Introduction to Multiple Linear Regression

Instructor: Dr. Sharandeep Singh Pandher*

1 Introduction

The multiple linear regression model (MLR) is an extension of a simple linear regression model. This model can be used to assess the relationship between two or more predictors and a single continuous response variable. For example, a public health researcher interested in social factors that influence heart disease. He survey 500 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease. We wish to build the multiple linear regression model that fits the data better than the simple linear regression model. This model is a foundation of statistical analysis in many areas such as biology, medical science, engineering, etc, because of its power and flexibility. Assumptions of multiple linear regression as below:

- **Homogeneity of variance (homoscedasticity)**: the size of the error in our prediction doesn't change significantly across the values of the independent variable
- **Independence of observations**: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among variables.
  In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (correlation coefficient > 0.6), then only one of them should be used in the regression model.
- **Normality**: The data follows a normal distribution.
- **Linearity:** the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

So the above assumptions can be explain mathematically in the Multiple linear regression model as below: In the multiple regression model, we assume that a linear relationship exists between the variable y, which

*Address: Department of Mathematics and Statistics, University of Alberta, Edomoton, AB,T6G 2G1, Canada, e-mail: sharand1@ualberta.ca;

we call the response variable, and p predictor variables, $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, ..., \boldsymbol{x_p}$. Consider the regression model

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{ip}\beta_2 + ... + x_{ip}\beta_p + \epsilon_i \quad (1)$$

where $i = 1, 2, , n$ is the index for n successive observations. Let us assume that the errors ϵ_i are distributed normally, independently and identically with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i. The equation above can be written as

$$y_i = \boldsymbol{x_i}\boldsymbol{\beta} + \epsilon_i \quad (2)$$

where $\boldsymbol{x_i} = (1, x_{i1}, x_{i2}, ...x_{ip})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$.The equation (1) can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ . & . & . & ... & . \\ . & . & . & ... & . \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

which can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (3)$$

where $\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ . & . & . & ... & . \\ . & . & . & ... & . \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix}$ is a matrix of size $n \times (p+1)$

$\boldsymbol{y} = (y_1, y_2, ..., y_n)^\top$,$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^\top$ and $\boldsymbol{\epsilon} \sim \boldsymbol{N}(0, \boldsymbol{I}\sigma^2)$, Now $E(\boldsymbol{y}) = E(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{X}\boldsymbol{\beta} + 0 = \boldsymbol{X}\boldsymbol{\beta}$ and $Var(\boldsymbol{y}) = Var(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$.Now the distribution of $\boldsymbol{y}$ is

$\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}\sigma^2)$

The first step in multiple linear regression is to determine the vector of **least squares estimators** or maximum likelihood estimators, $\hat{\boldsymbol{\beta}}$, which gives the linear combination $\hat{\boldsymbol{y}}$ that minimizes the length of the error vector. Basically the estimator $\hat{\boldsymbol{\beta}}$ provides the least possible value to sum of the squares difference between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$. Algebraically $\hat{\boldsymbol{\beta}}$ can be expressed as matrix notation. An important stipulation in multiple regression analysis is that the variables $\boldsymbol{x_1}, \boldsymbol{x_2}, , \boldsymbol{x_p}$ be linearly independent. This implies that the correlation between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ (with $i \neq j$) is small. Now, since the objective of multiple regression is to minimize the sum of the squared errors, the regression coefficients that meet this condition are determined by solving the normal equations, which can be obtained from the **least squares**/maximum likelihood method.

2 Estimation Methods

2.1 Ordinary Least Squares Method

The principle of least squares is again useful in estimating the regression parameters. For the model (3), we are required to estimate $\boldsymbol{\beta}$ simultaneously to minimize the sum of squared deviations. That is, we wish to find the vector of least squares estimators $\hat{\boldsymbol{\beta}}$, that minimizes

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

where $\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y}$ is a 1×1 or a scaler and its transpose$(\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y})^\top = \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta}$ is the same scaler. The least squares estimator must satisfy

$$\begin{aligned} (\frac{\partial S}{\partial \boldsymbol{\beta}})_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} &= -2\boldsymbol{X}^\top \boldsymbol{y} + 2[\boldsymbol{X}^\top \boldsymbol{X}]\boldsymbol{\beta} = 0 \\ \Rightarrow (\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{\beta} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\ \Rightarrow \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \end{aligned}$$

provided that the inverse matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists. The matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is always invertible if the predictors are linear independent, that is, if no column of the $\boldsymbol{X}$ matrix is a linear combination of the other columns.

2.2 Maximum likelihood method

Find the Maximum Likelihood estimators of regression parameter vector $\boldsymbol{\beta}$ and the error variance σ^2. The likelihood function $L(\boldsymbol{\beta}, \sigma^2|y)$ is the joint pdf of $f(y|\boldsymbol{\beta}, \sigma^2)$:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2|y) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{(y_i - \boldsymbol{x_i}\boldsymbol{\beta})^2}{2\sigma^2}] \\ &= (2\pi\sigma^2)^{-n/2} exp[-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}] \\ l(\boldsymbol{\beta}, \sigma^2|y) &= logL(\boldsymbol{\beta}, \sigma^2|y) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} \\ \frac{\partial l}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2} \frac{\partial [(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = \boldsymbol{0} \end{aligned}$$

which is a likelihood equation for $\boldsymbol{\beta}$.

$$
\begin{aligned}
-\frac{1}{2\sigma^2}\frac{\partial[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})]}{\partial\boldsymbol{\beta}} &= \mathbf{0} \\
\Rightarrow \frac{\partial[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})]}{\partial\boldsymbol{\beta}} &= \mathbf{0}, since(\boldsymbol{X}\boldsymbol{\beta})^\top = (\boldsymbol{\beta}^\top\boldsymbol{X}^\top) \\
\Rightarrow \frac{\partial[(\boldsymbol{y}^\top-\boldsymbol{\beta}^\top\boldsymbol{X}^\top)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})]}{\partial\boldsymbol{\beta}} &= \mathbf{0} \\
\Rightarrow \frac{\partial[\boldsymbol{y}^\top\boldsymbol{y}-\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y}+\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}]}{\partial\boldsymbol{\beta}} &= \mathbf{0} \\
\Rightarrow -\boldsymbol{X}^\top\boldsymbol{y}-\boldsymbol{X}^\top\boldsymbol{y}+[(\boldsymbol{X}^\top\boldsymbol{X})+(\boldsymbol{X}^\top\boldsymbol{X})^\top]\boldsymbol{\beta} &= 0 \\
\Rightarrow -2\boldsymbol{X}^\top\boldsymbol{y}+[2(\boldsymbol{X}^\top\boldsymbol{X})]\boldsymbol{\beta} &= 0, since(\boldsymbol{X}^\top\boldsymbol{X})^\top = (\boldsymbol{X}^\top\boldsymbol{X}) \\
\Rightarrow (\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{\beta} &= \boldsymbol{X}^\top\boldsymbol{y} \\
\Rightarrow (\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{\beta} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\
\Rightarrow \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned}
$$

The estimator $\hat{\boldsymbol{\beta}}$ is called least squares or maximum likelihood estimator of $\boldsymbol{\beta}$.
Also $\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y}$ are scalers and $\frac{\partial(\boldsymbol{X}^\top\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \frac{\partial(\boldsymbol{\beta}^\top\boldsymbol{X})}{\partial\boldsymbol{\beta}} = \boldsymbol{X}$

$$
\begin{aligned}
\frac{\partial(\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= \frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta})^\top(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{\beta}+\frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta})^\top[(\boldsymbol{X}^\top\boldsymbol{X})^\top\boldsymbol{\beta}] \\
&= (\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{\beta}+[(\boldsymbol{X}^\top\boldsymbol{X})^\top\boldsymbol{\beta}] \\
&= 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta},
\end{aligned}
$$

where $\boldsymbol{X}^\top\boldsymbol{X}$ is symmetric. Again

$$
\begin{aligned}
\frac{\partial l}{\partial\sigma^2} &= \frac{\partial}{\partial\sigma^2}[-\frac{n}{2}log(2\pi\sigma^2)]-\frac{\partial}{\partial\sigma^2}[\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}] \\
&= -\frac{n}{2\sigma^2}+\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^4} = 0,
\end{aligned}
$$

This likelihood equation for σ^2
Hence $\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}$
The fitted regression model is

$$
\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\
&= \boldsymbol{H}\boldsymbol{y}
\end{aligned}
$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is usually called the hat matrix. It plays an important role in regression analysis. The difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$ is the residual $\hat{e}_i = y_i - \hat{y}_i$. The n residuals can be conveniently written in vector form:

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

There are other ways to express the vector of residual $\boldsymbol{e}$:

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (1 - \boldsymbol{H})\boldsymbol{y}$$

Example:1 Swiss Data. Swiss data sets has Fertility as the response where as Agriculture, Examination, Education, Catholic and Infant.Mortality are predictors.

Fitting multiple linear regression model: We start with the scatter plot matrix. These plot provide a compact display of the relationship between a number of variable pairs. The advantage of this plot is that you can scan the plots for highly correlated variables and for outliers. We see from Figure 1 that Fertility has positive correlation with Agriculture and Infant.Mortality but Fertility has negative correlation with Examination and Education; Fertility has a curvature correlation with Catholic (ie The relationship between Fertility and Catholic may not be significant).

To prepare R-Code for Swiss data.

```
rm(list=ls())

library(faraway)

library(graphics)

data("swiss")

# Response variable

Fertility=Y=swiss[,1]

# Predictor variables

Agriculture=swiss[,2]

Examination=swiss[,3]

Education=swiss[,4]
```

```
Catholic=swiss[,5]

Infant.Mortality=swiss[,6]

CC=cbind(Agriculture,Examination,Education,Catholic,Infant.Mortality)

# Draw Matrix plot

pairs(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality, panel=panel.smooth)

# Form design matrix

X=matrix(cbind(rep(1,length(Agriculture)), CC), nrow=length(Agriculture), ncol=6)

betahat=solve(t(X)%*%X)%*%t(X)%*%Y
betahat
             [,1]
[1,] 66.9151817
[2,] -0.1721140
[3,] -0.2580082
[4,] -0.8709401
[5,]  0.1041153
[6,]  1.077048

# Fitting MLR

fit1=lm(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality)
summary(fit1)

Call:
lm(formula = Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
```

```
Agriculture       -0.17211    0.07030  -2.448  0.01873 *
Examination       -0.25801    0.25388  -1.016  0.31546
Education         -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic           0.10412    0.03526   2.953  0.00519 **
Infant.Mortality   1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10

# calculate Hat Matrix

H=X%*%solve(t(X)%*%X)%*%t(X)
```

The fitted regression line is:

$Fertility = 66.91518 - 0.17211 * Agriculture - 0.25801 * Examination - 0.87094 * Education + 0.10412 * Catholic + 1.07705 * Infant.Mortality$

- $\hat{\beta}_0 = 66.91518$. This is the intercept, the value of Fertility when all the predictors take the value zero.
- $\hat{\beta}_1$ = -0.17211. In this model, if the amount of Agriculture increases by 1 unit, the amount of Fertility will decrease (**on average**) by 0.17211. (assuming all other predictors are held constant).
- $\hat{\beta}_2$ = -0.25801. In this model, if the amount of Examination increases by 1 unit, the amount of Fertility will decrease (**on average**) by 0.25801. (assuming all other predictors are held constant).
- $\hat{\beta}_3$ = -0.87094. In this model, if the amount of Education increases by 1 unit, the amount of Fertility will decrease (**on average**) by 0.87094. (assuming all other predictors are held constant).
- $\hat{\beta}_4$ = 0.10412. In this model, if the amount of Catholic increases by 1 unit, the amount of Fertility will increase (**on average**) by 0.10412. (assuming all other predictors are held constant).
- $\hat{\beta}_5$ = 1.07705. In this model, if the amount of Infant.Mortality increases by 1 unit, the amount of Fertility will increase (**on average**) by 1.07705. (assuming all other predictors are held constant).

Adjusted-R^2 is 67.1 percentage indicating that 67.1 percentage of the variation in the Fertility is explained by the Agriculture, Examination, Education, Catholic, and Infant.Mortality, which says that this model fits the data pretty well.

What is difference between R^2 and Adjusted-R^2?

Answer: Both R^2 and the adjusted-R^2 give you an idea of how many data points fall around the fitted

explained by predictors (x_1, x_2, ...)

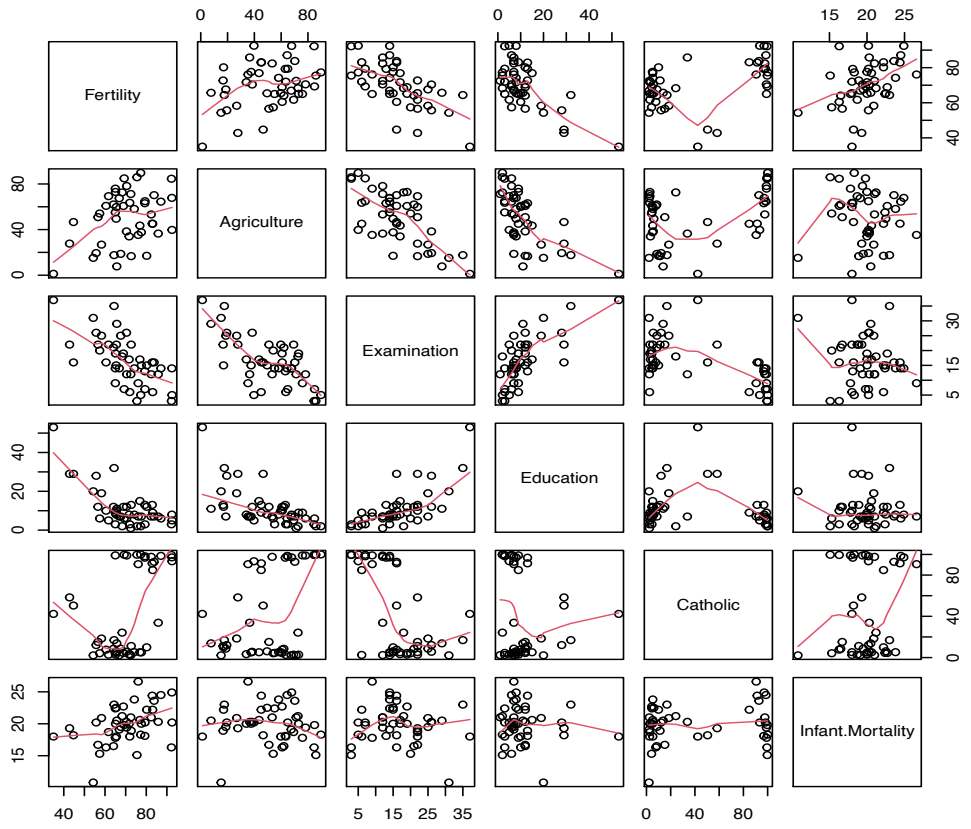

Figure 1: Matrix Plot

regression line. However, the main difference between R^2 and the adjusted-R^2 is that R^2 assumes that each predictor explains the percentage of variation in the response variable and the adjusted-R^2 tells the percentage of variation explained by only the predictors that actually affect the response variable.

The adjusted- R^2 compares the interpreting power of regression models that contain different numbers of predictors. Suppose you are comparing the R^2 of a seven-predictor model to a one-predictor model. **Does the Seven predictors model have a higher R^2 as its better? OR is R^2 higher because it has more predictors?** To answer this question, you need to compare adjusted R^2 values. The adjusted R^2 is a modified version of R^2 that has been adjusted for the number of predictors in the model. One could continue to add predictors to the model which would continue to improve the ability of the predictors to explain the response variable, although some of this increase in R^2 would be simply due to chance variation in that particular sample. It is always lower than the R^2.

The adjusted-R^2, attaches a small penalty to adding more predictors. If adding a predictor raises the R^2 for a regression, thats a better indication of improving the model. Otherwise it merely raises the unadjusted R^2. The formula for adjusted-R^2 is

$$Adjusted - R^2 = 1 - \frac{SS_{Res}/(n-p-1)}{SS_T/(n-1)}$$

2.3 Properties of estimators

- $\hat{\boldsymbol{\beta}}$ is unbiased estimate of $\boldsymbol{\beta}$

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top E(\boldsymbol{y}) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top E(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top E(\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta}. \end{aligned}$$

- Variance of $\hat{\boldsymbol{\beta}}$

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}) &= var((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top var(\boldsymbol{y})[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top]^\top \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\sigma^2 \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\sigma^2. \end{aligned}$$

```
# Calculate variance-covariance matrix of betahat manually

varbetahat= solve(t(X)%*%X)*51.34251 # variance covariance matrix
 round(varbetahat,3)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,] 114.619 -0.485 -1.203 -0.281 -0.022 -3.266
[2,]  -0.485  0.005  0.004  0.005 -0.001  0.007
[3,]  -1.203  0.004  0.064 -0.027  0.005  0.000
[4,]  -0.281  0.005 -0.027  0.033 -0.003  0.012
[5,]  -0.022 -0.001  0.005 -0.003  0.001 -0.003
[6,]  -3.266  0.007  0.000  0.012 -0.003  0.146

# variance-covariance matrix from fitted model(ie from lm function)
round(vcov(fit1),3)
                (Intercept) Agriculture Examination Education Catholic Infant.Mortality
(Intercept)         114.619      -0.485      -1.203    -0.281   -0.022          -3.266
Agriculture          -0.485       0.005       0.004     0.005   -0.001           0.007
Examination          -1.203       0.004       0.064    -0.027    0.005           0.000
Education            -0.281       0.005      -0.027     0.033   -0.003           0.012
```

Catholic	-0.022	-0.001	0.005	-0.003	0.001	-0.003
Infant.Mortality	-3.266	0.007	0.000	0.012	-0.003	0.146

- If σ is fixed then the log-likelihood is minimized when the term $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ is minimized. In this situation the least squares estimator $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ is the same as the maximum likelihood estimator of $\boldsymbol{\beta}$ where the errors are normally distributed.

- Estimation of σ^2: From simple linear regression we know that

$$SS_{Res} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}\hat{e}_i^{\,2} = \hat{\boldsymbol{e}}^\top\hat{\boldsymbol{e}}$$

since $\hat{\boldsymbol{e}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$

$$\begin{aligned}
SS_{Res} &= \hat{\boldsymbol{e}}^\top\hat{\boldsymbol{e}} = (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{y}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y})^\top(\boldsymbol{y}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}) \\
&= (\boldsymbol{y}-\boldsymbol{H}\boldsymbol{y})^\top(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{y}) \\
&= ((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y})^\top((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}) \\
&= \boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{H})^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y} \\
&= \boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y} \\
&= \boldsymbol{y}^\top\boldsymbol{y}-\boldsymbol{y}^\top\boldsymbol{H}\boldsymbol{y} \\
&= \boldsymbol{y}^\top\boldsymbol{y}-\boldsymbol{y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\
&= \boldsymbol{y}^\top\boldsymbol{y}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y})^\top\boldsymbol{X}^\top\boldsymbol{y} \\
&= \boldsymbol{y}^\top\boldsymbol{y}-\hat{\boldsymbol{\beta}}\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned}$$

The residual sum of squares has $n-p-1$ degrees of freedom associated with it since $p+1$ parameters are estimated in the regression model. The residual mean square for the model with $p+1$ parameters is

$$MS_{Res} = \frac{SS_{Res}}{n-p-1}$$

An unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = MS_{Res}$$

The least squares estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-p-1}$$

which is unbiased estimate of σ^2 but not the same as the maximum likelihood estimates (MLE) of σ^2. The MLE of $\hat{\sigma}^2$ is not unbiased estimate of σ^2.

```
> SSRES= t(Y)%*%Y - t(betahat)%*%t(X)%*%Y
> SSRES
         [,1]
[1,] 2105.043
> MSRES=SSRES/41
> MSRES
         [,1]
[1,] 51.34251
```

3 Test for significance of regression

Once we have estimated the parameters in the model, we face two immediate questions: 1) What is the overall adequacy of the model? and 2) Which specific predictors seem important? This section considers following cases: 1) Test for significance of regression (sometimes called the global test of model adequacy) 2) Tests on individual regression coefficients (or groups of coefficients)

3.1 Test for significance of regression

The test for significance of regression is a test to determine if there is a linear relationship between the response y and any of the predictor variables $x_1, x_2, ...x_p$. This procedure is often called the overall test of model adequacy. The appropriate hypotheses are

$$\begin{aligned} \boldsymbol{H_0}: \beta_1 &= \beta_2 = ... = \beta_p = 0 \\ \boldsymbol{H_1}: \beta_j &\neq 0, \end{aligned}$$

for atleast one j.

Rejection of this null hypothesis implies that at least one of the predictors $x_1, x_2, ...x_p$ contributes significantly to the model. The test procedure is a generalization of the ANOVA used in simple linear regression. The total sum of squares is partitioned in two parts:

$$SS_T = SS_R + SS_{Res}$$

This leads to an ANOVA procedure with the F_0 test statistic which follows the $F_{p,n-p-1}$ distribution. This test reject H_0 if $F_0 > F_{\alpha,p,n-p-1}$. The ANOVA table is given by

Table 1: ANOVA table for multiple linear regression

Source of Variation	sum of squares	DF	Mean Square	F_0
Regression	SS_R	p	$MS_R = \frac{SS_R}{p}$	$\frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$n-p-1$	$MS_{Res} = \frac{SS_{Res}}{n-p-1}$	
Total	SS_T	$n-1$		

The SS_T, SS_R, and SS_{Res} can be written as in vector form:

$$\begin{aligned} SS_T &= \boldsymbol{y}^\top \boldsymbol{y} - \frac{(\sum_{i=1}^n y_i)^2}{n} \\ SS_R &= \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{y} - \frac{(\sum_{i=1}^n y_i)^2}{n} \\ SS_{Res} &= \boldsymbol{y}^\top \boldsymbol{y} - \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{y} \end{aligned}$$

```
> anova(fit1)
Analysis of Variance Table

Response: Fertility
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Agriculture       1  894.84  894.84 17.4288 0.0001515 ***
Examination       1 2210.38 2210.38 43.0516 6.885e-08 ***
Education         1  891.81  891.81 17.3699 0.0001549 ***
Catholic          1  667.13  667.13 12.9937 0.0008387 ***
Infant.Mortality  1  408.75  408.75  7.9612 0.0073357 **
Residuals        41 2105.04   51.34
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
#### This ANOVA table can be collapsed into Analysis of Variance Table###
> Regression=CC
> fit2=lm(Fertility~Regression)
> anova(fit2)
Analysis of Variance Table

Response: Fertility
            Df Sum Sq Mean Sq F value    Pr(>F)
Regression   5 5072.9  1014.58  19.761 5.594e-10 ***
Residuals   41 2105.0    51.34
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

3.2 Tests on Individual Regression Coefficients

Once we have determined from the test of overall model significance that at least one of the predictors is important, a next step is to determine which ones are important.
Note:

- Adding a predictor always causes an increase in SS_R and a reduction in SS_{Res}.
- We must decide whether the increase in SS_R is sufficient to justify using the additional predictor.
- Adding an insignificant predictor may increase SS_{Res}, which may decrease the usefulness of the model

To determine the significance of an individual predictor x_j , $j = 1, 2, , p$, we test the hypotheses:

$$H_0 : \beta_j = 0 \quad \text{vs.} H_1 : \beta_j \neq 0.$$

If H_0 not is rejected, then predictor x_j can be deleted from the model. The test statistic for this test is:

$$t_0 = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{MS_{Res}C_{jj}}}$$

where C_{jj} be the j-th diagonal element of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ The test statistic t_0 has t -distribution with $n-p-1$ degrees of freedom. This is a partial or marginal test because any estimate of the regression coefficient depends on all of the other predictors. This test is a test of contribution of x_j given the other predictors in the model.

```
> summary(fit1)

Call:
lm(formula = Fertility ~ Agriculture + Examination + Education +
    Catholic + Infant.Mortality)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
```

```
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

We can build a confidence interval for β_j in the usual way, all we need to know the standard error of $\hat{\beta}_j$. As always, the margin of error is just the product of the standard error and a critical value. Here the critical value comes from the t-distribution on $n-p-1$ degrees of freedom. So a confidence interval for β_j is

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2} SE(\hat{\beta}_j)$$

A 95 percentage confidence interval for the coefficient of Education is β_3

$$\begin{aligned} \hat{\beta}_3 \pm t_{41,0.025} SE(\hat{\beta}_3) &= \hat{\beta}_3 \pm qt(0.975, 41) SE(\hat{\beta}_3) \\ &= -0.87094 \pm 2.021 * 0.18303 \\ &= -0.87094 \pm 0.36990 \\ &= [-1.24084, -0.50104]. \end{aligned}$$

4 Multicolinearity

A serious problem that may dramatically impact the usefulness of a regression model is multicollinearity, or near-linear dependence among the predictor variables. Multicollinearity implies near-linear dependence among the predictors. The predictors are the columns of the $\boldsymbol{X}$ matrix, so clearly an exact linear dependence would result in a singular $\boldsymbol{X}^\top \boldsymbol{X}$. The presence of multicollinearity can dramatically impact the ability to estimate regression coefficients and other uses of the regression model. Alternatively, multicollinearity occurs when the predictor variables in a regression are so highly correlated that it becomes difficult or impossible to distinguish their individual effects on the response variable.

- Multicollinearity provides redundant information about the response.

- Example of multicollinear predictors are height and weight of a person, years of education and income, and assessed value and square footage of a home.

- Consequences of high multicollinearity:

    1. Multicollinearity inflates the variances of the parameter estimates and hence this may lead to lack of statistical significance of individual predictor variables even though the overall model may be significant.

    2. The presence of multicollinearity can cause serious problems with the estimation of $\hat{\boldsymbol{\beta}}$ and the interpretation

Identify multicollinearity:
To identify the multicollinearity, we need to focus on the following points:

1. Examination of correlation matrix.

2. Variance inflation factor (VIF)

1. **Examination of correlation matrix:**

    - Large correlation coefficients in the correlation matrix of predictor variables indicate multicollinearity.

    - If there is a multicollinearity between any two predictor variables, then the correlation coefficient between these two variables will be near to unity. $|\rho| \approx 1$

2. **Variance inflation factor (VIF):**

    - The VIF quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

    - Calculate the variance inflation factors for each predictor x_j:

$$VIF_j = \frac{1}{1 - {R_j}^2},$$

where R_j^2 is the coefficient of determination of the model that includes all predictors except the jth predictor. Note that a VIF exists for each of the p predictors in a multiple regression model.

- The VIF is an index which measures how much variance of an estimated regression coefficient is increased because of multicollinearity.

- The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity and requiring correction.

```
library(olsrr)
> fit1=lm(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality)
>
> ols_vif_tol(fit1)
         Variables Tolerance      VIF
1      Agriculture 0.4378036 2.284129
2      Examination 0.2720778 3.675420
3        Education 0.3603678 2.774943
4         Catholic 0.5162195 1.937160
5 Infant.Mortality 0.9029001 1.107542
```

Note: The VIF of all predictors of Swiss data less than 4 that indicates no multicollinearity exist.

5 Robust regression

Least squares works well when there are normal errors, but performs poorly for long-tailed errors ie when data are contaminated with outliers or influential observations, and Robust regression can also be used for the purpose of detecting influential observations **but question arise here How to Perform Robust regression in R ?**. The following steps required to perform Robust regression:

1. To identity outliers of data (Swiss data) using Ordinary Least Squares Regression if outliers available that indicate to employ Robust regression. We will begin by running an OLS regression and looking at diagnostic plots examining residuals, fitted values, Cooks distance, and leverage. From figure 2, we can identify observations 6, 37, and 47 as possibly problematic(outliers) to our model and thus we may benefit from performing robust regression.

```
#########Finding outliers or influential observations###
  library(foreign)
  library(MASS)
> opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
> plot(fit1, las = 1)
```

2. **To Perform Robust Regression**: lets use the $rlm()$ function to fit a robust regression model.To determine if this robust regression model offers a better fit to the data compared to the OLS model, we can calculate the residual standard error(RSE) of each model. The residual standard error (RSE) is a way to measure the standard deviation of the residuals in a regression model. The lower the value for RSE, the more closely a model is able to fit the data.The following code shows how to calculate the RSE for each model:

```
#############################Robust regression######
fit1=lm(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality)
fit3=rlm(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality)
> summary(fit1)$sigma
[1] 7.165369
> summary(fit3)$sigma
[1] 6.63321
```

We can see that the RSE for the robust regression model is much lower than the ordinary least squares regression model, which tells us that the robust regression model offers a better fit to the data.

Note: **Variable Selection procedure** will discuss in the Next Chapter. **Transformation procedure** will discuss if required in the Next chapters.

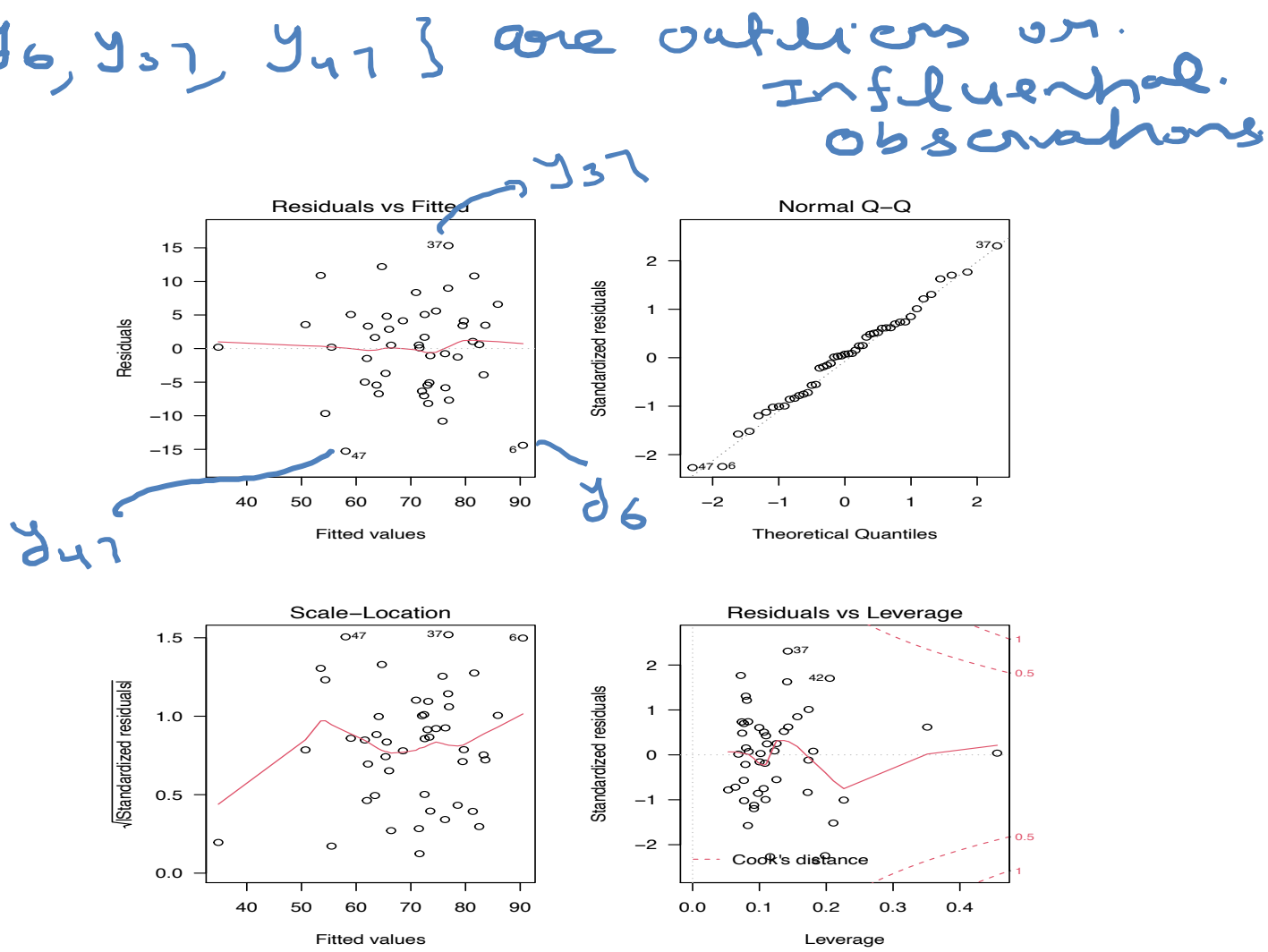
{y6, y37, y47} are outliers or influential observations
y37
y47
y6
Residuals vs Fitted
Residuals
Fitted values
Normal Q–Q
Standardized residuals
Theoretical Quantiles
Scale–Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Figure 2: