STAT 9610: Midterm Exam

Name

Fall 2021

1 Instructions

Setup. Clone this repository and open `midterm-fall-2021.tex` in your LaTeX editor. Use this document as a starting point for your writeup, adding your solutions between `\begin{sol}` and `\end{sol}`. Add R code for problem i in `problem-i.R` (rather than in your LaTeX report), saving your figures and tables to the `figures-and-tables` folder for LaTeX import.

Resources. Consult the getting started guide if you need to brush up on R, LaTeX, or Git, the preparing reports guide for guidelines on presentation quality, the sample homework for an example of a completed homework repository, and this webpage for more detailed instructions on using GitHub and Gradescope to complete and submit homework.

Allowed materials. The allowed materials are as stated on the Syllabus:

> Students may consult all course materials, including course textbooks, for all assignments and assessments. Students may also consult the internet (e.g. Stack Overflow) for help with general programming tasks (e.g. how to add a dashed line to a plot). Students may not search the internet for help with specific questions or specific datasets on any homework or exam. In particular, students may not use solutions to problems that may be available online and/or from past iterations of the course.

Collaboration. Students must complete the exam individually.

Programming. The `tidyverse` paradigm for data manipulation (`dplyr`) and plotting (`ggplot2`) is required; points will be deducted for using base R.

Grading. Each sub-part of each problem will be worth 3 points: 0 points for no solution or completely wrong solution; 1 point for some progress; 2 points for a mostly correct solution; 3 points for a complete and correct solution modulo small flaws. The presentation quality of the solution for each problem (see the preparing reports guide) will be evaluated out of an additional 3 points.

Submission. This is a practice midterm, so please do not submit it. For the actual exam, compile your LaTeX report to PDF and commit your work. Then, push your work to GitHub. Finally, submit your GitHub repository to Gradescope.

Problem 1. The consequences of model bias.

To study the effect of a predictor x_{p-1} on a response y, we collect an observational dataset of n samples; for each sample we measure y, x_{p-1}, and $p-1$ possible confounders $x_0, x_1, \ldots, x_{p-2}$. We then postulate the linear model model

$$y = \beta_0 x_0 + \cdots + \beta_{p-2} x_{p-2} + \beta_{p-1} x_{p-1} + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \tag{1}$$

based on which we construct $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$, test $H_0 : \beta_{p-1} = 0$, and construct a confidence interval for β_{p-1} as in Unit 2. Unfortunately, we forgot about one confounder, x_p! It turns out that that x_{p-1} actually has no effect on y, and that the true distribution of the data is

$$y = \beta_0 x_0 + \cdots + \beta_{p-2} x_{p-2} + \beta_{p-1} x_{p-1} + \beta_p x_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad \text{where } \beta_{p-1} = 0. \tag{2}$$

In this problem, we will investigate the consequences of this model bias. As usual, we view the predictors as fixed.

(a) What is the distribution of the least squares coefficient estimate $\widehat{\beta}_{p-1}$—defined based on the postulated linear model (1)—under the true data-generating model (2)? What is the bias of $\widehat{\beta}_{p-1}$?

(b) What is the expectation of the variance estimate $\widehat{\sigma}^2$—defined based on the postulated linear model (1)—under the true data-generating model (2)?

(c) What is the Type-I error of the right-sided level-α t-test of $H_0 : \beta_{p-1} = 0$—constructed based on the postulated model (1)—under the true data-generating model (2)? [For the sake of this question, you may ignore the sampling variability in $\widehat{\sigma}^2$ (i.e. assume $\widehat{\sigma}^2$ is always equal to its expectation) and approximate $t_{n-p} \approx N(0,1)$.]

(d) How do the bias found in part (a) and the Type-I error found in part (c) vary with β_p? Discuss the intuition for these results. [To discuss the dependency of the Type-I error on β_p, you may restrict your attention to $\beta_p \to \infty$.]

(e) Carry out the following numerical simulation to assess bias and Type-I error. Set $n = 100$, $p = 20$, $\sigma = 1$, $(\beta_0, \ldots, \beta_{p-1}) = \mathbf{0}$, $\beta_p \in \{0, 0.5, 1, \ldots, 4.5, 5\}$, and $\alpha = 0.05$. Draw $(x_{i,0}, \ldots, x_{i,p-1}, x_{i,p}) \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}(\rho))$ for $i = 1, \ldots, n$, where $\boldsymbol{\Sigma}(\rho)$ is the AR(1) covariance matrix with autocorrelation parameter $\rho \in \{0.05, 0.2\}$, i.e. $\boldsymbol{\Sigma}(\rho)_{j_1, j_2} = \rho^{|j_1 - j_2|}$ for $j_1, j_2 \in \{1, \ldots, p+1\}$. For each pair (β_p, ρ), compute the bias of $\widehat{\beta}_{p-1}$ computed with respect to the postulated model (1) and the Type-I error of the corresponding t-test, via 1000 draws of $\boldsymbol{y}$ based on the model (2), while keeping the predictors fixed. Plot the simulated bias and Type-I error as a function of β_p for each ρ, overlaying the theoretical predictions from parts (a) and (c), respectively. Add a dashed horizontal line on the Type-I error plot at the nominal level α. Comment on the agreement between the simulation and theoretical predictions, the shapes of the resulting curves, and how these connect to the discussion in part (d).

Solution 1.

Problem 2. Case study: Determinants of COVID case-fatality rate.

The coronavirus pandemic has had a disparate impact on different communities across the United States. A key measure of this impact is the *case-fatality rate*, defined as the ratio of the number of deaths to the number of cases, expressed as a percentage. The goal of the present analysis is to study the relationship between the case-fatality rate and a variety of health and socioeconomic factors at the county level in the year 2020, before vaccines became widely available.

To this end, we are given `covid_data.tsv`, compiled from case and death tracking data from The New York Times and 41 county-level health and socioeconomic factors compiled by the County Health Rankings and Roadmaps. Descriptions of these 41 socioeconomic factors are given in Appendix A. The data contain 935 counties out of about 3000 total in the US, for which the health and socioeconomic factors were available.

(a) Run a linear regression of `case_fatality_rate` on the 41 given predictors. What fraction of the variation in the response is explained by the predictors? Print a table containing the features whose t-test p-values pass the multiplicity-adjusted threshold of $\alpha' = 0.05/41 \approx 0.0012$, for each feature displaying the coefficient estimate, standard error, and p-value.

(b) Create the residuals-versus-fitted-values and residuals-versus-leverage diagnostic plots. Are there any apparent concerns regarding the independence and homoskedasticity assumptions? Are there any apparent outliers?

(c) To further probe the independence assumption, visualize the distributions of the standardized residuals grouped by state. [Hint: Use a box plot, with states on the vertical axis.] Are any departures from independence apparent in this plot? To assess statistically whether `state` is associated with `case_fatality_rate`, run a heteroskedasticity-robust test to determine whether the model with an indicator for state fits significantly better than the model run in part (a). What do you conclude?

(d) The effect of the `state` variable can be accounted for using two different robust analyses: (1) based on the linear regression in part (a) but with Liang-Zeger standard errors, clustering by `state` and (2) based on the linear regression in part (a) but with `state` as an additional predictor and with Huber-White standard errors. For both of these methods, print tables containing the features whose t-test p-values pass the multiplicity-adjusted threshold of $\alpha' = 0.05/41 \approx 0.0012$, for each feature displaying the coefficient estimate, standard error, and p-value.

(e) Discuss the pros and cons of the two analyses done in part (d). In what situations would analysis (1) be more appropriate, and in what situations would analysis (2) be more appropriate? Which analysis leads to greater standard error inflation? To address the latter question, for each robust analysis produce a histogram (across features) of the factor by which the standard error exceeds that obtained from the analysis in part (a). On the whole, which analysis would you recommend for this problem?

Solution 2.

A Descriptions of features in COVID data

Below are the 41 features we used for analysis. Words written in parentheses represent variable names. Unless noted otherwise, all variables are continuous.
Health behaviors:

- *Tobacco Use*
  - Adult smoking (`smoke_perc`): Percentage of adults who are current smokers.
- *Diet and Exercise*
  - Adult obesity (`obesity_perc`): Percentage of the adult population (age 20 and older) reporting a body mass index (BMI) greater than or equal to 30 kg/m2.
  - Food environment index (`food_environment`): Index of factors that contribute to a healthy food environment, from 0 (worst) to 10 (best).
  - Physical inactivity (`inactive_perc`): Percentage of adults age 20 and over reporting no leisure-time physical activity.
  - Access to exercise opportunities (`physical_exercise_opportunities`): Percentage of population with adequate access to locations for physical activity.
  - Food insecurity (`Food_Insecure_perc`): Percentage of population who lack adequate access to food.
  - Limited access to healthy foods (`limited_healthy_access`): Percentage of population who are low-income and do not live close to a grocery store.
- *Alcohol and Drug Use*
  - Excessive Drinking (`drinking_perc`): Percentage of adults reporting binge or heavy drinking.
- *Sexual Activity*
  - Sexually transmitted infections (`stis`): Number of newly diagnosed chlamydia cases per 100,000 population.
  - Teen births (`teen_births`): Number of births per 1,000 female population ages 15-19.
  - Low Birth Weight Percentage (`low_birthweight_percentage`): Percentage of live births with low birthweight (<2,500 grams).

Clinical care:

- *Access to Care*
  - Uninsured (`uninsured`): Percentage of population under age 65 without health insurance.
  - Primary care physicians (`primarycare_ratio`): Ratio of population to primary care physicians.
  - Dentists (`dentist_ratio`): Ratio of population to dentists.
  - Mental health providers (`mentalhealth_ratio`): Ratio of population to mental health providers.
  - Other primary care providers (`otherproviders_ratio`): Ratio of population to primary care providers other than physicians.
- *Quality of Care*

  - Preventable hospital stays (`preventable_hospitalization`): Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees.
  - Mammography screening (`mammogram_perc`): Percentage of female Medicare enrollees ages 65-74 that received an annual mammography screening.
  - Flu vaccinations (`flu_vaccine_perc`): Percentage of fee-for-service (FFS) Medicare enrollees that had an annual flu vaccination.
  - Teen births (`teen_births`): Number of births per 1,000 female population ages 15-19.

Social and economic factors:

- *Education*
  - High school completion (`HS_completion`): Percentage of adults ages 25 and over with a high school diploma or equivalent.
  - Some college (`some_college`): Percentage of adults ages 25-44 with some post-secondary education.
  - Disconnected youth (`disconnected_youth`): Percentage of teens and young adults ages 16-19 who are neither working nor in school.
- *Employment*
  - Unemployment (`unemployment`): Percentage of population ages 16 and older who are unemployed but seeking work.
- *Income*
  - Children in poverty (`children_poverty_percent`): Percentage of people under age 18 in poverty.
  - Income inequality (`income_inequality`): Ratio of household income at the 80th percentile to income at the 20th percentile.
  - Median household income (`median_income`): The income where half of households in a county earn more and half of households earn less.
  - Children eligible for free or reduced price lunch (`children_freelunches`): Percentage of children enrolled in public schools that are eligible for free or reduced price lunch.
- *Family and Social Support*
  - Children in single-parent households (`single_parent_households`): Percentage of children that live in a household headed by a single parent.
  - Social associations (`social_associations`): Number of membership associations per 10,000 residents.
  - Residential segregation—Black/White (`segregation_black_white`): Index of dissimilarity where higher values indicate greater residential segregation between Black and White county residents.
  - Residential segregation—non-White/White (`segregation_nonwhite_white`): Index of dissimilarity where higher values indicate greater residential segregation between non-White and White county residents.
- *Community Safety*
  - Violent crime rate (`Violent_crime`) Number of reported violent crime offenses per 100,000 residents.

Physical environment:

- *Air and Water Quality*
  - Air pollution - particulate matter (`air_pollution`): Average daily density of fine particulate matter in micrograms per cubic meter (PM2.5).
  - Drinking water violations (`water_violations`): Indicator of the presence of health-related drinking water violations. 1 indicates the presence of a violation, 0 indicates no violation.
- *Housing and Transit*
  - Housing overcrowding (`housing_overcrowding`): Percentage of households with overcrowding,
  - Severe housing costs (`high_housing_costs`): Percentage of households with high housing costs
  - Driving alone to work (`driving_alone_perc`): Percentage of the workforce that drives alone to work.
  - Long commute—driving alone (`long_commute_perc`): Among workers who commute in their car alone, the percentage that commute more than 30 minutes.
  - Traffic volume (`traffic_volume`): Average traffic volume per meter of major roadways in the county.
  - Homeownership (`homeownership`): Percentage of occupied housing units that are owned.
  - Severe housing cost burden (`severe_ownership_cost`): Percentage of households that spend 50% or more of their household income on housing.