Cap 1 Ghysel and Marcelino 2018

sábado, 31 de outubro de 2020 13:22

The Baseline Regression Model

1. Introduction

- Overview of linear regression models covering the following topics:
  - a) specification
  - b) Estimation
  - c) model diagnostic
  - d) Forecasting

1.1. Specification

- Goal: assessing the explanatory power of a set of k variables, grouped into X, for the variable y
- Both X and y are stochastic process
- If we assume a linear marginal contribution of X on y, then we can write:

$$y_t = x_1\beta_1 + \dots + x_{kt}\beta_k + \varepsilon_t \quad , \quad t = 1, \dots, T \qquad (1.2.1) \qquad x_t = [x_{1t} \dots x_{kt}]^T$$

β_i: parameter related to the explanatory power of x_i

- If $\mathbb{E}[\varepsilon_t] = 0 \overset{1.2.1}{\Rightarrow} \mathbb{E}[y_t | x_{1t}, \dots, x_{kt}] = x_{1t}\beta_1 + \dots + x_{kt}\beta_k$
- $\varepsilon_t = y_t - \mathbb{E}[y_t | x_{1t}, \dots, x_{kt}]$
- We can group / stack the T variables to get:

$$\underset{T\times 1}{y} = \underset{T\times k}{X}\,\underset{k\times 1}{\beta} + \underset{T\times 1}{\varepsilon} \qquad (1.2.2)$$

Assumptions (Linear Regression Assumptions) The linear regression model defined on (1.2.2) satisfies the following assumptions:

Weak assumptions:
1) $\mathbb{E}[\varepsilon] = 0$ — construction
2) $\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 I$ — Homoskedasticity
3) $X \perp\!\!\!\perp \varepsilon$
4) X^TX is non singular — lack of perfect multicolinearity
5) X is weak stationary — temporal persistence

Implicit assumptions
1) No omitted variables
2) The DGP is linear in the parameters β
3) The relationship between X and y does not change over time $\beta_{\cancel{t}}$

1.2 Estimation

- To use model (1.2.2) on forecasting we need to estimate the parameters β and σ^2 (variance of the residual)
- If we assume X is deterministic for a moment, from the Gauss-Markov theorem, we know that the best linear unbiased estimator (BLUE) of β is given by:

$$\hat{\beta} = (X^TX)^{-1}X^Ty \qquad (1.3.1)$$

which is called the OLS estimator of β.

- The OLS estimator is obtained by solving the following minimization problem:

$$\hat{\beta} = \underset{\beta}{\arg\min}\ \varepsilon^T\varepsilon = \underset{\beta}{\arg\min} \sum_{t=1}^{T} \varepsilon_t$$

where:

$$\varepsilon_t = \sum_{t=1}^{T}(y_t - x_t\beta)^2 = (y - X\beta)^T(y - X\beta)$$

- Characteristics of the OLS estimator:
  1) Unbiased: $\mathbb{E}[\hat{\beta}] = \beta$
  2) Minimal variance in the class of linear estimators
  3) Consistent: let $\mathrm{var}(\hat{\beta}) = \frac{1}{T}\left(\frac{X^TX}{T}\right)^{-1}$. Then $\lim_{T\to\infty} \mathrm{var}(\hat{\beta}) = 0$. This fact together with 1) means that:

$$\mathbb{P}(|\hat{\beta} - \beta| < \varepsilon) = 1 \qquad \text{(converges in probability)}$$

- Given the OLS estimator for β, the estimated residuals is:

$$\hat{\varepsilon}_t = y_t - x_t\hat{\beta} \Rightarrow \hat{\varepsilon} = y - X\hat{\beta} = y - X(X^TX)^{-1}X^Ty = \underbrace{y[I - X(X^TX)^{-1}X^T]y^T}_{\text{quadratic form}}$$

- We can use the residuals to construct the estimator for its variance:

$$\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon} / \underbrace{(T-k)}_{\text{degrees of freedom}} \Rightarrow \mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

- Until now, we have not assumed any distribution for ε, but its sometimes convenient to assume normality:

$$\varepsilon \sim N(0, \sigma^2 I) \Rightarrow \sqrt{T}(\hat{\beta} - \beta) \sim N\left(0, \left(\frac{X^TX}{T}\right)^{-1}\right)$$

$$\frac{(T-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{T-k}$$

- Adding normality to the assumptions above makes the set of assumptions strong
- Similar results can be derived when X are stochastic but uncorrelated to ε and the sample size $T \to \infty$

1.4. Measure of fit

- Its convenient to examine measures of in sample model fit prior to forecasting, though this should not be the unique forecast selection criteria
- A common in sample measure of fit is the R^2 which is given by:

$$R^2 = 1 - \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{y^Ty} = 1 - \frac{(y - \hat{y})^T(y - \hat{y})}{y^Ty} = \frac{y^Ty - (y-\hat{y})^T(y-\hat{y})}{y^Ty} = \frac{y^Ty - y^Ty + 2y^T\hat{y} + \hat{y}^T\hat{y}}{y^Ty} =$$

- The problem of the R^2 is that its a monotonically increasing function of the number of variables $\Rightarrow k\uparrow \Rightarrow R^2\uparrow$
- Adjusted R^2 correct for this fact:

$$\bar{R}^2 = 1 - \frac{\hat{\varepsilon}^T\hat{\varepsilon} / (T-k)}{y^Ty / (T-k)}$$

- Another alternative for in sample model evaluation are information criteria methods such as BIC and AIC

1.5. Constructing point forecasts

- Goal: forecast y_{t+h} with the parameters estimated as in (1.3.1)
- Assuming that X_{t+h} is known, we have that the BLUE forecast is given by:

$$y_{t+h} = \underset{1\times h}{X_{T+h}}\,\underset{k\times 1}{\hat{\beta}} \qquad (1.5.1)$$

in the sense of producing minimum forecast error variance and zero mean forecast error, where the forecast error is:

$$e_{t+h} = y_{t+h} - \hat{y}_{t+h}$$

- Let us consider the set of linear predictors characterized by $\underset{1\times T}{1 = [1 \dots 1]}$ applied to the data:

$$y_{T+h} - 1^Ty = (X_{T+h}\beta + \varepsilon_{T+h} - 1^TX\beta - \varepsilon) = (X_{T+h} - 1^TX)\beta + \varepsilon_{T+h} - 1^T\varepsilon$$

$$\Leftrightarrow \mathbb{E}[y_{T+h} - 1^Ty] = (X_{T+h} - 1^TX)\beta$$

- We want to find the vector 1 which minimizes this quantity:

$$\mathbb{E}[y_{T+h} - 1^Ty]^2 = \sigma^2(1 + 1^T1) \qquad \text{mean squared error}$$

$$\Rightarrow \hat{y}_{T+h} = 1^{*T}y = X_{T+h}(X^TX)^{-1}X^Ty = X_{T+h}\hat{\beta}$$

1 that minimizes the variance of the prediction error.

- The associated forecast error of the optimal predictor is given by:

$$e_{T+h} = X_{T+h}(\beta - \hat{\beta}) + \varepsilon_{T+h}$$

and its variance is given by:

$$\mathbb{E}[y_{T+h} - \hat{y}_{T+h}] = \mathbb{E}[X_{T+h}(\beta - \hat{\beta}) + \varepsilon_{T+h}]^2 = \sigma^2[\underbrace{1}_{\text{variance of the error}} + \underbrace{\underset{1\times k}{X_{T+h}}\,\underset{k\times k}{(X^TX)^{-1}}\,\underset{k\times 1}{X^T_{T+h}}}_{\text{variance of the parameter}}]$$

- Problems with this approach:
  1) X_{T+h} are typically not observed
  2) y_{T+h} are " " "
  3) The relationship between y and the marginal contributions of X could be non-linear
  4) We have derived the optimal y_{T+h} assuming that we want an unbiased forecast error with minimum variance $\Leftrightarrow$ particular loss function (symmetric)
  5) The optimality of y_{T+h} is valid if the underlying assumptions of the model holds

1.6 Interval and density forecasts

- Assume that assumptions 1-3 holds, the sample size T is large, and the error term has normal distribution
- This means that:

$$\left(\frac{y_{T+h} - \hat{y}_{T+h}}{\sqrt{\mathrm{var}(e_{T+h})}}\right) \sim N(0,1)$$

which implies that:

$$y_{T+h} \sim N(\hat{y}_{T+h}, \mathrm{var}(\hat{e}_{T+h})) \qquad (1.6.1)$$

theoretical density forecast of y_{T+h}

- The equation (1.6.1) is only an approximation for finite sample sizes
- If instead we use the estimators for β and σ^2 we have that the standardized forecast error has a student t distribution with T-k degrees of freedom
- With the correct interpretation, density forecast can be used to assign probabilities to specific events concerning the behaviour of y
- Another use of the forecast density is to construct confidence intervals (interval forecasts) for y_{T+h}
- A $[1-\alpha]\%$ forecast interval is represented as:

$$\left[\hat{y}_{T+h} - c_{\alpha/2}\sqrt{\mathrm{var}(e_{T+h})}\ ,\ \hat{y}_{T+h} + c_{\alpha/2}\sqrt{\mathrm{var}(e_{T+h})}\right] \qquad (1.6.2)$$

- The correct interpretation: suppose that we could generate a very large number of samples of y and X, each of size $T+h$, and for each sample construct the interval forecast for y_{T+h} as in 1.6.2. Then, in $[1-\alpha]\%$ of the samples the realization $\tilde{y}_{T+h}$ will fall into 1.6.2

1.7. Parameter Testing

- Recall the linear regression model in compact notation:

$$y = X\beta + e$$

and assume assumptions 1-5.

- We would like statistical tools to decide whether a particular variable should be included or not in the model
- One key instrument for this goal is hypothesis testing
- A testing problem has four ingredients:
  1. The hypothesis of interest: statement about what we believe true. Often called null hypothesis H_0.
  2. The alternative hypothesis: H_1 or H_a
  3. The test: random variable that must be informative about the null hypothesis, and whose distribution under the null is known.
  4. Decision rule: decision rule that tell us, based on the test statistic, whether to accept or reject the null hypothesis

Problem 1 – $H_0: \beta_i = c \quad H_1: \beta_i < c$

- The test statistics is given by:

$$t\text{-stat} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\mathrm{var}(\hat{\beta}_i)}} \quad , \quad \mathrm{var}(\hat{\beta}_i) = \hat{\sigma}^2\left(\frac{X^TX}{T}\right)^{-1}$$

- Under H_0 we have that:

$$t\text{-stat} = \frac{\hat{\beta}_i - c}{\sqrt{\mathrm{var}(\hat{\beta}_i)}} \sim t_{T-k}$$

- We will reject H_0 as far as the t-stat is a negative value in the far left side of the distribution
- To introduce a decision rule for the test we must define the significance level of the test, which is the maximum probability of rejecting H_0 when its true (controls type I error)
- For the test at hand, α would be a critical value on the left tail of the distribution under H_0
- α has typically values of 1%, 5%, and 10%
- The probability of type I error could be stated as follows:

$$\mathbb{P}_{H_0}(t\text{-stat} < a) = p\text{-value} \quad , \quad a \text{ is the realization of the t-stat}$$

which is the probability of commiting type I error.

- If this probability is under the significance level $\alpha \Rightarrow$ can reject H_0

Problem 2 – $H_0: \beta_i = c \quad H_1: \beta_i \neq c$

- The test statistics is given by:

$$t\text{-stat} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\mathrm{var}(\hat{\beta}_i)}}$$

which under the null is given by:

$$t\text{-stat} = \frac{\hat{\beta}_i - c}{\sqrt{\mathrm{var}(\hat{\beta}_i)}} \sim t_{T-k}$$

- We reject H_0 as the t-stat is far in the left or right side of the distribution
- Given a significance level α, the decision rule could be:

Reject H_0 if $p\text{-value} = \mathbb{P}_{H_0}(|t\text{-stat}| > a) < \alpha$

Problem 3 – $H_0: \beta = c \quad H_1: \beta \neq c$

- In this case, the test statistics is given by:

$$F\text{-stat} = \frac{(\hat{\beta} - c)^T X^TX (\hat{\beta} - c)}{k\sigma^2} = \frac{(\hat{\beta} - c)^T \mathrm{var}(\hat{\beta})^{-1}(\hat{\beta} - c)}{k}$$

$\Rightarrow F\text{-stat} \sim F_{k,T-k}$ (can only take positive values)

- We reject H_0 if:

$$p\text{-value} = \mathbb{P}_{H_0}(F\text{-stat} > a) < \alpha$$