Econometrics Cheat Sheet

Basic concepts

Definition of econometrics

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

1. Cross section: data taken at a given moment in time, an static "photo". Order does not matter.
2. Temporal series: observation of one/many variable/s across time. Order does matter.
3. Panel data: consist of a temporal series for each observation of a cross section.
4. Pooled cross sections: combines cross sections from different temporal periods.

Phases of an econometric model

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

Assumptions of the econometric model

Under this assumptions the estimators of the parameters will present "good properties". GAUSS MARKOV ASSUMPTIONS (EXTENDED)

- Parameters linearity.
- The sample of the population is random. Characteristics:
  - Independence: independence, that guarantees that all the co-variances between independents are zero.
  - Identical distribution: that guarantees that the n expected values and variances of the observations are the same.
- $E(u/X_1, X_2, ..., X_k) = 0$, guarantees that the estimations are unbiased, that have some implications:
  - $E(u) = 0$ there are none systematic errors.
  - $Cov(u, X_1) = Cov(u, X_2) = ... = Cov(u, X_k) = 0$ there are no relevant variables not included in the model.
  - $E(Y/X_1, X_2, ..., X_k) = \beta_0 + \beta_1 X_1 + \beta_k X_k$ the lineal relation between Y and $X_1, ..., X_k$ is fulfilled, at least in average.
- Homocedasticity: $Var(u_i/X_{1i}, X_{2i}, ..., X_{ki}) = \sigma^2$, the variability of the error is the same for all levels of x. Guarantees that the estimations are efficient. Implies that: $Var(Y_i/X_{1i}, X_{2i}, ..., X_{ki}) = \sigma^2$, the variability of the dependent variable is the same for all levels of x.
- No auto-correlation: $Cov(u_i, u_j) = 0 \rightarrow Cov(Y_i Y_j/X) = 0$ for every i different from j. The errors do not contain information about other errors.
- The distribution of the errors is normal (is not always necessary).
- No multicolineality: none of the independent variables is constant nor exist an exact (or approximate) linear relation between them, they are linearly independents.
- The number of available data is greater than $k+1$ (β parameters to estimate).

The homocedasticity and no auto-correlation assumptions can also be written in matrix form: $Var(u/X) = \sigma^2 I_n$

Interpretation of the coefficients

Model	Dependent	Independent	Interpretation β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$log(x)$	$\Delta y = (\beta_1/100)[1\%\Delta x]$
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	$x + x^2$	$\Delta y = (\beta_1 + 2\beta_2 x)\Delta x$

OLS estimation of the model

Simple regression model

$Y_i = \beta_0 + \beta_1 X_{1i} + u_i, i = 1, ..., n$

Definitions

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$

Objective is minimize the square sum of residuals:

$Min \sum_{i=1}^{n} \hat{u}_i^2 = Min \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2$

With

$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

$\hat{\beta}_1 = \frac{Cov(Y,X)}{Var(X)}$

Multiple regression model

$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_i, i = 1, ..., n$

$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i + ... + \hat{\beta}_k X_{ki})$

Objective:

$Min \sum_{i=1}^{n} \hat{u}_i^2$

Then

$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - ... - \hat{\beta}_k \overline{X}_k$

$\hat{\beta}_j = \frac{Cov(Y, resid(X_j))}{Var(resid(X_j))}$

Properties of OLS

- Linearity in Y.
- Normality: $Y/X \ N(\beta_0 + \beta_1 X, \sigma^2)$

- Expected value of the estimator: $E(\hat{\beta}_1/X_i) = \beta_1$, then $\hat{\beta}_1$ is an unbiased estimator of β_1
- Variance of the estimator: $Var(\hat{\beta}_1/X_i) = \frac{\sigma^2}{nVar(X_i)}$

Efficiency of OLS estimators, Gauss-Markov Theorem. In the context of the simple or multiple linear regression model, the OLS estimators of the parameters are those with the lowest variance between the lineal and unbiased estimators

Central Limit Theorem

Under the CLT, $\hat{\beta}_j$ is a consistent estimator of the population parameter β_i.

$plim\hat{\beta}_i = \beta_i$

The Central Limit Theorem allow us to obtain (asymptotically):

$\frac{\hat{\beta}_i - \beta_i}{s(\hat{\beta}_i)} \sim N(0,1)$

Goodness of the fit, R-Squared

The R2 is a measure of the goodness of the fit, how the OLS fit to the data.

Is the proportion of variability of the dependent variable explained by the regression line:

$R^2 = \frac{\sum_i (\hat{Y}_i - \overline{Y}_i)^2}{\sum_i (Y_i - \overline{Y}_i)^2} = 1 - \frac{\sum_i \hat{\epsilon}_i^2}{nS_y^2}$

The R^2 takes values between 0 (no lineal explanation of the variations of Y) and one (total explanation of the variations of Y)

Is a descriptive measure of the global fit of the model.

The R^2 measures the percentage of variation of Y that is linearly explained by the variations of X.

The R^2 increments it's value when increments the number of regressors, whatever they are relevant or not.

For eliminate the above phenomena, there is a R^2 corrected by degrees of freedom ($\overline{R}^2$).

$\overline{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{\sum_i \hat{\epsilon}_i^2}{\sum_i (Y_i - \overline{Y}_i)^2} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$

For big sample sizes:

$\overline{R}^2 \approx R^2$

Errors

Standard error of the regression is a measure of the goodness of the fit.

$\hat{\sigma} = \sqrt{\frac{\sum_i \hat{\epsilon}_i^2}{n-k-1}}$

It's value decreases as the number of regressors increase, so it have the same problem as the R^2

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 on 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 (Error type I). 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 significance contrasts

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

Supposing that the model's errors are distributed as a normal distribution.

$Y_i/X_i, ..., X_k \sim N(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \sigma^2)$

Then, under H_0:

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

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

Confidence intervals

Supposing normality of the residuals:

$Y_i/X_i, ..., X_k \sim N(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \sigma^2)$

Then, $t = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k-1}$

The confidence interval:

$P[\hat{\beta}_j - t_{n-k-1,\alpha/2} s(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{n-k-1,\alpha/2} s(\hat{\beta}_j)] = 1 - \alpha$

Regression Analysis

Study and predict the mean value of a variable regarding the base of fixed values of other variables. We usually use Ordinary Least Squares (OLS).

Correlation Analysis

The correlation analysis not distinguish between dependent and independent variables. **Simple Correlation** Measure the grade of lineal association between two variables.

Utilization

Interpretation of the model

Heterocedasticity

The residuals u_i of the population regression function don't have the same variance σ^2:

$$Var(u_i \mid x_i) = \sigma_i^2; i = 1, ..., n$$

Consequences

Under the Gauss-Markov Theorem assumptions, OLS estimators are not efficient. The estimations of the variance of the estimators are biased. The hyphotesis contrast and the confidence intervals are not reliable.

Detection

Plots (look for structures in plots with the square residuals) and contrasts: Park test, Goldfield-Quandt, Bartlett, Breush-Pagan, CUSUMQ, Spearman, White. White's null hypothesis:

$$H_0 = HOMOCEDASTICITY$$

Correction

- When the variance structure is known, use weighted least squares.
- When the variance structure is not known: make asumptions of the possible structure and apply weighted least squares
- Supposing that σ_i^2 is proportional to x_i^2, divide by x_i