Arizona State University

MSE Software Engineering

IEE 578 Applied Regression Analysis

Final Project

John Raphael Fox

Abstract

In the following paper, we present to you a linear regression model that tries to identify if mortality rate due to lung cancer is a function of income. We attempt to gain knowledge from different models that involve variables such as interest dividends and government subsidies. We analyze each variable individually, and the relation between each of them. An analysis of *collinearity* between each of the regressors is presented, and the identification of *influential* or *leverage* points is studied. We hope that the reasons for the decisions taken in this text are sufficient so that the reader concurs with our findings.

Does Income Help Survive Cancer?

In 2016, it was estimated that 1,685,210 new cases of cancer would be diagnosed in the United States and 595,690 people would die from the disease (Institute, 2018). This includes many types of cancer: breast cancer, leukemia, lung and bronchus cancer, pancreatic cancer, etc.

It is known that cancer mortality rate is higher among men than women (207.9 per 100,000 men and 145.4 per 100,00 women) (Institute, 2018). Race also plays a role in the susceptibility of obtaining cancer; it is highest in African American men (261.5 per 100,000) and lowest in Asian/Pacific Islanders women (91.5 per 100,000) (Institute, 2018). What if other variables besides gender and race exist, that may influence the mortality rate due to cancer? The primary roll of this paper is to analyze and discuss whether other possible factors such as *economic status* and/or *location* within the United States affects mortality rate due to cancer.

Two different resources were used to obtain the required data sets: National Cancer Institute (NIH) and the U.S. Census Bureau. For our **response** variable we downloaded, from the NIH, the **incidence** and **mortality rates** of Lung and Bronchus cancer datasets. This specific type of cancer was selected because of regulations on health data. After the cleaning, preparation, and exploration part of the process was finished, the decision to eliminate the incidence dataset was made. This was because of the large amount of data that was missing.

The US Census Bureau in partnership with data.world allowed an easy way of obtaining and querying the **regressor** variables that were used in the analysis. Since we want to relate lung cancer mortality rate with economic status, the following data was used:

1. All_Poverty – This was selected for the obvious reasons.
2. Med_Income_With_Interest_Dividends – If you are earning interest dividends, this probably means you are not considered poor.
3. Med_Income_With_SS – Are people that besides having an income, also have social security, which is more than $1,000 per month.

4. Med_Income_With_SSI – Are people with an income and Supplemental Social Security which is approximately $750 per month.
5. Med_Income_With_Public_Assitance – These are people that are assisted by the government with cash or vouchers.
6. Med_Income_With_Public_Assis_Food_Stamps – People that are assisted by the government with food vouchers
7. Retirement_Income – People that are retired.

The data used for the regressor are thought to include the information necessary to span from low income to higher income. It is important to note that the regressors are normalized per 100,000 people.

By using the Federal Information Processing Standards (FIPS) county codes, we were able to merge the datasets. This information also allowed us to easily extract the state in which each county was located, and with this, we constructed three **indicator** variables that dealt with the regions of the United States: West, Mid-West, and South; the North-East is implicit in the analysis.

The GitHub and data.world links are included in the reference (Fox, GitHub, 2018) (Fox, data.world, 2018). The first contains the Jupyter Notebook code that was utilized for the data gathering and cleansing process, and the second, the data that was used for the analysis. JMP was employed for the regression analysis.

Analysis

In the following section we will walk you through the analysis, and the logic behind the choices that were made. Like George Box said, "All models are wrong" (I read it in a comment posted on the discussion board), but we tried our best to choose an appropriate one.

Model Selection Process

The analysis started by viewing the regressors individually and the relation between them; this is illustrated in Appendix A. For the most part, the regressors seem to be normally distributed; some may seem a little skewed, but they conserve the same basic form. (See figure A.1) The *correlation matrix* seems to show a bit of *multicollinearity* between SSI and Food_Stamps, with a value of 0.8134. (See table A.1) This can be confirmed by how narrow the density ellipses are between these two variables in the *scatter plot matrix*. (See figure A.2)

We then fit a 1st order linear regression model with all the variables. (See Appendix B) The analysis of this model indicates that the parameters are significant. Furthermore, the residual analysis indicates that there may be constant variance (see figure B.1) and that the assumption of normality is correct (see figure B.2). Figures B.1 and B.2, also indicate that a *transformation* for the regressors or response variables is not required. It is important to notice, that the objective of this paper is to demonstrate that cancer mortality depends on income. That said, contrary to our belief, the sign of the parameter related to the poverty regressor is negative; which indicates negative correlation with mortality rate. This might be a consequence of multicollinearity being present in the dataset. Also, the parameters are small because of the normalization; they are normalized per 100,000 people. Possible influential and leverage points are also labeled on the residual plots. (See figures B.1 and B.2)

The whole process of model selection can be seen by looking over Appendices B-H. By using the mixed stepwise regression option (with $\alpha = 0.05$) in JMP; we decided to use a 2nd order polynomial regression model. For this data we selected a cutoff value for the VIF of 5. This will enable us to assume multicollinearity between All_Poverty, SSI, and Food_Stamps. The decision was made because of the authors life experience. So, we decided to eliminate SSI and Food_Stamps from the analysis. Observations 9 and 10 were also eliminated from the analysis. The reasons behind this were: their $h_{ii} > \frac{2p}{n}$ and that their Cook's D's are 3 orders of magnitude greater that the mean of all the Cook's D's. To further justify this decision, data regarding mortality rate and population was investigated. They had a mortality rate of 103.2 and 132.5 per 100,000 people, and population of 9846 and 9687, respectively. Which seems highly unlikely.

Residual Analysis

The overall analysis of the residual plots indicated that there seems to be *constant variance* and the assumption of *normality* is correct. Also, no transformations were required. The residual plots also helped us identify that observations 9 and 10 were outliers. With the decision that were taken, the residual plots even evolved to something more esthetic.

Conclusions

The signs and magnitudes of the final parameters indicate that cancer mortality does depend on income and even age (when referring to retirement income). People with public assistance are more likely to die from cancer than others, and people with interest dividend are less likely to die from this disease. The reason behind this might be the capability of acquiring state of the art treatments and better physicians. Finally, region is not a factor in cancer mortality.

Recommendations

It is recommended to further investigate the matter at hand. Creating a model that utilizes a piece of the data as training data, and the remainder as test data would be useful to *validate* the model. Using other available data for the regressors would be also necessary. We believed that 4 of the 7 regressors could be considered low income, in other words, they could be related. For instance, in the US if you are low income or you qualify for government help, it is very likely you can obtain food stamps.

References

Fox, J. (2018). *data.world*. Retrieved from data.world: https://data.world/johnfox/cancer-analysis-hackathon-challenge

Fox, J. (2018). *GitHub*. Retrieved from GitHub: https://github.com/johnfox17/IEE-578-Reg-Analysis-Project

Institute, N. C. (2018). *National Cancer Institute*. Retrieved from National Cancer Institute: https://www.cancer.gov/about-cancer/understanding/statistics

Appendix A

Initial analysis

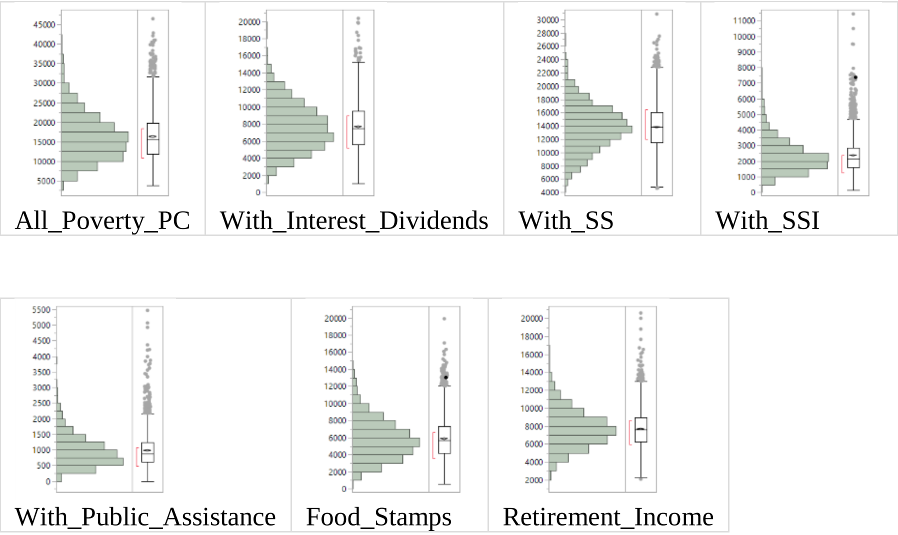
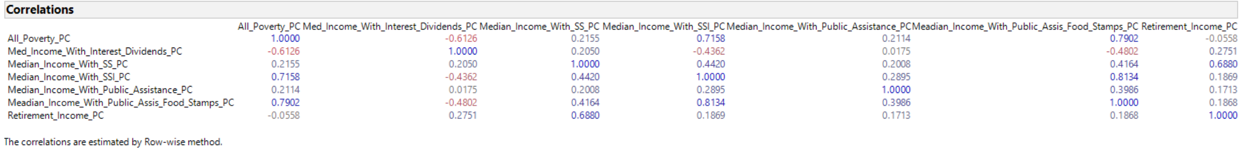


Figure A.1 – Histograms of each of the regressors.



Correlations

	All_Poverty_PC	Med_Income_With_Interest_Dividends_PC	Median_Income_With_SS_PC	Median_Income_With_SSI_PC	Median_Income_With_Public_Assistance_PC	Meadian_Income_With_Public_Assis_Food_Stamps_PC	Retirement_Income_PC
All_Poverty_PC	1.0000	-0.6126	0.2155	0.7158	0.2114	0.7902	-0.0558
Med_Income_With_Interest_Dividends_PC	-0.6126	1.0000	0.2050	-0.4362	0.0175	-0.4802	0.2751
Median_Income_With_SS_PC	0.2155	0.2050	1.0000	0.4420	0.2008	0.4164	0.6880
Median_Income_With_SSI_PC	0.7158	-0.4362	0.4420	1.0000	0.2895	0.8134	0.1869
Median_Income_With_Public_Assistance_PC	0.2114	0.0175	0.2008	0.2895	1.0000	0.3986	0.1713
Meadian_Income_With_Public_Assis_Food_Stamps_PC	0.7902	-0.4802	0.4164	0.8134	0.3986	1.0000	0.1868
Retirement_Income_PC	-0.0558	0.2751	0.6880	0.1869	0.1713	0.1868	1.0000

The correlations are estimated by Row-wise method.

Table A.1 – Correlation matrix.

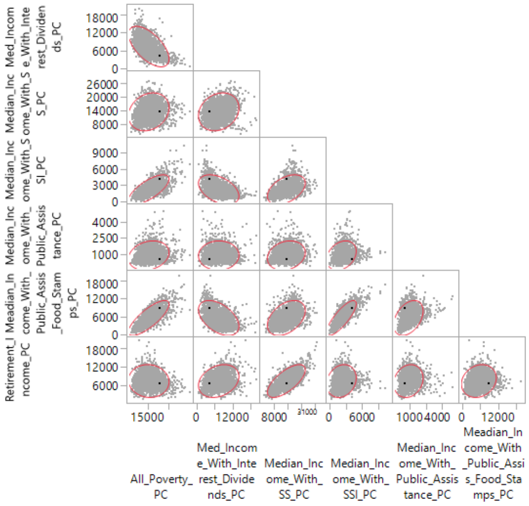
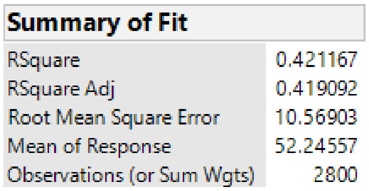


Figure A.2 – Scatter plot matrix.

Appendix B

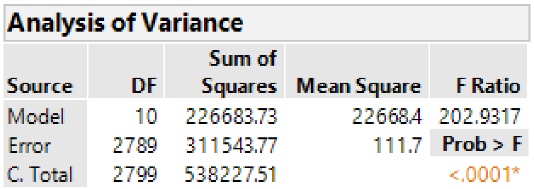
First order polynomial regression with all regressors and indicator variables.



Summary of Fit

RSquare	0.421167
RSquare Adj	0.419092
Root Mean Square Error	10.56903
Mean of Response	52.24557
Observations (or Sum Wgts)	2800

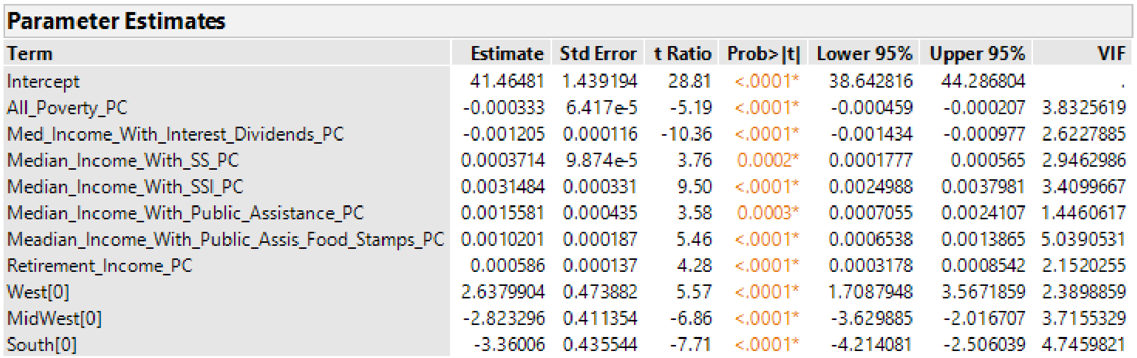
Table B.1 – Summary of fit 1st order polynomial linear regression.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	10	226683.73	22668.4	202.9317
Error	2789	311543.77	111.7	Prob > F
C. Total	2799	538227.51		<.0001*

Table B.2 – Analysis of variance 1st order polynomial linear regression.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	41.46481	1.439194	28.81	<.0001*	38.642816	44.286804	.
All_Poverty_PC	-0.000333	6.417e-5	-5.19	<.0001*	-0.000459	-0.000207	3.8325619
Med_Income_With_Interest_Dividends_PC	-0.001205	0.000116	-10.36	<.0001*	-0.001434	-0.000977	2.6227885
Median_Income_With_SS_PC	0.0003714	9.874e-5	3.76	0.0002*	0.0001777	0.000565	2.9462986
Median_Income_With_SSI_PC	0.0031484	0.000331	9.50	<.0001*	0.0024988	0.0037981	3.4099667
Median_Income_With_Public_Assistance_PC	0.0015581	0.000435	3.58	0.0003*	0.0007055	0.0024107	1.4460617
Meadian_Income_With_Public_Assis_Food_Stamps_PC	0.0010201	0.000187	5.46	<.0001*	0.0006538	0.0013865	5.0390531
Retirement_Income_PC	0.000586	0.000137	4.28	<.0001*	0.0003178	0.0008542	2.1520255
West[0]	2.6379904	0.473882	5.57	<.0001*	1.7087948	3.5671859	2.3898859
MidWest[0]	-2.823296	0.411354	-6.86	<.0001*	-3.629885	-2.016707	3.7155329
South[0]	-3.36006	0.435544	-7.71	<.0001*	-4.214081	-2.506039	4.7459821

Table B.3 – Parameter Estimates 1st order polynomial liner regression.

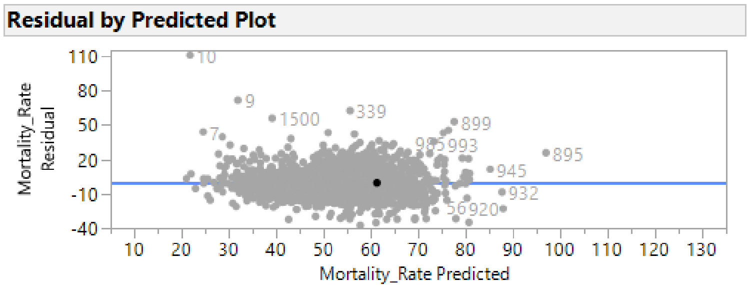


Figure B.1 – Residual by predicted plot 1st order polynomial linear regression.

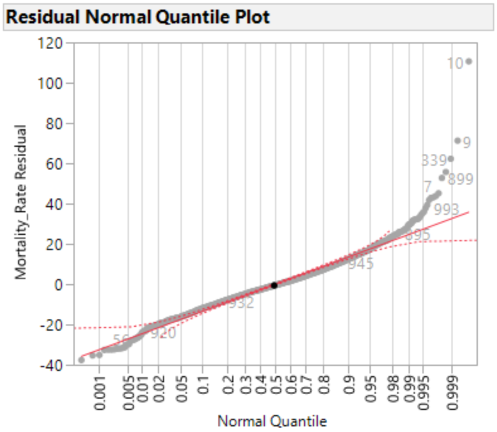
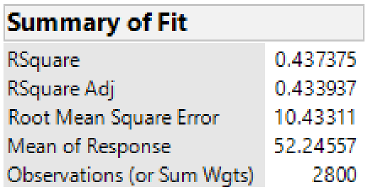


Figure B.2 – Residual normal quantile plot 1st order polynomial linear regression.

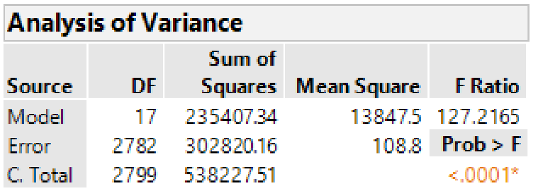
Appendix C

Second order polynomial regression with all regressors.



Summary of Fit	
RSquare	0.437375
RSquare Adj	0.433937
Root Mean Square Error	10.43311
Mean of Response	52.24557
Observations (or Sum Wgts)	2800

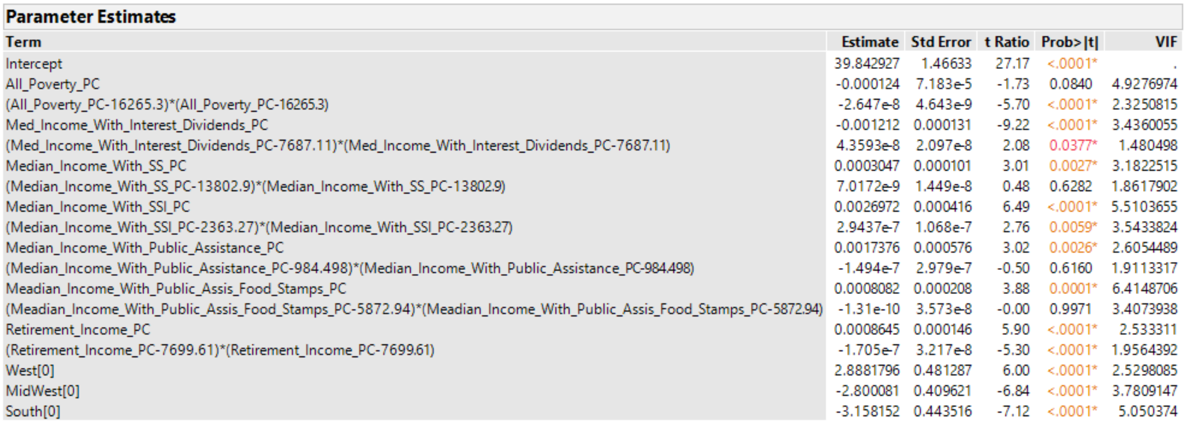
Table C.1 – Summary of fit 2nd order polynomial linear regression.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	17	235407.34	13847.5	127.2165
Error	2782	302820.16	108.8	Prob > F
C. Total	2799	538227.51		<.0001*

Table C.2 – Analysis of variance 2nd order polynomial linear regression.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	39.842927	1.46633	27.17	<.0001*	.
All_Poverty_PC	-0.000124	7.183e-5	-1.73	0.0840	4.9276974
(All_Poverty_PC-16265.3)*(All_Poverty_PC-16265.3)	-2.647e-8	4.643e-9	-5.70	<.0001*	2.3250815
Med_Income_With_Interest_Dividends_PC	-0.001212	0.000131	-9.22	<.0001*	3.4360055
(Med_Income_With_Interest_Dividends_PC-7687.11)*(Med_Income_With_Interest_Dividends_PC-7687.11)	4.3593e-8	2.097e-8	2.08	0.0377*	1.480498
Median_Income_With_SS_PC	0.0003047	0.000101	3.01	0.0027*	3.1822515
(Median_Income_With_SS_PC-13802.9)*(Median_Income_With_SS_PC-13802.9)	7.0172e-9	1.449e-8	0.48	0.6282	1.8617902
Median_Income_With_SSI_PC	0.0026972	0.000416	6.49	<.0001*	5.5103655
(Median_Income_With_SSI_PC-2363.27)*(Median_Income_With_SSI_PC-2363.27)	2.9437e-7	1.068e-7	2.76	0.0059*	3.5433824
Median_Income_With_Public_Assistance_PC	0.0017376	0.000576	3.02	0.0026*	2.6054489
(Median_Income_With_Public_Assistance_PC-984.498)*(Median_Income_With_Public_Assistance_PC-984.498)	-1.494e-7	2.979e-7	-0.50	0.6160	1.9113317
Meadian_Income_With_Public_Assis_Food_Stamps_PC	0.0008082	0.000208	3.88	0.0001*	6.4148706
(Meadian_Income_With_Public_Assis_Food_Stamps_PC-5872.94)*(Meadian_Income_With_Public_Assis_Food_Stamps_PC-5872.94)	-1.31e-10	3.573e-8	-0.00	0.9971	3.4073938
Retirement_Income_PC	0.0008645	0.000146	5.90	<.0001*	2.533311
(Retirement_Income_PC-7699.61)*(Retirement_Income_PC-7699.61)	-1.705e-7	3.217e-8	-5.30	<.0001*	1.9564392
West[0]	2.8881796	0.481287	6.00	<.0001*	2.5298085
MidWest[0]	-2.800081	0.409621	-6.84	<.0001*	3.7809147
South[0]	-3.158152	0.443516	-7.12	<.0001*	5.050374

Table C.3 – Parameter Estimates 2nd order polynomial liner regression.

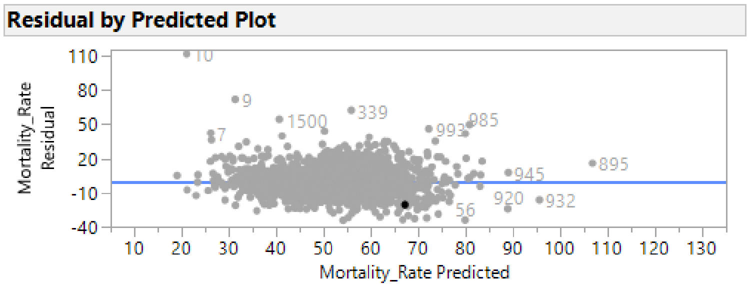


Figure C.1 – Residual by predicted plot 2nd order polynomial linear regression.

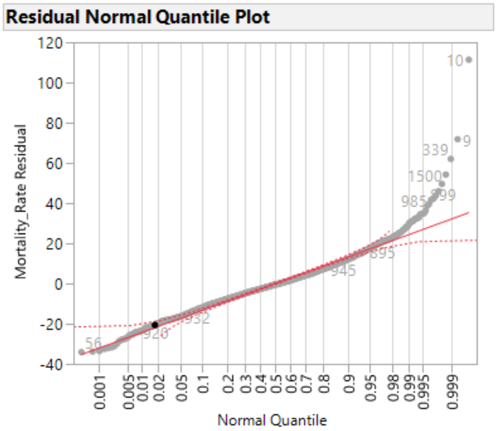
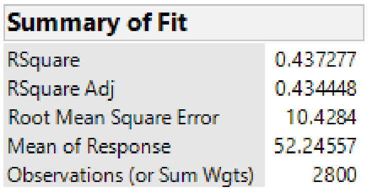


Figure C.2 – Residual normal quantile plot 2nd order polynomial linear regression.

Appendix D

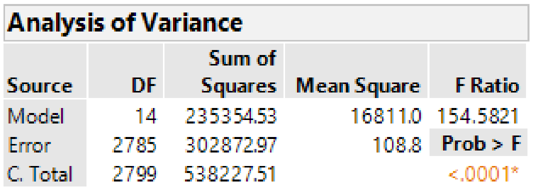
Second order polynomial regression using stepwise regression methods with $\alpha = 0.05$.



Summary of Fit

RSquare	0.437277
RSquare Adj	0.434448
Root Mean Square Error	10.4284
Mean of Response	52.24557
Observations (or Sum Wgts)	2800

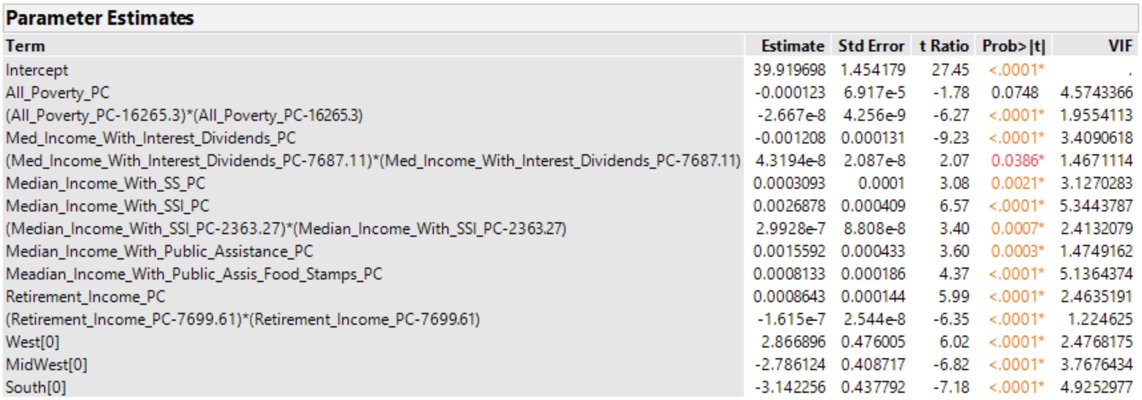
Table D.1 – Summary of fit 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	235354.53	16811.0	154.5821
Error	2785	302872.97	108.8	Prob > F
C. Total	2799	538227.51		<.0001*

Table D.2 – Analysis of variance 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	39.919698	1.454179	27.45	<.0001*	.
All_Poverty_PC	-0.000123	6.917e-5	-1.78	0.0748	4.5743366
(All_Poverty_PC-16265.3)*(All_Poverty_PC-16265.3)	-2.667e-8	4.256e-9	-6.27	<.0001*	1.9554113
Med_Income_With_Interest_Dividends_PC	-0.001208	0.000131	-9.23	<.0001*	3.4090618
(Med_Income_With_Interest_Dividends_PC-7687.11)*(Med_Income_With_Interest_Dividends_PC-7687.11)	4.3194e-8	2.087e-8	2.07	0.0386*	1.4671114
Median_Income_With_SS_PC	0.0003093	0.0001	3.08	0.0021*	3.1270283
Median_Income_With_SSI_PC	0.0026878	0.000409	6.57	<.0001*	5.3443787
(Median_Income_With_SSI_PC-2363.27)*(Median_Income_With_SSI_PC-2363.27)	2.9928e-7	8.808e-8	3.40	0.0007*	2.4132079
Median_Income_With_Public_Assistance_PC	0.0015592	0.000433	3.60	0.0003*	1.4749162
Meadian_Income_With_Public_Assis_Food_Stamps_PC	0.0008133	0.000186	4.37	<.0001*	5.1364374
Retirement_Income_PC	0.0008643	0.000144	5.99	<.0001*	2.4635191
(Retirement_Income_PC-7699.61)*(Retirement_Income_PC-7699.61)	-1.615e-7	2.544e-8	-6.35	<.0001*	1.224625
West[0]	2.866896	0.476005	6.02	<.0001*	2.4768175
MidWest[0]	-2.786124	0.408717	-6.82	<.0001*	3.7676434
South[0]	-3.142256	0.437792	-7.18	<.0001*	4.9252977

Table D.3 – Parameter Estimates 2nd order polynomial liner regression with mixed stepwise and $\alpha = 0.05$.

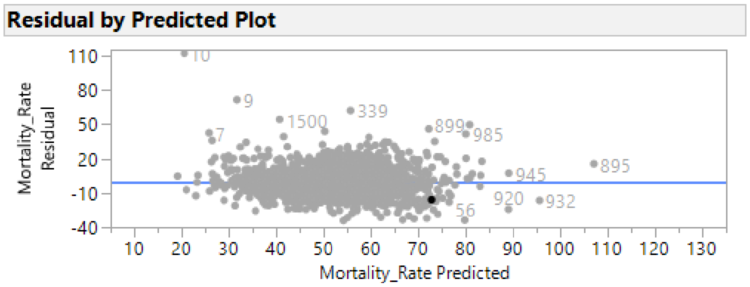


Figure D.1 – Residual by predicted plot 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$.

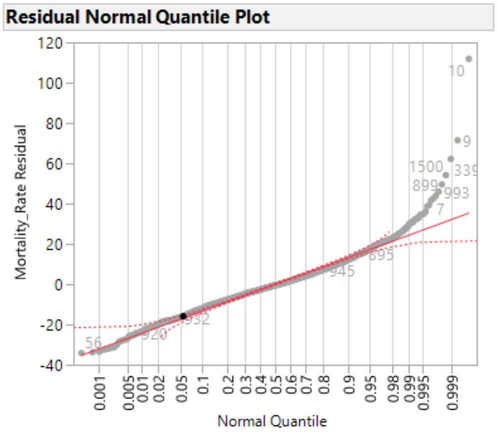
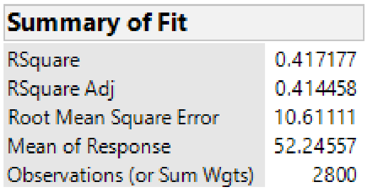


Figure D.2 – Residual normal quantile plot 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$.

Appendix E

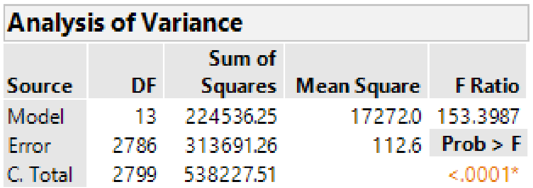
Second order polynomial regression using stepwise regression methods with $\alpha = 0.05$ and excluding Med_Income_With_SSI_PC.



Summary of Fit

RSquare	0.417177
RSquare Adj	0.414458
Root Mean Square Error	10.61111
Mean of Response	52.24557
Observations (or Sum Wgts)	2800

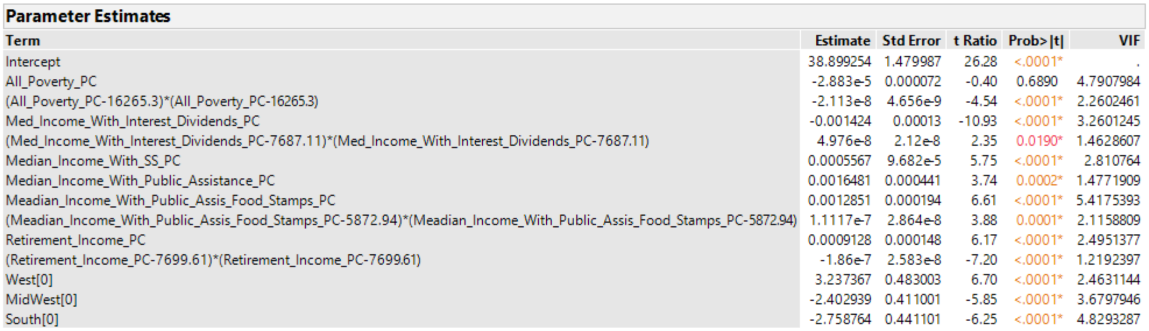
Table E.1 – Summary of fit 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_SSI_PC)



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	224536.25	17272.0	153.3987
Error	2786	313691.26	112.6	Prob > F
C. Total	2799	538227.51		<.0001*

Table E.2 – Analysis of variance 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_SSI_PC)



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	38.899254	1.479987	26.28	<.0001*	.
All_Poverty_PC	-2.883e-5	0.000072	-0.40	0.6890	4.7907984
(All_Poverty_PC-16265.3)*(All_Poverty_PC-16265.3)	-2.113e-8	4.656e-9	-4.54	<.0001*	2.2602461
Med_Income_With_Interest_Dividends_PC	-0.001424	0.00013	-10.93	<.0001*	3.2601245
(Med_Income_With_Interest_Dividends_PC-7687.11)*(Med_Income_With_Interest_Dividends_PC-7687.11)	4.976e-8	2.12e-8	2.35	0.0190*	1.4628607
Median_Income_With_SS_PC	0.0005567	9.682e-5	5.75	<.0001*	2.810764
Median_Income_With_Public_Assistance_PC	0.0016481	0.000441	3.74	0.0002*	1.4771909
Meadian_Income_With_Public_Assis_Food_Stamps_PC	0.0012851	0.000194	6.61	<.0001*	5.4175393
(Meadian_Income_With_Public_Assis_Food_Stamps_PC-5872.94)*(Meadian_Income_With_Public_Assis_Food_Stamps_PC-5872.94)	1.1117e-7	2.864e-8	3.88	0.0001*	2.1158809
Retirement_Income_PC	0.0009128	0.000148	6.17	<.0001*	2.4951377
(Retirement_Income_PC-7699.61)*(Retirement_Income_PC-7699.61)	-1.86e-7	2.583e-8	-7.20	<.0001*	1.2192397
West[0]	3.237367	0.483003	6.70	<.0001*	2.4631144
MidWest[0]	-2.402939	0.411001	-5.85	<.0001*	3.6797946
South[0]	-2.758764	0.441101	-6.25	<.0001*	4.8293287

Table E.3 – Parameter Estimates 2nd order polynomial liner regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_SSI_PC)

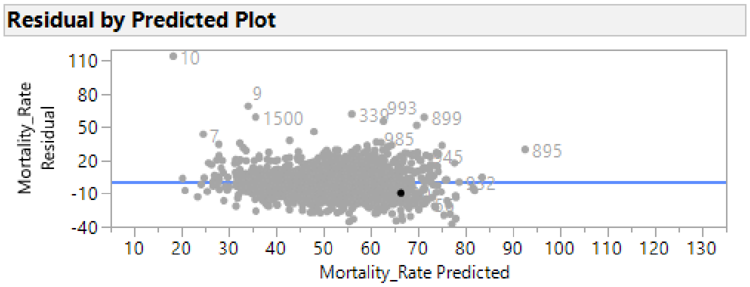


Figure E.1 – Residual by predicted plot 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_SSI_PC)

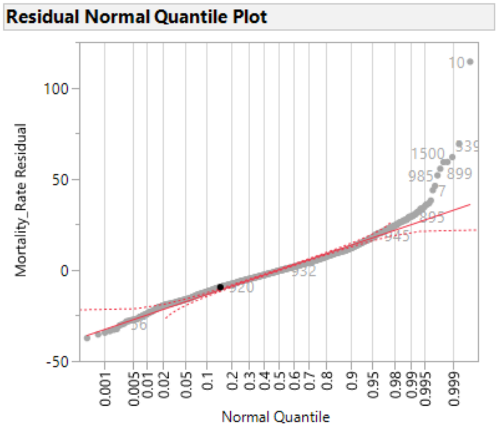
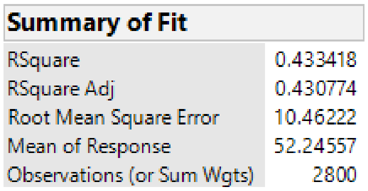


Figure E.2 – Residual normal quantile plot 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_SSI_PC)

Appendix F

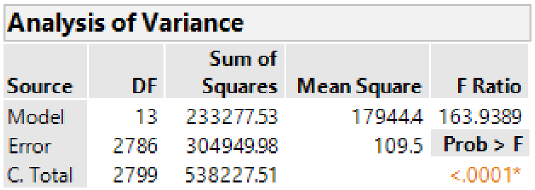
Second order polynomial regression using stepwise regression methods and excluding Med_Income_With_Public_Assis_Food_Stamps_PC.



Summary of Fit

RSquare	0.433418
RSquare Adj	0.430774
Root Mean Square Error	10.46222
Mean of Response	52.24557
Observations (or Sum Wgts)	2800

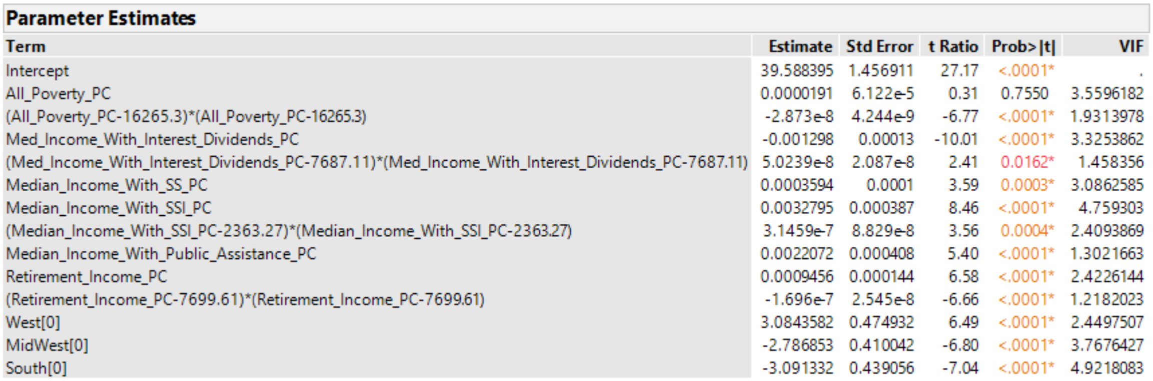
Table F.1 – Summary of fit 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC)



Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	13	233277.53	17944.4	163.9389
Error	2786	304949.98	109.5	Prob > F
C. Total	2799	538227.51		<.0001*

Table F.2 – Analysis of variance 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	39.588395	1.456911	27.17	<.0001*	.
All_Poverty_PC	0.0000191	6.122e-5	0.31	0.7550	3.5596182
(All_Poverty_PC-16265.3)*(All_Poverty_PC-16265.3)	-2.873e-8	4.244e-9	-6.77	<.0001*	1.9313978
Med_Income_With_Interest_Dividends_PC	-0.001298	0.00013	-10.01	<.0001*	3.3253862
(Med_Income_With_Interest_Dividends_PC-7687.11)*(Med_Income_With_Interest_Dividends_PC-7687.11)	5.0239e-8	2.087e-8	2.41	0.0162*	1.458356
Median_Income_With_SS_PC	0.0003594	0.0001	3.59	0.0003*	3.0862585
Median_Income_With_SSI_PC	0.0032795	0.000387	8.46	<.0001*	4.759303
(Median_Income_With_SSI_PC-2363.27)*(Median_Income_With_SSI_PC-2363.27)	3.1459e-7	8.829e-8	3.56	0.0004*	2.4093869
Median_Income_With_Public_Assistance_PC	0.0022072	0.000408	5.40	<.0001*	1.3021663
Retirement_Income_PC	0.0009456	0.000144	6.58	<.0001*	2.4226144
(Retirement_Income_PC-7699.61)*(Retirement_Income_PC-7699.61)	-1.696e-7	2.545e-8	-6.66	<.0001*	1.2182023
West[0]	3.0843582	0.474932	6.49	<.0001*	2.4497507
MidWest[0]	-2.786853	0.410042	-6.80	<.0001*	3.7676427
South[0]	-3.091332	0.439056	-7.04	<.0001*	4.9218083

Table F.3 – Parameter Estimates 2nd order polynomial liner regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC

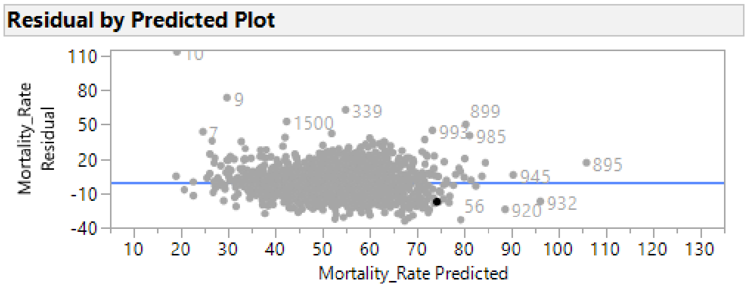


Figure F.1 – Residual by predicted plot 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC

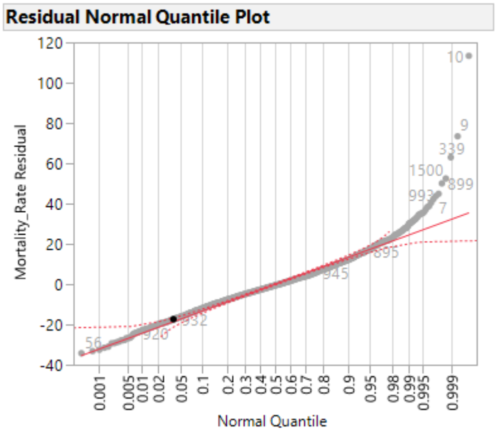
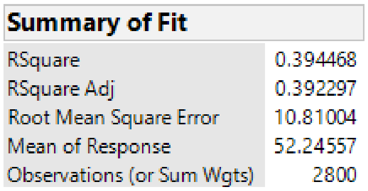


Figure F.2 – Residual normal quantile plot 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC

Appendix G

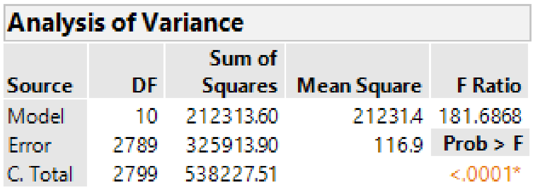
Second order polynomial regression using stepwise regression methods and excluding Med_Income_With_SSI_PC an Med_Income_With_Public_Assis_ Food_Stamps_PC.



Summary of Fit

RSquare	0.394468
RSquare Adj	0.392297
Root Mean Square Error	10.81004
Mean of Response	52.24557
Observations (or Sum Wgts)	2800

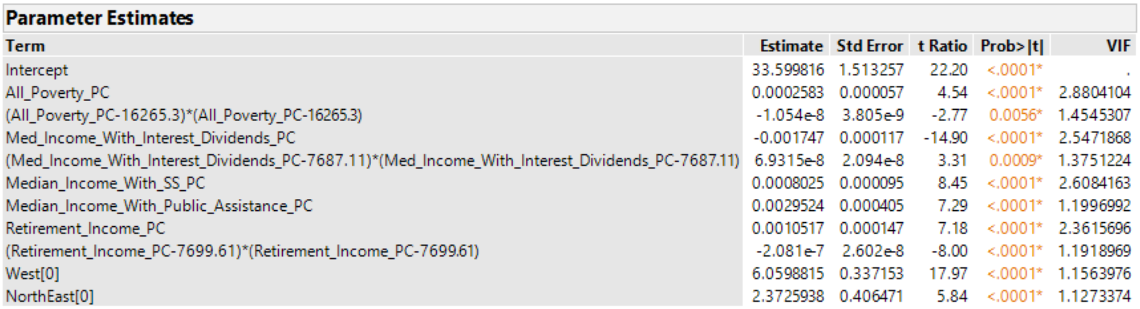
Table G.1 – Summary of fit 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC and With_Public_Assis_Food_Stamps_PC)



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	10	212313.60	21231.4	181.6868
Error	2789	325913.90	116.9	Prob > F
C. Total	2799	538227.51		<.0001*

Table G.2 – Analysis of variance 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC and With_Public_Assis_Food_Stamps_PC



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	33.599816	1.513257	22.20	<.0001*	.
All_Poverty_PC	0.0002583	0.000057	4.54	<.0001*	2.8804104
(All_Poverty_PC-16265.3)*(All_Poverty_PC-16265.3)	-1.054e-8	3.805e-9	-2.77	0.0056*	1.4545307
Med_Income_With_Interest_Dividends_PC	-0.001747	0.000117	-14.90	<.0001*	2.5471868
(Med_Income_With_Interest_Dividends_PC-7687.11)*(Med_Income_With_Interest_Dividends_PC-7687.11)	6.9315e-8	2.094e-8	3.31	0.0009*	1.3751224
Median_Income_With_SS_PC	0.0008025	0.000095	8.45	<.0001*	2.6084163
Median_Income_With_Public_Assistance_PC	0.0029524	0.000405	7.29	<.0001*	1.1996992
Retirement_Income_PC	0.0010517	0.000147	7.18	<.0001*	2.3615696
(Retirement_Income_PC-7699.61)*(Retirement_Income_PC-7699.61)	-2.081e-7	2.602e-8	-8.00	<.0001*	1.1918969
West[0]	6.0598815	0.337153	17.97	<.0001*	1.1563976
NorthEast[0]	2.3725938	0.406471	5.84	<.0001*	1.1273374

Table G.3 – Parameter Estimates 2nd order polynomial liner regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC and With_Public_Assis_Food_Stamps_PC

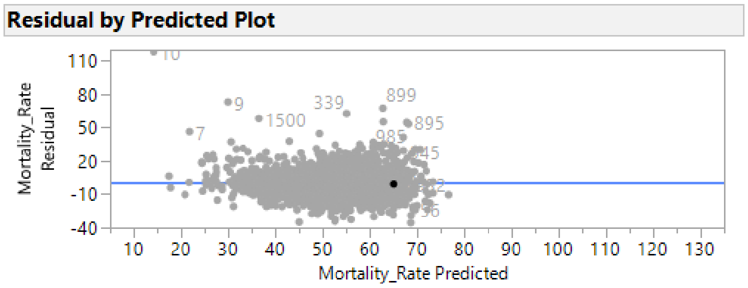


Figure G.1 – Residual by predicted plot 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC and With_Public_Assis_Food_Stamps_PC

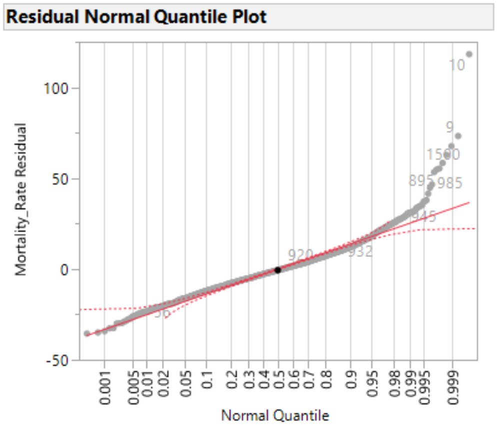
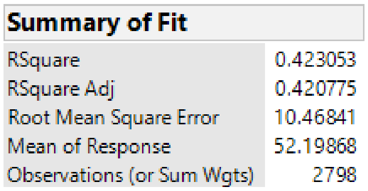


Figure G.2 – Residual normal quantile plot 2nd order polynomial linear regression with mixed stepwise and $\alpha = 0.05$. (excluding Med_Income_With_Public_Assis_Food_Stamps_PC and With_Public_Assis_Food_Stamps_PC

Appendix H

Final Model

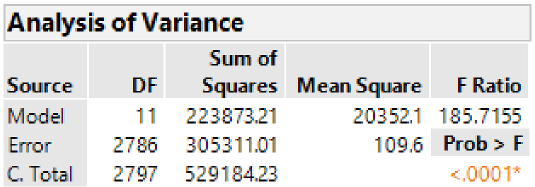
1. Second order polynomial regression.
2. Stepwise mixed regression method with $\alpha = 0.05$.
3. Excluding Med_Income_With_SSI_PC and Med_Income_With_Public_Assis_Food_Stamps_PC
4. Excluded observation 9 and 10.
5. Included NorthEast indicator variable instead of South.



Summary of Fit

RSquare	0.423053
RSquare Adj	0.420775
Root Mean Square Error	10.46841
Mean of Response	52.19868
Observations (or Sum Wgts)	2798

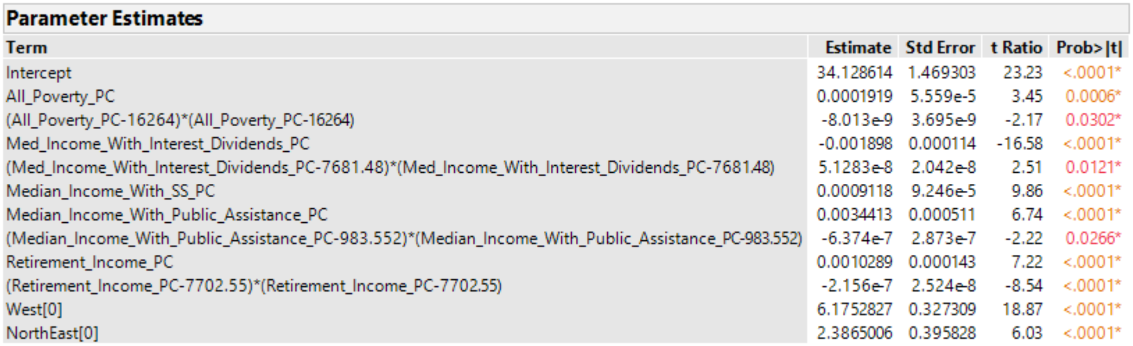
Table H.1 – Summary of fit of final model



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	223873.21	20352.1	185.7155
Error	2786	305311.01	109.6	Prob > F
C. Total	2797	529184.23		<.0001*

Table H.2 – Analysis of variance final model of final model



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	34.128614	1.469303	23.23	<.0001*
All_Poverty_PC	0.0001919	5.559e-5	3.45	0.0006*
(All_Poverty_PC-16264)*(All_Poverty_PC-16264)	-8.013e-9	3.695e-9	-2.17	0.0302*
Med_Income_With_Interest_Dividends_PC	-0.001898	0.000114	-16.58	<.0001*
(Med_Income_With_Interest_Dividends_PC-7681.48)*(Med_Income_With_Interest_Dividends_PC-7681.48)	5.1283e-8	2.042e-8	2.51	0.0121*
Median_Income_With_SS_PC	0.0009118	9.246e-5	9.86	<.0001*
Median_Income_With_Public_Assistance_PC	0.0034413	0.000511	6.74	<.0001*
(Median_Income_With_Public_Assistance_PC-983.552)*(Median_Income_With_Public_Assistance_PC-983.552)	-6.374e-7	2.873e-7	-2.22	0.0266*
Retirement_Income_PC	0.0010289	0.000143	7.22	<.0001*
(Retirement_Income_PC-7702.55)*(Retirement_Income_PC-7702.55)	-2.156e-7	2.524e-8	-8.54	<.0001*
West[0]	6.1752827	0.327309	18.87	<.0001*
NorthEast[0]	2.3865006	0.395828	6.03	<.0001*

Table H.3 – Parameter Estimates of final model

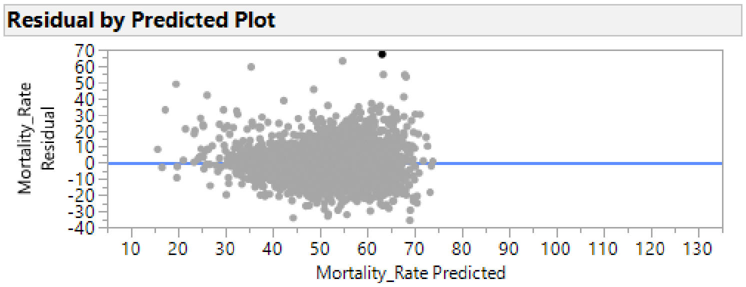


Figure H.1 – Residual by predicted of final model.

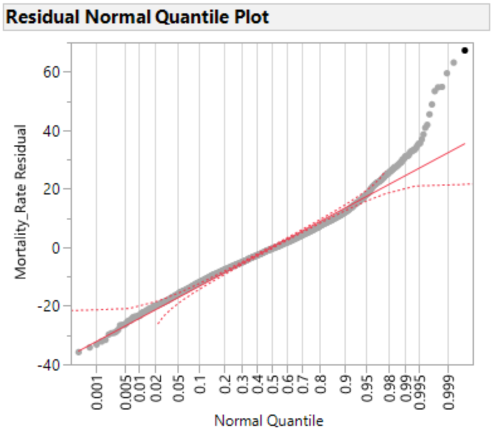


Figure H.2 – Residual normal quantile plot of final model