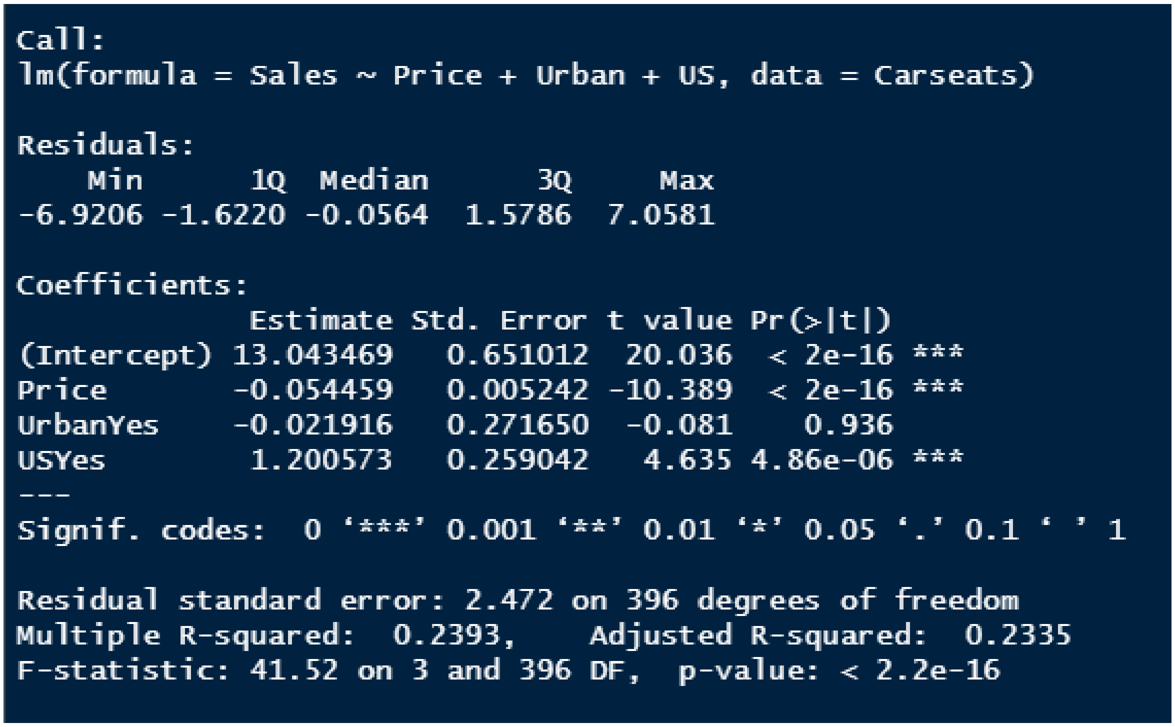
Applied Data Mining
Chapter 3, Linear Regression
Section 3.7 Questions 10 and 13

10. This question should be answered using the Carseats data set.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.



```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9206 -1.6220 -0.0564  1.5786  7.0581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936    
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

The very low p-value for Price indicates that Price has a significant (negative) effect on sales. Given a one-dollar shift in Price, Sales, would decrease by $54.46, provided that Urban and US remain fixed.

Urban indicates an average, unit Sales decrease by $21.90 if the store is located in the city versus in the country. However, the presence of a direct relationship between Urban and Sales is not indicated, given the very high value of p (0.936, or, close to “1”), provided that Price and US remain fixed.

Whether or not the store is in the US seems to be (positively) significant with relation to Sales, given the very low p-value; if the store is located in the US, Sales are $1,200.57 more compared to non-US locations, provided that Urban and Price stay fixed. [don’t need 1-unit increase for qualitative/categorical variable] baseline = non-US.

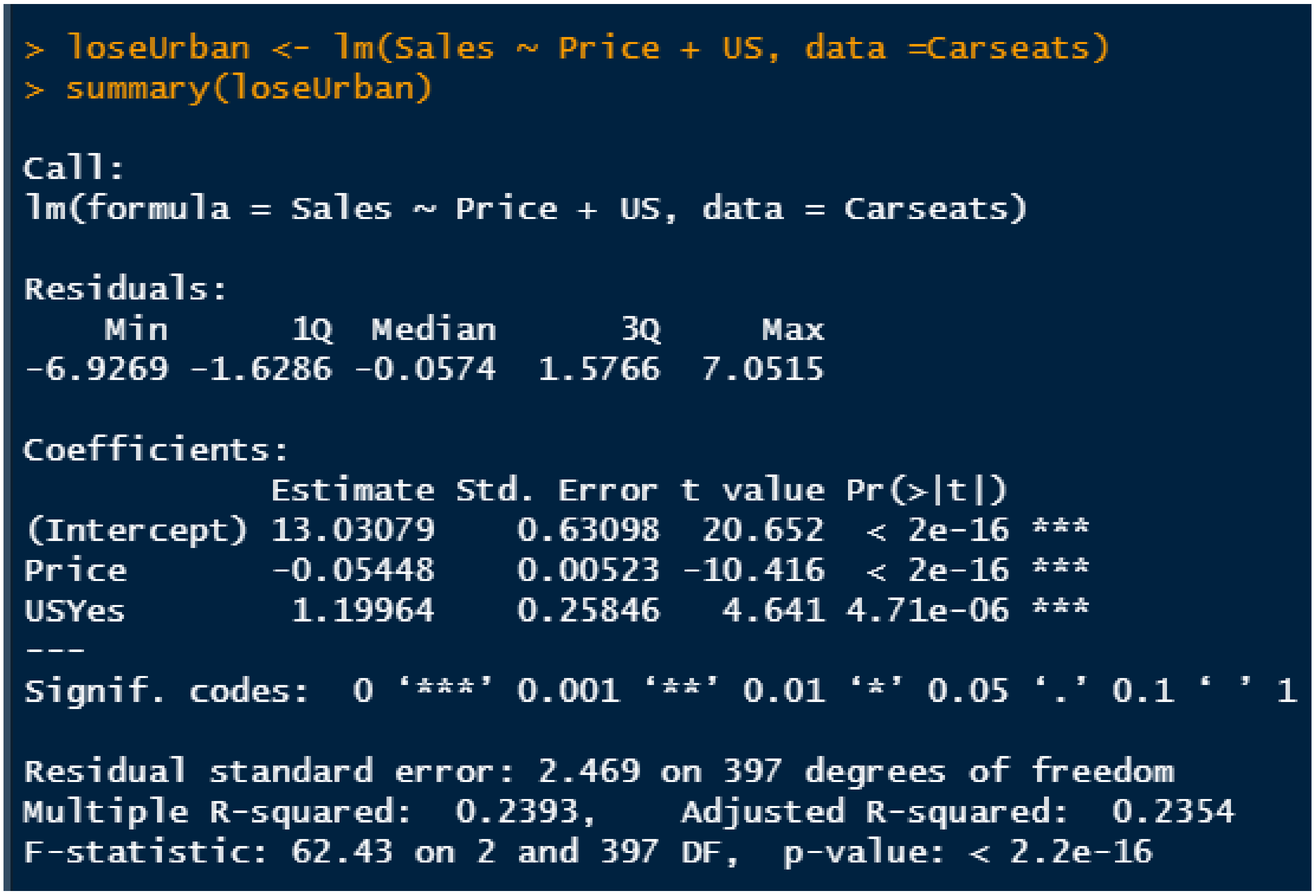
(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

Sales = 13.043469 + ((-0.054459)Price) + ((-0.021916)Urban) + ((1.200573)US) + ε

(d) For which of the predictors can you reject the null hypothesis H0 : βj = 0?

The null hypothesis can be rejected for the predictors Price and US because Urban doesn’t have a significant relationship with Sales, but Price and US both have nonzero values.

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.



```
> loseUrban <- lm(Sales ~ Price + US, data =Carseats)
> summary(loseUrban)

Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9269 -1.6286 -0.0574  1.5766  7.0515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2354 
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

(f) How well do the models in (a) and (e) fit the data?

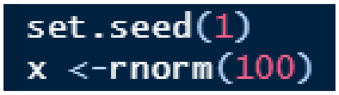
The model in (e) fits slightly better than the model in (a), but not by much (.19%) The Adjusted R-squared shows us that Price and US in the second model in (e) account for 23% of the variability of Sales. The same is true of the model in (a), but it is .0019 worse at accounting for the variability in Sales, given that it has a “noise predictor”, Urban.

F-stat tells us that at least one of the predictors is significant. Larger here the F-stat value indicates a greater level of comparison. There is more significance in the smaller model.

---

13. 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.

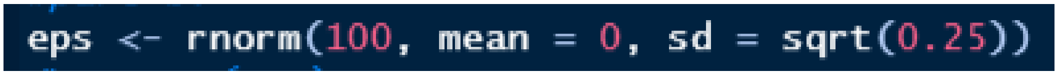
(a) Using the rnorm() function, create a vector, x, containing 100 observations drawn from a N(0, 1) distribution. This represents a feature, X.



```
set.seed(1)
x <-rnorm(100)
```

(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.

[**Standard deviation** is calculated as the **square root of variance**]

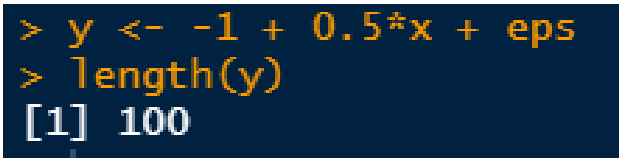


```
eps <- rnorm(100, mean = 0, sd = sqrt(0.25))
```

(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 β0 and β1 in this linear model?

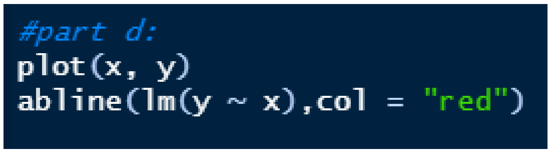


```
> y <- -1 + 0.5*x + eps
> length(y)
[1] 100
```

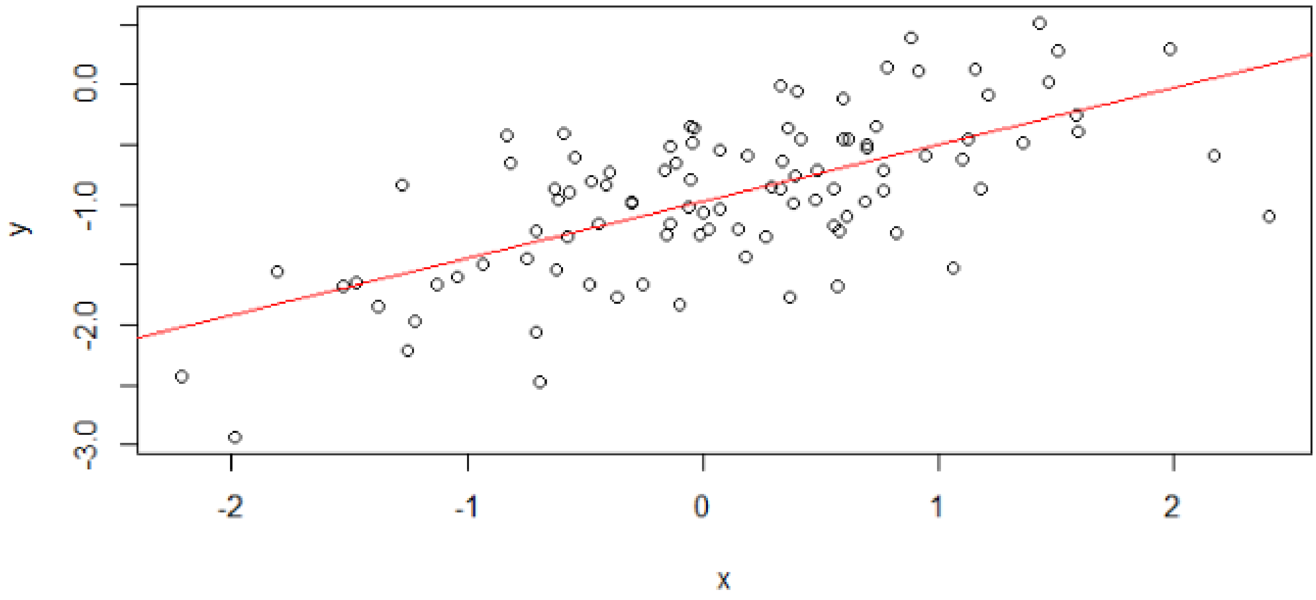
The length of vector y is 100.

β0 (intercept) is -1 and β1 (slope) is 0.5.

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

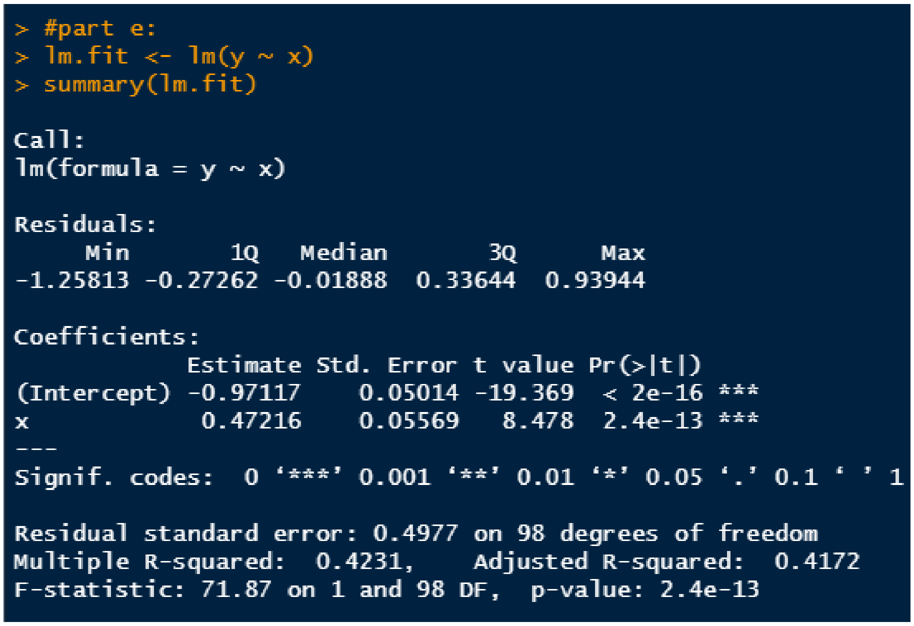


```
#part d:
plot(x, y)
abline(lm(y ~ x),col = "red")
```



The linear regression model seems to be an appropriate model, in that there is a positive correlation or direct relationship between x, the predictor, and y, the response variable, because of the way that the data points are scattered in a linear shape. (When x increases, y increases.) [exploratory data analysis – visualization to see the nature of the pattern between x and y]

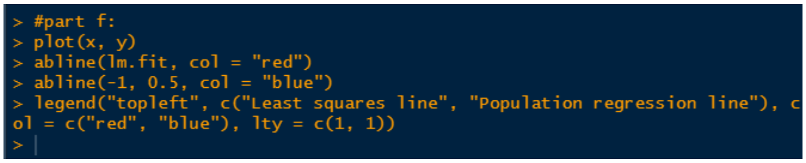
(e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do ˆ β0 and ˆ β1 compare to β0 and β1?



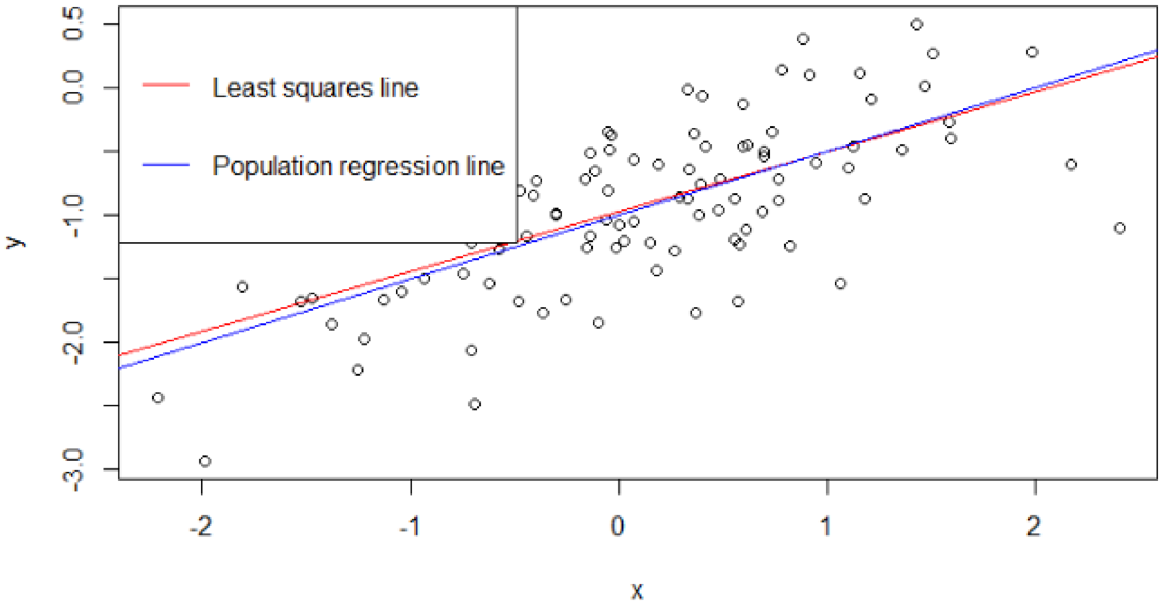
```
> #part e:
> lm.fit <- lm(y ~ x)
> summary(lm.fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.25813 -0.27262 -0.01888  0.33644  0.93944 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.97117    0.05014 -19.369  < 2e-16 ***
x            0.47216    0.05569   8.478  2.4e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4977 on 98 degrees of freedom
Multiple R-squared:  0.4231,    Adjusted R-squared:  0.4172 
F-statistic: 71.87 on 1 and 98 DF,  p-value: 2.4e-13
```

The p-value being low indicates that the predictor x has direct effect on the response variable y. β0-hat (-0.97117) and β1-hat (0.47216) are somewhat close to β0 (-1) and β1 (0.5). However, the model can only capture 41.72% of the variation in y. [smaller error variance is a better case for linear regression]. We are probably still missing a very significant predictor, one that is highly correlated with our response variable. [Some studies are satisfied with finding 20% variation. Not as satisfactory in a social science, for example.]

(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.

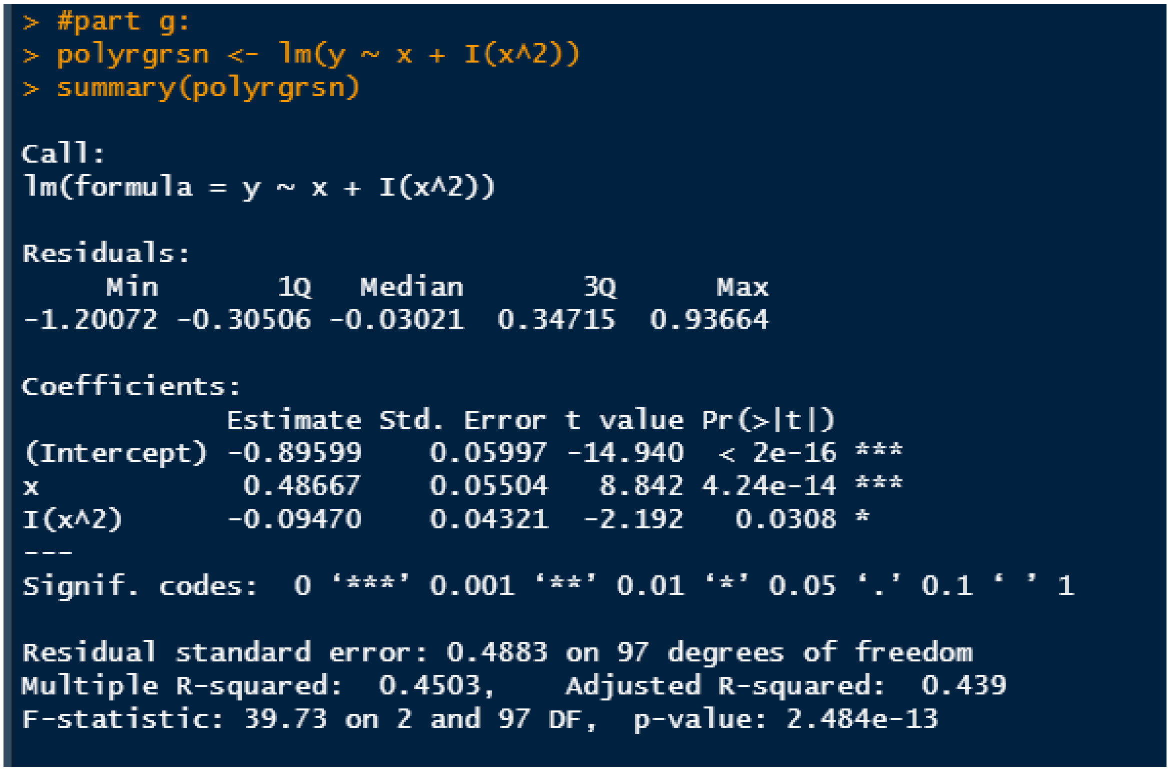


```
> #part f:
> plot(x, y)
> abline(lm.fit, col = "red")
> abline(-1, 0.5, col = "blue")
> legend("topleft", c("Least squares line", "Population regression line"), c
ol = c("red", "blue"), lty = c(1, 1))
> |
```



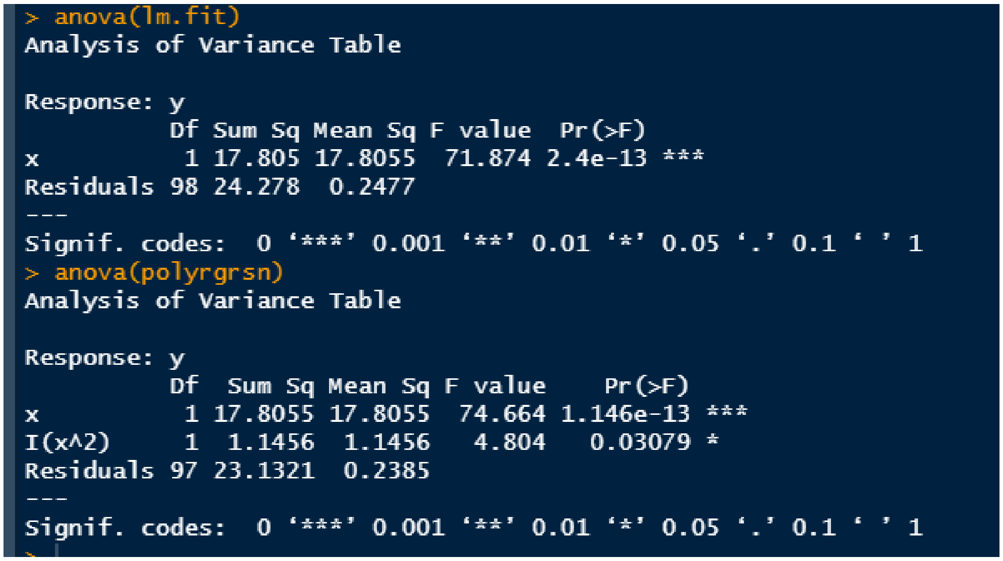
[Our goal with the linear regression model is to find the least squares line that is closest to the population regression line. In this case, a simulated model, the population regression line serves as a reference.]

(g) Now fit a polynomial regression model that predicts y using x and x2. Is there evidence that the quadratic term improves the model fit? Explain your answer.



```
> #part g:
> polyrgrsn <- lm(y ~ x + I(x^2))
> summary(polyrgrsn)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.20072 -0.30506 -0.03021  0.34715  0.93664 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.89599    0.05997 -14.940  < 2e-16 ***
x            0.48667    0.05504   8.842 4.24e-14 ***
I(x^2)      -0.09470    0.04321  -2.192   0.0308 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4883 on 97 degrees of freedom
Multiple R-squared:  0.4503,	Adjusted R-squared:  0.439 
F-statistic: 39.73 on 2 and 97 DF,  p-value: 2.484e-13
```

Since the Adjusted R-squared and p-values improved only slightly, the benefits of using polynomial regression in this case are nominal, though it helps a little bit. However, and probably most importantly, the p-value of the quadratic term of the variable, I(x^2), is quite low and less than 0.05, which most likely indicates that it should be accepted as a significant “predictor”. [Now, the relationship between y and x is quadratic. In terms of finding the least squares estimator, we just need to find the parameter that minimizes yi hat.]



```
> anova(lm.fit)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)    
x          1 17.805 17.8055  71.874 2.4e-13 ***
Residuals 98 24.278  0.2477                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(polyrgrsn)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x          1 17.8055 17.8055  74.664 1.146e-13 ***
I(x^2)     1  1.1456  1.1456   4.804   0.03079 *  
Residuals 97 23.1321  0.2385                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```