University of California, Berkeley Lecture: Peng Ding

STAT230A

Due at 3:00 p.m. on Feb 10, 2017 (@ Yuting Ye's mailbox at 367 Evans Hall)

Problem Set 2

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 (20 pts, ★★☆☆☆)

This problem is designed to show that if $Y \sim N(\mu, V)$, A is a symmetric matrix, if follows that

$$var(Y^T AY) = 2tr(AVAV) + 4\mu^T AVA\mu$$

In order to complete the proof, follow the below steps:

(a) If $X \sim N(0, 1)$, it follows that

$$\mathbb{E}X^k = \begin{cases} 0 & k \textit{ is odd} \\ (k-1)!! & k \textit{ is even} \end{cases}$$

(b) If $Z \sim N(\mathbf{0}, \mathbf{I}_p)$, $b \in \mathbb{R}^p$, $C \in \mathbb{R}^{p\times p}$ and C is symmetric, it follows that

$$\mathbb{E}(b^T Z + Z^T CZ)^2 = (\sum_i C_{ii})^2 + 2\sum_{i,j} C_{ij}^2 + \sum_i b_i^2$$

and thus

$$var(b^T + Z^T CZ) = 2tr(C^T C) + b^T b$$

(c) Show that

$$var(Y^T AY) = 2tr(AVAV) + 4\mu^T AVA\mu$$

(**hint:** $Y = \mu + V^{\frac{1}{2}} Z$)

Problem 2 (10 pts, ★☆☆☆☆)

Some practice with matrix manipulations.

(a) Consider a rectangular matrix $X \in \mathbb{R}^{n\times p}$ and $n > p$. Show that the right singular vectors of X are the eigenvectors of the matrix $X^T X$ and the the eigenvalues of $X^T X$ are the squares of the singular values of X. Also show that $X^T X$ is positive semi-definitive (which is good because $X^T X$ is essentially an empirical covariance matrix, up to scaling and shifting).

(b) Suppose $X \in \mathbb{R}^{n\times n}$ and assume you have already computed the eigendecomposition of X. How can you compute the eigenvalues of $Z = X + cI$ in $O(n)$ arithmetic calculations (including any additions or multiplications), where c is a scalar and I is the identity matrix.

Problem 3 (20 pts, ★★☆☆☆)

Suppose $\mathbf{X} \in \mathbb{R}^{n\times p}$, $p < n$ and $\mathbf{X}^T\mathbf{X}$ is invertible. Under the linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where $\beta \in \mathbb{R}^p$ and $\epsilon \sim N(0, \sigma^2 I_n)$. Let $\hat{\beta}$ be the Least Square solution. Answer the following questions

(a) Give the explicit form of $\hat{\beta}$. Show that $\hat{\beta}$ is a multivariate normal and

$$\mathbb{E}(\hat{\beta}) = \beta$$

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

(b) Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ and $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. Show that

$$\hat{\mathbf{Y}} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{H})$$

and

$$\hat{\mathbf{e}} \sim N(0, \sigma^2(\mathbf{I} - \mathbf{H}))$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the hat matrix.

(c) Show that $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are idempotent, that is

$$\mathbf{H}^2 = \mathbf{H}$$

and

$$(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - \mathbf{H}$$

(d) Show that the hat matrix $\mathbf{H}$ is a projection matrix which projects any vector in $\mathbb{R}^n$ to $C(\mathbf{X})$, where $C(\mathbf{X}) := \{\mathbf{X}\theta : \theta \in \mathbb{R}^p\}$. Then show that $\hat{\mathbf{Y}}$ and $\hat{\mathbf{e}}$ are independent.

(e) Give a condition such that the following equations hold, and prove them.

- $\sum_{i=1}^n \hat{e}_i = 0$
- $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$
- The sample covariance of $\hat{\mathbf{e}}$ with each precditor $\mathbf{X}_j$ and $\hat{\mathbf{Y}}$ is zero. Here the sample covariance bewteen $\mathbf{S} \in \mathbb{R}^n$ and $\mathbf{T} \in \mathbb{R}^n$ is defined as $\frac{1}{n}\sum_{i=1}^n (S_i - \bar{\mathbf{S}})(T_i - \bar{\mathbf{T}})$.

Problem 4 (20 pts, ★★☆☆☆)

Consider a regression version of the two-sample problem in which:

$$Y_i = \begin{cases} \beta_1 + \beta_2 x_i + \epsilon_i & i = 1, \dots, n_1 \\ \beta_3 + \beta_4 x_i + \epsilon_i & i = n_1 + 1, \dots, n_1 + n_2 = n \end{cases}$$

with $\epsilon_i, \dots, \epsilon_n$ i.i.d from $N(0, \sigma^2)$. Derive a $1 - \alpha$ confidence interval for $\beta_4 - \beta_2$, the difference between the two regression slopes.

Problem 5 (10 pts, ★☆☆☆☆)

Load the data *family.rda*. We will be using weight and BMI to predict height (BMI is body mass index, commonly used to measuer obesity).

(a) Using matrix algebra instead of existing R functions, create **betahat** and **residuals**, the vectors of linear model coefficients and residuals, respectively. Include an intercept term, so **betahat** should be a vector of length 3.

(b) Plot height and weight, as well as the regression line. Make the size of the points proportional to BMI (seems appropriate). Comment on the pattern you observed in the sizes of the points and how this related to the sign of the coefficient for BMI.

(c) Make a 3D plot and plot the regression plane through it (x = weight, y = BMI, z = height). You'll need the *rgl* package. Use the functions *plot3d()* and *planes3d()*.

Problem 6 (20 pts, ★★☆☆☆)

In the age of big data, data accumulates much faster than we can manipulate. In some occassions, the data is too redundant, that is, we might have more data than we want. For example, we want to do a linear regression on n samples, each with p predictors. If p is small but n is large, say $p = 10$,$n = 10^{12}$, a standard linear regression using all these data can be a waste of computation resources. A natural idea is to sampling r samples out of n, and hopefully these r subsamples can be representatives of the whole data.

To obtain the subsample, we use **Weighted Leveraging**

- **Step 1. Taking a random subsample of size r from the data.** Constructing sampling probability $\pi = \{\pi_1, \ldots, \pi_n\}$ for the data points. Draw a random subsample (with replacement) of size $r \ll n$, denoted as (X^*, y^*), from the full sample according to the probability π. Record the corresponding sampling probability matrix $\Phi^* = diag\{\pi_k^*\}$.
- **Step 2. Solving a weighted least squares (WLS) problem using the subsample.** Obtain lest squares estimate using the subsample. Estimate β via solving a WLS problem on the subsample to get estimate $\tilde{\beta}$, i.e., solve

$$arg\min_{\beta} ||(\Phi^*)^{-1/2}y^* - (\Phi^*)^{-1/2}X^*\beta||2$$

This problem is going to discuss how the hat matrix is used for measuring the levaraing values of the samples. You're supposed to play around a toy example to see how the hat matrix plays the role in sampling.

1. Data generation: Let $n = 500$, $r = 10$, $x_i \overset{i.i.d}{\sim} t(6),\ i = 1, \ldots, n$, $\epsilon \overset{i.i.d}{\sim} N(0, 1),\ i = 1, \ldots, n$, where $t(6)$ is the t distribution with degrees of freedom equal to 6. Let $y_i = -x_i + \epsilon_i,\ i = 1, \ldots, n$.
2. Use the built-in R function *lm()* to run the linear regression on the whole data. Report the coefficients (including the intercept) and Plot y against x as well as the regression line.

3. Apply the **Weighted Leveraging** with $\pi_i \equiv \frac{1}{n},\ i = 1, \ldots, n$. Repeat this step for 1000 times and report the average coefficients and the MSE.

4. Compute the hat matrix H (You need to consider the intercept). Plot y against x again, but color the top r points with the highest H_{ii} as red while coloring other points as black.

5. Apply the **Weighted Leveraging** with $\pi_i = \frac{H_{ii}}{\sum_{j=1} H_{jj}}$. Repeat this step for 1000 times and report the average coefficients and the MSE. (**Hint:** *lm()* has a *weights* option for **WLS**.)

6. Compare the above results, what conclusion can you draw?

(**Reference:** Leveraging for big data regression, P. Ma and X. Sun, 2014)