Esta aula

- Plano
  - R-quadrado ajustado
  - Testes de Hipóteses
  - Heterocedasticidade
  - Variáveis binárias

- Bibliografia
  - Wooldridge, J. M. Introductory Econometrics: A modern Approach, 6th Ed.

R-quadrado Ajustado

R-quadrado

Vimos que o R-quadrado amostral é definido por:

$$R^2 \equiv 1 - \frac{[SQR]}{[SQT]}$$

Já o R-quadrado polulacional é definido por:

$$R^2 \equiv 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

R-quadrado

- Problema: o R-quadrado sempre aumenta com a inclusão de uma variável adicional
- Já o R-quadrado ajustado é penalizado com o aumento do número de variáveis explicativas incluídas:

$$\bar{R}^2 \equiv 1 - \frac{\left[SSR/(n-k-1)\right]}{\left[SST/(n-1)\right]}$$

$$= 1 - \frac{\hat{\sigma}^2}{\left[SST/(n-1)\right]}$$

R-quadrado

- O R-quadrado ajustado pode ser utilizado para comparar o poder explicação de dois modelos diferentes com a mesma variável dependente y. Porém, apenas a teoria de RI pode predizer quais variáveis poderiam ser incluídas.
- Uma variável nunca deve ser incluída apenas porque aumenta o R-quadrado ajustado se não existe uma justificativa teórica para a sua inclusão.

Testes de Hipóteses

Regressão Linear - Inferência

- Para realizar testes de hipóteses, assumimos que u tem distribuição normal com média 0 e variância constante.

- Nesse caso, $$\left(\hat{\beta}_{\mathrm{j}} - \beta_j\right) \Big/ se\left(\hat{\beta}_j\right) \sim t_{n-k-1}$$

- IMPORTANTE: Mesmo que u não tenha distribuição normal, se a amostra for relativamente grande as estatística t e F usuais tem uma distribuição que converge para as distribuições t-student e F.

Regressão Linear - Inferência

- Podemos testar hipóteses com relação aos parâmetros populacionais

- Para isso, precisamos calcular o desvio-padrão associado aos estimadores.

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{SST_j\left(1 - R_j^2\right)}, \text{ onde}$$

Regressão Linear Simples - Inferência

- Estamos prontos para testar hipóteses com relação aos valores dos parâmetros populacionais

- Exemplo:

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

Regressão Linear - Inferência

- A estatística do teste é:

$$t = \left(\hat{\beta}_{\mathrm{j}}\right) \Big/ se\left(\hat{\beta}_{j}\right)$$

Teste de Hipótese

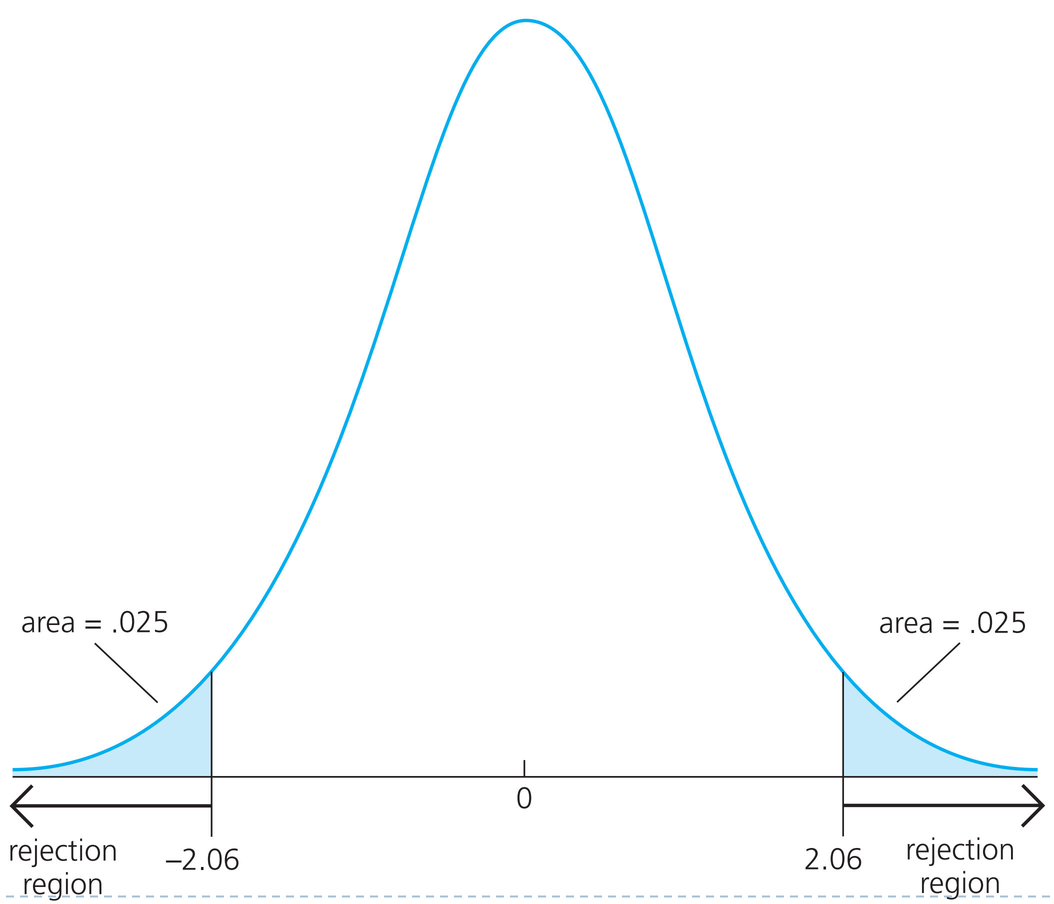
- As etapas do teste são:

1. Escrever as hipóteses alternativas e nulas
2. Escolher o nível de significância do teste α
3. Calcular a estatística t, conhecida como a **estatística do teste**
4. Encontrar o **valor crítico** do teste t^*,
5. Decidir: Se o valor absoluto de t for maior do que o de t^*, rejeitar H_0 com um nível de confiança de $1-\alpha$ **(teste bicaudal)**

Regressão Linear - Inferência

Valores Críticos

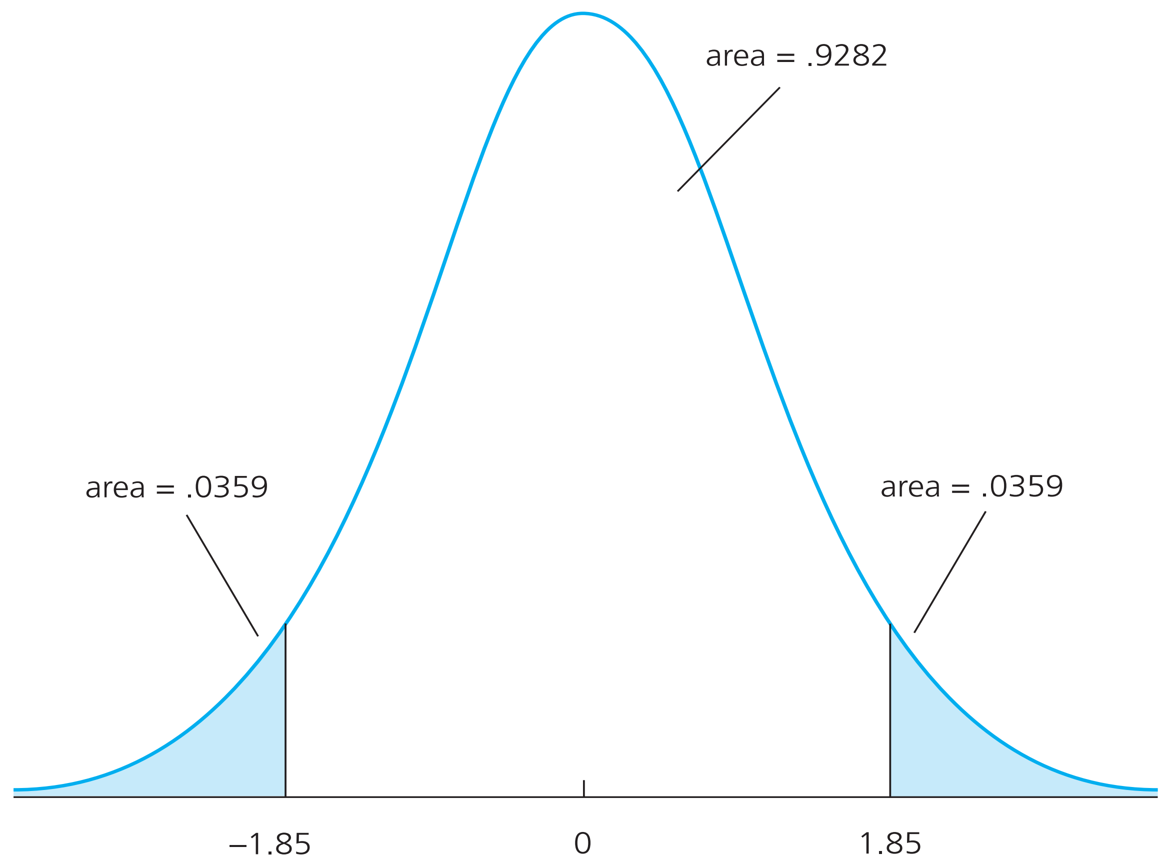
FIGURE 4.4 5% rejection rule for the alternative H_1: $\beta_j \neq 0$ with 25 *df*.



Regressão Linear - Inferência

P-value: $\mathrm{P}(|T| > |t|)$

FIGURE 4.6 **Obtaining the *p*-value against a two-sided alternative, when $t = 1.85$ and $df = 40$.**



Regressão Linear - Inferência

- Podemos também testar hipóteses conjuntas:

- Exemplo:

  H_0: $\beta_{k-q+1,\cdots..,}\beta_k = 0$

  H_1: *H_0 não é verdadeira*

  Ou, equivalentemente,

  H_0: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-q} x_{k-q} + u$
  H_1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-q-} x_{k-q} + \beta_{k-q+1} x_{k-q+1} \cdots + \beta_{k-} x_k + u$

Regressão Linear - Inferência

A estatística do teste é dada por:

$$F \equiv \frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-k-1)}$$

Onde:
SSR_r é a soma do quadrado dos resíduos do modelo restrito (sob H_0);
SSR_{ur} é a soma do quadrado dos resíduos do modelo irrestrito (sob H_1);
Q é o número de restrições;
K+1 é o número de parâmetros estimados.

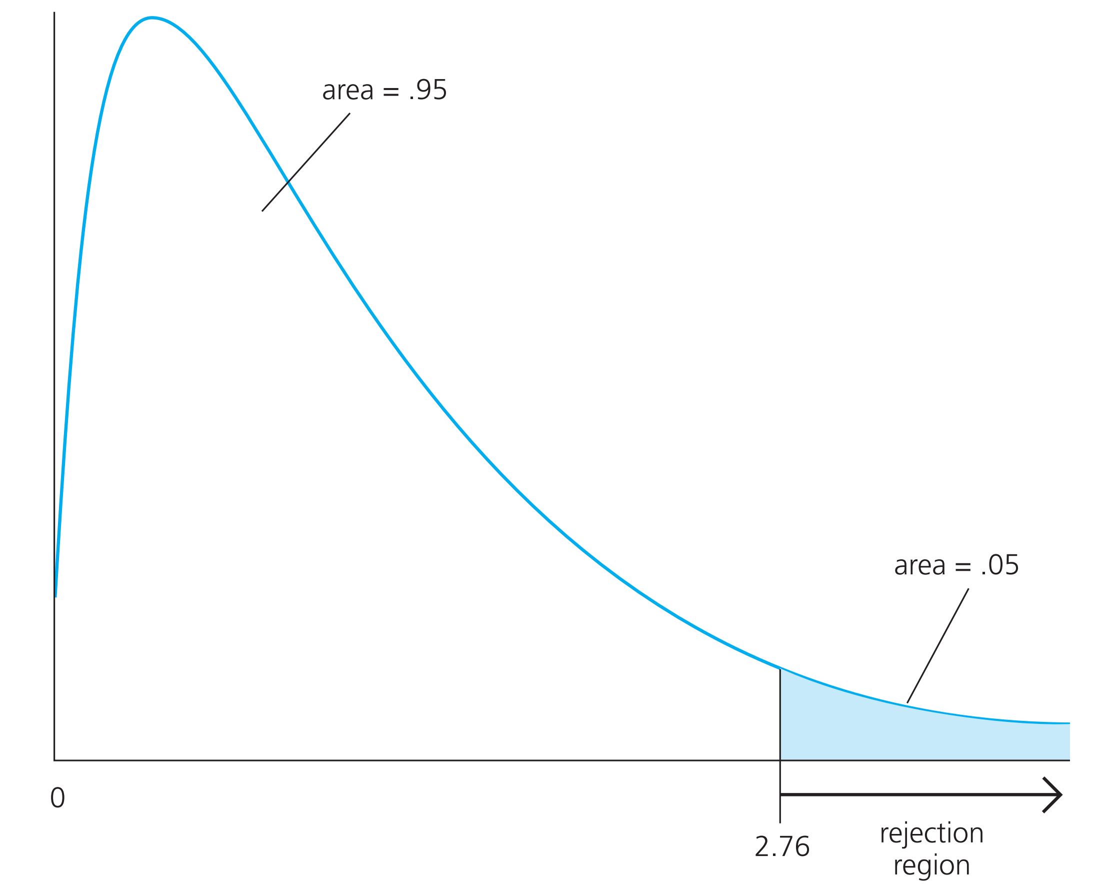
Teste de Hipótese

- As etapas do teste são:

1. Escrever as hipóteses alternativas e nulas
2. Escolher o nível de significância do teste α
3. Calcular a estatística F, conhecida como a **estatística do teste**
4. Encontrar o **valor crítico** do teste F^*,
5. Decidir: Se o valor de F for maior do que o de F^*, rejeitar H_0 com um nível de confiança de $1-\alpha$

Regressão Linear - Inferência

FIGURE 4.7 The 5% critical value and rejection region in an $F_{3,60}$ distribution.



Regressão Linear Simples - Inferência

- Se $SSR = SST*(1-R^2)$, a estatística F também pode ser reescrita como:

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)}$$

Regressão Linear Simples - Inferência

- E podemos testar o poder de explicação do modelo:

H_0 : todos os parâmetros (com exceção do intercepto) são equivalentes a zero

H_1 : pelo menos um dos parâmetros (com exceção do intercepto) é diferente de zero.

- Nesse caso, a estatística do teste é:

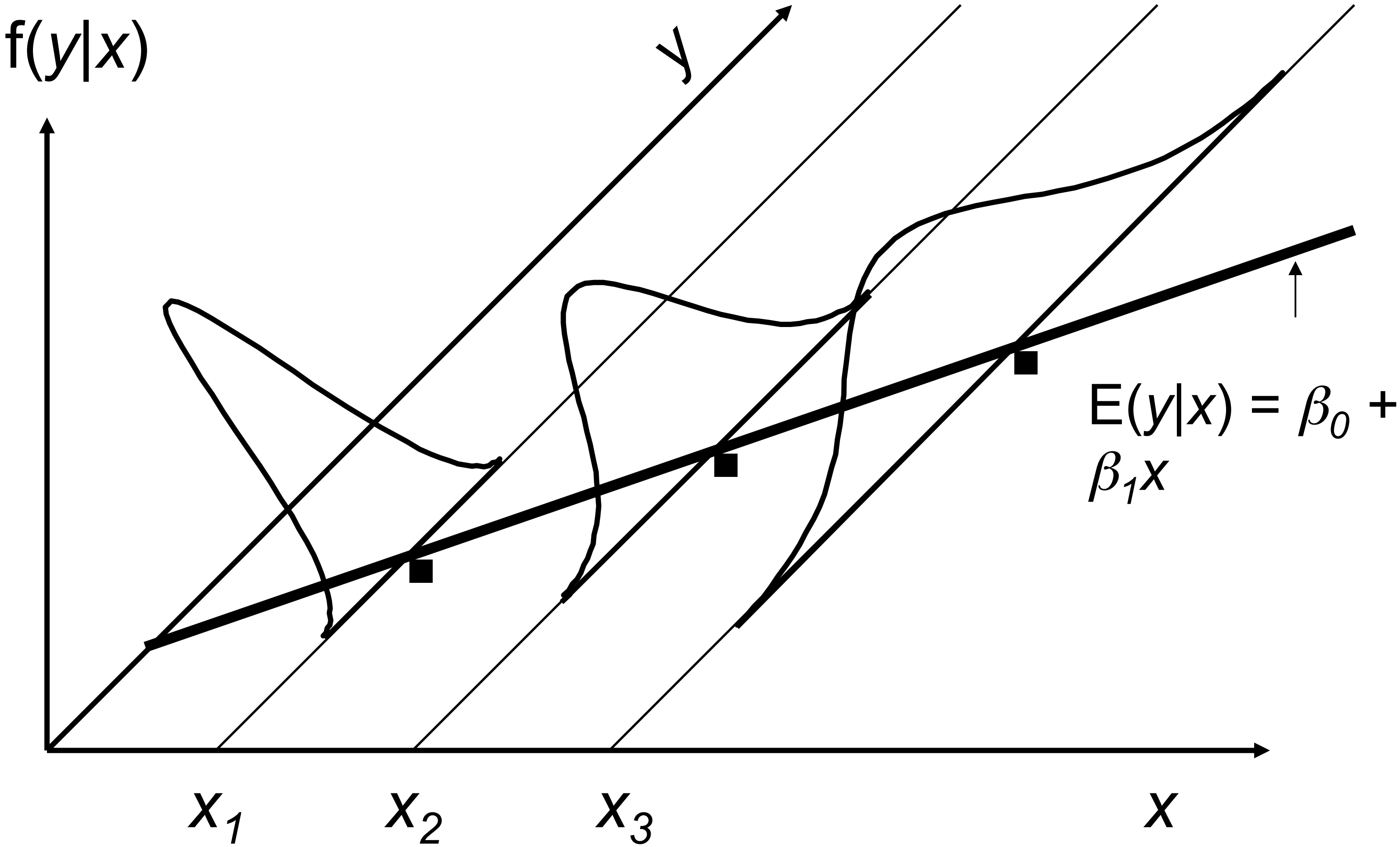
$$F = \frac{R^2 / k}{\left(1 - R^2\right) / \left(n - k - 1\right)}$$

Heterocedasticidade

Revisão: Homocedasticidade

- O pressuposto de homocedasticidade significa que a variância do erro não-observável é constante e independente do valor das variáveis explicativas

Exemplo de Heterocedasticidade



Consequências da Heterocedasticidade

- MQO é não-enviesado mesmo na presença de heterocedasticidade.
- Porém, nesse caso, os erros padrões são enviesados.
- Portanto, as estatisticas t e F não são válidas.

Variância com Homocedasticidade

$$Var\left(\hat{\beta}_1\right) = \sigma^2 \Big/ s_x^2$$

Variância com Homocedasticidade

- Quanto maior a variância do erro, maior a variância do estimador beta
- Quanto maior a variabilidade de x, menor essa variância.

Variância com Homocedasticidade

Um estimador não-enviesado de σ^2 é:

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum \hat{u}_i^2 = SQR/(n-2)$$

Variância com Homocedasticidade

Regressão Múltipla

Dados os pressupostos de Gauss-Markov, tem-se:

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{SST_j\left(1 - R_j^2\right)}, \text{ onde}$$

$$SST_j = \sum\left(x_{ij} - \overline{x}_j\right)^2 \text{ e } R_j^2 \text{ é o } R^2 da$$

$regressãode\ x_j$ nos demais x's

Heterocedasticidade

Importante: se o modelo tiver heterocedasticidade e os demais 4 pressupostos de Gauss-Markov continuarem válidos, o estimador MQO continua não-enviesado porém não é mais BLUE!

Além disso, testes de hipóteses baseadas nas variâncias dos parâmetros estimadas por MQO não são mais válidos

Variância com Heterocedasticidade

$$\text{Re}\,gressão\ \ Simples$$

$$Var\left(\hat{\beta}_1\right) = \frac{\sum\left(x_i - \overline{x}\right)^2 \sigma_i^2}{SST_x^2}, \text{ onde } SST_x = \sum\left(x_i - \overline{x}\right)^2$$

Um estimador convergente ou válido assintoticamente seria:

$$\frac{\sum\left(x_i - \overline{x}\right)^2 \hat{u}_i^2}{SST_x^2}$$

Variância com Heterocedasticidade

Regressão Múltipla

$$\widehat{Var}\left(\hat{\beta}_j\right) = \frac{\sum \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$, onde $\hat{r}_{ij}$ é o i^{th} resíduo da

regressão xj contra todas as outras variáveis explicativas e SSR_j é a soma do quadrado dos resíduos dessa regressão.

Variância com Heterocedasticidade

- A variância e desvio-padrão robusto somente levarão a estatísticas t e F válidas se as amostras forem grandes.

Teste para Heterocedasticidade: Breusch-Pagan

- Queremos testar H_0:$Var(u|x_1, x_2,\ldots, x_k) = \sigma^2$
- Se assumirmos uma relação linear entre u^2 e x_j:

$$u^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v$$

, podemos testar:

- H_0: $\delta_1 = \delta_2 = \ldots = \delta_k = 0$
- Esse é o teste Breusch-Pagan.

Teste White

- O Teste White permite que a heterocedasticidade dependa das variáveis explicativas ao quadrado e produtos cruzados.
- Para simplificar o teste, procedemos da seguinte forma:

Teste White

- Fazemos a regressão dos resíduos ao quadrado nos valores de $\hat{y}$ e $\hat{y}^2$ e fazemos um teste F no R-quadrado

Mínimos Quadrados Ponderados

- Apesar de conseguirmos obter erros padrões robustos, se soubermos a forma funcional da heterocedasticidade, podemos obter estimativas mais eficientes.
- O métodos dos mínimos quadrados ponderados (MQP) ou weighted least squares (WLS) permite transformar o modelo tal que ele tenha erros homocedásticos, soubermos a forma exata da heterocedasticidade. Nesse caso WLS is BLUE!
- Assim como em MQO, os testes t e F são assintoticamente válidos ou exatamente válidos se os erros possuírem distribuição Normal.

Mínimos Quadrados Ponderados

- Suponha que a heterocedasticidade seja da seguinte forma: $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$,
- Divindo todas as variáveis por $\sqrt{h}$, obtemos um modelo com erros homocedásticos:

$$y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} + \beta_1(x_{i1}/\sqrt{h_i}) + \beta_2(x_{i2}/\sqrt{h_i}) + \ldots$$
$$+ \beta_k(x_{ik}/\sqrt{h_i}) + (u_i/\sqrt{h_i})$$

- Nesse caso, $\mathrm{Var}(u_i/\sqrt{h_i}|\boldsymbol{x}) = \sigma^2$, visto que:

$$\mathrm{E}\left((u_i/\sqrt{h_i})^2\right) = \mathrm{E}(u_i^2)/h_i = (\sigma^2 h_i)/h_i = \sigma^2,$$

Mínimos Quadrados Ponderados

- Porém, o método dos mínimos quadrados ponderados já está programado no Stata.
- Apenas declare o peso 1/hi para efetuar a estimação (no stata declare [aweight=hi]).

FGLS

- Se não soubermos $h(\boldsymbol{x}_i)$, podemos estimar essa função
- Assume-se:

  $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k)$
- e estima-se δ!

FGLS

- Portanto:

  $u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k) v$

  onde $E(v|\boldsymbol{x}) = 1$

Equivalentemente:

$\ln(u^2) = \alpha_0 + \delta_1 x_1 + \ldots + \delta_k x_k + e$

FGLS

- Portanto, devemos adotar o seguinte procedimento:

1. Estimar o modelo por MQO
2. Obter os resíduos, $\hat{u}$
3. Gerar $\ln(\hat{u}^2)$
4. Fazer a regresão de $\ln(\hat{u}^2)$ nas variáveis independentes e obter os valores previstos $\hat{g}$. A variancia estimada é equivalente a $\exp(\hat{g})$.
5. Fazer a regressão pelo método dos mínimos quadrados ponderados com peso $\exp(\hat{g})$ no Stata. ($1/\exp(\hat{g})$ sendo o peso para os resíduos ao quadrado).

FGLS

- FGLS não é não-enviesado, mas ainda consistente e assintoticamente eficiente.

Variável Binária ou Dummy

Variável Binária ou "Dummy"

- Uma variável binária é aquela que toma dois valores possíveis, geralmente 0 e 1.
- No nosso banco de dados trabalhado nas últimas aulas, female é uma variável binária.

Variável Binária ou “Dummy”

- Considere o seguinte modelo com uma variável binária:
- $y = \beta_0 + \delta_0 d + \beta_1 x + u$
- Nesse caso, a variável d representa uma mudança de intercepto quando se passa de um grupo para o outro do banco de dados.
- If $d = 0$, $y = \beta_0 + \beta_1 x + u$
- If $d = 1$, $y = (\beta_0 + \delta_0) + \beta_1 x + u$

Variável Binária ou "Dummy"

- Considere agora o seguinte modelo:
- $y = \beta_0 + \delta_1 d + \beta_1 x + \delta_2 d^* x + u$
- If $d = 0$, $y = \beta_0 + \beta_1 x + u$
- If $d = 1$, $y = (\beta_0 + \delta_1) + (\beta_1 + \delta_2)\ x + u$
- Nesse modelo, a variável dummy permite uma mudança de intercepto, bem como uma mudança de inclinação.

Modelo Linear de Probabilidade

- Se a variável dependente for binária, então $P(y = 1|x) = E(y|x)$. Nesse caso:
- $P(y = 1|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- Nesse caso, cada intercepto beta i mede o impacto de variações marginais em xi na probabilidade do evento 1 ocorrer.
- O valor previsto de y nesse é a probabilidade estimada do evento 1 ocorrer.

Modelo Linear de Probabilidade

- Problema: nada impede que o y previsto não esteja no intervalo [0,1].
- Geralmente, esse modelo viola também o pressuposto de homocedasticidade.

Avaliação de Políticas/Programas

- Podemos utilizar variáveis dummy para avaliar o impacto de políticas/programas
- Por exemplo, qual o impacto da participação no programa Bolsa Família no nível educacional da família?

Avaliação de Políticas/Programas

- Problema: variáveis que influenciam a participação no Bolsa Família, como a renda dos ascendentes, também podem explicar o nível educacional.
- Isso levaria a viés nas estimações

Obrigada!