4.3 Establishing Causation

3 things to know:

1. ***Causation*** – A direct cause-and-effect link between the variables.
* Mothers/Daughters BMI - \*\*Even when direct causation is present, it is rarely a complete explanation of an association between two variables.
* Rats with bladder tumors - \*\*Even well-setablished causal relationships may not generalize to other settings.
1. ***Common Response*** – Both x and y change in response to lurking variable(z).
* High SAT scores and GPA
* Stock market
1. ***Confounding*** – Two variables effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory or lurking.
* Going to Church and Living longer
* Years of education and Income

Establishing CAUSATION…

The best method for establishing causation is to conduct a carefully designed experiment in which the effects of possible lurking variables are controlled. (Ch. 5) However, experiments are not always possible for ethical or practical reasons.

Criteria…….

* Association is strong
* Association is consistent
* Larger values of the response variable are associated with stronger responses
* The alleged cause precedes the effect in time
* The alleged cause is plausible

4.2 Categorical Variables

To analyze categorical data, we use counts or percents of individuals that fall into various categories. (no scatterplots, no correlation measure to help us summarize strength☹)

Two-way Table: College students by sex and age group, 2003 (thousands of persons)

|  |  |  |  |
| --- | --- | --- | --- |
| Age Group | Female | Male | **TOTAL** |
| 15 to 17 years | 89 | 61 | **150** |
| 18 to 24 years | 5,668 | 4,697 | **10,365** |
| 25 to 34 years | 1,904 | 1,589 | **3,494** |
| 35 years or older | 1,660 | 970 | **2,630** |
| **TOTAL** | **9,321** | **7,317** | **16,639** |

***Marginal Distributions***

The distributions of age alone and sex alone. Each marginal distribution from a two-way table is a distribution for a single categorical variable.

Marginal Distributions of AGE:

Marginal Distributions of SEX:

***Conditional Distributions***

When we compare one variable to another with “given” information. Comparing conditional distributions is one way to dexcribe the association between the two variables.

Conditional Distribution of SEX given that a student is “18 to 24 years old”.

Conditional Distribution of AGE given that a student is “male”.

Be careful of…***Simpson’s Paradox*** – an association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group.

4.1 Modeling Nonlinear Data

1. Create a scatterplot of the data (L1, L2)
	1. The scatterplot should not be linear; if it is then there is no need to do a data transformation!
	2. Consider the data in context to decide how to go about straightening it:
		* Sometimes the problem itself will tell you how to go about straightening the data; #15 told us to take squared time and the square-rooted distance in order to achieve linearity.
		* Exponential growth and decay which is (x, log y), or calculator-wise (L1, L3 = log L2): to check if something is changing exponentially, find the ratios of the y values by taking each y value and dividing it by the previous y value. The quotient for each division should be relatively the same (see #6 part b). If the data is exponential a scatterplot comparing (L1, L3) should show a relatively straight set of points
		* Power models which is (log x, log y), or calculator-wise (L3 = log L1, L4 = log L2): variables that have a dimensional relationship, for instance height (one-dimensional measure) vs. weight (three-dimensional measure). If the data fits this category a scatterplot comparing (L3, L4) should show a relatively straight set of points
		* Unsure of which model to use? Try both an exponential and a power transformation and determine which scatterplot shows a straighter set of points.
2. How to confirm that the transformation fit the data to a straight line:
	1. Exponential models:
		* Residual Plots: after the LSR equation (L1, L3) has been pasted into Y1, L4 = Y1(L1) and L5 = L3 – L4. Look at the scatterplot for L1 vs. L5(RESID). Remember that the data may show some type of pattern, but as long as the numerical summaries are strong, this isn’t an issue here.
		* Numerical summaries:
			1. r2: is the percent of variation in the log y data that is explained by the linear relationship with the x data.
			2. s =  and is the average length of the residuals; (average amount that the observed log y’s are from the predicted log y’s) Σ*X*2 is found in after performing 1-varstat L5 calculation.
	2. Power models
		* Residual Plots: after the LSR equation (L3, L4) has been pasted into Y1, L5 = Y1(L3) and L6 = L4 – L5. Look at the scatterplot for L3 vs. L6 (RESID). Remember that the data may show some type of pattern, but as long as the numerical summaries are strong, this isn’t an issue here.
		* Numerical summaries:
			1. r2: is the percent of variation in the log y data that is explained by the linear relationship with the log x data.
			2. s =  and is the average length of the residuals; (average amount that the observed log y’s are from the predicted log y’s) Σ*X*2 is found in after performing 1-varstat L6 calculation
3. More on the LSR equations:
	1. Exponential models:
		* Written log y = slope (x) + intercept
		* After an inverse transformation (see #5 part f or example 4.7 in the book when dealing with ln transformation) you should end up with a y = abx
	2. Power models:
		* Written log y = slope (log x) + intercept
		* After an inverse transformation (see #11 part c or the example 4.9 the book), you should end up with y = axp
	3. With either model:
		* Given the original scatterplot of the data (L1, L2), the equation you found after the inverse transformation should fit nicely over the scatterplot when you write the equation in Y1.
		* Predictions with your new equation are the easiest part! Overflow problems will happen sometimes because the numbers are so astronomically huge that our basic technology can’t even fathom.