Math 316

Section 12.6 Interpreting Regression and Correlation

There are various types of regression (fitting a function to data). Linear regression is sometimes (often?) not appropriate. See **two examples** on page 550. Also, sometimes a set of data is transformed before we work with it. For example, we might log values so the data are not as spread out. See **Class Example 1**.

If using a straight line $b_0 + b_1X$ to fit data, the quantities b_0, b_1, s_e and r can be used to describe a scatterplot that shows a linear trend:

- b_0 and b_1 describe the line that best fits the data.
- s_e and r tell us how well the data fit the line:
  - r tells us how well a line would fit the data. **See Figure 12.2.3**.
  - Residual standard deviation s_e is essentially the average distance of each data point to the line, and describes how spread out the given data are around the linear regression line. It also gives us an idea of how much data are within a certain range/window of the line: roughly 2/3 of points are within one s_e. See **Figure 12.3.8**.
  - And recall how r and s_e are related: $r^2 \approx 1 - \frac{s_e^2}{s_y^2}$.
    - s_y^2 measures total variation within y-values.
    - s_e^2 measures total variation due to line not perfectly fitting data.
    - s_e^2 / s_y^2 is percentage of total variation due to line not perfectly fitting data.
    - r^2 is percentage of variation due to movement (slope) in line.

Recall outliers can affect how well a line would fit data. See **Figures 12.6.3** and **12.6.4**.

From the book, on page 552:

1. *Design conditions.* We have discussed two sampling models for regression and correlation:
   (a) Random subsampling model: For each observed X, the corresponding observed Y is viewed as randomly chosen from the conditional population distribution of Y values for that X.*
   (b) Bivariate random sampling model: Each observed pair (X, Y) is viewed as randomly chosen from the joint population distribution of bivariate pairs (X, Y).
   In either sampling model, each observed pair (X, Y) must be independent of the others. This means that the experimental design must not include any pairing, blocking, or hierarchical structure.
2. *Conditions concerning parameters.* The linear model states that
   (a) $\mu_{Y|X} = \beta_0 + \beta_1 X$.
   (b) σ_e does not depend on X.
3. *Condition concerning population distributions.* The confidence interval and t test are based on the conditional population distribution of Y for each fixed X having a normal distribution.

Examples of:

1a	**Figure 12.1.1**
1b	**Figure 12.1.2**
2a/b, 3	**Figure 12.4.1**

We'll not cover Section 12.7.

Section 12.8 Perspective

If we have time, else on your own: in **Example 12.8.1** is how we normally compare two samples. In this example, $SE = \sqrt{66.12^2/6 + 69.64^2/5} \approx 41.195$. Just above Figure 12.8.1 is the value of t_s if SE were computed using $SE^2_{pooled} \approx 41.00$; t_s is *approximately* the same. We saw in Chapter 11, that the pooled standard deviation s^2_{pooled} was most useful to us when doing ANOVA, when we needed to combine (find the average of) the standard deviations of multiple samples. A third way to compare two samples is to see if the straight line between the two collections of sample data would have slope 0 or not. See **Figure 12.8.2** and just below is $\boldsymbol{t_s = b_1/SE_{b_1}}$. It's the same t_s as when using s^2_{pooled}. You can read the details on your own. Kind of interesting.

Example 12.8.2 shows us that more information is better, and that we should find a way to incorporate that additional information into our analysis. In this case, the additional information is the range of blood pressure levels within each group of *normal* vs. *high* blood pressure. Indeed, the transition between low and high blood pressure is arbitrary, since there is a wide range of values within each group and overall.

There is a whole world of other types of regression. There are two versions of this:

1. One variable X with more in the regression formula than simply $b_0 + b_1X$
2. More than one variable $X_1, X_2, etc.$

See just above and below **Example 12.8.3** for a few examples. Another example is Example 12.8.5, **Figures 12.8.6/7**. To me trying to fit this logistic function to these data is likely a bad idea. Another (more appropriate) example is seen in Figures **12.8.9/10**. Notice the ***Nonparametric*… note** just above **Example 12.8.4**. Statisticians continue to try to find and create better ways of working with data.

Let's look at Example 12.8.4, **Figures 12.8.4** and **12.8.5**. Notice in **Figure 12.8.4** the X and Y values of the data have been *natural logged*, which made the data less spread out (as mentioned today in Class Example 1). The book states that it should be useful to have three different (as proved by doing ANOVA) types of data (in this example, three different diets) in which the same general trend is apparent, that body weight and head weight are positively correlated: as one is larger, so is the other. The more different contexts in which we can prove correlation, the more convinced we are of the correlation.

Section 12.9 is a **summary of the formulas** from this chapter. **Notice that test statistic** t_s has two forms and is used to test either null hypothesis $H_0: \beta_1 = 0$ or $H_0: \rho = 0$. **Let's look at this in more detail in class**. Note that most of the values we could compute using these formulas are usually just given to us. Don't worry about the *Prediction* formulas, which come from Section 12.7, which we are not covering.