Announcements

- Midterm is Thursday, February 24 in class
- Midterm 2 covers chapters 5 through 8, lectures 1-20-11 through 2-10-11
- Don't forget a scantron sheet and a calculator
- Office hours this week: today 2pm-5pm, tomorrow 9am-noon

A Quick Review for the Midterm

A very broad outline of the midterm topics:

Graphical Representations of Bivariate Data

- Scatterplots
- Line graphs with multiple time series on them
- Residual plots

A Quick Review for the Midterm

Descriptive Statistics for Bivariate Data

- Covariance
- Correlation
- Regression results
- Goodness of fit

A Quick Review for the Midterm

Statistical Inference

- Population assumptions
- Distribution of slope coefficient and intercept
- Hypothesis testing for the slope coefficient and intercept
- Confidence intervals
- Statistical vs economic significance

A Quick Review for the Midterm

Prediction

- How to predict the actual value of y and the expected value of y
- Standard errors of these predictions
- What influences those standard errors

A Quick Review for the Midterm

Bivariate Data Transformation

- When to use logs
- Interpreting coefficients for log-log, linear-log, log-linear
- Polynomials
- Dummy variables

A Quick Review for the Midterm

Problems With Bivariate Regression

- Badly behaved residuals
- Sample selection bias
- Incorrect interpretation of coefficients (omitted variables, correlation vs. causality)

Quick Review of Multivariate Hypothesis Testing

Hypothesis testing for a single regressor:

$$H_o: \ \beta_j = \beta_j^*$$

$$H_a: \ \beta_j \neq \beta_j^*$$

$$t^* = \frac{b_j - \beta_j^*}{s_{b_j}}$$

$$p = Pr(T_{n-k} > t^*) = TDIST(|t^*|, n-k, 2)$$

$$c = t_{\frac{\alpha}{2}, n-k} = TINV(\alpha, n-k)$$

- Reject null hypothesis if $p < \alpha$ or $|t^*| > c$
- Can also do one-sided hypothesis tests

Quick Review of Multivariate Hypothesis Testing

Testing overall significance:

$$H_o: \ \beta_2 = 0, \beta_3 = 0, ..., \beta_k = 0$$

$$H_a: \ \text{at least one of } \beta_2, ..., \beta_k \neq 0$$

$$F^* = \frac{R^2}{1 - R^2} \frac{n - k}{k - 1}$$

$$p = Pr(F_{k-1,n-k} > F^*) = FDIST(F^*, k - 1, n - k)$$

$$c = F_{\alpha,k-1,n-k} = FINV(\alpha, k - 1, n - k)$$

- Reject null hypothesis if $p < \alpha$ or $F^* > c$

Testing the Significance of a Subset of Regressors

- Sometimes we don't want to test the overall significance of a regression, instead we want to test the significance of a particular subset of regressors
- For example, suppose we had a wage regression with lots of information on education, demographics, etc.
- We might be interested in testing whether including information on an individual's parents can improve our model
- Our hypotheses in this case are:

$$H_o: \ \beta_{g+1} = 0, ..., \beta_k = 0$$

$$H_a: \ \text{at least one of } \beta_{g+1}, ..., \beta_k \neq 0$$

Testing the Significance of a Subset of Regressors

- We call our model with all of the regressors in it the **unrestricted model**:

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_g x_g + \beta_{g+1} x_{g+1} + ... + \beta_k x_k + \varepsilon$$

- We call our model without the subset of regressors we are interested in the **restricted model**:

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_g x_g + \varepsilon$$

- We basically want to test whether the fit is significantly better for the unrestricted model compared to the restricted model

Testing the Significance of a Subset of Regressors

- To do that, we use the following test statistic:

$$F^* = \frac{ESS_r - ESS_u}{ESS_u} \frac{n-k}{k-g}$$

  where ESS_r is the error sum of squares for the restricted model and ESS_u is the error sum of squares for the unrestricted model

- We can also write this test statistic in terms of the R^2 of the two models:

$$F^* = \frac{R_u^2 - R_r^2}{1 - R_u^2} \frac{n-k}{k-g}$$

- Either way, it is clear that F^* is larger when the improvement in fit switching from the restricted to unrestricted model is bigger

Testing the Significance of a Subset of Regressors

- The test statistic is distributed according to an F distribution with $k - g$ and $n - k$ degrees of freedom
- To test the hypothesis, we can take either the p-value approach ($p = Pr(F_{k-g,n-k} > F^*)$) or the critical value approach ($c = F_{\alpha,k-g,n-k}$)
- If p is less than α or if F^* is greater than c, we will reject the null hypothesis
- Just like with overall significance, we can calculate p in Excel with FDIST() and c with FINV() only now we use $k - g$ instead of $k - 1$
- To Excel and some data on prisoners (prison-data.csv)...

Multivariate Data Transformation

- Just as with bivariate data, sometimes we will need to use data transformations with multivariate data
- We can use all of the transformations we have already talked about:
  - Taking the natural log of the dependent variable
  - Taking the natural log of the regressors
  - Using polynomials for particular regressors
- We also have a couple of new possibilities
  - Multiple dummy variables
  - Interaction terms

Logs and Multivariate Data

- We use logs with multivariate data for the same reasons as with bivariate data
- Changes in logs can be interpreted as percent changes (eg. elasticities)
- Logs help us deal with a variable for which different observations are on very different scales (eg. population, income)
- Logs can capture exponential growth (with log-linear models)
- It may make sense to take logs of just some variables or to take logs of all variables

A Classic Example of a Multivariate Log-log Model

- Consider the widely used Cobb-Douglas production function:

$$y = AK^{\alpha}L^{\beta}$$

- Suppose we want to get estimates of A, α and β using ordinary least squares
- We need to transform this into a linear model:

$$\ln y = \ln(AK^{\alpha}L^{\beta})$$

$$\ln y = \ln A + \ln K^{\alpha} + \ln L^{\beta}$$

$$\ln y = \ln A + \alpha \ln K + \beta \ln L$$

- So if we regress $\ln y$ on $\ln K$ and $\ln L$, the intercept will give us an estimate of $\ln A$ and the coefficients will give us estimates of α and β

Polynomials and Multivariate Data

- Polynomials offer a very flexible way to fit nonlinear trends
- Recall the example of income and age (the U-shaped curve meant we should use a quadratic in age):

$$\ln wage_i = \beta_1 + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 edu_i + \varepsilon_i$$

- If we think that there is a nonlinear relationship between y and a particular regressor x_j, we should consider including a polynomial in x_j in our regression $(x_j, x_j^2, x_j^3, ...)$

Dummy Variables and Multivariate Data

- We may want to use dummy variables to include categorical data in our regressions
- Recall that a dummy variable is either zero or one depending on the value of a particular categorical variable (eg. male equals one, female equals zero)
- When we considered categorical variables with more than two values, we split the values into two groups so that we could use a binary dummy variable
- If we are willing to use several regressors, we have another option available to us: multiple dummy variables

Using Multiple Dummy Variables

- Suppose we have a categorical variable for education (*edu*) that can take on any of the following values: some high school, high school graduate, some college, college graduate
- To include this variable in our regression, we can use several dummy variables
- Each dummy variable still needs to be either zero or one, for example the dummy variable for 'some high school' would be defined as:

$$d_{somehs} = 1 \text{ if } edu = \text{“some HS”}, 0 \text{ otherwise}$$

- We could define a dummy variable this way for each educational cateogory: $d_{somehs}, d_{hsgrad}, d_{somecol}, d_{colgrad}$

Using Multiple Dummy Variables

edu	d(somehs)	d(hsgrad)	d(somecol)	d(colgrad)
some college	0	0	1	0
high school graduate	0	1	0	0
college graduate	0	0	0	1
high school graduate	0	1	0	0
some high school	1	0	0	0
some college	0	0	1	0
college graduate	0	0	0	1

Using Multiple Dummy Variables

- Including these dummy variables as regressors will allow us to estimate average differences in the dependent variable across different groups
- For example, suppose we regress books read per year on our education dummies and age:

$$books = b_1 + b_2 age + b_3 d_{somehs} + b_4 d_{hsgrad} + b_5 d_{somecol}$$

- Notice that I didn't include $d_{colgrad}$, it is very important to exclude one of the dummy variables
- Why?

Using Multiple Dummy Variables

- Suppose we had included $d_{colgrad}$:

$$books = b_1 + b_2 age + b_3 d_{somehs} + b_4 d_{hsgrad} + b_5 d_{somecol} + b_6 d_{colgrad}$$

- But $d_{colgrad} = 1 - d_{somehs} - d_{hsgrad} - d_{somecol}$, so the above equation can be rewritten as:

$$books = (b_1 + b_6) + b_2 age + (b_3 - b_6) d_{somehs}$$

$$+ (b_4 - b_6) d_{hsgrad} + (b_5 - b_6) d_{somecol}$$

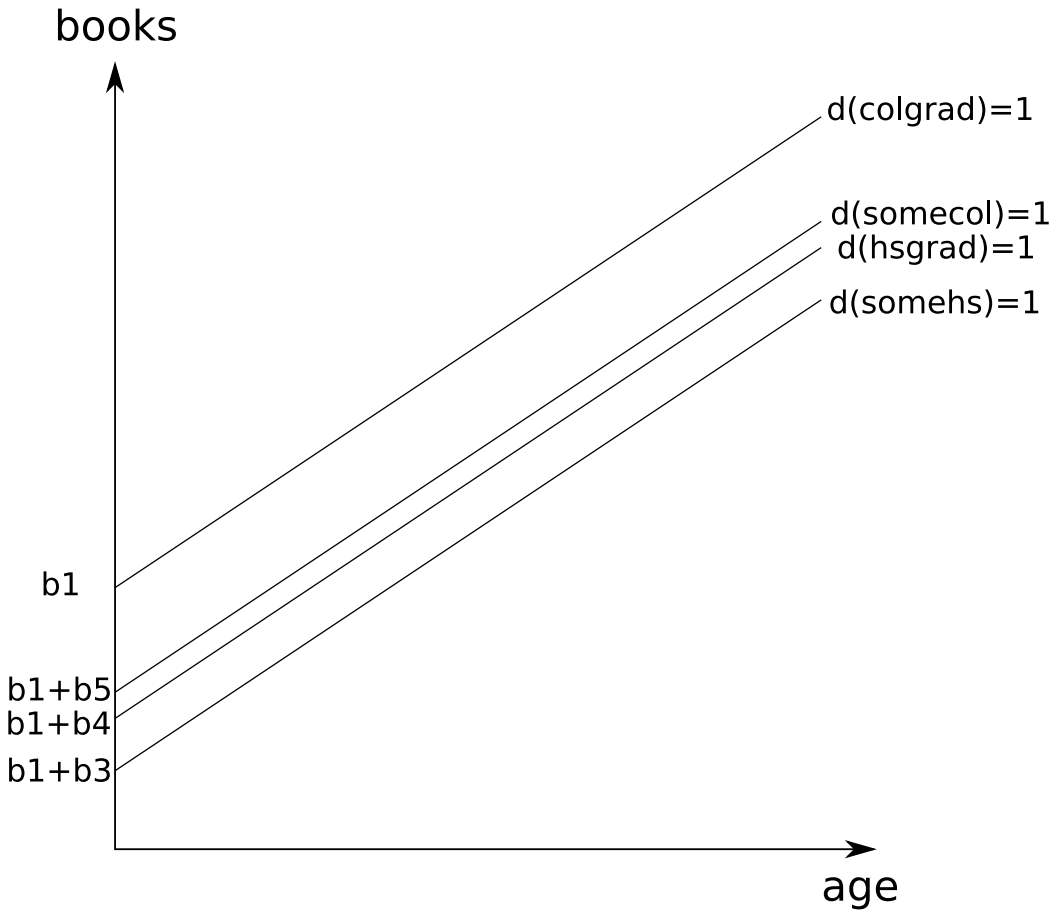
- There won't be a unique set of coefficients for the regression equation
- To get around this problem, we need to drop one of the dummy variables

Using Multiple Dummy Variables

$$books = b_1 + b_2 age + b_3 d_{somehs} + b_4 d_{hsgrad} + b_5 d_{somecol}$$

- The coefficient in front of each dummy variable can be interpreted as the difference between the average outcome for the dummy variable group and the average outcome for the omitted group
- For example, b_3 would be the difference between the average number of books read by high school dropouts and college graduates
- We can also compare coefficients to each other, $b_4 - b_3$ is the difference in the average number of books read by high school grads and high school dropouts

Using Multiple Dummy Variables



Dummy Variables and Interaction Terms

- What if the difference between educational groups isn't just the average number of books read but how that number changes with age?
- We can model a situation like this by including *interaction terms* in our regression
- An interaction term is when we multiply one variable by another and include the result as a regressor
- For example, suppose we think that college graduates may read more books as they get older while non-college grads read fewer books as they get older
- We could use the following regression to test this:

$$books = b_1 + b_2 age + b_3 d_{colgrad} + b_4 d_{colgrad} \cdot age$$

Dummy Variables and Interaction Terms

- How do we interpret the coefficients?

$$books = b_1 + b_2 age + b_3 d_{colgrad} + b_4 d_{colgrad} \cdot age$$

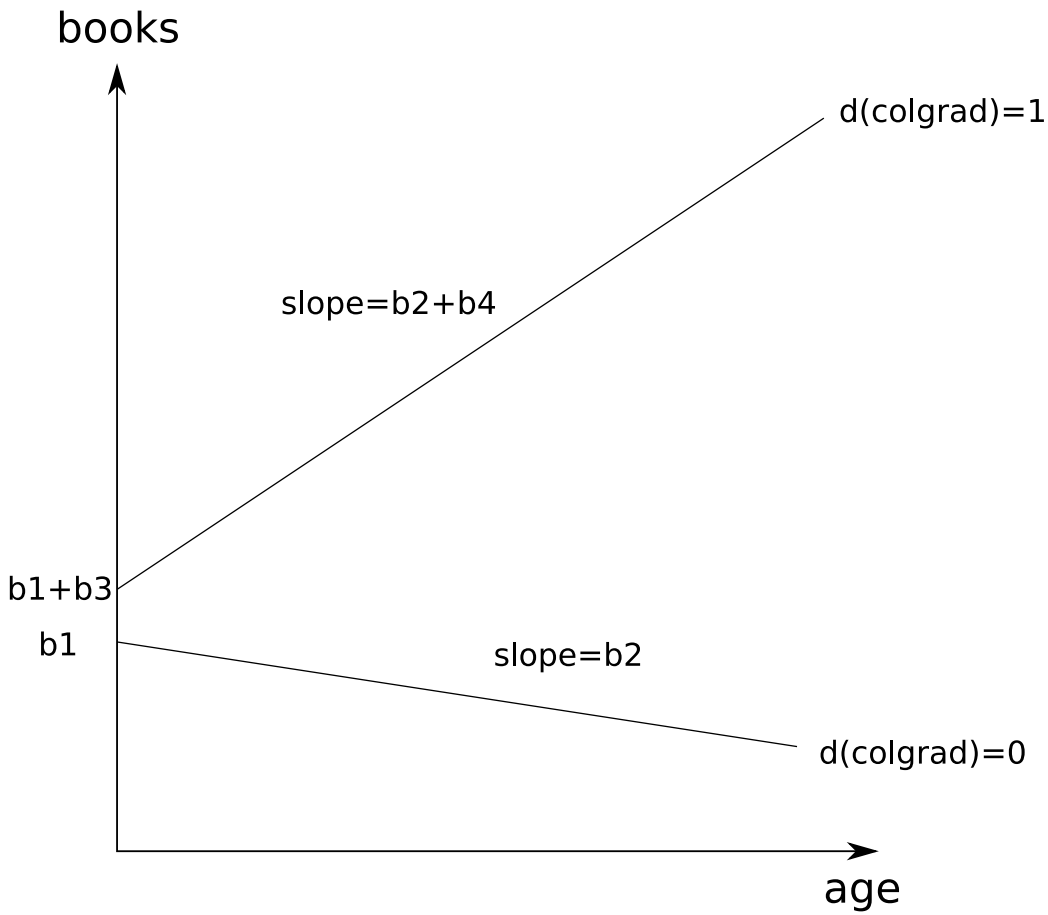
- The difference in the predicted number of books read between college grads and non-grads will depend on age now

$$\widehat{books}_{grad} = (b_1 + b_3) + (b_2 + b_4) age$$

$$\widehat{books}_{nongrad} = b_1 + b_2 age$$

$$\widehat{books}_{grad} - \widehat{books}_{nongrad} = b_3 + b_4 \cdot age$$

Dummy Variables and Interaction Terms



Interpreting Coefficients with Dummy Variables

- Suppose we wanted to investigate wage discrimination based on gender by regressing log wages on age, education, a dummy equal to one for males and and interaction between that dummy and education and get the following (standard errors are in parentheses):

$$\underset{}{\ln \text{wage}} = \underset{(.80)}{5} + \underset{(.001)}{0.03}\,\text{age} + \underset{(.002)}{0.08}\,\text{edu} + \underset{(.08)}{0.04}\,\text{male} + \underset{(.03)}{0.01}\,\text{male*edu}$$

- Notice the positive coefficient on the interaction term is not significant
- This does not mean that education has no effect on wage for males
- It means that education has no significant additional effect for males beyond the effect common to males and females