The Highest Earner: Factors that affect personal earnings in the United States

KHL

remove the pound sign and insert the file path to the image on your computer in the

Introduction

- Does a higher level of educational attainment generally increase personal earnings income across different states?
- Do personal earnings increase with an individual's health?
- Do older individuals generally earn more money than younger individuals?

Given the rise in inflation and cost of living, exploring the relationships connected to personal earnings across different states is fundamental to understanding how factors like education, health status, and age, influence income disparities at a regional level. This inquiry is grounded in the longstanding debate within economic and social research regarding the return on investment in education, health, and wellness. By analyzing U.S. Census data, we can gain detailed insights into the nation's political and economic structures, examining how local economies, policies, and opportunities influence each community. This analysis underscores the importance of ensuring that every community receives its fair share of resources, tailored to its unique needs (Bureau, 2021). In addition, it is crucial to the political sphere with its use in redrawing a multitude of political boundaries to ensure each district contains roughly equal numbers of people thereby addressing funding disparities (Mather & Scommegna, 2019).

The dataset was merged on a state basis, focusing on individuals 18 and older to better represent the adult population. It includes averages of education level, gender, work expenses, and age from the ASEC survey, combined with unemployment rates from the Bureau of Labor Statistics and sales tax rates from the Tax Foundation at the state level.

Methods and Analysis

Before the model building process began, the data was first subsetted to exclude the observation corresponding to the District of Columbia. Exploratory data analysis revealed that it was the sole territory with the level of "Bachelor's Degree" as the response to the variable H_ED, highest educational attainment. As H_ED was believed to be a significant predictor in the model, this observation was excluded from the model to prevent skew or extrapolation. Following the removal of this observation, a histogram of the response, personal earnings, appeared unimodal with moderate right skew and with minimal outliers. The first stage of the model building process then began: fitting the model with quantitative predictors. Quantitative variables were examined for evidence of multicollinearity through correlative

plots; no concerning relationships were found, but multicollinearity will be reassessed in the final model. Scatterplots of each quantitative variable with the response showed varying degrees of association. Moderate to strong relationships with the response existed with four variables: unemployment rate, tax rate, work expenses, and urban percentage. These four variables were used to build a model, resulting in a globally significant model. The model was reduced via the performance of individual t-tests to include only the two most significant quantitative predictors, urban percentage and work expenses, which had p-values of 0.00879 and 0.01667, respectively.

Next, qualitative predictors were added to the model. The examination of boxplots of the three qualitative variable with the response suggested that alternative levels of sex does not significantly impact earnings; the median personal earnings of states with a predominantly female workforce is approximately equivalent to those with a predominantly male workforce. However, differing levels of education and health status suggested significant different responses to personal earnings, indicated by non-overlapping interquartile ranges of differing levels for these variables. Therefore, these two quantitative variables, highest education level and health status were added to the model in the second stage of the model building process. The proposed model demonstrated significance by the global f-test. Individual t-tests were then performed to build a model with two predictors: urban percentage and highest educational attainment. The quantitative work expenses predictor became insignificant upon that addition of qualitative predictors, and health status did not demonstrate individual significance. No interactions were believed to be influencing the model, but an interaction between the two remaining main effects within the model was explored. A grouped by scatterplot, plotting PEARNVAL as the response, URB_PER as the quantitative explanatory variable, and H_ED as the qualitative explanatory variable demonstrated no difference in slope when a regression line was plotted through the grouped points.

From this process, the equation of the proposed model is: $PEARNVAL = 47415.20 + 191.57URB\_PER - 6997.66H\_EDvocationalassociates$. An examination of VIFs suggested no concern for multicollinearity with an average VIF and highest individual of 1.103. Analysis of residuals showed a lack of fit with no obvious pattern, a mean of zero, and no fanning patterns, suggesting constant variance. Slight deviations from normality were observed in the qqplot of the residuals of the model, suggested by deviations from linearity on the tail ends of the distribution. Three transformations, logarithmic, exponential, and square root, were attempt to resolve this deviation; however, they did not resolve the violation of normality. The original model, with an untransformed response, was kept as the model is assumed robust to violations of normality due to its sufficient sample size. Examination of

outliers was performed and 4 observations were removed on the basis of having excessive leverage and Cook's distance.

With data subsetted to exclude outliers, a new final model is proposed: $PEARNVAL = 47408.52 + 191.95URB\_PER - 7017.40H\_EDvocationalassociates$. After verifying the model was trained on data of sufficient size, the external validation technique of data splitting was used to examine the model. The data was split randomly into two subsets of equal size, one used to estimate model parameters and the other used to assess the model's predictive ability. The results from cross-splitting suggested a generally poor predictive ability of the model, as will be discussed in subsequent sections.

Results

$PEARNVAL = 47408.52 + 191.95URB\_PER - 7017.40H\_EDvocationalassociates$

This model is statistically significant with a p-value < 0.0001 and an adjusted r-squared value of 0.56, indicating that this model accounts for 56% of the variation in the data. Additionally, The R^2 value of 0.5792 suggests that approximately 57.92% of the variance in personal earnings can be explained by urban percentile and higher education.

Conclusions

In our final model, urban percentile and level of higher education are the most significant variables for predicting personal earnings by state. Our model shows that increase in urban percentile (URB_PER) leads to higher personal earnings whereas possession of a higher education vocational associate's degree (H_EDvassoc) results in lower personal earnings in the United States.

When estimating the average personal earnings of Virginia given an URB_PER of 75.5 and H_EDVassoc of 0, the model predicted that the personal earnings would be $61,900.74; the actual earnings were $71,818.63. The prediction was off by $9,917.89, a percent error of 13.8%.

While urban environment and educational attainment do influence personal income, they may not be the most robust predictors. Additional variables or more complex models could be necessary to capture the full dynamics of affecting personal income and increase its accuracy. Further research could include other variables not analyzed by our data, such as family size or occupation. Additionally, since we were using 2020 Census data, that could

cause a skew towards high unemployment and decrease in personal earnings due to the state of the economy during the global pandemic. Thus, increasing our dataset to beyond 2020 can be beneficial at creating more observations. We could also modify existing variables by treating gender as a continuous proportion instead of a binomial variable. In all, this model is a good starting point for understanding the relationship between urban environment, educational attainment, and personal earnings, but further research is needed to create a stronger model.

Appendix A: Data Dictionary

Reference Name	Variable Name	Description
State by FIPS Code	`STATEFIPS`	A qualitative measure that identifies the U.S. state (or D.C.) corresponding to the observation by a standardized numeric code. The 51 possible levels are discrete, ranging from 1-56, omitting 3, 7, 14, 43, and 52.
State	`State`	A qualitative measure that identifies the state corresponding to the observation. The 51 possible levels are names of the 50 U.S. states and the District of Columbia.
Educational Attainment	`H_ED`	A qualitative measure that identifies the average of highest education among adult residents of a given state. The three possible levels include a Vocational Associate's Degree, an Academic Associate's Degree, and a Bachelor's Degree.
Majority Sex	`SEX`	A qualitative measure that identifies the predominant sex among a state's adult residents. Two possible levels, male and female, indicate if the adult population of a state is predominately male or female.
Health Status	`HEA`	A qualitative measure that reports the average health status of a state's residents. Two levels, very good health and good health indicate the average health status of a state's residents.
Personal Earnings	`PEARNVAL`	A continuous quantitative measure that reports the average personal earnings of a state's residents, reported in U.S. Dollars. Possible values within the data range from $45096.53 to $95387.40.
Age	`AGE`	A continuous quantitative measure that reports the average age of a state's adult residents in years. Values range from 40.83460 to 46.39759.

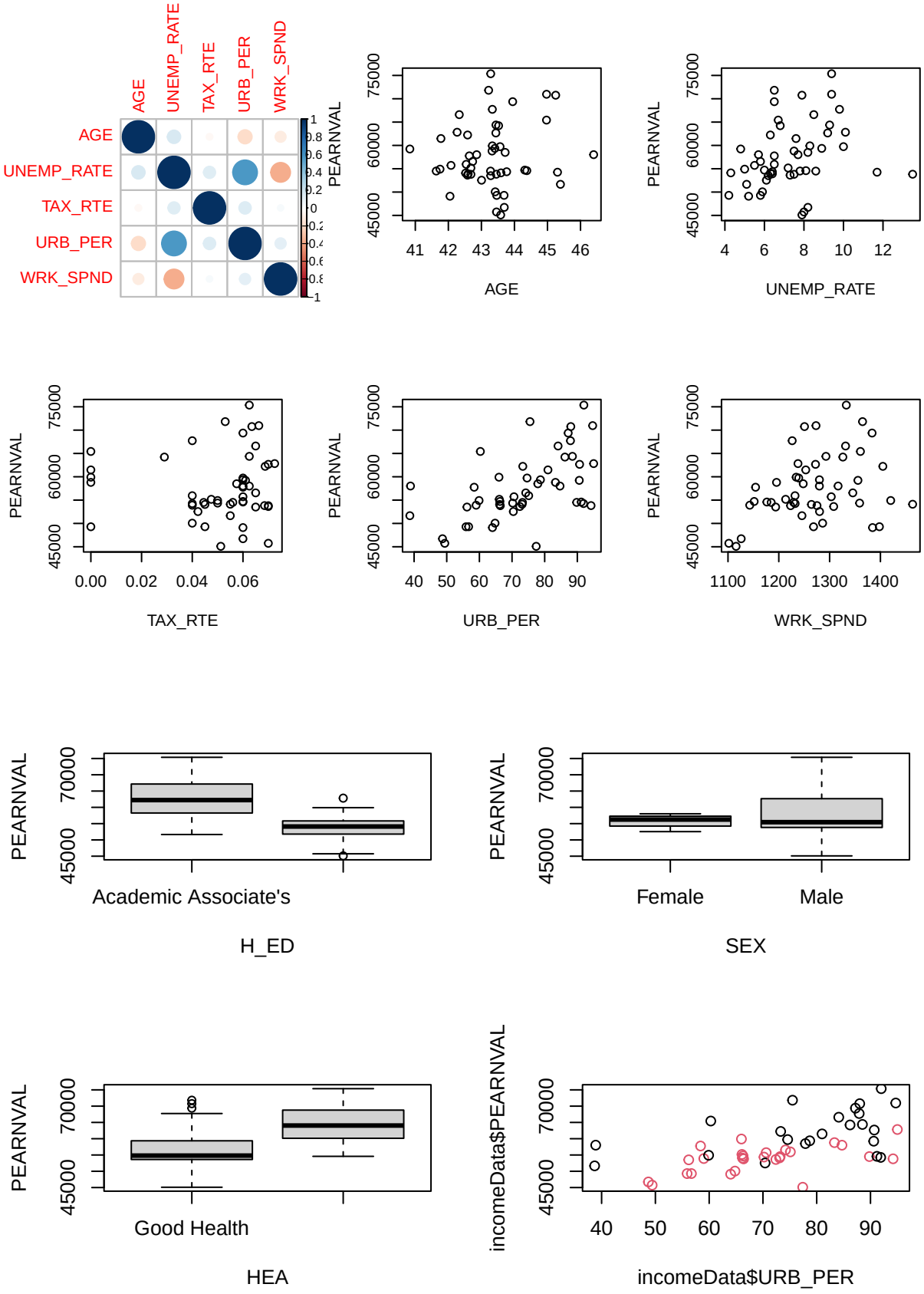
Reference Name	Variable Name	Description
Unemployment Rate	`UNEMP_RATE`	A continuous quantitative measure of a state's unemployment rate from 2020. Unemployment rate is reported as a percentage; the range of possible values within the data is from 4.2% to 13.5%.
Sales Tax Rate	`TAX_RTE`	A continuous quantitative measure of a state's sales tax. Sales Tax Rate is reported as a numerical figure; the range of possible values within the data is from 0.0% (0% sales tax) to 7.25% (7.25% sales tax).
Percentage of Urban Residents	`URB_PER`	A continuous quantitative measure of a state's proportion of urban residents to nonurban residents. This variable is reported as a percentage; the range of possible values within the data is from 38.7% to 100.0%.
Work Expenses	`WRK_SPND`	A continuous quantitative measure that identifies the average amount of money spent on work-related expenses among residents of a state, reported in U.S. Dollars. Possible values in the data range from $1101.676 to $1463.411.

Appendix B: Data Rows

	STATEFIPS	State	H_ED	SEX	HEA
1	1	Alabama	Vocational Associate's	Male	Good Health
2	2	Alaska	Vocational Associate's	Male	Good Health
3	4	Arizona	Vocational Associate's	Male	Good Health
4	5	Arkansas	Vocational Associate's	Male	Good Health
5	6	California	Vocational Associate's	Male	Good Health
6	8	Colorado	Academic Associate's	Male	Good Health
7	9	Connecticut	Academic Associate's	Male	Good Health
8	10	Delaware	Vocational Associate's	Male	Good Health
9	12	Florida	Academic Associate's	Male	Very Good Health
10	13	Georgia	Vocational Associate's	Female	Good Health
12	16	Idaho	Vocational Associate's	Male	Good Health
13	17	Illinois	Academic Associate's	Male	Good Health
14	18	Indiana	Vocational Associate's	Male	Good Health
15	19	Iowa	Vocational Associate's	Male	Good Health
16	20	Kansas	Vocational Associate's	Female	Good Health

	PEARNVAL	AGE	UNEMP_RATE	TAX_RTE	URB_PER	WRK_SPND
1	53905.05	42.60043	6.4	0.0400	59.0	1142.836
2	59908.18	43.33103	8.3	0.0000	66.0	1234.210
3	54509.31	41.63326	7.8	0.0560	89.8	1229.184
4	53513.54	43.28300	6.2	0.0650	56.2	1193.809
5	62824.72	42.26563	10.1	0.0725	95.0	1237.779
6	64224.86	43.52050	6.8	0.0290	86.2	1326.110
7	70758.66	45.24870	7.9	0.0635	88.0	1250.562
8	58795.20	43.32292	7.5	0.0000	83.3	1195.540
9	54585.91	44.37481	8.1	0.0600	91.2	1176.506
10	55946.86	42.54033	6.5	0.0400	75.1	1231.916
12	55717.66	42.08074	5.5	0.0600	70.6	1303.237
13	64375.88	43.44074	9.3	0.0625	88.5	1292.794
14	53621.19	42.57899	7.3	0.0700	72.4	1308.013
15	49106.65	42.04989	5.2	0.0600	64.0	1384.532
16	56551.28	42.73529	5.8	0.0650	74.2	1345.676

Appendix C: Tables and Figures

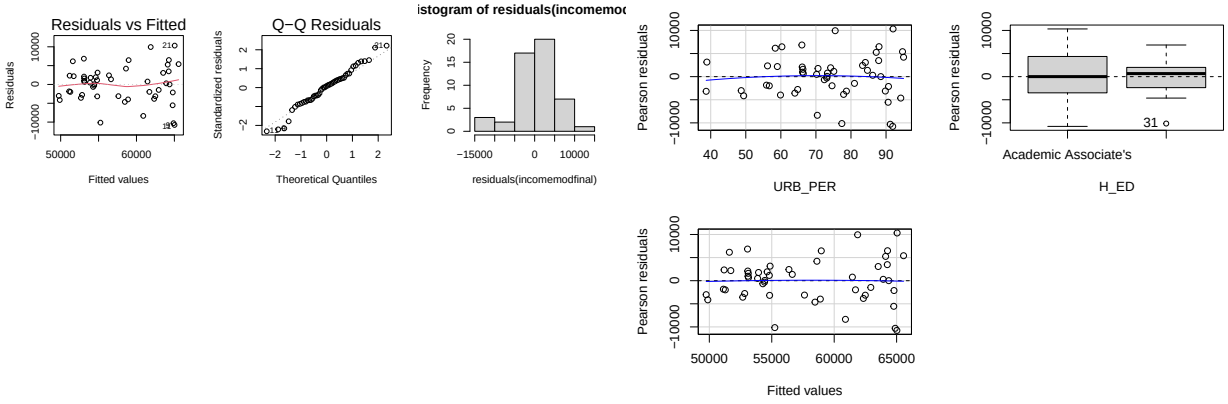


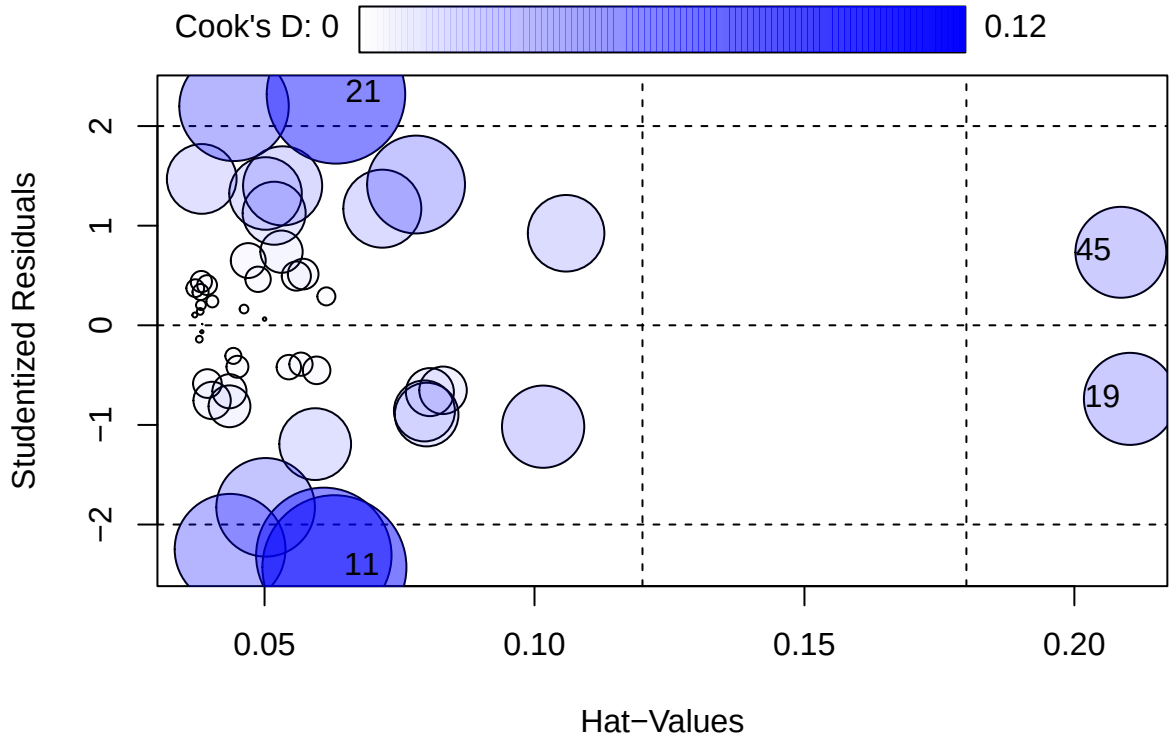
```
Call:
lm(formula = PEARNVAL ~ URB_PER + H_ED, data = subsetincomeData)

Residuals:
     Min       1Q   Median       3Q      Max
-10328.4  -2955.4    382.2   2298.4   9917.9

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 47408.53    4833.78   9.808 1.55e-12 ***
URB_PER                       191.95      58.28   3.294  0.00198 **
H_EDVocational Associate's -7017.40    1488.60  -4.714 2.57e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4433 on 43 degrees of freedom
Multiple R-squared:  0.5795,    Adjusted R-squared:   0.56
F-statistic: 29.63 on 2 and 43 DF,  p-value: 8.141e-09
```






```
      StudRes         Hat       CookD
11 -2.4286854 0.06293499 0.11958773
19 -0.7384973 0.21029459 0.04888333
21  2.3184308 0.06322337 0.11062508
45  0.7328573 0.20861622 0.04766255
```

```
     RMSE       R2      MAE
1 3861.12 0.580639 3076.044
```

Appendix D: References

Background

- Bureau, U. C. (2021, November 23). Why we conduct the decennial census of Population and Housing. Census.gov. https://tinyurl.com/5fdyh82c
- Mather, M., & Scommegna, P. (2019, March 15). Why is the U.S. Census so important?. Population Reference Bureau https://www.prb.org/resources/importance-of-u-s-census/
- Farley, R. (2020, January 31). The importance of census 2020 and the challenges of getting a complete count. Harvard Data Science Review. https://hdsr.mitpress.mit.edu/pub/rosc6trb/release/3

Data

- 2020 Unemployment Rates: U.S. Bureau of Labor Statistics. (2024). Unemployment rates for states. U.S. Bureau of Labor Statistics. https://www.bls.gov/lau/lastrk20.htm
- Urban percentage of the population for states, historical. Urban Percentage of the Population for States, Historical | Iowa Community Indicators Program. (2024.). https://www.icip.iastate.edu/tables/population/urban-pct-states
- State and local sales tax rates, 2020. Tax Foundation. (2024, February 22). https://taxfoundation.org/data/all/state/2020-sales-taxes/
- Bureau, U. C. (2022, October 27). 2020 annual social and economic supplements. Census.gov. https://www.census.gov/data/datasets/2020/demo/cps/cps-asec-2020.html
- ASEC 2020 Public Use Data Dictionary. (2020). https://tinyurl.com/3h8vexva

Supplemental Code and Analysis Help

- https://rpubs.com/muxicheng/1004550