HOW TO WRITE A CASE STUDY

When you write up a case study for this class, organize your writing into the following sections. Each section should have its own header (matching the headers given here), and within the section, the writing should be prose (meaning complete sentences, not lists of items, etc.).

Introduction (no section header)

The introduction is the only section that shouldn’t have a section header, since it comes first and doesn’t need one.

In the introduction, state the question(s) of interest that motivate the paper. Then state what the paper will cover and when. The first of these helps the reader know what the paper will be about, and the second helps the reader navigate the paper, jumping directly to specific parts.

Question of interest

State the question(s) that motivate the study clearly and unambiguously. These questions should arise from real-world situations. They should be phrased in real-world language, not statistical language.

Study design

Describe what you did in order to study your question(s) of interest.

If you collected your own data, then describe the process that you used for that. However, usually in this class, instead of collecting our own data we will use available data, which is data that was collected by someone else for a purpose besides our study. If you use available data, state that and describe the data collection process. If there are important parts of the process that you don’t know, state those and explain why that missing information would be helpful to your study.

Cite your data source with a bibliographic entry, whether the source is yourself or someone else. If the data is from a secondary source, cite both the primary source and the secondary source. If you don’t know the primary source, state that and cite only the secondary source.

Define the variables that you use. Explain the random process behind any random variables (such as the response variable). This should include a discussion of what the origins of variability in that variable are. (Such variability is necessary in order for statistical inference to be necessary.)

For each numerical variable that you observe, state the units that you will measure it in. Also, state the levels of each categorical variable that you observe. (It isn’t important how those levels are coded in your data file, but you should state what they are in real-world terms.)
Define the **parameters** that will be involved in your statistical inference. This includes giving the **model formula** and **true model equation** that you are postulating for your study.

In terms of these variables and parameters, state the specific **statistical task(s)** that you will do to investigate your question(s) of interest. Often this will include conducting a hypothesis test and/or computing a confidence interval for a parameter.

Also, in this section, state explicitly:

1. What an **individual** is,
2. To what **population** your study extends and why,
3. What **treatments**, if any, are being applied to individuals in the study,
4. Whether your results will carry much **causative weight**, and why.

For studies involving models with more than one explanatory variable, discuss any **relationships among variables** that might affect your study. In particular, this should include a discussion of whether you expect multicollinearity based on what the explanatory variables represent.

### Analysis

Cite R with the information from the `citation()` function, and cite any R packages that you use with the information from the `citation(package="myPackage")` function.

Analyze your data with the following steps:

1. **Explore the data** with graphical and/or numerical tools. If you see anything unusual or noteworthy, state that and investigate it. If you have more than one numerical explanatory variable, look for signs of multicollinearity. If you see such signs, discuss the severity of the apparent multicollinearity and the implications of it for your analysis.

2. Check the **model fitting assumptions**:
   
   (a) The **model is appropriate**, and
   
   (b) The **observations are independent**.

The appropriateness of the model is established by considering what the data represents: is the postulated model a reasonable description of the true real-world situation? Knowledge of the subject field that the data comes from is often helpful here.

Observations are seldom, if ever, truly independent. However, you should verify that there are no obvious relationships among the observations. For example, if you have several observations of each of several individuals, the observations of each individual are related in violation of this model fitting assumption.

If the model fitting assumptions are violated mildly, you should state that and proceed with the analysis. If they are violated severely, you should not fit the model, since it won’t give you any useful information. (The computer will be unaware of this problem and can fit the model as usual, but the model cannot validly be interpreted as usual in this case.)
3. **Fit the model** and give the fitted model equation. Be sure to include the units on all the coefficients (unless a variable has been log transformed, in which case you may omit the units at this point.)

If your model has few enough variables that you can display regression line(s), **produce a scatterplot with regression line(s).**

If any of the fitted model coefficients gave you useful information for your analysis that isn’t immediately apparent from them, **explain what you deduced from the fitted model coefficients.** (For example, in a 1-way ANOVA, the fitted model coefficients can be used to obtain point estimates of the response variable’s random variable mean within each level of the categorical explanatory variable.)

As a routine part of fitting the model, you should assess the goodness of fit by computing and interpreting the **coefficient of determination** $R^2$.

4. If there is more than one numerical explanatory variable, look for signs of multicollinearity. This should include computing the variance inflation factor for each variable’s coefficient in a model omitting the categorical explanatory variables. If you see any signs of multicollinearity, discuss their severity and whether or not they will impact your study.

5. Check the **sampling variability assumptions**:
   - (a) The conditional error terms are **normally distributed**,
   - (b) The conditional error terms all **have equal variances**,
   - (c) The conditional error terms all have **zero mean**, and
   - (d) The conditional error terms are all **independent**.

We check the first of these by inspecting a normal quantile plot of the residuals. In a normal quantile plot, points that are observations of a normally distributed random variable will tend to lie approximately in a line. If the residuals do not deviate strongly or systematically from a line in such a plot, then they don’t show evidence that the conditional error terms are non-normally distributed.

In most cases, we check the last three of these by inspecting a standardized residuals versus fitted plot. Violations of the equal variances assumption would show up as systematically different vertical ranges for different locations along the $x$ axis. Violations of the zero means assumption would show up if the residuals appear to have averages that are systematically different near different locations along the $x$ axis. Violations of the independence assumption may show up as systematic patterns among residuals (such as trending, etc.). In short, anything systematic in this plot is indicative of a violation of one of the last three sampling variability assumptions.

For a 2-sample $t$ test or 1-way ANOVA, to investigate the last three sampling variability assumptions, we instead use simultaneous density plots of the standardized residuals, grouped by levels of the categorical explanatory variable. The independence assumption isn’t ordinarily an issue (assuming that the observations are independent, as in the model fitting assumptions); nor is the zero means assumption. However, violations of the equal variances assumption will show up as the density plots having markedly different spreads.

If any mild violations of sampling variability assumptions are apparent, you should comment on them and proceed. If, however, severe violations of sampling variability assumptions are
apparent, you should not use the fitted model for statistical inference. Instead, you should look for another type of model, such as a nonparametric model (which we don’t cover in this class), if you want to conduct statistical inference.

6. **Conduct statistical inference.** Usually this will involve conducting a hypothesis test or computing a confidence interval.

If you conduct a hypothesis test, **interpret your p-value** by stating in precise statistical terms what you did or did not find statistically significant evidence of. As part of this, state in parentheses: the value of the test statistic, the degrees of freedom of the test statistic, and the p-value of the test.

If you compute a confidence interval or a prediction interval, state in precise statistical terms what parameter you are estimating and what your point estimate of that parameter is, including units. Then in parentheses give your confidence or prediction interval as well, again including units.

**Discussion**

Translate your results back into the real-world setting. Explain what you learned about your question(s) of interest. This should involve paraphrasing in less technical (but still correct!) terms what **statistical inference procedures** you carried out, and what their **results** were.

Again state the **population** to which your results apply, and whether or not they carry much **causal weight**.

**Discuss the limitations** of your study frankly and openly. The more that you are open and careful about the possible limitations of your study, the more credible your study is.

If your study suggests any **further questions**, state those as well.

**Bibliography**

Your paper should have **in-text citations** to entries in the bibliography. I am not particularly concerned with the exact format of the bibliography; just make sure that it is easy to find the sources that you use from the information given there.

However, be sure that all **website citations** have **access dates**. The internet changes, so stating when the website was accessed is important.