UNIVERSIDADE EDUARDO MONDLANE
Faculdade de Ciências
Departamento de Matemática e Informática

Trabalho de Licenciatura em Estatística

Tratamento de dados com Outliers

Uma aplicação à análise de Regressão

Autor: Eurico Fernandes José Cumbi

UNIVERSIDADE EDUARDO MONDLANE
Faculdade de Ciências
Departamento de Matemática e Informática

Trabalho de Licenciatura em Estatística

Tratamento de dados com Outliers

Uma aplicação à análise de Regressão

Autor: Eurico Fernandes José Cumbi
Supervisor: Dr. Alberto Mulenga

Maputo, Fevereiro de 2010

DECLARAÇÃO

Declaro que este trabalho é resultado da minha própria investigação, que não foi submetido para outro grau que não seja o indicado ï Licenciatura em Estatística, da Universidade Eduardo Mondlane.

Maputo, aos 11 de Fevereiro de 2010

O Autor

_____

(Eurico Fernandes José Cumbi)

AGRADECIMENTOS

Ao apresentar o presente trabalho não posso deixar de expressar os meus agradecimentos a todos que, de algum modo, contribuíram para que o presente trabalho se concretizasse.

- ✓ Aos docentes, colegas e os demais pelo incentivo e encorajamento que sempre me proporcionaram em todos os aspectos e em todos os momentos, o meu muito obrigado.

- ✓ Um agradecimento especial ao meu supervisor Dr. Alberto Mulenga pelo permanente acompanhamento, encorajamento e muita paciência dedicada na elaboração do presente trabalho.

- ✓ Aos meus amigos, meus irmãos e minha namorada Vanícia que sempre me encorajaram e respeitaram o tempo que eu dedicava à elaboração do presente, vai o meu muito obrigado.

- ✓ Finalmente, o meu agradecimento aos meus pais José Cumbi e Anastácia Bucuane pela compreensão, estímulo e apoio incondicional sempre manifestados.

RESUMO

Na análise de dados é frequente encontrar observações com características diferentes das demais e que parecem não pertencer ao padrão de variabilidade formado por outras observações, estas são chamadas de *Outliers*. A presença de outliers provoca distorção das estimativas tais como a média, o desvio padrão, etc. Por isso é importante detecta-los e procurar formas de corrigir ou minimizar os efeitos provocados pela sua presença na amostra. Assim, no presente trabalho pretende-se estimar modelos de regressão linear cujos dados amostrais possuem outliers. Para tal usou-se a base de dados TEMPENTREGA e estimou-se um modelo de regressão usando o método dos Mínimos Quadrados Ordinários (MQO), em seguida, para detectar a presença de outliers usou-se os resíduos estudantizados e os valores da diagonal principal da matriz de projecção e para avaliar se os outliers encontrados são ou não influentes no modelo de regressão usaram-se as medidas de influência da Distância de Cook, o SDFFITS e o SDFBETA, com os quais detectou-se duas observações como outliers influentes. Para minimizar o efeito das observações outliers usou-se dois métodos: o método da regressão robusta dos Mínimos Quadrados Aparados (MQA) e o método dos MQO sem as observações outliers. Do estudo concluiu-se que os modelos de regressão dados pelos métodos dos MQA e MQO sem as observações outliers eram os melhores para fazer previsões pois apresentavam melhores estimativas.

Palavras-chave: Outliers, Resíduos, Medidas de influência, Modelo de regressão.

ÍNDICE

LISTA DE FIGURAS E TABELAS

Figuras

Tabelas

LISTA DE ABREVIATURAS

DW: Durbin-Watson;

FIV: Factor de Inflação da Variância;

K-S: Kolmogorov-Smirnov;

MMQ: Mínima Mediana dos Quadrados;

MQA: Mínimos Quadrados Aparados;

MQO: Mínimos Quadrados Ordinários;

p_value: valor da probabilidade ou probabilidade exacta de cometer o erro do *tipo I*;

SQE: Soma dos Quadrados Explicados pela regressão;

SQR: Soma dos Quadrados dos Resíduos;

TEMPENTREGA: Tempo de Entrega.

I INTRODUÇÃO

Na análise de dados estatísticos é frequente encontrar observações discrepantes, normalmente designados *Outliers*, que segundo Draper e Smith (1998), são observações que em valor absoluto são muito distantes das demais observações na amostra. Para Johnson e Wichern (2007), outliers são observações estranhas que parecem não pertencer ao padrão de variabilidade produzido por outras observações; estas observações parecem ser inconsistentes com o restante das observações.

Normalmente, a real causa da existência dos outliers é desconhecida pelos analistas e utilizadores dos dados. Segundo Last e Kandel (2001), algumas vezes, um outlier é um valor erróneo, resultante da má qualidade do conjunto de dados, isto é, erro no registo/introdução ou conversão de dados; medições físicas, especialmente quando realizado com equipamento em mau estado, que pode produzir uma certa quantidade de dados distorcidos. Nestes casos nenhuma informação útil é trazida pelo outlier. Contudo, é também possível que um outlier constitua informação correcta, mas excepcional.

A presença de outliers nos dados de uma amostra provoca uma distorção nas estimativas, como é o caso da média aritmética, do desvio padrão (visto que estas sofrem maior influência dos valores extremos), nos modelos de regressão (que afecta a estimativa do coeficiente de correlação), o que pode provocar a não fiabilidade dos resultados nas inferências.

Assim, numa amostra com outliers colocam-se as seguintes perguntas:

- Quais os métodos usados para detectar a presença dos outliers?
- Quais as possíveis origens dos outliers? e,
- Quais as possíveis soluções para reduzir o efeito provocado pelos outliers?

Com base nas questões colocadas acima, cujas respostas são de grande importância tanto para os académicos como para os analistas de dados em geral, surgiu o interesse para a escolha do tema do presente trabalho de investigação. Como forma de procurar responder estas questões, são traçados os seguintes objectivos da investigação.

Objectivos

Objectivo geral

Estimar modelos de regressão linear cujos dados amostrais possuem outliers

Objectivos específicos

- Descrever os métodos utilizados para detectar outliers;
- Identificar as possíveis causas da presença de outliers;
- Estimar modelos de regressão linear com menor efeito possível dos outliers.

De modo a alcançar os objectivos acima traçados o presente trabalho aborda nos próximos capítulos a Revisão da Literatura onde são tratados os aspectos teóricos relacionados com o tratamento de dados com outliers, os Materiais e Métodos onde é indicada a metodologia usada para a análise dos dados, os Resultados e Discussões e as conclusões alcançadas após a análise dos resultados.

II REVISÃO DA LITERATURA

2.1 Conceitos e Definições

A análise de dados impõe o uso de técnicas estatísticas, sendo que para obter bons resultados os dados devem ter qualidade e deve-se também verificar alguns pressupostos tais como a *normalidade, a homocedasticidade e a linearidade*. A qualidade dos dados assim como os pressupostos são verificados fazendo uma análise exploratória aos dados antes da aplicação das técnicas.

O pressuposto da normalidade diz que as variáveis na análise devem ser normalmente distribuídas enquanto que o pressuposto da homocedasticidade diz que as variáveis dependentes devem ter variâncias iguais ao longo do domínio das variáveis independentes e o pressuposto da linearidade diz que as variáveis dependentes devem possuir relações lineares com as variáveis independentes.

A qualidade dos dados é geralmente afectada por *não respostas*, que consiste em informação não disponível de um indivíduo sobre o qual outra informação está disponível, e as observações *discrepantes* (geralmente designados por *outliers*). Assim, o presente trabalho aborda um dos problemas que afecta a qualidade dos dados, os *outliers*.

Outliers, segundo Draper e Smith (1998), são observações que em valor absoluto são muito distantes das demais observações na amostra. Esta definição embora correcta, tem o inconveniente de restringir o campo de abrangência; tem mais sentido para dados univariados e não para dados bivariados e multivariados.

Outliers, para Johnson e Wichern (2007), são observações estranhas que parecem não pertencer ao padrão de variabilidade produzido por outras observações e segundo Hair et al (2005), são observações com uma combinação única de características identificáveis como sendo notavelmente diferentes das outras observações. Estas definições abrangem não somente o sentido univariado dos dados mas também o sentido bivariado e o multivariado.

Como forma de ilustrar as definições acima, tem-se a figura 2.1, que é um diagrama de dispersão. O diagrama tem uma observação, envolvida por um círculo, que se encontra distante do padrão formado pelo restante das observações, podendo ser um outliers[1]. Analisando de forma univariada tanto em relação a variável *x* assim como a variável *y* esta observação não é mínima nem máxima, mostrando assim que a definição dada por Draper e Smith (1998) não é muito útil para casos bivariados assim como multivariado.

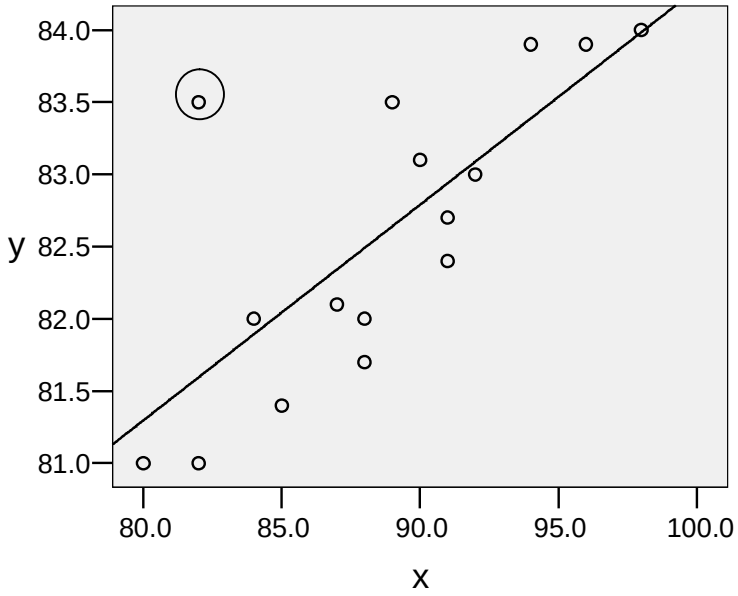


Figura 2.1: Representação de dados num espaço bidimensional com uma observação afastada

Para Dielman (2001) e Hair et al (2005), os outliers podem ser bons, quando fornecem informação sobre o comportamento em estudo que de contrário não seria acessível, e maus, quando não são representativos da população e podem distorcer seriamente os resultados dos testes.

Maus outliers podem ser devidos a erro no registo/introdução ou codificação/conversão de dados, por exemplo, numa amostra com dados referentes a idade, em anos, de alunos de uma escola primária, se se encontrar uma observação com a idade 78 anos, fica-se claro de que este não é um valor correcto, podendo ser um erro de introdução de dados.

[1] É necessário um teste formal para se poder afirmar com certeza de que este é um outlier.

Bons outliers podem ser devidos a eventos excepcionais, podendo ser o objecto principal da investigação, por exemplo, numa amostra com dados referentes ao tempo de internamento pós-parto das mulheres num hospital, onde o normal é ter um tempo de permanência de 2 à 3 dias, aparecer uma observação que indica 45 dias. Esta observação pode representar um valor correcto e dever-se a complicações de parto.

Um outro exemplo de outlier, é o de uma amostra de salários de trabalhadores de uma empresa, onde o normal é encontrar salários entre 5 000.00 à 18 000.00 meticais. Se se encontrar uma observação com o salário 60 000.00 meticais pode-se considerar como outlier. Para este caso, a priori não se sabe se este é um valor erróneo ou verdadeiro, podendo-se consultar à fonte dos dados para confirmar a sua veracidade, mas se a fonte dos dados não estiver acessível, torna-se difícil decidir a natureza deste evento.

A presença de outliers provoca distorção das estimativas, como é o caso da estimativa da média e do desvio padrão, que são facilmente enviesados por valores extremos, as estimativas dos Mínimos Quadrados Ordinários (MQO), que são afectadas por resíduos com maior valor absoluto, etc (Last e Kandel, 2001).

2.2 Detecção dos Outliers

Os outliers podem ser detectados segundo uma perspectiva *univariada*, *bivariada* e *multivariada*, sendo que deve-se usar tanto quanto possível essas perspectivas de forma a detectar de maneira consistente estas observações.

Detecção univariada

Nesta perspectiva, os outliers podem ser detectados fazendo um exame visual da distribuição de cada variável através de tabelas de frequência, gráficos como o histograma, diagrama de pontos e diagrama de caule-e-folha (Last e Kandel, 2001; Manly, 2000). Contudo, Last e Kandel (2001) alertam ao facto de que este tipo de detecção tem como desvantagens a *subjectividade* e a *ineficiência*, pois o analista dos dados deve aplicar sua própria percepção subjectiva para determinar se uma observação está ñmuitoò ou ñpoucoò afastada do padrão de variabilidade produzido por outras observações, assim, como o exame visual dos gráficos é uma tarefa extremamente demorada quando o número de variáveis é elevado.

Outro tipo de detecção gráfica, que é mais formal, é o uso do *box-plot* (caixa de bigodes), o qual é constituído por mediana ou 2° quartil (Q_2), 1° quartil (Q_1), 3° quartil (Q_3), o valor mínimo, dado pela expressão: $Q_1 - 1.5*(Q_3 - Q_1)$, e o máximo, dado pela expressão $Q_3 + 1.5*(Q_3 - Q_1)$. Assim, são considerados como outliers as observações com valores que se encontram abaixo do mínimo e acima do máximo (Triola, 1999; Last e Kandel, 2001; Silvestre, 2007). Para estes autores, as observações x_i com valores que se encontram nos intervalos dados pela expressão (2.1) são consideradas outliers *suaves* e são consideradas outliers *extremos* as observações x_i que se encontram nos intervalos dados pela expressão (2.2).

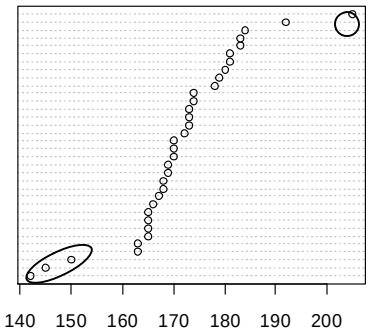
$$(Q_1 - 3IQ) < x_i < (Q_1 - 1.5IQ) \qquad \text{ou} \qquad (Q_3 + 1.5IQ) < x_i < (Q_3 + 3IQ) \tag{2.1}$$

$$x_i < (Q_1 - 3IQ) \qquad \text{ou} \qquad x_i > (Q_3 + 3IQ) \tag{2.2}$$

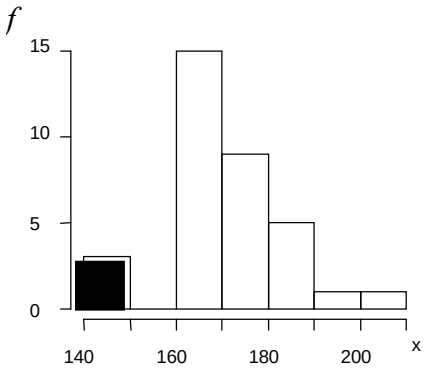
onde:

$IQ = Q_3 - Q_1$

Na figura 2.2 estão algumas representações gráficas de uma amostra hipotética referente a alturas de indivíduos. Na figura 2.2 (a), no diagrama de pontos, são consideradas possíveis outliers as primeiras três observações e a última, envolvidas por círculos, dado que estas observações encontram-se distantes do padrão formado por outras observações. Na figura 2.2 (b) temos um histograma onde o primeiro grupo de observações, entre 140-150 é constituído por três observações as quais encontram-se também distante das demais, sendo assim consideradas possíveis outliers. Na figura 2.2 (c), o diagrama de caule-e-folha, apresenta também três observações diferentes das demais em valor absoluto que são 142, 145 e 205, a negrido no diagrama. Na figura 2.2 (d), no box-plot, são notavelmente diferentes das demais observações os valores 142 e 205, sendo estes outliers suaves.



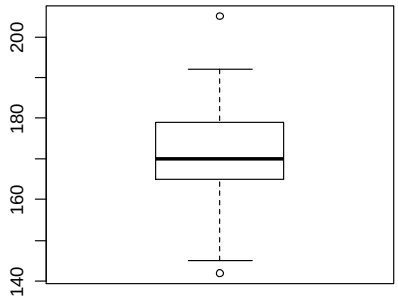
(a) Diagrama de pontos



(b) Histograma

14 | **25**
15 | 0
16 | 335555678899
17 | 00023334489
18 | 011334
19 | 2
20 | **5**
Legenda: 14|2 significa 142

(c) Diagrama de caule-e-folha



(d) Box-plot

Figura 2.2: Representação gráfica de uma amostra de alturas de indivíduos com observações outliers

Da análise gráfica, pode-se notar que as observações com os valores 142, 145 e 205 são as que mais aparecem como notavelmente diferentes das demais observações. Assim pode-se considerar estas observações como outliers nesta amostra.

Segundo Hair et al (2005), um outro método usado para detecção de outliers consiste em estandardizar os valores da variável *x* para *z*, onde a média passa para o valor 0 (zero) e o desvio padrão para 1 (um). Assim, considera-se outlier se o valor estandardizado *z*, em valor absoluto, estiver acima de 2.5, para amostras de tamanho menores que 80, ou 3 à 4, para amostras de tamanhos maiores que 80.

Detecção bivariada e multivariada

De acordo com Johnson e Wichern (2007), para certas situações de análise de dados usando procedimentos univariados alguns outliers não são detectados, porém podem ser detectados se for realizado um exame tomando em conta a associação entre duas ou mais variáveis.

Um método bastante usado para a detecção bivariada de outliers é o uso do diagrama de dispersão, onde são consideradas como sendo outliers as observações que se encontram distantes do padrão de variabilidade produzido pelas restantes observações (exemplo: figura 2.1). Porém, este método tem a desvantagem da subjectividade e ineficiência anteriormente descrita.

Johnson e Wichern (2007) e Silvestre (2007) propõem outro método mais formal usado para detectar casos que podem não ter sido detectados por métodos univariados, neste caso a *Distância de Mahalanobis*.

Para o cálculo da *Distância de Mahalanobis* toma-se em consideração as correlações entre as variáveis, isto é, pondera os quadrados das diferenças usando as covariâncias das variáveis, conforme ilustra a expressão (2.3).

$$D^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \tag{2.3}$$

onde:

x_i - i-ésimo vector das variáveis x com a observação i ou $x_i' = (x_{1i}, x_{2i}, \ldots, x_{ki})$;

$\bar{x}$ - vector das médias das variáveis x;

S ï matriz das covariâncias das variáveis x.

Segundo Johnson e Wichern (2007), se a população a ser considerada está distribuída de forma normal multivariada, então a *Distância de Mahalanobis* segue a distribuição de *Qui-quadrado*, com k graus de liberdade.

Uma variante da *Distância de Mahalanobis* é a *Distância Euclidiana Estandardizada* dada pela expressão (2.4).

$$d_i^2 = \left(x_i - \bar{x}\right)' \mathbf{D}^{-1} \left(x_i - \bar{x}\right) \quad (2.4)$$

onde:

D: é matriz das variâncias em que na diagonal principal tem-se as variâncias das variáveis e os restantes elementos da matriz são zeros

A desvantagem da *Distância Euclidiana Estandardizada* é que ela não toma em consideração as correlações entre as k variáveis.

2.3 Tratamento de Outliers

Após identificar os outliers, surge a importante questão de difícil resposta: *o que fazer com tais observações?* Não existe um método único para a solução deste problema.

Os investigadores e analistas dos dados têm em mente de que incluir observações discrepantes na análise distorce os resultados, assim, uma das primeiras opções tem sido excluir estas observações da análise. Por outro lado, excluir uma ou mais observações implica reduzir o tamanho da amostra, o que pode ser inadequado para algumas análises, porque a menores tamanhos da amostra as generalizações dos resultados são menos precisas.

Segundo Harper (1991), incluir erradamente um outlier na análise, é considerado um erro mais sério do que excluir erradamente esta observação. Mas Hair et al (2005) sustentam que estas observações devem ser mantidas, a menos que exista uma prova evidente de que estão verdadeiramente fora do normal e que não são representativas da população.

Se um outlier é assumido como uma observação com um valor completamente errado (por exemplo, valores muito fora do intervalo), pode-se estimar um novo valor usando métodos de estimação das ñ*não respostas*ò (Last e Kandel, 2001). Neste caso estar-se-á a considerar o outlier como um valor não obtido no processo de recolha de dados ou omitido no registo/introdução dos dados.

Entre os métodos da estimação das *não respostas* que podem ser usados para estimar um novo valor em substituição do outlier destacam-se os seguintes[2]:

- **Substituição por um caso**: consiste em substituir a observação por uma outra obtida fora da amostra;
- **Substituição pela média**: consiste em substituir o valor em causa pela média da variável;
- **Atribuição por regressão**: uso da análise de regressão para estimar o valor que é atribuído a posição do outlier com base na relação entre as variáveis; e
- **Atribuição múltipla**: consiste na combinação de vários métodos e geralmente usa-se a média das várias estimativas.

O método de substituição por um caso tem a vantagem de poder-se escolher um caso fora da amostra com as características semelhantes às do caso a substituir mas com um valor absoluto dentro dos padrões dos outros valores da amostra, porém a escolha do novo caso não garante a aleatoriedade.

O método da substituição pela média é amplamente usado, devido a sua vantagem de ser fácil de implementar, porém apresenta a desvantagem de subestimar a verdadeira variância dos dados, distorcer a real distribuição de valores e comprimir as correlações entre as observações.

O método da atribuição por regressão tem a desvantagem de reforçar as relações existentes, subestimar a variância da distribuição e pressupõe relações substanciais entre a variável com o valor outlier e as outras variáveis, e os valores previstos podem não estar nos intervalos válidos da variável. Contudo, quando as relações entre as variáveis são suficientemente estabelecidas este método não influencia bastante a generalidade dos resultados.

O método da atribuição múltipla procura minimizar as desvantagens de um método em particular, sendo que a composição dos métodos fornece a melhor estimativa.

[2] Adaptado de Hair et al (2005)

Dependendo da natureza do outlier e do objectivo da investigação, o investigador deve procurar encontrar o método que mais se adequa a situação.

2.4 Observações Discrepantes em Regressão

2.4.1 Conceitos de observações discrepantes em regressão

O método dos Mínimos Quadrados Ordinários é um método amplamente usado para a estimação dos modelos de regressão linear e está disponível na maioria dos pacotes estatísticos.

O método MQO determina as estimativas de regressão de modo que o somatório dos quadrados dos resíduos seja mínimo possível, ou seja

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \rightarrow \quad \min imo \tag{2.5}$$

onde:

y_i - é o valor real da variável dependente;

$\hat{y}_i$ - é o valor estimado da variável dependente.

Assim, este método dá, proporcionalmente, mais peso às observações extremas no somatório dos quadrados dos erros, o que significa que as estimativas dos MQO são muito sensíveis a observações extremas.

Hair et al (2005) e Mukherjee et al (1998) classificam as observações extremas em relação a recta de regressão como *outliers, leverage* e *influentes*, e as definem do seguinte modo:

Outliers, no contexto de regressão, são observações com grandes valores residuais, sendo que podem ser identificados apenas em relação a um modelo específico de regressão.

Leverages são observações diferentes das demais em relação aos valores das variáveis independentes. Seu impacto é particularmente perceptível nos coeficientes estimados para uma ou mais variáveis independentes.

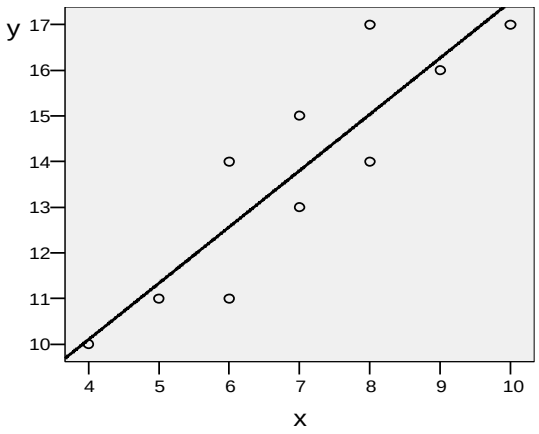
A diferença entre outlier e leverage é que o outlier é definido como um ponto dos dados com uma distância vertical excepcionalmente larga da linha de regressão, enquanto que a definição de leverage não envolve totalmente a linha de regressão, pois altos leverages apenas referem a um ponto que é desproporcionalmente distante dos outros pontos na direcção do eixo *X*. Pode-se também considerar um ponto com alto leverage numa regressão de uma variável *Y* por uma variável *X* como um ñoutlierò na distribuição da variável *X*.

Uma observação é considerada **influente** se ao remove-la da amostra muda consideravelmente a posição da linha de regressão. Portanto, as observações influentes puxam a linha de regressão em sua direcção.

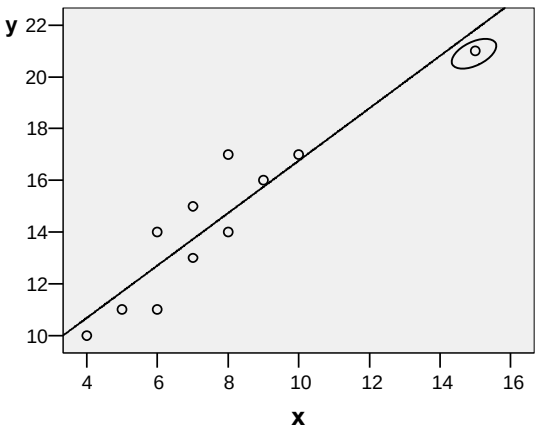
As observações influentes são a categoria mais ampla pois, incluem os outliers e os leverages. Contudo, deve-se realçar que nem todos outliers e leverages são necessariamente observações influentes.

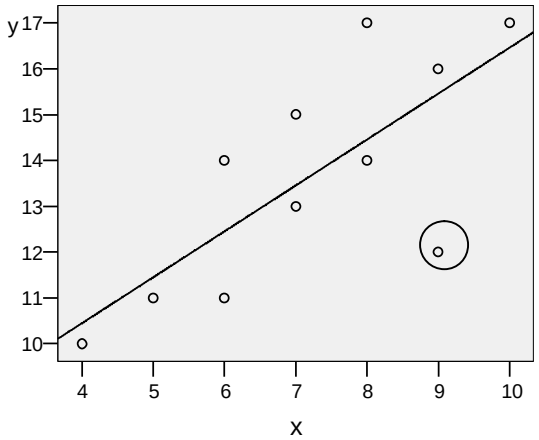
A figura 2.3 é a representação gráfica do efeito da presença de um outlier ou um leverage numa linha de regressão. A figura 2.3 (a) representa o ajustamento da linha de regressão sem valores extremos, cujos dados encontram-se na tabela ao lado, onde *x* é a variável independente e *y* a variável dependente.

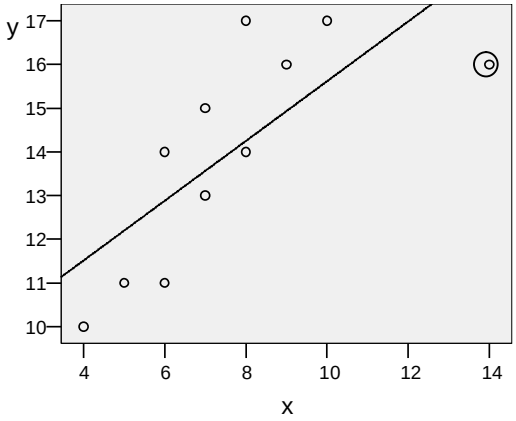
Na figura 2.3 (b) incluiu-se a observação (15, 21), a qual é um leverage, mas não é necessariamente uma observação influente, pois tem pouco efeito na inclinação da linha de regressão, apenas estende a amplitude da aplicação da análise, o que pode não ser ideal para as inferências. Na figura 2.3 (c) incluiu-se a observação (9, 12), que é um outlier, o qual tem também pouco efeito na inclinação da linha de regressão, não sendo assim uma observação influente. Na figura 2.3 (d) incluiu-se a observação (14, 16), a qual é um leverage. Esta observação fez com que a linha de regressão mudasse bastante a inclinação, o que indica que esta é uma observação influente. É também um outlier influente a observação (4, 16), que se encontra na figura 2.3 (e), pois também mudou bastante a inclinação inicial da linha de regressão.



X	Y
4	10
5	11
6	11
6	14
7	13
7	15
8	14
8	17
9	16
10	17







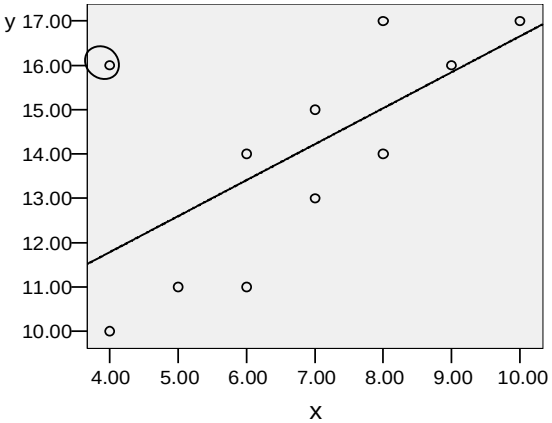


Figura 2.3: Representação gráfica do efeito de outlier e leverage numa linha de regressão[3]

Na tabela 2.1 está o resumo do comportamento dos coeficientes e do somatório dos quadrados dos resíduos de regressão para cada situação ilustrada pelos gráficos da figura 2.3.

[3] Adaptado de Mukherjee et al (1998)

Tabela 2.1: Resumo do comportamento do modelo na presença de valores discrepantes

Modelo	B_0	B_1	SSE	ponto	classificação
(a)	5.167	1.233	11.967	ï	ï
(b)	6.624	1.013	14.166	(15, 21)	Leverage não influente
(c)	6.435	1.003	26.727	(9, 12)	Outlier não influente
(d)	8.776	0.684	27.109	(14, 16)	Leverage influente
(e)	8.538	0.812	36.831	(4, 16)	Outlier influente

2.4.2 Identificação de outliers e leverage

A definição de outlier, em regressão, toma como base o valor do resíduo, sendo assim o principal meio de identificação de outliers.

O *resíduo* é a diferença entre o valor observado da variável dependente y_i e o se valor estimado $\hat{y}_i$, ou:

$$e_i = y_i - \hat{y}_i \tag{2.6}$$

onde:

e_i - é o resíduo.

Uma outra forma de resíduo é o *resíduo eliminado* $e_{i(i)}$, onde é eliminada a observação *i* quando se estima a equação de regressão usada para calcular o valor estimado da mesma observação, mas a forma de cálculo é a mesma com a dada na expressão 2.6. Assim, se elimina o impacto no seu próprio valor estimado.

Segundo Hair et al (2005), para se detectar um outlier constrói-se um diagrama de dispersão entre (y_i, e_i) ou $(y_i, e_{i(i)})$, onde serão consideradas outliers as observações com valores de resíduo elevados. Contudo, tanto os resíduos na sua forma normal assim como os resíduos eliminados estão na escala da variável dependente a qual é útil na interpretação, mas não dão a ideia de quão elevado é o valor do resíduo para se considerar a observação como outlier.

Uma forma de resolver as limitações dos resíduos na sua forma normal e os resíduos eliminados é o uso dos *resíduos estandardizados*, que segundo Dielman (2001), são obtidos mediante a divisão do resíduo na forma normal pelo erro padrão dos resíduos, dado pela expressão 2.7.

$$e_{is} = \frac{y_i - \hat{y}_i}{s} \quad \text{ou} \qquad e_{is} = \frac{e_i}{s} \tag{2.7}$$

onde:

e_{is} - é o resíduo estandardizado

s - é o erro padrão dos resíduos

Os resíduos estandardizados tem como média o valor 0 e desvio padrão 1. Assim, se a amostra está distribuída normalmente ou se o tamanho da amostra é maior ou igual a 50, então estes resíduos seguem a distribuição *t* de Student.

Para Mukherjee et al (1998), os resíduos estandardizados sofrem do facto de, se existir um outlier na amostra, este irá distorcer o erro padrão da regressão, podendo assim não ser detectado o outlier. Assim, uma forma de prevenir esta situação é o uso dos *resíduos estudantizados*, onde a i-ésima observação é omitida na determinação do erro padrão usado para estandardizar o resíduo. O resíduo estudantizado é dado pela seguinte expressão[4]:

$$t_i = \frac{e_i}{s_{(i)}\sqrt{(1-h_i)}} \qquad \text{para regressão simples} \tag{2.8.a}$$

ou

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \qquad \text{para regressão múltipla} \tag{2.8.b}$$

onde:

t_i - é o resíduo estudantizado;

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \tag{2.8.c}$$

ou

[4] Fonte: Draper e Smith (1998)

$h_{ii} \in H$ sendo H - *matriz de projecção*, $H = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix}$, dada por:

$$H = X\left(X'X\right)^{-1}X' \qquad (2.8.d)$$

X - é a matriz das variáveis independentes,

$s_{(i)}$ - é o erro padrão dos resíduos sem o impacto da observação *i*. Este é dado por,

$$s_{(i)} = \left[\frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}\right]^{1/2} \quad \text{e} \quad s = \left[\frac{\sum e_i^2}{n-p}\right]^{1/2}$$

p - é o número de variáveis independentes.

Os resíduos estudantizados seguem a distribuição *t* de *student* com $n-p-1$ graus de liberdade.

Para detectar se uma observação é ou não um leverage, basta avaliar o valor de h_i ou h_{ii}, o qual segundo Hair et al (2005), tem como limites aceitáveis de $3p/n$ (para $p<10$ *ou* $n<50$) ou $2p/n$ (para $p>10$ *ou* $n \geq 50$).

2.4.3 Identificação de observações influentes

De acordo com a figura 2.3, tanto os outliers como os leverages podem ou não exercer influência nos resultados de regressão, tornando assim necessário medir a sua influência. Assim, as medidas usadas para medir a influência de uma dada observação nos resultados de regressão são, nomeadamente o *DFBETA*, *DFFITS*, *Rácio de Covariância* e *Distância de Cook*.

(i) ***DFBETA***

Segundo Mukherjee et al (1998), esta estatística mede a sensibilidade em cada coeficiente de regressão com a eliminação da observação *i*, isto é, se a eliminação desta observação leva a

uma mudança drástica em cada coeficiente significa que esta observação é influente. Esta medida é dada pela expressão 2.9.

$$DFBETA_i = \frac{b_i - b_{i(i)}}{s_{b(i)}} \tag{2.9}$$

onde: $b_{i(i)}$ *e* $s_{b(i)}$ são o coeficiente e o erro padrão da estimativa do coeficiente de regressão depois da eliminação da observação *i*.

Hair et al (2005), Pestana e Gageiro (2005), sugerem uma versão estandardizada do *DFBETA*, que é o *SDFBETA*, o qual tem como ponto de corte $\pm\sqrt{n}$.

(ii) ***DFFITS***

Segundo Hocking (1996), esta estatística mede o impacto no valor estimado da observação *i*, causado pela eliminação deste caso e é dada por:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{s_{(i)}^2 h_{ii}}} \tag{2.10}$$

onde:

$\hat{Y}_i$ - é o valor estimado da observação *i*

$\hat{Y}_{i(i)}$ - é o valor estimado da observação *i*, depois da eliminação da observação *i*.

Esta medida, segundo Hair et al (2005), Pestana e Gageiro (2005), tem também uma versão estandardizada, o *SDFFITS*, o qual tem como valor de corte $2\sqrt{(p+1)/(n-p-1)}$.

(iii) ***Rácio de Covariância*** *(COV)*

Segundo Hair et al (2005), esta estatística mede a influência da eliminação da observação *i* nos erros padrão dos coeficientes de regressão. Esta toma valores elevados se h_{ii} for elevado e valores baixos se t_i for baixo. Assim, a observação *i* é considerada influente se $COV < 1-3p/n \quad ou \quad COV > 1+3p/n$.

(iv) ***Distância de Cook*** (D_i)

Segundo Hair et al (2005), a distância de Cook mede o impacto da observação *i*, tanto no tamanho das variações nos valores estimados quando a observação é eliminada, como a distância entre a observação em causa das outras observações. Esta também é tida como a medida mais representativa da influência sobre o ajuste geral. Ela é dada pela expressão 2.11 e tem como ponto de corte o valor $4/(n-p-1)$.

$$D_i = \left(\frac{e_i}{s(1-h_{ii})^{1/2}} \right)^2 \left(\frac{h_{ii}}{1-h_{ii}} \right) \frac{1}{p} \tag{2.11}$$

Hair et al (2005) observam que o processo de identificação de observações influentes requer o uso de vários métodos, procurando a convergência e consistência nos resultados, pois nenhuma medida representa totalmente todas as dimensões de influência.

Mukherjee et al (1998) alertam ao facto de que é possível que várias observações influentes tenham formado um grupo e conjuntamente puxar a linha de regressão em sua direcção. Nesta situação, a eliminação de uma por uma observação no cálculo das medidas não revela grande influência, daí que faz sentido a observação acima feita por Hair et al (2005). Nestes casos, se possível, é útil o uso de diagramas de dispersão para detectar tais casos e no cálculo as medidas de influência, ao invés de eliminar uma a uma observação, convém eliminar duas a duas ou mesmo três a três.

2.4.4 Tratamento de observações discrepantes em regressões

Segundo Dielman (2001), existem várias razões para que uma observação seja discrepante. Para este autor, se se tiver violado o pressuposto da *linearidade* ou da *homocedasticidade* pode provocar algumas observações discrepantes, que podem ser corrigidas escolhendo transformações adequadas aos dados; se for um erro de codificação ou introdução de dados, deve-se corrigir ou substituí-los pelos dados correctos. Porém, se não forem encontrados os valores correctos para a substituição, o mais adequado é excluir este tipo de observação da análise.

Se a observação discrepante não for devida à violação de algum pressuposto ou erro de codificação/introdução de dados e ser apenas uma observação estranha relativamente às outras observações da amostra, torna-se difícil ajuizar o problema. Em tais casos, Hair et al (2005), sugere o uso de técnicas de estimação mais *robustas*, entre elas a **regressão robusta**.

Qualquer que seja a medida a tomar deve-se ter como objectivo tornar os dados mais representativos da população de forma a garantir a validade e a generalização dos resultados.

Alguns conceitos da regressão robusta

A regressão robusta é uma ferramenta importante para analisar dados, pois fornece resultados estáveis mesmo na presença de observações discrepantes. Os métodos de regressão robusta permitem que as observações sejam ponderadas de forma desigual, isto é, as observações que produzem elevados resíduos recebem menores pesos no cálculo das estimativas de regressão tais como os erros padrão dos coeficientes e o erro padrão dos resíduos.

Um conceito importante quando se trabalha com métodos robustos é o *ponto de falha* que, segundo Rousseeuw e Leroy (2003), é o valor da proporção da contaminação por valores discrepantes à partir do qual o método começa a produzir estimativas destorcidas. Por exemplo, o método dos MQO tem como ponto de falha o valor $1/n$, o qual em amostras grandes tende para zero, o que reflecte a sua sensibilidade perante a presença de observações discrepantes.

Os métodos da regressão robusta comummente usados são nomeadamente, o método do mínimo valor absoluto, normalmente designado por estimador L_1, o estimador M, o método da mínima mediana dos quadrados (MMQ), o método dos mínimos quadrados aparados (MQA), o estimador MM e o estimador S.

O estimador L_1 minimiza o somatório dos valores absolutos dos resíduos, ou seja:

$$\min_{\beta} \quad \sum_{i=1}^{n} | e_i |$$

Este estimador é robusto em relação a outliers, sendo assim preferível em relação ao método MQO. Contudo, este não é robusto em relação aos leverages, sendo que na presença deste tipo de observação ele tem como ponto de falha o valor $1/n$.

O estimador *M* minimiza a soma da função dos resíduos ρ, ou seja:

$$\min_{\beta} \sum_{i=1}^{n} \rho(e_i)$$

onde ρ é uma função quadrática, simétrica e com um único mínimo em zero. Derivando esta expressão em relação aos coeficientes de regressão β_i, obtém-se um sistema de *p* equações, dado por:

$$\sum_{i=1}^{n} \psi(e_i) x_i = 0$$

o qual sua solução é obtida por métodos iterativos designados por mínimos quadrados reponderados. O estimador *M* é estatisticamente mais eficiente que o estimador L_1, mas também não é robusto em relação a presença de leverages.

O estimador MMQ minimiza a mediana dos resíduos com respeito aos elementos de β, que é o vector dos coeficientes de regressão, isto é,

$$\min_{\beta} \; mediana \; e_i^2$$

Este método procura encontrar o caminho mais curto que cubra a metade das observações. O método é robusto na presença tanto de outliers assim como na presença de leverages e pode atingir um ponto de falha de 50%.

O estimador dos MQA é semelhante ao método dos MQO pois consiste em minimizar a soma dos quadrados dos resíduos, sendo que a única diferença é que no método MQA os quadrados dos resíduos com valores absolutos muito grandes não são usados, permitindo que a linha de regressão se encontre distante dos valores extremos. O método MQA é robusto na presença tanto de outliers assim como na presença de leverages e com o ponto máximo de falha de 50%.

Os estimadores MMQ e MQA são definidos de modo a minimizar a medida robusta da dispersão dos resíduos. Contudo segundo Rousseeuw e Leroy (2003), embora ambos sejam robustos tanto para outliers como para leverages e tenham ponto máximo de falha de 50%, o método MQA tem a função objectiva mais suave, fazendo com que seja menos sensível a efeitos locais e é também estatisticamente mais eficiente.

Os estimadores S e MM são também estimadores eficientes e com um ponto de falha alto, sendo também robustos tanto em relação à presença de outliers assim como à de leverages.

Os métodos de estimação dos parâmetros do modelo de regressão usados no presente trabalho são o método dos mínimos quadrados, porque este apresenta óptimos resultados quando os pressupostos ao modelo são satisfeitos, e o método de estimação robusta dos mínimos quadrados aparados pois este é o método de estimação com melhores características nos pacotes estatísticos disponíveis. Para a detecção de outliers e leverages são usados os resíduos estudantizados e valores da diagonal principal da matriz de projecção, respectivamente e para as medidas de influência são usados o SDFFITS, o SDFBETA e a Distância de Cook.

III MATERIAL E MÉTODOS

3.1 Material

Para efeitos de aplicação são usados dados referentes ao Tempo de Entrega (TEMPENTREGA) extraídos de Rousseeuw e Leroy (2003). A base de dados TEMPENTREGA é composta por 25 observações completas, isto é, sem não respostas e três variáveis, nomeadamente:

- **NProduto**: é o número de produtos em stock no ponto de venda (x_1);
- **Distância**: é a distância em kilómetros percorrida para entregar o produto no ponto de venda pela pessoa que faz as entregas (x_2);
- **TEntrega**: é o tempo em horas requerido para a entrega do produto ao ponto de venda (y).

O processamento dos dados é feito usando os pacotes estatísticos SPSS na versão 13.0, EViews na versão 4.0 e o R na versão 2.10.0.

Na realização dos testes estatísticos é usado o nível de significância (α) de 5%. Em situações nas quais um outro nível de significância é usado é indicado no mesmo texto.

3.2 Métodos

Para a aplicação do modelo de regressão onde a variável TEntrega é a variável dependente e as variáveis NProduto e Distância são variáveis independentes é feita a exploração dos dados, a estimação e testes de hipóteses ao modelo de regressão e a avaliação do modelo estimado.

3.2.1 Métodos de exploração dos dados

Este estágio consiste em descrever as variáveis incluídas no modelo e analisar como as observações se distribuem. As medidas descritivas usadas são a média, o desvio padrão e os

valores mínimo e máximo. Para verificar como se distribuem as observações é usada a representação gráfica, através dos box-plot e diagramas de dispersão.

3.2.2 Métodos de estimação e testes de hipóteses ao modelo de regressão

O modelo de regressão usado no presente trabalho é o Modelo de Regressão Múltiplo dado pela seguinte expressão:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \text{ ou } \qquad Y = X\beta + \varepsilon \tag{3.1}$$

Onde:

Y : é o vector coluna das observações da variável dependente populacional;

β : é o vector dos parâmetros $\beta_0, \beta_1 \, e \, \beta_2$;

ε : é o vector coluna dos termos de erro.

Segundo Pestana e Gageiro (2005), para que se garanta a validade da inferência dos resultados amostrais para o universo é necessário que sejam satisfeitos os pressupostos inerentes à análise de regressão. Embora existam mais pressupostos, de acordo Hair et al (2005), os mais importantes de se verificar são os seguintes:

(i) ***Linearidade do fenómeno em estudo***: o valor médio de y para qualquer combinação específica das variáveis independentes deve ser uma função linear.

(ii) ***Normalidade***: os termos de erro ε_i devem estar normalmente distribuídos;

(iii) ***Homocedasticidade***: os termos de erro ε_i devem ter variância constante ao longo dos valores das variáveis independentes.

(iv) ***Independência dos termos de erro***: os termos de erro referentes a duas observações diferentes devem ser independentes, isto é, não devem estar correlacionados, sendo a sua covariância igual a zero.

(v) ***Independência das variáveis explicativas***: as variáveis explicativas não devem possuir uma relação linear exacta, isto é, não devem ser multicolineares.

Pelo facto de trabalhar-se com dados amostrais então o modelo de regressão é dado pela seguinte expressão:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i \tag{3.2}$$

onde: $\hat{y}$: é estimativa de $E(Y \mid X_1, X_2)$;

$\hat{\beta}_i$: é estimativa de β_i;

e: é a estimativa de ε ou simplesmente resíduo.

Os pressupostos são testados usando a estimativa do termo de erro, neste caso o resíduo. Assim, para verificar se os pressupostos ao modelo são satisfeitos é necessário antes de mais estimar o modelo para obter os resíduos e fazer os devidos testes.

O método usado para a estimação dos parâmetros do modelo de regressão é o método dos Mínimos Quadrados Ordinários, pois este apresenta bons resultados quando os pressupostos inerentes ao modelo de regressão são satisfeitos. Assim, para a obtenção dos coeficientes $\hat{\beta}_i$, que são as estimativas dos parâmetros β_i, é usada a seguinte expressão:

$$\underset{p\times 1}{\hat{\beta}_i} = \underset{p\times p}{(X'X)^{-1}} \quad \underset{p\times n}{X} \quad \underset{n\times 1}{y} \tag{3.3}$$

Após estimar o modelo testa-se a significância estatística dos coeficientes do modelo de regressão e o seu ajustamento. O teste da significância estatística do modelo é feito tanto para a significância geral do modelo, assim como para os coeficientes de regressão individuais.

O teste de significância geral do modelo consiste em testar se há relação linear da variável y com as variáveis x_1 e x_2, isto é, testar se os coeficientes β_1 e β_2 são conjuntamente iguais a zero. Assim, temos as seguintes hipóteses:

H_0: $\beta_1 = \beta_2 = 0$

H_1: $\exists \beta_i : \beta_i \neq 0, \quad i = 1,2$

A estatística de teste é: $$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2 / p}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n - p - 1)} \sim F(p, n - p - 1, 1 - \alpha) \qquad (3.4)$$

onde:

$\bar{y}$: é o valor médio observado de *y*.

O valor calculado de F é comparado com o F crítico e rejeita-se a hipótese nula se $F_{calculado} > F_{critico}$.

O teste de significância para os coeficientes de regressão individuais consiste em testar se existe uma relação linear entre a variável y e cada uma das variáveis x_i, o que é o mesmo que testar as hipóteses:

$H_{0:}$ $\beta_i = 0$

$H_{1:}$ $\beta_i \neq 0$

A estatística de teste é: $t = \frac{\beta_i}{ep(\beta_i)} \sim t(n - p - 1)$ (3.5)

Se ao comparar o valor calculado de *t* com o valor crítico e o valor calculado for maior que o crítico, ou seja se for maior que 2.0 rejeita-se a hipótese nula de que o coeficiente β_i é igual a zero.

3.2.3 Métodos de avaliação do modelo estimado

Na avaliação do modelo verifica-se se os pressupostos inerentes à análise de regressão são satisfeitos, assim como identificar observações discrepantes de forma a poder aplicar as medidas correctivas quando isso se demonstrar necessário.

Verificação dos pressupostos

(i) ***Linearidade do fenómeno em estudo***: é verificada através do diagrama de dispersão, onde verifica-se se as relações formadas entre a variável dependente y e as independentes x_1 e x_2 são ou não lineares. É também verificada através dos coeficientes de correlação de Pearson, onde verifica-se se os coeficientes de correlação entre a variável dependente y e as independentes x_1 e x_2 são significativos e para tal, testa-se as seguintes hipóteses:

H_0: $\rho = 0$

H_1: $\rho \neq 0$

a estatística de teste é: $$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \qquad (3.6)$$

Tendo-se comparado o T com o *t* crítico, se o T for maior que o *t* crítico rejeita-se a hipótese nula de que $\rho = 0$, o que é o mesmo que dizer que existe uma relação linear estatisticamente significativa.

(ii) ***Normalidade***: é verificada através da análise do gráfico Normal Q-Q, que se baseia na distribuição de probabilidade dos valores observados e esperados numa distribuição normal, onde se as duas distribuições são iguais, neste caso normais, os pontos se sobrepõem sobre a diagonal do gráfico. Como forma de confirmar os resultados da análise gráfica recorre-se ao teste de normalidade de Kolmogorov-Smirnov (K-S) com a correcção de significância de Lilliefors, onde se testam as seguintes hipóteses:

H_0: $e \sim N(0,\sigma^2)$

H_1: $e \overset{\text{não}}{\sim} N(0,\sigma^2)$

A estatística de teste é: $$D_{\max} = g_{\max} + \frac{1}{2n} \qquad (3.7)$$

onde: $g_{\max}$: maior valor calculado de *g*;
n : tamanho da amostra ou número de parcelas.

sendo: $g = |\check{F}(z_i) - F_{0.5}|$ e $F_{0.5} = \frac{i + 0.5}{n}$

$\check{F}(z_i)$: função de distribuição normal acumulada;
$F_{0.5}$: frequência relativa observada acumulada e ajustada;

Com base no valor de $D(1-\alpha, n)$ da tabela de Kolmogorov-Smirnov rejeita-se a hipótese nula de que os resíduos são normalmente distribuídos com média igual a zero e variância constante se o valor de $D_{\max}$ for maior que o valor crítico.

(iii) ***Homocedasticidade***: é verificada através da análise gráfica do diagrama de dispersão entre os resíduos estudantizados e os valores previstos estandardizados, onde é considerado haver homocedasticidade se as observações não apresentarem um padrão definido, isto é, distribuírem-se de forma aleatória em torno da média dos resíduos estudantizados. O método formal usado é o método de White, no qual testam-se as seguintes hipóteses:

H_0: $\sigma^2_{e|x_1, x_2} \equiv \sigma^2$

H_1: $\sigma^2_{e|x_1, x_2} \neq \sigma^2$

Para testar as hipóteses acima o método de White consiste em estimar uma regressão entre os resíduos ao quadrado e as variáveis independentes, dada pela seguinte expressão:

$$\check{e}_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{1i}^2 + \alpha_4 x_{2i}^2 + \alpha_5 x_1 x_2 + \nu \qquad (3.8)$$

onde os α_i são os coeficientes da regressão e ν é o termo aleatório da mesma regressão.

De acordo com Gujarati (2006), pode-se demonstrar que o tamanho da amostra multiplicado pelo coeficiente de determinação (R^2) segue assintoticamente a distribuição de Qui-quadrado com o número de graus de liberdade (gl) igual ao número de coeficientes estimados; para o presente caso $5\,gl$, isto é,

$$n * R^2 \underset{ass}{\sim} \chi^2_{gl}$$

Comparando o valor calculado e o valor $\chi^2_{critico}$ rejeita-se a hipótese nula de que os resíduos têm variância constante se $n * R^2 > \chi^2_{gl}$.

(iv) ***Independência dos resíduos***: é verificada através do teste de Durbin-Watson (DW) onde se testa as seguintes hipóteses:

H_0: $Cov(e_i, e_j) = 0, \quad i \neq j$

H_1: $Cov(e_i, e_j) \neq 0$

Para testar estas hipóteses calcula-se o valor do teste de DW pela expressão:

$$DW = \sum_{t=2}^{n} \frac{(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\hat{\varepsilon}_t^2} \tag{3.9}$$

o qual é comparado com os valores da tabela de Durbin-Watson[5] e que tem como regra de decisão as condições apresentadas na tabela 3.1.

Tabela 3.1: Regra de decisão do teste de Durbin-Watson

condição	Decisão
$0 < DW < d_L$	Rejeitar H_0
$d_L \leq DW \leq d_U$	Sem decisão
$d_U < DW < 4 - d_U$	Não rejeitar H_0
$4 - d_U \leq DW \leq 4 - d_L$	Sem decisão
$4 - d_L < DW < 4$	Rejeitar H_0

(v) ***Independência das variáveis explicativas***: é verificada através dos valores do factor de inflação da variância (FIV), dado pela expressão:

$$FIV_i = \frac{1}{1 - R_j^2} \tag{3.10}$$

[5] Tabela disponível em Gujarati (2006)

onde R_j^2 corresponde ao coeficiente de determinação entre a variável x_i e as restantes variáveis explicativas. Assim, segundo Pestana e Gageiro (2005), está-se perante uma situação de multicolinearidade se os valores de FIV forem maiores que 10.

Uma outra forma de decidir usada no presente trabalho quanto a rejeição ou não rejeição da hipótese nula dos testes feitos consiste em usar o *valor da probabilidade* (*p_value*) que é a probabilidade exacta de cometer o erro de tipo I ou o nível de significância observado e os pacotes estatísticos usados trazem este valor nos seus *outputs* (saídas). Assim, quando *p_value* for menor que o nível de significância fixado então rejeita-se a hipótese nula e não se rejeita a hipótese nula caso contrário.

Na possibilidade de não se atenderem os pressupostos inerentes ao modelo de regressão é necessário aplicar algumas medidas correctivas. A medida correctiva comummente usada quando os pressupostos da linearidade, normalidade e homocedasticidade dos resíduos não são satisfeitos é a transformação das variáveis, calculando-se a sua raiz quadrada, logaritmos ou o inverso da variável conforme o caso e a necessidade. Para a autocorrelação dos resíduos, segundo Hair (2005), a medida é identificar a variável não representada no modelo e acrescenta-la ao modelo, dado que este pressuposto é violado nos casos em que há erro de especificação. Contudo, a violação dos pressupostos pode dever-se à presença de observações discrepantes na amostra, pelo que, verifica-se se tais observações estão presentes.

Identificação de observações discrepantes

O primeiro passo para a identificação de observações discrepantes é verificar se existem outliers e/ou leverages.

Os outliers são verificados através dos resíduos estudantizados, dados pela expressão 2.8b, e são considerado outliers os resíduos com valor absoluto maior que 2.0. A identificação é feita através do diagrama de dispersão e pela inspecção aos resíduos dado que o tamanho da amostra é pequeno.

Os leverages são verificados através dos valores da diagonal principal da matriz de projecção, dada pela expressão 2.8.d, e são considerados leverages as observações com $h_{ii} > 3p/n$, neste caso $h_{ii} > 0.24 \ (=3*2/25)$.

Após identificar as observações outliers e/ou leverages verifica-se se estes são ou não influentes através das medidas de influência da Distância de Cook, SDFFITS e SDFBETA. A tabela 3.2 indica os intervalos para os quais se considera uma observação como um outlier/leverage influente.

Tabela 3.2: Medidas de influência, pontos de corte e regra de decisão

Medida de influência	**Ponto de corte**	**Decisão**	
		não influente	**influente**
Distância de Cook	$4/(n-p-1)$	$Cook's \leq 0.182$	$Cook's > 0.182$
SDFFITS	$\pm 2\sqrt{(p+1)/(n-p-1)}$	$\lvert SDFFITS \rvert \leq 0.739$	$\lvert SDFFITS \rvert > 0.739$
SDFBETA	$\pm 2/\sqrt{n}$	$\lvert SDFBETA \rvert \leq 0.4$	$\lvert SDFBETA \rvert > 0.4$

Dado que nem sempre as medidas de influência dão mesmo resultado para uma mesma observação, no presente trabalho são consideradas influentes as observações que nas três medidas remeterem à mesma decisão, neste caso, influente.

Se forem encontrados outliers e/ou leverages influentes tem-se como medida a estimação de um modelo de regressão com métodos robustos e a estimação de um modelo pelo método dos MQO eliminando tais observações.

O método robusto de estimação do modelo usado é o método dos Mínimos Quadrados Aparados, que consiste em minimizar a soma dos quadrados dos resíduos, isto é,

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{h} e_{(i)}^2$$

Para tal, os resíduos são ordenados de forma crescente, $e_{(1)}^2 \leq e_{(2)}^2 \leq \ldots \leq e_{(n)}^2$, $i = 1, \ldots, n$ e, segundo Chen (2002), h está definido entre $n/2+1 \leq h \leq (3n+p+1)/4$, onde se atinge o ponto de falha igual a $(n-h)/n$. O *output* deste método não traz consigo muita informação,

trazendo apenas as estimativas dos coeficientes, a estimativa do erro padrão da regressão que corresponde a s. De acordo com Kleinbaun et al (1998), o coeficiente de determinação numa regressão de Y dada por $X_1, X_2, \ldots, X_p$ é igual ao quadrado do coeficiente de correlação de Pearson, isto é, $R^2(Y \mid X_1, X_2, \ldots, X_p) = r^2(Y, \hat{Y})$, desta forma pode-se obter o R^2 das estimativas dos MQA. Contudo, para obter outras estatísticas é necessário estimar um modelo pelo método dos MQO com as observações discrepantes eliminadas.

Estimar um modelo de regressão eliminando as observações discrepantes, consiste em eliminar do modelo uma a uma observação, por regra deve-se eliminar primeiro a observação que tiver maiores valores tanto dos resíduos assim como de leverage, e observar o comportamento das estatísticas assim como dos pressupostos ao modelo depois de cada eliminação. Para o presente trabalho não são necessariamente eliminadas todas observações, sendo que o processo de eliminação termina quando os pressupostos forem satisfeitos dado que o tamanho da amostra é pequeno.

IV RESULTADOS E DISCUSSÕES

4.1 Exploração dos Dados

Tabela 4.1: Estatísticas descritivas

Variável	Média	Mediana	Desvio padrão	Mínimo	Máximo
Número de produtos (x_1)	8.8	7.0	6.9	2.0	30.0
Distância (x_2)	409.3	330.0	325.2	36.0	1460.0
Tempo de entrega (y)	22.38	18.11	15.52	8.00	79.24

A tabela 4.1 mostra as estatísticas descritivas da base de dados TEMPENTREGA. Os pontos de venda têm em média 8.8 unidades de produto em stock com um desvio em relação a média de 6.9 unidades, sendo 50% dos pontos de venda com menos de 7 unidades em stock. A quantidade mínima em stock verificada é de 2 unidades e máxima de 30 unidades. A distância percorrida varia entre 36 à 1460 km com uma média de 409 km e um desvio em relação a média de 325 km e para cerca de 50% dos pontos de venda a distância percorrida é menor que 330 km. O tempo médio de entrega dos produtos aos pontos de venda é de 22 horas e 23 minutos com um desvio em relação a média de 15 horas e 31 minutos. O tempo mínimo verificado de entrega é de 8 horas e o máximo de 79 horas e 14 minutos.

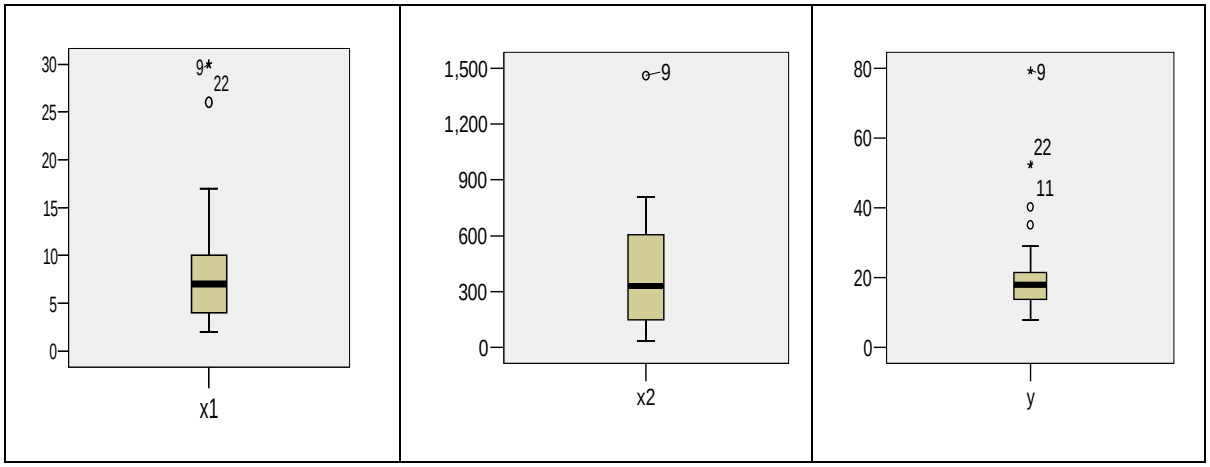


Figura 4.1: Representação gráfica da distribuição das observações por variável

A figura 4.1 apresenta como as observações estão distribuídas por variável, onde a variável número de produtos tem 2 observações que se encontram afastadas das restantes, as observações 9 e 22, sendo a observação 9 a que se encontra mais afastada. A variável distância tem uma observação afastada das restantes, que é a observação 9 e a variável tempo de entrega tem as observações 9, 11, 20 e 22 afastadas das restantes, sendo as observações 9 e 22 estão mais afastadas. Assim, pode-se suspeitar que as observações 9 e 22 são outliers o que é também reforçado pelos diagramas de dispersão da figura 4.2 que indicam estas duas observações como sendo as que mais se afastam da maior parte das observações.

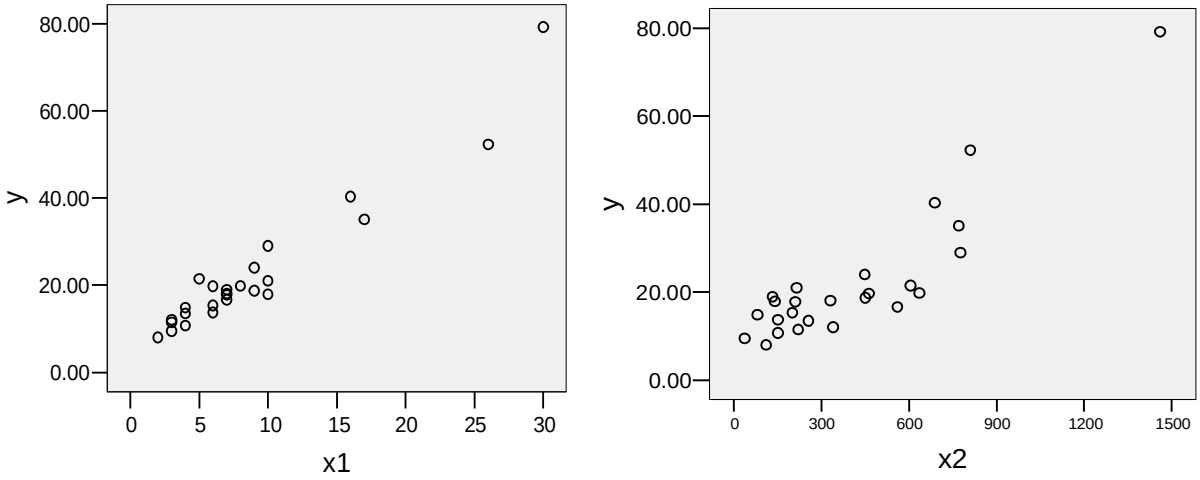


Figura 4.2: Diagramas de dispersão entre a variável dependente e as variáveis independentes.

Os diagramas de dispersão acima mostram que tanto para grandes distâncias assim como para maiores quantidades de produtos nos pontos de venda, maiores são os tempos de entrega.

4.2 Estimação e Testes de Hipóteses ao Modelo de Regressão

Pelo método dos MQO obteve-se os seguintes resultados, onde primeiro se destaca a análise de variância dada pela tabela 4.2 à seguir:

Tabela 4.2: Análise de variância do modelo de regressão com todas observações

Fonte de variação	Soma dos quadrados	gl	Quadrado médio	F	p_value
Regressão	5550.811	2	2775.405	261.235	0.000
Resíduo	233.732	22	10.624		
Total	5784.543	24			

A Soma dos Quadrados Explicados pela regressão (SQE) é muito maior que a Soma dos Quadrados dos Resíduos (SQR), o que é indicativo de que nem todos os coeficientes de regressão são iguais a zero, o que vem a ser confirmado pelo valor de F, calculado com base na expressão 3.4 ($F_{cal} = 261.2$), que é maior que o valor crítico dado pela tabela da distribuição F ($F_{2,\,22,\,0.95} = 3.4$), o que significa que rejeita-se a hipótese nula de que os coeficientes são simultaneamente iguais a zero (ou seja $p\_value < 0.05$).

$$\begin{aligned} \hat{y}_i &= 2.341 + 1.616x_{1i} + 0.014x_{2i} \\ &\quad (1.097)\ (0.171)\quad (0.004) \end{aligned} \tag{4.1}$$

Aplicando a expressão 3.5 obtém-se os valores da estatística *t* e observa-se que para todos os coeficientes de regressão esta estatística tem valores maiores que 2.0 (tabela B1, no anexo), o que leva a rejeitar para todos os coeficientes individuais a hipótese nula de que $\beta_i = 0$ e isto significa que a um nível de significância de 5% todos os coeficientes de regressão são estatisticamente significativos.

O número de produtos em stock nos pontos de venda e a distância percorrida até aos pontos de venda explicam 96.0% da variação do tempo de entrega e os restantes 4% do tempo de entrega são explicados por outros factores.

Mantendo constante a influência da distância e a cada aumento de uma unidade do número de produtos, o tempo de entrega aumenta em média cerca de 1 hora e 37 minutos enquanto que mantendo constante a influência do número de produtos e a cada aumento de 1 km de distância, o tempo de entrega aumenta em média aproximadamente 1 minuto. Através dos

coeficientes estandardizados de regressão dados na equação 4.2 constata-se que a variável x_1 tem maior importância relativa no modelo de regressão.

$$Z_{Yi} = 0.716Z_{X1i} + 0.301Z_{X2i} \tag{4.2}$$

4.3 Avaliação do Modelo Estimado

Tabela 4.3: Correlações de Pearson entre as variáveis

		x1	x2	y
x1	Correlação de Pearson	1	0.824	0.965
	p_value		0.000	0.000
x2	Correlação de Pearson	0.824	1	0.892
	p_value	0.000		0.000
y	Correlação de Pearson	0.965	0.892	1
	p_value	0.000	0.000	

Com base na inspecção visual ao diagrama de dispersão da figura 4.2 pode-se notar que a variável tempo de entrega se relaciona de forma linear com as variáveis independentes e da tabela 4.3 tem-se valores dos coeficientes de correlação de Pearson que indicam uma relação linear muito forte. Pode-se de outra forma confirmar através do teste dos coeficientes de correlação de Pearson, cujas estatísticas são calculadas usando a expressão 3.6 as quais dão como resultados $T(x_1, y) = 17.65$ e $T(x_2, y) = 9.46$ e são ambos maiores que o valor de *t* crítico que é 2.07 com o qual a um nível de significância de 5% há evidências suficientes para se rejeitar a hipótese nula de que os coeficientes de correlação de Pearson são iguais a zero, que é o mesmo que dizer que existe uma relação linear entre a variável dependente e as variáveis independentes.

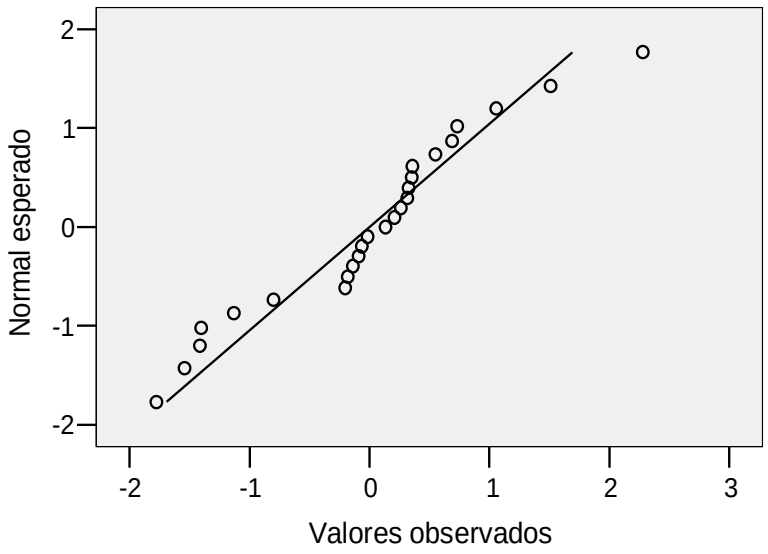


Figura 4.3: Gráfico normal Q-Q para análise da normalidade do modelo com outliers

Pela inspecção ao gráfico normal Q-Q da figura 4.3 nota-se que em geral as observações não se afastam muito da diagonal embora existam algumas e poucas observações que não se sobrepõem tanto à diagonal. Pelo teste de K-S com a correcção de significância de Lilliefors (tabela B4 nos anexos), a um nível de significância de 5% rejeita-se a hipótese nula ($D_{\max} = 0.176$ e $p\_value = 0.045$) de que os resíduos têm distribuição normal. Contudo, a um nível de significância de 4% não se rejeita a hipótese nula.

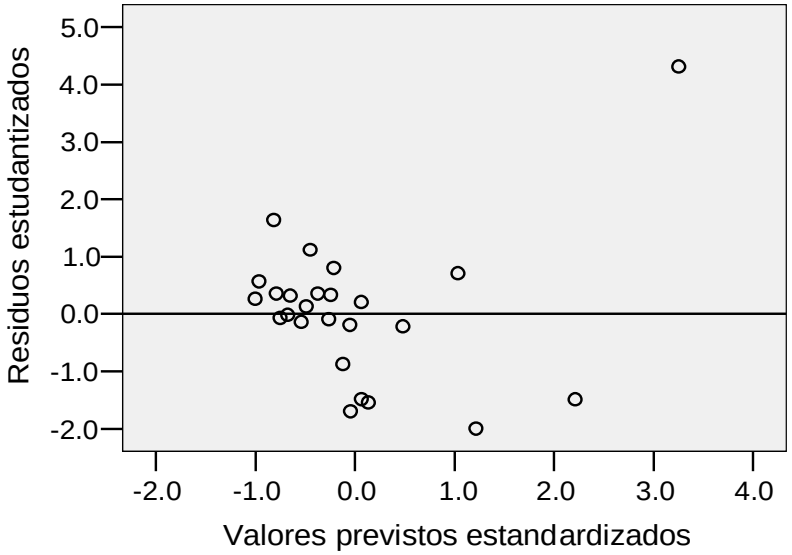


Figura 4.4: Diagrama de dispersão para análise da homocedasticidade do modelo com outliers

$$\begin{aligned} \hat{e}_i^2 &= 2.241+1.081x_{1i} - 0.017x_{2i} - 0.093x_{1i}^2 + 7.67*10^{-6}x_{2i}^2 + 0.003(x_{1i}x_{2i}) \\ R^2 &= 0.598 \end{aligned} \tag{4.3}$$

A figura 4.4 apresenta o diagrama de dispersão entre os resíduos estudantizados e os valores previstos estandardizados, mostra um padrão decrescente formado pelas observações, o que é um indicativo da não aleatoriedade da distribuição dos resíduos em torno do valor zero, isto é, sugere a heterocedasticidade. Para confirmar as suspeitas usou-se o teste de White, onde para se obter a estatística de teste e com base na equação 3.8 estimou-se o modelo 4.3 de onde se obteve o $R^2 = 0.598$, então, $n*R^2 = 25*0.598 = 14.95$ e este valor é maior que o $\chi^2_{critico} = 11.07$. Assim, há evidências suficientes para se rejeitar a hipótese nula de que os resíduos têm variância constante, isto é, há heterocedasticidade entre os resíduos.

Com base na equação 3.9 calculou-se estatística de Durbin-Watson do modelo dado pela equação 4.1 onde obteve-se o valor 1.17 e da tabela de DW para $p = 2$ e $n = 25$ têm-se os valores $d_L = 1.206$ e $d_U = 1.550$. Assim, com $DW < d_L$ (1.17 <1.206) há evidências suficientes para se rejeitar a hipótese nula de que os resíduos são independentes, podendo-se assim dizer que há autocorrelação entre os resíduos.

O valor de FIV calculado à partir da expressão 3.10 é, tanto para x_1 assim como para x_2, igual 3.118 que é inferior a 10, o que indica que as variáveis explicativas são independentes entre si, isto é, não há colinearidade.

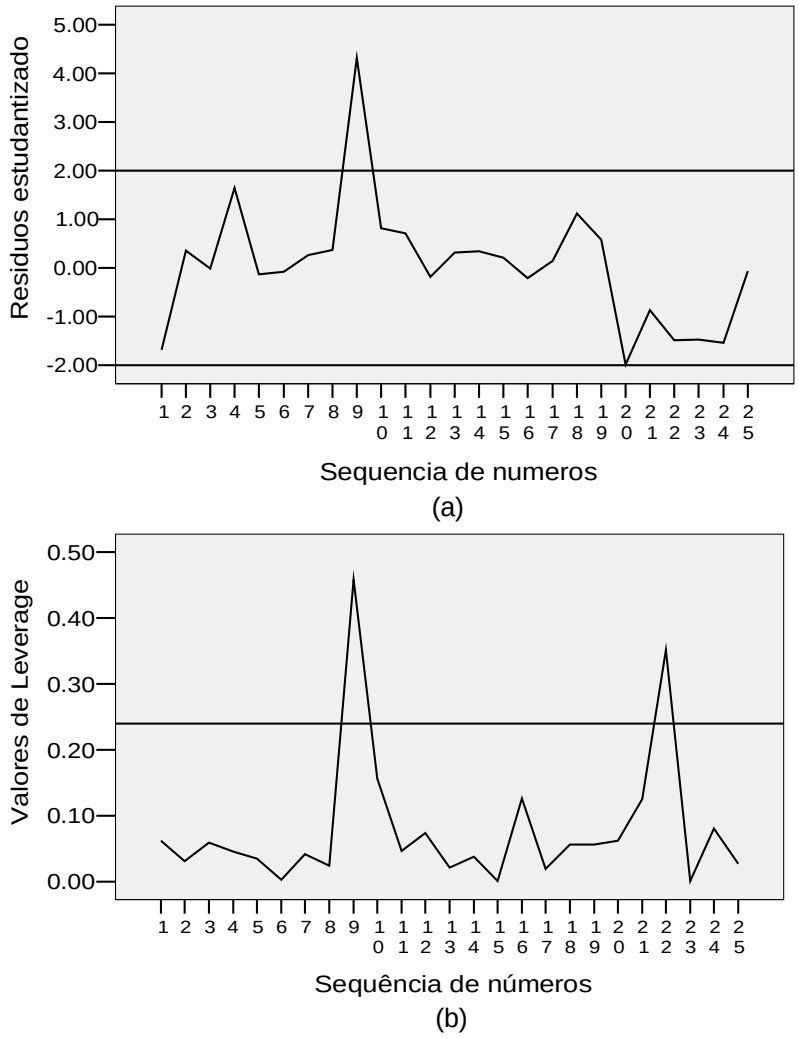


Figura 4.5: Identificação de outliers e leverages

O facto de os resíduos não seguirem a distribuição normal e haver heterocedasticidade pode ser devido à presença de observações discrepantes. Assim, a figura 4.5 têm os gráficos para a detecção de observações outliers e leverages, onde a figura 4.5 (a) dos resíduos estudantizados indica uma observação que se encontra fora do intervalo aceitável, [-2.0, 2.0], neste caso a observação 9 com $t_i = 4.318$, o que significa que esta observação é um outlier. A figura 4.5 (b) dos valores da matriz de projecção indica as observações 9 e 22 como sendo as que estão acima do limite aceitável ($= 0.24$) as quais tem como valores de h_{ii} iguais a 0.458 e 0.352, respectivamente. Assim, pode-se afirmar que estas duas observações são leverages. De notar que estas observações foram tidas como suspeitas na exploração dos dados

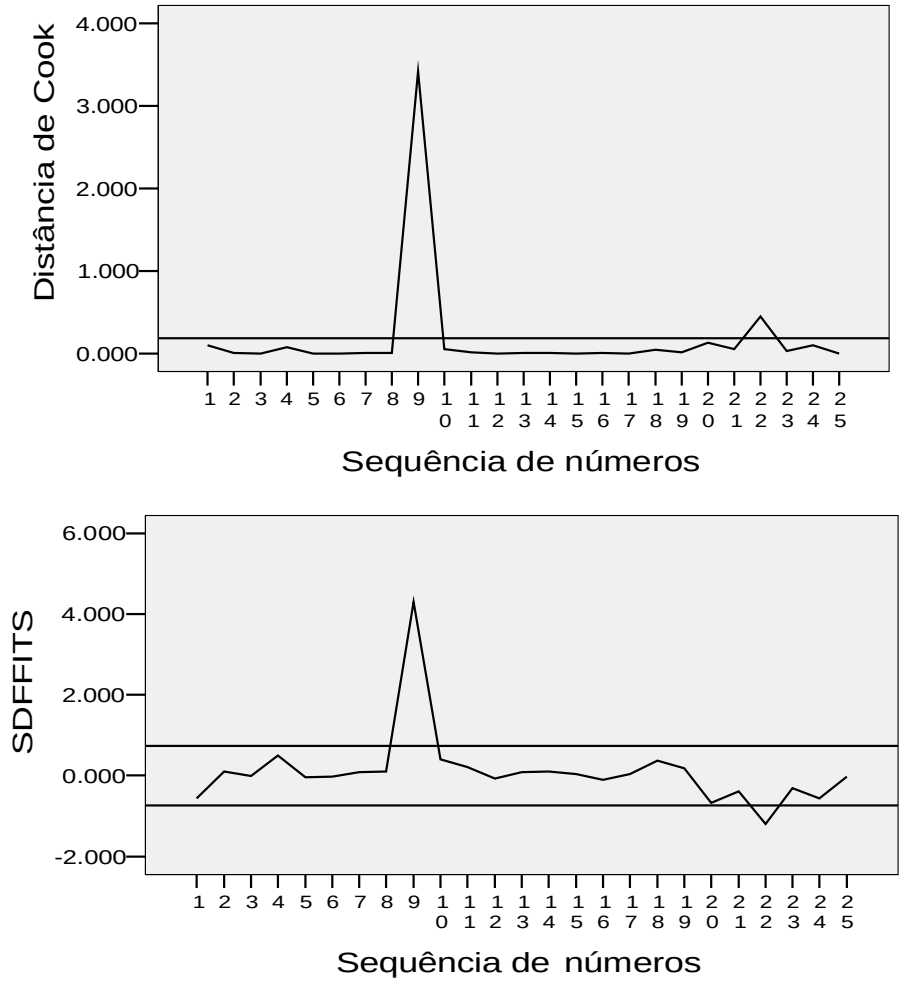


Figura 4.6: Identificação de observações influentes[6]

Como foram detectadas as observações outlier e leverages é necessário saber se estas exercem grande influência no modelo de regressão. Assim, a figura 4.6 têm gráficos com valores das medidas de influência da Distância de Cook, do SDFFITS e do SDFBETA, as quais indicam que as observações 9 e 22 têm valores fora dos limites aceitáveis, podendo-se assim considerar que estas observações são influentes no modelo de regressão, sendo a observação 9 a que em geral tem valores mais altos destas medidas.

[6] O gráfico com a medida SDFBETA encontra-se na Figura B1 nos anexos

4.4 Estimação de modelos com menor efeito dos valores discrepantes

Como forma de minimizar o efeito da presença de valores discrepantes (outlier e leverages influentes) na amostra estimou-se o modelo de regressão pelo método robusto dos MQA, o qual forneceu os seguintes resultados:

$$\hat{y}_i = 3.364 + 1.277x_{1i} + 0.02x_{2i} \qquad \text{erro padrao da regressao : }]1.301, 1.356[\qquad (4.4)$$

Desta regressão para além dos coeficientes de regressão pode-se obter o erro padrão da regressão dado sob forma de intervalo. O coeficiente de determinação é igual a 0.955. Assim, pode-se dizer que o número de produtos em stock nos pontos de venda e a distância percorrida até aos pontos de venda explicam 95.5.0% da variação do tempo de entrega no modelo de regressão estimado pelos MQA e os restantes 4.5% são explicados por outros factores.

Mantendo constante a influência da distância, a cada aumento de uma unidade do número de produtos, o tempo de entrega aumenta em média cerca de 1 hora e 17 minutos enquanto que mantendo constante a influência do número de produtos, a cada aumento de 1 km de distância, o tempo de entrega aumenta em média aproximadamente cerca de 1 minuto.

O modelo de regressão sem as observações discrepantes dado pelo método dos MQO é dado à seguir, onde a sua estimação é feita eliminando uma a uma observação, sendo a primeira a ser eliminada aquela que tem valores mais elevados das medidas de influência, neste caso a observação 9. A tabela 4.4 é referente a análise de variância para o modelo sem a observação outlier.

Tabela 4.4: Analise de variância do modelo sem a observação outlier

Fonte de variação	Soma dos quadrados	gl	Quadrado médio	F	p_value
Regressão	2293.244	2	1146.622	194.182	0.000
Resíduo	124.002	21	5.905		
Total	2417.246	23			

A tabela 4.4 tem como valor da estatística $F = 194.2$, calculado com base na expressão 3.4, que é muito maior que o valor crítico $F_{2,\,21,\,0.95} = 3.5$, assim, a um nível de significância de 5% há evidências suficientes para se rejeitar a hipótese nula de que os coeficientes de regressão são simultaneamente iguais a zero, isto é, pelo menos um dos coeficientes é significativamente diferente de zero.

$$\begin{array}{rl} \hat{y} = & 4.447 + 1.498x_1 + 0.010x_2 \\ & (0.952)\ \ (0.130)\quad (0.003) \end{array} \tag{4.5}$$

A partir da expressão 3.4 obteve-se os valores da estatística *t* (dados na tabela B3, nos anexos), os quais para todos os coeficientes de regressão são maiores que 2.0, o que indica que a um nível de significância de 5% os coeficientes de regressão são estatisticamente significativos. Ainda em relação a este modelo, o número de produtos em stock nos pontos de venda e a distância percorrida até aos postos de venda explicam 94.9% da variação do tempo de entrega e os restantes 5.1% são explicados por outros factores.

Mantendo constante a influência da distância, a cada aumento de uma unidade do número de produtos, o tempo de entrega aumenta em média cerca de 1 hora e 30 minutos enquanto que mantendo constante a influência do número de produtos, a cada aumento de 1 km de distância, o tempo de entrega aumenta em média aproximadamente meio minuto.

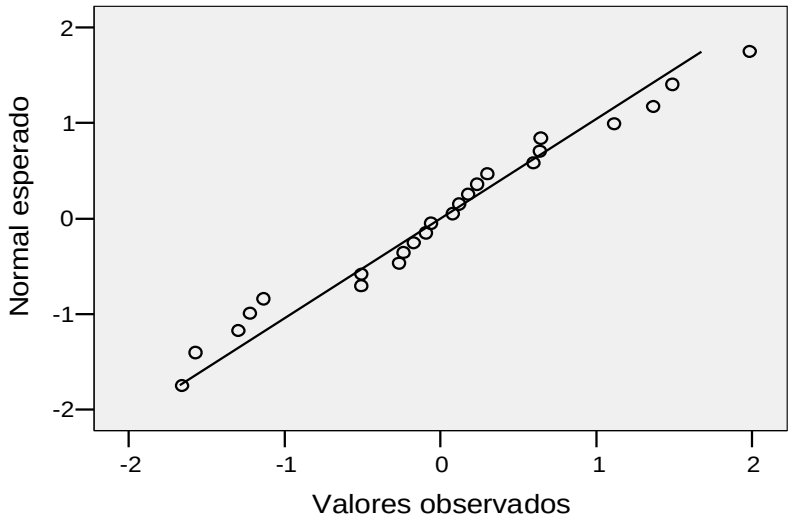


Figura 4.7: Gráfico normal Q-Q para análise da normalidade do modelo sem outliers

O gráfico normal Q-Q da figura 4.7 indica que as observações se sobrepõem relativamente mais à diagonal em relação ao modelo com a observação 9 incluída, o que sugere a normalidade dos resíduos. Tal é confirmado pelo teste K-S com a correcção de significância de Lilliefors (tabela B4 nos anexos), no qual não se rejeita a hipótese nula de que os resíduos seguem a distribuição normal ($D_{\max} = 0.099$ e $p\_value \geq 0.200$).

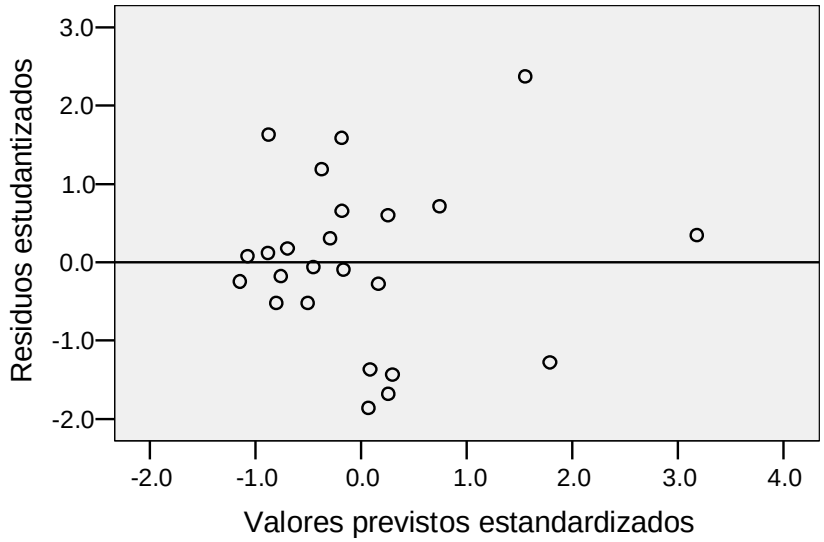


Figura 4.8: Diagrama de dispersão para análise da homocedasticidade do modelo sem outliers

$$\hat{\varepsilon}_i^2 = -2.815 + 1.508x_{1i} - 0.003x_{2i} - 0.070x_{1i}^2 + 9.15*10^{-6}x_{2i}^2 + 4.0*10^{-4}(x_{1i}x_{2i}) \qquad (4.6)$$
$$R^2 = 0.270$$

O diagrama de dispersão entre os resíduos estudantizados e os valores previstos estandardizados da figura 4.8 não indica nenhum padrão formado pelos resíduos, o que significa que eles se distribuem de forma aleatória e isto sugere a homocedasticidade dos resíduos, o que é confirmado pelo teste de White, onde temos $R^2 = 0.270$ na regressão entre os quadrados dos resíduos e as variáveis x_1, x_2, x_1^2, x_2^2 e x_1x_2 (equação 4.6) que na multiplicação com número de observações temos $n*R^2 = 24*0.270 = 6.48$ que é inferior ao valor $\chi^2_{critico} = 11.07$, logo pode-se afirmar que neste modelo os resíduos têm variância constante.

Com o valor de DW =1.331 do modelo dado pela equação 4.5 não se tem um resultado conclusivo já que este se encontra no intervalo $d_L = 1.188 < DW = 1.331 < d_U = 1.550$ no qual não se tem nenhuma decisão. Ainda em relação a este modelo, O valor de FIV tanto para x_1 assim como para x_2 é 1.914 que é muito inferior a 10, o que indica que as variáveis explicativas são independentes entre si, isto é, não há colinearidade.

Com a eliminação da observação 9 o modelo de regressão melhorou quanto ao atendimento aos pressupostos do modelo de regressão, pois verifica-se a normalidade e a homocedasticidade dos resíduos, a não colinearidade entre variáveis explicativas mantém-se, pois o valor do FIV baixou de 3.118 para 1.914, o valor de DW não está na região de rejeição nem de aceitação da hipótese nula, contudo, já não se pode afirmar que há autocorrelação entre os resíduos. Assim, já não se torna necessário eliminar a observação 22 pois isso reduziria mais o tamanho da amostra.

Tabela 4.5: Comparação da observação outlier com as estatísticas amostrais

	observação 9	Média	Desvio padrão
NProduto	30	8.8	6.9
Distância	1460	409.3	325.2
TEntrega	79.24	22.38	15.52

Á partir da tabela 4.5 é notável que a observação 9 é uma observação com valores completamente diferente das outras presentes na amostra (está a pelo menos 3 desvios padrão dos valores médios em todas as variáveis) e não há indícios de que seja resultado de um erro de registo ou introdução de dados, isto sugere que esta observação deveria ser usada num estudo semelhante mas numa amostra onde o número de produtos em stock, a distância e o tempo de entrega sejam maiores que os apresentados na presente amostra e com o padrão da observação 9.

O modelo de regressão estimado pelo método dos MQO sem a observação outlier apresenta melhores resultados que o modelo estimado pelo método dos MQO com todas observações, pois tem os erros padrão dos coeficientes de regressão e o erro padrão da regressão com valores mais baixos, os coeficientes de regressão são estatisticamente significativos e com os

valores da estatística *t* mais altos. Este modelo tem valores de R^2 e R^2 ajustado muito alto e os pressupostos do modelo são satisfeitos, então pode-se afirmar que este é o método mais adequado para fazer previsões. Contudo, se se pretender fazer previsões com um modelo estimado com todas observações o modelo mais adequado é o estimado pelo método dos MQA, porque este apresenta um coeficiente de determinação muito alto e sobretudo o erro padrão das estimativas é muito baixo.

V CONCLUSÕES E RECOMENDAÇÕES

5.1 Conclusões

Após a análise da base de dados e de acordo com os objectivos traçados para o presente trabalho chegou-se às seguintes conclusões:

1) Os métodos que se mostraram adequados para detectar outliers e leverages são o uso dos resíduos estudantizados e o uso dos valores da diagonal principal da matriz de projecção, respectivamente. Com estes métodos detectou-se a observação 9 como outlier e leverage ao mesmo tempo e 22 como leverage;

2) Os métodos que se mostraram adequados para a detecção de observações influentes são a Distância de Cook, o SDFFITS e o SDFBETA, os quais indicaram as observações 9 e 22 como influentes no modelo de regressão;

3) A observação 9 apresenta valores bastante altos para todas as variáveis o que leva a concluir que este não é devido a erro de registo ou introdução de dados mas sim uma inclusão por erro nesta amostra pois esta observação seria adequada a um estudo semelhante onde as observações tivessem o mesmo padrão de variabilidade. A observação 22 também não apresenta evidências de ser erróneo e ainda que tenha valores um pouco elevados pertence ao padrão de variabilidade formado pelas restantes observações; e

4) Os modelos que se mostraram adequados para fazer previsões e com menor efeito da presença dos outliers são os modelos estimados pelo método robusto dos Mínimos Quadrados Aparados e o modelo estimado pelo método dos Mínimos Quadrados Ordinários sem a observação 9.

5.2 Recomendações

Nos próximos trabalhos relacionados com o tratamento de dados com outliers em regressão recomenda-se que:

1) Se use várias bases de dados e com tamanhos de amostra mais elevados que o usado no presente trabalho;

2) Se aprofunde mais os aspectos relacionados com os métodos de regressão robusta usando pacotes estatísticos específicos para estes métodos; e

3) Sejam explorados mais métodos de correcção dos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

Chen, Colin (2002). Robust Regression and Outlier Detection with the ROBUSTREG Procedure. SUGI Paper 265-27. SAS Institute: Cary, NC. Disponivel em http://www2.sas.com/proceedings/sugi27/P265-27.pdf Acesso: 16 de Outubro de 2008.

Dielman, T. E. (2001). Applied Regression Analysis- for Business and Economics, 3rd edition. Thompson Learning, USA.

Draper, N. R. and H. Smith (1998). Applied Regression Analysis, 3rd edition. New York, John Wiley. USA.

Gujarati, D. (2006). Econometria Básica, 4ª edição. Rio de Janeiro, Elsevier Ltda, Brazil.

Hair, J. F., R. E. Anderson, R. L. Tathan and W. C. Black (2005). Análise Multivariada de Dados, 5ª edição. Porto Alegre, Bookman, Brazil.

Harper, W. M. (1991). Statistic, 6th edition. London, ELBS, England.

Hocking, R. R. (1996). Methods and Applications of Linear Models- regression and analysis of variance. New York, John Wiley & Sons, USA.

Johnson, R. A. and D. W. Wichern (2007). Applied Multivariate Statistical Analysis, 6th edition. Pearson Prentice Hall, USA.

Kleinbaum, D. G., L. L. Kupper, K. E. Muller and A. Nizam (1998). Applied Regression Analysis and Other Multivariable Methods, 3rd edition. Brooks/Cole Publishing Company, USA.

Last, M. and A. Kandel (2001). Automated Detection of Outliers in Real-World Data. Disponivel em: http://www.ise.bgu.ac.il/faculty/mlast/papper/outliers2.pdf Acesso: 24 de Agosto de 2009.

Manly, B. F.J. (2000). Multivariate Statistical Methods- A primer, 2nd edition. New York, Chapman & Hall/CRC, USA.

Mukherjee, C., H. White and M. Wuyts (1998). Econometrics and Data Analysis for Developing Countries. Routledge, New York

Pestana, M. H. e J. N. Gageiro (2005). Análise de Dados para Ciências Sociais, 4ª edição. Lisboa, Edições Sílabo, Lda, Portugal.

Rousseeuw, P. J. and A. M. Leroy (2003). Robust Regression and Outlier Detection. New Jersey, John Wiley and Sons, USA.

Silvestre, A. L. (2007). Análise de Dados e Estatística Descritiva. Lisboa, Editora Escolar, Portugal.

Triola, M. F. (1999). Introdução à Estatística, 7ª edição. Rio de Janeiro, LTC, Brazil.

ANEXOS

ANEXOS A: **Bases de dados**

Tabela A1: Base de dados da Figura 2.1

obs	X	Y
1	80.0	81.0
2	82.0	81.0
3	84.0	82.0
4	85.0	81.4
5	87.0	82.1
6	88.0	81.7
7	88.0	82.0
8	89.0	83.5
9	90.0	83.1
10	91.0	82.4
11	91.0	82.7
12	92.0	83.0
13	94.0	83.9
14	96.0	83.9
15	98.0	84.0
16	81.0	83.5

Fonte: Reis (2001)[7].

Tabela A2: Base de dados da Figura 2.2

142	145	150	163	163	165	165	165	165	166
167	168	168	169	169	170	170	170	172	173
173	173	174	174	178	179	180	181	181	183
183	184	192	205						

[7] Reis, E. A. (2001). Exercícios Resolvidos em introdução à Estatística para Ciências sociais. 1ª edição. Universidade Federal de Minas Gerais, Instituto de Ciências Exactas, Departamento de Estatística

Tabela A3: Base de dados TEMPENTREGA

obs	NProduto	Distância	TEntrega
1	7	560	16.68
2	3	220	11.5
3	3	340	12.03
4	4	80	14.88
5	6	150	13.75
6	7	330	18.11
7	2	110	8.00
8	7	210	17.83
9	30	1460	79.24
10	5	605	21.50
11	16	688	40.33
12	10	215	21.00
13	4	255	13.50
14	6	462	19.75
15	9	448	24.00
16	10	776	29.00
17	6	200	15.35
18	7	132	19.00
19	3	36	9.50
20	17	770	35.10
21	10	140	17.90
22	26	810	52.32
23	9	450	18.75
24	8	635	19.83
25	4	150	10.75

Fonte: Rousseeuw e Leroy (2003)

ANEXOS B: Resultados e Discussões

Tabela B1: Resultados de regressão com todas as observações

R múltiplo	R^2 múltiplo	R^2 ajustado	Erro padrão da estimativa
0.980	0.960	0.956	3.25947

Variáveis na equação

Variáveis	Coeficiente	Erro padrão do coeficiente	t	p_value
Constante	2.341	1.097	2.134	0.044
x1	1.616	0.171	9.450	0.000
x2	0.014	0.004	3.500	0.001

Tabela B2: Teste de normalidade de Kolmogorov-Smirnov do modelo com todas observações

	Kolmogorov-Smirnov(a)		
	estatística	gl	p_value
Standardized Residual	**0.099**	**24**	**0.200(*)**

* este é um limite inferior da real significância .
a Correcção Significância Lilliefors

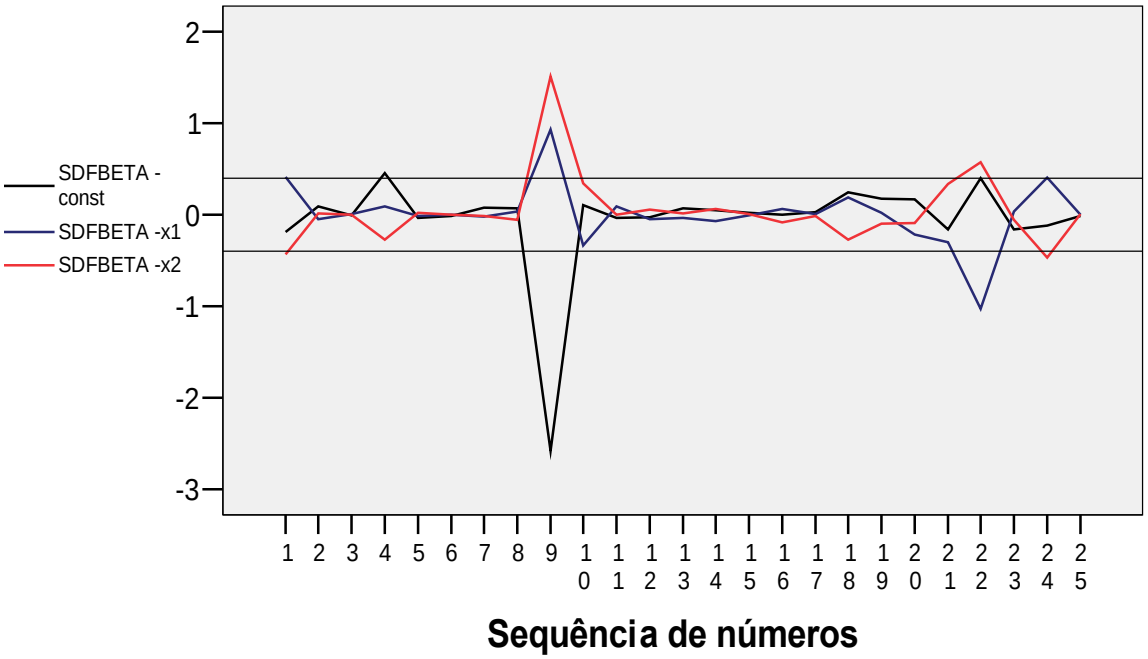


Figura B1: Identificação de observações influentes através de SDFBETA

Tabela B3: Resultados da Regressão sem a observação outlier

R múltiplo	R^2 múltiplo	R^2 ajustado	Erro padrão da estimativa
0.974	0.949	0.944	2.430

Variáveis na equação

Variáveis	Coeficiente	Erro padrão do coeficiente	t	p_value
Constante	4.447	0.952	4.671	0.000
x1	1.498	0.130	11.523	0.000
x2	0.010	0.003	3.333	0.002

Tabela B4: Teste de normalidade de Kolmogorov-Smirnov do modelo sem a observação outlier

	Kolmogorov-Smirnov(a)		
	estatística	gl	p_value
Standardized Residual	0.176	25	0.045

a Correcção Significância Lilliefors

Tabela B5: Valores previstos pelos três métodos usados

obs	Valores observados	Valores previstos por MQO (n = 25)	Valores previstos por MQA (n=25)	Valores previstos por MQA (n=24)
1	16.68	21.7081	23.2409	20.7125
2	11.50	10.3536	11.4917	11.2116
3	12.03	12.0798	13.8358	12.4505
4	14.88	9.9556	10.0340	11.2639
5	13.75	14.1944	13.9552	14.982
6	18.11	18.3996	18.7482	18.338
7	8.00	7.1554	8.0661	8.5783
8	17.83	16.6734	16.4041	17.0991
9	79.24	71.8203	70.1908	ï
10	21.50	19.1236	21.5661	18.1817
11	40.33	38.0925	37.2337	35.5133
12	21.00	21.5930	20.3326	21.6438
13	13.50	12.4730	13.4524	13.0706
14	19.75	18.6825	20.0497	18.2031
15	24.00	23.3288	23.6070	22.5516
16	29.00	29.6629	31.2910	27.4356
17	15.35	14.9136	14.9319	15.4982
18	19.00	15.5514	14.8805	16.2939
19	9.50	7.7068	7.8975	9.3120
20	35.10	40.8880	40.1124	37.8575
21	17.90	20.5142	18.8676	20.8695
22	52.32	56.0065	52.3861	51.7497
23	18.75	23.3576	23.6461	22.5723
24	19.83	24.4029	25.9829	22.9845
25	10.75	10.9626	11.4013	11.9866