**FINA375**

**Business Analytics**

**Name: Romario Medel-Estrada**

**Linear Probability Model (LPM) Mini Project**

**State: New York**

**Business Analytics Modeling on the Probability of Accepting Wage Income Less Than**

**$35,000**

We model a simplified linear probability model of earning low income by respondents in the

ACS 2022 data. The state analyzed is New York. The target or response variable is a binary

(zero-one) or dummy variable that equals to 1 if the respondent earns a wage income less than

$35,000 per year. The LPM estimated coefficients are interpreted as the partial effects of an

independent variable (calculated at the mean) on the probability of accepting a wage income of

$35,000 or less.

**The Business Questions and Findings**

This project seeks to answer two related business questions from the perspective of an analyst

working for CVS. The project seeks to assess female applicants with specific disabilities and

how they might be good fit in entry level jobs in the personal care retail industry.

This project finds that all disabilities contribute to higher probability of earning low income

(compared to no disability), but different disabilities have varying different impact on this

probability.

The LPM also predicts that female applicants would have a greater probability of earning low

income, and this effect would double as the number of people in her household goes up by one

above the average of 3 people.

The LPM predicts that disabilities involving physical, hearing, and memory abilities positively

contributes to higher probability of earning low income. Female candidates in New York are

about 3.8% more likely to accept low wage income, and this effect doubles as the number of

people in the household increases above 3. The findings can assist CVS in recruiting female

workers among large households, with outreach to those with some hearing and physical

disability for which accommodations are not too costly.

**Table 1: LPM Results (New York)**

**The Dependent (Target Variable):** a binary (dummy) variable coded 1 (one) if the respondent

earns less than $35,000 per year, or 0 (zero) otherwise.



**Model Assessment and Confusion Matrix**

The goodness of fit of binary models (where the target is a dummy variable) is assessed by their ability to correctly classify the target. In this modeling exercise, we cross-validate the model by using the hold-out method (and one fold) by splitting the data into a training set (60%) and a validation set (40%) so that the prediction model is less likely to overfit the data. We also round the predicted probability to 0s and 1s at the 0.5 cut-off.

A screenshot of a computer

Description automatically generated

As shown in the Confusion Matrix, the accuracy of our LPM is 0.73 (close to 1): the overall proportion of observations classified correctly (true positive and true negative) is 73%. Therefore, the misclassification rate is 27%. The specificity of the model is 0.7613. The precision is 0.7050. Sensitivity of the model is 0.6989. Overall, the model does a decent job of classifying the wage income of individuals and makes acceptable false positives and false negatives predictions.

**Interpretation of Coefficients**

The interpretation of the estimated coefficients (that are significant at the 5% level) is presented below. The omitted variables (dummies) that serve as benchmarks are Age 35 or younger, non married, non-citizen, male, no disabilities, high school or less, and working for self or family. Two variables are not significant at the 5% level (ambulatory and visual disabilities), so their interpretations are omitted.

**Other Income:** holding all other variables constant, as other income (non-wage) goes up by $1,000 above its average level, the probability of earning less than $35,000 goes up by 7.8%. This is very similar to the Minnesota coefficient of 7%.

**Private Employment:** holding all other variables constant, if the respondent works in the private sector, the probability of earning less than $35,000 goes down by 25.6% compared to those who work for family or self. New York coefficient is smaller than Minnesota coefficient of 28.8%, indicating a greater than 3% difference.

**Public Employment:** holding all other variables constant, if the respondent works in the private sector, the probability of earning less than $35,000 goes down by 38.3% compared to work for family or self**.**

**Age 26 to 35:** holding all other variables constant, if the respondent’s age is 25 to 35, the probability of earning less than $35,000 goes down by 34.5% compared to those who are younger than 25. The Minnesota coefficient is 32.2% so there is only a 2.3% difference compared to New York.

**Age 36 to 45:** holding all other variables constant, if the respondent’s age is 36 to 45, the probability of earnings less than $35,000 goes down by 36.4% compared to those who are younger than 25.

**Age 46 to 55:** holding all other variables constant, if the respondent’s age is 46 to 55, the probability of earnings less than $35,000 goes down by 36.2 compared to those who are younger than 25.

**Age 56 and older:** holding all other variables constant, if the respondent’s age is 56 and older, the probability of earning less than $35,000 goes down by 19.5% compared to those who are younger than 25.

**Number of people in a household (NP**): holding all other variables constant, if the number of people in a household goes up by one (above the average NP), the probability of earning less than $35,000 goes down by 1.3%. The Minnesota coefficient was 4.2%, this indicates an almost 3% difference compared to New York.

**Married:** holding all other variables constant, if the respondent is married the probability of earning less than $35,000 goes down by 6.4% compared to those non-married.

**Citizen**: holding all other variables constant, if the respondent is an American citizen, the probability of earning less than $35,000 goes down by 10.5% compared to non-citizens.

**Female**: holding all other variables constant, if the respondent is female, the probability of earning less than $35,000 goes up by3.8% compared to male.

**FEMNP (Female\*NP):** this interaction term indicates that among females, while holding other relevant variables constant, if the number of people in a household goes up by one (above the average NP), the probability of earning less than $35,000 goes up by 2.9%. This coefficient has a higher impact in Minnesota with a 4.2% probability increase compared to New York probability increase of 2.9%.

**Physical Disability:** holding all other variables constant if the respondent has a physical disability, the probability of earning less than $35,000 goes up by 16.7% compared to those without a disability.

**Hearing Disability:** holding all other variables constant if the respondent has a physical disability, the probability of earning less than $35,000 goes up by 3% compared to those without a disability.

**Memory Disability:** holding all other variables constant if the respondent has a physical disability, the probability of earning less than $35,000 goes up by 18.8% compared to those without a disability.

**Bachelor or less:** holding all other variables constant, if the respondent highest level of education completed is bachelor (or less but more than high school), the probability of earning less than $35,000 goes down by 19.9% compared to those who have completed high school or less. The Minnesota coefficient is 16.1%, which indicates that bachelor or less slightly more likely to accept $35,000 than New York.

**Master or higher:** holding all other variables constant, if the respondent highest level of education completed is master or higher, the probability of earning less than $35,000 goes down by 35% compared to those who have completed high school or less.

**Economic Significance**

The three variables with the largest (absolute) magnitude or economic significance (with p-value <5% are: Public employment (0.383), Age 46 to 55 (0.364), and Master Plus (0.35). The 3 economic significance in Minnesota were: Public employment (0.351), Age 46 to 55 (0.354), and Master Plus (0.304). They were the same variables with slightly different coefficient values. If these variables were similar in more states, this could indicate the largest factors to wage earnings.

**Predictions**

The model predicts a probability of 0.424 of earning low income (less than $35,000) for an applicant who is between 36 to 45 years old, earns $0 in other income, works in the private sector, male, American, married, four people in the household, and has a bachelor (more than high school) degree, and has some physical disability. We therefore predict this candidate **would not** accept a low wage offer of $35,000 for an entry level job at CVS. The model predicts a probability of 0.4998 of earning low income (less than $35,000) for an applicant who is younger than 26, earns $0 in other income, works in the public sector, female, American, not married, one person in her household, and has a master or higher degree, and no disability. We therefore predict this candidate **would not** accept a low wage offer of $35,000 for an entry level job at CVS.

**Predictions New York vs Minnesota**

The model predictions compared to Minnesota are that the applicant who is younger than 26 with a bachelor’s or less (but more than high school) would accept the low wage offer of $35,000 because the model predicted a probability of 0.5388 compared to the probability in New York of .4998 telling us they would not accept the job in New York. On the other hand, both applicants who are between 36 and 45 years old would not accept the low wage offer neither in Minnesota or in New York. The predicted probabilities were 0.3355 and 0.424 respectively.

**Unbiasedness and Robustness**

The OLS regression models have not been tested for violations in model assumptions (multicollinearity, non-constant variance, omitted variable bias) or for model misspecifications.

**Appendix**

Screenshots of the scripts and output are shown below.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated