Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Collection of problems for Spring 2022 (Version: 04/12/2022)

Instructions: Please round numeric responses to a reasonable and appropriate number of digits. The request to "perform an analysis" is a request for a write-up in which you state and interpret the point/interval estimates and summarize your conclusions with appropriate inferential measures. For all problems involving real data, the associated documentation has essential information and so reading it carefully is considered part of the problem. Unedited software code or output should not be included as part of your response under any circumstance. However, code should be attached as an appendix. Please submit your word-processed responses via e-mail to all three of us (siwei.zhang.1@vanderbilt.edu, julia.c.thome@vanderbilt.edu, and andrew.spieker@vumc.org) by the deadline (10:30a on due date indicated on the syllabus).

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

(a) Reporting a prior symptomatic SARS-CoV-2 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) Report point estimates and 95% CIs for each of the following quantities:

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

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

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

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

(c) Perform an analysis to evaluate whether the mean baseline IgG to RBD differs between SOT recipients and HCs.

(d) Perform an analysis to evaluate whether the mean IgG to RBD differs between SOT recipients and HCs 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) Report point estimates and 95% CIs for each of the following quantities:

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

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

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

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

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

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

(k) 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. You may have noticed in your exploration of the data that the distribution of IgG to RBD did not seem to follow an approximate normal distribution. Some will say that the t-test should only be used for approximately normal outcomes. In this problem, you will conduct a simulation study (e.g., in R) in which you illustrate that the level of the two-sample t-test is approximately valid in large sample sizes when outcomes aren't normally distributed. Let n denote the total sample size (with $n_0 = n_1 = n/2$). Let $X_1, \ldots, X_{n_0}, Y_1, \ldots, Y_{n_1} \sim \text{Uniform}(0, 1)$ denote i.i.d. random variables such that the null hypothesis of no mean difference is true. Generate data under this setup with total sample sizes of $n = 6, 10, 20, 50, 100, 200$, and 500 for a total of $M = 50{,}000$ simulations. Within each simulation, extract the p-value from a two-sample t-test (with equal variances, as this assumption is satisfied and not the focus of this problem). For each sample size considered, determine the proportion of p-values at or below the nominal level of $\alpha = 0.05$ (this proportion is generally referred to as the type 1 error rate). Plot these proportions as a function of n. What do you notice? Now, repeat this for $X_1, \ldots, X_{n_0}, Y_1, \ldots, Y_{n_1} \sim \text{Exponential}(\lambda = 1)$ and comment on what you observe. Keep in mind that this simulation focuses only on *one* operating characteristic in only *two* scenarios, so broad claims are not appropriate. Do not mathematically derive anything for this problem.

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 "RBD3" to denote IgG to RBD three weeks following the second dose (it is labeled `rbd3` in the data set).

(a) Using simple linear regression, perform an analysis to evaluate whether the mean RBD3 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 RBD3 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 RBD3 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 "RBD3" to denote IgG to RBD three weeks following the second dose.

(a) Using simple linear regression, perform an analysis to quantify the association between age and mean RBD3.

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

(c) Repeat part (a), restricting to HCs only.

(d) Identify the key way in which the result of part (a) seems not to align with the results of parts (b) and (c).

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

(f) Use the model of part (b) to determine a point estimate and 95% CI for the mean RBD3 among 72 year-old SOT recipients.

(g) Use the model of part (c) to construct a 95% prediction interval for RBD3 among 65 year-old HCs. Use diagnostics to evaluate how well assumptions seem to hold.

(h) Use the model of part (c) to determine a point estimate and 95% CI for the difference in mean RBD3 between 60 and 70 year-old HCs. Please do this "the easy way."

(i) To what degree would you trust the model of part (a) to reliably estimate the mean RBD3 among 55 year-old SOT recipients? Very briefly justify your response.

(j) Create a scatterplot with age on the x-axis and RBD3 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, to discuss whether these analyses support the blanket claim that the SARS-CoV-2 immunogenicity is weaker in older subjects.

B2. The Michaelis-Menten relationship is a well known model of enzyme kinetics whereby the reaction velocity, V, is related to the substrate concentration, S, as follows:

$$V = \frac{v_{\max} S}{k + S}.$$

In this formula, $v_{\max}$ and k are real-valued constants representing the terminal velocity and how rapidly the reaction rises to its maximum rate (respectively). For a particular enzyme, the values of $v_{\max}$ and k are not known exactly but can be estimated in an experimental context.

To that end, suppose a group of investigators conduct an experiment in which they observe a collection of independent reaction velocities for several fixed substrate concentrations (data set: `enzyme.csv`). The variables in the data set are as follows:

`concentration`	substrate concentration, S, in ppm
`rate`	reaction velocity, V, in $(\text{mol}/\text{m}^3)/\text{sec}$

Though the substrate concentration values are fixed and known in advance, the reaction velocities are observed with some degree of random noise that you may assume not to depend upon the substrate concentration. Interestingly, despite the nonlinear relationship between S and V, it is possible to estimate $v_{\max}$ and k by using SLR with transformations of S and V as an intermediate step. We will focus in this problem on this clever approach, called the Lineweaver-Burke linearization method.

(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 $\widehat{\beta_1}$ based on the `enzyme.csv` data (please also report the 2×2 sandwich-based covariance matrix).

(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.

(e) Is it possible to easily obtain a 95% CI for either $v_{\max}$ or k another way? Where it is easy to do so, do so and report your results. *Hint*: The CI need not be symmetric.

(f) Briefly comment on the importance of using the sandwich variance in the problem.

A4. 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 adjusted for baseline SBP (SBP_0, for ease of notation). Describe the primary purpose of this adjustment. How would you expect the point estimate and 95% CI of the adjusted model to compare to those of the unadjusted model? Verify your suspicions, and report the point estimate and 95% CI.

(c) Very briefly describe the real-world circumstances under which one should include an interaction term between SBP_0 and VERB. Fit this model to answer parts (d) and (e).

(d) Use the model of part (c) to evaluate whether the effect of VERB on mean 30-day SBP is modified by SBP_0.

(e) Use the model of part (c) to determine whether there is evidence of an overall effect of VERB on 30-day SBP.

(f) Use each of the three models of parts (a), (b), and (c) 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 (a), (b), and (c), and the columns as [i.] through [iv.]).

[i.] The mean 30-day SBP among control subjects with $\text{SBP}_0 = 140$ mm Hg.

[ii.] The mean 30-day SBP among control subjects with $\text{SBP}_0 = 150$ mm Hg.

[iii.] The effect of VERB on mean 30-day SBP among subjects with $\text{SBP}_0 = 140$ mm Hg.

[iv.] The effect of VERB on mean 30-day SBP among subjects with $\text{SBP}_0 = 150$ mm Hg.

A5. 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 "RBD3" denote IgG to RBD three weeks following the second dose. For reasons of sparsity, combine heart and lung into one category.

(a) Consider the following regression model (keep in mind that healthy controls are included):

$$\mathbf{E}[\texttt{rbd3}|\texttt{transplant group}] = \beta_0 + \beta_1 1(\texttt{Kidney}) + \beta_2 1(\texttt{Liver}) + \beta_3 1(\texttt{Heart/lung}).$$

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 RBD3 across kidney, liver, and heart/lung transplant recipients. Being very careful, describe and/or show the "regression math" that leads to your conclusions.

(c) For reasons of sparsity, re-categorize number of immunosuppressants as (0/1/ 2+). Report a cross-tabulation of re-coded immunosuppressant category and transplant type (including healthy controls as a transplant group). With this in mind, determine a saturated model that allows an interaction between organ transplant group and (re-categorized) immunosuppressant group—be certain that all of the coefficients in your model can be estimated; your model will have fewer coefficients than that produced by Stata's `##` feature. Showing or describing the "regression math" that leads to your answer, perform an analysis to determine the difference in mean RBD3 between liver transplant recipients on two immunosuppressants and kidney transplant recipients on one immunosuppressant.

B3. This problem seeks 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 $\mathbf{x}_1 = (1, 0)$, $\mathbf{x}_2 = (0, 1)$, and $\mathbf{x}_3 = (1, 1)$, and an outcome vector given by $\mathbf{y} = (3, 3, 0)$.

(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, $\mathbf{y}$ 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 $\text{col}(\mathbf{X})$". Without actually computing it, write an expression for the vector $\widehat{\mathbf{y}}$ in terms of $\widehat{\boldsymbol{\beta}}$ (which you need not yet determine—you will do so later), and argue that, therefore, $\widehat{\mathbf{y}} \in \text{col}(\mathbf{X})$.

(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 $(1, 1, 2)$. 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. Return to data from the VERB study (`verb.csv`). In problem A4, we glossed over the variability in follow-up times and referred to the first follow-up outcome loosely as "30-day SBP." Consider instead a "spline-interaction" model in which you include: (i) a treatment indicator (VERB), (ii) a natural cubic spline on time with knots at the 10th, 50th, and 90th percentiles of first follow-up time, (iii) interactions between VERB and each of the basis terms for follow-up time, and (iv) a natural cubic spline on baseline SBP with three knots (chosen at the default percentiles provided by Stata). To check your work, this model should have a total of eight coefficients (including the intercept).

(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 A4(a) and to A4(b).

(b) Carefully obtain a point estimate and 95% CI for the effect of VERB on SBP 38 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) Using the model described above, carefully obtain a point estimate and 95% CI for the effect of VERB on SBP 0.5 days post-baseline, ignoring (just for pedagogical purposes) the fact that this is severe extrapolation and not clinically meaningful. Why might the CI width be so different from that of part (a)? Support your answer heuristically based on the model's mathematical assumptions.

A7. Load the data set `hsct-flu.csv`. In this problem, we will focus on the A/H3N2 antigen.

(a) Perform an analysis to evaluate whether a two-dose course of HD influenza vaccination is associated with a higher geometric mean HAI titer to A/H3N2 in pediatric hematopoietic stem cell transplant recipients as compared to a two-dose course of SD vaccination.

(b) Repeat part (a) with two important modifications: (1) adjust for (log-transformed) baseline HAI titer to A/H3N2, and (2) include a natural cubic spline on time post-transplant with three knots (chosen as the default knots provided by Stata). Account for the differences you see as compared to part (a). Perform regression diagnostics as necessary—with the important caveat that you should only conduct/describe/present the diagnostics that would be essential for you to trust the conclusions of your analysis.

(c) Use the models from parts (a) and (b) to form two 95% prediction intervals for (post-vaccine 2) HAI titer to A/H3N2 for pediatric patients receiving high-dose with a baseline HAI titer of 1:80 12 months post-transplant. Perform regression diagnostics as necessary—with the caveat that you should only conduct/describe/present the diagnostics that would be essential for you to trust the conclusions of your analysis.

B4. Suppose you seek to model a process involving a positive-valued predictor X and an outcome Y such that $\mathbf{E}[Y|X = x] = f(x)$ is a differentiable function that is constant for $0 < x \leq c$ and quadratic for $x > c$ (treat c as a known value—i.e., as a knot). Write a basis expansion that would allow you to use simple linear regression to estimate $f(x)$, and then show that $f(x)$ is continuously differentiable but *not* twice-differentiable. Once you've derived the basis, reflect (and briefly comment on) why it makes intuitive sense for this function to only require two degrees of freedom. As a 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 $x = c$.

A8. Load the data from the Medicaid Work Requirements (MWR) survey experiment (`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.

(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 severe 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 mild 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.

A9. Load the data from the infertility study (`infert.csv`).

(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).

A10. 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 (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.

A11. Load the data set `mri.csv`. We will examine the association between certain aspects of the MRI (specifically, those pertaining to infarcts) and all-cause death. Let X denote number of infarcts (0=none, 1=one, 2=at least two), and let Z denote total infarct volume (cm^3). Consider the following Cox proportional hazards model:

$$\log \lambda(t|X = x, Z = z) = \log(\lambda_0(t)) + \beta_1 1(X = 1) + \beta_2 1(X = 2) + \beta_3 1(X = 1)Z + \beta_4 1(X = 2)Z$$

(a) Characterize the subgroup to whom the baseline hazard function applies. Examine the model—there is something unusual about it. Is this a mistake, or does the model represent a necessary simplification? Interpret each of its coefficients and estimate them in Stata.

(b) Determine a point estimate and 95% CI for the hazard ratio that compares the hazard of all-cause death between those with a two infarcts each of volume 2 cm^3 and those with no infarcts.

(c) Determine a point estimate and 95% CI for the hazard ratio that compares the hazard of all-cause death between those with a single infarct of volume of 4 cm^3 and those with two infarcts each of volume 2 cm^3.

(d) Suppose your model treated number of infarcts linearly and did not allow an interaction between number of infarcts and volume. How would you expect your answers to parts (b) and (c) to compare? Verify your suspicions.

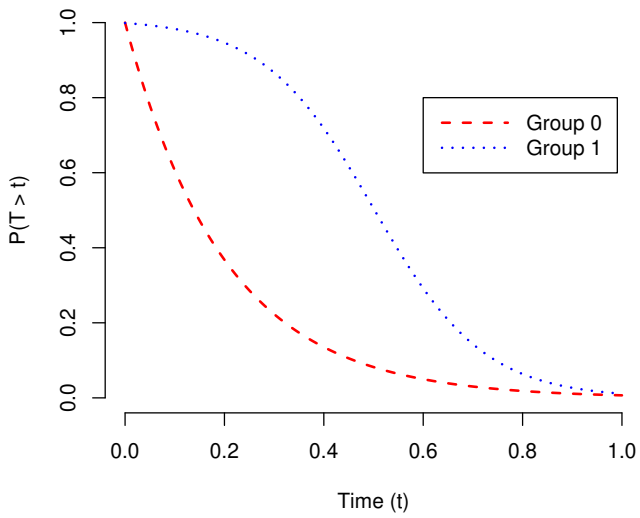
A12. Load the data set `prostate.csv`, which involves patients with prostatic adenocarcinoma.

(a) On a single plot, produce Kaplan-Meier curves for time to biochemical recurrence by subgroups defined by cribriform status. Use a non-parametric method to determine whether the distributions between groups are different, and summarize your conclusions.

(b) Report an estimate and 95% CI for the median survival time for those with and without intraductal carcinoma.

(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.

A13. Examine the survival curves below. Note: you needn't be fancy about your approximations in this problem.



(a) Argue heuristically/geometrically that the restricted mean survival time to time $t = 1$ is greater than 0.5 for Group 1, but less than 0.5 for Group 0.

(b) Approximate the median survival time in each group.

(c) Approximate 80th percentile of the survival distribution in each group.

(d) Rank the following quantities in order from highest to lowest (note that you need not compute or even approximate any of these quantities in order to answer this question).

(i) The instantaneous hazard rate for Group 0 at time $t = 0.001$.

(ii) The instantaneous hazard rate for Group 1 at time $t = 0.001$.

(iii) The instantaneous hazard rate for Group 0 at time $t = 0.05$.

(iv) The instantaneous hazard rate for Group 1 at time $t = 0.50$.

(e) Approximate the cumulative hazard in each group at time $t = 0.4$.

A14. Revisit problem A1. Read it through all the way in order to refresh your memory on the setup of the problem. Follow the instructions provided in part (a) 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 RBD 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. Comment on the degree to which any of the estimates match those of A1(f).

(c) Re-do problem A1(g) using the model of A14(b); compare your results to those of A1(g).

(d) Re-do problem A1(h) using the model of A14(b); compare your results to those of A1(h).

(e) Re-do problem A1(i) using the model of A14(b); compare your results to those of A1(i).

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

A15. Revisit problem A6. Read it through all the way in order 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 completely analogous to that of Problem A6, reflecting the longitudinal outcomes and updating the knots to reflect the availability of additional data.

(b) Use a random-intercepts model 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 A4(a), A4(b), and A6(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?

B5. In this problem, you will illustrate via simulation the limitations of including fixed intercepts for each subject. Let N denote the total number of subjects, each with subject-specific (time-stable) covariate $X_i \sim \mathcal{N}(0,1)$ and three outcomes, $Y_{it} = X_i + \gamma_i + \epsilon_{it}$, for $1 \leq t \leq 3$, where $\gamma_i \sim \mathcal{N}(0,1)$ is a subject-specific intercept, and $\epsilon_{it} \sim \mathcal{N}(0,1)$. Consider, in turn, the model $\mathbf{E}[Y_{it}|X_i = x_i] = \beta_0 + \sum_{j=2}^{N} \beta_j 1(i = j) + \beta_x x_i$, which can be fit with ordinary least squares linear regression. For various choices of N (I recommend $N = 5, 10, 20, 50$), generate $K = 1000$ replicates from the described data generating mechanism, fit the naive ordinary least squares model, and extract the sandwich variance estimate. For each sample size, determine the average estimate of β_x, the average estimated variance, and the empirical variance (i.e., the variance of $\widehat{\beta}_x$ across simulations. What conclusions do you draw?

A16. Load the data set `rrms.csv`. For this problem, 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. Any time there is an opportunity to set a seed in this problem, please set it to `6312` for reproducibility. Further, please log-transform each of the four T cell responses (treat the log of a zero as a zero), and then center/scale them to have mean zero and variance one. Consider the following three models (though do not fit them just yet):

(I) A logistic model with (transformed) IFNG-secreting cell response to MBP as a covariate.

(II) A logistic model with each of the four (transformed) T cell responses as covariates.

(III) A logistic model allowing a four-way interaction between the four transformed T cell responses (and all lower-order interactions) with a ridge penalty chosen by five-fold cross-validation.

(a) Split the data into a training set and a test set with a 1:1 ratio. Fit each of the three models 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.

B6. Let $\mathbf{X}$ denote an $N \times p$ design matrix in which the p 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 $\widehat{\boldsymbol{\beta}}_\lambda$ is equivalent to the ordinary least squares estimate based on an augmented data set $(\mathbf{X}', \mathbf{y}')$, where $\mathbf{X}'$ denotes the matrix $\mathbf{X}$ augmented with p additional rows defined by $\sqrt{\lambda}\mathbf{I}$, and $\mathbf{y}'$ is the outcome vector $\mathbf{y}$ augmented with p zeros.