Lista 1

Aluno: Deivison Venicio Souza
Professor: José Luiz Padilha da Silva

Questão 1:
Considere a função de produção de Cobb-Douglas:

$$Y_i = \beta_0 X_{i1}^{\beta_1} X_{i2}^{\beta_2} e^{\varepsilon_i}$$

Em que Y é a produção, X_1 é o insumo de trabalho, X_2 é o insumo de capital, e " é o termo de erro. A soma (β_1+ β_2) informa a respeito dos retornos de escala, a resposta do produto a uma variação proporcional nos insumos. Se essa soma for igual a 1, haverá retornos constantes de escala, isto é, se dobrarmos os insumos, a produção dobrará, se o triplicarmos, a produção triplicará.

a) Escreva a função de produção de Cobb-Douglas na forma do modelo linear geral.

R: $Y_i = \beta_0 + \beta_1 Ln_{X_{i1}} + \beta_2 Ln_{X_{i2}} + \epsilon_i$

b) Escreva as hipóteses nula e alternativa para testar se há retornos constantes de escala. Qual é o procedimento estatístico a ser usado?

R: As hipóteses para retornos constantes ficam:

H_0: $\beta_1 + \beta_2 = 1$ (hipótese de nulidade)
H_a: $\beta_1 + \beta_2 \neq 1$ (hipótese alternativa)

Para testar as hipóteses quanto aos retornos constantes pode-se utilizar de uma análise de variância (ANOVA).

c) Escreva o modelo reduzido referente ao teste em (b).

R: Considerando o teste de hipótese de retornos constantes o modelo ficaria:

Sabendo que = H_0: $\beta_1 + \beta_2 = 1$;
Têm-se que = $\beta_1 = 1 - \beta_2$

Então, fazendo a substituição:
R: $Y_i = \beta_0 X_{i1}^{1-\beta_2} X_{i2}^{\beta_2} e^{\epsilon_i}$ (modelo reduzido)

Questão 2: Considere uma amostra de tamanho n = 20 de uma variável aleatória Y e de um conjunto de variáveis explicativas X1, X2, X3 e X4, para a qual foram ajustados os seguintes modelos:

fit1 = lm(Y~X1+X2+X3+X4)
fit2 = lm(Y~X2+X1+X3+X4)

```
> summary(fit1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.07555    1.91636   1.605   0.1294
X1           0.83243    0.13337   6.242 1.58e-05 ***
X2           0.19218    0.09452   2.033   0.0601 .
X3           0.80688    0.31582   2.555   0.0220 *
X4          -0.23702    0.15195  -1.560   0.1396
> summary(fit2)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.07555    1.91636   1.605   0.1294
X4          -0.23702    0.15195  -1.560   0.1396
X2           0.19218    0.09452   2.033   0.0601 .
X1           0.83243    0.13337   6.242 1.58e-05 ***
X3           0.80688    0.31582   2.555   0.0220 *
```

a) Escreva o modelo e interprete os coeficientes das variáveis significativas para o primeiro ajuste (Assuma α = 5%).

R: Substituindo os parâmetros estimados para o modelo 1 (fit_1) têm-se:

$$fit_1 = 3.07555 + 0.83243X_1 + 0.19218X_2 + 0.80688X_3 - 0.23702X_4 + \epsilon_i$$

No primeiro ajuste (fit_1) apenas as variáveis X1 e X3 foram consideradas significativas a através do teste *t*-Student (α > 5%) para previsão da variável resposta (Y). Os parâmetros associados à estas variáveis possuíram valores de 0,83243 ($\hat{\beta}_1$) e 0,80688 ($\hat{\beta}_3$), respectivamente. Ambos os parâmetros indicam a alteração na resposta média de Y a cada mudança unitária nos valores de X_1 e X_3, mantendo-se tudo mais constante.

b) Foi também realizado o teste *F* para verificação da significância das variáveis potencialmente preditoras, sendo obtido os resultados a seguir:

```
> anova(fit1)
Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 59.687  59.687 66.8585 6.573e-07 ***
X2         1  2.724   2.724  3.0519  0.101082
X3         1 11.351  11.351 12.7147  0.002817 **
X4         1  2.172   2.172  2.4331  0.139645
Residuals 15 13.391   0.893
```

```
> anova(fit2)
Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
X4         1  5.358   5.358  6.0020   0.02705 *
X2         1  0.363   0.363  0.4062   0.53349
X1         1 64.386  64.386 72.1225 4.095e-07 ***
X3         1  5.827   5.827  6.5273   0.02199 *
Residuals 15 13.391   0.893
```

As conclusões obtidas pela ANOVA são bem distintas entre os ajustes fit_1 e fit_2. Por exemplo, para o primeiro ajuste vemos que X4 apresenta p = 0,1396 enquanto para o segundo ajuste temos p = 0,027. Compare com o p = 0,1396 do *teste t* para ambos os modelos. Explique a que se deve essa diferença.

R: Inicialmente, deve-se compreender que a análise de regressão linear tem como importante instrumento de teste de hipóteses o medida p-valor. Na RL, o p-valor pode ser usado para testar a hipótese H_0 em dois momentos: a) teste t-Student para significância dos parâmetros estimados; e b) teste F-Snedecor para significância da regressão.

No que diz respeito ao teste t-Student este é usado para avaliar indícios da existência ou não de associação linear entre a variável resposta Y e seus potencias preditores. A ideia geral é avaliar a significância de uma variável preditora qualquer condicional à todas as demais preditoras. Para o caso específico da preditora "x_4" poderíamos fazer a seguinte indagação: "Qual a importância que X_4 têm para explicar a resposta Y dada a inclusão das variáveis X_1, X_2 e X_3 no modelo?" Para as demais variáveis a pergunta seria a mesma, por exemplo: "Qual a importância que X_1 têm para explicar a resposta Y dado a inclusão de X_2, X_3 e X_4 no modelo?" Dessa forma, o teste t-Student sempre levará em consideração todas as variáveis preditoras inseridas no modelo. Assim, o score do p-valor não dependerá da posição em que a variável entra no modelo, mas fica condicionado a existência das mesmas preditoras, como pode-se percebe: fit1 e fit2 possuíram o mesmo p-valor para os parâmetros estimados, pois foram ajustados com as mesmas preditoras, embora com entradas diferenciadas no modelo. Usando o p-valor, rejeita-se H_0 se o p-valor for menor do que o nível de significância α estabelecido (p-valor < 0,05). Para o beta associado à variável X_4 não se rejeitou a hipótese H_0, isto é, β_4 foi considerado estatisticamente igual a zero ($\beta_4 = 0$).

A ANOVA da RL subsidia a aplicação da estatística F-Snedecor, a qual é utilizada para testar a hipótese de nulidade (H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$) contra uma hipótese alternativa ($\beta_j \neq 0$, para qualquer j = 1, ..., p). Diferentemente do teste t-Student para significância dos coeficientes da regressão, a posição em que cada variável entra no modelo afetada diretamente o p-valor. Assim, para a variável X_4 na ANOVA do modelo 1 (fit1), por exemplo, far-se-ia a seguinte indagação: "Dado que X_1, X_2 e X_3 já estão no modelo (fit1) qual a importância de X_4 para explicar Y?". Se replicarmos a pergunta para X_3 na ANOVA de fit1, teríamos: "Dado que X_1 e X_2 já estão no modelo (fit1) qual a importância de X_3 para explicar Y?". Portanto, o teste F-Snedecor tem um caráter condicional sequencial, ou seja, a ordem com que uma variável preditora quaisquer é inserida no modelo é levada em consideração na avaliação de sua importância para explicar a resposta Y.

Questão 3. Para os dados da questão anterior o software relata:

Residual standard error: 0.9448 on 15 degrees of freedom
Multiple R-squared: 0.8501, Adjusted R-squared: 0.8101
F-statistic: 21.26 on 4 and 15 DF, p-value: 4.86e-06

a) Interprete a estatística R2. Explique por que ele e maior que o R2 ajustado e se tal fato e esperado.

R: O R^2 (coeficiente de determinação) refere-se à capacidade das variáveis preditoras em explicar a variável resposta Y. Para o caso específico, poder-se-ia interpretar que o modelo ajustado foi capaz de explicar, aproximadamente, 94,48% das variações que ocorrem na reposta média de Y. Um fato importante é que a estatística R^2 sempre aumenta com a adição de termos ao modelo (Betas), portanto não se deve usar o R^2 enquanto medida comparativa modelos com diferentes números de variáveis preditoras. Neste caso, a alternativa é usar o R^2 ajustado (coeficiente de determinação ajustado) que é uma medida ajustada para o número de preditores do modelo em relação ao número de observações. Portanto, é de se esperar que o R^2 ajustado tenha um valor inferior do que R^2.

b) Escreva as hipóteses relacionadas com a estatística F. Com base no valor de p, qual é a conclusão?

R: A estatística F é utilizada como indicador da existência de significância da regressão, isto é atestar se há uma relação linear entre a variável resposta Y e algumas das variáveis regressoras (preditoras). Portanto, é um indicador da adequabilidade do modelo ajustado. As hipóteses testadas pelo teste F-Snedecor para o modelo específico são:

H_0: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
H_1: $\beta_j \neq 0$

A partir das hipóteses têm-se que:

- Rejeita-se H_0: Quando o valor de $F_{\text{calculado}} > F_{1;\, n-2;\, 1-\alpha}$ (em que α é o nível de significância considerado (no caso, $\alpha = 5\%$). De outro modo, a hipótese nula será rejeitada se o p-valor for menor do que o nível de significância estabelecido (α). Nos casos contrários, ter-se-á não rejeição da hipótese nula (H_0).

Para o caso específico, o valor da estatística F-Snedecor foi de 21,26 e o p-valor associado foi de $4{,}86e^{-06}$. Assim, haja vista que o p-valor foi menor do que o nível de significância estabelecido ($\alpha = 5\%$) têm-se evidências fortes pela rejeição da hipótese de nulidade (H_0), isto é, existe pelo menos uma variável regressora que está contribuindo significativamente para explicar as variações observáveis no modelo.

Questão 4. A Zarthan Company vende um creme para a pele exclusivamente através de lojas de moda. Estão disponíveis dados de venda para quinze distritos. Vendas (em lotes) e tratada como a variável dependente Y, e população alvo (em milhares de pessoas) e renda per capita (em dólares) são as variáveis independentes X1 e X2, respectivamente. Espera-se que o modelo $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, i = 1, \dots, n$ com erros normais seja adequado. Os resultados do ajuste foram:

```
> fit=lm(Y~X1+X2,data=dados)
> summary(fit)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.374572  11.786833   1.644    0.126
X1           0.479601   0.026852  17.861  5.2e-10 ***
X2           0.004429   0.004526   0.979    0.347
---
Residual standard error: 10.29 on 12 degrees of freedom
Multiple R-squared:  0.9734,    Adjusted R-squared:  0.9689
F-statistic: 219.3 on 2 and 12 DF,  p-value: 3.567e-10
```

a) O R^2 obtido é bastante alto. Neste caso, por que devemos realizar análise de resíduos e de observações influentes?

R: A estatística R^2 é uma medida de qualidade do ajuste de um modelo de regressão linear. No entanto, esta medida não deve ser utilizada **isoladamente** como critério para tomada de decisão da adequabilidade do modelo ajustado. No caso específico, se levarmos em consideração somente o valor do R^2 = 0,9734 seriamos impulsionados a concluir que o modelo foi é excelente (pois, foi capaz de explicar boa parte das variações ocorridas em Y) e, portanto, que foi bem especificado. No entanto, é bem provável que isso não seja totalmente verdade. Por exemplo, ao analisar o resultado do teste *t- Student* para os coeficientes da regressão (β_j), pode-se notar que o valor estimado para o parâmetro β_2 foi não significativo (isto é, não diferente de zero) indicando, portanto, que a variável associada a este parâmetro (no caso, a renda per capita) não é importante para explicar as variações da resposta Y (Vendas). Assim, pode-se perceber um erro de especificação do modelo. O diagnóstico das pressuposições da análise de regressão linear (normalidade, homocedasticidade e independência dos resíduos) é imprescindível para tomada de decisão final sobre a adequabilidade do modelo. Então, apesar do modelo possuir elevado R^2, deve-se observar se os resíduos são normalmente distribuídos, com variância constante e próximo da média zero e, ainda, se são independentes entre si. Caso as pressuposições não sejam atendidas pode-se ter erros de especificação do modelo, ou mesmo a presença de valor(es) influente(s) (outliers) podem estar alterando significativamente o ajuste do modelo e, por conseguinte, as predições.

b) Assuma, por ora, que o modelo é adequado, e suponha que a companhia tenha interesse em estimar vendas em um distrito com população alvo X_{h1} = 220 mil pessoas e renda per capita X_{h2} = 2500 dólares. A estimativa pontual encontrada foi $\hat{Y}_h$ = 135,96. Um intervalo de confiança de 95% calculado para a resposta média foi (129,21 - 142,71), e o intervalo obtida para uma nova observação foi (112,55 - 159,37). Explique a diferença entre eles.

R: A estimativa de um intervalo de confiança para a resposta média de Y (reta de regressão) e também para uma nova observação (predição) é de grande interesse. Assim, a construção dos ICs nos fornecem uma ideia da precisão do(s) nosso(s) beta(s) estimados. Particularmente, o IC para reposta média nos informar o quanto, provavelmente, os parâmetros estimados variariam caso fizéssemos um novo ajuste, considerando dados de uma mesma população. No caso dos ICs específico, poderíamos interpretá-los da seguinte forma:

- IC para reposta média: Indica que se realizados outros 100 ajustes dos parâmetros (β_s') com novos dados oriundos de uma mesma população, em 95% das vezes teríamos valores de betas estimados contidos no intervalo de 129,21 - 142,71.

- IC para uma nova predição: Indica que a predição de Y para uma nova observação deverá estar no intervalo de 112,55 - 159,37, com 95% de certeza.