# Predicting Median Income at the County Level using Multiple Regression

Ben Abraham

STATS 401: Applied Statistical Methods II

April 25, 2022

## Background

The data set used for this analysis contains data on a random sample of 570 counties throughout the United States. The data comes from a variety of sources including the American Community Survey, *clinicaltrials.gov*, and *cancer.gov*.

The variables included in the dataset originally pertained to predicting cancer mortality rates, but this project will seek to understand the relationship between these variables and median income. Variables of interest for this model include median income (in thousands of dollars), poverty rate, average household size (in number of people), percent of residents with a high school diploma, percent of residents with a bachelor's degree, unemployment rate, percent of residents who are married, percent of married households, the proportions of residents with private, public, or employer healthcare, population, and the region (relative to the United States).

## Analysis

### Initial Data Exploration

The initial model predicts the median income for a county given values for the percent of residents with bachelor's degrees, unemployment rate, average household size, and region. This model has a $R^2$ value of 0.6583 along with a residual standard error value of \$6740 (Figure 7). These variables are all slightly right-skewed, but they likely do not require a transformation (Figures 2 - 5). The percent of residents with a bachelor's degree is strongly correlated with the response, while the unemployment rate and average household size are moderately correlated with the response (Figure 1). The boxplot of the relationship between region and median income shows a noticeable difference for each of the different regions (Figure 6).

**Transformations**

Of the variables included in the final model, both the population variable and the poverty rate variables demonstrate a need for a transformation. As seen in Figure 8, the distribution of population is strongly skewed right. In Figure 9, a logarithmic transformation on population generates a more normal distribution. Additionally, when looking at the relationship between the poverty rate and median income, there appears to be a strong but nonlinear relationship (Figure 10). As seen in Figure 25, having a model with a quadratic term on poverty greatly improves the accuracy compared to a model without. Thus, it is best to continue with a quadratic term. However, it is important to be cautious of overfitting the model.

**Categorical Predictors**

In the initial model, no region had a p-value greater than 0.054, which suggests that there is relatively strong evidence that there is a difference between each region and the Midwest (Figure 7). Additionally, as mentioned earlier, the boxplot of the region variable demonstrates the practical significance of this relationship (Figure 6).

**Interaction Term**

Upon the investigation of an interaction between population and region, there was little evidence found to suggest that including an interaction would improve the model. There seems to be little variation among the scatterplot (Figure 11). In the model created with the interaction term, only the interaction between population and the northeast region demonstrated strong evidence in favor of an interaction (Figure 12). Therefore, no interaction was implemented in the final model.

**Model Comparison**

In addition to the variables included in the initial model, the proportion of residents with a high school diploma, the employment rate, the proportions of residents with private, employer,

and public healthcare, the proportion of residents who are married, and the proportion of married households all share relatively strong linear relationships with median income (Figures 13 and 14). Additionally, with a log transformation, a linear relationship emerges between population and the response (Figure 15). Additionally, as mentioned earlier, the relationship between poverty and median income might best be modeled with a quadratic term.

Of new models consisting of the initial model and one of the predictors listed above, those that included the proportions of residents with bachelor's degrees, public healthcare, private healthcare, employer healthcare, the proportions of residents that are married and are in poverty, the proportion of homes with married couples, the unemployment rate, and population demonstrated increases in predictive power (adjusted $R^2$ values) compared to the initial model (Figures 16 - 23). From here, a new model constructed with the variables from the initial model as well as those above had an adjusted R-squared value of 0.8913 and a residual standard error of 3.779 thousand dollars which is much improved from the original model (Figure 24).

**Diagnostic Plots and Assumptions**

To begin, as the fitted values increase, there seems to be a pattern in which the residuals become more positive (Figure 26). A model with an exponent to the third power of poverty, generates a pattern of residuals that is more linear and has a greater p-value (Figure 28). This model also has a higher adjusted $R^2$ value than the model with a second power exponent (Figure 27) Therefore, while not perfect, the assumption of linearity is reasonably met, although it is now important to be cautious of overfitting. Moreover, this same graph gives relatively sufficient evidence to suggest that the constant variance assumption is reasonably met. Additionally, in the QQ plot (Figure 29), the observations at the upper tail pull away from the line. The normality condition is reasonably met, but it is important to proceed with caution. The random assumption

is true since the data comes from a random sample. It is also reasonable to assume that the independence condition is satisfied, but it should be noted that it is possible that surrounding counties may have impacts on one another. Lastly, the zero mean assumption is satisfied because how Ordinary Least Squares Regression works.

**Multicollinearity and Overfitting**

While the model demonstrates high predictive power, it is possible that this model contains evidence of multicollinearity since several predictors may have strong correlations with each other such as the proportion of residents that are married and the proportion of homes which house a married couple. The VIF output for the current model demonstrates little evidence of multicollinearity (Figure 30). There are 16 predictor variables in the model, which for a sample of 570 observations is a reasonable amount. There is also a quadratic term, and it is to the third degree which could potentially lead to overfitting.

**Final Model**

The final model had a $R^2$ value of 0.9006 which implies that about 90.06% of the variation in median income is explained by this linear relationship and a residual standard error of 3.664 thousand dollars meaning that the average error of the model is approximately $3664 (Figure 24). Moreover, the final model created had the following conditional mean function:

$$E(Y_{medianIncome}|X) = \beta_0 + \beta_1 X_{pctBach} + \beta_2 X_{pctUnemployed} + \beta_3 X_{pctPublicHC} +$$
$$\beta_4 X_{pctPrivateHC} + \beta_5 X_{pctEmployerHC} + \beta_6 X_{pctMarried} + \beta_7 X_{pctMarriedHouse} + \beta_8 X_{avgHouse}$$
$$+ \beta_9 X_{pctPoverty} + \beta_{10} X^2_{pctPoverty} + \beta_{11} X^3_{pctPoverty} + \beta_{12} \log X_{population}$$
$$+ \beta_{13} Z_{northeast} + \beta_{14} Z_{southeast} + \beta_{15} Z_{southwest} + \beta_{16} Z_{west}$$

Looking at the employment rate variable, one will notice that in the final model, if the unemployment rate were to grow by 1%, then the model expects the median income for the county to increase by $197.05 on average (while holding all other variables constant). Additionally, the coefficient on the population variable (which has a logarithmic transformation) is 0.7619289 which suggests that on average if the population of a given county were to increase by 5%, the model estimates an average increase in median income of about $37.17 (while holding all other variables constant). Moreover, the model calculates that compared to counties in the Midwest, counties in the Northeast United States have a greater predicted intercept by 3.4398760 thousand dollars. This suggests that counties in Northeast United States have a total predicted intercept of 78.9334434 thousand dollars.

The proportion of residents that are married has a p-value of 0.15515, but a model without would experience a decrease in adjusted $R^2$ (Figure 32). Additionally, all the other predictors in the model have low p-values which suggest that they are all statistically significant. Figure 31 depicts the expected change (while holding all other predictors constant) in median income for an increase of the standard deviation for each quantitative variable in the model (aside from poverty rate). There is a noticeable change in the response for all variables, so these predictors are practically significant.

## Conclusion

Ultimately, with a $R^2$ value of 0.9006 and an average error of $3664, the model constructed is quite accurate at predicting median income (Figure 27). Its largest weakness is that as fitted values grow larger, it becomes less accurate. In future iterations of this model, a logarithmic transformation on median income may perhaps lead to a better relationship. Another issue this model is facing is overfitting, so testing on a larger sample size in the future may be beneficial.