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

Homework 4

Due date : August 07, 11:59 PM (PT)

For full credit, solutions (.pdf file, either typed or handwritten and scanned) must be uploaded to `Gradescope` by **11:59 pm PT on Friday August 07**. Any changes to the Homework will be posted on `Canvas`.

Collaboration on homework problems is fine, but your write up should be your own and the write up should mention the names of your collaborators.

From now on, we request that solutions to questions using `R` be prepared using `R Markdown`, e.g. within `R Studio`. Please convert your `.Rmd` file to a `.pdf` file before uploading.

1. (20 points) Faraway, Exercise 6.1 (a)-(d).

2. (20 points) [Simulation example to illustrate collinearity] Consider a linear model $Y_i = 1 + \sum_{j=1}^{4} \beta_j X_i^{(j)} + \epsilon_i$ for $i = 1, \ldots, 100$ with $\beta_1 = \beta_2 = 1$ and $\beta_3 = \beta_4 = 0$. Suppose that ϵ_i are i.i.d $N(0, 1)$ and that X_i are i.i.d. $N_4(0, \Sigma)$ independently of ϵ. Draw a sample in two cases:

   (a) $\Sigma = I$,

   (b) $\sigma_{ii} = 1$ for $i = 1, \ldots, 4$ and $\sigma_{13} = 0.95, \sigma_{24} = -0.95$.

   In each case, fit the regression model for Y on all $X^{(j)}$ with intercept, and report the standard errors of the coefficient estimates. Comment on why they are different, and compare the corresponding variance inflation factors.

3. (10 points) [Variable selection methods on `prostate` data, available in `library(faraway)`] Faraway, Exercise 10.1.

4. (8+7+10+5 points) [Ridge regression and lasso on the diabetes data.]

   The file `Homework 4/diascale.txt` on `Canvas` contains the "diabetes data". Ten baseline variables, age, sex, body mass index, average blood pressure and six blood serum measurements were obtained for each of $n = 442$ diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. In the data file, the response Y is already centered, and the ten predictors are centered and scaled to have mean zero and sum of squares equal to 1.

   (a) By modifying the commands in `Homework 4/ridge.Rmd` (or otherwise), produce analogs of the coefficient trace plots for ridge regression and the lasso and the table of least squares, ridge, and lasso coefficient estimates as in the class notes of Lecture 14 (No need to create a training subset of the data this time). What happens to the sign of the coefficient of S3 in the trace plots?

   (b) Compute the variance inflation factors for the matrix of predictors.

   (c) In the class notes, we saw that the ridge regression estimator has the form $\hat{\beta}_\lambda = (X'X + \lambda I)^{-1} X'y$. Using the singular value decomposition $X = UDV'$, derive a formula relating $\hat{\beta}_\lambda$ to the least squares estimator $\hat{\beta}_{LS}$.

(d) In part (a), notice that some of the ridge estimate coefficients are larger than the corresponding least squares estimates. Does this contradict the shrinkage property discussed in class? Explain briefly.

5. (10 points) [One way ANOVA, `infmort` data] Faraway, Exercise 15.4.

6. (10 points) Using block matrix inversion formulas prove the following formula which we stated in Lecture 12,

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the multiple R^2 in fitting $X^{(j)}$ with the other $X^{(k)}$'s.