Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Exercícios resolvidos em Análise de Regressão utilizando o MINITAB®

Giselle Silva de Carvalho
Ilka Afonso Reis

Relatório Técnico
RTE-01/2004
Série Ensino

Sumário

Introdução

Este relatório consiste de listas de exercícios de Análise de Regressão elaboradas pela professora Ilka Afonso Reis e resolvidas pela aluna, então no 4º período de Graduação em Estatística, Giselle Silva de Carvalho.

As listas estão divididas em teóricas (1º parte) e práticas (2º parte), sendo que as listas teóricas não estão resolvidas. Há também uma parte na qual se ensina de maneira resumida como usar o software Minitab® para se fazer análise de regressão. Os dados utilizados nos exercícios estão nas tabelas em anexo.

A intenção deste relatório é fazer com que alunos, não só da Estatística e Ciências Atuariais, mas outras pessoas interessadas nesta área, tenham um material (em português) para consultar.

1º Parte - Exercícios práticos

- **Questões:**

  ❖ *Regressão linear simples:*

  - Parte 1

1) Utilizando os dados da Tabela A.1 (página 51, Draper & Smith, 3 ed.) :

a) Faça o diagrama de dispersão.

b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.

c) Construa a Tabela de Análise de Variância e calcule o R^2.

d) Retire o par de observações nº. 16 (Y=5.9 ; X = 6.7) e refaça os itens de a) a c).

e) Comparando somente os valores de R^2, quais dos dois modelos é o melhor? O par de observações nº 16 influencia a qualidade do ajuste ?

2) Os dados deste exercício são do exercício K (Capítulo 3) do livro de Draper & Smith e estão na Tabela A.2 no Anexo. A variável resposta (Y.3K) representa a porcentagem de amendoins não-contaminados por certo fungo em um lote e a variável explicativa (X.3K) representa a quantidade média de uma substância química para evitar contaminação em cada 60 gramas de amendoins.

a) Faça o diagrama de dispersão.

b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.

c) Construa a Tabela de Análise de Variância e calcule o R^2.

d) Este conjunto de dados possui dois níveis de X com medidas repetidas (X = 18,8 e X = 46,8). Entretanto, alguns níveis de X tem valores “muito próximos” que, na prática, poderiam ser considerados “iguais” e, assim, os valores de Y nestes níveis poderiam ser considerados medidas repetidas. São eles:

X = 9,3 ; 9,9
X = 12,3 ; 12,5 e 12,6
X = 18,8 ; 18,8 ; 18,9
X = 21,7 ; 21,9
X = 46,8 ; 46,8 (estes são realmente medidas repetidas)
X = 70,6 ; 71,1 ; 71,3
X = 83,2 ; 83,6 .

e) Considere os valores de Y nestes níveis como sendo medidas repetidas e calcule a soma de quadrados do erro puro (SSErroPuro). Este valor é , claro, uma aproximação.

Encontre também os graus de liberdade desta soma.

f) Construa a nova Tabela de Análise de Variância, agora com a SSResidual desmembrada em SSErroPuro e o SSL (SS da falta-de-ajuste). Faça o teste F da falta-de-ajuste.

g) Caso o teste F da falta-de-ajuste seja não-significante, faça o teste F geral.

h) Interprete os coeficientes da reta de regressão.

- Parte 2

1) Utilizando os dados da Tabela A.3. (exercício **V**, capítulo 3, Draper & Smith, 3 ed., página 105) :

Variável Resposta: Y = tamanho da "linha da vida" da mão esquerda (em cm) ;
Variável Explicativa: X = idade da pessoa ao morrer (em anos);

a) Faça o diagrama de dispersão.

b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.

c) Construa a Tabela de Análise de Variância com a SSResidual desmembrada em SSErroPuro e o SSL (SS da falta-de-ajuste). Faça o teste F da falta-de-ajuste.

d) Faça o teste F da regressão (Escreva hipóteses nula e alternativa, faça o teste e conclua).

e) Calcule o valor de R^2 e o valor de max(R^2) e faça a interpretação de R^2.

f) Verifique a suposição de normalidade dos resíduos através do gráfico de probabilidade Normal.

g) Faça o teste da homogeneidade do erro puro (Bartllet e Levene).

h) Analise os gráficos de resíduos apropriados.

i) Reporte os possíveis problemas encontrados na análise dos resíduos (itens *f*, *g* e *h*) .

j) Faça o teste H_0: $\beta_0 = 0$ contra H_a: $\beta_0 \neq 0$.

k) A partir de suas análises nos itens anteriores, conclua sobre a relação entre Y e X.

- Parte 3 – Regressão simples e regressão inversa

1) Num estudo retrospectivo sobre a possível relação entre "o tempo de utilização de um plano de previdência" e o "tempo de contribuição do beneficiário" , ambos medidos em meses, uma amostra de 100 beneficiários de um plano de previdência tiveram essas duas variáveis registradas. Os dados estão na Tabela A.4 em anexo.

Variável resposta: Y = tempo de contribuição, em meses.
Variável explicativa: X = tempo de utilização do benefício, em meses (tempo entre a data da aposentadoria e a data do falecimento do beneficiário).

a) Faça o diagrama de dispersão.

b) Ajuste o modelo de regressão linear adequado, encontrando a reta estimada.

c) Construa a tabela de análise de variância com a SSresidual desmembrada em Sserropuro e o SSL (SS da falta de ajuste). E faça o teste da falta de ajuste.

d) Faça o teste F da regressão (escreva a hipótese nula e alternativa, faça o teste e conclua).

e) Calcule o valor de R^2 e do max(R^2) e faça a interpretação do R^2.

f) Verifique a suposição de normalidade dos resíduos através do gráfico de probabilidade Normal.

g) Faça o teste da homogeneidade do erro puro (Bartlet e Levene).

h) Analise os gráficos de resíduos apropriados.

i) Reporte os possíveis problemas encontrados na análise de resíduos.

j) Faça o teste H_o: $\beta_o = 0$ contra H_a : $\neq \beta_o$.

k) A partir das suas análises anteriores conclua sobre a relação entre Y e X.

l) Regressão inversa: como o estudo foi retrospectivo, a partir do falecimento do beneficiário foi possível estabelecer o valor da variável explicativa e, então o valor da resposta para aquele nível ad variável explicativa. Porém, na prática, gostaríamos de estudar a relação inversa, ou seja, a partir do tempo de contribuição gostaríamos de predizer o tempo de uso do benefício. Deste modo, usaremos a regressão inversa.

l.1) A partir da reta estimada em (b), estabeleça a equação da regressão inversa, Isto é, X como função de Y.

l.2) Dado o valor do tempo de contribuição igual a 348 meses, estime o valor médio do tempo de uso do benefício.

l.3) Estabeleça o intervalo a 95% de confiança para o tempo de uso do beneficiário quando o tempo de contribuição for igual a 355 meses.

- Parte 4

1) Um investigador deseja estudar a possível relação entre os salários e o tempo de experiência no cargo de gerente de agências bancárias de uma grande empresa. Além disto, gostaria de saber se há diferenças quando são levados em conta homens e mulheres separadamente. Os dados coletados estão disponíveis na Tabela A.5 em anexo, e a descrição do banco de dados segue abaixo.

Variável Resposta Y: Salário, em mil reais ;
Variáveis Explicativas X: Experiência = tempo de trabalho no cargo, em anos completos ;
Sexo = sexo do empregado (0 – feminino ; 1 – masculino)

a) Faça o diagrama de dispersão do salário versus experiência e avalie a possibilidade do ajuste de um modelo de regressão linear.

b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, sendo X a variável "experiência" , encontrando a reta estimada.

c) Construa a Tabela de Análise de Variância e calcule o valor de R^2.

d) Verifique a suposição de normalidade dos resíduos através do gráfico de probabilidade Normal.

e) Analise o gráfico *resíduos* versus *ajustados* (*preditos*). Os resíduos parecem se distribuir aleatoriamente em torno do valor zero?

f) Analisando as respostas aos itens d) e e), o modelo ajustado em b) parece ser adequado?

g) Analise o gráfico *resíduos* versus *sexo*. O que se pode concluir?

h) Para cada sexo separadamente, repita os itens de b) a e).

i) Para cada sexo separadamente, faça o teste F da regressão (escreva hipóteses nula e alternativa, faça o teste e conclua).

j) Compare os valores de R^2 dos modelos em separado com o valor calculado em c). O que se pode concluir?

k) Faça a mesma comparação usando o valor do MSResidual das tabelas ANOVA. Lembre-se de que o MSResidual é a estimativa da variância da resposta (Utilize o conceito de desvio-padrão, se achar mais fácil sua análise).

l) Interprete a reta de regressão estimada para cada sexo e tire suas conclusões sobre a relação entre "salário" e "experiência" para os gerentes de banco desta empresa.

- Parte 5 – Modelo sem intercepto e variáveis *Dummy*

1) Considere o conjunto de dados da Tabela A.6 no Anexo.

a) Ajuste o modelo de regressão $Y=\beta_0+\beta_2 X_2+\varepsilon$.

b) Construa a Tabela de Análise de Variância, calcule o valor de R^2, faça o teste de falta de ajuste (se possível)[1].

c) Caso não haja problemas com o teste da falta de ajuste , faça o teste F da regressão (escreva hipóteses nula e alternativa, faça o teste e conclua).

d) Teste a significância do intercepto do modelo (teste *t*-Student ou intervalo de confiança. Escreva hipóteses nula e alternativa, faça o teste e conclua).

e) Ajuste o modelo de regressão sem o intercepto. $Y=\beta_2 X_2+\varepsilon$.

f) Note que o MINITAB não calcula o R^2 para o modelo sem intercepto. Use então o valor do MSResidual para escolher entre os dois modelos (com intercepto e sem intercepto).

2) **Variáveis *Dummy***

Suponha que desejássemos estudar a renda (em R$) dos empregados de certo setor em função de sua experiência no cargo em que ocupa (anos) e de seu local de trabalho. Se tivéssemos 4 cidades (A, B , C e D), as variáveis *dummies* a serem criadas seriam :

	Local 1	**Local 2**	**Local 3**
Cidade A	0	0	0
Cidade B	1	0	0
Cidade C	0	1	0
Cidade D	0	0	1

a) Suponha que exista uma quinta cidade (Cidade E). Como ficaria a tabela de codificação das cidades com a introdução da Cidade E?

b) Considere agora a seguinte codificação:

	Local 1	**Local 2**	**Local 3**
Cidade A	0	0	1
Cidade B	0	1	0
Cidade C	1	0	0
Cidade D	0	0	0

[1] Por questões didáticas, estamos omitindo a etapa de análise dos resíduos, que viria antes da utilização de qualquer teste.

o modelo :
Salário = β + β_l experiência + β_{21} “local1” + β_{22} “local2” + β_{23} “local3” + erro

e seguinte equação de regressão estimada :
Salário = 2,50 + 0,099 experiência + 0,55 “local1” + 0,69 “local2” + 0,75 “local3”

Considerando a mesma experiência, qual é a diferença média entre os salários das pessoas da:

b.1) cidade A e B
b.2) cidade A e C
b.3) cidade A e D
b.4) cidade B e C
b.5) cidade B e D
b.6) cidade C e D

c) Considere a primeira codificação. Suponha que, ao fazermos o teste t-Student para os parâmetros do modelo:

A categoria de referência é a cidade A .
O parâmetro β_{21} refere à cidade B (local1).
O parâmetro β_{22} refere à cidade C (local2).
O parâmetro β_{23} refere à cidade D (local3).

c.1) a hipótese $\beta_{21} = 0$ não seja rejeitada. O que isto significa em termos da comparação entre as cidades?

c.2) a hipótese β_{22} = 0 não seja rejeitada. O que isto significa em termos da comparação entre as cidades?

c.3) a hipótese β_{23} = 0 não seja rejeitada. O que isto significa em termos da comparação entre as cidades?

d) Pense na primeira tabela de codificação (local 1, local 2 e local 3). Para representar a cidade E, uma alternativa à resposta em a) seria fazer “local 1” = 1; “local 2” = 1 e “local 3” =1 . Considerando os testes de hipóteses para os parâmetros descritos em c) , pense em por que este procedimento **não** pode ser adotado (pense na comparação entre as cidades quando apenas um parâmetro não for considerado significativo)

	β_{21}	β_{22}	β_{23}
	Local 1	**Local 2**	**Local 3**
Cidade A	0	0	0
Cidade B	1	0	0
Cidade C	0	1	0
Cidade D	0	0	1
Cidade E	1	1	1

❖ *Exercícios de Revisão de Regressão Linear Simples*

Considere o modelo de regressão linear simples, $Y = \beta_0 + \beta_1 X + \varepsilon$.

1) Qual é a variável dependente? E qual é a variável independente? Que outros nomes são usados para se referir a estas variáveis?

2) Qual é o método utilizado para estimar β_0 e β_1? Para utilizar esse método é necessário supor alguma distribuição para a variável resposta Y? Em caso positivo, qual é a distribuição?

3) Quais as suposições feitas pelo modelo de erros normais? O que estas suposições acarretam para Y?

4) O que significa "fazer extrapolação" no contexto de um modelo de regressão linear simples? Cite pelo menos dois riscos desta prática.

5) Defina o coeficiente de determinação (R^2) e explique quais valores ele pode assumir.

6) Em que situação é possível realizar um teste de falta de ajuste ("Lack-of-fit") e qual é o objetivo deste teste?

7) Quais os procedimentos gráficos podem ser usados para verificar as suposições enumeradas no item (2)?

8) Em que situação podemos utilizar um teste para a suposição de não auto-correlação entre os erros? Cito dois possíveis testes a serem usados nesta situação.

9) Quando é indicado o uso de transformação da variável resposta?

10) Que tipo de transformação é feita na variável resposta no método analítico de Box-Cox? Exemplifique.

11) Em que situação é usada a regressão inversa?

12) Por que o teste F da tabela ANOVA é equivalente ao teste t-student para as hipóteses H_o: $\beta_1 = 0$ contra H_a: $\beta_1 \neq 0$? (Mostre a equivalência entre as duas estatísticas de teste)

13) Na análise de resíduos, porque utilizamos o gráfico "resíduos" x "valores ajustados" e não o gráfico dos "resíduos"x "valores observados"?

❖ *Regressão Múltipla*

- Parte 1

1) **(Adaptação dos exercícios 3.LL e 6.H, Draper and Smith)** O gerente de um pequeno serviço de entregas contrata pessoal adicional sempre que o volume de serviço excede a carga de trabalho de seus usuais três empregados. Para verificar a eficácia desta idéia, ele registrou durante 13 dias seguidos as seguintes variáveis:

Variável Resposta: Y - Número de Entregas ;
Variáveis Explicativas: X - Número de Empregados (atuais mais extras) ;
Z - Número de Empregados que não estavam trabalhando em algum período do dia;

Os dados coletados estão disponíveis em na Tabela A.7 no anexo. Obs: nos três primeiros dias de coleta, alguns dos empregados usuais estavam de férias ou de licença médica.

a) Faça o diagrama de dispersão de Y versus X, Y versus Z e avalie a possibilidade do ajuste de um modelo de regressão linear.

b) Faça o gráfico em 3 dimensões de Y versus X e Z. (MINITAB: Graph > 3-D plot)

c) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.

d) Construa a Tabela de Análise de Variância.

e) Faça Análise dos Resíduos (considere o dia como ordem de coleta e faça também o gráfico dos resíduos versus a variável Z). Se existem problemas com as suposições do modelo de erros normais, quais são eles?

f) Caso não haja problemas com as suposições do modelo de erros normais, faça os testes F (Falta de Ajuste e Regressão) da Tabela Anova em (d).

g) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, encontrando a equação estimada.

h) Construa a Tabela de Análise de Variância, separando as SS seqüenciais.

i) Faça Análise dos Resíduos do modelo em (g) . Há algum problema?

j) Caso não haja problemas em (i), faça o teste da Falta de Ajuste da Tabela Anova em (h).

k) Caso não haja problemas no teste de falta de ajuste, faça os testes F seqüenciais da regressão (escreva as hipóteses nula e alternativa de cada teste).

l) Utilizando o teste t-Student, teste a significância de cada parâmetro individualmente. Os resultados concordam com os resultados dos testes F seqüenciais de (k)?

m) Interprete a equação de regressão estimada em (g).

n) Intervalo de Confiança para E[Y] dadas novas observações de X e Z : a matriz $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ pode ser armazenada no MINITAB (na janela Regression, botão Storage, marque a opção X'X inverse). Esta matriz será armazenada num objeto chamado m1. Para imprimir este objeto na janela Session , basta ir no menu Edit > Command Line Editor , digitar print m1 e pressionar Submit Commands . Esta é a matriz que será usada no cálculo do erro de estimação no intervalo de confiança para E[Y|(x,z)].

Considerando um número de empregados (X) igual a 5 e todos eles trabalhando todo o tempo (ou seja, Z = 0), construa um intervalo de 95% de confiança para E[Y], o número médio de entregas realizadas quando há 5 empregados trabalhando todo o tempo .

- Parte 2 – Detecção de pontos de influência

1) **Detectando pontos de influência -** Considere os seguintes exercícios das listas anteriores : 2 - parte 1; 1 – parte2; 1 – parte 3; 1 – parte 4; 2 – parte 5 e 1 – parte 6.

a) Faça a análise de resíduos à procura de pontos de influência. Use as medidas Hi, D-cook, *resíduos studentizados*.

b) Caso seja(m) detectado(s) ponto(s) de influência, ajuste o modelo sem este(s) ponto(s) e compare sua equação estimada com a equação estimada com todos os pontos para verificar o tamanho da influência deste(s) ponto(s).

- Parte 3 – modelo com ponderação

1) **(Adaptação dos dados da Tabela 3.8, Montgomery and Peck)** A renda mensal média de vendas de refeições (Y), assim como os gastos mensais com propaganda (X), foram registradas para 30 restaurantes. Um analista de vendas gostaria de encontrar uma relação entre as vendas e os gastos com propagandas.

Os dados coletados estão disponíveis em na Tabela A.8 no Anexo. (Os valores de Y e X foram arredondados para facilitar a resolução do problema)

a) Faça o diagrama de dispersão de Y versus X e avalie a possibilidade do ajuste de um modelo de regressão linear.

b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada

c) Faça Análise dos Resíduos do modelo em b). Se existem problemas com as suposições do modelo de erros normais, quais são eles?

d) Para corrigir o problema da heterocedasticidade, vamos proceder com a técnica dos

mínimos quadrados ponderados:

d.1) Calcule a estimativa do Erro Puro para cada nível de X com medidas repetidas (No MINITAB, use o comando Stat > Basics Statistics > Display Descriptive

d.2) Faça um gráfico de Var(Y|X), as estimativas do Erro Puro encontradas em d.1), versus nível de X . Existe relacionamento entre estas duas variáveis? Se sim, de que tipo?

d.3) Crie uma coluna de pesos e coloque o inverso da coluna X . Por que usar o inverso de X como peso? (Pense no relacionamento encontrado em d.2) e nos exemplos utilizados em sala).

d.4) Use os pesos construídos em f) para ajustar o modelo em b). (No MINITAB, na janela Regression, botão Options, selecionar a coluna com pesos no espaço weights. Não se esqueça de guardar os resíduos e os preditos).

e) Análise dos Resíduos: Crie uma coluna com a multiplicação da coluna de resíduos pela coluna da raiz quadrada dos pesos. Faça o mesmo com a coluna dos preditos e com a coluna dos valores de X .

f) Faça o gráfico de resíduos transformados versus preditos transformados. O problema da homocedasticidade foi resolvido?

g) Caso não haja problemas em i), construa a Tabela Anova e faça o teste da Falta de Ajuste da Tabela Anova.

h) Caso não haja problemas no teste de falta de ajuste, faça o teste F da regressão (escreva as hipóteses nula e alternativa de cada teste).

i) Utilize agora a transformação raiz quadrada em Y e ajuste o modelo de regressão linear, fazendo a análise de resíduos . Esta transformação resolve o problema da heterocedasticidade?

j) Analisando o valor do R^2, compare o ajuste do modelo em b) feito via mínimos quadrados ponderados com o ajuste feito via transformação “raiz quadrada” em Y. Por que não podemos comparar os valores do MSResidual?

- Parte 4 – Multicolinearidade e Análise de Variância via Análise de Regressão

1) **(Multicolinearidade)** Um grupo de estudantes participou de um experimento simples: cada estudante teve anotado sua altura (height), peso (weight), sexo (sex) , hábito de fumo (smokes), nível de atividade usual (activity) e pulso em repouso. Depois, eles correram no lugar durante um minuto e o pulso foi novamente medido. O objetivo é saber como predizer a medição do pulso depois da corrida através das variáveis medidas. Os dados estão na Tabela A.9 no Anexo.

Pulse1 - pulso antes da corrida (em batidas por minuto)
Pulse2 - pulso depois da corrida (em batidas por minuto)
Smokes - 1= fuma regularmente ; 2 = não fuma regularmente
Sex - 1 = homem 2 = mulher
Height - altura (em polegadas)
Weight - Peso (em libras)
Activity - Nível de atividade física : 1 = leve 2 = moderado 3 = intenso

a) Ajuste um modelo de regressão linear, entrando seqüencialmente com as variáveis: pulse1, Sex, height, weight, smokes, activity. A cada entrada de variável, faça o **teste F seqüencial**, avaliando a Soma de Quadrados Extra devida à variável que está entrando . Avalie os VIF´s (fatores de inflação da variância). (No MINITAB, janela Regression, botão Options).

b) Ajuste o modelo de regressão somente com as variáveis que deram contribuição significativa para a Soma de Quadrados de Regressão, avaliando também os VIF´s. Há indicação de problemas de multicolinearidade das variáveis explicativas?

c) Interprete o modelo ajustado em b).

2) **(Análise de Variância via Análise de Regressão)**

Pulse1 pulso antes da corrida (em batidas por minuto)
Activity Nível de atividade física : 1 = leve 2 = moderado 3 = intenso

Com os dados do exercício 1, vamos verificar se o pulso médio varia conforme o nível de atividade. Ou seja, devemos comparar a média do pulso em três grupos de indivíduos.
A hipótese nula é a de que o pulso médio é igual nos três grupos , e a hipótese alternativa é a de que pelo menos um dos grupos tem média diferente.
Estas são as hipóteses usadas na técnica de Análise de Variância, que pode ser realizada através de um modelo de regressão. Vejamos como:

a) Ajuste um modelo de regressão (com intercepto) da variável *pulse1* em função da variável *activity*. Lembre-se de que a variável activity é qualitativa e tem três níveis. Construa a Tabela Anova e teste a significância desta regressão, através do teste F. Em caso de rejeição de H_o, teste a significância de cada coeficiente em separado através do teste t.

b) Interprete o modelo ajustado. Qual é a diferença média entre o pulso de indivíduos do grupo de atividade física leve e o pulso de indivíduos do grupo de atividade física moderada ? E entre indivíduos do grupo de atividade física leve e os de atividade intensa? E entre os dos grupos moderada e intensa? (se a regressão não for considerada significante, essa interpretação servirá como prática).

c) Com o teste F em a), existem evidências estatísticas suficientes contra a hipótese de igualdade entre o pulso médio dos três grupos?

d) Utilizando a técnica da Análise de Variância, responda novamente a questão c).

e) Compare a tabela ANOVA de d) com a tabela ANOVA de a). O que se pode concluir?

- Parte 5 – Regressão Polinomial

1) (Adaptação de Montgomery and Peck, 2ª Edição : Modelos Polinomiais) O nível de carbonação (gás) de um refrigerante é afetado pela temperatura do produto e pela pressão da máquina que enche as garrafas. Para estudar este processo, foram coletados dados em 12 situações, que estão disponíveis na Tabela A.10 no Anexo

Y - carbonação da bebida
X - temperatura da bebida
Z - Pressão da máquina que enche a garrafa

a) Centralize as variáveis explicativas (X e Z) em torno de suas médias (No MINITAB, use o menu Calc ou o menu Edit > Command Line Editor com os seguintes comandos let c4 = c2-mean(c2) e let c5 = c3-mean(c3) , onde c2 e c3 são as colunas quem contém X e Z, respectivamente).

b) Faça um diagrama de dispersão de Y e X e outro para Y e Z, usando as variáveis centralizadas criadas em a). Com qual das duas variáveis (X ou Z) o relacionamento de Y parece ser mais forte? De que tipo parece ser este relacionamento?

c) Com a variável explicativa escolhida em b), ajuste um modelo de regressão linear simples. Faça o gráfico de resíduos versus preditos. Há algum problema com este gráfico?

d) Acrescente o termo quadrático ao modelo ajustado em c), guarde os resíduos e faça novamente o gráfico de resíduos versus preditos. O aspecto do gráfico melhora em relação ao do gráfico em c)?

e) Teste a contribuição do termo quadrático para a soma de quadrados de regressão através do teste F seqüencial.

f) Faça um gráfico dos resíduos do modelo em d) versus a variável explicativa (centralizada) que ficou de fora (X ou Z). Há algum padrão neste gráfico?

g) Acrescente a variável utilizada em f) (centralizada) ao modelo em d). Teste a contribuição desta variável para a soma de quadrados de regressão através do teste F seqüencial. Ela é significante? Em caso negativo, retire-a do modelo.

h) Ao modelo escolhido em g), acrescente o termo de interação entre X e Z (centralizado)(comando: let c10 = c4*c5, onde c4 e c5 são as colunas quem contém X e Z centralizadas, respectivamente). A contribuição do termo de interação para a soma de quadrados de regressão é significante (use o teste F seqüencial) ? Em caso negativo, retire-o do modelo.

i) Para o modelo escolhido em h), faça a análise de resíduos completa (gráficos de resíduos, probabilidade normal, testes, se possível, pontos de influência,

multicolinearidade (VIF´s)).

j) Faça o teste de falta de ajuste, se possível.

k) Caso o modelo passe pelo teste em j), faça o teste F da regressão e, em caso de significância estatística, faça o teste t individuais.

l) (Utilizando a equação escolhida) Para uma máquina operando a uma pressão de 23,5 e um produto à temperatura de 30, qual é o nível de carbonação esperado? (Lembre-se de que o modelo utiliza as variáveis centralizadas)

m) Construa um intervalo de 95% de confiança para o valor de Y, quando X e Z possuem os valores de l). Para calcular o erro de estimação, lembre-se de que será necessária a matriz $(X'X)^{-1}$. Para o modelo em h), ela pode ser armazenada em Storage, na janela Regression. Ela será armazenada no objeto m1. Para imprimí-lo, vá ate o menu Edit > Command Line Editor com o seguinte comando: print m1.

OBS: O MINITAB possui a janela do PROJECT MANAGER (gerenciador do projeto) onde estão as informações sobre todo o projeto: planilhas, colunas, objetos (constantes e matrizes). Além disto, é nesta janela onde podemos escrever informações sobre o projeto, descrições das colunas e objetos. Esta janela está sempre ativa no modo minimizado. Para vê-la, uma opção é minimizar todas as outras janelas, localizá-la e maximizá-la.

❖ *Exercícios de Revisão de Regressão Múltipla*

Considere o modelo de regressão linear múltipla, $\mathbf{Y = X\beta + \varepsilon}$, onde **Y, X,** β e ε são vetores ou matrizes.

1) Se dispomos de 100 "indivíduos" com observações em 5 variáveis consideradas explicativas, mais a variável resposta, quais são as dimensões de **Y, X,** β e ε ?

2) Qual é o método utilizado para estimar o vetor β? Para utilizar este método, é necessário supor alguma distribuição para a variável resposta Y? Em caso positivo, qual distribuição?

3) Quais são as suposições feitas pelo modelo de erros normais? O que estas suposições acarretam para Y?

4) Considerando o modelo de regressão linear múltipla, em que situação é possível realizar um teste de falta de ajuste ("lack-of-fit") e qual é objetivo deste teste?

5) Quais os procedimentos gráficos podem ser usados para verificar as suposições enumeradas no item (3) ? Que outros gráficos podem ser feitos na análise de resíduos?

6) Quais são as hipóteses nula e alternativa do teste F da tabela ANOVA ?

7) (Soma de Quadrados Extras ; Testes F seqüenciais). Pensando num modelo de regressão linear com três variáveis explicativas (X1, X2 e X3) e *n* observações, como montar a tabela ANOVA com a decomposição da soma de quadrados da regressão (SSReg) abaixo?

Explique como obter as SSReg´s da tabela, quais seriam os respectivos graus de liberdade (g.l.), como obter os MS (quadrados médios) e as respectivas estatísticas F.

Fonte	**SS**	**g.l**	**MS**	**F**
Regressão (X1, X2, X3)				
X1 **X2 | X1** **X3 | X1, X2**				
Resíduo (Erro)				
Total				

8) Quais as hipóteses nula e alternativa de cada um dos testes F da tabela ANOVA em (7)?

9) O que é multicolinearidade e o que este problema pode causar na análise de regressão?

10) Quais são os tipos de pontos de influência e como detectá-los?

11) Em qual(is) situação(ões) é indicado o uso do Método dos Mínimos Quadrados Ponderados (MQP) ao invés do Método dos Mínimos Quadrados Ordinários (MQO) na estimação da equação de regressão? Qual é a diferença entre os dois métodos? Quais são as conseqüências de se usar o MQO quando o MQP seria o método indicado?

12) Compare a transformação de Box-Cox e o MQP como alternativas para estabilizar a variância dos erros, citando vantagens e desvantagens.

13) Quais são as vantagens da centralização das variáveis explicativas em suas médias para a estimação dos parâmetros da regressão ? (Pense em termos da matriz **(X'X)**)

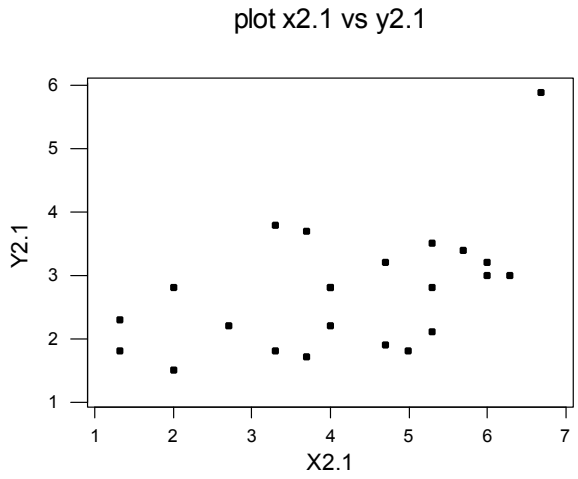
- **Respostas:**

❖ *Regressão linear simples:*

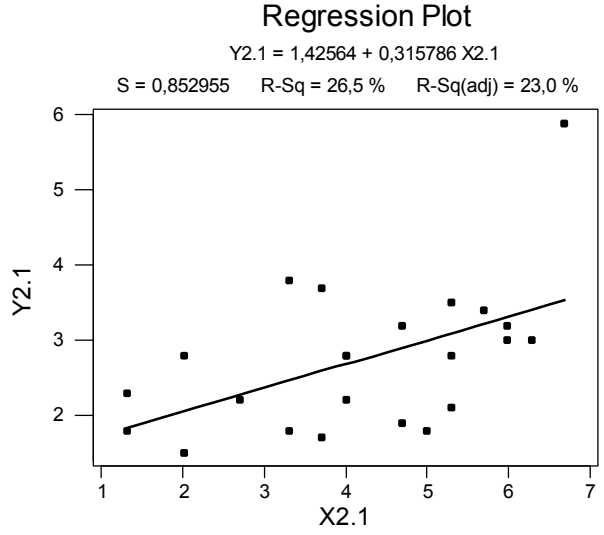
- <u>Parte 1</u>

1) Utilizando os dados da Tabela A.1 no Anexo. (página 51, Draper & Smith, 3 ed.) :

a) Faça o diagrama de dispersão.



b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.



c) Construa a Tabela de Análise de Variância e calcule o R^2.

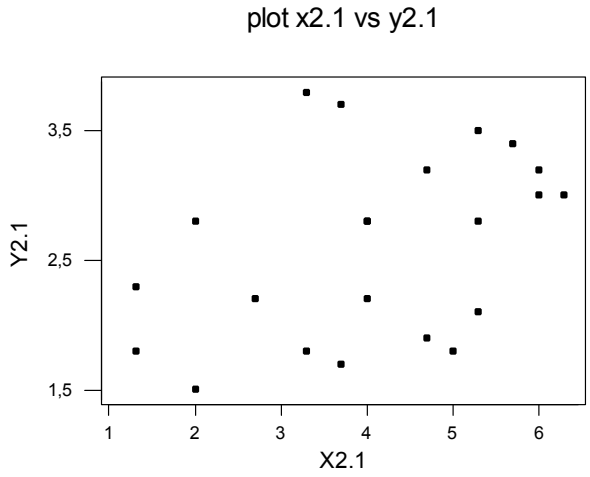
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	5,4992	5,4992	7,56	0,012
Residual Error	21	15,2782	0,7275		
Total	22	20,7774			

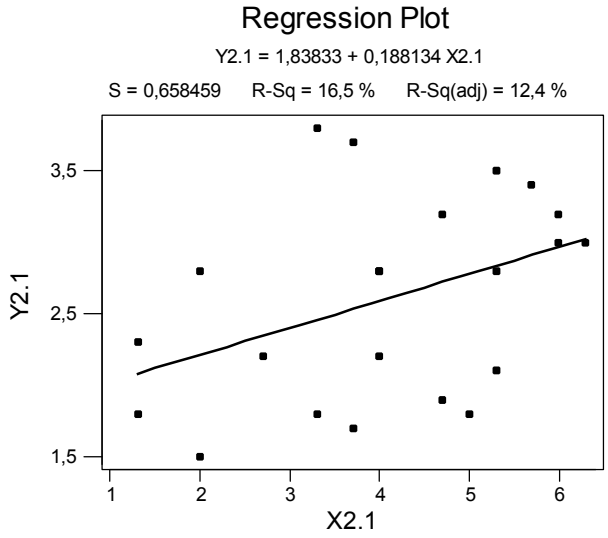
O valor de R^2 é: 26,5% .

d) Retire o par de observações no. 16 (Y=5.9 ; X = 6.7) e refaça os itens de a) a c).

a)



b)



c)

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1,7182	1,71818	3,96288	0,060
Error	20	8,6714	0,43357		
Total	21	10,3895			

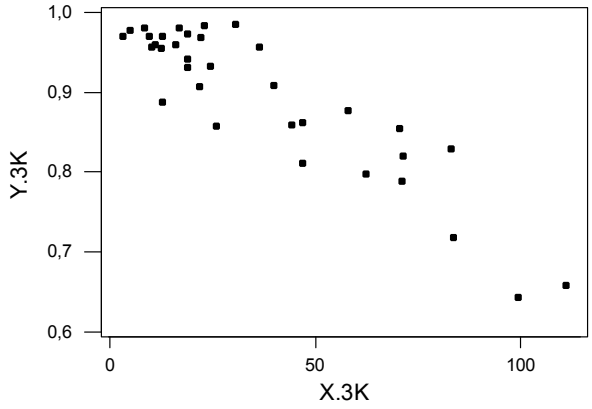
R^2 = *16,5%* .

e) Comparando somente os valores de R2, quais dos dois modelos é o melhor? O par de observações nº 16 influencia a qualidade do ajuste ?

Observando-se apenas os valores dos coeficientes de determinação dos dois modelos, vê-se que o modelo relativo à questão (b) é melhor, pois este apresenta maior R2 (26,5%). Pode-se notar ainda que, pelo fato de haver ocorrido mudanças significativas na regressão como um todo, a observação que foi retirada foi modelo estava influenciando o mesmo. Note que esta influencia é negativa, pois houve um decréscimo no valor do R2 e um aumento no valor P da regressão. Neste caso seria melhor estudar a possibilidade de se retirar a observação influente do modelo.

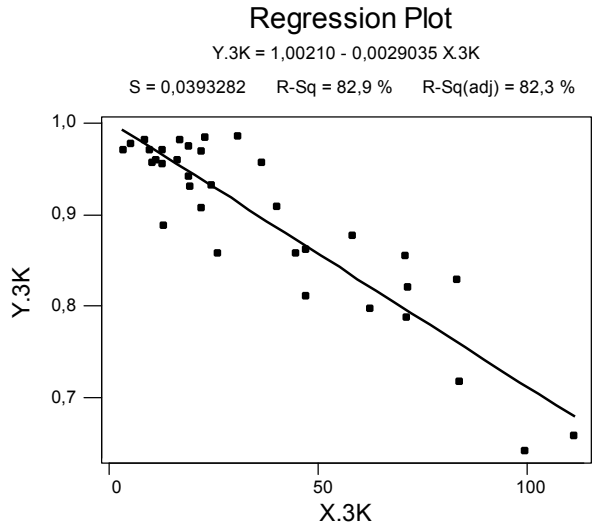
2) Os dados deste exercício são do exercício K (Capítulo 3) do livro de Draper & Smith e estão na tabela A.2 no Anexo. A variável resposta (Y.3K) representa a porcentagem de amendoins não-contaminados por certo fungo em um lote e a variável explicativa (X.3K) representa a quantidade média de uma substância química para evitar contaminação em cada 60 gramas de amendoins.

a) Faça o diagrama de dispersão.



b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.

A reta estimada é: Y.3K = 1,00 - 0,00290 X.3K



c) Construa a Tabela de Análise de Variância e calcule o R_2.

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	0,23915	0,23915	154,62	0,000
Residual Error	32	0,04949	0,00155		
Total	33	0,28864			

O valor do R_2 é: 82,9%

d) Este conjunto de dados possui dois níveis de X com medidas repetidas (X = 18,8 e X = 46,8). Entretanto, alguns níveis de X tem valores "muito próximos" que, na prática, poderiam ser considerados "iguais" e, assim, os valores de Y nestes níveis poderiam ser considerados medidas repetidas. São eles:

X = 9,3 ; 9,9
X = 12,3 ; 12,5 e 12,6
X = 18,8 ; 18,8 ; 18,9
X = 21,7 ; 21,9
X = 46,8 ; 46,8 (estes são realmente medidas repetidas)
X = 70,6 ; 71,1 ; 71,3
X = 83,2 ; 83,6 .

e) Considere os valores de Y nestes níveis como sendo medidas repetidas e calcule a soma de quadrados do erro puro (SSErroPuro). Este valor é , claro, uma aproximação. Encontre também os graus de liberdade desta soma

SSErroPuro = 0,01678
Graus de liberdade = 10

f) Construa a nova Tabela de Análise de Variância, agora com a SSResidual desmembrada em SSErroPuro e o SSL (SS da falta-de-ajuste). Faça o teste F da falta-de-ajuste.

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	0,23897	0,23897	153,95	0,000
Residual Error	32	0,04967	0,00155		
Lack of Fit	22	0,03289	0,00150	0,89	0,610
Pure Error	10	0,01678	0,00168		
Total	33	0,28864			

Teste de Falta de Ajuste:

H_o: Não há falta de ajuste
H_a: Há falta de ajuste

O valor observado de F foi de 0,89.

Região Crítica = {F: F > 2,7740} , Nível de significância = 0,05.

Como 0,89 não está na região crítica, então pode-se afirmar que o modelo não apresenta falta de ajuste.

g) Caso o teste F da falta-de-ajuste seja não-significante, faça o teste F geral.

H_o: $\beta 1 = 0$, isto é, o modelo não é razoável
H_a: $\beta 1 \neq 0$, ou seja o modelo é razoável .

O valor observado de F foi de 153,95.

Região Crítica = {F: F > 4,1709} , Nível de significância = 0,05.

Como não está na região crítica, então se pode afirmar que o $\beta 1$ é diferente de zero, logo o modelo parece descrever bem os dados.

h) Interprete os coeficientes da reta de regressão.

Caso a quantidade média de uma substância química para evitar contaminação em cada 60 gramas de amendoins seja igual a zero, teremos 100% de amendoins contaminados.

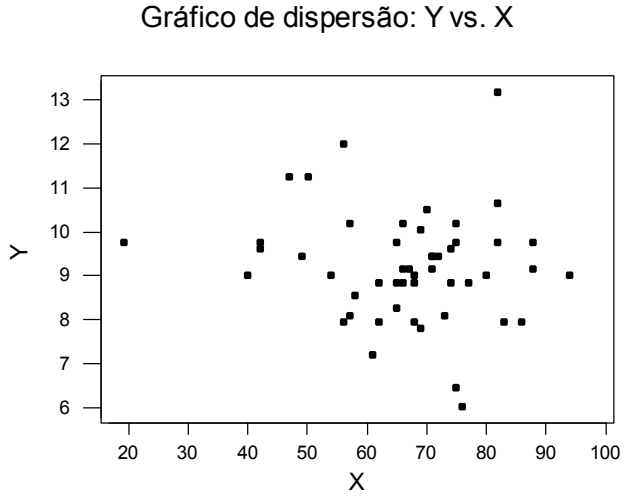
E para cada aumento de uma unidade na quantidade média da substância química para evitar contaminação haverá um decréscimo de 0,00290 na porcentagem de amendoins não contaminados em um lote.

- Parte 2

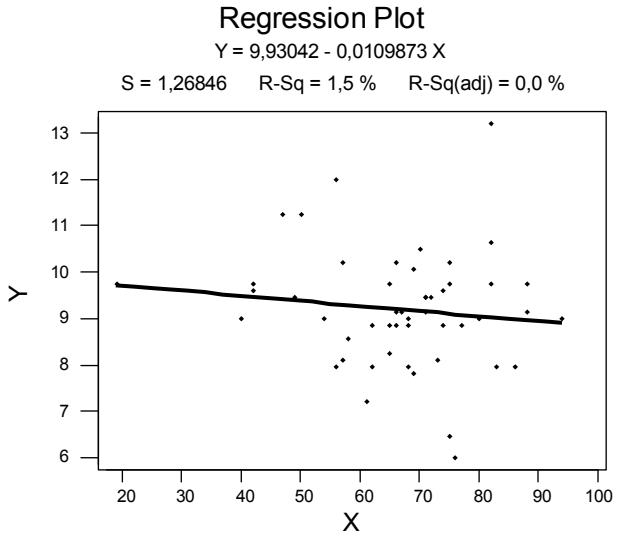
1) Utilizando os dados da Tabela A.3 do Anexo. (Exercício V, capítulo 3, Draper & Smith, 3 ed., página 105) :

Variável Resposta: Y = tamanho da "linha da vida" da mão esquerda (em cm) ;
Variável Explicativa: X = idade da pessoa ao morrer (em anos);

a) Faça o diagrama de dispersão.



b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.



c) Construa a Tabela de Análise de Variância com a SSResidual desmembrada em SSErroPuro e o SSL (SS da falta-de-ajuste). Faça o teste F da falta-de-ajuste.

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	1,178	1,178	0,73	0,397
Residual Error	48	77,232	1,609		
Lack of Fit	29	45,777	1,579	0,95	0,557
Pure Error	19	31,455	1,656		
Total	49	78,410			

Teste de Falta de Ajuste:

H_o: Não há falta de ajuste
H_a: Há falta de ajuste

Observando que o P-valor da falta de ajuste é de maior que 0,05 (0,557), conclui-se que o modelo não apresenta falta de ajuste.

d) Faça o teste F da regressão (Escreva hipóteses nula e alternativa, faça o teste e conclua).

H_o: $\beta_1 = 0$, isto é, o modelo não é razoável
H_a: $\beta_1 \neq 0$, ou seja o modelo é razoável .

Sendo o P-valor da regressão igual a 0,397, isto é, maior que 0,05, verifica-se que o modelo não é razoável, pois β_1, que é o parâmetro mais importante do modelo, é igual a zero.

e) Calcule o valor de R^2 e o valor de max(R^2) e faça a interpretação de R^2.

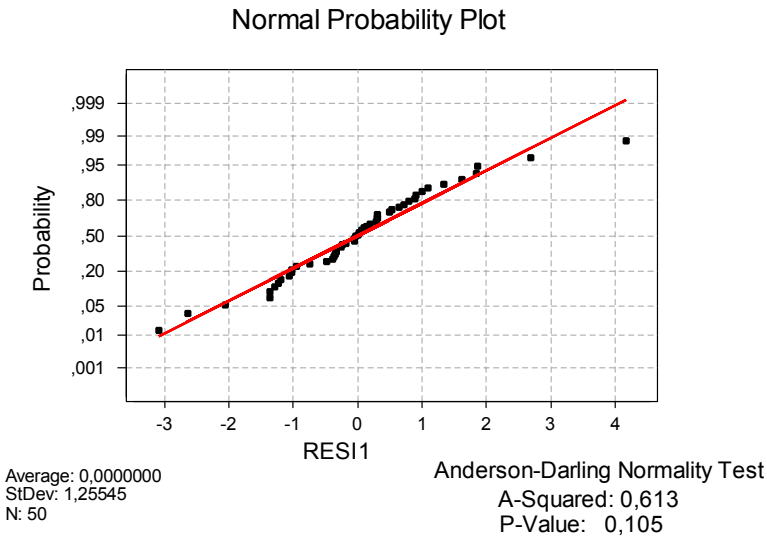
$R^2 = 1,5\%$

$$Max(R^2) = 1 - \frac{SSerropuro}{SStotal} = 0,4012$$

$R_2/Max(R_2) = 0,03739$

Através do valor do coeficiente de determinação vê-se que a variação de Y que explicada pela reta de regressão é muito pequena.

f) Verifique a suposição de normalidade dos resíduos através do gráfico de probabilidade Normal.

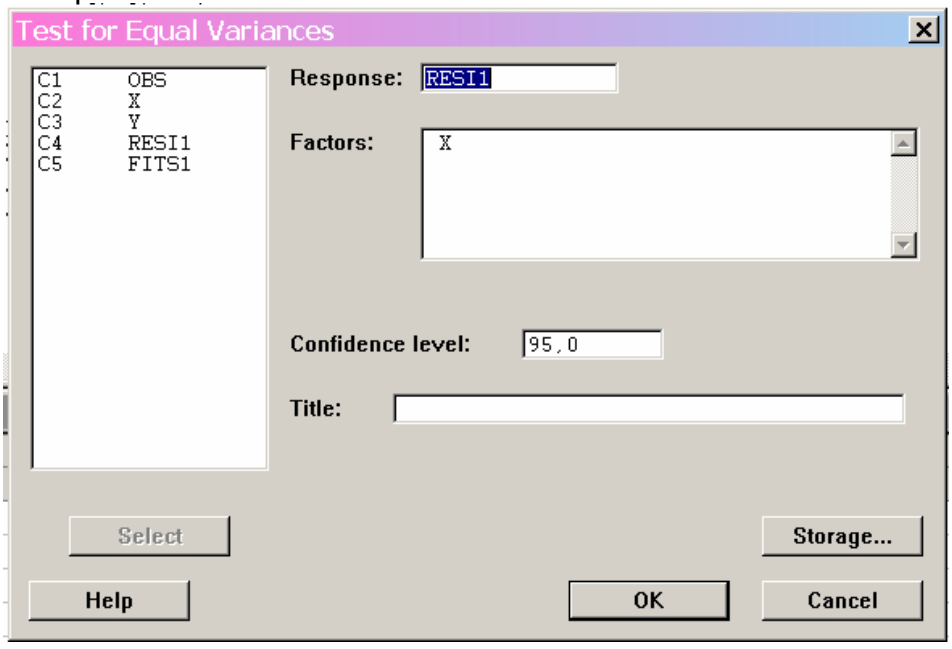


Teste de Normalidade:
H_o: Os resíduos seguem a distribuição normal
H_a: Os resíduos não seguem a distribuição normal

Como o P-valor do teste de Anderson –Darling foi maior que 0,05 pode-se admitir que os resíduos são normalmente distribuídos.

g) Faça o teste da homogeneidade do erro puro (Bartllet e Levene).

Como fazer o teste:
1º) Ir em : STAT > ANOVA> TEST FOR EQUAL VARIANCES
2º) em seguida aparecerá a Janela:



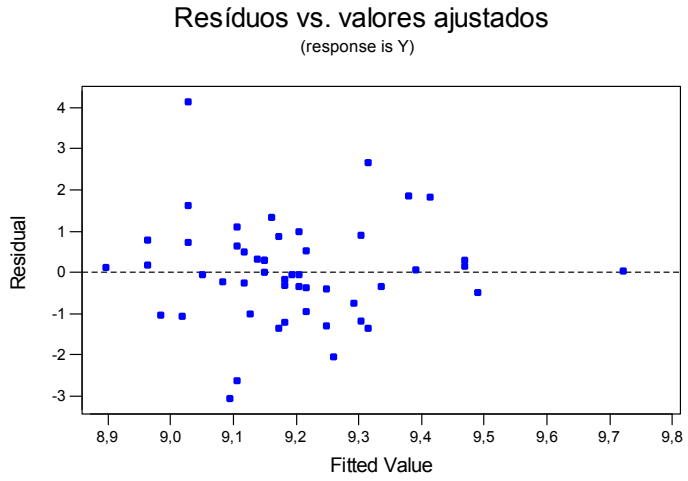
Na qual basta colocar a coluna dos resíduos no local escrito Response e selecionar a coluna com a variável X onde está escrito Factors.
A saída será parecida com a abaixo, porém com alguns detalhes a mais.

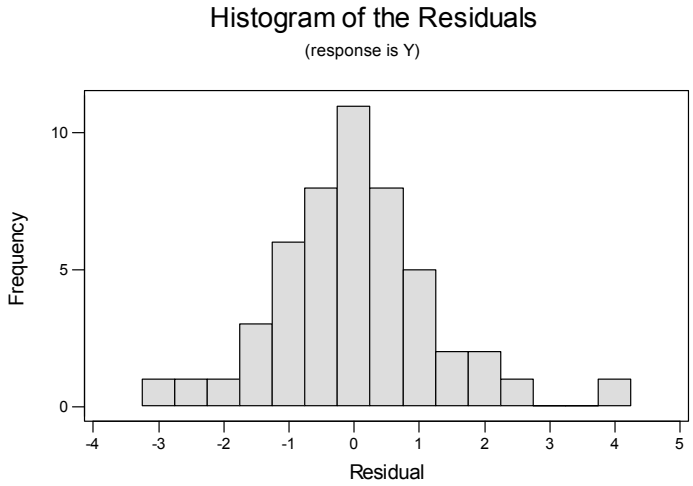
H_o: Os resíduos têm variância constante.
H_a: Os resíduos não têm variância constante.

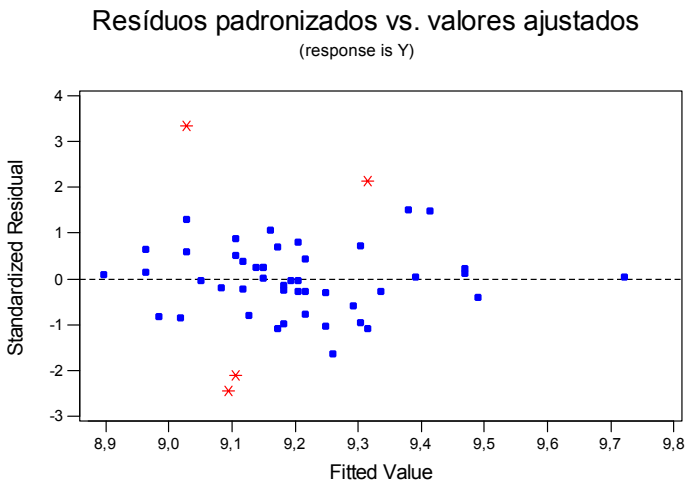
```
Bartlett's Test

Test Statistic: 16,228
P-Value       :  0,181

Levene's Test

Test Statistic: 1,239
P-Value       : 0,328
```

Como nos dois testes a probabilidade de significância foi maior que 0,05 a hipótese de que os erros possuem variância constante não foi rejeitada.

h) Analise os gráficos de resíduos apropriados.







Obs.: os asteriscos representam os pontos que estão fora do intervalo (-2, 2).

Analisando-se o gráfico dos resíduos versus os ŷ, vê-se que parece existir um dado atípico, o que pode estar influenciando a variância dos resíduos, fazendo com que esta pareça não ser constante. Ainda através da análise deste gráfico nota-se que existe uma tendência não linear dos resíduos. Pelo gráfico dos resíduos padronizados contra os ŷ percebe-se que existem 4 pontos (ou seja, 8% dos dados) que estão fora do intervalo (-2, 2), como esta porcentagem é maior que 5% isto poderia estar comprometendo a normalidade dos resíduos. O que não acontece, como pode ser averiguado pelo histograma dos resíduos (que está de acordo com o teste de normalidade realizado no item(f)). Sendo assim estas observações podem ser atípicas ou apresentarem algum outro problema.

i) Reporte os possíveis problemas encontrados na análise dos resíduos (itens *f*, *g* e *h*) .

Os resíduos não apresentaram grandes problemas, porém existem algumas observações que podem estar prejudicando o modelo, principalmente no que se trata à variância, como foi destacado no item anterior.

j) Faça o teste H_0: $\beta_0 = 0$ contra H_a: $\beta_0 \neq 0$.

H_0: $\beta_0 = 0$
H_a: $\beta_0 \neq 0$

```
The regression equation is
Y = 9,93 - 0,0110 X

Predictor        Coef     SE Coef          T        P
Constant       9,9304      0,8747      11,35    0,000
X           -0,01099      0,01284      -0,86    0,397
```

Como o P-valor de β_0 é aproximadamente zero pode-se dizer que esse parâmetro é significante para o modelo.

f) A partir de suas análises nos itens anteriores, conclua sobre a relação entre Y e X.

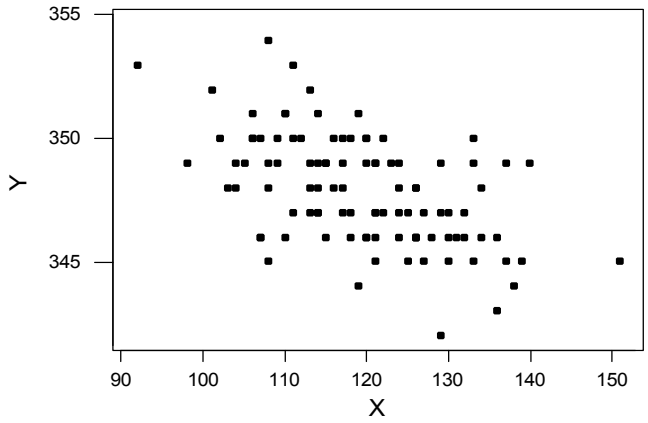
A relação entre X e Y não é claramente linear como pode ser visto no gráfico de dispersão. Isto pode ser explicado pelo fato de haver alguns dados muito afastados da nuvem de pontos. O que atrapalha também na detecção de uma relação clara entre as variáveis em questão. Na verdade, não parece existir relacionamento algum entre Y e X.

- Parte 3 – Regressão Simples e Regressão Inversa

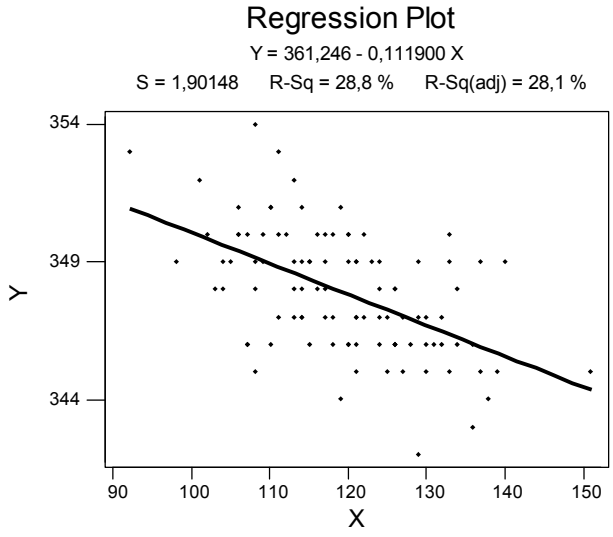
1) Num estudo retrospectivo sobre a possível relação entre "o tempo de utilização de um plano de previdência" e o "tempo de contribuição do beneficiário" , ambos medidos em meses, uma amostra de 100 beneficiários de um plano de previdência tiveram essas duas variáveis registradas.

Variável resposta: Y = tempo de contribuição, em meses.
Variável explicativa: X = tempo de utilização do benefício, em meses (tempo entre a data da aposentadoria e a data do falecimento do beneficiário).

a) Faça o diagrama de dispersão.



b) Ajuste o modelo de regressão linear adequado, encontrando a reta estimada.



c) Construa a tabela de análise de variância com a SSresidual desmembrada em Sserropuro e o SSL (SS da falta de ajuste). E faça o teste da falta de ajuste.

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	143,46	143,46	39,68	0,000
Residual Error	98	354,33	3,62		
Lack of Fit	40	89,03	2,23	0,49	0,991
Pure Error	58	265,30	4,57		
Total	99	497,79			

Teste de Falta de Ajuste:

H_o: Não há falta de ajuste
H_a: Há falta de ajuste

Como o Valor P do teste é maior que 0,05 pode-se dizer que o modelo não apresenta falta de ajuste.

d) Faça o teste F da regressão (escreva a hipótese nula e alternativa, faça o teste e conclua).

H_o: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

Observa-se que a probabilidade de significância deste teste é inferior a 0,05, o que nos possibilita afirmar que o modelo ajustado é razoável, pois a hipótese de que $\beta_1 = 0$ foi rejeitada.

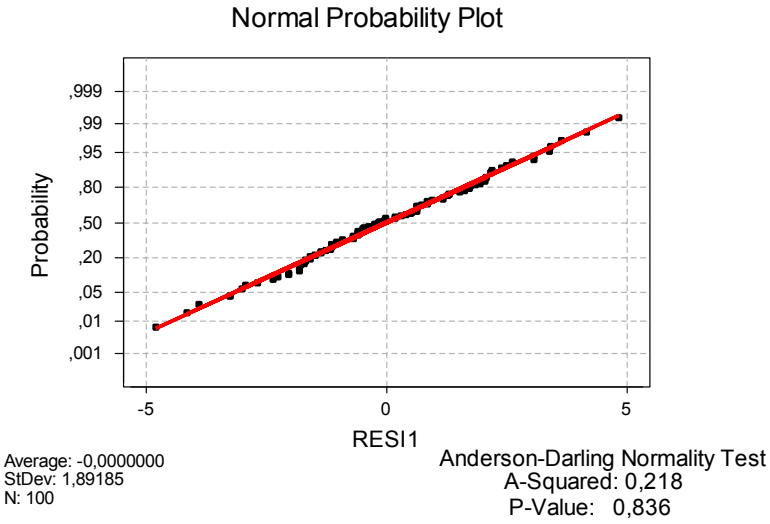
e) Calcule o valor de R^2 e do max(R^2) e faça a interpretação do R^2.

max(R^2) = 0,4670.
R^2 = 28,8%.

$R^2/max(R^2) = 0,288/0,4670 = 0,6166$

Apesar do modelo não apresentar falta de ajuste o valor do coeficiente de determinação é razoável, pois a porcentagem da variabilidade de Y que é possível de ser explicada por X vale 61,66%.

f) Verifique a suposição de normalidade dos resíduos através do gráfico de probabilidade Normal.



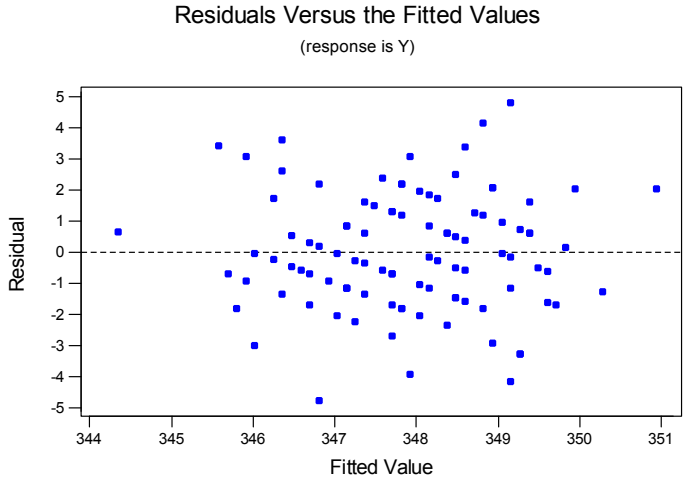
g) Faça o teste da homogeneidade do erro puro (Bartlet e Levene).

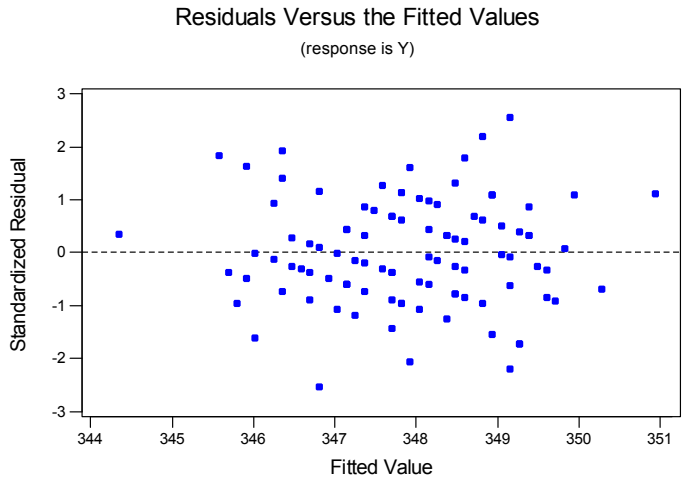
H_o: Os resíduos têm variância constante.
H_a: Os resíduos não têm variância constante.

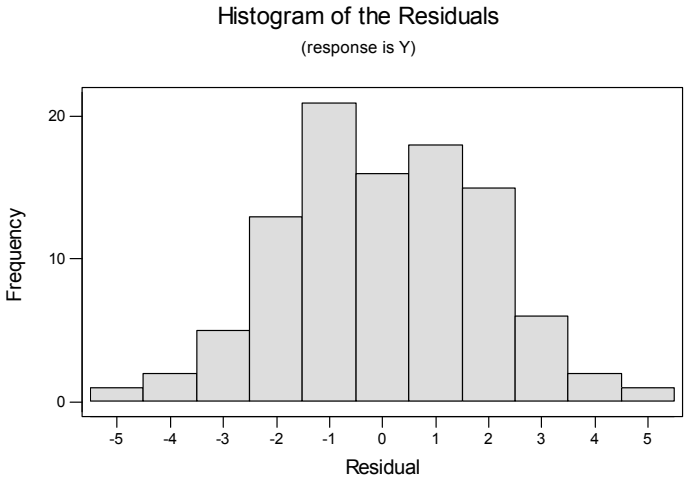
```
Bartlett's Test

Test Statistic: 19,981
P-Value       :  0,832

Levene's Test

Test Statistic: 0,631
P-Value       : 0,904
```

É possível afirmar que os resíduos possuem homocedasticidade, pois ambos P-valores, do teste de Bartlet e do teste de Levene, são maiores que 0,05.

h) Analise os gráficos de resíduos apropriados.







Através do histograma acima verifica-se que os resíduos são normalmente distribuídos. Pelo primeiro gráfico apresentado nesta questão pode-se considerar que os resíduos possuem uma variância razoavelmente constante.

i) Reporte os possíveis problemas encontrados na análise de resíduos.

Os resíduos não apresentaram problemas, pois as análises anteriormente feitas mostraram que eles são normalmente distribuídos, razoavelmente homocedásticos e aleatórios.

j) Faça o teste H_o: $\beta_o = 0$ contra H_a : $\neq \beta_o$.

H_0: $\beta_0 = 0$
H_a: $\beta_0 \neq 0$

Ao analisar-se o P-valor de β_0 vê-se que esse é aproximadamente zero, logo a hipótese de que β_0 é igual a zero pode ser refutada.

k) A partir das suas análises anteriores conclua sobre a relação entre Y e X.

A relação entre X e Y parece realmente ser linear, como pode ser verificado pelo gráfico de dispersão, porém não é uma relação muito forte.

l) Regressão inversa: como o estudo foi retrospectivo, a partir do falecimento do beneficiário foi possível estabelecer o valor da variável explicativa e, então o valor da resposta para aquele nível ad variável explicativa. Porém, na prática, gostaríamos de estudar a relação inversa, ou seja, a partir do tempo de contribuição gostaríamos de predizer o tempo de uso do benefício. Deste modo, usaremos a regressão inversa.

l.1) A partir da reta estimada em (b), estabeleça a equação da regressão inversa, Isto é, X como função de Y.

A equação de regressão inversa é: $X_0 = \dfrac{361{,}246 - Y_0}{0{,}112}$

l.2) Dado o valor do tempo de contribuição igual a 348 meses, estime o valor médio do tempo de uso do benefício.

O valor médio do tempo de uso do benefício ($\hat{X}$) é: 118,268

l.3) Estabeleça o intervalo a 95% de confiança para o tempo de uso do beneficiário quando o tempo de contribuição for igual a 355 meses.

IC = $\hat{X}_0 \pm t_{\frac{\alpha}{2},n-2}\left[\dfrac{QMR}{\beta_1^2}\left(1+\dfrac{1}{n}+\dfrac{(\hat{X}_0-\overline{X})^2}{Sxx}\right)\right]^{\frac{1}{2}}$

Sendo $\hat{X}_0$ *= 55,7678 , Sxx = 11457,04 e* $t_{\alpha/2\,,\,n-2}$ *= 1,96 , temos que:*

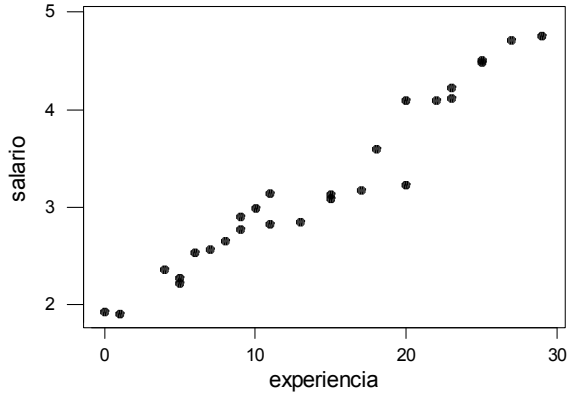
$IC_{95\%}$ = [16,8285; 94,7071]

- Parte 4

1) Um investigador deseja estudar a possível relação entre os salários e o tempo de experiência no cargo de gerente de agências bancárias de uma grande empresa. Além disto, gostaria de saber se há diferenças quando são levados em conta homens e mulheres separadamente. Os dados coletados estão disponíveis na Tabela A.5 no Anexo e a descrição do banco de dados segue abaixo.

Variável Resposta: - Salário, em mil reais ;
Variáveis Explicativas: - Experiência = tempo de trabalho no cargo, em anos completos ;
- Sexo = sexo do empregado (0 – feminino ; 1 – masculino) .

a) Faça o diagrama de dispersão do salário versus experiência e avalie a possibilidade do ajuste de um modelo de regressão linear.



b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, sendo X a variável "experiência" , encontrando a reta estimada.

The regression equation is
salário = 1,83 + 0,0998 experiência

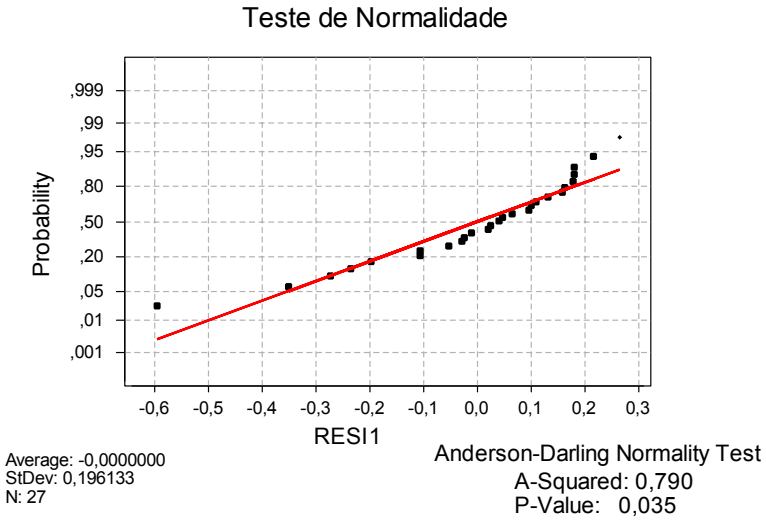
c) Construa a Tabela de Análise de Variância e calcule o valor de R^2.

```
Analysis of Variance

Source            DF          SS          MS          F         P
Regression         1      18,154      18,154    453,77    0,000
Residual Error    25       1,000       0,040
  Lack of Fit     18       0,560       0,031      0,49    0,892
  Pure Error       7       0,440       0,063
Total             26      19,154

S = 0,2000      R-Sq = 94,8%     R-Sq(adj) = 94,6%
```

Max(R^2) = 1 – (0,440/19,154) = 1 – 0,0229 = 0,977

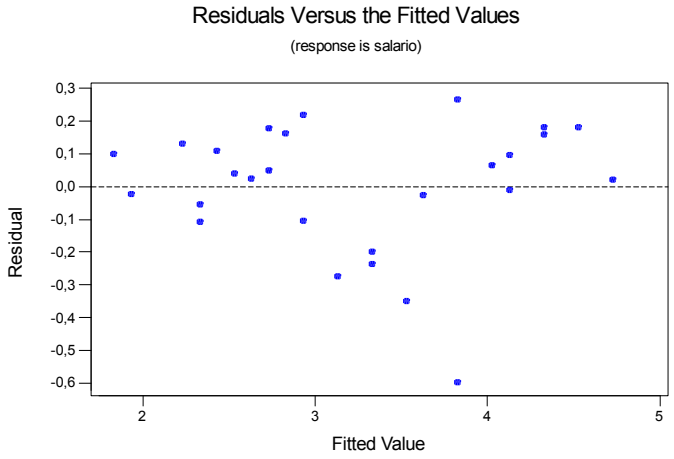
0,948/0,977 = 0,97 (a variável experiência explica 97% da variabilidade dos salários que pode ser explicada).

d) Verifique a suposição de normalidade dos resíduos através do gráfico de probabilidade Normal.



P-valor do teste Anderson-Darling = 0,035 (a hipótese de normalidade dos resíduos é rejeitada a 5%)

e) Analise o gráfico *resíduos* versus *ajustados* (*preditos*). Os resíduos parecem se distribuir aleatoriamente em torno do valor zero?

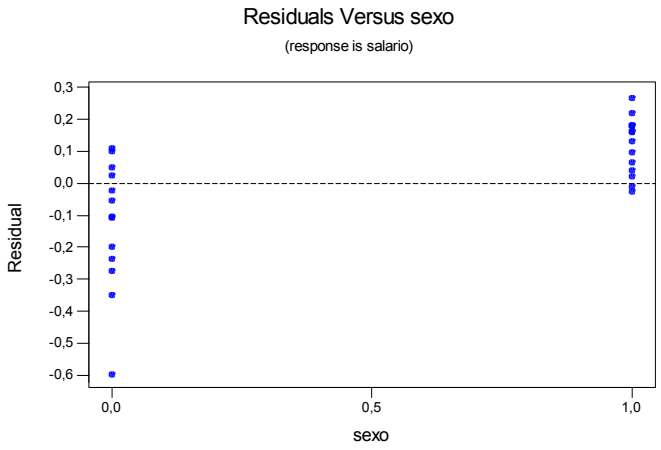


Não, há agrupamentos de resíduos, ora acima de zero, ora abaixo de zero.

f) Analisando as respostas aos itens d) e e), o modelo ajustado em b) parece ser adequado?

Não, pois a suposições de normalidade foi violada e há indícios de que os resíduos não se distribuem aleatoriamente em torno do valor zero, existindo relação entre eles e os valores ajustados.

g) Analise o gráfico *resíduos* versus *sexo*. O que se pode concluir?



Existe clara correlação entre os resíduos e a informação sobre o sexo do empregado.

h) Para cada sexo separadamente, repita os itens de b) a e).

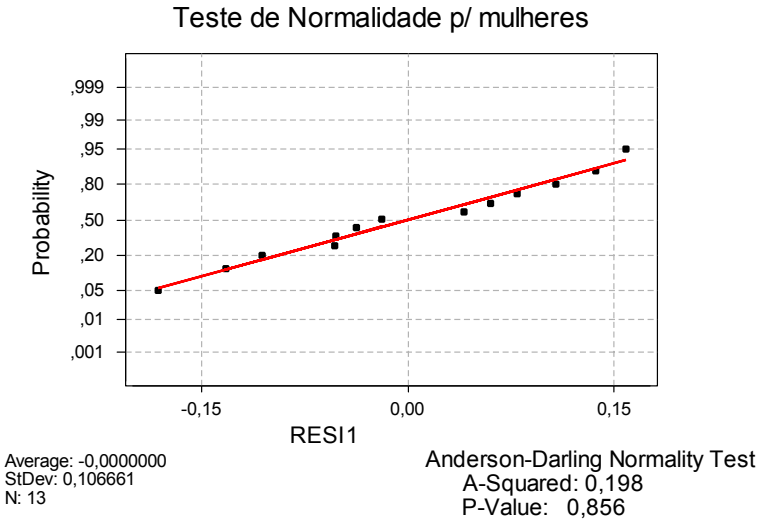
<u>Sexo feminino:</u>

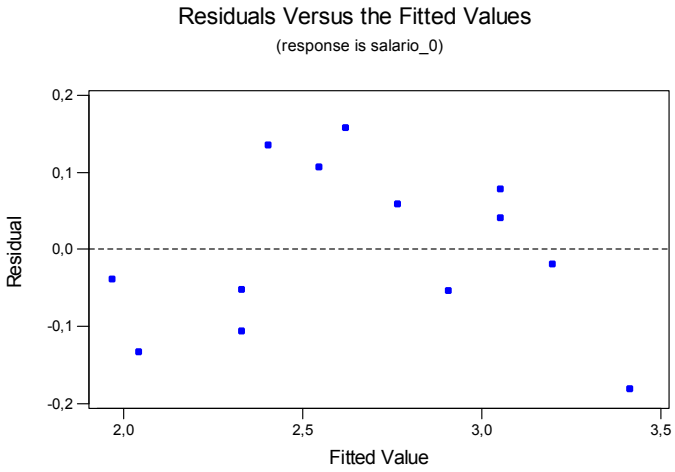
```
The regression equation is
salario_0 = 1,97 + 0,0722 experiencia_0

Predictor        Coef     SE Coef          T        P
Constant      1,96844     0,05877      33,49    0,000
experien     0,072199    0,005199      13,89    0,000

S = 0,1114      R-Sq = 94,6%     R-Sq(adj) = 94,1%
```

Max(R^2) = 1 – (0,0022/2,5296) = 0,999
0,946 / 0,999 = 0,9469

```
Analysis of Variance

Source            DF          SS          MS          F        P
Regression         1      2,3931      2,3931     192,82    0,000
Residual Error    11      0,1365      0,0124
  Lack of Fit      9      0,1343      0,0149      13,63    0,070
  Pure Error       2      0,0022      0,0011
Total             12      2,5296
```

<u>Comentários</u> : o problema da normalidade dos resíduos foi corrigido, mas ainda há problemas com a distribuição dos resíduos em torno do zero, que não parece ser aleatória.





<u>Sexo masculino</u> :

The regression equation is

salario_1 = 1,98 + 0,0983 experiencia_1

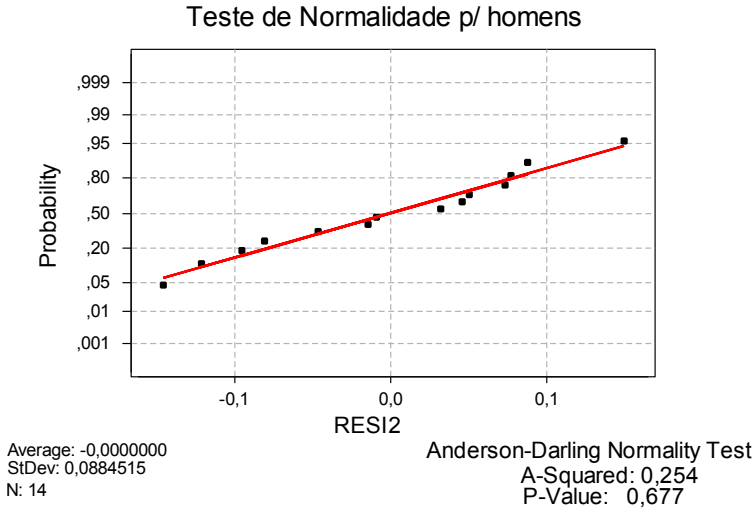
Predictor	Coef	SE Coef	T	P
Constant	1,97753	0,06122	32,30	0,000
experien	0,098261	0,003102	31,68	0,000

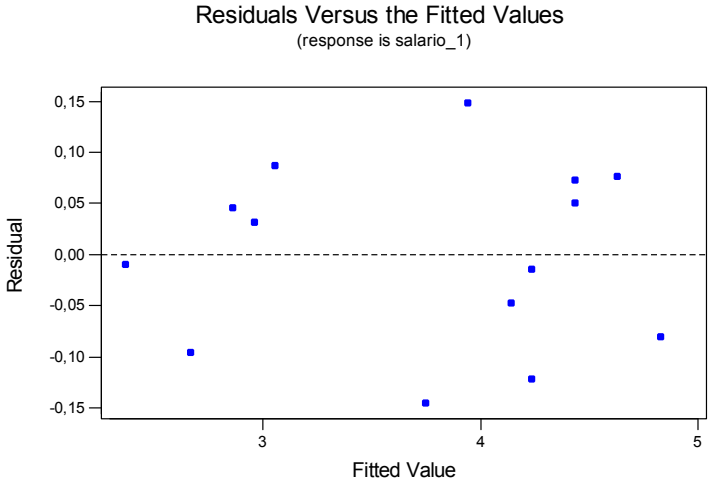
S = 0,09206 R-Sq = 98,8% R-Sq(adj) = 98,7%

Max(R^2) = 1 – (0,0061/8,6073) = 0,999
0,988 / 0,999 = 0,9887

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	8,5056	8,5056	1003,54	0,000
Residual Error	12	0,1017	0,0085		
Lack of Fit	10	0,0956	0,0096	3,15	0,265
Pure Error	2	0,0061	0,0030		
Total	13	8,6073			





<u>Comentários</u> : o problema da normalidade dos resíduos foi corrigido, mas ainda há problemas com a distribuição dos resíduos em torno do zero, que não parece ser aleatória.

i) Para cada sexo separadamente, faça o teste F da regressão (escreva hipóteses nula e alternativa, faça o teste e conclua).

<u>Sexo feminino</u> :

H_o: $\beta_1 = 0$ (A variável experiência não explica uma parte significativa da variabilidade dos salários entre as mulheres)
H_a: $\beta_1 \neq 0$ (A variável experiência explica uma parte significativa da variabilidade dos salários entre as mulheres)

Estatística F da ANOVA = 192,82 .
Comparar com o percentil 95 da $F_{1\,;\,11} = 4,8443$

Rejeitar a H_0, ou seja, existem evidências estatísticas de que a regressão dos salários na variável experiência é significante a 5% no grupo das mulheres.

Sexo masculino :

H_o: $\beta_1 = 0$ (A variável experiência não explica uma parte significativa da variabilidade dos salários entre os homens)
H_a: $\beta_1 \neq 0$ (A variável experiência explica uma parte significativa da variabilidade dos salários entre os homens)

Estatística F da ANOVA = 1003,54 .
Comparar com o percentil 95 da $F_{1\ ;\ 12}$ = 4,7472

Rejeitar a H_0, ou seja, existem evidências estatísticas de que a regressão dos salários na variável experiência é significante a 5% no grupo dos homens.

j) Compare os valores de R^2 dos modelos em separado com o valor calculado em c) O que se pode concluir?

Modelo	*R^2 / max(R^2)*
Geral	*0,977*
Homens	*0,989*
Mulheres	*0,947*

Em termos de R^2, houve um pequeno ganho no grupo de homens em relação ao modelo geral, mas uma pequena perda no grupo de mulheres.

k) Faça a mesma comparação usando o valor do MSResidual das tabelas ANOVA. Lembre-se de que o MSResidual é a estimativa da variância da resposta (Utilize o conceito de desvio-padrão, se achar mais fácil sua análise).

Modelo	*MSResidual (s)*
Geral	*0,040 (0,200)*
Homens	*0,0085 (0,092)*
Mulheres	*0,0124 (0,111)*

Os dois modelos, tanto para homens, quanto para mulheres, conseguiram uma redução na variância, evidenciando que uma parte da variância dos salários pode ser explicado pelo sexo do empregado. A redução foi maior entre os homens.

l) Interprete a reta de regressão estimada para cada sexo e tire suas conclusões sobre a relação entre "salário" e "experiência" para os gerentes de banco desta empresa.

As retas são

Sexo feminino
salario_0 = 1,97 + 0,0722 experiencia_0

Sexo masculino
salario_1 = 1,98 + 0,0983 experiencia_1

O intercepto é praticamente o mesmo para ambos os grupos, indicando que um empregado com menos de um 1 ano de experiência ganha, em média, 1,97 mil reais (mulheres) e 1,98 mil (homens) .

Já o coeficiente angular mostra uma maior inclinação da reta para o grupo de homens, indicando que, para um mesmo ganho na experiência, o aumento médio no salário dos homens é maior do que das mulheres. No grupo de mulheres, a cada ano de experiência, há um aumento médio no salário de R$72,20. No grupo de homens, este aumento é R$98,30.

Obs: ainda há problemas nos modelos separados, como vimos na análise de resíduos. Uma das maneiras de solucionar é tentar um modelo de regressão múltipla, onde iremos considerar a interação entre a experiência e o sexo do empregado.

$Y = \beta_0 + \beta_1(\text{experiência}) + \beta_2(\text{sexo}) + \beta_{12}(\text{sexo*experiência}) + \text{erro}$

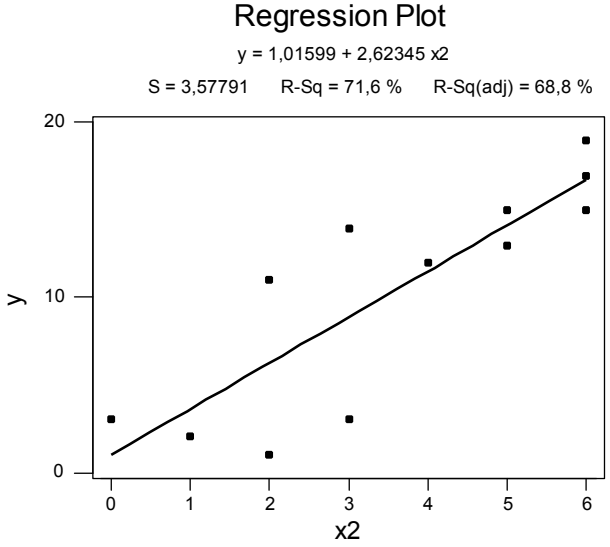
- Parte 5 – Modelo sem Intercepto e Variáveis *Dummy*

1) Considere o conjunto de dados da Tabela A.6 no Anexo.

Para se estudar a influência das variáveis “capital investido” e “gasto em publicidade” no lucro anual de empresas, foram observadas essas variáveis em doze empresas em um mesmo ano. Os seguintes resultados foram registrados, na unidade de 100 mil reais.

Variáveis:
- Y – Lucro anual
- X1 – Capital
- X2 – Publicidade

a) Ajuste o modelo de regressão $Y = \beta_0 + \beta_2 X_2 + \varepsilon$.



b) Construa a Tabela de Análise de Variância, calcule o valor de R^2, faça o teste de falta de ajuste (se possível)[2].

R^2= 71,6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	322,90	322,90	25,22	0,001
Residual Error	10	128,01	**12,80**		
Lack of Fit	5	7,51	1,50	0,06	0,996
Pure Error	5	120,50	24,10		
Total	11	450,92			

Teste de Falta de Ajuste:

H_o: Não há falta de ajuste
H_a: Há falta de ajuste

Como o P-valor da falta de ajuste é maior que 0,05 pode-se dizer que o modelo não apresenta falta de ajuste.

c) Caso não haja problemas com o teste da falta de ajuste , faça o teste F da regressão (escreva hipóteses nula e alternativa, faça o teste e conclua).

H_o: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$
Sendo a probabilidade de significância da regressão maior que 0,05, é possível afirmar que β_1 não é zero, isto é, o modelo de regressão ajustado é razoável.

d) Teste a significância do intercepto do modelo (teste *t*-Student ou intervalo de confiança. Escreva hipóteses nula e alternativa, faça o teste e conclua).

H_o: $\beta_0 = 0$
H_a: $\beta_0 \neq 0$

Estatística t-student = 0,48
Região Crítica = {$t \in R: t \geq 2,228$ ou $t \leq -2,228$}

À 5% de significância, pode-se afirmar que o intercepto do modelo é igual zero, ou seja, o mesmo não é importante para o modelo.

e) Ajuste o modelo de regressão sem o intercepto. $Y=\beta_2 X_2+\varepsilon$.

The regression equation is:
y = 2,84 x2

[2] Por questões didáticas, estamos omitindo a etapa de análise dos resíduos, que viria antes da utilização de qualquer teste.

f) Note que o MINITAB não calcula o R^2 para o modelo sem intercepto. Use então o valor do MSResidual para escolher entre os dois modelos (com intercepto e sem intercepto).

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1622,1	1622,1	136,30	0,000
Residual Error	11	130,9	**11,9**		
Total	12	1753,0			

Nota-se que o MSResidual do modelo sem intercepto é menor que este mesmo valor para o modelo com intercepto. Isto mostra que realmente foi melhor, neste caso, retirar β_0 do modelo.

2) Variáveis *Dummy*

Suponha que desejássemos estudar a renda (em R$) dos empregados de certo setor em função de sua experiência no cargo em que ocupa (anos) e de seu local de trabalho. No exemplo utilizado em sala, lidamos com 4 cidades (A, B , C e D) e as variáveis *dummies* criadas foram :

	Local 1	**Local 2**	**Local 3**
Cidade A	0	0	0
Cidade B	1	0	0
Cidade C	0	1	0
Cidade D	0	0	1

a) Suponha que exista uma quinta cidade (Cidade E). Como ficaria a tabela de codificação das cidades com a introdução da Cidade E?

	Local 1	**Local 2**	**Local 3**	***Local 4***
Cidade A	0	0	0	*0*
Cidade B	1	0	0	*0*
Cidade C	0	1	0	*0*
Cidade D	0	0	1	*0*
Cidade E	*0*	*0*	*0*	*1*

b) Considere agora a seguinte codificação:

	Local 1	**Local 2**	**Local 3**
Cidade A	0	0	1
Cidade B	0	1	0
Cidade C	1	0	0
Cidade D	0	0	0

o modelo :

Salário = β_0 + β_1 experiência + β_{21} "local1" + β_{22} "local2" + β_{23} "local3" + erro

e seguinte equação de regressão estimada :

Salário = 2,50 + 0,099 experiência + 0,55 "local1" + 0,69 "local2" + 0,75 "local3"

Considerando a mesma experiência, qual é a diferença média entre os salários das pessoas da:

b.1) cidade A e B = *0,06* *(0,75 – 0,69) = 0,06 (R$6,00 a mais)*
b.2) cidade A e C = *0,20* *(0,75 – 0,55) = 0,20 (R$20,00 a mais)*
b.3) cidade A e D = *0,75* *(0,75 – 0,00) = 0,75 (R$75,00 a mais)*
b.4) cidade B e C = *0,14* *(0,69 – 0,55) = 0,14 (R$14,00 a mais)*
b.5) cidade B e D = *0,69* *(0,69 – 0,00) = 0,69 (R$69,00 a mais)*
b.6) cidade C e D = *0,55* *(0,55 – 0,00) = 0,55 (R$55,00 a mais)*

c) Considere a primeira codificação. Suponha que, ao fazermos o teste t-Student para os parâmetros do modelo:

A categoria de referência é a cidade A .
O parâmetro β_{21} refere à cidade B (local1).
O parâmetro β_{22} refere à cidade C (local2).
O parâmetro β_{23} refere à cidade D (local3).

c.1) a hipótese $\beta_{21} = 0$ não seja rejeitada. O que isto significa em termos da comparação entre as cidades?

Significa que uma pessoa que mora na cidade B tem o mesmo salário de uma que mora na cidade A, com o mesmo tempo de experiência.

c.2) a hipótese $\beta_{22} = 0$ não seja rejeitada. O que isto significa em termos da comparação entre as cidades?

Significa que as pessoas, com o mesmo tempo de experiência, que residem nas cidades A e C ganham o mesmo salário.

c.3) a hipótese $\beta_{23} = 0$ não seja rejeitada. O que isto significa em termos da comparação entre as cidades?

Indivíduos que residem nas cidades A e D e que possuem o mesmo tempo de experiência tem salários iguais.

d) Pense na primeira tabela de codificação (local 1, local 2 e local 3). Para representar a cidade E, uma alternativa à resposta em a) seria fazer "local 1" = 1 ; "local 2" = 1 e "local 3" =1 . Considerando os testes de hipóteses para os parâmetros descritos em c) , pense em por que este procedimento não pode ser adotado (pense na comparação entre as cidades quando apenas um parâmetro não for considerado significativo)

	β_{21}	β_{22}	β_{23}
	Local 1	***Local 2***	***Local 3***
Cidade A	*0*	*0*	*0*
Cidade B	*1*	*0*	*0*
Cidade C	*0*	*1*	*0*
Cidade D	*0*	*0*	*1*
Cidade E	*1*	*1*	*1*

Por que não conseguimos comparar as cidades A e E. E ainda cada variável Local representa duas cidades, a cidade E e alguma outra.

❖ *Exercícios de Revisão de Regressão Linear Simples*

Considere o modelo de regressão linear simples, $Y = \beta_0 + \beta_1 X + \varepsilon$.

1) Qual é a variável dependente? E qual é a variável independente? Que outros nomes são usados para se referir a estas variáveis?

- *Variável independente ou variável resposta = Y*
- *Variável dependente ou explicativa ou preditora = X*

2) Qual é o método utilizado para estimar β0 e β1? Para utilizar esse método é necessário supor alguma distribuição para a variável resposta Y? Em caso positivo, qual é a distribuição?

O método utilizado para estimar β_0 e β_1 é chamado de métodos dos mínimos quadrados. Na verdade, para se usar o método de mínimos quadrados não é necessário supor distribuição para Y. A distribuição é necessária quando queremos fazer testes e construir intervalos.

3) Quais as suposições feitas pelo modelo de erros normais? O que estas suposições acarretam para Y?

Ë necessário supor que os erros são independentes, aleatórios e normalmente distribuídos com média zero e variância σ^2. Isto implica que os Y_i's tenham distribuição normal com médias $\beta_0 + \beta_1 X_i$ e variância constante σ^2.

4) O que significa "fazer extrapolação" no contexto de um modelo de regressão linear simples? Cite pelo menos dois riscos desta prática.

Fazer extrapolação significa inferir acerca de valores de X não contidos na amostra usada para ajustar o modelo de regressão. Ao se fazer extrapolação pode acontecer do valor estudado estar muito afastado dos valores da amostra e, desta maneira, ser descrito por outro modelo, isto é, ter outro comportamento diferente dos dados da amostra. Acontece também que a variância do valor predito fica grande à medida que nos afastamento do valor médio de X, ficando o intervalo de confiança muito largo e sem utilidade prática.

5) Defina o coeficiente de determinação (R^2) e explique quais valores ele pode assumir.

$R^2 = \frac{SQReg}{SQT}$; $0 \leq R^2 \leq 1$

O coeficiente de determinação representa a porcentagem da variabilidade de Y que é explicada pelo modelo de regressão ajustado.
Em caso de existência de medidas repetidas, o valor máximo de R^2 é 1 – (SSErroPuro/SQT).

6) Em que situação é possível realizar um teste de falta de ajuste ("Lack-of-fit") e qual é o objetivo deste teste?

É possível realizar o teste de falta de ajuste quando existem medidas de X repetidas. Este teste nos permite verificar se a reta de regressão ajustada se "ajusta" aos dados, ou seja, se o modelo é bom.

7) Quais os procedimentos gráficos podem ser usados para verificar as suposições enumeradas no item (2)?

- *o gráfico de probabilidade normal (p/ os erros) – para a verificação de normalidade dos resíduos (e assim dos Y_i's)*
- *Gráfico dos resíduos vs. a ordem (tempo) de coleta, quando disponível – para se constatar a aleatoriedade dos erros ;*
- *Gráfico dos resíduos vs. variável explicativa – para verificar suposição de variância constante (homocedasticidade) e aleatoriedade dos resíduos;*
- *Gráfico dos resíduos vs. Preditos – para verificar suposição de variância constante (homocedasticidade) e aleatoriedade dos resíduos;*

8) Em que situação podemos utilizar um teste para a suposição de não auto-correlação entre os erros? Cito dois possíveis testes a serem usados nesta situação.

Quando a ordem de coleta está disponível utiliza-se os seguintes testes:

- *Teste de Durbin-Watson*
- *Teste de corridas*

9) Quando é indicado o uso de transformação da variável resposta?

A transformação é necessária nos casos em os erros não possuem variância constante e/ou não são normalmente distribuídos. E ainda quando a relação entre X e Y não é linear.

10) Que tipo de transformação é feita na variável resposta no método analítico de Box-Cox? Exemplifique.

A transformação é a seguinte:

$$Y^{\lambda} = \begin{cases} Y^{\lambda} - 1 \Big/ \lambda \dot{Y}^{\lambda-1} & se \quad \lambda \neq 0 \\ \dot{Y} \ln Y & se \quad \lambda = 0 \end{cases}$$

Caso o valor de λ seja igual a ½, por exemplo, a transformação será $\sqrt{Y}$.

11) Em que situação é usada a regressão inversa?

A regressão inversa é feita quando surge a necessidade (por algum motivo) de se estimar valores para X a partir de em Y conhecido, além de saber os possíveis valores de

uma variável Y a partir dos valores de X.

12) Por que o teste F da tabela ANOVA é equivalente ao teste t-student para as hipóteses Ho: $\beta 1 = 0$ contra Ha: $\beta 1 \neq 0$? (Mostre a equivalência entre as duas estatísticas de teste)

$t = \dfrac{\hat{\beta}_1}{\sqrt{QMR / Sxx}} = \hat{\beta}_1 \sqrt{\dfrac{Sxx}{QMR}} = \sqrt{\dfrac{\hat{\beta}_1^2 Sxx}{QMR}} = \sqrt{\dfrac{SQ\,\mathrm{Re}\,g}{QMR}}$ *Elevando-se ambos lados ao quadrado temos:* $t^2 = \dfrac{SQ\,\mathrm{Re}\,g}{QMR} = \dfrac{QM\,\mathrm{Re}\,g}{QMR} = F$ *, sendo que uma variável que possui distribuição t-student com n graus de liberdade, quando elevada ao quadrado, passa a ter distribuição F com 1 grau de liberdade no numerador e n no denominador.*

13) Na análise de resíduos, porque utilizamos o gráfico "resíduos" x "valores ajustados" e não o gráfico dos "resíduos"x "valores observados"?

Porque a correlação entre os resíduos e os valores ajustados para Y é zero, mas existe correlação entre os resíduos e os valores observados para Y, mesmo que o modelo esteja bem ajustado.
Assim, se o modelo foi bem ajustado, não podemos observar padrões no gráfico resíduos vs valores ajustados.

❖ *Regressão Múltipla*

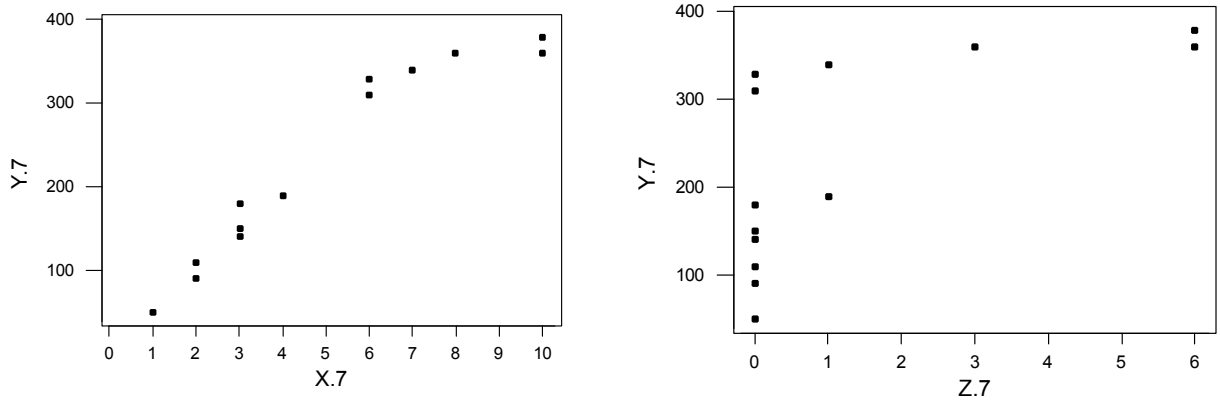
- Parte 1

1) **(Adaptação dos exercícios 3.LL e 6.H, Draper and Smith)** O gerente de um pequeno serviço de entregas contrata pessoal adicional sempre que o volume de serviço excede a carga de trabalho de seus usuais três empregados. Para verificar a eficácia desta idéia, ele registrou durante 13 dias seguidos as seguintes variáveis:

Variável Resposta: Y - Número de Entregas ;
Variáveis Explicativas: X - Número de Empregados (atuais mais extras) ;
Z - Número de Empregados que não estavam trabalhando em algum período do dia;

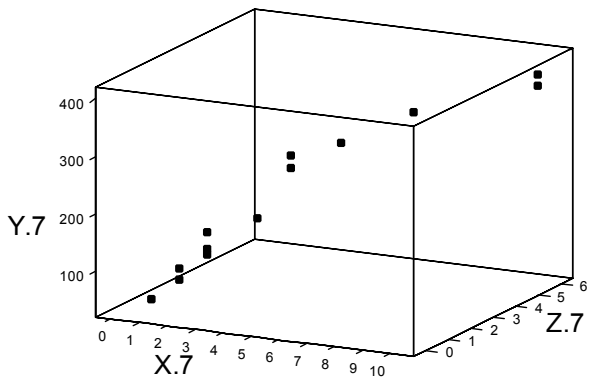
Os dados coletados estão disponíveis na Tabela A.7 em Anexo.
Obs: nos três primeiros dias de coleta, alguns dos empregados usuais estavam de férias ou de licença médica.

a) Faça o diagrama de dispersão de Y versus X, Y versus Z e avalie a possibilidade do ajuste de um modelo de regressão linear.



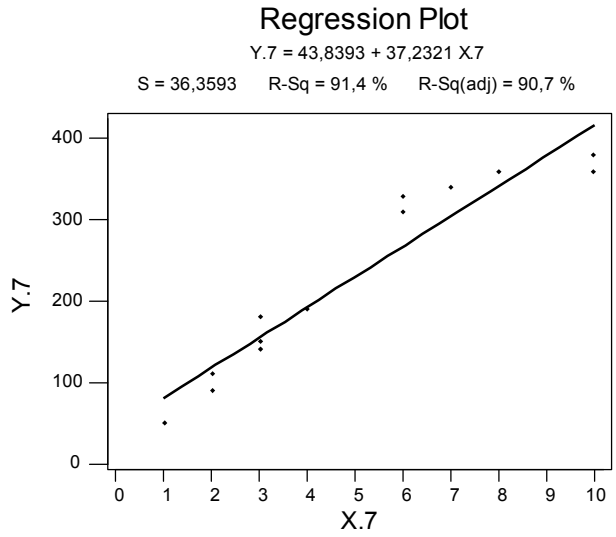
Ao se analisar os gráficos acima vê-se que há um relacionamento claro entre as variáveis Y e X, o que não o corre com a variável Z. Por isso o ajuste de um modelo de regressão linear seria mais aconselhável para as variáveis Y e X.

b) Faça o gráfico em 3 dimensões de Y versus X e Z. (MINITAB: Graph > 3-D plot)



Neste gráfico vê-se que quando se analisa as três variáveis juntas o relacionamento entre elas fica evidente.

c) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.

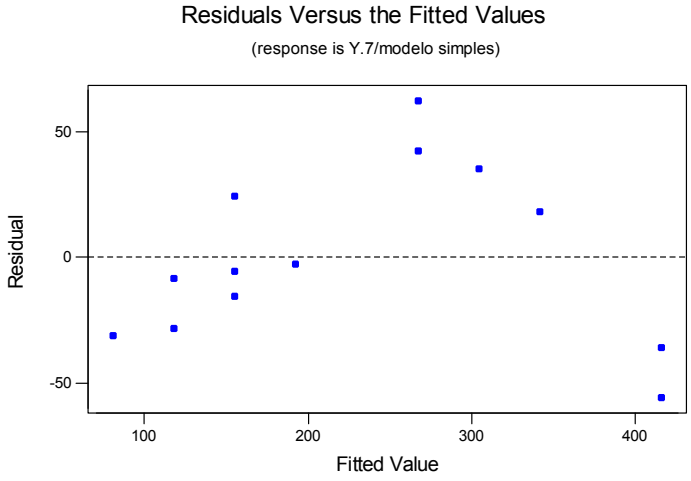


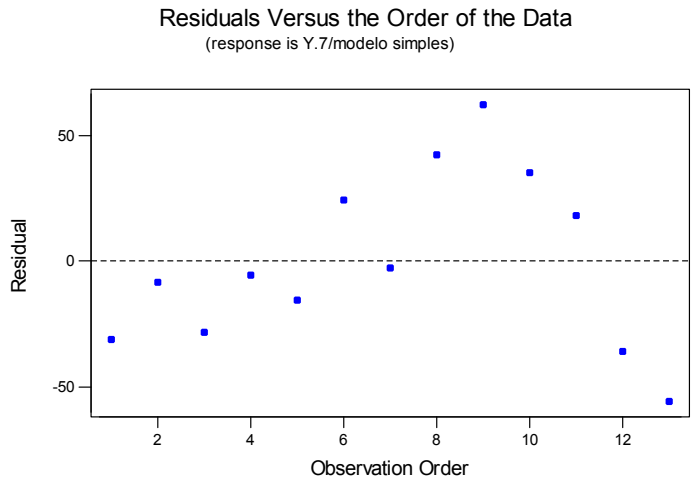
d) Construa a Tabela de Análise de Variância.

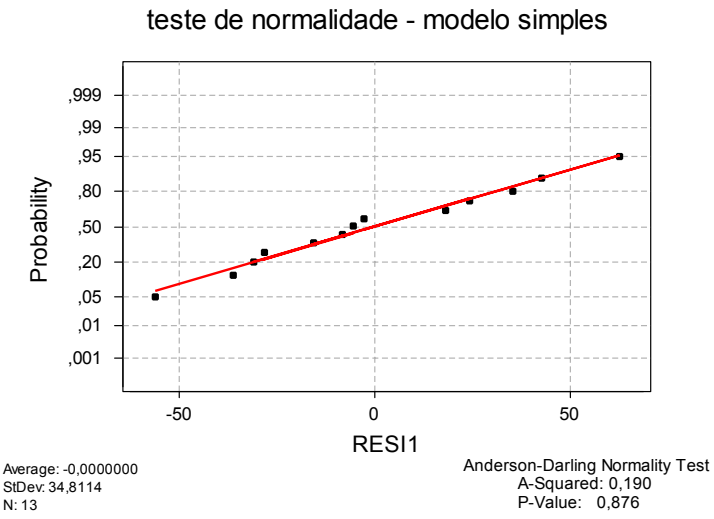
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	155258	155258	117,44	0,000
Residual Error	11	14542	1322		
Lack of Fit	6	13075	2179	7,43	0,022
Pure Error	5	1467	293		
Total	12	169800			

e) Faça Análise dos Resíduos (considere o dia como ordem de coleta e faça também o gráfico dos resíduos versus a variável Z). Se existem problemas com as suposições do modelo de erros normais, quais são eles?







<u>Teste de Durbin-Watson</u>

H_o: Os resíduos não são correlacionados
H_a: Os resíduos são correlacionados

D = 0,74 4 – D = 3,26 (Como D é mais próximo de zero, trabalhar com D)

dl = 0.95 du = 1.23

Como D < dl, há evidências de correlação serial positiva, como pode ser visualizado no gráfico de resíduos versus ordem de coleta.

<u>Teste de homogeneidade</u>

H_o: Os resíduos têm variância constante.
H_a: Os resíduos não têm variância constante.

```
Bartlett's Test
Test Statistic: 0,270
P-Value       : 0,966

Levene's Test
Test Statistic: 0,079
P-Value       : 0,968
```

Em ambos os testes a hipótese de variância dos resíduos constante não foi rejeitada.

Através das análises dos gráficos acima vê-se que os resíduos apresentam correlação e também não parecem ser aleatórios. Entretanto os mesmos apresentam distribuição normal (teste de normalidade) e variância constante, pelo teste de homogeneidade.

f) Caso não haja problemas com as suposições do modelo de erros normais, faça os testes F (Falta de Ajuste e Regressão) da Tabela Anova em (d).

Há problemas : padrão não esperado no gráfico resíduos versus ajustados e no gráfico resíduos versus ordem de coleta.
Não fazer testes F .

g) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, encontrando a equação estimada.

```
The regression equation is
Y.7 = - 5,95 + 54,4 X.7 - 27,4 Z.7
```

h) Construa a Tabela de Análise de Variância, separando as SS seqüenciais.

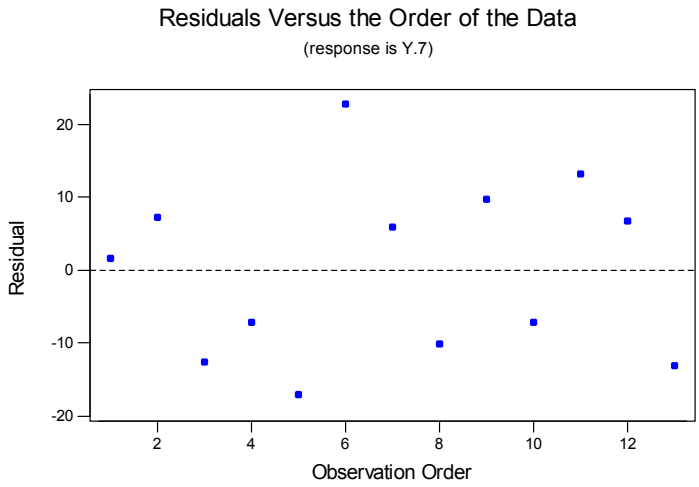
Analysis of Variance

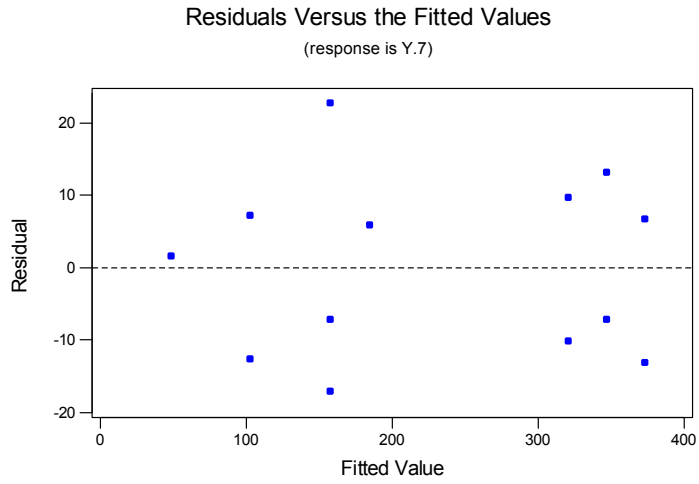
Source	DF	SS	MS	F	P
Regression	2	168031	84016	474,98	0,000
X.7	**1**	**155258**			
Z.7	**1**	**12773**			
Residual Error	10	1769	177		
Lack of Fit	5	302	60	0,21	0,946
Pure Error	5	1467	293		
Total	12	169800			

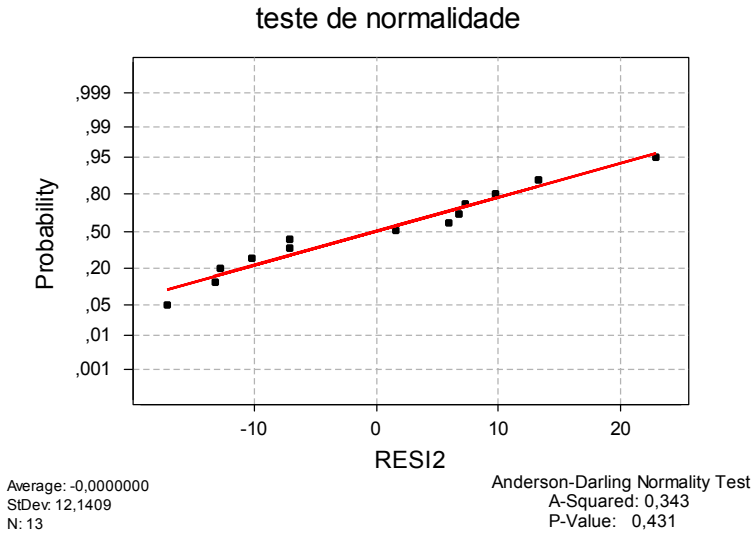
OBS.:

SS(Z.7) = SSReg – SS(X.7) = 168031 – 155258 = 12773

i) Faça Análise dos Resíduos do modelo em (g) . Há algum problema?







Teste de Durbin-Watson

H_o: Os resíduos não são correlacionados
H_a: Os resíduos são correlacionados

D = 2,41 4 – D = 1,59
dl = 0,83 du = 1,40

Como ambos D e 4-D são maiores que du pode-se afirmar que os resíduos não são correlacionados.

Analisando-se os gráficos acima nota-se que os resíduos não são correlacionados, possuem variância constante e são aleatórios. E ainda, através do teste de Anderson-Darling foi verificado que os resíduos não normalmente distribuídos.

j) Caso não haja problemas em (i), faça o teste da Falta de Ajuste da Tabela Anova em (h).

Teste de Falta de Ajuste:
H_o: Não há falta de ajuste
H_a: Há falta de ajuste

É possível afirmar que o modelo de regressão ajustado não apresenta falta de ajuste, pois o valor P da falta de ajuste mostrado na tabela de análise de variância é maior que 0,05 (0,946).

k) Caso não haja problemas no teste de falta de ajuste, faça os testes F seqüenciais da regressão (escreva as hipóteses nula e alternativa de cada teste).
H_o: A contribuição de β_1, dado β_0, não é significante ($\beta_1 = 0$)
H_a: A contribuição de β_1, dado β_0, é significante ($\beta_1 \neq 0$)

Estatística F = $\dfrac{SQ\,\mathrm{Re}\,g(X_1)/1}{QMR(X_1)}$ *= 155258 /1322 = 117,44*

Região Crítica = {F : F > $F_{1;11;0,05}$} , onde $F_{1;11;0,05}$ = 4,8443

H_o: A contribuição de β_2, dado β_1 e β_0, não é significante ($\beta_2 = 0$)
H_a: A contribuição de β_2, dado β_1 e β_0, é significante ($\beta_2 \neq 0$)

Estatística F = $\dfrac{SQ\,\mathrm{Re}\,g(X_2 \mid X_1)/1}{QMR(X_1X_2)}$ *= 12773 /177 = 72,164*

Região Crítica = {F : F > $F_{1;10;0,05}$} , onde $F_{1;10;0,05}$ = 4,9646

Em ambos os testes os valores de F estão na região crítica o que significa que os dois parâmetros são significativos.

l) Utilizando o teste t-Student, teste a significância de cada parâmetro individualmente. Os resultados concordam com os resultados dos testes F seqüenciais de (k)?

H_o: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

Estatística t = 22,89
Região Crítica = {t : t ≤ -2,201 ou t ≥ 2,201}

H_o: $\beta_2 = 0$
H_a: $\beta_2 \neq 0$

Estatística t = -8,50
Região Crítica = {t : t ≤ -2,201 ou t ≥ 2,201}

Nos dois testes a hipótese nula foi rejeitada, o que quer dizer que os dois parâmetros são importantes para o modelo.

m) Interprete a equação de regressão estimada em (g).

Para um número fixo de empregados que não estavam trabalhando em algum período do dia, a cada aumento de uma unidade no número de empregados há um aumento de 54,4 no numero de entregas. Já para um numero de empregados fixo, o número de entregas decresce de 27,4 a cada uma unidade aumentada no número de empregados que não estavam trabalhando em algum período do dia.

n) Intervalo de Confiança para E[Y] dadas novas observações de X e Z : a matriz $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ pode ser armazenada no MINITAB (na janela Regression, botão Storage, marque a opção X'X inverse). Esta matriz será armazenada num objeto chamado m1. Para imprimir este objeto na janela Session , basta ir no menu Edit > Command Line Editor , digitar print m1 e pressionar Submit Commands . Esta é a matriz que será usada no cálculo do erro de estimação no intervalo de confiança para E[Y|(x,z)].

Considerando um número de empregados (X) igual a 5 e todos eles trabalhando todo o tempo (ou seja, Z = 0), construa um intervalo de 95% de confiança para E[Y], o número médio de entregas realizadas quando há 5 empregados trabalhando todo o tempo .

$(X'X)^{-1}$ =

0,494189	-0,11138	0,10678
-0,11138	0,031881	-0,03672
0,10678	-0,03672	0,058757

*$QMR[x_0'(X'X)^{-1}x_0]$ = 177 * 0,174 = 30,798*
$t_{\alpha/2\ ;\ (n-p-1)}$ = 2,201

Logo,

$IC_{95\%} = (\hat{Y} \pm t_{\alpha/2,(n-p-1)}\sqrt{QMR[x_0'(X'X)^{-1}x_0]}) =$ *(253,83 ; 278,26)*

O número médio de entregas realizadas quando há 5 empregados trabalhando todo o tempo está entre 253 e 278 casos, com 95% de confiança.

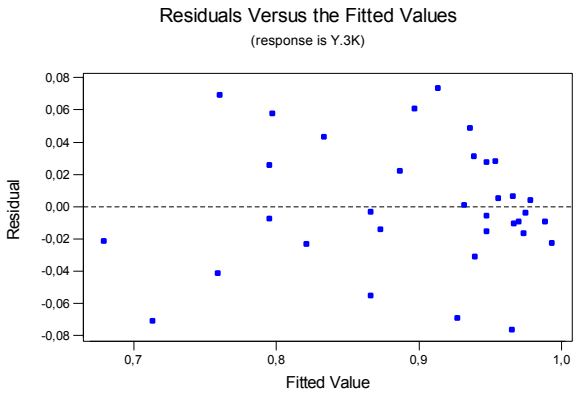
- Parte 2 – Detecção de Pontos de Influência

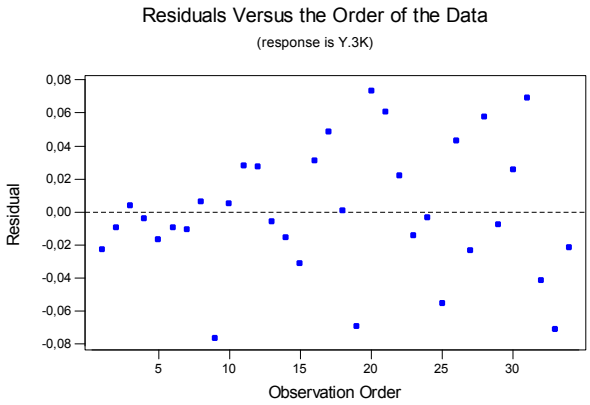
1) **Detectando pontos de influência -** Considere os seguintes exercícios das listas anteriores : 2 - parte 1; 1 – parte2; 1 – parte 3; 1 – parte 4; 2 – parte 5 e 1 – parte 6.

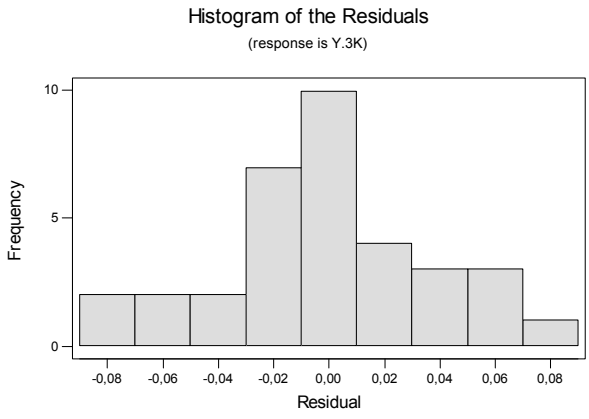
a) Faça a análise de resíduos à procura de pontos de influência. Use as medidas Hi, D-cook, *resíduos studentizados*.

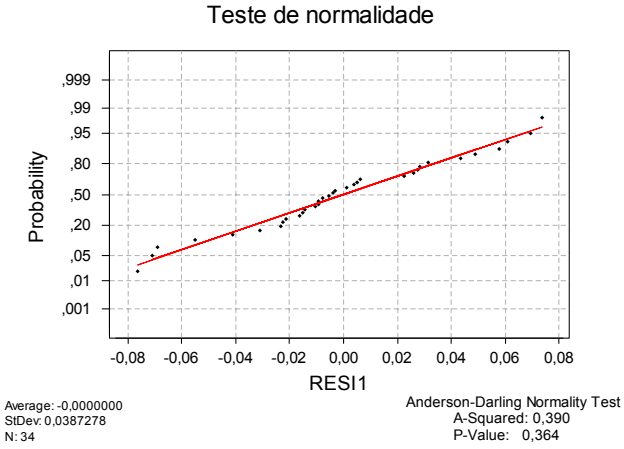
b) Caso seja(m) detectado(s) ponto(s) de influência, ajuste o modelo sem este(s) ponto(s) e compare sua equação estimada com a equação estimada com todos os pontos para verificar o tamanho da influência deste(s) ponto(s).

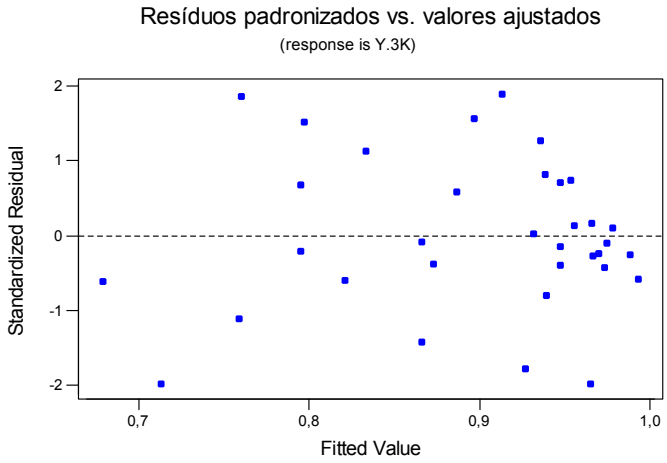
2 – parte 1)
Análise de resíduos







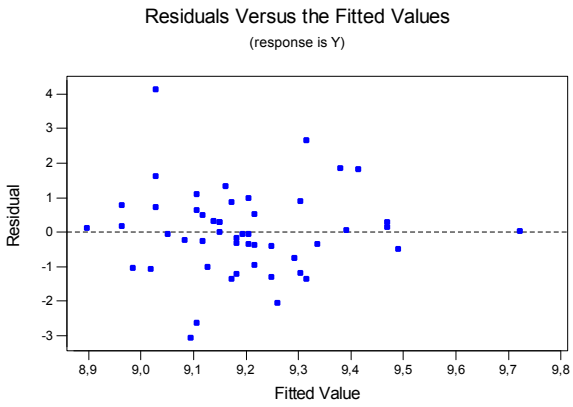


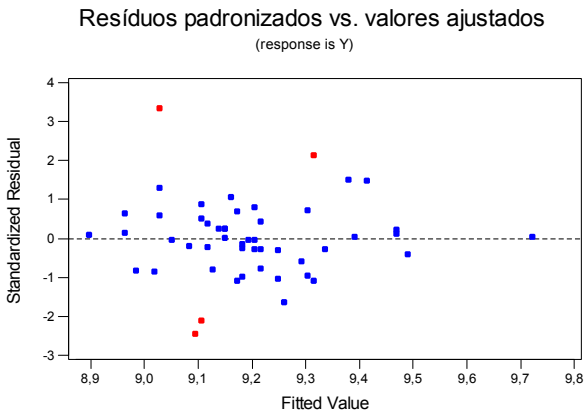


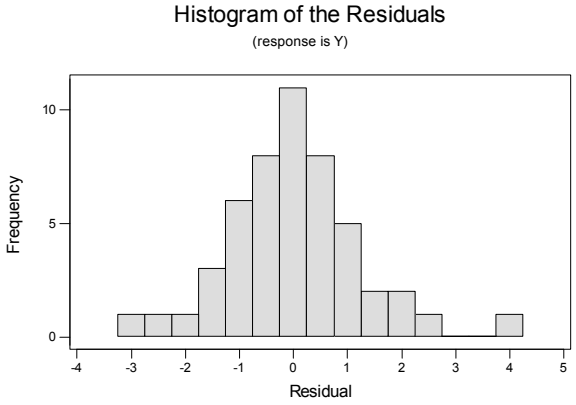
Obs	SRES1	HI1	COOK1	Obs	SRES1	HI1	COOK1
1	-0,59010	0,069216	0,012947	18	0,03014	0,034835	0,000016
2	-0,24864	0,065291	0,002159	19	-1,78955	0,033526	0,055545
3	0,10474	0,057650	0,000336	20	1,90461	0,030682	0,057412
4	-0,10723	0,055690	0,000339	21	1,57455	0,029417	0,037571
5	-0,42771	0,054548	0,005277	23	0,57977	0,029772	0,005157
6	-0,23933	0,052519	0,001588	24	-0,37399	0,031500	0,002275
7	-0,27101	0,050232	0,001942	25	-0,08317	0,033077	0,000118
8	0,16156	0,049891	0,000685	26	-1,42780	0,033077	0,034869
9	-1,99583	0,049722	0,104211	27	1,13468	0,045702	0,030830
10	0,13176	0,044521	0,000404	28	-0,60640	0,052690	0,010226
11	0,73801	0,043376	0,012348	29	1,52642	0,070155	0,087896
12	0,71350	0,040585	0,010767	30	-0,20214	0,071363	0,001570
13	-0,14315	0,040585	0,000433	31	0,68410	0,071850	0,018114
14	-0,39518	0,040459	0,003292	**32**	**1,86819**	**0,105953**	**0,206807**
15	-0,80578	0,037241	0,012558	33	-1,11325	0,107272	0,074460
16	0,81585	0,037032	0,012799	**34**	**-1,98586**	**0,168868**	**0,400629**
17	1,27164	0,036128	0,030305	35	-0,61342	0,225576	0,054803

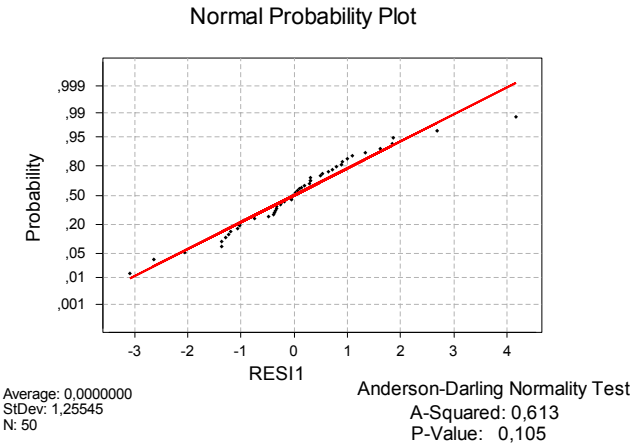
Pela análise da tabela acima percebe-se que as observações nº 32 e 34 possuem valores de COOKs um pouco maiores que as demais, porém os valores dos Hi's e dos resíduos studentizados não são muito discrepantes. Também pela análise gráfica dos resíduos vê-se que os pontos citados e nenhum outro ponto consistem num ponto influente.

1 – parte 2)
Análise de resíduos









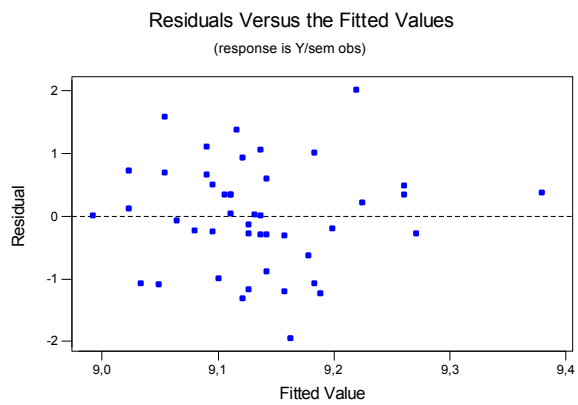
Obs	SRES1	HI1	COOK1	Obs	SRES1	HI1	COOK1
1	0,02585	0,252847	0,000113	26	-0,14597	0,020184	0,000219
2	-0,40635	0,092859	0,008451	27	-1,09315	0,020561	0,012543
3	0,10785	0,082337	0,000522	28	0,69917	0,020561	0,005131
4	0,23130	0,082337	0,002400	29	1,06671	0,021144	0,012289
5	**1,49259**	**0,059621**	**0,070624**	30	-0,00025	0,021931	0,000000
6	0,04693	0,051970	0,000060	31	0,23889	0,021931	0,000640
7	1,51044	0,048452	0,058084	32	0,23889	0,021931	0,000640
8	-0,27073	0,036430	0,001386	33	0,24778	0,022923	0,000720
9	-1,09365	0,031649	0,019546	34	-0,82065	0,024120	0,008323
10	**2,15094**	**0,031649**	**0,075605**	35	-0,21351	0,025523	0,000597
11	-0,96364	0,029566	0,014146	36	0,38545	0,025523	0,001946
12	0,71693	0,029566	0,007830	**37**	**-2,12316**	**0,027130**	**0,062854**
13	-0,59415	0,027688	0,005026	38	0,51444	0,027130	0,003690
14	-1,64340	0,023284	0,032192	39	0,87411	0,027130	0,010654
15	-1,03581	0,022226	0,012194	**40**	**-2,47635**	**0,028942**	**0,091387**
16	-0,31827	0,022226	0,001151	41	-0,18771	0,030960	0,000563
17	-0,76958	0,020282	0,006131	42	-0,04134	0,038242	0,000034

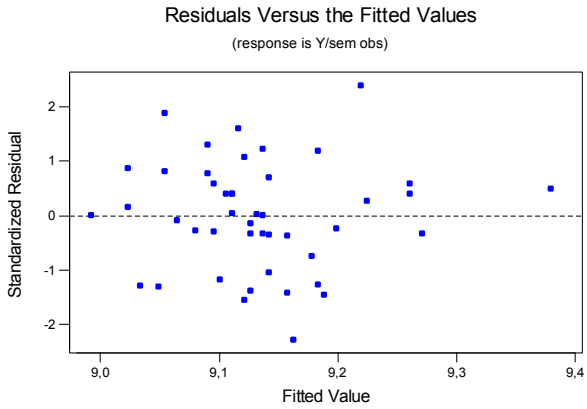
18	-0,29170	0,020282	0,000881	43	0,58101	0,044122	0,007791
19	0,42513	0,020282	0,001871	44	1,30672	0,044122	0,039408
20	-0,28291	0,020045	0,000819	**45**	**3,36289**	**0,044122**	**0,261005**
21	-0,04400	0,020045	0,000020	46	-0,86302	0,047370	0,018518
22	0,79219	0,020045	0,006418	47	-0,84125	0,058342	0,021924
23	-0,03525	0,020012	0,000013	48	0,15217	0,066682	0,000827
24	-0,98222	0,020184	0,009937	49	0,64178	0,066682	0,014714
25	-0,26543	0,020184	0,000726	50	0,08493	0,096623	0,000386

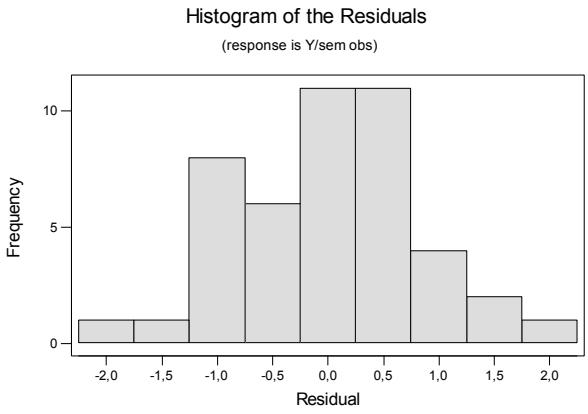
Através da análise dos gráficos nota-se que existem alguns pontos que podem estar influenciando o modelo. Esta suposição é confirmada pela tabela acima, que apresenta cinco pontos cujos valores dos COOKs se diferem das demais observações. Para avaliar se estes pontos realmente são pontos influentes vamos ajustar um modelo sem estes valores.

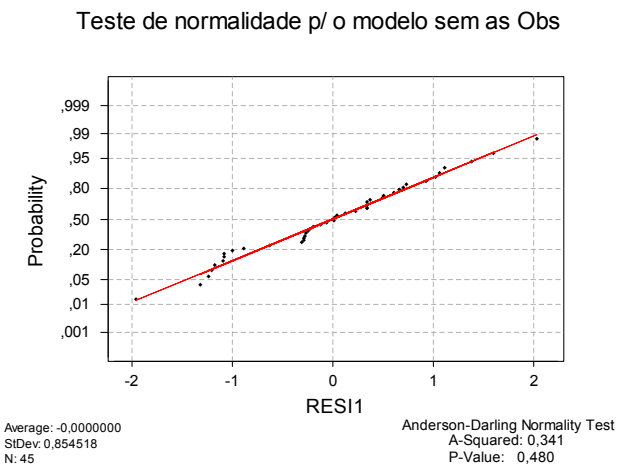
A equação estimada com todas as observações é: Y = 9,93042 - 0,0109873 X
Equação sem as observações influentes: Y = 9,48 - 0,00516 X

Com relação às retas estimadas a diferença entre elas não foi muito grande, entretanto, no que diz respeito aos resíduos, a melhora foi significativa, como pode ser verificado através dos gráficos abaixo:

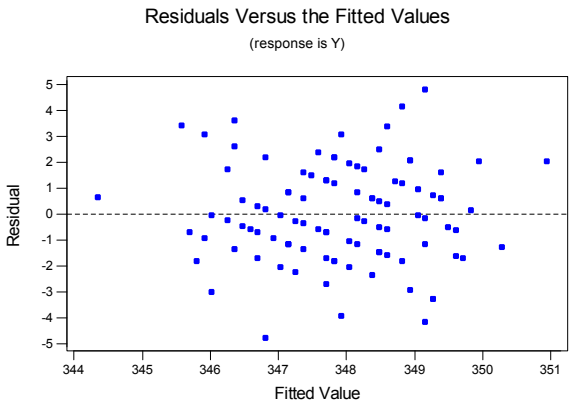


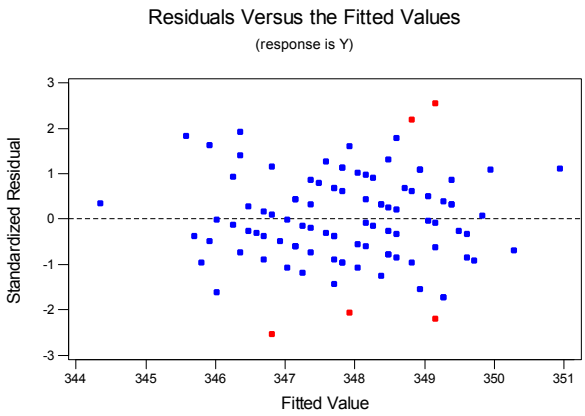


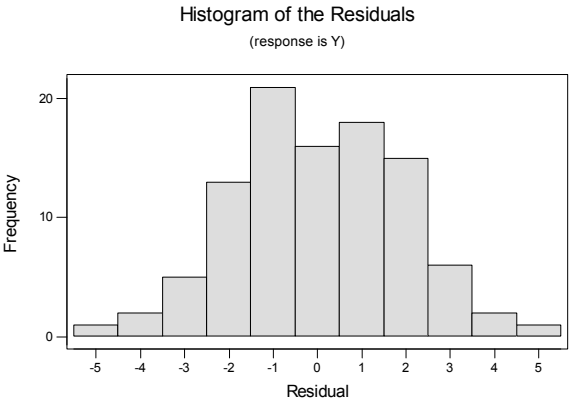


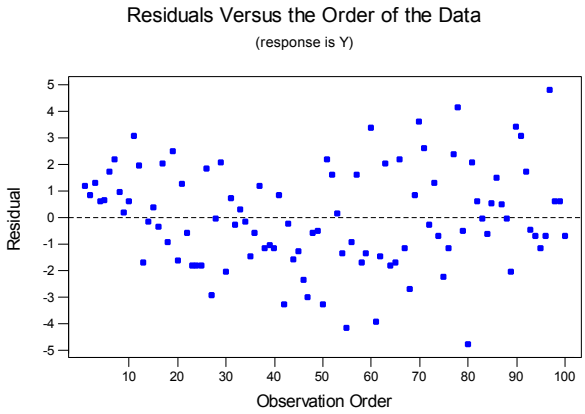


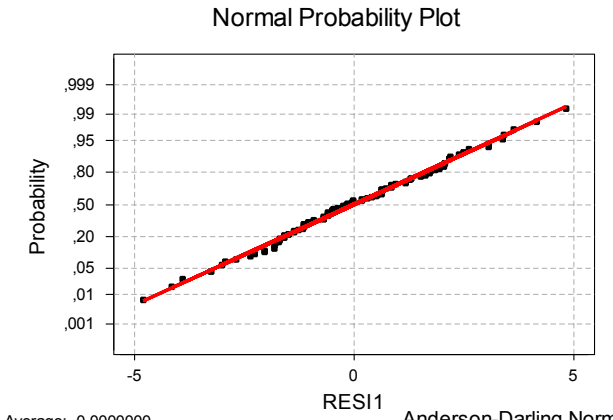
1 - parte 3)
Análise de Resíduos











Average: -0,0000000
StDev: 1,89185
N: 100

Anderson-Darling Normality Test
A-Squared: 0,218
P-Value: 0,836

Obs	SRES1	HI1	COOK1	Obs	SRES1	HI1	COOK1
1	0,62456	0,0100358	0,0019772	51	1,15313	0,0100358	0,0067399
2	0,45175	0,0138483	0,0014329	52	0,86043	0,0255790	0,0097170
3	0,68378	0,0102348	0,0024174	53	0,08969	0,0363043	0,0001515
4	0,32910	0,0116592	0,0006388	54	-0,72523	0,0118792	0,0031615
5	0,36009	0,0973777	0,0069942	55	-2,21204	0,0212638	0,0531533

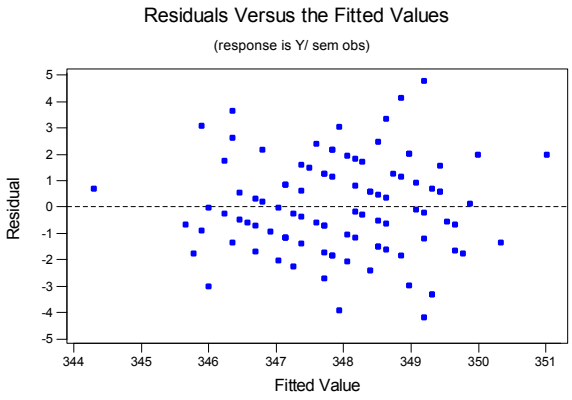
6	0,93289	0,0287072	0,0128608	56	-0,48957	0,0165156	0,0020125
7	1,16163	0,0181111	0,0124448	57	0,86195	0,0118792	0,0044659
8	0,50490	0,0193680	0,0025175	58	-0,92040	0,0333611	0,0146184
9	0,10016	0,0181111	0,0000925	59	-0,72677	0,0262389	0,0071164
10	0,33289	0,0118792	0,0006661	60	1,79941	0,0135305	0,0222057
11	1,65285	0,0371597	0,0527176	61	-2,07739	0,0100113	0,0218206
12	1,03490	0,0101614	0,0054974	62	-0,78843	0,0125076	0,0039368
13	-0,90274	0,0198812	0,0082653	63	1,10297	0,0394220	0,0249635
14	-0,08568	0,0212638	0,0000797	64	-0,96114	0,0100358	0,0046824
15	0,21091	0,0135305	0,0003051	65	-0,90208	0,0102348	0,0042073
16	-0,19617	0,0118792	0,0002313	66	1,15313	0,0100358	0,0067399
17	1,12031	0,0753371	0,0511292	67	-0,60743	0,0138483	0,0025907
18	-0,49099	0,0371597	0,0046519	68	-1,43070	0,0102348	0,0105831
19	1,32848	0,0125076	0,0111768	69	0,45175	0,0138483	0,0014329
20	-0,85932	0,0305925	0,0116516	70	1,93796	0,0262389	0,0506001
21	0,68157	0,0147281	0,0034720	71	1,40501	0,0262389	0,0265964
22	-0,31859	0,0135305	0,0006961	72	-0,13703	0,0127764	0,0001215
23	-0,96114	0,0100358	0,0046824	73	0,68378	0,0102348	0,0024174
24	-0,96857	0,0403263	0,0197103	74	-0,37346	0,0102348	0,0007211
25	-0,96786	0,0161001	0,0076643	75	-1,19563	0,0127764	0,0092503
26	0,97591	0,0104861	0,0050464	76	-0,60743	0,0138483	0,0025907
27	-1,55860	0,0176468	0,0218192	77	1,27179	0,0106083	0,0086712
28	-0,02617	0,0193680	0,0000068	78	2,21330	0,0161001	0,0400801
29	1,09445	0,0176468	0,0107586	79	-0,26505	0,0279985	0,0010118
30	-1,07950	0,0101614	0,0059814	**80**	**-2,55352**	**0,0181111**	**0,0601356**
31	0,38683	0,0233341	0,0017876	81	1,09445	0,0176468	0,0107586
32	-0,14066	0,0109854	0,0001099	82	0,32766	0,0255790	0,0014091
33	0,15969	0,0198812	0,0002586	83	-0,01859	0,0150946	0,0000026
34	-0,08146	0,0104861	0,0000352	84	-0,32517	0,0305925	0,0016684
35	-0,78843	0,0125076	0,0039368	85	0,27915	0,0239451	0,0009559
36	-0,31437	0,0106083	0,0005298	86	0,80246	0,0111565	0,0036325
37	0,62272	0,0161001	0,0031727	87	0,27002	0,0125076	0,0004617
38	-0,60743	0,0138483	0,0025907	88	-0,01498	0,0341676	0,0000040
39	-0,55090	0,0101614	0,0015578	89	-1,07844	0,0150946	0,0089123
40	-0,61727	0,0212638	0,0041390	**90**	**1,84239**	**0,0471832**	**0,0840447**
41	0,44722	0,0104861	0,0010598	91	1,62253	0,0100113	0,0133111
42	-1,74178	0,0233341	0,0362411	92	0,91698	0,0109854	0,0046699
43	-0,13436	0,0287072	0,0002668	93	-0,25316	0,0239451	0,0007862
44	-0,84809	0,0135305	0,0049327	94	-0,37152	0,0198812	0,0013999
45	-0,69068	0,0498226	0,0125068	95	-0,61015	0,0104861	0,0019726
46	-1,25790	0,0116592	0,0093331	96	-0,37346	0,0102348	0,0007211
47	-1,62036	0,0341676	0,0464417	97	2,57226	0,0212638	0,0718747
48	-0,31239	0,0218259	0,0010887	98	0,32766	0,0255790	0,0014091
49	-0,25921	0,0125076	0,0004255	99	0,32910	0,0116592	0,0006388
50	-1,74178	0,0233341	0,0362411	100	-0,37230	0,0436675	0,0031645

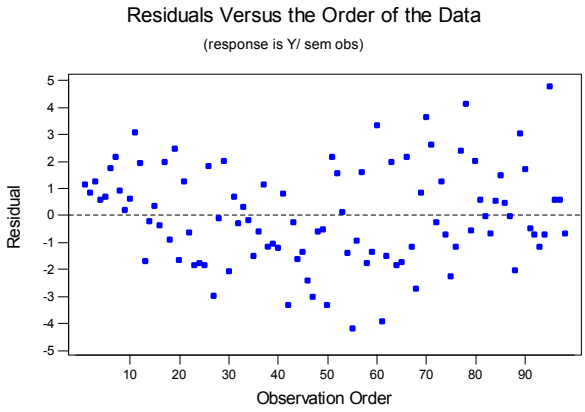
Apesar de existirem cinco pontos que, no gráfico dos resíduos padronizados vs. Valores ajustados, estão fora do intervalo (-2 ; 2), os mesmos parecem não serem pontos

de grande influencia no modelo. Apenas dois pontos de destacam um pouco dos demais, são eles: obs. nº 80 e nº 90.

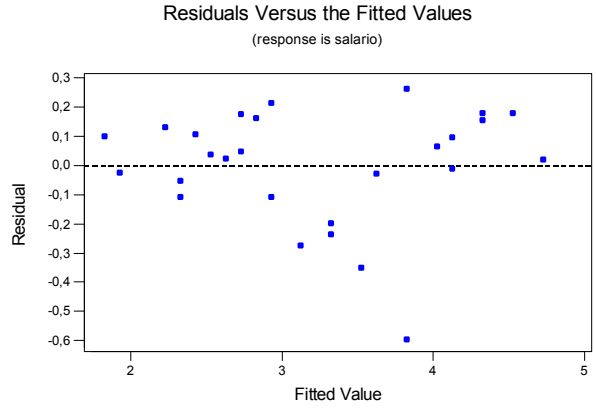
Ajustado o modelo com todas as observações temos: Y = 361,246 - 0,111900 X
A equação estimada sem os pontos influentes é: Y = 362 - 0,114 X

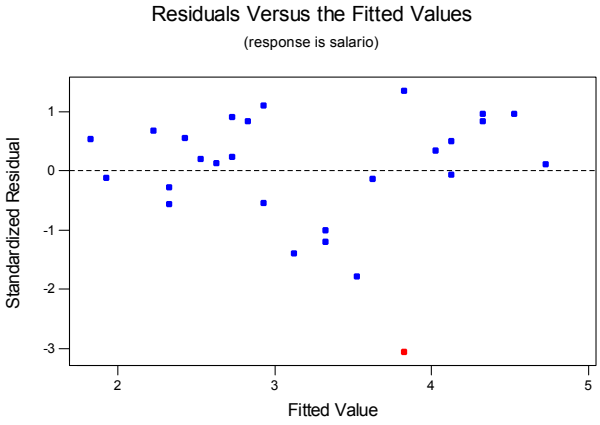
Não houveram diferenças significantes entres os dois modelos tanto quanto à reta estimada quanto aos resíduos.

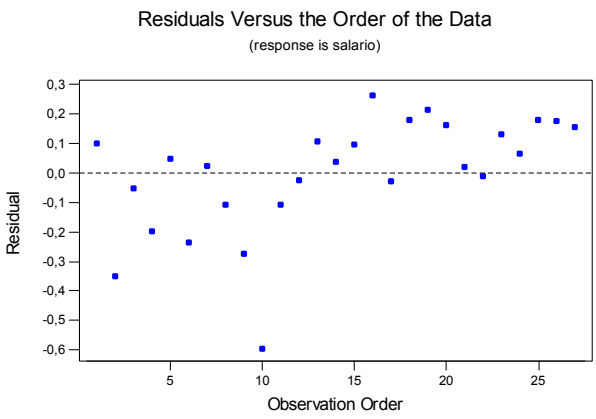


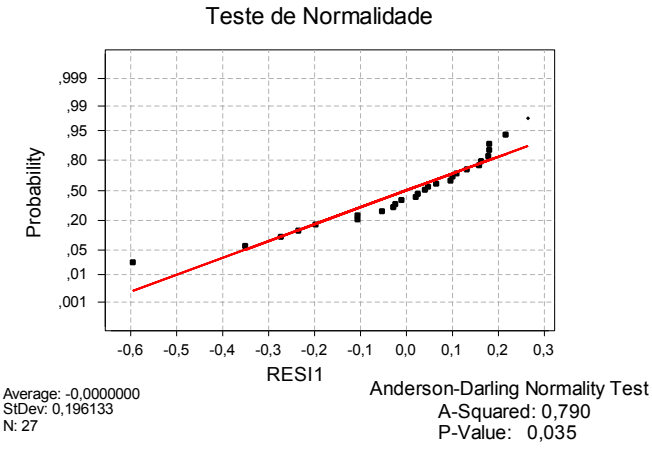


1 – parte 4)
Análise de Resíduos







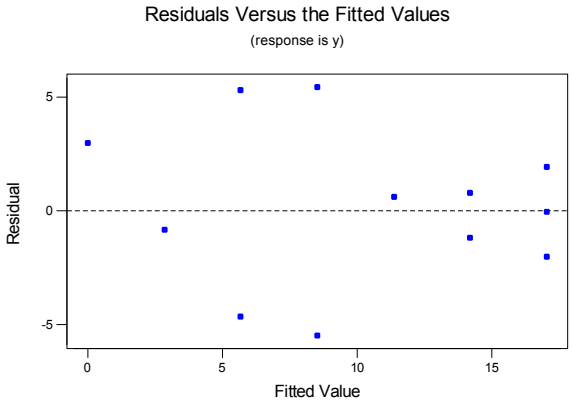


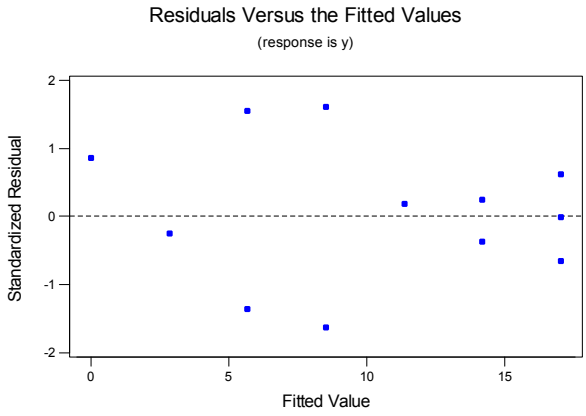
Obs	SRES1	HI1	COOK1	Obs	SRES1	HI1	COOK1
1	0,54059	0,144611	0,024703	15	0,50329	0,081494	0,011237
2	-1,79140	0,041977	0,070305	16	1,36539	0,056796	0,056130
3	-0,27590	0,081494	0,003377	17	-0,14039	0,045819	0,000473
4	-1,00536	0,037586	0,019737	18	0,97437	0,129792	0,070801
5	0,24548	0,050758	0,001611	19	1,11046	0,041977	0,027015
6	-1,20105	0,037586	0,028168	20	0,83641	0,045819	0,016796
7	0,12641	0,056796	0,000481	21	0,11277	0,160528	0,001216
8	-0,55708	0,081494	0,013767	22	-0,05854	0,081494	0,000152
9	-1,39911	0,037586	0,038224	23	0,69007	0,091922	0,024102
10	**-3,07003**	**0,056796**	**0,283768**	24	0,34047	0,072163	0,004508
11	-0,53991	0,041977	0,006386	25	0,95803	0,103448	0,052951
12	-0,12281	0,129792	0,001125	26	0,91616	0,050758	0,022441
13	0,56467	0,072163	0,012400	27	0,83658	0,103448	0,040377
14	0,20553	0,063931	0,001443				

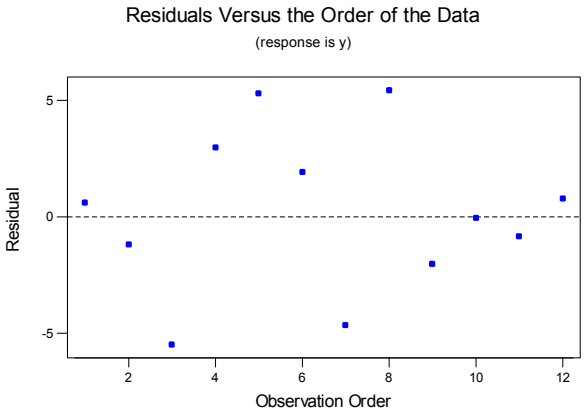
Nota-se que apenas um ponto, obs. 10, está muito afastada dos demais pontos (maior valor de resíduos studentizados) e pode estar influenciando o modelo de regressão (valor alto de COOK). Para verificar esta influencia é interessante ajustar um modelo sem esta observação.

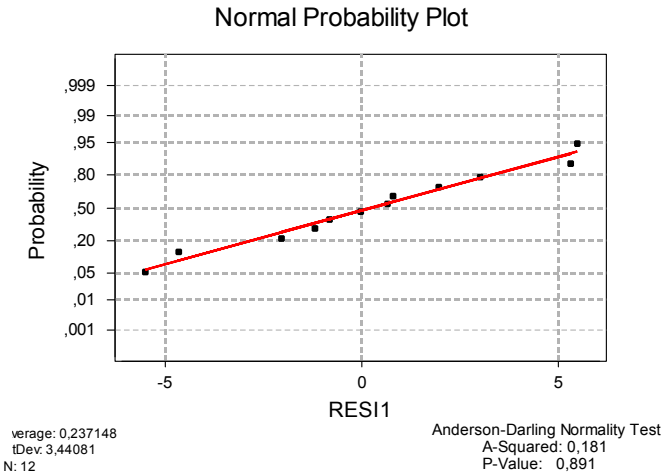
Equação estimada com todas observações: salário = 1,83070 + 0,0998186 experiência
Reta estimada sem a observação 10: salário = 1,82 + 0,102 experiência
Neste caso, assim como nos anteriores, não houve muitas diferenças entre os dois modelos (com e sem a obs. 10).

2 – parte 5)
Análise de Resíduos





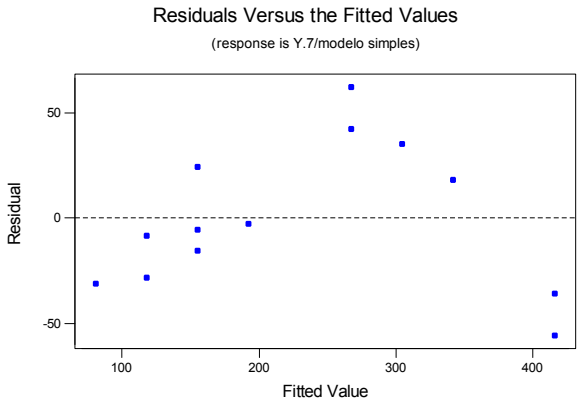


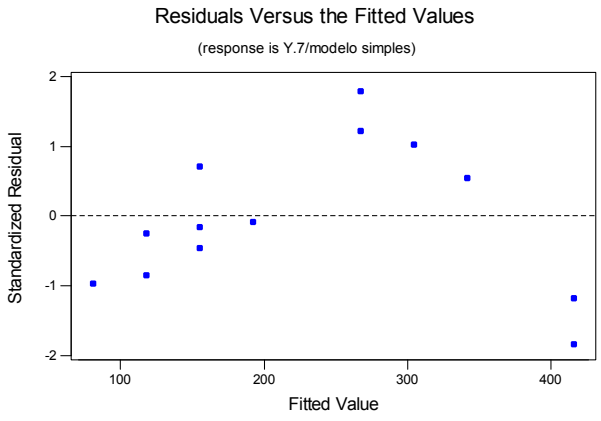


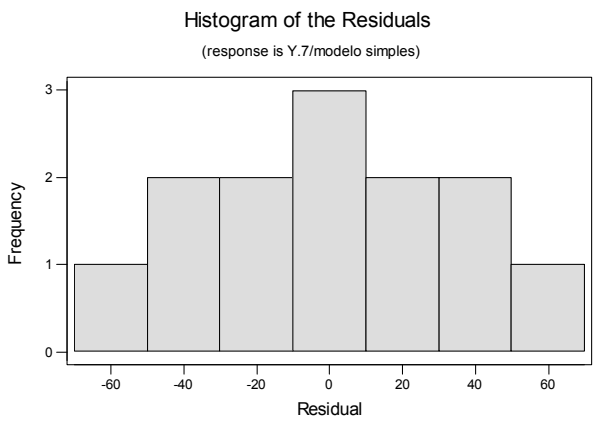
Obs	SRES2	HI2	COOK2	Obs	SRES2	HI2	COOK2
1	0,19242	0,079602	0,003202	7	-1,37081	0,019900	0,038154
2	-0,37297	0,124378	0,019760	8	1,62464	0,044776	0,123724
3	-1,63792	0,044776	0,125755	9	-0,65421	0,179104	0,093381
4	0,86964	0,000000	0,000000	10	-0,01433	0,179104	0,000045
5	1,55727	0,019900	0,049240	11	-0,24434	0,004975	0,000299
6	0,62556	0,179104	0,085380	12	0,24659	0,124378	0,008638

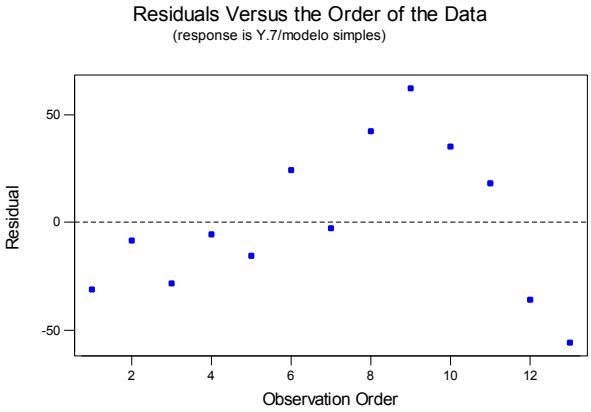
Tanto pela análise gráfica quanto analisando-se os valores da tabela acima vê-se que parece não existir nenhum ponto influente.

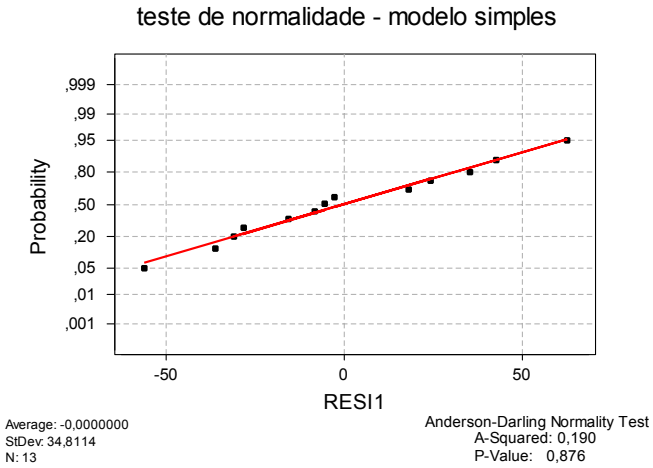
1 - parte 6)
Análise de Resíduos









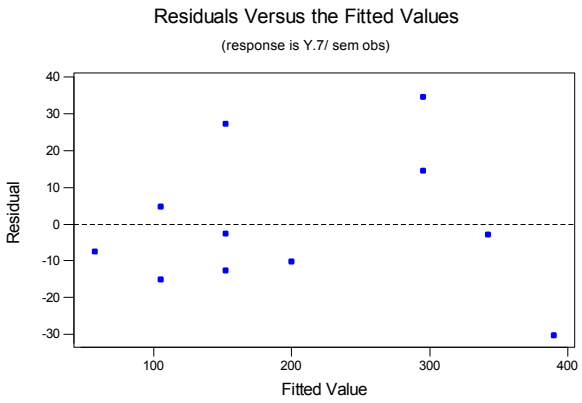


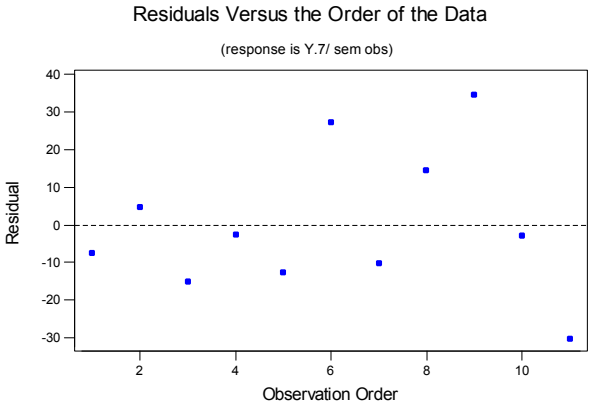
obs	SRES1	HI1	COOK1	obs	SRES1	HI1	COOK1
1	-0,96747	0,219780	0,131831	8	1,23025	0,085852	0,071071
2	-0,24878	0,157280	0,005775	9	1,80557	0,085852	0,153084
3	-0,84798	0,157280	0,067101	10	1,03753	0,112637	0,068320
4	-0,16162	0,112637	0,001658	11	0,54838	0,157280	0,028062
5	-0,45359	0,112637	0,013058	**12**	**-1,18882**	**0,300137**	**0,303046**
6	0,71428	0,112637	0,032381	**13**	**-1,84634**	**0,300137**	**0,730970**
7	-0,07962	0,085852	0,000298				

Tanto pela análise gráfica quanto póla tabela acima é possível notar que os pontos 12 e 13 são os de maior influencia no modelo. Para verificar esta influencia será ajustado o modelo sem estas observações.

A reta estimada com todas observações é: Y.7 = 43,8393 + 37,2321 X.7
A reta estimada sem as observações 12 e 13 é: Y.7 = 10,1 + 47,5 X.7

Neste caso foi verificado que realmente estas observações estavam influenciando muito o modelo, pois as duas retas estimadas acima são bastante diferentes.



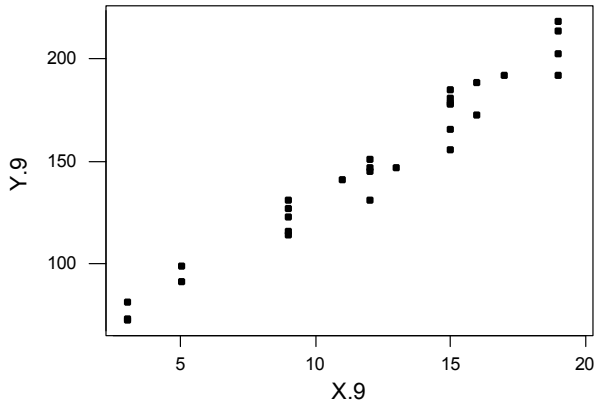


- Parte 3 – Modelo com Ponderação

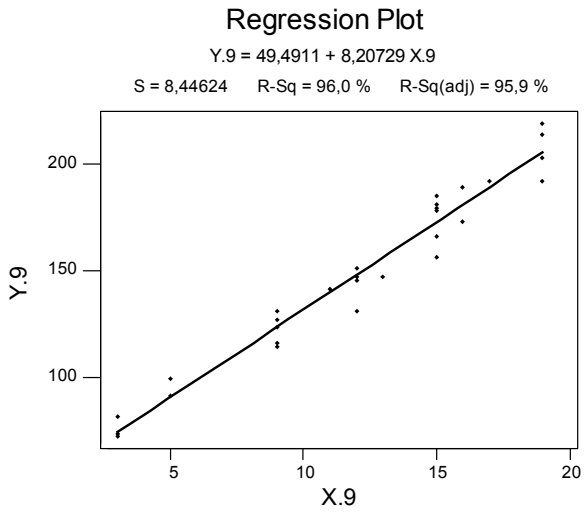
1) **(Adaptação dos dados da Tabela 3.8, Montgomery and Peck)** A renda mensal média de vendas de refeições (Y), assim como os gastos mensais com propaganda (X), foram registradas para 30 restaurantes. Um analista de vendas gostaria de encontrar uma relação entre as vendas e os gastos com propagandas.

Os dados coletados estão disponíveis em na Tabela A.8 em Anexo. (Os valore de Y e X foram arredondados para facilitar a resolução do problema)

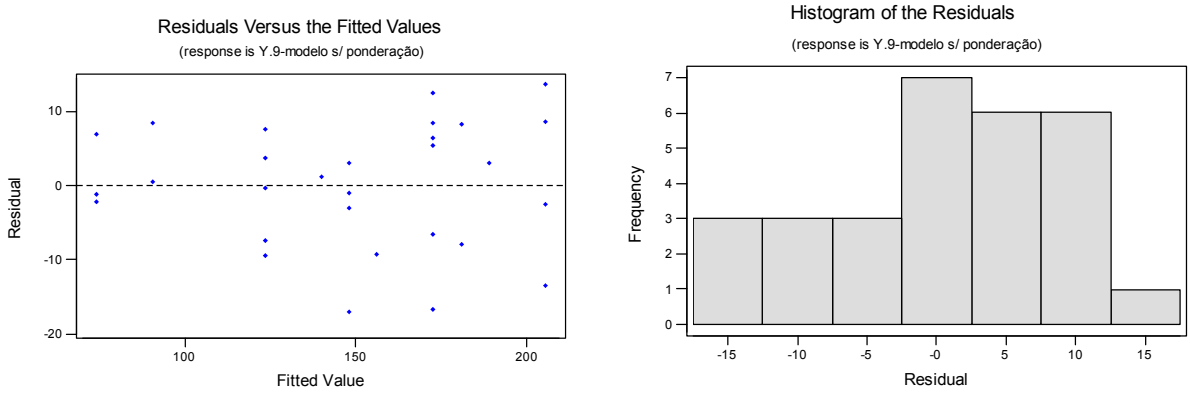
a) Faça o diagrama de dispersão de Y versus X avalie a possibilidade do ajuste de um modelo de regressão linear.

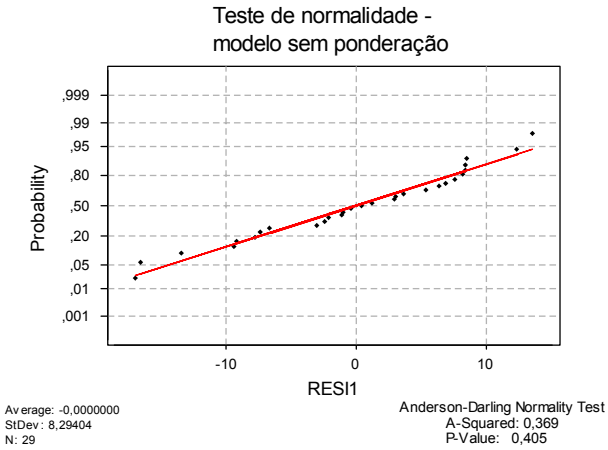


b) Ajuste o modelo de regressão $Y = \beta_0 + \beta_1 X + \varepsilon$, encontrando a reta estimada.



c) Faça Análise dos Resíduos do modelo em b). Se existem problemas com as suposições do modelo de erros normais, quais são eles?





Através dos gráficos acima vê-se que apesar dos resíduos serem normalmente distribuídos, os mesmos não possuem variância constante.

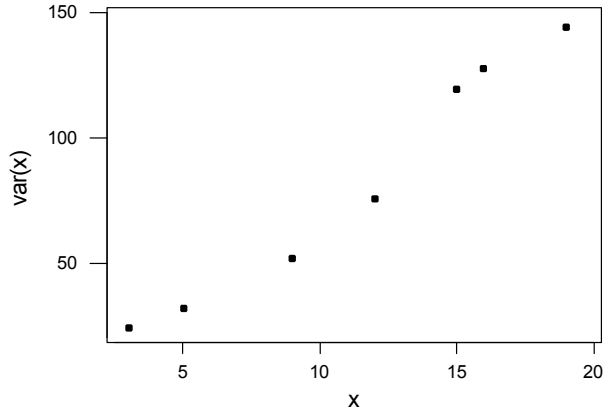
d) Para corrigir o problema da heterocedasticidade, vamos proceder com a técnica dos mínimos quadrados ponderados:

d.1) Calcule a estimativa do Erro Puro para cada nível de X com medidas repetidas (No MINITAB, use o comando `Stat > Basics Statistics > Display Descriptive`).

Variable	X.9	N	Mean	Median	TrMean	StDev
Y.9	3	3	75,33	73,00	75,33	4,93
	5	2	95,00	95,00	95,00	5,66
	9	5	122,20	123,00	122,20	7,19
	11	1	141,00	141,00	141,00	*
	12	4	143,50	146,00	143,50	8,70
	13	1	147,00	147,00	147,00	*
	15	6	174,17	178,50	174,17	10,94
	16	2	181,00	181,00	181,00	11,31
	17	1	192,00	192,00	192,00	*
	19	4	207,00	208,50	207,00	12,03

d.2) Faça um gráfico de Var(Y|X), as estimativas do Erro Puro encontradas em d.1), versus nível de X . Existe relacionamento entre estas duas variáveis? Se sim, de que tipo?

x	*var(x)*
3	*24,305*
5	*32,036*
9	*51,696*
12	*75,690*
15	*119,684*
16	*127,916*
19	*144,721*



Existe relacionamento entre as variáveis, e este é linear positivo (tipo Y = X).

d.3) Crie uma coluna de pesos e coloque o inverso da coluna X . Por que usar o inverso de X como peso? (Pense no relacionamento encontrado em d.2) e nos exemplos utilizados em sala).

Pesos	***Pesos***	***Pesos***	***Pesos***
0,333333	*0,076923*	*0,062500*	*0,090909*
0,333333	*0,066667*	*0,058824*	*0,083333*
0,333333	*0,066667*	*0,052632*	*0,083333*
0,200000	*0,066667*	*0,052632*	*0,083333*
0,200000	*0,066667*	*0,052632*	*0,083333*
0,111111	*0,066667*	*0,052632*	
0,111111	*0,066667*	*0,111111*	
0,111111	*0,062500*	*0,111111*	

Devido ao relacionamento linear entre a variância e X. Fazendo esta ponderação estamos dando um peso pequeno aos pontos com resíduos maiores e pesos maiores aos pontos com resíduos pequenos. Assim tornamos os pontos mais homogêneos, isto é, concentrados em torno de um só valor.

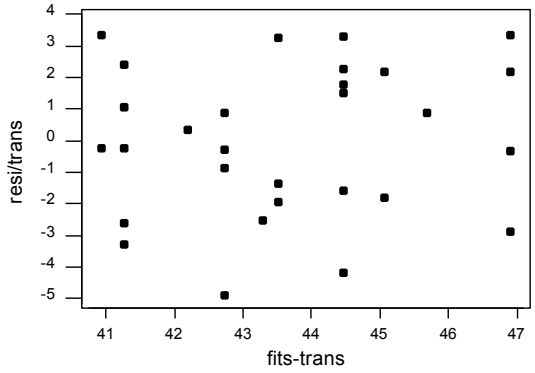
d.4) Use os pesos construídos em f) para ajustar o modelo em b). (No MINiTAB, na janela Regression, botão Options, selecionar a coluna com pesos no espaço weights. Não se esqueça de guardar os resíduos e os preditos.

```
The regression equation is
Y.9 = 51,2 + 8,07 X.9

Predictor         Coef     SE Coef           T        P
Constant       51,170       2,620       19,53    0,000
X.9            8,0686      0,2521       32,00    0,000

S = 2,422       R-Sq = 97,4%     R-Sq(adj) = 97,3%
```

e) Análise dos Resíduos: Crie uma coluna com a multiplicação da coluna de resíduos pela coluna da raiz quadrada dos pesos. Faça o mesmo com a coluna dos preditos e com a coluna dos valores de X .

resi/trans	***x-trans***	***fits-trans***	***resi/trans***	***x-trans***	***fits-trans***	***resi/trans***	***x-trans***	***fits-trans***
3,24709	*1,73205*	*43,5183*	*0,32432*	*3,31662*	*42,1888*	*3,30531*	*3,87298*	*44,4615*
-1,37171	*1,73205*	*43,5183*	*0,86807*	*3,46410*	*42,7219*	*-4,18246*	*3,87298*	*44,4615*
-1,94906	*1,73205*	*43,5183*	*-0,28663*	*3,46410*	*42,7219*	*-1,81679*	*4,00000*	*45,0668*
-0,22942	*2,23607*	*40,9259*	*-4,90543*	*3,46410*	*42,7219*	*2,18321*	*4,00000*	*45,0668*
3,34829	*2,23607*	*40,9259*	*-0,86398*	*3,46410*	*42,7219*	*0,88872*	*4,12311*	*45,6781*
1,07092	*3,00000*	*41,2624*	*-2,51320*	*3,60555*	*43,2837*	*-0,33789*	*4,35890*	*46,9093*
-3,26241	*3,00000*	*41,2624*	*1,75612*	*3,87298*	*44,4615*	*-2,86146*	*4,35890*	*46,9093*
-2,59574	*3,00000*	*41,2624*	*-1,60047*	*3,87298*	*44,4615*	*3,33276*	*4,35890*	*46,9093*
-0,26241	*3,00000*	*41,2624*	*2,27252*	*3,87298*	*44,4615*	*2,18569*	*4,35890*	*46,9093*
2,40426	*3,00000*	*41,2624*	*1,49792*	*3,87298*	*44,4615*			

f) Faça o gráfico de resíduos transformados versus preditos transformados. O problema da heterocesdasticidade foi resolvido?



O problema da heterocedasticidade foi resolvido.

g) Caso não haja problemas em i), construa a Tabela Anova e faça o teste da Falta de Ajuste da Tabela Anova.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	6009,9	6009,9	1024,11	0,000
Residual Error	27	158,4	5,9		
Lack of Fit	8	23,2	2,9	0,41	0,903
Pure Error	19	135,3	7,1		
Total	28	6168,4			

H_0 : o modelo não apresenta falta de ajuste
H_a: o modelo apresenta falta de ajuste

Como o P-valor da falta e ajuste é maior que 0,05, pode-se afirmar que o modelo de regressão ajustado não apresenta falta ajuste.

h) Caso não haja problemas no teste de falta de ajuste, faça o teste F da regressão (escreva as hipóteses nula e alternativa de cada teste).

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

Sendo a probabilidade de significância deste teste aproximadamente zero, é possível dizer que β_1 é diferente de zero, ou seja, o modelo ajustado é razoável.

i) Utilize agora a transformação raiz quadrada em Y e ajuste o modelo de regressão linear, fazendo a análise de resíduos . Esta transformação resolve o problema da heterocedasticidade?

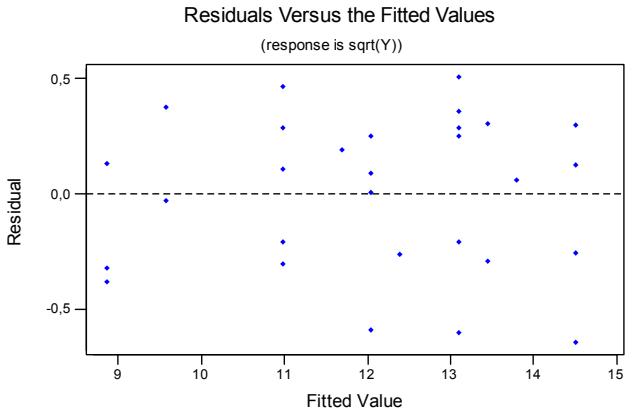
The regression equation is
sqrt(Y) = 7,81 + 0,352 X.9

Predictor	Coef	SE Coef	T	P
Constant	7,8104	0,1676	46,60	0,000
X.9	0,35216	0,01284	27,43	0,000

S = 0,3384 R-Sq = 96,5% R-Sq(adj) = 96,4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	86,152	86,152	752,33	0,000
Residual Error	27	3,092	0,115		
Lack of Fit	8	0,433	0,054	0,39	0,914
Pure Error	19	2,658	0,140		
Total	28	89,244			



Ao analisar-se o gráfico acima nota-se que o problema de heterocedasticidade dos resíduos foi resolvido.

j) Analisando o valor do R2, compare o ajuste do modelo em b) feito via mínimos quadrados ponderados com o ajuste feito via transformação "raiz quadrada" em Y. Por que não podemos comparar os valores do MSResidual?

Modelo ponderado: R^2 = 97,4%
Modelo transformado: R^2 = 96,5%

Apesar dos dois valores estarem próximos, o R^2 do modelo ponderado é maior, sendo assim este modelo parece ser o melhor.

Não pode-se comparar os MSResidual porque estamos tratando de escalas diferentes.

- Parte 4 – Multicolinearidade e Análise de Variância via Análise de Regressão

2) **(Multicolinearidade)** Um grupo de estudantes participou de um experimento simples: cada estudante teve anotado sua altura (height), peso (weight), sexo (sex) , hábito de fumo (smokes), nível de atividade usual (activity) e pulso em repouso. Depois, eles correram no lugar durante um minuto e o pulso foi novamente medido. O objetivo é saber como predizer a medição do pulso depois da corrida através das variáveis medidas .

Pulse1	pulso antes da corrida (em batidas por minuto)
Pulse2	pulso depois da corrida (em batidas por minuto)
Smokes	1= fuma regularmente ; 2 = não fuma regularmente
Sex	1 = homem 2 = mulher
Height	altura (em polegadas)
Weight	Peso (em libras)
Activity	Nível de atividade física : 1 = leve 2 = moderado 3 = intenso

a) Ajuste um modelo de regressão linear, entrando seqüencialmente com as variáveis: pulse1, Sex, height, weight, smokes, activity. A cada entrada de variável, faça o **teste F seqüencial**, avaliando a Soma de Quadrados Extra devida à variável que está entrando . Avalie os VIF´s (fatores de inflação da variância). (No MINITAB, janela Regression, botão Options).

Tabela ANOVA para os testes F seqüenciais :

Fonte de Variação	*g.l.*	*SS*	*MS*	*MSResi (g.l)*	*F*
Regressão (X1,...X6)	***7***	***8972,9***	***1281,8***	***119,6 (27)***	***10,72 ****
X1 (pulse1)	*1*	*4500,2*	*4500,2*	*233,4 (33)*	*19,28 **
X2 (Sex) | X1	*1*	*3332,9*	*3332,9*	*136,5 (32)*	*24,41 **
X3 (height) | X1, X2	*1*	*62,2*	*62,2*	*138,9 (31)*	*0,45*
X4 (weight) | X3, X2, X1	*1*	*156,0*	*156,0*	*138,3 (30)*	*1,13*
X5 (smokes) | X4,X3,X2,X1	*1*	*201,6*	*201,6*	*136,1 (29)*	*1,48*
X6 (activity) | X5,X4,X3,X2,X1	*2*	*720,1*	*360,1*	*119,6 (27)*	*3,01*
Resíduo (Erro) (X1,...,X6)	***27***	***3227,9***	***119,6***	***--------***	***------***
Total	***34***	***12200,7***	***----***	***----------***	***------***

<u>OBS</u>: as somas de quadrados não somam exatamente a SSTotal devido a erros de arredondamento, dado que cada SS veio do ajuste de modelos diferentes.

$F_{0.05\ ;\ 1;\ 30} = 4,1709$ $F_{0.05\ ;\ 1;\ 29} = 4,1830$ $F_{0.05\ ;\ 2;\ 27} = 3,3541$

Predictor	**Coef**	**SE Coef**	**P**	**VIF**
Constant	30,21	62,88	0,635	
Pulse1	0,6542	0,1906	0,002	1,4
Sex	14,353	6,417	0,034	2,6
Height	0,1565	0,8556	0,856	2,4
Weight	-0,1502	0,1363	0,280	2,7
Smokes	3,690	4,221	0,390	1,2
Act1	3,923	7,548	0,607	3,4
Act2	-8,806	7,861	0,273	2,9

b) Ajuste o modelo de regressão somente com as variáveis que deram contribuição significativa para a Soma de Quadrados de Regressão, avaliando também os VIF´s. Há indicação de problemas de multicolinearidade das variáveis explicativas?

Modelo com Pulse1 e Sex como explicativas:

The regression equation is

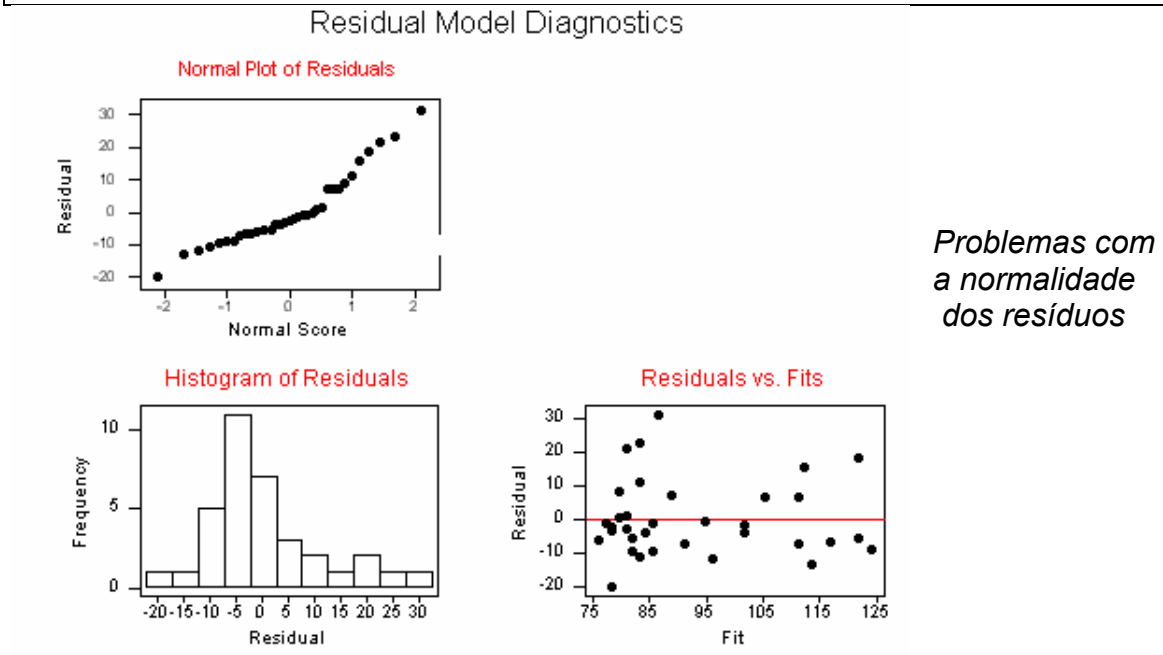
Pulse2 = 18,9 + 0,583 Pulse1 + 23,4 Sex

Predictor	Coef	SE Coef	T	P	VIF
Constant	18,86	13,05	1,45	0,158	
Pulse1	0,5830	0,1950	2,99	0,005	1,2
Sex	23,396	4,735	4,94	0,000	1,2

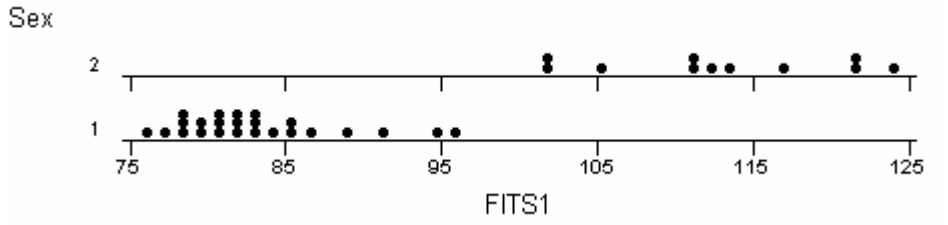
S = 11,68 R-Sq = 64,2% R-Sq(adj) = 62,0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	7833,1	3916,5	28,69	0,000
Residual Error	32	4367,7	136,5		
Total	34	12200,7			



Problemas com a normalidade dos resíduos

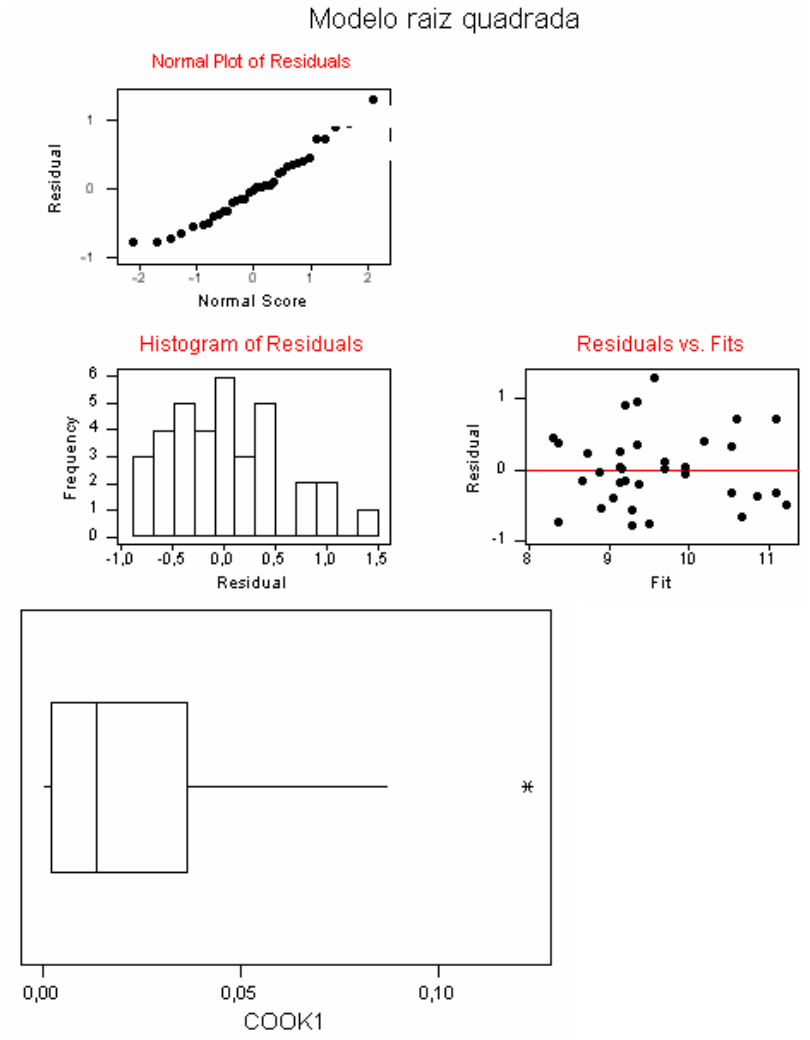


Tentativas para corrigir a normalidade:

✓ *Transformação raiz quadrada : (pulso são contagens, bpm)*

Modelo :
*raiz(pulse2) = b0 + b1 * raiz(pulse1) + b2* Sex + erro*
Sem sucesso.

Modelo :
*raiz(pulse2) = b0 + b1 * raiz(pulse1) + b2* Sex + b31*act1 +b32*act2 + erro*



Corrige o problema da normalidade sem causar outros problemas.

The regression equation is
Sqrt(pulse2) = 4,02 + 0,596 Sqrt(pulse1) + 0,907 Sex2 + + 0,332 Act1 - 0,368 Act2

Sex 2 0 , se masculino
1 , se feminino

```
Predictor      Coef     SE Coef         T         P       VIF
Constant       4,022       1,498      2,68     0,012
Sqrt(pul      0,5961      0,1674      3,56     0,001      1,3
Sex2          0,9069      0,2546      3,56     0,001      1,6
Act1          0,3323      0,3637      0,91     0,368      3,0
Act2         -0,3679      0,3875     -0,95     0,350      2,7

S = 0,5570      R-Sq = 71,0%     R-Sq(adj) = 67,1%

Analysis of Variance

Source            DF       SS          MS         F         P
Regression         4     22,7980      5,6995     18,37     0,000
Residual Error    30      9,3064      0,3102
Total             34     32,1043
```

c) Interprete o modelo ajustado em b).

```
The regression equation is
Sqrt(pulse2) = 4,02 + 0,596 Sqrt(pulse1) + 0,907 Sex2 +
               + 0,332 Act1 - 0,368 Act2
```

*Considerando pessoas de mesmo sexo e mesmo nível de atividade física, um aumento de 1 unidade na raiz quadrada do pulso em repouso leva a um aumento **médio** de 0,596 unidades na raiz quadrada do pulso após a corrida.*

*Considerando pessoas de mesmo pulso em repouso e mesmo nível de atividade física, a raiz quadrada do pulso **médio** de um indivíduo do sexo feminino tem 0,907 unidades a mais do que o pulso **médio** de um indivíduo do sexo masculino.*

*Considerando pessoas de mesmo pulso em repouso e mesmo sexo, a raiz quadrada do pulso **médio** de um indivíduo com nível de atividade física moderada tem 0,332 unidades a mais do que o pulso **médio** de um indivíduo com nível de atividade física leve.*

*Considerando pessoas de mesmo pulso em repouso e mesmo sexo, a raiz quadrada do pulso **médio** de um indivíduo com nível de atividade física moderada tem 0,332 unidades a mais do que o pulso **médio** de um indivíduo com nível de atividade física leve.*

*Considerando pessoas de mesmo pulso em repouso e mesmo sexo, a raiz quadrada do pulso **médio** de um indivíduo com nível de atividade física intensa tem 0,368 unidades a mais do que o pulso **médio** de um indivíduo com nível de atividade física leve.*

***3)* (Análise de Variância via Análise de Regressão)**

Pulse1 - pulso antes da corrida (em batidas por minuto)
Activity - Nível de atividade física : 1 = leve 2 = moderado 3 = intenso

Com os dados do exercício 1, vamos verificar se o pulso médio varia conforme o nível de atividade. Ou seja, devemos comparar a média do pulso em três grupos de

indivíduos.

A hipótese nula é a de que o pulso médio é igual nos três grupos , e a hipótese alternativa é a de que pelo menos um dos grupos tem média diferente.

Estas são as hipóteses usadas na técnica de Análise de Variância, que pode ser realizada através de um modelo de regressão. Vejamos como:

a) Ajuste um modelo de regressão (com intercepto) da variável *pulse1* em função da variável *activity*. Lembre-se de que a variável activity é qualitativa e tem três níveis. Construa a Tabela Anova e teste a significância desta regressão, através do teste F. Em caso de rejeição de H0, teste a significância de cada coeficiente em separado através do teste t.

Criação das variáveis dummies:

	Act 1	***Act 2***	*No MINITAB, menu Calc > Make Indicator Variables. Apagar a coluna com a indicadora da classe a ser referência*
Activity – 1	*0*	*0*	
Activity – 2	*1*	*0*	
Activity – 3	*0*	*1*	

The regression equation is

Pulse1 = 76,7 - 2,91 Act1 - 4,95 Act2

Predictor	Coef	SE Coef	T	P
Constant	76,667	6,764	11,33	0,000
Act1	-2,907	7,159	-0,41	0,687
Act2	-4,952	8,085	-0,61	0,545

S = 11,72 R-Sq = 1,2% R-Sq(adj) = 0,0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	53,7	26,9	0,20	0,823
Residual Error	32	4392,7	137,3		
Total	34	4446,4			

Teste F da Tabela ANOVA $H_0: \beta_1 = \beta_2 = 0$

H_1: *pelo um dos betas diferente de zero*

Teste T para o intercepto β_0 . $H_0: \beta_0 = 0$

$H_1: \beta_0 \neq 0$

b) Interprete o modelo ajustado. Qual é a diferença média entre o pulso de indivíduos do grupo de atividade física leve e o pulso de indivíduos do grupo de atividade física moderada ? E entre indivíduos do grupo de atividade física leve e os de atividade intensa? E entre os dos grupos moderada e intensa? (se a regressão não for considerada significante, essa interpretação servirá como prática).

<u>Interpretação de β_0</u> : representa o pulso médio dos indivíduos de atividade física leve (76,7 batidas por minuto, bpm).

Interpretação de β_1 : representa a diferença entre o pulso médio dos indivíduos de atividade física moderada e os de atividade física leve (2,91 bpm).

Interpretação de β_2 : representa a diferença entre o pulso médio dos indivíduos de atividade física intensa e os de atividade física leve (4,95 bpm).

Diferença entre β_2 e β_1 : representa a diferença entre o pulso médio dos indivíduos de atividade física intensa e os de atividade física moderada (4,95 - 2,91 = 2,04 bpm).

c) Com o teste F em a), existem evidências estatísticas suficientes contra a hipótese de igualdade entre o pulso médio dos três grupos?

Não, pois os coeficientes de regressão não foram considerados significantes.

d) Utilize a técnica da Análise de Variância, responda à questão em c).

Utilizando o menu `Stat > ANOVA > One- Way:`

Response : Pulse1 Factor: Activity

One-way ANOVA: Pulse1 versus Activity

Analysis of Variance for Pulse1

Source	DF	SS	MS	F	P
Activity	2	54	27	0,20	0,823
Error	32	4393	137		
Total	34	4446			

Level	N	Mean	StDev	Individual 95% CIs For Mean Based on Pooled StDev
1	3	76,67	12,22	(----------*----------)
2	25	73,76	11,55	(-----*-----)
3	7	71,71	12,19	(----------*----------)

Pooled StDev = 11,72

64,0 72,0 80,0 88,0

e) Compare a tabela ANOVA de d) com a tabela ANOVA de a). O que se pode concluir?

REGRESSÃO

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	53,7	26,9	0,20	0,823
Residual Error	32	4392,7	137,3		
Total	34	4446,4			

ANOVA

Analysis of Variance for Pulse1

Source	DF	SS	MS	F	P
Activity	2	54	27	0,20	0,823
Error	32	4393	137		
Total	34	4446			

A conclusão é : a técnica da Análise de Variância (ANOVA) para testar igualdade das médias de vários grupos é um caso particular de Análise de Regressão Linear, onde "as variáveis explicativas" são as variáveis dummies criadas a partir da variável indicadora de grupo.

- Parte 5 – Regressão Polinomial

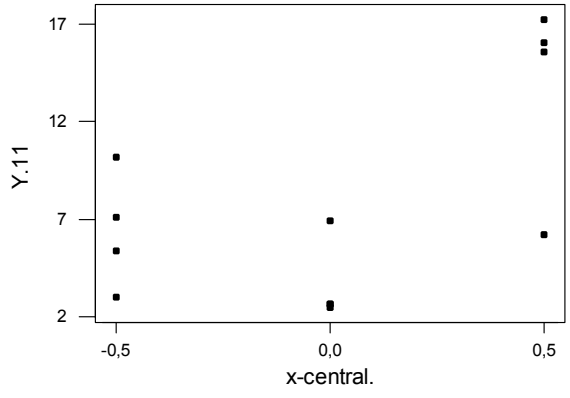
1) (Adaptação de Montgomery and Peck, 2ª Edição : Modelos Polinomiais) O nível de carbonação (gás) de um refrigerante é afetado pela temperatura do produto e pela pressão da máquina que enche as garrafas. Para estudar este processo, foram coletados dados em 12 situações, que estão disponíveis na Tabela A.10, no Anexo.

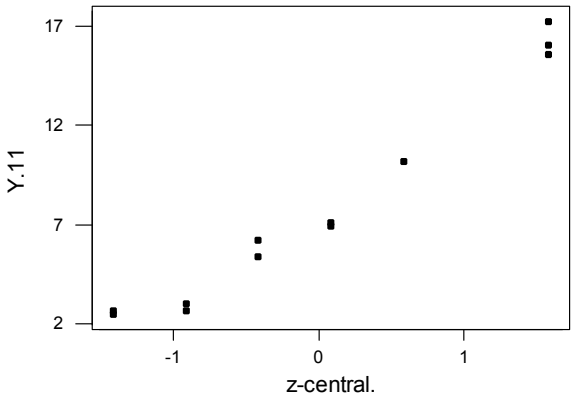
Y carbonação da bebida
X temperatura da bebida
Z Pressão da máquina que enche a garrafa

a) Centralize as variáveis explicativas (X e Z) em torno de suas médias (No MINITAB, use o menu Calc ou o menu Edit > Command Line Editor com os seguintes comandos let c4 = c2-mean(c2) e let c5 = c3-mean(c3) , onde c2 e c3 são as colunas quem contém X e Z, respectivamente).

x-centralizado	*z-centralizado*	*x-centralizado*	*z-centralizado*
-0,5	*-0,91667*	*0,0*	*-0,91667*
-0,5	*-0,41667*	*0,0*	*0,08333*
-0,5	*0,08333*	*0,5*	*-0,41667*
-0,5	*0,58333*	*0,5*	*1,58333*
0,0	*-1,41667*	*0,5*	*1,58333*
0,0	*-1,41667*	*0,5*	*1,58333*

b) Faça um diagrama de dispersão de Y e X e outro para Y e Z, usando as variáveis centralizadas criadas em a). Com qual das duas variáveis (X ou Z) o relacionamento de Y parece ser mais forte? De que tipo parece ser este relacionamento?

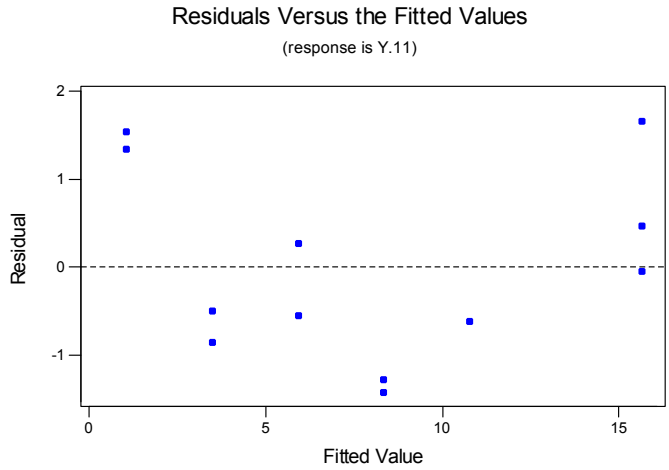




A variável Y aparenta ter uma relação mais forte com a variável Z, e essa relação parece ser linear.

c) Com a variável explicativa escolhida em b), ajuste um modelo de regressão linear simples. Faça o gráfico de resíduos versus preditos. Há algum problema com este gráfico?

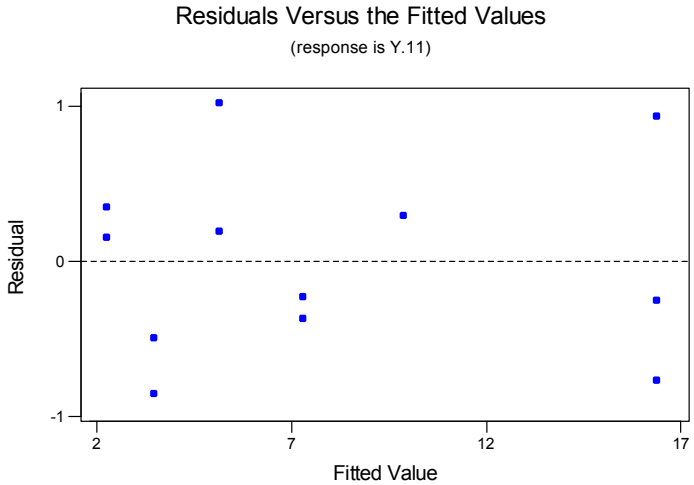
```
The regression equation is
Y.11 = 7,95 + 4,87 z-central.
```



Nota-se que o gráfico acima apresenta uma tendência na forma de uma parábola, o que nos leva a pensar no ajuste de um modelo de regressão quadrático.

d) Acrescente o termo quadrático ao modelo ajustado em c), guarde os resíduos e faça novamente o gráfico de resíduos versus preditos. O aspecto do gráfico melhora em relação ao do gráfico em c)?

```
The regression equation is
Y.11 = 6,91 + 4,56 z-central. + 0,896 (z/centra.)^2
```



O ajuste do modelo com o termo quadrático melhorou o aspecto do gráfico, pois agora o mesmo não apresenta nenhuma tendência e nenhum outro problema.

e) Teste a contribuição do termo quadrático para a soma de quadrados de regressão através do teste F seqüencial.

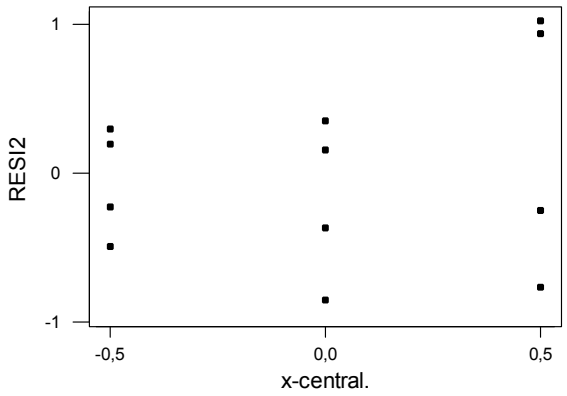
H_o: A contribuição de β_2, dado β_0 e β_1, não é significante ($\beta_2 = 0$)
H_a: A contribuição de β_2, dado β_0 e β_1, é significante ($\beta_2 \neq 0$)

Estatística F = 8,63 /0,45 = 19,178

Região Crítica = {F : F > $F_{1;9;0,05}$} , onde $F_{1;9;0,05}$ = 5,1174

Como F_{obs} está na região crítica estão é possível dizer que a influência de β_2 para o modelo é significante, ou seja, o termo quadrático tem contribuição para o modelo.

f) Faça um gráfico dos resíduos do modelo em d) versus a variável explicativa (centralizada) que ficou de fora (X ou Z). Há algum padrão neste gráfico?



Pelo gráfico acima nota-se que a medida em que o valor de x cresce a variância dos resíduos também aumenta.

g) Acrescente a variável utilizada em f) (centralizada) ao modelo em d). Teste a contribuição desta variável para a soma de quadrados de regressão através do teste F seqüencial. Ela é significante? Em caso negativo, retire-a do modelo.

```
The regression equation is
Y.11 = 7,05 + 4,50 z-central. + 0,775 (z/centra.)^2 + 0,561 x-
central.
```

H_o: A contribuição de β_3, dado β_0, β_1 e β_2, não é significante ($\beta_3 = 0$)
H_a: A contribuição de β_3, dado β_0, β_1 e β_2, é significante ($\beta_3 \neq 0$)

Estatística F = 0,33 /0,47 = 0,703

Região Crítica = {F : F > $F_{1;8;0,05}$} , onde $F_{1;8;0,05}$ = 5,3177

Há evidências a favor da hipótese de que a contribuição de β_3 para o modelo não é significante, isto é, x (centralizada) não é importante para o modelo de regressão ajustado, pois F_{obs} não está na região crítica.

h) Ao modelo escolhido em g), acrescente o termo de interação entre X e Z (centralizado)(comando: let c10 = c4*c5, onde c4 e c5 são as colunas quem contém X e Z centralizadas, respectivamente). A contribuição do termo de interação para a soma de quadrados de regressão é significante (use o teste F seqüencial) ? Em caso negativo, retire-o do modelo.

The regression equation is
Y.11 = 6,87 + 4,72 z-central. + 1,11 (z/centra.)^2 - 0,993 iteracao

H_o: A contribuição de β_3, dado β_0, β_1 e β_2, não é significante ($\beta_3 = 0$)
H_a: A contribuição de β_3, dado β_0, β_1 e β_2, é significante ($\beta_3 \neq 0$)

Estatística F = 0,54 /0,44 = 1,228

Região Crítica = {F : F > $F_{1;8;0,05}$} , onde $F_{1;8;0,05}$ = 5,3177

Não existem evidências a favor da hipótese de que β_3 é significante para o modelo, pois F_{obs} não está na região de rejeição.

i) Para o modelo escolhido em h), faça a análise de resíduos completa (gráficos de resíduos, probabilidade normal, testes, se possível, pontos de influência, multicolinearidade (VIF´s)).

O modelo escolhido em (h) é: Y.11 = 6,91 + 4,56 z-central. + 0,896 (z/centra.)^2

Predictor	Coef	SE Coef	T	P	VIF
Constant	6,9057	0,3072	22,48	0,000	
z-centra	4,5608	0,1934	23,58	0,000	1,1
(z/centr	0,8962	0,2052	4,37	0,002	1,1

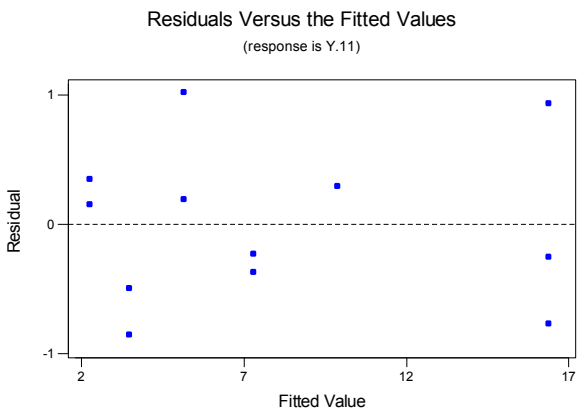
S = 0,6729 R-Sq = 98,8% R-Sq(adj) = 98,5%

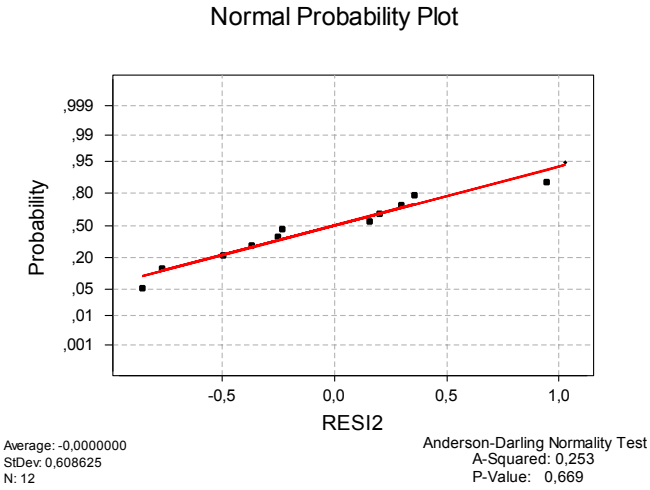
Analysis of Variance

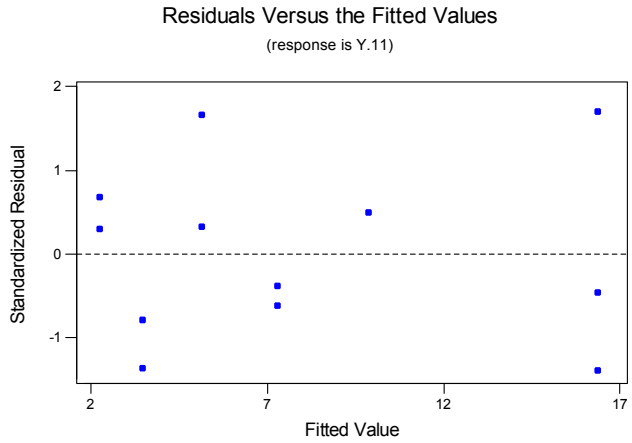
Source	DF	SS	MS	F	P
Regression	2	338,12	169,06	373,41	0,000
Residual Error	9	4,07	0,45		
Lack of Fit	3	2,08	0,69	2,08	0,204
Pure Error	6	2,00	0,33		
Total	11	342,19			

1 rows with no replicates

Source	DF	Seq SS
z-centra	1	329,48
(z/centr	1	8,63







Teste de Durbin-Watson

H_o: Os resíduos não são correlacionados
H_a: Os resíduos são correlacionados

D = 1,83 4 – D = 2,17
dl = 0,83 du = 1,40

Como ambos D e 4 – D são maiores que du pode-se afirmar que os resíduos não são correlacionados.

Obs	SRES1	HI1	COOK1	Obs	SRES1	HI1	COOK1
1	-0,79986	0,143718	0,035794	7	-1,37805	0,143718	0,106244
2	0,32317	0,162140	0,006737	8	-0,62343	0,213644	0,035199
3	-0,38880	0,213644	0,013690	9	1,67079	0,162140	0,180069
4	0,49799	0,204207	0,021212	**10**	**1,71116**	**0,324279**	**0,468394**
5	0,68011	0,391977	0,099398	**11**	**-1,39855**	**0,324279**	**0,312888**
6	0,29892	0,391977	0,019201	12	-0,45841	0,324279	0,033615

Pela análise dos gráficos de resíduos verifica-se que os mesmos possuem homocedasticidade, são normalmente distribuídos (teste de Anderson-Darling) e não são correlacionados, o que é confirmado pelo teste de Durbin-Watson. Nota-se ainda que o

modelo apresenta uma pequena multicolinearidade, mas esta que não é prejudicial ao modelo, pois os VIFs estão próximos de um (1,1).

Através da tabela acima vê-se que existem dois possíveis pontos influentes (obs. 10 e 11), porém ao analisar o gráfico dos resíduos padronizados percebe-se que isto não acontece, pois não há nenhum ponto fora do intervalo de –2 a 2.

j) Faça o teste de falta de ajuste, se possível.

H_o: Não há falta de ajuste
H_a: Há falta de ajuste

Existem evidências de que o modelo não apresenta falta de ajuste, pois o p-valor da falta de ajuste é maior que 0,05 (0,204).

k) Caso o modelo passe pelo teste em j), faça o teste F da regressão e, em caso de significância estatística, faça o teste t individuais.

H_o: $\beta_1 = \beta_2 = 0$
H_a: pelo menos um diferente de zero

Como o valor P da regressão é aproximadamente zero pode-se dizer que pelo menos um parâmetro do modelo é diferente de zero.

H_o: $\beta_1 = 0$ *H_o: $\beta_2 = 0$*
H_a: $\beta_1 \neq 0$ *H_a: $\beta_2 \neq 0$*

Como para os dois parâmetros os valores P são menores que 0,05, nos é permitido afirmar que ambos parâmetros são diferentes de zero.

l) (Utilizando a equação escolhida) Para uma máquina operando a uma pressão de 23,5 e um produto à temperatura de 30, qual é o nível de carbonação esperado? (Lembre-se de que o modelo utiliza as variáveis centralizadas)

O nível de carbonação esperado é de 12,898.

m) Construa um intervalo de 95% de confiança para o valor de Y, quando X e Z possuem os valores de l). Para calcular o erro de estimação, lembre-se de que será necessária a matriz $(X'X)^{-1}$. Para o modelo em h), ela pode ser armazenada em `Storage`, na janela `Regression`. Ela será armazenada no objeto m1. Para imprimí-lo, vá ate o menu `Edit > Command Line Editor` com o seguinte comando: `print m1`.

$(X'X)^{-1}$ =

```
 0,208480  0,036724 -0,107911
 0,036724  0,082633 -0,031666
-0,107911 -0,031666  0,093049
```

$IC_{95\%} = (\hat{Y} \pm t_{\alpha/2,(n-p-1)} \sqrt{QMR[x_0'(X'X)^{-1}x_0]}) = (12,254;13,543)$

❖ *Exercícios de Revisão de Regressão Múltipla*

Considere o modelo de regressão linear múltipla, $\mathbf{Y = X\beta + \varepsilon}$, onde **Y, X,** β e ε são vetores ou matrizes.

1) Se dispomos de 100 "indivíduos" com observações em 5 variáveis consideradas explicativas, mais a variável resposta, quais são as dimensões de **Y, X,** β e ε ?

Y tem dimensão: 100 x 1
X tem dimensão: 100 x 6
β *tem dimensão: 6 x 1*
ε *tem dimensão: 100 x 1*

2) Qual é o método utilizado para estimar o vetor β? Para utilizar este método, é necessário supor alguma distribuição para a variável resposta Y? Em caso positivo, qual distribuição?

O método utilizado para estimar β *é chamado de métodos dos mínimos quadrados. Na verdade, para se usar o método de mínimos quadrados não é necessário supor distribuição para Y. A distribuição é necessária quando queremos fazer testes e construir intervalos.*
Estimativa de $\beta = (X'X)^{-1}X'Y$

3) Quais são as suposições feitas pelo modelo de erros normais? O que estas suposições acarretam para Y?

É *necessário supor que os erros são independentes, aleatórios e normalmente distribuídos com média zero e variância* σ^2*. Isto implica que o vetor* ***Y*** *tenham distribuição normal com média* βX *e variância constante* σ^2*.*

4) Considerando o modelo de regressão linear múltipla, em que situação é possível realizar um teste de falta de ajuste ("lack-of-fit") e qual é objetivo deste teste?

Quando se tem medidas repetidas, lembrando que a repetição tem que acontecer em todas as variáveis para que dois observações sejam consideradas medidas repetidas. Este teste nos permite verificar se a reta de regressão ajustada se "ajusta" aos dados, ou seja, se o modelo é bom.

5) Quais os procedimentos gráficos podem ser usados para verificar as suposições enumeradas no item (3) ? Que outros gráficos podem ser feitos na análise de resíduos?

- *gráfico de probabilidade normal (p/ os erros) – para a verificação de normalidade dos resíduos (e assim dos Y)*
- *Gráfico dos resíduos vs. a ordem (tempo) de coleta, quando disponível – para se constatar a aleatoriedade dos erros ;*
- *Gráfico dos resíduos vs. variável explicativa – para verificar suposição de variância constante (homocedasticidade) e aleatoriedade dos resíduos;*

- *Gráfico dos resíduos vs. Preditos – para verificar suposição de variância constante (homocedasticidade) e aleatoriedade dos resíduos;*
- *E ainda o gráfico de resíduos vs. variáveis que não entraram no modelo – Para verificar se há relação entre os resíduos do modelo e as variáveis fora dele.*

6) Quais são as hipóteses nula e alternativa do teste F da tabela ANOVA ?

H_0: $\beta_1 = \beta_2 = \beta_3 = ... = \beta_p$
H_a: *pelo menos* β *um é diferente*

7) (Soma de Quadrados Extras ; Testes F seqüenciais). Pensando num modelo de regressão linear com três variáveis explicativas (X1, X2 e X3) e *n* observações, como montar a tabela ANOVA com a decomposição da soma de quadrados da regressão (SSReg) abaixo?

Fonte	**SS**	**g.l**	**MS**	**F**
Regressão(X1, X2, X3)	*SSReg(X1,X2,X3)*	*3*	*SSReg(X1,X2,X3) / 3*	*MSReg(X1,X2,X3) / MSRes(X1,X2,X3)*
X1	*SSReg(X1)*	*1*	*SSReg(X1) / 1*	*MSReg(X1) / MSRes(X1)*
X2 | X1	*SSReg(X2 | X1)*	*1*	*SSReg(X2 | X1) / 1*	*MSReg(X2| X1) / MSRes(X1,X2)*
X3 | X1, X2	*SSReg(X3 | X1, X2)*	*1*	*SSReg(X3 | X1, X2) / 1*	*MSReg(X3| X1,X2) / MSRes(X1,X2,X3)*
Resíduo (Erro)	*SSRes(X1,X2,X3)*	*n – 4*	*SSRes(X1,X2,X3) / n-4*	
Total	*SSTotal*	*n - 1*		

Explique como obter as SSReg´s da tabela, quais seriam os respectivos graus de liberdade (g.l.), como obter os MS (quadrados médios) e as respectivas estatísticas F.

As somas de quadrados das regressões são obtidas da seguinte forma:
Ex.: Calculo da SQReg(X2|X1) :

- Ajusta-se o modelo apenas com a variável X1 e depois faz-se outra regressão com X1 e X2. Dessa forma obtêm-se: SQReg(X1) e SQReg(X1,X2). Assim temos que: SQReg(X2|X1) = SQReg(X1,X2) - SQReg(X1)

Procede-se dessa maneira para todas as outras SQReg's.

Cada SS tem 1 grau de liberdade, se for adicionado 1 termos, 2 graus se forem adicionados 2 termos e assim por diante. Exemplo: os graus de liberdade da SQReg(X1|X2) é um e os g. l. de SQReg(X2,X3|X1) é 2.

Para se obter os quadrados médios basta dividir a soma de quadrados da regressão pelo seu respectivo grau de liberdade. E para calcular a estatística F dividi-se o MSReg seqüencial pelo MSRes da regressão "maior". Por exemplo: $F_{X3|X1,X2}$ *= MSReg(X3| X1,X2) / MSRes(X1,X2,X3) , onde MSRes(X1,X2,X3) é o MSRes da regressão com as três variáveis, X1, X2 e X3 .*

8) Quais as hipóteses nula e alternativa de cada um dos testes F da tabela ANOVA em (7)?

Ho: A contribuição de β_1, dado β_0, não é significante ($\beta_1 = 0$)
Ha: A contribuição de $\beta 1$, dado $\beta 0$, é significante ($\beta_1 \neq 0$)

Ho: A contribuição de β_2, dado β_1 e β_0, não é significante ($\beta_2 = 0$)
Ha: A contribuição de β_2, dado β_1 e β_0, é significante ($\beta_2 \neq 0$)

Ho: A contribuição de β_3, dado β_2, β_1 e β_0, não é significante ($\beta_3 = 0$)
Ha: A contribuição de β_3, dado β_2, β_1 e β_0, é significante ($\beta_3 \neq 0$)

9) O que é multicolinearidade e o que este problema pode causar na análise de regressão?

Multicolinearidade é a existência de correlação entre as variáveis explicativas. Caso ela exista a qualidade, (precisão) do modelo de regressão ajustado será afetada.

10)Quais são os tipos de pontos de influência e como detectá-los?

Os pontos de influência podem ser pontos de alavancas (outliers em X, mas não em Y) e pontos de influência propriamente ditos (outliers em X e Y). Os Hi´s servem para detectar os pontos de alavanca e o Dcooks os pontos de influência. Pode-se também detectar possíveis pontos de influência através da análise dos gráficos de resíduos.

11) Em qual(is) situação(ões) é indicado o uso do Método dos Mínimos Quadrados Ponderados (MQP) ao invés do Método dos Mínimos Quadrados Ordinários (MQO) na estimação da equação de regressão? Qual é a diferença entre os dois métodos? Quais são as conseqüências de se usar o MQO quando o MQP seria o método indicado?

O método de mínimos quadrados ponderados é indicado quando há indícios de que os erros não apresentam variância constante, verificado através da análise de resíduos. A diferença entre os dois métodos está no fato de que, ao fazer a ponderação, o MQP dá pesos diferentes às observações.

12)Compare a transformação de Box-Cox e o MQP como alternativas para estabilizar a variância dos erros, citando vantagens e desvantagens.

Uma desvantagem do método de box-cox é o fato de que para se fazer previsões é necessário fazer a transformação inversa. Outra desvantagem é que não está implementado em programas estatísticos conhecidos. Uma vantagem dele é que é semiautomático, bastando apenas escolher valores apropriados para os lambdas a serem testados. A desvantagem do MQP é que precisamos descobrir os pesos a serem usados, o que pode ser bastante trabalhoso. Uma vantagem é que já está implementado e fornece a estimativas dos betas diretamente.

13) Quais são as vantagens da centralização das variáveis explicativas em suas médias para a estimação dos parâmetros da regressão ? (Pense em termos da matriz **(X'X)**)

Com a centralização, a média das novas variáveis será igual a zero, zerando os elementos fora da diagonal da matriz (X´X). Isto ajuda na estimação dos betas, que passa a ter estimativas não correlacionadas e ajuda a evitar o problema da multicolinearidade na regressão polinomial.

2º Parte – Exercícios Teóricos

❖ *Regressão Simples*

1) Para o modelo de regressão $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$,
encontre os estimadores de β_0, β_1 e β_2 pelo método de mínimos quadrados dos erros.

2) Mostre que: $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$

Dica : usar o resultado: $\sum_{i=1}^{n} \hat{Y}_i * e_i = 0$

3) Considere o modelo $Y = \beta_0 + \beta_1 X + \varepsilon$.

Mostre que SSReg = $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = \beta_1^2 * S_{xx}$, onde $S_{XX} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

4)

a) Mostre que $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i\left(X_i - \overline{X}\right)$.

b) Usando o resultado de a) , mostre que $\hat{\beta}_1$ é um estimador não viciado para β_1.

5) Para o modelo $Y = \beta_0 + \beta_1 X + \varepsilon$, mostre que $R^2 = r^2_{\hat{y}y}$, onde $r^2_{\hat{y}y}$ é o quadrado do coeficiente de correlação entre Y e $\hat{Y}$.

❖ *Regressão Múltipla*

1) Considerando o modelo de regressão linear simples em termos matriciais: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

onde **β**' = (β_0 , β_1), Y é um vetor de *n* observações e X é a matriz definida

como $\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdots & \\ 1 & X_n \end{bmatrix}$

Mostre que, em termos matriciais, que $\hat{Y}'.e$ = 0 .

Lembre-se de que $\mathbf{e}$ é o vetor de resíduos e pode ser escrito como $\mathbf{e} = \mathbf{(I - H)Y}$, onde $\mathbf{I}$ é a matriz identidade e $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.

2) Considerando o modelo de regressão linear simples em termos matriciais: $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

onde $\boldsymbol{\beta}'$ = (β_1), **Y** é um vetor de *n* observações e **X** é a matriz definida como $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix}$. Note que não há intercepto no modelo.

Mostre, em termos matriciais, **que** $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$. Lembre-se que a solução das equações normais

é dada por $\hat{\boldsymbol{\beta}} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'Y}$. (Neste caso, há somente uma equação normal, pois há somente um parâmetro a ser estimado).

Análise de Regressão no Minitab®

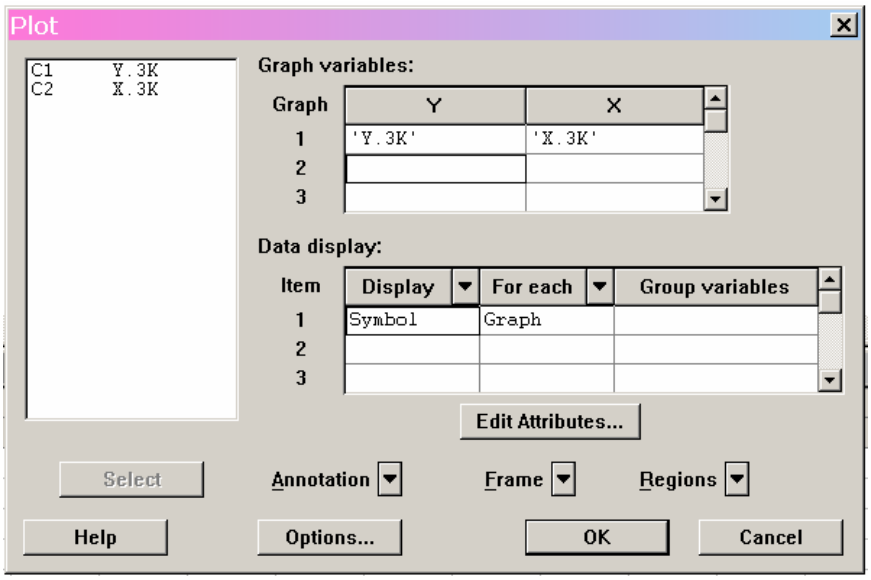
• Regressão Simples

- *Gráfico de dispersão*:

Acesse na barra de ferramentas:

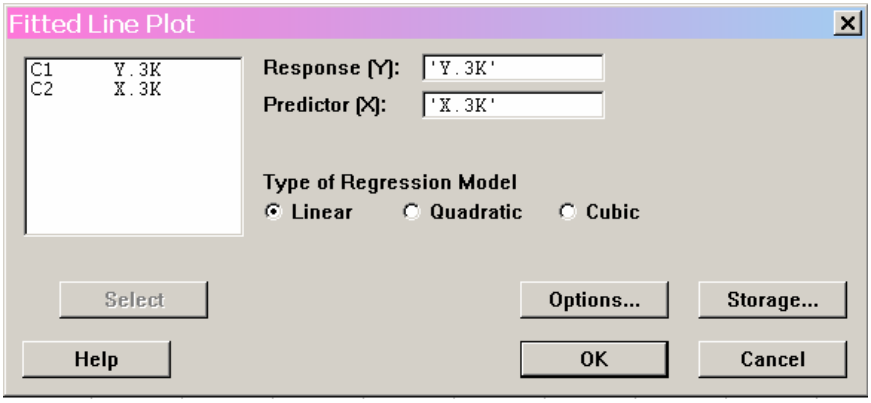
GRAPH > PLOT

Aparecerá a janela abaixo, na qual basta colocar no local correspondente à variável Y a coluna que contém esta variável e fazer o mesmo para X, como na figura abaixo. Para isto você deve clicar uma vez no local onde quer colocar a coluna, depois clicar no nome da coluna que você deseja mover (que está do lado esquerdo da janela) e então clicar em SELECT. Ou clicar uma vez no local onde quer colocar a coluna e depois clicar duas vezes no nome da coluna.



- *Ajustando a reta de regressão*:

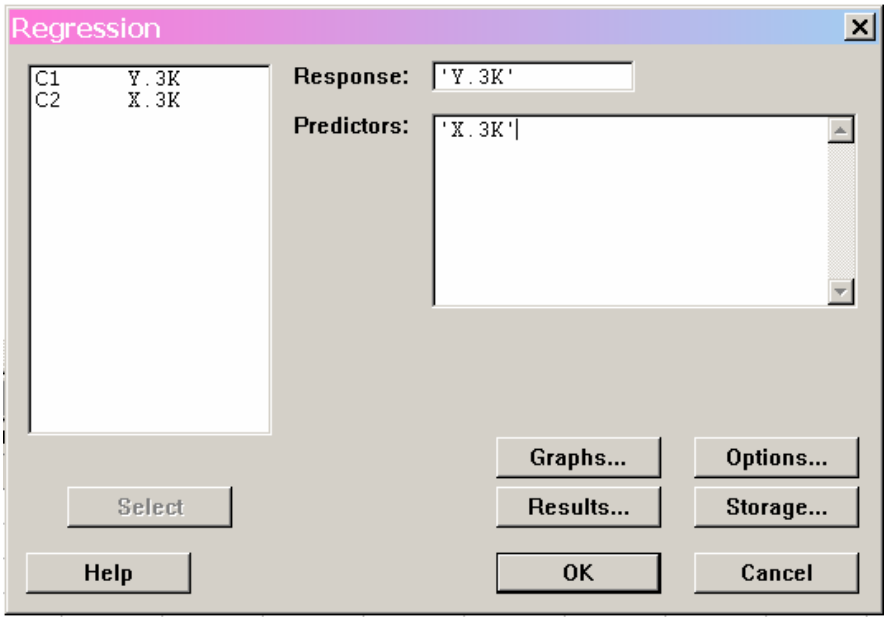
STAT > REGRESSION > FITTED LINE PLOT >



Basta selecionar as variáveis da mesma forma citada anteriormente.

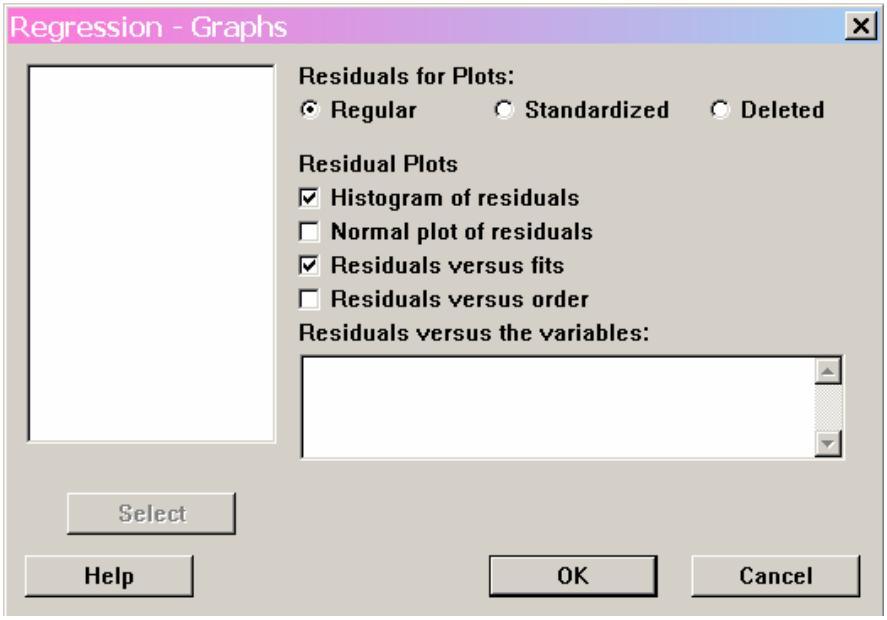
- *Encontrando a reta de regressão, a tabela ANOVA, gráficos de resíduos e etc :*

STAT > REGRESSION > REGRESSION... >

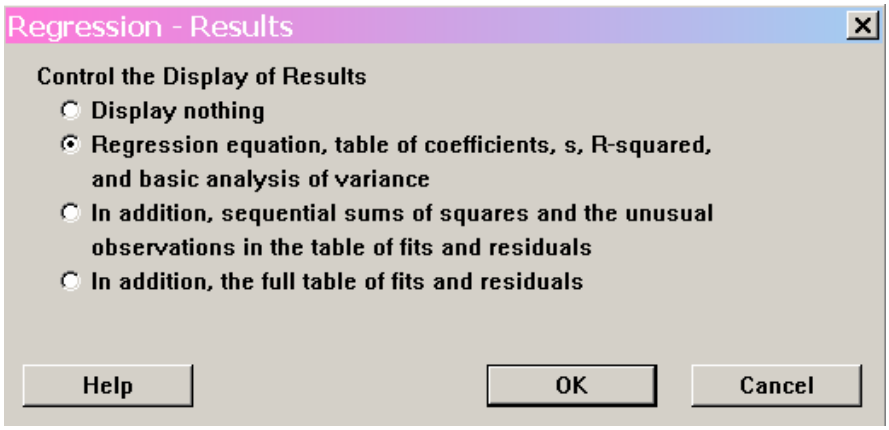


Após aparecer a janela acima basta colocar as colunas correspondentes à variável resposta e à variável explicativa nos locais, respectivamente, RESPONSE e PREDICTORS. Caso se deseje apenas a reta estimada e a tabela ANOVA é só parar por aqui e dar OK.

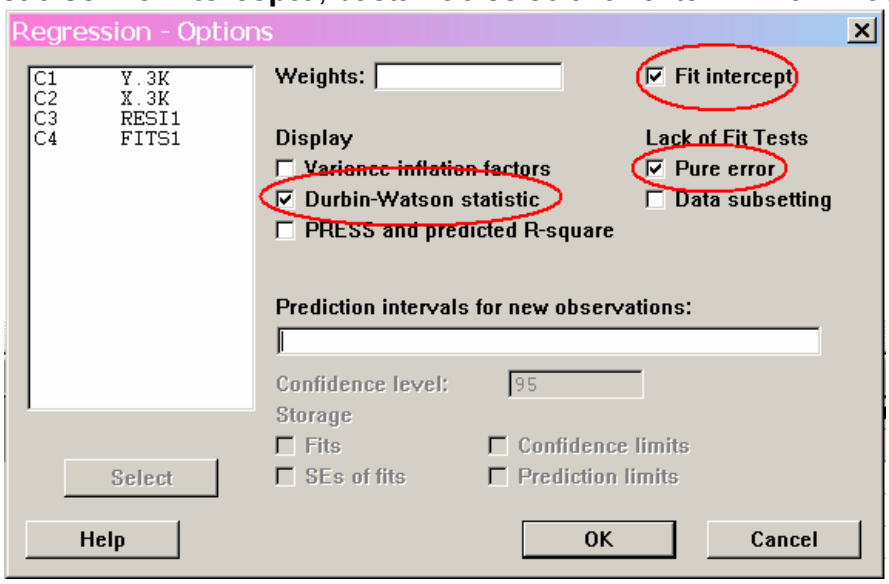
Para se obter os gráficos de resíduos separados clica-se no botão GRAPHS... (dentro na mesma janela mostrada acima) e então é só selecionar com um clique os gráficos desejados.



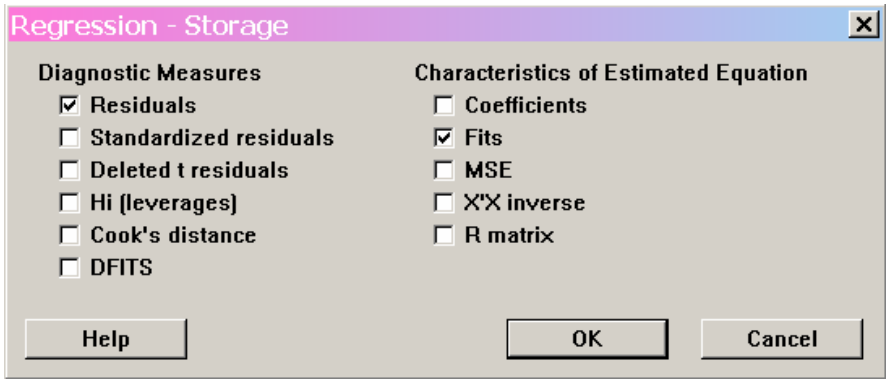
No botão RESULTS você poderá selecionar o tipo de informação que deseja obter juntamente com a tabela ANOVA. A seleção que normalmente já está selecionada é a mostrada na figura que segue.



Já no botão OPTIONS é possível selecionar valores como as estatísticas de teste do teste de Durbin Watson ("`Durbin-Watson statistic`") e do teste de Falta de Ajuste, assim como o valor do Erro Puro ("`Pure Error`"). Também pode-se ajustar uma reta de regressão **sem o intercepto**, basta não selecionar o item "`Fit intercept`".



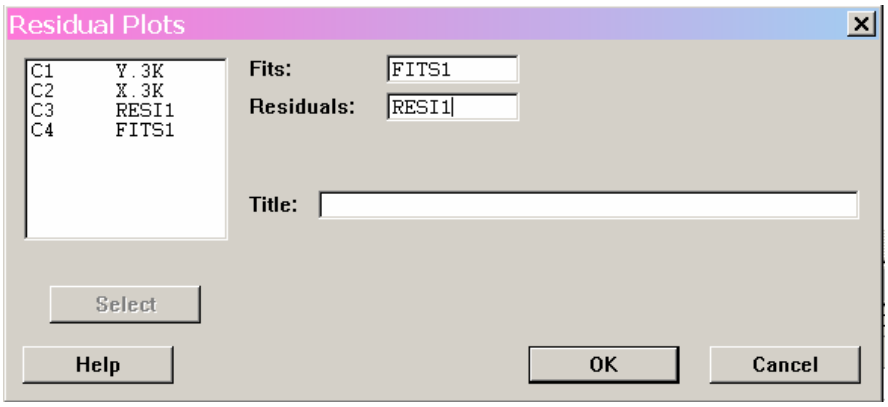
No botão STORAGE, clicando-se em algum dos itens obtém-se uma coluna com os valores solicitados , por exemplo os resíduos e valores preditos.



Caso deseje obter os gráficos de resíduos (mais usados) juntos deve-se seguir o seguinte caminho: STAT > REGRESSION > RESIDUAL PLOTS...

Porém, já deve existir uma coluna com os resíduos e outra com os valores preditos.

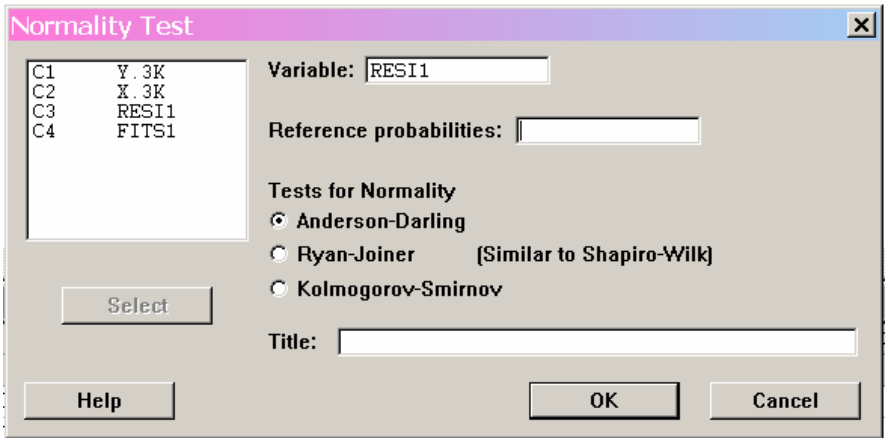
Isto pode ser obtido conforme explicado anteriormente. Assim, basta selecionar estas colunas nos locais mostrados abaixo.



- *Teste de Normalidade para os Resíduos*

STAT > BASICS STATISTICS > NORMALITY TEST >

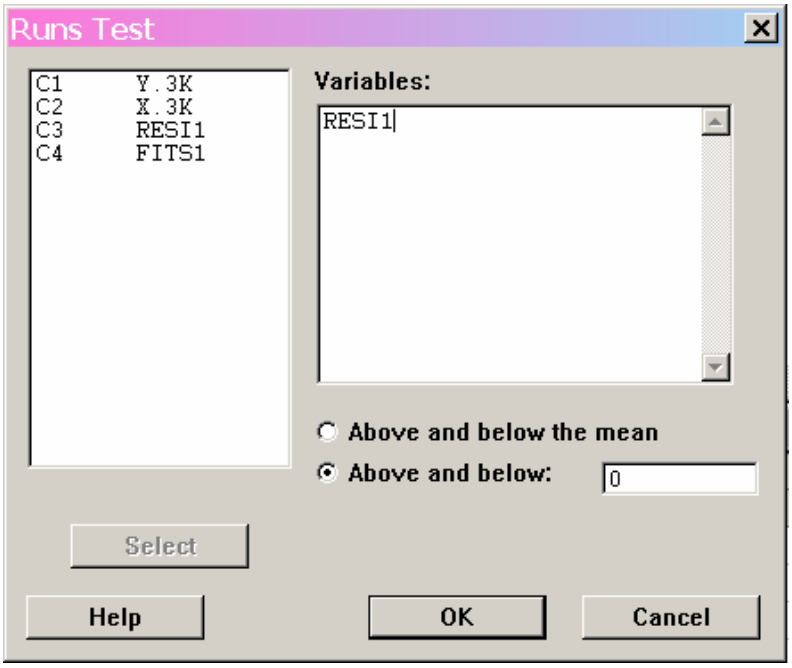
Onde está escrito `Variable` coloca-se a coluna correspondente aos resíduos a serem testados, e então é só escolher o teste de normalidade desejado.



- *Teste de Aleatoriedade (Corridas)*

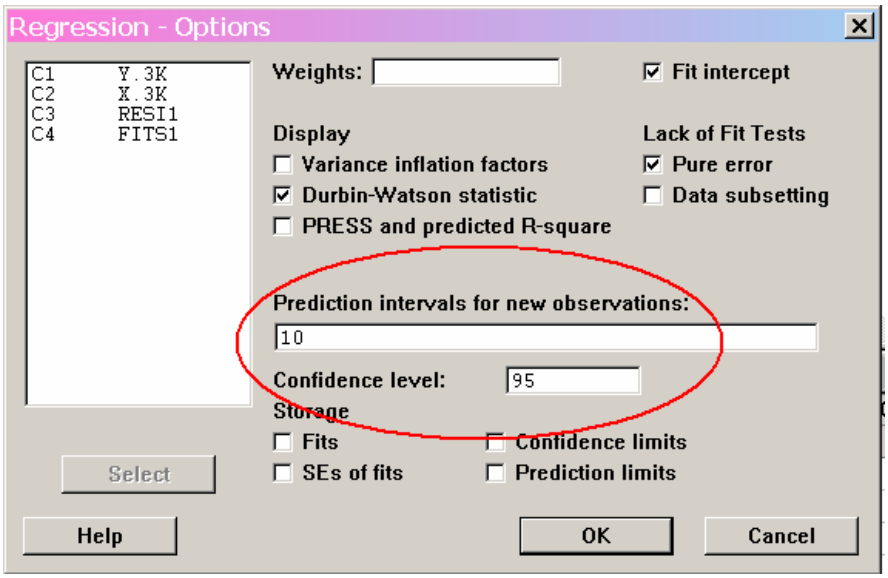
STAT > NONPARAMETRICS > RUNS TEST... >

Na janela mostrada na figura abaixo você deve selecionar a coluna que contém os resíduos e depois escolher entre as opções de fazer teste considerando valores acima e abaixo da média dos resíduos ("`Above and below the mean`"), ou acima e abaixo de um valor a ser escolhido por você ("`Above and below:____`").



- *Predição de valores*

STAT > REGRESSION > REGRESSION > OPTIONS >



Como mostrado na figura acima, deve-se colocar no local "`Prediction intervals for new observations`" o valor para o qual deseja-se fazer a previsão. E onde está "Confidence Level" coloca-se o nível de confiança, isto é 1 – nível de significância.

➢ *Interpretando os Resultados*

- *Tabela ANOVA e reta estimada:*

Regression Analysis: Y.3K versus X.3K

```
The regression equation is
Y.3K = 1,00 - 0,00290 X.3K
```

Reta de regressão estimada

Valores estimados para os coeficientes

Desvio-padrão dos coeficientes

Estatísticas de teste e valores P dos coeficientes

```
Predictor         Coef     SE Coef        T        P
Constant       1,00210     0,01089    92,04    0,000
X.3K        -0,0029035   0,0002335   -12,43    0,000
```

Desvio-padrão do erro

R^2

R^2 ajustado (regressão múltipla)

```
S = 0,03933      R-Sq = 82,9%      R-Sq(adj) = 82,3%
```

Tabela ANOVA:

Graus de liberdade

somas de quadrados

quadrado médio

estatísticas de teste

Valores P

```
Analysis of Variance

Source            DF        SS         MS         F        P
Regression         1     0,23915    0,23915    154,62    0,000
Residual Error    32     0,04949    0,00155
  Lack of Fit     30     0,04760    0,00159      1,67    0,443
  Pure Error       2     0,00190    0,00095
Total             33     0,28864
```

Falta de ajuste

```
30 rows with no replicates

Durbin-Watson statistic = 1,98
```

Estatística D do teste de Durbin Watson

```
Predicted Values for New Observations

New Obs      Fit      SE Fit          95,0% CI              95,0% PI
1        0,97307     0,00917    ( 0,95439; 0,99174)   ( 0,89081; 1,05532)

Values of Predictors for New Observations
```

Estimação pontual

Intervalo de confiança

Intervalo de predição

```
New Obs       X.3K
1             10,0
```

valor para o qual foi feita a predição

- *Teste de Aleatoriedade (corridas)*

Runs Test: RESI1

```
RESI1

  K =     0,0000
```

valor usado para fazer o teste (a média dos resíduos ou o valor que você escolheu)

```
The observed number of runs =  17
The expected number of runs =  17,7647
15 Observations above K   19 below
     The test is significant at  0,7870

     Cannot reject at alpha = 0,05
```

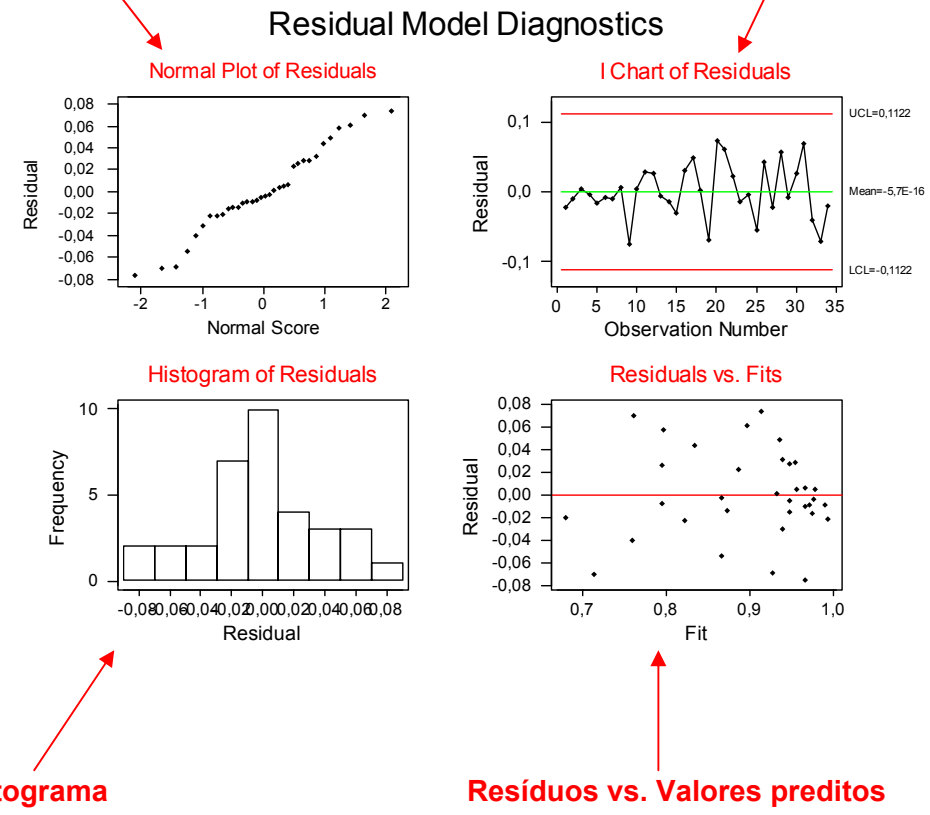
nível de significância para o qual se rejeitaria H_0

Conclusão do teste

- *Gráficos de resíduos (juntos)*

Gráfico de normalidade

Resíduos vs. Ordem de coleta



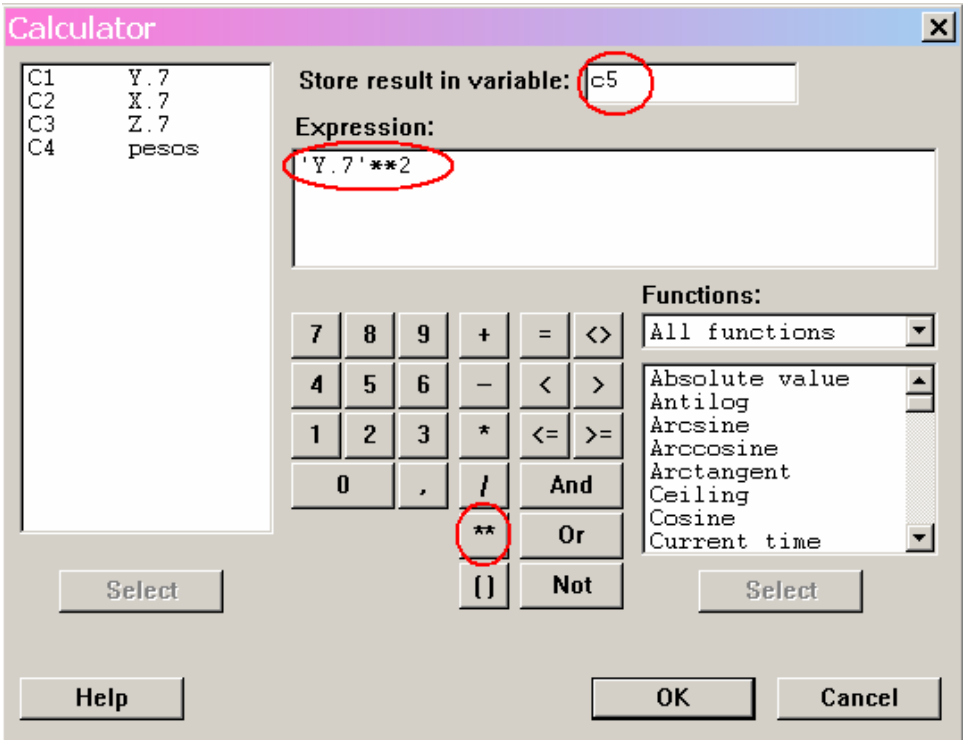
Histograma

Resíduos vs. Valores preditos

- **Transformação das variáveis**

O Minitab não faz as transformações automaticamente dentro do item Regression, para isto deve-se utilizar como auxílio do menu Calc (CALC > CALCULATOR) da seguinte forma:

1. Na janela abaixo você faz a transformação que deseja. Por exemplo, Y^2:



No caso da figura os valores do resultado serão colocados na coluna c5. Lembrando que potência é representada por : "**", como acima.

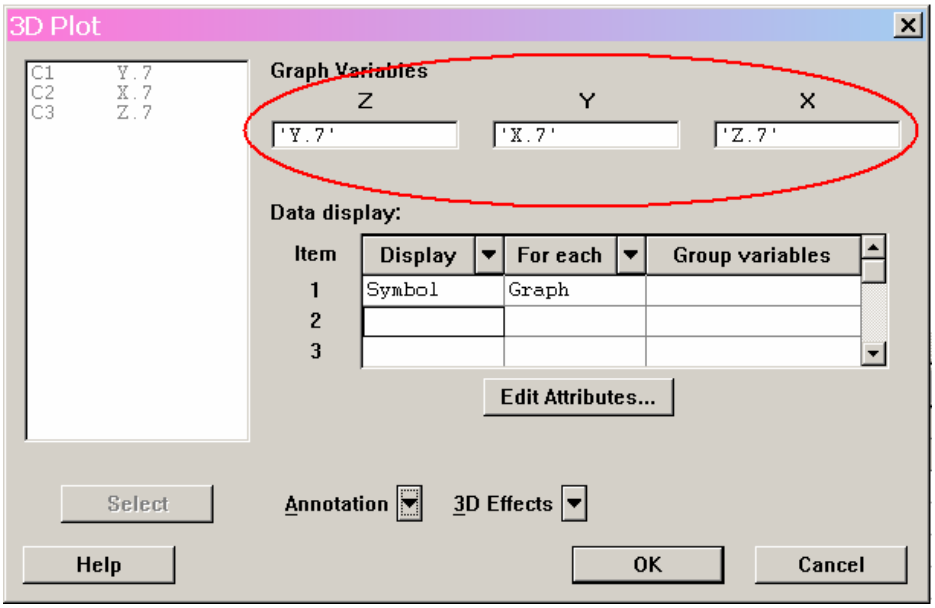
2. Agora é só ajustar o modelo de regressão usando esta coluna (neste caso, c5) como variável resposta.

A análise da saída do programa é a mesma das mostradas antes.

- **Regressão Múltipla**

  - *Diagrama de dispersão em 3D:*

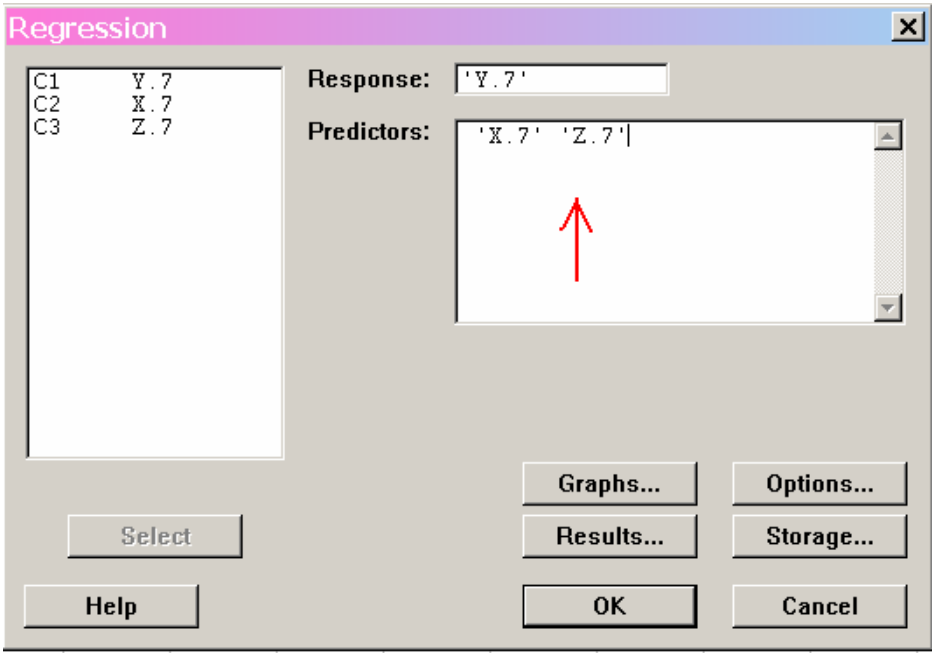
GRAPH > 3D PLOT >



Logo é só selecionar as colunas com as variáveis nos locais indicados na figura acima.

  - *Encontrando a reta de regressão, a tabela ANOVA, gráficos, VIFs, COOKs e etc.*

STAT > REGRESSION > REGRESSION… >

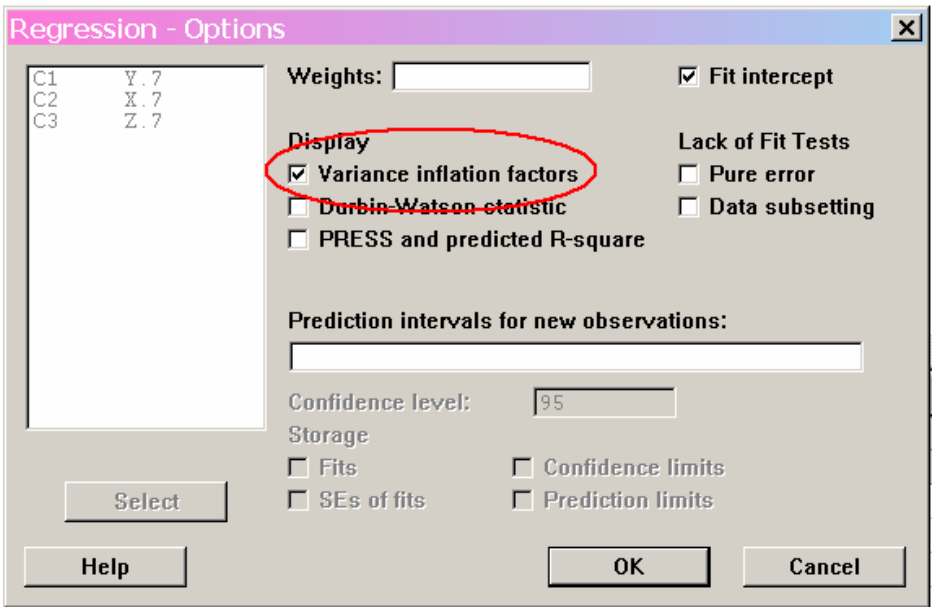


No caso da regressão múltipla o procedimento é o mesmo do utilizado na regressão simples, sendo que agora coloca-se no quadro "`Predictors`" todas as

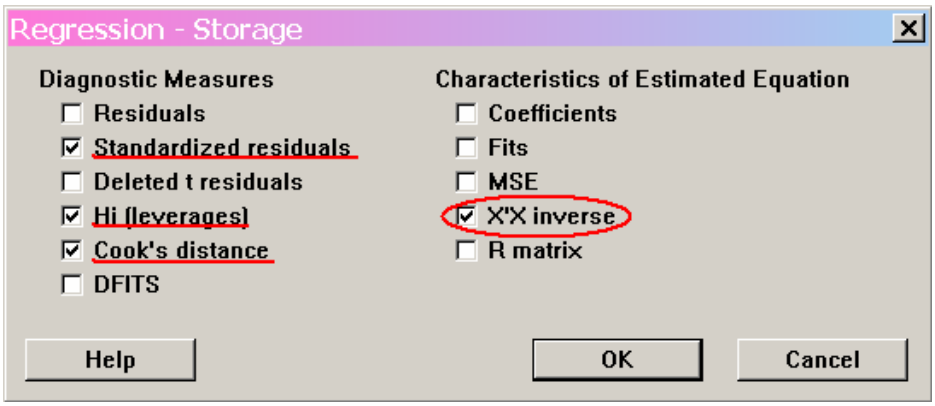
variáveis explicativas separadas por espaço, como mostrado na figura anterior.

Os botões GRAPHS, OPTIONS, RESULTS e STORAGE (da janela acima) continuam tendo as mesmas utilizações que na regressão simples, porém com mais algumas funções a serem usadas.

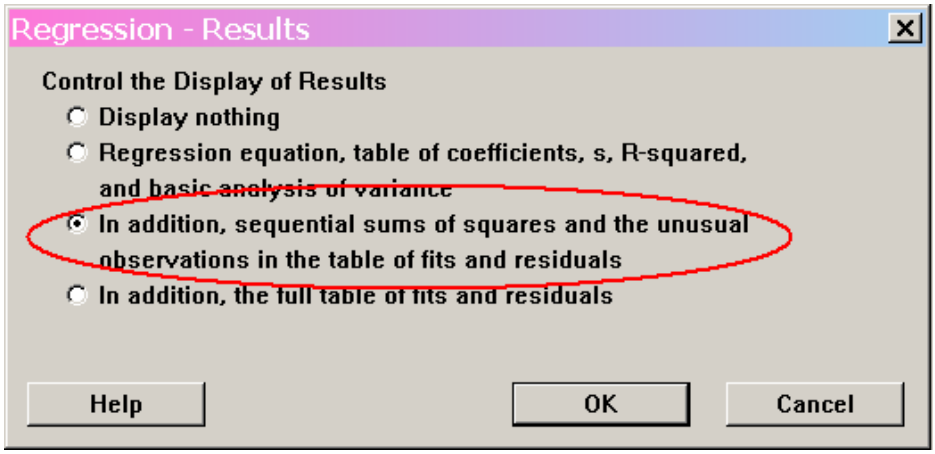
No botão OPTIONS além do que foi citado antes também é possível solicitar os valores dos VIFs ("Variance inflation factors"). Basta fazer a seleção mostrada na figura que segue.



Em STORAGE além dos resíduos e dos valores ajustados também pode-se solicitar os valores dos Hi's ("`Hi (leverages)`"), COOKs ("`Cook's distance`"), Resíduos Studentizados ("`Standardized residuals`") e ainda a matriz $(X'X)^{-1}$ ("`X'X inverse`"), entre outros. Para isto marca-se os itens mostrados na figura abaixo.



Deve-se atentar para o fato de que no caso da regressão múltipla há a necessidade de se obter as somas de quadrados seqüenciais, contudo se no botão RESULTS a opção marcada for a mesma mostrada na regressão simples não obteremos estes valores. Para resolvermos este problema basta acessar o botão RESULTS (na janela REGRESSION) , como mostrado anteriormente, e marcar a terceira opção da janela que aparecerá, como na figura abaixo.



➢ *Interpretando Resultados*

As interpretações são as mesmas que na regressão simples, porém com algumas adições.

Regression Analysis: Y.7 versus X.7; Z.7

VIFs

```
The regression equation is
Y.7 = - 5,95 + 54,4 X.7 - 27,4 Z.7

Predictor        Coef     SE Coef           T        P       VIF
Constant       -5,947       9,350      -0,64    0,539
X.7            54,354       2,375      22,89    0,000       3,6
Z.7           -27,395       3,224      -8,50    0,000       3,6

S = 13,30       R-Sq = 99,0%     R-Sq(adj) = 98,7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      168031       84016    474,98    0,000
Residual Error    10        1769         177
  Lack of Fit      5         302          60      0,21    0,946
  Pure Error       5        1467         293
Total             12      169800

4 rows with no replicates

Source       DF      Seq SS
X.7           1      155258
Z.7           1       12773
```

Somas de quadrados seqüenciais

```
Durbin-Watson statistic = 2,41
```

Os valores de COOKs, His e etc aparecem na planilha onde estão os dados, assim como os resíduos. Já a matriz $(X'X)^{-1}$ não aparece automaticamente, para obtê-la deve-se proceder da seguinte forma: Clique na janela de sessão (onde aparecem os resultados), então vá na barra de ferramentas e acesse o item "EDITOR" e aí clique no item "Enable Commands". Fazendo isto você poderá dar comandos digitando na janela de sessão, na qual aparecerá na última linha `MTB >` . Então quando você pedir a regressão a parecerá o seguinte:

```
MTB > Name m2 = 'XPXI2'
MTB > Regress 'Y.7' 2  'X.7' 'Z.7';
SUBC>   XPXInverse 'XPXI2';
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Pure;
SUBC>   Brief 2.
MTB >
```

É preciso fazer isto para saber como foi chamada a matriz $(X'X)^{-1}$, isto aparece na primeira linha, `MTB > Name m2 = 'XPXI2'`. Agora basta você digitar na última linha o comando: `print m2`.

Dessa forma aparecerá a matriz assim:

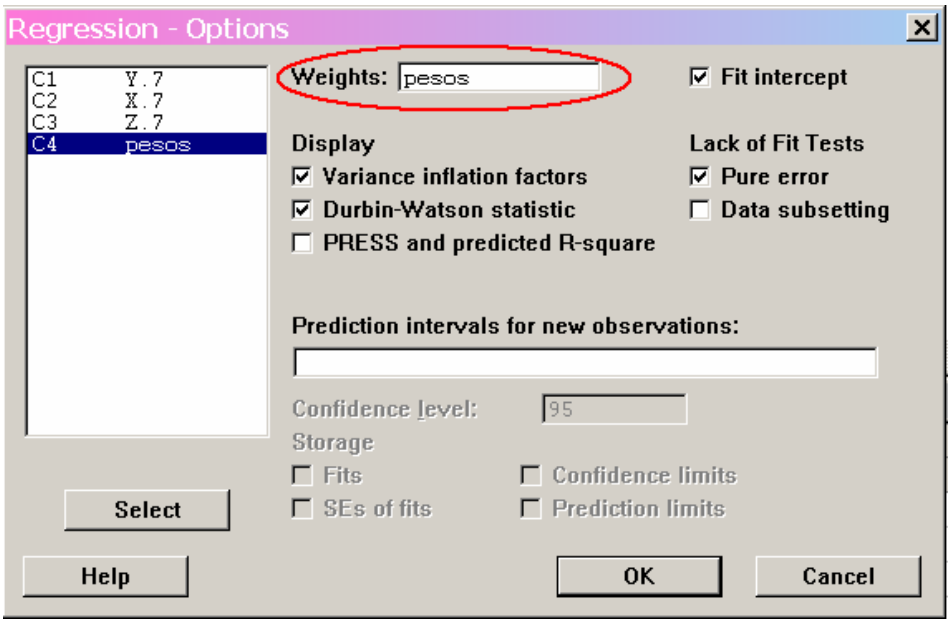
Data Display

```
 Matrix XPXI2

 0,494189 -0,111380  0,106780
-0,111380  0,031881 -0,036723
 0,106780 -0,036723  0,058757
```

- **Modelo Ponderado**

STAT > REGRESSION > REGRESSION... >

Aparecerá à mesma janela mostrada antes, e nesta janela acessa-se o botão OPTIONS e seleciona-se a coluna que contém os pesos no espaço ”Weights”.



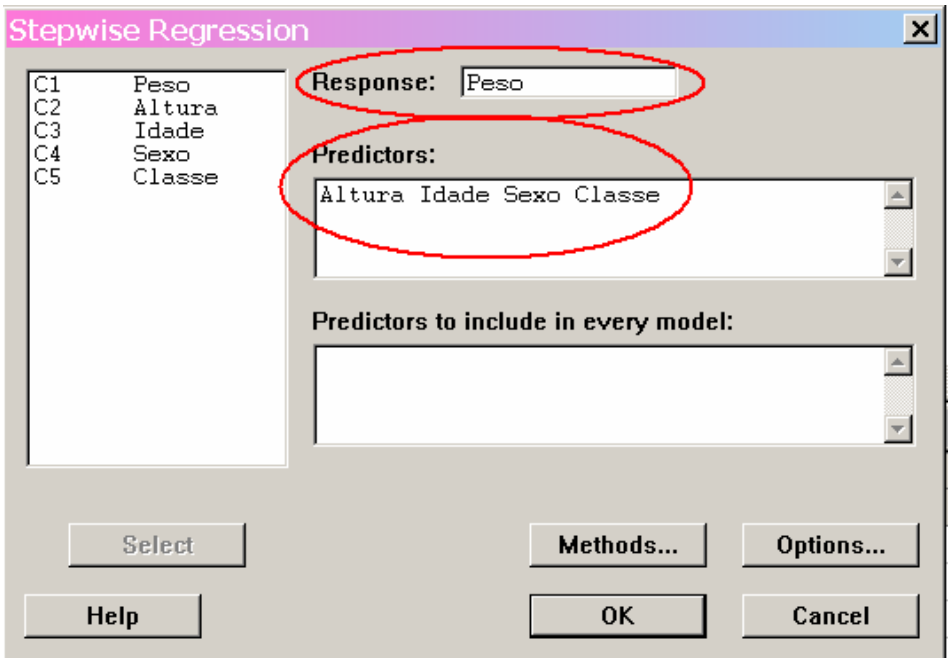
- **Modelo com Interação**

Primeiramente deve-se criar a variável interação, para isso basta multiplicar as variáveis que possuem interação (pode fazer isso no Excel). Então é só ajustar o modelo com mais esta variável. Os resultados serão os mesmos de antes.

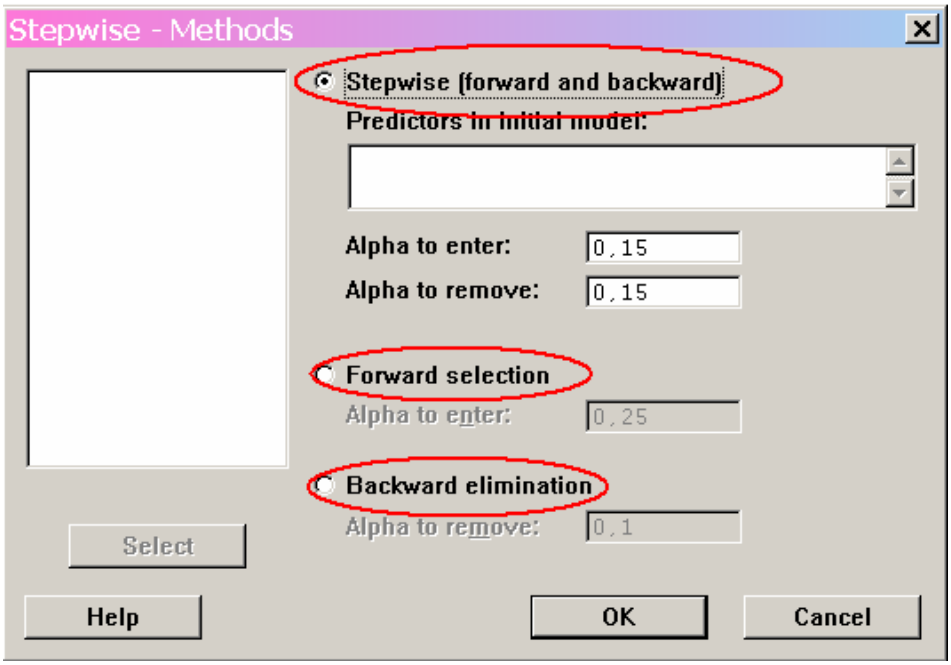
- **Seleção de variáveis**

    - *seleção automática de variáveis*

REGRESSION > STEPWISE... >

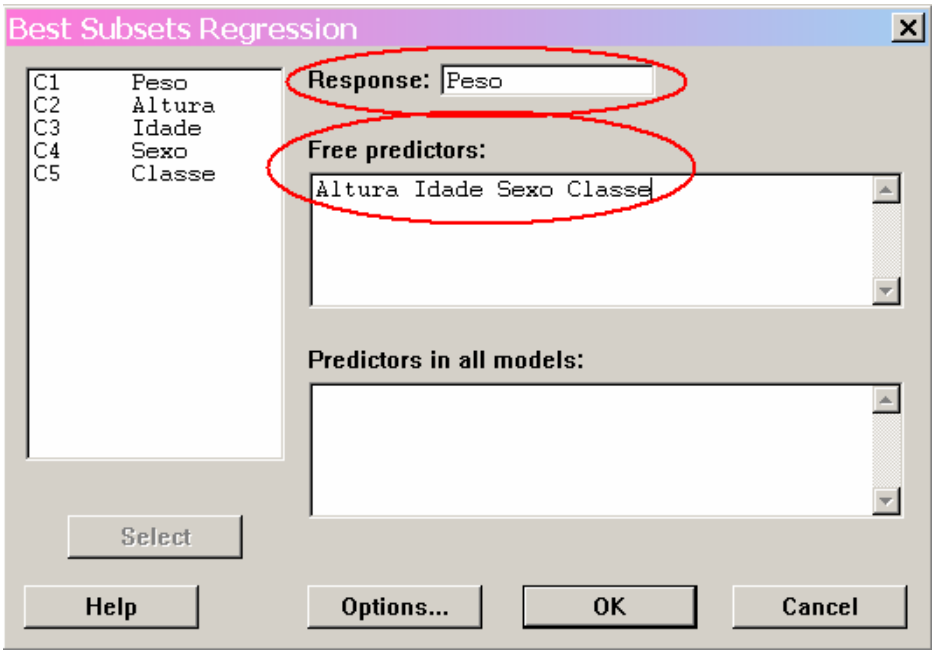


Na janela acima basta entrar com as colunas nos locais indicados. No botão "`Methods...`" escolhe-se o método de seleção desejado: Stepwise ("`Stepwise(forward and backward)`"), Forward ("`Forward selection`") ou Backward ("`Backward elimination`"). Os níveis de significância de entrada e saída de variáveis são colocados em: "`Alpha to enter`" e "`Alpha to remove`".



- *Método de ajuste de todos os modelos possíveis*

REGRESSION > BEST SUBSETS >



Coloca-se a coluna com variável resposta em "`Response`", e as variáveis explicativas em "`Free predictors`". Caso exista alguma variável que você queira que esteja em todos os modelos ajustados basta coloca-la no local "`Predictors in all models`".

➢ *Interpretando os resultados*

- *seleção automática de variáveis*

Método utilizado

Stepwise Regression: Peso versus Altura; Idade; Sexo; Classe

Backward elimination. Alpha-to-Remove: 0,1

Response is Peso on 4 predictors, with N = 34

Step	1	2	3	4
Constant	-107,39	-95,79	-92,26	-85,34
Altura	94	89	88	86
T-Value	4,69	6,62	6,72	6,64
P-Value	0,000	0,000	0,000	0,000
Idade	0,19	0,13		
T-Value	0,55	0,43		
P-Value	0,590	0,671		
Sexo	1,8			
T-Value	0,38			
P-Value	0,707			
Classe	1,8	1,8	2,3	
T-Value	0,72	0,76	1,03	
P-Value	0,475	0,456	0,313	
S	8,71	8,59	8,47	8,48
R-Sq	59,80	59,60	59,35	57,97
R-Sq(adj)	54,25	55,56	56,73	56,65
C-p	5,0	3,1	1,3	0,3

nº do passo

valor de β_0

Variáveis explicativas e seus respectivos P-valores e estatísticas de teste T.

erro-padrão, R^2 e R^2 ajustado e Cp de Mallows

A variável que restar no último passo é a variável escolhida para permanecer no modelo, que neste caso foi Altura.

Para os outros métodos a saída do Minitab é semelhante a esta.

- *Método de ajuste de todos os modelos possíveis*

Best Subsets Regression: Peso versus Altura; Idade; Sexo; Classe

Variáveis explicativas

Response is Peso

Nº de variáveis explicativas utilizadas

Vars	R-Sq	R-Sq(adj)	C-p	S	Altura	Idade	Sexo	Classe
1	58,0	56,7	0,3	8,4799	X			
1	26,1	23,8	23,3	11,244		X		
2	59,3	56,7	1,3	8,4728	X			X
2	58,8	56,2	1,7	8,5269	X	X		
3	59,6	55,6	3,1	8,5866	X	X		X
3	59,4	55,3	3,3	8,6092	X		X	X
4	59,8	54,2	5,0	8,7118	X	X	X	X

Em cada linha tem os valores de R^2, R^2 ajustado, C-p de Mallows e do erro-padrão do modelo ajustado com a(s) variável(eis) explicativa marcada com um X. Neste caso o usuário é quem decide qual o melhor modelo.

- **Validação do modelo**

O Minitab não faz a validação do modelo automaticamente, sendo assim segue um esquema de como faze-la manualmente:

1. Primeiramente deve-se colher uma pequena reamostra da amostra usada para ajustar o modelo de regressão escolhido. Então se separa esta reamostra da amostra original.
2. Em seguida ajusta-se o modelo de regressão com os dados que sobraram da amostra original.
3. Faz-se as predições pontuais e por intervalo (intervalo de predição) para os dados da reamostra que foi separada.
4. Agora basta verificar se os valores reais estão dentro dos intervalos de predição.

Caso se queira o erro quadrático de predição é só acessar o botão "OPTIONS" dentro da janela de regressão e selecionar do lado esquerdo a opção "PRESS and predicted R-square".

Bibliografia

- Norman R. Draper, Harry Smith, *Applied regression analysis,* 3º Ed. New York: Wiley, c1998.
- Douglas C. Montgomery, Elizabeth A. Peck, *Introduction to linear regression analysis*, 2º Ed. New York: J. Wiley, c1992.

Anexos

Tabela A.1

Y2.1	X2.1
1,8	3,3
2,2	4
3,5	5,3
3,4	5,7
2,8	4
2,8	5,3
2,8	2
1,5	2
3,2	6
2,1	5,3
3,7	3,7
2,3	1,3
3	6
3	6,3
1,9	4,7
5,9	6,7
2,2	2,7
1,8	5
1,7	3,7
2,8	4
3,2	4,7
3,8	3,3
1,8	1,3

Tabela A.2

Y.3K	X.3K
0,971	3
0,979	4,7
0,982	8,3
0,971	9,3
0,957	9,9
0,961	11
0,956	12,3
0,972	12,5
0,889	12,6
0,961	15,9
0,982	16,7
0,975	18,8
0,942	18,8
0,932	18,9
0,908	21,7
0,97	21,9
0,985	22,8
0,933	24,2
0,858	25,8
0,987	30,6
0,958	36,2
0,909	39,8
0,859	44,3
0,863	46,8
0,811	46,8
0,877	58,1
0,798	62,3
0,855	70,6
0,788	71,1
0,821	71,3
0,83	83,2
0,718	83,6
0,642	99,5
0,658	111,2

Tabela A.3

OBS	**X**	**Y**
1	19	9,75
2	40	9
3	42	9,6
4	42	9,75
5	47	11,25
6	49	9,45
7	50	11,25
8	54	9
9	56	7,95
10	56	12
11	57	8,1
12	57	10,2
13	58	8,55
14	61	7,2
15	62	7,95
16	62	8,85
17	65	8,25
18	65	8,85
19	65	9,75
20	66	8,85
21	66	9,15
22	66	10,2
23	67	9,15
24	68	7,95
25	68	8,85
26	68	9
27	69	7,8
28	69	10,05
29	70	10,5
30	71	9,15
31	71	9,45
32	71	9,45
33	72	9,45
34	73	8,1
35	74	8,85
36	74	9,6
37	75	6,45
38	75	9,75
39	75	10,2
40	76	6
41	77	8,85
42	80	9
43	82	9,75
44	82	10,65
45	82	13,2
46	83	7,95
47	86	7,95
48	88	9,15
49	88	9,75
50	94	9

Tabela A.4

X	Y	X	Y
120	349	120	350
126	348	106	351
121	349	102	350
115	349	124	346
151	345	108	345
134	348	128	346
129	349	124	349
109	350	103	348
129	347	133	345
124	348	113	352
137	349	119	344
118	350	114	347
130	345	101	352
108	349	120	346
113	349	121	346
124	347	120	350
92	353	126	346
137	345	121	345
114	351	126	348
104	348	133	350
112	350	133	349
113	348	125	347
120	346	121	349
138	344	121	347
111	347	125	345
117	350	126	346
110	346	122	350
109	349	111	353
110	351	105	349
118	346	129	342
107	350	110	351
116	348	106	350
130	347	127	347
117	348	104	349
114	347	132	347
122	347	123	349
111	350	114	349
126	346	136	346
118	347	127	345
108	348	140	349
117	349	119	351
107	346	116	350
134	346	132	346
113	347	130	346
98	349	117	347
115	346	121	347
136	343	108	354
131	346	106	350
114	348	115	349
107	346	139	345

Tabela A.5

salario	experiencia	sexo
1,9307	0	0
3,1769	17	0
2,2769	5	0
3,1307	15	0
2,7769	9	0
3,0923	15	0
2,6538	8	0
2,223	5	0
2,8538	13	0
3,2307	20	0
2,823	11	0
1,9076	1	0
2,5384	6	0
2,5692	7	1
4,223	23	1
4,0923	20	1
3,6	18	1
4,7076	27	1
3,1461	11	1
2,9923	10	1
4,7461	29	1
4,1153	23	1
2,3615	4	1
4,0923	22	1
4,5076	25	1
2,9076	9	1
4,4846	25	1

Tabela A.6

y	x1	x2
12	31	4
13	16	5
3	29	3
3	19	0
11	27	2
19	21	6
1	24	2
14	11	3
15	26	6
17	18	6
2	12	1
15	3	5

Tabela A.7

Dia	Y.7	X.7	Z.7
1	50	1	0
2	110	2	0
3	90	2	0
4	150	3	0
5	140	3	0
6	180	3	0
7	190	4	1
8	310	6	0
9	330	6	0
10	340	7	1
11	360	8	3
12	380	10	6
13	360	10	6

Tabela A.8

Y.9	X.9
81	3
73	3
72	3
91	5
99	5
127	9
114	9
116	9
123	9
131	9
141	11
151	12
147	12
131	12
145	12
147	13
179	15
166	15
181	15
178	15
185	15
156	15
173	16
189	16
192	17
203	19
192	19
219	19
214	19

Tabela A.9

Pulse1	Pulse2	Smokes	Sex	Height	Weight	Activity
64	88	2	1	66	140	2
58	70	2	1	72	145	2
62	76	1	1	73,5	160	3
66	78	1	1	73	190	1
64	80	2	1	69	155	2
74	84	2	1	73	165	1
84	84	2	1	72	150	3
68	72	2	1	74	190	2
62	75	2	1	72	195	2
76	118	2	1	71	138	2
90	94	1	1	74	160	1
80	96	2	1	72	155	2
92	84	1	1	70	153	3
68	76	2	1	67	145	2
60	76	2	1	71	170	3
62	58	2	1	72	175	3
66	82	1	1	69	175	2
70	72	1	1	73	170	3
68	76	1	1	74	180	2
72	80	2	1	66	135	3
70	106	2	1	71	170	2
74	76	2	1	70	157	2
66	102	2	1	70	130	2
70	94	1	1	75	185	2
96	140	2	2	61	140	2
62	100	2	2	66	120	2
78	104	1	2	68	130	2
82	100	2	2	68	138	2
100	115	1	2	63	121	2
68	112	2	2	70	125	2
96	116	2	2	68	116	2
78	118	2	2	69	145	2
88	110	1	2	69	150	2
62	98	1	2	62,75	112	2
80	128	2	2	68	125	2

Tabela A.10

Y.11	X.11	Z.11
2,98	30,5	21,5
5,36	30,5	22
7,06	30,5	22,5
10,17	30,5	23
2,6	31	21
2,4	31	21
2,62	31	21,5
6,92	31	22,5
6,19	31,5	22
17,32	31,5	24
15,6	31,5	24
16,12	31,5	24