MATH 423/533 - ASSIGNMENT 1

To be handed in not later than 11:59pm, 9th October 2017.
Please submit your solutions with relevant R code included as a pdf file via myCourses

Data stored in three data files on the course website contain x and y variables that are to be used for simple linear regression. The data files are `a1-1.txt`, `a1-2.txt` and `a1-3.txt`.

Code for Assignment 1

```
#Read in data set 1
file1<-"http://www.math.mcgill.ca/yyang/regression/data/a1-1.txt"
data1<-read.table(file1,header=TRUE)
plot(data1$x,data1$y,pch=18)
x1<-data1$x
y<-data1$y
```

(a) Perform a least squares fit of a simple linear regression model (including the intercept) in `R` for each of the three data sets. In particular, for each data set

(i) report the parameter estimates arising from a least squares fit; 20 Marks

(ii) produce a plot of the data with the line of best fit superimposed; 10 Marks

(iii) plot (against the x values) the residuals $e_i, i = 1, \ldots, n$, from the fit; 10 Marks

(iv) comment on the adequacy of the straight line model, based on the residuals plot – that is, comment on whether the assumptions of least squares fitting and how they relate to the residual errors ϵ_i are met by the observed data. 10 Marks

Note: the `R` functions `lm`, `coef` and `residuals` will be useful.

(b) Demonstrate both numerically and theoretically what happens to the least squares estimates if the predictor is

(i) subjected to a location shift: $x_{i1} \longrightarrow x_{i1} - m$ for some m; Compute $\hat{\beta}_0$ and $\hat{\beta}_1$ for the location shift data, and compare them with $\hat{\beta}_0$ and $\hat{\beta}_1$ for the original data. 20 Marks

(ii) rescaled: $x_{i1} \longrightarrow l x_{i1}$ for some $l > 0$; Compute $\hat{\beta}_0$ and $\hat{\beta}_1$ for the rescaled data, and compare them with $\hat{\beta}_0$ and $\hat{\beta}_1$ for the original data. 20 Marks

(iii) Compute $\mathbb{E}[\hat{\beta}_0|\mathbf{X}]$, $\mathbb{E}[\hat{\beta}_1|\mathbf{X}]$ and $\mathrm{Var}[\hat{\beta}_0|\mathbf{X}]$, $\mathrm{Var}[\hat{\beta}_1|\mathbf{X}]$ for the location shift data and the rescaled data respectively. Describe also how the properties of these corresponding estimators change, compared with the original data case.
10 Mark

EXTRA QUESTION FOR STUDENTS IN MATH 533

Consider the ridge regression problem

$$\hat{\beta}^{\text{ridge}} = \text{argmin}_{\beta_0,\beta_1} \left\{ \sum_{i=1}^{N} [y_i - \beta_0 - x_{i1}\beta_1]^2 + \lambda \beta_1^2 \right\} (*).$$

Show that this prob- lem is equivalent to the problem

$$\hat{\beta}^c = \text{argmin}_{\beta_0^c,\beta_1^c} \left\{ \sum_{i=1}^{N} [y_i - \beta_0^c - (x_{i1} - \bar{x}_1)\beta_1^c]^2 + \lambda \beta_1^{c2} \right\}$$

Give the correspondence between β^c and the original β in $(*)$. Characterize the solution to this modified criterion.
30 Marks

EXTRA CREDIT QUESTION FOR ALL STUDENTS

In the linear model, it is possible to use a different 'best-fit' criterion based on the Euclidean distance between modelled means vector $\mathbf{X}\beta$ and the observed data vector $\mathbf{y}$: that is, we choose β to minimize

$$\|\mathbf{y} - \mathbf{X}\beta\| = \sqrt{\|\mathbf{y} - \mathbf{X}\beta\|^2} = \sqrt{\sum_{i=1}^{n}(y_i - \mathbf{x}_i\beta)^2}.$$

For simple linear regression, derive the two equations that would need to be solved for the elements of $\beta = (\beta_0, \beta_1)^\top$ to find the minimizing values. Explain how the corresponding model assumptions (in terms of the properties of the residual errors and how they define the least squares procedure) differ from those behind least squares.

20 Marks

Another fitting criterion is based on the sum of absolute differences

$$\sum_{i=1}^{n} |y_i - \mathbf{x}_i\beta|$$

Does this criterion lead to estimators with different statistical characteristics to least squares? Justify your answer.

10 Marks