AP Statistics Chapter 4 – More about Two-Variable Relationships

4.1: Transformations to Achieve Linearity – Nonlinear Modeling

The Exponential Model

The equation: $y = A(B)^x$

Linear Transformation: $(x, \log y)$

Finding A & B from a & b: $A = 10^a$, $B = 10^b$

Example – The population of the U.S. is given below (in millions) from 1900-2000.

Year	Pop.
1900	76.2
1910	92.2
1920	106.0
1930	123.2
1940	132.2
1950	151.3
1960	179.3
1970	203.3
1980	226.5
1990	248.7
2000	281.4

Let x = years since 1900, so 0, 10, etc…
Take log of y (pop), then perform linear regression on $(x, \log y)$.

Result: $y = 1.90607 + .00556x$

So $A = 10^{1.90607} = 80.551$
and $B = 10^{.00556} = 1.013$

So the exponential model is $y = 80.551(1.013)^x$

This models says two things:

- Estimate 1900 population is 80.55
- Population is growing 1.3% per year

The Power Model

The equation: $y = Ax^B$

Linear Transformation: $(\log x, \log y)$

Finding A & B from a & b: $A = 10^a$, $B = b$

Example – Weight lifted in 2000 Olympics by athletes in various weight classes (in kg).

Class	Lifted
56	305
62	325
69	357.5
77	367.5
85	390
94	405
105	425

Take log of x (weight class) and log of y (weight lifted), then perform linear regression on $(\log x, \log y)$.

Result: $y = 1.58281 + .52025x$

So $A = 10^{1.58281} = 38.266$
and $B = .52025 = .520$ (to 3 decimal places)

So the power model is $y = 38.266x^{.520}$

Use the model to predict the weight lifted by an athlete in the 115 kg class:

$y = 38.266(115)^{.520} = 451.21$ kg

4.2: Relationships between Categorical Variables

Two-Way Tables
A ***two-way table*** organizes counts from two categorical variables into rows and columns. They are often used to summarize large amounts of data by grouping outcomes into categories.

> **Example**: Below is data on a random selection of people who were categorized as having low, moderate or high anger. Additionally, it was determined whether or not they had coronary heart disease (CHD) at the end of a 4-year study.

	Low anger	Moderate anger	High anger	**Total**
CHD	53	110	27	**190**
No CHD	3057	4621	606	**8284**
Total	**3110**	**4731**	**633**	**8474**

Marginal Distributions
The row totals and column totals give the ***marginal distributions*** of the two individual variables. These numbers tell us nothing about the relationship between the two variables. They are simply used to summarize the data.

> Looking at the marginal distributions of anger rating in percents, we get:

Low anger	**Moderate anger**	**High anger**
$\frac{3110}{8474} = 36.7\%$	$\frac{4731}{8474} = 55.8\%$	$\frac{633}{8474} = 7.5\%$

> Note that this tells us what percent of the total fell into each of the three groups in the anger rating variable. We see here that the majority of subjects were rated moderate, and a very small percent were rated as high.

Conditional Distributions
To find the ***conditional distribution*** of the row variable, begin by focusing on a single column. Then find each entry in the column as a percent of the column total. These percentages will tell us about the association between the two variables in the table.

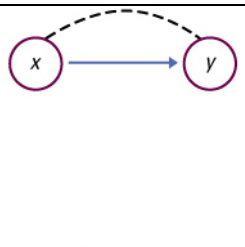
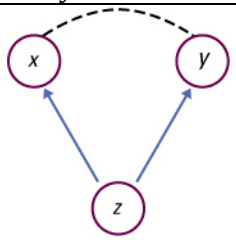
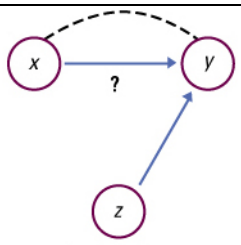
> Now we look at whether anger tells has an association with CHD by calculating the percent with CHD in each anger group.

Low anger	**Moderate anger**	**High anger**
$\frac{53}{3110} = 1.7\%$	$\frac{110}{4731} = 2.3\%$	$\frac{27}{633} = 4.3\%$

We observe that the rate of CHD increases with anger level. So we conclude that the angrier you are, the more likely you are to have coronary heart disease. The moral to this story is…chill out!

4.3: Establishing Causation

The Question of Causation
An association between two variables doesn't necessarily mean that *x causes y*. Consider 3 possibilities in this situation and diagrams of the relationship between variables.

Causation – *x* does indeed cause *y*	**Common Response** – a 3rd variable, *z*, causes changes in both *x* and *y*	**Confounding** – both *x* and a 3rd variable, *z*, cause changes in *y*
 Causation	 Common response	 Confounding

In the above diagrams, *z* would be a *lurking variable*

Also consider that sometimes it is believed that *x causes y*, when it is actually the other way around (*y causes x*). Call this **Reverse Causation.**

Confounding – definition
Two variables are said to be confounded when their effects on a response variable cannot be distinguished from each other.

Examples

1. You observe that when more people wear coats there are also more people with colds. Does wearing a coat cause the common cold? No – weather is a ***common response*** lurking variable here. Cold weather causes more people to wear coats and more people to get colds.

2. A study looks at the effects of after school tutoring and finds that students who get tutoring fail classes at a higher rate than those who do not go to tutoring. So does tutoring just cause students to do worse? No – this is ***reverse causation***. Students go to tutoring because they are doing poorly in class, not the other way around.

3. People who drink large amounts of alcohol tend to die earlier than those who do not. Does alcohol *cause* an early death? Certainly, but other factors are to blame as well. People who drink a lot probably also have other harmful behaviors such at poor diet, which can also lead to early death. So both behaviors have causal links to early death, but how much does each contribute? Hard to say, since these variables are ***confounded***.