Econometrics Cheat Sheet

By Marcelo Moreno - Universidad Rey Juan Carlos
The Econometrics Cheat Sheet Project

Basic concepts

Definitions

Econometrics - is a social science discipline with the objective of quantify the relationships between economic agents, test 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 doesn't matter.

Time series - observation of variables across time. Order does matter.

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

Pooled cross sections - combines cross section 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

Correlation analysis don't distinguish between dependent and independent variables.

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

Assumptions and properties

Econometric model assumptions

Under this assumptions, the OLS estimator will present good properties. **Gauss-Markov** assumptions:

1. **Parameters linearity** (and 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: $\text{Var}(x_j) \neq 0, \ \forall j = 1, \ldots, k$
   - There isn't an exact linear relation between independent variables.
4. **Conditional mean zero and correlation zero**.
   a. There aren't systematic errors: $\text{E}(u \mid x_1, \ldots, x_k) = \text{E}(u) = 0 \to$ **strong exogeneity** (a implies b).
   b. There are no relevant variables left out of the model: $\text{Cov}(x_j, u) = 0, \ \forall j = 1, \ldots, k \to$ **weak exogeneity**.
5. **Homoscedasticity**. The variability of the residuals is the same for all levels of x: $\text{Var}(u \mid x_1, \ldots, x_k) = \sigma_u^2$
6. **No auto-correlation**. Residuals don't contain information about any other residuals: $\text{Corr}(u_t, u_s \mid x_1, \ldots, x_k) = 0, \ \forall t \neq s$
7. **Normality**. Residuals are independent and identically distributed: $u \sim \mathcal{N}(0, \sigma_u^2)$
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

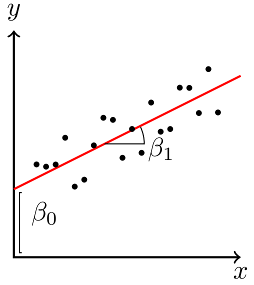
Under the econometric model assumptions and the Central Limit Theorem (CLT):

- Hold 1 to 4a: OLS is **unbiased**. $\text{E}(\hat{\beta}_j) = \beta_j$
- Hold 1 to 4: OLS is **consistent**. $\text{plim}(\hat{\beta}_j) = \beta_j$ (to 4b left out 4a, weak exogeneity, biased but consistent)
- Hold 1 to 5: **asymptotic normality** of OLS (then, 7 is necessarily satisfied): $u \underset{a}{\sim} \mathcal{N}(0, \sigma_u^2)$
- Hold 1 to 6: **unbiased estimate** of σ_u^2. $\text{E}(\hat{\sigma}_u^2) = \sigma_u^2$
- 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

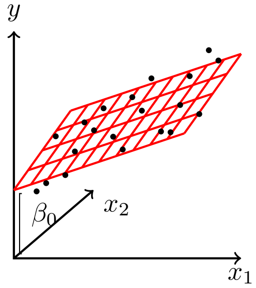
Objective - minimize the Sum of Squared Residuals (SSR):
$$\min \sum_{i=1}^n \hat{u}_i^2, \text{ where } \hat{u}_i = y_i - \hat{y}_i$$

Simple regression model



Equation:
$$y_i = \beta_0 + \beta_1 x_i + u_i$$
Estimation:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
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:
$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$$
Estimation:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$
where:
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}_1 - \cdots - \hat{\beta}_k \overline{x}_k$$
$$\hat{\beta}_j = \frac{\text{Cov}(y, \text{resid } x_j)}{\text{Var}(\text{resid } x_j)}$$
Matrix: $\hat{\beta} = (X^\intercal X)^{-1}(X^\intercal y)$

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 measurements

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: $\hat{\sigma}_u = \sqrt{\frac{\text{SSR}}{n-k-1}}$

Standard Error of the $\hat{\beta}$'s: $\text{se}(\hat{\beta}) = \sqrt{\hat{\sigma}_u^2 \cdot (X^\intercal X)^{-1}}$

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

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 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) **and 1** (total explanation).

When the number of regressors increases, the value of the R-squared also increases, whatever the new variables are relevant or not. To solve this problem, there is an **adjusted R-squared** by degrees of freedom (or corrected):

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

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

Hypothesis testing

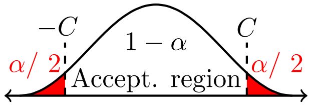
Definitions

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.

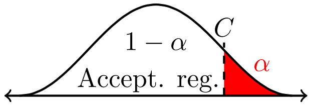
Elements of an hypothesis test:

- **Null hypothesis** (H_0) - is the hypothesis to be tested.
- **Alternative hypothesis** (H_1) - is the hypothesis that cannot be rejected when H_0 is rejected.
- **Test statistic** - is a random variable whose probability distribution is known under H_0.
- **Critical value** (C) - is the value against which the test statistic is compared to determine if H_0 is rejected or not. It sets the frontier between the regions of acceptance and rejection of H_0.
- **Significance level** (α) - is the probability of rejecting the null hypothesis being true (Type I Error). Is chosen by who conduct the test. Commonly is 10%, 5% or 1%.
- **p-value** - is the highest level of significance by which H_0 cannot be rejected.

Two-tailed test. H_0 dist.



One-tailed test. H_0 dist.



The rule is: if p-value $< \alpha$ holds, there is evidence to reject H_0, thus, there is evidence to accept H_1.

Individual tests

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

The F test

Simultaneously tests 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 test have been imposed.

Then, looking at the errors, there are:

- SSR_{UR} - is the SSR of the non restricted model.
- SSR_{R} - is the SSR of the restricted model.

$$\text{Under } H_0: \quad F = \frac{\text{SSR}_{\text{R}} - \text{SSR}_{\text{UR}}}{\text{SSR}_{\text{UR}}} \cdot \frac{n-k-1}{q} \sim F_{q,n-k-1}$$

where k is the number of parameters of the non restricted model and q is the number of linear hypothesis tested.

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

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

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

In this case, we can simplify the formula for the F statistic:

$$\text{Under } H_0: \quad F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} \sim F_{k,n-k-1}$$

If $F > F_{k,n-k-1}$, 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

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

- Takes the **value 1** in a given category and **0 in 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 variables (D) matters:

- **On the intercept** (additive effect) - represents the mean difference between the values produced by the structural change.
$$y = \beta_0 + \delta_1 D + \beta_1 x_1 + u$$
- **On the slope** (multiplicative effect) - represents the effect (slope) difference between the values produced by the structural change.
$$y = \beta_0 + \beta_1 x_1 + \delta_1 D \cdot x_1 + u$$

Chow's structural test - analyze the existence of structural changes in all the model parameters, it's a particular expression of the F test, where H_0: No structural change (all $\delta = 0$).

Changes of scale

Changes in the **measurement units** of the variables:

- In the **endogenous** variable, $y^* = y \cdot \lambda$ - affects all model parameters, $\beta_j^* = \beta_j \cdot \lambda$, $\forall j = 1, \ldots, k$
- In an **exogenous** variable, $x_j^* = x_j \cdot \lambda$ - only affect the parameter linked to said exogenous variable, $\beta_j^* = \beta_j \cdot \lambda$
- Same scale change on endogenous and exogenous - only affects the intercept, $\beta_0^* = \beta_0 \cdot \lambda$

Changes of origin

Changes in the **measurement origin** of the variables (endogenous or exogenous), $y^* = y + \lambda$ - only affects the model's intercept, $\beta_0^* = \beta_0 + \lambda$

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 between independent variables, $|r| > 0.7$.
- **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 - there might be multicollinearity problems.
  - Values > 10 - there are multicollinearity problems.

One typical characteristic of multicollinearity is that the regression coefficients of the model aren't 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

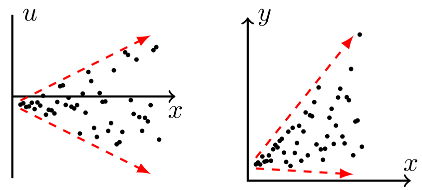
The residuals u_i of the population regression function do not have the same variance σ_u^2:
$$\mathrm{Var}(u \mid x_1, \ldots, x_k) = \mathrm{Var}(u) \neq \sigma_u^2$$
Is the **breaking of the fifth (5) econometric** model **assumption**.

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 testing is 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, H_0: No heteroscedasticity.

Correction

- Use OLS with a variance-covariance matrix estimator robust to heteroscedasticity (HC), 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):
  - Supposing that $\mathrm{Var}(u) = \sigma_u^2 \cdot x_i$, divide the model variables by the square root of x_i and apply OLS.
  - Supposing that $\mathrm{Var}(u) = \sigma_u^2 \cdot x_i^2$, divide the model variables by x_i (the square root of x_i^2) and apply OLS.
- 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 a new model specification, for example, logarithmic transformation (lower variance).

Auto-correlation

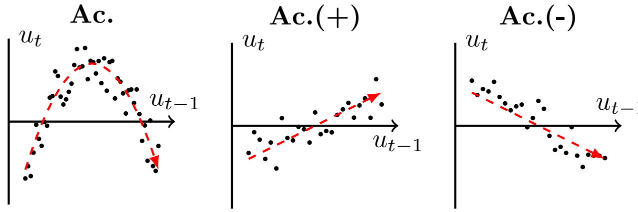
The residual of any observation, u_t, is correlated with the residual of any other observation. The observations are not independent.
$$\mathrm{Corr}(u_t, u_s \mid x_1, \ldots, x_k) = \mathrm{Corr}(u_t, u_s) \neq 0, \quad \forall t \neq s$$
The "natural" context of this phenomena is time series. Is the **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 testing is 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, H_0: No auto-correlation.

Correction

- Use OLS with a variance-covariance matrix estimator robust to heterocedasticity and auto-correlation (HAC), 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.