Please initial each line before you start

*XZ*__ The effort on this exam is entirely of my own.

*XZ*____ I will not discuss any part of this exam with anyone from the class or out of the class (roommates, classmates, friends, acquaintances, etc..)

*XZ*___ I will not seek any sort of assistance or advice from anyone other than the primary instructor.

Please sign below

I understand that if I violate the honor code or if there is any suspicion of violation, then I and others involved will be reported to the UC Davis Office of Judicial Affairs, which will then open a case for investigation. Until a decision is made, my grade will be on hold.

*Xingxuan Zhang 12/10/2019*_____

Sign and Date

In [13]:
```
install.packages('olsrr')
install.packages('gvlma')
```

executed in 3.37s, finished 16:59:59 2019-12-12

```
Warning message:
"package 'olsrr' is in use and will not be installed"

package 'gvlma' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\xingxuan_dell\AppData\Local\Temp\RtmpkBb4rU\downloaded_packages
```

In [14]:
```
library(stats)
library(PMCMR)
library(reshape2)
library(reshape)
library(tidyverse)
library(PASWR2)
library(pwr2)
library(olsrr)
library(gvlma)
```

executed in 36ms, finished 17:00:05 2019-12-12

1 (20 points) Question 1

In one of the class sessions on regression, we discussed the use of simple linear regression to estimate the relationship between the return on a stock and the market return. Think CAPM. The slope, called the beta coefficient, is used to measure how responsive a stock's price is to movements in the market. In financial investment parlance, the beta coefficient is used as a measure of a firm's systematic risk. In the file named Question 1, the return on the stock of three companies is provided: Dell, Sabre, and Walmart. Also provided is the return on the market (This is the value-weighted return computed by CRSP, the Center for Research on Security Prices). Five years of monthly returns are used so that there are a total of 60 observations for each company. Run the regression and answer the following questions:

In [1]:
```
setwd("D:/BAX441/Final Exam")
Q1<-read.csv('Question1.csv')
walmart<-Q1[,1]
dell<-Q1[,2]
sabre<-Q1[,3]
market<-Q1[,4]
```

executed in 89ms, finished 16:22:06 2019-12-13

1.1 (5 points) What are the beta coefficients for each of the three companies? Comment on the riskiness of each stock based on the beta coefficients.

In [28]:
```
####################WALMART#################################################
#### Correlation Coefficient and Scatterplot ####
#plot(walmart,market)
#abline(lm(market~walmart))
#cor(walmart,market)
#### Regression Model ####
lm.CAPM <- lm(walmart~market)
#anova(lm.CAPM)
summary(lm.CAPM)
```

executed in 29ms, finished 21:50:19 2019-12-13

```
Call:
lm(formula = walmart ~ market)

Residuals:
      Min        1Q    Median        3Q       Max
-0.212196 -0.057687  0.008831  0.043653  0.205777

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01977    0.01082   1.826 0.072961 .
market       0.73167    0.19121   3.826 0.000321 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08384 on 58 degrees of freedom
Multiple R-squared:  0.2016,    Adjusted R-squared:  0.1878
F-statistic: 14.64 on 1 and 58 DF,  p-value: 0.0003207
```

In [29]:
```
####################DELL################################################
#### Regression Model ####
lm.CAPM <- lm(dell~market)
summary(lm.CAPM)
```

executed in 24ms, finished 21:50:20 2019-12-13

```
Call:
lm(formula = dell ~ market)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23614 -0.12065 -0.00163  0.06605  0.40424

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02800    0.01846   1.517    0.135
market       1.66791    0.32605   5.116 3.69e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.143 on 58 degrees of freedom
Multiple R-squared:  0.3109,    Adjusted R-squared:  0.299
F-statistic: 26.17 on 1 and 58 DF,  p-value: 3.691e-06
```

In [30]:
```
####################SABRE################################################
#### Regression Model ####
lm.CAPM <- lm(sabre~market)
summary(lm.CAPM)
```

executed in 26ms, finished 21:50:21 2019-12-13

```
Call:
lm(formula = sabre ~ market)

Residuals:
      Min        1Q    Median        3Q       Max
-0.234055 -0.062216 -0.000204  0.048334  0.216329

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.002524   0.013899   0.182    0.857
market      1.470630   0.245524   5.990 1.42e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1077 on 58 degrees of freedom
Multiple R-squared:  0.3822,    Adjusted R-squared:  0.3715
F-statistic: 35.88 on 1 and 58 DF,  p-value: 1.417e-07
```

Comments

The beta coefficient for Walmart is 0.73167

The beta coefficient for Dell is 1.66791

The beta coefficient for Sabre is 1.470630

Walmart' beta coefficient is less than 1, so it is a conservative stock that are less risky than the market.

Dell's and Sabre's beta coefficient is more than 1, so they are more risky than the market.

According to the beta coefficients, Dell's stock is the riskest stock. Sabre is the second one. Walmart is the least risky stock among these three companies.

1.2 (5 points) Is there a linear relationship between the firm return and the market return for each of these three companies? Please state your hypotheses, the p-values, and your decision. Use a 5% level of significance.

$H_0 : \rho = 0$

$H_1 : \rho ! = 0$

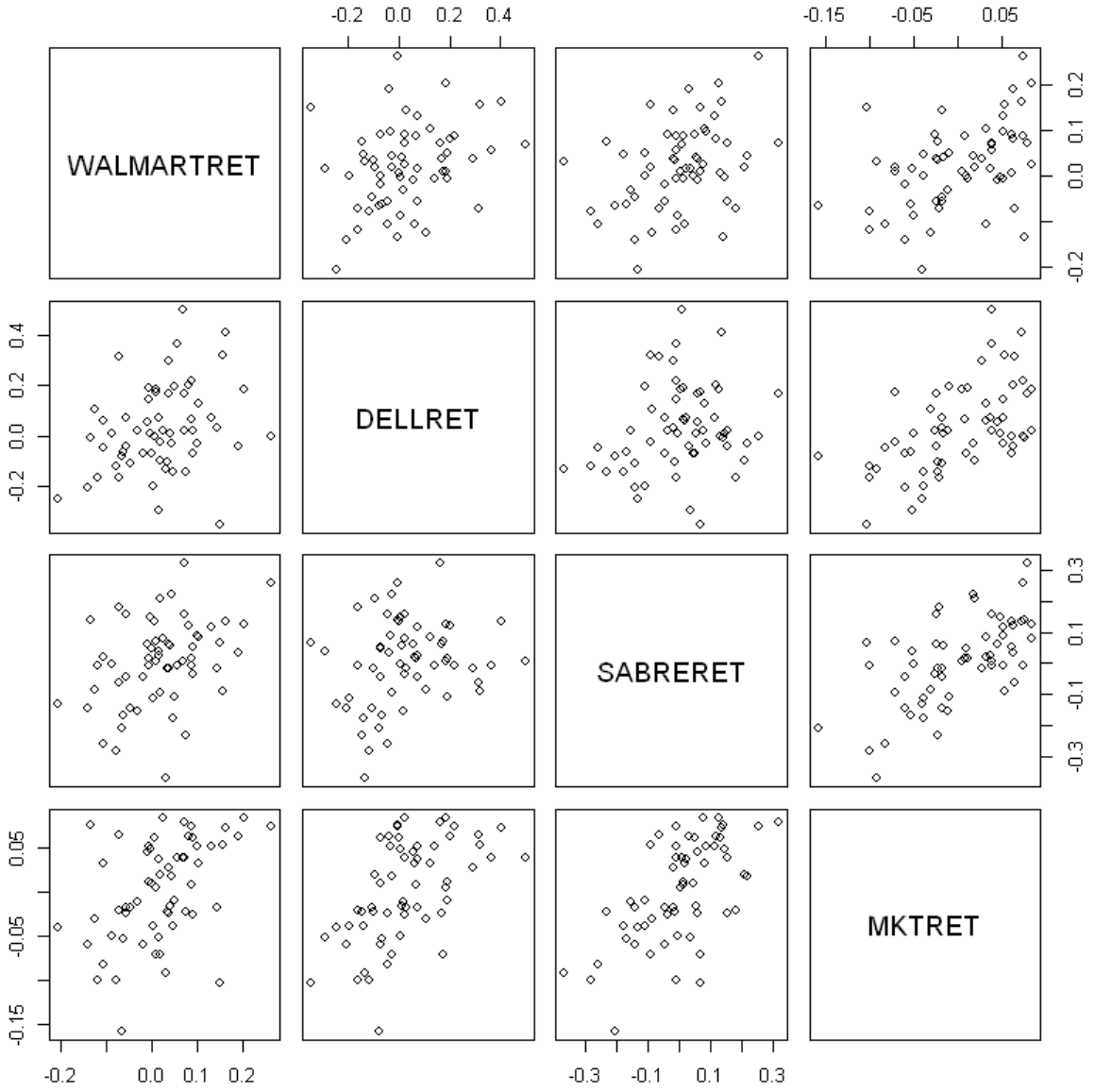
In [31]:
```
# Walmart
cor.test( Q1[,1],Q1[,4], alternative = "two.sided", method = "pearson")
# Dell
cor.test(Q1[,2],Q1[,4], alternative = "two.sided", method = "pearson")
# Sabre
cor.test( Q1[,3],Q1[,4], alternative = "two.sided", method = "pearson")

plot(Q1)
cor(Q1)

print('The p-value for walmart is 0.0003207. The p-value for Dell is 3.691e-06. The p-value for S
print('Since all companies have p-values that less than 0.05 for the correlation test, we can con
print('I also check the scatter plot, these companies do have linear relationship with the market
```

executed in 144ms, finished 21:54:33 2019-12-13

```
        Pearson's product-moment correlation

data:  Q1[, 1] and Q1[, 4]
t = 3.8264, df = 58, p-value = 0.0003207
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2201248 0.6309509
sample estimates:
      cor
0.4489551


        Pearson's product-moment correlation

data:  Q1[, 2] and Q1[, 4]
t = 5.1156, df = 58, p-value = 3.691e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3537575 0.7108695
sample estimates:
      cor
0.5575947


        Pearson's product-moment correlation

data:  Q1[, 3] and Q1[, 4]
t = 5.9898, df = 58, p-value = 1.417e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4321066 0.7537971
sample estimates:
      cor
0.6182008
```

	WALMARTRET	DELLRET	SABRERET	MKTRET
WALMARTRET	1.0000000	0.2998723	0.3693259	0.4489551
DELLRET	0.2998723	1.0000000	0.2138302	0.5575947
SABRERET	0.3693259	0.2138302	1.0000000	0.6182008
MKTRET	0.4489551	0.5575947	0.6182008	1.0000000

```
[1] "The p-value for walmart is 0.0003207. The p-value for Dell is 3.691e-06. The p-value for Sa
bre is 1.417e-07. "
[1] "Since all companies have p-values that less than 0.05 for the correlation test, we can conc
lude that, at 5% significant level, we can reject the null hypothesis. There is linear relations
hip between firms and the market return."
[1] "I also check the scatter plot, these companies do have linear relationship with the marke
t."
```



1.3 (5 points) Test to see if Dell's beta coefficient is greater than 1. Again, state your hypotheses, the p-values, and your decision. Use a 5% significance level. Reminder: This is a nonstandard case.

$H_0 : \beta \leq 1$

$H_1 : \beta > 1$

In [33]:
```
#### This is a non-standard case, so we cannot infer based on the T-Values and P-Values from R.
#### First, we need to find the t_critical value. We will use qt function in R.
#### Note the degrees of freedom (df) is n - k - 1 where k = number of predictors, which in this

t_critical <- qt(0.05, 60-1-1, lower.tail = FALSE)


#### Next, we need to find the t_test_statistic value.
#### t = (b - beta)/std. error

t_test_stat <- (1.66791 - 1)/ 0.32605


cat('P-value is ',pt(t_test_stat,58,lower.tail=FALSE))
```

executed in 27ms, finished 21:55:27 2019-12-13

```
P-value is  0.02252345
```

Decision.

Since the p-value is less than 0.05, we reject H0: beta <= 1.

We infer at 5% significance level that beta > 1.

Return on Dell stock is higher risky than the returns on the market.

1.4 (5 points) Test to see if Walmart's beta coefficient is less than 1. Again, state your hypotheses, the p-values, and your decision. Use a 5% significance level. Reminder: This is a nonstandard case.

$H_0 : \beta \geq 1$

$H_1 : \beta < 1$

In [35]:
```
#### This is a non-standard case, so we cannot infer based on the T-Values and P-Values from R.
#### First, we need to find the t_critical value. We will use qt function in R.
#### Note the degrees of freedom (df) is n - k - 1 where k = number of predictors, which in this

t_critical <- qt(0.05, 60-1-1, lower.tail = TRUE)


#### Next, we need to find the t_test_statistic value.
#### t = (b - beta)/std. error

t_test_stat <- (0.73167 - 1)/ 0.19121


pt(t_test_stat,58,lower.tail=TRUE)
```

executed in 28ms, finished 21:57:11 2019-12-13

```
0.0829253945767768
```

Decision.

Since the p-value is larger than 0.05, we cannot reject H0: beta >= 1.

We infer at 5% significance level that beta >= 1.

Return on Walmart stock is not less risky than the returns on the market.

2 (20 points) Question 2

In a large public agency in Sacramento, management is concerned that a supervisor has been reporting unusual costs for supplies and equipment within her cost center. The average value seems reasonable, but individual amounts swing wildly and just don't "look right." The management would like to have a systematic method to audit purchase amounts that signal when the amounts may be due to fraud. No one wants to accuse staff of fraud falsely, but no one wants to miss the problem either. In pursuit of this objective, the managers collected a sample of n=135 invoices approved by the supervisor. The data are available in the file Question 2. It is mathematically known that when numerical values range over several orders of magnitude (which our invoice data indicates), the first digit should not be uniformly distributed, but the leading digit should follow Benford's law. You are hired as a consultant to solve this case using a systematic, data-driven approach. Identify and run an appropriate statistical test to infer **if the purchase activities constitute evidence of fraud**. The management has asked for a report from your team. While you don't have to create a report for answering this final exam question, but I would love to hear your thoughts on **any strengths and shortcomings of your study that would go on the findings report**, if you were to create one. In addition, **what follow-up action items would you recommend on your report to the management of this public agency**?

In [17]:
```
setwd("D:/BAX441/Final Exam")
Q2<-read.csv('Question2.csv')
amount<-Q2[,2]
```

executed in 25ms, finished 17:10:16 2019-12-12

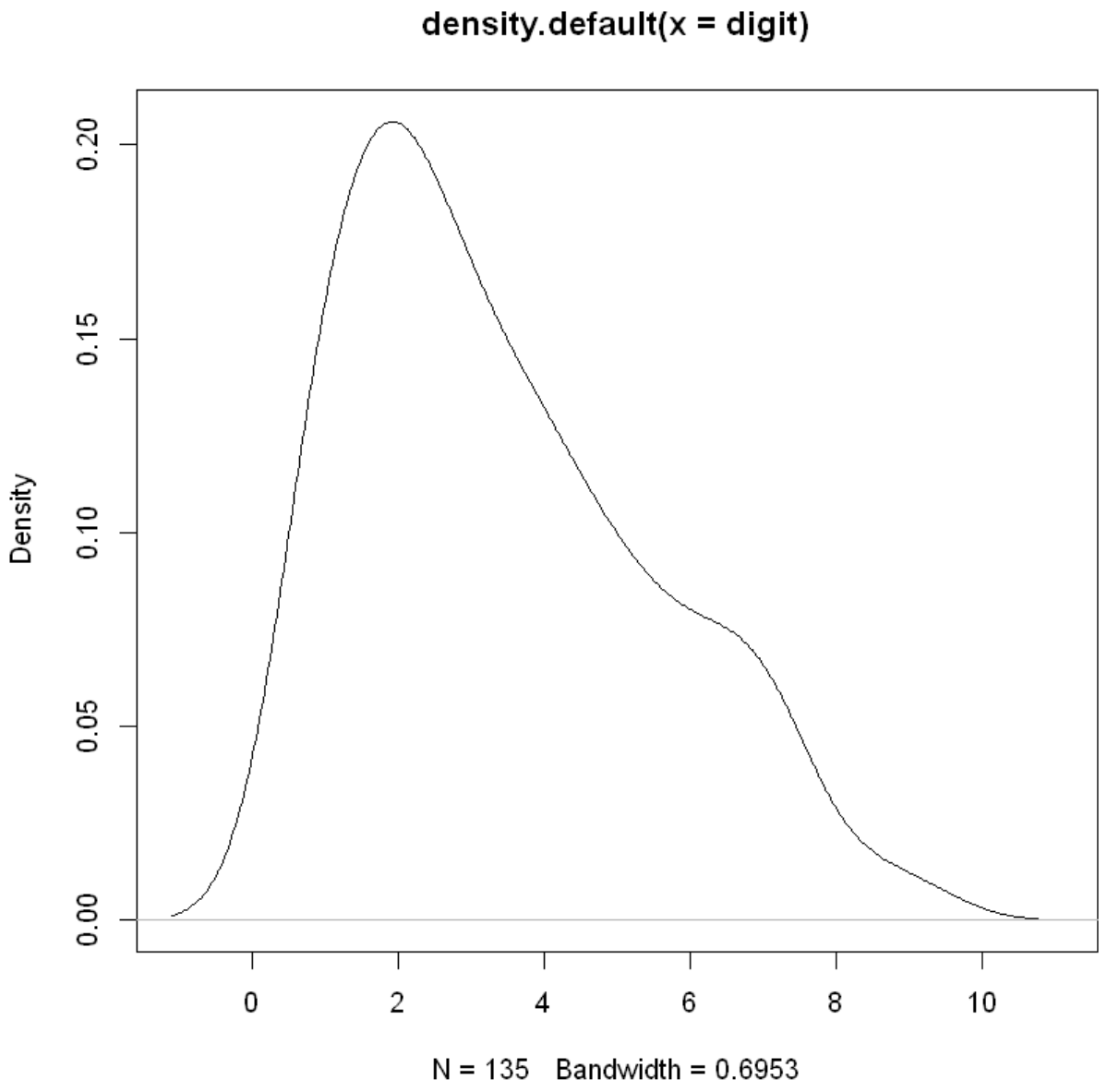
d	$P(d)$	Relative size of $P(d)$
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

In [27]:
```
bben <- function(k){
as.numeric(head(strsplit(as.character(k),'')[[1]],n=1))

}

# extract the first digit from all numbers computed

digit <- sapply(amount, bben)
```

executed in 31ms, finished 17:13:21 2019-12-12

In [34]:
```
plot(density(digit))
```

executed in 63ms, finished 17:14:32 2019-12-12



In [48]:
```
df<-data.frame(digit)
prop_1<-length(df[df$digit==1,])
prop_2<-length(df[df$digit==2,])
prop_3<-length(df[df$digit==3,])
prop_4<-length(df[df$digit==4,])
prop_5<-length(df[df$digit==5,])
prop_6<-length(df[df$digit==6,])
prop_7<-length(df[df$digit==7,])
prop_8<-length(df[df$digit==8,])
prop_9<-length(df[df$digit==9,])

prob<-rbind(prop_1,prop_2,prop_3,prop_4,prop_5,prop_6,prop_7,prop_8,prop_9)
num<-seq(1,9,by =1)
df_1<-cbind(num,prob)
```

executed in 33ms, finished 17:25:11 2019-12-12

HYPOTHESIS

$H_0 : p_1 = 0.301, p_2 = 0.176, \ldots, p_9 = 0.046$

H_1:At least one proportion differs from the Benford's Law

Significant Level: 5%

In [51]:
```
obs<-c(26,31,22,18,13,10,11,2,2)
exp<-c(0.301,0.176,0.125,0.097,0.079,0.067,0.058,0.051,0.046)
chisq.test(x=obs,p=exp)
```

executed in 32ms, finished 17:28:06 2019-12-12

```
        Chi-squared test for given probabilities

data:  obs
X-squared = 19.086, df = 8, p-value = 0.0144
```

CONCLUSION

At the 5% significant level, we can reject the null hypothesis that the data fit the Benford's Law. We have evidence that there is fraud in the data.

STRENGTH&SHORTCOMING

The chi-suqare test for the goodness of fit study is a powerful statistic technique, and the Benford's Law is a well-known rule for detecting fraud in financial industry. By doing this study, we can leverage the statistical tool to help the company with fraud detection. However, this test also has a disadvantage that it requires large size of sample to to chi-square approximation. If we cannot have enough data, we cannot do this test.

RECOMMENDATION

After this test, the company can sampling data again to repeat the teat and to see whether this study is robust. If it is, the company can do deeper investigation about this manager. In the future, this test can be apply to every case to help the company detect the fraud.

3 (5 points) Question 3

Let us say that you run a regression with two explanatory variables and notice that the p-value in the ANOVA table is extremely small but the p-values of both explanatory variables are larger than 0.05. What is the probable reason for this? Can you conclude that neither explanatory variable does a good job in predicting the dependent variable?

ANSWER

The anova table's p-value is significant, wich means the overall multiple linear regression does have meaning in this case. However, the insignificant p-values of both explanatory variables occurs, which means in this multiple linear refression, they do not jointly predict the the dependent variable well. One probable reson for this situation is that there may be multicollinearity between these two explanatory variables. The multicollinearity occurs when the two explantory variables have significant high correlation. In this case, we cannot say neither explantory variable does a good job in predicting the dependent variable. For the next step, I will detect the multicollinearity. If it do happens, I would remove one variable to see the result.

4 (5 points) Question 4

A biostatistician who designed a study for investigating the efficacy of a new medication at a leading U.S. pharmaceutical company was fired after the study. The tested null hypothesis stated that the drug is no better than a placebo. The t-test yielded a value of 20 in a study of 400 subjects. The null hypothesis was rejected showing that the drug has a beneficial effect. Why did the management fire the biostatistician?

ANSWER

The test for medication is very serious. In the question, the biostatistician only do 1 t-test on 400 subjects to get the conclusion. I think if the condition allows, the biostatistician should have more tests.

For the test method, the biostatistician only use a t-test on controlled group and the test group. However, he did not segment the subjects: whether they have the same level of the disease, whether they are the same age and gender. ANOVA test could be a better choice.

More questions arised with the distribution of the data of these 400 subjects. We do not know whether they are normal distributed, whether they have constant variation, and how these 400 subjects was seleced.

The question only say the t-statistic value is 20, but we do not know what is the t-critical value.

5 (10 points) Question 5

The file named Question 5 contains the following data for each of the 50 states: total expenditures on a state's criminal justice system (in millions of dollars) (EXPEND) total number of police employed in the state (POLICE)

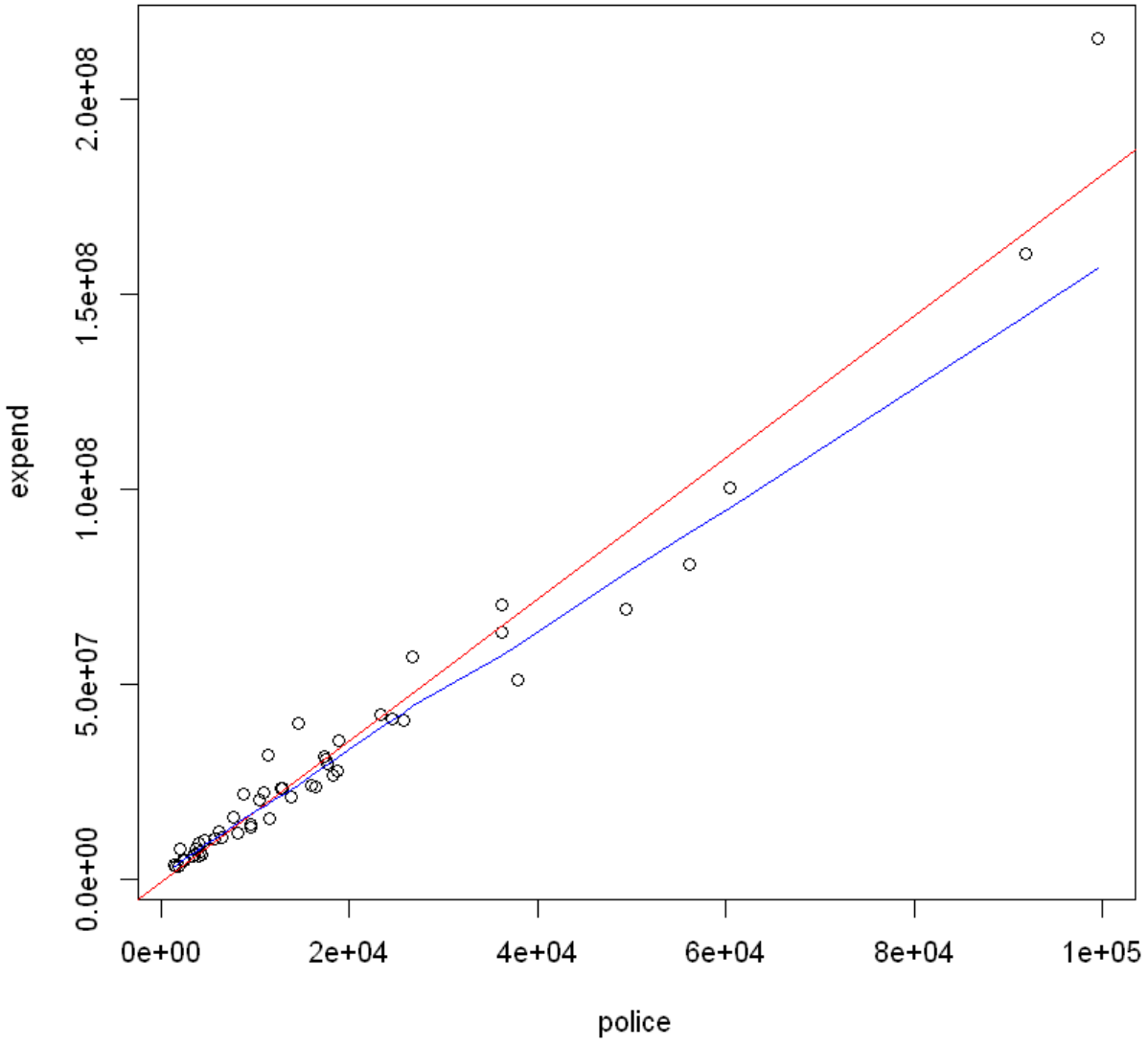
State governments must try to project spending in many areas. Expenditure on the criminal justice system is one area of continually rising cost. Your job is to build a model that can be used to forecast spending on a state's criminal justice system. Use one of the functional forms discussed in the class to build a model. Once your model is complete, predict expenditures for a state that plans to hire 10,000 police officers. Find a point prediction and a 95% prediction interval. The functional forms notes and solved examples should help. (Source: These data are a bit dated and were obtained from the U.S. Department of Criminal Justice website.)

In [16]:
```
setwd("D:/BAX441/Final Exam")
Q5<-read.csv('Question5.csv')
expend<-Q5[,2]
police<-Q5[,3]
```

executed in 28ms, finished 20:29:54 2019-12-13

In [11]:
```
plot(police, expend)
abline(lm(expend~police),col='red')
lines(lowess(police, expend),col='blue')
```

executed in 65ms, finished 15:16:59 2019-12-11



PLOT

The plot shows that the classical linear regression line (red) is very close to the local weighted line (blue). But, the scatter shows that the log-linear form may be better choices because we want to strech the left-down side.

In [14]:
```
###### LOG-LOG REGRESSION MODEL #################
###### Y = ln(Expenditure), X = ln(police) ############
###### ln(Expenditure) = beta0 + beta1*ln(police) + error #######
################################################
police_log<-log(police)
expend_log<-log(expend)
model_log_log<-lm(expend_log ~ police_log)
summary(model_log_log)
```

executed in 27ms, finished 15:28:31 2019-12-11

```
Call:
lm(formula = expend_log ~ police_log)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30248 -0.15516 -0.00623  0.08474  0.62813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.18519    0.24509   33.40   <2e-16 ***
police_log   0.92725    0.02632   35.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1999 on 49 degrees of freedom
Multiple R-squared:  0.962,     Adjusted R-squared:  0.9612
F-statistic:  1241 on 1 and 49 DF,  p-value: < 2.2e-16
```

MODEL

ln(expend)=8.18519+0.92725*ln(police)

In [28]:
```
point<-exp(8.18519+0.92725*log(10000))

sd_error_2<-(0.1999)^2
x_mean<-mean(police_log)
x_sd<-sd(police_log)
n<-length(police_log)
x_p<-log(10000)

s_p<-sqrt(sd_error_2*(1+1/n+((x_p-x_mean)^2/((n-1)*x_sd^2))))
t<-qt(0.975,n-2)
y_ln<-8.18519+0.92725*log(10000)

upper<-exp(y_ln+s_p*t)
lower<-exp(y_ln-s_p*t)

cat('The point estimate is ',point,'. ')
cat('The prediction intercal at 95% is (',lower,' , ',upper,').')
```

executed in 38ms, finished 15:50:38 2019-12-11

```
The point estimate is  18356203 . The prediction intercal at 95% is ( 12235342  ,  27539091 ).
```

6 (10 points) Question 6

The project manager at a construction company is evaluating how crew size affects the productivity of framing jobs. He has experimented with varying crew size (the number of workers) on a weekly basis over the past 27 weeks and has recorded productivity (jobs/week). The CSV file Question6 contains the data.

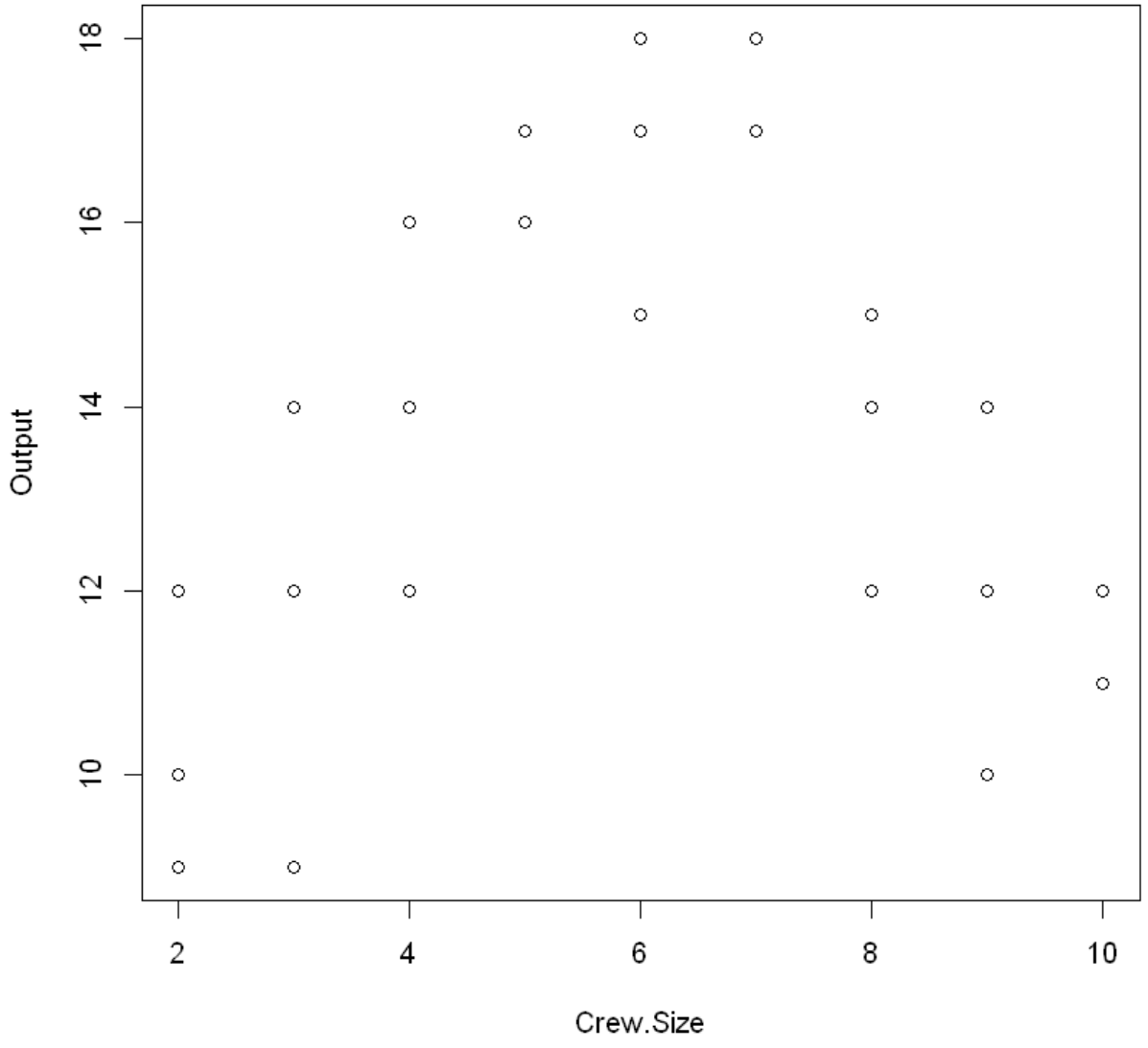
In [31]:
```
setwd("D:/BAX441/Final Exam")
Q6<-read.csv('Question6.csv')
size<-Q6[,1]
output<-Q6[,2]
```

executed in 23ms, finished 15:54:02 2019-12-11

6.1 (2 points) Create a scatterplot of the data. Based on the scatterplot alone, what crew size(s) seems optimal?

In [32]:
```
plot(Q6)
```

executed in 51ms, finished 15:54:47 2019-12-11



PLOT

From the plot, it seems size 6 or size 7 is the optimal size.

6.2 (5 points) Estimate the linear and the quadratic regression models. Evaluate the two models in terms of variable significance and adjusted R^2. Which model provides the best fit? Provide an intuitive justification for the chosen model.

In [36]:
```
summary(lm(output~size))
summary(lm(output~poly(size,2,raw=TRUE)))
```

executed in 32ms, finished 16:00:07 2019-12-11

```
Call:
lm(formula = output ~ size)

Residuals:
   Min     1Q Median     3Q    Max
-4.407 -2.130  0.037  2.426  4.259

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.0741     1.3666   9.567 7.77e-10 ***
size          0.1111     0.2092   0.531      0.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.807 on 25 degrees of freedom
Multiple R-squared:  0.01116,   Adjusted R-squared:  -0.0284
F-statistic: 0.2821 on 1 and 25 DF,  p-value: 0.6


Call:
lm(formula = output ~ poly(size, 2, raw = TRUE))

Residuals:
    Min      1Q  Median      3Q     Max
-3.5354 -0.8838  0.2525  1.2778  2.1919

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.11111    1.81800   1.161    0.257
poly(size, 2, raw = TRUE)1  4.59596    0.67574   6.801 4.92e-07 ***
poly(size, 2, raw = TRUE)2 -0.37374    0.05533  -6.754 5.50e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.682 on 24 degrees of freedom
Multiple R-squared:  0.6591,    Adjusted R-squared:  0.6307
F-statistic:  23.2 on 2 and 24 DF,  p-value: 2.462e-06
```

INTERPRET

From the results of these two model, we can see in the linear model, the size is not a significant variable to predict the output. The adjusted R^2 in the linear model is negative, which means the model is not fit for the data. However, in the quadratic model, the quadratic size is a significant variable to predict the output, and the adjusted R^2 became 0.6307 (better than linear's). Intuitively, the quadratic model is better than the linear model.

6.3 (1 point) Use the best-fitting model to predict how many jobs a crew of 5 could be expected to complete in a week.

MODEL

Output = 2.11111 + 4.59596*Size - 0.37374*Size^2

In [40]:
```
cat('The expected output in a week is ',2.11111 + 4.59596*5 - 0.37374*(5^2))
```

executed in 30ms, finished 16:10:28 2019-12-11

```
The expected output in a week is  15.74741
```

6.4 (2 points) Estimate the cubic regression model. Does it improve the fit as compared to the quadratic regression model?

In [41]:
```
summary(lm(output~poly(size,3,raw=TRUE)))
```

executed in 23ms, finished 16:10:50 2019-12-11

```
Call:
lm(formula = output ~ poly(size, 3, raw = TRUE))

Residuals:
    Min      1Q  Median      3Q     Max
-3.6178 -0.7374  0.1936  1.1380  2.3569

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.68519    4.22994   0.162   0.8727
poly(size, 3, raw = TRUE)1  5.54068    2.61205   2.121   0.0449 *
poly(size, 3, raw = TRUE)2 -0.55051    0.47483  -1.159   0.2582
poly(size, 3, raw = TRUE)3  0.00982    0.02619   0.375   0.7112
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.713 on 23 degrees of freedom
Multiple R-squared:  0.6612,    Adjusted R-squared:  0.617
F-statistic: 14.96 on 3 and 23 DF,  p-value: 1.288e-05
```

ANSWER

The cubic model does not improve.

7 (15 points) Question 7

Numerous attempts have been made to relate happiness to various factors. Since there is no unique way to quantify happiness, researchers generally rely on surveys to capture a subjective assessment of well-being. One study relates happiness with age and finds that holding everything else constant, people seem to be least happy when they are in their mid- to upper-40s (The Economist, December 16, 2010). Perhaps with greater age comes maturity that contributes to a better sense of overall well-being. With regard to the influence of money, a study from Princeton University's Woodrow Wilson School suggests that money does buy happiness, but its effect diminishes as incomes rise above $75,000 a year (Time Magazine, September 6, 2010). Perhaps people do not need more than 75,000 dollars to do what matters most to their emotional well-being, such as spending time with friends and family and meeting their basic food, health, and leisure needs. Nick Fisher is a young business school graduate who is fascinated by these reports. He decides to collect his own data to better comprehend and also verify the results of these studies. He surveys working adults in his hometown and inputs information on the respondent's self-assessed happiness on a scale of 0 to 100, along with age and family income. The data is shown on the CSV file Question7.

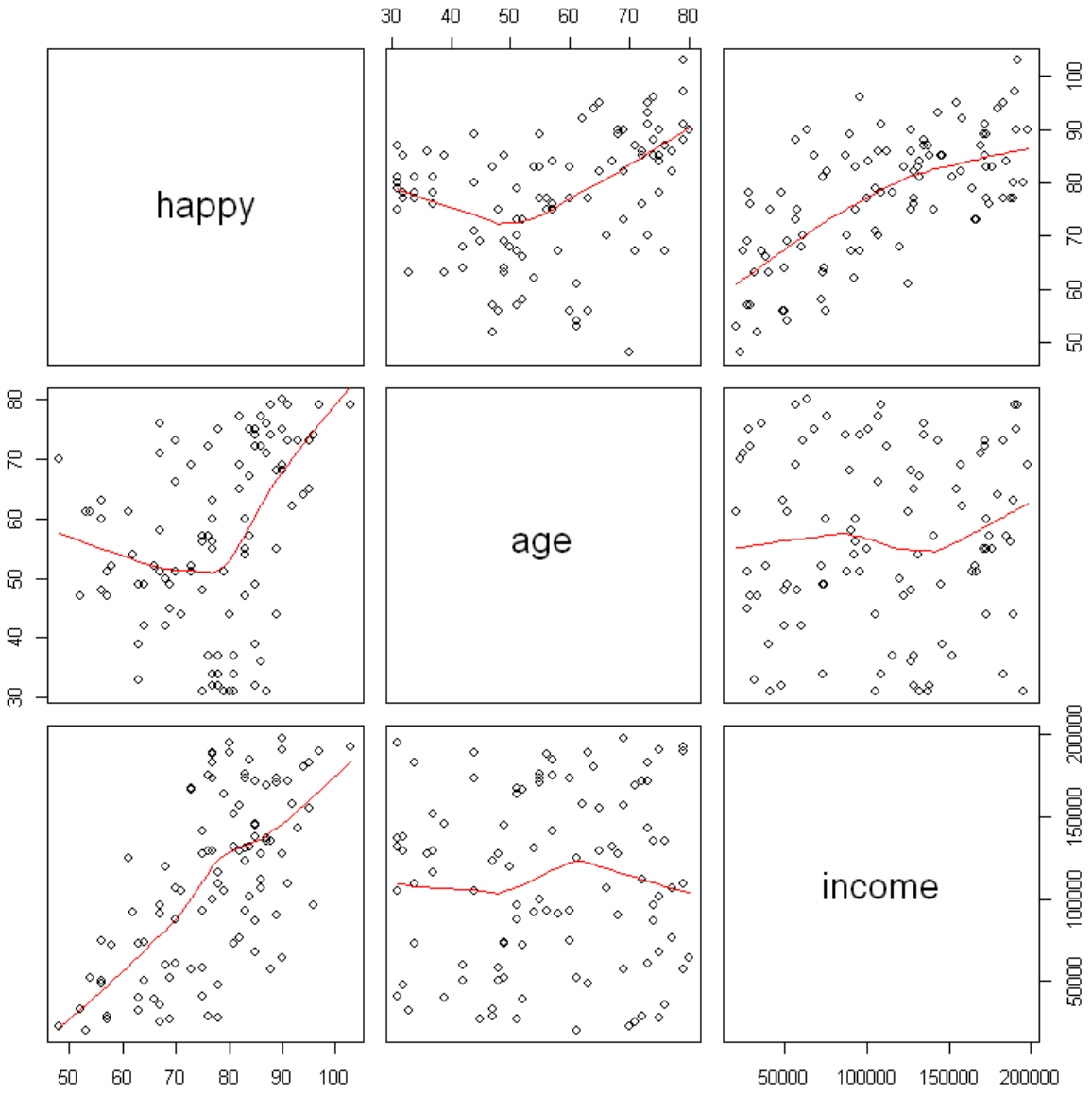
In [20]:
```
setwd("D:/BAX441/Final Exam")
Q7<-read.csv('Question7.csv')
happy<-Q7[,1]
age<-Q7[,2]
income<-Q7[,3]
```

executed in 27ms, finished 21:35:31 2019-12-13

7.1 a. (13 points) Run four models – linear, linear-log, log-linear, and log-log. Summarize the four models like my solved example shows. Compare the four functional forms and choose the one that fits the data the best.

In [21]:
```
pairs(cbind(happy,age,income), panel = panel.smooth)
```

executed in 104ms, finished 21:35:34 2019-12-13



In [22]:

```
summary(lm(happy~age+income))
summary(lm(happy~I(log(age))+I(log(income))))
summary(lm(I(log(happy))~I(log(age))+I(log(income))))
summary(lm(I(log(happy))~age+I(log(income))))
```

executed in 48ms, finished 21:37:49 2019-12-13

```
Call:
lm(formula = happy ~ age + income)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.908  -5.707   0.814   6.205  16.956 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.919e+01  3.709e+00  13.264  < 2e-16 ***
age         2.212e-01  5.714e-02   3.872 0.000196 ***
income      1.404e-04  1.593e-05   8.815 4.83e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.377 on 97 degrees of freedom
Multiple R-squared:  0.4967,    Adjusted R-squared:  0.4863 
F-statistic: 47.86 on 2 and 97 DF,  p-value: 3.466e-15


Call:
lm(formula = happy ~ I(log(age)) + I(log(income)))

Residuals:
     Min       1Q   Median       3Q      Max 
-20.8353  -4.8876   0.5392   5.9093  15.6381 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -107.318     19.611  -5.472 3.47e-07 ***
I(log(age))       9.735      2.972   3.275  0.00146 ** 
I(log(income))   12.707      1.380   9.206 6.93e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.355 on 97 degrees of freedom
Multiple R-squared:  0.4993,    Adjusted R-squared:  0.489 
F-statistic: 48.37 on 2 and 97 DF,  p-value: 2.684e-15


Call:
lm(formula = I(log(happy)) ~ I(log(age)) + I(log(income)))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.28533 -0.06502  0.01485  0.07624  0.20326 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.85176    0.26943   6.873 6.09e-10 ***
I(log(age))     0.11374    0.04084   2.785  0.00643 ** 
I(log(income))  0.17697    0.01896   9.332 3.71e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1148 on 97 degrees of freedom
Multiple R-squared:  0.4972,    Adjusted R-squared:  0.4869 
F-statistic: 47.96 on 2 and 97 DF,  p-value: 3.288e-15
```

```
Call:
lm(formula = I(log(happy)) ~ age + I(log(income)))

Residuals:
     Min       1Q   Median       3Q      Max
-0.28476 -0.05519  0.01050  0.07576  0.21054

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1542221  0.2156099   9.991  < 2e-16 ***
age            0.0027751  0.0007627   3.638 0.000442 ***
I(log(income)) 0.1765608  0.0184863   9.551 1.25e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1119 on 97 degrees of freedom
Multiple R-squared:  0.5222,    Adjusted R-squared:  0.5124
F-statistic: 53.01 on 2 and 97 DF,  p-value: 2.772e-16
```

	linear	lin-log	log-linear	log-log
intercept	4.92E+01	-107.3	1.8518	2.154222
age	2.21E-01	NA	NA	0.002775
income	1.40E-04	NA	NA	NA
ln(age)	NA	9.735	0.1137	NA
ln(income)	NA	12.707	0.177	0.176561
standard error	8.377	8.355	0.1148	0.1119
R^2	0.4967	0.4993	0.4972	0.5222
Adj. R^2	0.4863	0.489	0.4869	0.5124

According to the results, lin-log model is better than linear model, and log-log model is better than log-lin model.

In [49]:
```
#Lin-Log
pred<-predict(lm(happy~I(log(age))+I(log(income))))
cor<-cor(happy,pred)^2
1-((1-cor)*(length(happy)-1)/(length(happy)-2-1))
```

executed in 29ms, finished 17:23:37 2019-12-11

0.488994091896839

In [23]:
```
#Log-Log
pred<-predict(lm(I(log(happy))~age+I(log(income))))
cor<-cor(happy,exp(pred))^2
1-((1-cor)*(length(happy)-1)/(length(happy)-2-1))
```

executed in 31ms, finished 21:42:20 2019-12-13

0.523007513441615

ANSWER

The log-log model fits the data best.

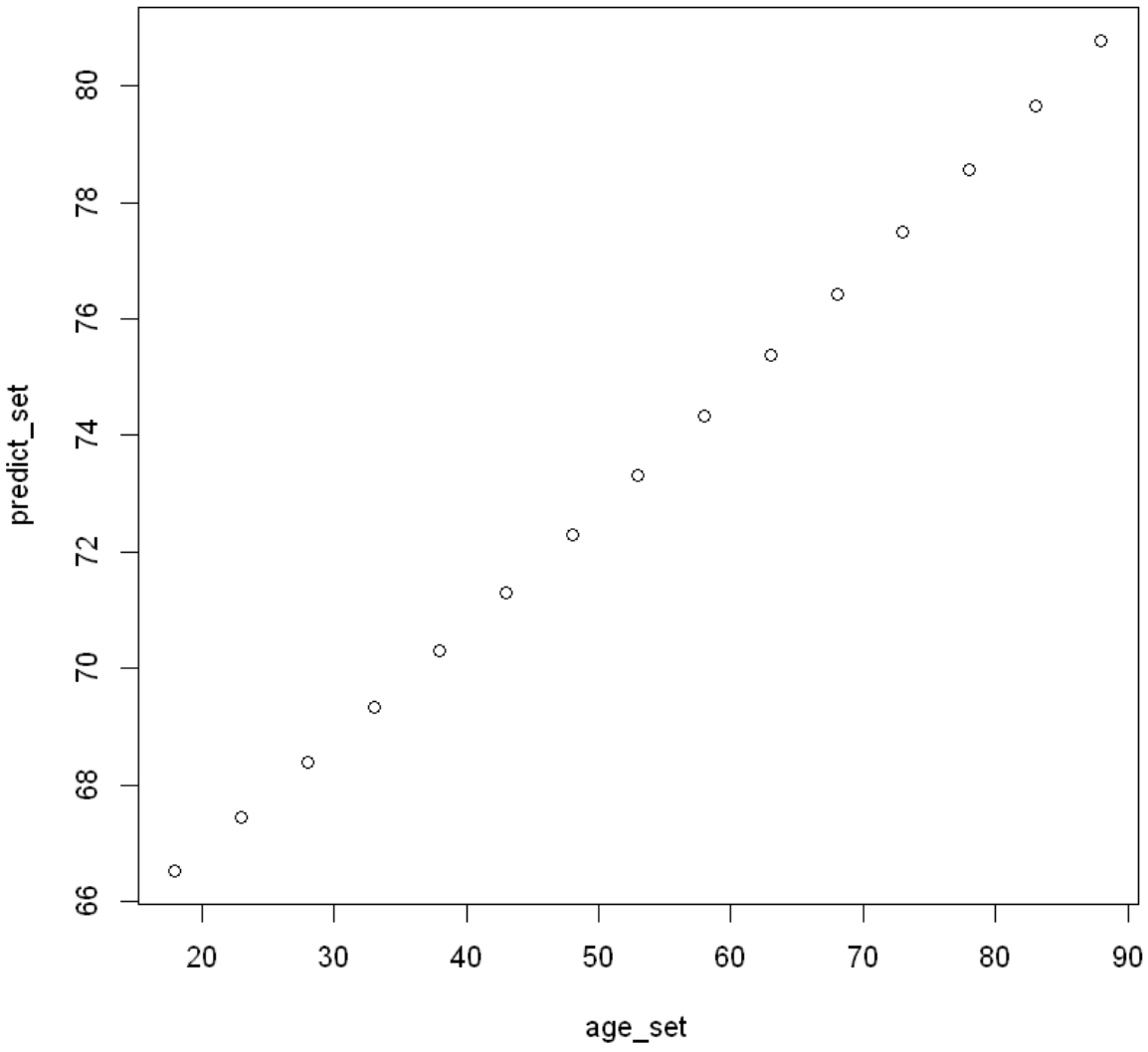
7.2 b. (1 point) Using the model chosen in part a, predict happiness associated with varying levels of age for a family with income of $80,000.

In [24]:

```
age_set<-seq(18,90,by=5)
predict_set<-exp(2.1542221+0.0027751*age_set+0.1765608*log(80000))
prediction<-cbind(age_set,predict_set)
prediction
plot(prediction)
```

executed in 73ms, finished 21:45:00 2019-12-13

age_set	predict_set
18	66.52028
23	67.44971
28	68.39213
33	69.34772
38	70.31667
43	71.29914
48	72.29535
53	73.30548
58	74.32972
63	75.36827
68	76.42133
73	77.48910
78	78.57180
83	79.66962
88	80.78278

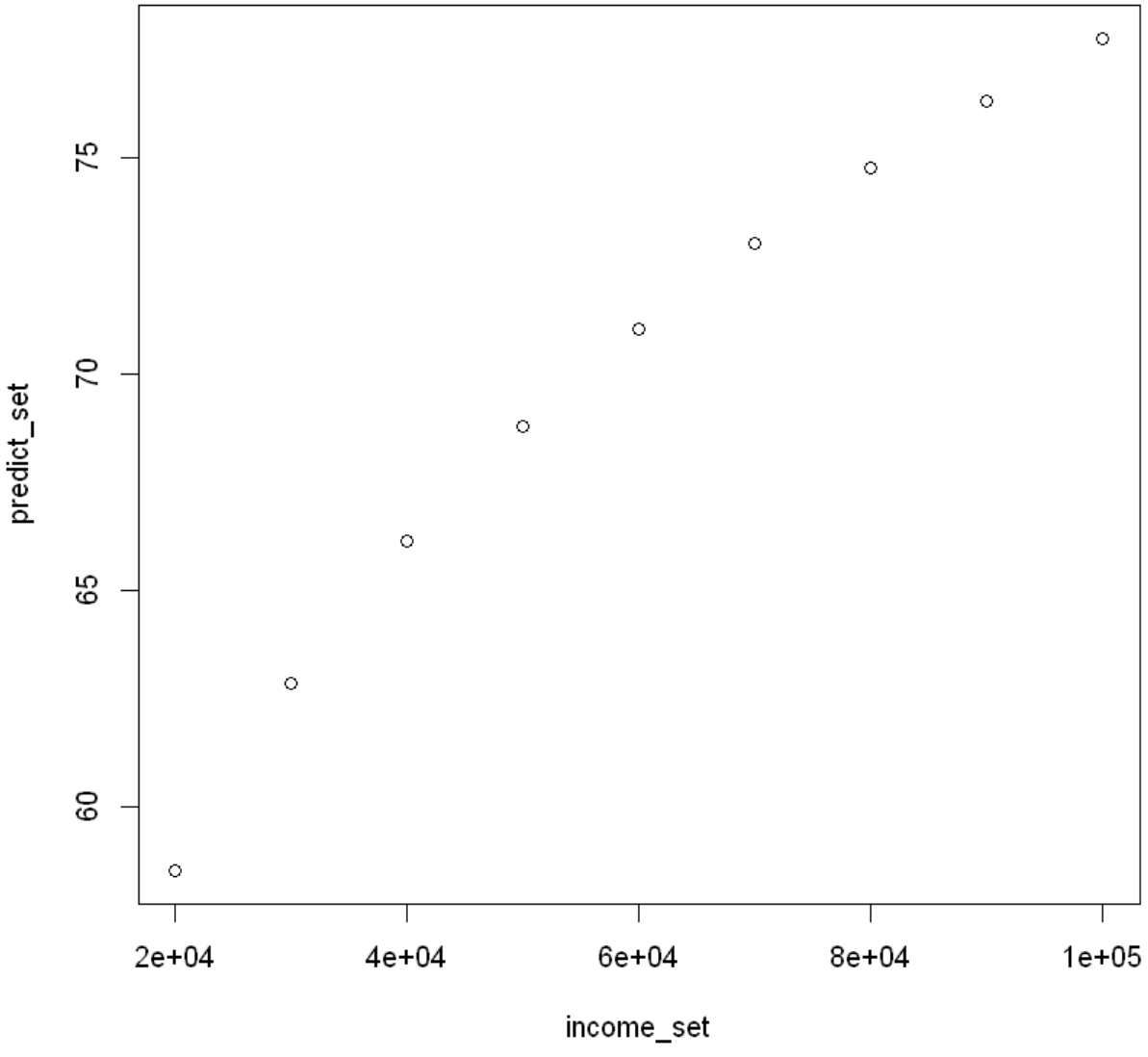


7.3 c. (1 point) Using the model chosen in part a, predict happiness associated with varying levels of family income for a 60-year-old working adult.

In [27]:
```
income_set<-seq(20000,100000,by=10000)
predict_set<-exp(2.1542221+0.0027751*60+0.1765608*log(income_set))
prediction<-cbind(income_set,predict_set)
prediction
plot(prediction)
```

executed in 63ms, finished 21:46:02 2019-12-13

income_set	predict_set
2e+04	58.51574
3e+04	62.85843
4e+04	66.13370
5e+04	68.79127
6e+04	71.04174
7e+04	73.00184
8e+04	74.74341
9e+04	76.31404
1e+05	77.74696



8 (10 points) Question 8

Traffic congestion on roads and highways costs industry billions of dollars annually as workers struggle to get to and from work. Several suggestions have been made about how to improve this situation, one of which is called flextime – workers are allowed to determine their own schedules (provided they work a full shift). Such workers will likely choose an arrival and departure time to avoid rush-hour traffic. In a preliminary experiment designed to investigate such a program, the general manager of a large company wanted to compare the times it took workers to travel from their homes to work at 8 a.m. with travel time under the flextime program. A random sample of 32 workers was selected. The employees recorded the time (in minutes) it took to arrive at work at 8 a.m. on Wednesday of one week. The following week, the same employees arrived at work at times of their own choosing. The travel time on Wednesday of that week was recorded. These results are listed in the CSV file Question8.

Run an appropriate statistical test to infer at the 5% significance level if travel times under the flextime program are different from travel times to arrive at work at 8 a.m.? Which conditions did you assess to select the statistical method?

In [18]:
```
setwd("D:/BAX441/Final Exam")
Q8<-read.csv('Question8.csv')
fix<-Q8[,1]
flex<-Q8[,2]
colnames(Q8)<-c('fix','flex')
df<-stack(Q8)
method<-as.factor(df[,2])
time<-as.numeric(df[,1])
```

executed in 33ms, finished 21:34:27 2019-12-13

In [76]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Unsatisfy, because the p-value of the normality test is less than 0.05
model_ex8 <- lm( time~  method)
resids_ex8 <- residuals(model_ex8)
preds_ex8 <- predict(model_ex8)

nortest::ad.test(resids_ex8)

#hist(resids_ex8)
#qqnorm(resids_ex8)
#qqline(resids_ex8)

#3. Variances are constant - Satisfy, because the p-value of the normality test is larger than 0.(
fligner.test( time~  method)
car::leveneTest(time~  method)
```

executed in 135ms, finished 18:29:52 2019-12-11

```
        Anderson-Darling normality test

data:  resids_ex8
A = 1.0586, p-value = 0.008208


        Fligner-Killeen test of homogeneity of variances

data:  time by method
Fligner-Killeen:med chi-squared = 0.036304, df = 1, p-value = 0.8489
```

	Df	F value	Pr(>F)
group	1	0.006560152	0.935707
	62	NA	NA

Wilcoxon Rank Sum Test for paired 2 sample interval data

$H_0 : \mu_{fix} = \mu_{flex}$

$H_1 : \mu_{fix} ! = \mu_{flex}$

Significant Leve: 5%

In [78]:
```
wilcox.test(time~  method,alt='two.sided',paired=TRUE,exacted=FALSE,conf.level=0.95)
```

executed in 29ms, finished 18:35:39 2019-12-11

```
Warning message in wilcox.test.default(x = c(34, 35, 43, 46, 16, 26, 68, 38, 61, :
"cannot compute exact p-value with ties"


        Wilcoxon signed rank test with continuity correction

data:  time by method
V = 367.5, p-value = 0.05262
alternative hypothesis: true location shift is not equal to 0
```

CONCLUSION

At 5% significant level, we cannot reject the null hypothesis. There is no different in travel times between fixtime and flextime program.

9 (5 points) Question 9

An auditor counted the number of filing errors associated with random samples of three types of mortgages at four banks. A filing error might, for instance, be a missing signature or incomplete paperwork. The mortgage types are traditional, refinancing, and home equity loan. What would it mean to find an interaction between the type of loan and the bank?

ANSWER

The interaction between the type of loan and the bank means every bank has its own shortness and advantages on filling errors. One bank has the least errors in one type of loans may have more errors in other type of loans compare with other two banks. No bank is better than others, and no type of loans is better than others. The effect of interaction will impact the result of the ANOVA test. This means that we need not to test the main effect.

10 (2 points) Question 10

What is the difference between a standard deviation and a standard error?

ANSWER

The standard deviation is a measure of the variability of a random variable.

A standard error is a measure of how precise an estimate is. The estimate may be a mean, or a regression coefficient. The standard error is the standard deviation of the estimator in many samples from the population.

11 (5 points) Question 11

Suppose that you have generated three alternative multiple regression models to explain the variation in a particular regressand. The regression output for each model can be summarized as follows:

	Model 1	Model 2	Model 3
No. of regressors	4	6	9
R^2	0.76	0.77	0.79
Adjusted R^2	0.75	0.74	0.73

All three models have the same regressand. Which of the models would you select and why? Be precise yet succinct.

ANSWER

I would choose model 1 because its adjusted R^2 is the highest one. The R^2 would increase when the number of regressors increases, so in this question we need not to check the R^2.

12 (10 points) Question 12

The WinterFun Company sells winter sports merchandise including skis, ice skates, sleds, and so on. Quarterly sales (in thousands of dollars) for the WinterFun company are shown on the CSV file Question12. The time period represented starts in the first quarter of 2008 and ends in the fourth quarter of 2017.

A linear regression containing only the time variable like shown below is called a **linear trend model.**

$y=\beta_1 + \beta_2 t + \varepsilon$

One of the common objectives of such analysis is if the linear trend model is sufficient for predicting sales or are the sales influenced by seasonality. For this question, follow the steps provided in the parts below to construct a final model:

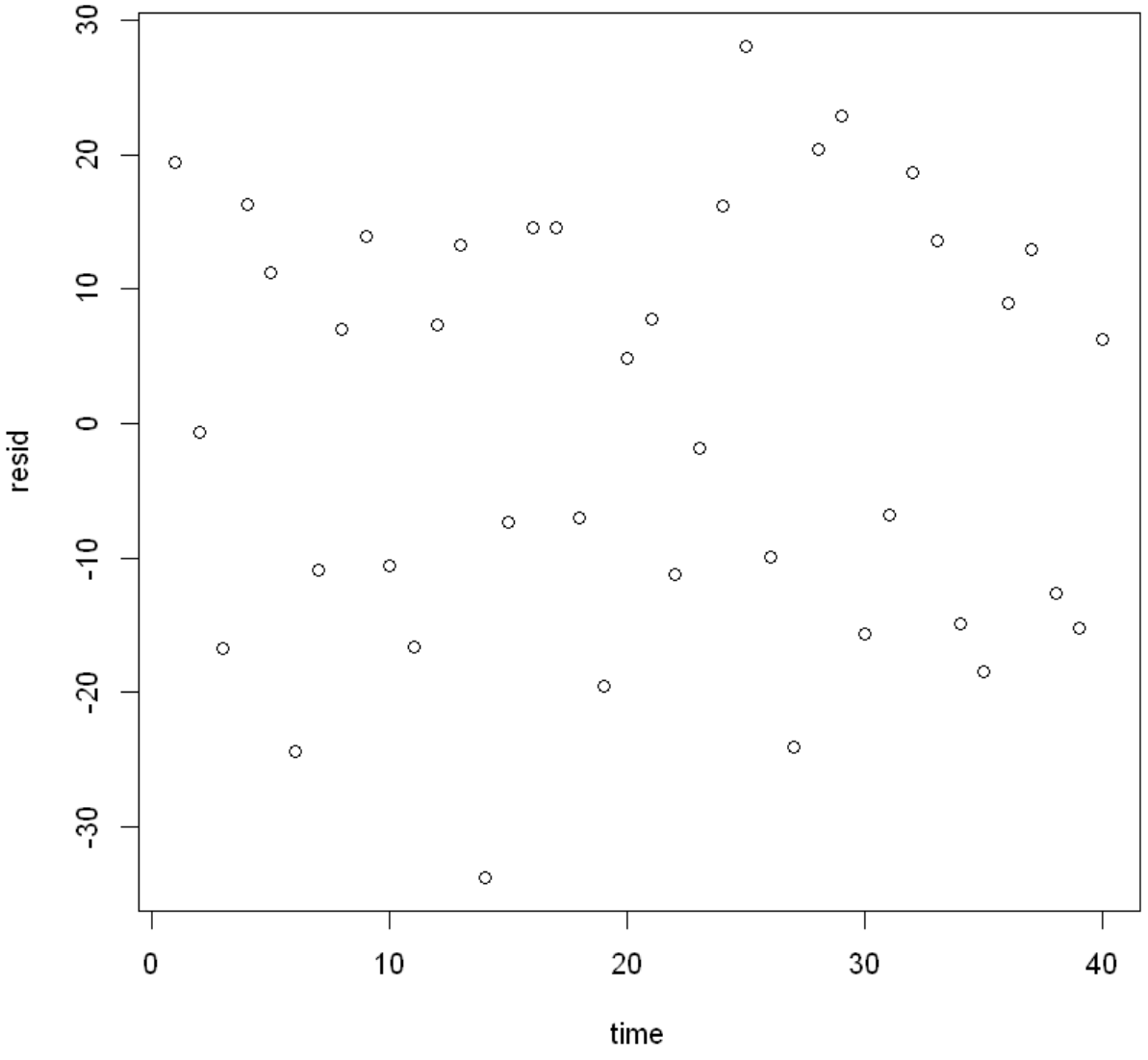
In [13]:
```
setwd("D:/BAX441/Final Exam")
Q12<-read.csv('Question12.csv')
```

executed in 48ms, finished 16:52:24 2019-12-13

12.1 a. (3 points) Create a linear trend model. Write the model. Examine the residual plot (versus fits) for this model. Does it show any patterns?

In [4]:
```
time<-Q12[,1]
sales<-Q12[,4]
model1<-lm(sales~time)
resid<-residuals(model1)
plot(time, resid)
```

executed in 67ms, finished 14:47:09 2019-12-12



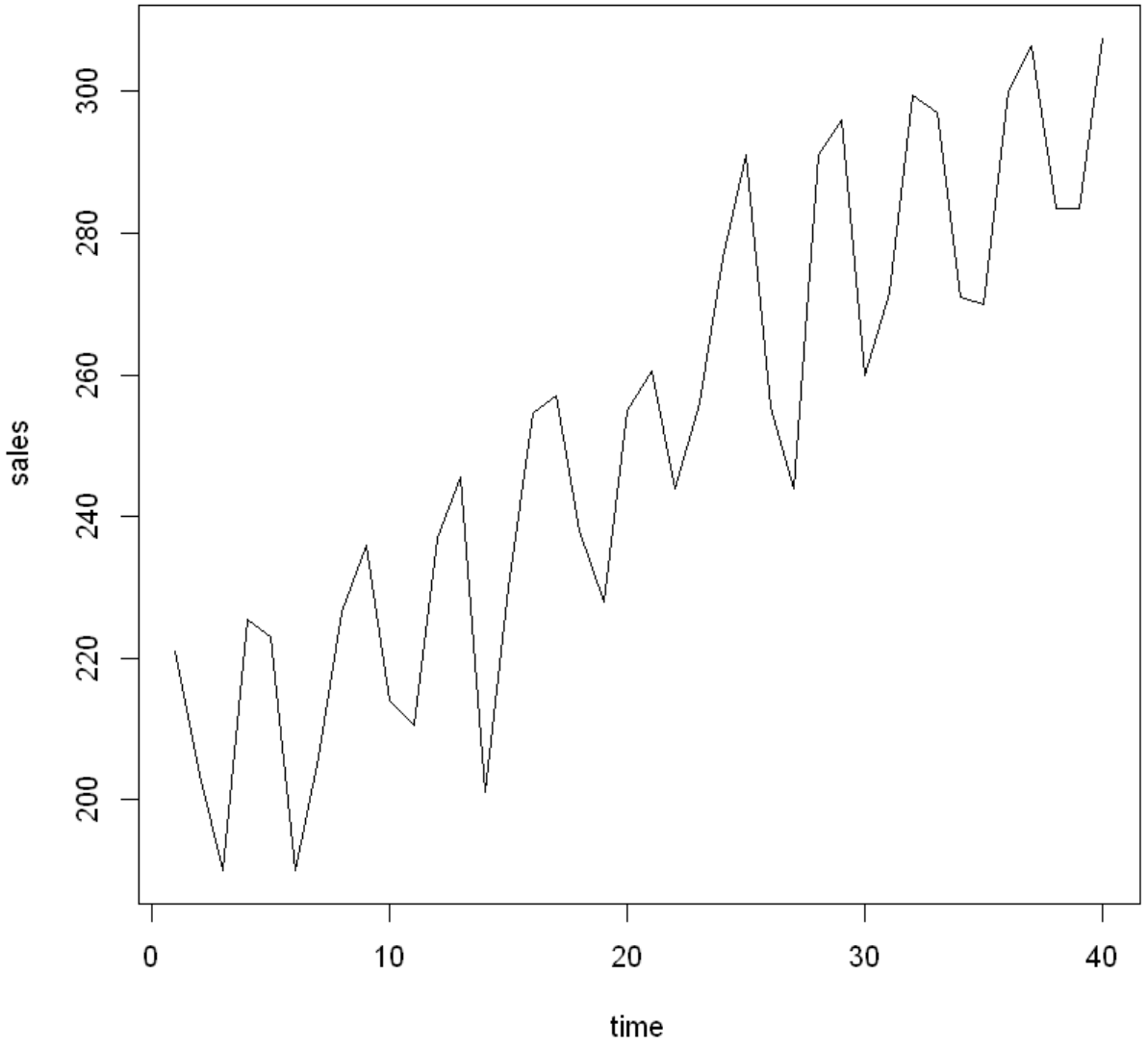
ANSWER

No, there is no pattern in the residual vs. fitted plot. Thus, the variance is constant.

12.2 b. (2 points) Now, you want to see if there is any seasonality. For this create a time plot of sales (Sales versus Time) and check if there are seasonal patterns. Examine evidence of seasonality visually.

In [6]:
```
plot(time,sales,type='l')
```

executed in 50ms, finished 14:49:07 2019-12-12



ANSWER

From the plot, I think there is seasonality in the data.

12.3 c. Create indicator variables for quarters. No need to show your R code.

In [8]:
```
q2<-ifelse(Q12[,3]==2, 1,0)
q3<-ifelse(Q12[,3]==3,1,0)
q4<-ifelse(Q12[,3]==4,1,0)

summary(lm(sales~q2+q3+q4+time))
```

executed in 43ms, finished 14:52:54 2019-12-12

```
Call:
lm(formula = sales ~ q2 + q3 + q4 + time)

Residuals:
    Min      1Q  Median      3Q     Max
-19.653  -4.478  -1.429   4.148  13.640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 214.59413    2.95039  72.734  < 2e-16 ***
q2          -29.86610    3.21712  -9.284 5.72e-11 ***
q3          -29.53220    3.22168  -9.167 7.85e-11 ***
q4           -3.74830    3.22927  -1.161    0.254
time          2.56610    0.09895  25.932  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.19 on 35 degrees of freedom
Multiple R-squared:  0.9593,    Adjusted R-squared:  0.9546
F-statistic: 206.1 on 4 and 35 DF,  p-value: < 2.2e-16
```

12.4 d. (5 points) Conduct a Partial F-test to assess if the seasonal indicator variables are necessary in the model. Write the full model, the reduced models, and the hypotheses clearly. The reduced model in this case is the linear trend model and the full model is the one with indicator variables. Write your interpretation.

MODELS

Reduced Model: $Y_{sales} = \beta_0 + \beta_1 * X_{time} + \varepsilon$

Full Model: $Y_{sales} = \beta_0 + \beta_1 * X_{time} + \beta_2 * X_{quarter2_dummy} + \beta_3 * X_{quarter3_dummy} + \beta_4 * X_{quarter4_dummy} + \varepsilon$

HYPOTHESIS

$H_0 : \beta_2 = \beta_3 = \beta_4$

H_1: At least 1 β is not 0

In [9]:
```
reduced_model<-lm(sales~time)
full_model<-lm(sales~q2+q3+q4+time)
anova(reduced_model,full_model)
```

executed in 31ms, finished 15:01:31 2019-12-12

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
38	9622.061	NA	NA	NA	NA
35	1809.507	3	7812.554	50.37089	8.72319e-13

INTERPRETATION

From the partial F-test, the p-value is less any significant level, so we reject the null hypothesis. At least 1 quarter dummy variable is meaningful. Further, we can test which dummy variable is meaningful to the model.

13 (5 points) Question 13

A bank encouraged customers who have its debit cards to use them more frequently. It collected a sample of customers and found the number of times that each used the debit card during the four-month promotion March – June. To learn whether the average number of charges over this period is different in any month, the bank plans to use a one-way ANOVA. Would this be the right analysis? If not, what should the bank do?

ANSWER

No, I do not think this is the right analysis. Since the bank is comparing the use times for each customers in the sample, I think the randomized block design should be should be considered.

14 (25 points) Question 14

A technology company in the bay area is interested in conducting a study of the factors that affect absenteeism among its employees. After reviewing the literature on absenteeism and interviewing a number of employees, the analyst in charge of the project defined the variables as shown below:

Absenteeism (ABSENT): The number of distinct occasions that the employee was absent during a year. Each occasion consists of one or more consecutive days of absence.

Job Complexity (COMPLX): An index ranging from 0 to 100.

Seniority (SENIOR): Number of complete years with the company on December 31st of the year in which the study was completed.

The CSV file Question14 contains data on 77 employees. The dependent variable is absenteeism (ABSENT). The possible explanatory variables are

COMPLX = measure of job complexity SENIOR = seniority SATIS = categorical (qualitative) variable – response to “How satisfied are you with your manager?”

In this question, use SENINV = 1/SENIOR and COMPLX as two of the explanatory variables. The SENINV is the reciprocal of the seniority variable and the variable SATIS should be transformed into indicator variables as follows:

FS1 = 1 if SATIS = 1 (very dissatisfied) = 0 otherwise

FS2 = 1 if SATIS = 2 (somewhat dissatisfied) = 0 otherwise

FS3 = 1 if SATIS = 3 (neither satisfied nor dissatisfied) = 0 otherwise

FS4 = 1 if SATIS = 4 (somewhat satisfied) = 0 otherwise

FS5 = 1 if SATIS = 5 (very satisfied) = 0 otherwise

Five indicator variables are created to represent all five supervisor satisfaction categories. Recall that only four need to be used in the regression.

In [10]:
```
setwd("D:/BAX441/Final Exam")
Q14<-read.csv('Question14.csv')
absent<-Q14[,1]
complex<-Q14[,2]
fs2<-ifelse(Q14[,3]==2,1,0)
fs3<-ifelse(Q14[,3]==3,1,0)
fs4<-ifelse(Q14[,3]==4,1,0)
fs5<-ifelse(Q14[,3]==5,1,0)
senior<-Q14[,4]
```

executed in 34ms, finished 15:11:05 2019-12-12

14.1 a. (3 points) Run the regression containing all the independent variables (continuous + indicator). This is your full model.

In [11]:
```
full_model<-lm(absent~complex+I(1/senior)+fs2+fs3+fs4+fs5)
summary(full_model)
```

executed in 24ms, finished 15:12:08 2019-12-12

```
Call:
lm(formula = absent ~ complex + I(1/senior) + fs2 + fs3 + fs4 +
    fs5)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7326 -0.8154 -0.0236  0.6371  3.6174

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.534709   0.608359   4.166  8.7e-05 ***
complex     -0.014070   0.006333  -2.222   0.0295 *
I(1/senior)  1.102841   0.453456   2.432   0.0176 *
fs2          0.700092   0.602709   1.162   0.2494
fs3         -0.493028   0.523048  -0.943   0.3491
fs4         -0.834578   0.583957  -1.429   0.1574
fs5         -0.904105   0.837699  -1.079   0.2842
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.252 on 70 degrees of freedom
Multiple R-squared:  0.3405,    Adjusted R-squared:  0.284
F-statistic: 6.024 on 6 and 70 DF,  p-value: 4.04e-05
```

14.2 b. (3 points) Run the regression after removing the indicator variables. This is your reduced model.

In [13]:
```
reduced_model<-lm(absent~complex+I(1/senior))
summary(reduced_model)
```

executed in 25ms, finished 15:13:06 2019-12-12

```
Call:
lm(formula = absent ~ complex + I(1/senior))

Residuals:
    Min      1Q  Median      3Q     Max
-2.3782 -1.1363 -0.0130  0.7584  3.9146

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.047420   0.428909   4.774  8.9e-06 ***
complex     -0.012060   0.006242  -1.932  0.05718 .
I(1/senior)  1.281891   0.469632   2.730  0.00792 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.332 on 74 degrees of freedom
Multiple R-squared:  0.2104,    Adjusted R-squared:  0.189
F-statistic: 9.857 on 2 and 74 DF,  p-value: 0.0001602
```

14.3 c. (5 points) Using Partial F-test, identify if the indicator variables (degree of satisfaction) are needed to determine the behavior of absenteeism in the employees. Use 5% level of significance.

In [14]:
```
anova(reduced_model,full_model)
```

executed in 30ms, finished 15:13:37 2019-12-12

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
74	131.3666	NA	NA	NA	NA
70	109.7132	4	21.65332	3.453849	0.01237738

HYPOTHESIS

$H_0 : \beta_2 = \beta_3 = \beta_4$

H_1: At least 1 β is not 0

ANSWER

At 5% significant level, we can reject the null hypothese. At least 1 indicator variable is needed to determine the behavior of absenteeism in the employees.

14.4 d. (3 points) Use the appropriate model (after determining the final model from the results of the partial F-test) and calculate the average absenteeism rate for all employees with COMPLX = 60 and SENIOR = 30 who were very dissatisfied with their managers.

In [15]:
```
full_model
```
executed in 30ms, finished 15:19:35 2019-12-12

```
Call:
lm(formula = absent ~ complex + I(1/senior) + fs2 + fs3 + fs4 +
    fs5)

Coefficients:
(Intercept)      complex  I(1/senior)          fs2          fs3          fs4
    2.53471     -0.01407      1.10284      0.70009     -0.49303     -0.83458
        fs5
   -0.90411
```

In [17]:
```
cat('The average absenteeism rate is ',2.53471-0.01407*60+1.10284*(1/30))
```
executed in 21ms, finished 15:20:58 2019-12-12

```
The average absenteeism rate is  1.727271
```

14.5 e. (3 points) Use the appropriate model (after determining the final model from the results of the partial F-test) and calculate the average absenteeism rate for all employees with COMPLX = 60 and SENIOR = 30 who were very satisfied with their managers.

In [18]:
```
cat('The average absenteeism rate is ',2.53471-0.01407*60+1.10284*(1/30) -0.90411)
```
executed in 21ms, finished 15:22:09 2019-12-12

```
The average absenteeism rate is  0.8231613
```

14.6 f. (3 points) Use the appropriate model (after determining the final model from the results of the partial F-test) and calculate the average absenteeism rate for all employees with COMPLX = 10 and SENIOR = 3 who were very dissatisfied with their managers.

In [19]:
```
cat('The average absenteeism rate is ',2.53471-0.01407*10+1.10284*(1/3))
```
executed in 19ms, finished 15:22:32 2019-12-12

```
The average absenteeism rate is  2.761623
```

14.7 g. (3 points) Use the appropriate model (after determining the final model from the results of the partial F-test) and calculate the average absenteeism rate for all employees with COMPLX = 10 and SENIOR = 3 who were very satisfied with their managers.

Please note that the variable used for seniority in the regression model is the reciprocal and NOT the seniority value. Keep this in mind while answering parts d, e, f, and g above.

In [20]:
```
cat('The average absenteeism rate is ',2.53471-0.01407*10+1.10284*(1/3) -0.90411)
```
executed in 24ms, finished 15:23:19 2019-12-12

```
The average absenteeism rate is  1.857513
```

14.8 h. (2 points) How could this study be used by the management to help identify employees who might be prone to absenteeism?

ANSWER

From this study, we found that employees who are doing less complex job with lower seniority and not satisfy with the manager might be prone to absenteeism.

15 (25 points) Question 15

The dataset on Question15 file presents data from the case of United States Department of the Treasury v. Harris Trust and Savings Bank (1981). The data includes the salary of 93 employees of the bank, their education level, and their gender.

In [14]:
```
setwd("D:/BAX441/Final Exam")
Q15<-read.csv('Question15.csv')
salary<-Q15[,1]
edu<-Q15[,2]
experience<-Q15[,3]
male<-ifelse(Q15[,5]=='MALE',1,0)
```

executed in 55ms, finished 17:01:42 2019-12-13

15.1 a. (3 points) Create a multiple regression model using education level and the male/female dummies.

In [22]:
```
summary(lm(salary~edu+male))
```

executed in 27ms, finished 15:29:30 2019-12-12

```
Call:
lm(formula = salary ~ edu + male)

Residuals:
     Min       1Q   Median       3Q      Max 
-1241.50  -364.29    16.41   378.50  1943.90 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4173.13     339.18  12.304  < 2e-16 ***
edu            80.70      27.67   2.916  0.00447 ** 
male          691.81     132.23   5.232 1.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 572.4 on 90 degrees of freedom
Multiple R-squared:  0.3634,    Adjusted R-squared:  0.3492 
F-statistic: 25.68 on 2 and 90 DF,  p-value: 1.498e-09
```

15.2 b. (5 points) Interpret the differential intercept coefficient and the parameter estimate of the education level. Is there evidence of employment discrimination?

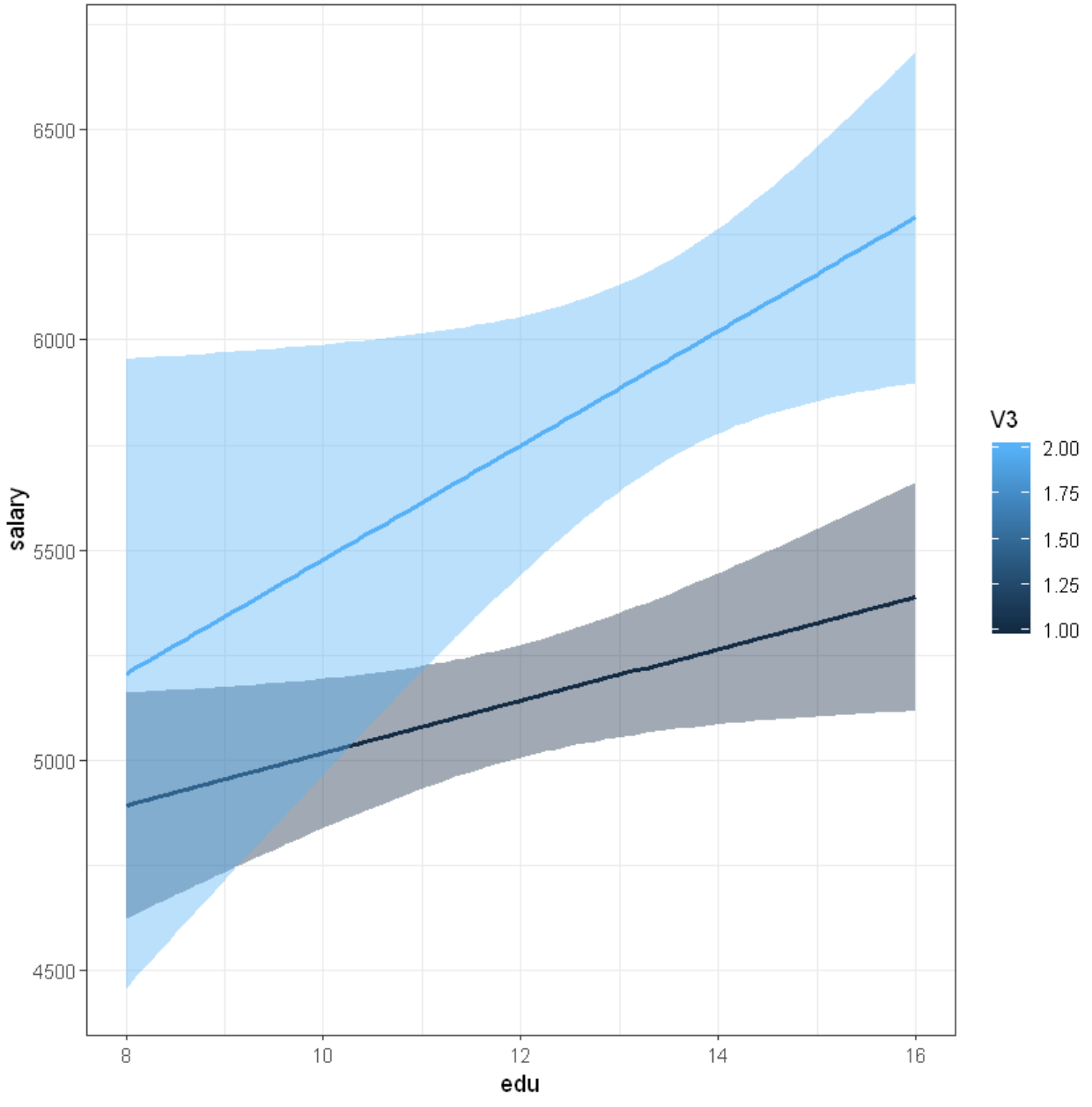
ANSWER

From the result, the parameter estimate of the education level is 80.7 and the p-value is significant , which means 1 unit higher in the education level the salary would increase 80.7 dollars on average. Holding the education level, male can earn 691.81 dollars on average more than female with the significant p-value of the differential intercept coefficient. Based on this model, it seems that there is evidence on gender discrimination. However, we need to test the effect of interaction.

15.3 c. (5 points) Create a plot with the two regressions – one for male and another for female. Are these two regressions parallel, coincident, dissimilar, or concurrent regressions? Refer to my Week 9 note under the topic of Comparing Two Regressions.

In [41]:
```
library(ggplot2)
df<-data.frame(cbind(salary,edu,Q15$GENDER))
ggplot(df, aes(edu, salary, group=V3, colour=V3, fill=V3)) +
  geom_smooth(method="lm") +
  theme_bw()
```

executed in 254ms, finished 16:06:37 2019-12-12



ANSWER

They are dissimilar.

15.4 d. (5 points) Now using the model from the above parts, determine if there is evidence of employment discrimination at Harris

Bank even after taking into account education (i.e. interaction). Hint: For this, you will have to create an interaction variable EDUC*MALES.

In [42]:
```
summary(lm(salary~edu+male+edu:male))
```

executed in 26ms, finished 16:08:40 2019-12-12

```
Call:
lm(formula = salary ~ edu + male + edu:male)

Residuals:
     Min       1Q   Median       3Q      Max 
-1240.89  -349.06    72.72   372.72  1808.08 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4395.32     389.21  11.293   <2e-16 ***
edu            62.13      31.94   1.945   0.0549 .  
male         -274.86     845.75  -0.325   0.7460    
edu:male       73.59      63.59   1.157   0.2503    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 571.4 on 89 degrees of freedom
Multiple R-squared:  0.3728,    Adjusted R-squared:  0.3516 
F-statistic: 17.63 on 3 and 89 DF,  p-value: 4.557e-09
```

ANSWER

The new model shows that the p-value of the interaction variable is not significant. From the plot in the previoud question, there is no interact between male regression and female regression. Thus, there is no effect of interaction. According to the model in question a, the p-value of male dummy variable is significant, so there is discremination.

15.5 e. (3 points) Is the interaction significant? Is the dummy variable significant?

ANSWER

The new model shows that the p-value of the interaction variable is not significant. The dummy variable in the new model also not significant.

15.6 f. (4 points) What is the difference between the model with interaction variable and the model with only EDUCATION and the dummy variable?

ANSWER

In the model with interaction variable, at 5% significant level, no explainor has significant p-value, but the p-value of the F-test is significant. While the model with interaction variable has an adjusted R^2 of 0.35, the model without interaction variable has an adjusted R^2 of 0.349 - they are very close.However, the model without the interaction variable's explainors' p-value are all significant. One reason is that there is no interaction between education level and gender, so the model with interaction variable may have multicollinearity.

16 (8 points) Question 16

The growth model that we covered in the note on functional forms was fitted to several U.S. economic time series and the following results were obtained:

Time series and period	B_1	B_2	r^2
Real GNP (1954–1987) (1982 dollars)	7.2492 $t = (529.29)$	0.0302 (44.318)	0.9839
Labor force participation rate (1973–1987)	4.1056 $t = (1290.8)$	0.053 (15.149)	0.9464
S&P 500 index (1954–1987)	3.6960 $t = (57.408)$	0.0456 (14.219)	0.8633
S&P 500 index (1954–1987 quarterly data)	3.7115 $t = (114.615)$	0.0114 (27.819)	0.8524

The B1 and B2 are the intercept β_1 and slope β_2.

16.1 a. (4 points) In each case find out the instantaneous rate of growth.

In [80]:
```
cat('The instantaneous rate of growth of real GNP is ',0.0302*100, '%.\n')
cat('The instantaneous rate of growth of labor force participation rate is ',0.053*100, '%.\n')
cat('The instantaneous rate of growth of S&P 500 index is ',0.0456*100, '%.\n')
cat('The instantaneous rate of growth of S&P 500 index (quarterly) is ',0.0114*100, '%.\n')
```

executed in 32ms, finished 18:59:47 2019-12-11

```
The instantaneous rate of growth of real GNP is  3.02 %.
The instantaneous rate of growth of labor force participation rate is  5.3 %.
The instantaneous rate of growth of S&P 500 index is  4.56 %.
The instantaneous rate of growth of S&P 500 index (quarterly) is  1.14 %.
```

16.2 b. (4 points) What is the compound rate of growth in each case?

In [82]:
```
cat('The compound rate of growth of real GNP is ',(exp(0.0302)-1)*100, '%.\n')
cat('The compound rate of growth of labor force participation rate is ',(exp(0.053)-1)*100, '%.\n
cat('The compound rate of growth of S&P 500 index is ',(exp(0.0456)-1)*100, '%.\n')
cat('The compound rate of growth of S&P 500 index (quarterly) is ',(exp(0.0114)-1)*100, '%.\n')
```

executed in 30ms, finished 19:01:52 2019-12-11

```
The compound rate of growth of real GNP is  3.066065 %.
The compound rate of growth of labor force participation rate is  5.442965 %.
The compound rate of growth of S&P 500 index is  4.665566 %.
The compound rate of growth of S&P 500 index (quarterly) is  1.146523 %.
```

17 (5 points) Question 17

XYZ Realty sells homes along the East Coast of the United States. One of the questions most frequently asked by prospective buyers is: If we purchase this home, how much can we expect to pay to heat it during the winter? The research department at XYZ has been asked to develop some guidelines regarding heating costs for single-family homes.

Four variables are thought to relate to the heating costs:

- the mean daily outside temperature,
- the number of inches of insulation in the attic,
- the age in years of the furnace, and
- whether the home has a garage or not

To investigate, XYZ's research department selected a random sample of 20 recently sold homes. It determined the cost to heat each home last January, as well as the January outside temperature in the region, the number of inches of insulation in the attic, the age of the furnace, and whether the home has a garage or not. The sample information is reported in CSV file Question17. Build a multiple regression model using the stepwise variable selection technique.

In [5]:
```
setwd("D:/BAX441/Final Exam")
Q17<-read.csv('Question17.csv')
cost<-Q17[,1]
temp<-Q17[,2]
insul<-Q17[,3]
age<-Q17[,4]
garage<-ifelse(Q17[,5]=='Yes',1,0)
```

executed in 29ms, finished 16:39:59 2019-12-12

In [6]:
```
### First run a regression model with all variables for a quick check
### for multicollinearity
model_sol_ex_17 <- lm(cost ~ temp + insul + age + garage )

car::vif(model_sol_ex_17) ### No evidence of multicollinearity
```

executed in 37ms, finished 16:40:00 2019-12-12

temp	1.59494797881867
insul	1.08313126161829
age	1.37866907691764
garage	1.51435731055554

In [10]:
```
### Store all vars + Dummy in a separate dataframe.
### Could have included in the original dataframe...
df_SolvedEx17 <- data.frame(cost, temp, insul, age, garage)

### Run Stepwise Regression
model <- lm(cost ~ ., data = df_SolvedEx17)
k <- olsrr::ols_step_both_p(model, prem = 0.10, pent = 0.10, details = FALSE)
#plot(k)
```

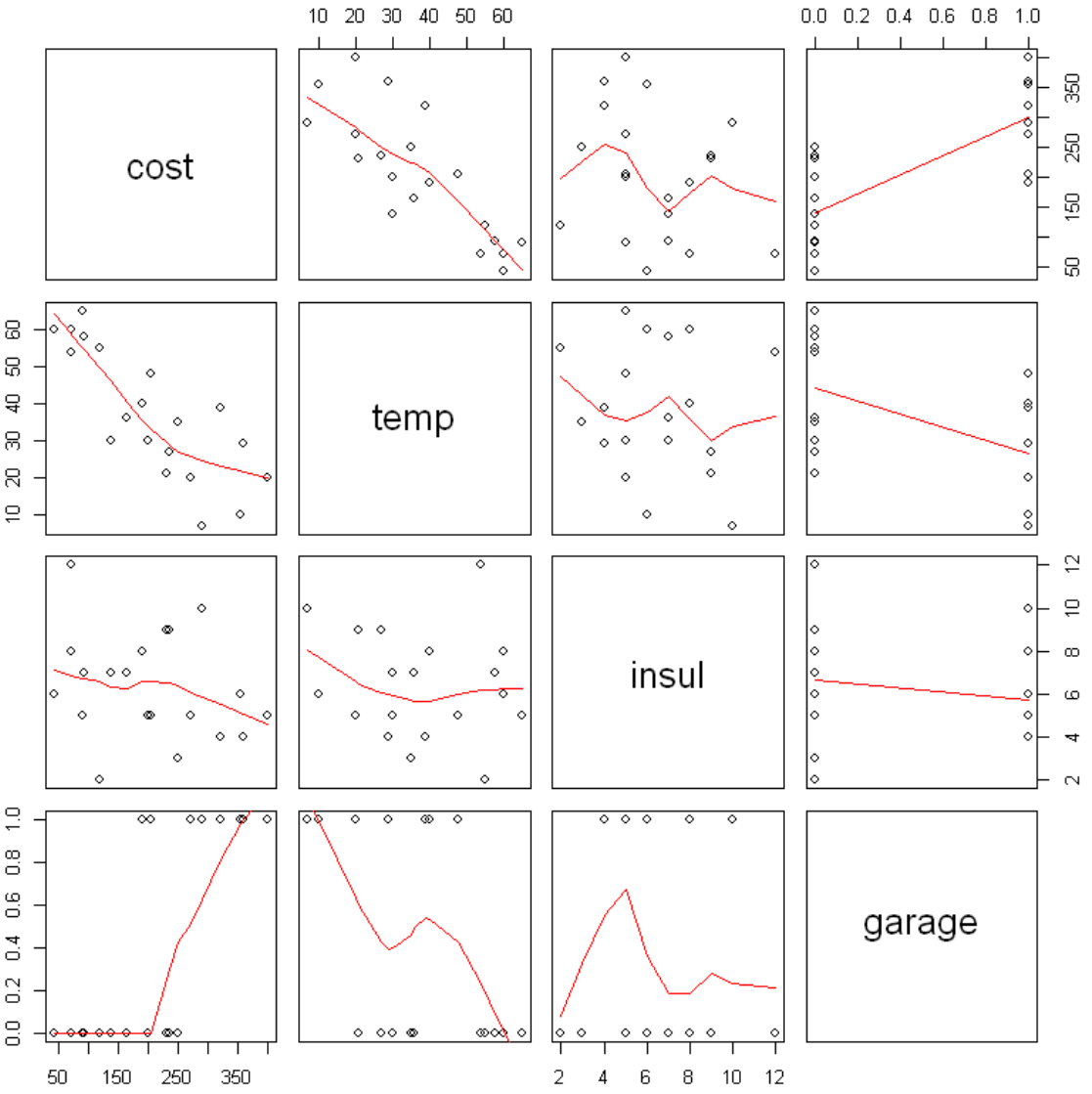
executed in 121ms, finished 16:46:03 2019-12-12

```
Stepwise Selection Method
---------------------------

Candidate Terms:

1. temp
2. insul
3. age
4. garage

We are selecting variables based on p value...

Variables Entered/Removed:

- temp added
- garage added
- insul added

No more variables to be added/removed.


Final Model Output
------------------

                         Model Summary
---------------------------------------------------------------
R                       0.933       RMSE                41.618
R-Squared               0.870       Coef. Var           20.277
Adj. R-Squared          0.845       MSE               1732.093
Pred R-Squared          0.800       MAE                 32.193
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------------
Regression    185202.269         3      61734.090    35.641    0.0000
Residual       27713.481        16       1732.093
Total         212915.750        19
--------------------------------------------------------------------------

                                  Parameter Estimates
-----------------------------------------------------------------------------------------------
      model       Beta    Std. Error    Std. Beta      t        Sig       lower      upper
-----------------------------------------------------------------------------------------------
(Intercept)    393.666        45.001                 8.748    0.000    298.267    489.064
       temp     -3.963         0.653       -0.652    -6.072    0.000     -5.346     -2.579
    garage     77.432        22.783        0.368     3.399    0.004     29.135    125.730
     insul    -11.334         4.002       -0.265    -2.832    0.012    -19.817     -2.851
-----------------------------------------------------------------------------------------------
```

In [9]:
```r
### Consider selected variables only and create a separate data frame
df_sel_var_17 <- data.frame(cost, temp, insul,garage)

### Quickly create a scatterplot to see if there is any evidence of nonlinearity.
pairs(df_sel_var_17, panel = panel.smooth)

model_sol_ex_17a <- lm(cost ~ temp + insul+garage)
summary(model_sol_ex_17a)
```

executed in 100ms, finished 16:44:32 2019-12-12

```
Call:
lm(formula = cost ~ temp + insul + garage)

Residuals:
    Min      1Q  Median      3Q     Max
-63.171 -32.194   0.056  29.111  64.829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 393.6657    45.0013   8.748 1.71e-07 ***
temp         -3.9628     0.6527  -6.072 1.62e-05 ***
insul       -11.3340     4.0015  -2.832  0.01201 *
garage       77.4321    22.7828   3.399  0.00367 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.62 on 16 degrees of freedom
Multiple R-squared:  0.8698,    Adjusted R-squared:  0.8454
F-statistic: 35.64 on 3 and 16 DF,  p-value: 2.586e-07
```


cost
temp
insul
garage
10 20 30 40 50 60
0.0 0.2 0.4 0.6 0.8 1.0
50 150 250 350
10 20 30 40 50 60
2 4 6 8 10 12
0.0 0.2 0.4 0.6 0.8 1.0
50 150 250 350
2 4 6 8 10 12

In [15]:
```
gvlma::gvlma(model_sol_ex_17a)
```

executed in 26ms, finished 17:00:41 2019-12-12

```
Call:
lm(formula = cost ~ temp + insul + garage)

Coefficients:
(Intercept)         temp        insul       garage
    393.666       -3.963      -11.334       77.432


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma::gvlma(x = model_sol_ex_17a)

                     Value p-value                Decision
Global Stat        2.229191  0.6937 Assumptions acceptable.
Skewness           0.002617  0.9592 Assumptions acceptable.
Kurtosis           1.031894  0.3097 Assumptions acceptable.
Link Function      0.128627  0.7199 Assumptions acceptable.
Heteroscedasticity 1.066053  0.3018 Assumptions acceptable.
```

18 (5 points) Question 18

In evaluating the performance of new hires, the HR division found that candidates with higher scores on its qualifying exam performed better. In a multiple regression that also used the education of the new hire as a regressor variable, the slope for the test score was zero. How can you explain this paradox to the HR manager?

ANSWER

Maybe there is multicollinearity between education level and the test score because people who obtain higher education degree may be smarter than who do not. Thus, in the multiple regression, the test score is explained by education level and cannot be used to predict the performance.

19 (5 points) Question 19

Managers in the HR department suspect that sick-day absentee rates are higher on some weekdays than others. What test cam they use to investigate this claim?

ANSWER

1-way ANOVA

20 (5 points) Question 20

A manager in the previous question thinks that the absentee rate is the same on Monday and Friday, but different from the rate on Tuesday and Thursday. What method should she use to test her suspicion?

ANSWER

Multiple regression with dummy variable.