Specification and Data Issues: Part I

Statistics and Econometrics

Jiahua Wu

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

Roadmap

- Regression analysis with cross-sectional data
  - Basics: estimation, inference, analysis with dummy variables
  - More involved: model specification and data issues
- Advanced topics
  - Binary dependent variable models
  - Panel data analysis
  - Time series analysis

Outline (Wooldridge, Chap. 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 to account for nonlinear functions of x and y
- Two common functional forms
  - Logarithmic form
  - Quadratic form

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

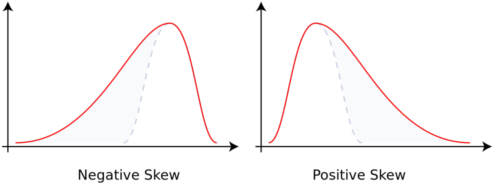
- If the model is

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

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

Log Form: When to Use Log Models?

- We use log transformation
  - when the distribution of residuals is skewed or heteroskedastic



  - for model interpretation (i.e., percentage change)
  - for multiplicative models
    - Eg., Cobb-Douglas production function: $Y = AL^{\beta}K^{\alpha}$, where $Y =$ total production, $L =$ labor input and $K =$ capital input
    - We can take log and estimate

$$\log(Y) = \log(A) + \beta \log(L) + \alpha \log(K) + u$$

Quadratic Form

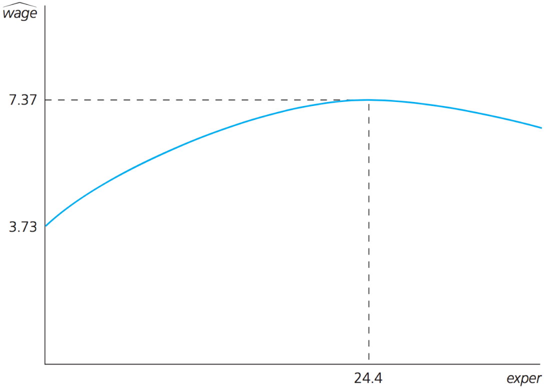
- For a model
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u,$$
we cannot interpret β_1 alone as measuring the change in y with respect to x
- We need to take into account β_2 as well, as
$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \quad \text{so } \frac{\Delta\hat{y}}{\Delta x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$$
- 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.



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 independent variables
  - Consequence: 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 to be fixed?
  - If the misspecification is caused by omitting a (nonlinear) function of independent variables, 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, 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. (hprice1.RData) Consider the model

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

$n = 88$

  - The RESET F stat is 4.67 ($F_{2,82}$ p-value $.012$)
  - A drawback with RESET: Provides no real direction on how to proceed!

- Remark: RESET has no power detecting omitted variables or heteroskedasticity

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 independent variables by using R^2
  - But R^2 always increases as more independent variables are added to the model
  - To compare different models, we need to take into account the model size (number of independent variables)

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

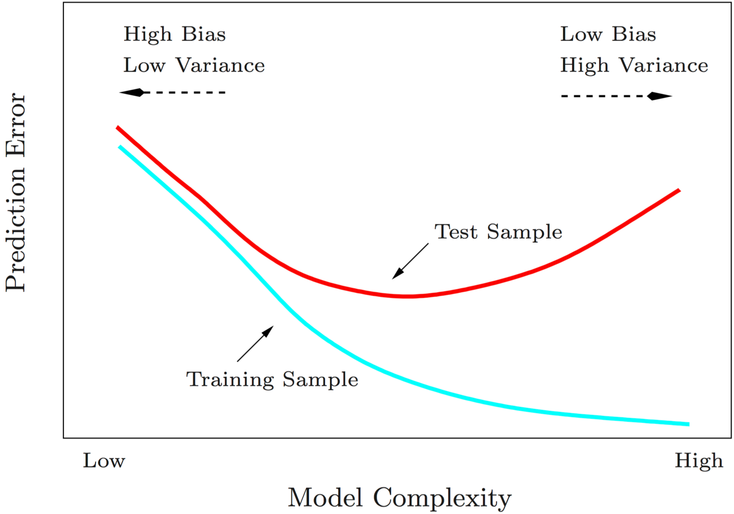
Goodness-of-Fit: Information Criteria

- 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



Source: The Elements of Statistical Learning: Data mining, inference and prediction by Hastie et al.

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 $\hat{\theta}_0$

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.(wage1.RData) $wage = \beta_0 + \beta_1 educ + u$
  - What is the expected wage of an average person with $educ = 12$?
  - Regression results are

$$\widehat{wage} = \underset{(.15)}{5.59} + \underset{(.05)}{.54}(educ - 12)$$

  - The 95% interval prediction $\approx 5.59 \pm 1.96 \cdot (.15) = [5.30, 5.89]$

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

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