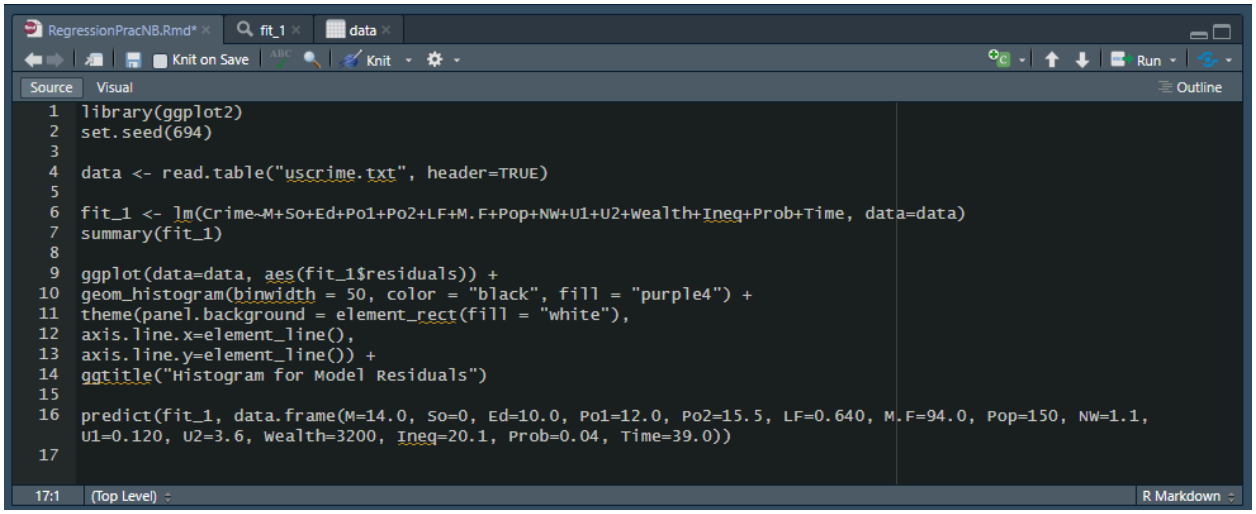
Question 8.1

In the field I intend to pursue, bioinformatics, it is not uncommon to encounter continuous variables that can be used to make inferences on future prospects. One such example is seen in the prediction of gene expression patterns. When predicting gene expression patterns, the intensity of expression of multiple related genes are measured over time and used as predictors for the linear regression model. Below, you will find 5 predictors used in gene expression pattern regression:

1. Height (value of the highest point corresponding to the midpoint of the peak)
2. Distance (midpoint of the peak and transcription start site)
3. Width (difference between the right and left boundaries of the peak)
4. Time (number of days gene expression was measured)
5. Gene Expression Relatedness (picking genes that are similar in function or location of predicted gene)

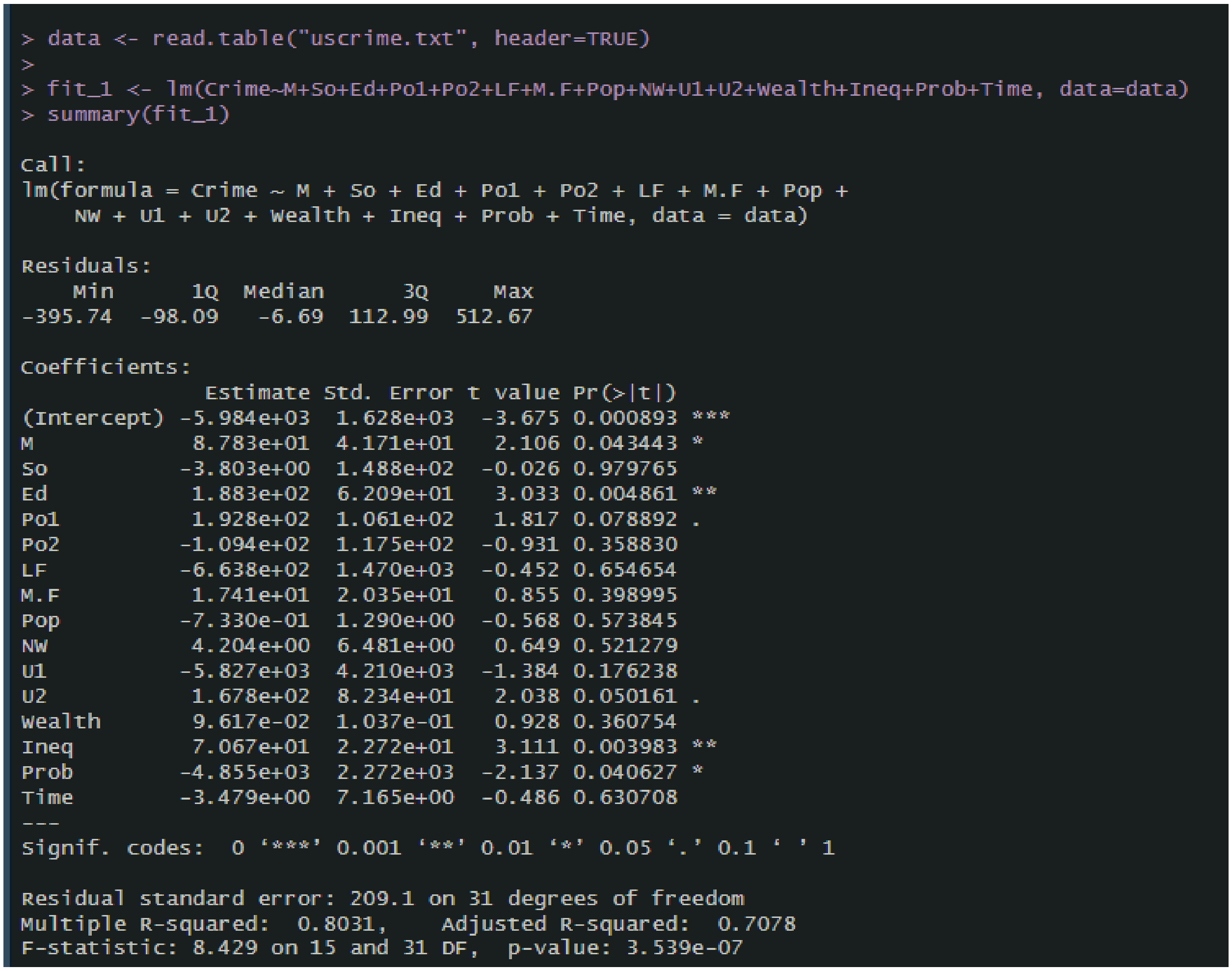
Question 8.2

In the following figures, you will observe exactly how I used the uscrime dataset to create and utilize a linear regression model. Figure 1 shows the code used to generate the results in Figures 2-4. The code itself is fairly self-explanatory as it essentially used every value in the uscrime dataset as a predictor to generate a result for Crime. I did incorporate a histogram of the residuals computed in Figure 2 so that the fitting of the model with the data could be visualized (per https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/). According to the source previously linked, the histogram should have a fairly symmetrical distribution around zero, and in my case this seems to be present. Additionally, they should also possess the quality of being normally distributed around zero, essentially meaning that there should be a bell curve type distribution. Again, the model ran does have this quality as seen in Figure 3. Figure 2 is a snapshot of the coefficients generated from the model. Finally, Figure 4 is the final predicted value of Crime for the given test data based on the aforementioned model. This value seems to be very off target of the other crime values in the dataset. This value is significantly lower than even the lowest data point of 342, with a value of 155. The main reason behind this is due to the overfitting the model is doing with the given dataset. With 16 variables and only 47 observations, such a problem is to be expected. Possible ways of combating this could include reducing the amount of parameters used to calculate the model, obtaining many more observations, or employ some kind of dimensionality reduction.



```
library(ggplot2)
set.seed(694)

data <- read.table("uscrime.txt", header=TRUE)

fit_1 <- lm(Crime~M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time, data=data)
summary(fit_1)

ggplot(data=data, aes(fit_1$residuals)) +
geom_histogram(binwidth = 50, color = "black", fill = "purple4") +
theme(panel.background = element_rect(fill = "white"),
axis.line.x=element_line(),
axis.line.y=element_line()) +
ggtitle("Histogram for Model Residuals")

predict(fit_1, data.frame(M=14.0, So=0, Ed=10.0, Po1=12.0, Po2=15.5, LF=0.640, M.F=94.0, Pop=150, NW=1.1,
U1=0.120, U2=3.6, Wealth=3200, Ineq=20.1, Prob=0.04, Time=39.0))

```

**Figure 1.** This figure depicts the code used to generate the results in the following figures.



```
> data <- read.table("uscrime.txt", header=TRUE)
>
> fit_1 <- lm(Crime~M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time, data=data)
> summary(fit_1)

Call:
lm(formula = Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop +
    NW + U1 + U2 + Wealth + Ineq + Prob + Time, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-395.74  -98.09   -6.69  112.99  512.67

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
M            8.783e+01  4.171e+01   2.106 0.043443 *
So          -3.803e+00  1.488e+02  -0.026 0.979765
Ed           1.883e+02  6.209e+01   3.033 0.004861 **
Po1          1.928e+02  1.061e+02   1.817 0.078892 .
Po2         -1.094e+02  1.175e+02  -0.931 0.358830
LF          -6.638e+02  1.470e+03  -0.452 0.654654
M.F          1.741e+01  2.035e+01   0.855 0.398995
Pop         -7.330e-01  1.290e+00  -0.568 0.573845
NW           4.204e+00  6.481e+00   0.649 0.521279
U1          -5.827e+03  4.210e+03  -1.384 0.176238
U2           1.678e+02  8.234e+01   2.038 0.050161 .
Wealth       9.617e-02  1.037e-01   0.928 0.360754
Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
Time        -3.479e+00  7.165e+00  -0.486 0.630708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.1 on 31 degrees of freedom
Multiple R-squared:  0.8031,	Adjusted R-squared:  0.7078
F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

**Figure 2.** This is a snapshot of the coefficients and residuals of the model. The sections outside of the coefficients detail how well the model fits the data, and the coefficients dictate how the hypothesis (relatedness of predictors to Crime) is supported.

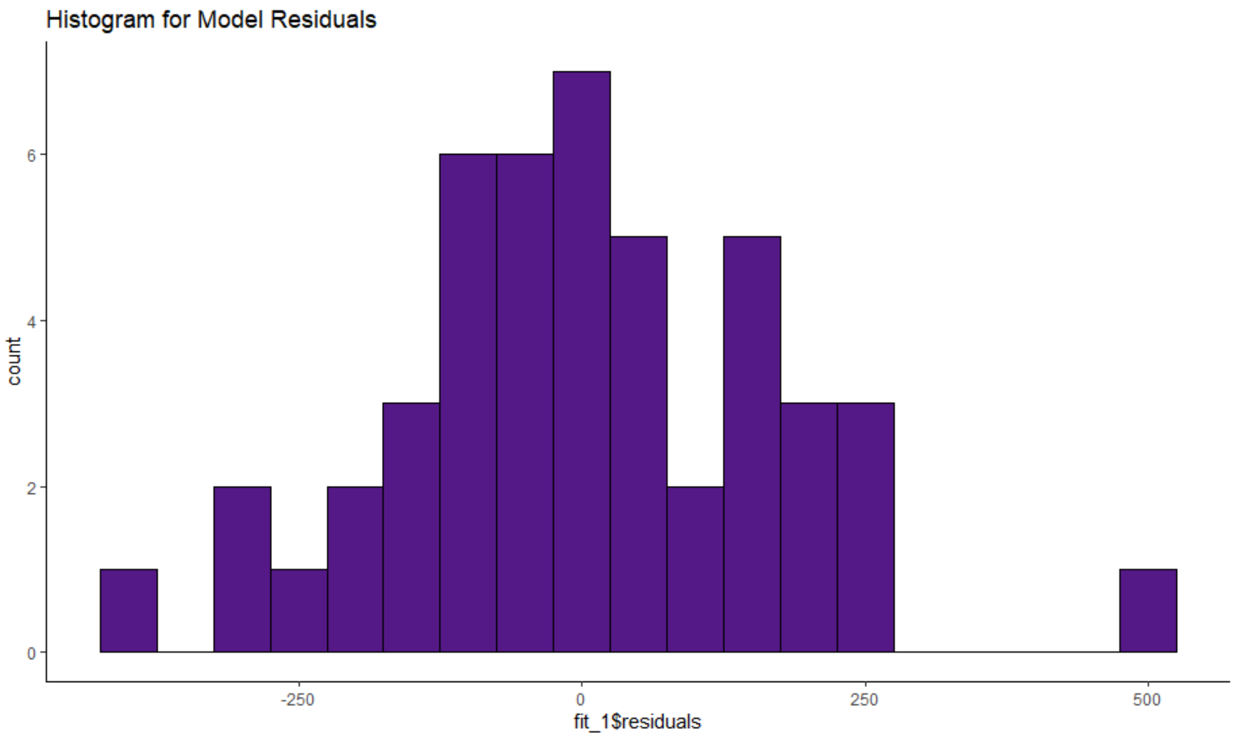
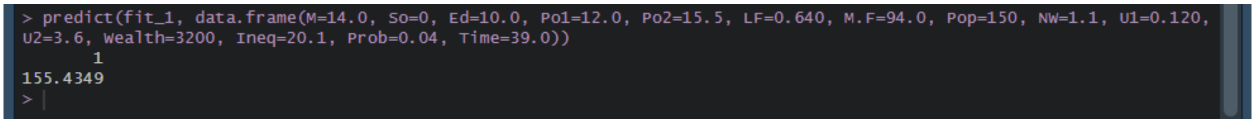


**Figure 3.** This is a histogram of the residuals calculated in Figure 2. The visualization of these values shows more easily how well the model fits the data.



```
> predict(fit_1, data.frame(M=14.0, So=0, Ed=10.0, Po1=12.0, Po2=15.5, LF=0.640, M.F=94.0, Pop=150, NW=1.1, U1=0.120,
U2=3.6, Wealth=3200, Ineq=20.1, Prob=0.04, Time=39.0))
       1
155.4349
> 
```

**Figure 4.** This is the final predicted value of the given test data based on the model generated in Figures 1 & 2.