Econometrics Cheat Sheet

By Marcelo Moreno - King Juan Carlos University

Basic concepts

Definitions

Econometrics - is a social science discipline with the objective of quantify the relationships between economic agents, contrast economic theories and evaluate and implement government and business policies.

Econometric model - is a simplified representation of the reality to explain economic phenomena.

Ceteris paribus - if all the other relevant factors remain constant.

Data types

Cross section - data taken at a given moment in time, an static *photo*. Order does not matter.

Time series - observation of one/many variable/s across time. Order does matter.

Panel data - consist of a time series for each observation of a cross section.

Pooled cross sections - combines cross sections from different time periods.

Phases of an econometric model

1. Specification.
2. Estimation.
3. Validation.
4. Utilization.

Regression analysis

Study and predict the mean value of a variable (dependent variable, y) regarding the base of fixed values of other variables (independent variables, x's). In econometrics it is common to use Ordinary Least Squares (OLS) for regression analysis.

Correlation analysis

The correlation analysis not distinguish between dependent and independent variables.

- The simple correlation measures the grade of linear association between two variables.
$$r = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}}$$
- The partial correlation measures the grade of linear association between two variables controlling a third.

Assumptions and properties

Econometric model assumptions

Under this assumptions, the estimators of the OLS parameters will present good properties. **Gauss-Markov assumptions extended**:

1. **Parameters linearity** (plus weak dependence in time series). y must be a linear function of the β's.
2. **Random sampling**. The sample from the population has been randomly taken. (Only when cross section)
3. **No perfect collinearity**.
   - There are no independent variables that are constant: $\text{Var}(x_j) \neq 0$
   - There is not an exact linear relation between independent variables.
4. **Conditional mean zero and correlation zero**.
   a. There are no systematic errors: $\text{E}(u|x_1, ..., x_k) = \text{E}(u) = 0 \rightarrow$ **strong exogeneity** (a implies b).
   b. There are no relevant variables left out of the model: $\text{Cov}(x_j, u) = 0$ for any $j = 1, ..., k \rightarrow$ **weak exogeneity**.
5. **Homoscedasticity**. The variability of the residuals is the same for all levels of x: $\text{Var}(u|x_1, ..., x_k) = \sigma^2$
6. **No auto-correlation**. The residuals do not contain information about other residuals: $\text{Corr}(u_t, u_s|x) = 0$ for any given $t \neq s$. (Typical of time series)
7. **Normality**. The residuals are independent and identically distributed: $u \sim N(0, \sigma^2)$
8. **Data size**. The number of observations available must be greater than $(k+1)$ parameters to estimate. (It is already satisfied under asymptotic situations)

Asymptotic properties of OLS

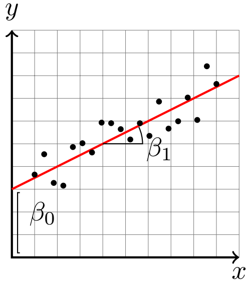
Under the econometric model assumptions and the Central Limit Theorem:

- Hold (1) to (4a): OLS is **unbiased**. $\text{E}(\hat{\beta}_j) = \beta_j$
- Hold (1) to (4): OLS is **consistent**. $\text{p.lim}(\hat{\beta}_j) = \beta_j$ (to (4b) left out (4a), weak exogeneity, biased but consistent)
- Hold (1) to (5): **asymptotic normality** of OLS (then, (7) is necessarily satisfied): $u \sim_a N(0, \sigma^2)$.
- Hold (1) to (6): **unbiased estimate of** σ^2. $\text{E}(\hat{\sigma}^2) = \sigma^2$
- Hold (1) to (6): OLS is BLUE (Best Linear Unbiased Estimator) or **efficient**.
- Hold (1) to (7): hypothesis testing and confidence intervals can be done reliably.

Ordinary Least Squares

Objective - minimize the Sum of Squared Residuals (SSR):
$$\text{Min} \sum_{i=1}^n \hat{u}_i^2, \text{ where } \hat{u}_i = y_i - \hat{y}_i$$

Simple regression model

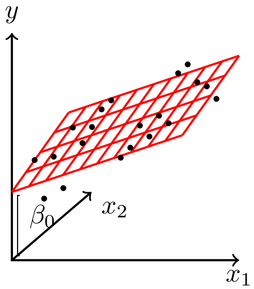


Equation:
$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

Estimation:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

Where:
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$
$$\hat{\beta}_1 = \frac{\text{Cov}(y,x)}{\text{Var}(x)}$$

Multiple regression model



Equation:
$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + u_i$$

Estimation:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki}$$

Where:
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}_1 - ... - \hat{\beta}_k \overline{x}_k$$
$$\hat{\beta}_j = \frac{\text{Cov}(y, \text{residualized } x_j)}{\text{Var}(\text{residualized } x_j)}$$

Matrix: $\hat{\beta} = (X^T X)^{-1}(X^T y)$

Interpretation of coefficients

Model	Dependent	Independent	β_1 interpretation
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y \approx (\beta_1/100)(\%\Delta x)$
Log-level	$\log(y)$	x	$\%\Delta y \approx (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y \approx \beta_1(\%\Delta x)$
Quadratic	y	$x + x^2$	$\Delta y = (\beta_1 + 2\beta_2 x)\Delta x$

Error measures

Sum of Sq. Residuals: $\text{SSR} = \sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Explained Sum of Squares: $\text{SSE} = \sum_{i=1}^n (\hat{y}_i - \overline{y})^2$

Total Sum of Sq.: $\text{SST} = \text{SSE} + \text{SSR} = \sum_{i=1}^n (y_i - \overline{y})^2$

Standard Error of the Regression: $\hat{\sigma} = \sqrt{\frac{\text{SSR}}{n-k-1}}$

Standard Error of the $\hat{\beta}$'s: $\text{se}(\hat{\beta}) = \hat{\sigma} \cdot \sqrt{(X^T X)^{-1}}$

Mean Squared Error: $\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Absolute Mean Error: $\text{AME} = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$

Mean Percentage Error: $\text{MPE} = \frac{\sum_{i=1}^n |\hat{u}_i / y_i|}{n} \cdot 100$

R-squared

Is a **measure of the goodness of the fit**, how the regression fits to the data:

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

- Measures the **percentage of variation of y that is linearly explained by the variations of x's**.
- Takes values **between 0** (no linear explanation of the variations of y) **and 1** (total explanation of the variations of y).

When the number of regressors increment, the value of the R-squared increments as well, whatever the new variables are relevant or not. To solve this problem, there is an **R-squared corrected by degrees of freedom**:

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\text{SSR}}{\text{SST}} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

For big sample sizes: $\overline{R}^2 \approx R^2$

Hypothesis testing

The basics of hypothesis testing

An hypothesis test is a rule designed to explain from a sample, if **exist evidence or not to reject an hypothesis** that is made about one or more population parameters.

Elements of an hypothesis contrast:

- **Null hypothesis (H_0)** - is the hypothesis to be tested.
- **Alternative hypothesis (H_1)** - is the hypothesis that cannot be rejected when the null hypothesis is rejected.
- **Statistic of contrast**: is a random variable whose probability distribution is known under the null hypothesis and is tabulated.
- **Significance level (α)** - is the probability of rejecting the null hypothesis being true (Type I error). Is chosen by who conduct the contrast. Commonly is 0.10, 0.05, 0.01 or 0.001.
- **Critic value** - is the value against which the statistic of contrast is compared to determine if the null hypothesis is rejected or not.
- **p-value** - is the highest level of significance by which the null hypothesis cannot be rejected (H_0).

The rule is: if the p-value is less than α, there is evidence that, at a given α, the null hypothesis is rejected (the alternative is accepted instead).

Individual contrasts

Tests if a parameter is significantly different from a given value, ϑ.

- $H_0 : \beta_j = \vartheta$
- $H_1 : \beta_j \neq \vartheta$

$$\text{Under } H_0: \quad t = \frac{\hat{\beta}_j - \vartheta}{\text{se}(\hat{\beta}_j)} \sim t_{n-k-1,\alpha/2}$$

If $\mid t \mid> t_{n-k-1,\alpha/2}$, there is evidence to reject H_0.

Individual significance test - tests if a parameter is significantly **different from zero**.

- $H_0 : \beta_j = 0$
- $H_1 : \beta_j \neq 0$

$$\text{Under } H_0: \quad t = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-k-1,\alpha/2}$$

If $\mid t \mid> t_{n-k-1,\alpha/2}$, there is evidence to reject H_0.

The F contrast

Simultaneously contrasts multiple (linear) hypothesis about the parameters. It makes use of a non restricted model and a restricted model:

- **Non restricted model** - is the model on which we want to test the hypothesis.
- **Restricted model** - is the model on which the hypothesis that we want to contrast have been imposed.

Then, looking at the errors, there are:

- $\sum_{i=1}^n \hat{u}_{\text{nr}}^2$ - is the Sum of Sq. Resid. of the non restricted model (SSR_{nr}).
- $\sum_{i=1}^n \hat{u}_{\text{r}}^2$ - is the Sum of Sq. Resid of the restricted model (SSR_{r}).

Under H_0:

$$F = \frac{\text{SSR}_{\text{r}} - \text{SSR}_{\text{nr}}}{\text{SSR}_{\text{nr}}} \frac{(n - k_{\text{nr}} - 1)}{q} \sim F_{q,n-k_{\text{nr}}-1}$$

Where k_{nr} is the number of parameters of the non restricted model and q is the number of linear hypothesis tested.

If $F_{q,n-k_{\text{nr}}-1} < F$, there is evidence to reject H_0.

Global significance test - tests if all the parameters associated to x's are **simultaneously equal to zero**.

- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
- $H_1 : \beta_1 \neq 0$ and/or $\beta_2 \neq 0$... and/or $\beta_k \neq 0$

In this case, we can simplify the formula for the F stat.

Under H_0:

$$F = \frac{R^2}{1-R^2}\frac{(n-k-1)}{k} \sim F_{k,n-k-1}$$

If $F_{k,n-k-1} < F$, there is evidence to reject H_0.

Confidence intervals

The confidence intervals at $(1 - \alpha)$ confidence level can be calculated:

$$\hat{\beta}_j \mp t_{n-k-1,\alpha/2} \cdot \text{se}(\hat{\beta}_j)$$

Dummy variables and structural change

Dummy (or binary) variables are used for qualitative information like sex, civil state, country, etc.

- Get the **value of 1 in a given category, and 0 on the rest**.
- Are used to analyze and modeling **structural changes** in the model parameters.

If a qualitative variable have m categories, we only have to include $(m - 1)$ dummy variables.

Structural change

Structural change refers to changes in the values of the parameters of the econometric model produced by the effect of different sub-populations. Structural change can be included in the model through dummy variables.

The location of the dummy variable matters:

- **On the intercept (β_0)** - represents the mean difference between the values produced by the structural change.
- **On the parameters that determines the slope of the regression line (β_j)** - represents the effect (slope) difference between the values produced by the structural change.

The Chow's structural contrast - when we want to analyze the existence of structural changes in all the model parameters, it is common to use a particular expression of the F contrast known as the Chow's contrast, where the null hypothesis is: H_0 : No structural change.

Predictions

Two types of prediction:

- Of the mean value of y for a specific value of x.
- Of an individual value of y for a specific value of x.

If the values of the variables (x) approximate to the mean values $(\overline{x})$, the confidence interval amplitude of the prediction will be shorter.

Multicollinearity

- **Perfect multicollinearity** - there are independent variables that are constant and/or there is an exact linear relation between independent variables. Is the **breaking of the third (3) econometric model assumption**.
- **Approximate multicollinearity** - there are independent variables that are approximately constant and/or there is an approximately linear relation between independent variables. It **does not break any econometric model assumption**, but has an effect on OLS.

Consequences

- **Perfect multicollinearity** - the equation system of OLS cannot be solved due to infinite solutions.
- **Approximate multicollinearity**
  - Small sample variations can induce to big variations in the OLS estimations.
  - The variance of the OLS estimators of the x's that are collinear, increments, thus the inference of the parameter is affected. The estimation of the parameter is very imprecise (big confidence interval).

Detection

- **Correlation analysis** - look for high correlations (greater than 0.7) between independent variables.
- **Variance Inflation Factor (VIF)** - indicates the increment of $\mathrm{Var}(\hat{\beta}_j)$ because of the multicollinearity.

  $$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1-R_j^2}$$

  Where R_j^2 denotes the R-squared from a regression between x_j and all the other x's.
  - Values between 4 to 10 suggest that it is advisable to analyze in more depth if there might be multicollinearity problems.
  - Values bigger than 10 indicates that there are multicollinearity problems.

One typical characteristic of multicollinearity is that the regression coefficients of the model are not individually different from zero (due to high variances), but jointly they are different from zero.

Correction

- Delete one of the collinear variables.
- Perform factorial analysis (or any other dimension reduction technique) on the collinear variables.
- Interpret coefficients with multicollinearity jointly.

Heteroscedasticity

The residuals u_i of the population regression function do not have the same variance σ^2:

$$\mathrm{Var}(u|x) = \mathrm{Var}(y|x) \neq \sigma^2$$

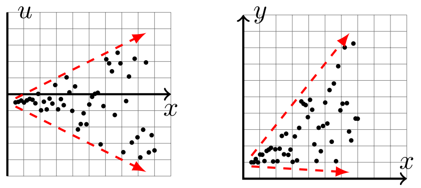
Is the **breaking of the fifth (5) econometric model assumption**.

Consequences

- OLS estimators still are unbiased.
- OLS estimators still are consistent.
- OLS is **not efficient** anymore, but still a LUE (Linear Unbiased Estimator).
- **Variance estimations of the estimators are biased**: the construction of confidence intervals and the hypothesis contrast are not reliable.

Detection

- **Graphs** - look for scatter patterns on x vs. u or x vs. y plots.



- **Formal tests** - White, Bartlett, Breusch-Pagan, etc. Commonly, the null hypothesis: H_0: Homoscedasticity.

Correction

- Use OLS with a variance-covariance matrix estimator robust to heteroscedasticity, for example, the one proposed by White.
- If the variance structure is known, make use of Weighted Least Squares (WLS) or Generalized Least Squares (GLS).
- If the variance structure is not known, make use of Feasible Weighted Least Squared (FWLS), that estimates a possible variance, divides the model variables by it and then apply OLS.
- Make assumptions about the possible variance:
  - Supposing that σ_i^2 is proportional to x_i, divide the model variables by the square root of x_i and apply OLS.
  - Supposing that σ_i^2 is proportional to x_i^2, divide the model variables by x_i and apply OLS.
- Make a new model specification, for example, logarithmic transformation.

Auto-correlation

The residual of any observation, u_t, is correlated with the residual of any other observation. The observations are not independent.

$$\mathrm{Corr}(u_t, u_s|x) \neq 0 \text{ for any } t \neq s$$

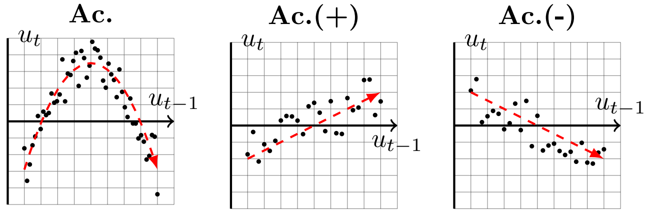
The "natural" context of this phenomena is time series. Is the **breaking of the sixth (6) econometric model assumption**.

Consequences

- OLS estimators still are unbiased.
- OLS estimators still are consistent.
- OLS is **not efficient** anymore, but still a LUE (Linear Unbiased Estimator).
- **Variance estimations of the estimators are biased**: the construction of confidence intervals and the hypothesis contrast are not reliable.

Detection

- **Graphs** - look for scatter patterns on u_{t-1} vs. u_t or make use of a correlogram.



- **Formal tests** - Durbin-Watson, Breusch-Godfrey, etc. Commonly, the null hypothesis: H_0: No auto-correlation.

Correction

- Use OLS with a variance-covariance matrix estimator robust to auto-correlation, for example, the one proposed by Newey-West.
- Use Generalized Least Squares. Supposing $y_t = \beta_0 + \beta_1 x_t + u_t$, with $u_t = \rho u_{t-1} + \varepsilon_t$, where $|\rho| < 1$ and ε_t is white noise.
  - If ρ is known, create a quasi-differentiated model where u_t is white noise and estimate it by OLS.
  - If ρ is not known, estimate it by -for example- the Cochrane-Orcutt method, create a quasi-differentiated model where u_t is white noise and estimate it by OLS.