Regressão Linear Simples e Múltipla

Universidade Federal da Paraíba

Sumário

Exemplo:

Desejamos prever o gasto mensal de uma pessoa com base em sua renda. Suponha que temos os seguintes dados:

Renda Mensal (em R$)	**Gasto Mensal (em R$)**
1000	300
2000	600
3000	900
4000	1600
5000	2000
6000	2400
7000	3500
8000	4000
9000	3600
10000	4200

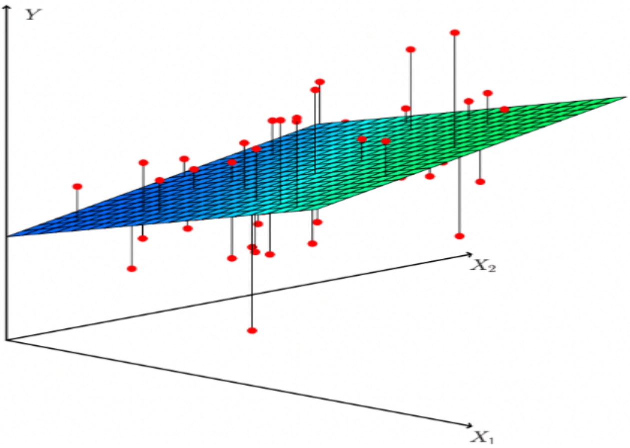
Regressão Linear Simples: Homocedasticidade

- Uma vez que N é pequeno, tanto o teste de Breusch-Pagan quanto o de White podem falhar.
- Ao tomar o mesmo exemplo com N suficientemente grande, os testes conseguem capturar a heterocedasticidade dos resíduos.
- Podemos concluir que o modelo de regressão linear para este caso não vai gerar boas previsões.

REGRESSÃO LINEAR MÚLTIPLA

Regressão Linear Múltipla: Construção do Modelo

- Modelo de Regressão Linear Múltipla



Regressão Linear Múltipla: Construção do Modelo

- Para o caso geral, em que temos d variáveis independentes, então o conjunto de pontos observados é o seguinte subconjunto de $\mathbb{R}^{d+1}$:

$$\mathcal{O} = \{(x_1^j, x_2^j, \ldots, x_d^j, y^j) : j = 1, \ldots, N\}$$

- Deseja-se encontrar o hiperplano de dimensão d que mais se aproxima deste conjunto de pontos. Neste caso, a regressão linear pode ser escrita da seguinte forma:

$$f(x_1, x_2, \ldots, x_d) = a_1.x_1 + a_2.x_2 + \ldots + a_d.x_d + b$$

Regressão Linear Múltipla: Construção do Modelo

- Problema de minimização associado:

$$d(f(x_1,\ldots,x_d),\mathcal{O}) = \underbrace{\sum_{j=1}^{N}(y^j - a_1.x_1^j - \ldots - a_d.x_d^j - b)^2}_{F(a_1,\ldots,a_d,b)}$$

- Pontos críticos de F:

$$\nabla F(a_1,\ldots,a_d,b) = 0 \quad \Rightarrow \quad \boxed{(\hat{a}_1,\ldots,\hat{a}_d,\hat{b}) \leftarrow \text{Parâmetros Ótimos}}$$

- Resíduos:

$$\boxed{R_j = y^j - \hat{a}_1.x_1^j - \ldots - \hat{a}_d.x_d^j - \hat{b}}$$

VALIDAÇÃO DA RLM

Regressão Linear Múltipla: Validação do Modelo

- Além das propriedades verificadas para validar o modelo de regressão linear simples, agora também precisamos verificar uma propriedade extra: a multicolinearidade.
- A multicolinearidade ocorre quando duas ou mais variáveis independentes estão altamente correlacionadas.
- Problemas que podem decorrer da multicolinearidade:
  - Coeficientes de regressão instáveis, ou seja, pequenas alterações no conjunto de dados podem resultar em grandes mudanças nos coeficientes estimados;
  - Significância estatística comprometida;
  - Aumento nas variâncias dos estimadores dos coeficientes (VIF);
  - Redução do poder preditivo;
  - Redundância das variáveis (redução de dimensionalidade);

Validação do Modelo: Multicolinearidade

Fator de Inflação da Variância (VIF):

- O Fator de Inflação da Variância é uma métrica usada para detectar a presença e o grau de multicolinearidade entre as variáveis independentes em um modelo de regressão múltipla.
- O VIF mensura o quanto a variância de um coeficiente de regressão é inflada devido à colinearidade com outras variáveis independentes.
- Para cada $i \in \{1, \ldots, N\}$, deve-se determinar uma regressão auxiliar, onde a variável x_i é considerada como dependente e as demais variáveis ($\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\}$) são consideradas como independentes.
- As atuais variáveis independentes serão rebatizadas:

$$x_1 \leftarrow x_1^i; \ldots; x_{i-1} \leftarrow x_{i-1}^i; x_{i+1} \leftarrow x_{i+1}^i; \ldots; x_N \leftarrow x_N^i$$

Validação do Modelo: Multicolinearidade

Fator de Inflação da Variância:

- O próximo passo consiste em calcular o coeficiente de determinação da regressão auxiliar R_i^2:

$$R_i^2 = 1 - \frac{\sum_{j=1}^{N}(x_j^i - \hat{x}^i)^2}{\sum_{j=1}^{N}(x_j^i - \bar{x}^i)^2}$$

- Cálculo do VIF referente à variável x_i

$$\text{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

- Interpretação dos valores de VIF:
  - $1 \leq VIF < 5$: Multicolinearidade baixa ou inexistente;
  - $5 \leq VIF < 10$: Multicolinearidade de moderada a alta;
  - $VIF \geq 10$: Multicolinearidade muito alta.

Validação do Modelo: Normalidade dos Resíduos

Testes de Shapiro-Wilk e Anderson-Darling:

- Estatísticas dos testes:

$$\begin{cases} W = \dfrac{\left(\sum_{i=1}^{N} a_i x_{(i)}\right)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \\ \\ A^2 = -N - \sum_{i=1}^{N} \frac{2i-1}{N}[\ln(F(X_i)) + \ln(1 - F(X_{n+1-i}))] \end{cases}$$

- Teste de hipóteses (Nível de Significância = 5%):

$$\begin{cases} H_0 : \text{A distribuição dos resíduos é normal} \\ H_1 : \text{A distribuição dos resíduos não é normal} \end{cases}$$

Validação do Modelo: Independência dos Resíduos

Teste de Durbin-Watson:

- Estatística do teste:

$$d = \frac{\sum_{t=2}^{N}(R_t - R_{t-1})^2}{\sum_{t=1}^{N}(R_t)^2}$$

- Teste de hipóteses (Nível de Significância = 5%):

$$\begin{cases} H_0 : \text{Não existe autocorrelação entre os resíduos} \\ H_1 : \text{Existe correlação entre os resíduos} \end{cases}$$

- Interpretação da Estatística Durbin-Watson:
  - $d = 2$ indica que não existe autocorrelação;
  - $d = 0$ indica autocorrelação positiva extrema;
  - $d = 4$ indica autocorrelação negativa extrema.

Validação do Modelo: Existência de Outliers

Distância de Cook:

É uma medida utilizada para identificar observações influentes em uma análise de regressão linear. A Distância de Cook combina informações sobre a magnitude dos resíduos com a posição da observação em relação aos valores ajustados para determinar sua influência no modelo.

- Uma regra prática comum é considerar observações com uma Distância de Cook maior que $\frac{4}{N}$ (sendo N o número de observações) como possivelmente influentes.

Validação do Modelo: Homocedasticidade

Testes de Breusch-Pagan e White:

- Ambos os testes são qui-quadrado. As estatísticas dos testes são avaliadas sob uma distribuição $N\chi^2$.
- A estatística do teste é obtida a partir de uma regressão dos quadrados dos resíduos nas variáveis independentes.
- Teste de hipóteses (Nível de Significância $= 5\%$):

$$\begin{cases} H_0 : \text{Ocorre a homocedasticidade} \\ H_1 : \text{Ocorre a heterocedasticidade} \end{cases}$$