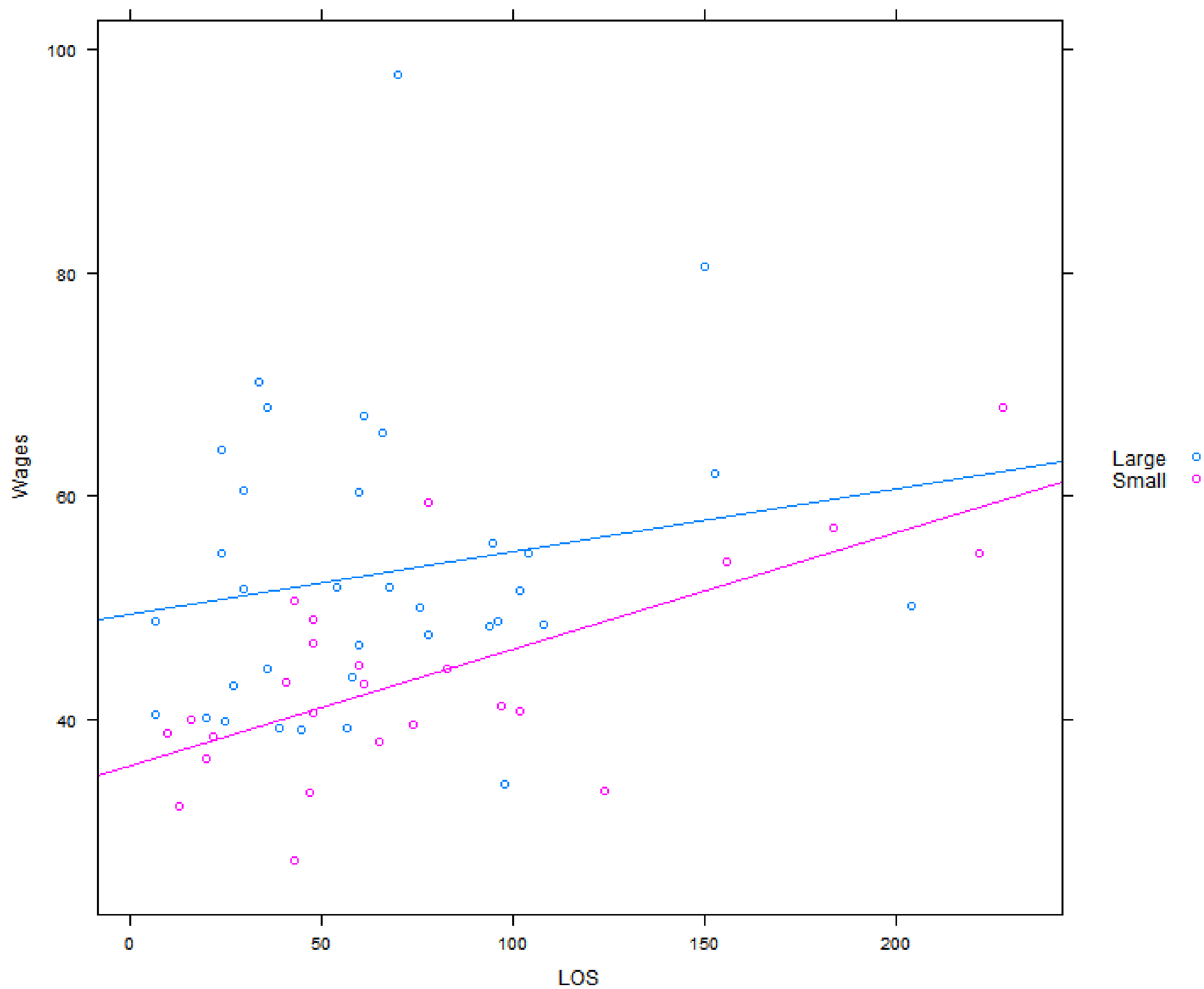
1)A)



B) *For Large bank,* $\hat{y}$ = 49.54532 + 0.05595x

For Small bank, $\hat{y}$ = 35.87192 + 0.10416x

C) For Large Bank, estimates of the intercept, β_0 = 49.54532

95% Confidence interval of intercept β_0

	fit	lwr	upr
$\widehat{\beta_0}$	49.54532	41.38071	57.70992

For Small Bank, estimates of the intercept, β_0 = 35.87192

95% Confidence interval of intercept β_0

	fit	lwr	upr
$\widehat{\beta_0}$	35.87192	31.15136	40.59248

Based on the confidence interval for the intercepts of both large and small banks, a brand new employee be better off at a large bank.

D)

Estimate of the slope for large bank $\widehat{\beta_1}$ = 0.05595

Estimate of the slope for small bank $\widehat{\beta_1}$ = 0.10416

From the figure of the slopes, we can interpret that although the starting salary of the large bank for a brand new employee is better, an employee will have a larger increase in salary in a small bank over time.

E)

For large bank,

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

p-value = 0.282 > 0.05

Conclusion: The null hypothesis can not be rejected with 95% confidence.

For small bank,

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

p-value = 0.000171 < 0.05

Conclusion: The null hypothesis can be rejected with 95% confidence, meaning we can be 95% confident that the slope of the true regression line is zero.

For the lack of Fit test, we make the following hypothesis:

H_0 : The linear regression model is appropriate

H_1 : The linear regression model is not appropriate

p-value for large bank = 0.2442 > 0.05, which means there is no evidence of lack of fit for the linear regression for the large bank.

p-value for small bank = 0.9095 > 0.05, which means there is no evidence of lack of fit for the linear regression for the small bank.

This means that we have evidence that LOS is (linearly) related to wages.

F)

For large bank, the prediction and the confidence interval are the following:

```
       fit      lwr      upr
1 54.91688 49.43465 60.39911
```

For small bank, the prediction and the confidence interval are the following:

```
       fit      lwr      upr
1 45.87139 42.83057 48.91221
```

Based on salary, an employee with 8 years of experience be better off at a Large bank.

G)

Prediction intervals will be wider than confidence interval.

H)

Outlier from the large bank has the following: LOS = 70, Wages = 97.6801

Rstudent residual for the outlier = 4.242492

Bonferoni adjusted p-value = 0.006173458 < 0.05

I)

Large Banks:

Estimated correlation for large banks , r = 0.1870392

p-value = 0.282 > 0.05

Null hypothesis that the population correlation is zero for large banks can not be rejected with 95% certainty.

Small Banks:

Estimated correlation for small banks, r = 0.682432

p-value = 0.001712

Null hypothesis that the population correlation is zero for small banks can not be rejected with 95% certainty.

Compared to part E, we see that we obtain the same p-values for large and small banks.

J)

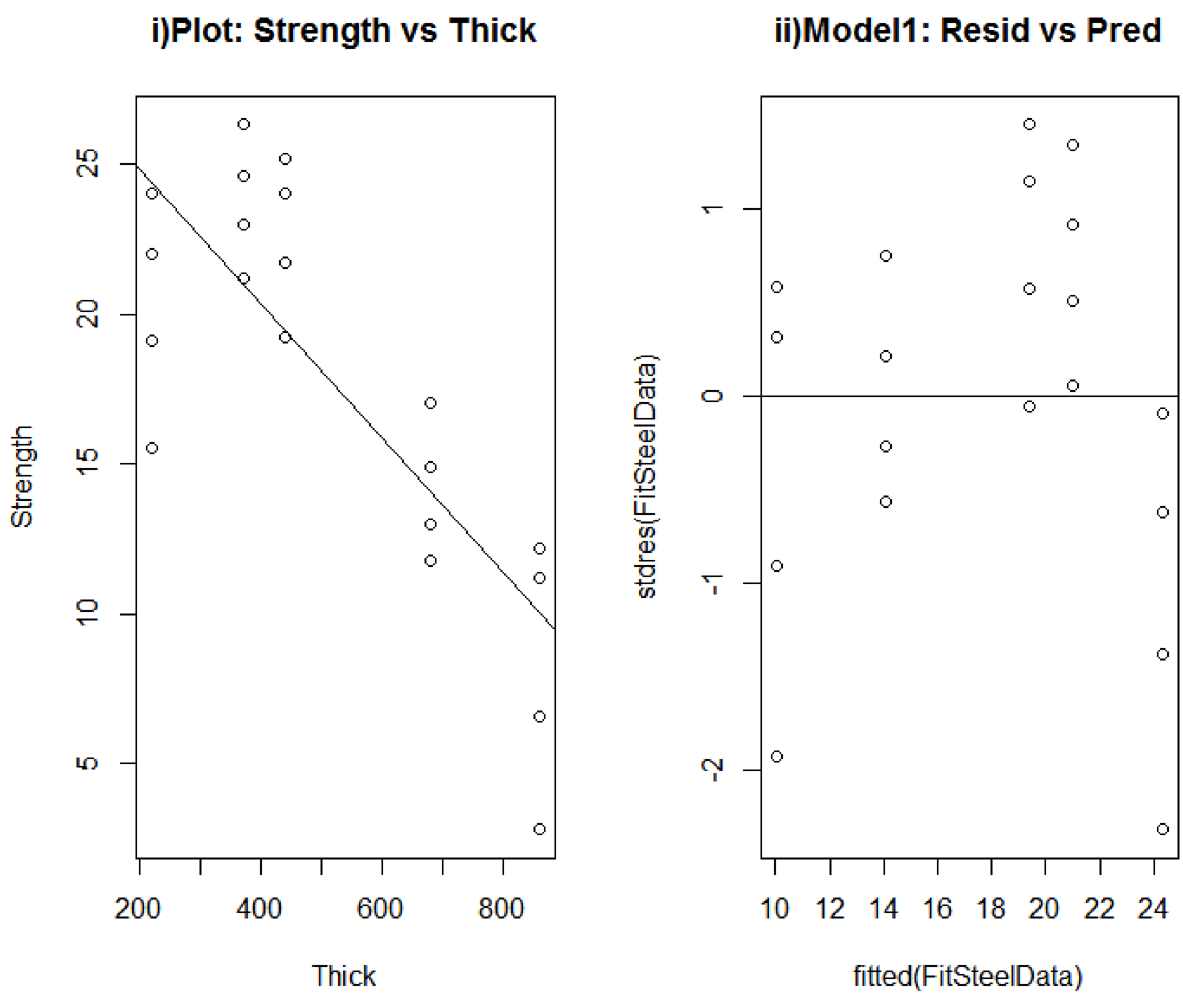
By testing variance of LOS of large bank and small bank, we get F-test p-value = 0.06389 > 0.05, which means we can not reject with 95% certainty that the null hypothesis that true ratio of variances in LOS of large and small bank is equal to 1.

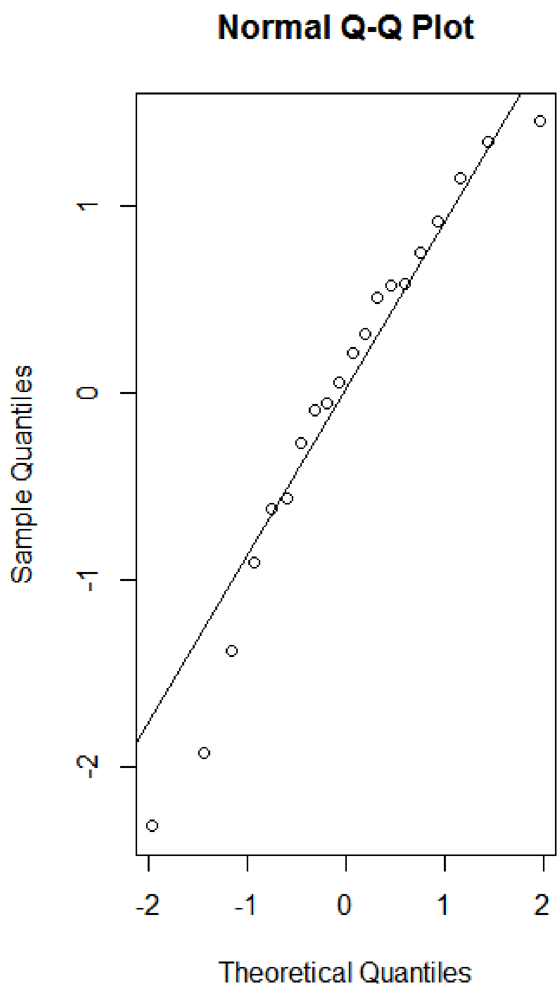
This means we operate assuming equal variances of LOS data for large and small banks.

Assuming equal variances, we run the two-sample t-test and get the p-value = 0.3915 > 0.05.

From this , we can conclude that we can not reject the assumption of equal means of LOS in larger and smaller bank with 95% certainty.

2)A)





The regression assumption of equal scatter in the plt of residuals vs fitted value does not seem to be met.

B)

Performing the F-test for lack of fit, we get the following:

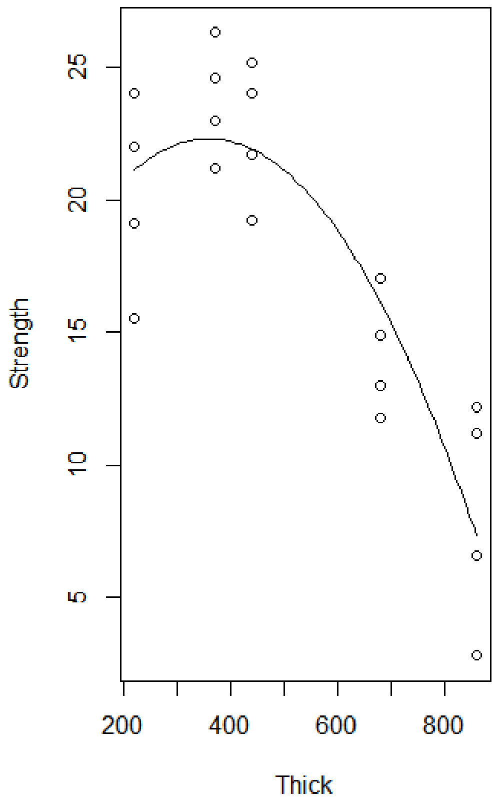
```
Analysis of Variance Table

Model 1: Strength ~ Thick
Model 2: Strength ~ as.factor(Thick)
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1     18 301.90
2     15 148.57  3    153.33 5.16 0.01195 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value = 0.01195 < 0.05 , which means we can reject the null hypothesis that the linear regression model is appropriate.

C)

:)Plot: Strength vs Thick with quadratic



Summary Table:

```
Call:
lm(formula = Strength ~ Thick + I(Thick^2), data = steelData)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6222 -2.1960  0.2443  2.4491  4.8763 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  1.452e+01  4.752e+00   3.057  0.00713 **
Thick        4.318e-02  1.980e-02   2.181  0.04354 * 
I(Thick^2)  -5.994e-05  1.786e-05  -3.357  0.00374 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.268 on 17 degrees of freedom
Multiple R-squared:  0.7796,	Adjusted R-squared:  0.7537 
F-statistic: 30.07 on 2 and 17 DF,  p-value: 2.609e-06
```

```
> #Question1
> bankSalary <- read.csv(file.choose())
> str(bankSalary)
'data.frame':	60 obs. of  3 variables:
 $ Wages: num  48.3 49 40.9 36.6 46.8 ...
 $ LOS  : int  94 48 102 20 60 78 45 39 20 65 ...
 $ Size : Factor w/ 2 levels "Large","Small": 1 2 2 2 1 2 1 1 1 2 ...
> head(bankSalary)
    Wages LOS  Size
1 48.3355  94 Large
2 49.0279  48 Small
3 40.8817 102 Small
4 36.5854  20 Small
5 46.7596  60 Large
6 59.5238  78 Small
> large <- subset(bankSalary, Size=="Large")
> small <- subset(bankSalary, Size=="Small")
> head(large)
     Wages LOS  Size
1  48.3355  94 Large
5  46.7596  60 Large
7  39.1304  45 Large
8  39.2465  39 Large
9  40.2037  20 Large
11 50.0905  76 Large
> head(small)
     Wages LOS  Size
2  49.0279  48 Small
3  40.8817 102 Small
4  36.5854  20 Small
6  59.5238  78 Small
10 38.1563  65 Small
12 46.9043  48 Small
> #1A: Create a scatterplot
> library(lattice)
> xyplot(Wages ~ LOS , data = bankSalary , groups = Size, type = c("p","r"), auto.key = l
ht"))
> #1B: Regressions
> FitLarge <- lm(Wages ~ LOS, data = large)
> summary(FitLarge)

Call:
lm(formula = Wages ~ LOS, data = large)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.688  -8.472  -3.691   5.767  44.218 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 49.54532    4.01305  12.346 6.46e-14 ***
LOS          0.05595    0.05116   1.094    0.282    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.02 on 33 degrees of freedom
Multiple R-squared:  0.03498,	Adjusted R-squared:  0.005741 
F-statistic: 1.196 on 1 and 33 DF,  p-value: 0.282

> FitSmall <- lm(Wages ~ LOS, data = small)
> summary(FitSmall)

Call:
lm(formula = Wages ~ LOS, data = small)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-15.0716  -4.4861   0.3944   2.8101  15.5273 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.87192    2.28194  15.720 8.53e-14 ***
LOS          0.10416    0.02326   4.478 0.000171 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.021 on 23 degrees of freedom
Multiple R-squared:  0.4657,	Adjusted R-squared:  0.4425 
F-statistic: 20.05 on 1 and 23 DF,  p-value: 0.0001712

> confint(FitLarge, level = 0.95)
                  2.5 %     97.5 %
(Intercept) 41.38071287 57.7099183
LOS         -0.04812646  0.1600341
> confint(FitSmall, level = 0.95)
                  2.5 %     97.5 %
(Intercept) 31.15135828 40.5924796
LOS          0.05603753  0.1522847
> anova(FitLarge)
Analysis of Variance Table

Response: Wages
          Df Sum Sq Mean Sq F value Pr(>F)
LOS        1  202.8  202.75  1.1963  0.282
Residuals 33 5592.9  169.48               
> anova(FitSmall)
Analysis of Variance Table

Response: Wages
          Df  Sum Sq Mean Sq F value    Pr(>F)    
LOS        1  988.32  988.32  20.048 0.0001712 ***
Residuals 23 1133.85   49.30                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> newdata <- data.frame(LOS = 0.0)
> predict(FitLarge, newdata, interval = "confidence", level = 0.95)
       fit      lwr      upr
1 49.54532 41.38071 57.70992
> predict(FitSmall, newdata, interval = "confidence", level = 0.95)
       fit      lwr      upr
1 35.87192 31.15136 40.59248
> ANOVAFitLarge <- lm(Wages ~ as.factor(LOS), data = large)
> ANOVAFitSmall <- lm(Wages ~ as.factor(LOS), data = small)
> #Lack of fit test for Large bank
> anova(FitLarge,ANOVAFitLarge)
Analysis of Variance Table

Model 1: Wages ~ LOS
Model 2: Wages ~ as.factor(LOS)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     33 5592.9                           
2      5  478.8 28    5114.1 1.9074 0.2442
> anova(FitSmall,ANOVAFitSmall)
Analysis of Variance Table

Model 1: Wages ~ LOS
Model 2: Wages ~ as.factor(LOS)
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
```

```
1      23 1133.85
2       3  307.54 20     826.3 0.403 0.9095
> #F LOS 96 months
> NewLOS <- data.frame(LOS = 96.0)
> predict(FitLarge, NewLOS, interval = "confidence", level = 0.95)
       fit      lwr      upr
1 54.91688 49.43465 60.39911
> predict(FitSmall, NewLOS, interval = "confidence", level = 0.95)
       fit      lwr      upr
1 45.87139 42.83057 48.91221
> #xyplot(Wages ~ LOS , data = bankSalary , groups = Size, type = c("p","r"), auto.key =
ght"))
> plot(Wages ~ LOS, data =bankSalary)
> identify(bankSalary$Wages ~ bankSalary$LOS , labels = bankSalary$Wages)
warning: nearest point already identified
warning: nearest point already identified
[1] 15
> plot(Wages ~ LOS, data =bankSalary)
> identify(bankSalary$Wages ~ bankSalary$LOS , labels = bankSalary$LOS)
warning: nearest point already identified
warning: nearest point already identified
warning: nearest point already identified
warning: nearest point already identified
warning: nearest point already identified
[1] 15
> #H residual and RStudent
> bankSalary
      Wages LOS  Size
1  48.33550  94 Large
2  49.02790  48 Small
3  40.88170 102 Small
4  36.58540  20 Small
5  46.75960  60 Large
6  59.52380  78 Small
7  39.13040  45 Large
8  39.24650  39 Large
9  40.20370  20 Large
10 38.15630  65 Small
11 50.09050  76 Large
12 46.90430  48 Small
13 43.18940  61 Small
14 60.56370  30 Large
15 97.68010  70 Large
16 48.57950 108 Large
17 67.15510  61 Large
18 38.78470  10 Small
19 51.89260  68 Large
20 51.83260  54 Large
21 64.10260  24 Large
22 54.94510 222 Small
23 43.80950  58 Large
24 43.34550  41 Small
25 61.98930 153 Large
26 40.01830  16 Small
27 50.71430  43 Small
28 48.84000  96 Large
29 34.34070  98 Large
30 80.58610 150 Large
31 33.71630 124 Small
32 60.37920  60 Large
33 48.84000   7 Large
34 38.55790  22 Small
35 39.27600  57 Large
36 47.65640  78 Large
```

```
37 44.68640  36 Large
38 44.57875  83 Small
39 65.62880  66 Large
40 33.57750  47 Small
41 41.20880  97 Small
42 67.90960 228 Small
43 43.09420  27 Large
44 40.70000  48 Small
45 40.57480   7 Large
46 39.68250  74 Small
47 50.17420 204 Large
48 54.94510  24 Large
49 32.38220  13 Small
50 51.71300  30 Large
51 55.83790  95 Large
52 54.94510 104 Large
53 70.27860  34 Large
54 57.23440 184 Small
55 54.11260 156 Small
56 39.86870  25 Large
57 27.47250  43 Small
58 67.95840  36 Large
59 44.93170  60 Small
60 51.56120 102 Large
> large_Subdata <- data.frame(large, Resid = resid(FitLarge), student = stdres(FitLarge)
student(FitLarge))
> large_Subdata[large_Subdata$LOS == 70,]
      Wages LOS  Size    Resid student RStudent
15 97.6801  70 Large 44.21802 3.44666 4.242492
> #Bonferoni Adjusted 2-sided p-value
> 2*35*(1-pt(4.242492,32))
[1] 0.006173458
> #I
> #Estimated Correlation
> large
      Wages LOS  Size
1  48.3355  94 Large
5  46.7596  60 Large
7  39.1304  45 Large
8  39.2465  39 Large
9  40.2037  20 Large
11 50.0905  76 Large
14 60.5637  30 Large
15 97.6801  70 Large
16 48.5795 108 Large
17 67.1551  61 Large
19 51.8926  68 Large
20 51.8326  54 Large
21 64.1026  24 Large
23 43.8095  58 Large
25 61.9893 153 Large
28 48.8400  96 Large
29 34.3407  98 Large
30 80.5861 150 Large
32 60.3792  60 Large
33 48.8400   7 Large
35 39.2760  57 Large
36 47.6564  78 Large
37 44.6864  36 Large
39 65.6288  66 Large
43 43.0942  27 Large
45 40.5748   7 Large
47 50.1742 204 Large
48 54.9451  24 Large
```

```
50 51.7130  30 Large
51 55.8379  95 Large
52 54.9451 104 Large
53 70.2786  34 Large
56 39.8687  25 Large
58 67.9584  36 Large
60 51.5612 102 Large
> cor.test(large$Wages, large$LOS)

	Pearson's product-moment correlation

data:  large$Wages and large$LOS
t = 1.0938, df = 33, p-value = 0.282
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1559263  0.4897590
sample estimates:
      cor
0.1870392

> small
      Wages LOS  Size
2  49.02790  48 Small
3  40.88170 102 Small
4  36.58540  20 Small
6  59.52380  78 Small
10 38.15630  65 Small
12 46.90430  48 Small
13 43.18940  61 Small
18 38.78470  10 Small
22 54.94510 222 Small
24 43.34550  41 Small
26 40.01830  16 Small
27 50.71430  43 Small
31 33.71630 124 Small
34 38.55790  22 Small
38 44.57875  83 Small
40 33.57750  47 Small
41 41.20880  97 Small
42 67.90960 228 Small
44 40.70000  48 Small
46 39.68250  74 Small
49 32.38220  13 Small
54 57.23440 184 Small
55 54.11260 156 Small
57 27.47250  43 Small
59 44.93170  60 Small
> cor.test(small$Wages, small$LOS)

	Pearson's product-moment correlation

data:  small$Wages and small$LOS
t = 4.4775, df = 23, p-value = 0.0001712
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3933745 0.8487086
sample estimates:
     cor
0.682432

> #J
> var.test(large$LOS,small$LOS)

	F test to compare two variances
```

```
data:  large$LOS and small$LOS
F = 0.50183, num df = 34, denom df = 24, p-value = 0.06389
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.230228 1.041119
sample estimates:
ratio of variances
         0.5018278

> t.test(large$LOS,small$LOS, var.equal = TRUE)

        Two Sample t-test

data:  large$LOS and small$LOS
t = -0.8634, df = 58, p-value = 0.3915
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -38.89185  15.45185
sample estimates:
mean of x mean of y
    65.60     77.32

> #2
> #i)
> library(MASS)
> steelData <- read.csv(file.choose())
> head(steelData)
  Thick Strength
1   220     24.0
2   220     22.0
3   220     19.1
4   220     15.5
5   370     26.3
6   370     24.6
> plot(Strength ~ Thick , data = steelData, main = "i)Plot: Strength vs Thick")
> FitSteelData <- lm(Strength ~ Thick , data = steelData)
> abline(coef(FitSteelData))
> #ii)
> plot(stdres(FitSteelData) ~ fitted(FitSteelData), main = "ii)Model1: Resid vs Pred")
> abline(h=0)
> #iii)
> qqnorm(stdres(FitSteelData))
> qqline(stdres(FitSteelData))
> ANOVAFitSteelData <- lm(Strength ~ as.factor(Thick), data = steelData)
> anova(FitSteelData)
Analysis of Variance Table

Response: Strength
          Df Sum Sq Mean Sq F value    Pr(>F)
Thick      1 522.04  522.04  31.125 2.699e-05 ***
Residuals 18 301.90   16.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(ANOVAFitSteelData)
Analysis of Variance Table

Response: Strength
                 Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(Thick)  4 675.37 168.843  17.047 1.881e-05 ***
Residuals        15 148.57   9.905
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(FitSteelData,ANOVAFitSteelData)
```

```
Analysis of Variance Table

Model 1: Strength ~ Thick
Model 2: Strength ~ as.factor(Thick)
  Res.Df    RSS Df Sum of Sq    F  Pr(>F)
1     18 301.90
2     15 148.57  3    153.33 5.16 0.01195 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #C) Quadratic term adding
> FitSteelData2 <- lm(Strength ~ Thick +I(Thick^2) , data = steelData)
> plot(Strength ~ Thick , data = steelData, main = "C)Plot: Strength vs Thick with quadra
> FitSteelData2$coefficients
  (Intercept)         Thick    I(Thick^2)
 1.452457e+01  4.317629e-02 -5.994113e-05
> curve((1.452457e+01) + (4.317629e-02)*x + (-5.994113e-05)*x^2 , add = TRUE)
> summary(FitSteelData2)

Call:
lm(formula = Strength ~ Thick + I(Thick^2), data = steelData)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6222 -2.1960  0.2443  2.4491  4.8763

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.452e+01  4.752e+00   3.057  0.00713 **
Thick        4.318e-02  1.980e-02   2.181  0.04354 *
I(Thick^2)  -5.994e-05  1.786e-05  -3.357  0.00374 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.268 on 17 degrees of freedom
Multiple R-squared:  0.7796,	Adjusted R-squared:  0.7537
F-statistic: 30.07 on 2 and 17 DF,  p-value: 2.609e-06

>
>
>
>
```