Day4

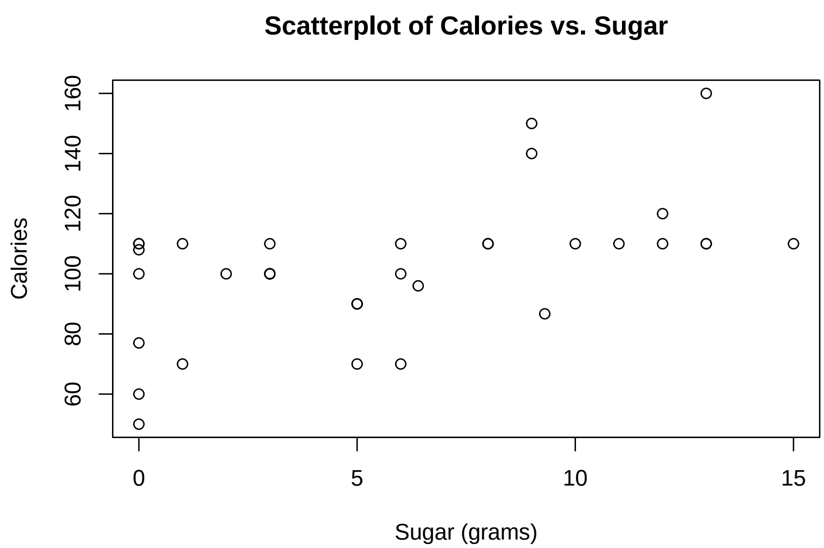
Olivia Wu

2024-02-15

Problem 1.19

a) The scatterplot of Calories vs. Sugar is pretty random and shows no particular trend.

```
plot(Calories~Sugar, data=Cereal, main="Scatterplot of Calories vs. Sugar", xlab="Sugar (grams)")
```



b) $\widehat{Calories} = 87.428 + 2.481 Sugar$

```
model <- lm(Calories~Sugar, data=Cereal)
summary(model)
```

```
##
## Call:
## lm(formula = Calories ~ Sugar, data = Cereal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.428  -9.832   0.245   8.909  40.322
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  87.4277     5.1627  16.935   <2e-16 ***
## Sugar         2.4808     0.7074   3.507   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.27 on 34 degrees of freedom
## Multiple R-squared:  0.2656, Adjusted R-squared:  0.244
## F-statistic:  12.3 on 1 and 34 DF,  p-value: 0.001296
```

c) On average, an increase of 1 gram of sugar also increases the total calories of a cereal by 2.481 calories.

Problem 1.21

a) $\widehat{Calories} = 87.428 + 2.481 * 10 = \boxed{112.238}$

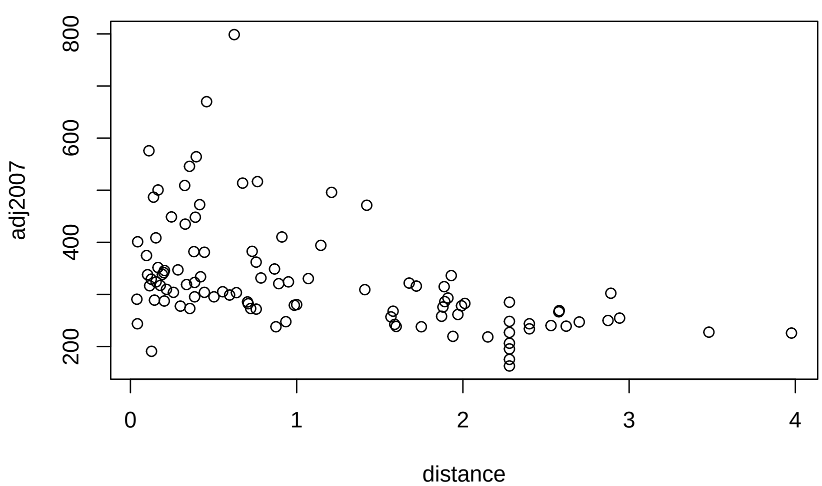
b)

$$\begin{aligned} \text{residual} &= \text{observed} - \text{predicted} \\ &= 110 - (87.428 + 2.481) \\ &= \boxed{20.091} \end{aligned}$$

c) The linear regression model does not appear to be a good summary because there are large residuals.

Problem 1.25

a) There is an overall weak negative relationship between *adj2007* and *distance*. As distance increases, the average value of a house tends to decrease.



b) $\widehat{Adj2007} = 388.204 - 54.427 Distance$

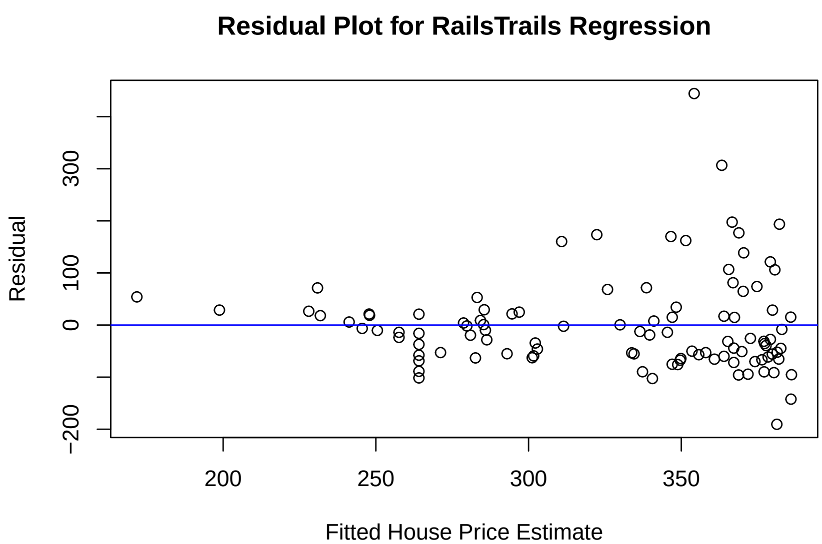
For every mile to the nearest entry point to the rail trail network, we expect the estimated prices of homes in 2007 to decrease by $54,427.

```
##
## Call:
## lm(formula = adj2007 ~ distance, data = RtoT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190.55  -58.19  -17.48   25.22  444.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  388.204     14.052  27.626  < 2e-16 ***
## distance     -54.427      9.659  -5.635 1.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.13 on 102 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2299
## F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

c) The regression standard error is 92.13. If the conditions for the model are met We expect on average, a house price differs from the estimate by $92,130.

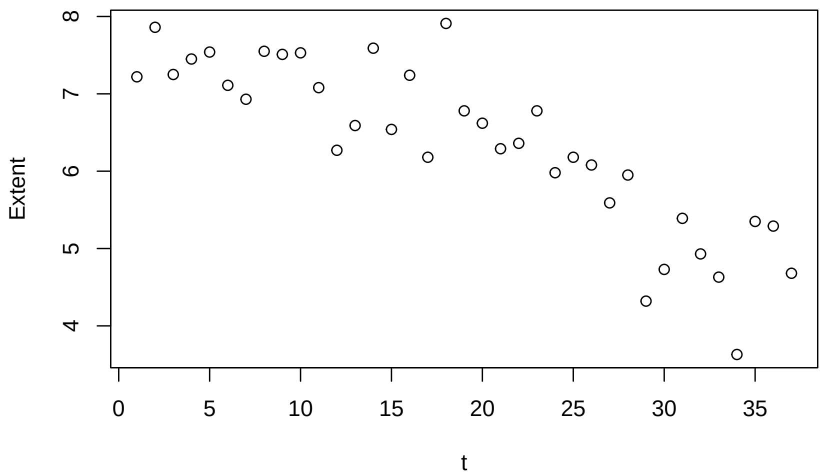
d) Linear: The residual plot below is not scattered randomly.

Constant Variance: Residuals increase as the predicted response increases; there is not a uniform spread.



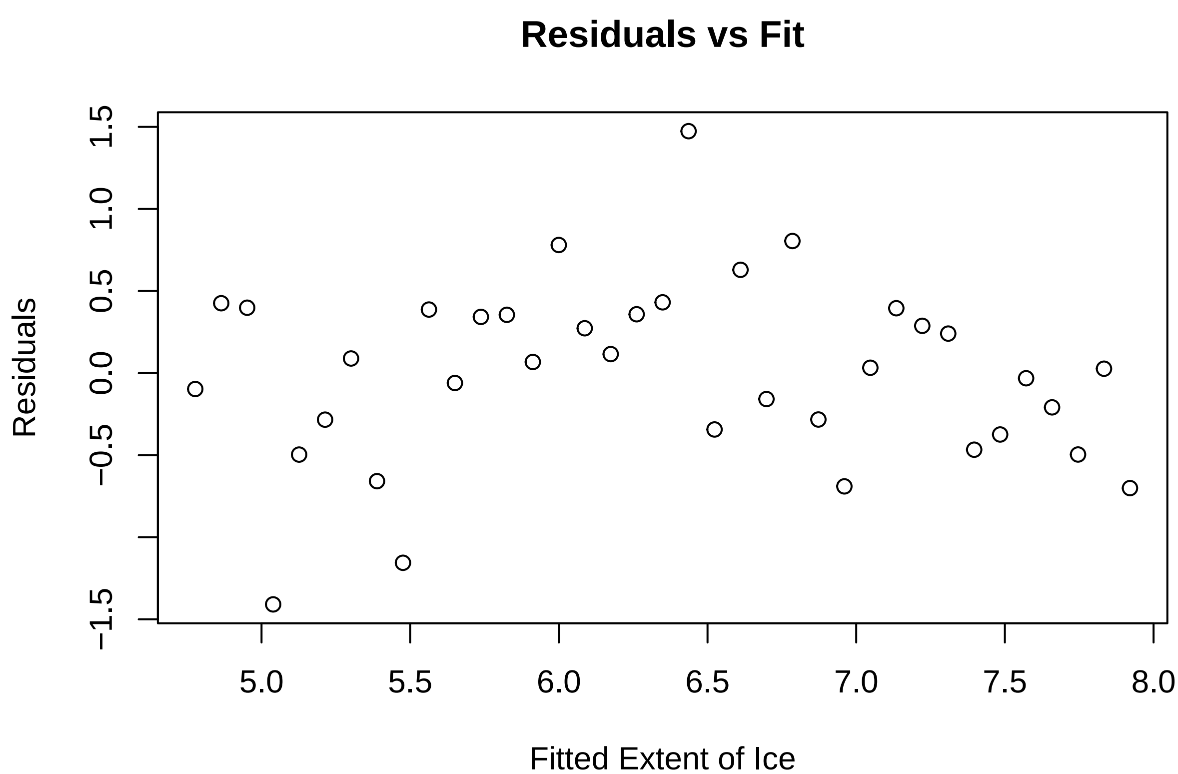
Problem 1.28

a) There is a moderately strong negative relationship between year and extent. As the year increases, the extent of sea ice decreases. There is a slight curve towards the end

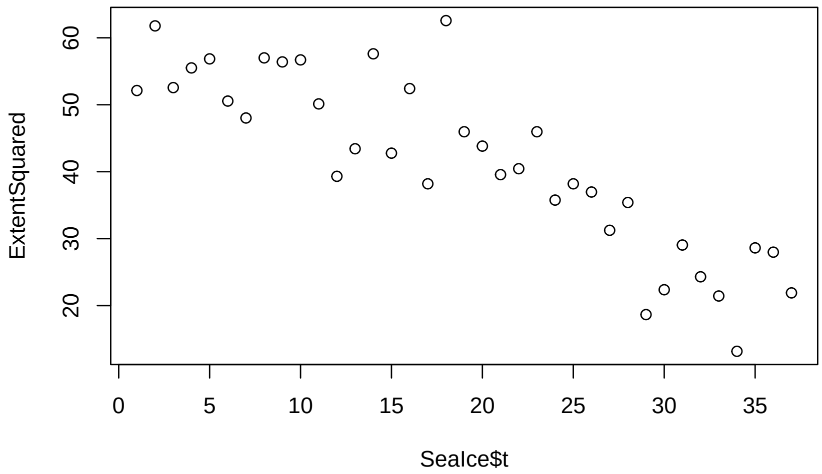



```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'main' will be disregarded
```

b) The residuals spread pretty randomly, so a linear model would be the best fit for this data. There is still some curve

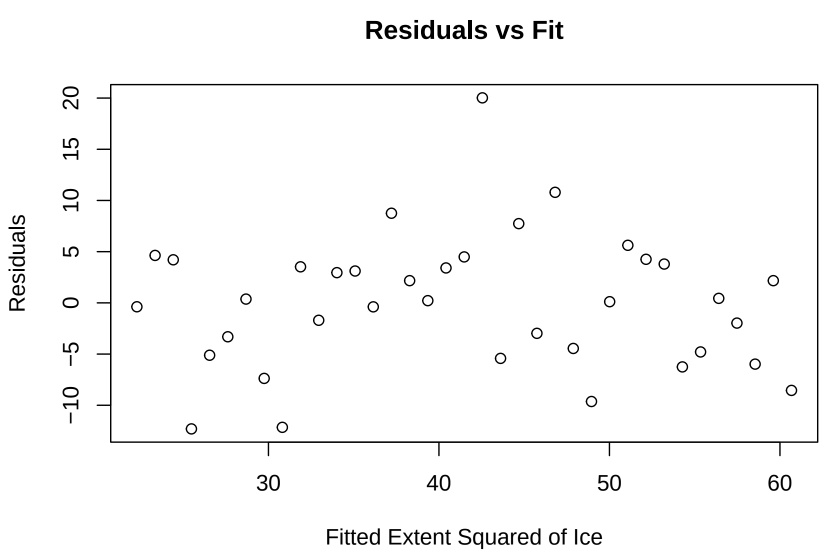


c) There is a strong negative relationship, so when the year increases, the extent of the sea ice decreases. There is less curvature than in part *a)*

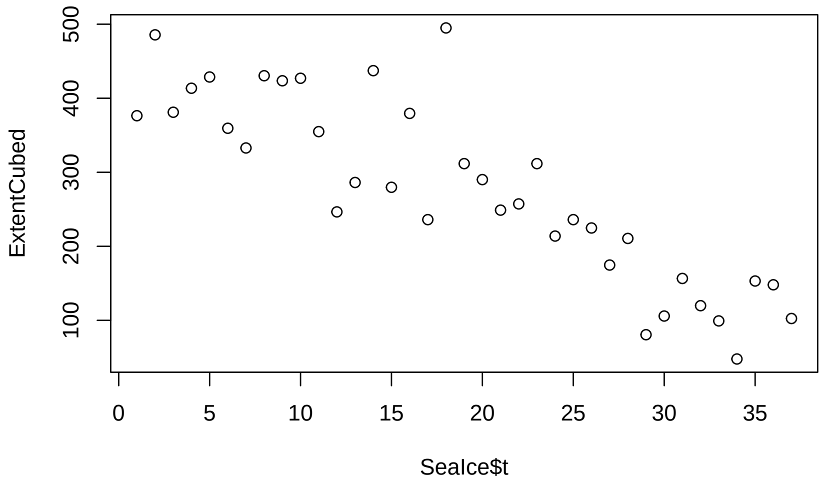



```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'main' will be disregarded
```

d) The residuals are more random and evenly scattered. This shows improvement from part b)

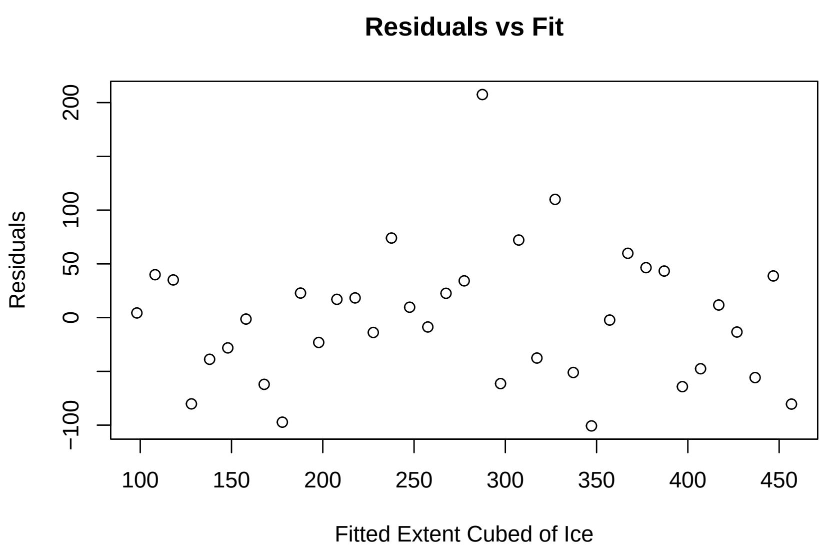


e) There appears to be a strong linear negative relationship between Extent Cubed and the year.




```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'main' will be disregarded
```

The residuals are still evenly and randomly scattered.



f) The model I would be most comfortable with using a linear model is the cubed one because its scatterplot is most linear.