Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis (Spring 2021)
Problem set collection: Last updated 4/09/2021 at 12:06p

General instructions: For problems referring to real data sets, you are expected to refer to the corresponding documentation. Round numeric responses to a reasonable number of digits. Problems asking you to perform an analysis are asking you to provide a three-four sentence write-up in which you state/interpret the point and interval estimates and summarize your conclusions with appropriate inferential measures. Unedited Stata output should not be included as part of your response. Code should be attached as an appendix. Please e-mail your responses to caroline.i.birdrow@vanderbilt.edu and jamie.g.joseph@vanderbilt.edu by the deadline.

---

1. The REACH study (`reach.csv`) was conducted to evaluate whether text messages pertaining to medication adherence and self-efficacy could reduce mean hemoglobin A1c (%) in patients with type 2 diabetes.

   (a) Perform an analysis to compare mean six-month A1c between treatment groups, assuming equal variances between groups.

   (b) Perform an analysis to compare mean six-month A1c between treatment groups, allowing unequal variances between groups. Compare your results to those of part (a). State which analysis—(a) or (b)—you would have chosen *a priori*, very briefly justifying your response.

   (c) Perform an analysis in which you compare geometric mean A1c between treatment groups at six months, very briefly justifying any choices you make in your analysis.

2. Load the data set `mri.csv`. We will consider several ways to examine the association between age and serum low-density lipoprotein (LDL). For this problem, define age groups in a dichotomous fashion as younger (< 74 years) and older (≥ 74 years).

   (a) Perform an analysis to compare mean serum LDL between age groups.

   (b) Perform an analysis to compare geometric mean serum LDL between age groups.

   (c) Perform an analysis to compare the proportion with serum LDL exceeding 160 mg/dL between age groups.

   (d) Which of these three methods would you have chosen *a priori*? Very briefly justify your response.

3. Student's t-test with nominal level α is conservative when the group with the smaller sample size has lower variance, meaning $\text{P}(\text{Reject } H_0 | H_0 \text{ true}) < \alpha$. When the group with the smaller sample size has higher variance, Student's t-test is anti-conservative, meaning $\text{P}(\text{Reject } H_0 | H_0 \text{ true}) > \alpha$. Conduct a simulation study (e.g., in R) in which you illustrate this. I suggest the following setup. Generate normally distributed data of mean zero in each group and use sample sizes of $N_0 = 50$ and $N_1 = 100$. Determine the proportion of times out of 100,000 independent replicates in which Student's t-test rejects H_0 at the $\alpha = 0.05$ level when $\sigma_0^2 = 400$ and $\sigma_1^2 = 4$; repeat when $\sigma_0^2 = 4$ and $\sigma_1^2 = 400$. No mathematical proofs are expected or required. Attach your code as part of your response to this problem.

4. A cross-sectional study of adults over 65 years old was conducted. Consider a linear model of the association between `height` (inches) and mean systolic blood pressure (`sbp`, mmHg):

$$\mathbf{E}[\texttt{sbp}|\texttt{height}] = \beta_0 + \beta_1 \texttt{height}$$

   (a) State literal, plain-language interpretations for β_0 and β_1.

   (b) Thinking in terms of both the data *and* the modeling assumptions necessary for validity, summarize the conditions under which you would trust an estimate of β_1 obtained from ordinary least squares simple linear regression with robust standard errors.

   (c) Suppose we estimate β_1 as $\widehat{\beta}_1 = 0.261$ (95% CI: $[0.065, 0.457]$; $p = 0.0091$). Explain at least two things wrong with deducing that systolic blood pressure increases as you grow.

   (d) Consider a simple linear model in which both `height` and `sbp` are *log-transformed*:

$$\mathbf{E}[\texttt{log(sbp)}|\texttt{height}] = \beta_0 + \beta_1 \log(\texttt{height})$$

   State a plain-language interpretation for $\exp(\beta_1 \log(2.5)) = 2.5^{\beta_1}$ based on this model.

5. Load the FEV data set (`fev.csv`), which we have considered previously in the course notes.

   (a) Use simple linear regression to form a prediction interval for FEV for those with a height of 50 inches. State the assumptions invoked and evaluate how well they appear to hold (include at most two diagnostic plots as part of your response).

   (b) Repeat part (a), but instead of using height as the predictor, use height^3 (please divide height^3 by 1000 to prevent regression coefficients from being too small).

   (c) While the assumptions may not be perfectly satisfied in either, which key assumption is clearly more closely satisfied in the model of part (b) as compared to part (a)?

6. Again consider the FEV data set (`fev.csv`). Consider a simple linear regression model with FEV as the outcome and age as the predictor. Using the results of an OLS fit:

   (a) State the estimated difference in mean FEV comparing subgroups of ages 8 and 12 years.

   (b) State the root mean squared error with an interpretation that assumes homoscedasticity.

   (c) State an interpretation of the the root mean squared error that would be considered proper even if you were *not* willing to assume error homoscedasticity.

7. Consider the MRI study (`mri.csv`), previously considered in Problem 2.

   (a) Using OLS simple linear regression, perform an analysis to compare mean serum LDL between age groups defined in a dichotomous fashion as younger (< 74 years) and older (≥ 74 years). Compare your results to those of Problem 2(a).

   (b) Perform an analysis to compare mean serum LDL across age groups (i.e., treating age *continuously*). What advantage(s) does this analysis possess over the analysis you performed in part (a) above and in Problem 2(a)?

   (c) Perform an analysis to compare geometric mean serum LDL across levels of age. Briefly justify any choices you make in your analysis.

8. An α-level test of β_1 from a simple linear regression model that assumes homoscedasticity is conservative when X is skewed and the outlying values have lower variance, and anti-conservative when the outlying values have higher variance. Conduct a simulation study in which you illustrate this. I suggest the following setup. Consider a sample size of $N = 100$, with $X \sim \text{Exponential}(1)$, and $Y \sim \mathcal{N}(\mu = 0, \sigma^2 = g(X))$. First, let $g(X) = X^{-2}$ (the case in which the outlying values have lower variance) and determine the proportion of times out of 10,000 independent replicates in which simple linear regression rejects H_0 at the $\alpha = 0.05$ level. Then, repeat when $g(X) = X^2$ (the case in which the outlying values have higher variance). Attach your code as part of your response to this problem.

9. Load the data set (`fev.csv`); we will examine the association between smoking and FEV.

   (a) Write a linear regression model with FEV as the outcome and smoking status as the predictor (be mindful of the coding of smoking status). Provide a plain-language interpretation for each coefficient; report point estimates and 95% confidence intervals for each (based on OLS with robust standard errors). Does the direction of the association align with what you might expect? Besides random variation, what might explain this?

   (b) Perform an analysis to compare mean FEV between smokers and nonsmokers of the same age. State any assumptions you invoke; use (and comment on) diagnostics to provide justification for reasonableness of and/or evidence of violations to those assumptions.

   (c) Perform an analysis to compare mean FEV between smokers and nonsmokers of the same age using weighted least squares, with weights proportional to age raised to the fifth power. Discuss the consequences of this unusual weighting scheme, and summarize the degree to which your conclusions change from those you derived in part (b).

10. We will use the MRI data (`mri.csv`) to model cognitive function, as measured by the digit/symbol substitution test (DSST).

    (a) Consider a regression model for mean DSST that includes stroke category and sex as categorical predictors of interest, allowing interaction between them. Write the linear regression model, and provide a plain-language interpretation for each of its six coefficients.

    (b) Which parameter(s) should be tested to evaluate the association between stroke category and DSST? Conduct this test and report the results.

    (c) Which parameter(s) should be tested to evaluate the association between sex and DSST? Conduct this test and report the results in plain language.

    (d) Which parameter(s) should be tested to evaluate whether the association between stroke category and DSST is modified by sex? Conduct this test and report the results.

    (e) Which parameter(s) should be tested to evaluate whether the association between sex and DSST is modified by stroke category? Conduct this test and report the results.

    (f) Which parameter(s) should be tested to evaluate the association between stroke category and DSST among females? Conduct this test and report the results.

    (g) Showing your work, which linear combination of parameters should be tested to evaluate the association between stroke category and DSST among *males*? Conduct this test and report the results in plain language. Then, in a maximum of one sentence, describe how this association could have been evaluated with a slightly modified model.

11. Consider the REACH study (`reach.csv`).

    (a) Perform an (unadjusted) analysis to evaluate whether receiving REACH has an effect on mean A1c twelve months post-baseline, relative to control.

    (b) Repeat problem (a), this time adjusting for baseline A1c. Compare your results to those from part (a). What is the major advantage of the adjusted model?

    (c) Write down a linear regression model that could be used to determine whether baseline A1c modifies the effect of REACH on mean twelve-month A1c. Then, evaluate whether there is evidence of an effect of REACH among individuals having a baseline A1c of 7.5% (there are two ways to approach this problem—I don't particularly care which one you use as long as you understand both).

    (d) Perform an analysis to determine the effect of REACH among subjects with a baseline A1c of *at least* 7.5%, adjusting for baseline A1c.

    (e) All subjects in the REACH group received the text-message based intervention; a subset received additional family-based coaching (FAMS). Suppose we are interested in evaluating whether, adjusted for baseline A1c, FAMS has an *additional* effect on mean twelve-month A1c beyond that of REACH alone. Consider two approaches to addressing this problem — one in which you use data from only the REACH subjects and one in which you use data from the whole sample.

$$\begin{aligned} \mathbf{E}[\texttt{a1c}_{12}|\texttt{reach} = 1, \texttt{fams}, \texttt{a1c}_0] &= \beta_0 + \beta_1 1(\texttt{fams} = 1) + \beta_2 \texttt{a1c}_0. \\ \mathbf{E}[\texttt{a1c}_{12}|\texttt{reach}, \texttt{fams}, \texttt{a1c}_0] &= \beta_0 + \beta_1 1(\texttt{reach} = 1) + \beta_2 1(\texttt{fams} = 1) + \beta_3 \texttt{a1c}_0. \end{aligned}$$

    Discuss how each of these models could be used to answer the question of interest, and comment on the relative advantages of each approach.

12. Conduct a simulation to illustrate the Gauss-Markov theorem. I suggest the following setup. Let $N = 500$; let $X \sim \mathcal{U}(0.5, 3)$, $\epsilon \sim \mathcal{N}(0, \sigma^2 = g(X))$, and let $Y = X + \epsilon$ (so that $\beta_0 = 0$ and $\beta_1 = 1$). Consider two ways of generating data: (1) $g(X) = 4$, and (2) $g(X) = 4X^2$. For each case, consider two ways to estimate β_1: (1): unweighted ordinary least squares, and (2) weighted least squares with weights given by $w(X) = X^{-2}$. Finally, consider two kinds of standard errors: robust and non-robust. Based on 10,000 simulation replicates, fill in the rightmost three columns of the table below.

Variance	Weights	Robust SE	$\mathbf{E}[\widehat{\beta}_1]$	$\text{SD}(\widehat{\beta}_1)$	$\mathbf{E}[\widehat{\text{SE}}(\widehat{\beta}_1)]$
$g(X) = 4$	$w(X) = 1$	N	–	–	–
$g(X) = 4$	$w(X) = 1$	Y	–	–	–
$g(X) = 4$	$w(X) = X^{-2}$	N	–	–	–
$g(X) = 4$	$w(X) = X^{-2}$	Y	–	–	–
$g(X) = 4X^2$	$w(X) = 1$	N	–	–	–
$g(X) = 4X^2$	$w(X) = 1$	Y	–	–	–
$g(X) = 4X^2$	$w(X) = X^{-2}$	N	–	–	–
$g(X) = 4X^2$	$w(X) = X^{-2}$	Y	–	–	–

Note: $\mathbf{E}[\widehat{\beta}_1]$ refers to the average coefficient estimate across simulations, $\text{SD}(\widehat{\beta}_1)$ refers to the empirical standard error (i.e., the standard deviation of the estimates across simulations), and $\mathbf{E}[\widehat{\text{SE}}(\widehat{\beta}_1)]$ refers to the average estimated standard error across simulations. Briefly comment on your findings. Attach your code as part of your response to this problem.

13. Once again, consider the REACH study (`reach.csv`).

    (a) Consider a logistic regression model that compares the odds of a six-month A1c exceeding 8.0% between those receiving and not receiving REACH. Provide a point estimate and 95% confidence interval for the odds ratio.

    (b) Is it possible to use the results of (a) to predict the odds of a six-month A1c exceeding 8.0% among those receiving REACH? If so, do so; if not, briefly explain why not.

    (c) Is it possible to use the results of (a) to predict the risk of a six-month A1c exceeding 8.0% among those not receiving REACH? If so, do so; if not, briefly explain why not.

    (d) Is it possible to use the results of part (a) to approximate the risk ratio that compares the risk of a six-month A1c exceeding 8.0% between those receiving and those not receiving REACH? If so, do so; if not, briefly explain why not.

    (e) Repeat part (a), this time adjusting for baseline A1c. Apart from random variation, what is the most plausible reason for any differences you see?

    (f) Perform an analysis to evaluate whether baseline A1c modifies the association between REACH and odds of a six-month A1c exceeding 8.0%.

14. Load the data set `esoph.csv`, which examines risk-factors for esophageal cancer.

    (a) Briefly justifying any choices you make, perform an analysis to evaluate the association between tobacco consumption and odds of esophageal cancer, adjusting for alcohol consumption. What do you expect to be the most likely advantage of adjusting for alcohol consumption?

    (b) Can the results of part (a) be used to predict the odds of esophageal cancer among those who consume neither tobacco nor alcohol? If so, do so; if not, briefly explain why not.

    (c) Is it possible to repeat part (a) on the *risk* scale instead of the odds scale? If so, do so; if not, briefly explain why not.

15. Load the data set `cac.csv`. This problem involves an investigation of whether kidney stone history (exposure) is associated with coronary artery calcification (CAC; outcome), *adjusting for age and race category*. You'll conduct it in three different general ways—one for each of parts (a), (b), and (c). However, you'll still have to make some choices. Therefore, for each of parts (a)-(c), begin your response with a clear, concise description of the model you're fitting in a way such that your results could be reproduced; include brief justifications for any choices you make.

    (a) Treat kidney stone history continuously and treat CAC as a binary variable.

    (b) Treat kidney stone history as a binary variable and treat CAC nominally (as an unordered categorical variable). Don't bother reporting all the odds ratios and confidence intervals.

    (c) Treat kidney stone history nominally and treat CAC as an ordinal variable.

    (d) Briefly summarize some relative advantages and disadvantages of each of the approaches above. You don't need to cover everything, but you *should* be able to identify at least one advantage and at least one limitation of each of the three approaches.

16. Load the data set `chemo.csv`. For this problem, act as if the addition of chemosensitizers is *not* a part of this experiment, and just focus on the variables `doxconc` and `count`. Moreover, please restrict your analysis to the set of experiments using 10% calf serum only. Consider the following simple Poisson regression model:

$$\log(\mathbf{E}[\texttt{count}|\texttt{doxconc}]) = \beta_0 + \beta_1 \texttt{doxconc}.$$

(a) Express the IC_{50} of doxorubicin (described in the documentation) in terms of β_1.

(b) Use Stata to estimate the parameters of this model and, in turn, the IC_{50} for doxorubicin. Form a 95% CI for the IC_{50} by transforming the endpoints of a 95% CI for β_1.

(c) Show that under a simple linear model, the IC_{50} depends on both β_0 and β_1. Why might it be more challenging to derive a confidence interval for IC_{50} this way? State an additional limitation of using simple linear regression to estimate the IC_{50}.

17. Conduct a simulation study to illustrate non-collapsibility of the logit link. I recommend the following setup. Let $X, Z \overset{\text{indep.}}{\sim} \text{Bernoulli}(p = 0.5)$ denote independent covariates, and $Y \sim \text{Bernoulli}(p = \text{expit}(-2.5 + 0.7X + 1.5z))$. Simulate under this setup *one* time under an extremely large sample size (e.g., $N = 10{,}000{,}000$). Then, fit two log-odds models, one including X and Z, and one including X only (it will take a minute for these models to run). How do the coefficients corresponding to X compare? Attach your code as part of your response to this problem.

18. Suppose we conduct a five-year study of advanced-stage cancer patients for which the *true* survival and censoring times are represented by the following distributions:

$$S(t) = \text{P}(T^0 > t) = \exp(-t/5)$$

$$C(t) = \text{P}(C^0 > t) = \begin{cases} 1 - t/20 & \text{if } t < 5 \\ 0 & \text{if } t \geq 5 \end{cases},$$

where t is time in years. The censoring distribution reflects uniform enrollment times and administrative censoring at five years; you may assume censoring to be non-informative. Suppose a Kaplan-Meier curve is fit to the study data. For each of the following quantities, state whether you would expect to be able to estimate it from the Kaplan-Meier curve under a sufficiently large sample size.

(a) The probability of surviving past one year.

(b) The probability of dying within six years.

(c) The median survival time.

(d) The 95th percentile of the survival distribution.

(e) The restricted mean survival time up to three years.

(f) The restricted mean survival time up to five years.

(g) The restricted mean survival time up to seven years.

(h) The overall mean survival time.

19. Load the data set `leuk.csv`. These data come from a randomized trial to compare two chemotherapy agents for leukemia (daunorubicin and idarubicin).

    (a) On one plot, provide Kaplan-Meier curves for survival among those receiving each treatment. Would you expect a point estimate of the hazard ratio ($\lambda(t|\texttt{tx}=1)/\lambda(t|\texttt{tx}=0)$) to be greater than or less than one? Briefly justify your response.

    (b) Perform an analysis in which you compare the hazard of death between the two treatment groups. Evaluate the proportional hazards assumption in this example.

    (c) Fit a Cox model (with time to death as the outcome) that allows an interaction between treatment and baseline white blood cell count. Using an appropriate post-estimation command, state a point estimate and 95% confidence interval for the a hazard ratio that compares the two treatment groups having a white blood cell count of 30,000 cells/mm^3. You may find the option `hr` useful.

    (d) Accounting for death as a competing risk, perform an analysis in which you compare the cumulative incidence of complete remission between treatment groups, adjusting for age and gender. Be careful about how you code the competing risk—the variable indicating death does not distinguish between (1) a death that follows a previously observed complete remission, and (2) a death that occurs with no prior complete remission.

    (e) Using the model you estimated in part (d), state a point estimate and 95% confidence interval for the subdistribution hazard ratio that compares subgroups of the same treatment and gender, but differing in age by two years.

20. Load the data set `uterine.csv`. You will investigate whether adjuvant radiation therapy improves survival in women with confirmed uterine cancer following hysterectomy.

    (a) Perform an analysis to evaluate whether adjuvant radiation therapy is associated with a reduced hazard of death.

    (b) Perform an analysis to evaluate whether cumulative dose of adjuvant radiation therapy is associated with a reduced hazard of death. You may find the following command useful: `bysort id (t) : gen cumulrt = sum(rt)`.

21. Earlier in the course, we noted that linear regression does not require *perfect* linearity in order to be useful. Conduct a simulation study in which you demonstrate that a Cox proportional hazards model is still able to correctly provide insights into an overall improvement in survival even under a pretty serious violation to the proportional hazards assumption. I suggest the following setup. Consider two groups, defined by $X = 0$ and $X = 1$, with respective survivor functions given by $S_0(t) = \sqrt{1-t^2}$ and $S_1(t) = 1 - \sqrt{1-(t-1)^2}$ (each defined for $0 \leq t \leq 1$). Graph these two functions in the same plot; convince yourself that (1) the proportional hazards assumption clearly violated, and (2) the group defined by $X = 0$ clearly has the higher mean survival time. You need not compute the hazards or the means for either group—convincing yourself on the basis of the plot is sufficient. Consider a Cox model, $\log \lambda(t|X = x) = \lambda_0(t)\exp(\beta x)$. Without even generating data to verify this, observation (2) alone should suggest that $\beta > 0$ (put another way, $\exp(\beta) > 1$). Now, verify this by generating $N = 500{,}000$ data points (no censoring) from each of these survivor distributions, and fit a simple Cox model to these data. Report your results. Attach your code as part of your response to this problem.

22. Consider the MRI study (`mri.csv`). We seek to compare the "predictive ability" across models for global brain atrophy, defined as an atrophy score of at least 36; let $Y = 1(\texttt{atrophy} \geq 36)$ denote the indicator of atrophy. Consider the following models for Y:

    (I) A logistic model with only an intercept.

    (II) A logistic model including a restricted cubic spline on age (with knots at 68, 74, and 81 years).

    (III) A logistic model including terms for age, gender, race category, weight, coronary heart disease category, stroke category, low-density lipoprotein, blood albumin, blood creatinine, and FEV, including a LASSO penalty selected by ten-fold cross-validation.

    (a) Split the data into random halves, the first being a "training" and the second being a "test" set, (please use a seed of $s = 6312$ at this stage for reproducibility and consistency). In turn, fit Models (I)-(III) to the training data (please use a seed of $s = 2021$ for any cross-validation-based tuning parameter selection). Determine and report the training and test AUC for each model. For the LASSO model, you will not be able to do this using the `lroc` command, but you may find the `roctab` command useful.

    (b) Briefly discuss your findings, commenting on how the results align with your expectations.

23. Load the data set `immunogenicity.csv`. Consider the following linear model of the log-transformed titer outcomes, with $t = 1, 2$ indexing the outcomes immediately prior to the second vaccination and 180 days following the second vaccination, respectively. Note that $t = 0$ indexes the baseline observations occuring immediately prior to the first vaccination:

    $$\mathbf{E}[\log(\texttt{titer}_t)|\texttt{titer}_0, \texttt{dose}] = \beta_0 + \beta_1 \log(\texttt{titer}_0) + \beta_2 1(\texttt{t} = 2) + \beta_3 \texttt{dose} + \beta_4 1(\texttt{t} = 2)\texttt{dose}.$$

    (a) State plain-language interpretations for each of the following parameters:

        (i) $\exp(\beta_1)$

        (ii) $\exp(\beta_3)$

        (iii) $\exp(\beta_4)$

    (b) Use GEE with working independence (and post-estimation commands such as `test`, `lincom`, etc. as appropriate) to form point estimates and 95% confidence intervals for geometric mean ratios that compare the following groups (adjusted for baseline titer); define the parameter being estimated in each comparison and express it in terms of $\boldsymbol{\beta}$.

        (i) One round of high dose to one round of standard dose.

        (ii) Two rounds of high dose to two rounds of standard dose.

        (iii) Two rounds of high dose to one round of high dose.

        (iv) Two rounds of standard dose to one round of high dose.

    (c) Repeat part (b) using GEE with a working exchangeable structure and comment on the results (you need not re-express the parameters in terms of $\boldsymbol{\beta}$). Your answer may surprise you!

    (d) Repeat part (b) using a mixed effects model that allows random intercepts and random slopes and comment on the results (you need not re-express the parameters in terms of $\boldsymbol{\beta}$). The random effects are coded as "`|| id: t`" in Stata.

24. Conduct a simulation study in which you demonstrate the utility of splines in accounting for confounding. I suggest the following setup. Let $N = 100$ denote your sample size. Let $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 = X_1 + U$, where $U \sim \mathcal{U}(5, 10)$. Let $\epsilon \sim \mathcal{N}(0, 1)$ denote the error term, and suppose $Y = 0.1X_1 + f(X_2) + \epsilon$, where $f(x) = 0.005(X_2/1.2 - 4)^4$. Consider three linear regression models estimated by ordinary least squares: one including a linear term for X_1 (but not X_2), one including linear terms for both X_1 and X_2, and one including a linear term for X_1 and a (regular) cubic spline on X_2 with knots at 5, 8, and 11. Determine the average estimate of β_1 across 10,000 simulation replicates, along with the standard deviation of the estimates. Comment on your results. Attach your code as part of your response to this problem.

25. Consider the primary goal of the REACH study (`reach.csv`), which was to investigate whether REACH had an effect on mean A1c six months post-baseline relative to control. Now, think of this problem like a Bayesian, and consider the following four choices of prior distributions:

    (I) $\beta_0 \sim \mathcal{N}(\mu = 0, \sigma^2 = 9^2)$; $\beta_1 \sim \mathcal{N}(\mu = 0, \sigma^2 = 9^2)$; $\tau^2 \sim \text{InvGamma}(0.2, 0.4)$.

    (II) $\beta_0 \sim \mathcal{N}(\mu = 0, \sigma^2 = 9^2)$; $\beta_1 \sim \mathcal{N}(\mu = 0, \sigma^2 = 2^2)$; $\tau^2 \sim \text{InvGamma}(0.2, 0.4)$.

    (III) $\beta_0 \sim \mathcal{N}(\mu = 0, \sigma^2 = 2^2)$; $\beta_1 \sim \mathcal{N}(\mu = 0, \sigma^2 = 9^2)$; $\tau^2 \sim \text{InvGamma}(0.2, 0.4)$.

    (IV) $\beta_0 \sim \mathcal{N}(\mu = 0, \sigma^2 = 2^2)$; $\beta_1 \sim \mathcal{N}(\mu = 0, \sigma^2 = 2^2)$; $\tau^2 \sim \text{InvGamma}(0.2, 0.4)$.

    Note that β_1, which corresponds to the treatment indicator, is the parameter of interest.

    (a) Briefly comment on the degree to which you believe the posterior variance for β_1 will vary across the above combinations of prior distributions. In particular, are there any that you would clearly expect to be narrower or wider?

    (b) For each of the four choices of prior distributions, determine and report the corresponding posterior means, standard deviations, and 95% credible intervals for β_1. Please add the Bayes option `rseed(6312)` for reproducibility. To what degree do the results align with your expectations?

    (c) How do the posterior means and 95% credible intervals you determined in part (b) compare to the point estimate and 95% confidence interval from a frequentist analysis?

    (d) How would you expect the posterior mean and 95% credible interval for the treatment effect under a Bayesian analysis that adjusts for baseline A1c to compare to those of an unadjusted analysis that utilizes the same priors for other parameters? Considering prior (I) only, add baseline A1c as a covariate, with $\beta_2 \sim \mathcal{N}(\mu = 0, \sigma^2 = 9^2)$; please add the Bayes option `rseed(6312)` for reproducibility. Report the corresponding posterior mean and 95% credible interval, and compare to the unadjusted model.

26. Conduct a simulation study in which you illustrate that the Bonferroni correction is conservative if the analyses being compared are highly correlated. This problem is deliberately open-ended. But don't go wild – you can accomplish this with a very simple setup. The goal is not to *prove*, but to *illustrate*. Attach your code as part of your response to this optional problem, should you choose to do it. *Hint*: Estimates from adjusted and unadjusted regression models are often highly correlated.