Econometrics of Policy Evaluation: Regression Review

Cristian Huse

Least Squares Problem

- Obtain the estimators for beta that minimize the sum of squared residuals

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Differentiating w.r.t. $\hat{\beta}_0, \hat{\beta}_1$ yields

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} \hat{u}_i = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = \sum_{i=1}^{n} \hat{u}_i x_i = 0$$

- Solving for $\hat{\beta}_0, \hat{\beta}_1$ yields

$$\hat{\beta}_1 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Assumptions

1. **[Linearity in parameters]** In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

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

   where β_o, β_1 are the population intercept and slope parameters, respectively
2. **[Random sampling]** We have a random sample of size n, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$, following the population model in (1)
3. **[Sample variation in the explanatory variable]** The sample outcomes on x, $\{x_i, i = 1, \ldots, n\}$, are not all the same value
4. **[Zero conditional mean]** The error u has an expected value of zero given any value of the explanatory variable, i.e. $E(u|x) = 0$
5. **[Homoskedasticity]** The error u has the same variance given any value of the explanatory variable, i.e. $Var(u|x) = \sigma^2$
6. **[Normality]** The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is distributed $u \sim N(0; \sigma^2)$

Gauss-Markov Theorem

- Under 1-4, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ are unbiased, i.e. on average they get it right
- Under 1-5, can obtain the OLS variance formulas,

$$Var(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$Var(\hat{\beta}_0) = \sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- GM Theorem: OLS estimator is BLUE (best linear unbiased estimator)
  - linear $\rightarrow$from A1
  - unbiased $\rightarrow$from A1-4
  - best = smallest variance

Distribution and Hypotheses Testing

- Under 1-6, the OLS estimator is Normally distributed around its population ("true") value β_0

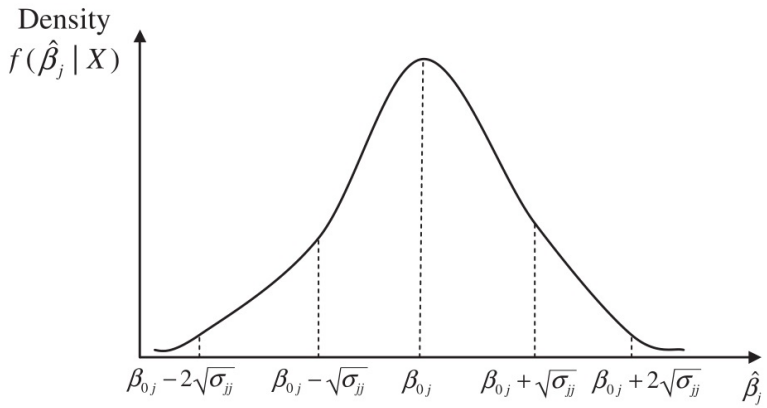

Figure 2.3. The distribution of an OLS estimator: $E[\hat{\beta}_j \mid X] = \beta_{0j}$ and $\text{Var}[\hat{\beta}_j \mid X] = \sigma_{jj}$.

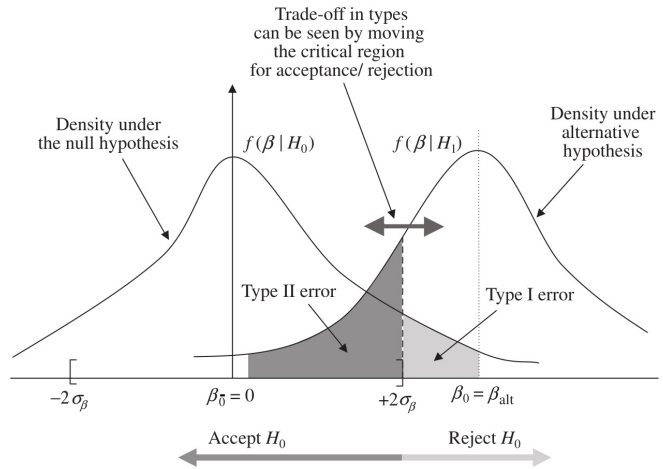

Figure 2.4. Hypothesis testing and the trade-off between type I and type II errors.

- This allows testing hypotheses about the parameters

Hypothesis Testing: The t-Test

- The t-test is used to test hypotheses about a single parameter
- Given a population model

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

and Assumptions 1-6, then

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

where $k+1$ is the number of unknown parameters in the population model and $n-k-1$ is the degrees of freedom (df)

- This allows us to test hypotheses of the form $H_0 : \beta_j = 0$
- Note: df of a model is $(\#observations - \#parameters)$

Hypothesis Testing: The F-Test

- The F-test is used to test multiple linear restrictions:
- Assume you estimated a model and want to test the joint null hypothesis

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$

- The original model is denoted *ur* and we will compare it to the restricted model *r* on which we estimate after imposing the above constraints
- The test statistic is given by

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

where $q = df_r - df_{ur}$ is the numerator degrees of freedom and $n - k - 1 = df_{ur}$ is the denominator degrees of freedom

Multiple Regression

- **Example:** Demand for a product relates to its price, but also to other variables, such as income, price of competing products etc
  - Thus, using only one regressor not enough ⇒ multiple regression

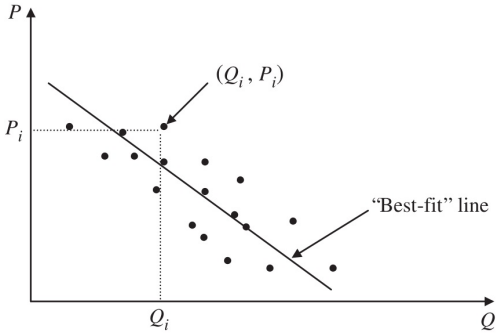

Figure 2.1. Scatter plot of the data and a "best-fit" line.

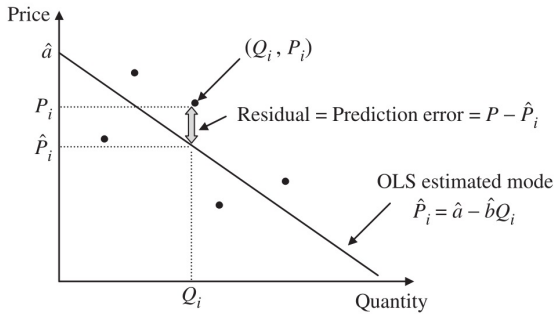

Figure 2.2. Estimated residuals in OLS regression.

- Multiple regression naturally generalizes simple regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k + e = X\beta + e$$

where matrix notation simplifies matters substantially
$[y(n \times 1), X(n \times k), \beta(k \times 1), e(n \times 1)]$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} \\ 1 & x_{12} & x_{22} & x_{32} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \beta + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Overview

- Classical model is based on a number of strong assumptions. Extensions include errors exhibiting
  - Heteroskedasticity
  - Autocorrelation
  - In either case, point estimates not affected, but standard errors need to be adjusted
- Classical model assumes continuous dependent variable
  - Discrete dependent variables can be estimated using logit and probit models (ordinary logit or probit if data is ranked)
- Classical model assumes exogeneity of regressors
  - Endogeneity (e.g. y, x simultaneously determined as in price and quantity in demand and supply equations)
  - Need to find an instrument z for endogenous x and perform IV estimation, e.g. 2SLS

Misspecification

- No model is correctly specified, but some are worse than others
- Misspecification occurs when a regression model cannot represent/approximate, for any parameter values, the true DGP
- Typical cases include
  - Incorrect functional form, e.g., linear when true model is nonlinear
  - Omission of relevant explanatory variable
- Problem
  - If the OV is important to explain the dependent variable and correlated with one of the explanatory variables included in the regression, the estimated parameters on the included regressors in the regression will be biased
    - (This leads to endogeneity, see below)
    - (Intuition: error term has random term + non-random term which wasn't included in regression)

Example

The quality of a product determines the demand for it. However, if not included in regressor, might generate bias since quality is typically correlated with price.

Exogeneity

- Exogeneity is a critical assumption in OLS specifications. It may appear in several related forms:

$$u_i|x_i \sim N(0, \sigma^2)$$

$$u_i \coprod x_i \text{ (independence)}$$

$$u_i \perp x_i \text{ or } E(u_i|x_i) = 0 \text{ (mean-independence)}$$

$$Cov(x_i, u_i)=0$$

- In words: "The explanatory variables contain no information about the error term."
  - If it were the case, there would be problems with the model
- In practice: Difficult to assume in many contexts. Why?

Exogeneity cont'd

- Why?
  - There may exist an (some? many?) unobserved variable that correlates with the explanatory variables (X) and with the error term (u)
  - Intuitively, instead of observing the "true", "clean" error term satisfying the classical assumptions, one observes a "dirty" error term which does not satisfy them (see example below)
- Implications
  - The estimated coefficient (β) does not have a causal interpretation because it contains (selection or omitted variables) bias(es)

Exogeneity cont'd

- Suppose true model is

$$y_i = \alpha + \beta x_i + \gamma w_i + u_i, E(u_i|x_i, w_i) = 0, E(w_i) = 0$$

  w not observed

- If w omitted, OLS will treat it as part of the error term:

$$y_i = \alpha + \beta x_i + v_i, v_i = \gamma w_i + u_i$$

- Problem if $Cov(x_i, w_i) \neq 0$ (x is endogenous):

$$\hat{\beta} = \beta + \gamma \left[Cov(x_i, w_i) / Var(x_i) \right]$$

- That is, sign of bias depends on product $\gamma Cov(x_i, w_i)$

Exogeneity Example

- Bennedsen et al (2006, QJE). Inside the family firm: The role of families in succession decisions and performance
- Question: What is the causal impact of families on corporate decisions and performance? (Are familly-managed firms better?)
- Motivation:
  - The majority of firms around the world are controlled by their founders or their founders' descendents (family-CEOs)
  - It is important to understand whether and how this organization structure is valuable
- Empirical challenge: CEO-family status is endogenous (e.g. to performance)!

Exogeneity Example cont'd

- Empirical strategy: Use the gender of the firstborn child of the exiting CEO as an instrument for family CEO status (implicit assumption quite realistic for a long time: male heirs are more likely to become CEO's, and this is not related to company performance)
- Results: Family successions have a large negative causal impact on performance

Exogeneity: Instrument

- Intuitively, want to "break" the correlation between x and u
- Idea: Search for a variable that is
  - Correlated with x
  - Not correlated with u
  - Not an explanatory variable of the original equation
  - a.k.a. INSTRUMENT
- Intuitively:
  - Instruments helps extract part of x that is not related to u
  - Amounts to artificially create exogenous shocks on x, as one would do in a random experiment (more on this later)

Exogeneity: Estimation

- Traditional estimation method is 2SLS (Two-stage least squares)
- Model is

$$y_i = \alpha + \beta x_i + \gamma v_i + u_i, Cov(u_i, x_i) \neq 0$$

- Step 1: Regress endogenous variable on instrument (z) and exogenous variables

$$x_i = \alpha_0 + \alpha_1 z_i + \alpha_2 v_i + e_i$$

and get its predicted value $\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i + \hat{\alpha}_2 v_i$ which captures the part of the variability of x captured by v and z

- Comments:
  - Ideally the relation between x and z should be strong; the stronger it is (t-stat of α_1, F-stat of first-stage) the better the instrument
  - i.e. a large change in z will generate a large change in x that is not related to the error term u
  - This exogenous variation in x can be used to identify a causal effect of x on y (more on this later)

Exogeneity: Estimation cont'd

- Step 2: Estimate $y_i = \alpha + \beta \hat{x}_i + \gamma v_i + u_i$
- Comments:
  - The resulting IV estimator is consistent and asymptotically Normal
  - The variance from OLS in Step 2 needs a correction because of use of $\hat{x}_i$
    - i.e. the fact that $\hat{x}_i$ is not data, but an estimated quantity (most software does it for you)
  - IV estimator has a larger s.e. than OLS (since uses only part of variation in x)

Methodology Corner

- Whenever you need to check whether an estimator is consistent, proceed as above:
  - Figure out what the "real error" looks like
  - Figure out whether the components of the real correlate with any of the regressors
  - If the correlation is non-zero, the estimator is inconsistent – how can you fix it?

Correlated Errors and Heteroskedasticity

- OLS standard errors valid under in particular cases: uncorrelated, homoskedastic errors
- Problems:
  - Autocorrelated errors: correlation over time
  - Clustered errors: students within a class, models produced by same carmaker
  - Heteroskedastic errors: variance not uniform, e.g., higher variance for extreme observations
- Consequence of wrong standard errors is on inference, e.g., hypotheses tests, not point estimates

Take-aways

- OLS as a starting point for empirical analysis
- Real data typically presents a number of additional challenges, thus departures from OLS
- Nevertheless, you should review OLS right now and become familiar with R using it
- Depending on your thesis topic, you will invest time in one specific empirical method, yet OLS is typically still the starting point

Next steps

- Refresh your Econometrics and check the R Tutorial

References

- Wooldridge (2013). Introductory Econometrics: A Modern Approach, 5th Edition (any edition should do)
  - Cross-sectional data: Chapters 2-9
  - Time series data: Chapters 10-12
  - Advanced topics: Chapters 13-19
- Introduction to Econometrics with R
  - https://bookdown.org/machar1991/ITER/or https://www.econometrics-with-r.org/ITER.pdf
- Florian Heiß's
  - Using R for Introductory Econometrics http://urfie.net/
  - YouTube channel: https://www.youtube.com/channel/UC1GroDuLOkwxsntDNwHvChg