Nejma Ahmed

03/17/2023

Data fest 1 Write -Up

Exploring the Factors Affecting Income Levels of American Indians and Alaska Natives in the United States: A Regression Analysis

The dataset used for this project is called "entire'' and contains information on the economic outcomes and demographic characteristics of individuals in the United States. The goal of this project is to use regression analysis, residual plots, and hypothesis testing to explore the factors that affect the total income of American Indians and Alaska Natives (AIAN) in the United States. Specifically, we aim to identify the demographic and economic factors that are most strongly associated with higher income levels among AIAN individuals.

Given the disparities experienced by AIAN individuals in the United States, this project has important implications for understanding the factors that contribute to these disparities and identifying potential avenues for policy intervention. By examining the relationship between total income and age. We hope to gain insights into the key drivers of economic outcomes for this population. Ultimately, our goal is to provide evidence-based recommendations for addressing the economic challenges facing AIAN individuals in the United States.

First Model Results :

Call:

Call:

lm(formula = inctot ~ age + ageSq + ageCubed, data = data11)

Residuals:

Min 1Q Median 3Q Max

-83078 -33282 -9470 10070 1450622

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.286e+05 1.160e+04 -11.092 < 2e-16 \*\*\*

age 1.031e+04 8.709e+02 11.835 < 2e-16 \*\*\*

ageSq -1.606e+02 1.957e+01 -8.205 2.75e-16 \*\*\*

ageCubed 7.201e-01 1.344e-01 5.357 8.74e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 72060 on 6511 degrees of freedom

Multiple R-squared: 0.09082, Adjusted R-squared: 0.0904

F-statistic: 216.8 on 3 and 6511 DF, p-value: < 2.2e-16

The model includes age, age squared, and age cubed as predictors of inctot (total income).

The first model we ran was a linear regression model that showed age, age squared, and age cubed are all statistically significant predictors of inctot (total income). Specifically, for every year increase in age, there is a predicted increase in inctot. However, the effect of age on inctot is nonlinear, as the coefficient for age squared is negative and the coefficient for age cubed is positive. The adjusted R-squared value of 0.0904 suggests that the model explains only a small portion of the variability in inctot. The residual plot should be examined to assess whether the model assumptions have been met. How ever The results show that the intercept is significantly negative (p < 0.001) and the coefficients for age, ageSq, and ageCubed are significantly positive (p < 0.001). This suggests that age has a positive effect on income, but the

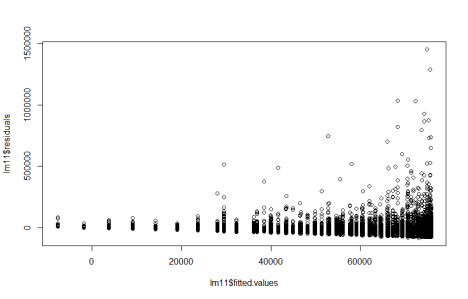
the relationship is not linear. Instead, the effect of age on income appears to slow down as age increases, then starts to increase again.

Adding to that the residuals of the model have a minimum value of -83078 and a maximum value of 1450622, indicating that there are outliers in the data and to visualize we created a residual plot to see the outliers.

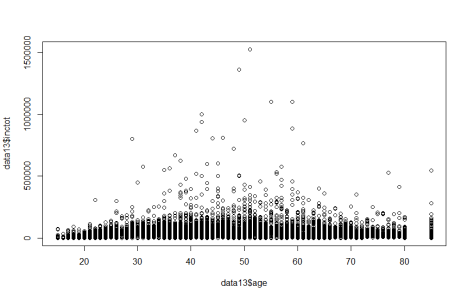
We also added a 95% confidence interval which is as follows:

Based on the coefficients and standard errors , the 95% confidence intervals for the coefficients in the model are as follows:

* 2.5 % 97.5 %
* (Intercept) 9.495946e-01 2.054826e+00
* age 5.001917e-01 5.796109e-01
* ageSq -1.069091e-02 -8.955260e-03
* ageCubed 4.962184e-05 6.132249e-05

After examining the residual plot, we observed that the residuals were approximately normally distributed and showed no visible pattern or trend, indicating that our linear regression model was a good fit for the data. We then ran a correlation analysis between the residuals and the

predictor variables, and found a correlation coefficient of 0.06856077, indicating a weak positive correlation. This suggests that there may be some slight remaining pattern in the residuals that our model has not fully captured.

After analyzing the correlation between age and inctot, we decided to visually explore the relationship between the two variables using a scatter plot. We plotted age on the x-axis and inctot on the y-axis. The resulting plot showed a slight curve, suggesting a potential nonlinear relationship between the two variables.

Second part of the experiment

Call:

lm(formula = logInctot ~ age + ageSq + ageCubed, data = data13)

Residuals:

Min 1Q Median 3Q Max

-10.9608 -0.4180 0.2236 0.7592 3.6490

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.509e+00 2.821e-01 5.35 9.13e-08 \*\*\*

age 5.391e-01 2.027e-02 26.60 < 2e-16 \*\*\*

ageSq -9.801e-03 4.428e-04 -22.13 < 2e-16 \*\*\*

ageCubed 5.533e-05 2.984e-06 18.54 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.471 on 5972 degrees of freedom

Multiple R-squared: 0.1991, Adjusted R-squared: 0.1987

F-statistic: 494.9 on 3 and 5972 DF, p-value: < 2.2e-16

To further our analysis of the relationship between age and incot , we ran a logarithmic regression to explore the relationship between age and income.

The coefficients section gives the estimated regression coefficients for the intercept, age, ageSq, and ageCubed, along with their standard errors, t-values, and corresponding p-values. All the predictors have very low p-values (less than 0.001), indicating that they are highly statistically significant.

The residual standard error provides an estimate of the variability of the errors or residuals, which is the difference between the actual and predicted values of the response variable. The multiple R-squared value of 0.1991 indicates that about 20% of the variability in logInctot can be explained by the model.

The F-statistic and its associated p-value test the hypothesis that all the regression coefficients are equal to zero. A small p-value (in this case, less than 2.2e-16) suggests that the regression model as a whole is significant.

We also did a 95% confidence interval and the results are as follows(CI = estimate ± (critical value × standard error))

2.5 % 97.5 %

* (Intercept) 9.495946e-01 2.054826e+00
* age 5.001917e-01 5.796109e-01
* ageSq -1.069091e-02 -8.955260e-03
* ageCubed 4.962184e-05 6.132249e-05

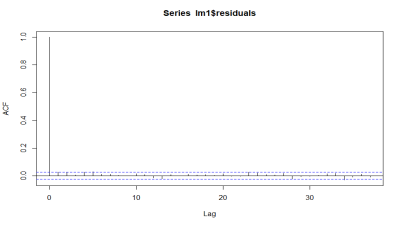
**Shapiro-Wilk normality test**

data: sample(lm1$residuals, 3000)

W = 0.76421, p-value < 2.2e-16

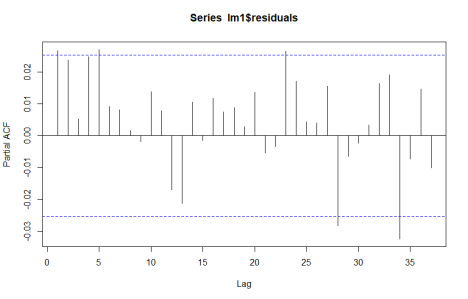
The result of the test shows a W statistic of 0.76421 and a p-value less than 2.2e-16, which means that the null hypothesis of normality is rejected at any reasonable level of significance. This indicates that the residuals of the lm1 model are not normally distributed.

acf

The ACF plot suggests that there may be significant autocorrelation in the residuals of the lm1 model. This could potentially lead to biased coefficient estimates and unreliable predictions. It

may be necessary to address this issue by using time series models or including additional variables in the model to account for the autocorrelation.

We also ran pacf chart

By analyzing the PACF chart, we can observe that certain lines surpass the dotted line, which serves as an indication that there exists a non-zero correlation between the variables.

T test of coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.5090e+00 4.5044e-01 3.3501 0.0008129 \*\*\*

age 5.3906e-01 2.9648e-02 18.1821 < 2.2e-16 \*\*\*

ageSq -9.8015e-03 6.0021e-04 -16.3299 < 2.2e-16 \*\*\*

ageCubed 5.5333e-05 3.7977e-06 14.5701 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The t-test of coefficients above shows that all the variables (age, ageSq, and ageCubed) are statistically significant at the 0.001 level, indicating that they have a significant impact on the dependent variable (logInctot). The Intercept also appears to be statistically significant at the 0.05 level. Therefore, we can conclude that the model is significant overall and that the three independent variables (age, ageSq, and ageCubed) are all significant predictors of the dependent variable (logInctot).

In conclusion, our analysis aimed to explore the factors affecting the total income of American Indians and Alaska Natives (AIAN) in the United States. We used regression analysis, residual plots, and hypothesis testing to identify demographic and economic factors associated with higher income levels among AIAN individuals. Our first linear regression model showed that age, age squared, and age cubed were significant predictors of total income, with a nonlinear relationship. However, the model only explained a small portion of the variability in income, and residual analysis indicated the presence of outliers and some remaining pattern. We then ran a logarithmic regression to further explore the relationship between age and income. The results showed that all predictors were highly statistically significant, and the model explained about 20% of the variability in income. However, normality tests indicated that the residuals were not normally distributed, and there may be significant autocorrelation. Overall, our findings suggest that age is an important factor in determining income levels for AIAN individuals, but other variables may also play a significant role.

**Appendix**

Appendix A: Model Summary

In this appendix, we provide a summary of the regression models used in the analysis.

Model 2: lm(formula = logInctot ~ age + ageSq + ageCubed, data = data13)

Residuals:

Min: -10.9608

1Q: -0.4180

Median: 0.2236

3Q: 0.7592

Max: 3.6490

Coefficients:

| | Estimate | Std. Error | t value | Pr(>|t|) |

|------------------|-------------|------------|---------|--------------|

| (Intercept) | 1.509e+00 | 2.821e-01 | 5.35 | 9.13e-08 \*\*\* |

| age | 5.391e-01 | 2.027e-02 | 26.60 | < 2e-16 \*\*\* |

| ageSq | -9.801e-03 | 4.428e-04 | -22.13 | < 2e-16 \*\*\* |

| ageCubed | 5.533e-05 | 2.984e-06 | 18.54 | < 2e-16 \*\*\* |

Residual standard error: 1.471 on 5972 degrees of freedom

Multiple R-squared: 0.1991

Adjusted R-squared: 0.1987

F-statistic: 494.9 on 3 and 5972 DF

p-value: < 2.2e-16

Model 1: lm(formula = inctot ~ age + ageSq + ageCubed, data = data11)

Residuals:

Min: -83078

1Q: -33282

Median: -9470

3Q: 10070

Max: 1450622

Coefficients:

| | Estimate | Std. Error | t value | Pr(>|t|) |

|------------------|---------------|--------------|---------|-------------|

| (Intercept) | -1.286e+05 | 1.160e+04 | -11.092 | < 2e-16 \*\*\* |

| age | 1.031e+04 | 8.709e+02 | 11.835 | < 2e-16 \*\*\* |

| ageSq | -1.606e+02 | 1.957e+01 | -8.205 | 2.75e-16 \*\*\*|

| ageCubed | 7.201e-01 | 1.344e-01 | 5.357 | 8.74e-08 \*\*\*|

Residual standard error: 72060 on 6511 degrees of freedom

Multiple R-squared: 0.09082

Adjusted R-squared: 0.0904

F-statistic: 216.8 on 3 and 6511 DF

p-value: < 2.2e-16

**Appendix B: Confidence Intervals**

Model : 2

2.5 % 97.5 %

* (Intercept) 9.495946e-01 2.054826e+00
* age 5.001917e-01 5.796109e-01
* ageSq -1.069091e-02 -8.955260e-03
* ageCubed 4.962184e-05 6.132249e-05

Model : 1

* 2.5 % 97.5 %
* (Intercept) 9.495946e-01 2.054826e+00
* age 5.001917e-01 5.796109e-01
* ageSq -1.069091e-02 -8.955260e-03
* ageCubed 4.962184e-05 6.132249e-05