Teste de White

Julio Manuel P Osório

FEA.Dev

2022

Conceitos preliminares

Considere um dataset com m observações e n features, e que regredimos uma certa variável y para cada observação $i \in \{1, \ldots, m\}$ com base nas n features. Considere que ajustamos a seguinte regressão linear:

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_n X_n \tag{1}$$

Daí, definimos o erro ε_i:

$$\varepsilon_i := y_i - \hat{y}_i \tag{2}$$

onde $\hat{y}_i$ é o valor predito pelo modelo para y com respeito à observação i e y_i é o valor observado para a observação i. Assim, já podemos definir o conceito que será protagonista da apresentação de hoje.

Homocedasticidade

Definição

Mantendo a notação que usamos em (1) e (2), dizemos que há **homocedasticidade** quando

$$Var(\varepsilon_i|x_1, ..., x_m) = \sigma^2 \tag{3}$$

$\forall i \in 1, ..., m.$

Sendo assim, a condição de homocedasticidade diz-nos que a variância condicional do erro da regressão linear é constante (i.e., não depende do conjunto de observações $x_1, ..., x_m$. Quando tivemos a situação contrária, ou seja, quando tivermos

$$Var(\varepsilon_i|x_1, ..., x_m) = \sigma_i^2 \tag{4}$$

dizemos que há **heterocedasticidade** no modelo.

Homocedasticidade

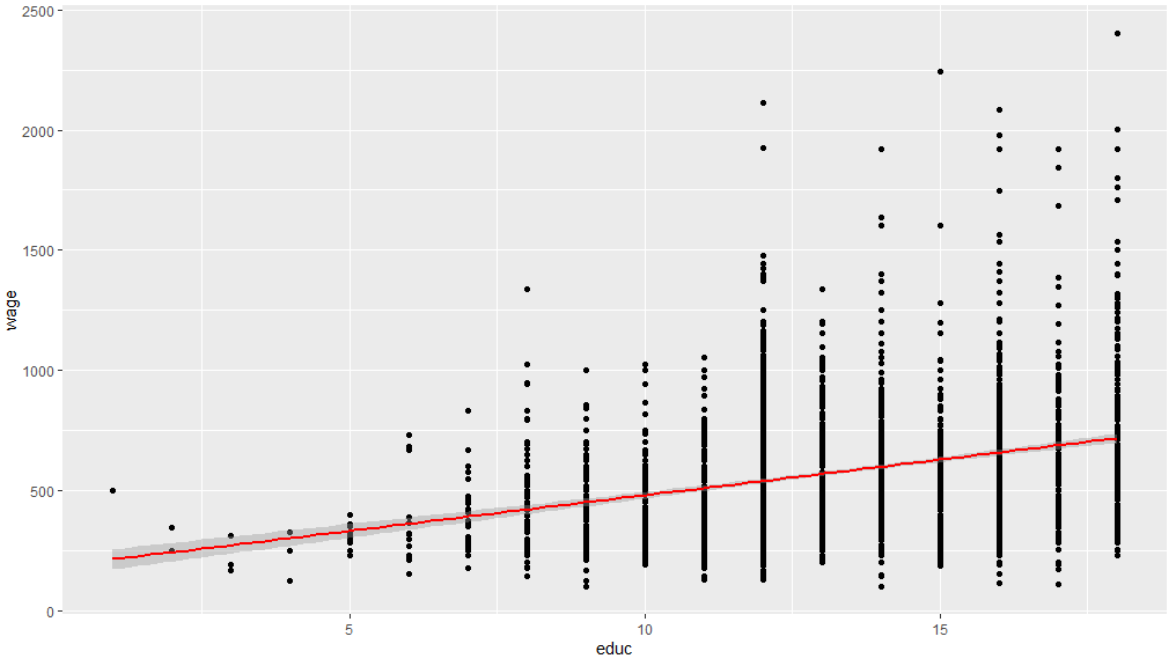
Por resultados prévios, sabemos que $Var(\varepsilon_i|x_1, ..., x_m) = Var(y_i|x_1, ..., x_m)$. Sendo assim, haverá heterocedasticidade sempre que $Var(y_i|x_1, ..., x_m)$ for uma função de $x_1, ..., x_m$.

Homocedasticidade importa pois ela garante que, num modelo linear, teremos:

- Os estimadores de MQO $\alpha_0, \alpha_1, \ldots, \alpha_n$ serão BLUE (melhores estimadores lineares não-viciados)
- A variância dos estimadores de MQO não será a menor de todos os estimadores
- O cálculo da variância dos estimadores levará em conta algum viés

Conclusão: por enviesar o erro-padrão dos estimadores de MQO, a presença de heterocedasticidade em um modelo linear pode levar a equívocos quando fazemos inferência.

Interpretação gráfica



Note como a variância do erro aumenta conforme a variável "educ" aumenta. Este é um exemplo claro de heterocedasticidade.

Teste de White

Como a homocedasticidade é uma hipótese muito importante para a qualidade de uma regressão linear, queremos poder diagnosticar se um modelo apresenta heterocedasticidade.
Nesse sentido, podemos usar o **Teste de White**. Relembre-se da equação (4). O Teste de White consiste justamente em testar se a variância do erro é uma função das observações.
Este teste permite detectar formas não-lineares de heteroscedasticidade (em oposição a outros testes, como Breusch-Pagan), e é muito bom em detectar formas de heterocedasticidade que invalidam os resultados garantidos pela homocedasticidade (que foram citados dois slides atrás).

Teste de White

Passo a passo do Teste de White:

1. Estime um modelo linear $Y = \alpha_0 + \alpha_1 X_1 + ... + \alpha_n X_n$
2. Calcule os erros ε_i
3. Estime por MQO o modelo

$$\varepsilon^2 = \delta_0 + \sum_{k=1}^{n} \delta_k X_k + \sum_{k=1}^{n} \delta_{n+k} X_k^2 + \delta_{2n+1} X_1 X_2 + ... + \delta_{2n+\frac{n(n-1)}{2}} X_{n-1} X_n \tag{5}$$

4. Teste a hipótese nula:
   H_0 :

$$\delta_1 = \cdots = \delta_{2n+\frac{n(n-1)}{2}} = 0 \tag{6}$$

   contra a hipótese alternativa:
   H_a:
   alguma das igualdades em (6) não vale.
5. Interprete o resultado do teste.

Teste de White

Note que a equação 5 tem três partes.

- Primeira parte: $\delta_1 X_1 + \cdots + \delta_n X_n$
- Segunda parte: $\delta_{n+1} X_1^2 + \cdots + \delta_{2n} X_n^2$
- Terceira parte: $\delta_{2n+1} X_1 X_2 + \cdots + \delta_{2n+\frac{n(n-1)}{2}} X_{n-1} X_n$

A primeira parte é uma combinação linear das variáveis preditoras
A segunda parte é uma combinação linear das variáveis preditoras ao quadrado
A terceira parte é uma combinação linear dos produtos cruzados das variáveis preditoras
Conclusão: o Teste de White testa diversas formas de dependência de ε^2 em $X_1, ..., X_n$

Teste de White: exemplo

Vamos usar o dataset presente no pacote wooldridge, no R. Ajustaremos um modelo de regressão linear para "lprice" em função de "crime" e "dist"

crime	dist	lprice
0.006	4.09	10.085809
0.027	4.97	9.980402
0.027	4.97	10.454495
0.032	6.06	10.416311
0.069	6.06	10.496787
0.030	6.06	10.264688
0.088	5.56	10.038892
0.145	5.95	10.207289
0.211	6.08	9.711116
0.170	6.59	9.846917

Teste de White: exemplo

Neste caso, nosso modelo para *lprice* será:

$$lprice = \alpha_0 + \alpha_1 X_{crime} + \alpha_2 X_{dist}$$

Nosso modelo para ε^2 será:

$$\varepsilon^2 = \delta_0 + \delta_1 X_{crime} + \delta_2 X_{dist} + \delta_3 X_{crime}^2 + \delta_4 X_{dist}^2 + \delta_5 X_{crime} X_{dist}$$