Atividade 5

Regressão linear múltipla

Paulo Ricardo Seganfredo Campana

16 de setembro de 2023

Problema:

Um engenheiro está estudando sobre o sistema de abastecimento de máquinas de venda automática de refrigerantes. Ele está interessado em prever a quantidade total de tempo necessário para o funcionário abastecer e fazer a manutenção de rotina das máquinas. Ele acredita que as duas variáveis mais importantes que afetam o tempo de abastecimento (y) são o número de pacotes de refrigerantes que serão estocados (x_1) e a distância percorrida pelo funcionário até a máquina (x_2). Estes dados são do livro do *Montgomery* podem ser encontrados no pacote *MPV* do **R** como o nome *softdrink*.

Considerando estes dados, fixando o nível de significância em 5% para os testes e considerando 95% de confiança para os intervalos, responda as questões abaixo:

a) Ajuste um modelo de regressão linear que relaciona o tempo de abastecimento (y) com o número de pacotes de refrigerantes que serão estocados (x_1) e a distância percorrida pelo funcionário até a máquina (x_2). Expresse o modelo estimado e interprete os parâmetros destes modelos.

```
data <- MPV::softdrink
fit1 <- lm(y ~ x1 + x2, data)
summary(fit1)
```

Tabela 1: Variáveis do modelo de regressão linear

Variáveis	Estimativa	Erro padrão	Estatística	p-valor
(Intercepto)	2.341	1.097	2.135	0.0441
x_1	1.616	0.171	9.464	3.25×10^{-9}
x_2	0.014	0.004	3.981	0.000631

O modelo estimado de regressão linear múltipla usando as variáveis x_1 e x_2 é expresso pelo plano $\hat{y} = 2.341 + 1.616 \times x_1 + 0.014 \times x_2$, desta maneira, o tempo de abastecimento aumenta em média 97 segundos ($\beta_1 = 1.616$ minutos) para cada pacote de refrigerante adicional a ser estocado e também em média 0.86 segundos ($\beta_2 = 0.014$ minutos) para cada pé de distância entre o funcionário e a máquina.

b) Através do teste F, você acha que o modelo adotado é razoável? Justifique sua resposta apresentando e analisando os resultados.

```
anova(fit1)
```

Tabela 2: Análise de variância

Variáveis	gl	SS	MS	Estatística	p-valor
x_1	1	5382.4	5382.4	506.6	1.11×10^{-16}
x_2	1	168.4	168.4	15.8	0.000631
(Regressão)	2	5550.8	2775.4	261.8	4.69×10^{-16}
(Residuos)	22	233.7	10.6		

Sim, o modelo é razoável pois o teste F mostra que, conjuntamente, as variáveis x_1 e x_2 são significantes com p-valor muito baixo (4.69×10^{-16}).

c) O R^2 e R^2_a sugerem que o modelo proposto explica razoavelmente os dados? Justifique sua resposta.

```
summary(fit1)$r.squared
## [1] 0.9595937
summary(fit1)$adj.r.squared
## [1] 0.9559205
```

Sim, o modelo explica acima de 95% da variabilidade dos dados como mostram os valores do R^2 e R^2_a.

d) Obtenha o intervalo de confiança para os coeficientes da regressão.

```
confint(fit1)
```

Tabela 3: Intervalo de confiança para os coeficientes da regressão ($\alpha = 5\%$)

Coeficientes	IC inferior	IC superior
(Intercepto)	0.066	4.615
x_1	1.261	1.969
x_2	0.006	0.021

e) Todas as variáveis regressoras contribuem significativamente para o modelo? Justifique sua resposta apresentando e analisando os resultados.

Sim, como visto da Tabela 1, o p-valor individual para a significância de cada variável está abaixo de 5%.

f) Qual a estimativa da variância dos erros?

```
summary(fit1)$sigma ^ 2
## [1] 10.62417
```

O estimador para a variância dos erros é a variância dos resíduos, dada por $\hat{\sigma}^2 = 10.624$.

g) Estime um modelo utilizando apenas a variável x_1, número de pacotes de refrigerantes que serão estocados. Compare o erro padrão estimado e o coeficiente de determinação ajustado dos dois modelos estimados. Baseado nessas medidas, qual dos dois modelos explica melhor o tempo de abastecimento?

```
fit2 <- lm(y ~ x1, data)

summary(fit1)$sigma ^ 2
## [1] 10.62417
summary(fit2)$sigma ^ 2
## [1] 17.48408

summary(fit1)$adj.r.squared
## [1] 0.9559205
summary(fit2)$adj.r.squared
## [1] 0.9274588
```

O primeiro modelo que utiliza as duas variáveis explica melhor o tempo de abastecimento pois o possui menor erro quadrático médio e maior coeficiente de determinação ajustado, as estimativas baseadas no primeiro modelo então serão mais precisas e mais correlacionadas com a realidade.

h) No gerenciamento do tempo para o funcionário abaster a máquina, o engenheiro afirma que um funcionário consegue estocar 7 pacotes de refrigerantes e percorrendo uma distância de 275 pés em um tempo de 13 minutos. Calcule o intervalo de predição para o tempo de abastecimento e verifique a afirmação do engenheiro.

```
predict(fit1, newdata = data.frame(x1 = 7, x2 = 275), interval = "prediction")
```

Tabela 4: Intervalo de confiança de predição

Estimativa	IC inferior	IC superior
17.608	10.688	24.528

Segundo o modelo, é possivel um funcionário realizar essa tarefa em 13 minutos pois este valor está contido no intervalo de predição, por mais que seja abaixo do esperado.

i) Estime o tempo médio de abastecimento se o funcionário for estocar 7 pacotes de refrigerantes e percorrer uma distância de 275 pés até a máquina e construa um intervalo de confiança para o valor médio do tempo de abastecimento para este caso.

```
predict(fit1, newdata = data.frame(x1 = 7, x2 = 275), interval = "confidence")
```

Tabela 5: Intervalo de confiança para a média

Estimativa	IC inferior	IC superior
17.608	16.126	19.089

Em média, o tempo gasto para abascetimento dos 7 pacotes de refrigerantes nesta distância está entre 16.1 a 19.1 minutos segundo o modelo.

j) Considere agora que o engenheiro necessite fazer novas predições para os seguintes casos:

1. Estocar 8 pacotes de refrigerantes em percorrendo uma distância de 300 pés;
2. Estocar 8 pacotes de refrigerantes em percorrendo uma distância de 1400 pés;
3. Estocar 25 pacotes de refrigerantes em percorrendo uma distância de 200 pés;
4. Estocar 25 pacotes de refrigerantes em percorrendo uma distância de 1300 pés;

Verifique se todos os pontos pertencem a região conjunta que contém os dados utilizados para estimar o modelo. Em quais casos o engenheiro não poderá utilizar o modelo ajustado para estimar o tempo de abastecimento? Justifique a resposta e apresente os resultados.

```r
X <- model.matrix(fit1)
H <- X %*% solve(t(X) %*% X) %*% t(X)
hmax <- max(diag(H))
hmax
## [1] 0.4982922
```

Tabela 6: Valores de h para os 4 casos acima

x_1	x_2	h	$h < h_{\max}$
8	300	0.048	Sim
8	1400	1.319	Não
25	300	0.948	Não
25	1400	0.429	Sim

Como $h_{\max} = 0.498$, Não é possivel usar o modelo para prever o tempo de abastecimento nos casos 2 e 3 pois apenas o primeiro e o último caso percentem a região conjunta dos dados, isso se dá pois há uma alta correlação entre as variáveis x_1 e x_2, os dados utilizados não apresentam situações em que a quantidade de pacotes é baixa mas a distância é grande ou vice versa.