Writing Appropriately for Linear Regression

Amanda Kay Montoya, akmontoya@ucla.edu, @AmandaKMontoya

My philosophy: While there are often many statistics courses required for both undergraduate and graduate psychology degrees, the reason for this requirement is often to ensure 1) appropriate implementation of data analysis techniques for research, and 2) appropriate interpretation and reporting of results of analysis. This second one is often overlooked in statistics courses, by assuming that if you did the statistics correctly you will report them correctly. I do not believe this is always the case, and I use this course as an opportunity to provide feedback on reporting and interpretation of statistical results, with the primary goal of improving your writing skills in this area for when you are writing research papers. As such I have a number of policies that should be observed when writing for this class (and for the rest of your research career). This document attempts to describe these issues but is not meant to describe all possible ways in which someone might incorrectly answer questions in the course.

Interpretation

- An interpretation is a translation of a numerical result to contextualize it's meaning within the example/data set/analysis model in which it has been estimated.
- Interpretations differ from definitions (i.e., a question which asks you to define a concept vs. one which asks you to interpret a numerical result) in that an interpretation should be specific to the example at hand, and a definition should apply to any possible example and does not typically involve the specific variables and concepts from an example.
- Interpretations should be specific to the number (not just the sign or vague magnitude) being interpreted. A good interpretation will embed the number in the interpretation.
    - Good example "Two people who differ by 1 year on education are expected to differ by $4,500 in yearly income, where the person with more education is expected to have greater income."
    - Bad example "People who are higher on years of education are expected to have higher yearly income".

***Interpretations of Regression Coefficients*:** This is a regression class, and by the end of it you will have mastered the interpretation of regression coefficients in a variety of contexts. There are many pitfalls that researchers often fall into in interpreting regression coefficients. This section will help you identify and avoid those pitfalls:

- **Avoid Within-Person Interpretations**: Interpretations which imply that we know what would happen if we changed an individual person (or case) when the data are completely cross-sectional (i.e., no repeated measurements of people or longitudinal data). These kinds of interpretations should be avoided.
    - Example: "We expect a change of $4,500 in yearly income, for each additional year of education." This could imply a within-person change, which is not appropriate if we are analyzing cross-sectional data.

    - Notice how awkward this sounds for a concept that we think of as unlikely to change: "For a one-unit change in gender, we expect a $10,000 change in income." If this study was examining individuals before and after gender transitions, this statement might be feasible, but it is not accurate when describing gender differences in income among individuals who have not *changed* gender.
    - Other words which often imply within-person change and thus should be avoided: increase, decrease, change, improve, decline
    - The term **difference** can often be used to replace change or other words that imply within-person differences. "For a one-unit difference in gender, we expect a $10,000 difference in income."
- **Avoid Causal Interpretations**: When predictor variables are not manipulated we cannot use causal language about their relationships. While language can be ambiguous about what is causal and what is not, I list here some words which should be avoided for the purposes of this class.
    - **Avoid:** influence, cause, impact, increase, decrease, affect
    - **Ambiguous (use carefully):** attenuate, modulate, additional
    - **Non-Causal:** association, relation, greater than, less than, higher, lower
    - **A special note on the word "effect":** The word effect is used in two contexts in statistics and it is important to differentiate these two in your writing. The difficulty is that one of these implies causality while the other is not necessarily causal. The first definition of effect focuses on causality: "change which is a result or consequence of an action or other cause." For example, in your writing, you might say something like "This study tells us about the effect of social media use on depression in teens." This, I would say, is inappropriate if you do not have an appropriate design to evaluate a causal claim. However, the second definition of effect is used in a more statistical context, for example in "effect size", which is defined as "a value measuring the strength of the relation between two variables". Here effect does not necessarily mean "cause" but rather "relation". In these cases, I believe it is reasonable to say things like "b_1 quantifies the effect of X_1 on Y controlling for X_2". Would it be less ambiguous to say "b_1 quantifies the relation between X_1 on Y controlling for X_2"? Yes, I think so. But is the first quote incorrect? No, I don't think so. Ultimately, be careful about using the word effect when you are not referring to a direct statistical estimate. Please also note that this is a topic on which others may not agree with me, so you might consider how to be careful about using this language as you move beyond this class. Do not be surprised if you find folks out there in the world who think it's never okay to use the word "effect" in a non-causal sense, and be prepared to defend your position if you choose to disagree with them.
        - An example where the word "effect" sounds causal to me: "There were no effects of socioeconomic background on endorsement of the follow-your-passions ideology". A recommended re-write would be

"There was no relation between socioeconomic background and endorsement of the follow-your-passions ideology."

- **Avoid Deterministic Interpretations**: Regression coefficients are an estimate from a sample, this means that they are descriptive of the sample, but not individual people. A deterministic interpretation often overlooks making clear that the coefficient is an estimate or an average. For example, "Two people who differ by one year on age will differ by $2,000 on income." This interpretation is **deterministic** because it does not allow for individual variability or uncertainty. A more appropriate interpretation would be, "Two people who differ by one year on age are estimated to differ by $2,000 on income" or "…are expected to differ by…". It is important not to ignore the role of individual variability in the interpretation.
- You should not use variable names (from the data) in your interpretations. For example, "Two people who differ by 1 on EDUC are expected to differ by 4.5 on YRLY.INCOME". Good interpretations should embed both the conceptual meaning of the variable and the units of the variable: "Two people who differ by 1 year on education are expected to differ by $4,500 on yearly income."
- Avoid unclear directionality for one variable but not the other. Examples include "gender is related to higher depression", "reaction time is correlated with high numeracy." In both of these examples it is unclear the directionality of one of the variables (gender, reaction time). These can be revised to include directionality: "Being a woman (compared to being a man) is related to higher depression" or "slower reaction time is linked with higher numeracy."
- In multiple regression when we interpret regression coefficients, we need to articulate what variables are being "controlled for" or "partialled out". There are two parts of this which can be difficult 1) articulating exactly which variables are being controlled for, and 2) using appropriate (but perhaps not overly repetitive language).
    - In your interpretations you should always say exactly which variables are being controlled for. You should not say….
        - "controlling for all other variables in the model" or any other similar articulation, because this is actually not true! The outcome variable is a variable "in the model" and you are not controlling for it. Additionally, this puts the cognitive load on your reader, as they have to remember which other variables are in the model rather than you telling them. If you have a lot of variables in the model and need to write many interpretations, you might consider explicitly defining a set of variables as the "control variables" or "demographic covariates" or something of the sort. If you did this it would then it may be appropriate to use language like the following:
            - *Example:* In the following regression models the set of demographic predictors gender, age, and ethnicity were included in all models. These will be referred to as "demographic covariates"…. "Two people who differ by 1 year of education and are equal on all demographic covariates

are expected to differ by $4,500 in yearly income, where the person with more education is expected to have greater income." You can see though how this can be risky.

- There is a variety of language which can be used to refer to partialing out the effect of another variable. Some of this language is accurate, while other examples we can find to be misleading or confusing.
  - Recommended language:
    - Controlling for….
    - Partialing out….
    - Holding constant….
    - For two cases with the same score on…
    - Adjusting for…
    - Among individuals with the same…
  - Non-recommended examples:
    - Regardless of…
    - Above and beyond…
    - Ignoring…

Inferences

In this class you will frequently be asked to report inferences separate from interpretations. Your inferences should refer to **a conclusion about a population parameter**, whereas your interpretation focuses on **a sample estimate**. You will be required to differentiate these two ideas in your writing throughout the class. All inferential statements should be accompanied by supporting statistical information. This section outlines some major characteristics that are important in your inferences.

- What is the null hypothesis you are rejecting?
  - There are many "null hypotheses" and it is not enough to say "we reject *the* null hypothesis". You should, in your writing, specify the null hypothesis that you are rejecting. For example, "We reject the null hypothesis that the true partial regression coefficient for social media use predicting depression is zero." Alternatively, you could write "We conclude that the true partial regression coefficient for social media use predicting depression is not zero."
  - Two null hypotheses which are important to differentiate are: 1) A and B are unrelated, compared to 2) A and B are unrelated when controlling for C. One does not imply the other. Rejected of one null hypothesis does not imply rejection of the other. Make clear which of these hypotheses you are rejecting.
- What does "statistical significance" apply to? What is non-zero?
  - Be careful that you are referring to the **sample statistic** and not the population parameter. The sample statistic will essentially always be non-zero, but we want to evaluate if it is sufficiently far from zero to suggest that the result is

implausible if the population parameter is zero. We do not need a hypothesis test to determine whether the sample correlation or sample regression coefficient is different from zero, we can make that determination based on the sample itself. Claiming that a sample statistic is "statistically significant" is not an inference (it is an observation about the sample). The inference would be that the population parameter is non-zero (not "significantly different from zero"). Inference is about the population parameter and should be discussed as such.

- Interpreting *p*-values
    - In this class, you are instructed to conduct tests at a level of 0.05, unless otherwise instructed. You should not interpret *p*-values which are higher than the cut off as "marginal", "approaching significance", "tentative evidence", or any other similar description. Similarly, *p*-values which are much smaller than the α-level should not be interpreted as "highly significant", "very significant", etc. Merely remark on whether the *p*-value is above or below the α-level to determine statistical significance.
- **Conclusions:** Mapping back to the problem at hand
    - Frequently, while writing students will finish an analysis by making a statement about inference. However, there should be an extra step where this is generalized and explained. Frequently this would involve mapping back to the original hypotheses of the study. For example, "Therefore we conclude that social media use plays an important role in the development of depression amongst teens." In this class you will often be asked to state your inferential information AND make a conclusion. This conclusion should map back to your original hypothesis/research question.
- Not accepting the null hypothesis (how to fail to reject)
    - In this class, you will be expected to appropriately interpret analyses where we fail to reject the null hypothesis. In this case, you should under no circumstances accept the null hypothesis. Rather the evidence is inconclusive. This will often come into play in the conclusion, where failing to reject the null hypothesis should **not** result in a conclusion like "Therefore we conclude that social media use does not play an important role in the development of depression amongst teens." This is accepting the null. Instead, one might write something like "Therefore the results of the study were inconclusive for determining whether social media use plays an important role in the development of depression amongst teens." Resist the urge to imply that *if only we had a larger sample, we would have found something*. You don't know that for sure.

Statistical Reporting

As psychology researchers you will be required to analyze data and report those analyses in published papers. In order to better prepare you for those experiences, we're asking that you practice reporting in APA style in this course. In particular, for statistical tests this means reporting the test statistic (e.g., *F*, *t*, *Z*), associated degrees of freedom, *p*-value, and confidence interval (when applicable) for each statistical test. Estimates of effects can either be integrated

into a sentence (e.g., a one-unit difference in *X* results in a b_1 unit difference in *Y*) or included in the inferential information (e.g., The effect of X on Y was statistically significant (b_1 = estimate, *t*(df) = t-value, *p* = p-value). The APA guidelines for reporting *p*-values are as follows:

> When reporting *p* values, report exact *p* values (e.g., *p* = .031) to two or three decimal places. However, report *p* values less than .001 as *p* < .001. The tradition of reporting *p* values in the form *p* < .10, *p* < .05, *p* < .01, and so forth, was appropriate in a time when only limited tables of critical values were available. However, in tables the "*p* <" notation may be necessary for clarity (see section 5.16)."
>
> *Publication Manual of the American Psychological Association*, Sixth Edition, Page 114 Section 4.35

Never report a *p*-value as being zero. The probability of finding a specific effect under the null hypothesis is never exactly zero, but rather too small to portray in 3 decimal points (i.e., p < .001). If you have additional questions about statistical reporting please consult the APA Manual or your instructors or TAs for assistance.