HW9

Daryl

3/10/2021

Instructions

Problem 1 is associated with your personal dataset project. Submit your work for all problems in a single .pdf compiled with knitr. The Homework Data folder has a starting .Rmd template with spaces for you to fill in your answers to the questions.

Problem 1

This question is related to your personal data set project. The goal for the end of this week is to have at least one column of data uploaded as a .csv to your github repository.

(A) Access at least one column of data (with at least 5 rows), either from one of the sources you identified in problem #1 of your week 7 assignment or another source relevant to your proposed topic.

(B) Write a brief paragraph describing why you decided on this data source, where the data comes from originally, how the chosen data will help you answer a question you are interested in (not necessarily one of the questions you wrote about in Week #7 but hopefully something related), and any processing steps you applied to the data. Submit the paragraph as a response to this question - you do not have to post this to github.

I chose this data as it was in a format that was easily accessible to those who don't wish to recklessly download strange programs onto their laptop, as many cybersecurity-related logs are in the .pcap format. It also consisted of data that was articulated in relatively simplistic terms, with columns that were clearly defined compared to many other cybersecurity logs. One of the most beneficial features of this data set is its inclusion of keywords that I can use to analyze what keywords are most commonly recurring in which type of event.

(C) Upload the data column to your github repo from this project (from week 5) and submit the url as a response to this problem.

https://github.com/notcaughtyet/DATA115/blob/main/CSRIC_Best_Practices.csv

Problem 2

In your own words, please write brief answers to the following:

(A) What are the assumptions for linear regression from a statistical perspective?

The first assumption we make is that it is actually a linear function. Next we assume that the errors are normally distributed, and are not affected by the function itself. Finally we assume that the correlation between the variables we are studying are not due to some other factor.

(B) What is the definition of a residual?

A residual is an error that does not follow the function.

(C) What is the coefficient of determination?

The coefficient of determination is a way of measuring how well our model fits with the data.

(D) What is the difference between the total sum of squares and the model sum of squares?

The total sum of squares is the model sum of squares plus the residual sum of squares.

(E) What do the slope and intercept of the best fit line tell us about the data?

The slope tells us tells us how much one variable changes in relation to another, and the intercept tells us what the base case is for calculating any variable.

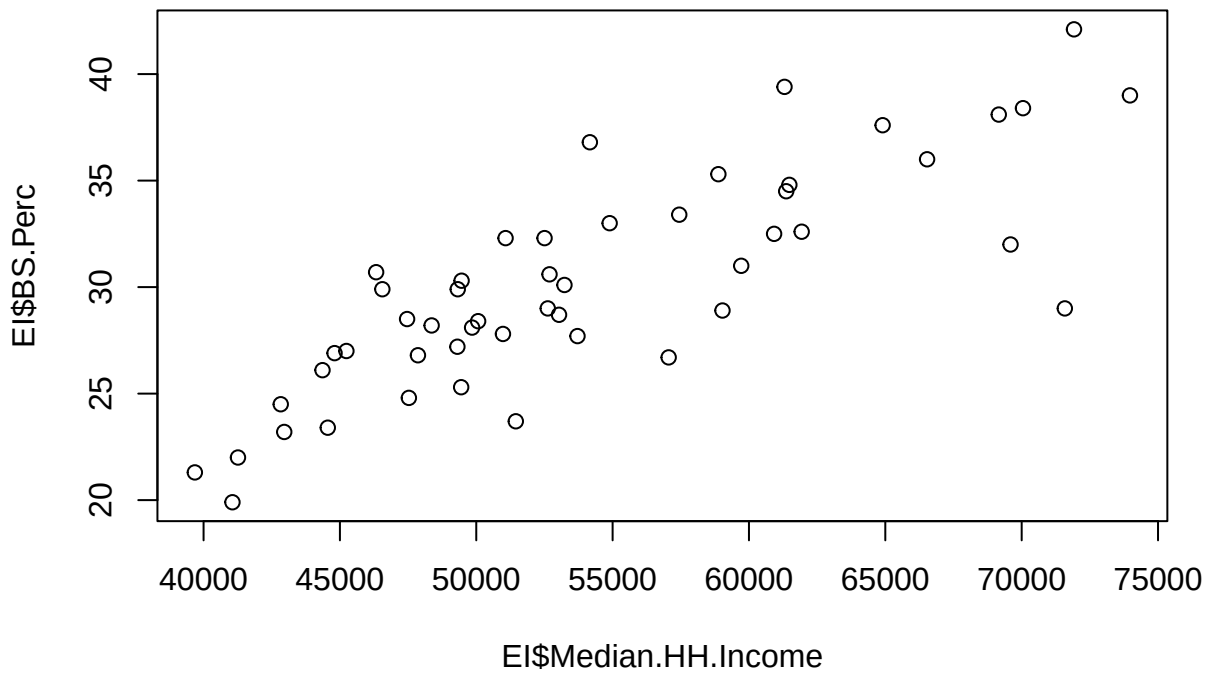
Problem 3

This problem checks your understanding of linear regression and diagnosis of these fits. Start by loading in the education_income.csv file as a dataframe.

```
EI <- read.csv("education_income.csv")
```

(A) Make a scatterplot of the percentage of BS holders by state against the median household (HH) income.

```
plot(EI$BS.Perc ~ EI$Median.HH.Income)
```



(B) Based on this plot, do you think linear regression is appropriate to attempt?

Although this plot isn't a perfect fit, I believe it does follow a linear pattern.

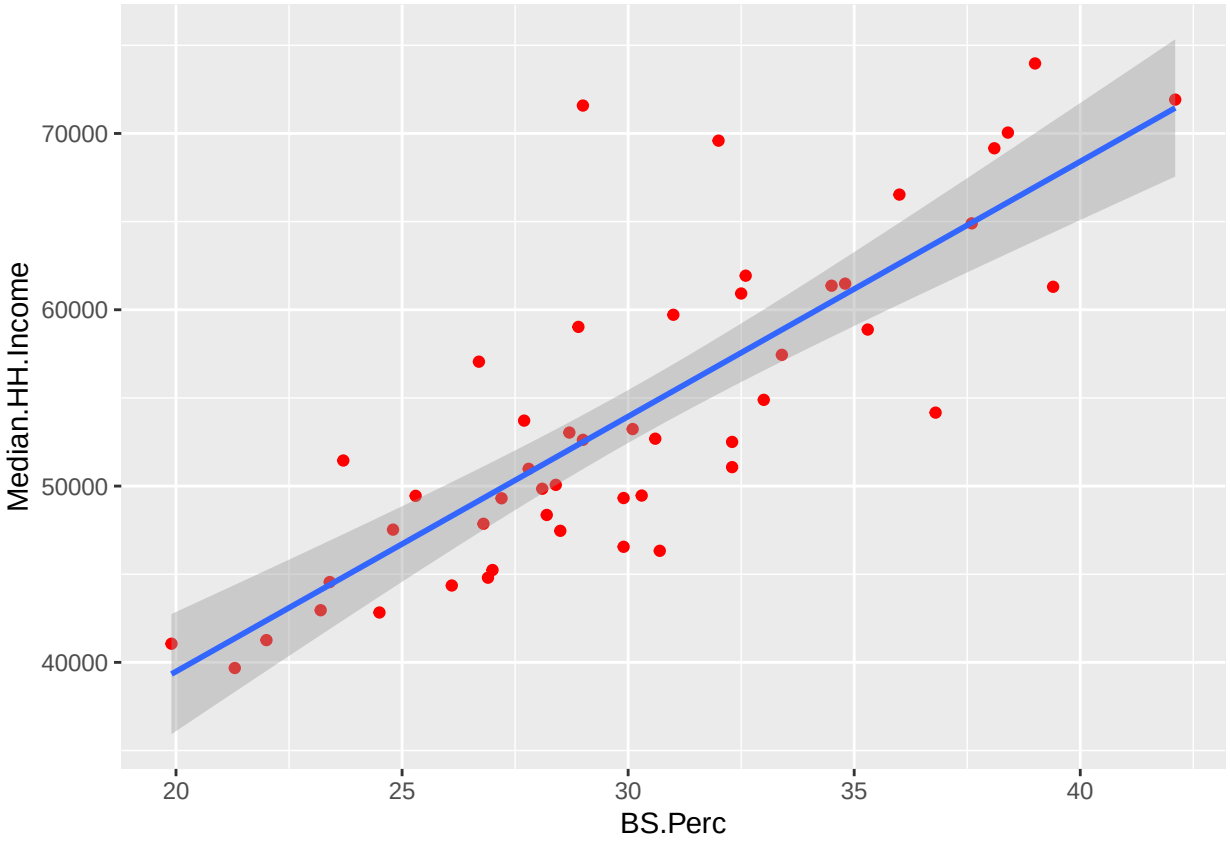
(C) Fit a simple linear model to this pair of columns.

```
linearEI <- lm(Median.HH.Income ~ BS.Perc, EI)
summary(linearEI)
```

```
##
## Call:
## lm(formula = Median.HH.Income ~ BS.Perc, data = EI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9614.5 -3334.7  -721.1  2895.3 19084.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10552       4548   2.320   0.0246 *
## BS.Perc         1446        149   9.709 6.61e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5273 on 48 degrees of freedom
## Multiple R-squared:  0.6626, Adjusted R-squared:  0.6556
## F-statistic: 94.26 on 1 and 48 DF,  p-value: 6.609e-13
```

(D) Overlay the best fit line on the scatterplot from part (A).

```
library(ggplot2)
ggplot(EI, aes(x=BS.Perc,y=Median.HH.Income)) +geom_point(color='red') + geom_smooth(method="lm",formula
```



(E) Write the linear equation estimated by your fit.

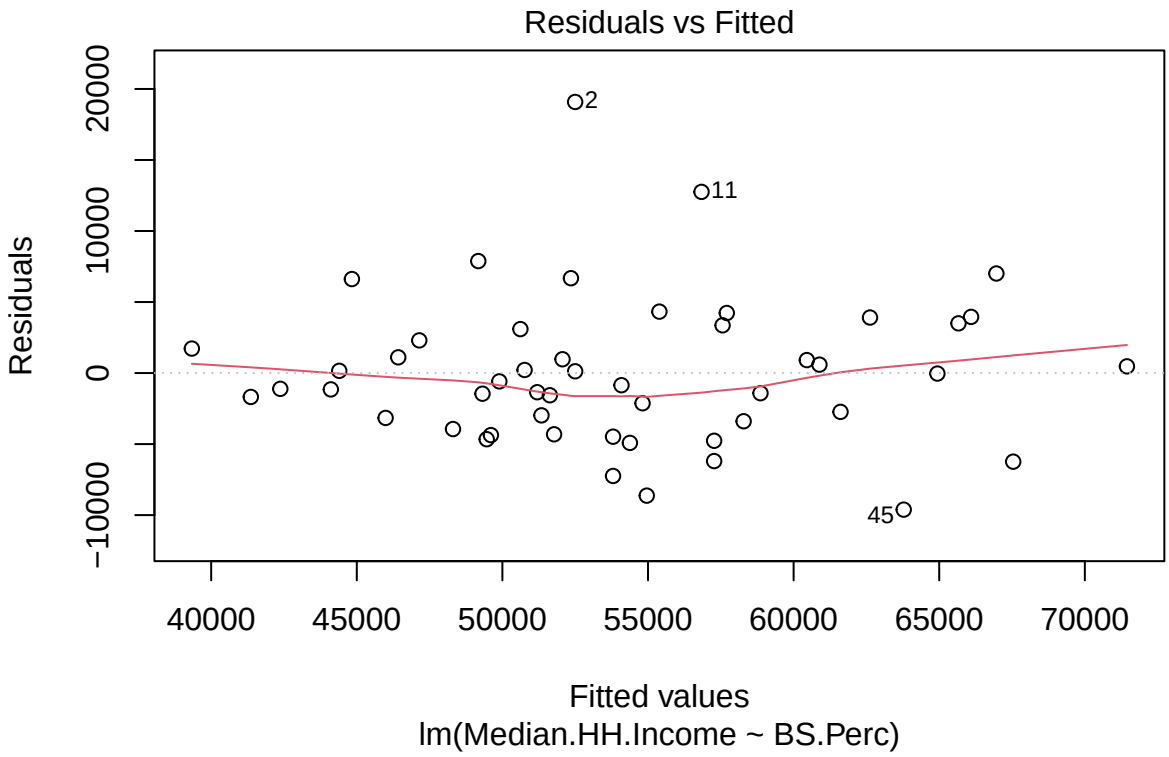
y = 10552 + 1446x

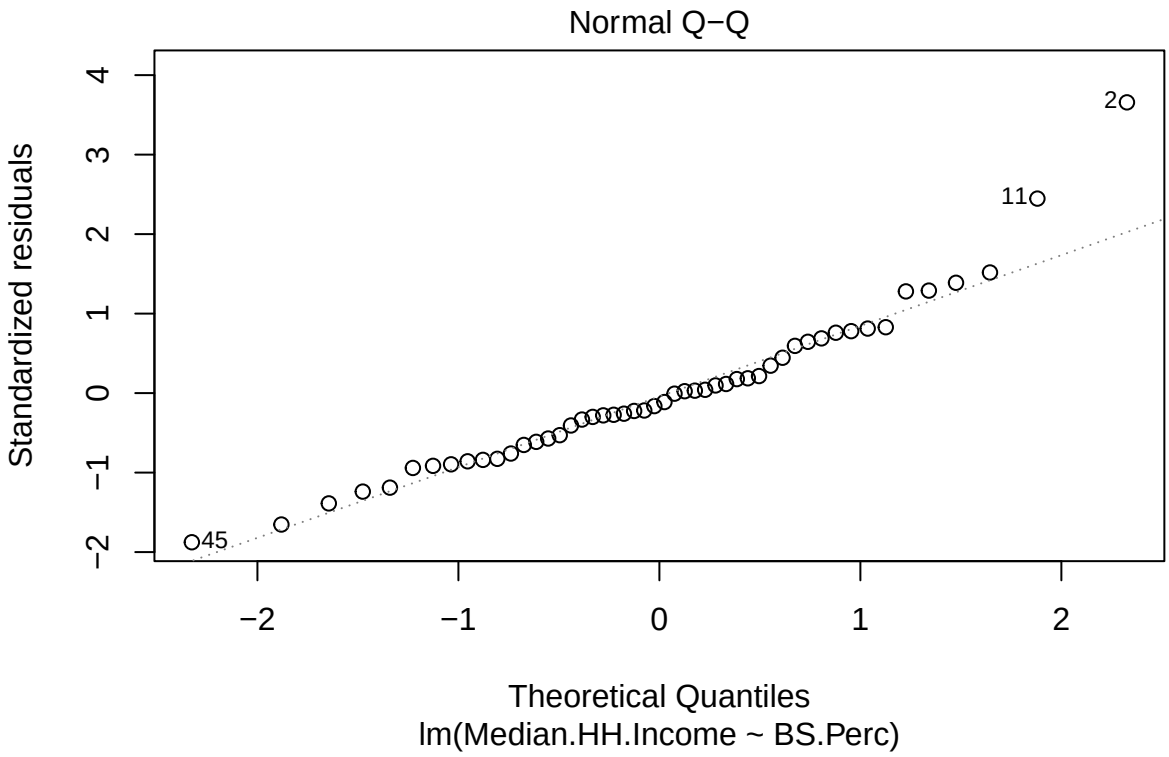
(F) Write the coefficient of determination for your fit.

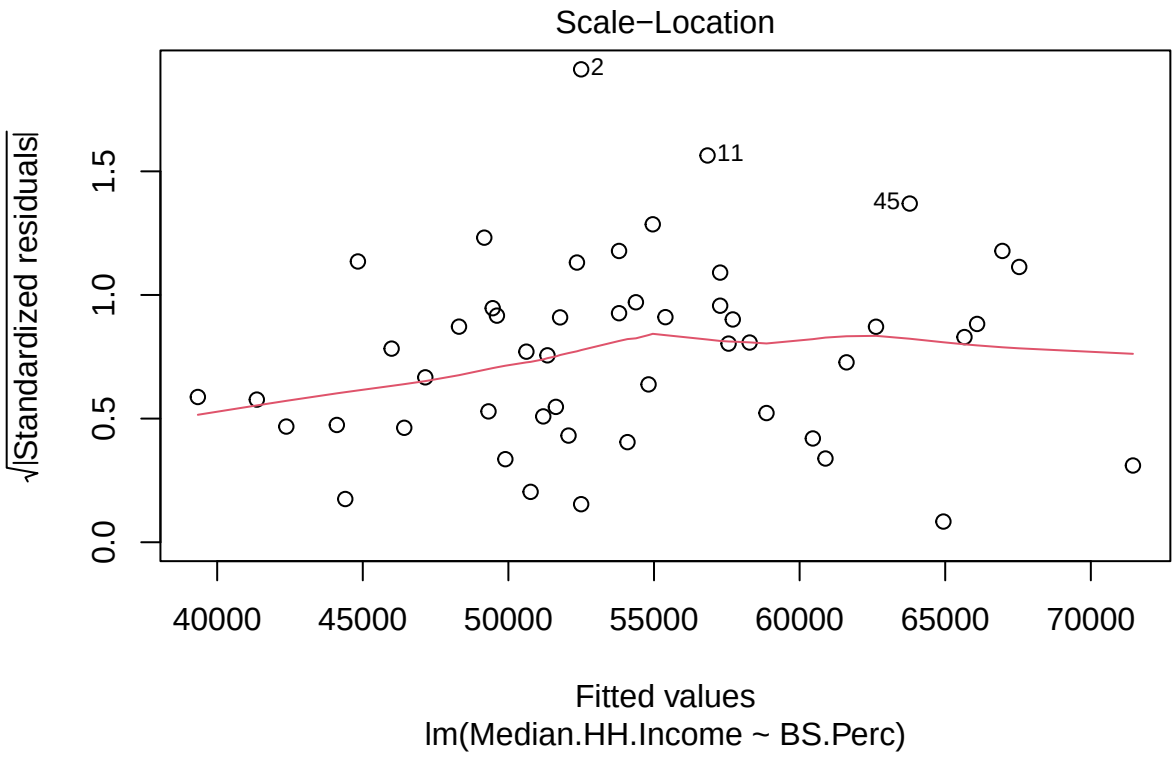
0.6626

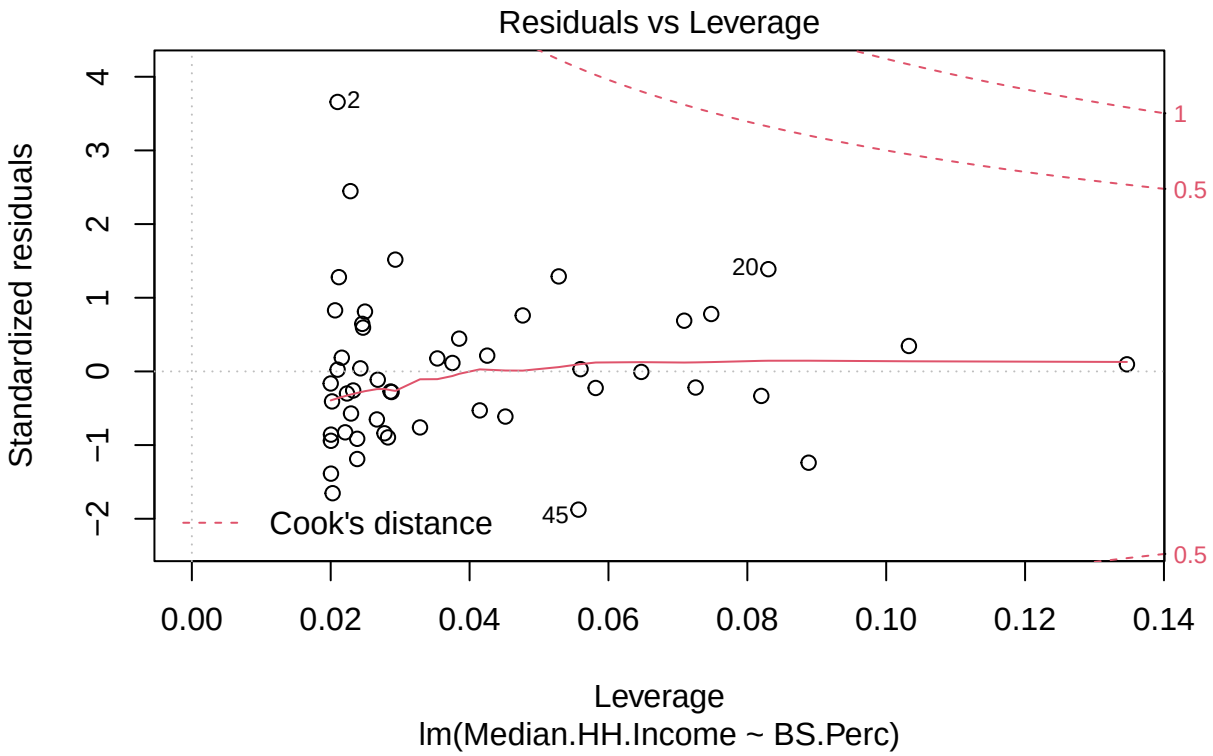
(G) Make a QQ plot of the residuals to check if they are normally distributed. What do you conclude?

```
plot(linearEI)
```


Residuals vs Fitted
Residuals
20000
10000
0
−10000
2
11
45
40000
45000
50000
55000
60000
65000
70000
Fitted values
lm(Median.HH.Income ~ BS.Perc)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Median.HH.Income ~ BS.Perc)
2
11
45
−2
−1
0
1
2
3
4


Scale−Location
2
11
45
√|Standardized residuals|
0.0
0.5
1.0
1.5
40000
45000
50000
55000
60000
65000
70000
Fitted values
lm(Median.HH.Income ~ BS.Perc)



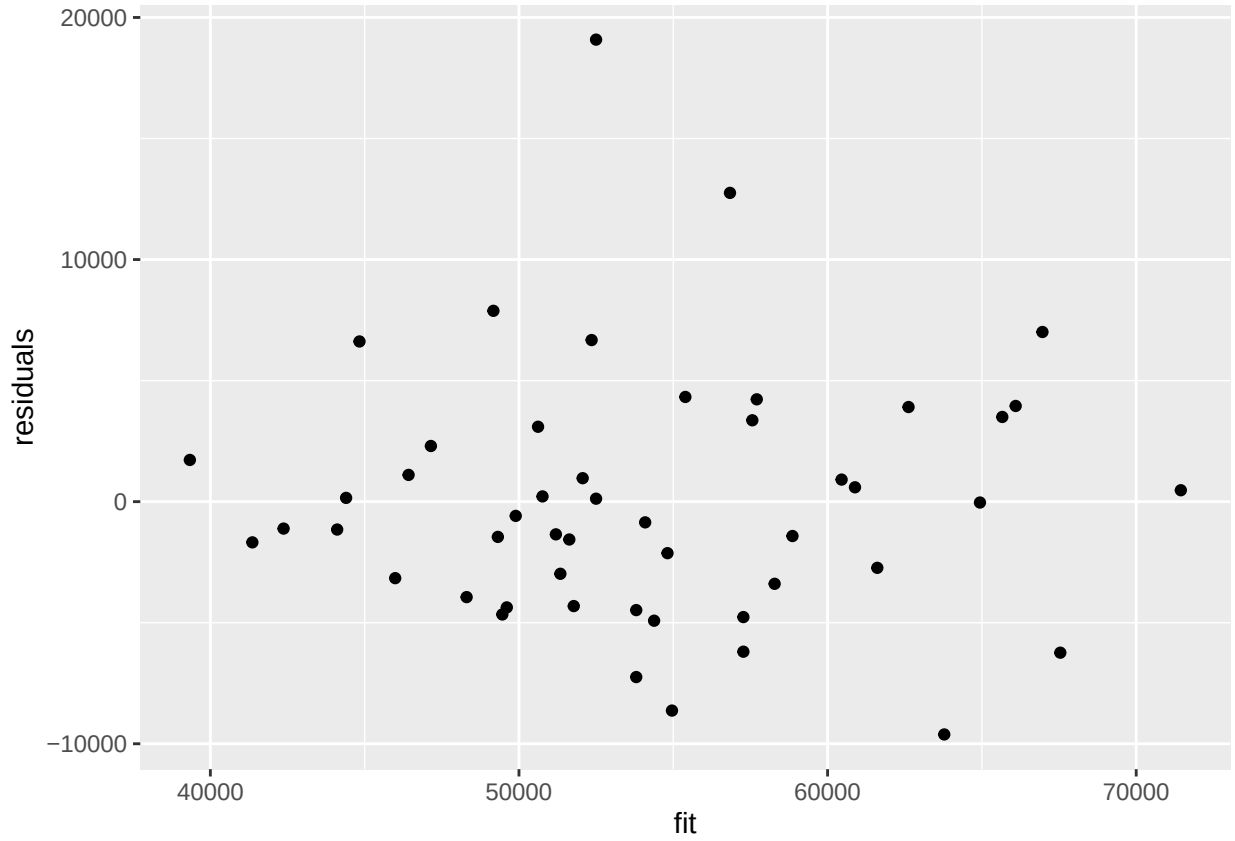
The residuals do seem to be relatively normally distributed up until around the quantile of 0.5. After that, the residuals begin to deviate from the normal line.

(H) Plot the residuals against the fitted values. Do you notice anything concerning?

```
residuals(linearEI)
```

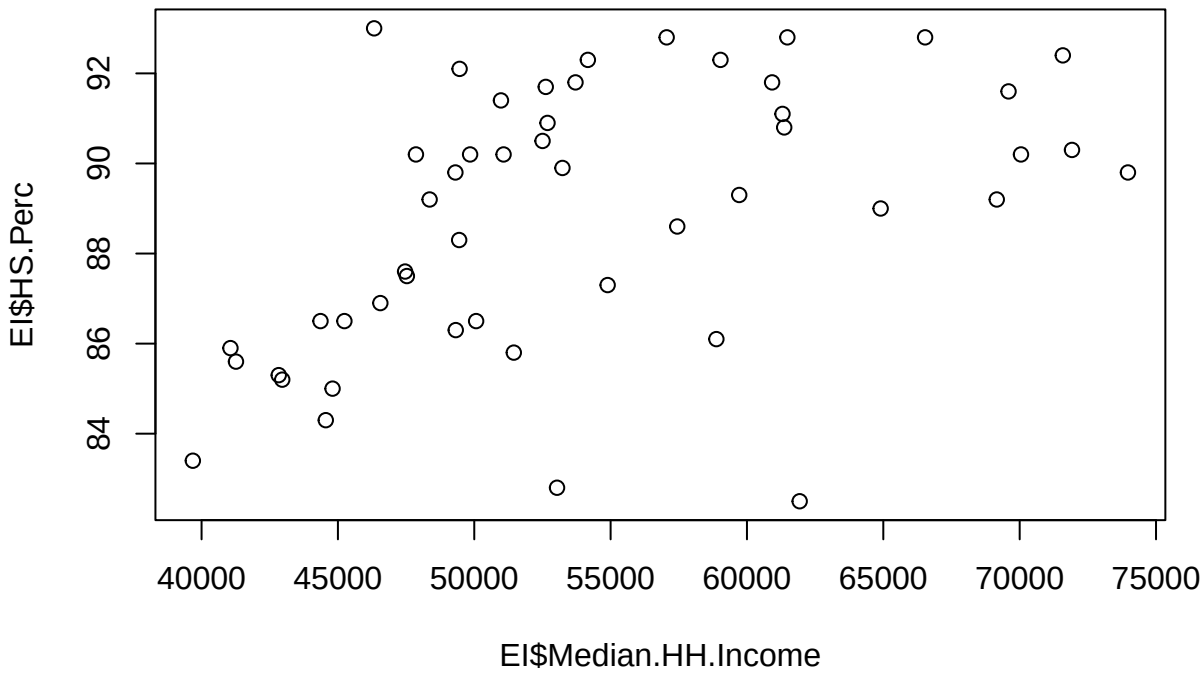
```
##           1           2           3           4           5           6
## -3159.46805 19084.60182 -1562.54083 -1111.39575  4227.45771 -6238.25894
##           7           8           9          10          11          12
##  3953.16998  4324.74398 -4312.18372 -4479.18421 12754.31506 -1455.25456
##          13          14          15          16          17          18
## -1418.68543  2299.38882  3093.95941 -4767.61362 -1151.11045   156.60376
##          19          20          21          22          23          24
## -4916.75578  7008.31263   472.38298 -1349.61216   593.31409 -1680.89551
##          25          26          27          28          29          30
## -2978.25505 -8629.32735 -2126.68445  6617.67509  3908.59938  3499.09865
##          31          32          33          34          35          36
## -4657.89745 -2732.90037 -7244.18421  6675.24471  -586.82613  1105.60327
##          37          38          39          40          41          42
## -6196.61362  -855.47000 -3393.11386 -4367.54035   216.31652 -3942.75432
##          43          44          45          46          47          48
##   970.53049  3361.10060 -9614.54375   -35.68689   912.24276  1723.10498
##          49          50
##   123.60182  7883.38833
```

```
EI$residuals <- linearEI$residuals
fit <- fitted.values(linearEI)
ggplot(EI, aes(x=fit, y=residuals)) +geom_point()
```



(I) Choose two other columns from the dataframe and repeat steps (a)-(h).

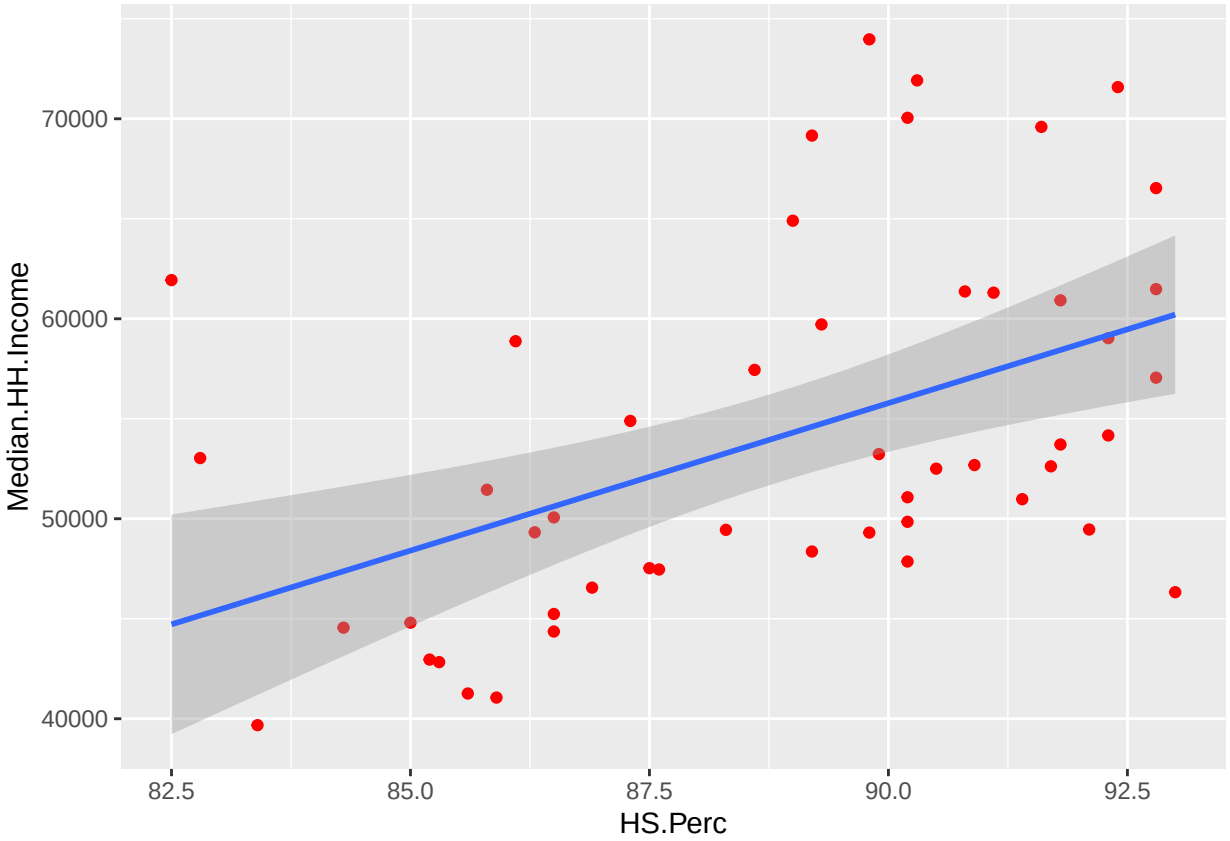
```
plot(EI$HS.Perc ~ EI$Median.HH.Income)
```



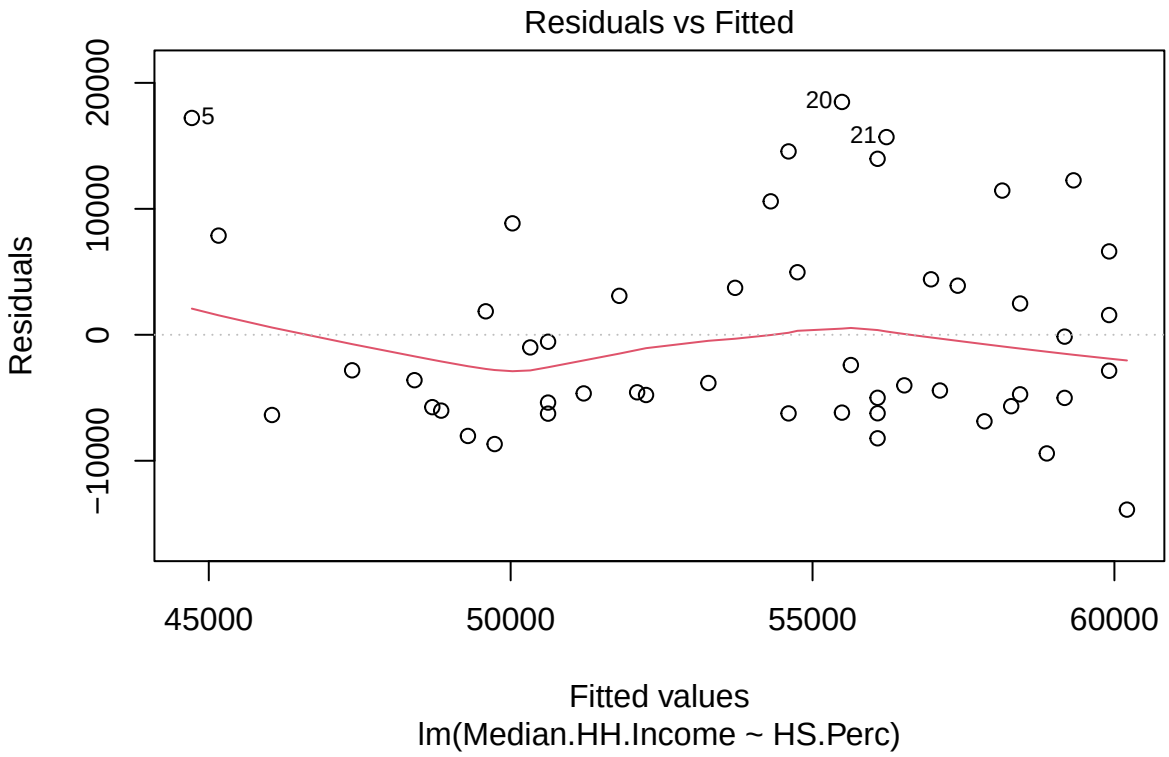

```
linearEIHS <- lm(Median.HH.Income ~ HS.Perc, EI)
summary(linearEIHS)
```

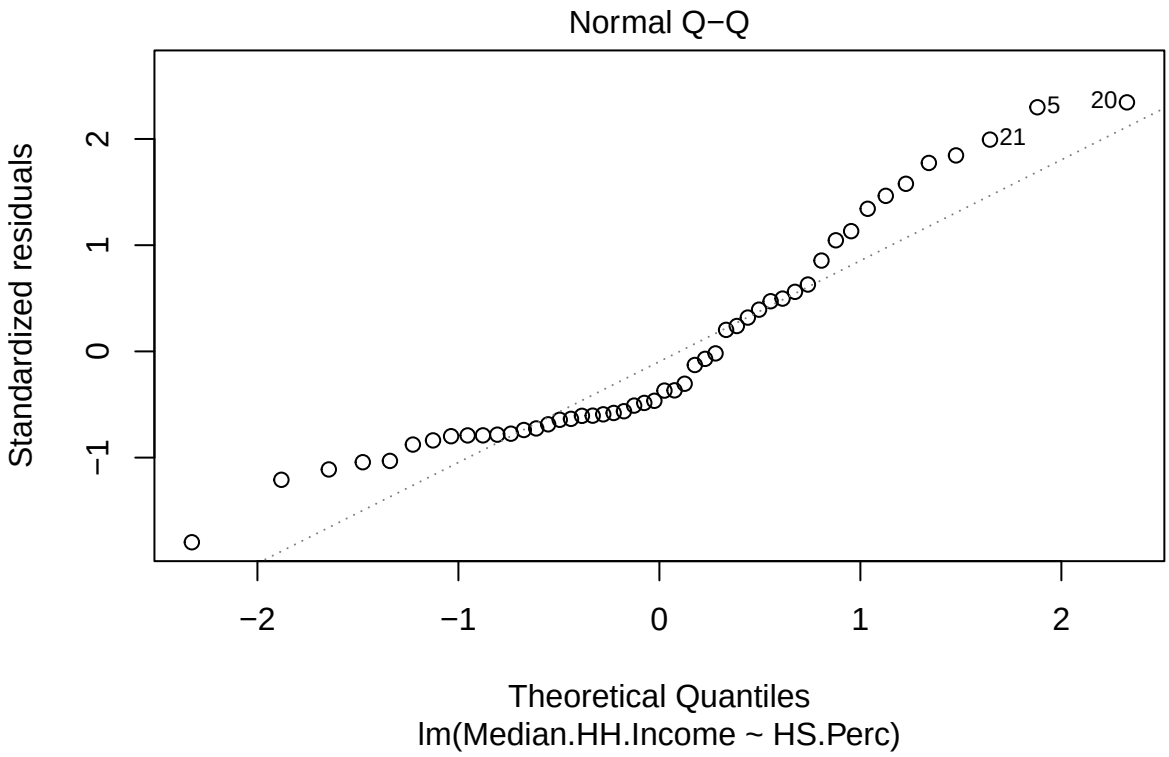
```
##
## Call:
## lm(formula = Median.HH.Income ~ HS.Perc, data = EI)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13880  -5725  -3231   4277  18484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -76973.6    34733.8  -2.216 0.031459 *
## HS.Perc       1475.1      390.7   3.776 0.000439 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7971 on 48 degrees of freedom
## Multiple R-squared:  0.229,  Adjusted R-squared:  0.2129
## F-statistic: 14.26 on 1 and 48 DF,  p-value: 0.0004393
```

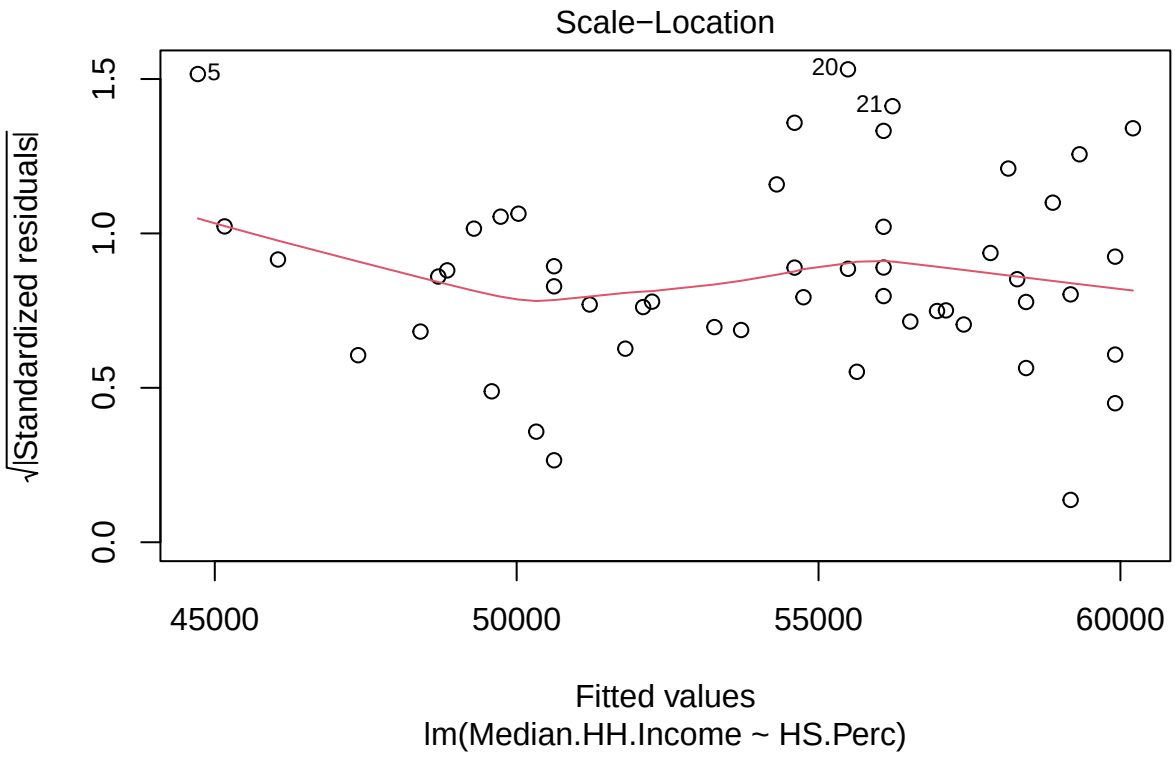
```
ggplot(EI, aes(x=HS.Perc,y=Median.HH.Income)) +geom_point(color='red') + geom_smooth(method="lm",formula
```

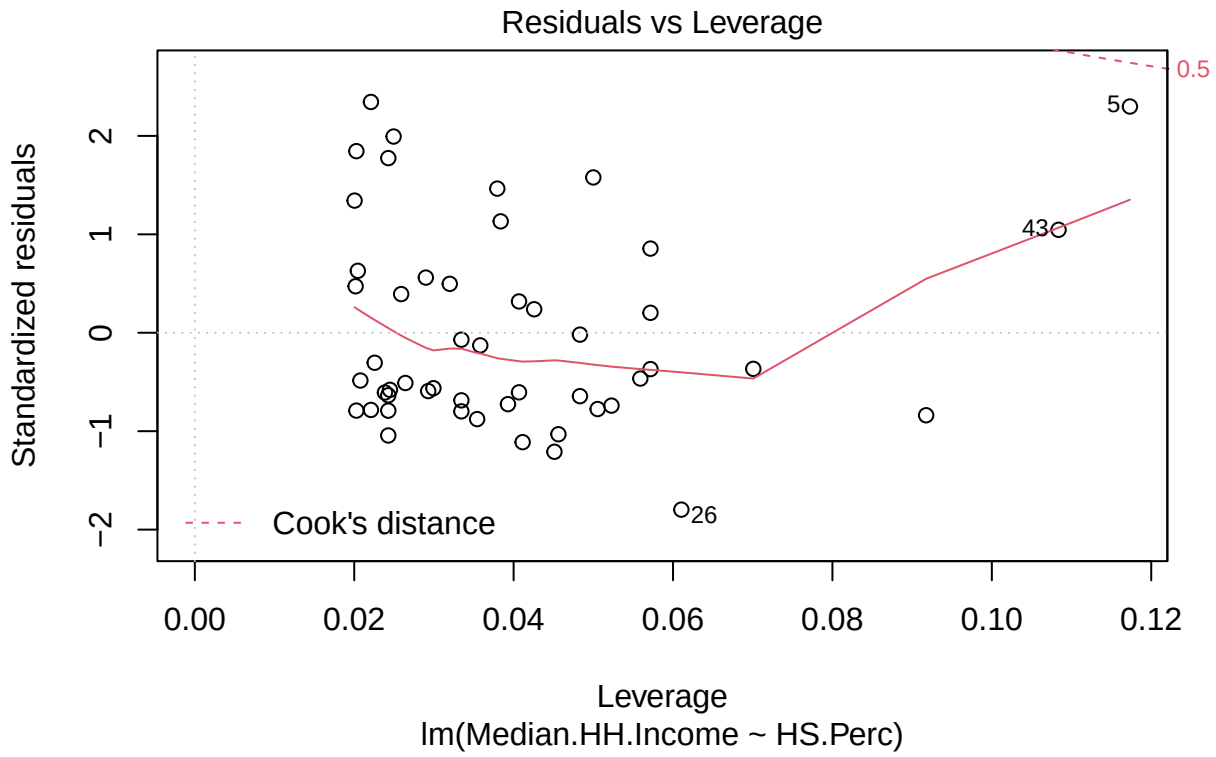



```
plot(linearEIHS)
```


Residuals vs Fitted
Residuals
20000
10000
0
−10000
5
20
21
45000
50000
55000
60000
Fitted values
lm(Median.HH.Income ~ HS.Perc)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Median.HH.Income ~ HS.Perc)
5
20
21
−2
−1
0
1
2

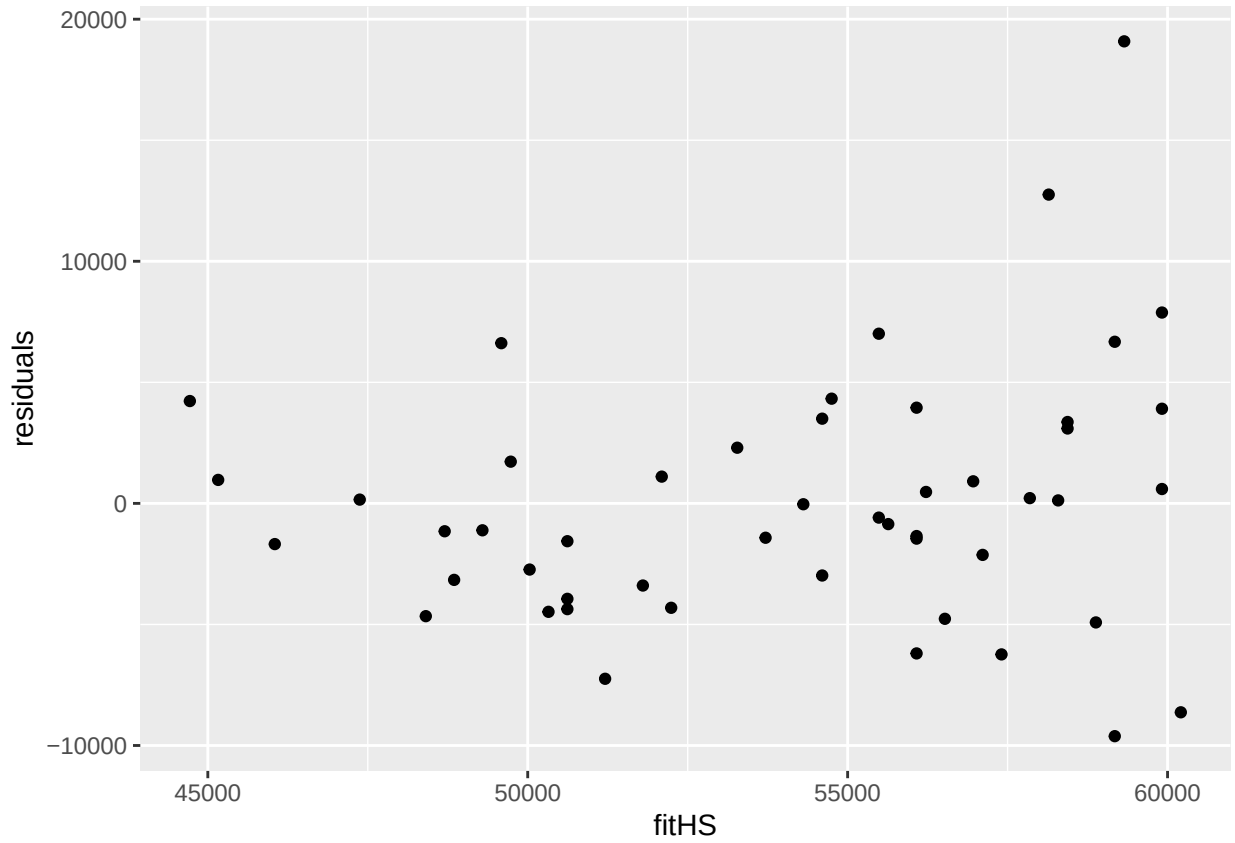

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
20
21
45000
50000
55000
60000
Fitted values
lm(Median.HH.Income ~ HS.Perc)



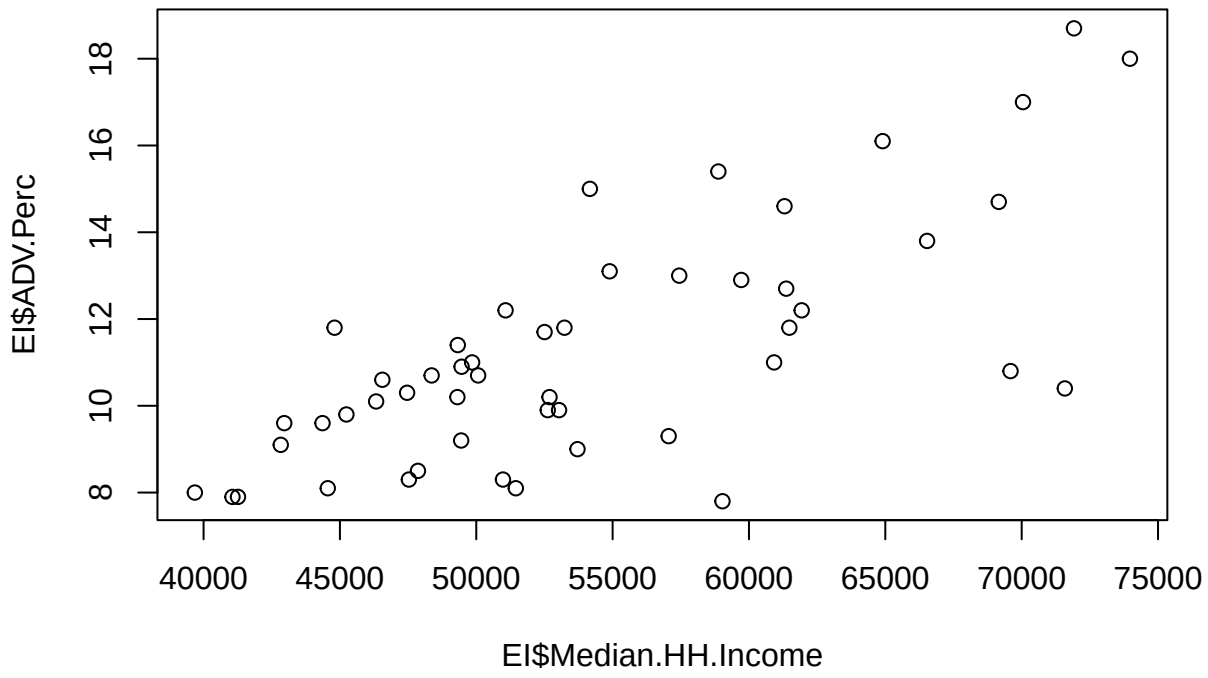

```
residuals(linearEIHS)
```

```
##           1           2           3           4           5           6
##  -6019.6291  12260.3913   -551.7102  -8030.1494  17213.5601   3897.9791
##           7           8           9          10          11          12
##  13970.5399   4966.1007  -4779.2845  -1003.6966  11449.4453  -8216.4601
##          13          14          15          16          17          18
##   3726.6480  -3828.8318  -4725.5682  -4015.9804  -5744.1223  -2819.5615
##          19          20          21          22          23          24
##  -9418.0885  18483.5669  15694.0331  -6230.4601   1568.3642  -6367.0007
##          25          26          27          28          29          30
##  -6239.3926 -13879.6493  -4424.0074   1862.8371   6619.3642  14557.6074
##          31          32          33          34          35          36
##  -3604.1088   8848.3169  -4653.7372   -146.1020  -6179.4331  -4565.7777
##          37          38          39          40          41          42
##  -5002.4601  -2400.9399   3091.2358  -5381.7102  -6868.5412  -6258.7102
##          43          44          45          46          47          48
##   7873.0398   2484.4318  -5009.1020  10594.6209   4403.4993  -8675.6696
##          49          50
##  -5668.0615  -2857.6358
```

```
fitHS <- fitted.values(linearEIHS)
ggplot(EI, aes(x=fitHS, y=residuals)) +geom_point()
```



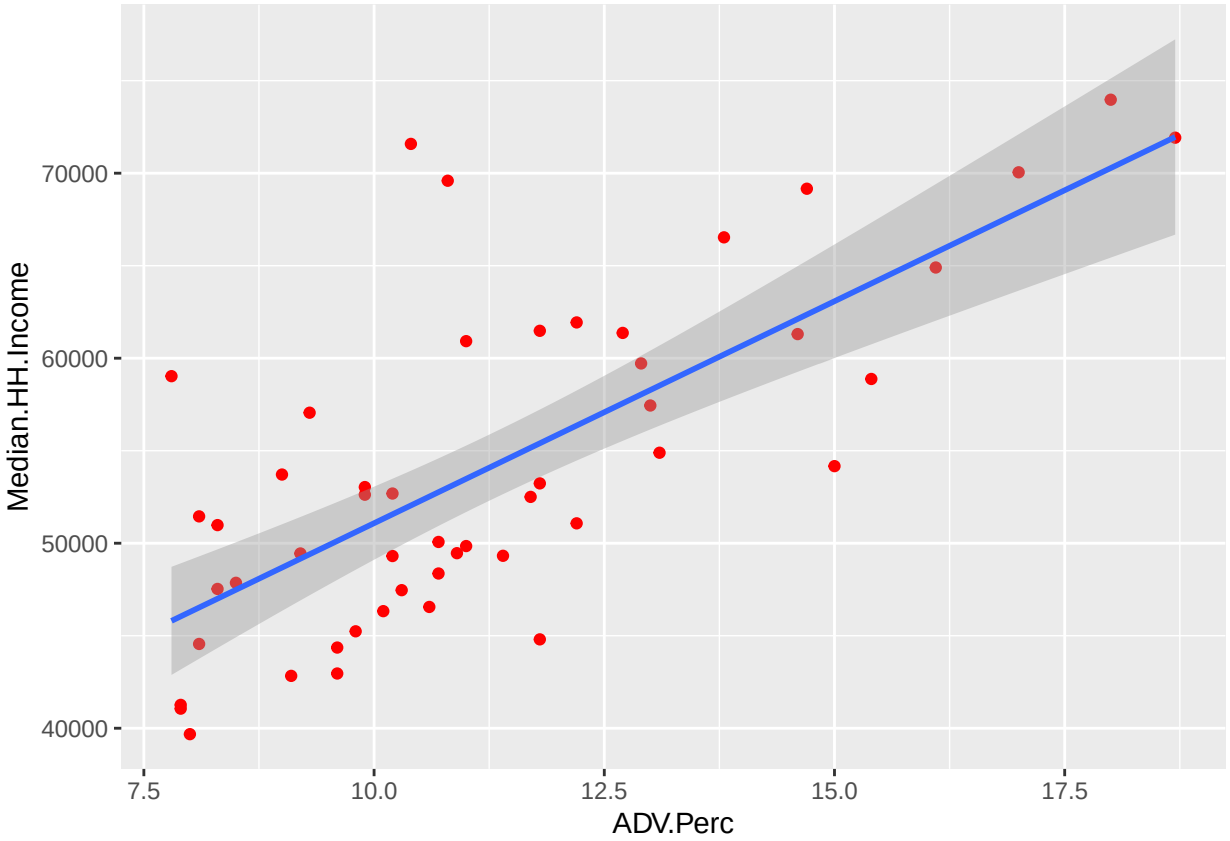

```
plot(EI$ADV.Perc ~ EI$Median.HH.Income)
```



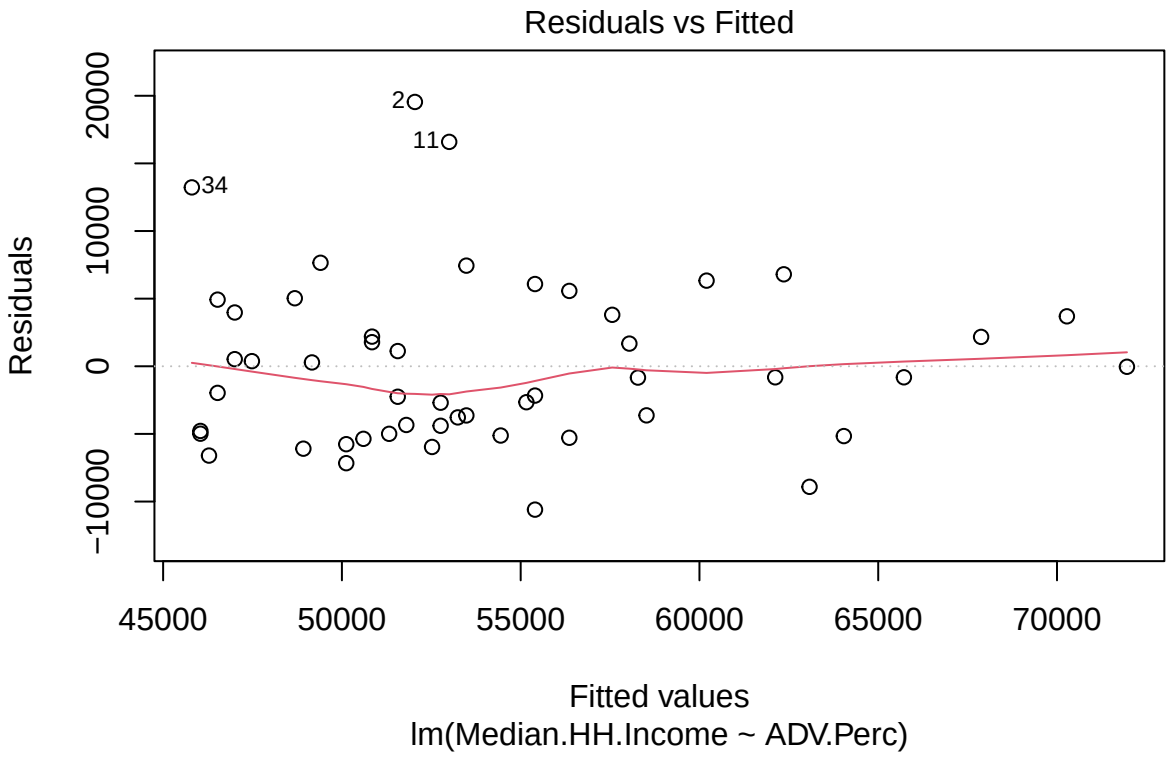

```
linearEIADV <- lm(Median.HH.Income ~ ADV.Perc, EI)
summary(linearEIADV)
```

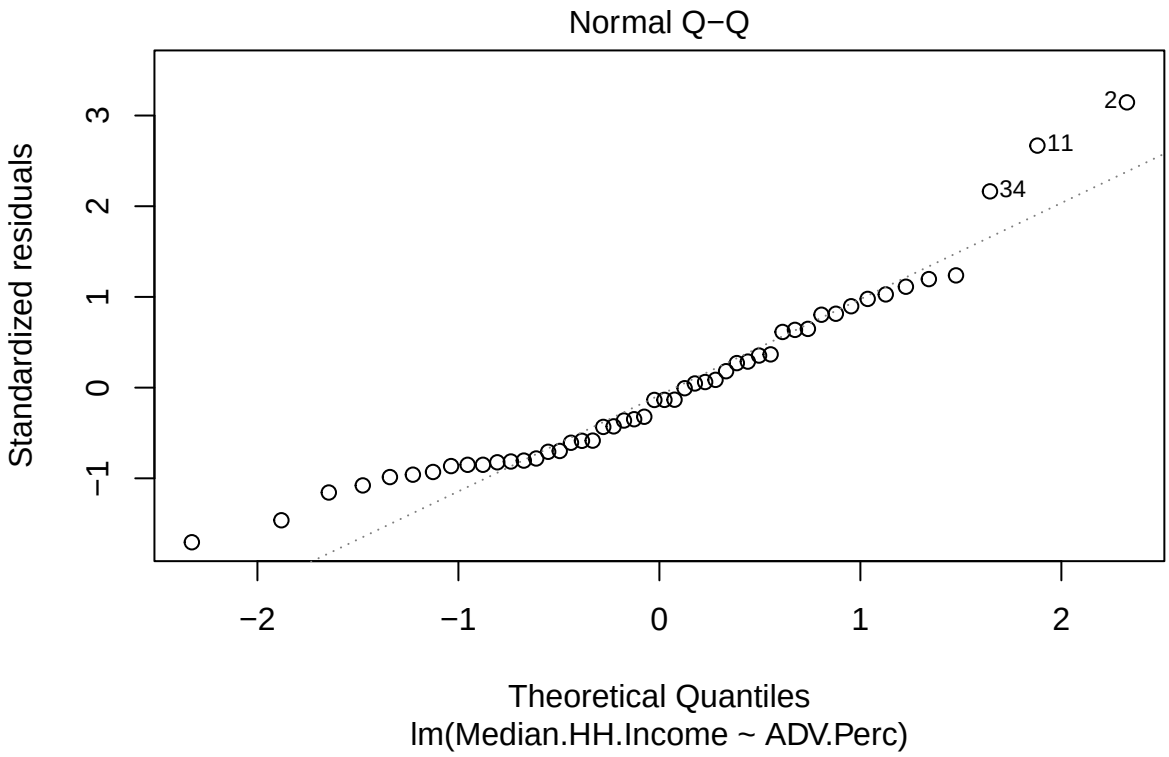
```
##
## Call:
## lm(formula = Median.HH.Income ~ ADV.Perc, data = EI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10597.6  -4933.2   -825.9   3778.0  19541.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    27088       3844   7.048 6.21e-09 ***
## ADV.Perc        2399        332   7.226 3.31e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6283 on 48 degrees of freedom
## Multiple R-squared:  0.521,  Adjusted R-squared:  0.5111
## F-statistic: 52.22 on 1 and 48 DF,  p-value: 3.311e-09
```

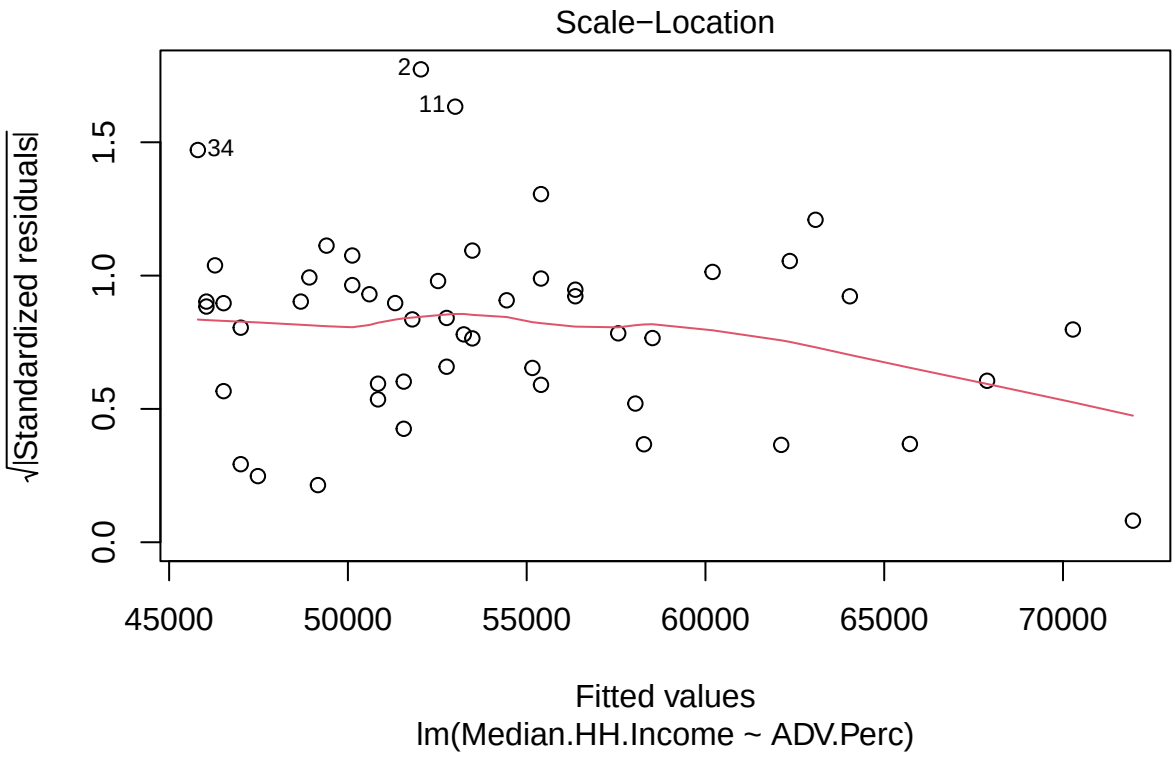
```
ggplot(EI, aes(x=ADV.Perc,y=Median.HH.Income)) +geom_point(color='red') + geom_smooth(method="lm",formul
```

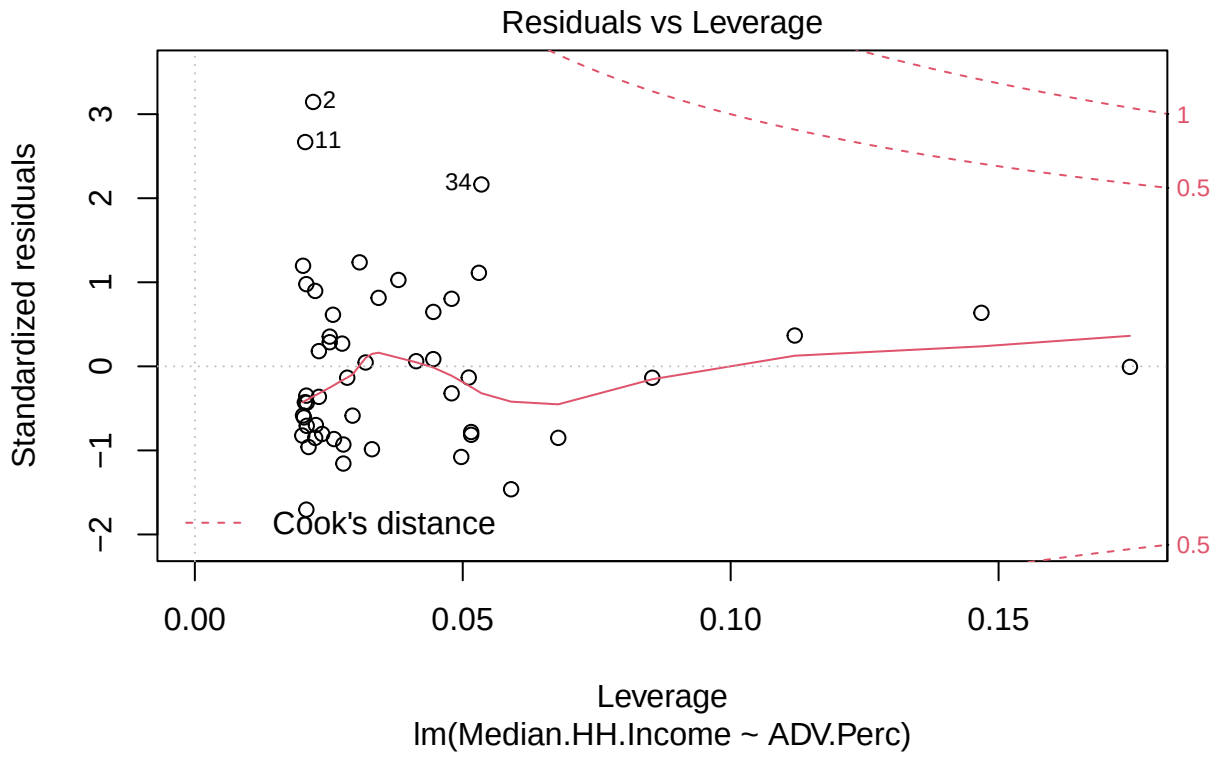



```
plot(linearEIADV)
```


Residuals vs Fitted
Residuals
20000
10000
0
−10000
2
11
34
45000
50000
55000
60000
65000
70000
Fitted values
lm(Median.HH.Income ~ ADV.Perc)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Median.HH.Income ~ ADV.Perc)
2
11
34
−2
−1
0
1
2
−1
0
1
2
3

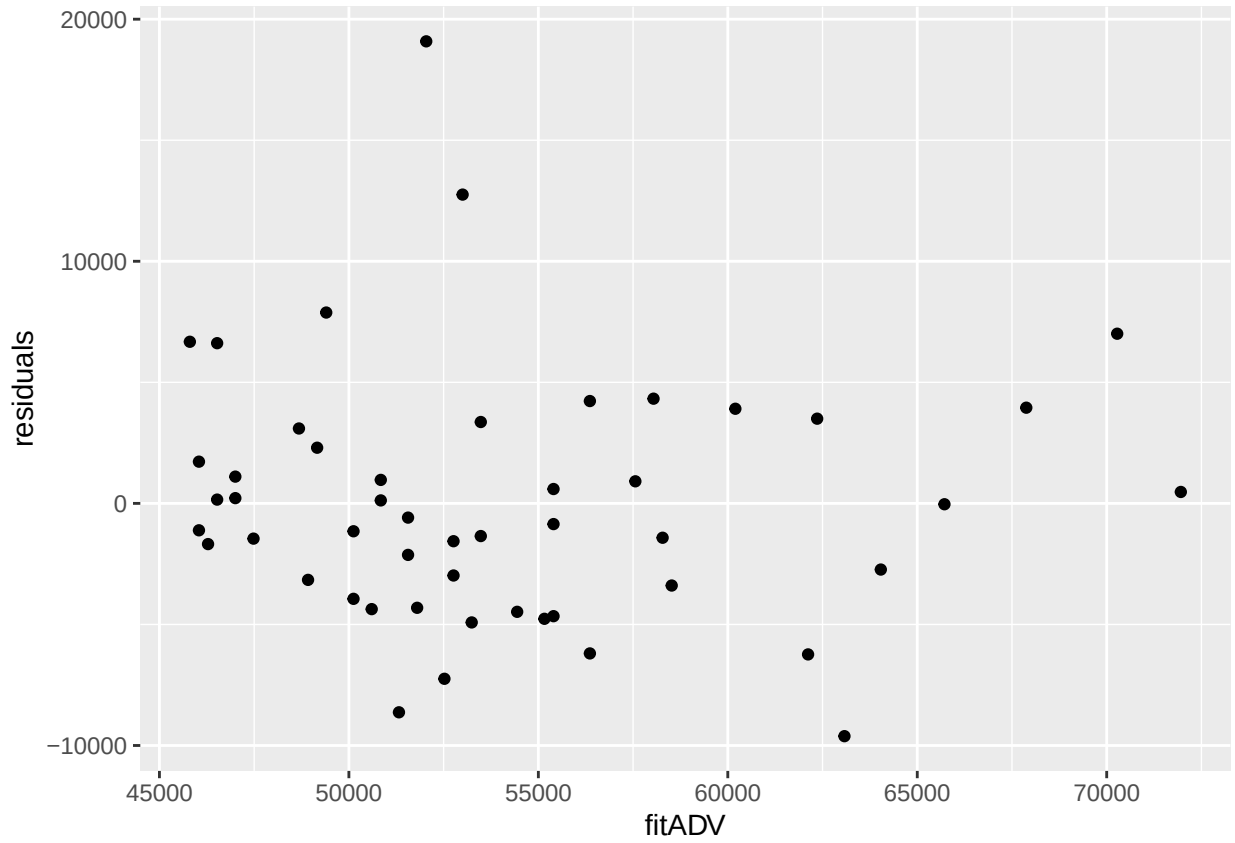

Scale−Location
2
11
34
√|Standardized residuals|
0.0
0.5
1.0
1.5
45000
50000
55000
60000
65000
70000
Fitted values
lm(Median.HH.Income ~ ADV.Perc)




```
residuals(linearEIADV)
```

```
##           1           2           3           4           5           6
##  -6092.24455  19541.55190  -2693.26431  -4780.97973   5572.65467   -815.87497
##           7           8           9          10          11          12
##   2170.59539   1676.08352  -4338.50937  -5119.83545  16590.79696    378.38786
##          13          14          15          16          17          18
##   -835.85521    283.81671   5029.69418  -2656.65166  -7163.93822  -1967.85720
##          19          20          21          22          23          24
##  -3779.14178   3694.20804    -37.36311  -3634.08051   6080.40961  -6602.91847
##          25          26          27          28          29          30
##  -4398.26431  -4993.63190   1124.42937   4927.14280   6332.63491   6801.18629
##          31          32          33          34          35          36
## -10597.59039  -5160.38485  -5965.32557  13225.95900  -2253.57063    526.26533
##          37          38          39          40          41          42
##  -5285.34533  -2166.59039  -3628.79395  -5363.81569   3976.26533  -5760.93822
##          43          44          45          46          47          48
##   2193.24557   7440.91949  -8912.62991   -815.95600   3805.96099  -4983.97973
##          49          50
##   1780.24557   7652.87798
```

```
fitADV <- fitted.values(linearEIADV)
ggplot(EI, aes(x=fitADV, y=residuals)) +geom_point()
```



(J) Do you observe any differences in the main results for these new columns?

The most obvious differences I saw was that, whereas for the Bachelor percentage, the residual veered off toward the latter quantiles, the High School percentage veered off during the entire graph, and the ADV Percentage veered off primarily during the beginning, with some notable outliers toward the end.

Problem 4

Read the following exploratory data analysis project about traffic and COVID: https://arxiv.org/pdf/2009.04612.pdf and write a paragraph summarizing your responses to the following questions.

- What research question is being addressed in this piece?
- Can you tell where the underlying data came from?
- Is the data publicly available? Do you think you could reproduce the cleaning and processing steps to regenerate the data set?
- What technique that we have discussed in our class do the authors use to evaluate the data?
- What do you think about the visualizations and examples in this piece?
- Which of the correlation heatmaps do you think is most effective or useful?
- Do you think the clusters of states in the *Analysis Results* section offer a fair comparison?
- The authors chose to use 2016 Presidential election results to divide up the states into red and blue. Does this seem like a reasonable choice? What else might they have used?
- How strong is the final evidence the authors provide for their conclusions?

In this piece, they are investigating the relationship between the level of transportation during COVID-19 and the presidential candidate chosen in that state during the 2016 election, with the hopes of finding some correlation between a state's overall political affiliation and their lockdown restrictions/willingness to comply with lockdown restrictions. The authors used data which is publicly available via the Department of Transportation. The way they processed and evaluated this data is, for the most part, fairly straightforward –

although some unfamiliar methods such as k-clustering are used, they focus on correlation analysis, which is fairly straightforward. Although the visualizations weren't terribly complex, some of them were low-quality and were screenshotted directly from a CNN broadcast, and many of them had a rather strange method of color-coding, where the colors did not correspond with the political party they were related to. I thought both of the correlation heat maps were rather difficult to read at first, particularly as they were divided by clusters which required referring to the previous page in order to understand which cluster denoted which grouping of states. The clusters themselves seemed to be a reasonable decision considering the complexity of the other factors that would influence travel; however because the states have such varied factors involved I feel the division could become somewhat arbitrary in some cases. I felt as if equating 2016 Presidential Election results to Republican vs Democratic states was a bit of a stretch. Even to my politically undeveloped eyes, I understand there were many factors involved that would have caused a particular state to vote in one direction, especially in an election as contentious as the one in question. For instance, the swing states may have been swayed one way during the election that does not reflect their overall political leanings. In addition, due to the winner-take-all approach that the electoral college uses in most states, and the dichotomous perspective of the article's author, a state may be categorized as being politically affiliated one way, when in actuality their leanings could be quite moderate. Many other facors could have affected the election results in a way that does not represent the state's actual population – the effects of gerrymandering and states such as Washington, who have a generally Democratic population in its metropolitan area, and more Republican-leaning residents in its other districts are other examples of confounding factors. Although again, the political leanings of states is too complex to put into such simple terms, perhaps instead they could have observed the popular vote for a particular (recent) amount of time, rather than basing their division off of one election. Although the authors' rather weak conclusion seemed to be reasonably obtained, despite the political shortcomings, I don't feel completely satisfied for a variety of reasons. For one, I do feel as if this statement is too complex and doesn't take into account many other factors that could influence the amount of travel and the rates of COVID. Although I don't disagree with the correlations found, I think there ought to be more research into the other factors that could affect the findings before any kind of conclusion could be drawn.