Regression Analysis Examples in R

Chapter 2 - Simple Linear Regression

Example. Airfreight Data

	1	2	3	4	5	6	7	8	9	10
Shipment Route (x)	1	0	2	0	3	1	0	1	2	0
Airfreight Breakage (y)	16	9	17	12	22	13	8	15	19	11

a) Compute the ANOVA table
b) Compute the confidence intervals for the parameters
c) Compute the confidence interval on the average (mean) response when $X = 1$.
d) What is the total variability in y explained by this model?

Solution.

Part a.

We can compute the anova table manually as follows,

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 20 - \frac{1}{10}(100) = 10$$

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = 182 - \frac{1}{10}(10)(142) = 40$$

Therefore,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{40}{10} = 4$$

Then, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, so

$$\hat{\beta}_0 = \frac{1}{10}(142) - 4 \cdot \frac{1}{10}(10) = 10.2$$

This gives us our linear model

$$\hat{y} = 10.2 + 4x$$

The sum of squares for regression is

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx} = 16 \cdot 10 = 160$$

The total sum of squares is

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 2194 - 10(14.2)^2 = 177.6$$

Then, the residual sum of squares is

$$SSE = SST - SSR = 177.6 - 160 = 17.6$$

Now we can construct the anova table

Source	Sum of Squares	DF	MS=SS/df	F = MSR/MSE
Regression	160	1	160	72.727
Error	17.6	8	2.2	
Total	177.6			

We conclude tbat the regression is highly significant since the F value is very large. We can also do this in R

```
x <- c(1,0,2,0,3,1,0,1,2,0)
y <- c(16,9,17,12,22,13,8,15,19,11)
model <- lm(formula = y ~ x)
print(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##        10.2          4.0
```

```
anova(model)
```

```
## 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 can see that we get the same results and reach the same conclusion, and we also get the p-value which is very small and we can conclude from there as well that the regression is highly significant.

Part b.

We can construct confidence intervals, first we need to compute $se(\hat{\beta}_1)$ and $se(\hat{\beta}_0)$.

$$se^2(\hat{\beta}_0) = MSE\left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}}\right) = 2.2\left(\frac{1}{10} + \frac{1}{10}\right) = 0.44 \implies se(\hat{\beta}_0) = \sqrt{0.44} = 0.6633$$

$$se^2(\hat{\beta}_1) = \frac{MSE}{S_{xx}} = \frac{2.2}{10} = 0.22 \implies se(\hat{\beta}_1) = \sqrt{0.22} = 0.490$$

Then, we have to compute $t_{\alpha/2,n-2} = t_{0.025,8}$, we either use a t look up table or in R,

```r
qt(0.025, 8, lower.tail=FALSE)
```

```
## [1] 2.306004
```

Thus, our confidence intervals are

$$\hat{\beta}_0 - t_{\alpha/2,n-2}se(\hat{\beta}_0) \leq \hat{\beta}_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}se(\hat{\beta}_0) \rightarrow 10.2 \pm 2.306(0.6633) = (8.6704, 11.7296)$$

$$\hat{\beta}_1 - t_{\alpha/2,n-2}se(\hat{\beta}_1) \leq \hat{\beta}_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}se(\hat{\beta}_1) \rightarrow 4 \pm 2.306(0.490) = (2.9392, 5.0608)$$

We can compute these confidence intervals in R as well

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

```
##                2.5 %    97.5 %
## (Intercept) 8.670370 11.729630
## x           2.918388  5.081612
```

Part c.

We want to compute first $E(y|x_0)$, where $x_0 = 1$. An unbiased estimator for $E(y|x_0)$ is

$$\widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 10.2 + 4(1) = 14.2$$

Then, the confidence interval is

$$\left[\hat{\mu}_{y|x_0} \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right] = \left[14.2 \pm 2.306\sqrt{2.2\left(\frac{1}{10} + \frac{(1-1)^2}{S_{xx}}\right)}\right] = (13.11839, 15.28161)$$

We can do this in R with

```r
predict(model, newdata = data.frame(x=1), interval = 'confidence', level=0.95)
```

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

Part d.

The total variability in y explained by the regressor x is measured by the coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{160}{177.6} = 0.9009$$

We can also see this in the summary of the model in R

```r
summary(model)
```

```
##
## 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
```

The R^2 value is 0.9009.

Chatper 3 - Multiple Linear Regression

Question 3.1

Consider the National Football League data in Table B.1

a. Fit a multiple linear regression model relating the number of games won to the team's passing yardage (x_2), the percentage of rushing plays (x_7), and the opponents' yards rushing (x_8).
b. Construct the analysis-of-variance table and test for significance of regression.
c. Calculate t statistics for the hypotheses $H_0 : \beta_2 = 0$, $H_0 : \beta_7 = 0$, and $H_0 : \beta_8 = 0$. What conclusions can you draw about the roles of variables in x_2, x_7, and x_8 play in the model?
d. Calculate R^2 and R^2_{Adj} for this model.
e. Using the partial F test, determine the contribution of x_7 to the model. How is the partial F statistic related to the t test for β_7 calculated in part c above?

Question 3.3

Refer to problem 3.1

a. Find a 95% CI for β_7.

b. Find a 95% CI on the mean number of games won by a team when $x_2 = 2300$, $x_7 = 56$, and $x_8 = 2100$.

Question 3.4

Reconsider the National Football League data from Problem 3.1. Fit a model to these data using only x_7 and x_8 as regressors.

a. Test for significance of regression.

b. Calculate R^2 and R^2_{Adj}. How do these quantities compare to the value computed for the model in Problem 3.1, which included an additional regressor (x_2)?

c. Calculate a 95% CI on β_7. Also find a 95% CI on the mean number of games won by a team when $x_7 = 56$, and $x_8 = 2100$. Compare the lengths of these CIs to the lengths of the corresponding CIs from Problem 3.3.

d. What conclusions can you draw from this problem about the consequences of omitting an important regressor from a model?

Solutions.

Question 3.1

Part a.

We can fit a linear model using the same R function,

```
# Table b1 was loaded ahead of time.
model <- lm(formula = y ~ x2 + x7 + x8, tableb1)
model
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = tableb1)
##
## Coefficients:
## (Intercept)           x2           x7           x8
##   -1.808372     0.003598     0.193960    -0.004815
```

This gives us our linear model with the estimates $\hat{\beta}_0 = -1.808$, $\hat{\beta}_2 = 0.00360$, $\hat{\beta}_7 = 0.194$, and $\hat{\beta}_8 = -0.00482$.

$$y = -1.808 + 0.00360x_2 + 0.194x_7 - 0.00482x_8$$

Part b.

We test for significance of regession using the hypotheses

$$H_0 : \beta_2 = \beta_7 = \beta_8 = 0, \;\; H_1 : \beta_j \neq 0, j = 2, 7, 8$$

We use the F-statistic

$$F_0 = \frac{SSR/k}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{k,n-p}$$

We reject the null hypothesis when $F > F_{\alpha,k,n-k-1}$, we compute these values in R. In this case, we have $k = 3$ regressors and coefficients, so $p = k + 1 = 4$, thus

```
n <- nrow(tableb1)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x2         1  76.193  76.193  26.172 3.100e-05 ***
## x7         1 139.501 139.501  47.918 3.698e-07 ***
## x8         1  41.400  41.400  14.221 0.0009378 ***
## Residuals 24  69.870   2.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
qf(0.05, 3, n-4, lower.tail=FALSE)
```

```
## [1] 3.008787
```

Source	Sum of Squares	DF	MS	F	P
x_2	76.193	1	76.193	26.172	$3.1 \cdot 10^{-5}$
x_7	139.501	1	139.501	47.918	$3.698 \cdot 10^{-7}$
x_8	41.400	1	41.400	14.221	0.0009378
Residuals	69.8790	24	2.911		

$$F_0 = \frac{SSR/k}{SSE/(n-p)} = \frac{(76.193 + 139.501 + 41.4)/3}{69.870/24} = 29.439$$

We can also obtain the F-statistic from the summary of the model,

```r
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = tableb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
## x7           0.193960   0.088233   2.198 0.037815 *
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

Therefore, we reject the null hypothesis $H_0 : \beta_2 = \beta_7 = \beta_8 = 0$, and conclude our regression is significant.

Part c.

We want to conduct tests on the indivudal coefficients, with the hypotheses $H_0 : \beta_2 = 0$, $H_0 : \beta_7 = 0$, and $H_0 : \beta_8 = 0$. We need the t-statistic

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

We reject the null hypothesis when $|t_0| > t_{\alpha/2, n-p}$,

```r
qt(0.025, n-4, lower.tail=FALSE)
```

```
## [1] 2.063899
```

We have all the information we need however in the summary of the model, we can see the estimate and the standard error for each coefficient, which tells us the t-value, but also the t-value is included. We can see that for all coefficents and their respective t-values, $|t_0| > t_{\alpha/2,n-p}$ so we reject all 3 of the null hypotheses.

Part d.

The R^2 value can be computed with

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{76.193 + 139.501 + 41.4}{76.193 + 139.501 + 41.4 + 69.8790} = 0.7863$$

We get these values from the ANOVA table above and also in the summaryof our model, we have $R^2 = 0.7863$, and the adjusted R^2 value is

$$R^2_{Adj} = 1 - \frac{SSE/(n-k)}{SST/(n-1)} = 1 - \frac{68.8790/24}{326.973/27} = 0.7596$$

This value is also in the summary R output above.

Part e.

To conduct the partial F test to determine the contribution of x_7, we want to tes tthe hypotheses

$$H_0 : \beta_7 = 0, \ H_1 : \beta_7 \neq 0$$

We fit the model assuming the null hypothesis is true to get the reduced model and obtain the anova table,

```r
reduced_model <- lm(formula = y ~ x8 + x2, data=tableb1)
anova(reduced_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x8         1 178.092 178.092  53.043 1.245e-07 ***
## x2         1  64.934  64.934  19.340 0.0001775 ***
## Residuals 25  83.938   3.358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then, we want to use the F statistic to test the hypotheses

$$F_0 = \frac{SSR(\beta_7|\beta_2, \beta_0)/r}{MSE}$$

So, we have

$$SRR(\beta_7|\beta_2, \beta_8) = SSR(\beta_7, \beta_2, \beta_8) - SSR(\beta_2, \beta_8) = 257.094 - (178.092 + 64.934) = 14.064$$

Therefore,

$$F_0 = \frac{SSR(\beta_7|\beta_2, \beta_8)}{MSE} = \frac{14.064}{2.911} \approx 4.831$$

we reject the null hypothesis if $F_0 > F_{\alpha,r,n-p}$,

```
qf(0.05, 1, 24, lower.tail=FALSE)
```

```
## [1] 4.259677
```

We conclude that x_7 contributed significantly to this model. We also notice that the F statistic is the square of the t test used in part c.

Question 3.3

Part a.

To consturct a confidence interval on β_7, we need to compute

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

We have the standard error for β_7 from the previous summary output from R, so $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} = 0.088233$. Then, the t-value was also obtained earlier and we have $t_{\alpha/2,n-p} = 2.063899$, and the coefficient estimator $\hat{\beta}_7 = 0.193960$. Therefore, the confidence interval is

$$0.193960 \pm 2.063899 \cdot (0.088233) = (0.011856, 0.376064)$$

This can be also obtained in R

```
confint(model, "x7")
```

```
##         2.5 %    97.5 %
## x7 0.01185532 0.3760651
```

Part b.

The mean response at $x_2 = 2300$, $x_7 = 56$ and $x_8 = 2100$ is

$$y_0 = \beta_0 + \beta_2(2300) + \beta_7(56) + \beta_8(2100) = 7.215188$$

Then, the confidence interval on the mean response is

$$\left[\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0}\right]$$

This can be calculated in R with

```
predict(model, newdata=data.frame(x2 = 2300, x7 = 56, x8 = 2100),
        interval = 'confidence', level=0.95)
```

```
##        fit      lwr      upr
## 1 7.216424 6.436203 7.996645
```

Question 3.4

Part a

We want to test the hypotheses for significance,

$$H_0 : \beta_7 = \beta_8 = 0,\ H_1 : \beta_j \neq 0,\ j = 7, 8$$

We can fit the model and get the anova table,

```
model <- lm(formula = y ~ x7 + x8, tableb1)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x7         1  97.238  97.238  16.437 0.000431 ***
## x8         1  81.828  81.828  13.832 0.001015 **
## Residuals 25 147.898   5.916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the anova table that the regression is highly significant, but we can also use the F test statistic,

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} = \frac{(97.238 + 81.828)/2}{147.898/25} = \frac{89.533}{5.916} = 15.134$$

Then, we can compute $F_{\alpha,k,n-k-1}$,

```
qf(0.05, 3, 25, lower.tail=FALSE)
```

```
## [1] 2.991241
```

We can see that $F_0 > F_{\alpha,k,n-k-1}$ so we reject the null hypotheses that $H_0 : \beta_7 = \beta_8 = 0$, and conclude that the regression is significant.

Part b.

We can obtain a summary for the model and look at the R^2 and R^2_{Adj} values similar to the previous questions.

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x7 + x8, data = tableb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7985 -1.5166 -0.5792  1.9927  4.5248
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.944319   9.862484   1.819  0.08084 .
## x7           0.048371   0.119219   0.406  0.68839
## x8          -0.006537   0.001758  -3.719  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.432 on 25 degrees of freedom
## Multiple R-squared:  0.5477, Adjusted R-squared:  0.5115
## F-statistic: 15.13 on 2 and 25 DF,  p-value: 4.935e-05
```

We get an R-squared value $R^2 = 0.5477$ and the adjusted R-squared is $R^2_{Adj} = 0.5115$, we can see that these values are lower than when we had x_2 in the model. So, the model with x_2 was able to better explain the variablity in y and this suggests x_2 may have been contributing significantly to the model.

Part c.

```
confint(model, "x7")
```

```
##         2.5 %   97.5 %
## x7 -0.1971643 0.293906
```

A 95% confidence interval on β_7 is

$$(-0.1971643, 0.293906)$$

```
new_data <- data.frame(x7 = 56, x8=2100)
predict(model, newdata=new_data, interval='confidence', level=0.95)
```

```
##        fit      lwr      upr
## 1 6.926243 5.828643 8.023842
```

Our 95% confidence interval on the mean number of games one when $x_7 = 56$ and $x_8 = 2100$ is

$$(5.828643, 8.023842)$$

We can see that the length of both confidence intervals are greater than when x_2 was included in the model. This suggests we were more confident with our estimates when x_2 was included.

d.

We can conclude that omitting an important regressor (x_2) affected our estimates and standard error of coeffficients, resulting in larger lenghts in the confidence intervals and lower values for R^2.

Chapter 4 - Model Adequecy

Question 4.1

Consider the simple regression model fit to the National Football League team performance data in Problem 2.1. (Same data as previous questions, with the model y $x8$.

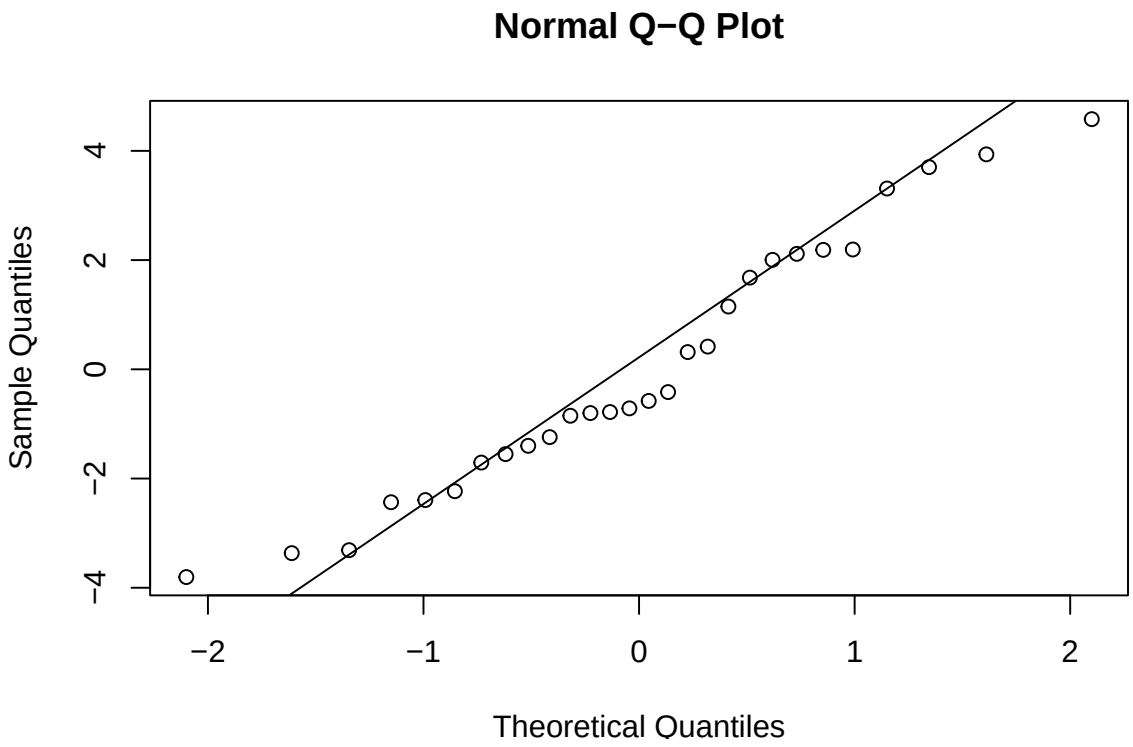
a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?

b. Construct an interpret a plot of the residuals versus the predicted repsonse.

c. Plot the residuals versus the team passing yardage, x_2. DOes this plot indicate that the model will be improved by adding x_2 to the model?

Part a.

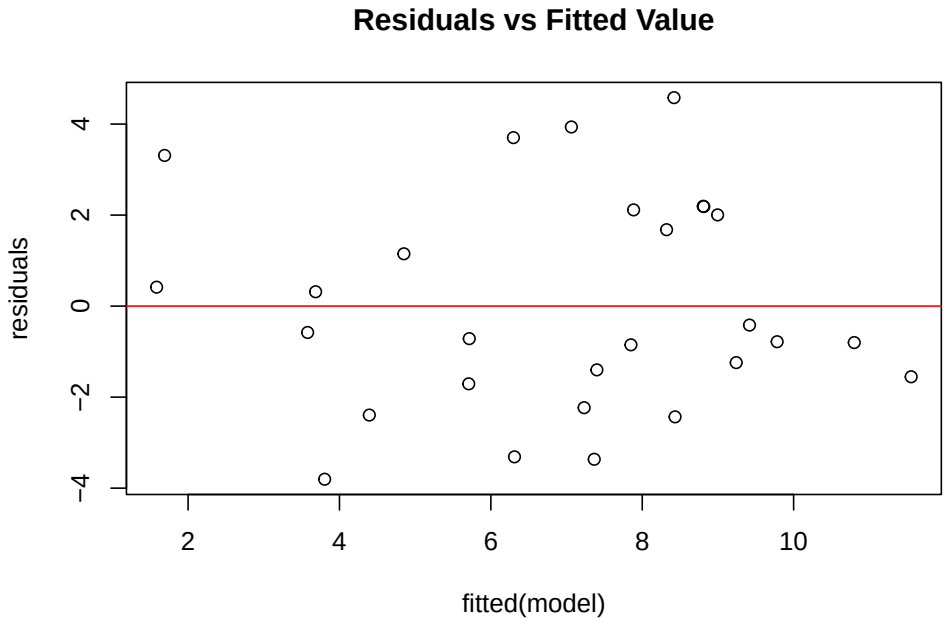
```
library(ggplot2)
model <- lm(formula = y ~ x8, tableb1)
residuals <- residuals(model)
qqnorm(residuals)
qqline(residuals)
```



It appears that the normality assumption is fine since the standardized residuals closely follow the theoretical quantiles for a normal distribution as observered in the quantile-quantile plot.

Part b.

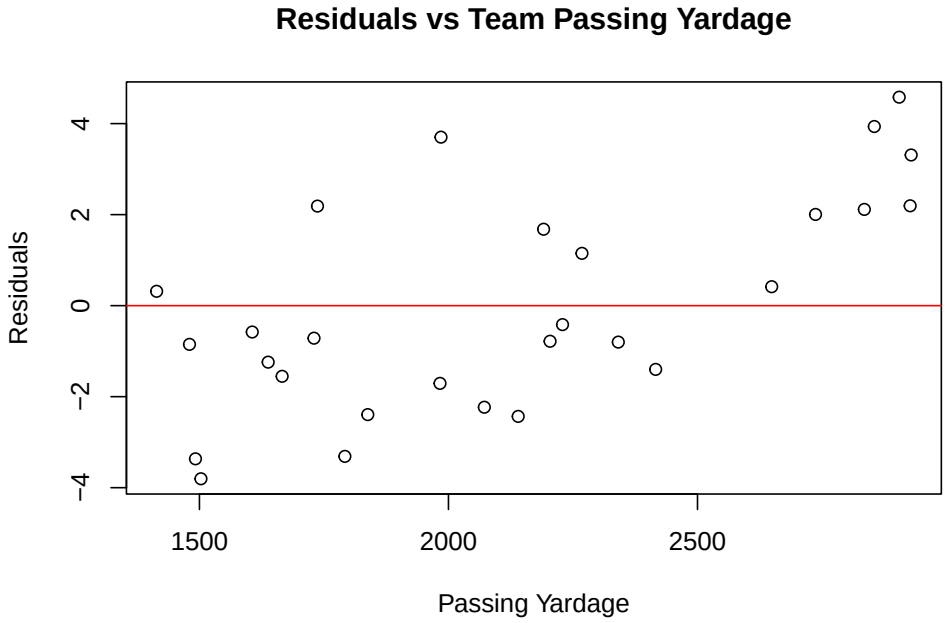
```
plot(fitted(model), residuals, main = "Residuals vs Fitted Value")
abline(h=0, col = 'red')
```



The model appears to be adequate since the residuals vs fitted values appear to be randomly distributed around 0, showing no patterns.

Part c.

```
residuals <- residuals(model)
plot(tableb1$x2, residuals, xlab="Passing Yardage",
     ylab="Residuals", main="Residuals vs Team Passing Yardage")
abline(h=0, col="red")
```



The model appears to be improved when adding x_2 since the residuals appear to be more close to 0.