MTH 4230 Lab 3

Due Wed., Feb. 19

1 Part A: Regression Through the Origin

1.1 Real Estate Data

A real estate agency provides data on rental properties for clients in a metropolitan area. The file **properties.txt** contains the real estate data on **age**, **operating expenses** and **taxes**, **vacancy rates**, **square footage**, and **rental rates** for each of $\boldsymbol{n = 81}$ properties.

1. Use `read.table()` to read the data into a *data frame* in R named, say, `my.data`.
2. Use `plot()` to make a scatterplot of `Rent.Rate` (y-axis) versus `Op.Expense` (x-axis). Use the arguments `xlim = c(0, 15)` and `ylim = c(0, 20)` so that the **origin** shows up in the plot.
3. We want to fit the ***no-intercept* model**, i.e.

   $$Y_i \;=\; \beta_1 X_i + \epsilon_i$$

   where `Rent.Rate` is the response and `Op.Expense` the predictor. Fit the model by typing something like:

   ```
   my.reg <- lm(Rent.Rate ~ -1 + Op.Expense, data = my.data)
   ```

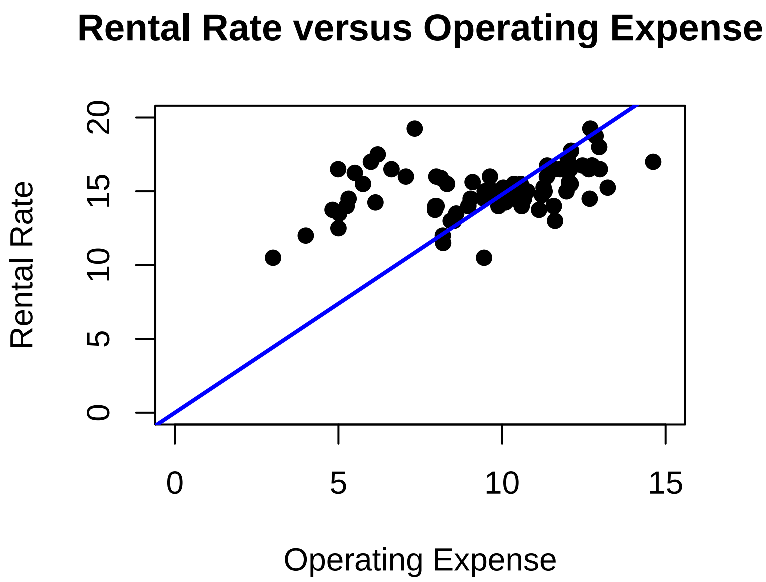
   then look at the results:

   ```
   summary(my.reg)
   ```

4. Use

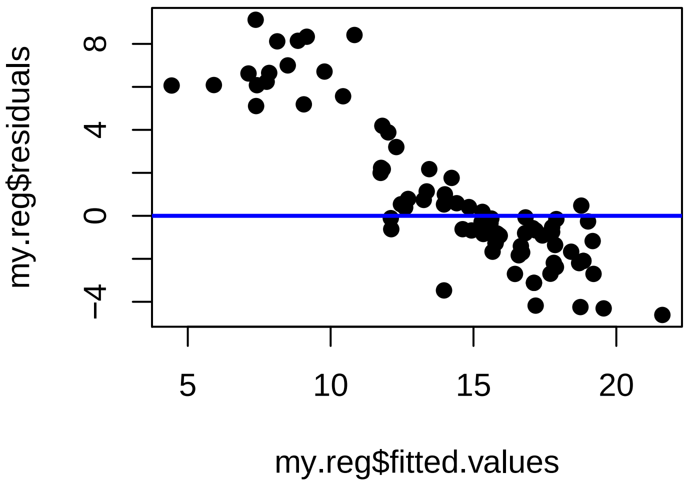
   ```
   abline(my.reg)
   ```

   to add the line to the scatterplot of Step 2. Your plot should look similar to this:



5. Make a plot of the residuals versus the fitted values:

```
plot(x = my.reg$fitted.values,
     y = my.reg$residuals,
     pch = 19)
abline(h = 0, lwd = 2, col = "blue")
```



6. Use `sum()` to obtain the sum of the residuals.

2 Part B: Matrix Approach to Regression

2.1 Real Estate Data (Cont'd)

1. Now use `lm()` to fit the model **with an intercept term**, i.e. the usual regression model

$$Y_i, \; = \; \beta_0 + \beta_1 X_i + \epsilon_i \, ,$$

where `Rent.Rate` is the response and `Op.Expense` the predictor.

Then use `model.matrix()` to obtain (and save) the **design matrix $\boldsymbol{X}$**, for example by typing:

```
X <- model.matrix(my.reg)
```

2. Now use the following matrix functions and operator

`t()`	**matrix transpose** function
`%*%`	**matrix multiplication** operator
`solve()`	**matrix inverse** function

and the design matrix $\boldsymbol{X}$ to compute the matrix

$$(\boldsymbol{X}^T \boldsymbol{X})^{-1} \, .$$

3. Create a vector $\boldsymbol{Y}$ containing the **response** variable `Rent.Rate`, for example by typing:

```
Y <- my.data$Rent.Rate
```

4. Use the matrices $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\boldsymbol{X}$ and the response vector $\boldsymbol{Y}$ to compute the estimated **coefficient vector** $\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$, which is given by

$$\boldsymbol{b} \;=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,.$$

Compare your coefficient vector $\boldsymbol{b}$ to the coefficients obtained by typing

```
summary(my.reg)
```

5. Now obtain the vector of **fitted values** $\hat{\boldsymbol{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$ using matrix operations and the relation

$$\hat{\boldsymbol{Y}} \;=\; \boldsymbol{X}\boldsymbol{b}$$

Compare your vector of fitted values $\hat{\boldsymbol{Y}}$ to those obtained by typing

```
my.reg$fitted.values
```

6. Use your vectors $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ and the relation

$$\boldsymbol{e} \;=\; \boldsymbol{Y} - \hat{\boldsymbol{Y}}$$

to obtain the vector $\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$ of **residuals**.

Compare your vector $\boldsymbol{e}$ of residuals to those obtained by typing

```
my.reg$residuals
```

7. Use the $\sqrt{\textbf{MSE}}$ from the output of

```
summary(my.reg)
```

and the design matrix $\boldsymbol{X}$ to compute the matrix

$$\textbf{MSE}\cdot(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,.$$

Compare the **square roots** of the **diagonal** elements of $\textbf{MSE}\cdot(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ to the **standard errors** obtained by typing

```r
summary(my.reg)
```

3 Part C: Linearly Dependent Design Matrix

3.1 Hypothetical Snakes Data

It can be shown that when the columns of the **design matrix** $\boldsymbol{X}$ are ***linearly dependent*** (e.g. one is a linear combination of the others), the matrix $\boldsymbol{X}^T\boldsymbol{X}$ will be ***singular*** (i.e. *not invertible*).

In this case, **estimates** for the **coefficients** in the model are **not uniquely defined**.

1. Consider the following *hypothetical* data on **lengths** and **weights** of $\boldsymbol{n = 9}$ snakes:

```r
snakes

##   Length Weight
## 1     63    136
## 2     63    198
## 3     63    194
## 4     63    140
## 5     63     93
## 6     63    172
## 7     63    116
## 8     63    174
## 9     63    145
```

After creating the (hypothetical) `snakes` data frame:

```r
length <- c(63, 63, 63, 63, 63, 63, 63, 63, 63)
weight<- c(136, 198, 194, 140, 93, 172, 116, 174, 145)
snakes <- data.frame(Length = length, Weight = weight)
```

try fitting the linear regression model:

```r
my.reg <- lm(Weight ~ Length, data = snakes)
```

Then look at the outputs of the following:

```r
summary(my.reg)
model.matrix(my.reg)
```