Model Mis-specification

Covered Topics

- Functional forms
- Underfitting
- Overfitting
- linearity vs non-linear (Ramsey's RESET)

Functional Forms (1)

- **Linear linear**
- linear log
- linear reciprocal
- **quadratic (polynomial)**
- **interaction (cross terms)**
- **log linear**
- log reciprocal
- log quadratic
- **log log**
- **logistic**

Functional Forms (2)

log-log: $\ln Y_i = \beta_1 + \beta_2 \ln X_i + \varepsilon_i$

log-linear: $\ln Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$

interaction:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{2i} X_{3i}) + \varepsilon_i$$

logistic:

$$\ln \frac{Y_i}{1 - Y_i} = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Note that all are linear in parameters.
==>OLS applies.

Functional Forms (3)

Assumption ==>form. For example,

- constant elasticity => double log (log log)
- constant rate of change =>log linear
- Y between 0 and 1 =>logistic
- variable slope => interaction or polynomial
- combination, e.g., log-log+interaction

There could be more than one that fit.

Box-Cox Transformation (1)

Box-Cox transformation for X

$$B(X,\lambda)=\frac{X^{\lambda}-1}{\lambda}$$

Note that $B(X,1)=X-1$ and $B(X,0)=\ln X$ <==Why?

<u>Model</u>

$$\frac{Y_i^{\lambda}-1}{\lambda}=\beta_1+\beta_2\frac{X_i^{\lambda}-1}{\lambda}+\varepsilon_i$$

λ=1 ==> lin-lin model

λ=0 ==> double log model

Otherwise, non-linear model (need NLS or MLE)

Box-Cox Transformation (2)

$$\frac{Y_i^{\lambda}-1}{\lambda}=\beta_1+\beta_2\frac{X_i^{\mu}-1}{\mu}+\varepsilon_i$$

- Use NLS or MLE to select the best value of (λ,μ).
- No need to pre-choose the specific functional form of the model.
- Require more computational effort. No big deal.

Explanatory Variables

The complete list of X's is purely based on theoretical reasons.

Underfitting = exclusion of relevant variable X's

Overfitting = inclusion of irrelevant variable X's

What are their effects?

Under-fitting (1)

Let X_K be the omitted variable with $\beta_K \neq 0$

$$Y_i = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \ldots + \gamma_{K-1} X_{K-1,i} + \nu_i$$

<u>Effects</u>

OLS estimator of γ will be a biased estimator of β. if the omitted variable is related to the remaining X's. True?

Under-fitting (2)

Let $X_{Ki} = \theta_1 X_{1i} + \theta_2 X_{2i} + \ldots + \theta_{K-1} X_{K-1,i} + \xi_i$

with $\theta_k \neq 0$ for some k=1,…,K-1

If $\theta_k \neq 0, \hat{\gamma}_k$ will be a biased estimator of β_k because $\gamma_k=\beta_k+\beta_K\theta_k$. Note that γ_k includes not only the effect of X_k but also that of the omitted variable(X_K).

Under-fitting (3)

Substitute into the exact model

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{K-1} X_{K-1,i}$$
$$+ \beta_K(\theta_1 X_{1i} + \theta_2 X_{2i} + \ldots + \theta_{K-1} X_{K-1,i} + \xi_i) + \varepsilon_i$$
$$= (\beta_1 + \beta_K \theta_1) X_{1i} + (\beta_2 + \beta_K \theta_2) X_{2i}$$
$$+ \ldots + (\beta_{K-1} + \beta_K \theta_{K-1}) X_{K-1,i}$$
$$+ (\varepsilon_i + \beta_K \xi_i)$$

Over-fitting (1)

Define Z = irrelevant variable

$$Y_i = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \ldots + \gamma_K X_{Ki} + \delta Z_i + \nu_i$$

Since Z is known to be irrelevant, the real δ=0.

Effects

- OLS estimator of γ is an unbiased estimator of β because γ =β.
- loss of accuracy. Why?

Mis-fitting (1)

<u>Rules</u>

- Include all the explanatory variables suggested by the underlying theories.
- Excluding them requires theoretical explanation.
- Even if the test statistic indicates insignificance, leave them in the model to avoid unbiasedness. Low statistic does not imply irrelevance. Data just can't reveal it.

Mis-fitting (2)

- Add variables to test their relevancy.

  Remove the insignificant variables.

  Further theories could be developed if the test indicates their significance.

Ramsey's RESET (1)

<u>RE</u>gression <u>S</u>pecification <u>E</u>rror <u>T</u>est

- test the linear (in X) model against unspecified non-linear model

<u>Concept</u>

Using Taylor series expansion, a non-linear model can be expressed as a polynomial model. If the exact model is non-linear, using a linear model is equivalent to omitting variables (high order terms)

Ramsey's RESET (2)

<u>Test Equation</u>

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \delta_2 \hat{Y}_i^2 + \delta_3 \hat{Y}_i^3 + \ldots + \delta_M \hat{Y}_i^M + \varepsilon_i$$

Perform an F-test or Chi-square test

$$H_0 : \delta_2 = \delta_3 = \ldots = \delta_M = 0$$

$$H_1 : \delta_2 \neq \delta_3 \neq \ldots \neq \delta_M \neq 0$$

EViews can do RESET.

Ramsey's RESET (3)

Notes

- Start with a square term
- The highest order *M* must be pre-selected. *M*-1 terms added.
- Reject H_0 => need a new model
- The test does not suggest the form of a new model. Try a polynomial order *M*.

Model Selection Criteria

Covered Topics

- Criteria
- Adding/dropping variables
- Double linear against double log
- Linear with different X's

Forecasting Error

Forecasting error = $Y_i - \hat{Y}_i$

Linear model $\hat{Y}_i = \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots \hat{\beta}_K X_{Ki}$

Double log

$$\widehat{\ln Y_i} = \hat{\beta}_1 \ln X_{1i} + \hat{\beta}_2 \ln X_{2i} + \ldots \hat{\beta}_K \ln X_{Ki}$$

$$\hat{Y}_i = \exp(\widehat{\ln Y_i})$$

General Selection Criteria (1)

Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}$$

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

General Selection Criteria (2)

Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|$$

Theil's inequality coefficient

$$U = \frac{\sqrt{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}}{\sqrt{\sum_{i=1}^{n}Y_i^2} + \sqrt{\sum_{i=1}^{n}\hat{Y}_i^2}}$$

Relevancy of Variables (1)

Should $Z_1, Z_2, \ldots, Z_M$ be added?

Test Equation

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \delta_1 Z_{1i} + \delta_2 Z_{2i} + \ldots + \delta_M Z_{Mi} + \varepsilon_i$$

Perform an F-test or χ^2-test on

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_M = 0$$

$$H_1 : \delta_1 \neq \delta 2 \neq \ldots \neq \delta_M \neq 0$$

Relevancy of Variables (2)

Note that $F_{cal} \sim F(M, n-(K+M))$

$$\chi^2_{cal} \sim \chi^2(M)$$

EViews gives the 2-run F-test and the LR test.

Reject H_0 implies that the variables may be relevant. It needs explanation. No harm to the unbiasedness of the estimator of parameters β which have been already included.

Another test equation is the old residuals against all the X's (old & new) w/ a constant term.

Redundancy of Variables (1)

Should $X_{\text{K-M+1}}, X_{\text{K-M+2}},\dots, X_{\text{K}}$ be dropped?

Test Equation

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{K-M} X_{K-M,i} + \beta_{K-M+1} X_{K-M+1,i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$$

Perform an F-test or χ^2-test on

$$H_0 : \beta_{K-M+1} = \beta_{K-M+2} = \ldots = \beta_K = 0$$

$$H_1 : \beta_{K-M+1} \neq \beta_{K-M+2} \neq \ldots \neq \beta_K \neq 0$$

Redundancy of Variables (2)

Note that
$$F_{cal} \sim F(M, n-K)$$
$$\chi^2_{cal} \sim \chi^2(M)$$

EViews also gives 2-run F-test and LR test results.

Accept H_0 does <u>not</u> imply that the variables are not relevant. It just says that they are redundant. With the other X's in the model, they are not needed to explain the variation of Y. If they are dropped, there might be estimation bias but no harm to prediction of Y.

MWD Test (1)

MacKinnon-White-Davidson Test

Choose between lin-lin and log-log with same X and Y

<u>Model A</u>

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$$

<u>Model B</u>

$$\ln Y_i = \gamma_1 \ln X_{1i} + \gamma_2 \ln X_{2i} + \ldots + \gamma_K \ln X_{Ki} + \varepsilon_i$$

MWD Test (2)

Note that R^2 cannot be used to judge.

Given $\hat{Y}_i$ the fitted value from Model A

$\widehat{\ln Y_i}$ the fitted value from Model B

Define

$$Z_{Ai} = \ln \hat{Y}_i - \widehat{\ln Y_i}$$

$$Z_{Bi} = \hat{Y}_i - \exp(\widehat{\ln Y_i})$$

MWD Test (3)

Model A'

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki\,i}$$
$$+ \delta_A Z_{Ai} + \varepsilon_i$$

t-Test on $H_0 : \delta_A = 0$

$H_1 : \delta_A \neq 0$

Accept H_0 => double lin “encompasses” double log

MWD Test (4)

Model B'

$$\ln Y_i = \gamma_1 \ln X_{1i} + \gamma_2 \ln X_{2i} + \ldots + \gamma_K \ln X_{K\,i}$$
$$+ \delta_B Z_{Bi} + + \varepsilon_i$$

t-Test on $H_0 : \delta_B = 0$

$H_1 : \delta_B \neq 0$

Accept H_0 =>double log “encompasses” double line

MWD Test (5)

<u>Conclusion</u>

	$\delta_B = 0$	$\delta_B \neq 0$
$\delta_A = 0$	No clear preference	Double linear is chosen
$\delta_A \neq 0$	Double log is chosen	Neither model good enough

J-test (1)

Davidson-MacKinnon's J-test

Choose between two linear models with different set of explanatory variables but same dependent variable.

R^2 cannot be used either.

<u>Model C</u> $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$

<u>Model D</u> $Y_i = \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \ldots + \gamma_L Z_{Li} + \varepsilon_i$

J-test (2)

<u>Model C'</u>

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki\,i} + \delta_C \hat{Y}_{Di} + \varepsilon_i$$

where $\hat{Y}_i^D$ is the fitted value from Model D

t-Test on $H_0 : \delta_C = 0$

$H_1 : \delta_C \neq 0$

Accept H_0 =>model C "encompasses" model D

J-test (3)

<u>Model D'</u>

$$Y_i = \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \ldots + \gamma_L Z_{Li} + \delta_D \hat{Y}_{Ci} + \varepsilon_i$$

where $\hat{Y}_{Ci}$ is the fitted value from Model C

t-Test on $H_0 : \delta_D = 0$

$H_1 : \delta_D \neq 0$

Accept H_0 =>model D "encompasses" model C

J-test (4)

<u>Conclusion</u>

	$\delta_D = 0$	$\delta_D \neq 0$
$\delta_C = 0$	No clear preference	Model C is chosen
$\delta_C \neq 0$	Model D is chosen	Neither model good enough