APM 630 Regression Analysis
HW6 – Weighted Least Squares
Name: Ruta Basijokaite

Project Report

1. Introduction

A study was conducted to investigate what factors influence perception of what specific acts constitute a crime, as there is considerable variation among individuals in their perception of crime. To get an idea of factors (age and family income) that influence this perception, a sample of 45 college students were given the following list of acts and asked how many of these they perceived as constituting a crime, including: aggravated assault, armed robbery, arson, atheism, auto-theft, burglary, civil disobedience, communism, drug addiction, embezzlement, forcible rape, gambling, homosexuality, land fraud, Nazism, payola, price fixing, prostitution, sexual abuse of child, sex discrimination, shoplifting, striking, strip mining, treason, vandalism, etc. The variables in the study included:

Y = CRIMES – number of items considered as a crime
X1 = AGE – student's age (year)
X2 = INCOME – student's family income ($1000)

The purpose of the study was to (1) apply Ordinary Least Squares model to fit the data, (2) obtain model residuals and conduct a residual analysis, (3) compute the variance of the model, (4) apply Weighted Least Squares model to the data, (5) analyze residuals from WLS model, (6) compare OLS and WLS models in terms of coefficient estimates, standard errors of the coefficient, and residual plots, (7) estimate the asymptotic covariance matrix of the estimates, (8) compare HCCM standard errors of the coefficients against the OLS standard errors of the coefficients.

2. Methods

Simple Ordinary Least Squares (OLS) model was created using Statistical Analysis System (SAS) to fit the data represented in the section 1 of this report. Model was developed based on the following linear regression formula:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \varepsilon \qquad [1]$$

where Y is number of items considered as a crime, β_0 - β_2 are regression coefficients to be estimated, X1 is student's age, X2 is student's family income, and ε is the model error.

The OLS regression method is based on the assumption that the errors are additive, normally distributed, and independent with a mean zero and common variance σ^2. To evaluate linear regression model [1], model residuals were analyzed to test for heterogeneous variance. The assumption of common variance implies that every observation on the dependent variable receives the same weight. The direct impact if heterogeneous variance is a loss of precision in the coefficient estimates. White test was used to test for heterogeneous variance of model residuals. The null hypothesis for this test maintains that the model residuals are homoscedastic, independent of the predictor variables and that several technical assumptions about the model specifications are valid.

The assumption of homogeneous error variance is often violated in practical situations. It is almost common that in scientific measurement, as number becomes larger, the variations of either response variable or predictor variable become larger. In that case, Weighted Least Squares (WLS) are used:

$$Y^* = X^* \beta + \varepsilon^* \quad [2]$$

where
$Y^* = WY$, $X^* = WX$, and $\varepsilon^* = W\varepsilon$. The usual assumption of equal variance is met, and the Ordinary Least Squares (OLS) can be used on Y^* and X^* to produce the weighted lest squares estimator of β:

$$\hat{\beta}_{WLS} = (X'V^{-1}X)^{-1}(X'V^{-1}Y) \quad [3]$$

WLS is equivalent to the OLS applied to the transformed variables to find the solution $\hat{\beta}_{WLS}$ that minimizes $(e^*)'(e^*)=e'V^{-1} e$. The fitted values and the residual on the transformed scale, Y^* and ε^*, are the appropriate quantities to inspect for behavior of the model. The transformation between scales for the fitted values and for the residuals is the same as the original transformation between Y and Y^*.

To explore whether the error variance has a simple relation to income, the observations were divided into three groups (Table 1).

Table 1. Differences between groups

Group	Income ($1000)	Sample Size	Estimated S_e^2	Estimated w_i
1	Under 40	17	2.1215	0.4714
2	Between 40 and 60	17	5.6250	0.1778
3	Over 60	11	16.3006	0.0613

Specific weight was used the same for all observations in the same group.

In the presence of heteroscedasticity, OLS estimates are unbiased, but become inefficient, which leads to inappropriate hypothesis testing. One way to deal with heteroscedasticity is to use Heteroscedasticity Consistent Covariance Matrix (HCCM), which provides a consistent estimator of the covariance matrix of the regression coefficients in the presence of heteroscedasticity of an unknown form, especially for small sample sizes.

3. Results and Discussion

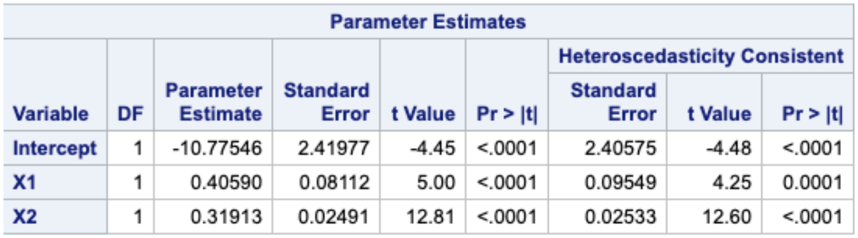
3.1 OLS Model

Equation 1 was edited to the data using least-square method. Model can be represented using the following equation:

$$\hat{Y} = -10.77546 + 0.4059*X1 + 0.31913*X2 \qquad [4]$$

Model produced R^2 value that was equal to 0.8083 indicating that over 80% of the total variation can be explained by this model. MSE was equal to 7.17648. In addition, p-value analysis revealed that slope coefficients (β_1 and β_2) were statistically significant at α=0.05 (Table 2).

Table 2. Estimated regression coefficients for linear model

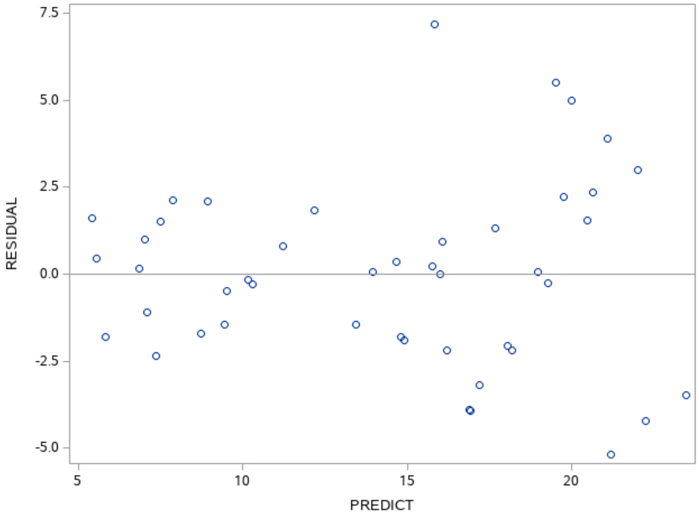


Parameter Estimates								
						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	-10.77546	2.41977	-4.45	<.0001	2.40575	-4.48	<.0001
X1	1	0.40590	0.08112	5.00	<.0001	0.09549	4.25	0.0001
X2	1	0.31913	0.02491	12.81	<.0001	0.02533	12.60	<.0001

SPEC (White) test produced p-value equal to 0.0174, which is lower that α=0.05, indicating that null hypothesis can be rejected and model residuals are not homoscedastic.

Residual plot is illustrated in Figure 1.

Figure 1. Relationship between residual and predicted values.



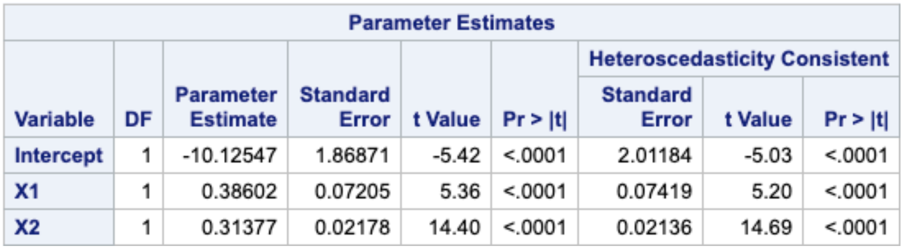
3.2 WLS Model

Equation 2 was edited to fit data using WLS model. Model can be represented using the following equation:

$$\hat{Y} = -10.12547 + 0.38602*X1 + 0.31377*X2 \qquad [5]$$

Model R^2 value was equal to 0.8583. MSE was equal to 1.03945. Slope coefficients were statistically significant at α=0.05 (Table 3).

Table 3. Estimated regression coefficients for WLS model

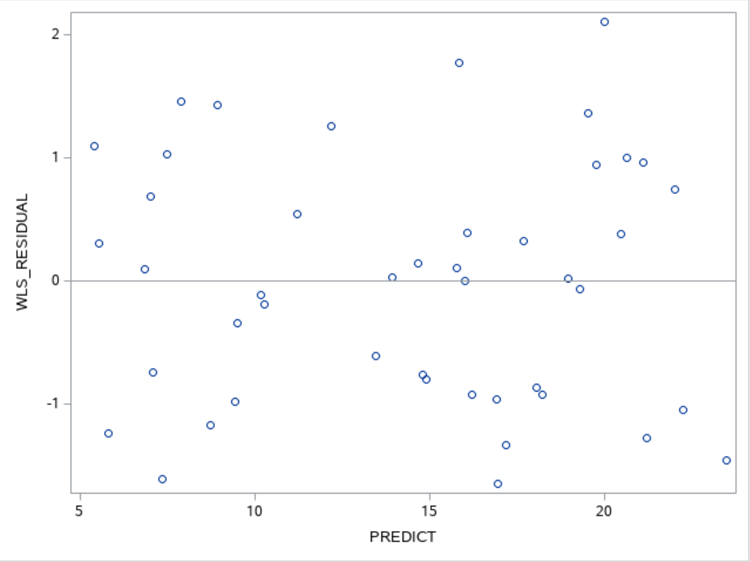


Parameter Estimates								
						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	-10.12547	1.86871	-5.42	<.0001	2.01184	-5.03	<.0001
X1	1	0.38602	0.07205	5.36	<.0001	0.07419	5.20	<.0001
X2	1	0.31377	0.02178	14.40	<.0001	0.02136	14.69	<.0001

SPEC (White) test produced p-value equal to 0.7782, which is higher that α=0.05, indicating that null hypothesis cannot be rejected and model residuals are homoscedastic.

WLS residual plot is illustrated in Figure 2.

Figure 2. Relationship between WLS residual and predicted values.



3.3 Model Comparison

OLS and WLS models produced different coefficient estimates and standard error of the coefficient (Table 3). Coefficient estimates produced by both models were close. β_0 estimate varied the most between the models, but the difference was just under 6%. On the other hand, standard errors of the coefficients varied more noticeably. Also, standard error of β_0 varied the most between the models. Standard error produced by OLD model was equal to 2.41977, while standard error produced by WLS model was 1.86871, which implies 22.8% reduction in standard error for β_0. Standard errors for β_1 and β_2 were also reduced with WLS model.

Table 3. OLS, WLS and HC model comparison

Coefficients	**OLS Model**		**WLS Model**		**HC**
	Estimate	**Ste**	**Estimate**	**Ste**	**Ste**
β_0	-10.77546	2.41977	-10.12547	1.86871	2.40575
β_1	0.4059	0.08112	0.38602	0.07205	0.09549
β_2	0.31913	0.02491	0.31377	0.02178	0.02533

Residual plots produced by OLS model (Figure 1) and WLS model (Figure 2) were different. Residual variance was more homogeneous with WLS model as compared to OLS model. Also, this was confirmed using SPEC command in SAS, where OLD model was concluded as non-homoscedastic, since null hypothesis (stating that the model residuals are homoscedastic, independent of the predictor variables and that several technical assumptions about the model specifications are valid) was rejected. On the other hand, residuals produced by WLS model, passed the heterogeneous variance test and WLS model residuals were concluded as homoscedastic.

The HCCM standard errors of the coefficients and OLS model produced standard errors were different (Table 3). Despite the fact that standard errors were obtained from the asymptotic covariance matrix, HC produced standard errors were not always smaller that standard errors produced by OLS model. Standard error for β_0 was smaller with HC, compared to OLD model. However, the two other standard errors for and were larger with HC. This indicates that HC was not able to deal with collection of errors in OLS model. HC only improves standard error values if it is able to deal with a collection of minor concerns about the failure to meet assumptions, in this case problems with heteroscedasticity.

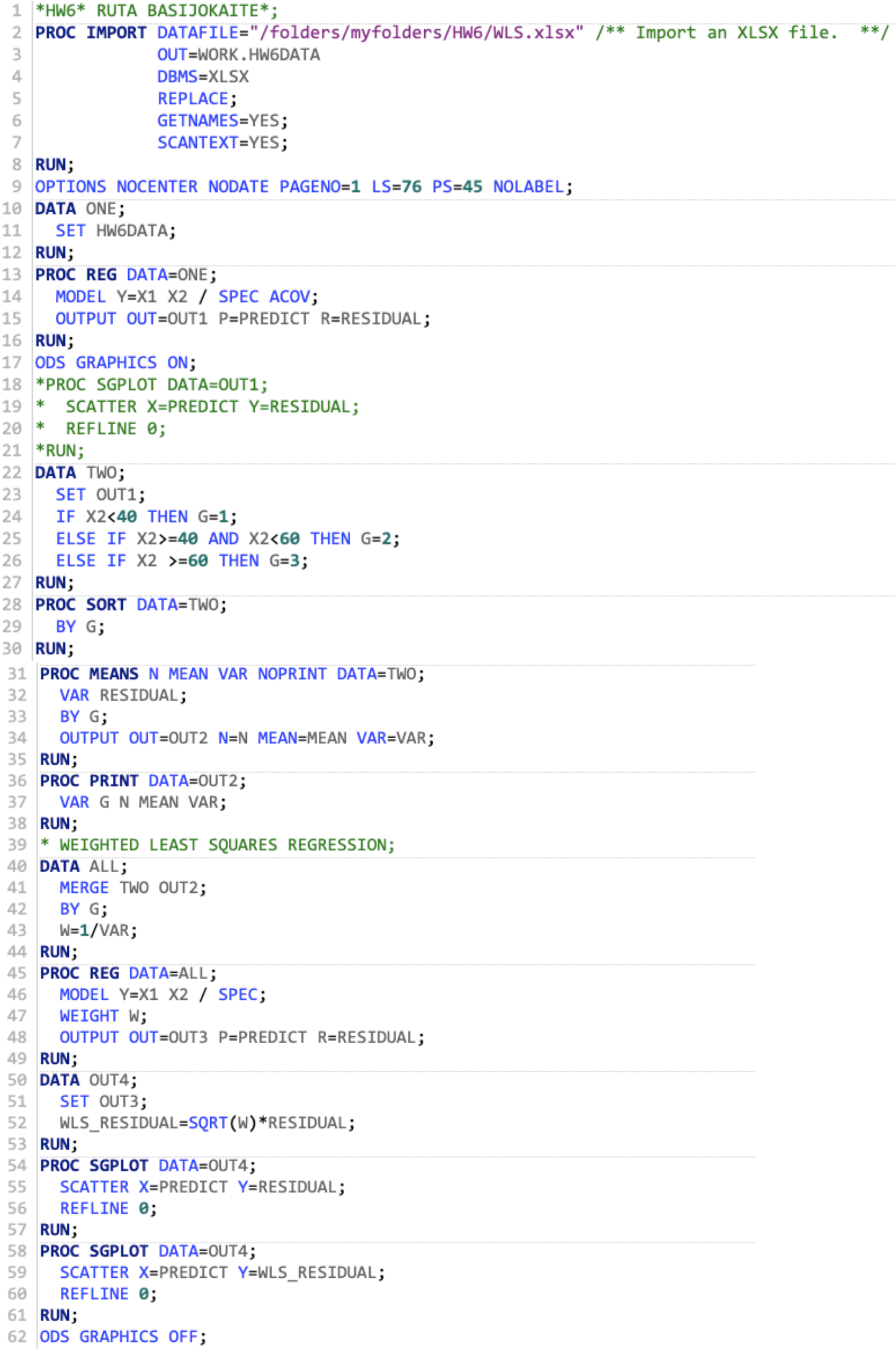
4. Summary

Relationship between perception of a crime and two factors (student's age and family income) was analyzed. Answers were collected from 45 students to develop a linear relationship between variables. First, linear regression model was created. Then linear regression model was modified to include weighted least squares. For that, data was divided into three groups based on student's family income, and new linear model was created giving each observation in a group the same weight. Results revealed that WLS model performed better than OLS model. Produced R^2 value was higher in WLS model. Also, residual analysis revealed that OLS model residuals were not homoscedastic, whereas this problem was resolved using WLS model. Although both models produced similar coefficient values, standard errors of coefficients were reduced as much as 22.8% in WLS model. Asymptotic covariance matrix estimated standard errors were not higher in

2 out of 3 coefficients compared to OLS model standard errors indicating that heteroscedasticity was not resolved using HCCM.

Further analysis needs to be performed to improve WLS model even further. For example, instead of assigning the same weight to the whole group, individual weight could be assigned for each observation. This would likely improve the model as assigned weight are calculated to fit each observation. In addition, more robust residual analysis could be done to identify outliers that are likely influencing the model.

5. SAS programs



```
*HW6* RUTA BASIJOKAITE*;
PROC IMPORT DATAFILE="/folders/myfolders/HW6/WLS.xlsx" /** Import an XLSX file.  **/
            OUT=WORK.HW6DATA
            DBMS=XLSX
            REPLACE;
            GETNAMES=YES;
            SCANTEXT=YES;
RUN;
OPTIONS NOCENTER NODATE PAGENO=1 LS=76 PS=45 NOLABEL;
DATA ONE;
  SET HW6DATA;
RUN;
PROC REG DATA=ONE;
  MODEL Y=X1 X2 / SPEC ACOV;
  OUTPUT OUT=OUT1 P=PREDICT R=RESIDUAL;
RUN;
ODS GRAPHICS ON;
*PROC SGPLOT DATA=OUT1;
*  SCATTER X=PREDICT Y=RESIDUAL;
*  REFLINE 0;
*RUN;
DATA TWO;
  SET OUT1;
  IF X2<40 THEN G=1;
  ELSE IF X2>=40 AND X2<60 THEN G=2;
  ELSE IF X2 >=60 THEN G=3;
RUN;
PROC SORT DATA=TWO;
  BY G;
RUN;
PROC MEANS N MEAN VAR NOPRINT DATA=TWO;
  VAR RESIDUAL;
  BY G;
  OUTPUT OUT=OUT2 N=N MEAN=MEAN VAR=VAR;
RUN;
PROC PRINT DATA=OUT2;
  VAR G N MEAN VAR;
RUN;
* WEIGHTED LEAST SQUARES REGRESSION;
DATA ALL;
  MERGE TWO OUT2;
  BY G;
  W=1/VAR;
RUN;
PROC REG DATA=ALL;
  MODEL Y=X1 X2 / SPEC;
  WEIGHT W;
  OUTPUT OUT=OUT3 P=PREDICT R=RESIDUAL;
RUN;
DATA OUT4;
  SET OUT3;
  WLS_RESIDUAL=SQRT(W)*RESIDUAL;
RUN;
PROC SGPLOT DATA=OUT4;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT4;
  SCATTER X=PREDICT Y=WLS_RESIDUAL;
  REFLINE 0;
RUN;
ODS GRAPHICS OFF;
```