Universidade de São Paulo

Escola de Artes, Ciências e Humanidades

Trabalho Final

Trabalho Final da disciplina ACH 2053 – Introdução à Estatística, do curso de Bacharelado em Sistemas de Informação.

Professor:
Fernando Fagundes Ferreira
Monitor:
Eder Lucio da Fonseca
Integrantes:
Melina Brilhadori – nº USP 6412313
Murilo Galvão Honório – nº USP 6411927
Thiago de Oliveira Shirata – nº USP 6412212

São Paulo

Junho de 2011

1 Resolução das Questões

Questão 01

a) *Comente um pouco sobre o coeficiente de correlação (ρ). Como ele pode ser interpretado? Existe algum caso em que ele não pode trazer informação sobre a relação entre as variáveis? Comente.*

O coeficiente de correlação populacional (ρ) serve para estudar o comportamento conjunto de duas variáveis quantitativas distintas. Ele quantifica a intensidade da associação linear entre duas variáveis da escala métrica (intervalo ou razão), bem como a direção linear dessa correlação. Pode assumir valores entre -1 e 1, sendo que o valor zero denota uma falta de linearidade. Um valor ρ = 1 indica correlação perfeita positiva entre duas variáveis. Em contrapartida, um valor ρ = -1 denota correlação perfeita negativa (duas variáveis inversamente proporcionais). O coeficiente de correlação tem como estimador o coeficiente de correlação amostral (*r*). Em amostras grandes, correlações pequenas podem ser significantes, mesmo que o diagrama de dispersão mostre pouca evidência de linearidade, ou seja, pode haver correlação significativa sem importância prática.

b) *Calcule o coeficiente de correlação amostral (r) entre as variáveis X1 e Y1 pertinentes ao seu grupo.*

O coeficiente de correlação amostral (*r*) é dado pela seguinte fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

No Excel, pode ser facilmente obtido com o uso da função *CORREL(Matriz1;Matriz2)*, em que a Matriz1 é o intervalo de valores de X e a Matriz2 o intervalo de valores de Y. Assim, o valor obtido foi r = -0,665.

c) *Teste se a correlação é igual a zero apresentando todos os passos necessários para a tomada de decisão (teste bilateral com α=0,04). O que aconteceria se utilizássemos α=0,01?*

- Passo um: formular as hipóteses

  Como não temos uma hipótese a priori, escolhemos um teste de significância bilateral com $\propto= 0{,}04$. As hipóteses são:

  $H_0: \rho = 0$

  $H_1: \rho \neq 0$

- Passo dois: obter o valor crítico

  Procuramos o valor crítico $t_\propto$ com $v = n - 2$ graus de liberdade para o nível de 4%. Para um teste bilateral usando $v = 30 - 2 = 28$ graus de liberdade, a tabela t fornece $t_{0,02} = 2{,}154$.

- Passo três: calcular a estatística de teste

  A estatística de teste para a hipótese $H_0: \rho = 0$ é

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -0{,}665\sqrt{\frac{30-2}{1-(-0{,}665)^2}} = -4{,}712$$

- Passo quatro: tomar a decisão

  Rejeitamos $\rho = 0$, pois $t = -4{,}712 < t_\propto = -2{,}154$

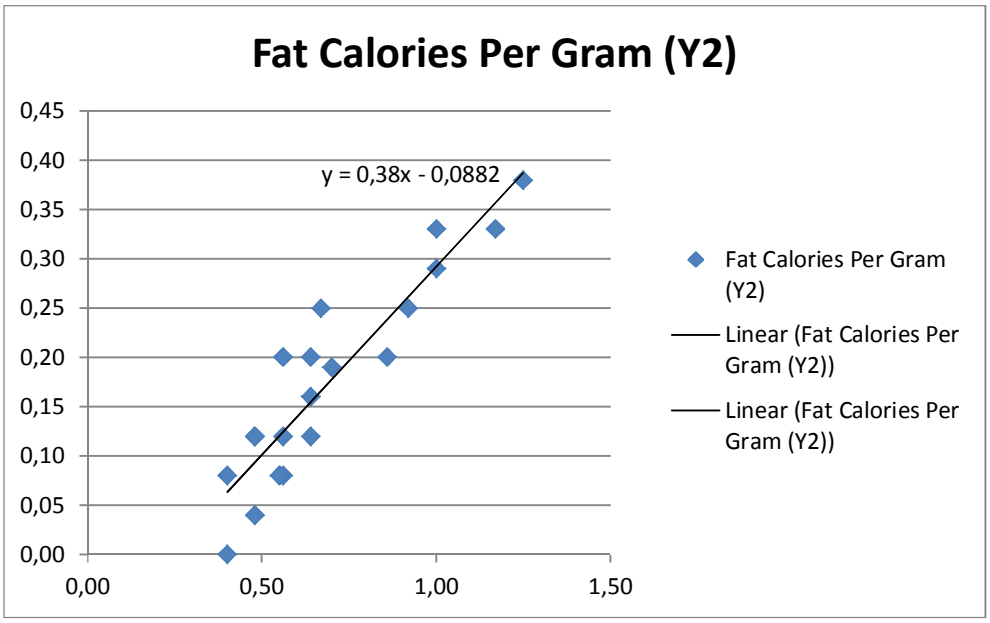
  Um nível descritivo pode ser calculado usando a função do Excel *DIST(t; graus_liberdade; caudas)*. Assim, o valor-p bilateral é DIST(-4,712;28;2) ≅ 0,0000.

  Para α=0,01, através do nível descritivo, uma vez que $p < 0{,}01$ também rejeitamos a hipótese nula.

Questão 02

Para o conjunto de dados pertinente ao seu grupo:

a) *Construa um gráfico de dispersão para X2 e Y2.*



Construímos o gráfico selecionando as colunas com rótulos X2 e Y2 adequadamente no Excel.

b) *Justifique a necessidade de utilizarmos uma regressão linear.*

É necessária a utilização de uma regressão, pois desejamos modelar uma relação entre duas variáveis quantitativas (X2, Y2) e desse modo fazer predições. Utilizamos a regressão linear porque é a que melhor se encaixa na dispersão dos dados analisados, além de ser simples de calcular. Estamos conjecturando se a quantidade de calorias de gordura de determinados molhos é uma função da quantidade total de calorias.

c) *Efetue um teste t unilateral para a nulidade do coeficiente de correlação, com $\alpha = 0,05$. Formule as hipóteses claramente.*

- Passo um: formular as hipóteses

  Através de inspeção visual podemos identificar uma relação direta (coeficiente positivo) entre as varáveis, portanto escolhemos um teste de significância

unilateral à direita com $\propto = 0{,}05$. As hipóteses são:

$H_0: \rho \leq 0$

$H_1: \rho > 0$

- Passo dois: obter o valor crítico

  Procuramos o valor crítico $t_\propto$ com $v = n - 2$ graus de liberdade para o nível de 5%. Para um teste bilateral usando $v = 20 - 2 = 18$ graus de liberdade, a tabela t fornece $t_{0,05} = 1{,}734$.

- Passo três: calcular a estatística de teste

  Obtemos diretamente do Excel o coeficiente de correlação r ≅ 0,918.

  A estatística de teste para a hipótese $H_0: \rho = 0$ é

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}918\sqrt{\frac{20-2}{1-(0{,}918)^2}} = 9{,}821$$

- Passo quatro: tomar a decisão

  Rejeitamos $\rho = 0$, pois $t = 9{,}821 > t_\propto = 1{,}734$. Um nível descritivo pode ser calculado usando a função do Excel DIST(9,821;20;2) ≅ 0,999999. Para α = 0,05, através do p-valor também rejeitamos a hipótese nula.

d) *Encontre a equação da reta ajustada ao conjunto de dados.*

Utilizamos o Excel para adicionar a reta e estimar os coeficientes da regressão de maneira a produzir um bom ajuste. Assim:

$$Y = 0{,}388X - 0{,}0882$$

e) *Interprete o coeficiente angular e o intercepto. O intercepto faz sentido?*

O coeficiente angular (β_1) indica que para cada grama adicional de caloria total em um molho, em média aproximadamente 0,38 das calorias são de gordura. Essa estimativa é estatística, pois uma amostra diferente poderia fornecer uma estimativa diferente do coeficiente angular. Não há sentido em interpretar o intercepto (β_0), pois ele sugere que mesmo um molho sem calorias terá uma quantidade mínima de calorias de gordura, o

que não tem sentido lógico. Além disso, X=0 está fora do intervalo dos dados observados.

f) *Construa um intervalo de confiança para o coeficiente angular com α = 0,01. Qual é a sua conclusão a respeito do coeficiente angular? Interprete o intervalo de confiança.*

Tomando α = 0,01, construiremos um intervalo para a confiança de 99%. Os graus de liberdade são $n = 20 - 2 = 18$.

Para a construção do intervalo de confiança demandamos o estimador erro padrão (s_{xy}). Através de operações realizadas no Excel (suplemento regressão da análise de dados), obtivemos os dados referentes aos ajustes, demonstrados na tabela abaixo:

Estatística de regressão	
R múltiplo	0,918034523
R-Quadrado	0,842787385
R-quadrado ajustado	0,834053351
Erro padrão	0,042295447
Observações	20

De posse do erro padrão (s_{yx}), calculamos o erro padrão do coeficiente angular S_{b_1}, conforme a fórmulas abaixo:

$$S_{b_1} = \frac{s_{yx}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{0{,}042295}{1{,}195520} = 0{,}035$$

Para os dados do problema, $s_{b_1} = 0{,}035$, obtidos através de cálculos do Excel. Desta forma, com $n - 2 = 20 - 2 = 18$ graus de liberdade, $t_\alpha = 2{,}567$ e, portanto o intervalo de confiança para o coeficiente angular é:

$$b_1 - t_{n-2} S_{b_1} \leq \beta_1 \leq b_1 + t_{n-2} S_{b_1}$$

$$0{,}353 \leq \beta_1 \leq 0{,}423$$

g) *Escolha dois valores de X2: um dentro do conjunto de dados e outro fora do conjunto de dados. Em seguida gere duas previsões com os valores escolhidos. Interprete os resultados.*

Assumimos os seguintes valores de X2:

$$X2_1 = 0{,}64\ (conjunto\ de\ dados); X2_2 = 0{,}99$$

Os respectivos valores de Y2 são:

$$Y2_1 = (0{,}38)(0{,}64) - 0{,}0882 = 0{,}155$$
$$Y2_2 = (0{,}38)(0{,}99) - 0{,}0882 = 0{,}288$$

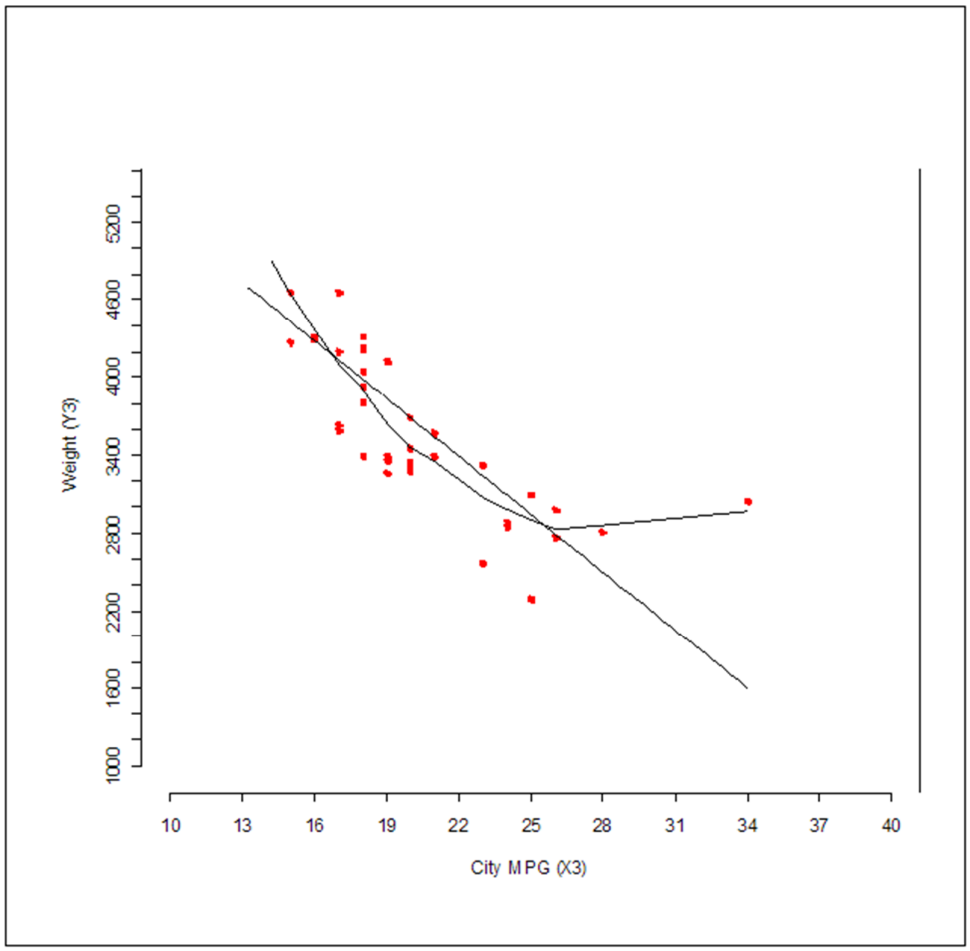
Essa previsão segue a interpretação do coeficiente angular o X2 é maior que o Y2, sendo que Y2 é em média 38% de X2.

Questão 03

Para o conjunto de dados pertinente ao seu grupo e utilizando a ferramenta Action:

a) *Construa um gráfico de dispersão para X3 e Y3.*

Estamos interessados na relação linear de apenas uma variável de entrada (X3) com a variável resposta (Y3), então temos o caso de Regressão Linear Simples. O gráfico de Scatter (scatterplot) permite a visualização de uma possível associação entre as variáveis quantitativas X3 e Y3.



b) *Escreva a equação da reta ajustada.*

Onde .

Através dos valores obtidos a partir da estimação, concluímos que a cada aumento da quilometragem (X3), há uma diminuição de 149,45 no peso (Y3),

Coeficientes				
Preditor	Estimativa	Desvio Padrão	Estat. T	P-valor
Intercepto	6680,37094998	317,82250589	21,01918784	0,00000000
C1	-149,45024054	15,97634300	-9,35447120	0,00000000

C1: City MPG (X3) C2: Weight (Y3)

c) *Interprete a estatística F obtida e o p-valor obtido.*

Como o valor do intervalo de confiança não foi estipulado pelo exercício, consideramos 95% de confiança, por ser este um valor comumente utilizado além de ser o valor padrão para o cálculo da análise de variância utilizado pelo Action. Obtemos um valor nível de significância 0,05. Com os graus de liberdade 1 para Variância Entre e 40 (valor mais próximo de 41) para Variância Dentro, o valor F tabelado é 4,08.

Tabela da Anova					
Fatores	G.L.	Soma Quad	Quadrado Médio	Estat. F	P-valor
C1	1	17923358,812488	17923358,812488	87,506132	0,000000
Residuals	41	8397785,373558	204824,033501		

Considerando as seguintes hipóteses para o teste F:

- H_0: β1 = β2 =... = βi = 0, ou seja, as i variáveis independentes não exercem influência na variável dependente, segundo o modelo proposto. Neste caso, como tratamos de modelo linear simples (1° grau), H_0: β1 = X3 = 0, X3 não exerce influência sobre Y3, para esta hipótese.
- H_A: βi ≠ 0, para pelo menos um i, o que significa dizer que pelo menos uma das i variáveis independentes exerce influência na variável dependente, segundo o modelo proposto. Neste caso, i=1, β1 = X3 ≠ 0, X3, que é a variável independente, exerce influência sobre Y3, variável dependente, para esta hipótese.

A regra de decisão para o teste F é:

- Se $F_{calculado} \geq F_{tabelado}$ rejeita-se H_0 ao nível de significância que foi realizado o teste. Pode-se inferir que o modelo proposto é adequado para descrever o fenômeno, X3 é importante para explicar a variabilidade em Y3;
- Se $F_{calculado} < F_{tabelado}$ não se rejeita H_0 ao nível de significância que foi realizado o teste. Neste caso, pode-se inferir que o modelo proposto não é adequado para descrever o fenômeno, não há relação linear entre X3 e Y3.

Baseado na regra de decisão do teste F, $F_{calculado}$(aprox.87,5) > $F_{tabelado}$ (aprox.4,08), portando rejeita-se H_0 ao nível de significância de 5%. O modelo proposto, através das

estimativas de 6680,37 para o intercepto e -149,45 para o coeficiente angular descreve satisfatoriamente o problema de peso e quilometragem fornecido pela amostra, X3 é importante para explicar a variabilidade em Y3.

O P-Valor é outra maneira de avaliar a significância da estatística F. A regra de decisão neste caso é (com α = nível de significância, neste caso α = 0,05):

- Se o P-Valor ≤ α, rejeita-se a hipótese H_0;
- Se o P-Valor > α, não se rejeita hipótese H_0, ou seja, não há evidências de diferenças significativas entre as variáveis de regressão, ao nível α de significância escolhido.

Baseado na regra de decisão do P-valor, P-valor (aprox.0) < α (0,05), portanto rejeita-se H_0 ao nível de significância de 5%, confirmando o anteriormente obtido, ou seja, que o modelo ajustado descreve satisfatoriamente o problema, X3 exerce influencia sobre Y3 na forma da reta ajustada.

d) *Qual a relação entre a estatística F, a Soma dos Quadrados da Regressão e Soma dos Quadrados dos Resíduos? Dica: Pense em como foram calculados os Quadrados Médios dos Resíduos e Quadrados Médios da Regressão (utilizando os graus de liberdade).*

A Análise de Variância (ANOVA) baseia-se na decomposição da variação total da variável resposta em partes que podem ser atribuídas as variáveis regressoras (Variância Entre) e ao resíduo (Variância Dentro). Essa variação pode ser medida por meio das somas de seus quadrados:

$$SQTotal = SQRegressão + SQResíduo$$

Tal que SQTotal é a variação total em Y e é calculada por:

$$SQTotal = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

SQRegressão é a variação em Y explicada pela regressão ajustada, calculada por:

$$SQRegressão = \beta_0 \sum_{i=1}^{n} Y_i + \beta_1 \sum_{i=1}^{n} Y_i X_i - \frac{\sum_{i=1}^{n} (Y_i)^2}{n}$$

Onde:

Y_i = valor observado para a variável dependente Y no i-ésimo nível da variável independente X.

β_0 = estimador da constante de regressão. Representa o intercepto da reta com o eixo dos Y.

β_1 = estimador do coeficiente de regressão. Representa a variação de Y em função da variação de uma unidade da variável X.

X_i = i-ésimo nível da variável independente X (i = 1,2,...,n)

n = número de observações

E SQResíduo é a variação não explicada pela regressão e que pode ser calculada pela diferença:

$$SQResíduo = SQTotal - SQRegressão$$

Tabela da Anova					
Fatores	G.L.	Soma Quad	Quadrado Médio	Estat. F	P-valor
	glRegressão	SQRegressão	QMRegressão	F_calculado	
C1	**1**	**17923358,81**	**17923358,81**	**87,50613151**	1,00E-11
	glResíduo	SQResíduo	QMResíduo		
Residuals	**41**	**8397785,374**	**204824,0335**		
		SQTotal			
		26321144,19			

Os quadrados médios são obtidos pela divisão da soma dos quadrados pelos seus respectivos graus de liberdade. Assim,

$$QMRegressão = \frac{SQRegressão}{glRegressão} \text{ e } QMResíduo = \frac{SQResíduo}{glResíduo}$$

Para testar a hipótese H_0, utiliza-se o teste F, que é a relação entre o QMRegressão e o QMResíduo:

$$F_{calculado} = \frac{QMRegressão}{QMResíduo}$$

e) *O ajuste obtido foi bom? Justifique utilizando a estatística R2.*

O valor de R^2 indica a proporção (ou porcentagem) da variação de Y que é "explicada" pela regressão, ou quanto da variação na variável dependente Y está sendo "explicada" pela variável independente X. Ele fornece uma informação auxiliar ao resultado da análise de variância da regressão, para se verificar se o ajuste é adequado ou não para descrever o problema abordado.

$$R^2 = {SQRegressão}/{SQTotal}$$

$0 \leq R^2 \leq 1$. Valores próximos de 1 indicam que o modelo proposto é adequado para descrever o problema.

Tabela da Anova					
Fatores	G.L.	Soma Quad	Quadrado Médio	Estat. F	P-valor
		SQRegressão			
C1	1	**17923358,81**	17923358,81	87,50613151	1,00198E-11
Residuals	41	8397785,374	204824,0335		
		SQTotal			
		26321144,19			

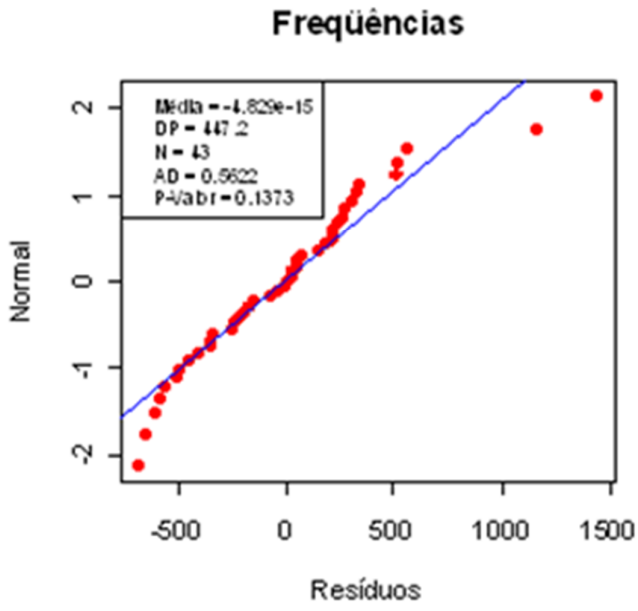
Desvio Padrão dos Resíduos	*Graus de Liberdade*	*R^2*	*R^2 Ajustado*
452,5748927	41	0,680949076	0,673167346

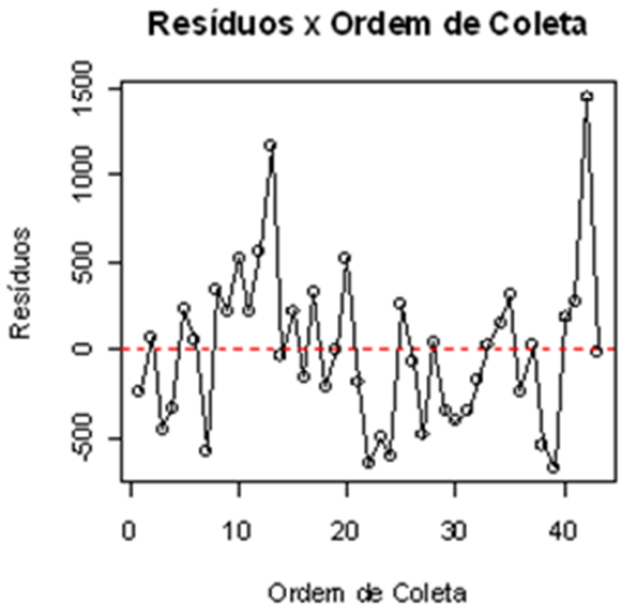
SQTotal: 26321144
SQRegressão: 17923359
R2: 0,680949

Portanto, o modelo proposto para ajuste é considerado bom, já que, baseado em R^2, aproximadamente 68% da variável dependente Y3 é explicada por X3.

O R^2 é uma medida descritiva da qualidade do ajuste obtido, entretanto o valor do coeficiente de determinação depende do número de observações n, tendendo a crescer quando n diminui. A magnitude de R^2 também depende da amplitude da variação da variável regressora X3. Geralmente, R^2 aumentará com maior amplitude de variação dos X3's e diminuirá em caso contrário.

f) *Teste as suposições adotadas ao ajustar uma regressão linear:*





- *Os erros são normalmente distribuídos?*

Para verificar se um conjunto de dados tem distribuição Normal, podem ser usados os seguintes testes: Testes Kolmogorov-Smirnov, Anderson-Darling e Shapiro-Wilk.

Tomamos como hipótese nula a normalidade dos resíduos e utilizamos a estatística de Anderson-Darling para testar a hipótese nula contra a alternativa de que os erros não seguem a normal.

Estatística: Anderson-Darling	0,5622
P-valor	0,1373

Valores podem ser verificados no gráfico de Frequências em AD e P-Valor

Como o P-valor é 13,73%, aceitamos a hipótese de normalidade. Assim, com nível de confiança de 95%, temos evidências estatísticas de que os dados seguem uma distribuição normal.

Através do gráfico Papel de Probabilidade, gráfico 2 (Frequências), é possível observar que os dados seguem uma distribuição normal, pois se apresentam distribuídos sobre a reta. Este método compara os quantis observados (resíduos) com os quantis esperados de uma distribuição normal. No caso de resíduos normalmente distribuídos esperamos que os pares (quantis observados, quantis esperados) estejam alinhados em uma linha reta.

Também se observa pelo Histograma de Resíduos, gráfico 1, que exibe sua dispersão e distribuição, ou seja, a forma da função densidade dos resíduos.

- *Os erros têm variância constante?*

O gráfico 3, Resíduos x Valores Ajustados, mostra a relação entre os resíduos e os valores esperados (ajustados) de Y3. Para haver homocedasticidade, variância constante, os pontos deveriam ficar distribuídos de forma equilibrada acima e abaixo da linha que passa pelo ponto de resíduo 0, formando uma nuvem retangular de pontos.

Neste caso, não há homocedasticidade. Quase se observa uma nuvem em forma de cone. A dispersão dos valores é maior na parte final da distribuição.

- *Os erros são independentes (não autocorrelacionados)? Utilize o teste de Durbin-Watson.*

  Observando o gráfico Resíduos x Ordem da coleta, gráfico 4, verifica-se a hipótese de independência dos resíduos. Se nós percebemos uma tendência dos pontos, ou seja, se os pontos tiverem um comportamento que se repete em determinado ponto do gráfico, temos indícios de dependência dos resíduos. Em nosso caso, os pontos

estão distribuídos aleatoriamente, não possuindo traços de comportamento tendencioso, assim, temos indícios de independência.

Podemos analisar a independência através do teste estatístico de Durbin-Watson, o qual é aplicado sobre os resíduos de uma análise de regressão e mostra o quanto os dados testados são correlacionados. Em outras palavras, exibe o quanto um evento é influenciado por outro e é calculado por:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}(e_i)^2}$$

Neste caso, testaremos a presença de autorrelação nos resíduos, ou seja, a correlação entre cada resíduo e o resíduo imediatamente anterior.

A partir do conjunto de dados Resíduos, gerados pelo Action, calculamos a estatística de teste através do Excel com as seguintes operações:

- SOMAXMY2(intervalo1;intervalo2), que soma o quadrado das diferenças dos dois intervalos, no nosso caso, intervalo1 = ($Resíduos_2$, $Resíduos_n$) e intervalo2 = ($Resíduos_1$,$Resíduos_{n-1}$).
- SOMAQUAD(intervalo), que é a soma dos quadrados dos termos do intervalo. Neste caso, intervalo = ($Resíduos_1$,$Resíduos_n$).

O resultado da estatística de teste de Durbin-Watson é:

SOMAXMY2	12491334
SOMAQUAD	8397785
DW	1,487456

$$\mathrm{DW} = \mathrm{SOMAXMY2} \,/\, \mathrm{SOMAQUAD}$$

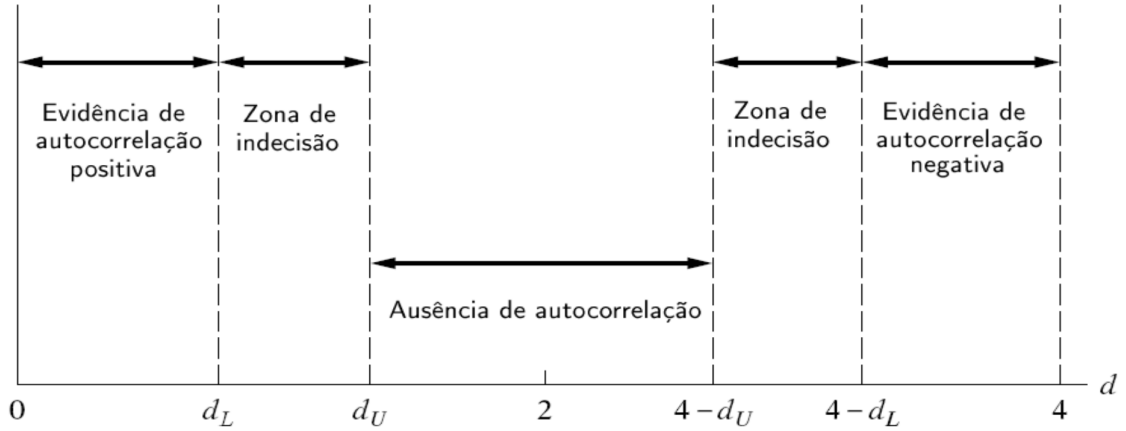
Onde DW é o cálculo de Durbin-Watson.

Para encontrar os valores críticos, DL e DU, verificamos na tabela de valores críticos de Durbin-Watson, no Anexo 1 (página 18), a partir de α/2 (0,05/2 = 0,025), quantidade de variáveis explicativas (números de X_i, i = 1) e quantidade de observações (n = 43, aprox.40). Assim,

DL	1,35
DU	1,45

Temos que o valor de DW e sua interpretação seguem:

- $0 < \mathrm{DW} < \mathrm{DL}$ Evidência de autocorrelação positiva - dependência



- DL < DW < DU Zona de indecisão – teste inconclusivo
- DU < DW < 4 - DU Ausência de autocorrelação - independência
- 4 - DL < DW < 4 Zona de indecisão – teste inconclusivo
- 4 - DL < DW < 4 Evidência de autocorrelação negativa - dependência

Assim,

$$DU \leq DW \leq 4\text{-}DU$$

$$1{,}45 \leq 1{,}49 \leq 2{,}55$$

Portanto não rejeitamos H_0 e, com nível de confiança de 95%, os resíduos são independentes.

Questão 04

Um grupo de 36 pacientes portadores de anemia foi dividido em três grupos (Homens, Mulheres e Crianças), e todos foram submetidos ao mesmo tratamento. A doença é caracterizada por uma baixa taxa de hemoglobina, ou seja, valores inferiores a 13 g/dl para homens, inferiores a 12 g/dl para mulheres e inferiores a 11g/dl para crianças. Para o conjunto de dados pertinente ao seu grupo, verifique, aos níveis de 5% e 1% de significância, se a melhoria do tratamento foi igual para os três grupos.

- Vamos comparar as medianas de c = 3 amostras independentes, supondo que as populações difiram somente em centralidade. Portanto, efetuaremos o teste de Kruskal-Wallis.

- As hipóteses a serem testadas são:

$$H_0: Todas\ as\ medianas\ das\ três\ populações\ são\ iguais$$
$$H_1: Nem\ todas\ as\ medianas\ das\ populações\ são\ iguais$$

- Para obter a estatística de teste, combinamos as amostras e atribuímos um posto para cada observação em cada grupo, conforme mostrado na tabela abaixo (a média dos postos é atribuída a cada observação repetida):

Obs	Posto	Taxa	Grupo
1	1	8,70	Mulheres
2	2	9,00	Mulheres
3	3	9,60	Mulheres
4	4	10,50	Crianças
5	5	10,70	Crianças
6	6	10,90	Crianças
7	8	11,20	Mulheres
8	8	11,20	Mulheres
9	8	11,20	Crianças
10	10,5	11,30	Mulheres
11	10,5	11,30	Crianças
12	12,5	11,40	Mulheres
13	12,5	11,40	Crianças
14	14,5	11,50	Homens
15	14,5	11,50	Crianças
16	16,5	11,60	Crianças
17	16,5	11,60	Crianças

18	18	11,80	Mulheres
19	19	11,90	Homens
20	22	12,00	Homens
21	22	12,00	Homens
22	22	12,00	Mulheres
23	22	12,00	Crianças
24	22	12,00	Crianças
25	25	12,10	Mulheres
26	26	12,20	Crianças
27	27	12,40	Mulheres
28	28	12,50	Mulheres
29	29	12,60	Homens
30	30	12,70	Homens
31	31	13,10	Homens
32	32	13,40	Homens
33	33	13,50	Homens
34	34	13,60	Homens
35	35	13,90	Homens
36	36	14,50	Homens

Tabela 1

- A seguir, ordenamos os dados por grupos, e somamos os postos para obter $T_1, T_2 \; e \; T_3$, conforme a tabela 2 abaixo:

Homens	Posto	Mulheres	Posto	Crianças	Posto
11,50	14,5	8,70	1	10,50	4
11,90	19	9,00	2	10,70	5
12,00	22	9,60	3	10,90	6
12,00	22	11,20	8	11,20	8
12,60	29	11,20	8	11,30	10,5
12,70	30	11,30	10,5	11,40	12,5
13,10	31	11,40	12,5	11,50	14,5
13,40	32	11,80	18	11,60	16,5
13,50	33	12,00	22	11,60	16,5
13,60	34	12,10	25	12,00	22
13,90	35	12,40	27	12,00	22
14,50	36	12,50	28	12,20	26
Soma dos Postos (T_1)	337,5	Soma dos Postos (T_2)	165	Soma dos Postos (T_3)	163,5
Tamanho da amostra	$n_1 = 12$	Tamanho da amostra	$n_2 = 12$	Tamanho da amostra	$n_3 = 12$

Tabela 2

- O valor da estatística de teste é dado por

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} - 3(n+1)$$

Em que

$n = n_1 + n_2 + n_3$

$n_j = número\ de\ observações\ no\ grupo\ j$

$T_j = soma\ dos\ postos\ no\ grupo\ j$

Então,

$$H = \frac{12}{36(36+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} - 3(36+1)$$

$$= \frac{12}{36(36+1)} \left[\frac{337{,}5^2}{12} + \frac{165^2}{12} + \frac{163{,}5^2}{12} \right] - 3(36+1) \cong 15{,}078$$

O mesmo resultado pode ser obtido utilizando a ferramenta *Action*, no menu teste Kruskal-Wallis, selecionando adequadamente as colunas taxa e grupo da tabela 1.

- A estatística de teste *H* segue uma distribuição de qui-quadrado com $v = c - 1 = 3 - 1 = 2$ graus de liberdade. Para $v = 2$ a tabela χ^2 fornece os seguintes valores:

  $$\chi^2_{0,05} = 5{,}991;\ \chi^2_{0,01} = 9{,}210$$

  O valor-p pode ser obtido através da ferramenta *Action* ou da função do Excel *DIST.QUI(x;graus_liberdade)*, então DIST.QUI(15,078;2) = 0,000531934.

- Decisão: como *H* é um teste unilateral à direita, rejeitaremos a hipótese nula de igualdade das medianas se *H* exceder seu valor crítico.

  Nesse caso, como $H > \chi^2_{0,05}$ e $H > \chi^2_{0,01}$ rejeita-se H_0tanto ao nível de 5% como ao nível de 1% de significância. Pelo nível descritivo, o valor-p está próximo de zero o que corrobora nossa decisão de rejeitar H_0.

2 Anexos

Tabela de valores críticos de Durbin-Watson

		Número de variáveis explicativas									
		1		2		3		4		5	
n	Nível de significância	DL	DU	DL	DU	DL	DU	DL	DU	DL	DU
	0,01	0,81	1,07	0,7	1,25	0,59	1,46	0,49	1,7	0,39	1,96
15	0,025	0,95	1,23	0,83	1,4	0,71	1,61	0,59	1,84	0,48	2,09
	0,05	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
	0,01	0,95	1,15	0,86	1,27	0,77	1,41	0,63	1,57	0,6	1,74
20	0,025	1,08	1,28	0,99	1,41	0,89	1,55	0,79	1,7	0,7	1,87
	0,05	1,2	1,41	1,1	1,54	1	1,68	0,9	1,83	0,79	1,99
	0,01	1,05	1,21	0,98	1,3	0,9	1,41	0,83	1,52	0,75	1,65
25	0,025	1,13	1,34	1,1	1,43	1,02	1,54	0,94	1,65	0,86	1,77
	0,05	1,2	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
	0,01	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
30	0,025	1,25	1,38	1,18	1,46	1,12	1,54	1,05	1,63	0,98	1,73
	0,05	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
	0,01	1,25	1,34	1,2	1,4	1,15	1,46	1,1	1,52	1,05	1,58
40	**0,025**	**1,35**	**1,45**	1,3	1,51	1,25	1,57	1,2	1,63	1,15	1,69
	0,05	1,44	1,54	1,39	1,6	1,34	1,66	1,29	1,72	1,23	1,79
	0,01	1,32	1,4	1,28	1,45	1,24	1,49	1,2	1,54	1,16	1,59
50	0,025	1,42	1,5	1,38	1,54	1,34	1,59	1,3	1,64	1,26	1,69
	0,05	1,5	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,7