Violet Chang || chan1300 || 5197617

Stat 3022 Homework 3

```
> ## Problem 2.25
> library(Stat2Data)
> data(Pines)
> head(Pines)
  Row Col Hgt90 Hgt96 Diam96 Grow96 Hgt97 Diam97 Spread.97 Needles97 Deer95
1   1   1    NA    NA     NA     NA    NA     NA        NA        NA     NA
2   1   2    14   284    4.2     96   362    6.6       162        66      0
3   1   3    17   387    7.4    110   442    9.3       250        77      0
4   1   4    NA    NA     NA     NA    NA     NA        NA        NA     NA
5   1   5    24   294    3.9     70   369    7.0       176        72      0
6   1   6    22   310    5.6     84   365    6.9       215        76      0
  Deer97 Cover95 Fert Spacing
1     NA       0    0      15
2      1       2    0      15
3      0       1    0      15
4     NA       0    0      15
5      0       2    0      15
6      0       1    0      15
> lm1=lm(Hgt97~Hgt90, data=Pines)
> summary(lm1)
Call:
lm(formula = Hgt97 ~ Hgt90, data = Pines)

Residuals:
     Min       1Q   Median       3Q      Max 
-261.886  -44.343    7.308   55.114  196.114 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   307.439      9.841  31.239  < 2e-16 ***
Hgt90           2.322      0.492   4.721 2.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78.79 on 807 degrees of freedom
  (191 observations deleted due to missingness)
Multiple R-squared:  0.02687,	Adjusted R-squared:  0.02567 
F-statistic: 22.28 on 1 and 807 DF,  p-value: 2.772e-06
```

#2.25(a): The regression equation is hat(Hgt97) = 307+2.32*Hgt90. From summary, the t value is 4.72 and the p value is 2.77e-06 which is very smaller than zero. Thus we accept the hypothesis H1: Beta1 != 0.

```
> summary(lm1)$adj.r.squared
[1] 0.02566567
> summary(lm1)$r.squared
[1] 0.02687153
```

#2.25(b): The R-squared is about 2.7%, which means only 2.7% of variability in Hgt97 can be explained by Hgt90.

```
> anova(lm1)
Analysis of Variance Table

Response: Hgt97
           Df  Sum Sq Mean Sq F value    Pr(>F)    
Hgt90       1  138344  138344  22.284 2.772e-06 ***
Residuals 807 5010010    6208                      
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#2.25(c): From the anova table shows above.

```
> 138344/(5010010+138344)
[1] 0.0268715
```

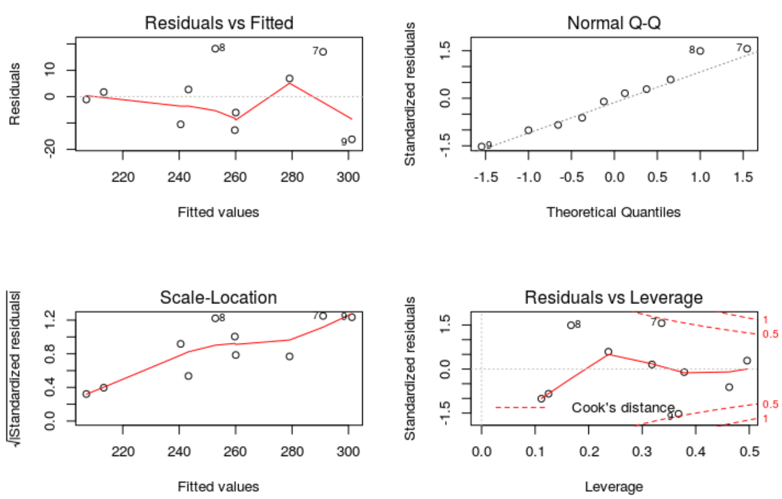
#2.25(d): The coefficient value of determination computed from the anova table is about 2.7%.

#2.25(e): No. Because from the r-squared value we know that only small part of the dataset Hgt97 can be explained by this model, it is not that acceptable.

```
> ## Problem 3.13
> data(MathEnrollment)
> head(MathEnrollment)
  Ayear Fall Spring
1  2001  259    246
2  2002  301    206
3  2003  343    288
4  2004  307    215
5  2005  286    230
6  2006  273    247
> mathenroll = data.frame(MathEnrollment)
> newMathenroll = subset(mathenroll, Ayear!=2003)
> lm2=lm(formula = Spring~Fall+Ayear, data=newMathenroll)
> summary(lm2)
Call:
lm(formula = Spring ~ Fall + Ayear, data = newMathenroll)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.1945  -9.3982   0.3212   5.8503  18.2036 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) -1.172e+04  2.686e+03  -4.361  0.00331 **
Fall        -1.007e+00  2.041e-01  -4.933  0.00169 **
Ayear        6.107e+00  1.337e+00   4.566  0.00258 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.37 on 7 degrees of freedom
Multiple R-squared:  0.871,     Adjusted R-squared:  0.8342 
F-statistic: 23.64 on 2 and 7 DF,  p-value: 0.0007704
```

#3.13(a): The regression equation is hat(Spring) = -11720-1.007*Fall+6.107*Ayear

```
> par(mfrow=c(2,2))
> plot(lm2)
```



#3.13(b): The residuals vs fitted value plot shows that the points around the zero line is random and the zero mean assumption holds. The normal Q-Q plot shows a general linear pattern while there is one point shows significant difference from the linear line. Generally, this model is acceptable.

```
> ## Problem 3.14
> summary(lm2)$r.squared
[1] 0.8710292
> summary(lm2)$adj.r.squared
[1] 0.8341804
```

#3.14(a): The R-squared value is about 87.1% which means that there is 87.1% of Spring enrollment can be explained by this model.

```
> summary(lm2)$sigma
[1] 13.36684
```

#3.14(b): The standard error is about 13.37 which is the size of typical error for this multiple regression.

```
> anova(lm2)
Analysis of Variance Table

Response: Spring
          Df Sum Sq Mean Sq F value   Pr(>F)
Fall       1 4721.1  4721.1  26.423 0.001338 **
Ayear      1 3725.8  3725.8  20.852 0.002585 **
Residuals  7 1250.7   178.7
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

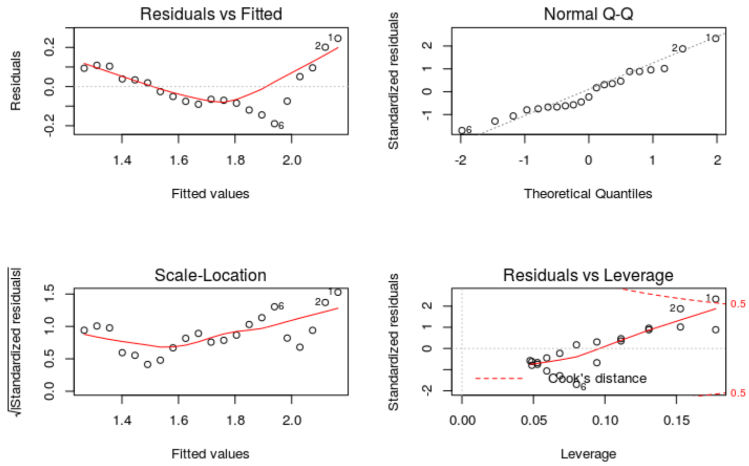
#3.14(c): From the anova table shows the p-values of F-test for both Fall and Ayear, which are 0.001338 and 0.002585, are smaller than 0.05, thus we say that these two variables are significant.

#3.14(d): The hypotheses for Fall are H0: Beta1 = 0 and H1: Beta1 != 0. The t statistic value is -4.933 and the p-value is 0.00169.
The hypotheses for Ayear are H0: Beta1 = 0 and H1: Beta1 != 0. The t statistic value is 4.566 and the p-value is 0.00258.
For both Ayear and Fall, the p-value is smaller than 0.05 which means we reject the H0 and the coefficient for Fall and Ayear are different from zero.

```
> ## Problem 3.19
> data(Speed)
> head(Speed)
  Year FatalityRate StateControl
1 1987         2.41            0
2 1988         2.32            0
3 1989         2.17            0
4 1990         2.08            0
5 1991         1.91            0
6 1992         1.75            0
> lm3=lm(FatalityRate~Year, data=Speed)
> summary(lm3)
Call:
lm(formula = FatalityRate ~ Year, data = Speed)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.18959 -0.07550 -0.02576  0.09346  0.24606 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 91.320887   8.374227    10.9 1.28e-09 ***
Year        -0.044870   0.004193   -10.7 1.75e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1164 on 19 degrees of freedom
Multiple R-squared:  0.8577,	Adjusted R-squared:  0.8502 
F-statistic: 114.5 on 1 and 19 DF,  p-value: 1.75e-09
```

#3.19(a): The regression equation is hat(FatalityRate) = 91.32-0.045*Year, and the slope is -0.045 which shows this is a decline line.

```
> anova(lm3)
Analysis of Variance Table

Response: FatalityRate
          Df  Sum Sq Mean Sq F value   Pr(>F)    
Year       1 1.55026 1.55026  114.49 1.75e-09 ***
Residuals 19 0.25726 0.01354                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> par(mfrow=c(2,2))
> plot(lm3)
```



#3.19(b): The residuals vs fitted value plot is in V shape from the plot which is not that random and doesn't against the zero mean assumption.

```
> lm4=lm(formula = FatalityRate~Year+StateControl+Year*StateControl, data=Speed)
> summary(lm4)
Call:
lm(formula = FatalityRate ~ Year + StateControl + Year * StateControl,
    data = Speed)

Residuals:
       Min        1Q    Median        3Q       Max
-0.103571 -0.020769  0.004048  0.022473  0.091667

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.162e+02  1.303e+01   16.59 6.19e-12 ***
Year             -1.076e-01  6.548e-03  -16.44 7.19e-12 ***
StateControl     -1.614e+02  1.447e+01  -11.15 3.07e-09 ***
Year:StateControl  8.097e-02  7.264e-03   11.15 3.08e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04243 on 17 degrees of freedom
Multiple R-squared:  0.9831,    Adjusted R-squared:  0.9801
F-statistic:   329 on 3 and 17 DF,  p-value: 2.998e-15
> anova(lm4)
Analysis of Variance Table

Response: FatalityRate
                  Df  Sum Sq Mean Sq  F value    Pr(>F)
Year               1 1.55026 1.55026 860.9841 5.288e-16 ***
StateControl       1 0.00292 0.00292   1.6211    0.2201
Year:StateControl  1 0.22373 0.22373 124.2562 3.082e-09 ***
Residuals         17 0.03061 0.00180
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#3.19(c): The regression equation for this model is hat(FatalityRate) = 216.2-0.1076*Year-161.4*StateControl+0.08097*Year*StateControl. The p-value for all state control, year and the year*state control are very small (nearly equal to 0). This shows that all these three factors are very important for this model and the relationship between fatality rate and year is different before and after 1995.

#3.19(d): Plug the value 0 and 1 for state control into the regression before, we

```
get two different equations:
1.hat(FatalityRate) = 216.2-0.1076*Year for Year<1995
2.hat(FatalityRate) = 54.8-0.02663*Year for Year>1995
```

```
> ## Problem 3.21
> data(BritishUnions)
> head(BritishUnions)
    Date AgreePct DisagreePct NetSupport Months Late Unemployment
1 Oct-75       75          16        -59      2    0          4.9
2 Aug-77       79          17        -62     23    0          5.7
3 Sep-78       82          16        -66     36    0          5.5
4 Sep-79       80          16        -64     48    0          5.4
5 Jul-80       72          19        -53     58    0          6.8
6 Nov-81       70          22        -48     74    0         10.2
> BritishUnions$Late = as.factor(BritishUnions$Late)
> lm5=lm(formula = NetSupport~Months+Late+Months*Late, data=BritishUnions)
> summary(lm5)
Call:
lm(formula = NetSupport ~ Months + Late + Months * Late, data = BritishUnions)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.6724 -5.3454 -0.1211  3.6432 14.7972 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -66.62827    4.94880 -13.464  5.2e-09 ***
Months          0.21037    0.07392   2.846   0.0138 *  
Late1          13.11464   21.57377   0.608   0.5537    
Months:Late1    0.17398    0.12761   1.363   0.1959    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.241 on 13 degrees of freedom
Multiple R-squared:  0.9752,    Adjusted R-squared:  0.9695 
F-statistic: 170.4 on 3 and 13 DF,  p-value: 1.102e-10
> anova(lm5)
Analysis of Variance Table

Response: NetSupport
             Df  Sum Sq Mean Sq  F value    Pr(>F)    
Months        1 25734.8 25734.8 490.8652  1.04e-11 ***
Late          1   965.7   965.7  18.4200 0.0008764 ***
Months:Late   1    97.5    97.5   1.8588 0.1959052    
Residuals    13   681.6    52.4                       
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

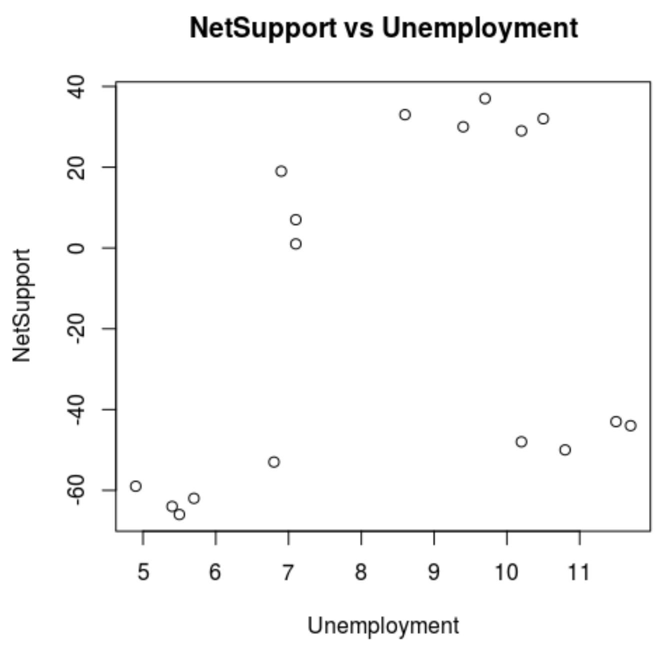
#3.21(a): The regression equation is hat(NetSupport) = -66.628+0.2104*Months+13.115*Late1+0.1740*Months*Late1

#3.21(b): The t statistic value is 1.363 for Months:Late1 and the p-value is 0.1959 which is vary larger than 0.05. So we say that parallel lines are adequate for describing the relationship between NetSupport and Months and the interaction can be dropped from this model. We can't drop Late factor because we don't know what the model will look like when interaction dropped.

```
> lm6=lm(formula = NetSupport~Months, data=BritishUnions)
> anova(lm6,lm5)
Analysis of Variance Table

Model 1: NetSupport ~ Months
Model 2: NetSupport ~ Months + Late + Months * Late
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     15 1744.73
2     13  681.56  2    1063.2 10.139 0.002221 **
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#3.21(c): The F test statistic value is 10.139 from the anova table and the p-value is 0.002221 which is very small. The hypotheses are H0: Beta2=Beta3=0 and H1: Beta2!=0 and Beta3!=0. Since the p-value is very small, we reject the H0 and the Late factor is important for the model.

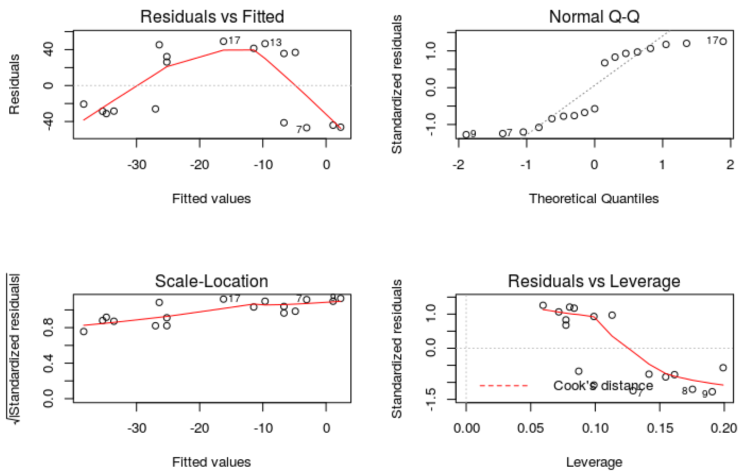
```
> ## Problem 3.22
> par(mfrow=c(1,1))
> plot(NetSupport~Unemployment, data=BritishUnions, main="NetSupport vs Unemployment")
```



#3.22(a): The scatter plot shows above. The plot shows that unemployment rate vs net support is random and can’t find a specific pattern for it.

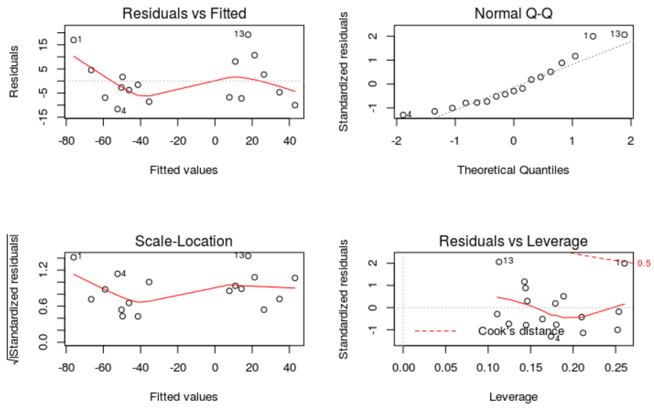
```
> lm7=lm(formula = NetSupport~Unemployment, data=BritishUnions)
> summary(lm7)

Call:
lm(formula = NetSupport ~ Unemployment, data = BritishUnions)

Residuals:
    Min      1Q Median      3Q     Max 
-46.93 -31.23 -20.64  36.87  49.23 
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -67.660     37.862  -1.787   0.0942 .
Unemployment     5.980      4.379   1.366   0.1921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.37 on 15 degrees of freedom
Multiple R-squared:  0.1106,	Adjusted R-squared:  0.05132
F-statistic: 1.865 on 1 and 15 DF,  p-value: 0.1921
> par(mfrow=c(2,2))
> plot(lm7)
```



#3.22(b): From the residuals vs fitted value plot, we know that the points around the zero line are not very random and the zero mean assumption is againsted, so this model is not that acceptable. The p-value from the summary is 0.192 which is very larger than 0.05 so we can say that unemployment is not that significant for the net support.

```
> lm8=lm(formula = NetSupport~Unemployment+Months, data=BritishUnions)
> summary(lm8)
Call:
lm(formula = NetSupport ~ Unemployment + Months, data = BritishUnions)

Residuals:
    Min      1Q  Median      3Q     Max
-11.628  -6.924  -2.717   4.554  19.202

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -65.51220    9.27541  -7.063 5.66e-06 ***
Unemployment   -2.35767    1.20207  -1.961     0.07 .
Months          0.53898    0.03508  15.362 3.71e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.887 on 14 degrees of freedom
Multiple R-squared:  0.9502,	Adjusted R-squared:  0.9431
F-statistic: 133.5 on 2 and 14 DF,  p-value: 7.603e-10
> plot(lm8)
```



#3.22(c): From the summary, the p-value for months is very small which indicate that this factor is important for this model and the residuals vs fitted value plot shows that the variances around the zero line is acceptable and doesn’t against the zero mean assumption. As compare at the level of 0.10, p-value for both unemployment and months are smaller than 0.10. So the model fitted is acceptable and we can reject the H0: Beta1=0.

#3.22(d): The p-value of unemployment is smaller in part c than part b which indicate that its importance becomes higher in the second model. The coefficient is positive in the first model and negative in the second model. Thus, we can conclude that after adjust months, unemployment rate can be associated with net support in better way.

```
> ## Problem 3.30
> data(Pollster08)
> head(Pollster08)
        PollTaker   PollDates MidDate Days    n Pop McCain Obama Margin Charlie
1       Rasmussen  8/28-30/08    8/29    1 3000  LV     46    49      3       0
2           Zogby  8/29-30/08    8/30    2 2020  LV     47    45     -2       0
3 Diageo/Hotline  8/29-31/08    8/30    2  805  RV     39    48      9       0
4             CBS  8/29-31/08    8/30    2  781  RV     40    48      8       0
5             CNN  8/29-31/08    8/30    2  927  RV     48    49      1       0
6       Rasmussen 8/30-9/1/08    8/31    3 3000  LV     45    51      6       0
  Meltdown
1        0
2        0
3        0
4        0
5        0
6        0
> lm9=lm(formula = Margin~Days+I(Days^2),data=Pollster08)
> summary(lm9)
Call:
lm(formula = Margin ~ Days + I(Days^2), data = Pollster08)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7496  -2.0461  -0.1227   1.9297   6.8969

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.477958   1.095676   4.087 8.89e-05 ***
Days        -0.604426   0.138598  -4.361 3.18e-05 ***
I(Days^2)    0.021129   0.003776   5.595 1.97e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.014 on 99 degrees of freedom
Multiple R-squared:  0.3495,    Adjusted R-squared:  0.3363
F-statistic: 26.59 on 2 and 99 DF,  p-value: 5.711e-10
> anova(lm9)
Analysis of Variance Table

Response: Margin
          Df Sum Sq Mean Sq F value    Pr(>F)
Days       1 198.74 198.736  21.879 9.205e-06 ***
I(Days^2)  1 284.34 284.345  31.304 1.966e-07 ***
Residuals 99 899.24   9.083
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#3.30(a): The R-squared value is 34.95% and the SSE is 899.24 from summary and the anova table.

```
> Pollster08$Charlie = as.factor(Pollster08$Charlie)
> lm10=lm(formula = Margin~Days+Charlie, data=Pollster08)
> summary(lm10)
Call:
lm(formula = Margin ~ Days + Charlie, data = Pollster08)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7871  -2.1513  -0.1123   1.7988   9.0684

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.31282    0.78692  -0.398   0.6918
Days         0.12222    0.06774   1.804   0.0742 .
Charlie1     0.63640    1.30386   0.488   0.6266
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.454 on 99 degrees of freedom
Multiple R-squared:  0.1458,    Adjusted R-squared:  0.1286
F-statistic: 8.451 on 2 and 99 DF,  p-value: 0.0004089
> anova(lm10)
Analysis of Variance Table

Response: Margin
          Df  Sum Sq Mean Sq F value    Pr(>F)
Days       1  198.74 198.736 16.6630 9.056e-05 ***
Charlie    1    2.84   2.841  0.2382    0.6266
Residuals 99 1180.75  11.927
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#3.30(b): The R-squared value is 14.58% and the SSE is 1180.75 from summary and the anova table.

```
> Pollster08$Meltdown = as.factor(Pollster08$Meltdown)
> lm11=lm(formula = Margin~Days+Meltdown, data=Pollster08)
> summary(lm11)
Call:
lm(formula = Margin ~ Days + Meltdown, data = Pollster08)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7480  -2.5448   0.0408   2.0390   8.2618

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.735333   0.881351   0.834   0.4061
Days        0.001412   0.072015   0.020   0.9844
Meltdown1   3.187183   1.340626   2.377   0.0194 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.363 on 99 degrees of freedom
Multiple R-squared:   0.19,     Adjusted R-squared:  0.1736
F-statistic: 11.61 on 2 and 99 DF,  p-value: 2.949e-05
> anova(lm11)
Analysis of Variance Table
```

```
Response: Margin
           Df  Sum Sq Mean Sq F value    Pr(>F)
Days       1  198.74 198.736  17.572 6.024e-05 ***
Meltdown   1   63.92  63.922   5.652   0.01936 *
Residuals 99 1119.67  11.310
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#3.30(c): The R-squared value is 19% and the SSE is 1119.67 from summary and the anova table.

#3.30(d): After comparing these three model I would choose model in part a because it has the largest R-squared value which means it can explain the most data for margin on days among these three models and the SSE for the model in part a is the smallest one among all three models which means its predicting error is the smallest in these three.