General Comments on Statistical Inference

1 Hypothesis Statements

Let's consider a t test for the regression parameter, β_1. Depending on context, the following could be null and alternative hypotheses

- $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$.
- $H_0 : \beta_1 = 0, H_a : \beta_1 > 0$.
- $H_0 : \beta_1 = 0, H_a : \beta_1 < 0$.

The null hypothesis should be stated as a statement of **equality**. This concept holds true for hypothesis tests in general. Some other books / resources might state them as

- $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$.
- $H_0 : \beta_1 \leq 0, H_a : \beta_1 > 0$.
- $H_0 : \beta_1 \geq 0, H_a : \beta_1 < 0$.

I prefer using the equality statement for the null hypothesis for the following reasons (theoretical, pedagogical, practical):

1. The null hypothesis being an equality aligns with the definition of the p-value.
    - The p-value is the probability of observing our sample estimate (or a value more extreme), if the null hypothesis is true (i.e. β_1 is truly 0). This is what we are assuming in the calculation for the test statistic.
2. People tend to get confused between the null and alternative hypotheses if both involve inequalities (the alternative is the hypothesis you are trying to support).
3. Conclusions are made in terms of supporting (or not supporting) the alternative hypothesis.

2 Sample Size and Statistical Inference

Generally speaking, there is a relationship between sample size and statistical inference (assuming other characteristics remain the same and our sample was randomly obtained or representative of the population of interest):

- Larger sample sizes (typically) lead to narrower confidence intervals (more precise intervals).
- Sample estimates based on larger samples are more likely to be closer to the true parameters.
- Larger sample (typically) lead to more evidence against the null hypothesis.
  - This means a larger sample size leads to a more powerful test. The power of a test is the probability a hypothesis test is able to correctly reject the null hypothesis.

2.1 Small Sample Sizes

Small sample sizes tend to result in:

- Confidence intervals that are wide.
- Sample estimates that are more likely to be further away from the true parameters.
- Hypothesis tests that are more likely to incorrectly fail to reject the null hypothesis when the alternative hypothesis is true.

While larger sample sizes have their advantages, there are also some disadvantages with sample sizes that are extremely large.

2.2 Large Sample Sizes

A "statistically significant" result does not necessarily mean that the result has practical consequences. Suppose a 95% confidence interval for β_1 is $(0.001, 0.002)$. The interval excludes 0, so it is "statistically significantly" different from 0 (because it is!), but does this result have practical consequences? A narrow CI that barely excludes the null value can happen when we have a large sample size.

If one was to conduct the corresponding hypothesis test, we would reject the null hypothesis that $\beta_1 = 0$. With large sample sizes, hypothesis tests are sensitive to small departures from the null hypothesis.

In such instances, it may be worth considering hypothesis tests involving a different value in the null hypothesis, one that makes sense for your question. For example, a practically significant slope may need to be greater than a specific numerical value for a certain context.

- Statistical inference to assess statistical significance.
- Subject area knowledge to assess practical significance.

2.2.1 Questions

Are the following results statistically significant? If so, are the results also practically significant? Assume a two-sided test with a null value of 0 (These are made up examples):

1. In assessing if studying more is associated with better test scores, a SLR is carried out with test scores (out of 100 points) against study time (in hours). The 95% confidence interval for the slope β_1 is (5.632, 7.829).

2. A SLR is carried out to explore the linear relationship between number of years in school with income (in thousands of dollars). The 95% confidence interval for the slope β_1 is (0.051, 0.243).

3 Cautions using SLR and Correlation

Simple linear regression and correlation are meant for assessing **linear** relationships. If the relationship is not linear, we will need to transform the variable(s) (so the transformed variables have a linear relationship. Will explore this in Module 5).

- Always verify via a scatterplot that the relationship is at least approximately linear.

- A high correlation or a significant estimated slope by themselves do not prove that we have a strong linear relationship between the variables. Conversely, a correlation close to 0 or an insignificant estimated slope is also not proof that we do not have a relationship between the variables. (Please see the video on Module 3.6 in Canvas for some examples.)

3.1 Outliers

SLR and correlation are sensitive to outliers / influential observations. Generally speaking, these are data that are "far away" or very different from the rest of the observations. These data points can be visually inspected from a scatterplot. (Please see the video on Module 3.6 in Canvas for some examples. We will explore measures on how to detect these numerically in Module 10.) Some potential considerations when dealing with such data points:

- Investigate these observations. There is usually something that is making them "stand out" from the rest of the data.

- Data entry errors that can be corrected. Be sure to mention in the report.

- Revisit how the data were sampled. Perhaps the data point is is not part of the population of interest. If so, data point can be removed (this is legitimate), but be sure to mention in the report.

With regards to regression analysis:

- Exclusion of data points must be clearly documented.

- Fit the regression with and without the data points in question, and see how similar or different the conclusions become.

- If the data points have large value(s) on the predictor and/or response, a log transformation on the variable can pull in the large values.

- Consider subsetting your data and create separate models for each subset; or focus on a subset and make it clear your analysis is for a subset.

- Knowing your data and context can help a lot in these decisions.

3.2 Association and Causation

Two correlated variables do not mean that one variable causes the other variable to change. For example, consider a plot of ice cream consumption and deaths by drowning during various months. There may be some positive correlation, and clearly, eating more ice cream does not cause more drownings. The correlation can be explained by a third (lurking) variable: the weather.

A **lurking variable** is a variable that has an impact on the relationship between the variables being studied, but is itself not studied.

A carefully designed **randomized experiment** can control for lurking variables, and causal relationships can be established. Typically, such experiments include:

- A control group and a treatment group.

- Random assignment of large number of observations into the treatment and control groups. Due to the random assignment, the general characteristics of of subjects in each group are similar.

Lurking variables are always an issue with **observational studies**. Researchers in observational studies do not intervene with the observations and simply observe the data that the observations generate. Causal relationships are much more difficult to establish with observational studies.

3.2.1 Questions

1. Consider the `palmerpenguins` dataset that we have been working on. The data contain size measurements for three different species of penguins on three islands in the Palmer Archipelago, Antarctica over three years. Is this an observational study or randomized experiment?

2. A fertilizer company wishes to evaluate how effective a new fertilizer is in terms of improving the yield of crops. A large field is divided into many smaller plots, and each smaller plot is randomly assigned to receive either the new fertilizer or the standard fertilizer. Is this an observational study or randomized experiment?

3. A professor wishes to evaluate the effectiveness of various teaching methods (traditional vs flipped classroom). The professor uses the traditional approach for a section that meets on Mondays, Wednesdays, and Fridays from 9 to 10am and uses the flipped classroom approach for a section that meets on Mondays, Wednesdays, and Fridays from 2 to 3pm. Students were free to choose whichever section that wanted to register for, with no knowledge of the teaching method being used. What are some potential lurking variables in this study?