Regressão Linear Múltipla

Introdução

Formalização

Estimação

Interpretação

Propriedades

Tópicos adicionais

Efeito parcial

Variáveis omitidas

Variáveis irrelevantes

Estatística II

1º Sem. 2015/2016

Lic. Economia e Finanças

Modelo de Regressão Linear Múltipla

1 Introdução

2 Formalização do modelo

3 O estimador OLS

4 Interpretação dos coeficientes e das suas estimativas

5 Propriedades estatísticas

6 Alguns tópicos adicionais
- Efeito parcial
- Variáveis omitidas
- Variáveis irrelevantes

Introdução ao modelo

- A regressão múltipla permite controlar explicitamente o efeito na variável dependente de múltiplos factores, ajudando a construir melhores modelos para prever Y (R^2 mais elevado).
- Permite medir melhor o efeito *ceteris paribus*, porque introduz explicitamente vários factores que afectam Y, evitando colocá-los no erro u.
- Permite introduzir formas funcionais mais gerais e flexíveis (ex. quadrados, etc.)

Introdução ao modelo

Regressão Linear Múltipla
Introdução
Formalização
Estimação
Interpretação
Propriedades
Tópicos adicionais
Efeito parcial
Variáveis omitidas
Variáveis irrelevantes

- Considere-se o seguinte modelo

$$\begin{aligned} E\left(sal|educ, exper\right) &= \beta_0 + \beta_1 educ + \beta_2 exper \\ sal &= \beta_0 + \beta_1 educ + \beta_2 exper + u \end{aligned}$$

- Agora podemos ter em consideração explicitamente os efeitos de duas variáveis importantes.

Introdução ao modelo

- Mesmo que o interesse seja apenas conhecer β_1 este modelo é vantajoso pois permite ver o que acontece ao salário para um indivíduo que troca um ano de educação por um ano adicional de experiência.
- Note-se que

$$E\left(u|educ, exper\right) = 0.$$

- Como podemos interpretar isto?

Introdução ao modelo

- Suponha-se que *sal* é uma função de *educ* e *exper*, mas optamos pela regressão simples.

- A regressão simples não permite na maioria das situações uma análise *ceteris paribus* do efeito da variável *educ*, já que

$$corr(educ, exper) < 0 \Longrightarrow (\Delta educ > 0 \Longrightarrow \Delta exper < 0)$$

Introdução ao modelo

Regressão Linear Múltipla
Introdução
Formalização
Estimação
Interpretação
Propriedades
Tópicos adicionais
Efeito parcial
Variáveis omitidas
Variáveis irrelevantes

- Portanto no modelo simples

$$sal = \beta_0 + \beta_1 educ + u, \qquad E(u|educ) = 0$$

$$\Delta educ = 1 \Longrightarrow \Delta E(sal|educ) = \beta_1$$

mas o salário também é influenciado pela experiência através da escolaridade, e por isso β_1 reflecte o efeito líquido das *duas* variações.

Introdução ao modelo

- Com a regressão múltipla já é possível apurar o resultado de variações independentes

$$E(sal|educ, exper) = \beta_0 + \beta_1 educ + \beta_2 exper$$

$$\Delta educ = 1,\ \Delta exper = 0$$
$$\Rightarrow \Delta E(sal|educ, exper) = \beta_1$$

$$\Delta educ = 1,\ \Delta exper = 1$$
$$\Rightarrow \Delta E(sal|educ, exper) = \beta_1 + \beta_2$$

Introdução ao modelo

- Quanto mais variáveis forem incluídas no modelo, mais "pura" se torna a interpretação do coeficiente dos regressores.
- Mas não queremos introduzir todas as variáveis possíveis: por exemplo, não queremos incluir informação sobre a profissão do indivíduo pois a maior parte do efeito da educação é através da escolha da profissão.

Forma funcional

- Além disso, a regressão múltipla permite **flexibilizar a forma funcional:**

$$sal = \beta_0 + \beta_1 edu + \beta_2 exper + \beta_3 exper^2 + u$$

- Neste modelo, β_2 não tem a interpretação usual, vista no modelo anterior.

- Com efeito,

$$\frac{\Delta E(sal|educ, exper)}{\Delta exper} \approx \beta_2 + 2\beta_3 exper$$

Formalização do modelo

- Em geral, o modelo pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- Na amostra:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i, \quad (i = 1, \ldots, n)$$

Formalização do modelo

- **Forma vectorial:**

$$y_i = x_i\beta + u_i \quad (i = 1, \ldots, n)$$

com

$$x_i = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{ik} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Formalização do modelo

- **Forma matricial:**

$$y = X\beta + u$$

com

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Formalização do modelo

- $\beta_0, \beta_1, \ldots, \beta_k$ são parâmetros desconhecidos
- β_0 é o termo independente e β_j $(j = 1, \ldots, k)$ os coeficientes de declive do modelo
- y é o regressando, variável dependente, explicada ou de resposta
- x_j $(j = 1, \ldots, k)$ são os regressores, variáveis explicativas, independentes ou de controle
- $E(y|X) = X\beta$ (com β desconhecido)
- $u = y - E(y|X)$ não observável

O estimador OLS

- Se $\hat{\beta}$ é uma estimativa para β então

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

é um valor ajustado para y.

- Resíduo: $\hat{u} = y - \hat{y}$.

O estimador OLS

- Exemplo:

$$\widehat{sal} = -3.39 + 0.64educ + 0.07exper$$

- Se $(educ, exper, sal) = (12, 5, 5)$, como $\widehat{sal} = 4.64$ tem-se $\hat{u} = 0.36$.

O estimador OLS

- **Minimização da soma dos quadrados dos resíduos**:

  O estimador OLS, $\hat{\beta}$, para β é a solução do **problema de minimização**

$$\begin{aligned} \hat{\beta} : \min SSR &= \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right)^2 \\ &= \sum \hat{u}_i^2 \end{aligned}$$

Condições necessárias

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = 0 \iff -2\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}\right) = 0$$
$$\iff \sum \hat{u}_i = 0$$

$$\frac{\partial SSR}{\partial \hat{\beta}_j} = 0 \iff -2\sum x_{ij}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}\right) = 0$$
$$\iff \sum x_{ij}\hat{u}_i = 0 \qquad (j = 1, \ldots, k)$$

O estimador OLS

- Em **notação matricial** o problema escreve-se

$$\hat{\beta} : \min SSR = \sum \left(y_i - x_i \hat{\beta} \right)^2$$

e tem **solução**

$$\begin{aligned} \frac{\partial SSR}{\partial \hat{\beta}} = 0 \quad & \Longleftrightarrow \quad -2 \sum x_i' \left(y_i - x_i \hat{\beta} \right) = 0 \\ & \Longleftrightarrow \quad -2X' \left(y - X\hat{\beta} \right) = 0 \\ & \Longleftrightarrow \quad X'y = X'X\hat{\beta} \\ & \Longleftrightarrow \quad \hat{\beta} = \left(X'X \right)^{-1} X'y \end{aligned}$$

se $\left(X'X \right)^{-1}$ existir.

O estimador OLS

- O OLS é um **estimador do método dos momentos:**

$$E\left(u|x_1,\ldots,x_k\right)=0 \implies \begin{cases} E(u)=0 \\ E(x_j u)=0 \quad (j=1,\ldots,k) \end{cases}$$

e construímos o estimador que satisfaça

$$\begin{cases} \frac{1}{n}\sum \hat{u}_i = 0 \\ \frac{1}{n}\sum x_{ij}\hat{u}_i = 0, \quad (j=1,\ldots,k) \end{cases}$$

Propriedades dos resíduos

A partir das equações $\partial SSR/\partial\hat{\beta}=0$ podem retirar-se as seguintes propriedades para os resíduos:

1. Se houver termo independente, β_0, então

$$\sum \hat{u}_i = 0 \Longleftrightarrow \bar{\hat{u}} = 0$$

2. $\sum x_{ij}\hat{u}_i = 0 \quad (j = 1, 2, \ldots, k)$: os resíduos são ortogonais aos regressores.

Propriedades dos resíduos

Tal como no modelo de regressão simples, isto implica

3. $\bar{y} = \bar{x}\hat{\beta}$: a regressão estimada passa pelos pontos médios
4. $\bar{\hat{y}} = \bar{y}$
5. $\hat{y}'\hat{u} = \left(x\hat{\beta}\right)'\hat{u} = \hat{\beta}'x'\hat{u} = 0$

Coeficiente de determinação

- Medida da qualidade do ajustamento (goodness-of-fit).

1. $SST = \sum (y_i - \bar{y})^2$: **variação total**
2. $SSE = \sum (\hat{y}_i - \bar{y})^2$: **variação explicada**
3. $SSR = \sum \hat{u}_i^2$: **variação residual**

Coeficiente de determinação

- Com termo independente: $SST = SSE + SSR$.
- Esta decomposição permite avaliar a "qualidade do ajustamento" através do coeficiente de determinação:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \qquad \text{e} \qquad 0 \leq R^2 \leq 1$$

Coeficiente de determinação

- No modelo de regressão múltipla tem-se $R^2 = r^2_{y,\hat{y}}$, onde $r_{y,\hat{y}}$ é o coeficiente de correlação simples entre y e $\hat{y}$
- Note-se que o R^2 não diminui quando se aumenta o número de regressores.
- Se o modelo não tiver termo independente, o R^2 só pode ser calculado como $R^2 = r^2_{y,\hat{y}}$.

Coeficiente de determinação

Estatística II
1º Sem. 2015/2016

Lic. Economia e Finanças

Regressão Linear Múltipla
Introdução
Formalização
Estimação
Interpretação
Propriedades
Tópicos adicionais
Efeito parcial
Variáveis omitidas
Variáveis irrelevantes

- Exemplo:

$$\widehat{sal} = -3.39 + 0.64 \quad educ + 0.07 exper$$
$$R^2 = 0.2252$$

educ e *exper* explicam 22.52% da variação total de *sal* na amostra com a regressão linear OLS.

Interpretação dos coeficientes e das suas estimativas

- Consideremos o **modelo lin-lin**

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

estimado por

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

- Tem-se

$$\Delta x_j = 1 \underset{Cp}{\Longrightarrow} \Delta E(y|x_1, \ldots, x_k) = \beta_j \quad \text{e} \quad \Delta \hat{y} = \hat{\beta}_j$$

Interpretação dos coeficientes e das suas estimativas

- Exemplo:

$$\widehat{sal} = -3.39 + 0.64educ + 0.07exper$$

$$\Delta educ = 1 \underset{Cp}{\Longrightarrow} \Delta\widehat{sal} = 0.64$$

Interpretação dos coeficientes e das suas estimativas

- **Modelo log-lin**

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- **Semi-elasticidade** (aproximada):

$$\Delta x_1 = 1 \underset{Cp}{\Longrightarrow} \%\Delta\hat{y} \approx \hat{\beta}_1 \times 100\%$$

- Semi-elasticidade (exacta):

$$\Delta x_1 = 1 \underset{Cp}{\Longrightarrow} \%\Delta\hat{y} = \left(e^{\hat{\beta}_1} - 1\right) \times 100\%$$

Interpretação dos coeficientes e das suas estimativas

Regressão Linear Múltipla
Introdução
Formalização
Estimação
Interpretação
Propriedades
Tópicos adicionais
Efeito parcial
Variáveis omitidas
Variáveis irrelevantes

- Exemplo:

$$\widehat{\log(sal)} = 0.22 + 0.10 \quad educ + 0.01 exper$$

$$\Delta educ = 1 \underset{Cp}{\Longrightarrow} \%\Delta\widehat{sal} \approx 10\%$$

$$\%\Delta\widehat{sal} = \left(e^{0.1} - 1\right) \times 100\% = 10.52\%$$

Estatística II
1º Sem.
2015/2016

Lic.
Economia e
Finanças

Regressão Linear Múltipla
Introdução
Formalização
Estimação
Interpretação
Propriedades
Tópicos adicionais
Efeito parcial
Variáveis omitidas
Variáveis irrelevantes

- **Modelo log-log**

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2$$

$\beta_1 \approx$ **elasticidade**

$$\%\Delta x = 1\% \underset{Cp}{\Longrightarrow} \%\Delta y \approx \beta_1\%$$

Interpretação dos coeficientes e das suas estimativas

- Exemplo:

$$\widehat{\log(price)} = 0.22 + 0.70\log(sqrft) + 0.17\log(lotsize) + 0.04bdrms$$

$$\%\Delta sqrft = 1\% \underset{Cp}{\Longrightarrow} \%\Delta\widehat{price} \approx 0.70\%$$

Interpretação dos coeficientes e das suas estimativas

- **Modelo lin-log**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2$$

Se $\%\Delta x = 1\%$ então $\Delta \hat{y} = \frac{\beta_1}{100}$

Interpretação dos coeficientes e das suas estimativas

- Exemplo:

$$\widehat{price} = -2026.42 + 224.97\log(sqrft) + 61.46\log(lotsize) + 19.35\,bdrms$$

$$\%\Delta sqrft = 1\% \underset{Cp}{\Longrightarrow} \Delta\widehat{price} \approx 2.2497$$

Hipóteses de Gauss-Markov

- **MLR.1: Modelo linear nos parâmetros**

  $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

  (ou, na forma matricial, $y = X\beta + u$)

  $\beta_0, \beta_1, \ldots, \beta_k$ são parâmetros desconhecidos

  u é o erro ou variável residual (v.a. não observada)

Hipóteses de Gauss-Markov

- **MLR.2: Amostragem aleatória**

  As observações (y_i, x_i) $(i = 1 \dots n)$ são *i.i.d.*

- **MLR.3: Ausência de multicolinearidade perfeita**

  Nenhuma das variáveis explicativas é constante, e nenhuma é combinação linear das restantes.

  $r(X) = k + 1$

  Note que $r(X) = k + 1 \Longleftrightarrow X'X$ invertível, $n \geq k + 1$

Hipóteses de Gauss-Markov

- **MLR.4: Exogeneidade**

  $E(u|x_1, x_2, \ldots, x_k) = 0$

  (ou, na forma matricial, $E(u|X) = 0$)

- **MLR.5: Homocedasticidade**

  $Var(u|x_1, \ldots, x_k) = \sigma^2$

  (ou, na forma matricial, $Var\,(u|X) = E\,(uu'|X) = \sigma^2 I$)

Resultados importantes

Regressão Linear Múltipla

Introdução

Formalização

Estimação

Interpretação

Propriedades

Tópicos adicionais

Efeito parcial

Variáveis omitidas

Variáveis irrelevantes

- **Teorema 1:** Se as hipóteses MLR.1–MLR.4 estão satisfeitas então $E(\hat{\beta}) = \beta$.

- **Teorema 2:** Se as hipóteses MLR.1–MLR.5 estão satisfeitas então $Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$.

Exercício 7 Cap. 3

7 Which of the following can cause OLS estimators to be biased?

(i) Heteroskedasticity.

(ii) Omitting an important variable.

(iii) A sample correlation coefficient of .95 between two independent variables both included in the model.

Resultados importantes

- **Teorema 3 (Teorema de Gauss-Markov):**
  Satisfeitas as hipóteses MLR.1–MLR.5, o estimador dos mínimos quadrados é o estimador de variância mínima na classe dos estimadores lineares não enviesados.

- Esta propriedade pode ser expressa como:
  "$\hat{\beta}$ OLS é **BLUE** (Best Linear Unbiased Estimator)".

Resultados importantes

Regressão Linear Múltipla
Introdução
Formalização
Estimação
Interpretação
Propriedades
Tópicos adicionais
Efeito parcial
Variáveis omitidas
Variáveis irrelevantes

- **Corolário:** O estimador de variância mínima para a combinação linear $c\beta$ é dado por $c\hat{\beta}$.

- Exemplo: $\widehat{\beta_1 - \beta_2} = \hat{\beta}_1 - \hat{\beta}_2$ é o estimador mais eficiente para $\beta_1 - \beta_2$.

Exercício 6 Cap. 3

6 Consider the multiple regression model containing three independent variables, under Assumptions MLR.1 through MLR.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You are interested in estimating the sum of the parameters on x_1 and x_2; call this $\theta_1 = \beta_1 + \beta_2$.

(i) Show that $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ is an unbiased estimator of θ_1.

(ii) Find $\mathrm{Var}(\hat{\theta}_1)$ in terms of $\mathrm{Var}(\hat{\beta}_1)$, $\mathrm{Var}(\hat{\beta}_2)$, and $\mathrm{Corr}(\hat{\beta}_1, \hat{\beta}_2)$.

Resultados importantes

Regressão Linear Múltipla

Introdução

Formalização

Estimação

Interpretação

Propriedades

Tópicos adicionais

Efeito parcial

Variáveis omitidas

Variáveis irrelevantes

- **Teorema 4:** Satisfeitas as hipóteses MLR.1–MLR.5,

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n-k-1} = \frac{\sum \hat{u}_i^2}{n-k-1} = \frac{SSR}{n-k-1}$$

é estimador centrado de σ^2.

- Em consequncia deste resultado, o estimador de $Var(\hat{\beta}|X)$

$$\widehat{Var(\hat{\beta}|X)} = \hat{\sigma}^2 \left(X'X\right)^{-1}$$

é centrado.

Estimativa $\hat{\beta}_j$ como efeito parcial

O parâmetro $\hat{\beta}_j$ pode ser escrito como o coeficiente na regressão linear simples de y_i em $\hat{r}_{ij}$

$$\hat{\beta}_j = \frac{\sum \hat{r}_{ij} y_i}{\sum \hat{r}_{ij}^2} \qquad (j = 1, \ldots, k)$$

onde $\hat{r}_j$ é o resíduo da regressão de x_j nos outros regressores

$$\hat{r}_{ij} = x_{ij} - \hat{x}_{ij} \qquad (i = 1, \ldots, n)$$

com

$$\hat{x}_{ij} = \hat{\gamma}_0 + \hat{\gamma}_1 x_{i1} + \cdots + \hat{\gamma}_{j-1} x_{ij-1} + \hat{\gamma}_{j+1} x_{ij+1} + \cdots + \hat{\gamma}_k x_{ik}$$

($\hat{r}_{ij}$ é o que resta em x_{ij} depois de expurgar o efeito das outras variáveis)

Estimativa $\hat{\beta}_j$ como efeito parcial

- Isto salienta que o sinal de um parâmetro num modelo de regressão múltipla nem sempre é o que se espera, pois depende do *coeficiente de correlação parcial* entre y e x_j dados os restantes regressores, e não do coeficiente de correlação simples entre elas.

- Exemplo:

$$\widehat{sal} = -3.39 + 0.64educ + 0.07exper$$

Estimativa $\hat{\beta}_j$ como efeito parcial

A estimativa $\hat{\beta}_1 = 0.64$ pode ser obtida de forma equivalente seguindo-se os passos:

1. Regressão OLS de *educ* sobre o termo independente e *exper* obtendo-se os respectivos resíduos

$$\begin{aligned} \widehat{educ} &= 13.60 - 0.06 exper \\ &\Downarrow \\ \hat{r}_{educ} &= educ - \widehat{educ} \end{aligned}$$

2. Regressão OLS de *sal* sobre os resíduos da regressão anterior obtendo-se

$$\widehat{sal} = 0.64 \hat{r}_{educ}$$

Variáveis omitidas

- Considere-se um modelo em que x está particionada em z e w

$$E\left(y|z,w\right)=z\beta_1+w\beta_2$$

- Se omitirmos w e estimarmos β_1, obtemos

$$\hat{\beta}_1=\left(z'z\right)^{-1}z'y$$

Variáveis omitidas

- O seu valor esperado é dado por:

$$
\begin{aligned}
E\left(\hat{\beta}_1|z,w\right) &= \left(z'z\right)^{-1} z' E\left(y|z,w\right) \\
&= \left(z'z\right)^{-1} z' \left[z\beta_1 + w\beta_2\right] \\
&= \beta_1 + \left(z'z\right)^{-1} z' w\beta_2 = \beta_1 + \hat{\gamma}\beta_2
\end{aligned}
$$

onde $\hat{\gamma}$ é o declive da regressão de w sobre z.

Variáveis omitidas

Consequência: $\hat{\beta}_1$ é um **estimador enviesado** para β_1 excepto se uma ou outra das condições seguintes se verificar

1. $\beta_2 = 0$, o que significaria que, afinal, o verdadeiro modelo não inclui o regressor w
2. $\hat{\gamma} = 0$, caso em que as variáveis z e w não estão correlacionadas.

Variáveis omitidas: Exemplo

$$
\begin{aligned}
wage &= \beta_0 + \beta_1 educ + \beta_2 exper + u \\
\widehat{wage} &= -0.9 + 0.54 educ \\
\widehat{exper} &= 35.46 - 1.47 educ \\
env(\hat{\beta}_1) &= -1.47\beta_2
\end{aligned}
$$

O sinal da estimativa $\hat{\gamma} = -1.47$ é o expectável e, como o sinal espectável de β_2 é positivo, admite-se que $env(\hat{\beta}_1) < 0$. E, na verdade, tem-se

$$
\widehat{wage} = -3.39 + 0.64 educ + 0.07 exper
$$

Variáveis irrelevantes

- Suponha a seguinte situação:

  Verdadeiro modelo (verifica MLR.1 a MLR.4):

  $$y = \beta_0 + \beta_1 x_1 + u$$

  Modelo (incorrectamente) especificado:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v \xrightarrow{OLS} \tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2$$

- Tem-se $E(y|x_1, x_2) = E(y|x_1)$, isto é, dado x_1, o valor da variável x_2 não afecta y.

Variáveis irrelevantes

- Portanto, o modelo especificado também verifica as hipóteses MLR.1 a MLR.4, $\widetilde{\beta}$ é BLUE, $E(\widetilde{\beta}_1) = \beta_1$ e $\widetilde{\beta}_2 \approx 0$.

- Conclusão: a inclusão de variáveis irrelevantes **não enviesa o estimador**.

- Todavia, a inclusão de variáveis irrelevantes **pode afectar a variância**.

Variáveis irrelevantes

Regressão Linear Múltipla
Introdução
Formalização
Estimação
Interpretação
Propriedades
Tópicos adicionais
Efeito parcial
Variáveis omitidas
Variáveis irrelevantes

Num modelo com constante, a igualdade

$$Var\left(\hat{\beta}|X\right) = \sigma^2\left(X'X\right)^{-1}$$

pode ser expressa, para $j = 1, \ldots, k$, na forma

$$Var\left(\hat{\beta}_j|X\right) = \frac{\sigma^2}{\left(1 - R_j^2\right)\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

onde R_j^2 é o coeficiente de determinação da regressão de x_j nos restantes regressores.

Variáveis irrelevantes

- Como se sabe, **adicionar regressores não pode fazer descer o R^2**, que geralmente aumenta.

- A inclusão de variáveis irrelevantes não reduz σ^2 mas geralmente aumenta R_j^2, **reduzindo e eficiência do estimador**.

Variáveis irrelevantes

Esta fórmula mostra claramente os factores que influenciam a variância dos estimadores dos coeficientes do modelo, $Var(\hat{\beta}_j|x)$. A variância depende

1. directamente de σ^2
2. inversamente de $\sum_{i=1}^{n}(x_{ji}-\bar{x}_j)^2$, isto é, da variabilidade de x_j na amostra e da dimensão n
3. directamente de R_j^2

Variáveis irrelevantes

- Note que incluir variáveis não correlacionadas com x_j mas que tenham coeficiente não nulo reduz a variância do estimador pois reduz a variância do erro, σ^2.

- Note que $R_j^2 = 1$ significa que existe **multicolinearidade**.

Novamente o coeficiente de determinação

Se adicionarmos variáveis explicativas ao modelo, mesmo sem poder explicativo significativo, o coeficiente de determinação

$$R^2 = 1 - \frac{SSR}{SST}$$

nunca diminui e tem tendência a aumentar.

Novamente o coeficiente de determinação

Daí o interesse numa medida que reflicta mais adequadamente o poder explicativo do modelo, o **coeficiente de determinação ajustado**

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{SST/\left(n-1\right)} = 1 - \frac{SSR/\left(n-k-1\right)}{SST/\left(n-1\right)}$$

$\bar{R}^2$ pode tomar **valores negativos**.

Novamente o coeficiente de determinação

Veja-se que R^2 pode ser definido como

$$R^2 = 1 - \frac{SSR \;/\, (n-1)}{SST \;/\, (n-1)} = 1 - \frac{s_{\hat{u}}^{\prime 2}}{s_y^{\prime 2}}$$

enquanto que

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{s_y^{\prime 2}}$$

Novamente o coeficiente de determinação

- Ambos são estimadores do parâmetro da população $1 - \sigma_u^2/\sigma_y^2$, tendo a mesma interpretação.

- Todavia $\bar{R}^2$ é **melhor estimador** porque considera $\hat{\sigma}^2$ em vez de $s_{\hat{u}}^{\prime 2}$.

Exercício 5 Cap. 3

5 In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

(i) In the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

does it make sense to hold *sleep*, *work*, and *leisure* fixed, while changing *study*?

(ii) Explain why this model violates Assumption MLR.3.

(iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?