How Race, Income, and Education Relate to Internet Access in US Counties

Abstract

Lack of internet access is an important question of equality during a time when readily available internet use is increasingly critical for daily life. Interested in the demographic factors that contribute to internet access, we use the 2016 American Community Survey to examine the relationship between lack of internet access and race, education, and economic characteristics. Using multivariate regression with minimization of mean squares error with backward selection, we identified a multivariate polynomial model using the median household income, the population proportion of Native American residents, and the proportion of residents who have attended some college . From the model, we find that median household income and proportion of residents who have attended some college is negatively related to the percentage of households without internet access while the population proportion of Native American residents is positively related to the percentage of households without internet access.

Introduction

The COVID-19 pandemic spurred a major shift in our daily activities as we transitioned our interactions from in-person to online. We have seen a large increase in people working from home, attending classes remotely, socializing over zoom calls, and even ordering groceries from their phones. With seemingly everything happening online, internet access has become especially important. [2] However, not everyone is fortunate enough to have reliable internet access in their home, and the pandemic has amplified the disadvantages felt by those who don't have access. [3] This heightened need for and disparity concerning internet access spurred by the pandemic makes it all the more important to study which parts of our population live without internet and what community factors might contribute toward lower rates of household internet access. Investigating these types of questions could help pinpoint specific parts of our population that are most affected by a lack of internet access, which could then inform potential assistance programs or policies seeking to minimize the proportion of homes without internet access.

In 2010, researchers at the University of Washington explored the best ways to track the "digital divide" across different communities, focusing on comparing Canada and the U.S. [5] They found that while the overall inequity in internet access has declined over the years, the digital divide "remains the most pronounced in the U.S." with respect to education. This means that internet access is much more concentrated among highly educated U.S. citizens than among those with less educational experience. This correlation between internet access and education level exposes the importance of internet access for both achieving in school and pursuing higher education.

Our research seeks to find which factors or community traits are the strongest indicators of a high proportion of residents without internet access. We are specifically looking at the influence of race, education level, and income. We predict that areas with higher minority populations, lower education levels, and lower median income levels will have higher rates of households without internet access. Answering this question of which community traits affect a county's proportion of households without internet access will help identify communities most in need of assistance. It will also shed light on how socioeconomic factors and internet access are related, which could be useful in trying to minimize these inequalities.

Methods

Data. The internet data was collected through the 2016 American Community Survey (ACS) conducted by the US Census Bureau. [5] The ACS samples approximately 295,000 households monthly. The annual consolidation of the ACS results in information from 3.5 million sampled households. The ACS is released to the public every year with the data gathered from counties with populations greater than or equal to 65,000 people. The unit-response rate measures the amount of households that ultimately respond to the survey after being contacted. The lower the unit-response rate, the more biased the sample data may be. [8] The 2016 ASC has a unit-response rate of 94.7 percent.[7]

Variables. We are interested in understanding the response variable of the percent of households within each country that do not have internet access. The predictor variables we will be examining define the racial, educational, and economic profile of each country. To understand the racial profiles we use the population proportions of White residents, Black residents, Asian residents, Native residents, and Hawaiian residents, respectively. Each proportion is calculated by dividing the population of the racial or ethnic group by the sum of the populations of the given racial groups. For example, prop_white = P_white/(P_white + P_black + P_asian + P_native + P_hawaiian + P_others). To understand the economic profiles, we use the variables of the county's gini index, median household income, and the median percent of income spent on rent. The gini index is a measure of statistical dispersion that indicates the level of income inequality within each county. The coefficient ranges from 0 to 1 where 0 indicates perfect income equality and 1 indicates perfect income inequality. The larger the coefficient, the more dispersed incomes are in that county. [4] The median rent income gives a percent of income spent on rent. This accounts for different costs of living across different counties. Finally, to understand the educational profiles we are looking at the population proportions for varying levels of education attained for each county. The levels of attainment are having education at or below the 8th grade level, having some high school education but no diploma, having a high school diploma or the equivalent, having some college education, or having at least a bachelor's degree. The population proportion for each respective education level is calculated by taking the population with that level of education and dividing it by the sum of the populations for each level. For example, prop_somecollege = P_some_college / (P_below_middle_schoool + P_some_high_school + P_high_school_eq + P_some_college + P_bachelor_or_above).

Analytic Methods. We are interested in whether the racial, economic, and economic profile of a county are related to internet access in the county.

We use a multivariate regression with minimization of mean squares error to identify the variables that most strongly influence internet access. We began with all 13 variables included in a multivariate regression model and then used backward selection to remove variables from the model. For backward selection, we took a nuanced approach that examined the statistical significance of each predictor, the maximization of the adjusted R^2 value, and the minimization of the Akaike Information Criterion (AIC) value. We began backward selection by eliminating the least significant predictor until all predictors were statistically significant using the threshold of a p-value $\leq 2*10^{-16}$. Then we checked for multicollinearity within our model. Using a threshold of 2.5 for multicollinearity, two of the predictors: proportion with an education level of highschool or equivalent and proportion with an education level of some college, were found to have collinearity. We confirmed the collinearity by finding the correlation coefficient between the two predictors, which was found to be 0.82. We then made two models, each one used only one of the correlated variables. From here, we used the maximization of the adjusted R^2 value and minimization of the AIC value to guide our modeling. After obtaining a multivariate model with three predictors, we added polynomial parameters to see if it would increase the accuracy of the model. We continued to evaluate the additional parameters based on the maximization of the adjusted R^2 value and minimization of the AIC value. We stopped fitting polynomial parameters once the measures of R^2 and AIC indicated a worse model.

Results

To describe the data from the 2016 American Community Survey, the median county population is 158,071. The median age is 38.4. The median county percent of households without internet access is 14.711%, with county proportions ranging from 2.661 to 54.011%. The median proportion of residents living below the poverty line is 0.13022 with the lowest county being 0.03583 and the highest being 0.36522. The median proportion of residents who have attained at least a bachelor's degree is 0.28243, with a range from 0.09377 to 0.74903.

Through a multivariate regression with backward selection, we find that the percent of a county's population without access to internet can be modeled by

$$y = 60.44 + (-1.484*10^{-3} + 1.488*10^{-8}\, x_1 - 5.263*10^{-14}\, x_1^2)\, x_1 + 21.32\, x_2 - 1.307\, x_3\,,$$

where x_1 is the median household income, x_2 is the proportion of native residents, x_3 is the proportion of residents with an education level of some college, and y is the percent of the county residents without access to the internet. This model has an adjusted R^2 value of 0.6299 and an AIC value of 3154.071.

Discussion/Conclusion

Based on our model, the percent of county residents without internet access can be predicted using the median household income, the proportion of Native American residents, and the proportion of residents with some college experience. The coefficient for median household income indicates that if all other variables are constant, for every dollar increase in median household income, the percent of residents without internet access changes by $(-1.484*10^{-3} + 1.488*10^{-8}\, x_1 - 5.263*10^{-14}\, x_1^2)$, where x_1 is the median household income. This suggests that as median household income increases, the percentage of households without internet access decreases, however as median household income continues to increase, the rate of decrease in households without internet access is decreasing. If we consider internet access to be a commodity, it makes sense that the rate of decrease in households without internet access is decreasing as median household income increases. The coefficient for the proportion of native residents indicates that if all other variables are constant, as the proportion increases by 0.01, the percent of residents without internet access increases 0.2307 percent. This suggests that there are lower levels of internet access in counties with larger Native American populations. Finally, the coefficient for the proportion of residents with an education level of some college indicates that if all other variables are constant, as the proportion increases by 0.01, the percent of residents without internet access decreases 0.01307 percent. This suggests that there are higher levels of internet access in counties with higher proportions of residents that have education levels of some college.

Our objective was to find the community variables that most strongly influence the percent of households without internet within a county. We predicted that counties with higher minority populations, lower education levels, and lower income levels would have higher rates of residents living without internet access. As expected, our model shows that poorer communities face higher levels of households without internet access. Additionally, a larger proportion of Native American residents results in a larger proportion of residents without internet access. The proportion of Native Americans was a much more significant variable than the proportion of any other minority race. Finally, as we expected, counties with higher proportions of residents who have attended some college reflects counties that are more educated and also have lower levels of residents without access to the internet. We can interpret that the "digital divide," that the 2010 University of Washington study [6] examined with respect to education, is also strong when considering median household income or the proportion of Native American residents.

Limits. The focus of this model is on the relationship between the input variables and response variable, as opposed to prediction. This model was created based on data from larger counties, so it would not be appropriate for assessing internet access in counties with populations less than 65,000. As seen in the Residual vs. Leverage plot in the appendix, county 36, Apache County in Arizona, is an outlier. Apache County has a much higher proportion of Native American residents than other counties and 54% of residents do not have access to internet. It seems our model does not work as well for counties with significantly high Native American populations.

Future research. Although we examined the counties' lack of internet access using race, education, and economic characteristics, it is important to recognize that we did not actually find the relationship between a particular household's lack of internet access using these demographic characteristics. In other words there could be intermediate county dynamics implicitly indicated by these characteristics that are influencing the counties' lack of internet access.To understand more about how race, education, and economic characteristics relate to lack of internet access, research should examine internet access at the household level, rather than county level.

Additionally, there are more dimensions of county demographics to be explored. A potential predictor that could be significant in predicting internet access is location. Future research could include qualitative predictors based on region as well as quantitative predictors such as distance from the capital city or physical size of the county to understand the relationship of location and lack of internet access.

References

[1] https://en.wikipedia.org/wiki/American_Community_Survey

"American Community Survey." *Wikipedia*, Wikimedia Foundation, 10 Feb. 2021, en.wikipedia.org/wiki/American_Community_Survey.

[2] https://www.policylab.chop.edu/broadband-internet-access

Capozzi, Lindsay. "Broadband Internet Access, Education & Child Health: From Differences." *Broadband Internet Access, Education & Child Health: From Differences to Disparities, Part 1*, 2 Nov. 2020, policylab.chop.edu/blog/broadband-internet-access-education-child-health-differences-disparities-part-1.

[3] https://www.tandfonline.com/education_during_pandemic

Geeta Verma, Todd Campbell, Wayne Melville & Byung-Yeol Park (2020) Science Teacher Education in the Times of the COVID-19 Pandemic, Journal of Science Teacher Education, 31:5, 483-490, DOI: 10.1080/1046560X.2020.1771514

[4] https://en.wikipedia.org/wiki/Gini_coefficient

"Gini Coefficient." *Wikipedia*, Wikimedia Foundation, 4 Mar. 2021, en.wikipedia.org/wiki/Gini_coefficient.

[5] https://www.kaggle.com/madaha/people-without-internet?select=kaggle_internet.csv

GL_Li. "People without Internet." *Kaggle*, 11 Jan. 2018, www.kaggle.com/madaha/people-without-internet?select=kaggle_internet.csv.

[6] https://www.academia.edu/Comparing_Digital_Divides

Howard, Philip N., et al. "Comparing Digital Divides: Internet Access and Social Inequality in Canada and the United States." *Canadian Journal of Communication*, vol. 35, no. 1, 2010, doi:10.22230/cjc.2010v35n1a219

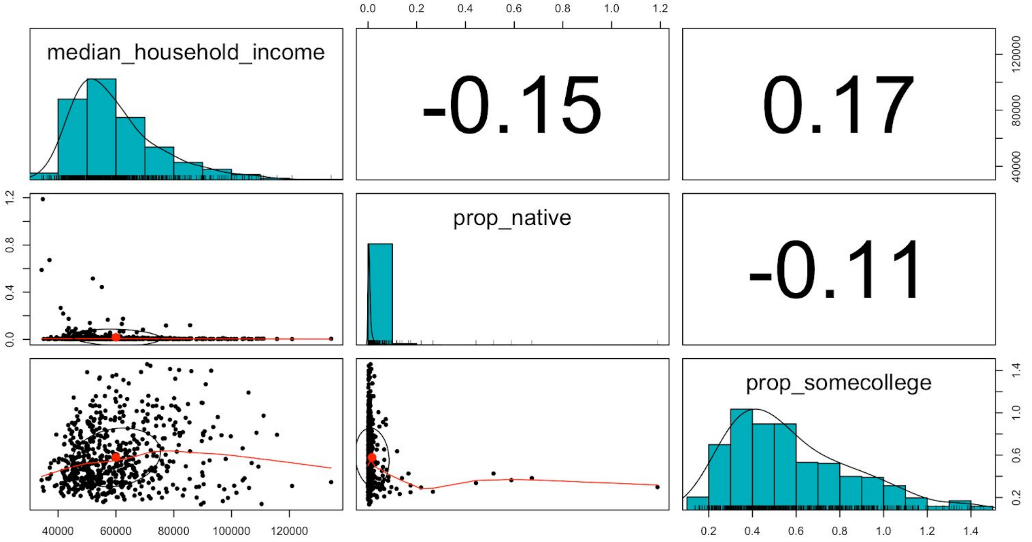
[7]
https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/index.php

Office, American Community Survey. “Response Rates.” *Response Rates | American Community Survey | U.S. Census Bureau*, 16 Jan. 2015.

[8]
https://www.census.gov/content/dam/Census/library/publications/2020/acs/acs_general_handbook_2020.pdf

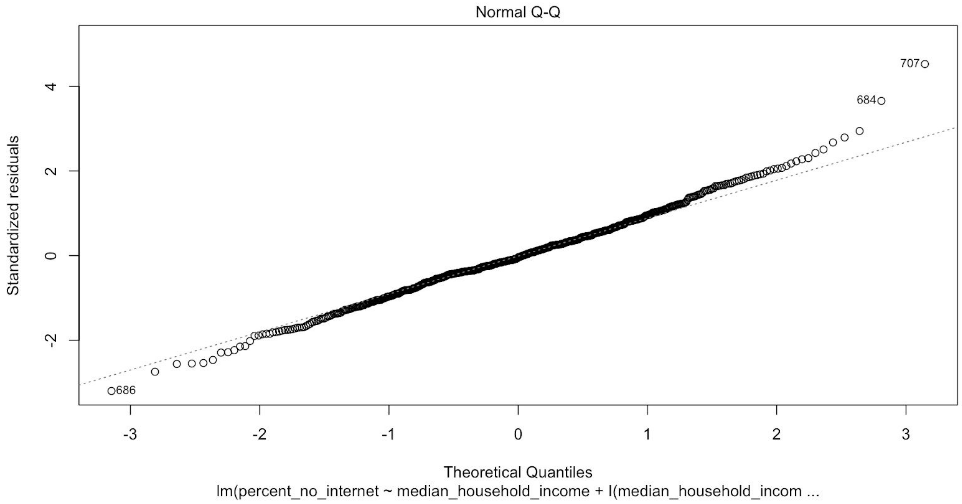
Understanding and Using American Community Survey Data, US Department of Commerce, (67-69).

Appendix

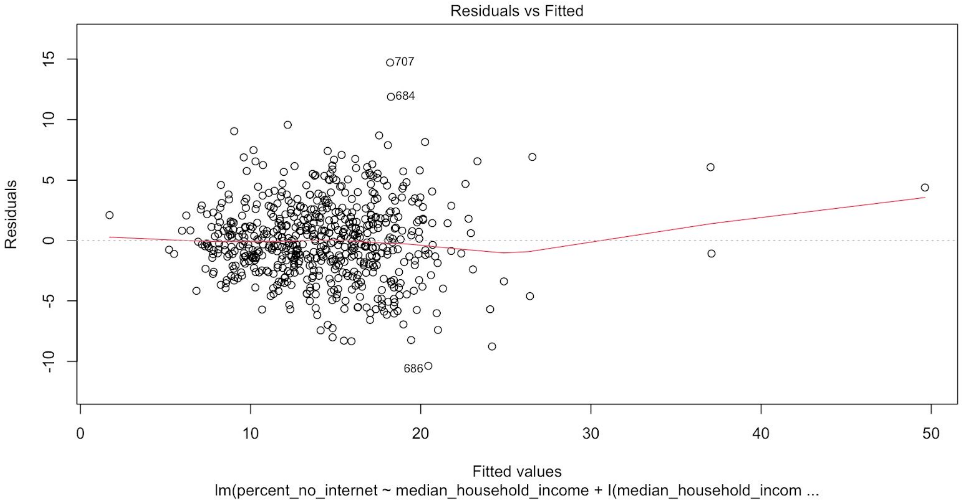


The correlation coefficients shown in the top right of the matrix above indicates there is little correlation between the predictors–the median household income, native population proportion, and proportion of native residents with an education level of some college. Additionally, the histograms featured diagonally in the matrix show the distribution of each predictor variable. The median household income and proportion with some colleges both have distributions that are skewed slightly to the right, and the Native American population proportion distribution is skewed drastically to the right. Finally, the bivariate scatterplot in the bottom left of the matrix shows that there is little relationship between the variables.

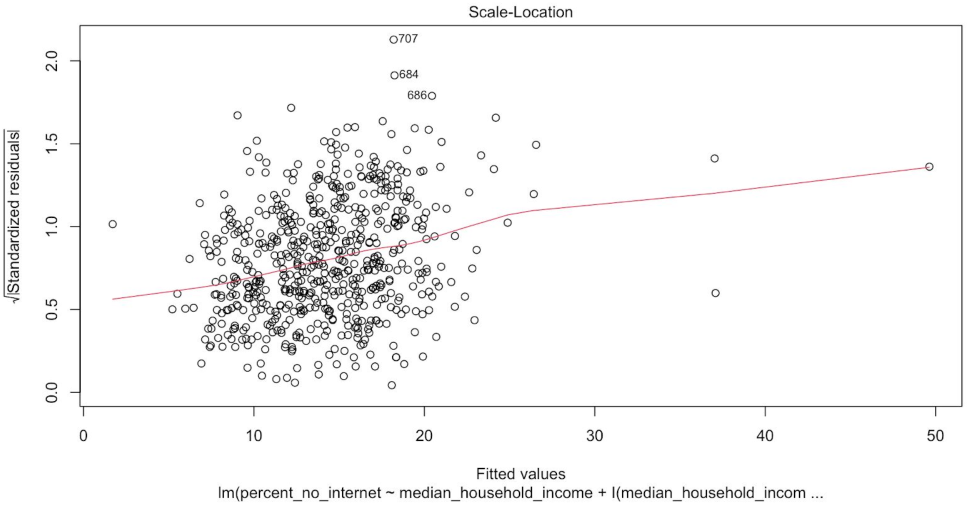
Residual Analysis



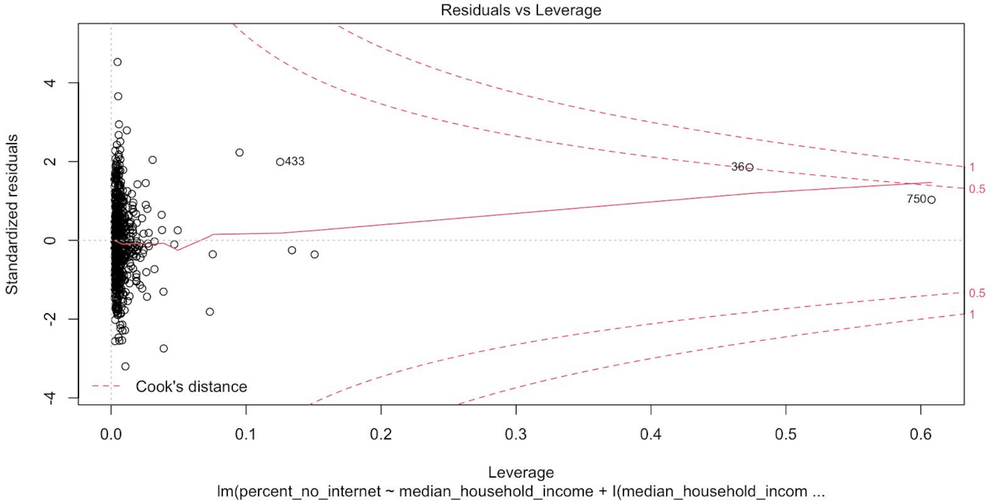
The QQ-plot indicates that the residuals are relatively normally distributed.



The Residuals vs. Fitted plot confirms that the residuals are normally distributed.



The Scale-Location plot shows that the residuals are not correlated with each other.



The Residuals vs. Leverage plot indicates potential outliers. The most significant outlier is case 36 which is Apache County in Arizona. This county is an outlier because of its high native population proportion and high percentage of households without internet access. Additionally, the 750th county in the dataset is Loudoun County in Virginia, which has one of the highest median income levels of all counties.

R Code

```
View(kaggle_internet)

kaggle_internet<- cbind(kaggle_internet,race_total = rowSums(kaggle_internet[,c(12, 13, 14, 15, 16, 17)]))
```

```
kaggle_internet<- cbind(kaggle_internet,educ_total = rowSums(kaggle_internet[,c(18, 19, 20, 21, 22, 23)]))

attach(kaggle_internet)

View(race_total)
View(educ_total)


kaggle_internet$prop_white = P_white/race_total
kaggle_internet$prop_black = P_black/race_total
kaggle_internet$prop_asian = P_asian/race_total
kaggle_internet$prop_native = P_native/race_total
kaggle_internet$prop_hawaiian = P_hawaiian/race_total
kaggle_internet$prop_other = P_others/race_total

kaggle_internet$P_white <- NULL
kaggle_internet$P_black <- NULL
kaggle_internet$P_asian <- NULL
kaggle_internet$P_native <- NULL
kaggle_internet$P_hawaiian <- NULL
kaggle_internet$P_others <- NULL


kaggle_internet$prop_middleschool_below = P_below_middle_school/educ_total
kaggle_internet$prop_somehighschool = P_some_high_school/educ_total
kaggle_internet$prop_highschooleq = P_high_school_equivalent/educ_total
kaggle_internet$prop_somecollege = P_some_college/educ_total
kaggle_internet$prop_bachelor_above = P_bachelor_and_above/educ_total

kaggle_internet$P_below_middle_school <- NULL
kaggle_internet$P_some_high_school <- NULL
kaggle_internet$P_high_school_equivalent <- NULL
kaggle_internet$P_some_college <- NULL
kaggle_internet$P_bachelor_and_above <- NULL

kaggle_internet$prop_belowpoverty = P_below_poverty/P_total
kaggle_internet$P_below_poverty <- NULL

View(kaggle_internet)

internet = kaggle_internet[, -c(1, 2, 3, 4, 5, 12, 13)]
summary(na.omit(internet))

internet = na.omit(internet)
dim(internet)

View(internet)

save(internet, file = "internet.rdata")

## CREATING CORRELATION MATRIX

education = sum(internet[, c(13,14,15,16,17)])
View
```

```
View(education)

cor(internet) >= .50

## MODELING W/ BACKWARD SELECTION
model1 = lm(percent_no_internet ~ . , data = internet)
summary(model1)
#R2adj = 0.7316

model2 = lm(percent_no_internet ~ . -prop_white, data = internet)
summary(model2)
#R2adj = 0.732

model3 = lm(percent_no_internet ~ . - prop_white - gini_index, data = internet)
summary(model3)
#R2adj = 0.7325

model4 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other, data = internet)
summary(model4)
#R2adj = 0.7329

model5 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - prop_somehighschool, data = internet)
summary(model5)
#R2adj = 0.7323

model6 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - prop_somehighschool - prop_asian, data = internet)
summary(model6)
#R2adj = 0.7322

model7 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - prop_somehighschool - prop_asian -
prop_hawaiian, data = internet)
summary(model7)
#R2adj = 0.7324

model8 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - P_total - prop_asian - prop_somehighschool -
prop_hawaiian, data = internet)
summary(model8)
#R2adj = 0.732
#* all co-effs are stat sig. at 0.01 level

model9 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - P_total - prop_asian - prop_somehighschool -
prop_hawaiian - prop_bachelor_above, data = internet)
summary(model9)
#R2adj = 0.7283

model10 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - P_total - prop_asian - prop_somehighschool -
prop_hawaiian - prop_bachelor_above - prop_black, data = internet)
summary(model10)
#R2adj = 0.7156

model11 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - P_total - prop_asian - prop_somehighschool -
prop_hawaiian - prop_bachelor_above - prop_black - prop_middleschool_below, data = internet)
summary(model11)
#R2adj = 0.7026
```

```
model12 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - P_total - prop_asian - prop_somehighschool -
prop_hawaiian - prop_bachelor_above - prop_black - prop_middleschool_below - median_age, data = internet)
summary(model12)
#R2adj = 0.6846

model13 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - P_total - prop_asian - prop_somehighschool -
prop_hawaiian - prop_bachelor_above - prop_black - prop_middleschool_below - median_age - median_rent_per_income, data =
internet)
summary(model13)
#R2adj = 0.667

model14 = lm(percent_no_internet ~ . - prop_white - gini_index - prop_other - P_total - prop_asian - prop_somehighschool -
prop_hawaiian - prop_bachelor_above - prop_black - prop_middleschool_below - median_age - median_rent_per_income -
prop_belowpoverty, data = internet)
summary(model14)
AIC(model14)
#R2adj = 0.65
## all parameters are significant with p-values < 2e-16

## checking for multicolinearity
install.packages("car")
library(car)
car::vif(model14)

## checking for correlation between predictors
cor(internet[, c(4, 10, 15, 16)])
summary(model16)
plot(model16)

## because there is a correlation between prop_somecollege and prop_highschooleq,
## testing models using only one of the two predictors

model15 = lm(percent_no_internet ~  median_household_income + prop_native + prop_somecollege, data = internet)
summary(model15)
#R2adj = 0.5862
AIC(model15) #3219.61

model16 = lm(percent_no_internet ~  median_household_income + prop_native + prop_highschooleq, data = internet)
summary(model16)
#R2adj = 0.5816
AIC(model16) #3226.221

model17 = lm(percent_no_internet ~ median_household_income + prop_native, data = internet)
summary(model17)
#R2adj = 0.5782
AIC(model17) #3230.229


# =============LOOKING AT POLYNOMIAL REGRESSIONS====================
model18 = lm(percent_no_internet ~ median_household_income + I(median_household_income^2)+ prop_native, data =
internet)
summary(model18)
AIC(model18) #3170.708
```

```
#R2adj = 0.6183


model19 = lm(percent_no_internet ~ median_household_income + I(median_household_income^2)+ prop_native +
prop_somecollege, data = internet)
summary(model19)
AIC(model19) #3166.36
#R2adj = 0.6217

model19b = lm(percent_no_internet ~ median_household_income + I(median_household_income^2)+ prop_native +
prop_somecollege + I(prop_somecollege^2), data = internet)
summary(model19b)
AIC(model19b) #3167.94
#R2adj = 0.6213

model20 = lm(percent_no_internet ~ median_household_income + I(median_household_income^2)+
I(median_household_income^3) + prop_native + prop_somecollege, data = internet)
summary(model20)
AIC(model20) #3154.071
#R2adj = 0.6299

model21 = lm(percent_no_internet ~ median_household_income + I(prop_native^2)+ prop_native, data = internet)
summary(model21)
AIC(model21) #3220.798
#R2adj = 0.5854

model22 = lm(percent_no_internet ~  median_household_income + I(median_household_income^2) + prop_native +
prop_highschooleq + I(prop_highschooleq^2), data = internet)
summary(model22)
#R2adj = 0.6263

# ================= WINNER =================

model20 = lm(percent_no_internet ~ median_household_income + I(median_household_income^2)+
I(median_household_income^3) + prop_native + prop_somecollege, data = internet)
summary(model20)
AIC(model20) #3154.071
#R2adj = 0.6299

#=============== RESIDUALS =====================
plot(model20)
cor(internet[, c(4, 10, 16)]) #we can say predictors are not related

install.packages('psych')
library(psych)

pairs.panels(internet[, c(4, 10,16)],
        method = "pearson", # correlation method
        hist.col = "#00AFBB",
        density = T,  # show density plots
        ellipses = T # show correlation ellipses
        )
```