The University of British Columbian —Okanagan

DATA 310

Assignment 3

Due at 11:59 pm on Nov 29, 2024. Submit it on Canvas.
Use package `MPV` in R to find all data in the text book.

```
install.packages("MPV") # if you did not install before, otherwise skip this
library(MPV)
help(package="MPV") # opens the MPV help file
```

Scroll down the help file to find the name of the data frame specified in the problems at the end of each chapter. For example, Problem 2.10 in the textbook uses data frame `p2.10`. The data in a table in a textbook example is also included. For example, data in Example 3.1 used Table B.1. This is called `table.b1` in the R package.

Question 1

Consider the linear regression model in the form $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$, where $\mathrm{E}[\underset{\sim}{\epsilon}] = 0$ and $\mathrm{E}[\underset{\sim}{\epsilon}\underset{\sim}{\epsilon}^{\mathbf{T}}] = \sigma^2 I$, and $\underset{\sim}{\epsilon}$ is normally distributed. The least-squares estimator for $\underset{\sim}{\beta}$ is

$$\underset{\sim}{\widehat{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\underset{\sim}{\mathbf{y}}$$

This estimator requires the matrix $(\mathbf{X}^{\mathbf{T}}\mathbf{X})$ to be invertible.

(a) (3 points) Suppose $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ is a singular 2×2 matrix. That is, $(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$ does not exist and the determinant is

$$\begin{aligned} \det(\mathbf{X}^{\mathbf{T}}\mathbf{X}) &= \det\left(\begin{bmatrix} a & c \\ b & d \end{bmatrix}\right) \\ &= ad - bc = 0 \end{aligned}$$

Note that a and d are nonnegative. Show $\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{I}$ is invertible, that is $\det(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{I}) > 0$, when $\lambda \neq 0$ and $\mathbf{I}$ is an identity matrix.

(b) (5 points) In many situations, some of the explanatory variables are multicollinear, which means the variables are highly correlated and the matrix $(\mathbf{X}^{\mathbf{T}}\mathbf{X})$ is not invertible (or close, i.e. with a very very small determinant). In that case there is a problem to estimate $\underset{\sim}{\beta}$. Ridge regression is one of solutions for this kind problem.

Instead of minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_k x_k)^2$$

i.e.

$$(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\beta})^{\mathbf{T}}(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\beta})$$

ridge regression minimizes

$$(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\beta})^{\mathbf{T}}(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\beta}) + \lambda\underset{\sim}{\beta}^{\mathbf{T}}\underset{\sim}{\beta}$$

where λ is a known positive constant.
Expand the above expression algebraically, and use what you learned about minimizing $(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\beta})^{\mathbf{T}}(\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\beta})$ to obtain the $\underset{\sim}{\beta}$ estimator for ridge regression. The estimator is in terms of $\mathbf{X}$, $\underset{\sim}{\mathbf{y}}$ and λ.

(c) (2 points) Is the estimator of $\underset{\sim}{\beta}$ an unbiased estimator?

Question 2

When we use several variables to build a multiple regression model, it may be necessary to perform all kind of tests, for example, $H_0 : \beta_2 = \beta_6 = 0$, $H_0 : \beta_2 = \beta_6, \beta_3 = \beta_4$. Based on the different scenarios you want to test, this kind of testing problem can be written in the form of a General Linear Hypothesis, i.e.

$$H_0 : T\underset{\sim}{\beta} = 0$$

where we suppose

- T is an $r \times p$ matrix. $(r \leq p)$
- $T\underset{\sim}{\beta}$ is estimated by $T\widehat{\underset{\sim}{\beta}}$.

(The values of r for the examples above are 1 and 2, respectively.)

(a) (4 points) Build a matrix T to represent the null hypotheses $H_0 : \beta_2 = \beta_6 = 0$ and $H_0 : \beta_2 = \beta_6, \beta_3 = \beta_4$

(b) (3 points) Under $H_0 : T\underset{\sim}{\beta} = 0$, show $\text{Var}(T\widehat{\underset{\sim}{\beta}}) = \sigma^2 T(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}T^{\mathbf{T}}$

(c) (1 point) Under $H_0 : T\underset{\sim}{\beta} = 0$, show $T\widehat{\underset{\sim}{\beta}} = T(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\underset{\sim}{\epsilon}$

(d) (6 points) Under $H_0 : T\underset{\sim}{\beta} = 0$, show $\widehat{\underset{\sim}{\beta}}^{\mathbf{T}}T^{\mathbf{T}}\Sigma^{-1}T\widehat{\underset{\sim}{\beta}} \sim \chi^2_{(r)}$, where Σ^{-1} denotes $\text{Var}(T\widehat{\underset{\sim}{\beta}})$. **(Sorry: Σ should be $\text{Var}(T\widehat{\underset{\sim}{\beta}})$.)**

(e) (2 points) Verify that $(I - \mathbf{H})[\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}T^{\mathbf{T}}C^{-1}T(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}] = 0$

(f) (3 points) Under $H_0 : T\underset{\sim}{\beta} = 0$, show $F_0 = \dfrac{\widehat{\underset{\sim}{\beta}}^{\mathbf{T}}T^{\mathbf{T}}C^{-1}T\widehat{\underset{\sim}{\beta}}/r}{MSE} \sim F_{r,n-p}$ where MSE is computed for the full model (with p parameters).

(g) (3 points) Find the matrix T for the following hypothesis: $\beta_0 = \beta_1 = \beta_2 = \cdots = \beta_k$.

Question 3

(6 points) Problem 3.25 on page 130

Use `lm` function to answer the following questions.

Question 4

(6 points) Problem 3.1 on page 125

Question 5

(2 points) Problem 3.2 on page 125

Question 6

(4 points) Problem 3.3 on page 125

Question 7

(6 points) Problem 3.4 on page 126

Question 8

(bonus) Consider the multiple linear regression model $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$, where $\mathrm{E}[\underset{\sim}{\epsilon}] = 0$ and $\mathrm{E}[\underset{\sim}{\epsilon}\underset{\sim}{\epsilon}^{\mathbf{T}}] = \sigma^2 I$, and $\underset{\sim}{\epsilon}$ is normally distributed and I is $n \times n$ identity matrix. There are k predictors and an intercept in the model. Suppose $\underset{\sim}{\beta} = \begin{bmatrix} \underset{\sim}{\beta_1} \\ \underset{\sim}{\beta_2} \end{bmatrix}$ and $\underset{\sim}{\beta_2}$ contains r coefficients that we want to test, i.e. $H_0 : \underset{\sim}{\beta_2} = \underset{\sim}{\mathbf{0}}$. Partition $\mathbf{X}$ accordingly: $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2]$. The relevant extra sum of squares is:

$$SSR(\underset{\sim}{\beta_2}|\underset{\sim}{\beta_1}) = \hat{\underset{\sim}{\beta}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\underset{\sim}{\mathbf{y}} - \hat{\underset{\sim}{\beta_1}}^{\mathbf{T}}\mathbf{X}_1^{\mathbf{T}}\underset{\sim}{\mathbf{y}}$$

(a) (2 points) Show $SSR(\underset{\sim}{\beta_2}|\underset{\sim}{\beta_1}) = \underset{\sim}{\epsilon}^{\mathbf{T}}(\mathbf{H} - \mathbf{H}_1)\underset{\sim}{\epsilon}$ if $H_0 : \underset{\sim}{\beta_2} = \underset{\sim}{\mathbf{0}}$ is true.

(b) (1 point) Show $\frac{SSR(\underset{\sim}{\beta_2}|\underset{\sim}{\beta_1})}{\sigma^2}$ is distributed as $\chi^2_{(r)}$.

(c) (1 point) Show $\frac{SSR(\underset{\sim}{\beta_2}|\underset{\sim}{\beta_1})/r}{MSE}$ is distributed as $F(r, n-p)$.