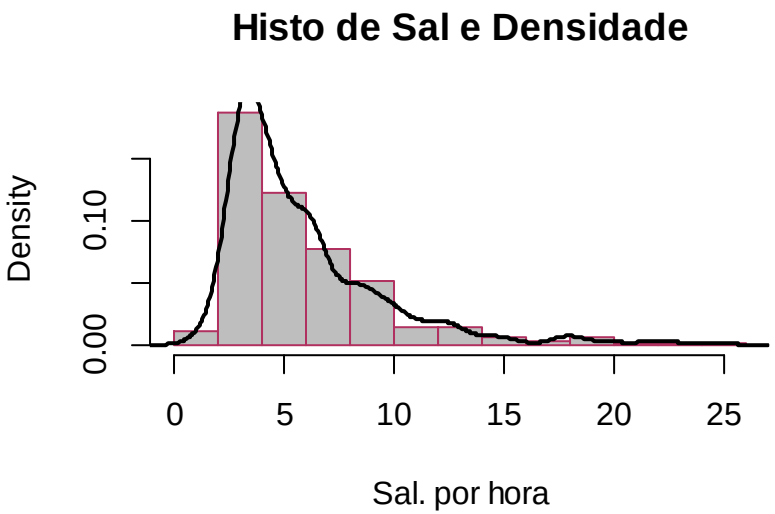
Testando em um Modelo

- Suponha que,
- $Salário = b_0 + b_1 educ_i + e_i$
- Fazendo diagnósticos...
- Pelos sinais vermelhos no eixo y, podemos perceber que os salários estão muito concentrados em torno de 5 USD por hora, com cada vez menos observações em taxas mais altas; e segundo, parece que o salário por hora parece aumentar com níveis educacionais mais elevados.



Testando em um Modelo

- Suponha que,
- $Salário = \ b_0 + b_1 educ_i + e_i$
- Fazendo diagnósticos...
- O histograma reforça o primeiro ponto, mostrando que a pdf (função de densidade de probabilidade) estimada, mostrada como uma linha preta, tem uma cauda direita muito longa: há sempre valores cada vez menores, mas sempre maiores, de salário por hora nos dados.



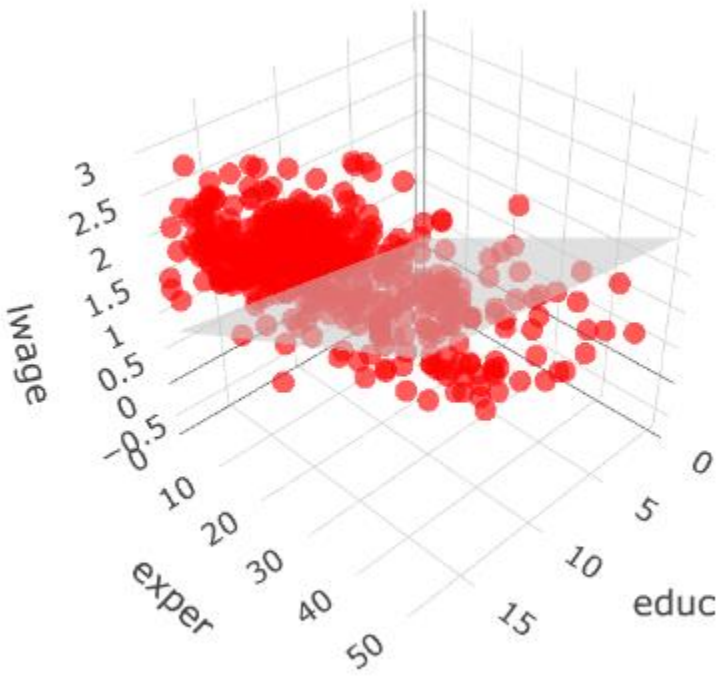
Testando em um Modelo

- Resultados,
- Com zero ano de educação, o salário-hora é de aproximadamente -0,9 dólares por hora (linha denominada (Intercepto).
- Cada ano adicional de educação aumenta o salário-hora em 54 centavos. (linha denominada educ)
- Por exemplo, para 15 anos de educação, prevemos aproximadamente -0,9 + 0,541 * 15 = 7,215 dólares/h.
- Todavia, nosso R^2 é baixo o que demonstra um ajuste fraco no modelo. Com isso, talvez existem variáveis omitidas e coisas que devemos investigar.

```
## Call:
## lm(formula = wage ~ educ, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3396 -2.1501 -0.9674  1.1921 16.6085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.90485    0.68497  -1.321    0.187
## educ         0.54136    0.05325  10.167   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.378 on 524 degrees of freedom
## Multiple R-squared:  0.1648, Adjusted R-squared:  0.1632
## F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

Regressão Múltipla

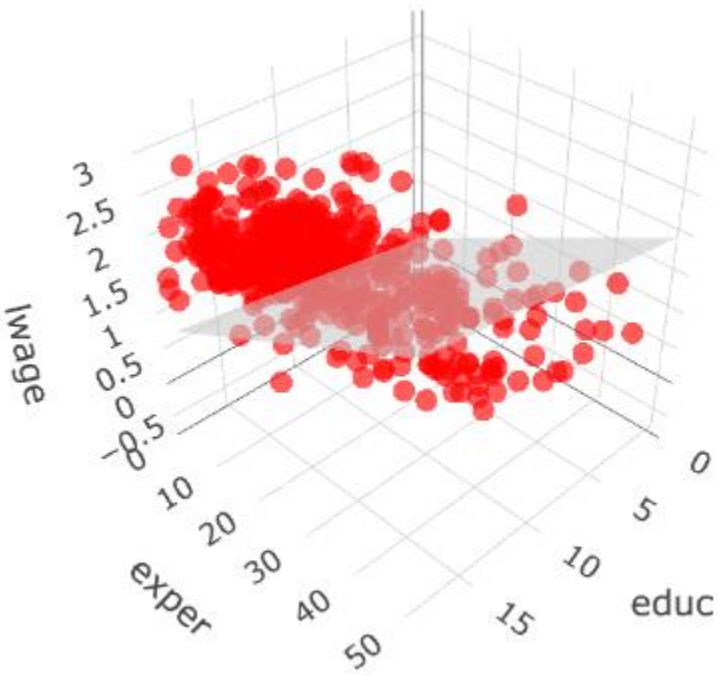
- Tudo o que aprendemos para o caso de variável única também se aplica aqui. Em vez de uma reta de regressão, agora temos um plano de regressão, ou seja, um objeto representável em três dimensões...
- Por exemplo, (x_1, x_2, y).



Fonte: Econometrics with R (Sciences Po, 2020)

Regressão Múltipla

- Considerando o modelo anterior:
- $Salário = \ b_0 + b_1 educ_i + e_i$
- Agora adicionamos mais uma variável e deixamos salário em log.
- $Log.Sal. = \ b_0 + b_1 educ_i + b_2 Exper_i + e_i$



Fonte: Econometrics with R (Sciences Po, 2020)

Regressão Múltipla (Ceteris Paribus)

- $Log.Sal. = \ b_0 + b_1 educ_i + b_2 Exper_i + e_i$
- Tudo o mais sendo igual...
- Mantendo o valor de $Exper_i$ fixo, qual seria o impacto em $Log.Sal.$ se aumentássemos apenas $educ_i$? Em outras palavras, mantendo tudo o mais constante, qual seria o impacto de alterar $educ_i$ em $Log.Sal.$?
- Qual é o impacto de cada variável isoladamente?
- Na verdade, o tipo de pergunta feita aqui é tão comum que tem seu próprio nome: diríamos "ceteris paribus, qual é o impacto da educação nos Salários?"

Regressão Múltipla (Multicolinearidade)

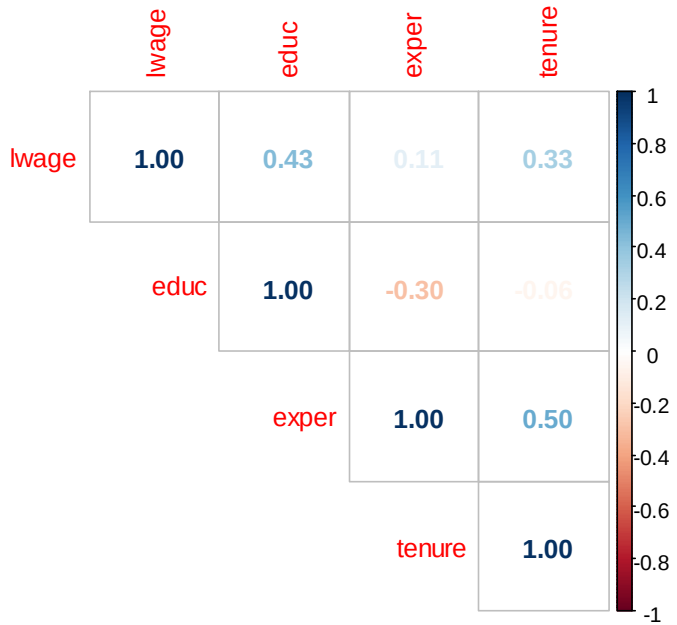
- Um requisito importante para a regressão múltipla é que os dados não sejam linearmente dependentes: cada variável deve fornecer pelo menos alguma informação nova para o resultado e não pode ser replicada como uma combinação linear de outras variáveis.
- Não podemos ter variáveis linearmente dependentes ou perfeitamente colineares. Isso é conhecido como condição de **classificação**.

Regressão Múltipla (Multicolinearidade)

- Em particular, a condição determina que precisamos de pelo menos $N \geq K + 1$, ou seja, mais observações do que coeficientes.
- Quanto maior o grau de dependência linear entre nossas variáveis explicativas, menos informações podemos extrair delas, e nossas estimativas se tornam menos precisas.

Regressão Múltipla (Exemplo)

- $Log.Sal. = b_0 + b_1 educ_i + b_2 Exper_i + e_i$
- Em nosso modelo, Podemos perceber que a correlação entre as nossas variáveis são moderadas ou baixas.
- Ou seja, não temos problemas de multi-colinearidade.



Regressão Múltipla (Exemplo)

- Podemos perceber que ao adicionar mais regressores, o ajuste do nosso modelo melhora.
- Por exemplo, $Log.Sal. = b_0 + b_1 educ_i + b_2 Exper_i$
- Temos um R^2 um pouco maior em 0.249.
- $Log.Sal. = b_0 + b_1 educ_i + b_2 Exper_i + b_3 Ten_3$
- O ajuste melhora ainda mais...

	Dependent variable:		
	lwage		
	(1)	(2)	(3)
educ	0.083***	0.098***	0.092***
	(0.008)	(0.008)	(0.007)
exper		0.010***	0.004**
		(0.002)	(0.002)
tenure			0.022***
			(0.003)
Constant	0.584***	0.217**	0.284***
	(0.097)	(0.109)	(0.104)
Observations	526	526	526
R^2	0.186	0.249	0.316
Adjusted R^2	0.184	0.246	0.312
Residual Std. Error	0.480 (df = 524)	0.461 (df = 523)	0.441 (df = 522)
F Statistic	119.582*** (df = 1; 524)	86.862*** (df = 2; 523)	80.391*** (df = 3; 522)
Note:			*p<0.1; **p<0.05; ***p<0.01

Regressão Múltipla (Graus de Liberdade)

- Graus de Liberdade = Degrees of Freedom (df).
- Quando você ajusta mais regressores (preditores) em um modelo de regressão, você perde graus de liberdade (gl).
- Cada regressor adicional consome um (ou mais, no caso de variáveis categóricas) grau de liberdade para estimar seu coeficiente associado.

	Dependent variable:		
	lwage		
	(1)	(2)	(3)
educ	0.083***	0.098***	0.092***
	(0.008)	(0.008)	(0.007)
exper		0.010***	0.004**
		(0.002)	(0.002)
tenure			0.022***
			(0.003)
Constant	0.584***	0.217**	0.284***
	(0.097)	(0.109)	(0.104)
Observations	526	526	526
R^2	0.186	0.249	0.316
Adjusted R^2	0.184	0.246	0.312
Residual Std. Error	0.480 (df = 524)	0.461 (df = 523)	0.441 (df = 522)
F Statistic	119.582*** (df = 1; 524)	86.862*** (df = 2; 523)	80.391*** (df = 3; 522)
Note:			*p<0.1; **p<0.05; ***p<0.01

Regressão Múltipla (Graus de Liberdade)

- Para um conjunto de dados com n observações, o total de graus de liberdade é n-1 (usado para estimar a variância da variável dependente em torno de sua média).
- Se você incluir K regressores (excluindo o intercepto), usará até k graus de liberdade para estimar seus coeficientes.
- Os graus de liberdade restantes são n-k-1 (onde -1 representa o intercepto).
- Adicionar mais regressores reduz o GL, o que significa que restam menos informações independentes para estimar a variância do erro.

	Dependent variable:		
	lwage		
	(1)	(2)	(3)
educ	0.083***	0.098***	0.092***
	(0.008)	(0.008)	(0.007)
exper		0.010***	0.004**
		(0.002)	(0.002)
tenure			0.022***
			(0.003)
Constant	0.584***	0.217**	0.284***
	(0.097)	(0.109)	(0.104)
Observations	526	526	526
R^2	0.186	0.249	0.316
Adjusted R^2	0.184	0.246	0.312
Residual Std. Error	0.480 (df = 524)	0.461 (df = 523)	0.441 (df = 522)
F Statistic	119.582*** (df = 1; 524)	86.862*** (df = 2; 523)	80.391*** (df = 3; 522)
Note:			*p<0.1; **p<0.05; ***p<0.01

Regressão Múltipla (Graus de Liberdade)

- Cada regressor adicionado poderia melhorar o ajuste (R^2 mais alto), mas se for espúrio, desperdiça df e infla a variância.

- A penalização desencoraja essa compensação, a menos que o regressor forneça valor explicativo suficiente.

	Dependent variable:		
	lwage		
	(1)	(2)	(3)
educ	0.083***	0.098***	0.092***
	(0.008)	(0.008)	(0.007)
exper		0.010***	0.004**
		(0.002)	(0.002)
tenure			0.022***
			(0.003)
Constant	0.584***	0.217**	0.284***
	(0.097)	(0.109)	(0.104)
Observations	526	526	526
R^2	0.186	0.249	0.316
Adjusted R^2	0.184	0.246	0.312
Residual Std. Error	0.480 (df = 524)	0.461 (df = 523)	0.441 (df = 522)
F Statistic	119.582*** (df = 1; 524)	86.862*** (df = 2; 523)	80.391*** (df = 3; 522)
Note:			*p<0.1; **p<0.05; ***p<0.01

Inferência (P-valores e Teste-T)

- **P-valores (as estrelinhas):** Quantificar a probabilidade de observar o coeficiente estimado (ou mais extremo) se a hipótese nula (H_0).
- **Interpretação:**
- p < 0,05 (limiar típico): Rejeitar H0; evidência de que o coeficiente é significativo.
- p > 0,05: Falha ao rejeitar H0; nenhuma evidência forte contra H0.

	Dependent variable:		
	lwage		
	(1)	(2)	(3)
educ	0.083^{***}	0.098^{***}	0.092^{***}
	(0.008)	(0.008)	(0.007)
exper		0.010^{***}	0.004^{**}
		(0.002)	(0.002)
tenure			0.022^{***}
			(0.003)
Constant	0.584^{***}	0.217^{**}	0.284^{***}
	(0.097)	(0.109)	(0.104)
Observations	526	526	526
R^2	0.186	0.249	0.316
Adjusted R^2	0.184	0.246	0.312
Residual Std. Error	0.480 (df = 524)	0.461 (df = 523)	0.441 (df = 522)
F Statistic	119.582^{***} (df = 1; 524)	86.862^{***} (df = 2; 523)	80.391^{***} (df = 3; 522)
Note:			*p<0.1; **p<0.05; ***p<0.01

Inferência (P-valores e Teste-T)

- $teste\ t = \frac{Coeficiente\ Estimado}{Erro\ Padrão\ do\ coeficiente}$
- Um grande valor t absoluto sugere que é improvável que o coeficiente seja zero por acaso.
- Os testes t avaliam se o efeito de um regressor é estatisticamente distinguível de zero.

	Dependent variable:		
	lwage		
	(1)	(2)	(3)
educ	0.083***	0.098***	0.092***
	(0.008)	(0.008)	(0.007)
exper		0.010***	0.004**
		(0.002)	(0.002)
tenure			0.022***
			(0.003)
Constant	0.584***	0.217**	0.284***
	(0.097)	(0.109)	(0.104)
Observations	526	526	526
R^2	0.186	0.249	0.316
Adjusted R^2	0.184	0.246	0.312
Residual Std. Error	0.480 (df = 524)	0.461 (df = 523)	0.441 (df = 522)
F Statistic	119.582*** (df = 1; 524)	86.862*** (df = 2; 523)	80.391*** (df = 3; 522)
Note:			*p<0.1; **p<0.05; ***p<0.01

Bibliografia

- Wooldridge, J.M. (2013) Introductory econometrics: a modern approach. 5th ed. Michigan State University.