HW10

Yigao Li

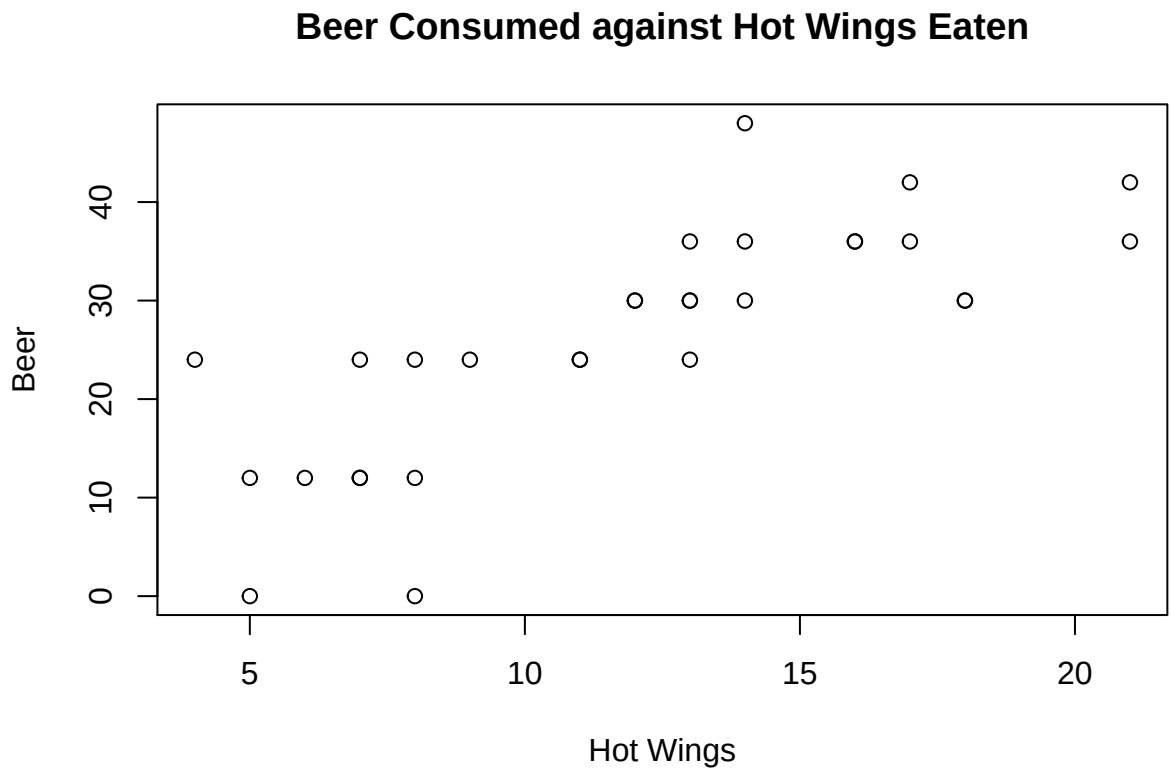
December 11, 2017

Part 1

Problem 1

(a)

```
beerwing <- read.csv("D:/Courses/ANLY 511/Beerwings.csv")
plot(beerwing$Hotwings, beerwing$Beer, xlab = "Hot Wings", ylab = "Beer",
     main = "Beer Consumed against Hot Wings Eaten")
```




```
cor(beerwing$Hotwings, beerwing$Beer)
```

```
## [1] 0.7841224
```

The correlation between beer and hot wings is 0.7841224.

(b)

```
model.1 <- lm(Beer ~ Hotwings, data = beerwing)
summary(model.1)
```

```
##
## Call:
## lm(formula = Beer ~ Hotwings, data = beerwing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.566  -4.537  -0.122   3.671  17.789
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0404     3.7235   0.817    0.421
## Hotwings      1.9408     0.2903   6.686 2.95e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.479 on 28 degrees of freedom
## Multiple R-squared:  0.6148, Adjusted R-squared:  0.6011
## F-statistic:  44.7 on 1 and 28 DF,  p-value: 2.953e-07
```

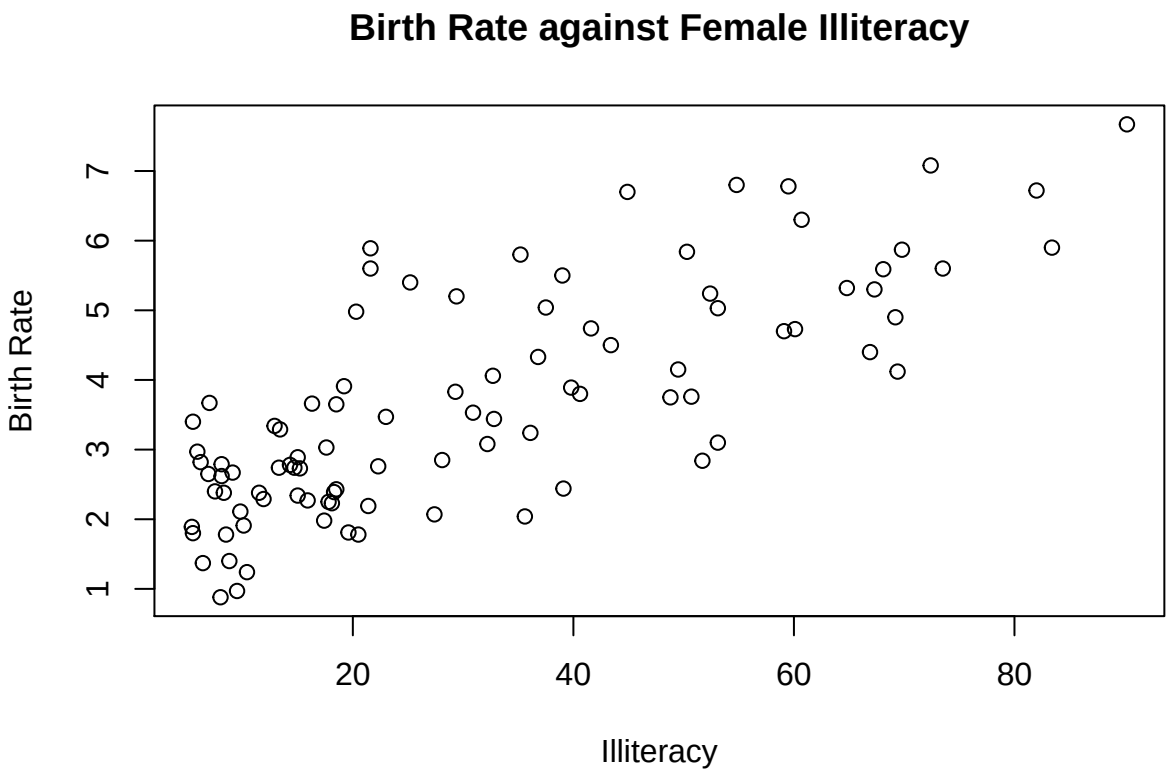
Beer $= 3.0404 + 1.9408 \times$ Hotwing $+ \epsilon$
The slope means that as you eat 1 more hot wing, the estimated average amount of beer consumed increases by 1.9408.

(c) R-squared is 0.6148. We can explain 61.48% of the variability in beer consumption by using hot wings in a regression model.

Problem 2

(a)

```
illiteracy <- read.csv("D:/Courses/ANLY 511/Illiteracy.csv")
plot(illiteracy$Illit, illiteracy$Births, xlab = "Illiteracy", ylab = "Birth Rate",
     main = "Birth Rate against Female Illiteracy")
```



Birth rate and female literacy tend to be linearly related.

(b)

```
model.2 <- lm(Births ~ Illit, data = illiteracy)
summary(model.2)
```

```
##
## Call:
## lm(formula = Births ~ Illit, data = illiteracy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92762 -0.62924 -0.08767  0.53068  2.76355
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.94874    0.18227   10.69   <2e-16 ***
## Illit        0.05452    0.00473   11.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 92 degrees of freedom
## Multiple R-squared:  0.5908, Adjusted R-squared:  0.5864
## F-statistic: 132.9 on 1 and 92 DF,  p-value: < 2.2e-16
```

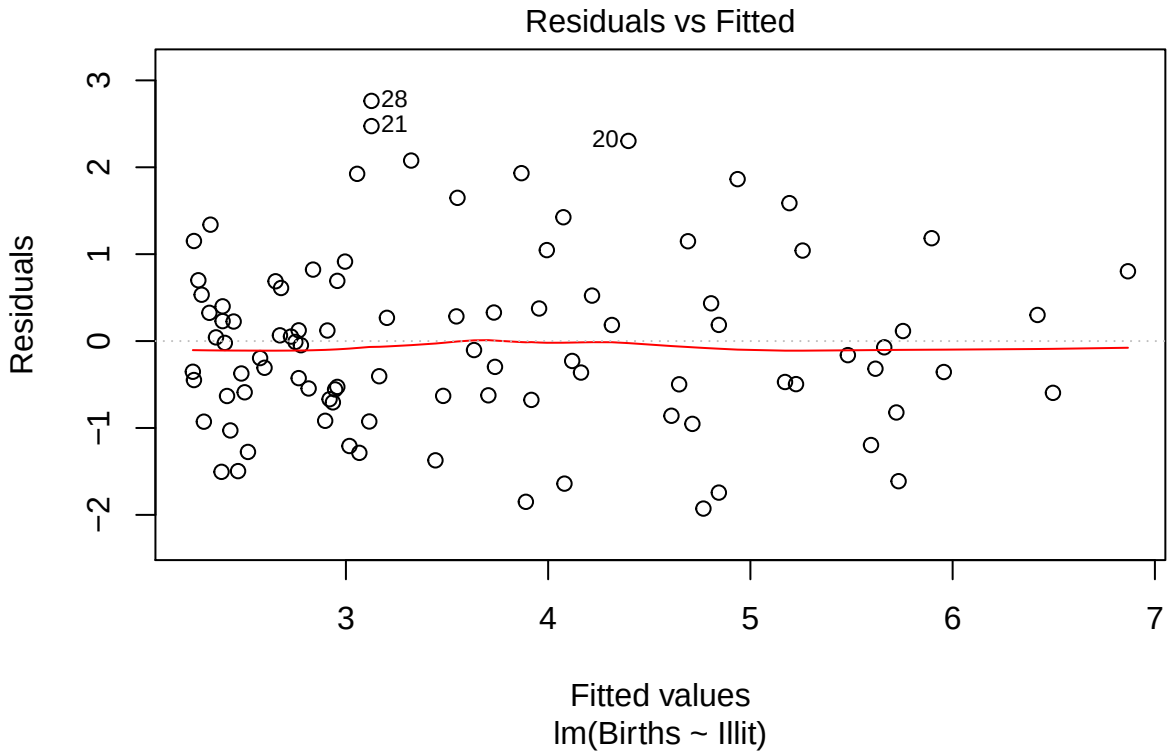
$\text{Births } = 1.94874 + 0.05452 \times \text{ Illit } + \epsilon$

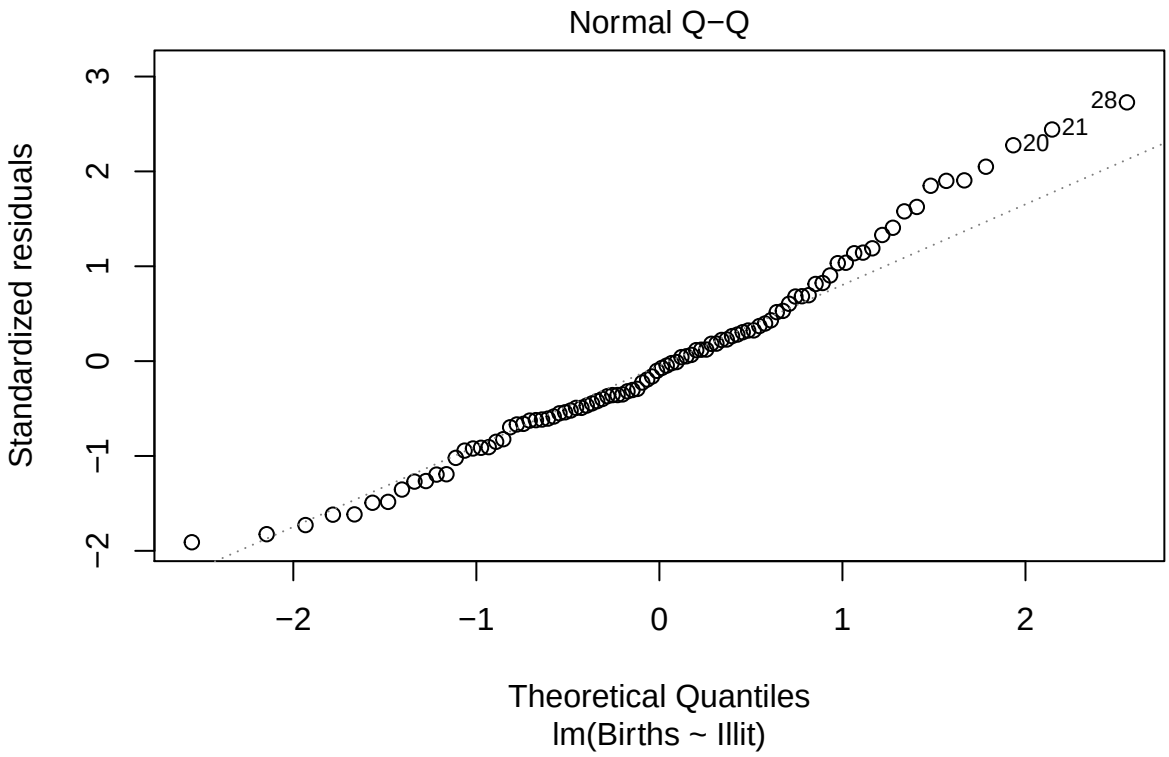
The slope means that as female illiteracy goes up by 1%, the estimated average birth rate goes up by 0.05452%.

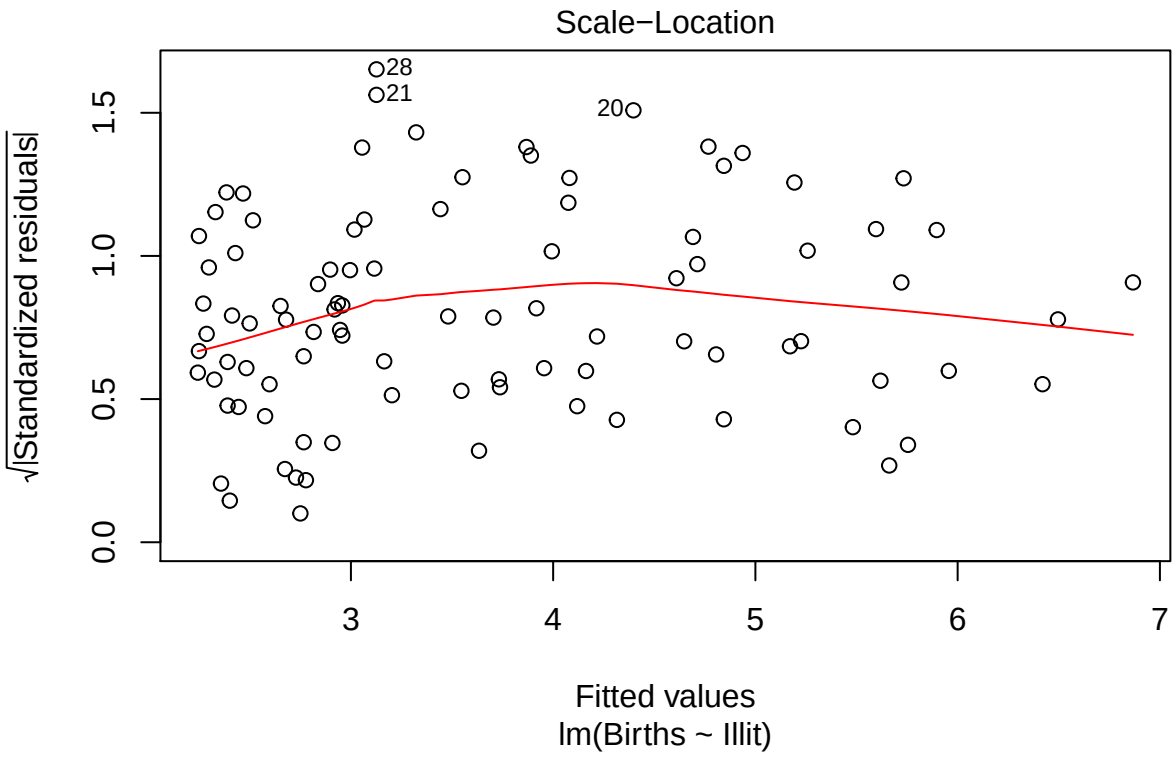
R-squared is 0.5908. We can explain 59.08% of the variability in birth rate by using female illiteracy in a regression model.

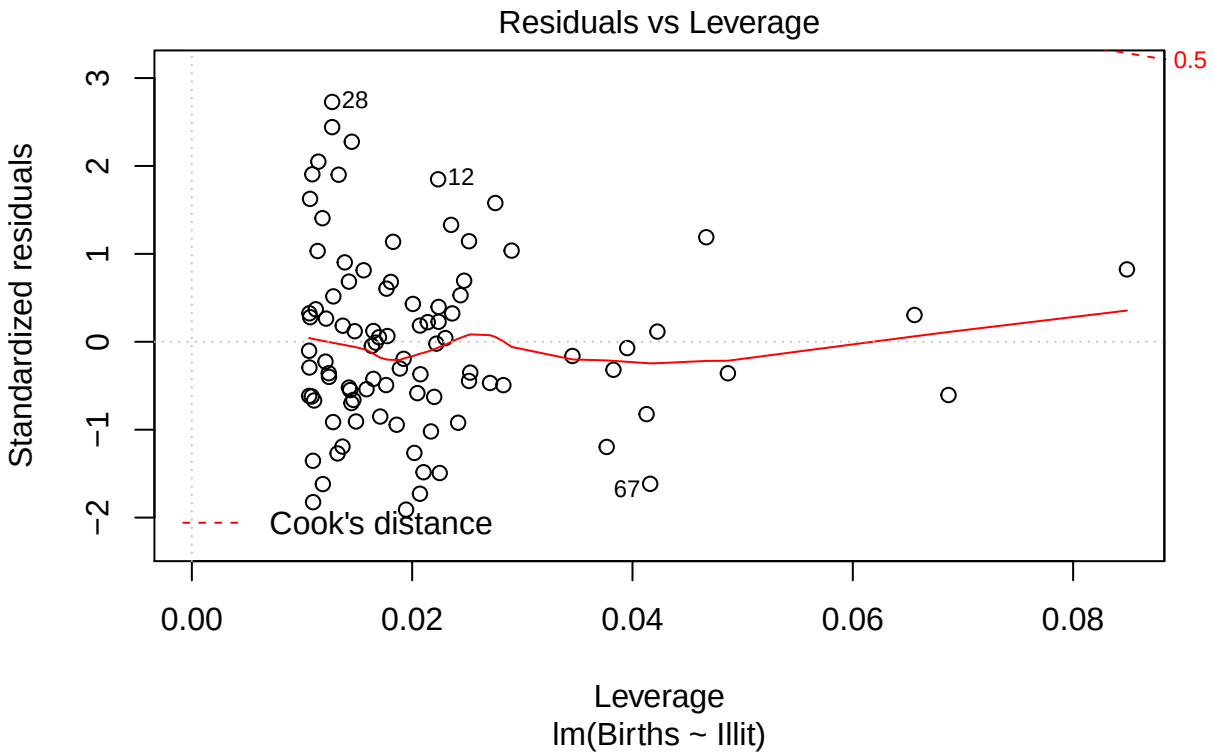
(c)

```
plot(model.2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Births ~ Illit)
28
21
20
−2
−1
0
1
2
3


Scale−Location
28
21
20
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
4
5
6
7
Fitted values
lm(Births ~ Illit)



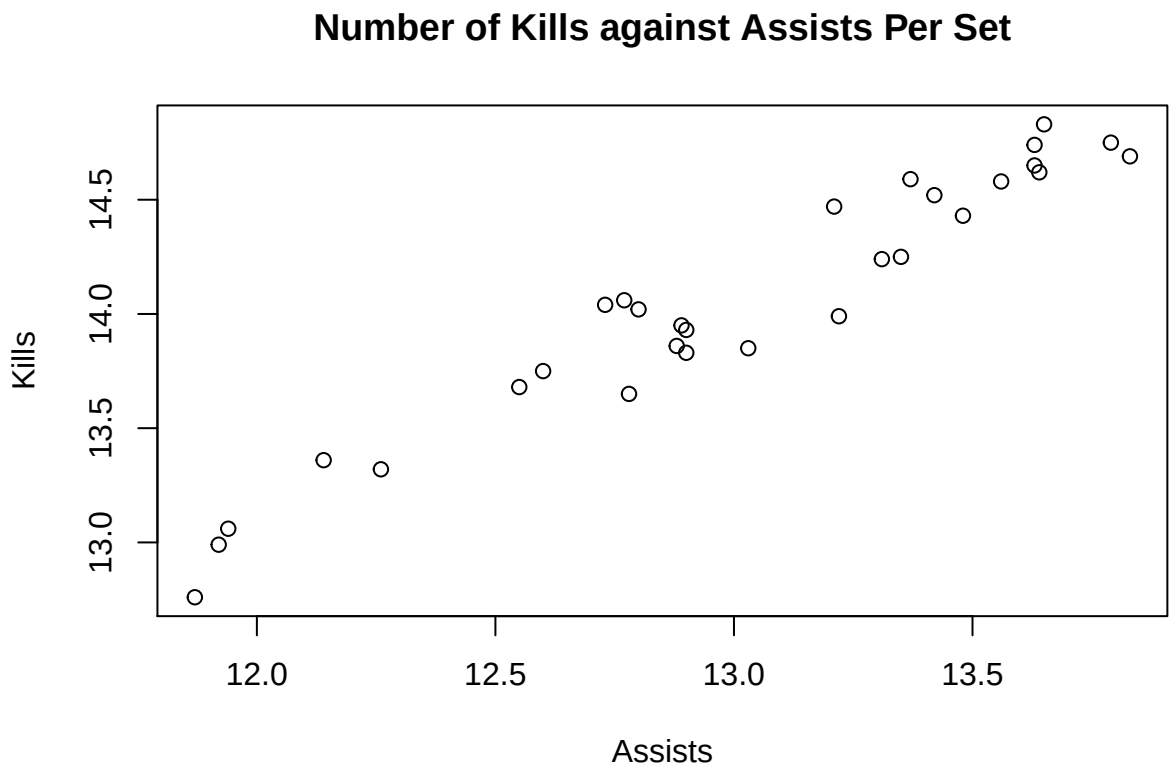
From the Residuals vs Fitted plot, we can see that the linear model is appropriate.

(d) We can conclude that improving literacy will decrease birth rate because p-value for slope coefficient is very small.

Problem 3

(a)

```
volleyball <- read.csv("D:/Courses/ANLY 511/Volleyball2009.csv")
plot(volleyball$Assts, volleyball$Kills, xlab = "Assists", ylab = "Kills",
     main = "Number of Kills against Assists Per Set")
```



The number of kills increases as the number of assists goes up.

(b)

```
model.3 <- lm(Kills ~ Assts, data = volleyball)
summary(model.3)
```

```
##
## Call:
## lm(formula = Kills ~ Assts, data = volleyball)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.265426 -0.093853 -0.009895  0.092659  0.248598
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.73626    0.60523   2.869  0.00775 **
## Assts        0.94699    0.04651  20.362  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.144 on 28 degrees of freedom
## Multiple R-squared:  0.9367, Adjusted R-squared:  0.9345
## F-statistic: 414.6 on 1 and 28 DF,  p-value: < 2.2e-16
```

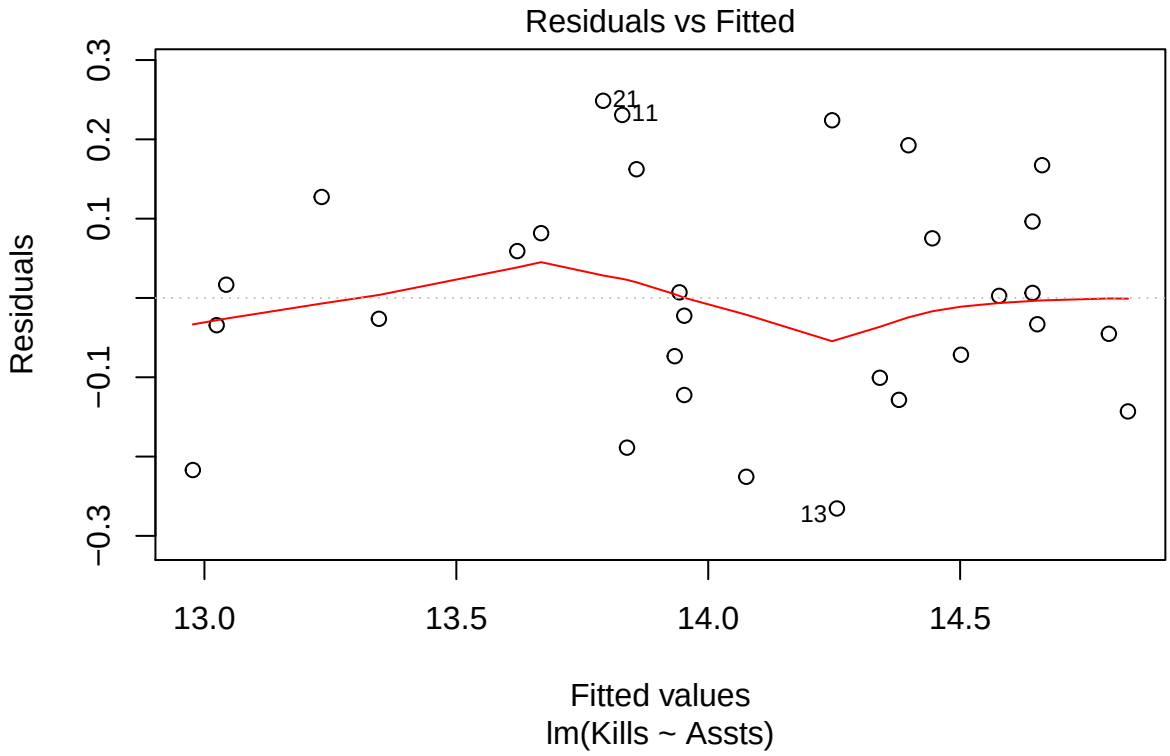
$\text{Kills } = 1.73626 + 0.94699 \times \text{ Assts } + \epsilon$

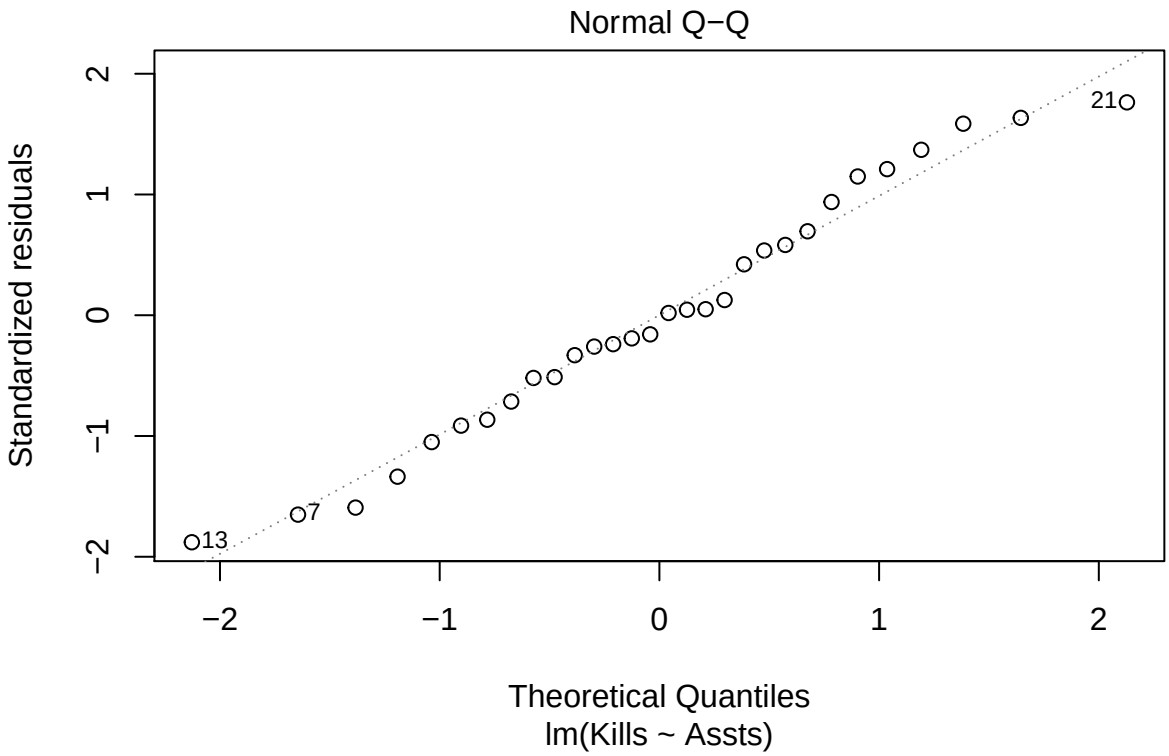
The slope means that as the number of assists goes up by 1, the estimated number of kills goes up by 0.94699

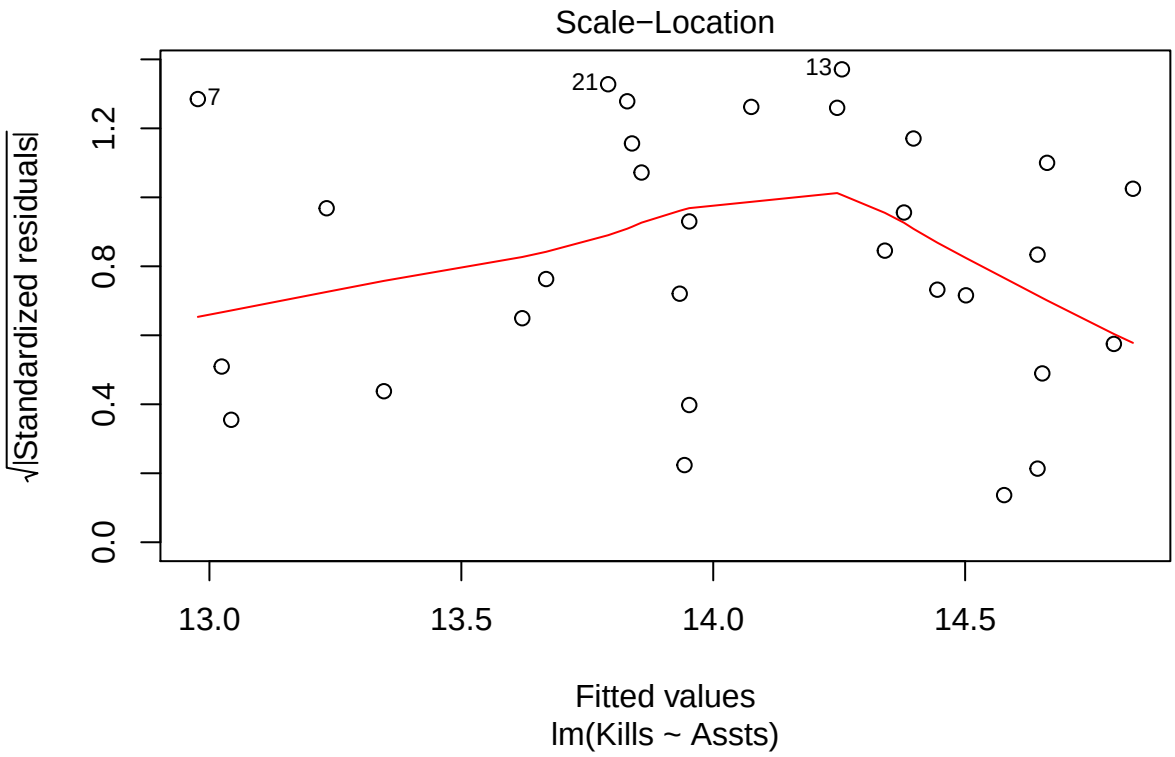
in average.
R-squared is 0.9367. We can explain 93.67% of the variability in birth rate by using female illiteracy in a regression model.

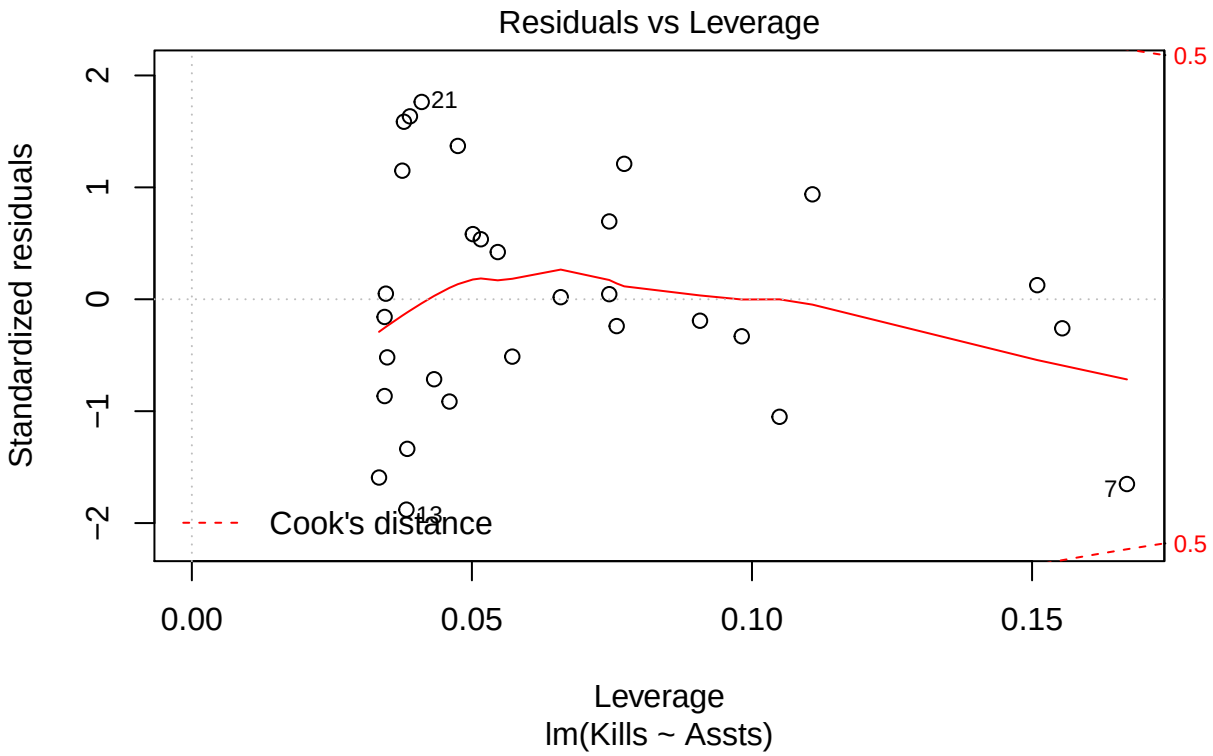
(c)

```
plot(model.3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Kills ~ Assts)
13
7
21
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
7
21
13
13.0
13.5
14.0
14.5
Fitted values
lm(Kills ~ Assts)

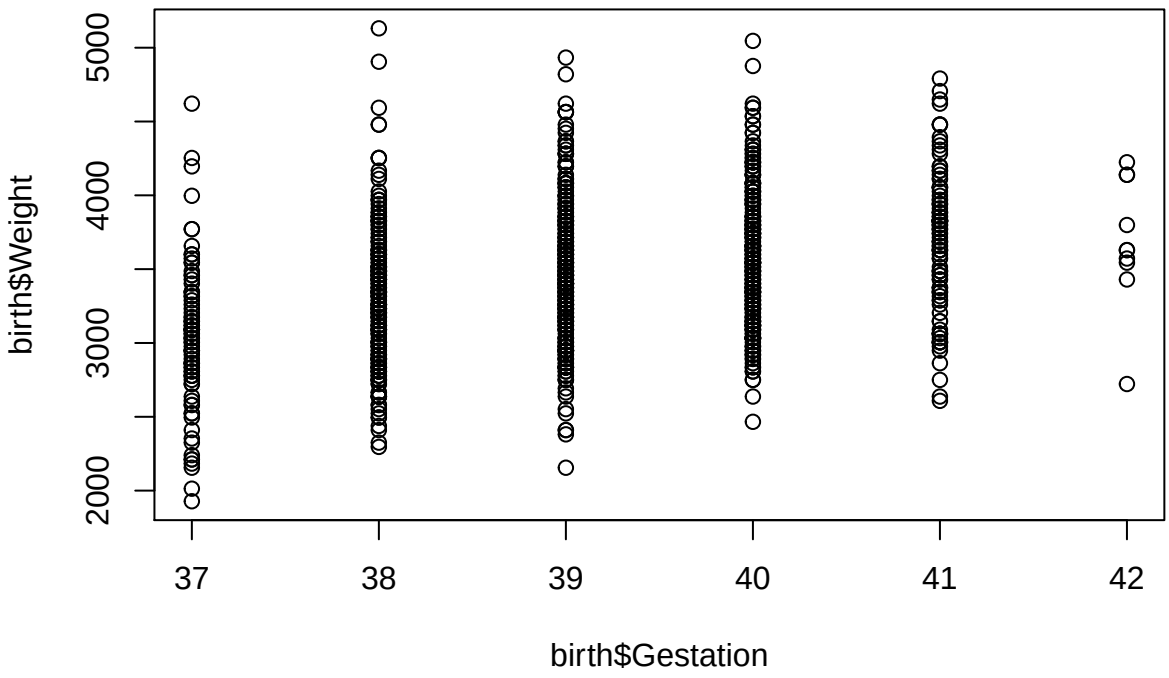


This linear model is appropriate.

Problem 4

(a)

```
birth <- read.csv("D:/Courses/ANLY 511/NCBirths2004.csv")
plot(birth$Gestation, birth$Weight)
```




```
cor(birth$Gestation, birth$Weight)
```

```
## [1] 0.3486057
```

The correlation is 0.3486057.

(b)

```
model.4 <- lm(Weight ~ Gestation, data = birth)
summary(model.4)
```

```
##
## Call:
## lm(formula = Weight ~ Gestation, data = birth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1276.13  -312.13   -22.13   267.88  1848.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2379.69     493.99  -4.817 1.68e-06 ***
## Gestation     149.00      12.62  11.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 457.4 on 1007 degrees of freedom
## Multiple R-squared:  0.1215, Adjusted R-squared:  0.1207
```

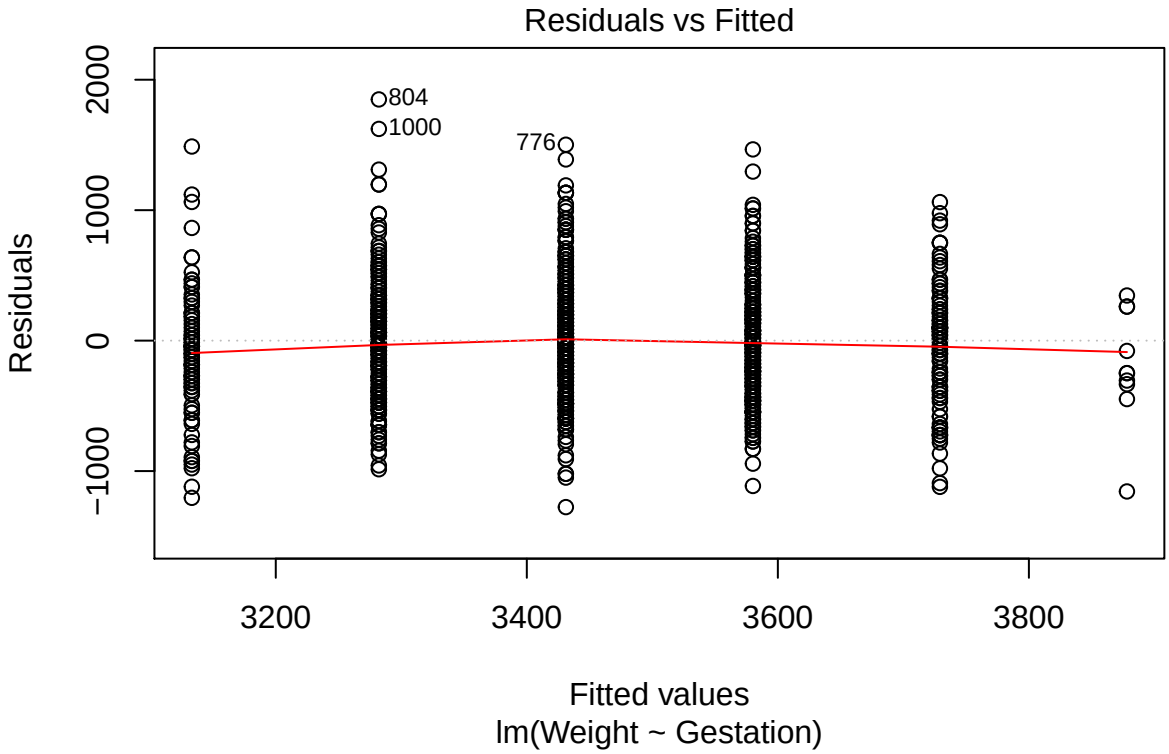
```
## F-statistic: 139.3 on 1 and 1007 DF,  p-value: < 2.2e-16
```

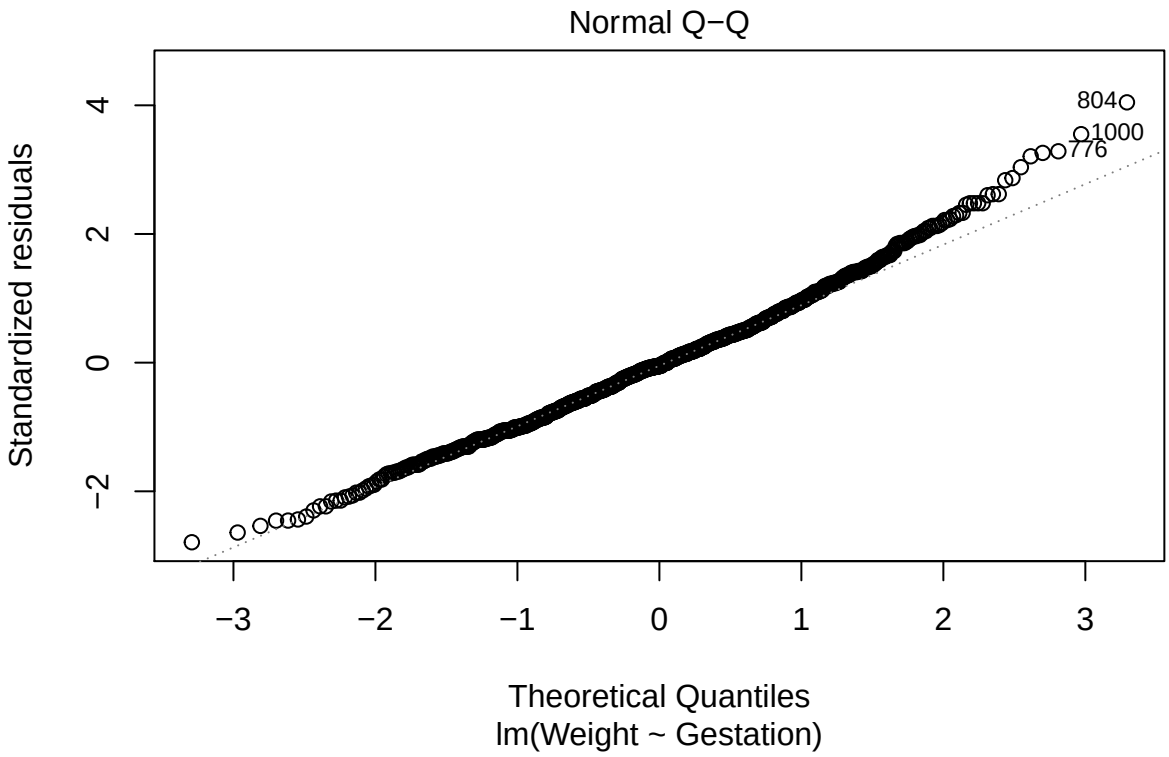
Weight $= -2379.69 + 149 \times$ Gestation $+ \epsilon$

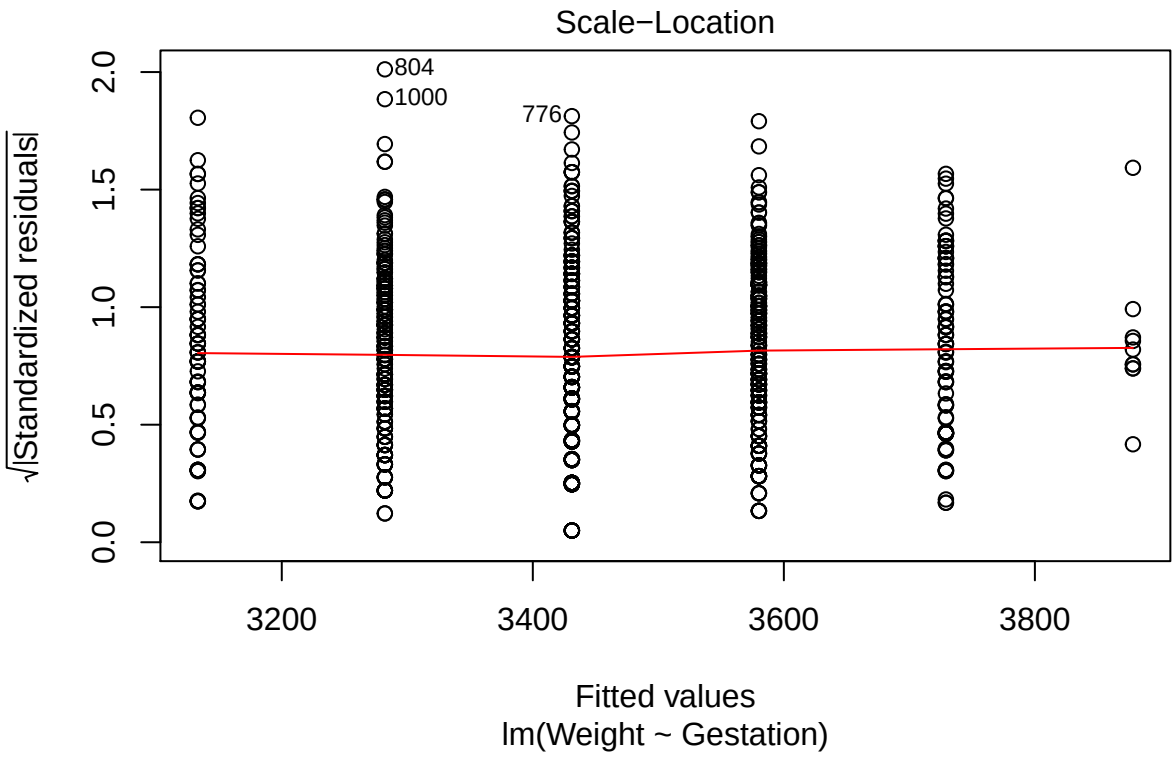
(c) The slope means that as gestation period goes up by 1, the estimated weight goes up by 149 in average. R-squared is 0.1215. We can explain 12.15% of the variability in birth rate by using female illiteracy in a regression model.

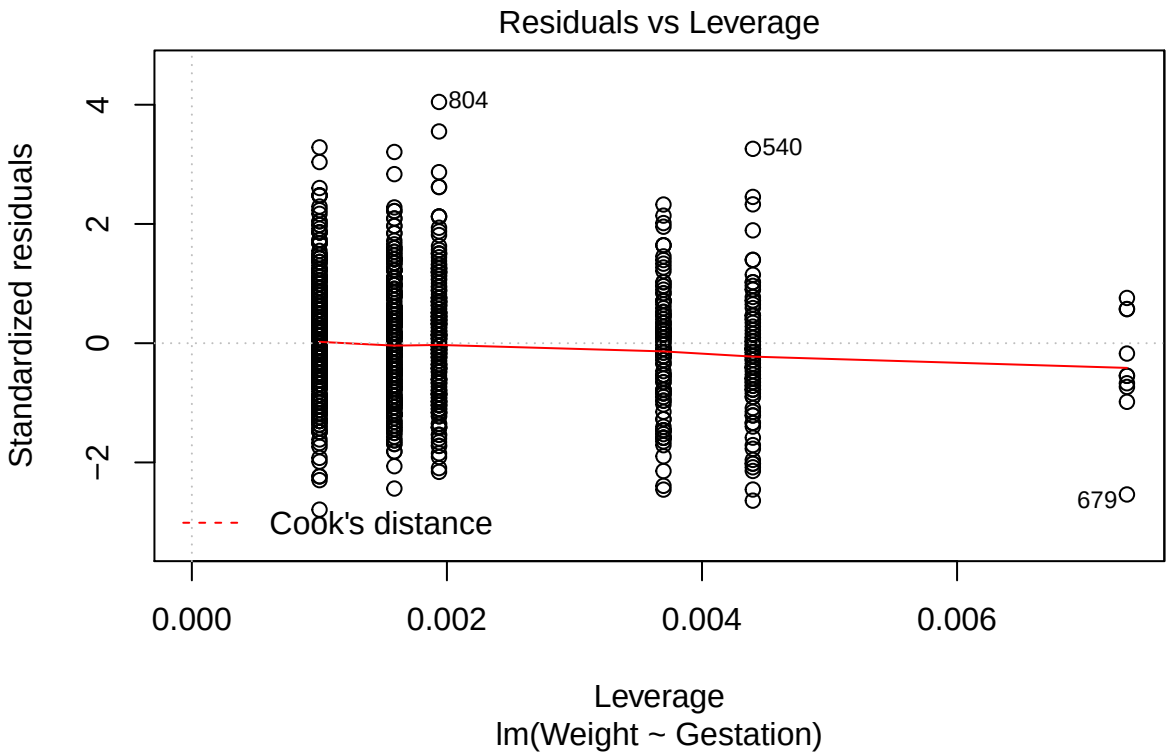
(d)

```
plot(model.4)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
804
1000
776
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Weight ~ Gestation)


Scale−Location
804
1000
776
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
3200
3400
3600
3800
Fitted values
lm(Weight ~ Gestation)



Equal variance may be a problem because every group of points has different variance.

Problem 5

(a)

$$10 \times 1.45 = 14.5$$

Test score will increase 14.5 points.

(b)

$$\begin{aligned} SD(Score) &= \sqrt{Var(Score)} \\ &= \sqrt{Var(502.7 + 1.45Hours)} \\ &= \sqrt{1.45^2 Var(Hours)} \\ &= \sqrt{1.45^2 (SD(Hours))^2} \\ &= 1.45 \times SD(Hours) \\ &= 39.875 \end{aligned}$$

(c)

```
n <- 100
beta.1_hat <- 1.45
SE <- beta.1_hat/sqrt(n-2)
lower.bound <- beta.1_hat - qt(0.975, n-2) * SE
upper.bound <- beta.1_hat + qt(0.975, n-2) * SE
cat("95% confidence interval for the true slope is [", lower.bound, ",", upper.bound, "].")
```

```
## 95% confidence interval for the true slope is [ 1.159331 , 1.740669 ].
```

(d)

```
x_bar <- 55
x_star <- 50
SSE <- 16.54
SDx <- 27.5
SE_mean_Y <- SSE*sqrt(1/n + (x_star - x_bar)^2/SDx)
lower.bound <- 502.7+1.45*x_star - qt(0.975, n-2) * SE_mean_Y
upper.bound <- 502.7+1.45*x_star + qt(0.975, n-2) * SE_mean_Y
cat("95% confidence interval for the mean score when tutored 50 h is [", lower.bound, ",", upper.bound,
```

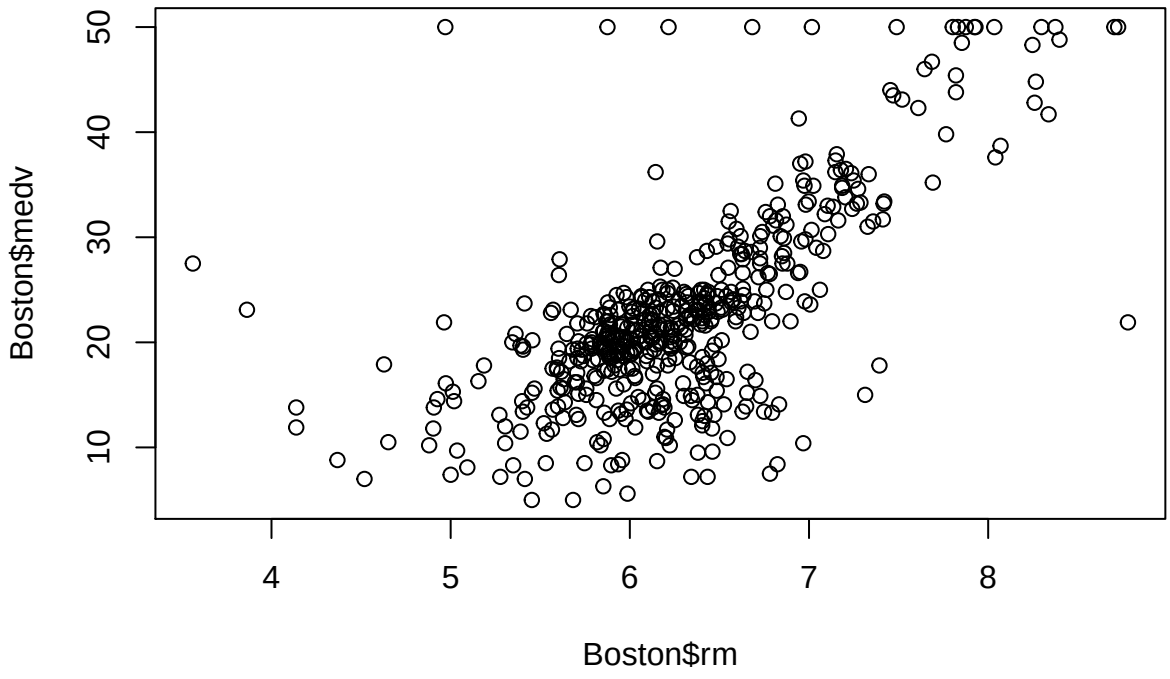
```
## 95% confidence interval for the mean score when tutored 50 h is [ 543.7328 , 606.6672 ].
```

Part 2

Single linear model

(a)

```
library(MASS)
data(Boston)
plot(Boston$rm, Boston$medv)
```



(b) The slope should be positive.

It will not pass 0 because both variable "medv" is the median value of owner-occupied homes in $1000s. It cannot be negative.

(c)

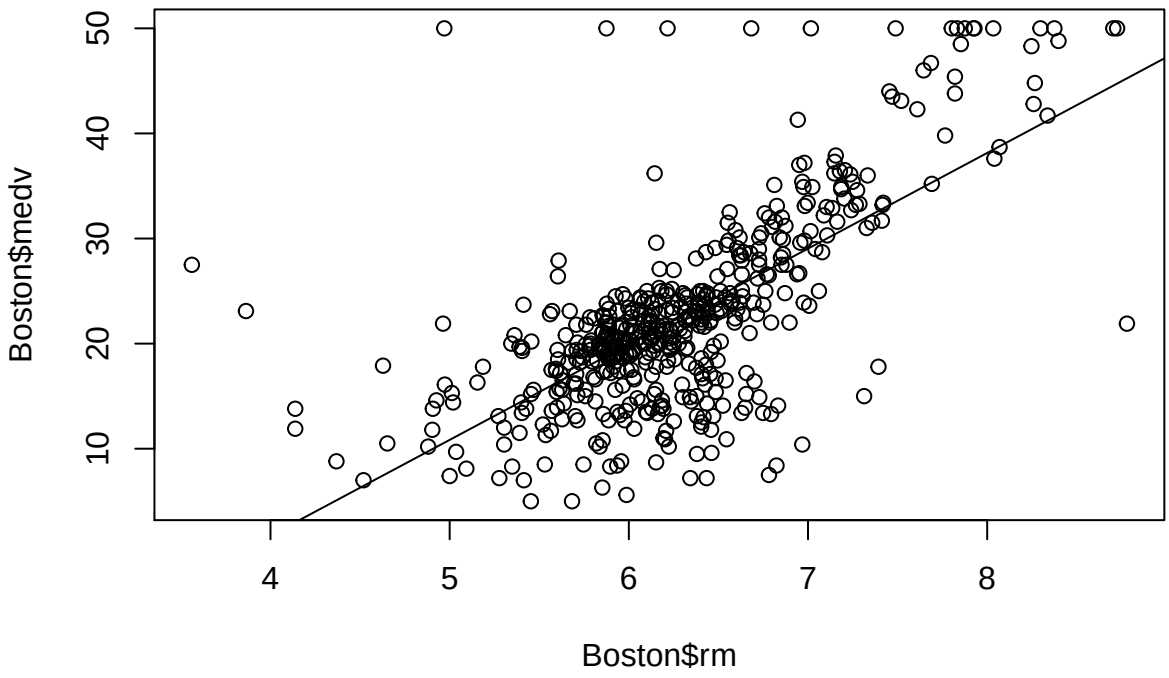
i.

```
medv_model<- lm(medv~rm,data=Boston)
summary(medv_model)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

ii.

```
plot(Boston$rm, Boston$medv)
abline(coef = c(-34.671, 9.102))
```



iii.
Slope means that as average number of rooms per dwelling increases by 1, the estimated median value of owner-occupied homes increases $\$9,102$.
Intercepts means that when there is no room per dwelling. The estimated median value of owner-occupied homes is $\$-34,671$, which is in this case is meaningless.

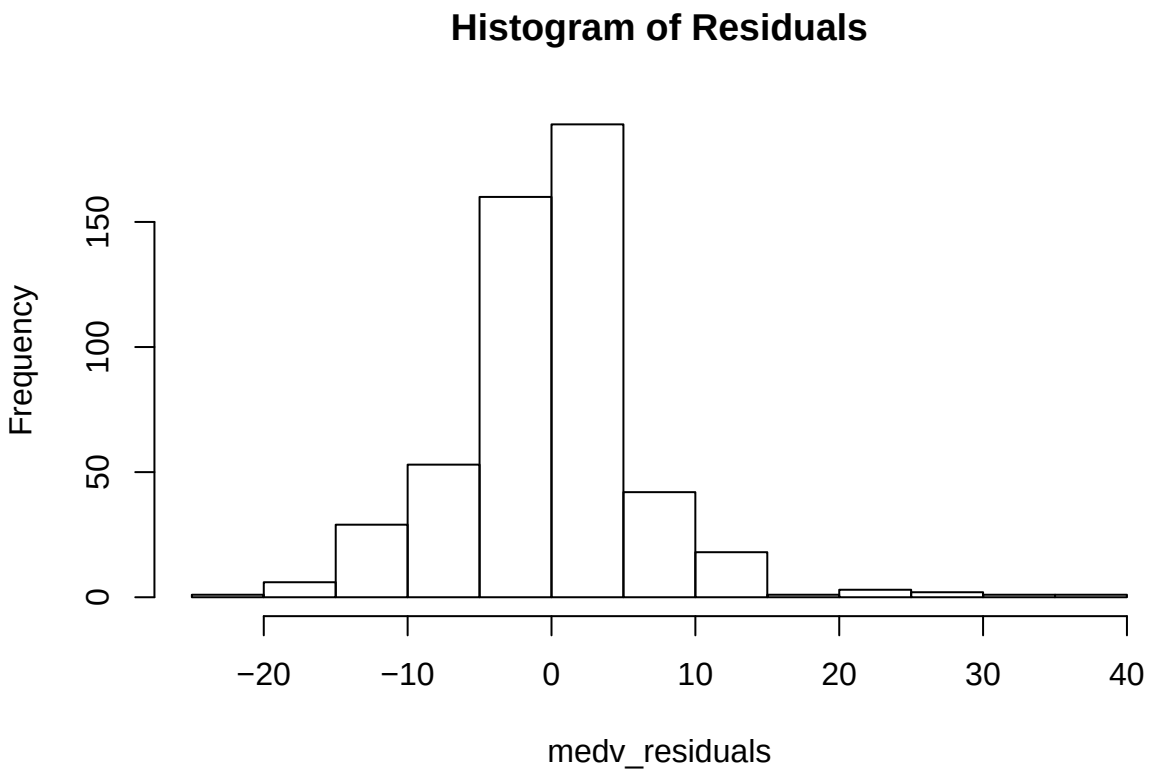
(d)

i.

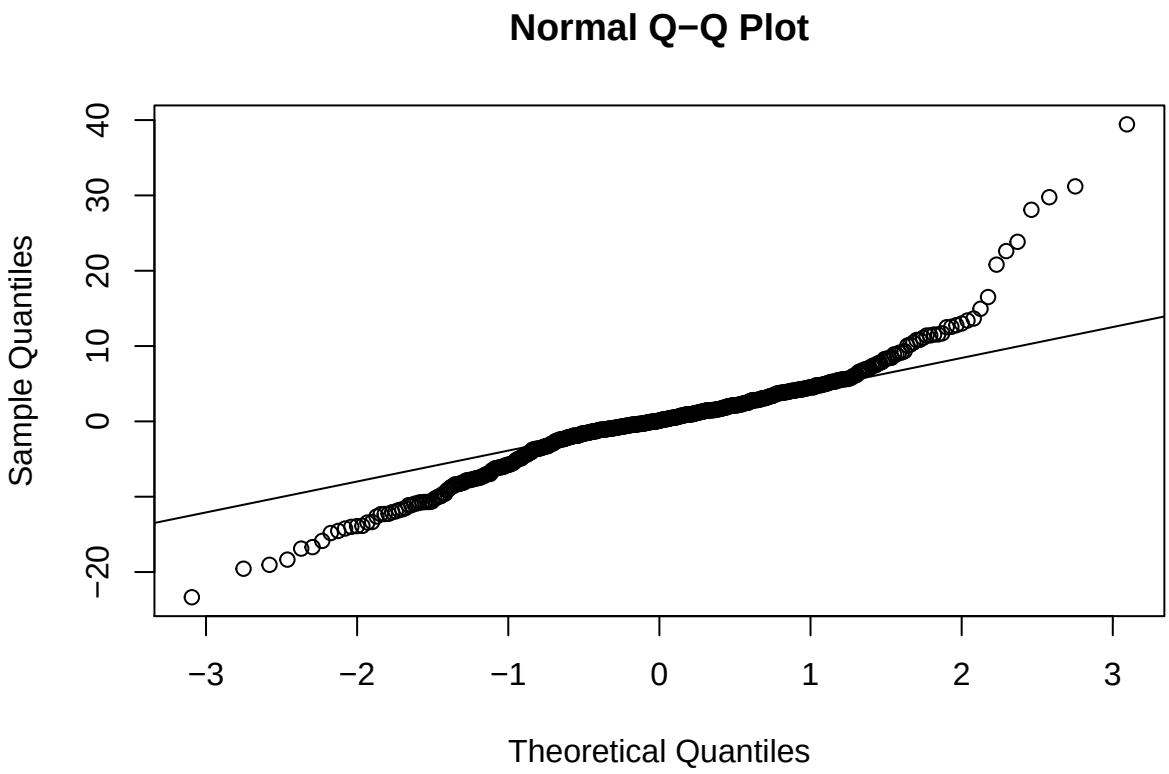
```
medv_residuals <- residuals(medv_model)
```

ii.

```
hist(medv_residuals, main = "Histogram of Residuals")
```




```
qqnorm(medv_residuals)
qqline(medv_residuals)
```



iii.
Residuals is normally distributed within ± 1 standard deviation.

(e) Summary information is shown in part(c)i.

i. P-value for intercept and slope are both less than 2×10^{-16}, and they are statistically significant.

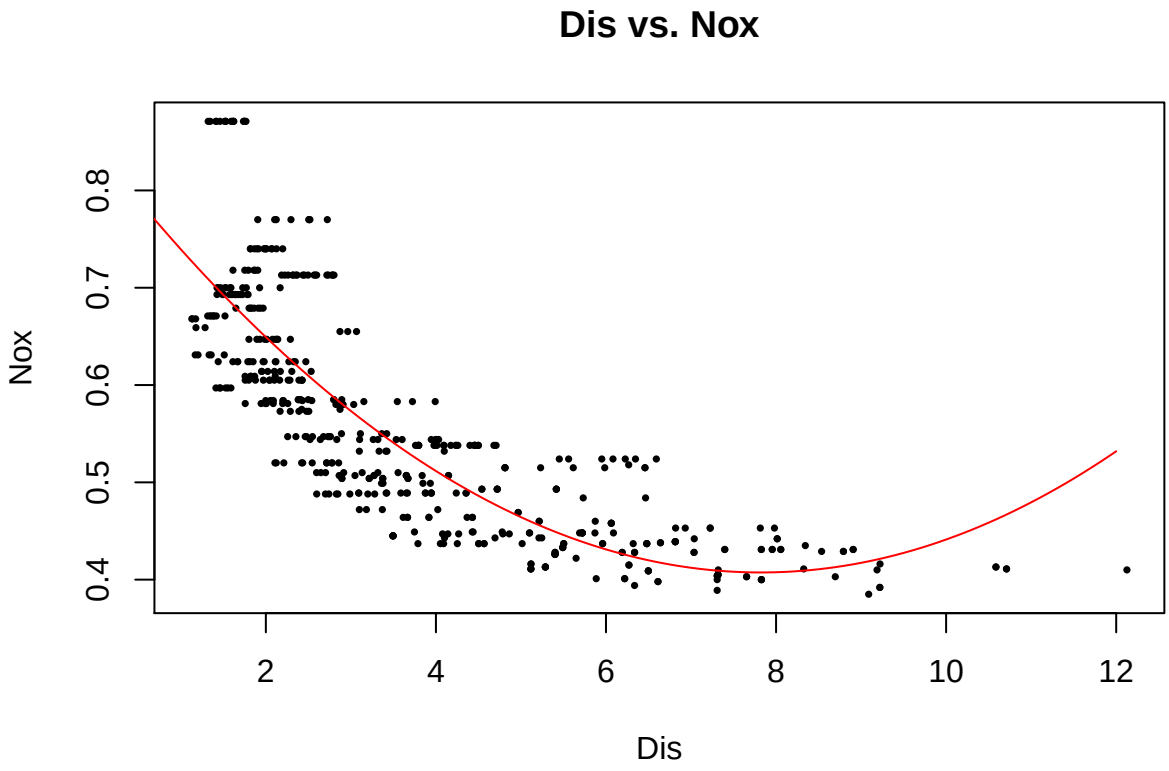
ii. Multiple R-squared is 0.4835. We can explain 48.35% of the variability in median value of owner-occupied homes by using average number of rooms per dwelling in a regression model.

iii. F-statistics is 471.8 with degrees of freedom 1 and 504. P-value is also less than 2×10^{-16}. But it does not contradict with Multiple R-squared, because r^2 means the variance of data points is high around the regression line.

Polynomial Regression

(a)

```
plot(Boston$dis, Boston$nox, main = "Dis vs. Nox",
     xlab='Dis', ylab='Nox', pch=16, cex=0.5)
m2 <- lm(nox ~ poly(dis,2), data=Boston)
newdata <- data.frame(dis=seq(0,12,by=0.1))
predicted_value <- predict(m2, newdata = newdata)
points(unlist(newdata),predicted_value, col='red', type='l')
```




```
summary(m2)
```

```
##
## Call:
## lm(formula = nox ~ poly(dis, 2), data = Boston)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.129559 -0.044514 -0.007753  0.025778  0.201882
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.554695   0.002828  196.16   <2e-16 ***
## poly(dis, 2)1 -2.003096   0.063610  -31.49   <2e-16 ***
## poly(dis, 2)2  0.856330   0.063610   13.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06361 on 503 degrees of freedom
## Multiple R-squared:  0.6999, Adjusted R-squared:  0.6987
## F-statistic: 586.4 on 2 and 503 DF,  p-value: < 2.2e-16
```

It is not linear. But it looks like negative logarithm function.

(b)

i. It looks like quadratic when Dis is between 2 and 10.

ii. Intercept: When weighted mean distance to five Boston employment centers is 0, the nitrogen oxides concentration is 0.554695 parts per 10 million.
There's no interpretation for other 2 coefficients due to quadratic equation.

iii. P-values for all parameters are less than 2×10^{-16}. They are statistically significant.

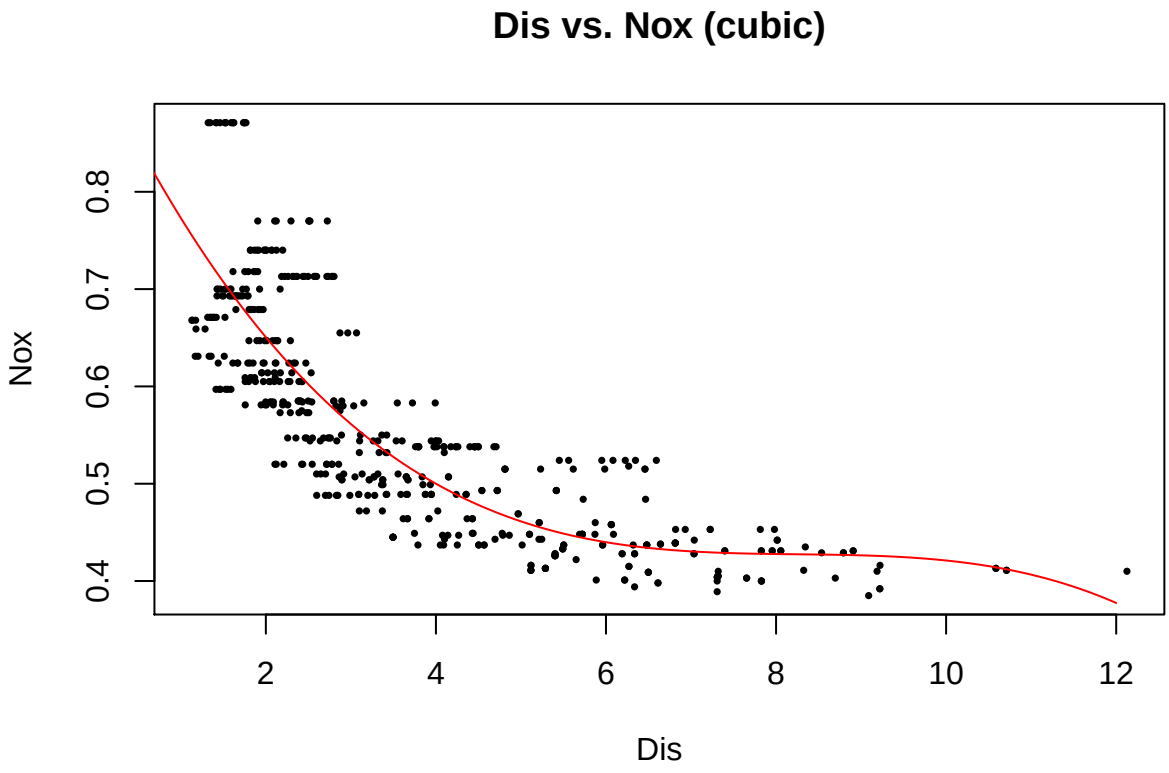
iv. Multiple R-squared is 0.6999 and F statistics is 586.4 with 2 and 503 degrees of freedom.

v. Like I said, this only looks like quadratic when dis is below 10. When dis is greater than 12, this model fails.

(c)

i.

```
plot(Boston$dis, Boston$nox, main = "Dis vs. Nox (cubic)",
     xlab='Dis', ylab='Nox', pch=16, cex=0.5)
m3 <- lm(nox ~ poly(dis,3), data=Boston)
newdata <- data.frame(dis=seq(0,12,by=0.1))
predicted_value <- predict(m3, newdata = newdata)
points(unlist(newdata),predicted_value, col='red', type='l')
```



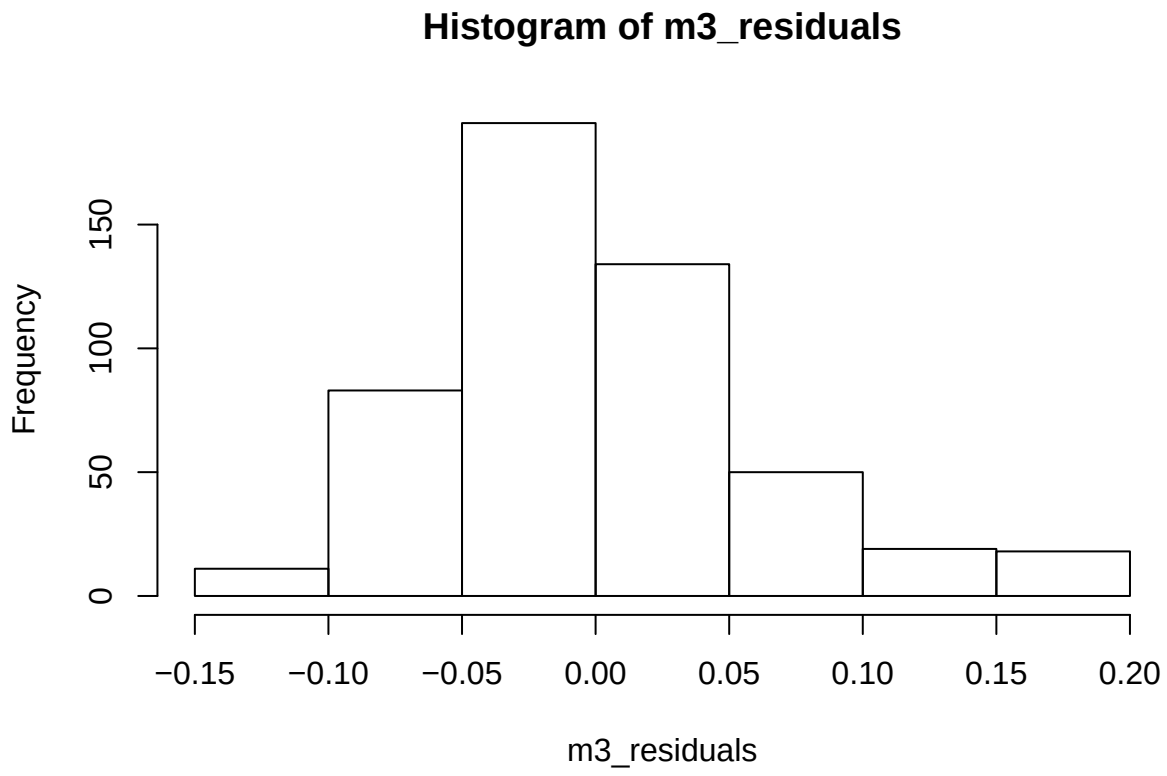

```
summary(m3)
```

```
##
## Call:
## lm(formula = nox ~ poly(dis, 3), data = Boston)
##
```

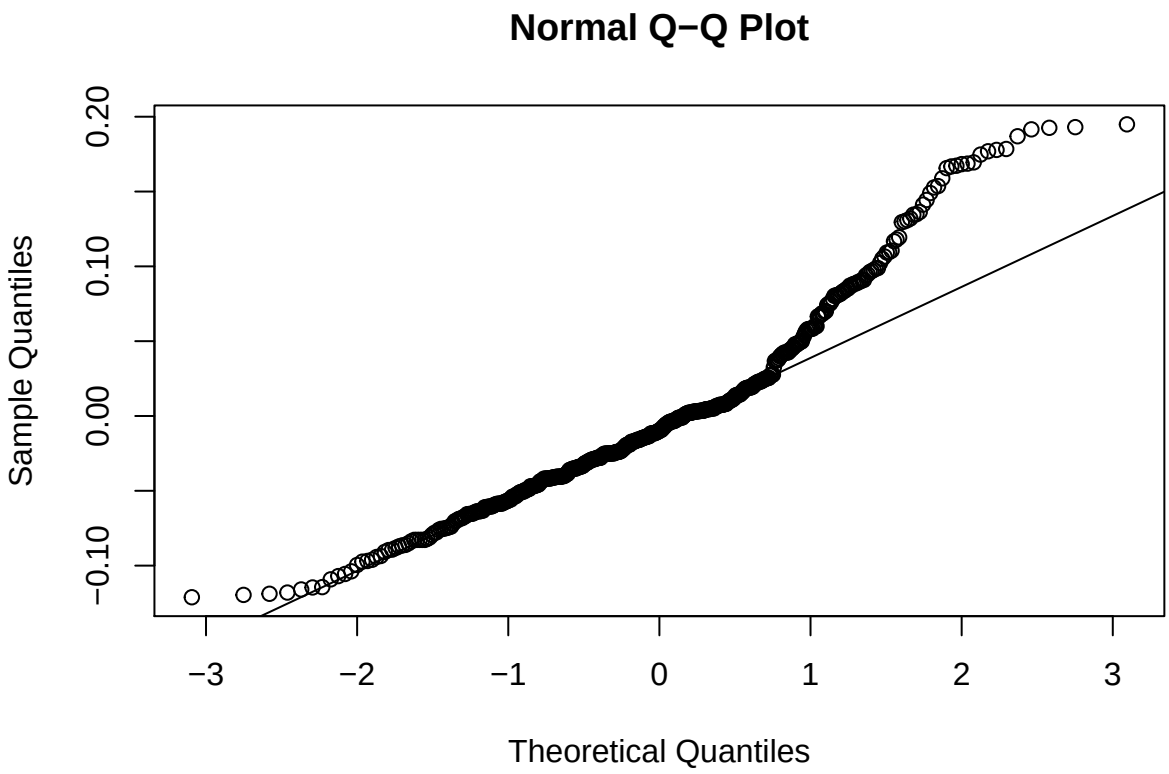
```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.121130 -0.040619 -0.009738  0.023385  0.194904
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.554695   0.002759 201.021  < 2e-16 ***
## poly(dis, 3)1 -2.003096   0.062071 -32.271  < 2e-16 ***
## poly(dis, 3)2  0.856330   0.062071  13.796  < 2e-16 ***
## poly(dis, 3)3 -0.318049   0.062071  -5.124 4.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06207 on 502 degrees of freedom
## Multiple R-squared:  0.7148, Adjusted R-squared:  0.7131
## F-statistic: 419.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

Based on Multiple R-squared, cubic regression is better.
ii.

```
m3_residuals <- residuals(m3)
hist(m3_residuals)
```



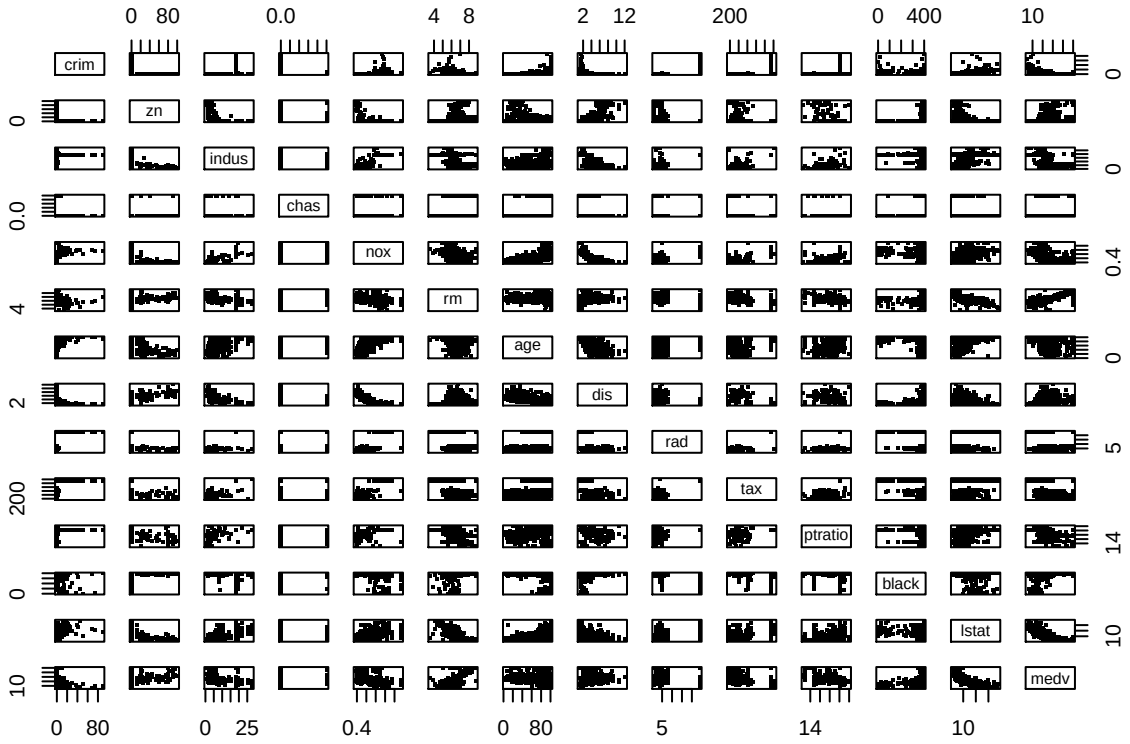

```
qqnorm(m3_residuals)
qqline(m3_residuals)
```



Multiple regression

(a)

```
pairs(Boston, cex=0.4, pch=15)
```



"nox" and "lstat" look correlated with "age".

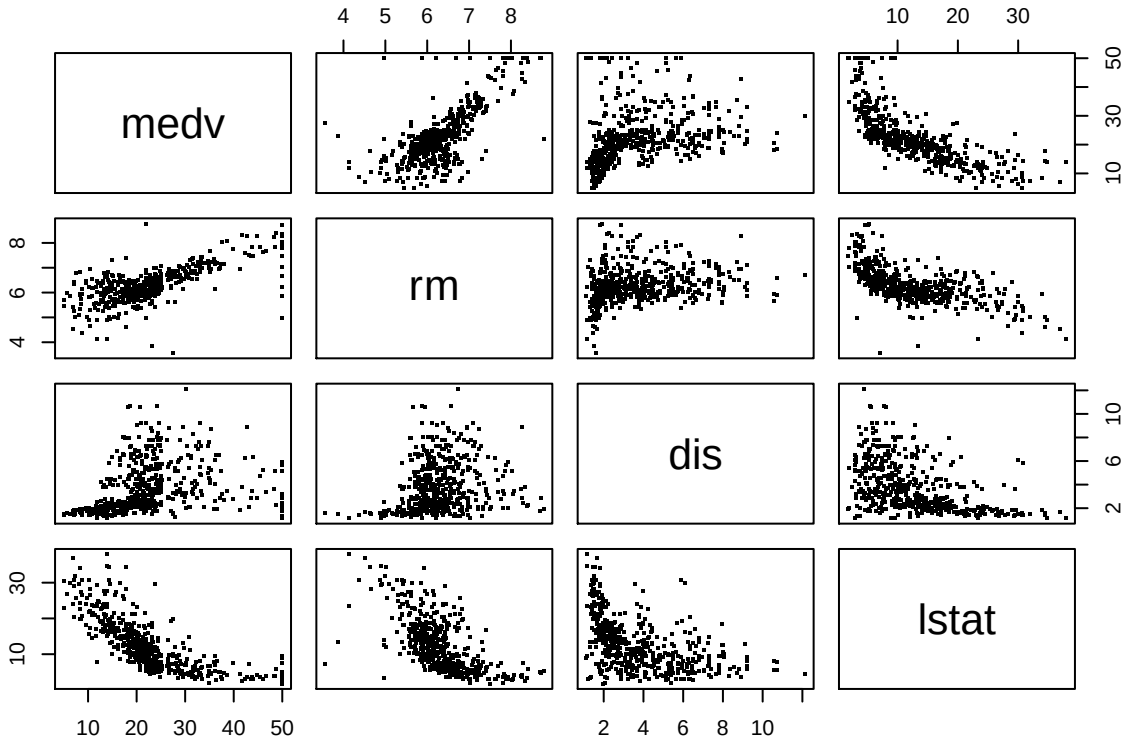
(b)

i.

```
cor(Boston$rm, Boston$lstat)
```

```
## [1] -0.6138083
```

```
medv_subset <- Boston[,c("medv", "rm", "dis", "lstat")]
pairs(medv_subset, cex=0.4, pch=15)
```




```
medv_model<- lm(medv~rm+lstat+dis, data=medv_subset)
summary(medv_model)
```

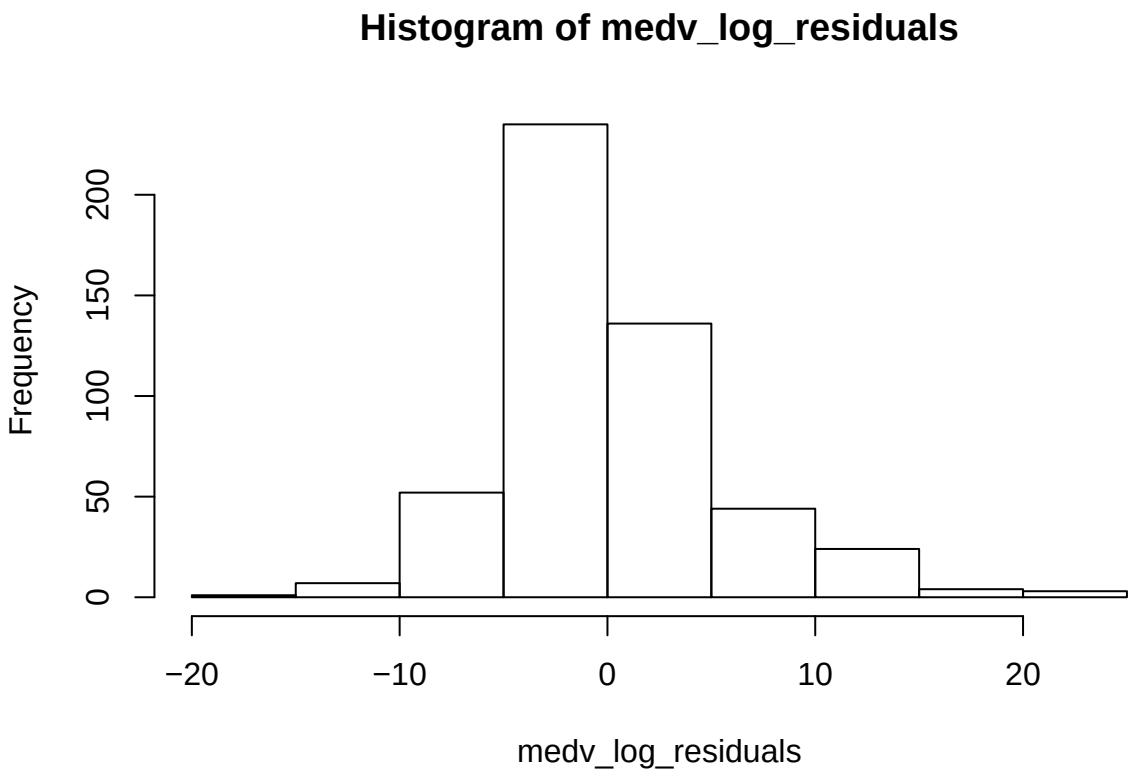
```
##
## Call:
## lm(formula = medv ~ rm + lstat + dis, data = medv_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.992  -3.133  -0.871   1.910  25.944
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.80829    3.36800   0.834 0.404781
## rm           4.87339    0.44456  10.962  < 2e-16 ***
## lstat       -0.72333    0.04933 -14.662  < 2e-16 ***
## dis         -0.46128    0.13495  -3.418 0.000682 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.482 on 502 degrees of freedom
## Multiple R-squared:  0.6468, Adjusted R-squared:  0.6447
## F-statistic: 306.4 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
medv_model_log<- lm(medv~rm+lstat+log(dis), data=medv_subset)
summary(medv_model_log)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat + log(dis), data = medv_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4962  -3.0956  -0.9204   2.0185  24.7448
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.25839    3.41301   1.248    0.213
## rm           4.84833    0.44181  10.974  < 2e-16 ***
## lstat       -0.75284    0.05084 -14.808  < 2e-16 ***
## log(dis)    -2.24726    0.54960  -4.089 5.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.456 on 502 degrees of freedom
## Multiple R-squared:  0.6502, Adjusted R-squared:  0.6481
## F-statistic: 311.1 on 3 and 502 DF,  p-value: < 2.2e-16
```

"rm" and "lstat" are negative correlated. When building models with these two variables, it will cause Multicollinearity. Estimated coefficients may change dramatically in response to small changes in data.
ii.
Adjusted R-squared is 0.6447.
iii.
Using "log(dis)", adjusted R-squared does not improve much from 0.6447 to 0.6481.

```
medv_log_residuals <- residuals(medv_model_log)
p1 <- hist(medv_log_residuals)
```



Residuals of medv model with log(dis) are normally distributed.

(c)

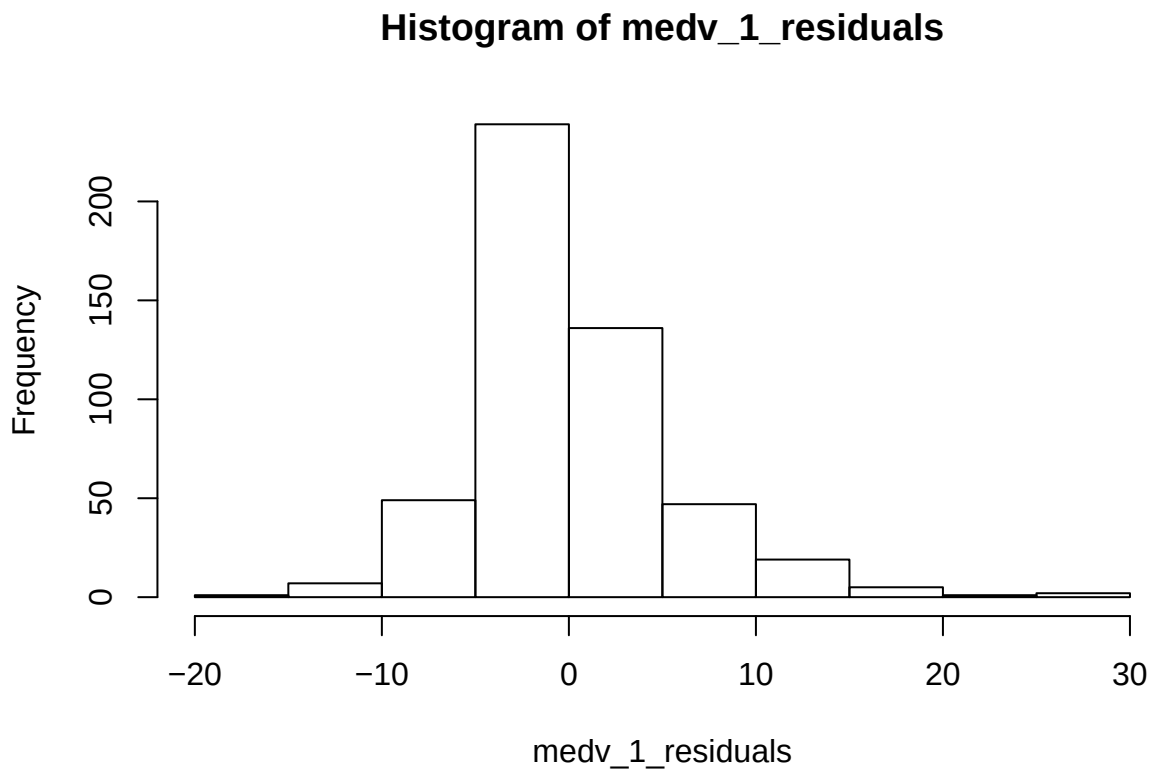
i.

```
medv_model_1<- lm(medv~rm+lstat+dis - 1, data=medv_subset)
summary(medv_model_1)
```

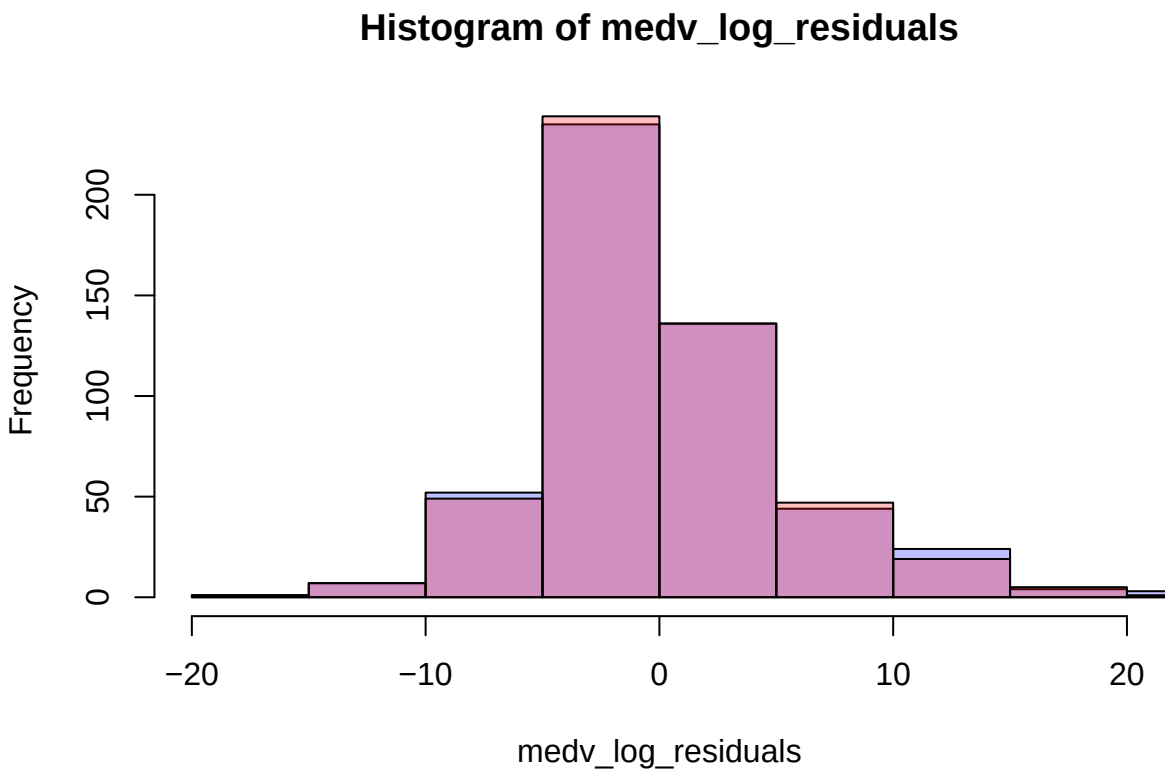
```
##
## Call:
## lm(formula = medv ~ rm + lstat + dis - 1, data = medv_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5614  -3.0458  -0.8712   1.8787  26.8212
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## rm     5.23050    0.11914  43.902  < 2e-16 ***
## lstat -0.69214    0.03215 -21.526  < 2e-16 ***
## dis   -0.42055    0.12577  -3.344 0.000888 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.481 on 503 degrees of freedom
## Multiple R-squared:  0.9496, Adjusted R-squared:  0.9493
## F-statistic:  3157 on 3 and 503 DF,  p-value: < 2.2e-16
```

In the previous model, p-value for intercept is 0.404781, which is not statistically significant.
ii.
Adjusted R-squared is 0.9493.
iii.

```
medv_1_residuals <- residuals(medv_model_1)
p2 <- hist(medv_1_residuals)
```




```
plot(p1, col = rgb(0,0,1,1/4), xlim = c(-20,20))
plot(p2, col = rgb(1,0,0,1/4), xlim = c(-20,20), add = TRUE)
```



Residuals of model without intercept have less variance than model using log(dis).