Validation

Econometrics. BAM.

Validation

Validation is the process of deciding whether the numerical results are acceptable as descriptions of the data and are in accordance with economic theory.

Validation is based on the use of

- Inference analysis
- Goodness-of-fit Measurements
- Comparison with alternative estimated models
- Interpretation of the results in economics terms

3.2. Inference

Sometimes it is useful to test whether the parameters take a particular value. We can find some examples in economic theory

-Constant returns to scale in a production function

- Wagner’s law

- Fisher effect

3.2. Inference

We can also test the significance of a parameter in a estimated model.

$$y_t = 2 + 1.5\, x_{1t} - 0.001\, x_{2t}$$

If we can conclude that the parameter that accompanies x_{2t} is 0, then we can remove it from the model.

This can help us to improve the final specification.

3.2. Inference

Thus, we need some statistics that allow us to test for several hypothesis.

To that end, we should take into account that:

-We are assuming that u follows a N(0,s2I) distribution

- If this restriction would not be imposed, we could then consider that X'u goes asymptotically towards a N(0, Σ) distribution and, therefore, we could prove that the vector of estimator goes asymptotically towards a Normal distribution.

3.2. Inference

Under the previously presented assumptions of the GLM, we know that:

$$u \sim N\left(0, \sigma^2 I\right)$$

Then, it is true that:

$$\frac{u}{\sigma} \sim N\left(0, I\right)$$

3.2. Inference

Consequently, we can prove that:

$$\hat{\beta} \sim N\left[\beta, \sigma^2 \left(X'X\right)^{-1}\right]$$

Then, if we want to test for a null hypothesis, we could use the following statistic:

3.2. Inference

- Null hypothesis: $H_o : \beta_i = \beta_i^0$
- Alternative hypothesis; $H_A : \beta_i \neq \beta_i^0$
- The test statistics would be as follows:

$$\frac{\hat{\beta}_i - \beta_i^0}{\sqrt{\sigma^2 \left(X'X \right)^{-1}}} \sim N(0,1)$$

- If we know the value of the variance, we can obtain this statistic and compare it with the critical value of a N(0,1) distribution.

3.2. Inference

Unfortunately, we do not know this value. But, we can circumvent this problem by taking into account that:

$$\frac{N(0,1)}{\sqrt{\frac{\chi_m^2}{m}}} \sim t_m$$

Where the normal and the chi-squared distributions are independent.

3.2. Inference

If we take into account that:

$$\hat{u} = Y - X\hat{\beta} = Y - X\left(X'X\right)^{-1} X'Y =$$

$$= \left[I - X\left(X'X\right)^{-1} X' \right] Y = MY$$

With M being a matrix of residuals

3.2. Inference

Matrix M has the following properties:

- Symmetry:

$$M' = [I - X (X'X)^{-1} X']' = [I - X (X'X^{-1}X'] = M$$

- Idempotency

$$M M = [I - X (X'X)^{-1} X'] [I - X (X'X)^{-1} X'] =$$

$$= I - X (X'X)^{-1} X' - X (X'X)^{-1} X' + X (X'X)^{-1} X' X (X'X)^{-1}X' =$$

$$= I - X (X'X)^{-1} X' - X (X'X)^{-1} X' + X (X'X)^{-1} X' =$$

$$= I - X (X'X)^{-1} X' = M$$

3.2. Inference

Thus, if we take into account that:

$$\frac{\hat{u}'\hat{u}}{\sigma^2} = \frac{u'M\,u}{\sigma^2} = \frac{u'M'Mu}{\sigma^2} = \left(\frac{u}{\sigma}\right)' M \left(\frac{u}{\sigma}\right) = \nu' M \nu$$

Where we should note that the vector ν follows a N(0,I) distribution.

$$\frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2_{\text{T-k}}$$

3.2. Inference

Where we should note that the vector ν follows a N(0,I) distribution. Thus, we have a quadratic form, based on matrix M, which implies that we only have T-k independent components. This is equivalent, in scalar terms, to have T-k independent squared N(0,1). Then, we can conclude that:

$$\frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2_{\text{T-k}}$$

3.2. Inference

Then, we have that:

$$\frac{\dfrac{\hat{\beta}_i - \beta_i^0}{\sqrt{\sigma^2 \left(X'X \right)^{-1}}}}{\dfrac{\hat{u}'\hat{u}}{(T-k)\sigma^2}} = \frac{\hat{\beta}_i - \beta_i^0}{\sqrt{\hat{\sigma}^2 \left(X'X \right)^{-1}}} \sim t_{T-k}$$

3.2. Inference

- Null hypothesis: $H_o : \beta_i = \beta_i^0$
- Alternative hypothesis; $H_A : \beta_i \neq \beta_i^0$
- The test statistic can be stated as follows:

$$\frac{\hat{\beta}_i - \beta_i^0}{\sqrt{\hat{\sigma}^2 \left(X'X \right)^{-1}}} \sim t_{T-k}$$

- Then, for a given significance level (α), the we will reject the null hypothesis whenever this statistic is greater than its correspondent critical value $t_{T-k}^{\alpha/2}$

3.2. Inference

- We cannot reject the null hypothesis whenever the value of the statistic is lower or equal the critical value $t_{T-k}^{\alpha/2}$

- We can alternatively use the p-value. The p-value can be defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed. Then, we can reject the null hypothesis, for a given significance level α, whenever the p-value is lower than this α.

3.2. Inference

- We cannot reject the null hypothesis whenever the value of the statistic is lower or equal the critical value $t_{T-k}^{\alpha/2}$

- We can alternatively use the p-value. The p-value can be defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed. Then, we can reject the null hypothesis, for a given significance level α, whenever the p-value is lower than this α.

3.2. Inference

- We could be interest in testing for multiple hypothesis. For instance, we can test for constant returns to scale in a production function.
- We can state a general hypothesis as follows:

$$H_o : R\beta = R\beta^0$$

- With R being a mxk matrix, where m are the number of hypothesis for being tested.

3.2. Inference

For instance, if we have this model:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

Where we want to test the joint null hypothesis β_1=0 and β_2+β_3 =1, then the matrix R should adopt the following form:

$$R = \begin{pmatrix} 1 0 0 \\ 0 1 1 \end{pmatrix}$$

3.2. Inference

Thus, we need a statistic which is valid for several situations, not just for single hypothesis. To get it, we should note that:

$$R\hat{\beta} \sim N\left[R\beta, \sigma^2 R'\left(X'X\right)^{-1} R\right]$$

And, consequently, it can be proved that:

$$\left(R\hat{\beta} - R\beta\right)'\left[\sigma^2 R'\left(X'X\right)^{-1} R\right]^{-1}\left(R\hat{\beta} - R\beta\right) \sim \chi_r^2$$

3.2. Inference

Again, the variance of the perturbation is unknown and, therefore, we cannot use this statistic. But, we can remove it, by simply considering that the ratio of two independent chi-squared distributions follows a F distribution:

$$\frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}} \sim F_{m,n}$$

3.2. Inference

Then, we have that:

$$F = \frac{\dfrac{\left(R\hat{\beta} - R\beta\right)' \left[\sigma^2 R' \left(X'X\right)^{-1} R\right]^{-1} \left(R\hat{\beta} - R\beta\right)}{m}}{\dfrac{\hat{u}'\hat{u}}{(T-k)\sigma^2}} =$$

$$= \frac{\left(R\hat{\beta} - R\beta\right)' \left[R' \left(X'X\right)^{-1} R\right]^{-1} \left(R\hat{\beta} - R\beta\right)}{m\hat{\sigma}^2} \sim F_{m,T-k}$$

3.2. Inference

Then, in order to test for a joint hypothesis, we should carry out the following steps:

- To explicit the null hypothesis $H_o : R\beta = R\beta^o$

- To explicit the alternative hypothesis $H_A : R\beta \neq R\beta^o$

- To estimate the model and obtain the statistic F

- The null hypothesis is rejected whenever $F > F^{\alpha}_{m,T-k}$ or, alternatively, whenever p-value lower than the significance level.

3.2. Inference

Some very important statistics.

1. We will always test for the individual significance of the parameters of explanatory variables.

$$H_o : \beta_i = 0, \ \forall i = 1, 2, \ldots, k$$

$$H_A : \beta_i \neq 0, \ \forall i = 1, 2, \ldots, k$$

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 \left(X'X \right)^{-1}_{ii}}} = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim t_{T-k}$$

3.2. Inference

Some very important statistics.

2. It is also advisable to test for the joint significance of the parameters of explanatory variables. This statistic is referred to as ANOVA, an acronym that stands for ANalysis Of the VAriance.

$$H_o : \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$$H_A : No\, H_o$$

3.3. Goodness-of-fit

- We can consider that an estimated model is good if it can explain a high proportion of the variation of the dependent variable.
- We then need some statistics that can provide a measure of this ability of explanation
- We need to define when this statistic takes a value that allows us to consider that the model is good.

3.3. Goodness-of-fit

- The most employed statistic is the R^2.

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

$$SSR = \hat{u}'\hat{u} = \sum_{i=1}^{T} \hat{u}_i^2$$

$$SST = y'y - T\ \bar{y}^2 = \sum_{i=1}^{T} \left(y_i - \bar{y} \right)^2$$

3.3. Goodness-of-fit

If we take into account the theorem of Pythagoras, we can prove that:

$$y'y = \hat{y}'\hat{y} + \hat{u}'\hat{u} \Rightarrow$$

$$\Rightarrow y'y - T\,\overline{y}^2 = \hat{y}'\hat{y} - T\,\overline{y}^2 + \hat{u}'\hat{u} \Rightarrow$$

$$\Rightarrow SST = SSE + SSR$$

Then, this statistic is bounded between 0 and 1.

$$0 \leq R^2 \leq 1$$

3.3. Goodness-of-fit

REMARK 1

- If the model does not included an intercept, then the former is not longer true and, consequently, the R^2 can take negative values.

- To avoid this problem, amongst other reasons, I suggest you to always include an intercept in your empirical specification.

3.3. Goodness-of-fit

REMARK 2

- The R^2 is very related to the F-ANOVA statistic. Specifically, this can be stated as follows:

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

- The higher the value of the R^2 , the lower the p-value and, therefore, the easier the rejection of the null hypothesis of non-significance of the parameters of the explanatory variables.

3.3. Goodness-of-fit

REMARK 3

- Let us consider the following specification:

$$Model\ A: \qquad y_t = \beta_1 + \beta_2\, x_{2t} + u_t$$

Then, let us name the R2 of model A as follows:

$$R_A^2 = 1 - \frac{SSR_A}{SST_A} = 1 - \frac{\hat{u}'\hat{u}}{y'y - T\,\overline{y}^2}$$

3.3. Goodness-of-fit

REMARK 3

Alternatively, we consider the following model

$$Model\ B: \quad y_t = \beta_1 + \beta_2\, x_{2t} + \beta_3\, x_{3t} + v_t$$

Then, let us name the R2 of model A as follows:

$$R_B^2 = 1 - \frac{SSR_B}{SST_B} = 1 - \frac{\hat{v}'\hat{v}}{y'y - T\,\bar{y}^2}$$

3.3. Goodness-of-fit

REMARK 3

Given that

$$SSR_A \geq SSR_B$$

Then, we have that:

$$R_A^2 \leq R_B^2$$

3.3. Goodness-of-fit

REMARK 4

- So, you can increment the value of the R^2 by including in the model very stupid variables, that cannot explain the evolution of the dependent variable in economic terms.
- Then, the R^2 cannot be used for comparative purposes when the number of regressors varies.

3.3. Goodness-of-fit

REMARK 4

- In these cases, we can use the adjusted R^2, which is defined as follows:

$$\bar{R}^2 = 1 - \frac{\frac{SSR}{T-k}}{\frac{SST}{T-1}} = 1 - \frac{SSR}{SST}\frac{T-1}{T-k}$$

3.3. Goodness-of-fit

REMARK 4

Then, if we include a new variable in the model then the value of k and the SSR vary. So, the statistic only increases its value if the reduction of the SSR balances the reduction of the degrees of freedom (T-K).

$$\bar{R}^2 = 1 - \frac{\dfrac{SSR}{T-k}}{\dfrac{SST}{T-1}} = 1 - \frac{SSR}{SST}\frac{T-1}{T-k}$$

3.3. Goodness-of-fit

REMARK 4

Consequently, this statistic may take negative values, even if the model includes an intercept.

$$\bar{R}^2 = 1 - \frac{\dfrac{SSR}{T-k}}{\dfrac{SST}{T-1}} = 1 - \frac{SSR}{SST}\frac{T-1}{T-k}$$

3.4. Model Selection

Akaike Information Criterion

$$AIC(i) = \ln\left(\tilde{\sigma}_i^2\right) + \frac{2}{T} k_i$$

3.4. Model Selection

Schwarz Bayesian Information Criterion

$$SBIC(i) = \ln\left(\tilde{\sigma}_i^2\right) + \frac{\ln\left(T\right)}{T} k_i$$

3.4. Model Selection

Hannan-Quinn statistic

$$HQ(i) = T\ln\left(\tilde{\sigma}_i^2\right) + 2k_i \ln\left(\ln\left(T\right)\right)$$

Final Remarks

1. Interval estimation
2. Restricted least squares
3. F-test as a function of the SSR

1. Interval estimation

- We have seen how we can estimate the values of the parameters of the model.
- We have used a point estimator to that end
- But, we should recall that estimators are random variables that follow a distribution and we just obtain a particular value. Instead of this, we can give a range of values.
- This is the interval estimation.

1. Interval estimation

If we take into account that the vector of estimators follows a $N(\beta,\sigma^2(X'X)^{-1})$ distribution, then it the interval estimation of the parameter β_i is stated as:

$$\hat{\beta}_i \pm t_{T-k}^{\alpha/2}\, \hat{\sigma}_{\hat{\beta}_i}$$

FR1. Interval estimation

This implies that if we could repeat the estimation with 100 different samples, the true value of the parameter would belong to the following interval

$$\beta_i \in \left(\hat{\beta}_i - t_{T-k}^{\alpha/2}\, \hat{\sigma}_{\hat{\beta}_i} , \hat{\beta}_i + t_{T-k}^{\alpha/2}\, \hat{\sigma}_{\hat{\beta}_i} \right)$$

In an (1-α)% of these replications.

FR1. Interval estimation

- All the values included in the interval are statistically equivalent.
- Thus, if the interval includes the value 0, this implies that we cannot reject the null hypothesis that the parameter is 0.
- The wider is the interval, the less informative is the estimation, e.g. $(-\infty, \infty)$.

FR2. Restricted Least Squares

- Our model strategy is based on the analysis of the individual significance of the parameters, in order to remove all of them that cannot reject this hypothesis.
- Then, this can be interpreted as the inclusion of some restrictions in our model.
- How can this modify the estimators of the model?

FR2. Restricted Least Squares

Analytically, the new scenario can be defined as follows:

$$\min_{\hat{\beta}} \hat{u}'\hat{u} = \min_{\hat{\beta}} \left(y - X\hat{\beta} \right)' \left(y - X\hat{\beta} \right)$$

$$s.c. \;\; R\beta = r$$

FR2. Restricted Least Squares

Thus, we should optimize the following Lagrangean function:

$$L = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right) + 2\lambda'(R\beta - r)$$

FR2. Restricted Least Squares

If we derivate with respect to b and the we do some algebra, it can be proved that:

$$\hat{\beta}_R = \hat{\beta} + (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} \left(r - R\hat{\beta} \right)$$

FR2. Restricted Least Squares

It can be proved that:

- The RLS and the OLS coincide whenever the OLS exactly holds the restriction.
- If the restriction is true, then the RLS is unbiased.
- In any event, we can prove that:

$$Var\left(\hat{\beta}_R\right) = Var\left(\hat{\beta}\right) - \sigma^2 \left(X'X\right)^{-1} R' \left[R\left(X'X\right)^{-1} R'\right]^{-1} R\left(X'X\right)^{-1}$$

FR2. Restricted Least Squares

- Then, we have that

$$Var\left(\hat{\beta}_R\right) \leq Var\left(\hat{\beta}\right)$$

- Accordingly, if we include an incorrect restriction, the new estimators are biased but show lower variance.
- Then, if we define the MSE = bias^2 + Var, the inclusion of a incorrect restriction may provide us better estimators than the OLS.

FR3. F-test as function of the SSR

If we take into account the definition of the RLS, we can use this equation in order to re-define the F-test.

$$Var\left(\hat{\beta}_R\right) \leq Var\left(\hat{\beta}\right)$$

Let us consider that we want to test for the restriction Rβ=r,where R is mxk matrix, then the F-test can be defined as follows:

$$F = \frac{SSR_R - SSR}{SSR}\frac{T-k}{m} \sim F_{m,\,T-k}$$

FR3. F-test as function of the SSR

In particular, if we what to test whether a subset of the coefficients is 0, this statistic can be stated as function of the R2 of two models, provided that the SST do not vary

$$F = \frac{SSR_R - SSR}{SSR}\frac{T-k}{m} = \frac{\frac{SSR_R}{SST} - \frac{SSR}{SST}}{\frac{SSR}{SST}}\frac{T-k}{m}$$

$$= \frac{\left(1-R_R^2\right)-\left(1-R^2\right)}{\left(1-R^2\right)}\frac{T-k}{m} = \frac{R^2 - R_R^2}{\left(1-R^2\right)}\frac{T-k}{m} \sim F_{m,\,T-k}$$