



DCP098

Fundamentos para Avaliação Quantitativa de Políticas Públicas

**Regressão multivariada: precisão.
Multicolinearidade.**

Aula 18
01 de junho de 2022

Ana Paula Karruz

Agenda da aula anterior e desta aula

1. Grau de ajuste
2. **Regressão multivariada: precisão**
3. Multicolinearidade

Dois desafios à análise estatística: aleatoriedade e endogeneidade

Fontes de incerteza quanto ao efeito estimado de X sobre Y

Sampling randomness: amostras de diferentes tamanhos geram coeficientes estimados diferentes; amostras diferentes de um mesmo tamanho também geram coeficientes estimados diferentes; na estatística frequentista, coeficiente populacional é fixo)

Modeled randomness: aleatoriedade e complexidade na formação de Y redundam em variáveis omitidas; nota: aqui não estamos falando de variáveis omitidas correlacionadas com X

Variáveis omitidas correlacionadas com X: existência dessas variáveis implica espuriedade

Aleatoriedade
(compromete a
precisão)

Como a regressão múltipla aumenta a precisão das estimativas?

Endogeneidade
(compromete a
acurácia)

$\hat{\beta}_1$

(ou qualquer outro coeficiente de inclinação estimado)

Numa regressão bivariada, a variância de $\beta_1\text{hat}^*$ é dada por:

[Obs.: A variância de $\beta_0\text{hat}$ é dada por outra fórmula.]

Refresher

Erro padrão de $\beta_1\text{hat}$, o $\text{se}(\beta_1\text{hat})$ = Raiz quadrada da $\text{var}(\beta_1\text{hat})$

$$\text{var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{N \times \text{var}(X)}$$

Variância da regressão = (Residual standard error)². Obtenha o residual standard error no output da regressão.

$\hat{\sigma}^2$

Variância da regressão

- Mede quão bem o modelo explica a variação de Y
- Seu cálculo baseia-se nos resíduos
- j = número de parâmetros estimados (incluindo o intercepto)
- É também uma estimativa da variância de ϵ
- **Intuição:** média do quadrado da distância entre valores observados e previstos de Y

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{N - j}$$

$$= \frac{\sum_{i=1}^N \hat{\epsilon}_i^2}{N - j}$$

N

Tamanho da amostra

- **Intuição:** mais dados implicam menor variância, pois a chance de o acaso nos levar às caudas da distribuição de $\beta_1\text{hat}$ é menor em amostras maiores (i.e., menor sampling randomness)

$\text{var}(X)$

Variância amostral de X

- Quanto mais X variar, mais precisa será a distribuição de $\beta_1\text{hat}$
- **Intuição:** se X varia pouco, não temos muita informação para estimar o efeito da variação de X sobre a variação de Y

$$\text{var}(X) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

Em amostras pequenas, divida o somatório por N - 1

* Fórmula de $\text{var}(\beta_1\text{hat})$ é mais complicada quando erros são correlacionados ou heteroscedásticos, mas as intuições sobre variância da regressão, tamanho da amostra e $\text{var}(X)$ se aplicam. Voltaremos a esse ponto em aulas futuras.

Na regressão múltipla, $\text{var}(\hat{\beta})$ é influenciada também pela correlação entre as variáveis independentes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, \dots, X_k \text{ são mantidos constantes})$

- Para erros homoscedásticos e independentes entre si*, seja X_j uma variável explicativa:

$$\text{var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{N \text{var}(X_j)(1 - R_j^2)}$$

Observe o numerador: regressão múltipla tende a gerar estimativas mais precisas

- $(1 - R_j^2)$ is the new kid on the block
- Wait... Hold on... We already know his face! Or, at least, the face of a close relative of his:
 - R_j^2 é o R^2 (coeficiente de determinação) de uma **regressão auxiliar** em que X_j é a variável dependente; todas as outras variáveis independentes do modelo principal são as variáveis independentes da regressão auxiliar; **cada X_j produz uma regressão auxiliar diferente**



NEW KIDS ON THE BLOCK
Boy band americana formada em 1986; separou-se em 1994, retornando em 2008.

Regressões auxiliares não são modelos causais; servem apenas para apurar o grau de correlação entre as variáveis independentes do modelo principal.

* Fórmula de $\text{var}(\hat{\beta})$ é mais complicada quando erros são correlacionados ou heteroscedásticos, mas as intuições sobre variância da regressão, tamanho da amostra, $\text{var}(X)$ e multicolinearidade se aplicam. 5

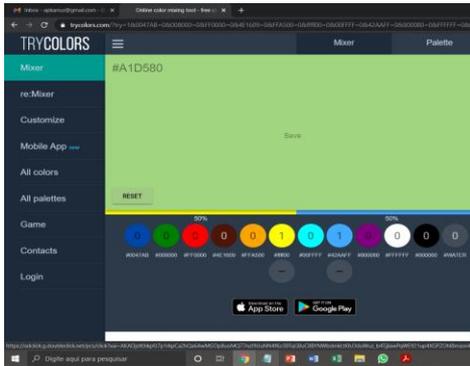
var($\hat{\beta}_j$) é maior quando X_j é bastante correlacionado com uma ou mais das outras variáveis independentes

$$\text{var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{N\text{var}(X_j)(1 - R_j^2)}$$

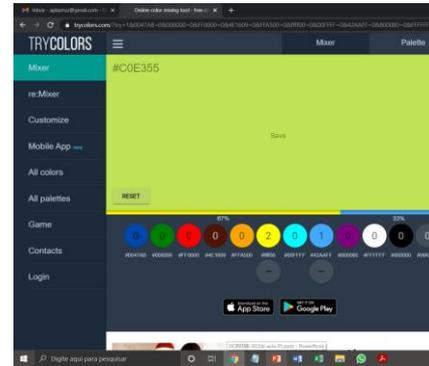
Esses R_j^2 nos dizem o **quanto as outras variáveis independentes explicam X_j** . Se as outras variáveis explicarem X_j muito bem, o R_j^2 será alto e – aqui está a sacada principal – o denominador será menor. Observe que o denominador da fórmula da $\text{var}(\hat{\beta}_j)$ é $(1 - R_j^2)$. Lembre-se de que qualquer R^2 está entre 0 e 1, então, **à medida que R_j^2 fica maior, $1 - R_j^2$ diminui** o que por sua vez **faz $\text{var}(\hat{\beta}_j)$ aumentar**. A intuição é que se a variável X_j for **praticamente indistinguível** das outras variáveis independentes, faz sentido que seja **difícil dizer quanto X_j afeta Y** e teremos, portanto, uma maior $\text{var}(\hat{\beta}_j)$.

Bailey (2016: 227)

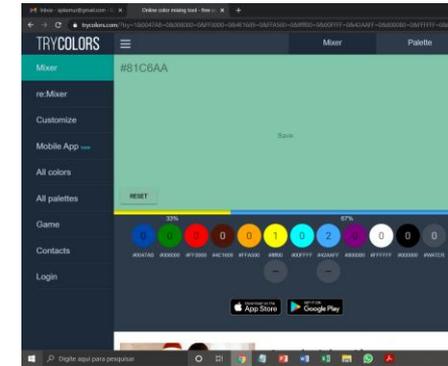




$Y_{hat} = B_0_{hat} + B_1_{hat}Amarelo + B_2_{hat}AzulRoyal$

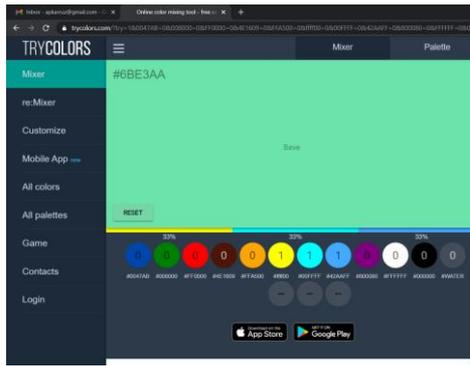


Aumenta Amarelo

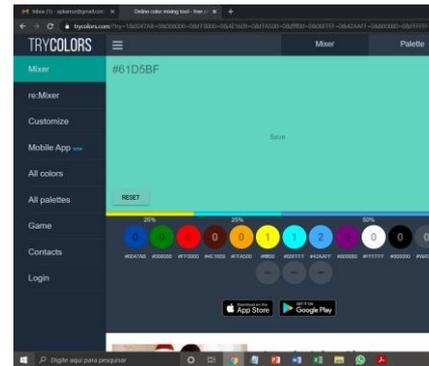


Aumenta AzulRoyal

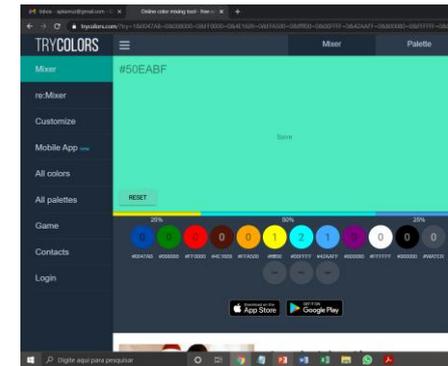
- ↪
 Neste primeiro modelo, as duas cores usadas para prever Y (Y = tons de verde) são bastante distintas; o efeito da intensificação do Amarelo é bem diferente do efeito da intensificação do AzulRoyal



$Y_{hat} = B_0_{hat} + B_1_{hat}Amarelo + B_2_{hat}AzulRoyal + B_3_{hat}AzulClaro$



Aumenta AzulRoyal



Aumenta AzulClaro

- ↪
 No segundo modelo, duas das três cores usadas para prever Y são muito semelhantes entre si; tal semelhança traz imprecisão para a estimativa do efeito de cada uma dessas cores na composição de Y
- Ao adicionarmos AzulClaro, estamos demandando que o modelo estime mais um parâmetro, porém a informação “nova” trazida pela variável adicional é pouca: mantendo-se constante AzulRoyal, sobra pouca variação em AzulClaro (e vice-versa)

Agenda da aula anterior e desta aula

1. Grau de ajuste
2. Regressão multivariada: precisão
3. **Multicolinearidade**

As propriedades mais bacanas de MQO

Veja, propriedade é diferente de premissa



Se as premissas de MQO forem atendidas,
MQO é **BLUE!**

- **Best** (variância mínima, i.e., máxima precisão)
 - **Linear**
 - **Unbiased** (livre de vieses)
 - **Estimator**
-
- Se as premissas de MQO estiverem atendidas, MQO é “melhor” em relação às alternativas, quais sejam: adaptações de MQO (e.g., Mínimos Quadrados Generalizados) e estimadores de Máxima Verossimilhança (um algoritmo iterativo que permite não linearidades nos β s)
 - As propriedades BLUE do MQO são provadas pelo teorema de Gauss-Markov

Para que o estimador MQO seja **BLUE**, é preciso atender às chamadas “premissas clássicas”

O que eventualmente ficou de fora não é correlacionado com as variáveis incluídas – se for, teremos o problema do viés de variável omitida

- O modelo de regressão é **linear nos parâmetros** (i.e., coeficientes são adicionados e com expoente = 1)

- ε , o termo de erro aleatório, tem **média populacional = 0**
- Não há correlação entre cada uma das variáveis explicativas e o termo de erro – esta é a premissa de **exogeneidade: correlação $(X_j, \varepsilon) = 0$**

Premissa da média condicional zero é resultado destas duas premissas

- As variáveis explicativas não são uma função linear das outras (i.e., **não há multicolinearidade**)
- O termo de erro tem variância constante (i.e., é **não há heteroscedasticidade**)
- Os erros não são correlacionados entre si (i.e., **não existe autocorrelação** serial ou espacial)
- O **erro** tem distribuição **normal**: $\varepsilon \sim N(0, \sigma^2)$

Premissa requerida para teste de hipótese e intervalo de confiança, não para estimação de MQO

Para que o estimador MQO seja **BLUE**, é preciso atender às chamadas “premissas clássicas”

O que eventualmente ficou de fora não é correlacionado com as variáveis incluídas – se for, teremos o problema do viés de variável omitida

- O modelo de regressão é **linear nos parâmetros** (i.e., coeficientes são adicionados e com expoente = 1)

- ε , o termo de erro aleatório, tem **média populacional = 0**
- Não há correlação entre cada uma das variáveis explicativas e o termo de erro – esta é a premissa de **exogeneidade: correlação $(X_j, \varepsilon) = 0$**

Premissa da média condicional zero é resultado destas duas premissas

- **As variáveis explicativas não são uma função linear das outras (i.e., não há multicolinearidade)**

Premissa requerida para teste de hipótese e intervalo de confiança, não para estimação de MQO

- O termo de erro tem variância constante (i.e., é **não há heteroscedasticidade**)
- Os erros não são correlacionados entre si (i.e., **não existe autocorrelação** serial ou espacial)
- O **erro** tem distribuição **normal**: $\varepsilon \sim N(0, \sigma^2)$

Temos multicolinearidade quando há associação linear entre variáveis independentes

- É **improvável** que encontremos variáveis explicativas **ortogonais** (i.e., que apresentem correlação = 0); portanto, conviveremos com algum grau de multicolinearidade
- **Não há um limiar** para definir o que é um nível “**aceitável**” ou “**preocupante**” de multicolinearidade
- Mas é possível medir **quanto** a correlação existente **infla a $\text{var}(\hat{\beta}_j)$** , em comparação com um **cenário de zero correlação** entre variáveis explicativas; fazemos isso com base no **variance inflation factor (VIF)**

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

VIF é parte da fórmula da variância do coeficiente de inclinação estimado na regressão múltipla

$$\begin{aligned} \text{var}(\hat{\beta}_j) &= \frac{\hat{\sigma}^2}{N * \text{var}(X_j) * (1 - R_j^2)} \\ &= \underbrace{\frac{\hat{\sigma}^2}{N * \text{var}(X_j)}}_{\text{fórmula da } \text{var}(\hat{\beta}_j) \text{ na regressão simples}} * \underbrace{\frac{1}{(1 - R_j^2)}}_{\text{VIF}} \end{aligned}$$

VIF no

```
> if (! "haven" %in% installed.packages()) install.packages("haven", dep = T) # for reading .dta files
> if (! "car" %in% installed.packages()) install.packages("car", dep = T) # for vif e lht (F test)
> library(haven)
> library(car)

> dados <- read_dta('auto.dta')
> dados = as.data.frame(dados)

> reg_main = lm(price ~ mpg + trunk + foreign, data = dados)
> summary(reg_main)
```

Call:
lm(formula = price ~ mpg + trunk + foreign, data = dados)

Residuals:

	Min	1Q	Median	3Q	Max
	-3461.1	-1704.2	-873.2	1014.3	10287.3

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10033.08	2256.68	4.446	3.21e-05	***
mpg	-261.99	64.91	-4.036	0.000137	***
trunk	83.65	86.50	0.967	0.336871	
foreign	1887.46	711.42	2.653	0.009861	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2532 on 70 degrees of freedom
Multiple R-squared: 0.2933, Adjusted R-squared: 0.263
F-statistic: 9.683 on 3 and 70 DF, p-value: 1.996e-05

```
> vif(reg_main)
      mpg      trunk  foreign
1.605833 1.558684 1.220334
```

Como interpretar os VIFs?

Comumente, a interpretação de VIFs baseia-se em regras de bolso. O'Brien (2007, 684-685) recomenda que sejam considerados outros fatores capazes de influenciar a variância das estimativas:

Regras de bolso para valores de VIF têm aparecido na literatura: **regra do 4, regra do 10, etc.** Segundo essas regras, quando o VIF excede esses valores, a multicolinearidade é considerada muito alta e **pairam dúvidas sobre os resultados da análise de regressão.**

[...]

Demonstramos que as regras práticas associadas ao VIF (e à tolerância*) precisam ser interpretadas no contexto de outros fatores [e.g., tamanho da amostra, variância da variável independente] que influenciam a estabilidade das estimativas do coeficiente de regressão em questão. Esses efeitos podem facilmente reduzir a variância dos coeficientes de regressão muito mais do que o VIF infla essas estimativas, mesmo quando o VIF é 10, 20, 40 ou mais. É importante ressaltar que a preocupação com os efeitos da inflação de variância é diferente em situações em que rejeitamos a hipótese nula [...] em relação às situações em que a hipótese nula não é rejeitada [...]. No primeiro caso, encontramos um resultado estatisticamente significativo, [...] mesmo com inflação da variância. No segundo caso, podemos ter sido prejudicados pelo aumento da variância associada ao coeficiente de regressão.

Referência

O'Brien, R. M., 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), pp. 673-690. <https://doi.org/10.1007/s11135-006-9018-6>

* Tolerância = 1/VIF.

Multicolinearidade não causa viés, e não exige correção da $\text{var}(\hat{\beta}_j)$

*É muito importante entender o que a multicolinearidade faz. **Não causa viés. Nem mesmo faz com que os erros padrão de $\hat{\beta}_j$** [ou, genericamente falando, de $\hat{\beta}_j$] **sejam incorretos.** Simplesmente faz com que os erros padrão sejam maiores do que seriam se não houvesse multicolinearidade. Em outras palavras, [em caso de multicolinearidade] o MQO [...produz] estimativas não enviesadas e com a incerteza [i.e., o erro padrão] adequadamente calculada. [A consequência da multicolinearidade é que] quando as variáveis [independentes] estão fortemente relacionadas entre si, teremos mais incerteza – as distribuições de $\hat{\beta}_j$ [ou, genericamente falando, de $\hat{\beta}_j$] serão mais dispersas, o que significa que será mais difícil aprender com os dados.*

Bailey (2016: 228)

Multicolinearidade não parece ser um grande problema. Ainda assim, devemos fazer algo sobre ela? Depende.

- Se a $\text{var}(\hat{\beta}_j)$ for pequena, será **possível distinguir o efeito de diferentes variáveis explicativas** – neste caso, **deixe estar**; exemplos:
 - Table 5.2 (Bailey, 2016: 201): $\text{corr}(\text{adult height, adolescent height}) = 0,86$
 - Table 5.3 (Bailey, 2016: 216): $\text{corr}(\text{years of school, test score}) = 0,81$
- Se a **multicolinearidade for substancial**, não sendo possível distinguir o efeito de diferentes variáveis explicativas, **conduza o test F de múltiplas restrições** e apresente seus resultados
 - O teste F de múltiplas restrições indicará **se, tomadas em conjunto, as variáveis colineares parecem importar para a explicação de Y**, ainda que não possamos apurar os efeitos individuais dessas variáveis

Não caia na tentação de descartar (drop) uma das variáveis colineares: se você tinha uma boa razão teórica para ter essas variáveis na equação, faltará uma boa razão teórica para remover uma delas.

Vide ANEXO: Sobre testes F



Dose letal

*Uma dose letal de multicolinearidade é chamada de **multicolinearidade perfeita**, que ocorre quando **uma variável independente é completamente explicada por outras variáveis independentes**. Se isso acontecer, $R_j^2 = 1$ e $\text{var}(\hat{\beta}_j)$ não pode ser calculada, pois tem $(1 - R_j^2)$ no denominador (no sentido de que o denominador se torna zero, o que causa indefinição). Nesse caso, o software estatístico se recusará a estimar o modelo ou excluirá [will drop] automaticamente variáveis independentes até que não haja multicolinearidade perfeita. Um exemplo bobo de multicolinearidade perfeita é quando alguém inclui a mesma variável duas vezes em um modelo.*

Ou quando incluímos dummies para todas as categorias possíveis (e.g., uma dummy para brasileiro e outra para estrangeiro no mesmo modelo)

Bailey (2016: 230)

ANEXO: Sobre testes F
Este conteúdo não será cobrado.

Sobre testes F

- **Teste F de múltiplas restrições**

- Teste F de significância geral da regressão

Teste F : Teste de múltiplas restrições (ou teste de restrições de exclusão)

- Testar se um grupo de variáveis tem efeito sobre a variável dependente.
- A hipótese nula é que um conjunto de variáveis não tem efeito sobre y (β_3 , β_4 e β_5 , por exemplo), quando o outro conjunto de variáveis foi controlado (β_1 e β_2 , por exemplo).
- Esse é um exemplo de restrições múltiplas.
- $H_0: \beta_3=0, \beta_4=0, \beta_5=0$.
- $H_1: H_0$ não é verdadeira.
- Quando pelo menos um dos betas for diferente de zero, rejeitamos a hipótese nula.

Estatística F (ou razão F)

- Precisamos saber o quanto SQR aumenta quando retiramos as variáveis que estamos testando.
- Modelo restrito terá β_0, β_1 e β_2 .
- Modelo irrestrito terá $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ e β_5 .
- A estatística F é definida como:

$$F \equiv \frac{(SQR_r - SQR_{ir})/q}{SQR_{ir}/(n - \boxed{k} - 1)}$$

Quanto poder explicativo perdemos ao remover β_3, β_4 e β_5 ?

Do modelo irrestrito

- SQR_r é a soma dos resíduos quadrados do modelo restrito.
- SQR_{ir} é a soma dos resíduos quadrados do modelo irrestrito.
- q é o número de variáveis independentes retiradas (neste caso temos três: β_3, β_4 e β_5), ou seja, $q = gl_r - gl_{ir}$.

Regras de rejeição de F

- O valor crítico (c) depende de:
 - Nível de significância (10%, 5% ou 1%, por exemplo).
 - Graus de liberdade do numerador ($q = gl_r - gl_{ir}$).
 - Graus de liberdade do denominador ($n - k - 1$).
 - Quando os gl do denominador chegam a 120, a distribuição F não é mais sensível a eles (usar $gl = \infty$).

- Uma vez obtido o F_c , rejeitamos H_0 , em favor de H_A , ao nível de significância escolhido se: $F \geq F_c$

Distribuição F
assume apenas
valores positivos

- Se H_0 ($\beta_3 = 0, \beta_4 = 0, \beta_5 = 0$) é rejeitada, β_3, β_4 e β_5 são **estatisticamente significantes conjuntamente**.
- Se H_0 ($\beta_3 = 0, \beta_4 = 0, \beta_5 = 0$) não é rejeitada, β_3, β_4 e β_5 são **conjuntamente não significantes**.

Forma R^2 da estatística F

$$F \equiv \frac{(SQR_r - SQR_{ir})/q}{SQR_{ir}/(n - k - 1)}$$

- O teste F pode ser calculado usando os R^2 s dos modelos restrito e irrestrito.
- É mais fácil utilizar números entre zero e um (R^2) do que números que podem ser muito grandes (SQR).

$$F \equiv \frac{(R_{ir}^2 - R_r^2)/q}{(1 - R_{ir}^2)/(n - k - 1)}$$

Para chegar à forma R^2 da estatística F , basta multiplicar numerador e denominador por $1/SQT$

$$F \equiv \frac{(SQR_r - SQR_{ir})/q}{SQR_{ir}/(n-k-1)}$$

$$\equiv \frac{1/SQT \cdot (\text{numerador})}{1/SQT \cdot (\text{denominador})}$$

$$\boxed{\text{numerador}} \left(\frac{1}{SQT} * SQR_r - \frac{1}{SQT} * SQR_{ir} \right) / q$$

$$= \left((1 - R_r^2) - (1 - R_{ir}^2) \right) / q$$

$$(R_{ir}^2 - R_r^2) / q$$

$$\boxed{\text{denominador}}$$

$$\left(\frac{1}{SQT} * SQR_{ir} \right) / (n-k-1)$$

$$(1 - R_{ir}^2) / (n-k-1)$$

Teste F para múltiplas restrições em R

```
> summary(spec1)
```

```
Call:
```

```
lm(formula = my_formula, data = womenlabor)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-5.5874 -1.7729 -0.2061  1.7253  8.5357
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.7767655   8.4815645    6.930 5.88e-10 ***
yf            0.0004277   0.0004080    1.048 0.297292
ym           -0.0003552   0.0003104   -1.144 0.255502
educ          0.3973625   0.0570042    6.971 4.87e-10 ***
ue           -0.9404541   0.2536938   -3.707 0.000360 ***
mr           -0.3109530   0.1329970   -2.338 0.021577 *
dr            0.1776552   0.1739730    1.021 0.309883
urb          -0.0108626   0.0243890   -0.445 0.657096
wh           -0.1158179   0.0311566   -3.717 0.000348 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.835 on 91 degrees of freedom
Multiple R-squared:  0.7575,          Adjusted R-squared:
0.7362
F-statistic: 35.53 on 8 and 91 DF,  p-value: < 2.2e-16
```

```
> lht(spec1, c("ym = 0", "mr = 0", "dr = 0"))
```

```
Linear hypothesis test
```

```
Hypothesis:
```

```
ym = 0
mr = 0
dr = 0
```

```
Model 1: restricted model
```

```
Model 2: wlfp ~ yf + ym + educ + ue + mr + dr + urb + wh
```

```
      Res.Df    RSS Df Sum of Sq    F Pr(>F)
1         94 795.91
2         91 731.25  3     64.664 2.6823 0.05141 .
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Teste F para múltiplas restrições em R

Residual standard error = Raiz quadrada de σ_{hat}^2 ,
que é a variância estimada da regressão

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{n - k - 1}$$

SQR = Residual standard error² * (n - k - 1)

$$\text{SQR}_r = 796,0014$$

$$\text{SQR}_{ir} = 731,3875$$

$$F \equiv \frac{(796,0014 - 731,3875)/3}{731,3875/91}$$

$$F \equiv \frac{64,6139/3}{8,0372}$$

$$F \equiv 2,6798$$

Com $\alpha = 5\%$, $F_{(2,60)} = 2,76$ e $F_{(2,100)} = 2,70$

Sobre testes F

- Teste F de múltiplas restrições

- **Teste F de significância geral da regressão**

Teste F para significância geral da regressão

- No modelo com k variáveis independentes, podemos escrever a hipótese nula como:
 - $H_0: x_1, x_2, \dots, x_k$ não ajudam a explicar y .
 - $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$.
- Modelo restrito: $y = \beta_0 + \varepsilon$
- Modelo irrestrito: $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$
- Número de variáveis independentes retiradas ($q =$ graus de liberdade do numerador) é igual ao próprio número de variáveis independentes (k):

$$F \equiv \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

- Mesmo com R^2 pequeno, podemos ter teste F significativo para o conjunto, por isso não podemos olhar somente o R^2 .

Passo a passo do teste F para significância geral da regressão

- **Passo 1:** Especifique suas hipóteses; note que H_0 pode ser escrita em termos do R^2 :

$$H_0: R^2 = 0 \text{ (i.e., } \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0)$$

$$H_A: R^2 > 0$$

- **Passo 2:** Escolha um nível de significância (α) e apure o respectivo F crítico. A distribuição F é dada pela combinação de dois tipos de graus de liberdade. Na tabela F , devemos procurar o F -crítico $F_{c(k, n-k-1)}$:
 - k graus de liberdade na coluna
 - $n - k - 1$ graus de liberdade na linha

- **Passo 3:** Obtenha a estatística F

$$F \equiv \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

- **Passo 4:** Compare o F -crítico com a estatística F e aplique a regra de decisão

Se $F \geq F_c$, então rejeite H_0

Se $F < F_c$, então não rejeite H_0

Teste F de significância geral em R

```
> summary(spec1)
```

```
Call:
```

```
lm(formula = my_formula, data = womenlabor)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max  
-5.5874 -1.7729 -0.2061  1.7253  8.5357
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 58.7767655  8.4815645   6.930 5.88e-10 ***  
yf           0.0004277  0.0004080   1.048 0.297292  
ym          -0.0003552  0.0003104  -1.144 0.255502  
educ        0.3973625  0.0570042   6.971 4.87e-10 ***  
ue          -0.9404541  0.2536938  -3.707 0.000360 ***  
mr          -0.3109530  0.1329970  -2.338 0.021577 *  
dr           0.1776552  0.1739730   1.021 0.309883  
urb         -0.0108626  0.0243890  -0.445 0.657096  
wh          -0.1158179  0.0311566  -3.717 0.000348 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.835 on 91 degrees of freedom
```

```
Multiple R-squared:  0.7575,          Adjusted R-squared:  
0.7362
```

```
F-statistic: 35.53 on 8 and 91 DF, p-value: < 2.2e-16
```

```
> lht(spec1, c("ym = 0", "mr = 0", "dr = 0"))
```

```
Linear hypothesis test
```

```
Hypothesis:
```

```
ym = 0
```

```
mr = 0
```

```
dr = 0
```

```
Model 1: restricted model
```

```
Model 2: wlfp ~ yf + ym + educ + ue + mr + dr + urb + wh
```

```
      Res.Df    RSS Df Sum of Sq    F Pr(>F)  
1         94 795.91  
2         91 731.25  3    64.664 2.6823 0.05141 .
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



DCP098

Fundamentos para Avaliação Quantitativa de Políticas Públicas

**Regressão multivariada: precisão.
Multicolinearidade.**

Aula 18
01 de junho de 2022

Ana Paula Karruz