Matematické Metody v Ekonometrii 2.

The Simple Regression Model

Blanka Šedivá

KMA

zimní semestr 2016/2017

Simple linear regression model

- We assume a model $Y_i = \beta_1 + \beta_2\ x_i + \varepsilon_i, i = 1, 2, \ldots, n$
- Population regression line $f(x_i) = \beta_1 + \beta_2\ x_i$,
- The regression coefficients; β_1 and β_2 are also known as intercept and slope coefficients,
- ε_i stochastic disturbance or stochastic error term,
- $\mathrm{E}\,(Y|X = x) = \beta_1 + \beta_2\ x$ population regression function
- Samples $y_1, y_2, \ldots, y_n \Rightarrow$ the sample regression function
- Estimators of β_1, β_2 depend on samples $y_1, y_2, \ldots, y_n$
- The hat symbol denotes an estimated value $\hat{\beta}_1, \hat{\beta}_2 \Rightarrow$

$$\hat{\beta}_j = \hat{\beta}_j\,(y_1, y_2, \ldots, y_n)$$

- The sample regression function is based on the sample $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2\ x_i$

The classical linear regression model (CLRM) I

- **Assumption 1**: Linear regression model. The regression model is linear in the parameters
- **Assumption 2**: Values taken by the regressor x are considered fixed in repeated samples.
  More technically, x is assumed to be nonstochastic or the average value of ε does not depend on the value of x or $\mathrm{E}(\varepsilon|x) = \mathrm{E}(\varepsilon)$.
  Zero covariance between ε_i and x_i.
  Symbolically $\mathrm{cov}(\varepsilon_i, x_i) = 0$ or $\mathrm{E}(\varepsilon_i x_i) = 0$.
- **Assumption 3**: Zero mean value of disturbance ε_i. Given the value of x, the mean, or expected, value of the random disturbance term ε_i is zero. Technically, the conditional mean value of ε_i is zero.
  Symbolically, we have $\mathrm{E}(\varepsilon_i|x_i) = 0$

The classical linear regression model (CLRM) II

- **Assumption 4**: Homoscedasticity or equal variance of ε_i. Given the value of x, the variance of ε_i is the same for all observations. That is, the conditional variances of ε_i are identical.
  Symbolically, we have $\mathrm{var}\left(\varepsilon_i | x_i\right) = \sigma^2$
- **Assumption 5**: No autocorrelation between the disturbances. Given any two x values, x_i and x_j $(i \neq j)$, the correlation between any two ε_i and ε_j is zero.
  Symbolically, $\mathrm{cov}\left(\varepsilon_i, \varepsilon_j | x_i, x_j\right) = 0$
- **Assumption 7**: The number of observations n must be greater than the number of parameters to be estimated. Alternatively, the number of observations n must be greater than the number of explanatory variables.
- **Assumption 8**: Variability in x values. The x values in a given sample must not all be the same.
  Technically, $\mathrm{var}\left(x\right)$ must be a finite positive number

Gauss–Markov Theorem

Given the assumptions (A1-A8) of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.

- (**L**inearity of OLS) b_1, b_2 are linear combination of $y_1, y_2, \ldots, y_n$
- (**U**nbiasedness of OLS) $\mathrm{E}(b_1) = \beta_1$, $\mathrm{E}(b_2) = \beta_2$
- (Variance) $\operatorname{var} b_1 = \sigma^2 \frac{\sum x_i^2}{n \sum (x_i - \overline{x})^2}$ $\qquad \operatorname{var} b_2 = \sigma^2 \frac{1}{\sum (x_i - \overline{x})^2}$
- (**B**est) means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators
- The estimating of error variance $s^2 = \dfrac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}$ is unbiased estimator of σ^2, in other words $\mathrm{E}(s^2) = \sigma^2$

Gauss–Markov Theorem for combination of parameters

Given the assumptions of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.

- (BLUES) of $\Theta = a_1\beta_1 + a_1\beta_2$ is $\widehat{\Theta} = a_1 b_1 + a_2 b_2$
- $\operatorname{var}\widehat{\Theta} = \sigma^2 v^2(a_1, a_2)$, where

$$v^2(a_1, a_2) = \frac{a_0^2 \frac{1}{n}\sum x_i^2 - 2\,\overline{x}\,a_1\,a_2 + a_2^2}{\sum (x_i - \overline{x})^2};$$

- especially

$$\operatorname{var}\widehat{y_j} = \operatorname{var}\,(b_1 + b_2 x_j) = \sigma^2 \frac{\sum x_i^2 - 2\left(\sum x_i\right) x_j + x_j^2}{n\sum (x_i - \overline{x})^2}$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum x_i^2 - n\,(\overline{x})^2}\right)$$

Limit Gauss–Markov Theorem

Given the assumptions of the classical linear regression model, and assumption $\limsup_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} |x_i|^3 < +\infty$,
than the least-squares estimators

$$\lim_{n\to+\infty} P\left(\frac{\widehat{\Theta}_n - \Theta}{\sigma \cdot v_n(a_1, a_2)} \leq x\right) = \Phi(x)$$

Classical Linear Normal Regression Model (CLRM),

- The classical normal linear regression model assumes that each ε_i is **distributed normally** with zero mean and constant variance and non-correlation,
- $\varepsilon \sim N_n(0, \sigma^2 \boldsymbol{I})$

Given the assumptions of the classical linear normal regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUES.

Distribution of estimated parameters

(i) Random variable $\widehat{\Theta} = a_1 b_1 + a_2 b_2$ is normal distributed with expected value
$\Theta = a_1\beta_1 + a_2\beta_2$ and variance $\mathrm{var}\,(\widehat{\Theta}) = \sigma^2 \cdot v^2(a_1, a_2)$;

(ii) Random variable $\dfrac{s^2(n-2)}{\sigma^2}$ is χ^2 distributed with $n-2$ degree of freedom;

(iii) Random variable $\widehat{\Theta}$ and s^2 are independent;

(iv) Random variable $\dfrac{\widehat{\Theta} - \Theta}{s \cdot v(a_1, a_2)}$ is t-distributed with $n-2$ degree of freedom.

Interval estimation and testing hypothesis

- Standard error of $\hat{\beta}_1 = b_1$ is $se(b_1) = \sqrt{\mathrm{var}\,(b_1)} = \left(\frac{s^2}{\sum(x_i-\overline{x})^2}\right)^2$
- Standard error of $\hat{\beta}_2 = b_2$ is $se(b_2) = \sqrt{\mathrm{var}\,(b_2)} = \left(s^2 \frac{\sum x_i^2}{n\sum(x_i-\overline{x})^2}\right)^2$
- From Gauss-Markov theorem we know, that $\frac{b_j-\beta_j}{se(b_j)} \sim t - distribution$ for $j = 1, 2$
- Confidence interval for b_j

$$b_j \pm t_{1-\alpha/2,(\nu=n-2)} \cdot se(b_j)$$

- Testing hypothesis, for example $H_0 : \beta_1 = 0$ and $H_1 : nonH_0$ the test criterion is

$$T = \frac{b_1}{se(b_1)} \sim t - distribution$$

Interval estimation

Parameter	Lower boundary		Upper boundary
Θ :	$\widehat{\Theta} - t_{1-\alpha/2,(\nu=n-2)}\, s\, v\left(a_0, a_1\right)$	$-$	$\widehat{\Theta} + t_{1-\alpha/2,(\nu=n-2)}\, s\, v\left(a_0, a_1\right)$
β_1 :	$b_0 - t_{1-\alpha/2,(\nu=n-2)}\, s \sqrt{\frac{\sum x_i^2}{n\sum(x_i-\overline{x})^2}}$	$-$	$b_0 + t_{1-\alpha/2,(\nu=n-2)}\, s \sqrt{\frac{\sum x_i^2}{n\sum(x_i-\overline{x})^2}}$
β_2 :	$b_1 - t_{1-\alpha/2,(\nu=n-2)}\, s \sqrt{\frac{1}{\sum(x_i-\overline{x})^2}}$	$-$	$b_1 + t_{1-\alpha/2,(\nu=n-2)}\, s \sqrt{\frac{1}{\sum(x_i-\overline{x})^2}}$
σ^2 :	$\frac{(n-2)s^2}{\chi^2_{1-\alpha/2,(\nu=n-2)}}$	$-$	$\frac{(n-2)s^2}{\chi^2_{\alpha/2,(\nu=n-2)}}$

Hypothesis tests for parameters of model

Hypothesis H_0	Hypotesis H_1	region of rejection
$\Theta = \Theta_0$	$\Theta \neq \Theta_0$	$\dfrac{\left\lvert \widehat{\Theta} - \Theta_0 \right\rvert}{s\, v(a_0, a_1)} \geq t_{1-\alpha/2,(\nu=n-2)}$
$\Theta = \Theta_0$	$\Theta > \Theta_0$	$\dfrac{\left(\widehat{\Theta} - \Theta_0\right)}{s\, v(a_0, a_1)} \geq t_{1-\alpha,(\nu=n-2)}$
$\Theta = \Theta_0$	$\Theta < \Theta_0$	$\dfrac{\left(\widehat{\Theta} - \Theta_0\right)}{s\, v(a_0, a_1)} \leq t_{\alpha,(\nu=n-2)}$
$\beta_2 = 0$	$\beta_2 \neq 0$	$\dfrac{\lvert b_1 \rvert \sqrt{\sum (x_i - \overline{x})^2}}{s} \geq t_{1-\alpha/2,(\nu=n-2)}$
$\beta_2 = 0$	$\beta_2 > 0$	$\dfrac{b_1 \sqrt{\sum (x_i - \overline{x})^2}}{s} \geq t_{1-\alpha,(\nu=n-2)}$

Confidence intervals for $\hat{y} = b_1 + b_2 x$

- For mean $(b_1 + b_2\zeta) \pm t_{1-\alpha/2,(\nu=n-2)} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(\zeta-\overline{x})^2}{\sum x_i^2 - n(\overline{x})^2}}$
- For population function
  $(b_1 + b_2\zeta) \pm \sqrt{2F_{1-\alpha/2,(\nu_1=2,\nu_2=n-2)}} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(\zeta-\overline{x})^2}{\sum x_i^2 - n\overline{x}^2}}$
- For individual prediction (new observation)
  $(b_1 + b_2\zeta_P) \pm t_{1-\alpha/2,(\nu=n-2)} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(\zeta_P-\overline{x})^2}{\sum x_i^2 - n(\overline{x})^2}}.$

Maximum Likelihood Estimation of Classical Linear Normal Regression Model

(i) ML estimators β_1, β_2 are OLS estimators

(ii) ML estimator σ^2 is $\dfrac{\sum_i (y_i - \widehat{y_i})^2}{n}$ and is asymptotic equivalent with $s^2 = \frac{\sum_i (y_i - \widehat{y_i})^2}{n-2}$

Proof: Joint partial distribution function for vector $y_1, y_2, \ldots, y_n$ is

$$f(y_1, y_2, \ldots, y_n, \beta_1, \beta_2, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\Big(- \frac{1}{2\sigma^2} \sum (y_i - \beta_1 - \beta_2 x_i)^2 \Big)$$

The likelihood function $L(\boldsymbol{y}|\beta_1, \beta_2, \sigma^2) = f(y_1, y_2, \ldots, y_n, \beta_1, \beta_2, \sigma^2)$
The maximum likelihood estimator of parameters β_1 is the value of β_1 that maximizes the likelihood function.