Exam 2 – Math 150

Ethan Ashby

due after 48 hours ~ May 2, 2021

PLEASE

Knit this file before doing anything! Make sure that the data load properly and that the packages work.

Guidelines

The exam is not be timed. You will pick up the exam sometime between Wednesday and Friday (from Gradescope), and it will be due 48 hours later. The general idea for what is allowed and what isn't is: all the class materials are allowed; all the non-class materials aren't allowed. Don't talk to anyone except the professor about the exam.

You may access the following:

- talking to **me**
- course textbook
- online course notes
- any notes you personally have written
- course videos
- applets
- HW / WU assignments & solutions
- the help files in R, e.g., `?geom_point()`
- previous Discord conversations
- posted sample exam and solutions
- graded assignments (on Gradescope)

You may not access the following:

- talking to **others**
- Google / online searching
- other textbooks / online materials
- other videos on the content

If you have any questions about R or any other types of questions whatsoever (even if they seem silly!), please, please, please email or DM me on Discord. I **love** the questions, and I'll try to answer as quickly as possible. If it doesn't seem like a fair question to ask, I'll just say that I can't answer it. No problem. But please ask!

Q1. Fruitflies

Consider the fruitfly data used in a few of the survival analysis homework assignments.

(a) (+10 pts) Address the question pertaining to whether or not `Partners` should be used as a linear explanatory variable in the Cox PH model. You will need to run two separate models (one with `Partners` coded numerically, one with `Partners` coded as a factor variable, ask if you need hints for running the R code!). [This question is not about the significance of the coefficient, so feel free to ignore the p-value here. Although you are free to add the idea of significance to your answer if you would like.]

Should `Partners` be coded as numeric or as categorical in the Cox PH model?

Hint: in order to answer the question, you will need to make (at least) two numerical comparisons of the coefficients.

I'll preface my solution by saying that `Partners` only has three unique values (0,1,8), which may be an insufficient number of unique values to determine with high confidence if a variable should be encoded continously or categorically. But to get an idea of what we should do, I fit two Cox PH models, one with the `Partners` variable encoded as a continous (numeric) variable and the other with `Partners` encoded as a factor variable.

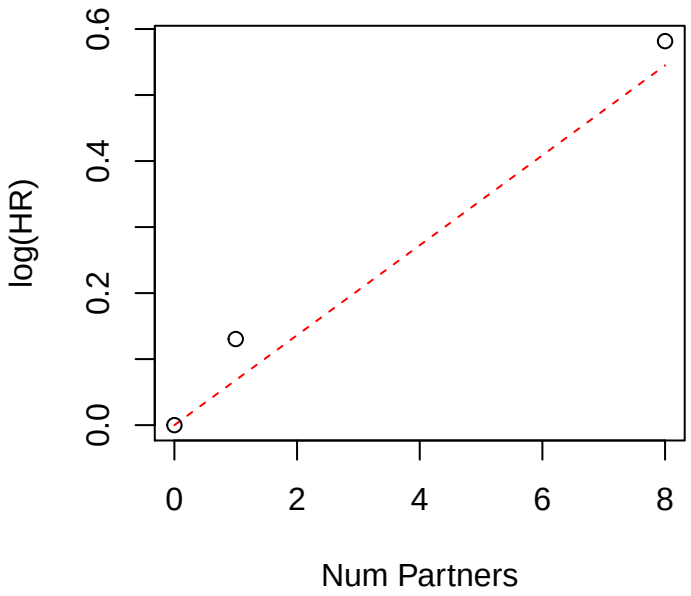
```
fruitfly <- read_csv("https://pomona.box.com/shared/static/qnsl0sp0twdutz6azidxb5yt37boee7v", na="*")
#see what values of partners we're working with
#fruitfly$Partners %>% unique()

#make factor variable
fruitfly$Partners_fac<-factor(fruitfly$Partners, levels=c("0","1","8"))

#run coxph model with factors
coxfac<-coxph(Surv(Longevity, Censor)~Partners_fac, data=fruitfly)
coxfac %>% tidy() %>% kable()
```

term	estimate	std.error	statistic	p.value
Partners_fac1	0.1303901	0.2472203	0.5274249	0.5978986
Partners_fac8	0.5816500	0.2503367	2.3234710	0.0201539

```
#run coxph model with continuous
coxcont<-coxph(Surv(Longevity, Censor)~Partners, data=fruitfly)
coxcont %>% tidy() %>% kable()
```

term	estimate	std.error	statistic	p.value
Partners	0.0681148	0.0252824	2.694157	0.0070567

```
plot(x=c(0,1,8), y=c(0, coxfac %>% tidy() %>% .$estimate), ylab="log(HR)", xlab="Num Partners")
lines(x=seq(0, 8, by=0.01), y=coxcont %>% tidy() %>% .$estimate*seq(0, 8, by=0.01), col="red", lty=2)
```



If partners should be encoded continously, the β estimates produced by the factor model to match the β estimates associated with the continous model. In other words, the log(HR) associated with the factor levels 1 and 8 would match the log(HR) of 1 and 8 unit increases (relative to the baseline) in the number of partners based on the continuous model.

Shown in the plot above are the log(HR) for the factor model (black dots), compared against the log(HR) for the continous model (dotted red line) for various values of partners. In general, the black dots roughly track the dotted red line, indicating that we should encode the `Partners` variable as a **continous** variable.

(b) (+4 pts) Notice that the fruitfly data are all complete (that is, none are censored). Give two reasons why one might want to use a survival approach to the dataset, even in the setting of complete data.

1. The fruitfly dataset contains time-to-event data (encoded as `Longevity` variable), and survival analysis is designed precisely for the analysis of time to event data. Survival analysis methods have no requirement that observations be incomplete/censored. So survival analysis is a theoretically tenable approach to analyzing the fruitfly dataset.
2. A survival approach offers a principled way to determine which covariates are important for determining risk of death in flies, determine the effect size of each covariate relative to the baseline risk, and can help test particular scientific questions (for example, like whether flies with more partners live longer than flies with fewer partners).

Q2. VA Lung

Consider the VA Lung Cancer data used in a few of the survival analysis homework assignments.

```
VAlung <- read_csv("https://pomona.box.com/shared/static/r6hoo1gawopkt0526xvwwze5fl3245de", na="*")
```

(a) (+6 pts) Recall that the censoring indicator refers to survival / not survival (with an associated time).

- Provide one advantage to using Cox PH model to address whether treatment is significantly associated with survival.

- Provide one advantage to using logistic regression to address whether treatment is significantly associated with survival.

The Cox PH model is well-suited for survival (time to event) data with missing event times (censored observations). In contrast, logistic regression only considers a binary response variable, cannot account for the time-to-event format of the data, and cannot handle missing/censored data.

A logistic regression model $\left(\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right)$, can estimate the baseline probability of an event (i.e. of survival/death in this case) by fitting the β_0 parameter. In contrast, Cox PH does not explicitly model the baseline hazard.

(b) (+3 pts) Can you use a likelihood ratio test to compare a logistic regression model (on survival) with the proportional hazards survival analysis model to see which model is a better fit to the data? If so, write down the null and alternative hypotheses. If not, explain why not.

No I cannot make this comparison, because LRTs are typically used for comparing nested models, but Cox PH and the logistic regression are fundamentally different models with different response variables. Logistic regression models a binary (0/1) outcome variable, while Cox PH models time-to-event data (i.e., the event times + a censoring indicator). Trying to compare Cox PH and logistic regression using a LRT is an apples and oranges comparison that should not be performed.

(c) (+10 pts) At one point Karnofsky score was broken down into "low" (≤ 60) and "high" (> 60). By calculating appropriate HR (using a Cox PH model, find the HR, not the log of the HR), argue that Karnofsky score (coded as binary) and treatment **interact** (in predicting survival). In your argument, you should be clear what you mean by interaction. For each HR that is calculated, specify exactly what the group comparison is (that is, what is in the numerator and what is in the denominator).

To start, I encode both the `karno` and `trt` variables as categorical variables: `karno_c` and `trt_c` respectively.

```
VAlung$karno_c<-factor(as.integer(VAlung$karno>60), levels=c(0,1))
VAlung$trt_c<-factor(VAlung$trt, levels=c(1,2))
coxp2<-coxph(Surv(time, status)~karno_c*trt_c, data=VAlung)
coxp2 %>% tidy() %>% kable()
```

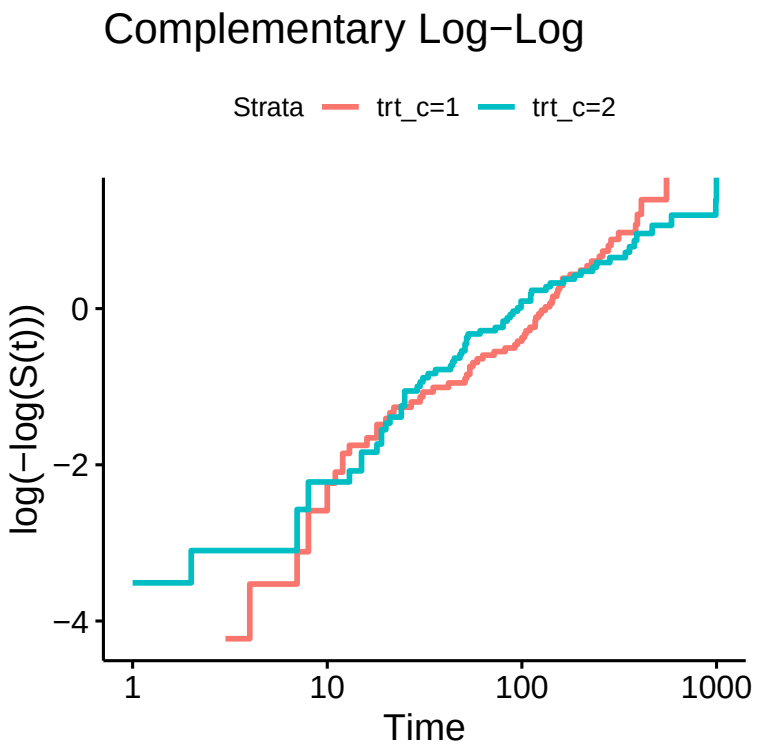
term	estimate	std.error	statistic	p.value
karno_c1	-0.4606063	0.2561043	-1.798510	0.0720962
trt_c2	0.4599960	0.2355719	1.952677	0.0508578
karno_c1:trt_c2	-0.7610580	0.3704291	-2.054531	0.0399244

The output of the interactive model tells us that the interaction between `trt_c` and `karno_c` is significant (per p-value=0.03992). The hazard ratio associated with being in the treatment (test therapy) group (relative to not being in the control/standard chemo group) is $\exp(0.4600) = 1.584$. The hazard ratio associated with being in upper karno score group (relative to being in the lower karno score group) is $\exp(-0.4606) = 0.6309$. The hazard ratio associated with being in the experimental treatment group AND being in the upper karno score group (relative to being in the control and lower karno score group) is $\exp(0.4600 - 0.4606 - 0.7611) = 0.4669$. This illustrates that in isolation, high karno score reduces the risk of death relative to the baseline and the experimental treatment increases death risk relative to baseline. However, in people with high karno score who recieved experimental treatment, death risk was greatly attenuated relative to the baseline risk.

(d) (+5 pts) Do the survival curves (in a **transformed** log-log space) indicate that a Cox PH model was

the appropriate model to use to measure interaction? Be sure that your plot addresses the question of using Cox PH to assess the interaction model (again, both treatment and Karnofsky score as binary variables). Explain.

```
ggsurvplot(survfit(Surv(time,status) ~ trt_c, data=VAlung),
           censor=F, conf.int=F, fun="cloglog") + ggtitle("Complementary Log-Log")
```



A pre-requisite for using the Cox model is proportional hazards, meaning that covariates exert a constant relative risk over time. The PH assumption is violated if survival curves cross. Since the log-log transformation is monotonic, crossing of survival curves in the log-log space also denotes a violation of PH. The survival curves for the treatment groups (shown above) cross, indicating that the effect of treatment interacts with time, and the proportional hazards assumption is violated. Since the PH assumption is violated, using the Cox model to assess the significance/effect size of any term related to treatment, including the interaction between treatment and Karno score, is theoretically untenable.

Q3. FWER and FDR

(a) (+5 pts) When hoping for high power, would you prefer to adjust your p-values using Holm or Benjamini-Hochberg? Explain.

Benjamini-Hochberg is a higher power method for p-value adjustment, as it is more likely to identify true (alternative) relationships than the Holm method. This is because the Holm method controls the Family Wise Error Rate, which is the probability of making *at least one* type 1 error. Benjamini-Hochberg controls the false discovery rate, or the proportion of significant tests that yield type 1 errors. Thus, the Holm method is much more conservative when it comes to type 1 errors, and may therefore miss out on some true relationships that Benjamini-Hochberg will identify. So Benjamini-Hochberg has higher power.

(b) (+4 pts) Given your choice, and after adjusting your p-values + writing up your results, what did you "control"? Explain.

By choosing Benjamini-Hochberg, we controlled the False Discovery Rate, i.e. the proportion of significant tests with a true null hypothesis. So if we controlled the FDR at 0.05, this would mean that we would expect 5% of the reported "significant"" results to actually be false discoveries, i.e., erroneous rejections of true null hypotheses.

Q4. Survival curves

Assume that you have a Cox Proportional Hazards model, but that you happen to know the baseline hazard, which is constant. That is:

$$h_0(t) = 0.2$$

Recall that the survival function for a set of explanatory variables (X) can be written as a function of the hazard function. Here, the baseline hazard is not a function of t, so I did the integration for you (feel free to check my work or not). For context, let's say that time is measured in years after surgery (ranging from 0 to 10 years).

$$S_x(t) = e^{-h_x(t)\cdot t}$$

Please ask if you don't understand the notation above.

(a) (+10 pts) Given the following Cox PH regression output, find the survival function (as a function of time) for someone who is 47 years old and is on drug 1 (as opposed to drug 0):

	coef	exp(coef)
age	0.00942	1.0095
drug	0.11859	1.1259
age:drug	-0.00670	0.9933

age is measured in years

First, let's calculate the hazard for this fine individual:

$$h_X(t) = h_0(t)e^{\sum \beta_i x_i} = 0.2e^{47*0.00942+1*0.11859-0.00670*47*1} = 0.2e^{0.24643} = 0.2558899$$

Then let's find the survival function by plugging in the formula for the hazard:

$$S_x(t) = e^{-0.2558899\cdot t}$$

(b) (+5 pts) Find the median survival time for 47 year-olds on drug 1. That is, the time at which half the people have survived.

To do this, we set the survival function = 0.5 and solve for t:

$$S_x(t) = 0.5 = e^{-0.2558899\cdot t}$$

$$\frac{\ln(0.5)}{-0.2558899} = t$$

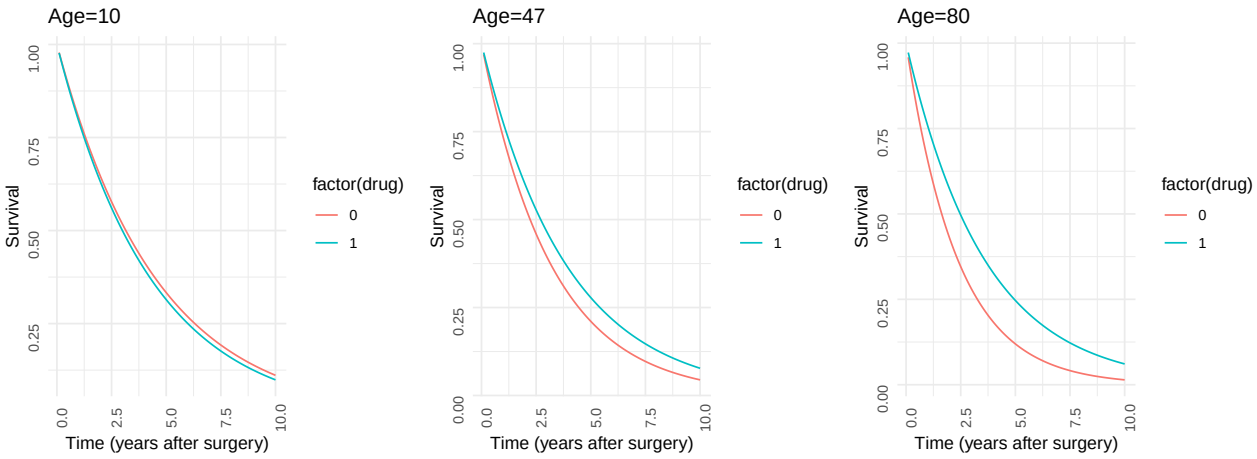
$$2.708 = t$$

So the median survival time for a 47 yo on drug 1 is 2.708 years!

(c) (+8 pts) Using the survival curve, explain which treatment (drug=1 or drug=0) is associated with longer survival times. (Note: your answer should include the context of the interaction of drug with age). Feel free to use Wolfram-alpha (or if you tell me what you want to plot in R, I can help you do it); let me know if you want to include a screenshot.

Let's get plotting! In the tibble below, I include a variable `time` denoting the time after surgery. `age` is varied between 10, 47, and 80 for the different plots. `drug` is a 0/1 binary variable. `hazard` denotes the hazard at each age and drug combination. `surv` denotes the survival probability associated with each person

```
responses<-tibble(
  time=rep(seq(0.1, 10, by=0.1), 2),
  #age=rep(seq(1, 100, by=1), 2),
  age=10,
  drug=rep(c(0, 1), each=100),
  hazard=0.2*exp(age*0.00942+drug*0.11859-0.00670*age*drug),
  surv=exp(-hazard*time)
)

p0<-responses %>% ggplot(aes(y=surv, x=time, color=factor(drug)))+
  geom_line()+theme_minimal()+
  xlab("Time (years after surgery)")+ylab("Survival")+
  theme(axis.text=element_text(angle=90))+ggtitle("Age=10")

responses<-tibble(
  time=rep(seq(0.1, 10, by=0.1), 2),
  #age=rep(seq(1, 100, by=1), 2),
  age=47,
  drug=rep(c(0, 1), each=100),
  hazard=0.2*exp(age*0.00942+drug*0.11859-0.00670*age*drug),
  surv=exp(-hazard*time)
)

p1<-responses %>% ggplot(aes(y=surv, x=time, color=factor(drug)))+geom_line()+
  theme_minimal()+xlab("Time (years after surgery)")+
  ylab("Survival")+
  theme(axis.text=element_text(angle=90))+ggtitle("Age=47")

responses<-tibble(
  time=rep(seq(0.1, 10, by=0.1), 2),
  #age=rep(seq(1, 100, by=1), 2),
  age=80,
  drug=rep(c(0, 1), each=100),
  hazard=0.2*exp(age*0.00942+drug*0.11859-0.00670*age*drug),
  surv=exp(-hazard*time)
)

p2<-responses %>% ggplot(aes(y=surv, x=time, color=factor(drug)))+geom_line()+
  theme_minimal()+xlab("Time (years after surgery)")+ylab("Survival")+
  theme(axis.text=element_text(angle=90))+ggtitle("Age=80")

grid.arrange(p0, p1, p2, nrow=1)
```



It looks like for people aged 10, survival is very similar regardless of drug treatment (slightly higher survival in those who didn't recieve the drug). For people aged 47, the survival probabilities begin to diverge; survival is higher in the 47 year olds who recieved the drug relative to the control group. For people aged 80, there is a marked improvement in survival in those who recieved the drug as compared to the control group. Thus, the drug appears to offer survival benefits for older patients.

Q5. Semester Reflection

The grading for each of the parts below is 8 points for an answered question, 10 points for a thoughtful answer (out of max 10 pts each).

(a) **Connections** Find a topic from this class in the wild (must be publicly available [through the library is fine], within the last 5 years, something you find interesting). Answer the following:

- provide the citation / URL (the source does not need to be an academic journal, it could be a media article or even someone's blog). The article must not be something I have provided for you already as part of the course materials.
- provide the quote / result which is related to the class work
- give your own interpretation of the statistical idea (what does it mean?)
- reflect on the wild citing: was it done well? did the authors misinterpret something? was it out of context? did it provide needed information for the article?

This paper (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1626238/) is a study that showed that longer telomeres (repetitive sequence caps on the end of chromosomes) were associated with higher survival in tree swallows. The authors sampled 1 year old female birds and put them into 4 groups depending on their telomere lengths. The authors defined a "death" as a female bird not returning to nest at the same location (since swallows are known to return to the same nesting site at high frequencies). Survival curves were fit to each group of birds and differences between the curves were assessed using the Wilcoxon rank test:

"...individuals with longer TRFs had a higher survival probability than those individuals with shorter TRFs (p=0.010)."

The survival curves described above were merely the proportions of birds tagged in 2001 that returned to the nesting site near Ithaca, New York over the subsequent 3 years. The Wilcoxon test is a less powerful but more flexible test that doesn't require the proportional hazards assumption, although it requires survival curves to never cross. I agree with the way they implemented the test; according to the the survival curves presented in the paper, the curves did not cross, making the Wilcoxon test appropriate in this situation.

While the results of this study were intriguing, I have two concerns about this study. The first involves the small sample size: in total, only 22 birds were studied, and survival curves were built using only a handful of birds. With small sample sizes, stochastic effects (e.g., a severe weather event, local food shortages, etc.) could produce differences in between-group survival that could yield Type 1 errors. Additionally, the small sample size and restriction to one nesting population limits the generalizability of this study. My second concern is that the authors determined survival/death by whether the tree swallows were reobserved at the Ithaca nesting site in subsequent years. The authors used "returning to nest site" as a surrogate outcome for survival. Surrogate outcomes are often used in the biomedical setting as a shortcut for demonstrating a difficult-to-measure outcome: for example, measuring tumor regression as opposed to 5 year cancer survival. The authors relied on previous literature stating that birds that do not return to nesting sites are "probably dead". A more rigorous approach to this problem (like an experiment demonstrating the high correlation between absence at nesting site and death OR using a survival analysis method that can account for uncertainty in the time to event values) would increase my confidence in their findings.

(b) **Persistence** Find one homework problem you have worked on this semester that you struggled to understand and solve (include the problem so that I can see it). Explain how the struggle itself was valuable. In the context of this question, describe the struggle and how you overcame the struggle. You might also discuss whether struggling built aspects of character in you (e.g., endurance, self-confidence, competence to solve new problems) and how these virtues might benefit you in later ventures.

The problem: HW9 Q2. Chp 9, E11. The gist of the problem was to check the PH assumption using log-log plots, use a nested models approach to identify the "best" model, and then interpret the coefficients in the final model in the fruitfly dataset.

This problem captured a lot of different ideas that we worked with in the course (checking assumptions, model building, hypothesis testing, interpreting results), and so there was a LOT of things to keep track of. After checking the PH assumptions, I was really unsure if we should be encoding variables as categorical or continous (esp. variables like `Thorax` and `Sleep`) and how I should go about discretizing continous measurements (like thorax length and hours slept) for the categorical models. However, I just reminded myself that "model building is an art", and after some exploratory plots of the distrubutions of these covariates, I chose the categorical cutoffs as best I could. Then, by comparing the log(HR) estimates associated with each category, I could assess is the HR was roughly linear with respect to each covariate (and whether or not the variables should be encoded as continous).

After encoding my variables, I went through the process of carefully stepping through nested models and using the LRT to assess whether variables should be left in/excluded from the model. Again, this was a tricky task. I needed to keep track of how variables were encoded to determine the degrees of freedom to calculate the Chi-Square p-value associated with the two models. After stepping through each model, I got down to a model with a single variable `Thorax`. My mentor pod reminded me that we needed to assess whether we should include `Thorax` in the model, and helped me construct a model with no covariates as the null model. In the end, after many steps, I determined that the Cox PH model with only `Thorax` was best model.

Doing the above problem offered some immediate benefits. After completing Q2, the subsequent problem (Q3) was much easier to solve, as my pod and I understood the process better. More broadly, working on this problem reminded me of the complexities of model building and that model building isn't a perfect science; it is a process underpinned by decisions made on part of the statistician. For instance, when I encoded continous measurements (like Thorax length) as categorical variables, I had to **choose** how to break the measurements into discrete categories. I made my decision with the help of some diagnostic plots, but the decision was ultimately my own. I think there's a tendency to think of statistical methods/models as perfect and unimpeachable, but a **true** appreciation for statistics lies in understanding the underlying assumptions, limitations, and choices that go into building any model of the real world. Lastly, the experience working on this problem highlighted the importance of working in a group when it comes to challenging problems. This problem had a lot of moving parts, and it was super valuable for me to work with, check my solutions, and bounce ideas off my groupmates. Plus, they made working on homeworks a TON of fun as well!

(c) **Community** Provide an example (based on the course material) of something that someone other than the instructor taught you this semester. The example might have come from a question during office hours, a debate in your mentor pod, a breakout room work session, or a start-of-class discussion. What did you learn from your community member? Why do you think you remember learning it (that is, what struck you about it)?

On HW 4 Q6, one of my podmates helped explain how to find the inflection point on an S curve/logistic curve. Inspecting the second derivative of the logit function (a kinda nasty expression: $-\frac{\beta_1^2(e^{\beta_1 x+\beta_0}-1)e^{\beta_1 x+\beta_0}}{(e^{\beta_1 x+\beta_0}+1)^3}$) reveals that the expression reaches 0 when $\beta_1 x+\beta_0 = 0$. When $\beta_1 x+\beta_0 = 0$, then the logit(p)=$\frac{\exp(0)}{1+\exp(0)} = \frac{1}{2}$. Since the inflection point is achieved by $\beta_1 x + \beta_0 = 0$, the x value associated with the inflection point is therefore $x = -\beta_0/\beta_1$. I initially approached the problem by eyeballing the proobability associated with the inflection point on the S curve. However, my podmate approached it in a totally different way, and, though I usually shy away from value judgements, her result was WAY better! Her work showed the WHY behind the results that I had heuristically observed. It was really satisfying to understand why we were seeing certain relationships in the graphs of the different logistic curves. Lastly, my podmate posted her work in our mentorless mentor pod Discord chat, which was super generous of her! The experience affirmed my love of working with fun-loving and insightful peers. I will miss my pod dearly next year!

```
praise()
```

```
## [1] "You are spectacular!"
```