Predicting Per Capita Income by County (2021)

STATS 101A, Summer 2023

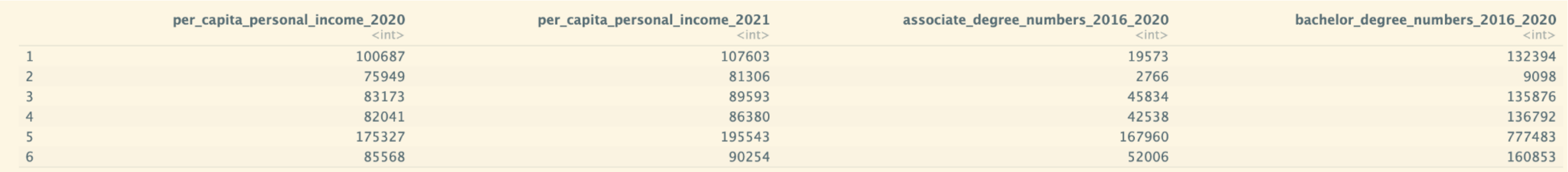
Professor Cha

Presented by: Vanda Suklar, Janice Chung, Neha Jonnalagadda, Liandong Zou, Gelare modiri

INTRODUCTION

a. Background

We are interested in modeling the Per Capita Income by County (2021) given a variety of factors that are believed to contribute to its value. Throughout this report we will verify that a linear model is appropriate for predicting the output and that all variables in the dataset should be considered as part of the final model. The chart below shows the first 6 rows of our data set. The first column is Per Capita Income by County (2020). The second column is our response variable we are predicting, Per Capita Income by County (2021). The third column is the sum of associate degrees in the county from 2016 to 2020. The fourth column is the sum of bachelor degrees in the county from 2016 to 2020. The first, third, and fourth columns (column 1, 3, 4) make up our three predictor variables for the model.



	per_capita_personal_income_2020	per_capita_personal_income_2021	associate_degree_numbers_2016_2020	bachelor_degree_numbers_2016_2020
	<int>	<int>	<int>	<int>
1	100687	107603	19573	132394
2	75949	81306	2766	9098
3	83173	89593	45834	135876
4	82041	86380	42538	136792
5	175327	195543	167960	777483
6	85568	90254	52006	160853

b. Source of Data

The United States per capita personal income by county is from bea.gov, the U.S. Bureau of Economic Analysis. The education level data is from usda.gov, the U.S. Department of Agriculture.

c. Overview of Paper

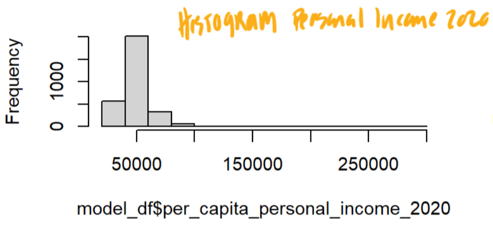
This paper will start by providing a preliminary analysis of the data including summary statistics and a visual representation of the relationship amongst the variables, including their correlation. Next it will report the linear model used to predict and analyze the relationship between per capita income for 2020, total associate degrees, total bachelor degrees, and per capita income of the year 2021. In our endeavor to attempt to understand if any relationship exists among these variables, several models were attempted, but only the best fitted model was utilized. The assessment of our validity test and goodness of fit will be reported and delved into further. Lastly, an overall discussion and limitations of the analysis.

DATA DESCRIPTION

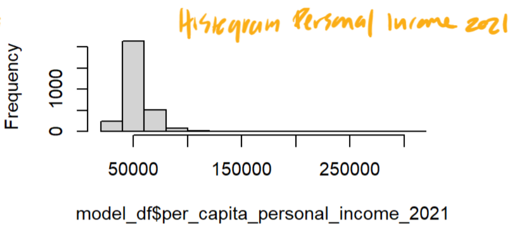
a. Summary Statistics

As a preliminary step to our research, we computed and analyzed summary statistics about the data, and particularly about the variables that our model computation will be based on. Upon careful analysis of these variables we can gain a better understanding of their general characteristics and behavior.

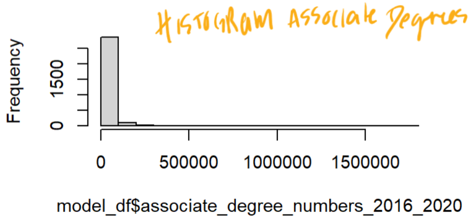
Variables	Mean	Standard Deviation
Per capita personal income for the year 2021	53308.48	14690.29
Per capita personal income for the year 2020	49513.36	13592.47
Total Associate degrees per county, from 2016 to 2020	21363.98	65240.51
Total Bachelors Degrees per county, from 2016 to 2020	24012.55	87380.29



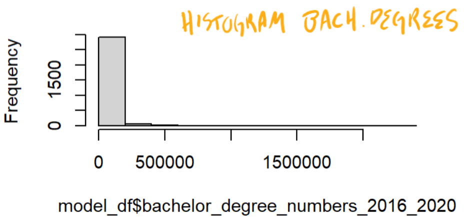
Histogram Personal Income 2020



Histogram Personal Income 2021



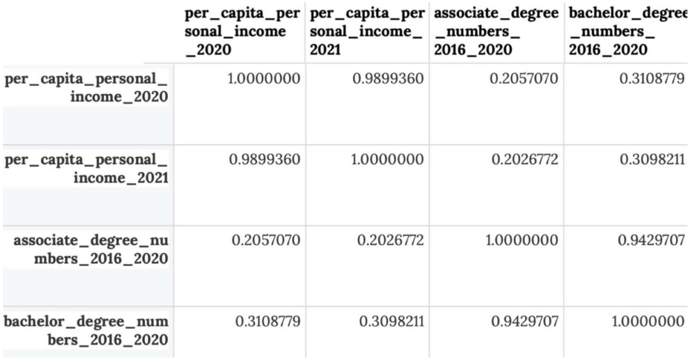
Histogram Associate Degrees



HISTOGRAM BACH. DEGREES

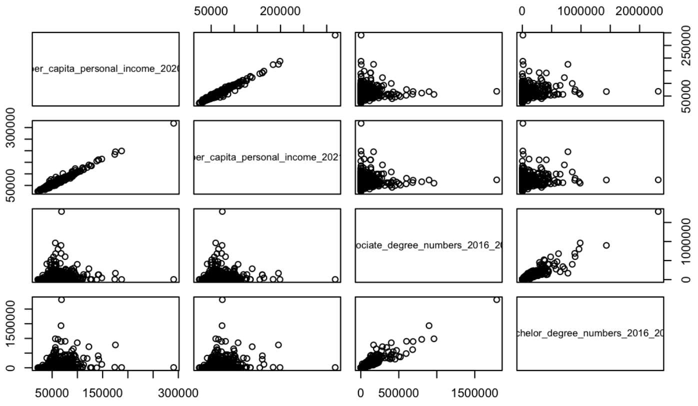
Relationship Amongst the Variables

Visually the variables relate to each other, as can be seen in the table of correlations coefficients and the scatter plot printed below.



	per_capita_personal_income_2020	per_capita_personal_income_2021	associate_degree_numbers_2016_2020	bachelor_degree_numbers_2016_2020
per_capita_personal_income_2020	1.0000000	0.9899360	0.2057070	0.3108779
per_capita_personal_income_2021	0.9899360	1.0000000	0.2026772	0.3098211
associate_degree_numbers_2016_2020	0.2057070	0.2026772	1.0000000	0.9429707
bachelor_degree_numbers_2016_2020	0.3108779	0.3098211	0.9429707	1.0000000

Scatter Plot of Variables



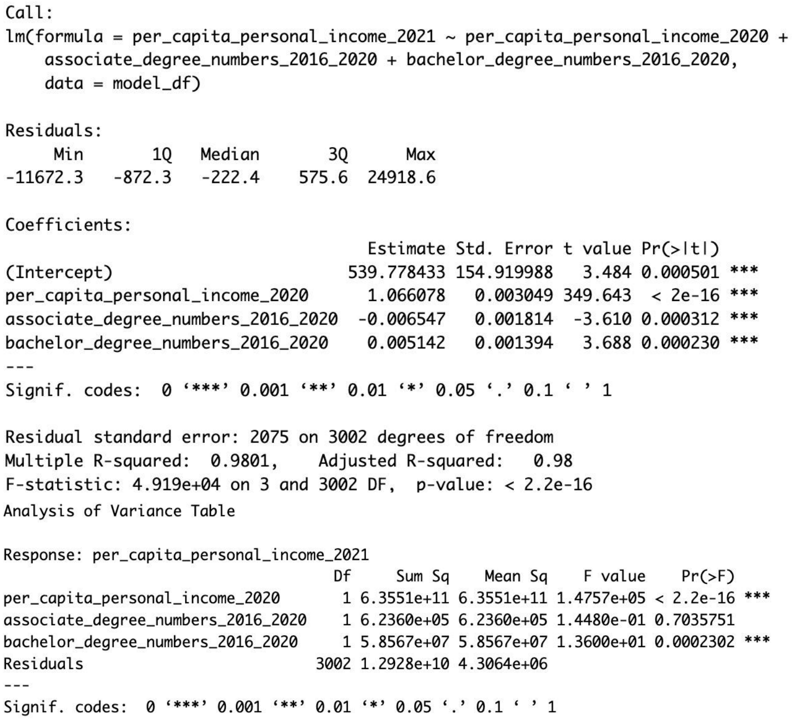
RESULTS AND INTERPRETATION

Goodness

We first check if Per Capita Income by County (2021) is linearly associated with the predictors.

As the scatter plot matrix shown in our preliminary analysis, Per Capita Income by County (2021) is somewhat linearly associated with the predictor Per Capita Income by County (2020), but the other two predictors: the sum of associate degrees in the county from 2016 to 2020, and the sum of bachelor degrees in the county from 2016 to 2020 don’t affect it much.

To take a closer look at the relationships we model per capita income of 2021 in terms of the other three variables.

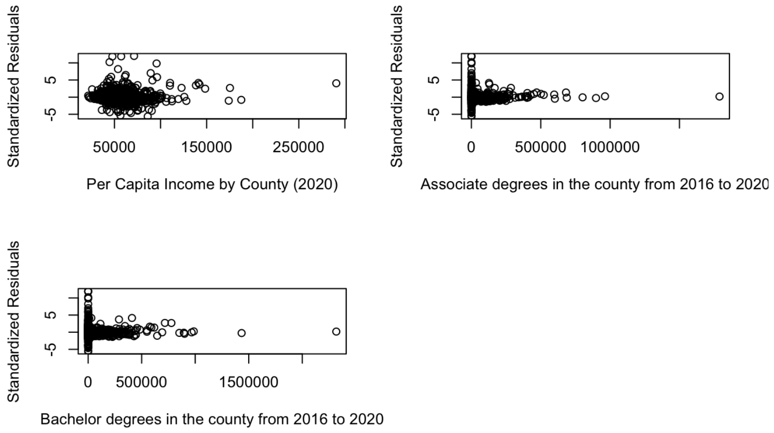


```
Call:
lm(formula = per_capita_personal_income_2021 ~ per_capita_personal_income_2020 +
    associate_degree_numbers_2016_2020 + bachelor_degree_numbers_2016_2020,
    data = model_df)

Residuals:
     Min       1Q   Median       3Q      Max
-11672.3   -872.3   -222.4    575.6  24918.6

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                        539.778433 154.919988   3.484 0.000501 ***
per_capita_personal_income_2020      1.066078   0.003049 349.643  < 2e-16 ***
associate_degree_numbers_2016_2020  -0.006547   0.001814  -3.610 0.000312 ***
bachelor_degree_numbers_2016_2020    0.005142   0.001394   3.688 0.000230 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2075 on 3002 degrees of freedom
Multiple R-squared:  0.9801,	Adjusted R-squared:   0.98
F-statistic: 4.919e+04 on 3 and 3002 DF,  p-value: < 2.2e-16
Analysis of Variance Table

Response: per_capita_personal_income_2021
                                     Df     Sum Sq    Mean Sq    F value    Pr(>F)
per_capita_personal_income_2020       1 6.3551e+11 6.3551e+11 1.4757e+05 < 2.2e-16 ***
associate_degree_numbers_2016_2020    1 6.2360e+05 6.2360e+05 1.4480e-01 0.7035751
bachelor_degree_numbers_2016_2020     1 5.8567e+07 5.8567e+07 1.3600e+01 0.0002302 ***
Residuals                          3002 1.2928e+10 4.3064e+06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In our first attempted model, we fit the multiple linear regression model with all three predictors. All p-values are less than 0.05, so the slope is significant, and the variables have a significant linear association. Also, the multiple R-squared is close to 1 which indicates that the model explains all the variability of the response data around its mean. However, the p-value of associate degree number income 2016-2020 in anova is larger than 0.05 so we would like to see if this model could be further improved.

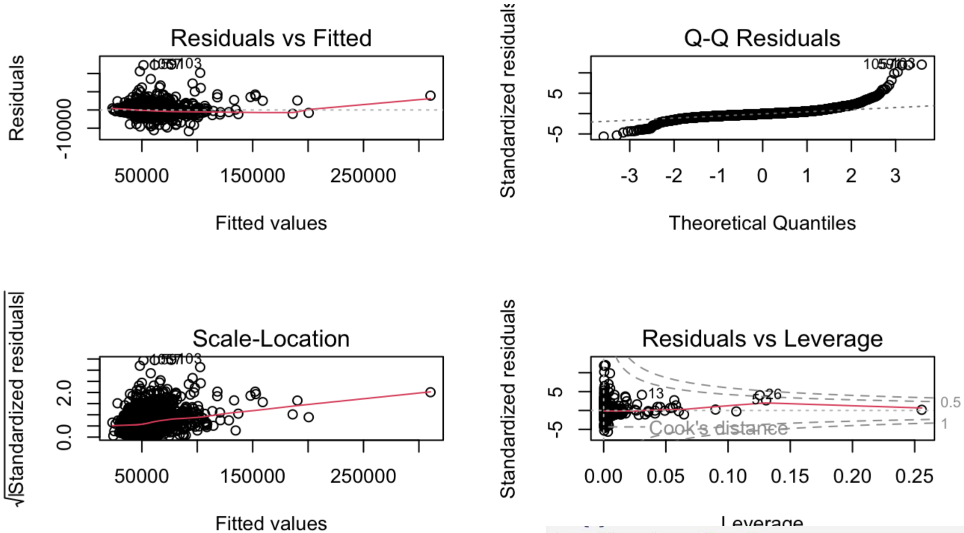
Validity

To check the model assumptions, we plot the standardized residual vs. each of the three predictors.



A random scatter of points around the horizontal axis, signify the mean function of the e_i is zero when a correct model has been fit. Thus, the plots of standardized residuals vs. three predictors show that the model is valid.

Additionally we considered the four diagnostic plots below:

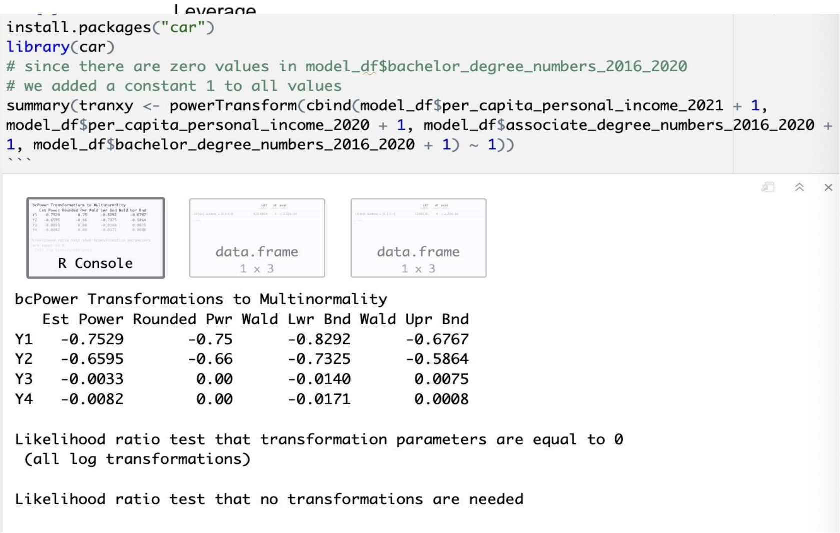


In the plot of Residual vs. Fitted, the red line shows the relationship is linear but not very straight, and the average of the error is close to 0.

The points in the Normal Q-Q plot are aligned to the straight line, but there are still a lot of outliers.

The standardized residual plot indicates a violation of the constant variance of the residuals assumption.

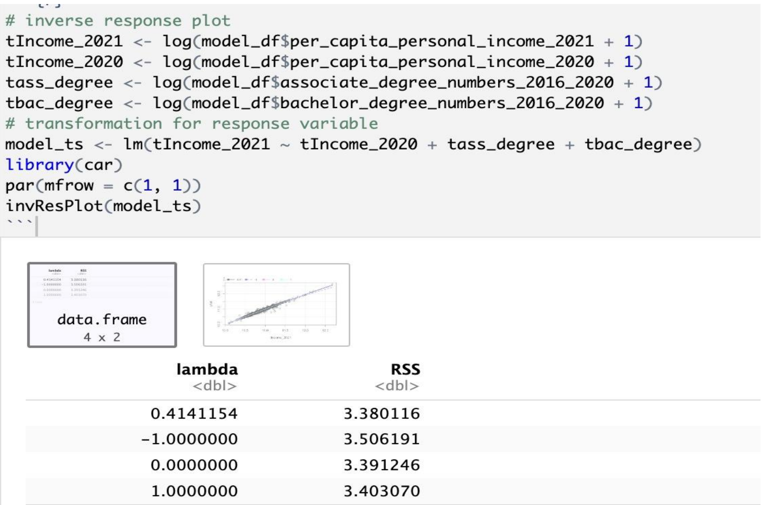
Considering Other Models



```
install.packages("car")
library(car)
# since there are zero values in model_df$bachelor_degree_numbers_2016_2020
# we added a constant 1 to all values
summary(tranxy <- powerTransform(cbind(model_df$per_capita_personal_income_2021 + 1,
model_df$per_capita_personal_income_2020 + 1, model_df$associate_degree_numbers_2016_2020 +
1, model_df$bachelor_degree_numbers_2016_2020 + 1) ~ 1))
```

```
bcPower Transformations to Multinormality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.7529       -0.75      -0.8292      -0.6767
Y2   -0.6595       -0.66      -0.7325      -0.5864
Y3   -0.0033        0.00      -0.0140       0.0075
Y4   -0.0082        0.00      -0.0171       0.0008

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

The diagnostic output above demonstrates the original model could be improved . Several approaches were tested in search of the appropriate model. Firstly, a log transformation to the independent variables was applied. This approach slightly improved the violation of linearity. Secondly, we applied log transformation to the dependent variable exclusively. This approach also improved the random scatter of the residuals vs.predicted values, thus further addressing our violation of linearity. Another approach pursued was adding the constant 1 to all the values to address the zero values. Lastly, we applied the log transformation consistently to all the variables. This approach produced the best response in our model diagnostics. To begin the process we used the box-cox method to find and verify the transformation for the predictors.



```
# inverse response plot
tIncome_2021 <- log(model_df$per_capita_personal_income_2021 + 1)
tIncome_2020 <- log(model_df$per_capita_personal_income_2020 + 1)
tass_degree <- log(model_df$associate_degree_numbers_2016_2020 + 1)
tbac_degree <- log(model_df$bachelor_degree_numbers_2016_2020 + 1)
# transformation for response variable
model_ts <- lm(tIncome_2021 ~ tIncome_2020 + tass_degree + tbac_degree)
library(car)
par(mfrow = c(1, 1))
invResPlot(model_ts)
```

lambda <dbl>	RSS <dbl>
0.4141154	3.380116
-1.0000000	3.506191
0.0000000	3.391246
1.0000000	3.403070

Since all lambdas imply small RSS, we took lambda = 0 which is log transformation, and our final regression model is:

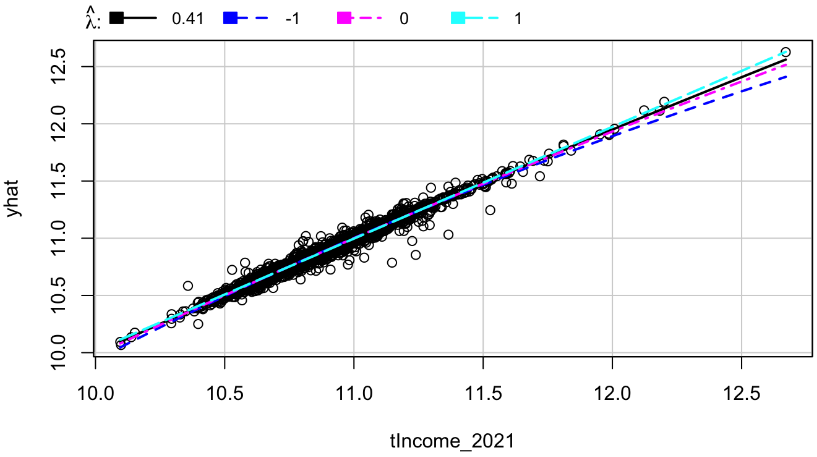
log(model_df$per_capita_personal_income_2021 + 1) = 0.199124 + 0.989634 * log(model_df$per_capita_personal_income_2020 + 1) + 0.005902 *

log(model_df$associate_degree_numbers_2016_2020 + 1) - 0.007765 *
log(model_df$bachelor_degree_numbers_2016_2020 + 1)

Finally, we got the numerical measures of the final model and p-value of ANOVA:

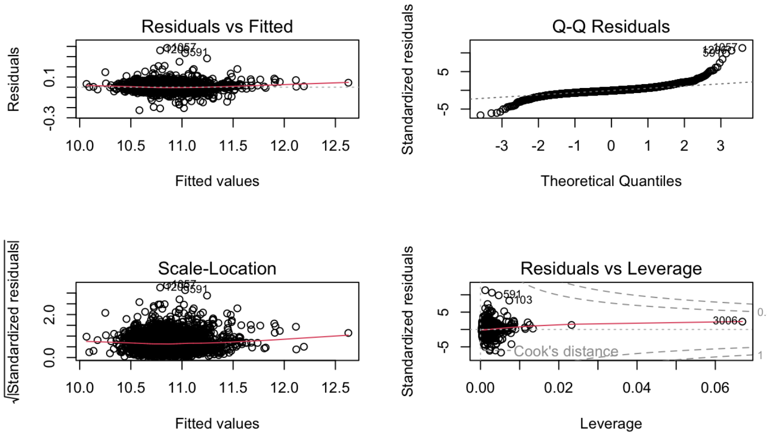
According to the slope estimates of tIncome_2020, we estimate that for every 1% increase in per capita personal income in 2020, there will be approximately a 0.99 increase in per capita personal income in 2021. Since the slope estimates of the other two predictors are very small and close to zero, they don't really affect our response variable.

They make sense because we first assumed that Per Capita Income by County (2021) is somewhat linearly associated with the predictor Per Capita Income by County (2020), but the other two predictors: the sum of associate degrees in the county from 2016 to 2020, and the sum of bachelor degrees in the county from 2016 to 2020 don’t affect it much.



The numerical output suggests that the model is significant. Also, the diagnostic plots for the final full model below suggests meaningful improvement in model assumptions.

tIncome p.value = <2.2e-16 tass_degree p.value = 4.424e-10 tbac_degress p.value = 0.0001166



SUMMARY

The project aims to predict the per capita income of counties in the United States for the year 2021 based on a set of predictor variables, including the per capita income for the year 2020, the sum of associate degrees in the county from 2016 to 2020, and the sum of bachelor degrees in the county from 2016 to 2020. The data is obtained from official sources, such as the U.S. Bureau of Economic Analysis and the U.S. Department of Agriculture. The project starts with a preliminary analysis that includes calculating summary statistics and visualizing the relationships between the variables, including their correlation.

In their initial multiple linear regression model, they find that the per capita income for 2021 is somewhat linearly associated with the per capita income for 2020, but the other two predictors, the sum of associate degrees, and the sum of bachelor degrees, do not have a significant impact on the response variable. To improve the model, we explore different transformations for the predictor variables and decide to use the log transformation (Box-Cox) for all three predictors. The final regression model is log(model_df$per_capita_personal_income_2021 + 1) = 0.199124 + 0.989634 * log(model_df$per_capita_personal_income_2020 + 1) + 0.005902 * log(model_df$associate_degree_numbers_2016_2020 + 1) - 0.007765 * log(model_df$bachelor_degree_numbers_2016_2020 + 1)

The final model's numerical measures and ANOVA p-value show that the per capita income for 2020 is a significant predictor of the per capita income for 2021. A 1% increase in 2020's per capita income is associated with approximately a 0.99 increase in 2021's per capita income. However, the other two predictor variables have minimal impact on the response variable.

The result that the previous year's per capita income is a strong predictor of the current year's per capita income aligns with economic intuition. Economic growth and development are often gradual processes, and the income level of a region tends to have a significant carryover effect from one year to the next. Higher income in one year usually translates to higher income in the following year due to factors like job stability, business growth, and investment patterns.

The finding that educational attainment variables, such as the number of associate and bachelor degrees, have limited impact on per capita income might also make sense. While education is generally associated with higher income and improved economic outcomes, the model's result might indicate that the variation in educational attainment levels among counties is not substantial enough to significantly influence per capita income variations independently. Also it's essential to consider that other factors not included in the model could also play significant roles in determining per capita income in different counties.

To ensure the model's accuracy and practicality in a real-world situation, ongoing validation and testing against new data are crucial. Economic conditions and influencing factors can change over time, so regularly updating and refining the model based on the latest data and economic trends will enhance its reliability for policy planning and decision-making.

DISCUSSION

There are a couple limitations to consider, such as the limited predictor variables, assumption of linearity, temporal assumptions, and geographical variability. In terms of limited predictor variables, the analysis primarily focuses on a small set of predictor variables, including per capita income for the previous year and educational attainment levels. While these variables are important, they might not capture all the factors that influence per capita income variations among counties. In the future, expanding the set of predictor variables to include other socio-economic indicators (e.g., unemployment rate, industry composition, cost of living) could enhance the model's explanatory power.

Another limitation was the assumption of linearity, and even though the paper addresses violations of linearity through log transformations, there might still be underlying non-linear relationships between variables that the transformations do not fully capture. Exploring more flexible regression techniques, such as polynomial regression or spline functions, could improve the model's ability to capture non-linear associations.

In terms of temporal assumptions, the analysis assumes that the relationships between predictor variables and per capita income are consistent over time. Economic conditions and influencing factors can change, so considering time-varying relationships or incorporating time-series analysis techniques could provide more accurate predictions.

Finally, geographic variability is lacking as well, since all counties in the United States are treated as a homogeneous group. However, there could be significant regional variations that impact per capita income differently. Incorporating spatial analysis techniques, such as spatial autocorrelation or considering regional clusters, could help account for spatial dependencies.