Intro. Econometria Usando R
- Aula 3-

Prof. Mestre. Omar Barroso Khodr

Instituto Brasileiro de Educação, Pesquisa e Desenvolvimento

Tópicos

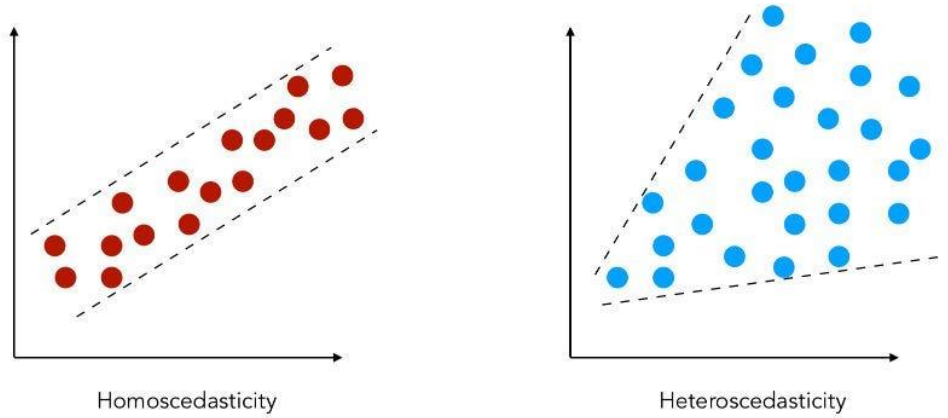
- Problemas em modelos de Regressão (heterocedasticidade e formas de detecção).
- Correções: testes de Breusch-Pagan e White.
- Modelos com variáveis categóricas (dummy) e interações.

Heterocedasticidade

- A heterocedasticidade é um fenômeno que viola uma das principais premissas do Modelo de Regressão Linear Clássico (MRLC), especificamente a suposição de homocedasticidade (variância constante dos erros). Seu impacto afeta a **eficiência** dos estimadores e a validade de **inferências estatísticas**.

Heterocedasticidade

- A heterocedasticidade ocorre quando a variância do termo de erro ε_i não é constante para todas as observações, ou seja:
- $Var(\varepsilon_i|x_i) = \sigma_i^2 \neq \sigma^2$ (variação constante)
- $Var(\varepsilon_i|x_i) = \sigma^2$ (variância constante)



Fonte: Yemelyanov, 2020

Heterocedasticidade

- Dois casos que consideraremos em detalhes são heterocedasticidade e autocorrelação. Distúrbios são heterocedasticidade quando têm variâncias diferentes.
- A heterocedasticidade geralmente surge em **dados de séries temporais voláteis** de alta frequência, como observações diárias em mercados financeiros e em dados transversais, onde a escala da variável dependente e o poder explicativo do modelo tendem a variar entre as observações.

Heterocedasticidade

- Omissão de variáveis relevantes ou forma funcional incorreta (ex: usar modelo linear quando a relação é quadrática).
- Observações extremas (outliers) podem inflar a variância em certas regiões dos dados.

Heterocedasticidade (testes de correção)

- Teste Breusch-Pagan (BP).
- Pergunta: *"Os resíduos ao quadrado (erro²) têm relação com as variáveis explicativas do modelo?"*
- Se os erros são homocedásticos (variância constante), os resíduos² não devem ter padrão sistemático.
- Se forem heterocedásticos, os resíduos² podem aumentar/diminuir com alguma variável (ex: tempo e renda).

Heterocedasticidade (testes de correção)

- Teste Breusch-Pagan (BP).
- Passos do Teste (simplificado):
- Estimar o modelo original (ex: y = xb) e obter os resíduos.
- Regredir os resíduos² contra as variáveis explicativas originais (ex: resíduos² =x).
- Testar significância:
- Se as **variáveis explicativas forem significativas (Pv < 0.05)** para prever resíduos², há evidência de heterocedasticidade.

Heterocedasticidade (testes de correção)

- Teste White.
- Pergunta: *"Os resíduos² têm relação não só com as variáveis originais, mas também com seus quadrados e interações?"*
- Lógica:
- O teste de White é uma extensão do BP que captura não-linearidades (ex: variância que aumenta com o quadrado de x).
- Inclui termos como x^2 e x1 * x2 na regressão dos resíduos².

Heterocedasticidade (testes de correção)

- Teste White.
- Estimar o modelo original (ex: y = x1 + x2).
- Regredir os resíduos² contra:
- Variáveis originais (x1, x2).
- Seus quadrados ($x1^2$, $x2^2$).
- Interações (x1 * x2).
- Se pelo menos um termo for **significativo**, há heterocedasticidade.

Intro. Econometria Usando R
- Aula 4-

Prof. Mestre. Omar Barroso Khodr

Instituto Brasileiro de Educação, Pesquisa e Desenvolvimento

- **Variáveis Categóricas**
- **Variáveis Dummy**

Variáveis Categóricas

- Até agora, encontramos apenas exemplos com *variáveis contínuas*.
- Existem muitas situações em que faz sentido pensar nos dados em termos de categorias, em vez de números contínuos. Por exemplo, se uma observação i é masculina ou feminina, se um pixel em uma tela é preto ou branco e se um produto foi produzido na França, Alemanha, Itália, China ou Espanha, são todas classificações categóricas de dados.
- Provavelmente, o tipo mais simples de variável categórica é a **binária**, booleana ou simplesmente variável *dummy*. Como o nome sugere, ela pode assumir apenas dois valores: 0 e 1, ou VERDADEIRO e FALSO.
- Representamos que uma determinada observação i é membro de uma determinada categoria j.

Variáveis Categóricas

- Por exemplo,
- $m_i = \begin{cases} 1;\ se\ i\ \text{é}\ m \\ 0\ caso\ contrário \end{cases}$
- $f_i = \begin{cases} 1;\ se\ i\ \text{é}\ f \\ 0; c.c. \end{cases}$
- Por definição, acabamos de introduzir uma dependência linear em nosso conjunto de dados. Sempre será verdade que m + f = 1. Isso ocorre porque variáveis fictícias são baseadas em dados categorizados mutuamente de forma exclusiva — aqui, você é m ou f.

Variáveis Categóricas

- Por exemplo,
- $m_i = \begin{cases} 1;\ se\ i\ \text{é}\ m \\ 0\ caso\ contrário \end{cases}$
- $f_i = \begin{cases} 1;\ se\ i\ \text{é}\ f \\ 0; c.c. \end{cases}$
- Por definição, acabamos de introduzir uma dependência linear em nosso conjunto de dados. Sempre será verdade que m + f = 1. Isso ocorre porque variáveis fictícias são baseadas em dados categorizados mutuamente de forma exclusiva — aqui, você é m ou f.

Variáveis Categóricas

- Suponha que queremos ver a relação de gênero e renda (y).
- $y_i = b_0 + b_1 f_i + b_2 m_i + e_i$
- Nesse contexto, essa relação seria inválida devido à colinearidade perfeita entre f e m.
- Em regressões com variáveis fictícias, removemos uma categoria da regressão (por exemplo, aqui: *m*) e a chamamos de categoria de referência. O efeito de ser *m* é **absorvido no intercepto**.
- O coeficiente nas categorias restantes mede a diferença no resultado médio em relação à categoria de referência.

Variáveis Categóricas

- Desta maneira, consideramos apenas...
- $y_i = b_0 + b_1 f_i + e_i$
- $f_i = \begin{cases} 1;\ se\ i\ \text{é}\ f \\ 0; c.c. \end{cases}$
- Vamos testar pelo R com exemplo distinto...

Variáveis Dummy (Exemplo)

- Suponha que somos analistas econômicos em uma empresa de consultoria publica e privada.
- Nessa semana, estamos trabalho em um projeto que analisa os hábitos de consumo de combustível de automóveis entre homens e mulheres.
- Nosso principal método de medida é ‘Milhas por Galão’ e também vamos considerar o peso dos veículos de acordo com marca e modelo.

Variáveis Dummy (Exemplo)

- Nosso principal método de medida é ‘Milhas por Galão’ e também vamos considerar o peso dos veículos de acordo com marca e modelo. Assim, nossa regressão fica como:
- $mpg_i = b_0 + b_1 wt_i + b_2 f_i + b_2 m_i + e_i$
- Queremos especificamente saber os hábitos femininos de consumo.
- $mpg_i = b_0 + b_1 wt_i + b_2 f_i + e_i$

Variáveis Dummy (Exemplo)

- O coeficiente para generoMulher mostra a diferença no consumo entre mulheres e homens, mantendo o peso constante.
- Na primeira regressão (1), generoHomem é absorvido pelo intercepto.
- Nesse contexto, generoMulher apresenta um coeficiente de 3.15 com $p < 0.05$, concluiríamos que mulheres têm em média 3.15 mpg a mais que homens para carros de mesmo peso.

	Dependent variable:	
	mpg	
	(1)	(2)
wt	-4.443***	-4.443***
	(0.613)	(0.613)
generoHomem		33.004***
		(2.355)
generoMulher	3.154**	36.159***
	(1.191)	(1.766)
Constant	33.004***	
	(2.355)	
Observations	32	32
R^2	0.801	0.984
Adjusted R^2	0.787	0.982
Residual Std. Error (df = 29)	2.780	2.780
F Statistic	58.361*** (df = 2; 29)	596.072*** (df = 3; 29)
Note:		*p<0.1; **p<0.05; ***p<0.01

Variáveis Dummy (Exemplo)

- Na segunda regressão (2) considerando os Homens...
- O modelo estimado é:
- mpg = -4.4428*wt + 33.0042*generoHomem + 36.1586*generoMulher
- Mulheres têm um intercepto 3.1544 mpg maior que homens (36.1586 - 33.0042).

	Dependent variable:	
	mpg	
	(1)	(2)
wt	-4.443***	-4.443***
	(0.613)	(0.613)
generoHomem		33.004***
		(2.355)
generoMulher	3.154**	36.159***
	(1.191)	(1.766)
Constant	33.004***	
	(2.355)	
Observations	32	32
R^2	0.801	0.984
Adjusted R^2	0.787	0.982
Residual Std. Error (df = 29)	2.780	2.780
F Statistic	58.361*** (df = 2; 29)	596.072*** (df = 3; 29)
Note:		*p<0.1; **p<0.05; ***p<0.01

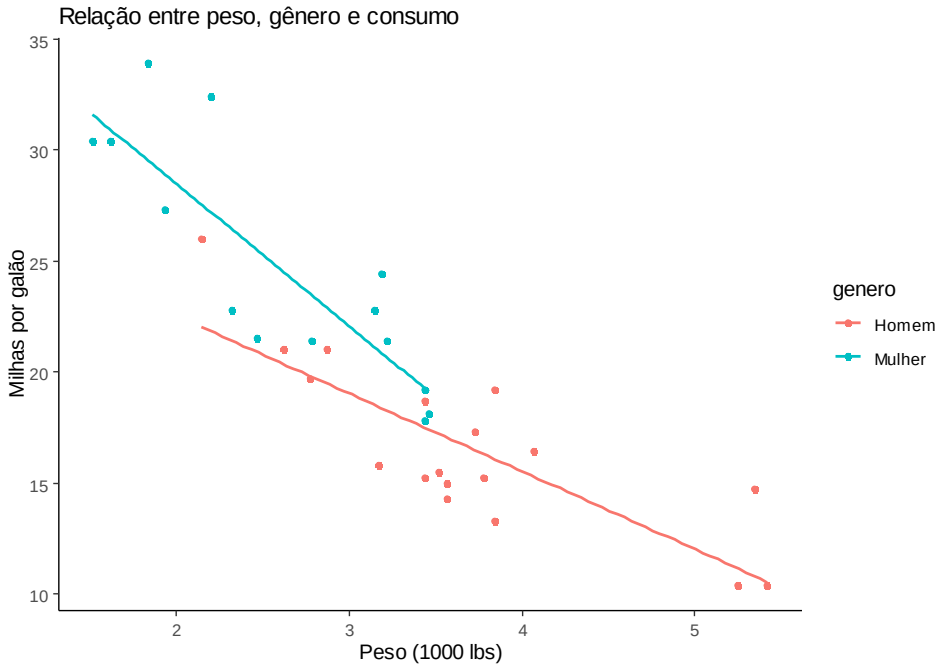
Variáveis Dummy (Exemplo)

- Isso sugere que, para um peso fixo, mulheres tendem a ter um consumo de combustível ligeiramente melhor que homens.
- Efeito do peso (wt) é o mesmo para ambos os grupos (o coeficiente de wt é compartilhado).
- Mulheres têm um consumo ligeiramente melhor que homens para um mesmo peso.
- O modelo explica 98.4% da variabilidade em mpg ($R^2 = 0.984$), indicando um excelente ajuste.

	Dependent variable:	
	mpg	
	(1)	(2)
wt	-4.443***	-4.443***
	(0.613)	(0.613)
generoHomem		33.004***
		(2.355)
generoMulher	3.154**	36.159***
	(1.191)	(1.766)
Constant	33.004***	
	(2.355)	
Observations	32	32
R^2	0.801	0.984
Adjusted R^2	0.787	0.982
Residual Std. Error (df = 29)	2.780	2.780
F Statistic	58.361*** (df = 2; 29)	596.072*** (df = 3; 29)
Note:		*p<0.1; **p<0.05; ***p<0.01

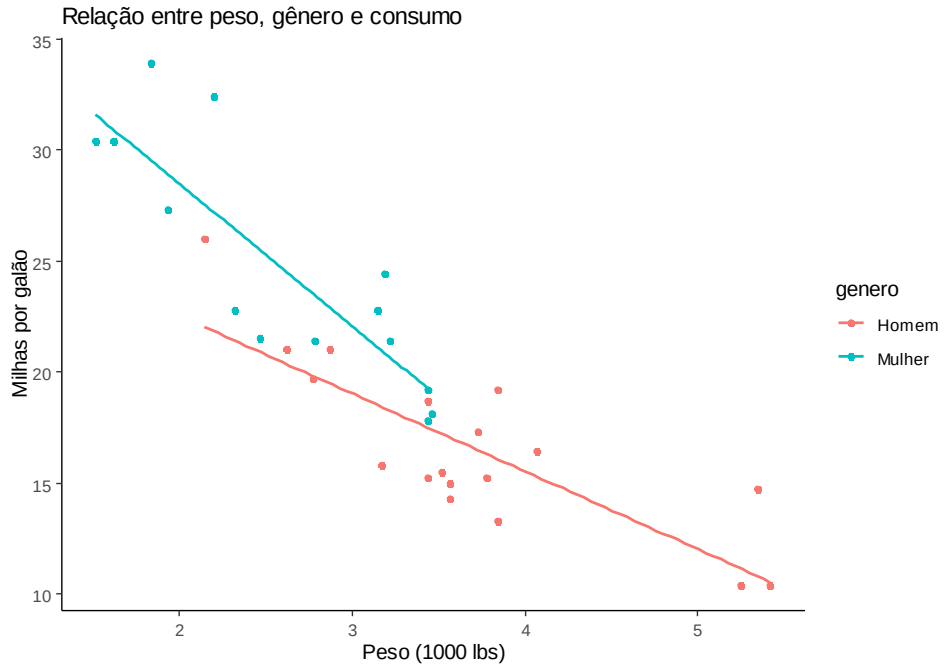
Variáveis Dummy (Exemplo)

- Neste modelo, cada categoria de gênero tem seu próprio intercepto (Homem e Mulher).
- Isso permite comparar diretamente os dois grupos sem depender de uma categoria de referência.



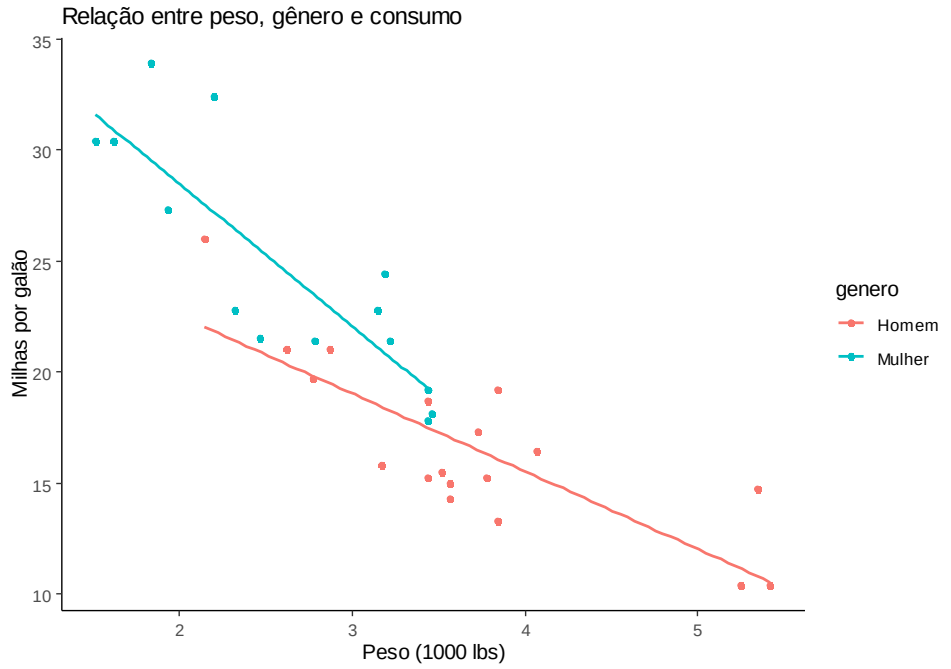
Como o R remedia colinearidade entre Dummies

- Em modelos COM intercepto, o R automaticamente remove uma das dummies para evitar colinearidade.
- Ex.: Se genero tem níveis Homem e Mulher, o R usará apenas generoMulher (considerando Homem como referência).
- A equação seria:
- $mpg = \beta_0 + \beta_1 wt + \beta_2 generoMulher$



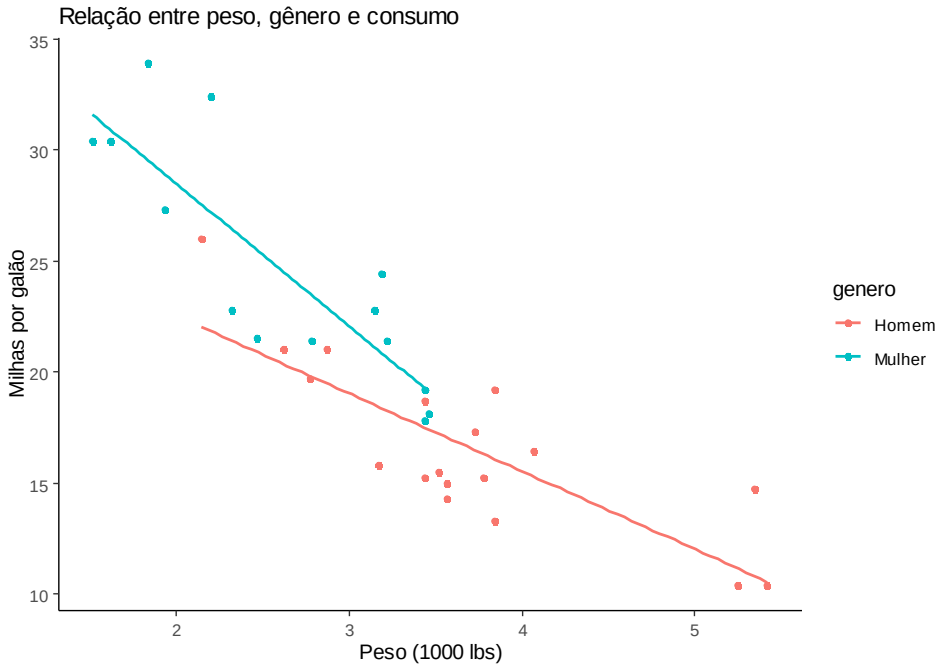
Como o R remedia colinearidade entre Dummies

- Em modelos COM intercepto, o R automaticamente remove uma das dummies para evitar colinearidade.
- Ex.: Se gênero tem níveis Homem e Mulher, o R usará apenas generoMulher (considerando Homem como referência).
- A equação seria:
- $mpg = \beta_0 + \beta_1 wt + \beta_2 generoMulher$



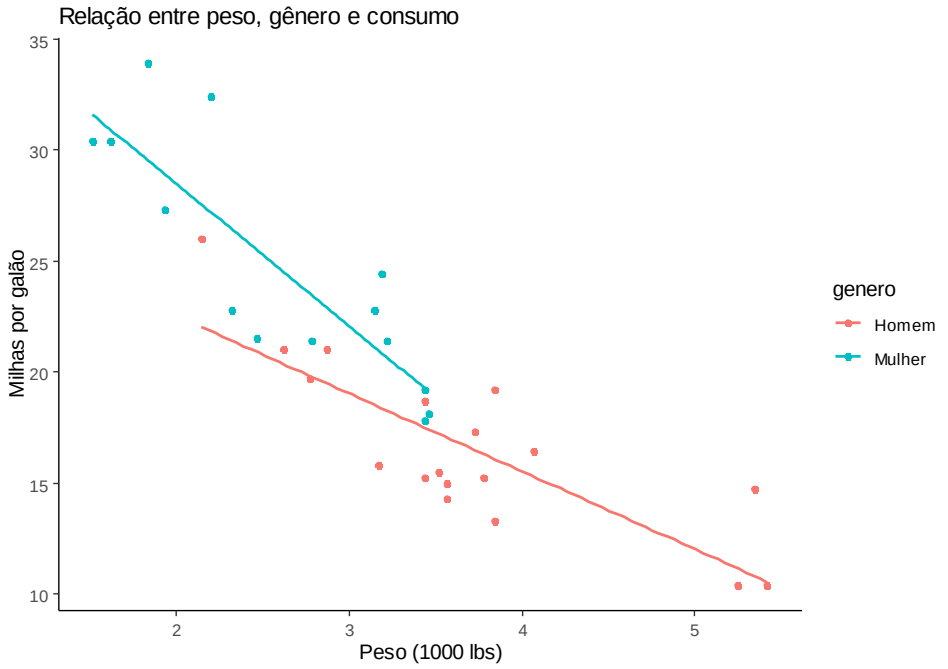
Como o R remedia colinearidade entre Dummies

- Em modelos SEM intercepto (- 1), o R mantém todas as dummies, pois não há intercepto para causar colinearidade.
- Nesse caso, cada dummy representa o intercepto específico da sua categoria.
- $mpg = \beta_1 wt + \beta_2 Homem + \beta_3 Mulher$



Como o R remedia colinearidade entre Dummies

- Matematicamente:
- A matriz de design (X) do modelo tem:
- Uma coluna para wt
- Uma coluna para generoHomem (1 se Homem, 0 se Mulher)
- Uma coluna para generoMulher (1 se Mulher, 0 se Homem)
- Essas colunas não são linearmente dependentes porque não há uma coluna de "1"s (intercepto) para vinculá-las.



Exemplo em Salários

- Suponha o modelo,
- $lnw_i = b_0 + b_1 educ_i + b_2 f_i + e_i$
- Voltando ao nosso exemplo salarial, queremos saber se o salário de um trabalhador é explicado pela sua educação e também do seu gênero.

Exemplo em Salários

- Suponha o modelo,
- $lnw_i = b_0 + b_1 educ_i + b_2 f_i + e_i$
- Voltando ao nosso exemplo salarial, queremos saber se o salário de um trabalhador é explicado pela sua educação e também do seu gênero.

	Dependent variable:	
	lwage	
	(1)	(2)
educ	0.083***	0.077***
	(0.008)	(0.007)
female1		-0.361***
		(0.039)
Constant	0.584***	0.826***
	(0.097)	(0.094)
Observations	526	526
R^2	0.186	0.300
Adjusted R^2	0.184	0.298
Residual Std. Error	0.480 (df = 524)	0.445 (df = 523)
F Statistic	119.582*** (df = 1; 524)	112.189*** (df = 2; 523)
Note:		*p<0.1; **p<0.05; ***p<0.01

Exemplo em Salários

- Já conhecemos muito bem os resultados da coluna (1). Como a relação muda se incluirmos o indicador feminino? Lembre-se do que foi dito acima que feminino é um fator com dois níveis, 0 e 1, onde 1 significa que é uma mulher.

	Dependent variable:	
	lwage	
	(1)	(2)
educ	0.083^{***}	0.077^{***}
	(0.008)	(0.007)
female1		-0.361^{***}
		(0.039)
Constant	0.584^{***}	0.826^{***}
	(0.097)	(0.094)
Observations	526	526
R^2	0.186	0.300
Adjusted R^2	0.184	0.298
Residual Std. Error	0.480 (df = 524)	0.445 (df = 523)
F Statistic	119.582^{***} (df = 1; 524)	112.189^{***} (df = 2; 523)
Note:		*p<0.1; **p<0.05; ***p<0.01

Exemplo em Salários

- Já conhecemos muito bem os resultados da coluna (1). Como a relação muda se incluirmos o indicador feminino? Lembre-se do que foi dito acima que feminino é um fator com dois níveis, 0 e 1, onde 1 significa que é uma mulher.

	Dependent variable:	
	lwage	
	(1)	(2)
educ	0.083^{***}	0.077^{***}
	(0.008)	(0.007)
female1		-0.361^{***}
		(0.039)
Constant	0.584^{***}	0.826^{***}
	(0.097)	(0.094)
Observations	526	526
R^2	0.186	0.300
Adjusted R^2	0.184	0.298
Residual Std. Error	0.480 (df = 524)	0.445 (df = 523)
F Statistic	119.582^{***} (df = 1; 524)	112.189^{***} (df = 2; 523)
Note:		*p<0.1; **p<0.05; ***p<0.01

Exemplo em Salários

- Mas você também pode observar que há uma mudança paralela para baixo da linha masculina para a feminina. A estimativa de $b^2 = -0,36$ é a magnitude da mudança para baixo.

	Dependent variable:	
	lwage	
	(1)	(2)
educ	0.083***	0.077***
	(0.008)	(0.007)
female1		-0.361***
		(0.039)
Constant	0.584***	0.826***
	(0.097)	(0.094)
Observations	526	526
R^2	0.186	0.300
Adjusted R^2	0.184	0.298
Residual Std. Error	0.480 (df = 524)	0.445 (df = 523)
F Statistic	119.582*** (df = 1; 524)	112.189*** (df = 2; 523)
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.