STAT 308 – Homework 2

For the problems in which calculations are needed, please include your `R` code with your answers, otherwise you will not be given full credit. Please upload your assignment by Thursday, September 15, 11:59 pm in a pdf file to Sakai.

- 1. In a simple linear regression problem where $n = 30$, we obtain

$$\sum_{i=1}^{n} x_i = 75, \sum_{i=1}^{n} y_i = 660, \sum_{i=1}^{n} x_i^2 = 240, \sum_{i=1}^{n} y_i^2 = 18000, \sum_{i=1}^{n} x_i y_i = -1200.$$

– a. Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$.

```
# Write your code below this line
n <- 30
sum_x <- 75
sum_y <- 660
sum_x2 <- 240
sum_y2 <- 18000
sum_xy <- -1200
num <- sum_xy - sum_x*sum_y/n
den <- sum_x2 - sum_x^2/n
beta1 <- num/den
beta0 <- sum_y/n - beta1*sum_x/n
beta1
```

```
## [1] -54.28571
```

```
beta0
```

```
## [1] 157.7143
```

– b. Calculate SSE and $s^2_{Y|X}$.

```
# Write your code below this line
SSE <- sum_y2 + n*beta0^2 + beta1^2*sum_x2 - 2*beta0*sum_y - 2*beta1*sum_xy + 2*beta0*beta1*sum_x
s2 <- SSE/(n-2)
SSE
```

```
## [1] -151234.3
```

```
s2
```

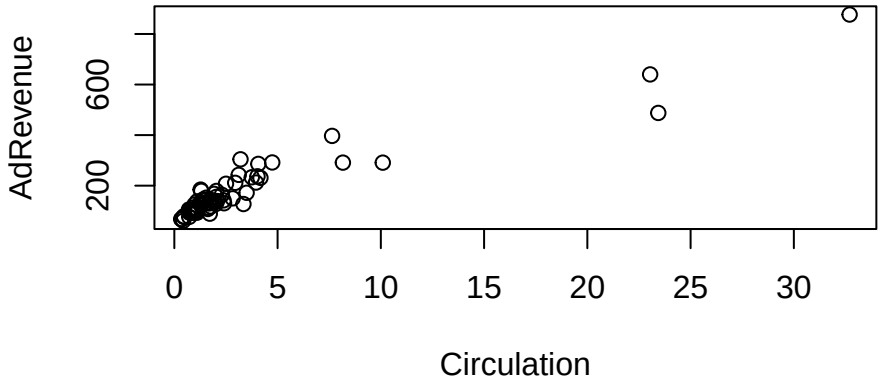
```
## [1] -5401.224
```

NOTE: In the real world, SSE is always positive!!!! I just wanted to use this example to show you the methods of these calculations.

- 2. Consider the dataset `AdRevenue.csv` on the course webpage. Suppose we are interested in modelling the ad revenue (in millions of dollars) of magazines based on the number of magazines in circulation (in millions).

– a. Draw a scatterplot of AdRevenue vs. Circulation. Comment on the four aspects of a scatterplot.

```
# Write your code below this line
AdRev <- read.csv("../Data/AdRevenue.csv")
plot(AdRevenue ~ Circulation,AdRev)
```



There appears to be a positive linear relationship with moderate strength and no apparent outliers.

– b. Do you think a linear relationship between Circulation and AdRevenue is appropriate? Justify your response.

Because the scatterplot appears to be linear, a linear relationship between Circulation and AdRevenue seems appropriate

– c. Using `R`, find the equation of the least squares regression line.

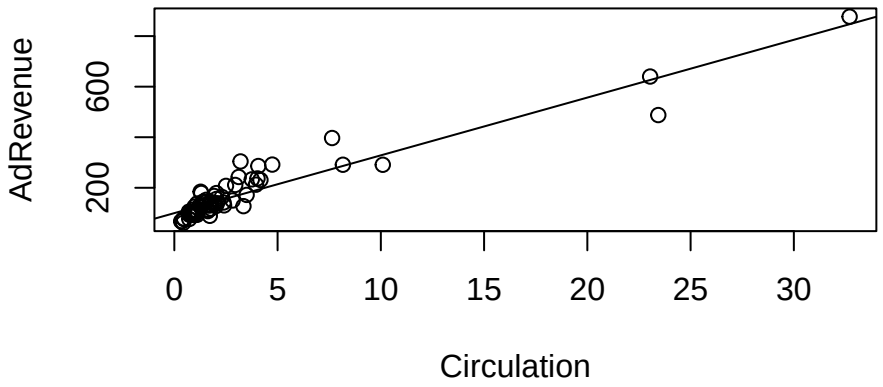
```
# Write your code below this line
mod <- lm(AdRevenue ~ Circulation,AdRev)
mod
```

```
##
## Call:
## lm(formula = AdRevenue ~ Circulation, data = AdRev)
##
## Coefficients:
## (Intercept)  Circulation
##       99.81        22.85
```

$\hat{Y}_x = 99.81 + 22.85 * x$

– d. Add the least squares regression line to the scatterplot in (a).

```
# Write your code below this line
plot(AdRevenue ~ Circulation,AdRev)
abline(mod)
```



– e. Interpret the slope of the regression line in the context of the given problem.

For a 1 million magazine increase in circulation, the expected ad revenue of the magazine increases by 22.85 million dollars.

– f. Interpret the intercept of the regression line in the context of the given problem. Does this interpretation make sense? Why or why not?

When there are no magazines in circulation, the expected amount of ad revenue of the magazine is 99.81 million dollars. This does not make sense because if there are no magazines in circulation, there will be no ad revenue.

– g. What do we expect the amount of ad revenue to be when there are 4 million magazines in circulation?

```
# Write your code below this line
newdata <- data.frame(Circulation=4)
predict(mod,newdata)
```

```
##        1
## 191.2231
```

$\hat{Y}_{x=4} = 191.22$ million dollars

– h. Find the value of SSE and $s^2_{Y|X}$ for the least squares regression line.

```
# Write your code below this line
SSE <- deviance(mod)
df <- df.residual(mod)
s2 <- SSE/df
SSE
```

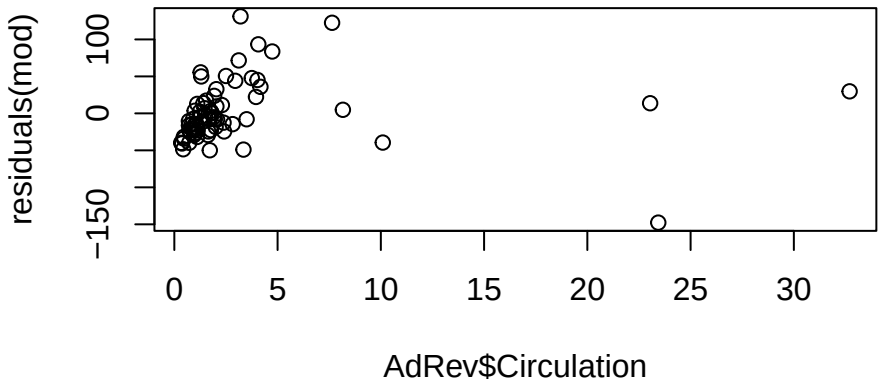
```
## [1] 121223.3
```

```
s2
```

```
## [1] 1782.696
```

– i. Determine if the assumption of homoscedasticity is violated.

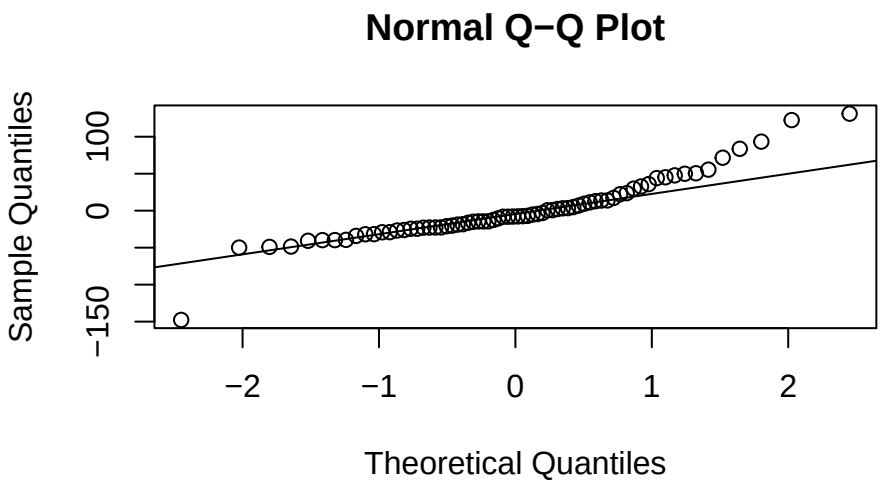
```
# Write your code below this line
plot(AdRev$Circulation,residuals(mod))
```



There are multiple possible ways to answer this question. Some may see a positive slope in the residuals for the lower levels of circulation, suggesting that there is evidence that there is not homoscedasticity. Some may see a random scattering of points around zero with a similar spread, suggesting the assumption of homoscedasticity is not violated.

– j. Determine if the assumption of normally distributed residuals is violated.

```
# Write your code below this line
qqnorm(residuals(mod))
qqline(residuals(mod))
```



Again, there are multiple possible answers to this question. Some of you may see the majority of the points in the middle fall closely to the reference line, suggesting the normality assumption is not violated. Others may pay closer attention to the points in the upper tail deviate dramatically from the reference line, suggesting there is evidence that the normality assumption is violated.