Experimental Statistics for Engineers I

Kamala Dadashova

Problem 1: The article "Parametrical Optimization of Laser Surface Alloyed NiTi Shape Memory Alloy with Co and Nb by the Taguchi Method" (J. of Engr. Manuf., 2012: 969–979) described an investigation to see whether the percent by weight of nickel in the alloyed layer is affected by carbon monoxide powder paste thickness (C, at three levels), scanning speed (B, at three levels), and laser power (A, at three levels). One observation was made at each factor-level combination [Note: Thickness column headings were incorrect in the cited article]:

a) Assuming the absence of three factor interactions (as did the investigators), $SSE = SSABC$ can be used to obtain an estimate of σ^2. Construct an ANOVA table based on this data.

```
surface <- read.csv("surface.csv")
# factors
surface$power <- as.factor(surface$power)
surface$speed<- as.factor(surface$speed)
surface$thickness <- as.factor(surface$thickness)

# fit model
fit <- aov(X.weight~ power + speed + thickness + power:speed + power:thickness+ speed:thickness,data=su
summary(fit)  # ANOVA table
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## power            2  124.6   62.30   4.849 0.0417 *
## speed            2   20.6   10.30   0.802 0.4815
## thickness        2  356.9  178.47  13.892 0.0025 **
## power:speed      4   57.5   14.37   1.119 0.4118
## power:thickness  4   61.4   15.35   1.195 0.3834
## speed:thickness  4   11.1    2.76   0.215 0.9226
## Residuals        8  102.8   12.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) Use the appropriate F ratios to show that none of the two-factor interactions is significant at a $\alpha = 0.05$?

```
falpha.AB <- qf(0.05,4,8,lower.tail=F)
falpha.AB
```

```
## [1] 3.837853
```

```
falpha.AC <- qf(0.05,4,8,lower.tail=F)
falpha.AC
```

```
## [1] 3.837853
```

```
falpha.BC <- qf(0.05,4,8,lower.tail=F)
falpha.BC
```

```
## [1] 3.837853
```

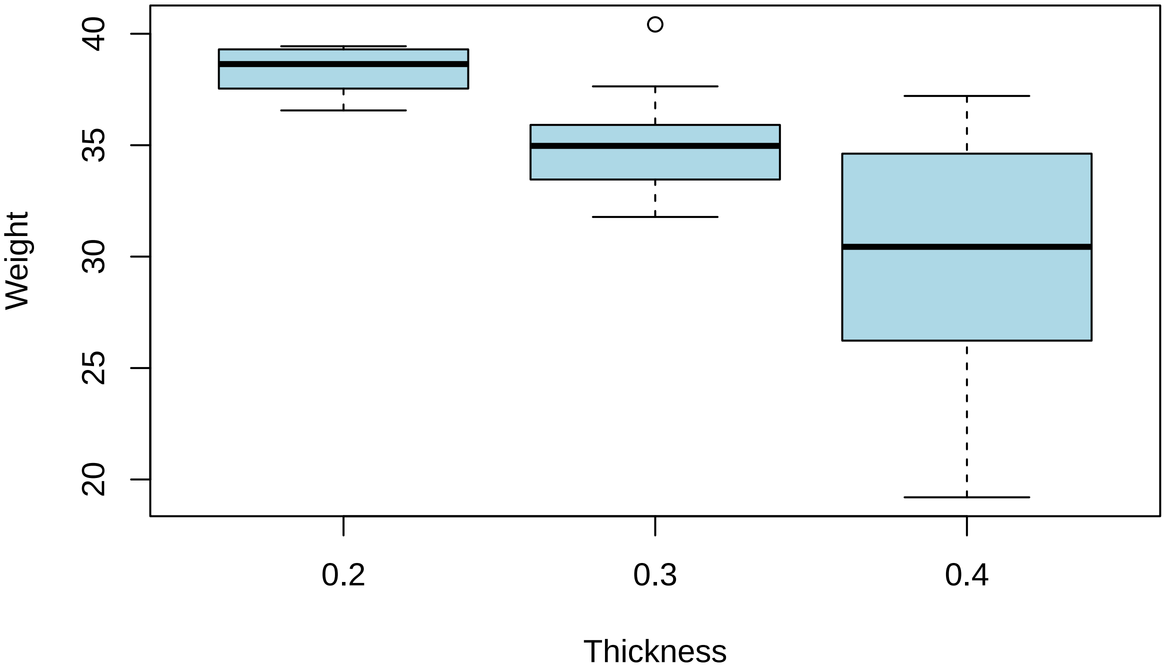
Since rejection area for AB interaction is $(3.837853, \infty)$ does not cover F value 1.119 also respective P value on Anova table is greater than 0.05, we say AB interaction term is not statistically significant. Since rejection area for AC interaction is $(3.837853, \infty)$ does not cover F value 1.195 also respective P value on Anova table is greater than 0.05, we say AC interaction term is not statistically significant. Since rejection area for BC interaction is $(3.837853, \infty)$ does not cover F value 0.215 also respective P value on Anova table is greater than 0.05, we say BC interaction term is not statistically significant.

c) Which main effects are significant at a $\alpha = 0.05$?

According to the P-values, one can say the factor A=power and C=thickness($0.0417 < 0.05$,$0.0025 < 0.05$) main effects are statistically significant at the .05 level while the factor B=speed ($0.4815 > 0.05$) main effect is not statistically significant.

d) Use Tukey's procedure with a simultaneous confidence level of 95% to identify significant differences between levels of paste thickness.

```
boxplot(X.weight~thickness,data=surface,col="light blue",xlab="Thickness",ylab="Weight")
```



```
TukeyHSD(fit)$thickness
```

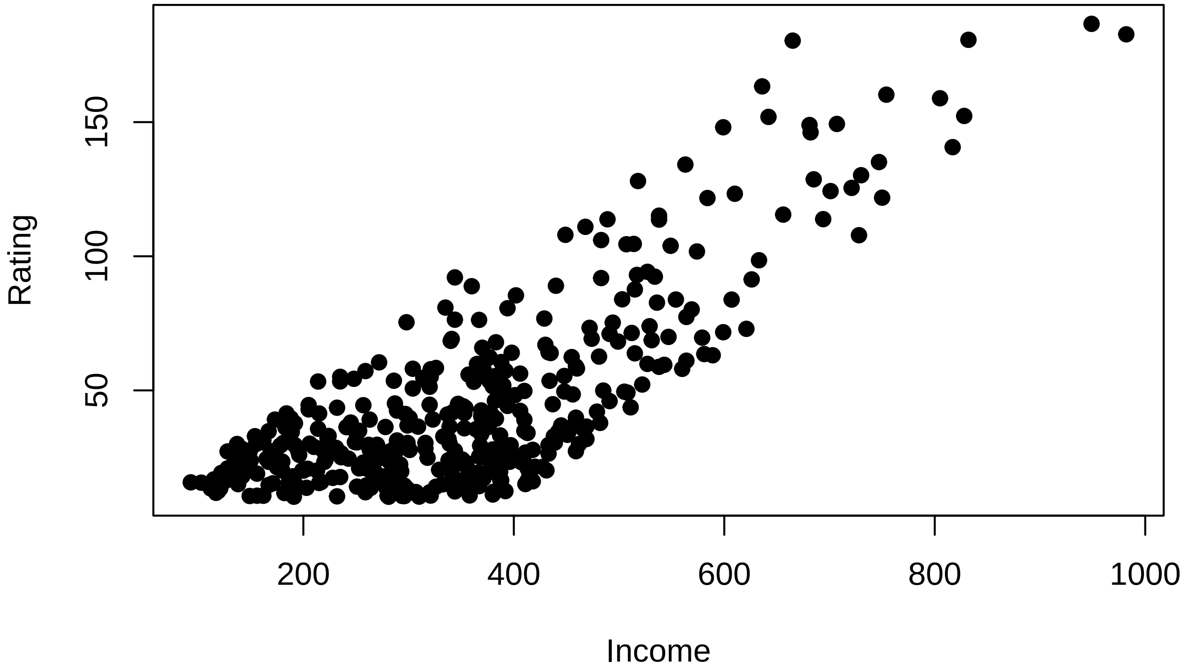
```
##               diff        lwr        upr       p adj
## 0.3-0.2 -3.172222  -8.000324  1.6558792 0.206795175
## 0.4-0.2 -8.793333 -13.621435 -3.9652319 0.002080907
## 0.4-0.3 -5.621111 -10.449213 -0.7930097 0.025229515
```

We see that for difference 0.3-0.2 p value 0.206 is bigger than 0.05 so it does not have that much significance. However other two levels p values 0.002080907 and 0.025229515, which both are less than 0.05. They are significant.

Problem 2: This exercise we will compare simple linear regression models using the Credit data to understand which variables are more important contributors to credit rating.

a) Use the lm() function to perform a simple linear regression with Rating as the response and Income as the predictor. Use the plot() function to examine residuals to ensure we are not violation assumptions for least squares. Use the summary() function to print the results. Comment on the output.

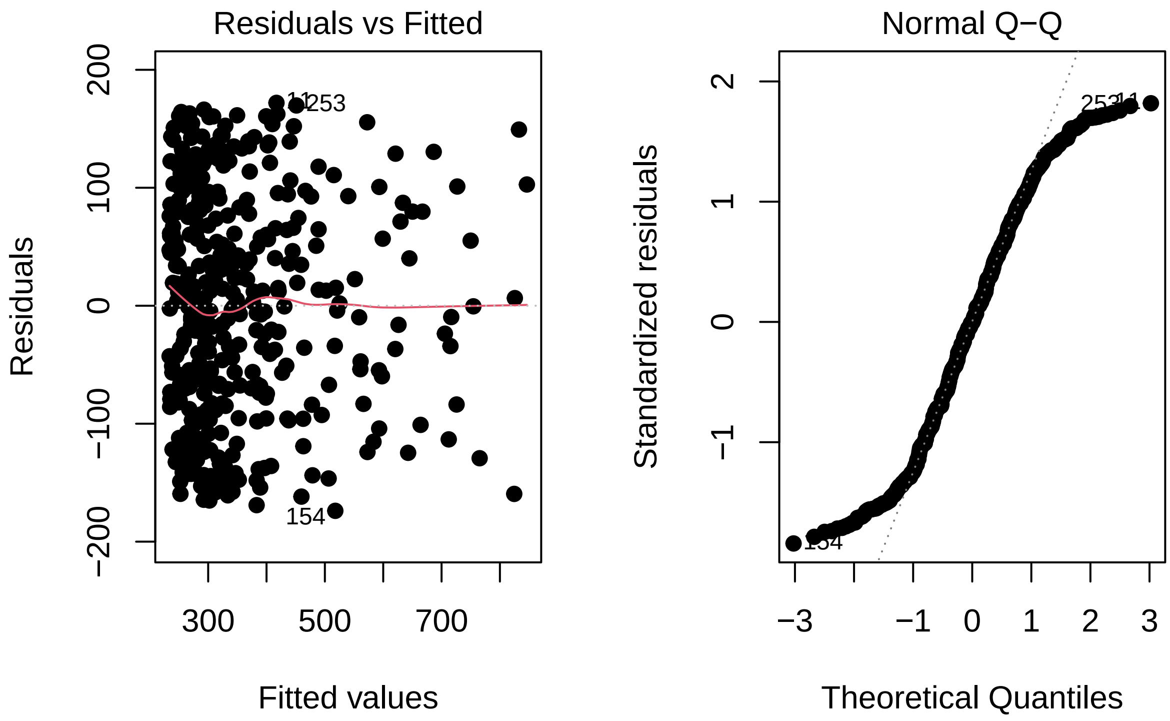
```
Credit <- read.csv("Credit.csv")
par(mfrow=c(1,1))
plot(Credit$Rating,Credit$Income,pch=19,
     xlab="Income", ylab="Rating")
```



```
cor(Credit$Rating,Credit$Income)
```

```
## [1] 0.7913776
```

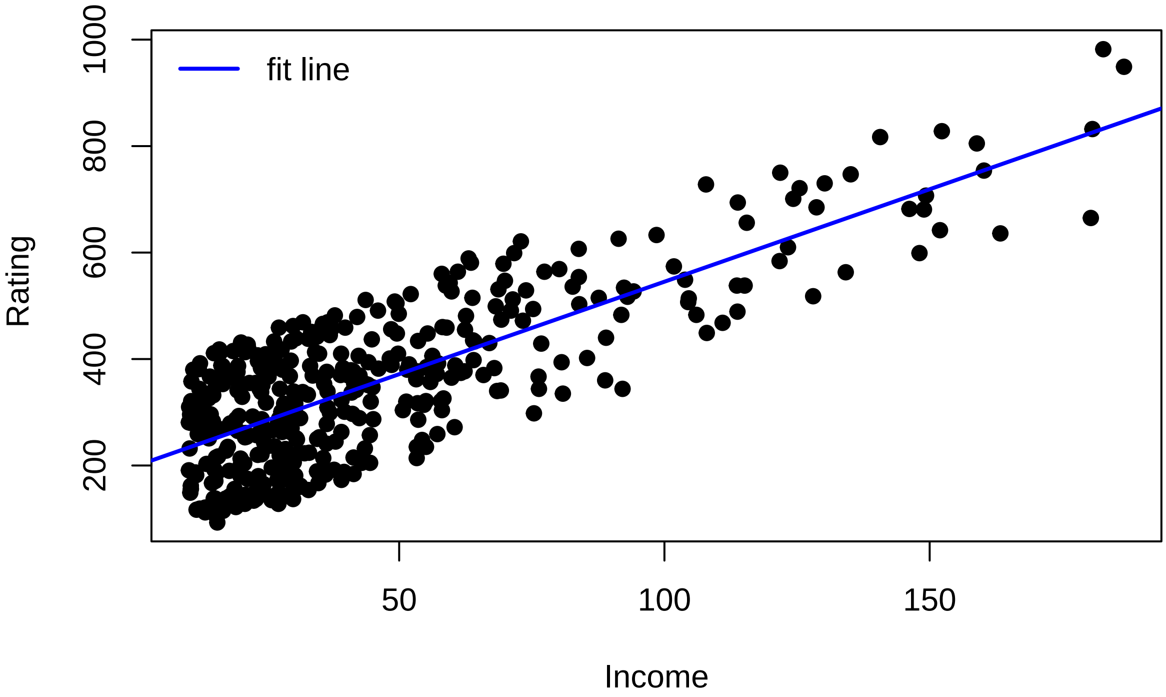
```
# fit model
fit <- lm(Rating~Income,Credit)
par(mfrow=c(1,2))
plot(fit,1:2,pch=19)  # residual diagnostics
```



```
summary(fit) # fit summary
```

```
##
## Call:
## lm(formula = Rating ~ Income, data = Credit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -173.855  -79.417   -0.384   79.747  171.955
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 197.8411     7.7089   25.66   <2e-16 ***
## Income        3.4742     0.1345   25.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.71 on 398 degrees of freedom
## Multiple R-squared:  0.6263, Adjusted R-squared:  0.6253
## F-statistic:   667 on 1 and 398 DF,  p-value: < 2.2e-16
```

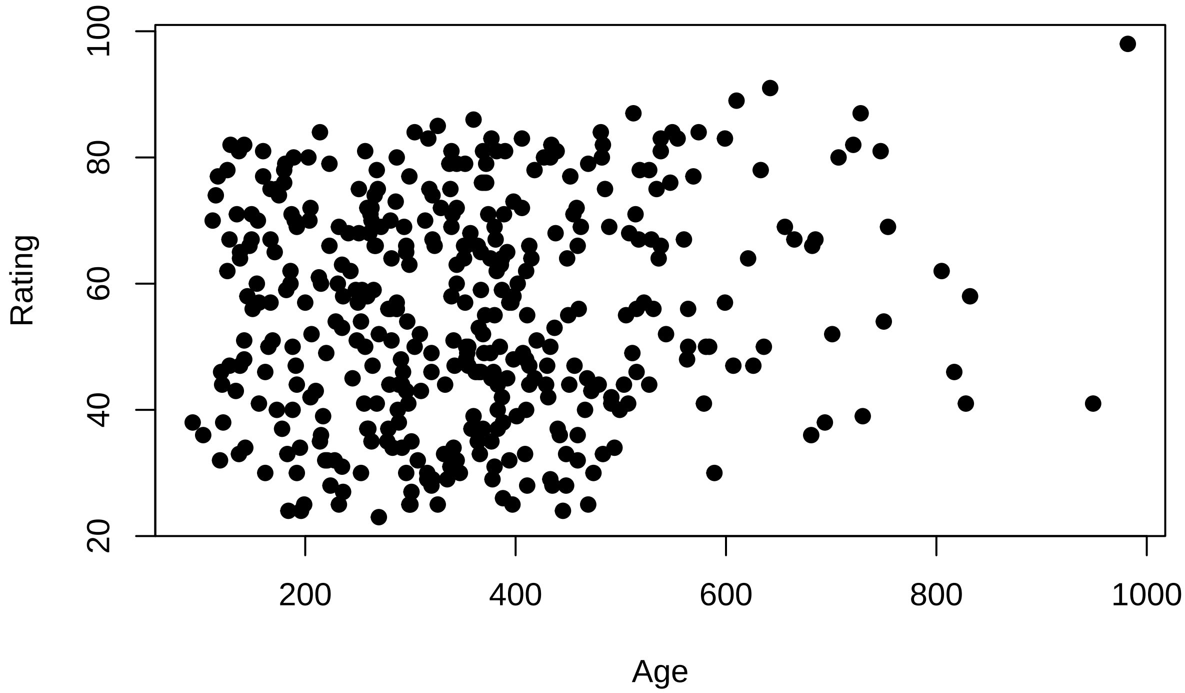
```
# scatterplot
par(mfrow=c(1,1))
plot(Credit$Income,Credit$Rating,pch=19,
     xlab="Income", ylab=" Rating")
abline(fit,lwd=2,col="blue")
legend("topleft",legend="fit line",lwd=2,col="blue",bty="n")
```



Looking at the plots and summary of this fit, we can suppose the data is reliable as the residuals are centered around zero and while the tail ends of the normal plot seem to deviate from the normal line. Looking at the summary there does appear to be a relationship between the predictor and response as the coefficients show us there is a statistically significant positive correlation with a slope of 3.47. We can say this is a strong correlation as the R Squared value is 0.62.

b) Repeat (a) using Age as the predictor.

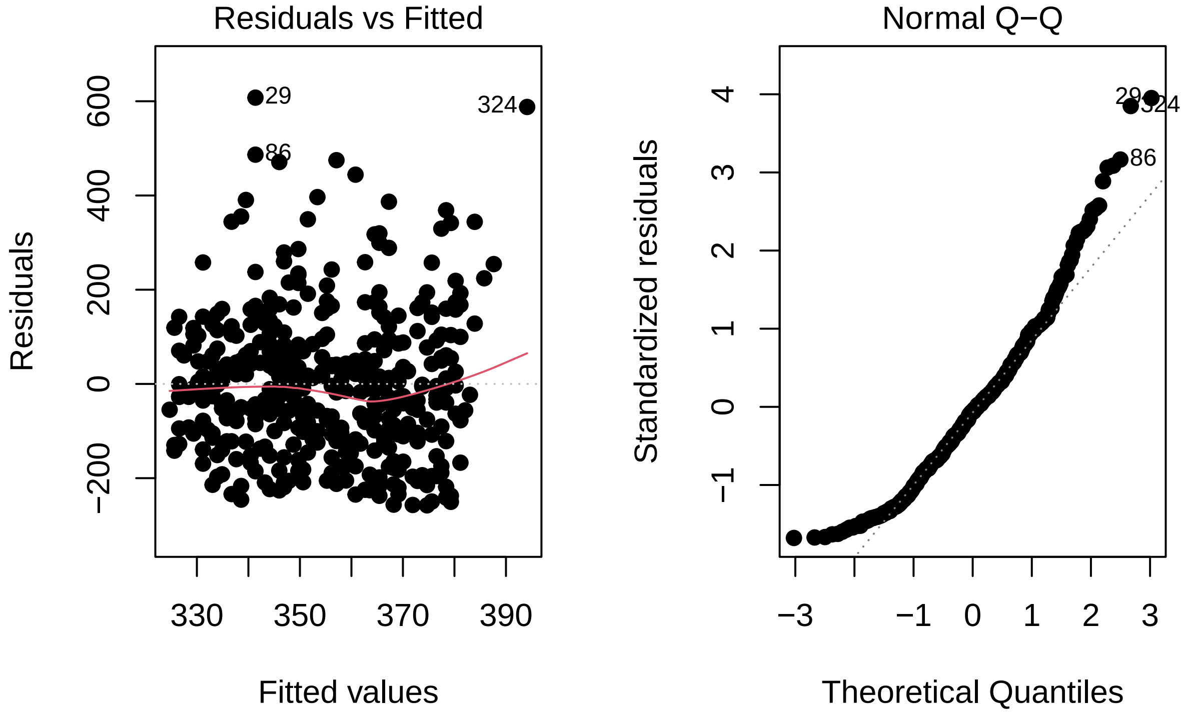
```
Credit <- read.csv("Credit.csv")
par(mfrow=c(1,1))
plot(Credit$Rating,Credit$Age,pch=19,
     xlab="Age", ylab="Rating")
```



```
cor(Credit$Rating,Credit$Age)
```

```
## [1] 0.103165
```

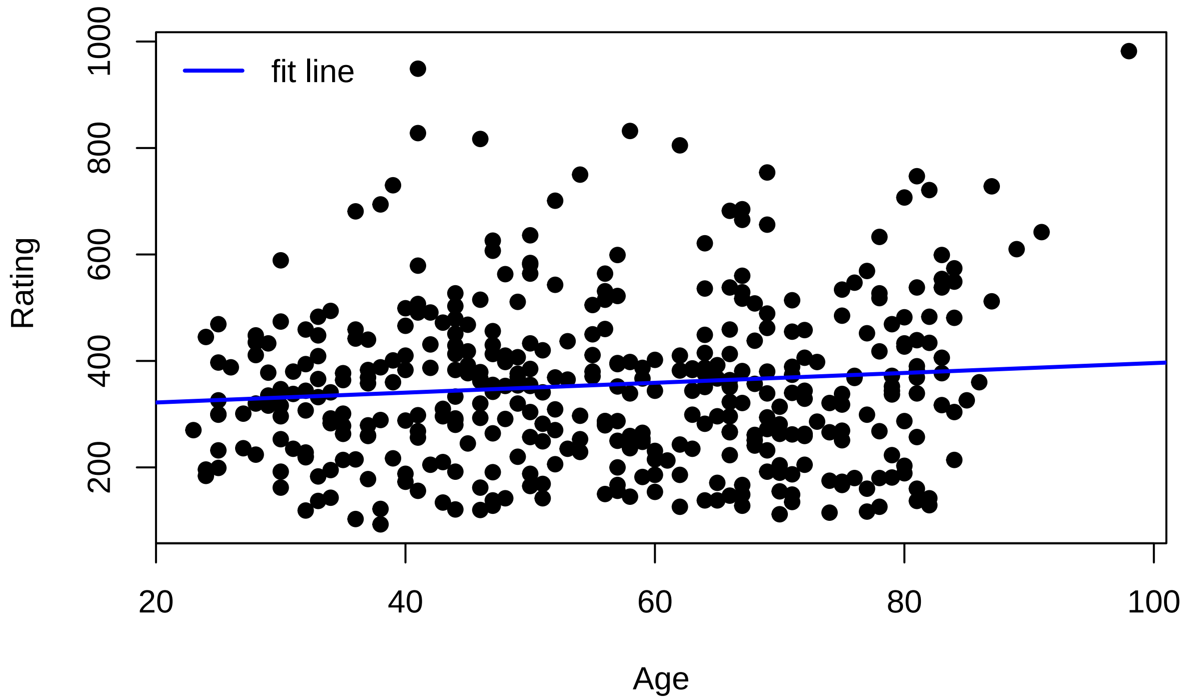
```
# fit model
fit <- lm(Rating~Age,Credit)
par(mfrow=c(1,2))
plot(fit,1:2,pch=19)   # residual diagnostics
```



```
summary(fit) # fit summary
```

```
##
## Call:
## lm(formula = Rating ~ Age, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -257.68 -107.25   -9.37   85.11  607.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 303.4281    26.0599  11.643   <2e-16 ***
## Age           0.9254     0.4472   2.069   0.0392 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 154.1 on 398 degrees of freedom
## Multiple R-squared:  0.01064,    Adjusted R-squared:  0.008157
## F-statistic: 4.281 on 1 and 398 DF,  p-value: 0.03917
```

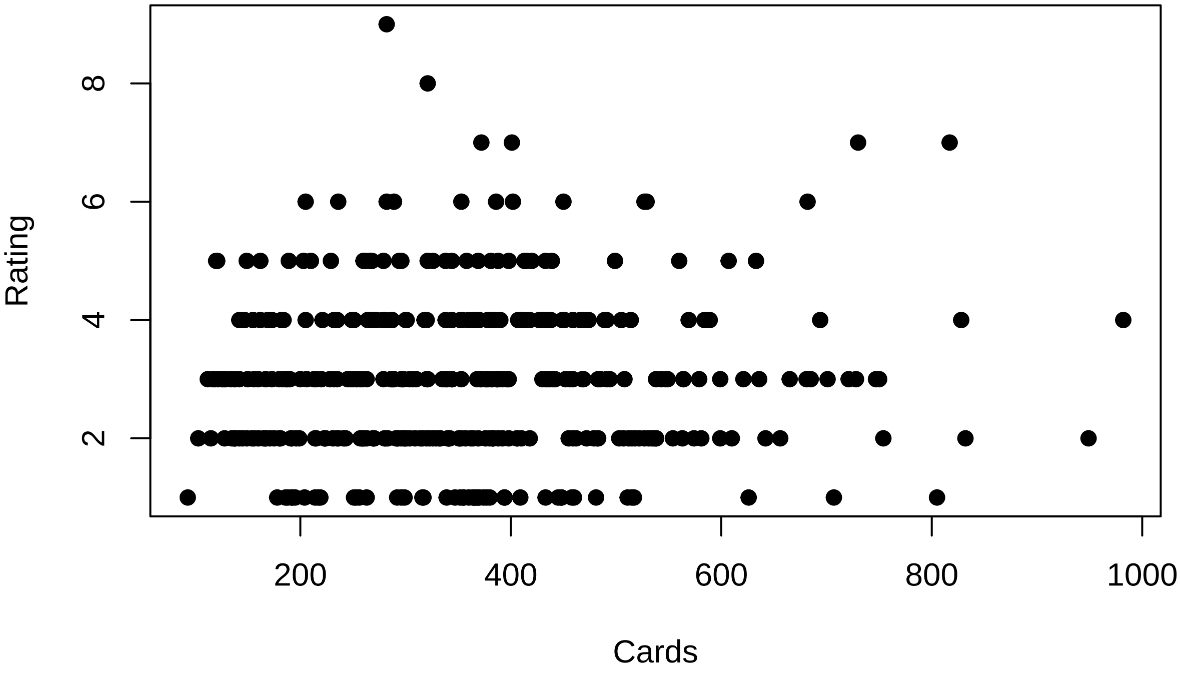
```
# scatterplot of Gore vs. Clinton with fit line overlay
par(mfrow=c(1,1))
plot(Credit$Age,Credit$Rating,pch=19,
     xlab="Age", ylab=" Rating")
abline(fit,lwd=2,col="blue")
legend("topleft",legend="fit line",lwd=2,col="blue",bty="n")
```



The summary tells us this correlation has a slope of 0.925 with a significance level below 0.05. We see that it is a fairly week correlation as the R squared value is only 0.01064.

c) Repeat (a) using Cards (number of credit cards the individual possesses) as the predictor; be sure to use the as.factor() function so that R treats Cards as a categorical variable.

```
Credit <- read.csv("Credit.csv")
Cards<-as.factor(Credit$Cards)
par(mfrow=c(1,1))
plot(Credit$Rating,Credit$Cards,pch=19,
     xlab="Cards", ylab="Rating")
```

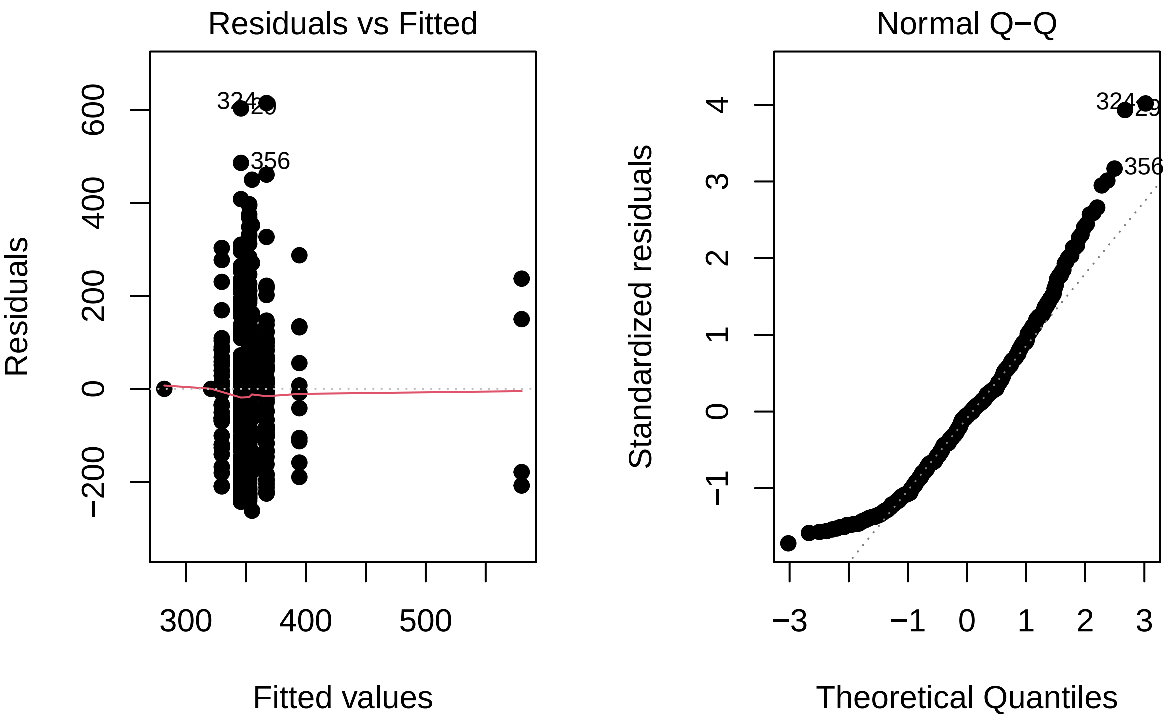


```
cor(Credit$Rating,Credit$Cards)
```

```
## [1] 0.05323903
```

```
# fit model
fit <- lm(Rating~as.factor(Cards),Credit)
par(mfrow=c(1,2))
plot(fit,1:2,pch=19)  # residual diagnostics
```

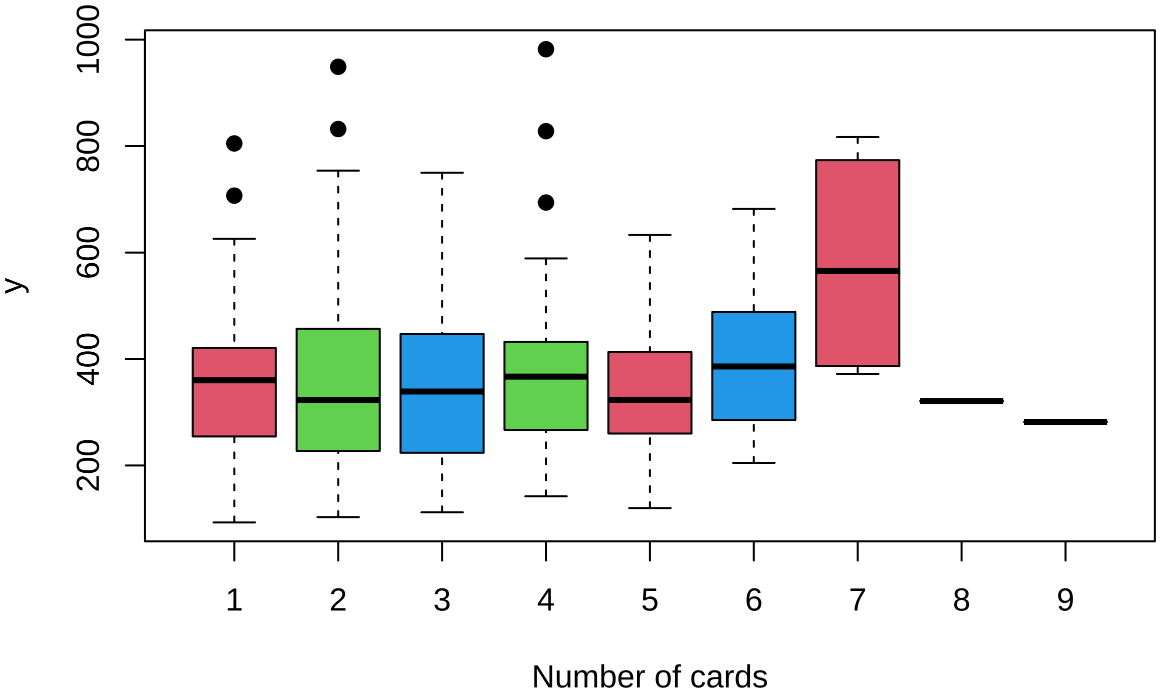
```
## Warning: not plotting observations with leverage one:
##   206, 384
```



```
summary(fit) # fit summary
```

```
##
## Call:
## lm(formula = Rating ~ as.factor(Cards), data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -262.14 -110.15   -8.39   83.86  614.75
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          355.137     21.576  16.460  < 2e-16 ***
## as.factor(Cards)2     -9.233     25.923  -0.356  0.72191
## as.factor(Cards)3     -2.326     26.066  -0.089  0.92893
## as.factor(Cards)4     12.113     28.201   0.430  0.66779
## as.factor(Cards)5    -25.225     34.115  -0.739  0.46010
## as.factor(Cards)6     39.499     51.225   0.771  0.44112
## as.factor(Cards)7    224.863     80.007   2.811  0.00519 **
## as.factor(Cards)8    -34.137    155.590  -0.219  0.82645
## as.factor(Cards)9    -73.137    155.590  -0.470  0.63857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 154.1 on 391 degrees of freedom
## Multiple R-squared:  0.02811,    Adjusted R-squared:  0.008226
## F-statistic: 1.414 on 8 and 391 DF,  p-value: 0.1888
```

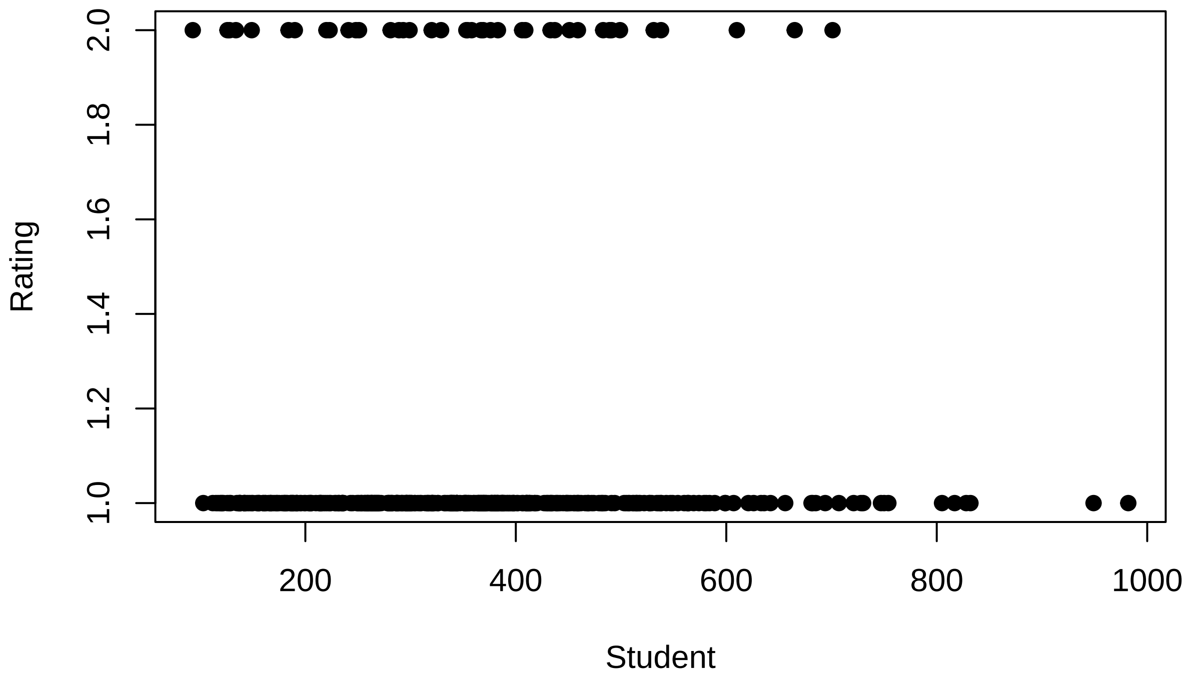
```
par(mfrow=c(1,1))
plot(Cards,Credit$Rating,col=Cards,pch=19,xlab="Number of cards")
```



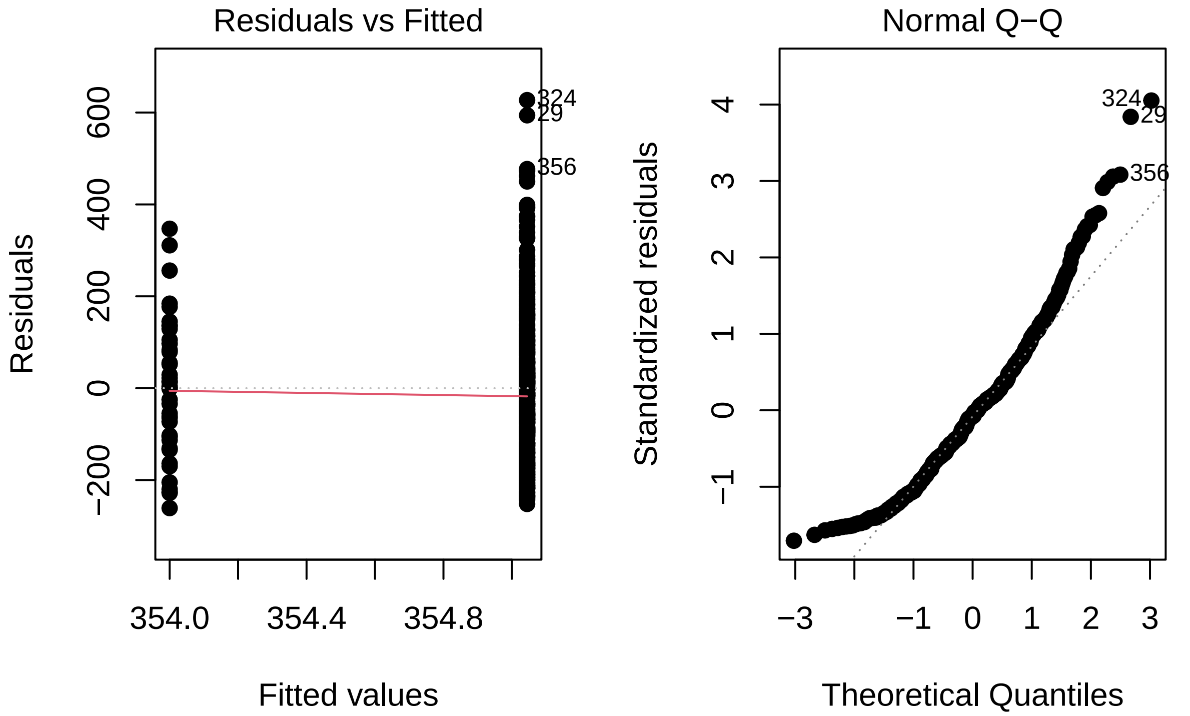
We observe relatively constant variance versus fitted values and Q-Q plot shows reasonable normality. F-statistic is 1.414 and overall model is not significant since p=0.1888 is bigger than 0.05. $\hat{\beta_0}$= 355.137 with p= 2e-16, so predictor is significant at 0.05 level. Conclusion $\hat{\beta_0} \neq 0$. $\hat{\beta_1} = -9.233$, $\hat{\beta_2} = -2.326$, $\hat{\beta_3} = 12.113$, $\hat{\beta_4} = -25.225$, $\hat{\beta_5} = 39.499$, $\hat{\beta_7} = -34.137$, $\hat{\beta_8} = -73.137$ and corresponding p values are all bigger than 0.05 level. So, we cannot conclude either are different from β_0. $\hat{\beta_6} = 224.863$ and its p value $0.00519 > 0.05$ so it is marginally significant at 0.05 level.

d) Repeat (a) using Student as the predictor.

```
Credit <- read.csv("Credit.csv")
par(mfrow=c(1,1))
plot(Credit$Rating,as.factor(Credit$Student),pch=19,
     xlab="Student", ylab="Rating")
```



```
# fit model
fit <- lm(Rating~as.factor(Student),Credit)
par(mfrow=c(1,2))
plot(fit,1:2,pch=19)  # residual diagnostics
```

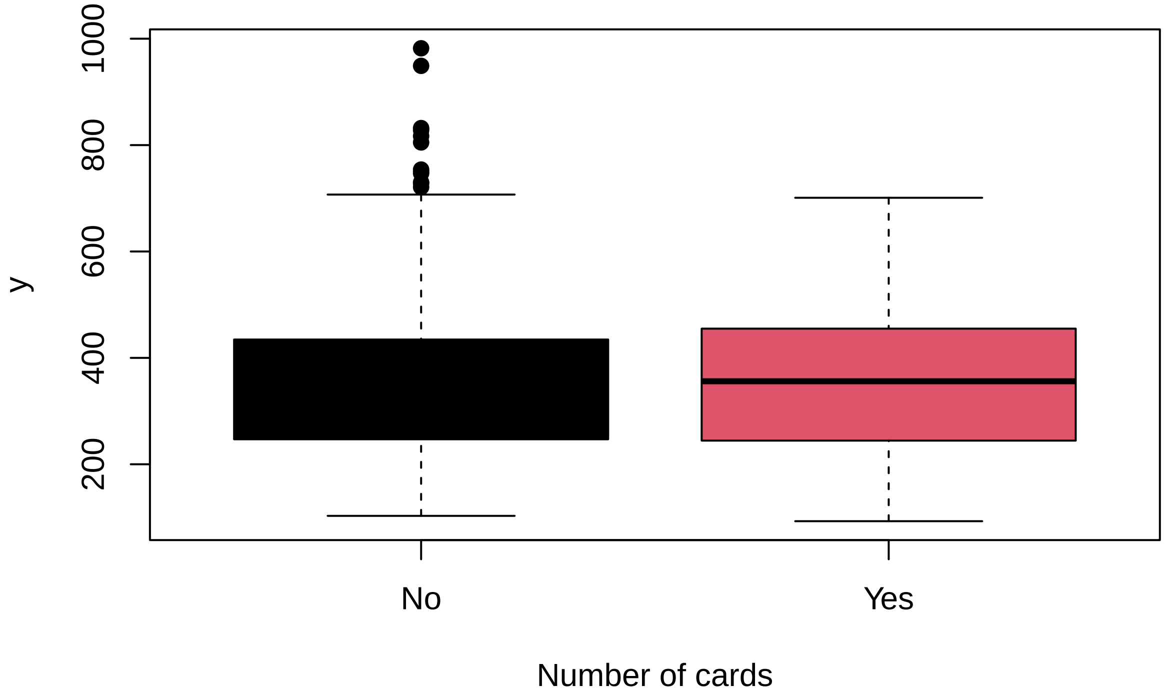


```
summary(fit) # fit summary
```

```
##
## Call:
## lm(formula = Rating ~ as.factor(Student), data = Credit)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -261.00 -107.04  -11.04   82.97  626.96
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            355.044      8.165   43.48   <2e-16 ***
## as.factor(Student)Yes   -1.044     25.820   -0.04    0.968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 154.9 on 398 degrees of freedom
## Multiple R-squared:  4.111e-06,  Adjusted R-squared:  -0.002508
## F-statistic: 0.001636 on 1 and 398 DF,  p-value: 0.9678
```

```
par(mfrow=c(1,1))
plot(as.factor(Credit$Student),Credit$Rating,col=as.factor(Credit$Student),pch=19,xlab="Number of cards
```



This summary shows us that there is no relationship between the predictor and response. While there is a slope coefficient of -1 along with the P value of 0.97 means, we can disregard it as statistically insignificant. Furthermore, the extremely small R-squared value of 4.11e-06 confirms that there is no correlation.

e) Based on this analysis, which of the predictor variables above has the strongest relationship with an individual's credit rating?

To sum up, according to these summaries, it seems that income is the strongest predictor of credit rating with the highest R squared rating and the most statistically significant coefficients.

Problem 3: In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results.

a) Using the rnorm() function, create a vector, x, containing 100 observations drawn from a N(0, 1) distribution. This represents a feature, X.

```
set.seed(1)
X=rnorm(100,0,1)
```

b) Using the rnorm() function, create a vector, eps, containing 100 observations drawn from a N(0, 0.25) distribution i.e. a normal distribution with mean zero and variance 0.25.

```
eps=rnorm(100,0,0.25)
```

c) Using x and eps, generate a vector y according to the model $Y = -1 + 0.5X + \epsilon$. What is the length of the vector y? What are the values of β_0 and β_1 in this linear model?

$\beta_0 = -1$ and $\beta_1 = 0.5$.

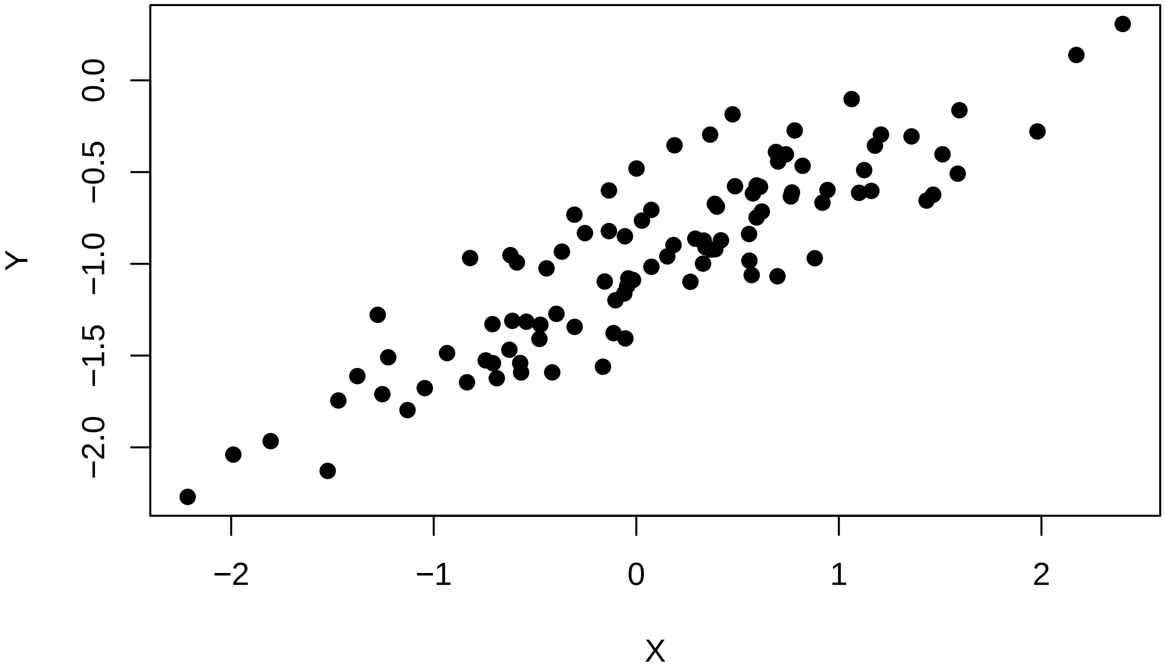
```
Y=-1+0.5*X+eps
length(Y)
```

```
## [1] 100
```

d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

Correlation number 0.88 is close to 1 and it seems there is a strong positive correlation.

```
par(mfrow=c(1,1))
plot(X,Y,pch=19)
```

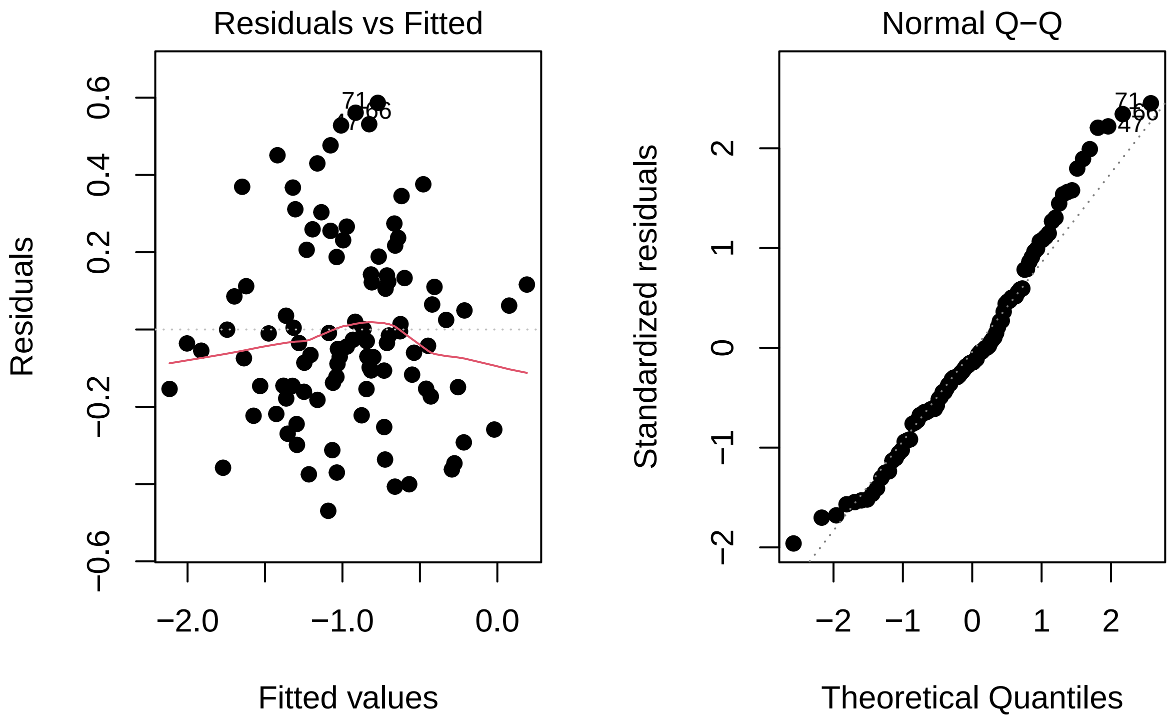


```
cor(X,Y)
```

```
## [1] 0.8822902
```

e) Fit a least squares linear model to predict y using x. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

```
fit1 <- lm(Y~X)
par(mfrow=c(1,2))
plot(fit1,1:2,pch=19)   # residual diagnostics
```



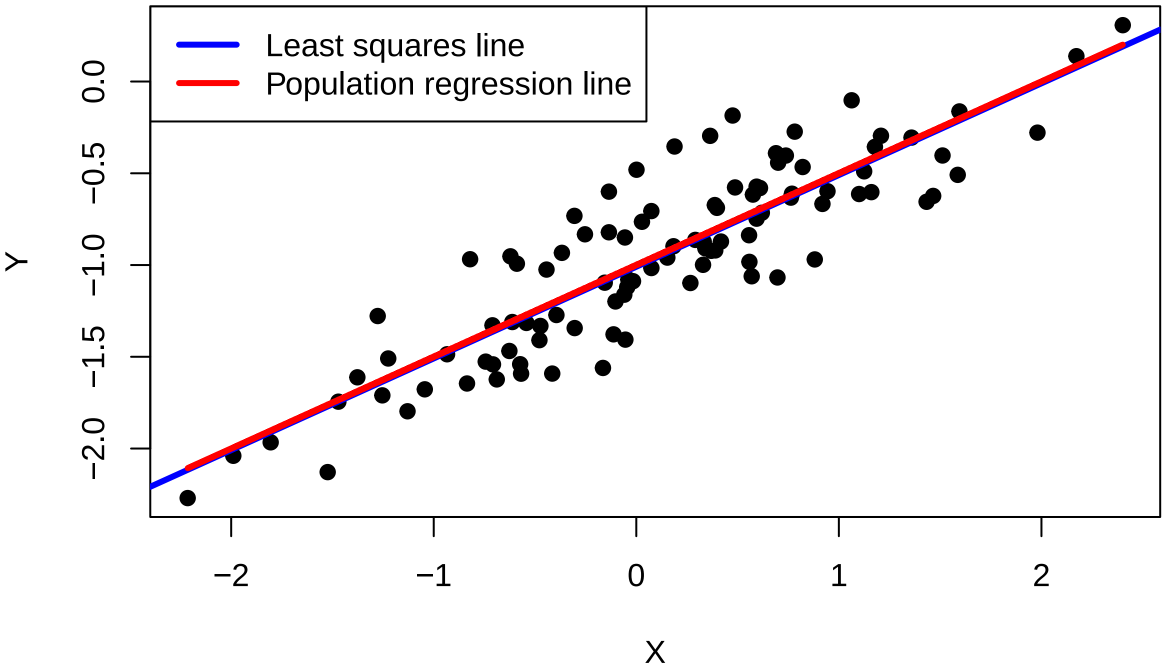
```
summary(fit1)   # fit summary
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46921 -0.15344 -0.03487  0.13485  0.58654
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.00942    0.02425  -41.63   <2e-16 ***
## X            0.49973    0.02693   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2407 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

We see on the summary table that estimates for β_0 and β_1 are -1.00942 and 0.49973 which are close to true value of $\beta_0 = -1$ and $\beta_1 = 0.5$.

f) Display the least squares line on the scatterplot obtained in `(d)`. Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.

```
# scatterplot with fit line
par(mfrow=c(1,1))
plot(X,Y,pch=19)
abline(fit1,lwd=3,col="blue")
lines(X,-1+0.5*X,lwd =3, col = "red")
legend("topleft",legend=c("Least squares line","Population regression line"),lwd=3,col=c("blue", "red")
```

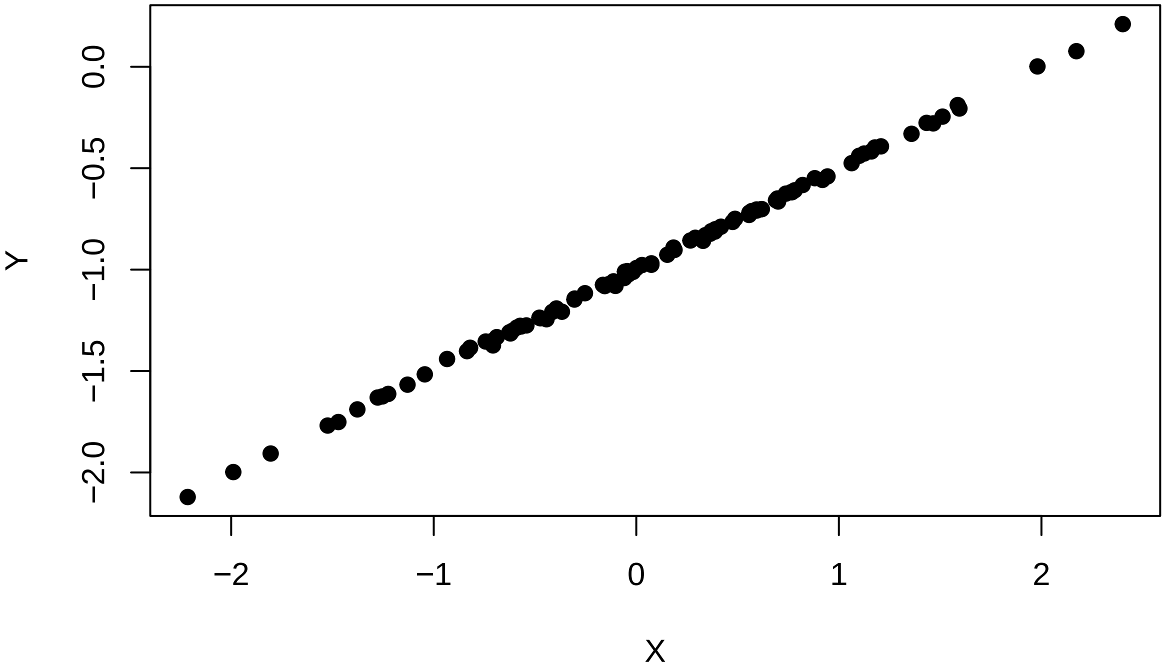


g) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The model should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

```
eps=rnorm(100,0,0.01)
Y=-1+0.5*X+eps
length(Y)
```

```
## [1] 100
```

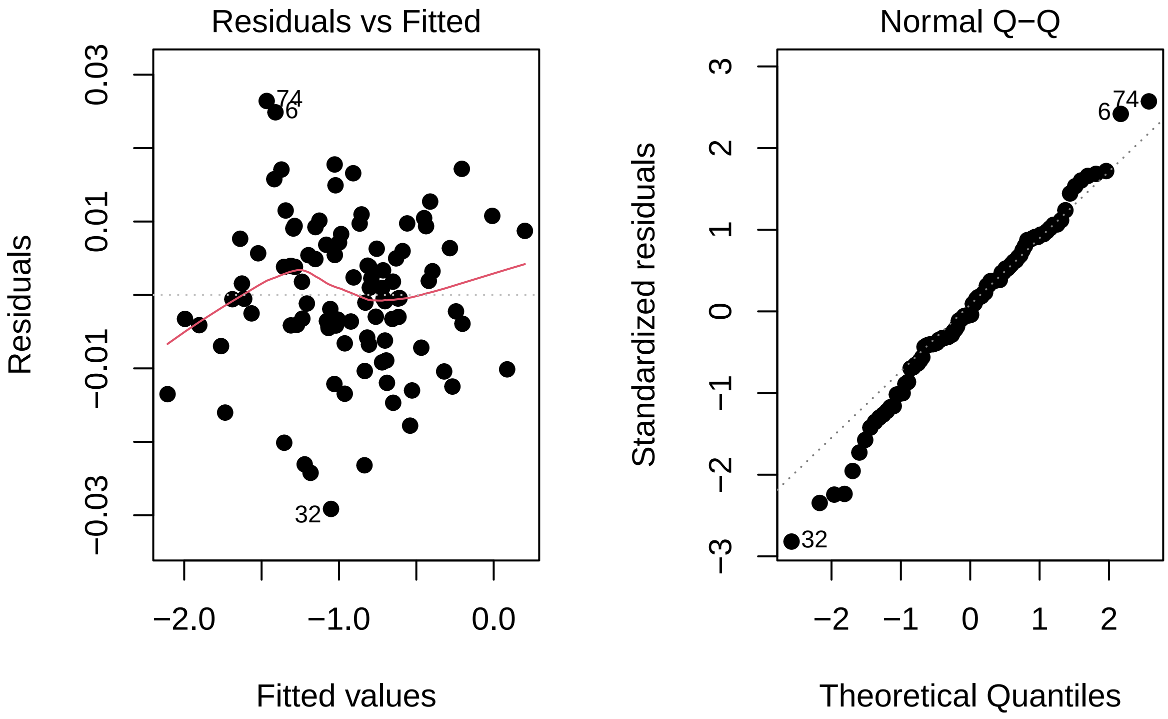
```
par(mfrow=c(1,1))
plot(X,Y,pch=19)
```



```
cor(X,Y)
```

```
## [1] 0.9997352
```

```
fit2 <- lm(Y~X)
par(mfrow=c(1,2))
plot(fit2,1:2,pch=19)   # residual diagnostics
```

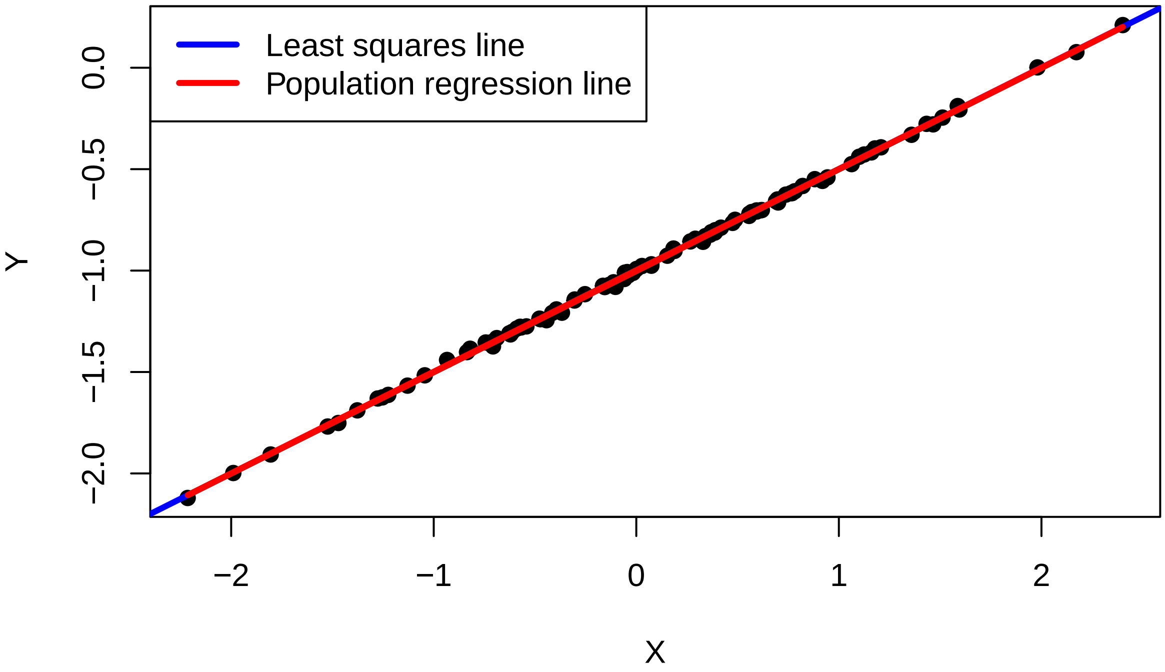


```
summary(fit2)    # fit summary
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0291411 -0.0048230 -0.0004533  0.0064924  0.0264157
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.999726   0.001047  -954.8   <2e-16 ***
## X            0.500212   0.001163   430.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01039 on 98 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 1.85e+05 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,1))
plot(X,Y,pch=19)
abline(fit2,lwd=3,col="blue")
```

```
lines(X,-1+0.5*X,lwd =3, col = "red")
legend("topleft",legend=c("Least squares line","Population regression line"),lwd=3,col=c("blue", "red")
```



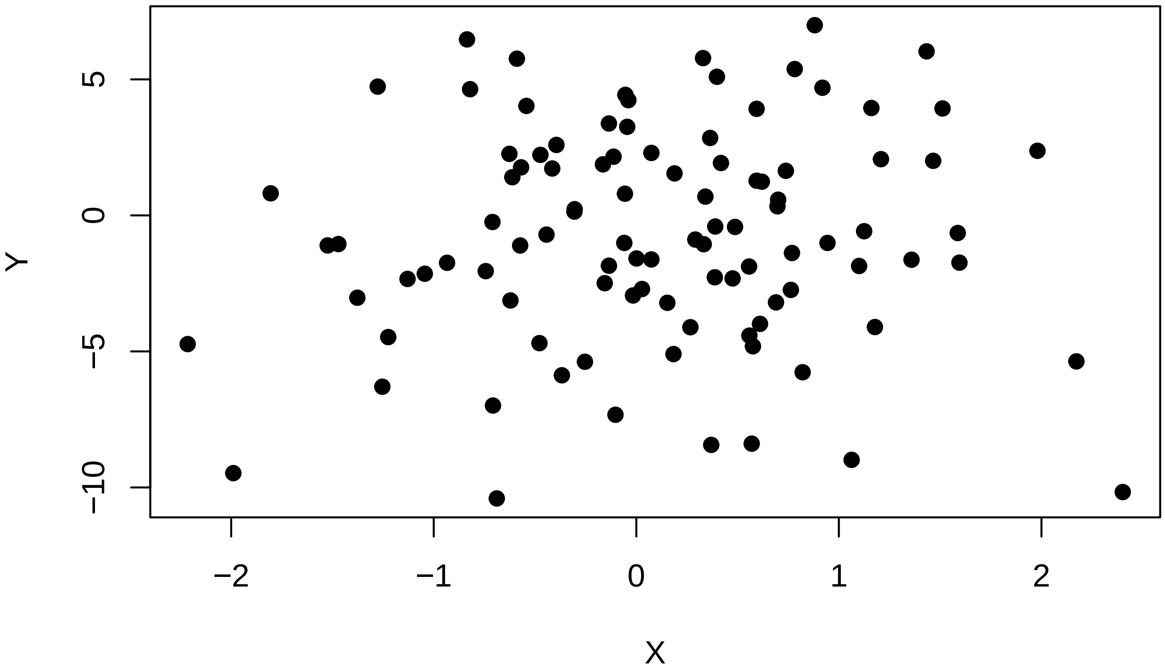
We take smaller variance and observe that least square line and population regression line overlap. It is clear even from the summary table since $R^2 = 0.9995$.

h) Repeat `(a)-(f)` after modifying the data generation process in such a way that there is more noise in the data. The model should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term in `(b)`. Describe your results.

```
eps=rnorm(100,0,4)
Y=-1+0.5*X+eps
length(Y)
```

```
## [1] 100
```

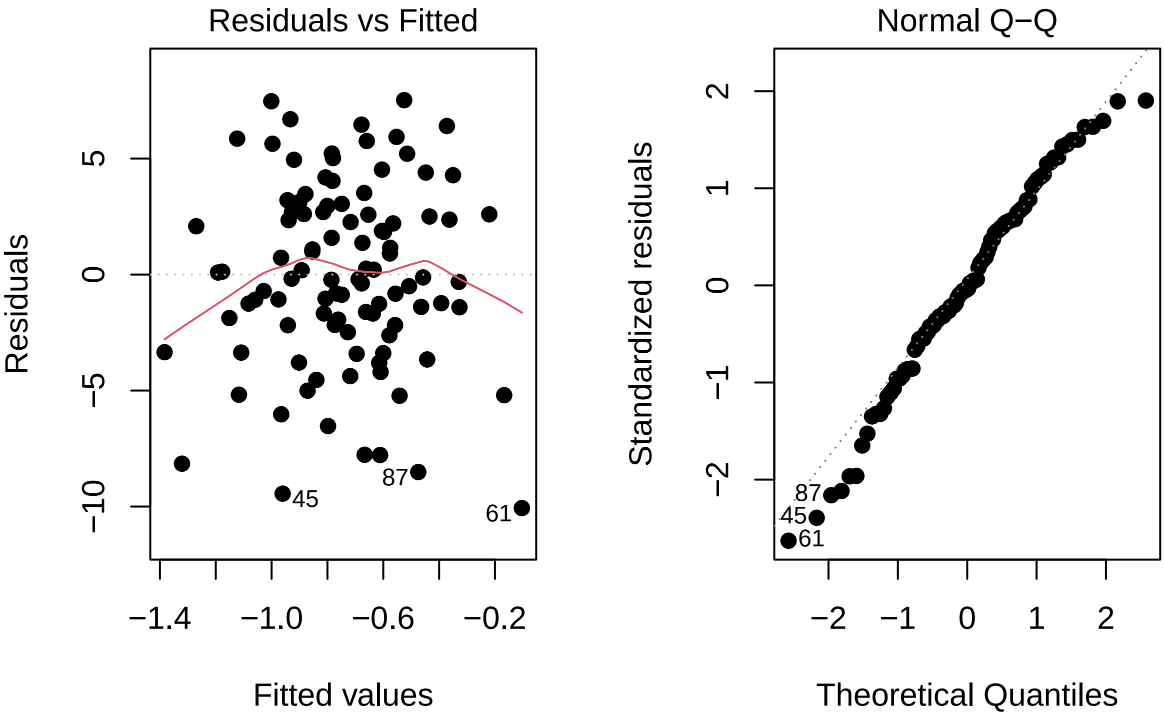
```
par(mfrow=c(1,1))
plot(X,Y,pch=19)
```



```
cor(X,Y)
```

```
## [1] 0.06273656
```

```
fit3 <- lm(Y~X)
par(mfrow=c(1,2))
plot(fit3,1:2,pch=19)   # residual diagnostics
```

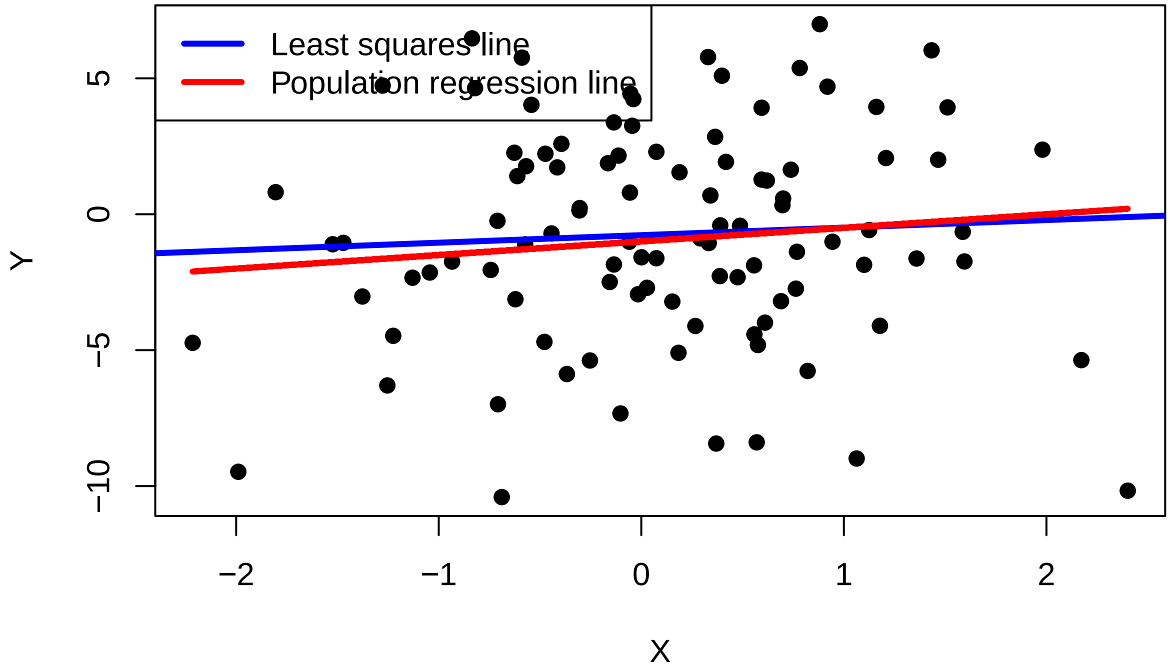


```
summary(fit3)    # fit summary
```

```
##
```

```
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.065  -2.181  -0.151   2.692   7.516
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.7693     0.4011  -1.918    0.058 .
## X             0.2773     0.4455   0.622    0.535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.982 on 98 degrees of freedom
## Multiple R-squared:  0.003936,   Adjusted R-squared:  -0.006228
## F-statistic: 0.3872 on 1 and 98 DF,  p-value: 0.5352
```

```
par(mfrow=c(1,1))
plot(X,Y,pch=19)
abline(fit3,lwd=3,col="blue")
lines(X,-1+0.5*X,lwd =3, col = "red")
legend("topleft",legend=c("Least squares line","Population regression line"),lwd=3,col=c("blue", "red"))
```



Now, we observe that that least square line and population regression line do not overlap. It is because we have smaller $R^2 = 0.9995$ value.

i) What are the confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$ based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

```
confint(fit1)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0575402 -0.9613061
## X            0.4462897  0.5531801
```

```
confint(fit2)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0018041 -0.9976485
## X            0.4979038  0.5025196
```

```
confint(fit3)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.5653598 0.02668273
## X           -0.6069117 1.16142321
```

This results are consistent since we also expect with the low noise trial have the narrowest confidence interval and the highest noise trial with the widest interval.