Stat 218 - Day 34
Least Squares Regression (cont.)

We have begun studying the technique known as least squares regression for modeling the relationship between two quantitative variables with a straight line. Last time we examined:

- Residuals
- Least squares criterion
- Prediction
- Interpretation of slope coefficient
- Influential observations
- Coefficient of determination r^2
- Residual standard deviation

One important issue that we have yet to consider the question of how to calculate the slope and intercept coefficients of the least squares line. Let the equation of a generic least squares line be $\hat{y} = b_0 + b_1 x$, so b_0 is the intercept coefficient and b_1 is the slope coefficient.

- Two ways to find the least squares value of the slope coefficient are: $b_1 = r\frac{s_y}{s_x}$ and $b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$.
- The least squares value of the intercept coefficient is found as: $b_0 = \bar{y} - b_1 \bar{x}$.

Example: New car data (cont.)
Reconsider the new car data, and consider predicting a car's highway MPG rating from its weight.

(a) Examine and comment on a scatterplot of these data (`cars99.mtw`). Remember to put the response variable on the vertical axis.

Calculate the following descriptive statistics:

Mean highway MPG	Std dev highway MPG	Mean weight	Std dev weight	Correlation coefficient

(b) Use these statistics to determine the least squares line for predicting a car's highway MPG rating from its weight. [Hint: Be sure to write this as an *equation*.]

(c) Use Minitab to confirm these calculations and to superimpose the regression line on the scatterplot (`Stat> Regression> Fitted Line Plot`).

(d) Interpret the value of the slope coefficient of this line.

(e) What highway MPG rating would the least squares line predict for a car weighing 3600 pounds?

(f) What proportion of the variability in highway MPG ratings is explained by the least squares line with weight?

Example: House prices

The data in the Minitab worksheet `RealEstate.mtw` are the prices (in thousands of dollars), sizes (in square feet), number of bedrooms, and number of bathrooms of a sample of houses sold in the Bakersfield, California area for one week in April 2003.

(a) Which of these variables is most reasonable to consider as the response variable?

(b) Make a guess for which of the three explanatory variables would be the best predictor of the response variable.

(c) For the explanatory variable that you guessed in (b), produce and comment on a scatterplot, with the regression line superimposed. Also report and interpret the value of the slope coefficient. Also report and interpret the value of r^2. Also report the value of the residual standard deviation. Does this explanatory variable appear to do a reasonable job of predicting the response? Explain.

(d) Repeat (c) for one of the other explanatory variables.

(e) Repeat (c) for the remaining explanatory variable.

(f) Which explanatory variable appears to be most helpful for predicting the response variable? Explain.

Example: Televisions and life expectancy (cont.)
Reconsider the data on the life expectancy and number of people per television set in a sample of 22 countries (`TVlife.mtw`).

(a) Create a scatterplot of the data, with the regression line for predicting life expectancy from number of people per television superimposed. Does the regression line appear to summarize the relationship well? Explain.

(b) Transform the "people per television" variable by taking the log base 10 (`MTB> let c5=logt(c4)`). Examine a scatterplot of life expectancy vs. this transformed variable. Does the relationship appear to be more linear?

(c) Determine the regression equation for predicting life expectancy from this transformed variable.

(d) Use this regression equation to predict the life expectancy for a country with 10 people per television. Then repeat for a country with 100 people per television.

(e) Calculate the difference in these predictions. Does the result look familiar? Explain why this makes sense.