Homework 4

Nikhil Guruji

December 3, 2017

1-

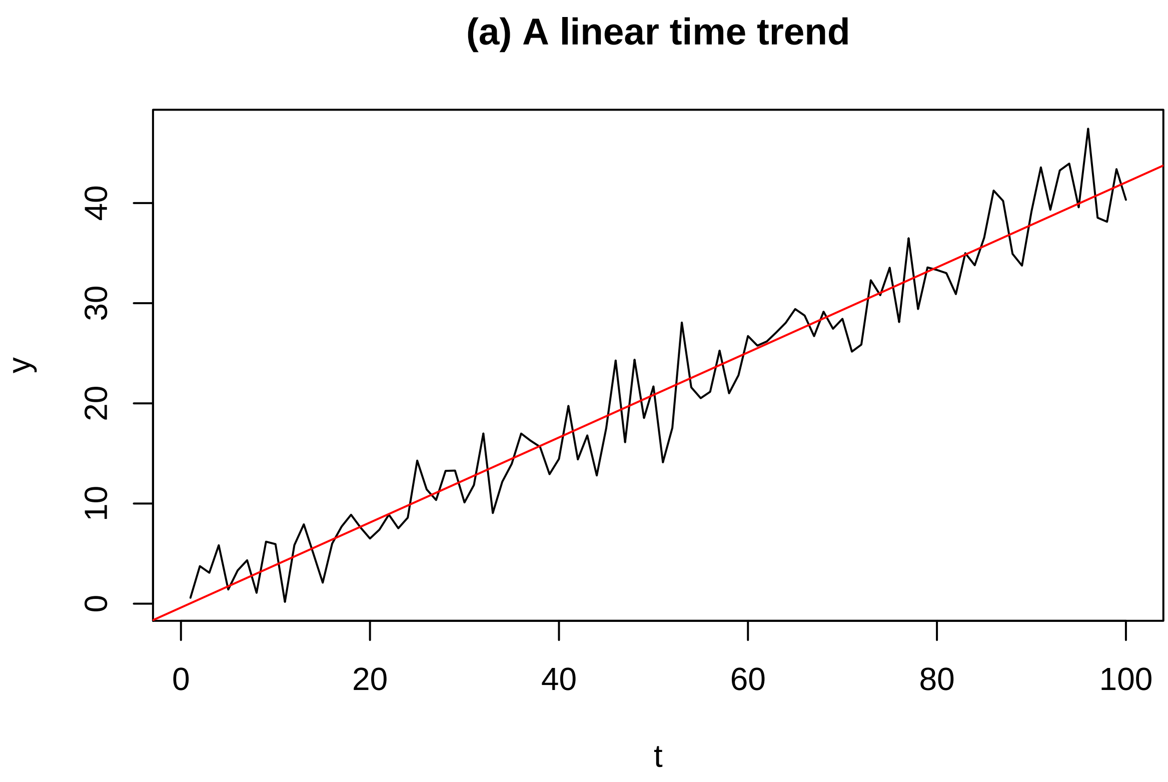
Assumptions:

σ=3

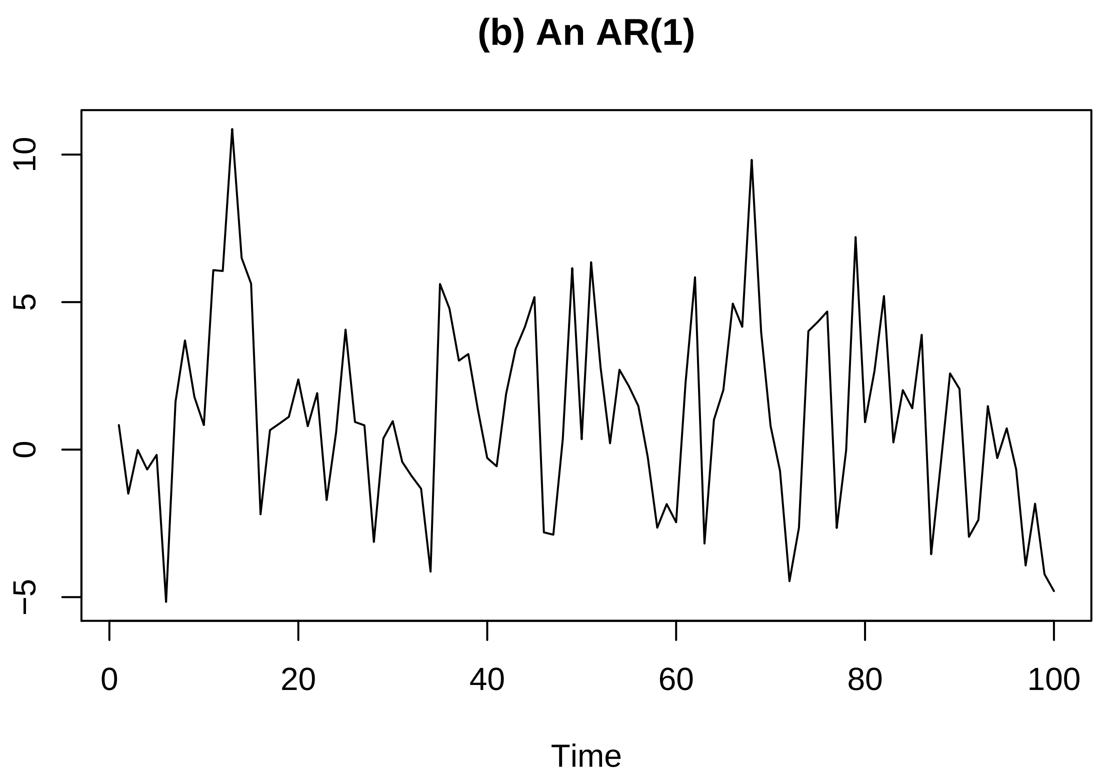
β=0.4

α=0.5

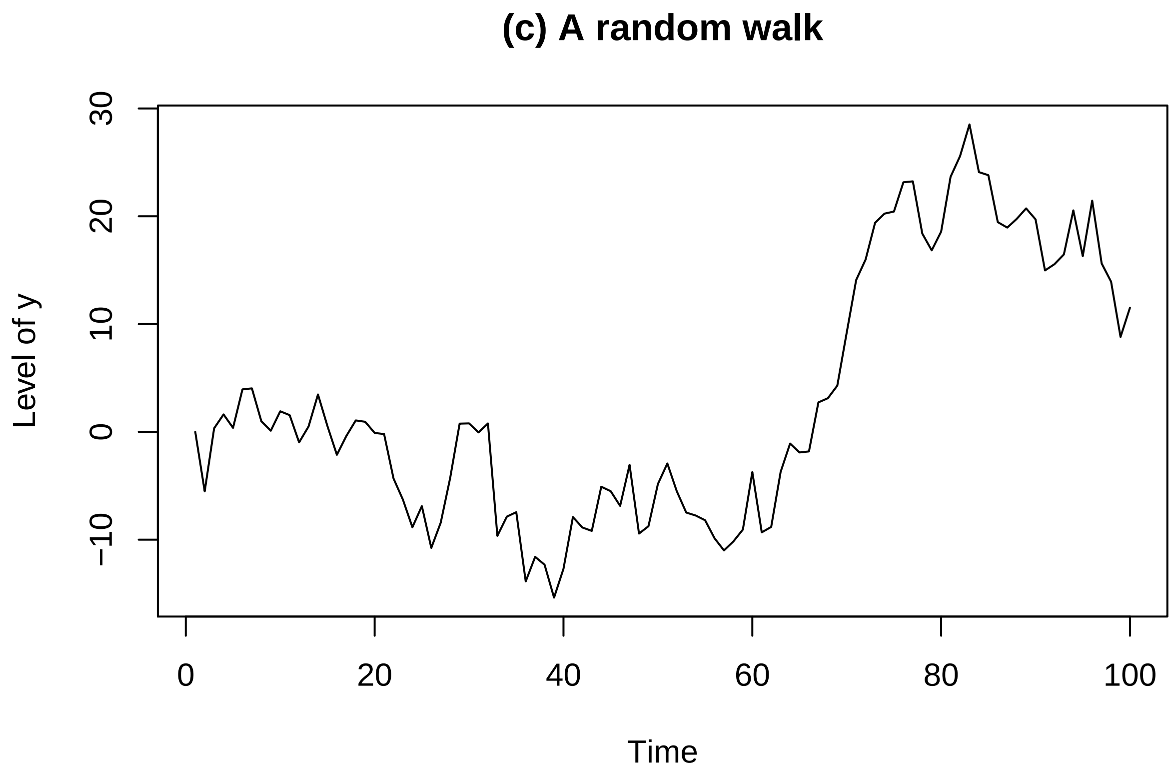
```
# part a: for linear time trend
library(DataAnalytics)
alpha=0.5
beta=0.4
std_dev=3
epsilon=rnorm(100,0,std_dev)
t=seq(1:100)
y= alpha + beta*t + epsilon
plot(t,y,type="l",main="(a) A linear time trend")
abline(lm(y~t)$coef,col="red")
```



```r
#part b: for AR(1)

yar1=double(length(t))
yar1[1]=alpha/(1-beta)
for(i in 2:length(t)){
  yar1[i]= alpha + beta*yar1[i-1] + rnorm(1,sd=std_dev)
}
plot(yar1,type="l",xlab="Time",ylab="",main="(b) An AR(1)")
```

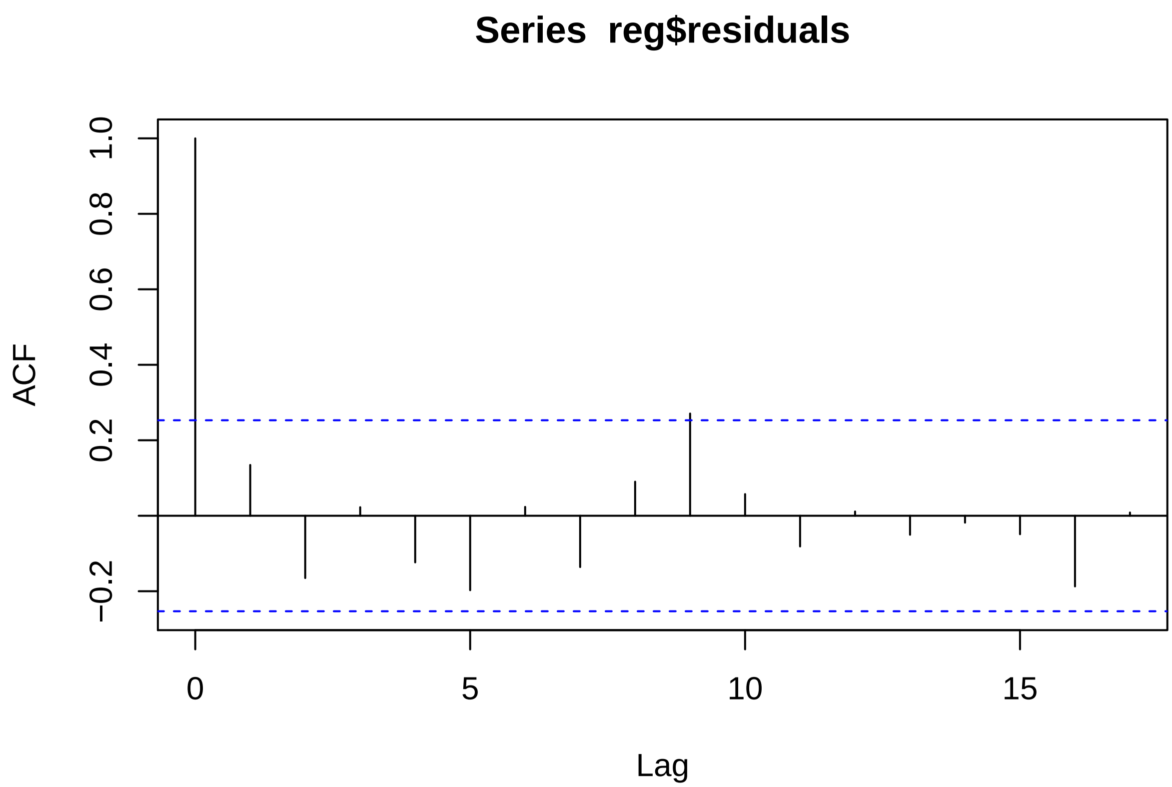



```r
#part c: for random walk

yrw=double(length(t))
for(i in 2:length(yrw)){
  yrw[i]= yrw[i-1] + rnorm(1,sd=std_dev)
}
plot(y=yrw,type="l",x=c(1:100),xlab="Time",ylab="Level of y",main="(c) A random walk")
```



2-

```
data("beerprod")
#part a
reg=lm(b_prod~back(b_prod,noperiods = 1)+back(b_prod,noperiods = 6)
       +back(b_prod,noperiods = 12),data=beerprod)

#part b
acf(reg$residuals)
```



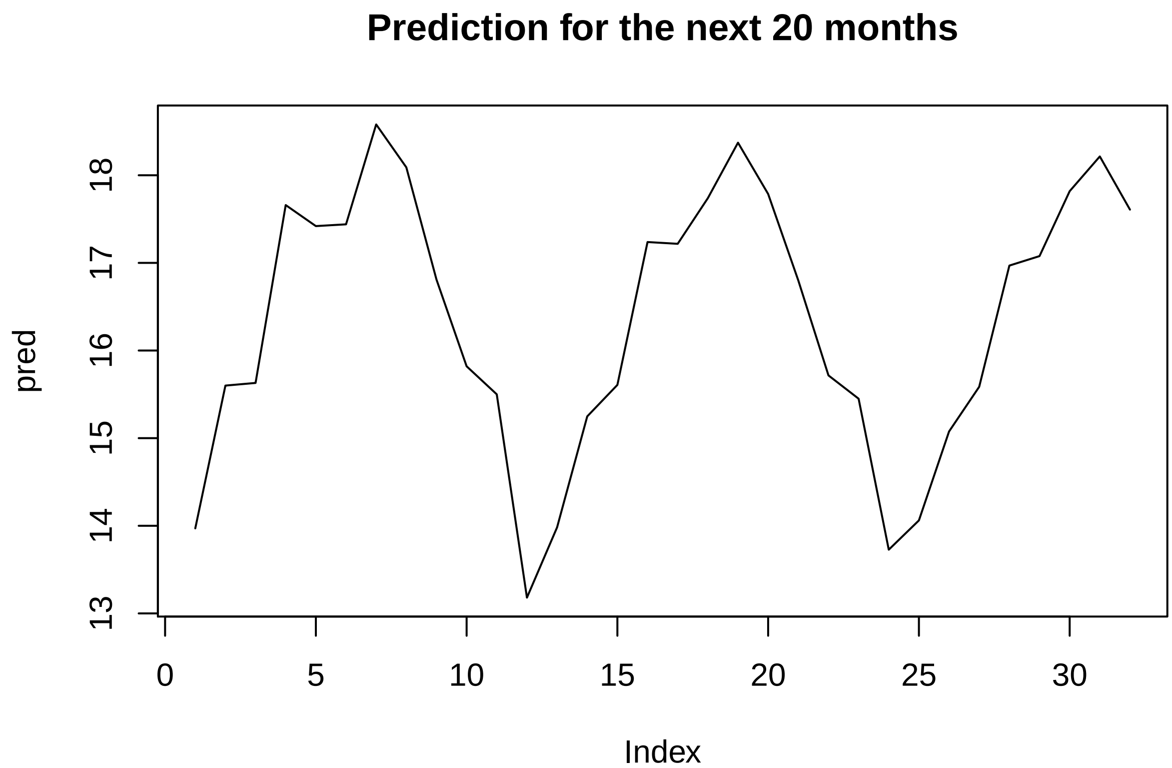

```
Box.test(reg$residuals,type="Ljung")
```

```
##
##  Box-Ljung test
##
## data:  reg$residuals
## X-squared = 1.1417, df = 1, p-value = 0.2853
```

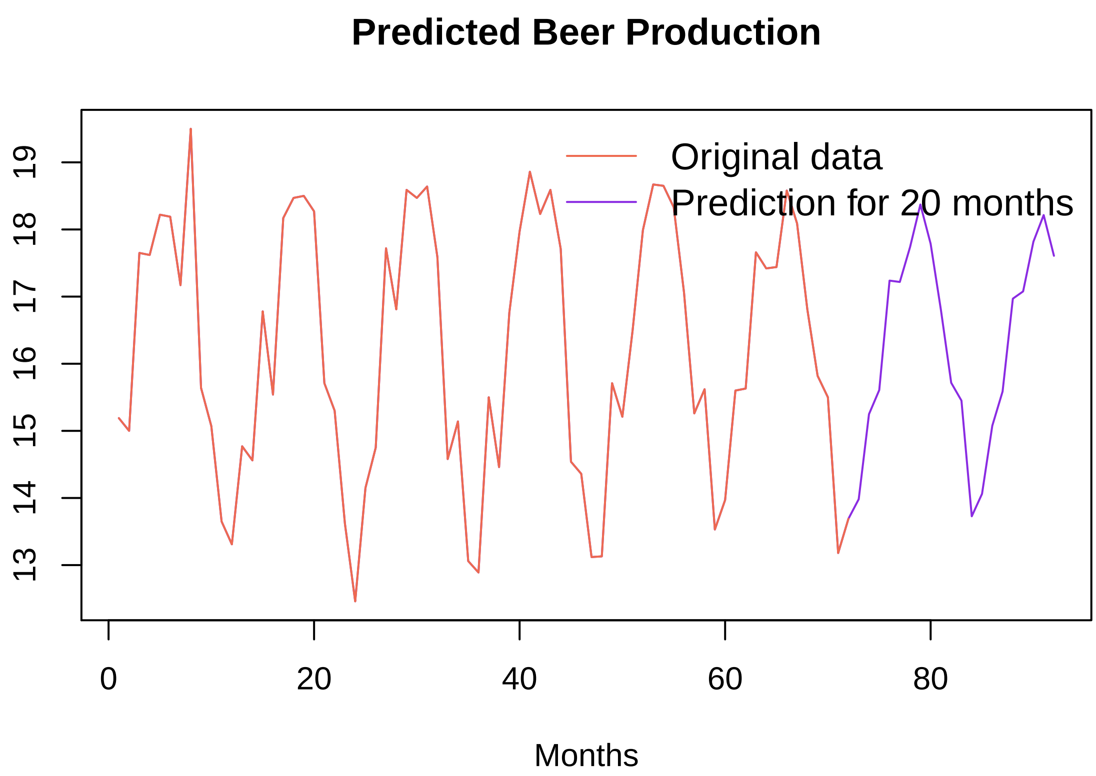
(b) As seen from the auto correlation plot, there is no significant autocorrelation left for most of the periods. This happens because we have considered lags upto 12 periods.

Also, the p-value from the Box-Ljung test is large which indicates that the null hypothesis is true which says that the auto correlation is zero.

```
#part c
n=20
pred=double(n+12)
pred[1:12]=beerprod$b_prod[(nrow(beerprod)-12):nrow(beerprod)]

for(i in 13:(n+12)){
  pred[i]=reg$coef[1]+reg$coef[2]*pred[i-1]+reg$coef[3]*pred[i-6]+reg$coef[4]*pred[i-12]
  }

plot(pred,type="l",main="Prediction for the next 20 months")
```




```
bpred=c(beerprod$b_prod,pred[13:32])
plot(bpred,type="l",col="blueviolet",main="Predicted Beer Production",ylab="",xlab="Months")
lines(beerprod$b_prod,col="coral2")
legend("topright", legend = c("Original data", "Prediction for 20 months"), bty = "n",
       lwd = 1, cex = 1.2, col = c("coral2", "blueviolet"), lty = c(1, 1))
```



3-

(a)

coefficient from regression of Y on X is given by:

$\beta = \frac{Cov(X,Y)}{\sigma_X^2}$

Correlation is given by:

$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Therefore,

$\beta = \rho\frac{\sigma_Y}{\sigma_X}$

In this case, $\sigma_X = \sigma_Y$ because we are regressing the same dependent variable on it's lagged values and assuming stationarity.

Hence, $\beta = \rho$

Verification:

```
a=beerprod$b_prod
b=back(a)
per1=lm(a~b)

beta=per1$coef[2]
```

```
acomplete=a[complete.cases(a,b)]
bcomplete=b[complete.cases(a,b)]

rho=cor(acomplete,bcomplete)

#coefficient on the lagged dependent variable
beta
```

```
##         b
## 0.7042857
```

```
#correlation between the dependent variable and its lag
rho
```

```
## [1] 0.6969658
```

(b)

We know that $s^2_{b_1} = \frac{\sigma^2}{(N-1)s^2_x}$

where b_1 is the coefficient of the independent variable, s_x is the standard deviation of the independent variable, σ is the standard deviation of the dependent variable and $N-1$ are the degrees of freedom (or the number of observations - 1).

In this case, the independent variable is the dependent variable lagged one period. So, $\sigma = s_x$ because variance remains constant by stationarity assumption. And the number of observations $= T-1$ (because the t=0 observation will not be counted as there is no observation for t=-1).

Therefore,

$s^2_{b_1} = \frac{1}{(T-1)-1}$

$s^2_{b_1} = \frac{1}{T-2}$

$s_{b_1} = \sqrt{\frac{1}{T-2}}$

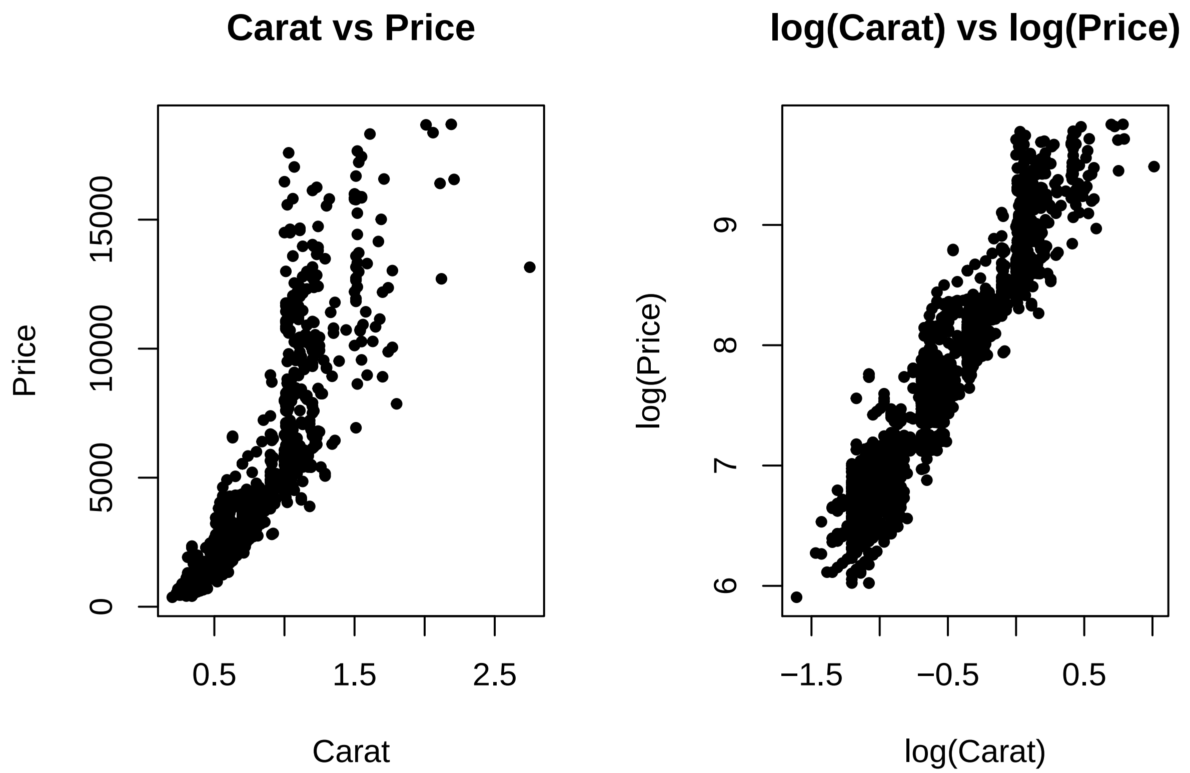
For large number of observations, $T \approx T-2$

Hence,

$s_{b_1} = \frac{1}{\sqrt{T}}$

4-

```
library(ggplot2)
data("diamonds")
subdata=diamonds[diamonds$cut=="Ideal" & diamonds$color=="D",]

#part a

par(mfrow=c(1,2))
plot(subdata$carat,subdata$price,xlab="Carat",ylab="Price",main="Carat vs Price",pch=20)
plot(log(subdata$carat),log(subdata$price),xlab="log(Carat)"
     ,ylab="log(Price)",main="log(Carat) vs log(Price)",pch=20)
```



(b)

For IF1:

$log(price_{IF}) = \beta_0 + \beta_1 log(carat) + \beta_{IF}1$

and

$log(price_{SI2}) = \beta_0 + \beta_1 log(carat) + \beta_{SI2}1$

Subtracting the equations, we get:

$log\left[\frac{price_{IF}}{price_{SI2}}\right] = \beta_{IF} - \beta_{SI2}$

Therefore,

$\frac{price_{IF}}{price_{SI2}} = e^{\beta_{IF} - \beta_{SI2}}$

percentage premium of "IF" relative to "SI2" is:

$\frac{price_{IF}}{price_{SI2}} - 1 = e^{(\beta_{IF} - \beta_{SI2})} - 1$

```
#part b
cl=factor(subdata$clarity,ordered = FALSE)
reg=lm(log(price)~log(carat)+cl,data=subdata)
summary(reg)
```

```
##
## Call:
## lm(formula = log(price) ~ log(carat) + cl, data = subdata)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55673 -0.09503 -0.00961  0.09408  0.43722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.092421   0.038888 208.095  < 2e-16 ***
## log(carat)  1.892054   0.006054 312.517  < 2e-16 ***
## clSI2       0.321977   0.039568   8.137 5.99e-16 ***
## clSI1       0.506748   0.039259  12.908  < 2e-16 ***
## clVS2       0.703740   0.039253  17.928  < 2e-16 ***
## clVS1       0.739629   0.039662  18.648  < 2e-16 ***
## clVVS2      0.923933   0.039833  23.195  < 2e-16 ***
## clVVS1      1.063408   0.040725  26.112  < 2e-16 ***
## clIF        1.440083   0.047056  30.603  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1401 on 2825 degrees of freedom
## Multiple R-squared:  0.9728, Adjusted R-squared:  0.9728
## F-statistic: 1.265e+04 on 8 and 2825 DF,  p-value: < 2.2e-16
```

```r
levels(subdata$clarity)
```

```
## [1] "I1"   "SI2"  "SI1"  "VS2"  "VS1"  "VVS2" "VVS1" "IF"
```

```r
ratio_price=exp(reg$coefficients[9]-reg$coefficients[3])
names(ratio_price)=""
percent_rel_prem=ratio_price-1
ratio_price
```

```
##
## 3.059056
```

```r
percent_rel_prem
```

```
##
## 2.059056
```

Thus, diamonds with clarity “IF” have a price premium 3.06 or 205.9% times more than those with clarity “SI2”

(c)

Regression line for log(price) for “IF” is:

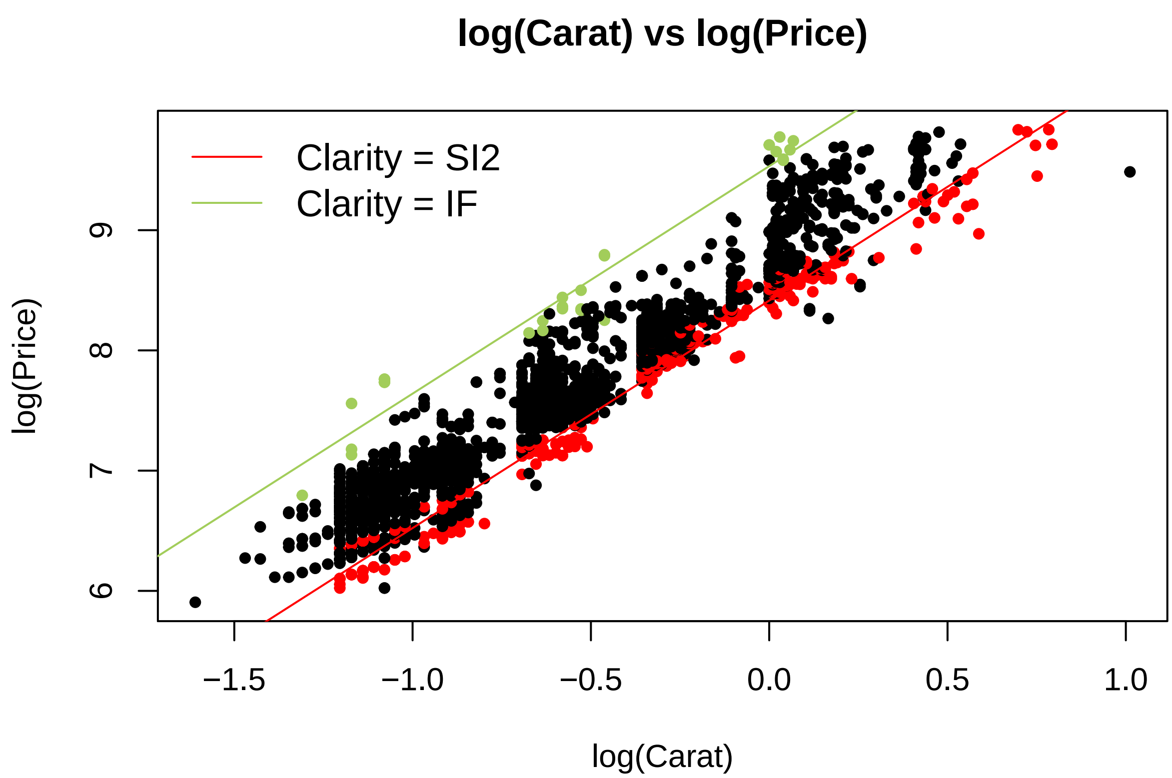
$log(price) = (\beta_0 + \beta_{IF} \times 1) + \beta_1 log(carat)$

Similarly, regression line for log(price) for “SI2” is:

$log(price) = (\beta_0 + \beta_{SI2} \times 1) + \beta_1 log(carat)$

```r
#part c
subdata$colourofpoints="black" #default colour
subdata$colourofpoints[subdata$clarity=="IF"]="darkolivegreen3"
subdata$colourofpoints[subdata$clarity=="SI2"]="red"
plot(log(subdata$carat),log(subdata$price),xlab="log(Carat)",
     ylab="log(Price)",main="log(Carat) vs log(Price)", col=subdata$colourofpoints,pch=20)
abline(a=(reg$coefficients[1]+reg$coefficients[9]),b=reg$coefficients[2],col="darkolivegreen3")
abline(a=(reg$coefficients[1]+reg$coefficients[3]),b=reg$coefficients[2],col="red")
```

```r
legend("topleft", legend = c("Clarity = SI2", "Clarity = IF"), bty = "n",
       lwd = 1, cex = 1.2, col = c("red", "darkolivegreen3"), lty = c(1, 1))
```



5-

```r
data("mtcars")
corr_act=cor(mtcars$mpg,mtcars$disp)

# construction bootstrap confidence interval for slope coefficient
Reg_data=na.omit(mtcars[,c(1,3)])
N=nrow(Reg_data)

B=10000
BS_coefs=rep(0,B)
for(b in 1:B){
  BS_sample=Reg_data[sample(1:N,size=N,replace=TRUE),]
  BS_coefs[b]=cor(BS_sample$mpg,BS_sample$disp)
}

int=quantile(BS_coefs,probs=c(.975,.025))
CI.pivotal.bootstrap = c(2*corr_act-int[1],
                         2*corr_act-int[2])
CI.pivotal.bootstrap
```

```
##      97.5%       2.5%
## -0.9342998 -0.7842833
```

```r
corr_act
```

```
## [1] -0.8475514
```

```r
mean(BS_coefs)
```

```
## [1] -0.8480177
```

```r
hist(BS_coefs,col="darkslategray1")
abline(v=mean(BS_coefs),lwd=1,col="darkorchid1")
abline(v=corr_act,lwd=1,col="forestgreen")
abline(v=CI.pivotal.bootstrap,lwd=1,col="red")
legend("right",legend=c("Mean Correlation from bootstrap samples","Actual Correlation"
  ,"95% Confidence intervals"),col=c("darkorchid1","forestgreen","red"),lty=c(1,1,1),lwd=1,bty='n')
```