MSDS 6372 Project 1

Intelligence and Class as Predictors of Future Income (Males only)

Team members:

Killion, Kyle

Martos, Mike

Taylor, Celia

Introduction

Our group decided to work on problem 12.23, "Intelligence and Class as Predictors of Future Income (Males only)," and the corresponding data set in our textbook, Ramsey, Fred; Schafer, Daniel. The Statistical Sleuth: A Course in Methods of Data Analysis (Page 380). Cengage Textbook. Kindle Edition. This data set is based on a widely publicized and controversial book, The Bell Curve: Intelligence and Class Structure in American Life, by psychologist Richard Hernstein and political scientist Charles Murray. The book argues that intelligence is a better predictor of financial income than other factors such as socioeconomic status or education. The controversy occurred because some contended the ASVAB intelligence score, AFQT, which was used in the study, was not an accurate intelligence test, the book was racist due to the book's statements about race and intelligence, and the data from The National Longitudinal Survey of Youth (NLSY79) did not support the authors' claims. This paper is a statistical analysis of the data set to understand and substantiate if the data set does or does not support the authors' claims that intelligence is a better predictor of income and to search for the best possible model of predictors of income based on the given flawed or unflawed data set.

Descriptive Statistics

To launch this analysis of whether future income is based on intelligence, we first set our income variable, Income2005, to be our response variable. The other 30 variables in the data set became potential explanatory variables. These explanatory variables are in six categories, which are AFQT and its components, Math, Arith, Word, and Parag, non-AFQT intelligence scores, socioeconomic status, family literacy in 1979, personal demographics, and ten esteem variables. We processed the data by selecting only the male observations and removing the ten esteem variables, which were from a self-assessment participants took in 2006. Also, we only used the Educ variable from the demographic group of variables.

Each of the remaining explanatory variables was put in many different groupings and compared to Income2005. An example of this is in Figure 01, where we are utilizing the final modeled variables to depict the preprocessing methods and issues encountered within the model. There is a noticeable skewness to almost all of the variables, so a transformation seemed warranted.

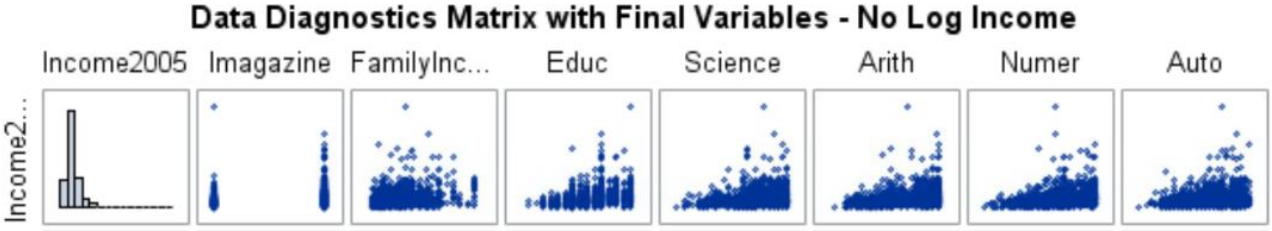

Figure 01

Analysis

Assumptions:

Once we had a subset of the data, we initialized our filters to review and ensure the assumptions for a multiple regression model.

- The Assumption of Linearity of residuals was examined using scatterplots and histograms. We transformed the response variable, Income2005, by taking the log of this variable and calling it loginc05. As Figure 02 depicts a banding of data along a line, loginc05 showed the assumption of the linearity of residuals was true.

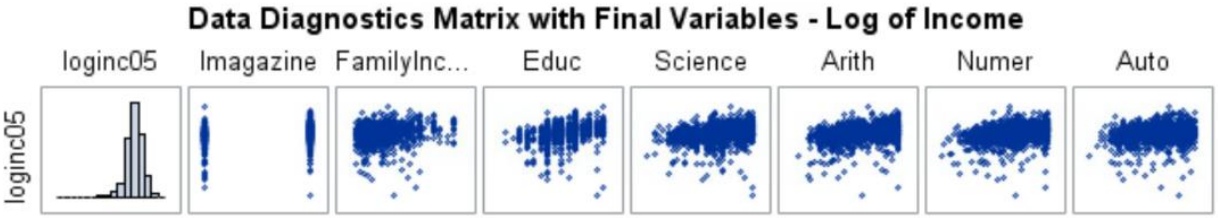

Figure 02

- We studied additional plots to understand the effect of loginc05 on outliers and leverage points. Figure 03 shows the data points of 233, 250, and 633 as outliers and leverage points, but no known, sound reason exists to exclude them from the model.

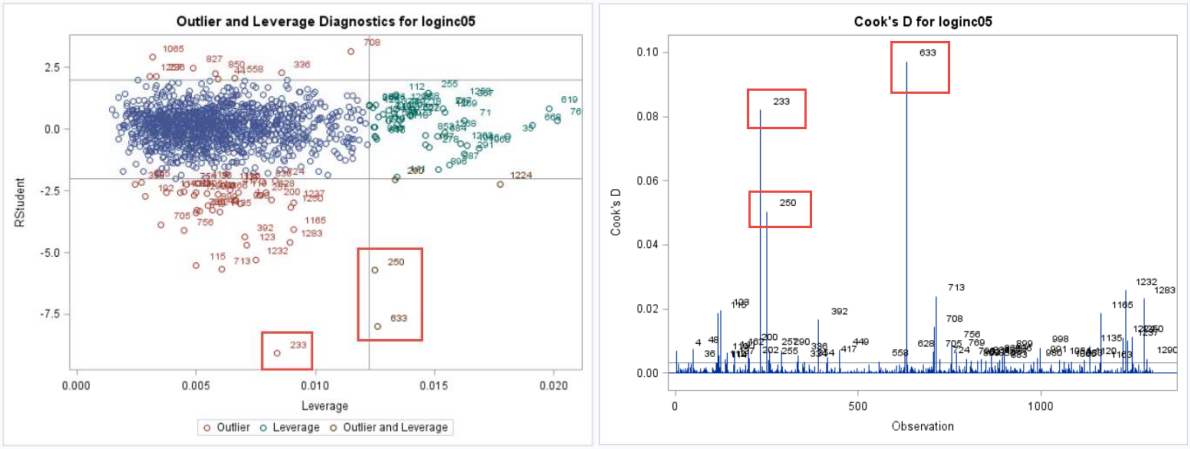

Figure 03

- The Assumptions of Normality and Independence were investigated under the multiple regression model. Figure 04 demonstrates that residuals are normally distributed and that errors are independently distributed with a slight negative skew which didn't violate this assumption.

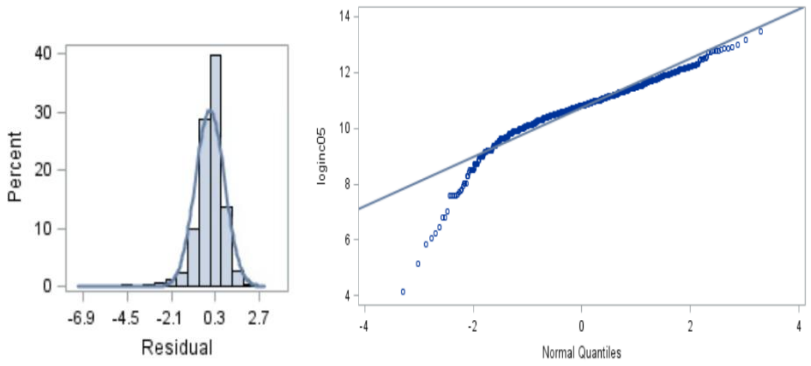

Figure 04

- The Assumption of Constant Variance within the residuals was then assessed. A stable level of variance is displayed in Figure 05 with no discernible pattern with the variables in the regression.

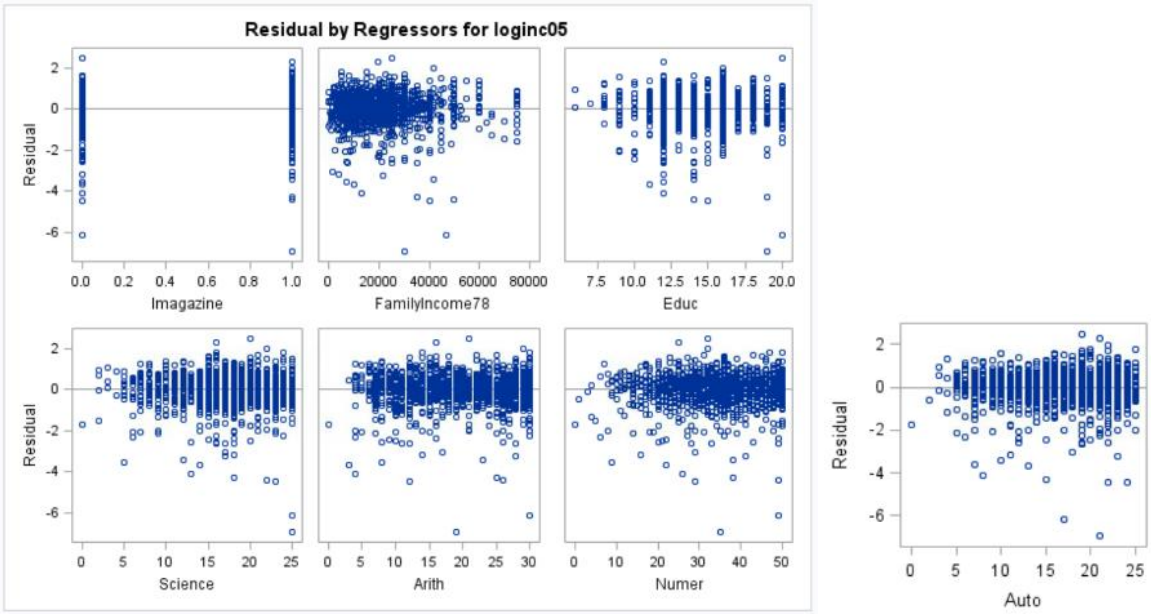

Figure 05

Finally, the variables within the data set are all quantifiable because they are either an integer or decimal number.

Build a Model:

In our efforts to reduce the number of variables, we check the variables for collinearity with other variables. We start by looking at variables that should have collinearity with each other because of their definition. Under suspicion are groups of variables such as Arith, Math, and Numer, or Word and Parag or AFQT since it is a combination of Word, Parag, Math, and Arith.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	9.00417	0.18799	47.90	<.0001	0
Imagazine	1	0.24741	0.05517	4.48	<.0001	1.24816
Inewspaper	1	-0.12202	0.07106	-1.72	0.0862	1.22547
Ilibrary	1	-0.00174	0.05382	-0.03	0.9743	1.11253
MotherEd	1	-0.01345	0.01143	-1.18	0.2396	1.85502
FatherEd	1	0.01640	0.00836	1.96	0.0500	1.90026
FamilyIncome78	1	0.00000483	0.00000174	2.78	0.0055	1.22044
Educ	1	0.06749	0.01195	5.65	<.0001	1.98110
Science	1	-0.01971	0.00935	-2.11	0.0353	4.36623
Arith	1	0.01963	0.00713	2.75	0.0060	5.67693
Word	1	-0.00391	0.00714	-0.55	0.5837	5.42845
Parag	1	-0.00283	0.01452	-0.19	0.8457	4.79743
Numer	1	0.01028	0.00336	3.06	0.0023	2.53951
Coding	1	-0.00008077	0.00216	-0.04	0.9702	2.12160
Auto	1	0.01652	0.00712	2.32	0.0204	2.77060
Math	1	-0.00206	0.00865	-0.24	0.8118	6.58973
Mechanic	1	0.00024491	0.00744	0.03	0.9737	3.03368
Elec	1	0.00549	0.01022	0.54	0.5910	3.55645
AFQT	1	0.00017390	0.00329	0.05	0.9578	18.23461

Figure 06

After running the PROC REG command to give the output in Figure 06, the AFQT variable shows a variance of inflation (VIF) value of 18. We decide to exclude the AFQT variable since the inflation is higher than 10. The other variables we thought might show collinearity didn't with VIF values of less than 7, so we decide to keep them in the model.

We conducted a variable selection procedure using the LASSO model with the cross-validation splitting the data randomly into five groups (4 used to test the variables and 1 to validate) and using AIC to stop the process. The process selected variables which affect the loginc05.

The result of the variable selection process is:

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	9.037202
Imagazine	1	0.216867
Inewspaper	1	-0.066227
FatherEd	1	0.008146
FamilyIncome78	1	0.000003968
Educ	1	0.056821
Science	1	-0.007910
Arith	1	0.015934
Numer	1	0.008616
Auto	1	0.010065

Root MSE	0.79340
Dependent Mean	10.74656
R-Square	0.1959
Adj R-Sq	0.1903
AIC	713.47507
AICC	713.67908
BIC	-592.35697
C(p)	9.11416
SBC	-542.77769
CV PRESS	824.28916

Stop Details				
Candidate For	Effect	Candidate AIC		Compare AIC
Entry	Word	712.5747	>	711.8073

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	9	198.69630	22.07737	35.07
Error	1296	815.81250	0.62948	
Corrected Total	1305	1014.50880		

Figure 07

Fit the model:

To make sure the variables above fit the model we ran a regression analysis. Per the regression analysis the variables FatherEd (p-value = 0.1224) and lnewspaper (p-value = 0.058) (Code 01) are not significant. We remove these variables, and the final model is as follows:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	8.94299	0.13850	64.57	<.0001	0
Imagazine	1	0.23661	0.05345	4.43	<.0001	1.17472
FamilyIncome78	1	0.00000478	0.00000167	2.86	0.0043	1.12954
Educ	1	0.06695	0.01104	6.06	<.0001	1.69600
Science	1	-0.02086	0.00755	-2.76	0.0058	2.85543
Arith	1	0.01889	0.00500	3.78	0.0002	2.80526
Numer	1	0.00941	0.00285	3.30	0.0010	1.83037
Auto	1	0.01597	0.00579	2.76	0.0059	1.83793

Figure 08

The above information shows the significant variables with associated parameter estimate, p-values, and the variance of inflation.

Check the Model Fit (lack of fit test):

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	198.56036	28.36577	45.12	<.0001
Error	1298	815.94844	0.62862		
Lack of Fit	1298	815.94844	0.62862	.	.
Pure Error	0	0	.		
Corrected Total	1305	1014.50880			

Root MSE	0.79286	R-Square	0.1957
Dependent Mean	10.74656	Adj R-Sq	0.1914
Coeff Var	7.37776		

Figure 09

The model is a good fit (p-value < 0.0001) with a very small Adjusted R^2 = 0.1914

The final model equation is as follows:

$$Loginc05 = 8.94 + (0.24)Imagazine + (0.000005)FamilyIncome78 + (0.07)Educ + (-0.02)Science + (0.02)Arith + (0.01)Numer + (0.02)Auto$$

Interpretation and Conclusion

There is significant evidence to suggest that the model is a good fit for the data set because of the p-value <.0001 for the model fit test, the very small p-values <.006 for all of the variables, and the very small Adjusted R^2 = 0.19, however, the Adjusted R^2 value <0.20 is not a statistically significant and confidence-building value. The closer the Adjusted R^2 value is to 1.0, the better the model fit is. No matter what variable grouping we did or what manual or automated selection procedures we tried, we never got an Adjusted R^2 value >0.20. The final model that was created was the best model for the data set with the largest Adjusted R^2 value = 0.20.

The authors, Hernstein and Murray, of The Bell Curve: Intelligence and Class Structure in American Life purported that the AFQT intelligence test was the best indicator for the future income of male subjects in the study. Yet, the AFQT VIF value > 18 in our tests showed that it was the worst and consequently eliminated it at the start of the analysis.

If we just look at intelligence as the predictor of income, the intelligence associated variables, Educ, Science, Arith, and Numer, are a part of the model. However, the largest coefficient (0.24) goes with the variable Imagazine, which is curious because the magazine could have been at any reading level for any subject's household member. The other variables are FamilyIncome78 which has such a tiny coefficient, and Auto which measures automotive and shop information. The sum of all of the variables' coefficients besides Imagazine is about half the value of Imagazine's coefficient. So the model does not weight intelligence as much as the regular presence of a magazine of unknown reading level in the household.

Because this is an observational study data set, no causations can be made. The best model from this statistical analysis does not seem to support the original authors' position that intelligence is the best predictor of future income. The data set seems more likely to support anyone who questioned and disbelieved the authors' controversial claims.

Appendix

```
/*************************************************************************
*************************************************************************

        Read in the Data set as ex1223 with all variables with Gender as characters
*************************************************************************

*************************************************************************/

data ex1223;
infile '\\Client\C$\Users\hb13316\Documents\Data Science\Experimental Stats 1\StatisticalSleuthData
sets_2\CSV\ex1223.csv'
firstobs=2 dlm=',';
input Subject Imagazine Inewspaper Ilibrary MotherEd FatherEd
FamilyIncome78 Race Gender $ Educ Science Arith Word Parag Numer
Coding Auto Math Mechanic Elec AFQT Income2005 Esteem1
Esteem2 Esteem3 Esteem4 Esteem5 Esteem6 Esteem7 Esteem8 Esteem9 Esteem10;

DROP Esteem1 Esteem2 Esteem3 Esteem4 Esteem5 Esteem6 Esteem7 Esteem8 Esteem9 Esteem10;
run;

/*************************************************************************
*************************************************************************

        Subset data to only males and log the response Variable Income as 'loginc05'

*************************************************************************

*************************************************************************/

data proj;
set ex1223;
if Gender='male';
loginc05 = log(Income2005);
run;

/*************************************************************************
*************************************************************************

        Explorative data analytics with the variables within the Final Model.
        No Log of Income to show non Linearity

*************************************************************************

*************************************************************************/

proc sgscatter data=proj;
matrix Income2005 Imagazine FamilyIncome78 Educ Science Arith Numer Auto  / diagonal = (histogram);
TITLE 'Data Diagnostics Matrix with Final Variables - No Log Income';
Footnote 'Figure 1';
run;quit;
```

```
/*************************************************************************
*************************************************************************

        Explorative data analytics with the variables within the Final Model.
                Log of Income to show Linearity (loginc05)

*************************************************************************

*************************************************************************/

/* proc sgscatter data=proj;
matrix loginc05 Imagazine FamilyIncome78 Educ Science Arith Numer Auto  / diagonal = (histogram);
TITLE 'Data Diagnostics Matrix with Final Variables - Log of Income';
Footnote 'Figure 2';
run;quit;

/*************************************************************************
*************************************************************************
```

Data points to be investigated were observations 233, 250 and 633, however, couldn't determine the healthy reason to eliminate the outliers from the model.

significant outliers of the population are:

obs 233 -> Significant Outlier

obs 250 -> Outlier and Leverage

obs 633 -> Significant Outlier and Leverage

```
*************************************************************************

*************************************************************************/

proc reg data=proj corr plots(label)=(RSTUDENTBYLEVERAGE cooksd);
model loginc05 = Imagazine FamilyIncome78 Educ Science Arith Numer Auto;
TITLE 'Diagonistics Breakdown - Regression of Income utilizing Imagazine FamilyIncome78 Educ
Science Arith Numer Auto';
FOOTNOTE 'Figure 3';
run;
quit;

/*************************************************************************
*************************************************************************

        Here we begin to analyze the Residuals for Normality and Independence

*************************************************************************

*************************************************************************/

proc reg data = proj;
model loginc05 = Imagazine FamilyIncome78 Educ Science Arith Numer Auto;
TITLE 'Data Diagnostics - Regression Residuals';
FOOTNOTE 'Figure 4 / Figure5';
run; quit;
```

```
/*************************************************************************
*************************************************************************

                    Conducted VIF analysis and verify as a whole within the dataset
*************************************************************************

*************************************************************************/

proc reg data=proj;
model loginc05 = Imagazine Inewspaper Ilibrary MotherEd FatherEd Educ Science Arith Word Parag
Numer Coding Auto Math Mechanic Elec AFQT/ VIF;
TITLE 'Verify the VIF for all variables';
FOOTNOTE 'Figure 6';
run;quit;

#Options – VIF (Variance Inflation Factor) for multicollinearity (correlation between predictors)


/*************************************************************************
*************************************************************************

                    Conducted LASSO variable selection process
*************************************************************************

*************************************************************************/

proc glmselect data=proj;
model loginc05 = Imagazine Inewspaper Ilibrary MotherEd FatherEd FamilyIncome78 Educ Science Arith
Word Parag Numer Coding Auto Math Mechanic Elec /selection = LASSO (choose = cv stop = aic)
cvmethod = random(5) stats = (adjrsq cp bic sbc sl);
TITLE 'Variable selection with LASSO';
FOOTNOTE 'Figure 7';
run;quit;

#Options –
Variable selection criteria = LASSO
Variable stop criteria = AIC
Break up data = Cross Validation / Method = random in 5 parts
    • Adjusted R2 = linear correlation
    • cp = Mallows C(p)
    • sl = significance level of the F-Stat for entering and exiting effects
    • sbc = Schwarz  Bayesian data measure
    • bic = Sawa Bayesian data measure

/*************************************************************************
*************************************************************************

                    Code 1 - Model Lack of Fit, not used in document, but p-values are referenced
 *************************************************************************

*************************************************************************/

proc reg data=proj;
model loginc05 = Imagazine Inewspaper FatherEd FamilyIncome78 Educ Science Arith Numer Auto;
run;
```

```
/*************************************************************************
*************************************************************************

                        Final Model with VIF and lackfit

 *************************************************************************

*************************************************************************/

proc reg data=proj;
model loginc05 = Imagazine FamilyIncome78 Educ Science Arith Numer Auto/ VIF lackfit;
TITLE 'FINAL Model with lack of fit - Regression of Income utilizing Imagazine FamilyIncome78 Educ
Science Arith Numer Auto';
FOOTNOTE 'Figure 8, Figure 9';
run;quit;

#Options – VIF (Variance Inflation Factor) for multicollinearity (correlation between predictors) and lackfit
for gauging the fitted model
```