

Universidade Estadual de Campinas

ME613 - Análise de Regressão

Mínimos Quadrados Ponderados

Alunos:

Lorena Baquete Marini - 198483

Brenda Luiza Correa - 216037

Vitória Linda da Silva Oliveira - 212826

Paula Liserre Calabrez - 242782

Docente:

Tatiana Benaglia

Campinas - SP

Novembro de 2023

1 Introdução

O método de Mínimos Quadrados Ponderados (MQP) é uma técnica utilizada para estabilizar a variância de um modelo de regressão em situações onde as observações são independentes e apresentam variâncias desiguais. Ele permite que observações com variâncias mais altas tenham um peso menor no processo de estimativa. Esse método se torna especialmente útil quando a natureza da heterocedasticidade é conhecida, demonstrando ser mais eficiente na obtenção de estimativas de erro do que o método de Mínimos Quadrados Ordinário (MQO), uma vez que realiza a minimização dos resíduos através de uma ponderação.

Na Seção 2 será discutida a motivação por trás da escolha do método MQP. Apresentando as situações em que o método se destaca, delineando a sua eficácia ao lidar com a falta de homocedasticidade dos resíduos do modelo.

A Seção 3 abordará a metodologia de aplicação do Método dos Mínimos Quadrados Ponderados MQP em três cenários distintos para a estimativa de regressão. Na primeira parte (Seção 3.1), serão considerados casos em que as variâncias dos erros são conhecidas, empregando o método da Máxima Verossimilhança para estimar os coeficientes de regressão. Em seguida (Seção 3.2), o foco estará nas situações em que as variâncias dos erros são conhecidas até uma constante de proporcionalidade, exigindo o cálculo do MSE_w , um estimador para a constante k desconhecida. A terceira parte (Seção 3.3) abordará o cenário em que a variância dos erros é desconhecida. Nessa seção, serão exploradas duas abordagens distintas: na primeira (Seção 3.3.1), será detalhada a obtenção de estimativas das variâncias, descrevendo a relação de i^2 com as variáveis preditoras relevantes; na segunda (Seção 3.3.2), será discutido o uso de replicatas em experimentos planejados para cada combinação de níveis das variáveis preditoras, permitindo a obtenção de pesos para cada combinação de níveis das variáveis X . Essa seção será fundamental para compreender os métodos estatísticos e matriciais empregados em cada cenário, visando a obtenção precisa dos coeficientes de regressão diante da variabilidade das variâncias dos erros nos modelos de regressão.

Na Seção 4 é feita a aplicação do MQP em um conjunto de dados. A Seção 4.1 descreve o banco de dados obtido no site da PennState Eberly College of Science. A Seção 4.2 apresenta a análise dos resíduos antes e depois da aplicação do MQP pela criação de diferentes modelos de regressão. Por fim, a seção 5 apresenta a bibliografia utilizada para a realização deste relatório.

2 Motivação para o Uso dos Mínimos Quadrados Ponderados (MQP) na Regressão Linear

A escolha do método de Mínimos Quadrados Ponderados (MQP) na análise de regressão é motivada por uma necessidade crucial: lidar com a heteroscedasticidade, um desafio recorrente na modelagem estatística. A heteroscedasticidade, caracterizada pela variação não constante dos erros de

um modelo de regressão, compromete a precisão das estimativas dos parâmetros, levando a inferências imprecisas e potencialmente incorretas.

O método de Mínimos Quadrados Ponderados oferece uma abordagem eficaz ao atribuir pesos diferenciados às observações com base na variância de seus erros. Essa estratégia inteligente compensa a disparidade na variabilidade dos dados: observações com maior variância de erro recebem pesos menores, enquanto aquelas com menor variância recebem pesos maiores. Essa diferenciação na atribuição de pesos melhora significativamente a precisão das estimativas dos parâmetros do modelo.

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.

Portanto, a adoção do método de Mínimos Quadrados Ponderados se justifica pela sua capacidade de corrigir a heterocedasticidade, preservando a interpretação dos parâmetros do modelo original. Sua aplicabilidade generalizada, aliada à sua eficácia na presença de variâncias heterogêneas nos dados, reforça sua posição como uma escolha prudente na análise de regressão.

3 Metodologia

Como introduzido na Seção 2, o Método dos Mínimos Quadrados Ponderados aborda casos em que os erros têm variância heterogênea, ou seja, a variância não é constante. Denota-se, então, a variância do termo de erro ε_i como σ_i^2 , $i = 1, \dots, n$, representando que os erros podem ter variâncias diferentes. Desse modo, o modelo de regressão múltipla é apresentado como: $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$, no qual $\beta_0, \beta_1, \dots, \beta_{p-1}$ são os parâmetros; $X_{i1}, \dots, X_{i,p-1}$ são constantes conhecidas; $\varepsilon_i \sim N(0, \sigma_i^2)$ e são independentes; $i = 1, \dots, n$. Ademais, a matriz de variância e covariância dos termos de erro para o modelo de regressão múltipla é descrita como:

$$\sigma^2\{\varepsilon\} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Neste contexto, serão considerados três casos para as estimativas de regressão: Variâncias dos erros σ_i^2 são conhecidas (Seção 3.1), Variâncias dos erros conhecidas até uma constante de proporcionalidade (Seção 3.2) e Variância dos erros desconhecida (Seção 3.3).

3.1 Variância dos Erros Conhecidas:

Quando as variâncias dos erros σ_i^2 são conhecidas, usa-se o Método da Máxima Verossimilhança para obter estimadores dos coeficientes de regressão no modelo de regressão, maximizando $L(\beta)$ em relação a $\beta_0, \beta_1, \dots, \beta_{p-1}$:

$$L(\beta) = \left[\prod_{i=1}^n \left(\frac{w_i}{2\pi} \right)^{1/2} \right] \exp \left[-\frac{1}{2} \sum_{i=1}^n w_i (Y_i - \beta_0 X_{i1} - \dots - \beta_p X_{ip-1})^2 \right], \text{ em que } w_i = \frac{1}{\sigma_i^2}.$$

Desse modo, w_i está inversamente relacionado com a variância σ_i^2 . Uma observação Y_i que possui uma variância maior recebe menos peso do que outra observação que possui uma variância menor. Assim, quanto mais precisa for Y_i , ou seja, menor for σ_i^2 , mais informação Y_i fornece sobre $E\{Y_i\}$ e, portanto, mais peso ela deve receber ao ajustar a função de regressão.

Como w_i depende das variâncias dos erros, ele é assumido como conhecido, desse modo, maximizar $L(\beta)$ é o mesmo que maximizar o termo exponencial, apresentado a seguir, em relação aos coeficientes de regressão:

$$Q_w = \sum_{i=1}^n w_i (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_{p-1} X_{ip-1})^2$$

Para facilitar, vamos expressar Q em forma matricial: $Q = (Y - Xb)' W(Y - Xb)$, em que W é uma matriz diagonal contendo os pesos w_i :

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix}$$

Assim, $Q = (Y - Xb)' W(Y - Xb) = (Y'WY - Y'WXb - b'X'WY + b'X'WXb)$.

Derivando em relação a β e igualando a zero, obtemos o vetor dos coeficientes de regressão estimados:

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

Enfim, para definir a Esperança e Variância de $\hat{\beta}$, consideramos que: $E(\varepsilon) = 0$ e $Var(\varepsilon) = \sigma_i^2$. Em

que, $\hat{\beta} = KY$ onde $K = (X'WX)^{-1}X'W$. Então $\hat{\beta} = KY = K(X\beta + \varepsilon) = \beta + K\varepsilon$, desse modo, $E[\hat{\beta}] = \beta$ e $Var(\hat{\beta}) = K\sigma_i^2K^T$. Então,

$$Var(\hat{\beta}) = K\Sigma K^T = [(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]\Sigma[\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}] = (X^T\Sigma^{-1}X)^{-1}$$

Note que, o estimador não é viesado.

Como $w_i = \frac{1}{\sigma_i^2}$, é possível inferir que a matriz de variância-covariância dos coeficientes de regressão, estimados por meio dos MQP, é determinada por $\sigma_{pxp}^2\{\beta_w\} = (X'WX)^{-1}$, uma vez que as variâncias são assumidas como conhecidas.

Portanto, os estimadores dos Mínimos Quadrados Ponderados e da Máxima Verossimilhança dos Coeficientes de Regressão são não tendenciosos, consistentes e têm variância mínima entre os estimadores lineares não tendenciosos. Assim, quando os pesos são conhecidos, o vetor β_w geralmente exibe menos variabilidade do que o estimador dos mínimos quadrados ordinários β .

3.2 Variâncias dos Erros Conhecidas até uma Constante de Proporcionalidade:

Considera-se agora, o caso em que apenas são conhecidas as magnitudes relativas das variâncias. Por exemplo, se σ_1^2 é duas vezes maior que σ_2^2 , pode-se usar os pesos $w_1 = 1$ e $w_2 = \frac{1}{2}$. Nesse caso, os pesos relativos w_i são um múltiplo constante dos pesos verdadeiros desconhecidos $\frac{1}{\sigma_i^2}$, sendo $w_i = k(\frac{1}{\sigma_i^2})$ e k a constante de proporcionalidade. Onde W continua a ser a matriz diagonal contendo os pesos w_i , e portanto, esses pesos e W são conhecidos. Se assumirmos ainda que $k=1$, obteremos o modelo com variâncias conhecidas e erros independentes conforme descrito acima, mas nesta seção, k é outra incógnita que deve ser estimada.

Os parâmetros de regressão continuam sendo estimados através da equação o $\hat{\beta}_w = (X'WX)^{-1}X'WY$. Isso ocorre porque os Estimadores de Mínimos Quadrados Ponderados e de Máxima Verossimilhança não são afetados pela constante k , tendo em vista que k aparece em ambos os lados das equações e se cancela. Isso significa que $\hat{\beta}_w$ é o mesmo, independentemente de as variações serem ou não tratadas como conhecidas. Porém, a matriz de variância-covariância dos coeficientes de regressão dos MQP passa a ser: $\sigma_{pxp}^2\{\beta_w\} = k(X'WX)^{-1}$. A constante de proporcionalidade k é desconhecida, mas é convencionalmente estimada como o erro quadrático médio ponderado (MSE):

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

Em seguida, substituímos a estimativa de k em $\sigma_{pxp}^2\{\beta_w\}$ para estimar $Var(\hat{\beta})$. Como resultado geral obtido comparando a estatística do teste $\hat{\beta}/\sqrt{Var(\hat{\beta})}$ com valores críticos da distribuição t com

$n - p$ graus de liberdade ao realizar testes de hipóteses para $H_0: \beta_w = 0$, um Intervalo de Confiança

para $\hat{\beta}_w$ é calculado como $\hat{\beta}_w \pm t_{n-p,\alpha/2} \sqrt{Var(\hat{\beta})}$.

3.3 Variância dos Erros Desconhecidas

Quando as variâncias dos erros são desconhecidas é necessário usar estimativas das variâncias, que podem ser obtidas de diversas maneiras. Serão abordados dois métodos de obtenção de estimativas das variâncias de σ_i^2 :

3.3.1 - Estimativa da Função de Variância ou Função de Desvio Padrão.

O primeiro método é baseado em descobertas empíricas que indicam que as magnitudes de σ_i^2 e σ_i muitas vezes variam de maneira regular com uma ou várias variáveis preditoras X_k ou com a resposta média $E\{Y_i\}$. A variância do termo de erro ε_i^2 , representada por σ_i^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, $E\{\varepsilon_i\} = 0$. Portanto, o resíduo ao quadrado (ε_i^2) é um estimador de σ_i^2 . Assim, podemos estimar a função de variância descrevendo a relação de σ_i^2 com as variáveis preditoras relevantes, primeiro ajustando o modelo de regressão usando mínimos quadrados não ponderados e, em seguida, regressando os resíduos ao quadrado em relação à variável preditora apropriada.

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 ao quadrado em relação a X_3 exibe uma tendência ascendente. Regresse os resíduos ao quadrado contra X_3 .
- 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 .

Após a estimação da função de variância ou da função de desvio padrão, os valores ajustados dessa função são usados para obter os pesos estimados:

$$w_i = \frac{1}{(\hat{s}_i)^2} \quad \text{em que } \hat{s}_i \text{ é o valor ajustado da função de desvio padrão.}$$

$$w_i = \frac{1}{\hat{v}_i} \quad \text{em que } \hat{v}_i \text{ o valor ajustado da função de variância}$$

Esses pesos constituem a matriz diagonal W, usada na expressão do estimador ponderado dos mínimos quadrados: $\hat{\beta}_w = (X'WX)^{-1}(X'WY)$.

3.3.2 - Uso de Repetições ou Quase Repetições.

Um segundo método para obter estimativas das variâncias dos termos de erro σ_i^2 pode ser 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 w_i podem ser obtidos diretamente das variâncias amostrais das observações de Y em cada combinação dos níveis das variáveis X. Caso contrário, as variâncias amostrais ou desvios padrão amostrais devem ser primeiramente regredidos contra variáveis preditoras apropriadas para estimar a função de variância ou desvio padrão, a partir da qual os pesos podem então ser obtidos.

4. Aplicação a um conjunto de dados

4.1 Dados escolhidos

O conjunto de dados "Home Price" foi obtido no site da PennState Eberly College of Science, que fornece recursos e dados para análises estatísticas e pesquisas acadêmicas. O conjunto de dados contém 522 observações, onde cada observação inclui um identificador (ID), o preço de venda (Sales Price), a metragem quadrada da casa (home ft²) e a metragem quadrada do lote. Ou seja,

Y = Preço de venda da casa (Variável resposta)

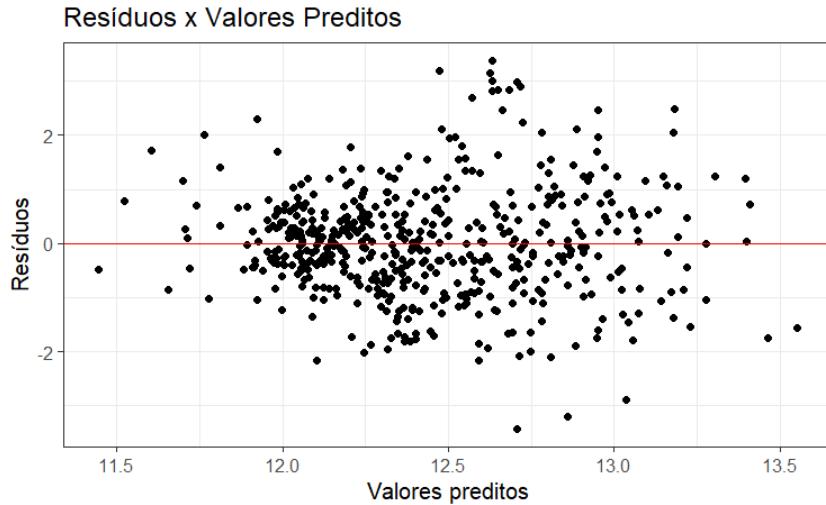
X_1 = Metragem quadrada da casa (Variável preditora)

X_2 = Metragem quadrada do lote (Variável preditora)

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

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.96361	0.31286	6.276	7.33×10^{-10}
X_1	1.21976	0.03401	35.867	$< 2 \times 10^{-16}$
X_2	0.11034	0.02412	4.575	5.97×10^{-06}

Foi obtido o Gráfico de dispersão dos resíduos pelos valores preditos:



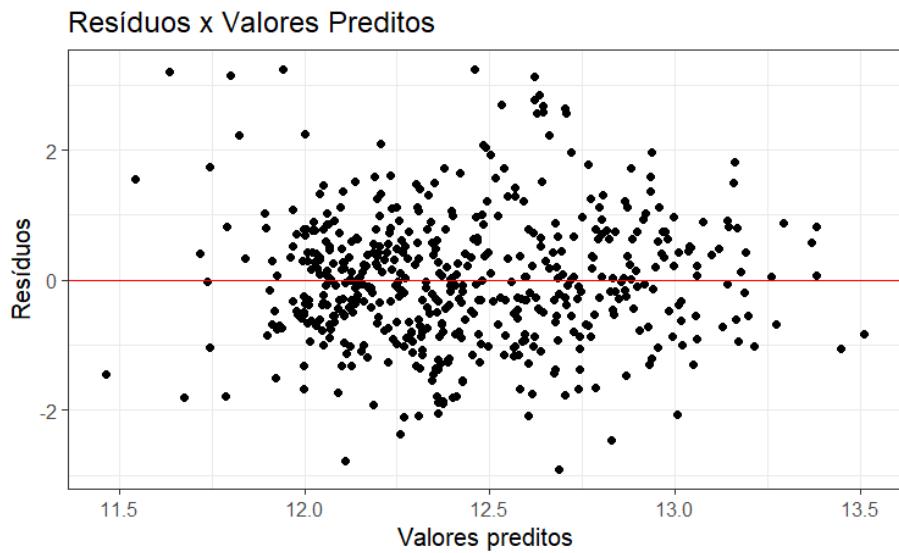
Nota-se um "formato de megafone" ou "cônico" dos resíduos no gráfico, o que sugere a presença de heterocedasticidade, ou seja, a dispersão dos resíduos parece mudar à medida que os valores preditos aumentam ou diminuem.

Desse modo, o modelo foi ajustado novamente, seguindo os seguintes passos:

1. Calculando os valores absolutos dos resíduos do mínimos quadrados ordinários;
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;
3. Calculando pesos iguais a $1/fits^2$, sendo "fits" os valores ajustados da regressão na última etapa;
4. Em seguida, reajustando o modelo de regressão original, mas desta vez usando esses pesos em uma regressão de mínimos quadrados ponderados (MQP).

Desse modo, foi obtido um novo modelo e gráfico de dispersão dos resíduos pelos valores preditos:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.37671	0.28367	8.379	5.04×10^{-16}
X_1	1.20136	0.03363	35.721	$< 2 \times 10^{-16}$
X_2	0.08308	0.02169	3.83	0.000144



Nota-se que os pontos à esquerda se dispersaram e o gráfico não apresenta mais um formato cônico, ou seja, os pontos não apresentam mais um padrão, dando indício assim, que os erros são homocedásticos. Esse resultado é um indício positivo de que o novo modelo ajustado com mínimos quadrados ponderados oferece uma melhor adequação aos dados, corrigindo a heterocedasticidade presente no modelo original.

5. Bibliografia

WOOLDRIDGE, Jeffrey M. Introdução à econometria: uma abordagem moderna. São Paulo: Cengage Learning, 2008. Capítulo 8.

KUTNER, M. H. et al. Applied linear statistical models. 5th ed. [S.l.]: [s.n.], [s.d.]. Disponível em: https://users.stat.ufl.edu/~winner/sta4211/ALSM_5Ed_Kutner.pdf. Acesso em: [03/10/2023].

STATISTICAL OFFICE AT CARNEGIE MELLON UNIVERSITY. Lecture notes. [S.l.], [s.d.]. Disponível em: <https://www.stat.cmu.edu/~larry/stat401/lecture-24.pdf>. Acesso em: [03/10/2023].

PENNSYLVANIA STATE UNIVERSITY. Dataset: Home Prices. [S.l.], [s.d.]. Disponível em: https://online.stat.psu.edu/onlinecourses/sites/stat501/files/data/home_price.txt. Acesso em: [03/10/2023].