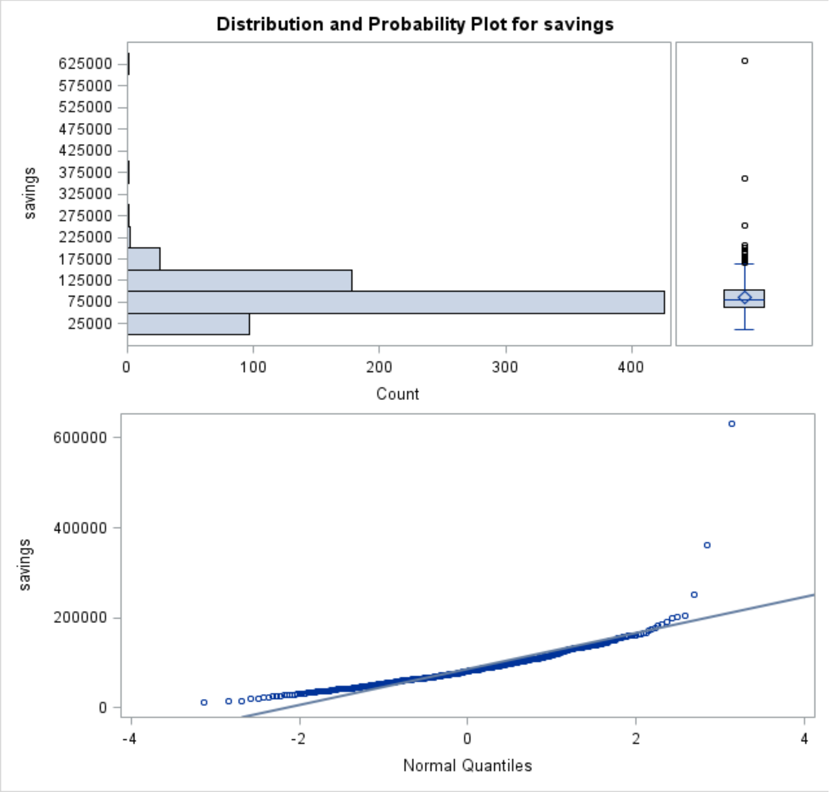
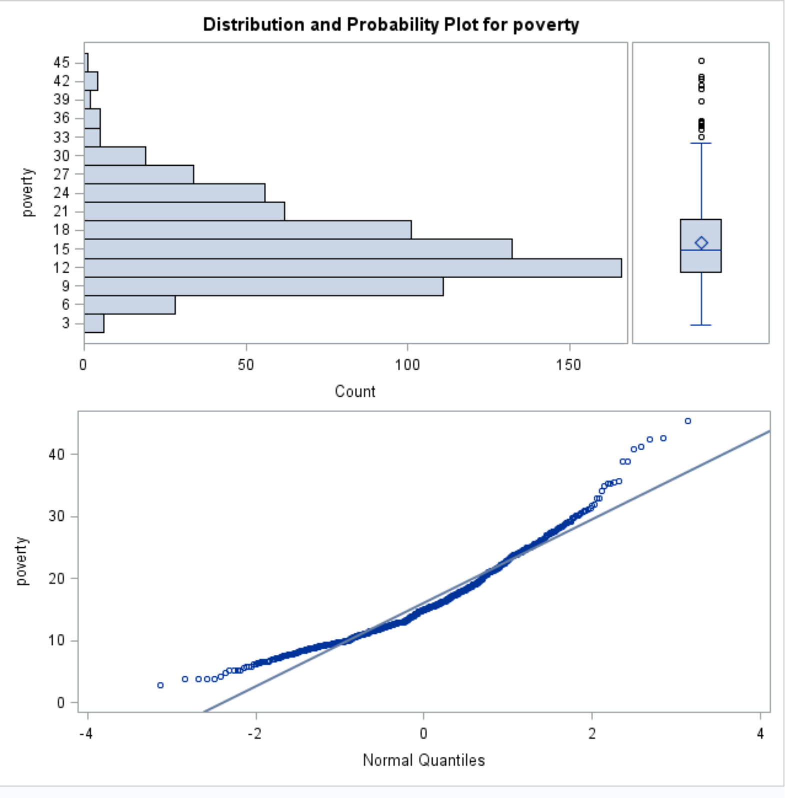
Data Exploration And Multiple Linear Regression Using SAS

Kevin Ovendorf

DSBA 6201

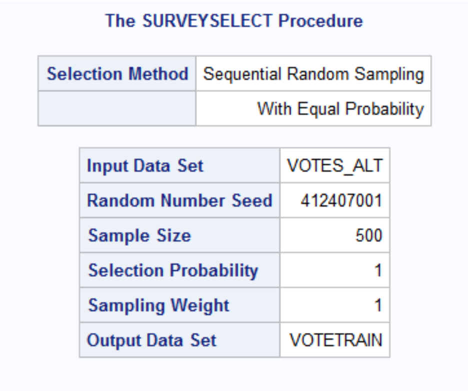
1. Generate box-plots of the savings (Mean Savings in $) and poverty (% in poverty) attributes and identify/removethe cutoff values for outliers.


Distribution and Probability Plot for savings


Distribution and Probability Plot for poverty

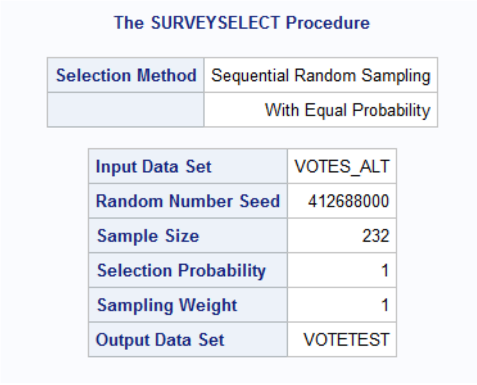
Savings Outlier: Q3 + 1.5 IQR → 103658.5 + 1.5(41300) = 165608.5
Poverty Outlier: Q3 + 1.5 IQR → 19.75 + 1.5(8.55) = 32.575

2. Try to fit an MLR to this dataset, with VOTES as thedependent variable. INCOME has somewhat longish tail, so we will take a log transform, (use LINCOME = log(INCOME)) and then use LINCOME as one of predictor. Keep the first 500 records as a training set (call it VOTETRAIN) which you will use to fit the model; the remaining 232 will be used as a test set (VOTETEST). Use only the following variables in your model:
VOTES =LINCOME + SAVINGS + FEMALE +DENSITY +POVERTY + VETERANS


The SURVEYSELECT Procedure

Selection Method	Sequential Random Sampling
	With Equal Probability

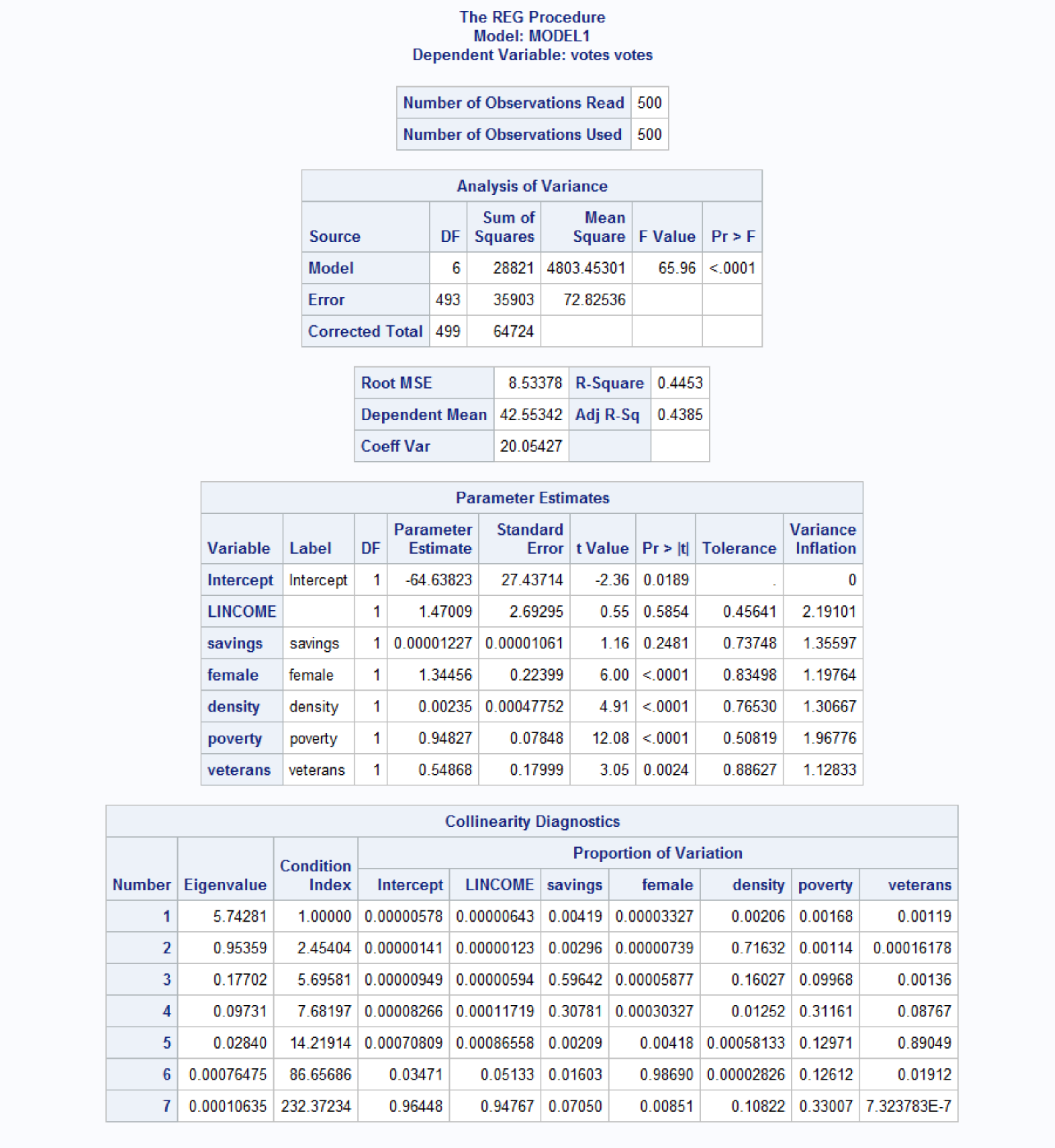
Input Data Set	VOTES_ALT
Random Number Seed	412407001
Sample Size	500
Selection Probability	1
Sampling Weight	1
Output Data Set	VOTETRAIN


The SURVEYSELECT Procedure

Selection Method	Sequential Random Sampling
	With Equal Probability

Input Data Set	VOTES_ALT
Random Number Seed	412688000
Sample Size	232
Selection Probability	1
Sampling Weight	1
Output Data Set	VOTETEST

(a) Report the coefficients obtained by your model. Would you drop any of the variables used in your model (based on the t-scores or p-values)?



The REG Procedure
Model: MODEL1
Dependent Variable: votes votes

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	28821	4803.45301	65.96	<.0001
Error	493	35903	72.82536		
Corrected Total	499	64724			

Root MSE	8.53378	R-Square	0.4453
Dependent Mean	42.55342	Adj R-Sq	0.4385
Coeff Var	20.05427		

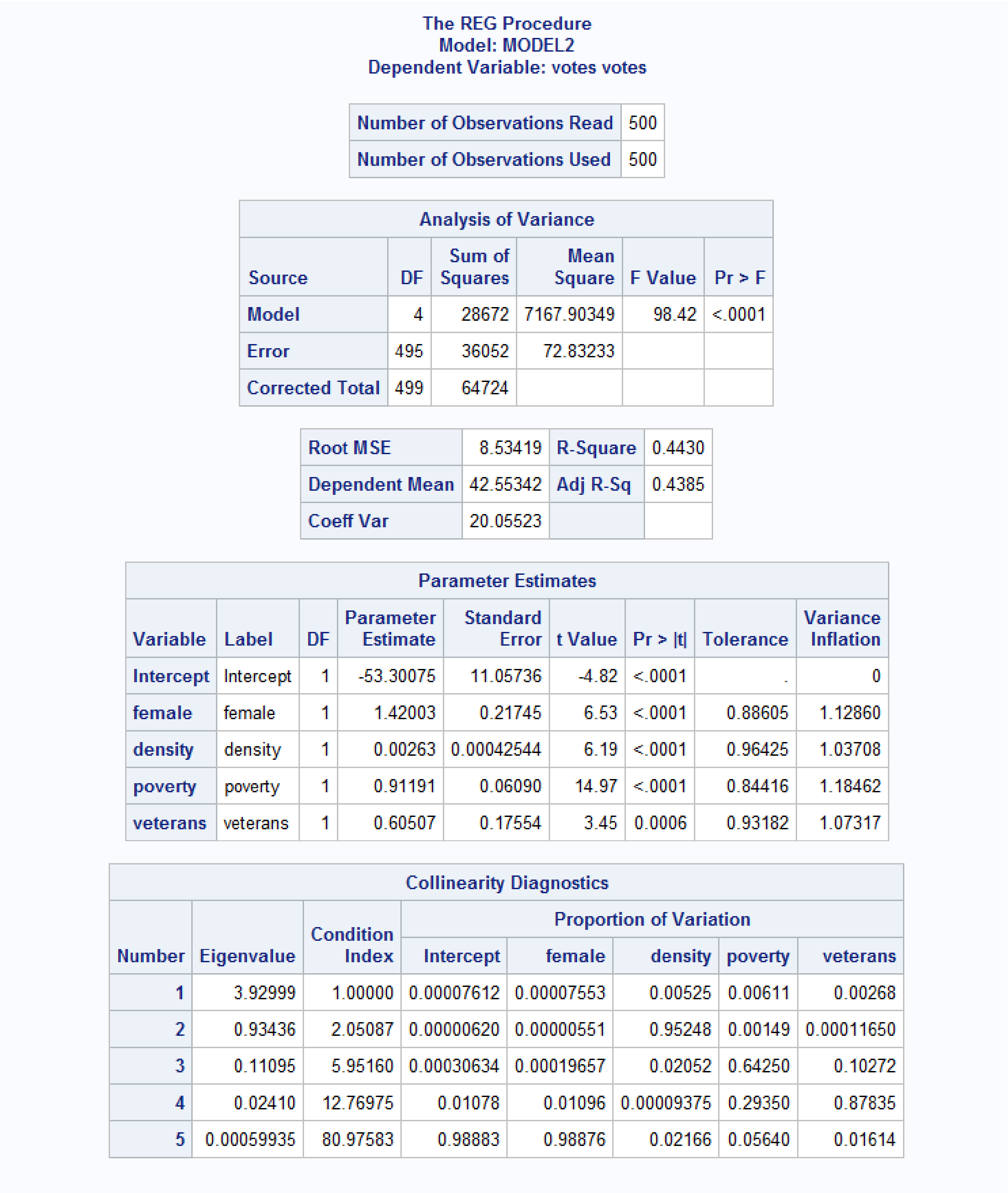
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-64.63823	27.43714	-2.36	0.0189	.	0
LINCOME		1	1.47009	2.69295	0.55	0.5854	0.45641	2.19101
savings	savings	1	0.00001227	0.00001061	1.16	0.2481	0.73748	1.35597
female	female	1	1.34456	0.22399	6.00	<.0001	0.83498	1.19764
density	density	1	0.00235	0.00047752	4.91	<.0001	0.76530	1.30667
poverty	poverty	1	0.94827	0.07848	12.08	<.0001	0.50819	1.96776
veterans	veterans	1	0.54868	0.17999	3.05	0.0024	0.88627	1.12833

Collinearity Diagnostics

			Proportion of Variation						
Number	Eigenvalue	Condition Index	Intercept	LINCOME	savings	female	density	poverty	veterans
1	5.74281	1.00000	0.00000578	0.00000643	0.00419	0.00003327	0.00206	0.00168	0.00119
2	0.95359	2.45404	0.00000141	0.00000123	0.00296	0.00000739	0.71632	0.00114	0.00016178
3	0.17702	5.69581	0.00000949	0.00000594	0.59642	0.00005877	0.16027	0.09968	0.00136
4	0.09731	7.68197	0.00008266	0.00011719	0.30781	0.00030327	0.01252	0.31161	0.08767
5	0.02840	14.21914	0.00070809	0.00086558	0.00209	0.00418	0.00058133	0.12971	0.89049
6	0.00076475	86.65686	0.03471	0.05133	0.01603	0.98690	0.00002826	0.12612	0.01912
7	0.00010635	232.37234	0.96448	0.94767	0.07050	0.00851	0.10822	0.33007	7.323783E-7

All of the coefficients obtained are within the parameter estimate. Based on Pr > |t| of a value greater than 0.05 we can drop savings and Lincome because they have a confidence interval of less than 95%. So of course the model is ran again and here are the following results below:



The REG Procedure
Model: MODEL2
Dependent Variable: votes votes

Number of Observations Read	500
Number of Observations Used	500

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	28672	7167.90349	98.42	<.0001
Error	495	36052	72.83233		
Corrected Total	499	64724			

Root MSE	8.53419	R-Square	0.4430
Dependent Mean	42.55342	Adj R-Sq	0.4385
Coeff Var	20.05523		

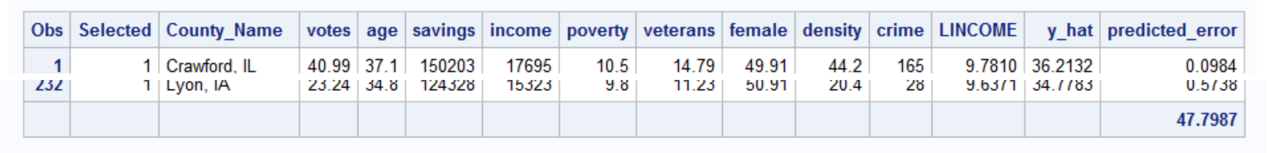
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-53.30075	11.05736	-4.82	<.0001	.	0
female	female	1	1.42003	0.21745	6.53	<.0001	0.88605	1.12860
density	density	1	0.00263	0.00042544	6.19	<.0001	0.96425	1.03708
poverty	poverty	1	0.91191	0.06090	14.97	<.0001	0.84416	1.18462
veterans	veterans	1	0.60507	0.17554	3.45	0.0006	0.93182	1.07317

Collinearity Diagnostics

			Proportion of Variation				
Number	Eigenvalue	Condition Index	Intercept	female	density	poverty	veterans
1	3.92999	1.00000	0.00007612	0.00007553	0.00525	0.00611	0.00268
2	0.93436	2.05087	0.00000620	0.00000551	0.95248	0.00149	0.00011650
3	0.11095	5.95160	0.00030634	0.00019657	0.02052	0.64250	0.10272
4	0.02410	12.76975	0.01078	0.01096	0.00009375	0.29350	0.87835
5	0.00059935	80.97583	0.98883	0.98876	0.02166	0.05640	0.01614

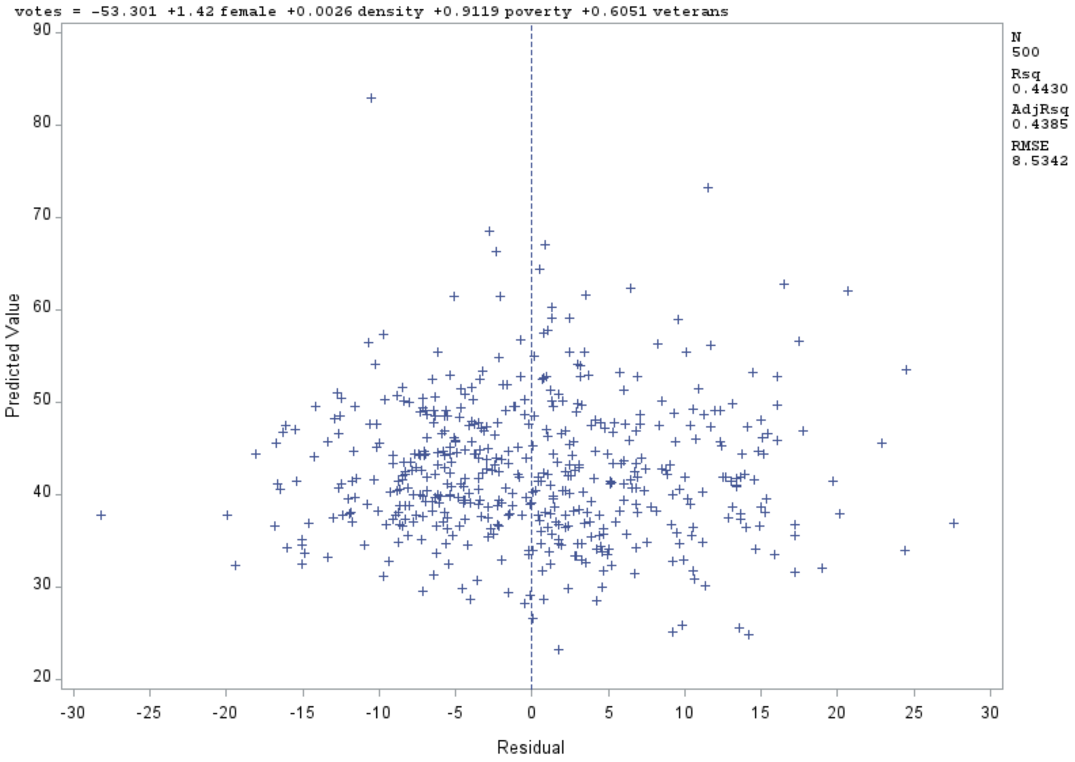
(b) Report the MSE obtained on VOTETRAIN. How much does this increase when you score your model on VOTETEST?



Obs	Selected	County_Name	votes	age	savings	income	poverty	veterans	female	density	crime	LINCOME	y_hat	predicted_error
1	1	Crawford, IL	40.99	37.1	150203	17695	10.5	14.79	49.91	44.2	165	9.7810	36.2132	0.0984
232	1	Lyon, IA	23.24	34.8	124328	15323	9.8	11.23	50.91	20.4	28	9.6371	34.7783	0.5738
														47.7987

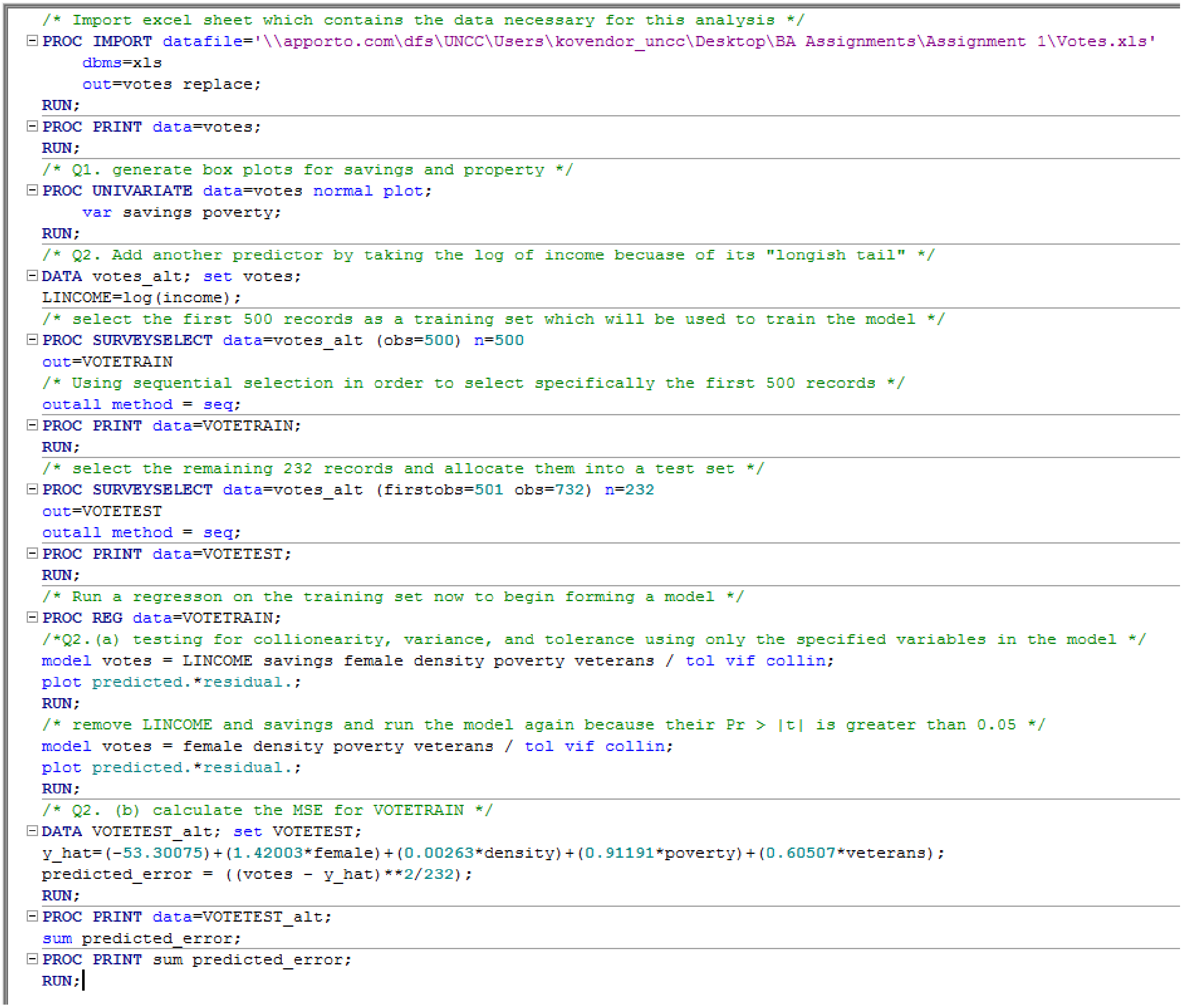
VOTETRAIN MSE: 72.8254 VOTETEST MSE: 47.7987
So there is a decrease in the MSE of 25.0267

(c) (Bonus 2 points). Do you think your MLR model is reasonable for this problem? You may look at the distribution of residuals to provide an informed answer.



Based upon the residual plot I believe this MLR is reasonable for this problem, there is an ever so slight trend on the upper right hand of the graph but overall it is uneven and cloud like and random. Also another valid reason for this being a reasonable MLR is that the adjusted R-squared: 0.4385, only has a difference of .0045 from the R squared value: 0.4430.

Not sure if I am supposed to also share my code through this assignment as it is not stated but I am doing so nonetheless just in case.



```
/* Import excel sheet which contains the data necessary for this analysis */
PROC IMPORT datafile='\\apporto.com\dfs\UNCC\Users\kovendor_uncc\Desktop\BA Assignments\Assignment 1\Votes.xls'
    dbms=xls
    out=votes replace;
RUN;
PROC PRINT data=votes;
RUN;
/* Q1. generate box plots for savings and property */
PROC UNIVARIATE data=votes normal plot;
    var savings poverty;
RUN;
/* Q2. Add another predictor by taking the log of income becuase of its "longish tail" */
DATA votes_alt; set votes;
LINCOME=log(income);
/* select the first 500 records as a training set which will be used to train the model */
PROC SURVEYSELECT data=votes_alt (obs=500) n=500
out=VOTETRAIN
/* Using sequential selection in order to select specifically the first 500 records */
outall method = seq;
PROC PRINT data=VOTETRAIN;
RUN;
/* select the remaining 232 records and allocate them into a test set */
PROC SURVEYSELECT data=votes_alt (firstobs=501 obs=732) n=232
out=VOTETEST
outall method = seq;
PROC PRINT data=VOTETEST;
RUN;
/* Run a regresson on the training set now to begin forming a model */
PROC REG data=VOTETRAIN;
/*Q2.(a) testing for collionearity, variance, and tolerance using only the specified variables in the model */
model votes = LINCOME savings female density poverty veterans / tol vif collin;
plot predicted.*residual.;
RUN;
/* remove LINCOME and savings and run the model again because their Pr > |t| is greater than 0.05 */
model votes = female density poverty veterans / tol vif collin;
plot predicted.*residual.;
RUN;
/* Q2. (b) calculate the MSE for VOTETRAIN */
DATA VOTETEST_alt; set VOTETEST;
y_hat=(-53.30075)+(1.42003*female)+(0.00263*density)+(0.91191*poverty)+(0.60507*veterans);
predicted_error = ((votes - y_hat)**2/232);
RUN;
PROC PRINT data=VOTETEST_alt;
sum predicted_error;
PROC PRINT sum predicted_error;
RUN;
```