Estatística II

1º Sem. 2015/2016

Lic. Economia e Finanças

Inferência no Modelo de Regressão Linear

Inferência estatística no Modelo de Regressão Linear

O objectivo agora é fazer inferência sobre os parâmetros do modelo. Exemplo:

$$colgpa = \beta_0 + \beta_1 hsgpa + \beta_2 skipped + u$$

- *colgpa*: nota média ensino superior
- *hsgpa*: nota média ensino secundário
- *skipped*: faltas semanais nas aulas

Será $\beta_2 = 0$? Será $\beta_1 = 1$?

Inferência estatística no Modelo de Regressão Linear

Partindo dos dados, poderá ter interesse:

1. Estimar os coeficientes pelo OLS e estimar as variâncias;
2. Construir intervalos de confiança para os coeficientes;
3. Testar hipóteses sobre os parâmetros ou suas combinações.

Vários tipos de hipóteses

- $H_0 : \beta_j = 0$, $H_1 : \beta_j \neq 0$: significância individual (testa se cada regressor é individualmente relevante)
- $H_0 : \beta_1 = \beta_2$, $H_1 : \beta_1 < \beta_2$: combinações
- $H_0 : \beta_3 = 0$, $\beta_4 = 0$: significância conjunta (restrições de exclusão)
- $H_0 : \beta_j = 0 \;\; \forall j > 0$: significância global (teste a todos os coeficientes de declive).

Inferência estatística no Modelo de Regressão Linear

- É necessário encontrar uma estatística de teste com distribuição conhecida.
- Para isso, há que introduzir a distribuição da população.

Distribuição amostral do estimador OLS

- **MLR.6: Normalidade**

$$y|x \sim N(x\beta, \sigma^2 I) \Leftrightarrow u \sim N(0, \sigma^2 I)$$

(sob homocedasticidade e independência)

- Sob esta hipótese o OLS não só é BLUE como é MVUE.
- Isto resulta de sob estas hipóteses o OLS ser o MLE (desigualdade de **Fréchet-Cramér-Rao**).

Distribuição amostral do estimador OLS

- A hipótese de normalidade é muito forte e mais tarde vamos ver em que condições a podemos dispensar.
- Casos em que não se verifica:
  - Salários e despesa (usualmente logaritmizadas);
  - Variáveis dependentes discretas que tomam poucos valores: binárias, contagem

    (exemplos: ser ou não fumador, ser ou não repetente, número de detenções de uma pessoa, número de disciplinas em atraso, número de idas ao cinema num mês).

Distribuição amostral do estimador OLS

Notando que

$$\hat{\beta} = \left(x'x\right)^{-1} x'y = \beta + \left(x'x\right)^{-1} x'u$$

segue que o estimador OLS é uma combinação linear de variáveis aleatórias normais independentes, e portanto tem também distribuição Normal.

Distribuição amostral do estimador OLS

Esta hipótese implica que

$$\hat{\beta}|x \sim N(\beta, \sigma^2 \left(x'x\right)^{-1});$$

$$\hat{\beta}_j|x \sim N(\beta_j, Var\left(\hat{\beta}_j\right));$$

$$\left(\hat{\beta}_j - \beta_j\right) \Big/ sd\left(\hat{\beta}_j\right) \sim N(0,1);$$

onde $sd\left(\hat{\beta}_j\right)$ é a raiz quadrada do j-ésimo elemento da diagonal principal de $\sigma^2 \left(x'x\right)^{-1}$.

Distribuição amostral do estimador OLS

- Já que σ^2 é desconhecido, utilizamos a sua estimativa $\hat{\sigma}^2 = SSR/(n-k-1)$.

- O seguinte resultado vale:

$$\frac{\sum_{i=1}^{n} \widehat{u}_i^2}{\sigma^2} = \frac{(n-k-1)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k-1)}$$

- Portanto: $\left(\hat{\beta}_j - \beta_j\right) \Big/ se\left(\hat{\beta}_j\right) \sim t_{(n-k-1)}$

  onde $se\left(\hat{\beta}_j\right)$ é a raiz quadrada do j-ésimo elemento da diagonal principal de $\hat{\sigma}^2\,(x'x)^{-1}$.

Testes de hipóteses sobre um parâmetro

- Considere-se a seguinte hipótese simples:

$$H_0 : \beta_j = 0$$

  (significância estatística individual).

- Esta hipótese significa que, dadas as restantes variáveis, x_j não tem efeito em y.

Testes de hipóteses sobre um parâmetro

- A estatística de teste a usar é

$$t = \left(\hat{\beta}_j - 0\right) \Big/ se\left(\hat{\beta}_j\right) \sim t_{(n-(k+1))}$$

- A alternativa pode ser unilateral ou bilateral.
- Geralmente os testes em econometria são bilaterais.

Testes de hipóteses sobre um parâmetro

Podem, obviamente, considerar-se hipóteses mais gerais:

$$H_0 \quad : \quad \beta_j = a_j$$

$$t \quad = \quad \frac{\hat{\beta}_j - a_j}{se\left(\hat{\beta}_j\right)} \sim t_{(n-(k+1))}$$

Significância económica e significância estatística

- Com amostras muito grandes, os parâmetros são estimados com muita precisão pelo que mesmo que um parâmetro seja ínfimo pode ter um t-ratio relativamente grande.
- Portanto, significância estatística pode não implicar significância prática ou económica.
- O oposto também é verdadeiro: com amostras pequenas, os parâmetros são estimados com pouca precisão pelo que os t-ratios tendem a ser baixos.
- Portanto, variáveis com grande significância económica podem ser estatisticamente não significantes.

Intervalos de confiança

- Usando estes resultados é fácil obter intervalos de confiança para os parâmetros:

$$\hat{\beta}_j \pm t_{\alpha/2}\,(n-k-1)\,se\left(\hat{\beta}_j\right)$$

- Daqui podem fazer-se testes de hipóteses, rejeitando-se os valores a_j que não estão contidos no intervalo.

Hipóteses sobre uma combinação linear de parâmetros

- Considere-se o seguinte modelo

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

e suponha-se que a hipótese de interesse é

$$H_0 : \beta_2 = \beta_3 \Leftrightarrow H_0 : \beta_3 - \beta_2 = 0$$

Hipóteses sobre uma combinação linear de parâmetros

- A estatística de teste terá a forma

$$t = \frac{\hat{\beta}_3 - \hat{\beta}_2 - 0}{se\left(\hat{\beta}_3 - \hat{\beta}_2\right)} \sim t\,(n - k - 1)$$

- Como calcular este desvio padrão?

Hipóteses sobre uma combinação linear de parâmetros

$$
\begin{aligned}
Var\left(a\hat{\beta}_2 + b\hat{\beta}_3\right) &= E\left\{\left[\left(a\hat{\beta}_2 + b\hat{\beta}_3\right) - \left(a\beta_2 + b\beta_3\right)\right]^2\right\} \\
&= E\left\{\left(a\hat{\beta}_2^* + b\hat{\beta}_3^*\right)^2\right\} \\
&= E\left\{a^2\left(\hat{\beta}_2^*\right)^2 + b^2\left(\hat{\beta}_3^*\right)^2 + 2ab\left(\hat{\beta}_2^*\hat{\beta}_3^*\right)\right\} \\
&= a^2 Var\left(\hat{\beta}_2\right) + b^2 Var\left(\hat{\beta}_3\right) + 2abCov\left(\hat{\beta}_2, \hat{\beta}_3\right)
\end{aligned}
$$

onde $\hat{\beta}_j^* = \hat{\beta}_j - \beta_j$. Segue que:

$$
se\left(\hat{\beta}_3 - \hat{\beta}_2\right) = \sqrt{Var\left(\hat{\beta}_2\right) + Var\left(\hat{\beta}_3\right) - 2Cov\left(\hat{\beta}_2, \hat{\beta}_3\right)}
$$

Hipóteses sobre uma combinação linear de parâmetros

- Em geral, dada a combinação linear $R\hat{\beta}$, tem-se

$$Var\left(R\hat{\beta}\right) = RVar\left(\hat{\beta}\right)R' = \sigma^2 R\left(x'x\right)^{-1}R'$$

- No caso anterior, considerando $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]'$ e $R = [0, 0, -1, 1]$, tem-se $R\hat{\beta} = \hat{\beta}_3 - \hat{\beta}_2$.

Hipóteses sobre uma combinação linear de parâmetros

- Alternativamente, pode-se definir $\theta = \beta_3 - \beta_2$ e substituir $\beta_3 = \theta + \beta_2$:

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (\theta + \beta_2) x_3 + u \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + x_3) + \theta x_3 + u
\end{aligned}
$$

- Estimando este modelo posso testar $H_0 : \beta_2 = \beta_3$ como um simples teste para a significância individual de θ.

Hipóteses sobre uma combinação linear de parâmetros

- **Nota:** Este método pode ser utilizado também para calcular intervalos de confiança para θ, uma vez que a estimação OLS fornece o $se(\hat{\theta})$:

$$\hat{\theta} \pm t_{\alpha/2}\,(n-k-1)\,se\left(\hat{\theta}\right)$$

- Exemplo: $H_0 : \beta_1 - 3\beta_2 = 1$

  Define-se $\theta = \beta_1 - 3\beta_2$, substitui-se no modelo

  $\beta_1 = \theta + 3\beta_2$ e depois testa-se $H_0 : \theta = 1$.

Hipóteses sobre combinações lineares de parâmetros

- Suponha-se que se quer testar um conjunto de q restrições lineares sobre β que pode ser escrito como:

$$H_0 : R\beta = r$$

em que R é uma matriz $q \times (k+1)$ e r um vector $q \times 1$.

- Exemplo $H_0 : \beta_1 = \beta_2, \beta_3 = 0, \beta_4 = 1$

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Hipóteses sobre combinações lineares de parâmetros

- Uma medida natural da validade da hipótese $H_0 : R\beta = r$ é dada por $\left(R\hat{\beta} - r\right)$.

- É fácil ver que **se** H_0 **for verdadeira**,

$$E\left(R\hat{\beta} - r\right) = 0.$$

- Portanto, se $\left(R\hat{\beta} - r\right)$ está "próximo" de zero, tendo a não rejeitar, se está "longe" tendo a rejeitar.

Hipóteses sobre combinações lineares de parâmetros

- Como medir a distância?
- Já sabemos que $Var\left(R\hat{\beta}\right) = \sigma^2 R\left(x'x\right)^{-1} R'$, pelo que

$$\left(R\hat{\beta} - r\right)' \left[\sigma^2 R\left(x'x\right)^{-1} R'\right]^{-1} \left(R\hat{\beta} - r\right) \sim \chi^2(q)$$

- Como σ^2 não é conhecido, tem de se fazer a habitual passagem à distribuição F.

Hipóteses sobre combinações lineares de parâmetros

Lembrando que

$$(n-k-1)\,\frac{\hat{\sigma}^2}{\sigma^2} = \left(\frac{\hat{u}}{\sigma}\right)'\left(\frac{\hat{u}}{\sigma}\right) \sim \chi^2\,(n-k-1)$$

e fazendo a razão das χ^2 (independentes) divididas pelos respectivos graus de liberdade, obtém-se:

$$F = \left[\frac{n-k-1}{q}\right]\frac{\left(R\hat{\beta}-r\right)'\left[R\,(x'x)^{-1}R'\right]^{-1}\left(R\hat{\beta}-r\right)}{\hat{u}'\hat{u}}$$

$$F \sim F\,(q, n-k-1)$$

Hipóteses sobre combinações lineares de parâmetros

- Esta fórmula é sempre aplicável, mas geralmente pode ser escrita de forma muito mais conveniente:

$$F = \left[\frac{n-k-1}{q} \right] \frac{SSR_r - SSR_u}{SSR_u} \sim F\left(q, n-k-1\right)$$

em que SSR_u é o SSR do modelo sem restrições, e SSR_r é o SSR do modelo em que foram impostas as restrições.

- Quando $q = 1$ a estatística F pode ser obtida como o quadrado de uma t.

Hipóteses sobre combinações lineares de parâmetros

Se as variáveis dependentes dos modelos restrito e não restrito são idênticas $(r = 0)$, então

$$F = \left[\frac{n-k-1}{q}\right]\frac{R_u^2 - R_r^2}{1 - R_u^2} \sim F\left(q, n-k-1\right),$$

em que R_u^2 é o R^2 do modelo sem restrições, e R_r^2 é o R^2 do modelo em que foram impostas as restrições.

Exemplo de caso em que $r \neq 0$

Considere-se o modelo:

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + u$$

- *salary*: salário total de um jogador de baseball em 1993
- *years*: número de anos que joga na liga
- *gamesyr*: número médio de jogos que jogou por ano
- *bavg*: média em toda a carreira de "Battings"

Seja: $H_0 : \beta_1 - \beta_2 = 0, \beta_3 = 1$

Exemplo de caso em que $r \neq 0$

- Impondo as restrições:

$$H_0 : \beta_1 - \beta_2 = 0, \beta_3 = 1$$

obtém-se o modelo:

$$\log(salary) = \beta_0 + \beta_2 \left(gamesyr + years\right) + bavg + u$$

- O modelo restrito é o seguinte:

$$\log(salary) - bavg = \beta_0 + \beta_2 \left(gamesyr + years\right) + u$$

Teste de significância global da regressão

- Uma hipótese que é sempre testada é

$$H_0 : \beta_j = 0, \forall j > 0$$

- Neste caso, $r = 0$ e

$$R = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$F = \left[\frac{n-k-1}{k}\right] \frac{R_u^2}{1-R_u^2} \sim F\left(k, n-k-1\right),$$

Teste de restrições de exclusão

- Mais útil é geralmente testar a nulidade conjunta de parâmetros cuja nulidade individual não é rejeitada.

- Neste caso, supondo que os parâmetros cuja nulidade conjunta se quer testar são os últimos q,

$$R = [0|I_q]$$