Trabalho Final

ME613 - Análise de Regresssão

Professora: Tatiana Benaglia

Lorena Baquete Marini - 198483
Brenda Luiza Correa - 216037
Vitória Linda da Silva Oliveira - 212826
Paula Liserre Calabrez - 242782

Mínimos Quadrados Ponderados

- Introdução
- Motivação
- Metodologia
- Aplicação a um conjunto de dados
- Bibliografia

Introdução

O método de Mínimos Quadrados Ponderados (MQP) é uma técnica para estabilizar a variância de um modelo de regressão quando as observações são independentes e as variâncias desiguais.

É adequado para minimizar a soma dos quadrados quando um modelo de regressão linear apresenta uma grande variabilidade dos erros.

Compreendendo o uso dos Mínimos Quadrados Ponderados

Presuposto da Regressão Linear Múltipla

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon i$$

A variância da variável dependente, *Y*, deve ser constante para os valores das variáveis preditoras *X1*, *X2*, ..., *Xp-1*.

Os erros do modelo (resíduos) devem ser constantes, independente dos valores da variável resposta (Y).

Homocedasticidade

Heterocedasticidade

Ou seja, quando não acontece, diz-se que os dados apresentam heterocedasticidade

A HOMOSCEDASTICIDADE É NECESSÁRIA PARA ESTIMAR OS TESTES DE T E F, ALÉM DOS INTERVALOS DE CONFIANÇA.

Então, vamos discorrer sobre as consequências da heterocedasticidade para a estimação de Mínimos Quadrados Ordinários e trazer soluções para a ocorrência de heterocedasticidade utilizando Mínimos Quadrados Ponderados.

HETEROSCEDASTIDADE

No cenário de heteroscedastidade, os estimadores de variância são viesados.

Então, como os erros-padrão dos estimadores de Mínimos Quadrados Ordinários (MQO) dependem diretamente dessas variâncias, eles não são mais apropriados para a construção de testes e intervalos de confiança estatísticas t e F.

Precisamos, portanto, obter estimadores de variância que sejam válidos na presença de heterocedasticidade. Para isso, usaremos o método dos Mínimos Quadrados Ponderemos (MQP)

Comparativamente a outros métodos de correção da heterocedasticidade, como a transformação de Box-Cox, o MQP se destaca pela sua eficiência sem demandar uma transformação específica dos dados. Ao preservar a interpretação dos coeficientes do modelo original, o MQP oferece resultados mais diretos e interpretáveis.

MQP PRESERVA A INTERPRETAÇÃO DOS COEFICIENTES DO MODELO ORIGINAL

Metodologia

Como vimos anteriormente, o Método dos Mínimos Quadrados Ponderados aborda casos em que os erros têm variância heterogênea, ou seja, a variância não é constante.

- Modelo de regressão múltipla:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon i$$

- Matriz de variância e covariância: $\underset{n \times n}{\sigma^2\{\varepsilon\}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$

Vamos considerar três casos para as estimativas de regressão: Variâncias dos erros σi^2 são conhecidas, Variâncias dos Erros Conhecidas até uma Constante de Proporcionalidade e Variância dos erros desconhecida.

Metodologia

1. Variância dos Erros conhecida

Neste caso, usa-se o Método da Máxima Verossimilhança para obter estimadores dos coeficientes no modelo de regressão

$$L(\beta) = \left[\prod_{i=1}^{n}\left(\frac{w_i}{2\pi}\right)^{1/2}\right] exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i (Y_i - \beta_0 X_{i1} - \ldots - \beta_p X_{i,p-1})^2\right]$$

em que $wi = 1/\sigma i^2$. Desse modo, wi está inversamente relacionado com a variância σi^2.

Metodologia

1. Variância dos Erros conhecida

Como wi depende das variâncias dos erros, ele é conhecido, desse modo, maximizar L(β) é o mesmo que maximizar o termo exponencial em relação aos coeficientes de regressão: $Q_w = \sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 - \beta_{p-1} X_{i\,p-1})^2$ sendo $Q = (Y \quad Xb)'\,W(Y \quad Xb)$ em que W é uma matriz diagonal contendo os pesos wi.

Vetor dos coeficientes de regressão estimados:

$$(X'WX)\,\hat{\beta}_w = X'WY, \text{ portanto } \hat{\beta}_{w_{px1}} = (X'WX)^{-1} X'WY.$$

Metodologia

1. Variância dos Erros conhecida

Temos que a esperança e variância são: $E[\hat{\beta}] = \beta$ e $Var(\hat{\beta}) = K\sigma_i^2 K^T$. tomamos $\Sigma = \sigma i^2$ e $K = (X'WX)^{-1}X'W$, então $Var(\hat{\beta}) = K\Sigma K^T = [(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]\Sigma[\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}] = (X^T\Sigma^{-1}X)^{-1}$

Como $wi=1/\sigma i^2$, é possível inferir que a matriz de variância-covariânci determinada por $\sigma^2_{pxp}\{\beta_w\} = (X'WX)^{-1}$

- Os estimadores dos mínimos quadrados ponderados e da máxima verossimilhança são: não tendenciosos; consistentes e têm variância mínima entre os estimadores lineares.

Metodologia

2. Variâncias dos Erros Conhecidas até uma Constante de Proporcionalidade

São conhecidas as magnitudes relativas das variâncias, assim os pesos relativos wi são um múltiplo constante dos pesos verdadeiros desconhecidos $1/\sigma_i^2$, sendo $w_i = k(1/\sigma_i^2)$. W continua a ser a matriz diagonal contendo os pesos *wi*, e portanto, esses pesos e W são conhecidos.

Os parâmetros de regressão continuam sendo estimados através da equação:

$$\hat{\beta}_w = (X'WX)^{-1}X'WY$$

Isso ocorre porque estimadores dos mínimos quadrados ponderados e de máxima verossimilhança não são afetados pela constante *k*

Metodologia

2. Variâncias dos Erros Conhecidas até uma Constante de Proporcionalidade

A matriz de variância-covariância passa a ser: $\sigma^2_{pxp}\{\beta_w\} = k(X'WX)^{-1}$ e é desconhecida, porém pode ser estimada como o erro quadrático médio ponderado (MSE):

$$\hat{k} = \frac{\Sigma w_i (Y_i - \hat{Y}_i)^2}{n-p} = \frac{\Sigma w_i e_i^2}{n-p}$$

Então, basta substituir a estimativa de *k em* $\sigma^2_{pxp}\{\beta_w\}$ *para estimar* $Var(\hat{\beta})$

Metodologia

3. Variância Desconhecida

Para o último caso é necessário usar estimativas das variâncias, que podem ser obtidas de diversas maneiras. Abordaremos dois métodos de obtenção de estimativas das variâncias de σ_i^2:

Metodologia

Estimativa da Função de Variância ou Função de Desvio Padrão

A variância do termo de erro εi^2 pode ser expressa da seguinte forma:

$\sigma_i^2 = E\{\varepsilon_i^2\} - (E\{\varepsilon_i\})^2 = E\{\varepsilon_i^2\}$,uma vez que $E\{\varepsilon i\} = 0$.

Assim, podemos estimar a função de variância descrevendo a relação de σi^2 com as variáveis preditoras relevantes, primeiro ajustando o modelo usando MQO, em seguida regressando os resíduos ao quadrado em relação à variável preditora apropriada.

Metodologia

Estimativa da Função de Variância ou Função de Desvio Padrão

Uso de algumas possíveis funções de variância e desvio padrão:

- Um gráfico residual em relação a X_1 exibe uma forma de megafone. Regresse os resíduos absolutos contra X_1.
- Um gráfico residual em relação a $\hat{Y}$ exibe uma forma de megafone. Regresse os resíduos absolutos contra $\hat{Y}$.
- Um gráfico dos resíduos em relação a X_2 sugere que a variância aumenta rapidamente com o aumento de X_2 até um ponto e depois aumenta mais lentamente. Regresse os resíduos absolutos contra X_2 e X_2^2.
- Obtemos os pesos estimados: $w_i = \frac{1}{(\hat{s}_i)^2}$ ou $w_i = \frac{1}{\hat{v}_i}$ que constituem a matriz diagonal W .

Metodologia

Uso de Repetições ou Quase Repetições

Utilizado em experimentos planejados onde observações replicadas são feitas em cada combinação dos níveis das variáveis preditoras.

Se o número de replicatas for grande, os pesos Wi podem ser obtidos diretamente das variâncias amostrais.

Caso contrário, as variâncias ou desvios padrão amostrais devem ser regredidos contra variáveis preditoras apropriadas para estimar a função de variância ou desvio padrão, a partir da qual os pesos podem então ser obtidos.

Dados

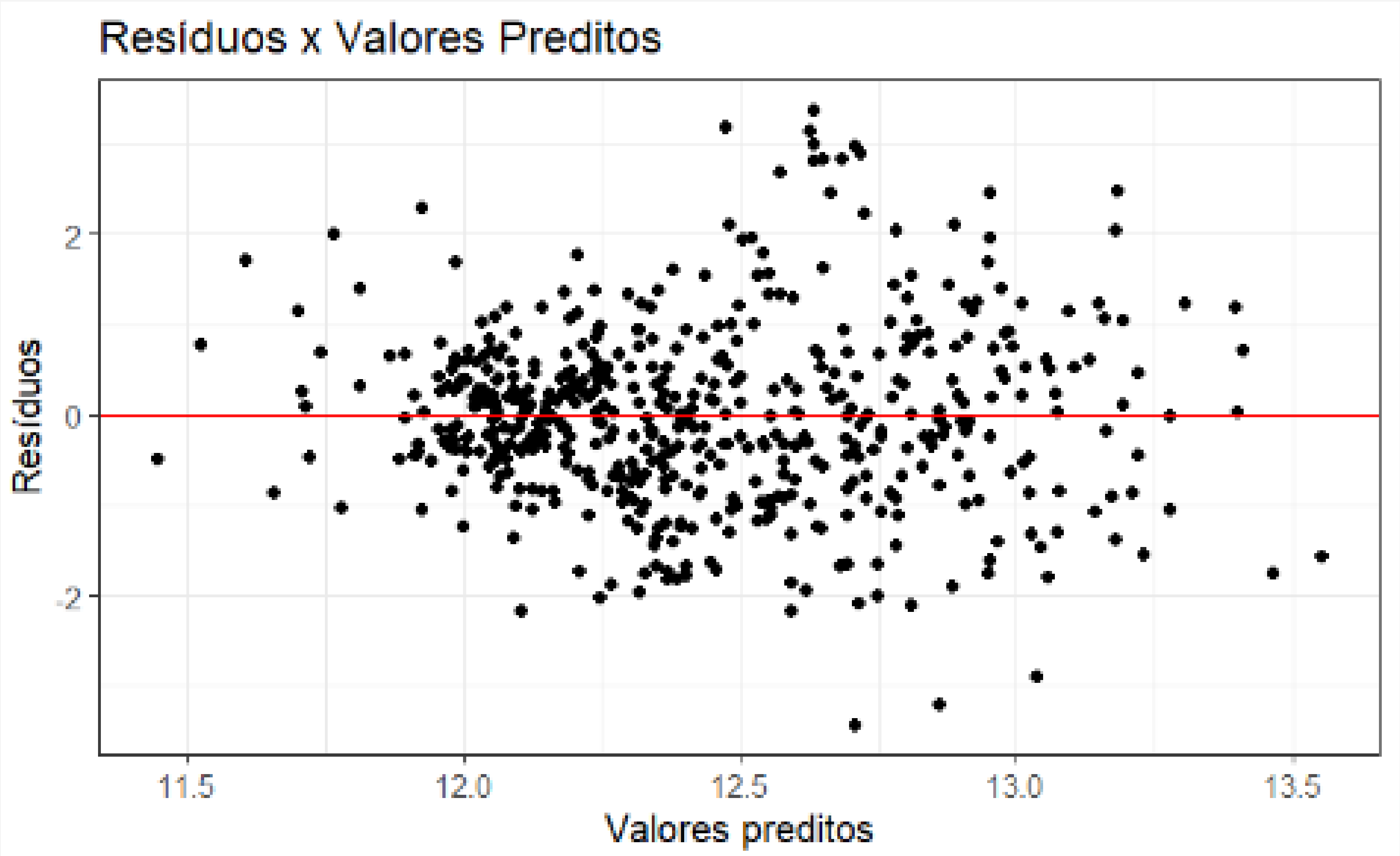
O conjunto de dados Preço de Casa que foi avaliado possui as seguintes variáveis:

- Y = preço de venda de uma casa
- X1= metragem quadrada da casa
- X2 = metragem quadrada do lote

Uma transformação de variável foi aplicada às três características por meio do logaritmo natural, resultando no modelo subsequente.

Aplicação dos dados

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.96361    0.31286   6.276 7.33e-10 ***
x1           1.21976    0.03401  35.867  < 2e-16 ***
x2           0.11034    0.02412   4.575 5.97e-06 ***
---
```



Nota-se um formato "cônico" dos resíduos no gráfico, o que sugere a presença de heterocedasticidade.

Aplicação dos dados

O modelo foi ajustado novamente:

1- Calcule os valores absolutos dos resíduos do MQO;

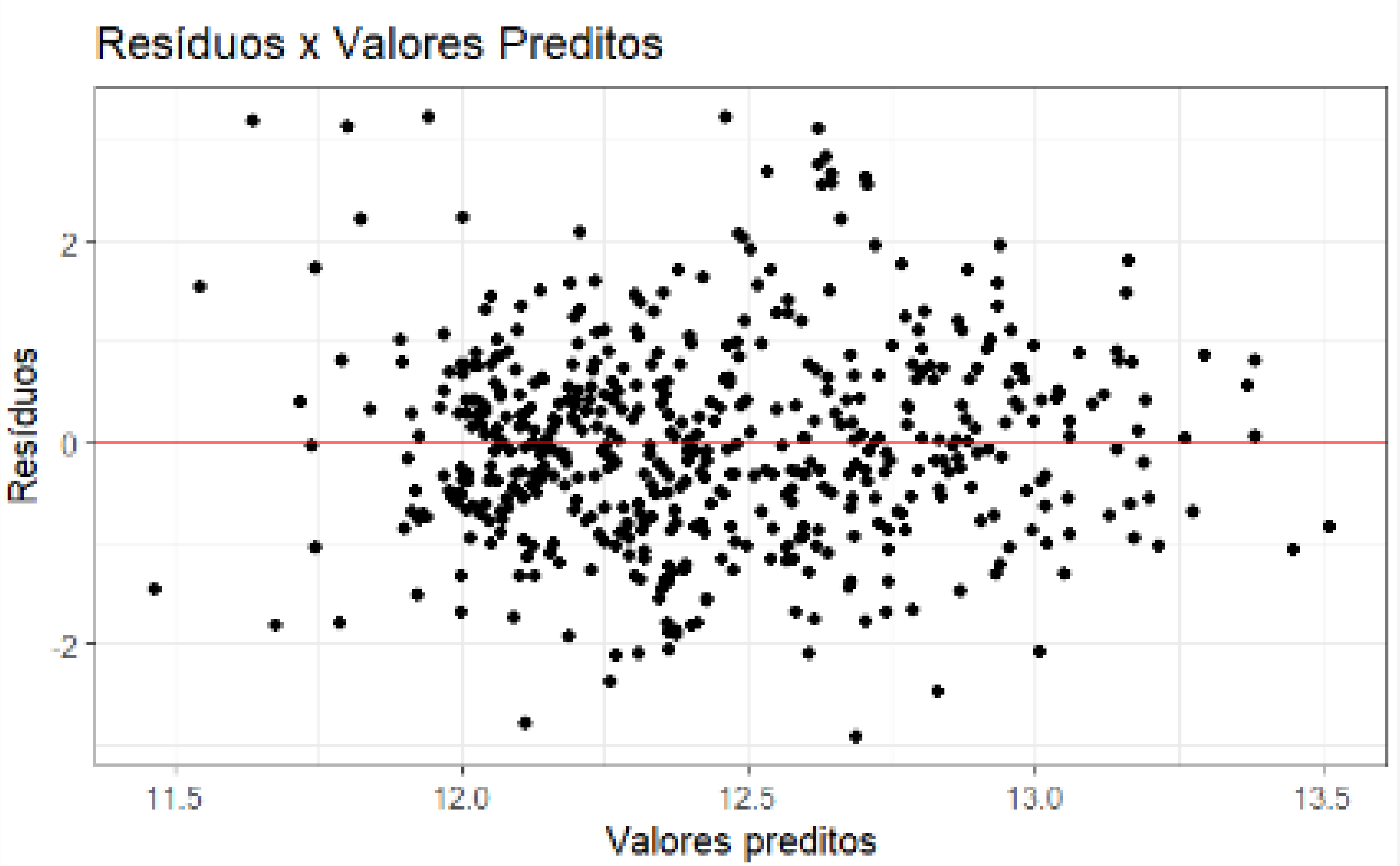
2 - Regresse os valores absolutos dos resíduos de MQO versus os valores ajustados de MQO;

3 - Calcule pesos iguais a $1/\text{fits}^2$, onde "fits" são os valores ajustados da regressão na última etapa.

4 - Em seguida, reajuste o modelo de regressão original, mas desta vez use esses pesos em uma regressão de mínimos quadrados ponderados (MQP).

Aplicação dos dados

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.37671    0.28367   8.379 5.04e-16
x1           1.20136    0.03363  35.721  < 2e-16
x2           0.08308    0.02169   3.830 0.000144
```



Neste novo gráfico podemos ver que os resíduos não seguem mais nenhuma tendência.

Aplicação dos dados

```
fit <- lm(formula = log(y) ~ x1 + x2)
#1 - Calculando os valores absolutos dos resíduos do mínimos
quadrados ordinários
fit_res <- residuals(fit)
abs_res <- abs(fit_res)

#2 - Regressando os valores absolutos dos resíduos de Mínimos
quadrados ordinários (MQO) versus os valores ajustados de MQO -
esses valores ajustados são estimativas dos desvios padrão dos
erros
fit_values <- fitted(fit)
abs_resid_model <- lm(abs_res ~ fit_values)

#3 - Calculando pesos iguais a 1/fits², sendo "fits" os valores
ajustados da regressão na última etapa;
fitted_values_abs_resid <- fitted(abs_resid_model)
weights <- 1/(fitted_values_abs_resid)^2

#4 - Em seguida, reajustando o modelo de regressão original, mas
desta vez usando esses pesos em uma regressão de mínimos
quadrados ponderados (MQP).
fit2 <- lm(formula = log(y) ~ x1 + x2, weights = weights)
residuals_wls <- resid(fit2)
```

Obrigada pela atenção!

Lorena Baquete Marini - 198483
Brenda Luiza Correa - 216037
Vitória Linda da Silva Oliveira - 212826
Paula Liserre Calabrez - 242782