A4

James Tran

April 4, 2018

## First review the assignment.

This assignment is to produce a predictive model of median household income using other variables in the countyComplete dataframe, which is in the openintro package. You may not include per capita income. I have included one model for discussion, but you need to create a model with different choices. You may construct new variables based on those included in countyComplete.

## Load the required libraries.

library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1 v purrr 0.2.4  
## v tibble 1.4.2 v dplyr 0.7.4  
## v tidyr 0.7.2 v stringr 1.2.0  
## v readr 1.1.1 v forcats 0.2.0

## Warning: package 'ggplot2' was built under R version 3.4.4

## -- Conflicts ---------------------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(openintro)

## Please visit openintro.org for free statistics materials

##   
## Attaching package: 'openintro'

## The following object is masked from 'package:ggplot2':  
##   
## diamonds

## The following objects are masked from 'package:datasets':  
##   
## cars, trees

library(broom)

## Problem 1

Run the commands glimpse() and summary() on the dataframe to understand the meaning of the variables and to verify the integrity of the data. The documentation in the openintro package is useful for understanding the variables. Avoid variables with many missing values when you construct your model.

glimpse(countyComplete)

## Observations: 3,143  
## Variables: 53  
## $ state <fct> Alabama, Alabama, Al...  
## $ name <fct> Autauga County, Bald...  
## $ FIPS <dbl> 1001, 1003, 1005, 10...  
## $ pop2010 <dbl> 54571, 182265, 27457...  
## $ pop2000 <dbl> 43671, 140415, 29038...  
## $ age\_under\_5 <dbl> 6.6, 6.1, 6.2, 6.0, ...  
## $ age\_under\_18 <dbl> 26.8, 23.0, 21.9, 22...  
## $ age\_over\_65 <dbl> 12.0, 16.8, 14.2, 12...  
## $ female <dbl> 51.3, 51.1, 46.9, 46...  
## $ white <dbl> 78.5, 85.7, 48.0, 75...  
## $ black <dbl> 17.7, 9.4, 46.9, 22....  
## $ native <dbl> 0.4, 0.7, 0.4, 0.3, ...  
## $ asian <dbl> 0.9, 0.7, 0.4, 0.1, ...  
## $ pac\_isl <dbl> NA, NA, NA, NA, NA, ...  
## $ two\_plus\_races <dbl> 1.6, 1.5, 0.9, 0.9, ...  
## $ hispanic <dbl> 2.4, 4.4, 5.1, 1.8, ...  
## $ white\_not\_hispanic <dbl> 77.2, 83.5, 46.8, 75...  
## $ no\_move\_in\_one\_plus\_year <dbl> 86.3, 83.0, 83.0, 90...  
## $ foreign\_born <dbl> 2.0, 3.6, 2.8, 0.7, ...  
## $ foreign\_spoken\_at\_home <dbl> 3.7, 5.5, 4.7, 1.5, ...  
## $ hs\_grad <dbl> 85.3, 87.6, 71.9, 74...  
## $ bachelors <dbl> 21.7, 26.8, 13.5, 10...  
## $ veterans <dbl> 5817, 20396, 2327, 1...  
## $ mean\_work\_travel <dbl> 25.1, 25.8, 23.8, 28...  
## $ housing\_units <dbl> 22135, 104061, 11829...  
## $ home\_ownership <dbl> 77.5, 76.7, 68.0, 82...  
## $ housing\_multi\_unit <dbl> 7.2, 22.6, 11.1, 6.6...  
## $ median\_val\_owner\_occupied <dbl> 133900, 177200, 8820...  
## $ households <dbl> 19718, 69476, 9795, ...  
## $ persons\_per\_household <dbl> 2.70, 2.50, 2.52, 3....  
## $ per\_capita\_income <dbl> 24568, 26469, 15875,...  
## $ median\_household\_income <dbl> 53255, 50147, 33219,...  
## $ poverty <dbl> 10.6, 12.2, 25.0, 12...  
## $ private\_nonfarm\_establishments <dbl> 877, 4812, 522, 318,...  
## $ private\_nonfarm\_employment <dbl> 10628, 52233, 7990, ...  
## $ percent\_change\_private\_nonfarm\_employment <dbl> 16.6, 17.4, -27.0, -...  
## $ nonemployment\_establishments <dbl> 2971, 14175, 1527, 1...  
## $ firms <dbl> 4067, 19035, 1667, 1...  
## $ black\_owned\_firms <dbl> 15.2, 2.7, NA, 14.9,...  
## $ native\_owned\_firms <dbl> NA, 0.4, NA, NA, NA,...  
## $ asian\_owned\_firms <dbl> 1.3, 1.0, NA, NA, NA...  
## $ pac\_isl\_owned\_firms <dbl> NA, NA, NA, NA, NA, ...  
## $ hispanic\_owned\_firms <dbl> 0.7, 1.3, NA, NA, NA...  
## $ women\_owned\_firms <dbl> 31.7, 27.3, 27.0, NA...  
## $ manufacturer\_shipments\_2007 <dbl> NA, 1410273, NA, 0, ...  
## $ mercent\_whole\_sales\_2007 <dbl> NA, NA, NA, NA, NA, ...  
## $ sales <dbl> 598175, 2966489, 188...  
## $ sales\_per\_capita <dbl> 12003, 17166, 6334, ...  
## $ accommodation\_food\_service <dbl> 88157, 436955, NA, 1...  
## $ building\_permits <dbl> 191, 696, 10, 8, 18,...  
## $ fed\_spending <dbl> 331142, 1119082, 240...  
## $ area <dbl> 594.44, 1589.78, 884...  
## $ density <dbl> 91.8, 114.6, 31.0, 3...

summary(countyComplete)

## state name FIPS   
## Texas : 254 Washington County: 30 Min. : 1001   
## Georgia : 159 Jefferson County : 25 1st Qu.:18178   
## Virginia: 134 Franklin County : 24 Median :29177   
## Kentucky: 120 Jackson County : 23 Mean :30390   
## Missouri: 115 Lincoln County : 23 3rd Qu.:45082   
## Kansas : 105 Madison County : 19 Max. :56045   
## (Other) :2256 (Other) :2999   
## pop2010 pop2000 age\_under\_5 age\_under\_18   
## Min. : 82 Min. : 67 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 11104 1st Qu.: 11210 1st Qu.: 5.500 1st Qu.:21.40   
## Median : 25857 Median : 24608 Median : 6.200 Median :23.30   
## Mean : 98233 Mean : 89623 Mean : 6.261 Mean :23.42   
## 3rd Qu.: 66699 3rd Qu.: 61766 3rd Qu.: 6.800 3rd Qu.:25.10   
## Max. :9818605 Max. :9519338 Max. :12.600 Max. :41.60   
## NA's :3   
## age\_over\_65 female white black   
## Min. : 3.50 Min. :27.90 Min. : 2.70 Min. : 0.000   
## 1st Qu.:13.10 1st Qu.:49.60 1st Qu.:75.25 1st Qu.: 0.500   
## Median :15.60 Median :50.50 Median :89.10 Median : 2.000   
## Mean :15.88 Mean :50.03 Mean :82.89 Mean : 8.931   
## 3rd Qu.:18.20 3rd Qu.:51.10 3rd Qu.:95.50 3rd Qu.:10.200   
## Max. :43.40 Max. :56.80 Max. :99.20 Max. :85.700   
## NA's :17   
## native asian pac\_isl two\_plus\_races   
## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.100   
## 1st Qu.: 0.200 1st Qu.: 0.300 1st Qu.: 0.000 1st Qu.: 1.100   
## Median : 0.400 Median : 0.500 Median : 0.000 Median : 1.600   
## Mean : 2.026 Mean : 1.166 Mean : 0.154 Mean : 1.976   
## 3rd Qu.: 0.800 3rd Qu.: 1.000 3rd Qu.: 0.100 3rd Qu.: 2.300   
## Max. :96.000 Max. :43.900 Max. :48.900 Max. :29.500   
## NA's :4 NA's :24 NA's :1697   
## hispanic white\_not\_hispanic no\_move\_in\_one\_plus\_year  
## Min. : 0.000 Min. : 2.70 Min. : 51.6   
## 1st Qu.: 1.600 1st Qu.:66.95 1st Qu.: 83.2   
## Median : 3.300 Median :85.80 Median : 86.3   
## Mean : 8.284 Mean :78.29 Mean : 85.8   
## 3rd Qu.: 8.200 3rd Qu.:94.20 3rd Qu.: 89.0   
## Max. :95.700 Max. :99.20 Max. :100.0   
##   
## foreign\_born foreign\_spoken\_at\_home hs\_grad bachelors   
## Min. : 0.000 Min. : 0.000 Min. :47.90 Min. : 3.70   
## 1st Qu.: 1.200 1st Qu.: 2.800 1st Qu.:78.40 1st Qu.:13.10   
## Median : 2.400 Median : 4.800 Median :84.60 Median :16.90   
## Mean : 4.372 Mean : 9.057 Mean :83.11 Mean :19.03   
## 3rd Qu.: 5.300 3rd Qu.:10.000 3rd Qu.:88.60 3rd Qu.:22.60   
## Max. :72.200 Max. :96.000 Max. :99.30 Max. :71.00   
##   
## veterans mean\_work\_travel housing\_units home\_ownership   
## Min. : 0 Min. : 4.30 Min. : 50 Min. : 0.00   
## 1st Qu.: 958 1st Qu.:19.00 1st Qu.: 5416 1st Qu.:69.50   
## Median : 2180 Median :22.40 Median : 12162 Median :74.60   
## Mean : 7207 Mean :22.73 Mean : 41904 Mean :73.26   
## 3rd Qu.: 5944 3rd Qu.:26.10 3rd Qu.: 30574 3rd Qu.:78.40   
## Max. :368128 Max. :44.20 Max. :3445076 Max. :91.30   
##   
## housing\_multi\_unit median\_val\_owner\_occupied households   
## Min. : 0.00 Min. : 0 Min. : 22   
## 1st Qu.: 6.10 1st Qu.: 80200 1st Qu.: 4260   
## Median : 9.70 Median : 105900 Median : 9868   
## Mean :12.33 Mean : 132545 Mean : 36346   
## 3rd Qu.:15.90 3rd Qu.: 152950 3rd Qu.: 25358   
## Max. :98.50 Max. :1000001 Max. :3217889   
##   
## persons\_per\_household per\_capita\_income median\_household\_income  
## Min. :1.100 Min. : 7772 Min. : 19351   
## 1st Qu.:2.370 1st Qu.:19030 1st Qu.: 36952   
## Median :2.490 Median :21773 Median : 42445   
## Mean :2.513 Mean :22505 Mean : 44270   
## 3rd Qu.:2.630 3rd Qu.:24814 3rd Qu.: 49142   
## Max. :4.470 Max. :64381 Max. :115574   
##   
## poverty private\_nonfarm\_establishments private\_nonfarm\_employment  
## Min. : 0.0 Min. : 0 Min. : 0   
## 1st Qu.:11.0 1st Qu.: 229 1st Qu.: 2109   
## Median :14.7 Median : 551 Median : 6351   
## Mean :15.5 Mean : 2362 Mean : 35656   
## 3rd Qu.:19.0 3rd Qu.: 1484 3rd Qu.: 19436   
## Max. :53.5 Max. :245523 Max. :3703233   
##   
## percent\_change\_private\_nonfarm\_employment nonemployment\_establishments  
## Min. :-83.2000 Min. : 21   
## 1st Qu.:-12.0000 1st Qu.: 729   
## Median : -2.0000 Median : 1594   
## Mean : 0.5338 Mean : 6720   
## 3rd Qu.: 9.8000 3rd Qu.: 4130   
## Max. :386.5000 Max. :821177   
## NA's :67 NA's :5   
## firms black\_owned\_firms native\_owned\_firms asian\_owned\_firms  
## Min. : 27 Min. : 0.200 Min. : 0.200 Min. : 0.300   
## 1st Qu.: 1074 1st Qu.: 2.100 1st Qu.: 0.525 1st Qu.: 1.400   
## Median : 2350 Median : 5.700 Median : 0.900 Median : 2.200   
## Mean : 9301 Mean : 9.806 Mean : 3.785 Mean : 3.422   
## 3rd Qu.: 6034 3rd Qu.:13.350 3rd Qu.: 2.300 3rd Qu.: 3.700   
## Max. :1046940 Max. :66.700 Max. :71.800 Max. :56.600   
## NA's :176 NA's :2376 NA's :2653 NA's :2408   
## pac\_isl\_owned\_firms hispanic\_owned\_firms women\_owned\_firms  
## Min. : 0.0000 Min. : 0.300 Min. : 6.50   
## 1st Qu.: 0.1000 1st Qu.: 1.400 1st Qu.:22.70   
## Median : 0.1000 Median : 2.800 Median :26.20   
## Mean : 0.7171 Mean : 6.811 Mean :25.96   
## 3rd Qu.: 0.3000 3rd Qu.: 6.800 3rd Qu.:29.20   
## Max. :10.5000 Max. :78.000 Max. :56.20   
## NA's :3073 NA's :2363 NA's :970   
## manufacturer\_shipments\_2007 mercent\_whole\_sales\_2007 sales   
## Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 0 1st Qu.: 42125 1st Qu.: 79988   
## Median : 238180 Median : 138930 Median : 257667   
## Mean : 1680613 Mean : 1794262 Mean : 1262270   
## 3rd Qu.: 1161878 3rd Qu.: 562056 3rd Qu.: 791435   
## Max. :169275136 Max. :205478751 Max. :119111840   
## NA's :488 NA's :1022 NA's :42   
## sales\_per\_capita accommodation\_food\_service building\_permits  
## Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 6993 1st Qu.: 9349 1st Qu.: 5   
## Median : 9793 Median : 31065 Median : 32   
## Mean :10375 Mean : 211181 Mean : 192   
## 3rd Qu.:12980 3rd Qu.: 110695 3rd Qu.: 123   
## Max. :80800 Max. :24857836 Max. :15039   
## NA's :42 NA's :272   
## fed\_spending area density   
## Min. : 0 Min. : 2.0 Min. : 0.0   
## 1st Qu.: 102922 1st Qu.: 430.7 1st Qu.: 16.9   
## Median : 214994 Median : 615.6 Median : 45.2   
## Mean : 944376 Mean : 1123.7 Mean : 259.3   
## 3rd Qu.: 522228 3rd Qu.: 924.0 3rd Qu.: 113.8   
## Max. :80457156 Max. :145504.8 Max. :69467.5   
## NA's :4

## Problem 2

Use lm() to create a model to predict median household income using 5 other variables. Do not include per capita income. Display a summary of the model.

Model = lm(median\_household\_income ~ median\_val\_owner\_occupied + age\_under\_5 +bachelors + poverty + persons\_per\_household, data=countyComplete )  
summary(Model)

##   
## Call:  
## lm(formula = median\_household\_income ~ median\_val\_owner\_occupied +   
## age\_under\_5 + bachelors + poverty + persons\_per\_household,   
## data = countyComplete)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21032 -2410 -133 2090 34408   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.722e+04 8.798e+02 19.573 < 2e-16 \*\*\*  
## median\_val\_owner\_occupied 3.313e-02 1.338e-03 24.763 < 2e-16 \*\*\*  
## age\_under\_5 4.361e+02 8.902e+01 4.899 1.01e-06 \*\*\*  
## bachelors 4.003e+02 1.336e+01 29.972 < 2e-16 \*\*\*  
## poverty -1.060e+03 1.439e+01 -73.660 < 2e-16 \*\*\*  
## persons\_per\_household 1.144e+04 4.358e+02 26.244 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4463 on 3137 degrees of freedom  
## Multiple R-squared: 0.8508, Adjusted R-squared: 0.8506   
## F-statistic: 3579 on 5 and 3137 DF, p-value: < 2.2e-16

## Problem 3

Which 2 numbers in the summary output describe the overall performance of the model. Use these two numbers in appropriate sentences to describe how well your model performed.

The two numbers that discribe the overall performance of the model is the Residual Standard Error and R-Squared. The Residual Standard error within Model is: 4463 Multiple R-squared is: 0.85 Ajusted R-squared is: 0.8506

## Problem 4

Examine the p-values for the individual coefficients of the model. Can you reject the hypothesis that the true coefficient value is zero in every case?

Probablity if the true coefficient is zero,we are observing the probability of zero. The P values of every coefficients is close to 0.

## Problem 5

Look at the signs of the coefficients. Do all of them have the signs that you would expect? Note any exceptions.

From the slected variables, they will positively impact the median household income. The negative value of poverty shows it has a negative impact on household income, as poverty goes up the median houshold income will decreases.

## Problem 6

Consider the forecasts of median household income for three different countys. Choose the counties you want. Use the augment() function from broom. Describe how the forecasts compare with the actuals for these counties.

am = augment(Model, data= countyComplete)  
str(am)

## 'data.frame': 3143 obs. of 60 variables:  
## $ state : Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ name : Factor w/ 1877 levels "Abbeville County",..: 83 90 101 151 166 227 237 250 298 320 ...  
## $ FIPS : num 1001 1003 1005 1007 1009 ...  
## $ pop2010 : num 54571 182265 27457 22915 57322 ...  
## $ pop2000 : num 43671 140415 29038 20826 51024 ...  
## $ age\_under\_5 : num 6.6 6.1 6.2 6 6.3 6.8 6.5 6.1 5.7 5.3 ...  
## $ age\_under\_18 : num 26.8 23 21.9 22.7 24.6 22.3 24.1 22.9 22.5 21.4 ...  
## $ age\_over\_65 : num 12 16.8 14.2 12.7 14.7 13.5 16.7 14.3 16.7 17.9 ...  
## $ female : num 51.3 51.1 46.9 46.3 50.5 45.8 53 51.8 52.2 50.4 ...  
## $ white : num 78.5 85.7 48 75.8 92.6 23 54.4 74.9 58.8 92.7 ...  
## $ black : num 17.7 9.4 46.9 22 1.3 70.2 43.4 20.6 38.7 4.6 ...  
## $ native : num 0.4 0.7 0.4 0.3 0.5 0.2 0.3 0.5 0.2 0.5 ...  
## $ asian : num 0.9 0.7 0.4 0.1 0.2 0.2 0.8 0.7 0.5 0.2 ...  
## $ pac\_isl : num NA NA NA NA NA NA 0 0.1 0 0 ...  
## $ two\_plus\_races : num 1.6 1.5 0.9 0.9 1.2 0.8 0.8 1.7 1.1 1.5 ...  
## $ hispanic : num 2.4 4.4 5.1 1.8 8.1 7.1 0.9 3.3 1.6 1.2 ...  
## $ white\_not\_hispanic : num 77.2 83.5 46.8 75 88.9 21.9 54.1 73.6 58.1 92.1 ...  
## $ no\_move\_in\_one\_plus\_year : num 86.3 83 83 90.5 87.2 88.5 92.8 82.9 86.2 88.1 ...  
## $ foreign\_born : num 2 3.6 2.8 0.7 4.7 1.1 1.1 2.5 0.9 0.5 ...  
## $ foreign\_spoken\_at\_home : num 3.7 5.5 4.7 1.5 7.2 3.8 1.6 4.5 1.6 1.4 ...  
## $ hs\_grad : num 85.3 87.6 71.9 74.5 74.7 74.7 74.8 78.5 71.8 73.4 ...  
## $ bachelors : num 21.7 26.8 13.5 10 12.5 12 11 16.1 10.8 10.5 ...  
## $ veterans : num 5817 20396 2327 1883 4072 ...  
## $ mean\_work\_travel : num 25.1 25.8 23.8 28.3 33.2 28.1 25.1 22.1 23.6 26.2 ...  
## $ housing\_units : num 22135 104061 11829 8981 23887 ...  
## $ home\_ownership : num 77.5 76.7 68 82.9 82 76.9 69 70.7 71.4 77.5 ...  
## $ housing\_multi\_unit : num 7.2 22.6 11.1 6.6 3.7 9.9 13.7 14.3 8.7 4.3 ...  
## $ median\_val\_owner\_occupied : num 133900 177200 88200 81200 113700 ...  
## $ households : num 19718 69476 9795 7441 20605 ...  
## $ persons\_per\_household : num 2.7 2.5 2.52 3.02 2.73 2.85 2.58 2.46 2.51 2.22 ...  
## $ per\_capita\_income : num 24568 26469 15875 19918 21070 ...  
## $ median\_household\_income : num 53255 50147 33219 41770 45549 ...  
## $ poverty : num 10.6 12.2 25 12.6 13.4 25.3 25 19.5 20.3 17.6 ...  
## $ private\_nonfarm\_establishments : num 877 4812 522 318 749 ...  
## $ private\_nonfarm\_employment : num 10628 52233 7990 2927 6968 ...  
## $ percent\_change\_private\_nonfarm\_employment: num 16.6 17.4 -27 -14 -11.4 -18.5 2.1 -5.6 -45.8 5.4 ...  
## $ nonemployment\_establishments : num 2971 14175 1527 1192 3501 ...  
## $ firms : num 4067 19035 1667 1385 4458 ...  
## $ black\_owned\_firms : num 15.2 2.7 NA 14.9 NA NA NA 7.2 NA NA ...  
## $ native\_owned\_firms : num NA 0.4 NA NA NA NA NA NA NA NA ...  
## $ asian\_owned\_firms : num 1.3 1 NA NA NA NA 3.3 1.6 NA NA ...  
## $ pac\_isl\_owned\_firms : num NA NA NA NA NA NA NA NA NA NA ...  
## $ hispanic\_owned\_firms : num 0.7 1.3 NA NA NA NA NA 0.5 NA NA ...  
## $ women\_owned\_firms : num 31.7 27.3 27 NA 23.2 38.8 NA 24.7 29.3 14.5 ...  
## $ manufacturer\_shipments\_2007 : num NA 1410273 NA 0 341544 ...  
## $ mercent\_whole\_sales\_2007 : num NA NA NA NA NA ...  
## $ sales : num 598175 2966489 188337 124707 319700 ...  
## $ sales\_per\_capita : num 12003 17166 6334 5804 5622 ...  
## $ accommodation\_food\_service : num 88157 436955 NA 10757 20941 ...  
## $ building\_permits : num 191 696 10 8 18 1 3 107 10 6 ...  
## $ fed\_spending : num 331142 1119082 240308 163201 294114 ...  
## $ area : num 594 1590 885 623 645 ...  
## $ density : num 91.8 114.6 31 36.8 88.9 ...  
## $ .fitted : num 52863 52137 30568 47711 45755 ...  
## $ .se.fit : num 140 110 147 276 145 ...  
## $ .resid : num 392 -1990 2651 -5941 -206 ...  
## $ .hat : num 0.000978 0.000609 0.001083 0.003819 0.001062 ...  
## $ .sigma : num 4464 4464 4464 4463 4464 ...  
## $ .cooksd : num 1.26e-06 2.02e-05 6.38e-05 1.14e-03 3.76e-07 ...  
## $ .std.resid : num 0.088 -0.4461 0.5943 -1.3336 -0.0461 ...

View(am)  
 am[c("2743", "2981", "2584"),]

## state name FIPS pop2010 pop2000 age\_under\_5  
## 2743 Texas Tarrant County 48439 1809034 1446219 7.9  
## 2981 Washington Pierce County 53053 795225 700820 7.0  
## 2584 Texas Denton County 48121 662614 432976 7.5  
## age\_under\_18 age\_over\_65 female white black native asian pac\_isl  
## 2743 28.0 8.9 51.0 66.6 14.9 0.7 4.7 0.2  
## 2981 24.9 11.0 50.6 74.2 6.8 1.4 6.0 1.3  
## 2584 27.5 6.9 50.8 75.0 8.4 0.7 6.6 0.1  
## two\_plus\_races hispanic white\_not\_hispanic no\_move\_in\_one\_plus\_year  
## 2743 3.0 26.7 51.8 80.9  
## 2981 6.8 9.2 70.3 80.2  
## 2584 2.9 18.2 64.4 78.5  
## foreign\_born foreign\_spoken\_at\_home hs\_grad bachelors veterans  
## 2743 15.6 26.8 83.8 28.7 121190  
## 2981 9.4 13.7 89.8 23.4 87888  
## 2584 13.6 21.0 91.2 39.5 39355  
## mean\_work\_travel housing\_units home\_ownership housing\_multi\_unit  
## 2743 25.7 714803 63.4 27.7  
## 2981 28.4 325375 63.3 24.8  
## 2584 27.5 256139 66.2 27.0  
## median\_val\_owner\_occupied households persons\_per\_household  
## 2743 134900 632518 2.72  
## 2981 269300 295554 2.57  
## 2584 178300 224840 2.73  
## per\_capita\_income median\_household\_income poverty  
## 2743 27333 55306 13.4  
## 2981 27446 57869 11.6  
## 2584 32538 70622 8.0  
## private\_nonfarm\_establishments private\_nonfarm\_employment  
## 2743 36935 677390  
## 2981 16768 228905  
## 2584 11225 157762  
## percent\_change\_private\_nonfarm\_employment  
## 2743 5.7  
## 2981 9.9  
## 2584 42.6  
## nonemployment\_establishments firms black\_owned\_firms  
## 2743 133171 158104 10.3  
## 2981 37592 53296 2.9  
## 2584 50692 58767 5.3  
## native\_owned\_firms asian\_owned\_firms pac\_isl\_owned\_firms  
## 2743 0.7 6.1 NA  
## 2981 NA 7.9 NA  
## 2584 0.6 6.2 NA  
## hispanic\_owned\_firms women\_owned\_firms manufacturer\_shipments\_2007  
## 2743 11.8 30.0 43337529  
## 2981 3.5 29.4 4958643  
## 2584 8.6 30.1 4008259  
## mercent\_whole\_sales\_2007 sales sales\_per\_capita  
## 2743 25801512 24931407 14582  
## 2981 6090535 9741617 12585  
## 2584 12358953 7668326 12465  
## accommodation\_food\_service building\_permits fed\_spending area  
## 2743 3763516 5089 8195152 863.61  
## 2981 1172468 1900 10024949 1669.51  
## 2584 797850 2096 1914745 878.43  
## density .fitted .se.fit .resid .hat .sigma .cooksd  
## 2743 2094.7 53525.34 189.9705 1780.655 0.001811591 4463.898 4.823155e-05  
## 2981 476.3 55656.53 178.6080 2212.470 0.001601362 4463.837 6.579192e-05  
## 2584 754.3 64950.88 277.7841 5671.116 0.003873491 4462.859 1.050382e-03  
## .std.resid  
## 2743 0.3993166  
## 2981 0.4961000  
## 2584 1.2730785

The median\_household\_income and fitted values for the selected countys are very close and their difference is less than Residual standard error.

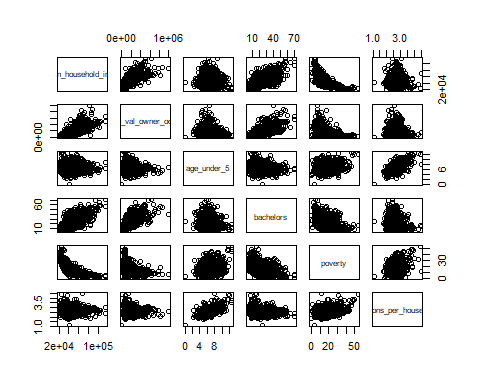
## Problem 7

First get a more convenient datframe, myvars, with only the variables you used.

# Insert your code here.  
  
myvars = countyComplete %>%  
 select(median\_household\_income,median\_val\_owner\_occupied, age\_under\_5,bachelors, poverty, persons\_per\_household)

Now look at the relationships between the pairs of variables in myvars.

plot(myvars)



cor(myvars)

## median\_household\_income  
## median\_household\_income 1.00000000  
## median\_val\_owner\_occupied 0.70773560  
## age\_under\_5 0.04925028  
## bachelors 0.68409243  
## poverty -0.74162718  
## persons\_per\_household 0.11749759  
## median\_val\_owner\_occupied age\_under\_5  
## median\_household\_income 0.70773560 0.04925028  
## median\_val\_owner\_occupied 1.00000000 -0.07236442  
## age\_under\_5 -0.07236442 1.00000000  
## bachelors 0.68867305 -0.07469412  
## poverty -0.39492737 0.18493018  
## persons\_per\_household 0.07714010 0.62080199  
## bachelors poverty persons\_per\_household  
## median\_household\_income 0.68409243 -0.7416272 0.1174976  
## median\_val\_owner\_occupied 0.68867305 -0.3949274 0.0771401  
## age\_under\_5 -0.07469412 0.1849302 0.6208020  
## bachelors 1.00000000 -0.4128962 -0.1186152  
## poverty -0.41289620 1.0000000 0.2428883  
## persons\_per\_household -0.11861518 0.2428883 1.0000000

What do you see that might suggest a reformulation of your model. Considering Age\_under\_5 and person per household to compute the relation with Median house hold income might not contribute greatly.

## Create the new model and produce a summary. Does the new model work better?

Model = lm(median\_household\_income ~ median\_val\_owner\_occupied + women\_owned\_firms + bachelors + poverty + persons\_per\_household, data=countyComplete )  
summary(Model)

##   
## Call:  
## lm(formula = median\_household\_income ~ median\_val\_owner\_occupied +   
## women\_owned\_firms + bachelors + poverty + persons\_per\_household,   
## data = countyComplete)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23937 -2412 -210 2057 25365   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.260e+04 1.118e+03 11.274 < 2e-16 \*\*\*  
## median\_val\_owner\_occupied 2.648e-02 1.393e-03 19.014 < 2e-16 \*\*\*  
## women\_owned\_firms 1.013e+02 1.707e+01 5.935 3.42e-09 \*\*\*  
## bachelors 4.269e+02 1.429e+01 29.879 < 2e-16 \*\*\*  
## poverty -1.165e+03 1.723e+01 -67.639 < 2e-16 \*\*\*  
## persons\_per\_household 1.416e+04 4.170e+02 33.954 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4280 on 2167 degrees of freedom  
## (970 observations deleted due to missingness)  
## Multiple R-squared: 0.8815, Adjusted R-squared: 0.8812   
## F-statistic: 3223 on 5 and 2167 DF, p-value: < 2.2e-16

## We can see see that median house hold income increases as the contribution of women in every household.