# L10: Model specification and diagnosis

**Feng Li**
feng.li@cufe.edu.cn

**School of Statistics and Mathematics**
**Central University of Finance and Economics**

Revision: December 3, 2014

# Today we are going to learn...

1. **Model specification**

2. **Model selection criteria**

# Model selection criteria

- Be data admissible: the logical prediction
- Be consistent with theory
- Have weakly exogenous regressors: $\mathrm{cor}(X_i, X_j) = 0, \mathrm{cor}(X_i, u) = 0$
- Exhibit parameter constancy
- Exhibit data coherency: white noise of the data.
- Be encompassing: have the **best** model (if possible)

# Specification errors
## ↪ Omitting a relevant variable

- Suppose the true model is

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

  but you fit the following model Instead

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i$$

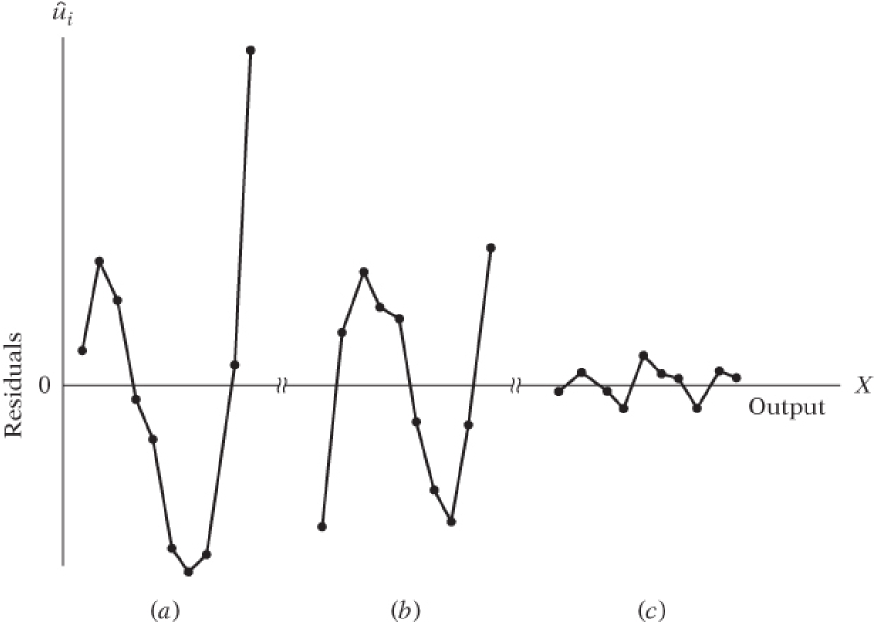
- The model is **underfitted**.
- The consequences:
  - If $X_2$ and $X_3$ are correlated, $E(\hat{\alpha}_1) \neq \beta_1$, $E(\hat{\alpha}_2) \neq \beta_2$ **see Appendix 13A**
  - Even though $E(\hat{\alpha}_2) = \alpha_2$ but $E(\hat{\alpha}_1) \neq \alpha_1$
  - $\hat{\sigma}^2$ incorrectly estimates $\sigma^2$
  - $var(\hat{\alpha}_2) \neq var(\hat{\beta}_2)$
  - Forecasting confidence intervals will be unreliable.

# Specification errors
## ↛ Omitting a relevant variable

- Use residuals to test for omitted variables.
  Consider the three models (a),(b) and (c)

$$
\begin{aligned}
Y_i &= \lambda_1 + \lambda_2 X_i + u_i \ (a) \\
Y_i &= \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^3 + u_i \ (b) \\
Y_i &= \beta_1 + \beta_2 X_i + \beta_3 X_i^3 + \beta_4 X_i^4 + u_i + u_i \quad (c)
\end{aligned}
$$

  which is more likely be the right model?
- Use Durbin–Watson d statistic to detect model specification errors.


$\hat{u}_i$
Residuals
0
$X$
Output
$(a)$
$(b)$
$(c)$

# Specification errors
## ↪ Omitting a relevant variable

- Use Durbin–Watson d statistic to detect model specification errors.
  - Run the assumed model and obtain OLS residuals.
  - You want check if a variable $Z$ was omitted in the previous model, order the previous residual according to increasing values of $Z$.
  - Compute the d statistic with the ordered residuals in the previous step as

$$d = \frac{\sum_{t=2}^{n} \left(\hat{u}_t - \hat{u}_{t-1}\right)^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

  - The decision rule: if $d$ is significant, then there is model misspecification (omitting $Z$).

# Specification errors

## ↛ Including an unnecessary or irrelevant variable

- Suppose the true model is

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

but you fit the following model Instead

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \nu_i$$

- The model is **overfitted**.
- The consequences:
  - The OLS will be unbiased and consistent, $E(\hat{\alpha}_1) = \beta_1$, $E(\hat{\alpha}_2) = \beta_2$ and $E(\hat{\alpha}_3) = 0$
  - $\hat{\sigma}^2$ incorrectly estimates $\sigma^2$ (same as previous)
  - Confidence interval, and hypothesis testings are still **valid**.
  - $var(\hat{\alpha})$ is usually greater than $var(\hat{\beta})$. **See section 13A.2**

- Detecting overfitting
  - **Bottom-up approach**: start with a simple model and expand it until you find it overfitted.
  - **Data mining**: **take home read pp. 475-476**

# Specification errors
## $\hookrightarrow$ Regression specification error test (REST)

- Let the model to be

$$Y_i = \lambda_1 + \lambda_2 X_i + u_i$$

- The idea: If the model is correctly specified, the residuals $(\hat{u}_i)$ should be uncorrelated with $\hat{Y}_i$
- The testing procedure
  - Obtain the fitted value $\hat{Y}_i$ and $R^2_{old}$
  - Run the auxiliary regression

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + u_i$$

    and obtain $R^2_{new}$. (**Note**: we have $Y_i^2$ and $Y_i^3$ as two new regressors.)
  - Carry out the $F$ test

$$F_{obs} = \frac{(R^2_{new} - R^2_{old})/\{\text{no. of new regressors}\}}{(1 - R^2_{new})/\{n - \text{no. parameters in the new model}\}}$$

  - The decision rule: if $F$ statistic is significant, the model is misspecified.

# Specification errors
## ↪ Lagrange Multiplier Test

- Take home read **pp. 481-482** together with **pp. 249-250**.

# Model selection criteria

## ↪ The $R^2$ criterion

- $R^2$ measure **in-sample** (forecasting data is same as data used for modeling) fitting.
- Good in-sample fitting does not necessarily mean **out-of sample fitting** (forecasting data is different from the data used for modeling).
- Compare with two or more $R^2$, the regressand (response) must be the same.
- $R^2$ will grow when more variables are used in the model. Use adjusted $R^2$ instead.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

which adds penalty to the original $R^2$.

# Model selection criteria
## ↪ The Akaike's Information Criterion (AIC)

- The AIC also penalizes the residual sum squared

$$\mathrm{AIC} = \exp(2k/n)\frac{\sum \hat{u}_i^2}{n} = \exp(2k/n)\frac{\mathrm{RSS}}{n}$$

  where $k$ is the number of regressors (including intercept) and $2k/n$ is the penalty factor.
- AIC can be used in both **nested models** (Model A is nested in Model B when Model A is a special case of model B) and unrested models.
- AIC can also be used for out-of-sample forecasting performance.
- AIC can tell nothing about the quality of the model in an absolute sense. If all the candidate models fit poorly, AIC will not give any warning of that.
- **Implement this in R**

# Model selection criteria
## ↛ The CP criterion

- The CP criterion is defined as follows

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p)$$

- Notice that $E(C_p) = p$ because that $E(RSS) = (n - p)\sigma^2$.
- Compare two models with CP. Model with $C_p$ closed to $p$ should be preferred.

# Take home questions

- **13.2, 13.3, 13.11, 13.19, 13.20,**
- Read topic based on out-of-sample model comparison criterion.