Lista de Exercícios 7

1) Em um estudo foi utilizada, erroneamente, uma amostra de apenas 3 observações para se estimarem os coeficientes de uma equação de regressão. **Obteve-se R^2 = 0,96**. A título de "brincadeira", foi dito ao analista responsável que, se ele quisesse melhorar os resultados, bastaria eliminar uma observação e ficar com apenas **n = 2**. Faça uma crítica sobre o uso de amostras muito pequenas em modelos de regressão.

*O uso de amostras muito pequenas pode afetar os resultados de um modelo de regressão por que quanto menor o **N** menor são os resíduos (Distância entre cada observação e o dado esperado pelo modelo) e maior será o R^2 (que representa a porcentagem de variância explicada pelo modelo) Assim amostras pequenas em modelos de regressão tendem a superestimar o valor do R^2, fazendo com que o modelo pareça mais relevante do que de fato é (Aumenta a chance de erro tipo I) Falta de validade externa.*

2) Observe o banco de dados "Dummies" a seguir. Neles estão descritos os dados referentes a empresa Dummies S.A. A tabela a seguir apresenta os dados correspondentes às **vendas de produtos** de determinada categoria, **ao preço** e às **horas de treinamento** dos vendedores.

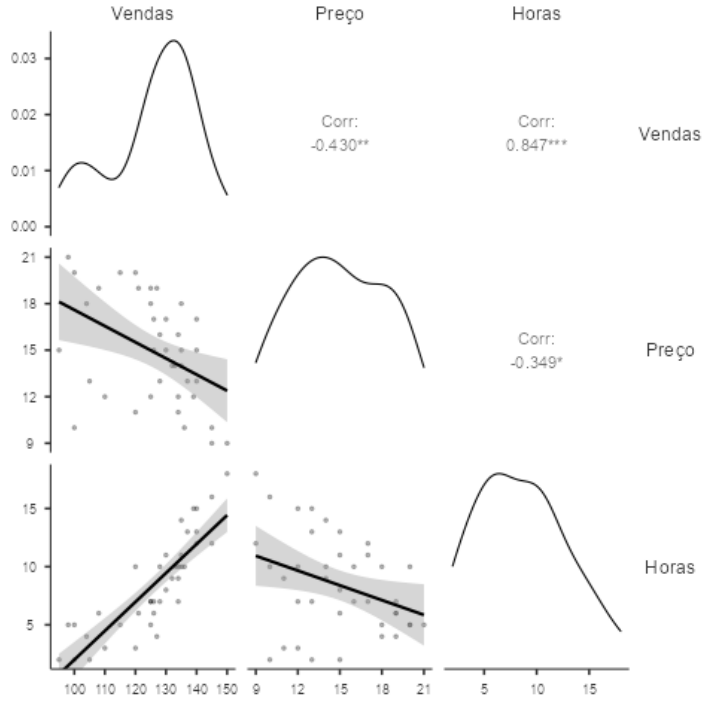
a) Obtenha a matriz de correlação de todas as variáveis deste estudo. Obtenha a reta de regressão múltipla de vendas sobre preço e horas de treinamento. Analise as tabelas do modelo.

Matriz de Correlações

		Vendas	Preço	Horas
Vendas	R de Pearson	—		
	p-valor	—		
	Limite superior do IC a 95%	—		
	Limite inferior do IC a 95%	—		
	Rho de Spearman	—		
	p-valor	—		
	N	—		
Preço	R de Pearson	−0.430 **	—	
	p-valor	0.006	—	
	Limite superior do IC a 95%	−0.133	—	
	Limite inferior do IC a 95%	−0.657	—	
	Rho de Spearman	−0.473 **	—	
	p-valor	0.002	—	
	N	39	—	
Horas	R de Pearson	0.847 ***	−0.349 *	—
	p-valor	< .001	0.029	—
	Limite superior do IC a 95%	0.918	−0.038	—
	Limite inferior do IC a 95%	0.726	−0.599	—
	Rho de Spearman	0.902 ***	−0.316	—
	p-valor	< .001	0.050	—
	N	39	39	—

Nota. * p < .05, ** p < .01, *** p < .001

Gráfico



A correlação entre vendas e preço é R= -0.430, que representa uma correlação média e negativa, ou seja o aumento do preço do produto diminui as vendas.

Já a correlação entre vendas e horas de treinamento R=0.847, que representa uma relação preditiva e positiva, ou seja quanto maior o número de horas de treinamento maiores são as vendas do produto.

Por fim a correlação entre Preço e Horas R=-0.349, que representa uma correlação fraca e negativa, assim quanto maior o preço do produto menor o número de horas de treinamento (essa relação não tem sentido prático).

b) Estime as vendas ao preço de **$25 e 6 horas de treinamento.**

Medidas de Ajustamento do Modelo

Modelo	R	R^2	R^2 Ajustado	AIC	BIC	RMSE
1	0.859	0.739	0.724	272	279	7.16

Coeficientes do Modelo - Vendas

Preditor	Estimativas	Erro-padrão	t	p
Intercepto	112.469	7.053	15.95	< .001
Preço	−0.631	0.375	−1.68	0.102
Horas	2.706	0.310	8.73	< .001

Teste de autocorrelação de Durbin-Watson

Autocorrelação	Estatística DW	p
0.433	1.10	0.002

[3]

Estatísticas de Colinearidade

	VIF	Tolerância
Preço	1.14	0.878
Horas	1.14	0.878

[3]

Teste à Normalidade (Shapiro-Wilk)

Estatística	p
0.970	0.365

Vendas = B + A * Preço +A1*Horas de Treinamento

Vendas = 112.469 - 0.631 * Preço + 2.706 * Horas de Treinamento

Vendas = 112.469 - 0.631 * 25 + 2.706 * 6

Vendas = 112.93

Segundo a Regressão Linear realizada o Valor de Teste T não foi significativo para preço (P=0.102) mas foi significativo para as horas de treinamento (P=0.001). Durbin-Watson foi aceitável (DW= 1.10, P=0.002), e a tolerância foi de 0.878.

c) Faça um novo modelo de regressão tendo como variável dependente as vendas e como variáveis independentes – preço do produto, horas de treinamento, roupa do vendedor (terno ou havaiana) e sexo do vendedor (M ou F). Verifique a tabela dos coeficientes e interprete TODOS os coeficientes (independentemente de sua significância). Lembre-se de considerar a interpretação quando a variável independente é contínua ou categórica. Utilize método ENTER

Regressão Linear

Medidas de Ajustamento do Modelo

Modelo	R	R^2	R^2 Ajustado
1	0.890	0.793	0.768

Coeficientes do Modelo - Vendas

Preditor	Estimativas	Erro-padrão	t	p
Intercepto [a]	116.984	6.718	17.412	< .001
Preço	−0.615	0.377	−1.630	0.112
Horas	2.294	0.334	6.865	< .001
Sexo_Vendedor:				
F – M	2.476	2.626	0.943	0.353
Roupa_Vendedor:				
Terno – Havaianas	−6.709	2.857	−2.348	0.025

[a] Representa o nível de referência

Segundo o modelo de regressão temos que o R2 foi de 0.793, (79% da variância dos dados foi explicado pelo modelo) o Durbin-Watson foi de 1.14 (Baixa evidência de homocedasticidade - efeito de variável externa).

Sobre as variáveis preditoras temos que:

-O preço não foi considerado um preditor (p = 0.112) mas podemos interpretar que a cada R$1,00 a mais no preço do produto, diminui 0.615 nas vendas do produto.

-Horas é um preditor (p=0.001) e podemos dizer que a cada 1 Hora a mais de treinamento do funcionário, aumenta 2.294 nas vendas do produto.

-Sexo do vendedor não é uma variavel preditora mas podemos dizer que as mulheres vendem em média 2.476 produtos a mais que os homens.

-Roupa do vendedor é um preditor e podemos dizer que quando o vendedor usa terno ele vende 6.709 produtos a menos que quando o vendedor usa havaianas.

(OBS: o valor do coeficiente no caso da variável dummy representa a diferença entre os grupos(Diferença nas vendas entre homens e mulheres, Terno e Havaianas))

Em relação a tolerância é baixa em todas as variáveis indicando problemas de multicolinearidade

d) Observe o mesmo modelo da questão c e verifique os atributos relacionados a multicolinearidade (Tolerância) e Homocedasticidade (Durbin-Watson). Eles são aceitáveis? Caso não seja, refaça o modelo utilizando o método Stepwise e veja se os indicadores melhoraram.

Durbin-Watson está aceitável 1.14, mas o modelo não apresenta boa tolerância indicando problemas de multicolinearidade (Alta correlação entre as VI (X).

Medidas de Ajustamento do Modelo

Modelo	R	R^2
1	0.881	0.776

Coeficientes do Modelo - Vendas

Preditor	Estimativas	Erro-padrão	t	p
Intercepto [a]	109.75	3.695	29.70	< .001
Horas	2.34	0.324	7.21	< .001
Roupa_Vendedor:				
Terno – Havaianas	−8.18	2.688	−3.04	0.004

[a] Representa o nível de referência

Apesar do modelo stepwise as variáveis preditoras foram horas de treinamento e roupa do vendedor no entanto a tolerância continua baixa 0.63 o que indica que essas variáveis possuem relação significante e isso pode induzir a um viés de interpretação dos resultados.

e) Observe agora o modelo com stepwise da questão d). Como fazer para lidar com a multicolinearidade presente no modelo? Qual seria uma possível explicação para ela?

Dados: Nível de significância adotado – 5%.

Para verificar a multicolinearidade entre horas e roupas do vendedor foi feito um teste T independente:

Teste t para amostras independentes

		Estatística	gl	p	Diferença média	Erro-padrão da Diferença
Horas	t de Student	4.09	37.0	< .001	4.63	1.13

Descritivas de Grupo

	Grupo	N	Média	Mediana	Desvio-padrão	Erro-padrão
Horas	Havaianas	22	10.5	10.0	3.78	0.805
	Terno	17	5.82	5.00	3.11	0.754

O teste t mostra que existe diferença significativa entre horas de treinamento entre vendedores que usam terno ou havaianas (T=4.094, P=0.001) sendo que os vendedores com havaianas foram treinados por 10.4 horas em média e os de terno por apenas 5.8 horas. Isso indica um viés de seleção onde não sabemos qual variável de fato impacta as vendas (a roupa do vendedor ou as horas de treinamento).