Inference in the multiple regression model

Susan Thomas
IGIDR, Bombay

3 November, 2008

Recap: a three-variable regression model

- $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$
  - Independence: $Y_i, X_{1,i}, X_{2,i}$ are independent across i
  - Normality of Y_i conditional on $X_{1,i}, X_{2,i}$: $Y_i \sim N[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}, \sigma^2]$
  - $X_{1,i}, X_{2,i}$ is exogenous
- Parameters: $\beta_0, \beta_1, \beta_2, \sigma^2$
- Reparameterised model: $Y_i = \delta_0 + \delta_1 X_{1.0,i} + \delta_2 X_{2.0.1,i} + u_i$
  Parameters: $\delta_0, \delta_1, \delta_2, \sigma^2$

Solution space for coefficients

- 2-variables – $\delta_1 = \frac{\sum_i Y_i X_{1.0,i}}{\sum_i X^2_{1.0,i}}$
  which is related to the correlation between Y, X_1.
- 3-variable model – δ_1, δ_2:

$$\hat{\beta}_2 = \hat{\delta}_2 = \frac{\sum_i Y_i X_{2.0.1,i}}{\sum_i X^2_{2.0.1,i}}$$

  which is related to the partial correlation between $Y, X_{2.0.1}$, where $X_{2.0.1}$ is X_2 net of the effects of X_0, X_1.
- New concept: partial correlation.

$$r_{y.x_1,x_2} = \frac{r_{y,x_2} - r_{y,x_1} * r_{x_1,x_2}}{\sqrt{(1 - r^2_{y,x_1})(1 - r^2_{x_1,x_2})}}$$

- New aspect of the estimation problem: Collinearity. If $r^2_{x_1,x_2} = 1$, then $\hat{\delta}_1, \hat{\delta}_2$ cannot be estimated.

Solution space for $\hat{\beta}_0$

- 2–variables: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1$
- 3–variables: $\hat{\delta}_0 = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$
- **HW**: Work out what is $\hat{\beta}_0$ in terms of $\hat{\beta}_1, \hat{\beta}_2, \bar{X}_1, \bar{X}_2, \bar{Y}$.

Solution space for $\hat{\sigma}^2$

- 2–variables: $Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\epsilon}_i^2}{N-2}$$

- 3–variables: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$

$$\hat{\sigma}^2 = \frac{\sum_i \hat{u}_i^2}{N-3}$$

Inference for $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$

1. Test the estimated coefficient $\hat{\beta}_i$ against H_0 individually. Distribution of $\hat{\beta}_i$ is N(β_i, σ_i^2).
2. Test whether the model is significant. Benchmark model: $Y_i = \beta_0 (= \bar{Y}) + \epsilon_i$
   Tests used: LR test, R^2 (F-distribution)
3. Test sets of coefficients jointly. For example, $H_0 : \beta_1 = \beta_2, H_A : \beta_1 \neq \beta_2$
   Tests used: LR test after *reparameterising the model* and re-estimating it.

Testing single coefficient estimates

Distribution of $\hat{\beta}_1$

- $\beta_1 = \frac{\sum_i Y_i X_{1.0.2,i}}{\sum_i X^2_{1.0.2,i}}$
- By expanding $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$, it can be shown that

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i X_{1.0.2,i} u_i}{\sum_i X^2_{1.0.2,i}}$$

1. $E(\hat{\beta}_1) = \beta_1$
2. $var(\hat{\beta}_1) = \sigma^2_u / \sum_i X^2_{1.0.2,i}$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2_u}{\sum_i X^2_{1.0.2,i}}\right)$$

- 2-variable model:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2_\epsilon}{\sum_i X^2_{1.0,i}}\right)$$

- With the distribution of $\hat{\beta}_i$ we can test H_0 for $\hat{\beta}_1$.

HW No. 1

What is the distribution for the estimator for β_0?

Link between β_i and δ_i

- Model: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$
- Reparameterised model: $Y_i = \delta_0 + \delta_1 X_{1.0,i} + \delta_2 X_{2.0.1,i} + u_i$
- $\text{cov}(\delta_0, \delta_1) = \text{cov}(\delta_0, \delta_2) = \text{cov}(\delta_1, \delta_2) = 0$
- $\text{cov}(\beta_2, \beta_1) = \frac{-\sigma_u^2 r_{2,1.0}}{(1-r_{2,1.0}^2)\sqrt{\sum_i X_{2.0,i}^2 \sum_i X_{1.0,i}^2}}$
- Again: if $r_{2,1.0} \sim 1$, then $(1 - r_{2,1.0}^2) \sim 0$ and the estimator variances will be very large.

Testing the overall model

The form of the Liklihood Ratio (LR) test

- The LR test takes the form:

$$-2 \log Q = -2 \log(L_{\text{restricted}} / L_{\text{unrestricted}})^{-\frac{N}{2}} = -N \log \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_R^2} \right) \sim \chi^2[m$$

  where m is the number of restrictions.

- For example, if we want to test $H_0 : \beta_2 = 0$, then
  1. $\sigma_R^2 = \sigma_w^2$ from model $Y_i = \beta_0 + \beta_1 X_1 + w_i$
  2. $\sigma_U^2 = \sigma_u^2$ from model $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$ and the test statistic is:
  3. $\chi^2[1]$

The R^2

- $R^2 = \frac{\text{ESS}}{\text{TSS}}$
- $N\sigma_u^2$ = RSS = TSS - ESS = (1 - R^2) TSS
- In 2-variable model: $R^2 = r^2_{Y,X_1}$
- In 3-variable model:

$$\begin{aligned} N\sigma_u^2 &= (1 - r^2_{Y,X_{2.0.1}})(1 - r^2_{Y,X_{1.0}})TSS \\ (1 - R^2) &= (1 - r^2_{Y,X_{2.0.1}})(1 - r^2_{Y,X_{1.0}}) \end{aligned}$$

- **Note:** for $Y_i = \beta_0 + \beta_1 X_1 + w_i$, RSS is

$$N\sigma_w^2 = (1 - r^2_{Y,X_{1.0}})\text{TSS}$$

LR, R^2 for restriction $\beta_2 = 0$

- Restricted model: $Y_i = \beta_0 + \beta_1 X_1 + w_i$
  $\sigma^2_R = \sigma^2_w = (1 - r^2_{Y,X_{1.0}})\text{TSS}$
- LR = $-N \log\left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2_R}\right) \sim \chi^2[m = 1]$

$$
\begin{aligned}
LR_{\beta_2=0} &= -N\frac{(1 - r^2_{Y,X_{2.0.1}})(1 - r^2_{Y,X_{1.0}})\text{TSS}}{(1 - r^2_{Y,X_{1.0}})\text{TSS}} \\
&= -N(1 - r^2_{Y,X_{2.0.1}}) \sim \chi^2[1]
\end{aligned}
$$

LR, R^2 for restriction $\beta_1 = \beta_2 = 0$

- Restricted model: $Y_i = \beta_0 + \epsilon_i$
  $\sigma_R^2 = \sigma_\epsilon^2 =$TSS
- LR = $-N \log\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_R^2}\right) \sim \chi^2[m = 2]$

$$\begin{aligned} LR_{\beta_1=\beta_2=0} &= -N\frac{(1 - r^2_{Y,X_{2.0.1}})(1 - r^2_{Y,X_{1.0}})\text{TSS}}{\text{TSS}} \\ &= -N(1 - R^2) \sim \chi^2[2] \end{aligned}$$

Some assumptions in the above tests

- The test for $\beta_1 = \beta_2 = 0$ is done as if:
  1. Test for $\beta_1 = 0$ **given that** β_2 has been shown to be 0.

     $$\mathrm{LR}_{\beta_2=\beta_1=0} = \mathrm{LR}_{\beta_2=0} + \mathrm{LR}_{\beta_1=0|\beta_2=0}$$

  2. Therefore, the assumption is that these models are "correct" when we do a joint test.
     It may not be so: topic of "misspecification of models" while accepting/rejecting null hypothesis about the model.
- Within this framework, the most robust test is to do with testing the significance of any one coefficient.

Further tests of parameters

Linear hypothesis with > 1 parameter

- For example: $H_0 : \beta_1 = \beta_2$
- Here, the degrees of freedom is 1 for the test, but there are two parameters involved.
- Such tests are done by reparameterising the model to reflect the restriction.
- For $H_0 : \beta_1 = \beta_2$ in model: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$.
  Reparameterised model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_1 + X_2) + (\beta_2 - \beta_1)X_2 + u_i \\ &= \delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + u_i \end{aligned}$$

  Test: $H_0 : \delta_2 = 0$.
- The framework of the LR test statistic remains the same, except that $m = 1$.

Test whether the model is better as $H_0 : \beta_1 = 1$

- Test with the LR framework and the hypothesis set as a model restriction.
- This also starts with reparameterising the model as .

$$
\begin{aligned}
Y_i - X_{2,i} &= \beta_0 + \beta_1 X_{1,i} + (\beta_2 - 1) X_{2,i} + u_i \\
T_i &= \delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + u_i
\end{aligned}
$$

Again: $H_0 : \delta_2 = 0$.

- The framework of the LR test statistic remains the same, except that $m = 1$.

HW No. 2

- Work through the examples on Pages 116 and 117, in "Econometric Modeling" by Hendry and Nielsen.
- Work through Chapter 8, in the same book