Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis (Spring 2022)
Exam #2

Name (Printed): _____

Instructions: Please adhere to the following guidelines:

- The in-class component of this exam features six required problems and one optional problem that is optional for all students. The take-home component of this exam features one required problem and one optional problem that I will provide you once you complete the in-class portion of the exam.
- Please read the questions carefully and answer no more or less than what you are being asked to answer.
- My recommendation is to provide your responses to the problems you find easiest first, and then return to the more challenging ones.
- This in-class portion of the exam is closed-everything, and is an individual effort. You will, however, be permitted the use of a scientific calculator.
- The take-home portion of the exam is open-everything, but is still an individual effort.
- For the in-class portion of the exam, please indicate on the first page 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. When you turn in your response to the take-home portion of the exam, please confirm by e-mail whether you agree with the integrity statement.
- Please round any final calculations to a reasonable number of significant digits!
- 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 this semester.

In-class component:

#	Score	Points
1		10
2		20
3		15
4		20
5		25
6		10
Total:		100
7 (Optional)		

Take-home component:

#	Score	Points
8		20
Total:		20
9 (Optional)		

Signature for integrity statement: _____

1. [10 pts] Below are ten true-or-false questions (1 pt. each). Circle your choice (**TRUE** or **FALSE**) for each question. Please read the statements *carefully*. **There is no need to provide a written justification for your response**.

(a) **TRUE** or **FALSE** Logistic regression is *only* useful in case-control studies.

(b) **TRUE** or **FALSE** A multinomial logistic regression model with a three-level outcome and a single binary predictor is an example of a saturated model.

(c) **TRUE** or **FALSE** The specific advantage carried by the robust standard error depends partly upon the type of regression model to which it is applied (e.g., linear regression of continuous outcomes, logistic regression for binary outcomes, etc.).

(d) **TRUE** or **FALSE** If two Kaplan-Meier curves cross, that means that the proportional hazards assumption is violated and the Kaplan-Meier curves are useless.

(e) **TRUE** or **FALSE** If I conclude from a longitudinal study that older people have lower mean cholesterol values as compared to younger people, I can also logically conclude that cholesterol values tend to decrease as people age.

(f) **TRUE** or **FALSE** A model's training error tends to be more optimistic than its test error.

(g) **TRUE** or **FALSE** The frequentist paradigm has no merits because it's easy to misinterpret p-values.

(h) **TRUE** or **FALSE** The Bayesian paradigm has no merits because it requires prior distributions.

(i) **TRUE** or **FALSE** Multiple imputation is the magic solution that cures all problems associated with missing data.

(j) **TRUE** or **FALSE** The bootstrap procedure may be able to help you form a 95% CI in settings where the sampling distribution of an estimator is not approximately normal.

2. [20 pts] A study was conducted to understand the relationship between smoking and esophageal cancer in adults. A total of $n_0 = 250$ patients with esophageal cancer and $n_1 = 250$ healthy controls were sampled in a case-control fashion. Enrolled subjects responded to a survey regarding smoking history at the time of diagnosis. The variables evaluated in this study are as follows:

`esoph`	esophageal cancer (0 = no; 1 = yes)
`smk`	smoking history (0 = never-smoker; 1 = former smoker; 2 = current smoker)
`age`	age (years)

The study investigators fit a model in which they allowed an interaction between age and smoking status:

$$\text{logit}(\text{P}(\texttt{esoph}|\texttt{smk}, \texttt{age})) = \beta_0 + \beta_1 \texttt{age} + \beta_2 1(\texttt{smk=1}) + \beta_3 1(\texttt{smk=2}) + \beta_4 1(\texttt{smk=1}) \times \texttt{age} + \beta_5 1(\texttt{smk=2}) \times \texttt{age}.$$

(a) [3 pts] Express (in terms of the model coefficients) the odds of esophageal cancer among 70 year-old former smokers.

(b) [2 pts] Which of the following statements is true regarding this study's capacity to estimate the quantity described in part (a)? Circle the number corresponding to your choice.

(i.) This quantity can be approximately estimated if the prevalence of esophageal cancer is rare.

(ii.) This quantity can be estimated provided there are enough individuals around the age of 70 years.

(iii.) The outcome-dependent nature of the study design does not allow us to estimate this quantity without some external information.

(c) [3 pts] Express (in terms of the model coefficients) the odds ratio that compares the odds of esophageal cancer between current smokers differing age by one year.

(d) [2 pts] Which of the following statements is true regarding this study's capacity to estimate the quantity described in part (c)? Circle the number corresponding to your choice.

(i.) This quantity cannot be estimated unless the prevalence of esophageal cancer is rare.

(ii.) The outcome-dependent nature of the study design does not allow us to estimate this quantity without some external information.

(iii.) An estimate of this quantity may serve as an approximate estimate of the corresponding risk ratio if the prevalence of esophageal cancer is rare.

(e) [2 pts] Suppose you seek to evaluate whether there is an overall association between smoking history and esophageal cancer. Express the null hypothesis, H_0, in terms of a suitable subset or combination of the model parameters. **You need not show the "regression math" in your response.**

(f) [2 pts] Suppose you seek to evaluate whether there is evidence that age modifies the association between smoking history and esophageal cancer. Express the null hypothesis, H_0, in terms of a subset or combination of the model parameters. **You need not show the "regression math" in your response.**

(g) [3 pts] What does it mean in practical terms if both $\beta_2 = \beta_3$ and $\beta_4 = \beta_5$? **You need not show the "regression math" in your response.**

(h) [3 pts] Suppose you seek to evaluate whether there is evidence that the odds of esophageal cancer differs between 75 year-old current smokers and 85 year-old former smokers. Express the null hypothesis, H_0, in terms of a subset or combination of the model parameters. **Please show the "regression math" that leads to your answer.**

3. [15 pts] A preliminary laboratory study was conducted to evaluate doxorubicin as a potential chemotherapy agent. In this study, doxorubicin was applied to $N = 279$ independent cell cultures of the same size at one of the following concentrations, X: 0.00, 0.01, 0.05, 0.10, 0.50, 1.00, and 5.00 µmol/L (`doxconc`). After a commonly timed incubation period, the total number of colonies, Y (`count`), was measured. The post-incubation colony frequencies ranged from about 0 to 250. Consider the following Poisson regression model:

$$\log \mathbf{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

Although we've discussed in class that this process is almost assuredly better described by the Michaelis-Menten equation, you may disregard that for now and assume the Poisson regression model above to be correctly specified in this problem. The Stata output for this model is provided below.

```
. poisson count doxconc, robust nolog

Poisson regression                                      Number of obs =    279
                                                        Wald chi2(1)  = 106.13
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -4249.1357                       Pseudo R2     = 0.6349

------------------------------------------------------------------------------
             |               Robust
       count | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     doxconc |  -1.409016    .136769   -10.30   0.000    -1.677078   -1.140954
       _cons |   5.114238   .0336559   151.96   0.000     5.048274    5.180203
------------------------------------------------------------------------------
```

(a) [3 pts] Because the doxorubicin concentrations considered in this study were discrete in nature, a 95% CI for the mean post-incubation colony frequency associated with a concentration of 1.00 µmol/L could be formed using just the observations evaluated at that concentration (for instance, via the Stata command `ci means count if doxconc == 1`, whereby we obtain a simple 95% CI of [25.2, 40.2]). If instead you formed a *model-based* 95% CI for this quantity, would you expect it to be wider or narrower compared to the simple 95% CI? Very briefly state the reason for your conjecture.

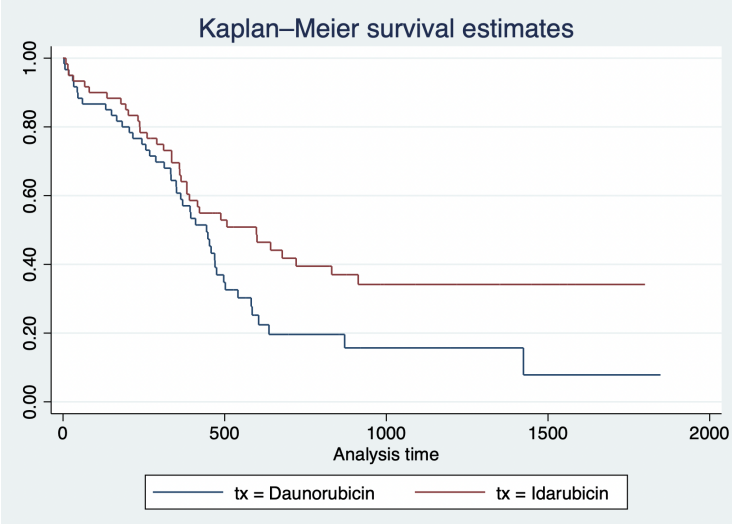
(b) [3 pts] Is it possible to use the Stata output above to form a model-based 95% CI for the post-incubation colony frequency at a concentration of 1.00 µmol/L? If so, do so. If not, briefly explain why not.

(c) [3 pts] Determine the concentration of doxorubicin corresponding to an expected post-incubation colony frequency of 80.

(d) [4 pts] The half-maximal inhibitory concentration (denoted as IC_{50}) is, by definition, the level of doxorubicin at which the expected post-incubation colony frequency is half the expected frequency under no exposure to doxorubicin. Determine a point estimate and a 95% CI for the IC_{50} in this study based on the Poisson regression model.

(e) [2 pts] Provide an example of a modification to study's design that, if implemented, would require us to include an offset term in the Poisson regression model.

4. [20 pts] A randomized controlled trial of $N = 120$ independently sampled adults was conducted to compare daunorubicin (`tx=0`) and idarubicin (`tx=1`) for treatment of acute myelogenous leukemia. Kaplan-Meier curves for time (days) to all-cause death are shown below.



(a) [2 pts] In the space provided, state a rough approximation of the estimated 60^{th} percentile of the survival distribution in each group.

- Daunorubicin: _____
- Idarubicin: _____

(b) [2 pts] In the space provided, state a rough approximation of the estimated proportion of patients in each group who die within three years.

- Daunorubicin: _____
- Idarubicin: _____

(c) [2 pts] Which treatment group has a higher estimated restricted mean survival time to $t = 1000$ days? Select your choice below.

- Daunorubicin: ◯
- Idarubicin: ◯

(d) [2 pts] Which treatment group appears to have a higher estimated hazard rate around the 500-day mark? Select your choice below.

- Daunorubicin: ◯
- Idarubicin: ◯

Parts (e)-(h) pertain to the following abridged output of a Cox model for all-cause death:

```
. stcox tx age, robust nolog

No. of subjects =    120                                 Number of obs =    120
No. of failures =     79

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. ratio   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          tx |    .630039    .143793    -2.02   0.043     .4028086    .9854534
         age |   1.018673   .0083252     2.26   0.024     1.002486    1.035121
------------------------------------------------------------------------------
```

(e) [3 pts] Is it possible to use the Stata output above to estimate the instantaneous hazard rate of all-cause death among 50 year-old adults receiving idarubicin? If so, do so. If not, explain briefly why not.

(f) [3 pts] Suppose that age had not been included as a covariate. If you had to guess, would you expect the (unadjusted) hazard ratio for treatment to be larger, smaller, or about the same as 0.630039? Briefly explain your response.

(g) [4 pts] State and briefly describe the two fundamental assumptions invoked in this time-to-event analysis.

(h) [2 pts] Briefly describe the extent to which the number of study subjects (N = 120) and the number of deaths (L = 79) play a role in the statistical power of this analysis.

5. 25 pts A group of investigators sought to compare the degree to which a SARS-CoV-2 booster elicited antibody (Ab) responses between healthy controls and patients receiving dialysis. The Ab test was performed at baseline on a large number of independently sampled healthy controls and dialysis patients, after which the SARS-CoV-2 booster was administered. The Ab test was then conducted six weeks and twelve weeks post-booster. When produced in the "wide" format, the data set comprises the following variables:

`id`	subject ID
`group`	(0 = healthy control; 1 = dialysis patient)
`Ab0`	SARS-CoV-2 antibody response at baseline
`Ab1`	SARS-CoV-2 antibody response six weeks post-booster
`Ab2`	SARS-CoV-2 antibody response twelve weeks post-booster

For simplicity, you may assume baseline Ab to be uncorrelated with group. Consider the following different approaches to analyzing/comparing the post-booster Ab levels:

(i) Two linear regression models applied separately to `Ab1` and `Ab2`, each with `group` as the sole predictor.

(ii) An approach similar to (i), adjusting for baseline Ab level (`Ab0`) with a linear term in each model.

(iii) Generalized estimating equations with `Ab1` and `Ab2` as outcomes (using a working independence structure), including `group`, time (the indicator function `1(t=2)`, for example), and a group-time interaction.

(iv) An approach similar to (iii), adjusting for baseline Ab level (`Ab0`) with a single linear term.

(a) 4 pts In the space provided, state the number of degrees of freedom used by each approach.

- Approach (i): _____
- Approach (ii): _____
- Approach (iii): _____
- Approach (iv): _____

(b) 4 pts Which approach(es) involve(s) saturated models? Check all that apply.

- Approach (i): ☐
- Approach (ii): ☐
- Approach (iii): ☐
- Approach (iv): ☐

(c) 4 pts Which approach(es) would produce both an unbiased estimate and a valid 95% CI for the difference in mean Ab between dialysis patients and healthy controls at each time? Check all that apply.

- Approach (i): ☐
- Approach (ii): ☐
- Approach (iii): ☐
- Approach (iv): ☐

(d) [4 pts] Which approach(es) would produce both an unbiased estimate and a valid 95% CI for the mean change in mean Ab from time $t = 1$ to $t = 2$ within dialysis patients? Check all that apply.

- Approach (i): ☐
- Approach (ii): ☐
- Approach (iii): ☐
- Approach (iv): ☐

(e) [3 pts] Compared to approach (i), which approach(es) more efficiently estimate(s) the difference in mean Ab between dialysis patients and healthy controls at each time? Check all that apply.

- Approach (ii): ☐
- Approach (iii): ☐
- Approach (iv): ☐

(f) [3 pts] Briefly justifying your response, would you expect the difference in mean Ab between dialysis patients and healthy controls at each time to be more efficiently estimated by approach (ii) or (iii)?

(g) [3 pts] Briefly describe a circumstance in which approach (iv) may be preferable to approach (ii) if the goal is to estimate the difference in mean Ab between dialysis patients and healthy controls at each time.

6. [10 pts] A study was conducted to evaluate whether T cell responses to myelin proteins could distinguish between patients with brain-predominant multiple sclerosis (n = 40) and healthy controls (n = 40). The T cell responses are given by MBP-IFNG, MPB-IL17, MOG-IFNG, and MOG-IL17, the details of which you do not need to worry about (but may recall from a prior problem set). The study investigators fit a logistic regression model to their full data set that includes a four-way interaction between all predictors (along with all lower-order terms). This model uses sixteen degrees of freedom.

(a) [4 pts] The investigators are very pleased to obtain an estimated area under the receiver operating characteristic curve (AUC) of 0.90. To what extent does this estimate of predictive performance reflect the quantity that the investigators truly care about? Briefly explain your response.

(b) [3 pts] If the investigators want to evaluate the out-of-sample predictive performance of their model, briefly outline (bullet form is fine) a simple strategy that the investigators could use to do so.

(c) [3 pts] Suppose that once the steps you've outlined in part (b) are implemented, the investigators obtain a far less impressive out-of-sample AUC of 0.55. Briefly describe why a penalized regression model (e.g., ridge or LASSO) would likely improve the out-of-sample predictive performance over the investigators' proposed model.

7. ***Optional problem***: This is an optional problem — please do not attempt it until you have completed and are satisfied with your answers to the required problems. A small amount of credit can be earned for correct responses.

Suppose the proportional hazards assumption is satisfied in the setting of a two-group comparison of a time-to-event outcome, T (that is, suppose it is known that $\lambda_1(t) = \lambda_0(t) \times e^{\beta}$ for some β). Show that the survival curves are parallel after undergoing the complementary log-log transformation (that is, show that $\log(-\log[S_1(t)]) = \log(-\log[S_0(t)]) + \beta$).

Pages 12 constitutes the required part of the take-home portion of the exam (open-notes and individual-effort). Please e-mail your final responses to me (e-mail: andrew.spieker@vumc.org) by the deadline (April 27 at 5:00p CDT). Further, please submit your Stata code as an appendix. The exam is not to be considered submitted until I confirm receipt by e-mail.

8. [20 pts] You're researching treatment strategies for patients with substance abuse. Patients are randomly and evenly allocated in a 2×3 factorial design, with `type` denoting specific type of treatment (0 = behavior modification therapy, 1 = psychotherapy) and `setting` denoting the setting in which the treatment is provided (0 = outpatient, 1 = day-treatment, 2 = inpatient). The outcome is an ordinal substance abuse severity score (SASS) measured from 1 to 6, with higher values indicating greater severity.

(a) [3 pts] Write a cumulative-logit proportional odds model that allows the effects of treatment setting and treatment type to interact. Is this a saturated model? Briefly explain your response.

The data set has been provided to you by email (`substance.csv`), and should include data from $N = 264$ participants. Use Stata to estimate the parameters of the model of part (a). This is the model you should use to answer each of remaining questions. **You don't need to report the results of this model fit, although all code should be included in your appendix. Further, you do not need to show any of your "regression math" for any of these problems.** Your responses should be very brief.

(b) [3 pts] Conduct a test to evaluate whether there is an overall association between treatment type and SASS (simply state your conclusion and report the p-value).

(c) [3 pts] Conduct a test to evaluate whether there is an overall association between treatment setting and SASS (simply state your conclusion and report the p-value).

(d) [3 pts] Conduct a test to evaluate whether treatment setting and type interact in their effects on SASS (simply state your conclusion and report the p-value).

(e) [3 pts] Conduct a test to evaluate whether patients receiving day-treatment and patients receiving in-patient therapy experience differential effects on SASS (simply state your conclusion and report the p-value).

(f) [3 pts] Perform (and briefly summarize the results of) an analysis in which you compare SASS between patients receiving outpatient psychotherapy and patients receiving day-treatment behavior modification therapy. Please include the proper measure of association, a 95% CI, and a proper statement regarding statistical strength of evidence.

(g) [2 pts] Obtain a model-based point estimate of the proportion of patients receiving day-treatment behavior modification therapy with a SASS of at least 4. Compare it to the actual proportion of patients in this subgroup who meet this condition in the data set. How does this square with your response to part (a)?

9. ***Optional problem***: This is an optional problem — please do not attempt it until you have completed and are satisfied with your answers to the required problems. A small amount of credit can be earned for correct responses. You have until Friday, April 29 to submit your response to this problem.

We have covered a wide range of topics this semester pertaining to regression analysis. There are certain ideas and themes that we have circled back to several times, one of which I describe below as follows:

Sometimes, two models/approaches/methods could reasonably be used to answer a scientific question. The choice between one method or another will often be guided by a trade-off, and how you choose to navigate that trade-off is typically very situation-dependent (for instance: you may need to consider how big your sample size is, how many covariates are in your model, how interpretable you want your results to be, and so on).

Identify **one** example of this phenomenon that we have learned in class (there are many, but I ask that you choose one only). Write a brief paragraph in which you do the following:

- Describe the setting, along with the two competing approaches you're thinking of.
- Describe the trade-off between the competing approaches you have in mind. State what information about the study and/or data would help you navigate that trade-off and how it would do so.

Because there are so many examples to choose from, I'm providing a sample response to help give you a sense of what I'm looking for (and to demonstrate that your response does not need to be terribly long).

Sample response: Suppose you seek to estimate the association between a predictor and a continuous outcome in a cross-sectional study, and that the error variance is proportional to the predictor. Ordinary least squares could be used to estimate the association, and is valid so long as the robust variance is used and linearity is closely approximated. On the other hand, weighted least squares (weights inversely proportionally to the predictor), could provide a more *efficient* estimate. There is reason to be cautious if the relationship between the exposure and outcome is non-linear, whereby the weighted least squares method changes the value of the quantity being estimated by placing more weight on the observations corresponding to lower values of the predictor (recall discussion on Slide 462). If linearity is satisfied, the degree of heteroscedasticity is great, *and* I am confident in mean-variance relationship, then efficiency gains from weighting can be substantive. Otherwise, I may be better off using ordinary least squares.