Stats 203V: Introduction to Regression Models and ANOVA Summer 2020

Final

Due date : August 15, 11:59 PM (PT)

Solutions should be uploaded to Gradescope before Saturday, August 15 at 11:59 p.m, PT. The deadline is firm: to avoid technical glitches, please start early.

There are 5 problems. The datasets `binge.txt` along with `survey203.csv`, `survey203.Rmd` and `pima.Rmd` are in the directory Final on `Canvas`.

In writing solutions to the data problems, `R Markdown` is highly recommended. Include only the pieces of output (and plots) needed for your answer. Readability of solutions is important. Poorly organized presentation and/or errors in additional plots/analyses may be penalized. For each question, as appropriate, explain clearly (i) your objectives, (ii) any hypotheses that you are testing, (iii) the statistical procedure, and (iv) the statistical support for your conclusions. An example (solution for HW2-Q2) is given in the files `sample-solution.Rmd` and `sample-solution.pdf`.

Please respect the honor code in completing this exam. You can use books, computers and the internet, but not other people.

1. (15 points) Some years back, the Centers for Disease Control issued a report on binge drinking that received national attention. The data set `binge.txt` contains data for 48 states (no data for South Dakota and Tennessee) on the age-adjusted prevalence of binge drinking (as a percentage of adults responding to a telephone survey).

   The CDC article stated "Overall, states with the highest age-adjusted prevalence of adult binge drinking were in the Midwest and New England, and included Alaska and Hawaii.".

   A question of interest might be whether the variation in binge drinking was associated with climate, in particular the depth of winters. The file `binge.txt` includes columns with the average winter temperature (degrees Celsius) and the state population (in millions). Investigate the relation between prevalence of binge drinking and these predictors. For example, can the regional variation be ascribed to differences in climate? Summarize your findings.

2. (5+5+5+5 points) The file `survey203.csv` contains the survey data on lecture attendance and Freedman practice problems studied (variables Lectures and Practice, with NA used for no-response cases) merged with the midterm Scores, for course Stats203 during Spring, 2019. The R Markdown file `survey203.Rmd` pre-processes the data to reorder the levels of the factors. You should add your answers in your copy of this file.

   (a) Create a variable that indicates whether the case (i.e. row) contains a missing value. Is non-response to the survey associated with the mid-term scores?

   Use `na.omit` to create a data frame with no missing values for the rest of this question.

   (b) Assess whether the factor `Practice` have a significant effect on the midterm scores, both via ANOVA and by pairwise comparison of means. Summarize your conclusions.

   (c) Now consider both factors and construct an ANOVA table with main effects and interactions and interpret your conclusions.

   (d) Use the function `as.numeric()` to "coerce" each of the factors to numeric variables `PracticeN`, `LecturesN`. Is there an association between these (coerced) variables? Fit a linear model with `Score` as response and the numeric variables plus interaction as

predictors, and summarize your conclusions, including a comparison with the results from (c).

3. (5+5+5 points)

   (a) Consider a one way ANOVA model

   $$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i.$$

   Suppose that the design is balanced, $n_i \equiv n_1$ for all i. Consider the design matrices corresponding to "treatment" and "sum" contrasts. In each of the two cases, is the intercept column orthogonal to the factor columns? What if the design is unbalanced? Explain.

   (b) Now consider the coefficient differences $\alpha_i - \alpha_j$ in the balanced one way ANOVA model. Do their estimates $\hat{\alpha}_i - \hat{\alpha}_j$ depend on whether treatment or sum contrasts are chosen? Explain.

   (c) In question 2(c), do the sums of squares in the ANOVA table change if the order in which `Practice` and `Lectures` appear is switched? Can you explain briefly in words (i.e. without detailed mathematical argument) why this might be?

4. (5+5+5+5+5+5 points) The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima` in `library(faraway)`. See also `pima.Rmd` in which some pre-processing is done: it creates a factor version of the test results. And as discussed in Chapter 1 of Faraway, the zero values for variables `diastolic`, `glucose, triceps, insulin` and `bmi` in fact seem to be missing values, so those are set to `NA`.

   (a) Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting?

   (b) Refit the model but now without the `insulin` and `triceps` predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question and report your conclusion.

   [Hint: use `na.omit()` to create a data frame with no missing values.]

   (c) Use AIC via the function `step()`to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?

   (d) Create a variable that indicates whether the case contains a missing value. Use this variable as a predictor of the test result. Is missingness associated with the test result? Refit the selected model from (c), but now using as much of the data as is feasible. Explain why it is appropriate to do this.

   (e) Using the last fitted model of the previous question, what is the ratio of the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this ratio.

   (f) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

5. (10+10 points) Suppose that we wish to find the weights $\beta_i (i = 1, 2, \ldots, k)$ of k objects. One method is to weigh each object r times and take the average; this requires a total of kr weighings, and the variance of each average is σ^2/r (σ^2 being the variance of the weighing error). Another method is to weigh the objects in combinations; some of the objects are distributed between the two pans and weights are placed in one pan to achieve equilibrium. The regression model for such a scheme is

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where $x_i = 0, 1$, or -1 according as the i-th object is not used, placed in the left pan or in the right pan, ε is the weighing error, and Y is the weight required for equilibrium (Y is regarded as negative if placed in the left pan). We assume that weighing errors for different weighing operations are uncorrelated with mean zero and constant variance σ^2. After n such weighing operations we can find the least squares estimates $\hat{\beta}_i$ of the weights.

(a) Show that the estimates of the weights have maximum precision (i.e.,minimum variance) when each entry in the design matrix X is ± 1 and the columns of X are mutually orthogonal.

(b) If the objects are weighed individually, show that at least kn weighings are required to achieve the same or more precision than that given by the optimal design with n weighings. (Here more precision implies more precision for each of the k estimates. So for example, you are not allowed to use all the weighing operations to weigh only one object and thus improving the precision for the estimate of its weight while ignoring the others.)

Remark To make the above problem more mathematically precise, let us write the question in the following way. Let $\mathcal{D}$ be the design (here design refers to the design matrix) and $\hat{\beta}_{\mathcal{D}}$ be the estimate according to this design. The design $\mathcal{D}$ is said to have optimal precision of size n if the design $\mathcal{D}$ needs exactly n weighings and $\mathrm{Var}(\hat{\beta}_{\mathcal{D}}) \preceq \mathrm{Var}(\hat{\beta}_{\mathcal{D}'})$, for any other design $\mathcal{D}'$ which needs exactly n weighings. Here $\preceq$ refers to co-ordinatewise inequality.

The first problem asks you to show that any optimal design $\mathcal{D}$ of size n should have each entries of the design matrix to be ± 1 and mutually orthogonal columns.

The second question asks you to show that if $\mathcal{D}'$ is some design which only allows the objects only to be weighted individually and $\mathrm{Var}(\hat{\beta}_{\mathcal{D}}) \succeq \mathrm{Var}(\hat{\beta}_{\mathcal{D}'})$ for some optimal design $\mathcal{D}$ of size n, then the design $\mathcal{D}'$ needs at least kn weighings.