Econometrics Cheat Sheet

By Marcelo Moreno - King Juan Carlos University

Version 1.2-en

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.

Data types

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

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

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

Pooled cross sections - combines cross sections from different temporal 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.
- The partial correlation measures the grade of linear association between two variables controlling a third variable.

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**. y must be a linear function of the β's
2. **Random sampling**. The sample from the population has been randomly taken. (ONLY makes sense when data is cross section)
3. **No perfect collinearity**.
   - There are no independent variables that are constant: $Var(X) \neq 0$.
   - There is not an exact linear relation between independent variables.
4. **Conditional mean zero and correlation zero**.
   - There are no systematic errors: $E(u|x_1, ..., x_k) = E(u) = 0$.
   - There are no relevant variables left out the model: $Cov(x_j|u) = 0$ for any $j = 1, ..., k$.
5. **Homoscedasticity**. The variability of the residual is the same for all levels of x: $Var(u|x_1, ..., x_k) = \sigma^2$.
6. **No auto-correlation**. The residuals do not contain information about other residuals: $Corr(u_t, u_s|X) = 0$ for any $t \neq s$. (ONLY makes sense when data is temporal 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. (Do NOT makes sense under asymptotic situations)

Asymptotic properties of OLS

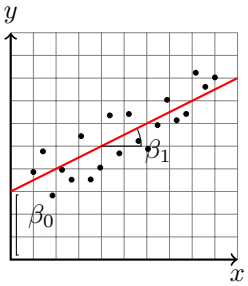
Under the econometric model assumptions and the Central Limit Theorem:

- Hold 1 to 4: OLS is **unbiased**. $E(\hat{\beta}_j) = \beta_j$
- Hold 1 to 4: OLS is **consistent**. $plim(\hat{\beta}_j) = \beta_j$
- Hold 1 to 5: **asymptotic normality** of OLS (then, 7 is necessarily satisfied): $u \sim_a N(0, \sigma^2)$.
- 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): $Min \sum_{i=1}^{n} \hat{u}_i^2$, where $\hat{u}_i = y_i - \hat{y}_i$.

Simple regression model



Equation:

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

Estimation:

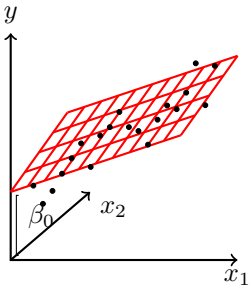
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

Where:

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

$$\hat{\beta}_1 = \frac{Cov(y,x)}{Var(x)}$$

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:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}_1 - ... - \hat{\beta}_k \overline{x}_k$$

$$\hat{\beta}_j = \frac{Cov(y, resid(x_j))}{Var(resid(x_j))}$$

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

Interpretation of coefficients

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

Error measures

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

Expl. Sum of Sq.: $SSE = \sum_{i=1}^{n} (\hat{y}_i - \overline{y}_i)^2$

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

Standard error (se) of the regression: $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1}}$

Sqrt. of the Quadratic Mean Error: $\sqrt{\frac{\sum_{i=1}^{n} (\hat{u}_i - \overline{u})^2}{n}}$

Absolute Mean Error: $\frac{\sum_{i=1}^{n} |\hat{u}_i|}{n}$

R-squared

Is a measure of the goodness of the fit (how the OLS fits to the data):

$R^2 = \frac{\sum_i^n (\hat{y}_i - \overline{y}_i)^2}{\sum_i^n (y_i - \overline{y}_i)^2} = 1 - \frac{\sum_i^n \hat{u}_i^2}{nS_y^2}$

- 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 eliminate the last point, there is an r-squared corrected by degrees of freedom:

$\overline{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \overline{y}_i)^2} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$

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 of one or more population parameters.

Elements of an hypothesis contrast:

- Null hypothesis (H_0): is the hypothesis that you want to contrast.
- Alternative hypothesis (H_1): is the hypothesis that cannot be rejected when the null hypothesis is rejected.
- Statistic of contrast: is a random variable with a known distribution that allow us to see if we reject (or not) the null hypothesis.
- 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 that, for a determined value of α, determines the reject (or not) of the null hypothesis.
- p-value: is the highest level of significance for what we do not reject (accept) the null hypothesis (H_0).

The rule is: if p-value is lower than α, there is evidence at that given α to reject the null hypothesis (accept the alternative instead).

Individual contrasts

Under the premise of normality of the residuals, contrast if a given parameter is significantly different from a given value.

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

Under H_0:

$t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1,\alpha/2}$

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

Individual significance contrasts - contrast if a given parameter is significantly different from zero.

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

Under H_0:

$t = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} \sim t_{n-k-1,\alpha/2}$

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

Confidence intervals

Under the normality of the residuals requirement, the confidence intervals at $(1-\alpha)$ confidence can be calculated:

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

The F contrast

It uses a non restricted model and a restricted model to do assumptions about the parameters.

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

Then, looking at the errors, there are:

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

Then: $F = \frac{SSR_r - SSR_{nr}}{SSR_{nr}} \frac{(n-K-1)}{q} \sim F_{q,n-K-1}$

Where K is the number of parameters of the non restricted model and q is the number of linear hypothesis.

When $F_{q,n-K-1} < F$, there is evidence to reject the null hypothesis.

Dummy variables and structural change

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

- The dummy variables get the value of 1 in a given category, and 0 on the rest.
- Dummy variables 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

We denominate structural changes to the modifications in the value of the parameters of the models for different subpopulations.

The position of the dummy variable matters:

- On the constant, their associate parameter represents the difference in mean between the values.
- On the parameters that determines the slope of the regression line, the associate parameter represents the difference in the effect between the values.

The Chow's structural contrast

When we want to analyze the existence of structural changes in all the model parameters, is more common to use a particular expression of the F contrast known as the Chow's contrast.

It defines two non restricted models (with structural change):

$y_i = \beta_0^A + \beta_1^A x_{1i} + ... + \beta_k^A x_{ki} + u_i$ from sub-sample A

$y_i = \beta_0^B + \beta_1^B x_{1i} + ... + \beta_k^B x_{ki} + u_i$ from sub-sample B

Restricted model (without structural change):

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

With the restriction:

$H_0 : \beta_j^A = \beta_j^B$ for $j = 0, 1, ..., k$, there is no structural difference.

- Be SSR_{nr} the sum of the OLS square residuals of the non restricted model: $SSR_{nr} = SSR_A + SSR_B$
- Be SSR_r the sum of the OLS square residuals of the restricted model.

Then:

$F = \frac{SSR_r - SSR_{nr}}{SSR_{nr}} \frac{n-2(k+1)}{k+1} \sim F_{k+1,n-2(k+1)}$

If $F_{q,n-K-1} < F$, there is evidence to reject the null hypothesis.

Multicollinearity

If there is exact multicolineality, the equation system of OLS cannot be solved due to infinite solutions.

- Approximate multicollinearity: when one or more variables are almost a constant or there is a linear relation between them. In this context, there is not a problem, given the classic requirements of OLS, and the inference is valid. But, there are some empiric consequences of this:
  - Small sample variations can induce to big variations in the OLS estimations.
  - The variance of the OLS estimator of the x's that are collinear $Var(\hat{\beta}_j)$ increments, then, the inference of the parameter is affected $\rightarrow$ The estimation of the parameter is very imprecise (big confidence interval).

Calculating the Variance Inflation Factor to analyze multicollinearity problems:
$VIF(\hat{\beta}_j) = \frac{1}{1-R_j^2}$

Indicates the increment of $Var(\hat{\beta}_j)$ because of the multicollinearity.

- If it is bigger than 10, indicates that there are multicollinearity problems.
- From 4 onwards, it is advisable to analyze in more detail if there might be multicollinearity.

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

Heteroscedasticity

The residuals u_i of the population regression function do not have the same variance σ^2:
$Var(u|x) = Var(y|x) \neq \sigma^2$

Consequences

- The estimators still are unbiased.
- The estimators still are consistent.
- The variance estimations of the estimators is biased: the construction of confidence intervals and the hypothesis contrast are not reliable.

In this context, OLS is not an unbiased lineal estimator of minimum variance. There is an alternate unbiased lineal estimator of minimum variance denominated estimator of least weighted squares (OLWS) or least generalized squares (LGS).

Detection

Plots. Look for structures in plots with the square residuals) and contrasts: Park test, Goldfield-Quandt, Bartlett, Breush-Pagan, CUSUMQ, Spearman, White.
White test null hypothesis:
$H_0 = Homoscedasticity$

Correction

- When the variance structure is known, use weighted least squares.
- When the variance structure is not known: make assumptions of the possible structure and apply weighted least squares (factible weighted least squares).
- Supposing that σ_i^2 is proportional to x_i^2, divide by x_i.
- New model specification, for example, logarithmic transformation.
- Standard errors with heteroscedasticity corrected by the White's method.

Auto-correlation

The "natural" context of this phenomena is in temporal series.
The residual of any observation, u_i is correlated with the residual of any other observation. The observations are not independent $E(u_i, u_j) \neq 0; i \neq j$

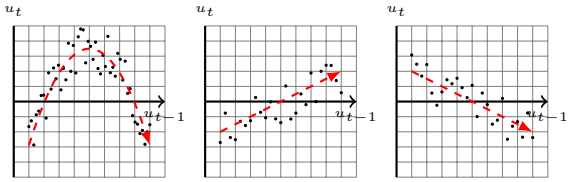
Consequences

OLS estimators are not lineal, not efficient, and are biased. Because OLS estimators are not efficient, variance estimations of the estimators are biased, hypothesis contrast and confidence intervals are not reliable.

Detection

- Graphic residual analysis. There are auto-correlation structures that can be identified in a plot. For example:

AR: AR(+): AR(-):



- Formal contrasts: Breusch-Godfrey. It allows:
  - Dynamic models.
  - u_t that follows an auto-regressive model or ρ order.
  - Moving averages of the error term.

$H_0 : Noauto - correlation$
$H_1 : u_t \sim AR(\rho)$ or $u_t \sim MA(q)$

Correction

Prediction

Two types of prediction:

- Prediction of the mean value of y for a specific value of x.
- Prediction 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 will be less.