Statistical Inference

- Considering the simple model

$$Y_i = \alpha + \beta X_i + u_i$$

with

$$\begin{aligned} E(u_i) &= 0 \\ var(u_i) &= \sigma^2 \quad \forall\ i \\ u_i \text{ and } u_j \text{ independent for } i &\neq j \\ & X \text{ non stochastic} \end{aligned}$$

- Make additional assumption that errors are normally distributed and can test hypotheses about $\widehat{\alpha}$ and $\widehat{\beta}$

Consider $\widehat{\beta}$ we know

$$E(\widehat{\beta}) = \beta \text{ and } \text{var}(\widehat{\beta}) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$$

but σ^2 is unknown and we have to estimate it.:

$$\widehat{\sigma}^2 = \frac{\sum (Y_i - \overline{Y})^2}{n-k}$$

which is an unbiased estimator.

- Now we can show that for k degrees of freedom

$$\begin{aligned} \frac{\widehat{\beta} - \beta}{\sqrt{var(\widehat{\beta})}} &\sim t_{n-k} \\ \Rightarrow \quad \frac{\widehat{\beta} - \beta}{se(\widehat{\beta})} &\sim t_{n-k} \end{aligned}$$

which holds generally for multiple regression when testing one coefficient
So we can derive conficence intervals. For example for 30 observations and k=2

$$\Pr ob \left[-2.048 < \frac{\widehat{\beta} - \beta}{se(\widehat{\beta})} < 2.048 \right] = 0.95$$

if we want upper or lower limits for β we can construct one sided intervals:

$$\Pr ob \left[t < 1.70 \right] = 0.95$$

Testing Hypotheses

Usually we test the hypthesis that $\beta = 0$ and this is what is reported in the t ratio in the Microfit regression output

$$\frac{\widehat{\beta} - 0}{se(\widehat{\beta})} = \frac{\widehat{\beta}}{se(\widehat{\beta})} \sim t_{n-k}$$

We can test $\beta = 1$ or some other value„ but will need to calculate this ourselves

$$\frac{\widehat{\beta} - 1}{se(\widehat{\beta})} \sim t_{n-k}$$

Note:

For 28 degrees of freedome 5% probability points are ±2.048 for the two sided (tailed) test and 1.70 for the one sided. If both high and low t values are considered as evidence against the hypothesis then we reject if the observed t is greater than 2.048 or less than -0,2048. We could consider only very high or very low values and use t$\lessgtr$ 1.7 depending on the hypothesis sign ie use a one tail tes, which discounts either the negative or the postive as impossible .

For 5% probability points at 30degrees of freedom t becomes 2.042 and is approximately 2.0. For most of our samples we expect to have 30 or more dofs and so we use the rule of thumb that $|t| > 2$

Microfit reports the significance level:, esentially the area under the tails of the curve at the t value that is calculated> So if the value of t is 2.0 with more than thirty dgrees of freedom the level will be 0.05. If the t value were bigger the level would decline and vice versa. So if the significance level is less than 0.05 we can reject the null at the 5% level.

The chosen significance level is not 'God given' it is simply accepted practice and can be adjusted to suit the purpose it was required for.

An important concern is the implication of the chosen significance level:

Type 1 error: rejecting the null when it is true — Prob(Type1) = chosen significance level

Type 2 error: failing to reject the null when it is false — Prob(Type2) depends on what β actually is

Note that Type 2 errors will decline as the sample increase

Type 2 error: failing to reject the null when it is false

F Test and the R Squared

Analysis of Variance: Have seen that we can break down the total sum of squares

$$\sum (Y - \overline{Y})^2 = \sum (\widehat{Y} - \overline{Y})^2 + \sum (Y - \widehat{Y})^2$$

that is the total sum of squares (TSS) is equal to the explained sum of squares (ESS) + the residual sum of squares (RSS)

So

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\widehat{Y} - \overline{Y})^2}{\sum (Y - \overline{Y})^2} = \frac{\sum (Y - \overline{Y})^2 - \sum (Y - \widehat{Y})^2}{\sum (Y - \overline{Y})^2} 1 - \frac{\sum (Y - \widehat{Y})^2}{\sum (Y - \overline{Y})^2} = 1 - \frac{RSS}{TSS}$$

Assuming $\beta = 0$

If Y_i are independent samples from a normal distribution $Y \sim N(\mu, \sigma^2)$ then

$$\begin{aligned} (Y - \mu)/\sigma &\sim N(0,1) \\ \sum_i \frac{(Y_i - \widehat{Y})^2}{\sigma^2} &\sim \varkappa^2_{n-1} \end{aligned}$$

losing one dof because we use the sample mean.

Now for

$$\begin{aligned} Y_i &= \alpha + \beta X_i + u_i \\ u_i &\sim N(0, \sigma^2) \\ &\Rightarrow Y_i - (\widehat{\alpha} + \widehat{\beta} X_i) = \widehat{u}_i = (Y_i - \widehat{Y}_i) \sim N(0, \sigma^2) \\ &\Rightarrow \sum_i \frac{(Y_i - \widehat{Y}_i)^2}{\sigma^2} = \frac{\sum \widehat{u}_i}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \varkappa^2_{n-2} \end{aligned}$$

more generally

$$\frac{RSS}{\sigma^2} \sim \varkappa^2_{n-k-1}$$

where k is the number of explanatory variables

similarly we can show that under the null that the explanatory variables are all insignificant, ie have coefficients that are not significantly different to zero.

$$\frac{ESS}{\sigma^2} = \frac{\sum (\widehat{Y}_i - \overline{Y})^2}{\sigma^2} \sim \varkappa^2_k$$

Now if we take 2 independent chi-squared distributed variables $X_1$and X_2 then the ratio of the two variables divided by their degrees of freedom is an F distribution

$$\frac{X_1/n_1}{X_2/n_2} \sim F_{(n_1, n_2)}$$

hence

$$\frac{\left(\frac{ESS}{\sigma^2}\right)/k}{\left(\frac{RSS}{\sigma^2}\right)/n-k-1} \sim F_{(k, n-k-1)}$$

meaning

$$\frac{ESS/k}{RSS/n-k-1} \sim F_{(k, n-k-1)}$$

this is now a test that the explanatory variables coefficients (apart from the constant) are not significant jointly and we can compare with the F distribution for 95%. If the F values exceeds the critical value we can reject the hypothesis

NB It is possible for a set of variables to be jointly significant even if they are individually insignificant.

Can generalise this to create a test of when a subset of the variables in a model are insignificant

Writing RRSS as the RSS obtained from the restricted model (when the r restrictions are imposed) and URSS as the RSS obtained from the full untestricted model, it can be shown that under the null

$$\frac{(URSS - RRSS)/r}{\sigma^2} \sim \varkappa_r^2$$

Meaning

$$\frac{(RRSS - URSS)/r}{URSS/(n-k-1)} \sim F_{(r,n-k-1)}$$

This is available in Microfit as a variable deletion test in the Post Regression Menu.

We could use this to test more than one restriction as long as they are line -an example might be $\beta_1 + \beta_2 = 0$ and $\beta_3 = 1$. We just impose the restrictions to get RRSS.

If there was only one coefficent being tested (eg $\beta_3 = 0$) then we would get

$$\frac{(RRSS - URSS)}{URSS/(n-k-1)} \sim F_{(1,n-k-1)}$$

But the t ratio we discussed above is much easier to use in this case -in fact $F = t$

Can show where t test comes from:

$$\begin{aligned} \widehat{\beta} &\sim N(\beta, var(\widehat{\beta})) \\ Var(\widehat{\beta}) &= \frac{\sigma^2}{\sum (X_i - \overline{X})^2} \end{aligned}$$

thus

$$\frac{\left(\widehat{\beta} - \beta\right)}{\sqrt{\frac{\sigma^2}{\sum (X_i - \overline{X})^2}}} \sim N(0,1)$$

Now we know

$$\frac{\sum \widehat{u}_i}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \varkappa_{n-2}^2$$

Now a normal distribution divided by chi-squared will give a t distribution so:

$$\frac{\frac{\left(\widehat{\beta}-\beta\right)}{\sqrt{\frac{\sigma^2}{\sum(X_i-\overline{X})^2}}}}{\frac{\sum\widehat{u}_i}{\sigma^2}} = \frac{\left(\widehat{\beta}-\beta\right)}{\sqrt{\frac{\sigma^2}{\sum(X_i-\overline{X})^2}}}\cdot\frac{\sigma^2}{\sum\widehat{u}_i} = \frac{\left(\widehat{\beta}-\beta\right)}{\sqrt{\frac{\sum\widehat{u}_i^2/n-2}{\sum(X_i-\overline{X})^2}}} = \frac{\left(\widehat{\beta}-\beta\right)}{\sqrt{\frac{\sum\widehat{\sigma}^2}{\sum(X_i-\overline{X})^2}}}$$

$$= \frac{\left(\widehat{\beta}-\beta\right)}{\sqrt{var\widehat{\beta}}} = \frac{\left(\widehat{\beta}-\beta\right)}{se(\widehat{\beta})} \sim t_{n-2}$$

There is also a clear link between the t test and the r squared, which is the squared correlation coefficient when we only have one explanatory variable. We can show

$$t^2 = \frac{(n-2)r^2}{1-r^2}$$

giving a relation between the F test of $\beta = 0$ and r^2 and

$$r^2 = \frac{t^2}{t^2+(n-2)}$$

giving a relation between the t test of $\beta = 0$ and r^2