Notes on Regression Analysis

November 18, 2013

Bivariate Regression

$$y = \beta_0 + \beta_1 x + u$$

- The coefficient estimates are given by

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$b_1 = \frac{s_{xy}^2}{s_x^2} = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}.$$

- Predicted values are

$$\hat{y} = b_0 + b_1 x,$$

  which is also the regression line.

- Residuals are given by the difference between actual observation and predicted value,

$$e = y - \hat{y}$$

- Residuals are useful for postregression analysis, e.g. residual plots and computing the R^2,

$$R^2 = 1 - \frac{Var(e)}{Var(y)}$$

- Standard errors: How close are the estimated coefficients to the true betas?
  - Run hypothesis tests:

$$z = \frac{obs - exp}{SE}$$

  - Test the null of $\beta = 0$
  - Think of a confidence interval around b: Does it include 0? Then we cannot reject the null that $\beta = 0$.

- Basics of interpreting coefficients

$$y = a + bx + u$$

- y level - x level: Increasing x by one unit changes y by b units

$$y = a + blog(x) + u$$

- y level - x log: Increasing x by one percent changes y by $b/100$ units

$$log(y) = a + bx + u$$

- y log - x level: Increasing x by one unit changes y by $100 * b$ percent

$$log(y) = a + blog(x) + u$$

- y log - x log: Increasing x by one percent changes y by b percent

Multiple Regression and Model building

- Bias-Variance Tradeoff:
  - Endogeneity bias from omitted variables
  - Imprecise SE from including irrelevant variables
- Think of theoretical arguments which regressors you want to include. Keep them if they are practically important (even if they are not statistically significant).
- Dummies and interaction effects
  - A dummy is used to get separate intercepts for each group, i.e. we shift the regression line.
    - The actual intercept is the constant plus the coefficient on the group dummy.
    - The coefficient on the group dummy measures the difference between this group and the omitted baseline group.
  - Dummy variable trap
    - If you include variables that are collinear (linear combinations), then we cannot tell apart which variable caused the change in the outcome.
    - Example dummy variables for seasons...
  - Interaction coefficient tests an additional effect of satisfying both criteria as opposed to just one or the other.
  - The overall intercept for
- Joint hypothesis testing
  - Can test more complicated hypotheses: $\beta_1 = 0$ AND $\beta_2 = 0$

    - The main problem is that the estimates for different coefficients are not independent. Instead they will be correlated.
    - Compare the fit of two different models: Restricted (under the null hypothesis) and unrestricted
    - Stata command "test" reports F-statistic

$$F = \frac{\left(R^2 - R_*^2\right)/J}{\left(1 - R^2\right)/\left(n - k\right)}$$

  - Can also compare effects of different variables, e.g. $\beta_1 = \beta_3$ etc.

Common Problems with Running Regressions

- Basic requirements for regression to work:

$$E\left(u|x\right) = 0$$
$$Var(u|x) = \sigma_u^2$$

  - The average across error terms for a given value of x is zero.
  - The variation in error terms for a given value of x is the same as for another value of x.

- Endogeneity: There are some omitted variables that are correlated with both regressors on the RHS and the LHS variable

$$E\left(u|x\right) \neq 0$$

  - Example 1: Shoe purchases, ad exposure and preferences for shoes.
  - Example 2: Demand curve, prices, advertising
  - We can sign the direction of the bias generated by omitted variables if we know how they affect the RHS and LHS variable!
  - How do experiments help to avoid this problem?
    - The key RHS variable is a dummy whether a person receives treatment. Think about the example of seeing an ad. If we could randomly select people who see the ad, then there will be no correlation between ad exposure and shoe preferences.
    - How is a natural experiment different from a lab experiment?

- Multicollinearity

  - If some of the independent variables are highly correlated but both have a direct effect on the dependent variable, then the coefficients will be estimated with more noise because it's harder to attribute the effect to one or the other variable given that the move together most of the time.
  - It is important to include both variables though. Otherwise there might be omitted variable bias since the variables are correlated and both affect the dependent variable.

- Heteroskedasticity: The error variance differs across different values of x

  - The problem is that the standard errors that we compute are wrong.
  - More on this and other issues in advanced courses.