HW5

Sam Olson

Assignment 5

Q1

Cell Means versus Additive Model

In class, we talked extensively about the two types of models when analyzing two treatment factors. For this question, consider your audience to be a first-year graduate student in a different discipline. Their understanding of statistics includes what you learned in Stat 5000 but they have not seen much Stat 5100 materials. The student is asking you for help trying to better understand:

a)

The two types of statistical models themselves.

At a high-level, first bear in mind nothing is fundamentally changing about the underlying data. Instead, these are different parametrizations of the data. So these are not diametrically opposed, and will actually provide the same estimates under certain conditions. Just different ways of modelling the relationship between variables/factors and the response/thing you're trying to predict.

Cell Means:

- We are focused on estimating the average response for each unique combination of factor levels.
- Every combination of factors has their own mean, with a focus on estimating individual effects.

Additive Model:

- Instead of estimating for each combination of factor levels, we are more interested in estimating the effects of the factors themselves.
- To do that, we assume the effect of a factor is consistent across levels, i.e. no unique interactions, interactions are "additive", hence the name.

b)

The difference between both types of statistical models.

Cell Means:

- Directly estimates separate means for each combination of factor levels, i.e. the "individual, cell means".

- Estimating parameters equal to the number of unique factor-level combinations.
- No assumptions about the relationship between treatment means.

- "More flexible", as we do not impose constraints like the Additive Model.

Additive Model:

- Decomposes what you're estimating (the β's) into an overall mean, plus main effects (plus interactions effects too).
- Estimating fewer parameters due to imposed structure (constraints).
- Assumes additive relationship between factors and their interactions.
- "Less flexible" (requires constraints such as sum-to-zero constraints).

c)

Which one they should use for their own experiment that they plan on carrying out studying the effect of two treatments on some response y.

Both *can* work, at least in theory! They *can* also lead to very similar results, especially if the researcher is concerned with a more general, *"is there an effect?"* style questions, and if some assumptions are met (consistent interaction effects, for example). This does not mean that the two different models are naturally interchangeable though!

There are some key distinctions to think about to determine if one seems more appropriate for the study. Also bear in mind, that this is a "study" rather vaguely. There is inherently a lot of subject matter expertise to consider, not the least of which being how the study is designed and whether the things being measured or treatments being used make sense or are interpretable.

That being said, some considerations for the researcher, a.k.a. questions I'd ask:

- Prior to collecting any data or running any experiment, what you believe about the underlying relationship you're trying to test?
- Do any of the underlying assumptions of the two models readily incorrect?
- Are you hoping to predict things, e.g. we expect an individual with particular characteristics to have y response, or are you more interested in the broad characteristics of your population of interest?

Q2

The Surprising Power of Reflection

a)

Watch the following video: The surprising power of reflection

b)

Statistical Flaw in the First Study

The video showcases two studies: one immediately in the beginning within the first minute of the video and a second one introduced in the last 45 seconds of minute two. Listen carefully to the descriptions of each study. What *flaw*, statistically speaking, does the first study suffer from that does not show up in the second study? Briefly explain.

Overall, I think the statistical concept is a *randomized study design*, which does not exist in the first and is being used in the second.

Details:

In the study shown in the beginning of the video participants selected which group they belong to, i.e. chose to practice or to reflect. By contrast, participants in the second one were assigned their group, i.e. "treatments" were assigned and the sample under study was divided into a "control" group and an "experimental" group.

c)

Reflection on Learning

In 2012, I took a workshop at ISU with Dr. Jan Wiersema called Project LEA/RN. Dr. Wiersema shared the following advice with us during the workshop – it has stuck with me ever since:

> **It's the thinking about the doing that does the learning.**

Reflect on, and briefly summarize how *you* best learn new things. Share one or two tips on how you deal with challenging course material to ensure you learn it. You can reference experiences you have made as a student since joining our program but you can also reference an experience at some other time in your life.

Overall: I do best by having a mix of "alone" and "together" time. My general approach, if I have one, is to first give it a try without any assistance, references, or guidance. After that, I give myself some time to process, identify loose-ends/dead-ends, or walls I came up against and potential resources that could help, e.g. *"Wasn't there a formula on a slide we recently covered for that?"*. After reflecting, I then revisit, give it another go. Only after then do I feel comfortable discussing/sharing with others, with a focus on trying to *"knowing what I don't know"*. Sometimes I "rubber duck" it too, where my collaboration is to try explaining a topic, question, solution to an inanimate object.

None of this is prescriptive though, not even for myself. Had I to think of an approach I use and tends more often than not to work for me, the above would be it.

Tips:

- Have a constructive relationship with failure. I remember more problems I got wrong than problems I got right.

- Confusion and difficulty are the norm, especially when learning something considered "advanced" like grad school.
- Related, remember: A minority of people know how to calculate a derivative. I'd also wager even more people were ecstatic when they finished their last math class. . . in high school, sometimes undergrad.

- Sometimes the best thing to do is step away. It doesn't mean you're abandoning it, just that you need some time before trying again.

Q3

MS Exam Repository

Go to the old **MS exam repository** and look at the Methods I and II questions. Familiarize yourself with the questions in Methods I – you should have an idea on how to answer most questions that are part of Methods I. Familiarize yourself with the questions in Methods II – these you cannot answer yet, for the most part. Select one Methods II question you find intriguing and download the question document.

If you are in the PhD program in Statistics, in addition, pick and download a Methods II question from the old **PhD exam repository.**

a)

Selected Questions

Which questions did you pick? Answer by following this format:

YEAR Methods II MS repository (and *YEAR Methods II PhD repository* if you are a PhD student).

MS: 2020 Part IV

PhD: 2020 Part I

b)

Submission of Selected Questions

Submit your question(s) as part of the homework. Note that I am not asking you to solve the question(s) (*yet*); I just want you to familiarize with them so you have an idea about expectations. For the new PhD qualifying exam, expect the level of difficulty to fall approximately in between old MS exam and old PhD exam questions.

Part IV:

Crude oil is a mixture of many different hydrocarbon compounds. Some compounds have large effects on fish growth rate; other compounds have small or no effect. Different samples of crude oil can have very different compositions. The investigators collected 60 different crude oil samples and dissolved each in water. The concentration of 52 hydrocarbon compounds was measured in each of the 60 water samples. Each water sample was divided into 4 containers and one fish added to each container. As before, initially all fish were the same age and similar size. After 10 days, each fish was weighed. No fish died. There are 240 observations in the data set.

The 52 measured compounds are named X1, X2, $\cdots$ X52. All concentrations were ln transformed before model fitting. All subsets regression was used to find models (one, perhaps more) that will predict fish growth in new observations. Model selection statistics from some models are in Table 4.

11. The goal is to predict fish weight for new observations. Based on the information in Table 4, which model is most appropriate model? Briefly explain your choice.

Table 4: Model selection statistics for some models to predict fish weight from ln transformed hydrogen concentrations.

Model #	Variables in model	# variables	R^2	AIC	BIC
1	X4 X7 X11 X24 X27 X31 X50	7	0.667	-244.46	-220.09
2	X4 X7 X24 X27 X31 X40 X50	7	0.666	-243.98	-219.61
3	X4 X7 X11 X27 X44 X50	6	0.659	-240.49	-219.35
4	X4 X7 X11 X27 X50	5	0.650	-243.54	-219.18
5	X4 X5 X7 X9 X10 X11 X20 X21 X24 X27 X31 X50	12	0.690	-251.87	-210.11
6	X4 X7 X11 X17 X20 X24 X27 X31 X35 X40 X44 X50	12	0.689	-251.12	-209.35
7	all 52 variables	52	0.726	708.67	896.63

12. Based on their knowledge of the 52 hydrocarbon compounds, the investigators know that X15 strongly reduces fish growth and X25 has no effect. They expected Table 4 to include models with X15 but not X25. List two potential reasons why X15 might not "discovered" by an all-subsets model selection algorithm.

13. The results in Table 4 are based on 240 observations, i.e. 4 fish for each of the 60 crude oil samples. Is this appropriate or is there an issue such as lack of independence or overfitting? Briefly explain your answer.

Figure 1: MS

Part I

Researchers conducted an experiment to learn about the effects of two feed types (1 and 2) on weight gain in pigs. Ten pens, each containing four pigs, were used for the experiment. Each pen contained a single container, called a feeder, in which researchers placed feed for distribution to the four pigs in a pen. A completely randomized design was used to assign feed type 1 to five pens and feed type 2 to five pens. Feed of the assigned type was placed in each pen's feeder each day for a three-week period. Each pig was weighed at the beginning and end of the three-week period. Let y_{ijk} be the weight gained by kth pig in the jth pen treated with feed type i ($i = 1, 2; j = 1, \ldots, 5; k = 1, \ldots, 4$). Assume the model

$$y_{ijk} = \mu + \phi_i + p_{ij} + e_{ijk}, \tag{1}$$

where μ, ϕ_1, and ϕ_2 are unknown parameters, $p_{ij} \sim N(0, \sigma_p^2)$ for some unknown variance parameter σ_p^2, $e_{ijk} \sim N(0, \sigma_e^2)$ for some unknown variance parameter σ_e^2, and all p_{ij} and e_{ijk} terms are independent.

The weight gain data were stored in a vector `y` in `R`, along with information about the feed treatment and pen provided in factors `feed` and `pen`. The code and output on page 2 contains information useful for answering the problems **1** through **4** below. Note that the end of page 2 provides quantiles of t distributions.

1. Model (1) may be written in the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$, where $\boldsymbol{y}$ is the vector of y_{ijk} values (ordered as in the `R` code), $\boldsymbol{\beta} = (\mu, \phi_1, \phi_2)'$, $\boldsymbol{u} = (p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{21}, p_{22}, p_{23}, p_{24}, p_{25})'$, and $\boldsymbol{e}$ is the vector of e_{ijk} values (ordered to match the order of $\boldsymbol{y}$). Using Kronecker product notation, provide expressions for $\boldsymbol{X}$ and $\boldsymbol{Z}$.

2. Provide the value of an unbiased estimator for σ_p^2.

3. Determine the value of the F statistic you would use to test $H_0 : \phi_1 = \phi_2$.

4. Find a 95% confidence interval for $\phi_1 - \phi_2$.

Figure 2: PhD Pt1

R Code and Output for Part I

```
> length(y)
[1] 40

> y[c(1:4, 37:40)]
[1] 30.6 23.6 30.6 29.6 29.6 28.0 32.5 29.1

> feed
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
[32] 2 2 2 2 2 2 2 2
Levels: 1 2

> pen
 [1] 1  1  1  1  2  2  2  2  3  3  3  3  4  4  4  4  5  5  5  5
[21] 6  6  6  6  7  7  7  7  8  8  8  8  9  9  9  9  10 10 10 10
Levels: 1 2 3 4 5 6 7 8 9 10

> anova(lm(y ~ feed + pen))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
feed       1 119.02 119.025 11.8894 0.0016944 **
pen        8 418.02  52.252  5.2195 0.0003843 ***
Residuals 30 300.33  10.011

> mean(y[feed == 1])
[1] 31.69

> mean(y[feed == 2])
[1] 35.14

> round(qt(.975, 1:40), 3)
 [1] 12.706  4.303  3.182  2.776  2.571  2.447  2.365  2.306
 [9]  2.262  2.228  2.201  2.179  2.160  2.145  2.131  2.120
[17]  2.110  2.101  2.093  2.086  2.080  2.074  2.069  2.064
[25]  2.060  2.056  2.052  2.048  2.045  2.042  2.040  2.037
[33]  2.035  2.032  2.030  2.028  2.026  2.024  2.023  2.021
```

Figure 3: PhD Pt2

c)

Reflection on Learning

Reflect on your Fall semester; what has helped you learn and why? Note that it is the *why-part* of your answer that I am most interested in.

What helped: Sharing my work, asking for feedback, and proactively starting discussions with members of the cohort. I have a number of people I email drafts of my problemset, with the hopes of seeing how they compare.

Why: I think there are two components to the why of this. Incomplete as an answer, as it will ever be.

(1): In order to explain an answer, I need to know more than it takes to have *an answer*. I sometimes forget a key concept, have a typo causing massive downstream issues, or simply enough just misunderstood what a question was asking. Improving is collaborative and goes both ways. Getting other perspectives helps me not only with the end goal, getting an answer, but also with the journey, trite as I feel it may be to say. The process affirms my own knowledge and my confidence.

(2): My primary struggle, an ongoing one, is having started graduate school "late". I was in industry for roughly 7 years after completing undergrad. I don't regret having spent that time away from academia as it was formative, but that is a source of imposter syndrome and FOMO. These anxieties are not unique to me though, and realizing that has been very relieving. Sharing something as direct as, *"Major hand waving at Q5, anyone have any better ideas, or whether I'm wrong there?"*, and hearing back... well, anything, lets me know I'm not alone. I'm not the only one struggling, and I'm not the only one looking for help.

References

- Di Stefano, Giada and Gino, Francesca and Pisano, Gary and Staats, Bradley R., Learning by Thinking: How Reflection Can Spur Progress Along the Learning Curve (February 6, 2023). Harvard Business School NOM Unit Working Paper No. 14-093, Kenan Institute of Private Enterprise Research Paper No. 2414478, Available at https://dx.doi.org/10.2139/ssrn.2414478
- Learning by Thinking: How Reflection Can Spur Progress Along the Learning Curve