Regression Using MINITAB

Example: Poplar tree example (from lecture 1) - poplar2.mtw

- We are interested in whether D^2H can be used to predict the weight of trees.
- Regression of Weight on D^2H: Weight = a + bD^2H
  - Select **Stat-Regression-Regression**
  - Response: Weight
  - Predictors: D^2H
  - Output: regression equation, r^2, possible outliers and influential observations

- Residuals and Fitted Values:
  - Select **Stat-Regression-Regression** and click on **Storage**. Select **Residuals** and **Fits.**
- Residual Plots:
  1. Use the residuals and fitted values in the worksheet to make them.
  2. Select **Stat-Regression-Regression** and click on **Graphs**. Choose **Regular** residuals and put D2H in the Residuals Versus the Variables box.

  Other types of regression plots (such as those discussed in lecture notes 9) are sometimes useful and can be made in Minitab.

- Removing the possible *influential observation*.
  - First, identify it: Select **Editor-Brush** and click on it on one of the plots.
  - It turns out the weight of tree 15 is 0.7 instead of 0.07. Find the regression equation with this correction made.

- Making a scatterplot of the data with the regression line:
  - Select **Stat-Regression-Fitted Line Plot**
  - Specify the response and predictor variables and select "linear" as the type of regression model.
- Notice that this plot also give the r^2 value and the equation of the regression line.

Limitations of Correlation and the Regression Model:

1. Correlation measures only *linear association*, and fitting a straight line makes sense only when the overall pattern of the relationship is linear. Always plot your data before calculating it.
2. *Extrapolation* often produces unreliable predictions.
3. Correlation and least-squares regression are *not resistant*. Always plot your data and look for potentially influential observations.
4. *Lurking variables…*

Association Does Not Imply Causation

An association between an explanatory variable x and a response variable y, even if it very strong, is not itself good evidence that changes in x actually cause changes in y.

Why? *Lurking variables* can create "nonsense correlations."

Example: flu cases and ice-cream sales - correlated, but is there a causal relationship?

Example: nicotine use and lung cancer - correlated, but is there a causal relationship?

Possible *lurking variables* => causal association may not be real:

- **Common variables** such as a genetic factor that predisposes people both to nicotine addiction and to lung cancer (hard to observe).
- **Confounding variables** (diet, alcohol use, lack of exercise) interact with smoking to prevent us from drawing conclusions about the association between nicotine use and cancer.

Verifying Causal Relationship

- Very hard!
- Need to do an experiment that *controls* for lurking variables. (Chapter 3)

Ecological Fallacy / Simpson's Paradox

A correlation based on averages over many individuals is usually higher than the correlation between the same variables based on data for individuals.

The Restricted Range Problem

When data are only observed on a restricted range, r and r^2 are lower than they would be if the full range of data could be observed.