Lab for Spring 2016 *(worth a total of 100 points)*
Due Date: Thursday, Feb.18th, beginning of class period
GEO 4167/GEO 6161--Intermediate Quantitative Methods (Fik)

Name: _____ Score: _____

Instructions: Complete this lab to the best of your abilities. Attach your work to this cover sheet, including relevant computer output, results, and write-up. Try to minimize paper usage (so edit/cut/paste…and minimize accordingly). Limit your output (< 12 pages total). Be sure to discuss your findings in detail.

Regression Model Estimation & Diagnostics

Using a statistical software package or method of your choice, estimate a viable multivariate Least Squares regression model that ***best explains the variation in a dependent variable*** -- namely, ***monthly mortgage payment*** Y (in $) using the data in Table 1. The dependent variable Y (the monthly mortgage payment in $, does not including escrow payments, insurance, or taxes). Assume you wish to model variability in Y as a function of household disposable (after-tax) income X1 (x$1,000; rounded to the nearest thousand), square footage of the housing units in question X2 (assumed to be a second-order polynomial as suggested by recent empirical evidence), mortgage type X3 (15-year versus 30-year notes, fixed-rate loans at fixed interest rate between 3.75% and 4.25% APR), and the housing unit's Age (in years, rounded to the nearest integer value).

The data represent a random sample of n=45 housing loans issued between November 1st and December 31st of 2015 with 10% of the purchase price down at closing, for single-family detached housing units sold in the city of Athens, Georgia (in a given geographic sub-market defined by housing units between 1,300 and 1,800 sq. ft. in size, built after Jan.1, 2001).

Using a statistical software package or method of your choice and the sample data provided in Table 1, estimate a "best-fit" regression model. Provide evidence to support what you identify as a "best model", and discuss the results. In addition, (a) test to see if the assumption of normality of error is valid; and (b) if the error terms are independent of the dependent variable, predicted values, and the regressors/covariates in your final model. Using your best model, recover estimates of Y (in $) and the mean prediction confidence band for observation #25. Note that you may consider variable transformations (e.g., logarithms), variable interactions, non-linear estimation, and step-wise procedures to help you identify a "best-fit" model. Also, highlight any extreme error terms or outliers. Be sure to briefly discuss your modeling strategy and what criteria you used to identify the final model.

Table 1. Sample Database for Lab#1 (see Excel spreadsheet for Lab1)

Obs.	Y	X1	X2	X3	A
1	450	24	1050	0	2
2	390	26	1000	0	3
3	385	28	1148	0	8
4	440	29	1100	0	4
5	530	34	1110	0	4
6	420	19	1270	0	3
7	435	22	1200	0	7
8	550	44	1280	1	5
9	600	40	1300	1	2
10	570	39	1240	0	2
11	455	29	1200	0	3
12	505	40	1300	0	1
13	655	45	1320	1	2
14	475	35	1090	0	8
15	480	30	1140	0	8
16	680	42	1400	1	3
17	435	29	1000	0	10
18	400	30	1040	0	11
19	390	25	1000	0	9
20	380	28	1090	0	9
21	445	30	1180	0	6
22	440	40	1200	0	8
23	500	36	1300	0	4
24	615	42	1400	0	1
25	605	45	1370	0	5
26	590	46	1320	0	5
27	545	44	1380	0	5
28	380	21	1200	0	12
29	520	45	1240	0	4
30	395	38	1160	0	12
31	400	35	1100	0	9
32	710	46	1480	0	4
33	755	44	1450	0	3
34	800	41	1400	0	3
35	855	46	1450	1	4
36	875	45	1500	1	2
37	955	44	1510	1	3
38	1000	49	1600	0	2
39	1100	50	1590	0	6
40	1235	48	1650	1	6
41	1250	51	1700	1	2
42	1300	52	1660	1	3
43	1360	48	1590	1	2
44	1400	55	1790	0	1
45	1325	45	1700	0	1

Dummy: (X3=1 if 15 yr note; X3=0 if 30 yr note); A = Age of Unit in Years (from year built, rounded to the nearest integer value).