Regression_Modeling_Analysis

Sejun Song

2022-10-24

```
library(latexpdf)
library(tinytex)
library(xfun)
```

```
##
## Attaching package: 'xfun'
```

```
## The following objects are masked from 'package:base':
##
##     attr, isFALSE
```

```
library(xfun)
library(readr)
library(ggplot2)
library(readxl)
require(rsq)
```

```
## Loading required package: rsq
```

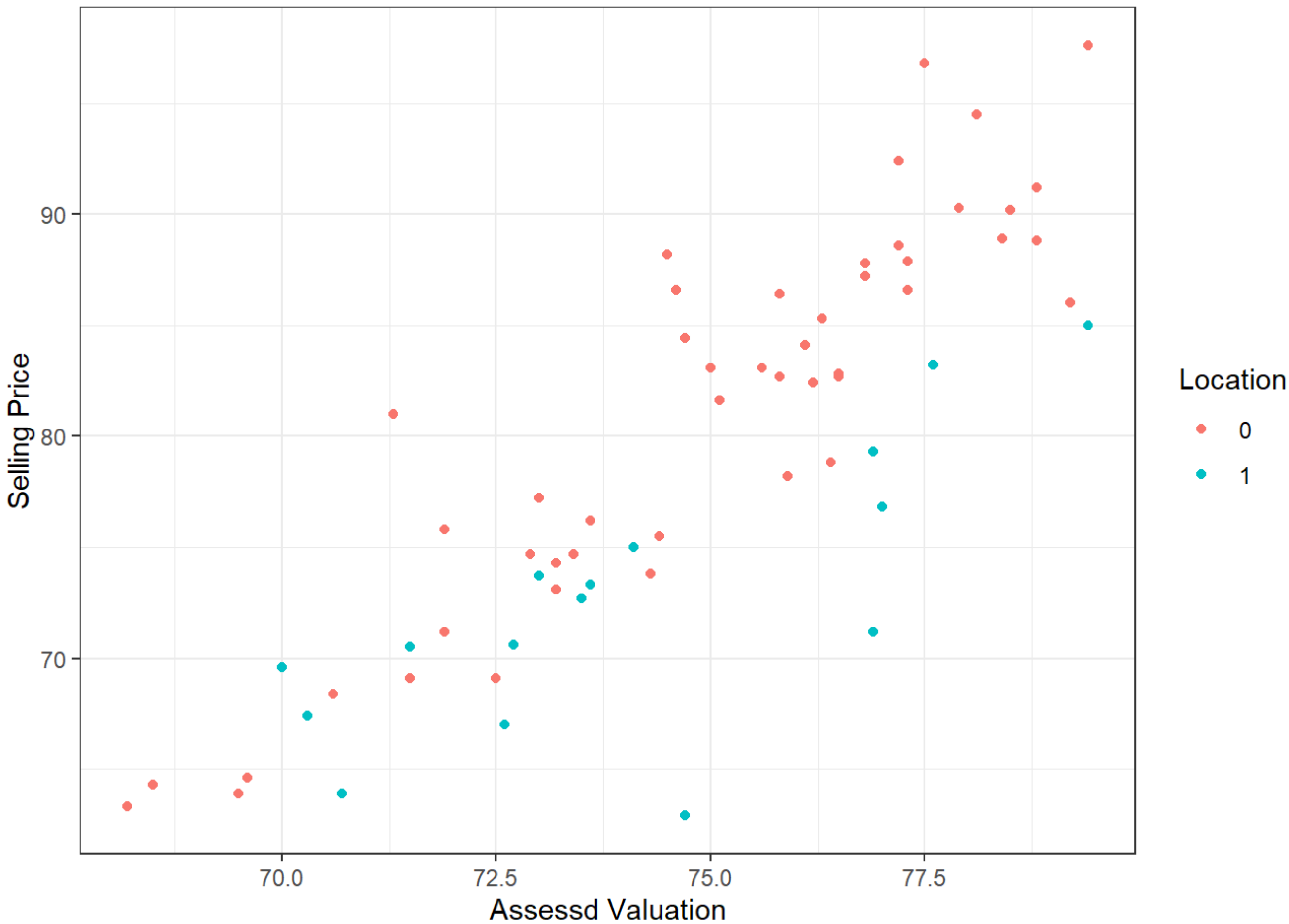
```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

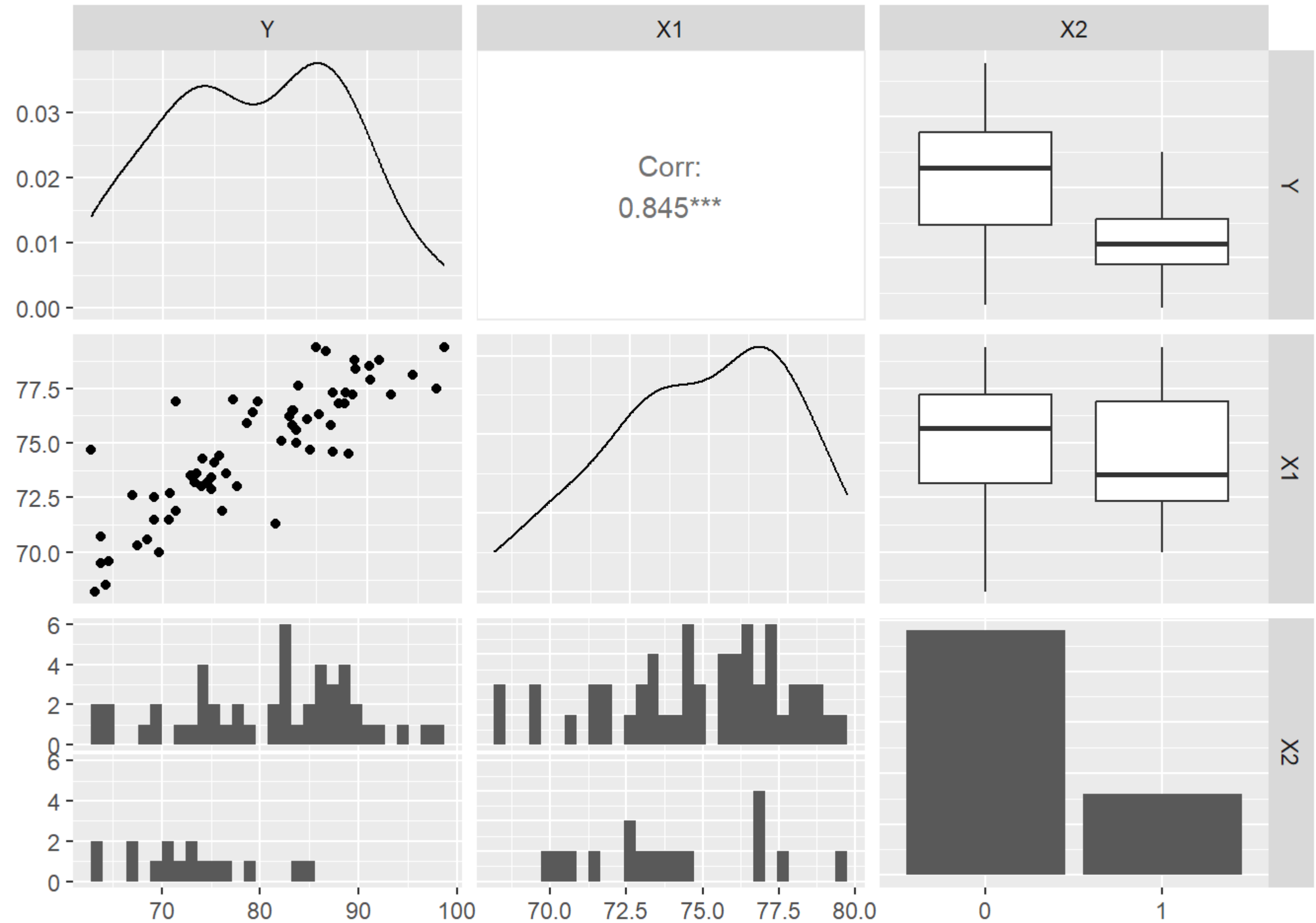
```
## The following object is masked from 'package:latexpdf':
##
##     wrap
```

```r
## Read data
tax <- read.table(file = "http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatas
ets/KutnerData/Chapter%20%208%20Data%20Sets/CH08PR24.txt")
names(tax) <- c("Y","X1","X2")
tax$X2 <- factor(tax$X2)
## plot the variables
par(mfrow=c(2,2))
taxplot <- ggplot(data = tax)
taxplot+geom_point(mapping = aes(x = X1 ,y = Y, colour = X2))+
  theme_bw() +
  labs(x = "Assessd Valuation",y = "Selling Price",colour = "Location")
```




```r
ggpairs(tax)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



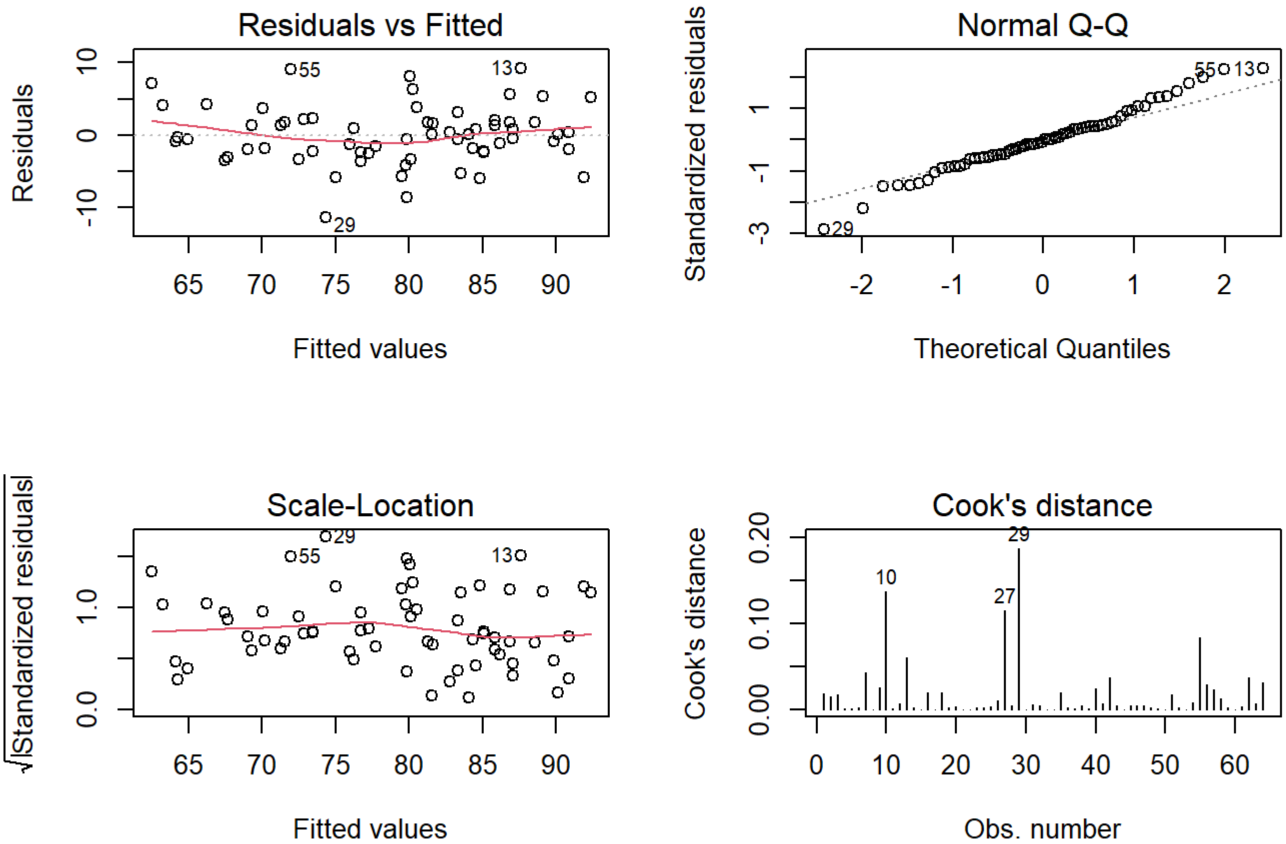

```
## Created an additive linear model to find signifcant variables
taxmod <- lm(Y~X1+factor(X2), data=tax)
summary(taxmod)
```

```
##
## Call:
## lm(formula = Y ~ X1 + factor(X2), data = tax)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4141  -2.2927  -0.1456   1.8678   9.2341
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -107.4597    13.5509   -7.93 5.80e-11 ***
## X1             2.5165     0.1806   13.93  < 2e-16 ***
## factor(X2)1   -6.2057     1.1933   -5.20 2.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.093 on 61 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7949
## F-statistic: 123.1 on 2 and 61 DF,  p-value: < 2.2e-16
```

```r
taxmod2 <- lm(Y~X1+X2, data=tax)
summary(taxmod2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = tax)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4141  -2.2927  -0.1456   1.8678   9.2341
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -107.4597    13.5509   -7.93 5.80e-11 ***
## X1             2.5165     0.1806   13.93  < 2e-16 ***
## X21           -6.2057     1.1933   -5.20 2.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.093 on 61 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7949
## F-statistic: 123.1 on 2 and 61 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
## Created the residual plots to check all 4 assumptions of linear model
plot(taxmod,1:4)
```




```
## Created an interaction linear model to compare the linearity with the additive model & checked the residuals to confirm that the variables are significant with less than p-value of 0.05
par(mfrow=c(2,2))
interactmod <- lm(Y~X1*factor(X2), data=tax)
summary(interactmod)
```

```
##
## Call:
## lm(formula = Y ~ X1 * factor(X2), data = tax)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8470  -2.1639   0.0913   1.9348   9.9836
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -126.9052    14.7225  -8.620 4.33e-12 ***
## X1                2.7759     0.1963  14.142  < 2e-16 ***
## factor(X2)1      76.0215    30.1314   2.523  0.01430 *
## X1:factor(X2)1   -1.1075     0.4055  -2.731  0.00828 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.893 on 60 degrees of freedom
## Multiple R-squared:  0.8233, Adjusted R-squared:  0.8145
## F-statistic: 93.21 on 3 and 60 DF,  p-value: < 2.2e-16
```

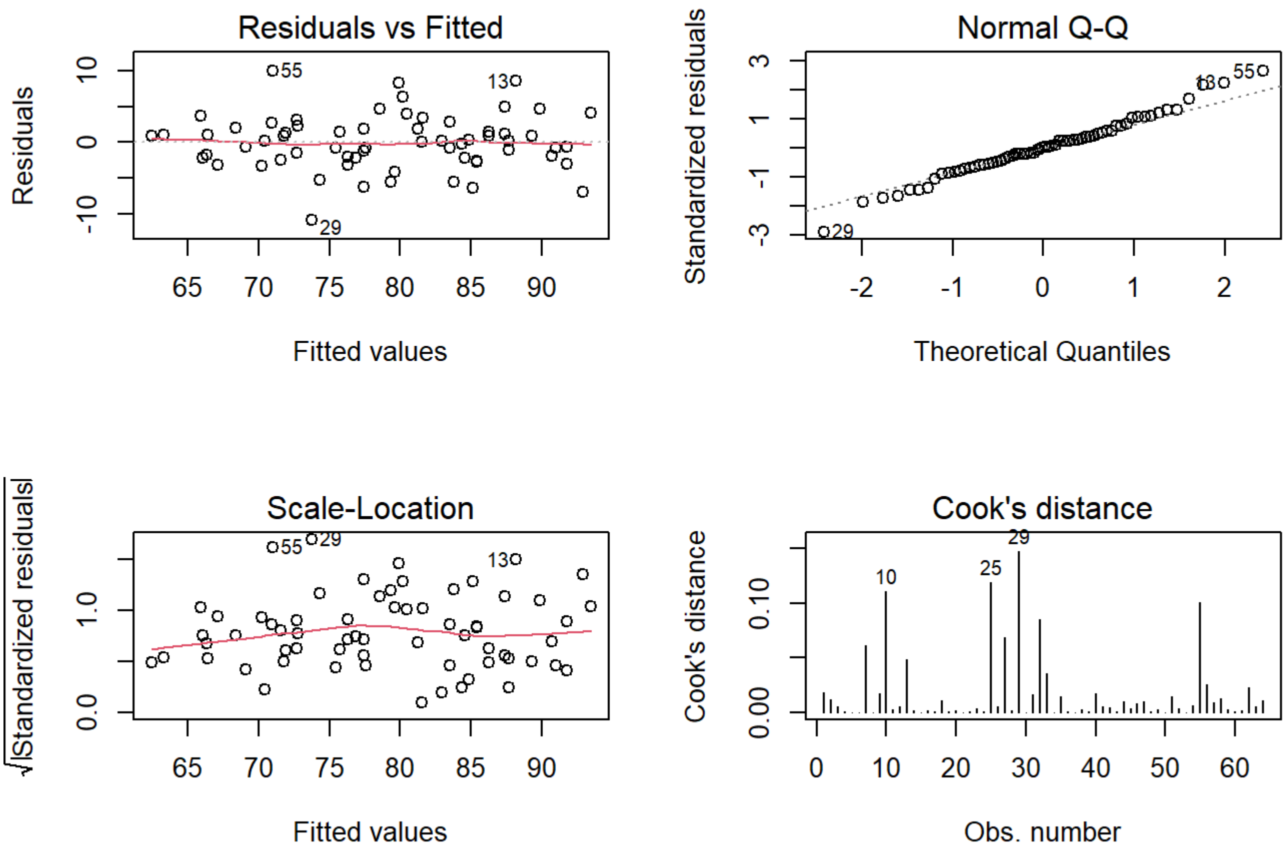
```
interactmod2 <- lm(Y~X1*X2, data=tax)
summary(interactmod2)
```

```
##
## Call:
## lm(formula = Y ~ X1 * X2, data = tax)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8470  -2.1639   0.0913   1.9348   9.9836
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -126.9052    14.7225  -8.620 4.33e-12 ***
## X1             2.7759     0.1963  14.142  < 2e-16 ***
## X21           76.0215    30.1314   2.523  0.01430 *
## X1:X21        -1.1075     0.4055  -2.731  0.00828 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.893 on 60 degrees of freedom
## Multiple R-squared:  0.8233, Adjusted R-squared:  0.8145
## F-statistic: 93.21 on 3 and 60 DF,  p-value: < 2.2e-16
```

```
anova(taxmod,interactmod)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + factor(X2)
## Model 2: Y ~ X1 * factor(X2)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     61 1022.1
## 2     60  909.1  1       113 7.4578 0.008281 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(interactmod, 1:4)
```




```
## Identified the expected price for one-family residential dwelling by using predict function i
n the model
to.predict1 <- data.frame(X1=73,X2=1)
predict.lm(object = interactmod,newdata = to.predict1,interval = "confidence")
```

```
##       fit      lwr      upr
## 1 70.9107 68.83105 72.99034
```

```
round(68.83105,4)
```

```
## [1] 68.831
```

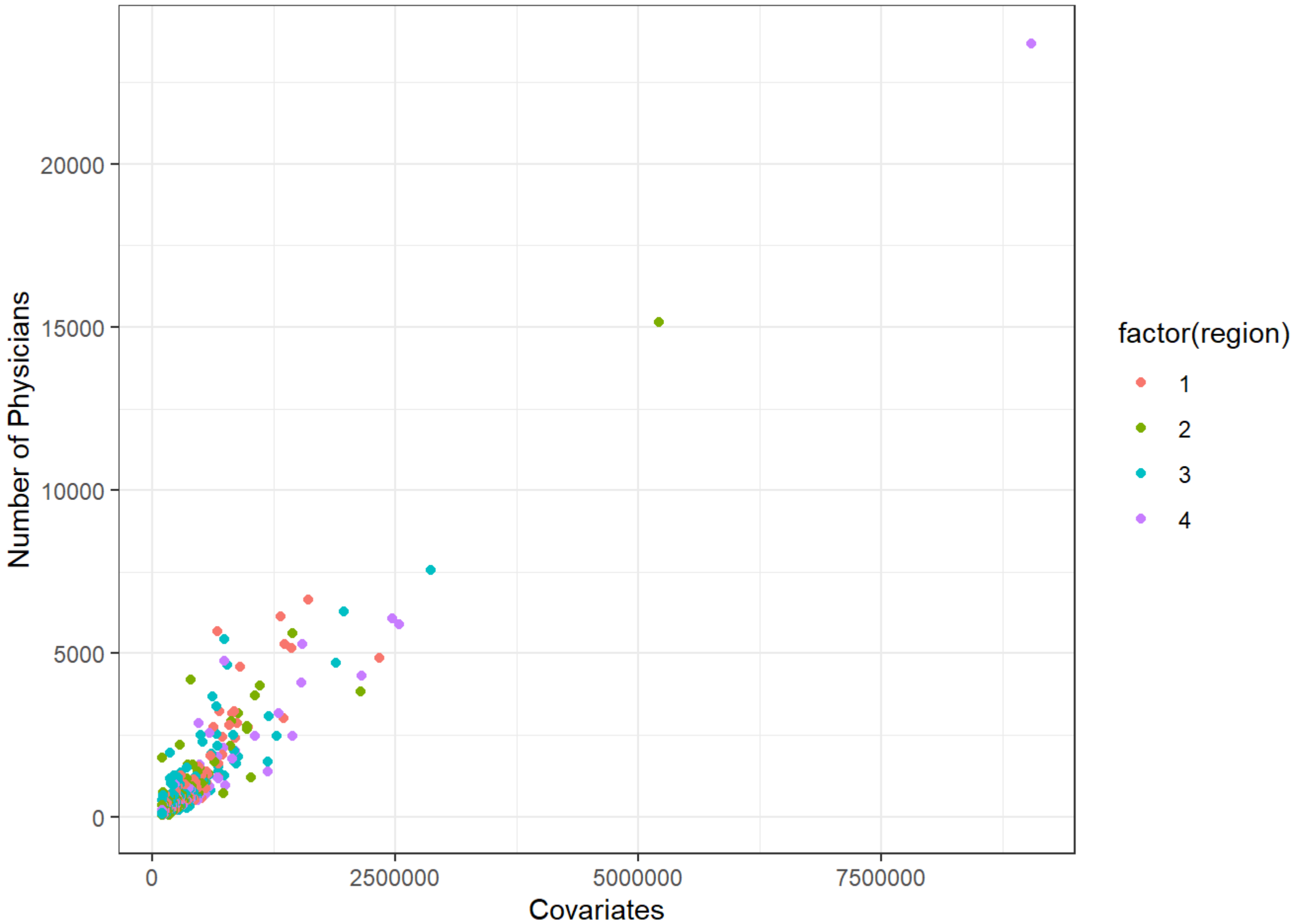
```
to.predict2 <- data.frame(X1=73,X2=1)
predict.lm(object = interactmod,newdata = to.predict2,interval = "prediction")
```

```
##       fit      lwr      upr
## 1 70.9107 62.85154 78.96985
```

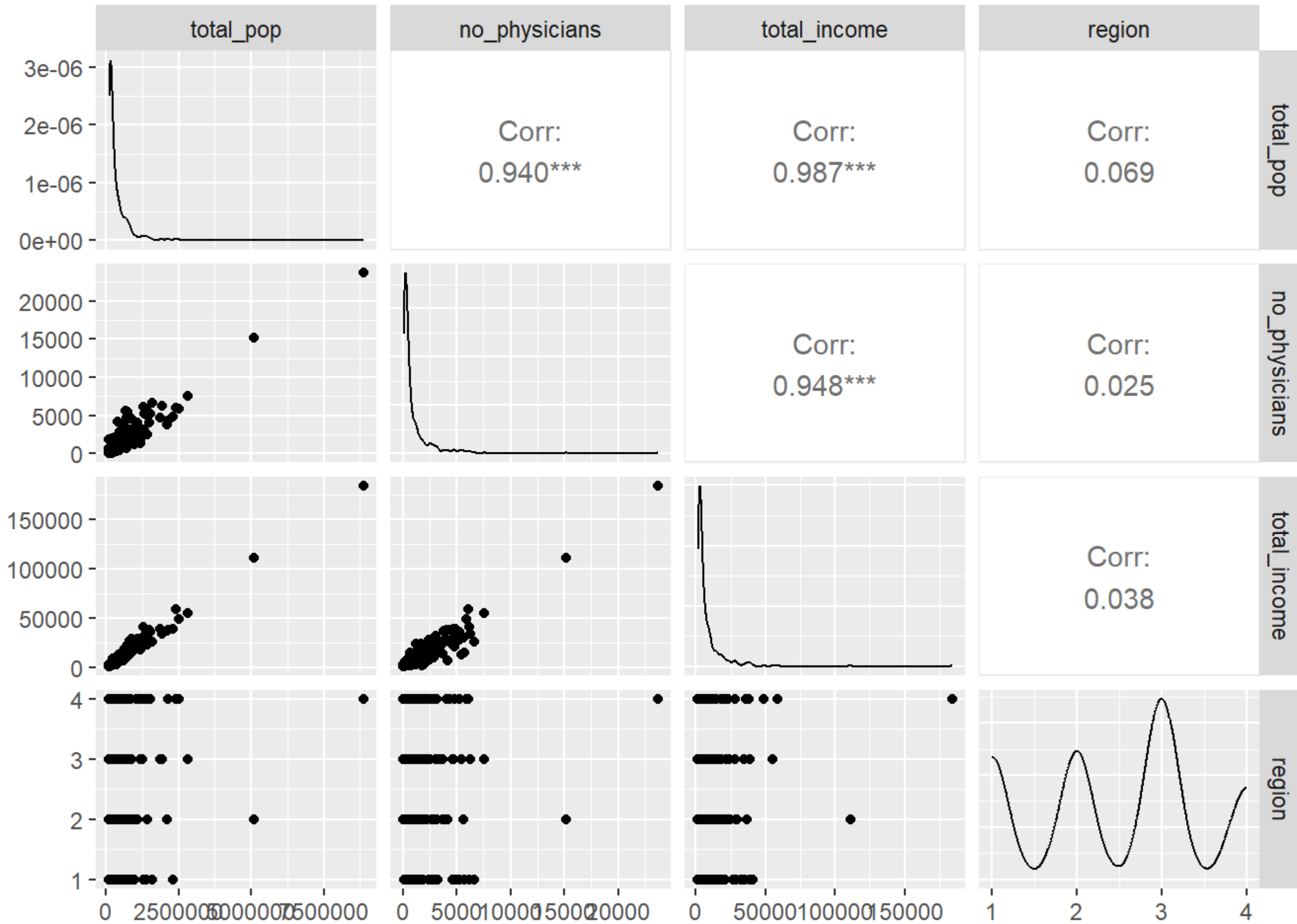
```
round(78.96985 ,4)
```

```
## [1] 78.9698
```

```
## Created an Explanatory Data Analysis (EDA) to check the linearity
cdi <- read.table(file = "C:\\Users\\yss06\\Desktop\\STAT 462\\APPENC02.txt")
names(cdi) <- c("identity","county",
                "state","land_area",
                "total_pop","percent_18_34","percent_65_older", "no_physicians",
                "no_beds","total_crime","percent_highschool","percent_bachelor","percent_below",
                "percent_unemployment","per_capita_income","total_income","region")
## Removed variables that are not relevant to the response variable
cdi$identity <- NULL
cdi$county <- NULL
cdi$state <- NULL
cdi$land_area <- NULL
cdi$percent_18_34 <- NULL
cdi$percent_65_older <- NULL
cdi$no_beds <- NULL
cdi$total_crime <- NULL
cdi$percent_highschool <- NULL
cdi$percent_bachelor <- NULL
cdi$percent_below <- NULL
cdi$percent_unemployment <- NULL
cdi$per_capita_income <- NULL
cdiplot <- ggplot(data = cdi)
cdiplot+geom_point(mapping = aes(x = total_pop + total_income,y =  no_physicians,colour= factor
(region)))+
  theme_bw() +
  labs(x = "Covariates",y = "Number of Physicians")
```


Number of Physicians
20000
15000
10000
5000
0
0
2500000
5000000
7500000
Covariates
factor(region)
1
2
3
4

```
ggpairs(cdi)
```




```
## created a second-order model (quadratic model) using the total population as a covariate

secondfull <- lm(no_physicians~I(total_pop)+I(total_pop^2),data = cdi)
summary(secondfull)
```

```
##
## Call:
## lm(formula = no_physicians ~ I(total_pop) + I(total_pop^2), data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2161.9  -201.3   -59.6    48.1  3875.4
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.674e+02  4.214e+01  -3.971 8.36e-05 ***
## I(total_pop)    2.983e-03  9.313e-05  32.031  < 2e-16 ***
## I(total_pop^2) -3.295e-11  1.400e-11  -2.353   0.0191 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 606.9 on 437 degrees of freedom
## Multiple R-squared:  0.8855, Adjusted R-squared:  0.885
## F-statistic:  1690 on 2 and 437 DF,  p-value: < 2.2e-16
```

```
reduced <- lm(no_physicians~I(total_pop),data = cdi)
summary(reduced)
```

```
##
## Call:
## lm(formula = no_physicians ~ I(total_pop), data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1969.4  -209.2   -88.0    27.9  3928.7
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.106e+02  3.475e+01  -3.184  0.00156 **
## I(total_pop)  2.795e-03  4.837e-05  57.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841, Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

```
## Utilized the general F-test to test if the quadratic term can be dropped from the model
anova(reduced,secondfull)
```

```
## Analysis of Variance Table
##
## Model 1: no_physicians ~ I(total_pop)
## Model 2: no_physicians ~ I(total_pop) + I(total_pop^2)
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1    438 163025135
## 2    437 160985454  1   2039681 5.5368 0.01906 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## fitted 3 different models to make a model that is the most fitted
mod1 <- lm(no_physicians~total_pop+total_income+factor(region), data=cdi)
summary(mod1)
```

```
##
## Call:
## lm(formula = no_physicians ~ total_pop + total_income + factor(region),
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.848e+01  5.882e+01  -0.994   0.3207
## total_pop        5.515e-04  2.835e-04   1.945   0.0524 .
## total_income     1.070e-01  1.325e-02   8.073  6.8e-15 ***
## factor(region)2 -3.493e+00  7.881e+01  -0.044   0.9647
## factor(region)3  4.220e+01  7.402e+01   0.570   0.5689
## factor(region)4 -1.490e+02  8.683e+01  -1.716   0.0868 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

```
mod2 <- lm(formula = no_physicians~I(total_pop)+I(total_pop^2)+I(total_income)+region,data = cd
i)
summary(mod2)
```

```
##
## Call:
## lm(formula = no_physicians ~ I(total_pop) + I(total_pop^2) +
##     I(total_income) + region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1917.0  -194.0   -54.2    69.8  3713.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.439e+01  7.513e+01  -0.458   0.6474
## I(total_pop)     7.708e-04  2.991e-04   2.577   0.0103 *
## I(total_pop^2)  -2.341e-11  1.312e-11  -1.785   0.0750 .
## I(total_income)  1.023e-01  1.322e-02   7.742  6.9e-14 ***
## region          -3.001e+01  2.670e+01  -1.124   0.2615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.4 on 435 degrees of freedom
## Multiple R-squared:  0.9008, Adjusted R-squared:  0.8999
## F-statistic: 987.3 on 4 and 435 DF,  p-value: < 2.2e-16
```

```
mod3 <- lm(formula = no_physicians ~ I(total_pop)+I(total_pop^2)+I(total_income)+factor(region)+
I(total_pop)*factor(region)+I(total_pop^2)*factor(region)+I(total_income)*factor(region),data=cd
i)
summary(mod3)
```

```
##
## Call:
## lm(formula = no_physicians ~ I(total_pop) + I(total_pop^2) +
##     I(total_income) + factor(region) + I(total_pop) * factor(region) +
##     I(total_pop^2) * factor(region) + I(total_income) * factor(region),
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1951.0  -177.9   -53.3    90.6  3490.8
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -3.984e+02  1.242e+02  -3.208 0.001440 **
## I(total_pop)                    2.068e-03  8.477e-04   2.439 0.015118 *
## I(total_pop^2)                 -3.356e-10  2.570e-10  -1.306 0.192339
## I(total_income)                 8.841e-02  2.333e-02   3.790 0.000173 ***
## factor(region)2                 3.115e+02  1.504e+02   2.070 0.039029 *
## factor(region)3                 2.659e+02  1.514e+02   1.756 0.079811 .
## factor(region)4                 4.019e+02  1.564e+02   2.569 0.010541 *
## I(total_pop):factor(region)2    -1.611e-03  1.203e-03  -1.339 0.181203
## I(total_pop):factor(region)3     4.754e-05  1.105e-03   0.043 0.965711
## I(total_pop):factor(region)4    -2.552e-03  1.041e-03  -2.452 0.014621 *
## I(total_pop^2):factor(region)2   3.354e-10  2.625e-10   1.278 0.201985
## I(total_pop^2):factor(region)3   2.323e-10  2.919e-10   0.796 0.426576
## I(total_pop^2):factor(region)4   3.721e-10  2.578e-10   1.444 0.149610
## I(total_income):factor(region)2  2.703e-02  4.554e-02   0.594 0.553156
## I(total_income):factor(region)3 -4.239e-02  3.788e-02  -1.119 0.263708
## I(total_income):factor(region)4  4.761e-02  3.567e-02   1.335 0.182704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 552.4 on 424 degrees of freedom
## Multiple R-squared:  0.908,  Adjusted R-squared:  0.9048
## F-statistic:   279 on 15 and 424 DF,  p-value: < 2.2e-16
```

```
## Examined if the interactions in the models are significant by using the general F-tests
reduced1 <- lm(no_physicians~I(total_pop)+I(total_pop^2)+I(total_income)+factor(region)+I(total_
pop)*factor(region)+I(total_pop^2)*factor(region), data = cdi)
reduced2 <- lm(no_physicians~I(total_pop)+I(total_pop^2)+I(total_income)+factor(region)+I(total_
income)*factor(region), data = cdi)
reduced3 <- lm(no_physicians~I(total_pop)+I(total_pop^2)+I(total_income)+factor(region),data=cd
i)
anova(reduced1,mod3)
```

```
## Analysis of Variance Table
##
## Model 1: no_physicians ~ I(total_pop) + I(total_pop^2) + I(total_income) +
##     factor(region) + I(total_pop) * factor(region) + I(total_pop^2) *
##     factor(region)
## Model 2: no_physicians ~ I(total_pop) + I(total_pop^2) + I(total_income) +
##     factor(region) + I(total_pop) * factor(region) + I(total_pop^2) *
##     factor(region) + I(total_income) * factor(region)
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    427 130993130
## 2    424 129358977  3   1634153 1.7854 0.1492
```

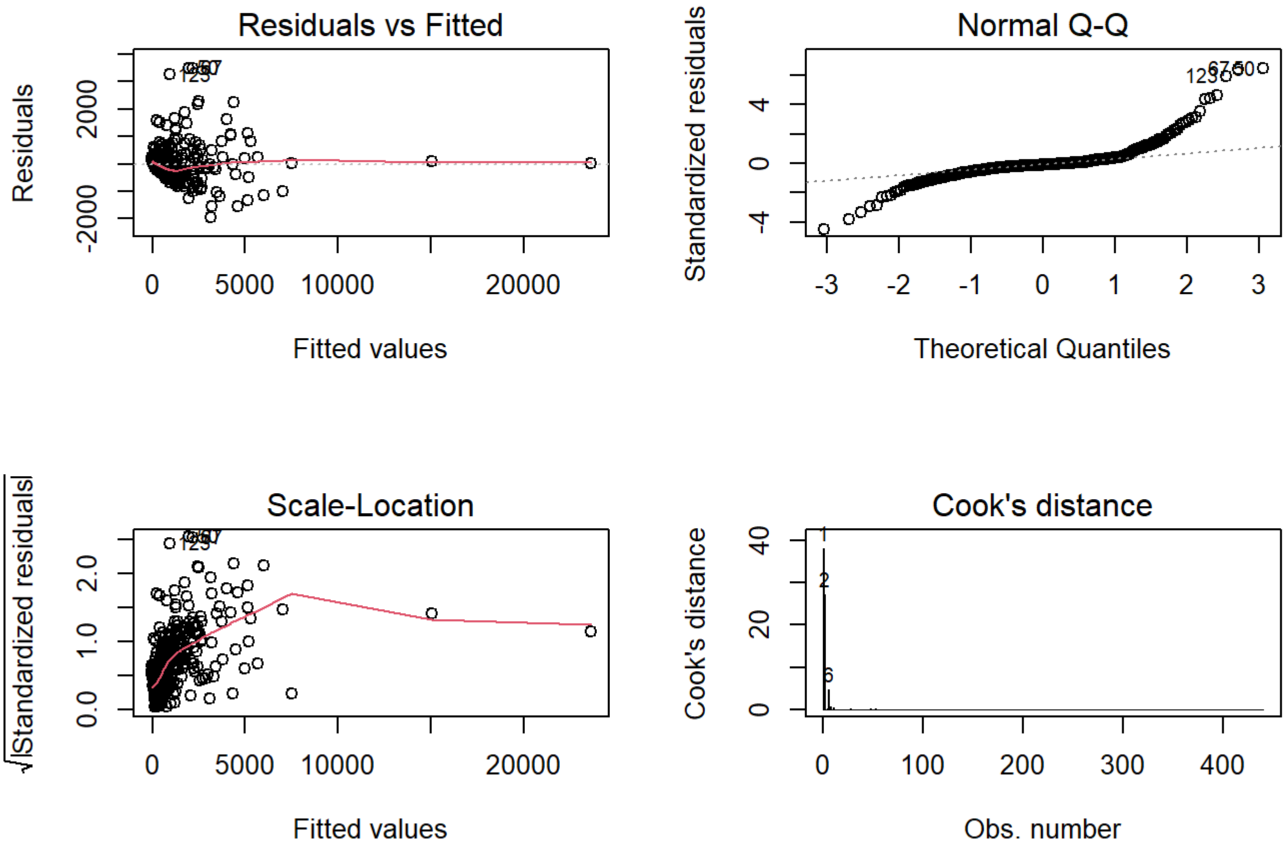
```
anova(reduced2, mod3)
```

```
## Analysis of Variance Table
##
## Model 1: no_physicians ~ I(total_pop) + I(total_pop^2) + I(total_income) +
##     factor(region) + I(total_income) * factor(region)
## Model 2: no_physicians ~ I(total_pop) + I(total_pop^2) + I(total_income) +
##     factor(region) + I(total_pop) * factor(region) + I(total_pop^2) *
##     factor(region) + I(total_income) * factor(region)
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    430 132599254
## 2    424 129358977  6   3240277 1.7701 0.1037
```

```
anova(reduced3, mod3)
```

```
## Analysis of Variance Table
##
## Model 1: no_physicians ~ I(total_pop) + I(total_pop^2) + I(total_income) +
##     factor(region)
## Model 2: no_physicians ~ I(total_pop) + I(total_pop^2) + I(total_income) +
##     factor(region) + I(total_pop) * factor(region) + I(total_pop^2) *
##     factor(region) + I(total_income) * factor(region)
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1    433 137980691
## 2    424 129358977  9   8621714 3.1399 0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

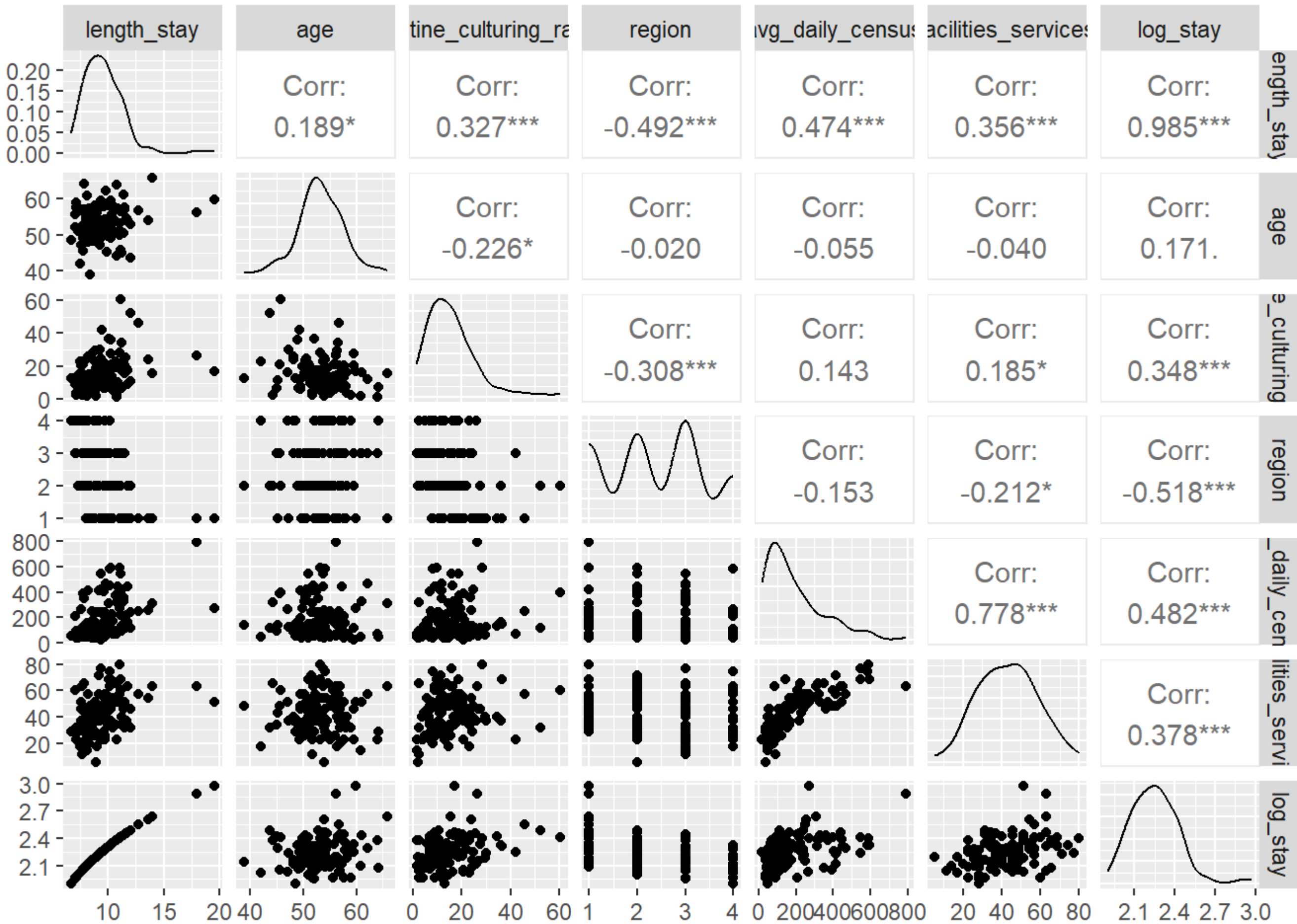
```
## Analyzed the residuals of a model to check three assumptions of linear model: Linearity, Equa
l Variance, Normality
par(mfrow=c(2,2))
plot(mod3,1:4)
```



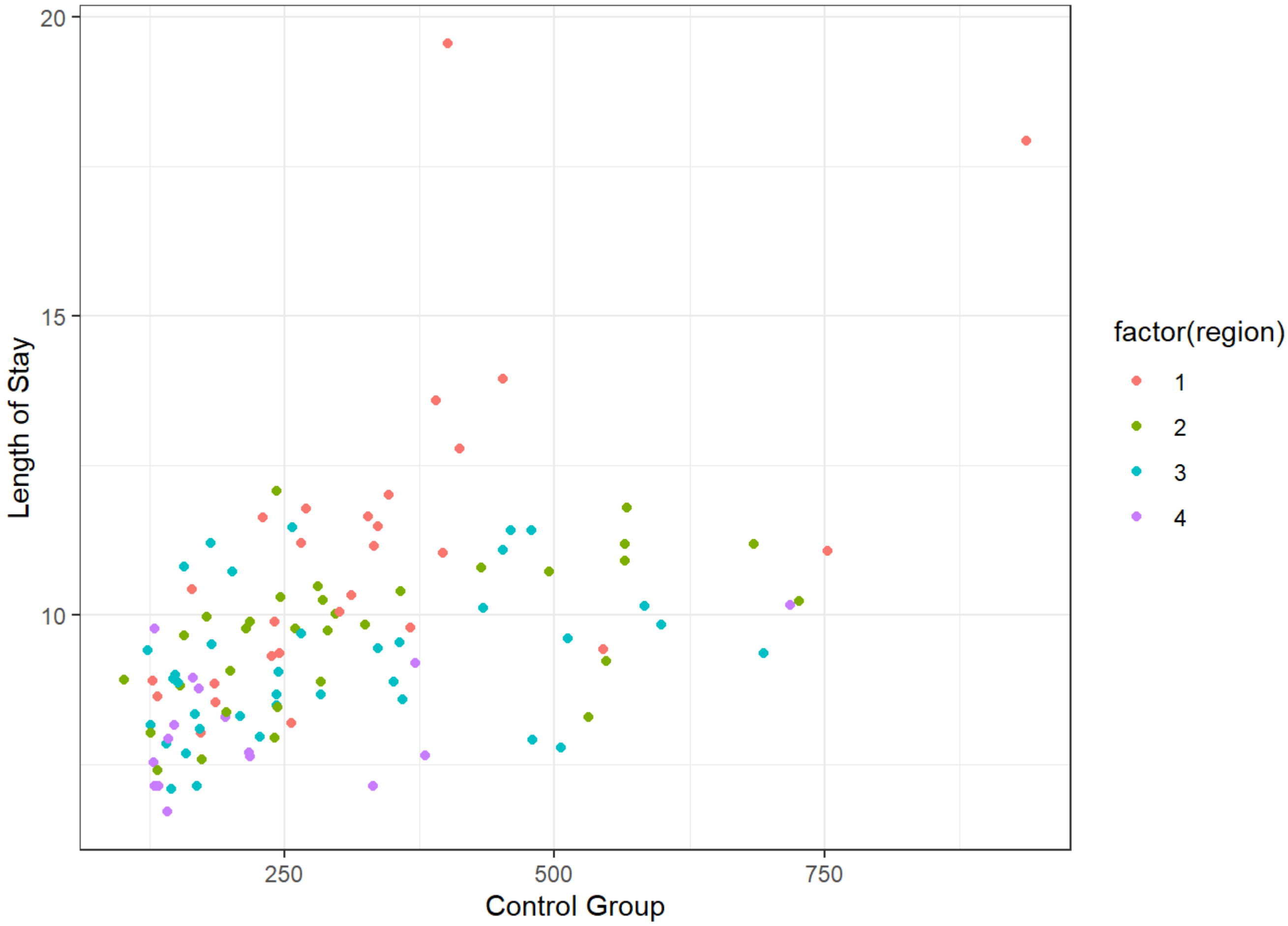

```
shapiro.test(residuals(mod3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod3)
## W = 0.76149, p-value < 2.2e-16
```

```
## Loaded data set :
senic <- read.table(file = "C:\\Users\\yss06\\Desktop\\STAT 462\\APPENC01.txt")
senic$V1 <- NULL
senic$V4 <- NULL
senic$V6 <- NULL
senic$V7 <- NULL
senic$V8 <- NULL
senic$V11 <- NULL
names(senic) <- c("length_stay","age","routine_culturing_ratio","region","avg_daily_census","fac
ilities_services")
senic$log_stay <- log(senic$length_stay)
ggpairs(senic)
```




```
## Plot the data set
senicplot <- ggplot(data = senic)
senicplot+geom_point(mapping = aes(x = age + routine_culturing_ratio+avg_daily_census+facilities
_services,y = length_stay,colour= factor(region)))+
  theme_bw()+
  labs(x = "Control Group",y = "Length of Stay")
```

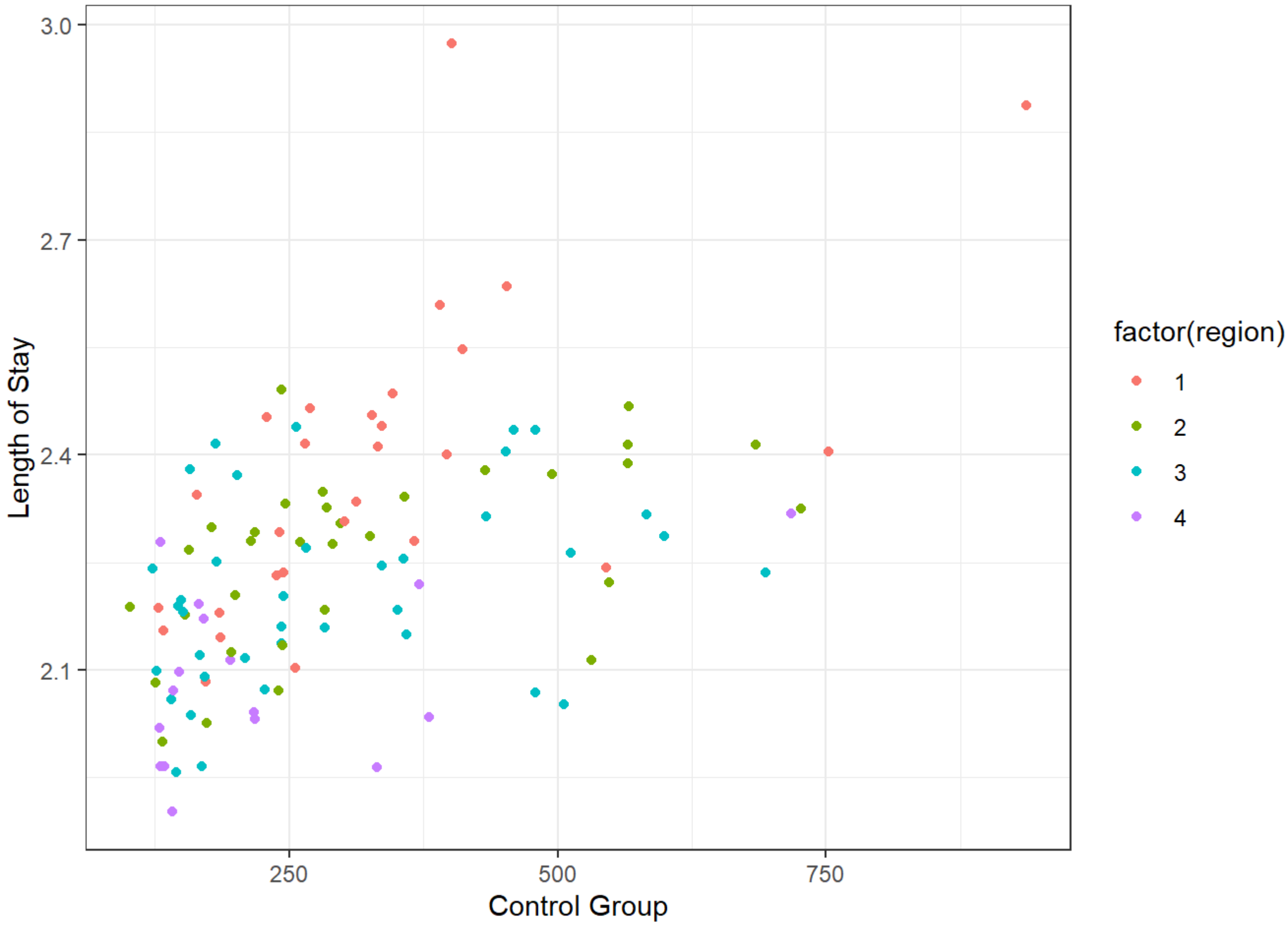



```
senicmod <- lm(formula = length_stay~age+routine_culturing_ratio+factor(region)+avg_daily_census
+facilities_services,data = senic)
summary(senicmod)
```

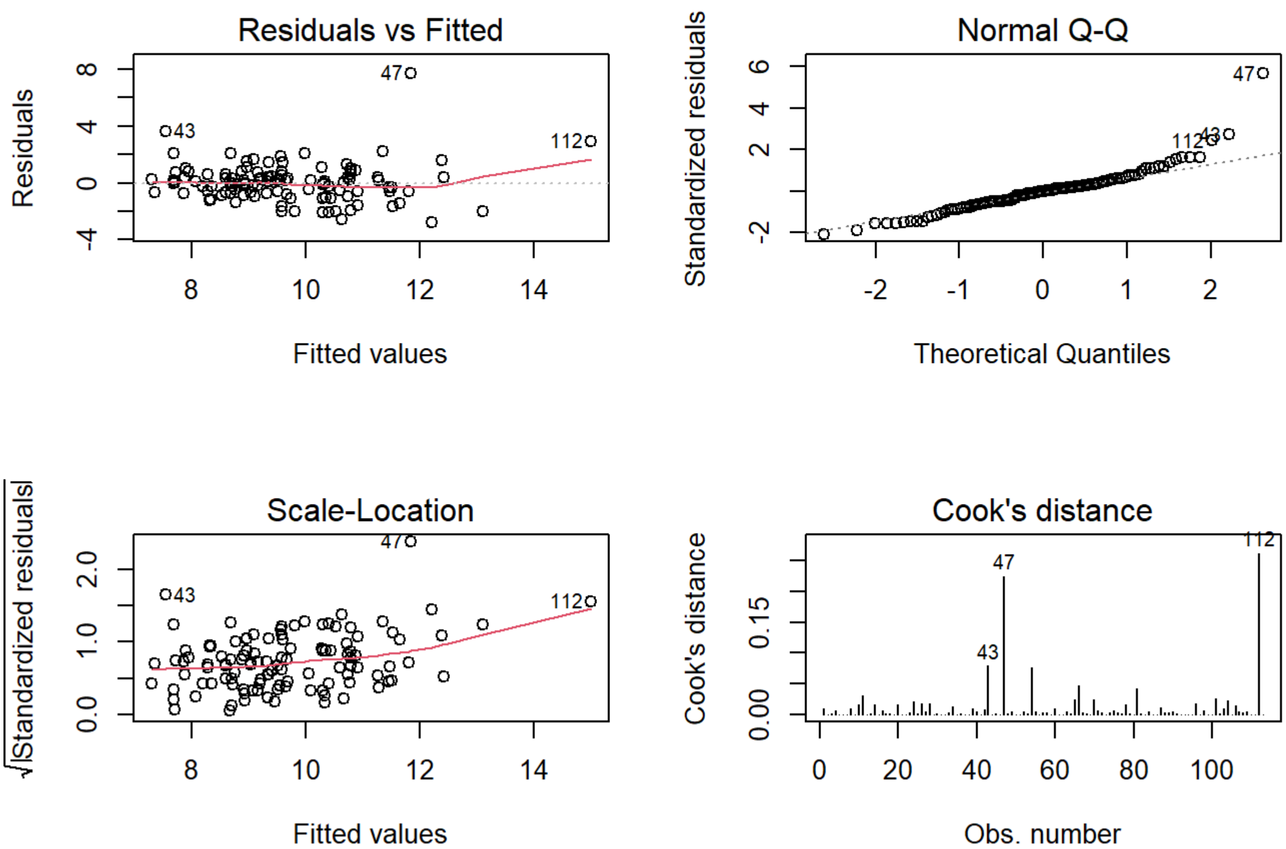
```
##
## Call:
## lm(formula = length_stay ~ age + routine_culturing_ratio + factor(region) +
##     avg_daily_census + facilities_services, data = senic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7938 -0.7304  0.0037  0.5388  7.7231
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.197818   1.878025   2.235 0.027519 *
## age                      0.103691   0.031459   3.296 0.001338 **
## routine_culturing_ratio  0.040302   0.014303   2.818 0.005781 **
## factor(region)2         -0.959655   0.381722  -2.514 0.013454 *
## factor(region)3         -1.516510   0.380092  -3.990 0.000123 ***
## factor(region)4         -2.149988   0.461517  -4.659 9.37e-06 ***
## avg_daily_census         0.006600   0.001404   4.700 7.92e-06 ***
## facilities_services     -0.020761   0.014369  -1.445 0.151477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.399 on 105 degrees of freedom
## Multiple R-squared:  0.4981, Adjusted R-squared:  0.4647
## F-statistic: 14.89 on 7 and 105 DF,  p-value: 2.283e-13
```

```
## Used log on the variables to make the model to have more linearity
log_senicplot <- ggplot(data = senic)
senic$log_age <- log(senic$age)
senic$log_ratio <- log(senic$routine_culturing_ratio)
senic$log_avgcensus <- log(senic$avg_daily_census)
senic$log_facilities <- log(senic$facilities_services)


log_senicplot+geom_point(mapping = aes(x = age + routine_culturing_ratio+avg_daily_census+facili
ties_services,y = log_stay, colour = factor(region)))+
  theme_bw()+
  labs(x = "Control Group",y = "Length of Stay")
```




```
## Created a residual plot
par(mfrow=c(2,2))
plot(senicmod, 1:4)
```




```
## Conducted parameter testing to check if each factor is significant
senicreduced1 <- lm(length_stay~routine_culturing_ratio+factor(region)+avg_daily_census+faciliti
es_services, data = senic)
anova(senicreduced1,senicmod)
```

```
## Analysis of Variance Table
##
## Model 1: length_stay ~ routine_culturing_ratio + factor(region) + avg_daily_census +
##     facilities_services
## Model 2: length_stay ~ age + routine_culturing_ratio + factor(region) +
##     avg_daily_census + facilities_services
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    106 226.61
## 2    105 205.36  1    21.248 10.864 0.001338 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
senicreduced2 <- lm(length_stay~age+factor(region)+avg_daily_census+facilities_services, data =
senic)
anova(senicreduced2,senicmod)
```

```
## Analysis of Variance Table
##
## Model 1: length_stay ~ age + factor(region) + avg_daily_census + facilities_services
## Model 2: length_stay ~ age + routine_culturing_ratio + factor(region) +
##     avg_daily_census + facilities_services
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    106 220.89
## 2    105 205.36  1    15.528 7.9392 0.005781 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
senicreduced3 <- lm(length_stay~age+routine_culturing_ratio+avg_daily_census+facilities_service
s, data = senic)
anova(senicreduced3,senicmod)
```

```
## Analysis of Variance Table
##
## Model 1: length_stay ~ age + routine_culturing_ratio + avg_daily_census +
##     facilities_services
## Model 2: length_stay ~ age + routine_culturing_ratio + factor(region) +
##     avg_daily_census + facilities_services
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    108 255.74
## 2    105 205.36  3    50.378 8.586 3.771e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
senicreduced4 <- lm(length_stay~age+routine_culturing_ratio+factor(region)+facilities_services,d
ata=senic)
anova(senicreduced4, senicmod)
```

```
## Analysis of Variance Table
##
## Model 1: length_stay ~ age + routine_culturing_ratio + factor(region) +
##     facilities_services
## Model 2: length_stay ~ age + routine_culturing_ratio + factor(region) +
##     avg_daily_census + facilities_services
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    106 248.57
## 2    105 205.36  1     43.21 22.093 7.921e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
senicreduced5 <- lm(length_stay~age+routine_culturing_ratio+factor(region)+avg_daily_census,data
=senic)
anova(senicreduced5,senicmod)
```

```
## Analysis of Variance Table
##
## Model 1: length_stay ~ age + routine_culturing_ratio + factor(region) +
##     avg_daily_census
## Model 2: length_stay ~ age + routine_culturing_ratio + factor(region) +
##     avg_daily_census + facilities_services
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    106 209.45
## 2    105 205.36  1     4.083 2.0876 0.1515
```

```
shapiro.test(residuals(senicmod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(senicmod)
## W = 0.89069, p-value = 1.361e-07
```