Exploratory Data Analysis & Statistical Consulting Homework

The aim of the final project, continuing the first homework, is to make an `R` function which is able to fulfill the multiple linear regression when you get a response variable ($\mathbf{y}$) and a set of explanatory variables ($\mathbf{X}_1$). First, you should read the following explanations carefully and, then write a code by yourself.

Suppose you have a response vector of size n as

$$\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$$

and a matrix consisting of p explanatory variables as

$$\mathbf{X}_1 = (\mathbf{x}_1, \cdots, \mathbf{x}_p) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

We would like to build a linear model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

for $i = 1, 2, \ldots, n$. We assume that ϵ_i's are independently and identically distributed from $N(0, \sigma^2)$. To turn this model into a matrix form, we may write it again as a simple form of

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{X} = (\mathbf{1}, \mathbf{X}_1) = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)^T.$$

Using this vector-matrix form, from the theory of linear regression, we know that the least-squares or maximum likelihood estimate of the model parameters is given as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

And the corresponding the fitted (or predicted) values are

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)^T = \mathbf{X}\hat{\boldsymbol{\beta}}$$

and the residuals are

$$\mathbf{e} = (e_1, e_2, \cdots, e_n)^T = \mathbf{y} - \hat{\mathbf{y}} \quad \text{or} \quad e_i = y_i - \hat{y}_i.$$

And the unbiased estimate of the variance component is known as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-p-1} = \mathbf{e}^T\mathbf{e}/(n-p-1).$$

Other necessary statistics for the project are:

- SST: The sum of squares of total is defined as

  $$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

  and its degree of freedom is $df_{SST} = n - 1$.

- SSR and MSR: The sum of squares of regression and mean squares of regression are defined as

  $$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad MSR = SSR/df_{SSR}$$

  and its degree of freedom is $df_{SSR} = p$.

- SSE and MSE: The sum of squares of errors and mean squares of errors are defined as

  $$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad MSE = SSE/df_{SSE}$$

  and its degree of freedom is $df_{SSE} = n-p-1$. Necessarily, it holds that $SST = SSR+SSE$ and $df_{SST} = df_{SSR} + df_{SSE}$.

- F statistic and P-value: F-statistic to testing the null hypothesis $\beta_0 = \cdots = \beta_p = 0$ is defined as

  $$F = MSR/MSE$$

  which follows F distribution with degrees of freedom, df_{SSR} and df_{SSE}. i.e.,

  $$F \sim F(df_{SSR}, df_{SSE}).$$

  Thus, P-value of F statistic is obtained by the right-tail probability from the above distribution.

- R square: *R square* is the proportion of response variation explained by the assumed regression model. That is defined as

  $$R^2 = SSR/SST$$

- Standard error for parameter estimates: Let, first, us define the matrix $\mathbf{C}$ as

  $$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}.$$

  Then, the standard error of $\hat{\beta}_j$ ($j = 0, 1, \cdots, p$) is given as

  $$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

  where C_{jj} is the jth diagonal element of $\mathbf{C}$.

- t statistic and its p-value: t statistic is defined as

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad \text{for } j = 0, 1, \cdots, p.$$

They follow t distribution with degree of freedom df_{SSE}. i.e.,

$$t \sim t(df_{SSE}).$$

Thus, P-value for t statistic is obtained by the sum of the right probability of $|t|$ and the left probability of $-|t|$ from the above distribution.

- Studentized residuals: We define the studentized residuals, r_i $(i = 1, \cdots, n)$, as

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - H_{ii})}} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\sigma}^2(1 - H_{ii})}}$$

where H_{ii} is the ith diagonal element of the *hat matrix* $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Write a program the following directions:

1. The function name should be `MyReg`. And this function has only a single argument, which is a data frame having the response variable in the first element (column) and the explanatory variables following. For example, suppose you have a response vector as y and a set of explanatory variables $x_1, \ldots, x_p$ in a data frame `X` in `R` program. Then, in the console window,

   ```
   > res <- MyReg(X)
   ```

   should give the desired results. Note that the `R` object `X` contains the response variable in the first column of `X` and p explanatory variables appear from the second column. DO NOT include a vector of 1 in `X`.

2. The function `MyReg` should return a value in the form of "list". This "list" object should have the following components:

   - `res$beta`: (*vector-valued*) parameter estimates of $\boldsymbol{\beta}$
   - `res$sig2`: (*scalar-valued*) parameter estimate of σ^2
   - `res$pred`: (*vector-valued*) fitted values, $\hat{y}$, from the model assessed
   - `res$residuals`: (*vector-valued*) residuals, $e = y - \hat{y}$, from the model
   - `res$stdresid`: (*vector-valued*) studentized residuals, r, from the model

- `res$SSR`: (*scalar-valued*) sum of squares of regression from the model
- `res$SSE`: (*scalar-valued*) sum of squares of errors from the model
- `res$F`: (*scalar-valued*) F statistic from the model
- `res$P.value`: (*scalar-valued*) p-value from F stat testing $H_0 : \beta_1 = \cdots = \beta_p = 0$

3. After `MyReg()` execution, some regression analysis results (ANOVA table, parameter estimations) are printed on the console window as the below format:

```
> res <- MyReg(dat)

 == ANALYSIS OF VARIANCE ==

     Source          df          SS          MS           F    P-value
-------------------------------------------------------------------
 Regression           6       195.3       32.55        8.72           0 ***
      Error          93      347.11        3.73
      Total          99      542.41
-------------------------------------------------------------------
 Estimated error variance : 3.7323
 R-squares : 0.3601


 == PARAMETER ESTIMATES ==

                Estimate   Std.Error     t value    Pr(>|t|)
 (Intercept)     -0.7866       1.182     -0.6655      0.5058
         age      0.0198      0.1951      0.1012      0.9194
      height      0.5986      0.1953      3.0653      0.0022 **
      weight     -0.0803      0.6073     -0.1322      0.8948
     smoking      0.2927      0.4105       0.713      0.4758
     therapy      1.2801      0.2794      4.5809           0 ***
     surgery      1.0944      0.2018      5.4219           0 ***
>
```

Note that all elements are well aligned to the right and the explanatory variable names are correctly printed in the result. For the text print and alignment, you may use `cat()` and `encodeString()` functions. On the right side of tables, asterisks (*) appear for some lines. Meaning of this is: for the corresponding p-value, p,

- '`.`' if $0.05 \leq p < 0.1$
- '`*`' if $0.01 \leq p < 0.05$
- '`**`' if $0.001 \leq p < 0.1$
- '`***`' if $p < 0.001$

4. In addition, this function should be able to create four plots. They should be plotted in separate four graphic windows.

- scatterplots between all variables using `pairs()` function : outliers are highlighted by color and size. We claim an observation as an outlier when $|r_i| > 2$.
- scatterplot of y versus $\hat{y}$: circles on this plot should be proportional to the absolute value of r_i. Outliers should be represented by different color and corresponding observation numbers should be placed in the center of circles. Reference line of 0 intercept and 1 slope should be provided.
- residual plot with boxplot : draw a scatterplot of r_i versus $\hat{y}_i$ with 3 reference horizontal lines at $-2, 0, 2$ on the vertical axis. The result from `rug()` function should show on the right vertical axis. Outliers should be shown with difference color, size and shape with indicating the observation numbers. Boxplot of r_i should accompany on the right side of the residual plot. 5 summary numbers appear near boxplot.
- absolute studentized residual plot: draw a scatterplot of $|r_i|$ versus $\hat{y}_i$. Points on this plot represent their sign of r_i by difference shape. Illustrating legend should be placed on the top right corner. In addition, provide 2 smoothing reference curves on the scatterplot, one of which is kernel density plot (KDE) and the other is moving average with order 10 (MA10)
  - KDE: for any point x on x-axis, compute

    $$f(x) = \frac{\sum_{i=1}^{n} w_i |r_i|}{\sum_{i=1}^{n} w_i}$$

    with $w_i = \frac{1}{\sqrt{2\pi}} \exp(-(\hat{y}_i - x)^2/2)$. Then draw a curve of $(x, f(x))$ on the plot.
  - MA(10): for any point x on x-axis, compute the average of $|r_i|$ of 10 nearest neighbors from x, saying $f(x)$. Then draw a curve of $(x, f(x))$ on the plot.

Try to make plots as **close to those in the sample file** as you can!

To test your code, you should apply this function to (1) the simulation data as described in the sample file and (2) any data set provided in `R` program. Write the report using the results from two datasets.

You should submit 2 files in `eClass` webpage: (1) report and (2) R code file. You can use `Hangul`, `MS Word`, or any other word processors for writing the report. However, you should convert the file into **pdf format!** R function must be placed in the separate file, which should be directly usable in my R program by just "copy-and-paste." So, do not include '>' or '+' in the front of command lines. They will severely hinder me to test your code on my computer.