15th Class **6/29/10**

"Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house."--Jules Henri Poincare

consider interpretation of coefficients and other info on computer output; go back to discussion of multiple regression and omitted variable bias

consider baseball players problem on the problem set: what are important omitted variables, important in the sense that leaving them out causes the coefficient on experience to be too large (in absolute value)? Need measures of ability; after all, the better players, barring injury, play for longer in the major leagues. think also about problem of extrapolation given that few players (e.g., Nolan Ryan) play for long periods of time

discuss how to read regression output: show examples from prob. set #9 [handout]

$$s = \text{S.E. of regression} = \text{Root MSE} = \sqrt{\frac{\text{sum of squared residuals}}{\text{degrees of freedom}}}$$

(where degrees of freedom = number of observations - number of variables, including the constant term)

don't really need this though because the standard errors of the coefficients are already calculated for you in the output, along with the t-statistics and p-values for the two-tailed test where the null is that the coefficient is 0 (so note these are two-tailed p-values--to get the one-tailed p-values just divide them by 2)

test statistics for the whole regression rather than for a particular coefficient:
we will learn about the F-statistic, R-squared and adjusted R-squared today

finish discussion of nonlinear models that we started at the end of last class

nonlinear specifications--can create new variables which "trick" the regression package into estimating Y as a nonlinear function of X

e.g. suppose we think returns to experience start to decrease after a certain point, so the earnings function flattens out

may prefer to estimate:

$$\text{earnings} = b_0 + b_1 * \text{experience} + b_2 * (\text{experience})^2$$

create a new variable $\text{expsq} = (\text{experience})^2$

and run the equation:

$$\text{earnings} = b_0 + b_1 * \text{experience} + b_2 * \text{expsq}$$

can use this technique to estimate any degree of polynomial; then can use t-tests to see if the coefficients on the higher-order terms are significantly different from zero to see if they should be included or not.

can also use this for a number of other functional forms; basically any form where the dependent variable can be expressed as a linear function of the <u>coefficients</u>:

$$Y = \sqrt{X} \quad , Y = \frac{1}{X} \quad , Y = \log X$$

in particular, economists often express formulas in terms of natural logs in order to estimate them:

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

The Cobb-Douglas function is not directly estimatable using linear regression, but take the log of the Cobb-Douglas production function:

$$\log Q = \log A + \alpha \log K + \beta \log L$$

now this equation can be estimated using linear regression:

$$\log Q = b_0 + b_1 \log K + b_2 \log L$$

note that $\varepsilon = \frac{d\log Q}{d\log K} = \alpha$, so can recover elasticities directly as the coefficients of an equation in logs

discuss plotting residuals to discover systematic patterns

--better than plotting Y against X once you are in a multiple regression framework: reference line is the X-axis (0) rather than having to draw a regression line in the plot

make point about how ANOVA is a special case of multiple regression
where all the independent variables are dummy variables

what if dependent variable is a dummy/binary variable?
then we can get estimates of ranges for the proportion of people in the population that fall into each category, as well as trying to predict for an individual whether they will answer yes or no
need a different estimation technique that doesn't allow the predicted range for the variable to deviate from the 0-1 range
probit and logit techniques: still get coefficients, standard errors, t-tests, p-values

Ch. 15:
the art of inclusion vs. exclusion of variables: depends on your theory/prior if the t ratios are kind of low

interpretation of each coefficient: the effect on the dependent variable if all other variables are held constant

--it is the direct effect only

discuss tradeoff between adding a statistically insignificant variable and deleting it
depends on your prior (based on theory or prior knowledge)

this chapter builds up to presenting a criterion for choosing between regressions

refresh our minds about the correlation coefficient r

consider how b and r are related for a simple regression:

$$b = \frac{\sum xy}{\sum x^2} \text{ and } r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

$$\text{so } b = r\frac{s_Y}{s_X}$$

so if either b or r is 0, so is the other one, and testing for either one to be equal to 0 is the equivalent test for a lack of a relationship between X and Y

total sum of squares = explained + unexplained (as in 15-8)
(show geometric interpretation as in figure 15-5)

mention F-test briefly, where $F = \frac{\text{explained SS}}{\text{unexplained SS}}$

can show that $r^2 = \frac{\text{explained SS}}{\text{total SS}}$

and $s^2 = (1 - r^2)s_Y^2$

going to the multiple regression framework,

$R = r_{\hat{Y}Y}$

then $R^2 = \frac{\text{SS explained by all the regressors}}{\text{total SS}}$

we use corrected, or adjusted R^2 as our criterion to correct for irrelevant regressors:

$$\overline{R}^2 = \frac{(n - 1)R^2 - k}{n - k - 1}$$

and $s^2 = (1 - \overline{R}^2)s_Y^2$

show the class the horsebetting regression: good R-squared, but some evidence of nonlinearity in the error pattern
[handout]

return to discussion of the art of inclusion vs. exclusion of variables: depends on your theory/prior if the t ratios are kind of low

discuss tradeoff between adding a statistically insignificant variable and deleting it depends on your prior (based on theory or prior knowledge)

technique of stepwise regression

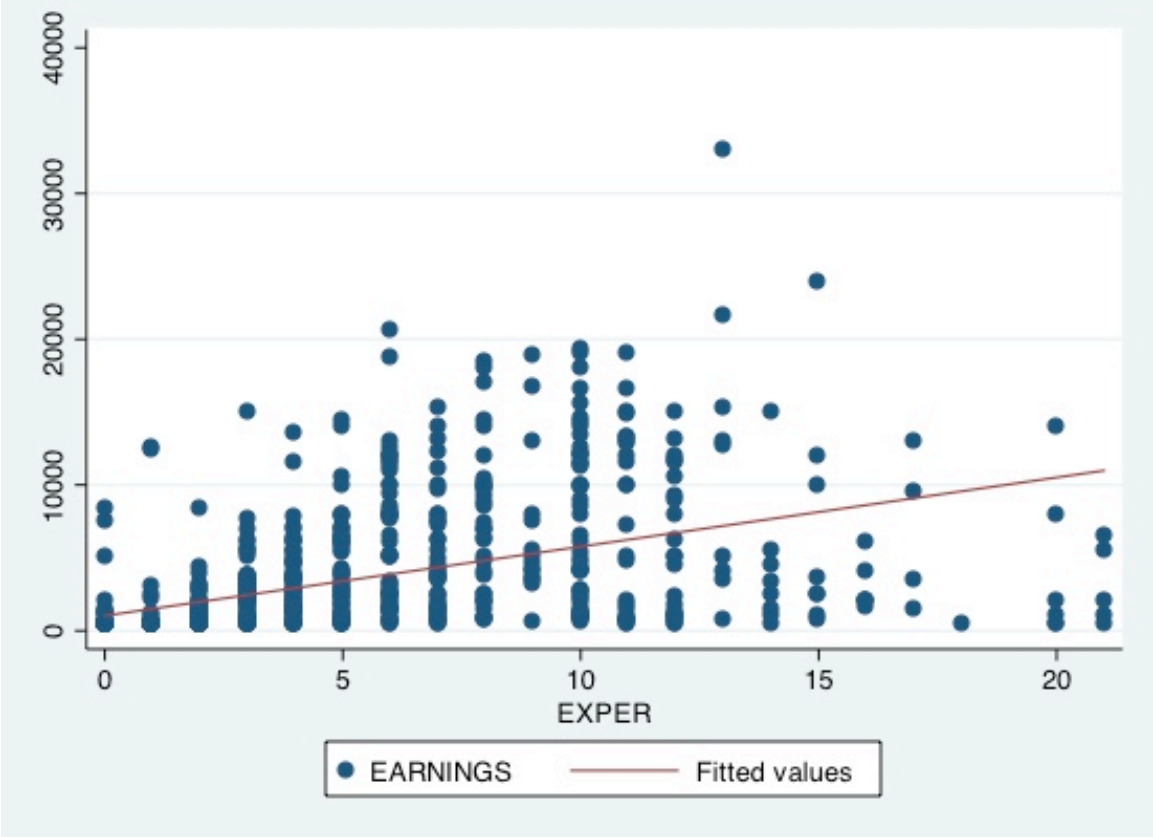
this is one of a number of techniques which economists often refer to as "fishing expeditions" or "data mining"

discuss Table 15-3

a problem with using this method is that it alters the interpretation of confidence intervals on the coefficients (you are kind of cheating by looking for the best fit from a number of possible sets of X variables)
this technique shows what an art rather than a science regression analysis is: how to gauge the "best" regression equation? Brevity (Occam's Razor), accuracy of prediction, high $\overline{R}^2$?, high t-statistics on some or all included coefficients?

problem of multicollinearity
--this is a problem with the data (X variables are overly correlated), not with the OLS estimation technique

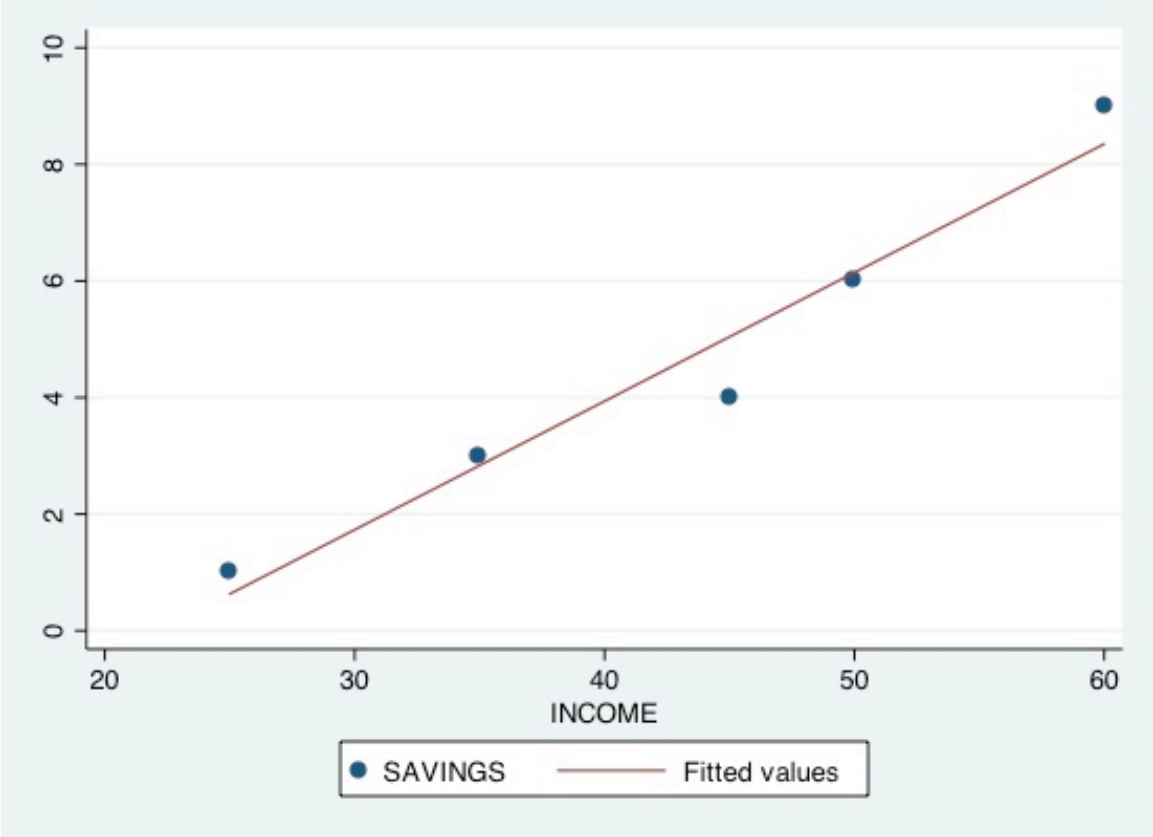
next class: finish Ch. 15



. reg EARNINGS EXPER

Source	SS	df	MS	Number of obs =	818
Model	3.6827e+09	1	3.6827e+09	F(1, 816) =	253.02
Residual	1.1877e+10	816	14555035.4	Prob > F =	0.0000
Total	1.5560e+10	817	19044849.4	R-squared =	0.2367
				Adj R-squared =	0.2357
				Root MSE =	3815.1

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
EXPER	474.1865	29.8106	15.91	0.000	415.672	532.701
_cons	1032.355	193.4111	5.34	0.000	652.7131	1411.997



. reg SAVINGS INCOME

Source	SS	df	MS	Number of obs =	5
Model	35.5082192	1	35.5082192	F(1, 3) =	62.97
Residual	1.69178082	3	.563926941	Prob > F =	0.0042
				R-squared =	0.9545
Total	37.2	4	9.3	Adj R-squared =	0.9394
				Root MSE =	.75095

SAVINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
INCOME	.2205479	.0277939	7.94	0.004	.1320952	.3090006
_cons	-4.883562	1.241428	-3.93	0.029	-8.834339	-.9327846