Census Income Analysis

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

Abstract

With this adult census income dataset extracted by Ronny Kohavi and Barry Becker from the 1994 Census bureau database (UCI Machine Learning, 2016), we have 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. This will be analyzed with a logistic regression analysis, for which the final model in one case was able to predict 37.98% of the variability. Logistic regression analysis methodology included exploratory analysis, residual analysis, multicollinearity analysis, examining significant variables, and model selection. The test set performance of the final model was a sensitivity of 88%, an accuracy of 78.67%, a precision of 54.72%, a specificity of 75.53%, and an f-metric of 67.48%. The most influential independent variable was capital_gain. Since the model's performance and explanatory power can be increased, the recommendation is to explore different modeling techniques in the future, including naive Bayes and a decision tree model.

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. This will be done using a full linear regression analysis, including exploratory plots, examining correlation, and model selection in order to determine the most significant factors. The final model for hours per week was only able to predict 10.77% of the variability. There must be many outlying conditions other than the variables used here to determine how many hours per week a person will work. Likely, we need more details about their job, their financial situation, and their work ethic. 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.

Introduction

Amy

When considering how many hours per week a person will have to work to make over $50k/year, some hypotheses come to mind: 1. The individuals native country will be a determining factor, and those from countries with a lower GDP will have to work more hours in order to meet the same income goal as their counterparts (Classora, 2016), 2. The subjects sex will also be a determining factor, as it relates to education and occupation due to the gender wage gap (Hartmann, 2005), and 3. The income bracket over $50,000 will be dominated by those with higher education. Exploring this subject is important to researching wealth inequality between genders,

races, and comparing how immigrants fare after moving to the United States. It is likely that due to the slow wage growth beginning in the 1970s, young people are having to work longer hours in order to make the same money as an older generation (Levy 1998). Validating these theories gives credibility to social and political movements that are fighting for higher wages and public access to systems such as job training, transportation, and education all in the name of equality.

Jenny

The goal of this analysis was to see which independent variable has the greatest effect on whether an individual made more than $50,000 a year, based on a dataset of Census income data from 1994. The null hypothesis is that none of the independent variables had a significant effect on the dependent variable, whether an individual made more than $50,000 a year. Hypothesis 1 is that at least one of the independent variables has a significant effect on income outcomes. It is important to determine which factors individuals can focus on, if they can affect those factors at all, in order to earn a high income.

Hypothesis 2 is that the country of origin will have a significant effect on income outcomes. This is informed by a paper suggesting that immigrants earn higher wages since they are "a self-selected group, and as a result, [they] may be more able and highly motivated" (Borjas 1987). Hypothesis 3 is that number of years of education will have a significant effect on income outcomes. This makes sense as education affects employment types and related incomes, presumably.

A cross-country study found "sizeable differences in...years of schooling between the US and low inequality EU countries" (Devroye and Freeman 2001). This means that since inequality is more prevalent in the US, years of education has an effect on outcomes, whereas it is not as strong a factor in countries with less inequality. Another cross-country study found that "returns to schooling increase over the wage distribution...the earnings increment associated to schooling is higher for those individuals whose unobservable characteristics place them at the top of the conditional wage distribution" (Martins and Pereira 2003). This suggests that increased education helps those who are already naturally advantaged, perhaps by gender or age, suggesting an interaction variable.

Danyang

The goal is to find what affects the number of working hour per week. I choose this as my research subject since I’m interested in knowing the distribution of working hours among different people. To be more specific, to know whether education level, working class, occupation, age, marital status, race, and relationship have effect on working hour. My hypothesis is higher education level, higher working class may have great influence on the

working hour since I believe that those smart people tend to spend more hours working in the modern society. They are eager to learn new things, and they usually have creative thought to manage company's operation effectively. My second hypothesis is married man works harder than married women since men tend to have greater responsibility on their family. Economic Sociology (1983) argues "But since this means that married man work about 25% more hours per week for pay than married women do" .

Persid

The goal of this project is to determine a regression model that predicts whether the income of a person exceeds $50,000 per year based on census data. The income levels in the dataset were classified in two categories, less than or equal than $50,000 and more than $50,000. Therefore, a logistic regression analysis will be performed. The results of this project may serve to understand better the contributing factors on the income level, and investments can be made from government agencies or private entities to target directly the problem of poverty and inequality of the population (Hartmann 2005) .

Carl

The goal of this analysis was to see which independent variables have the most significance on whether an individual made more than $50,000 a year, based on a dataset of Census income data from 1994. For the goodness of fit, the null hypothesis is that none of the independent variables will have a significant effect on the dependent variable. The alternate hypothesis is that at least one of the independent variables will have a significant effect on income.

I believe that number of years of education will have a significant effect on income. This makes sense as education affects occupation and status. Also, I believe as age increases, probability of an individual making more than $50,000 a year goes up. I Also think that a married man works harder than married women since men tend to have greater responsibility on their family. Economic Sociology (1983) argues "But since this means that married man work about 25% more hours per week for pay than married women do".

I will test my believes while analyzing the income data. Compared to my group members, I am only considering observations which have native country as United States since comparing other countries is not part of my hypothesis.

Alice

The goal is to see if I can predict the chances of someone making >50k with certain attributes.

H0: none of the predictors can predict the y-variable (d_income)
H1: the more education a person has the greater their chances are of making >50k
H2: single people have a greater chance of making >50k than their married counterparts

According to a recent meta-analysis, the effect of educational expenditure on growth is positive. A 0.2-0.3% increase in growth for an increase in education expenditure by 1% of GDP (Carmignani). This is important because if I can predict how the amount of education a person has affects his/her income, then the broader more important question to ask is if the government should spend more money on its educational to grow their GDP and increase its people's living conditions.

A study by W. Bradford Wilcox, director of the National Marriage Project at the University of Virginia, and Robert Lerman, an economics professor at American University, insinuates that single women make more than their married counterparts (Wilcox 36). Even though this is not statistically significant, they found that single women between 28 and 30 years old make $1,349 more per year in individual income. This is a widespread question that people have. Are single people more successful than married people, in monetary terms?

Methodology

Amy

Kohavi and Becker extracted records from the 1994 census with the following conditions:((age > 16) && (average gross income > 100) && (fnlwght > 1) && (hours per week > 0)). This dataset was retrieved from the UCI Machine Learning Repository via Kaggle [dot] com (UCI MAchine Learning, 2016). I continued to refine this selection to gain a cleaner data set. Using R studio, I began by noting columns with many missing values and I wrote a query to remove those from my data. Then I filtered to only look at cases where income is >50K because that is the selection relevant to my question. Finally, I took a random sample of 27% of the total data in order to get a selection of 2,027 cases.

Next, the data was loaded in to SAS for exploratory analysis. The first steps here are to visualize the data using boxplots, histograms, and scatterplots and to describe the data with statistics and frequency matrices. Based on these, I concluded that the model for hours per week would be a linear. The text variables sex, race, occupation, education, marital-status, relationship, workclass, and native-country were all converted into dummy variables in order to consider them in the model. Based on the histograms of capital-gain and capital-loss, those variables were logarithmically transformed to have a normal distribution (Appendix A.2).

I fit a basic model to get a starting point and examine Pearson Correlation Values. There were no issues there. I took a look at the residual plots – most were dummy variables and therefore not relevant. The other residual plots showed no outstanding issues. The normality plot showed a roughly linear trend. Due to the large number of dummy variables, in order to determine interaction variables I used glmselect with various model selection techniques. Based on the models given, I included all of the interaction variables suggested into further analysis.

I analyzed multicollinearity by using the vif function of proc reg. Where vif > 10, there are multicollinearity issues. I recursively removed variables where vif > 10 and re-ran the model with vif in descending order. I then fit the base model to examine outliers and influential variables, using outliers = Pearson Residual > 3 and influential points = |DfBeta > 2/sqrt(2027)|. After recursively removing cases that were outliers and influential, the adjusted R squared value did not change at all. Therefore, I decided to use the original set of data for the rest of the analysis.

I then split the data into training and testing sets of 75% and 25%, respectively. I will use this tool in order to test how accurate the final model will be. I performed model selection with the training set. For the models I tried adjusted R squared, CP, Stepwise, Forwards, and Backwards. Once I had the variables selected for each model, I analyzed those for significance. The hypothesis test states that a p-value must be than alpha = 0.05 for it to be statistically significant. Upon distilling the 5 models, I determined that the model gleaned from Forward selection had too many variables, making it unreasonable to use. The model for Adjusted R squared and CP ended up as the same. The final models to compare were Adjusted R squared, Backwards, and Stepwise. When comparing the models, the goal is to minimize RMSE, MAE, and CV while maximizing R squared and adjRsq. I will compare the model residuals, looking for constant variance, independence, normality, and linearity. With the final model, we can write the model equation and make predictions.

Jenny

The Census Income dataset was obtained from http://mlr.cs.umass.edu/ml/datasets/Census+Income, and was extracted by Barry Becker from the 1994 Census database. The original dataset had 48, 842 observations, but a random sample of 2000 observations was taken (Appendix B.Figure 1) (SAS Institute, 2010). Originally, there were 6 numerical variables and 8 text variables. 2 independent variables were dropped from the analysis, fnlwgt and relationship. Fnlwgt was the Census's estimate of approximately how many people fit the characteristics for that observation, and would not have been useful to the analysis. Relationship was a field that displayed an individual's relationship status within a family, i.e. Wife, Own-child, Husband, Not-in-family, Other-relative, and Unmarried. This status also seemed unnecessary to include as not having relevance to income outcomes, as well as having overlap with the original marital-status variable.

Since the dataset originally came with many categorical text variables, many dummy variables had to be created. The dependent variable for income was coded as dincome where 1 indicated an income greater than $50,000, and 0 indicating an income less than $50,000. As for independent variables, dsex was 1 for male and the baseline was 0 for female. 3 dummy variables were created to represent levels of workclass, with the baseline of 0 being reserved for an unemployment category of "Without-pay" and "Never-worked." dworkclass2 grouped both self-employed categories, "Self-emp-not-inc" and "Self-emp-inc", into one category. dworkclass3 covered government employment: "Federal-gov", "Local-gov", and "State-gov". The last dummy variable, dworkclass4, was set to "Private" to represent employment in private industry.

There were 6 dummy variables created for various levels of education, with the baseline of 0 being less than a high school education, which covers "11th", "9th", "7th-8th", "12th", "1st-4th", "10th", "5-6th", and "preschool." In ascending order, the education dummy variables represented "HS-grad", "Some-college", "Prof-school", an Associates' category, "Masters", and "Doctorate." The Associates' category represented "Assoc-acdm" and "Assoc-voc." Marital status was split into 4 dummy variables, with the baseline of 0 representing currently married individuals, so the "Married-civ-spouse" and "Married-AF-spouse" designations. In ascending order, the marital status dummy variables represent: "Divorced", "Never-married", a separated category, and "Widowed." The separated category contains both "Married-spouse-absent" and "Separated" itself.

Occupation was a bit tricky to categorize, and best judgment was used in grouping occupations together. The baseline is "Tech-support". In ascending order, the occupation dummy variables represent a mechanical category ("Craft-repair" or "machine-op-inspect"), a service-related category ("Other-service", "Sales", and "Adm-clerical"), "Exec-managerial", "Prof-specialty", a menial work category ("Handlers-cleaners" and "Priv-house-serv"), out-door related work ("Farming-fishing" and "Transport-moving"), and security-related work ("Armed-Forces" and "Protective-serv"). Race was split into 3 dummy variables, with the baseline being "White." "Asian-Pac-Islander" was the second level. "Amer-Indian-Eskimo" and "Other" were grouped together in an other-race bucket for the third level. "Black" was the fourth level.

Native-country presented a categorical challenge as well, and roughly continental regions were decided on as the categorical levels for the dummy variables. Another possibility would have been by the country's level of development."United States" was the baseline so that the native population could be measured against all immigrant populations. The baseline also included "Outlying-US(Guam-USVI-etc)" to stand for U.S. territories. Subsequent dummy variables represented Asia, Europe, Central and South America, and the rest of North America. The second level of Asia included Cambodia, India, Japan, China, the Philippines, Vietnam, Laos, Taiwan, Thailand, Hong and Iran. The Europe level included England, Germany, Greece, Italy, Poland, Portugal, Ireland, France, Hungary, Scotland, Yugoslavia, and Holand-Netherlands. The Central and South American region included "South", Cuba, Honduras, Jamaica, Dominican-Republic, Ecuador, Haiti, Columbia, Guatemala, Nicaragua, El-Salvador, Trinidad&Tobago, and Peru. The other North American region included Canada and Mexico.

There was also an age and sex interaction variable included in the analysis informed by the aforementioned Martins and Pereira study.

After recoding the variables, histogram and scatterplot analysis were conducted, which along with the binary nature of the dependent income variable, suggested a logistic regression analysis. Scatterplots were created to examine the independent variables' relationships with the income variable. A full logistic regression model was fitted and the independent variables were examined for significance. Residual plots were analyzed and observations that were both influential and outliers were removed from the dataset. The model was analyzed for collinearity.

Next, the dataset was split into testing and training datasets. The training dataset was fitted with stepwise and backwards selection methods, and a fitted model was selected after the resulting models were compared. Residual and multicollinearity analysis was conducted on the fitted model and a final model was decided upon. The final model was used to make two predictions on the pre-split dataset that only varied in years of education to see if there was a difference between the two predicted observations' incomes. The training dataset was used to analyze the test dataset's performance. The results are explained in detail below.

Danyang

I found the data from https://www.kaggle.com/uciml/adult-census-income. There are over 30,000 observations, so I organized the data by randomly choosing the first 3,125 observations as my dataset. I deleted all missing values in the original dataset, and I kept the United-States as "native.country" variable, and I dropped other countries because I only did the research for the U.S. people since most people chosen in the dataset come from the U.S.. Then I dropped "native.country" since the rest of data are from the research objects of the United States. I also dropped "fnlwgt" and "education.num" variable. The ratio of train set to the test set is 75:25. The dependent variable is "hours_per_week" (the number of working hour per week). The independent variables are: "age", "capital_gain", "capital_loss", "workclass1" – "workclass5", "education1" – "education15", "marital_status1" – "marital_status5", "occupation1" – "occupation13", "relationship1"- "relationship5", "race1" – "race4", "sex1", "sex_income", and "sex_age". I created "sex_income" and "sex_age" as the interaction variables. My methodologies are regression model, testing performance, 5-folder cross validation method, and calculating prediction intervals.

Persid

Rows with missing data were deleted prior to selecting a simple random sample of 2000 people on which the logistic regression analysis will be performed.

Variables "fnlwgt" (sample weight) and "education_num" were dropped in the analysis. The sample weight is deleted for not being relevant to the model. Education is represented by the qualitative variable "education". Qualitative variables were grouped in fewer categories where possible, and coded as dummy variables (Table 1 in Appendix D) . Two interaction terms are investigated in the model, age*marital_status, and hrs_per_week*sex1.

A full model will be fitted with all the variables. The model will be checked for multicollinearity, outliers, and influential points. Appropriate steps will be taken regarding the dataset and the number of the variables analyzed in the final model, after these checks will be performed.

The final model will be selected based on the comparisons among the results of different selection procedure. The selection procedures will be run on the training data that is obtained after splitting the original dataset. The data splitting is done, based on a ratio 60/40, where 60% of the data is in the training set, and 40% is in the test set. The validation and the predictions are made using test set. Two predictions are made in the model and the results will be shown.

Carl

The Census Income dataset was obtained from UCI Repository, and was extracted by Barry Becker from the 1994 Census database. The original dataset had 48,842 observations. I attempted to use the data set in its raw form but the values at the end of each line were not being read. I coded a Python program to replace every new line character with a comma and new line character. Once I got that working, I knew that I only wanted to use observations where the native country is the United States since it would be more relevant to me and comparing other countries would not come into play. I also right away coded age and hours per week into managable sections: age(Under18[0-17], youngAdult[18-29], adult30s[30-39], adult40s[40-49], adult50s[50-59], adult60+[60-99]) hours(lowPartTime[0-19], highPartTime[20-29], fullTime[30-40], overTime[41-99]). I attained these values by research into typical ranges. So now I took a random sample of 2000 observations (Appendix F. Figure 1) (SAS Institute, 2010). Originally, there were 6 numerical variables and 8 text variables. 2 independent variables were dropped from the analysis, fnlwgt and native country. Fnlwgt was the Census's estimate of approximately how many people fit the characteristics for that observation, and would not have been useful to the analysis.

Since the dataset originally came with many categorical text variables, many dummy variables had to be created. The dependent variable for income was coded as dIncome where 1 indicated an income greater than $50,000, and 0 indicating an income less than or equal to $50,000. As for independent variables, workclass(dWork) has Private(0), Self-emp-not-inc(1), Federal-gov(2), Local-gov(3), State-gov(4), Self-emp-inc(5), Without-pay(6), and Never-worked(7). For marital_status(dMstatus) Married-civ-spo(0), Never-married(1), Divorced(2), Separated(3), Widowed(4), Married-AF-spou(5), and Married-spouse(6). For occupation(dJob) Exec-managerial(0), Sales(1), Adm-clerical(2), Craft-repair(3), Prof-specialty(4), Machine-op-inspect(5), Transport-moving(6), Tech-support(7), Other-service(8), Handlers-cleaners(9), Farming-fishing(10), Protective-serv(11), Priv-house-serv(12), and Armed-forces(13). For relationship(dRelation) Husband(0), Wife(1), Own-child(2), Unmarried(3), Not-in-family(4), and Other-relative(5). For race(dRace) White(0), Asian-P(1), Amer-In(2), Black(3), and Other(4). For sex(dSex), 1 is for male and 0 is for female.

Dummy variable value for each independent variable is based on running a test on frequency, and ordered in little endian.

After recoding the variables, histogram and scatterplot analysis were conducted, which along with the binary nature of the dependent income variable, suggested a logistic regression analysis. Scatterplots were created to examine the independent variables relationships with the income variable. A full logistic regression model was fitted and the independent variables were examined for significance. The model was analyzed for collinearity and there was none. 64 outliers were removed.

Then the dataset was split into testing and training datasets. The training dataset was fitted with stepwise and backwards selection methods, and a fitted model was selected after the resulting models were compared. Analysis was conducted on the fitted model and a final model was decided upon. The final model was used to make predictions on specific data using education_num, dMstatus, dJob, dRelation, and dRace.The testing dataset was used to analyze the performance of the dataset.

Alice

The dataset came from Kaggle (https://www.kaggle.com/uciml/adult-census-income/data). First, I read the adult.cvs file into SAS. During this process, I made sure that none of the data was truncated and that everything was read in properly. Then, any observations with missing data values were deleted from the dataset. 2000 observations were sampled randomly from the dataset and giving a seed number. Now that I had a working dataset, I began making dummy variables. For the variable “workclass", I made two dummy variables. d_work_private included the nested categories: “Private,” "Self-emp-not-inc,” and "Self-emp-inc.” I grouped these three categories together because they made the most business sense to be together. d_work_gov included the categories: "Federal-gov," "Local-gov," and "State-gov.” These were grouped together because they were are related to working in the government. The nested categories "Without-pay" and "Never-worked” were set as the base case. Even though it would not really matter what I picked as my base case, I picked the least important categories to be in my “workclass” base case so that analyzing the model later on would be intuitive. "Without-pay" and "Never-worked” had the least amount of observations. For the “education” variable, I made four dummy variables. d_edu_primary includes the nested categories: “Preschool,” "1st-4th," "5th-6th,” and "7th-8th.” d_edu_primary is grouped by everything before high school. d_edu_secondary includes the nested categories: "9th," "10th," "11th," "12th," and "HS-grad.” d_edu_secondary is grouped by high school because there is a transition from middle school to high school and to college. So, the high school level of education makes another category. d_edu_college includes the nested categories: "Some-college," "Bachelors," "Assoc-acdm," and "Assoc-voc.” d_edu_college is a category for people who have a Bachelor degree and or some sort of college experience. d_edu_higher includes the following nest categories: "Prof-school" and “Masters.” d_edu_higher includes people who have a little more schooling than people who just have a Bachelor degree. The base case is "Doctorate.” People who have doctorates have reached the pinnacle of their education. This is the highest level of education. Essentially, I

grouped the education dummy variables by the transitions a student makes throughout their academic career. For the variable "marital_ status," I felt that it would be most interesting to compare married people not not married people. So, I only made one dummy variable for married people called "d_marital" that includes "Married-AF-spouse," "Married-civ-spouse," and "Married-spouse-absent." For "occupation," I made two dummy variables. d_occup_service included "Farming-fishing," "Transport-moving," "Craft-repair," "Other-service," "Priv-house-serv," "Protective-serv," "Armed-Forces," "Handlers-cleaners," and "Sales." These are all service jobs. d_occup_exec included "Exec-managerial" and "Prof-specialty." These are more advanced jobs with managerial positions and higher skill sets. The base case are the technology jobs: "Tech-support," "Machine-op-inspct," and "Adm-clerical." For "relationship," I made one dummy variable separating the people with a family and the people who do not. d_relat_family includes "Husband," "Wife," "Other-relative," and "Own-child." The base case includes "Unmarried" and "Not-in-family." For race, I simply made four dummy variables, one for each race ("White," "Asian-Pac-Islander," "Amer-Indian-Eskimo," and "Black") except for the base case ("Other"). For the "sex" variable, I made one dummy variable for "Male" with "Female" being the base case. I also had to recode the y-variable "income" because it was binary. I made the dummy variable d_income to include the category ">50k" because I want to predict how much of a chance a person has of making over 50k based on the certain qualities he or she possesses described in the x-variables. Finally I tried two interaction terms. One of the was thrown out by the model because it was not significant. I thought that that age and education_num might have some correlation o I grouped those two categories to predict who makes over 50k. However, this interaction term was thrown out of the model. So, I tried another interaction term: "interaction2." This interaction term is between age and hours_per_week (how many hours a person works per week). Interaction2 was eventually found to be significant in the model selection process. I dropped fnlwgt because it is a weighted variable of how many people fall into that observation. After all of the needed variables were recoded. I began the data exploration phase, making frequency tables, descriptive statistics, and box plots to understand the data better. Then, I fit the full model with logistic regression because my y-variable "d_income" is binary. I fit the model with the standardized estimates, r-square, correlation matrix, and residual plots to check for multicollinearity and to exclude any outliers or significantly influential points. After the multicollinearity and outliers/influential points were handled, I fit the full model one more time. Then, I split the dataset into training and testing with a sample rate of 75% to 25%, respectively. The new y-variable for the training was created "train_y" for the values that were selected for the training set. Three model selection methods were run with r-square to find the best model: stepwise, backward, and forward. The with the best model, the full model was fitted with residuals once more to check for outliers and influential points. After this was handled, the full model was fitted once more. Finally, the classification matrix was produced to measure the sensitivity/recall, accuracy, precision, specificity, and F-metric of the model. If the values for these categories were unsatisfactory, then various factors were changed: how dummy variables are created, how outliers/influential points are handled, new sample rate, new seed number, etc. Then, through the selection model process, a new best model for that trial is created. All of the "best" models produced are then compared to see which one of them has the highest values in these categories produced by the classification matrix. Finally, two prediction are made based on the best model.

Analysis, Results, and Findings

Amy

In the first steps of exploratory analysis, a scatterplot of all the continuous variables was produced (See Appendix A.2 Figure 1), which did not show any trends in their relationship. From this point on, it became clear that there would not be a very accurate way of predicting how many hours per week one will have to work to make over $50,000 a year. Box plots were produced to examine hours_per_week vs the categorical variables (appendix A.2 Figures 2-9). In summary, they show that men work slightly more hours than women. Those who are self employed have a higher average of hours per week than those in the government or private sectors, though the 25th percentile of all work classes are approximately the same. Marital-Status appears to have no effect on hours worked per week and they have very similar values for the IQR, mean and median. However, being married with a civilian spouse offered the greatest variability with the highest maximum and lowest minimum. The occupation predictor also has similar mean and median hours worked per week, appearing to be between 40 and 50 hours. Notably, members of the armed forces only work 40 hours per week. Many people from certain native-countries are shown to only work a specific number of hours per week, but this may be due to lack of people from those countries in the data rather than a trend. Overall, relationship seems to have an effect on hours worked per week if the relationship variable equals wife. When it comes to education, if a person has not completed high school, there is much less variability in the hours they could work per week. Here I'd hypothesis that perhaps those people are working lower-skilled jobs because they have less education and their jobs are less flexible. A final histogram of hours-per-week confirmed that the best model would be linear, because the distribution is roughly normal (Appendix A.2 Figure 14).

I created dummy variables for the categorical variables. A complete table of the transformations can be found in Appendix A.2 Table 1 and Table 2. I kept many of the variables ungrouped because I wanted to see how relevant that specific occupation or educational attainment was. For the native-country categorical variable, I labeled the dummy variables 1-40, in order of increasing GDP (Classora, 2016) for ease of use.

In order to test interaction variables, I used the glmselect class option. This allowed me to split the dummy variables in the model education rather than multiply each dummy variable by each other dummy variable manually. I ran the glmselect model four times, using stepwise(cp), stepwise(adjrsq), forwards and backwards selections. The only interactions that made it into the model outputs were

education_Doctorate*occupation_Prof-specialty, education_Assoc-voc*occupation_Transport-moving, education_Bachelors*occupation_Prof-specialty, marital-status_married*relationship_Wife, and marital-status_married*relationship_Child.

I then included these interaction variables into the rest of my model analysis, and treated them as a regular variable. The relationship between education and occupation appears to be highly significant to predict working week hours. As does the relationship between marital-status and relationship, which makes sense because they are speaking to similar criteria.

Next, correlation was examined among the independent variables using the Pearson Correlation Coefficient. There were no issues, and in fact, the variables seemed highly uncorrelated. I next looked at multicollearity and recursively removed variables that were the worst problem and re-ran the model. For this I used the VIF value, and if VIF > 10, there is multicollinearity, but because the model changes each time, I removed the largest VIF each time I ran the model. See appendix A.2 for the final multicollearity model (Figure 27) as well as the residuals (Figures 15-26) associated with that model. Most of the residuals are dummy variables, which do not show accurate assumptions of constant variance, independence, or normality. Those residual plots of the continuous variables (fnlwgt, capital-loss, capital-gain, and age) do not show an even distribution of points around the zero line, and they all show some outliers. This is not enough to deter the analysis, however, because the QQ plot of hours-per-week shows it to be almost linear, except in the extremes. Finally, in order to better represent the data, I examined the outliers and influencers. I ran the model a few times, deleting some of the most egregious outliers and influences each time, and in the end the adjusted R^2 value stayed the same as it was before the deletions (9.97%). Because the outliers and influencers may actually be very valuable to the model, as opposed to mistakes, and the model did not get better, I decided to continue the analysis with the original complete data.

At this point none of the model variables are collinear and model selection for hours-per-week can begin. I split the data into a training and testing set of 75%, 25% respectively in order to test the model I find on unseen data. There are 5 selection methods we will use to determine the best model. Selection Method 1 uses Adj-R^2 to find a model with the highest R^2 value. Method 2 uses the backwards elimination method which begins with the full model and recursively removes the least significant variable until they are all significant. Selection Method 3 is stepwise, which alternates between dropping a variable and re-considering the addition of all previously dropped variables at each step. This method usually gives the smallest number of variables compared to other methods. The 4th method is CP selection method for models of least squares. The smaller the value of C(p) the more precise the model. The last method is forwards selection, which starts with no variables and adds all that are deemed significant. The forward selection result gave a model with twice the amount of variables as any other so we will not mention it again.

With models 1, 2, 3 and 4 from adjusted R squared, backwards elimination, stepwise, and C(p) selection methods, respectively I then examined for variable significance. The hypothesis test states that a p-value must be than alpha = 0.05 for it to be statistically significant.The model for Adjusted R squared and CP ended up as the same. The final models to compare were Adjusted R squared, Backwards, and Stepwise. When comparing the models, we want to minimize RMSE, MAE, and CV while maximizing R squared and adjRsq. We will compare the model residuals, looking for constant variance, independence, normality, and linearity. (see Table 3 and Table 4 in appendix A.2 for training and testing comparisons for Models 1, 2, and 3). Model 1 did better than the other two models in all of the categories, so we will use that to fit our final model on all of the data. The result is:

Hours_per_week = 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 strongest predictor in this model is the interaction between education_Assoc-Voc and occupation_Transport-Moving. If both these categories are true for a person, they will work an extra 32 hours, holding all of the other variables constant. The model's Adj-R^2 is only 10.77%. This means that this model only explains 10.77% of the variability in hours worked per week, and, in other words, is not a very good predictor. The F-value is 18.47 and the associated p-value is <.0001. This means that the model is better than the intercept-only model! There may be many outside factors that are better at explaining how many hours per week one will work. These factors might include greater details on occupation and work ethic, as well as a better understanding of one's financial situation.

For a person to have a doctorate degree and an occupation of specialty professor, holding all other variables constant, their hours worked per week will increase by 5.52. If a person has the education of professional school and the occupation of a specialty professor, all else constant, they will work 7.568 more hours per week. These are interesting because a doctorate degree is higher than a professional one. In the same job position of specialty professor, those with a higher degree work less.

Age plays a small part, where the older you are, the less you work. For each year, all else constant, a person will work 0.154 less hours per week. If a person's race is Asian-Pac-Islander, all else constant, they will work 2.253 less hours per week. They'll also work less if your occupation of Admin-Clerical, Craft Repair, Machine-Op-Instruct, Prof-Specialty, or Tech-support. Holding each other variable constantly, they would work 3.785, 2.91, 4.93, 3.28, or 5.4696 less hours per week, respectively. However, if their occupation is Farming-Fishing and all else is constant, they will work 5.9 more hours per week. This makes sense because outdoor labor is typically a job with long hours for a specific season, and then a break in the winter. A technique not discussed in class, that would be interesting to implement on this subset is Time Series Forecasting. Blue collar occupations such as farming and fishing have seasons so in order to accurately predict how much they will work, knowing the trend of the year would be valuable (Roubinchtein, 2017).

The last significant variables are relationship Not-in-Family and relationship Husband. Those not-in-family will work 6.18 more hours a week, all else constant and husbands will work 6.8 more hours a week under the same conditions. This could be due to men often being the breadwinner of a family, though it's interesting that sex was not a contributing variable to the model. In 1993, the year before this data was collected, men worked 8 more hours a week than women (Rones, 1997).

Nevertheless, with the final model acquired (appendix A.2 Figure 28), we can use it to make predictions. For example, the two situations I devised for practice are:

1. Husband working for himself incorporated,, 35 y/o his race is black and he is from England. His capital loss on the company is $2000. Has a bachelor's degree. how many hours per week will he work?

2. 28 y/o woman working in the federal government with a masters degree. Never married in tech-support. She is from the US. How many hours per week will she work?
For problem 1 - the man is predicted to work 49 hours per week with a 95% confidence interval that the actual hours worked are in the range of 48.2191 hrs - 49.7814 hrs. The 95% prediction interval for a future response is 29.3971 - 68.6032 hours. For problem 2 - the women is predicted to work 38.3077 hours per week with a 95% confidence interval that the actual hours worked are in the range of 35.5968 hrs - 41.0187 hrs. The 95% prediction interval for a future response is 18.5336 - 58.0818 hours (results shown in appendix A.2). Based on these predictions, the women is actually working less hours than the man for both of them to make over $50k a year (appendix A.2 Figure 29). This invites further study to see if her master's degree is the reason she is working less hours per week or if there is a gender disparity. Unfortunately in this data set, it is unknown which of the two people are making more money.

Jenny

Exploratory analysis was conducted on the dataset with a histogram as shown in Appendix B.Figure 2. Due to the binary nature of the dependent variable income, which had been recoded as dincome, the histogram did not show much. The mean was .244, which is not very high, and it seemed from the distribution of the 0's and 1's on the histogram that there were far more people with incomes below $50,000 than incomes above $50,000. This makes sense as $50,000 is a high income in 1994 and a smaller portion of the population can be expected to make a high income. The histogram suggests a logistic regression.

A scatterplot matrix was attempted but failed to generate meaningful output with 33 independent variables. The scatterplots of dincome against age and number of years of education (Appendix B.Figures 3 and Appendix B. Figure 4) were examined, but again they failed to reveal much since most of the points were clustered either at the top or the bottom. The case is looking stronger for logistic regression analysis.

Boxplot analysis was conducted on several independent variables against dincome: age, capital_gain, dmarital3, doccup4, doccup5, edu_num, and hours_per_week. Appendix B.Figure 5 for age shows that, somewhat surprisingly, the IQRs of individuals with incomes above and below $50,000 are similar, although it is larger for those with incomes below $50,000. The average ages also seemed similar at around 40. The main difference was in the ranges, as the below $50,000 income had a very wide range that extended to all ages. It makes sense that the age for those making above $50,000 is higher as people need time to start their careers and earn more money.

Appendix B.Figure 6 shows a large range of capital gains for those with incomes above $50,000 and a small range for those with incomes below $50,000, indicating that capital gains probably do not have much effect on income. It makes sense that those with a smaller income would not have much in capital gains. The average capital gains seemed similar for both income

groups, at a few thousand dollars. This indicates that most of the people in the dataset earned their money primarily through their income.

Appendix B.Figure 7 shows a large range for those who have never married for both income levels, though the IQR for the lower income level spans the entire range. The average for the lower income level is higher than the average for the income level above $50,000. Dmarital3 is a dummy variable so it is a bit more difficult to interpret, but it seems as if remaining single does not have a strong impact on income; if anything, more single people remain at the lower income level. This is a bit surprising since single status would enable one theoretically to focus more on one's career.

Appendix B.Figure 8 shows the opposite effect from figure 7; there is a large range for both income levels for those working in executive-managerial occupations, but the IQR spans the entire range for the higher income level and the average for the higher income level is also higher than the average for those with incomes below $50,000. This indicates that there are more people with higher incomes working in executive-managerial positions, which makes sense as those positions tend to be higher paid. Appendix B.Figure 9 is similar to figure 8, which makes sense as it is a boxplot of individuals with professional specialty occupations, which also tend to be higher paid.

Appendix B.Figure 10 shows high averages for both men and women, though the average for men is higher. The range for both sexes is quite large but the IQR for women spans the entire range, indicating there are more women in the observations who have an income of $50,000 or less. Since there is a gender-based pay gap, it is not surprising that women have a lower chance of having a higher income.

Appendix B.Figure 11 shows a slightly higher average for those with higher incomes when it comes to years of education, though both are around 10 years of education. The higher income level also had a larger IQR whose upper bound is at about the 12-year mark, so people with higher incomes tend to have more years of education. This makes sense as education is often an investment in obtaining a higher income. Both years of education ranges top out at around 15, so there are lower paying jobs that require the maximum years of education.

Appendix B.Figure 12 shows a slightly higher but similar average, at around 40 hours, for those with higher and those with lower income levels. However, note that the IQR for those with incomes greater than $50,000 extends up to 50 hours a week, while the IQR for those with lower income levels tops out at the average. The range of the lower income level does extend up to almost a 100, so there are individuals who are working long hours for low pay. It makes sense that on average those with higher incomes tend to work more hours.

Next, a Pearson correlation matrix was generated to check for multicollinearity but it is ineffective for a categorical dependent variable like dincome (Appendix B.Figure 13). Logistic regression was run for a full model with all variables. The full model has really high AIC and SC values, of 1433.308 and 1623.738, respectively, which is unfortunate (Appendix B.Figure 14). The r-squared value is also a bit low at 0.3735. However, the goodness of fit test looks good with a likelihood ratio of 935.2461 and a p-value of less than 0.001, so the null hypothesis is rejected.

However, there seem to be many outliers and influential points according to the Pearson residuals exceeding +-3, and dfBetas greater than |Dfbeta| > 2/($\sqrt{n}$), n = 2000, so |Dfbeta| > 0.0447 (Appendix B.Figure 15).

The initial analysis of independent variables' significance based on p-values greater than 0.05 suggests insignificant independent variables of: dsex, dworkclass2, dedu2, dedu3, dedu4, dedu5, dedu6, dedu7, doccup2, ooccup3, doccup6, doccup7, doccup8, drace2, drace3, drace4, dnativec2, dnativec3, dnativec4, dnativec5, and the interaction variable, age * dsex (Appendix B.Figure 16). The occupation related dummy variables correspond to the occupation categories the boxplot analysis excluded, which is a good sign for the executive-managerial and professional-occupation based variables. Race and native-country are insignificant except for one variable, which seems unfavorable for hypothesis 2, that being an immigrant would have a significant impact on achieving a higher income level. Dsex is insignificant, which probably influences the interaction variable's insignificance, and is encouraging for gender-based salary differences.

Multicollinearity was low overall, except for the interaction variable age * dsex, which had a -0.88 correlation with age, and a -.9524 correlation with dsex. (Appendix B.Figure 17). Dworkclass2 and dworkclass3 had a correlation of .8892, and dworkclass4 and dworkclass2 had a correlation of .9347. Dworkclass4 and dworkclass3 have a correlation of .9256. Multicollinearity for these work class dummy variables can be safely ignored because they are levels for a categorical variable with 3 or more levels.

Returning to the analysis of influential points and outliers, there were 12 observations found to be both influential points and outliers with Pearson residuals greater than +-3 and dfBeta values greater than 0.0447 , and these were removed first (Appendix B.Figure 18). The model improved to have an r-square of 0.3904 and an improved likelihood ratio of 983.9645 (Appendix B.Figure 19). AIC and SC have also decreased to 1356.354 and 1552.175, respectively.

Next, outliers were removed in order to improve the model further, but since they only marginally improved the model, they were kept in the dataset (Appendix B.Figure 20). The rejected model had an r-square of .4061 which is only about a .01 improvement (Appendix B.Figure 21). The likelihood ratio only increased slightly as well to 1029.7460, and there were marginal improvements to AIC and SC, respectively, at 1284.306 and 1479.915. The largest deciding factor in rejecting this model was the too-small improvement in r-square.

The dataset without observations that are both influential points and outliers was split into a training and a testing set, and stepwise and backwards selection were used on the training set. Both models resulted in the same 16 remaining variables with r-square of .3798, AIC of 1031.280, SC of 1121.503, and likelihood ratio of 712.2861 (Appendix B.Figures 22 and 23).

The correlation of coefficients matrix analysis reveals the same multicollinearity that can be ignored because it affects dummy variables on a category with 3 or more levels, namely dworkclass2, dworkclass3, and dworkclass4 (Appendix B.Figure 24). The correlation value between dworkclass2 and dworkclass3 was .8823, and the correlation value between

dworkclass2 and dworkclass4 was .9315. In addition, there was a correlation value of .9182 between dworkclass3 and dworkclass4. The other variables did not display multicollinearity.

However, the fitted model still seems to have outliers and influential points according to the Pearson residuals exceeding +-3, and dfBetas greater than |Dfbeta| > 2/($\sqrt{n}$), n = 1491, so |Dfbeta| > 0.0517 (Appendix B.Figure 25). There were 11 observations that were both outliers and influential points found (Appendix B.Figure 26) and 9 observations found that were outliers (Appendix B.Figure 27). Unfortunately, removing observations that were both outliers and influential points did not significantly improve the model, so no outliers or influential points were removed. The rejected model had an r-square of .3970, which is again only a ~0.01 difference from the previous r-square (Appendix B.Figure 28). The AIC, SC, and likelihood ratio were 971.493, 1061.590, and 748.6328, respectively.

So, the final model is 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 (Appendix B.Figure 29). The most significant variable is capital gain, as per the standardized estimates. All of the marital categories have negative parameter estimates. The independent variables will be discussed as odds ratios below (Appendix B.Figure 30):

Age: Increasing age by one year increases the average odds of having a salary greater than $50,000 by 2.7%, and with a 95% confidence that the average increase is between 1.2% and 4.3%.

Edu_num: Increasing the number of years of education by one year increases the average odds of having a salary greater than $50,000 by 32.4%, and with a 95% confidence that the average increase is between 22.1% and 43.6%

Capital_gain: Increasing capital gain by one dollar does not increase the chance of having a salary greater than $50,000. There is an equal chance that the salary will be less than $50,000.

Capital_loss: Increasing capital loss by one dollar increases the average odds of having a salary greater than $50,000 by 1%, and with a 95% confidence that the average increase is between 0% and 1%.

Hours_per_week: Increasing the number of hours per week by one hour increases the average odds of having a salary greater than $50,000 by 2.1%, and with a 95% confidence that the average increase is between .8% and 3.4%

dworkclass2: Being self-employed, as opposed to not working, increases the average odds of having a salary greater than $50,000 by 241.3%, and with a 95% confidence that the average increase is between 3.4% and 1026.9%.

dworkclass3: Working for the government, as opposed to not working, increases the average odds of having a salary greater than $50,000 by 435.4%, and with a 95% confidence that the average increase is between 58.2% and 1711.7%.

dworkclass4: Working in private industry, as opposed to not working, increases the average odds of having a salary greater than $50,000 by 465.2%, and with a 95% confidence that the average increase is between 78.6% and 1688.2%.

dmarital2: Being divorced, as opposed to being married, decreases the average odds of having a salary greater than $50,000 by 9.7%, and with a 95% confidence that the average decrease is between 22.1% and 43.6%

dmarital2: Being divorced, as opposed to being married, decreases the average odds of having a salary greater than $50,000 by 9.7%, and with a 95% confidence that the average decrease is between 5.5% and 17.1%

dmarital3: Not having been married, as opposed to being married, decreases the average odds of having a salary greater than $50,000 by 7.9%, and with a 95% confidence that the average decrease is between 4.6% and 13.3%.

dmarital4: Being separated, as opposed to being married, decreases the average odds of having a salary greater than $50,000 by 7.2%, and with a 95% confidence that the average decrease is between 2.3% and 22.9%.

dmarital5: Being widowed, as opposed to being married, decreases the average odds of having a salary greater than $50,000 by 1.1%, and with a 95% confidence that the average decrease is between less than 0.1% and 21.6%.

doccup2: Being employed in a repair-related industry, as opposed to being employed as a tech support worker, increases the average odds of having a salary greater than $50,000 by 85%, and with a 95% confidence that the average increase is between 16.4% and 193.8%.

doccup4: Being employed in an executive-managerial role, as opposed to being employed as a tech support worker, increases the average odds of having a salary greater than $50,000 by 235.1%, and with a 95% confidence that the average increase is between 114.7% and 423.2%.

doccup5: Being employed in a professional specialty, as opposed to being employed as a tech support worker, increases the average odds of having a salary greater than $50,000 by 167.9%, and with a 95% confidence that the average increase is between 62.6% and 341.5%.

dnativec3: Being from Europe, as opposed to being a native born U.S. citizen, increases the average odds of having a salary greater than $50,000 by 227.5%, and with a 95% confidence that the average increase is between 10.9% and 866.9%.

As per the fitted model analysis, the final model has no significant multicollinearity issues. However, it does have unaddressed outliers and influential points. The likelihood ratio of the final model is 712.2861 at a significance of <0.001, so the null hypothesis that none of the independent variables account for the dependent variable of whether income is greater than $50,000 is rejected. Hypothesis 1, that at least one of the independent variables has a significant effect on the dependent variable, is accepted as a better model. Hypothesis 2, that country of origin has a significant effect on the dependent variable, is only partly accepted since it is only true for Europe. Hypothesis 3, that number of years of education has a significant effect on the dependent variable, is accepted, since number of years of education remained in the final model.

Hypothesis 3 was further explored with two predictions added to the pre-split dataset. Two males, both 43 years old, both have a capital gain of $10,000 and a capital loss of $0, both work 40 hours per week, both are self-employed, both are from Europe, and both work in the repair industry. The only difference is that the first prediction has 16 years of education, while the second one has 10 years of education. Both are expected to have a high probability of making over $50,000, although the male with 16 years of education should have a slightly higher probability.

The predicted probability for a 43 year old male with 16 years of education is computed as p = .98999 with a 95% prediction interval of (0.97183, 0.99648) (Appendix B.Figure 31). The odds of having an income greater than $50,000 for the 43-year old male will increase between 164.27% and 170.87% when he has a capital gain of $10000 and a capital loss of $0, works 40 hours a week, is self-employed, is from Europe, and works in the repair industry. 95% of the time, the predicted probability will fall between 0.97183 and 0.97183.

The predicted probability for a 43-year old male with 10 years of education is computed as p = .94589 with a 95% prediction interval of (0.87326, 0.97795) (Figure 31). The odds of having an income greater than $50,000 for the 43-year old male will increase between 139.47% and 165.89% when he has a capital gain of $10000 and a capital loss of $0, works 40 hours a week, is self-employed, is from Europe, and works in the repair industry. 95% of the time, the predicted probability will fall between 0.87326 and 0.97795. These results align with expected predicted probability results.

Finally, the predicted probability for Y in the training set was computed. The classification table was analyzed to identify a threshold value (Appendix B.Figure 32) of 0.2, as that is where sensitivity and specificity are both high, at 87.1 and 74.5, respectively, for a combined value of 161.6. The threshold value was used to generate the classification matrix table (Appendix B.Figure 33) with a true negative frequency of 281, a true positive frequency of 110, a false negative frequency of 15, and a false positive frequency of 91. The performance statistics computed are: a sensitivity/recall value of .88, an accuracy value of .78672, a precision value of .54726, a specificity of .75538, and an f-metric of .67485.

The final model seems to have a high sensitivity value, so a high proportion of correctly classified positives. However, its precision value is low, indicating that it has a low proportion of true positives. The accuracy, or proportion of correctly classified positives and negatives, is good enough, and the same goes for the specificity value, or the proportion of correctly classified negatives.

Danyang

From the dependent variable, most data are focus on 20-60, and there are many outliers over studentized residual 3 or less than -3. There isn’t a nonlinear relationship, and, from Q-Q plot, we don’t need to transform the dependent variable. “education3”, “education5”, “education6”,

“education8”-“education15”, “marital_status1”, “relationship1”,”relationship3”-“relationship4”, “sex1”, and “sex_age” have collinearity issue:

Capital_loss has p-value 0.7727>0.05;
Workclass1 has p-value 0.616>0.05;
Workclass2 has p-value 0.9056>0.05;
Workclass3 has p-value 0.0728>0.05;
Workclass5 has p-value 0.2918>0.05;
Education3 has TOL 0.08605<0.1, and VIF 11.62156>10;
Education5 has TOL 0.05202<0.1, and VIF 19.22499>10;
Education6 has TOL 0.04931<0.1 and VIF 20.27875>10;
Education8 has TOL 0.00449<0.1, and VIF 222.57717>10;
Education9 has TOL 0.01502<0.1, and VIF 66.569>10;
Education10 has TOL 0.02629<0.1, and VIF 38.03019>10;
Education11 has TOL 0.02102<0.1, and VIF 47.59718>10;
Education12 has TOL 0.00574<0.1, and VIF 174.18336>10;
Education13 has TOL 0.00437<0.1, and VIF 228.94588>10;
Education14 has TOL 0.00793<0.1, and VIF 126.04257>10;
Education15 has TOL 0.02411<0.1, and VIF 41.47615>10;
Marital_status1 has p-value 0.6107>0.05, TOL 0.01667<0.1, and VIF 59.98052>10;
Marital_status4 has p-value 0.3798>0.05;
Occupation1 has p-value 0.7429>0.05;
Occupation5 has p-value 0.9565>0.05;
Occupation6 has p-value 0.6545>0.05;
Occuaption7 has p-value 0.1244>0.05;
Occupation8 has p-value 0.5468>0.05;
Occupation9 has p-value 0.0543>0.05;
Occupation10 has p-value 0.2272>0.05;
Occupaton12 has p-value 0.9674>0.05;
Relationship1 has p-value 0.458>0.05, TOL 0.02231<0.1, and VIF 44.82751>10;
Relationship2 has p-value 0.5593>0.05;
Relationship3 has p-value 0.2492>0.05, TOL 0.08708<0.1, and VIF 11.4834>10;
Relationship4 has p-value 0.8013>0.05, TOL 0.06163<0.1, and VIF 16.22672>10;
Race2 has p-value 0.606>0.05;
Race3 has p-value 0.6947>0.05;
Race4 has p-value 0.8602>0.05;
Sex1 has TOL 0.07046<0.1, and VIF 14.19254>10;
Sex_income has p-value 0.5614>0.05;
Sex_age has TOL 0.06899<0.1, and VIF 14.49569>10.

The model approach is using the regression model to find and drop variables whose p-value is less or equal to 0.05, whose tolerance value is bigger than 0.01, and whose VIF is less than 10. When I get Income_Train_new data, some variables became invalid, and I had to drop them:

Workclass4 has p-value 0.1005>0.05;

Education2 has p-value 0.3065>0.05;

Education4 has p-value 0.3725>0.05;

Education7 has p-value 0.2738>0.05.

Then drop all 238 outliers and some influential points to get the final model. The final model is:

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.

F-value was increased from 10.09 to 41.82. Root MSE was decreased from 10.52 to 6.89. The adjusted-R square is 0.1624. This took great time to finish since I needed to drop each outlier each time to see whether the adjusted-R square had improved.

I tested the performance of the test set using the regression model. There are 2344 observations for train set, and 781 observations for test set. Although it had a new final model basing on validation method, it has lower adjusted-R square (0.119) compared to the previous one (0.1624). CV R^2 is 0.036 after calculating the difference between $yhat^2$ (0.155) and R^2 Train (0.119). The model is good although there is a low adjusted-R square. The mean of yhat is 44.39; standard deviation is 3.94; minimum is 31.55; maximum is 55.23; RMSE is 10.2792, which is less than the one (6.89) of the final model. MAE is 7.24677, and MAPE is 0.21709. Moreover, sex has the biggest effect on the dependent variable, and people who has sales job showed the least importance to the working hour per week.

Then I did 5-folder Cross validation using two selection methods: stepwise and backward selection. To compare ASE (Train) and ASE (Test), backward method has close ASE (Train) and ASE (Test), and the same is to stepwise method. Although models seem better than what I did since they have higher adjusted-R square (0.1832 for stepwise selection and 0.1856 for backward selection), they have lower F-values (35.77 for stepwise selection and 17.72 for backward selection), bigger RMSE (10.19 for stepwise selection and 10.37 for backward selection), and greater size of variables (15 variables for stepwise method and 32 variables for backward method). Hence, they are not better than my final model.

Finally, I used two random examples to predict the number of working hours per week. The first example is a person, who is never married, who is 50, and whose occupation is farming-fishing tends to work 42.38 hours per week. The second example is a person, who is 30, whose occupation is sales, and who is a wife tends to work 38.13 hours per week. Prediction intervals for two people are (28.62, 56.15) and (24.53,51.73), respectively.

Learning SAS by Example – A Programmers Guide (2007) argues there is a new way to select random sample.

```
proc surveyselect data=learn.blood
out=subset
method=srs
sampsize=100;
Run;
```

SRS is called simple random variable. Sampsize allows me to choose the size of the sample.

There are 5 more models from the reference I chose. It limited the p-value in the code:

First three SAS code used forward, backward, and stepwise method. The fourth SAS code calculates the RMSE for each possible subset model, sorts the models from smallest to largest RMSE and then prints the best 10 models. Specifying adjrsq in the option selection=adjrsq is not crucial since the goal is to minimize RMSE. Other choices for the selection option are rsquare or CP. The model diagnostics are output into the data sets est4 and est5.

Some terms in the code:
Outset: output name of the dataset
Slstay: minimum p-value that a variable must have
Slentry: maximum p-value that a variable must have
Noint: fits a model without the intercept term

After running five regressions, I fond the last one is the best since it had the highest Adjrsq, lowest RMSE, relatively low F-value (although it is higher than my final model), and had 35 variables. The final model of the reference I chose is:

hours_per_week= 24.72 - 0.16 age +0.00003 capital_gain + 1.70 workclass3 - 1.80 workclass4 +26.63 education2 + 20.48 education3 + 27.55 education4 + 23.67 education5 + 18.24 education6 + 22.17 education7 22.70 education8 + 28.18 education9 + 22.54 education10 + 23.05 education11 + 22.08 education12 + 23.61 education13 + 24.54 education14 + 27.65 education15 - 2.43 marital_status3 - 4.31 marital_status5 + 2.45 occupation2 + 5.13 occupation3 + 8.72 occupation4 - 2.44 occupation7 + 1.83 occupation9 + 1.97 occupation10 + 4.08 occupation11 + 5.61 occupation13 + 1.40 relationship1 - 3.22 relationship2 - 5.30 relationship3 - 5.38 relationship5 - 6.74 sex1 + 2.01 income1 +0.10 sex_age

Moreover, for the selection of random observation, There is a new way to select random sample from Learning SAS by Example – A Programmer's Guide (2007):

```
proc surveyselect data=learn.blood
out=subset
method=srs
sampsize=100;
run;
```

SRS is called simple random variable. Sampsize allows me to choose the size of the sample.

Persid

Observations

The majority of the people surveyed are white and from USA (Figure 1 in Appendix D). The model will not make good predictions regarding the race and the native country. Married people, and especially married males tend to earn more income. Also, males earn more than females. Educated people and Professionals have greater chances to earn above $50K.

Model Estimation and Interpretation

A full model (with all independent variables) is fitted to predict income. No multicollinearity is detected among variables (|corr>Value| < 0.9) (Table 2 in Appendix D). Several outliers (Pearson Residual > 3) and influential points (|DfBeta > 2/sqrt(2000) = 0.0447) are in the model (Table 3 in Appendix D. However, these data points are not removed or modified as no anomalies are noticed. More data would explain better the context (e.g., the person in observation 286 may be the owner of a cleaning business, even though he is included in the working-class category, or he may have other investments outside his everyday job).

Final Model

Stepwise selection procedure (Table 4 in Appendix D) and backward selection procedure (Table 5 in Appendix D) are run and the results are compared. Backward selection procedure has better parameters, and more predictors which are useful to create a full picture of the contributing factors on the income.

The fitted regression equation:

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

For this model all the independent variable have p-values less than 0.05 (implying that we would reject H0: □1=□2= ... =□14=0 for α = 0.05). The chi-square (test statistic for the overall adequacy of the logistic model) is 556.7638 with observed significance level p < .0001. Based on the p-value of the test, we can reject H0 and conclude that at least one of the □ coefficients is nonzero. The the model is adequate for predicting the income.

From the equation it can be concluded that , Males (sex1), Employees in Federal Government (workclass3), Professionals (occupation3), Executives (occupation5), Wifes (relationship1), longer hours per week (hrs_per_wek), capital gain, capital loss have positive association with p and odds to earn an average income more than $50K get increased.

High-school graduates (edLevel1), Not-married people (maritalStatus1), working class people (occupation1), people working in services (occupation2), people from Latin America have negative association with p and odds to earn an average income more than $50K get decreased.

The coefficient beta for a certain variable X represents the change in log(odds) for any 1-unit increase in X. Exp(beta) is the estimated odds ratio for a unit increase of X, and is equal to the odds at X+1 divided by the odds at X: odds_ratio=odds(x+1)/odds(x).

Thus, exp(beta) represents the change (increase or decrease) in odds for any 1-unit increase in X. If exp(beta)>1 (OR beta >0) the odds increase, if exp(beta)<1 (OR beta < 0), the odds decrease . E.g., the average odds earning an income above $50K are 163% higher in males than females. E.g., the average odds earning an income above $50K are 53% lower for a person working in a working-class job than a person working in an administrative-clerical job.

The predicted probability for a white male, 40 years old, a professional working in private sector, not-married, born in Europe, that works 40 hrs per week is p = 0.171 with a 95% prediction interval between 0.107 and 0.26.

The predicted probability for a white female, 55 years old, a professional working in federal government, married, born in USA, that works 40 hours per week is p = 0.812 with a prediction interval between 0.588 and 0.929.

Performance metrics are computed based on the best cutoff value found to be 0.35 (Table 6 in Appendix D), and are as follows:

Sensitivity =	0.80
Specificity =	0.86
Accuracy =	0.84
Precision =	0.65
F-metric =	0.72

The performance measures indicate that the model is good, but the model needs improvement.

Carl

I started out trying to get a feel for the way the data is shaped and looks. I ran frequency to variables (age, workclass, education, marital_status, occupation, relationship, race, sex, hours,

income) with names to visualise what the demographic of the population is at (Appendix F. Figure 2). I then ran a histogram and a matrix scatterplot. These were not useful since the dependent variable is binary. Scatterplots revealed patterns, but this was expected since I re-coded the categorical variables. The boxplots, like the frequencies, revealed insight into the data (Appendix F. Figure 3) by comparing dIncome to age, dIncome toeducation_num, and dIncome to hours_per_week. I then tested the correlation, which showed no high correlation between any of the variables (Appendix F. Figure 4). I then performed a regression analysis even though my goal is to use binary logistic regression. My initial regression had all the variables, and I integrated two interaction terms (age*education_num dSex*dJob). SAS rejected them, so I ran another regression without them and tested multicollinearity. There was no issues with multicollinearity, and I think it helped that in my data cleaning stage I decided to only use education_num instead of also education, which I would have had to transform to dummy variables. I removed dWork because it had the highest p-value for testing purposes. Now I got to the Logistic regression model, where I also checked for outliers and influential points. I ended up removing 65 outliers based on the DFBeats. I was surprised how much data cleaning was required in this data set. I know that it should take up most of my time, and it did.

For model selection I chose to compare Stepwise and Backward, since they should get different results. However, both eliminated the same independent variables (dSex, dWork, and hours_per_week) (Appendix F. Figure 5). I then split the data into training and testing (60, 40). I fit the model based on Stepwise selection and where above $50,000 a year someone makes is the response. Running predictions showed significance to higher education and age as I predicted. The model passed goodness of fit and test performance, but was low R2 and difficult to fit into useful predictions. The accuracy was good, but the precision was low, and the most significant predictor I observed was education (Appendix F. Figure 6).

Alice

In the data exploratory stage, I found that 75% of the observations made equal or less than 50k (Figure 1). The descriptive statistics showed that the median number of hours a person works (hours_per_week) in the dataset is 40 hours (Figure 2). The minimum and maximum for hours_per_week is 1 hour and 99 hours, respectively. This suggests that there will be outliers since the range is this large. The median age is 37 years old. The median number of years of education is 10 years. The box plot for d_income and age, shows that the median (44 years old) and mean (44 years old) for age are higher for people who make more than 50k that people who make less than 50k (Figure 3). The median and mode for people who make equal or less than 50k is 38 years old and 39 years old. This suggest that age will have a positive correlation with how much a person makes. The box plot for d_income and hours_per_week also seems to suggest a positive correlation between how many hours a week a person works and the amount of money they make (Figure 4). The mean for people who make over 50k is slightly higher than people who make less. The capital_gains and d_income box plot shows that the capital gains a person making over 50k has is severely skewed to the right (Figure 5). The mean for people

making over 50k is higher which suggests that people who make over 50k have more capital gains. However, the box plot for capital_loss and d_income also shows that the mean for the amount of capital loss the people making over 50k experience is slightly more than the capital loss people making <=50 have which suggests that the people who make >50 also have more capital loss (Figure 6). The counts show that there are much more male observations than females (Figure 12). The counts for race shows that there are much more white people in the dataset than people of other races (Figure 11). The counts for race also show that white people have the highest ratio (0.36) of all the white people who make >50k and white people who make <=50k (Figure 10). Asian-Pac-Islander have the next highest ratio (0.27). The counts also show that most of the data has around nine years of education (Figure 7). Also, most of the data shows the people are working in private sectors than in any other sector (Figure 8).

I tried two interaction terms. One of the was thrown out by the model because it was not significant. I thought that age and education_num might have some correlation because as a person gets older, they go up a grade. So I grouped those two categories to predict who makes over 50k. However, this interaction term was thrown out of the model. So, I tried another interaction term: "interaction2." This interaction term is between age and hours_per_week (how many hours a person works per week). Interaction2 was eventually found to be significant in the model selection process. Since my y-variable d_income is binary, I do not have to check for normality.

First I fit the full model and saw that the AIC and SC were 1382.100 and 1510.921, respectively (Figure 13). The AIC and SC should be minimized as much as possible. One way of doing this is to get rid of the insignificant predictors in the model which I do later on in the model selection process. The r-square is 0.3624 which should be maximized as much as possible. This means that 36.24% of the variation in the dataset is explained by the full model. The likelihood ratio test or goodness-of-fit test tells me to reject the null hypothesis that none of the predictors have a significant effect on d_income. This model is better than the null model with no covariates. The likelihood ratio (899.9606) is high, and the p-value associated with the likelihood ratio is almost 0 (<.0001). The likelihood ratio should be maximized as much as possible and the p-value associated with it should be as close to zero as much as possible.

After I checked the first model's performance indicators, I checked the correlation matrix for multicollinearity (Figure 14). Any two variables' correlation values above 0.9 should be flagged as having multicollinearity and should be studied further to see if action should be taken. Hours_per_week and age have a correlation value of 0.9168. However, action should not be taken because I made a interaction variable called interaction2 that is hours_per_week*age. This multicollinearity will not have adverse effects on the model; the p-value will not be affected. D_edu_secondary and d_edu_college have a correlation value of 0.9383, but no action should be taken because these are dummy variables that represent a categorical variable with three or more categories. This is the same deal with d_race_white and d_race_asian (0.9508). In the end, no action was taken over multicollinearity because of these reasons.

After the multicollinearity was reviewed, I checked for outliers and influential points. I checked the Pearson Residual graph and Deviance Residual graph for values that greater than three or less than three (Figure 15). There was one significant outlier (#1914) (Figure 31). The hours_per_week was 25 whereas the 25th percentile of hours_per_week is 40. The Pearson Residual graph indicated that there were outliers, but the Deviance Residual graph did not. Most

of the outliers fell towards the end of the dataset. I did calculate the Dfbeta value to check for influential points (2/sqrt(2000) = 0.0447), but did not delete any influential points because all of the data for hours_per_week, age, education_num, capital_gains, capital_loss, and interaction2 are relevant to predicting what type of person makes over 50k (Figure 16-18). Also, the DfBeta graphs for the dummy variables can be ignored. I deleted the outliers.

Then, I fit my model again (Figure 19). AIC (1228.393) and SC (1356.947) decreased. R-square increased to 0.3942. The p-value associated with the likelihood ratio is still <.0001; the likelihood ratio (990.8429) increased. After I split the dataset into training and testing (75%, 25%), I executed the model selection process. I used all of the model selections that I could use for logistic regression: stepwise, backward, and forward. All three models were the same. They had the same variables, standard errors, r-square (0.3988), AIC (925.482), SC (967.896), likelihood ratio (754.5484), p-value associated with the likelihood ratio (<.0001) (Figure 20-22). Then, the full model (M1) was fitted again and outliers were handled (Figure 23-25). The final model's r-square (0.415) is better than the original model's. The AIC (865.474) and SC (906.829) has also lowered quite a bit. The likelihood ratio (789.2297) has decreased compared to the original model's likelihood ratio, but the p-value associated with the likelihood ratio (<.0001) is still almost 0.

Then, I produced the classification matrix to calculate the sensitivity/recall, accuracy, precision, specificity, and f-metric values to compare M1 (the model discussed above) and M2 (Figure 29-30). M2 is another model that I produced by changing a few things. Several independent variables that were a direct copy of each other. For example, education and num_of_edu are a direct copy of each other. So, I deleted education from the model from the beginning. I also used different groupings for dummy variables, seed number, and different handling on outliers and influential points (deleting influential points) to try to improve M1. For, M2 I used several references. For M1, the sensitivity/recall (88%) accuracy (78%), precision (49%), specificity (75%), and f-metric (63%) values in general were better/larger than M2's the sensitivity/recall (89%) accuracy (73%), precision (42%), specificity (69%), and f-metric (57%) values with the exception of sensitivity/recall. Therefore, M1 is the best model. The sensitivity/recall (88%) means that when the person actually makes >50k, M1 predicts that a person will make >50k correctly 88% of the time. The specificity (75%) means that when the person actually makes <=50k, M1 predicts that a person will make <=50k correctly 75% of the time. The accuracy (78%) means that overall in classifying whether a person will make >50k or not, M1 will predict the right income for the person 78%. The precision (49%) means that when M1 classifies that the person will make >50k, it will be correct 49%. The closer these five values are to 100%, the better the model is. The precision is a little low.

Calculations:
M1
Sensitivity/Recall = 91/(91+12) = 88%
Accuracy = (91+295)/(91+295+96+12)= 78%
Precision = 91/(91+96) = 49%
Specificity = 295/(295+96)= 75%
F-metric = 2(0.49*0.88)/(0.49+0.88)= 63%
M2

Sensitivity/Recall = 71/(71+9) = 89%
Accuracy = (71+222)/(71+222+98+9) = 73%
Precision = 71/(71+98) = 42%
Specificity = 222/(222+98) = 69%
F-metric = 2(0.42*0.89)/(0.42+0.89) = 57%

I calculated two predictions. For the first one, I predicted the probability of a person making >50k if they have these certain attributes: number of education (10), amount of capital gains ($3000), amount of capital loss ($1000), married, white, lives in the U.S, 36 years old, and works 40 hours per week (Figure 27). The probability of a person making >50k with these attributes is 0.71523. 95% of the time, the predicted probability will fall within 0.59789 and 0.80926. For the second one, I predicted the probability of a person making >50k if they have these certain attributes: number of education (9), amount of capital gains ($2000), amount of capital loss ($0), not married, other race, does not live in the U.S, 26 years old, and works 38 hours per week (Figure 28). The probability of a person making >50k with these attributes is 0.00326. 95% of the time, the predicted probability will fall within 0.00120 and 0.00881.

Model Equation (Figure 26):

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

The odds of a person making >50k increases by 56.9% ((exp(0.4507) -1) * 100) when number of years of education he or she has increases by one.
The odds of a person making >50k increases by 0.04% ((exp(0.000442) -1) * 100) when the amount of capital gain he/she has increases by one.
The odds of a person making >50k increases by 0.07% ((exp(0.000671) -1) * 100) when the amount of capital loss he/she has increases by one.
The odds of a person making >50k increases by 1788.4% ((exp(2.9383) -1) * 100) if he/she is married and not single (base case).
The odds of a person making >50k increases by 66.5% ((exp(0.5097) -1) * 100) if he/she is white and not other (base case).
The odds of a person making >50k increases by 229.1% ((exp(1.1912) -1) * 100) if he/she is from the U.S and not from any of the other countries (base case).
The odds of a person making >50k increases by 0.07% ((exp(0.000977) -1) * 100) when his/her age increases by one and the amount of hours he/she works per week is held constant, and vise versa.

The most influential predictor of Model 1 is capital_gains because it has the highest standardized estimate (2.4421). d_marital is the next strongest predictor with a standardized estimate of 0.8100.

The null hypothesis can be rejected because the p-value associated with the likelihood ratio (<.0001) is very close to 0.
H1 is correct. The more education a person has, the greater their chances are of making >50k
H2 is incorrect. Single people do not have a greater chance of making >50k than their married counterparts.

Future Work

Amy

In conclusion, the final model needs a great deal of performance improvement. The model suggests that education and occupation are the most significant contributors to determining how many hours a person works per week. Performance may be improved by grouping the education and occupation variables differently. For example, jobs in the technical field could all be counted the same, or even broken down by job-sector class like “white collar” and “blue collar”. I also think further research into income trends throughout time would help us create a better model.

An entire other area of research can could influence the hours worked per week is geographical area. This model did not include any native country variables as a predictor, so where a person comes from must not be as important as other life factors. One of these factors could be where they live, ie. suburb vs city. In 1997, the median family income was higher in the suburbs in the city (Levy, 1998) leaving questions of how hard or how much per week one has to work to earn a living wage in these different geographical regions.

Jenny

In conclusion, the final logistic model seems to have a decent performance on the test set, but it still has outliers and influential points, high AIC and SC values, and a low r-square value. Its performance could be improved upon in the future. There seem to be more significant independent variables missing from the analysis since the model has low explanatory power, even though all of the variables it contains are significant. Two of the significant variables, capital gain and capital loss, also seem to have a very small effect on the dependent variable despite the former's high standardized estimate. More domain research should be conducted to explore other viable independent variables.

Modeling techniques not covered in this class could be more suitable for the dataset, as indicated by the results of a paper which compared logistic regression to the performance of a naïve Bayes, and a decision tree model. The paper claims that the decision tree model had the highest accuracy, with a test error of 14.778% (Lemon 2015). Considering that the logistic regression model was the best fit in this class for a binary dependent variable and several logistic models were analyzed and considered, the results of this paper are a good starting point for future exploration of more advanced models.

Danyang

My final model is not great since the adjusted-R square is very low (0.1624 of my model and 0.1819 of the hybrid model). The future work is to study other resources to know which aspects have greater influence on the working hours. Hours per hour may relate to personality, confidence level, power of execution, market fluctuation and even company's policy. Hence, Hours per hour may be simultaneously influenced by national and personal aspects. Although they are assumed by my personal experience, I need to do more research on this topic.

Persid

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.

Alice

The final model's r-square (0.415) is a little low and could be improved. The AIC (865.474) and SC (906.829) is still a bit high. The p-value associated with the likelihood ratio (<.0001) looks good though. This model could be improved if the data was not so skewed. Gathering more data on females, other races other than white people, more people who have different educational levels other than around 9 years of education, and people from other countries besides the U.S. Also, gathering more data about people who do not work in the private sector would benefit the model as well. So, to improve the model, more data should be gathered.

Carl

There is a lot of future work needed. Since the data has a high frequency in white males in America, demographics could be explored more. Even though I did a lot of data cleaning, more will probably increase diagnostics. Also breaking down the categorical variables to sections that make more sense would benefit the model. So basically more data from a wider range of income situations would help in the future.

References

UCI Machine Learning. 2016. *Adult Census Income*. Available at: https://www.kaggle.com/uciml/adult-census-income. Accessed 1 May 2018.

Amy

Classora. 2016. *Ranking of the World's Richest Countries by GDP (1994).* Available at: http://en.classora.com/reports/t24369/ranking-of-the-worldsrichest-countries-by-gdp?edition=1994. Accessed 23 May 2018.

Hartmann, Heidi. 2005. *The gender wage gap is real.* Available at: https://www.epi.org/publication/webfeatures_snapshots_20050914/. Accessed 30 May 2018.

Levy, Frank, and Nicholas Lemann. 1998. *New Dollars and Dreams, The: American Incomes in the Late 1990s.* New York, NY: Russell Sage Foundation.

Roubinchtein, Alex; Nimmo, Bruce. 2017. *2017 Employment Projections Technical Report.* Washington. Available at: https://fortress.wa.gov/esd/employmentdata/docs/industry-reports/2017-employment-projections-technical-report.pdf. Accessed 3 June 2018.

Rones, Phillip L, et al. "Trends in Hours of Work Since the Mid-1970s." *Monthly Labor Review*, 1997, pp. 1–3., https://stats.bls.gov/opub/mlr. Accessed 1 June 2018.

Jenny

Borjas, George. 1987. "Self-selection and the earnings of immigrants". *NBER Working Paper Series.* 2248.

Devroy, Dan and Freeman, Richard. 2009. "Does Inequality in Skills Explain Inequality in Earnings Across Advanced Countries?" *NBER Working Paper Series*. 8140.

Martins, Pedro, and Pereira, Pedro. 2003. "Does education reduce wage inequality? Quantile regression evidence from 16 countries." *Labour Economics*. 11: 355-371.

SAS Institute. 2010. Simple Random Sampling. *SAS/STAT(R) 9.22 User's Guide.* https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_surveyselect_sect003.htm. Accessed June 1, 2018.

Lemon, Chet, Zezalo, Chris, and Mulakalu, Kesav. 2015. "Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques." *University of San Diego Computer Science and Engineering*. https://cseweb.ucsd.edu/~jmcauley/cse190/reports/sp15/048.pdf. Accessed June 1, 2018.

Danyang

Arthur L. Stinchcombe. 1983. *Economic Sociology*: 173.

Ron Cody. 2007. *Learning SAS® by Example: A Programmer's Guide*:200 Cary, NC:SAS Institute Inc.

Dennis J. Beal, Science Applications International Corporation, Oak Ridge, TN. *SAS Code to Select the Best Multiple Linear Regression Model for Multivariate Data Using Information Criteria*

Persid

Hartmann, Heidi. 2005. *The gender wage gap is real.* Available at: https://www.epi.org/publication/webfeatures_snapshots_20050914/.

Alice

Carmignani, Fabrizio. 29 May 2018. “Does Government Spending on Education Promote Economic Growth?” *The Conversation*, The Conversation, theconversation.com/does-government-spending-on-education-promote-economic-growth-60229. Accessed May 30, 2018.

Gillett, Rachel. 8 July 2016. “7 Ways Being Single Makes You More Successful.” *Business Insider*, Business Insider, www.businessinsider.com/how-being-single-makes-you-more-successful-2016-7#single-men-tend-to-work-fewer-hours-than-married-men-6. Accessed May 30, 2018.

Carl

Ron Cody. 2007. *Learning SAS® by Example: A Programmer's Guide*:200 Cary, NC:SAS Institute Inc.

Arthur L. Stinchcombe. 1983. *Economic Sociology*: 173.

Carmignani, Fabrizio. 29 May 2018. "Does Government Spending on Education Promote Economic Growth?" *The Conversation*, The Conversation, theconversation.com/does-government-spending-on-education-promote-economic-growth-60229. Accessed May 30, 2018.

Appendix A - Amy

Appendix A.1 - non SAS Code

Preprocessing R Studio Code:
First I loaded the data into R Studio using File > Import Dataset. The name of data is "adult". Then I began to refine with the following code:

```
> adult50K = adult[adult$workclass != "?",]
> adult50K = adult50K[adult50K$native.country != "?",]
> adult50K = adult50K[adult50K$income == ">50K",]

(2000/7508 * 100 = 26.6 %, so I'll use 27 percent to get a random sample over 2000 as per the extra credit)

> smp_size <- floor(0.27 * nrow(adult50K))
> set.seed(7853926)
> train_ind <- sample(seq_len(nrow(adult50K)), size = smp_size)
> train <- adult50K[train_ind,]
```

Then I saved the cleaned up data set named "train" in order to begin analyzing in SAS.

Appendix A.2 - Relevant Output

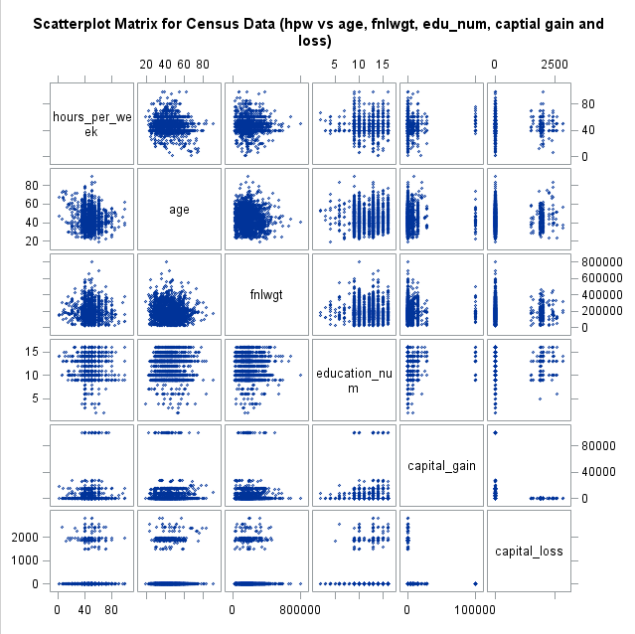


Figure 1. Scatter plot matrix

Figure 2. Boxplot, hours-per-week vs sex

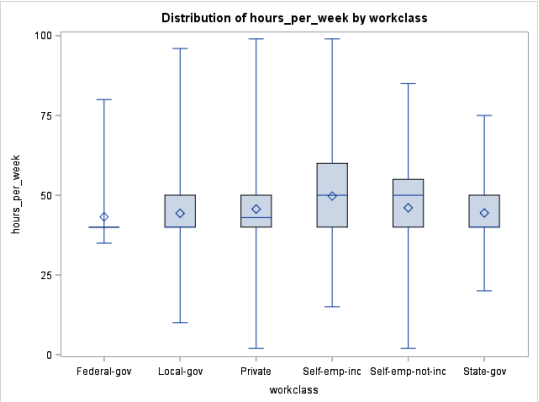


Figure 3. Boxplot, hours-per-week vs workclass

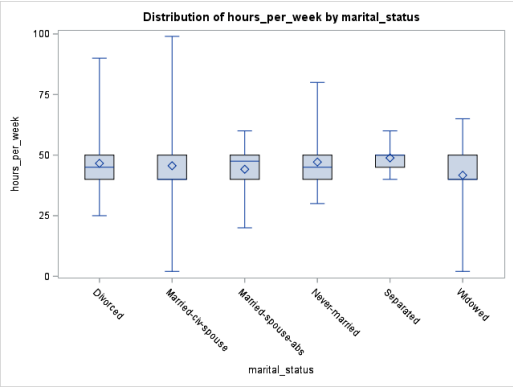


Figure 4. Boxplot, hours-per-week vs marital-status

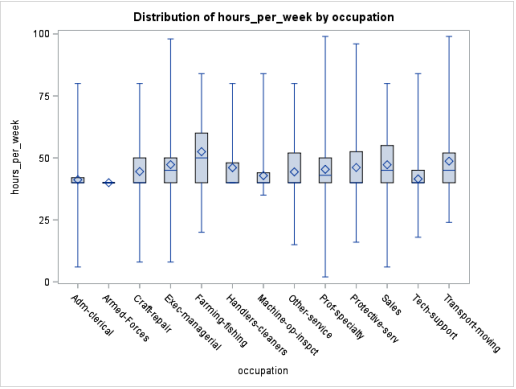


Figure 5. Boxplot, hours-per-week vs occupation

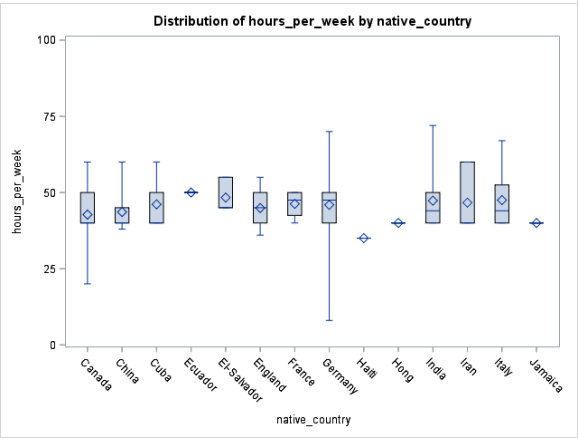


Figure 6. Boxplot, hours-per-week vs native-country

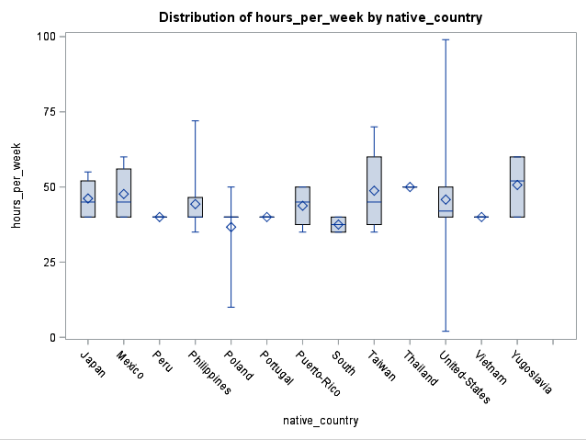


Figure 7. Boxplot, hours-per-week vs native-country pt 2

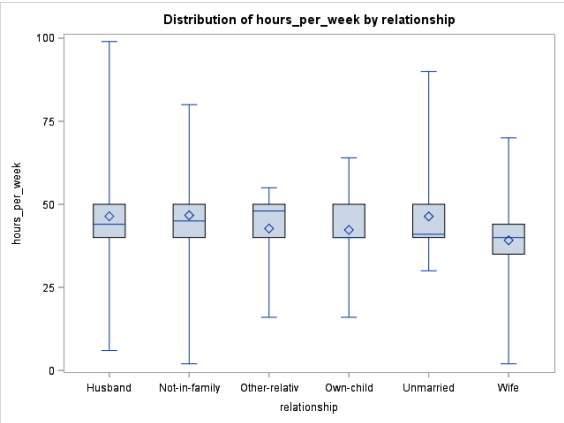


Figure 8. Boxplot, hours-per-week vs relationship

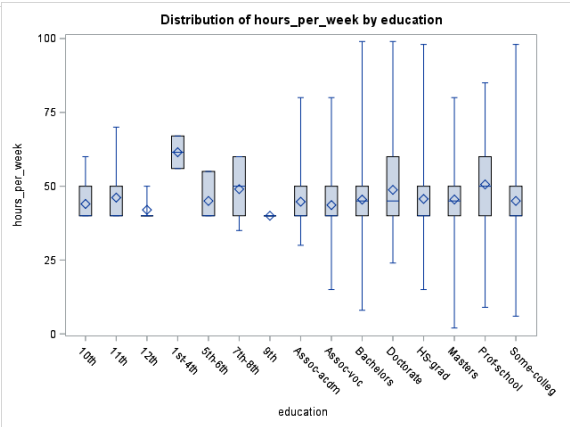


Figure 9. Boxplot, hours-per-week vs education

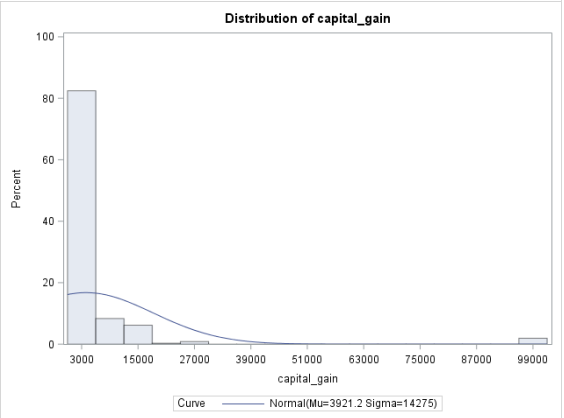


Figure 10. Histogram of skewed capital gains

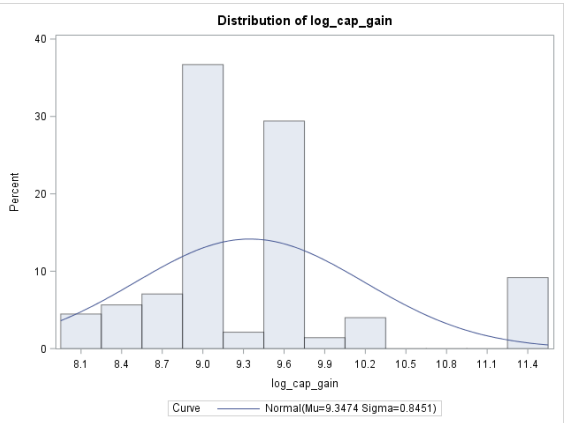


Figure 11. Histogram of log transformed capital gains

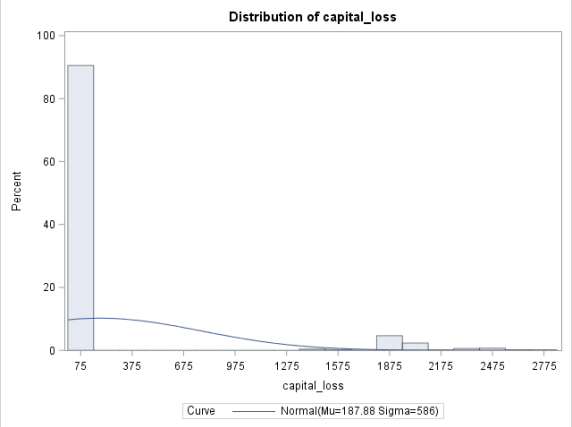


Figure 12. Histogram of skewed capital loss

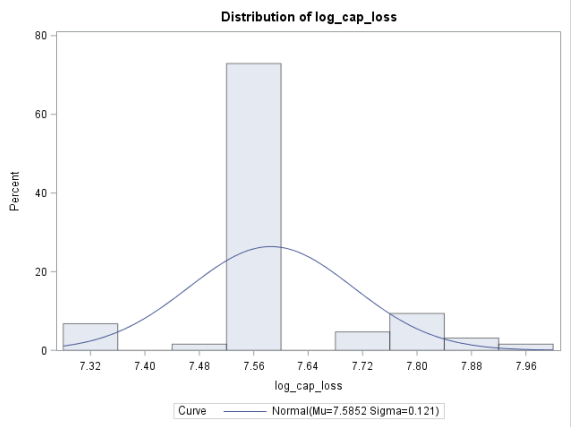


Figure 13. Histogram of log transformed capital loss

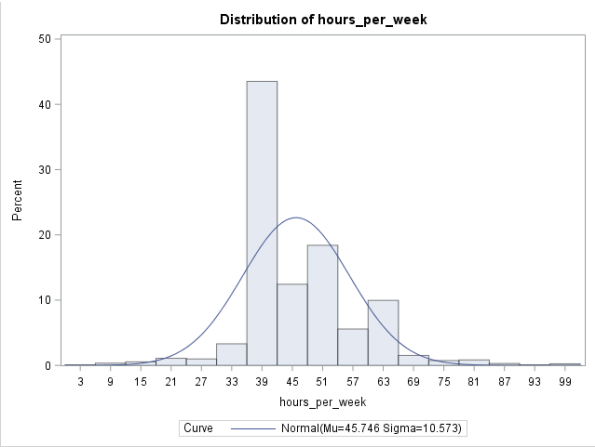


Figure 14. Histogram of dependent variable, hours per week worked

Residual Plots:

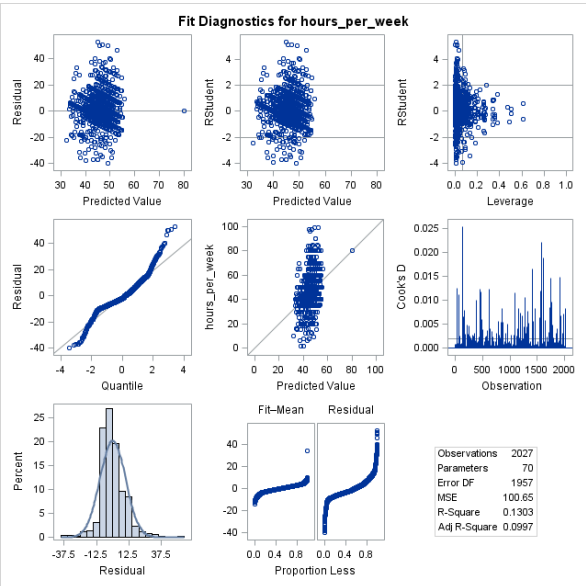


Figure 15.

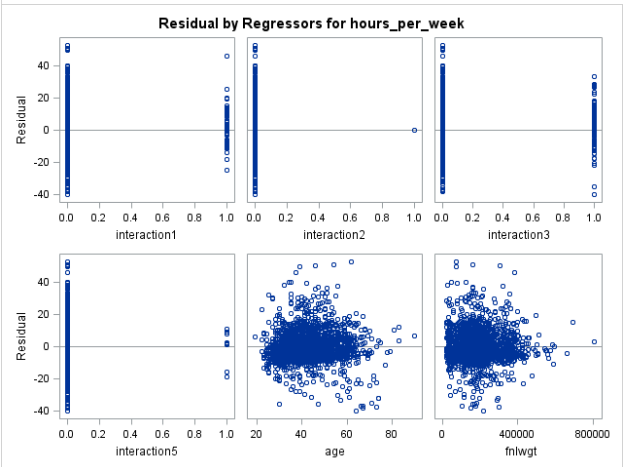


Figure 16.

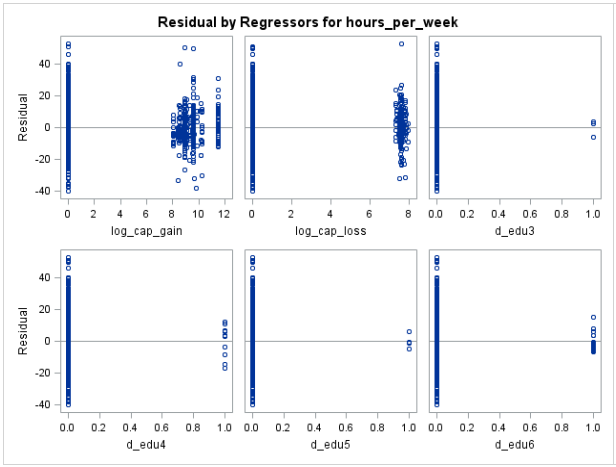


Figure 17.

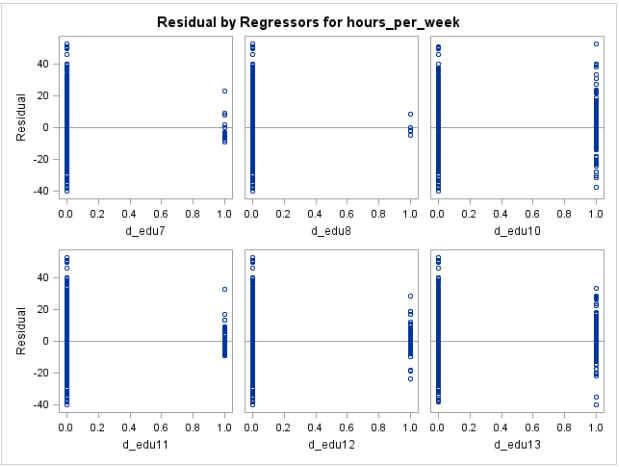


Figure 18.

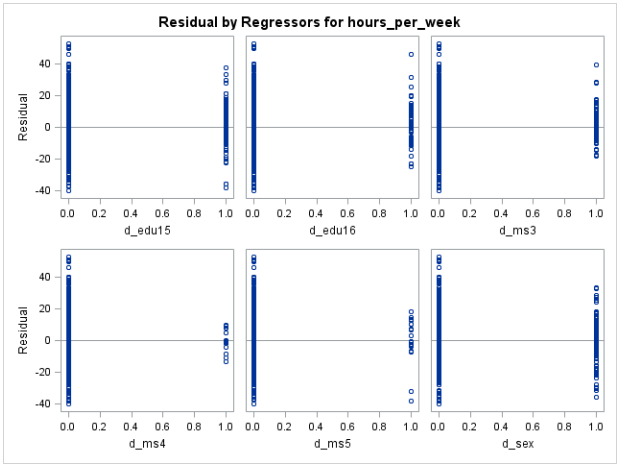


Figure 19.

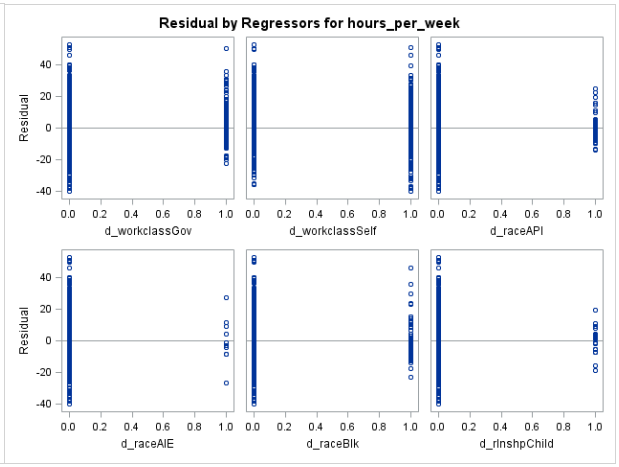


Figure 20.

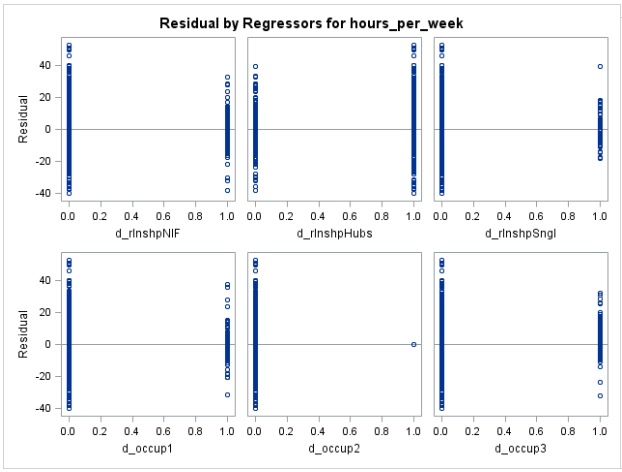


Figure 21.

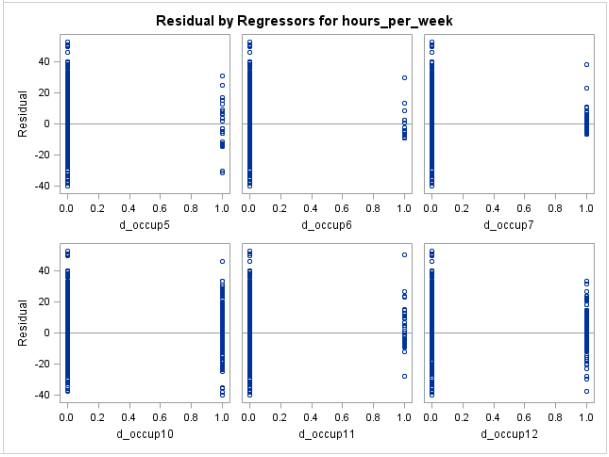


Figure 21.

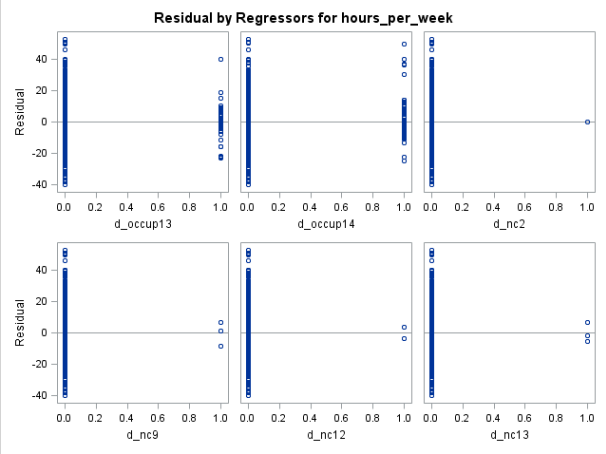


Figure 22.

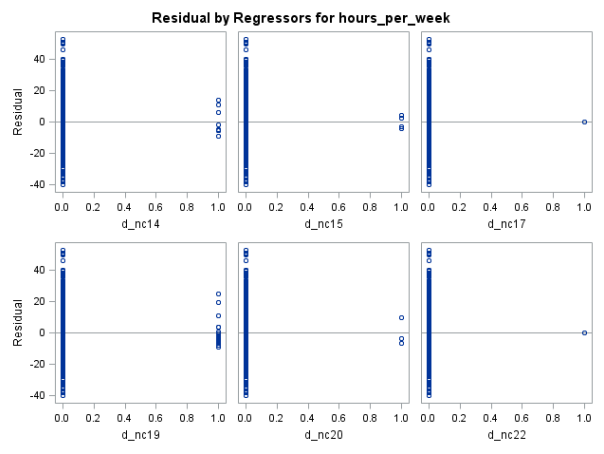


Figure 23.

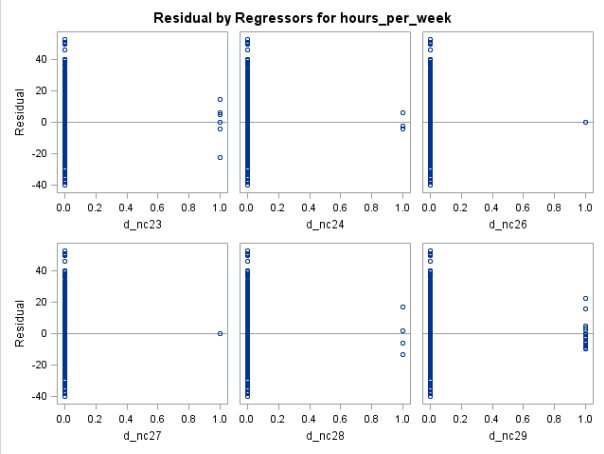


Figure 24.

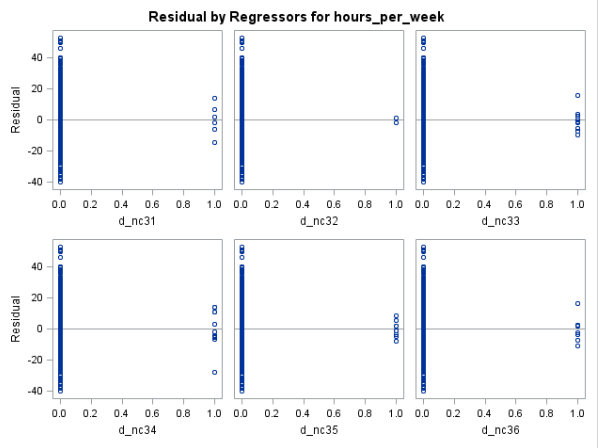


Figure 25.

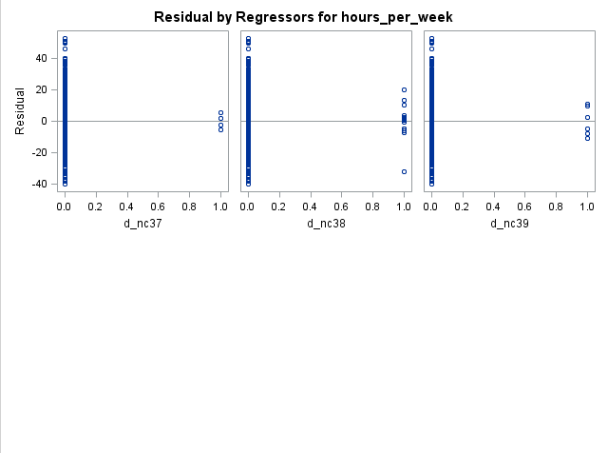


Figure 26.

Table 1. Dummy Variables			
Original Variable	Dummy Variable	Original Variable	Dummy Variable

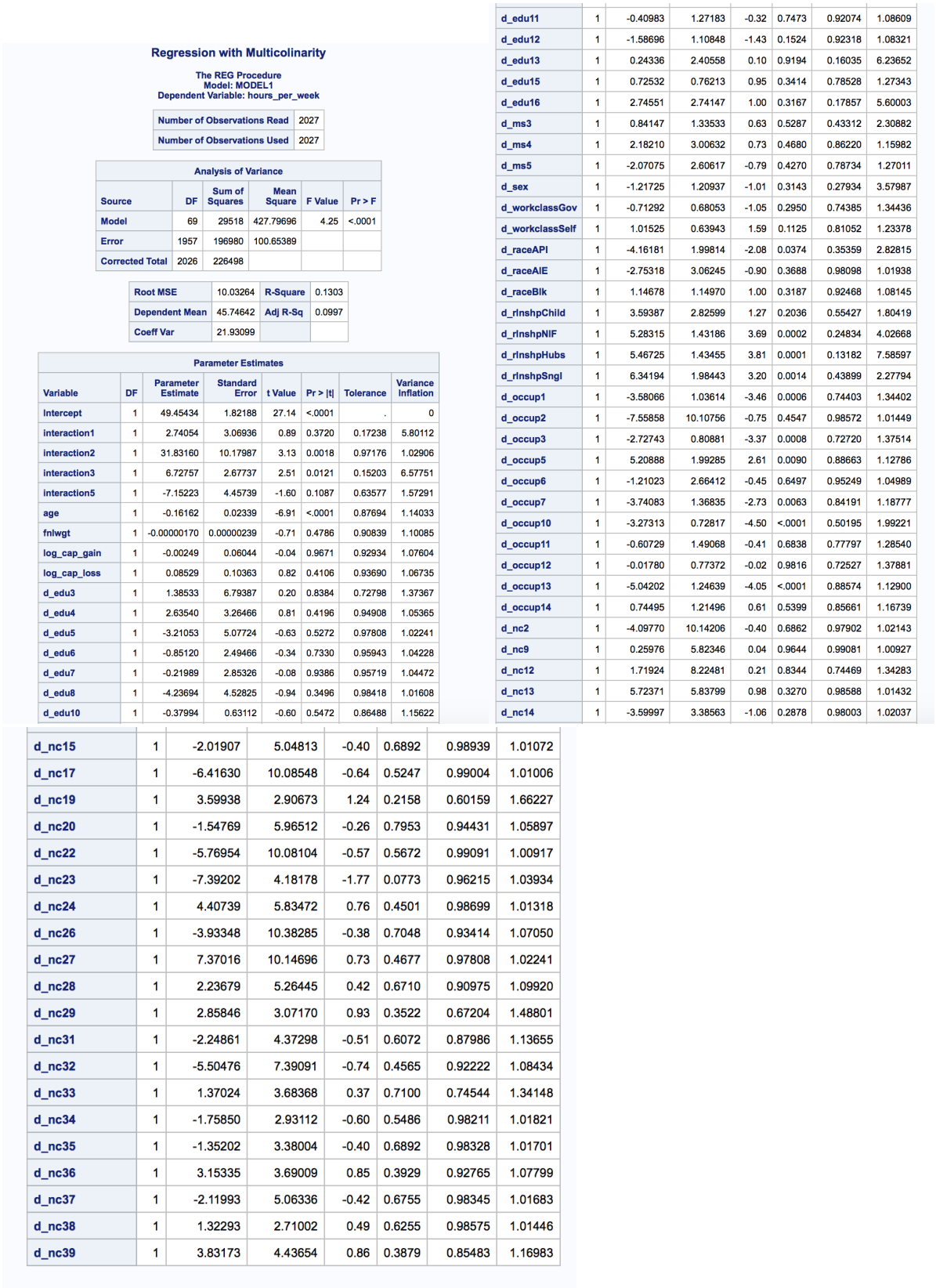
Workclass		Relationship	
Private	0 (base)	Other-relative	0 (base)
Local-gov, Federal-gov, State-gov	d_workclassGov	Wife	d_rlnshipWife
Self-emp-inc, Self-emp-not-inc	d_workclassSelf	Own-child	d_rlnshpChild
Without-pay	d_workclassNoPay	Husband	d_rlnshpHubs
Never-worked	d_workclassNeverWork	Not-in-family	d_rlnshpNIF
Race		Unmarried	d_rlnshpSngl
White	d_raceWht	Education	
Asian-Pac-Islander	d_raceAPI	Preschool	d_edu1
Amer-Indian-Eskimo	d_raceAIE	1st-4th	0 (base)
Black	d_raceBlk	5th-6th	d_edu3
Other	0 (base)	7th-8th	d_edu4
Sex		9th	d_edu5
Male	0 (base)	10th	d_edu6
Female	d_sex	11th	d_edu7
Occupation		12th	d_edu8
Adm-clerical	d_occup1	HS-grad	d_edu9
Armed-Forces	d_occup2	Some-college	d_edu10
Craft-repair	d_occup3	Assoc-acdm	d_edu11
Exec-managerial	d_occup4	Assoc-voc	d_edu12
Farming-fishing	d_occup5	Prof-school	d_edu13
Handlers-cleaners	d_occup6	Bachelors	d_edu14
Machine-op-inspect	d_occup7	Masters	d_edu15
Other-service	0 (base)	Doctorate	d_edu16
Priv-house-serv	d_occup9		

Prof-specialty	d_occup10		
Protective-serv	d_occup11		
Sales	d_occup12		
Tech-support	d_occup13		
Transport-Moving	d_occup14		

Table 2: Dummy variables for native-country by increasing GDP

country	*gdp*	*dummy variable*
Laos	*1,543,606,400*	*d_nc1*
Haiti	*2,378,749,950*	*d_nc2*
Cambodia	*2,791,000,000*	*d_nc3*
Nicaragua	*2,977,433,090*	*d_nc4*
Honduras	*3,432,356,610*	*d_nc5*
Jamaica	*4,938,132,500*	*d_nc6*
Trinadad&Tobago	*4,947,206,100*	*d_nc7*
Outlying-US(Guam-USVI-etc)	*6,193,000,400*	*d_nc8*
El-Salvador	*8,085,554,200*	*d_nc9*
Dominican-Republic	*10,927,265,800*	*d_nc10*
Guatemala	*12,983,235,600*	*d_nc11*
Vietnam	*16,286,434,300*	*d_nc12*
Ecuador	*18,581,907,500*	*d_nc13*
Cuba	*28,450,000,000*	*d_nc14*
Puerto-Rico	*39,690,600,000*	*d_nc15*
Hungary	*41,521,517,000*	*d_nc16*
Peru	*44,909,998,000*	*d_nc17*

Ireland	*55,417,197,000*	*d_nc18*
Philippines	*64,084,541,000*	*d_nc19*
Iran	*67,128,218,000*	*d_nc20*
Columbia	*81,709,449,000*	*d_nc21*
Portugal	*95,332,368,000*	*d_nc22*
Poland	*98,515,952,000*	*d_nc23*
Yugoslavia	*120,100,000,000*	*d_nc24*
Greece	*128,788,201,000*	*d_nc25*
Hong	*135,542,440,000*	*d_nc26*
Thailand	*144,526,557,000*	*d_nc27*
Taiwan	*256,438,000,000*	*d_nc28*
India	*323,506,143,000*	*d_nc29*
Holand-Netherlands	*351,190,221,000*	*d_nc30*
Mexico	*421,725,045,000*	*d_nc31*
South Korea	*423,434,453,000*	*d_nc32*
China	*559,225,900,000*	*d_nc33*
Canada	*564,494,010,000*	*d_nc34*
United Kingdom(Eng + Scot)	*1,042,206,950,000*	*d_nc35*
Italy	*1,053,924,260,000*	*d_nc36*
France	*1,364,286,370,000*	*d_nc37*
Germany	*2,146,293,780,000*	*d_nc38*
Japan	*4,760,416,600,000*	*d_nc39*
United-States	*7,017,500,000,000*	*0 (base)*



Regression with Multicolinearity

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2027
Number of Observations Used	2027

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	69	29518	427.79696	4.25	<.0001
Error	1957	196980	100.65389		
Corrected Total	2026	226498			

Root MSE	10.03264	R-Square	0.1303
Dependent Mean	45.74642	Adj R-Sq	0.0997
Coeff Var	21.93099		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	49.45434	1.82188	27.14	<.0001	.	0
interaction1	1	2.74054	3.06936	0.89	0.3720	0.17238	5.80112
interaction2	1	31.83160	10.17987	3.13	0.0018	0.97176	1.02906
interaction3	1	6.72757	2.67737	2.51	0.0121	0.15203	6.57751
interaction5	1	-7.15223	4.45739	-1.60	0.1087	0.63577	1.57291
age	1	-0.16162	0.02339	-6.91	<.0001	0.87694	1.14033
fnlwgt	1	-0.00000170	0.00000239	-0.71	0.4786	0.90839	1.10085
log_cap_gain	1	-0.00249	0.06044	-0.04	0.9671	0.92934	1.07604
log_cap_loss	1	0.08529	0.10363	0.82	0.4106	0.93690	1.06735
d_edu3	1	1.38533	6.79387	0.20	0.8384	0.72798	1.37367
d_edu4	1	2.63540	3.26466	0.81	0.4196	0.94908	1.05365
d_edu5	1	-3.21053	5.07724	-0.63	0.5272	0.97808	1.02241
d_edu6	1	-0.85120	2.49466	-0.34	0.7330	0.95943	1.04228
d_edu7	1	-0.21989	2.85326	-0.08	0.9386	0.95719	1.04472
d_edu8	1	-4.23694	4.52825	-0.94	0.3496	0.98418	1.01608
d_edu10	1	-0.37994	0.63112	-0.60	0.5472	0.86488	1.15622
d_edu11	1	-0.40983	1.27183	-0.32	0.7473	0.92074	1.08609
d_edu12	1	-1.58696	1.10848	-1.43	0.1524	0.92318	1.08321
d_edu13	1	0.24336	2.40558	0.10	0.9194	0.16035	6.23652
d_edu15	1	0.72532	0.76213	0.95	0.3414	0.78528	1.27343
d_edu16	1	2.74551	2.74147	1.00	0.3167	0.17857	5.60003
d_ms3	1	0.84147	1.33533	0.63	0.5287	0.43312	2.30882
d_ms4	1	2.18210	3.00632	0.73	0.4680	0.86220	1.15982
d_ms5	1	-2.07075	2.60617	-0.79	0.4270	0.78734	1.27011
d_sex	1	-1.21725	1.20937	-1.01	0.3143	0.27934	3.57987
d_workclassGov	1	-0.71292	0.68053	-1.05	0.2950	0.74385	1.34436
d_workclassSelf	1	1.01525	0.63943	1.59	0.1125	0.81052	1.23378
d_raceAPI	1	-4.16181	1.99814	-2.08	0.0374	0.35359	2.82815
d_raceAIE	1	-2.75318	3.06245	-0.90	0.3688	0.98098	1.01938
d_raceBlk	1	1.14678	1.14970	1.00	0.3187	0.92468	1.08145
d_rlnshpChild	1	3.59387	2.82599	1.27	0.2036	0.55427	1.80419
d_rlnshpNlF	1	5.28315	1.43186	3.69	0.0002	0.24834	4.02668
d_rlnshpHubs	1	5.46725	1.43455	3.81	0.0001	0.13182	7.58597
d_rlnshpSngl	1	6.34194	1.98443	3.20	0.0014	0.43899	2.27794
d_occup1	1	-3.58066	1.03614	-3.46	0.0006	0.74403	1.34402
d_occup2	1	-7.55858	10.10756	-0.75	0.4547	0.98572	1.01449
d_occup3	1	-2.72743	0.80881	-3.37	0.0008	0.72720	1.37514
d_occup5	1	5.20888	1.99285	2.61	0.0090	0.88663	1.12786
d_occup6	1	-1.21023	2.66412	-0.45	0.6497	0.95249	1.04989
d_occup7	1	-3.74083	1.36835	-2.73	0.0063	0.84191	1.18777
d_occup10	1	-3.27313	0.72817	-4.50	<.0001	0.50195	1.99221
d_occup11	1	-0.60729	1.49068	-0.41	0.6838	0.77797	1.28540
d_occup12	1	-0.01780	0.77372	-0.02	0.9816	0.72527	1.37881
d_occup13	1	-5.04202	1.24639	-4.05	<.0001	0.88574	1.12900
d_occup14	1	0.74495	1.21496	0.61	0.5399	0.85661	1.16739
d_nc2	1	-4.09770	10.14206	-0.40	0.6862	0.97902	1.02143
d_nc9	1	0.25976	5.82346	0.04	0.9644	0.99081	1.00927
d_nc12	1	1.71924	8.22481	0.21	0.8344	0.74469	1.34283
d_nc13	1	5.72371	5.83799	0.98	0.3270	0.98588	1.01432
d_nc14	1	-3.59997	3.38563	-1.06	0.2878	0.98003	1.02037
d_nc15	1	-2.01907	5.04813	-0.40	0.6892	0.98939	1.01072
d_nc17	1	-6.41630	10.08548	-0.64	0.5247	0.99004	1.01006
d_nc19	1	3.59938	2.90673	1.24	0.2158	0.60159	1.66227
d_nc20	1	-1.54769	5.96512	-0.26	0.7953	0.94431	1.05897
d_nc22	1	-5.76954	10.08104	-0.57	0.5672	0.99091	1.00917
d_nc23	1	-7.39202	4.18178	-1.77	0.0773	0.96215	1.03934
d_nc24	1	4.40739	5.83472	0.76	0.4501	0.98699	1.01318
d_nc26	1	-3.93348	10.38285	-0.38	0.7048	0.93414	1.07050
d_nc27	1	7.37016	10.14696	0.73	0.4677	0.97808	1.02241
d_nc28	1	2.23679	5.26445	0.42	0.6710	0.90975	1.09920
d_nc29	1	2.85846	3.07170	0.93	0.3522	0.67204	1.48801
d_nc31	1	-2.24861	4.37298	-0.51	0.6072	0.87986	1.13655
d_nc32	1	-5.50476	7.39091	-0.74	0.4565	0.92222	1.08434
d_nc33	1	1.37024	3.68368	0.37	0.7100	0.74544	1.34148
d_nc34	1	-1.75850	2.93112	-0.60	0.5486	0.98211	1.01821
d_nc35	1	-1.35202	3.38004	-0.40	0.6892	0.98328	1.01701
d_nc36	1	3.15335	3.69009	0.85	0.3929	0.92765	1.07799
d_nc37	1	-2.11993	5.06336	-0.42	0.6755	0.98345	1.01683
d_nc38	1	1.32293	2.71002	0.49	0.6255	0.98575	1.01446
d_nc39	1	3.83173	4.43654	0.86	0.3879	0.85483	1.16983

Figure 27. Regression analysis after analyzing for multicollinearity was finished

Table 3. Model comparison

	Training		
	Model 1	Model 2	Model 3
RMSE	9.95023	9.98377	9.9856
R^2	0.1253	0.1194	0.1191
adjR^2	0.1171	0.1131	0.1109
GOF	ok	ok	ok
residuals	ok	ok	ok

Table 4. Model comparison

	Testing		
	Model 1	Model 2	Model 3
RMSE	10.1650	10.1934	10.2748
MAE	7.47623	7.54007	7.62972
R^2	0.0733	0.0682	0.0581
adjR^2	0.0647	0.0595	0.0493
CV	0.052	0.0512	0.061
N	1521	1521	1521
K	14	14	14

final model - Model 1

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2027
Number of Observations Used	2027

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	25791	1842.22770	18.47	<.0001
Error	2012	200706	99.75471		
Corrected Total	2026	226498			

Root MSE	9.98773	R-Square	0.1139
Dependent Mean	45.74642	Adj R-Sq	0.1077
Coeff Var	21.83281		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	48.09033	1.17731	40.85	<.0001	0	0
interaction1	1	5.52059	1.36407	4.05	<.0001	0.09132	1.15608
interaction2	1	32.07802	9.99365	3.21	0.0013	0.06739	1.00069
interaction3	1	7.56840	1.15276	6.57	<.0001	0.15283	1.23033
age	1	-0.15404	0.02217	-6.95	<.0001	-0.14827	1.03415
d_raceAPI	1	-2.25285	1.19016	-1.89	0.0585	-0.03997	1.01241
d_rlnshpNIF	1	6.18041	0.99822	6.19	<.0001	0.18259	1.97466
d_rlnshpHubs	1	6.30116	0.78645	8.01	<.0001	0.25503	2.30048
d_rlnshpSngl	1	6.82175	1.48356	4.60	<.0001	0.10937	1.28463
d_occup1	1	-3.78532	0.95760	-3.95	<.0001	-0.08928	1.15834
d_occup3	1	-2.91193	0.72110	-4.04	<.0001	-0.08900	1.10292
d_occup5	1	5.90149	1.88877	3.12	0.0018	0.06630	1.02227
d_occup7	1	-4.09311	1.27150	-3.22	0.0013	-0.06872	1.03482
d_occup10	1	-3.48081	0.63504	-5.48	<.0001	-0.14223	1.52886
d_occup13	1	-5.46961	1.19049	-4.59	<.0001	-0.09830	1.03929

Figure 28. Final Linear Regression Model Output.

perform 2 predictions

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	49.0002	0.3983	48.2191	49.7814	29.3973	68.6032	.
2	.	38.3077	1.3823	35.5968	41.0187	18.5336	58.0818	.
3	45	45.4907	0.7764	43.9680	47.0134	25.8442	65.1371	-0.4907

Figure 29. Prediction output

Appendix B - Jenny

Figures

Simple Random Sample

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	CENSUS
Random Number Seed	12085001
Sample Size	2000
Selection Probability	0.061425
Sampling Weight	16.28
Output Data Set	CENSUSSAMPLE

Figure 1. Random sample of 2000 observations with a seed of 12085001.

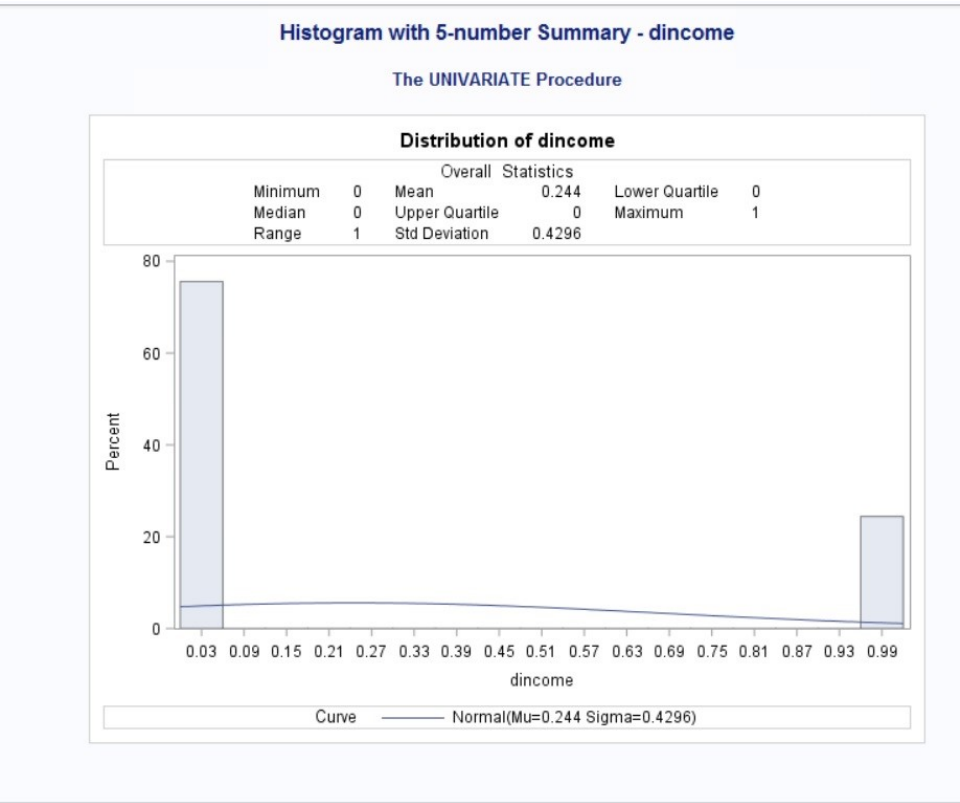


Figure 2. Histogram with 5-number summary

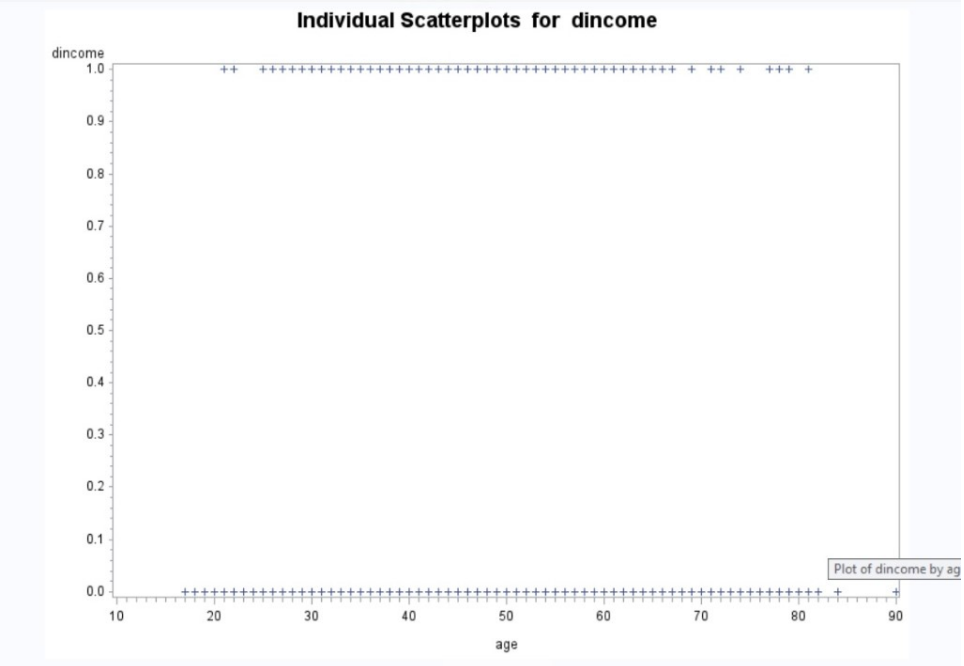


Figure 3. Scatterplot of dincome by age

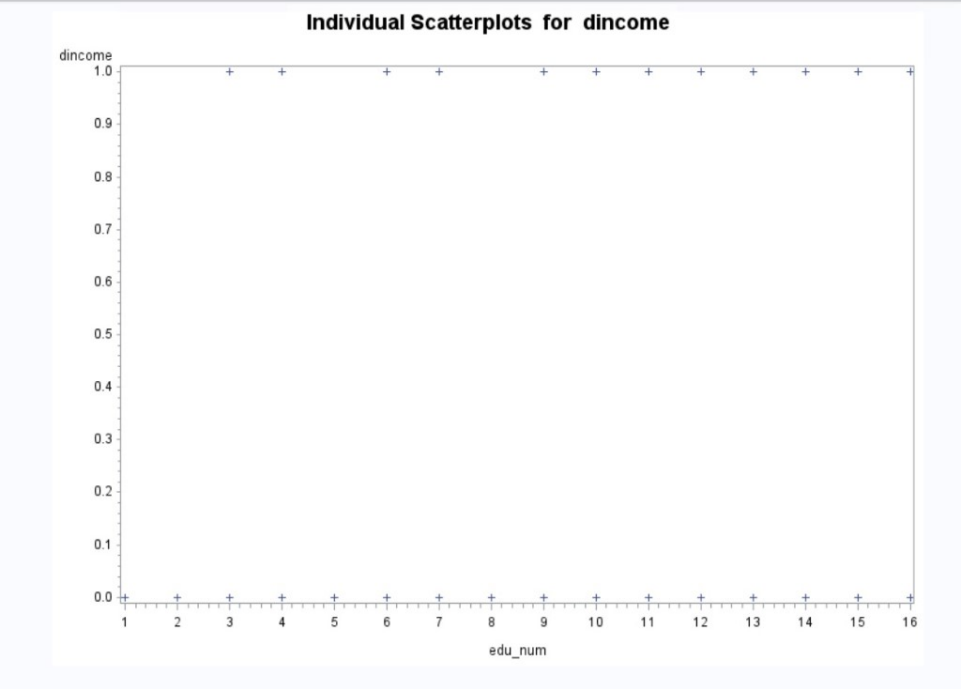


Figure 4. Scatterplot of dincome by edu_num

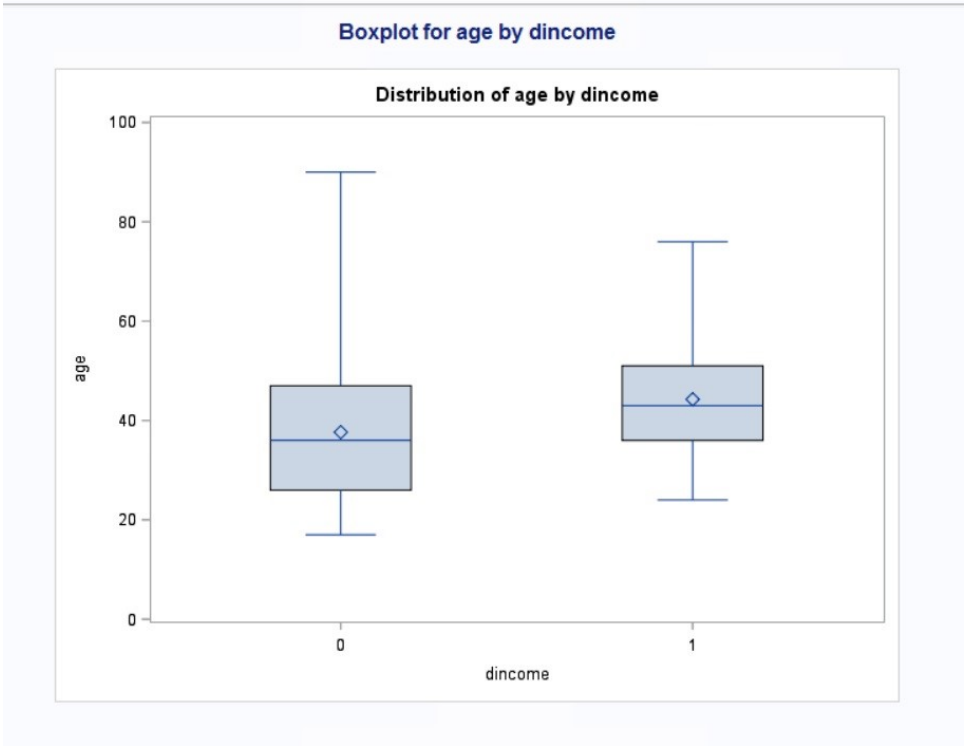


Figure 5. Boxplot of age by dincome

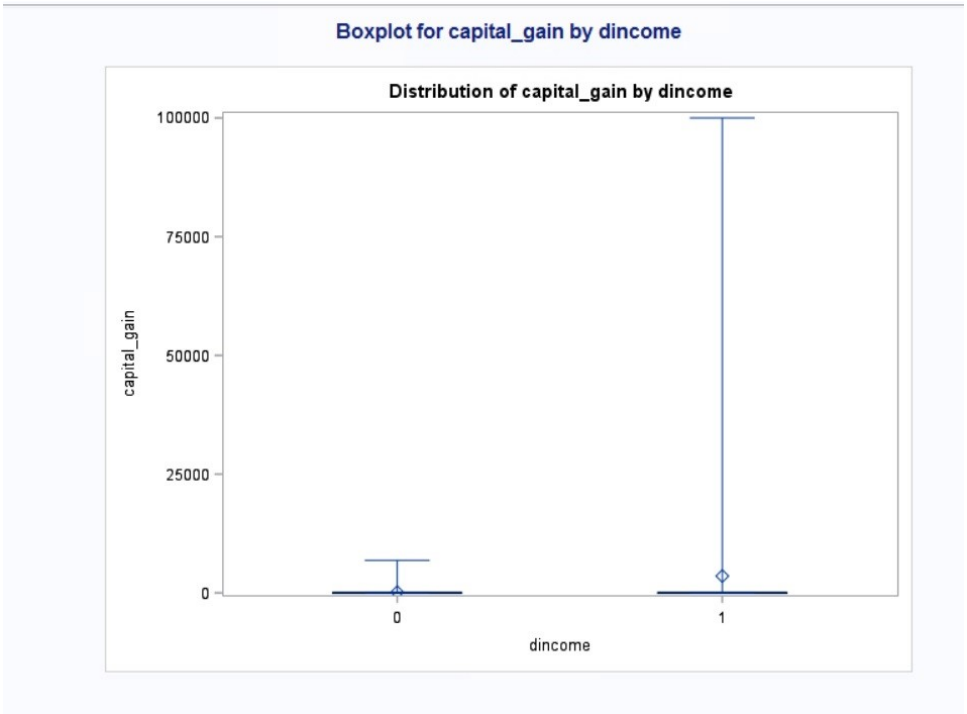


Figure 6. Boxplot of capital_gain by dincome

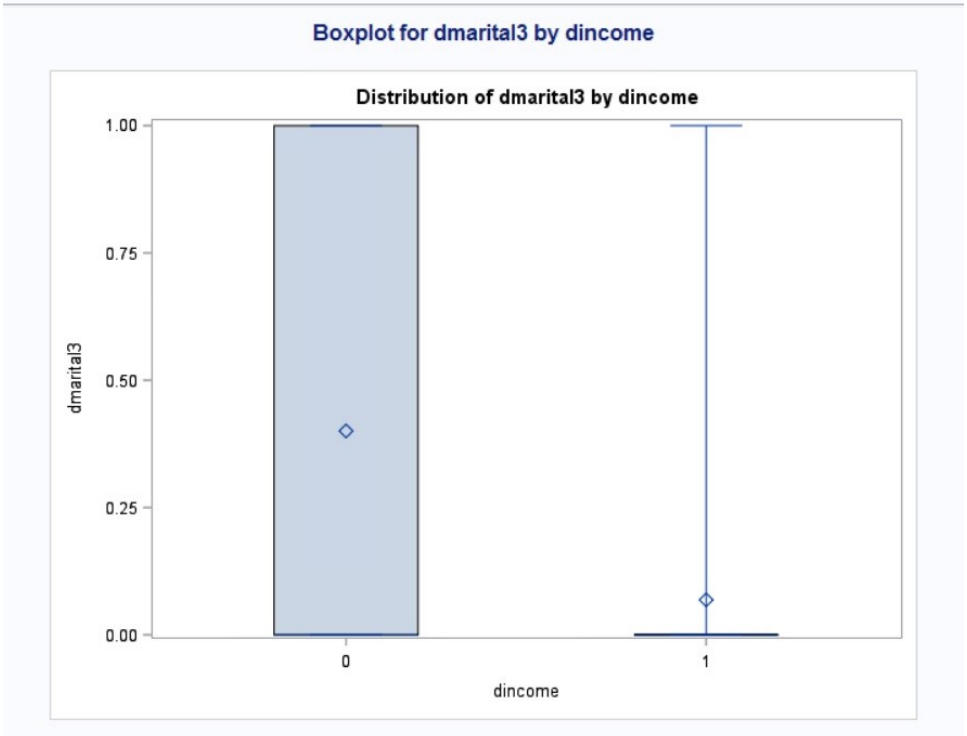


Figure 7. Boxplot of dmartial3 by dincome

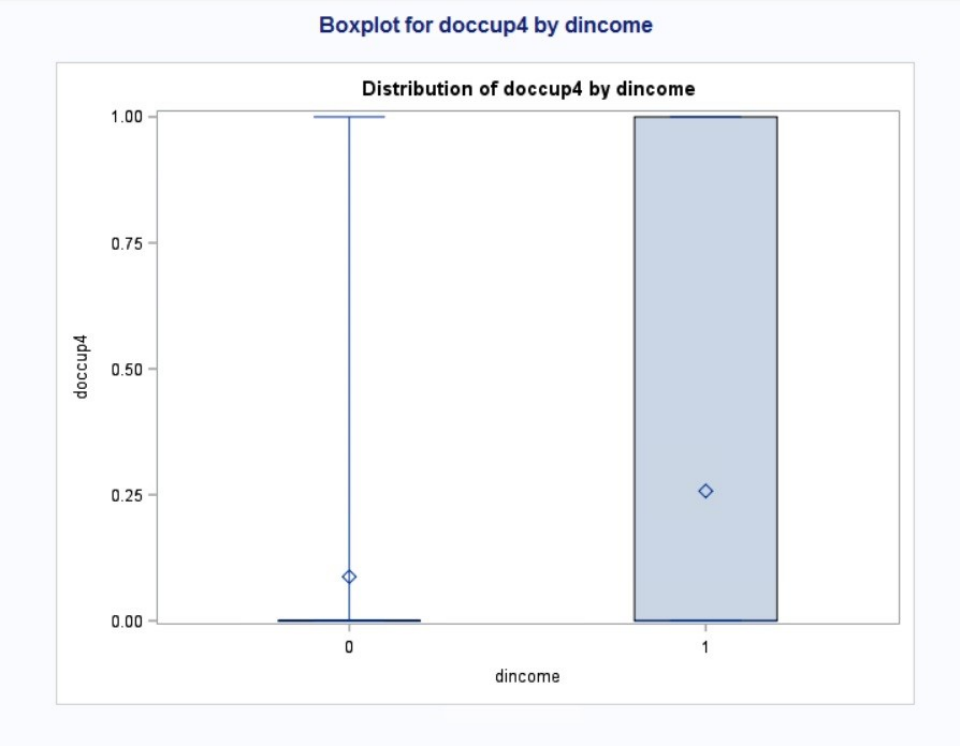


Figure 8. Boxplot of doccup4 by income

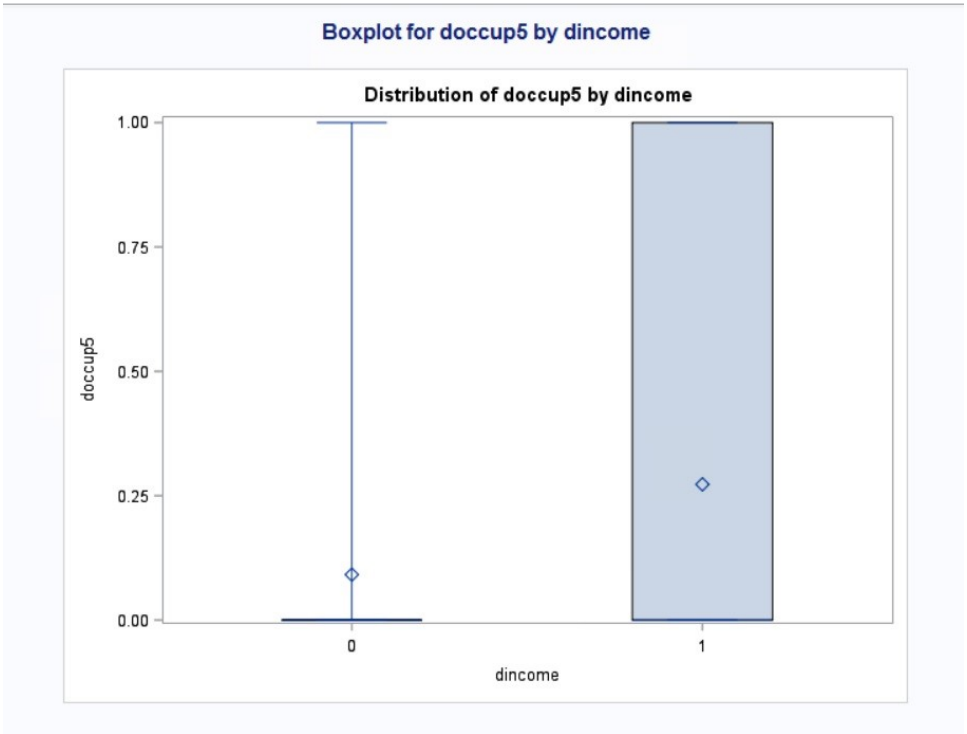


Figure 9. Boxplot of doccup5 by income.

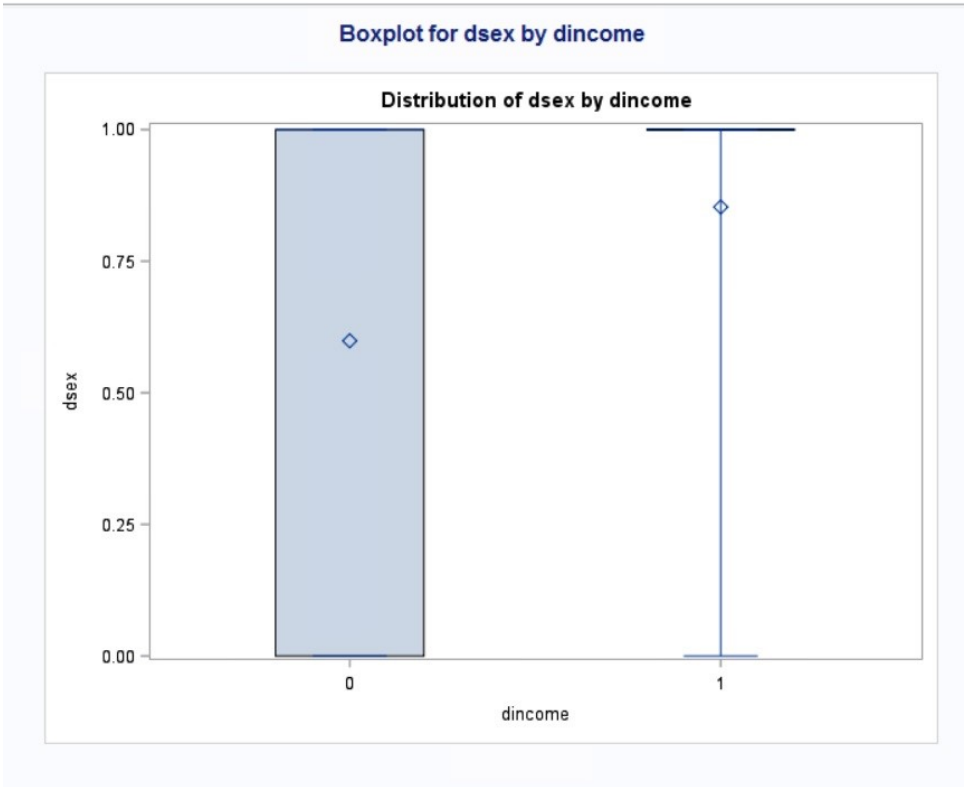


Figure 10. Boxplot of dsex by dincome

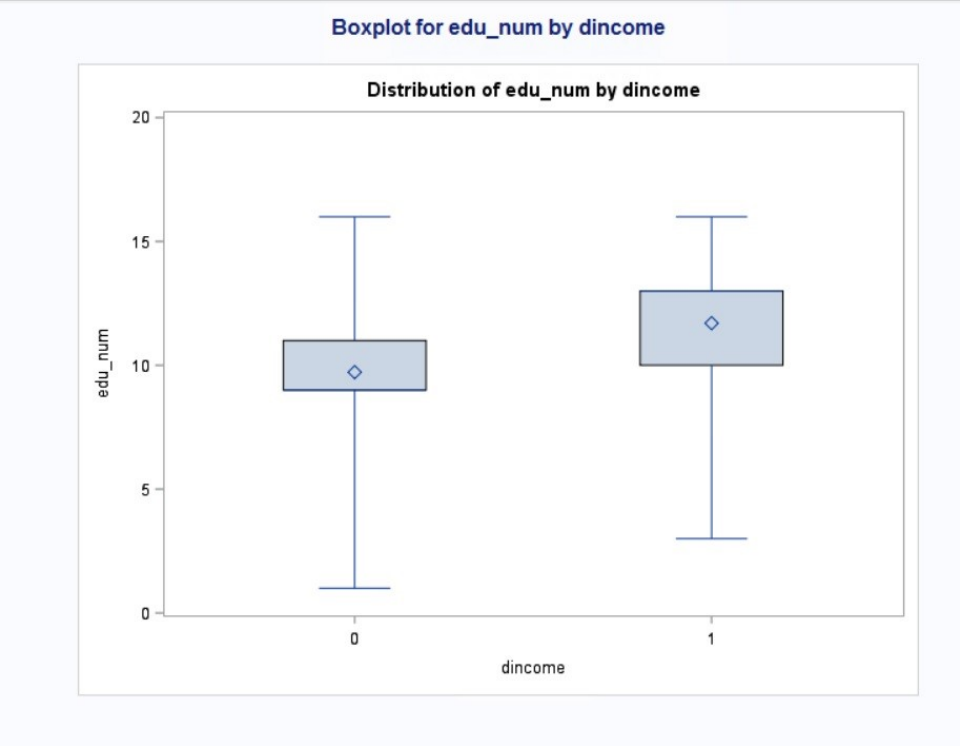


Figure 11. Boxplot of edu_num by dincome

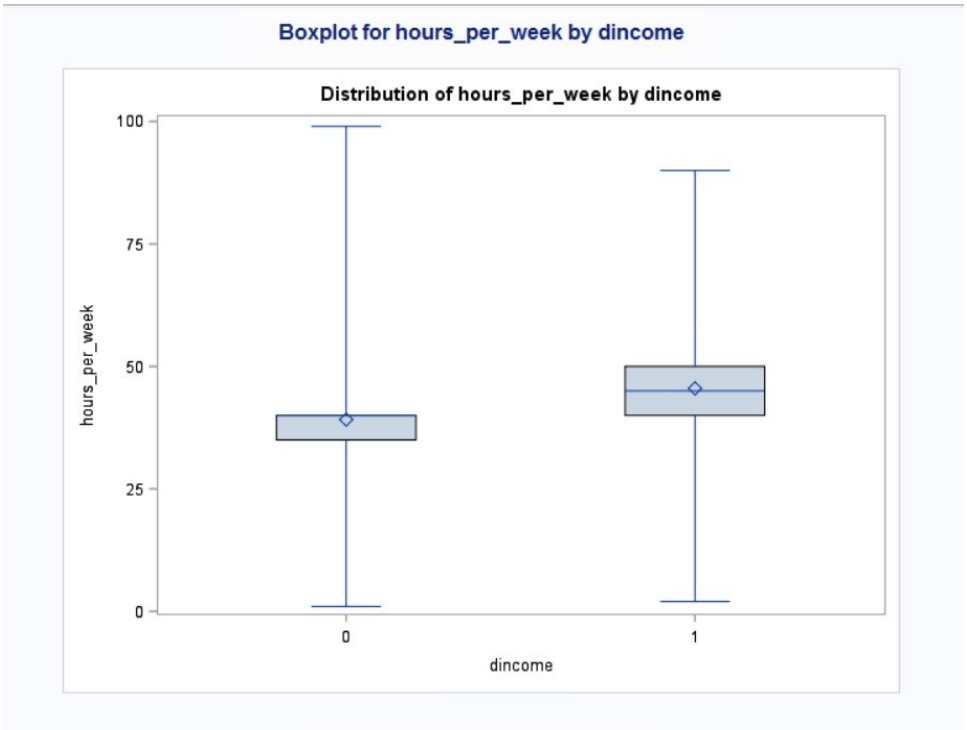


Figure 12. Boxplot of hours_per_week by dincome

	dincome	age	edu_num	capital_gain	capital_loss	hours_per_week	dsex	dworkclass2	dworkclass3	dworkclass4	dedu2	dedu3	dedu4	dedu5	dedu6	dedu7	dmarital2	dmarital3	dmarital4	dmarital5
dincome	1.00000	0.21649 <.0001	0.33901 <.0001	0.21938 <.0001	0.14698 <.0001	0.21725 <.0001	0.23685 <.0001	0.10335 <.0001	0.06662 0.0029	-0.06903 0.0020	-0.13839 <.0001	-0.03782 0.0908	0.16255 <.0001	-0.02914 0.1927	0.15852 <.0001	0.13705 <.0001	-0.12554 <.0001	-0.31441 <.0001	-0.09644 0.0001	-0.08041 0.0003
age	0.21649 <.0001	1.00000	0.01609 0.4720	0.08570 0.0001	0.06521 0.0035	0.05398 0.0158	0.07672 0.0006	0.14383 <.0001	0.07725 0.0005	-0.20168 <.0001	0.02320 0.2997	-0.10052 <.0001	0.03883 0.0826	-0.03297 0.1405	0.09899 <.0001	0.08011 0.0003	0.12264 <.0001	-0.51765 <.0001	-0.00541 0.8088	0.29832 <.0001
edu_num	0.33901 <.0001	0.01609 0.4720	1.00000	0.14282 <.0001	0.09502 <.0001	0.13641 <.0001	0.05050 0.0239	0.08570 0.0001	0.16089 <.0001	-0.13028 <.0001	-0.32886 <.0001	-0.04975 0.0261	0.27503 <.0001	0.13477 <.0001	0.35526 <.0001	0.29082 <.0001	0.00813 0.7164	-0.04164 0.0626	-0.05512 0.0137	-0.12627 <.0001
capital_gain	0.21938 <.0001	0.08570 0.0001	0.14282 <.0001	1.00000	-0.03460 0.1219	0.08811 <.0001	0.07455 0.0008	0.11611 <.0001	-0.00999 0.6551	-0.06481 0.0037	-0.05127 0.0218	-0.04094 0.0672	0.13077 <.0001	-0.01511 0.4995	0.01221 0.5853	0.19045 <.0001	-0.00587 0.7932	-0.07902 0.0004	-0.01644 0.4625	-0.01008 0.6523
capital_loss	0.14698 <.0001	0.06521 0.0035	0.09502 <.0001	-0.03460 0.1219	1.00000	0.02556 0.2533	0.06231 0.0053	0.02301 0.3036	0.03825 0.0873	-0.04683 0.0363	-0.01570 0.4830	-0.00309 0.8901	0.07467 0.0008	-0.04946 0.0270	0.06467 0.0038	0.00785 0.7255	-0.01069 0.6328	-0.07885 0.0004	-0.02468 0.2699	-0.04023 0.0721
hours_per_week	0.21725 <.0001	0.05398 0.0158	0.13641 <.0001	0.08811 <.0001	0.02556 0.2533	1.00000	0.25517 <.0001	0.16824 <.0001	-0.01918 0.3912	-0.00740 0.7410	0.04323 0.0532	-0.08529 0.0001	0.04706 0.0354	0.04629 0.0384	0.04839 0.0305	0.06457 0.0039	0.00397 0.8593	-0.18989 <.0001	-0.01413 0.5278	-0.11856 <.0001
dsex	0.23685 <.0001	0.07672 0.0006	0.05050 0.0239	0.07455 0.0008	0.06231 0.0053	0.25517 <.0001	1.00000	0.16140 <.0001	-0.05226 0.0194	-0.06152 0.0059	0.02372 0.2891	-0.08581 0.0001	0.05395 0.0158	-0.01509 0.5001	0.00523 0.8152	0.05023 0.0247	-0.25118 <.0001	-0.14224 <.0001	-0.11472 <.0001	-0.22369 <.0001
dworkclass2	0.10335 <.0001	0.14383 <.0001	0.08570 0.0001	0.11611 <.0001	0.02301 0.3036	0.16824 <.0001	0.16140 <.0001	1.00000	-0.15108 <.0001	-0.56474 <.0001	0.00487 0.8278	-0.03287 0.1417	0.06538 0.0034	0.03123 0.1627	0.00334 0.8815	0.05469 0.0144	-0.07260 0.0012	-0.12102 <.0001	-0.00624 0.7805	0.00103 0.9632
dworkclass3	0.06662 0.0029	0.07725 0.0005	0.16089 <.0001	-0.00999 0.6551	0.03825 0.0873	-0.01918 0.3912	-0.05226 0.0194	-0.15108 <.0001	1.00000	-0.56847 <.0001	-0.08396 0.0002	0.00066 0.9766	0.02367 0.2901	0.00225 0.9200	0.16999 <.0001	0.05385 0.0160	0.08723 <.0001	-0.04383 0.0500	-0.02844 0.2036	0.01650 0.4609
dworkclass4	-0.06903 0.0020	-0.20168 <.0001	-0.13028 <.0001	-0.06481 0.0037	-0.04683 0.0363	-0.00740 0.7410	-0.06152 0.0059	-0.56474 <.0001	-0.56847 <.0001	1.00000	0.06741 0.0026	-0.00218 0.9225	-0.06089 0.0064	-0.01231 0.5822	-0.10992 <.0001	-0.07102 0.0015	-0.00063 0.9775	0.10314 <.0001	0.01605 0.4732	-0.04282 0.0555
dedu2	-0.13839 <.0001	0.02320 0.2997	-0.32886 <.0001	-0.05127 0.0218	-0.01570 0.4830	0.04323 0.0532	0.02372 0.2891	0.00487 0.8278	-0.08396 0.0002	0.06741 0.0026	1.00000	-0.35993 <.0001	-0.10075 <.0001	-0.19770 <.0001	-0.16476 <.0001	-0.08804 <.0001	-0.01504 0.5016	-0.02812 0.2087	0.03738 0.0946	0.08612 0.0001
dedu3	-0.03782 0.0908	-0.10052 <.0001	-0.04975 0.0261	-0.04094 0.0672	-0.00309 0.8901	-0.08529 0.0001	-0.08581 0.0001	-0.03287 0.1417	0.00066 0.9766	-0.00218 0.9225	-0.35993 <.0001	1.00000	-0.07849 0.0004	-0.15402 <.0001	-0.12836 <.0001	-0.06859 <.0001	0.04267 0.0564	0.06955 0.0019	-0.05760 0.0100	-0.03263 0.1446
dedu4	0.16255 <.0001	0.03883 0.0826	0.27503 <.0001	0.13077 <.0001	0.07467 0.0008	0.04706 0.0354	0.05395 0.0158	0.06538 0.0034	0.02367 0.2901	-0.06089 0.0064	-0.10075 <.0001	-0.07849 0.0004	1.00000	-0.04311 0.0539	-0.03593 0.1082	-0.01920 0.3908	-0.02743 0.2202	-0.07045 0.0016	0.00072 0.9743	-0.02781 0.2138
dedu5	-0.02914 0.1927	-0.03297 0.1405	0.13477 <.0001	-0.01511 0.4995	-0.04946 0.0270	0.04629 0.0384	-0.01509 0.5001	0.03123 0.1627	0.00225 0.9200	-0.01231 0.5822	-0.19770 <.0001	-0.15402 <.0001	-0.04311 0.0539	1.00000	-0.07051 0.0016	-0.03767 0.0921	0.08981 <.0001	-0.03579 0.1095	-0.01877 0.4015	-0.03399 0.1286
dedu6	0.15852 <.0001	0.09899 <.0001	0.35526 <.0001	0.01221 0.5853	0.06467 0.0038	0.04839 0.0305	0.00523 0.8152	0.00334 0.8815	0.16999 <.0001	-0.10992 <.0001	-0.16476 <.0001	-0.12836 <.0001	-0.03593 0.1082	-0.07051 0.0016	1.00000	-0.03140 0.1604	-0.01100 0.6230	-0.06025 0.0070	-0.00053 0.9811	-0.02138 0.3393
dedu7	0.13705 <.0001	0.08011 0.0003	0.29082 <.0001	0.19045 <.0001	0.00785 0.7255	0.06457 0.0039	0.05023 0.0247	0.05469 0.0144	0.05385 0.0160	-0.07102 0.0015	-0.08804 <.0001	-0.06859 <.0001	-0.01920 0.3908	-0.03767 0.0921	-0.03140 0.1604	1.00000	-0.00431 0.8472	-0.02830 0.2058	0.00939 0.6748	-0.02430 0.2774
dmarital2	-0.12554 <.0001	0.12264 <.0001	0.00813 0.7164	-0.00587 0.7932	-0.01069 0.6328	0.00397 0.8593	-0.25118 <.0001	-0.07260 0.0012	0.08723 <.0001	-0.00063 0.9775	-0.01504 0.5016	0.04267 0.0564	-0.02743 0.2202	0.08981 <.0001	-0.01100 0.6230	-0.00431 0.8472	1.00000	-0.26410 <.0001	-0.08533 0.0001	-0.07332 0.0010
dmarital3	-0.31441 <.0001	-0.51765 <.0001	-0.04164 0.0626	-0.07902 0.0004	-0.07885 0.0004	-0.18989 <.0001	-0.14224 <.0001	-0.12102 <.0001	-0.04383 0.0500	0.10314 <.0001	-0.02812 0.2087	0.06955 0.0019	-0.07045 0.0016	-0.03579 0.1095	-0.06025 0.0070	-0.02830 0.2058	-0.26410 <.0001	1.00000	-0.14754 <.0001	-0.12678 <.0001
dmarital4	-0.08644	-0.00541	-0.05512	-0.01644	-0.02468	-0.01413	-0.11472	-0.00624	-0.02844	0.01605	0.03738	-0.05760	0.00072	-0.01877	-0.00053	0.00939	-0.08533	-0.14754	1.00000	-0.04096

Figure 13. Pearson correlation matrix

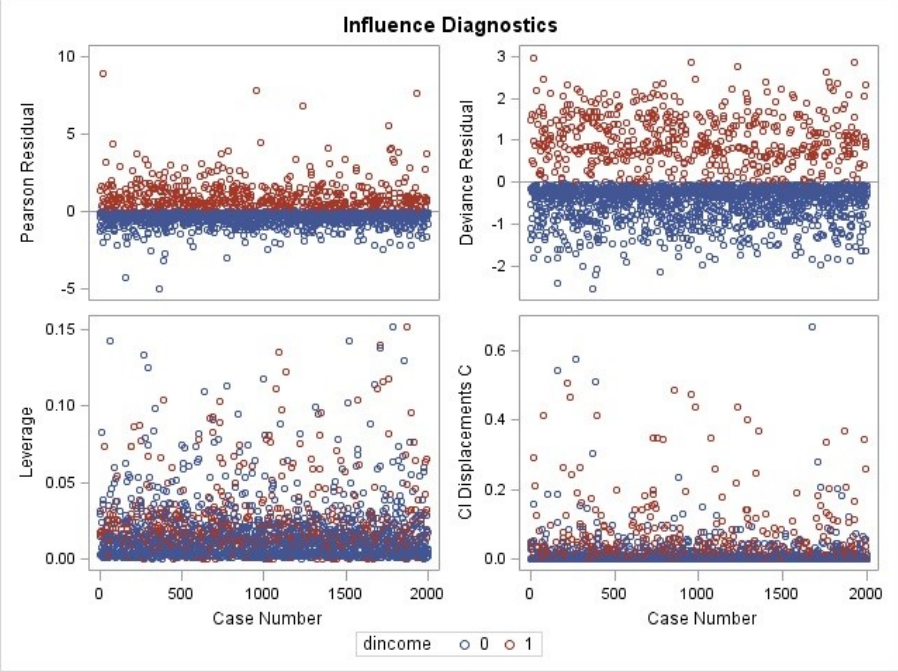
Convergence criterion (GCONV=1E-8) satisfied.

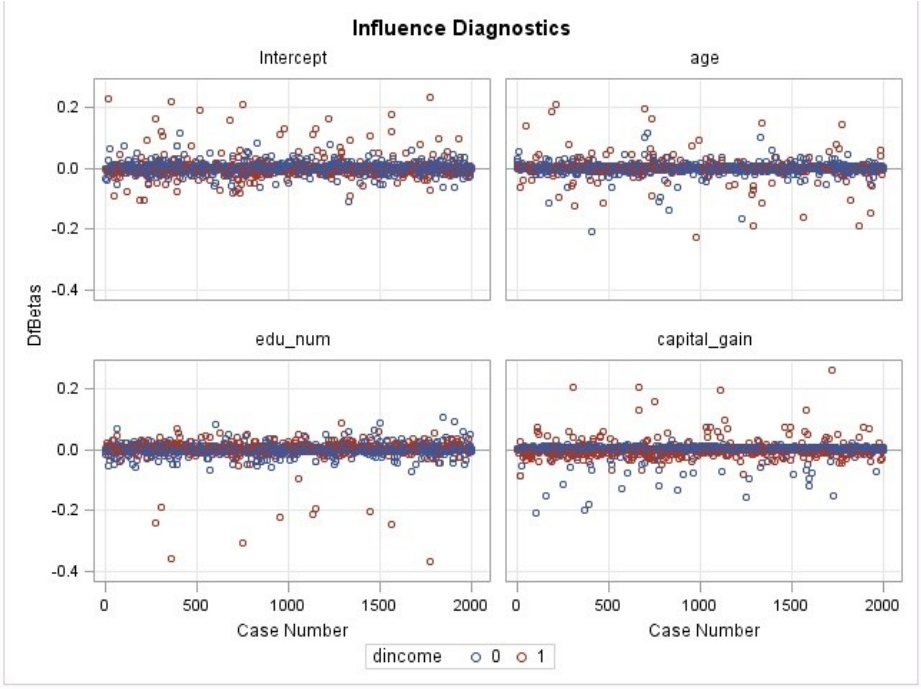
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2302.554	1433.308
SC	2308.155	1623.738
-2 Log L	2300.554	1365.308

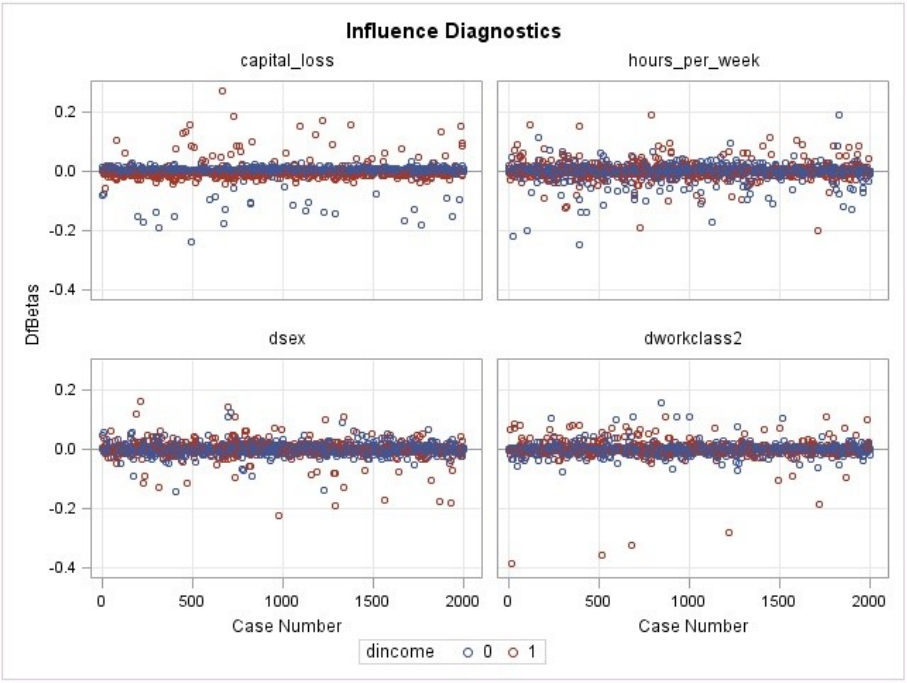
R-Square	0.3735	Max-rescaled R-Square	0.5465

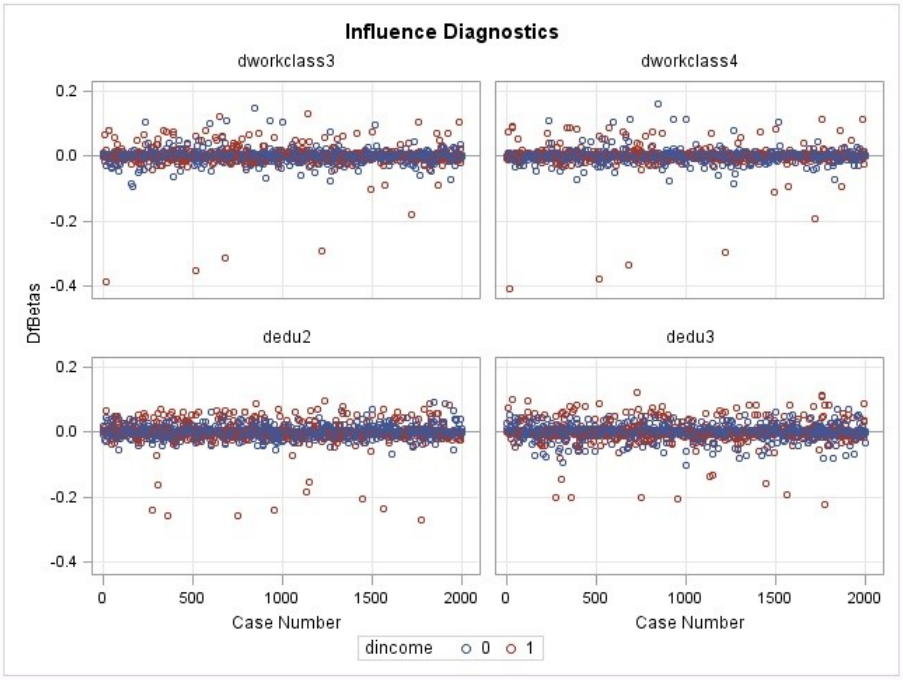
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	935.2461	33	<.0001
Score	723.5431	33	<.0001
Wald	420.3221	33	<.0001

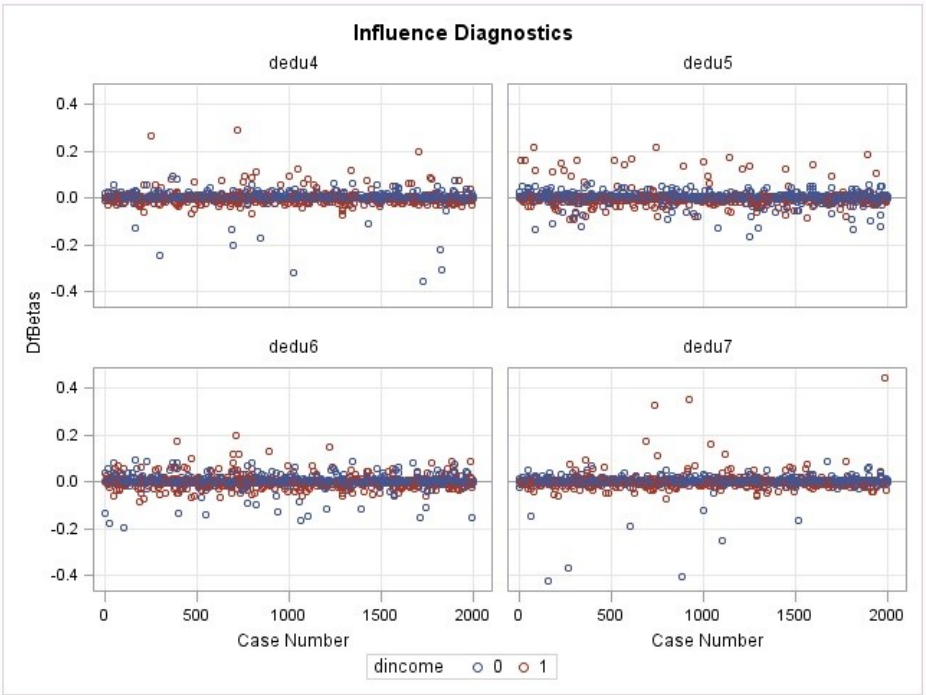
Figure 14. Full model results

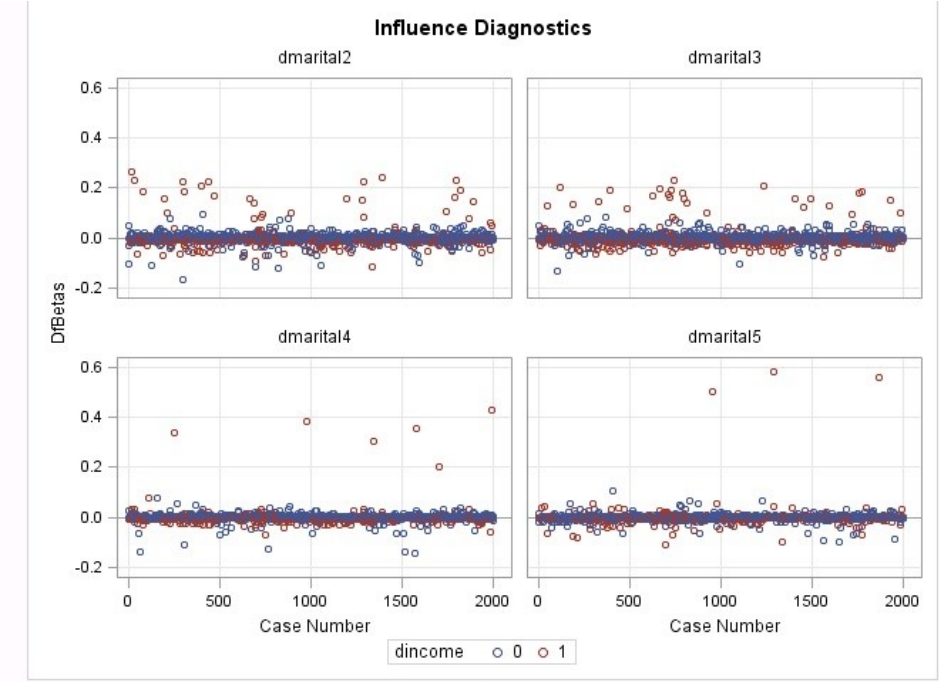


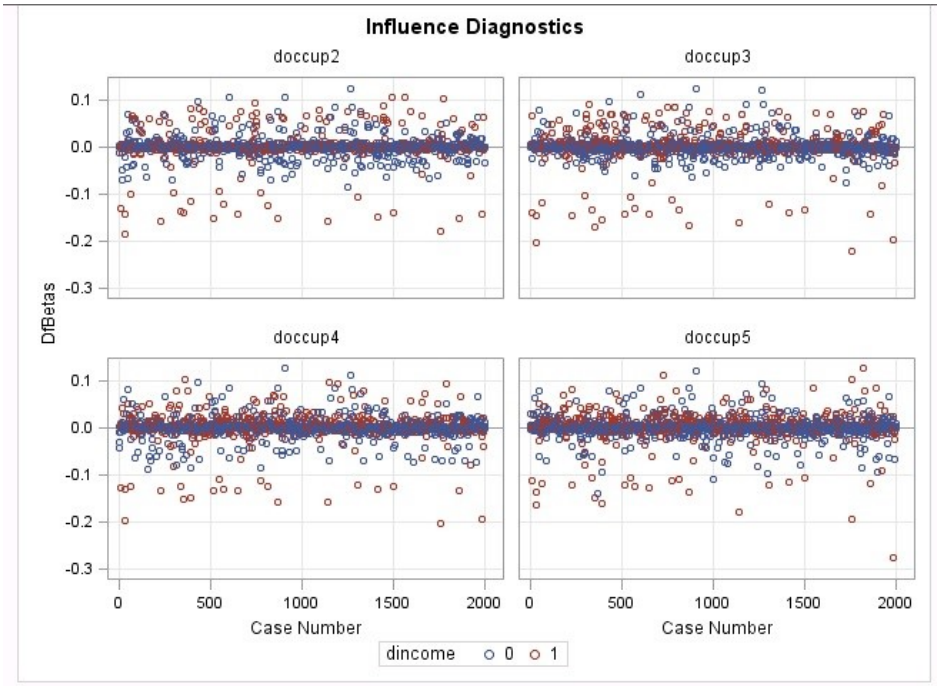

Influence Diagnostics
Intercept
age
edu_num
capital_gain
DfBetas
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Case Number
dincome
0
1

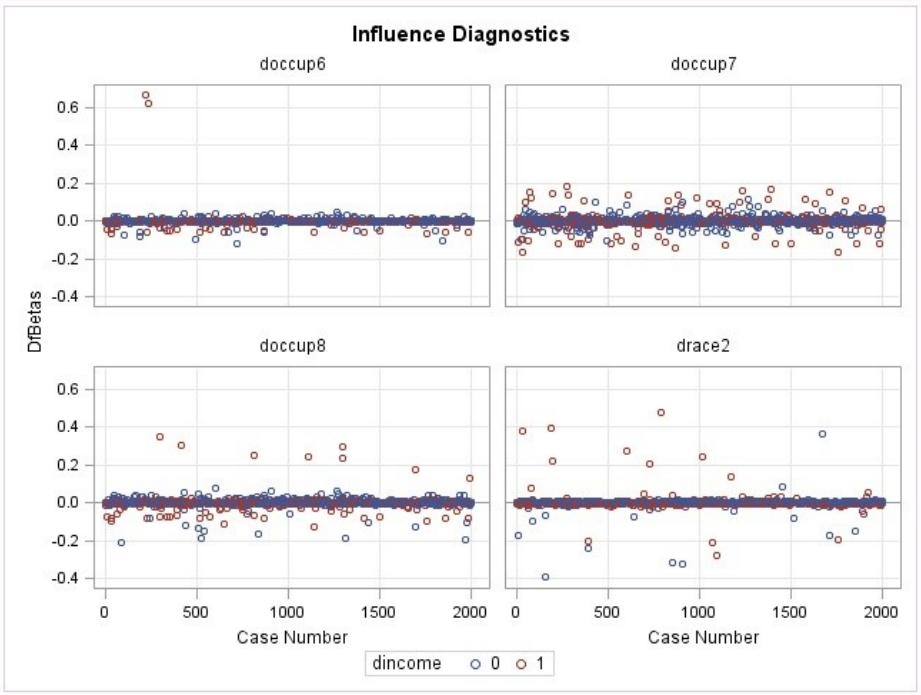

Influence Diagnostics
capital_loss
hours_per_week
dsex
dworkclass2
DfBetas
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Case Number
dincome
0
1

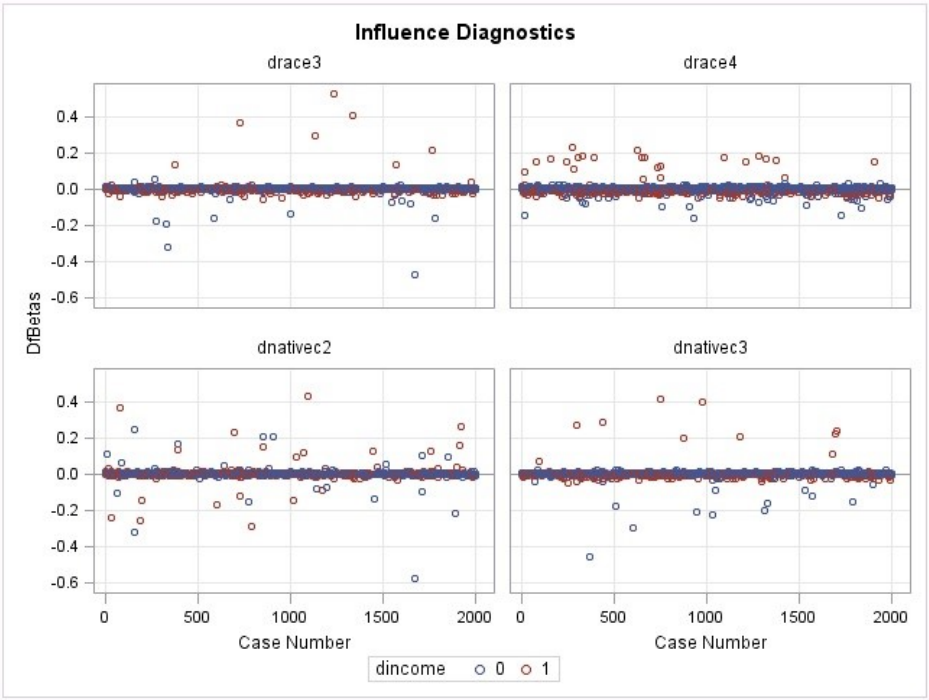

Influence Diagnostics
dworkclass3
dworkclass4
dedu2
dedu3
DfBetas
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Case Number
dincome 0 1


Influence Diagnostics
dedu4
dedu5
dedu6
dedu7
DfBetas
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Case Number
dincome 0 1


Influence Diagnostics
dmarital2
dmarital3
dmarital4
dmarital5
DfBetas
0.6
0.4
0.2
0.0
-0.2
0
500
1000
1500
2000
Case Number
dincome 0 1


Influence Diagnostics
doccup2
doccup3
doccup4
doccup5
DfBetas
0.1
0.0
-0.1
-0.2
-0.3
0
500
1000
1500
2000
Case Number
dincome 0 1


Influence Diagnostics
doccup6
doccup7
doccup8
drace2
DfBetas
0.6
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Case Number
dincome 0 1


Influence Diagnostics
drace3
drace4
dnativec2
dnativec3
DfBetas
0.4
0.2
0.0
-0.2
-0.4
-0.6
0
500
1000
1500
2000
Case Number
dincome 0 1

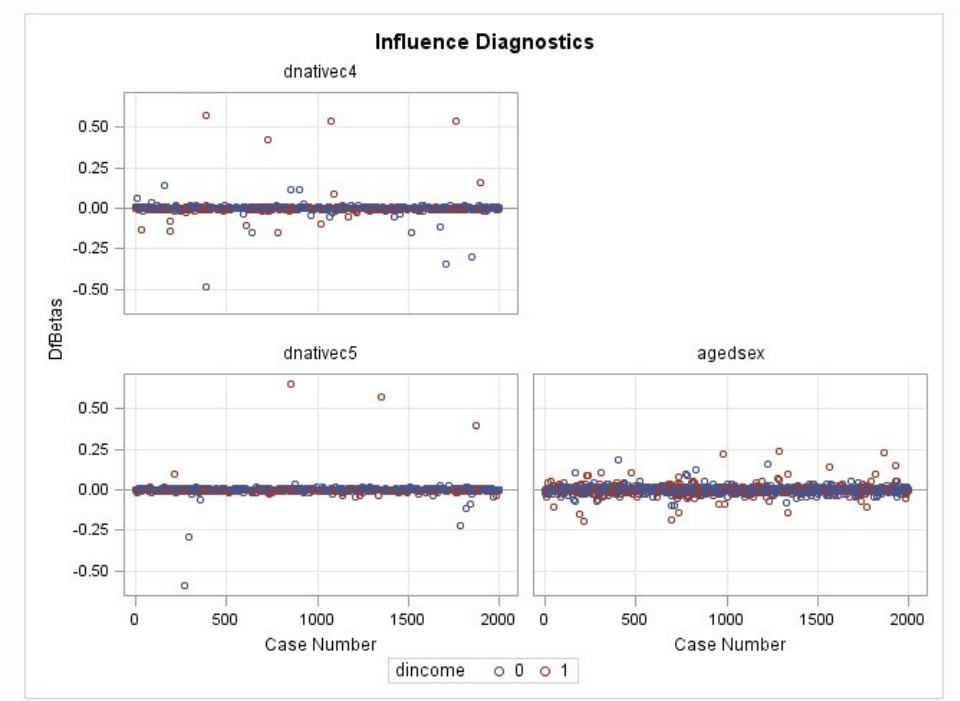


Figure 15. Full model Pearson residuals and dfbetas

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-8.0698	1.0101	63.8299	<.0001	
age	1	0.0319	0.0143	4.9654	0.0259	0.2367
edu_num	1	0.2995	0.0507	34.8441	<.0001	0.4237
capital_gain	1	0.000343	0.000041	70.6295	<.0001	1.2872
capital_loss	1	0.000521	0.000143	13.3654	0.0003	0.1178
hours_per_week	1	0.0204	0.00601	11.5446	0.0007	0.1453
dsex	1	0.5594	0.6586	0.7216	0.3956	0.1456
dworkclass2	1	1.0660	0.5487	3.7747	0.0520	0.1980
dworkclass3	1	1.4050	0.5542	6.4266	0.0112	0.2623
dworkclass4	1	1.4877	0.5239	8.0647	0.0045	0.3827
dedu2	1	0.1257	0.2333	0.2906	0.5899	0.0322
dedu3	1	0.3404	0.2194	2.4071	0.1208	0.0776
dedu4	1	-0.0542	0.4780	0.0129	0.9096	-0.00434
dedu5	1	-0.4023	0.2780	2.0945	0.1478	-0.0595
dedu6	1	-0.3424	0.3054	1.2571	0.2622	-0.0432
dedu7	1	-0.00280	0.5806	0.0000	0.9962	-0.00020
dmarital2	1	-2.0392	0.2634	59.9539	<.0001	-0.3813
dmarital3	1	-2.3455	0.2328	101.5037	<.0001	-0.6001

dmarital3	1	-2.3455	0.2328	101.5037	<.0001	-0.6001
dmarital4	1	-2.0444	0.4653	19.3054	<.0001	-0.2350
dmarital5	1	-2.3082	0.6595	12.2485	0.0005	-0.2307
doccup2	1	0.4758	0.2907	2.6784	0.1017	0.0836
doccup3	1	0.1581	0.2765	0.3267	0.5676	0.0407
doccup4	1	1.1950	0.2947	16.4383	<.0001	0.2231
doccup5	1	1.1405	0.3151	13.1031	0.0003	0.2176
doccup6	1	-1.0982	0.7668	2.0514	0.1521	-0.1158
doccup7	1	-0.1919	0.3420	0.3149	0.5747	-0.0291
doccup8	1	0.1843	0.5211	0.1251	0.7235	0.0130
drace2	1	0.3148	0.5080	0.3840	0.5354	0.0303
drace3	1	0.4721	0.5384	0.7690	0.3805	0.0337
drace4	1	0.0431	0.2701	0.0255	0.8731	0.00729
dnativec2	1	-1.0637	0.6795	2.4506	0.1175	-0.0769
dnativec3	1	0.8679	0.4762	3.3212	0.0684	0.0619
dnativec4	1	-0.9851	0.6951	2.0081	0.1565	-0.0761
dnativec5	1	-0.7438	0.7067	1.1077	0.2926	-0.0602
age*dsex	1	-0.00431	0.0156	0.0762	0.7826	-0.0516

Figure 16. Full model variables

Parameter	Intercept	age	edu_num	capital_gain	capital_loss	hours_per_week	dsex	dworkclass2	dworkclass3	dworkclass4	dedu2	dedu3	dec
Intercept	1.0000	-0.5974	-0.5926	-0.0815	-0.0060	-0.2251	-0.4470	-0.3944	-0.4153	-0.4522	-0.3854	-0.3123	0.1(
age	-0.5974	1.0000	0.0118	-0.0339	-0.0249	0.0707	0.8300	0.0146	0.0231	0.0328	-0.0286	-0.0094	0.0:
edu_num	-0.5926	0.0118	1.0000	0.0473	-0.0337	-0.0177	-0.0298	0.0369	0.0408	0.0530	0.5415	0.3790	-0.2
capital_gain	-0.0815	-0.0339	0.0473	1.0000	0.0771	-0.0274	-0.0342	0.0966	0.0922	0.1128	0.0233	0.0307	-0.0(
capital_loss	-0.0060	-0.0249	-0.0337	0.0771	1.0000	0.0258	-0.0236	0.0534	0.0443	0.0600	-0.0280	-0.0158	0.0(
hours_per_week	-0.2251	0.0707	-0.0177	-0.0274	0.0258	1.0000	-0.0667	-0.0848	-0.0300	-0.0418	-0.0650	-0.0209	-0.0
dsex	-0.4470	0.8300	-0.0298	-0.0342	-0.0236	-0.0667	1.0000	-0.0763	-0.0679	-0.0787	-0.0336	-0.0238	0.0:
dworkclass2	-0.3944	0.0146	0.0369	0.0966	0.0534	-0.0848	-0.0763	1.0000	0.8892	0.9347	0.0056	0.0249	0.0
dworkclass3	-0.4153	0.0231	0.0408	0.0922	0.0443	-0.0300	-0.0679	0.8892	1.0000	0.9256	0.0113	0.0213	0.0:
dworkclass4	-0.4522	0.0328	0.0530	0.1128	0.0600	-0.0418	-0.0787	0.9347	0.9256	1.0000	0.0120	0.0316	0.0(
dedu2	-0.3854	-0.0286	0.5415	0.0233	-0.0280	-0.0650	-0.0336	0.0056	0.0113	0.0120	1.0000	0.6132	-0.0
dedu3	-0.3123	-0.0094	0.3790	0.0307	-0.0158	-0.0209	-0.0238	0.0249	0.0213	0.0316	0.6132	1.0000	0.0(
dedu4	0.1010	0.0330	-0.2989	-0.0082	0.0014	-0.0128	0.0377	0.0499	0.0703	0.0614	-0.0486	0.0074	1.0(
dedu5	-0.1226	0.0141	0.0674	0.0107	0.0519	-0.0299	-0.0019	0.0068	0.0197	0.0189	0.3322	0.3280	0.1(
dedu6	0.1609	-0.0588	-0.3126	0.0152	-0.0207	-0.0329	-0.0166	0.0350	-0.0215	0.0278	0.0282	0.1005	0.2
dedu7	0.2015	-0.0223	-0.3240	0.0000	0.0186	-0.0236	0.0042	-0.0196	0.0034	-0.0064	-0.1018	-0.0425	0.2(
dmarital2	0.1007	-0.2069	-0.0435	-0.0293	-0.0465	-0.0632	-0.0767	0.0016	-0.0398	-0.0235	0.0016	-0.0071	0.0:
dmarital3	-0.1226	0.1291	-0.0752	-0.1078	0.0206	0.0456	0.0737	0.0440	0.0375	0.0313	0.0382	0.0505	0.0
dmarital4	-0.0009	-0.0290	-0.0224	-0.0137	0.0079	-0.0418	0.0141	0.0022	0.0019	0.0062	0.0157	0.0308	-0.0

dmarital3	-0.1226	0.1291	-0.0752	-0.1078	0.0206	0.0456	0.0737	0.0440	0.0375	0.0313	0.0382	0.0505	0.0
dmarital4	-0.0009	-0.0290	-0.0224	-0.0137	0.0079	-0.0418	0.0141	0.0022	0.0019	0.0062	0.0157	0.0308	-0.0
dmarital5	0.1470	-0.3469	0.0149	0.0125	0.0309	0.0341	-0.2319	0.0084	-0.0017	-0.0008	-0.0317	-0.0147	-0.0
doccup2	0.0324	-0.0098	0.0018	0.0104	-0.0351	-0.0068	-0.0193	-0.3634	-0.3482	-0.3580	-0.0655	-0.0827	-0.0
doccup3	0.0416	-0.0027	-0.0677	0.0041	-0.0503	-0.0227	0.0496	-0.4056	-0.3839	-0.3918	0.0286	-0.0639	0.0
doccup4	0.0765	0.0048	-0.1215	0.0224	-0.0279	-0.0716	0.0538	-0.3927	-0.3704	-0.3719	0.0408	-0.0225	0.0
doccup5	0.0743	-0.0082	-0.1511	0.0369	-0.0241	-0.0319	0.0384	-0.3715	-0.4003	-0.3564	0.0855	0.0430	-0.1
doccup6	0.0105	-0.0183	0.0083	0.0185	0.0096	0.0119	-0.0326	-0.1182	-0.1169	-0.1281	-0.0322	-0.0210	-0.0
doccup7	0.0607	-0.0139	0.0057	-0.0292	-0.0519	-0.1346	0.0420	-0.3487	-0.3274	-0.3206	-0.0247	-0.0676	0.0
doccup8	0.0546	-0.0212	-0.0429	0.0267	-0.0208	-0.0222	-0.0212	-0.1845	-0.2626	-0.1911	-0.0419	-0.0484	0.0
drace2	0.0052	-0.0168	-0.0261	0.0018	0.0350	0.0014	-0.0006	0.0024	0.0125	0.0159	-0.0107	-0.0114	0.0
drace3	-0.0983	0.0169	0.0899	-0.0373	0.0181	-0.0017	0.0283	0.0352	0.0323	0.0392	0.0541	0.0525	-0.0
drace4	-0.0780	0.0480	0.0320	0.0382	0.0067	0.0428	0.0532	-0.0158	-0.0553	-0.0213	-0.0186	-0.0386	0.0
dnativec2	0.0304	-0.0211	-0.0325	-0.0131	-0.0452	0.0207	-0.0263	0.0073	-0.0073	-0.0238	0.0330	0.0473	0.0
dnativec3	-0.0329	0.0047	0.0172	0.0103	0.0307	0.0229	-0.0009	-0.0330	-0.0326	-0.0317	-0.0232	0.0337	-0.0
dnativec4	-0.0331	0.0098	-0.0089	0.0107	-0.0450	0.0206	0.0194	0.0140	0.0448	0.0293	0.0341	0.0312	0.0
dnativec5	-0.0297	-0.0015	0.0229	0.0109	-0.0228	0.0262	-0.0124	0.0150	-0.0027	-0.0008	0.0345	0.0213	-0.0
agedsex	0.4379	-0.8838	0.0252	0.0405	0.0135	0.0126	-0.9524	0.0674	0.0677	0.0762	0.0458	0.0396	-0.0

Figure 17. Full model correlation of coefficients matrix

0	8.8983	2.9612	0.00364	0.2287	-0.0647	-0.0220	-0.0845	-0.0559	0.0221	-0.00677	-0.3839
	0.7942	0.9890	0.0323	0.00392	0.00573	0.00468	0.00236	0.00983	0.0117	0.0134	0.00617

Observation 15

	3.3603	2.2400	0.0164	0.1615	0.0106	-0.2397	-0.0295	-0.0151	-0.0401	0.0447	-0.0372

Observation 277

	-3.2162	-2.2040	0.0450	0.0742	-0.00906	-0.0482	-0.1780	0.0211	0.00623	0.0155
	2.1977	1.8778	0.0281	-0.0751	0.0287	0.00416	-0.0369	-0.0283	0.1513	-0.00703

Observation 389

	3.0646	2.1638	0.0146	-0.0724	0.0445	0.0538	-0.0329	-0.0250	0.0125	0.0363	0.00118	0.0574

Observation 623

0	7.8301	2.8748	0.00759	0.1111	0.0715	-0.2208	-0.0396	-0.00381	0.0200	0.0733

Observation 955

	6.7760	2.7743	0.00940	-0.0335	0.0197	0.0310	-0.0831	-0.0200	-0.0853	0.0982	-0.0128

Observation 1233

	2.9492	2.1316	0.0158	0.0695	-0.0854	-0.0202	-0.0140	-0.0250	-0.00616	-0.0795	-0.0125

Observation 1281

0	3.3328	2.2333	0.00749	-0.0313	0.0255	0.0589	-0.0361	-0.0145	-0.0202

Observation 1491

0	5.5655	2.6325	0.00429	-0.0374	0.0650	0.00498	-0.0433	0.000752	0.0300	-

Observation 1758

	3.2145	2.2035	0.0108	-0.00271	-0.0428	-0.0161	-0.0289	-0.0332	0.0872	-0.0277	0.0368	-0.0116

Observation 1795

	3.8425	2.3485	0.00744	0.0945	-0.1052	-0.0425	-0.0154	-0.0218	-0.0210	-0.1038	-0.0188

Observation 1820

0	7.6303	2.8570	0.00195	0.0953	-0.1477	-0.0244	-0.0408	-0.00747	0.0589	-0.1794	-0.0121
	0.4556	0.6142	0.0275	0.0183	0.0509	0.00059	0.00447	0.00559	0.0341	0.0395	0.00143

Observation 1928

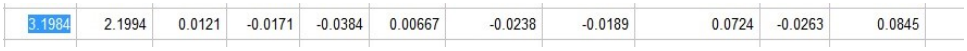
Figure 18. Observations that are influential points and outliers, in order: 15, 277, 389, 623, 955, 1233, 1281, 1491, 1758, 1795, 1820, 1928

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2272.318	1356.354
SC	2277.913	1552.175
-2 Log L	2270.318	1286.354

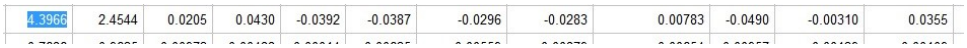
R-Square	0.3904	Max-rescaled R-Square	0.5734

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	983.9645	34	<.0001
Score	749.3297	34	<.0001
Wald	412.9877	34	<.0001

Figure 19. Full model after removing observations that are influential points and outliers



Observation 33



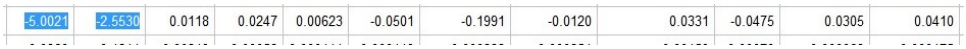
Observation 77

Observation 120

Observation 157

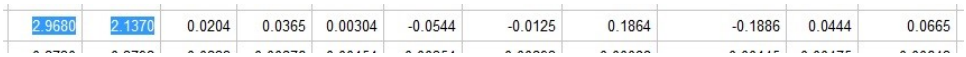
Observation 222

Observation 294

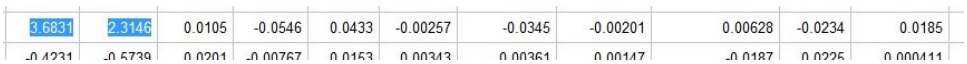


Observation 366

Observation 434



Observation 726



Observation 742

	-3.0347	-2.1556	0.00574	0.0259	-0.00443	-0.0506	-0.1172	0.00728	0.00222

Observation 774

)	3.8984	2.3600	0.0217	-0.0474	0.0268	-0.0488	-0.0490	0.00493	0.1928

Observation 786

	4.4618	2.4658	0.0210	0.1312	-0.2278	-0.00978	-0.0159	0.00677	0.0366	-0.2226

Observation 977

0	3.3437	2.2360	0.00760	0.0364	-0.00662	-0.0296	-0.0229	-0.0242	-0.0242	-0.0213

Observation 1196

)	4.1207	2.4039	0.00756	0.0216	-0.0503	0.00760	-0.0330	-0.0373	-0.0309

Observation 1394

0	3.0981	2.1729	0.00747	-0.0568	0.0358	0.0396	-0.0401	-0.0152	0.0881	0.0584

Observation 1403

	4.0339	2.3871	0.0122	0.2330	-0.00126	-0.3649	-0.0354	-0.0165	-0.0701

Observation 1772

	4.0685	2.3938	0.00739	-0.0738	0.1417	0.00654	-0.0419	-0.0113	0.0218

Observation 1774

	3.3975	2.2491	0.00743	0.0444	-0.0199	-0.0299	-0.0225	-0.0239	-0.0252	-0.0323

Observation 1786

)	3.7108	2.3206	0.0182	0.00563	-0.0184	-0.00646	-0.0284	-0.0138	0.0420

Observation 1992

Figure 20: Outliers, in order: 33, 77, 120, 157, 222, 294, 366, 434, 726, 742, 774, 786, 977, 1196, 1394, 1403, 1772, 1774, 1786, 1992

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2246.052	1284.306
SC	2251.641	1479.915
-2 Log L	2244.052	1214.306

R-Square	0.4061	Max-rescaled R-Square	0.5983

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1029.7460	34	<.0001
Score	770.7129	34	<.0001
Wald	401.3651	34	<.0001

Figure 21: Full model after removing more outliers in addition to previously removed observations

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1711.567	1031.280
SC	1716.874	1121.503
-2 Log L	1709.567	997.280

R-Square	0.3798	Max-rescaled R-Square	0.5567

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	712.2861	16	<.0001
Score	543.8560	16	<.0001
Wald	306.8555	16	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
19.2225	18	0.3782

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.5738	0.9028	70.3820	<.0001
age	1	0.0270	0.00764	12.5111	0.0004
edu_num	1	0.2809	0.0413	46.3667	<.0001
capital_gain	1	0.000366	0.000050	53.8642	<.0001
capital_loss	1	0.000660	0.000165	16.0365	<.0001
hours_per_week	1	0.0208	0.00671	9.5698	0.0020
dworkclass2	1	1.2277	0.6094	4.0593	0.0439
dworkclass3	1	1.6779	0.6219	7.2784	0.0070
dworkclass4	1	1.7319	0.5877	8.6846	0.0032
dmarital2	1	-2.3294	0.2883	65.2966	<.0001
dmarital3	1	-2.5443	0.2690	89.4596	<.0001
dmarital4	1	-2.6342	0.5911	19.8608	<.0001
dmarital5	1	-4.4672	1.4967	8.9089	0.0028
doccup2	1	0.6150	0.2362	6.7807	0.0092
doccup4	1	1.2094	0.2273	28.3113	<.0001
doccup5	1	0.9855	0.2548	14.9574	0.0001
dnativec3	1	1.1862	0.5524	4.6112	0.0318

Figure 22: Stepwise selection model result

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1711.567	1031.280
SC	1716.874	1121.503
-2 Log L	1709.567	997.280

R-Square	0.3798	Max-rescaled R-Square	0.5567

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	712.2861	16	<.0001
Score	543.8560	16	<.0001
Wald	306.8555	16	<.0001

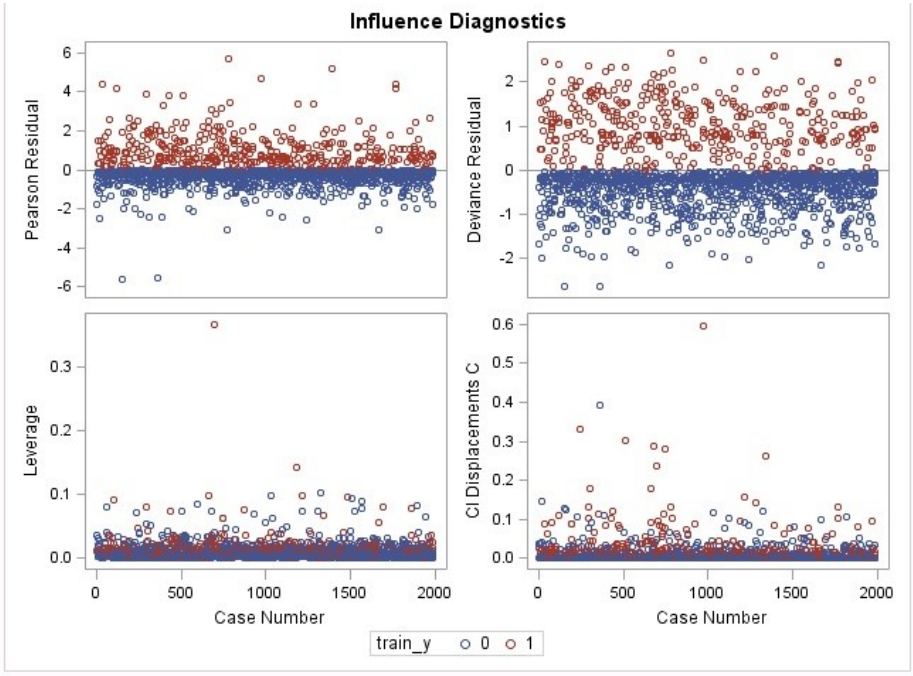
Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
19.2225	18	0.3782

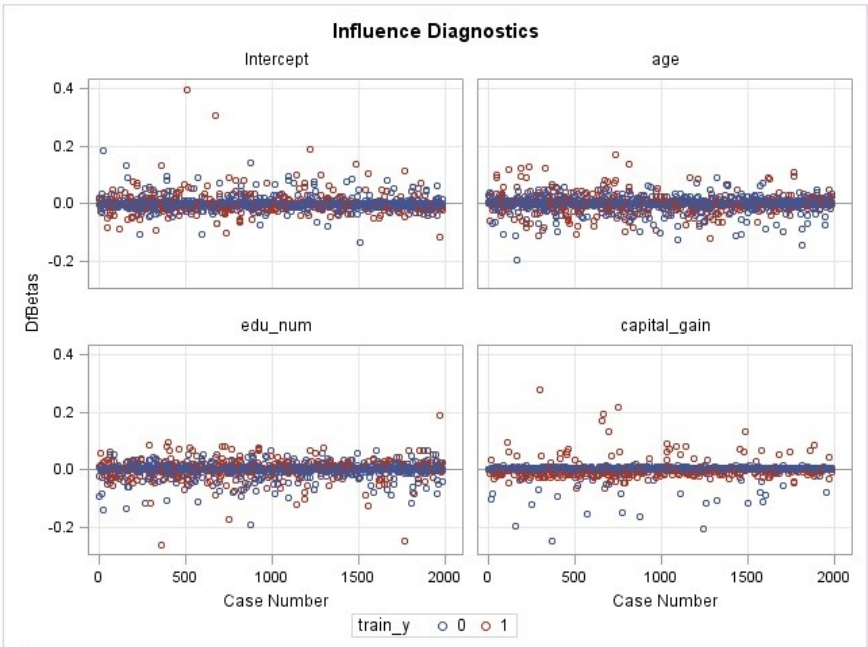
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.5738	0.9028	70.3820	<.0001
age	1	0.0270	0.00764	12.5111	0.0004
edu_num	1	0.2809	0.0413	46.3667	<.0001
capital_gain	1	0.000366	0.000050	53.8642	<.0001
capital_loss	1	0.000660	0.000165	16.0365	<.0001
hours_per_week	1	0.0208	0.00671	9.5698	0.0020
dworkclass2	1	1.2277	0.6094	4.0593	0.0439
dworkclass3	1	1.6779	0.6219	7.2784	0.0070
dworkclass4	1	1.7319	0.5877	8.6846	0.0032
dmarital2	1	-2.3294	0.2883	65.2966	<.0001
dmarital3	1	-2.5443	0.2690	89.4596	<.0001
dmarital4	1	-2.6342	0.5911	19.8608	<.0001
dmarital5	1	-4.4672	1.4967	8.9089	0.0028
doccup2	1	0.6150	0.2362	6.7807	0.0092
doccup4	1	1.2094	0.2273	28.3113	<.0001
doccup5	1	0.9855	0.2548	14.9574	0.0001
dnativec3	1	1.1862	0.5524	4.6112	0.0318

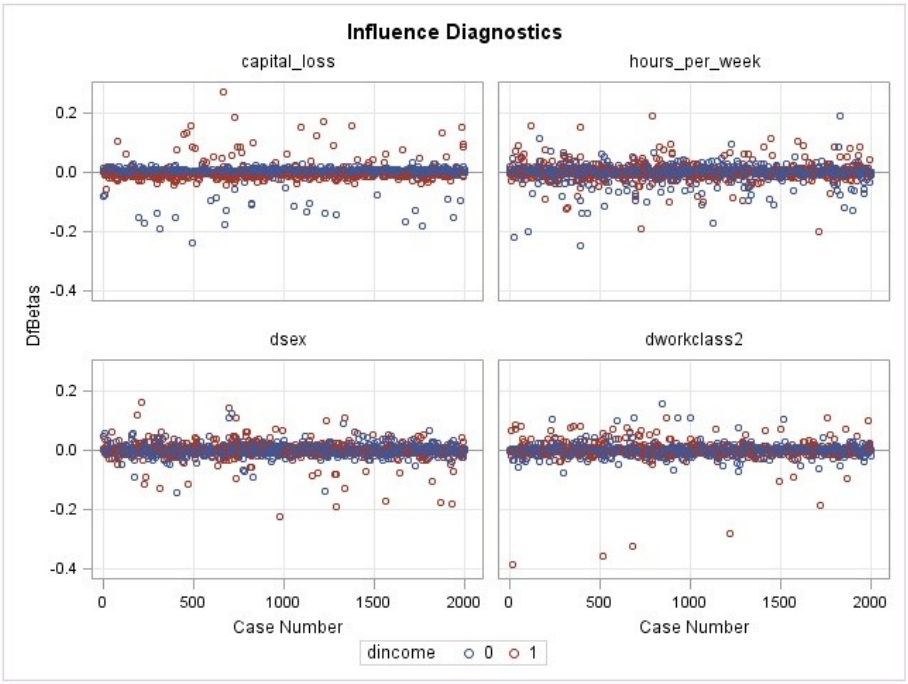
Figure 23: Backward selection model result

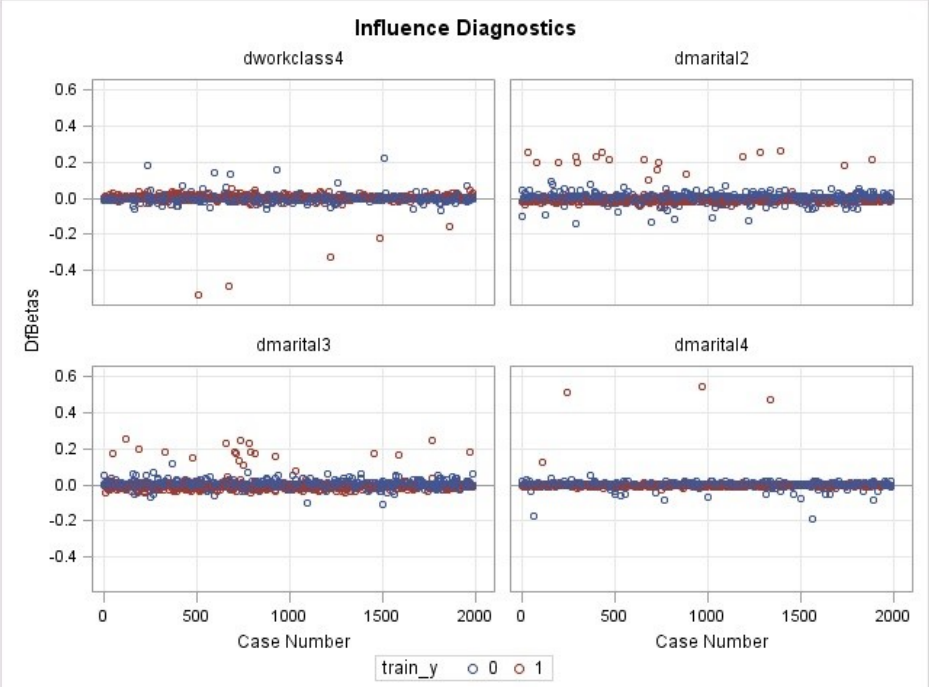
Parameter	Intercept	age	edu_num	capital_gain	capital_loss	hours_per_week	dworkclass2	dworkclass3	dworkclass4	dmarital2	dmarital3	dmarita
Intercept	1.0000	-0.5380	-0.5225	-0.0920	-0.0651	-0.3217	-0.6249	-0.6463	-0.7046	0.0745	-0.0790	0.03
age	-0.5380	1.0000	0.0129	-0.0044	-0.0365	0.1336	0.1453	0.1689	0.2008	-0.0786	0.2417	0.00
edu_num	-0.5225	0.0129	1.0000	0.0417	-0.0220	0.0034	0.0686	0.0692	0.0944	-0.0769	-0.1360	-0.06
capital_gain	-0.0920	-0.0044	0.0417	1.0000	0.0818	0.0026	0.0528	0.0731	0.0766	-0.0443	-0.0990	-0.02
capital_loss	-0.0651	-0.0365	-0.0220	0.0818	1.0000	0.0213	0.0648	0.0708	0.0841	-0.0022	0.0194	0.00
hours_per_week	-0.3217	0.1336	0.0034	0.0026	0.0213	1.0000	-0.1552	-0.0798	-0.0932	0.0061	0.0916	-0.01
dworkclass2	-0.6249	0.1453	0.0686	0.0528	0.0648	-0.1552	1.0000	0.8823	0.9315	-0.0197	0.0059	-0.02
dworkclass3	-0.6463	0.1689	0.0692	0.0731	0.0708	-0.0798	0.8823	1.0000	0.9182	-0.0881	-0.0005	-0.02
dworkclass4	-0.7046	0.2008	0.0944	0.0766	0.0841	-0.0932	0.9315	0.9182	1.0000	-0.0593	-0.0115	-0.01
dmarital2	0.0745	-0.0786	-0.0769	-0.0443	-0.0022	0.0061	-0.0197	-0.0881	-0.0593	1.0000	0.1324	0.06
dmarital3	-0.0790	0.2417	-0.1360	-0.0990	0.0194	0.0916	0.0059	-0.0005	-0.0115	0.1324	1.0000	0.07
dmarital4	0.0340	0.0079	-0.0681	-0.0237	0.0001	-0.0138	-0.0203	-0.0264	-0.0195	0.0646	0.0730	1.00
dmarital5	0.0272	-0.0841	0.0211	-0.2395	-0.0020	0.0600	-0.0443	-0.0338	-0.0366	0.0377	0.0261	0.01
doccup2	-0.0837	0.0470	0.0998	0.0170	-0.0223	0.0357	-0.0939	-0.0809	-0.0983	0.0171	0.0464	0.00
doccup4	0.1000	0.0054	-0.1987	0.0470	0.0353	-0.0213	-0.1385	-0.1225	-0.1067	-0.0922	-0.0365	-0.00
doccup5	0.1991	0.0262	-0.4350	0.0317	0.0439	0.0292	-0.1220	-0.1918	-0.1119	-0.0473	-0.0565	-0.00
dnativec3	-0.0626	0.0669	0.0282	0.0185	0.0349	0.0435	-0.0258	-0.0222	-0.0142	-0.0482	-0.0100	-0.05

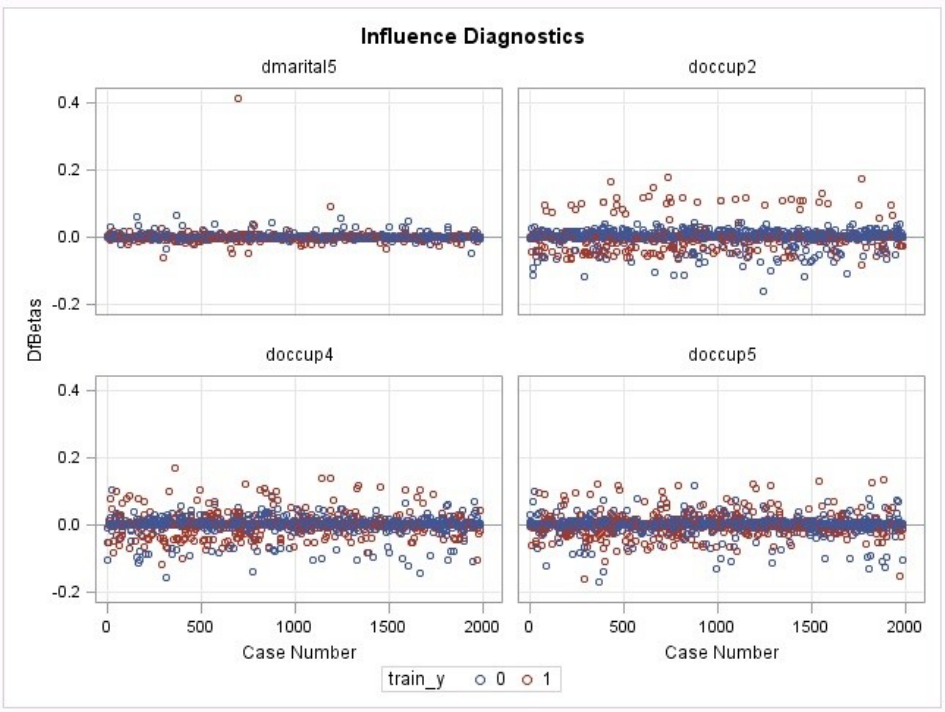
Figure 24: Fitted model correlation of coefficients matrix


Influence Diagnostics
Pearson Residual
Deviance Residual
Leverage
CI Displacements C
Case Number
Case Number
train_y 0 1


Influence Diagnostics
Intercept
age
edu_num
capital_gain
DfBetas
Case Number
Case Number
train_y 0 1


Influence Diagnostics
capital_loss
hours_per_week
dsex
dworkclass2
DfBetas
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Case Number
dincome 0 1


Influence Diagnostics
dworkclass4
dmarital2
dmarital3
dmarital4
DfBetas
0.6
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Case Number
train_y 0 1



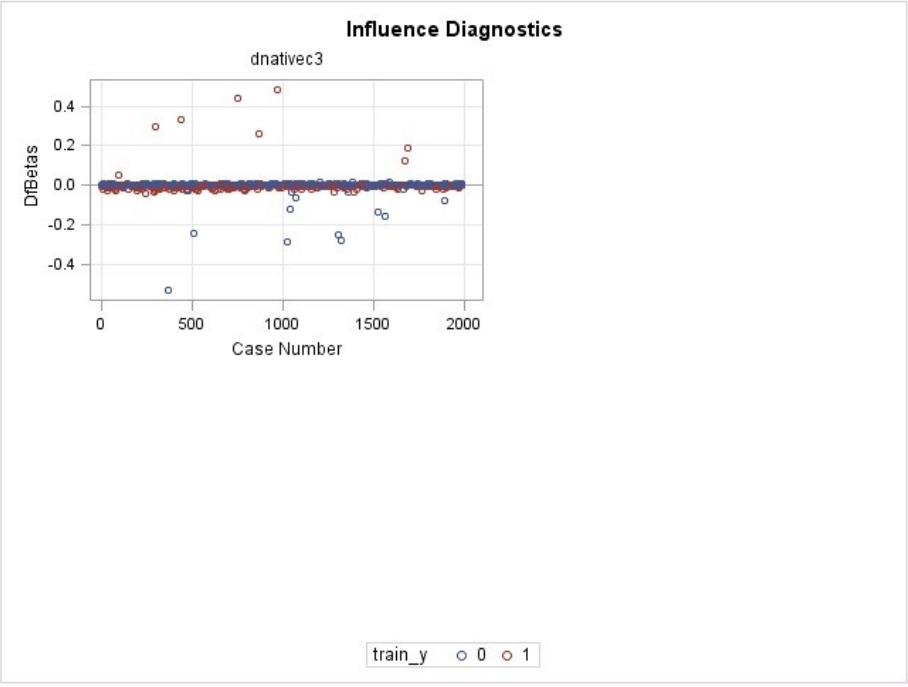


Figure 25: Fitted model Pearson residuals and dfbetas

	3.8540	2.3509	0.00867	0.00640	-0.1087	0.0848	-0.0265	-0.0270	0.0289	-0.00192	0.0599

Observation 292

	-5.5564	-2.6313	0.0124	0.00378	0.0598	-0.0383	-0.2475	-0.0269	0.0131	0.0211	0.0330

Observation 364

	3.3079	2.2271	0.00768	-0.0495	0.0174	0.0957	-0.0271	-0.0271	-0.0224	0.0240	0.00767	0.0312

Observation 396

)	3.7985	2.3393	0.0201	0.3946	-0.1067	-0.0514	-0.0634	-0.0700	0.0541	-0.5106

Observation 509

	3.0771	2.1672	0.0287	0.3056	0.0803	-0.0862	-0.0643	-0.0742	0.0774	-0.4740
	-0.3816	-0.5215	0.0159	-0.00097	0.00645	0.00146	0.00422	0.00149	-0.0127	0.00541

Observation 675

0	5.7406	2.6552	0.00399	-0.00591	-0.0270	-0.0751	-0.0474	-0.00786	0.2377	-0.0430	-0.0269	-0.0259
0	.	.	.	.	.	.	.	.	.	.	.	.

Observation 782

)0	4.6542	2.4983	0.0261	0.0723	-0.0915	-0.0649	-0.0257	0.00230	0.0313	-0.0421

Observation 972

)	3.3906	2.2474	0.00811	0.0704	-0.0187	-0.0993	-0.0235	-0.0151	-0.0323	-0.0207	-0.034

Observation 1191

0	3.3994	2.2495	0.0121	0.0997	-0.1217	-0.0211	-0.0313	-0.0227	-0.0993	0.0412	-0.0564

Observation 1281

0	-3.1169	-2.1779	0.00995	0.0795	-0.0753	-0.0691	-0.0139	-0.2169	0.0474	-0.0275	-0.0330	-0.0571

Observation 1664

	4.3761	2.4508	0.00678	0.1152	0.0933	-0.2470	-0.0435	-0.0341	-0.1157	0.0889	0.0269

Observation 1763

Figure 26: Fitted Model outliers and influential points, in order: 292, 364, 396, 509, 675, 782, 972, 1191, 1281, 1664, 1763

)	4.4146	2.4576	0.00452	-0.0108	0.00274	-0.0192	-0.0325	-0.0220	0.0879

Observation 32

	4.1675	2.4127	0.00622	-0.0878	0.1182	-0.0323	-0.0454	-0.0155	0.1888	-0.00723	0.00728

Observation 119

	-5.6642	-2.6454	0.00392	0.1330	-0.1093	-0.1321	-0.1942	0.0116	0.00229	-0.0313	-0.1007	-0.0470

Observation 156

0	3.8286	2.3456	0.00697	0.00671	-0.0186	0.00269	-0.0284	-0.0275	0.0159	-0.0176	-0.0298	-0.0147

Observation 431

)	.	.	.	.	.	.	.	.	.	.	.
)	3.1762	2.1936	0.0111	-0.1011	0.1705	0.0287	-0.0394	-0.0271	0.0146	0.0142

Observation 738

	-3.0689	-2.1650	0.00669	0.0345	0.0325	-0.0790	-0.1499	0.00251	-0.00481	0.00941	-0.00033

Observation 770

)	3.4191	2.2543	0.00868	0.0449	-0.0558	-0.0345	-0.0437	-0.0142	0.0120	0.0292

Observation 790

0	5.1692	2.5777	0.00316	0.0258	-0.0354	-0.0201	-0.0329	-0.0225	0.0201	0.000271

Observation 1387

4.2141	2.4214	0.00574	-0.0448	0.1088	-0.0105	-0.0422	-0.0251	-0.00650	0.00443	0.00664

Observation 1765

Figure 27: Fitted model outliers, in order: 32, 119, 156, 431, 738, 770, 790, 1387, 1765

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1688.126	971.493
SC	1693.426	1061.590
-2 Log L	1686.126	937.493

R-Square	0.3970	Max-rescaled R-Square	0.5839

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	748.6328	16	<.0001
Score	560.7776	16	<.0001
Wald	301.3970	16	<.0001

Figure 28: Fitted model with outliers and influential points removed

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-7.5738	0.9028	70.3820	<.0001	
age	1	0.0270	0.00764	12.5111	0.0004	0.1982
edu_num	1	0.2809	0.0413	46.3667	<.0001	0.3976
capital_gain	1	0.000366	0.000050	53.8642	<.0001	1.4609
capital_loss	1	0.000660	0.000165	16.0365	<.0001	0.1520
hours_per_week	1	0.0208	0.00671	9.5698	0.0020	0.1495
dworkclass2	1	1.2277	0.6094	4.0593	0.0439	0.2308
dworkclass3	1	1.6779	0.6219	7.2784	0.0070	0.3086
dworkclass4	1	1.7319	0.5877	8.6846	0.0032	0.4428
dmarital2	1	-2.3294	0.2883	65.2966	<.0001	-0.4406
dmarital3	1	-2.5443	0.2690	89.4596	<.0001	-0.6497
dmarital4	1	-2.6342	0.5911	19.8608	<.0001	-0.2988
dmarital5	1	-4.4672	1.4967	8.9089	0.0028	-0.4305
doccup2	1	0.6150	0.2362	6.7807	0.0092	0.1064
doccup4	1	1.2094	0.2273	28.3113	<.0001	0.2302
doccup5	1	0.9855	0.2548	14.9574	0.0001	0.1938
dnativec3	1	1.1862	0.5524	4.6112	0.0318	0.0856

Figure 29. Final model with standardized estimates

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.027	1.012	1.043
edu_num	1.324	1.221	1.436
capital_gain	1.000	1.000	1.000
capital_loss	1.001	1.000	1.001
hours_per_week	1.021	1.008	1.034
dworkclass2	3.413	1.034	11.269
dworkclass3	5.354	1.582	18.117
dworkclass4	5.652	1.786	17.882
dmarital2	0.097	0.055	0.171
dmarital3	0.079	0.046	0.133
dmarital4	0.072	0.023	0.229
dmarital5	0.011	<0.001	0.216
doccup2	1.850	1.164	2.938
doccup4	3.351	2.147	5.232
doccup5	2.679	1.626	4.415
dnativec3	3.275	1.109	9.669

Figure 30. Final model odds ratios

ccup8	drace2	drace3	drace4	dnativec2	dnativec3	dnativec4	dnativec5	dincome	_FROM_	_INTO_	IP_0	IP_1	_LEVEL_	phat	lcl	ucl
0	0	0	0	0	0	0	0	.	.	1	0.01001	0.98999	1	0.98999	0.97183	0.99648
0	0	0	0	0	0	0	0	.	.	1	0.05411	0.94589	1	0.94589	0.87326	0.97795

Figure 31. Final model predictions

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	90.4	Somers' D	0.807
Percent Discordant	9.6	Gamma	0.807
Percent Tied	0.0	Tau-a	0.311
Pairs	427964	c	0.904

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.200	338	822	281	50	77.8	87.1	74.5	45.4	5.7
0.250	318	867	236	70	79.5	82.0	78.6	42.6	7.5
0.300	305	906	197	83	81.2	78.6	82.1	39.2	8.4
0.350	289	943	160	99	82.6	74.5	85.5	35.6	9.5
0.400	266	963	140	122	82.4	68.6	87.3	34.5	11.2
0.450	248	992	111	140	83.2	63.9	89.9	30.9	12.4
0.500	232	1011	92	156	83.4	59.8	91.7	28.4	13.4
0.550	214	1026	77	174	83.2	55.2	93.0	26.5	14.5
0.600	199	1046	57	189	83.5	51.3	94.8	22.3	15.3

Figure 32. Final model classification table

Fit Model on Training Set

The FREQ Procedure

Frequency	Table of dincome by pred_dis		
	pred_dis		
dincome	0	1	Total
0	281	91	372
1	15	110	125
Total	296	201	497

Figure 33. Final model test classification matrix

Appendix C - Danyang

Original Income_Train (not fitted and the ratio of train and test is 75 : 25)

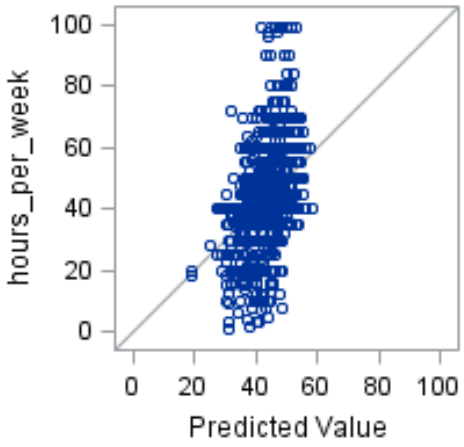
Test and Train Sets for Income

The SURVEYSELECT Procedure

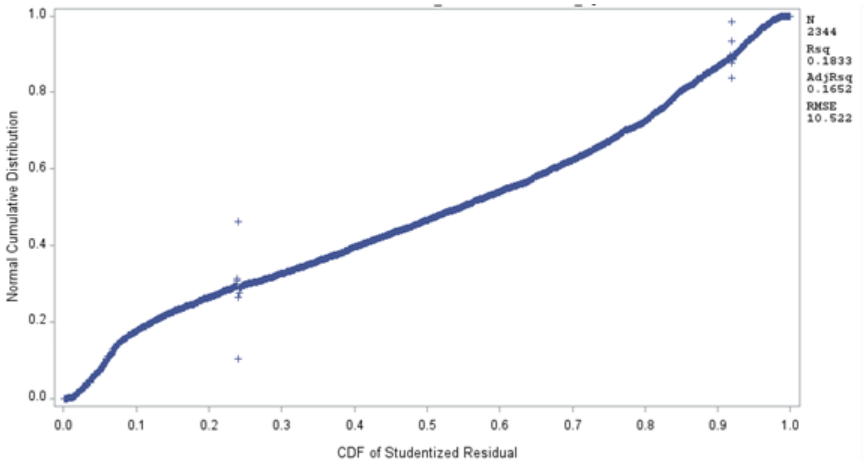
Selection Method	Simple Random Sampling

Input Data Set	INCOME
Random Number Seed	73051001
Sampling Rate	0.75
Sample Size	2344
Selection Probability	0.75008
Sampling Weight	0
Output Data Set	TRAIN

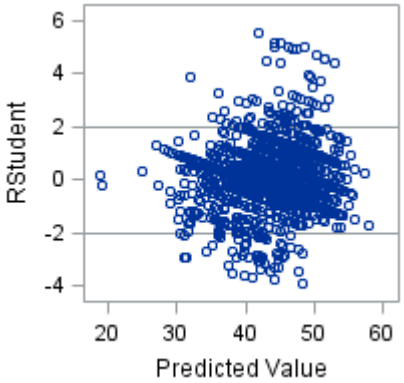
Linearity



Q-Q plot



Residuals



Before dropping insignificant variables

Test and Train Sets for Income

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2344
Number of Observations Used	2344

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	51	56967	1116.99896	10.09	<.0001
Error	2292	253771	110.72052		
Corrected Total	2343	310738			

Root MSE	10.52238	R-Square	0.1833
Dependent Mean	44.37671	Adj R-Sq	0.1652
Coeff Var	23.71150		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	21.87812	8.74518	2.50	0.0124	.	0
age	1	-0.15779	0.02304	-6.85	<.0001	0.61839	1.61709
capital_gain	1	0.00003142	0.00001222	2.57	0.0102	0.74232	1.34713
capital_loss	1	0.00007385	0.00025564	0.29	0.7727	0.79779	1.25346
workclass1	1	-0.58153	1.15945	-0.50	0.6160	0.91279	1.09555
workclass2	1	-0.09976	0.84095	-0.12	0.9056	0.80850	1.23685
workclass3	1	1.59027	0.88608	1.79	0.0728	0.84862	1.17838
workclass4	1	-1.91934	0.80831	-2.37	0.0177	0.82013	1.21932
workclass5	1	-1.24946	1.18487	-1.05	0.2918	0.90166	1.10906
~~education1~~	~~0~~	~~0~~	~~.~~	~~.~~	~~.~~	~~.~~	~~.~~
education2	1	26.88038	8.49727	3.16	0.0016	0.21972	4.55125
education3	1	20.64210	7.86305	2.63	0.0087	0.08605	11.62156
education4	1	27.75608	8.44117	3.29	0.0010	0.19490	5.13076
education5	1	23.78152	7.74920	3.07	0.0022	0.05202	19.22499
education6	1	18.47935	7.74983	2.38	0.0172	0.04931	20.27875
education7	1	22.49153	8.15299	2.76	0.0058	0.13952	7.16735

education8	1	22.94861	7.56272	3.03	0.0024	0.00449	222.57717
education9	1	28.60355	7.63801	3.74	0.0002	0.01502	66.56900
education10	1	22.89277	7.66528	2.99	0.0029	0.02629	38.03019
education11	1	23.33152	7.64237	3.05	0.0023	0.02101	47.59718
education12	1	22.35501	7.57241	2.95	0.0032	0.00574	174.18336
education13	1	23.93346	7.58592	3.15	0.0016	0.00437	228.94588
education14	1	24.88675	7.60642	3.27	0.0011	0.00793	126.04257
education15	1	28.18915	7.66315	3.68	0.0002	0.02411	41.47615
marital_status1	1	1.85357	3.64033	0.51	0.6107	0.01667	59.98052
marital_status2	0	0	.	.	.	.	.
marital_status3	1	-2.00347	0.95616	-2.10	0.0363	0.37638	2.65689
marital_status4	1	1.45124	1.65223	0.88	0.3798	0.82886	1.20648
marital_status5	1	-3.90696	1.81318	-2.15	0.0313	0.79784	1.25338
occupation1	1	-3.48752	10.63215	-0.33	0.7429	0.97988	1.02054
occupation2	1	2.51824	1.07110	2.35	0.0188	0.40792	2.45144
occupation3	1	5.20952	0.94985	5.48	<.0001	0.28864	3.46453
occupation4	1	8.88254	1.67633	5.30	<.0001	0.68506	1.45973
occupation5	1	-0.09636	1.76502	-0.05	0.9565	0.74081	1.34987
occupation6	1	0.61534	1.37481	0.45	0.6545	0.64267	1.55600

occupation7	1	-2.33150	1.51660	-1.54	0.1244	0.69917	1.43026
occupation8	1	6.40894	10.63370	0.60	0.5468	0.97959	1.02084
occupation9	1	1.92050	0.99725	1.93	0.0543	0.27364	3.65449
occupation10	1	2.20525	1.82561	1.21	0.2272	0.75269	1.32856
occupation11	1	4.13984	1.01254	4.09	<.0001	0.39607	2.52483
occupation12	1	-0.06143	1.50276	-0.04	0.9674	0.74255	1.34671
occupation13	1	5.67351	1.37173	4.14	<.0001	0.62058	1.61140
relationship1	1	2.69122	3.62575	0.74	0.4580	0.02231	44.82751
relationship2	1	-2.27688	3.89852	-0.58	0.5593	0.33427	2.99156
relationship3	1	-4.04883	3.51302	-1.15	0.2492	0.08708	11.48340
relationship4	1	0.95248	3.78432	0.25	0.8013	0.06163	16.22672
relationship5	1	-5.94087	1.28280	-4.63	<.0001	0.42216	2.36876
race1	0	0	.	.	.	.	.
race2	1	1.07754	2.08869	0.52	0.6060	0.19025	5.25613
race3	1	0.74538	1.89895	0.39	0.6947	0.18667	5.35695
race4	1	1.00654	5.71503	0.18	0.8602	0.84893	1.17795
sex1	1	-6.88621	1.97886	-3.48	0.0005	0.07046	14.19254
income1	1	1.82436	0.64715	2.82	0.0049	0.50896	1.96480
sex_income	1	0.70667	1.21646	0.58	0.5614	0.33385	2.99540
sex_age	1	0.10556	0.04700	2.25	0.0248	0.06899	14.49569

Income_Train_new

Test and Train Sets for Income

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2344
Number of Observations Used	2344

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	38833	2588.89863	22.17	<.0001
Error	2328	271905	116.79763		
Corrected Total	2343	310738			

Root MSE	10.80730	R-Square	0.1250
Dependent Mean	44.37671	Adj R-Sq	0.1193
Coeff Var	24.35353		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45.27415	1.01680	44.53	<.0001
age	1	-0.10141	0.02070	-4.90	<.0001
capital_gain	1	0.00005237	0.00001127	4.65	<.0001
workclass4	1	-1.30403	0.79370	-1.64	0.1005
education2	1	4.21031	4.11605	1.02	0.3065
education4	1	3.42535	3.84014	0.89	0.3725
education7	1	-1.96961	3.13882	-0.63	0.5304
marital_status3	1	-3.33937	0.68021	-4.91	<.0001
marital_status5	1	-5.10817	1.70255	-3.00	0.0027
occupation2	1	0.82578	0.75435	1.09	0.2738
occupation3	1	4.04963	0.57958	6.99	<.0001
occupation4	1	6.87139	1.52040	4.52	<.0001
occupation11	1	2.54837	0.69987	3.64	0.0003
occupation13	1	3.47084	1.14570	3.03	0.0025
relationship5	1	-8.16291	0.87973	-9.28	<.0001
income1	1	3.87902	0.53285	7.28	<.0001

Income_Train_New (after dropping insignificant variables)

Test and Train Sets for Income

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2344
Number of Observations Used	2344

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	38142	3814.15048	32.64	<.0001
Error	2333	272597	116.84392		
Corrected Total	2343	310738			

Root MSE	10.80944	R-Square	0.1227
Dependent Mean	44.37671	Adj R-Sq	0.1190
Coeff Var	24.35836		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45.58567	0.98841	46.12	<.0001
age	1	-0.10442	0.02057	-5.08	<.0001
capital_gain	1	0.00005109	0.00001125	4.54	<.0001
marital_status3	1	-3.45867	0.67488	-5.12	<.0001
marital_status5	1	-5.10093	1.69891	-3.00	0.0027
occupation3	1	3.88150	0.55915	6.94	<.0001
occupation4	1	6.01072	1.44548	4.16	<.0001
occupation11	1	2.30344	0.68066	3.38	0.0007
occupation13	1	3.31412	1.13128	2.93	0.0034
relationship5	1	-8.29624	0.87210	-9.51	<.0001
income1	1	3.79095	0.53049	7.15	<.0001

The final model(Dropping 238 outliers and influential points):

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2106
Number of Observations Used	2106

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	19859	1985.94011	41.82	<.0001
Error	2095	99488	47.48808		
Corrected Total	2105	119347			

Root MSE	6.89116	R-Square	0.1664
Dependent Mean	44.17711	Adj R-Sq	0.1624
Coeff Var	15.59894		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	43.47034	0.68102	63.83	<.0001
age	1	-0.04900	0.01444	-3.39	0.0007
capital_gain	1	0.00004216	0.00000837	5.04	<.0001
marital_status3	1	-1.63970	0.45422	-3.61	0.0003
marital_status5	1	-3.61086	1.38168	-2.61	0.0090
occupation3	1	3.91889	0.37362	10.49	<.0001
occupation4	1	3.00390	1.30252	2.31	0.0212
occupation11	1	2.36954	0.45711	5.18	<.0001
occupation13	1	1.69361	0.79774	2.12	0.0339
relationship5	1	-6.24055	0.59910	-10.42	<.0001
income1	1	2.67823	0.35649	7.51	<.0001

Test performance of test set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	3125
Number of Observations Used	2344
Number of Observations with Missing Values	781

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	38142	3814.15048	32.64	<.0001
Error	2333	272597	116.84392		
Corrected Total	2343	310738			

Root MSE	10.80944	R-Square	0.1227
Dependent Mean	44.37671	Adj R-Sq	0.1190
Coeff Var	24.35836		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45.58567	0.98841	46.12	<.0001
age	1	-0.10442	0.02057	-5.08	<.0001
capital_gain	1	0.00005109	0.00001125	4.54	<.0001
marital_status3	1	-3.45867	0.67488	-5.12	<.0001
marital_status5	1	-5.10093	1.69891	-3.00	0.0027
occupation3	1	3.88150	0.55915	6.94	<.0001
occupation4	1	6.01072	1.44548	4.16	<.0001
occupation11	1	2.30344	0.68066	3.38	0.0007
occupation13	1	3.31412	1.13128	2.93	0.0034
relationship5	1	-8.29624	0.87210	-9.51	<.0001
income1	1	3.79095	0.53049	7.15	<.0001

Obs	_TYPE_	_FREQ_	rmse	mae	mape
1	0	781	10.2792	7.24677	0.21709

The CORR Procedure

2 Variables:	hours_per_week yhat

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
hours_per_week	781	44.79001	11.17171	34981	2.00000	99.00000	
yhat	781	44.38713	3.94177	34666	31.54650	55.23432	Predicted Value of new_y

Pearson Correlation Coefficients, N = 781 Prob > |r| under H0: Rho=0		
	hours_per_week	yhat
hours_per_week	1.00000	0.39379 <.0001
yhat Predicted Value of new_y	0.39379 <.0001	1.00000

5-folder Cross validation: stepwise and backward

5-fold crossvalidation + 25% testing set

The GLMSELECT Procedure

Data Set	WORK.INCOME
Dependent Variable	hours_per_week
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	658332001

Number of Observations Read	3125
Number of Observations Used	3125
Number of Observations Used for Training	2315
Number of Observations Used for Testing	810

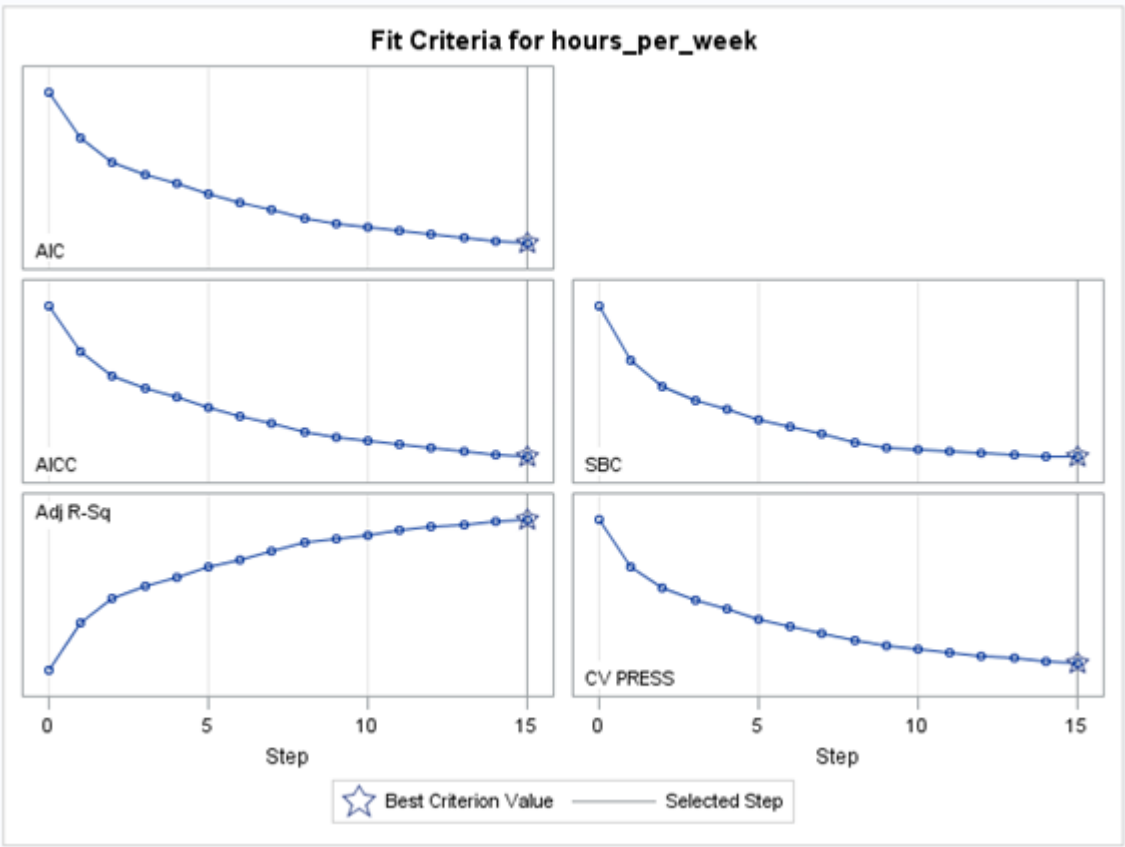
Dimensions	
Number of Effects	53
Number of Parameters	53

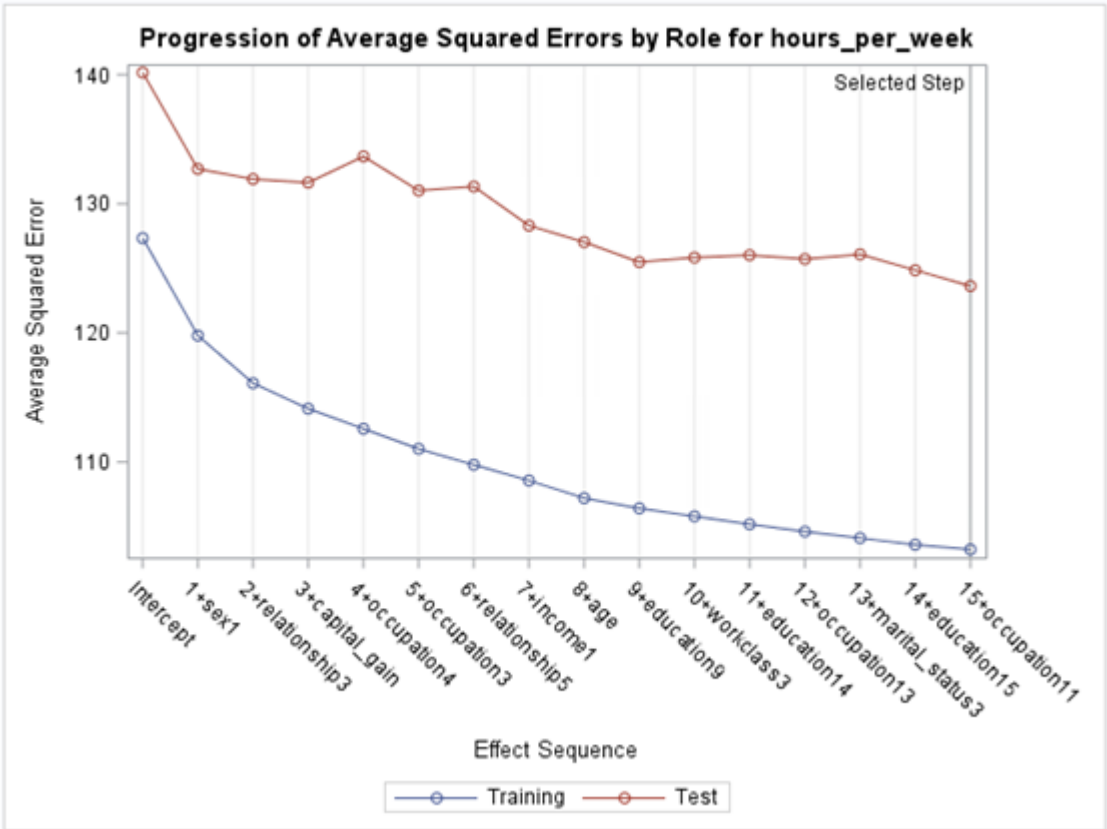
The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0	Intercept		1	11228.1105	127.3334	140.1525	295631.806
1	sex1		2	11094.6693	119.7996	132.6934	278409.283
2	relationship3		3	11030.0422	116.1123	131.9086	270247.474
3	capital_gain		4	10998.0566	114.1364	131.6402	265901.506
4	occupation4		5	10974.1474	112.5863	133.6751	262454.761
5	occupation3		6	10949.5103	111.0223	131.0256	258988.287
6	relationship5		7	10931.3955	109.7889	131.3353	256494.305
7	income1		8	10913.3078	108.5705	128.3160	253727.039
8	age		9	10891.9359	107.2134	127.0380	250821.683
9	education9		10	10882.3872	106.4154	125.4804	249202.575
10	workclass3		11	10876.7019	105.7997	125.8440	247820.760
11	education14		12	10870.9619	105.1851	126.0290	246481.504
12	occupation13		13	10866.2148	104.6189	125.7253	245397.998
13	marital_status3		14	10862.6814	104.1104	126.0872	244439.011
14	education15		15	10859.2407	103.6085	124.8591	243469.724
15	occupation11		16	10858.7268*	103.2394	123.6323	242848.212*

* Optimal Value of Criterion


Fit Criteria for hours_per_week
AIC
AICC
SBC
Adj R-Sq
CV PRESS
0
5
10
15
Step
Best Criterion Value
Selected Step


Progression of Average Squared Errors by Role for hours_per_week
Selected Step
Average Squared Error
140
130
120
110
Intercept
1+sex1
2+relationship3
3+capital_gain
4+occupation4
5+occupation3
6+relationship5
7+income1
8+age
9+education9
10+workclass3
11+education14
12+occupation13
13+marital_status3
14+education15
15+occupation11
Effect Sequence
Training
Test

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	15	55778	3718.51409	35.77
Error	2299	238999	103.95790	
Corrected Total	2314	294777		

Root MSE	10.19597
Dependent Mean	44.29028
R-Square	0.1892
Adj R-Sq	0.1839
AIC	13084
AICC	13084
SBC	10859
ASE (Train)	103.23940
ASE (Test)	123.63234
CV PRESS	242848

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1852	463	45357.193
2	1852	463	48354.479
3	1852	463	47611.465
4	1852	463	47064.109
5	1852	463	54460.966
Total			242848.212

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	48.116227	0.979089	49.14	4.77E+01	4.87E+01	4.84E+01	4.78E+01	4.80E+01
age	1	-0.145435	0.020028	-7.26	-1.30E-01	-1.62E-01	-1.58E-01	-1.31E-01	-1.46E-01
capital_gain	1	0.000042820	0.000011150	3.84	4.06E-05	3.21E-05	4.43E-05	5.14E-05	4.56E-05
workclass3	1	3.079154	0.862884	3.57	2.80E+00	3.18E+00	3.26E+00	3.52E+00	2.58E+00
education9	1	5.038957	0.997388	5.05	5.30E+00	5.23E+00	5.63E+00	4.70E+00	4.32E+00
education14	1	2.970630	0.687105	4.32	3.23E+00	2.90E+00	3.04E+00	3.12E+00	2.57E+00
education15	1	4.675465	1.310796	3.57	4.87E+00	4.46E+00	4.95E+00	5.08E+00	4.17E+00
marital_status3	1	-2.496087	0.716226	-3.49	-2.52E+00	-2.77E+00	-2.85E+00	-2.09E+00	-2.32E+00
occupation3	1	3.459690	0.557902	6.20	3.33E+00	3.75E+00	3.42E+00	3.59E+00	3.21E+00
occupation4	1	10.748843	1.382302	7.78	1.02E+01	1.08E+01	1.07E+01	1.07E+01	1.12E+01
occupation11	1	1.871456	0.652800	2.87	1.92E+00	2.34E+00	1.62E+00	1.56E+00	1.93E+00
occupation13	1	4.247796	1.079936	3.93	4.06E+00	5.39E+00	4.53E+00	3.82E+00	3.56E+00
relationship3	1	-8.953471	1.178913	-7.59	-9.64E+00	-8.32E+00	-7.71E+00	-9.82E+00	-9.23E+00
relationship5	1	-7.136155	1.038351	-6.87	-6.72E+00	-7.23E+00	-6.27E+00	-7.91E+00	-7.50E+00
sex1	1	-2.344300	0.671881	-3.49	-2.48E+00	-2.09E+00	-2.54E+00	-2.09E+00	-2.57E+00
income1	1	2.043855	0.517045	3.95	1.74E+00	2.21E+00	2.21E+00	1.82E+00	2.21E+00

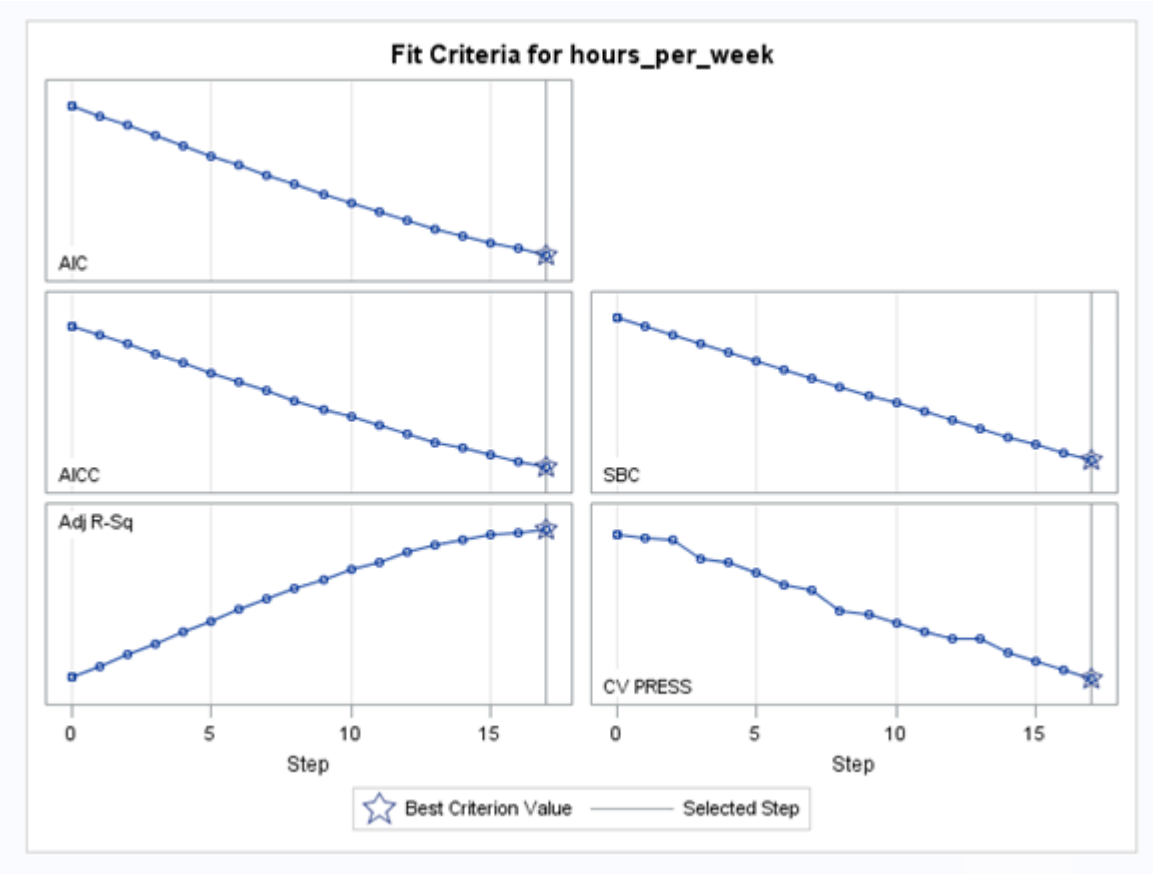
5-fold crossvalidation + 25% testing set

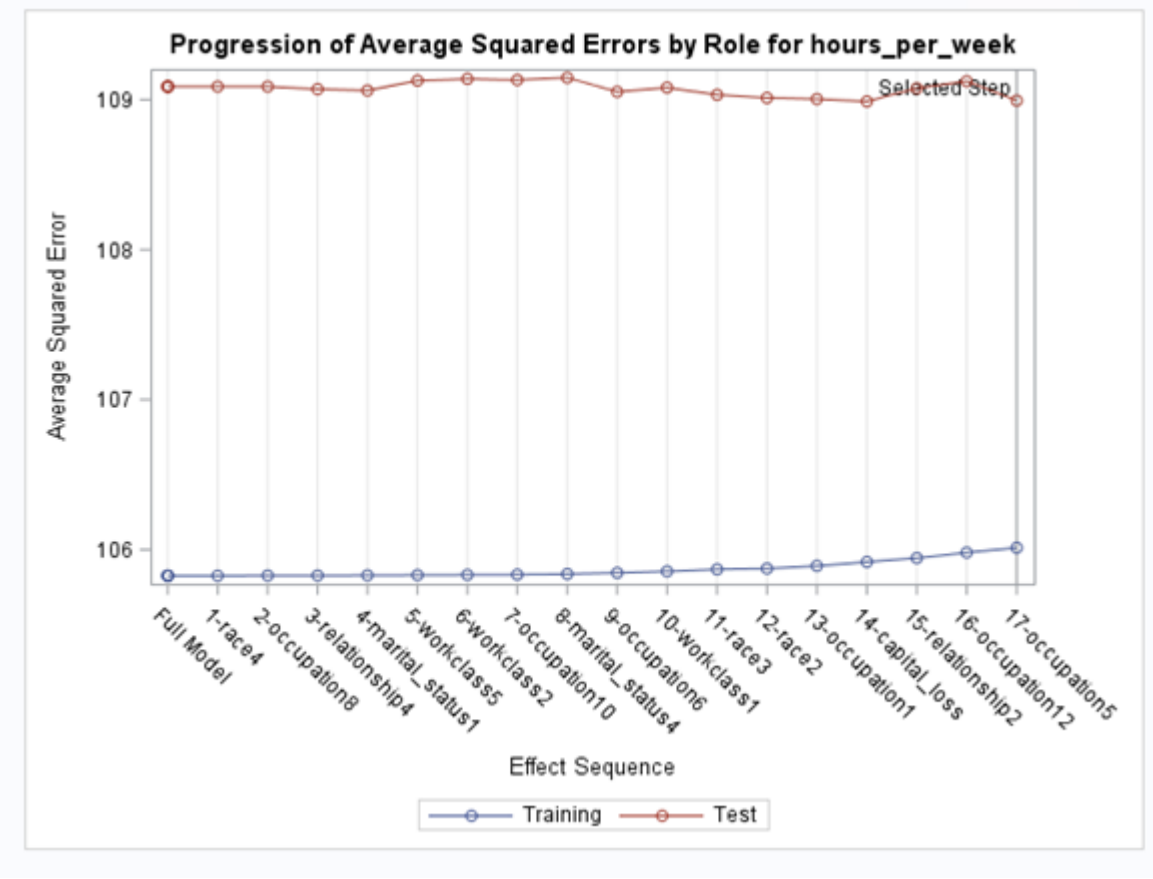
The GLMSELECT Procedure

Data Set	WORK.INCOME
Dependent Variable	hours_per_week
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	312677001

Number of Observations Read	3125
Number of Observations Used	3125
Number of Observations Used for Training	2348
Number of Observations Used for Testing	777

Dimensions	
Number of Effects	53
Number of Parameters	53


Fit Criteria for hours_per_week
AIC
AICC
SBC
Adj R-Sq
CV PRESS
0
5
10
15
Step
0
5
10
15
Step
Best Criterion Value
Selected Step


Progression of Average Squared Errors by Role for hours_per_week
Selected Step
Average Squared Error
109
108
107
106
Full Model
1-race4
2-occupation8
3-relationship4
4-marital_status1
5-workclass5
6-workclass2
7-occupation10
8-marital_status4
9-occupation6
10-workclass1
11-race3
12-race2
13-occupation1
14-capital_loss
15-relationship2
16-occupation12
17-occupation5
Effect Sequence
Training
Test

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	32	60962	1905.07755	17.72
Error	2315	248918	107.52406	
Corrected Total	2347	309881		

Root MSE	10.36938
Dependent Mean	44.57240
R-Square	0.1967
Adj R-Sq	0.1856
AIC	13366
AICC	13367
SBC	11206
ASE (Train)	106.01287
ASE (Test)	108.99427
CV PRESS	258504

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1878	470	53600.071
2	1878	470	48529.639
3	1878	470	48871.891
4	1879	469	50170.953
5	1879	469	57331.264
Total			258503.819

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	23.915117	7.542937	3.17	2.25E+01	2.61E+01	2.51E+01	2.36E+01	22.349258
age	1	-0.144713	0.020329	-7.12	-1.40E-01	-1.45E-01	-1.50E-01	-1.45E-01	-0.143852
capital_gain	1	0.000035642	0.000011421	3.12	3.75E-05	4.02E-05	3.73E-05	3.54E-05	0.000028
workclass3	1	1.909105	0.868868	2.20	2.32E+00	1.78E+00	1.11E+00	2.51E+00	1.927766
workclass4	1	-0.788331	0.778910	-1.01	-8.13E-01	-8.88E-01	-8.82E-01	-3.73E-01	-1.003699
education2	1	27.890136	8.236484	3.39	3.16E+01	2.34E+01	2.71E+01	2.87E+01	28.931952
education3	1	22.049711	7.721184	2.86	2.41E+01	1.93E+01	2.19E+01	1.95E+01	25.765143
education4	1	24.485098	8.130661	3.01	2.61E+01	2.20E+01	2.22E+01	2.95E+01	23.079539
education5	1	24.225583	7.650200	3.17	2.48E+01	2.28E+01	2.37E+01	2.56E+01	23.464952
education6	1	21.419501	7.599489	2.82	2.30E+01	1.85E+01	2.05E+01	2.16E+01	23.273538
education7	1	22.881324	7.916230	2.89	2.40E+01	2.17E+01	2.28E+01	2.00E+01	24.466538
education8	1	23.047169	7.435566	3.10	2.39E+01	2.11E+01	2.25E+01	2.31E+01	24.722434
education9	1	27.079141	7.518032	3.60	2.92E+01	2.46E+01	2.56E+01	2.77E+01	28.410195
education10	1	22.819742	7.534559	3.03	2.39E+01	2.10E+01	2.21E+01	2.24E+01	24.814109
education11	1	23.672095	7.509096	3.15	2.52E+01	2.20E+01	2.21E+01	2.41E+01	24.882476
education12	1	22.086924	7.445177	2.97	2.37E+01	1.97E+01	2.11E+01	2.23E+01	23.612717
education13	1	23.372269	7.457537	3.13	2.46E+01	2.16E+01	2.23E+01	2.36E+01	24.732889
education14	1	25.746166	7.474024	3.44	2.77E+01	2.36E+01	2.48E+01	2.53E+01	27.448500
education15	1	26.691583	7.537458	3.54	2.71E+01	2.64E+01	2.50E+01	2.63E+01	28.785859
marital_status3	1	-3.231566	0.855550	-3.78	-2.75E+00	-3.36E+00	-4.03E+00	-3.38E+00	-2.687046
marital_status5	1	-4.084830	1.705240	-2.40	-3.89E+00	-3.95E+00	-5.58E+00	-4.14E+00	-2.852411
occupation2	1	1.688785	0.822394	2.05	1.68E+00	1.87E+00	1.90E+00	1.50E+00	1.486736
occupation3	1	4.518608	0.729158	6.20	4.14E+00	4.14E+00	5.09E+00	4.67E+00	4.426322
occupation4	1	12.733686	1.638217	7.77	1.18E+01	1.32E+01	1.55E+01	1.04E+01	12.780311
occupation7	1	-2.134623	1.380623	-1.55	-1.33E+00	-2.21E+00	-2.18E+00	-1.69E+00	-3.308107
occupation9	1	1.609437	0.783990	2.05	9.41E-01	1.48E+00	1.86E+00	1.95E+00	1.721221
occupation11	1	3.125214	0.785487	3.98	3.22E+00	2.62E+00	3.61E+00	2.90E+00	3.172748
occupation13	1	4.957743	1.144299	4.33	4.48E+00	4.51E+00	4.15E+00	5.08E+00	6.574884
relationship1	1	1.828088	0.725229	2.52	1.86E+00	1.55E+00	2.15E+00	2.19E+00	1.356797
relationship3	1	-5.876703	1.295173	-4.54	-6.26E+00	-6.08E+00	-5.07E+00	-4.50E+00	-7.300939
relationship5	1	-5.693377	1.079256	-5.28	-6.18E+00	-5.04E+00	-6.21E+00	-4.96E+00	-6.170164
sex1	1	-2.600087	0.728631	-3.57	-2.69E+00	-2.51E+00	-2.46E+00	-3.42E+00	-1.926783
income1	1	2.761569	0.550681	5.01	2.94E+00	2.60E+00	2.72E+00	2.90E+00	2.696336

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	34.7799	0.6502	33.5048	36.0549	21.2056	48.3541	.
2	.	42.0004	0.3629	41.2887	42.7121	28.4674	55.5333	.
3	40	40.8244	0.3689	40.1009	41.5479	27.2908	54.3580	-0.8244
4	40	41.4614	0.3224	40.8292	42.0936	27.9324	54.9904	-1.4614
5	45	41.8044	0.3403	41.1371	42.4716	28.2737	55.3351	3.1956
6	40	41.6084	0.3263	40.9684	42.2484	28.0790	55.1378	-1.6084
7	35	43.9436	0.2677	43.4187	44.4686	30.4192	57.4681	-8.9436
8	55	48.4995	0.3873	47.7399	49.2591	34.9639	62.0351	6.5005
9	40	43.8946	0.2685	43.3680	44.4213	30.3701	57.4191	-3.8946

Models from reference:

Forward

Scatterplot

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2344
Number of Observations Used	2344

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	24	54627	2276.12590	20.61	<.0001
Error	2319	256111	110.44043		
Corrected Total	2343	310738			

Root MSE	10.50906	R-Square	0.1758
Dependent Mean	44.37671	Adj R-Sq	0.1673
Coeff Var	23.68149		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	1	47.48723	1.15382	41.16	<.0001	187069
age	1	-0.16131	0.02262	-7.13	<.0001	5614.38463
capital_gain	1	0.00003200	0.00001133	2.82	0.0048	881.04707
workclass3	1	1.68887	0.87490	1.93	0.0537	411.52709
workclass4	1	-1.90725	0.79252	-2.41	0.0162	639.62558
education6	1	-4.34738	1.74789	-2.49	0.0129	683.20863
education9	1	5.67139	1.14239	4.96	<.0001	2721.94544
education13	1	1.10654	0.59758	1.85	0.0642	378.67063
education14	1	2.03512	0.80933	2.51	0.0120	698.33064
education15	1	5.11070	1.31031	3.90	<.0001	1680.11279
marital_status3	1	-2.68540	0.86532	-3.10	0.0019	1063.62203
marital_status5	1	-4.30060	1.75572	-2.45	0.0144	662.63438
occupation2	1	2.29697	0.83858	2.74	0.0062	828.62120
occupation3	1	4.79617	0.72945	6.58	<.0001	4774.54542
occupation4	1	8.57421	1.54273	5.56	<.0001	3411.44719
occupation7	1	-2.58151	1.35525	-1.90	0.0569	400.71552
occupation9	1	1.55350	0.79113	1.96	0.0497	425.85038
occupation11	1	3.80751	0.79630	4.78	<.0001	2524.96706
occupation13	1	4.91551	1.18263	4.16	<.0001	1907.96755
relationship1	1	1.63591	0.75128	2.18	0.0295	523.65724
relationship3	1	-5.09735	1.26397	-4.03	<.0001	1796.14201
relationship5	1	-5.37829	1.08164	-4.97	<.0001	2730.57058
sex1	1	-6.58454	1.94127	-3.39	0.0007	1270.60151
income1	1	2.14097	0.55629	3.85	0.0001	1635.88676
sex_age	1	0.09648	0.04466	2.16	0.0308	515.45188

Backward:

Scatterplot

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2344
Number of Observations Used	2344

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	50200	3346.65101	29.90	<.0001
Error	2328	260539	111.91521		
Corrected Total	2343	310738			

Root MSE	10.57900	R-Square	0.1615
Dependent Mean	44.37671	Adj R-Sq	0.1561
Coeff Var	23.83908		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	1	45.58096	0.98379	46.33	<.0001	240242
age	1	-0.12861	0.01965	-6.54	<.0001	4791.45406
capital_gain	1	0.00003686	0.00001130	3.26	0.0011	1191.16095
workclass4	1	-2.26048	0.78574	-2.88	0.0041	926.24857
education9	1	4.85377	1.05914	4.58	<.0001	2350.40302
education15	1	4.18510	1.25229	3.34	0.0008	1249.93806
occupation2	1	2.59188	0.81699	3.17	0.0015	1126.38771
occupation3	1	5.89719	0.67984	8.67	<.0001	8420.98512
occupation4	1	9.13626	1.53343	5.96	<.0001	3972.79202
occupation9	1	2.74683	0.72173	3.81	0.0001	1621.09741
occupation11	1	4.58485	0.76953	5.96	<.0001	3972.74240
occupation13	1	5.04339	1.17410	4.30	<.0001	2065.03511
relationship3	1	-7.51234	1.11297	-6.75	<.0001	5098.88785
relationship5	1	-5.07958	1.01848	-4.99	<.0001	2783.83844
sex1	1	-2.99915	0.67255	-4.46	<.0001	2225.54081
income1	1	2.65012	0.53654	4.94	<.0001	2730.32858

Stepwise:

Scatterplot

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2344
Number of Observations Used	2344

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	49314	3522.45374	31.38	<.0001
Error	2329	261424	112.24732		
Corrected Total	2343	310738			

Root MSE	10.59468	R-Square	0.1587
Dependent Mean	44.37671	Adj R-Sq	0.1536
Coeff Var	23.87443		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	1	47.27537	0.93502	50.56	<.0001	286947
age	1	-0.12963	0.01969	-6.58	<.0001	4865.91705
capital_gain	1	0.00003729	0.00001131	3.30	0.0010	1219.40366
workclass4	1	-2.02433	0.78482	-2.58	0.0100	746.78268
education9	1	5.58808	1.00602	5.55	<.0001	3463.27342
education15	1	4.79173	1.22674	3.91	<.0001	1712.59814
occupation3	1	4.05708	0.55820	7.27	<.0001	5929.49225
occupation4	1	7.25127	1.48778	4.87	<.0001	2666.42372
occupation7	1	-4.18099	1.31446	-3.18	0.0015	1135.63641
occupation11	1	2.78479	0.67864	4.10	<.0001	1890.09465
occupation13	1	3.26658	1.11847	2.92	0.0035	957.43843
relationship3	1	-7.53477	1.11534	-6.76	<.0001	5122.75579
relationship5	1	-5.14392	1.01988	-5.04	<.0001	2855.38682
sex1	1	-3.01399	0.66020	-4.57	<.0001	2339.38048
income1	1	2.80977	0.52974	5.30	<.0001	3157.87158

Adjrsq:

Scatterplot

The REG Procedure
Model: MODEL1
Dependent Variable: hours_per_week

Number of Observations Read	2344
Number of Observations Used	2344

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	56527	1615.06205	14.66	<.0001
Error	2308	254211	110.14350		
Corrected Total	2343	310738			

Root MSE	10.49493	R-Square	0.1819
Dependent Mean	44.37671	Adj R-Sq	0.1695
Coeff Var	23.64963		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	1	24.72155	7.66000	3.23	0.0013	1147.23509
age	1	-0.15925	0.02272	-7.01	<.0001	5408.97971
capital_gain	1	0.00003116	0.00001133	2.75	0.0060	833.74364
workclass3	1	1.69785	0.87531	1.94	0.0525	414.40913
workclass4	1	-1.79846	0.79321	-2.27	0.0235	566.22462
education2	1	26.63010	8.46432	3.15	0.0017	1090.23662
education3	1	20.48278	7.83305	2.61	0.0090	753.13806
education4	1	27.54598	8.40192	3.28	0.0011	1183.90850
education5	1	23.66552	7.71979	3.07	0.0022	1035.09433
education6	1	18.24416	7.71672	2.36	0.0181	615.66036
education7	1	22.16590	8.12134	2.73	0.0064	820.49217
education8	1	22.69838	7.53408	3.01	0.0026	999.74172
education9	1	28.18229	7.60925	3.70	0.0002	1510.87407
education10	1	22.54411	7.63523	2.95	0.0032	960.24217
education11	1	23.05067	7.61069	3.03	0.0025	1010.36417
education12	1	22.08802	7.54389	2.93	0.0034	944.23896

education13	1	23.61287	7.55666	3.12	0.0018	1075.46575
education14	1	24.53660	7.57494	3.24	0.0012	1155.65862
education15	1	27.65144	7.63324	3.62	0.0003	1445.36348
marital_status3	1	-2.42509	0.88433	-2.74	0.0061	828.29051
marital_status5	1	-4.31371	1.76193	-2.45	0.0144	660.20944
occupation2	1	2.45083	0.85946	2.85	0.0044	895.63369
occupation3	1	5.12867	0.75218	6.82	<.0001	5120.62982
occupation4	1	8.72021	1.55351	5.61	<.0001	3470.44176
occupation7	1	-2.43574	1.38129	-1.76	0.0780	342.49171
occupation9	1	1.82969	0.81122	2.26	0.0242	560.32109
occupation10	1	1.97160	1.66173	1.19	0.2356	155.05127
occupation11	1	4.07936	0.81653	5.00	<.0001	2749.16289
occupation13	1	5.60827	1.21007	4.63	<.0001	2365.89587
relationship1	1	1.40364	0.77032	1.82	0.0686	365.70257
relationship2	1	-3.22120	2.36408	-1.36	0.1732	204.48915
relationship3	1	-5.29551	1.28114	-4.13	<.0001	1881.83903
relationship5	1	-5.37948	1.08708	-4.95	<.0001	2697.21420
sex1	1	-6.73872	1.95137	-3.45	0.0006	1313.51166
income1	1	2.01073	0.56293	3.57	0.0004	1405.25652

sex_age	1	0.10263	0.04501	2.28	0.0227	572.78009

Appendix D - Persid

Data obtained was cleaned from the rows with missing values and imported in SAS as .csv file (data1.csv). A sample of 2000 data points were selected by simple random sampling technique. The data created was then exported in a second .csv file (Random Sample.cvs) and was used for the regression analysis.

```
proc import datafile="S:\Final_Project\data1.csv" out=censusIncome replace;
delimiter=',';
getnames=yes;
Run;

*Randomly Selected Sample;
proc surveyselect data=censusIncome
method=srs n=2000 out=censusSample;
run;

PROC PRINT data=censusSample;
RUN;
```

a

workclass	Dummy variables
Private	Reference Level
Self-emp-not-inc	workclass1 = 1
Self-emp-inc	workclass2 = 1
Federal-gov	workclass3 = 1
Local-gov	workclass4 = 1
State-gov	workclass5 = 1
Without-pay	workclass6 = 1
Never-worked	workclass7 = 1

b

education	education_cat	Dummy variable
Preschool	No_Education	Reference Level
1st-4th		
5th-6th		
7th-8th		
9th		
10th		
11th		
12th		
HS-grad	High_School	edLevel1 = 1
Some-college	Post_Secondary	edLevel2 = 1
Assoc-acdm		
Assoc-voc		
Prof-school		
Bachelors	Bachelors	edlevel3 = 1
Masters	Masters	edLevel4 = 1
Doctorate	Doctorate	edLevel5 = 1

c

marital_status	civil-status	Dummy variable
Married-civ-spouse	Married	Reference Level
Married-AF-spouse		
Divorced	Not-Married	maritalStatus1 = 1
Widowed		
Separated		
Married-spouse-absent		
Never-married		

d

occupation	occupation_cat	Dummy variable
Adm-clerical	Adm-clerical	Reference Level
Craft-repair	Working_class	occupation1 = 1
Farming-fishing		
Handlers-cleaners		
Machine-op-inspct		
Transport-moving		
Priv-house-serv	Services	occupation2 = 1
Other-service		
Protective-serv		
Tech-support	Professional	occupation3 = 1
Prof-specialty		
Sales	Sales	occupation4 = 1
Exec-managerial	Exec-managerial	occupation5 = 1
Armed-Forces	Armed-Forces	occupation6 = 1

e

relationship	Dummy variable
Not-in-family	Reference Level
Wife	relationship1 = 1
Own-child	relationship2 = 1
Husband	relationship3 = 1
Other-relative	relationship4 = 1
Unmarried	relationship5 = 1

f

race	Dummy variable
White	Reference Level
Asian-Pac-Islander	race1 = 1
Amer-Indian-Eskimo	race2 = 1
Black	race3 = 1
Other	race4 = 1

g

gender	Dummy variable
Female	Reference Level
Male	sex1 = 1

h

native_country	Continent	Dummy variable
USA	USA	Reference Level
England	Europe	continent1 = 1
Canada		
Germany		
.		
.		
.		
Puerto-Rico	Latin_America	continent2 = 1
Cuba		
Honduras		
Jamaica		
Mexico		
.		
.		
.		
Cambodia	Asia	continent3 = 1
India		
Japan		
.		
.		
.		

i

class	Dummy variable
<=50K	income = 0
>50K	income = 1

Table 1. Grouping and Coding of the Qualitative Variables

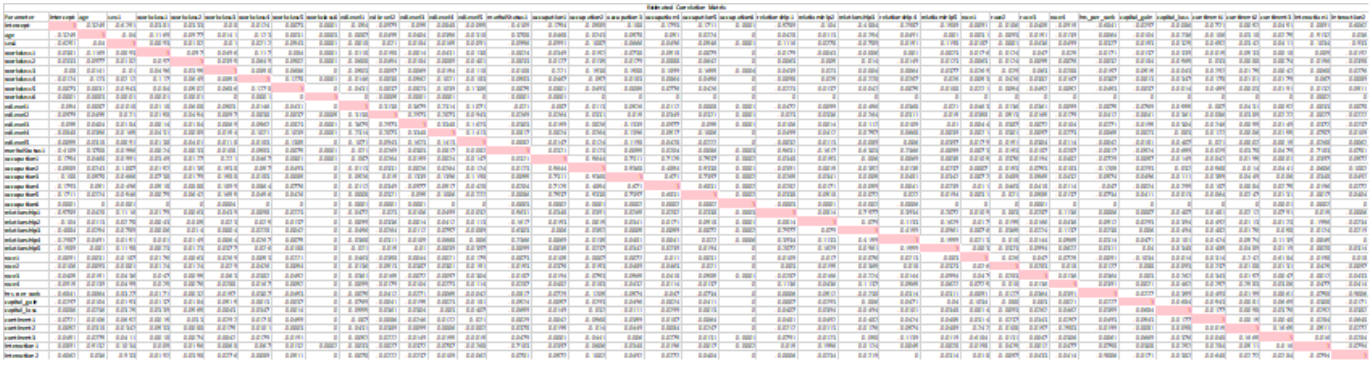
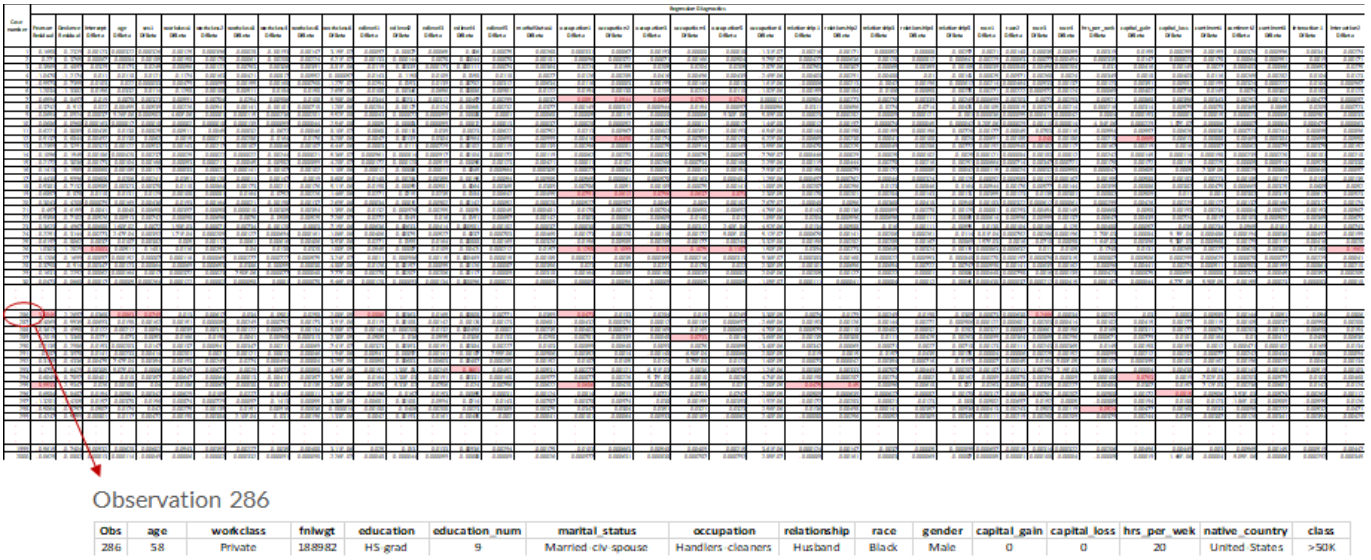


Table 2. Estimated Correlation Matrix.



Observation 286

Obs	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	gender	capital_gain	capital_loss	hrs_per_wek	native_country	class
286	58	Private	188982	HS-grad	9	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	20	United-States	>50K

Table 3. Influential Points and Outliers.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1333.740	815.822
SC	1338.830	866.723
-2 Log L	1331.740	795.822

R-Square	0.3602	Max-rescaled R-Square	0.5373

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	535.9186	9	<.0001
Score	398.4208	9	<.0001
Wald	235.4524	9	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
39.5298	27	0.0567

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-1.6260	0.3992	16.5877	<.0001	
edLevel1	1	-0.8439	0.2087	16.3565	<.0001	-0.2183
maritalStatus1	1	-4.6052	0.6106	56.8755	<.0001	-1.2651
occupation1	1	-1.2038	0.2030	35.1824	<.0001	-0.3160
occupation2	1	-1.8575	0.4210	19.4693	<.0001	-0.3473
hrs_per_wek	1	0.0466	0.00868	28.7574	<.0001	0.2960
capital_gain	1	0.000387	0.000056	47.9829	<.0001	1.4596
capital_loss	1	0.000618	0.000189	10.7262	0.0011	0.1360
continent2	1	-2.5953	1.0558	6.0426	0.0140	-0.2748
interaction1	1	0.0539	0.0139	15.0218	0.0001	0.5786

Table 4. Stepwise Selection Procedure Summary

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1333.740	804.977
SC	1338.830	881.328
-2 Log L	1331.740	774.977

R-Square	0.3712	Max-rescaled R-Square	0.5538

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	556.7638	14	<.0001
Score	418.1033	14	<.0001
Wald	238.2962	14	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
21.0835	22	0.5156

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-2.9433	0.5293	30.9261	<.0001	
sex1	1	0.9673	0.3590	7.2616	0.0070	0.2524
workclass3	1	0.9373	0.4705	3.9689	0.0463	0.0870
edLevel1	1	-0.7879	0.2144	13.5001	0.0002	-0.2038
maritalStatus1	1	-4.4091	0.6359	48.0758	<.0001	-1.2112
occupation1	1	-0.7531	0.2737	7.5702	0.0059	-0.1977
occupation2	1	-1.3206	0.4565	8.3708	0.0038	-0.2470
occupation3	1	0.7088	0.2896	5.9898	0.0144	0.1488
occupation5	1	0.7557	0.3011	6.2978	0.0121	0.1390
relationship1	1	1.0489	0.4697	4.9875	0.0255	0.1309
hrs_per_wek	1	0.0428	0.00878	23.7830	<.0001	0.2723
capital_gain	1	0.000384	0.000055	48.1858	<.0001	1.4458
capital_loss	1	0.000587	0.000191	9.4944	0.0021	0.1292
continent2	1	-2.5824	1.0609	5.9255	0.0149	-0.2735
interaction1	1	0.0597	0.0148	16.2056	<.0001	0.6404

Table 5. Backward Selection Procedure Summary

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensitivity	Specificity	False POS	False NEG
0.2	247	691	217	45	78.2	84.6	76.1	46.8	6.1
0.25	240	722	186	52	80.2	82.2	79.5	43.7	6.7
0.3	223	782	126	69	83.8	76.4	86.1	36.1	8.1
0.35	219	798	110	73	84.8	75	87.9	33.4	8.4
0.4	208	818	90	84	85.5	71.2	90.1	30.2	9.3
0.45	191	828	80	101	84.9	65.4	91.2	29.5	10.9
0.5	181	843	65	11	85.3	62	92.8	26.4	11.6
0.55	169	854	54	123	85.3	57.9	94.1	24.2	12.6
0.6	162	859	49	130	85.1	55.5	94.6	23.2	13.1

SUM = Sensitivity + Specificity
160.7
161.7
162.5
162.9
161.3
156.6
154.8
152
150.1

Table of income by pred_y			
	pred_y		
income	0	1	Total
0	512	86	598
1	40	162	202
Total	552	248	800

TN =	512
FN =	40
FP =	86
TP =	162

Sensitivity =	0.80
Specificity =	0.86
Accuracy =	0.84
Precision =	0.65
F-metric =	0.72

Table 6. Classification Table

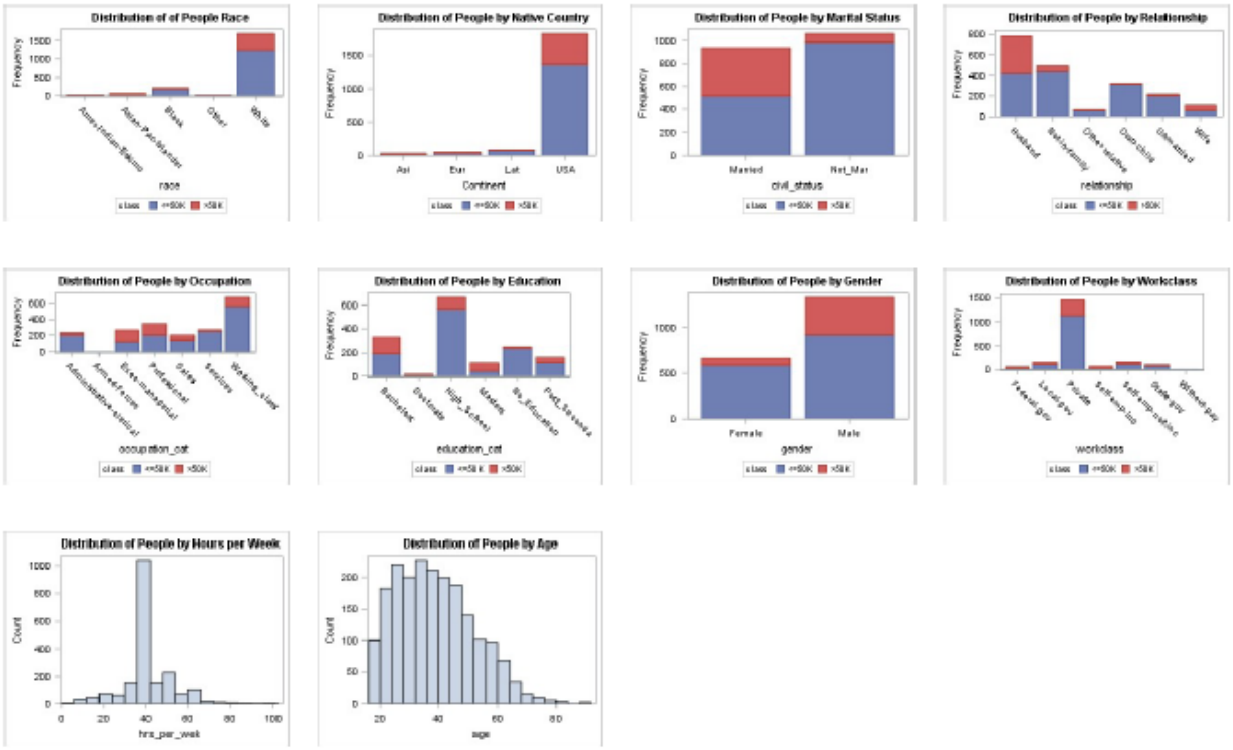


Figure 1. Distribution of Variables

Appendix E - Alice

Data Exploration

The FREQ Procedure

d_income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1506	75.30	1506	75.30
1	494	24.70	2000	100.00

Figure 1

DESCRIPTIVE STATS.

The MEANS Procedure

Variable	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum
hours_per_week	40.9675000	11.9432728	0.2670597	40.4437555	41.4912445	1.0000000	40.0000000	40.0000000	45.0000000	99.0000000
age	38.3370000	13.0308608	0.2913789	37.7655618	38.9084382	17.0000000	28.0000000	37.0000000	47.0000000	90.0000000
education_num	10.2195000	2.5885653	0.0578821	10.1059845	10.3330155	1.0000000	9.0000000	10.0000000	13.0000000	16.0000000
capital_gains	1408.30	9243.11	206.6823105	1002.97	1813.64	0	0	0	0	99999.00
capital_loss	79.2700000	382.8569652	8.5609420	62.4806964	96.0593036	0	0	0	0	2444.00
interaction2	1587.55	717.2457076	16.0381016	1556.10	1619.00	54.0000000	1082.50	1533.00	2015.50	5742.00

Figure 2

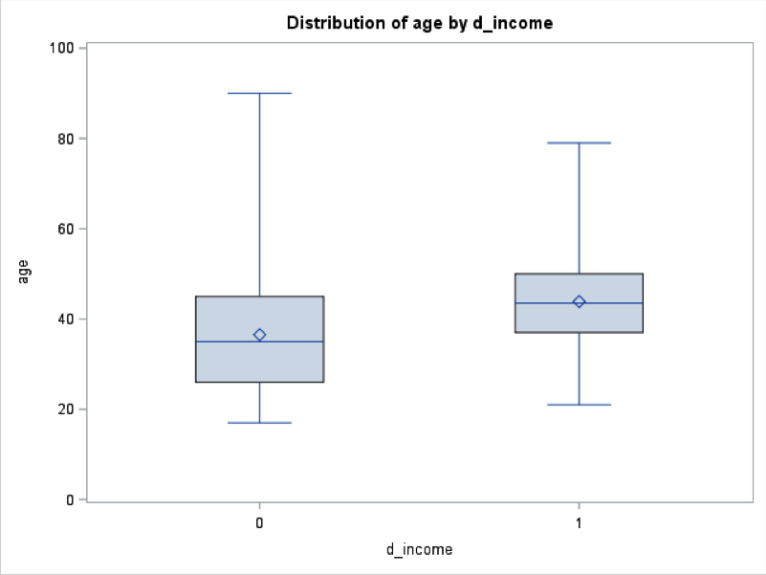


Figure 3

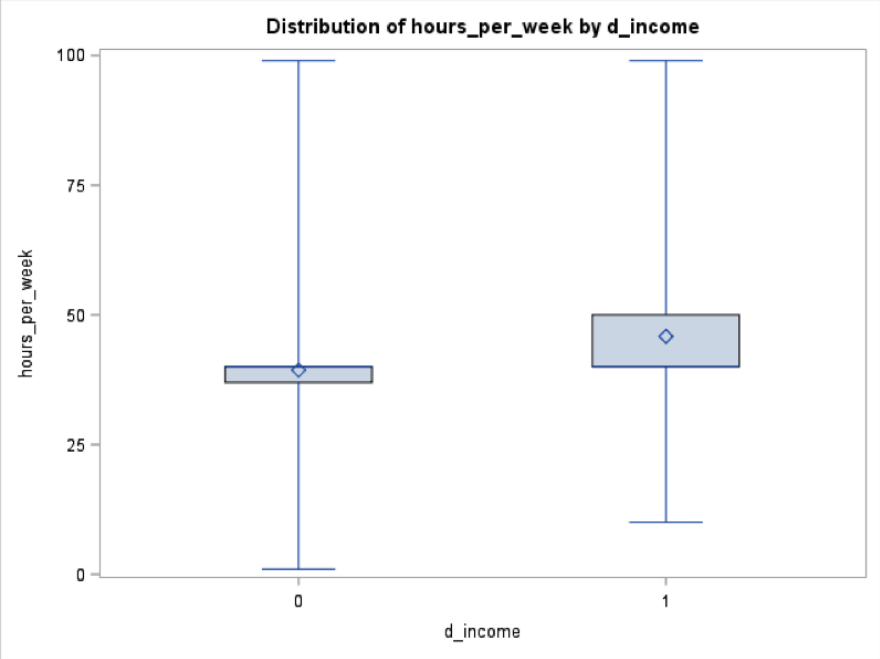


Figure 4

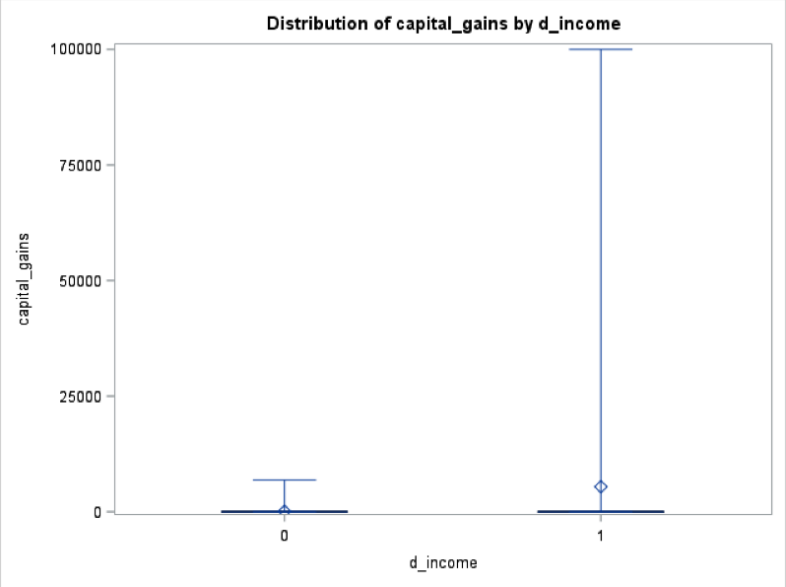


Figure 5

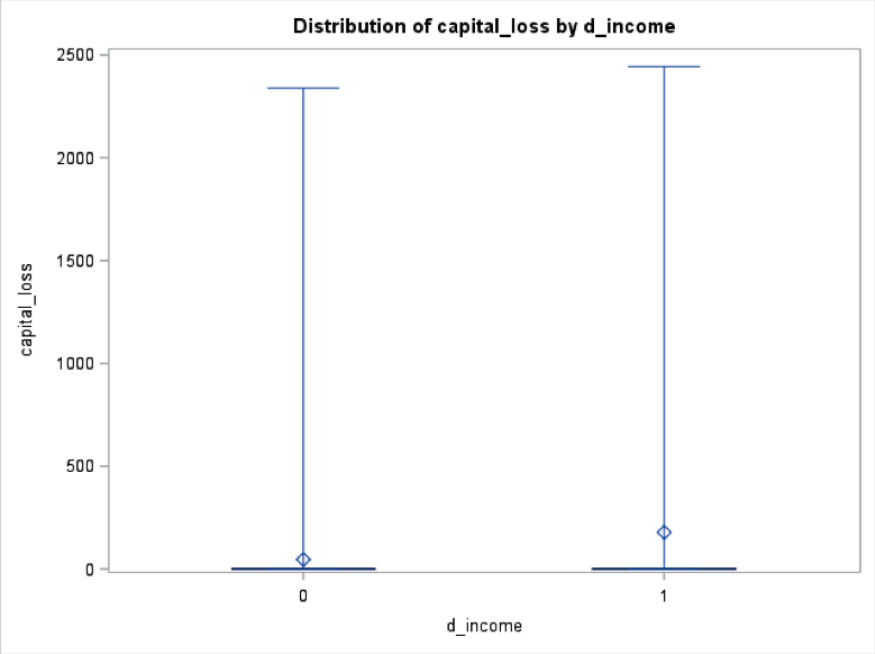


Figure 6

Counts

	education_num	income	COUNT
1	1	<=50K	1
2	2	<=50K	9
3	3	<=50K	19
4	3	>50K	2
5	4	<=50K	28
6	4	>50K	1
7	5	<=50K	35
8	5	>50K	2
9	6	<=50K	54
10	6	>50K	5
11	7	<=50K	55
12	7	>50K	4
13	8	<=50K	23
14	9	<=50K	539
15	9	>50K	102
16	10	<=50K	348
17	10	>50K	85
18	11	<=50K	72
19	11	>50K	18
20	12	<=50K	46
21	12	>50K	20
22	13	<=50K	216
23	13	>50K	132
24	14	<=50K	43
25	14	>50K	58
26	15	<=50K	8
27	15	>50K	34
28	16	<=50K	10
29	16	>50K	31

Figure 7

	workclass	income	COUNT
1	Federal-gov	<=50K	39
2	Federal-gov	>50K	27
3	Local-gov	<=50K	103
4	Local-gov	>50K	39
5	Private	<=50K	1155
6	Private	>50K	314
7	Self-emp-inc	<=50K	30
8	Self-emp-inc	>50K	44
9	Self-emp-not-inc	<=50K	119
10	Self-emp-not-inc	>50K	41
11	State-gov	<=50K	59
12	State-gov	>50K	29
13	Without-pay	<=50K	1

Figure 8

	occupation	income	COUNT
1	Adm-clerical	<=50K	230
2	Adm-clerical	>50K	28
3	Armed-Forces	<=50K	1
4	Craft-repair	<=50K	197
5	Craft-repair	>50K	70
6	Exec-managerial	<=50K	140
7	Exec-managerial	>50K	115
8	Farming-fishing	<=50K	60
9	Farming-fishing	>50K	4
10	Handlers-cleaners	<=50K	94
11	Handlers-cleaners	>50K	4
12	Machine-op-inspct	<=50K	94
13	Machine-op-inspct	>50K	16
14	Other-service	<=50K	202
15	Other-service	>50K	9
16	Priv-house-serv	<=50K	6
17	Prof-specialty	<=50K	157
18	Prof-specialty	>50K	130
19	Protective-serv	<=50K	26
20	Protective-serv	>50K	16
21	Sales	<=50K	177
22	Sales	>50K	71
23	Tech-support	<=50K	41
24	Tech-support	>50K	19
25	Transport-moving	<=50K	81
26	Transport-moving	>50K	12

Figure 9

race	income	COUNT	Ratio (>50k/<=50k)
Amer-Indian-Eskimo	<=50K	15	0.066666667
Amer-Indian-Eskimo	>50K	1	
Asian-Pac-Islander	<=50K	49	0.265306122
Asian-Pac-Islander	>50K	13	
Black	<=50K	181	0.121546961
Black	>50K	22	
Other	<=50K	4	0.25
Other	>50K	1	
White	<=50K	1257	0.363564041
White	>50K	457	

Figure 10

	race	income	COUNT
1	Amer-Indian-Eskimo	<=50K	15
2	Amer-Indian-Eskimo	>50K	1
3	Asian-Pac-Islander	<=50K	49
4	Asian-Pac-Islander	>50K	13
5	Black	<=50K	181
6	Black	>50K	22
7	Other	<=50K	4
8	Other	>50K	1
9	White	<=50K	1257
10	White	>50K	457

Figure 11

	sex	income	COUNT
1	Female	<=50K	584
2	Female	>50K	62
3	Male	<=50K	922
4	Male	>50K	432

Figure 12

Full Model

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2238.061	1382.100
SC	2243.662	1510.921
-2 Log L	2236.061	1336.100

R-Square	0.3624	Max-rescaled R-Square	0.5384

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	899.9606	22	<.0001
Score	698.9840	22	<.0001
Wald	381.2768	22	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-19.3163	326.3	0.0035	0.9528	
hours_per_week	1	-0.0100	0.0214	0.2201	0.6389	-0.0660
age	1	-0.00621	0.0202	0.0941	0.7590	-0.0446
education_num	1	0.2515	0.0622	16.3644	<.0001	0.3589
capital_gains	1	0.000407	0.000051	63.2556	<.0001	2.0760
capital_loss	1	0.000663	0.000157	17.8614	<.0001	0.1399
d_work_private	1	12.0783	326.3	0.0014	0.9705	2.3685
d_work_gov	1	11.8989	326.3	0.0013	0.9709	2.3301
d_edu_primary	1	-0.4351	1.0737	0.1642	0.6853	-0.0409
d_edu_secondary	1	-0.1579	0.6409	0.0607	0.8054	-0.0428
d_edu_college	1	0.0532	0.5323	0.0100	0.9204	0.0146
d_edu_higher	1	0.3982	0.5240	0.5775	0.4473	0.0566
d_marital	1	2.3468	0.3268	51.5705	<.0001	0.6469
d_occup_service	1	-0.0227	0.2048	0.0123	0.9117	-0.00626
d_occup_exec	1	0.3540	0.2194	2.6032	0.1066	0.0868
d_relat_family	1	-0.3699	0.3336	1.2294	0.2675	-0.0966
d_race_white	1	-0.0639	1.3612	0.0022	0.9625	-0.0123
d_race_asian	1	-0.3252	1.4052	0.0536	0.8170	-0.0311
d_race_amer_ind	1	-1.9653	1.9283	1.0387	0.3081	-0.0965
d_race_black	1	-0.6297	1.3889	0.2056	0.6503	-0.1049
d_sex	1	0.4907	0.2192	5.0089	0.0252	0.1265
d_country	1	0.7264	0.3209	5.1252	0.0236	0.1155
interaction2	1	0.000890	0.000473	3.5447	0.0597	0.3521

Figure 13
Correlation Matrix

Estimated Correlatic											
Parameter	Intercept	hours_per_week	age	education_num	capital_gains	capital_loss	d_work_private	d_work_gov	d_edu_primary	d_edu_secondary	d_edu_college
Intercept	1.0000	-0.0020	-0.0021	-0.0029	-0.0001	0.0000	-1.0000	-1.0000	-0.0026	-0.0031	-0.0028
hours_per_week	-0.0020	1.0000	0.9168	-0.0212	-0.0073	0.0017	-0.0005	-0.0005	-0.0005	0.0144	0.0178
age	-0.0021	0.9168	1.0000	0.0073	-0.0002	0.0066	-0.0004	-0.0004	-0.0049	0.0321	0.0348
education_num	-0.0029	-0.0212	0.0073	1.0000	0.0291	-0.0093	0.0001	0.0001	0.6725	0.6529	0.4679
capital_gains	-0.0001	-0.0073	-0.0002	0.0291	1.0000	0.0552	-0.0000	-0.0001	0.0185	0.0394	0.0618
capital_loss	0.0000	0.0017	0.0066	-0.0093	0.0552	1.0000	-0.0000	-0.0001	-0.0568	-0.0027	-0.0131
d_work_private	-1.0000	-0.0005	-0.0004	0.0001	-0.0000	-0.0000	1.0000	1.0000	0.0001	0.0002	0.0001
d_work_gov	-1.0000	-0.0005	-0.0004	0.0001	-0.0001	-0.0001	1.0000	1.0000	0.0000	0.0001	0.0001
d_edu_primary	-0.0026	-0.0005	-0.0049	0.6725	0.0185	-0.0568	0.0001	0.0000	1.0000	0.7585	0.6870
d_edu_secondary	-0.0031	0.0144	0.0321	0.6529	0.0394	-0.0027	0.0002	0.0001	0.7585	1.0000	0.9383
d_edu_college	-0.0028	0.0178	0.0348	0.4679	0.0618	-0.0131	0.0001	0.0001	0.6870	0.9383	1.0000
d_edu_higher	-0.0020	0.0453	0.0490	0.1987	0.0648	-0.0231	0.0001	0.0000	0.4931	0.7411	0.8292
d_marital	0.0001	-0.1352	-0.1829	0.0249	0.0121	-0.0104	-0.0000	-0.0000	0.0252	0.0053	0.0044
d_occup_service	-0.0004	-0.0447	-0.0339	0.0261	0.0044	0.0184	0.0001	0.0001	-0.0136	0.0078	0.0227
d_occup_exec	0.0001	-0.0881	-0.0619	-0.1428	0.0131	0.0358	0.0000	-0.0000	-0.0304	0.0323	0.0483
d_relat_family	-0.0003	0.1021	0.1282	0.0213	0.0073	-0.0178	0.0001	0.0001	-0.0106	-0.0066	-0.0028
d_race_white	-0.0038	-0.0336	-0.0432	-0.0656	-0.0141	-0.0088	0.0000	0.0000	-0.0379	-0.0280	-0.0148
d_race_asian	-0.0038	-0.0261	-0.0402	-0.0677	-0.0061	-0.0388	0.0000	0.0000	-0.0248	-0.0225	-0.0107
d_race_amer_ind	-0.0027	-0.0298	-0.0340	-0.0239	-0.0852	-0.0013	0.0000	0.0000	-0.0115	-0.0065	-0.0091
d_race_black	-0.0038	-0.0286	-0.0394	-0.0487	-0.0203	-0.0058	0.0000	0.0000	-0.0258	-0.0206	-0.0114
d_sex	0.0000	-0.0305	0.0160	-0.0125	0.0321	0.0591	-0.0001	-0.0000	-0.0411	-0.0434	-0.0490
d_country	-0.0005	-0.0194	-0.0375	0.0695	0.0095	-0.0077	0.0000	-0.0000	0.0990	0.0179	0.0058
interaction2	0.0017	-0.9501	-0.9521	0.0089	0.0081	-0.0021	0.0006	0.0006	0.0037	-0.0050	-0.0006

on Matrix										
d_edu_higher	d_marital	d_occup_service	d_occup_exec	d_relat_family	d_race_white	d_race_asian	d_race_amer_ind	d_race_black	d_sex	d_country	interaction2
-0.0020	0.0001	-0.0004	0.0001	-0.0003	-0.0038	-0.0038	-0.0027	-0.0038	0.0000	-0.0005	0.0017
0.0453	-0.1352	-0.0447	-0.0881	0.1021	-0.0336	-0.0261	-0.0298	-0.0286	-0.0305	-0.0194	-0.9501
0.0490	-0.1829	-0.0339	-0.0619	0.1282	-0.0432	-0.0402	-0.0340	-0.0394	0.0160	-0.0375	-0.9521
0.1987	0.0249	0.0261	-0.1428	0.0213	-0.0656	-0.0677	-0.0239	-0.0487	-0.0125	0.0695	0.0089
0.0648	0.0121	0.0044	0.0131	0.0073	-0.0141	-0.0061	-0.0852	-0.0203	0.0321	0.0095	0.0081
-0.0231	-0.0104	0.0184	0.0358	-0.0178	-0.0088	-0.0388	-0.0013	-0.0058	0.0591	-0.0077	-0.0021
0.0001	-0.0000	0.0001	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	-0.0001	0.0000	0.0006
0.0000	-0.0000	0.0001	-0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	-0.0000	-0.0000	0.0006
0.4931	0.0252	-0.0136	-0.0304	-0.0106	-0.0379	-0.0248	-0.0115	-0.0258	-0.0411	0.0990	0.0037
0.7411	0.0053	0.0078	0.0323	-0.0066	-0.0280	-0.0225	-0.0065	-0.0206	-0.0434	0.0179	-0.0050
0.8292	0.0044	0.0227	0.0483	-0.0028	-0.0148	-0.0107	-0.0091	-0.0114	-0.0490	0.0058	-0.0006
1.0000	0.0116	0.0026	-0.0152	-0.0102	-0.0048	-0.0025	-0.0029	-0.0053	-0.0279	-0.0114	-0.0213
0.0116	1.0000	0.0240	0.0210	-0.8131	0.0769	0.0807	0.0540	0.0745	-0.1538	0.0763	0.1224
0.0026	0.0240	1.0000	0.6551	-0.0012	-0.0006	0.0172	-0.0192	-0.0052	-0.2148	-0.0395	0.0245
-0.0152	0.0210	0.6551	1.0000	-0.0047	0.0266	0.0314	0.0099	0.0282	-0.1070	-0.0720	0.0541
-0.0102	-0.8131	-0.0012	-0.0047	1.0000	-0.0781	-0.0836	-0.0590	-0.0738	-0.1134	-0.0680	-0.0711
-0.0048	0.0769	-0.0006	0.0266	-0.0781	1.0000	0.9508	0.7046	0.9770	-0.0068	-0.1186	0.0467
-0.0025	0.0807	0.0172	0.0314	-0.0836	0.9508	1.0000	0.6697	0.9303	-0.0135	0.0158	0.0384
-0.0029	0.0540	-0.0192	0.0099	-0.0590	0.7046	0.6697	1.0000	0.6915	-0.0044	-0.0885	0.0381
-0.0053	0.0745	-0.0052	0.0282	-0.0738	0.9770	0.9303	0.6915	1.0000	0.0055	-0.1213	0.0412
-0.0279	-0.1538	-0.2148	-0.1070	-0.1134	-0.0068	-0.0135	-0.0044	0.0055	1.0000	-0.0125	-0.0195
-0.0114	0.0763	-0.0395	-0.0720	-0.0680	-0.1186	0.0158	-0.0885	-0.1213	-0.0125	1.0000	0.0199
-0.0213	0.1224	0.0245	0.0541	-0.0711	0.0467	0.0384	0.0381	0.0412	-0.0195	0.0199	1.0000

Figure 14
Outliers/Influential points: First time

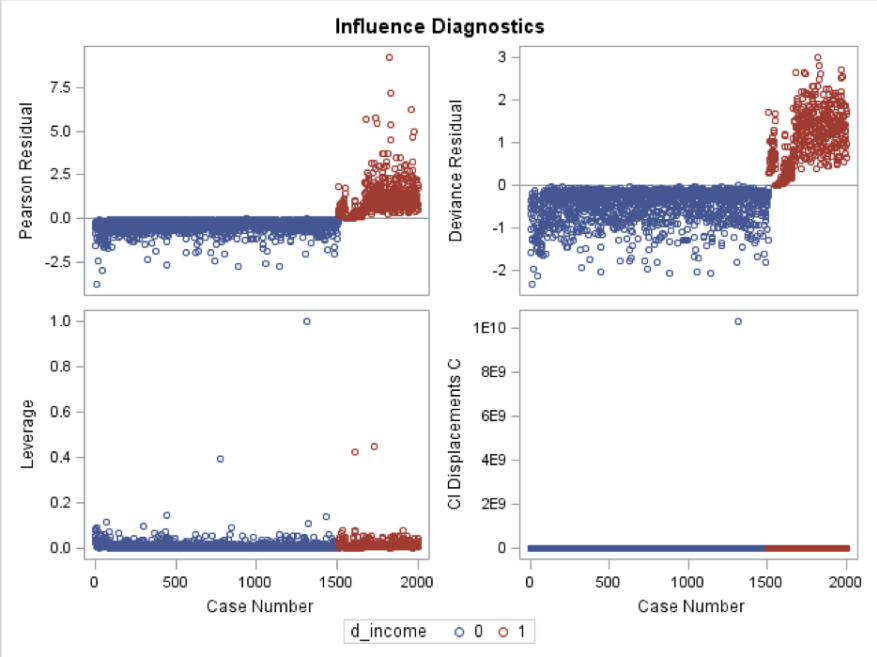


Figure 15

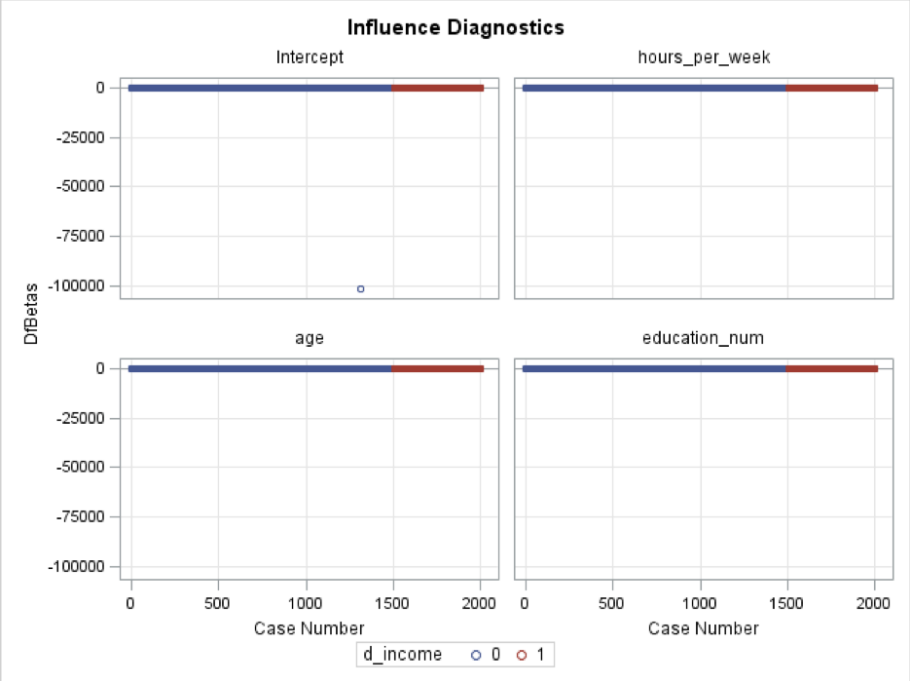


Figure 16

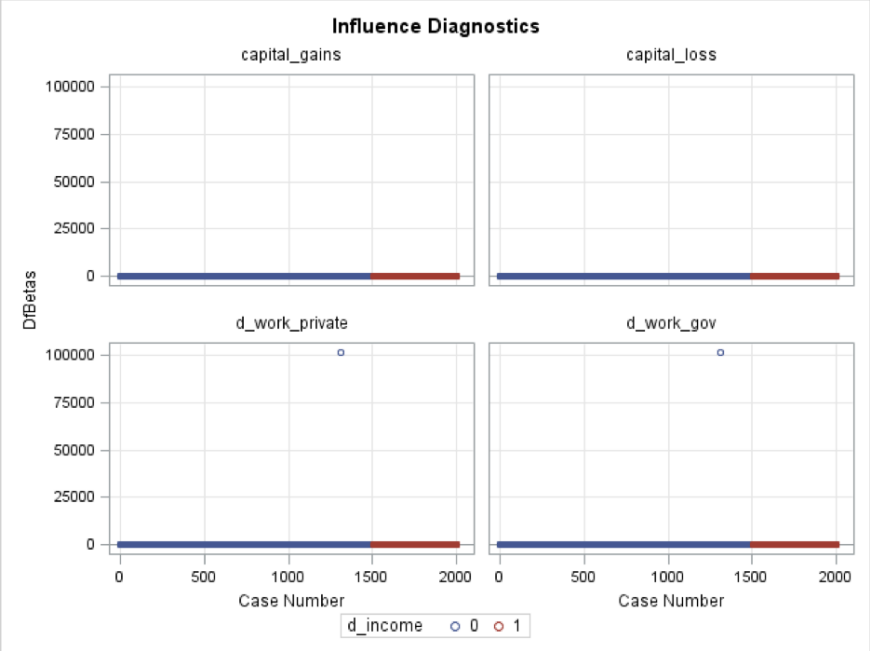


Figure 17

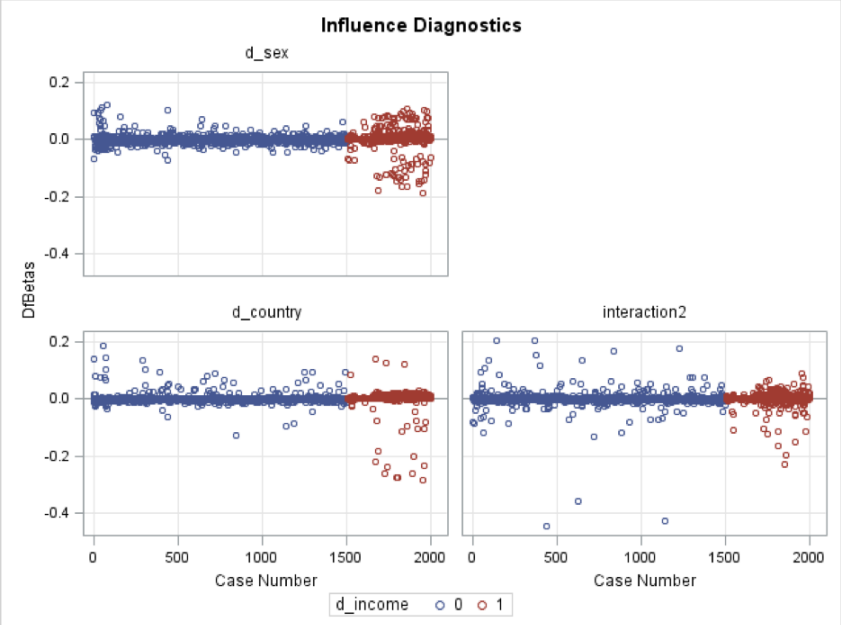


Figure 18
Model Fit after Outliers are handled

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2175.236	1228.393
SC	2180.825	1356.947
-2 Log L	2173.236	1182.393

R-Square	0.3942	Max-rescaled R-Square	0.5911

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	990.8429	22	<.0001
Score	748.3699	22	<.0001
Wald	366.9232	22	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-20.0005	200.6	0.0099	0.9206	
hours_per_week	1	-0.0242	0.0232	1.0877	0.2970	-0.1598
age	1	-0.0156	0.0219	0.5071	0.4764	-0.1123
education_num	1	0.3559	0.0677	27.6500	<.0001	0.5059
capital_gains	1	0.000447	0.000055	65.9252	<.0001	2.2885
capital_loss	1	0.000814	0.000171	22.5747	<.0001	0.1718
d_work_private	1	11.2664	200.5	0.0032	0.9552	2.2075
d_work_gov	1	10.9674	200.5	0.0030	0.9564	2.1459
d_edu_primary	1	-6.7947	21.4337	0.1005	0.7512	-0.6270
d_edu_secondary	1	0.3467	0.6846	0.2565	0.6126	0.0940
d_edu_college	1	0.3607	0.5675	0.4040	0.5250	0.0993
d_edu_higher	1	0.5525	0.5572	0.9832	0.3214	0.0789
d_marital	1	2.6139	0.3653	51.2037	<.0001	0.7206
d_occup_service	1	-0.0695	0.2179	0.1017	0.7499	-0.0192
d_occup_exec	1	0.3735	0.2322	2.5868	0.1078	0.0917
d_relat_family	1	-0.1896	0.3740	0.2570	0.6122	-0.0494
d_race_white	1	-0.5002	1.4673	0.1162	0.7332	-0.0969
d_race_asian	1	-0.7027	1.5114	0.2162	0.6420	-0.0670
d_race_amer_ind	1	-2.3228	2.0696	1.2597	0.2617	-0.1148
d_race_black	1	-0.9156	1.4950	0.3751	0.5402	-0.1533
d_sex	1	0.5093	0.2404	4.4884	0.0341	0.1314
d_country	1	1.1537	0.3775	9.3408	0.0022	0.1821
interaction2	1	0.00121	0.000513	5.5558	0.0184	0.4791

Figure 19
Stepwise

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1666.030	925.482
SC	1671.332	967.896
-2 Log L	1664.030	909.482

R-Square	0.3988	Max-rescaled R-Square	0.5913

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	754.5484	7	<.0001
Score	556.5623	7	<.0001
Wald	283.7173	7	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
14.7726	14	0.3939

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.5988	0.7184	217.6858	<.0001
education_num	1	0.4126	0.0386	114.2989	<.0001
capital_gains	1	0.000419	0.000061	46.9515	<.0001
capital_loss	1	0.000638	0.000177	12.9569	0.0003
d_marital	1	2.6551	0.2196	146.2252	<.0001
d_race_white	1	0.5345	0.2683	3.9692	0.0463
d_country	1	0.9183	0.3667	6.2721	0.0123
interaction2	1	0.000929	0.000136	46.6556	<.0001

Figure 20

Backward

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1666.030	925.482
SC	1671.332	967.896
-2 Log L	1664.030	909.482

R-Square	0.3988	Max-rescaled R-Square	0.5913

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	754.5484	7	<.0001
Score	556.5623	7	<.0001
Wald	283.7173	7	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
14.7726	14	0.3939

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.5988	0.7184	217.6858	<.0001
education_num	1	0.4126	0.0386	114.2989	<.0001
capital_gains	1	0.000419	0.000061	46.9515	<.0001
capital_loss	1	0.000638	0.000177	12.9569	0.0003
d_marital	1	2.6551	0.2196	146.2252	<.0001
d_race_white	1	0.5345	0.2683	3.9692	0.0463
d_country	1	0.9183	0.3667	6.2721	0.0123
interaction2	1	0.000929	0.000136	46.6556	<.0001

Figure 21

Forward

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1666.030	925.482
SC	1671.332	967.896
-2 Log L	1664.030	909.482

R-Square	0.3988	Max-rescaled R-Square	0.5913

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	754.5484	7	<.0001
Score	556.5623	7	<.0001
Wald	283.7173	7	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
14.7726	14	0.3939

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.5988	0.7184	217.6858	<.0001
education_num	1	0.4126	0.0386	114.2989	<.0001
capital_gains	1	0.000419	0.000061	46.9515	<.0001
capital_loss	1	0.000638	0.000177	12.9569	0.0003
d_marital	1	2.6551	0.2196	146.2252	<.0001
d_race_white	1	0.5345	0.2683	3.9692	0.0463
d_country	1	0.9183	0.3667	6.2721	0.0123
interaction2	1	0.000929	0.000136	46.6556	<.0001

Figure 22

Outliers/Influential points: 2nd time

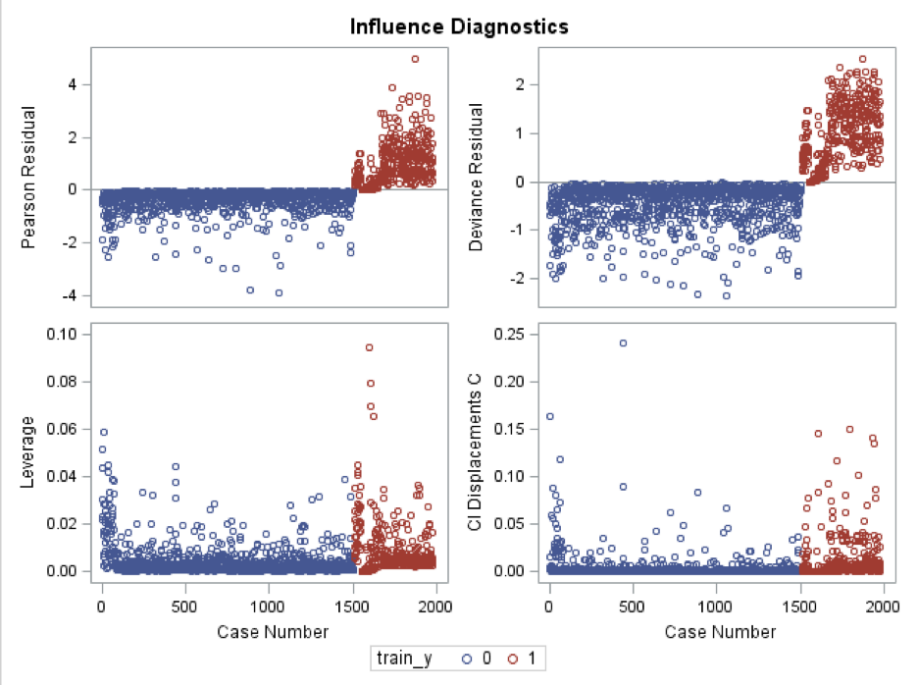


Figure 23

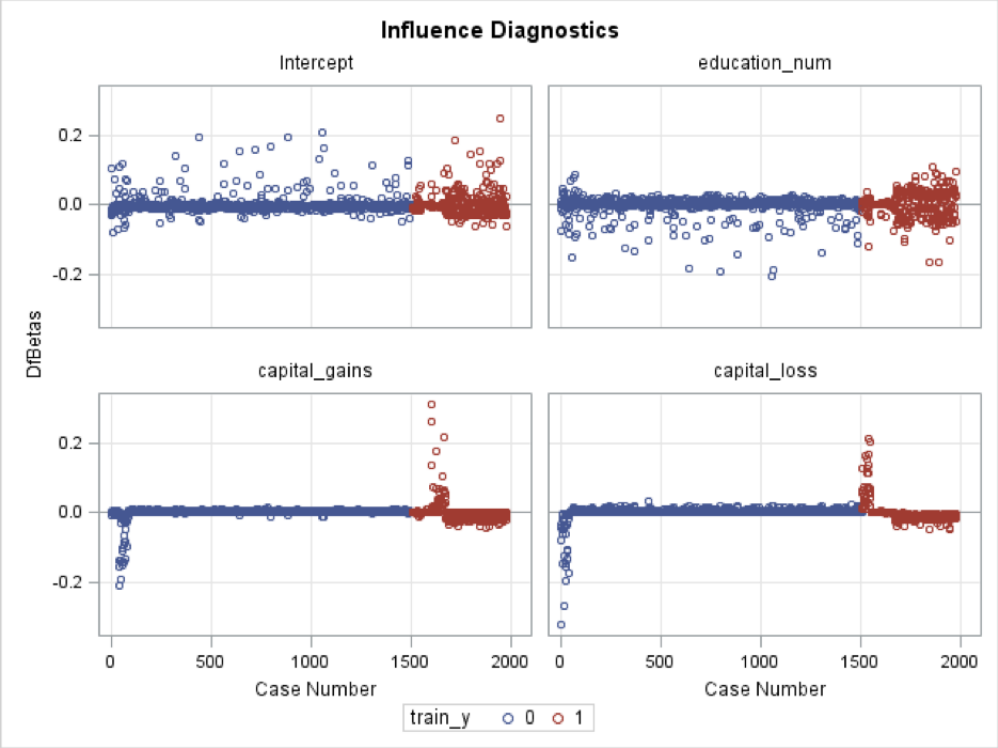


Figure 24

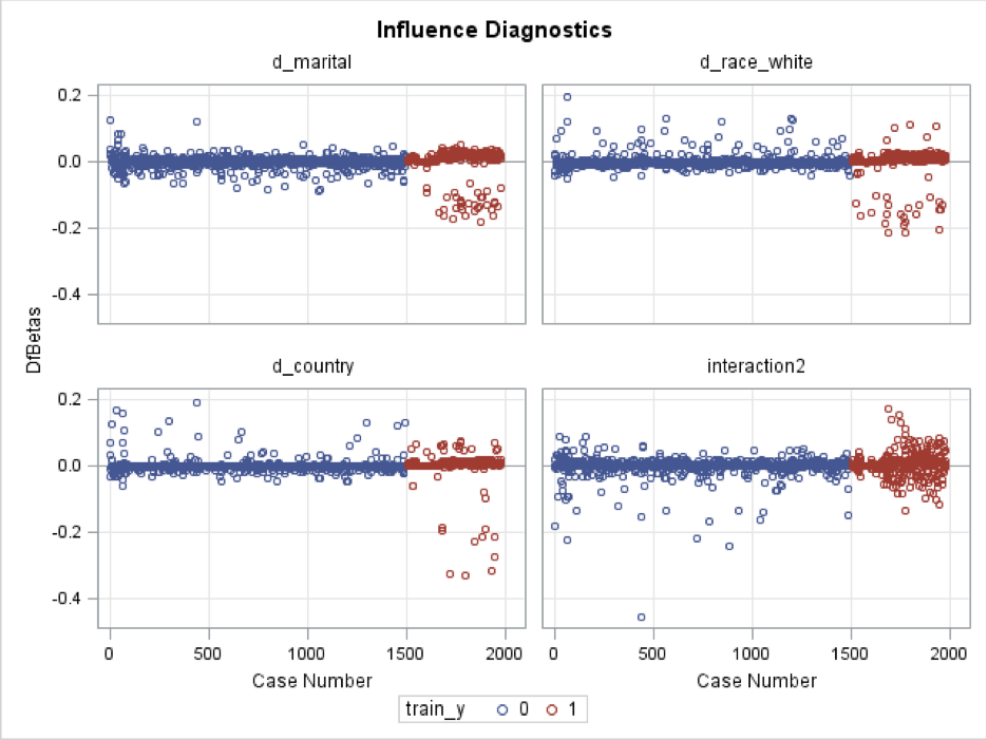


Figure 25

Final Model

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1639.704	864.474
SC	1644.998	906.829
-2 Log L	1637.704	848.474

R-Square	0.4150	Max-rescaled R-Square	0.6182

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	789.2297	7	<.0001
Score	572.2943	7	<.0001
Wald	273.7541	7	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-11.6287	0.7925	215.2837	<.0001	
education_num	1	0.4507	0.0411	120.3379	<.0001	0.6415
capital_gains	1	0.000442	0.000064	48.3252	<.0001	2.4421
capital_loss	1	0.000671	0.000184	13.3691	0.0003	0.1508
d_marital	1	2.9383	0.2420	147.4073	<.0001	0.8100
d_race_white	1	0.5097	0.2776	3.3713	0.0663	0.1030
d_country	1	1.1912	0.3984	8.9406	0.0028	0.1864
interaction2	1	0.000977	0.000143	46.9333	<.0001	0.3834

Figure 26

1st Prediction

Obs	education_num	capital_gains	capital_loss	d_marital	d_race_white	d_country	interaction2
1	10	3000	1000	1	1	1	1440

phat	lcl	ucl
0.71523	0.59789	0.80926

Figure 27

2nd Prediction

Obs	education_num	capital_gains	capital_loss	d_marital	d_race_white	d_country	interaction2
1	9	2000	0	0	0	0	988

phat	lcl	ucl
0.00326	0.00120	0.00881

Figure 28

Classification Table

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.100	350	747	365	10	74.5	97.2	67.2	51.0	1.3
0.150	341	804	308	19	77.8	94.7	72.3	47.5	2.3
0.200	324	865	247	36	80.8	90.0	77.8	43.3	4.0
0.250	307	915	197	53	83.0	85.3	82.3	39.1	5.5
0.300	284	953	159	76	84.0	78.9	85.7	35.9	7.4
0.350	264	977	135	96	84.3	73.3	87.9	33.8	8.9
0.400	242	1007	105	118	84.9	67.2	90.6	30.3	10.5
0.450	226	1021	91	134	84.7	62.8	91.8	28.7	11.6
0.500	213	1040	72	147	85.1	59.2	93.5	25.3	12.4
0.550	202	1053	59	158	85.3	56.1	94.7	22.6	13.0
0.600	193	1062	50	167	85.3	53.6	95.5	20.6	13.6

BOXPLOT - by d_income

The FREQ Procedure

Frequency	Table of d_income by pred_y		
	pred_y		
d_income	0	1	Total
0	295	96	391
1	12	91	103
Total	307	187	494

Figure 29

Model 2 Classification Matrix

Table of d_income by pred_y			
	pred_y		
d_income	0	1	Total
0	222	98	320
1	9	71	80
Total	231	169	400

Figure 30

Number						
	hours_per_week	age	education_num	capital_gains	capital_loss	d_work_private
10	50	59	13	0	2002	1
1674	40	40	4	0	0	1
1714	45	63	9	0	0	1
1728	40	56	10	0	0	1
1733	44	40	10	0	0	1
1743	40	53	3	0	0	1
1752	65	31	10	0	0	1
1781	50	49	9	0	0	1
1789	40	40	11	0	0	1
1800	40	28	9	0	0	1
1816	60	30	13	0	0	0
1818	43	31	5	0	0	1
1829	45	30	7	0	0	1
1830	40	36	9	0	0	0
1833	40	42	9	0	0	0
1867	50	53	10	0	0	1
1891	45	37	11	0	0	1
1914	25	66	5	0	0	1
1926	40	32	6	0	0	1
1961	50	36	3	0	0	1
1966	38	48	10	0	0	1
1967	50	28	13	0	0	1
1975	40	31	10	0	0	0

Figure 31

Appendix F - Carl

Figure 1

```
title "Import Income Data";
data Income;
proc import datafile = "Income_Carl.csv"
        out = Income replace;
getnames = yes;
delimiter = ',';
run;
data Income;
set Income;
if native_country = 'United-States';
if workclass = '?' then delete;
if occupation = '?' then delete;
if age in (0:17) then ages = 'under18';
if age in (18:29) then ages = 'youngAdult';
if age in (30:39) then ages = 'adult30s';
if age in (40:49) then ages = 'adult40s';
if age in (50:59) then ages = 'adult50s';
if age in (60:99) then ages = 'adult60+';
if hours_per_week in (0:19) then hours = 'lowPartTime';
if hours_per_week in (20:29) then hours = 'highPartTim';
if hours_per_week in (30:40) then hours = 'fullTime';
if hours_per_week in (41:99) then hours = 'overTime';
; run;
proc surveyselect data = Income
        method = srs n = 2000
        out = SampleIncome
        seed = 568212;
run;
proc print data = SampleIncome;
run;
```

Figure 2

ages	Frequency	Percent	Cumulative Frequency	Cumulative Percent
adult30	567	28.35	567	28.35
adult40	453	22.65	1020	51.00
adult50	262	13.10	1282	64.10
adult60	146	7.30	1428	71.40
under18	27	1.35	1455	72.75
youngAd	545	27.25	2000	100.00

workclass	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Federal-gov	67	3.35	67	3.35
Local-gov	139	6.95	206	10.30
Private	1479	73.95	1685	84.25
Self-emp-inc	75	3.75	1760	88.00
Self-emp-not-inc	164	8.20	1924	96.20
State-gov	76	3.80	2000	100.00

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10th	59	2.95	59	2.95
11th	69	3.45	128	6.40
12th	28	1.40	156	7.80
1st-4th	4	0.20	160	8.00
5th-6th	8	0.40	168	8.40
7th-8th	25	1.25	193	9.65
9th	22	1.10	215	10.75
Assoc-acdm	66	3.30	281	14.05
Assoc-voc	88	4.40	369	18.45
Bachelors	300	15.00	669	33.45
Doctorate	24	1.20	693	34.65
HS-grad	649	32.45	1342	67.10
Masters	130	6.50	1472	73.60
Prof-school	29	1.45	1501	75.05
Some-college	499	24.95	2000	100.00

marital_status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Divorced	284	14.20	284	14.20
Married-AF-spou	1	0.05	285	14.25
Married-civ-spo	968	48.40	1253	62.65
Married-spouse-	15	0.75	1268	63.40
Never-married	629	31.45	1897	94.85
Separated	57	2.85	1954	97.70
Widowed	46	2.30	2000	100.00

occupation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Adm-clerical	251	12.55	251	12.55
Craft-repair	282	14.10	533	26.65
Exec-managerial	264	13.20	797	39.85
Farming-fishing	57	2.85	854	42.70
Handlers-cleaners	83	4.15	937	46.85
Machine-op-inspct	142	7.10	1079	53.95
Other-service	190	9.50	1269	63.45
Priv-house-serv	6	0.30	1275	63.75
Prof-specialty	278	13.90	1553	77.65
Protective-serv	34	1.70	1587	79.35
Sales	242	12.10	1829	91.45
Tech-support	52	2.60	1881	94.05
Transport-moving	119	5.95	2000	100.00

relationship	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Husband	859	42.95	859	42.95
Not-in-family	503	25.15	1362	68.10
Other-relative	47	2.35	1409	70.45
Own-child	291	14.55	1700	85.00
Unmarried	204	10.20	1904	95.20
Wife	96	4.80	2000	100.00

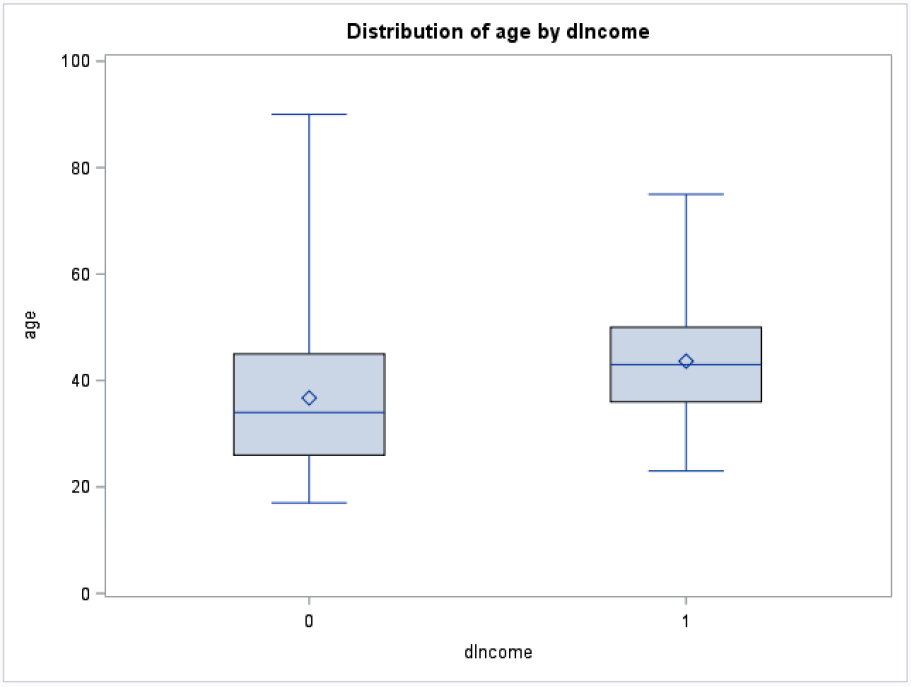
race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Amer-In	20	1.00	20	1.00
Asian-P	22	1.10	42	2.10
Black	190	9.50	232	11.60
Other	10	0.50	242	12.10
White	1758	87.90	2000	100.00

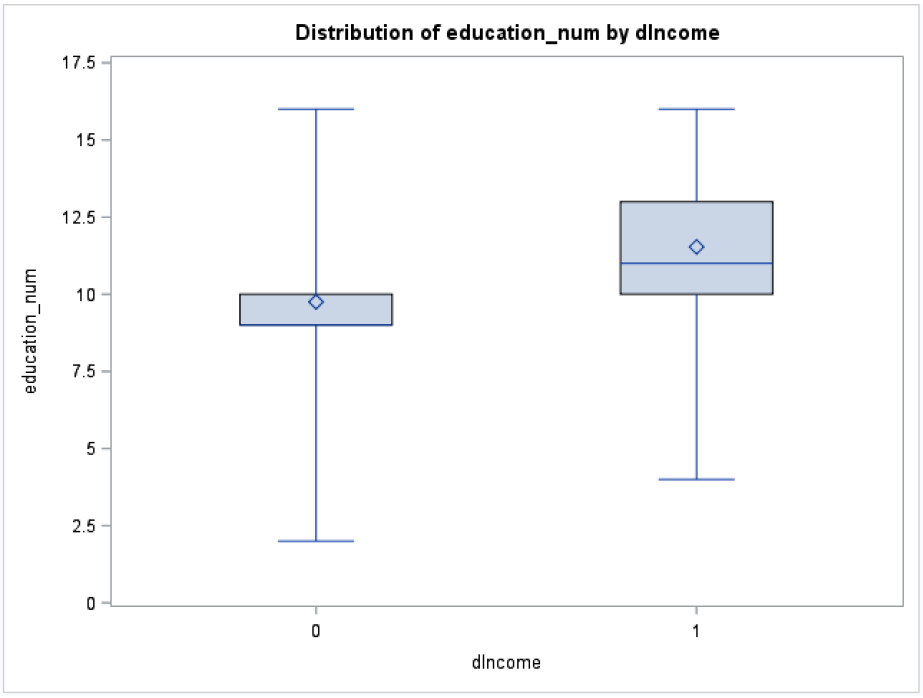
sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	631	31.55	631	31.55
Male	1369	68.45	2000	100.00

hours	Frequency	Percent	Cumulative Frequency	Cumulative Percent
fullTime	1170	58.50	1170	58.50
highPartTim	129	6.45	1299	64.95
lowPartTime	86	4.30	1385	69.25
overTime	615	30.75	2000	100.00

income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<=50K	1480	74.00	1480	74.00
>50K	520	26.00	2000	100.00

Figure 3


Distribution of age by dIncome
100
80
60
40
20
0
age
0
1
dIncome


Distribution of education_num by dIncome
17.5
15
12.5
10
7.5
5
2.5
0
education_num
0
1
dIncome

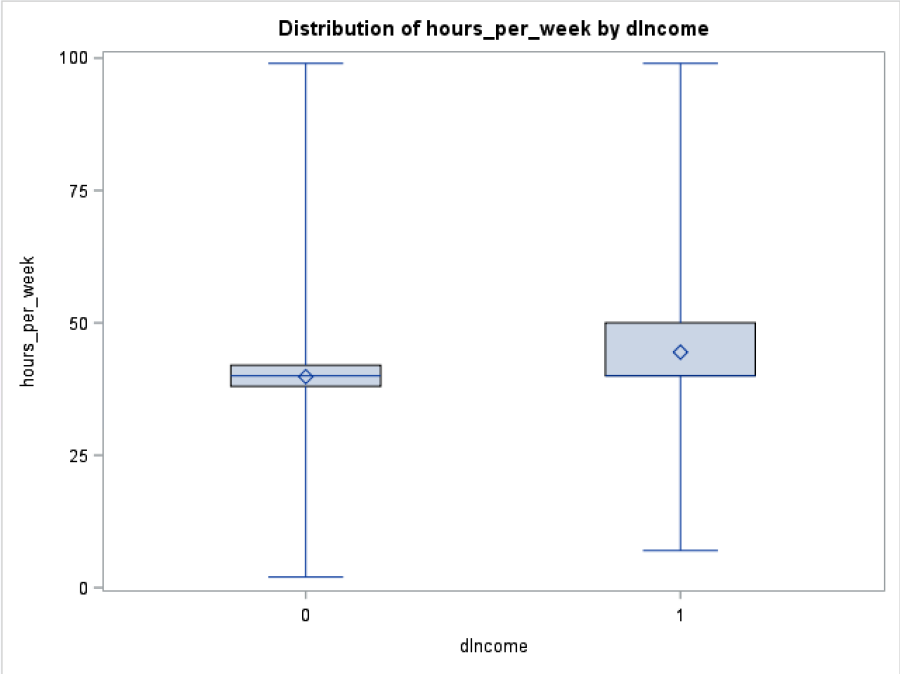


Figure 4

Estimated Correlation Matrix											
Parameter	Intercept	age	education_num	dMstatus	dJob	dRelation	dRace	dSex	capital_gain	capital_loss	hours_per_week
Intercept	1.0000	-0.5193	-0.6597	0.0115	-0.2610	-0.1385	-0.0546	-0.2733	-0.0207	-0.0544	-0.3691
age	-0.5193	1.0000	0.1091	-0.1112	0.0484	0.1544	0.0070	-0.0548	-0.0263	-0.0141	0.0201
education_num	-0.6597	0.1091	1.0000	0.0120	0.1449	-0.0987	0.0194	0.0518	0.0645	0.0558	-0.1349
dMstatus	0.0115	-0.1112	0.0120	1.0000	-0.0036	-0.7628	-0.0006	0.0509	-0.0186	0.0430	0.0221
dJob	-0.2610	0.0484	0.1449	-0.0036	1.0000	0.0050	-0.0554	-0.0476	-0.0027	-0.0102	0.0139
dRelation	-0.1385	0.1544	-0.0987	-0.7628	0.0050	1.0000	-0.0117	0.2733	-0.0625	-0.0699	-0.0050
dRace	-0.0546	0.0070	0.0194	-0.0006	-0.0554	-0.0117	1.0000	0.0527	-0.0783	0.0313	0.0007
dSex	-0.2733	-0.0548	0.0518	0.0509	-0.0476	0.2733	0.0527	1.0000	-0.0426	-0.0121	-0.1587
capital_gain	-0.0207	-0.0263	0.0645	-0.0186	-0.0027	-0.0625	-0.0783	-0.0426	1.0000	0.0611	-0.0278
capital_loss	-0.0544	-0.0141	0.0558	0.0430	-0.0102	-0.0699	0.0313	-0.0121	0.0611	1.0000	0.0030
hours_per_week	-0.3691	0.0201	-0.1349	0.0221	0.0139	-0.0050	0.0007	-0.1587	-0.0278	0.0030	1.0000

Figure 5

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	dRelation		1	1	362.8151		<.0001
2	education_num		1	2	220.3370		<.0001
3	capital_gain		1	3	78.2616		<.0001
4	capital_loss		1	4	31.7413		<.0001
5	age		1	5	28.6128		<.0001
6	dMstatus		1	6	15.4919		<.0001
7	dJob		1	7	7.0025		0.0081
8	dRace		1	8	4.0883		0.0432

Figure 6

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-6.0430	0.4980	147.2524	<.0001	
age	1	0.0310	0.00591	27.5999	<.0001	0.2221
education_num	1	0.4361	0.0347	158.2498	<.0001	0.5572
dMstatus	1	-0.8528	0.2160	15.5892	<.0001	-0.4437
dJob	1	-0.0615	0.0246	6.2319	0.0125	-0.1014
dRelation	1	-0.4736	0.0949	24.8769	<.0001	-0.4539
dRace	1	-0.1991	0.0991	4.0409	0.0444	-0.1028
capital_gain	1	0.000420	0.000051	67.1762	<.0001	0.4816
capital_loss	1	0.000823	0.000162	25.9720	<.0001	0.1747

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.032	1.020	1.044
education_num	1.547	1.445	1.655
dMstatus	0.426	0.279	0.651
dJob	0.940	0.896	0.987
dRelation	0.623	0.517	0.750
dRace	0.819	0.675	0.995
capital_gain	1.000	1.000	1.001
capital_loss	1.001	1.001	1.001

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.200	429	1076	381	52	77.7	89.2	73.9	47.0	4.6
0.250	405	1151	306	76	80.3	84.2	79.0	43.0	6.2
0.300	366	1201	256	115	80.9	76.1	82.4	41.2	8.7
0.350	338	1260	197	143	82.5	70.3	86.5	36.8	10.2
0.400	311	1288	169	170	82.5	64.7	88.4	35.2	11.7
0.450	287	1319	138	194	82.9	59.7	90.5	32.5	12.8
0.500	266	1345	112	215	83.1	55.3	92.3	29.6	13.8
0.550	247	1365	92	234	83.2	51.4	93.7	27.1	14.6
0.600	232	1372	85	249	82.8	48.2	94.2	26.8	15.4