Predicting UFO Sightings

Shane Caldwell, Fanny Chow, Mackenzie Gray, Noah Johnson

11/30/2017

Contents

1 Introduction

Despite the U.S. government's explicit expression of disinterest in U.F.O. research, U.F.O. Sightings have sparked interest in Americans from coast-to-coast. Avid U.F.O. enthusiasists have collected data on U.F.O. sightings throughout the years at the National U.F.O. Reporting Center. Our data analysis explores what factors or demographics make an American more likely to report a U.F.O. sighting? And what common characteristics do U.F.O. sighters share?

Our group set out to create a model to predict UFO sightings by a variety of cofactors. This was a case where the project and the analysis were created by the data at hand rather than a starting from a question.

The National UFO Reporting Center, established in 1974, has put together a database for citizens to submit reports of unidentified flying object sightings. One can submit reports online, but they also have a hotline that can be called at any time, and the information from the report will be placed into the database. Regardless of the medium the report is received in, it will be filtered to confirm it isn't a "hoax".

The database is moderately sized, with over 80,000 observations of reports. Reports include a timestamp, a description of the aircraft, the state and city the report came from, and the duration of the report.

Our interest was in finding covariates to help us predict the number of reports for a specific state during a specific year. This would involve cleaning the web-scraped UFO dataset and mutating it into something closer to what we needed. It also involved carefully choosing cofactors - not a lot of literature on UFOs sightings and their trends to lean on!

2 The Data

For independent variables, we ended up choosing number of breweries in a state per year, the number of alien movies released per year, per capita gdp per year, state population, portion of the population with internet access, and number of US Air Force bases in a state. Below, we argue for the "theoretical" inclusion of each variable and explain both why we chose it, where we found the data, and how we went about cleaning it.

2.1 Breweries Per State Per Year

Alcohol inhibits our perceptions. If you've been out drinking, perhaps you're more likely to think you saw a UFO instead of stopping at a more reasonable explanation. If there are more breweries in your state, it's more likely you'll be drinking.

2.2 Alien Movies Per Year

How does culture affect people's perceptions of the "supernatural"? When trailers for alien movies have been playing on TV particularly frequently for a given year, are you more likely to have aliens playing in the back of your mind when you see something in the sky you can't explain? Does watching sci-fi movies make it more likely that you turn to an organization like the UFO Reporting Center when something bright is in the sky?

Scraping for the number of alien movies released a year would be difficult, but luckily wikipedia has a list already compiled. While there is no way the list could be complete, it features movies from years as far back as 1902. The wikipedia table was scraped and turned into a csv file and a simple python script was written to bin the frequency of movies by year to create the final dataset.

2.3 Military Bases per State

A common explanation for unidentified flying objects is of a more terrestrial, but no less secretive, origin—top secret military aircraft. Perhaps the most famous Air Force base in the world, Area 51 in Nevada, is the rumored location of alien remains or technology. Or mabye it's just classified experimental aviation testing. Either way, to what extent can we explain U.F.O. sightings by their distance to a military base?

A number of immediate problems with obtaining this data spring to mind. For one, there are almost certainly research outposts that are unknown to the general public, so these by definition will be left out of any otherwise complete list. Additionally, having to sum over the distance to every one of the 58 Air Force bases in the US for each of the 80,000+ sightings, and then averaging those by state may not provide that much more information than simply how many bases there are per state. From this line of thinking, we decided that the number of bases per state was a reasonable substitute for a more complicated per-sighting calculation.

The list of bases and their locations was obtained from the aptly named list of United States Air Force installations on Wikipedia.

2.4 Per Capita GDP

Is "seeing things in the sky" a poor man's game? Does the wealth of a particular state influence how they spend their time? Perhaps those with less money have nothing to do but sit outside and look into the sky?

2.5 Portion of population with internet access by year

To even know about the UFO reporting center's existence, you need internet access. If you see a UFO and don't have access to the internet, it's unlikely you'd ever find the hotline to call to report it. It would just end up as a story you told your friends rather than a report in the UFO database.

3 EDA

3.1 Breweries Per State

```
beer.path <- "../../data/raw/brew_count_by_state_1984_2017.csv"
sightings.path <- "../../data/raw/ufo_sightings.csv"

beer.raw <- fread(file.path(beer.path), header=TRUE, na.strings=c("*", ""))
sightings.raw <- fread(file.path(sightings.path), header = TRUE, na.strings = c("", "Unknown",
```

```
# clean up junk at bottom file
beer.db <- beer.raw %>%
  filter(!is.na(STATE)) %>%
  filter(STATE != "Total") %>%
  filter(STATE != "Other") %>%
  filter(STATE != "* No reportable data") %>%
  filter(STATE != "«This list will be updated quarterly.")
```

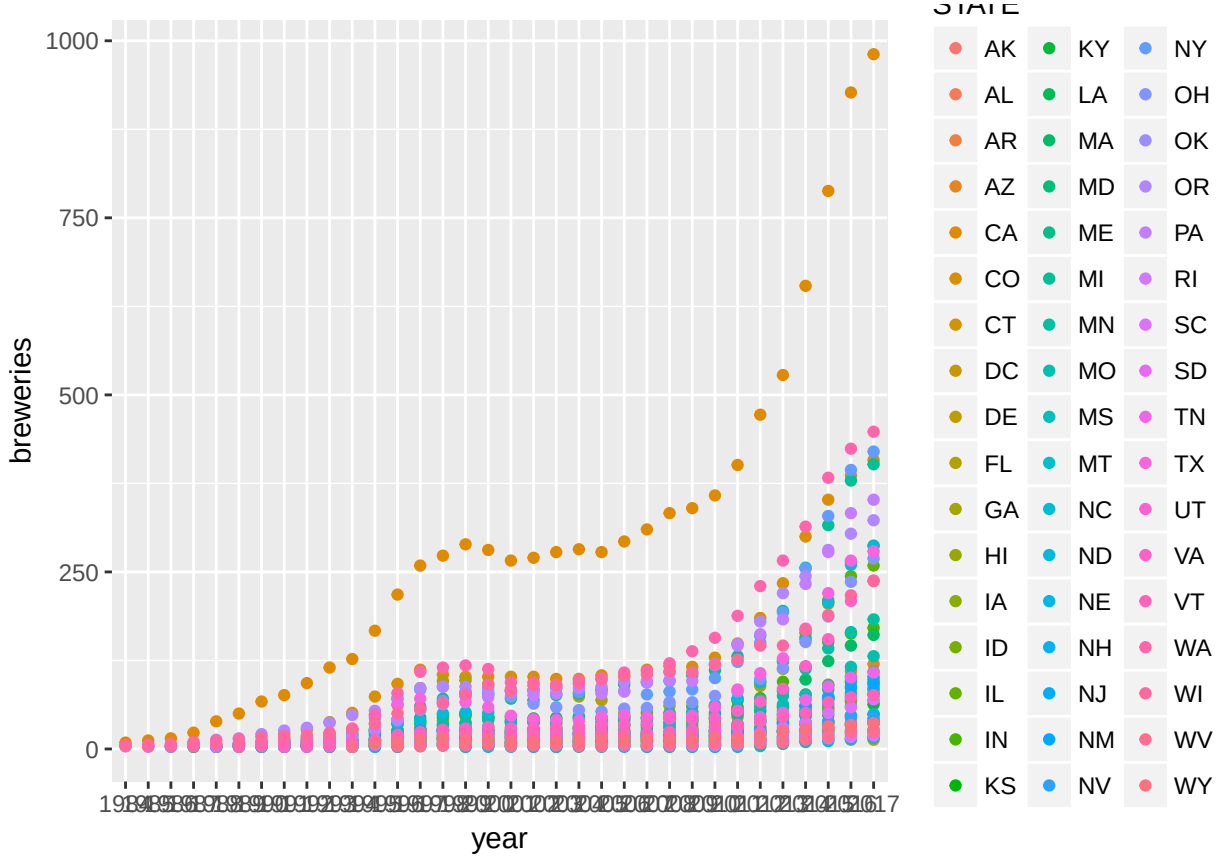
```
# create proper data types
beer.db$STATE <- as.factor(beer.db$STATE)

# wide to long format
olddata_wide <- beer.db
keycol <- "year"
valuecol <- "breweries"
gathercols <- as.character(seq(1984, 2017))

beer.df <- gather_(olddata_wide, keycol, valuecol, gathercols)
```

At a high-level glance, the general trend is increase number of breweries through the years for each s tate. Note that the number of breweries in 2005 will be contingent on the number of breweries in 2004, and there will be autocorrolation through years.

```
beer.df$breweries <- as.numeric(beer.df$breweries)
```

```
(ggplot() +
  geom_point(data=beer.df, aes(year, breweries, color=STATE)))
```

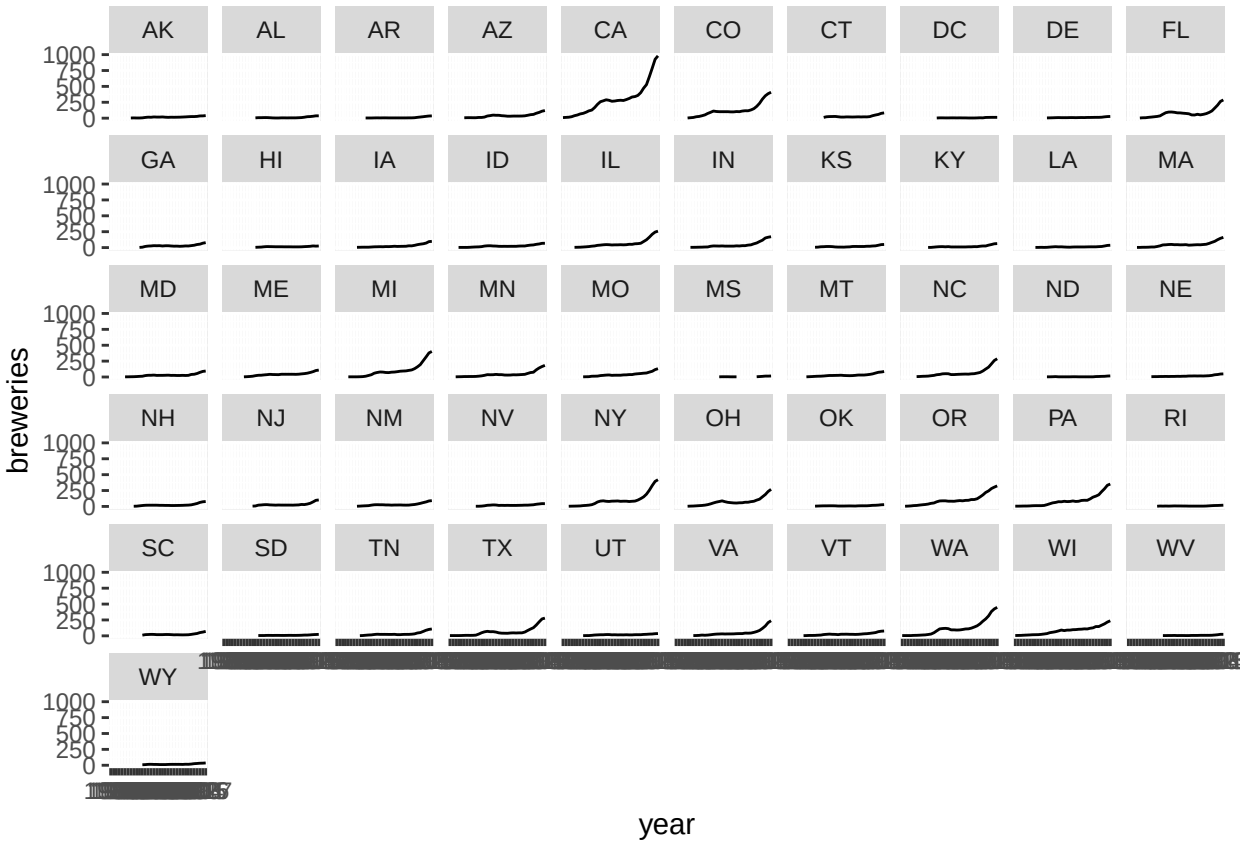
```
## Warning: Removed 350 rows containing missing values (geom_point).
```



Let's take a look at the number of breweries in each state through the years sorted by states. Since there's over 50 states we're looking at, it's challenging to discern trends from looking at all the states at once.

```
breweries.year <- ggplot(beer.df, aes(x = year, y = breweries, group=1))
(p2 <- breweries.year + geom_line() +
   facet_wrap(~STATE, ncol = 10))
```

```
## Warning: Removed 7 rows containing missing values (geom_path).
```



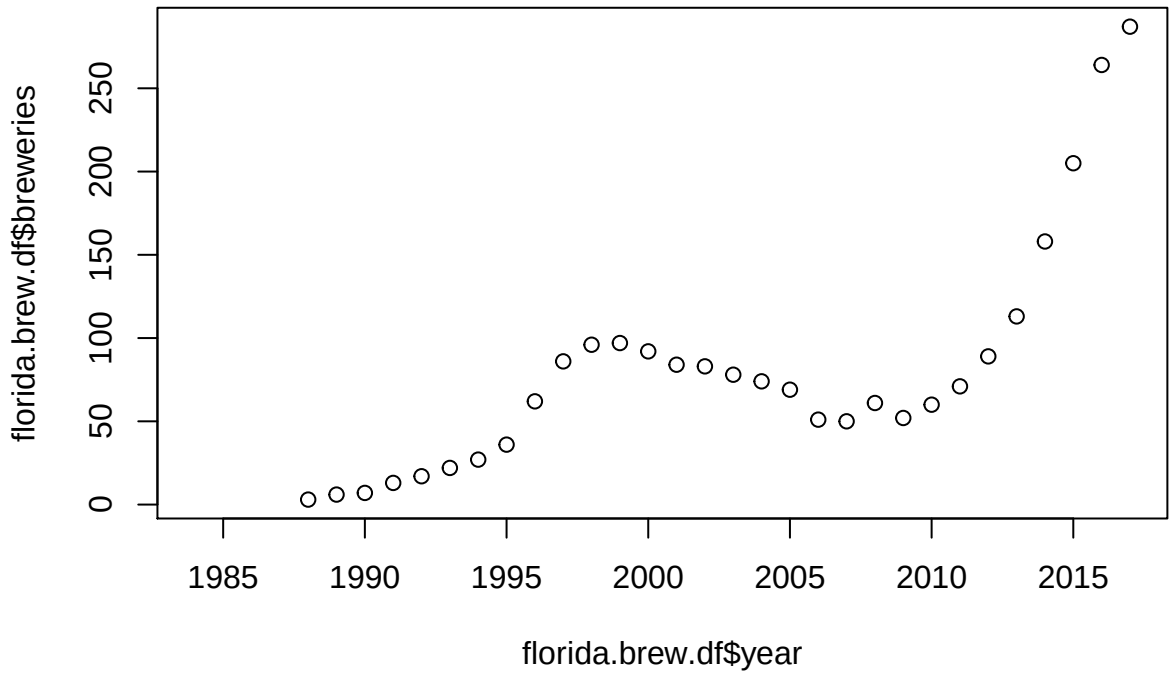
Let's focus on the state of Florida through the years. We observe an upward trend and then a sudden dip from the late 90's to 2010.

```
florida.brew.df <- beer.df %>%
  filter(STATE == "FL")
florida.brew.df
```

```
##    STATE year breweries
## 1     FL 1984        NA
## 2     FL 1985        NA
## 3     FL 1986        NA
## 4     FL 1987        NA
## 5     FL 1988         3
## 6     FL 1989         6
## 7     FL 1990         7
## 8     FL 1991        13
## 9     FL 1992        17
## 10    FL 1993        22
## 11    FL 1994        27
## 12    FL 1995        36
## 13    FL 1996        62
## 14    FL 1997        86
## 15    FL 1998        96
## 16    FL 1999        97
## 17    FL 2000        92
```

```
## 18    FL 2001        84
## 19    FL 2002        83
## 20    FL 2003        78
## 21    FL 2004        74
## 22    FL 2005        69
## 23    FL 2006        51
## 24    FL 2007        50
## 25    FL 2008        61
## 26    FL 2009        52
## 27    FL 2010        60
## 28    FL 2011        71
## 29    FL 2012        89
## 30    FL 2013       113
## 31    FL 2014       158
## 32    FL 2015       205
## 33    FL 2016       264
## 34    FL 2017       287
```

```r
plot(florida.brew.df$year, florida.brew.df$breweries)
```



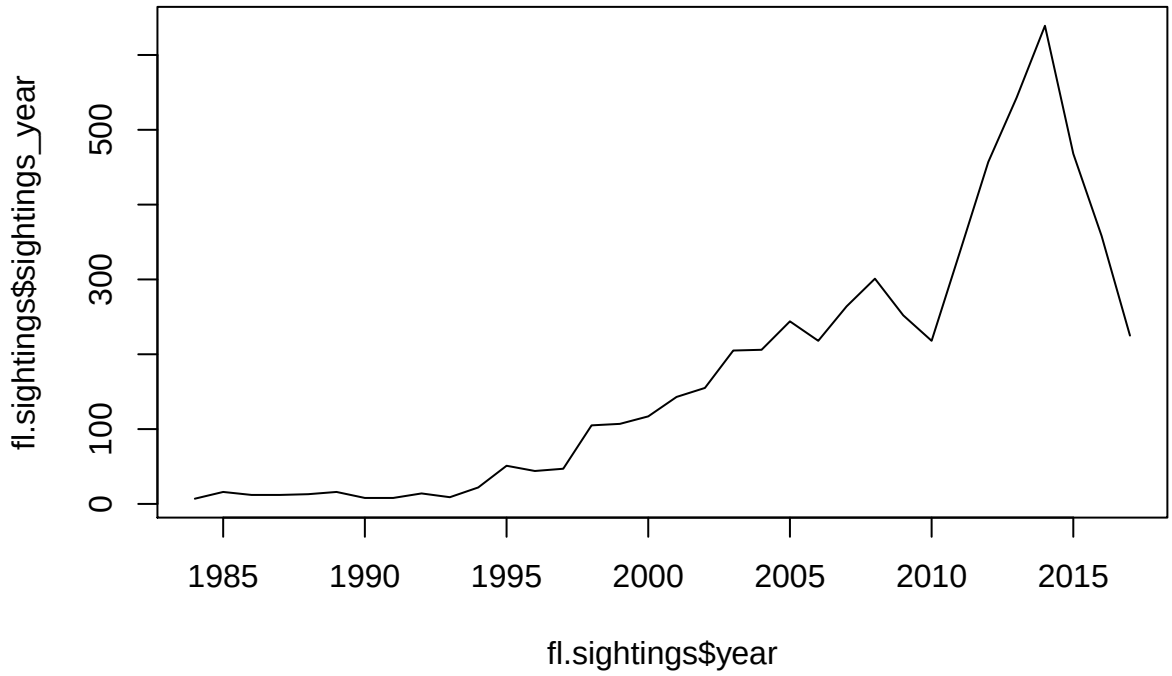
Since the range of time in the breweries data is from 1984-2017, let's subset the equivalent years from the sightings data.

```r
# clean up sightings data
fl.sightings <- sightings.raw %>%
  filter(state == 'FL') %>%
  mutate(year = as.numeric(format(as.Date(date_time, format="%m/%d/%y"),"%Y"))) %>%
  filter(year >= 1984) %>%
  filter(year <= 2017) %>%
  #group_by(year) %>%
```

```
  count(year) %>%
  rename(sightings_year = n)
```

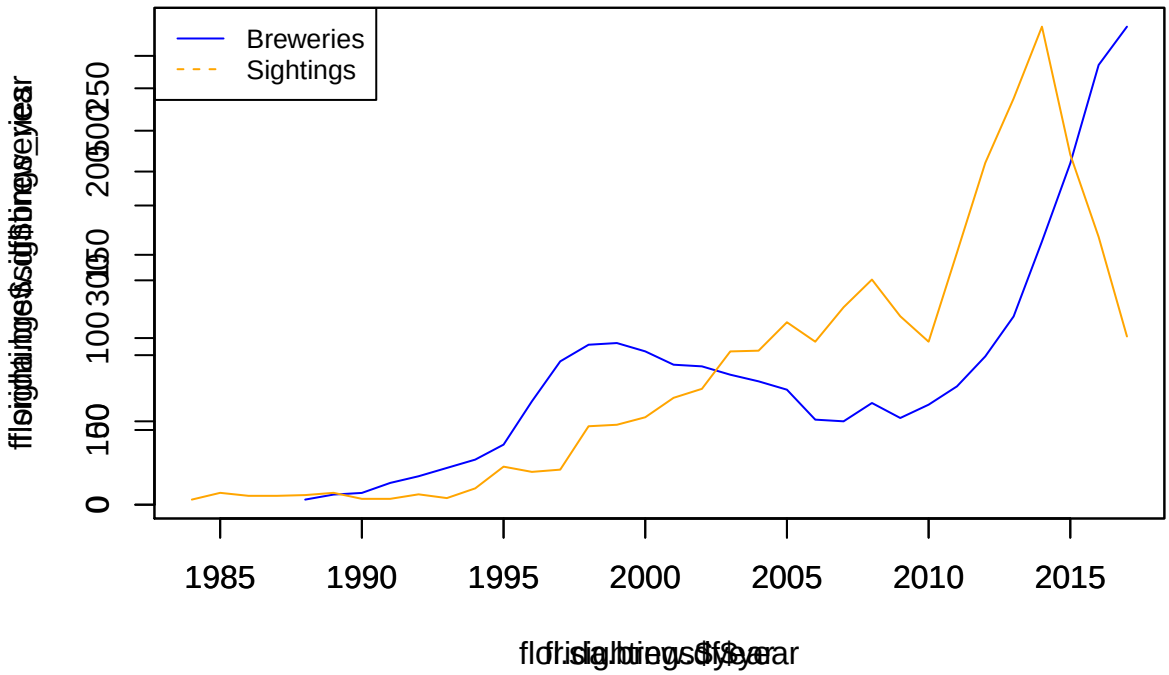
Let's take a snapshot of sightings per year in Florida.

```
plot(fl.sightings$year, fl.sightings$sightings_year, type="l")
```



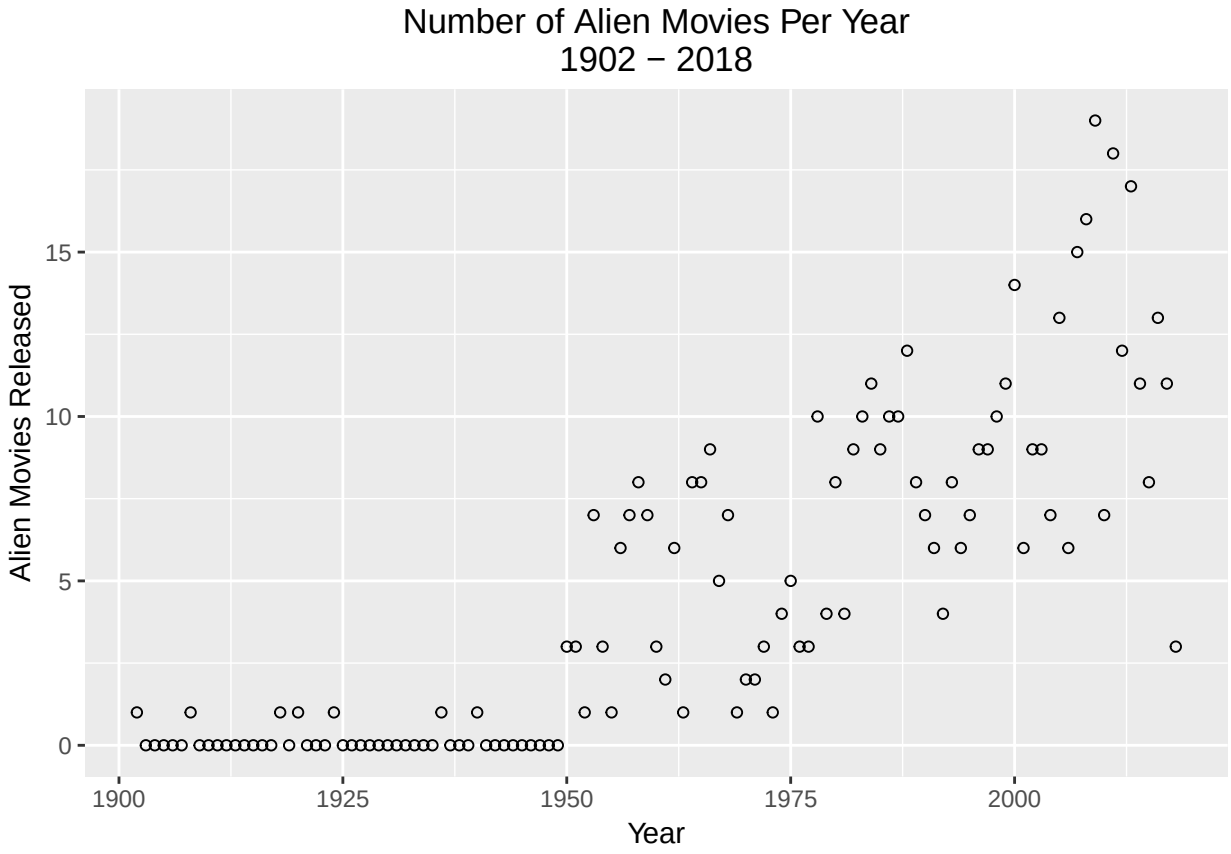
Let's compare the 2 plots at once. Interesting how the 2 plots follow the same shape until the early 2000s and then diverge drastically afterwards. It makes sense to see an overall trend of increased interest in breweries over time.

```
plot(florida.brew.df$year, florida.brew.df$breweries, type = "l", col="blue")
par(new=TRUE)
plot(fl.sightings$year, fl.sightings$sightings_year, type="l", col="orange")
legend("topleft",legend=c("Breweries", "Sightings"),
       col=c("blue", "orange"), lty=1:2, cex=0.8)
```



3.2 Alien Movies Per Year

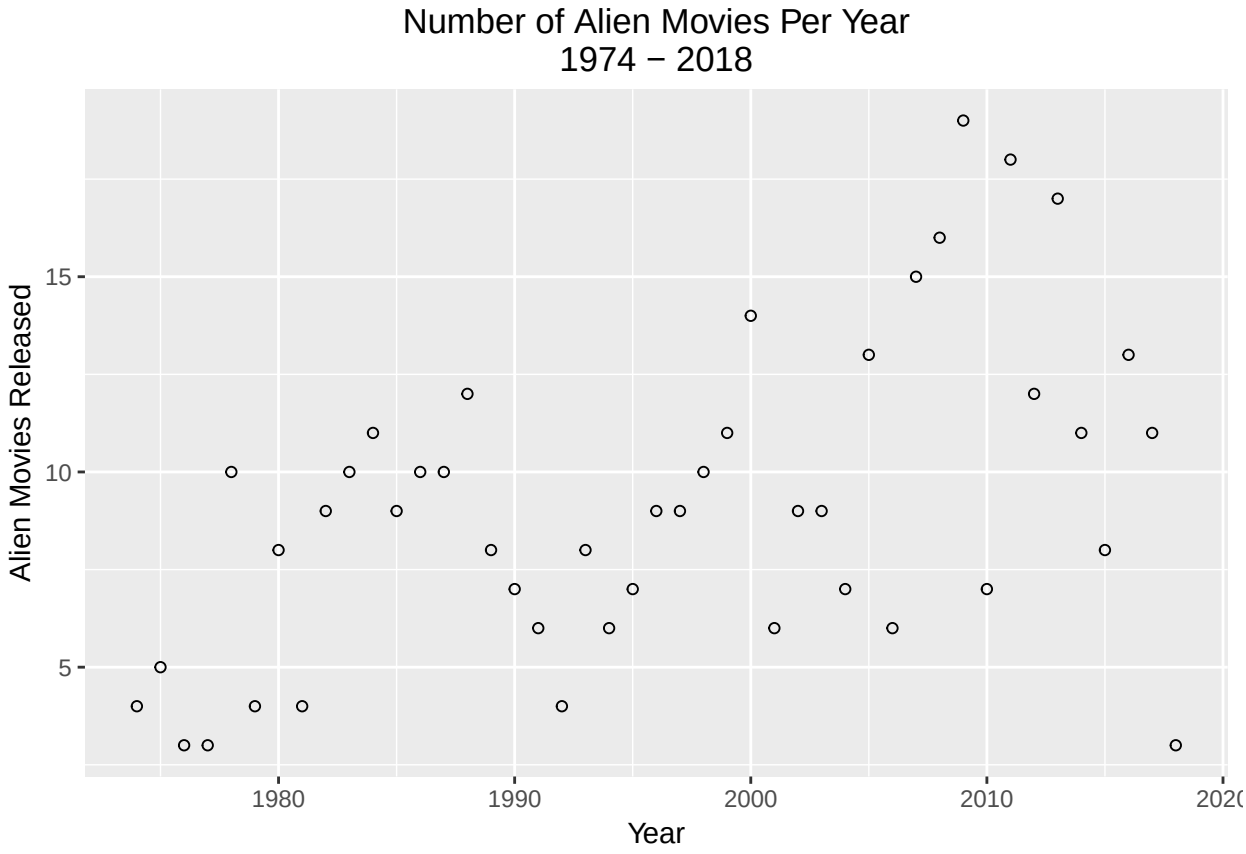
```
#pull in alien movies
alien_movies <- read.csv(file = "../../data/raw/alien_movies_per_year.csv", header = TRUE)
sight <- read.csv(file = "../../data/raw/ufo_sightings.csv", header = TRUE)

ggplot(alien_movies, aes(x =year, y =num_movies )) + geom_point(shape=1) + labs( title = "Numbe
```



One big problem right off the bat - a lot of the early years are only zeros. After the 1902 film (a silent film known as "A Trip to the Moon", if you were curious) is followed by a large number of empty slots.

Luckily, since the UFO reporting society was founded in 1974, we decided to remove all variables prior to this year. Reports before this year appear in the dataset, but as discussed above we determined these reports would not be live records. They would be "popular" events that were recorded after the fact, or incidents reported years after they happened. Let's see if that improves the look of our scatterpplot.

```
alien_movies_new <- filter(alien_movies, year >= 1974)


ggplot(alien_movies_new, aes(x =year, y =num_movies )) + geom_point(shape=1) + labs( title = "
```



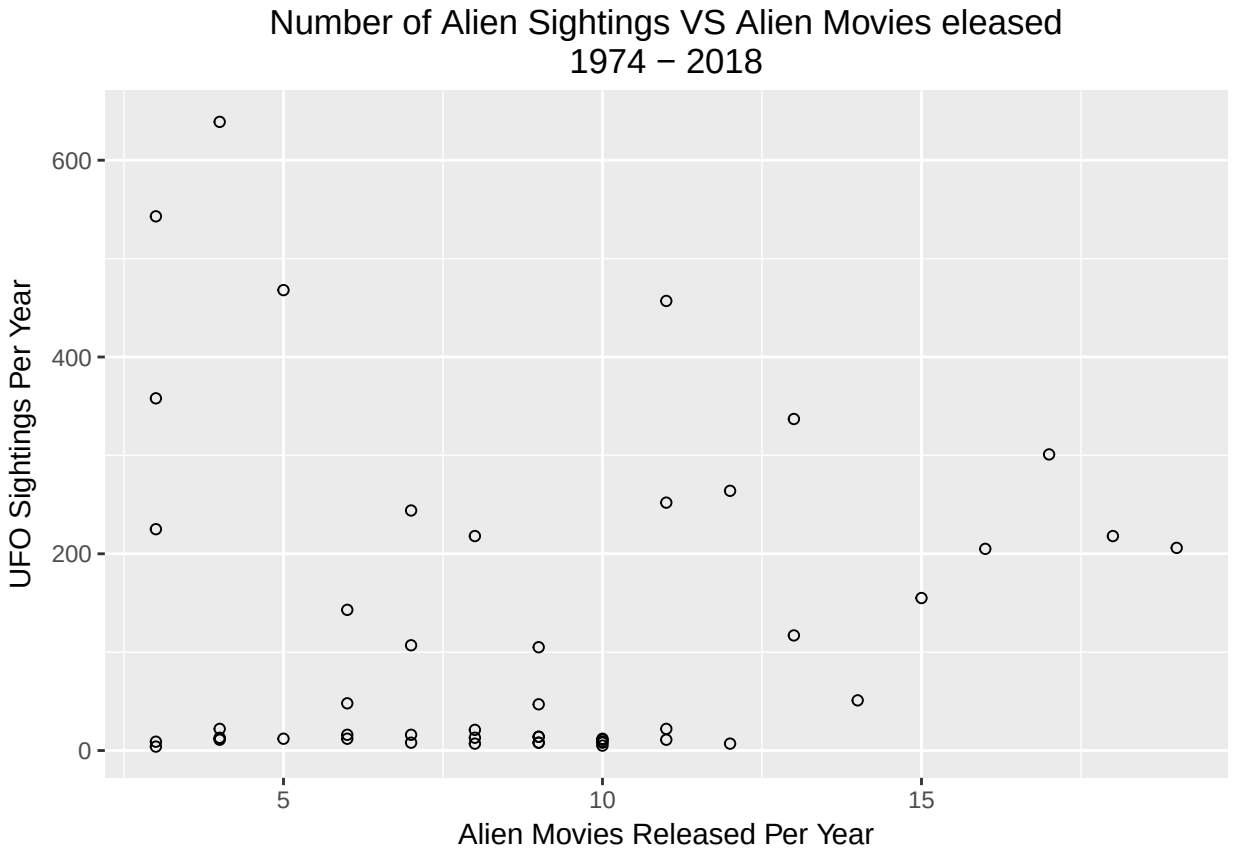
Still a definite trend upwards. That could have to do with the number of movies released increasing in general each year moreso than an increase in the genre alien movies specifically.

Anyway, we have to filter the sightings dataset to find more specific information about our dataset. Currently we can't compare the number of movies to our number of sightings. Let's filter the dataset down to sightings taking place in Florida since 1974 and see how they compare to the number of movies released each year.

```
fl_sightings <- filter(sight, tolower(state) == 'fl' )
fl_sightings$date_time <- as.Date(fl_sightings$date_time, "%m/%d/%y")
tmp <- lapply(strsplit(as.character(fl_sightings$date_time), "-"), `[[`, 1)
tmp2 <- sapply(tmp, "[[", 1)
fl_sightings$year <- as.numeric(tmp2)
fl_sightings_75 <- filter(fl_sightings, year >= 1975)
fl_sightings_75 <- filter(fl_sightings, year < 2018)

sightings_per_year_fl <- as.data.frame(table(fl_sightings_75$year))

#sightings_per_year_fl_75$Freq
movies_sightings <- cbind(sightings_per_year_fl$Freq, alien_movies_new$num_movies)
```

```
## Warning in cbind(sightings_per_year_fl$Freq, alien_movies_new$num_movies):
## number of rows of result is not a multiple of vector length (arg 2)
```

```r
movies_sightings <- as.data.frame(movies_sightings)

colnames(movies_sightings) <- c("Sightings_Per_Year", "Alien_Movies_Per_Year")

ggplot(movies_sightings, aes(x =Alien_Movies_Per_Year, y =Sightings_Per_Year )) + geom_point(sh
```



There seem to be several significant outliers. It appears one year there were over 600 alien sightings, but almost no alien movies released.

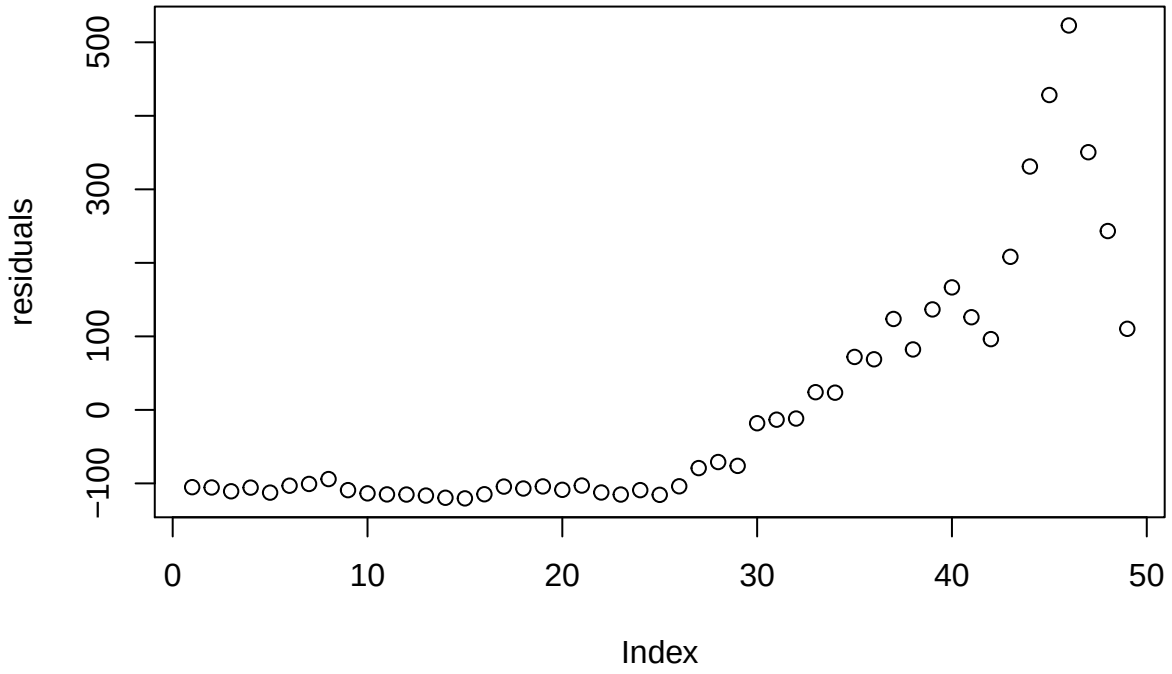
```r
lm.movies <- lm(Sightings_Per_Year~Alien_Movies_Per_Year, data = movies_sightings)
summary(lm.movies)
```

```
##
## Call:
## lm(formula = Sightings_Per_Year ~ Alien_Movies_Per_Year, data = movies_sightings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.36 -109.14 -100.74   82.22  522.88
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            110.505     54.624   2.023   0.0488 *
## Alien_Movies_Per_Year    1.404      5.689   0.247   0.8061
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162.3 on 47 degrees of freedom
## Multiple R-squared:  0.001295,   Adjusted R-squared:  -0.01995
## F-statistic: 0.06095 on 1 and 47 DF,  p-value: 0.8061
```

Results here are interesting. Our R-squared isn't incredible, only capturing 19% of the variance in our data. Without other variables in the model, it's difficult to say if the alien movies per year are useful on their own.

```
residuals <- resid(lm.movies)
plot(residuals)
```



This definitely does not appear to have no pattern! We could easily draw a curve fitting this data. The variance increases the further we go through the function. Because of the pattern in the error, believe there are other variables in the true model that we're currently missing. This isn't terrible news, because we certainly plan on adding more!

3.3 Number of USAF Bases by State

4 Loading and cleaning

First import all of the necessary packages

```
library(lmtest)
library(plyr)
library(tidyverse)
library(ggplot2)
```

```
library(ggrepel)
library(GGally)
library(lubridate)
library(zoo)
```

Next, read in the necessary data. `sightings` is the ufo sighting records from NUFORC, `base_locs` location information about USAF bases in the US, `name_to_abbr` a mapping of US state names to abbreviations, and `stat_pops` US Census population data per state per decade from 1960 to 2010.

```
sightings <- read_csv("../../data/raw/ufo_sightings.csv")
base_locs <- read_csv("../../data/raw/usaf_base_locs.csv")
state_pops <- read_csv("../../data/raw/state_pops.csv")
```

Next, transform and select the relevant data.

```
sightings <- sightings  %>%
    filter(X1 < 109088) %>%
    filter(state %in% state.abb)

state_pops$Name <- mapvalues(state_pops$Name, state.name, state.abb)
state_pops <- state_pops %>%
    filter(Name %in% state.abb) %>%
    select(one_of(c("1970", "1980", "1990", "2000", "2010"))) %>%
    apply(1, mean)

sightings <- sightings  %>%
    group_by(state) %>%
    count()

base_locs <- base_locs  %>%
    group_by(State) %>%
    count() %>%
    merge(state.abb, by.x="State", by.y=1, all=T) %>%
    dplyr::mutate(n=replace(n, is.na(n), 0)) %>%
    filter(!is.na(State))

df <- data.frame(sightings=sightings$n, base_count=base_locs$n, pop=state_pops, name=state.abb
```

4.1 Statistical analysis

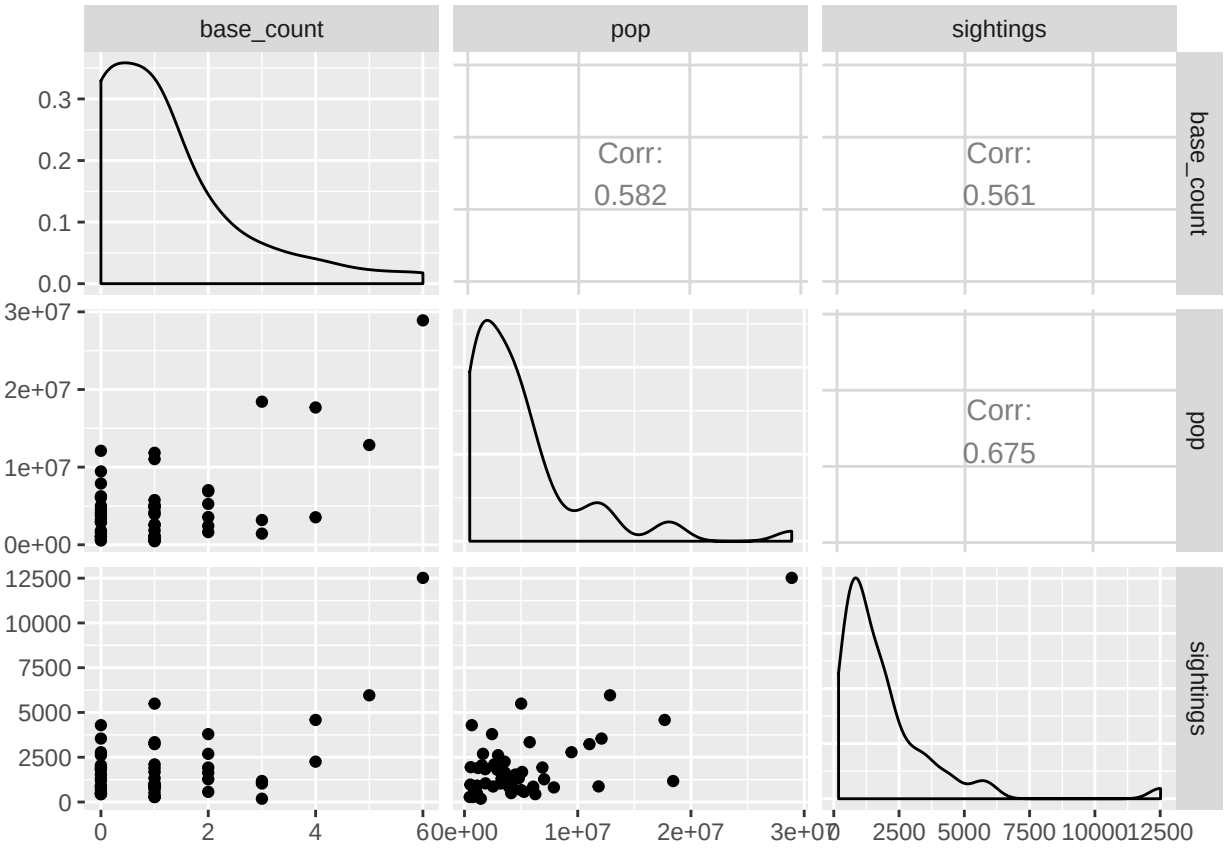
Now that the data is all clean and together, we can begin analysis.

4.1.1 Initial EDA

We can create a pairplot of the quantitative variables to get a rough look at how population and number of Air Force bases relate to the number of UFO sightings. It is reasonable to assume that

states with a higher population will have more bases and more sightings just by virtue of having more people living there. This can be confirmed using `ggpairs` from the `GGally` package.

```
ggpairs(df[,c("base_count", "pop", "sightings")])
```



From this we can make a few initial observations. It seems as though `pop` is relatively correlated with both `sightings` and `base_count` ($R^2 = 0.456$ and 0.338, respectively). The relationship between `base_count` and `pop` does not appear entirely linear however. Additionally, it seems as though this relationship may have some degree of hetereoscedasticity. It also appears, however, that `base_count` and `sightings` are correlated with $R^2 = 0.315$. This suggests multicolinearlity between the predictor variables `base_count` and `pop`, which will complicate matters further on.

4.1.2 State populations vs number of sightings

Although these two variables seems obviously related, it is good practice to double check assumptions.

```
(s1 <- summary(fit1 <- lm(sightings ~ pop, data=df)))
```

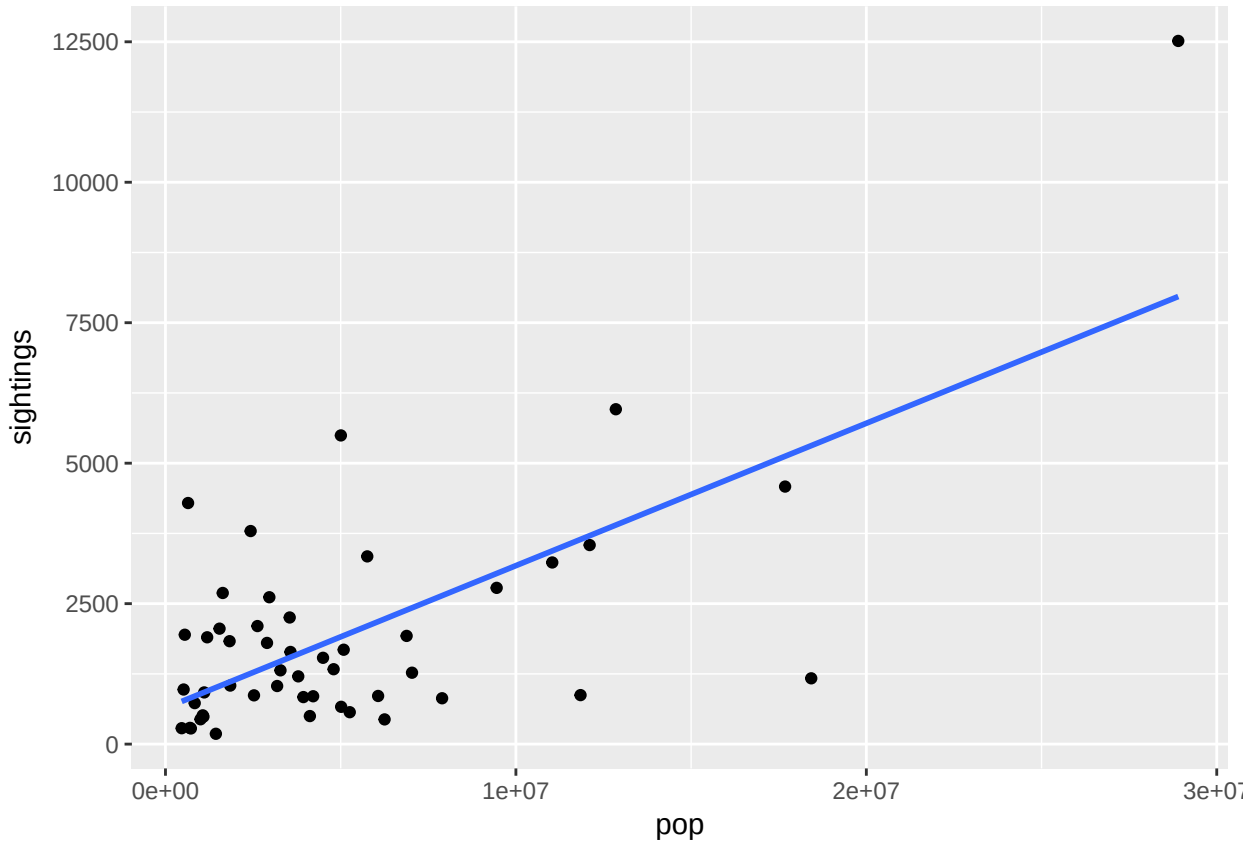
```
##
## Call:
## lm(formula = sightings ~ pop, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4141.2  -545.0  -253.7   722.8  4548.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.424e+02  2.958e+02   2.172   0.0348 *
## pop         2.534e-04  3.999e-05   6.337 7.65e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1525 on 48 degrees of freedom
## Multiple R-squared:  0.4555, Adjusted R-squared:  0.4442
## F-statistic: 40.16 on 1 and 48 DF,  p-value: 7.645e-08
```

Given $R^2 = 0.456$, $\beta_{pop} = 2.5 \times 10^{-4}$, and a 95% confidence interval for $\beta_{pop} = (1.7 \times 10^{-4}, 3.3 \times 10^{-4})$, we can be reasonably confident that a positive correlation does exist between the number of reported UFO sightings in a state since 1974 and the mean population of that state since 1970.

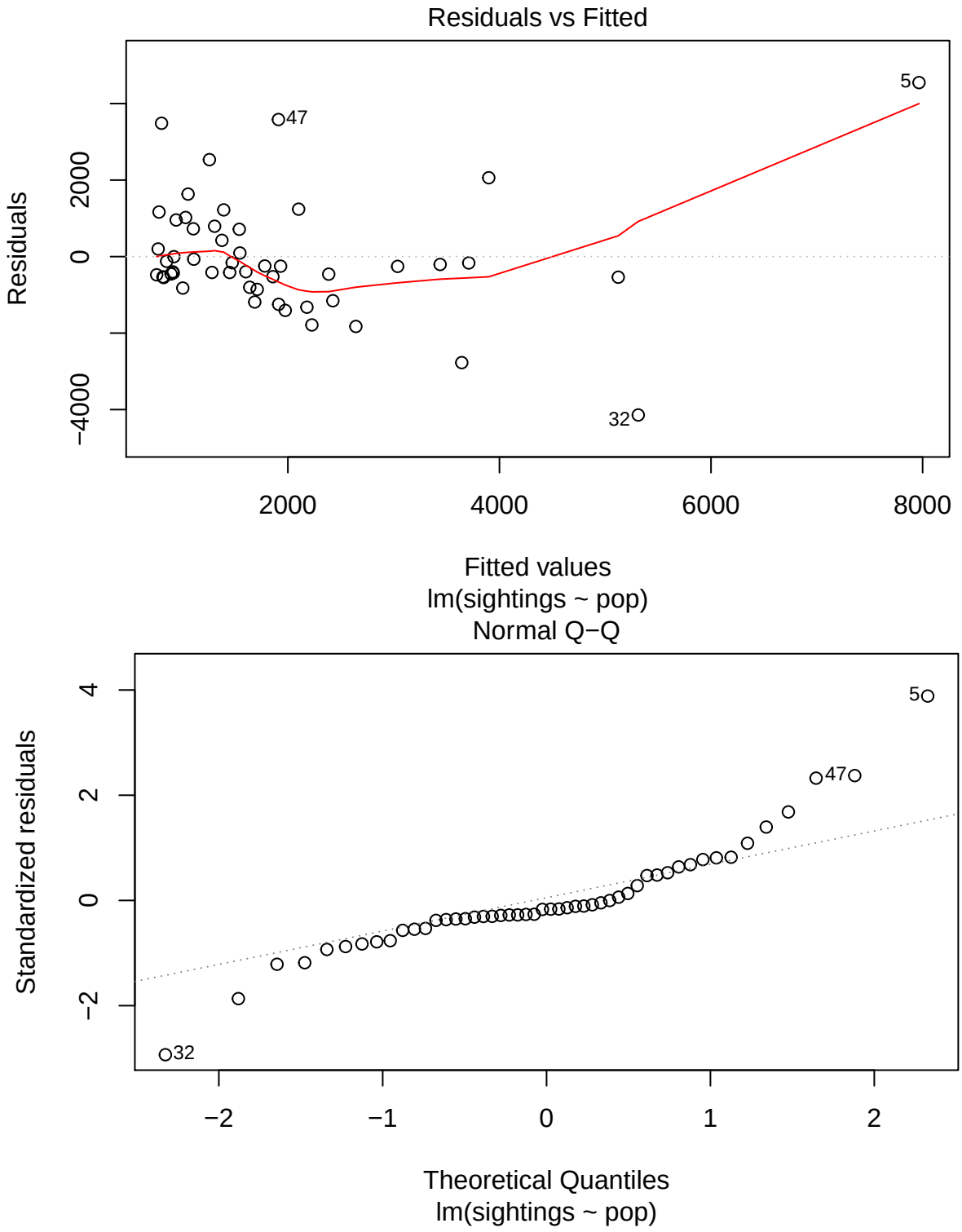
We can plot this relationship using `ggplot`.

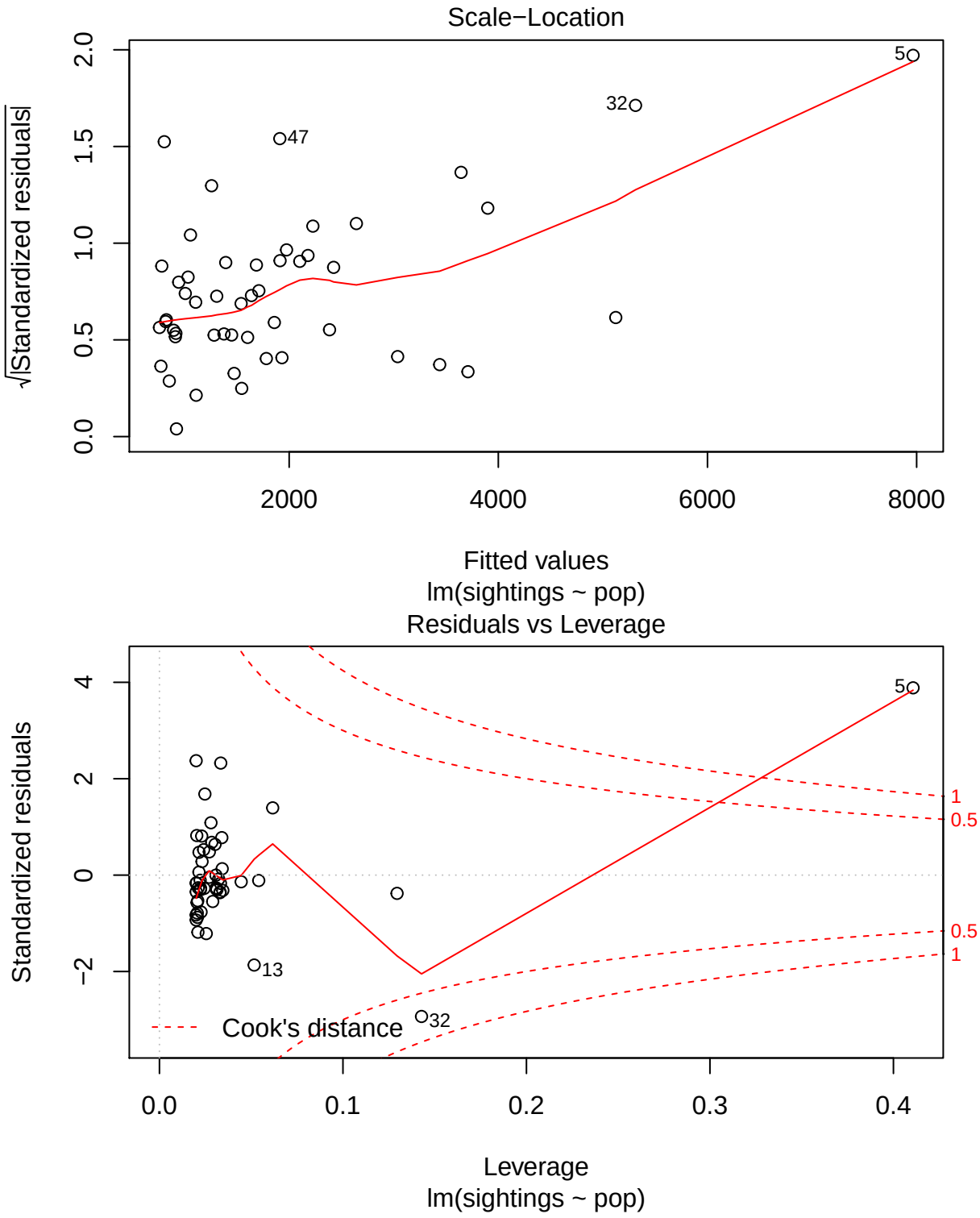
```
ggplot(data=df, aes(x=pop, y=sightings)) + geom_point() + stat_smooth(method='lm', se=F)
```



A number of obvious outliers are noticeable, but otherwise the relationship appears linear. We can graphically futher verify the assumptions of the least squares regression we performed.

```
plot(fit1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(sightings ~ pop)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sightings ~ pop)
5
47
32

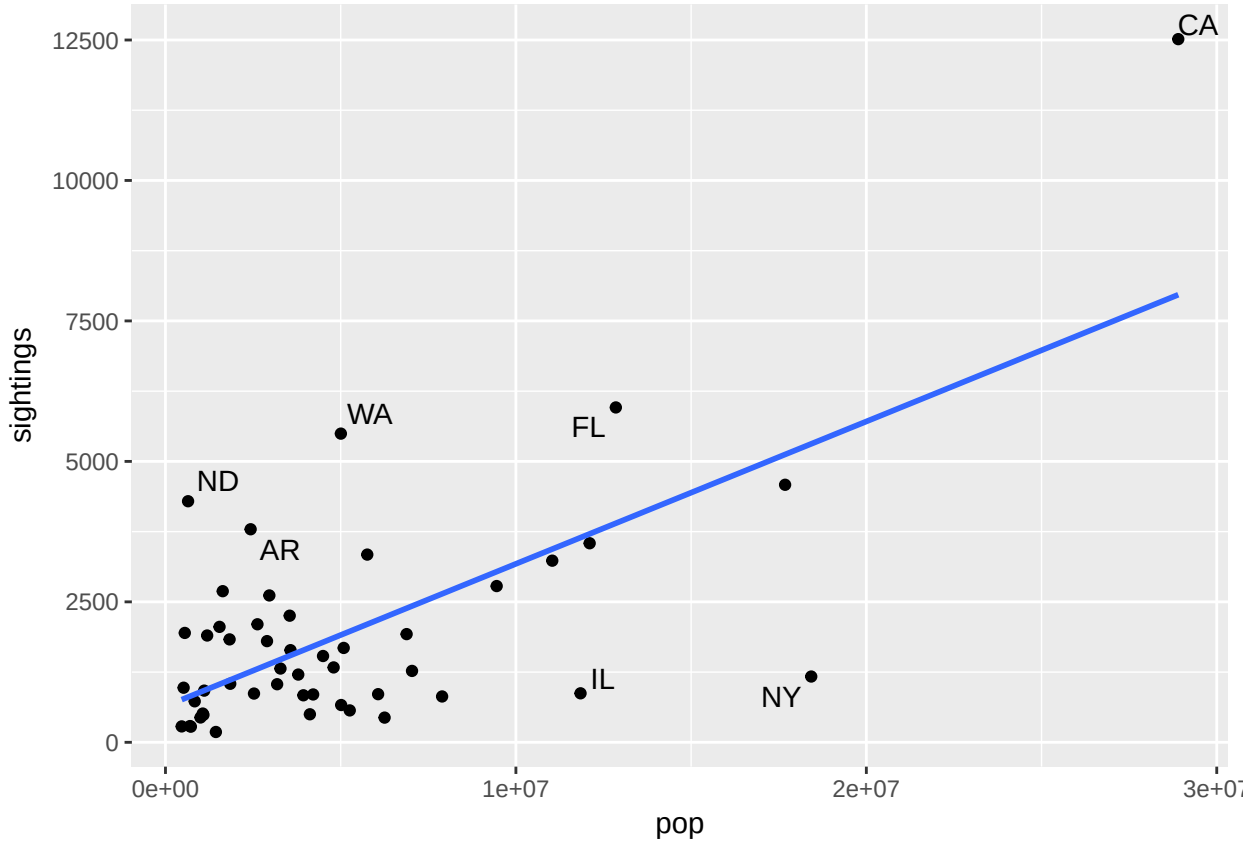


From the Q-Q plot, we can see that the errors do not appear to be normally distributed, with very large deviations from normal at either end. Additionally, two major outliers with IDs 5 and 32 are visible on the plot of residuals vs leverage. We can identify which states have the largest prediction error.

```
(top_resid <- fit1$residuals  %>%
    abs() %>%
    order(decreasing=T) %>%
    head(7) %>%
    df[., "name"])
```

```
## [1] CA NY WA ND IL AR FL
## 50 Levels: AK AL AR AZ CA CO CT DE FL GA HI IA ID IL IN KS KY LA MA ... WY
```

```
ggplot(data=df, aes(x=pop, y=sightings)) +
    geom_point() +
    stat_smooth(method='lm', se=F) +
    geom_text_repel(aes(label=name), data=df[df$name %in% top_resid,])
```



From this we can clearly see California (point number 5 on the diagnostic plots) is a major outlier. TODO: More analysis of this here

4.1.3 Number of US Air Force bases vs. population

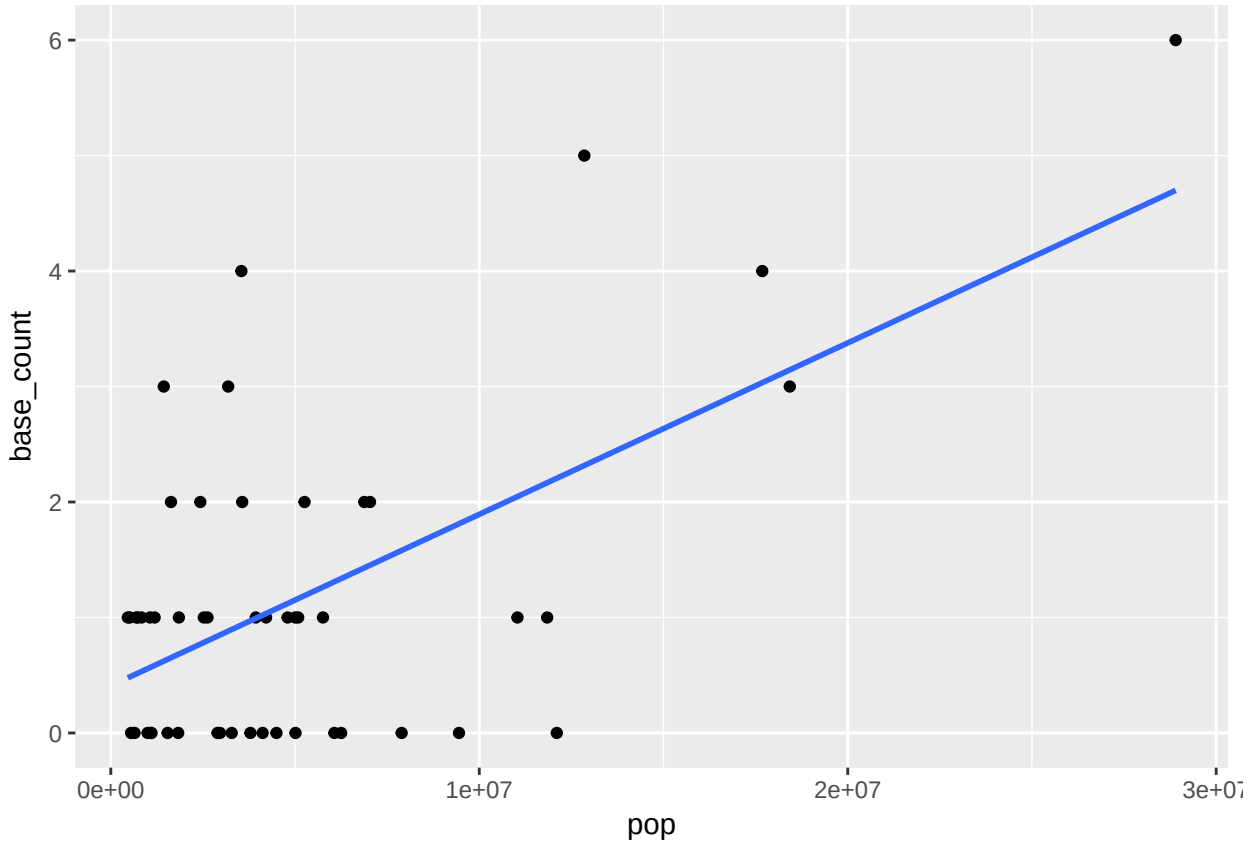
```
(s2 <- summary(fit2 <- lm(base_count ~ pop, data=df)))
```

```
##
## Call:
## lm(formula = base_count ~ pop, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2054 -0.8461 -0.1325  0.5207  3.0652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.086e-01  2.216e-01   1.844   0.0714 .
## pop         1.484e-07  2.996e-08   4.955  9.4e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.143 on 48 degrees of freedom
## Multiple R-squared:  0.3384, Adjusted R-squared:  0.3246
## F-statistic: 24.55 on 1 and 48 DF,  p-value: 9.402e-06
```

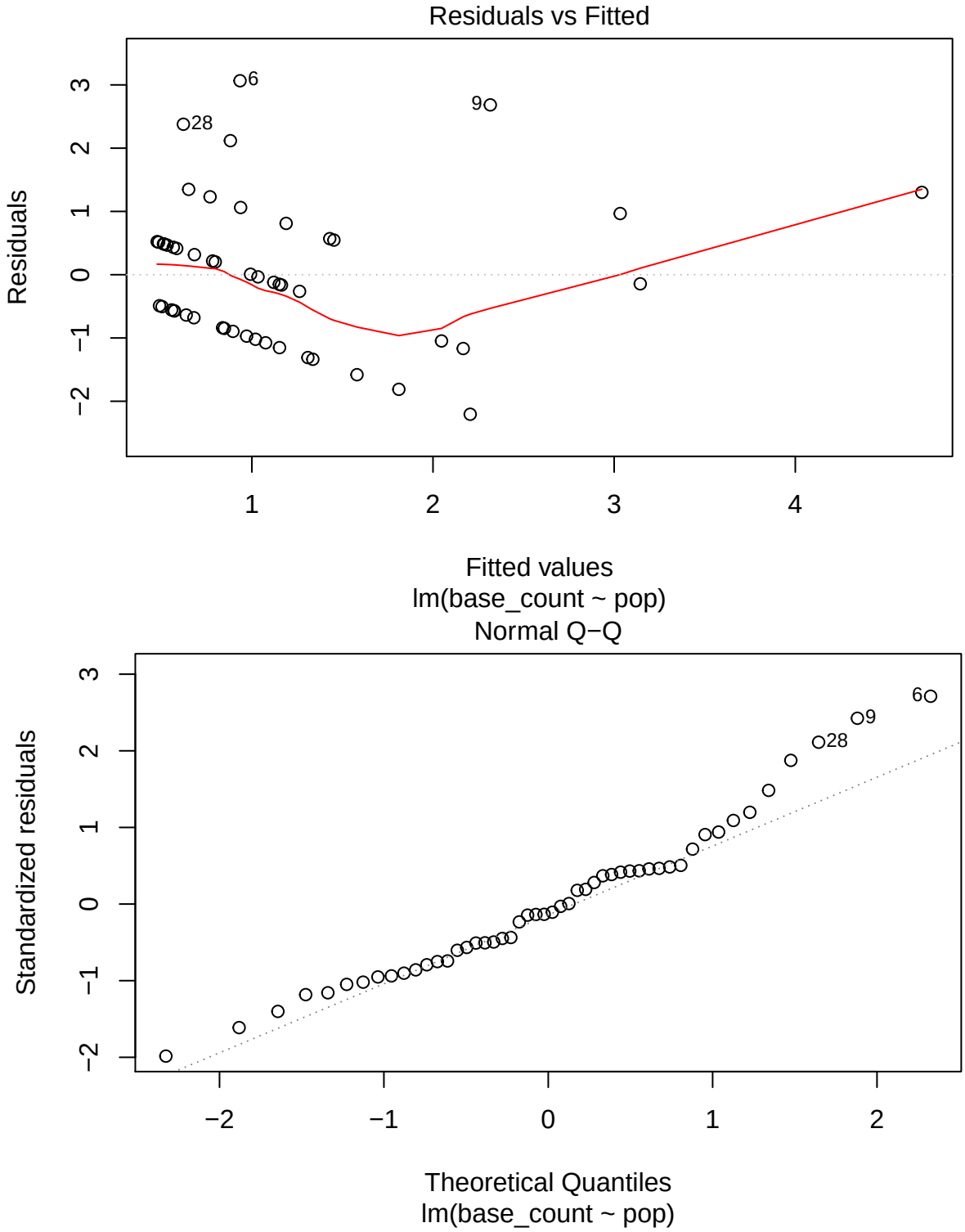
With an 95% confidence interval for $\beta_{pop} = (8.82 \times 10^{-8}, 2.087 \times 10^{-7})$, we can confidently state that there exists a positive relationship between the number of US military bases in a state and the mean population of that state from 1970-2010.

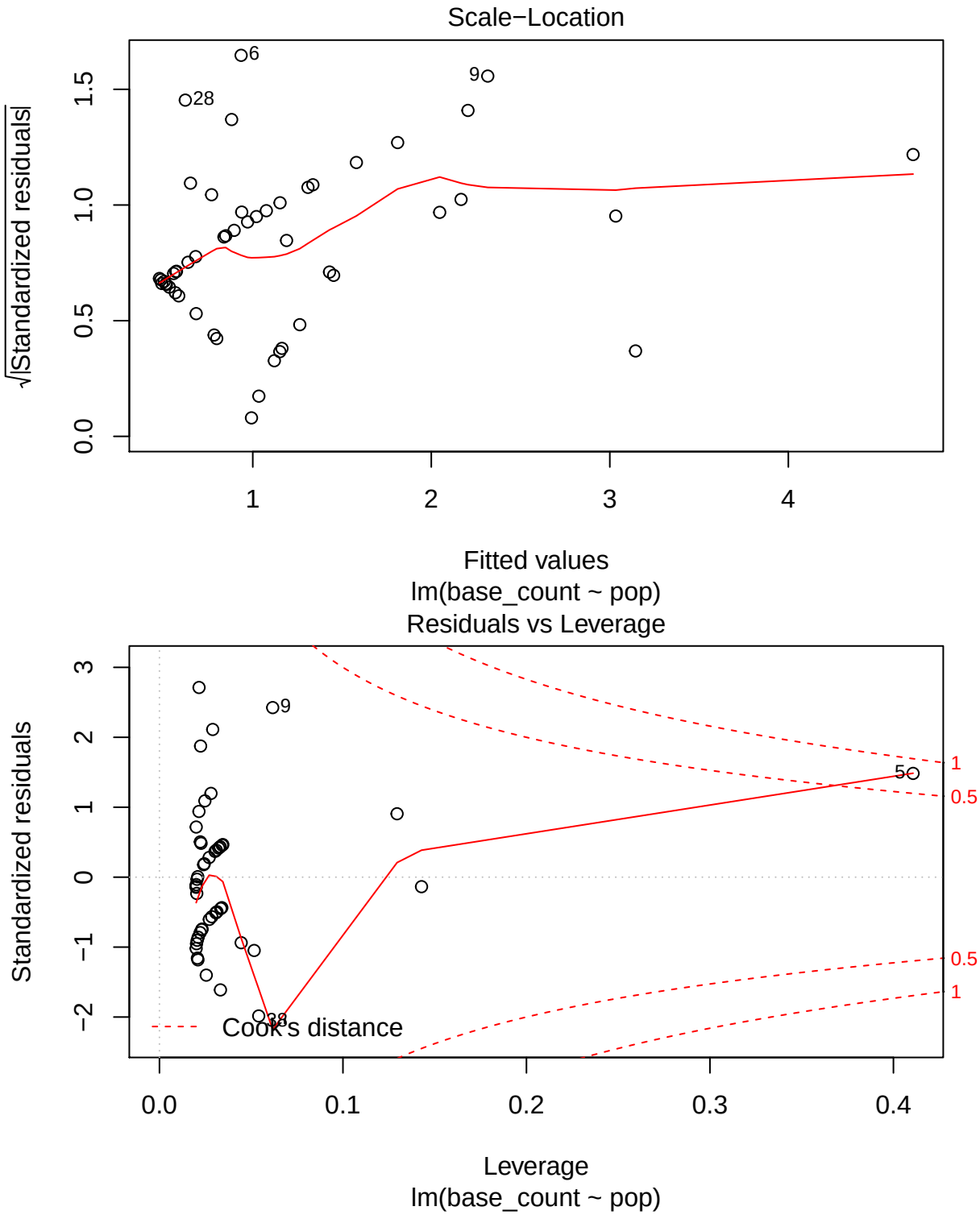
```
ggplot(data=df, aes(x=pop, y=base_count)) + geom_point() + stat_smooth(method='lm', se=F)
```



From this graph we can see a loose but definite positive relationship between the two variables.

```
plot(fit2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(base_count ~ pop)
6
9
28
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(base_count ~ pop)



There is a clear pattern in the plot of residuals vs fitted values due to the small number of discrete values which `base_count` can take on. The normal Q-Q plot is almost a straight line, except at the high end where the residuals are larger than would be expected for a normal distribution. In the plot of residuals vs leverage, California (point number 5) once again stands out as a significant outlier. TODO: Analyze more

Additionally, it appears as though the variance in this relationship may be heteroscedastic. We can test for this using the Breusch-Pagan test

```
bptest(fit2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit2
## BP = 2.1312, df = 1, p-value = 0.1443
```

Although the error appears heteroscedastic, according to the test results there is no significant evidence of heteroscedasticity.

Knowing that a relationship exists between `pop` and `sightings`, as well as `base_count` and `pop`, we can normalize the number of sightings by the population before fitting against `base_count` to account for these relationships. (TODO: can we?)

4.1.4 Number of USAF bases per state vs sightings per thousand people per state

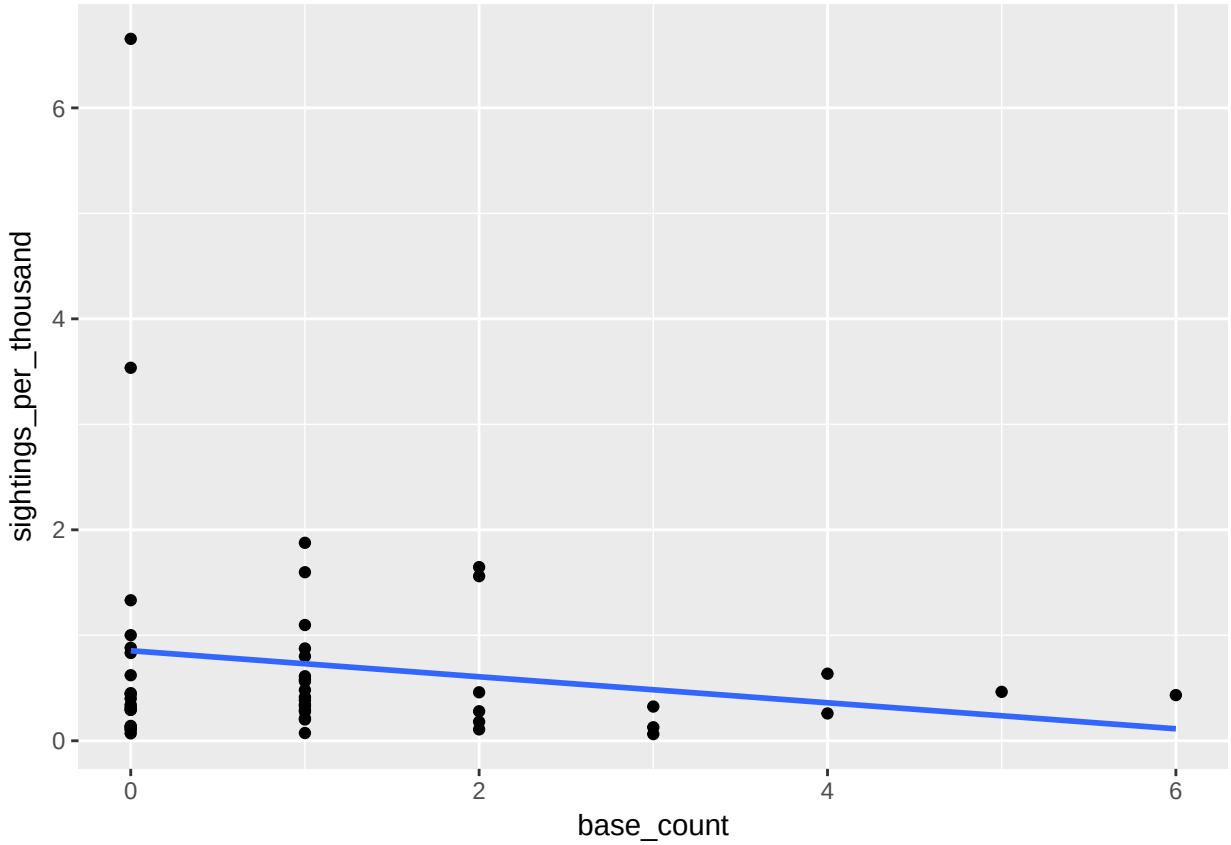
```
### Normalize the number of sightings by the population of each state
df$sightings_per_thousand <- (df$sightings / df$pop) * 1000

(s3 <- summary(fit3 <- lm(sightings_per_thousand ~ base_count, data=df)))
```

```
##
## Call:
## lm(formula = sightings_per_thousand ~ base_count, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7833 -0.4876 -0.3217  0.1260  5.8010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8537     0.1943   4.393 6.14e-05 ***
## base_count   -0.1233     0.1079  -1.142    0.259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.051 on 48 degrees of freedom
## Multiple R-squared:  0.02647,    Adjusted R-squared:  0.006191
## F-statistic: 1.305 on 1 and 48 DF,  p-value: 0.2589
```

With an 95% confidence interval for $\beta_{base\_count} = (-0.34, 0.094)$, we can confidently say there is no relationship between the number of US Air Force bases in a state and the population-adjusted number of reported UFO sightings in that state since 1970.

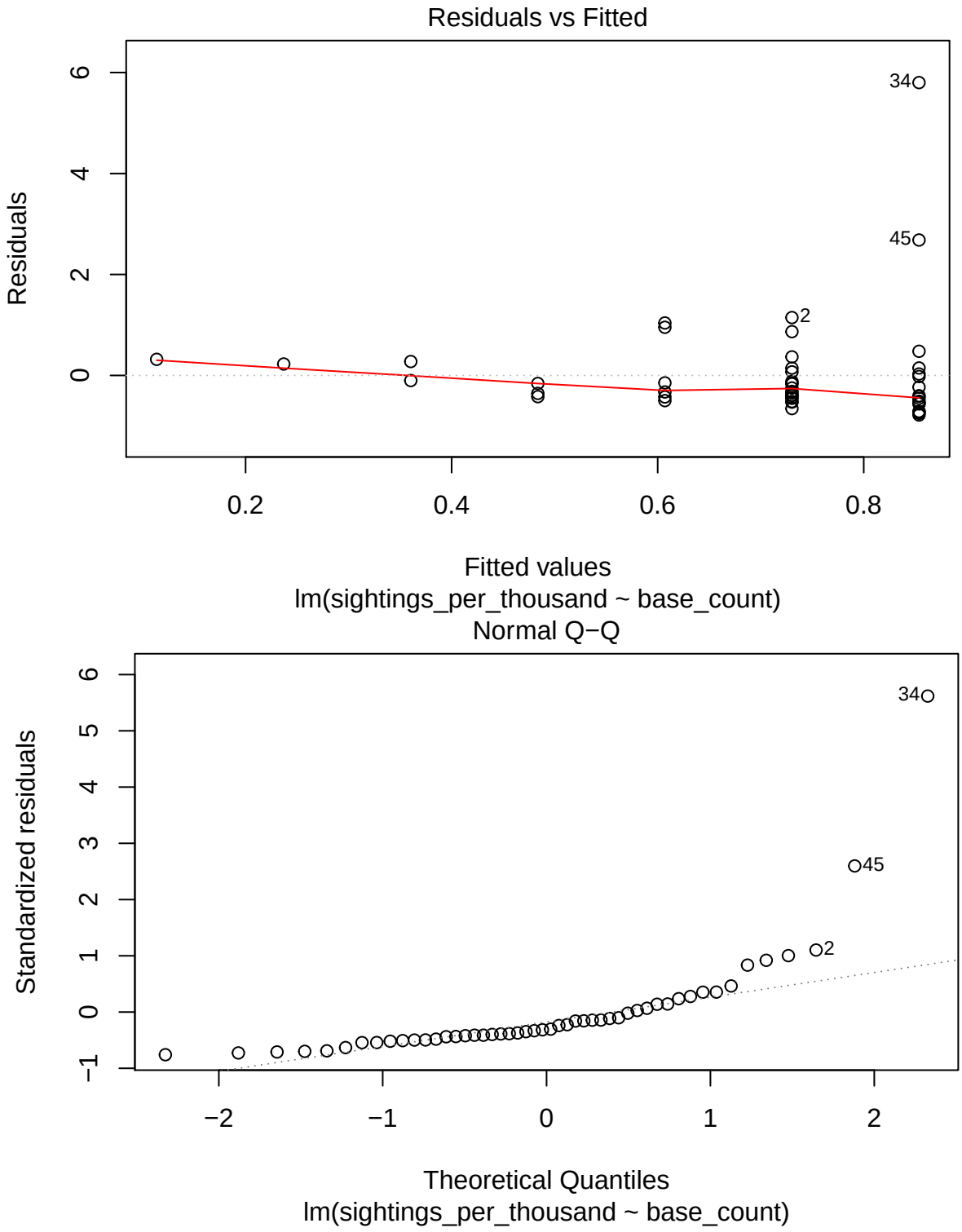
TODO: Analyze this relationship further

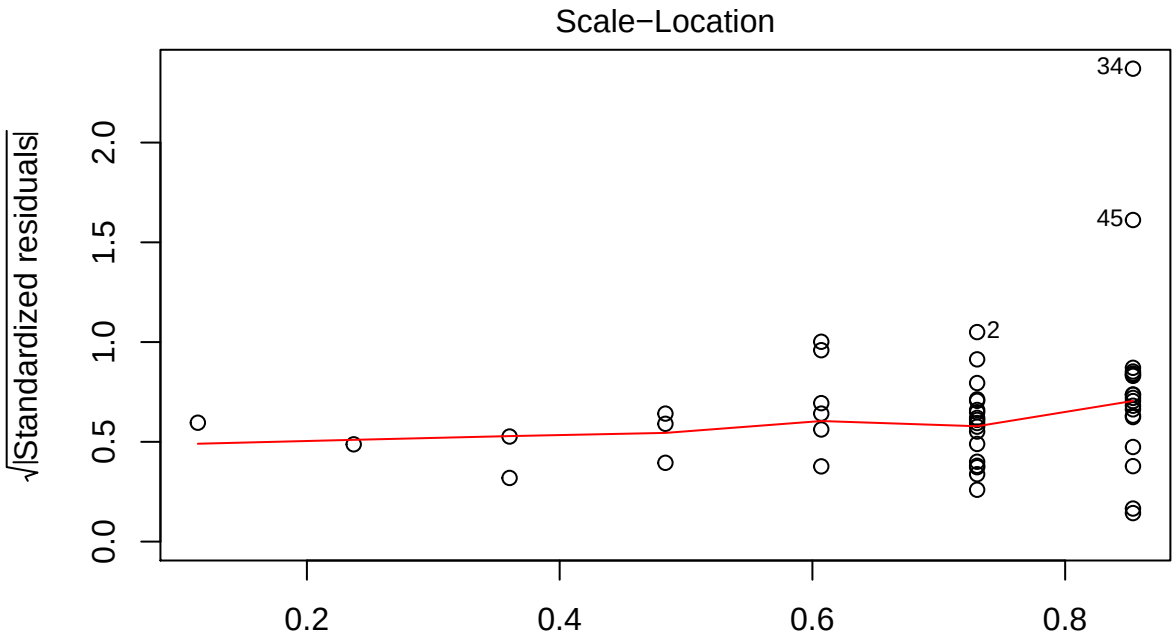
```
ggplot(data=df, aes(x=base_count, y=sightings_per_thousand)) + geom_point() + stat_smooth(metho
```

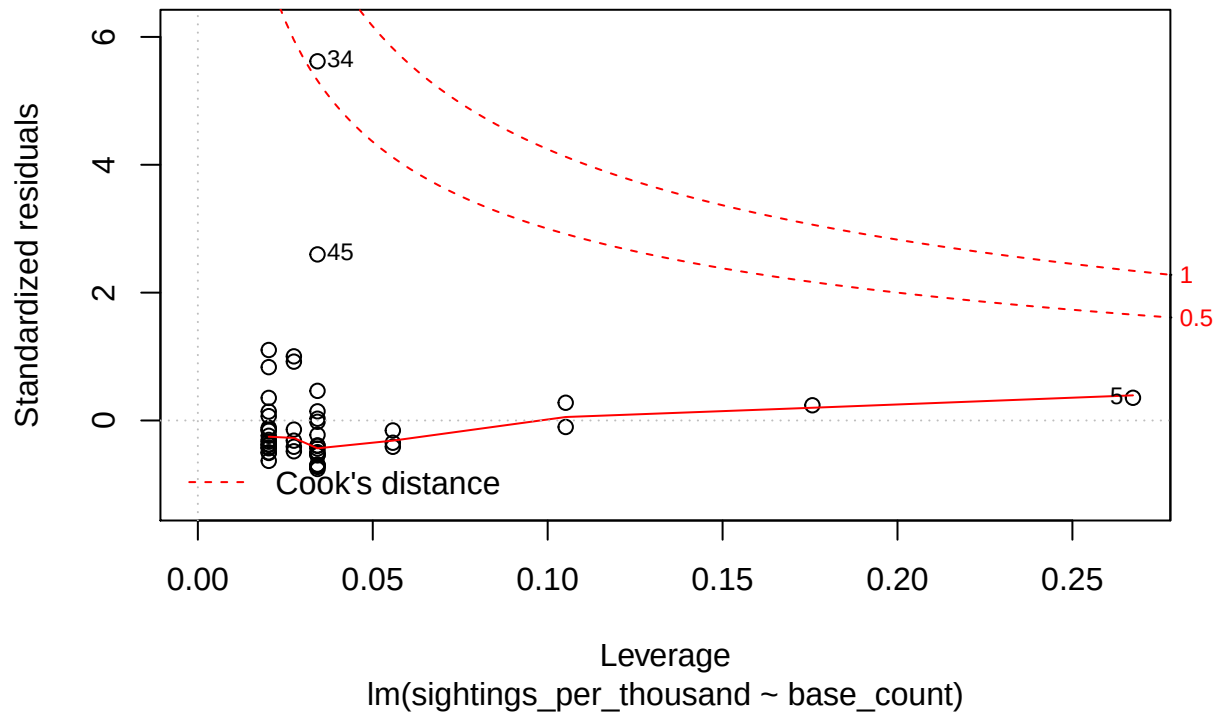


Graphically we can confirm the F-test results that there is little relationship between the two variables.

```
plot(fit3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(sightings_per_thousand ~ base_count)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sightings_per_thousand ~ base_count)
34
45
2





4.2 Per Capita GDP

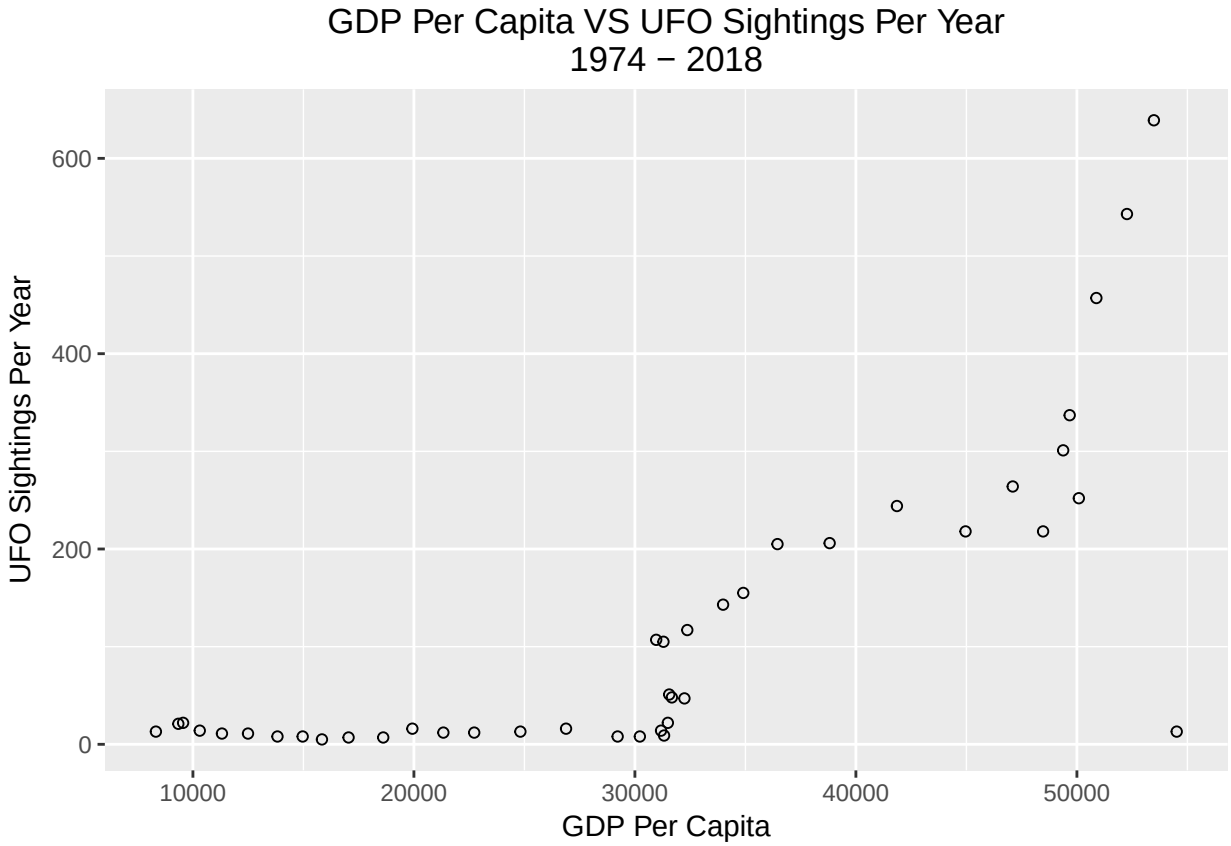
Per Capita GDP for the United States was pulled from Open Data Networks. Real GDP is an inflation-adjusted measure of each State's gross product that is based on national prices for the

goods and services produced within that State. This price is presented in current dollars at the time of the dataset's creation (last updated August 2016). Estimate of the population is based on midyear measurements from the Censue Bureau.

```
gdp_per_capita <- read.csv(file = "../../data/raw/gdp_per_capita_per_year.csv", header = TRUE)

fl_sightings <- filter(sight, tolower(state) == 'fl' )
fl_sightings$date_time <- as.Date(fl_sightings$date_time, "%m/%d/%y")
tmp <- lapply(strsplit(as.character(fl_sightings$date_time), "-"), `[[`, 1)
tmp2 <- sapply(tmp, "[[", 1)
fl_sightings$year <- as.numeric(tmp2)
fl_sightings_75 <- filter(fl_sightings, year >= 1974)
fl_sightings_75 <- filter(fl_sightings_75, year <= 2014)

sightings_per_year_fl <- as.data.frame(table(fl_sightings_75$year))

gdp_per_capita <- filter(gdp_per_capita, Year.and.category >= 1974)
gdp_sightings <- cbind(sightings_per_year_fl$Freq,gdp_per_capita$Per.capita.GDP..current....)
```

```
## Warning in cbind(sightings_per_year_fl$Freq, gdp_per_capita
## $Per.capita.GDP..current....): number of rows of result is not a multiple
## of vector length (arg 1)
```

```
gdp_sightings <- as.data.frame(gdp_sightings)

colnames(gdp_sightings) <- c("Sightings_Per_Year", "GDP_Per_Capita")

gdp.results <- lm(Sightings_Per_Year ~ GDP_Per_Capita, data = gdp_sightings)

ggplot(gdp_sightings, aes(x =GDP_Per_Capita, y =Sightings_Per_Year )) + geom_point(shape=1) + l
```



The scatterplot suggests that some kind of polynomial might fit the data best - a cubic term comes to mind. However, it's too soon for us to worry about modeling this term before we get a look at the a full model.

```
lm.fit <- lm(Sightings_Per_Year ~ GDP_Per_Capita,data = gdp_sightings)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Sightings_Per_Year ~ GDP_Per_Capita, data = gdp_sightings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -307.81  -53.68   -5.14   37.28  327.00
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.442e+02  3.839e+01  -3.756  0.00055 ***
## GDP_Per_Capita  8.530e-03  1.142e-03   7.472 4.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.3 on 40 degrees of freedom
## Multiple R-squared:  0.5826, Adjusted R-squared:  0.5721
```

```
## F-statistic: 55.83 on 1 and 40 DF,  p-value: 4.167e-09
```

R^2 is looking much better than it did with our aliens_movies data, capturing over 50% of the variation.

4.3 Portion of population with internet access by year

5 Data Analysis

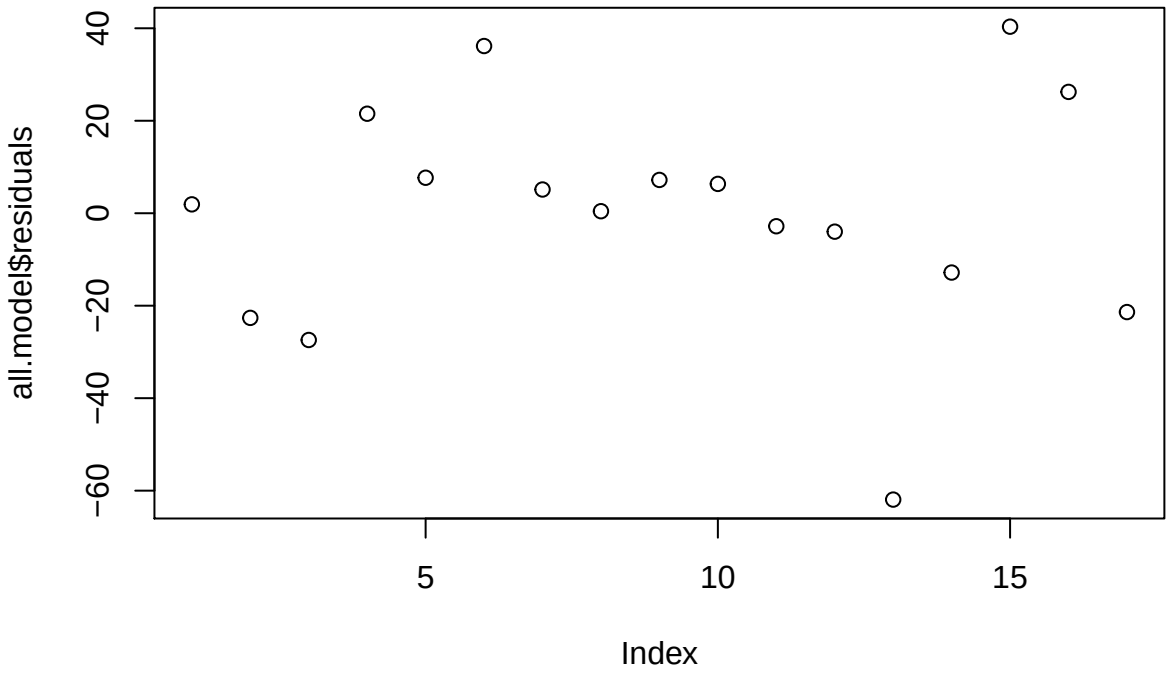
5.1 Model Selection

All the data discussed in the EDA section was pulled into one dataset to be imported into R.

```
all_florida <- read.csv(file = "../../data/raw/all_florida.csv", header = TRUE)
all_florida <- filter(all_florida, year >= 1998)

internet_rurality <- read.csv(file = "../../data/raw/internet_by_rurality.csv", header = TRUE)

all.model <- lm(sightings_year ~ per_capita_gdp_current + breweries + num_movies, data = all_fl
summary(all.model)
```

```
##
## Call:
## lm(formula = sightings_year ~ per_capita_gdp_current + breweries +
##     num_movies, data = all_florida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.932 -12.831   1.921   7.676  40.347
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -6.876e+02  4.464e+01 -15.402 9.96e-10 ***
## per_capita_gdp_current  1.557e-02  9.562e-04  16.286 4.99e-10 ***
## breweries               3.047e+00  2.594e-01  11.750 2.69e-08 ***
## num_movies              1.597e+00  1.812e+00   0.882    0.394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.75 on 13 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9669
## F-statistic: 156.7 on 3 and 13 DF,  p-value: 1.873e-10
```

```
plot(all.model$residuals)
```



Unfortunately, there are several NAs in our database which makes it difficult to use all the variables in our model. Internet by rurality only exists for 2000 to 2015, which is a very small sample size.

Since we're choosing to only model Florida as a time series, the base distance from an air force base will also not make sense to include in the model.

```
allpossreg <- regsubsets(sightings_year ~ per_capita_gdp_current + breweries + num_movies, nbes

aprout <- summary(allpossreg)

with(aprout,round(cbind(which,rsq,adjr2,cp,bic),3))
```

```
##   (Intercept) per_capita_gdp_current breweries num_movies   rsq adjr2
## 1           1                       1         0          0 0.687 0.666
## 1           1                       0         1          0 0.275 0.227
## 1           1                       0         0          1 0.130 0.072
## 2           1                       1         1          0 0.971 0.967
## 2           1                       1         0          1 0.687 0.643
## 2           1                       0         1          1 0.424 0.342
## 3           1                       1         1          1 0.973 0.967
##        cp     bic
## 1 138.169 -14.082
## 1 336.964   0.188
## 1 407.192   3.297
## 2   2.777 -51.970
## 2 140.055 -11.262
## 2 267.221  -0.879
## 3   4.000 -50.124
```

5.1.1 R-Squared

R^2 is a measure of variability in the dependent variable captured with the variables. As mentioned in the lecture, regardless of any other measures of quality, a low r-squared would indicate that our independent variables were just not appropriate. Luckily, many of our models clear .60 with and two are even over .90. This gives us some level of confidence in our model going forward. However, r-squared does not penalize multiple variables, so we cannot just choose the highest r-squared and believe we have a flexible model to approach other states with.

5.1.2 Adjusted R-Squared

Adjusted R-Squared is better at punishing variables included in a model that do not add much to its predictive power. Still, we have two r-squareds with over .90. Both those models, one including number of alien movies per year (featured as num_movies here) and one without, have almost equal adjusted r-squareds. Based only on this information, it would be difficult to choose a model. However, we do not want to stop at R^2 regardless because it is too lenient with unnecessary additions to the model.

5.1.3 BIC

BIC (Bayesian Information Criterion) does well with judging the best model for the observed data if one of the models up for consideration is the true model. However, the true model for how reports are submitted to the National UFO Reporting Center's website is most likely tossed up with it's ranking in Google Search Results, the number of employees/volunteers the organization has, and other factors not measured by our particular data. We are more focused on creating a working approximate model rather than the "true" model and for this reason with favor AIC.

5.1.4 Mallow's CP

A small value of CP is associated with a relatively precise model. The smallest CP here is associated with our model that drops the number of alien movies.

This isn't surprising. The movies only model we created earlier did not seem to have impressive predictability. It is also worth noting it only addresses the raw number of movies coming out, and they are in no way weighted by their popularity in the popular consciousness. Due to these flaws, we do not see a strong theoretical reason these values should be included in the model.

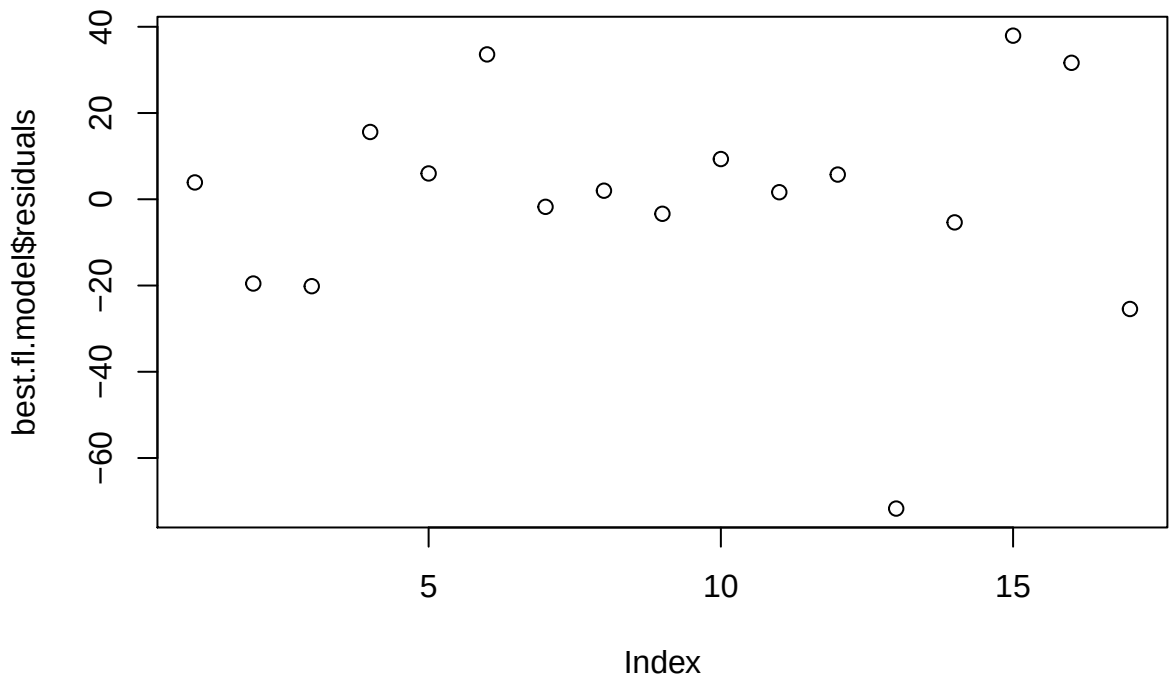
Because of this, we feel it is acceptable to drop the number of movies from the model and select the model with the lowest value for Mallow's CP.

```
best.fl.model <- lm(sightings_year ~ per_capita_gdp_current + breweries, data = all_florida)

summary(best.fl.model)
```

```
##
## Call:
## lm(formula = sightings_year ~ per_capita_gdp_current + breweries,
```

```
##     data = all_florida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.720  -5.361   1.985   9.330  37.937
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -6.835e+02  4.405e+01  -15.52 3.25e-10 ***
## per_capita_gdp_current  1.592e-02  8.616e-04   18.48 3.12e-11 ***
## breweries               3.037e+00  2.570e-01   11.82 1.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.53 on 14 degrees of freedom
## Multiple R-squared:  0.9715, Adjusted R-squared:  0.9674
## F-statistic: 238.4 on 2 and 14 DF,  p-value: 1.536e-11
```

```
plot(best.fl.model$residuals)
```



5.2 Diagnostics

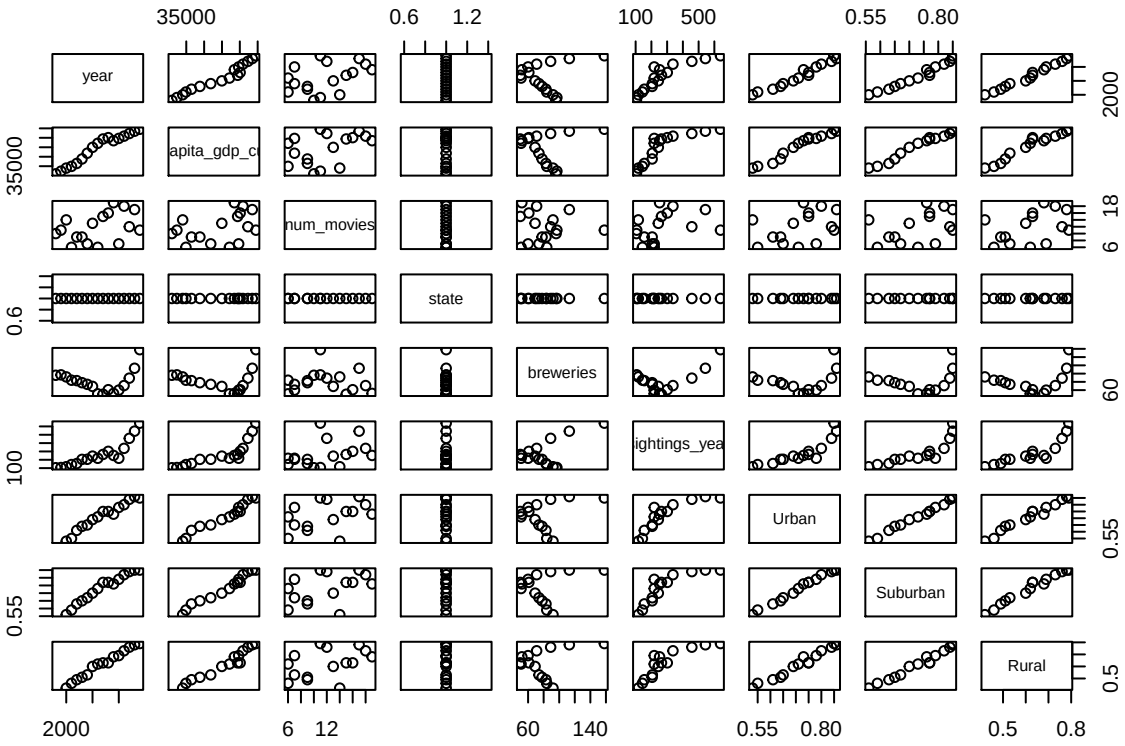
Just because we've use model selection to choose the best model doesn't mean we have a perfect model.

In fact, for us to put faith in our least squares regression model there are a variety of assumptions we've made. The error terms are supposed to be independent. The error should be normally distributed. The errors should have a constant variance. We don't have a way to know the true error, so we will use the residuals for the model as an approximation of the error.

If these assumptions are violated, we could have multicollinearity in our models (indicating our independent variables are correlated), we could have heteroskedasticity, or patterns in the residuals suggesting missing independent variables in the models.

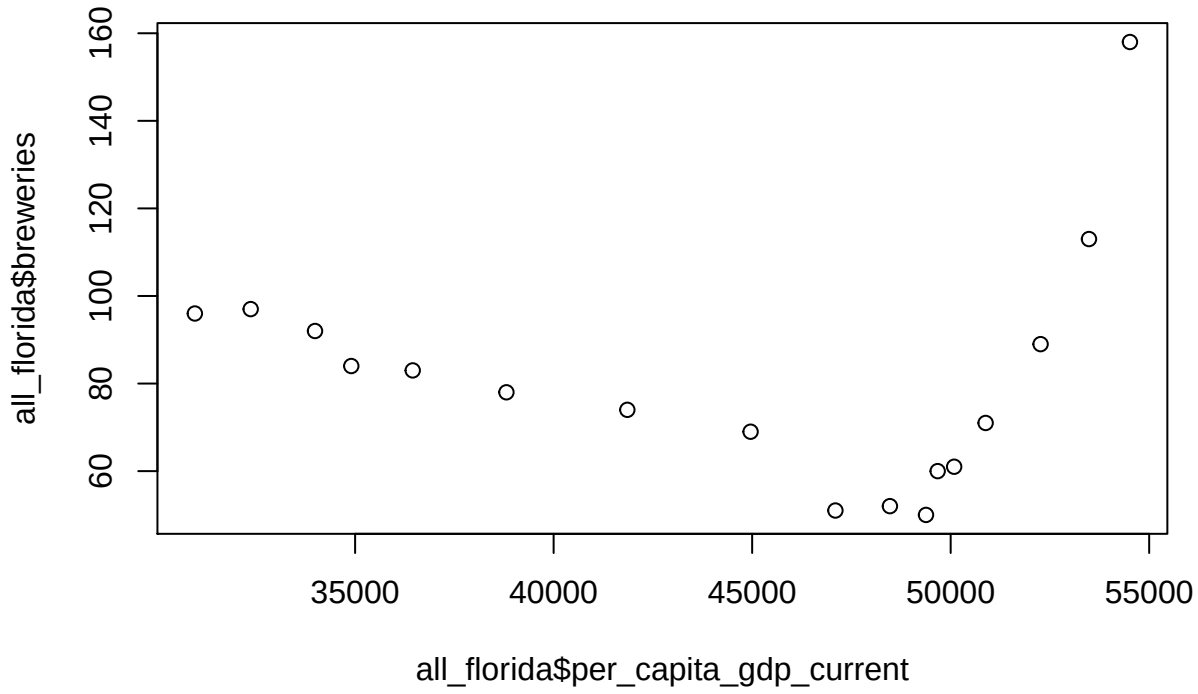
5.2.1 Pairwise Scatterplots

```
plot(all_florida)
```



There do appear to be some non-linear associations between our variables breweries and per-capita gdp. Since these are the only two variables in our model, this creates some concern there may be some multi-collinearity. Let's see that more closely.

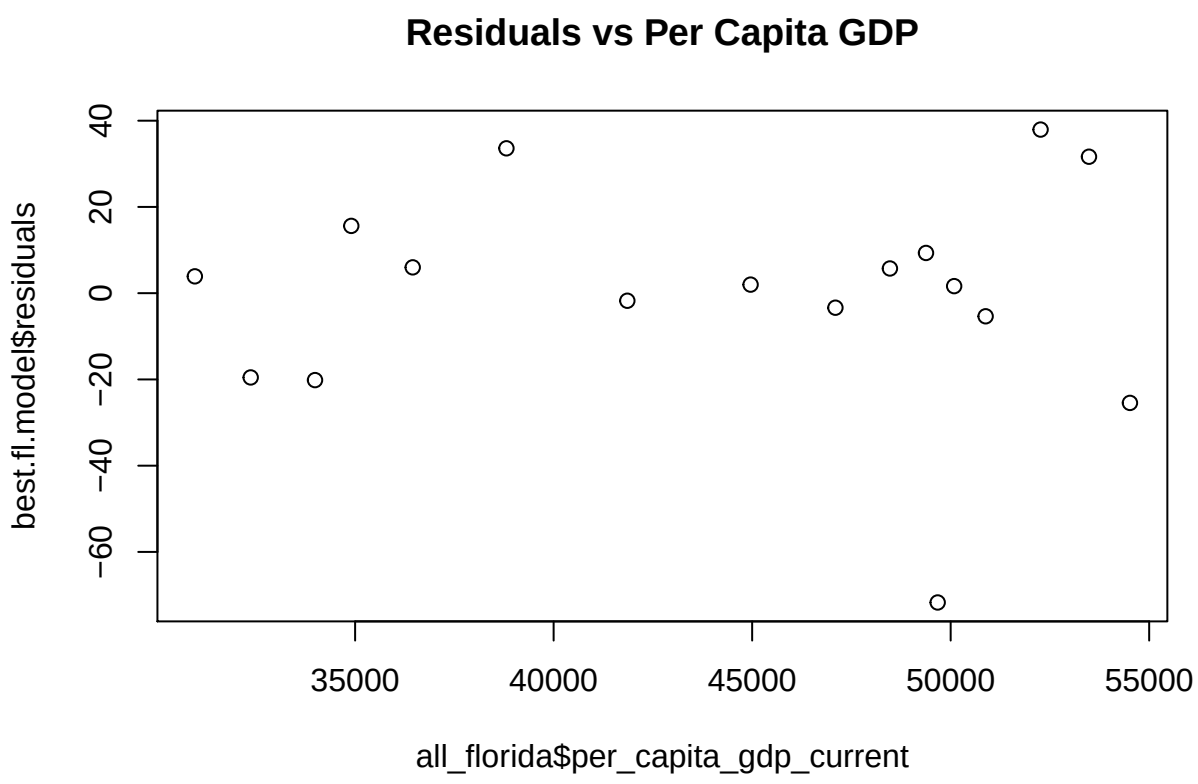
```
plot(all_florida$per_capita_gdp_current, all_florida$breweries)
```



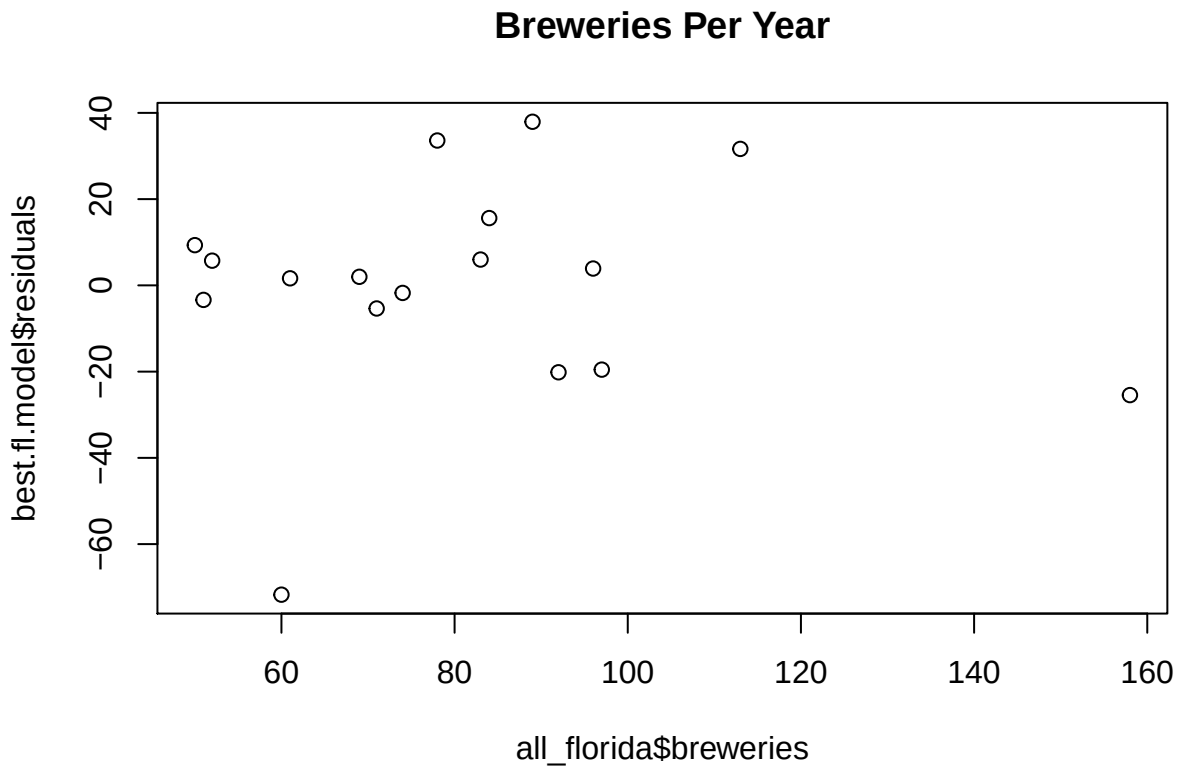
There definitely appears to be a pattern here. With 3 turns, it looks like it may be some cubic predictor.

Below, we plot the residuals against both our predictors.

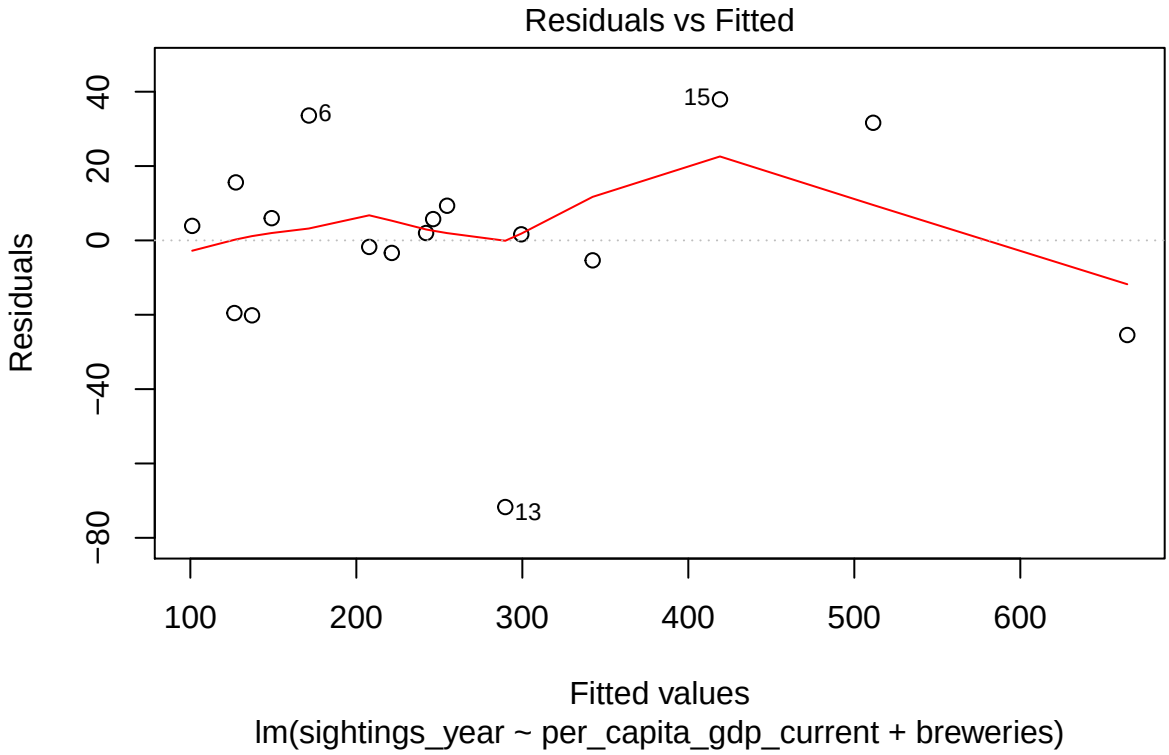
```
plot(all_florida$per_capita_gdp_current,best.fl.model$residuals, main = "Residuals vs Per Capi
```



```
plot(all_florida$breweries, best.fl.model$residuals, main = "Breweries Per Year")
```




```
plot(best.fl.model, which = 1)
```



The residuals vs the fitted values show that we have a few definite outliers at point 6, 13, and 15.

5.2.2 Heteroskedasticity

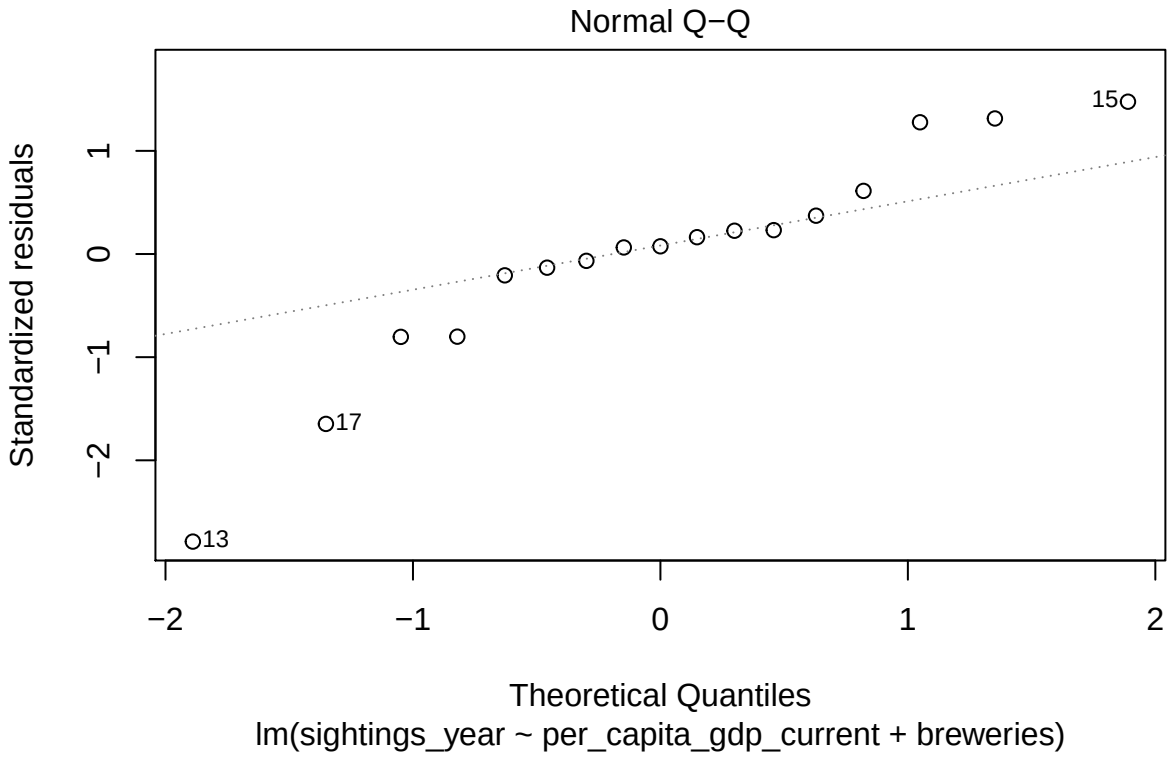
```
ols_bp_test(best.fl.model)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                   Data
##  -----------------------------------------
##  Response : sightings_year
##  Variables: fitted values of sightings_year
##
##         Test Summary
##  ---------------------------
##  DF            =    1
##  Chi2          =    1.277793
##  Prob > Chi2   =    0.2583098
```

The test for Breusch Pagan test fails to reject the null hyptothesis that our variance is constant, suggesting we don't have heteroskedasticity in our model. While the formal mechanical method fails to catch it, our plots suggest there may be non-constant variance in our model.

5.2.3 Normality

```
plot(best.fl.model, which = 2)
```




```
shapiro.test(residuals(best.fl.model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(best.fl.model)
## W = 0.90185, p-value = 0.07295
```

The QQ-plot does not look good. The Shapiro-Wilk test just barely fails to reject the null hypothesis that the data are sampled from a normally distributed population. So the test suggests the data are normal, but what we observe in the Normal Q-Q plot. As we have a small sample, it makes sense that our test might fail to reject the null hypothesis due to random variation expected in the sample anyway.

5.3 Cross-Validation

(TBD - using the model we've made to predict the values of sightings for other states as part of cross-validation)

6 Conclusion

6.1 Results

It is likely that the true value we are measuring is population. A greater number of breweries is most likely capturing the actual value of interest, which is the population of the state. A state

with a higher population is more likely to more breweries. This does not necessarily capture the "amount of alchohol consumed per capita" that we were interested in capturing.

Still, understanding we may not have created a model allowing for statistical inference, we still have some predictive power.

6.2 Future Work

6.2.1 Implementing Advanced Spatial-Temporal & Time-Series Models

With the limited scope of our understanding of statistics, we are unable to compare all five explanatory factors with the predicted variable, sightings per year. As we learn more advanced statistical techniques, we can come back to this analysis and improve on modelling with spatial and time dependencies.

6.2.2 Greater Sample Size

The number of sightings was healthy, over 80,000 reports filed. However, the other data was highly constricted in the amount of years it contained, some having as few as 15 observations. This is insufficient to create the kind of robust model that would be effective for all 50 states.

6.2.3 Investigate Other States

Here, we identified 5 states that were "super states" in their number of sightings as identified as the density of sightings per state. It may be the case that the number of sightings in these super states and lesser states are fundamentally different, requiring a different model. Or, the "super state" identification may actually be a variable that needs to be considered with a dummy variable to change the intercept for as oppossed to "regular" states.

6.2.4 Investigate Most Popular Alien Movies

Instead of using alien movies per year, we might use the amount of money grossed in aliens films per year. This would give a greater indicator of a film's affect on the popular consciousness. A movie like "Aliens" surely had a greater effect than "Mars Needs Moms", for example. Our current scraping of the data, relying on each movie as a equally effective unit and simply counting the number of movies per year is likely missing a lot of that information.

7 Works Cited/ Data Sources

- Blumenthal, R. (2017, April 24). People Are Seeing U.F.O.s Everywhere, and This Book Proves It. New York Times. https://www.nytimes.com/2017/04/24/science/ufo-sightings-book.html
- National UFO Reporting Center. (2017). [Website data] (http://www.nuforc.org/webreports.html)

- Open Data Network. (2012). [Data and codebook] Available from Opendatanetwork.gov Website (https://www.opendatanetwork.com/dataset/data.hawaii.gov/qnar-gix3)
- Quinton Mason. (2017). Brewery Count by State (1984 - March 31, 2017) [Data file and code book]. Available from Data.gov Website:https://catalog.data.gov/dataset/brewery-count-by-state-1984-march-31-2017
- Wikipedia. (2017). List of Films Featuring Extraterrestrials [Website data] Available from Wikipedia Website (https://en.wikipedia.org/wiki/List_of_films_featuring_extraterrestrials)
- Wikipedia. (2017). List of United States Air Force Installations [Website data] Available from Wikipedia Website (https://en.wikipedia.org/wiki/List_of_United_States_Air_Force_installations)