Introduction to econometrics

II. Non-technical introduction to econometrics

Content

1. Simple regression model
   - Ordinary least squares method
   - Basic statistical concepts
2. Multiple regression
3. Dummy variables

Introduction

- Topics: how to use gretl + non-technical introduction to regression.
- Reading for the next week: Koop (2008), chapters 1 and 2.

Content

1. Simple regression model
   - Ordinary least squares method
   - Basic statistical concepts
2. Multiple regression
3. Dummy variables

Regression

- Relationships among variables (linear, non-linear, two or more variables).
- Dependent variables and explanatory variables.
- $E(Y|X)$ as a function of x (Y dependent variable, X explanatory variables, x realizations of explanatory variables).
- Interesting examples: Gujarati, Porter (2009).

Example – costs of production

- Replicate example in Koop (2008)

Linear regression model

- Linear relationship between costs, Y, and output, X:

$$Y = \alpha + \beta X.$$

- Unknown parameters of the model: α ... intercept, β ... slope parameter (effect of variable X on Y).
- Error term, ϵ – measurement error, ommited explanatory variables, unobserved variables $\Rightarrow$ observations do not lie exactly on the line.

$$Y = \alpha + \beta X + \epsilon.$$

Content

1 Simple regression model
- Ordinary least squares method
- Basic statistical concepts

2 Multiple regression

3 Dummy variables

How to estimate parameters?

- Parameters estimates: $\widehat{\alpha}$, $\widehat{\beta}$.
- Best fitting line?

Error terms and residuals

- Observations:

$$Y_i = \alpha + \beta X_i + \epsilon_i.$$

- Error term:

$$\epsilon_i = Y_i - \alpha - \beta X_i.$$

- Residual:

$$\widehat{\epsilon}_i = Y_i - \widehat{\alpha} - \widehat{\beta} X_i.$$

- Fitted regression line:

$$\widehat{Y_i} = \widehat{\alpha} + \widehat{\beta} X_i.$$

- Fitted values:

$$\widehat{Y}_i.$$

Regression – best fitting line

- Sum of squared residuals (SSR).

$$
\begin{aligned}
SSR &= \sum_{i=1}^{N} \widehat{\epsilon}_i^{\,2} \\
&= \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i \right)^2 \\
&= \sum_{i=1}^{N} \left(Y_i - \widehat{Y}_i \right)^2 .
\end{aligned}
$$

- Ordinary least squares method – OLS.

Interpreting OLS estimates

- Intercept – sometimes economic interpretations.
- $\widehat{\alpha} = 2.19$ ... fixed costs of the industry.
- Slope parameter:

$$\frac{\mathrm{d}\widehat{Y}_i}{\mathrm{d}\widehat{X}_i} = \widehat{\beta}.$$

- $\widehat{\beta} = 4.79$ ... estimated marginal costs of the industry.

Interpreting OLS estimates – review

Tabulka: Interpreting parameters regarding functional relationship.

Model	Dependent	Explanatory	Interpretation of β
Level-Level	Y	X	$\Delta Y = \beta \Delta X$
Level-Log	Y	$\ln X$	$\Delta Y = (\beta/100)\% \Delta X$
Log-Level	$\ln Y$	X	$\% \Delta Y = (100\beta) \Delta X$
Log-Log	$\ln Y$	$\ln X$	$\% \Delta Y = \beta \% \Delta X$

Measuring the fit

- *Total sum of squares*:

$$TSS = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 .$$

- *Regression sum of squares*:

$$RSS = \sum_{i=1}^{N} \left(\widehat{Y}_i - \overline{Y}\right)^2 .$$

- Total variability Y:

$$TSS = RSS + SSR.$$

- Coefficient of determination, R^2 $(0 \leq R^2 \leq 1)$:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{SSR}{TSS}.$$

Content

1. Simple regression model
   - Ordinary least squares method
   - Basic statistical concepts
2. Multiple regression
3. Dummy variables

Confidence intervals

- Confidence interval for a parameter – a measure of uncertainty of the point estimate.

$$Pr(Int_D < \beta < Int_H) = 0.95$$

- Confidence level (e.g. 95 %).
- Usually 0.99 = 99 %, 0.95 = 95 %, 0.90 = 90 %.

Hypothesis testing

- „Does education increase an individual's earning potential?"
- „Will a certain advertising strategy increase sales?"
- „Will a new governemnt training scheme lower unemployment?"
- Mostly: „Does the explanatory variable have an effect on the dependent variable?", or „Is $\beta \neq 0$ in the regression of Y on X?".

Hypothesis testing involving a parameter

- Null and alternative hypothesis: $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$.
- Test statistics:

$$t = \frac{\widehat{\beta}}{s_b}.$$

- Level of significanse: usually 0.01, 0.05, 0.10 $\Rightarrow$ (1-confidence level) $=$ a probability needed to not to reject null hypothesis (using observations).
- Critical value of the test – based on significance level; define critical region $\rightarrow$ value that a test statistic must exceed in order for the the null hypothesis to be rejected.
- *p*-value: compare with significance level; the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.
- Hypothesis testing using confidence intervals.

Using computer software.

- $\widehat{\beta}$: point estimate.
- 95% confidence interval.
- Standard error of parameter estimate $(\widehat{\beta})$, s_b.
- t-statistics for $H_0 : \beta = 0$.
- p-value for $H_0 : \beta = 0$.
- Example – electric utility industry (see Koop and replicate example).

Hypothesis testing involving R^2.

- $H_0 : R^2 = 0$, $H_1 : R^2 \neq 0 \rightarrow X$ does not have any explanatory power for Y.
- F-statistics:

$$F = \frac{(N-2)R^2}{1-R^2}.$$

- Compare with critical value or use p-value.

Content

1 Simple regression model
- Ordinary least squares method
- Basic statistical concepts

2 Multiple regression

3 Dummy variables

Model and OLS estimates

- Model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

- Sum of squared residuals:

$$SSR = \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i} - \ldots - \widehat{\beta}_k X_{ki} \right)^2 .$$

- R^2 – effect of the all variables.
- F-test – test of whether the regression explains anything at all:

$$F = \frac{(N-k-1)R^2}{1-R^2}.$$

Interpreting OLS estimates

- Parameter – marginal effect of the explanatory variable on the dependent variable holding the other explanatory variables constant.
- Example – house prices (page 45).

Choosing explanatory variables

- Important consideration pulling in opposite directions:
  1. To include as many variables as possible (all variables that help explain the dependent variable).
  2. To include as few explanatory variables as possible (including irrelevant variables, statistically insignificant, can raduce the statistical significance of all the explanatory variables).
- Why not to exclude important explanatory variables? (omitted variables bias) – example (see Koop (2008)).

Practical guide

- Not possible to include all relevant variables.
- Start with the most variables $\rightarrow$ sequential elimination of insignificant variables.
- Final regression – statistically significant variables only $+$ intercept.
- Competing models $\rightarrow$ R^2.

Multicollinearity

- If some or all of the explanatory
- Some consequences – high R^2 $\times$ all parameters statistically insignificant (high std. errors).
- Perfect collinearity – estimation impossible (intuition based on parameters interpretation).
- Solution – exclude appropriate variables.
- Testing – correlation matrix.

Content

1. Simple regression model
   - Ordinary least squares method
   - Basic statistical concepts
2. Multiple regression
3. Dummy variables

Working with dummy variables

- Qualitative „1 or 0“ variables).
- Interpreted such as „ordinary“ variables.
- Regression with „variable“ intercepts or slopes (for each category).
- Dummy dependent variable = another kind of models (logit, probit)!
- Examples – see Koop (2008), pages 51–55.

Exercises

- Koop (2008) – exercises 1, 2 and 3 (chapter 2).