Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis (Spring 2021)
Exam #2 Key: Median=83 (IQR=[71-92])

Instructions: Please adhere to the following guidelines:

- Your responses to this exam are due by email at **12:00p** on **Wednesday, May 5**. I intend for this cutoff to be strict. You should not need anywhere near the full amount of time to complete the exam, but I am building in a generous cushion to allow for, e.g., technology issues. For your convenience—and to minimize the time and effort you need to spend on formatting—I have provided a template for your responses on which you should word-process your solutions and send to andrew.spieker@vumc.org by the deadline. Please label your solutions according to the format "LASTNAME-EXAM.docx." I will confirm receipt of your exam. If you want to be cautious, you are welcome to e-mail exam drafts along the way (I will not start grading until after 12:00p on May 5, so there is no risk in sending drafts. Don't worry about spamming my inbox; it won't bother me at all—I'll just grab the most recent copy at the deadline). I intend to score the exams, post the key, and provide detailed feedback to you by Friday, May 7.

- There are four required problems (each with multiple sub-questions of varying length and difficulty) and two optional problems; there are three pages of appendix material. You are *not* expected to utilize software for any of the required problems—all software output provided in this exam is sufficient to answer them.

- In a couple of problems on this exam, I have deliberately replaced part of the output with "%%%." The idea is that you should either be able to fill in the blank on the basis of other things in the output or by employing concepts and ideas covered in this course.

- Please read the instructions very carefully; answer no more and no less than what you are being asked to answer. You'll notice throughout the exam that I repeatedly implore you to be concise. There are a lot of sub-questions on this exam, many of which should go relatively quickly and should only require very brief responses. My strong recommendation is to provide your responses to the problems you find easiest first, and then return to the more challenging ones.

- This exam is open book/notes/calculator, but is an **individual effort**. You are not permitted to collaborate with individuals inside or outside the class, in-person, electronically, telepathically, or otherwise. You of course *may* consult any course materials including those on the course webpage. I cannot stop you from consulting other online resources, although doing so should not be necessary.

- All questions regarding the exam should be directed to Andrew (andrew.spieker@vumc.org). If I am able to respond to your question, I will provide my response to the whole group (and I will of course anonymize your question). If I am unable to respond, I will let you know.

- Upon completion of your exam, please indicate on the first page of the template whether you agree with the following statement: "On my honor, I have neither given nor received unauthorized aid on this exam." If you have concerns about your ability to answer this in the affirmative, please turn in your exam anyway, and send me an email so we can discuss.

- Please round any final calculations to a reasonable number of significant digits!

- As is always the case in this class, any reference to logarithmic transformations are based on the *natural* logarithm (i.e., having base e).

- **Importantly**: Take a deep breath — you've got this! This is an opportunity to showcase all of the hard work you've done in this class.

---

1. $\boxed{\text{30 pts}}$ A phase-II randomized controlled trial was conducted with the goal of comparing two treatments in a population of patients diagnosed with advanced non small-cell lung cancer. A total of $N = 188$ patients were randomized in a blinded fashion to receive either docetaxel plus a placebo or docetaxel plus an experimental receptor tyrosine kinase (RTK) blocker; patients were followed for a maximum of two years. The investigators hypothesized that the addition of the RTK blocker would make it more difficult for cancer cells to divide and that its receipt in conjunction with docetaxel could in turn improve survival time relative to docetaxel alone. There was reason to suspect *a priori* that any benefit derived from the addition of the experimental RTK blocker would be less pronounced among subjects with highly advanced disease at the time of initial diagnosis (in this study, advanced disease was defined by the presence of malignant pleural effusion, which is the build up of fluid and cancer cells between the chest wall and the lung). The variables measured in this study were as follows:

`tx`	treatment assignment (0 = docetaxel + placebo; 1 = docetaxel + RTK blocker)
`mpe`	(0 = no malignant pleural effusion; 1 = malignant pleural effusion)
`age`	age at time of diagnosis (years)
`obstime`	time from randomization to either death or censoring
`death`	status at last follow-up time (0 = alive; 1 = dead)

The investigators fit the following Cox proportional hazards model:

$$\log\left(\lambda(t|\texttt{tx},\texttt{mpe},\texttt{age})\right) \;=\; \log\left(\lambda_0(t)\right) + \beta_1 1(\texttt{tx=1}) + \beta_2 1(\texttt{mpe=1}) + \beta_3 1(\texttt{tx=1}) \times 1(\texttt{mpe=1}) + \beta_4 \texttt{age}.$$

Appendix I provides the Stata output for this model, along with the results of a specific hypothesis test (note that one of the hazard ratios has been deliberately replaced with a `%%%` symbol). **You do not need to do the "long" write-up for any of these problems.**

(a) Determine a point estimate, 95% CI, and p-value for the hazard ratio that compares the hazard of death between subgroups differing in their randomized treatment but of the same age and with no malignant pleural effusion at time of diagnosis.

Ans: $\log\left(\lambda(t|\texttt{tx=1},\texttt{mpe=0},\texttt{age})\right) - \log\left(\lambda(t|\texttt{tx=0},\texttt{mpe=0},\texttt{age})\right) = \beta_1$; therefore, the hazard ratio is estimated to be 0.491 (95% CI: [0.249, 0.9709]).

(b) Determine a point estimate, 95% CI, and p-value for the hazard ratio that compares the hazard of death between subgroups differing in their randomized treatment but of the same age and with malignant pleural effusion at time of diagnosis.

Ans: $\log\left(\lambda(t|\texttt{tx=1},\texttt{mpe=1},\texttt{age})\right) - \log\left(\lambda(t|\texttt{tx=0},\texttt{mpe=1},\texttt{age})\right) = \beta_1 + \beta_3$. It seems that the point estimate is blocked — there are a couple of ways that you could recover it (either is acceptable). One approach utilizes information in the output:

$$\widehat{\beta}_1 + \widehat{\beta}_3 = \log(0.4912413) + \log(2.013515) = -0.01094.$$

Another approach is to note that the point estimate should lie halfway between the endpoints of the confidence interval once log-transformed:

$$\widehat{\beta}_1 + \widehat{\beta}_3 = \frac{1}{2}\left(\log(0.68421) + \log(1.429915)\right) = -0.01094.$$

Exponentiating, we obtain an estimated hazard ratio of 0.989 (95% CI: [0.684, 1.430]).

(c) In one sentence, summarize the degree to which this study provides evidence of a differential treatment effect between groups defined by malignant pleural effusion status at time of diagnosis, adjusting for age (be certain to include a measure of statistical strength of evidence as part of your response).

Ans: First, note that

$$\begin{aligned}&(\log(\lambda(t|\texttt{tx=1},\texttt{mpe=1},\texttt{age})) - \log(\lambda(t|\texttt{tx=0},\texttt{mpe=1},\texttt{age})))\\ &-(\log(\lambda(t|\texttt{tx=1},\texttt{mpe=0},\texttt{age})) - \log(\lambda(t|\texttt{tx=0},\texttt{mpe=0},\texttt{age}))) = \beta_3.\end{aligned}$$

Therefore, this can be evaluated by examining the interaction term. This study does not provide sufficient evidence that the corresponding "ratio of hazard ratios" is different form one ($p = 0.077$).

(d) Despite your answer to part (c), briefly summarize the major advantage of the proposed model over a model that does not accommodate a differential treatment effect between groups defined by their malignant pleural effusion status at time of diagnosis.

Ans: The investigators had suspected that there would be a differential treatment effect between groups defined by advanced disease. The results of this study may not specifically confirm this hypothesis in the sense of providing statistically significant evidence of a nonzero interaction term—however, the model *did* provide statistically significant evidence of a treatment benefit in the group with no malignant pleural effusion at the time of diagnosis. This finding, if true, is important and informative, and would likely not have been uncovered with a model that did not include an interaction term.

(e) Briefly describe the most likely advantage of having included age as a covariate in this model.

Ans: Age is a key predictor of time to death. Hazard ratios are non-collapsible; failing to include age as a covariate in the model will attenuate the hazard ratios of interest toward one, thereby limiting the study's ability to uncover clinically important associations.

(f) In no more than three sentences, characterize the most essential assumptions invoked in this analysis. Note that I have not provided you with sufficient information to specifically *evaluate* the degree to which they appear to hold, but I expect you to at least name and/or describe the assumptions.

Ans: One key assumption invoked is the assumption of non-informative censoring; this means that (conditional on variables in the model), subjects who are censored at a given time are no more or less likely to die in the immediate future as compared to subjects who remain in the sample. Another key assumption is the assumption of proportional hazards; this means that the hazard (as a function of time) can be expressed as a multiple of the baseline hazard function for any subgroup defined by their treatment and MPE status. Further, we require subjects to be sampled independently.

(g) Suppose you fit the proposed Cox model under the Bayesian paradigm using non-informative priors (meaning, as close to "flat" as possible) on each coefficient. If the purpose of a flat prior is to mitigate the influence of the prior on the conclusions, explain why this choice may not be ideal.

Ans: The coefficients are defined on the log-scale, whereas parameters of interest are almost uniformly expressed as hazard ratios. What is considered "flat" on the log-scale will not be "flat" once exponentiated, and so the prior may be more informative than intended. In sufficiently large samples, this distinction will not matter as much as the (partial) likelihood will dominate the prior.

(h) Suppose a secondary analysis was conducted to study time to cancer-specific death. In no more than four sentences, describe the two possible approaches to such an analysis that we learned about in class. Be certain to summarize how they differ in their treatment of outcomes other than cancer-specific death.

Ans: We learned of two possible approaches: the "cause-specific hazard" approach and the "subdistribution hazard" approach. The former approach treats any non-cancer death as a censoring event, removing subjects from the "at-risk" set when they are either censored or when they have a non-cancer death. The second approach realizes non-cancer deaths as competing events that block the possibility of cancer-related deaths; in this approach, we do not remove subjects with the competing event from the "at-risk" set even though they cannot have a cancer-related death (although the approach sounds a little bit bizarre, it turns out to be the mathematically correct trick to induce a direct and identifiable correspondence between the subdistribution hazard and the cumulative incidence).

---

2. $\boxed{\text{20 pts}}$ You're working as part of an investigative team to construct a polygenic risk score for several classes of related cardiovascular diseases. In this study, $N = 730$ subjects were evaluated and two-hundred individual nucleotides with two known variants were assessed. The variables included in these data were as follows:

`snp1`	single-nucleotide polymorphism 1 (0 = more common variant; 1 = less common variant)
`snp2`	single-nucleotide polymorphism 2 (0 = more common variant; 1 = less common variant)
⋮	⋮
`snp200`	single-nucleotide polymorphism 200 (0 = more common variant; 1 = less common variant)
`cvd`	cardiovascular disease (0 = no; 1 = yes)

Appendix II shows the step-by-step procedure by which the data were analyzed. First, the data were split into (equally sized) random halves (the training set, `sample = 1`, and the test set, `sample = 2`). Then, a logistic model with a LASSO penalty was then fit on the training set with the tuning parameter selected via five-fold cross-validation. Subject-specific risk scores were generated for each subject in the whole data set based on the penalized coefficients, which were then summarized and assessed in a number of ways (note that certain information has been deliberately replaced with a `%%%` symbol). You may assume for the purposes of this problem that each measured single-nucleotide polymorphism (SNP) has at least one subject with each variant in *each* of the training and test sets.

(a) Appendix II presents the estimated absolute prediction error rates in the training set and the test set based on a specific cut-off of "> 0.5." Although the "`%%%`" symbols block you from seeing which estimate comes from the training set and which comes from the test set, take an educated guess as to which is which and very briefly state your reasoning.

Ans: The absolute prediction error rate is expected to be higher in the test set as compared to the training set. Therefore, we suspect that the estimated absolute prediction error rate is 0.206 in the training set and 0.228 in the test set.

(b) Four SNPs were selected into the model; therefore (since there are two possible values for each SNP) there are $2^4 = 16$ possible values that the predicted risk scores can take on based on this model. Determine the combination of SNPs that produces the highest risk score in these data; what is the risk score for this group? *Hint*: You don't need to compute all sixteen scores; you can figure out the right combination by examining the values of the penalized coefficients.

Ans: The risk score will be maximal for subjects with the more common variant of SNPS 60 and 78 and the less common variants of SNPS 93 and 164. In this case, the predicted risk score is given by:

$$\begin{aligned}\widehat{\text{P}}(\texttt{cvd=1}|\texttt{snp60=0}, \texttt{snp78=0}, \texttt{snp93=1}, \texttt{snp164=1}) &= \text{expit}\left(\widehat{\beta}_0^\lambda + \widehat{\beta}_{93}^\lambda + \widehat{\beta}_{164}^\lambda\right) \\ &= \text{expit}\left(-1.217392 + 0.0341118 + 1.802917\right) \\ &= \text{expit}(0.6196368) = \frac{e^{0.6196368}}{1 + e^{0.6196368}} = 0.650.\end{aligned}$$

(c) The cut-off choice of "> 0.5" is admittedly somewhat arbitrary. Briefly explain why using a cut-off of choice of "> 0.9" instead would nevertheless be completely uninformative about the model's predictive ability in this example. *Hint*: appeal to your response to part (b).

Ans: The highest possible predicted value is given by 0.650. Using a cut-off of "> 0.9" would produce predicted values for CVD of zero for all subjects in the test set. This is not helpful.

(d) Appendix II presents the estimated area under the ROC curve in the training set and the test set. Although the "%%%" symbols block you from seeing which estimate comes from the training set and which comes from the test set, take an educated guess as to which is which and very briefly state your reasoning.

Ans: The area under the ROC curve is expected to be lower (closer to 0.5) in the test set as compared to the training set. Therefore, we suspect that the estimated AUC is 0.715 in the training set and 0.656 in the test set.

(e) Briefly describe a key advantage of using the area under the ROC curve as a metric of predictive ability over absolute prediction error.

Ans: The area under the ROC curve aggregates information on prediction error rates across all possible cut-off points. Therefore, the key advantage of this metric is that it eliminates the need to arbitrarily select a cut-off, as we did in producing estimates of absolute prediction error rates.

(f) One of your collaborators working on the study is not familiar with penalized regression and is therefore reluctant to implement it. They propose that you use the full data set to test each of the two-hundred SNPs individually for an association with CVD and then build a multivariate logistic regression model based on the SNPs that show up as statistically significant (i.e., $p < 0.05$). Briefly summarize the most fundamental limitations of this approach.

Ans: The first, and I would argue most fundamental, limitation of this approach is that it cannot produce valid estimates of out-of-sample prediction error rates. The second major challenge of this approach is that selection by p-values is not a principled variable selection approach.

(g) Another collaborator working on the study is familiar with penalized regression and its purposes. They come to your defense and reject the approach described in part (f), but then challenge you on your choice to include a LASSO penalty instead of a ridge penalty (you can never please everyone!). Although each penalty may have its advantages this setting, briefly describe an important practical advantage that the LASSO penalty has over the ridge penalty.

Ans: The LASSO penalty can be thought of as a variable selection approach. From a geometric standpoint, the absolute value function has "sharp corners" that occur when a subset of the parameters are exactly equal to zero. In contrast, the ridge penalty will shrink parameters *toward* zero but not to zero exactly. The variable selection nature of the LASSO approach is appealing in that it produces a simpler model that depends upon fewer predictors (that is, a model that is said to be more *sparse*).

3. $\boxed{\text{20 pts}}$ The **V**anderbilt **E**mergency **R**oom **B**lood Pressure (VERB) study was a pilot study that sought to evaluate the degree to which a text-message intervention could reduce systolic blood pressure (SBP) in subjects admitted to the Vanderbilt emergency department with hypertension. Subjects were randomized to receive either a control condition (standard of care) or a text-message intervention (VERB) pertaining to antihypertensive medication adherence. The study investigators intended for subjects to follow up approximately thirty days following emergency room discharge. However, due to concerns regarding participant retention, many participants had their follow-up scheduled back-to-back with another pre-booked appointment at Vanderbilt Medicine—which was as early as $t = 15$ days post-discharge and as late as $t = 48$ days post-discharge. Therefore, the investigators decided to model the association between VERB and (mean) SBP as a continuous function of time. The variables under consideration are summarized in the table below:

`verb`	treatment assignment (0 = control; 1 = VERB)
`t`	time of follow-up measurement post-discharge (days)
`sbp`	systolic blood pressure at time of follow-up (mm Hg)

The investigators propose the following model, which you may freely assume for the purposes of this problem to be correctly specified:

$$\mathbf{E}[\texttt{sbp}|\texttt{verb},\texttt{t}] \quad = \quad \beta_0 + \beta_1\texttt{verb} + \beta_2\texttt{t} + \beta_3\texttt{t}^2 + \beta_4\texttt{verb}\times\texttt{t} + \beta_5\texttt{verb}\times\texttt{t}^2$$

This problem pertains specifically to different aspects of this proposed model; there is no software output for this problem.

(a) Embedded in the model is a reduced model that characterizes the relationship between time post-discharge and mean SBP specifically in the control group (`verb = 0`). Write down this reduced model.

Ans: The reduced model is given by $\mathbf{E}[\texttt{sbp}|\texttt{verb=0},\texttt{t}] = \beta_0 + \beta_2\texttt{t} + \beta_3\texttt{t}^2$.

(b) Embedded in the model is a reduced model that characterizes the relationship between time post-discharge and mean SBP specifically in the VERB group (`verb = 1`). Write down this reduced model.

Ans: The reduced model is given by $\mathbf{E}[\texttt{sbp}|\texttt{verb=1},\texttt{t}] = (\beta_0 + \beta_1) + (\beta_2 + \beta_4)\texttt{t} + (\beta_3 + \beta_5)\texttt{t}^2$.

(c) Show that this model implies that the difference in mean SBP between treatment groups is a quadratic function of time (*Hint*: Just subtract your answer to part (a) from your answer to part (b)).

Ans: The treatment effect is given by $\Delta(\texttt{t}) = \beta_1 + \beta_4\texttt{t} + \beta_5\texttt{t}^2$.

(d) Use your response to part (c) to characterize the effect of VERB on mean SBP at 30 days in terms of the model parameters.

Ans: This is given by $\Delta(30) = \beta_1 + 30\beta_4 + 30^2\beta_5 = \beta_1 + 30\beta_4 + 900\beta_5$.

(e) Use your response to part (c) to determine which model parameter(s) should be tested to evaluate whether there is an an effect of VERB on mean SBP at any time over the range of follow-up times observed.

Ans: Since the treatment effect depends upon β_1, β_4, and β_5, the correct hypothesis test would be $H_0 : \beta_1 = \beta_4 = \beta_5 = 0$ vs. $H_1 : \text{not } H_0$.

(f) Use your response to part (c) to determine an expression for the degree to which the effect of VERB on mean SBP differs between times $t = 20$ and $t = 40$ days post-discharge.

Ans: This is given by $\Delta(40) - \Delta(20) = (\beta_1 + 40\beta_4 + 40^2\beta_5) - (\beta_1 + 20\beta_4 + 20^2\beta_5) = 20\beta_4 + 1200\beta_5$.

(g) Use your response to part (c) to determine which parameter(s) should be tested to evaluate whether the effect of VERB on mean SBP is constant over range of follow-up times observed.

Ans: A constant treatment effect over time would suggest $\Delta(t) = c$ for some constant c that does not depend on `t`. For this to be the case, we would need that $\beta_4 = 0$ and $\beta_5 = 0$. Therefore, the appropriate hypothesis test would $H_0 : \beta_4 = \beta_5 = 0$ vs. $H_1 :$ not H_0.

---

4. $\boxed{\text{30 pts}}$ SARS-CoV-2 possesses four main structural proteins. The spike protein, located on the viral surface, is highly immunogenic (meaning it produces strong immune response) and is hence a focus of COVID-19 research. The SARS-CoV-2 receptor binding domain (RBD) immunoglobulin-G (IgG) test evaluates the extent of spike protein antibodies, and is measured in absorbance units (AU) per milliliter. We will henceforth refer to this as the RBD test. A group of investigators sought to compare the degree to which vaccination for SARS-CoV-2 elicited spike protein antibody production between healthy controls and patients taking immunosuppressant drugs (IDs) for treatment of multiple sclerosis (MS). A vaccine study of $N = 50$ subjects was conducted; twenty-five ID-treated MS patients were enrolled, along with twenty-five healthy controls (HCs). All subjects received SARS-CoV-2 vaccination doses six weeks apart. The RBD test was performed six weeks after the first vaccine dose (immediately prior to the second dose) and then again eight weeks after the second vaccine dose. You may assume that subjects are independent of one another. When produced in the "long" format, the data set comprises the following variables:

`id`	subject ID (1, 2, . . . , 50)
`t`	time (1 = six weeks after vaccine dose 1; 2 = eight weeks after vaccine dose 2)
`status`	(0 = HC; 1 = ID-treated MS patient)
`rbd`	SARS-CoV-2 RBD IgG response at time `t` (AU/ml)

The investigators propose fitting the following mean model using generalized estimating equations with a working independence correlation structure:

$$\mathbf{E}[\texttt{rbd}_\texttt{t}|\texttt{status}] \;=\; \beta_0 + \beta_1 1(\texttt{status=1}) + \beta_2 1(\texttt{t=2}) + \beta_3 1(\texttt{status=1}) \times 1(\texttt{t=2}).$$

The output for this model is presented in Appendix III, along with some additional results. **You do not need to do the "long" write-up for any of these problems.**

(a) Very briefly explain why standard multiple linear regression is not a valid approach to this problem.

Ans: Standard multiple linear regression requires errors to be pairwise independent, an assumption that cannot reasonably be justified in this setting due to repeated measures within subjects over time.

(b) Determine a point estimate and a 95% CI for the mean RBD among HCs eight weeks after dose 2.

Ans: $\{\mathbf{E}[\texttt{rbd}_2|\texttt{status=0}] = \beta_0 + \beta_2\}$ 1.39 AU/ml (95% CI: [1.26, 1.51]).

(c) Determine a point estimate and a 95% CI for the mean RBD among ID-treated MS patients six weeks after dose 1.

Ans: $\{\mathbf{E}[\texttt{rbd}_1|\texttt{status=1}] = \beta_0 + \beta_1\}$ 1.80 AU/ml (95% CI: [1.67, 1.92]).

(d) Determine a point estimate and a 95% CI for the difference in mean RBD between ID-treated MS patients and HCs six weeks after dose 1.

Ans: $\{\mathbf{E}[\texttt{rbd}_1|\texttt{status=1}] - \mathbf{E}[\texttt{rbd}_1|\texttt{status=0}] = \beta_1\}$ -0.214 AU/ml (95% CI: [-0.392, -0.354]).

(e) Determine a point estimate and a 95% CI for the difference in mean RBD between ID-treated MS patients and HCs eight weeks after dose 2.

Ans: $\{\mathbf{E}[\texttt{rbd}_2|\texttt{status=1}] - \mathbf{E}[\texttt{rbd}_2|\texttt{status=0}] = \beta_1 + \beta_3\}$ 0.0248 AU/ml (95% CI: [-0.166, 0.215]).

(f) Determine a point estimate and a 95% CI for the change in mean RBD from six weeks after dose 1 to eight weeks after dose 2 within HCs.

Ans: $\{\mathbf{E}[\texttt{rbd}_2|\texttt{status=0}] - \mathbf{E}[\texttt{rbd}_1|\texttt{status=0}] = \beta_2\}$ -0.622 AU/ml (95% CI: [-0.740, -0.504]).

(g) Determine a point estimate and a 95% CI for the change in mean RBD from six weeks after dose 1 to eight weeks after dose 2 within ID-treated MS patients.

Ans: $\{\mathbf{E}[\texttt{rbd}_2|\texttt{status=1}] - \mathbf{E}[\texttt{rbd}_1|\texttt{status=1}] = \beta_2 + \beta_3\}$ -0.384 AU/ml (95% CI: [-0.533, -0.234]).

(h) In one sentence, summarize the degree to which this study provides evidence that the difference in mean RBD between ID-treated MS patients and HCs changes from six weeks after dose 1 to eight weeks after dose 2 (be certain to include a measure of statistical strength of evidence as part of your response).

Ans: $\{(\mathbf{E}[\texttt{rbd}_2|\texttt{status=1}] - \mathbf{E}[\texttt{rbd}_2|\texttt{status=0}]) - (\mathbf{E}[\texttt{rbd}_1|\texttt{status=0}] - \mathbf{E}[\texttt{rbd}_1|\texttt{status=0}]) = \beta_3\}$; therefore, this study provides sufficient evidence of such a change ($p = 0.015$).

(i) In one sentence, summarize the degree to which this study provides evidence that the change in mean RBD between from six weeks after dose 1 to eight weeks after dose 2 differs between ID-treated MS patients and HCs (be certain to include a measure of statistical strength of evidence as part of your response).

Ans: $\{(\mathbf{E}[\texttt{rbd}_2|\texttt{status=1}] - \mathbf{E}[\texttt{rbd}_1|\texttt{status=1}]) - (\mathbf{E}[\texttt{rbd}_2|\texttt{status=0}] - \mathbf{E}[\texttt{rbd}_1|\texttt{status=0}]) = \beta_3\}$ (this exemplifies the symmetry property of interaction terms that we have discussed many times in this class); therefore, the answer is the same as that of part (h).

(j) In one sentence, summarize the degree to which this study provides evidence of overall differential vaccine immunogenicity (as measured by mean RBD) between ID-treated MS patients and HCs.

Ans: The hypothesis to test is $H_0 : \beta_1 = \beta_3 = 0$ vs. $H_1 :$ not H_0; this study provides sufficient evidence of such an association ($p < 0.001$).

---

5. ***Optional problem 1***: This is an optional problem — it was distributed prior to the official distribution of Exam 2; your response is due with the rest of the exam, although you may hand it in any time prior. Credit can be earned based a thoughtful and well-constructed response; however, please do not spend an extraordinary amount of time on this problem (recommended maximum: about ten minutes per part, or thirty minutes total).

Evidence shows that reflection is a key component of learning. This problem is about reflecting on the material covered in BIOS 6312. If you choose to complete this problem, your response to each question should be about four to ten sentences. If you need some inspiration, refer to the course learning objectives as stated in the syllabus. There is no one right answer to these questions; the goal is to give you an open-ended opportunity to reflect.

(a) Briefly summarize and discuss a general concept, theme, or underlying principle that has tied together some of the methods covered in this course (please go a bit deeper than simply naming some of the methods we've covered).

Ans: There are many correct answers!

(b) Consider one of the *many* examples in this class where you have learned of (at least) two competing methods, frameworks, or approaches to the address the same problem. State the two competing approaches in the example you have identified. Then, acknowledge at least one *advantage* and at least one *limitation* of each.

Ans: There are many examples!

(c) The following statement is attributed to Dr. George Box, a statistician who made great contributions to time-series analysis, Bayesian statistics, and experimental design:

"*All models are wrong, but some are useful.*"

Briefly discuss a couple of examples that we have covered in this class that very clearly align with the fundamental message of this quote.

Ans: There are many examples!

---

6. ***Optional problem 2***: This is an optional problem — I recommend not attempting it until you have completed and are satisfied with your answers to the required problems. A small amount of credit can be earned for attempting the problem, and an additional small amount of credit can be earned for a correct response.

Let $Y_1, \ldots, Y_N \sim \text{Poisson}(x_i\theta)$ denote independent count variables, each with mass function given by:

$$p_\theta(y|x_i) = \frac{(x_i\theta)^y}{y!}e^{-\theta x_i}.$$

Here, $\theta > 0$ is a fixed, unknown parameter to be estimated. You may assume for the purposes of this problem that $x_1, \ldots, x_N$ are fixed and known positive numbers that are bounded between $[1/c, c]$ for some $c \geq 1$.

(a) Obtain the maximum likelihood estimator, $\widehat{\theta}_N$, for θ and show that it is unbiased for θ.

Ans: To obtain the MLE, we first need to write down the likelihood based on the N observations:

$$\mathcal{L}(\theta; \mathbf{y}) = \prod_{i=1}^{N} \frac{(x_i\theta)^{y_i}}{y_i!}e^{-\theta x_i}.$$

In turn, the log-likelihood is given by

$$\ell(\theta; \mathbf{y}) = \log\left(\mathcal{L}(\theta; \mathbf{y})\right) = \sum_{i=1}^{N} y_i \log(\theta) - \sum_{i=1}^{N} x_i\theta + c(\mathbf{x}, \mathbf{y}).$$

The score function is given by:

$$\dot{\ell}(\theta; \mathbf{y}) = \frac{\partial}{\partial\theta}\ell(\theta; \mathbf{y}) = \frac{1}{\theta}\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i.$$

Setting equal to zero reveals that the maximum is achieved at $\widehat{\theta}_N = \left(\sum_{i=1}^{N} y_i\right) / \left(\sum_{i=1}^{N} x_i\right)$, which has expectation given by

$$\mathbf{E}\left[\widehat{\theta}_N\right] = \frac{1}{\sum_{i=1}^{N} x_i} \times \sum_{i=1}^{N} \mathbf{E}[Y_i] = \frac{1}{\sum_{i=1}^{N} x_i} \times \sum_{i=1}^{N} x_i\theta = \theta.$$

(b) Determine (and name) the posterior distribution $\pi(\theta|Y_1, \ldots, Y_N)$ under the prior $\theta \sim \text{Gamma}(\alpha, \beta)$. To be clear, the prior density function is given by:

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta},$$

with mean $\mathbf{E}[\theta] = \alpha/\beta$. It may help you to know that this choice is the "conjugate prior" for the Poisson family, meaning that the posterior distribution should also be in the Gamma family with parameters α^* and β^* that you are to determine.

Ans: The posterior distribution can be determined by identifying the Gamma kernel:

$$\begin{aligned}\pi(\theta|\mathbf{y}) &\propto_\theta f(\mathbf{y}|\theta)\pi(\theta) = \prod_{i=1}^{N} \frac{(x_i\theta)^{y_i}}{y_i!}e^{-\theta x_i} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\ &\propto_\theta \theta^{\alpha-1}e^{-\beta\theta}\theta^{\sum_{i=1}^{N} y_i}e^{-\theta\sum_{i=1}^{N} x_i} = \theta^{(\alpha+\sum_{i=1}^{N} y_i)-1}e^{-(\beta+\sum_{i=1}^{N} x_i)\theta}.\end{aligned}$$

This can be recognized as the kernel of a Gamma distribution with hyperparameters $\alpha^* = \alpha + \sum_{i=1}^{N} y_i$ and $\beta^* = \beta + \sum_{i=1}^{N} x_i$.

(c) State the posterior mean $\widetilde{\theta}_N = \mathbf{E}[\theta|Y_1,\ldots,Y_N]$. Show that it can be expressed in the following form:

$$\widetilde{\theta}_N \quad = \quad w_N \times \mathbf{E}[\theta] + (1 - w_N) \times \widehat{\theta}_N,$$

Specifically determine the value of w_N as part of your response.

Ans: The posterior mean is given by α^*/β^*, and we can factor it as follows:

$$\begin{aligned}\widetilde{\theta}_N &= \mathbf{E}[\theta|\mathbf{y}] = \frac{\alpha^*}{\beta^*} = \frac{\alpha + \sum_{i=1}^N y_i}{\beta + \sum_{i=1}^N x_i} = \frac{\alpha}{\beta + \sum_{i=1}^N x_i} + \frac{\sum_{i=1}^N y_i}{\beta + \sum_{i=1}^N x_i} \\ &= \frac{\beta}{\beta + \sum_{i=1}^N x_i} \times \frac{\alpha}{\beta} + \frac{\sum_{i=1}^N x_i}{\beta + \sum_{i=1}^N x_i} \times \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i}.\end{aligned}$$

It is clear from this factorization that the proposition of this problem is true, with weights given by

$$w_N \quad = \quad \frac{\beta}{\beta + \sum_{i=1}^N x_i}.$$

(d) Use your response to part (c) to argue that the posterior mean $\widetilde{\theta}_N$ is a biased but nevertheless consistent estimator for θ. A formal proof is *not* required.

Ans: Since $\widehat{\theta}_N$ is unbiased for θ, the only prior distributions that will result in an unbiased posterior mean $\widetilde{\theta}_N$ are ones for which $\alpha/\beta = \theta$. Since θ is unknown and unknowable, we can assume this occurs with probability zero. However, examine the weights, w_N. Since $1/c < x_i < c$ for all i, we have that

$$w_N \quad = \quad \frac{\beta}{\beta + \sum_{i=1}^N x_i} < \frac{\beta}{\beta + \sum_{i=1}^N (1/c)} = \frac{\beta}{\beta + N/c} \xrightarrow{N\to\infty} 0.$$

In large samples, the discrepancy between the MLE and the posterior mean will therefore go to zero. Since we know the MLE to be consistent from standard likelihood theory, it stands to reason that the posterior mean will also be consistent (this is not a formal proof, although a formal proof could be done). This aligns with the idea that the likelihood will dominate the prior "in the long run."

THE END! THANK YOU FOR A FABULOUS SEMESTER!

Appendix Material for Exam 2

APPENDIX I: Stata output for Problem 1

```
. stset obstime death

. stcox i.tx##i.mpe age, robust nolog

         failure _d:  death
   analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          188             Number of obs    =         188
No. of failures      =          140
Time at risk         =        73085
                                                Wald chi2(4)     =       14.29
Log pseudolikelihood =    -641.0026             Prob > chi2      =      0.0064

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        1.tx |   .4912413   .1707506    -2.04   0.041     .2485568    .9708768
       1.mpe |   1.287315   .3640708     0.89   0.372     .7395257     2.24087
             |
      tx#mpe |
        1 1  |   2.013515   .7975258     1.77   0.077     .9264154    4.376271
             |
         age |    .998936   .0163712    -0.06   0.948      .967359    1.031544
------------------------------------------------------------------------------


. lincom 1.tx + 1.tx#1.mpe, hr

 ( 1)  1.tx + 1.tx#1.mpe = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |        %%%   .1859951    -0.06   0.954       .68421    1.429915
------------------------------------------------------------------------------

** NOTE: %%% means the output has been deliberately withheld **
```

APPENDIX II: Stata output for Problem 2

```
. splitsample, generate(sample) nsplit(2) rseed(2021)

. lasso logit cvd snp1-snp200 if sample == 1, rseed(6312) folds(5)

. lassocoef, display(coef, penalized)

------------------------
             |    active
-------------+----------
       snp60 | -.0147246
       snp78 | -.0395375
       snp93 |  .0341118
      snp164 |  1.802917
       _cons | -1.217392
------------------------

. predict riskscore
(options pr penalized assumed; Pr(cvd) with penalized coefficients)

. gen cvdhat = 0

. replace cvdhat = 1 if riskscore > 0.5
(71 real changes made)

. gen abserror = abs(cvd - cvdhat)

. summarize abserror if sample == %%%

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    abserror |        355    .2056338    .4047345          0          1


. summarize abserror if sample == %%%

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    abserror |        355     .228169    .4202444          0          1


. roctab cvd riskscore if sample == %%%

                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------------------------------------
           355     0.6558       0.0342        0.58888     0.72278

. roctab cvd riskscore if sample == %%%

                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------------------------------------
           355     0.7146       0.0332        0.64941     0.77974

** NOTE: %%% means the output has been deliberately withheld **
```

APPENDIX III: Stata output for Problem 4

```
. regress rbd i.status##i.t, cluster(id) robust

Linear regression                               Number of obs     =        100
                                                F(3, 49)          =      46.39
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4083
                                                Root MSE          =      .3228

                                  (Std. Err. adjusted for 50 clusters in id)
------------------------------------------------------------------------------
             |               Robust
         rbd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.status |  -.2136516   .0887238    -2.41   0.020    -.3919487   -.0353544
         2.t |  -.6221736   .0586934   -10.60   0.000    -.7401224   -.5042249
             |
    status#t |
        1 2  |   .2384203   .0946806     2.52   0.015     .0481526     .428688
             |
       _cons |   2.011314    .064205    31.33   0.000     1.882289    2.140339
------------------------------------------------------------------------------

. testparm i.status status#t

 ( 1)  1.status = 0
 ( 2)  1.status#2.t = 0

       F(  2,    49) =    4.14
            Prob > F =    0.0219

. lincom _cons + 1.status

 ( 1)  1.status + _cons = 0

------------------------------------------------------------------------------
         rbd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.797662   .0612342    29.36   0.000     1.674608    1.920717
------------------------------------------------------------------------------

. lincom 2.t + 1.status#2.t

 ( 1)  2.t + 1.status#2.t = 0

------------------------------------------------------------------------------
         rbd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.3837533   .0742933    -5.17   0.000    -.5330513   -.2344554
------------------------------------------------------------------------------

. lincom _cons + 2.t

 ( 1)  2.t + _cons = 0

------------------------------------------------------------------------------
         rbd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    1.38914   .0624017    22.26   0.000     1.263739    1.514541
------------------------------------------------------------------------------

. lincom 1.status + 1.status#2.t

 ( 1)  1.status + 1.status#2.t = 0

------------------------------------------------------------------------------
         rbd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0247688   .0947101     0.26   0.795    -.1655583    .2150958
------------------------------------------------------------------------------
```