Final Project

Summer and Ben

5/12/2022

Introduction

Reference to Original Data Source

The dataset we will use throughout this report for our "client" comes from openintro, which is an R package for data and supplemental functions for openintro resources. Specifically, the package holds many datasets about a wide range of topics. A list of available datasets can be found here: https://www.openintro.org/data/. In addition, openintro includes open-source textbooks and resources for introductory statistics. More on the non-dataset resources in openintro can be found here: https://www.openintro.org/.

Our dataset is titled "county_complete" and contains economic information about every county in the United States. The dataset is relatively large and gathers data from a variety of sources, including: the American Community survey, tidycensus (another R package), the Census Bureau, United States Department of Agriculture, and the Bureau of Labor Statistics. Through this range of sources, the county_complete dataset includes vast amounts of data spanning more than ten years. For our purposes, we will scale down the data to the variables taken from the American Community Survey's five-year average for the years 2015-2019.

From this point forth, we may refer to our dataset as county2019. There are 3142 observations corresponding to every county in the United States, which includes about 100 boroughs and regions not usually categorized as counties.

Note to Client

You have provided us with an economic data set that contains 47 demographic variables about every county in the United States in the year 2019. Overall, we will aim to use statistical models in order to show the varying importance of these variables in predicting the mean household income in a county. For example, we might use a linear model to show the relationship between the average age of those in a county and the mean household income of those in a county.

We understand that you have been given a grant by the government and were asked to feed money into certain institutions in such a way that is most likely to increase the mean household income for a given county. Throughout this report we will present our findings to you based on the statistical models we choose to adopt. If you would like deeper explanations on the models and how they are created, you may look at the code file where the model is created and explained using coding comments.

Sparse and Smooth Linear Models

Ridge Regression

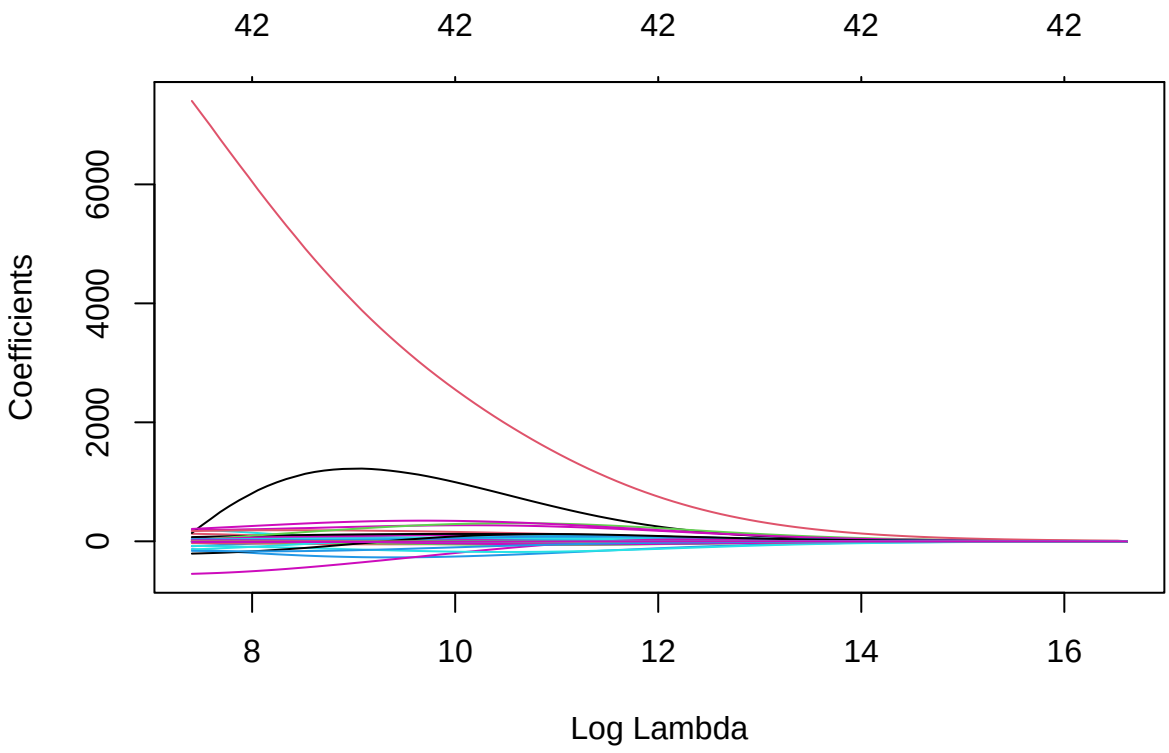
As a reminder, we are looking at 47 demographic variables to try and predict mean household income. We will refer to the variables we use to predict income as predictor variables.

Ridge regression is a model that is used for data sets that contain many predictor variables. Specifically, if there are a great deal of factors such as average age in a county, poverty level of a county, percent of firms that are women-owned in a county, etc., ridge regression is the kind of model that would simplify how these the predictor variables impact the model. For clarity, "many" factors would be hundreds or thousands of factors. Ridge regression might shrink, for example, the effect that average age of people in a county has on the model if that variable does not impact mean household income as much as other variables.

In addition to the matter of predictor variables, ridge regression models work best for data sets that contain a small amount of observations with respect to the amount of predictor variables. Your observations are the individual counties.

In short, your county dataset does not require the simplifying tools that ridge regression offers, since there is not a vast amount of predictor variables and since there are plenty of observations, but we will try it anyway and compare it to a model without the simplifying agent in case ridge regression ends up being a better model for your purposes.

The graph below shows the value of the modeled relationship between predictors and mean household income vs. the value of the model's penalty. As the penalty increases, the variable relationships approach zero.



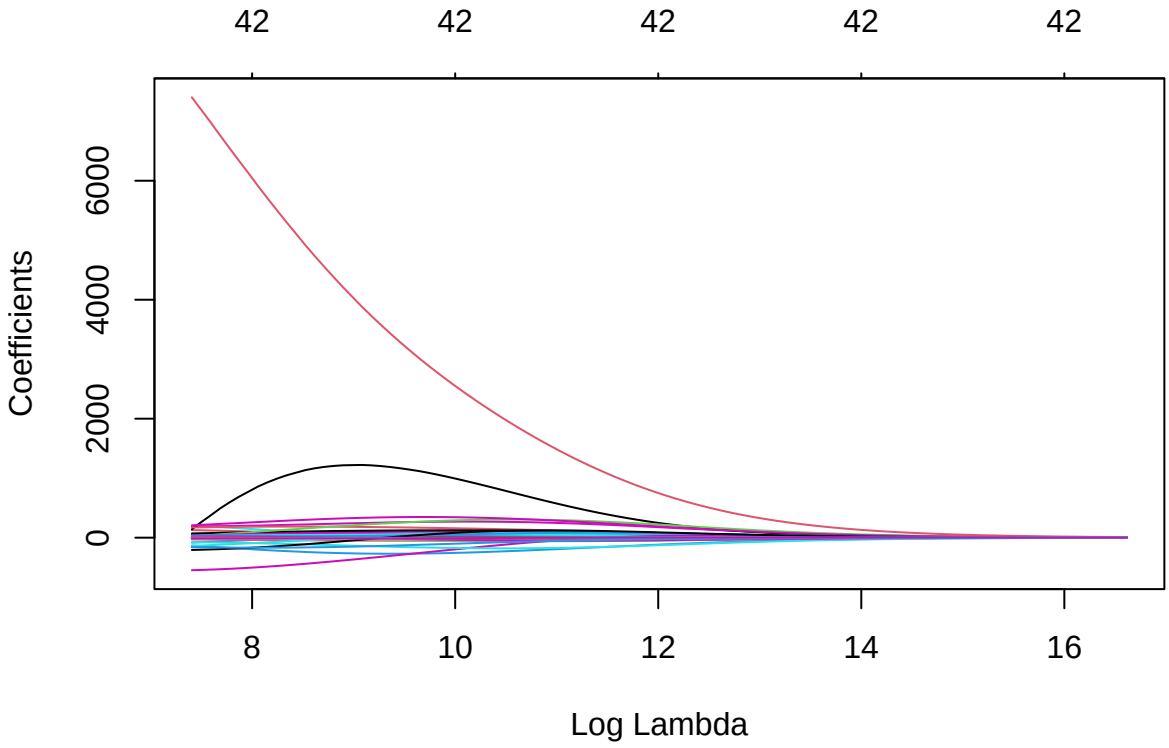
Ridge regression involves a "penalty". This penalty is the mathematical tool that allows us to gauge how much we need to simplify your predictor variables. We calculated the penalty, and it came out to .00005.

The fact that the penalty is so small indicates that a ridge regression model fitted to your data will perform as well as a statistical model that does not involve such a penalty. For clarity, the smaller a penalty is, the less it shrinks the impact the predictor variables have on the model. The software suggested a consistently small penalty for every predictor variable.

We will attempt a similar model to ridge regression below.

LASSO

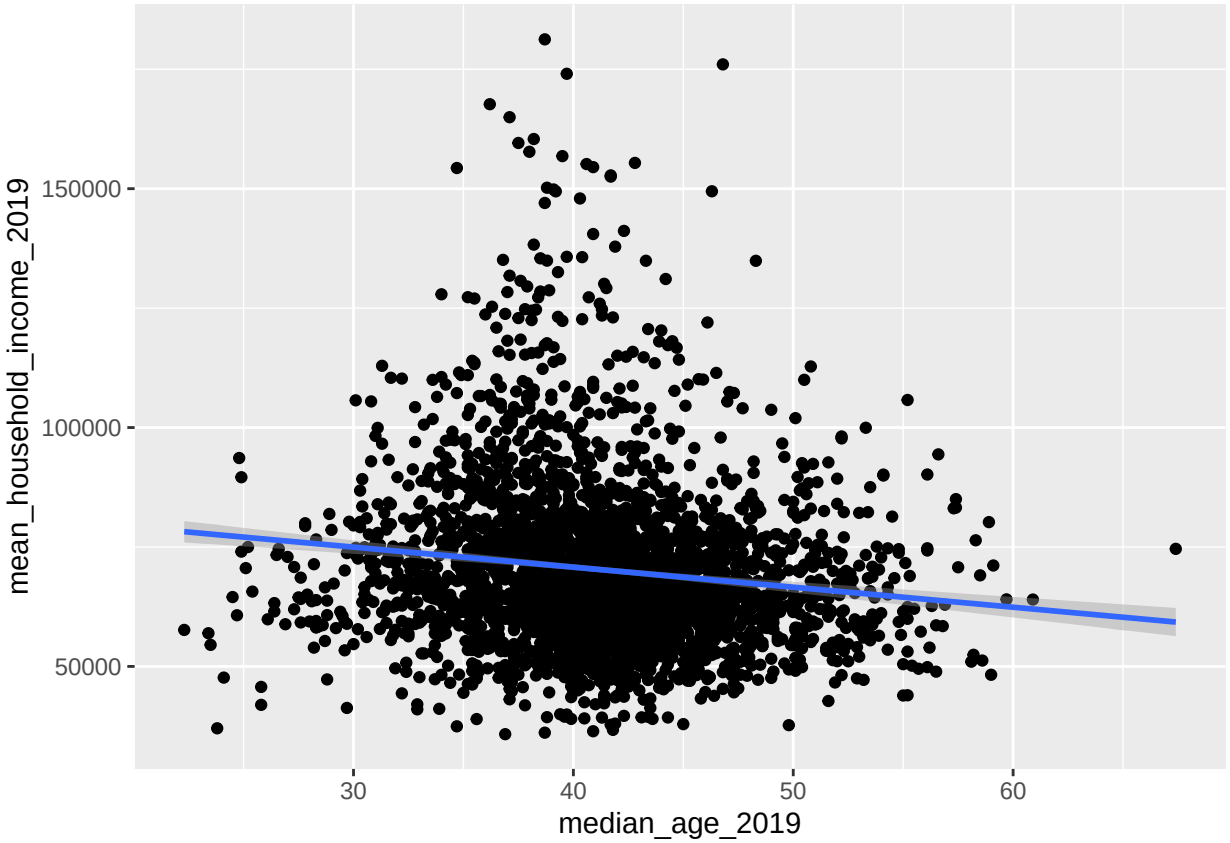
LASSO is similar to ridge regression. It is a stronger way to shrink predictor variable impact, allowing for the penalty to equal exactly zero if the software so suggests.



We received the same results for LASSO as we did for ridge regression. The software has given us a penalty of .00005, indicating that LASSO will perform as well as a model that does not involve a penalty. One such model is multiple linear regression, which is described and used below.

Multiple Linear Regression

Linear regression models a linear relationship between predictor variables and the chosen outcome variable, which you have chosen to be mean household income. Imagine we only seek to find the relationship between the median age in a county and the mean household income in a county. Plotting this data on a graph, we would see a swarm of dots signifying th salary for one county, given a specific value of median age. Please observe below.



The plot above does not paint a particularly obvious relationship between the variables but it shows the general process of linear regression.

Our goal is to build a linear model, like above, but with far more dimensions. We are able to include more x-variables, which will allow us to identify relationships among the many predictor variables with respect to mean household income. The beauty of multiple linear regression is that not only can we assess many relationships at once, but when we build a large model we can control every other predictor variable involved except the one we want to assess. For example, if our model contained median age, percentage of households with a computer, and percentage of women-owned firms, we could assess the impact of the perentage of women-owned firms on mean household income while keeping median age and percentage of households with a computer constant. This will provide us with excellent insight for answering your overall question: which institution has those most impact on mean household income?

The MLR model creates its linear relationship by setting up a line that is the closest to all the data points overall. It does this by finding the squared error between each point and a line, and chooses the line that minimizes this value.

An additional nuance to MLR, beyond the way it handles error, is the method used to choose which variables are best for the model. Not all predictor variables are beneficial for predicting a certain outcome variable. We have over forty predictor variables available to us, and we want to pick the ones that best predict mean household income. One method of doing this is beginning with a model that has no predictor variables and adding variables one by one based on which most critically and beneficially impact the linear model. This is the method we choose. We call it forward stepwise regression, and it is carried out below.

This method uses a particular measure of error to assess how much each variables impacts the model and adds the most impactful predictors to the model one-by-one. We have provided the model with every possible predictor variable and we instructed the model to begin adding variables one-by-one.

The forward stepwise regression method ultimately provided us with the following list:

- bachelors_2019
- household_has_computer_2019
- household_has_broadband_2019
- poverty_2019
- poverty_under_18_2019
- household_has_smartphone_2019
- asian_2019
- households_speak_asian_or_pac_isl_2019
- housing_one_unit_structures_2019
- households_speak_other_indo_euro_lang_2019
- hs_grad_2019
- poverty_65_and_over_2019
- mean_work_travel_2019
- unemployment_rate_2019
- uninsured_2019
- households_2019
- pop_2019
- age_over_65_2019
- uninsured_under_19_2019
- households_speak_limited_english_2019
- veterans_2019
- uninsured_under_6_2019
- two_plus_races_2019
- other_single_race_2019
- native_2019
- persons_per_household_2019
- black_2019
- uninsured_65_and_older_2019

- age_under_5_2019
- avg_family_size_2019
- hispanic_2019
- pac_isl_2019
- white_not_hispanic_2019
- white_2019
- households_speak_spanish_2019
- age_over_18_2019
- break
- housing_two_unit_structures_2019
- housing_mobile_homes_2019
- age_over_85_2019
- median_age_2019
- households_speak_other_2019

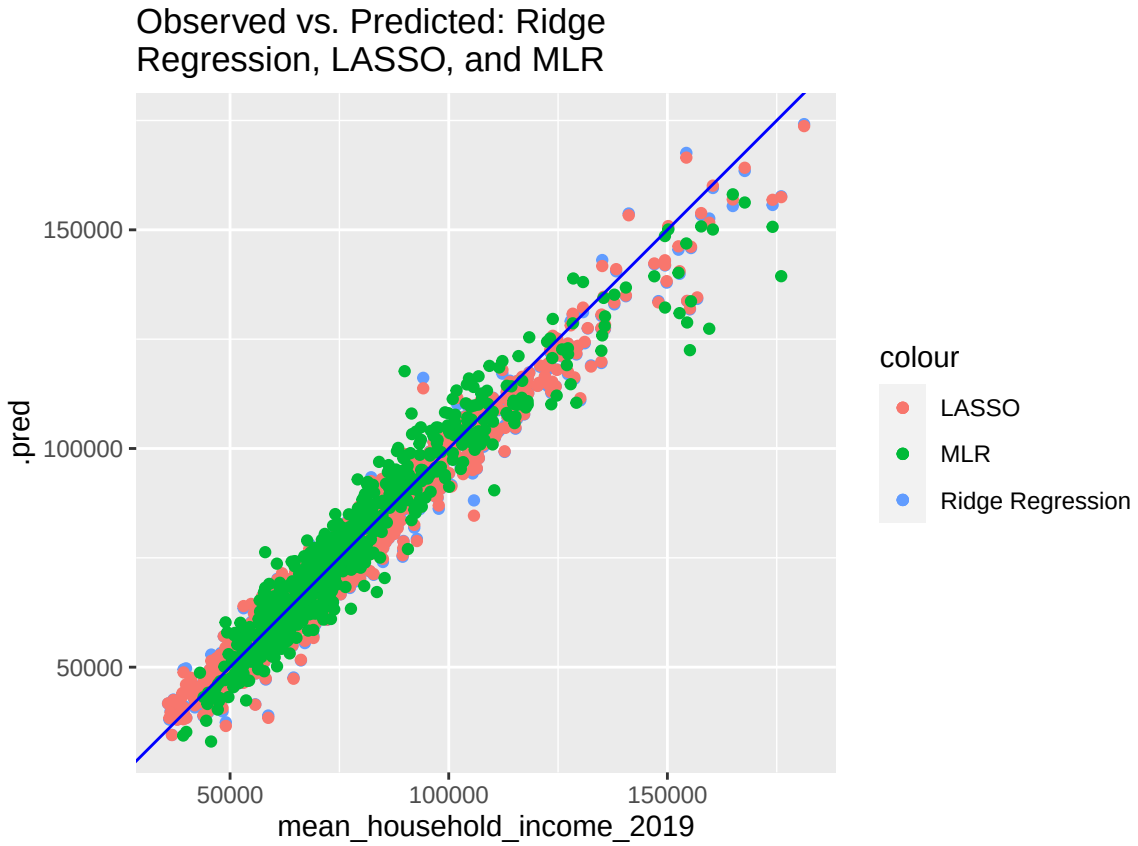
You can see exactly which variables had the most impact on the model. Where the list says "break" is where the variables that are being added begin to have a negative impact on the model. In other words, adding the percentage of houses with two-unit structures, mobile homes, and percent of population over the age of 85 makes a worse model.

Please not that the most important measures of predicting mean household income are percentage of the county with a bachelors' degree, measure of available technology, and poverty levels.

Compare and Contrast: RR, LASSO, MLR

We will now compare the three models we have used: ridge regression, LASSO, and MLR.

Plotting Observed vs. Predicted



As mentioned when assessing the first two models, the penalty our software gave us showed that we do not have the type of data that makes having a penalty beneficial. We described earlier how penalties are meant to make modeling easier for datasets that contain a great deal of predictor variables with respect to the amount of availabe observations. Our data, rather, has plenty of observations and a healthy amount of predictors. MLR provides us with the best model. Even if we used one of the other two models, the model would turn out roughly the same as that with MLR because the penalty is close to zero.

Regression Splines

We now move on to models that are not linear, beginning with regression splines. You are likely familiar with polynomial models, which can have any degree and curve through the data points in order to capture trends. Certain statistical methods allow us to break up our data into equidistant chunks and apply polynomial models inside each chunk. These models are called smoothing methods, as they build from piece-wise models and smooth out the model's shape in order to create a model from which we can draw inference.

For ease of interpretation, we will only assess the impact of median age in a county on mean household income. We are using this variable because it is not important in predicting mean household income. If we choose a variable that is very important, like bachelors' degrees in a county, it would be uninformative: the relationship is very obviously linear and would not have any indicative peaks and valleys no matter how we altered the parameters.

Before showing a model, let's discuss the key elements involved in a regression spline–there are three. The easiest to interpret is degree, which is the degree of the polynomial that will be fit inside each chosen division of the data. Say we decided to divide the data into four equal sections. There will thus be four polynomials

that make up the full model. If we choose degree three, each polynomial will be a cubic polynomial, and will have two "humps". The next element is actually the choice of how many divisions there will be in the model, called the amount of "knots". The knots signify the division between segments, so if there are four regions there will be three knots. Softwares place knots that are equidistant, so we will not worry at this time about where the knots are placed, only how many are placed. (Briefly: intuitively, knots should be placed where the most change occurs in a dataset, but in order to combat the problem of a software's tendency to place equidistant knots, a researcher can simply increase the amount of knots or the degree of the polynomial inside those knots in order to have a model that more closely fits the data.)

The relationship between all elements is:

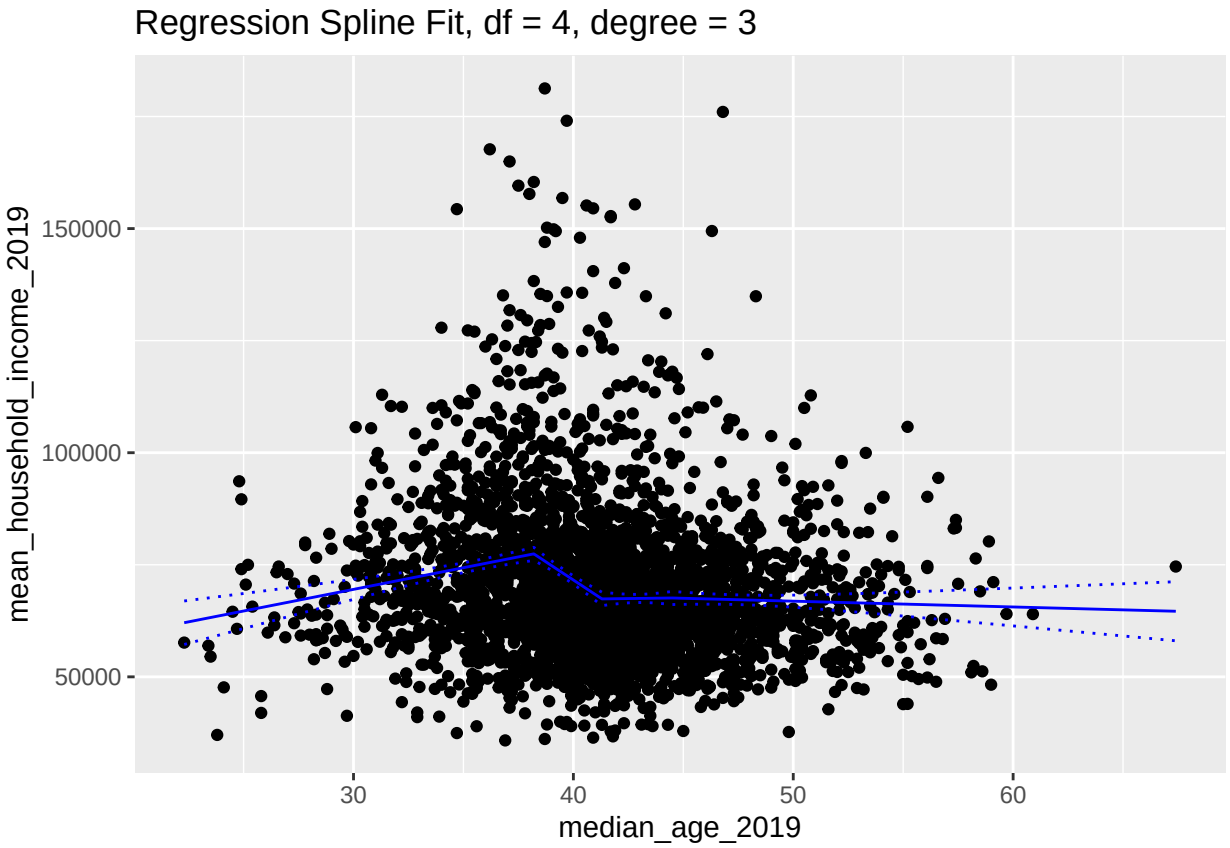
degrees of freedom = number of knots + degree.

If two elements–like degree and number of knots–are stipulated, the third is forged automatically. For simplicity, we will only discuss number of knots and degree; it is easier to think about these two than degrees of freedom when considering how we build a regression spline.

We will build and display a number of these models, showing the impact of altering the number of knots and the internal polynomial degree.

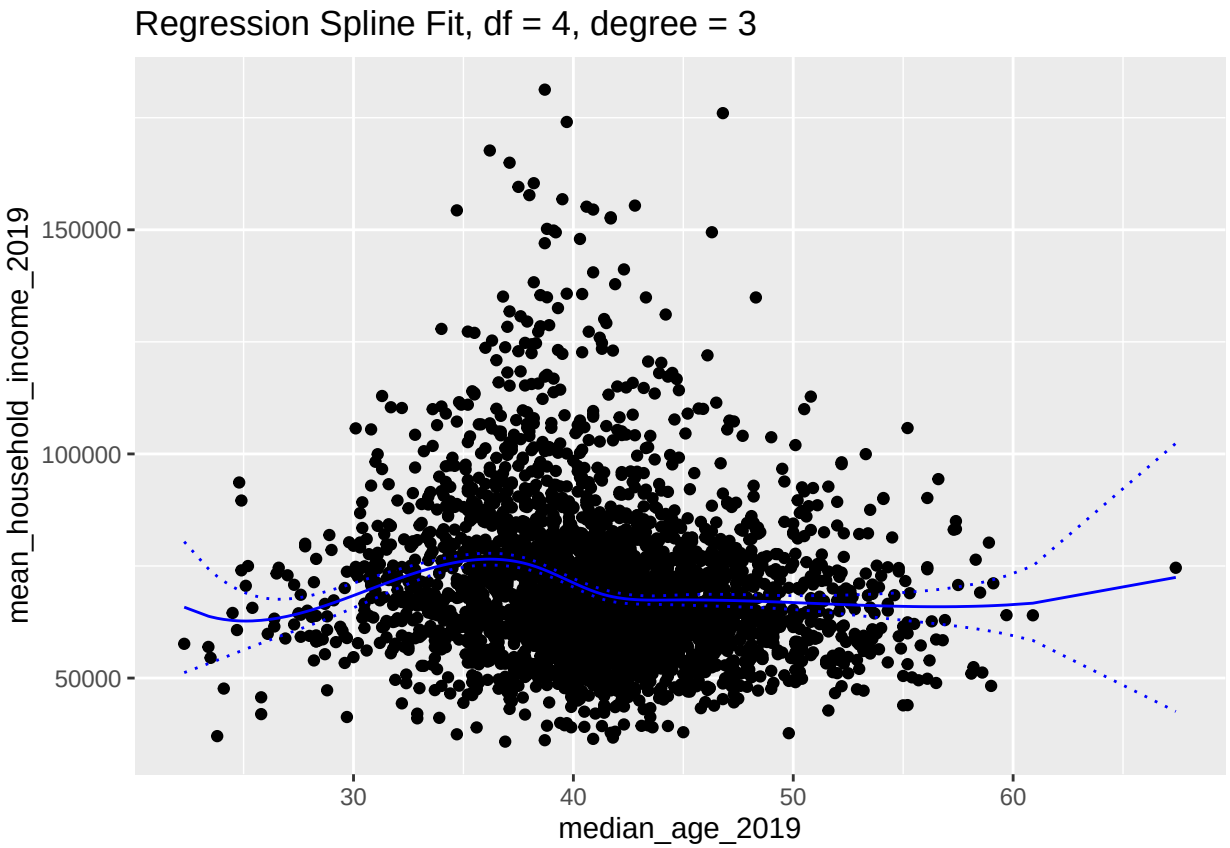
First, we will build a model with three knots and degree one. This will create a model with four lines stitched together.

Model 1



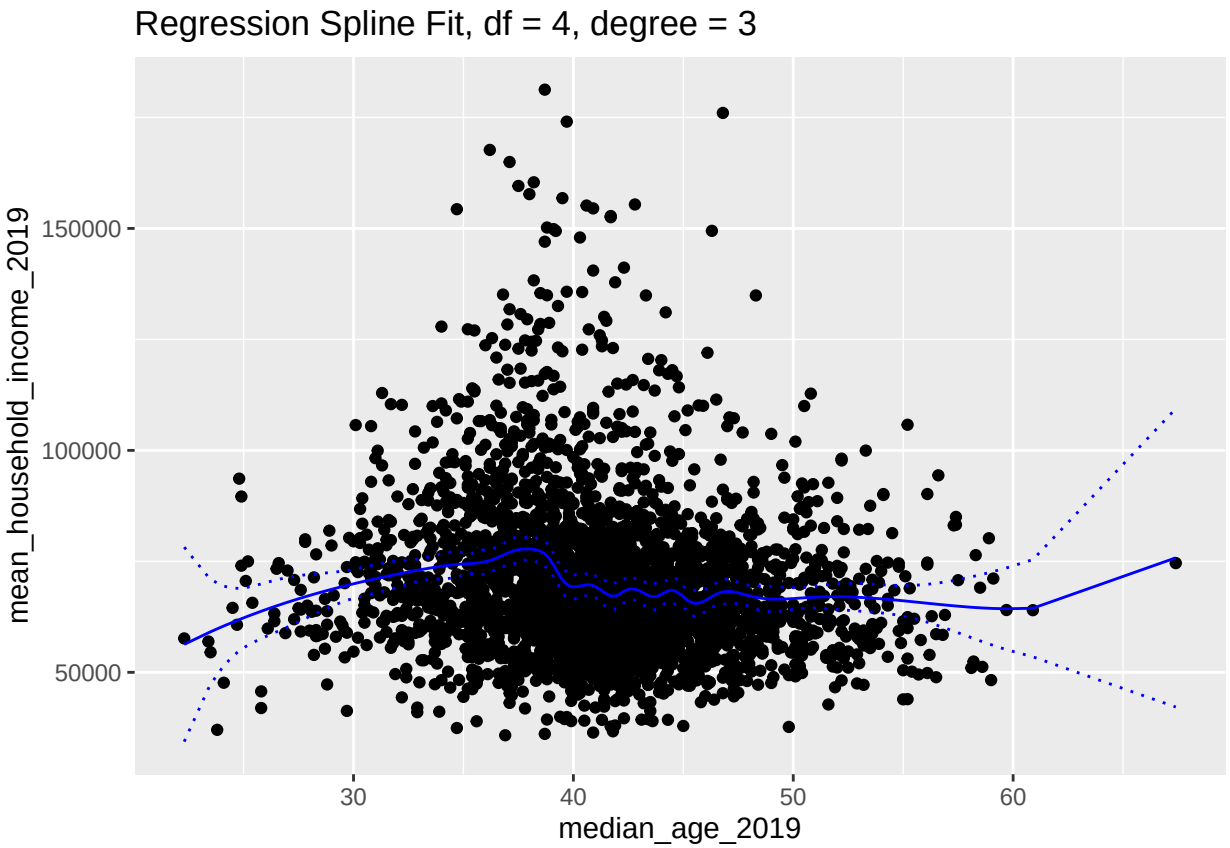
For the next model we will use the same amount of knots but use the polynomial degree that is typically used with splines: three.

Model 2



In order to demonstrate what a researcher can do to make a spline fit the data extremely well, the next model will have the typical cubics but will contain 19 knots.

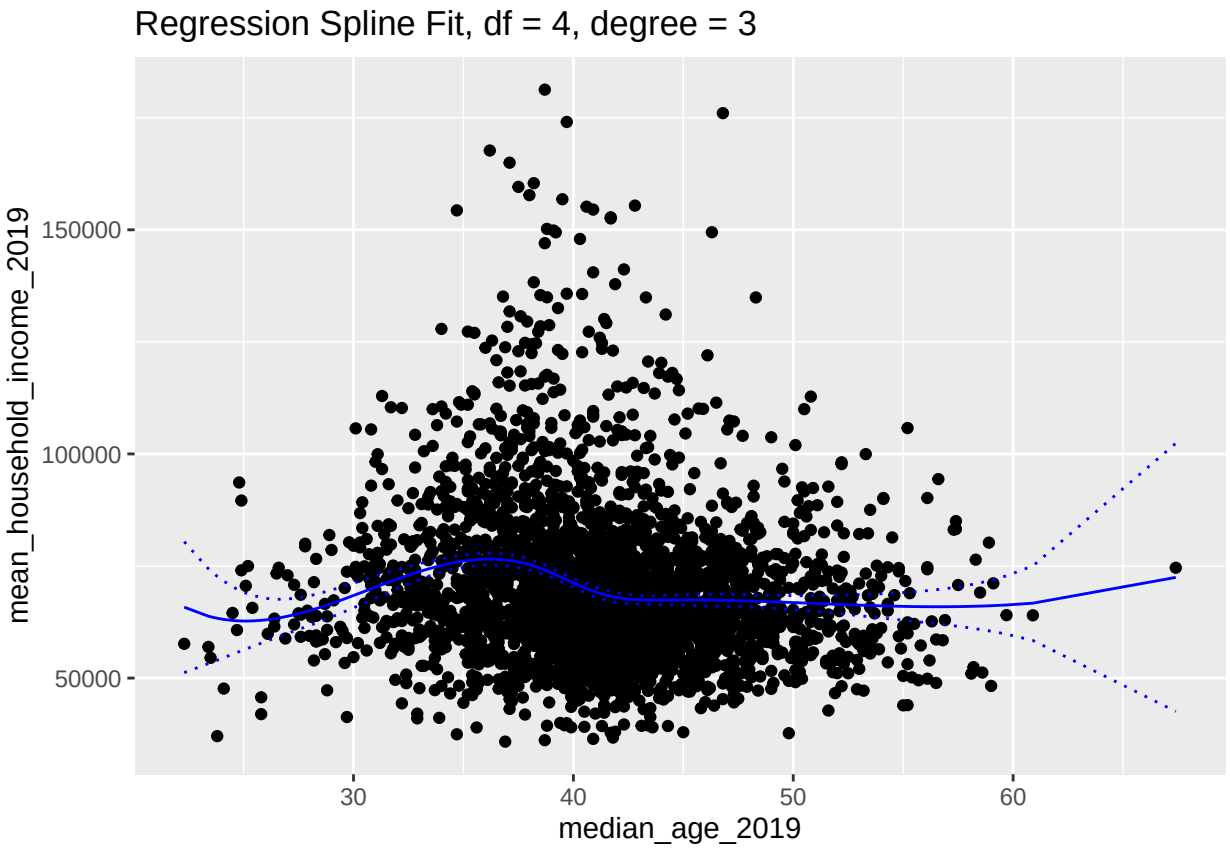
Model 3



There are methods that we can use similar to the methods used with ridge regression and lasso that will allow us to pick the ideal number of knots. For simplicity, we are just going to choose a couple of models that appear to follow the data well. We will use the standard degree of three and pick between three and four knots.

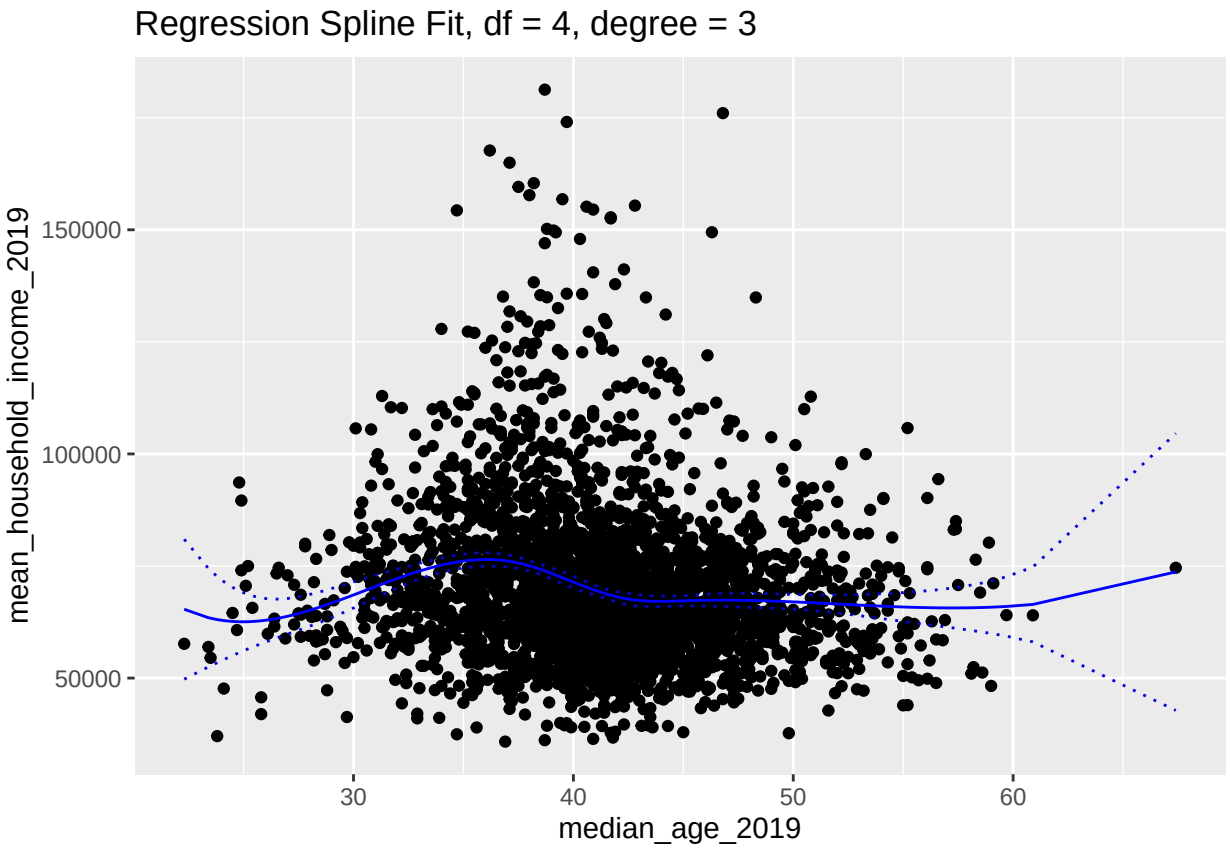
Three knots:

Model 4



Four knots:

Model 5



We find that four knots with a degree of three paints an interpretable picture of the impact median age has on mean household income.

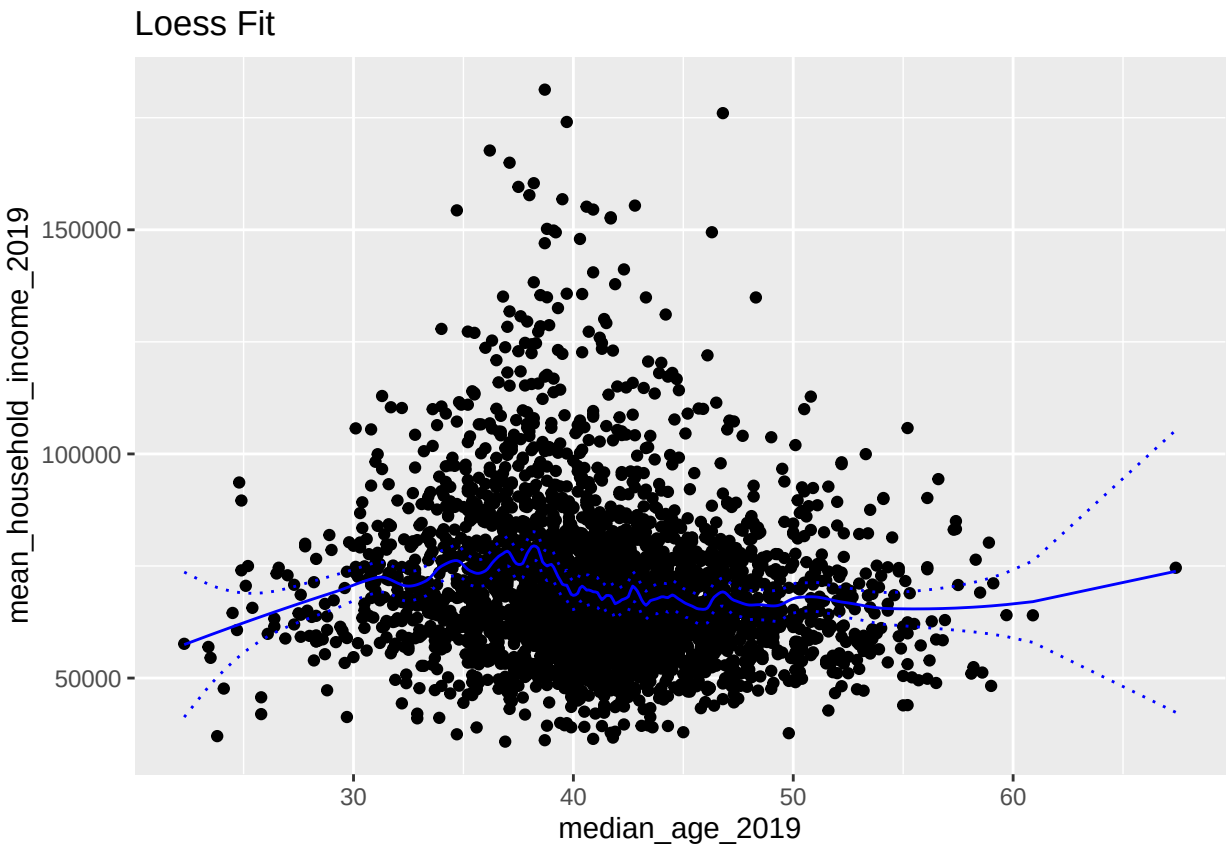
Local Regression

Local regression, or loess (or lowess), is another smoothing method. It is centered around using only local data points at a certain point in order to build on a model. Specifically, at a mid-level x-value for a data set, loess considers the data points near that mid-level x-value when adding to the model, not all available data points.

Whereas with splines we concerned ourselves with degree and knot number, for loess we are concerned with exactly how many data points are considered when building on each x-value level in the model. In other words, we adjust the range of x-values. If our range is .3, then at each point we assess the model and build on it, we are using 30 percent of the data.

In order to display the power of loess, we will first use a very small span of .1:

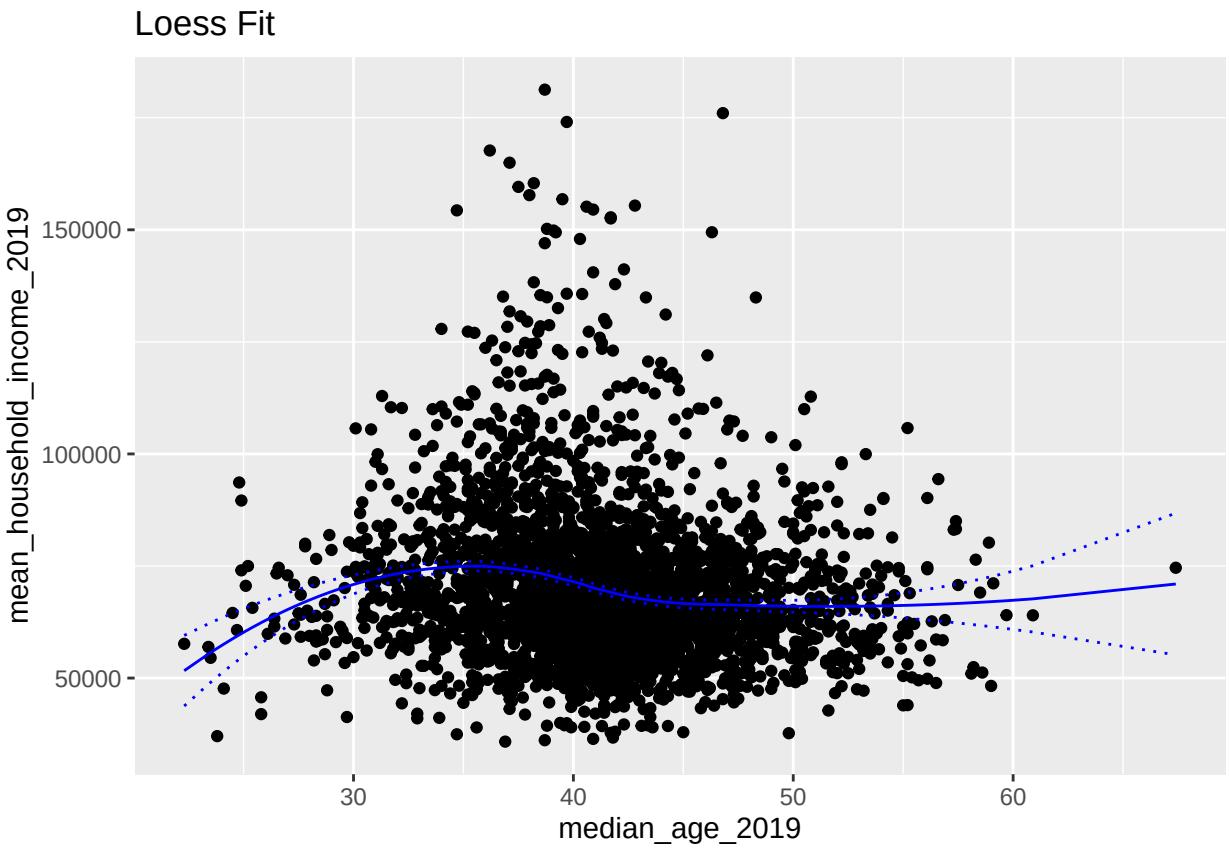
Model 1



As you can see, using 10 percent of data points at each assessment creates a model that fits the data incredibly well.

Next we will show the effect of using a very large range, .9:

Model 2

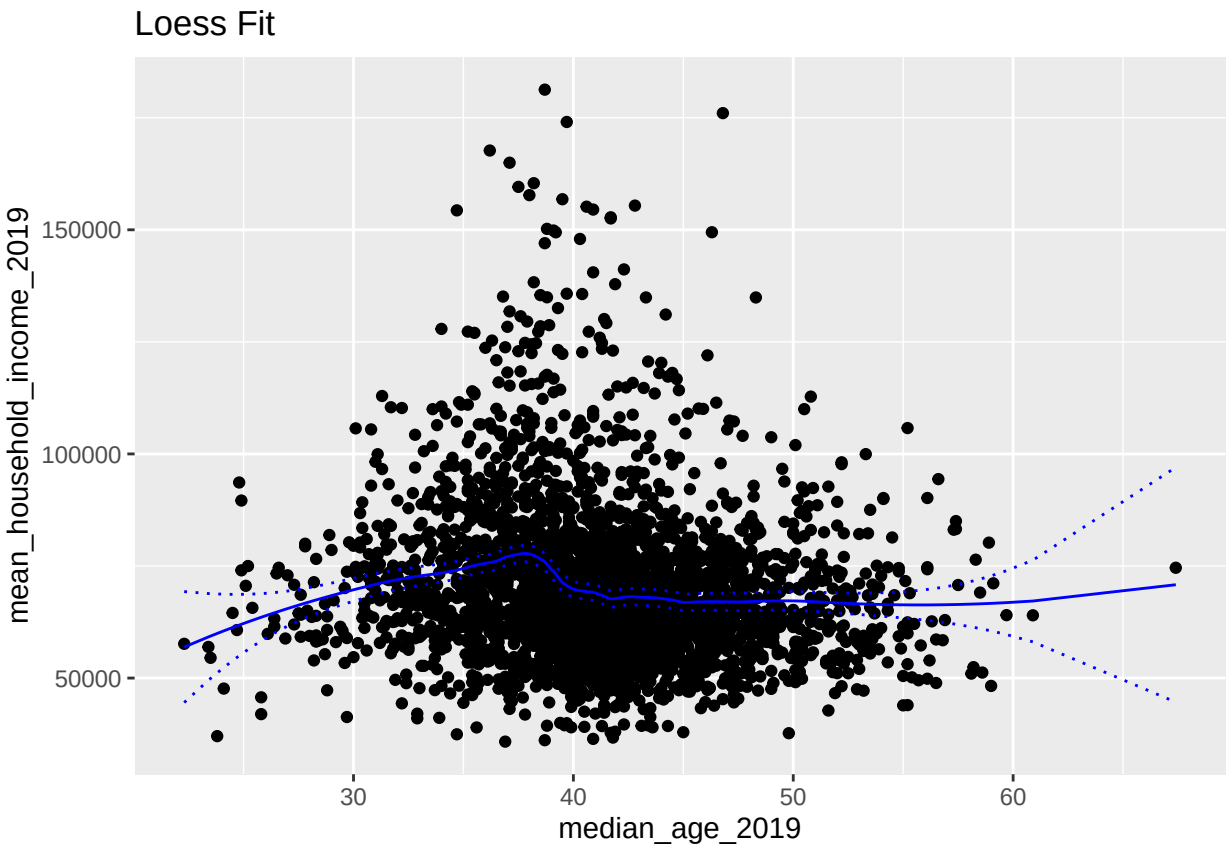


Using a large range, as you may have expected, creates a more general model.

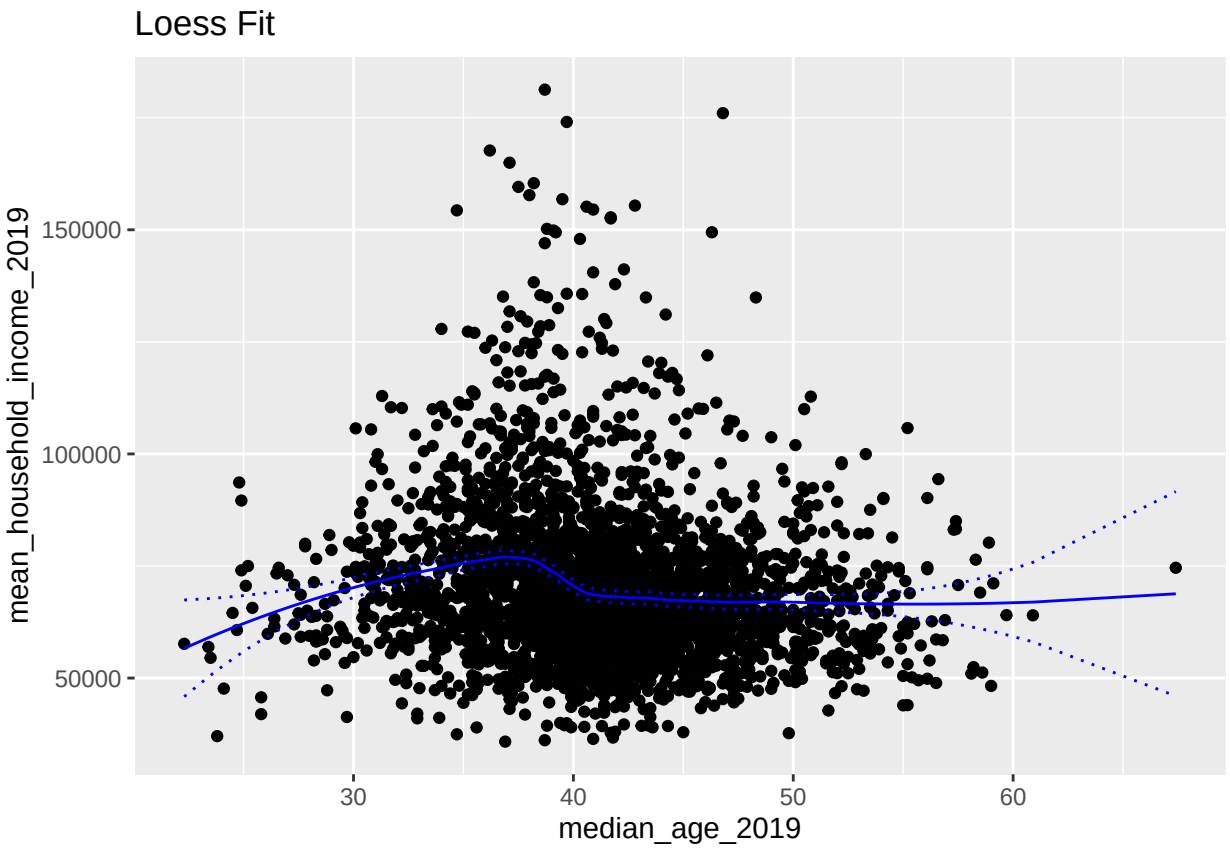
As with splines, there are methods for finding the ideal level of span. For our purposes, we will try out levels of span and visually assess which level seems to provide us with a solid interpretation of median age vs. mean household income.

Let's look at 30 and 50 percent. 30:

Model 3


Loess Fit
mean_household_income_2019
150000
100000
50000
30
40
50
60
median_age_2019

Model 4



A range of 50 percent paints a solid picture of the trends between these two variables.

We now move on to a vastly different type of analysis, which involves categorical variables.

Logistic Regression

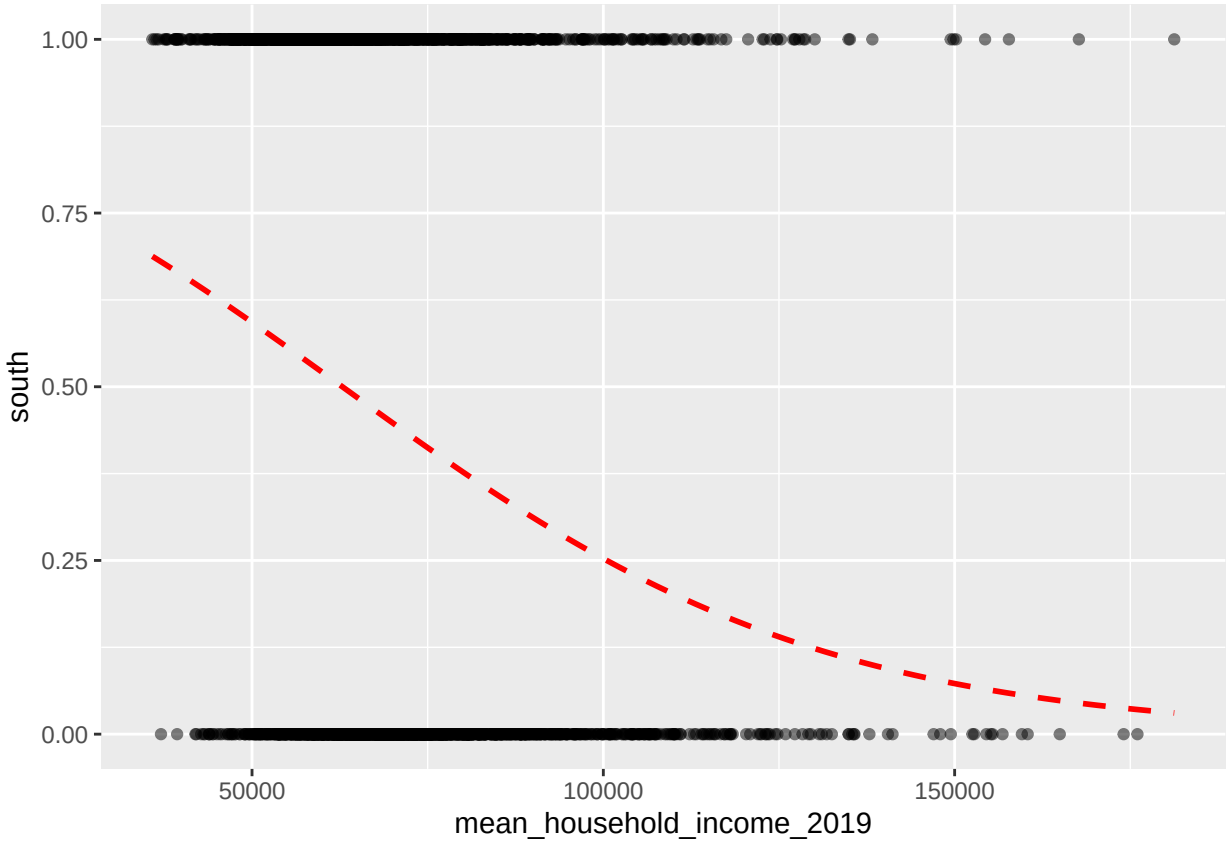
Creating a Categorical Variable

Categorical variables are those that concern a value that is not associated with an order of numbers. The percentage of those in a county with a bachelors' degree is quantitative: the ultimate value can be any level of number and interpretation comes from assessing increasing or decreasing (but above all, orderly) levels of this variable.

The data you provided us is almost entirely quantitative–such is the nature of census-type data. We have created a variable that tells us the region of the United States that a county is in: south, north east, north central, and west.

We ultimately want to create a model that tells us–given a certain mean household income of a county–the likelihood that a county is in a certain region of the United States. We will start with the south by manipulating our categorical variable to be one that equals 1 when a county is southern and 0 otherwise (these types of variables are called indicator variables).

The type of analysis we will be doing is called logistic regression. It is a model type that uses data manipulation and probability in order answer our question: given an increase in mean household income, what is the probability that a county is in the south? The logistic regression is given below.



Our model tells us that as mean household income decreases, a county is less and less likely to exist in the south. In other words, southern counties have less mean household income than counties in other regions.

This model type is useful because it is easily interperetable and there are not many other ways to answer the question that was asked. Additionally, the model is intriguing because it flips the roles of the original x and y. Logistic regression allows us to assess the probability of a categorical variable given certain levels of what we chose as our outcome variable.

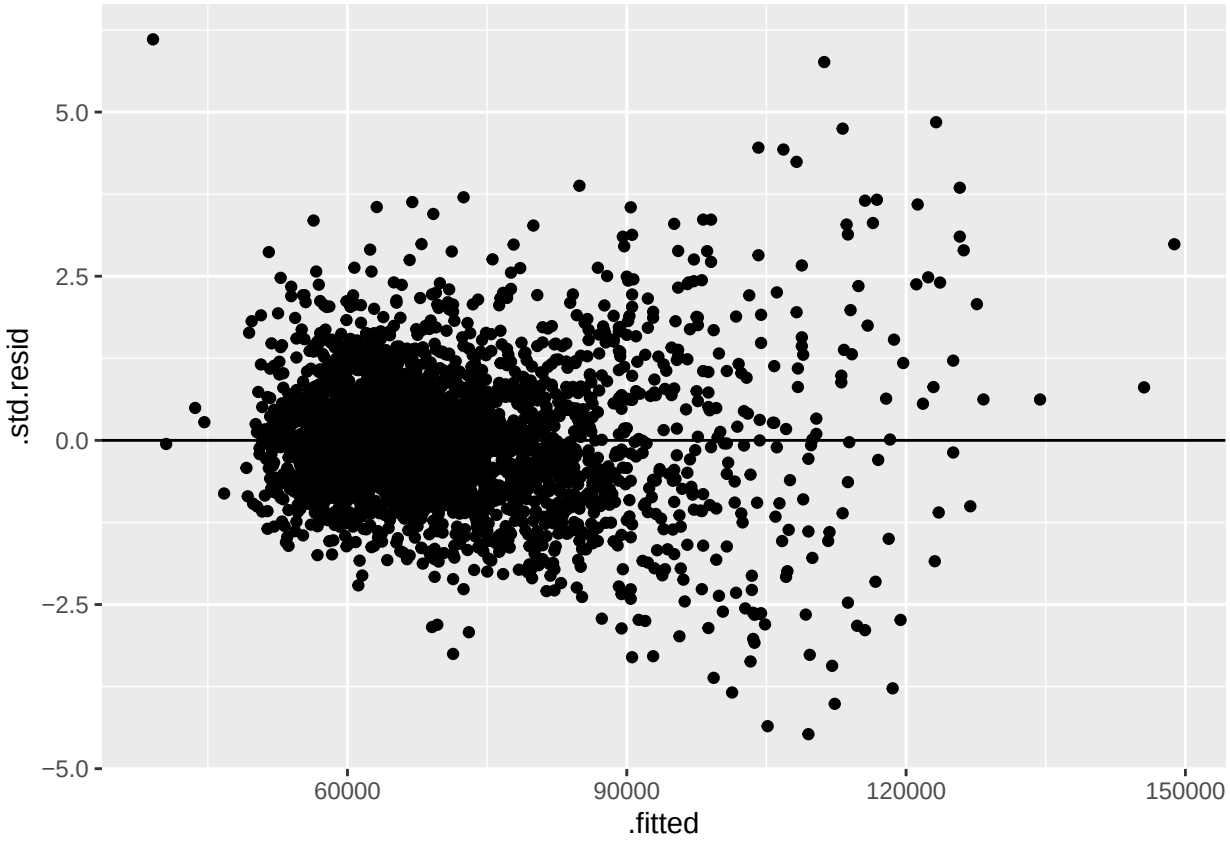
Because it allows us to assess our y variable in useful ways and allows us to use categorical variables easily, logistic regression is an incredibly valuable tool for analysis. If we were to expand this analysis, we would create plots that assess the west or north central in the same way the plot above assessed the south. For now, this completes our assessment of new statistical models.

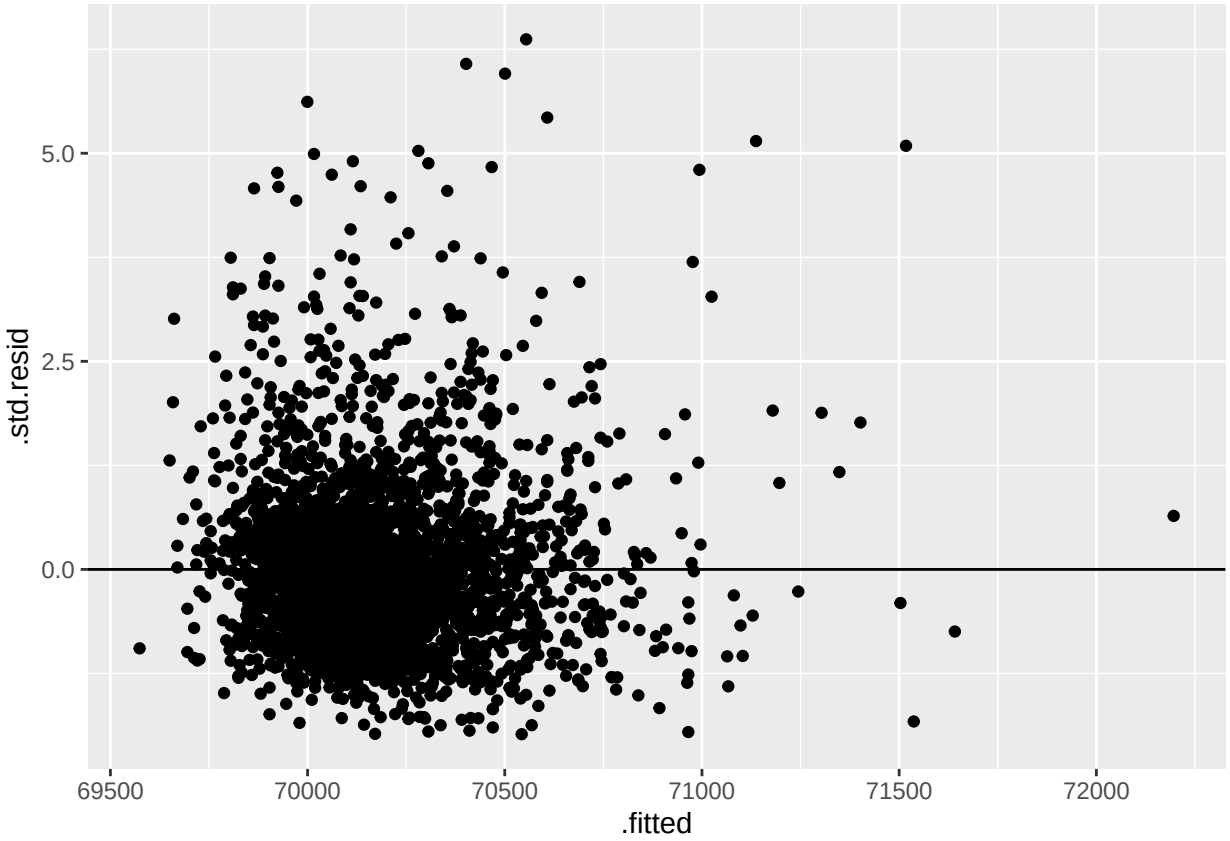
Summary

Residual Plot

The following residual plots show that our MLR model did a good job picking important variables. The first plot shows the residual plot for bachelors vs. mean household income. The second shows mobile homes vs. mean household income which was notably unnecessary for our model.

These plot depict a juxtaposition between "good" and "bad" variables for our chosen method of modelling predictors vs. an outcome. The software is able to discern which variables negatively impact the model, and the residual plots can lead us to inference on why those variables are not included in the forward stepwise model. The mobile home variable does not meet line conditions, since the errors are neither normal nor constant.


5.0
2.5
0.0
−2.5
−5.0
.std.resid
60000
90000
120000
150000
.fitted



Analysis

By far the most interesting part of this process was completing the full forward stepwise regression model.

Due to the nature of our data, we were unable to complete certain types of analysis that would have been formative for datasets that do not contain an entire population. For example, we had too much data to require any kind of imputation. In addition, none of our variables had any kind of observable non-linear relationship that would require us to make creative data transformations.

Since the data is economic, however, the name of the games is to throw all the spaghetti at the wall and see which noodles impacted the wall most significantly. Forward stepwise regression was a satisfyingly organized way to do this. At first when we put every single possible predictor into the data, the stepwise regression cheekily told us that median household income and per capita income were the most indicative predictors of mean household income. In order to perform the true analysis, we removed these predictors. We ended up dropping several variables that were obvious indicators of income.

As discussed in the MLR section, forward stepwise regression told us that the percentage of a county that has a bachelors' degree, followed by several measures of available technology in a county, followed by measures of poverty, were the variables that most impacted mean household income.

A summary of the full linear model is provided below.

```
## # A tibble: 37 x 5
##    term                                 estimate std.error statistic  p.value
##    <chr>                                   <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                          -809054.   368395.     -2.20  2.85e- 2
##  2 bachelors_2019                          1363.       60.3    22.6   1.61e-82
```

```
##  3 household_has_computer_2019               -316.      118.     -2.67  7.83e- 3
##  4 household_has_broadband_2019                97.4     106.      0.919 3.58e- 1
##  5 poverty_2019                             -2268.      178.    -12.7   4.72e-33
##  6 poverty_under_18_2019                      482.      119.      4.07  5.41e- 5
##  7 household_has_smartphone_2019              -75.8      99.4    -0.763 4.46e- 1
##  8 asian_2019                                8651.     3738.      2.31  2.10e- 2
##  9 households_speak_asian_or_pac_isl_2019    1976.      522.      3.78  1.69e- 4
## 10 housing_one_unit_structures_2019           -92.7      52.5    -1.77  7.80e- 2
## # ... with 27 more rows
```

We have discussed many types of statistical models and the ways we can assess their accuracy. Data can be created, manipulated, and destroyed, all in a myriad of creative ways, but by the nature of our data we did not have to go through much of any of those three processes.

Overall, statistical models can be complicated and can require involved manipulation, but ultimately our favorite and most useful model was the simplest there is: linear regression.

We hope that this model floors you. It paints an easily understandable yet moderately telling picture of the impact that a wide range of demographic variables has on mean household income in a given county. The model summary shows how the model conducted tests on each variable in order to determine how significant its relationship is to the outcome variable.

An additional aspect of this report we found exciting was the prospect of error removal via variable addition. We start with a great deal of discrepancy between prediction and actual value when only one predictor is involved, and gradually explain out that error by adding more predictors. The ability of a variable to explain out error is assessed by f-tests, which were not covered in this report. The AIC assessment used in the forward stepwise regression had a similar idea, where variables were judged by their ability to reduce the error created by fitting a model to data.

Implications and Applications

We appreciate the linear regression model especially because it answers your question, o client. Which institution has the strongest relationship with mean household income, holding many other institutions constant? The answer: college education. While this may seem obvious, the model we have given you provides a list of institutions in order of their magnitude of impact. If you are unable to put funds in one area, put it into the next most impactful.

This research is preliminary. The variable immediately following bachelors is an indication of the fact that more work needs to be done in order to create analysis of value with variables like this: whether or not a household has a computer is obviously correlated with wealth. So is poverty. We would need to perform more manipulation or account for different predictor variables in order to take care of this issue. We have a hunch that the relationship between education is not as obvious as it seems.