Final Project

Dina Stretiner

Exploratory Data Analysis - Data Cleaning and Transformations

```
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(glmtoolbox)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
library(leaps)

# Upload data
setwd = setwd('/Users/dvstretiner/Documents/CU Denver - Stats/MATH 5387 Regression Analysis/Project/')
energy = read.csv('energy_burden_for_project.csv')

#Preview data
str(energy)
```

```
## 'data.frame':    3108 obs. of  24 variables:
##  $ county_fips                                : int  1001 1003 1005 1007 1009 1011 1013 1015 1017 101
##  $ county                                     : chr  "Autauga" "Baldwin" "Barbour" "Bibb" ...
##  $ state                                      : chr  "AL" "AL" "AL" "AL" ...
##  $ county_pop                                 : int  55200 208107 25782 22527 57645 10352 20025 11509
##  $ bev_2018_reference_vehicle_counts          : int  151 446 40 28 72 32 28 222 52 115 ...
##  $ hev_gasoline_2018_reference_vehicle_counts : int  619 2591 286 132 454 136 109 1056 260 296 ...
##  $ phev_2018_reference_vehicle_counts         : int  37 189 17 12 31 6 11 125 37 39 ...
##  $ icev_gasoline_2018_reference_vehicle_counts: int  47791 167169 20669 17660 50846 8471 16081 101555
##  $ energyburden_indicator                     : num  0.0458 0 0.1222 0 0 ...
##  $ energyburden_1_prop                        : num  0.583 0.742 0 0 0 ...
##  $ energyburden_2_prop                        : num  0.333 0.258 0.778 1 1 ...
##  $ energyburden_3_prop                        : num  0.0833 0 0.2222 0 0 ...
##  $ energyburden_4_prop                        : num  0 0 0 0 0 ...
##  $ unemprate2020                              : num  4.9 5.6 7 6.6 4.1 5.5 8.8 7.1 6.8 4.6 ...
```

```
##  $ pctempmining                              : num  0.355 0.258 0 1.896 0.634 ...
##  $ ruralurbancontinuumcode2013               : int  2 3 6 1 1 6 6 3 6 6 ...
##  $ type_2015_farming_no                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ type_2015_mining_no                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ population_loss_2015_update               : int  0 0 0 0 0 0 1 0 1 0 ...
##  $ perpov_1980_0711                          : int  0 0 1 0 0 1 1 0 0 0 ...
##  $ minority_indicator                        : num  0 0 0.4087 0.0667 0 ...
##  $ lowincome_indicator                       : num  0.0813 0.0234 0.6087 0.2667 0.0857 ...
##  $ lessthanhs_indicator                      : num  0.0813 0 1.0104 0.4867 0.6257 ...
##  $ cancer_indicator                          : num  0.8 0.0473 0.8 0.8 0.4571 ...
```

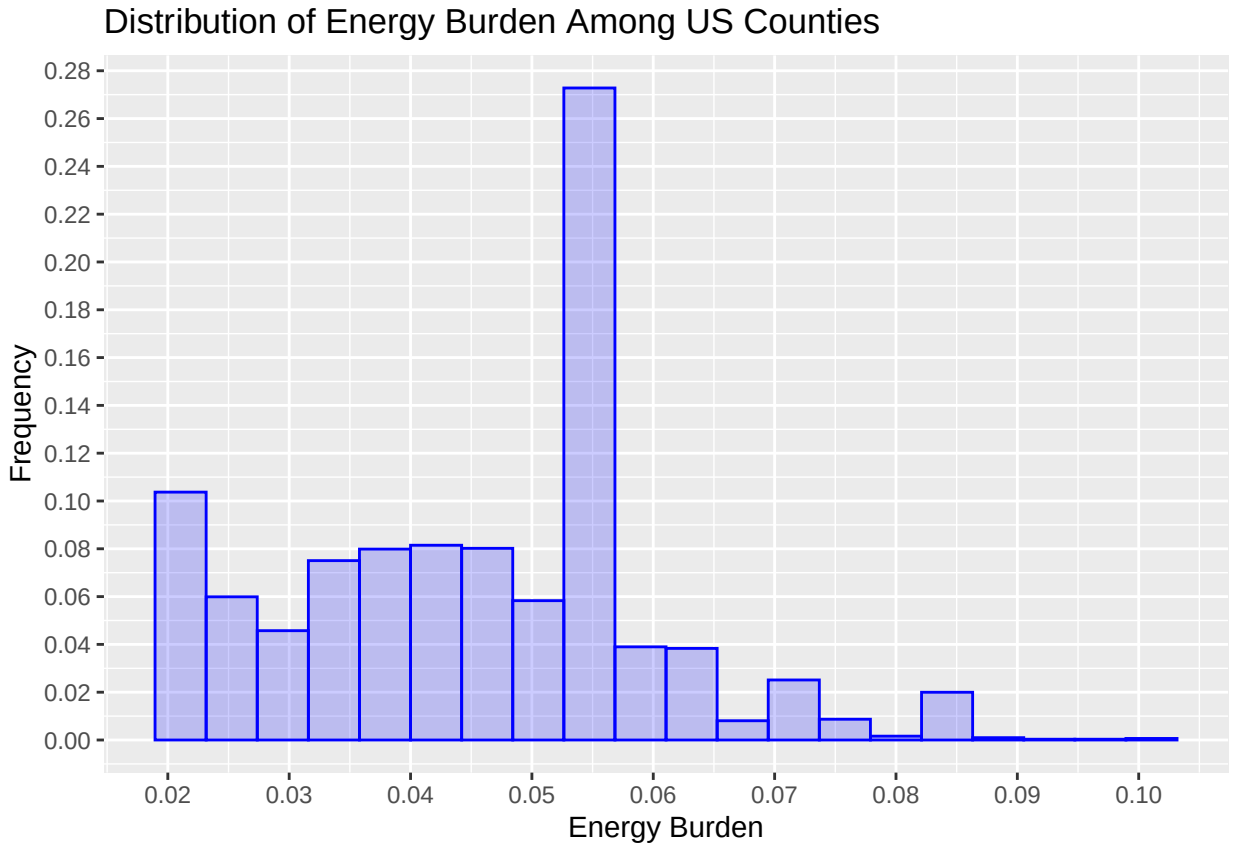
```
# Numerical Summaries
summary(energy)
```

```
##   county_fips       county             state              county_pop
##  Min.   : 1001   Length:3108        Length:3108        Min.   :     102
##  1st Qu.:19044   Class :character   Class :character   1st Qu.:   11204
##  Median :29212   Mode  :character   Mode  :character   Median :   25878
##  Mean   :30672                                         Mean   :  103199
##  3rd Qu.:46008                                         3rd Qu.:   67371
##  Max.   :56045                                         Max.   :10098052
##
##  bev_2018_reference_vehicle_counts hev_gasoline_2018_reference_vehicle_counts
##  Min.   :     0.0                  Min.   :     0
##  1st Qu.:    24.0                  1st Qu.:   153
##  Median :    75.0                  Median :   387
##  Mean   :   661.5                  Mean   :  2068
##  3rd Qu.:   209.8                  3rd Qu.:  1104
##  Max.   :147766.0                  Max.   :376448
##  NA's   :2                         NA's   :2
##  phev_2018_reference_vehicle_counts icev_gasoline_2018_reference_vehicle_counts
##  Min.   :     0.00                  Min.   :     99
##  1st Qu.:    12.00                  1st Qu.:   9455
##  Median :    37.00                  Median :  21416
##  Mean   :   311.86                  Mean   :  74209
##  3rd Qu.:    99.75                  3rd Qu.:  53762
##  Max.   :107281.00                  Max.   :6463113
##  NA's   :2                          NA's   :2
##  energyburden_indicator energyburden_1_prop energyburden_2_prop
##  Min.   :0.00000        Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.00000        1st Qu.:0.0000      1st Qu.:0.3125
##  Median :0.00000        Median :0.2500      Median :0.6000
##  Mean   :0.04289        Mean   :0.3432      Mean   :0.5791
##  3rd Qu.:0.01946        3rd Qu.:0.6211      3rd Qu.:0.9231
##  Max.   :0.60000        Max.   :1.0000      Max.   :1.0000
##  NA's   :1              NA's   :1           NA's   :1
##  energyburden_3_prop energyburden_4_prop unemprate2020     pctempmining
##  Min.   :0.00000     Min.   :0.00000     Min.   : 1.700   Min.   : 0.00000
##  1st Qu.:0.00000     1st Qu.:0.00000     1st Qu.: 5.200   1st Qu.: 0.07504
##  Median :0.00000     Median :0.00000     Median : 6.500   Median : 0.28767
##  Mean   :0.07431     Mean   :0.00337     Mean   : 6.706   Mean   : 1.49915
##  3rd Qu.:0.03030     3rd Qu.:0.00000     3rd Qu.: 8.000   3rd Qu.: 1.18002
##  Max.   :1.00000     Max.   :1.00000     Max.   :22.500   Max.   :45.41642
##  NA's   :1           NA's   :1
##  ruralurbancontinuumcode2013 type_2015_farming_no type_2015_mining_no
```

```
##  Min.   :1.000              Min.   :0.0000       Min.   :0.00000
##  1st Qu.:2.000              1st Qu.:0.0000       1st Qu.:0.00000
##  Median :6.000              Median :0.0000       Median :0.00000
##  Mean   :4.987              Mean   :0.1429       Mean   :0.07014
##  3rd Qu.:7.000              3rd Qu.:0.0000       3rd Qu.:0.00000
##  Max.   :9.000              Max.   :1.0000       Max.   :1.00000
##
##  population_loss_2015_update perpov_1980_0711 minority_indicator
##  Min.   :0.0000              Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000              1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000              Median :0.0000   Median :0.00000
##  Mean   :0.1689              Mean   :0.1129   Mean   :0.10723
##  3rd Qu.:0.0000              3rd Qu.:0.0000   3rd Qu.:0.09753
##  Max.   :1.0000              Max.   :1.0000   Max.   :1.36000
##
##  lowincome_indicator lessthanhs_indicator cancer_indicator
##  Min.   :0.0000      Min.   :0.0000       Min.   :0.00000
##  1st Qu.:0.0000      1st Qu.:0.0000       1st Qu.:0.00000
##  Median :0.1381      Median :0.1311       Median :0.00000
##  Mean   :0.2115      Mean   :0.2747       Mean   :0.07696
##  3rd Qu.:0.3389      3rd Qu.:0.4429       3rd Qu.:0.00000
##  Max.   :1.2000      Max.   :1.6800       Max.   :0.80000
##
```

All vehicle count variables have large outliers, judging by the max value compared to the 3rd quartile. Cancer indicator variable is mostly zero. Factor variables, such as farming and population loss need to be converted to such. Energy burden response variable needs to be calculated based on a weighted average

```
# Calculate weighted average energy burden variable based on the midpoint of the interval:
# < 4% for energy_burden_1: avg 2%
# 4 - 7% for energy_burden_2: avg 5.5%
# 7 - 10% for energy_burden_3: avg 8.5%
# >10% for energy_burden_4: assume 10% to be conservative
energy$energy_burden = 0.02 * energy$energyburden_1_prop + 0.055 * energy$energyburden_2_prop + 0.085 *

# Remove N/A's
energy = energy[complete.cases(energy),]

# Look at the distribution of energy burden
ggplot(energy, aes(x = energy_burden)) + geom_histogram(aes(y = ..count.. / sum(..count..)), color = 'bl
scale_x_continuous(breaks = seq(min(energy$energy_burden), max(energy$energy_burden), by = 0.01)) +
scale_y_continuous(breaks = seq(0, 0.4, by = 0.02)) + xlab("Energy Burden") + ylab("Frequency")
```




```
# Look at the numeric summary of energy_burden
summary(energy$energy_burden)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02000 0.03400 0.04750 0.04537 0.05500 0.10000
```

```
# Find mode of the distribution
library(modeest)
mlv(energy$energy_burden, method = 'mfv')
```

```
## [1] 0.055
```

75% of counties have average energy burden of <= 5.5%. In fact, this is the mode of the distribution, as seen on the histogram plot. This is potentially concerning, given that 5.5% was our manufactured estimate for the 4-7% bucket of the energy burden. This could lead to a biased model.

```
# Perform necessary transformations on variables

# Make copy of the dataframe
energy2 = energy

# Convert categorical variables to factors
energy2$rural_urban_flag = as.factor(energy$ruralurbancontinuumcode2013)
energy2$farming_flag = as.factor(energy$type_2015_farming_no)
energy2$pop_loss_flag = as.factor(energy$population_loss_2015_update)
energy2$poverty_flag = as.factor(energy$perpov_1980_0711)
energy2$state = as.factor(energy2$state)

# Take log of the large value variables: population and vehicle counts
```

```r
energy2$log_pop = log10(energy$county_pop)
energy2$log_battery_electric_vehicles = log10(energy$bev_2018_reference_vehicle_counts)
energy2$log_hybrid_vehicles = log10(energy$hev_gasoline_2018_reference_vehicle_counts)
energy2$log_plugin_hybrid_vehicles = log10(energy$phev_2018_reference_vehicle_counts)
energy2$log_internal_combustion_vehicles = log10(energy$icev_gasoline_2018_reference_vehicle_counts)
```

```r
# Remove duplicate factor variables
energy2 = energy2[, !(names(energy2) %in% c("ruralurbancontinuumcode2013", "type_2015_farming_no", "popu
```

```r
# Rename columns using simple names
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
energy2 = energy2 %>%
  rename(
     battery_electric_vehicles = bev_2018_reference_vehicle_counts,
     hybrid_vehicles = hev_gasoline_2018_reference_vehicle_counts,
     plugin_hybrid_vehicles = phev_2018_reference_vehicle_counts,
     internal_combustion_vehicles = icev_gasoline_2018_reference_vehicle_counts,
     unemployment_rate = unemprate2020,
     percent_employed_mining = pctempmining
  )
```

```r
# It appears that log10(0) = -Inf, this needs to be corrected. Replace -Inf with 0
energy2$log_battery_electric_vehicles[which(energy2$log_battery_electric_vehicles == -Inf)] = 0
energy2$log_hybrid_vehicles[which(energy2$log_hybrid_vehicles == -Inf)] = 0
energy2$log_plugin_hybrid_vehicles[which(energy2$log_plugin_hybrid_vehicles == -Inf)] = 0
energy2$log_internal_combustion_vehicles[which(energy2$log_internal_combustion_vehicles == -Inf)] = 0
```

```r
# Check numerical summary again for any irregularities
summary(energy2)
```

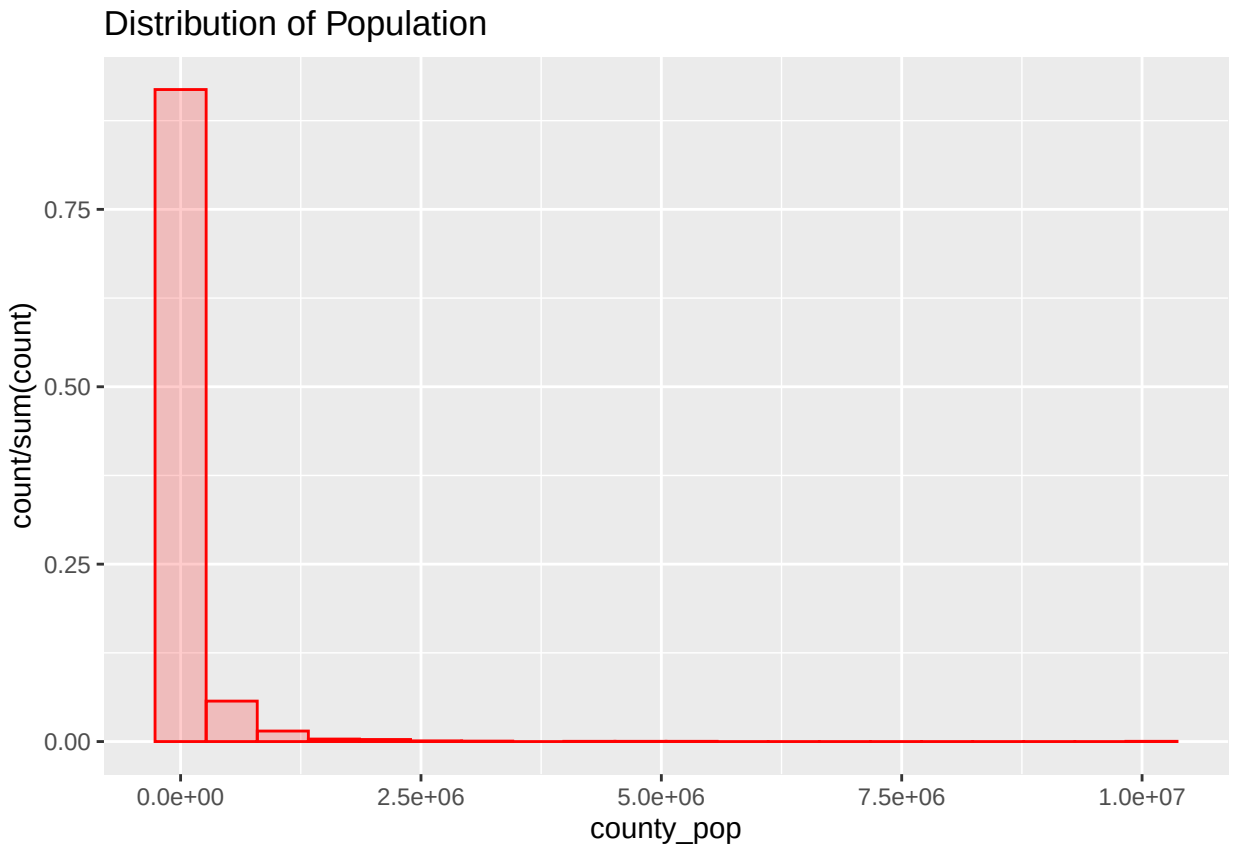
```
##   county_fips       county              state        county_pop
##  Min.   : 1001   Length:3105        TX     : 252   Min.   :     228
##  1st Qu.:19043   Class :character   GA     : 159   1st Qu.:   11215
##  Median :29209   Mode  :character   VA     : 133   Median :   25890
##  Mean   :30659                      KY     : 120   Mean   :  103284
##  3rd Qu.:46005                      MO     : 115   3rd Qu.:   67587
##  Max.   :56045                      KS     : 105   Max.   :10098052
##                                     (Other):2221
##  battery_electric_vehicles hybrid_vehicles  plugin_hybrid_vehicles
##  Min.   :     0.0          Min.   :     0   Min.   :     0.0
##  1st Qu.:    24.0          1st Qu.:   153   1st Qu.:    12.0
```

```
##  Median :    75.0          Median :   387   Median :    37.0
##  Mean   :   661.6          Mean   :  2068   Mean   :   311.9
##  3rd Qu.:   210.0          3rd Qu.:  1104   3rd Qu.:   100.0
##  Max.   :147766.0          Max.   :376448   Max.   :107281.0
##
##  internal_combustion_vehicles energyburden_indicator energyburden_1_prop
##  Min.   :     99              Min.   :0.00000        Min.   :0.0000
##  1st Qu.:   9449              1st Qu.:0.00000        1st Qu.:0.0000
##  Median :  21413              Median :0.00000        Median :0.2500
##  Mean   :  74224              Mean   :0.04292        Mean   :0.3431
##  3rd Qu.:  53762              3rd Qu.:0.01950        3rd Qu.:0.6200
##  Max.   :6463113              Max.   :0.60000        Max.   :1.0000
##
##  energyburden_2_prop energyburden_3_prop energyburden_4_prop unemployment_rate
##  Min.   :0.0000      Min.   :0.00000     Min.   :0.000000    Min.   : 1.700
##  1st Qu.:0.3125      1st Qu.:0.00000     1st Qu.:0.000000    1st Qu.: 5.200
##  Median :0.6000      Median :0.00000     Median :0.000000    Median : 6.500
##  Mean   :0.5792      Mean   :0.07436     Mean   :0.003372    Mean   : 6.707
##  3rd Qu.:0.9231      3rd Qu.:0.03030     3rd Qu.:0.000000    3rd Qu.: 8.000
##  Max.   :1.0000      Max.   :1.00000     Max.   :1.000000    Max.   :22.500
##
##  percent_employed_mining type_2015_mining_no minority_indicator
##  Min.   : 0.00000        Min.   :0.00000     Min.   :0.0000
##  1st Qu.: 0.07502        1st Qu.:0.00000     1st Qu.:0.0000
##  Median : 0.28713        Median :0.00000     Median :0.0000
##  Mean   : 1.48887        Mean   :0.07021     Mean   :0.1069
##  3rd Qu.: 1.17783        3rd Qu.:0.00000     3rd Qu.:0.0963
##  Max.   :45.41642        Max.   :1.00000     Max.   :1.3600
##
##  lowincome_indicator lessthanhs_indicator cancer_indicator  energy_burden
##  Min.   :0.0000      Min.   :0.0000       Min.   :0.00000   Min.   :0.02000
##  1st Qu.:0.0000      1st Qu.:0.0000       1st Qu.:0.00000   1st Qu.:0.03400
##  Median :0.1389      Median :0.1309       Median :0.00000   Median :0.04750
##  Mean   :0.2117      Mean   :0.2741       Mean   :0.07704   Mean   :0.04537
##  3rd Qu.:0.3393      3rd Qu.:0.4425       3rd Qu.:0.00000   3rd Qu.:0.05500
##  Max.   :1.2000      Max.   :1.6800       Max.   :0.80000   Max.   :0.10000
##
##  rural_urban_flag farming_flag pop_loss_flag poverty_flag    log_pop
##  6      :593      0:2662       0:2580        0:2754       Min.   :2.358
##  1      :432      1: 443       1: 525        1: 351       1st Qu.:4.050
##  7      :425                                              Median :4.413
##  9      :407                                              Mean   :4.467
##  2      :375                                              3rd Qu.:4.830
##  3      :352                                              Max.   :7.004
##  (Other):521
##  log_battery_electric_vehicles log_hybrid_vehicles log_plugin_hybrid_vehicles
##  Min.   :0.000                 Min.   :0.000       Min.   :0.000
##  1st Qu.:1.380                 1st Qu.:2.185       1st Qu.:1.079
##  Median :1.875                 Median :2.588       Median :1.568
##  Mean   :1.864                 Mean   :2.610       Mean   :1.529
##  3rd Qu.:2.322                 3rd Qu.:3.043       3rd Qu.:2.000
##  Max.   :5.170                 Max.   :5.576       Max.   :5.031
##
##  log_internal_combustion_vehicles
```

```
##  Min.   :1.996
##  1st Qu.:3.975
##  Median :4.331
##  Mean   :4.374
##  3rd Qu.:4.730
##  Max.   :6.810
##
```

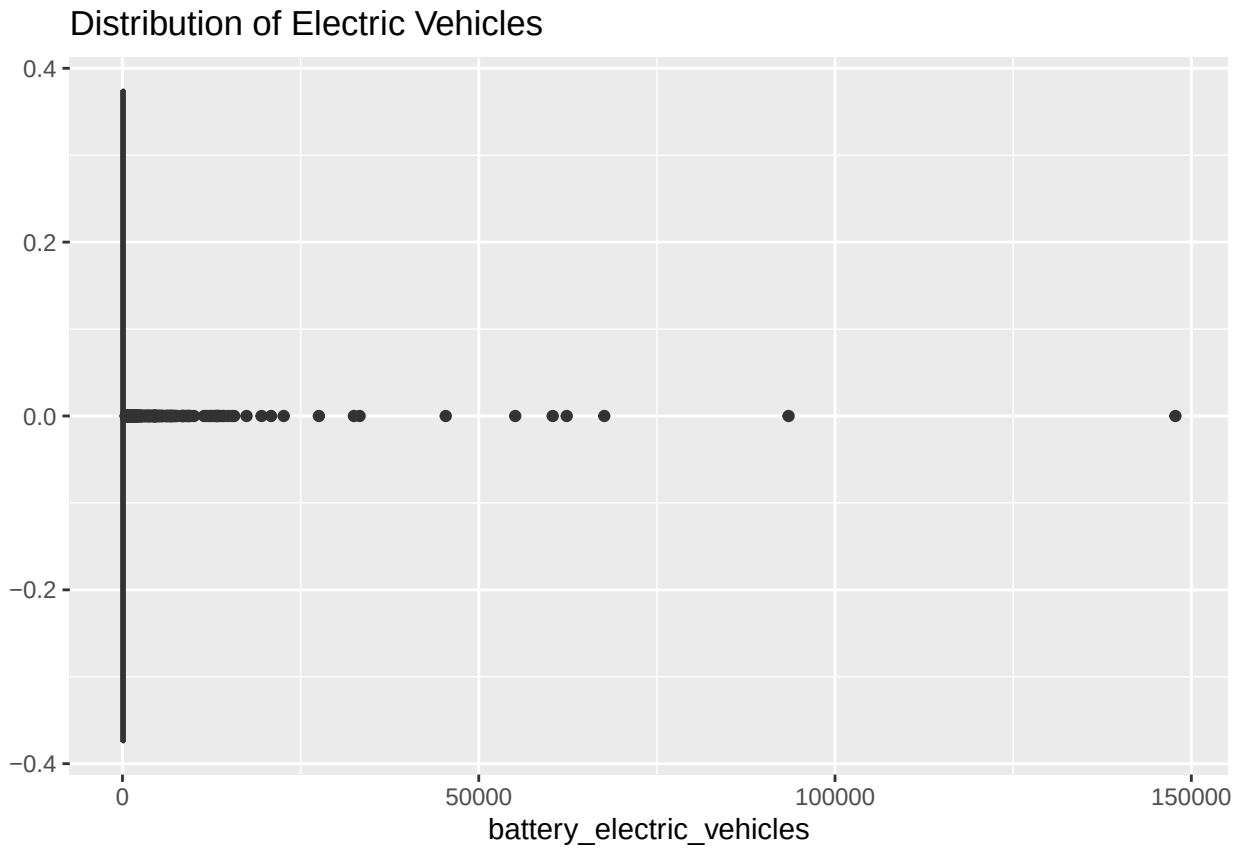
Exploratory Data Analysis - Univariate Distributions

```
# Population Distribution - Original
ggplot(energy2, aes(x = county_pop)) + geom_histogram(aes(y = ..count.. / sum(..count..)), color = 'red
```



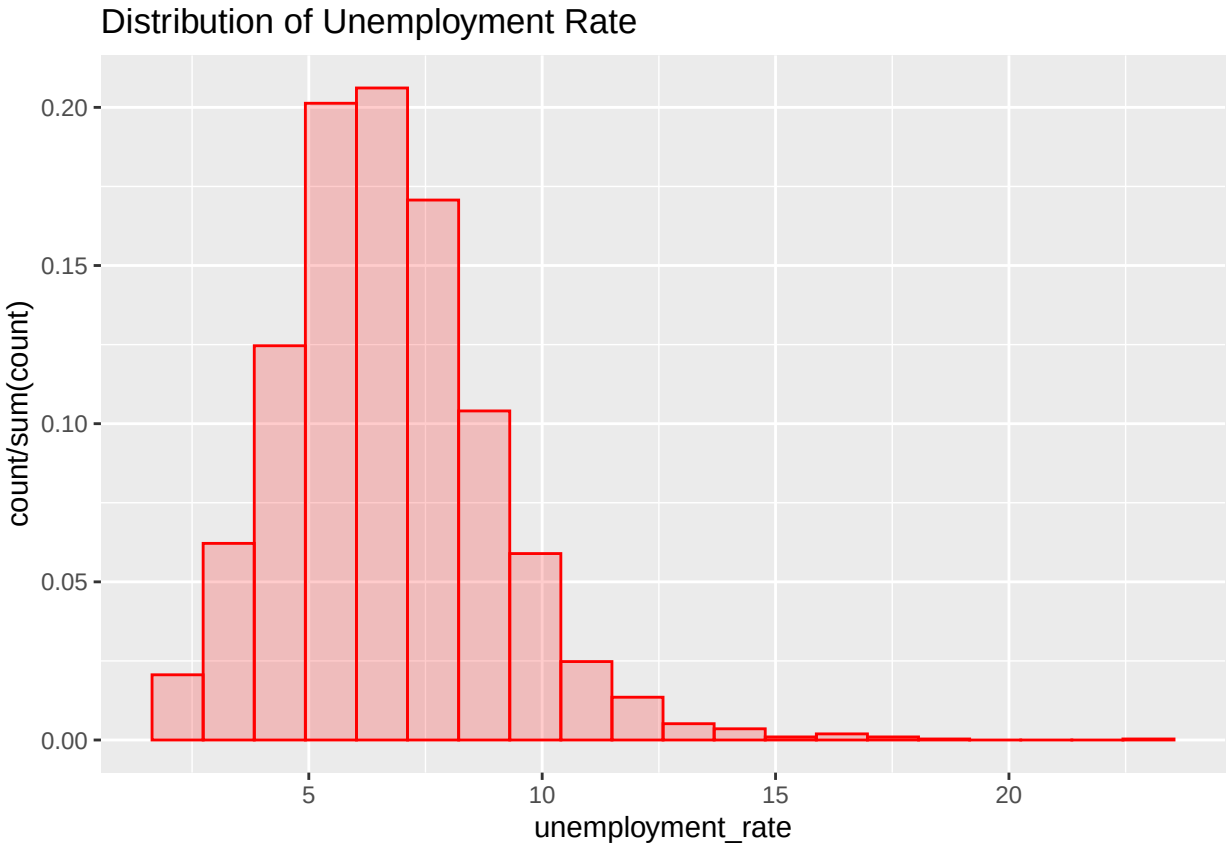

```
# Incredibly right skewed, as expected for population. Log transform was a good idea.
```

```
# Battery Electric Vehicles - Original
ggplot(energy2, aes(x = battery_electric_vehicles)) + geom_boxplot() + ggtitle("Distribution of Electri
```




```
# Lots of outliers, transformation was a good idea. From numerical summaries we can see that other vehi
```

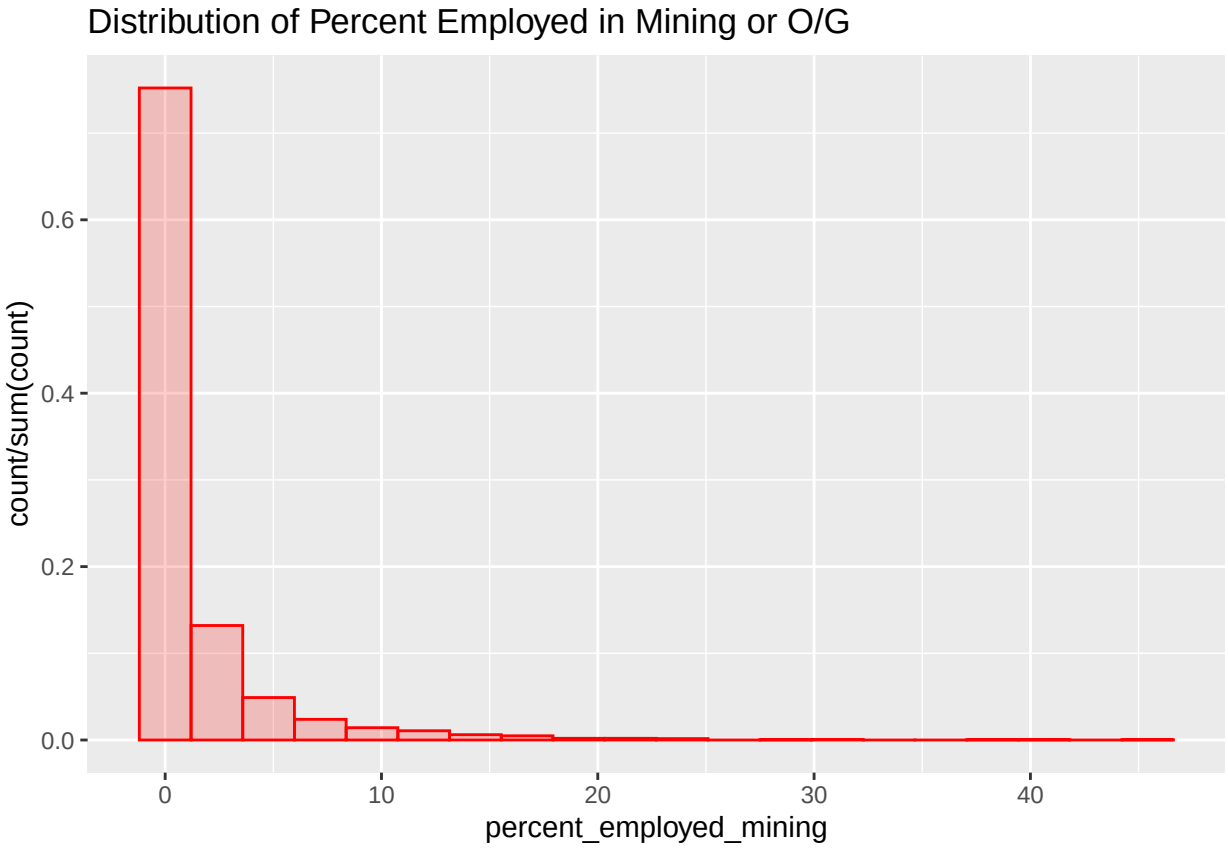
```
# Unemployment Rate - 2020 distribution
ggplot(energy2, aes(x = unemployment_rate)) + geom_histogram(aes(y = ..count.. / sum(..count..)), color
```



75% of the counties have unemployment rate less than 8% (from the numerical summary). Right-skewed, some counties really struggle with unemployment. That is to be expected, if for example, they are over-reliant on one specific industry. It might make sense to convert unemployment rate to a a percentage term.

```
# Transform unemployment to a percent
energy2$pct_unemployed = energy2$unemployment_rate/100
```

```
# Percent Employed Mining or Oil and Gas
ggplot(energy2, aes(x = percent_employed_mining)) + geom_histogram(aes(y = ..count.. / sum(..count..)),
```



Majority of counties >75% have 0-2% employed in mining.

```
# Transform percent mining to a percent
energy2$pct_mining_oil_gas = energy2$percent_employed_mining/100
```

Rural-urban continuum code variable should track closely with the population variable.

```
# Summarize percent of counties with population loss
energy2 %>%
  group_by(pop_loss_flag) %>%
  dplyr::summarise(cnt = n()) %>%
  dplyr::mutate(pct = round(cnt / sum(cnt),2))
```

```
## # A tibble: 2 x 3
##   pop_loss_flag   cnt   pct
##   <fct>         <int> <dbl>
## 1 0              2580  0.83
## 2 1               525  0.17
```

```
# Summarize percent of counties that are farming dependent
energy2 %>%
  group_by(farming_flag) %>%
  dplyr::summarise(cnt = n()) %>%
  dplyr::mutate(pct = round(cnt / sum(cnt),2))
```

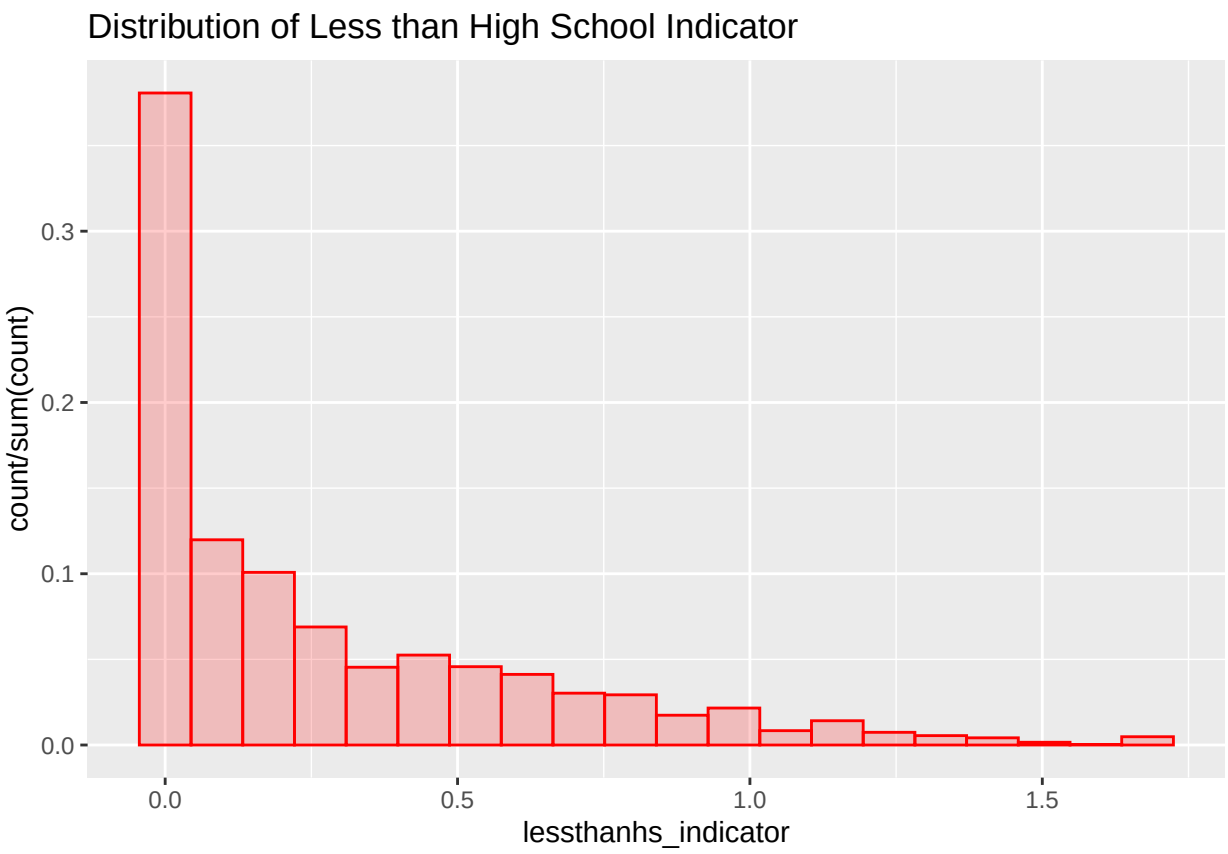
```
## # A tibble: 2 x 3
##   farming_flag   cnt   pct
##   <fct>        <int> <dbl>
## 1 0             2662  0.86
## 2 1              443  0.14
```

```r
# Summarize percent of counties that have persistent poverty
energy2 %>%
  group_by(poverty_flag) %>%
  dplyr::summarise(cnt = n()) %>%
  dplyr::mutate(pct = round(cnt / sum(cnt),2))
```

```
## # A tibble: 2 x 3
##   poverty_flag   cnt   pct
##   <fct>        <int> <dbl>
## 1 0             2754  0.89
## 2 1              351  0.11
```

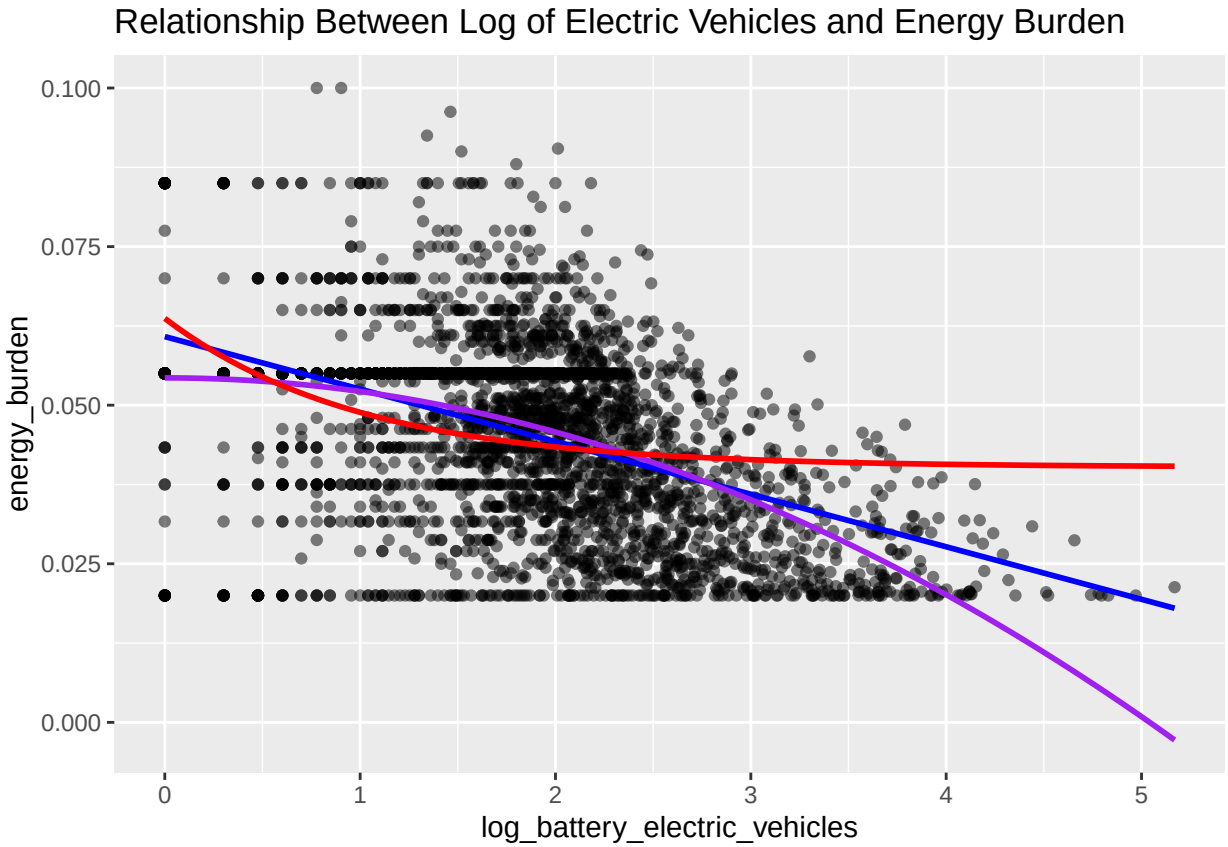
Regarding minority indicator, low income indicator and less than high school indicator variables: these represent proportion of census blocks that are in the 60 - 80th and 80-100th percentile minority, less than hs or low income. Due to the nature of these variables, we expect the data to be highly right skewed.

```r
# Example - less than HS indicator
ggplot(energy2, aes(x = lessthanhs_indicator)) + geom_histogram(aes(y = ..count.. / sum(..count..)), co
```



Exploratory Data Analysis: Bi-variate Relationships

```r
# Explore bivariate relationship between electric vehicle counts and energy_burden
ggplot(energy2, aes(x = log_battery_electric_vehicles, y = energy_burden)) + geom_point(alpha = 0.5) + g
  geom_smooth(method = "lm", formula =  y ~ I(x^2), se=FALSE, color="purple") +
  geom_smooth(method = 'lm', formula = y ~ I(exp(-x)), se=FALSE, color = "red")+
  ggtitle("Relationship Between Log of Electric Vehicles and Energy Burden")
```



Relationship is negative. A quadratic or exponential do not appear to be a good fit. Linear with a negative slope is better.

```
# Check linear relationship strength with log electric vehicles
lm_electric = lm(energy_burden ~ log_battery_electric_vehicles, data = energy2)
summary_lm_electric = summary(lm_electric)
summary_lm_electric$adj.r.squared
```

```
## [1] 0.2103065
```

21% of energy burden is explained by the linear model with log of electric vehicle counts.

```
# Check linear relationship with log hybrid vehicles
lm_hybrid = lm(energy_burden~ log_hybrid_vehicles, data = energy2)
summary_lm_hybrid = summary(lm_hybrid)
summary_lm_hybrid$adj.r.squared
```

```
## [1] 0.2401187
```

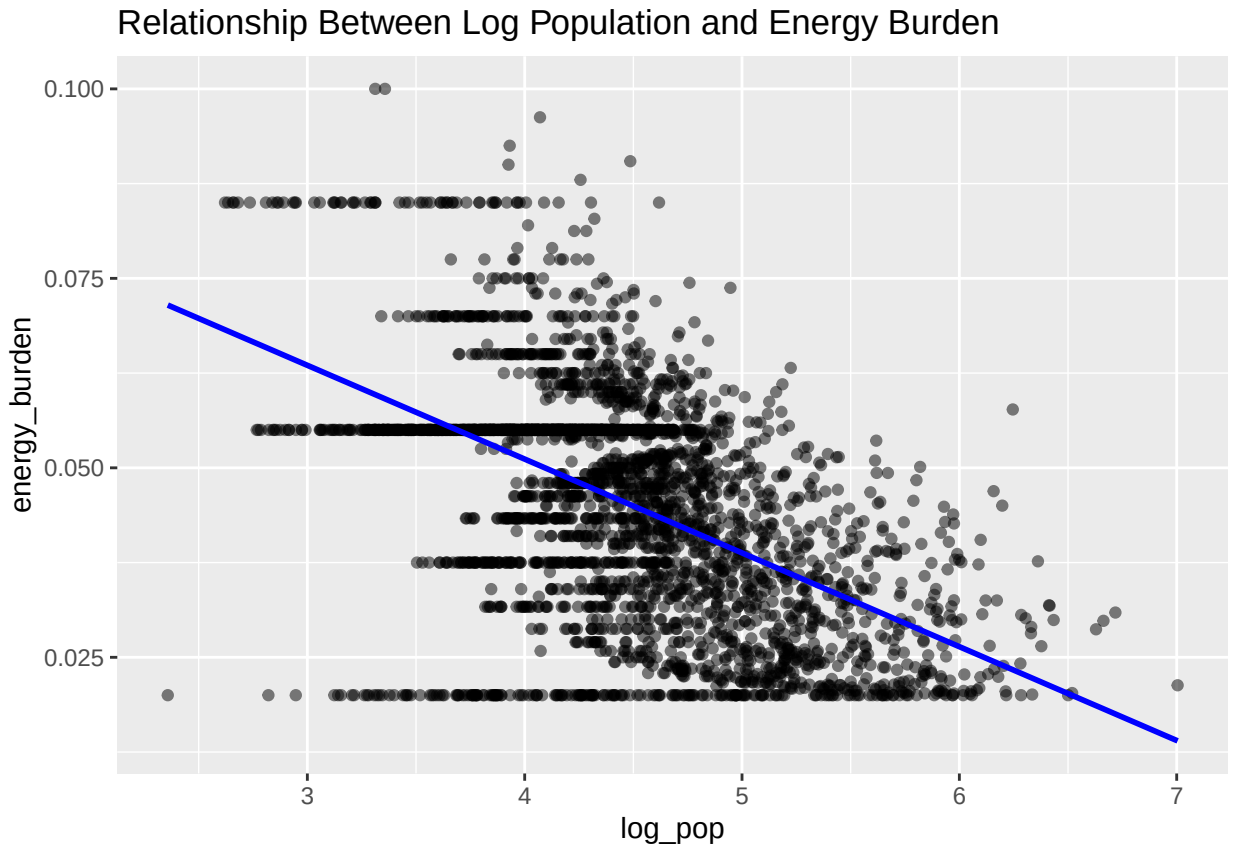
24% of energy burden is explained by the linear model with log hybrid vehicle counts. Hybrid vehicles have more of an influence on energy burden than electric vehicles.

```
# Check linear relationship with internal combustion vehicles
lm_ic = lm(energy_burden~ log_internal_combustion_vehicles, data = energy2)
summary_lm_ic = summary(lm_ic)
summary_lm_ic$adj.r.squared
```

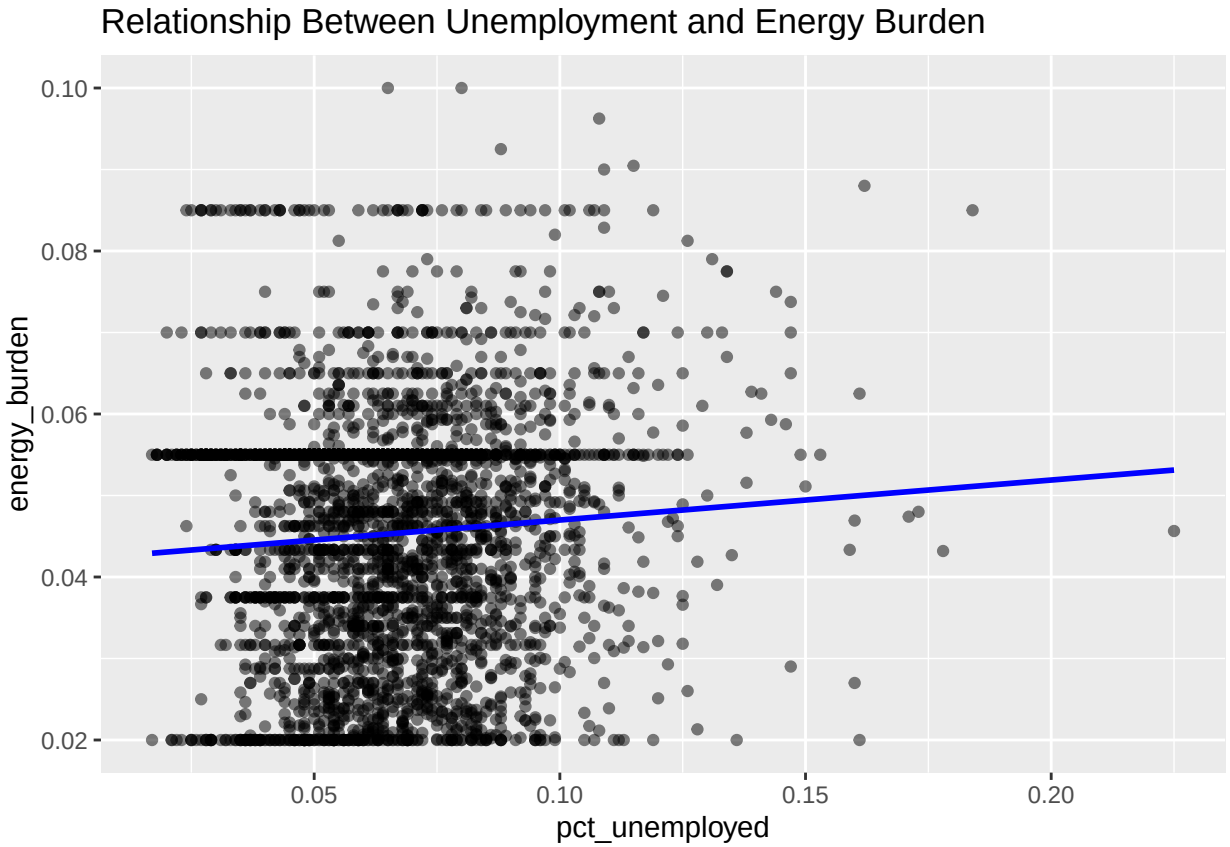
```
## [1] 0.2761065
```

27% of energy burden is explained by the linear model with log internal combustion counts. This is the best explanatory variable among all vehicles.

```
# Explore bivariate relationship between log of population and energy_burden variable
ggplot(energy2, aes(x = log_pop, y = energy_burden)) + geom_point(alpha = 0.5) + geom_smooth(method = "
```

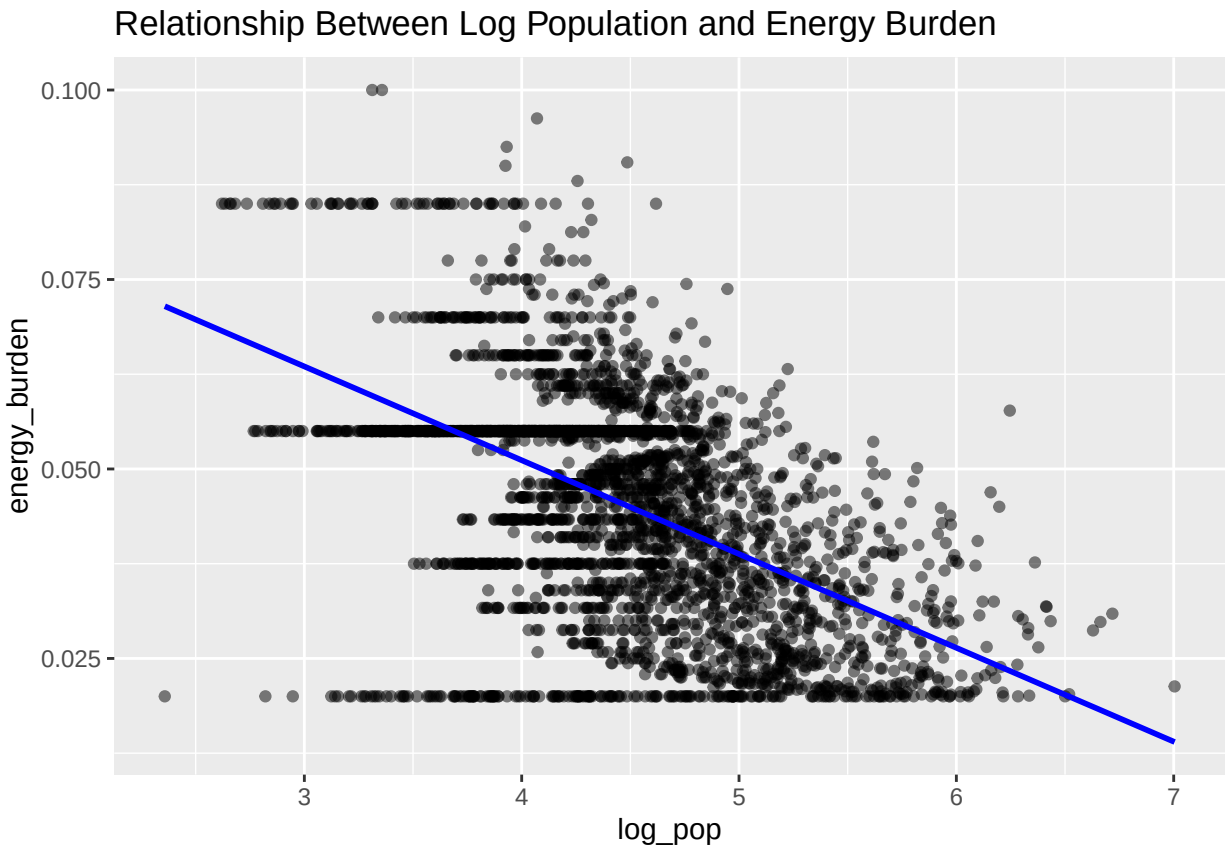



```
# Explore bivariate relationship between unemployment rate and energy_burden variable
ggplot(energy2, aes(x = pct_unemployed, y = energy_burden)) + geom_point(alpha = 0.5) + geom_smooth(meth
```



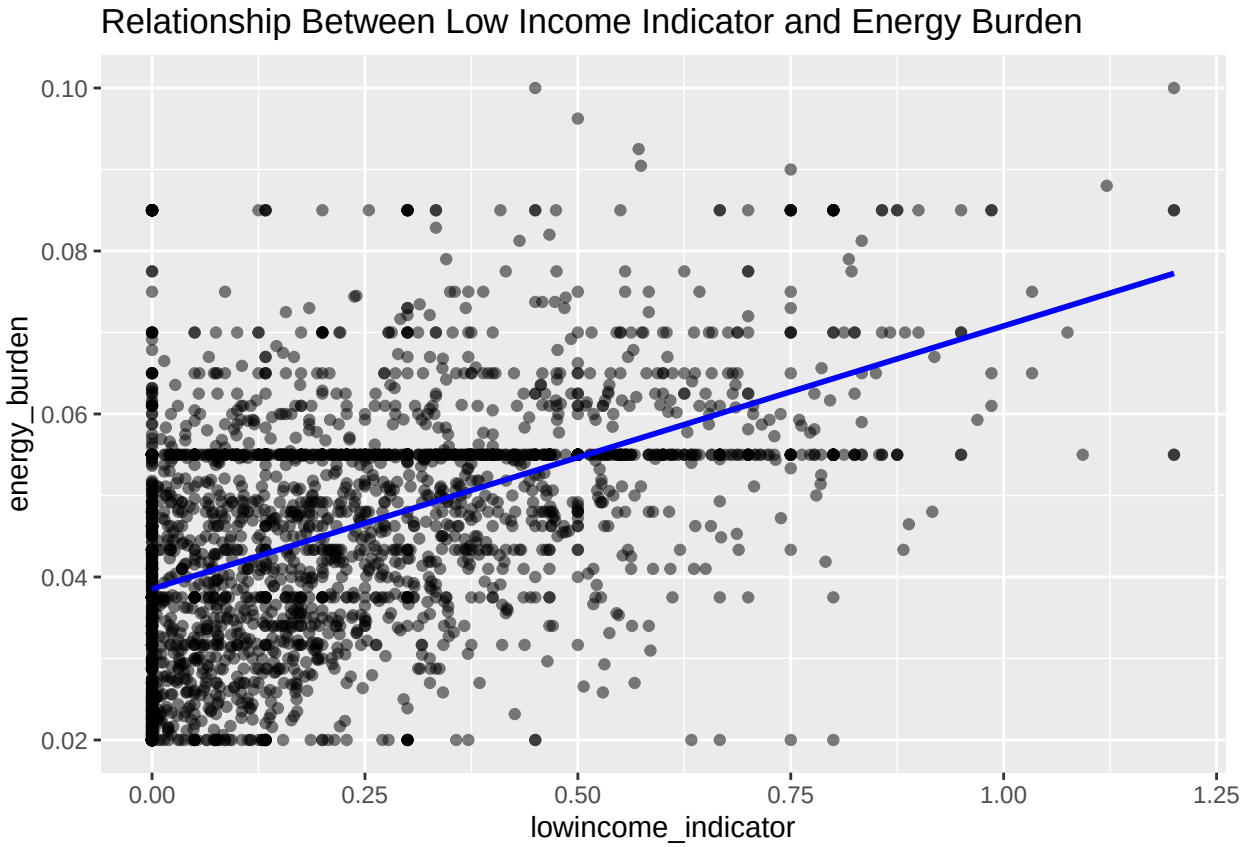
The linear relationship is not strong. Also the line of best fit is pulled by outliers on the right (high unemployment). Otherwise, the slope would be steeper.

```
# Explore bivariate relationship between population and energy_burden variable
ggplot(energy2, aes(x = log_pop, y = energy_burden)) + geom_point(alpha = 0.5) + geom_smooth(method = "
```



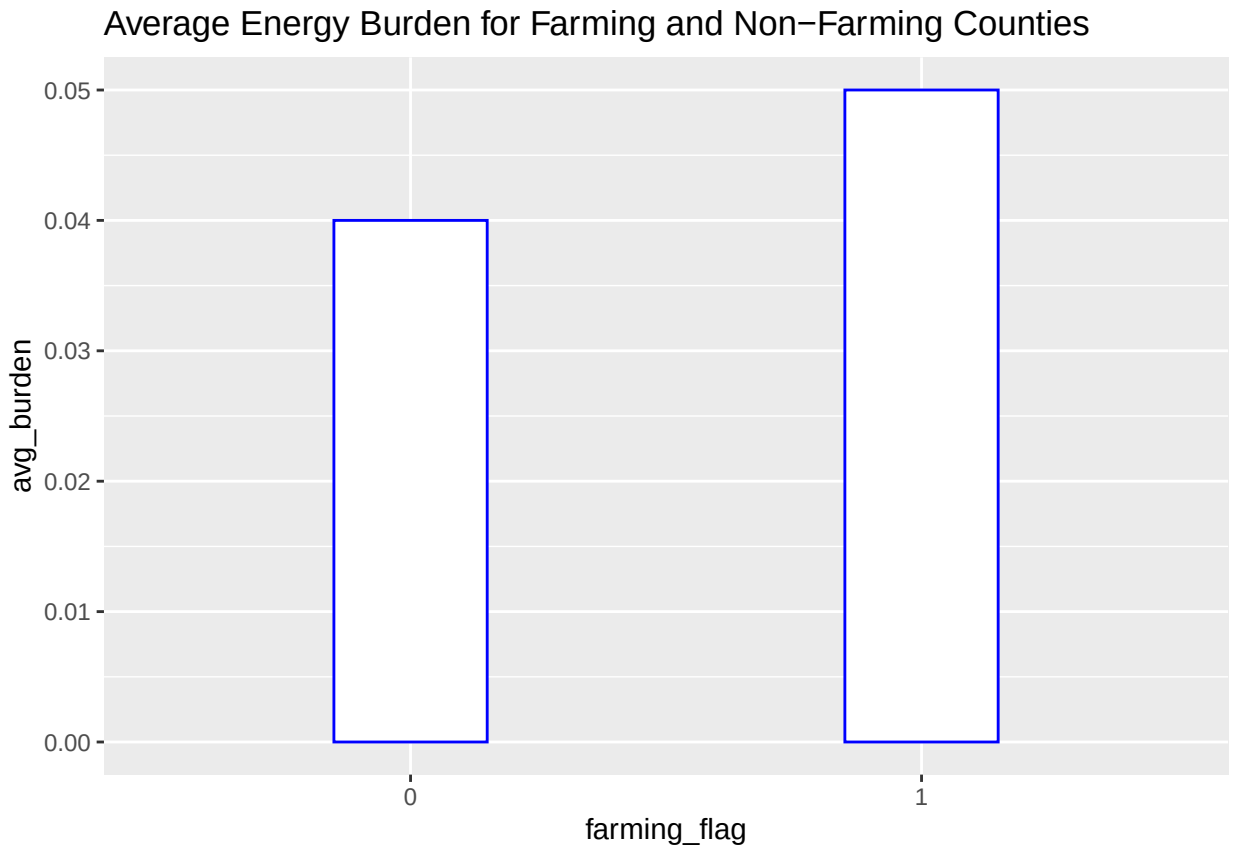
There are lots of instances, where energy burden doesn't vary based on population change. Overall, there is somewhat of a linear (negative) relationship.

```
# Explore bivariate relationship between low income indicator and energy_burden variable
ggplot(energy2, aes(x = lowincome_indicator, y = energy_burden)) + geom_point(alpha = 0.5) + geom_smoothl
```

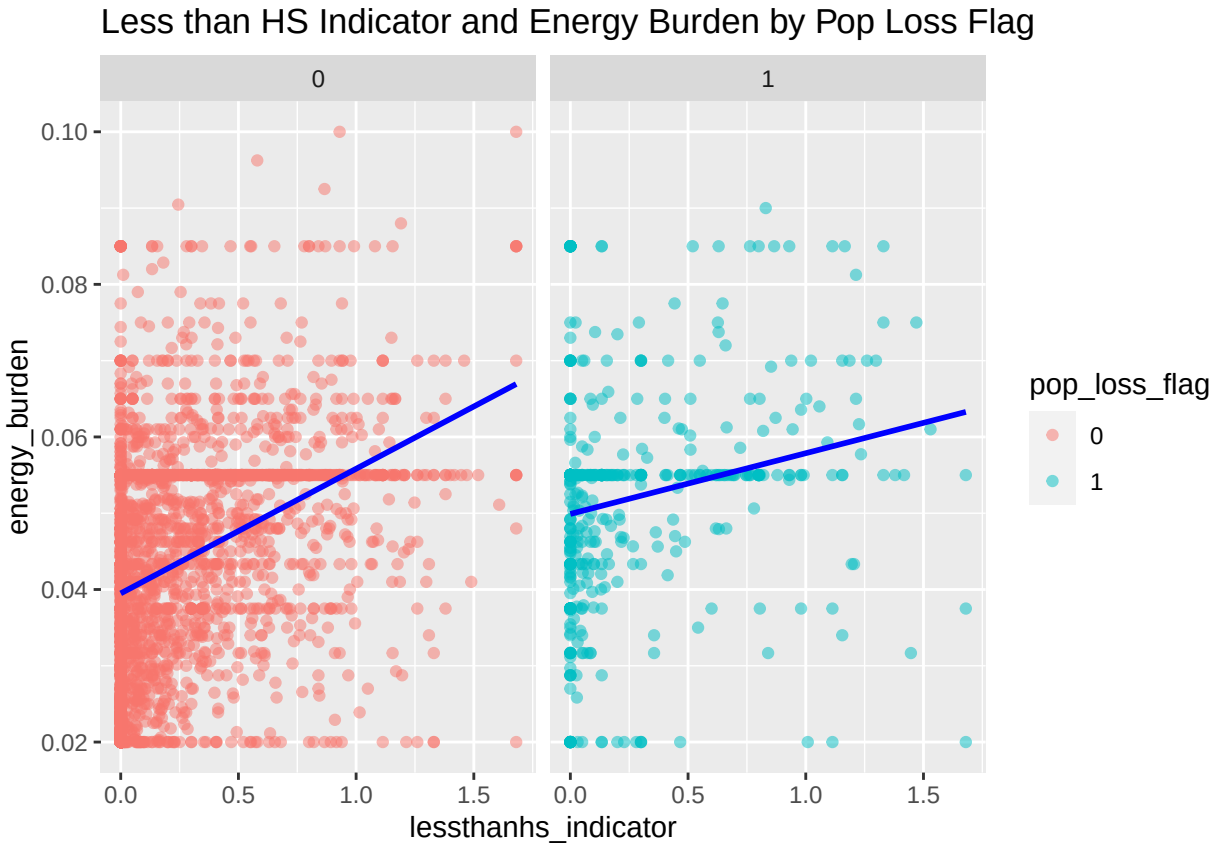


The lower the income, the higher the energy burden, as expected.

```
# Explore bivariate relationship between farming flag and energy burden
energy2 %>%
  dplyr::group_by(farming_flag) %>%
  dplyr::summarise(avg_burden = round(mean(energy_burden),2)) %>%
  ggplot(aes(x = farming_flag, y = avg_burden))  + geom_col(width = 0.3, color = "blue", fill = 'white'
```

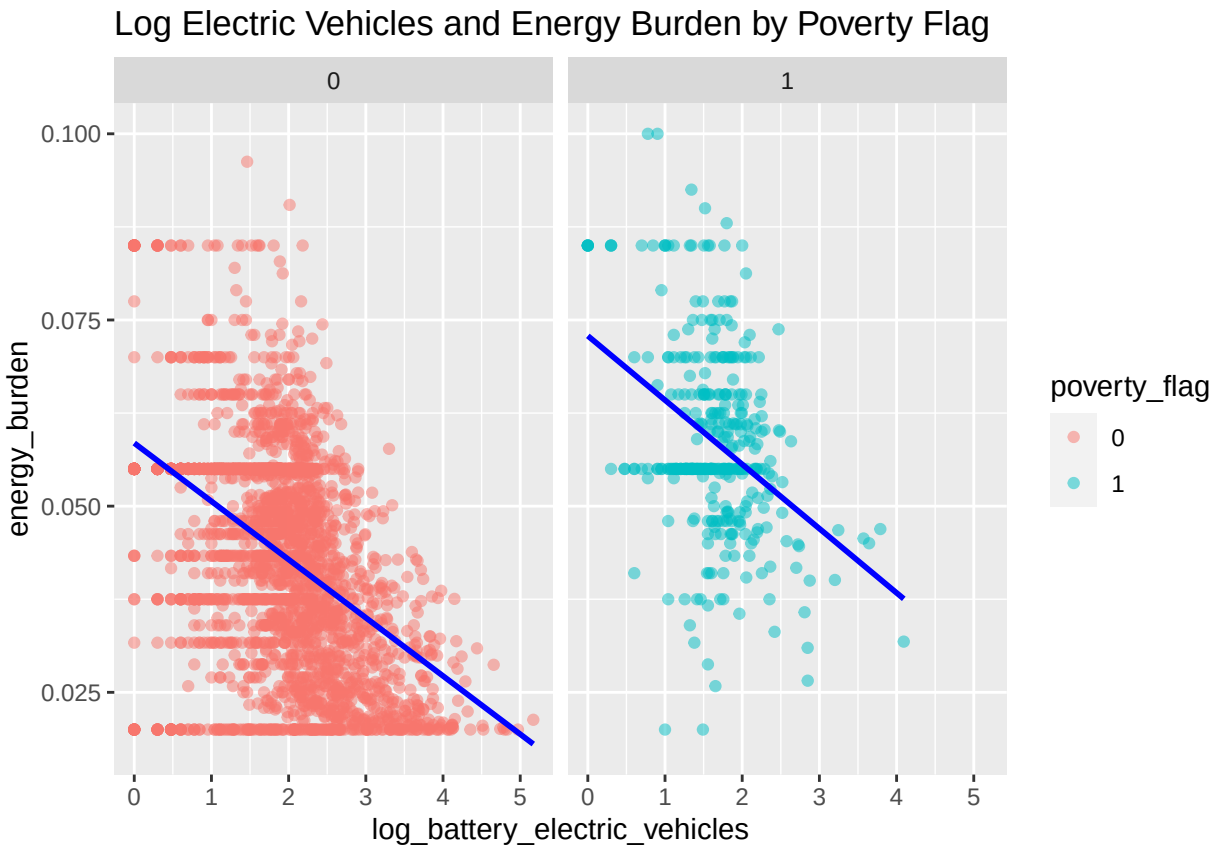



```
# Check for possible interactions
# Population loss flag and less than hs indicator

ggplot(energy2, aes(x = lessthanhs_indicator, y = energy_burden, col = pop_loss_flag)) + geom_point(alp
  ggtitle("Less than HS Indicator and Energy Burden by Pop Loss Flag")
```



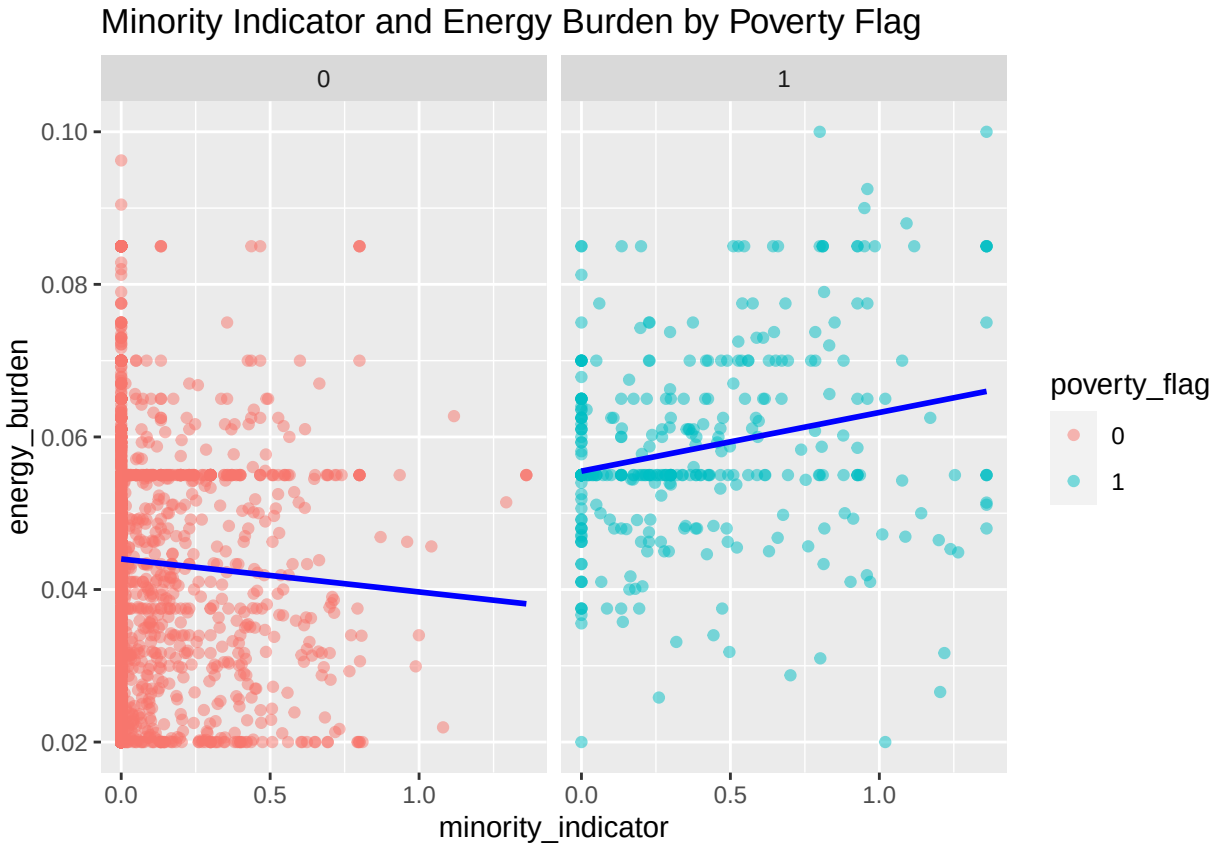
Different slopes - interaction is possible here.

```
# Check for possible interactions
# Poverty flag and electric vehicles

ggplot(energy2, aes(x = log_battery_electric_vehicles, y = energy_burden, col = poverty_flag)) + geom_p
  ggtitle("Log Electric Vehicles and Energy Burden by Poverty Flag")
```



These slopes are probably different, but they don't look significantly different.

```
# Check for possible interactions
# Poverty flag and minority indicator

ggplot(energy2, aes(x = minority_indicator, y = energy_burden, col = poverty_flag)) + geom_point(alpha =
  ggtitle("Minority Indicator and Energy Burden by Poverty Flag")
```



These slopes are significantly different (they also show that there isn't a great linear relationship between minority indicator and energy_burden). This is saying that if we have a persistent poverty county, then a higher minority population would result in a higher average energy burden.

```
# Check for possible interactions
# Farming Flag and Unemployment Rate

ggplot(energy2, aes(x = pct_unemployed, y = energy_burden, col = farming_flag)) + geom_point(alpha = 0.5
  ggtitle("Unemployment Rate and Energy Burden by Farming Flag")
```

These slopes are different, but pulled by outliers. Also the relationship between unemployment rate and energy burden is pretty weak.

Modeling

```
# Create the initial model
mod0 = lm(formula = energy_burden ~ state + log_pop + log_battery_electric_vehicles + log_hybrid_vehicle
            log_internal_combustion_vehicles + pct_unemployed + pct_mining_oil_gas + rural_urban_flag +
            pop_loss_flag + poverty_flag + minority_indicator + lowincome_indicator + lessthanhs_indicat
            poverty_flag * minority_indicator + pop_loss_flag * lessthanhs_indicator + cancer_indicator

summary(mod0)
```

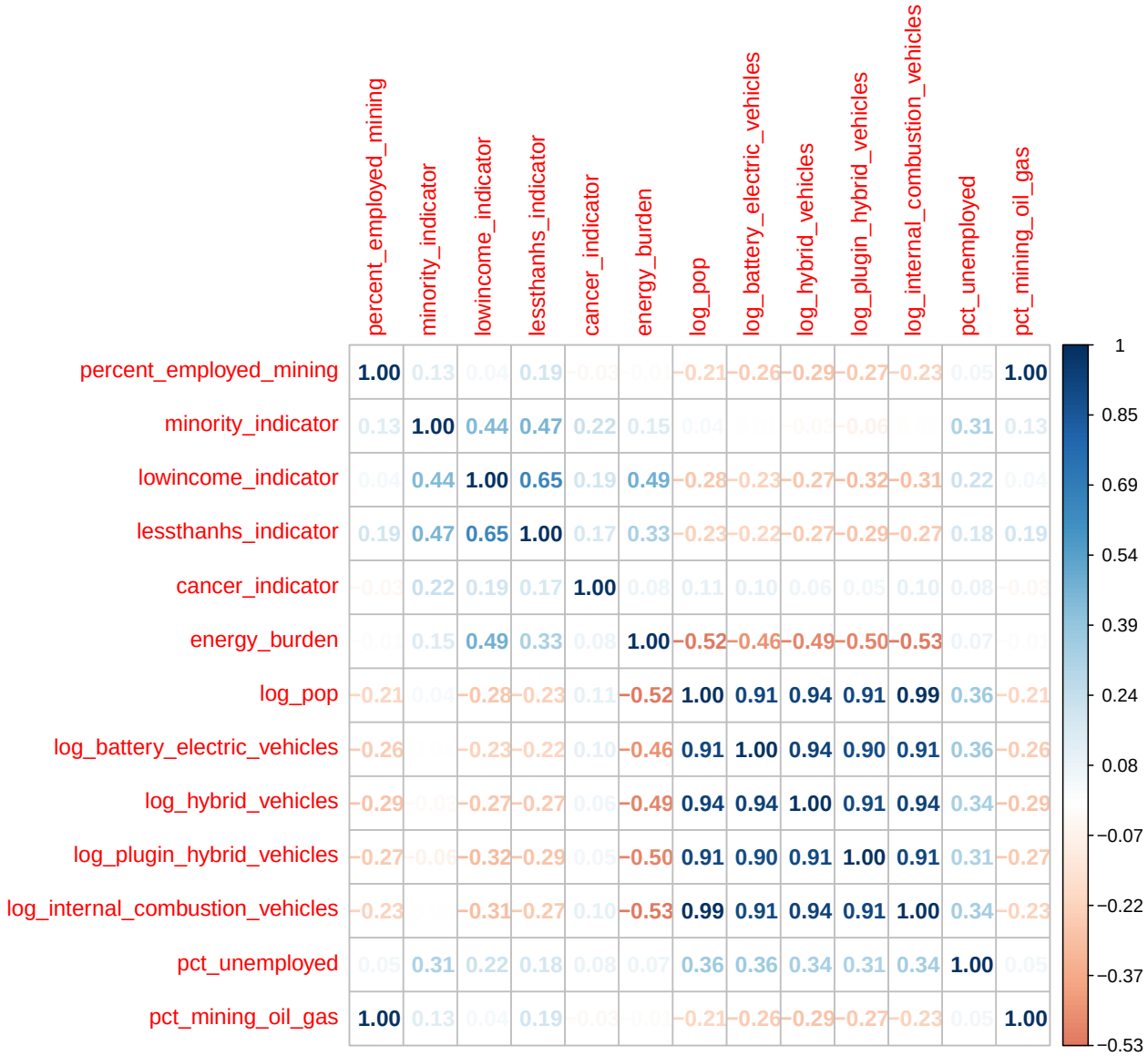
```
##
## Call:
## lm(formula = energy_burden ~ state + log_pop + log_battery_electric_vehicles +
##     log_hybrid_vehicles + log_plugin_hybrid_vehicles + log_internal_combustion_vehicles +
##     pct_unemployed + pct_mining_oil_gas + rural_urban_flag +
##     farming_flag + pop_loss_flag + poverty_flag + minority_indicator +
##     lowincome_indicator + lessthanhs_indicator + cancer_indicator +
##     poverty_flag * minority_indicator + pop_loss_flag * lessthanhs_indicator +
##     cancer_indicator, data = energy2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.043152 -0.006007  0.000213  0.006027  0.038565
##
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.406e-02  4.096e-03  20.522  < 2e-16 ***
## stateAR                        -7.702e-03  1.810e-03  -4.256 2.15e-05 ***
## stateAZ                        -7.165e-03  3.008e-03  -2.382 0.017257 *
## stateCA                        -1.225e-02  2.130e-03  -5.752 9.69e-09 ***
## stateCO                        -7.845e-03  2.004e-03  -3.914 9.29e-05 ***
## stateCT                         3.847e-03  3.902e-03   0.986 0.324158
## stateDC                        -2.179e-03  1.026e-02  -0.212 0.831771
## stateDE                        -4.166e-03  6.016e-03  -0.693 0.488623
## stateFL                        -6.241e-03  1.911e-03  -3.266 0.001102 **
## stateGA                        -4.777e-03  1.509e-03  -3.165 0.001566 **
## stateIA                        -4.754e-03  1.837e-03  -2.588 0.009696 **
## stateID                        -1.097e-02  2.137e-03  -5.134 3.02e-07 ***
## stateIL                        -7.721e-03  1.853e-03  -4.167 3.17e-05 ***
## stateIN                        -2.566e-03  1.839e-03  -1.395 0.163072
## stateKS                        -6.963e-03  1.821e-03  -3.822 0.000135 ***
## stateKY                        -1.040e-02  1.753e-03  -5.929 3.39e-09 ***
## stateLA                        -3.334e-03  1.841e-03  -1.811 0.070255 .
## stateMA                         3.656e-03  3.159e-03   1.157 0.247260
## stateMD                        -5.108e-03  2.591e-03  -1.972 0.048736 *
## stateME                         1.316e-02  2.953e-03   4.458 8.56e-06 ***
## stateMI                         4.159e-04  1.952e-03   0.213 0.831278
## stateMN                        -7.973e-03  1.868e-03  -4.268 2.03e-05 ***
## stateMO                        -3.630e-03  1.755e-03  -2.068 0.038727 *
## stateMS                        -7.424e-03  1.730e-03  -4.291 1.84e-05 ***
## stateMT                        -1.882e-03  2.036e-03  -0.924 0.355333
## stateNC                        -5.948e-03  1.793e-03  -3.317 0.000919 ***
## stateND                        -8.108e-03  2.116e-03  -3.831 0.000130 ***
## stateNE                        -5.991e-03  1.885e-03  -3.178 0.001498 **
## stateNH                         6.817e-03  3.542e-03   1.925 0.054360 .
## stateNJ                        -9.219e-03  2.768e-03  -3.330 0.000878 ***
## stateNM                        -1.160e-02  2.409e-03  -4.814 1.55e-06 ***
## stateNV                        -1.580e-02  2.973e-03  -5.313 1.16e-07 ***
## stateNY                         1.803e-04  2.045e-03   0.088 0.929751
## stateOH                        -2.085e-03  1.870e-03  -1.115 0.265036
## stateOK                        -5.107e-03  1.912e-03  -2.672 0.007590 **
## stateOR                        -1.903e-02  2.314e-03  -8.225 2.87e-16 ***
## statePA                         1.356e-03  1.993e-03   0.680 0.496254
## stateRI                        -2.331e-03  4.800e-03  -0.486 0.627353
## stateSC                        -1.043e-03  2.007e-03  -0.520 0.603334
## stateSD                        -1.257e-02  1.987e-03  -6.325 2.90e-10 ***
## stateTN                        -1.102e-02  1.787e-03  -6.164 8.03e-10 ***
## stateTX                        -9.094e-03  1.646e-03  -5.523 3.61e-08 ***
## stateUT                        -1.745e-02  2.430e-03  -7.181 8.69e-13 ***
## stateVA                        -8.212e-03  1.741e-03  -4.718 2.49e-06 ***
## stateVT                         7.026e-03  3.132e-03   2.243 0.024937 *
## stateWA                        -1.876e-02  2.296e-03  -8.170 4.49e-16 ***
## stateWI                        -9.125e-03  1.933e-03  -4.721 2.45e-06 ***
## stateWV                        -1.469e-02  2.084e-03  -7.051 2.19e-12 ***
## stateWY                        -1.175e-02  2.665e-03  -4.408 1.08e-05 ***
## log_pop                        -2.082e-02  3.647e-03  -5.709 1.24e-08 ***
## log_battery_electric_vehicles   1.629e-03  7.192e-04   2.265 0.023613 *
## log_hybrid_vehicles            -6.445e-05  9.265e-04  -0.070 0.944550
```

```
## log_plugin_hybrid_vehicles          -5.085e-04  6.348e-04  -0.801 0.423191
## log_internal_combustion_vehicles     9.380e-03  3.620e-03   2.591 0.009613 **
## pct_unemployed                       1.437e-01  1.277e-02  11.250  < 2e-16 ***
## pct_mining_oil_gas                  -5.002e-02  7.124e-03  -7.021 2.72e-12 ***
## rural_urban_flag2                    1.458e-03  7.455e-04   1.956 0.050531 .
## rural_urban_flag3                    2.972e-03  7.821e-04   3.801 0.000147 ***
## rural_urban_flag4                    4.408e-03  9.045e-04   4.873 1.16e-06 ***
## rural_urban_flag5                    2.288e-03  1.241e-03   1.843 0.065412 .
## rural_urban_flag6                    6.302e-03  7.846e-04   8.031 1.37e-15 ***
## rural_urban_flag7                    6.336e-03  8.791e-04   7.207 7.17e-13 ***
## rural_urban_flag8                    8.099e-03  1.058e-03   7.652 2.64e-14 ***
## rural_urban_flag9                    8.924e-03  1.041e-03   8.570  < 2e-16 ***
## farming_flag1                       -5.658e-04  6.988e-04  -0.810 0.418178
## pop_loss_flag1                       1.120e-03  7.181e-04   1.560 0.118925
## poverty_flag1                       -2.430e-04  9.429e-04  -0.258 0.796668
## minority_indicator                  -6.431e-04  1.511e-03  -0.426 0.670433
## lowincome_indicator                  1.715e-02  1.332e-03  12.874  < 2e-16 ***
## lessthanhs_indicator                 1.471e-03  8.694e-04   1.692 0.090729 .
## cancer_indicator                     1.547e-03  1.168e-03   1.324 0.185691
## poverty_flag1:minority_indicator     4.742e-03  1.947e-03   2.436 0.014921 *
## pop_loss_flag1:lessthanhs_indicator -3.311e-03  1.393e-03  -2.376 0.017549 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01006 on 3032 degrees of freedom
## Multiple R-squared:  0.5709, Adjusted R-squared:  0.5607
## F-statistic: 56.03 on 72 and 3032 DF,  p-value: < 2.2e-16
```

R^2 of 57% is pretty good. Let's look at the correlation table and VIF results to check for collinearity.

```
# Correlation for all numeric variables

corrplot(cor(energy2[,c(15, 17, 18, 19, 20, 21, 26, 27, 28, 29, 30, 31, 32)]), is.corr = FALSE, method =
```



Population and vehicle counts have significant correlation - in the 90% range. So do vehicle counts among each other (i.e., electric vehicles correlate with internal combustion vehicles etc.)

```
# Check Variance Inflation Factors for Collinearity
car::vif(mod0, type = 'predictor')
```

```
## GVIFs computed for predictors
```

```
##                                        GVIF Df GVIF^(1/(2*Df))
## state                             62.818220 48        1.044071
## log_pop                          167.454999  1       12.940440
## log_battery_electric_vehicles     11.217079  1        3.349191
## log_hybrid_vehicles               15.113114  1        3.887559
## log_plugin_hybrid_vehicles         9.214537  1        3.035546
## log_internal_combustion_vehicles 154.362151  1       12.424257
## pct_unemployed                     2.474088  1        1.572923
## pct_mining_oil_gas                 1.734324  1        1.316937
## rural_urban_flag                   4.477982  8        1.098228
```

```
## farming_flag                        1.831582  1        1.353360
## pop_loss_flag                       4.255413  3        1.272986
## poverty_flag                        3.758186  3        1.246895
## minority_indicator                  3.758186  3        1.246895
## lowincome_indicator                 2.846081  1        1.687033
## lessthanhs_indicator                4.255413  3        1.272986
## cancer_indicator                    2.015682  1        1.419747
##                                         Interacts With
## state                                                --
## log_pop                                              --
## log_battery_electric_vehicles                        --
## log_hybrid_vehicles                                  --
## log_plugin_hybrid_vehicles                           --
## log_internal_combustion_vehicles                     --
## pct_unemployed                                       --
## pct_mining_oil_gas                                   --
## rural_urban_flag                                     --
## farming_flag                                         --
## pop_loss_flag                       lessthanhs_indicator
## poverty_flag                          minority_indicator
## minority_indicator                          poverty_flag
## lowincome_indicator                                  --
## lessthanhs_indicator                      pop_loss_flag
## cancer_indicator                                     --
##
## state                               log_pop, log_battery_electric_vehicles, log_hybrid_vehicles, log_plu
## log_pop                               state, log_battery_electric_vehicles, log_hybrid_vehicles, log_plu
## log_battery_electric_vehicles                                state, log_pop, log_hybrid_vehicles, log_plu
## log_hybrid_vehicles                             state, log_pop, log_battery_electric_vehicles, log_plu
## log_plugin_hybrid_vehicles                             state, log_pop, log_battery_electric_vehicles,
## log_internal_combustion_vehicles                             state, log_pop, log_battery_electric_vehi
## pct_unemployed                                state, log_pop, log_battery_electric_vehicles, log_hybrid_v
## pct_mining_oil_gas                                 state, log_pop, log_battery_electric_vehicles, log_hybi
## rural_urban_flag                                 state, log_pop, log_battery_electric_vehicles, log_hybrid
## farming_flag                                   state, log_pop, log_battery_electric_vehicles, log_hybrid_vel
## pop_loss_flag                                                  state, log_pop, log_battery_electric_v
## poverty_flag                                                 state, log_pop, log_battery_electric_vehi
## minority_indicator                                           state, log_pop, log_battery_electric_vehi
## lowincome_indicator                              state, log_pop, log_battery_electric_vehicles, log_hyb
## lessthanhs_indicator                                           state, log_pop, log_battery_electric_v
## cancer_indicator                                 state, log_pop, log_battery_electric_vehicles, log_hybrid
```

Clearly, log_pop (population) is highly correlated with other variables with GVIF of 167. Internal combustion vehicle variable has GVIF of 154. It's expected that regular gasoline vehicles would correlate highly with population. Let's take out log_pop first and checkt the results.

```
mod1=update(mod0,.~.-log_pop)
summary(mod1)
```

```
##
## Call:
## lm(formula = energy_burden ~ state + log_battery_electric_vehicles +
##     log_hybrid_vehicles + log_plugin_hybrid_vehicles + log_internal_combustion_vehicles +
##     pct_unemployed + pct_mining_oil_gas + rural_urban_flag +
##     farming_flag + pop_loss_flag + poverty_flag + minority_indicator +
```

```
##     lowincome_indicator + lessthanhs_indicator + cancer_indicator +
##     poverty_flag:minority_indicator + pop_loss_flag:lessthanhs_indicator,
##     data = energy2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.043975 -0.006156  0.000247  0.006156  0.039799
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         0.0807060  0.0040748  19.806  < 2e-16 ***
## stateAR                            -0.0085464  0.0018130  -4.714 2.54e-06 ***
## stateAZ                            -0.0076438  0.0030220  -2.529 0.011477 *
## stateCA                            -0.0130672  0.0021366  -6.116 1.08e-09 ***
## stateCO                            -0.0078387  0.0020148  -3.891 0.000102 ***
## stateCT                             0.0030271  0.0039193   0.772 0.439964
## stateDC                            -0.0075910  0.0102648  -0.740 0.459653
## stateDE                            -0.0050560  0.0060449  -0.836 0.402990
## stateFL                            -0.0080630  0.0018936  -4.258 2.13e-05 ***
## stateGA                            -0.0054475  0.0015126  -3.601 0.000322 ***
## stateIA                            -0.0050251  0.0018458  -2.722 0.006518 **
## stateID                            -0.0114494  0.0021465  -5.334 1.03e-07 ***
## stateIL                            -0.0089213  0.0018505  -4.821 1.50e-06 ***
## stateIN                            -0.0034012  0.0018427  -1.846 0.065019 .
## stateKS                            -0.0074458  0.0018290  -4.071 4.80e-05 ***
## stateKY                            -0.0111933  0.0017570  -6.371 2.17e-10 ***
## stateLA                            -0.0049788  0.0018276  -2.724 0.006483 **
## stateMA                             0.0022181  0.0031654   0.701 0.483519
## stateMD                            -0.0064835  0.0025931  -2.500 0.012460 *
## stateME                             0.0125247  0.0029659   4.223 2.48e-05 ***
## stateMI                            -0.0006884  0.0019522  -0.353 0.724409
## stateMN                            -0.0084335  0.0018760  -4.495 7.20e-06 ***
## stateMO                            -0.0046445  0.0017554  -2.646 0.008190 **
## stateMS                            -0.0077023  0.0017386  -4.430 9.75e-06 ***
## stateMT                            -0.0017345  0.0020467  -0.847 0.396816
## stateNC                            -0.0063312  0.0018010  -3.515 0.000446 ***
## stateND                            -0.0085409  0.0021261  -4.017 6.03e-05 ***
## stateNE                            -0.0063945  0.0018936  -3.377 0.000742 ***
## stateNH                             0.0061912  0.0035584   1.740 0.081984 .
## stateNJ                            -0.0109414  0.0027661  -3.956 7.81e-05 ***
## stateNM                            -0.0116966  0.0024210  -4.831 1.42e-06 ***
## stateNV                            -0.0162162  0.0029875  -5.428 6.15e-08 ***
## stateNY                            -0.0019905  0.0020195  -0.986 0.324384
## stateOH                            -0.0029135  0.0018744  -1.554 0.120197
## stateOK                            -0.0067159  0.0019005  -3.534 0.000416 ***
## stateOR                            -0.0195213  0.0023246  -8.398  < 2e-16 ***
## statePA                             0.0002828  0.0019944   0.142 0.887253
## stateRI                            -0.0035273  0.0048206  -0.732 0.464393
## stateSC                            -0.0011347  0.0020173  -0.562 0.573822
## stateSD                            -0.0130393  0.0019951  -6.535 7.41e-11 ***
## stateTN                            -0.0113757  0.0017956  -6.335 2.72e-10 ***
## stateTX                            -0.0108907  0.0016244  -6.704 2.40e-11 ***
## stateUT                            -0.0190390  0.0024264  -7.847 5.87e-15 ***
## stateVA                            -0.0084553  0.0017491  -4.834 1.40e-06 ***
```

```
## stateVT                                0.0065503  0.0031469   2.082 0.037470 *
## stateWA                               -0.0193957  0.0023052  -8.414  < 2e-16 ***
## stateWI                               -0.0094551  0.0019419  -4.869 1.18e-06 ***
## stateWV                               -0.0160565  0.0020808  -7.716 1.61e-14 ***
## stateWY                               -0.0127988  0.0026727  -4.789 1.76e-06 ***
## log_battery_electric_vehicles          0.0009958  0.0007143   1.394 0.163394
## log_hybrid_vehicles                   -0.0007139  0.0009243  -0.772 0.439947
## log_plugin_hybrid_vehicles            -0.0005879  0.0006380  -0.921 0.356909
## log_internal_combustion_vehicles      -0.0102387  0.0011446  -8.945  < 2e-16 ***
## pct_unemployed                         0.1477830  0.0128148  11.532  < 2e-16 ***
## pct_mining_oil_gas                    -0.0523149  0.0071500  -7.317 3.24e-13 ***
## rural_urban_flag2                      0.0014062  0.0007493   1.877 0.060653 .
## rural_urban_flag3                      0.0028187  0.0007857   3.588 0.000339 ***
## rural_urban_flag4                      0.0043805  0.0009092   4.818 1.52e-06 ***
## rural_urban_flag5                      0.0020938  0.0012471   1.679 0.093273 .
## rural_urban_flag6                      0.0065439  0.0007875   8.309  < 2e-16 ***
## rural_urban_flag7                      0.0064332  0.0008835   7.282 4.18e-13 ***
## rural_urban_flag8                      0.0083694  0.0010628   7.875 4.72e-15 ***
## rural_urban_flag9                      0.0090759  0.0010463   8.674  < 2e-16 ***
## farming_flag1                         -0.0003376  0.0007013  -0.481 0.630278
## pop_loss_flag1                         0.0014736  0.0007191   2.049 0.040522 *
## poverty_flag1                         -0.0003998  0.0009474  -0.422 0.673042
## minority_indicator                    -0.0024483  0.0014852  -1.648 0.099361 .
## lowincome_indicator                    0.0164134  0.0013326  12.317  < 2e-16 ***
## lessthanhs_indicator                   0.0012419  0.0008730   1.423 0.154967
## cancer_indicator                       0.0014818  0.0011745   1.262 0.207164
## poverty_flag1:minority_indicator       0.0048327  0.0019570   2.469 0.013585 *
## pop_loss_flag1:lessthanhs_indicator -0.0031963  0.0014005  -2.282 0.022544 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01011 on 3033 degrees of freedom
## Multiple R-squared:  0.5663, Adjusted R-squared:  0.5561
## F-statistic: 55.78 on 71 and 3033 DF,  p-value: < 2.2e-16
```

R^2 stayed relatively the same. Battery electric vehicles are no longer statistically significant.

```
# Re-check VIFs
vif(mod1, type = 'predictor')
```

```
## GVIFs computed for predictors
```

```
##                                       GVIF Df GVIF^(1/(2*Df))
## state                           49.412796 48        1.041464
## log_battery_electric_vehicles   10.950634  1        3.309174
## log_hybrid_vehicles             14.885298  1        3.858147
## log_plugin_hybrid_vehicles       9.210122  1        3.034818
## log_internal_combustion_vehicles 15.272575  1        3.908014
## pct_unemployed                   2.466143  1        1.570396
## pct_mining_oil_gas               1.728792  1        1.314835
## rural_urban_flag                 4.418965  8        1.097318
## farming_flag                     1.825588  1        1.351143
## pop_loss_flag                    4.184884  3        1.269445
## poverty_flag                     3.486470  3        1.231396
## minority_indicator               3.486470  3        1.231396
## lowincome_indicator              2.819574  1        1.679159
```

```
## lessthanhs_indicator              4.184884  3        1.269445
## cancer_indicator                  2.015491  1        1.419680
##                                       Interacts With
## state                                         --
## log_battery_electric_vehicles                 --
## log_hybrid_vehicles                           --
## log_plugin_hybrid_vehicles                    --
## log_internal_combustion_vehicles              --
## pct_unemployed                                --
## pct_mining_oil_gas                            --
## rural_urban_flag                              --
## farming_flag                                  --
## pop_loss_flag                    lessthanhs_indicator
## poverty_flag                      minority_indicator
## minority_indicator                      poverty_flag
## lowincome_indicator                           --
## lessthanhs_indicator                    pop_loss_flag
## cancer_indicator                              --
##
## state                            log_battery_electric_vehicles, log_hybrid_vehicles, log_plugin_hybri
## log_battery_electric_vehicles                              state, log_hybrid_vehicles, log_plugin_hybri
## log_hybrid_vehicles                              state, log_battery_electric_vehicles, log_plugin_hybri
## log_plugin_hybrid_vehicles                              state, log_battery_electric_vehicles, log_hybri
## log_internal_combustion_vehicles                              state, log_battery_electric_vehicles, log
## pct_unemployed                                 state, log_battery_electric_vehicles, log_hybrid_vehicles,
## pct_mining_oil_gas                                state, log_battery_electric_vehicles, log_hybrid_vehicl
## rural_urban_flag                                 state, log_battery_electric_vehicles, log_hybrid_vehicles
## farming_flag                                   state, log_battery_electric_vehicles, log_hybrid_vehicles, l
## pop_loss_flag                                                    state, log_battery_electric_vehicles,
## poverty_flag                                                    state, log_battery_electric_vehicles, log
## minority_indicator                                              state, log_battery_electric_vehicles, log
## lowincome_indicator                               state, log_battery_electric_vehicles, log_hybrid_vehi
## lessthanhs_indicator                                             state, log_battery_electric_vehicles,
## cancer_indicator                                 state, log_battery_electric_vehicles, log_hybrid_vehicles
```

GVIF for internal combustion vehicles has gone down after we took out population. Overall, vehicle counts are still highly correlated. It seems reasonable to take-out internal combustion engine vehicles as the predictor because of collinearity.

```
# Remove internal combustion vehicles
mod2 = update(mod1,.~.-log_internal_combustion_vehicles)
summary(mod2)
```

```
##
## Call:
## lm(formula = energy_burden ~ state + log_battery_electric_vehicles +
##     log_hybrid_vehicles + log_plugin_hybrid_vehicles + pct_unemployed +
##     pct_mining_oil_gas + rural_urban_flag + farming_flag + pop_loss_flag +
##     poverty_flag + minority_indicator + lowincome_indicator +
##     lessthanhs_indicator + cancer_indicator + poverty_flag:minority_indicator +
##     pop_loss_flag:lessthanhs_indicator, data = energy2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.042880 -0.006282  0.000048  0.006272  0.039225
```

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.975e-02  2.180e-03  22.823  < 2e-16 ***
## stateAR                        -7.145e-03  1.830e-03  -3.905 9.61e-05 ***
## stateAZ                        -6.637e-03  3.059e-03  -2.170 0.030110 *
## stateCA                        -8.229e-03  2.094e-03  -3.930 8.68e-05 ***
## stateCO                        -4.578e-03  2.007e-03  -2.281 0.022639 *
## stateCT                         4.761e-03  3.965e-03   1.201 0.229954
## stateDC                        -1.674e-03  1.038e-02  -0.161 0.871817
## stateDE                        -3.689e-03  6.121e-03  -0.603 0.546774
## stateFL                        -6.870e-03  1.913e-03  -3.590 0.000335 ***
## stateGA                        -3.949e-03  1.523e-03  -2.593 0.009557 **
## stateIA                        -2.972e-03  1.855e-03  -1.602 0.109284
## stateID                        -9.743e-03  2.166e-03  -4.499 7.09e-06 ***
## stateIL                        -6.545e-03  1.855e-03  -3.528 0.000424 ***
## stateIN                        -1.185e-03  1.850e-03  -0.641 0.521774
## stateKS                        -5.865e-03  1.844e-03  -3.180 0.001485 **
## stateKY                        -9.348e-03  1.767e-03  -5.289 1.32e-07 ***
## stateLA                        -4.057e-03  1.848e-03  -2.195 0.028225 *
## stateMA                         4.966e-03  3.191e-03   1.556 0.119779
## stateMD                        -3.594e-03  2.606e-03  -1.379 0.167992
## stateME                         1.521e-02  2.989e-03   5.088 3.85e-07 ***
## stateMI                         9.286e-04  1.969e-03   0.472 0.637238
## stateMN                        -6.975e-03  1.893e-03  -3.684 0.000233 ***
## stateMO                        -2.856e-03  1.766e-03  -1.617 0.106062
## stateMS                        -7.000e-03  1.759e-03  -3.979 7.08e-05 ***
## stateMT                        -6.207e-04  2.069e-03  -0.300 0.764221
## stateNC                        -4.387e-03  1.811e-03  -2.422 0.015485 *
## stateND                        -8.258e-03  2.153e-03  -3.835 0.000128 ***
## stateNE                        -4.890e-03  1.911e-03  -2.559 0.010534 *
## stateNH                         8.290e-03  3.597e-03   2.305 0.021243 *
## stateNJ                        -9.395e-03  2.796e-03  -3.360 0.000790 ***
## stateNM                        -9.251e-03  2.437e-03  -3.797 0.000150 ***
## stateNV                        -1.300e-02  3.004e-03  -4.328 1.55e-05 ***
## stateNY                         7.344e-04  2.022e-03   0.363 0.716494
## stateOH                        -1.691e-03  1.894e-03  -0.893 0.371878
## stateOK                        -5.256e-03  1.918e-03  -2.740 0.006178 **
## stateOR                        -1.524e-02  2.304e-03  -6.613 4.43e-11 ***
## statePA                         1.187e-03  2.018e-03   0.588 0.556378
## stateRI                        -4.983e-04  4.871e-03  -0.102 0.918519
## stateSC                        -1.028e-04  2.040e-03  -0.050 0.959834
## stateSD                        -1.208e-02  2.018e-03  -5.986 2.40e-09 ***
## stateTN                        -1.007e-02  1.813e-03  -5.554 3.03e-08 ***
## stateTX                        -9.503e-03  1.638e-03  -5.802 7.24e-09 ***
## stateUT                        -1.658e-02  2.442e-03  -6.788 1.36e-11 ***
## stateVA                        -5.699e-03  1.744e-03  -3.268 0.001096 **
## stateVT                         1.066e-02  3.153e-03   3.382 0.000729 ***
## stateWA                        -1.489e-02  2.279e-03  -6.535 7.45e-11 ***
## stateWI                        -7.629e-03  1.956e-03  -3.900 9.82e-05 ***
## stateWV                        -1.363e-02  2.090e-03  -6.523 8.05e-11 ***
## stateWY                        -1.102e-02  2.700e-03  -4.081 4.60e-05 ***
## log_battery_electric_vehicles   9.749e-05  7.164e-04   0.136 0.891761
## log_hybrid_vehicles            -4.764e-03  8.162e-04  -5.837 5.88e-09 ***
```

```
## log_plugin_hybrid_vehicles          -2.898e-03  5.909e-04  -4.904 9.89e-07 ***
## pct_unemployed                       1.348e-01  1.290e-02  10.449  < 2e-16 ***
## pct_mining_oil_gas                  -5.782e-02  7.216e-03  -8.014 1.57e-15 ***
## rural_urban_flag2                    1.627e-03  7.586e-04   2.144 0.032100 *
## rural_urban_flag3                    3.469e-03  7.924e-04   4.378 1.24e-05 ***
## rural_urban_flag4                    4.851e-03  9.195e-04   5.276 1.41e-07 ***
## rural_urban_flag5                    2.476e-03  1.263e-03   1.961 0.049948 *
## rural_urban_flag6                    7.886e-03  7.831e-04  10.070  < 2e-16 ***
## rural_urban_flag7                    7.816e-03  8.811e-04   8.871  < 2e-16 ***
## rural_urban_flag8                    1.040e-02  1.052e-03   9.888  < 2e-16 ***
## rural_urban_flag9                    1.115e-02  1.034e-03  10.786  < 2e-16 ***
## farming_flag1                       -2.648e-04  7.103e-04  -0.373 0.709350
## pop_loss_flag1                       7.778e-04  7.241e-04   1.074 0.282842
## poverty_flag1                        5.587e-06  9.585e-04   0.006 0.995350
## minority_indicator                  -3.918e-03  1.495e-03  -2.620 0.008825 **
## lowincome_indicator                  1.758e-02  1.343e-03  13.089  < 2e-16 ***
## lessthanhs_indicator                 1.146e-03  8.842e-04   1.296 0.195072
## cancer_indicator                     1.428e-03  1.190e-03   1.201 0.229997
## poverty_flag1:minority_indicator     5.729e-03  1.980e-03   2.894 0.003830 **
## pop_loss_flag1:lessthanhs_indicator -1.664e-03  1.408e-03  -1.182 0.237396
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01025 on 3034 degrees of freedom
## Multiple R-squared:  0.5549, Adjusted R-squared:  0.5446
## F-statistic: 54.02 on 70 and 3034 DF,  p-value: < 2.2e-16
```

R^2 has gone down slightly to 55.49%. Plug-in hybrid vehicles and regular hybrid vehicles are now statistically significant. Based on an earlier correlation matrix, among all vehicles - internal combustion cars are the most correlated with energy burden, followed by hybrid and electric. However, all of the vehicle counts are highly correlated with population and log_pop is a great predictor of energy burden: 27% of variability in energy burden is explained by log of population. Therefore, it makes sense to remove all vehicle predictors and just keep the population variable. That should also take care of collinearity.

```
# Log population -> Energy burden.
lm_pop = lm(energy_burden ~ log_pop, data = energy2)
summary_lm_pop = summary(lm_pop)
summary_lm_pop$adj.r.squared
```

```
## [1] 0.2726938
```

```
# Create another model without any vehicle predictors, but with log of population.
mod3 = update(mod0,.~.-log_battery_electric_vehicles - log_hybrid_vehicles - log_internal_combustion_veh
summary(mod3)
```

```
##
## Call:
## lm(formula = energy_burden ~ state + log_pop + pct_unemployed +
##     pct_mining_oil_gas + rural_urban_flag + farming_flag + pop_loss_flag +
##     poverty_flag + minority_indicator + lowincome_indicator +
##     lessthanhs_indicator + cancer_indicator + poverty_flag:minority_indicator +
##     pop_loss_flag:lessthanhs_indicator, data = energy2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.043417 -0.006037  0.000222  0.006174  0.039255
```

```
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.0819967  0.0030658  26.745  < 2e-16 ***
## stateAR                    -0.0081416  0.0017992  -4.525 6.27e-06 ***
## stateAZ                    -0.0071447  0.0030036  -2.379 0.017435 *
## stateCA                    -0.0122852  0.0020235  -6.071 1.43e-09 ***
## stateCO                    -0.0076220  0.0019756  -3.858 0.000117 ***
## stateCT                     0.0036509  0.0038916   0.938 0.348236
## stateDC                    -0.0045363  0.0101985  -0.445 0.656494
## stateDE                    -0.0044997  0.0060170  -0.748 0.454619
## stateFL                    -0.0070100  0.0018794  -3.730 0.000195 ***
## stateGA                    -0.0050179  0.0014993  -3.347 0.000827 ***
## stateIA                    -0.0049756  0.0018243  -2.727 0.006420 **
## stateID                    -0.0114867  0.0021241  -5.408 6.88e-08 ***
## stateIL                    -0.0083165  0.0018259  -4.555 5.45e-06 ***
## stateIN                    -0.0029990  0.0018202  -1.648 0.099532 .
## stateKS                    -0.0073825  0.0018106  -4.077 4.67e-05 ***
## stateKY                    -0.0107345  0.0017376  -6.178 7.36e-10 ***
## stateLA                    -0.0040944  0.0018173  -2.253 0.024327 *
## stateMA                     0.0032119  0.0031286   1.027 0.304685
## stateMD                    -0.0055641  0.0025585  -2.175 0.029723 *
## stateME                     0.0128772  0.0029300   4.395 1.15e-05 ***
## stateMI                    -0.0001441  0.0019357  -0.074 0.940650
## stateMN                    -0.0082601  0.0018620  -4.436 9.49e-06 ***
## stateMO                    -0.0040997  0.0017360  -2.362 0.018262 *
## stateMS                    -0.0076497  0.0017290  -4.424 1.00e-05 ***
## stateMT                    -0.0019481  0.0020372  -0.956 0.338999
## stateNC                    -0.0059964  0.0017780  -3.373 0.000754 ***
## stateND                    -0.0084977  0.0021156  -4.017 6.05e-05 ***
## stateNE                    -0.0063589  0.0018804  -3.382 0.000730 ***
## stateNH                     0.0065212  0.0035328   1.846 0.065007 .
## stateNJ                    -0.0098328  0.0027461  -3.581 0.000348 ***
## stateNM                    -0.0116478  0.0023962  -4.861 1.23e-06 ***
## stateNV                    -0.0155361  0.0029452  -5.275 1.42e-07 ***
## stateNY                    -0.0008854  0.0019812  -0.447 0.654987
## stateOH                    -0.0024071  0.0018612  -1.293 0.196017
## stateOK                    -0.0058900  0.0018857  -3.123 0.001804 **
## stateOR                    -0.0190957  0.0022593  -8.452  < 2e-16 ***
## statePA                     0.0009049  0.0019820   0.457 0.648000
## stateRI                    -0.0029744  0.0047888  -0.621 0.534566
## stateSC                    -0.0010081  0.0020054  -0.503 0.615220
## stateSD                    -0.0130279  0.0019830  -6.570 5.91e-11 ***
## stateTN                    -0.0111804  0.0017833  -6.269 4.14e-10 ***
## stateTX                    -0.0100225  0.0016110  -6.221 5.61e-10 ***
## stateUT                    -0.0180421  0.0024015  -7.513 7.57e-14 ***
## stateVA                    -0.0082281  0.0017198  -4.784 1.80e-06 ***
## stateVT                     0.0068828  0.0030912   2.227 0.026048 *
## stateWA                    -0.0187425  0.0022315  -8.399  < 2e-16 ***
## stateWI                    -0.0092964  0.0019224  -4.836 1.39e-06 ***
## stateWV                    -0.0151883  0.0020522  -7.401 1.74e-13 ***
## stateWY                    -0.0121703  0.0026554  -4.583 4.76e-06 ***
## log_pop                    -0.0105947  0.0005133 -20.642  < 2e-16 ***
## pct_unemployed              0.1465215  0.0127252  11.514  < 2e-16 ***
```

```
## pct_mining_oil_gas                  -0.0518325  0.0070142  -7.390 1.89e-13 ***
## rural_urban_flag2                    0.0013167  0.0007441   1.769 0.076926 .
## rural_urban_flag3                    0.0027427  0.0007782   3.525 0.000431 ***
## rural_urban_flag4                    0.0042152  0.0009024   4.671 3.13e-06 ***
## rural_urban_flag5                    0.0019889  0.0012384   1.606 0.108364
## rural_urban_flag6                    0.0062219  0.0007842   7.934 2.97e-15 ***
## rural_urban_flag7                    0.0061286  0.0008763   6.994 3.28e-12 ***
## rural_urban_flag8                    0.0080603  0.0010549   7.641 2.87e-14 ***
## rural_urban_flag9                    0.0087384  0.0010394   8.407  < 2e-16 ***
## farming_flag1                       -0.0004908  0.0006952  -0.706 0.480292
## pop_loss_flag1                       0.0012508  0.0007085   1.765 0.077583 .
## poverty_flag1                       -0.0001839  0.0009396  -0.196 0.844834
## minority_indicator                  -0.0015130  0.0014795  -1.023 0.306562
## lowincome_indicator                  0.0169150  0.0013200  12.814  < 2e-16 ***
## lessthanhs_indicator                 0.0012740  0.0008664   1.470 0.141555
## cancer_indicator                     0.0016095  0.0011692   1.377 0.168733
## poverty_flag1:minority_indicator     0.0047655  0.0019491   2.445 0.014544 *
## pop_loss_flag1:lessthanhs_indicator -0.0030666  0.0013782  -2.225 0.026154 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01008 on 3036 degrees of freedom
## Multiple R-squared:  0.5692, Adjusted R-squared:  0.5596
## F-statistic:    59 on 68 and 3036 DF,  p-value: < 2.2e-16
```

R^2 has improved again to 56.92%. Need to check VIFs

```
# Check VIFs again
vif(mod3, type = 'predictor')
```

```
## GVIFs computed for predictors
```

```
##                            GVIF Df GVIF^(1/(2*Df))       Interacts With
## state                 32.962126 48        1.037081                   --
## log_pop                3.308895  1        1.819037                   --
## pct_unemployed         2.450803  1        1.565504                   --
## pct_mining_oil_gas     1.676760  1        1.294898                   --
## rural_urban_flag       4.278368  8        1.095103                   --
## farming_flag           1.808307  1        1.344733                   --
## pop_loss_flag          4.048427  3        1.262451 lessthanhs_indicator
## poverty_flag           3.426759  3        1.227856   minority_indicator
## minority_indicator     3.426759  3        1.227856         poverty_flag
## lowincome_indicator    2.788414  1        1.669855                   --
## lessthanhs_indicator   4.048427  3        1.262451        pop_loss_flag
## cancer_indicator       2.012945  1        1.418783                   --
##
## state                 log_pop, pct_unemployed, pct_mining_oil_gas, rural_urban_flag, farming_flag, pop
## log_pop                 state, pct_unemployed, pct_mining_oil_gas, rural_urban_flag, farming_flag, pop
## pct_unemployed                state, log_pop, pct_mining_oil_gas, rural_urban_flag, farming_flag, pop
## pct_mining_oil_gas                state, log_pop, pct_unemployed, rural_urban_flag, farming_flag, pop
## rural_urban_flag                state, log_pop, pct_unemployed, pct_mining_oil_gas, farming_flag, pop
## farming_flag                 state, log_pop, pct_unemployed, pct_mining_oil_gas, rural_urban_flag, pop
## pop_loss_flag                                      state, log_pop, pct_unemployed, pct_mining_oil_gas
## poverty_flag                                     state, log_pop, pct_unemployed, pct_mining_oil_gas, i
## minority_indicator                               state, log_pop, pct_unemployed, pct_mining_oil_gas, i
## lowincome_indicator               state, log_pop, pct_unemployed, pct_mining_oil_gas, rural_urban_fl
```

```
## lessthanhs_indicator                                state, log_pop, pct_unemployed, pct_mining_oil_gas
## cancer_indicator               state, log_pop, pct_unemployed, pct_mining_oil_gas, rural_urban_flag
```

VIFs are now under an acceptable level. This model shows a stronger associative relationship with the response and does not have excessive collinearity.

Further Model Selection

We can now remove statistically insignificant predictor variables.

```
summary(mod3)
```

```
##
## Call:
## lm(formula = energy_burden ~ state + log_pop + pct_unemployed +
##     pct_mining_oil_gas + rural_urban_flag + farming_flag + pop_loss_flag +
##     poverty_flag + minority_indicator + lowincome_indicator +
##     lessthanhs_indicator + cancer_indicator + poverty_flag:minority_indicator +
##     pop_loss_flag:lessthanhs_indicator, data = energy2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.043417 -0.006037  0.000222  0.006174  0.039255
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        0.0819967  0.0030658  26.745  < 2e-16 ***
## stateAR                           -0.0081416  0.0017992  -4.525 6.27e-06 ***
## stateAZ                           -0.0071447  0.0030036  -2.379 0.017435 *
## stateCA                           -0.0122852  0.0020235  -6.071 1.43e-09 ***
## stateCO                           -0.0076220  0.0019756  -3.858 0.000117 ***
## stateCT                            0.0036509  0.0038916   0.938 0.348236
## stateDC                           -0.0045363  0.0101985  -0.445 0.656494
## stateDE                           -0.0044997  0.0060170  -0.748 0.454619
## stateFL                           -0.0070100  0.0018794  -3.730 0.000195 ***
## stateGA                           -0.0050179  0.0014993  -3.347 0.000827 ***
## stateIA                           -0.0049756  0.0018243  -2.727 0.006420 **
## stateID                           -0.0114867  0.0021241  -5.408 6.88e-08 ***
## stateIL                           -0.0083165  0.0018259  -4.555 5.45e-06 ***
## stateIN                           -0.0029990  0.0018202  -1.648 0.099532 .
## stateKS                           -0.0073825  0.0018106  -4.077 4.67e-05 ***
## stateKY                           -0.0107345  0.0017376  -6.178 7.36e-10 ***
## stateLA                           -0.0040944  0.0018173  -2.253 0.024327 *
## stateMA                            0.0032119  0.0031286   1.027 0.304685
## stateMD                           -0.0055641  0.0025585  -2.175 0.029723 *
## stateME                            0.0128772  0.0029300   4.395 1.15e-05 ***
## stateMI                           -0.0001441  0.0019357  -0.074 0.940650
## stateMN                           -0.0082601  0.0018620  -4.436 9.49e-06 ***
## stateMO                           -0.0040997  0.0017360  -2.362 0.018262 *
## stateMS                           -0.0076497  0.0017290  -4.424 1.00e-05 ***
## stateMT                           -0.0019481  0.0020372  -0.956 0.338999
## stateNC                           -0.0059964  0.0017780  -3.373 0.000754 ***
## stateND                           -0.0084977  0.0021156  -4.017 6.05e-05 ***
## stateNE                           -0.0063589  0.0018804  -3.382 0.000730 ***
## stateNH                            0.0065212  0.0035328   1.846 0.065007 .
```

```
## stateNJ                                -0.0098328  0.0027461  -3.581 0.000348 ***
## stateNM                                -0.0116478  0.0023962  -4.861 1.23e-06 ***
## stateNV                                -0.0155361  0.0029452  -5.275 1.42e-07 ***
## stateNY                                -0.0008854  0.0019812  -0.447 0.654987
## stateOH                                -0.0024071  0.0018612  -1.293 0.196017
## stateOK                                -0.0058900  0.0018857  -3.123 0.001804 **
## stateOR                                -0.0190957  0.0022593  -8.452  < 2e-16 ***
## statePA                                 0.0009049  0.0019820   0.457 0.648000
## stateRI                                -0.0029744  0.0047888  -0.621 0.534566
## stateSC                                -0.0010081  0.0020054  -0.503 0.615220
## stateSD                                -0.0130279  0.0019830  -6.570 5.91e-11 ***
## stateTN                                -0.0111804  0.0017833  -6.269 4.14e-10 ***
## stateTX                                -0.0100225  0.0016110  -6.221 5.61e-10 ***
## stateUT                                -0.0180421  0.0024015  -7.513 7.57e-14 ***
## stateVA                                -0.0082281  0.0017198  -4.784 1.80e-06 ***
## stateVT                                 0.0068828  0.0030912   2.227 0.026048 *
## stateWA                                -0.0187425  0.0022315  -8.399  < 2e-16 ***
## stateWI                                -0.0092964  0.0019224  -4.836 1.39e-06 ***
## stateWV                                -0.0151883  0.0020522  -7.401 1.74e-13 ***
## stateWY                                -0.0121703  0.0026554  -4.583 4.76e-06 ***
## log_pop                                -0.0105947  0.0005133 -20.642  < 2e-16 ***
## pct_unemployed                          0.1465215  0.0127252  11.514  < 2e-16 ***
## pct_mining_oil_gas                     -0.0518325  0.0070142  -7.390 1.89e-13 ***
## rural_urban_flag2                       0.0013167  0.0007441   1.769 0.076926 .
## rural_urban_flag3                       0.0027427  0.0007782   3.525 0.000431 ***
## rural_urban_flag4                       0.0042152  0.0009024   4.671 3.13e-06 ***
## rural_urban_flag5                       0.0019889  0.0012384   1.606 0.108364
## rural_urban_flag6                       0.0062219  0.0007842   7.934 2.97e-15 ***
## rural_urban_flag7                       0.0061286  0.0008763   6.994 3.28e-12 ***
## rural_urban_flag8                       0.0080603  0.0010549   7.641 2.87e-14 ***
## rural_urban_flag9                       0.0087384  0.0010394   8.407  < 2e-16 ***
## farming_flag1                          -0.0004908  0.0006952  -0.706 0.480292
## pop_loss_flag1                          0.0012508  0.0007085   1.765 0.077583 .
## poverty_flag1                          -0.0001839  0.0009396  -0.196 0.844834
## minority_indicator                     -0.0015130  0.0014795  -1.023 0.306562
## lowincome_indicator                     0.0169150  0.0013200  12.814  < 2e-16 ***
## lessthanhs_indicator                    0.0012740  0.0008664   1.470 0.141555
## cancer_indicator                        0.0016095  0.0011692   1.377 0.168733
## poverty_flag1:minority_indicator        0.0047655  0.0019491   2.445 0.014544 *
## pop_loss_flag1:lessthanhs_indicator -0.0030666  0.0013782  -2.225 0.026154 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01008 on 3036 degrees of freedom
## Multiple R-squared:  0.5692, Adjusted R-squared:  0.5596
## F-statistic:    59 on 68 and 3036 DF,  p-value: < 2.2e-16
```

Farming flag, population loss flag, poverty flag, minority indicator, less than hs indicator and cancer indicator are not statistically signficant. However, population loss flag, poverty flag, minority indicator and less than hs indicator are main effects to the two interactions in the model, therefore it is wise to include them nonetheless.

```
# Remove not statistically significant variables, excluding main effects of interactions (at the 0.05 a
mod4 = update(mod3,.~.-farming_flag - cancer_indicator)
summary(mod4)
```

```
##
## Call:
## lm(formula = energy_burden ~ state + log_pop + pct_unemployed +
##     pct_mining_oil_gas + rural_urban_flag + pop_loss_flag + poverty_flag +
##     minority_indicator + lowincome_indicator + lessthanhs_indicator +
##     poverty_flag:minority_indicator + pop_loss_flag:lessthanhs_indicator,
##     data = energy2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.043638 -0.006061  0.000226  0.006158  0.039221
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         0.0825028  0.0029418  28.045  < 2e-16 ***
## stateAR                            -0.0089299  0.0017112  -5.219 1.92e-07 ***
## stateAZ                            -0.0082087  0.0029222  -2.809 0.005000 **
## stateCA                            -0.0133205  0.0019072  -6.984 3.50e-12 ***
## stateCO                            -0.0087794  0.0018216  -4.819 1.51e-06 ***
## stateCT                             0.0024433  0.0038081   0.642 0.521167
## stateDC                            -0.0051855  0.0101907  -0.509 0.610894
## stateDE                            -0.0056686  0.0059665  -0.950 0.342153
## stateFL                            -0.0079742  0.0017611  -4.528 6.18e-06 ***
## stateGA                            -0.0053654  0.0014797  -3.626 0.000292 ***
## stateIA                            -0.0060893  0.0016616  -3.665 0.000252 ***
## stateID                            -0.0126285  0.0019859  -6.359 2.33e-10 ***
## stateIL                            -0.0094644  0.0016555  -5.717 1.19e-08 ***
## stateIN                            -0.0040812  0.0016578  -2.462 0.013876 *
## stateKS                            -0.0085070  0.0016485  -5.160 2.62e-07 ***
## stateKY                            -0.0118001  0.0015807  -7.465 1.08e-13 ***
## stateLA                            -0.0044845  0.0018003  -2.491 0.012796 *
## stateMA                             0.0019875  0.0030222   0.658 0.510815
## stateMD                            -0.0067132  0.0024330  -2.759 0.005828 **
## stateME                             0.0118005  0.0028306   4.169 3.15e-05 ***
## stateMI                            -0.0012989  0.0017730  -0.733 0.463861
## stateMN                            -0.0094276  0.0016951  -5.562 2.90e-08 ***
## stateMO                            -0.0051963  0.0015737  -3.302 0.000971 ***
## stateMS                            -0.0081362  0.0017000  -4.786 1.78e-06 ***
## stateMT                            -0.0030581  0.0018945  -1.614 0.106593
## stateNC                            -0.0070772  0.0016138  -4.386 1.20e-05 ***
## stateND                            -0.0097235  0.0019636  -4.952 7.75e-07 ***
## stateNE                            -0.0075343  0.0017192  -4.382 1.21e-05 ***
## stateNH                             0.0054036  0.0034503   1.566 0.117422
## stateNJ                            -0.0110948  0.0026171  -4.239 2.31e-05 ***
## stateNM                            -0.0128670  0.0022531  -5.711 1.23e-08 ***
## stateNV                            -0.0166298  0.0028517  -5.832 6.07e-09 ***
## stateNY                            -0.0019824  0.0018390  -1.078 0.281110
## stateOH                            -0.0035370  0.0016948  -2.087 0.036970 *
## stateOK                            -0.0070218  0.0017291  -4.061 5.01e-05 ***
## stateOR                            -0.0202333  0.0021263  -9.516  < 2e-16 ***
## statePA                            -0.0002407  0.0018268  -0.132 0.895192
## stateRI                            -0.0041399  0.0047227  -0.877 0.380779
## stateSC                            -0.0017257  0.0019423  -0.888 0.374349
## stateSD                            -0.0142381  0.0018231  -7.810 7.82e-15 ***
```

```
## stateTN                                -0.0121852  0.0016381  -7.439 1.32e-13 ***
## stateTX                                -0.0110886  0.0014491  -7.652 2.63e-14 ***
## stateUT                                -0.0191117  0.0022837  -8.369  < 2e-16 ***
## stateVA                                -0.0092757  0.0015468  -5.997 2.25e-09 ***
## stateVT                                 0.0058262  0.0030003   1.942 0.052244 .
## stateWA                                -0.0199377  0.0020919  -9.531  < 2e-16 ***
## stateWI                                -0.0104045  0.0017643  -5.897 4.10e-09 ***
## stateWV                                -0.0162731  0.0019137  -8.504  < 2e-16 ***
## stateWY                                -0.0132542  0.0025515  -5.195 2.19e-07 ***
## log_pop                                -0.0105058  0.0005050 -20.803  < 2e-16 ***
## pct_unemployed                          0.1490969  0.0124562  11.970  < 2e-16 ***
## pct_mining_oil_gas                     -0.0510505  0.0069262  -7.371 2.18e-13 ***
## rural_urban_flag2                       0.0013480  0.0007437   1.813 0.069998 .
## rural_urban_flag3                       0.0027588  0.0007781   3.546 0.000397 ***
## rural_urban_flag4                       0.0042059  0.0009001   4.673 3.10e-06 ***
## rural_urban_flag5                       0.0020285  0.0012352   1.642 0.100654
## rural_urban_flag6                       0.0062351  0.0007820   7.973 2.17e-15 ***
## rural_urban_flag7                       0.0061368  0.0008729   7.031 2.53e-12 ***
## rural_urban_flag8                       0.0080543  0.0010540   7.641 2.86e-14 ***
## rural_urban_flag9                       0.0086777  0.0010375   8.364  < 2e-16 ***
## pop_loss_flag1                          0.0011976  0.0006963   1.720 0.085550 .
## poverty_flag1                          -0.0001307  0.0009391  -0.139 0.889284
## minority_indicator                     -0.0013421  0.0014687  -0.914 0.360890
## lowincome_indicator                     0.0169251  0.0013200  12.822  < 2e-16 ***
## lessthanhs_indicator                    0.0011130  0.0008600   1.294 0.195679
## poverty_flag1:minority_indicator        0.0046350  0.0019447   2.383 0.017213 *
## pop_loss_flag1:lessthanhs_indicator -0.0031069  0.0013769  -2.256 0.024111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01008 on 3038 degrees of freedom
## Multiple R-squared:  0.5689, Adjusted R-squared:  0.5595
## F-statistic: 60.74 on 66 and 3038 DF,  p-value: < 2.2e-16
```

Select the best model variation based on the AIC score

```
# Starting with the full model (backward)
step(mod4,direction=c('backward'))
```

```
## Start:  AIC=-28484.67
## energy_burden ~ state + log_pop + pct_unemployed + pct_mining_oil_gas +
##     rural_urban_flag + pop_loss_flag + poverty_flag + minority_indicator +
##     lowincome_indicator + lessthanhs_indicator + poverty_flag:minority_indicator +
##     pop_loss_flag:lessthanhs_indicator
##
##                                       Df Sum of Sq     RSS    AIC
## <none>                                             0.30845 -28485
## - pop_loss_flag:lessthanhs_indicator  1  0.000517 0.30897 -28482
## - poverty_flag:minority_indicator     1  0.000577 0.30903 -28481
## - pct_mining_oil_gas                  1  0.005516 0.31397 -28432
## - rural_urban_flag                    8  0.010532 0.31898 -28396
## - pct_unemployed                      1  0.014547 0.32300 -28344
## - lowincome_indicator                 1  0.016691 0.32514 -28323
## - log_pop                             1  0.043940 0.35239 -28073
## - state                              48  0.066045 0.37450 -27978
```

```
##
## Call:
## lm(formula = energy_burden ~ state + log_pop + pct_unemployed +
##     pct_mining_oil_gas + rural_urban_flag + pop_loss_flag + poverty_flag +
##     minority_indicator + lowincome_indicator + lessthanhs_indicator +
##     poverty_flag:minority_indicator + pop_loss_flag:lessthanhs_indicator,
##     data = energy2)
##
## Coefficients:
##                          (Intercept)                              stateAR
##                            0.0825028                            -0.0089299
##                              stateAZ                              stateCA
##                           -0.0082087                            -0.0133205
##                              stateCO                              stateCT
##                           -0.0087794                             0.0024433
##                              stateDC                              stateDE
##                           -0.0051855                            -0.0056686
##                              stateFL                              stateGA
##                           -0.0079742                            -0.0053654
##                              stateIA                              stateID
##                           -0.0060893                            -0.0126285
##                              stateIL                              stateIN
##                           -0.0094644                            -0.0040812
##                              stateKS                              stateKY
##                           -0.0085070                            -0.0118001
##                              stateLA                              stateMA
##                           -0.0044845                             0.0019875
##                              stateMD                              stateME
##                           -0.0067132                             0.0118005
##                              stateMI                              stateMN
##                           -0.0012989                            -0.0094276
##                              stateMO                              stateMS
##                           -0.0051963                            -0.0081362
##                              stateMT                              stateNC
##                           -0.0030581                            -0.0070772
##                              stateND                              stateNE
##                           -0.0097235                            -0.0075343
##                              stateNH                              stateNJ
##                            0.0054036                            -0.0110948
##                              stateNM                              stateNV
##                           -0.0128670                            -0.0166298
##                              stateNY                              stateOH
##                           -0.0019824                            -0.0035370
##                              stateOK                              stateOR
##                           -0.0070218                            -0.0202333
##                              statePA                              stateRI
##                           -0.0002407                            -0.0041399
##                              stateSC                              stateSD
##                           -0.0017257                            -0.0142381
##                              stateTN                              stateTX
##                           -0.0121852                            -0.0110886
##                              stateUT                              stateVA
##                           -0.0191117                            -0.0092757
##                              stateVT                              stateWA
```

```
##                           0.0058262                           -0.0199377
##                             stateWI                              stateWV
##                          -0.0104045                           -0.0162731
##                             stateWY                              log_pop
##                          -0.0132542                           -0.0105058
##                       pct_unemployed                    pct_mining_oil_gas
##                           0.1490969                           -0.0510505
##                     rural_urban_flag2                     rural_urban_flag3
##                           0.0013480                            0.0027588
##                     rural_urban_flag4                     rural_urban_flag5
##                           0.0042059                            0.0020285
##                     rural_urban_flag6                     rural_urban_flag7
##                           0.0062351                            0.0061368
##                     rural_urban_flag8                     rural_urban_flag9
##                           0.0080543                            0.0086777
##                       pop_loss_flag1                        poverty_flag1
##                           0.0011976                           -0.0001307
##                    minority_indicator                  lowincome_indicator
##                          -0.0013421                            0.0169251
##                  lessthanhs_indicator     poverty_flag1:minority_indicator
##                           0.0011130                            0.0046350
## pop_loss_flag1:lessthanhs_indicator
##                          -0.0031069
```

The smallest AIC of 27978 corresponds to model without state. Let's test that

```
# Remove state from model and check AIC
mod5 = update(mod4,.~.-state)
ols_aic(mod5, method = 'SAS')
```

```
## [1] -27978.24
```

This corresponds to the value calculated in the step function.

We can now try other metrics to determine the best model.

```
# Use regsubsets using model call from mod4 (without removing the state yet)
T = 20

reg_subsets=regsubsets(energy_burden ~ state + log_pop + pct_unemployed + pct_mining_oil_gas +
    rural_urban_flag + pop_loss_flag + poverty_flag + minority_indicator +
    lowincome_indicator + lessthanhs_indicator + poverty_flag:minority_indicator +
    pop_loss_flag:lessthanhs_indicator, data = energy2, really.big = T)

reg_summary = summary(reg_subsets)
print(reg_summary)
```

```
## Subset selection object
## Call: regsubsets.formula(energy_burden ~ state + log_pop + pct_unemployed +
##     pct_mining_oil_gas + rural_urban_flag + pop_loss_flag + poverty_flag +
##     minority_indicator + lowincome_indicator + lessthanhs_indicator +
##     poverty_flag:minority_indicator + pop_loss_flag:lessthanhs_indicator,
##     data = energy2, really.big = T)
## 66 Variables  (and intercept)
##                                    Forced in Forced out
## stateAR                                FALSE      FALSE
## stateAZ                                FALSE      FALSE
```

```
## stateCA                    FALSE      FALSE
## stateCO                    FALSE      FALSE
## stateCT                    FALSE      FALSE
## stateDC                    FALSE      FALSE
## stateDE                    FALSE      FALSE
## stateFL                    FALSE      FALSE
## stateGA                    FALSE      FALSE
## stateIA                    FALSE      FALSE
## stateID                    FALSE      FALSE
## stateIL                    FALSE      FALSE
## stateIN                    FALSE      FALSE
## stateKS                    FALSE      FALSE
## stateKY                    FALSE      FALSE
## stateLA                    FALSE      FALSE
## stateMA                    FALSE      FALSE
## stateMD                    FALSE      FALSE
## stateME                    FALSE      FALSE
## stateMI                    FALSE      FALSE
## stateMN                    FALSE      FALSE
## stateMO                    FALSE      FALSE
## stateMS                    FALSE      FALSE
## stateMT                    FALSE      FALSE
## stateNC                    FALSE      FALSE
## stateND                    FALSE      FALSE
## stateNE                    FALSE      FALSE
## stateNH                    FALSE      FALSE
## stateNJ                    FALSE      FALSE
## stateNM                    FALSE      FALSE
## stateNV                    FALSE      FALSE
## stateNY                    FALSE      FALSE
## stateOH                    FALSE      FALSE
## stateOK                    FALSE      FALSE
## stateOR                    FALSE      FALSE
## statePA                    FALSE      FALSE
## stateRI                    FALSE      FALSE
## stateSC                    FALSE      FALSE
## stateSD                    FALSE      FALSE
## stateTN                    FALSE      FALSE
## stateTX                    FALSE      FALSE
## stateUT                    FALSE      FALSE
## stateVA                    FALSE      FALSE
## stateVT                    FALSE      FALSE
## stateWA                    FALSE      FALSE
## stateWI                    FALSE      FALSE
## stateWV                    FALSE      FALSE
## stateWY                    FALSE      FALSE
## log_pop                    FALSE      FALSE
## pct_unemployed             FALSE      FALSE
## pct_mining_oil_gas         FALSE      FALSE
## rural_urban_flag2          FALSE      FALSE
## rural_urban_flag3          FALSE      FALSE
## rural_urban_flag4          FALSE      FALSE
## rural_urban_flag5          FALSE      FALSE
## rural_urban_flag6          FALSE      FALSE
```

```
## rural_urban_flag7                       FALSE      FALSE
## rural_urban_flag8                       FALSE      FALSE
## rural_urban_flag9                       FALSE      FALSE
## pop_loss_flag1                          FALSE      FALSE
## poverty_flag1                           FALSE      FALSE
## minority_indicator                      FALSE      FALSE
## lowincome_indicator                     FALSE      FALSE
## lessthanhs_indicator                    FALSE      FALSE
## poverty_flag1:minority_indicator        FALSE      FALSE
## pop_loss_flag1:lessthanhs_indicator     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          stateAR stateAZ stateCA stateCO stateCT stateDC stateDE stateFL
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
##          stateGA stateIA stateID stateIL stateIN stateKS stateKY stateLA
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
##          stateMA stateMD stateME stateMI stateMN stateMO stateMS stateMT
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     "*"     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     "*"     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     "*"     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     "*"     " "     " "     " "     " "     " "
##          stateNC stateND stateNE stateNH stateNJ stateNM stateNV stateNY
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
##          stateOH stateOK stateOR statePA stateRI stateSC stateSD stateTN
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
```

```
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     "*"     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     "*"     "*"     " "     " "     " "     " "
##          stateTX stateUT stateVA stateVT stateWA stateWI stateWV stateWY
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     "*"     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     "*"     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     "*"     " "     " "     " "
##          log_pop pct_unemployed pct_mining_oil_gas rural_urban_flag2
## 1  ( 1 ) "*"     " "             " "                " "
## 2  ( 1 ) "*"     " "             " "                " "
## 3  ( 1 ) "*"     "*"             " "                " "
## 4  ( 1 ) "*"     "*"             "*"                " "
## 5  ( 1 ) "*"     "*"             "*"                " "
## 6  ( 1 ) "*"     "*"             "*"                " "
## 7  ( 1 ) "*"     "*"             "*"                " "
## 8  ( 1 ) "*"     "*"             "*"                " "
##          rural_urban_flag3 rural_urban_flag4 rural_urban_flag5
## 1  ( 1 ) " "               " "               " "
## 2  ( 1 ) " "               " "               " "
## 3  ( 1 ) " "               " "               " "
## 4  ( 1 ) " "               " "               " "
## 5  ( 1 ) " "               " "               " "
## 6  ( 1 ) " "               " "               " "
## 7  ( 1 ) " "               " "               " "
## 8  ( 1 ) " "               " "               " "
##          rural_urban_flag6 rural_urban_flag7 rural_urban_flag8
## 1  ( 1 ) " "               " "               " "
## 2  ( 1 ) " "               " "               " "
## 3  ( 1 ) " "               " "               " "
## 4  ( 1 ) " "               " "               " "
## 5  ( 1 ) " "               " "               " "
## 6  ( 1 ) " "               " "               " "
## 7  ( 1 ) " "               " "               " "
## 8  ( 1 ) " "               " "               " "
##          rural_urban_flag9 pop_loss_flag1 poverty_flag1 minority_indicator
## 1  ( 1 ) " "               " "            " "           " "
## 2  ( 1 ) " "               " "            " "           " "
## 3  ( 1 ) " "               " "            " "           " "
## 4  ( 1 ) " "               " "            " "           " "
## 5  ( 1 ) " "               " "            " "           " "
## 6  ( 1 ) " "               " "            " "           " "
## 7  ( 1 ) " "               " "            " "           " "
## 8  ( 1 ) " "               " "            " "           " "
##          lowincome_indicator lessthanhs_indicator
## 1  ( 1 ) " "                 " "
## 2  ( 1 ) "*"                 " "
## 3  ( 1 ) "*"                 " "
## 4  ( 1 ) "*"                 " "
## 5  ( 1 ) "*"                 " "
```

```
## 6  ( 1 ) "*"                 " "
## 7  ( 1 ) "*"                 " "
## 8  ( 1 ) "*"                 " "
##          poverty_flag1:minority_indicator pop_loss_flag1:lessthanhs_indicator
## 1  ( 1 ) " "                              " "
## 2  ( 1 ) " "                              " "
## 3  ( 1 ) " "                              " "
## 4  ( 1 ) " "                              " "
## 5  ( 1 ) " "                              " "
## 6  ( 1 ) " "                              " "
## 7  ( 1 ) " "                              " "
## 8  ( 1 ) " "                              " "
```

Note: this ran for a very long time and produced non-sensical output.

Confidence Interval

Since we have an explanatory model and no new data, a confidence interval, which reflects uncertainty around the mean prediction is more appropriate than a prediction interval.

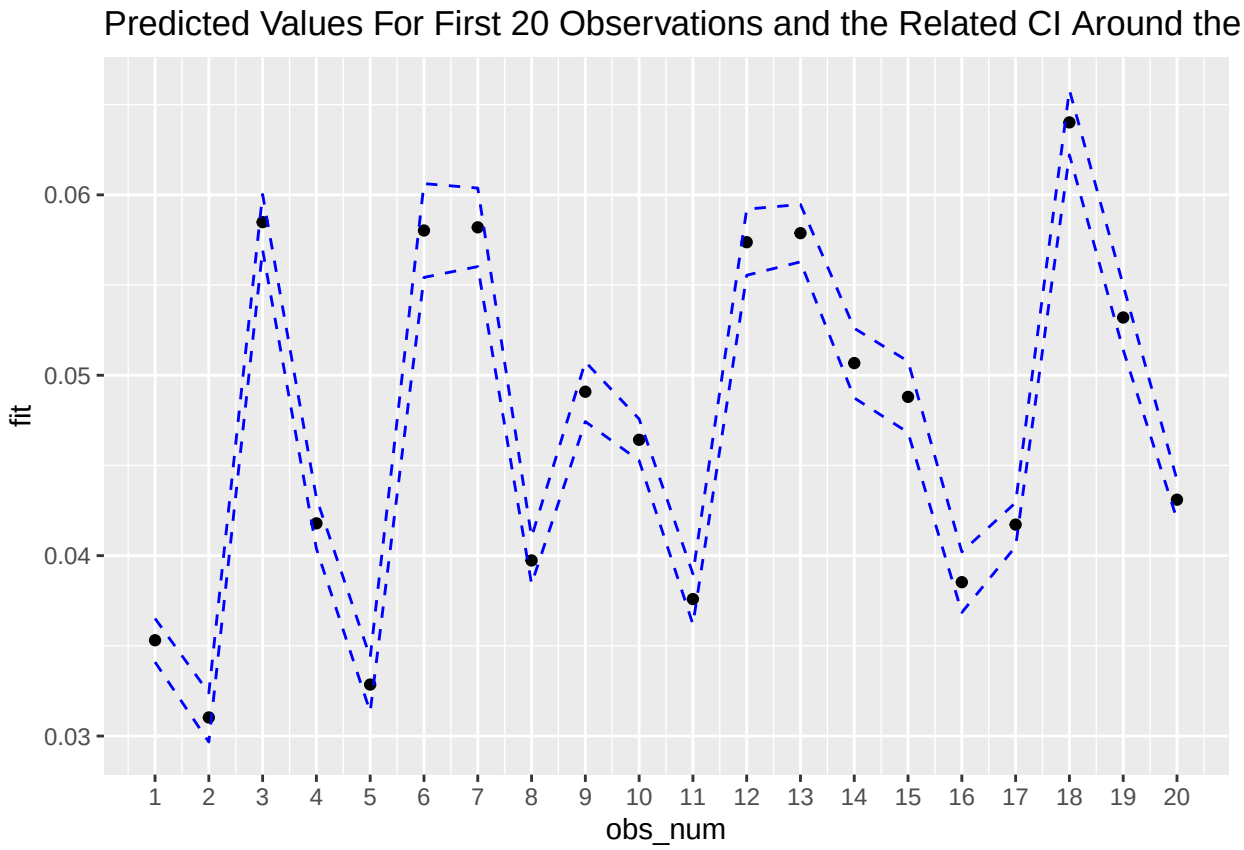
```
# For each observation, obtain confidence interval around the mean
confidence_mean = predict(mod5, interval = 'confidence')
head(confidence_mean)
```

```
##          fit        lwr        upr
## 1 0.03530980 0.03410203 0.03651758
## 2 0.03103016 0.02967077 0.03238955
## 3 0.05848826 0.05695232 0.06002421
## 4 0.04178962 0.04038738 0.04319186
## 5 0.03285292 0.03134429 0.03436155
## 6 0.05802395 0.05542489 0.06062300
```

Because we have multiple linear regression, it would be difficult to visualize confidence intervals around the mean in 2D space. However, we can plot the observation number on the x-axis, the predicted y value and the interval around it.

```
# Tag observation numbers onto the confidence data frame
confidence_mean = cbind(confidence_mean, seq(1, 3105, 1))
confidence_mean = data.frame(confidence_mean)
names(confidence_mean) = c('fit', 'lwr_ci', 'upr_ci', 'obs_num')
```

```
# Plot the predicted values and confidence interval around the mean for the first 20 observations
ggplot(confidence_mean[1:20,], aes(x = obs_num, y = fit)) + geom_point()+ geom_line(aes(y = lwr_ci), co
    geom_line(aes(y = upr_ci), color = "blue", linetype = "dashed") + ggtitle("Predicted Values For Firs
    scale_x_continuous(breaks = seq(1, 20, 1))
```



Prediction Interval (Illustration)

If we pretend that our data is new and we want to obtain a 95% prediction interval for each "new" observation. Here is how we do it:

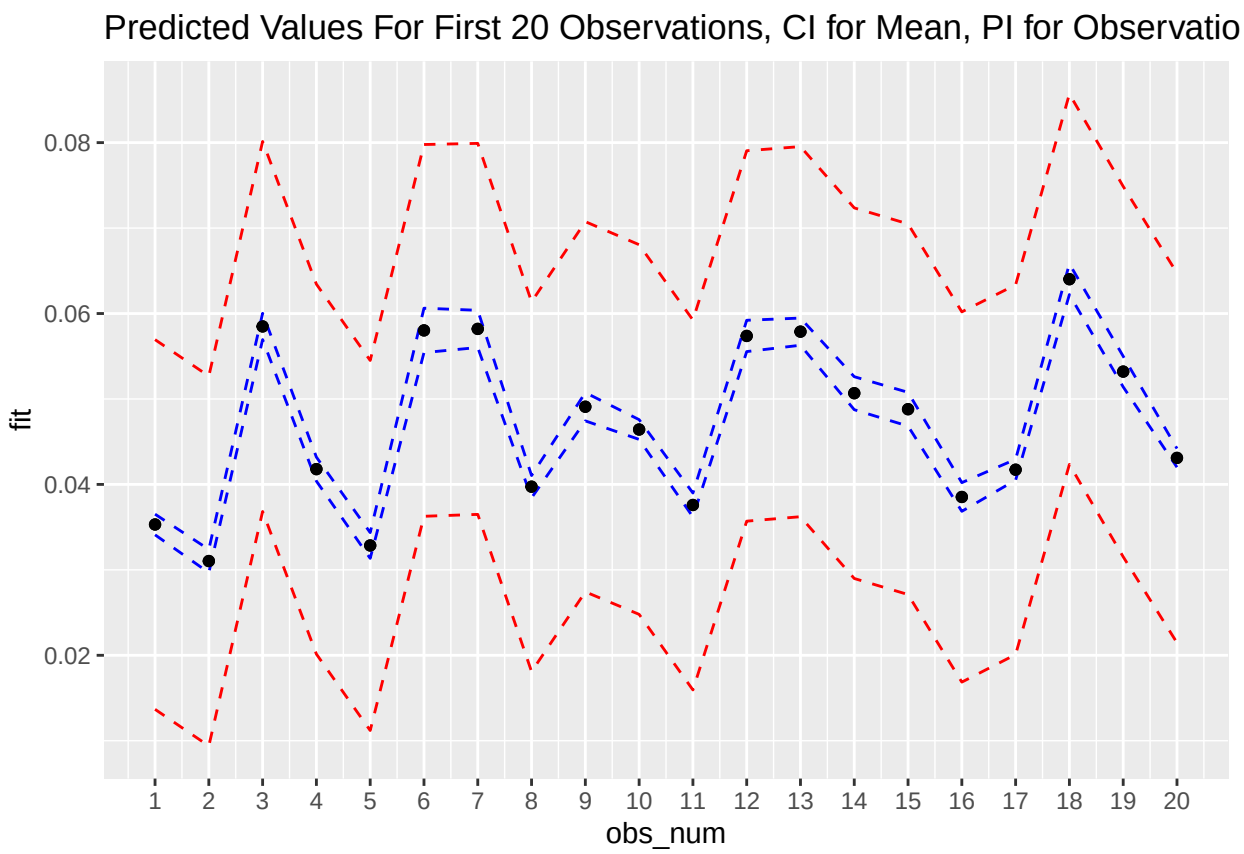
```
# Append the prediction interval
confidence_mean = cbind(confidence_mean, predict(mod5, interval = "prediction"))
```

```
## Warning in predict.lm(mod5, interval = "prediction"): predictions on current data refer to _future_ i
```

```
names(confidence_mean) = c("fit", "lwr_ci", "upr_ci", "obs_num", "fit2", "lwr_pi", "upr_pi")
```

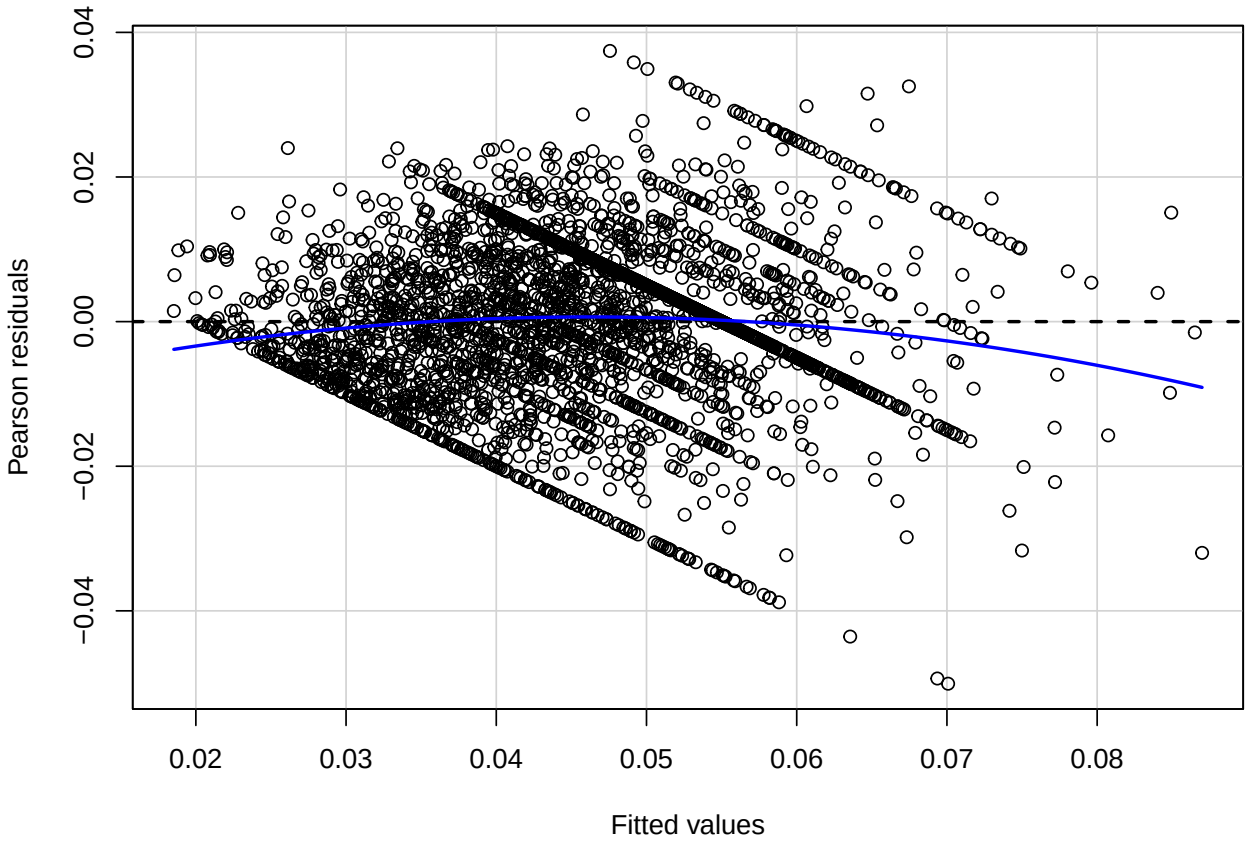
Just to illustrate how much wider the prediction intervals are, we will plot both the 95% confidence interval around the mean, as well as the 95% prediction interval for each observation (first 20)

```
# Plot the predicted values, confidence interval around the mean (blue) and prediction interval for the
ggplot(confidence_mean[1:20,], aes(x = obs_num, y = fit)) + geom_point()+ geom_line(aes(y = lwr_ci), co
    geom_line(aes(y = upr_ci), color = "blue", linetype = "dashed") + ggtitle("Predicted Values For Firs
    scale_x_continuous(breaks = seq(1, 20, 1)) + geom_point()+ geom_line(aes(y = lwr_pi), color = "red"
    geom_line(aes(y = upr_pi), color = "red", linetype = "dashed")
```



Check Model Assumptions

```
# Check for Equal Variance, Linear Relationship
residualPlot(mod5)
```

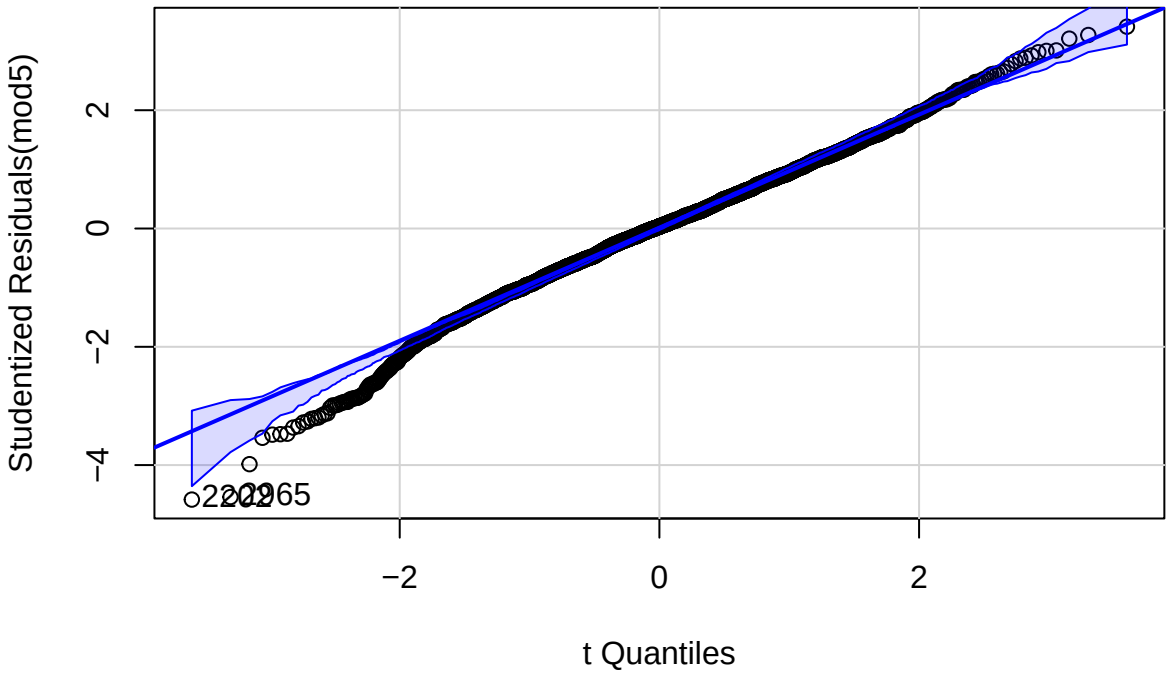


This plot shows that the model is extremely biased. This can be attributed to the biased response variable (energy burden), which was put together based on biased assumptions.

Equation for Model 5:

$$0.068 - 0.00912 * LogPop + 0.1418 * PctUnemployed - 0.06840 * PctMiningOilGas + 0.0017121 * I(RuralUrbanFlag = 2) + 0.0026$$

```
qqPlot(mod5)
```




```
## 2202 2965
## 2201 2962
```

The left tail significantly deviates from the line.

```
# K-S test for normality
library(nortest)
lillie.test(residuals(mod5))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuals(mod5)
## D = 0.030409, p-value = 5.829e-07
```

The p-value is extremely low, meaning we have to reject the null hypothesis that the underlying distribution is normal.

Switch to logic regression.

```
# Create a categorical response based on whether or not the energy burden is >4%
energy2$energy_burden_greater_4 = ifelse(energy2$energyburden_1_prop == 0 & energy2$energyburden_2_prop
```

```
# Check the proportion of counties with energy burden > 4%
energy2 %>%
  group_by(energy_burden_greater_4) %>%
  dplyr::summarise(cnt = n()) %>%
  dplyr::mutate(pct = round(cnt / sum(cnt),2))
```

```
## # A tibble: 2 x 3
##   energy_burden_greater_4   cnt   pct
##   <chr>                   <int> <dbl>
## 1 < 4%                     2004  0.65
## 2 >= 4%                    1101  0.35
```

65% of counties have energy burden < 4%, 35% have energy burden > 4%.

```
# Start with the initial model, excluding the vehicle count variables, because we know that they are co
log_mod0 = glm(ifelse(energy_burden_greater_4 == '>= 4%', 1, 0) ~ state + log_pop + pct_unemployed + pc
```

```
# Check the summary
summary(log_mod0)
```

```
##
## Call:
## glm(formula = ifelse(energy_burden_greater_4 == ">= 4%", 1, 0) ~
##     state + log_pop + pct_unemployed + pct_mining_oil_gas + rural_urban_flag +
##         farming_flag + pop_loss_flag + poverty_flag + minority_indicator +
##         lowincome_indicator + lessthanhs_indicator + cancer_indicator +
##         poverty_flag * minority_indicator + pop_loss_flag * lessthanhs_indicator +
##         cancer_indicator, family = "binomial", data = energy2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1001  -0.6267  -0.1985   0.6912   2.5788
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           9.76172    0.99284   9.832  < 2e-16 ***
## stateAR                              -0.66648    0.45400  -1.468 0.142102
## stateAZ                             -15.49031  521.41829  -0.030 0.976300
## stateCA                              -2.07510    0.71933  -2.885 0.003917 **
## stateCO                              -1.70461    0.54277  -3.141 0.001686 **
## stateCT                             -13.00252  802.85954  -0.016 0.987079
## stateDC                             -11.14618 2399.54477  -0.005 0.996294
## stateDE                             -13.27300 1330.82458  -0.010 0.992042
## stateFL                              -0.91143    0.54318  -1.678 0.093357 .
## stateGA                               0.15063    0.39037   0.386 0.699596
## stateIA                              -0.58041    0.46771  -1.241 0.214618
## stateID                              -0.92854    0.55515  -1.673 0.094408 .
## stateIL                              -1.30117    0.49074  -2.651 0.008014 **
## stateIN                               0.03400    0.46964   0.072 0.942289
## stateKS                              -0.54854    0.47584  -1.153 0.249003
## stateKY                              -1.39713    0.45502  -3.070 0.002137 **
## stateLA                              -0.62673    0.47116  -1.330 0.183465
## stateMA                               1.15038    1.03211   1.115 0.265027
## stateMD                               0.06902    0.76684   0.090 0.928281
## stateME                               2.90046    0.80565   3.600 0.000318 ***
## stateMI                              -0.66616    0.52336  -1.273 0.203070
## stateMN                              -1.52113    0.50988  -2.983 0.002851 **
## stateMO                              -0.26678    0.45474  -0.587 0.557428
## stateMS                              -0.47541    0.44232  -1.075 0.282463
## stateMT                              -1.98729    0.53647  -3.704 0.000212 ***
## stateNC                              -0.49410    0.47741  -1.035 0.300684
## stateND                              -1.20360    0.53614  -2.245 0.024772 *
## stateNE                              -0.92348    0.49046  -1.883 0.059717 .
## stateNH                               1.96638    0.89838   2.189 0.028611 *
## stateNJ                             -13.65492  466.02959  -0.029 0.976625
## stateNM                              -1.71400    0.67777  -2.529 0.011443 *
## stateNV                              -2.26134    0.91174  -2.480 0.013129 *
```

```
## stateNY                                  0.15991    0.55258   0.289 0.772286
## stateOH                                 -0.30577    0.50426  -0.606 0.544270
## stateOK                                 -0.58915    0.49328  -1.194 0.232339
## stateOR                                 -3.14193    0.81485  -3.856 0.000115 ***
## statePA                                 -0.32340    0.56168  -0.576 0.564767
## stateRI                                -14.06684 1034.15846  -0.014 0.989147
## stateSC                                  0.61895    0.52362   1.182 0.237188
## stateSD                                 -2.06353    0.51054  -4.042 5.30e-05 ***
## stateTN                                 -1.98168    0.48744  -4.065 4.79e-05 ***
## stateTX                                 -1.30700    0.43346  -3.015 0.002568 **
## stateUT                                 -2.26004    0.76236  -2.965 0.003032 **
## stateVA                                 -0.30466    0.45029  -0.677 0.498676
## stateVT                                  1.94856    0.79153   2.462 0.013825 *
## stateWA                                 -3.12346    0.83884  -3.724 0.000196 ***
## stateWI                                 -1.13453    0.53707  -2.112 0.034648 *
## stateWV                                 -2.60617    0.55969  -4.656 3.22e-06 ***
## stateWY                                 -1.72114    0.74430  -2.312 0.020755 *
## log_pop                                 -2.75084    0.19448 -14.145  < 2e-16 ***
## pct_unemployed                          19.71125    3.57049   5.521 3.38e-08 ***
## pct_mining_oil_gas                      -3.09863    1.82666  -1.696 0.089821 .
## rural_urban_flag2                        0.13012    0.27328   0.476 0.633962
## rural_urban_flag3                        0.33555    0.26250   1.278 0.201154
## rural_urban_flag4                       -0.03310    0.31035  -0.107 0.915061
## rural_urban_flag5                       -1.62272    0.75710  -2.143 0.032085 *
## rural_urban_flag6                        0.98173    0.23104   4.249 2.15e-05 ***
## rural_urban_flag7                        0.91343    0.24927   3.664 0.000248 ***
## rural_urban_flag8                        1.38871    0.28685   4.841 1.29e-06 ***
## rural_urban_flag9                        0.92935    0.28072   3.311 0.000931 ***
## farming_flag1                            0.07619    0.16948   0.450 0.653054
## pop_loss_flag1                          -0.08127    0.18046  -0.450 0.652442
## poverty_flag1                            0.20088    0.23764   0.845 0.397933
## minority_indicator                      -1.34934    0.43825  -3.079 0.002077 **
## lowincome_indicator                      1.83622    0.34621   5.304 1.13e-07 ***
## lessthanhs_indicator                     0.60242    0.23094   2.609 0.009092 **
## cancer_indicator                        -0.04388    0.32475  -0.135 0.892519
## poverty_flag1:minority_indicator        -0.85861    0.55680  -1.542 0.123064
## pop_loss_flag1:lessthanhs_indicator     -0.31394    0.35131  -0.894 0.371519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4038.0  on 3104  degrees of freedom
## Residual deviance: 2557.6  on 3036  degrees of freedom
## AIC: 2695.6
##
## Number of Fisher Scoring iterations: 15
```

Statistically insignificant variables: both interaction terms, cancer indicator, poverty flag, population loss flag, farming flag, percent mining & oil_gas.

```
# Obtain fitted values
energy2$fitted_probabilities = fitted.values(log_mod0)
energy2$fitted_energy_burden_greater_4 = ifelse(predict.glm(log_mod0, type = "response") > 0.5, 1, 0)
```

```
# Look at the confusion matrix
table(energy2$energy_burden_greater_4,energy2$fitted_energy_burden_greater_4)
```

```
##
##            0    1
##   < 4%  1721  283
##   >= 4%  342  759
```

```
# Calculate Accuracy: 79.9%
(1721 + 759) / 3105
```

```
## [1] 0.7987118
```

```
# Calculate True Positive Rate: 68.9%
759 / (759 + 342)
```

```
## [1] 0.6893733
```

```
# Calculate False Positive Rate: 14.12%
283 / (283 + 1721)
```

```
## [1] 0.1412176
```

Accuracy is fairly high. So is the TPR.

```
library(MLmetrics)
```

```
##
## Attaching package: 'MLmetrics'
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

```
# Check F1, Precision and Recall
F1_Score(ifelse(energy2$energy_burden_greater_4 == '>= 4%', 1, 0), energy2$fitted_energy_burden_greater
```

```
## [1] 0.8463241
```

```
Precision(ifelse(energy2$energy_burden_greater_4 == '>= 4%', 1, 0), energy2$fitted_energy_burden_greate
```

```
## [1] 0.834222
```

```
Recall(ifelse(energy2$energy_burden_greater_4 == '>= 4%', 1, 0), energy2$fitted_energy_burden_greater_4
```

```
## [1] 0.8587824
```

F1 = 2 * ((Precision * Recall) / (Precision + Recall))

Recall = TP / (TP + FN). Recall measures classifier completeness: how sensitive our model is to identifying true positive measures.

Precision = TP / (TP + FP). Precision measures classifier exactness: did the model correctly identify true positive values, or did it overdo it and label most observations as positive?

In our case, because precision and recall are fairly close together, F1 score is basically their average.

```
# Check Area Under the ROC Curve
AUC(ifelse(energy2$energy_burden_greater_4 == '>= 4%', 1, 0), energy2$fitted_energy_burden_greater_4)
```

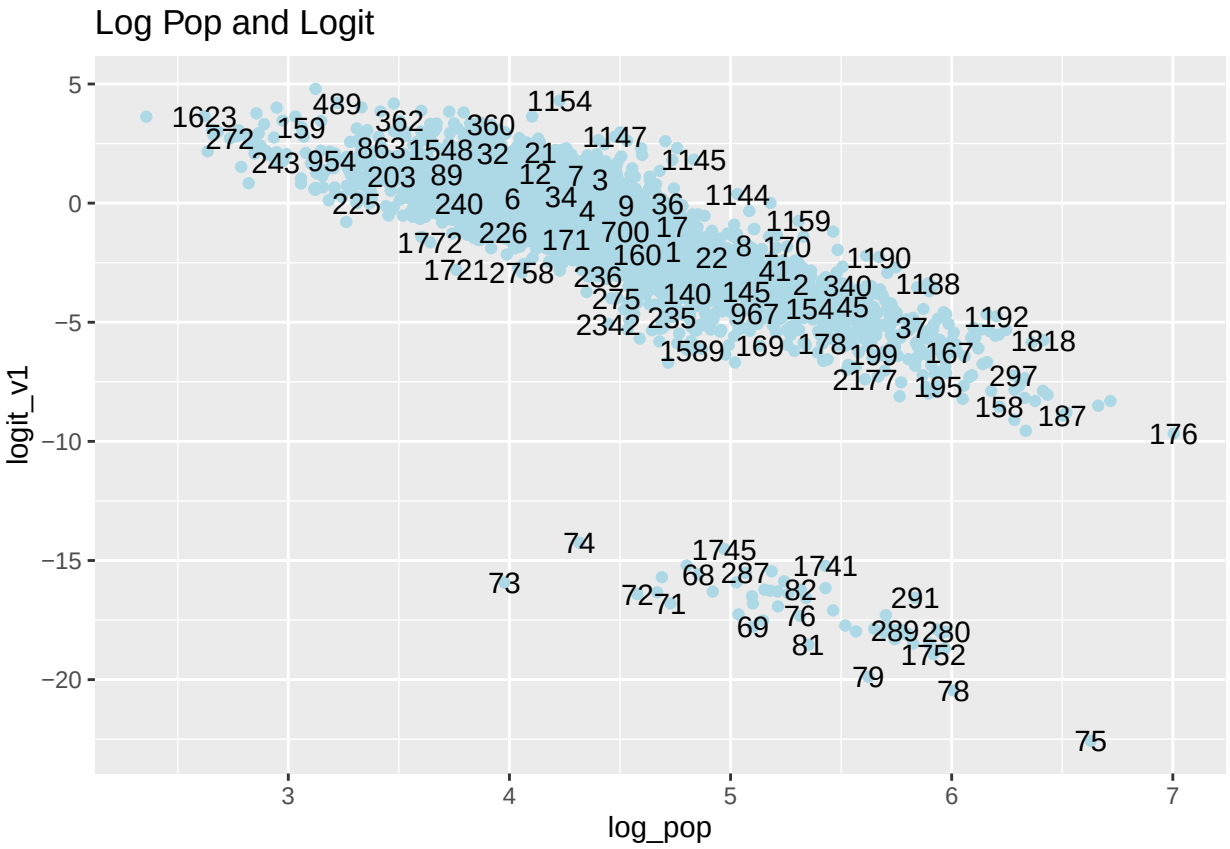
```
## [1] 0.7813145
```

78% is pretty good.

Check Assumptions for Logistic Regression

```
# Check linear relationship between logit and continuous predictors

# Calculate the logit variable = log(p / (1-p))
energy2$logit_v1 = log(energy2$fitted_probabilities / ( 1 - energy2$fitted_probabilities))
```

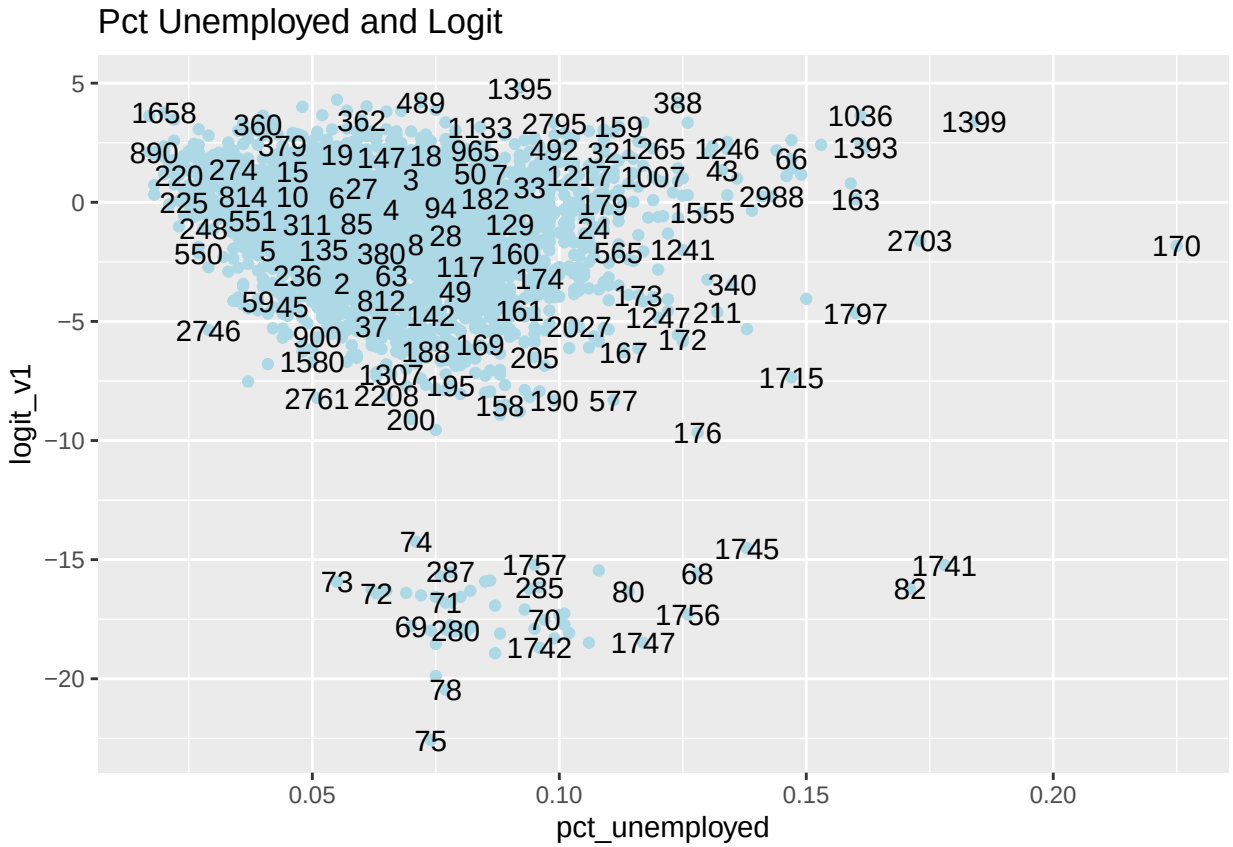
```
# Graph Relationships, label points to ID outliers

# Logit and log_pop
ggplot(energy2, aes(x = log_pop, y = logit_v1, label = rownames(energy2))) + geom_point(color = 'light
```



This looks fairly linear with the exception of the scattered points at the bottom. Maybe those are influential observations.

```
# Record outliers at the bottom
log_pop_outliers = c(73, 74, 72, 71, 68, 1745, 287, 1750, 69, 1741, 82, 76, 81, 291, 289, 79, 280, 1752
```

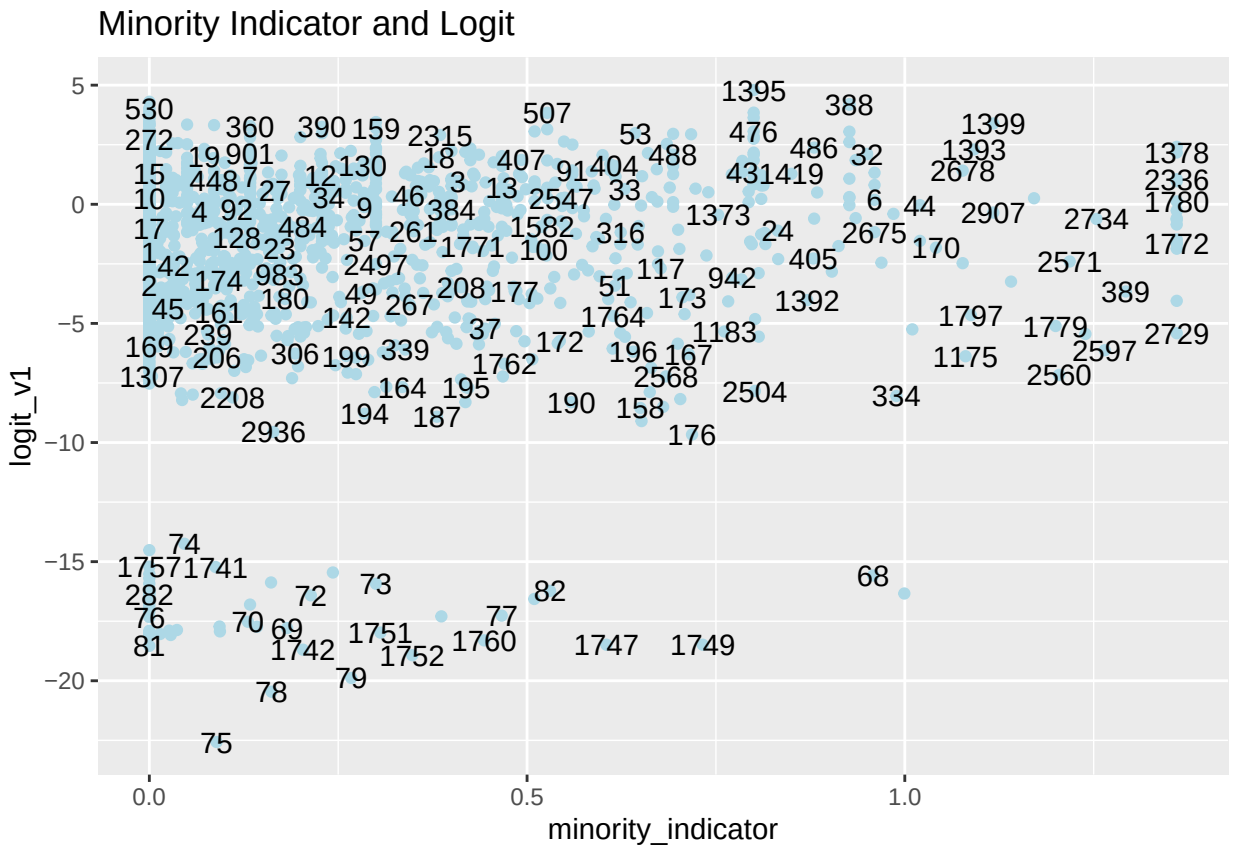
```
# Logit and Pct Unemployed
ggplot(energy2, aes(x = pct_unemployed, y = logit_v1, label = rownames(energy2))) + geom_point(color =
```



There is definitely more drift here at the top. There are outliers at the bottom.

```
# Record outliers at the bottom
unemployed_outliers = c(73, 72, 74, 287, 71, 69, 280, 78, 75, 1757, 285, 70, 1742, 1746, 80, 1747, 1756
```

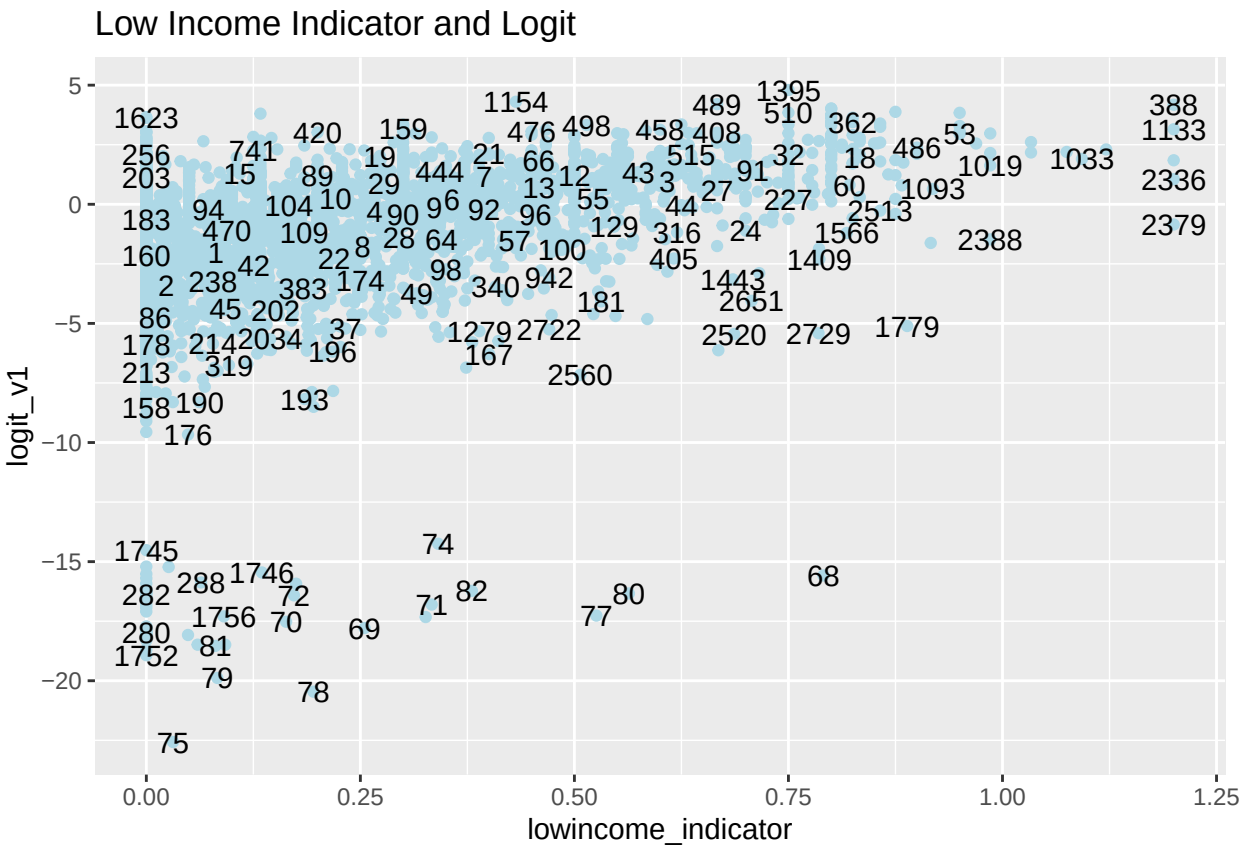
```
# Logit and Minority Indicator
ggplot(energy2, aes(x = minority_indicator, y = logit_v1, label = rownames(energy2))) + geom_point(colo
```



This does not look very linear. However, based on literature, minorities experience higher energy burdens, so it makes sense to keep this variable.

```
# Record outliers at the bottom
minority_outliers = c(774, 1757, 282, 76, 81, 1741, 70, 288, 69, 72, 1742, 78, 75, 1746, 73, 1751, 79,
```

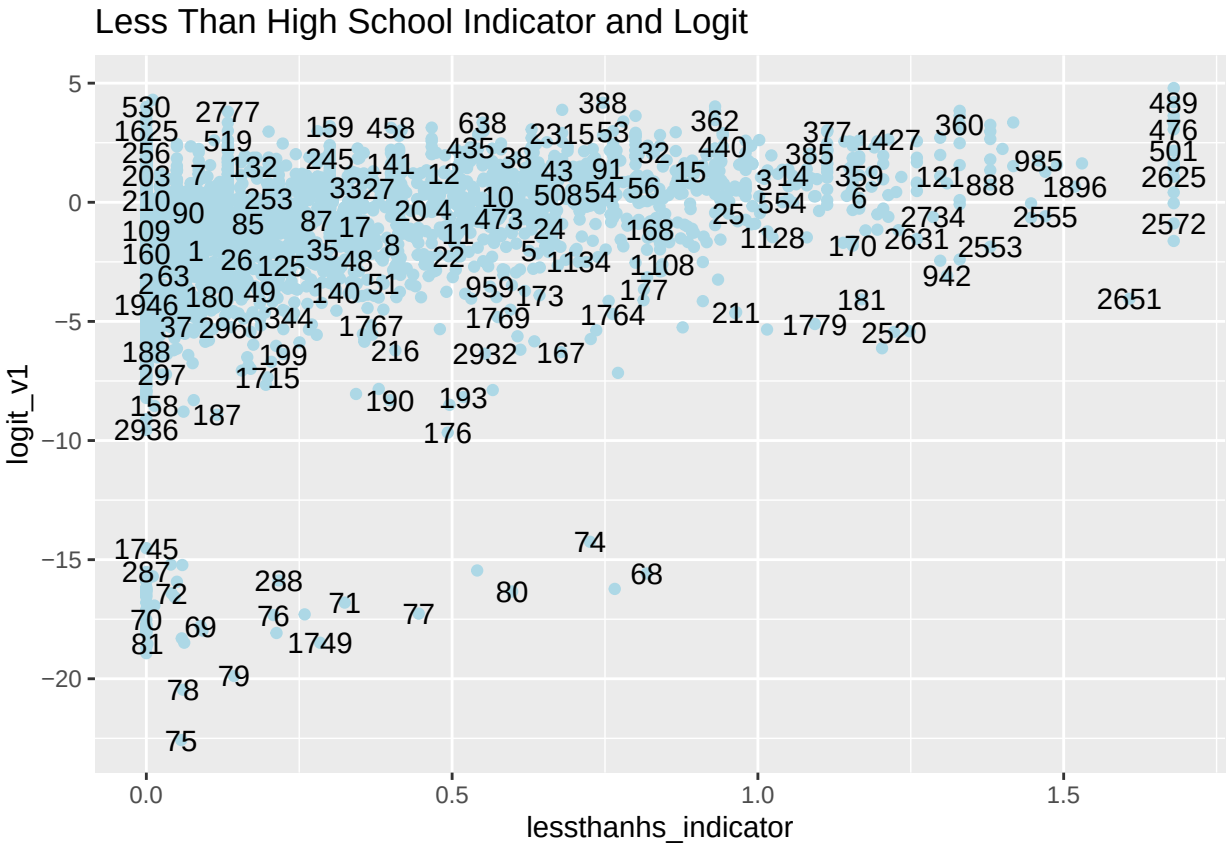
```
# Logit and Low Income Indicator
ggplot(energy2, aes(x = lowincome_indicator, y = logit_v1, label = rownames(energy2))) + geom_point(col
```



This looks more linear than the minority indicator. Again, with a caveat of noise at the bottom.

```
# Record outliers at the bottom
lowincome_outliers = c(1745, 282, 280, 1752, 288, 1756, 81, 79, 1746, 70, 72, 78, 69, 71, 74, 82, 77, 8
```

```
# Logit and Less Than HS Indicator
ggplot(energy2, aes(x = lessthanhs_indicator, y = logit_v1, label = rownames(energy2))) + geom_point(co
```



Similarly for the low income indicator, this could pass for a linear relationship.

```
# Record outliers at the bottom
lesshs_outliers = c(1745, 287, 72, 70, 81, 69, 79, 78, 75, 288, 76, 1749, 71, 77, 1746, 80, 74, 82, 68)
```

```
# Check for overlap between Unemployed and Minority Indicator
chart_outliers = Reduce(intersect, list(unemployed_outliers, minority_outliers))
chart_outliers
```

```
##  [1]   73   72   69   78   75 1757   70 1742 1746   80 1747 1756   68   82 1741
```

There is considerable overlap between outliers at the bottom of charts for different variables. It makes sense to remove them.

```
# Check some of the common bottom outliers
energy2[chart_outliers, c("county", "state", "log_pop", "pct_unemployed", "minority_indicator", "lesstha
```

```
##          county state  log_pop pct_unemployed minority_indicator
## 73     Greenlee    AZ 3.977906          0.055         0.30000000
## 72       Graham    AZ 4.578399          0.063         0.21379310
## 69      Cochise    AZ 5.101331          0.070         0.18235294
## 78         Pima    AZ 6.008482          0.077         0.16174603
## 75     Maricopa    AZ 6.628789          0.074         0.08902196
## 1757      Salem    NJ 4.801651          0.095         0.00000000
## 70     Coconino    AZ 5.146801          0.097         0.13040816
## 1742     Bergen    NJ 5.968482          0.096         0.20328700
## 1746 Cumberland    NJ 5.185825          0.108         0.24300000
## 80   Santa Cruz    AZ 4.668237          0.114         0.99939394
## 1747      Essex    NJ 5.899577          0.117         0.60500739
## 1756    Passaic    NJ 5.702466          0.126         0.38657541
```

```
## 68       Apache    AZ 4.854440          0.128         0.95818182
## 82         Yuma    AZ 5.317706          0.171         0.53042254
## 1741   Atlantic    NJ 5.429007          0.178         0.08771734
##      lessthanhs_indicator lowincome_indicator
## 73             0.05000000          0.17500000
## 72             0.04137931          0.17241379
## 69             0.08882353          0.25490196
## 78             0.06000000          0.19523810
## 75             0.05700599          0.03113773
## 1757           0.04000000          0.00000000
## 70             0.00000000          0.16326531
## 1742           0.00000000          0.00000000
## 1746           0.54100000          0.13500000
## 80             0.59818182          0.56363636
## 1747           0.06196722          0.05991060
## 1756           0.25890417          0.09041100
## 68             0.81818182          0.79090909
## 82             0.76591549          0.38028169
## 1741           0.05880435          0.02608695
```

These feature high values in two or more columns simultaneously. For example, Santa Cruz, Arizona has high unemployment, high minority indicator, high less than hs indicator and high low income indicator. It is a true outlier observation.

```
# Check for multicollinearity
# We already did this for the linear model, check again
vif(log_mod0)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                                       GVIF Df GVIF^(1/(2*Df))
## state                            51.871596 48        1.041991
## log_pop                           2.992246  1        1.729811
## pct_unemployed                    2.568067  1        1.602519
## pct_mining_oil_gas                1.765706  1        1.328799
## rural_urban_flag                  3.172594  8        1.074827
## farming_flag                      1.854889  1        1.361943
## pop_loss_flag                     2.294036  1        1.514608
## poverty_flag                      2.620071  1        1.618663
## minority_indicator                3.774422  1        1.942787
## lowincome_indicator               2.507950  1        1.583651
## lessthanhs_indicator              2.738644  1        1.654885
## cancer_indicator                  2.077609  1        1.441391
## poverty_flag:minority_indicator   4.049293  1        2.012285
## pop_loss_flag:lessthanhs_indicator  1.965434  1        1.401939
```

All variables have GVIF below 5.

```
# Check for influential observations using Cook's Distance
# Cook's Distance captures how much the model changes if one observation is removed.
cooks_d = cooks.distance(log_mod0)

# Influential if a given Cook's D is 10 times greater than the mean
influential = cooks_d[(cooks_d > 10 * mean (cooks_d, na.rm = TRUE))]
length(influential)
```

```
## [1] 33
```

```
influential[1:10]
```

```
##        <NA>         534        1133        1149        1155        1182
##          NA 0.004592634 0.004048980 0.003465624 0.005336060 0.004019554
##        1189        1399        1461        1583
## 0.010916975 0.004149033 0.007898638 0.004134109
```

There are 32 counties that fit that description. Note, I picked 10x the mean to only check on the most extreme outliers.

```
# Let's look at these 33 counties to see how they are different
# Only consider variables that are stat sig and also that have a linear relationship with the logit
energy2[names(influential)[2:33],c("county", "state","log_pop", "pct_unemployed", "lowincome_indicator"
```

```
##             county state  log_pop pct_unemployed lowincome_indicator
## 534          Clark    ID 3.032216          0.040           0.8000000
## 1133        Tensas    LA 3.668945          0.084           1.2000000
## 1149      Kennebec    ME 5.084737          0.050           0.0233010
## 1155     Sagadahoc    ME 4.547492          0.047           0.0000000
## 1182     Worcester    MD 4.712347          0.112           0.0000000
## 1189      Franklin    MA 4.850861          0.074           0.0000000
## 1399     Jefferson    MS 3.866051          0.184           0.8571429
## 1461        Butler    MO 4.630763          0.063           0.4083333
## 1583 Golden Valley    MT 2.859739          0.047           0.8000000
## 1724       Mineral    NV 3.648165          0.056           0.1250000
## 1726      Pershing    NV 3.820267          0.049           0.0000000
## 1731       Belknap    NH 4.782759          0.069           0.0000000
## 1773       Harding    NM 2.661813          0.043           0.8000000
## 1778          Luna    NM 4.384962          0.159           0.8823530
## 1791          Taos    NM 4.517037          0.116           0.4772727
## 1933      Richmond    NC 4.655033          0.092           0.5365853
## 2175         Baker    OR 4.203685          0.072           0.3294118
## 2193          Lake    OR 3.894482          0.056           0.2222222
## 2202       Sherman    OR 3.205475          0.061           0.8000000
## 2206       Wallowa    OR 3.840357          0.071           0.0857143
## 2209       Wheeler    OR 3.154120          0.043           0.7500000
## 2553        Dimmit    TX 4.027879          0.064           0.7857143
## 2576     Glasscock    TX 3.155336          0.036           0.3000000
## 2756          Kane    UT 3.866287          0.054           0.0000000
## 2777         Essex    VT 3.792952          0.065           0.1333333
## 2784    Washington    VT 4.766985          0.048           0.0000000
## 2786       Windsor    VT 4.743721          0.053           0.0000000
## 2926      Columbia    WA 3.602169          0.073           0.0000000
## 2965       Calhoun    WV 3.868997          0.161           0.6333333
## 3091         Crook    WY 3.869818          0.039           0.0000000
## 3101        Platte    WY 3.938169          0.050           0.0222222
## 3103      Sublette    WY 3.997867          0.072           0.0000000
##      lessthanhs_indicator minority_indicator
## 534             1.6800000          0.8000000
## 1133            1.3299999          0.5266666
## 1149            0.0000000          0.0000000
## 1155            0.0000000          0.0000000
## 1182            0.0000000          0.0000000
## 1189            0.0000000          0.0000000
```

```
## 1399            0.7800001          1.1171429
## 1461            0.4550001          0.0000000
## 1583            0.0000000          0.0000000
## 1724            0.3000000          0.0500000
## 1726            0.4200000          0.0000000
## 1731            0.0000000          0.0000000
## 1773            0.0000000          0.8000000
## 1778            1.3094117          0.8129412
## 1791            0.0727273          0.5918182
## 1933            0.5653659          0.1902439
## 2175            0.0000000          0.0000000
## 2193            0.3555556          0.0000000
## 2202            0.0000000          0.0000000
## 2206            0.0000000          0.0000000
## 2209            0.0000000          0.0000000
## 2553            1.3800001          1.3600000
## 2576            1.6800000          0.8000000
## 2756            0.0857143          0.0000000
## 2777            0.1333333          0.0000000
## 2784            0.0000000          0.0000000
## 2786            0.0000000          0.0000000
## 2926            0.0000000          0.0000000
## 2965            0.5799999          0.0000000
## 3091            0.0000000          0.0000000
## 3101            0.0000000          0.0000000
## 3103            0.0000000          0.0000000
```

```
# Check against the summary of relevant variables
summary(energy2[,c("log_pop", "pct_unemployed", "lowincome_indicator", "lessthanhs_indicator", "minority
```

```
##     log_pop      pct_unemployed    lowincome_indicator lessthanhs_indicator
##  Min.   :2.358   Min.   :0.01700   Min.   :0.0000      Min.   :0.0000
##  1st Qu.:4.050   1st Qu.:0.05200   1st Qu.:0.0000      1st Qu.:0.0000
##  Median :4.413   Median :0.06500   Median :0.1389      Median :0.1309
##  Mean   :4.467   Mean   :0.06707   Mean   :0.2117      Mean   :0.2741
##  3rd Qu.:4.830   3rd Qu.:0.08000   3rd Qu.:0.3393      3rd Qu.:0.4425
##  Max.   :7.004   Max.   :0.22500   Max.   :1.2000      Max.   :1.6800
##  minority_indicator
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1069
##  3rd Qu.:0.0963
##  Max.   :1.3600
```

After a quick glance, looks like many of these counties have high unemployment (over 6%). In addition to that, they may have high values of other variables. Due to their high Cook's Distance, it makes sense to remove them.

Make changes to the model

```
# Remove extreme outliers based on Cook's D, as well as outliers spotted in the graphs
energy3 = energy2[!seq(nrow(energy2)) %in% c(chart_outliers, names(influential)),]
```

```
# Fit logistic regression model, excluding variables not statistically significant
log_mod1 = glm(ifelse(energy_burden_greater_4 == '>= 4%', 1, 0) ~ state + log_pop + pct_unemployed + ru
```

```r
# Obtain fitted values
energy3$fitted_probabilities = fitted.values(log_mod1)
energy3$fitted_energy_burden_greater_4 = ifelse(predict.glm(log_mod1, type = "response") > 0.5, 1, 0)

# Check model summary
summary(log_mod1)
```

```
##
## Call:
## glm(formula = ifelse(energy_burden_greater_4 == ">= 4%", 1, 0) ~
##     state + log_pop + pct_unemployed + rural_urban_flag + minority_indicator +
##         lowincome_indicator + lessthanhs_indicator, family = "binomial",
##     data = energy3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1734  -0.6243  -0.1812   0.6731   2.6486
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             10.17458    0.95671  10.635  < 2e-16 ***
## stateAR                 -0.66118    0.42701  -1.548 0.121528
## stateAZ                -15.72686  819.60773  -0.019 0.984691
## stateCA                 -1.84352    0.68874  -2.677 0.007436 **
## stateCO                 -1.63528    0.49881  -3.278 0.001044 **
## stateCT                -12.79680  799.75130  -0.016 0.987234
## stateDC                -10.74680 2399.54477  -0.004 0.996427
## stateDE                -13.04742 1331.06723  -0.010 0.992179
## stateFL                 -0.81787    0.52380  -1.561 0.118423
## stateGA                  0.22351    0.38474   0.581 0.561273
## stateIA                 -0.50800    0.41950  -1.211 0.225913
## stateID                 -0.73216    0.51698  -1.416 0.156711
## stateIL                 -1.23801    0.44369  -2.790 0.005267 **
## stateIN                  0.12941    0.42176   0.307 0.758961
## stateKS                 -0.54249    0.43046  -1.260 0.207575
## stateKY                 -1.30570    0.40695  -3.209 0.001334 **
## stateLA                 -0.53025    0.46228  -1.147 0.251366
## stateMA                  0.42370    1.26049   0.336 0.736770
## stateMD                 -0.14550    0.79759  -0.182 0.855247
## stateME                  3.13767    0.87217   3.598 0.000321 ***
## stateMI                 -0.50004    0.47678  -1.049 0.294277
## stateMN                 -1.42027    0.46063  -3.083 0.002047 **
## stateMO                 -0.23534    0.40601  -0.580 0.562160
## stateMS                 -0.33670    0.44063  -0.764 0.444779
## stateMT                 -1.92869    0.49364  -3.907 9.34e-05 ***
## stateNC                 -0.31682    0.42800  -0.740 0.459154
## stateND                 -1.30952    0.48527  -2.699 0.006964 **
## stateNE                 -0.88503    0.44515  -1.988 0.046794 *
## stateNH                  1.76132    0.97529   1.806 0.070929 .
## stateNJ                -13.20704  550.72417  -0.024 0.980868
## stateNM                 -1.36052    0.65523  -2.076 0.037857 *
## stateNV                 -4.45489    1.45905  -3.053 0.002264 **
## stateNY                  0.30351    0.51022   0.595 0.551939
```

```
## stateOH                -0.18114    0.45907  -0.395 0.693150
## stateOK                -0.65101    0.44163  -1.474 0.140454
## stateOR                -2.99238    0.81097  -3.690 0.000224 ***
## statePA                -0.19822    0.51983  -0.381 0.702971
## stateRI               -13.94228 1030.51789  -0.014 0.989205
## stateSC                 0.73209    0.50461   1.451 0.146834
## stateSD                -2.07809    0.46504  -4.469 7.87e-06 ***
## stateTN                -1.84925    0.44200  -4.184 2.87e-05 ***
## stateTX                -1.36800    0.37677  -3.631 0.000283 ***
## stateUT                -2.30204    0.74217  -3.102 0.001924 **
## stateVA                -0.16590    0.39909  -0.416 0.677634
## stateVT                 2.06985    0.77013   2.688 0.007195 **
## stateWA                -3.32192    0.92712  -3.583 0.000340 ***
## stateWI                -1.00098    0.49165  -2.036 0.041756 *
## stateWV                -2.70583    0.51822  -5.221 1.78e-07 ***
## stateWY                -1.63769    0.71371  -2.295 0.021756 *
## log_pop                -2.83902    0.19357 -14.667  < 2e-16 ***
## pct_unemployed         18.25318    3.50590   5.206 1.93e-07 ***
## rural_urban_flag2       0.08429    0.27707   0.304 0.760955
## rural_urban_flag3       0.34225    0.26412   1.296 0.195034
## rural_urban_flag4      -0.15090    0.31996  -0.472 0.637189
## rural_urban_flag5      -2.39445    1.03703  -2.309 0.020946 *
## rural_urban_flag6       0.90670    0.23114   3.923 8.76e-05 ***
## rural_urban_flag7       0.79336    0.24711   3.211 0.001325 **
## rural_urban_flag8       1.32809    0.29032   4.575 4.77e-06 ***
## rural_urban_flag9       0.80638    0.27986   2.881 0.003959 **
## minority_indicator     -1.74140    0.32623  -5.338 9.40e-08 ***
## lowincome_indicator     2.06682    0.33562   6.158 7.36e-10 ***
## lessthanhs_indicator    0.53167    0.21835   2.435 0.014892 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3982.4  on 3057  degrees of freedom
## Residual deviance: 2490.6  on 2996  degrees of freedom
## AIC: 2614.6
##
## Number of Fisher Scoring iterations: 15
```

```
# Look at the confusion matrix
table(energy3$energy_burden_greater_4,energy3$fitted_energy_burden_greater_4)
```

```
##
##            0    1
##   < 4%  1695  274
##   >= 4%  331  758
```

```
# Calculate Accuracy: 80.21%
Accuracy(energy3$fitted_energy_burden_greater_4, ifelse(energy3$energy_burden_greater_4 == '>= 4%', 1, 0
```

```
## [1] 0.8021583
```

```
# Calculate True Positive Rate: 69.6%
758 / (758 + 331)
```

```
## [1] 0.6960514
```

```
# Calculate False Positive Rate: 13.9%
274 / (274 + 1695)
```

```
## [1] 0.1391569
```

```
# Check Other metrics
F1_Score(ifelse(energy3$energy_burden_greater_4 == '>= 4%', 1, 0), energy3$fitted_energy_burden_greater
```

```
## [1] 0.8485607
```

```
Precision(ifelse(energy3$energy_burden_greater_4 == '>= 4%', 1, 0), energy3$fitted_energy_burden_greate
```

```
## [1] 0.8366239
```

```
Recall(ifelse(energy3$energy_burden_greater_4 == '>= 4%', 1, 0), energy3$fitted_energy_burden_greater_4
```

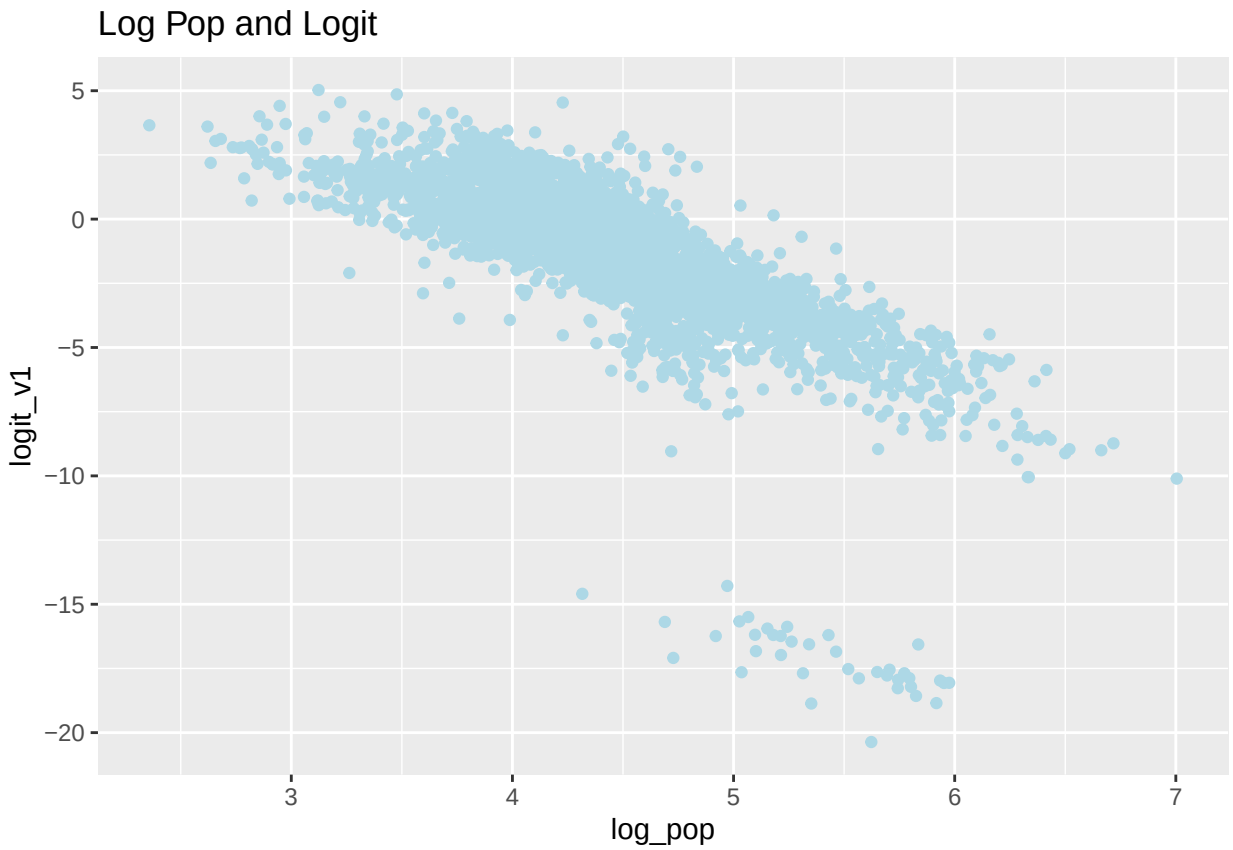
```
## [1] 0.8608431
```

```
AUC(ifelse(energy3$energy_burden_greater_4 == '>= 4%', 1, 0), energy3$fitted_energy_burden_greater_4)
```

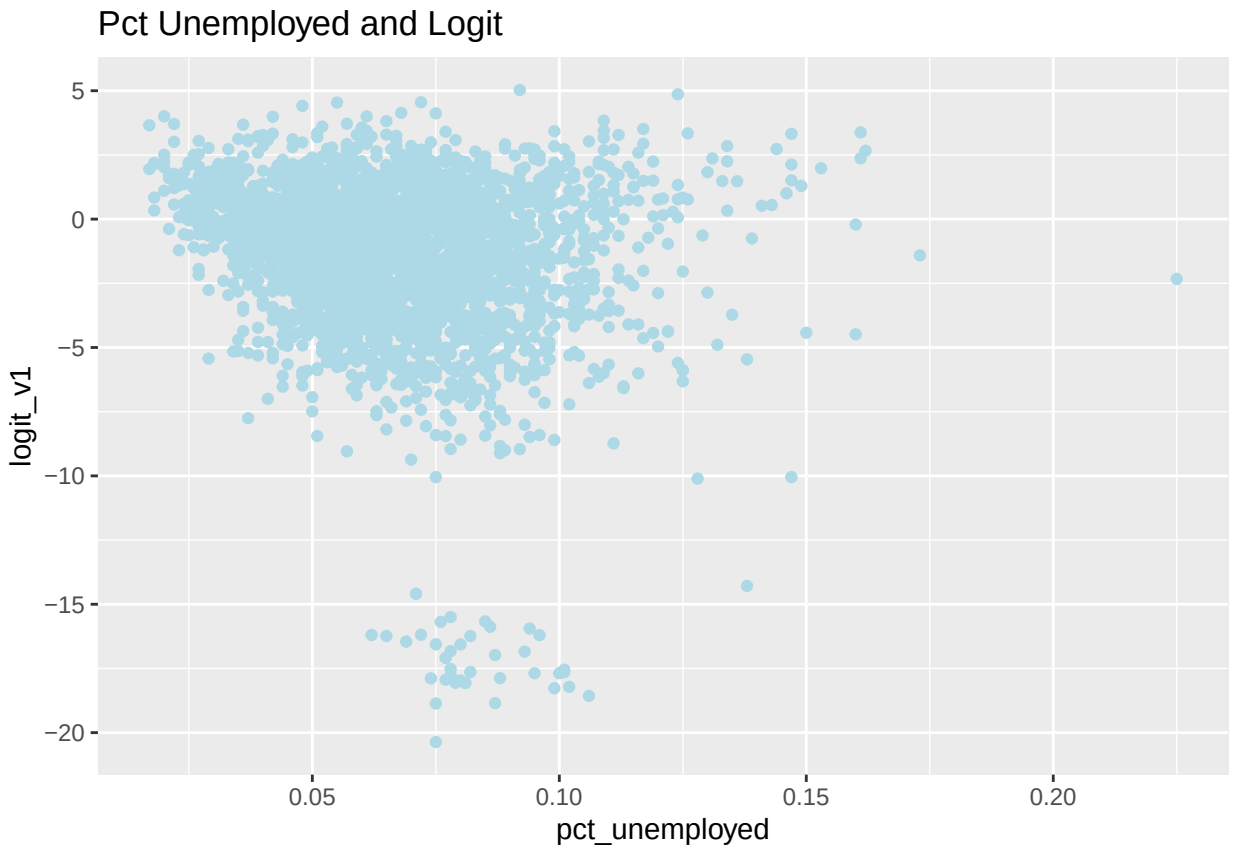
```
## [1] 0.78556
```

The metrics have largely stayed the same. Some Improvement within 0.5 - 1%.

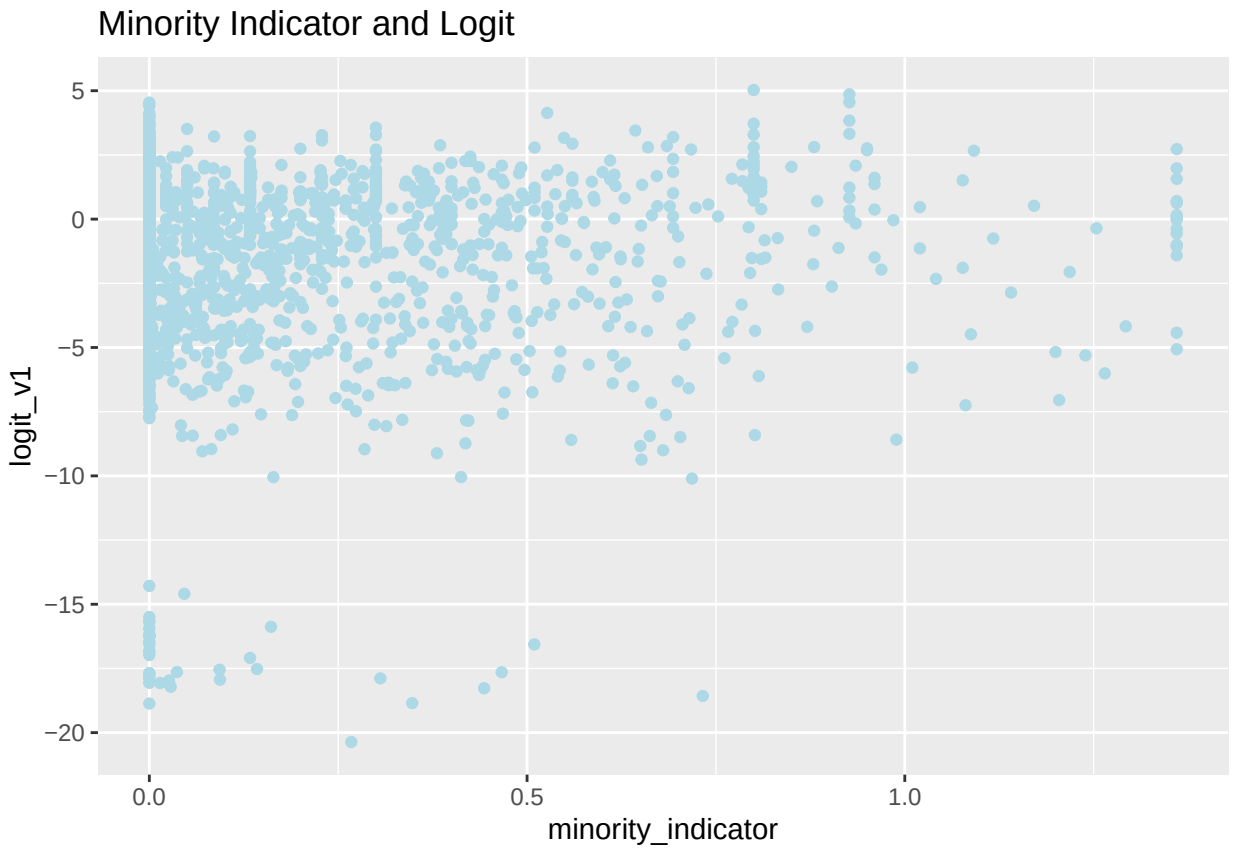
```
# Check linear relationship between logit and continuous predictors
library(ggplot2)

# Calculate the logit variable = log(p / (1-p))
energy3$logit_v1 = log(energy3$fitted_probabilities / ( 1 - energy3$fitted_probabilities))

# Graph Relationships

# Logit and log_pop
ggplot(energy3, aes(x = log_pop, y = logit_v1)) + geom_point(color = 'light blue') + ggtitle("Log Pop a
```



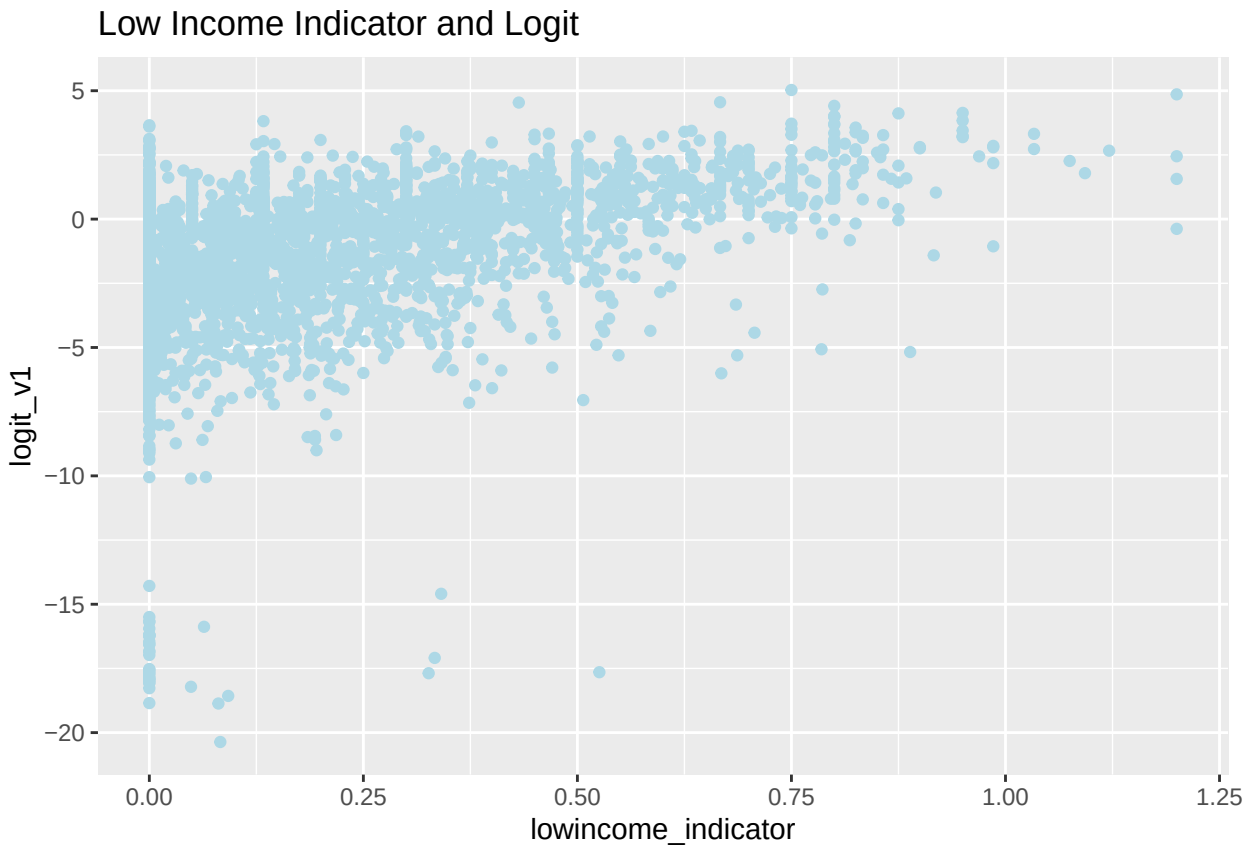

```
ggplot(energy3, aes(x = pct_unemployed, y = logit_v1)) + geom_point(color = 'light blue') + ggtitle("Pc
```



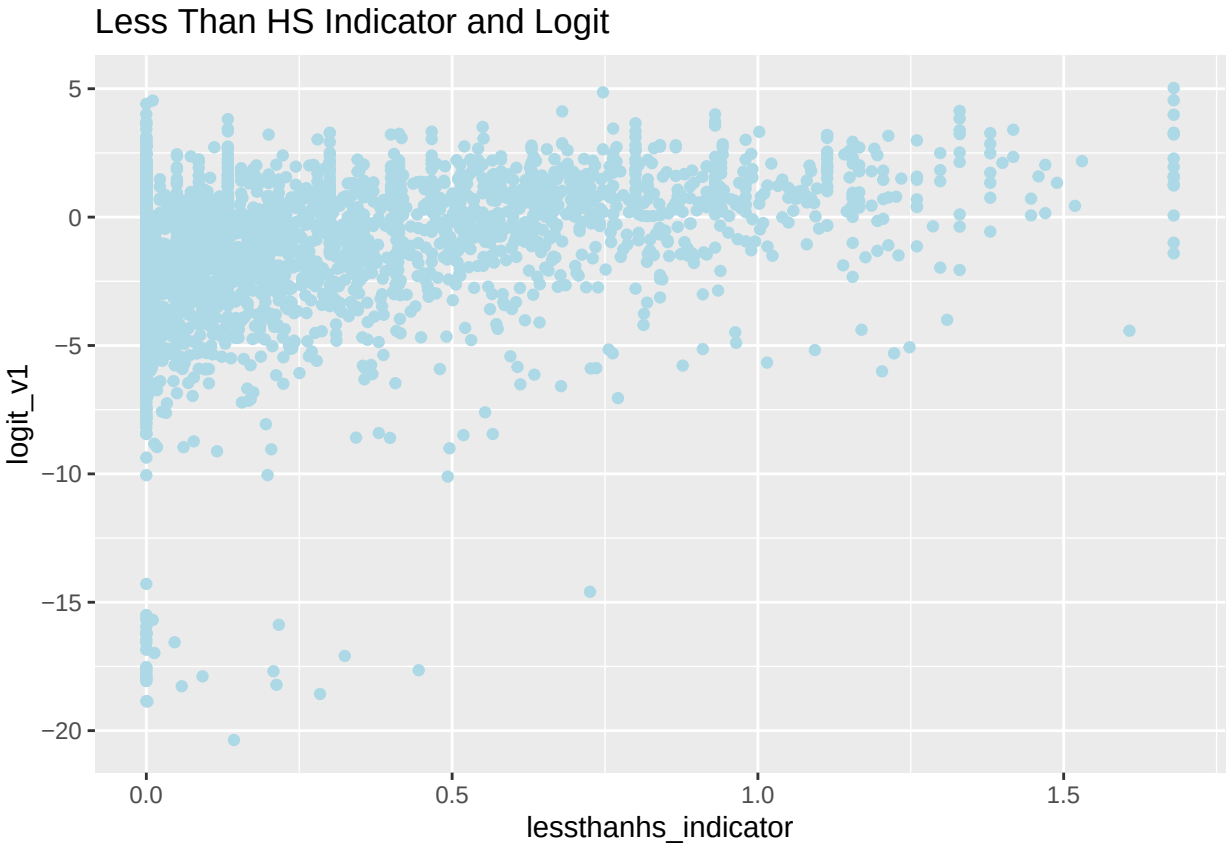

```
ggplot(energy3, aes(x = minority_indicator, y = logit_v1)) + geom_point(color = 'light blue') + ggtitle
```




```
ggplot(energy3, aes(x = lowincome_indicator, y = logit_v1)) + geom_point(color = 'light blue') + ggtitl
```




```
ggplot(energy3, aes(x = lessthanhs_indicator, y = logit_v1)) + geom_point(color = 'light blue') + ggtitl
```



The nature of relationships has largely stayed the same. Less outliers at the bottom.

Confidence Intervals and Interpretation

This is the full model

$$Odds\ of\ Energy\ Burden\ Being >= 4\% = EXP\left[Coef * I(State \neq AL) - 2.839 * LogPopulation + 18.25 * PctUnemployed + 0.08\right.$$

```
summary_log = summary(log_mod1)
```

```
# Find Confidence Intervals for odds

# Colorado odds compared to baseline - Alabama. Lower Bound of 95% confidence
exp(summary_log$coefficients[5,1] - qt(0.975, summary_log$df[2]) * summary_log$coefficients[5,2])
```

```
## [1] 0.07329038
```

```
# Colorado odds compared to baseline - Alabama. Higher Bound of 95% confidence
exp(summary_log$coefficients[5,1] + qt(0.975, summary_log$df[2]) * summary_log$coefficients[5,2])
```

```
## [1] 0.5182803
```

```
# Colorado odds
exp(summary_log$coefficients[5,1])
```

```
## [1] 0.1948973
```

Interpretation: The odds of a county having average energy burden over 4% is 81% lower (1 - 0.19 = 0.81) for a county in Colorado, compared to one in Alabama (baseline). These odds are an estimate, we are 95% sure that the true odds fall between 49% (1- 0.51) lower and 93% (1 - 0.07).

```
# Log Population odds

# 95% confidence: lower bound
exp(summary_log$coefficients[50,1] - qt(0.975, summary_log$df[2]) * summary_log$coefficients[50,2])
```

```
## [1] 0.04001228
```

```
# 95% confidence: upper bound
exp(summary_log$coefficients[50,1] + qt(0.975, summary_log$df[2]) * summary_log$coefficients[50,2])
```

```
## [1] 0.08547957
```

```
# Coefficient Odds

exp(summary_log$coefficients[50,1])
```

```
## [1] 0.05848275
```

Interpretation: If log10 of population increases by 1 (this happens if the population increases by factor of 10), then the odds of a county having average energy burden above 4% decrease by 94.1% (1 - 0.058). These odds are an estimate, we are 95% sure that the true odds fall between 91.4% (1- 0.085) and 96% (1 - 0.04) lower.

```
# Pct Unemployed Odds

# 95% confidence: lower bound
exp(summary_log$coefficients[51,1]/100 - qt(0.975, summary_log$df[2]) * summary_log$coefficients[51,2]/
```

```
## [1] 1.120516
```

```
# 95% confidence: upper bound
exp(summary_log$coefficients[51,1]/100 + qt(0.975, summary_log$df[2]) * summary_log$coefficients[51,2]/
```

```
## [1] 1.285662
```

```
# Coefficient Odds
exp(summary_log$coefficients[51,1]/100)
```

```
## [1] 1.200252
```

Interpretation: note - unemployment rate is written as decimal, not a whole number (e.g., 0.05 for 5%). If unemployment rate goes up by 0.01 (1%), then the odds of a high energy burden go up by 1.2. 95% confidence interval spans a factor of 1.12 and 1.29.

```
# Rural Urban Flag = 3

# 95% confidence: lower bound
exp(summary_log$coefficients[53,1] - qt(0.975, summary_log$df[2]) * summary_log$coefficients[53,2])
```

```
## [1] 0.8389384
```

```
# 95% confidence: upper bound
exp(summary_log$coefficients[53,1] + qt(0.975, summary_log$df[2]) * summary_log$coefficients[53,2])
```

```
## [1] 2.363444
```

```
# Coefficient Odds

exp(summary_log$coefficients[53,1])
```

```
## [1] 1.408114
```

Interpretation: Compared to a metro county with a metro population of 1 million or more (baseline, urban flag = 1), a metro county with a metro population of less than 250K residents has 40% higher odds (1.4 - 1) of having average energy burden over 4%. 95% confidence interval spans 17% lower odds and 136% higher odds.

```
# Minority Indicator

# 95% confidence: lower bound
exp(summary_log$coefficients[60,1] - qt(0.975, summary_log$df[2]) * summary_log$coefficients[60,2])
```

```
## [1] 0.09245351
```

```
# 95% confidence: upper bound
exp(summary_log$coefficients[60,1] + qt(0.975, summary_log$df[2]) * summary_log$coefficients[60,2])
```

```
## [1] 0.33229
```

```
# Coefficient Odds

exp(summary_log$coefficients[60,1])
```

```
## [1] 0.1752751
```

Interpretation: When minority indicator goes up by 1 (score based on the weighted average of percent of census tracts within 60-80th and 80-100th national percentiles), then the odds of having higher average energy burden go down by 83% (1 - 0.17). Note: this is inconsistent with prior findings. 95% confidence interval includes 67% lower odds and 91% lower odds.

```
# Low Income Indicator

# 95% confidence: lower bound
exp(summary_log$coefficients[61,1] - qt(0.975, summary_log$df[2]) * summary_log$coefficients[61,2])
```

```
## [1] 4.090856
```

```
# 95% confidence: upper bound
exp(summary_log$coefficients[61,1] + qt(0.975, summary_log$df[2]) * summary_log$coefficients[61,2])
```

```
## [1] 15.25477
```

```
# Coefficient Odds

exp(summary_log$coefficients[61,1])
```

```
## [1] 7.899689
```

Interpretation: When low income indicator goes up by 1, the odds of the county having energy burden over 4% go up by a factor of 7.9. 95% confidence interval spans factor of 4 to factor of 15.

```
# Less than HS Indicator

# 95% confidence: lower bound
exp(summary_log$coefficients[62,1] - qt(0.975, summary_log$df[2]) * summary_log$coefficients[62,2])
```

```
## [1] 1.1091
```

```
# 95% confidence: upper bound
exp(summary_log$coefficients[62,1] + qt(0.975, summary_log$df[2]) * summary_log$coefficients[62,2])
```

```
## [1] 2.611153
```

```r
# Coefficient Odds

exp(summary_log$coefficients[62,1])
```

```
## [1] 1.701772
```

Interpretation: When less than high school indicator goes up by 1, the odds of the county having energy burden over 4% go up by a factor of 1.7. The 95% confidence interval spans factor 1.1 to 2.6.