**Lista 1**

**Aluno:** Deivison Venicio Souza

**Professor:** José Luiz Padilha da Silva

**Questão 1:**

Considere a função de produção de Cobb-Douglas:

C:\Users\Deivison\OneDrive\Imagens\Capturas de tela\2018-09-06.png

Em que Y é a produção, X1 é o insumo de trabalho, X2 é o insumo de capital, e " é o termo de erro. A soma (+ ) 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.

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

**R:**

1. 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:

(hipótese de nulidade)

(hipótese alternativa)

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

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

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

Sabendo que =;

Têm-se que =

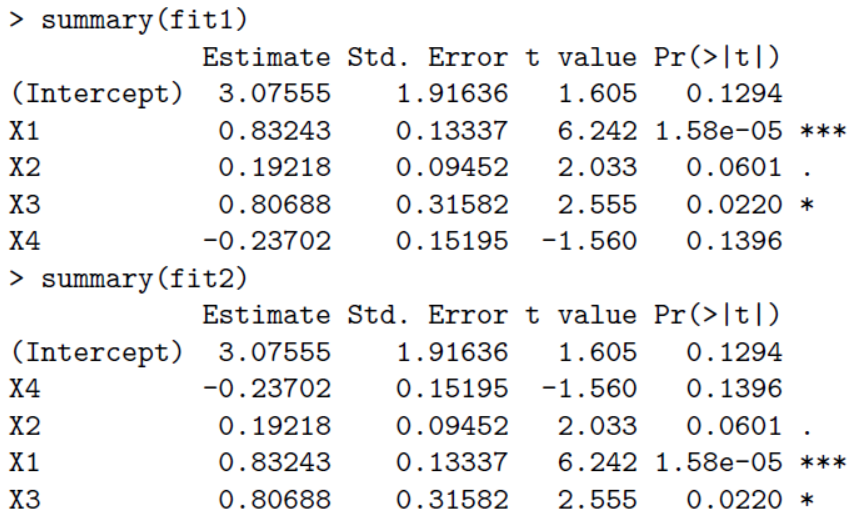
Então, fazendo a substituição:

**R:**  (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)

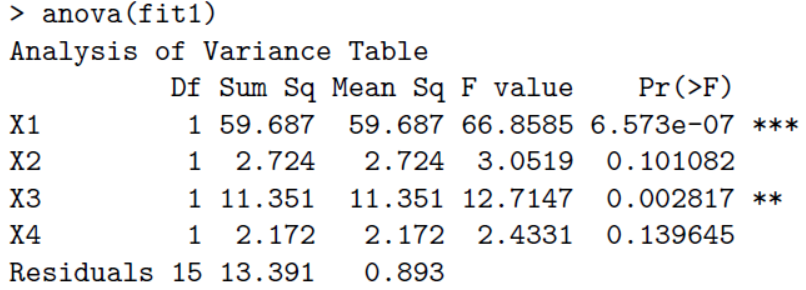


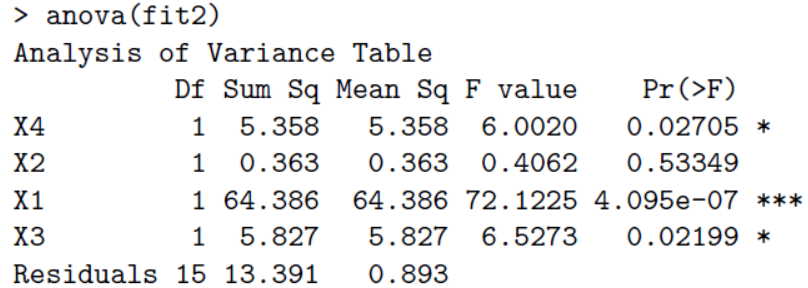
1. Escreva o modelo e interprete os coeficientes das variáveis significativas para o primeiro ajuste (Assuma α = 5%).

**R:** Substituindo os parâmetros estimados para o modelo 1 (fit1) têm-se:

No primeiro ajuste (fit1) 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 () e 0,80688 (), respectivamente. Ambos os parâmetros indicam a alteração na resposta média de Y a cada mudança unitária nos valores de X1 e X3, mantendo-se tudo mais constante.

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





As conclusões obtidas pela ANOVA são bem distintas entre os ajustes *fit1* e *fit2*. 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 *H0* 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 “x4” poderíamos fazer a seguinte indagação: “Qual a importância que X4 têm para explicar a resposta Y dada a inclusão das variáveis X1, X2 e X3 no modelo?” Para as demais variáveis a pergunta seria a mesma, por exemplo: “Qual a importância que X1 têm para explicar a resposta Y dado a inclusão de X2, X3 e X4 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 *H0* se o *p*-valor for menor do que o nível de significância α estabelecido (*p*-valor < 0,05). Para o beta associado à variável X4 não se rejeitou a hipótese *H0*, isto é, foi considerado estatisticamente igual a zero (.

A ANOVA da RL subsidia a aplicação da estatística *F*-Snedecor, a qual é utilizada para testar a hipótese de nulidade ( contra uma hipótese alternativa (, 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 X4 na ANOVA do modelo 1 (fit1), por exemplo, far-se-ia a seguinte indagação: “Dado que X1, X2 e X3 já estão no modelo (fit1) qual a importância de X4 para explicar Y?”. Se replicarmos a pergunta para X3 na ANOVA de fit1, teríamos: “Dado que X1 e X2 já estão no modelo (fit1) qual a importância de X3 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

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

**R:** O R² (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 R² sempre aumenta com a adição de termos ao modelo (Betas), portanto não se deve usar o R² enquanto medida comparativa modelos com diferentes números de variáveis preditoras. Neste caso, a alternativa é usar o R² 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 R² ajustado tenha um valor inferior do que R².

1. 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:

**H0:**

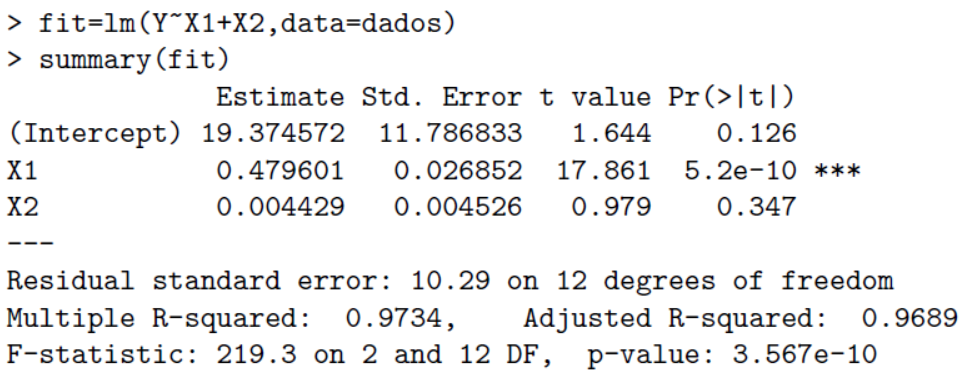
**H1:**

A partir das hipóteses têm-se que:

**- Rejeita-se H0:** Quando o valor de *F*calculado > F1; n-2; 1-α (em que é o nível de significância considerado (no caso, ). De outro modo, a hipótese nula será rejeitada se o *p*-valor for menor do que o nível de significância estabelecido (). Nos casos contrários, ter-se-á não rejeição da hipótese nula (H0).

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 () têm-se evidências fortes pela rejeição da hipótese de nulidade (H0), 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 com erros normais seja adequado. Os resultados do ajuste foram:



1. O R² obtido é bastante alto. Neste caso, por que devemos realizar análise de resíduos e de observações influentes?

**R:** A estatística R² 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 R² = 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 (, pode-se notar que o valor estimado para o parâmetro 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 R², 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.

1. Assuma, por ora, que o modelo é adequado, e suponha que a companhia tenha interesse em estimar vendas em um distrito com população alvo Xh1 = 220 mil pessoas e renda per capita Xh2 = 2500 dólares. A estimativa pontual encontrada foi = 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 () 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.