University of California, Berkeley Lecture: Peng Ding

STAT230A

Due at 3:00 p.m. on March 17, 2017 (@ Yuting Ye's mailbox at 367 Evans Hall)

Problem Set 4

Homework Policy:
You're encouraged to discuss with your fellows but need to finish your own write-up. Delay is not allowed. The write-up is encouraged to be a pdf version, either by Latex, Markdown or other formats. A tidy and clean write-up deserves a higher grade when there are mistakes. When there are real-data problems, R is required and only allowed. R Markdown and R Sweave (used in Rstudio) are recommended to report the results of the real-data problems.

Problem 1 (10 pts, ★☆☆☆☆)

Consider the linear regression model for which $\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n] = \mathbf{X}_n\beta$ and $Cov[\mathbf{Y}_n|X_n] = \sigma^2\mathbf{I}_n$, where $\mathbf{Y}_n \in \mathbb{R}^n$, $\mathbf{X}_n \in \mathbb{R}^{n\times p}$. Please derive the closed-form solutions for the following optimization problems.

1. Ordinary Least Squares (OLS).
$$\min_\beta ||\mathbf{Y}_n - \mathbf{X}_n\beta||_2^2 \tag{1}$$
2. Ridge.
$$\min_\beta ||\mathbf{Y}_n - \mathbf{X}_n\beta||_2^2 + \lambda||\beta||_2^2 \tag{2}$$
3. LASSO under the orthonormal covariates, i.e., $\mathbf{X}_n^T\mathbf{X}_n = \mathbf{I}_n$.
$$\min_\beta ||\mathbf{Y}_n - \mathbf{X}_n\beta||_2^2 + \lambda||\beta||_1 \tag{3}$$

Problem 2 (20 pts, ★☆☆☆☆)

For Table 1, fit the regression model of y to x_2, x_7, x_8 and answer the following questions:

a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?

b. Construct and interpret a plot of the residuals versus the predicted response.

c. Construct plots of the residuals versus each of the regressor variables. Do these plots imply that the regressor is correctly specified?

d. Construct the partial regression plots for this model. Compare the plots with the plots of residuals versus regressors from part c above. Discuss the type of information provided by these plots.

e. Compute the studentized residuals and the R-student residuals for this model. What information is conveyed by these scaled residuals?

Problem 3 (15 pts, ★☆☆☆☆)

This exercise is designed to show how to check for heteroscedasticity of residuals once you build the linear regression model. Use the cars dataset, which is already stored in R and you can directly type in *cars* to get access to it.

a. Fit *dist* to *speed* and store the result object as *lm_model*. Plot the plots of residuals v.s. fitted values and square root of studentized residuals v.s. fitted values by the command *plot(lm_model)*. Does heteroscedasticity exist?

b. Use Breush Pagan Test to check for heteroscedasticity. What's your conclusion?

c. If there exists heteroscedasticity, use Box-cox transformation rectification. Check the results after the transformation.

Problem 4 (30 pts, ★★☆☆☆)

This problem is designed to see the effect of heteroscedasticity on the linear regression model. The linear model can be written as

$$y = X\beta + \epsilon$$

where $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon\epsilon') = \Phi$, a positive definite matrix. Under this specification, the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$ is best linear unbaised with

$$var(\hat{\beta}) = (X'X)^{-1}X'\Phi X(X'X)^{-1} \tag{4}$$

An appropriate estimation of Eq. (4) is important to the t test for each entry of β. If the errors are homoscedastic, that is $\Phi = \sigma^2 I$, Eq. (4) simplifies to

$$var(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Defining the resiguals $e_i = y_i - x_i\hat{\beta}$, where x_i is the i_{th} row of X, we can estimate the OLS covariance matrix (OLSCM) of estimates as

$$OLSCM = \frac{\sum e_i^2}{N-K}(X'X)^{-1} \tag{5}$$

where N is the sample size and K is thenumber of elements in β. The OLSCM is appropriate for hypothesis testing and computing confidence intervals when the standard assumptions of the regression model, including homoscedasticity, hold. If there is heterocedasticity, four types of heteroscedasticity consistent covariance matrix (HCCM), referred as White, Eicker, or Huber estimator, are used.

$$\begin{aligned}
HC0 &= (X'X)^{-1}X' \cdot \text{diag}[e_i^2] \cdot X(X'X)^{-1} & (6)\\
HC1 &= \frac{N}{N-K}(X'X)^{-1}X' \cdot \text{diag}[e_i^2] \cdot X(X'X)^{-1} & (7)\\
HC2 &= (X'X)^{-1}X' \cdot \text{diag}[\frac{e_i^2}{1-h_{ii}}] \cdot X(X'X)^{-1} & (8)\\
HC3 &= (X'X)^{-1}X' \cdot \text{diag}[\frac{e_i^2}{(1-h_{ii})^2}] \cdot X(X'X)^{-1} & (9)
\end{aligned}$$

Team	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Washington	10	2113	1985	38.9	64.7	+4	868	59.7	2205	1917
Minnesota	11	2003	2855	38.8	61.3	+3	615	55.0	2096	1575
New England	11	2957	1737	40.1	60.0	+14	914	65.6	1847	2175
Oakland	13	2285	2905	41.6	45.3	−4	957	61.4	1903	2476
Pittsburgh	10	2971	1666	39.2	53.8	+15	836	66.1	1457	1866
Baltimore	11	2309	2927	39.7	74.1	+8	786	61.0	1848	2339
Los Angeles	10	2528	2341	38.1	65.4	+12	754	66.1	1564	2092
Dallas	11	2147	2737	37.0	78.3	−1	761	58.0	1821	1909
Atlanta	4	1689	1414	42.1	47.6	−3	714	57.0	2577	2001
Buffalo	2	2566	1838	42.3	54.2	−1	797	58.9	2476	2254
Chicago	7	2363	1480	37.3	48.0	+19	984	67.5	1984	2217
Cincinnati	10	2109	2191	39.5	51.9	+6	700	57.2	1917	1758
Cleveland	9	2295	2229	37.4	53.6	−5	1037	58.8	1761	2032
Denver	9	1932	2204	35.1	71.4	+3	986	58.6	1709	2025
Detroit	6	2213	2140	38.8	58.3	+6	819	59.2	1901	1686
Green Bay	5	1722	1730	36.6	52.6	−19	791	54.4	2288	1835
Houston	5	1498	2072	35.3	59.3	−5	776	49.6	2072	1914
Kansas City	5	1873	2929	41.1	55.3	+10	789	54.3	2861	2496
Miami	6	2118	2268	38.2	69.6	+6	582	58.7	2411	2670
New Orleans	4	1775	1983	39.3	78.3	+7	901	51.7	2289	2202
New York Giants	3	1904	1792	39.7	38.1	−9	734	61.9	2203	1988
New York Jets	3	1929	1606	39.7	68.8	−21	627	52.7	2592	2324
Philadelphia	4	2080	1492	35.5	68.8	−8	722	57.8	2053	2550
St. Louis	10	2301	2835	35.3	74.1	+2	683	59.7	1979	2110
San Diego	6	2040	2416	38.7	50.0	0	576	54.9	2048	2628
San Francisco	8	2447	1638	39.9	57.1	−8	848	65.3	1786	1776
Seattle	2	1416	2649	37.4	56.3	−22	684	43.8	2876	2524
Tampa Bay	0	1503	1503	39.3	47.0	−9	875	53.5	2560	2241

y: Games won (per 14-game season)
x_1: Rushing yards (season)
x_2: Passing yards (season)
x_3: Punting average (yards/punt)
x_4: Field goal percentage (FGs made/FGs attempted 2season)
x_5: Turnover differential (turnovers acquired–turnovers lost)
x_6: Penalty yards (season)
x_7: Percent rushing (rushing plays/total plays)
x_8: Opponents' rushing yards (season)
x_9: Opponents' passing yards (season)

Figure 1: National Football League 1976 Team Performance

where $h_{ii} = x_i(X'X)^{-1}x_i'$ is the leverage. Next, we're going to explore the empirical size and the power of Eq. (5)(6)(7)(8)(9) for the t test of each entry in β.

a. **Data Generation**. Simulate the model based on the model

$$y_i = 1 + 1 \cdot x_{1i} + 1 \cdot x_{2i} + 1 \cdot x_{3i} + 0 \cdot x_{4i} + \epsilon_i \tag{10}$$

$x_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})$ have a variety of distributions for diffent i's. Specifically,

$$x_{1i} \sim N(0, \tilde{\sigma}^2)$$

where $\tilde{\sigma}^2 = 1, 2, 3$;

$$x_{2i} \sim Unif[-b, b]$$

where $b = 1, 3, 5$;

$$x_{3i} \sim \chi^2_{df}$$

where $df = 1, 2, 3$;

$$x_{4i} \in \{0, 1\}$$

where $\mathbb{P}(x_{4i} = 1) = 0.2, 0.5, 0.8$. There are 81 equally likely combinations of the independent variables (x's). To sample one $x_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})$, we first uniformly sample an integer r.v. from $\{1, 2, \ldots, 81\}$, and use it to select the corresponding combination of distributions to sample x_i. As for the error term, we use

$$\begin{aligned}
\epsilon_i &= \epsilon_i^* \text{ for the homoscedastic case} \\
\epsilon_i &= \sqrt{x_{i3} + 1.6} \cdot \epsilon_i^* \text{ for the first type of heteroscedastic case} \\
\epsilon_i &= \sqrt{x_{i3}}\sqrt{x_{i4} + 2.5} \cdot \epsilon_i^* \text{ for the second type of heteroscedastic case}
\end{aligned}$$

where $\epsilon_i^* \overset{i.i.d}{\sim} \chi^2_5$. Generate $100,000$ observations by Eq. (10) for the homoscedastic case and the two types of heteroscedastic cases respectively. Regard the $100,000$ observations as the population, and for $N = 25, 50, 100, 250, 500, 1000$, uniformaly sample N points from the population without replacement for 1000 replications.

b. **Homoscedastic Case**. Use the homoscedastic set of data for the following experiment.

- To evaluate empirical size, the null hypothesis is $H_0 : \beta_k = \beta_k^*$, where β_k^* is the population value determined by a regression using all $100,000$ observations. For each of $\beta_1, \ldots, \beta_4$, draw curves for OLSCM, HC0, HC1, HC2, HC3 at a given sample size ($N = 25, 50, 100, 250, 500, 1000$) in one plot, for the proportion of times that the correct H_0 is rejected (significance level is 0.05) over the $1,000$ replications, which is called the empirical size.
- The epmirical power is the proportion of times the false hypothesis $H_0 : \beta_k = 0$ is rejected (significance level is 0.05) over $1,000$ replications. Draw four plots for empirical power similarly to the four plots for empirical size.

Which covariance estimator is favarable? And what can you conclude from these plots?

c. **Heteroscedastic Case**. Using the first type of heteroscedastic set of data, draw the four plots for empirical size and four plots for empirical power similarly to part b above. Which covariance estimator is favarable in this setup. And what can you conclude?

d. **Screening for Heteroscedasticity**. We propose a new procedure to tackel the heteroscedasticity.

Step 1 Use Breusch and Pagan (BP) test to determine if there is heteroscedasticity.

Step 2 If there is no heteroscedasticity, apply OLSCM, otherwise apply one variant of HC's.

Now, we have 9 methods, i.e., standard OLSCM test, HCm test regardless of the results of Breusch and Pagan test, $m = 0, 1, 2, 3$, and HCm test with BP test, $m = 0, 1, 2, 3$. Using the second type of heteroscedastic set of data, only draw the plots for empirical size. Is the BP procedure better than others? What can you conclude?

(**Hint:** See J. Long & L. Ervin 2000 for more details.)

Problem 5 (25 pts, ★★★☆☆)

This following example is designed to show the power of vectorization of R. The goal is to compute the likelihood for an overdispersed binomial random variable with the following probability mass function (pmf):

$$\mathbb{P}(Y = y) = \frac{f(y; n, p, \phi)}{\sum_{k=0}^{n} f(k; n, p, \phi)}$$

$$f(k; n, p, \phi) = \binom{n}{k} \frac{k^k (n-k)^{n-k}}{n^n} \left(\frac{n^n}{k^k (n-k)^{n-k}}\right)^{\phi} p^{k\phi} (1-p)^{(n-k)\phi}$$

where the denominator serves as a normalizing constant to ensure this is a valid probability mass function. Your job is to write code to evaluate the denominator of $\mathbb{P}(Y = y)$. You may need to evaluate $\mathbb{P}(Y = y)$ many many times, so efficient calculation of the denominator is important. For our purposes here you can take $p = 0.3$ and $\phi = 0.5$ when you need to actualy run your function.

1. First, write code to evaluate the denominator using *apply()*/*lapply()*/*sapply()*. Make sure to calcualte all the terms in $f(k; n, p, \phi)$ on the log scale to avoid numerical issues, before exponentiating and summing. Describe briefly what happens if you don't do the calculation on the log scale.
(**Hint:** *?Special* in R will tell you about a number of useful functions. Also, recall that $0^0 = 1$.)

2. Now wirte code to do the calculation in a fully vectorized fashion with no loops or *apply()* functions. Using the timing functino *benchmark()* in the *rbenchmark* package, compare the relative timing (a) and (b) for $n = 100, 500, 1000, 2000$. Note that for *benchmark()*, you need multiple replications (100 or 1000) in order to obtain a robust timing.

3. Please evaluate your The credit is given based on whether your code is as fast as my solution. When doing 100 replications for *benchmark()* with $n = 2000$, I got about 0.049s elapsed time, which was 20 times faster than the result of (a). You are supposed to get at least 15 times speeding up.