Lecture 17:
Finish Review of EXAM I Chapter 16, 1-5

6 weeks left...12 lectures.

I will cover **at least** chpt 6-11. Any spare time will be used in the lab.

Lecture on Chapter 6

**Specification:**

1. Choosing the correct independent variables
2. choosing the correct functional form
3. Choosing the correct for of the error.

Specification error occurs when an error occurs in the three steps above.
**Omitted variables**
True regression
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$
estimated
$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^*$$
where

$$\varepsilon_i^* = \varepsilon_i + \beta_2 X_{2i}$$

so
$$E\left(\hat{\beta}_0^*\right) \neq \beta_0$$

and
if $r_{12} = r_{21} = \dfrac{\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{\sqrt{\sum (x_1 - \bar{x}_1)^2 \sum (x_2 - \bar{x}_2)^2}} \neq 0$
then

$$E\left(\hat{\beta}_1^*\right) \neq \beta_1$$

Solution and identification?

**Irrelevant variables**

True regression
$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$
estimated
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i^{**}$$

where

$$\varepsilon_i^{**} = \varepsilon_i - \beta_2 X_{2i}$$

Four Important Specification Criteria
1. Theory
2. T-test
3. Rbar squared
4. Bias (do variables coefficients change significantly when variables are added)

**Specification Searches**

Data Mining
http://www.absoluteastronomy.com/topics/Testing_hypotheses_suggested_by_the_data

Stepwise regressions
http://www.stata.com/support/faqs/stat/stepwise.html

Sequential searches

Using T-tests to choose included variables

Scanning and Sensitivity analysis

So how do we choose a model?

Lecture 18: October 31

**Lagged independent variables**

Ramsey Regression Specification Error Test (RESET)

A test for misspecification and sometimes, rather mistakenly referred t as a test for omitted variables

Using OLS estimate

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \qquad \text{eq 1}$$

then generate $\hat{Y}^2{}_i, \hat{Y}^3{}_i, \hat{Y}^4{}_i$

re-estimate the original equation augmenting it with the polynomials of the fitted values.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \hat{Y}^2{}_i + \beta_4 \hat{Y}^3{}_i + \beta_5 \hat{Y}^4 + \varepsilon_i \qquad \text{eq 2}$$

$$F = \frac{(RSS_m - RSS)/M}{RSS/(n-(k+1))}$$

where RSS_m is from eq 1 and RSS is from eq 2.

Ramsey’s Regression Specification Error Test (RESET)
http://faculty.chass.ncsu.edu/garson/PA765/assumpt.htm

 Ramsey's RESET test (regression specification error test). Ramsey's general test of specification error of functional form is an F test of differences of R2 under linear versus nonlinear assumptions. It is commonly used in time series analysis to test whether power transforms need to be added to the model. For a linear model which is properly specified in functional form, nonlinear transforms of the fitted values should not be useful in predicting the dependent variable. While STATA and some packages label the RESET test as a test to see if there are "no omitted variables," it is a linearity test, not a general specification test. It tests if any nonlinear transforms of the specified independent variables have been omitted. It does not test whether other relevant linear or nonlinear variables have been omitted.

1. Run the regression to obtain Ro2, the original multiple correlation.
2. Save the predicted values (Y's).
3. Re-run the regression using power functions of the predicted values (ex., their squares and cubes) as additional independents for the Ramsey RESET test of functional form where testing that none of the independents is nonlinearly related to the dependent. Alternatively, re-run the regression using power functions of the independent variables to test them individually.
4. Obtain Rn2, the new multiple correlation.
5. Apply the F test, where F = ( Rn2 - Ro2)/[(1 - Rn2)/(n-p)], where n is sample size and p is the number of parameters in the new model.
6. Interpret F: For an adequately specified model, F should be non-significant.

Apparently some stats programs have rounding errors/computational problems that appear as multicollinearity. http://en.wikipedia.org/wiki/Multicollinearity

4) Mean-center the predictor variables. Mathematically this has no effect on the results from a regression. However, it can be useful in overcoming problems arising from rounding and other computational steps if a carefully designed computer program is not used.

But really, it shouldn't truly matter. http://www.bauer.uh.edu/jhess/papers/JMRMeanCenterPaper.pdf

But now that I do some digging I see that stata actually does this normalization as well, before taking the powers.
http://www.stata.com/statalist/archive/2004-06/msg00264.html

Akaike Information Criterion (AIC)

Minimize AIC=Log(RSS/n)+2(K+1)/n

Schwarz Criterion, or Schwarz Bayesian Criterion (SC, SBC)

Minimize SBC=Log(RSS/n)+Log(n)(K+1)/n

Lecture 19: November 4

The use and interpretation of the constant term

Don't do it.  There is an inherent identification problem, as the constant includes the true constant, means of omitted variables, and

Alternative functional forms
Linear Form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Double log form

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \varepsilon_i$$

Semi-log form

Log – Lin

$$\ln Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Lin-Log

$$Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \varepsilon_i$$

Polynomial functional form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

Inverse functional Form

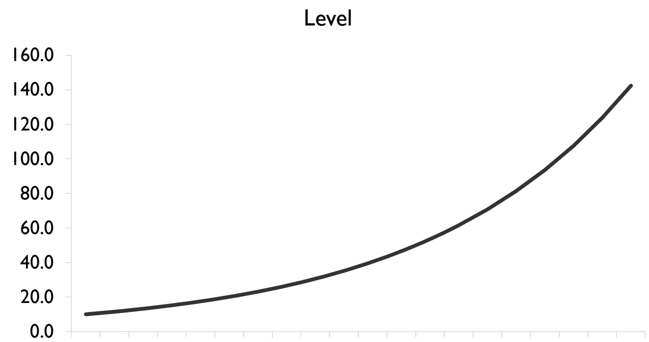
$$Y_i = \beta_0 + \beta_1 (1/X_{1i}) + \beta_2 X_{2i} + \varepsilon_i$$

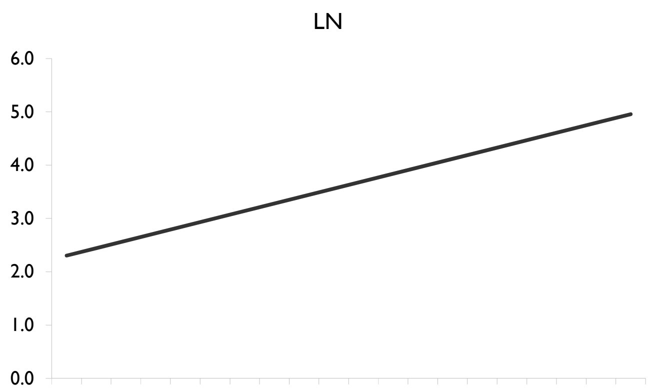
Be sure to appropriately interpret the marginal effects. Elasticities, percentage changes etc.

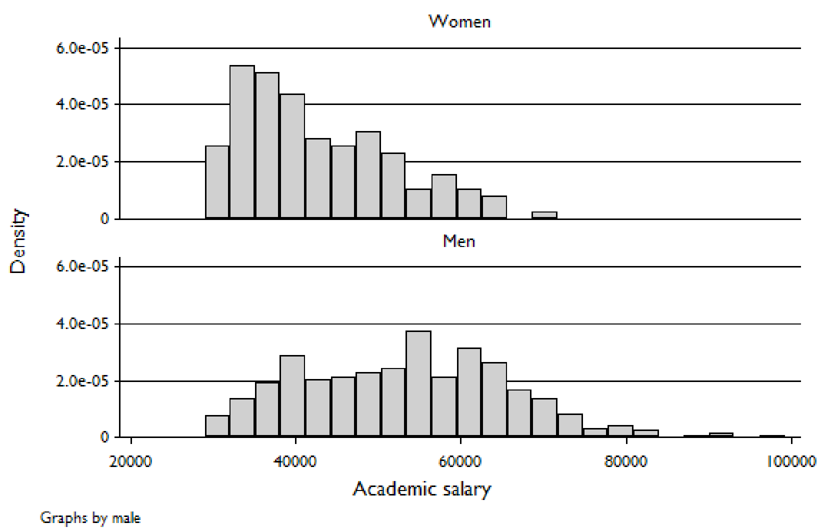
Never take the log of a dummy variable. Almost always take the log of a dollar value.

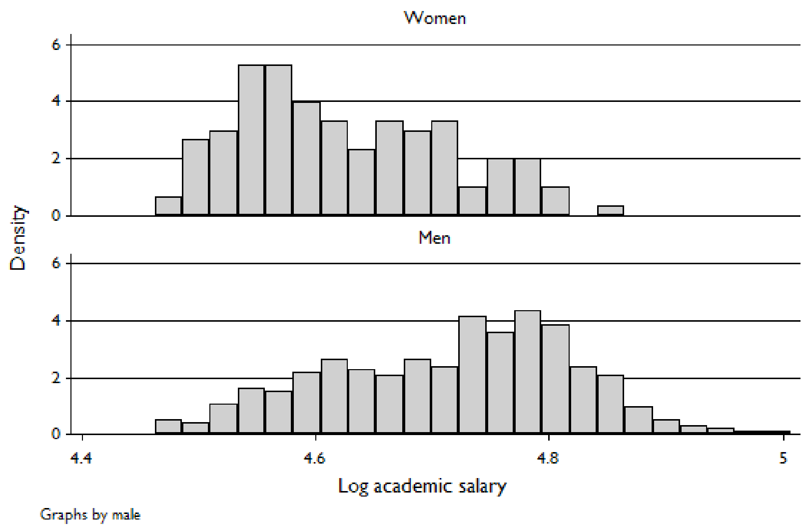
Problems with incorrect functional form.

Some pictures of alternative forms.


Level
160.0
140.0
120.0
100.0
80.0
60.0
40.0
20.0
0.0


LN
6.0
5.0
4.0
3.0
2.0
1.0
0.0


Women
Men
Density
6.0e-05
4.0e-05
2.0e-05
0
20000
40000
60000
80000
100000
Academic salary
Graphs by male




Rsquared are difficult to compare when transformed

Incorrect functional forms

Estimate

Lecture 20 :November 7

Using dummy variables

Intercept dummy

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Where:

Y is salary

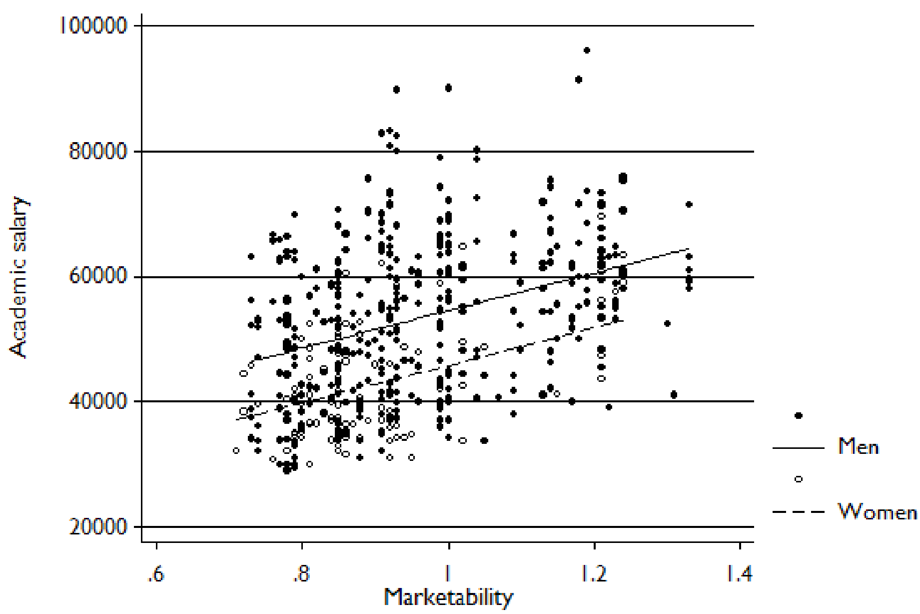
X1 is a dummy variable for male x2=1 for male, 0 for female.

X2 is marketability

```
. regress salary male marketc

      Source |       SS       df       MS              Number of obs =     514
-------------+------------------------------           F(  2,   511) =   85.80
       Model |  2.0711e+10     2  1.0356e+10           Prob > F      =  0.0000
    Residual |  6.1676e+10   511   120696838           R-squared     =  0.2514
-------------+------------------------------           Adj R-squared =  0.2485
       Total |  8.2387e+10   513   160599133           Root MSE      =   10986

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   8708.423   1139.411     7.64   0.000     6469.917    10946.93
     marketc |    29972.6   3301.766     9.08   0.000     23485.89     36459.3
       _cons |   44324.09   983.3533    45.07   0.000     42392.17       46256
------------------------------------------------------------------------------
```




As a follow up from the previous section, I re-run the regression using the log of salary as the dependent variable. Notice a few things, the R-squared is different, but remember that should not be used to decide on models as the dependent variable has a different total sum of squares. Do notice that the coefficient on male is quantitatively different. Now its interpretation is the effect of being male not on salary, but the log of salary, or the percentage change. So being male means a 7.6% increase in salary relative to females holding market constant, but not other excluded/omitted variables.

```
. regress lsalary male marketc

      Source |       SS       df       MS              Number of obs =     514
-------------+------------------------------           F(  2,   511) =   91.29
       Model |  1.5890749     2  .79453745           Prob > F      =  0.0000
    Residual | 4.44763545   511 .008703788           R-squared     =  0.2632
-------------+------------------------------           Adj R-squared =  0.2604
       Total | 6.03671035   513 .011767467           Root MSE      =  .09329

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0762761   .0096758     7.88   0.000     .0572669    .0952853
     marketc |   .2625476   .0280384     9.36   0.000      .207463    .3176323
       _cons |   4.635698   .0083506   555.14   0.000     4.619292    4.652104
------------------------------------------------------------------------------
```

Slope dummy (interaction terms)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} \varepsilon_i$$

Where:

Y is salary

X1 is a dummy variable for male x2=1 for male, 0 for female.

X2 is marketability

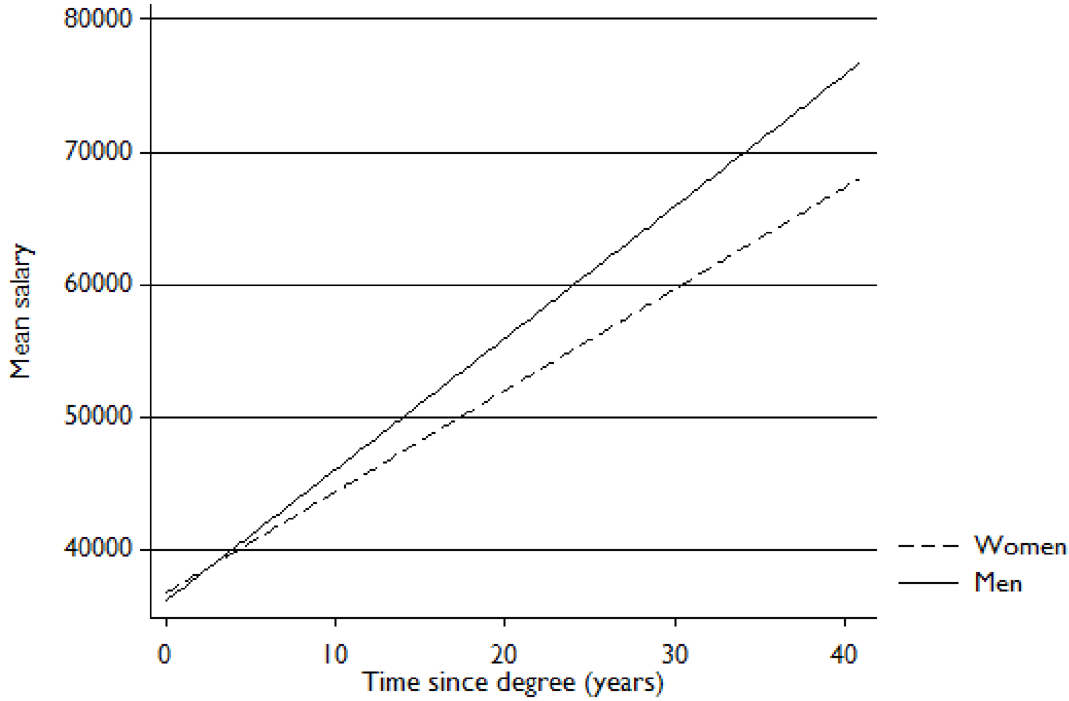
X3 is years to degree

X4 is m_years which is just X4=(X1*X3)

```
. regress salary male marketc yearsdg m_years

      Source |       SS       df       MS              Number of obs =     514
-------------+------------------------------           F(  4,   509) =  279.95
       Model |  5.6641e+10     4  1.4160e+10           Prob > F      =  0.0000
    Residual |  2.5746e+10   509  50581607.4           R-squared     =  0.6875
-------------+------------------------------           Adj R-squared =  0.6850
       Total |  8.2387e+10   513   160599133           Root MSE      =  7112.1

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -593.3088   1320.911    -0.45   0.654    -3188.418      2001.8
     marketc |   38436.65   2160.963    17.79   0.000     34191.14    42682.15
     yearsdg |   763.1896    83.4169     9.15   0.000     599.3057    927.0734
     m_years |   227.1532   91.99749     2.47   0.014     46.41164    407.8947
       _cons |   36773.64   1072.395    34.29   0.000     34666.78    38880.51
------------------------------------------------------------------------------
```




More uses for the F test.

Chow test

$$F = \frac{(RSS_T - (RSS_1 + RSS_2))/(k+1)}{(RSS_1 + RSS_2)/(N_1 + N_2 - (2k+2))}$$

where RSS_t is the residual sum of squares restricted equation and the others are from the individual unrestricted equations. It has an F(K+1, n1+n2-2k-2) distribution.

Lecture 21: November 12
Multicollinearity

Lecture 22: November 14th

Remedies for MC

Do nothing

Drop redundant Variable

Transform variables

Increase sample size

Example