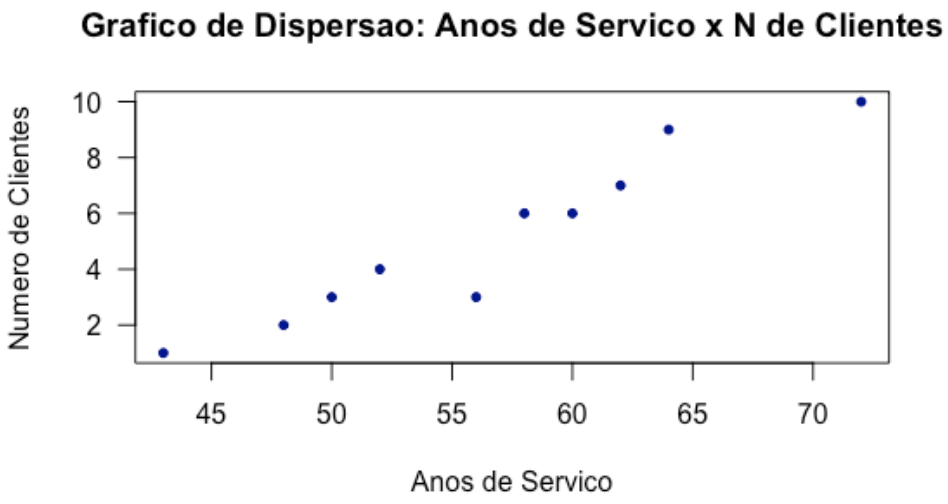
Estatística Avançada

Aula de Regressão

- Regressão Linear Simples
- Regressão Linear Múltipla
- Diagnóstico do modelo

4.1 Introdução a Regressão Linear Simples

Primeiramente, qual é o objetivo de uma análise de regressão linear? Vamos tomar como exemplo um gráfico de dispersão.



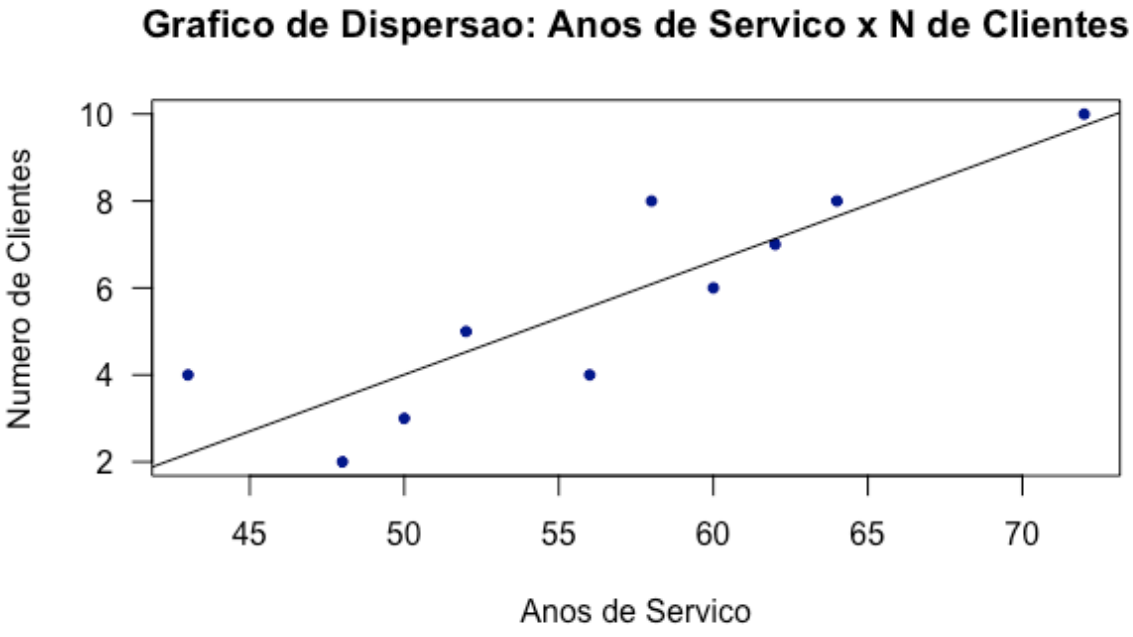
Observe que aparentemente existe uma relação linear entre o número de clientes e os anos de serviço para o nosso funcionário hipotético.

Esse comportamento linear muito nos interessa pois, em primeiro lugar, facilita a nossa interpretação quando observamos os dados, ou seja, eu consigo em poucas

palavras descrever a relação entre as variáveis. Em segundo lugar, nos ajuda a realizar uma possível previsão para dados que não foram observados. Em outras palavras, suponha que exista um funcionário com 20 anos de serviço, ou 10, ou 80. Esses valores não foram observados nos nossos dados, porém ainda assim, dado o caráter linear do que observamos, é possível fazer uma previsão.

Se pudermos traçar uma reta que represente esse comportamento, poderemos estimar o número de clientes de um funcionário qualquer, dada apenas a informação dos anos de serviço. E é para isso que a análise de regressão linear foi criada.

O objetivo da análise de regressão é encontrar uma reta ótima, como a reta abaixo:



Vamos relembrar a equação de uma reta?

$$Y = a + bx$$

Mas no nosso caso, sabemos que os dados não se dispõem de maneira perfeitamente linear. Eles se dispõem de maneira mais ou menos aleatória de modo que todo o conjunto de dados exiba um comportamento linear, mas os dados em si não se alinham perfeitamente. Em outras palavras, o nosso modelo irá prever esse comportamento geral, mas há um erro devido a essa variação dos dados. Por esse motivo, o modelo linear geral para a análise de regressão se dá por:

$$Y = a + bx + e,$$
$$e \sim N(0, \sigma^2)$$

Supõe-se que os erros sejam não-correlacionados e possuem uma distribuição Normal com média 0 e variância σ^2.

Como se obtém essa reta ótima?

O melhor modelo que pudermos ajustar é justamente o modelo que reduz o erro ao máximo.

A reta de uma regressão linear é obtida através do método dos mínimos quadrados.

Esse método consiste em minimizar a soma dos quadrados dos erros.

Do nosso modelo, isolando o erro obtemos que:

$$e = Y - a - bx$$

e, é claro:

$$e^2 = (Y - a - bx)^2$$

Para se encontrar estimativas para os parâmetros a e b, deve-se minimizar os erros em relação aos parâmetros a e b. Para isto, derivamos a somatória dos erros quadráticos em relação a cada um dos parâmetros.

Se quiser acompanhar com detalhes os cálculos, acesse:

http://www.portalaction.com.br/analise-de-regressao/12-estimacao-dos-parametros-do-modelo

Por fim, os estimadores para a e b serão dados por:

$$\hat{b} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{Y}\bar{X}}{(n-1)S_x^2}$$

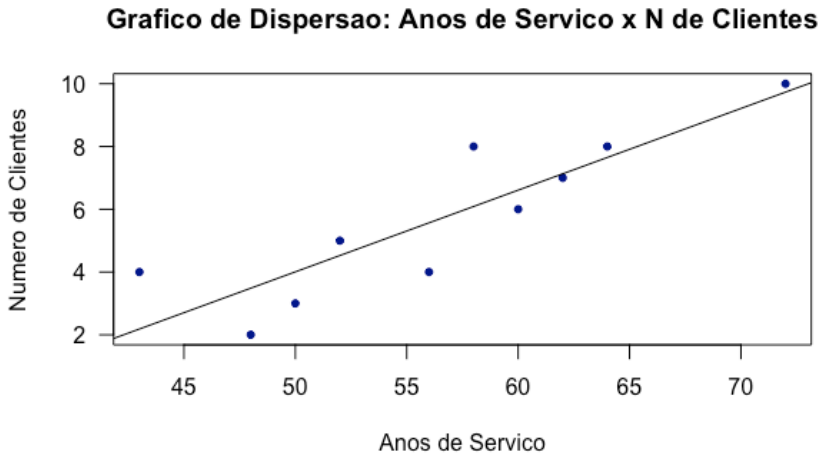
$$\hat{a} = \bar{y} - b\bar{x}$$

Interpretação dos coeficientes:

Para um aumento de uma unidade na variável x, a variável y aumenta em média b unidades.

Quando a variável x é igual a zero, a variável y é a.

No nosso exemplo...



A reta acima foi obtida a partir do método dos mínimos quadrados com coeficientes:

$$\hat{b} = 0.2604$$
$$\hat{a} = -9.014$$

Interprete-os.

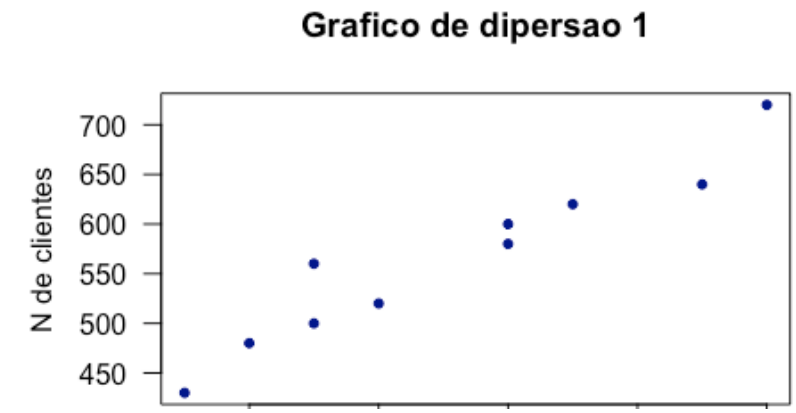
4.2 Introdução a Regressão Linear Múltipla

No caso da análise de regressão linear simples, estamos preocupados em estimar uma variável resposta (Y – Nº de clientes) nos baseando somente em uma variável explicativa (X – Anos de serviço). Mas frequentemente temos muitas outras variáveis explicativas envolvidas. Poderíamos ter, por exemplo, o seguinte banco de dados:

Funcionario	n_clientes	anos_servico	contatos_facebook
A	480	2	200
B	500	3	367
C	560	3	454
D	520	4	387
E	430	1	562
F	600	6	721
G	620	7	864
H	580	6	923
I	640	9	1027
J	720	10	1234

Então se quisermos prever o número de clientes de um funcionário, seria interessante levar em consideração não somente os anos de serviço desse funcionário, mas também o número de contatos que ele tem no facebook.

Como temos três variáveis envolvidas, é possível construirmos 3 gráficos de dispersão diferentes:



Observe que é possível avaliar a relação entre cada variável explicativa e a variável resposta e também a relação entre as duas variáveis explicativas.

Modelo de regressão linear múltipla:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + e$$
$$e \sim N(0, \sigma^2)$$

Supõe-se que os erros sejam não-correlacionados e possuem uma distribuição Normal com média 0 e variância σ^2.

Notação matricial (para generalizarmos para qualquer número de variáveis:

$$Y = X\beta + e$$
$$e \sim N(0, \sigma^2 I)$$

Supõe-se que os erros sejam não-correlacionados e possuem uma distribuição Normal com média 0 e variância σ^2.

Interpretação: Y é uma matriz nx1 de respostas, X é uma matriz nxp (cada explicativa para cada variável) e $\boldsymbol{\beta}$ é uma matriz px1 (a matriz de parâmetros do nosso modelo) e $\boldsymbol{I}$ é a matriz identidade nxn.

A estimativa para $\boldsymbol{\beta}$ é obtida de maneira similar pelo método dos mínimos quadrados, porém, é claro, adaptado para matrizes (mínimos quadrados generalizados).

$$\hat{\beta} = (X^t X^{-1})(X^t Y)$$

Pelo teorema de Gauss-Markov, temos que $\hat{\beta}$ é o melhor estimador linear não-viciado para β. (BLUE – Best Linear Unbiased Estimator)

4.3 Análise de Diagnóstico (resumo)

Após uma análise descritiva cautelosa um bom modelo candidato, é necessário realizarmos uma análise de diagnóstico para conferirmos se o nosso modelo proposto se ajusta adequadamente aos nossos dados e se as suposições estão sendo respeitadas.

Após rodar o modelo proposto no R, podemos observar o seguinte quadro:

```
> fit <- lm(n_clientes ~ anos_servico + contatos_facebook, data=exemplo)
> summary(fit)

Call:
lm(formula = n_clientes ~ anos_servico + contatos_facebook, data = exemplo)

Residuals:
    Min      1Q  Median      3Q     Max
-32.422 -15.501  -4.922   7.959  52.331

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       428.17150   20.32989  21.061 1.37e-07 ***
anos_servico       29.09997    6.92105   4.205  0.00401 **
contatos_facebook  -0.01719    0.06223  -0.276  0.79039
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.12 on 7 degrees of freedom
Multiple R-squared:  0.9218,    Adjusted R-squared:  0.8995
F-statistic: 41.27 on 2 and 7 DF,  p-value: 0.0001336
```

Primeiramente observamos que o R gerou estimativas para cada parâmetro. Cada parâmetro também possui

uma variância. Os valores exibidos em seguida são testes de significância (para testarmos se podemos “descartar” variáveis da nossa análise).

F é o teste para significância da regressão. É um teste para determinar se há uma relação linear entre a variável resposta Y e algumas das variáveis explicativas. Consideremos as hipóteses:

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ H_1: & \beta_j \neq 0 \quad \text{para qualquer } j = 1, \cdots, p \end{cases}$$

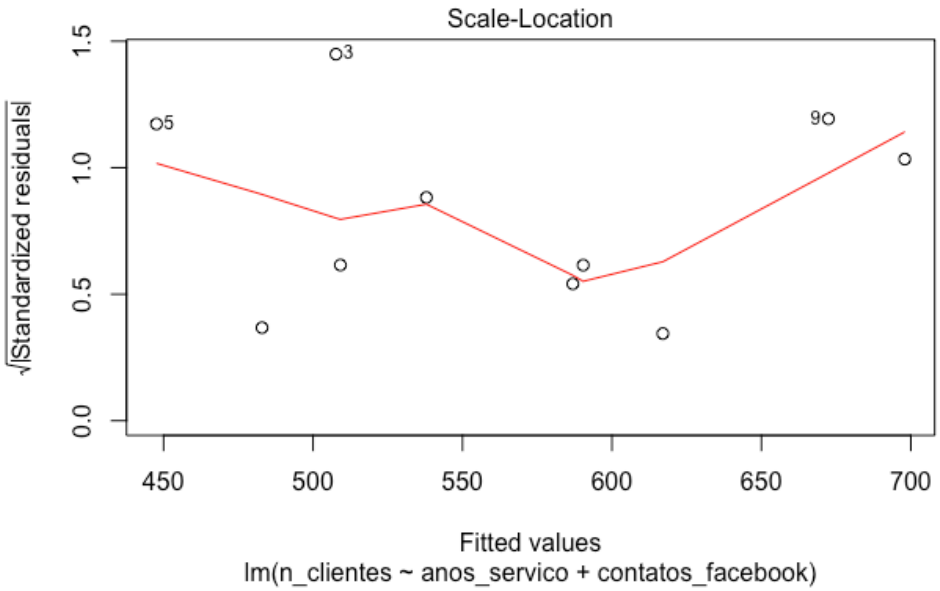
Se rejeitamos H_0, temos que ao menos uma variável explicativa $x_1, x_2, \ldots, x_p$ contribui significativamente para o modelo. Geralmente, adota-se um nível de significância de 5%. No nosso caso, a Estatística F mostrou p-valor bem menor que isso. Portanto não rejeitamos a nossa hipótese.

O R^2 múltiplo representa a proporção da variabilidade de Y explicada pelas variáveis explicativas. Assim, quanto mais próximo O R^2 estiver de 1, maior é a explicação da variável resposta pelo modelo ajustado. No nosso modelo obtivemos um valor muito próximo a 1.

Agora, vamos analisar os resíduos. Precisamos verificar se a suposição de não-correlacionalidade está sendo respeitada, se eles aparentam ter distribuição normal, etc.

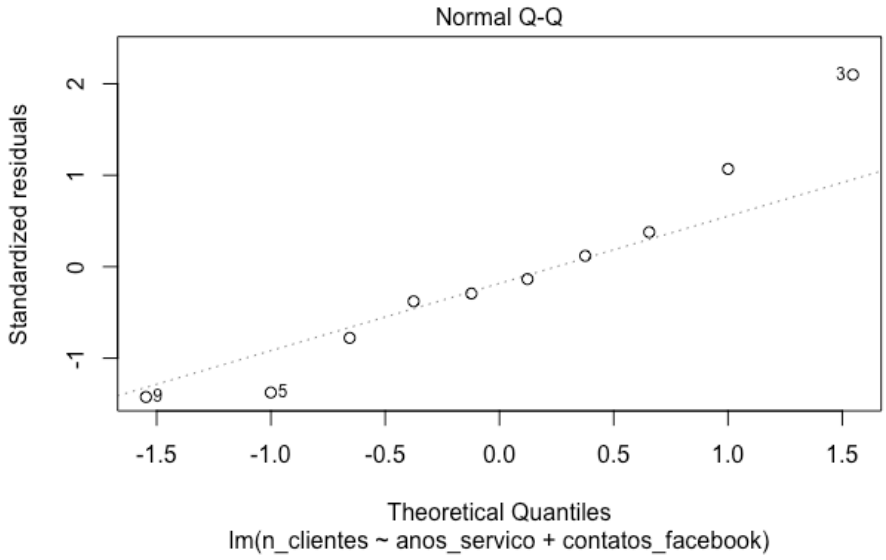
Gráficos de diagnóstico:

Primeiramente, analisaremos o gráfico de resíduos studentizados ou padronizados. Se o modelo for razoável, espera-se que 95% dos resíduos estejam variando entre 2 e -2.

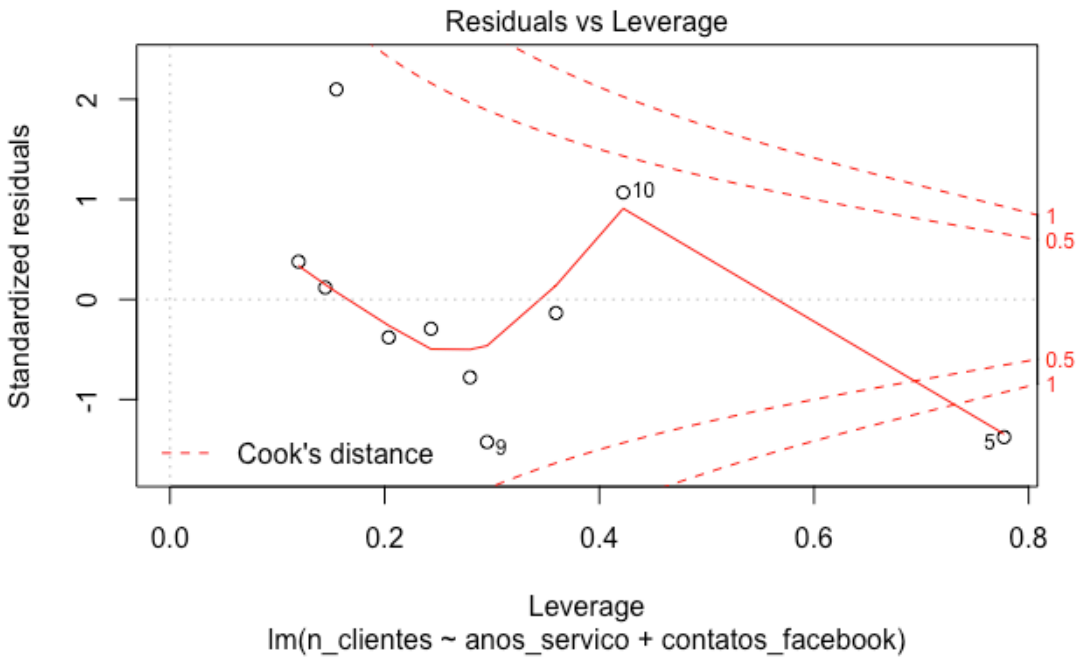


É o nosso caso.

Também precisamos verificar se a suposição de normalidade se adequa aos nossos resíduos. Para isso, gráfico QxQ.



Por fim, um gráfico com a distância de Cook, que utilizamos apenas para verificar pontos outliers.



Também é muito importante avaliarmos o nosso modelo quanto à suposição de erros não correlacionados. Para isso, utilizamos o teste Durbin-Watson. Espera-se que o DW observado esteja em torno de 2 para que essa suposição seja plausível e próximo a 0 se os erros apresentarem correlação.

Vamos analisar a nossa saída no R:

```
> dwtest(fit)

        Durbin-Watson test

data:  fit
DW = 2.5748, p-value = 0.7825
alternative hypothesis: true autocorrelation is greater than 0
```

Como vemos, o nosso DW apresentou valor em torno 2. Também vemos um p-valor bem alto, indicando que uma hipótese não foi rejeitada. Abaixo a hipótese alternativa: autocorrelação maior que zero, ou seja, a hipótese alternativa seria: há correlação entre os erros. Portanto a nula seria o contrário: Não há correlação entre os erros.

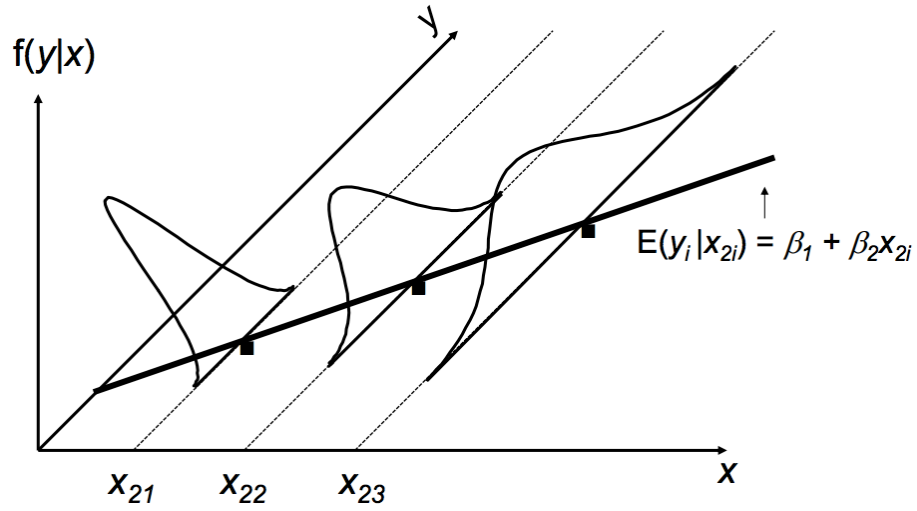
Se a nossa análise nos revelar que há uma possível correlação entre os erros, pode-se optar por uma modelagem utilizando o método de Cochrane Orcutt*.

Suposição de Homocedasticidade:

A suposição de homocedasticidade implica que, condicional às variáveis explicativas, a variância do erro é constante.

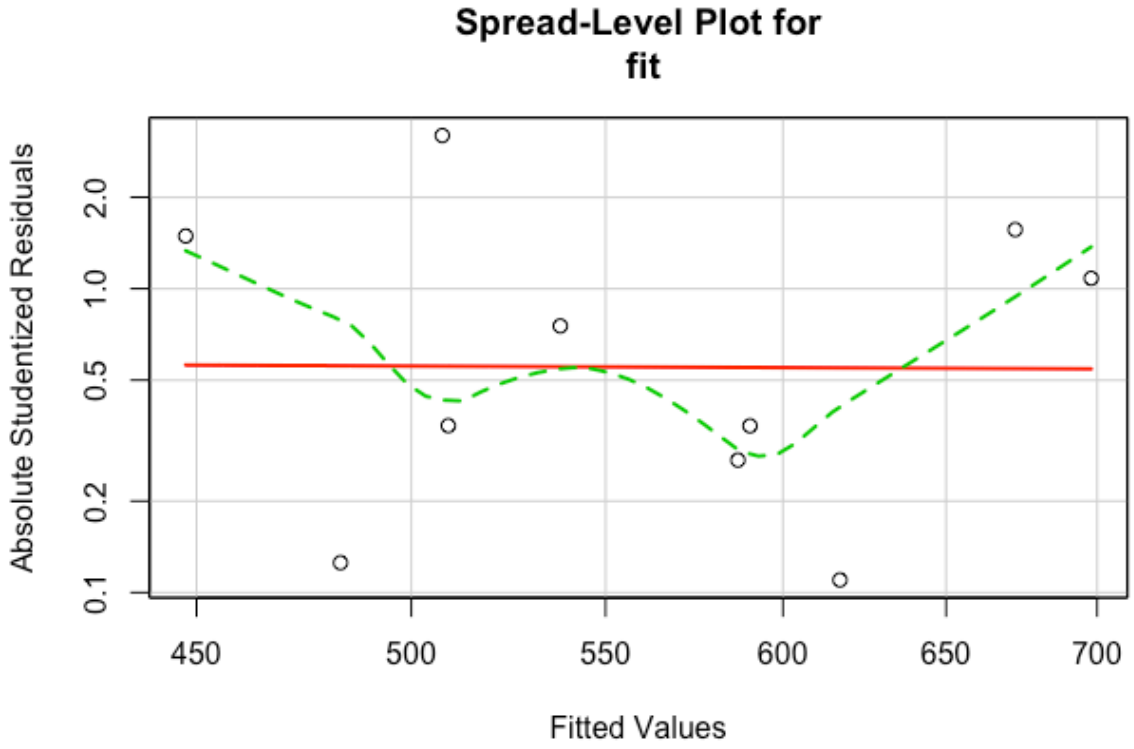
A homocedasticidade não se verifica sempre que a variância dos fatores não observáveis muda ao longo de diferentes segmentos da população, nos quais os segmentos são determinados pelos diferentes valores das variáveis explicativas.

Exemplo



A nossa análise diagnóstica tem que indicar que a variância não se altera para todos os erros, ou seja, eles têm que aparecer aleatoriamente bem distribuídos.

Vejamos o nosso exemplo:

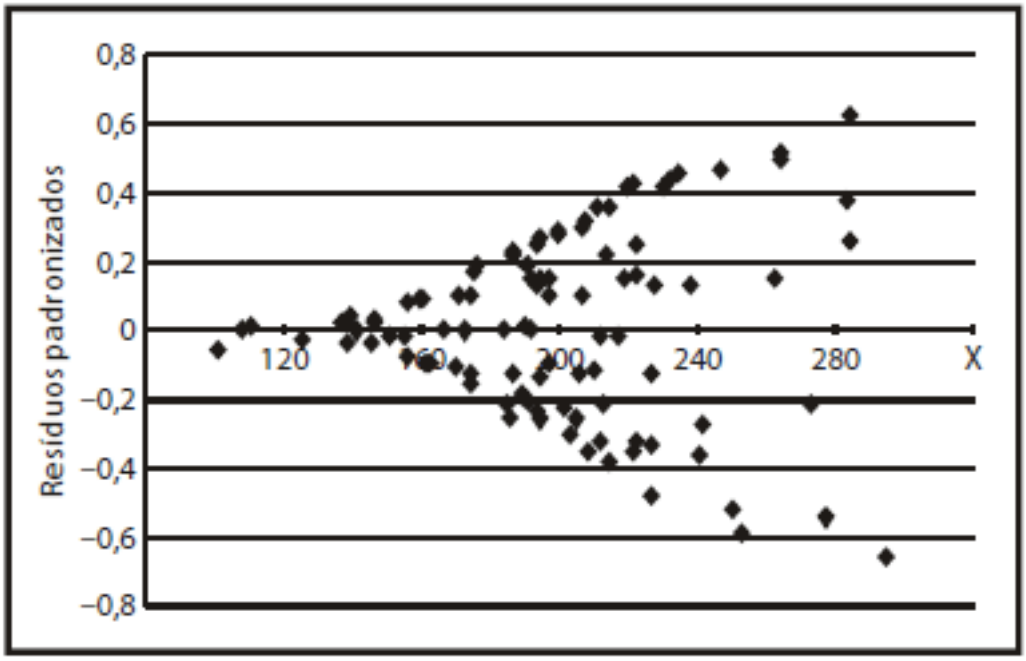


E também um teste estatístico:

```
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.02128767    Df = 1     p = 0.883998
```

A hipótese nula desse teste é: Variância constante dos erros versus alternativa: variância do erro se altera de acordo com o nível da variável resposta ou com uma combinação linear das variáveis explicativas. Vemos no nosso exemplo que a nossa hipótese não foi rejeitada.

Exemplo dessa suposição sendo violada:



Observe que neste caso, conforme x aumenta, a variância dos erros aumenta também. Isso viola a suposição de variância constante.

Conclusão:

No nosso caso, verificamos que o nosso modelo se ajustou muito bem e passou nos nossos testes diagnósticos.