16th Class **6/30/10**

"It is the mark of a truly intelligent person to be moved by statistics."
George Bernard Shaw

go back to the correlation coefficient and do more discussion of regressions

one can distinguish between a population correlation and the sample correlation and construct a confidence interval for the sample correlation r, but only if you assume that both X and Y are random variables (can still calculate r and b even if Xs are selected at specified levels, but then cannot talk about an underlying population correlation ρ between X and Y).

long interest in relationships between physical characteristics and mental characteristics [show link to article about longer limbs and dementia; notice this is also an a causes b and c example]

Francis Galton (1822-1911) the first to discuss regression toward the mean and the correlation coefficient. Started on meteorology and moved into heredity issues
[see handout]
note that regression (method of least squares) and even the use of the term "regression" predates the use of the correlation coefficient
Stigler quote: "Galton's own use of regression in *Natural Inheritance* shows either great naiveté or great optimism. Having found that the slope of the regression of child's height upon a midparent was $\frac{2}{3}$ (as shown in the figure on the handout) , he blithely supposed that the same value held for all other characteristics." For example, he based calculations for artistic ability on this same ratio.

another historical sidenote: Edgeworth was a distant cousin of Galton's and his work in statistics was influenced by Galton's work

consider the different ways one can predict Y (or X):

e.g., X = midterm score, Y = final score

1) use the average: $Y_i = \overline{Y}$
2) use the previous score: $Y_i = X_i$
3) use the previous score plus the average improvement: $Y_i = X_i + D$
4) use the regression line: $Y_i = a + bX_i$

these methods are better as we move from (1) to (4), where we know that OLS (4) is the best linear method (is BLUE), but is also the most "costly"
What about predicting X from Y?
Would you want to use four to predict X also?

compare directionality: $Y = f(X)$ or $X = g(Y)$

the correlation coefficient is symmetric, but the slopes and intercepts of these lines are different; however, if you know the slope of one of these lines, you can derive the slope (and intercept) of the other:

$$Y = a + bX$$
$$X = a' + b'Y$$

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

$$b = \frac{\sum xy}{\sum x^2} = r\frac{s_Y}{s_X} \text{ and } a = \overline{Y} - b\overline{X}$$

$$b' = \frac{\sum xy}{\sum y^2} = r\frac{s_X}{s_Y} \text{ and } a' = \overline{X} - b'\overline{Y}$$

contrast regression of Y against X to X against Y and to the equivalent performance line (also to the equal performance line and the average performance lines)
[draw graph]

show how both lines exhibit regression toward the mean, and away from the equivalent performance line (which is a 45 degree line through the origin if $\overline{X} = \overline{Y}$)

do Problems 15-7 and 15-9

15-7 a. $r = \frac{11000}{\sqrt{18000}\sqrt{21000}} = 0.57$

b. $X_2 = a + bX_1$

$$b = \frac{11000}{18000} = 0.61 = 0.57 * \frac{\sqrt{\frac{21000}{80}}}{\sqrt{\frac{18000}{80}}} \quad \text{and } a = 62 - 0.61*62 = 24$$

so $X_2 = 24 + 0.61X_1$

c. $X_2 = 24 + 0.61(90) = 79$; $X_2 = 24 + 0.61(40) = 48$

f. True with the qualification that there was a <u>tendency</u> to move closer to the mean

15-9 a. cannot distinguish effects of praise vs. punishment from phenomenon of regression to the mean
b. need randomized experimental design to decide which regime is better

discuss some more 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?

a few more examples of regression problems (partial models for project)

mention sports data sites, can try to explain why sports stars make the salaries that they do based on their personal characteristics and their team characteristics
show Lee's movie info site: how to derive predictions of revenue?

talk about causality again [use unemployment and well-being discussion]

problem of multicollinearity

--this is a problem with the data (X variables are overly correlated), not with the OLS estimation technique
--diagnosis of this problem: high correlation coefficients between various X-variables, high standard errors leading to low t-statistics on the variables where there is a problem
--this is because the correlation between the X variables increases the standard errors on the affected variables, which are an increasing function of the correlation of each X_i with all the other regressors:

the standard error of coefficient b_i, $SE_i = \frac{s}{\sqrt{\sum x_i^2}\sqrt{1 - R_i^2}}$

where s^2 = the usual residual variance (the sum of squared residuals divided by degrees of freedom)
and R_i is the multiple correlation of X_i with all the other regressors
so as R_i increases, the denominator shrinks and blows up the standard error of b_i
--can be helped by getting more data, particularly data in which there is more spread in the X variables (or at least in the one or more that are problematic)
--this problem can be avoided in experimental design by choosing the X values so that they are completely uncorrelated (e.g. in an agricultural experiment, set a number of values for land quality and levels of fertilizer and assign the pairs of values randomly to each other, or make sure that each possible cell is filled where the number of cells is the number of values for land quality times the number of levels of fertilizer)
--unfortunately, economists rarely have this luxury available since they use observational data

--sometimes there just isn't enough data to tell
[show death penalty discussion; also brings up the problem of statistical vs. practical significance--and plausible significance]

We'll let Galton have the last word today:
Some people hate the very name of statistics but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary.