# MTH 4230 R Notes 4

## 1 Multiple Regression

- Suppose we've created a *data frame* (e.g. using `read.table()`) called `my.data`, shown below.

```
my.data

##    response x1  x2  x3
## 1        24 10 2.4 100
## 2        22 11 1.9  75
## 3        25 13 2.6  68
## 4        22 12 1.8  77
## 5        26 13 1.3  80
## 6        26 16 0.9  81
## 7        25 15 1.2  72
## 8        27 20 1.0  55
## 9        30 18 1.3  39
## 10       33 19 1.0  67
## 11       29 19 0.8  77
## 12       30 22 0.7  94
```

We fit a multiple regression model using `lm()` and view the results using `summary()`. For example, to fit the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

we type:

```
my.reg <- lm(response ~ x1 + x2 + x3, data = my.data)
```

and to view the results, we type:

```
summary(my.reg)

##
## Call:
## lm(formula = response ~ x1 + x2 + x3, data = my.data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.8744 -1.0259 -0.3707  1.4058  4.0007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.01050    8.82022   1.588   0.1508
## x1           0.78376    0.31312   2.503   0.0368
## x2           0.60720    1.87980   0.323   0.7550
## x3          -0.00761    0.04393  -0.173   0.8668
##
## Residual standard error: 2.187 on 8 degrees of freedom
## Multiple R-squared:  0.6936,Adjusted R-squared:  0.5787
## F-statistic: 6.036 on 3 and 8 DF,  p-value: 0.01884
```

From the output, we conclude that

- The ***least squares estimates*** of the true (unknown) regression coefficients $\beta_0, \beta_1, \beta_2$, and $\beta_3$ are $\boldsymbol{b_0 = 14.0105}$, $\boldsymbol{b_1 = 0.7838}$, $\boldsymbol{b_1 = 0.6072}$, and $\boldsymbol{b_1 = -0.0076}$. Thus the fitted model is

$$\hat{Y} = 14.0105 + 0.7838X_1 + 0.6072X_2 - 0.0076X_3.$$

- The ***standard errors*** of the estimates are

$$\begin{aligned} \boldsymbol{s\{b_0\}} &= \boldsymbol{8.8202} \\ \boldsymbol{s\{b_1\}} &= \boldsymbol{0.3131} \\ \boldsymbol{s\{b_2\}} &= \boldsymbol{1.8798} \\ \boldsymbol{s\{b_3\}} &= \boldsymbol{0.0439} \end{aligned}$$

- The test statistics and p-values for the ***t tests*** of hypotheses of the form

$$\begin{aligned} H_0 : \beta_k &= 0 \\ H_a : \beta_k &\neq 0 \end{aligned}$$

  are

  * $\boldsymbol{t = b_0/s\{b_0\} = 1.588}$ and the p-value is **0.1508** (from the $t$ distribution with $\boldsymbol{n - 2 = 8}$ degrees of freedom). Thus we fail to reject $H_0$ and conclude that $\beta_0$ isn't different from 0.
  * $\boldsymbol{t = b_1/s\{b_1\} = 2.503}$ and the p-value is **0.0368** (from the $t$ distribution with $\boldsymbol{n - 2 = 8}$ degrees of freedom). Thus we reject $H_0$ and conclude that $\beta_1$ is different from 0.
  * $\boldsymbol{t = b_2/s\{b_2\} = 0.323}$ and the p-value is **0.755** (from the $t$ distribution with $\boldsymbol{n - 2 = 8}$ degrees of freedom). Thus we fail to reject $H_0$ and conclude that $\beta_2$ isn't different from 0.
  * $\boldsymbol{t = b_3/s\{b_3\} = -0.173}$ and the p-value is **0.8668** (from the $t$ distribution with $\boldsymbol{n - 2 = 8}$ degrees of freedom). Thus we fail to reject $H_0$ and conclude that $\beta_3$ isn't different from 0.

- The square root of the ***mean squared error*** is labeled `Residual standard error`, and its value is $\sqrt{\mathbf{MSE}} = \mathbf{2.187}$.
- The ***coefficient of determination*** is $\boldsymbol{R^2} = \mathbf{0.6936}$, labeled `Multiple R-squared`. Thus 69.36% of the $Y$ variation is explained by the regression model containing these three predictors.
- The adjusted $R^2$ is $\boldsymbol{R^2_a} = \mathbf{0.5787}$, labeled `Adjusted R-squared`.
- The $F$ statistic for the ***regression model F test*** of the hypotheses

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_2 = \beta_3 = 0 \\ H_a &: \quad \text{not all of } \beta_1, \beta_2, \text{ and } \beta_3 \text{ equal } 0 \end{aligned}$$

  is $\boldsymbol{F = \mathbf{MSR/MSE} = 6.036}$. From the $F$ distribution with numerator and denominator degrees of freedom **3** and **8**, respectively, the ***p-value*** is **0.01884**. Thus we reject $H_0$ and conclude that at least one of $\beta_1$, $\beta_2$, or $\beta_3$ is different from 0.

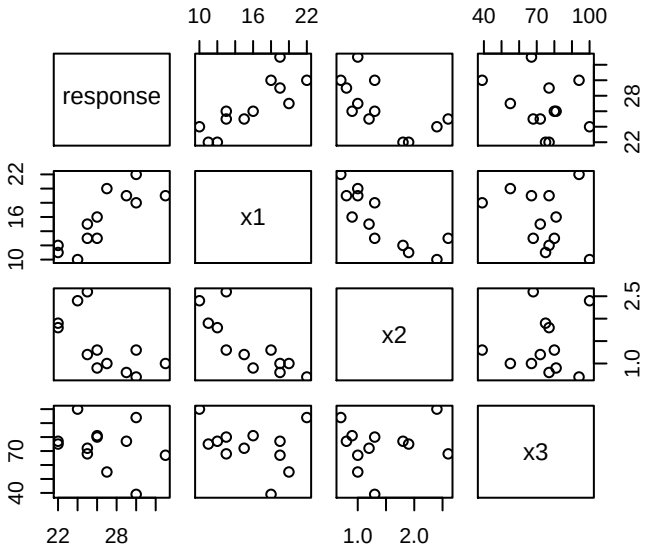
## 1.1 Scatterplot Matrices and Correlation Matrices

- We can make a ***scatterplot matrix*** using the function

```
pairs()      # Creates a scatterplot matrix from the columns of a
             # data frame or matrix.
```

The `pairs()` function takes a *data frame* (or a *matrix*) as an argument and produces a scatterplot matrix from its columns. For example, using the *data frame* `my.data` from above:

```
pairs(my.data)
```




- We can compute the (estimated) ***correlation matrix*** or ***covariance matrix*** by passing a *data frame* containing only numeric columns (or a *matrix*) to the `cor()` and `cov()` functions:

```
cor()      # Computes the correlation matrix of the variables stored as
           # as columns of a data frame or matrix.
cov()      # Computes the covariance matrix of the variables stored
           # as columns of a data frame or matrix.
```

- For example, using the `my.data` *data frame* from above the (estimated) **covariance matrix** is:

```
cov(my.data)

##            response         x1         x2         x3
## response  11.356061  10.939394 -1.3507576 -16.113636
## x1        10.939394  15.333333 -2.0151515 -19.090909
## x2        -1.350758  -2.015152  0.3935606   1.356818
## x3       -16.113636 -19.090909  1.3568182 259.477273
```

The diagonal elements are the **variances** and the off-diagonal elements are the **covariances**. Thus,

  - The ***variance*** of `response` is $\boldsymbol{S^2\{Y\} = \text{Var}(Y) = 11.3561}$.
  - The ***covariance*** between `response` and `x1` is $\boldsymbol{S^2\{Y, X_1\} = 10.9394}$.

- The ***covariance*** between `x1` and `x2` is $\boldsymbol{S^2\{X_1, X_2\} = -2.0152}$.
- Etc.

The (estimated) **correlation matrix** is

```
cor(my.data)

##            response         x1         x2         x3
## response  1.0000000  0.8290126 -0.6389366 -0.2968452
## x1        0.8290126  1.0000000 -0.8203206 -0.3026625
## x2       -0.6389366 -0.8203206  1.0000000  0.1342660
## x3       -0.2968452 -0.3026625  0.1342660  1.0000000
```

The diagonal elements of the **correlation matrix** are all one and the off-diagonal elements are the **correlations**.