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?

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

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 $\beta_K \neq 0$

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

Effects

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

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

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 $\gamma = \beta$.
- loss of accuracy. Why?

Mis-fitting (1)

Rules

- 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)

REgression Specification Error Test

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

Concept

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)

Test Equation

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

Perform an F-test or Chi-square test

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

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

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*.