How to make over 50k (in 1994):
Andrew Lee

I intend to use the 1994 Census bureau database to model the probability of earning over $50,000. This analysis should determine the important features which significantly impact the probability of income.

The data used for this analysis comes from the 1994 Census bureau database, so is slightly out of date. However, it contains data on 15 features of over 3000 individuals. The dependent feature this model will be exploring is the income variable. This is a binary classification of an income over $50,000 a year or less than or equal to $50,000 a year. This variable has been converted to a dummy variable for this analysis. It's important to note that these two classes are imbalanced in the dataset, with 75.9% of the individuals making less than or equal to $50,000. To help gain better intuition behind this data, the average annual inflation rate of 2.19% per year since 1994, this $50,000 would be equivalent to about $87,800 in 2020.

After some basic data exploration, it's clear some of these features do not have good data. For example, capital.gain and capital.loss both have medians of 0 but high means due to few outlier data points. These will be excluded from my analysis along with fnlwgt (an outdated variable attempting to control for demographic statistics). The relationship feature can be represented through sex and marital status, and education_num is better captured by the discrete variable education. These features will not be considered, but no individuals are removed from the dataset.

This leaves nine features that can be considered for the analysis: age, workclass, education, sex, occupation, marital status, native country, hours per week, and race. Each of these features were graphed to gain some intuition behind the breakdown of the various features. Then a linear regression of income and the feature was conducted to find the statistical significance of a simple linear relationship. This revealed more features which have insignificant data and may not be relevant enough for this analysis. For example, even though race has a highly significant linear relationship, it contains 27816 data points for white and only 4745 for all other races. Although race is an important feature, the disproportionately large sample size of White will lead to an inaccurate analysis. The native country feature suffers from this same problem. The data set contains 29170 individuals native to the United States and only 3391 for all other countries. Removing these two features will also remove any relationship between the

feature and income. However, the insignificant data would likely have led to an inaccurate model. This generalizes the model to all races and native countries, however, will be significantly more accurate for white people from the United States as that's where most of the data lies.

Categorical features with few data points in some categories give insignificant coefficients. For example, there are only 9 individuals with occupation Armed-Forces, so this model will not be accurate for this occupation. Priv-house-sev is another occupation which will give a high standard error with only 149 data points. Education has a similar problem, with only 51 individuals with education Preschool and 168 with education 1st - 4th. Married-AF-spouse is another category not represented well enough to create an accurate prediction with only 23 data points. These features also had a large impact on the marginal effect's graphs, so all 400 individuals were removed. This reduces the dataset from 32,561 individuals to 32,161.

Using the glm function in R, a binomial logit model was created. This model used the dummy variable income_Y as the independent variable. This dummy was set to 1 for income greater than $50,000, so a high probability predicts a higher income. The dependent variables used are education, sex, age, hours per week, marital status, and occupation. The model summary shows the coefficient estimate, standard error, z-score, and p-value for each feature. The summary shows the model is fairly accurate for a lot of features. Of the 34 coefficients estimated 27 have a p value of <0.5.

To gain a better understand of what these coefficients mean, predictions were made for various made-up individuals. First, a control individual was predicted upon. This individual was made using the median of numeric features like hours and age, and the mode of categorical date like occupation. This individual is a male high school graduate, age 37, married to a civilian, works 40 hours per week, and occupation is in a professional specialty. This control gave a probability of earning greater than $50,000 of 0.366. For each feature, a new individual was created changing only that feature from the control. The numeric features age and hours worked were changed from the median to the 3rd quartile value. For the categorical features occupation and education, the category with the lowest coefficient was chosen (lower the probability of earning less). For the binary classification of gender, the feature was simply switched to female. This led to a slight increase in probability of earning less. For marital status, the mode Married-civ-spouse is the highest coefficient, so the category with the lowest coefficient was chosen.

Figure 1 below is the graph of these predications and a summary of these individuals and their distinct features.

With so many categorical features it's difficult to capture the marginal effect of each feature. The sjPlot package was used to graph the marginal effect of each feature. The marginal effects of these numeric features were much easier to visualize. These graphs also show a 95% confidence interval of the marginal effect. The other plots also show a 95% confidence interval of marginal effect of each categorical feature. However, with so many categories they are hard to read. They can still help show the intuition behind which features are important. These are Figure 2 – 7 below.

Various functions were calculated to determine the accuracy of this model. First, a simple count R2 was calculated. This uses the model to create a prediction of income_Y for each individual in the dataset and compares the model predictions to the true income_Y value. A count R2 of 0.8193 was found, this model correctly classified 81.93% of individuals.

Next, a McFadden R2 was calculated. This creates a "dummy" model for which all the feature coefficients are zero. This model will always return a probability of 0.2431, or the percentage of the dataset earning greater than $50,000. This gives all individuals an income_Y of 0. The log likelihood of the multinomial logistic regression model and this dummy model are both calculated. The McFadden R2 is defined using the ratio between these two log likelihoods. A McFadden R2 of 0.3509 was found. This pseudo R2 calculation can be difficult to interpret, but it shows a significant difference between the calculated model and the dummy model.

Finally, the likelihood-ratio test was used to determine the significance between the two models. The chi squared test was used between the log likelihood of these two models which a null hypothesis of the models being equivalent. A p-value equivalent to zero was found, the null hypothesis can be rejected. There is significance between these two tests.

This binomial logit model using education, sex, age, hours per week, marital status, and occupation as independent variables and income of greater than $50,000 as dependent variable gives a training set accuracy greater than the dependent variable ratio. This means the model could lead to significant estimations more accurate than guessing. This analysis also showed surprising results from the marginal effect of different feature categorizations. I was particularly surprised by the significant difference being married made. I was also surprised that a bachelor's degree did not make a larger difference on the probability compared to other educational

achievements. I would be very interested to look at data from 2020 using $90,000 to see how this has changed. When doing this analysis again, I would split up my data into test and train sets. This would give a better estimate of the model's accuracy and reduce the chance of overfitting. Then a confusion matrix could be used to display the model accuracy.

Graphs:

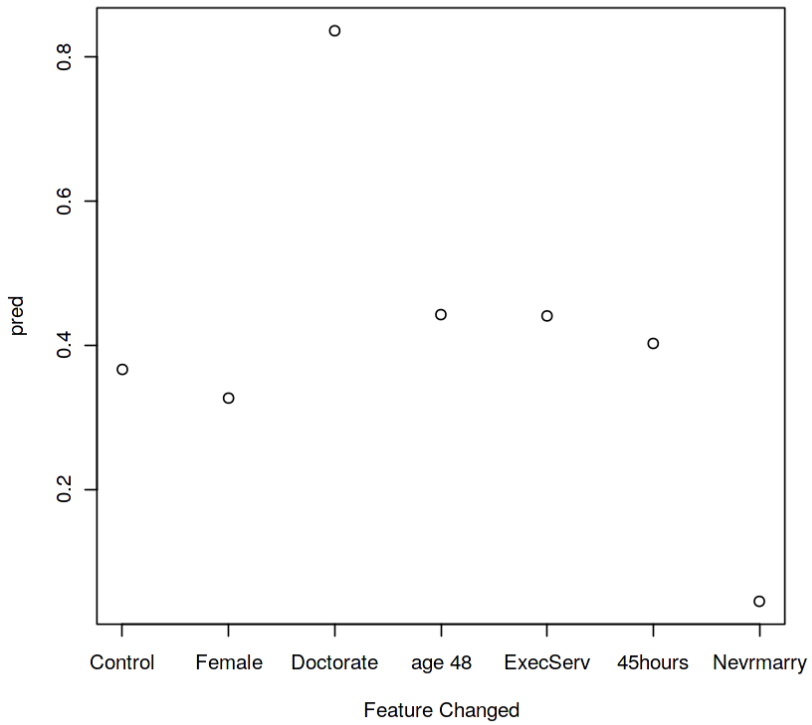

Figure 1 Predictions summary

- Control: education = HS-grad, sex = Male, age = 37, marital status = Married-civ-spouse, hours per week = 40, occupation = Prof-specialty
- Female: sex = Female
- Doctorate: education = Doctorate
- age 48: age = 48
- ExecServ: occupation = Exec-managerial
- 45hours: hours per week = 45
- Nevrmarry: marital status = Never-married

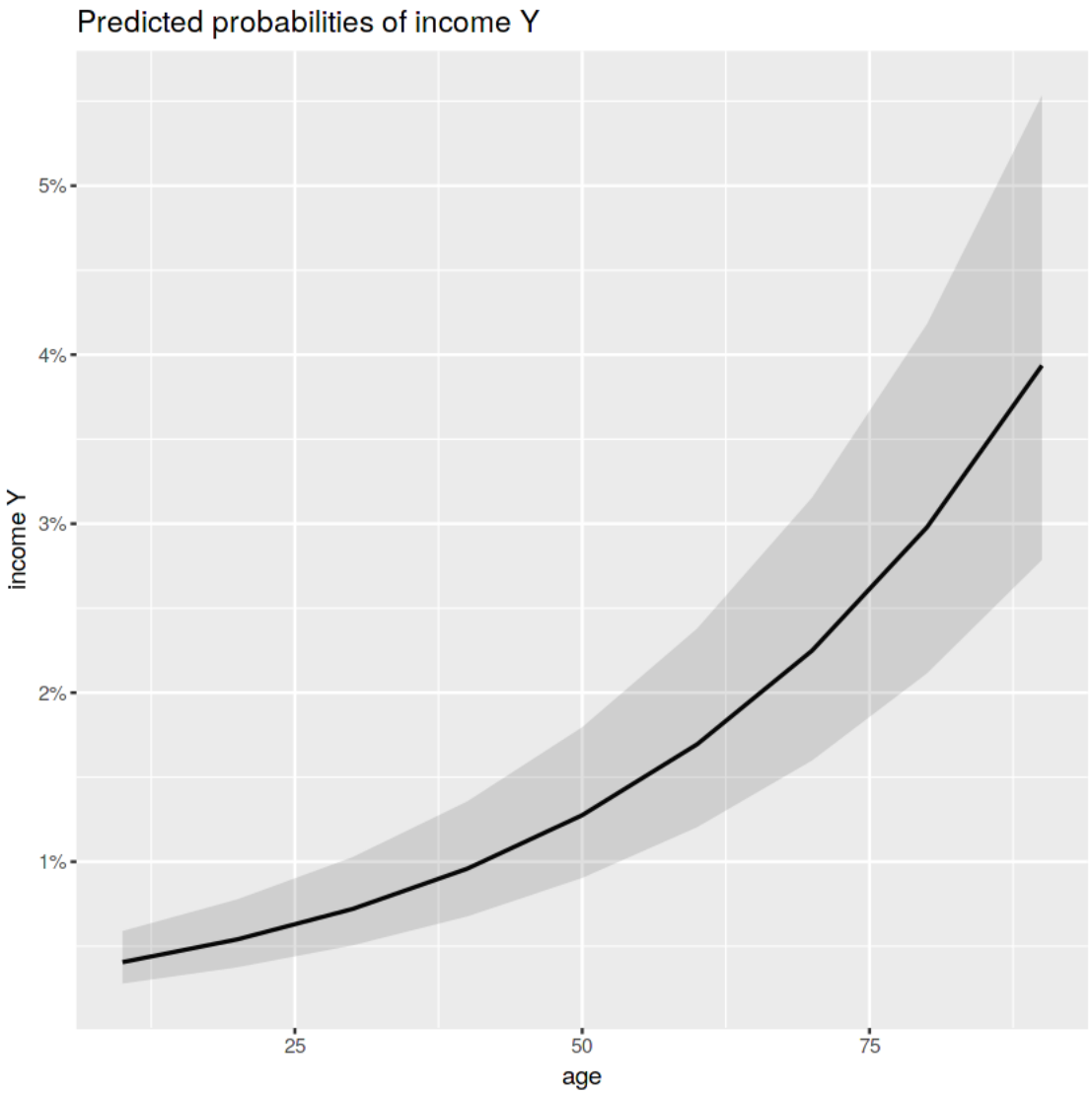

Figure 2 CI of marginal effect of age

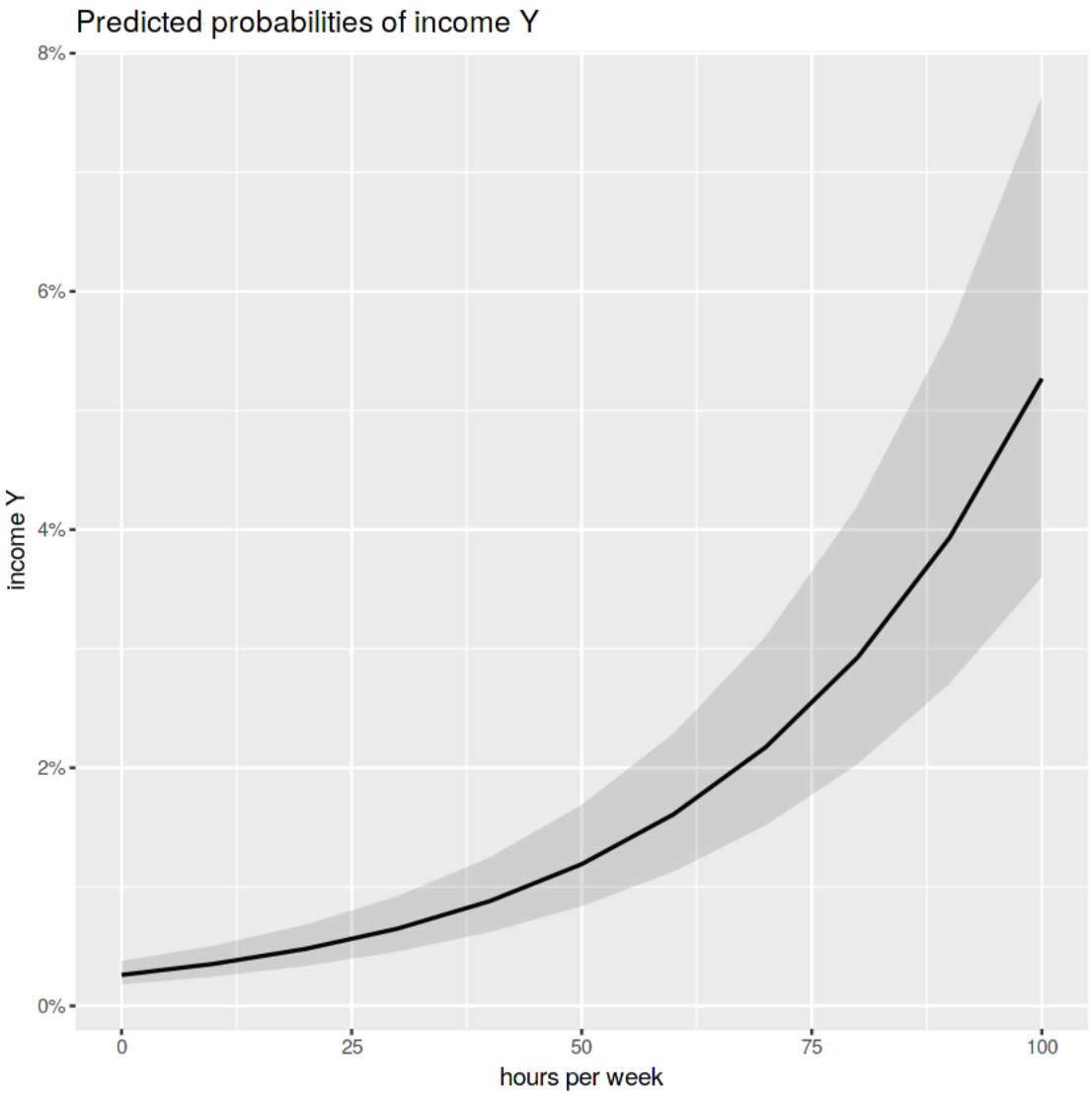

Figure 3 CI of marginal effect of hours per week

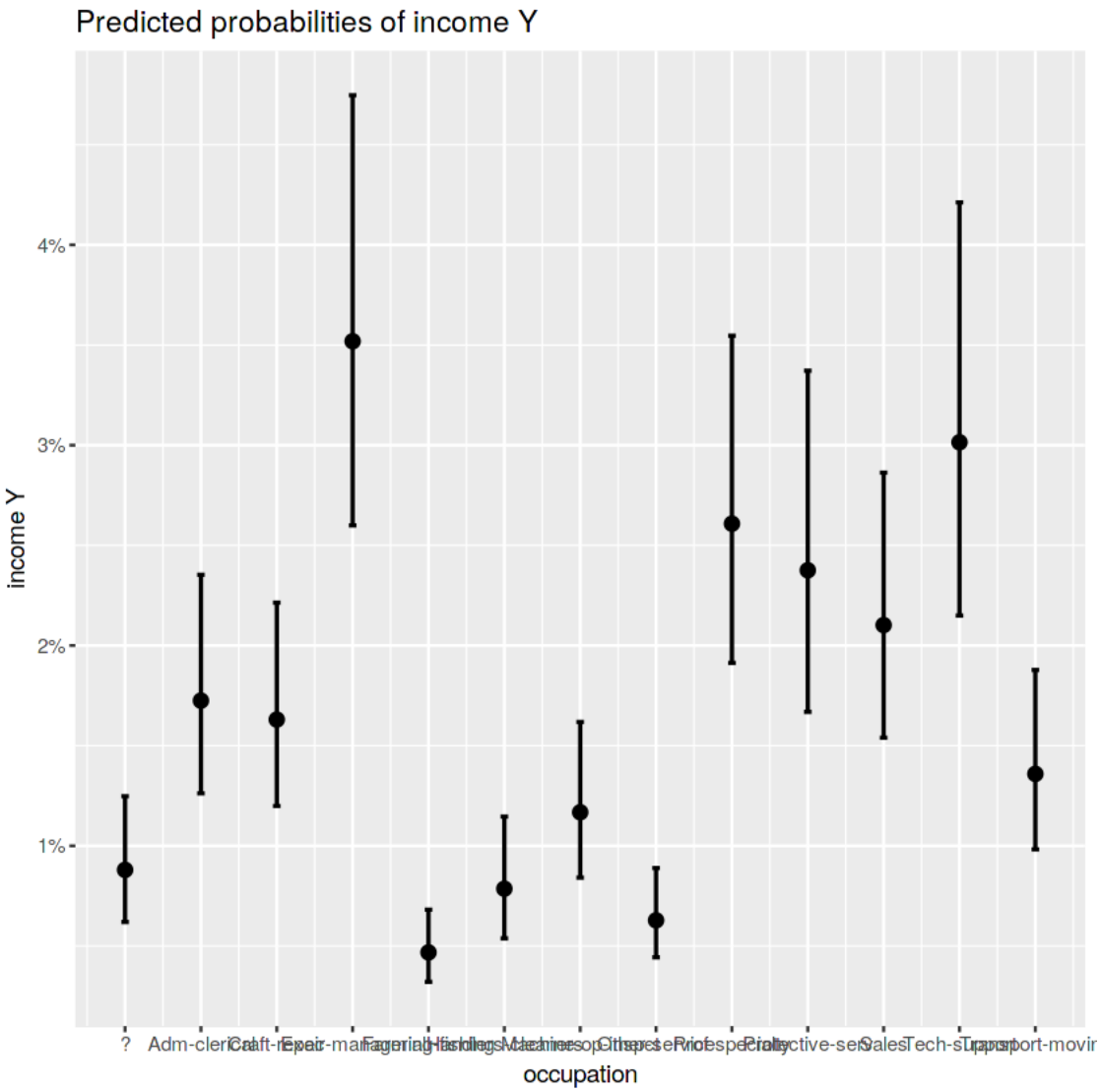

Figure 4 CI of marginal effect of occupations

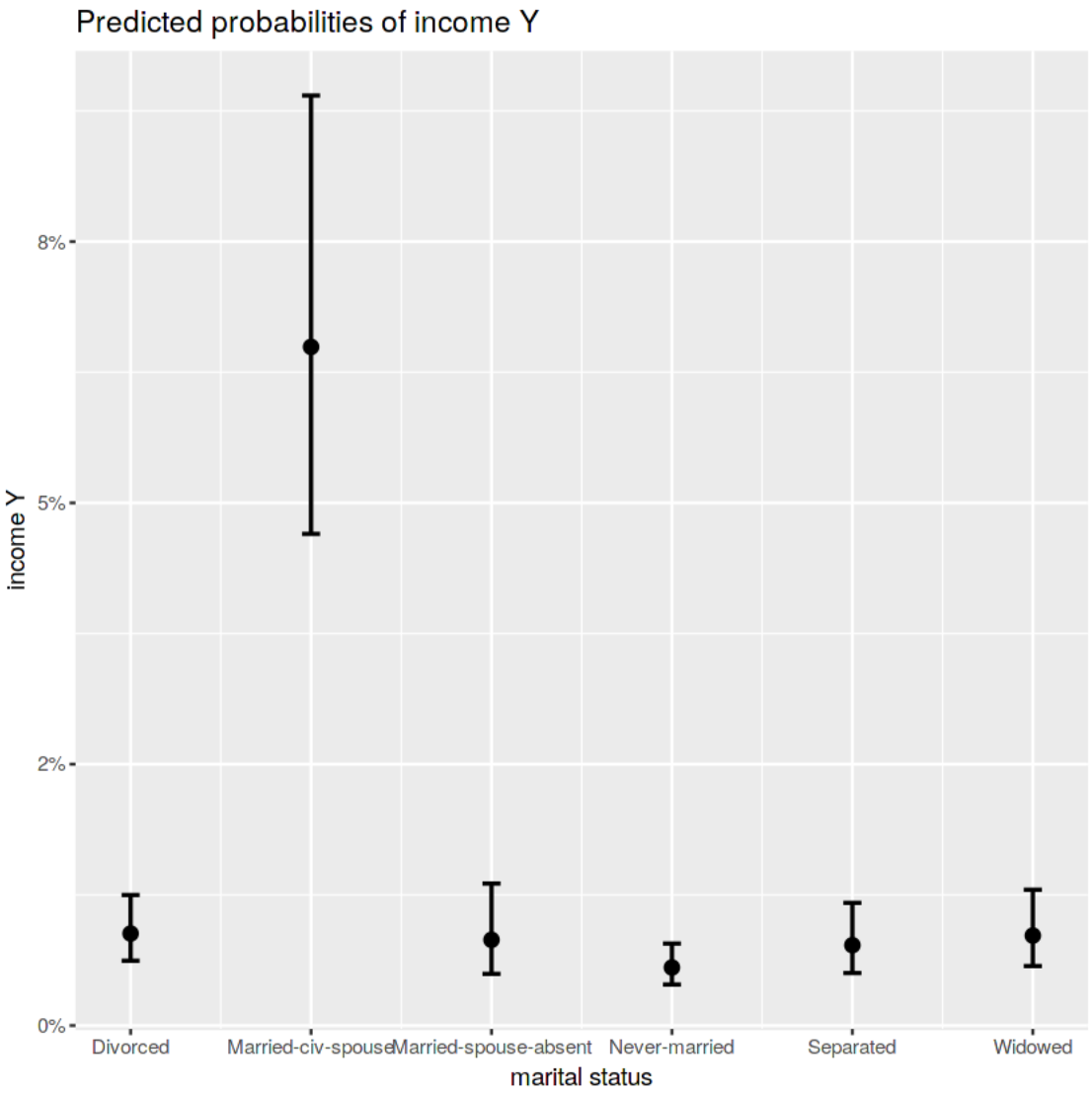

Figure 5 CI of marginal effect of marital status

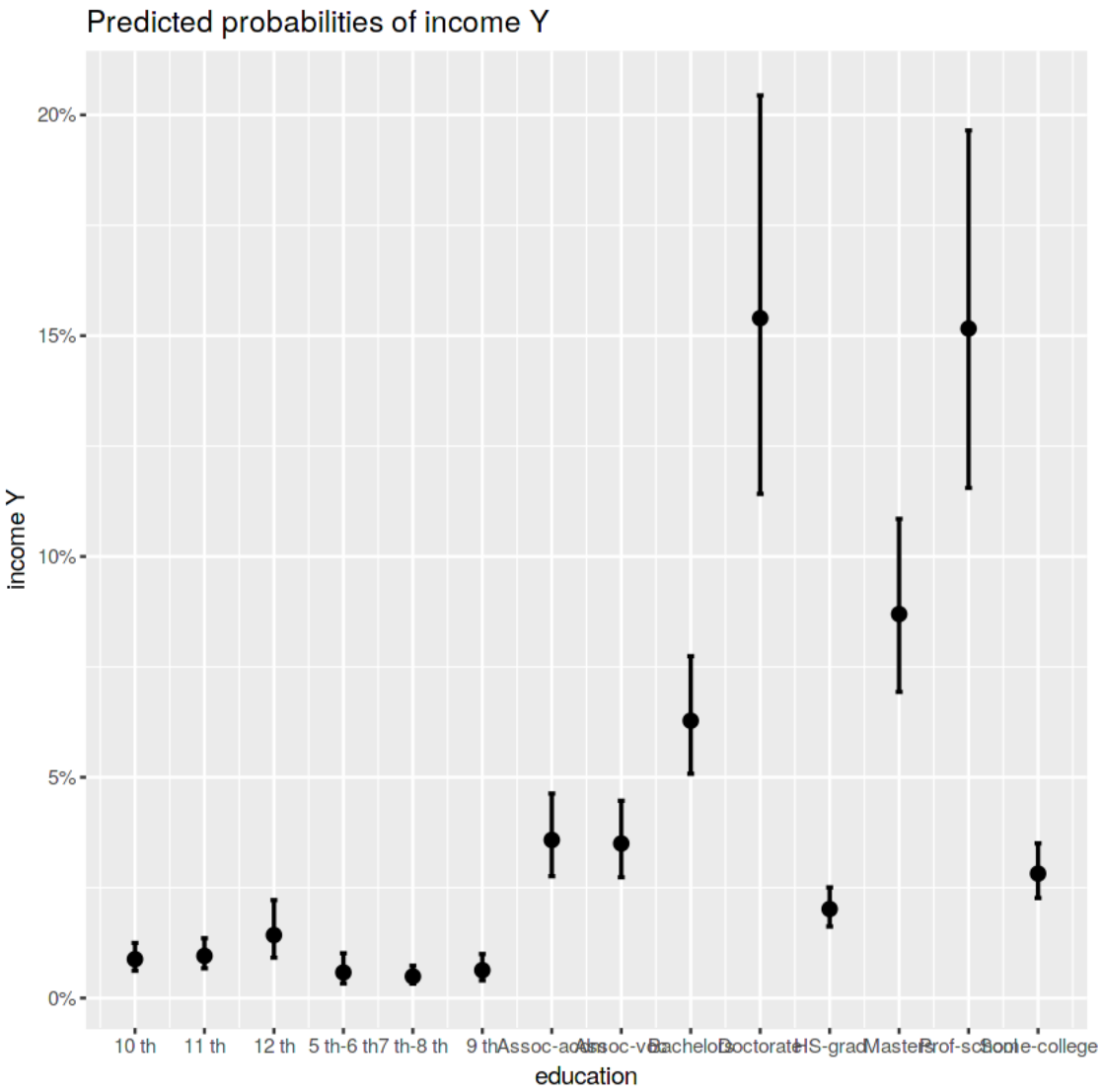

Figure 6 CI of marginal effect of education

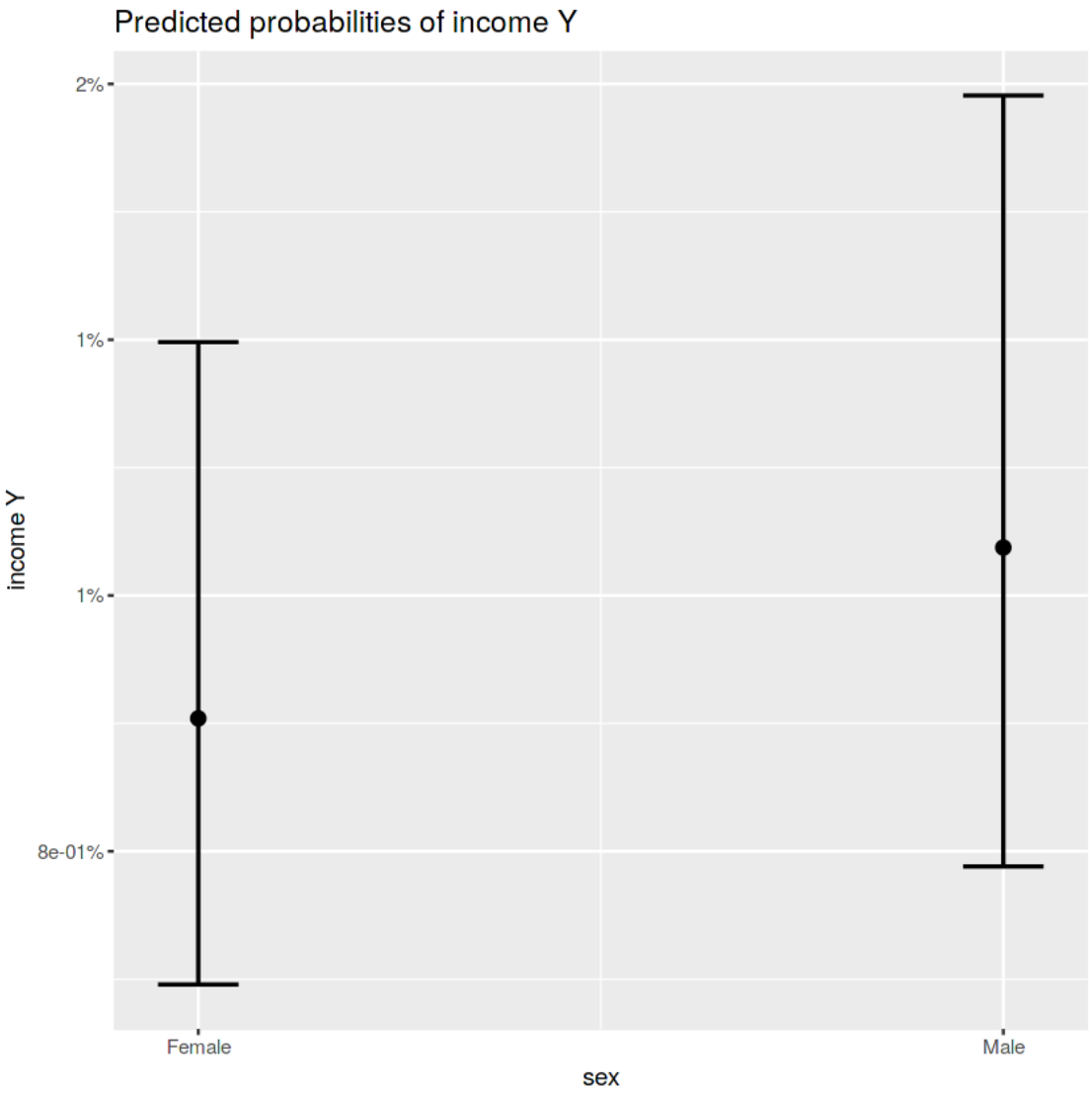

Figure 7 CI of marginal effect of gender