Census Income Analysis (Non-Technical)

Amy Edwards, Alice Fu, Persid Koci, Carl Snow, Jenny Tong, & Danyang Xiong

The group analyzed an adult census income dataset extracted from the 1994 Census bureau database (UCI Machine Learning, 2016), with a couple of goals. One goal is to use the data to determine if we can predict if a person will make over $50k per year.

Using a logistic regression analysis with the full equation of p = Pr(dincome = 1) = -7.5738 + .027 * age + .2809 * edu_num + .000366 * capital_gain + .000660 * capital_loss + .0208 * hours_per_week + 1.2277 * dworkclass2 + 1.6779 * dworkclass3 + 1.7319 * dworkclass4 – 2.3294 * dmarital2 – 2.5443 * dmarital3 -2.6342 * dmarital4 * -4.4672 * dmarital5 + 0.615 * doccup2 + 1.2094 * doccup5 + 1.1862 * dnativec3, Jenny found that the final model could explain 37.98% of the variability and was 78.67% accurate at predicting negatives and positives, i.e. whether an individual earned more than $50k a year, on test data. The most influential variable on earning more than $50k per year was capital gain, or income from investments other than a salary.

Carl's final model statement on analysing logistic regression on income is Pr (dIncome = 1) = 0.498 + age * 0.006 + education_num * 0.035 + dMstatus * 0.216 + dJob * 0.025 + dRelation * 0.095 + dRace * 0.099 + capital_gain * 0.0005 + capital_loss * 0.00016. I found that education and capital gain has the greatest influence on someone making more than $50,000 a year. Most of the data are white males, and for me, native country is only United States, so Most demographics did not show up for predicted higher income.

Alice's final model statement is Log(d_income = 1/d_income = 0) = -11.6287 + 0.4507education_num + 0.000442capital_gains + 0.000671capital_loss + 2.9383d_marital + 0.5097d_race_white + 1.1912d_country + 0.000977interaction2 (When d_marital = 1, d_race_white = 1, d_country = 1). It was found that the more capital a person loses the greater his/her chances are of making over 50k (0.000671capital_loss). Capital gains affects the chances that a person makes >50k the most. The more education someone has, the more income they have. People who are married, white, or come from the U.S, increases their chances of making over 50k also. Also, the older the person is and the more hours they work per week, the greater their chances are of making >50k.

Persid's final model: Log(p/(1-p))= -2.9433 + 0.9673*sex1 + 0.9373*workclass3 – 0.7879*edLevel1 – 4.4091*maritalStatus1 – 0.7531*occupation1 – 1.3206*occupation2 + 0.7088*occupation3 + 0.7557*occupation5 + 1.0489*relationship1 + 0.0428*hrs_per_week + 0.000384*capital_gain + 0.000587*capital_loss – 2.5824*continent2 + 0.0597*age*maritalStatus1

From the equation it can be concluded that males, employees in federal government, professionals, executives, wives, longer hours per week, capital gain, and capital loss have a positive association and odds to earn an average income more than $50K get increased. High-school graduates, not-married people, working class people, people working in services, and people from Latin America have negative association and odds to earn an average income more than $50K get decreased.

The second goal is to predict how many hours per week a person will have to work in order to make over $50k per year. Amy's full linear regression analysis found that the final model for hours per week was only able to predict 10.77% of the variability. The final equation to determine hours worked per week in order to make more than $50k a year is

48.09 + 5.52*(edu_doctorate*occup_ProfSpecialty) + 32.078 *(edu_AssocVoc*occup_TransportMoving) + 7.568*(edu_ProfSchool*occup_ProfSpecialty) - 0.154*age - 2.253*d_raceAPI + 6.18*d_rlnshpNIF + 6.3*d_rlnshpHubs + 6.82*d_rlnshpSngl - 3.785*occup_AdmClerical - 2.91*occup_CraftRepair + 5.9*occup_FarmingFish- 4.93*occup_MachineOpIns - 3.48*occup_ProfSpecialty - 5.4696*occup_TechSupport. The most influential variables determining hours worked per week are education_Assoc-voc and occupation_Transport-Moving. If a person as an Associate Vocational degree and also works in Transport Moving, they will work 32 additional hours per week more than a person who does not have those characteristics, holding all other variables constant.

Danyang's final model was: hours_per_week = 43.47034-0.049 age + 0.00004216 capital_gain - 1.6397 marital_status3 -3.61086 marital_status5 + 3.91889 occupation3 + 3.0039 occupation4 + 2.36954 occupation11 + 1.69361 occupation13 - 6.24055 relationship5 + 2.67823 income1. Working hours per week may influenced by several factors, such as the power of execution, confidence, personality, and so on. If a person is more energetic and efficient, it will take less time to finish the work, and have a higher chance to get promotion. In the dataset, sex, race, marital status may impose few impact on the working hour per week, and education level, occupation, and work class may have great impact on the dependent variable.

The end results definitely lend themselves to further research. Surprisingly, sex and native-country did not play big parts in the final models. It's quite possible that there are more details that need to be taken into consideration to find an accurate model for either goal. Some of the future work could involve learning more about the people in the model - specifically their occupations and specific financial situations. Capital-gains and occupation were two significant predictors for the models in both goals.

Answering these questions about income and work hours can lead to understanding how people make money, who makes the most money, and how hard they have to work to do so. There are endless avenues to take this research such as public advocacy for equality in education, public spaces, and wages. It could help governmental or private investments decide where they could best target inequality. It could even be used for economic studies if the data is gathered for many years in row.

Future work could be made in terms of getting more details in the dataset, like the location of each person, or if the person owns a business vs. being employed. Also, the model even if not improved could include more variables to create a full picture how different variables affect the income.