Predicting the Proportion of Residents of Toronto Neighbourhoods with High Income

Gavin Pu

2022-12-20

Introduction

A common metric of individual success is employment income. As earning a high income is a common goal for many, studies have been conducted to find what high income is most correlated with. Commonly cited factors are higher education and pursuing careers in high-paying fields. A simulation conducted in 2020 using data from the U.S. Census Bureau showed that university education has been found to correlate with higher individual earnings.[1] The results from an analysis conducted in 2021 suggest that upper social classes invest in higher education for their children to preserve their wealth, suggesting that education is positively associated with higher income.[2] The Occupational Outlook Handbook published by the U.S. Bureau of Labor Statistics reports the median annual wage in the U.S. for science occupations as $72,740 and for management occupations as $102,450 in May 2021.[3,4] The medians of both occupation groups are higher than the average median wage of $45,760 across all fields.[3,4] Migration after graduation from university is a less commonly studied factor showing correlation with higher income.[5]

Whether a combination of multiple of the aforementioned characteristics is correlated with higher income is not often explored within existing literature. Using data on the neighbourhoods of Toronto, this analysis aims to find a multiple linear regression (MLR) model that can most accurately predict the proportion of individuals in an area who have a high employment income.

Methods

Data Collection and Cleaning

2016 census data on Toronto's neighbourhoods were obtained from the City of Toronto's Open Data Portal. The dataset was cleaned by extracting the six characteristics of interest for each of the 140 neighbourhoods in the census. To calculate the proportion of residents of a neighbourhood who fit each characteristic, observations for each neighbourhood were divided by the total number of residents within the neighbourhood where the observation was taken. This step was performed because the neighbourhoods did not have the same number of residents.

Variables

Table 1 lists the six variables considered in this analysis. Each variable was measured as a proportion of residents within a neighbourhood. Excluding knowledge of English and French, proportions only included individuals aged 15 years and older.

Exploratory Data Analysis

An exploratory data analysis was conducted by performing a simple linear regression (SLR) between each predictor and the response. Residual plots and Normal Q-Q plots were created to visually diagnose the assumptions of linearity, homoscedasticity, independence of errors, and Normality of errors. Box-Cox transformations were performed on each model where assumptions were not satisfied. Influential points for

Table 1: Box-Cox transformations performed on variables and summary statistics after transformation ($n = 140$)

Variable	Transformation	Mean	SD
Predictors			
University Certificate, Diploma, or Degree at Bachelor Level or Above	Square Root	0.537	0.123
Work in Management Occupations	Logarithm	-2.910	0.491
Work in Professional, Scientific, or Technical Services	Logarithm	-2.908	0.564
Knowledge of Both English and French	Logarithm	-2.492	0.520
Location of Study Different Than Province or Territory of Residence	Logarithm	-3.761	0.858
Response			
Employment Income of $100,000 or Greater	Logarithm	-3.105	0.914

each model were found using Cook's distance. Models were refit after removing all influential points to measure their effects on the least squares estimates. ANOVA tests were then ran on the original models.

Variable Selection

An MLR model was fit using all five predictors. Residual plots and Normal Q-Q plots were used to visually diagnose the assumptions of linearity, homoscedasticity, independence of errors, and Normality of errors. A Box-Cox transformation was performed on all variables to correct the model for these assumptions. Using stepwise selection, two models were found: one with the lowest Akaike information criterion (AIC) and one with the lowest Bayesian information criterion (BIC). A third model was derived from the original MLR model with all predictors using the least absolute shrinkage and selection operator (LASSO). To handle the presence of multicollinearity, models were rejected if the variance inflation factor (VIF) of any predictor was greater than 5.

Model Diagnostics and Validation

Residual plots and Normal Q-Q plots were constructed for the remaining models to diagnose the assumptions of MLR. Box-Cox transformations were performed on models that violated one or more MLR assumptions. Using Cook's distance, influential points were found for the remaining models. Calibration plots were created to visually assess a model's prediction accuracy. On each model, an ANOVA test was conducted and the adjusted R-squared was calculated.

Results

Data Description

Box-Cox transformations were necessary for all variables to form a linear relationship between the response and each predictor. The transformation of each variable is shown in Table 1. Since no untransformed variables were used in any valid model in the analysis, the means and standard deviations of variables were calculated after their respective Box-Cox transformations were applied (Table 1).

Exploratory Data Analysis

Figure 1 displays the SLR models found. The Box-Cox transformation was used to fit each model because SLR could not be performed using the untransformed variables due to violations of linearity and homoscedasticity. Points shown in red are influential points determined by whether Cook's distance was greater than a threshold of $4/(n-2)$. The least squares estimate that exhibited the greatest change after removing all influential points from a model was the intercept parameter of the model show in Figure 1C, with an increase of 0.308.

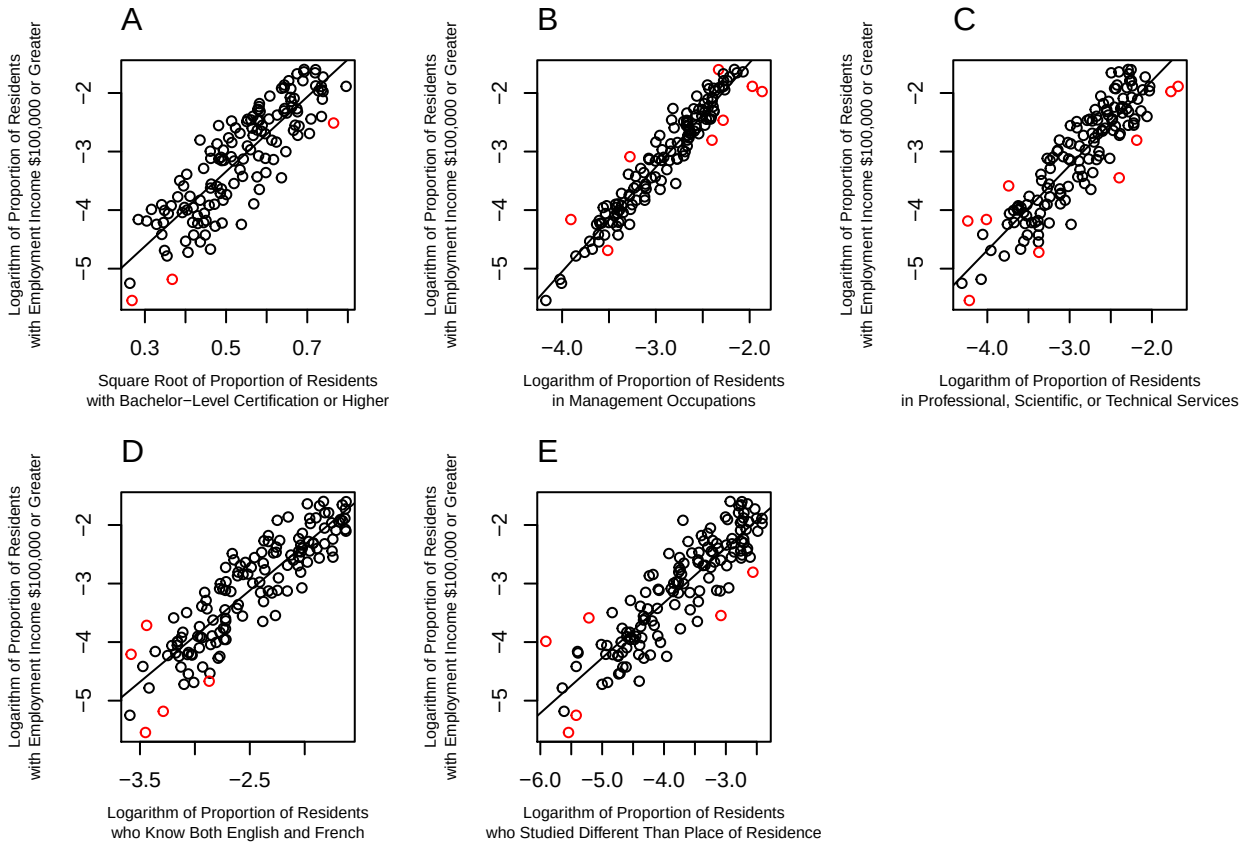

Figure 1: SLR models with Box-Cox transformations applied and influential points indicated in red ($n = 140$)

Model Selection

A Box-Cox transformation was performed to correct the initial MLR model with all 5 predictors so that it would satisfy the assumptions of linearity, homoscedasticity, independence of errors, and Normality of errors. The transformations applied to each variable are shown in Table 1. Table 2 displays the variables selected using each of the 3 methods. Stepwise selection of variables showed that the model with the lowest AIC had 3 predictors while the model with the lowest BIC had 2 predictors. The model selected using LASSO was the same as the model with the lowest AIC. The model with the lowest BIC was chosen over the model with the lowest AIC and found using LASSO because all VIFs of the model with the lowest BIC were less than 5 (Table 2), indicating that the model with the lowest BIC had less multicollinearity. The equation for this model is

$$\log \hat{y} = 0.953 + 0.979\sqrt{x_1} + 1.575 \log x_2 \tag{1}$$

where $\hat{y}$ is the predicted proportion of residents in a neighbourhood with an employment income of \$100,000 or greater, x_1 is the proportion of residents in a neighbourhood with a univerisity certificate, diploma, or degree at the bachelor level or higher, and x_2 is the proportion of residents in a neighbourhood working in management occupations.

Model Diagnostics and Validation

The standardized residuals were used to create a a standardized residuals versus fitted values plot (Figure 2A) and a Normal Q-Q plot (Figure 2B). Due to the lack of any systematic pattern or clustering in the residual plot, the assumptions of linearity, homoscedasticity, and independence of errors were fulfilled. Since there is a roughly one-to-one relationship between the theoretical quantiles of the standard Normal distribution and the sample quantiles of the standardized residuals, the assumption of Normality of errors was also met. The

Table 2: VIFs for each predictor of MLR models found using AIC, BIC, and LASSO ($n = 140$)

Variable Selection Method	Predictor	VIF
AIC/LASSO		
	Bachelor Level or Above	4.181
	Management Occupations	6.410
	English and French	5.922
BIC		
	Bachelor Level or Above	3.750
	Management Occupations	3.750

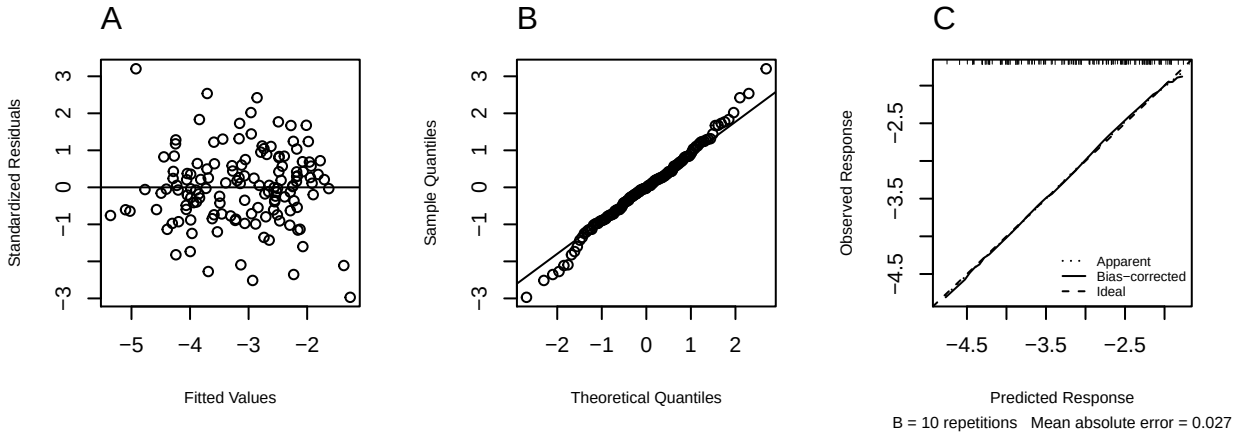

Figure 2: Standardized residuals versus fitted values plot (A), Normal Q-Q plot (B), and calibration plot (C) for model with lowest BIC ($n = 140$)

calibration plot obtained from performing cross validation demonstrates that the model has strong prediction accuracy (Figure 2C). No influential points were found in the model when comparing Cook's distance to the 0.5 quantile of the F distribution with $p + 1$ and $n - p - 1$ degrees of freedom ($p = 2$, the number of predictors).

Model Significance

The p-values from the ANOVA test for the model are shown in Table 3. Since both p-values are extremely small, there is strong evidence against the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ for an ANOVA test with 2 predictors. From the ANOVA results in Table 3, the adjusted R-squared was found to be $R^2_{adj} = 0.999$.

Discussion

Equation (1) can be expressed as

$$\hat{y} = k_1 k_2^{\sqrt{x_1}} x_2^{1.575} \tag{2}$$

where $k_1 = e^{0.953}$ and $k_2 = e^{0.979}$. Equation (2) indicates that both a greater proportion of residents having at least a bachelor-level education and a greater proportion of residents working in management occupations are correlated with a greater proportion of residents earning a high income within a Toronto neighbourhood. More broadly, the model suggests that higher education and careers in management are positively correlated with higher incomes.

The main shortcoming of this model is its limited generalizability. Since the data came exclusively from Toronto's 2016 census, whether the model is still reasonably applicable to Toronto in 2022 is questionable.

Table 3: ANOVA results for model with lowest BIC ($n = 140$)

Source	DF	Sum Squares	Mean Squares	F value	p-value
Bachelor Level or Above	1	85.194	85.194	1353.17	$< 2.2 \times 10^{-16}$
Management Occupations	1	22.203	22.203	352.66	$< 2.2 \times 10^{-16}$
Residuals	137	8.625	0.063		

It also cannot be assumed that the model accurate describes other locations, although it may generalize somewhat to places like Toronto containing a mix of urban and suburban areas. The potentially narrow applicability of this model shows that the findings from this analysis are quite limited in scope. Hence, the model could be improved by expanding it to include data from other areas.

Another weakness is that the transformations applied to each variable make the model difficult to interpret intuitively. For example, the least squares estimates in Equation (1) indicate the the estimated intercept is 0.953. A classical interpretation of this coefficient would be that the expected logarithm of the proportion of residents with an employment income of $100,000 or greater is 0.953 when both the square root of the proportion of residents with a university certificate, diploma, or degree at the bachelor level or above and the logarithm of the proportion of residents working in management occupations are 0. This interpretation, however, is awkward and does not convey an easily understandable meaning of the intercept in context.

References

1. Hershbein, B. (2020). College Attainment, Income Inequality, and Economic Security: A Simulation Exercise. *AEA Papers and Proceedings*, *110*, 352–355.
2. Prettner, K. (2021). The U-Shape of Income Inequality over the 20th Century: The Role of Education. *The Scandinavian Journal of Economics*, *123*(2), 645–675.
3. U.S. Bureau of Labor Statistics. (2022, September 14). *Life, Physical, and Social Science Occupations*. Occupational Outlook Handbook. Retrieved December 20, 2022, from https://www.bls.gov/ooh/life-physical-and-social-science/home.htm
4. U.S. Bureau of Labor Statistics. (2022, September 8). *Management Occupations*. Occupational Outlook Handbook. Retrieved December 20, 2022, from https://www.bls.gov/ooh/management/home.htm
5. Mitze, T. (2020). Graduate Migration and Early-career Labor Market Outcomes: Do Education Programs and Qualification Levels Matter? *Labour*, *34*(4), 477–503.