STAT302 Assignment 1 Solution

5/29/2023

Q1.4 (1 point)

$\hat{\beta}_1 = 0.01251$

Q1.6 (1 point)

$\hat{\beta}_0 = -16.47$

Q1.8 (1 point)

Every year has a 0.01251 increase in winning jump length compared with previous year.

Q1.10 (1 point)

$\hat{\sigma}_e = 0.259522$

Q1.12 (1 point)

$df = 26$

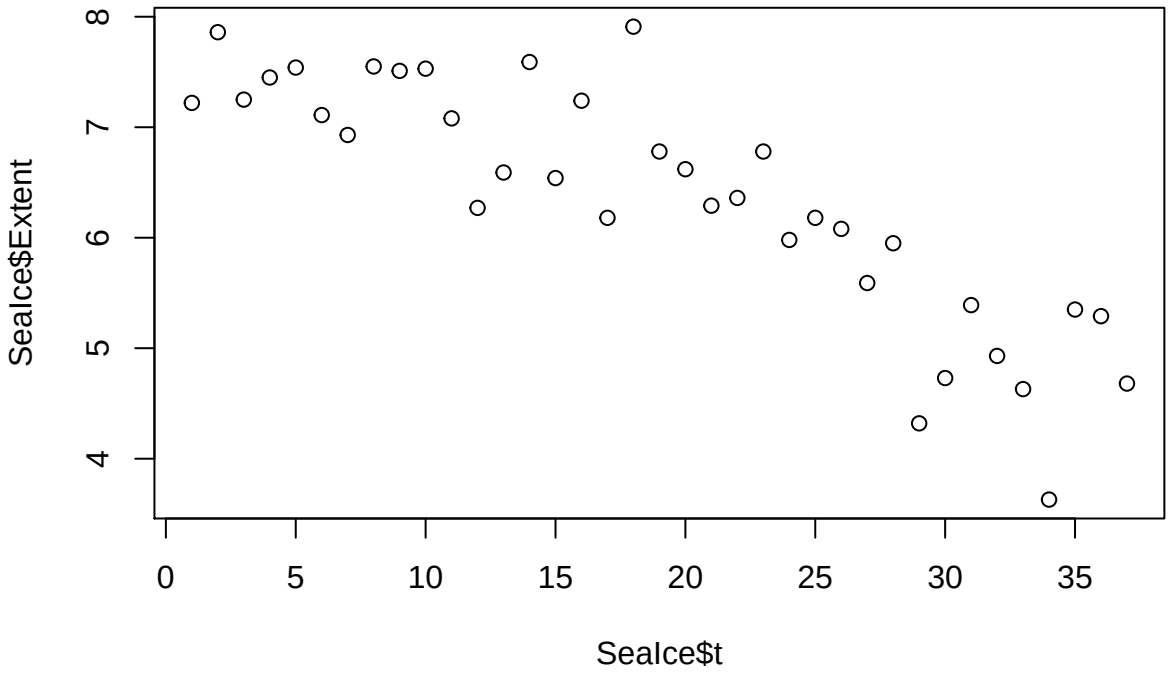
Q1.14 (2 points)

$$\hat{y} = 78 - 0.5x_1 = 78 - 0.5 \times 30 = 63 \text{ (1 point)} y_1 - \hat{y} = 60 - 63 = -3 \text{ (1 point)}$$

Q1.28 (12 points)

a. (2 points: 1 point for plot, 1 point for comment)

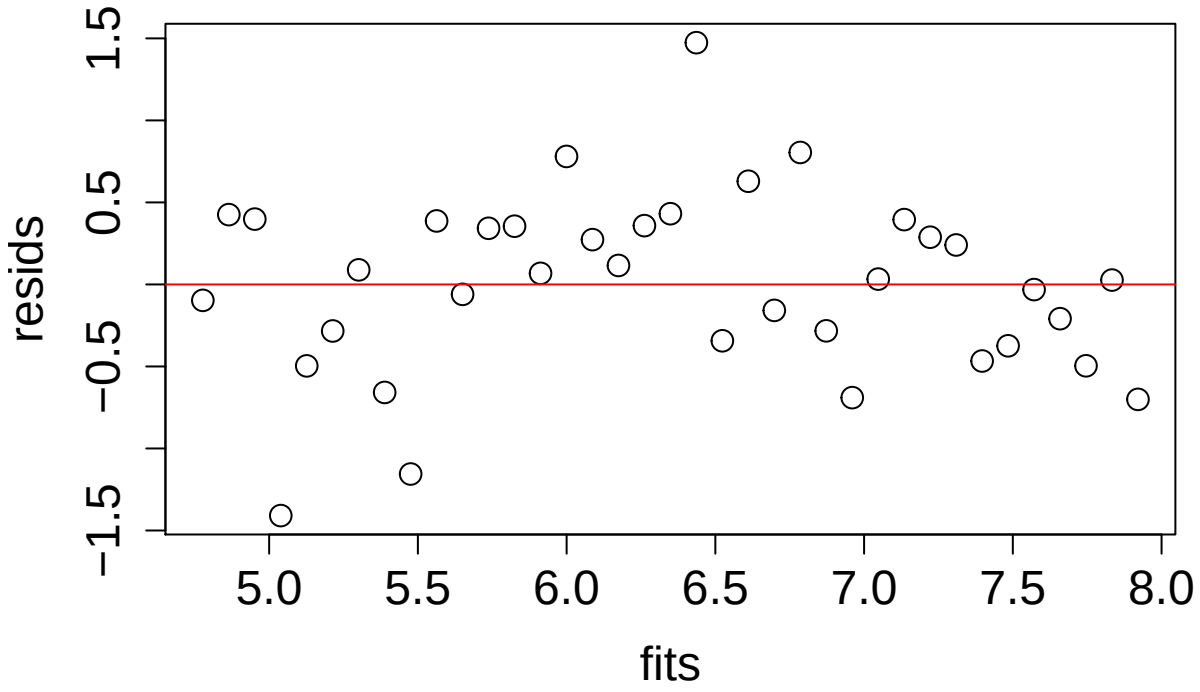
```
library(Stat2Data)
data("SeaIce")
plot(SeaIce$Extent~SeaIce$t)
```



The scatterplot shows that t is a fairy good predictor for $Extent$, and there is a negative relationship between these two variables.

b. (2 points: 1 point for plot, 1 point for comment)

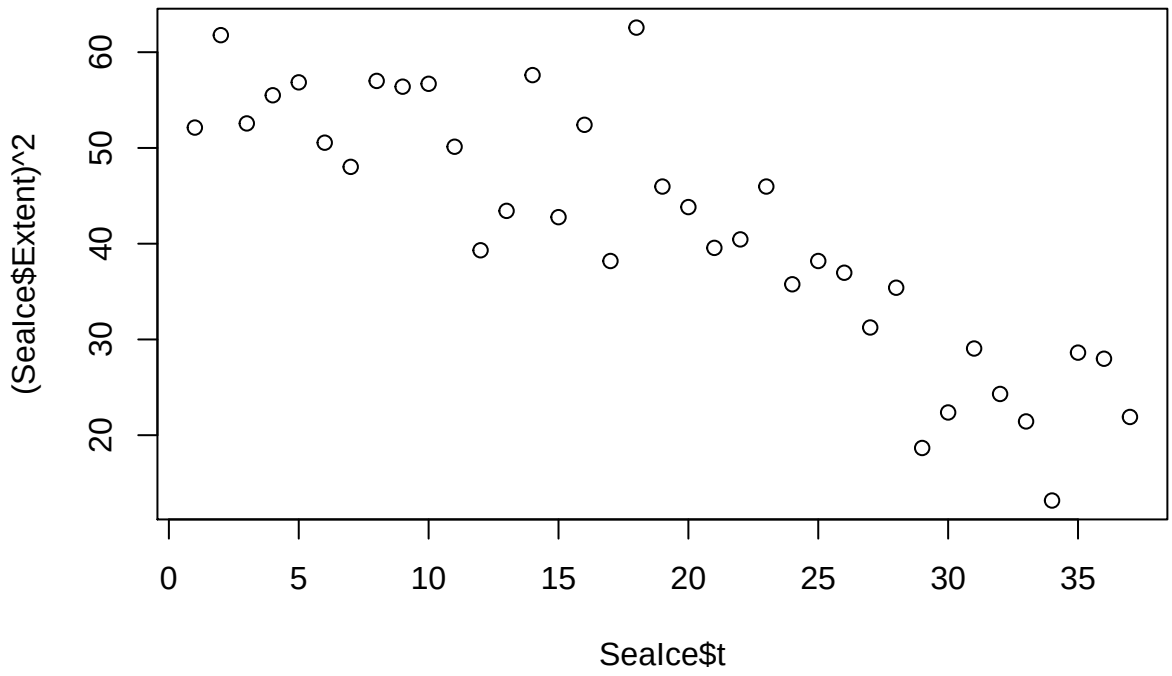
```
model <- lm(Extent~t, data=SeaIce)
fits <- model$fitted.values
resids <- model$residuals
plot(resids~fits,cex.lab=1.5,cex.axis=1.5,cex=1.5)
abline(0,0,col="red")
```



Residuals versus fitted values plot show linearity and constant variance conditions hold.

c. (2 points: 1 point for plot, 1 point for comment)

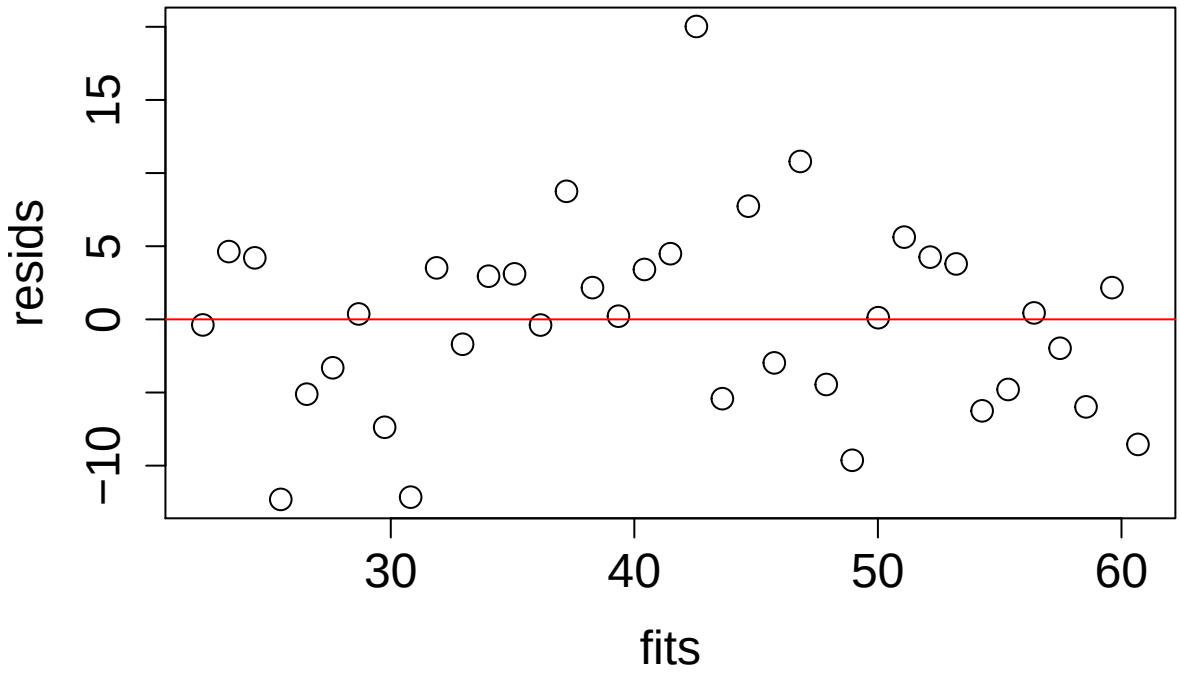
```
plot((SeaIce$Extent)^2~SeaIce$t)
```



The plot does not show significant difference after we squaring *Extent*.

d. (2 points: 1 point for plot, 1 point for comment)

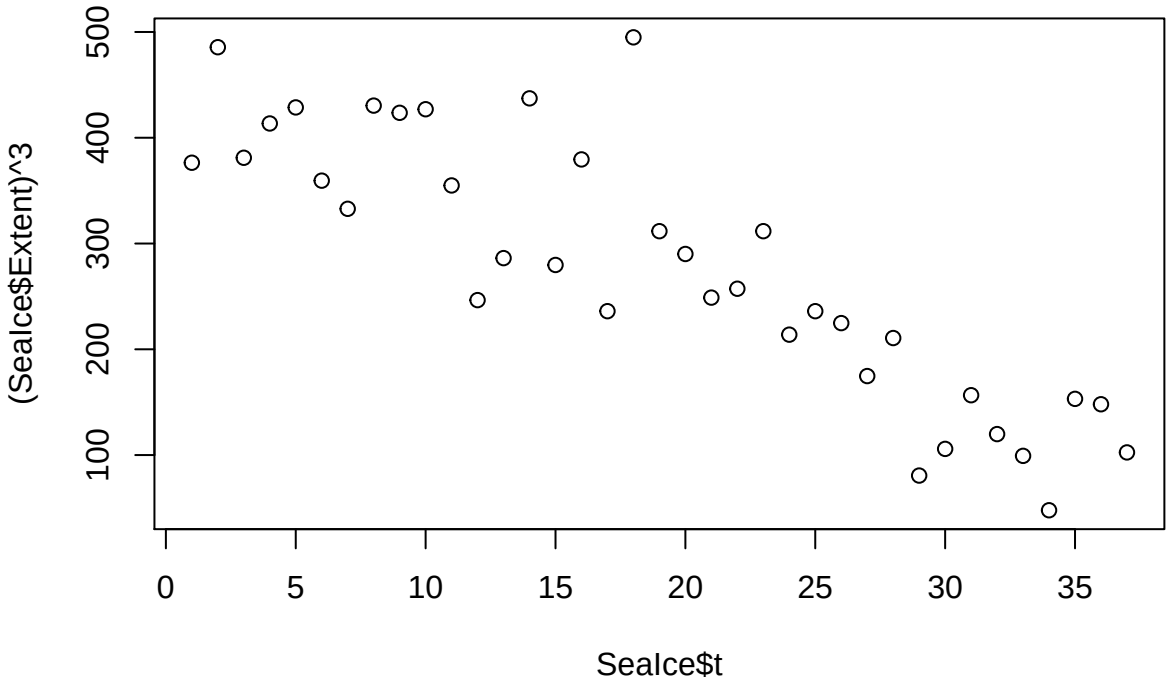
```
model <- lm(Extent^2~t, data=SeaIce)
fits <- model$fitted.values
resids <- model$residuals
plot(resids~fits,cex.lab=1.5,cex.axis=1.5,cex=1.5)
abline(0,0,col="red")
```



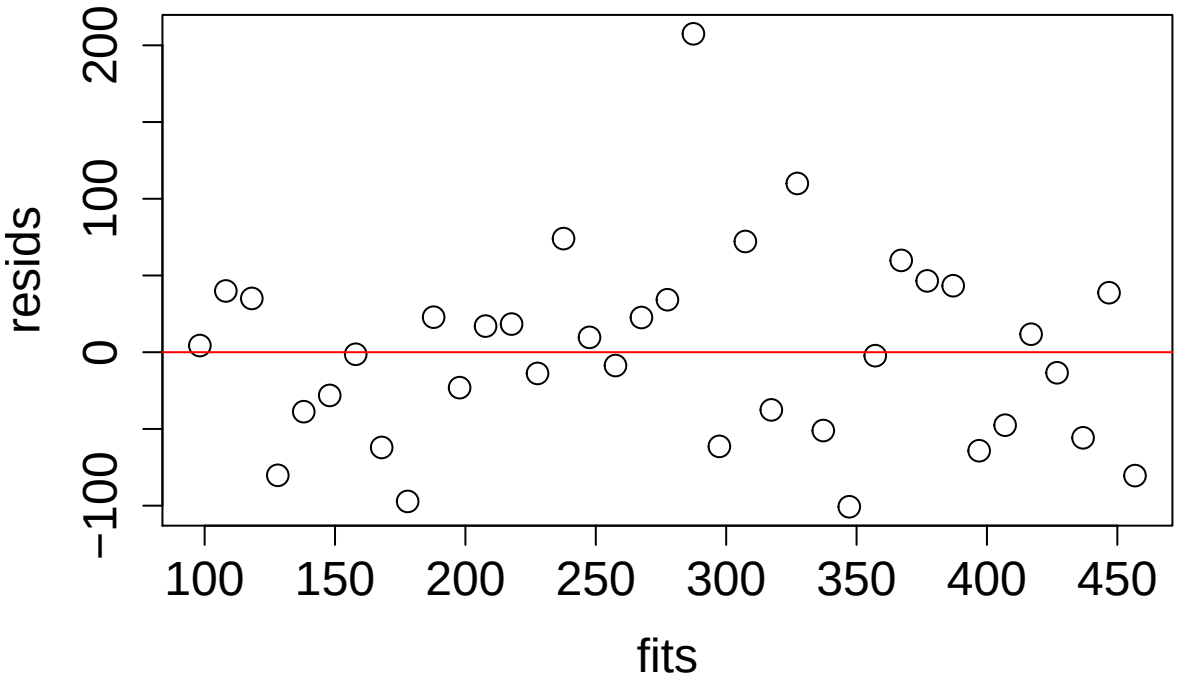
There are larger residuals in the plot compared with the one we created in part (b).

e. (2 points: 1 point for scatterplot, 1 point for residual plot)

```
plot((SeaIce$Extent)^3~SeaIce$t)
```




```
model <- lm(Extent^3~t, data=SeaIce)
fits <- model$fitted.values
resids <- model$residuals
plot(resids~fits,cex.lab=1.5,cex.axis=1.5,cex=1.5)
abline(0,0,col="red")
```



f. (2 point)

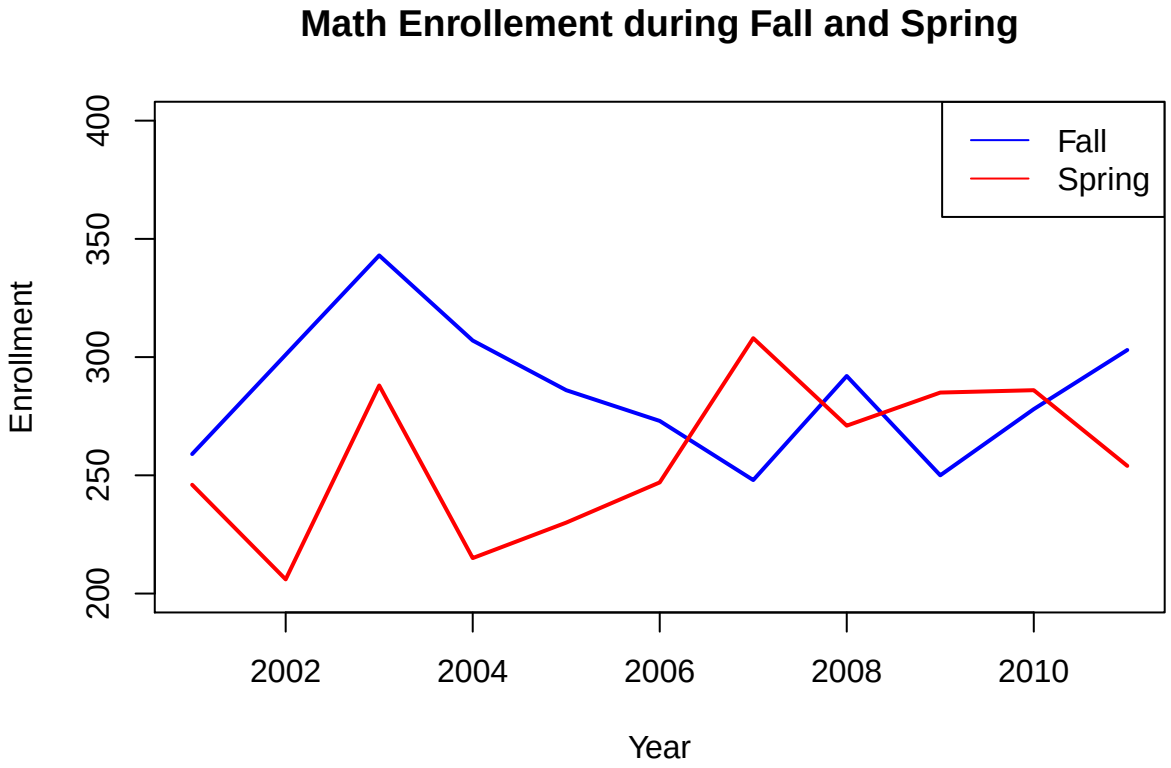
Transforming the responsible variable does not improve the scatterplot, but we are getting larger residuals, showing that the model deviates from the data. We are comfortable using a linear regression for the original

dataset.

Q1.34 (10 points)

a. (3 points: 1 point for plot, 1 point for labels, 1 point for comment)

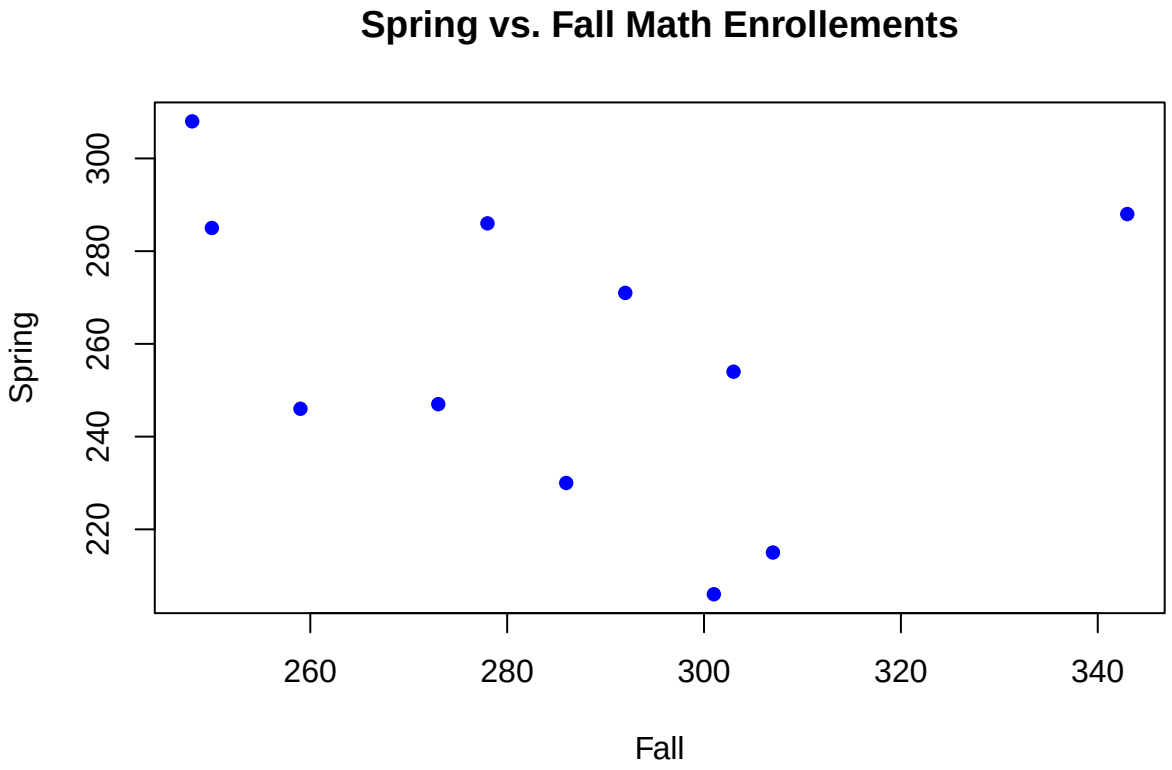
```
data("MathEnrollment")
plot(MathEnrollment$AYear, MathEnrollment$Fall, type='l', col='blue', lwd=2, ylim=c(200,400),
    xlab='Year', ylab='Enrollment', main='Math Enrollement during Fall and Spring')
lines(MathEnrollment$AYear, MathEnrollment$Spring, lwd='2', col='red')
legend('topright', col=c('blue', 'red'), legend=c("Fall", "Spring"), lty=c(1,1))
```



From an overall context the trend is the same. However, one could say that Fall enrollments has a slightly decreasing trend. Spring enrollements has a slightly increasing trend.

b. (2 points: 1 point for plot, 1 point for comment)

```
plot(MathEnrollment$Fall, MathEnrollment$Spring,
     main='Spring vs. Fall Math Enrollements', pch=16, col='blue',ylab='Spring', xlab='Fall')
```



No, the overall association is negative, but weak. Fitted model confirms it too.

c. (1 point)

Point on the top right hand corner seems to be an influential point since, the Fall enrollment is much higher compared to Spring enrollment.

d. (3 points: 1 for each fitted model, 2 point for the comment)

```
# All points
m1 <- lm(Spring~Fall, data=MathEnrollment)
summary(m1)
```

```
##
## Call:
## lm(formula = Spring ~ Fall, data = MathEnrollment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.740 -24.050   1.913  20.674  48.978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 351.0585   106.4710   3.297  0.00927 **
## Fall         -0.3266     0.3713  -0.880  0.40195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 33.09 on 9 degrees of freedom
## Multiple R-squared:  0.07916,    Adjusted R-squared:  -0.02315
## F-statistic: 0.7737 on 1 and 9 DF,  p-value: 0.4019
```

```r
# Without influential point
MathEnrollment2 <- MathEnrollment[MathEnrollment$AYear!=2003,]
m2  <- lm(Spring~Fall, data=MathEnrollment2)
summary(m2)
```

```
##
## Call:
## lm(formula = Spring ~ Fall, data = MathEnrollment2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.500 -17.353  -6.058  22.711  29.418
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 548.0094   106.7236   5.135 0.000891 ***
## Fall         -1.0483     0.3805  -2.755 0.024870 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.94 on 8 degrees of freedom
## Multiple R-squared:  0.4868, Adjusted R-squared:  0.4227
## F-statistic: 7.589 on 1 and 8 DF,  p-value: 0.02487
```

With all points

$$\hat{\text{Spring}} = 351.0585 - 0.3266 \times \text{Fall}$$

After removing outliner:

$$\hat{\text{Spring}} = 548.0094 - 1.0483 \times \text{Fall}$$

When we remove the data point associated with year 2003, we can see a stronger relationship and there is significant change in slope estimate. Thus, we could consider 2003 as influential point.