Chapter 3 HW

Nhi Vu

February 28,2021

Conceptual Questions

1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV,radio, and newspaper, rather than in terms of the coefficients of the linear model.

The hypotheses are $\beta_1 = 0$,$\beta_2 = 0$ and $\beta_3 = 0$, which means in table 3.4 the null hypotheses is that advertising budgets of "TV", "radio" or "newspaper" do not have any influences on sale.

As we know that "A p-value higher than 0.05 is not statistically signicant and indicates strong evidence for the null hypothesis", the p-value of TV and radio is statistically highly significant as less than 0.001 so we reject null hypotheses which includes that the advertising budgets of TV and radio affect on Sale. However, the p-value of newpaper is higher than 0.05, then we accept the null hypothesis to include that there is no relationship between sale and the advertising budgets of newpaper.

3. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ,X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\beta_0 = 50, \beta_1 = 20, \beta_2 = 0.07, \beta_3 = 35, \beta_4 = 0.01, \beta_5 = -10.$

(a) Which answer is correct, and why?

We get the general least square line for both gender:

$$\hat{y} = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA \times IQ - 10GPA \times Gender$$

which for female ($Gender = 1$):

$$\hat{y}_f = 85 + 10GPA + 0.07IQ + 0.01GPA \times IQ$$

and for male ($Gender = 0$) :

$$\hat{y}_m = 50 + 20GPA + 0.07IQ + 0.01GPA \times IQ$$

i. For a fixed value of IQ and GPA, males earn more on average than females.

We don't have enough information to include that this statement is true, because:

$$\hat{y}_m > \hat{y}_f$$

$$50 + 20GPA > 80 + 10GPA$$

$$GPA > 3.5$$

Therefore, if males earn more on average than females, the fixed GPA is greater than 3.5.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

Similar to above question, if females earn more on average than males, the fixed GPA is less than 3.5, so the statement is impossible to be correct.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

That's correct because if males earn more on average than females, the fixed GPA is higher than 3.5.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

That's is wrong because it happens the opposite way, which females earn more on average than males, the fixed GPA is lower than 3.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

$$\hat{y}_f(IQ = 110, GPA = 4) = 85 + 10GPA + 0.07IQ + 0.01GPA \times IQ$$

$$\hat{y}_f(IQ = 110, GPA = 4) = 85 + 10(4) + 0.07(110) + 0.01(4)(110) = 137.1$$

which gives us a starting salary of 137100$.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

I believe it is false because to evaluate the effects from interaction between GPA and IQ on the model, we need to check $H_0 : \beta_4 = 0$ by looking at its p-value, not its coefficient.

Applied questions:

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

```
library("ISLR")
data(Carseats)
?Carseats
```

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

```
lm.fit1=lm(Sales~Price+Urban+US, data=Carseats)
summary(lm.fit1)
```

```
##
## 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
```

Therefore,

$$Sale = 13.043469 - 0.054459Price - 0.021916Urban + 1.200573US + \epsilon$$

Urban: Yes=1, No=0.

US: Yes=1, No=0.

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

- The coefficient of the “Price” variable may be interpreted by saying that the average effect of a price increase of 1 dollar is a decrease of 54.4588492 units in sales all other predictors remaining fixed.
- The coefficient of the “Urban” variable may be interpreted by saying that on average the unit sales in urban location are 21.9161508 units less than in rural location all other predictors remaining fixed.
- The coefficient of the “US” variable may be interpreted by saying that on average the unit sales in a US store are 1200.5726978 units more than in a non US store all other predictors remaining fixed.

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

$$Sale = 13.043469 - 0.054459Price - 0.021916Urban + 1.200573US + \epsilon$$

Urban: Yes=1, No=0.

US: Yes=1, No=0.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

Based on the output of “summary(lm.fit)”, the p-value of Price and US are both very very small, less than 0.000001 so we can reject the null hypothesis of predictors “Price” and “US”.While the predictors “Urban” has a highh p-value(>0.01), so it proves that there are no relationship between Sale and Urban or it has very very less effects on Sale.

(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.

Like explanation in question c), we get a smaller multiple regression model to predict Sales using Price and US because Urban has insignificant effect on Sale.

```
lm.fit2=lm(Sales~Price+US, data=Carseats)
summary(lm.fit2)
```

```
##
## 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 R^2 of smaller model is higher and its residual standard error is lower than so it indicates a better fit for the data than the models in (a).

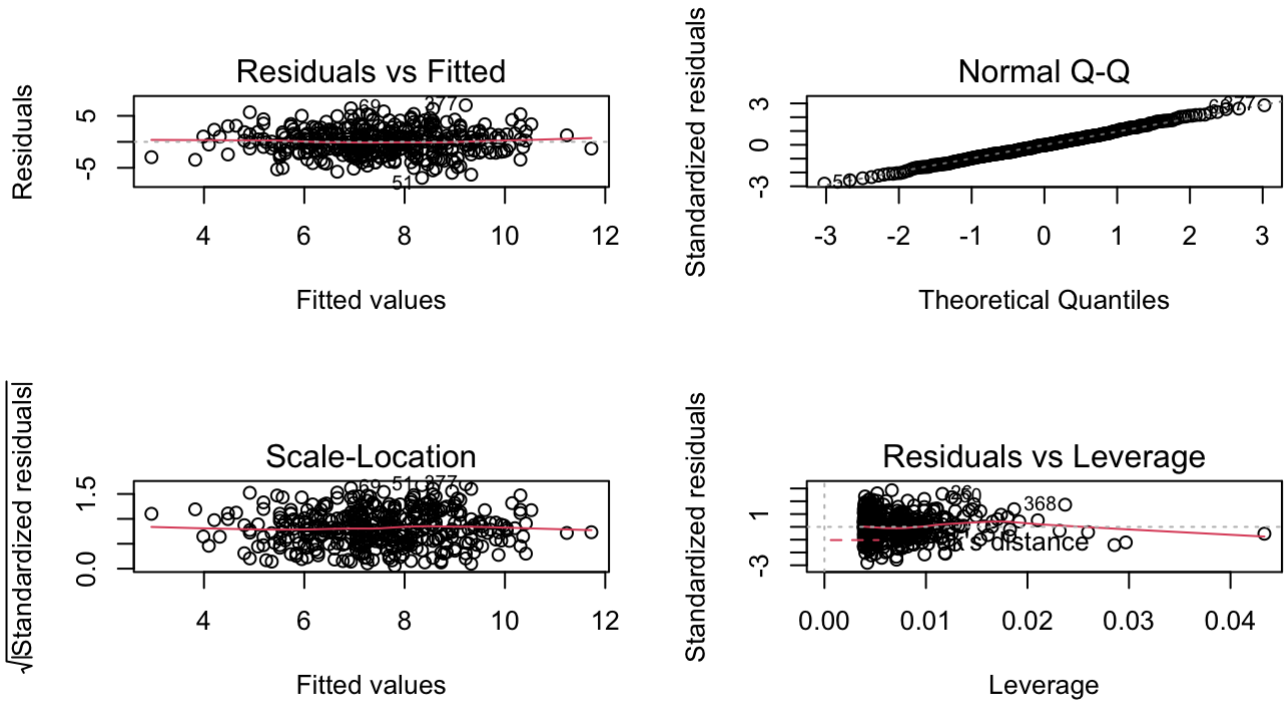
(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

```
confint(lm.fit2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

(h) Is there evidence of outliers or high leverage observations in the model from (e)?

```
par(mfrow = c(2, 2))
plot(lm.fit2)
```



In the Scale-Location graph does not show any highlighted outlier. In the Residuals vs Leverage graph notes a very high leverage observation.

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.

```
set.seed(1)
```

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

```
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.

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

(c) Using x and eps, generate a vector y according to the model $Y = -1 + 0.5X + \epsilon$ (3.39).

```
y=-1+0.5*x+eps
```

What is the length of the vector y?

Because y is generated by vector x with n=100, the length of the vector y is 100. Or we can use fuction length() to check:

```
length(y)
```

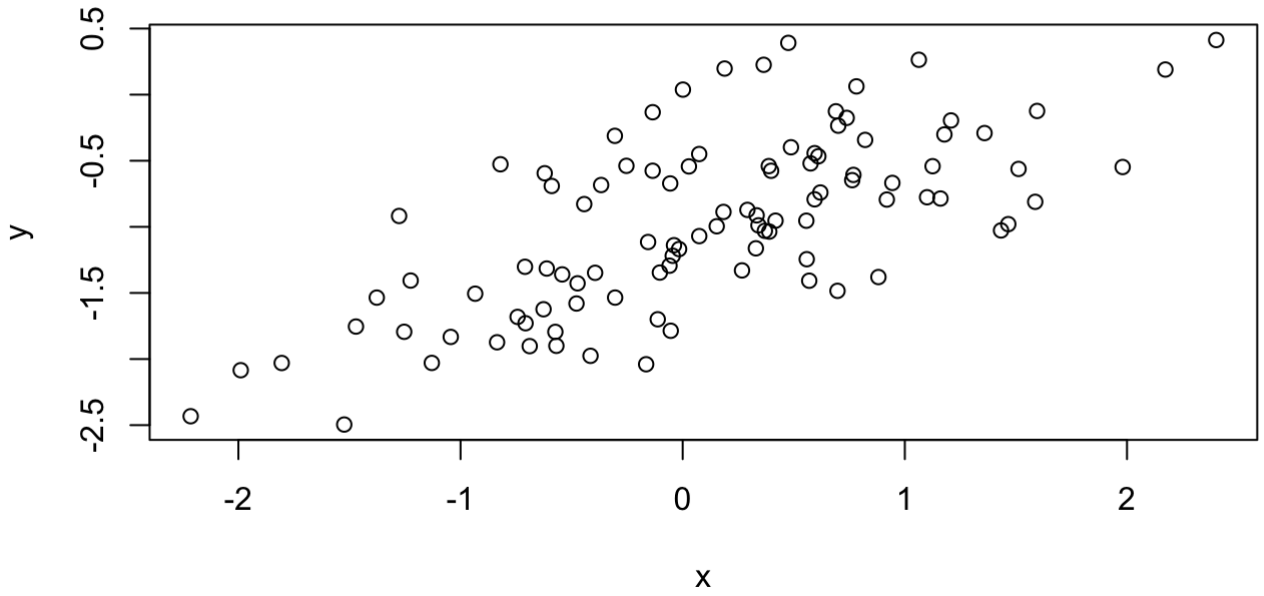
```
## [1] 100
```

What are the values of β_0 and β_1 in this linear model?

$$\beta_0 = -1, \beta_1 = 0.5$$

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

```
plot(x,y)
```



We have a clearly linear relationship between x and y, and a presence of variance, var(ϵ), in this distribuition represented by ϵ.

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

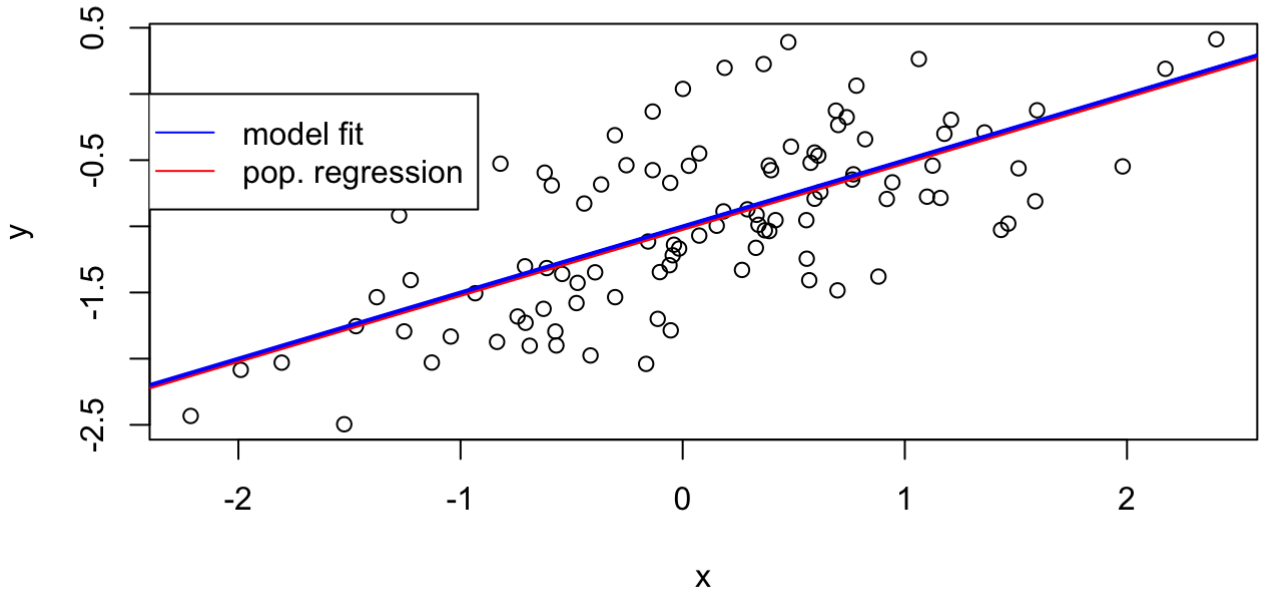
```
lm.fit.original=lm(y~x)
summary(lm.fit.original)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

$\hat{\beta}_0$ and $\hat{\beta}_1$ are very closely from the original ones. So the linear relationship is very closely of the true form of $\hat{f}$.

(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.

```
plot(x,y)
abline(lm.fit.original, col="red",lw=2)
abline(-1, 0.5, col="blue",lwd=2)
legend(-2.5,0, legend = c("model fit", "pop. regression"), col=c("blue", "red"),lwd=1)
```



(g) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

```
z<-x*x
lm.fit.poly=lm(y~x+z)
summary(lm.fit.poly)
```

```
##
## Call:
## lm(formula = y ~ x + z)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98252 -0.31270 -0.06441  0.29014  1.13500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
## x            0.50858    0.05399   9.420  2.4e-15 ***
## z           -0.05946    0.04238  -1.403    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 97 degrees of freedom
## Multiple R-squared:  0.4779, Adjusted R-squared:  0.4672
## F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

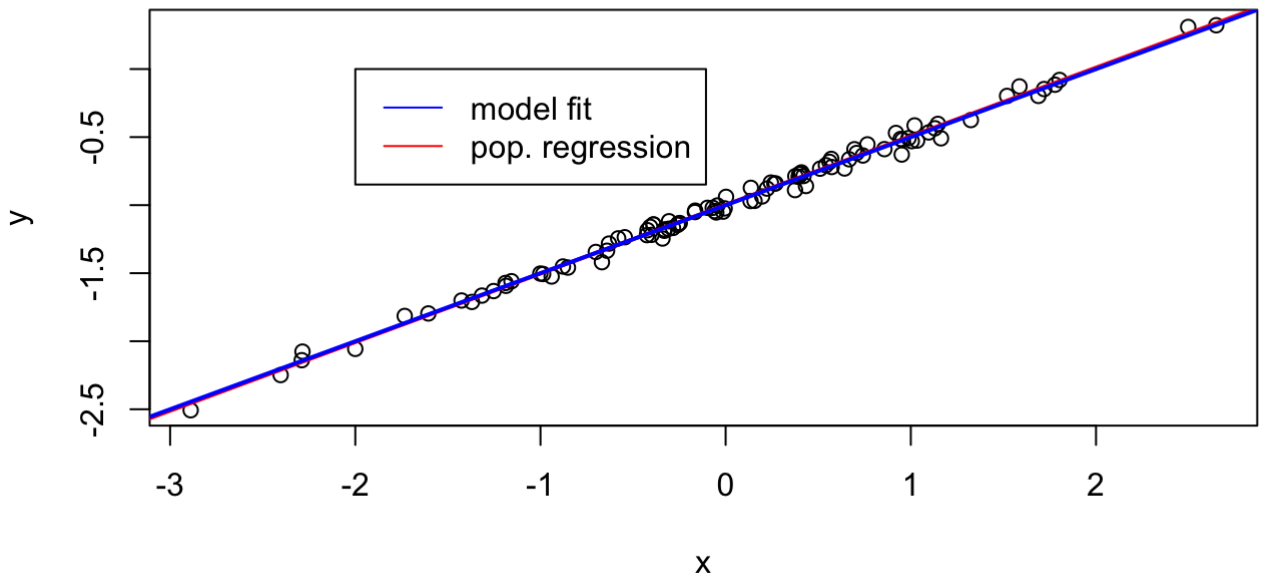
Because the p-value of x^2 is higher than 0.01, it indicates that x^2 has not significant effects on y.Thus, the quadratic term does not improves the model fit even though its R^2 is higher than of the linear model.

(h) Repeat (a)–(f) after modifying the data generation process in such a way that there is less noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error termin (b). Describe your results.

```
x <- rnorm(100)
eps <- rnorm(100, 0, .04)
y = -1 + .5*x + eps
lm.fit.less= lm(y ~ x)
summary(lm.fit.less)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10967 -0.02246 -0.00070  0.02719  0.07394
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.998062   0.003964  -251.8   <2e-16 ***
## x            0.504249   0.003850   131.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03962 on 98 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.9943
## F-statistic: 1.715e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(x,y)
abline(lm.fit.less, col="red",lw=2)
abline(-1, 0.5, col="blue",lwd=2)
legend(-2,0, legend = c("model fit", "pop. regression"), col=c("blue", "red"),lwd=1)
```



By decreasing the variance of the normal distribution used to generate the error term ϵ, the line regression seems overlap with the line of model (3.39).The R^2 is much higher, nearly to 100% and RSE is much lower as we have very little noise,which shows that this model is very close.

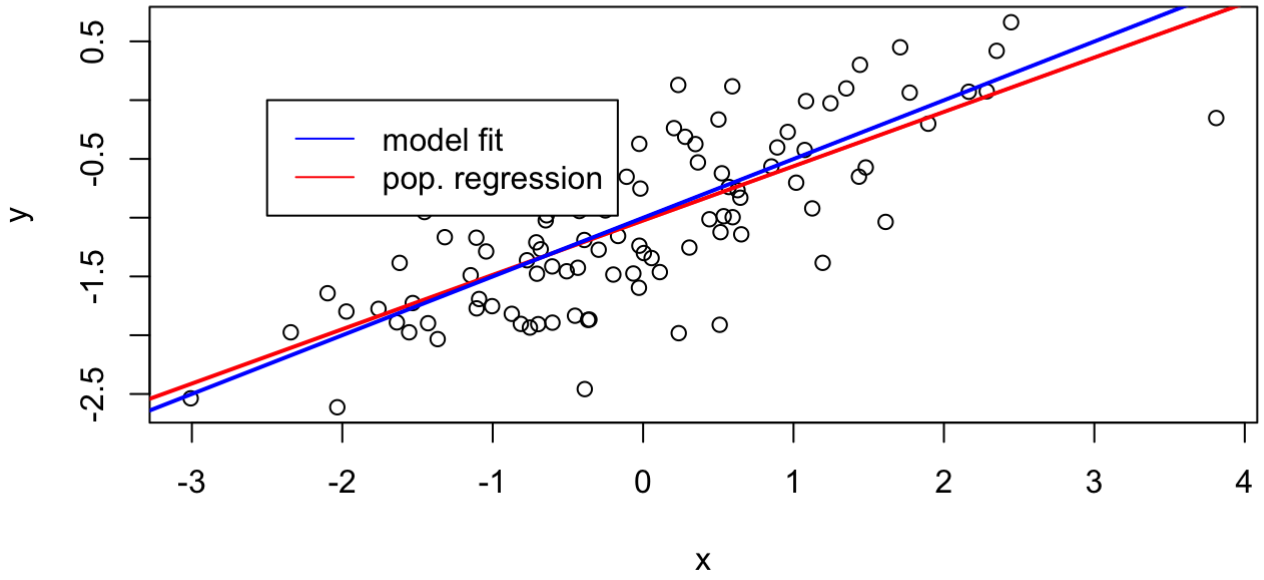
(i) Repeat (a)–(f) after modifying the data generation process in such a way that there is more noise in the data. The model (3.39) should remain the same. You can do this by increasing the

variance of the normal distribution used to generate the error term in (b). Describe your results.

```
x <- rnorm(100)
eps <- rnorm(100, 0, .5)
y = -1 + .5*x + eps
lm.fit.more= lm(y ~ x)
summary(lm.fit.more)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25507 -0.30275  0.01032  0.35241  1.04490
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.02373    0.04838  -21.16   <2e-16 ***
## x            0.46253    0.04155   11.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4835 on 98 degrees of freedom
## Multiple R-squared:  0.5584, Adjusted R-squared:  0.5539
## F-statistic: 123.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(x,y)
abline(lm.fit.more, col="red",lw=2)
abline(-1, 0.5, col="blue",lwd=2)
legend(-2.5,0, legend = c("model fit", "pop. regression"), col=c("blue", "red"),lwd=1)
```



Increasing more noise in the data makes RSE is higher and R^2 decreases even though the the coefficients are very close to the previous ones, the two lines are wider apart but are still really close to each other as we have a fairly large data set. ##### (j) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

```
#original
confint(lm.fit.original)
```

```
##                 2.5 %     97.5 %
## (Intercept) -1.1150804 -0.9226122
## x            0.3925794  0.6063602
```

```
#less noisy data.
confint(lm.fit.less)
```

```
##                 2.5 %     97.5 %
## (Intercept) -1.0059284 -0.9901956
## x            0.4966078  0.5118897
```

```
#more noisy data.
confint(lm.fit.more)
```

```
##                 2.5 %     97.5 %
## (Intercept) -1.1197386 -0.9277138
## x            0.3800695  0.5449816
```

It is very noted that noisiest data sets cause a wider confidential interval while with less noise, there is more predictability in the data set. All intervals seem to be centered on approximately 0.5.