ESTADO DE MATO GROSSO
SECRETARIA DE ESTADO DE CIÊNCIA E TECNOLOGIA
UNIVERSIDADE DO ESTADO DE MATO GROSSO
CAMPUS UNIVERSITÁRIO DE SINOP
CURSO DE CIÊNCIAS ECONÔMICAS

Econometria I: Aplicações

Laboratório 4 - Análise de regressão no Microsoft Excel

Profº Lindomar Pegorini Daniel[1]

[1] Professor Adjunto da Universidade do Estado de Mato Grosso (UNEMAT) – Campus de Sinop.

LABORATÓRIO 4 – ANÁLISE DE REGRESSÃO NO MICROSOFT EXCEL

Visão geral

Nos laboratórios anteriores, você explorou e analisou um conjunto de dados contendo detalhes das vendas de limonada. Neste laboratório, você aplicará a técnica de análise de regressão para prever a quantidade esperada de vendas para dias com características específicas. Além da previsão de valores de variáveis econômicas, a técnica de análise de regressão permite estimar relações entre variáveis econômicas, testar hipóteses da teoria econômica, avaliar a efetividade de políticas, dentre outras aplicações.

Do que você vai precisar

Para completar este laboratório você irá precisar:

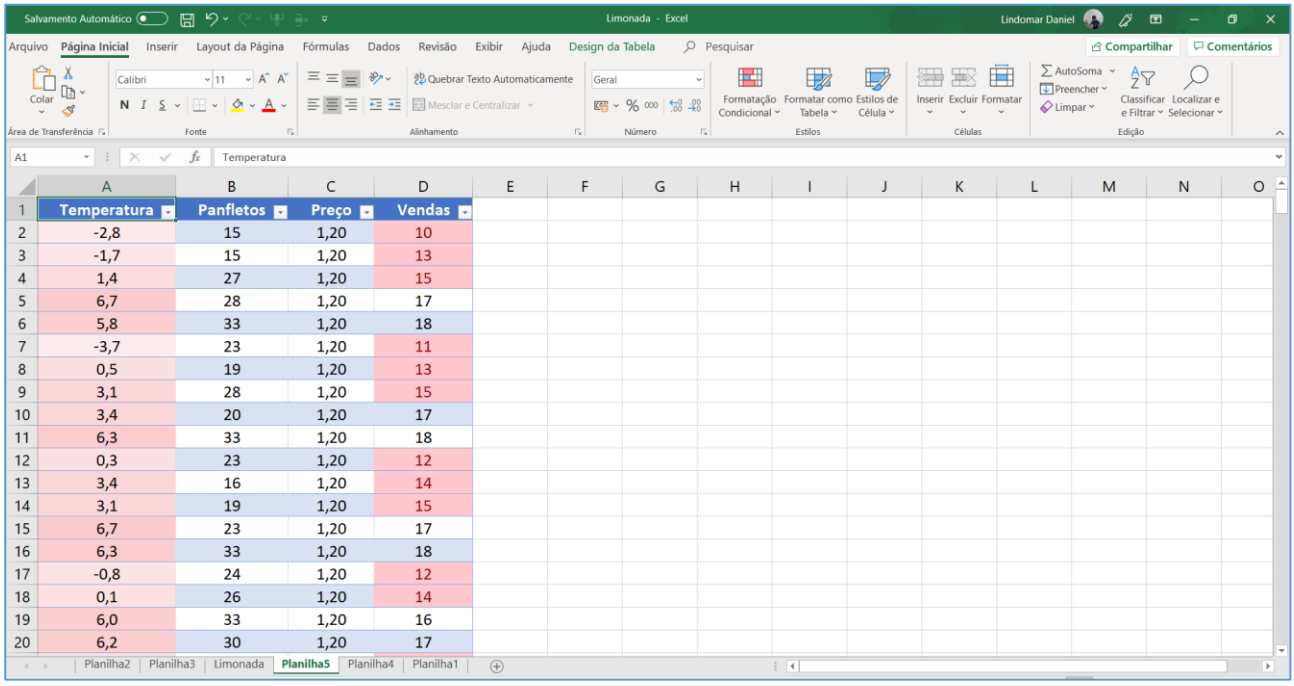
- Dos arquivos de dados **Limonada.xlsx**, **Salário por hora.xls** e **Função pulmonar.xls**;
- De um computador com um aplicativo gerenciador de planilhas compatível com a extensão xlsx, ou, de um computador com acesso à internet e uma conta Microsoft (hotmail.com, live.com ou outlook.com) para acessar o Excel Online de forma gratuita.

EXERCÍCIO 1: Inferência estatística e análise de regressão

Outra ferramenta estatística comumente utilizada para realizar estudos e previsões é a análise de regressão. A análise de regressão consiste na análise de dependência de uma variável em relação a outras e, a partir dessa relação de dependência, é possível prever valores da variável de interesse tendo como base os valores fixados das variáveis explicativas. Por exemplo, Rosie pode estar interessada em fazer previsões do número de vendas com base nas informações registradas sobre preço, temperatura e panfletos.

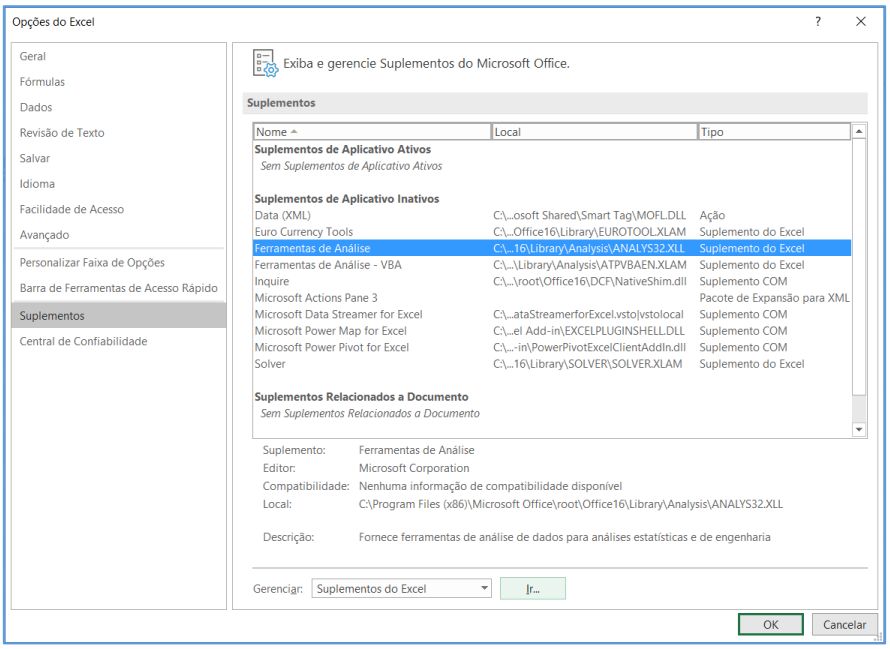
Estime uma regressão

1. Volte para a planilha **Limonada** e selecione a célula **A1** (o cabeçalho da coluna **Data**), utilize o comando CTRL + T para selecionar toda a tabela, copie e cole em uma nova planilha.
2. Na nova planilha, exclua colunas de forma que restem apenas as informações sobre **Temperatura**, **Panfletos**, **Preço** e **Vendas** para que sua planilha seja semelhante a esta:

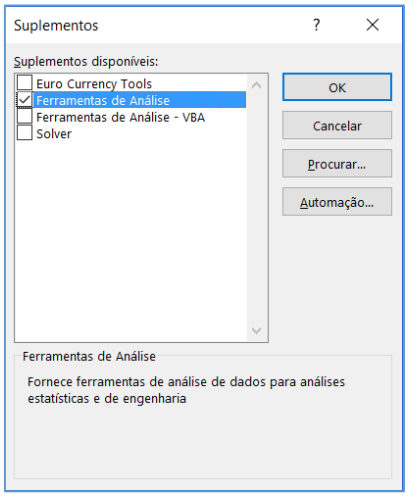


Temperatura	Panfletos	Preço	Vendas
-2,8	15	1,20	10
-1,7	15	1,20	13
1,4	27	1,20	15
6,7	28	1,20	17
5,8	33	1,20	18
-3,7	23	1,20	11
0,5	19	1,20	13
3,1	28	1,20	15
3,4	20	1,20	17
6,3	33	1,20	18
0,3	23	1,20	12
3,4	16	1,20	14
3,1	19	1,20	15
6,7	23	1,20	17
6,3	33	1,20	18
-0,8	24	1,20	12
0,1	26	1,20	14
6,0	33	1,20	16
6,2	30	1,20	17

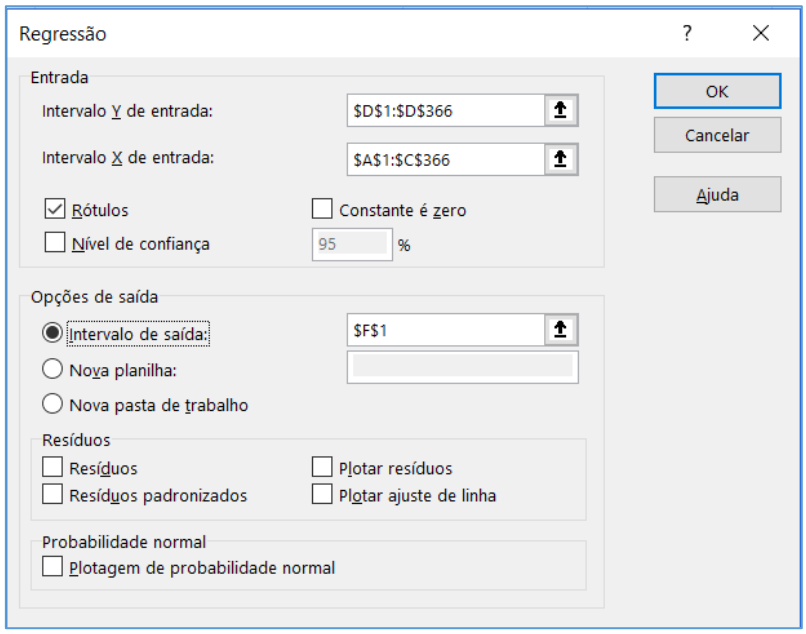
3. A aplicação de **Análise de Regressão** normalmente não está disponível no Excel e tem de ser habilitada. Na guia **Arquivo** da faixa de opções, clique em **Opções** e na caixa de diálogo no menu **Suplementos** marque a opção **Ferramentas de Análise** e, então, clique em **Ir**.



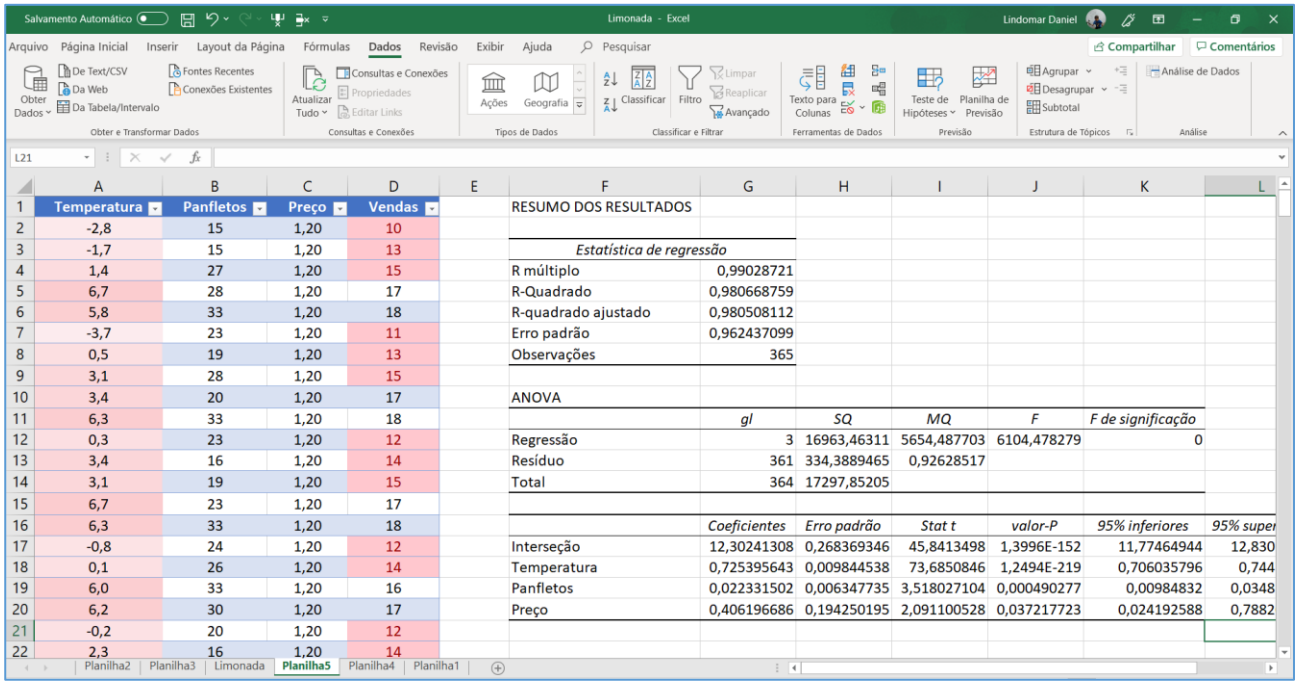
4. Na nova caixa de diálogo marque a opção **Ferramentas de Análise** e então clique em **OK**. Isso habilitará a opção **Análise de Dados** na guia **Dados** da faixa de opções.



5. Agora na guia **Dados** da faixa de opções, clique em **Análise de Dados**, selecione a opção **Regressão** e clique em **OK**. Isso abrirá uma caixa de diálogo.
6. Preencha o **Intervalo <u>Y</u> de entrada:** com a coluna onde estão as informações sobre a variável de interesse **Vendas**:
   D1:D366
7. Preencha o **Intervalo <u>X</u> de entrada:** com as colunas onde estão as informações sobre **Temperatura**, **Panfletos** e **Preço**:
   A1:C366
8. Marque a opção **Rótulos** para considerar o cabeçalho das variáveis e marque também a opção **Intervalo de saída:** e preencha com F1 para que os resultados sejam mostrados a partir dessa célula.
9. Por fim, clique em **OK**.



10. Sua planilha agora deve ser parecida com a seguinte:



Temperatura	Panfletos	Preço	Vendas
-2,8	15	1,20	10
-1,7	15	1,20	13
1,4	27	1,20	15
6,7	28	1,20	17
5,8	33	1,20	18
-3,7	23	1,20	11
0,5	19	1,20	13
3,1	28	1,20	15
3,4	20	1,20	17
6,3	33	1,20	18
0,3	23	1,20	12
3,4	16	1,20	14
3,1	19	1,20	15
6,7	23	1,20	17
6,3	33	1,20	18
-0,8	24	1,20	12
0,1	26	1,20	14
6,0	33	1,20	16
6,2	30	1,20	17
-0,2	20	1,20	12
2,3	16	1,20	14

RESUMO DOS RESULTADOS

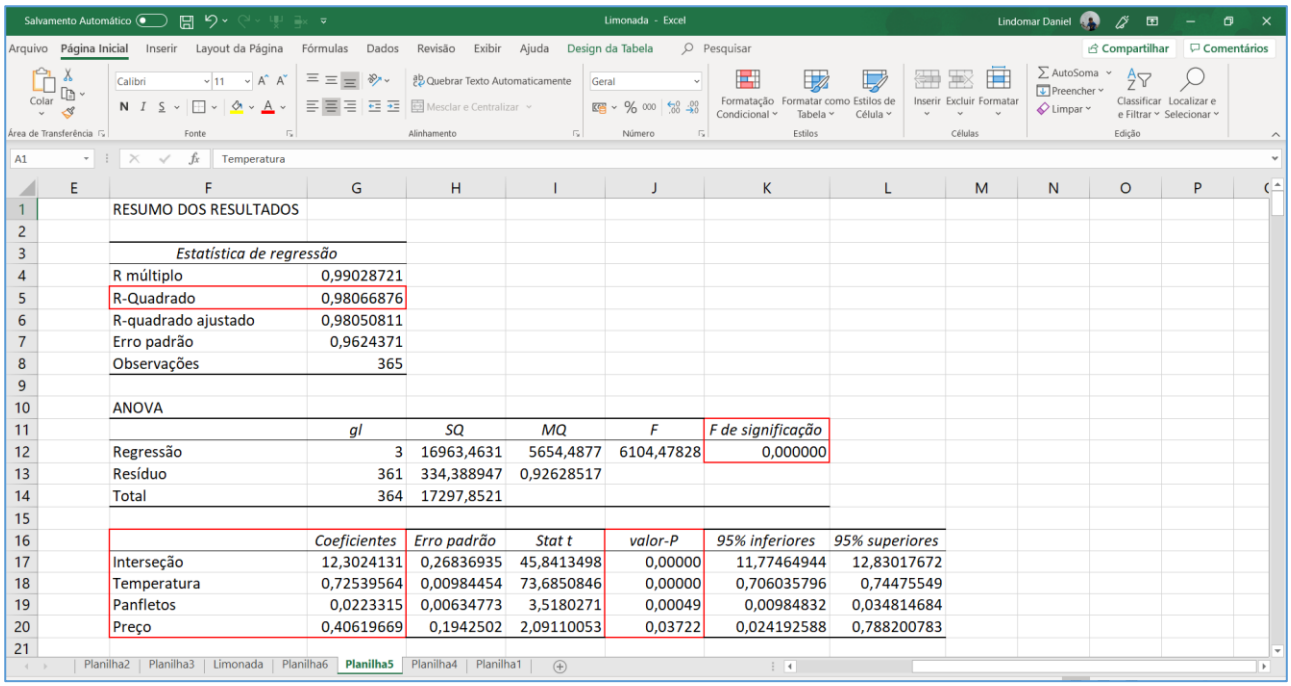
Estatística de regressão	
R múltiplo	0,99028721
R-Quadrado	0,980668759
R-quadrado ajustado	0,980508112
Erro padrão	0,962437099
Observações	365

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	3	16963,46311	5654,487703	6104,478279	0
Resíduo	361	334,3889465	0,92628517		
Total	364	17297,85205			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% super*
Interseção	12,30241308	0,268369346	45,8413498	1,3996E-152	11,77464944	12,830
Temperatura	0,725395643	0,009844538	73,6850846	1,2494E-219	0,706035796	0,744
Panfletos	0,022331502	0,006347735	3,518027104	0,000490277	0,00984832	0,0348
Preço	0,406196686	0,194250195	2,091100528	0,037217723	0,024192588	0,7882

Os resultados da análise de regressão são apresentados em 3 tabelas. A primeira apresenta as estatísticas da regressão, são informações sobre a qualidade de ajuste do modelo, ou, em outras palavras, se o modelo é bom para fazer previsões sobre as vendas de limonada, nesse caso. A segunda tabela é conhecida como análise de variância ou ANOVA, ela contém informações sobre a variância do modelo e um teste de significância geral do modelo. Já a última tabela apresenta os resultados das relações entre vendas e as demais variáveis. A planilha a seguir apresenta as informações mais relevantes para nossa análise.



	RESUMO DOS RESULTADOS						
	Estatística de regressão						
	R múltiplo	0,99028721					
	R-Quadrado	0,98066876					
	R-quadrado ajustado	0,98050811					
	Erro padrão	0,9624371					
	Observações	365					
	ANOVA						
		gl	*SQ*	*MQ*	*F*	*F de significação*	
	Regressão	3	16963,4631	5654,4877	6104,47828	0,000000	
	Resíduo	361	334,388947	0,92628517			
	Total	364	17297,8521				
		Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*
	Interseção	12,3024131	0,26836935	45,8413498	0,00000	11,77464944	12,83017672
	Temperatura	0,72539564	0,00984454	73,6850846	0,00000	0,706035796	0,74475549
	Panfletos	0,0223315	0,00634773	3,5180271	0,00049	0,00984832	0,034814684
	Preço	0,40619669	0,1942502	2,09110053	0,03722	0,024192588	0,788200783

O primeiro destaque é a medida de qualidade de ajuste **R-Quadrado** ou $\mathbf{R^2}$, ela varia de 0 a 1. Quanto mais próximo de 1 melhor é a qualidade de previsão do modelo, um valor acima de 0,80 é considerado bom. O nosso modelo para previsão de vendas apresenta um $\mathbf{R^2}$ de 0,98, ou seja, o modelo é muito bom para realizar previsões de vendas.

O segundo destaque, na tabela ANOVA, é o **F de significação**, conhecido como **teste F**, testamos se o modelo que acabamos de estimar é melhor do que a média aritmética para fins de previsão do volume de vendas. Se o valor do **F de significação** for menor que 0,05 o modelo é bom ou significativo para fazer previsões, caso seja maior que 0,05 indica que utilizar a média para prever as vendas é melhor, em outras palavras, nesse caso, as variáveis que utilizamos para prever as vendas não teriam nenhuma relação com o volume de vendas. Como observamos, o valor é menor que 0,05, ou seja, o modelo é melhor que a média para prever as vendas. O **teste F** é um complemento ao $\mathbf{R^2}$.

O terceiro destaque, na tabela de resultados, são os coeficientes das variáveis. Eles nos dizem qual é a relação quantitativa entre vendas e cada uma das variáveis que utilizamos, quais sejam, **Temperatura**, **Panfletos** e **Preço**. Podemos montar a seguinte expressão a partir do resultado:

$$\text{Vendas} = 12{,}3 + 0{,}72\text{xTemperatura} + 0{,}02\text{xPanfletos} + 0{,}40\text{xPreço}$$

O coeficiente da **Temperatura** nos indica que para cada 1ºC a mais no dia são vendidas 0,72 limonadas a mais, ou seja, se a temperatura sobe 10ºC de um dia para o outro é provável que sejam vendidas 7,2 limonadas a mais. O raciocínio é o mesmo para as demais variáveis. No caso dos **Panfletos**, para cada panfleto distribuído espera-se um aumento de 0,02 nas vendas de limonada, ou seja, a distribuição de 100 panfletos gera um aumento na expectativa de vendas da ordem de 2 limonadas.

Já o aumento do **Preço** em R$ 1,00 está associado com aumento de vendas de 0,40 limonadas, esse resultado não faz muito sentido econômico, pois o preço na maioria dos casos se relaciona de forma negativa com a quantidade vendida, no entanto, no caso da limonada caseira o preço parece responder ao aumento das vendas e não o contrário. O valor da **Interseção** (12,3) não está associado a nenhuma variável, ele indica a média de vendas de limonada caso filtrássemos a influência das demais variáveis.

De posse do modelo, Rosie ao preparar a limonada poderia levar em consideração a demanda esperada para o dia. Por exemplo, se o preço praticado no dia for R$ 2,00, a previsão do tempo indica uma temperatura média de 36ºC e ela pretende distribuir 80 panfletos, ela poderia substituir esses valores no modelo e obter a previsão de vendas para o dia:

$$\text{Vendas} = 12{,}3 + 0{,}72 \text{ x } 36 + 0{,}02 \text{ x } 80 + 0{,}40 \text{ x } 2{,}00 = 40{,}62$$

Rosie tem uma previsão de vendas de 40,62 limonadas em um dia com essas características. Por fim, cabe destacar o **valor-P**, na análise de regressão ele também é conhecido como teste de significância individual, é o mesmo do teste de hipótese. Se o **valor-P** for maior que 0,05 isso mostra que a variável em questão não possui relevância para explicar as vendas, ou, em outras palavras, que a variável é não significativa. Por outro lado, se o **valor-P** for menor que 0,05 evidencia que a variável em questão é importante para prever as vendas, ou, em outras palavras, que a variável é estatisticamente significativa. Quando uma variável é significativa indica que quando ela se altera muito provavelmente isso causará impacto/alterações na variável de interesse. Por exemplo, se a **Temperatura** aumenta é provável que as vendas de limonada também aumentem.

Desafio: Estime uma regressão com a variável chuva

1. Estime uma regressão utilizando a variável **Chuva** no lugar da variável **Temperatura** e observe os resultados.
2. Responda as questões do Enade a seguir:

Considere o modelo de regressão linear múltipla, com variável dependente **y** e variáveis explicativas $\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_k}$, que pode ser expresso como

$$y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + \varepsilon_t$$

no qual ε_t significa o fator de erro e t = 1, 2, …, no índice relativo às observações amostrais.

É CORRETO afirmar que o modelo clássico de Gauss de regressão linear supõe que

(A) a relação linear entre pelo menos duas variáveis explicativas seja exata.

(B) a variância dos erros varie na amostra: $E(\varepsilon_t^2) \neq E(\varepsilon_z^2)$ para $t \neq z$.

(C) o valor esperado do fator de erro seja diferente de zero: $E(\varepsilon_t) \neq 0$.

(D) os erros não sejam correlacionados $E(\varepsilon_t \varepsilon_z) = 0$ para $t \neq z$.

(E) os valores das variáveis explicativas, $X_1, X_2, \dots, X_k$ variem de amostra para amostra.

Um órgão regulatório, a fim de reduzir os impactos da poluição atmosférica causada por organizações, decidiu aplicar multas conforme a quantidade de poluentes emitidos na atmosfera pelas empresas, de modo que, quanto maior fosse essa quantidade, maior seria o valor da multa aplicada.

Ciente dessa decisão, determinada empresa contratou um consultor para realizar um levantamento da relação entre a quantidade de poluentes liberados por ela na atmosfera, o valor das multas aplicadas pelo órgão e os lucros da empresa (em 10 000 u.m.). O objetivo desse levantamento era avaliar o impacto das multas (variável independente) nos lucros da empresa (variável dependente). Os resultados desse estudo estão resumidos nas tabelas a seguir.

RESUMO DOS RESULTADOS	
Estatística de regressão	
R múltiplo	0,90
R-Quadrado	0,81
R-quadrado ajustado	0,79
Erro padrão	11,84

	Coeficientes	Erro padrão	Estatística T	P-valor
Interseção	151,44	9,23	16,41	1,47 E-08
Multas devido aos poluentes	- 0,37	0,06	- 6,47	7,15 E-05

Considerando a situação hipotética apresentada, avalie as afirmações a seguir.

I. Rejeitando-se a hipótese nula (Ho: β = 0), evidencia-se uma relação linear negativa entre as multas e o lucro da empresa.
II. A não rejeição da hipótese nula (Ho: β = 0) sugere que multas podem incentivar a redução da poluição, uma vez que diminuem o lucro da firma.
III. Ao nível de significância de 0,05, rejeita-se a hipótese nula (Ho: β = 0), podendo-se fazer estimativas do lucro da empresa por meio da reta Y = 151,44 - 0,37 X.

É correto o que se afirma em

(A) II, apenas.

(B) III, apenas.

(C) I e II, apenas.

(D) I e III, apenas.

(E) I, II e III.

A decisão do Comitê de Política Monetária (Copom) do Banco Central do Brasil, divulgada em 29/07/2015, de elevar a taxa de juros básica da economia em 0,5% para o nível de 14,25% ao ano, corresponde a uma contração monetária e implica menor nível de atividade econômica nos próximos períodos. A relação entre o nível de atividade econômica, medido pela taxa de desemprego, e a taxa de inflação é estudada por meio da versão moderna da Curva de Phillips. Os dados mensais no período no período de novembro de 2012 a abril de 2015 (30 observações) para essas duas variáveis foram obtidos na página do Instituto de Pesquisa Econômica Aplicada (IPEA). A componente cíclica do desemprego foi obtida utilizando-se o filtro Hodrick-Prescott.

A Curva de Phillips aumentada pelas expectativas pode ser especificada como:

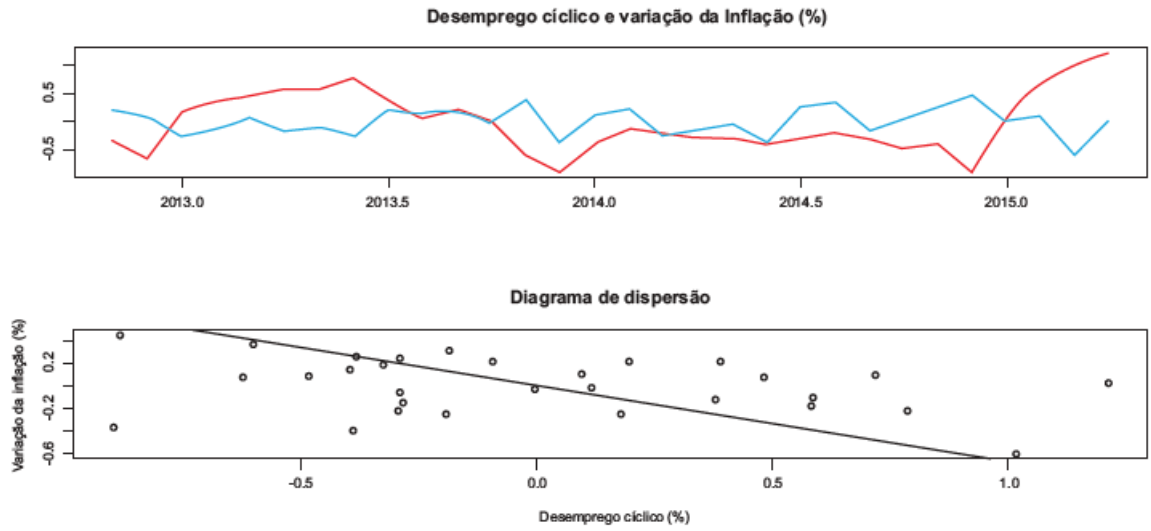
$$\pi_t = \pi^e - \varepsilon(u_t - u_n)$$

em que se presume que as expectativas de inflação são a própria inflação passada $\pi^e = \pi_{t-1}$, o que resulta em

$$\Delta\pi = -\varepsilon(u_t - u_n)$$

em que $\Delta\pi = \pi_t - \pi_{t-1}$ é a variação da inflação, u_t é o nível de desemprego, u_n é a taxa natural de desemprego e ε é a resposta da inflação aos desvios do desemprego em relação à taxa natural de desemprego.

Utilizando-se um modelo de regressão simples entre a variação da inflação e o ciclo de desemprego, o coeficiente ε foi estimado em -0,7, com uma probabilidade exata do teste (p-valor ou nível empírico de significância) associado de 0,0865. Nas figuras abaixo, estão dispostos a evolução da variação da inflação (em azul) e do ciclo de desemprego (em vermelho) e também o diagrama de dispersão para essas duas variáveis, com reta de regressão ajustada (em preto) para os dados da economia brasileira.



Fonte: IPEA.

Considerando a teoria da Curva de Phillips e os resultados obtidos para a economia brasileira, avalie as afirmações a seguir.

I. Se a decisão de contração monetária do Banco Central do Brasil for crível, a Curva de Phillips será deslocada paralelamente para baixo.
II. Para um nível de significância de 0,05 é possível inferir que há evidência de relação entre inflação e desemprego no Brasil.
III. Se a taxa natural de desemprego for de 4,5% ao mês, espera-se que a inflação seja 1,75% menor, se o nível de desemprego atingir 7% ao mês.
IV. Se o Banco Central desejar manter a taxa de inflação constante (sem variação), para uma taxa natural de desemprego de 4,5%, o nível de desemprego deve ser de 6,5%.

É correto o que se afirma em

(A) I.

(B) II.

(C) I e III.

(D) II e IV.

(E) III e IV.

Um pesquisador resolveu estimar uma versão da Lei de Okun para determinado país X. O resultado é apresentado na equação a seguir.

$$u_t = u_n - 0{,}5gy_t + e_t$$

em que u_t é a taxa de desemprego observada para o ano t ; u_n é a taxa de desemprego natural; gy_t é a taxa de crescimento do produto no ano t ; e_t é o termo de resíduo. O país apresenta uma taxa de desemprego natural igual a 10%.

Com o objetivo de analisar a predição do modelo, esse pesquisador utilizou os dados a seguir, para alguns anos selecionados.

Dados anuais selecionados do país X

Ano	Taxa de Crescimento do Produto	Taxa de Desemprego Observada
2013	4%	6%
2014	8%	5%
2015	4%	7%
2016	2%	10%
2017	10%	5%

Considerando as informações apresentadas, assinale a opção correta.

(A) Para o ano de 2013, o modelo previu uma taxa de desemprego inferior à observada.

(B) Para o ano de 2014, a taxa de desemprego estimada foi igual à observada.

(C) Para o ano de 2015, o modelo superestimou a taxa de desemprego.

(D) Para o ano de 2016, o erro de previsão do modelo foi igual a zero.

(E) Para o ano de 2017, o erro de previsão do modelo foi positivo.

3. Lembre-se de que você vai precisar dos resultados para responder os exercícios posteriormente.

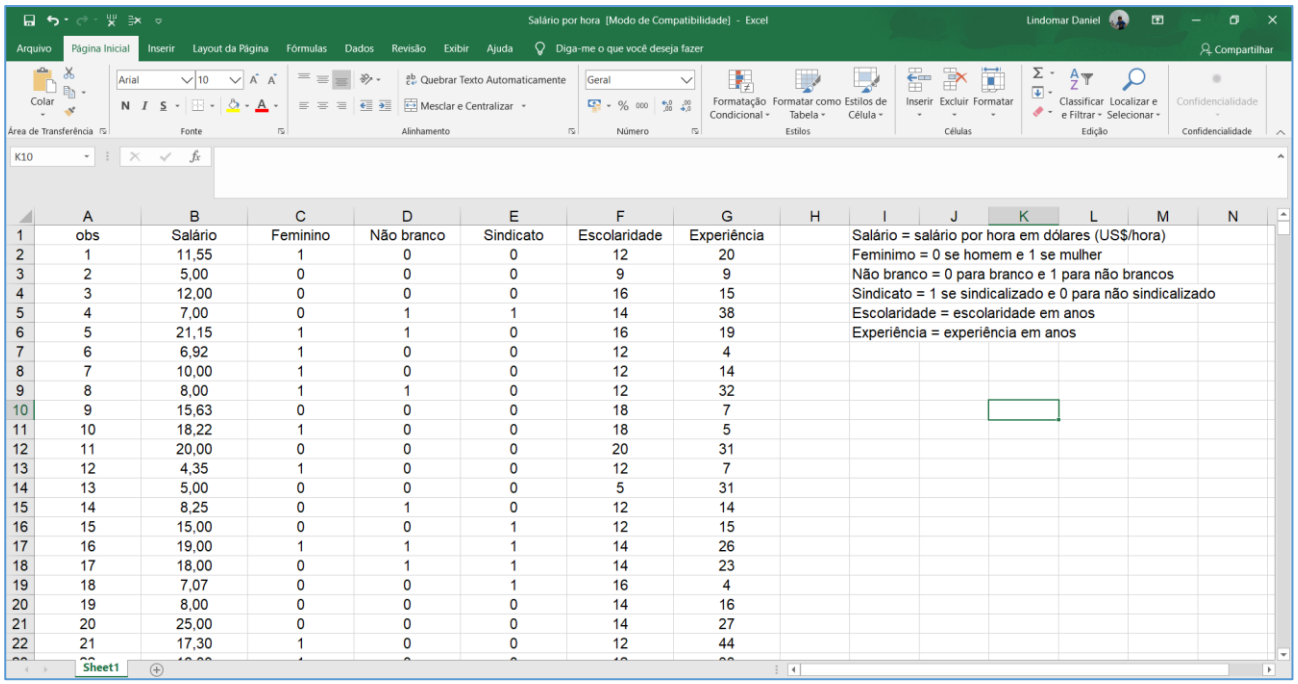
EXERCÍCIO 2: Determinantes do salário por hora

Como vimos no exercício anterior podemos utilizar a análise de regressão para fazer projeções. A regressão ajuda você a entender como as variáveis estão relacionadas e usar a informação para entender o comportamento de determinado fenômeno. Uma relação muito estudada na economia é a da educação com salários. Segundo a teoria do capital humano, quanto maior o nível de escolaridade do indivíduo, maior tende a ser a sua produtividade, portanto, maior será o seu salário.

Nesse exemplo exploraremos com maiores detalhes os resultados da regressão.

Estime uma regressão dos determinantes do salário por hora

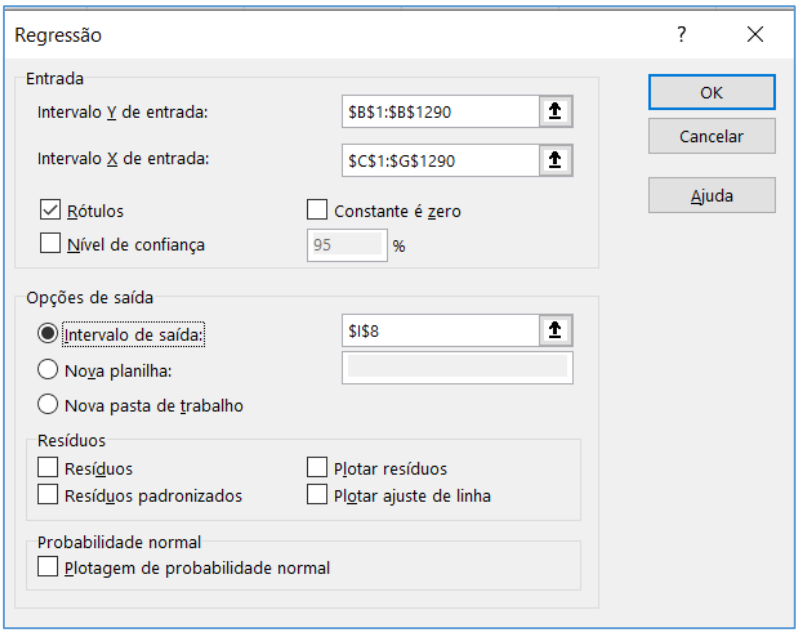
1. O arquivo **Salário por hora.xls** apresenta dados sobre salário e escolaridade de 1289 pessoas nos Estados Unidos em 1995.
2. Ao abrir o arquivo **Salário por hora.xlsx**, ele deveria parecer com o seguinte:



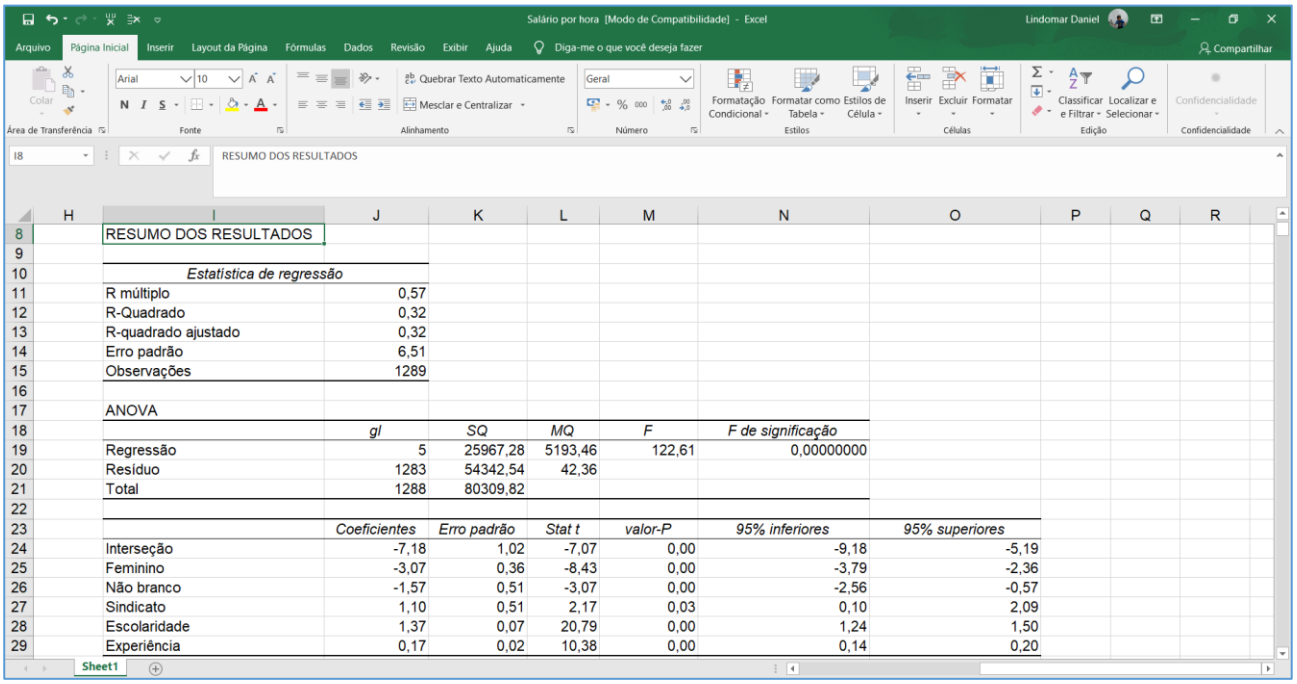
obs	Salário	Feminino	Não branco	Sindicato	Escolaridade	Experiência
1	11,55	1	0	0	12	20
2	5,00	0	0	0	9	9
3	12,00	0	0	0	16	15
4	7,00	0	1	1	14	38
5	21,15	1	1	0	16	19
6	6,92	1	0	0	12	4
7	10,00	1	0	0	12	14
8	8,00	1	1	0	12	32
9	15,63	0	0	0	18	7
10	18,22	1	0	0	18	5
11	20,00	0	0	0	20	31
12	4,35	1	0	0	12	7
13	5,00	0	0	0	5	31
14	8,25	0	1	0	12	14
15	15,00	0	0	1	12	15
16	19,00	1	1	1	14	26
17	18,00	0	1	1	14	23
18	7,07	0	0	1	16	4
19	8,00	0	0	0	14	16
20	25,00	0	0	0	14	27
21	17,30	1	0	0	12	44

Salário = salário por hora em dólares (US$/hora)
Feminino = 0 se homem e 1 se mulher
Não branco = 0 para branco e 1 para não brancos
Sindicato = 1 se sindicalizado e 0 para não sindicalizado
Escolaridade = escolaridade em anos
Experiência = experiência em anos

3. Agora na guia **Dados** da faixa de opções, clique em **Análise de Dados**, selecione a opção **Regressão** e clique em **OK**. Isso abrirá uma caixa de diálogo.
4. Preencha o **Intervalo <u>Y</u> de entrada:** com a coluna onde estão as informações sobre a variável de interesse **Salário**:
   B1:B1290
5. Preencha o **Intervalo <u>X</u> de entrada:** com as colunas onde estão as variáveis que explicarão o salário por hora: **Feminino**, **Não branco**, **Sindicato**, **Escolaridade** e **Experiência**:
   C1:G1290
6. Marque a opção **Rótulos** para considerar o cabeçalho das variáveis e marque também a opção **Intervalo de saída:** e preencha com I8 para que os resultados sejam mostrados a partir dessa célula.
7. Por fim, clique em **OK**.



8. Sua planilha agora deve ser parecida com a seguinte, talvez seja necessário rolar para o lado e para baixo para visualizar:



RESUMO DOS RESULTADOS						
	Estatística de regressão					
	R múltiplo	0,57				
	R-Quadrado	0,32				
	R-quadrado ajustado	0,32				
	Erro padrão	6,51				
	Observações	1289				
	ANOVA					
		gl	*SQ*	*MQ*	*F*	*F de significação*
	Regressão	5	25967,28	5193,46	122,61	0,00000000
	Resíduo	1283	54342,54	42,36		
	Total	1288	80309,82			
		Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*
	Interseção	-7,18	1,02	-7,07	0,00	-9,18	-5,19
	Feminino	-3,07	0,36	-8,43	0,00	-3,79	-2,36
	Não branco	-1,57	0,51	-3,07	0,00	-2,56	-0,57
	Sindicato	1,10	0,51	2,17	0,03	0,10	2,09
	Escolaridade	1,37	0,07	20,79	0,00	1,24	1,50
	Experiência	0,17	0,02	10,38	0,00	0,14	0,20

No Excel, os resultados da análise de regressão são apresentados em 3 tabelas. Vejamos agora em detalhes cada uma delas.

A primeira apresenta as estatísticas da regressão, são informações sobre a qualidade de ajuste do modelo, ou, em outras palavras, se o modelo é bom para explicar o **Salário por hora**, a variável dependente, a partir das variáveis explicativas.

Estatística de regressão	
R múltiplo	0,57
R-Quadrado	0,32
R-quadrado ajustado	0,32
Erro padrão	6,51
Observações	1289

-**R múltiplo**: varia entre -1 e 1, representa a correlação existente entre a variável Y e as variáveis X, no caso de apenas uma variável X ele é equivalente ao coeficiente de correlação. Não possui utilidade prática para a análise.

-**R-Quadrado** ou $\mathbf{R^2}$: varia de 0 a 1. Quanto mais próximo de 1 melhor é a qualidade de previsão do modelo. O nosso modelo para previsão de salário por hora apresenta um $\mathbf{R^2}$ de 0,32 valor considerado baixo, no entanto, comum para grandes amostras de corte transversal como essa.

-**R-Quadrado ajustado**: é uma medida usada para comparar modelos diferentes, mas que explicam a mesma variável dependente.

-Erro padrão: é o erro padrão da regressão, quanto menor melhor. Indica quanto em média o valor previsto pelas variáveis explicativas se distancia do valor verdadeiro.

-Observações: é o volume de informações que possuímos, nesse caso 1289 observações ou trabalhadores.

A segunda tabela é conhecida como análise de variância ou ANOVA, ela contém informações sobre a variância do modelo e um teste de significância geral do modelo.

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	5	25967,28	5193,46	122,61	0,0000000000
Resíduo	1283	54342,54	42,36		
Total	1288	80309,82			

A tabela ANOVA apresenta as estatísticas da análise de variância. A primeira coluna apresenta as fontes de variação: **Regressão** é a soma dos quadrados explicados (SQE), **Resíduo** é a soma dos quadrados dos resíduos (SQR) e **Total** é a soma total dos quadrados (SQT). Sendo que $\mathrm{SQT} = \mathrm{SQE} + \mathrm{SQR}$

-gl: são os graus de liberdade, é o número de observações independentes de cada soma dos quadrados. Para SQE $(\mathrm{k}-1)$, para SQR $(\mathrm{n}-\mathrm{k})$ e para SQT $(\mathrm{n}-1)$, onde k é o número de coeficientes da regressão, nesse exemplo são 6.

-SQ: é o resultado da soma dos quadrados.

-MQ: são os quadrados médios, é o resultado da divisão entre SQ e gl, ou seja, (SQ/gl).

-F: é a estatística do teste F. Testamos se o modelo que acabamos de estimar é melhor do que a média aritmética para fins de previsão do salário por hora.

- **F de significação:** é a probabilidade exata de encontrarmos em uma amostra qualquer todos os coeficientes angulares iguais a zero. Se o valor do **F de significação** for menor que 0,05 o modelo é bom ou significativo para fazer previsões, caso seja maior que 0,05 indica que utilizar a média para prever o salário por hora é melhor, em outras palavras, nesse caso, as variáveis que utilizamos não teriam nenhuma relação com o salário por hora. Como observamos, o valor é menor que 0,05, ou

seja, o modelo é melhor que a média para prever o salário por hora. O **teste F** é um complemento ao $\mathbf{R^2}$.

Já a terceira tabela apresenta os resultados das relações entre salário por hora e as demais variáveis.

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*
Interseção	-7,18	1,02	-7,07	0,00	-9,18	-5,19
Feminino	-3,07	0,36	-8,43	0,00	-3,79	-2,36
Não branco	-1,57	0,51	-3,07	0,00	-2,56	-0,57
Sindicato	1,10	0,51	2,17	0,03	0,10	2,09
Escolaridade	1,37	0,07	20,79	0,00	1,24	1,50
Experiência	0,17	0,02	10,38	0,00	0,14	0,20

A primeira coluna da tabela de resultados apresenta o intercepto e o nome das variáveis que utilizamos para explicar, nesse caso, o salário por hora. **Feminino**, **Não branco** e **Sindicato** são variáveis binárias, categóricas ou qualitativas, pois indicam uma qualidade ou uma condição e não um valor quantitativo.

-Coeficientes: os coeficientes da regressão nos dizem qual é a relação quantitativa entre salário por hora e cada uma das variáveis que utilizamos. Eles são calculados pelas fórmulas de MQO. Podemos montar a seguinte expressão a partir do resultado:

$$\text{Salário por hora} = -7{,}18 - 3{,}07\text{Feminino} - 1{,}57\text{Nãobranco} + 1{,}10\text{Sindicato} + 1{,}37\text{Escolaridade} + 0{,}17\text{Experiência} + e_i$$

O coeficiente de **Feminino** nos indica que as mulheres ganham em média 3,07 dólares por hora a menos que os homens. No caso de **Não branco**, os trabalhadores que declaram-se não brancos recebem em média 1,57 dólares por hora a menos que os brancos. Já os trabalhadores sindicalizados ganham em média 1,10 dólares a mais que os não sindicalizados.

A **Escolaridade** apresenta um coeficiente de 1,37, ou seja, em média cada ano de escolaridade promove um retorno de 1,37 dólares a mais no salário médio. Podemos esperar, portanto, que 10 anos a mais de estudo causariam um aumento médio no salário de 13,70 dólares por hora. A **Experiência** com coeficiente de 0,17 indica que cada ano de experiência gera um retorno extra de 0,17 dólares por hora no salário por hora.

O valor da **Interseção** (-7,18) não está associado a nenhuma variável, ele indica a média de salário por hora caso todas as variáveis explicativas sejam zero. Nesse caso, a interseção sugere um salário negativo para mulheres, não brancas, não sindicalizadas, sem escolaridade e sem experiência. Um alerta, a constante nem sempre possui uma interpretação econômica, como é o caso nessa análise.

-Erro padrão: é o erro padrão de cada coeficiente. Indica quanto o valor do estimador varia em média de amostra para amostra.

-Stat t: é a estatística do teste t, também é conhecido como teste de significância individual. Se **t** calculado for menor que 1,96 (o sinal não importa, a análise é feita em módulo) isso mostra que a variável em questão não possui relevância para explicar o salário por hora, ou, em outras palavras, que a variável é não significativa. Por outro lado, se o **t** calculado for maior que 1,96 evidencia que a variável em questão é importante para prever o salário por hora, ou, em outras palavras, que a variável é estatisticamente significativa. Quando uma variável é significativa indica que quando ela se altera

muito provavelmente isso causará impacto/alterações na variável de interesse. Por exemplo, se a **Escolaridade** aumenta é muito provável que o salário por hora também aumente.

-Valor-P: é a probabilidade exata de encontrarmos um valor zero para o coeficiente em questão. Estatística **t** maior que 1,96 gera, necessariamente, um **Valor-p** menor que 0,05, ou seja, a variável é significativa.

-95% inferiores e **95% superiores:** são os limites inferior e superior, respectivamente, do intervalo de confiança com 95% de nível de confiança. Podemos usá-los de duas formas: para avaliar o efeito das variáveis, por exemplo, o aumento de um ano de **Escolaridade** acarreta um aumento no salário por hora entre 1,24 e 1,50 com 95% de confiança. E para fazer testes de significância individual, como o valor zero não está contido no intervalo podemos rejeitar a hipótese nula de que a escolaridade não afeta o salário por hora.

Desafio: Estime uma regressão para determinantes da função pulmonar

Agora estamos prontos para analisar relações entre variáveis com a econometria. Faremos um modelo para explicar os determinantes da função pulmonar

1. O arquivo **Função pulmonar.xls** apresenta dados sobre a função pulmonar de 654 crianças e jovens com idades entre 3 e 19 anos, moradores da região de East Boston, no final dos anos 1970.
2. Quais variáveis você usaria em um modelo para explicar a capacidade pulmonar?
3. A priori, antes de estimar, qual a relação (positiva ou negativa) você espera encontrar entre as variáveis dependente e explicativas?
4. Estime o modelo de regressão, quais variáveis são significativas? Quais são os valores-P?
5. Variáveis com valor-P maior que 5% indicam que um regressor relevante não tem relevância prática?
6. Idade e altura são correlacionadas? Isso causaria um problema de multicolinearidade?
7. Você rejeitaria a hipótese de que todos os coeficientes angulares são insignificantes?
8. Qual o valor do R^2? Como você interpretaria esse valor?
9. Você concluiria, com base nesse exemplo, que fumar afeta negativamente a capacidade pulmonar?
10. Responda as questões do Enade a seguir:

Para analisar as diferenças salariais entre homens e mulheres, ou entre outros grupos populacionais, de maneira simples, estima-se uma regressão de remuneração em função de um conjunto de variáveis. Nestes termos, um pesquisador estimou as equações apresentadas na tabela, separadamente para homens e mulheres, usando o salário por hora trabalhada (em R$) em função da experiência (em anos de trabalho) e do nível de educação (em anos completos de estudo), para uma amostra de trabalhadores da Pesquisa Nacional por Amostra de Domicílio (PNAD) de 2016.

Resultados das regressões, com as indicações de erro-padrão

	Homens	Mulheres
Constante	5,012*	4,821*
	(0,4346)	(0,4783)
Experiência	0,1383*	0,1121*
	(0,009462)	(0,009801)
Educação	0,6675*	0,5703*
	(0,02492)	(0,02605)
N	2934	2066
R^2 Ajustado	0,2117	0,2048

* significância ao nível de 5%

IBGE. **Pesquisa Nacional por Amostra de Domicílio (PNAD)**, 2016 (adaptado).

Tomando como válidas as hipóteses clássicas de regressão e considerando os resultados fornecidos, assinale a opção correta.

(A) As mulheres da amostra apresentam menor nível de escolaridade média.

(B) O retorno marginal da educação para as mulheres é maior do que para os homens.

(C) Cada ano de experiência adicional faz a remuneração das mulheres aumentar em R$ 0,1121.

(D) Um aumento de 1% na educação dos homens gera um aumento de 0,6675% em sua remuneração.

(E) Aumentos na educação geram maior impacto sobre a remuneração dos homens com mais experiência.

Sabe-se que o aumento de anos de experiência em certas atividades profissionais acarreta acréscimos salariais. Porém, acredita-se que esses acréscimos sejam decrescentes ao longo doas anos. Para estudar esse problema, foi obtida, a partir de uma amostra aleatória de 526 indivíduos, os dados de salário por hora (w), medidos em Reais (R$), e a experiência ($x$), medida em anos de exercício da profissão.

O modelo econométrico foi especificado como:

$$w = \beta_0 + \beta_1 x + \beta_2 x^2 + u; \qquad u \sim N(0, \sigma^2)$$

Os resultados encontrados para a estimação foram:

$$\widehat{w} = 3{,}73 + 0{,}298x - 00061x^2$$

$$(0{,}35)(0{,}041)\ (0{,}0009)$$

Nessa expressão, a probabilidade exata do teste t para cada parâmetro estimado encontra-se, respectivamente, entre parênteses (p-valor). Considere as seguintes hipóteses:

H0: A experiência não tem efeito sobre o salário ao longo dos anos;

H1: A experiência tem efeito sobre o salário ao longo dos anos.

Considerando o comportamento do salário em relação à experiência, tendo em conta os resultados encontrados, avalie as afirmações a seguir.

I. Não é possível rejeitar H0 ao nível de significância de 5%.
II. Em face dos resultados, ao nível de significância de 1%, rejeita-se H0.
III. Ao serem representados graficamente os resultados acima, em que o salário por hora é uma função da experiência, observa-se que, inicialmente, a experiência pode exercer uma influência crescente sobre o salário, porém, após alguns anos, passa a ser decrescente.

É correto o que se afirma em

(A) I, apenas.

(B) III, apenas.

(C) I e II, apenas.

(D) II e III, apenas.

(E) I, II e III.

11. Lembre-se de que você vai precisar dos resultados para responder os exercícios posteriormente.

REFERÊNCIAS

GUJARATI, D. **Econometria: princípios, teoria e aplicações práticas.** São Paulo: Saraiva, 2019. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788553131952/pageid/4 .

GUJARATI, D. N. **Econometria básica.** 5. ed. Porto Alegre: AMGH, 2011. Disponível em: https://integrada.minhabiblioteca.com.br/books/9788580550511.

MICROSOFT PROFESSIONAL PROGRAM. **Introduction to data Science.** 2018. Disponível em: https://academy.microsoft.com/en-us/professional-program/.

SARTORIS, A. **Estatística e introdução à econometria.** 2. ed. São Paulo: Saraiva, 2013. Disponível em: https://integrada.minhabiblioteca.com.br/books/9788502199835.

WOOLDRIDGE, J. M. **Introdução à econometria: uma abordagem moderna.** São Paulo: Cengage Learning, 2016. Disponível em: https://integrada.minhabiblioteca.com.br/books/9788522126996.

EXERCÍCIOS

EXERCÍCIO LABORATÓRIO 4

No laboratório 4, você estimou algumas regressões:

1) Na regressão de vendas de limonada você incluiu a variável **Chuva** no lugar da variável **Temperatura**, informe os valores do:
a) R-Quadrado:
b) F de significação:
c) Coeficiente e valor-P de Interseção:
d) Coeficiente e valor-P de Chuva:
e) Coeficiente e valor-P de Panfletos:
f) Coeficiente e valor-P de Preço:

2) Na regressão da capacidade pulmonar informe os valores do:
a) R-Quadrado:
b) F de significação:
c) Coeficiente e valor-P de Interseção:
d) Coeficiente e valor-P de Idade:
e) Coeficiente e valor-P de Altura:
f) Coeficiente e valor-P de Fumante:
g) Coeficiente e valor-P de Masculino:
h) Coeficiente de correlação entre as variáveis Idade e Altura:

3) Considere o modelo de regressão linear múltipla, com variável dependente **y** e variáveis explicativas $\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_k}$, que pode ser expresso como

$$y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + \varepsilon_t$$

no qual ε_t significa o fator de erro e t = 1, 2, …, no índice relativo às observações amostrais.

É CORRETO afirmar que o modelo clássico de Gauss de regressão linear supõe que

(A) a relação linear entre pelo menos duas variáveis explicativas seja exata.

(B) a variância dos erros varie na amostra: $E(\varepsilon_t^2) \neq E(\varepsilon_z^2)$ para $t \neq z$.

(C) o valor esperado do fator de erro seja diferente de zero: $E(\varepsilon_t) \neq 0$.

(D) os erros não sejam correlacionados $E(\varepsilon_t \varepsilon_z) = 0$ para $t \neq z$.

(E) os valores das variáveis explicativas, $X_1, X_2, \dots, X_k$ variem de amostra para amostra.

4) Um órgão regulatório, a fim de reduzir os impactos da poluição atmosférica causada por organizações, decidiu aplicar multas conforme a quantidade de poluentes emitidos na atmosfera

pelas empresas, de modo que, quanto maior fosse essa quantidade, maior seria o valor da multa aplicada.

Ciente dessa decisão, determinada empresa contratou um consultor para realizar um levantamento da relação entre a quantidade de poluentes liberados por ela na atmosfera, o valor das multas aplicadas pelo órgão e os lucros da empresa (em 10 000 u.m.). O objetivo desse levantamento era avaliar o impacto das multas (variável independente) nos lucros da empresa (variável dependente). Os resultados desse estudo estão resumidos nas tabelas a seguir.

RESUMO DOS RESULTADOS	
Estatística de regressão	
R múltiplo	0,90
R-Quadrado	0,81
R-quadrado ajustado	0,79
Erro padrão	11,84

	Coeficientes	Erro padrão	Estatística T	P-valor
Interseção	151,44	9,23	16,41	1,47 E-08
Multas devido aos poluentes	- 0,37	0,06	- 6,47	7,15 E-05

Considerando a situação hipotética apresentada, avalie as afirmações a seguir.

IV. Rejeitando-se a hipótese nula (Ho: $\beta = 0$), evidencia-se uma relação linear negativa entre as multas e o lucro da empresa.

V. A não rejeição da hipótese nula (Ho: $\beta = 0$) sugere que multas podem incentivar a redução da poluição, uma vez que diminuem o lucro da firma.

VI. Ao nível de significância de 0,05, rejeita-se a hipótese nula (Ho: $\beta = 0$), podendo-se fazer estimativas do lucro da empresa por meio da reta $Y = 151,44 - 0,37\ X$.

É correto o que se afirma em

(A) II, apenas.

(B) III, apenas.

(C) I e II, apenas.

(D) I e III, apenas.

(E) I, II e III.

5) A decisão do Comitê de Política Monetária (Copom) do Banco Central do Brasil, divulgada em 29/07/2015, de elevar a taxa de juros básica da economia em 0,5% para o nível de 14,25% ao ano, corresponde a uma contração monetária e implica menor nível de atividade econômica nos próximos períodos. A relação entre o nível de atividade econômica, medido pela taxa de desemprego, e a taxa de inflação é estudada por meio da versão moderna da Curva de Phillips. Os dados mensais no período no período de novembro de 2012 a abril de 2015 (30 observações) para

essas duas variáveis foram obtidos na página do Instituto de Pesquisa Econômica Aplicada (IPEA). A componente cíclica do desemprego foi obtida utilizando-se o filtro Hodrick-Prescott.

A Curva de Phillips aumentada pelas expectativas pode ser especificada como:

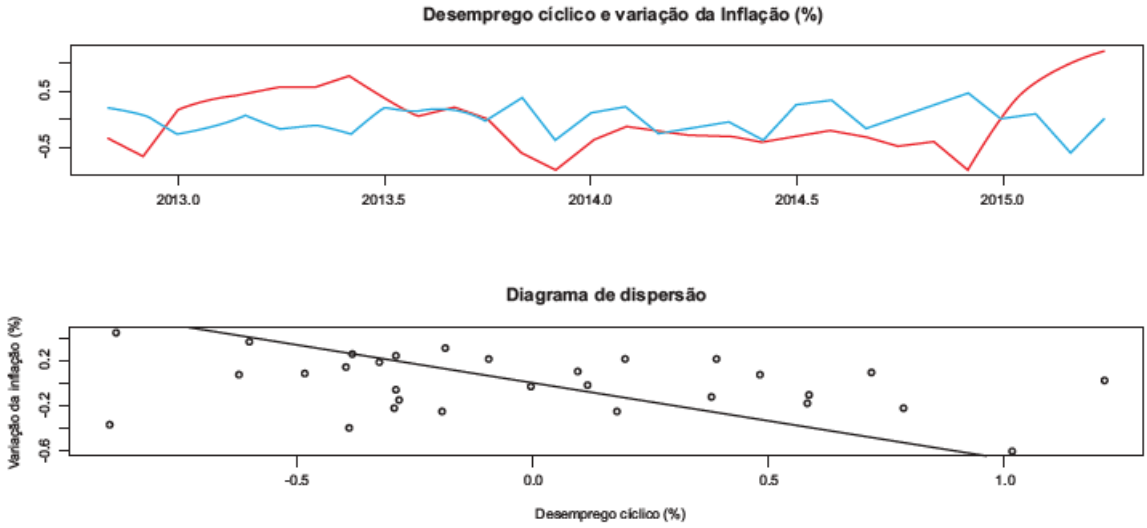
$$\pi_t = \pi^e - \varepsilon(u_t - u_n)$$

em que se presume que as expectativas de inflação são a própria inflação passada $\pi^e = \pi_{t-1}$, o que resulta em

$$\Delta\pi = -\varepsilon(u_t - u_n)$$

em que $\Delta\pi = \pi_t - \pi_{t-1}$ é a variação da inflação, u_t é o nível de desemprego, u_n é a taxa natural de desemprego e ε é a resposta da inflação aos desvios do desemprego em relação à taxa natural de desemprego.

Utilizando-se um modelo de regressão simples entre a variação da inflação e o ciclo de desemprego, o coeficiente ε foi estimado em -0,7, com uma probabilidade exata do teste (p-valor ou nível empírico de significância) associado de 0,0865. Nas figuras abaixo, estão dispostos a evolução da variação da inflação (em azul) e do ciclo de desemprego (em vermelho) e também o diagrama de dispersão para essas duas variáveis, com reta de regressão ajustada (em preto) para os dados da economia brasileira.



Fonte: IPEA.

Considerando a teoria da Curva de Phillips e os resultados obtidos para a economia brasileira, avalie as afirmações a seguir.

V. Se a decisão de contração monetária do Banco Central do Brasil for crível, a Curva de Phillips será deslocada paralelamente para baixo.
VI. Para um nível de significância de 0,05 é possível inferir que há evidência de relação entre inflação e desemprego no Brasil.
VII. Se a taxa natural de desemprego for de 4,5% ao mês, espera-se que a inflação seja 1,75% menor, se o nível de desemprego atingir 7% ao mês.
VIII. Se o Banco Central desejar manter a taxa de inflação constante (sem variação), para uma taxa natural de desemprego de 4,5%, o nível de desemprego deve ser de 6,5%.

É correto o que se afirma em

(A) I.

(B) II.

(C) I e III.

(D) II e IV.

(E) III e IV.

6) Um pesquisador resolveu estimar uma versão da Lei de Okun para determinado país X. O resultado é apresentado na equação a seguir.

$$u_t = u_n - 0{,}5gy_t + e_t$$

em que u_t é a taxa de desemprego observada para o ano t ; u_n é a taxa de desemprego natural; gy_t é a taxa de crescimento do produto no ano t ; e_t é o termo de resíduo. O país apresenta uma taxa de desemprego natural igual a 10%.

Com o objetivo de analisar a predição do modelo, esse pesquisador utilizou os dados a seguir, para alguns anos selecionados.

Dados anuais selecionados do país X

Ano	Taxa de Crescimento do Produto	Taxa de Desemprego Observada
2013	4%	6%
2014	8%	5%
2015	4%	7%
2016	2%	10%
2017	10%	5%

Considerando as informações apresentadas, assinale a opção correta.

(A) Para o ano de 2013, o modelo previu uma taxa de desemprego inferior à observada.

(B) Para o ano de 2014, a taxa de desemprego estimada foi igual à observada.

(C) Para o ano de 2015, o modelo superestimou a taxa de desemprego.

(D) Para o ano de 2016, o erro de previsão do modelo foi igual a zero.

(E) Para o ano de 2017, o erro de previsão do modelo foi positivo.

7) Para analisar as diferenças salariais entre homens e mulheres, ou entre outros grupos populacionais, de maneira simples, estima-se uma regressão de remuneração em função de um

conjunto de variáveis. Nestes termos, um pesquisador estimou as equações apresentadas na tabela, separadamente para homens e mulheres, usando o salário por hora trabalhada (em R$) em função da experiência (em anos de trabalho) e do nível de educação (em anos completos de estudo), para uma amostra de trabalhadores da Pesquisa Nacional por Amostra de Domicílio (PNAD) de 2016.

Resultados das regressões, com as indicações de erro-padrão

	Homens	Mulheres
Constante	5,012*	4,821*
	(0,4346)	(0,4783)
Experiência	0,1383*	0,1121*
	(0,009462)	(0,009801)
Educação	0,6675*	0,5703*
	(0,02492)	(0,02605)
N	2934	2066
R^2 Ajustado	0,2117	0,2048

* significância ao nível de 5%

IBGE. **Pesquisa Nacional por Amostra de Domicílio (PNAD)**, 2016 (adaptado).

Tomando como válidas as hipóteses clássicas de regressão e considerando os resultados fornecidos, assinale a opção correta.

(A) As mulheres da amostra apresentam menor nível de escolaridade média.

(B) O retorno marginal da educação para as mulheres é maior do que para os homens.

(C) Cada ano de experiência adicional faz a remuneração das mulheres aumentar em R$ 0,1121.

(D) Um aumento de 1% na educação dos homens gera um aumento de 0,6675% em sua remuneração.

(E) Aumentos na educação geram maior impacto sobre a remuneração dos homens com mais experiência.

8) Sabe-se que o aumento de anos de experiência em certas atividades profissionais acarreta acréscimos salariais. Porém, acredita-se que esses acréscimos sejam decrescentes ao longo doas anos. Para estudar esse problema, foi obtida, a partir de uma amostra aleatória de 526 indivíduos, os dados de salário por hora (w), medidos em Reais (R$), e a experiência ($x$), medida em anos de exercício da profissão.

O modelo econométrico foi especificado como:

$$w = \beta_0 + \beta_1 x + \beta_2 x^2 + u; \qquad u \sim N(0, \sigma^2)$$

Os resultados encontrados para a estimação foram:

$$\widehat{w} = 3{,}73 + 0{,}298x - 00061x^2$$

$$(0{,}35)(0{,}041)\ (0{,}0009)$$

Nessa expressão, a probabilidade exata do teste t para cada parâmetro estimado encontra-se, respectivamente, entre parênteses (p-valor). Considere as seguintes hipóteses:

H0: A experiência não tem efeito sobre o salário ao longo dos anos;

H1: A experiência tem efeito sobre o salário ao longo dos anos.

Considerando o comportamento do salário em relação à experiência, tendo em conta os resultados encontrados, avalie as afirmações a seguir.

I. Não é possível rejeitar H0 ao nível de significância de 5%.
II. Em face dos resultados, ao nível de significância de 1%, rejeita-se H0.
III. Ao serem representados graficamente os resultados acima, em que o salário por hora é uma função da experiência, observa-se que, inicialmente, a experiência pode exercer uma influência crescente sobre o salário, porém, após alguns anos, passa a ser decrescente.

É correto o que se afirma em

(A) I, apenas.

(B) III, apenas.

(C) I e II, apenas.

(D) II e III, apenas.

(E) I, II e III.