MAT 3375 Summary

Joe Zhang

Fall 2023

1 Introduction

We want to model Y in terms of X. We let $X_1, \ldots, X_p$ be the explanatory variables and Y be the response variable. We want to see how Y changes with $X_1, \ldots, X_p$. The relationship between the explanatory variables and the response variable can also be used for prediction the new value of Y given new value of the explanatory variables. The primary goal in regression is to develop a model that relates the response to the explanatory variables, to test it, and ultimately to use it for inference and prediction.

2 Simple Linear Regression

2.1 The Model

We collect a set of paired data. We plot the n paired data Y_i vs. X_i. If it seems reasonable to fit a straight line to the points, we then postulate the following simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

In the model, ϵ represents an unobserved random error term, β_0 is the intercept, and β_1 is the slope of the line.

Both β_0 and β_1 are labeled parameters. They need to be estimated usually from the observed data.

Alternatively, the model may be expressed in terms of $(X_i - \overline{X})$

$$Y_i = (\beta_0 + \beta_1 \overline{X}) + \beta_1 (X_i - \overline{X}) + \epsilon_i \tag{2}$$

where $\overline{X}$ represents the average of the X_i.

The proposed model is linear in the parameters β_0 and β_1.

The model would still be referred to as linear if instead we had X_i^2 instead of X_i.
(i.e. The model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$ is still linear in the parameters).

2.2 Model Assumptions

We assume the following: The random error terms are uncorrelated, have mean equal to 0, and common variance equal to σ^2. This assumption leads to the following:

- $E[Y_i] = \beta_0 + \beta_1 X_i$

- $Var[Y_i] = \sigma^2$

Caution: A well fitting regression model does not imply causation.

2.3 Least Squares Estimates

We define Q as the sum of square errors

$$\begin{aligned} Q &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2 \end{aligned}$$

Then we need to find β_0 and β_1 such that they minimize Q. We do this by differentiating with respect to β_0 and β_1 and then setting the partial derivatives equal to 0. We get that the partial derivatives are:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i] = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i] X_i = 0$$

By rearranging, we get the following equations:

$$\sum_{i=1}^{n} [Y_i] = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} [X_i Y_i] = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

Solving the system of linear equations, we let b_0 and b_1 represent the solutions to β_0 and β_1, respectively. We get

$$b_0 = \overline{Y} - b_1 \overline{X} \tag{3}$$

$$b_1 = \frac{\sum (X_i - \overline{X}) Y_i}{\sum (X_i - \overline{X})^2} \tag{4}$$

We can also express the equation of b_1 as

$$b_1 = \sum_{i=1}^{n} k_i Y_i$$

where $k_i = \frac{(X_i - \overline{X})}{\sum (X_i - \overline{X})^2}$

We have the following properties of the k_i:

- $$\sum k_i = 0$$
- $$\sum k_i X_i = 1$$
- $$\sum k_i^2 = \frac{1}{\sum (X_i - \overline{X})^2}$$

To show the properties, we have that

$$\begin{aligned}
\sum k_i &= \frac{\sum (X_i - \overline{X})}{\sum (X_i - \overline{X})^2} \\
&= \frac{(\sum X_i) - n\overline{X}}{\sum (X_i - \overline{X})^2} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\sum k_i X_i &= \frac{\sum (X_i - \overline{X}) X_i}{\sum (X_i - \overline{X})^2} \\
&= \frac{\sum X_i^2 - \overline{X} \sum X_i}{\sum (X_i - \overline{X})^2} \\
&= \frac{\sum X_i^2 - n\overline{X}}{\sum (X_i - \overline{X})^2} \\
&= 1
\end{aligned}$$

$$\begin{aligned}
\sum k_i^2 &= \frac{\sum (X_i - \overline{X})^2}{(\sum (X_i - \overline{X})^2)^2} \\
&= \frac{1}{\sum (X_i - \overline{X})^2}
\end{aligned}$$

After finding the least squares estimate for β_0 and β_1, which we denote as b_0 and b_1, respectively, the line that fits the data is:

$$\hat{Y} = b_0 + b_1 X \tag{5}$$

Alternatively, we can also have

$$\begin{aligned}
\hat{Y} &= (b_0 + b_1\overline{X}) + b_1(X - \overline{X}) \\
&= \overline{Y} - b_1\overline{X} + b_1\overline{X} + b_1(X - \overline{X}) \\
&= \overline{Y} + b_1(X - \overline{X})
\end{aligned}$$

It is also important to note that the point $(\overline{X}, \overline{Y})$ is on the line.

We can predict Y using X and the line.

2.4 The Gauss-Markov Theorem

The Gauss-Markov Theorem states that the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

Recall: An estimator is unbiased if its expected value is the value of its parameter.

To show that b_1 is an unbiased estimator of β_1, we need to show that $E[b_1] = \beta_1$

$$\begin{aligned}
E[b_1] &= \sum k_i E[Y_i] \\
&= \beta_0 \sum k_i + \beta_1 \sum k_i X_i \\
&= \beta_0 \cdot 0 + \beta_1 \cdot 1 \\
&= \beta_1
\end{aligned}$$

To show that b_0 is an unbiased estimator of β_0, we need to show that $E[b_0] = \beta_0$

$$\begin{aligned}
E[b_0] &= E[\overline{Y} - b_1\overline{X}] \\
&= E[\overline{Y}] - E[b_1\overline{X}] \\
&= \frac{1}{n}\sum E[Y_i] - \beta_1\overline{X} \\
&= \frac{1}{n}\sum(\beta_0 + \beta_1 X_i) - \beta_1\overline{X} \\
&= \beta_0 + \beta_1\overline{X} - \beta_1\overline{X} \\
&= \beta_0
\end{aligned}$$

Now, we want to show that b_0 and b_1 have minimum variance among all unbiased linear estimators.

Consider an unbiased estimator for β_1, say, $\hat{\beta_1} = \sum c_i Y_i$, it must satisfy

$$\begin{aligned}
\beta_1 &= E[\hat{\beta_1}] \\
&= \sum c_i E[Y_i] \\
&= \sum c_i[\beta_0 + \beta_1 X_i]
\end{aligned}$$

From this, we must have that $\sum c_i = 0$, $\sum c_i X_i = 1$, and $Var[\hat{\beta_1}] = \sigma^2 \sum c_i^2$.

We set $c_i = k_i + d_i$ for arbitrary d_i. Then we get

$$\begin{aligned}
\sum k_i d_i &= \sum k_i(c_i - k_i) \\
&= [\sum c_i \frac{(X_i - \overline{X})}{\sum(X_i - \overline{X})^2}] - \frac{1}{\sum(X_i - \overline{X})^2} \\
&= [\frac{1}{\sum(X_i - \overline{X})^2} - 0] - \frac{1}{\sum(X_i - \overline{X})^2} \\
&= 0
\end{aligned}$$

If we define the vectors $\mathbf{c}^T = [c_1, c_2, ..., c_n]$, $\mathbf{k}^T = [k_1, k_2, ..., k_n]$, and $\mathbf{d}^T = [d_1, d_2, ..., d_n]$, we get that $\mathbf{k}^T\mathbf{d} = 0$. This shows that $\mathbf{k}$ and $\mathbf{d}$ have inner product 0 and are orthogonal vectors.

Since we have $c_i = k_i + d_i$, we get that $\mathbf{c} = \mathbf{k} + \mathbf{d}$. Since $\mathbf{k}$ and $\mathbf{d}$ are orthogonal, we have that by the Pythagorean theorem, $||\mathbf{c}||^2 = ||\mathbf{k}||^2 + ||\mathbf{d}||^2$. Then, we get that

$$Var[\hat{\beta_1}] = \sigma^2(\sum k_i^2 + \sum d_i^2)$$

The variance is minimized when d_i are all 0. Then $\hat{\beta_1} = b_1$ since $c_i = k_i$.

2.5 Summary of estimates

We may write $\hat{Y} = b_0 + b_1 X$ for the estimated or fitted line, $e_i = Y_i - \hat{Y}_i$ for the estimated ith residual, and we estimate the variance σ^2 by

$$\hat{\sigma^2} = \frac{\sum e_i^2}{n-2}$$

This is also known as the mean square error or MSE.

We have

$$b_1 = \sum k_i Y_i$$

$$\begin{aligned} b_0 &= \overline{Y} - b_1 \overline{X} \\ &= \frac{\sum Y_i}{n} - \overline{X} \sum k_i Y_i \\ &= \sum (\frac{1}{n} - k_i \overline{X}) Y_i \end{aligned}$$

We also have the following propeties of the residuals:

- $\sum e_i = 0$
- $\sum X_i e_i = 0$

To prove the properties, we have:

$$\begin{aligned} \sum e_i &= \sum Y_i - \sum [\overline{Y} + b_1(X_i - \overline{X})] \\ &= \sum (Y_i - \overline{Y}) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \sum X_i e_i &= \sum X_i Y_i - \overline{Y} \sum X_i - b_1 \sum X_i (X_i - \overline{X}) \\ &= [\sum X_i Y_i - n\overline{YX}] - \frac{\sum X_i Y_i - n\overline{YX}}{\sum (X_i - \overline{X})^2} \sum X_i (X_i - \overline{X}) \\ &= 0 \end{aligned}$$

2.6 The Geometry of Estimation

We let $\boldsymbol{X} = (X_1, ..., X_n)^T$, $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\hat{\boldsymbol{Y}} = (\hat{Y_1}, \hat{Y_2}, ..., \hat{Y_n})^T$

We let $\boldsymbol{e} = (e_1, e_2, ..., e_n)^T$ and $\boldsymbol{1_n} = (1, 1, ..., 1)$. Then we can find that $(\boldsymbol{X} - \overline{X}\boldsymbol{1_n})\boldsymbol{e} = 0$. From this, we know that the vector $\boldsymbol{e}$ is orthogonal to the vectors $\boldsymbol{1_n}$ and $\boldsymbol{X} - \overline{X}\boldsymbol{1_n}$. Since $\hat{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1_n} + b_1(\boldsymbol{X} - \overline{X}\boldsymbol{1_n})$. From this, we get that $\boldsymbol{e}$ is orthogonal to $\hat{\boldsymbol{Y}}$.

Using this, we get the following result:

$$||\boldsymbol{Y}||^2 = ||\hat{\boldsymbol{Y}}||^2 + ||\boldsymbol{e}||^2$$

Since we have that $\hat{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1_n} + b_1(\boldsymbol{X} - \overline{X}\boldsymbol{1_n})$, we get that

$$\begin{aligned} ||\hat{\boldsymbol{Y}}||^2 &= ||\overline{Y}\boldsymbol{1_n}||^2 + ||b_1((X) - \overline{X}\boldsymbol{1_n})||^2 \\ &= \overline{Y}^2\boldsymbol{1_n}^T\boldsymbol{1_n} + b_1^2\sum(X_i - \overline{X})^2 \end{aligned}$$

Then we get that

$$\sum Y_i^2 = n\overline{Y}^2 + b_1^2\sum(X_i - \overline{X})^2 + \sum(Y_i - \hat{Y_i})^2$$

From that, we get

$$\sum(Y_i - \overline{Y})^2 = b_1^2\sum(X_i - \overline{X})^2 + \sum(Y_i - \hat{Y_i})^2 \tag{6}$$

We call $\sum(Y_i - \overline{Y})^2$ the total sum of squares, $b_1^2\sum(X_i - \overline{X})^2$ the regression sum of squares, and $\sum(Y_i - \hat{Y_i})^2$ the error sum of squares. This can be used for inferences in regression, which we will talk about in the next section.

2.7 Inference in regression

Remark: If we assume that the random errors $\epsilon_i \sim N(0, \sigma^2)$, then we get that the likelihood function is

$$L(\beta_0, \beta_1, \sigma^2) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{\frac{1}{2\sigma^2}\sum\epsilon_i^2}$$

Maximizing this function is equivalent to minimizing $Q = \sum epsilon_i^2$, we get the same results for β_0 and β_1.

We can also obtain an estimate for σ^2. It can be estimated by $MSE = \frac{\sum e_i^2}{n-2}$.

Suppose we have the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, ..., n$. Then we have

- $\frac{b_1 - \beta_1}{s(b_1)} \sim t_{n-2}$ where $s^2(b_1) = \frac{MSE}{\sum(X_i - \overline{X})^2}$
- $\frac{b_0 - \beta_0}{s(b_0)} \sim t_{n-2}$ where $s^2(b_0) = MSE(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2})$
- MSE is an unbiased estimate of σ^2 and $\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$

We can use the properties above to construct confidence intervals for the parameters and test hypotheses. We get that

- $100(1-\alpha)\%$ CI for $\beta_1 : b_1 \pm t_{n-2}(\frac{\alpha}{2})s(b_1)$
- $100(1-\alpha)\%$ CI for $\beta_0 : b_0 \pm t_{n-2}(\frac{\alpha}{2})s(b_0)$

We can also test hypotheses such as $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ using the test statistic $T = \frac{b_1}{s(b_1)} \sim t_{n-2}$.

2.8 Example for regression

We consider the following example on grade point averages at the end of the freshman year (Y) as a function of the ACT test scores (X).

- We plot the data
- We obtain the least squares estimates
- We plot the estimated regression function and estimate Y when $X = 30$

The R code below will complete the actions

```r
data = read.table("/Users/joezhang/Downloads/Grade point average.txt", header = TRUE, sep = '\t')
names(data)
```

```
## [1] "GPA" "ACT"
```

```r
GPA = data$GPA
ACT = data$ACT
fit = lm(GPA~ACT, data = data)
fit
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = data)
##
## Coefficients:
## (Intercept)          ACT
##     2.14596      0.03735
```

The number under (Intercept) is the least squares estimate for β_0 and the number under ACT is the least squares estimate for β_1.

The code below constructs a 95% confidence interval for both β_0 and β_1.

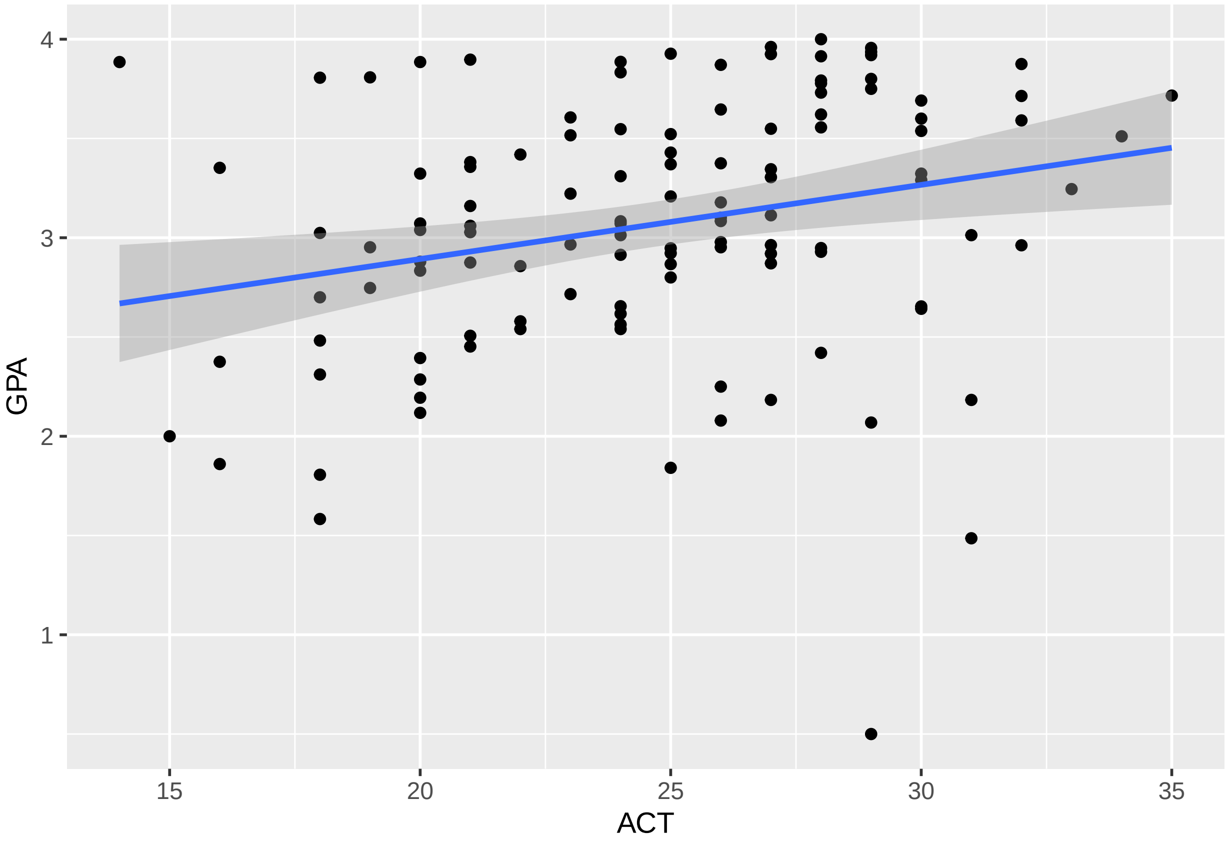
```r
confint(fit, level = 0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 1.5059161 2.786008
## ACT         0.0118145 0.062880
```

The code below plots the data and also constructs a 95% confidence interval and 95% prediction interval for the average of Y.

```
library(ggplot2)
ggplot(data, aes(x = ACT, y = GPA)) +
  geom_point()+
  geom_smooth(method = lm, se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

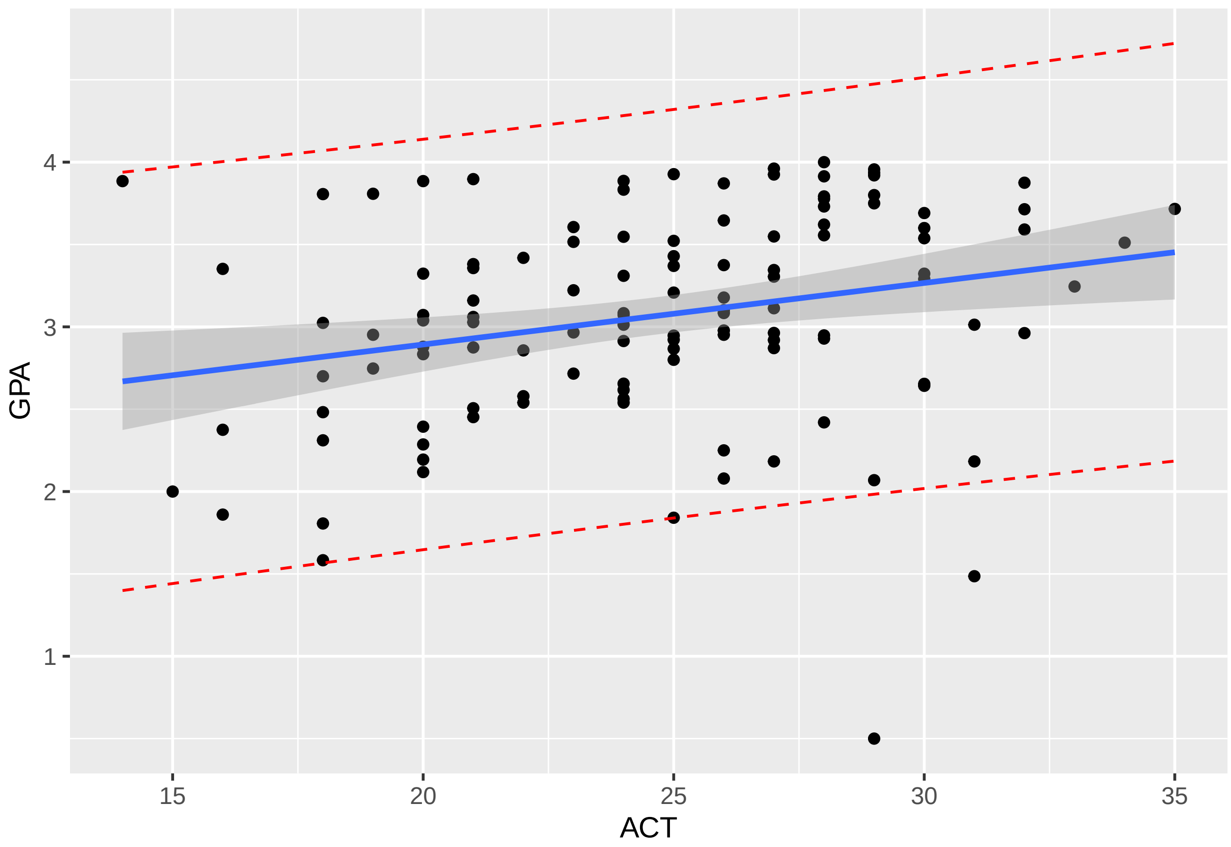


```
temp_var = predict(fit, interval = 'prediction')
```

```
## Warning in predict.lm(fit, interval = "prediction"): predictions on current data refer to _future_ r
```

```
new_df = cbind(data, temp_var)
ggplot(new_df, aes(ACT, GPA))+
  geom_point()+
  geom_line(aes(y = lwr), color = 'red', linetype = 'dashed')+
  geom_line(aes(y = upr), color = 'red', linetype = 'dashed')+
  geom_smooth(method = lm, se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



2.9 Analysis of Variance (ANOVA)

Below is the typical format of an analysis of variance (ANOVA) table (for this part, we use $p = 2$):

Table 1: ANOVA Table

Source	Sum of Squares (SS)	df	Mean Square (MS = SS/df)	F statistic	E[MS]
Regression	$SSR = b_1^2 \sum(X_i - \overline{X})^2$	$p-1$	$MSR = \frac{SSR}{p-1}$	$\frac{MSR}{MSE}$	$\sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-p$	$MSE = \frac{SSE}{n-p}$		σ^2
Total	$SSTO = \sum(Y_i - \overline{Y})^2$	$n-1$			

Each of the sums of squares is a quadratic form where the rank of the corresponding matrix is the degrees of freedom indicated. Cochran's theorem applies and we conclude that the quadratic forms are independent and have Chi-Square distributions. It is well known that the ratio of the two independent Chi-Square divided by their degrees of freedom has a F-distribution (To be seen in section 3 of the notes).

We get that

- $\frac{SSR}{\sigma^2} \sim \chi^2(p-1)$
- $\frac{SSE}{\sigma^2} \sim \chi^2(n-p)$

Then, we get that the F statistic is

$$F = \frac{SSR/(\sigma^2(p-1))}{SSE/(\sigma^2(n-p))} = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F(p-1, n-p)$$

The degrees of freedom are determined by how much data is required to calculate a particular expression.

$\sum(Y_i - \overline{Y})^2$ has $n-1$ degrees of freedom because of the constraints that $\sum(Y_i - \overline{Y}) = 0$

$b_1^2 \sum(X_i - \overline{X})^2$ has one degree of freedom because it is a function of b_1

$\sum(Y_i - \hat{Y}_i)^2$ has $n-2$ degrees of freedom because it is a function of two parameters.

We'll prove all these using matrices in section 3.

2.10 Testing with ANOVA table

We can use the ANOVA table to test the hypotheses $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. The null hypothesis states that the slope of the line is equal to 0. Under the null hypothesis, the expected mean square for regression and the expected mean square error are separate independent estimates of the variance σ^2. Hence, if the null hypothesis is true, the F-ratio should be small. On the other hand, if the alternative hypothesis H_1 is true, then the numerator of the F ration will be expected to be large. Consequently, large values of the F statistic are consistent with the alternative. We reject the null hypothesis for large values of F.

In other words, under the null hypothesis, we have that $E[MSR] = \sigma^2$ and $E[MSE] = \sigma^2$. Then the F ratio $F = \frac{MSR}{MSE}$ would be close to 1. Under the alternative hypothesis, $E[MSE] = \sigma^2$. However, $E[MSR] = \sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$ since $\beta_1 \neq 0$. Therefore, the F ratio is expected to be large. This is why we reject H_0 for large values of the F ratio.

2.11 Back to GPA data

If we consider the GPA data, we can construct an ANOVA table. We do this using R.

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: GPA
##            Df Sum Sq Mean Sq F value Pr(>F)
## ACT         1  3.264  3.2642  8.3917 0.0045 **
## Residuals 117 45.510  0.3890
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

This shows that the F value is large and the p-value is small. We can reject H_0 in this case. This means that there is convincing evidence that the slope is not 0 and there is a relationship between the ACT score and GPA.

Now, we want to construct a 95% confidence interval for β_0 and β_1 for the GPA data using the data summary.

```
summary(fit)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7290 -0.3524  0.0407  0.4362  1.2162
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.14596    0.32318   6.640 1.03e-09 ***
## ACT          0.03735    0.01289   2.897   0.0045 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6237 on 117 degrees of freedom
## Multiple R-squared:  0.06692,    Adjusted R-squared:  0.05895
## F-statistic: 8.392 on 1 and 117 DF,  p-value: 0.0045
```

We get that a confidence interval for β_0 can be calculated the following way:

CI for β_0: $b_0 \pm t_{\alpha/2}117 \cdot s(b_0) = 2.14596 \pm 1.98(0.32318) = (1.5059, 2.7860)$

We get that a confidence interval for β_1 can be calculated the following way:

CI for β_1: $b_1 \pm t_{\alpha/2}117 \cdot s(b_1) = 0.03735 \pm 1.98(0.01289) = (0.01181, 0.06288)$

We can do hypothesis testing using t statistics on both β_0 and β_1.

If we test $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$, we can use the R output and we find that $t = 6.640$, which is significant. We can then reject H_0. Similar with β_1.

However, if we want to test $H_0 : \beta_0 = \beta_{0_1}$ versus $H_1 : \beta_0 \neq \beta_{0_1}$ for some $\beta_{0_1} \neq 0$, then we can't use R. We have to use the test statistic $t = \frac{b_0 - \beta_{0_1}}{s(b_0)} \sim t_{n-2}$ to test and this cannot be computed using R. Similar for β_1.

2.12 Confidence Interval for mean of Y for a given X

We want to construct a confidence interval for the mean of Y^* at a given X^*, or $E[Y^*]$.

To estimate $E[Y^*]$, we know that $E[Y^*] = \beta_0 + \beta_1 X^*$. We can estimate $E[Y^*]$ by

$$\hat{Y}^* = b_0 + b_1 X^* = \sum(\frac{1}{n} + k_i(X^* - \overline{X}))Y_i$$

for a given value of X^*. The estimator is unbiased and has a normal distribution.

We also get that

$$\begin{aligned}
Var[\hat{Y}^*] &= \sigma^2 \sum(\frac{1}{n} + k_i(X^* - \overline{X}))^2 \\
&= \sigma^2 \sum((\frac{1}{n})^2 + k_i^2(X^* - \overline{X}) + 2(\frac{1}{n})k_i(X^* - \overline{X})) \\
&= \sigma^2(\frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2})
\end{aligned}$$

The variance of $\hat{Y}^*$ can be estimated by $s^2[\hat{Y}^*] = MSE(\frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2})$

We can then use the fact that $\frac{\hat{Y}^* - E[Y^*]}{s[\hat{Y}^*]} \sim t_{n-2}$ to make inference on $E[Y]$. We can then construct a $100(1-\alpha)\%$ confidence interval for $E[Y^*]$ by $\hat{Y}^* \pm t_{\alpha/2, t-2} s[\hat{Y}^*]$.

The width of the confidence interval is different at different values of X^*. In fact, the interval is the narrowest at $X^* = \overline{X}$ and gets wider as it deviates from $\overline{X}$.

2.13 Prediction Interval for Y for a given X

For prediction, we want to find a confidence interval for a new value of Y^* for a given X^*.

Note: Alvo's explanations don't make sense. I used the textbook, internet resources, and Boily's notes to make this section. Please let me know if there's anything I need to correct.

We consider the random variable $Y^* - \hat{Y}^*$ for a given X^*. We can use this to make inferences on the predicted value of Y^*.

We have that $E[Y^* - \hat{Y}^*] = 0$. To show this, we have that

$$\begin{aligned}
E[Y^* - \hat{Y}^*] &= E[Y^*] - E[\hat{Y}^*] \\
&= \beta_0 + \beta_1 X^* - E[b_0 + b_1 X^*] \\
&= \beta_0 + \beta_1 X^* - E[b_0] - E[b_1] X^* \\
&= \beta_0 + \beta_1 X^* - \beta_0 - \beta_1 X^* \\
&= 0
\end{aligned}$$

We also have that $Var[Y^* - \hat{Y}^*] = \sigma^2(1 + \frac{1}{n} + \frac{X^* - \overline{X}}{\sum(X_i - \overline{X})^2})$. To show this, we have

$$\begin{aligned}
Var[Y^* - \hat{Y}^*] &= Var[Y^*] + Var[\hat{Y}^*] \\
&= \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(X^* - \overline{X})}{\sum(X_i - \overline{X})}) \\
&= \sigma^2(1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2})
\end{aligned}$$

Then we have that $Y^* - \hat{Y}^* \sim N(0, \sigma^2(1 + \frac{1}{n} + \frac{(X^* - \overline{X})}{\sum(X_i - \overline{X})}))$

We estimate the variance of $Y^* - \hat{Y}^*$ by

$$s^2[Y^* - \hat{Y}^*] = MSE(1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum(X_i - \overline{X})^2})$$

Then we get that

$$\frac{(Y^* - \hat{Y}^*) - 0}{s[Y^* - \hat{Y}^*]} \sim t_{n-2}$$

Then we can construct a prediction interval for Y^*. The prediction interval is $\hat{Y}^* \pm t_{\alpha/2;n-2} s(Y^* - \hat{Y}^*)$

2.14 Example: Airfreight Data

```r
data = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Airfreight Data.txt", header=TRUE, sep =
kable(data)
```

Shipment.Route	Airfreight.breakage
1	16
0	9
2	17
0	12
3	22
1	13
0	8
1	15
2	19
0	11

a. Compute the ANOVA table.
b. Compute confidence intervals for the parameters.
c. Compute a confidence interval for the average response when X = 1.

To compute an ANOVA, table, we simply use the r command

```
x = data$Shipment.Route
y = data$Airfreight.breakage
fit = lm(y~x)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1  160.0   160.0  72.727 2.749e-05 ***
## Residuals  8   17.6     2.2
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We conclude that the regression is highly significant since the F statistic has a value of 72.73.

We now want to compute a confidence interval for the coefficients, we do this using the following R command:

```
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## x             4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

We get that for β_0, a $100(1-\alpha)\%$ confidence interval is $10.2000 \pm t_{\alpha/2,8} \cdot 0.6633$. For β_1, a $100(1-\alpha)\%$ confidence interval is $4.0000 \pm t_{\alpha/2,8} \cdot 0.4690$. In addition, we get that $\hat{\sigma^2} = 2.2$ on 8 degrees of freedom.

To compute a 95% confidence interval for the average response when $X = 1$, we can use the following R commands:

```
new.dat = data.frame(x=1)
predict(fit, newdata=new.dat, interval="confidence")
```

```
##    fit      lwr      upr
## 1 14.2 13.11839 15.28161
```

To compute a 95% prediction interval for Y at $X = 1$, we can use the following R commands:

```
new.dat = data.frame(x=1)
predict(fit, newdata=new.dat, interval='prediction')
```

```
##    fit     lwr     upr
## 1 14.2 10.6127 17.7873
```

2.15 Correlation Coefficient

The sample correlation coefficient is defined the following way:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})}} \tag{7}$$

The correlation coefficient is related to b_1. We can rewrite the equation as

$$r = b_1 (\frac{\sum(X_i - \overline{X})^2}{\sum(Y_i - \overline{Y})^2})^{\frac{1}{2}}$$

The population correlation coefficient is denoted by ρ. It is

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var[X]Var[Y]}}$$

We use r to estimate ρ.

Under $H_0 : \rho = 0$, we have that

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

We can perform a test for ρ using the R command:

```r
x = p2.10$sysbp
y = p2.10$weight
cor.test(x, y, NULL, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 5.9786, df = 24, p-value = 3.591e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5513214 0.8932215
## sample estimates:
##       cor
## 0.7734903
```

If we test $H_0 : \rho = \rho_0$, then we use the following fact to make inference:

$$Z = arctanh(r) = \frac{1}{2} \cdot ln(\frac{1+r}{1-r}) \sim N(arctanh(\rho), \frac{1}{n-3})$$

So if we want to test the hypothesis, we use the test statistic: $Z = (arctanh(r) - arctanh(\rho_0))\sqrt{n-3}$.

We reject H_0 for large values of the test statistic.

To compute a confidence interval of ρ, we use the following formula: $[\tanh(arctanh(r) - z_{\alpha/2}(n-3)^{\frac{1}{2}}), tanh(arctanh(r) + z_{\alpha/2}(n-3)^{\frac{1}{2}})]$

3 Matrix Approach to Regression

3.1 Matrix Notations

If we let $\boldsymbol{Y} = [Y_1, ..., Y_n]^T$ be the transpose of the column data vector, then we define the expectation by $\boldsymbol{E}[\boldsymbol{Y}] = [E[Y_1], ..., E[Y_n]]^T$.

Proposition: If $\boldsymbol{Z} = \boldsymbol{AY} + \boldsymbol{B}$ for some matrix of constants $\boldsymbol{A}$, $\boldsymbol{B}$, then we have $\boldsymbol{E}[\boldsymbol{Z}] = \boldsymbol{AE}[\boldsymbol{Y}] + \boldsymbol{B}$.

To prove this, we let $\boldsymbol{Z} = [Z_1, ..., Z_n]^T$, a_{ij} be the element of the matrix $\boldsymbol{A}$ in the i-th row and j-th column. Let $\boldsymbol{B} = [b_1, ..., b_n]$. Then we get

$$\begin{aligned} E[Z_i] &= E\{[\sum_j a_{ij} Y_j + b_i]\} \\ &= [\sum_j a_{ij} E[Y_j]] + b_i \end{aligned}$$

We define the covariance of $\boldsymbol{Y}$, or the variance-covariance matrix of Y, denoted by $Cov[\boldsymbol{Y}]$, by

$$Cov[\mathbf{Y}] = E\{[\mathbf{Y} - E[\mathbf{Y}]][\mathbf{Y} - E[\mathbf{Y}]]^T\}$$

. We denote this by $\boldsymbol{\Sigma}$

We have the following property of the variance-covariance matrix:

$$Cov[\boldsymbol{AY}] = \boldsymbol{A\Sigma A^T}$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\boldsymbol{Y}$

To prove this, we have that

$$\begin{aligned} Cov[\boldsymbol{AY}] &= E\{[\boldsymbol{AY} - \boldsymbol{E}[\boldsymbol{AY}]][\boldsymbol{AY} - \boldsymbol{E}[\boldsymbol{AY}]]^T\} \\ &= E\{[\boldsymbol{AY} - \boldsymbol{AE}[\mathbf{Y}]][\boldsymbol{AY} - \boldsymbol{AE}[\mathbf{Y}]]^T\} \\ &= E\{[\boldsymbol{A}[\boldsymbol{Y} - \boldsymbol{E}[\mathbf{Y}]][\boldsymbol{Y} - \boldsymbol{E}[\mathbf{Y}]]^{\boldsymbol{T}}\boldsymbol{A^T}\} \\ &= \boldsymbol{A}E\{[\boldsymbol{Y} - \boldsymbol{E}[\mathbf{Y}]][\boldsymbol{Y} - \boldsymbol{E}[\mathbf{Y}]]^{\boldsymbol{T}}\}\boldsymbol{A^T} \\ &= \boldsymbol{A\Sigma A^T} \end{aligned}$$

3.2 Multivariate Normal Distribution

A random vector $\boldsymbol{Y}$ has a multivariate normal distribution if its density is given by

$$f(y_1, \dots, y_n) = \frac{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \cdot exp(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}))$$

where $\boldsymbol{y} = [y_1, ..., y_n]^T$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_n]$, and $\boldsymbol{\Sigma} = Cov[\mathbf{Y}]$. We denote this by $Y \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

If we consider the special case where $n = 1$, we have that $\boldsymbol{\Sigma} = \sigma^2$ and $|\boldsymbol{\Sigma}|^{\frac{1}{2}} = \frac{1}{\sigma}$. Then the density function is

$$f(y_1) = \frac{1}{\sigma\sqrt{2\pi}} \cdot exp(-\frac{1}{2}\frac{(y_1 - \mu_1)^2}{\sigma^2})$$

we get back the univariate normal distribution.

Theorem: Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be an arbitrary $p \times n$ matrix of constants, then we have that

$$\boldsymbol{Z} = AY + B \sim N_p(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A^T})$$

Now, if we consider an example where we let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and we let $\boldsymbol{A} = [1, ..., 1]^T$, then we have that

$$\boldsymbol{AY} \sim N_1(\boldsymbol{A\mu}, A\Sigma A^T)$$

where $\boldsymbol{A\mu} = \sum_{i=1}^n \mu_i$, $\boldsymbol{A\Sigma A^T} = \sum \sigma_j^2 + 2\sum_{i \neq j} \sigma_{ij}$.

3.3 Matrix Approach to Linear Regression

If we use the matrix representation in regression, it makes it easy to generalize to fitting several independent variables. This would go beyond 1 independent variable. This approach is also known as Multiple Linear Regression.

We use vectors and matrices to denote the observations of the independent variables, the dependent variable, the coefficients, and the random term.

- We let $\boldsymbol{Y} = \begin{bmatrix} Y_1 & \dots & Y_n \end{bmatrix}^T$ be the transpose of the column vector of obervations of the dependent variable

- We let $\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \dots & \beta_n\end{bmatrix}^T$ be the transpose of the column vector of coefficients
- We let $\boldsymbol{\epsilon} = \begin{bmatrix}\epsilon_1 & \dots & \epsilon_n\end{bmatrix}^T$ be the transpose of the column vectors of the error terms
- We let $\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,p-1} \\ 1 & X_{2,1} & \dots & X_{2,p-1} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n,1} & \dots & X_{n,p-1} \end{pmatrix}$ be the matrix which incorporates the $p-1$ explanatory variables.

If $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 \boldsymbol{I_n})$, then the regression model may be expressed as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I_n})$$

where $\boldsymbol{I_n}$ is the $n \times n$ identity matrix and N_n is the multivariate normal distribution.

The above is the same as saying that if $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, ..., n$, then we have that

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i \sim N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_{p-1} X_{i,p-1}, \sigma^2)$$

for $i = 1, ..., n$.

The matrix approach is much nicer because it is more compact and it's can compute more values easily.

3.4 Least Squares Estimations

We want to find an estimate for the vector $\boldsymbol{\beta}$. To do this, we use the least squares approach. However, we're no longer using just scalars. We're instead dealing with vectors and matrices. We need formulas to take derivatives. Below are some facts for taking derivatives in matrix notation.

- If $z = \boldsymbol{a}^T \boldsymbol{y}$, then we have $\frac{\partial z}{\partial \boldsymbol{y}} = \boldsymbol{a}$
- If $z = \boldsymbol{y}^T \boldsymbol{y}$, then we have $\frac{\partial z}{\partial \boldsymbol{y}} = 2\boldsymbol{y}$
- If $z = \boldsymbol{a}^T \boldsymbol{A} \boldsymbol{y}$, then we have $\frac{\partial z}{\partial y} = \boldsymbol{A}^T \boldsymbol{a}$
- If $z = \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$, then we have $\frac{\partial z}{\partial \boldsymbol{y}} = \boldsymbol{A}^T \boldsymbol{y} + \boldsymbol{A} \boldsymbol{y}$
- If $z = \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$, and $\boldsymbol{A}$ is symmetric, then we have $\frac{\partial z}{\partial \boldsymbol{y}} = 2\boldsymbol{A}^T \boldsymbol{y}$

Using the derivative formulas above, we can derive the least squares estimate of the vector $\boldsymbol{\beta}$. To do this, we need to minimize the function

$$\begin{aligned} Q &= \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \\ &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= (\boldsymbol{Y} - \boldsymbol{X}\beta)^T (\boldsymbol{Y} - \boldsymbol{X}\beta) \end{aligned}$$

We can differentiate Q and then obtain the estimate for $\boldsymbol{\beta}$. If we differentiate Q, we get

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

. We then set the equation to 0. Then after we solve the equation, we get that a solution for $\boldsymbol{\beta}$ is $\boldsymbol{b} = \boldsymbol{(X^TX)^{-1}X^TY}$.

Therefore, the least squares estimate for $\boldsymbol{\beta}$ is $\boldsymbol{b} = \boldsymbol{(X^TX)^{-1}X^TY}$ if the matrix $(\boldsymbol{X^TX})^{-1}$ exists.

We have that $\boldsymbol{b} = \boldsymbol{(X^TX)^{-1}X^TY}$ is an unbiased estimator of $\boldsymbol{\beta}$. To prove this, we have that

$$\begin{aligned} E[\boldsymbol{b}] &= \boldsymbol{(X^TX)^{-1}X^TE[Y]} \\ &= \boldsymbol{(X^TX)^{-1}X^TX}\beta \\ &= \boldsymbol{\beta} \end{aligned}$$

This means that the least squares estimates of all the parameters are unbiased estimators of their respective parameters.

Now, we want to find the variance-covariance matrix of $\boldsymbol{b}$.

If we let $\boldsymbol{b} = \boldsymbol{AY}$, where $\boldsymbol{A} = \boldsymbol{(X^TX)^{-1}X^T}$. Then we get

$$\begin{aligned} Cov(\boldsymbol{b}) &= \boldsymbol{A\Sigma A} \\ &= \sigma^2\boldsymbol{AA^T} \\ &= \sigma^2(\boldsymbol{(X^TX)^{-1}(X^TX)(X^TX)^{-1}}) \\ &= \sigma^2(\boldsymbol{X^TX})^{-1} \end{aligned}$$

We get that $Cov(\boldsymbol{b}) = \sigma^2(\boldsymbol{X^TX})^{-1}$. We have therefore computed the variances of all the least-squares estimates of the parameters and the covariances between them. This is the nice thing about matrix notation, we can compute more values in one shot.

Now that we have computed the expectation and variance of $\boldsymbol{b}$, we can now determine its distribution. We get that $\boldsymbol{b} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X^TX})^{-1})$.

3.5 The Hat Matrix and its Properties

The predicted value of $\boldsymbol{Y}$ is written as

$$\boldsymbol{\hat{Y}} = \boldsymbol{Xb}$$

. We can rewrite the equation as

$$\boldsymbol{\hat{Y}} = \boldsymbol{HY}$$

where $\boldsymbol{H} = \boldsymbol{X(X^TX)^{-1}X}$. We call $\boldsymbol{H}$ the “hat” matrix.

We have that the hat matrix $\boldsymbol{H}$ is a projection matrix onto the estimation space. It projects $\boldsymbol{Y}$ onto the estimation space, leading to $\boldsymbol{\hat{Y}} = \boldsymbol{HY}$. The hat matrix is also idempotent. To show this, we have that

$$\begin{aligned} \boldsymbol{HH} &= \boldsymbol{X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T} \\ &= \boldsymbol{XI_n(X^TX)^{-1}X^T} \\ &= \boldsymbol{X(X^TX)^{-1}X^T} \\ &= \boldsymbol{H} \end{aligned}$$

The hat matrix is also symmetric, which means that $\boldsymbol{H^T} = \boldsymbol{H}$. To show this, we have

$$\begin{aligned}\boldsymbol{H^T} &= \boldsymbol{(X(X^TX)^{-1}X^T)^T} \\ &= \boldsymbol{X(X^TX)^{-1}X^T} \\ &= \boldsymbol{H}\end{aligned}$$

We also have that the matrix $\boldsymbol{(I-H)}$ is idempotent ($\boldsymbol{I}$ is the identity matrix). To show this, we have

$$\begin{aligned}\boldsymbol{(I-H)(I-H)} &= \boldsymbol{II-IH-HI+HH} \\ &= \boldsymbol{I-H-H+H} \\ &= \boldsymbol{I-H}\end{aligned}$$

We have that the matrix $\boldsymbol{H}$ and the matrix $\boldsymbol{I-H}$ are orthogonal. To show this, we have

$$\begin{aligned}\boldsymbol{H(I-H)} &= \boldsymbol{HI-HH} \\ &= \boldsymbol{H-H} \\ &= \boldsymbol{0}\end{aligned}$$

We can express the residual vector as $\boldsymbol{e}=\boldsymbol{(I-H)Y}$. To show this, we have

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{Y}-\boldsymbol{\hat{Y}} \\ &= \boldsymbol{Y-HY} \\ &= \boldsymbol{(I-H)Y}\end{aligned}$$

Putting all the properties together, we have that $\boldsymbol{\hat{Y}}=\boldsymbol{HY}$, $\boldsymbol{e}=\boldsymbol{(I-H)Y}$, and $\boldsymbol{Y}=\boldsymbol{HY}+\boldsymbol{(I-H)Y}$. We get that by the Pythagorean theorem, we have

$$||\boldsymbol{Y}||^2 = ||\boldsymbol{HY}||^2 + ||\boldsymbol{(I-H)Y}||^2$$

We also get that $Cov[\boldsymbol{e}]=\sigma^2(\boldsymbol{I-H})$, which is estimated by $s^2[\boldsymbol{e}]=MSE(\boldsymbol{I-H})$.

Now, we want to consider the special case where $p=2$. This is the case with 1 predictor variable, which goes back to simple linear regression. We want to compute the hat matrix for this case.

We let $\boldsymbol{X}=\begin{pmatrix}1 & (X_1-\overline{X}) \\ \dots & \dots \\ 1 & (X_n-\overline{X})\end{pmatrix}$. Then we have that $\boldsymbol{X^TX}=\begin{pmatrix}n & 0 \\ 0 & \sum(X_i-\overline{X})^2\end{pmatrix}$. Then we get that $\boldsymbol{(X^TX)^{-1}}=\begin{pmatrix}\sum(X_i-\overline{X})^2 & 0 \\ 0 & n\end{pmatrix}\frac{1}{n\sum(X_i-\overline{X})}$.

Now, we can compute the hat matrix.

$$\begin{aligned}\boldsymbol{H} &= \boldsymbol{(X^TX)^{-1}X^T} \\ &= \begin{pmatrix}\sum(X_i-\overline{X})^2+n(X_1-\overline{X})^2 & \dots & \sum(X_i-\overline{X})^2+n(X_1-\overline{X})^2+n(X_1-\overline{X})(X_n-\overline{X}) \\ \dots & \dots & \dots \\ \sum(X_i-\overline{X})^2+n(X_1-\overline{X})(X_n-\overline{X}) & \dots & \sum(X_i-\overline{X})^2+n(X_n-\overline{X})^2\end{pmatrix}\cdot\frac{1}{n\sum(X_i-\overline{X})^2} \\ &= \begin{pmatrix}\frac{1}{n} & \dots & \frac{1}{n} \\ \dots & \dots & \dots \\ \frac{1}{n} & \dots & \frac{1}{n}\end{pmatrix}+\begin{bmatrix}X_1-\overline{X} \\ \dots \\ X_n-\overline{X}\end{bmatrix}\begin{bmatrix}X_1-\overline{X} & \dots & X_n-\overline{X}\end{bmatrix}\frac{1}{\sum(X_i-\overline{X})^2} \\ &= \frac{1}{n}\boldsymbol{J}+\begin{bmatrix}X_1-\overline{X} \\ \dots \\ X_n-\overline{X}\end{bmatrix}\begin{bmatrix}k_1 & \dots & k_n\end{bmatrix}\end{aligned}$$

Note: $\boldsymbol{J}$ is a matrix of 1s.

Now that we have computed the hat matrix for 2 predictor variables, we can compute the least squares regression line in matrix form.

$$\begin{aligned}
\hat{\boldsymbol{Y}} &= \boldsymbol{HY} \\
&= \frac{1}{n}\boldsymbol{JY} + \begin{bmatrix} X_1 - \overline{X} \\ \dots \\ X_n - \overline{X} \end{bmatrix} \begin{bmatrix} k_1 & \dots & k_n \end{bmatrix} \boldsymbol{Y} \\
&= \begin{bmatrix} \overline{Y} \\ \dots \\ \overline{Y} \end{bmatrix} + \begin{bmatrix} X_1 - \overline{X} \\ \dots \\ X_n - \overline{X} \end{bmatrix} b_1 \\
&= \overline{Y}\boldsymbol{1_n} + b_1 \begin{bmatrix} X_1 - \overline{X} \\ \dots \\ X_n - \overline{X} \end{bmatrix}
\end{aligned}$$

Now, we can find the trace and rank of the hat matrix $\boldsymbol{H}$ and we show that it is equal to 2.

$$\begin{aligned}
Rank(\boldsymbol{H}) &= Trace(\boldsymbol{H}) \\
&= \frac{n\sum(X_i - \overline{X})^2 + n\sum(X_i - \overline{X})^2}{n\sum(X_i - \overline{X})^2} \\
&= 2
\end{aligned}$$

3.6 Quadratic Forms

We now want to look at the theory behind the relationship between sums of squares. We first need to look at a fundamental concept.

If we let $Y_1, ..., Y_n$ be a random sample from $N(\mu, \sigma^2)$. A quadratic form in the Y's is defined to be the real quantity $\boldsymbol{Q} = \boldsymbol{Y^T AY}$, where $\boldsymbol{A}$ is a symmetric positive definite matrix. The singular decomposition of $\boldsymbol{A}$ implies that there exists an orthogonal matrix $\boldsymbol{P}$ such that if $\boldsymbol{\Lambda} = (\lambda_i)$ is the diagonal matrix of eigenvalues of $\boldsymbol{A}$, we have $\boldsymbol{A} = \boldsymbol{P^T \Lambda P}$.

Proportion: $E[\boldsymbol{Y^T AY}] = Trace[\boldsymbol{A\Sigma}] + E[\boldsymbol{Y}]^T \boldsymbol{A} E[\boldsymbol{Y}]$.

To show this, we have

$$\begin{aligned}
\boldsymbol{Y^T AY} &= \boldsymbol{Y^T P^T \Lambda PY} \\
&= \boldsymbol{(PY)^T \Lambda (PY)} \\
&= \sum \lambda_i ||(\boldsymbol{PY})_i||^2
\end{aligned}$$

where $(\boldsymbol{PY})_i$ is the i-th element in the vector PY. The second moment of $(\boldsymbol{PY})_i$ is

$$\begin{aligned}
E[||(\boldsymbol{PY})_i||^2] &= Var[||(\boldsymbol{PY})_i||] + (E[(\boldsymbol{PY})_i])^2 \\
&= (\boldsymbol{P\Sigma P^T})_{ii} + [(\boldsymbol{PE[Y]})_i]^2
\end{aligned}$$

Now, we get

$$\begin{aligned}E[\sum \lambda_i||(PY)_i||^2] &= \sum \lambda_i(\boldsymbol{P\Sigma P^T})_{ii} + \sum \lambda_i[(\boldsymbol{PE}[\mathbf{Y}])_i]^2\\ &= Trace(\Lambda \boldsymbol{P\Sigma P^T}) + \boldsymbol{\mu^T A \mu}\\ &= Trace(\boldsymbol{P^T \Lambda P \Sigma}) + \boldsymbol{\mu^T A \mu}\\ &= Trace(\boldsymbol{A\Sigma}) + \boldsymbol{\mu^T A \mu}\end{aligned}$$

Lemma: The mean squared error is an unbiased estimate of σ^2.

To prove this, we have that the residual sum of squares (SSE) is

$$\sum e_i^2 = \sum (Y_i - \hat{Y})^2$$

This can be written in matrix notation as

$$(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$$

We also know for a fact that $\boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{(I - H)Y}$ and $\boldsymbol{I - H}$ is idempotent. We get that

$$(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y^T(I - H)Y}$$

Then we have that

$$\begin{aligned}E[\boldsymbol{Y^T(I-H)Y}] &= Trace(\boldsymbol{(I-H)\Sigma}) + \boldsymbol{\mu^T(I-H)\mu}\\ &= \sigma^2 Trace(\boldsymbol{I-H}) + (\boldsymbol{X\beta})^T(\boldsymbol{I-H})(\boldsymbol{X\beta})\\ &= \sigma^2(n-p) + \boldsymbol{\beta^T X^T}(\boldsymbol{I - X(X^T X)^{-1} X^T})\boldsymbol{X\beta}\\ &= \sigma^2(n-p) + \boldsymbol{\beta^T(X^T - X^T X(X^T X)^{-1}X^T)X\beta}\\ &= \sigma^2(n-p) + \boldsymbol{\beta^T(X^T - X^T)X\beta}\\ &= \sigma^2(n-p) + 0\\ &= \sigma^2(n-p)\end{aligned}$$

Consequently, we get that

$$\begin{aligned}E[MSE] &= E[\frac{SSE}{n-p}]\\ &= \frac{E[SSE]}{n-p}\\ &= \frac{\sigma^2(n-p)}{n-p}\\ &= \sigma^2\end{aligned}$$

3.7 Chi-Squared distribution and F distribution

A random variable U has a chi-squared χ^2_ν distribution with ν degrees of freedom if its density is given by

$$f(u;\nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\nu/2)} u^{(\nu/2)-1} e^{-u/2}$$

for $u > 0, \nu > 0$. The mean of U is ν and the variance of U is 2ν.

A random variable U has a non-central chi-squared distribution $\chi^2_\nu(\lambda)$ with ν degrees of freedom and non-centrality parameter λ if its density is given by

$$f(u; \nu, \lambda) = \sum_{i=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^i}{i!} f(u; \nu + 2i)$$

with $u > 0, \nu > 0$. The mean of U is $\nu + \lambda$ and the variance of U is $2\nu + 4\lambda$.

If we let $U_1 \sim \chi^2_{\nu_1}$ and $U_2 \sim \chi^2_{\nu_2}$, then we have that

$$F = \frac{U_1/\nu_1}{U_2/\nu_2} \sim F(\nu_1, \nu_2)$$

If the numerator has a non-central chi-squared distribution, then F has a non-central F distribution.

3.8 Cochran's Theorem

Cochran's Theorem states that if we let Y be a random vector with a multivariate normal distribution $N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ and suppose that we have the decomposition

$$\boldsymbol{Y^T Y} = Q_1 + \cdots + Q_k$$

where $Q_i = \boldsymbol{Y^T A_i Y}$ and $rank(A_i) = n_i$. Then $\frac{Q_i}{\sigma^2}$ are independent and have a non-central chi-squared distribution with n_i degrees of freedom and non-centrality parameter λ_i, where $\lambda_i = \boldsymbol{\mu^T A_i \mu}$.

We have some examples of quadratic forms that are particularly important for analysis.

We let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I_n})$ be the response vector. We can decompose $\boldsymbol{Y^T Y}$ the following way:

$$\boldsymbol{Y^T Y} = \boldsymbol{Y^T A Y} + \boldsymbol{Y^T} \frac{\boldsymbol{1_n 1_n^T}}{n} \boldsymbol{Y}$$

where $\boldsymbol{A} = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \cdots & -\frac{1}{n} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1 - \frac{1}{n} \end{pmatrix}$ (an $n \times n$ matrix with $1 - \frac{1}{n}$ on the diagonals and $-\frac{1}{n}$ on the off-diagonals) and $\boldsymbol{1_n} = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix}$ (n-dimensional column vector with all 1s).

We can rewrite $\boldsymbol{Y^T A Y}$ the following way:

$$
\begin{aligned}
\boldsymbol{Y^T A Y} &= \begin{bmatrix} Y_1 & \dots & Y_n \end{bmatrix} \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & \frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \dots & \dots & \dots & \dots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{pmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \\
&= \begin{bmatrix} Y_1-\overline{Y} & Y_2-\overline{Y} & \dots & Y_n-\overline{Y} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \\
&= Y_1(Y_1-\overline{Y}) + Y_2(Y_2-\overline{Y}) + \dots + Y_n(Y_n-\overline{Y}) \\
&= \sum Y_i(Y_i-\overline{Y}) \\
&= \sum Y_i^2 - \overline{Y}\sum Y_i \\
&= \sum Y_i^2 - n\overline{Y} \\
&= \sum (Y_i-\overline{Y})^2
\end{aligned}
$$

We can also rewrite $\boldsymbol{Y}^T \frac{\boldsymbol{1_n^T 1_n}}{n}\boldsymbol{Y}$ the following way:

$$
\begin{aligned}
\boldsymbol{Y}^T \frac{\boldsymbol{1_n^T 1_n}}{n}\boldsymbol{Y} &= \boldsymbol{Y}^T \frac{\boldsymbol{J_n}}{n}\boldsymbol{Y} \\
&= \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \dots & \dots & \dots & \dots \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \\
&= \begin{bmatrix} \overline{Y} & \overline{Y} & \dots & \overline{Y} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \\
&= n\overline{Y}
\end{aligned}
$$

So now we get that $\boldsymbol{Y^T Y} = \sum (Y_i-\overline{Y})^2 + n\overline{Y}$. We can now look at the degrees of freedom of $\sum (Y_i-\overline{Y})^2$ and $n\overline{Y}$.

We get that $\sum (Y_i-\overline{Y})^2 = \boldsymbol{Y^T A Y}$, where $\boldsymbol{A} = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \dots & \dots & \dots & \dots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{pmatrix}$. We know that A is idempotent and symmetric. Then we get that $rank(A) = trace(A) = n(1-\frac{1}{n}) = n-1$. This explains why $\frac{\sum (Y_i-\overline{Y})^2}{\sigma^2}$ has a chi-squared distribution with $n-1$ degrees of freedom.

We also know that $\frac{\boldsymbol{1_n 1_n^T}}{n}$ is an $n \times n$ matrix with $\frac{1}{n}$ as all its entries. This makes it an idempotent and symmetric matrix. It has $rank = trace = n(\frac{1}{n}) = 1$.

Therefore, we get that the ranks sum up to n and we have proven that $\frac{\sum (Y_i-\overline{Y})^2}{\sigma^2}$ has a chi-squared distribution with $n-1$ degrees of freedom.

4 Multiple Linear Regression

4.1 Linear models with 2 or more predictors

We usually see models with 2 or more predictor variables rather than 1 in the case of simple linear regression. For instance, with 2 predictors, we have models in the form

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$. This model displays a plane in 3 dimensions and β_1 represents the rate of change in a unit increase in X_1 when X_2 is fixed and vice versa for β_2.

In a model with $p-1$ predictors, we have that the model is in the form

$$Y_i = \beta_0 + \sum_{i=1}^{p-1} \beta_k X_{ik} + \epsilon_i$$

where β_k is the rate of change in a unit increase in X_k when all other explanatory variables are held fixed.

4.2 Matrix Approach: Review

To make the equation of the linear model more compact, we use the matrix notation discussed in section 2.

- We let $\boldsymbol{Y} = \begin{bmatrix} Y_1 & \dots & Y_n \end{bmatrix}^T$ be the transpose of the column vector of obervations of the dependent variable
- We let $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \dots & \beta_n \end{bmatrix}^T$ be the transpose of the column vector of coefficients
- We let $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & \dots & \epsilon_n \end{bmatrix}^T$ be the transpose of the column vectors of the error terms
- We let $\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,p-1} \\ 1 & X_{2,1} & \dots & X_{2,p-1} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n,1} & \dots & X_{n,p-1} \end{pmatrix}$ be the matrix which incorporates the $p-1$ explanatory variables.

Then we can write the model as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 \boldsymbol{I_n})$.

Recall that the least squares estimator for $\boldsymbol{\beta}$ is $\boldsymbol{b} = \boldsymbol{(X^T X)^{-1} X^T Y}$ and the fitted values are $\boldsymbol{\hat{Y}} = \boldsymbol{Xb} = \boldsymbol{HY}$, where $\boldsymbol{H} = \boldsymbol{X(X^T X)^{-1} X^T}$.

We also recall that the variance-covariance matrix of the residuals is $Cov(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, which can be estimated by $s^2[\boldsymbol{e}] = (MSE)(\boldsymbol{I} - \boldsymbol{H})$. We also have $s^2[\boldsymbol{b}] = (MSE)(\boldsymbol{X^T X})^{-1}$.

We can perform ANOVA on multiple regression models. To do this, it is similar to simple linear regression, except that we need to use matrix notation to write the sums of squares. We have that

- $SSTO = \boldsymbol{Y^T Y} - \frac{1}{n}\boldsymbol{Y^T J Y} = \boldsymbol{Y^T}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$
- $SSE = \boldsymbol{e^T e} = \boldsymbol{Y^T (I - H) Y}$
- $SSR = \boldsymbol{b^T X^T Y} - \frac{1}{n}\boldsymbol{Y^T J Y} = \boldsymbol{Y^T}(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$

These values are all in quadratic form. We get that $SSTO = SSR + SSE$. By Cochran's theorem, we get that SSR, SSE, and SSTO have a chi-squared distribution with degrees of freedom $p-1$, $n-p$, and $n-1$, respectively. This can be shown by computing the ranks of the matrices $\boldsymbol{H} - \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{J}$ and $\boldsymbol{I} - \boldsymbol{H}$. The ANOVA table is the same as for simple linear regression except that SSR has degree of freedom $p-1$ and SSE has degree of freedom $n-p$. This is not restricted to $p = 2$. Below is the typical structure of an ANOVA table (this is the same as for simple linear regression).

Table 3: ANOVA Table

Source	Sum of Squares (SS)	df	Mean Square (MS = SS/df)	F statistic	E[MS]
Regression	$SSR = b_1^2 \sum(X_i - \overline{X})^2$	$p-1$	$MSR = \frac{SSR}{p-1}$	$\frac{MSR}{MSE}$	$\sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-p$	$MSE = \frac{SSE}{n-p}$		σ^2
Total	$SSTO = \sum(Y_i - \overline{Y})^2$	$n-1$			

We can use this to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ against H_1 : not all $\beta_k = 0$. We do this by looking at the F statistic $F = \frac{MSR}{MSE}$ and we reject H_0 for large values of F.

We can also do hypothesis tests for individual coefficients. Say we test $H_0 : \beta_k = 0$ against $H_1 : \beta_k \neq 0$, then we can compute the test statistic

$$t = \frac{b_k}{s[b_k]} \sim t_{n-p}$$

and we reject H_0 for large values of t. ($s^2[b_k] = MSE(\boldsymbol{X^T X})^{-1}_{kk}$).

4.3 Extra Sums of Squares Principle

We can use a more general approach to regression to test if we can fit a reduced model rather than a full model to model the data. We first illustrate the case for $p = 2$.

To do this, we let the full model (F) be the model $Y = \beta_0 + \beta_1 X + \epsilon$ and the reduced model (R) be the model $Y = \beta_0 + \epsilon$. We compute the error sum of squares for each model. We get that $SSE(F) = \sum(Y_i - \hat{Y}_i)^2$ for the full model and $SSE(R) = \sum(Y_i - \hat{Y}_i)^2$ for the reduced model. This way we can test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ by computing the following statistic:

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

We reject H_0 for large values of F^*, which has an F distribution with degrees of freedom $df_R - df_F$ and df_F, respectively. In other words, $F^* \sim F(df_R - df_F, df_F)$

An immediate application of this approach is to the situation where there are repeat observations at the same values of X (i.e. when there are multiple observed Y values at the same X value). Suppose that the full model is given by

$$Y_{ij} = \mu_j + \epsilon_{ij}$$

where $i = 1, \ldots, n_j$ and $j = 1, \ldots, c$ and $\epsilon_{ij} \sum N(0, \sigma^2)$. The μ_j values are unrestricted parameters when $X = X_j$. To derive their least squares estimates, we want to minimize the following: $Q = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \epsilon_{ij}^2 = \sum_{j=1}^{c} \sum_{i=1}^{n_j} [Y_{ij} - \mu_j]^2$. We take the derivative and we get

$$\begin{aligned} \frac{\partial Q}{\partial \mu_j} &= \frac{\partial \sum_{i=1}^{n_j} (Y_{ij} - \mu_j)^2}{\partial \mu_j} \\ &= -2 \sum_{i=1}^{n_j} (Y_{ij} - \mu_j) \end{aligned}$$

We set the above equal to 0. Then we get that

$$\begin{aligned}
-2\sum_{i=1}^{n_j}(Y_{ij} - \mu_j) &= 0 \\
\sum_{i=1}^{n_j} Y_{ij} - \sum_{i=1}^{n_j} \mu_j &= 0 \\
n_j \mu_j &= \sum_{i=1}^{n_j} Y_{ij} \\
\hat{\mu} &= \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j}
\end{aligned}$$

So we have that the least squares estimators of μ_j are

$$\overline{Y_j} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j}$$

Therefore, the error sum of squares for this full unrestricted model is

$$SSE(F) = \sum_{ij}(Y_{ij} - \overline{Y_j})^2$$

The corresponding degrees of freedom are

$$df_F = \sum_{j=1}^{c}(n_j - 1) = n - c$$

If all $n_j = 1$, then $df_F = 0$ and $SSE(F) = 0$, and the analysis cannot proceed any further.

Now, we have the reduced model

$$Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$$

which has error sum of squares equal to

$$SSE(R) = \sum_{ij}(Y_{ij} - \hat{Y_{ij}})^2$$

where $\hat{Y_{ij}} = b_0 + b_1 X_j$. The degrees of freedom are $df_R = (n - 2)$

Now, we can test the hypotheses

$$H_0 : E[Y] = \beta_0 + \beta_1 X$$
$$H_1 : E[Y] \neq \beta_0 + \beta_1 X$$

by computing the ratio

$$F^* = \frac{[\frac{SSE(R) - SSE(F)}{df_R - df_F}]}{[\frac{SSE(F)}{df_F}]}$$

The test is on whether a linear model is justified at all. This is different from just testing that the slope is 0. The main purpose of this test is to see if we can use a linear model instead of a complex model.

We can gain some insight into the components of the F^* ratio. We have that

$$(Y_{ij} - \hat{Y_{ij}}) = (Y_{ij} - \overline{Y_j}) - (\overline{Y_j} - \hat{Y_{ij}})$$

We then get the relationship

$$\sum_{ij}(Y_{ij} - \hat{Y_{ij}})^2 = \sum_{ij}(Y_{ij} - \overline{Y_j})^2 + \sum_{ij}(\overline{Y_j} - \hat{Y_{ij}})^2$$

The components are broken down as follows:

- $SSE(R) = \sum_{ij}(Y_{ij} - \hat{Y_{ij}})^2$ is the error sum of squares for the reduced model
- $SSPE = \sum_{ij}(Y_{ij} - \overline{Y_j})^2$ is the pure error sum of squares
- $SSLF = \sum_{ij}(\overline{Y_j} - \hat{Y_{ij}})^2$ is the error sum of squares due to lack of fit which is independent of i

We also have that the the degrees of freedom of the pure error sum of squares is $df_{PE} = n - c$ and the degrees of freedom of the lack of fit sums of squares is $df_{LF} = c - 2$. An ANOVA table summarizes the analysis:

Table 4: ANOVA Table for Lack of Fit Test

Source	Sum of Squares (SS)	df	Mean Square (MS = SS/df)	F statistic	E[MS]
Regression	$SSR = \sum_{ij}(\hat{Y_{ij}} - \overline{Y})^2$	1	$MSR = SSR$	$\frac{MSR}{MSE}$	$\sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$
Error	$SSE(R) = \sum(Y_{ij} - \hat{Y_{ij}})^2$	$n-2$	$MSE = \frac{SSE(R)}{n-2}$		σ^2
Lack of Fit	$SSLF = \sum_{ij}(\overline{Y_j} - \hat{Y_{ij}})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$	$F^* = \frac{MSLF}{MSPE}$	$\sigma^2 + \frac{\sum n_i(\mu_i - \beta_0 - \beta_1 X_i)^2}{c-2}$
Pure Error	$SSPE = \sum_{ij}(Y_{ij} - \overline{Y_j})^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$		
Total	$SSTO = \sum(Y_i - \overline{Y})^2$	$n-1$			

The F^* ratio tests for lack of fit with a simple linear regression model. If there is no lack of fit, the the ratio should be closer to 1 since both the pure error sums of squares and the error sum of squares due to lack of fit are unbiased estimators of σ^2 under H_0. Otherwise, we would expect the ratio to be large.

We can also extend this concept to multiple linear regression. We consider the case where we have 2 predictors.

We define

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

to be the reduction in the error sum of squares when after X_1 is included, an additional variable X_2 is added to the model. We can rewrite the expression as

$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$$

Similarly, when we have 3 predictors, we have that

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2)$$

This decomposition enables us to judge the effect an added variable has on the sum of squares due to regression.

We can use this process to test a full model with all predictors and a reduced model with only selected predictors by obtaining the error sum of squares of the full and reduced models.

4.4 Example: Delivery Time Data

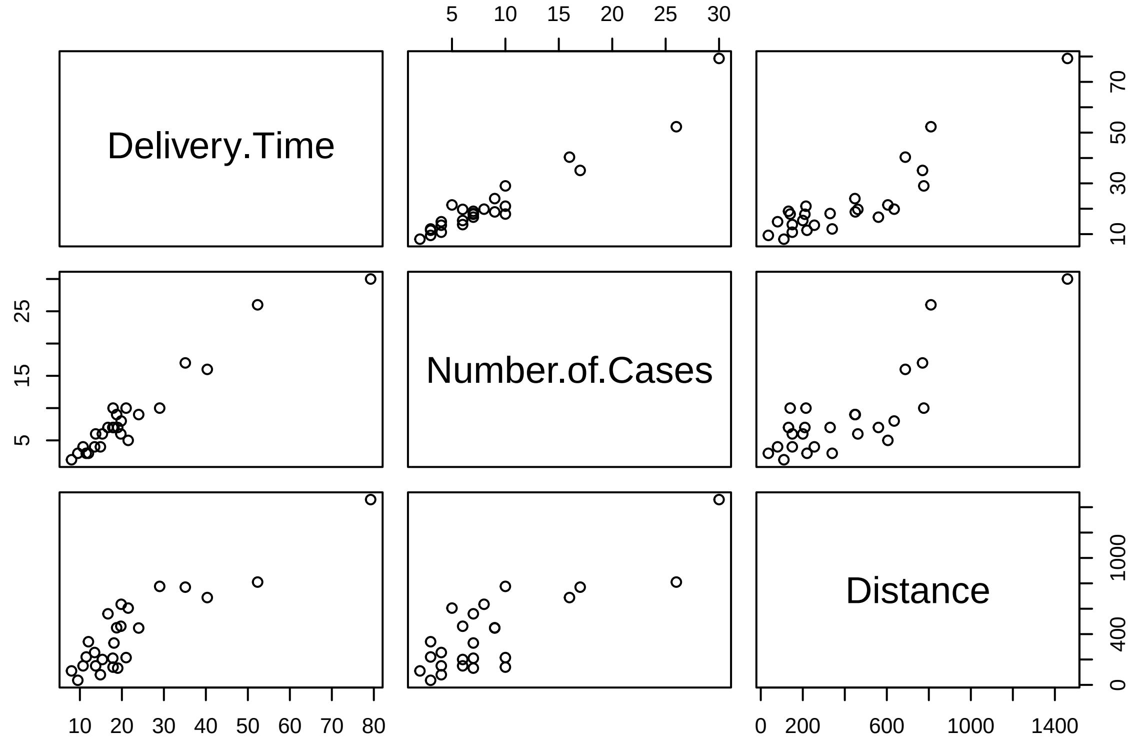
A soft drink bottler is interested in predicting the time required by the route driver to deliver the vending machines in an outlet. We let Y be the delivery time, X_1 be the number of cases of product stocked, and X_2 be the distance walked by the route driver in feet.

We first want to create a scatterplot matrix of the data

```
delivery = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Delivery Time.txt", header =TRUE, sep
names(delivery)
```

```
## [1] "Delivery.Time"   "Number.of.Cases" "Distance"
```

```
plot(delivery)
```



Now, we want to fit a multiple linear regression model for the data.

```
X_1 = delivery$Number.of.Cases
X_2 = delivery$Distance
Y = delivery$Delivery.Time
model = lm(Y~X_1+X_2, data=delivery)
model
```

```
##
## Call:
## lm(formula = Y ~ X_1 + X_2, data = delivery)
```

```
##
## Coefficients:
## (Intercept)          X_1          X_2
##     2.34123      1.61591      0.01438
```

```
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X_1 + X_2, data = delivery)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7880 -0.6629  0.4364  1.1566  7.4197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.341231   1.096730   2.135 0.044170 *
## X_1         1.615907   0.170735   9.464 3.25e-09 ***
## X_2         0.014385   0.003613   3.981 0.000631 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.259 on 22 degrees of freedom
## Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
## F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

From the summary above, we find that the least squares multiple linear regression function is $\hat{Y} = 2.341231 + 1.615907X_1 + 0.014385X_2$. If we want to test for regression, we see that the F statistic is 261.2 and the p-value is very small. This means that we can reject $H_0 : \beta_1 = \beta_2 = 0$, which means that there is enough evidence to suggest that at least one of X_1 and X_2 have influence on Y. If we look at the t statistics, we have that we can reject the null hypotheses $H_0 : \beta_0 = 0$, $H_0 : \beta_1 = 0$, and $H_0 : \beta_2 = 0$. This is because the t statistics are all statistically significant.

We can also do an ANOVA test on the model

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X_1        1 5382.4  5382.4 506.619 < 2.2e-16 ***
## X_2        1  168.4   168.4  15.851 0.0006312 ***
## Residuals 22  233.7    10.6
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We have found that the F statistic to test $H_0 : \beta_1 = 0$ is 506.619 and the F statistic to test $H_0 : \beta_2 = 0$ is 15.851, which are both significant. There is convincing evidence that both X_1 and X_2 have influence on Y.

We can also conduct a test using the extra sum of squares principle. We’re testing the full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ against the reduced model $Y = \beta_0 + \beta_1 X_1$.

```r
Full = lm(Y~X_1+X_2, data=delivery)
Reduced = lm(Y~X_1)
anova(Reduced, Full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X_1
## Model 2: Y ~ X_1 + X_2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     23 402.13
## 2     22 233.73  1     168.4 15.851 0.0006312 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We find that the F statistic is 15.851, which is statistically significant. Therefore, we can reject the null hypothesis that $H_0 : \beta_2 = 0$. This means that we cannot use the reduced model for this data.

4.5 Example: Bank Data

We want to illustrate for testing for lack of fit. For this, we'll use the Bank data. We'll compare the reduced model $Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$ and the full model $Y_{ij} = \mu_j + \epsilon_{ij}$ for a data with repeat observation at the same predictor values.

```r
#We first load the data
bank = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Bank Data.txt", header=TRUE, sep='\t')
names(bank)
```

```
## [1] "Minimum.Deposit"     "Number.New.accounts"
```

```r
#Now, we fit the reduced and full models for the data
x = bank$Minimum.Deposit
y = bank$Number.New.accounts

reduced = lm(y~x)
full = lm(y~0 + as.factor(x))
anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ 0 + as.factor(x)
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1      9 14742
## 2      5  1148  4     13594 14.801 0.005594 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The F statistic is 14.801, which is statistically significant. Therefore, we reject the null hypothesis that $E[Y] = \beta_0 + \beta_1 X$ and a linear model is not a good fit for the data.

4.6 Simultaneous Confidence Intervals

We have learned to construct a confidence interval for one specific parameter (i.e. confidence intervals for β_0 and β_1 in a simple linear regression model). However, sometimes we want to calculate simultaneous or joint confidence intervals for the entire set of parameters. For example, we may want to construct simultaneous confidence intervals for all the coefficients in a linear regression model (i.e. simultaneous confidence intervals that contain both the intercept and slope in a simple linear regression model). However, the confidence level decreases as we include more parameters to estimate.

For example, we consider 2 parameters: β_0 and β_1 of a simple linear regression model, and we want to construct simultaneous $100(1-\alpha/2)\%$ confidence intervals for the parameters. We can obtain a $100(1-\alpha/2)\%$ for each parameter. However, if we let A_1 be the event that β_0 is in its confidence interval and A_2 be the event that β_1 is in its confidence interval, then if we assume that A_1 and A_2 are independent, then we have that

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

Say if we have $P(A_1) = 0.95$ and $P(A_2) = 0.95$, then $P(A \cap B) = (0.95)^2 = 0.9025 < 0.95$. However, the events are never independent, so $P(A_1 \cap A_2)$ is even less.

One strategy is to use the Bonferroni's procedure. Bonferroni's inequality states that for 2 events $\overline{A_1}$, $\overline{A_2}$, we have that

$$P(\overline{A_1} \cap \overline{A_2}) = P(\overline{A_1}) + P(\overline{A_2}) - P(\overline{A_1} \cap \overline{A_2}) \leq P(\overline{A_1}) + P(\overline{A_2})$$

and then we use DeMorgan's identity:

$$P(A_1 \cap A_2) = 1 - P(\overline{A_1} \cup \overline{A_2}) \geq 1 - P(\overline{A_1}) - P(\overline{A_2})$$

We define A_1 as the event that β_0 is contained in its $100(1-\alpha)\%$ confidence interval and A_2 is the event that β_1 is contained in its $100(1-\alpha)\%$ confidence interval. In this case, we have that

$$P(\overline{A_1}) = P(\overline{A_2}) = \alpha$$

and hence, we get that

$$P(A_1 \cap A_2) \geq 1 - P(A_1) - P(A_2) \geq 1 - 2\alpha$$

Now the event $A_1 \cap A_2$ is the event that the intervals $b_0 \pm t(\alpha/2; n-2) \cdot s[b_0]$ and $b_1 \pm t(\alpha/2; n-2) \cdot s[b_1]$ simultaneously cover β_0 and β_1, respectively. The probability of such event is $1 - 2\alpha$. If we have $\alpha = 0.05$, then we get that $1 - 2\alpha = 0.90$. There would be a 0.90 probability that β_0 and β_1 simultaneously fall into the intervals. We would then be 90% confident that the intervals simultaneously cover β_0 and β_1.

On the other hand, if we want to be 95% confident that the intervals simultaneously cover β_0 and β_1, then we need $1-2\alpha = 0.95$, which means we need $\alpha = 0.025$. Which means that we need to compute $t(0.025/2; n-2)$. Then we can construct the simultaneous confidence intervals with the t critical value calculated.

In general, when there's p parameters, the probability that they all fall in their respective confidence intervals is $1 - p\alpha$. If we want to be 95% confident that the intervals simultaneously cover the parameters, then we need that $1 - p\alpha = 0.95$ and $\alpha = 0.05/p$. Then we need to compute $t(0.05/2p; n-p)$, which may not be possible without a computer.

4.7 Example of Simultaneous Confidence Intervals

We examine the rocket propellant data. We do this using R.

```
rocket = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Rocket .txt", header=TRUE, sep='\t')
names(rocket)
```

```
## [1] "Shear.strength"    "Age.of.Propellant"
```

```
y = rocket$Shear.strength
x = rocket$Age.of.Propellant
fit = lm(y~x)
confint(fit, level = 1-0.05/2)
```

```
##                  1.25 %    98.75 %
## (Intercept) 2519.79245 2735.85227
## x             -44.21747  -30.08971
```

This way, we have computed simultaneous 95% confidence intervals for β_0 and β_1. The intervals are $[2519.79245, 2735.85227]$ and $[-44.21747, -30.08971]$ for β_0 and β_1, respectively.

Alternatively, we can compute the critical value for computing the confidence intervals

```
qt(0.9875, nrow(rocket) - 2)
```

```
## [1] 2.445006
```

Then we can use this along with the summary data of the model to find the confidence interval.

```
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215.98  -50.68   28.74   66.61  106.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2627.822     44.184   59.48  < 2e-16 ***
## x            -37.154      2.889  -12.86 1.64e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.11 on 18 degrees of freedom
## Multiple R-squared:  0.9018, Adjusted R-squared:  0.8964
## F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
```

Then the simultaneous confidence intervals are $2627.822 \pm 2.44(44.184)$ for β_0 and $-37.154 \pm 2.44(2.889)$ for β_1.

5 Model Adequacy Checking

In the previous sections, we talked about simple and multiple linear regression. We recall that the assumptions that we make about the model are

- ϵ_i are normally distributed

- $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$
- ϵ_i are independent

We now need to check the assumptions. If the assumptions are not met, then the analysis we do with linear models may not closely reflect the actual model. In this section, we talk about how to check for normality of the error terms and the constancy of variance.

The basic tool that we use to check for model adequacy is analyzing the residuals $e_i = Y_i - \hat{Y}_i$.

5.1 Checking for Normality

To check for normality, we can do this using several ways. We do this by examining the shape of the distribution of the data collected. There are 3 ways in which we can do this:

- We can construct boxplots of residuals. Under the normality assumption, the boxplot should show a symmetric box around the median of approximately 0
- We can construct a histogram of the residuals. It provides a graphical check on normality because if it shows a bell shape centered at 0 approximately, then we can assume a normal distribution
- We can construct a quantile-quantile plot. This plot compares the quantiles of the residual data collected (sample quantiles) with the quantiles from a normal distribution. This is the plot of the ranked residuals against the expected value under normality. We let $e_{(k)}$ be the residual with rank k, and E_k be the expected value of the residual with rank k under normality. We have $E_k = \sqrt{MSE}\Phi^{-1}(\frac{k-0.375}{n+0.25})$ for $k = 1, ..., n$. We plot $e_{(k)}$ against E_k. Under normality, one would expect a straight line pattern in the plot.

5.2 Checking for Constancy of Variance

The other assumption we made about our models is that the variance of the random error terms remain constant. We need to use the residuals to verify this assumption.

We note that the variance of the residuals is $Var[e_i] = \sigma^2(1 - h_{ii})$ and the covariance of residuals is $Cov(e_i, e_j) = \sigma^2(1 - h_{ij})$, where h_{ij} is the element of the hat matrix on the ith row and jth column. If we look at the variance of each residual, we notice that the values are different. This means that the residuals have different variances. So, we want to look at the Studentized or standardized residuals

$$e_i^* = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

where $s^2 = MSE$. The semi-studentized residuals are defined as

$$\frac{e_i}{\sqrt{1 - h_{ii}}}$$

which also has constant variance.

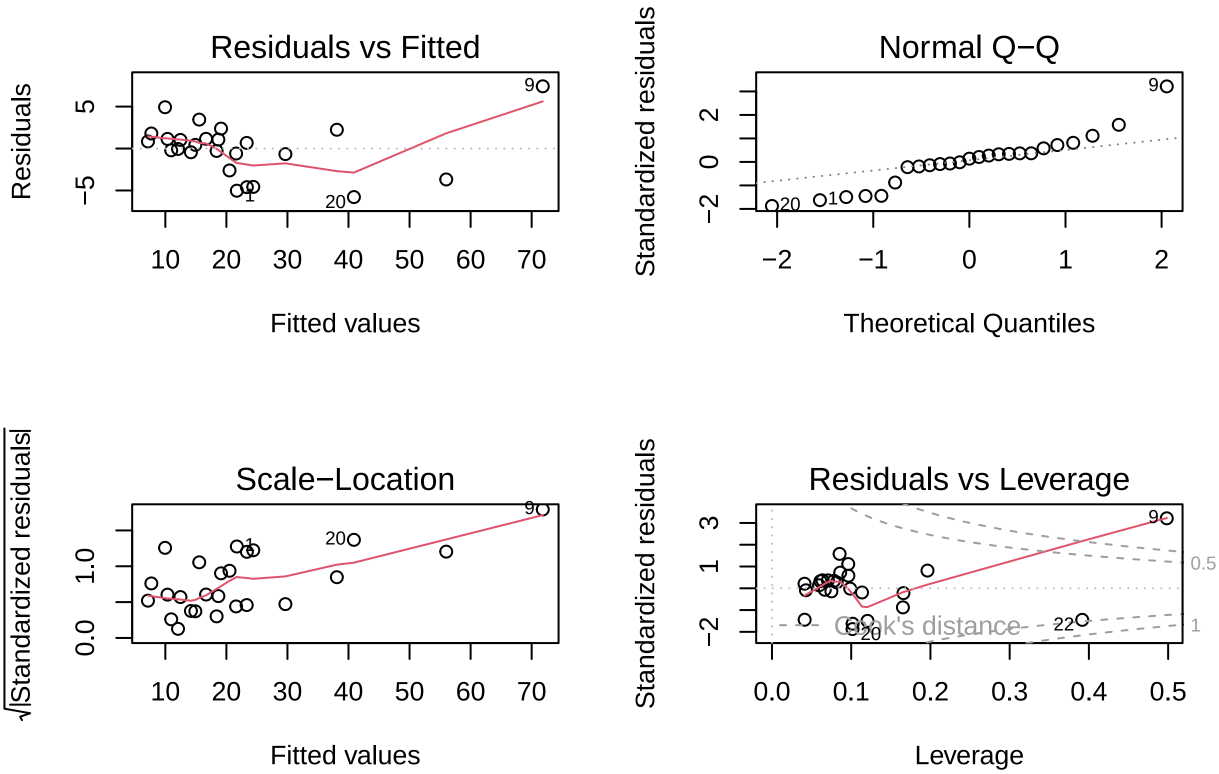
We can check for constancy of variance by making a plot of the Studentized residuals against the fitted values. If the plot shows a random distribution of the points, then we can assume that the variance of the error terms are constant. However, if we see a telescoping increasing or decreasing pattern among the points, then there is evidence of non-constancy of variance. Then the constancy of variance assumption would be violated.

We can also construct a scale-location plot to examine the homogeneity of the variance of the residuals. This plots the square roots of the absolute values of the Studentized residuals vs. the fitted values.

5.3 Example of checking for model adequacy

We look at the delivery time data. We let X_1 represent the delivery number of cases and X_2 represent the delivery distance, and we let Y represent the delivery time.

```
X1 = delivery$Number.of.Cases
X2 = delivery$Distance
Y = delivery$Delivery.Time

fit = lm(Y~X1+X2, data=delivery)
par(mfrow=c(2,2))
plot(fit)
```



From looking at the plots above, we have that the qq-plot follows a straight line pattern in the middle, but not so much on the sides. Therefore, there may be a problem with the normality assumption for this particular dataset. From looking at the residuals vs. fitted plot, we see that the points are randomly distributed on the plot. This means that we can assume constancy of variance in the dataset.

6 Remedial Measures and Transformations

There are times when the assumptions for fitting and testing a linear model are violated, but we don't want to immediately discard the model. Instead, we can do some transformations to the response or predictor variables. This may help linearize the model and bring it in line with the assumptions made.

6.1 Variance Stabilizing Transformations

When the assumption that the variance of the error terms is constant is not reasonable, we can then transform the data to make the assumption more reasonable. Below are some examples of transformations we apply to response variables from various distributions.

6.1.1 Poisson Distribution

In the case that the response variable Y follows a Poisson distribution, then its variance is equal to its mean. If Y is distributed as a Poisson random variable with mean λ, then $\sqrt{Y}$ is distributed more nearly normally with variance approximately $\frac{1}{4}$ if λ is large. In this case, we can regress $\sqrt{Y}$ against X and fit a linear regression model.

6.1.2 Binomial Distribution

In the case that the response variable Y follows a Binomial distribution $B(n, p)$, we have that its mean is $E[Y] = np$. We use the transformation

$$\sin^{-1}\sqrt{\frac{Y+c}{n+2c}}$$

where the optimal value of c is $\frac{3}{8}$ if $E[Y]$ and $n - E[Y]$ are large. The variance is approximately $\frac{1}{4}(n + \frac{1}{2})^{-1}$.

6.2 Transformations to Linearize the Model

Sometimes when we plot Y against X, the plot does not look linear. Then we can apply transformations to the response variable Y to make the plot look more linear. Below are some examples:

6.2.1 Exponential Model

If the true model is

$$Y = \beta_0 e^{\beta_1 X}\epsilon$$

then we can transform it by taking the logarithms. Here is what the model would look like:

$$\ln(Y) = \ln(\beta_0) + \beta_1 X + \ln(\epsilon)$$

We still have to make the usual assumptions for a linear model and then verify them.

6.2.2 Reciprocal Model

The model $Y = \beta_0 + \beta_1 X^{-1} + \epsilon$ can be linearized using the trainsformation $X^* = X^{-1}$. This is a transformation on the predictor variable.

The model $\frac{1}{Y} = \beta_0 + \beta_1 X + \epsilon$ can be linearized using the reciprocal transformation $Y^* = Y^{-1}$. This is a transformation of the response variable.

The model $Y = \frac{X}{\beta_0 + \beta_1 X}$ can be linearized using the reciprocal transformation in 2 steps. First, we use $Y^* = Y^{-1}$ and $X^* = X^{-1}$. Then we have

$$\begin{aligned} Y &= \frac{X}{\beta_0 + \beta_1 X} \\ &= \frac{1}{\beta_0 X^{-1} + \beta_1} \end{aligned}$$

Consequently, we get that $Y^{-1} = \beta_0 X^{-1} + \beta_1$. Then we obtain $Y^* = \beta_1 + \beta_0 X^*$.

6.3 Box-Cox Transformations

The data may not appear to be normally distributed sometimes. Then Box and Cox suggested a power transformation of the type

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda \dot{Y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{Y} \ln(Y) & \lambda = 0 \end{cases}$$

where

$$\dot{Y} = \ln^{-1}[\frac{\sum \ln(Y_i)}{n}]$$

The value of λ can be estimated using trial and error. We fit a model for $Y^{(\lambda)}$ for various values of λ and selecting the one which mimimizes the error sum of squares from a graphical plot. We can also construct a confidence interval for λ. Using R, we can estimate λ using maximum likelihood.

The theory behind the transformation is as follows:

The original model was assumed to be $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$

Then the transformed model is $Y^{(\lambda)}$ and has likelihood function given by

$$\frac{1}{(2\pi)^{n/2}\sigma^n} exp(-\frac{(\boldsymbol{y}^{(\lambda)} - \boldsymbol{X\beta})^T(\boldsymbol{y}^{(\lambda)} - \boldsymbol{X\beta})}{2\sigma^2})J(\lambda; \boldsymbol{y})$$

with parameters $\boldsymbol{\beta}$ and the Jacobian for the transformation

$$J(\lambda; \boldsymbol{y}) = \prod_{i=1}^{n} |\frac{\partial y_i^{(\lambda)}}{\partial y_i}|$$

The maximum likelihood estimator of the variance is given by

$$\hat{\sigma^2} = \frac{\boldsymbol{(Y^{(\lambda)})^T[I - X(X^TX)^{-1}X^T]Y^{(\lambda)}}}{n}$$

The maximum log likelihood for fixed λ is

$$L_{max} = -\frac{n}{2}\log(\hat{\sigma^2}) + \log(J(\lambda; \boldsymbol{y})) = -\frac{n}{2}\log(\hat{\sigma^2}) + (\lambda - 1)\sum \log(y_i)$$

We then plot $L_{max}(\lambda)$ against λ to find the value of λ which yields the maximum.

6.4 Weighted Least Squares

This approach is for when the constancy of variance assumption is not reasonable. We assume that $Var[e_i] = \sigma_i^2$. Instead of minimizing the sum of the squared errors, we can minimize the sum of the weighted squared errors

$$\sum w_i e_i^2$$

where the weights satisfy

$$Var[\sqrt{w_i}e_i] = \sigma^2$$

. In other words, we're to minimize

$$\sum w_i[Y_i - \beta_0 - \beta_1 X_i]$$

with respect to β_0 and β_1.

6.4.1 Estimation and Fitting

With ordinary least squares regression, we have that the linear regression model is

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

We define the matrix $\boldsymbol{W}$ as the matrix of weights

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_n \end{pmatrix}$$

Then the original model goes to

$$\boldsymbol{W^{1/2}Y} = \boldsymbol{W^{1/2}X\beta} + \boldsymbol{W^{1/2}\epsilon}$$

, where

$$\boldsymbol{W}^{1/2} = \begin{pmatrix} w_1^{1/2} & 0 & \dots & 0 \\ 0 & w_2^{1/2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_n^{1/2} \end{pmatrix}$$

Then we get that the least squares estimate for $\boldsymbol{\beta}$ is

$$\begin{aligned} \boldsymbol{b_W} &= ((\boldsymbol{W^{1/2}X})^{\boldsymbol{T}}(\boldsymbol{W^{1/2}X}))^{\boldsymbol{-1}}(\boldsymbol{W^{1/2}X})(\boldsymbol{W^{1/2}Y}) \\ &= (\boldsymbol{X^TW^{1/2}W^{1/2}X})^{\boldsymbol{-1}}\boldsymbol{X^TW^{1/2}W^{1/2}Y} \\ &= (\boldsymbol{X^TWX})^{\boldsymbol{-1}}\boldsymbol{X^TWY} \end{aligned}$$

The MSE becomes

$$\begin{aligned} MSE_W &= \frac{\sum w_i(Y_i - \hat{Y}_i)^2}{n-p} \\ &= \frac{\sum w_i(e_i)^2}{n-p} \end{aligned}$$

6.4.2 Choosing weights

We want to now choose the appropriate weights for weighted least squares. There are many ways we can do this. For example, we can take $w_i = \sqrt{X_i}$ or $w_i = \sqrt{Y_i}$. However, there is one way that we can commonly use. We first divide up the data into clusters using the X variable. Then, we estimate the sample variances of the Y_i for each group, denote them as s_i^2. Then we compute the averages of the X_i values in each cluster. We fit a regression model of the variances of Y_i against the averages of X_i in each cluster. We then estimate the variances by substituting all the X_i values into the equation of the regression function of the variances against the averages. The weights are the inverses of the estimated variances.

For example, we use the turkey data. We group the data into clusters, then we obtain the averages of the X_i in each group and the variances of Y_i in each group. We obtain the regression function $s^2 = 1.5329 - 0.7334\overline{X} + 0.0883\overline{X}^2$. Then we substitute the individual X_i in the data to find $\hat{s_i^2} = 1.5329 - 0.7334X_i + 0.0883X_i^2$. Then we obtain the weights as $\frac{1}{s_i^2}$. With the usual linear regression model, we get that the regression function is

$$Y = -0.579 + 1.14X$$

and the weighted regression is

$$Y = -0.892 + 1.16X$$

6.5 Breusch-Pagan Test

The Breusch-Pagan Test is used for checking for constancy of variance. It assumes the model

$$\log(\sigma_i^2) = \gamma_0 + \gamma_1 X_i$$

. We want to test the hypothesis $H_0 : \gamma_1 = 0$. To do this, we regress the squared residuals e_i^2 against the predictors X_i. Then we can calculate the regression sums of squares of the model SSR^*. Let SSE represent the error sum of squares when we fit a model for Y against X. The test statistic is then

$$\chi^2_{BP} = \frac{\frac{SSR^*}{2}}{\frac{SSE}{n}} \sim \chi^2(1)$$

We reject H_0 for large values of the test statistic.

We can also use R to do the test. For the purpose, we use the Turkey data. Then we get that

```
turkey = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Turkey data.txt", header=TRUE , sep='\
names(turkey)
```

```
## [1] "Age"    "Weight" "Origin" "Z1"     "Z2"
```

```
y <- turkey$Weight
x <- turkey$Age
fit <- lm(y~x)
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 2.5466, df = 1, p-value = 0.1105
```

We obtain a test statistic of 2.5466 and a p-value of 0.11. This means that we fail to reject H_0. We can assume that the variance of error terms are constant.

6.6 Example: Electricity Data

In this example, we look at model adequacy checking and the Box-Cox transformation

We want to analyze the electricity data. We first load the data

```
electric = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Electric Utility Data.txt", header=T
names(electric)
```

```
## [1] "Demand.Y" "Usage.X"
```

Now, we fit a linear regression model and look at the summary and ANOVA table

```
y = electric$Demand.Y
x = electric$Usage.X

model = lm(y~x, data=electric)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 293.62 293.622  109.39 2.731e-14 ***
## Residuals 51 136.89   2.684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x, data = electric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1383 -0.8449 -0.3135  1.1529  3.3221
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6427556  0.4481253  -1.434    0.158
## x            0.0035502  0.0003394  10.459 2.73e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.638 on 51 degrees of freedom
## Multiple R-squared:  0.682,  Adjusted R-squared:  0.6758
## F-statistic: 109.4 on 1 and 51 DF,  p-value: 2.731e-14
```

From looking at the summary table, we find that a simple linear regression model for this data is $\hat{Y} = -0.6427556 + 0.0035502X$, where Y is the electric demand and X is the electric usage. From looking at both the t statistics and the ANOVA table, we can reject the null hypothesis $H_0 : \beta_1 = 0$. There is evidence of relationship between electric usage and demand.

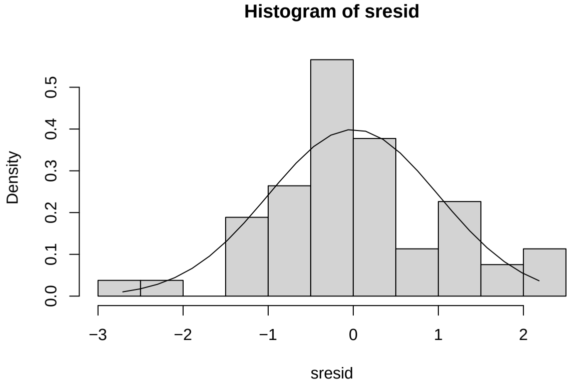
Now, we want to actually verify that the model we fitted actually make sense. We can now do the model adequacy checking. We first create a histogram of the residuals and a normal curve.

```
library(MASS)
```

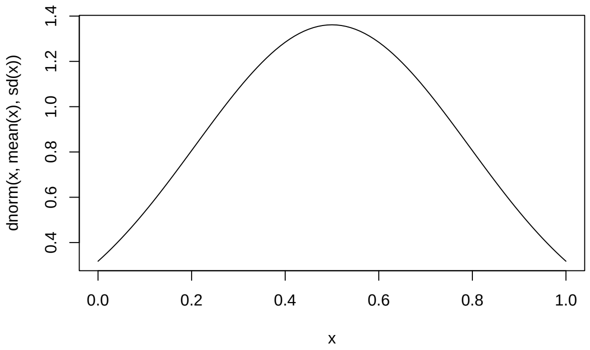
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:MPV':
##
##     cement
```

```
sresid = studres(model)
hist(sresid, freq=FALSE)
smodel=seq(min(sresid),max(sresid),length=25)
ymodel=dnorm(smodel)
lines(smodel, ymodel)
```



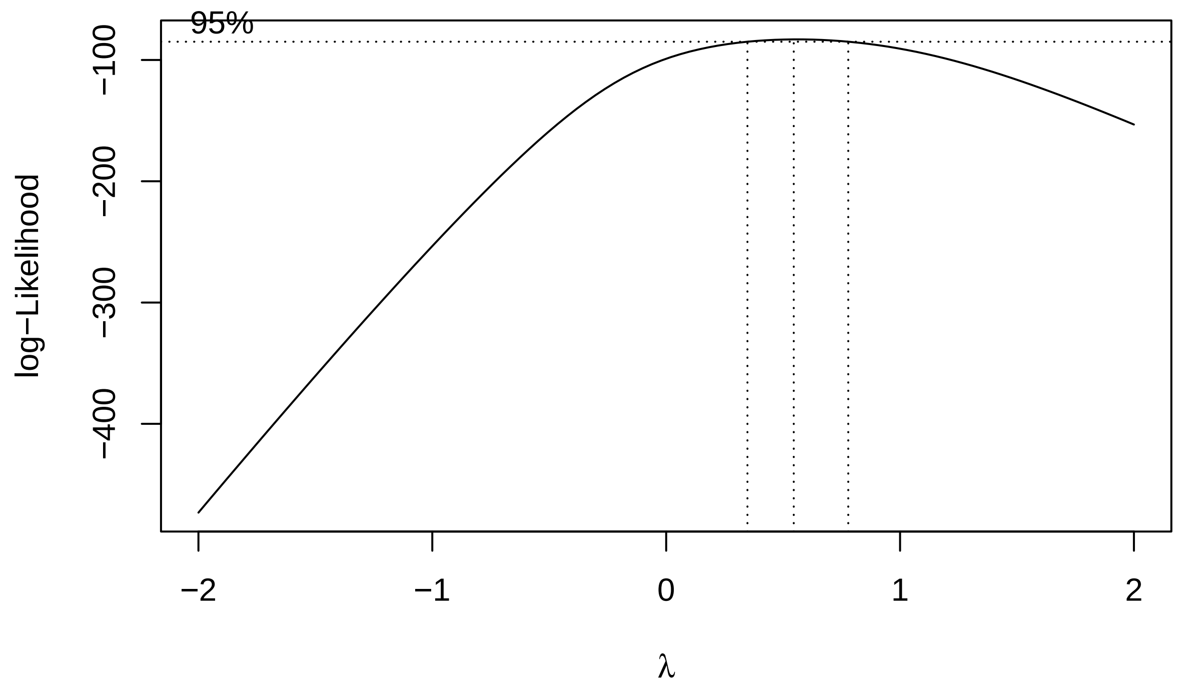
```
curve(dnorm(x, mean(x), sd(x)))
```



If we look at the histogram of the residuals and the normal curve, we see that the residuals are approximately normally distributed with mean 0. This means that we can assume that the random error terms follow a normal distribution.

Usually the Box-Cox transformation is done on data that doesn't seem normal, but we can still try to find a value for λ for Box-Cox transformation.

```
b=boxcox(model)
```



```
b
```

```
## $x
##   [1] -2.00000000 -1.95959596 -1.91919192 -1.87878788 -1.83838384 -1.79797980
##   [7] -1.75757576 -1.71717172 -1.67676768 -1.63636364 -1.59595960 -1.55555556
##  [13] -1.51515152 -1.47474747 -1.43434343 -1.39393939 -1.35353535 -1.31313131
##  [19] -1.27272727 -1.23232323 -1.19191919 -1.15151515 -1.11111111 -1.07070707
##  [25] -1.03030303 -0.98989899 -0.94949495 -0.90909091 -0.86868687 -0.82828283
##  [31] -0.78787879 -0.74747475 -0.70707071 -0.66666667 -0.62626263 -0.58585859
##  [37] -0.54545455 -0.50505051 -0.46464646 -0.42424242 -0.38383838 -0.34343434
##  [43] -0.30303030 -0.26262626 -0.22222222 -0.18181818 -0.14141414 -0.10101010
##  [49] -0.06060606 -0.02020202  0.02020202  0.06060606  0.10101010  0.14141414
##  [55]  0.18181818  0.22222222  0.26262626  0.30303030  0.34343434  0.38383838
##  [61]  0.42424242  0.46464646  0.50505051  0.54545455  0.58585859  0.62626263
##  [67]  0.66666667  0.70707071  0.74747475  0.78787879  0.82828283  0.86868687
##  [73]  0.90909091  0.94949495  0.98989899  1.03030303  1.07070707  1.11111111
##  [79]  1.15151515  1.19191919  1.23232323  1.27272727  1.31313131  1.35353535
##  [85]  1.39393939  1.43434343  1.47474747  1.51515152  1.55555556  1.59595960
##  [91]  1.63636364  1.67676768  1.71717172  1.75757576  1.79797980  1.83838384
##  [97]  1.87878788  1.91919192  1.95959596  2.00000000
##
## $y
##   [1] -473.20774 -463.95766 -454.72933 -445.52373 -436.34182 -427.18453
##   [7] -418.05282 -408.94780 -399.87060 -390.82242 -381.80455 -372.81833
##  [13] -363.86523 -354.94678 -346.06464 -337.22060 -328.41653 -319.65454
##  [19] -310.93681 -302.26578 -293.64412 -285.07467 -276.56072 -268.10571
##  [25] -259.71369 -251.38907 -243.13677 -234.96267 -226.87292 -218.87515
##  [31] -210.97770 -203.19007 -195.52374 -187.99077 -180.60599 -173.38543
##  [37] -166.34684 -159.51058 -152.89750 -146.53023 -140.43201 -134.62508
##  [43] -129.13034 -123.96732 -119.14913 -114.68662 -110.58416 -106.83875
##  [49] -103.44507 -100.38948  -97.65597  -95.22598  -93.07829  -91.19184
##  [55]  -89.54646  -88.12229  -86.90194  -85.87017  -85.01290  -84.31958
##  [61]  -83.78054  -83.38845  -83.13783  -83.02396  -83.04407  -83.19584
##  [67]  -83.47786  -83.88941  -84.42988  -85.09904  -85.89665  -86.82227
##  [73]  -87.87535  -89.05501  -90.35983  -91.78820  -93.33780  -95.00593
##  [79]  -96.78954  -98.68493 -100.68824 -102.79516 -105.00115 -107.30155
##  [85] -109.69148 -112.16606 -114.72043 -117.34976 -120.04932 -122.81458
##  [91] -125.64108 -128.52469 -131.46142 -134.44747 -137.47934 -140.55366
##  [97] -143.66751 -146.81816 -150.00278 -153.21847
```

From the graph above, we find that the optimal value for λ is between 0.5 and 1. We can find the actual value using the following command:

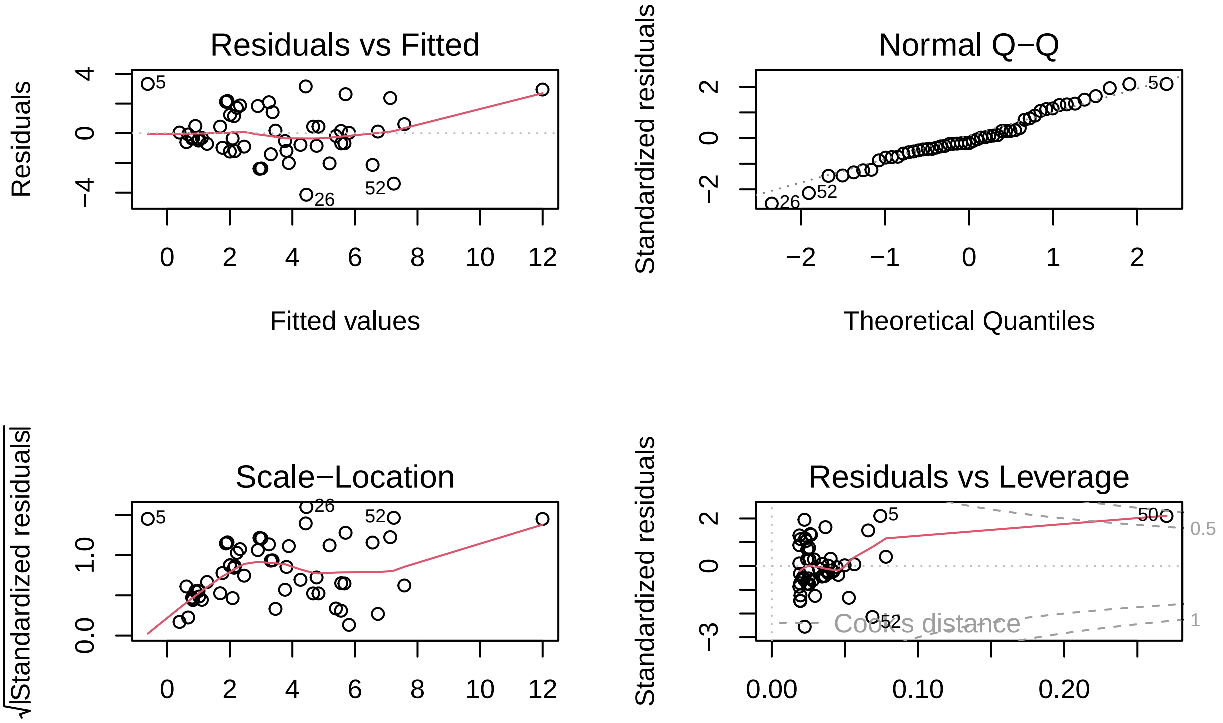
```
b$x[which.max(b$y)]
```

```
## [1] 0.5454545
```

Therefore, the optimal value for λ for the Box-Cox transformation is 0.5454545.

Now, we create diagnostic plots for the data

```
par(mfrow=c(2,2))
plot(model)
```



From looking at the normal probability plot (qq plot), we find that the normality assumption is reasonable. However, from looking at the scale-location plot, there seems to be a pattern among the points. This means that there may be a problem with the constancy of variance assumption.

We try a transformation by taking the square root of the response variable and regressing it against the predictor.

```
model1 = lm(sqrt(y)~x, data=electric)
summary(model1)
```

```
##
## Call:
## lm(formula = sqrt(y) ~ x, data = electric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38819 -0.33269 -0.02776  0.25926  1.01655
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6212411  0.1336872   4.647 2.41e-05 ***
## x           0.0009231  0.0001013   9.116 2.75e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4888 on 51 degrees of freedom
## Multiple R-squared:  0.6197, Adjusted R-squared:  0.6122
## F-statistic:  83.1 on 1 and 51 DF,  p-value: 2.747e-12
```

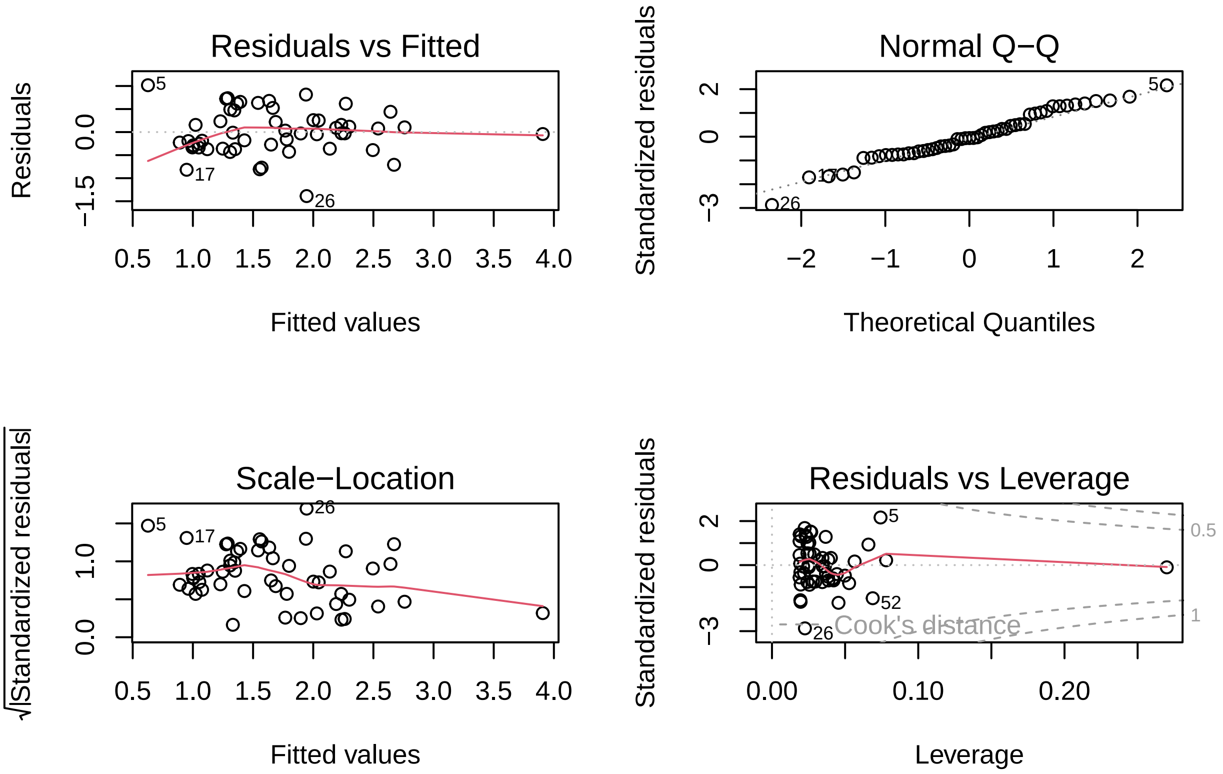
```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: sqrt(y)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 19.851 19.8506  83.097 2.747e-12 ***
## Residuals 51 12.183  0.2389
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the summary and ANOVA table above, we find that the regression model is $\sqrt{\hat{Y}} = 0.06212411 + 0.0009231X$. The t statistics and the ANOVA table all suggest that we can reject $H_0 : \beta_1 = 0$. This means that there is significant evidence that there is a relationship between $\sqrt{Y}$ and X and hence, there is a relationship between Y and X.

We now create the diagnostic plots and see if there is an improvement.

```
par(mfrow=c(2,2))
plot(model1)
```



We see that the residual plot and the scale-location plots now show a more random pattern and the normal qq plot still follows a straight-line pattern. Therefore, the transformation does improve the model.

Now, we want to show the Breusch-Pagan test. We do this using the following command:

```r
library(lmtest)
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 4.1108, df = 1, p-value = 0.04261
```

```r
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 0.869, df = 1, p-value = 0.3512
```

From looking at the Breusch-Pagan results, the original model has a p-value of 0.04261, which means that we reject the null hypothesis that there is no relationship between the variance of the error terms and the predictor. Therefore, there is convincing evidence to suggest that the variance of the error terms is not constant. The transformed model has a p-value of 0.3512, which means that we fail to reject the null hypothesis. Therefore, we can assume that the random error terms have constant variance.

7 Regression Diagnostics for Leverage and Measures of Influence

7.1 Leverages

There are times when a single observation may influence the results of a regression analysis. Hence, the detection of such influential observations is important. For this purpose, the hat matrix plays a very important role.

We begin with the minimized sum of squares:

$$\begin{aligned}
R(\boldsymbol{b}) &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) \\
&= \boldsymbol{Y^T Y} - \boldsymbol{Y^T X (X^T X)^{-1} X^T Y} \\
&= \boldsymbol{Y^T (I - X(X^T X)^{-1} X^T) Y} \\
&= \boldsymbol{Y^T (I - H) Y}
\end{aligned}$$

where $\boldsymbol{H} = \boldsymbol{X(X^T X)^{-1} X^T}$ is the hat matrix.

If we let $\boldsymbol{x_i}^T$ be the i-th row of X. Then the i-th diagonal of $\boldsymbol{H}$ is for $i = 1, ..., n$

$$h_{ii} = \boldsymbol{x_i}^T(\boldsymbol{X^T X})^{-1}\boldsymbol{x_i}$$

In the case of simple linear regression with $p = 2$, we have $\boldsymbol{x_i}^T = [1, X_i]$. Then we get

$$\begin{aligned}
h_{ii} &= \frac{\boldsymbol{x_i}^T \begin{pmatrix} \sum X_j^2 & -\sum X_j \\ -\sum X_j & n \end{pmatrix} \boldsymbol{x_i}}{n\sum X_j^2 - (\sum X_j)^2} \\
&= \frac{\sum X_j^2 - 2X_i \sum X_j + nX_i^2}{n\sum X_j^2 - (\sum X_j)^2} \\
&= \frac{\sum X_j^2 - n\overline{X}^2 + n\overline{X}^2 - 2nX_i\overline{X} + nX_i^2}{n\sum X_j^2 - (\sum X_j)^2} \\
&= \frac{\sum(X_j - \overline{X})^2 + n\overline{X}^2 + 2nX_i\overline{X} + nX_i^2}{n\sum(X_j - \overline{X})^2} \\
&= \frac{\sum(X_j - \overline{X})^2 + n(X_i - \overline{X})^2}{n\sum(X_j - \overline{X})^2} \\
&= \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum(X_j - \overline{X})^2}
\end{aligned}$$

The quantity h_{ii} is called the leverage of the i-th observation.

A further insight is gained by writing the mean $\overline{X}$ in terms of the mean $\overline{X}_{(i)}$ when the i-th observation is deleted. We can show

$$\overline{X} = \frac{\sum X_j}{n} = \frac{X_i + \sum_{j \neq i} X_j}{n} = \frac{X_i + (n-1)\overline{X}_{(i)}}{n}$$

so that

$$X_i - \overline{X} = X_i - \frac{1}{n}(X_i + (n-1)\overline{X}_{(i)}) = \frac{n-1}{n}(X_i - \overline{X}_{(i)})$$

Hence, we get

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum(X_j - \overline{X})^2} = \frac{1}{n} + (\frac{n-1}{n})^2 \frac{(X_i - \overline{X}_{(i)})^2}{\sum(X_j - \overline{X})^2}$$

This shows that the leverage of the i-th observation will be large if X_i is far from the mean of the other observations.

The leverage can be used to flag influential observations. This follows from the fact that

$$\begin{aligned} Trace[\boldsymbol{H}] &= Tr[\boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}] \\ &= Tr[(\boldsymbol{X^T X})^{-1} X^T X] \\ &= Tr[\boldsymbol{I_p}] \\ &= p \end{aligned}$$

Hence, the average $\frac{\sum h_i i}{n} = \frac{p}{n}$

Consequently, observations with a value of the leverage greater than twice the average should be flagged. (i.e. $h_{ii} > 2(\frac{p}{n})$).

7.2 DFFITS

A useful measure of the influence that case i has on the fitted value $\hat{Y}$ is given by

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i(\frac{h_{ii}}{1-h_{ii}})^{1/2}$$

where

$$t_i = e_i(\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2})^{1/2}$$

where we have $\hat{Y}_i$ is the fitted value of Y at X_i, $\hat{Y}_{(i)}$ 0s the fitted value at X_i with the i-th observation removed, $MSE_{(i)}$ is the MSE calculated with teh i-th case removed, ahd h_{ii} is the i-th diagonal of the hat matrix.

This shows that it can be calculated from the residuals, the error sum of squares, and the hat matrix values. The value of $DFFITS_i$ represents the number of estimated standard deviations of $\hat{Y}_i$ that the fitted value increases or decreases with the inclusion of the i-th case in fitting the regression model. If $DFFITS_i$ is high, then it is influential on the fitted value.

If case i is an X outlier and has high leverage, then $(\frac{h_{ii}}{1-h_{ii}})^{1/2} > 1$ and $DFFITS$ will be large in absolute value. As a guideline, influential cases are flagged if

$$|DFFITS_i| > 1$$

for small to medium data sets and

$$|DFFITS_i| > 2\sqrt{\frac{p}{n}}$$

for large data sets. This is a rule of thumb we use to flag influential values using DFFITS.

7.3 Cook's Distance

The next metric we'll be talking about is the Cook's distance. The difference between Cook's distance and DFFITS is that Cook's distance considers the influence of the i-th case on the entire collection of n fitted values instead of just one fitted value. The formula for cook's distance is

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_i - \hat{Y}_{j(i)})^2}{pMSE} = \frac{e_i^2}{pMSE}(\frac{h_{ii}}{(1-h_{ii})^2})$$

Cook's distance is a function of the residual e_i and the leverage h_{ii}. It can be large if either the residual is large and the leverage moderate, or if the residual is moderate and the leverage is large, or both are large. It can be shown that $D_i \sim F(p, n-p)$. Since $F_{0.50}(p, n-p) \approx 1$, we consider points for which $D_i > 1$ to be influential. Ideally, we want the estimated $\hat{\beta}_{(i)}$ to be within the boundary of the 10 - 20% confidence region. In R these regions are indicated in red.

7.4 DFBETAS

DFBETAS are a measure for the influence that case i has on each of the regression coefficients b_k, $k = 0, 1, ..., p-1$.

$$DFBETAS_{(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{ii}}}$$

where c_{ii} is the i-th diagonal element of $(\boldsymbol{X^T X})^{-1}$.

A large value of $DFBETAS_{(i)}$ indicates a large impact of the i-th case on the k-th regression coefficient. As a guideline, we flag the observation as influential if $DFBETAS_{(i)} > \frac{2}{\sqrt{n}}$ for large data sets and $DFBETAS_{(i)} > 1$ for small datasets.

7.5 Example: Body Fat Data Analysis

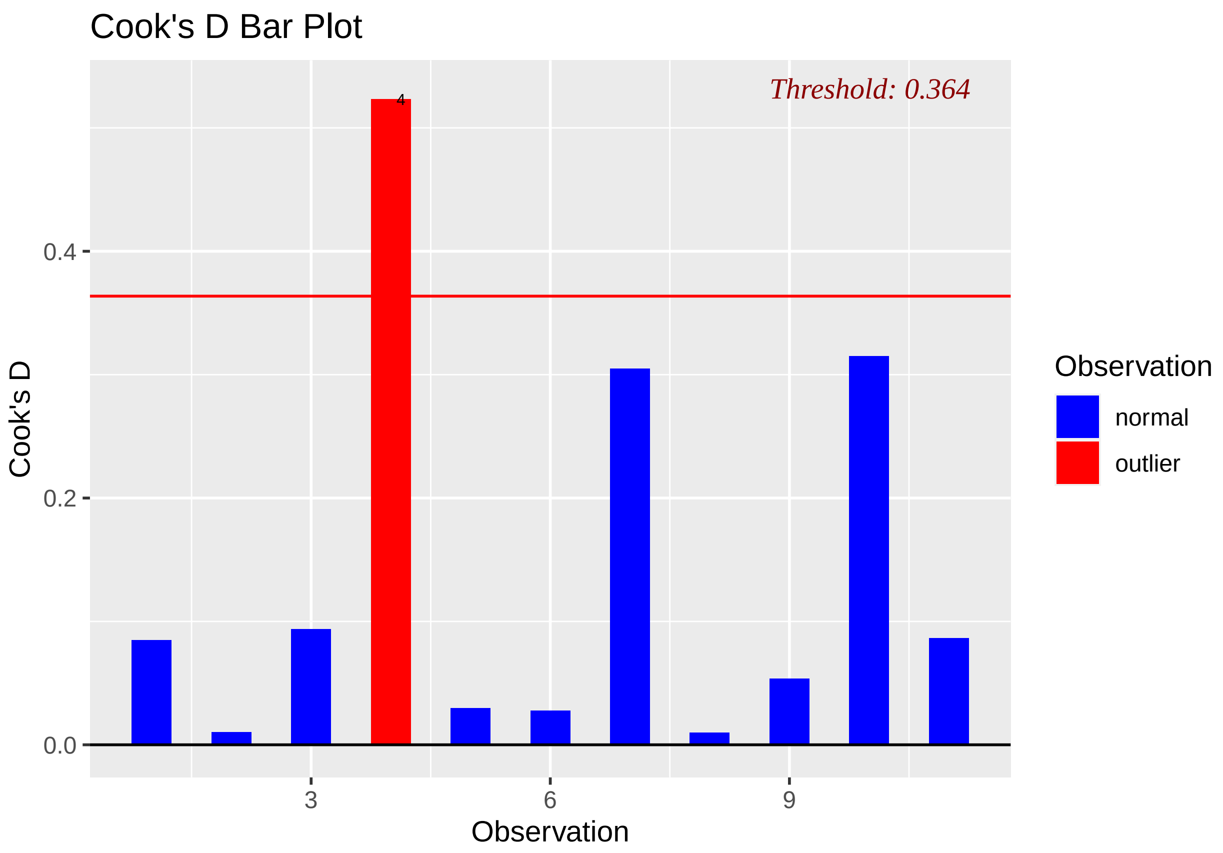
We now look at the Bank data set. We first load the data:

```
bank = read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Bank Data.txt", header=TRUE, sep="\t")
names(bank)
```

```
## [1] "Minimum.Deposit"     "Number.New.accounts"
```

Now, we look at the diagnostic measures on each observation and determine which ones are influential.

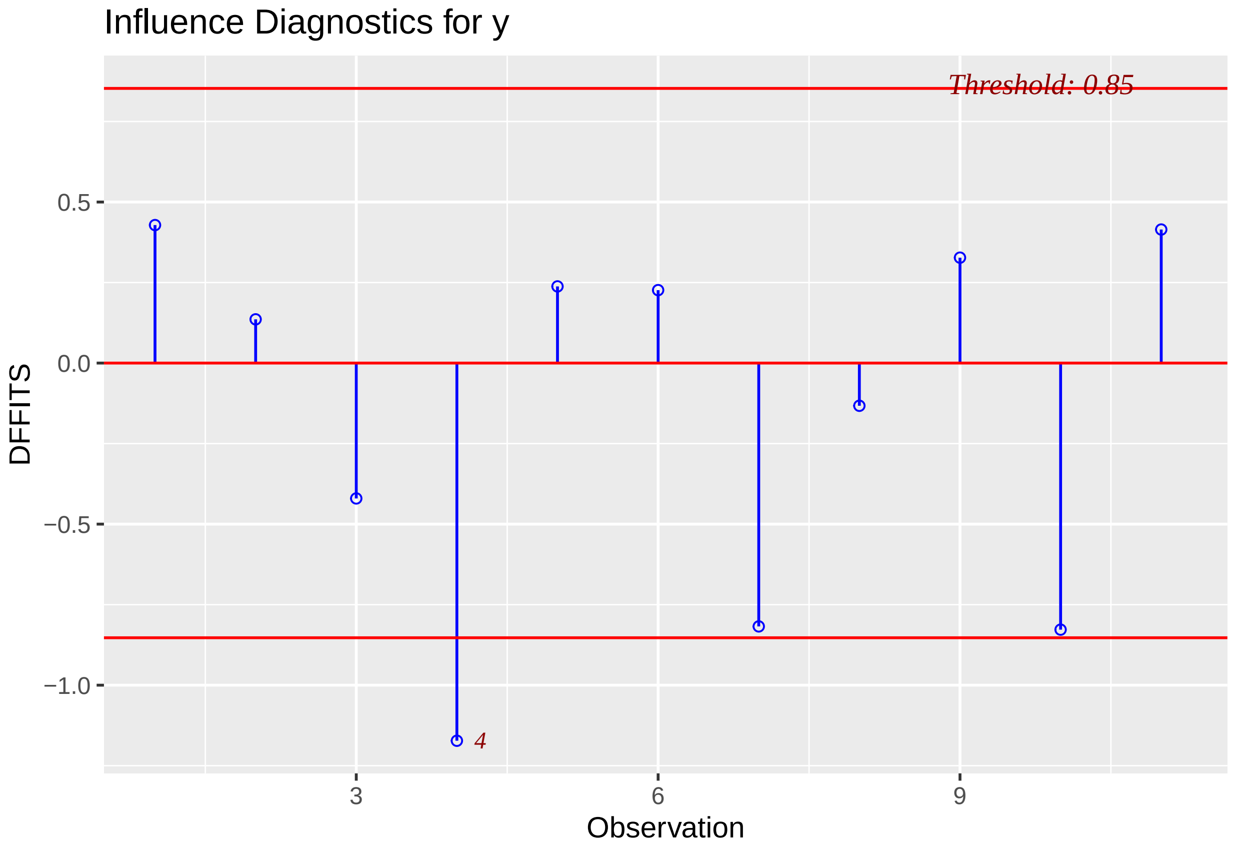
```
y <- bank$Number.New.accounts
x <- bank$Minimum.Deposit

fit <- lm(y~x)

# Cook's Distance vs. Observations
ols_plot_cooksd_bar(fit)
```



From looking at the cook’s distance chart, we find that the 4th observation is influential on the fitted values of the model at all X values.

Now, we construct a plot for DFFITS

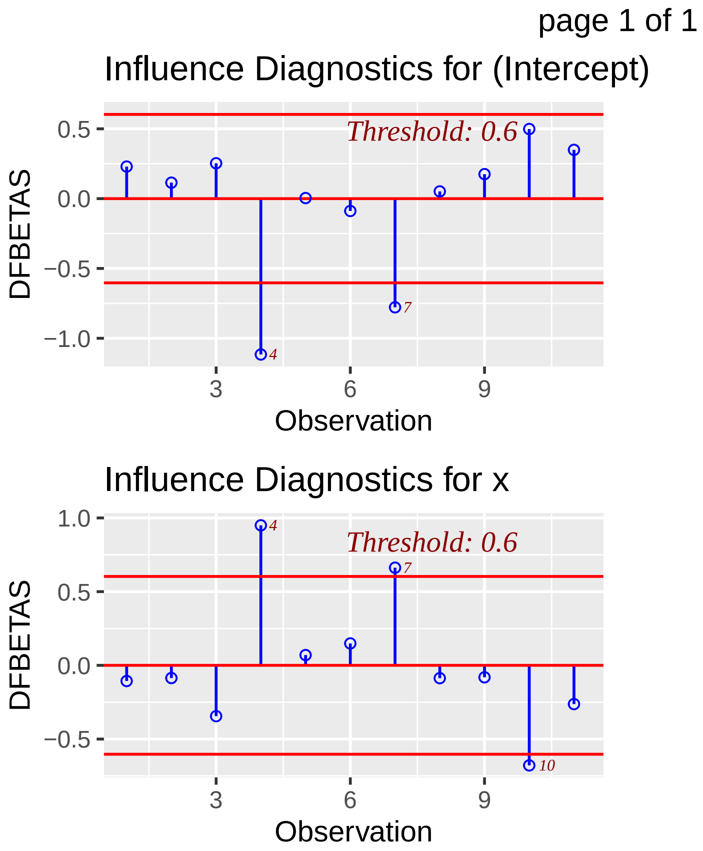
```
ols_plot_dffits(fit)
```



From looking at the plot of DFFITS vs. Observations, we find that the 4th observation is influential on the 4th fitted value of the model.

Now, we construct plots of DFBETAS vs. Observations for each regression coefficient:

```
ols_plot_dfbetas(fit)
```



From the DFBETAS vs. Observations plots, we find that the 4th and 7th observations are influential on the intercept, and the 4th, 7th, and 10th observations are influential on the slope.

We can also look at the diagnostic metrics themselves using the following command:

```r
influence.measures(fit)
```

```
## Influence measures of
##   lm(formula = y ~ x) :
##
##      dfb.1_   dfb.x  dffit cov.r  cook.d    hat inf
## 1   0.22952 -0.1062  0.428 0.951 0.08506 0.0969
## 2   0.11433 -0.0860  0.136 1.454 0.01024 0.1518
## 3   0.25317 -0.3446 -0.420 1.566 0.09395 0.2775
## 4  -1.11603  0.9498 -1.173 0.787 0.52318 0.2644   *
## 5   0.00406  0.0698  0.238 1.241 0.02993 0.0995
## 6  -0.08838  0.1486  0.226 1.409 0.02790 0.1597
## 7  -0.77818  0.6623 -0.818 1.133 0.30507 0.2644
## 8   0.05173 -0.0870 -0.133 1.472 0.00977 0.1597
## 9   0.17533 -0.0811  0.327 1.108 0.05356 0.0969
## 10  0.49853 -0.6786 -0.828 1.171 0.31504 0.2775
## 11  0.34910 -0.2626  0.415 1.189 0.08634 0.1518
```

From looking at the above table, it indicates that the 4th observation is the most influential on the regression model.

8 Different Regression Models

In this section, we talk about different regression models other than the linear regression models discussed in previous sections. We talk about 2 specific models in this section: Polynomial and Indicator Variables

8.1 Polynomial Regression

Sometimes a linear model may not be the appropriate fit for a dataset. A polynomial regression can sometimes be a better fit.

A k-order polynomial regression model in one variable takes the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_k X^k + \epsilon$$

which may be fitted using the matrix approach. It is important to keep in mind that in considering such a model, the order k should be as low as possible. If the value of k gets too high, then the inversion of the matrix $X^T X$ may be inaccurate or not even exist, which would result in poor estimates of the parameters and their variances.

Often, orthogonal polynomials are used in the modeling because they simplify the fitting process:

$$Y_i = \beta_0 P_0(X_i) + \beta_1 P_1(X_i) + \beta_2 P_2(X_i) + \cdots + \beta_k P_k(X_i) + \epsilon_i$$

where P_j is a j order orthogonal polynomial satisfying

$$\sum_{i=1}^{n} P_j(X_i) P_l(X_i) = 0$$

for all $j \neq l$ and we have $P_0(X_i) = 1$.

Such polynomials have been tabulated. The least squares estimates are given by

$$\hat{\beta_j} = \frac{\sum_{i=1}^{n} P_j(X_i) Y_i}{P_j^2(X_i)}$$

for $j = 0, 1, ..., k$.

The principal advantage of using orthogonal polynomials is that the model can be fitted sequentially. This specific advantange is less important today in the age of high speed computing compared to the times when much of the modeling was done using calculators.

When two or more variables are involved, interaction terms are included as in the following model involving two variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon$$

Such models are called response surfaces. They are often used in control theory problems to optimize the selection of control settings of the variables.

8.2 Indicator Regression Models

Regression analysis allows the use of indicator variables which are qualitative or categorical in nature. Such variables are labeled dummy variables. For example, to take into account gender we may define

$$X_2 = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$$

An interesting application is to the case where one wishes to fit a simple linear model as a function of gender. Set

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

In that case, we have

$$Y = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon & \text{male} \\ \beta_0 + \beta_1 X_1 + \epsilon & \text{female} \end{cases}$$

So here the lines are parallel.

This can also be generalized to 2 or more dummy variables.

8.3 Example: Turkey Data Analysis

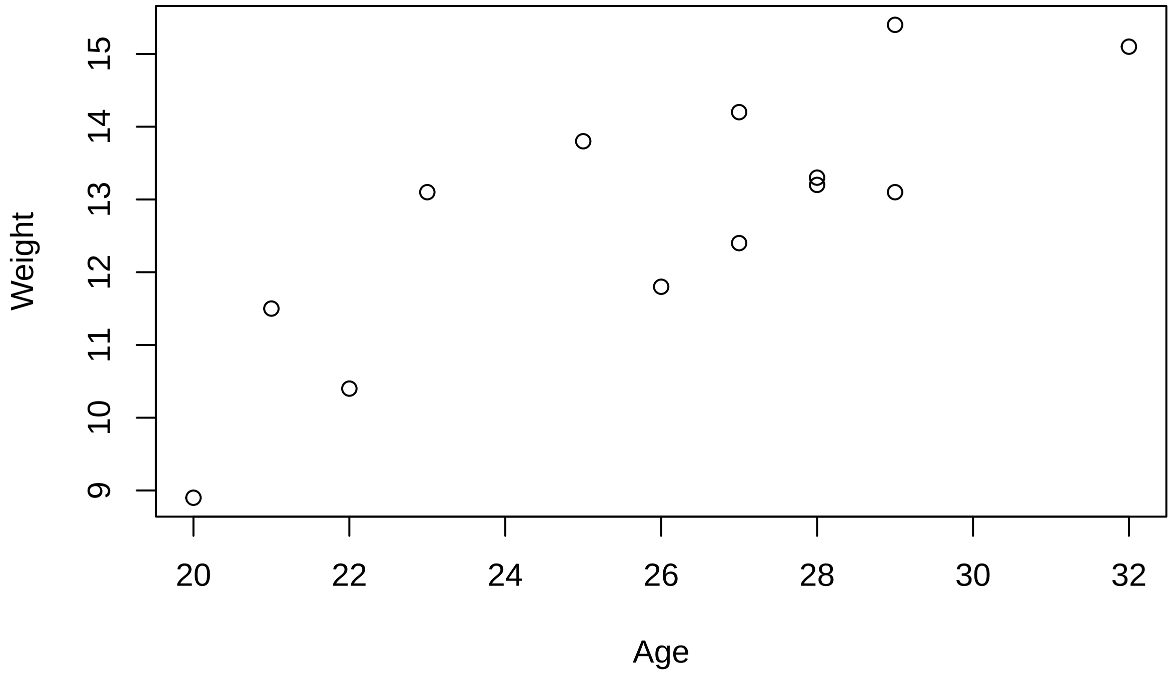
We now consider the turkey data. We first load the data into a dataframe.

```
Turkey <- read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Turkey data.txt", header=TRUE, sep="\
Turkey
```

```
##    Age Weight Origin Z1 Z2
## 1   28   13.3      G  1  0
## 2   20    8.9      G  1  0
## 3   32   15.1      G  1  0
## 4   22   10.4      G  1  0
## 5   29   13.1      V  0  1
## 6   27   12.4      V  0  1
## 7   28   13.2      V  0  1
## 8   26   11.8      V  0  1
## 9   21   11.5      W  0  0
## 10  27   14.2      W  0  0
## 11  29   15.4      W  0  0
## 12  23   13.1      W  0  0
## 13  25   13.8      W  0  0
```

We want to first fit a regression model without considering the origin of the data. We only consider the turkey weights and ages.

```
Age <- Turkey$Age
Weight <- Turkey$Weight
Origin <- Turkey$Origin
Z1 <- Turkey$Z1
Z2 <- Turkey$Z2
plot(Age, Weight)
```



```
model = lm(Weight~Age)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Age        1 26.202 26.2019   21.81 0.0006824 ***
## Residuals 11 13.215  1.2014
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

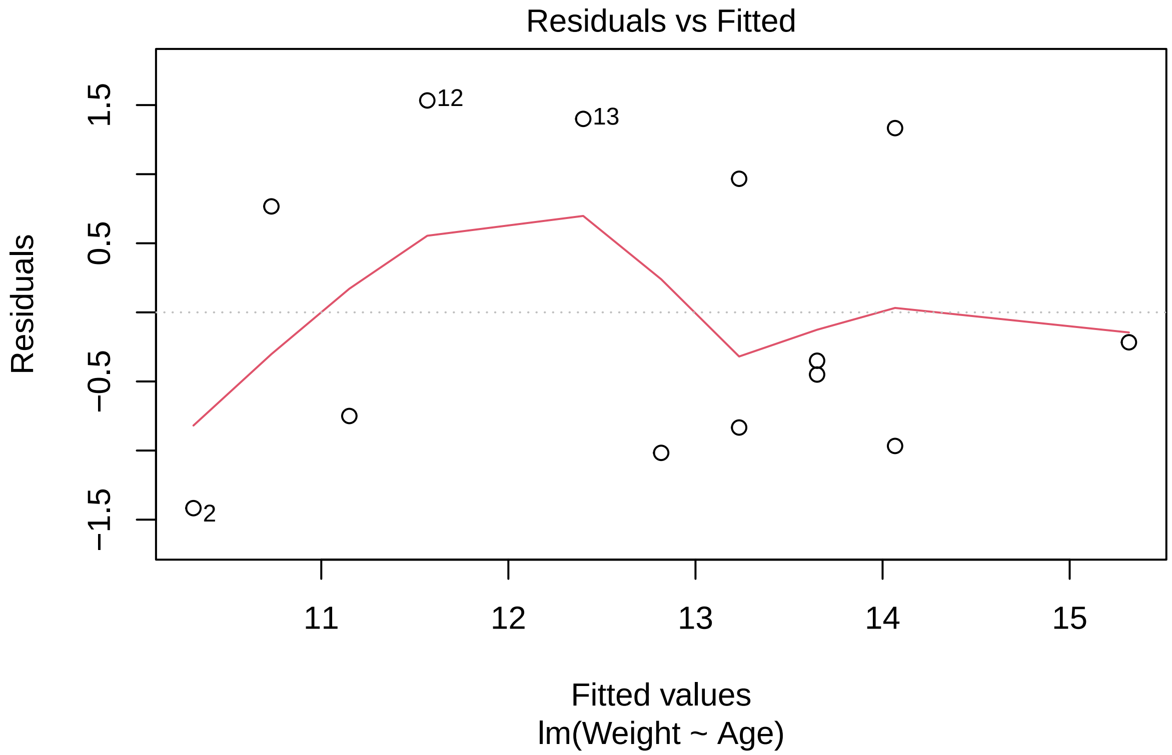
```
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4167 -0.8333 -0.3500  0.9667  1.5333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.98333    2.33273    0.85 0.413327
## Age          0.41667    0.08922    4.67 0.000682 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.096 on 11 degrees of freedom
## Multiple R-squared:  0.6647, Adjusted R-squared:  0.6343
## F-statistic: 21.81 on 1 and 11 DF,  p-value: 0.0006824
```

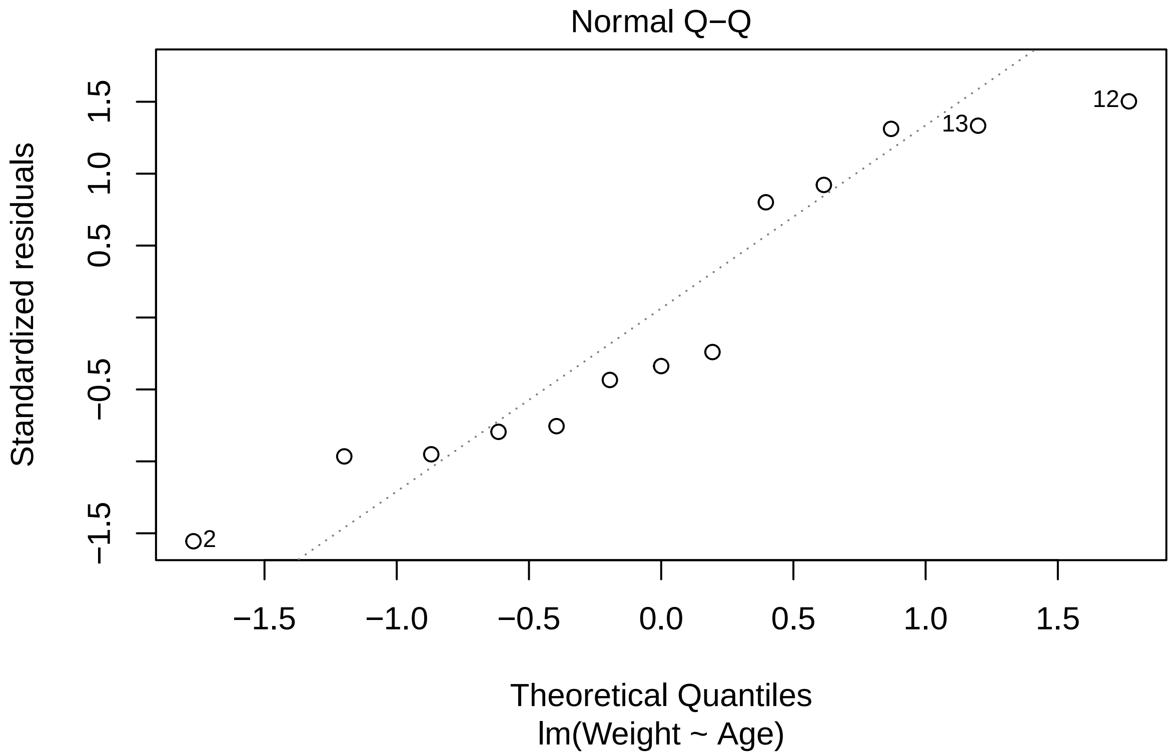
From looking at the summary above, we find that there is a significant relationship between Turkey age and weight as shown by the F statistic in the ANOVA table and the t statistic in the summary table. They both yield low p-values, which means we can reject $H_0 : \beta_1 = 0$. We also find that the R^2 value is 0.6647, which means that around 66.47% of variability is explained by the model.

Now, we create diagnostic plots to check for model adequacy:

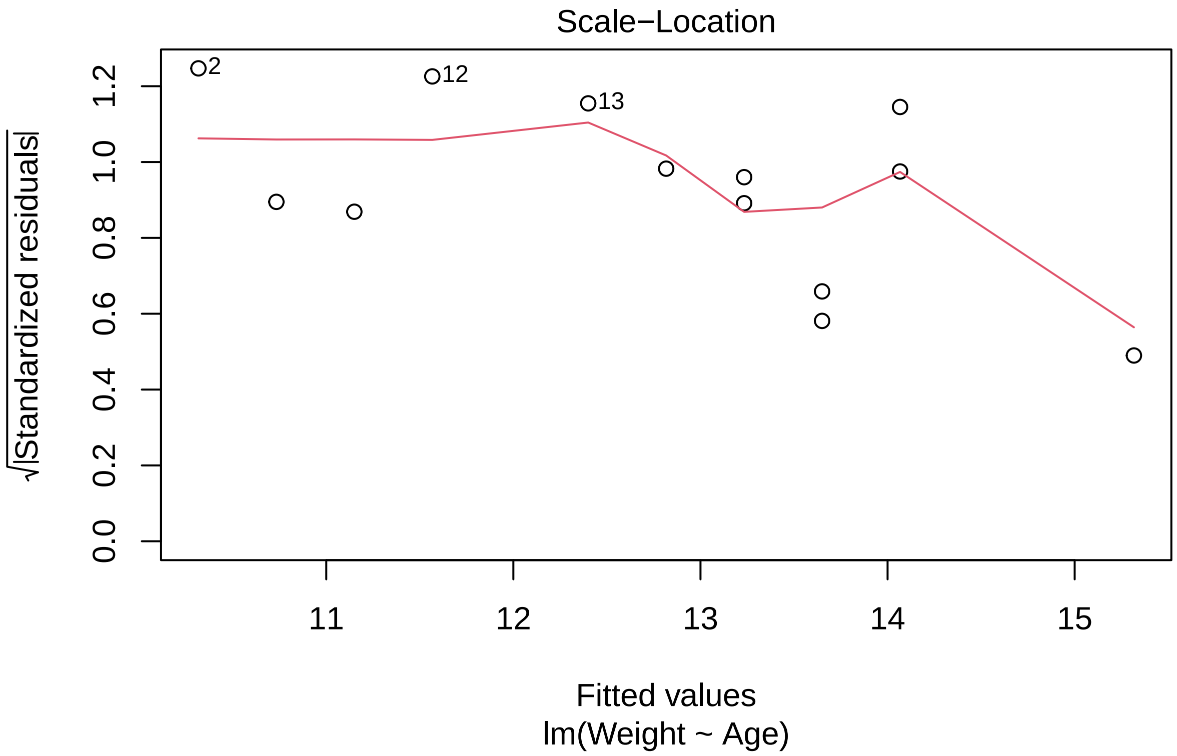
```
Residuals <- model$residuals
plot(model,1)
```



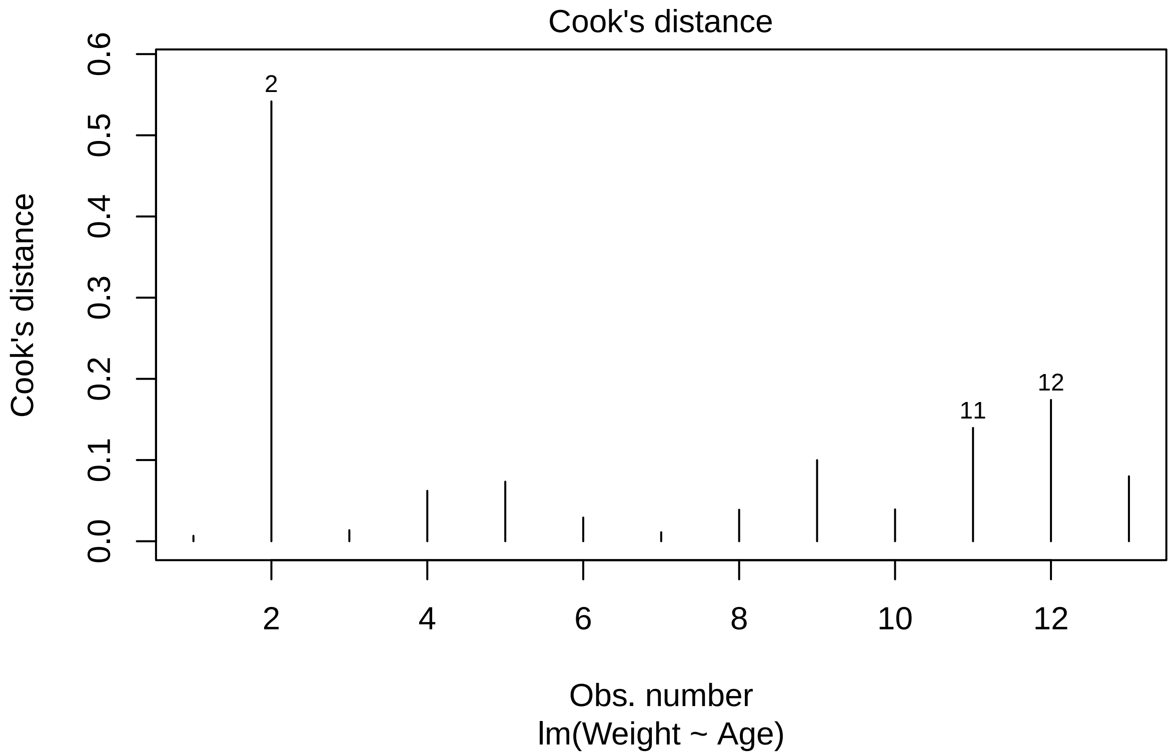
```
plot(model,2)
```


Normal Q−Q
Standardized residuals
−1.5
−0.5
0.5
1.0
1.5
2
13
12
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(Weight ~ Age)

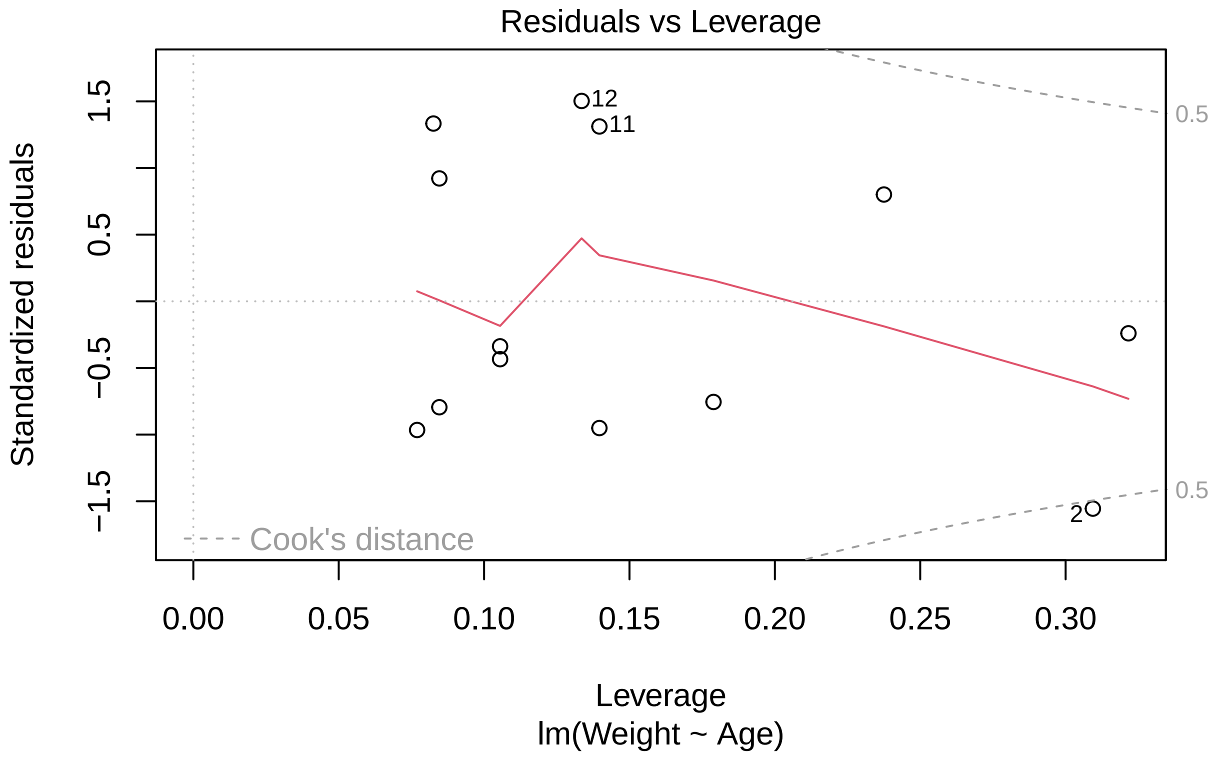
```
plot(model,3)
```


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
11
12
13
14
15
Fitted values
lm(Weight ~ Age)
2
12
13

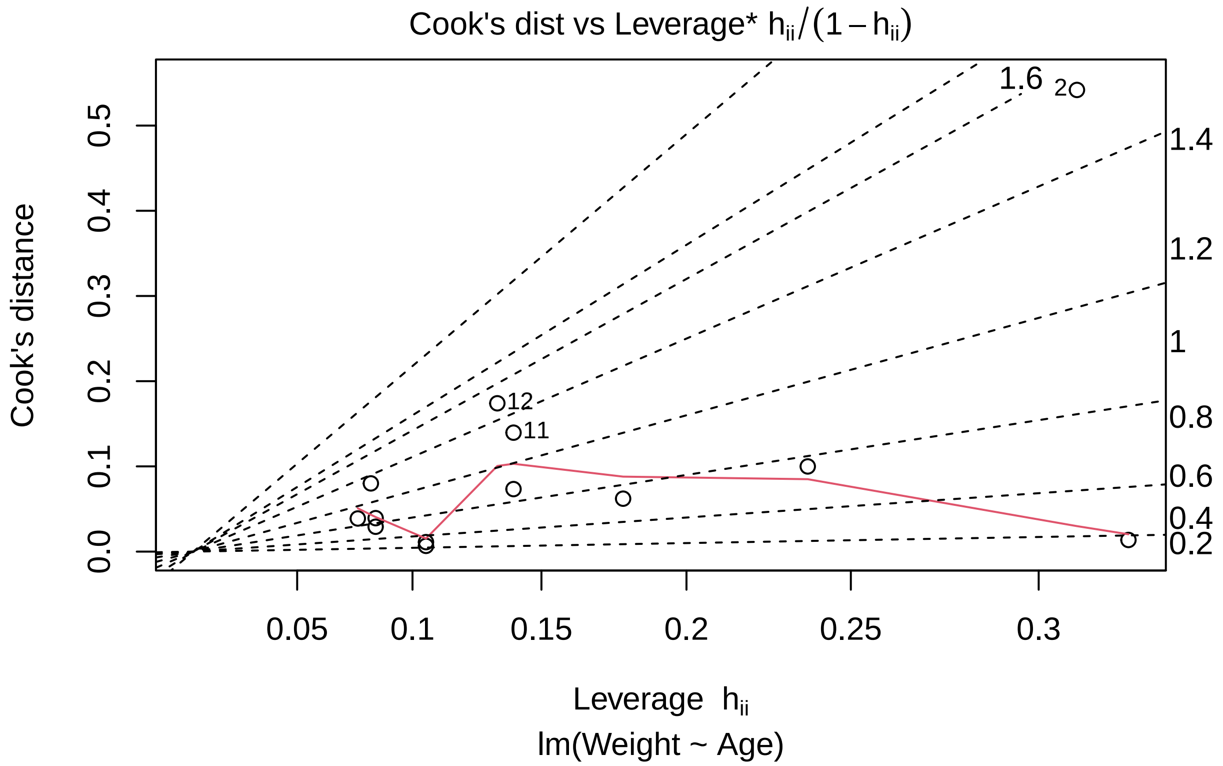
```
plot(model,4)
```


Cook's distance
Cook's distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2
11
12
2
4
6
8
10
12
Obs. number
lm(Weight ~ Age)

```
plot(model,5)
```


Residuals vs Leverage
Standardized residuals
1.5
0.5
−0.5
−1.5
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(Weight ~ Age)
12
11
2
0.5
0.5
Cook's distance

```
plot(model,6)
```



From looking at the normal qq-plot, we find that theere is some variability among the points, so the normality assumption may not be reasonable in this model. From looking at the residuals vs. fitted plot, there seems to be a curve pattern among the points so a linear model may not be the best fit. We also find that the 2nd observation is influential on the fitted values from looking at the Cook's distance plot.

Now, we fit a model with weight, age, and origin. To add origin as a predictor, we add the dummy variables Z_1 and Z_2 into the model.

```
model1 <- lm(Weight~Age+Z1+Z2, data=Turkey)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: Weight
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Age        1 26.2019 26.2019  290.71 3.691e-08 ***
## Z1         1  2.7165  2.7165   30.14 0.0003852 ***
## Z2         1  9.6873  9.6873  107.48 2.648e-06 ***
## Residuals  9  0.8112  0.0901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model1)
```

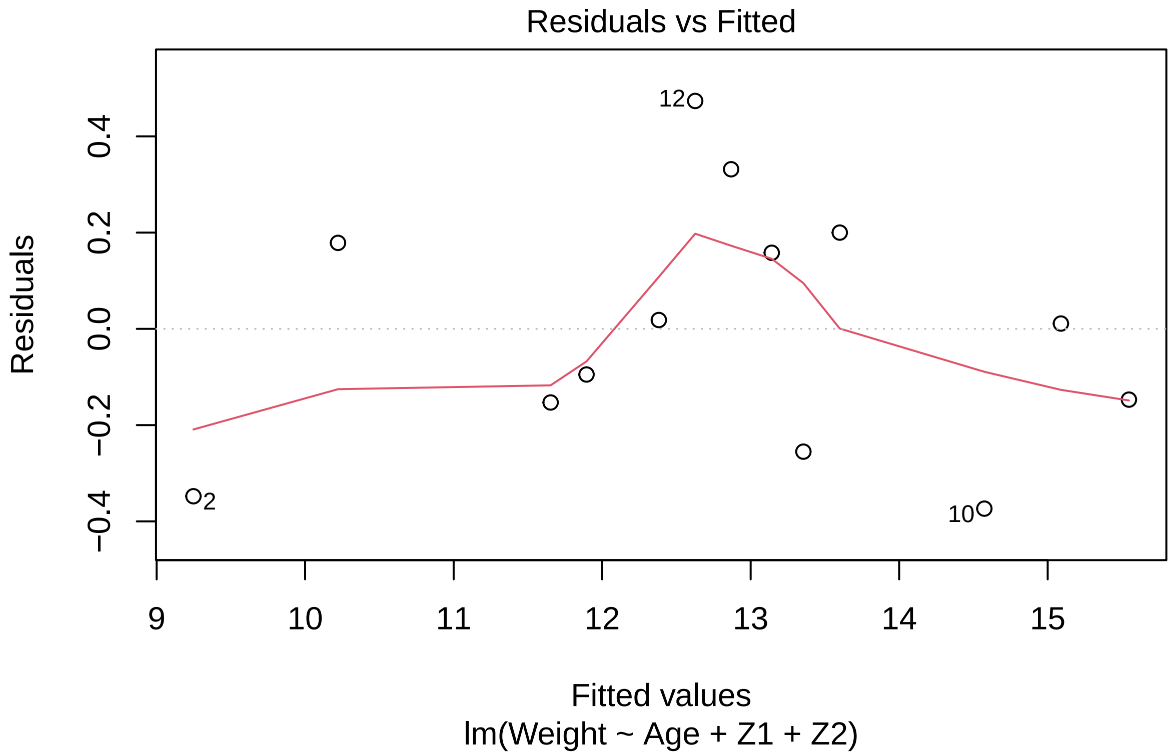
```
##
## Call:
```

```
## lm(formula = Weight ~ Age + Z1 + Z2, data = Turkey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37353 -0.15294  0.01103  0.17868  0.47353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.43088    0.65744   2.176   0.0575 .
## Age          0.48676    0.02574  18.908 1.49e-08 ***
## Z1          -1.91838    0.20180  -9.506 5.45e-06 ***
## Z2          -2.19191    0.21143 -10.367 2.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3002 on 9 degrees of freedom
## Multiple R-squared:  0.9794, Adjusted R-squared:  0.9726
## F-statistic: 142.8 on 3 and 9 DF,  p-value: 6.6e-08
```

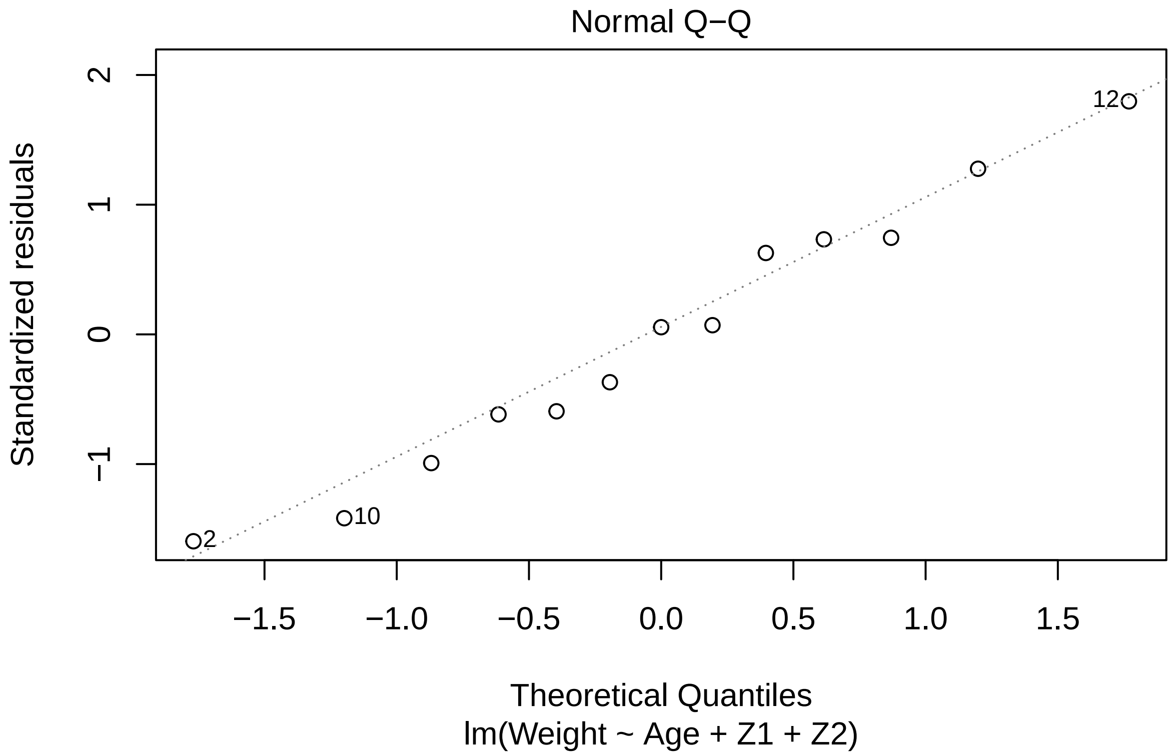
From looking at both the ANOVA table and the summary table, we find that there is a significant relationship between Turkey weight and age, Z_1, and Z_2, which means that there is a significant relationship between weight and age and between weight and origin. The R^2 value is 0.9794, which is a significant improvement compared to the model without origin as a variable. This means that 97.94% of variability is explained by the model.

Now, we create plots to check for model adequacy

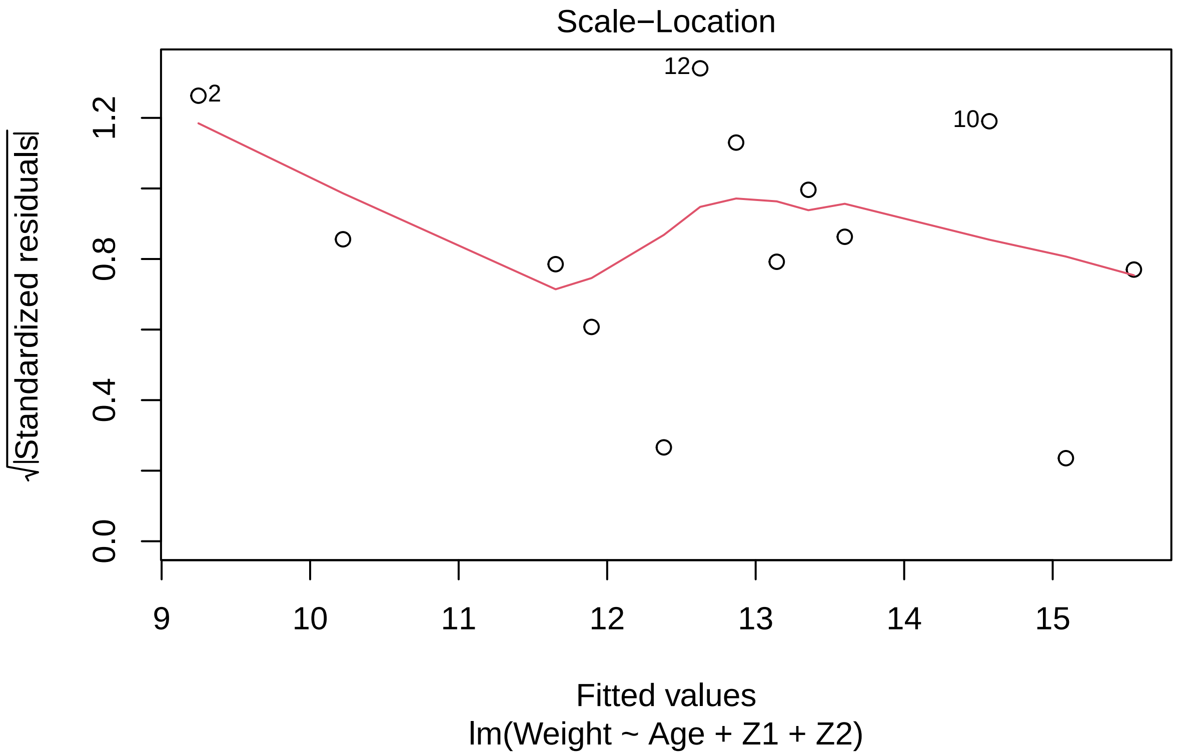
```r
Residuals1 <- model1$residuals
plot(model1,1)
```



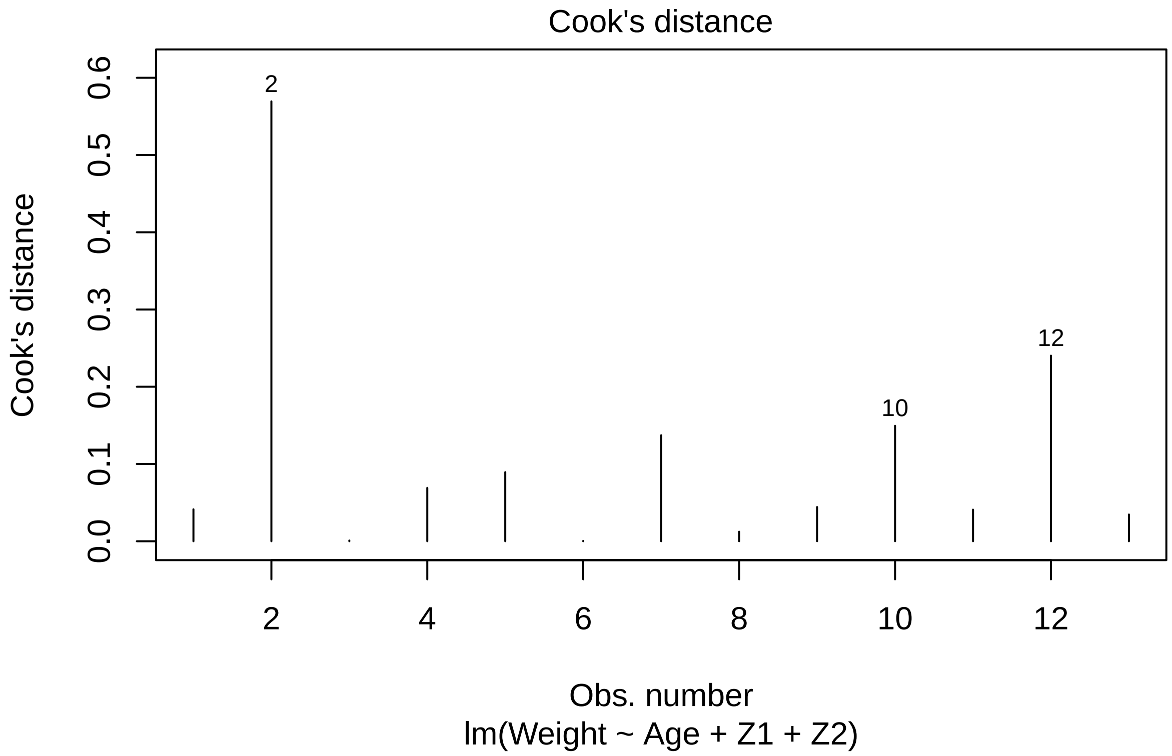
```
plot(model1,2)
```



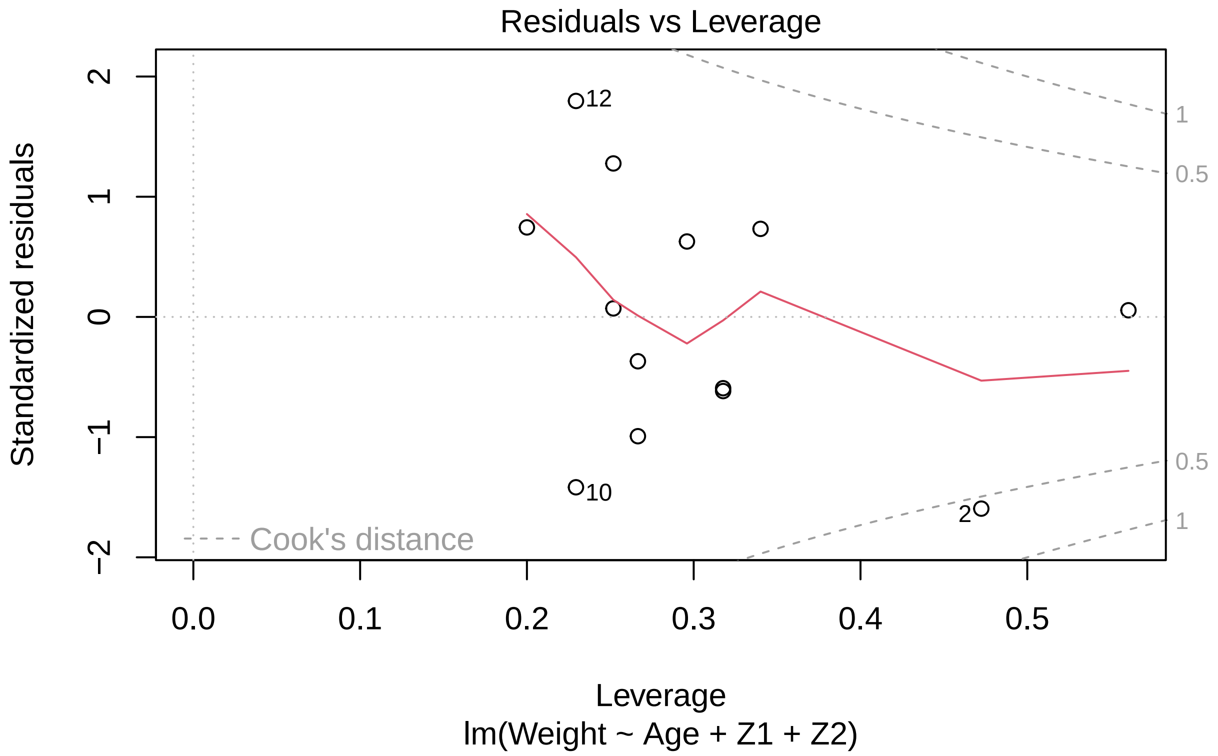
```
plot(model1,3)
```



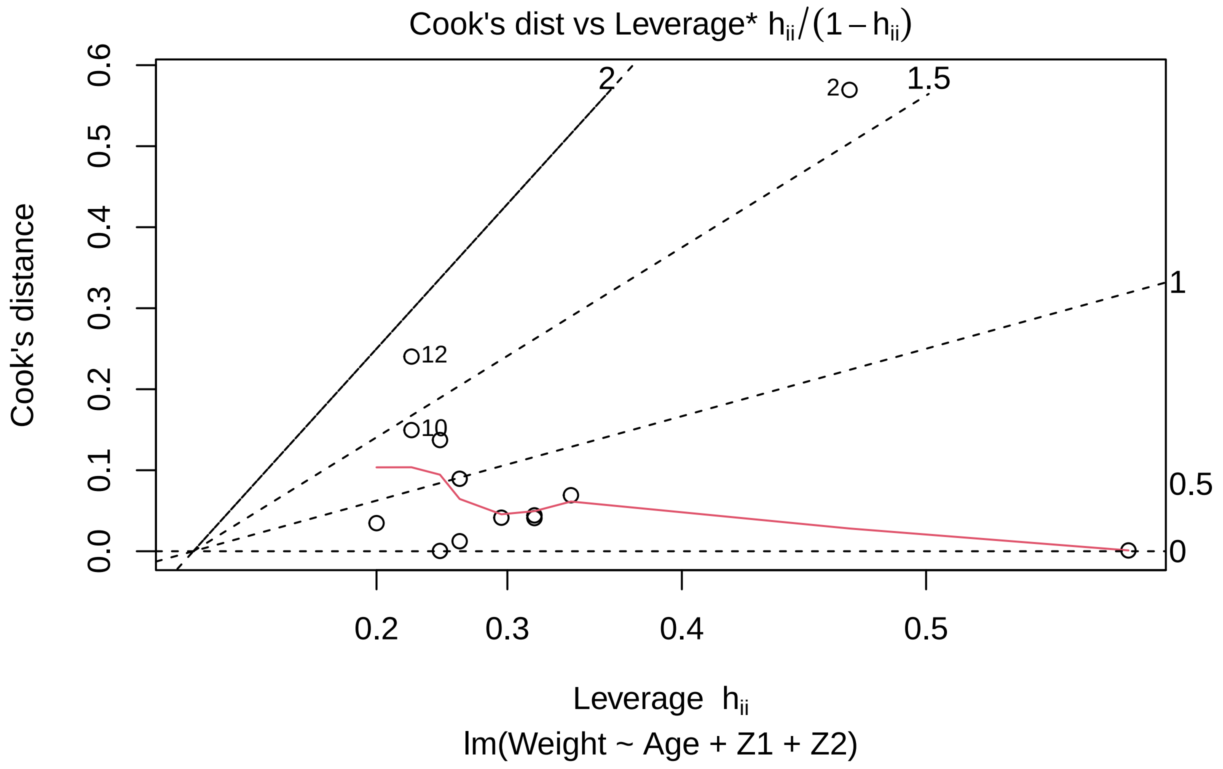
```r
plot(model1,4)
```



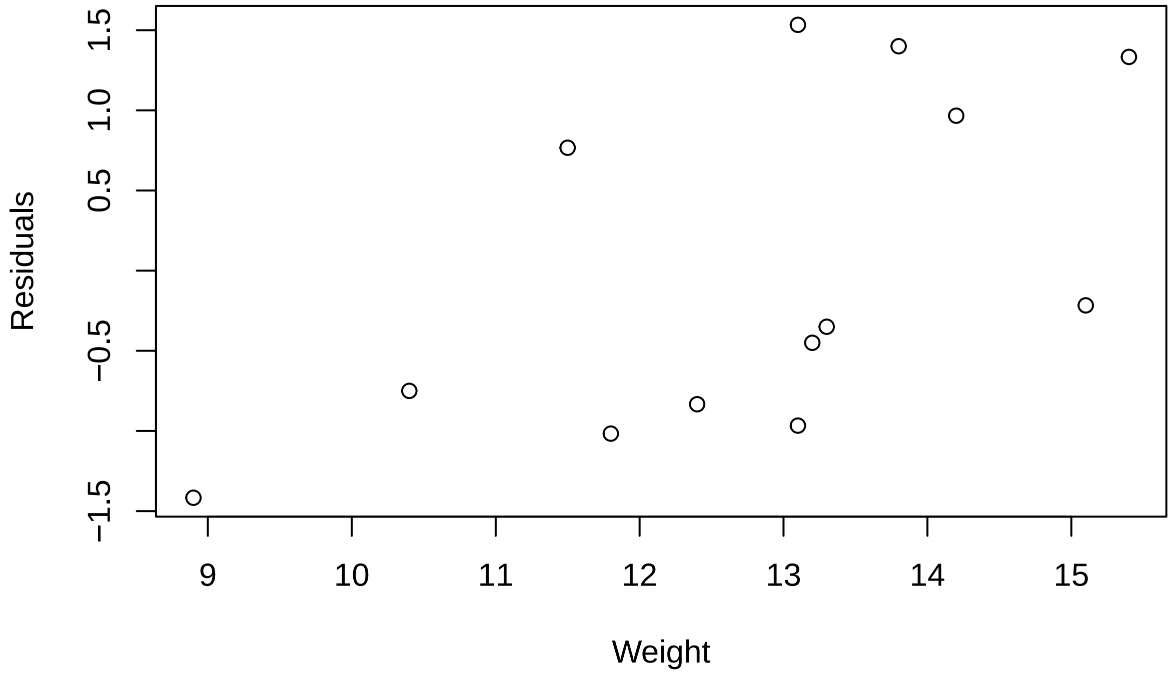
```
plot(model1,5)
```



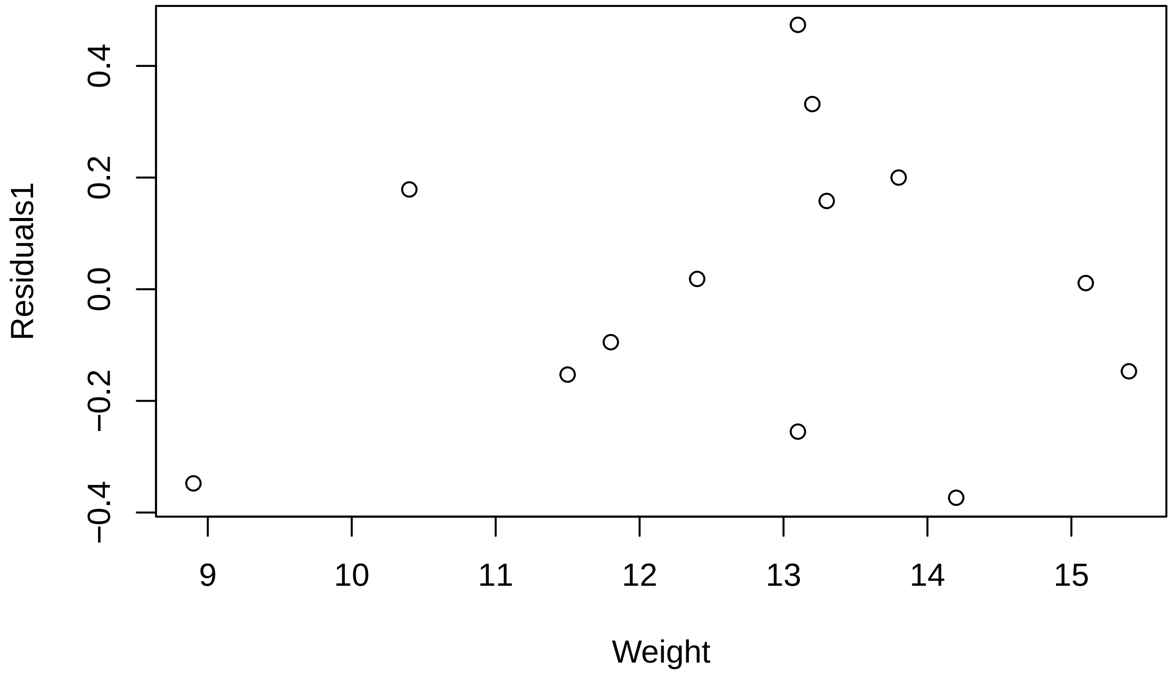
```
plot(model1,6)
```



```r
plot(Weight,Residuals)
```



```
plot(Weight,Residuals1)
```



The normal qq-plot is significantly improved compared to the previous model. The points follow a straight-line pattern closely. This means that the normality assumption is reasonable. The residuals vs. fitted plot indicates that a linear model may not be the best fit since there appears to be a slight curve pattern. From the Cook's distance plot, we find that the second observation is influential on the fitted values.

8.4 Example: Hardwood Data

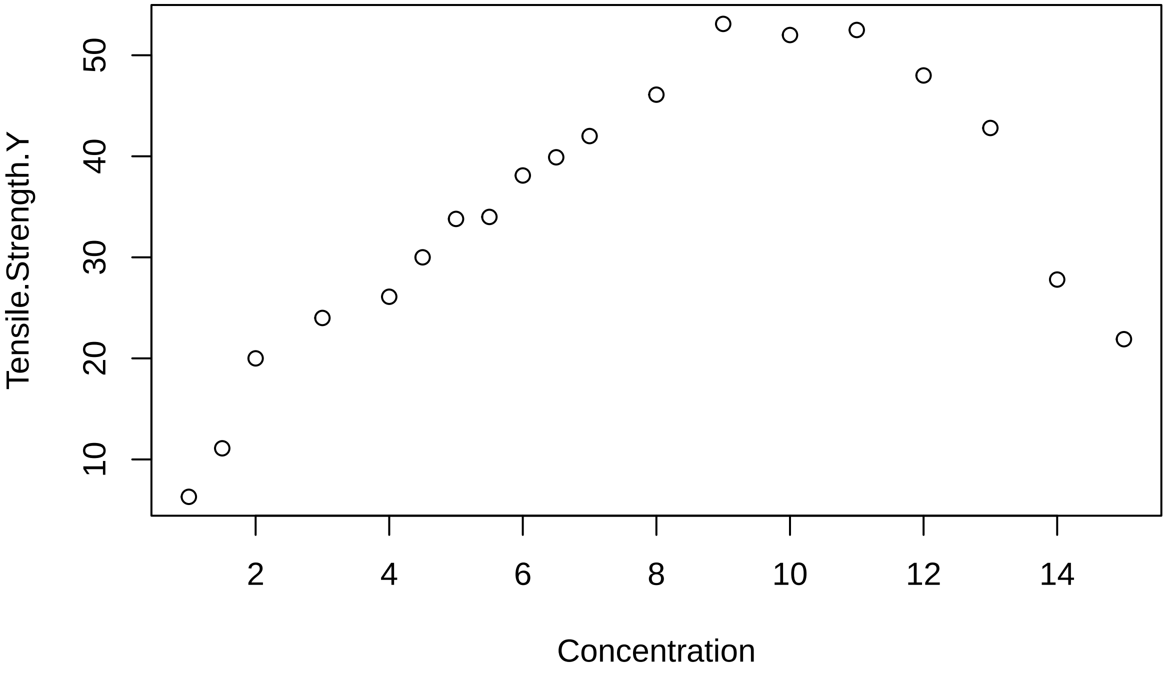
Now, we analyze the Hardwood data. We look at the polynomial regression model in this case. We first load the data.

```
Hardwood <- read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Hardwood Data.txt",
                       header=TRUE, sep="\t")
Hardwood
```

```
##    Concentration Tensile.Strength.Y
## 1            1.0                6.3
## 2            1.5               11.1
## 3            2.0               20.0
## 4            3.0               24.0
## 5            4.0               26.1
## 6            4.5               30.0
## 7            5.0               33.8
## 8            5.5               34.0
## 9            6.0               38.1
## 10           6.5               39.9
## 11          7.0               42.0
## 12          8.0               46.1
## 13          9.0               53.1
## 14         10.0               52.0
## 15         11.0               52.5
## 16         12.0               48.0
## 17         13.0               42.8
## 18         14.0               27.8
## 19         15.0               21.9
```

We first fit a linear model of tensile strength vs. concentration.

```
x <- Hardwood$Concentration
y <- Hardwood$Tensile.Strength.Y
x2 = x*x
plot(Hardwood)
```



From looking at the plot, it seems like the points do not follow a linear pattern, so a linear model likely won't be a good fit for the data. It looks more like a quadratic pattern. We first try fitting a linear model, then we try fitting a quadratic model.

```
model <- lm(y~x, data=Hardwood)
anova(model)
```

```
## Analysis of Variance Table
```

```
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## x          1 1043.4 1043.43  7.4736 0.01414 *
## Residuals 17 2373.5  139.62
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x, data = Hardwood)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.986  -3.749   2.938   7.675  15.840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.3213     5.4302   3.926  0.00109 **
## x             1.7710     0.6478   2.734  0.01414 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.82 on 17 degrees of freedom
## Multiple R-squared:  0.3054, Adjusted R-squared:  0.2645
## F-statistic: 7.474 on 1 and 17 DF,  p-value: 0.01414
```

From looking at the ANOVA table and the summary of the model, we find that there is a significant relationship between tensile strength and concentration. However, the relationship is only significant at the 0.05 level. The R^2 value is 0.3054, which means that only 30.54% of variability is explained by the model, which is low.

Now, we fit a polynomial regression model and see if we can improve the relationship:

```
model2 <- lm(y~x+x2)
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 1043.43 1043.43   53.40 1.758e-06 ***
## x2         1 2060.82 2060.82  105.47 1.894e-08 ***
## Residuals 16  312.64   19.54
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(model2)
```

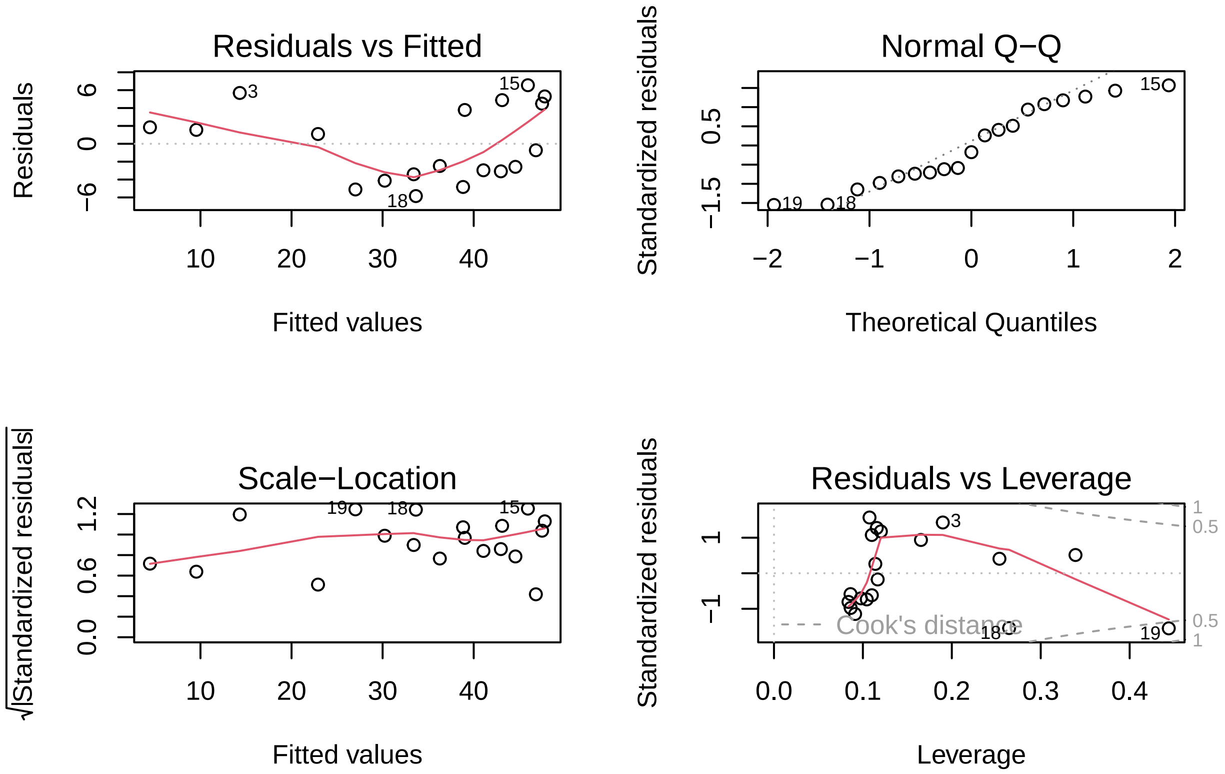
```
##
```

```
## Call:
## lm(formula = y ~ x + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8503 -3.2482 -0.7267  4.1350  6.5506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.67419    3.39971  -1.963   0.0673 .
## x           11.76401    1.00278  11.731 2.85e-09 ***
## x2          -0.63455    0.06179 -10.270 1.89e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.42 on 16 degrees of freedom
## Multiple R-squared:  0.9085, Adjusted R-squared:  0.8971
## F-statistic: 79.43 on 2 and 16 DF,  p-value: 4.912e-09
```

The quadratic model is a much better fit. There seems to be a highly significant relationship between tensile strength and concentration. The R^2 value is 0.9085, which means that around 90.85% of variability in Y is explained by the model. This is significantly higher than the original linear model without the quadratic term.

Now, we check for model adequacy:

```
par(mfrow = c(2,2))
plot(model2)
```



There is evidence of a nonlinear relationship between X and Y since there is a curve pattern in the residuals vs. fitted plot. It also means that the constancy of variance assumption is not always reasonable. We also have that the normal qq-plot doesn't follow a straight line pattern, so the normality assumption also doesn't seem reasonable with the model.

9 Multicollinearity and Ridge Regression

When there is correlation among the predictors themselves in a linear regression model, then we say that multicollinearity exists. In this section, we talk about some potential signs of multicollinearity, diagnostics for multicollinearity, and remedial measures to overcome this problem.

9.1 Multicollinearity overview

Some symptoms of multicollinearity include large variation in the estimated coefficients when a new variable is either added or deleted, when there are non significant results in individual tests on the coefficients of important variables, when there's large coefficients of simple correlation between pairs of variables, and when the confidence intervals for the regression coefficients of important variables are too wide.

When the variables are correlated among themselves, we say that multicollinearity exists. The principal difficulty is that the matrix $\boldsymbol{X^T X}$ may not be invertible. Multicollinearity also affects the interpretation of the coefficients as they may vary in value. Consider the case of two predictor variables X_1, X_2. If the variables are standardized, then we have the matrix

$$\boldsymbol{X^T X} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

where r_{12} is the correlation between the two variables. Then we get that

$$\boldsymbol{(X^T X)^{-1}} = \frac{1}{1-r_{12}^2}\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

and consequently, the variance-covariance matrix of the coefficients is

$$\sigma^2(X^T X)^{-1} = \sigma^2\frac{1}{1-r_{12}^2}\begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

As a result, as $|r_{12}| \to 1$, we have that the variance of the individual coefficients approach ∞. As $r_{12} \to \pm 1$, we have that the covariance of the coefficients approach $\pm\infty$. This means that the variance and covariance of the estimates of the coefficients grows larger when the correlation between individual coefficients gets closer to 1. In other words, multicollinearity causes the variance of the coefficients to become larger.

In general, the diagonal elements of $\boldsymbol{(X^T X)^{-1}}$ are $C_{jj} = \frac{1}{1-R_j^2}$ where R_j^2 is the R-squared value obtained from the regression of X_j on the other $p-1$ variables. If there is strong multicollinearity between X_j and the other $p-1$ variables, then $R_j^2 \approx 1$ and $Var[\hat{\beta}_j] \approx \infty$.

Under multicollinearity, the values of the estimates will also be large. We set

$$L = ||\boldsymbol{\hat{\beta}} - \boldsymbol{\beta}||^2$$

Then,

$$\begin{aligned} E[\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2] &= \sum_{j=1}^{p} Var[\hat{\beta}_j] \\ &= \sigma^2 Trace[\boldsymbol{(X^T X)^{-1}}] \\ &= \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \end{aligned}$$

where $\{\lambda_j\}$ are the eigenvalues of $\boldsymbol{X^T X}$.

Under multicollinearity, some of these eigenvalues will be small and hence their inverses will be large.

We expand the loss function L, then we get $L = \boldsymbol{\hat{\beta}^T \hat{\beta} - 2\hat{\beta}\beta + \beta^T \beta}$. Taking the expectation, we get

$$\begin{aligned} E[L] &= E[||\boldsymbol{\hat{\beta}}||^2] - ||\boldsymbol{\beta}||^2 \\ &= \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \end{aligned}$$

Hence, $E[||\boldsymbol{\hat{\beta}}||^2] = ||\boldsymbol{\beta}||^2 + \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$. This shows that the estimators are large when the inverses of the eigenvalues are large as a result of multicollinearity. In conclusion, multicollinearity causes the estimates of coefficients and the variances of the coefficient estimates to be large.

9.2 Multicollinearity Diagnostics

We start by standardizing the variables.

$$Y_i^* = \frac{1}{\sqrt{n-1}}(\frac{Y_i - \overline{Y}}{s_Y})$$

$$X_{ik}^* = \frac{1}{\sqrt{n-1}}(\frac{X_{ik} - \overline{X_k}}{s_k})$$

where s_Y is the standard deviation of the observations of Y_i and s_k is the standard deviation of the observations of X_{ik}.

Then we get that

$$Y_i^* = \sum_{k=1}^{p-1} \beta_i^* X_{ik}^* + \epsilon_i^*$$

$$\beta_k = (\frac{s_Y}{s_k})\beta_k^*$$

$$\beta_0 = \overline{Y} - \sum_{i=1}^{p-1} \beta_i \overline{X}_i$$

$$r_{XX} b^* = r_{YX}$$

where r_{XX} is the correlation matrix

$$r_{XX} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1,p-1} \\ r_{12} & 1 & \dots & r_{2,p-1} \\ \dots & \dots & \dots & \dots \\ r_{1,p-1} & \dots & \dots & 1 \end{pmatrix}$$

and

$$r_{YX} = \begin{pmatrix} r_{Y1} \\ r_{Y2} \\ \dots \\ r_{Y,p-1} \end{pmatrix}$$

We also have $r_{ij} = Corr(X_i, X_j)$ and $r_{Yi} = Corr(Y, X_i)$.

We now talk about diagnostics for multicollinearity. We first talk about the variance inflation factors (VIFs). Suppose that the regression is fitted using standardized predictor variables. Then we have $Var[\boldsymbol{b}] = \sigma^2 r_{XX}^{-1}$, where r_{XX} is the matrix of pairwise correlation coefficients among the predictors. We define the variance inflation factor (VIF)

$$(VIF)_k = (1 - R_k^2)^{-1}$$

where R_k^2 is the coefficient of multiple determination where X_k is regressed on the $p-2$ other X variables. Hence, $Var[b_k] = \sigma^2(1 - R_k^2)^{-1}$. This means that the VIFs can inflate the variances of the estimates of the coefficients, which is a sign of multicollinearity.

For each variable X_k, we have that $(VIF)_k = 1$ when $R_k^2 = 0$. Whenever X_k is not linearly related to the other X variables in the model. Under perfect correlation ($R_k^2 = 1$), the variance is unbounded. As a rule of thumb, a value $(VIF)_k > 10$ indicates that multicollinearity exists. Alternatively, we can compute the average

$$\overline{VIF} = \frac{\sum (VIF)_k}{p-1}$$

Mean values much greater than 1 point to serious multicollinearity.

We now talk about some other diagnostic measures for multicollinearity. As mentioned previously, the eigenvalues λ_i of the matrix $\boldsymbol{X^T X}$ can be used to measure the extend of multicollinearity in the system. If one or more are small, then there are near linear dependencies in the columns of $\boldsymbol{X^T X}$.

The condition number κ and condition indices κ_j of $\boldsymbol{X^T X}$ are defined to be

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

and

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j}$$

We use the following rule of thumb: we say that there is no serious multicollinearity if $\kappa < 100$, there is moderate to strong multicollinearity if $100 < \kappa < 1000$, and there is severe multicollinearity if $\kappa > 1000$.

9.3 Ridge Regression

Ridge regression is considered as a remedial measure to multicollinearity. The theory is as follows. We first transform the normal equation using standardized variables, so it goes from $(X^T X)b = X^T Y$ to $r_{XX} b = r_{YX}$.

Instead of solving the original normal equation, we instead solve the equation

$$(r_{XX} + cI)b^R = r_{YX}$$

where $c > 0$ is a constant and b^R is the ridge regression estimate of the parameters. The standardized ridge regression coefficients become

$$b^R = (r_{XX} + cI)^{-1} r_{YX}$$

The constant c reflects the fact that the ridge estimators will be biased (i.e. $E[b^R] \neq \beta^*$) but they tend to be more stable or less variable than the ordinary least squares estimators.

The constant c is usually chosen in such a way that the estimators of b_k^R are stable in value. Alternatively, whenever the $(VIK)_k$ are stable in value. A plot of coefficients against c is called the ridge trace and this helps in the selection of c.

Ridge regression can also be obtained from the method of penalized regression. We have the following system of equations:

$$\begin{aligned}
(1+c)b_1^R + r_{12}b_2^R + \cdots + r_{1,p-1}b_{p-1}^R &= r_{Y1} \\
r_{21}b_1^R + (1+c)b_2^R + \cdots + r_{2,p-1}b_{p-1}^R &= r_{Y2} \\
&\cdots \\
r_{p-1,1}b_1^R + r_{p-1,2}b_2^R + \cdots + (1+c)b_{p-1}^R &= r_{Y,p-1}
\end{aligned}$$

This is also the same as solving the penalized least squares

$$Q = \sum [Y_i - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1}]^2 + c \sum_{j=1}^{p-1} \beta_j^2$$

If we differentiate the equation above with respect to each of the parameters, we get the system of equations above. In the above equation, the penalty function is $c\sum_{j=1}^{p-1} \beta_j^2$

The major drawback of ridge regression is that the ordinary inference procedures are no longer applicable (i.e. ANOVA, t-tests, correlation tests). We need to use resampling methods to obtain the precision of the estimators.

Another approach is to use a penalty function

$$c \sum_{j=1}^{p-1} |\beta_j|$$

which permits some regression coefficients to be 0. This is know as LASSO regression, which stands for Least Absolute Shrinkage and Selection Operator.

9.4 Example: Body Fat Data Analysis

Consult the example Alvo did in class. He did not upload the Body Fat data to Brightspace.

10 Building the regression model

In this section, we talk about selecting the best model when given several predictor variables. We discuss how to select a subset of all predictors such that the model is a good fit for the data.

If $p-1$ predictors are available, then there will be $2^p - 1$ possible models which can be constructed.

10.1 All possible regressons

This method is self explanatory. We consider all $2^p - 1$ possible models.

We talk about criteria for model selection. There are several criteria to consider when we pick a model.

10.1.1 R^2

The R^2 criteria chooses the model with the largest value of explained variation. A plot of the mean square residual vs. the number of variables in the model will appear as a parabola with the last entry "curving" up a bit. This is the one with all variables in. One may draw a horizontal line parallel to the x-axis. The point where it meets the parabola determines the best fitting model since it will be as good as when all variables are included.

10.1.2 Adjusted R^2

An adjusted R^2 takes into account the values of n and p

$$\begin{aligned} R^2_{a,p} &= 1 - (\frac{n-1}{n-p})\frac{SSE}{SSTO} \\ &= 1 - \frac{MSE(p)}{SSTO/(n-1)} \end{aligned}$$

The criteria minimum $MSE(p)$ and maximum adjusted R^2 are equivalent. The adjusted R^2 does not necessarily increase as additional explanatory variables are added into the model. This value is similar to the R^2 value except that it can be negative.

10.1.3 Mallow's C_p

Now, we talk about another criteria that is used for model selection. We derive the Mallow's C_p criterion first.

Suppose that the true model has q predictor variables

$$\boldsymbol{Y} = \boldsymbol{X_q \beta_q} + \boldsymbol{\epsilon}$$

Suppose instead we fit a model using only p predictor variables. Let $\boldsymbol{H_p}$ be the hat matrix using only p variables. The bias for the i-th fitted value is

$$E[\hat{Y}_i] - \mu_i$$

where μ_i is the true mean. Consequently,

$$E[(\hat{Y}_i - \mu_i)^2] = (E[\hat{Y}_i] - \mu_i)^2 + Var[\hat{Y}_i]$$

The total mean squared error for all the fitted values divided by σ^2 is

$$\Gamma_p = \frac{1}{\sigma^2}(\sum_i (E[\hat{Y}_i] - \mu_i)^2 + \sum_i Var[\hat{Y}_i])$$

We may estimate σ^2 by the MSE when all the variables are included.

The vector of residuals becomes

$$\boldsymbol{e_p} = \boldsymbol{(I - H_p)Y}$$

and the error sum of squares is

$$SSE_p = \boldsymbol{e_p^T e_p}$$

It follows that

$$bias = E[\boldsymbol{e_p}] = \boldsymbol{(I - H_p)E[Y]} = E[\boldsymbol{Y}] - E[\boldsymbol{\hat{Y}}]$$

since $E[\boldsymbol{H_p Y}] = \boldsymbol{H_p} E[\boldsymbol{Y}] = E[\boldsymbol{\hat{Y}}]$.

When $p = q$, $bias = E[\boldsymbol{Y}] - E[\boldsymbol{\hat{Y}}] = 0$

Now, using the idempotency of $\boldsymbol{(I - H_p)}$, we get

$$\begin{aligned}
E[SSE_p] &= E[\boldsymbol{e_p^T e_p}] \\
&= E[\boldsymbol{Y^T (I - H_p) Y}] \\
&= E[\boldsymbol{Y^T (I - H_p)(I - H_p) Y}] \\
&= \sigma^2 Trace[\boldsymbol{I - H_p}] + (bias)^T (bias) \\
&= \sigma^2 (n - p) + \sum_i (E[\hat{Y}_i] - \mu_i)^2
\end{aligned}$$

The total mean squared error for all the fitted values divided by σ^2 is then

$$\begin{aligned}
\Gamma_p &= \frac{1}{\sigma^2} E[\sum_i (E[\hat{Y}_i] - \mu_i)^2 + \sum_i Var[\hat{Y}_i]] \\
&= \frac{1}{\sigma^2}(E[SSE_p] - \sigma^2(n - p) + \sum_i Var[\hat{Y}_i]) \\
&= \frac{1}{\sigma^2} E[SSE_p] - (n - p) + \frac{1}{\sigma^2} \sum_{i=1}^{n} Var[\hat{Y}_i] \\
&= \frac{1}{\sigma^2} E[SSE_p] - (n - p) + \frac{1}{\sigma^2} \sigma^2 \sum_{i=1}^{n} [\boldsymbol{x_i^T (X^T X)^{-1} x_i}] \\
&= \frac{1}{\sigma^2} E[SSE_p] - (n - p) + \frac{1}{\sigma^2} \sigma^2 Trace[\boldsymbol{H_p}] \\
&= \frac{1}{\sigma^2} E[SSE_p] - (n - p) + \frac{1}{\sigma^2} \sigma^2 p \\
&= \frac{1}{\sigma^2} E[SSE_p] - (n - p) + p
\end{aligned}$$

Hence, we get that

$$\Gamma_p = \frac{1}{\sigma^2} E[SSE_p] - (n - 2p)$$

If we estimate σ^2 by MSE with all predictors and $E[SSE_p]$ by SSE_p then the Mallow's criteria becomes

$$C_p = \frac{SSE_p}{MSE} - (n - 2p)$$

If the p-term model has negligible bias, then $E[SSE_p] \approx (n-p)\sigma^2$ and $C_p \approx p$. Mallows proposed a graphical technique to find the optimal subset. Plot C_p vs. p for all possible regressions. Models with small bias will be close to the line $C_p = p$ while those with large bias will be above the line. Values below the line are considered to have no bias.

10.1.4 Akaike Information Criterion

Akaike proposed a criteria based on minimizing the expected entropy of the model, which is essentially a penalized likelihood measure. In the case of ordinary least squares regression, it becomes

$$AIC_p = n \ln(SSE_p) - n \ln(n) + 2p$$

As more variables are included, AIC_p decreases and the issue becomes whether or not the decrease justifies the inclusion of more variables.

10.1.5 Schwartz's Bayesian Criterion

A Bayesian extension of the Akaike criterion was proposed by Schwartz

$$BIC_{Sch} = n \ln(SSE_p) - n \ln(n) + p(\ln(n))$$

This criterion places a greater penalty than the Akaike criterion and it is the one used by R.

10.1.6 Prediction Sum of Squares Criterion (PRESS)

Sometimes regression equations are used to predict future values. A different criteria used is to select the model which minimizes

$$PRESS_p = \sum_i [Y_i - \hat{Y}_{(i)}]^2$$

where $\hat{Y}_{(i)}$ is the fitted value when the i-th observation is deleted.

10.2 Forward Selection

In the previous section, we talked about all possible regressions and how to use different criteria to select the best model out of all. In this section and the next 2 sections, we talk about iterative models that can be helpful in selecting the best model. These methods work better with many predictors since using all possible regressions can be computationally difficult with $2^p - 1$ models when p is large.

We first talk about Forward Selection. We start with no regressors in the model (i.e. intercept only). We compute the standardized student t statistic for each variable and choose the one with the greatest absolute value to include in the model. This is also the variable that has the largest simple correlation with the response. We also choose a pre-selected critical F value, denote it by F_{in}.

With the variable selected, we choose the next variable using the same criteria as the previous step after adjusting for the effect of the first variable selected. The criteria makes use of partial correlations which are computed between the residuals from the previous step and the residuals from the regressions of the other regressors on X_j, that is residuals from $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ and residuals from $\hat{X}_j = \hat{\alpha_{0j}} + \hat{\alpha_{1j}} X_1$ for $j = 2, ..., k$. In short, we plot the residuals from the 2 models and we choose X_j such that it maximizes the partial correlation.

If X_2 is selected, it implies that the largest partial F statistic is

$$F = \frac{SSR(X_2|X_1)}{MSE(X_1, X_2)}$$

If $F > F_{in}$, then X_2 is entered into the model. Then we keep going until either all variables are included or when the largest partial F statistic no longer exceeds F_{in}.

One drawback to this method is that once a variable is entered, it doesn't get eliminated. This means that the variable can turn out to not be significant in the chosen model but it doesn't get removed.

10.3 Backward Elimination

This model is the opposite of Forward Selection discussed in the section above. We start with all regressors in the model (i.e. the full model with all predictors included). Compute the partial F statistic for each regressor as if it were the last one to enter the model. We compare the smallest partial F with a pre-selected F_{out} value. If it is smaller, then that variable is removed from the model. The procedure is repeated until the smallest partial F statistic is not less than F_{out} or when all variables are removed. Backward elimination is often preferred to forward selection because it begins with all the variables in the model.

One drawback is that once a variable is eliminated, it cannot be added back into the model. This can be a problem if that variable is significant in the reduced model chosen.

10.4 Stepwise Selection

Stepwise Selection combines the 2 approaches Forward Selection and Backward Elimination. It is a modification of forward selection in that it reassesses each of the regressors already in the model to see if it has become redundant. We need to pre-select F_{in} and F_{out}. Usually, we choose $F_{in} > F_{out}$ so that it becomes more difficult to add a variable than to remove it.

10.5 Example: Hald Cement Data Analysis

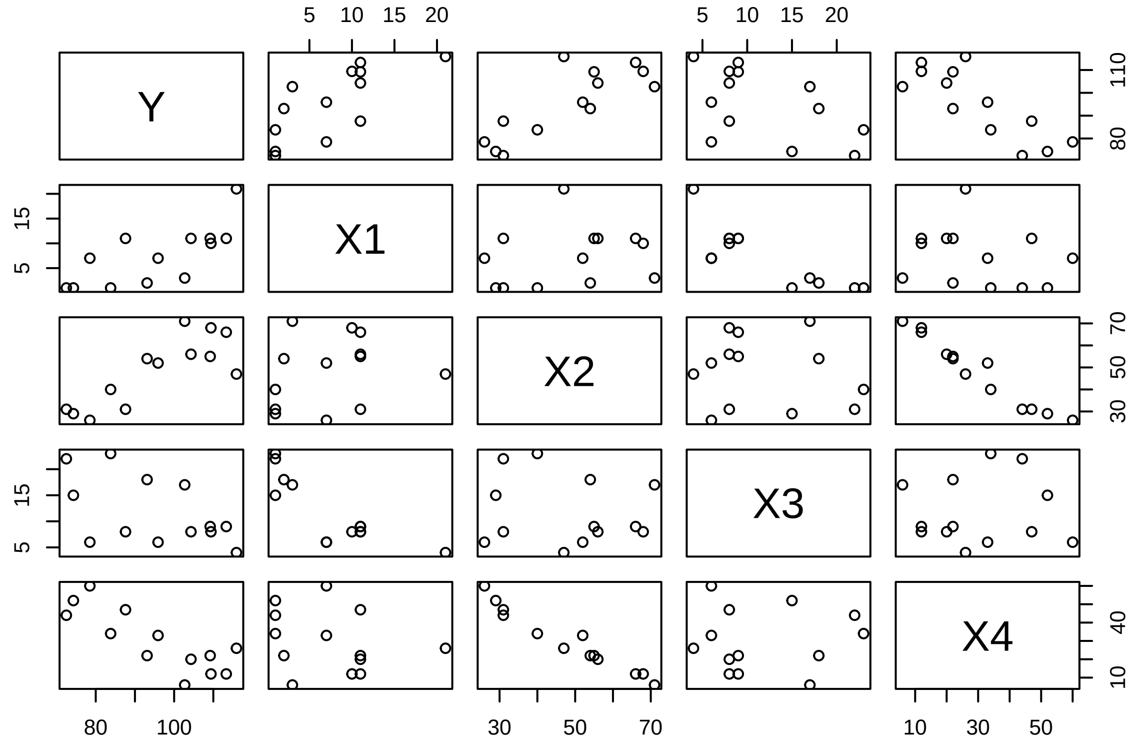
We now look at an example that examines the Hald Cement data. We first load the data:

```
Cement <- read.table("/System/Volumes/Data/MAT 3375/Summary Sheet/Hald Cement Data.txt", header=TRUE, s
```

```
names(Cement)
```

```
## [1] "Y"  "X1" "X2" "X3" "X4"
```

```
plot(Cement)
```



```r
cor(Cement)
```

```
##             Y         X1         X2         X3         X4
## Y   1.0000000  0.7307175  0.8162526 -0.5346707 -0.8213050
## X1  0.7307175  1.0000000  0.2285795 -0.8241338 -0.2454451
## X2  0.8162526  0.2285795  1.0000000 -0.1392424 -0.9729550
## X3 -0.5346707 -0.8241338 -0.1392424  1.0000000  0.0295370
## X4 -0.8213050 -0.2454451 -0.9729550  0.0295370  1.0000000
```

From looking at the correlation matrix and the plot matrix, we find that Y is strongly correlated with X_1, X_2, and X_4. There also appears to be multicollinearity between X_1 and X_3 and between X_2 and X_4.

Now, we look at multicollinearity diagnostics:

```r
model <- lm(Y~., data=Cement)
ols_coll_diag(model)
```

```
## Tolerance and Variance Inflation Factor
## ---------------------------------------
##   Variables   Tolerance       VIF
## 1        X1 0.025976582  38.49621
## 2        X2 0.003930460 254.42317
## 3        X3 0.021336344  46.86839
## 4        X4 0.003539662 282.51286
##
```

```
## 
## Eigenvalue and Condition Index
## ------------------------------
##     Eigenvalue Condition Index    intercept           X1           X2
## 1 4.119699e+00        1.000000 5.508939e-06 0.0003688922 1.832878e-05
## 2 5.538943e-01        2.727214 8.812348e-08 0.0100384607 1.264739e-05
## 3 2.887021e-01        3.777529 3.060952e-07 0.0005755116 3.198053e-04
## 4 3.763830e-02       10.462074 1.267880e-04 0.0574472804 2.783994e-03
## 5 6.613815e-05      249.578252 9.998673e-01 0.9315698551 9.968652e-01
##            X3           X4
## 1 0.000210219 3.640648e-05
## 2 0.002658636 1.007001e-04
## 3 0.001592054 1.680306e-03
## 4 0.045693504 8.837332e-04
## 5 0.949845587 9.972989e-01
```

From looking at the VIFs, we have that there is strong multicollinearity with X_2 and X_4 when all predictors are fitted. This means that we should pick a model that doesn’t necessarily include all variables.

We first try running all possible regressions:

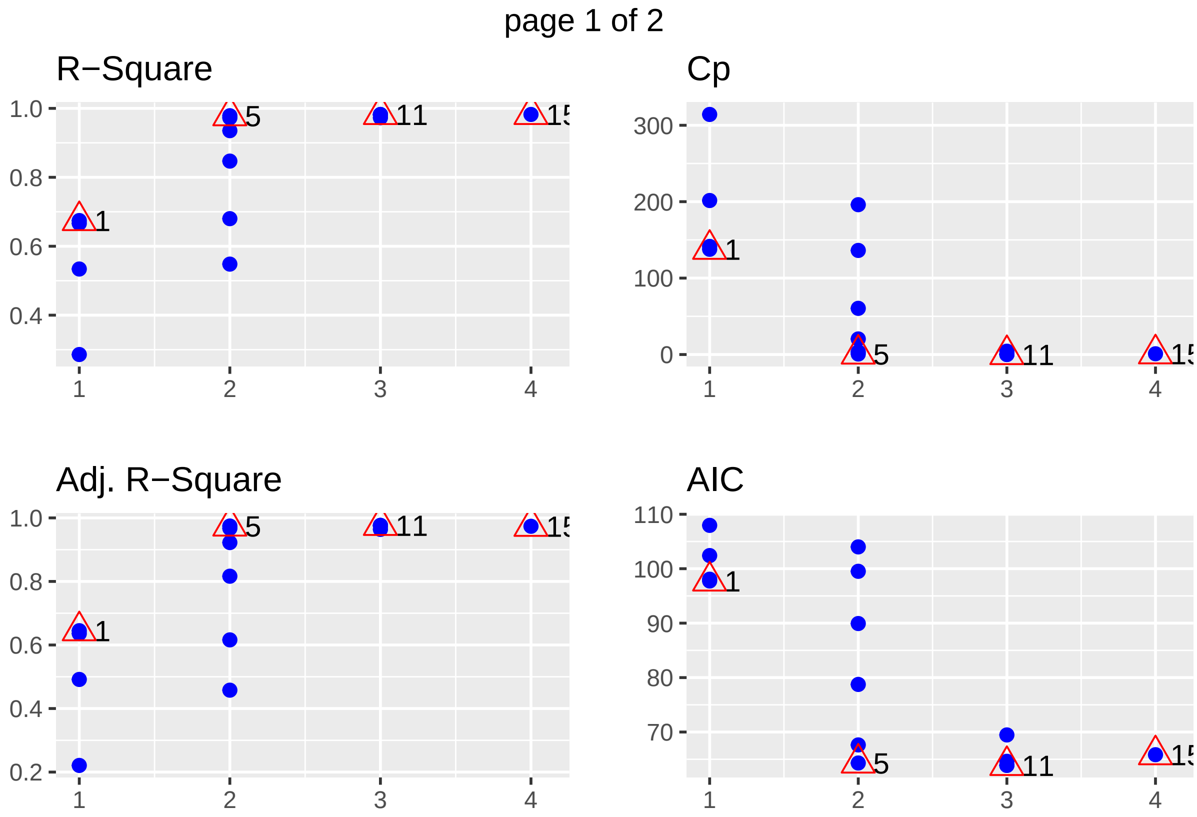
```
k = ols_step_all_possible(model)
k
```

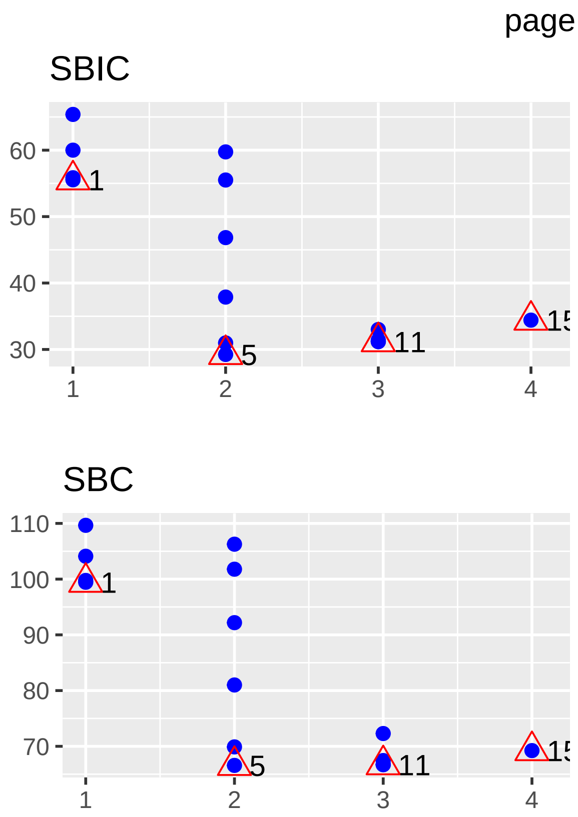
```
##    Index N  Predictors  R-Square Adj. R-Square Mallow’s Cp
## 4      1 1          X4 0.6745420     0.6449549  138.730833
## 2      2 1          X2 0.6662683     0.6359290  142.486407
## 1      3 1          X1 0.5339480     0.4915797  202.548769
## 3      4 1          X3 0.2858727     0.2209521  315.154284
## 5      5 2       X1 X2 0.9786784     0.9744140    2.678242
## 7      6 2       X1 X4 0.9724710     0.9669653    5.495851
## 10     7 2       X3 X4 0.9352896     0.9223476   22.373112
## 8      8 2       X2 X3 0.8470254     0.8164305   62.437716
## 9      9 2       X2 X4 0.6800604     0.6160725  138.225920
## 6     10 2       X1 X3 0.5481667     0.4578001  198.094653
## 12    11 3    X1 X2 X4 0.9823355     0.9764473    3.018233
## 11    12 3    X1 X2 X3 0.9822847     0.9763796    3.041280
## 13    13 3    X1 X3 X4 0.9812811     0.9750415    3.496824
## 14    14 3    X2 X3 X4 0.9728200     0.9637599    7.337474
## 15    15 4 X1 X2 X3 X4 0.9823756     0.9735634    5.000000
```

From looking at the R^2, adjusted R^2, and the Mallow’s C_p criterion, we find that the model involving X_1, X_2, and X_4 is the best model. However, we have seen that there is strong multicollinearity when both X_2 and X_4 are included in the model. So we can choose a further reduced model. The model with X_1, X_2, and X_3 also looks like a good model since it has a high R^2 value and the C_p value is close to the number of predictors. However, from looking at the correlation matrix, there’s high correlation between X_1 and X_3. We see that the model using only X_1 and X_2 has a high R^2 and adjusted R^2 and the C_p criterion is also close to the number of predictors. Therefore, the model involving only X_1 and X_2 is a good model to choose.

```
plot(k)
```

```
## Warning: The ‘guide‘ argument in ‘scale_*()‘ cannot be ‘FALSE‘. This was deprecated in
## ggplot2 3.3.4.
## i Please use "none" instead.
## i The deprecated feature was likely used in the olsrr package.
##   Please report the issue at <https://github.com/rsquaredacademy/olsrr/issues>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```





Now, we run best subset selection and choose the best model using it.

```
k = ols_step_best_subset(model)
k
```

```
##  Best Subsets Regression
## --------------------------
## Model Index    Predictors
## --------------------------
##      1         X4
##      2         X1 X2
##      3         X1 X2 X4
##      4         X1 X2 X3 X4
## --------------------------
##
##                                                     Subsets Regression Summary
## ------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model     R-Square   R-Square    R-Square      C(p)         AIC        SBIC        SBC          MSEP
## ------------------------------------------------------------------------------------------------------
##   1         0.6745     0.6450      0.5603    138.7308    97.7440    55.5401    99.4389    1047.0423
##   2         0.9787     0.9744      0.9654      2.6782    64.3124    29.2437    66.5722      76.2162
##   3         0.9823     0.9764      0.9686      3.0182    63.8663    31.1723    66.6910      71.0365
##   4         0.9824     0.9736      0.9594      5.0000    65.8367    34.4130    69.2264      81.0000
## ------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

We find that the best subset is the subset X_1, X_2, and X_4. However, as mentioned previously, there may be multicollinearity issues with the model. The next best model is the model with 2 predictors X_1 and X_2, which has less multicollinearity issues. This means that it's better to choose the subset with X_1 and X_2.

Now, we run forward selection

```
ols_step_forward_p(model, details=TRUE)
```

```
## Forward Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. X1
## 2. X2
## 3. X3
## 4. X4
##
## We are selecting variables based on p value...
##
##
## Forward Selection: Step 1
##
## - X4
##
##                         Model Summary
## --------------------------------------------------------------
## R                        0.821       RMSE                8.964
## R-Squared                0.675       Coef. Var           9.394
## Adj. R-Squared           0.645       MSE                80.352
## Pred R-Squared           0.560       MAE                 6.894
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression    1831.896         1       1831.896    22.799     6e-04
## Residual       883.867        11         80.352
## Total         2715.763        12
## --------------------------------------------------------------------
##
##                                     Parameter Estimates
## -------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## -------------------------------------------------------------------------------------------
## (Intercept)    117.568         5.262                 22.342    0.000    105.986    129.150
##          X4     -0.738         0.155       -0.821    -4.775    0.001     -1.078     -0.398
## -------------------------------------------------------------------------------------------
##
##
##
## Forward Selection: Step 2
##
## - X1
##
##                        Model Summary
## -------------------------------------------------------------
## R                       0.986       RMSE               2.734
## R-Squared               0.972       Coef. Var          2.865
## Adj. R-Squared          0.967       MSE                7.476
## Pred R-Squared          0.955       MAE                2.015
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ---------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    2641.001         2       1320.500    176.627    0.0000
## Residual        74.762        10          7.476
## Total         2715.763        12
## ---------------------------------------------------------------------
##
##                                     Parameter Estimates
## -------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta       t        Sig      lower      upper
## -------------------------------------------------------------------------------------------
## (Intercept)    103.097         2.124                  48.540    0.000    98.365    107.830
##          X4     -0.614         0.049       -0.683    -12.621    0.000    -0.722     -0.506
##          X1      1.440         0.138        0.563     10.403    0.000     1.132      1.748
## -------------------------------------------------------------------------------------------
##
##
##
## Forward Selection: Step 3
##
## - X2
##
##                        Model Summary
## -------------------------------------------------------------
## R                       0.991       RMSE               2.309
## R-Squared               0.982       Coef. Var          2.419
## Adj. R-Squared          0.976       MSE                5.330
## Pred R-Squared          0.969       MAE                1.606
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    2667.790         3        889.263    166.832    0.0000
## Residual        47.973         9          5.330
## Total         2715.763        12
## ---------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## -----------------------------------------------------------------------------------------
## (Intercept)    71.648        14.142                  5.066    0.001    39.656    103.641
##          X4    -0.237         0.173       -0.263    -1.365    0.205    -0.629      0.155
##          X1     1.452         0.117        0.568    12.410    0.000     1.187      1.717
##          X2     0.416         0.186        0.430     2.242    0.052    -0.004      0.836
## -----------------------------------------------------------------------------------------
##
##
##
## No more variables to be added.
##
## Variables Entered:
##
## + X4
## + X1
## + X2
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.991       RMSE               2.309
## R-Squared               0.982       Coef. Var          2.419
## Adj. R-Squared          0.976       MSE                5.330
## Pred R-Squared          0.969       MAE                1.606
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                Sum of
##                 Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    2667.790         3        889.263    166.832    0.0000
## Residual        47.973         9          5.330
## Total         2715.763        12
## ---------------------------------------------------------------------
##
##                                     Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## -----------------------------------------------------------------------------------------
## (Intercept)    71.648        14.142                  5.066    0.001    39.656    103.641
##          X4    -0.237         0.173       -0.263    -1.365    0.205    -0.629      0.155
##          X1     1.452         0.117        0.568    12.410    0.000     1.187      1.717
##          X2     0.416         0.186        0.430     2.242    0.052    -0.004      0.836
## -----------------------------------------------------------------------------------------
```

```
##
##                              Selection Summary
## -------------------------------------------------------------------------
##         Variable                  Adj.
## Step    Entered     R-Square    R-Square      C(p)        AIC        RMSE
## -------------------------------------------------------------------------
##    1    X4            0.6745      0.6450    138.7308    97.7440    8.9639
##    2    X1            0.9725      0.9670      5.4959    67.6341    2.7343
##    3    X2            0.9823      0.9764      3.0182    63.8663    2.3087
## -------------------------------------------------------------------------
```

We have seen that the forward selection method selected the variables X_1, X_2, and X_4 in the model. It selected X_4 first, then X_1, finally X_2. This is the same as the best model selected using all possible regressions. Again, there is multicollinearity issues if this model is selected.

Now, we run backward elimination:

```
ols_step_backward_p(model, details = TRUE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . X1
## 2 . X2
## 3 . X3
## 4 . X4
##
## We are eliminating variables based on p value...
##
## - X3
##
## Backward Elimination: Step 1
##
##  Variable X3 Removed
##
```

```
##                         Model Summary
## -------------------------------------------------------------
## R                       0.991       RMSE               2.309
## R-Squared               0.982       Coef. Var          2.419
## Adj. R-Squared          0.976       MSE                5.330
## Pred R-Squared          0.969       MAE                1.606
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    2667.790         3        889.263    166.832    0.0000
## Residual        47.973         9          5.330
## Total         2715.763        12
## ---------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## -----------------------------------------------------------------------------------------
## (Intercept)    71.648        14.142                  5.066    0.001    39.656    103.641
##          X1     1.452         0.117        0.568    12.410    0.000     1.187      1.717
##          X2     0.416         0.186        0.430     2.242    0.052    -0.004      0.836
##          X4    -0.237         0.173       -0.263    -1.365    0.205    -0.629      0.155
## -----------------------------------------------------------------------------------------
##
##
##
## No more variables satisfy the condition of p value = 0.3
##
##
## Variables Removed:
##
## - X3
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.991       RMSE               2.309
## R-Squared               0.982       Coef. Var          2.419
## Adj. R-Squared          0.976       MSE                5.330
## Pred R-Squared          0.969       MAE                1.606
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
```

```
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    2667.790         3        889.263    166.832    0.0000
## Residual        47.973         9          5.330
## Total         2715.763        12
## ---------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## -----------------------------------------------------------------------------------------
## (Intercept)    71.648        14.142                  5.066    0.001    39.656    103.641
##          X1     1.452         0.117        0.568    12.410    0.000     1.187      1.717
##          X2     0.416         0.186        0.430     2.242    0.052    -0.004      0.836
##          X4    -0.237         0.173       -0.263    -1.365    0.205    -0.629      0.155
## -----------------------------------------------------------------------------------------
```

```
##
##
##                           Elimination Summary
## -----------------------------------------------------------------------
##         Variable                  Adj.
## Step    Removed     R-Square    R-Square     C(p)       AIC       RMSE
## -----------------------------------------------------------------------
##    1    X3            0.9823      0.9764    3.0182    63.8663    2.3087
## -----------------------------------------------------------------------
```

We can see that backward elimination has eliminated the variable X_3 from the model, so we have that the method selected X_1, X_2, and X_4 in the model.

Finally, we run stepwise selection:

```
ols_step_both_p(model, details = TRUE)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. X1
## 2. X2
## 3. X3
## 4. X4
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - X4 added
```

```
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.821       RMSE                8.964
## R-Squared               0.675       Coef. Var           9.394
## Adj. R-Squared          0.645       MSE                80.352
## Pred R-Squared          0.560       MAE                 6.894
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression    1831.896         1       1831.896    22.799     6e-04
## Residual       883.867        11         80.352
## Total         2715.763        12
## --------------------------------------------------------------------
##
##                                     Parameter Estimates
## -------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta      t        Sig       lower      upper
## -------------------------------------------------------------------------------------------
## (Intercept)    117.568         5.262                 22.342    0.000    105.986    129.150
##          X4     -0.738         0.155       -0.821    -4.775    0.001     -1.078     -0.398
## -------------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 2
##
## - X1 added
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.986       RMSE               2.734
## R-Squared               0.972       Coef. Var          2.865
## Adj. R-Squared          0.967       MSE                7.476
## Pred R-Squared          0.955       MAE                2.015
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    2641.001         2       1320.500    176.627    0.0000
## Residual        74.762        10          7.476
## Total         2715.763        12
## ----------------------------------------------------------------------
##
##                                     Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta       t        Sig      lower      upper
## ------------------------------------------------------------------------------------------
## (Intercept)    103.097         2.124                  48.540    0.000    98.365    107.830
##          X4     -0.614         0.049       -0.683    -12.621    0.000    -0.722     -0.506
##          X1      1.440         0.138        0.563     10.403    0.000     1.132      1.748
## ------------------------------------------------------------------------------------------
##
##
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.986       RMSE               2.734
## R-Squared               0.972       Coef. Var          2.865
## Adj. R-Squared          0.967       MSE                7.476
## Pred R-Squared          0.955       MAE                2.015
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression    2641.001         2       1320.500    176.627    0.0000
## Residual        74.762        10          7.476
## Total         2715.763        12
## ----------------------------------------------------------------------
##
##                                     Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta       t        Sig      lower      upper
## ------------------------------------------------------------------------------------------
## (Intercept)    103.097         2.124                  48.540    0.000    98.365    107.830
##          X4     -0.614         0.049       -0.683    -12.621    0.000    -0.722     -0.506
##          X1      1.440         0.138        0.563     10.403    0.000     1.132      1.748
## ------------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 3
##
## - X2 added
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.991       RMSE               2.309
## R-Squared               0.982       Coef. Var          2.419
## Adj. R-Squared          0.976       MSE                5.330
## Pred R-Squared          0.969       MAE                1.606
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    2667.790         3        889.263    166.832    0.0000
## Residual        47.973         9          5.330
## Total         2715.763        12
## ---------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## -----------------------------------------------------------------------------------------
## (Intercept)    71.648        14.142                  5.066    0.001    39.656    103.641
##          X4    -0.237         0.173       -0.263    -1.365    0.205    -0.629      0.155
##          X1     1.452         0.117        0.568    12.410    0.000     1.187      1.717
##          X2     0.416         0.186        0.430     2.242    0.052    -0.004      0.836
## -----------------------------------------------------------------------------------------
##
##
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.991       RMSE               2.309
## R-Squared               0.982       Coef. Var          2.419
## Adj. R-Squared          0.976       MSE                5.330
## Pred R-Squared          0.969       MAE                1.606
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ---------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression    2667.790         3        889.263    166.832    0.0000
## Residual        47.973         9          5.330
## Total         2715.763        12
## ---------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## -----------------------------------------------------------------------------------------
## (Intercept)    71.648        14.142                  5.066    0.001    39.656    103.641
##          X4    -0.237         0.173       -0.263    -1.365    0.205    -0.629      0.155
##          X1     1.452         0.117        0.568    12.410    0.000     1.187      1.717
##          X2     0.416         0.186        0.430     2.242    0.052    -0.004      0.836
## -----------------------------------------------------------------------------------------
##
##
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.991       RMSE               2.309
## R-Squared               0.982       Coef. Var          2.419
## Adj. R-Squared          0.976       MSE                5.330
## Pred R-Squared          0.969       MAE                1.606
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression    2667.790         3        889.263    166.832    0.0000
## Residual        47.973         9          5.330
## Total         2715.763        12
## ----------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## -----------------------------------------------------------------------------------------
## (Intercept)    71.648        14.142                  5.066    0.001    39.656    103.641
##          X4    -0.237         0.173       -0.263    -1.365    0.205    -0.629      0.155
##          X1     1.452         0.117        0.568    12.410    0.000     1.187      1.717
##          X2     0.416         0.186        0.430     2.242    0.052    -0.004      0.836
## -----------------------------------------------------------------------------------------
```

```
##
##                              Stepwise Selection Summary
## -------------------------------------------------------------------------------------
##                       Added/                   Adj.
## Step    Variable    Removed     R-Square    R-Square      C(p)        AIC       RMSE
## -------------------------------------------------------------------------------------
##    1       X4       addition       0.675       0.645    138.7310    97.7440    8.9639
##    2       X1       addition       0.972       0.967      5.4960    67.6341    2.7343
##    3       X2       addition       0.982       0.976      3.0180    63.8663    2.3087
## ---------------------------------------------------------------------------------
```

We see that stepwise selection has added the variables X_1, X_2, and X_4. It didn't eliminate any variables. Therefore, it also selected X_1, X_2, and X_4 to be used in the model.

11 Logistic Regression

In all the sections above, we talked about ordinary linear regression, where we assume that the model is linear in the parameters. In this section, we consider the case where the response variable Y_i is binary and the appropriate model and inferences.

11.1 Logistic Regression Model

Sometimes the response variable is discrete. For example, we may wish to model gender or to estimate the likelihood that a person is wearing a life jacket.

We first consider the linear regression model

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

where

$$Y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i \end{cases}$$

Then we have $E[Y_i] = \pi_i$ since $Y_i \sim Ber(\pi_i)$.

In this case, the usual least squares fitting approach is problematic because the assumptions are no longer reasonable:

- the variance of Y_i, where $Var[Y_i] = \pi_i(1 - \pi_i)$ is not constant since it changes based on π_i
- the error terms are not normally distributed (since Y_i follow a Bernoulli distribution)
- there is no guarantee that the fitted model will force the estimate $\hat{Y_i}$ to be in the interval $(0, 1)$

The logistic distribution has density

$$f(x) = \frac{e^x}{(1 + e^x)^2}, \; -\infty < x < \infty$$

and cumulative distribution function

$$F(t) = \frac{e^t}{1 + e^t}$$

We can show $E[X] = 0$ and $Var[X] = \frac{\pi}{3}$.

Suppose that a random variable Y is binary with

$$Y_i = \begin{cases} 1 & \text{if } \beta_0^* + \beta_1^* X_i + \epsilon_i^* < k \\ 0 & \text{if } \beta_0^* + \beta_1^* X_i + \epsilon_i^* > k \end{cases}$$

for some constant k where ϵ_i^* has a logistic distribution.

Then we get

$$
\begin{aligned}
\pi_i &= P(Y_i = 1) \\
&= P(\beta_0^* + \beta_1^* X_i + \epsilon_i^* < k) \\
&= F(k - \beta_0^* - \beta_1^* X_i) \\
&= F(\beta_0 + \beta_1 X_i) \\
&= \frac{exp(\beta_0 + \beta_1 X_i)}{(1 + exp(\beta_0 + \beta_1 X_i))}
\end{aligned}
$$

where $\beta_0 = k - \beta_0^*$ and $\beta_1 = -\beta_1^*$

So we get that

$$E[Y_i] = \pi_i = \frac{exp(\beta_0 + \beta_1 X_i)}{(1 + exp(\beta_0 + \beta_1 X_i))}$$

. This is the structure of the logistic regression model.

It is common practice to model the logarithm of the odds ratio

$$\log(\frac{\pi_i}{1 - \pi_i}) = \log(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}) = \beta_0 + \beta_1 X_i$$

The estimation of the parameters is based on maximizing the likelihood:

$$
\begin{aligned}
\prod_i f(y_i) &= \prod_i \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} \\
&= \prod_i [(\frac{\pi_i}{1 - \pi_i})^{y_i} (1 - \pi_i)]
\end{aligned}
$$

Then we take the log likelihood.

$$logL = \sum y_i \log(\frac{\pi_i}{1 - \pi_i}) + \sum \log(1 - \pi_i) = \sum y_i(\beta_0 + \beta_1 X_i) - \sum \log(1 + exp(\beta_0 + \beta_1 X_i))$$

There is no closed form solution. Instead, numerical methods are used to obtain a solution for β_0 and β_1, denoted by b_0 and b_1 and we get

$$\hat{\pi_i} = \frac{exp(b_0 + b_1 X_i)}{(1 + exp(b_0 + b_1 X_i))}$$

To interpret the parameters in the logistic regression model, let us consider the fitted value at a specific value of X, say X_0. Then the difference between the log odds at $X_0 + 1$ and the log odds at X_0 is

$$logodds(X_0 + 1) - logodds(X_0) = \hat{\beta_1}$$

meaning that for every unit increase in X_0, the log odds $\log(\frac{\pi_i}{1 - \pi_i})$ at X_0 increases by an average of $\hat{\beta_1}$ units.

Taking the antilogarithms, we obtain the odds ratio

$$\hat{O}_R = e^{\hat{\beta_1}}$$

The odds ratio is the estimated increase in the probability of successes associated with a one unit change in the value of the predictor variable. For a change of d units, the odds ratio becomes

$$\hat{O}_R = e^{d\hat{\beta_1}}$$

11.2 Repeat Observations

In the section above, we looked at when we took individual observations at each X value. Now, we look at repeat observations. Suppose that we have repeat observations at each of the leels of the X variables and set Y_i to be the number of 1s observed for the i-th observation. Let n_i be the number of trials at each observation. Then $Y_i \sim Bin(n_i, \pi_i)$. In that case, estimation is done by maximizing

$$\begin{aligned}\log L(\beta_0, \beta_1) &= \log \prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \\ &= \sum_{i=1}^{n} (\log(\binom{n}{y_i})) + y_i(\log(\pi_i)) + (n_i - y_i)\log(1-\pi_i))\end{aligned}$$

11.3 Multiple Logistic Models

We can also involve multiple predictors in the logistic regression model. Specifically, we have

$$\boldsymbol{X^T\beta} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

and

$$E[Y] = \frac{exp(\boldsymbol{X^T\beta})}{1 + exp(\boldsymbol{X^T\beta})}$$

so that

$$\log(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}) = \boldsymbol{X^T\beta}$$

11.4 Inference on model parameters

Now, we talk about inference on the parameters of the model. There are several diagnostics that we can use to make inferences. The maximum likelihood estimators are approximately normally distributed with variances and covariances that are functions of the second order partial derivatives of the likelihood function. Let $\boldsymbol{G} = (\frac{\partial^2 L(\beta)}{\partial\beta_i \partial\beta_j}) \equiv (g_{ij})$ lebeled the Hessian where

$$\log L(\beta) = \sum_{i=1}^{n} Y_i(X_i^T \beta) - \sum_{i=1}^{n} \log(1 + e^{X_i^T \beta})$$

The linear predictor is $\boldsymbol{X^T\hat{\beta}}$ and the fitted value is $\hat{Y}_i = \hat{\pi}_i = \frac{exp(X_i^T\hat{\beta})}{(1+exp(X_i^T\hat{\beta}))}$.

It can be shown that

$$E[\boldsymbol{b}] = \boldsymbol{\beta}$$

The variance estimate is given by

$$Var[\boldsymbol{b}] = (\boldsymbol{X^T V X})^{-1}$$

where $\boldsymbol{V}$ is a diagonal matrix with $V_{ii} = n_i\hat{\pi}_i(1-\hat{\pi}_i)$. Moreover, we have that

$$\frac{b_k - \beta_k}{s[b_k]} \sim N(0,1)$$

for $k = 0, ..., p-1$, which is used for testing and constructing confidence intervals. The test statistic to test $H_0 : \beta_k = 0$ is

$$\frac{b_k}{s[b_k]} \sim N(0,1)$$

and a confidence interval is

$$b_k \pm z_{\alpha/2} \cdot s[b_k]$$