# STAT4355HW5

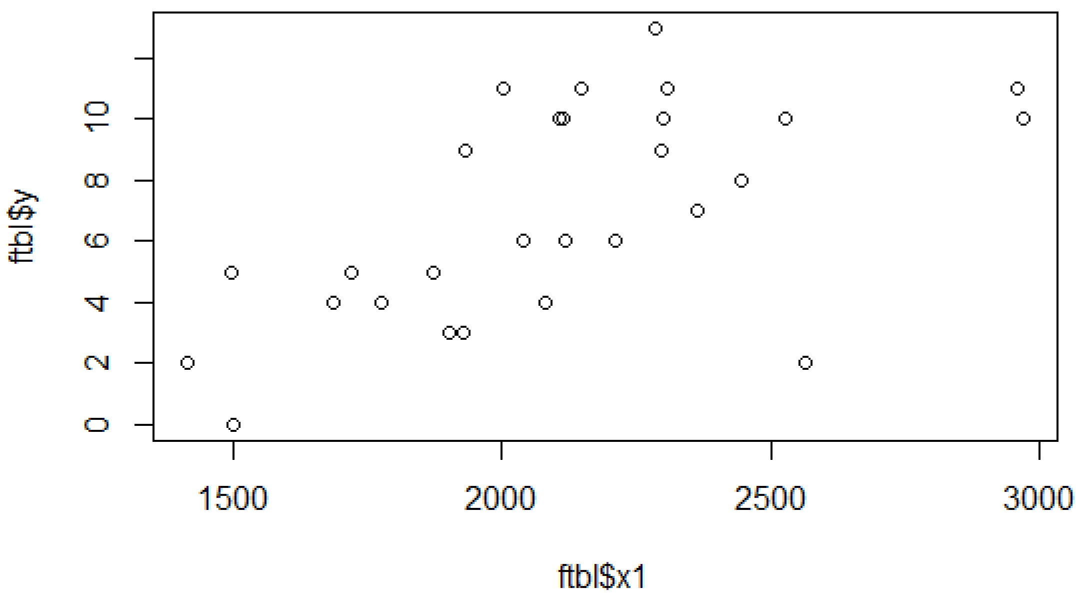
Code

## This is an R script with the purpose of running multiple linear regression on football data

## (a)

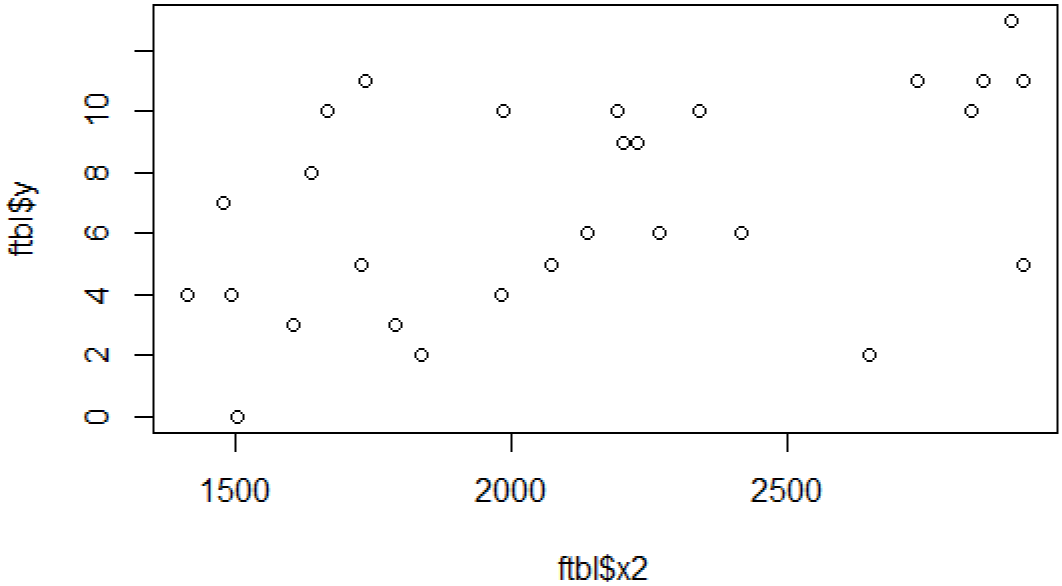
Hide

```
#load the data
ftbl <- read.csv(file = 'football.csv')
#9 scatterplots for number of games won (y) against other vars
plot(ftbl$x1, ftbl$y)
```



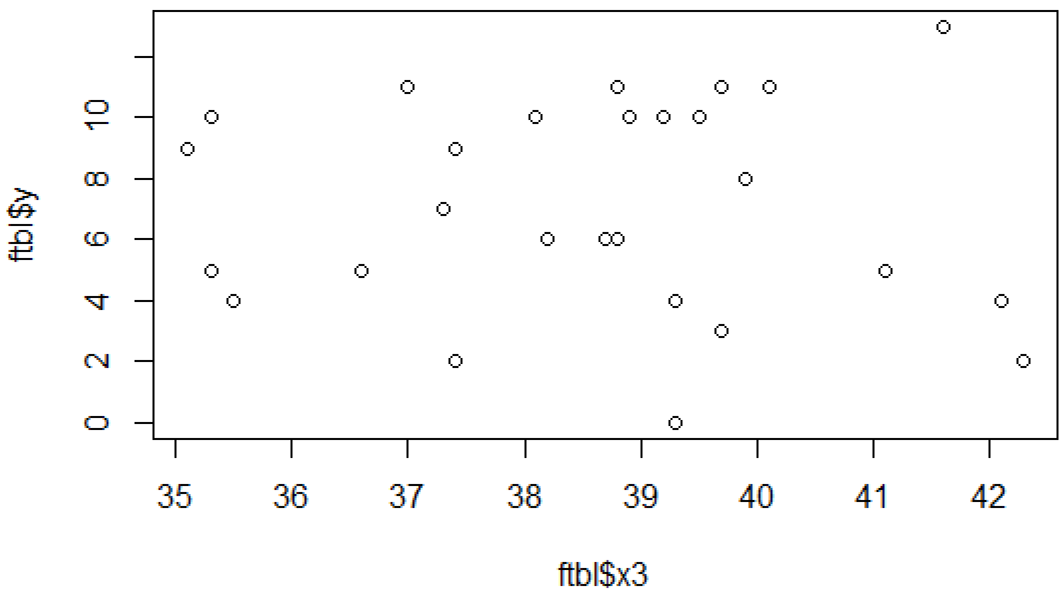

Hide

```
plot(ftbl$x2, ftbl$y)
```



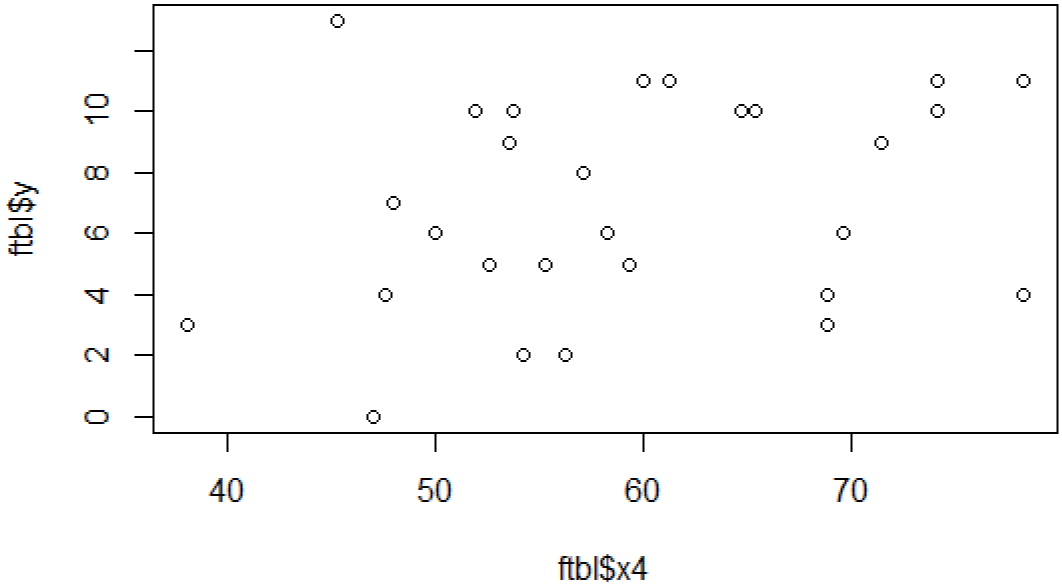

Hide

```
plot(ftbl$x3, ftbl$y)
```



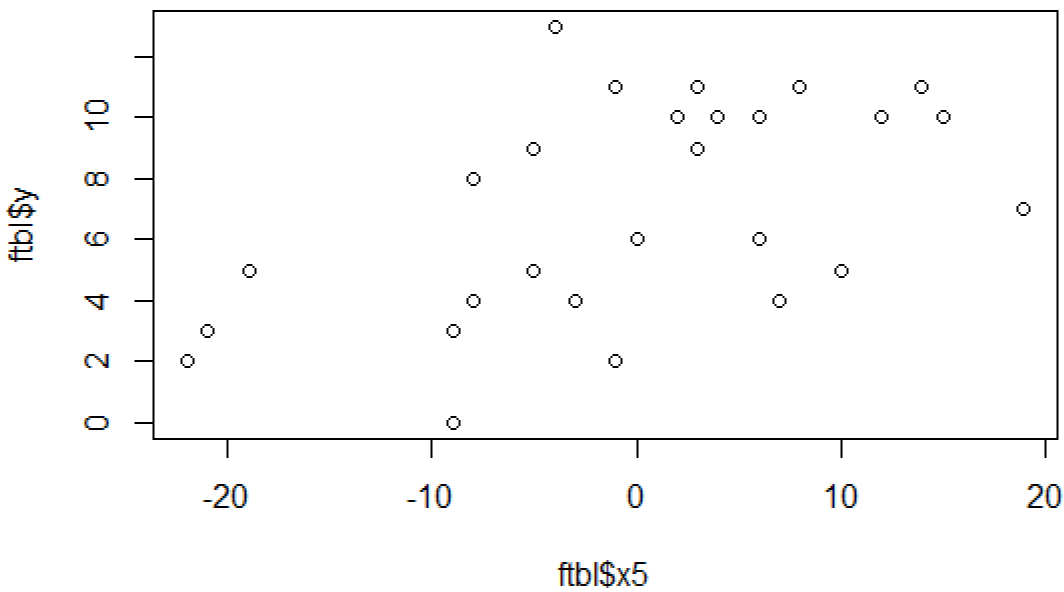

Hide

```
plot(ftbl$x4, ftbl$y)
```



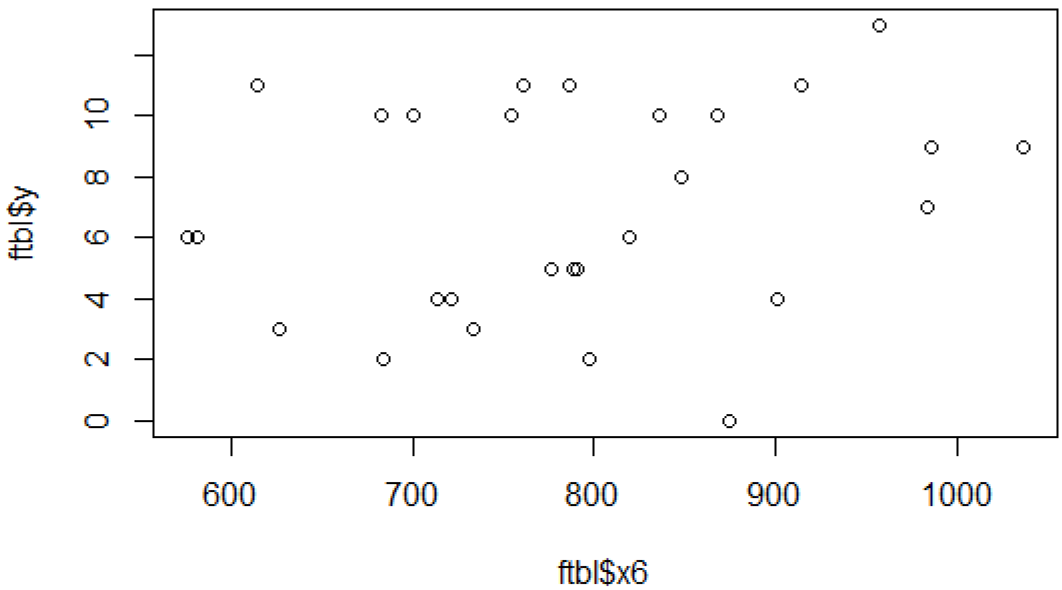

Hide

```
plot(ftbl$x5, ftbl$y)
```



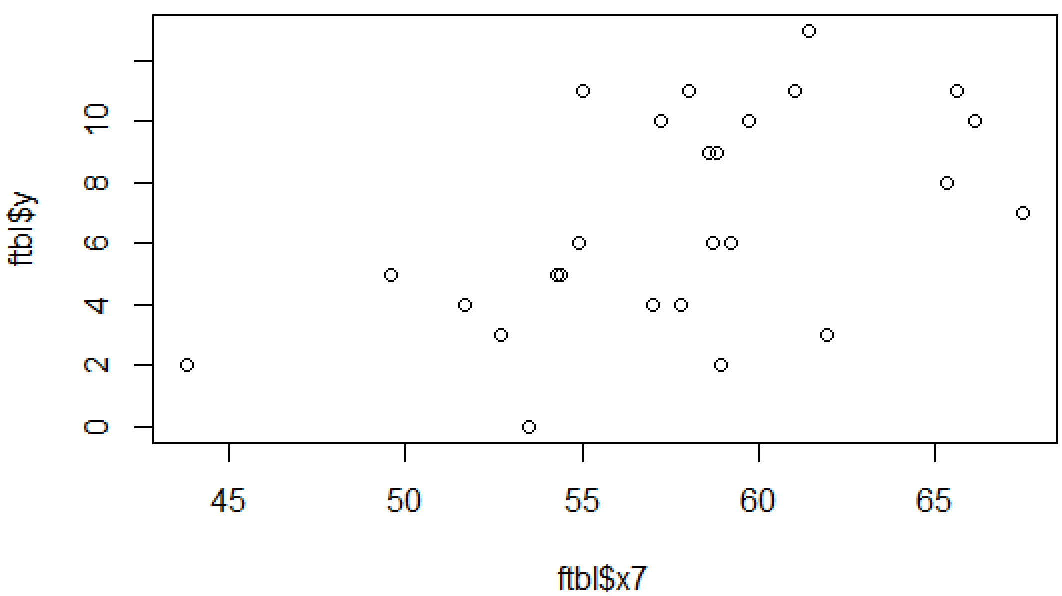

Hide

```
plot(ftbl$x6, ftbl$y)
```



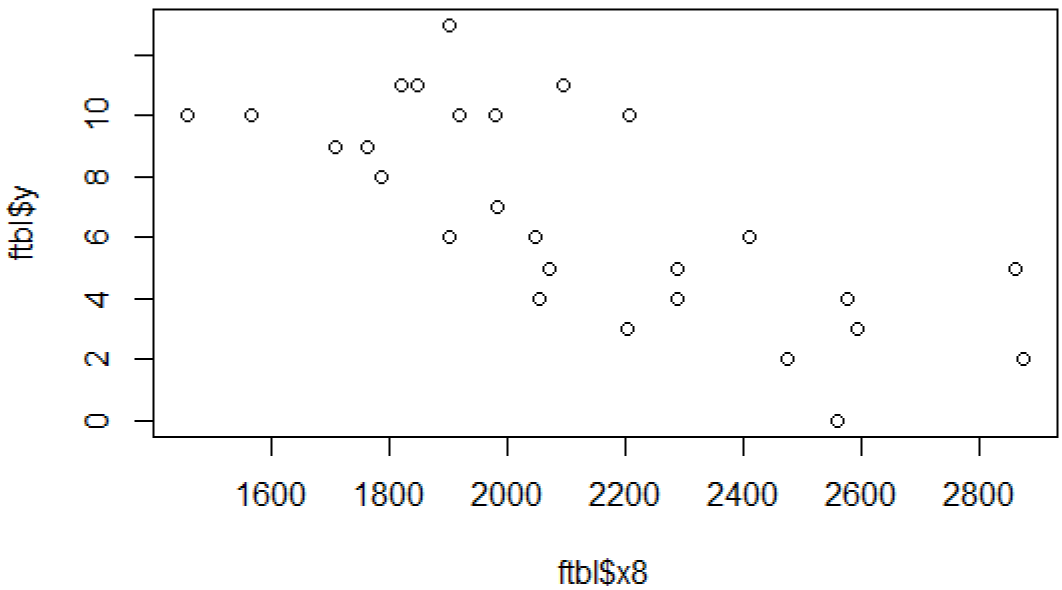

Hide

```
plot(ftbl$x7, ftbl$y)
```



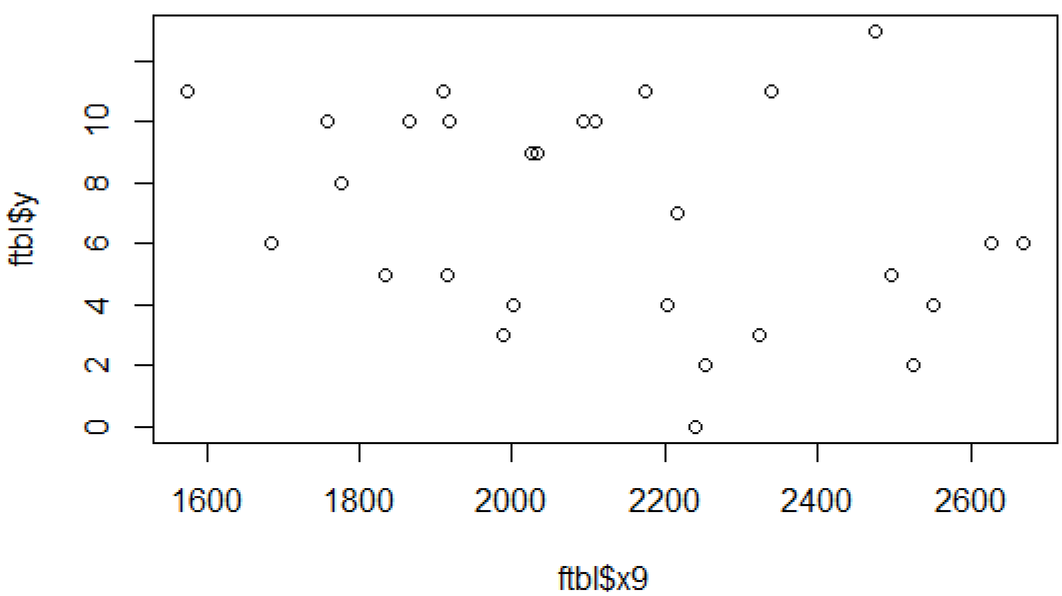

Hide

```
plot(ftbl$x8, ftbl$y)
```




Hide

```
plot(ftbl$x9, ftbl$y)
```




# (b)

Hide

```
#build model with most associated vars
lm1 <- lm(y~x1+x2+x5+x7+x8, data = ftbl)
#model summary
summary(lm1)
```

```
Call:
lm(formula = y ~ x1 + x2 + x5 + x7 + x8, data = ftbl)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7152 -0.6710 -0.2079  1.2287  3.7252 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0516532  9.0751149  -0.006 0.995510    
x1           0.0008480  0.0017008   0.499 0.623008    
x2           0.0034689  0.0007673   4.521 0.000169 ***
x5           0.0103765  0.0440046   0.236 0.815766    
x7           0.1336928  0.1375471   0.972 0.341628    
x8          -0.0047048  0.0013412  -3.508 0.001986 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.768 on 22 degrees of freedom
Multiple R-squared:  0.7896,	Adjusted R-squared:  0.7418 
F-statistic: 16.51 on 5 and 22 DF,  p-value: 8.363e-07
```

Hide

```
sigma(lm1)^2
```

```
[1] 3.126997
```

Fitted linear model:

y-hat = -0.0516532 + 0.0008480x1 + 0.0034689x2 + 0.0103765x5 + 0.1336928x7 + -0.0047048x8

i. (σ^)2 = 3.126997
ii. R^2 = 0.7896
iii. adjusted R^2 = 0.7418

# (c)

βˆ1: For a one unit increase in rushing yardage, number of games won is associated with a 0.0008480 increase on average.

βˆ2: For a one unit increase in passing yardage, number of games won is associated with a 0.0034689 increase on average.

βˆ5: For a one unit increase in turnover differential, number of games won is associated with a 0.0103765 increase on average.

βˆ7: For a one unit increase in percentage of rushing plays, number of games won is associated with a 0.1336928 increase on average.

βˆ8: For a one unit increase in opponents’ rushing yardage, number of games won is associated with a 0.0047048 decrease on average.

(σ^)2: The model on average has a squared error of 3.126997.

# (d)

Hide

```
#anova from pg.46
fit <- lm(y~x1+x2, data=ftbl)
anova(fit)
```

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x1         1 115.068 115.068  22.378 7.492e-05 ***
x2         1  83.343  83.343  16.208 0.0004635 ***
Residuals 25 128.553   5.142                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
#Source of Variation   SS      DF      MS      F0
#Regression          198.41    2     99.205   19.2926
#Residual            128.553   25    5.14212
#Total               326.963   27
```

ANOVA Table commented in code above^

# (e)

Hide

```
#anova from pg.47
fit0<-lm(y~1, data=ftbl)
anova(fit0, fit)
```

```
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     27 326.96                                  
2     25 128.55  2    198.41 19.293 8.556e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
#Source of Variation   SS      DF      MS      F0
#Regression          198.41    2     99.205   19.293
#Residual            128.55    25     5.142
#Total               326.96    27
```

ANOVA Table commented in code above^

# (f)

Hide

```
# test for significance of regression at significance level of 0.05
qf(0.05, 2, 25, lower.tail=FALSE)
```

```
[1] 3.38519
```

The two ANOVA tables are the same.

# (g)

Hide

```
#Test the value of passing yardage given all the other 4 predictors at significance level of 0.05
reduced <- lm(y~x1+x5+x7+x8, data = ftbl)
anova(reduced, lm1)
```

```
Analysis of Variance Table

Model 1: y ~ x1 + x5 + x7 + x8
Model 2: y ~ x1 + x2 + x5 + x7 + x8
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     23 132.703
2     22  68.794  1    63.909 20.438 0.000169 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i.

H0: $\hat{\beta}_2 = 0$

H1: $\hat{\beta}_2 \neq 0$

ii. F-distribution

iii. F = 20.438

p-value = 0.000169

p-value is very small, much less than .05 and F is quite large; we do reject the null hypothesis and conclude that x2, passing yardage, is a significant contributor.

# (h)

Hide

```
#investigate the contribution of the 3 predictors
reduced1 <- lm(y~x2+x8, data = ftbl)
anova(reduced1, lm1)
```

```
Analysis of Variance Table

Model 1: y ~ x2 + x8
Model 2: y ~ x1 + x2 + x5 + x7 + x8
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     25 83.938
2     22 68.794  3    15.144 1.6144 0.2147
```

i.

H0: $\hat{\beta}_1 = 0, \hat{\beta}_5 = 0, \hat{\beta}_7 = 0$

H1: $\hat{\beta}_1 != 0, \hat{\beta}_5 != 0, \hat{\beta}_7 != 0$

ii. F-distribution

iii.

F = 1.6144

p-value = 0.2147

p-value is not less than .05 and F is quite small; we do not reject the null hypothesis and conclude that x1, x5, and x7 are not significant contributors.

# (i)

I prefer the reduced model with predictors x2 and x8 as it has a higher RSS and is more significant.

# (j)

R code is within this notebook pdf