Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Collection of problems for Spring 2023 (Version: 04/10/2023)

Instructions: Please round numeric responses to a reasonable number of digits. A request to "perform an analysis" is a request for a write-up in which you state and interpret the point/interval estimates and summarize conclusions with appropriate inferential measures. For problems involving real data, the associated documentation has essential information; reading it carefully is considered part of the problem. Unedited software code/output should never be included in your response. However, code should be attached as an appendix. Please submit your word-processed responses via e-mail to us (that is, to marisa.h.blackman@vanderbilt.edu, siwei.zhang.1@vanderbilt.edu, and andrew.spieker@vumc.org) by 10:30a on the due date.

A1. Load the data set `sot-covid.csv` and read the documentation. For this problem, any proteins referenced were measured by ELISA, and *not* the bead-based immunoassay.

(a) Reporting a prior symptomatic SARS-CoV-2 infection was a study exclusion criterion. Nevertheless, there is always a possibility of asymptomatic infection. For reasons I won't go into now, a baseline (pre-vaccination) IgG to nucleocapsid of 0.4 ELISA units (EU) or higher was considered indicative of prior infection. For how many study participants was this threshold achieved? Make note of any obvious characteristics these subjects have in common (you're only looking to describe any "hit-you-in-the-face" similarities—you need not dig too hard and you should not perform formal statistical analysis). Drop these subjects from the data for the remainder of this problem.

(b) Obtain point estimates and 95% CIs for each of the following quantities (you don't need to do the long write-up for them):

[i.] Mean baseline IgG to ECD among SOT recipients.

[ii.] Mean baseline IgG to ECD among HCs.

[iii.] Mean IgG to ECD among SOT recipients three weeks following the second dose.

[iv.] Mean IgG to ECD among HCs three weeks following the second dose.

(c) Perform an analysis to evaluate whether the mean baseline IgG to ECD differs between SOT recipients and HCs. Please make certain your choice of an analysis does not involve any unnecessary assumptions.

(d) Repeat part (c), instead comparing ECD three weeks following the second dose.

(e) Briefly summarize parts what parts (c)-(d) suggest with respect to differences in humoral immunogenicity of the SARS-CoV-2 vaccine series between SOT recipients and HCs.

(f) Briefly summarize the assumptions essential for the results of part (c) and (d) to be valid.

(g) Report point estimates and 95% CIs for each of the following quantities:

[i.] Mean change in IgG to ECD from baseline to three weeks following the second dose among SOT recipients.

[ii.] Mean change in IgG to ECD from baseline to three weeks following the second dose among HCs.

(h) Perform an analysis to evaluate whether the mean change in IgG to ECD from baseline to three weeks following the second dose differs from zero among SOT recipients.

(i) Perform an analysis to evaluate whether the mean change in IgG to ECD from baseline to three weeks following the second dose differs from zero among HCs.

(j) Perform an analysis to evaluate whether the mean change in IgG to ECD from baseline to three weeks following the second dose differs between SOT recipients and HCs.

(k) How have parts (g)-(j) augmented what you've learned in parts (b)-(e)?

(l) Construct a figure capturing the key messages on which you've already commented. You need not mark statistical significance, but you should mark key summary measures (e.g., a median (IQR), or mean (95% CI)—be sure to clarify what you're doing).

B1. In this problem, you will conduct a simulation study (e.g., in R) to investigate what goes wrong with the usual formulation of a 95% CI for a mean difference with paired data (i.e., when independence does not hold). Let $N = 500$ denote the total sample size (and $n = N/2 = 250$ the sample size in each group). Let $(X_1, Y_1) \ldots, (X_n, Y_n)$ denote i.i.d. bivariate normal random variables, with $\mathbf{E}[X_i] = \mathbf{E}[Y_i] = 0$ and $\text{Var}[X_i] = \text{Var}[Y_i] = 1$ for $1 \leq i \leq n$. Note that $\delta = \mu_Y - \mu_X = 0$ regardless of the value of $\varrho = \text{Cor}(X_i, Y_i)$ $(0 \leq \varrho < 1)$. You need not mathematically derive anything for this problem.

(a) For $\varrho = 0.3$, generate data under this setup for $M = 50{,}000$ simulations (you will find the function `rmvnorm` helpful). Within each simulation, extract a 95% CI for the δ (do this using the CI from the function `t.test`; further note that the choice of equal or unequal variances is immaterial to this problem—please briefly explain why in your response). Determine the proportion of these CIs across simulations that include the true mean difference of zero (this proportion is generally referred to as the coverage probability). How does the coverage probability compare to would you would *hope* it to be?

(b) How would you expect your findings to change when $\varrho = 0.1$? Confirm your expectations by redoing this simulation with $\varrho = 0.1$ and explain why the result makes sense.

(c) How would you expect your findings to change when $\varrho = 0.5$? Confirm your expectations by redoing this simulation with $\varrho = 0.5$ and explain why the result makes sense.

(d) Comment briefly on the implications of this problem. For a scientifically savvy researcher using statistics as a tool in his or her research, what is the bottom line?

A2. Load the data set `sot-covid.csv`. Just as with problem A1, exclude participants with a baseline IgG to nucleocapsid of 0.4 EU or higher. Moreover, for ease of reading, let us use the shorthand notation "ECD3" to denote IgG to ECD three weeks following the second dose (it is labeled `ecd3` in the data set).

(a) Using simple linear regression, perform an analysis to evaluate whether the mean ECD3 differs between SOT recipients and HCs. Compare your answer to that of problem A1(d).

(b) Use the model of part (a) to construct a 95% CI for the mean ECD3 among SOT recipients. Compare your answer to that of problem A1(b)[iii].

(c) Use the model of part (a) to construct a 95% CI for the mean ECD3 among HCs. Compare your answer to that of problem A1(b)[iv].

A3. Again consider the data set `sot-covid.csv`. Just as with problem A1, exclude participants with a baseline IgG to nucleocapsid of 0.4 EU or higher. Again use the shorthand notation "ECD3" to denote IgG to ECD three weeks following the second dose.

(a) Create a scatter plot with age on the x-axis and ECD3 on the y-axis. Take an educated guess at the associated correlation (i.e., r). Then, determine the correlation and compare your answer to your educated guess (you will not be penalized for an incorrect guess, though your guess should at least share the same sign—positive or negative—as the truth).

(b) Using simple linear regression, perform an analysis to quantify the association between age and mean ECD3.

(c) Briefly summarize the assumptions necessary for your analysis of part (b) to be valid.

(d) Repeat part (b), restricting to HCs only.

(e) Use the model of part (d) to estimate the proportion of variability in ECD3 attributable to something other than age among HCs.

(f) The sample mean age among HCs is (approximately) 61.7. Use this together with the model of part (d) to determine the sample mean ECD3 across all HCs. Specifically, what key fact about simple linear regression allows you to do this?

(g) Use the model of part (d) to determine a point estimate and 95% CI for the difference in mean ECD3 between 60 and 61 year-old HCs.

(h) Repeat part (b), restricting to SOT recipients only.

(i) Determine the MSE associated with the model of part (h). Provide two interpretations: one that would be valid only under homoscedasticity, and another that would be valid even under heteroscedasticity.

(j) Use the model of part (h) to determine a point estimate and 95% CI for the difference in mean ECD3 between 68 and 72 year-old SOT recipients. Please do this "the easy way."

B2. This problem is about using simulations to understand sources of variability in estimating the slope coefficient of a simple linear regression model. Let $i = 1, \ldots, 4n$ index independent study subjects, and consider two study design scenarios and two outcome generation scenarios (for a total of four possible data generating mechanisms):

Study design (under each study design, note that $\mathbf{E}[X] = 0$ and $\text{Var}[X] = 1$):

i. X is generated randomly, as per an observational study: $X_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$.

ii. X is fixed, as per an experimental design: $X_i = \begin{cases} -\sqrt{2} & \text{for } i = 1, \ldots, n \\ 0 & \text{for } i = n+1, \ldots, 3n \\ \sqrt{2} & \text{for } i = 3n+1, \ldots, 4n \end{cases}$

Outcome generation mechanisms:

i. $Y_i = X_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 10)$.

ii. $Y_i = X_i^2 + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 10)$.

For each of the four possible data generating mechanisms:

- Set the seed using the command `set.seed(6312)` for reproducibility (you should do this for each of the four data generating mechanisms).
- Let $n = 60$ so that $4n = 240$ denotes the sample size of a data set, and generate ten-thousand such data sets under the given data generating mechanism.
- For each replicated data set, fit a simple linear regression model $\mathbf{E}[Y|X = x] = \beta_0 + \beta_1 x$ using OLS. Extract the estimated slope coefficient, $\widehat{\beta}_1$.
  - To make this go faster, consider programming this by hand instead of using the command `lm()`. To do this, define the design matrix (`X`) and outcome (`Y`) accordingly, and let your estimate of β_1 be given by the second element of the OLS estimator, `OLS <- solve(t(X) %*% X) %*% t(X) %*% Y`.
- Determine the variance of the ten-thousand slope estimates.

(a) Under which data generating mechanisms is the regression model correctly specified?

(b) Present a 2×2 table cross-tabulating the values of $\text{Var}[\widehat{\beta}_1]$ according to the study design (random vs. experimental) and outcome generation mechanism (linear vs. quadratic).

(c) Within each study design, compare the variances between the two outcome generation mechanisms and use heuristic arguments to account for any differences/similarities.

(d) Within each outcome generation mechanism, compare the variances between the two study designs and use heuristic arguments to account for any differences/similarities.

A4. This problem is a continuation of problem A3, which utilizes the data set `sot-covid.csv`. As noted in previous problems, please exclude participants with a baseline IgG to nucleocapsid of 0.4 EU or higher, and again use the shorthand notation "ECD3" to denote IgG to ECD three weeks following the second dose.

(a) Identify the key way in which the result of A3(b) does not align with A3(d) and A3(h).

(b) To what degree would you trust the model fit in A3(b) to reliably estimate the mean ECD3 among 52 year-old SOT recipients? Very briefly justify your response.

(c) Use the model in part (d) of problem A3 to determine a point estimate and 95% CI for the mean ECD3 among 65 year-old HCs. Use diagnostics to evaluate how well any necessary assumptions seem to hold. You should be explicit about what those assumptions are.

(d) Use the model in A3(h) to construct a 95% prediction interval for ECD3 among among 72 year-old SOT recipients. Use diagnostics to evaluate how well any necessary assumptions seem to hold. You should be explicit about what those assumptions are.

(e) Create a scatterplot with age on the x-axis and ECD3 on the y-axis, using distinct colors to distinguish SOT recipients from HCs. Use this, along with other key elements of your responses in this problem and problem A3, to discuss whether these analyses support the blanket claim that the SARS-CoV-2 immunogenicity is weaker in older subjects.

(f) Perform an analysis (using multiple linear regression) in which you determine the age-adjusted association between SOT status and ECD3 (specifically comparing SOT recipients to HCs). Note that this analysis turns what we've been talking about so far in this problem on its head by considering SOT status as the predictor of interest.

A5. Load the data set `verb.csv`, which is based on the Vanderbilt Emergency Room Bundle trial.

(a) Use simple linear regression to obtain a point estimate and 95% CI for the effect of VERB on SBP at the first follow-up (i.e., approximately 30 days post-baseline).

(b) Consider a model analogous to that of part (a), but this time adjusted for baseline SBP. The coefficient for VERB can be interpreted in two ways. State both of these interpretations and describe why they are both valid.

(c) Provide the literal interpretations of each of the other two coefficients from the baseline-adjusted model. Which, if either of them, possess meaningful, real-world interpretations? Briefly explain.

(d) What sort of advantage do you anticipate would come about from adjustment for baseline SBP? Briefly explain your response.

(e) Report the point estimate and 95% CI for the VERB coefficient as obtained from the baseline-adjusted model. Do you see evidence of the advantage you stated in part (d)?

(f) Use the estimated baseline-adjusted model to determine a point estimate and 95% CI for the mean SBP at the first follow-up among patients receiving VERB having a baseline SBP of 130 mm Hg.

(g) Use the estimated baseline-adjusted model to determine a point estimate and 95% CI for effect of VERB on mean SBP at the first follow-up among those having a baseline SBP of 130 mm Hg.

B3. This problem aims to enrich your geometric understanding of OLS with a setting simple enough to be visualized and done by brute force. Consider a "no-intercept" regression model $\mathbf{E}[Y|X_1 = x_1, X_2 = x_2] = \beta_1 x_1 + \beta_2 x_2$. Imagine you have $N = 3$ observations with covariate vectors $\mathbf{x}_1 = (2, 0)$, $\mathbf{x}_2 = (0, 2)$, and $\mathbf{x}_3 = (1, 2)$, and an outcome vector given by $\mathbf{y} = (6, 6, 0)$.

(a) Write the 3×2 design matrix, $\mathbf{X}$ (remember *not* to include the usual column of ones as there is no intercept in this model). What is the dimension of $\text{col}(\mathbf{X})$? Note: "$\text{col}(\mathbf{X})$" serves as shorthand notation for the linear subspace spanned by the columns of $\mathbf{X}$.

(b) Argue that $\mathbf{y} \notin \text{col}(\mathbf{X})$—that is, explain why $\mathbf{y}$ cannot be expressed as a linear combination of the columns of $\mathbf{X}$.

(c) Recall that we presented $\widehat{\mathbf{y}}$ as the "projection of $\mathbf{y}$ onto $\text{col}(\mathbf{X})$". Noting that $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, argue that $\widehat{\mathbf{y}} \in \text{col}(\mathbf{X})$—although you needn't compute it yet.

(d) Find a vector $\mathbf{x}_o$ such that $\mathbf{x}_o \perp \text{col}(\mathbf{X})$. *Hint*: Recall that the cross-product of two vectors ($\mathbf{a}$ and $\mathbf{b}$) gives a third vector ($\mathbf{c}$) that is orthogonal to both $\mathbf{a}$ and $\mathbf{b}$.

(e) Use part (d) to compute $\widehat{\mathbf{y}}$ by "brute force." *Hint*: The orthogonal projection of $\mathbf{y}$ onto $\text{col}(\mathbf{X})$ can be expressed as $\widehat{\mathbf{y}} = \mathbf{y} - [(\mathbf{x}_o^T\mathbf{y})/(\mathbf{x}_o^T\mathbf{x}_o)]\mathbf{x}_o$.

(f) Verify that your answer to part (e) is correct by computing $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.

(g) Verify that your answers to parts (e) and (f) are correct by loading the observations into R (for instance) and performing simple linear regression via ordinary least squares.

(h) Characterize all possible values, $\{\mathbf{y}^*\}$ of $\mathbf{y}$ such that $\widehat{\mathbf{y}}$ would be given by $(4, 2, 4)$. Which of these vectors gives the smallest possible variance for $\widehat{\boldsymbol{\beta}}$? Note: a description of such vectors is acceptable, and you need not mathematically prove your answers.

A6. This problem is a continuation of problem A5, which utilizes the data set `verb.csv`.

(a) Consider a model allowing baseline SBP (SBP_0 for ease of notation) to modify the effect of VERB on mean 30-day SBP. Provide a literal interpretation for each coefficient.

(b) Use the model of A6(a) to evaluate whether the effect of VERB on mean 30-day SBP is modified by SBP_0.

(c) Use the model of A6(a) to summarize evidence of an overall effect of VERB on mean 30-day SBP.

(d) Use each of the three models of A5(a), A5(b), and A6(a) to determine point estimates and 95% CIs for each of the following quantities (please present this in a 3×4 table, labeling the rows as "unadjusted model," "adjusted model," and "interaction model;" and label the columns as [i.] through [iv.]).

[i.] The mean 30-day SBP among control subjects with $SBP_0 = 135$ mm Hg.

[ii.] The mean 30-day SBP among control subjects with $SBP_0 = 145$ mm Hg.

[iii.] The effect of VERB on mean 30-day SBP among subjects with $SBP_0 = 135$ mm Hg.

[iv.] The effect of VERB on mean 30-day SBP among subjects with $SBP_0 = 145$ mm Hg.

A7. Load the data set `sot-covid.csv` yet again, excluding participants with a baseline IgG to nucleocapsid of 0.4 EU or higher, and letting "ECD3" denote IgG to ECD three weeks following the second dose. For reasons of sparsity, combine heart and lung into one category.

(a) Consider the following saturated regression model, which includes healthy controls:

$$\mathbf{E}[\texttt{ecd3}|\texttt{transplant group}] = \beta_0 + \beta_1 1(\texttt{Kidney}) + \beta_2 1(\texttt{Liver}) + \beta_3 1(\texttt{Heart/lung}).$$

Provide plain-language interpretations for each of the coefficients in the model, and then provide point estimates and 95% CIs in a table based on the fitted model.

(b) Use the model of part (a) to determine whether the study provides evidence of a difference in mean ECD3 between kidney and liver transplant recipients. Show the "regression math" that leads to your conclusions.

(c) Use the model of part (a) to determine whether the study provides evidence of a difference in mean ECD3 across kidney, liver, and heart/lung transplant recipients. Being very careful, show the "regression math" that leads to your conclusions.

(d) For reasons of sparsity, re-categorize number of immunosuppressants as (0/1/2+). Report a 4×3 cross-tabulation of re-coded immunosuppressant category and transplant type (including healthy controls as a transplant group).

(e) With part (d) in mind, present a saturated linear regression model that allows an interaction between organ transplant group and (re-categorized) immunosuppressant group. Interpret each coefficient, of which there should be seven (*fewer* than the number produced by Stata's `##` feature). You needn't estimate the parameters until part (f).

(f) Showing the "regression math" that leads to your answer, use the model you proposed in part (e) to determine a point estimate and 95% CI for the difference in mean ECD3 between heart/lung transplant recipients on one immunosuppressant and kidney transplant recipients on two immunosuppressants.

B4. Fill in the gaps associated with the collapsibility proof on Slide 267, justifying each step. Illustrate by simulation that this extends to the case in which X and Z are linearly uncorrelated. Consider predictors $X \sim \mathcal{N}(\mu = 0, \sigma^2 = 4)$ and $Z = X^2/3 + \gamma$, where $\gamma \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. Let $Y = X + Z + \epsilon$ denote the outcome, with $\epsilon \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. Generate a single replicate from this setup under sample sizes of $n = 10^k$ i.i.d. observations, where $k = 1, \ldots, 5$. Do $\widehat{\beta}_1^\star$ (unadjusted model) and $\widehat{\beta}_1$ (Z-adjusted model) appear consistent for the same quantity?

A8. Return to the VERB study (`verb.csv`). In problems A5 and A6, we glossed over the variability in follow-up times. Consider now a "spline-interaction" model in which you include: (i) a treatment indicator (VERB), (ii) a natural cubic spline on first follow-up time with knots at 20, 30, and 40 days, (iii) interactions between VERB and each basis term for follow-up time, and (iv) a linear term for baseline SBP. The model should have seven coefficients.

(a) Carefully obtain a point estimate and 95% CI for the effect of VERB on SBP 30 days post-baseline. Compare your answer to those of A5(a) and to A5(b).

(b) Carefully obtain a point estimate and 95% CI for the effect of VERB on SBP 10 days post-baseline. Why might the CI width be so different from that of part (a)? Support your answer with an exploration of the data.

(c) If you used a model that instead assumes the effect of VERB on mean SBP to vary linearly with follow-up time, would you expect the discrepancy between the intervals you obtained in parts (a) and (b) to be so large? Verify and heuristically justify your answer.

A9. Load the data set `hsct-flu.csv`. In this problem, we will focus on the A/H1N1 antigen.

(a) Perform an (unadjusted) analysis to evaluate whether a two-dose course of HD influenza vaccination produces a higher geometric mean HAI titer to A/H1N1 (visit 3) in pediatric HSCT recipients as compared to a two-dose course of SD vaccination.

(b) Repeat part (a), this time including a log-transformed baseline HAI titer to A/H1N1 as a covariate and placing a natural cubic spline on time post-transplant having three knots (chosen as the default knots provided by Stata). Why might it *not* be necessary to perform a whole bunch of in-depth diagnostics in order to trust the validity of the conclusions from this model?

(c) Use the models of parts (a) and (b) to form two 95% prediction intervals for (visit 3) HAI titer to A/H1N1 for pediatric patients receiving high-dose 14 months post-transplant and having a baseline titer of 1:160. Perform regression diagnostics as necessary and comment on the degree to which you believe the prediction intervals you formed are valid.

B5. Suppose you seek to model a process involving a predictor $X > 0$ and an outcome Y such that $\mathbf{E}[Y|X = x] = f(x)$ is a differentiable function that is constant for $x \in (0, c]$ and quadratic for $x > c$ (treat c as known). Write a basis expansion that would allow you to use simple linear regression to estimate f, writing $f(x) = \beta_0 + \beta_1 h(x; c)$ for some $h(x; c)$ that you determine. Show that f is continuously differentiable but *not* twice-differentiable. Once you've derived the basis, reflect (and comment on) why it is intuitive for this function to require two degrees of freedom. *Hint*: begin by writing down the most general form of the function that does not have specific constraints imposed. Then, solve for specific parameters by imposing a continuity and a differentiability constraint at $x = c$.

A10. Load the data from the Medicaid Work Requirements (MWR) data set, which comes from a survey experiment to evaluate physician attitudes toward work requirements for Medicaid eligibility as implemented in four U.S. states (`mwr.csv`). Note that is a substantive degree of missingness in these data. Please do not attempt to address the missing data, but instead use Stata's default approach of "available-case" analyses for each question. Note that at certain points in this problem, you are asked to justify your choices; what I mean by this is that you should describe *why* you made the choices you did at the outset (this is different than using the data to provide evidence in favor of or against an assumption, which I am not asking you to do in this problem).

(a) Construct a 2×2 cross-tabulation of randomized scenario (severity of depression presented in the vignette) and vignette response (recommendation regarding exemption). Note that you will need to group scenarios 0 and 2 together and scenarios 1 and 3 together (i.e., totally ignore the randomized duration of the patient-PCP relationship). Use logistic regression to obtain an estimated odds ratio and 95% CI. Confirm that the estimated odds ratio aligns with what you compute using the 2×2 table.

(b) Is it possible to use the model you fit in part (a) to estimate the odds of recommending an exemption among PCPs assigned to the mild depression scenario? If so, do so; if not, briefly explain why not.

(c) Is it possible to use the model you fit in part (a) to estimate the proportion recommending an exemption among PCPs assigned to the severe depression scenario? If so, do so; if not, briefly explain why not.

(d) Fit a model analogous to that of part (a), but adjusting for self-reported approval of MWR policy. You will need to make choices in your approach to this model. Carefully describe your choices in a way that your approach could be reproduced by someone who does not have your code; briefly justify your choices (you can list the choices and their justifications in bullet-form). In addition, identify the major scientific rationale for and mathematical consequence of adjustment for degree of approval.

(e) Perform an analysis to evaluate whether self-reported informedness regarding MWR policy modifies the association between severity of depression and odds of recommending an exemption. You will need to make choices in your approach to this analysis. Carefully describe your choices in a way that your approach could be reproduced by someone who does not have your code; briefly justify your choices (you can list the choices and their justifications in bullet-form).

(f) Develop a model to identify predictors of the odds of recommending an exemption for patients with severe depression only. Include age, gender, state, percent of patients receiving Medicaid, self-reported political affiliation, and self-reported approval of MWR policy as predictors, clearly stating and justifying any choices you make. Present and summarize your findings from this analysis.

(g) Develop a model to identify predictors of perceived degree of appropriateness of exemption for patients with mild depression only. Include age, gender, state, percent of patients receiving Medicaid, self-reported political affiliation, and self-reported approval of MWR policy as predictors, clearly stating and justifying any choices you make. Present and summarize your findings from this analysis.

A11. Load the data from the infertility study (`infert.csv`). Note that at certain points in this problem, you are asked to justify your choices; this means the same thing as in problem A10.

(a) Use logistic regression to determine whether this study provides sufficient evidence of an association between number of miscarriages (treated nominally) and odds of secondary infertility. Describe specifically how you are testing your hypothesis.

(b) Is it possible to use the model of part (a) to estimate the odds of secondary infertility among those with no prior miscarriages? If so, do so; if not, briefly explain why not.

(c) Is it possible to use the model of part (a) to estimate the odds ratio that compares the odds of secondary infertility between those with two prior miscarriages and those with one? If so, do so; if not, briefly explain why not.

(d) Is it possible to use the model of part (a) estimate the risk of secondary infertility among those with one prior miscarriage? If so, do so; if not, briefly explain why not.

(e) Is it possible to use the model of part (a) in order to "approximately" estimate the risk ratio that compares the risk of secondary infertility between those with one prior miscarriage and those with none? If so, do so; if not, briefly explain why not.

(f) Repeat part (a), this time adjusting for gravidity. You will need to make choices in your approach to this model. Carefully describe your choices in a way that your approach could be reproduced by someone who does not have your code; briefly justify your choices (you can list the choices and their justifications in bullet-form).

(g) Identify the major scientific rationale for and mathematical consequences of adjustment for gravidity. Take the study design into account when answering this question (read the documentation carefully).

A12. Load the data set `squamous.csv`. Perform an analysis to determine the association between (log-transformed) tumor volume and lymph node positivity rate per node removed (among those with at least one lymph node removed); adjust for age, gender, and p16 expression. Carefully describe your choices in a way that your approach could be reproduced by someone who does not have your code; briefly justify your choices (this means the same thing as in problem A10; you can list the choices and their justifications in bullet-form). Be very careful in your interpretation of the results; I recommend backing out of the log-transformation by comparing subgroups via base 2.

A13. Revisit problem A8. Read it through to refresh your memory on the setup of the problem. Now, we seek to leverage the availability of the second follow-up time as well, which was supposed to occur around the 90-day mark.

(a) Write down a mean model that is analogous to that of A8, reflecting the longitudinal outcomes and updating the knots to reflect the availability of additional data.

(b) Use GEE with working independence to estimate the parameters of the model. Using this model, carefully obtain a point estimate and 95% CI for the effect of VERB on SBP 30 days post-baseline. Compare your answer to those of A5(a), A5(e), and A8(a).

(c) Carefully obtain a point estimate and 95% CI for the effect of VERB on SBP 90 days post-baseline. After reading the documentation, does this finding surprise you?

A14. Revisit problem A1. Read it through to refresh your memory on the setup of the problem, and follow the instructions regarding dropping patients with evidence of prior infection.

(a) Write down a saturated linear model that encodes each of the quantities of A1(b), but also allows you to estimate mean group-specific responses to the *first* vaccine dose. Use GEE with working independence to produce point estimates and 95% CIs for its coefficients. Comment on the degree to which any of the estimates match those of A1(b).

(b) The model of part (a) is unusual as the left-hand side involves a mixture of pre- and post- exposure values. Consider instead a model with outcomes given by the changes in IgG to ECD from baseline to three weeks following each dose (so that each subject has two outcomes in the model instead of three). Write down the corresponding (saturated) linear model. Report the point estimates and corresponding 95% CIs for each coefficient (based on GEE with working independence) in a table.

(c) Re-do problems A1(h) and A1(i) using the model of A14(b); compare your results to those of A1(h) and A1(i), respectively.

(d) Using the model of A14(b), perform an analysis to evaluate whether the mean change in IgG to ECD from baseline to three weeks following the second dose among SOT recipients differs from the mean change in IgG to ECD from baseline to three weeks following the first dose among HCs.

B6. Let $i = 1, \ldots, N$ denote independent study subjects and let $t = 1, \ldots, T$. Let x_{it} denote age of subject i at time t and let Y_{it} denote SBP at the corresponding time. Suppose the outcomes are generated according to the form $Y_{it} = f(x_{i1}) + \beta(x_{it} - x_{i1}) + \epsilon_{it}$, where ϵ_{it} denote independent errors. The unknown function $f(\cdot)$ represents a general cohort effect, and β some longitudinal effect. Now consider the following mean model (which could be fit, for instance, using GEE):

$$\mathbf{E}[Y_{it} | \mathbf{X}_i = \mathbf{x}_i] \quad = \quad \alpha_0 + \alpha_1 x_{i1} + \alpha_2 (x_{it} - x_{i1}).$$

(a) State a sufficient condition on the function $f(\cdot)$ so that $\widehat{\alpha}_2$ from a GEE model would be unbiased for β. You do not need to prove your result; just argue the ideas.

(b) Suppose the condition you identified in part (a) is not met. State a sufficient condition on the spacing of follow-up times so that $\widehat{\alpha}_2$ from a GEE model would be unbiased for β. You do not need to prove your result; just argue the ideas. Support your statement with a brief simulation study.

A15. Load the data set `prostate.csv`, which involves patients with prostatic adenocarcinoma.

(a) Produce Kaplan-Meier curves for time to biochemical recurrence by cribriform status. Test the hypothesis of whether the distributions between groups are different.

(b) Report an estimate/95% CI for the median survival for those with and without IDC.

(c) Use a Cox proportional hazards model to evaluate morphology-related risk factors for biochemical recurrence. The following markers are of primary interest: percent of sample represented by Gleason Pattern 4; percent represented by Gleason Pattern 5; and indicators of cribriform, poorly formed, and glomeruloid patterns (all of which are sub-patterns of Gleason Pattern 4). Adjust for age and pathological stage. Summarize the most salient conclusions from this analysis.

A16. A randomized controlled trial is conducted to compare the survival rates in leukemia patients receiving either an experimental chemotherapy ($X = 1$) or standard of care ($X = 0$). In this example, we're pretending that we know the *true* survivor functions in each group, given as follows (there is no censoring in this problem, and we are not estimating anything):

$$\begin{aligned} S(t|X = 0) &= \mathrm{P}(T > t|X = 0) = \exp(-2t) \\ S(t|X = 1) &= \mathrm{P}(T > t|X = 1) = \exp(-t), \end{aligned}$$

where t is time in years. **You will not need calculus for any of the below problems**.

(a) Determine the probability of surviving beyond six months in each treatment group.

(b) Determine the probability of dying within nine months in each treatment group.

(c) Determine the median survival time in each group.

(d) Which group has greater one-year restricted mean survival time? How do you know?

(e) State the cumulative hazard functions, $\Lambda(t|X = 0)$ and $\Lambda(t|X = 1)$.

(f) State the hazard functions, $\lambda(t|X = 1)$ and $\lambda(t|X = 0)$ (these are given by the slopes of the cumulative hazard functions).

(g) Is the proportional hazards assumption satisfied? If so, state the hazard ratio (comparing group $X = 1$ to group $X = 0$).

B7. Repeat problem A14, but with the following two general survivor functions:

$$\begin{aligned} S(t|X = 0) &= \mathrm{P}(T > t|X = 0) = \exp(-(t/\lambda_0)^k) \\ S(t|X = 1) &= \mathrm{P}(T > t|X = 1) = \exp(-(t/\lambda_1)^k), \end{aligned}$$

where t is time in years, and λ_0, λ_1, and k are general positive constants.

A17. Load the data set `rrms.csv`. We seek to evaluate the extent to which certain T cell responses distinguish brain-predominant and spinal cord-predominant multiple sclerosis (MS) patients. Therefore, drop the healthy controls from this analysis. Please set any seeds to `2023` for reproducibility. Further, please first log-transform each of the T cell responses (treat $\log(0)$ as 0). Consider the following three models (though do not fit them just yet):

(I) A logistic model with IFNG-secreting cell response to MOG as a predictor.

(II) A logistic model with all T cell responses as predictors.

(III) A logistic model allowing a four-way interaction between the four T cell responses with a LASSO penalty chosen by five-fold cross-validation.

(a) Split the data into a training and test set with a 1:1 ratio. Fit each model on the training set. Report the training and test AUC for each model. Comment on the degree to which your findings square with what you might have anticipated *a priori*.

(b) Suppose a collaborator suggests to you that Model (III) is far too complicated and that it would be easier to just test the difference in means between groups for each of the four T cell responses. In a brief paragraph, propose a counterargument to this point of view, keeping the scientific goal in mind (however, *do* concede at least one merit to their perspective). This is an exercise in good collaboration practices.

B8. Let $\mathbf{X}$ denote an $N \times K$ design matrix in which the K variables have been centered and scaled (such that there is no leading column of ones for an intercept). Let $\mathbf{y}$ denote an $N \times 1$ vector of outcomes. The ridge regression estimator, $\widehat{\boldsymbol{\beta}}_\lambda$, is defined as the minimizer of the quantity $L_\lambda(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ for a fixed λ. Show that the problem of minimizing $L_\lambda(\boldsymbol{\beta})$ is equivalent to the problem of minimizing the objective function $\|\mathbf{y}' - \mathbf{X}'\boldsymbol{\beta}\|^2$ where $\mathbf{X}'$ denotes the matrix $\mathbf{X}$ augmented with K additional rows defined by $\sqrt{\lambda}\mathbf{I}$ and $\mathbf{y}'$ is the outcome vector $\mathbf{y}$ augmented with K zeros.

A18. Load the data set `enzyme.csv` and read the corresponding documentation.

(a) Use nonlinear least squares to obtain point estimates and 95% CIs for $v_{\max}$ and k.

(b) Obtain a point estimate and 95% confidence interval for the rate at a concentration of 0.1 ppm.

B9. This is a continuation of problem A18. You will find it helpful to conduct the software component (part (b), in particular) using R. Interestingly, despite the nonlinear relationship between S and V, it is possible to estimate $v_{\max}$ and k by using simple linear regression with suitable transformations of S and V as an intermediate step. We will focus in this problem on one such clever approach.

(a) Show that for some constants β_0 and β_1 (depending on $v_{\max}$ and k), the Michaelis-Menten model implies the following relationship:

$$\frac{1}{V} = \beta_0 + \beta_1 \times \frac{1}{S}.$$

Express $v_{\max}$ and k in terms of β_0 and β_1 (you need not use the data for this problem).

(b) Now, consider a simple linear regression model in which you let $Y = 1/V$ denote the outcome and let $X = 1/S$ denote the predictor:

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

Use simple linear regression to obtain point estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (please also report the 2×2 sandwich-based covariance matrix based on the `vcovHC()` function in the `sandwich` package with variance type `HC0`).

(c) Keeping in mind the relationship between $(v_{\max}, k)$ and (β_0, β_1) you noted in part (a), use your results of part (b) to obtain point estimates $\widehat{v}_{\max}$ and $\widehat{k}$.

(d) Use the delta method to obtain variance estimates, $\widehat{\text{Var}}(\widehat{v}_m)$ and $\widehat{\text{Var}}(\widehat{k})$. In turn, create 95% symmetric Wald-based CIs for $v_{\max}$ and k.