MSDS 596 Regression & Time Series Fall 2018

Homework 1

Due: Tues 09/18/18 @ 6:40pm rutgers.instructure.com/courses/17597

Problem 1. *Proposition 6.2 of Review of Matrix Algebra (I).* Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a real symmetric and idempotent matrix, and $\{\lambda_1, ..., \lambda_n\}$ its eigenvalues. Prove that:

(a) λ_i is either 0 or 1 for all $1 \le i \le n$;

(b) $\mathrm{tr}(\boldsymbol{A}) = \mathrm{rank}(\boldsymbol{A})$;

(c) $\mathrm{rank}(\boldsymbol{A}) + \mathrm{rank}(\boldsymbol{I} - \boldsymbol{A}) = n$.

Problem 2. Prove that $\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ is a projection matrix, and identify the vector to which it projects an arbitrary $\boldsymbol{y}$. In your own words, what does the projection matrix do?

Problem 3. Suppose we have a set of observations $(x_1, y_1), \ldots, (x_n, y_n)$ coming from the simple linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad 1 \le i \le n. \tag{1}$$

Least square method finds the minimizer of

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

(a) Find the expressions of

$$\frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_0}, \quad \text{and} \quad \frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_1}.$$

Define

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

(b) Show that the least square estimates are $\hat{\beta}_1 = S_{xy}/S_{xx}$, and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Be sure to check the conditions under which the estimates are indeed minimizers.

(c) Show that $y_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x}) + (y_i - \hat{y}_i)$.

(d) Show that $S_{yy} = \sum_{i=1}^{n}\hat{\beta}_1^2(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

(e) Show that R^2 equals to the sample correlation between $\boldsymbol{y}$ and $\boldsymbol{x}$, i.e.

$$1 - \frac{\mathrm{RSS}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

Problem 4. *Maximum likelihood estimate.* Consider the linear model (1). Assume x_i's are fixed, and ϵ_i's are independent and identically distributed as $N(0, \sigma^2)$. The *likelihood function* is defined as

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{N}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right]\right\}.$$

The estimate $(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$ which maximizes the likelihood function $L(\boldsymbol{\beta}, \sigma^2)$ is called *maximum likelihood estimate.*

(a) Show that $\tilde{\boldsymbol{\beta}}$ is the same as the least square estimate $\hat{\boldsymbol{\beta}}$.

(b) Find an expression for $\tilde{\sigma}^2$.

Problem 5. The dataset `teengamb` concerns a study of teenage gambling in Britain. Install the `R` package `faraway` and then use the `data()` function to load the dataset `teengamb`, which is stored in the data frame `teengamb`.

```
> install.packages("faraway")   # installed the faraway package if needed
> library(faraway)
> data(teengamb)
> head(teengamb)
  sex status income verbal gamble
1   1     51   2.00      8    0.0
2   1     28   2.50      8    0.0
3   1     37   2.00      6    0.0
4   1     28   7.00      4    7.3
5   1     65   2.00      8   19.6
6   1     61   3.47      6    0.1
```

Fit a regression model with the expenditure on gambling (`gamble`) as the response, and the sex, status, income and verbal scores as predictors.

(a) Report the coefficients and variance estimates.

(b) What percentage of variation in the response is explained by these predictors. In other words, what is the value of R^2?

(c) Which observations has the largest and smallest residual? Give the case numbers. (Suppose `out` is the fitted model object, use `out$residuals` to extract the vector consisting of all the residuals.)

(d) Compute the mean and median of the residuals.

(e) Compute the sample correlation of the residuals with the fitted values.

(f) Compute the sample correlation of the residuals with the income.

(g) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for male compared to a female? Be sure to find out how male vs female are respectively encoded in this dataset.

Problem 6. The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Load the dataset the same way as you did the previous problem.

(a) Fit a model with `lpsa` as the response and `lcavol` as the predictor. Display the scatterplot and the regression line. Report the residual standard error and the R^2.

(b) Now add `lweight`, `svi`, `lbph`, `age`, `lcp`, `pgg45` and `gleason` to the model one at a time. For each model record the residual standard error and R^2. Plot the trends in these two statistics and comment on them.