Review - Week 4

Read: Chapters 10-13

Review: Re-expressing Data

Often working with a re-expression of the data can greatly simplify statistical analysis. In the context of regression the goals of re-expression are to make:

- the form of a scatterplot more linear, and,
- the scatter in a scatterplot spread out evenly rather than follow a fan shape.

The most common transformations belong to the family of power transformations $y^p$, for some constant $p$, and $\log(y)$.

Linear growth increases by a fixed amount in each equal time period. Exponential growth increases by a fixed percentage of the previous total.

The exponential growth model $y = ab^x$ becomes linear when we plot $\log(y)$ against $x$.

Many laws of physics follow the power law model $y = ax^p$. It becomes linear when we plot $\log(y)$ against $\log(x)$.

Exercise 1: For the models below, predict $y$ for $x=4$.

(a) $\hat{y} = 3.2 + 1.2x$
(b) $\frac{1}{\hat{y}} = 1.2 + 3.2x$
(c) $\frac{1}{\sqrt{\hat{y}}} = 0.5 + 0.25x$
(d) $\log(\hat{y}) = 0.5 + 0.25x$
(e) $\log(\hat{y}) = 0.5 + 0.25\log(x)$

Exercise 2: A car manufacturer tested the stopping distance in feet at a variety of speeds for their new car model. Using the data they constructed a model for predicting stopping distance, $y$, for a car traveling $x$ mph. The model took the form:

$\sqrt{\hat{y}} = 3.303 + 0.235x$

(a) Estimate the stopping distance for a car traveling 55 mph.
(b) Estimate the stopping distance for a car traveling 70 mph.
**Review:** Sample Surveys

We collect data through simulations, sample surveys and experiments.

The idea behind sampling is to gain information about a large number of individuals by examining a smaller group of the individuals in question.

**Population** - Entire group of items/individuals we want information about.
**Sample** - The part of the population we actually examine in order to gather information.
**Strata** - subpopulations

The best way to choose a representative sample of a population is through randomization.

The sample size determines how well the sample represents the population, not the fraction of the population sampled.

A **parameter** is a number that describes the population. A parameter is a fixed number, but we do not know its actual value. A **statistic** is a number that describes a sample. The value of a statistic is known after we take a sample, but it can vary from sample to sample. The **sampling variability** is the variability in a statistic from sample to sample.

The design of a sample refers to the method used to choose a sample from the population. We want the sample to be representative of the population, i.e. we want the statistics computed from the sample to reflect the corresponding parameters accurately.

The best way to get a representative sample is to select the individuals in the sample randomly. To select a sample at random we need a **sampling frame**, which is a list of individuals from which the sample is drawn. There are a number of ways to randomly choose a sample of the population.

- **Simple random sample (SRS)** - choose \( n \) individuals from the population in such a way that every set of \( n \) individuals has an equal chance to be selected.
- **Stratified random sample** - divides the population into strata and choose a simple random sample from each sample.
- **Cluster sampling** – divides the population into similar groups and selects an entire group, or cluster, at random.
- **Multistage samples** - A combination of simple random samples and stratified random samples.
- **Systematic samples** – a sample drawn by selecting individuals systematically from the sampling frame.

Our goal in sampling design is to reduce bias, or the systematic favoring of a certain outcome. Bias can result from:

- **Voluntary response** - overrepresentation of people with strong opinions
- **Convenience sampling** – includes individuals who are on hand. This may not be representative of the population.
- **Undercoverage** - when some group of the population is given either no chance or a much smaller chance than other groups to be in the sample.
- **Nonresponse bias** – when individuals who are selected to be in the survey do not participate or cannot be contacted.
• Response bias – when individuals do participate but are not responding truthfully.
• Wording of questions. Never trust the results of a sample survey until you have read the exact question posed.

**Exercise 1:** A politician is running for office in a city with 100,000 registered voters. To gauge voter interest, he commissions a survey of 500 registered voters. The results show that 56% of those interviewed answer that they plan on voting on him.

(a) What is the population of interest?
(b) What is the sample?
(c) Is the value 56% a parameter or a statistic?

**Exercise 2:** A company employs 10 managers, 25 foreman and 400 laborers. To determine the level of employee discontent, the company plans to conduct job satisfaction interviews with a sample of the employees.

(a) Propose a sampling strategy that uses simple random sampling. How would you pick the people to interview?
(b) Is there any danger of bias in the company’s plan?
(c) Suppose the company wants to ensure that each of the three levels of employees is represented in the sample. How would you design this sampling strategy?

**Exercise 3:** Out of a group of 12 students (call them A, B, C, D, E, F, G, H, I, J, K and L), three are to be chosen at random to be interviewed about their school experience. Use the random digits below to select a sample of three students.

04905 83852 29350 91397 19994 65142 05087 11232