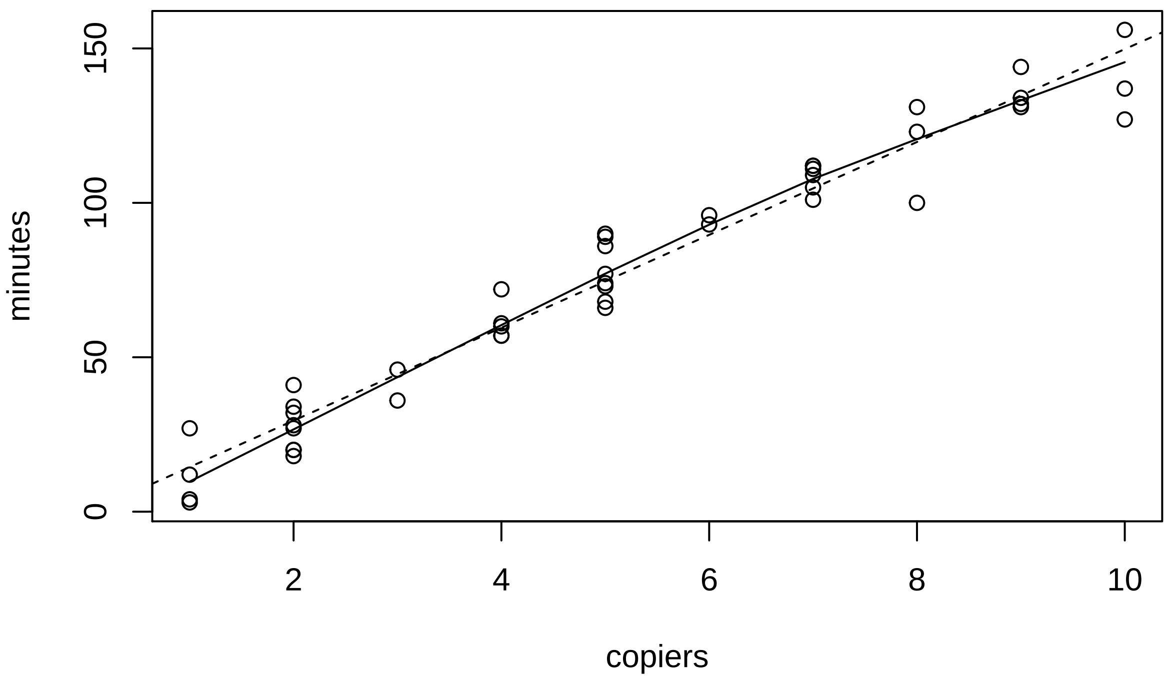
STAT GR5205 – Section 005 HW1 Bo Rong br2498

```r
#1.
filename <- "~/Downloads/copier_maintenance.txt"
Data <- read.table(file=filename, header=T)
#(a)
plot(minutes ~ copiers, data=Data)
low.fit <- lowess(x=Data$copiers, y=Data$minutes)
lines(low.fit)

#The simple linear regression model seems appropriate,since average minutes
#of service time appears to be a linear function of number of copiers serviced.


#(b)
ls.fit <- lm(minutes ~ copiers, data=Data)
abline(ls.fit, lty=2)
```




```r
summary(ls.fit)
```

```
##
## Call:
## lm(formula = minutes ~ copiers, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## copiers      15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

```
#Hence the regression line is Y^=-0.5802+15.0352x,the estimated regression function
#fit good to the data.


#(c)The service time increases b1 = 15.0352 minutes by each copier services.

#(d)b0 = -0.5802 maybe the starting time,which is the time before works.

#(e)estimate of the mean service time for calls with x = 5 copiers by
#Y^=0.5802+15.0352(5)=74.596 minutes

#(f)predict of the service time for a single call with  x = 5 copiers by
#Y^=0.5802+15.0352(5)=74.596 minutes


#(g)
e <- residuals(ls.fit)
e
```

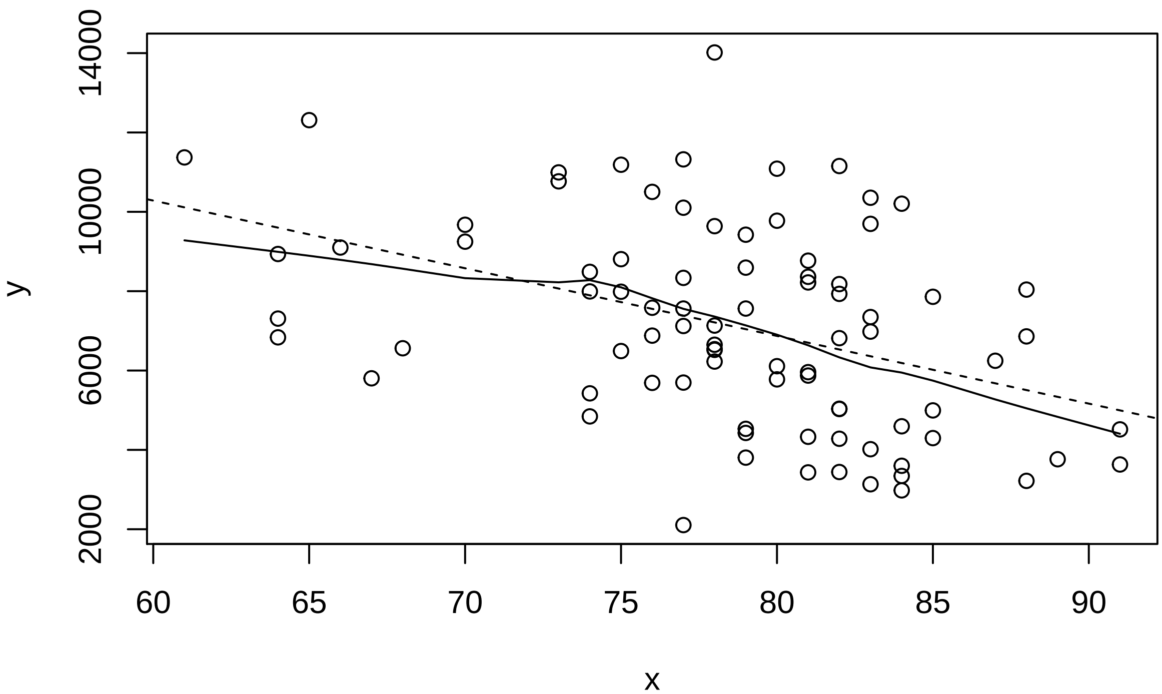
```
##           1           2           3           4           5           6
##  -9.4903394   0.4391645   1.4744125  11.5096606  -2.4550914 -12.7723238
##           7           8           9          10          11          12
##  -6.5960836  14.4039164 -10.4550914   2.5096606   9.2629243   6.2276762
##          13          14          15          16          17          18
##   3.3686684  -8.5255875  12.4391645 -19.7018277   0.3334204  11.2981723
##          19          20          21          22          23          24
## -22.7723238  -2.5608355  -8.5960836  -3.6665796   4.3334204  -0.5960836
##          25          26          27          28          29          30
##  -0.7370757   7.3334204 -11.4903394  -1.5960836   6.3334204   6.3686684
##          31          32          33          34          35          36
##   3.2981723  15.4039164  -9.4903394  -1.4903394 -11.4550914  -2.5608355
##          37          38          39          40          41          42
##  11.4039164  -2.7370757   7.3334204  12.5449086  -3.7370757   4.5096606
##          43          44          45
##  -2.4903394   1.4391645   2.4039164
```

```
sum(e)
```

```
## [1] -1.176836e-14
```

```
#sum(e)=-1.176836e-14 which is zero.The sum of squared residuals is the
#min value of quantity Q.
```

```r
#(h)Estimate variance(error) by the Residual standard error= 8.914 minutes.



#2.
#(a)
filename <- "~/Downloads/crime_rates.txt"
Data <- read.table(file=filename, header=T)
plot(y ~ x, data=Data)
low.fit <- lowess(x=Data$x, y=Data$y)
lines(low.fit)

#yes,the simple linear regression model seems appropriate for these data.
#Because the annual crime rate appears to be a linear function of the percentage of
#individuals which has at least a high-school diploma.

#(b)
ls.fit <- lm(y ~ x, data=Data)
abline(ls.fit, lty=2)
```




```r
summary(ls.fit)
```

```
##
## Call:
## lm(formula = y ~ x, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## x            -170.58      41.57  -4.103 9.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

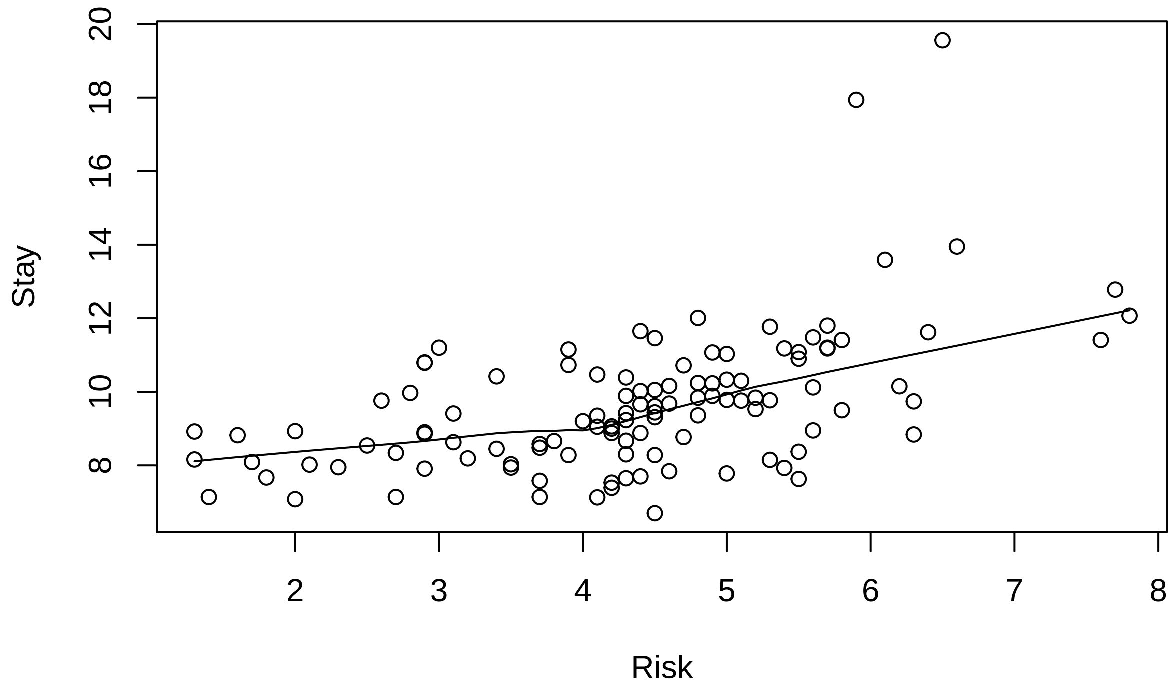
```r
#regression line is Y^=20517.60-170.58x

#(c)
#(i) High-school graduation rates lower by 5 percentage points  will have
#a higher crime rate by 5(170.58) = 853 crimes per 100,000.

#(ii)Estimation of the crime rate for a county with an 80% high-school
#graduation rate will be 20517.60 - 170.58(80)= 6872 crimes per 100,000.

#(iii)
Data[6,]
```

```
##      y  x
## 6 9100 66
```

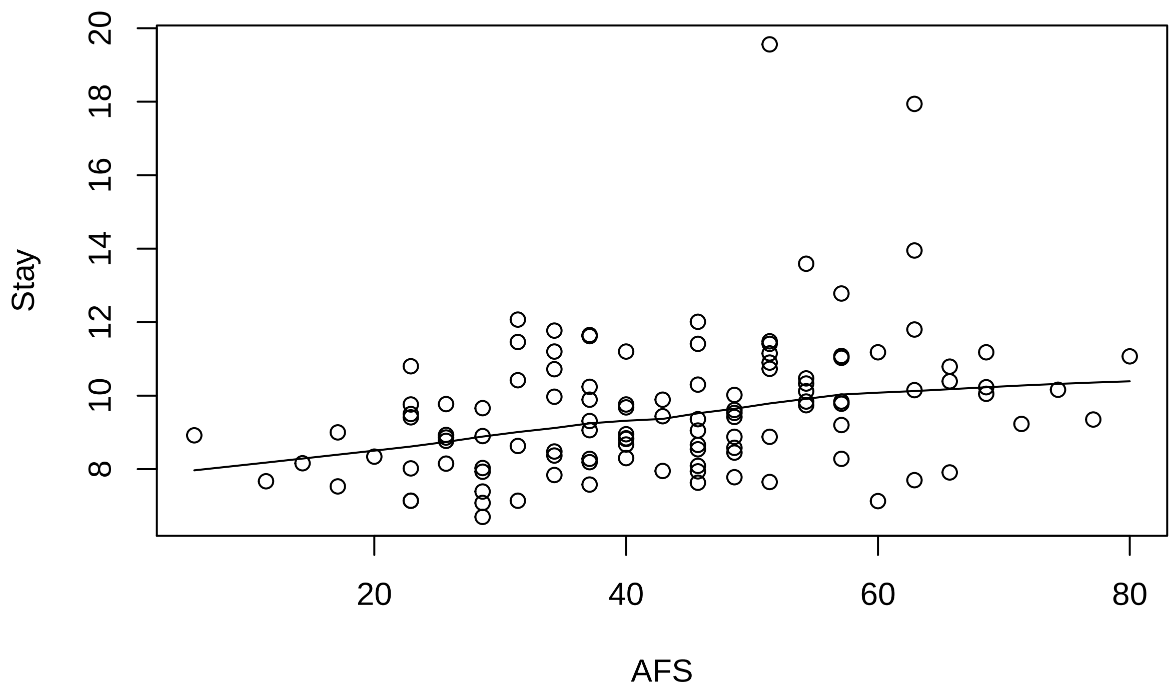
```r
#So the random error term for the sixth county is e6 =9100 - [20517.60 - 170.58(66)]
#= -159.64.

#(iv)
anova(ls.fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## x          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

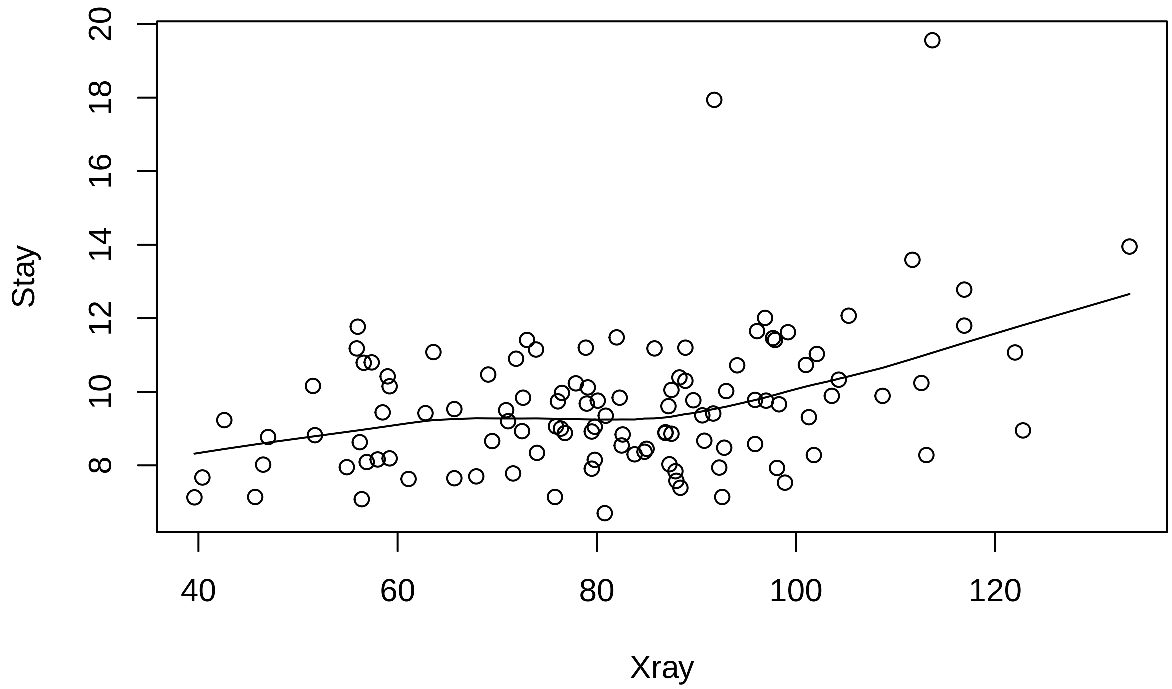
```r
# the error term = MSE = 5,552,112.



#3.
#(a)
filename <- "~/Downloads/SENIC.txt"
Data <- read.table(file=filename, header=T)
attach(Data)
plot(Risk, Stay)
lines(lowess(Risk, Stay))
```



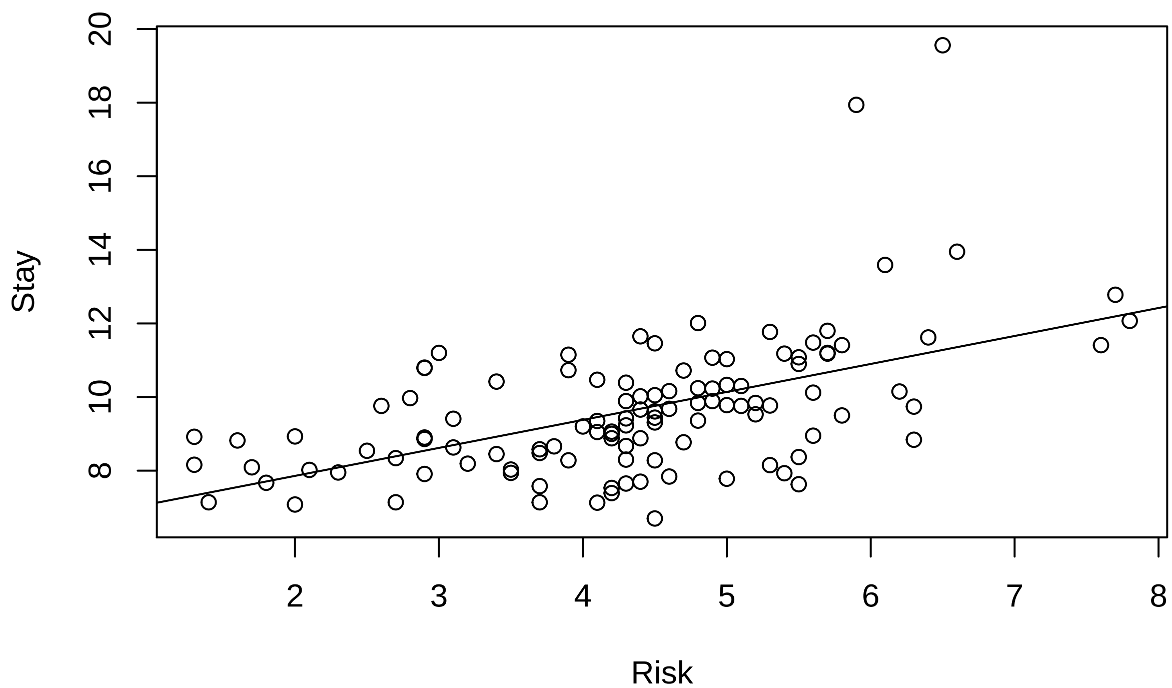

```
plot(AFS, Stay)
lines(lowess(AFS, Stay))
```



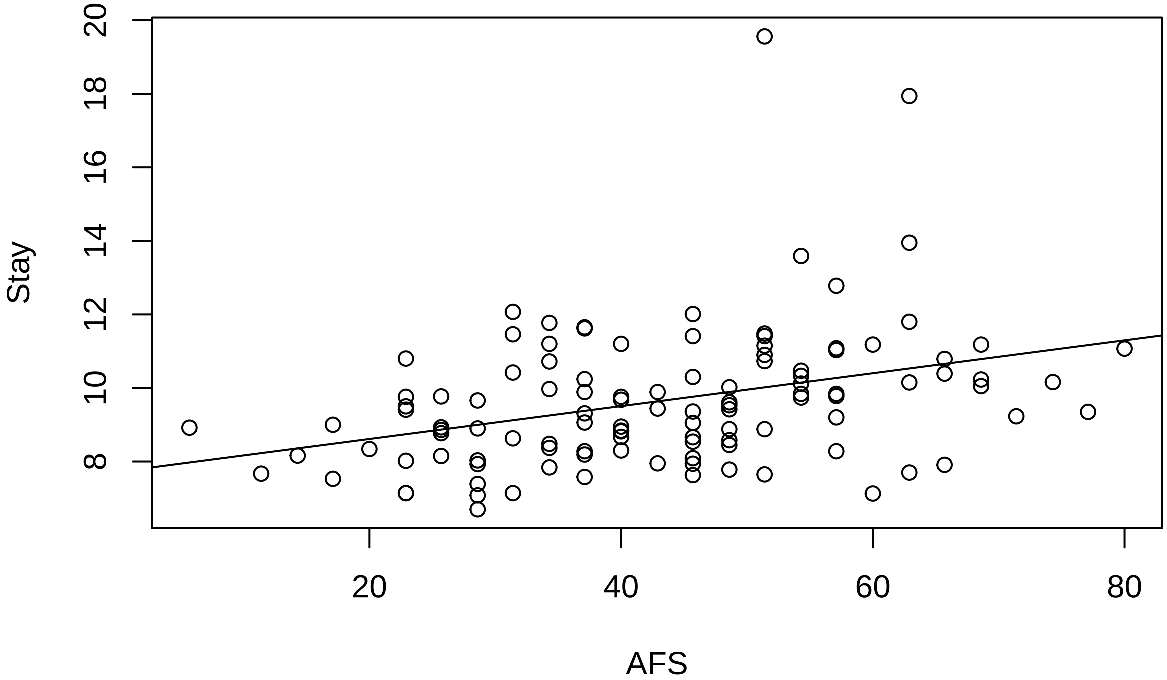

```
plot(Xray, Stay)
lines(lowess(Xray, Stay))
```



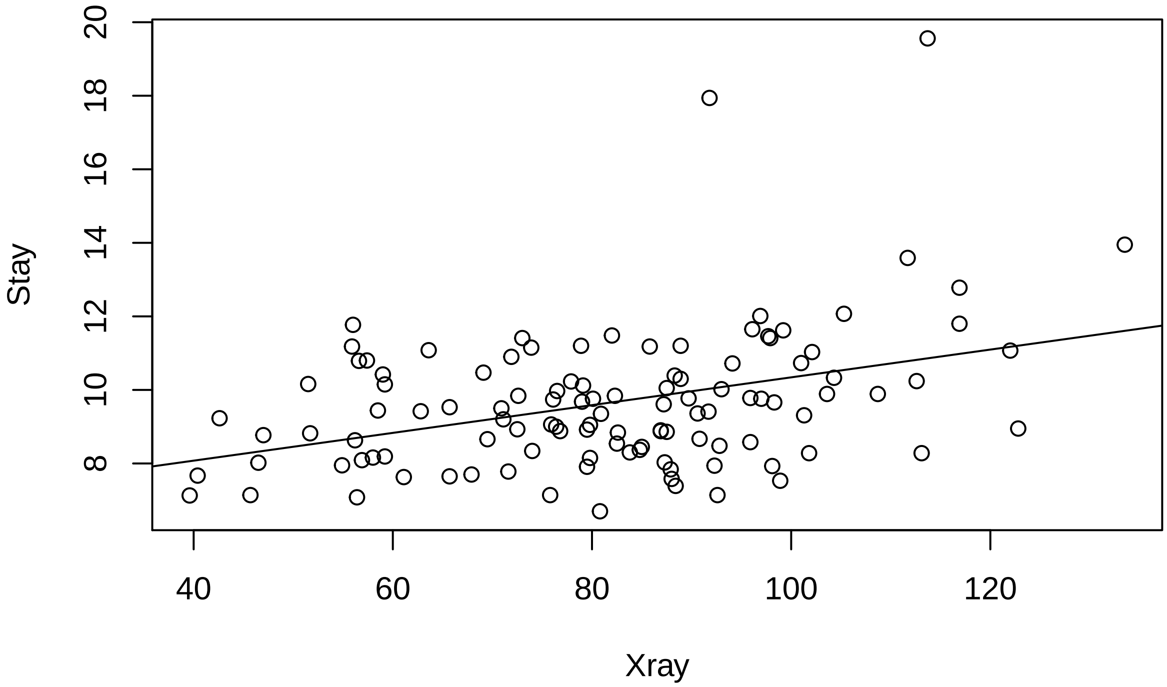

```
#The mean function does appear linear in each case.we can see a "straight line"
#from bottom left to middle right going through the center of the points.

#(b)
plot(Stay ~ Risk)
abline(lm(Stay ~ Risk))
```



```r
plot(Stay ~ AFS)
abline(lm(Stay ~ AFS))
```




```r
plot(Stay ~ Xray)
abline(lm(Stay ~ Xray))
```




```r
#The simple linear regression model seems plausible in each case.
#we can see a "straight line"from bottom left to middle right going through
```

```r
#the center of the points.

#(c)
anova(lm(Stay ~ Risk))
```

```
## Analysis of Variance Table
##
## Response: Stay
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Risk        1 116.45 116.446   44.15 1.177e-09 ***
## Residuals 111 292.76   2.638
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(Stay ~ AFS))
```

```
## Analysis of Variance Table
##
## Response: Stay
##            Df Sum Sq Mean Sq F value    Pr(>F)
## AFS         1  51.73  51.727  16.061 0.0001113 ***
## Residuals 111 357.48   3.221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(Stay ~ Xray))
```

```
## Analysis of Variance Table
##
## Response: Stay
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Xray        1  59.86  59.864  19.021 2.906e-05 ***
## Residuals 111 349.35   3.147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#MSE(Risk) = 2.638, MSE(AFS) = 3.221 and MSE(Xray) =3.147. Risk leads to
#the smallest variability around the fitted regression line.This result is
#apparent from plots in parts(a) and (b).
```