Yaniv Bronshtein HW1 Regression and Time Series

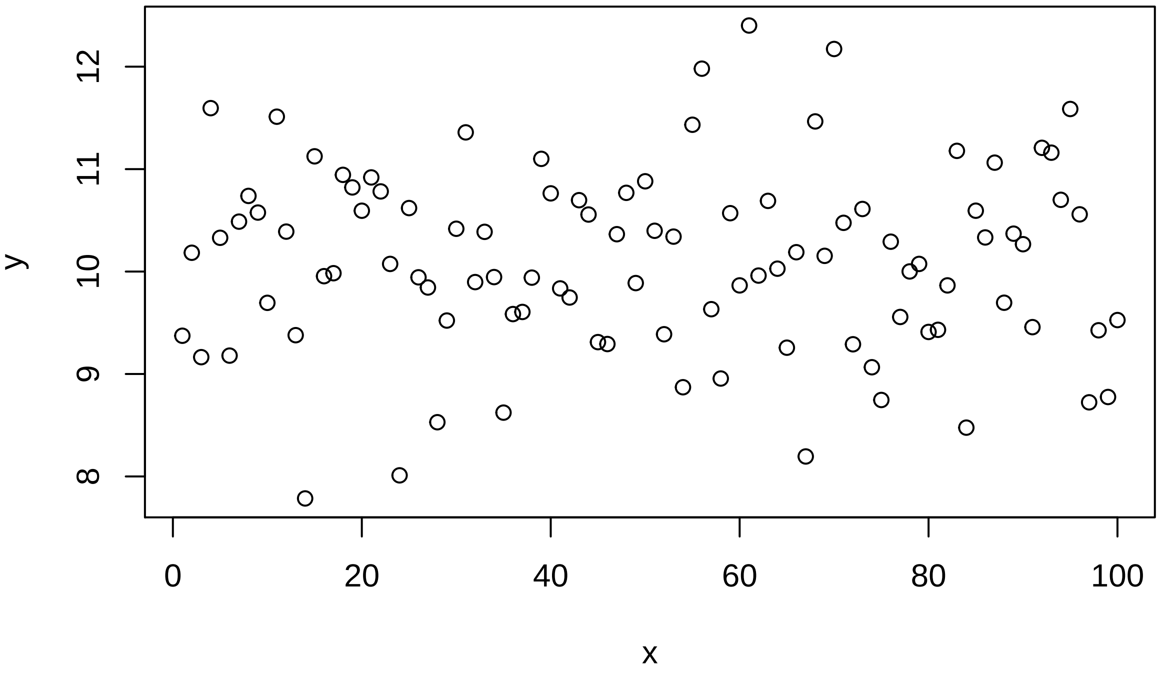
Yaniv Bronshtein

9/15/2021

#Question 1

(1)Create a column of y that includes 100 random sample from a normal distribution with mean=10 and sd=1

```
set.seed(1)
y <- matrix(rnorm(n=100,mean=10, sd=1))
#(2) Create a column of x that includes numbers 1 to 100
x <- matrix(1:100)
#(3) Draw a scatter plot using R. Based on the figure, do you think
#x and y are correlated?
plot(x,y)
```



```
corr <- cor(x,y)
corr
```

```
##              [,1]
## [1,] -0.01456896
```

Based on the plot and the result of cor(), I do not believe that x and y are correlated. The value is -0.01456896 which is very close to 0

(4) Fit a regression model y = B__0 + B__1 * x + epsilon and obtain the R^2 value.

```
lm_model <- lm(formula=y~x)
summary <- summary(lm_model)
r_sq <- summary$r.squared
r_sq
```

```
## [1] 0.0002122545
```

(5)Fit a regression model y=??x + epsilon and obtain the R^2 value.*
Compare this R2 value with the one you have in (4) and** **explain which one is more reasonable. Which model you will recommend?**

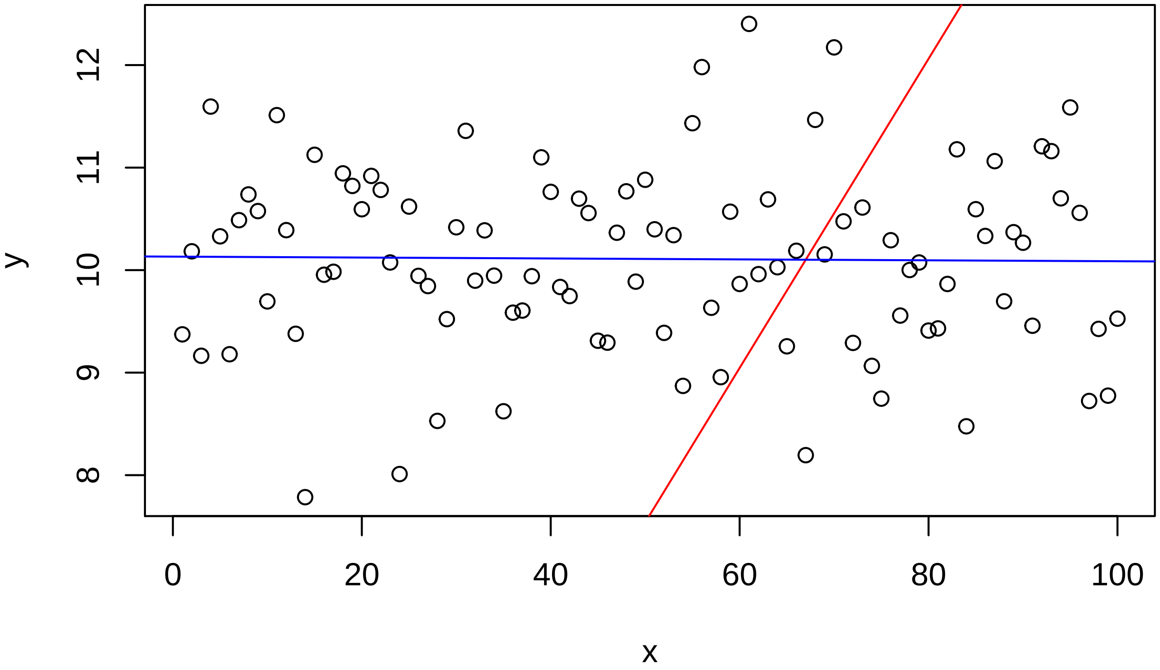
```
lm_model_2 <- lm(formula = y~x+0)
summary2 <- summary(lm_model_2)
r_sq2 <- summary2$r.squared
r_sq2
```

```
## [1] 0.7467852
```

The new R^2 is 0.7467852. The new R^2 is much higher and according to standards, is in a good range. If the R^2 had been significantly higher, it would signal likely over-fitting. I would recommend the second model as the first one suffers from under-fitting which explains why the R^2 is so low. The sum of squared distance between the points and the regression line is very high.

(6) Based on the model that you recommend in (5), is x an important factor to y? Is this conclusion consistent with the **one you have in (3)?

```
plot(x,y)
#Better model
abline(lm_model_2, col='red')
#Worse model
abline(lm_model, col='blue')
```



Based on the model I recommended in (5), we still cannot tell if x is an important factor to y because R^2 for the model is high while the correlation calculated mathematically and observed visually in the plot shows inconsistency between (3) and (5)