Specification and Data Issues: Part I

BS1802 Statistics and Econometrics

Jiahua Wu

382 Business School
j.wu@imperial.ac.uk

Roadmap

- Regression analysis with cross-sectional data
  - The multiple regression analysis
    - Basics: estimation, inference, analysis with dummy variables
    - More technically involved: asymptotics, heteroskedasticity, specification and data issues
- Advanced topics
  - Limited dependent variable models
  - Panel data analysis
  - Regression analysis with time series data

Outline (Wooldridge, Ch. 6.2 - 6.4, 9.1)

- Functional form
- Goodness-of-fit and variable selections
- Prediction

Outline

- Functional form
- Goodness-of-fit and variable selections
- Prediction

Functional Forms

- OLS can be used for relationships that are not strictly linear in x and y by taking into account nonlinear functions of x and y
- Three common functional forms
  - Logarithmic form
  - Quadratic form
  - Interaction terms

Log Form: Interpretation of Log Models

- If the model is

$$\log(y) = \beta_0 + \beta_1 \log(x) + u,$$

β_1 is approximately the percentage change in y given 1 percent increase in x

- If the model is

$$\log(y) = \beta_0 + \beta_1 x + u,$$

$100\beta_1$ is approximately the percentage change in y given 1 unit increase in x

- If the model is

$$y = \beta_0 + \beta_1 \log(x) + u,$$

$\beta_1/100$ is approximately the unit change in y given 1 percent increase in x

Log Form: Why Use Log Models?

- Log models are invariant to the scale of variables since measuring percent changes
- For models with $y > 0$, the conditional distribution is often heteroskedastic or skewed, while taking the log can mitigate, if not eliminate, both problems
- Taking the log of a variable often narrows it range, limiting the effect of outliers

Log Form: Some Rules of Thumb

- What types of variables are often used in log form?
  - Dollar amounts (wages, salaries, firm sales and firm market value) that must be positive
  - Very large variables (population, total number of employees, and school enrollment)
- What type of variables are often used in level form?
  - Variables measured in years (education, experience, term of employment, and age)
  - Variables that are proportions (the unemployment rate, the participation rate in a pension plan)

Quadratic Form

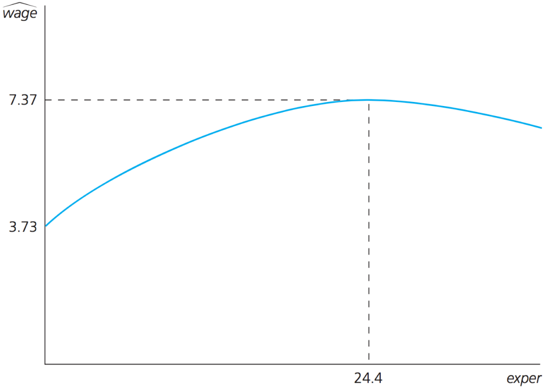
- For a model
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u,$$
we cannot interpret β_1 alone as measuring the change in y with respect to x
- We need to take into account β_2 as well, as
$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \quad \text{so } \frac{\Delta\hat{y}}{\Delta x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$$
- For $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$,
  - y is increasing in x at first, but will eventually turn around and be decreasing in x
  - the turning point will be at $x^* = |\hat{\beta}_1/(2\hat{\beta}_2)|$
- How about $\hat{\beta}_1 < 0$ and $\hat{\beta}_2 > 0$?

Quadratic Form

- Eg. Wage model (wage1.RData)

$$\widehat{wage} = 3.73 + .298exper - .0061exper^2$$

As *exper* increases, *wage* is predicted to go up, when *exper* is less than 24.4, and go down afterwards.



Interaction Terms

- For a model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u,$$

  we cannot interpret β_1 alone as measuring the change in y with respect to x_1

- We need to take into account β_3 as well, as

$$\Delta y = (\beta_1 + \beta_3 x_2)\Delta x_1$$

Interaction Terms: An Example

- Eg. House price (hprice1.RData)

  $$\textit{price} = \beta_0 + \beta_1 \textit{sqrft} + \beta_2 \textit{bdrms} + \beta_3 \textit{sqrft} \cdot \textit{bdrms} + u$$

  where

  - *price*: house price
  - *sqrft*: square footage
  - *bdrms*: number of bedrooms

- If $\beta_3 > 0$, it implies that an additional bedroom yields a higher increase in housing price for larger houses.

Functional Form Misspecification

- A regression is misspecified when its functional form is incorrect and fails to properly account for the relation between the dependent variable and observable explanatory variables
- Functional form misspecification generally causes bias in estimating parameters
- Eg. Suppose the true model is

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u.$$

Omitting $exper^2$ leads to biased estimation in

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + v,$$

as it misspecifies how *exper* affects $\log(wage)$

Functional Form Misspecification

- How do we know if we have gotten the right functional form of our model?
  - Use theory or common sense to guide you - think about the interpretation
    - Does it make more sense for x to affect y in percentage (use logs) or absolute terms?
    - Does it make more sense for the derivative of x_1 to vary with x_1 (quadratic) or with x_2 (interactions) or to be fixed?
  - If the misspecification is caused by omitting a (nonlinear) function of the regressors, we have tests for that.

REgression Specification Error Test (RESET)

- Key idea: when the model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ is correct (i.e., satisfies ZCM), no functions of x's should be significant when added to the model
- Similar to the White test, the squared and cubed fitted values, which are functions of x's, should be insignificant when added to the correct model
- Procedure of RESET
  1. OLS original model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ and save the fitted values $\hat{y}$
  2. Test $H_0 : \delta_1 = 0, \delta_2 = 0$ in the expanded model
     $$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error.$$
     The F stat follows $F_{2,n-k-3}$ distribution under the null
  3. Reject H_0 when F stat $> c$ ($F_{2,n-k-3}$ critical value)

REgression Specification Error Test (RESET)

- Example 9.2. Consider the two models

$$price = \beta_0 + \beta_1 lotsize + \beta_1 sqrft + \beta_3 bdrms + u$$

and

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + v,$$

$n = 88$

  - For the *price* model, the RESET F stat is 4.67 ($F_{2,82}$ p-value $.012$)
  - For the $\log(price)$ model, the RESET F stat is 2.56 ($F_{2,82}$ p-value $.084$)
  - The log-log model is preferred

- Note: It is possible that RESET rejects both or neither

Tests against Nonnested Models

- Nested vs. nonnested
  - "Model A nests Model B" = "Model B is a restricted version of Model A"
  - Nested models can be tested using exclusion F test
  - For nonnested models, the usual exclusion test is not applicable
- Two tests for nonnested models
  - Test exclusions within a comprehensive model that nests both of the two nonnested models
  - Davidson-MacKinnon test: use $\hat{y}$ from one model as a regressor in the second model and test for its significance

Comprehensive Model Approach

- Eg. We are unsure whether x variables should be in log or not. Thus, we have two competing models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \qquad \text{and}$$

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u$$

  - The comprehensive model is

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log(x_1) + \gamma_4 \log(x_2) + u$$

  - The acceptance of "$H_0 : \gamma_1 = 0, \gamma_2 = 0$" supports the second model
  - The acceptance of "$H_0 : \gamma_3 = 0, \gamma_4 = 0$" supports the first model

Davidson-MacKinnon Test

- Eg. Still the two competing models are

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \qquad \text{and}$$

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u.$$

  - Suppose the fitted values of the two models are g and h, respectively.
  - If the first model is correct, then h should be insignificant in

    $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta h + u$$

    Rejecting "$H_0 : \theta = 0$" is a rejection of the first model
  - Similarly, we can test the significance of g in the second model

- Note: A clear winner may not emerge

Davidson-MacKinnon Test: An Example

- Eg. Consider the two competing models (hprice1.RData)

$$\log(price) = \beta_0 + \beta_1 lotsize + \beta_1 sqrft + \beta_3 bdrms + u$$

and

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_1 \log(sqrft) + \beta_3 bdrms + u.$$

  - Suppose the fitted values of the two models are g and h, respectively.
  - For g in the log-log model, the t stat is 0.77 (p-value $.444$)
  - For h in the log-level model, the t stat is 2.34 (p-value $.022$)
  - The log-log model is preferred

Outline

- Functional form
- Goodness-of-fit and variable selections
- Prediction

Goodness-of-Fit: Adjusted R-Squared

- R^2 is the proportion of variation in y that is explained by x's - a measure of goodness-of-fit
  - It is tempting to compare models with different regressors by using R^2
  - But R^2 always increases as more regressors are added to the model
  - To compare different models, we need to take into account the model size (number of regressors)

Goodness-of-Fit: Adjusted R-Squared

- $R^2 = 1 - SSR/SST$
- The *df* in SSR is $n - k - 1$. The *df* in SST is $n - 1$.
- A fair measure is based on the sums of squares, adjusted for the degrees of freedom

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

  known as the adjusted R-squared, which is also routinely reported in OLS output
  - You can compare the fit of 2 models (with the same y) by comparing the adj-R^2
  - You cannot use the adj-R^2 to compare models where y are in different function forms

Goodness-of-Fit: Information Criteria

- Akaike Information Criteria (AIC) in selecting a model tries to balance the conflicting demand of accuracy (fit) and simplicity (small number of variables)

$$AIC = n \ln(SSR/n) + 2k$$

- AIC for a single model is not very meaningful - mainly used to rank multiple models
  - Models with smaller AIC are preferred
  - Rule of thumb: Models with AIC not differing by 2 should be treated as equally adequate. Larger differences in AIC indicate significant differences between the quality of models

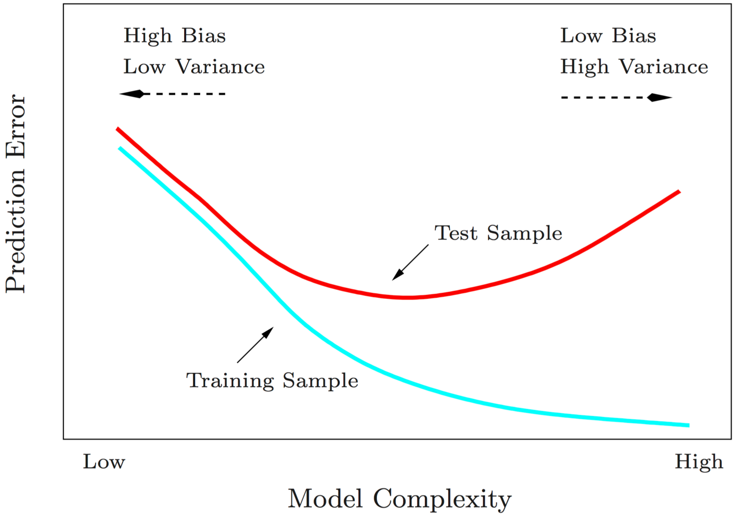
Goodness-of-Fit: Information Criteria

- Several modifications of AIC have been suggested
- One popular variation is Bayes Information Criterion (BIC)

$$BIC = n\ln(SSR/n) + k\ln(n)$$

- Difference between AIC and BIC is in the severity of penalty for k
  - The penalty is far more severe in BIC when $n > 8$
  - Tends to control the overfitting tendency of AIC

Bias-Variance Tradeoff



Goodness-of-Fit: Information Criteria

- Another modification of AIC to avoid overfitting is AIC$_c$

$$AIC_c = AIC + \frac{2(k+2)(k+3)}{n-k-3}$$

- Typically used for small samples
  - Correction to AIC is small for large n and moderate k
  - Correction is large when n is small and k is large

Variable Selection

- When the number of variables is small
  - We can evaluate all possible equations
  - The total number of equations fitted is 2^k with k variables
  - R function: `regsubsets()` in the library `leaps`
- When the number of variables is large
  - Forward- and backward-stepwise selection
  - With k variables these procedures will involve evaluation of at most $k+1$ equations
  - R function: `step()`
- An example with `bwght.RData`

Outline

- Functional form
- Goodness-of-fit and variable selections
- Prediction

Confidence Intervals for Predictions

- Suppose we have an estimated model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k,$$

  and we want an estimate of

$$\theta_0 = E(y|x_1 = c_1, \ldots, x_k = c_k) = \beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k$$

- This is easy to obtain by substituting the x's in our estimated model with c's, i.e.,

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k c_k$$

- What about a confidence interval of $\hat{\theta}_0$?
  - We need to know the standard error of $\hat{\theta}_0$
  - Follow the same approach when we test a linear combination of OLS estimators (see Inference slide deck)

Confidence Intervals for Predictions

- We can write $\beta_0 = \theta_0 - \beta_1 c_1 - \cdots - \beta_k c_k$
- Plug it into the model to obtain

$$y = \theta_0 + \beta_1(x_1 - c_1) + \cdots + \beta_k(x_k - c_k) + u$$

- The OLS estimator of θ_0 and its standard error are the intercept and its standard error in the regression of y_i on $(x_{i1} - c_1), \ldots, (x_{ik} - c_k)$
- Eg. The wage model: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$
  - What is the expected wage of an average person with $educ = 12$, $exper = 8$?
  - Regression results are

$$\widehat{wage} = \underset{(.18)}{4.90} + \underset{(.05)}{.64}\,(educ - 12) + \underset{(.01)}{.07}\,(exper - 8)$$

  - The 95% interval prediction $\approx 4.90 \pm 1.96 \cdot (.18) = [4.55, 5.25]$

Confidence Intervals for Predictions

- What if we want to predict y rather than $E(y|x)$?
  - The standard error for the average value of y is not the same as a standard error for a particular outcome of y
  - We must account for another very important source of variation: the variance in the unobserved error
  - Let the prediction error be $\hat{e}$. The standard error of $\hat{e}$ is given by $se(\hat{e}) = [se(\hat{\theta}_0)^2 + \hat{\sigma}^2]^{1/2}$
  - The 95% interval prediction (for large sample) is given by

$$\hat{\theta}_0 \pm 1.96 \cdot [se(\hat{\theta}_0)^2 + \hat{\sigma}^2]^{1/2}$$

Predicting y in a Log Model

- Model: $logy \equiv \log(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
- What is the predicted value $\hat{y}$?
  - $\hat{y} = \exp(\widehat{logy})$?
  - Need to scale this up by an estimate of the expected value of $\exp(u)$
  - Can use $n^{-1} \sum_{i=1}^{n} \exp(\hat{u}_i)$ as a sample estimate of $E(\exp(u))$, and thus

$$\hat{y} = n^{-1} \sum_{i=1}^{n} \exp(\hat{u}_i) \exp(\widehat{logy})$$