NEAS Regression Analysis – Student Project
By David Broomhead

Loss Reserving

Please refer to the excel file "Regression Analysis Project – David Broomhead.xls" for details, but I will attempt to describe my approach as completely as possible in this document. ***Throughout**, **I will deem P-values lower than 0.005 to be significant.***

Excel file sheets:

Sheet Name	Description
Stable_Generator	Loss Triangle generator with stable accident year trend and development pattern
Stable_Fixed	Copy of Stable_Generator, with static random numbers
Stable_Residuals	Regression analysis
Unstable_Generator	Loss Triangle generator with unstable accident year trend and unstable development pattern, as well as reform effects
Unstable_Fixed	Copy of Unstable_Generator, with static random numbers
Unstable_Residuals1	Initial regression analysis
Unstable_Residuals2	Subsequent regression analysis with additional explanatory variables
Unstable_Residuals3	Subsequent regression analysis with additional explanatory variables
Unstable_Residuals4	Subsequent regression analysis with additional explanatory variables
Unstable_Residuals5	Subsequent regression analysis with additional explanatory variables
Unstable_Residuals6	Subsequent regression analysis with additional explanatory variables
Unstable_Fixed no Variability	Copy of Unstable_Fixed, but with no variability (0 standard error)
Unstable_Residuals no Variab	Regression analysis

Stable Scenario

The Stable_Generator sheet generates a random Log(Incremental Paid), Accident Year/Development Year triangle according to a selected intercept and explanatory variable coefficients, along with a selected standard error (similar to the Loss Reserving template). The intercept sets a base level. The explanatory variables are Accident Year (x1) and Development year (x2), and the coefficients are accident year trend and development year decay.

The data produced by this process is linear by Accident Year and Development Year, so effectively the entire triangle is homogeneous as it is produced by consistent assumptions throughout.

The selected coefficients are:

sigma	0.05	standard error of the regression
alpha	8	intercept of the regression equation
beta1	0.01	accident period trend (inflation)
beta2	-0.3	development period decay

The Stable_Fixed sheet is a static (fixed) version of the Stable_Generator sheet, the only difference being that the random numbers are fixed. The regression will be performed on this dataset. Effectively a dataset has been generated, and I will pretend that I do not know the process used to generate the dataset and use techniques from this course to fit a regression model.

It is common to start with an incremental loss development triangle such as the one below (found in sheet Stable_Fixed).

Incremental Paid	Development Year														
Accident Yr	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	3,133	2,307	1,491	1,203	914	653	514	409	269	197	138	102	78	59	45
1	3,037	2,140	1,675	1,221	888	670	470	373	282	213	139	103	85	69	
2	2,930	2,282	1,543	1,258	887	731	496	372	281	208	155	115	77		
3	3,185	2,263	1,804	1,276	936	685	518	380	292	200	155	110			
4	2,819	2,136	1,694	1,385	1,032	707	509	379	263	213	153				
5	2,894	2,476	1,840	1,288	1,063	742	515	351	285	227					
6	3,472	2,243	1,951	1,346	982	665	515	413	294						
7	3,312	2,578	1,711	1,290	954	701	537	382							
8	3,144	2,388	1,686	1,225	984	719	517								
9	3,303	2,334	1,990	1,411	1,017	715									
10	3,295	2,628	1,768	1,311	955										
11	3,175	2,422	1,994	1,447											
12	3,462	2,306	1,892												
13	3,261	2,756													
14	3,518														

Since insurance losses are typically right-skewed, the log transform seems appropriate in order to normalize the distribution of each cell (normalize the errors). This triangle is shown below, but it can also be found in Stable_Fixed.

Log(Incremental Paid)	Development Year														
Accident Yr	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	8.050	7.744	7.307	7.092	6.818	6.482	6.242	6.013	5.595	5.282	4.927	4.628	4.355	4.076	3.801
1	8.019	7.668	7.424	7.107	6.789	6.507	6.153	5.921	5.641	5.363	4.931	4.635	4.441	4.236	
2	7.983	7.733	7.342	7.137	6.788	6.595	6.207	5.919	5.637	5.336	5.044	4.749	4.339		
3	8.066	7.724	7.498	7.151	6.841	6.529	6.250	5.940	5.677	5.297	5.042	4.700			
4	7.944	7.666	7.435	7.233	6.940	6.561	6.232	5.937	5.571	5.360	5.029				
5	7.970	7.814	7.518	7.161	6.969	6.610	6.244	5.860	5.654	5.424					
6	8.153	7.715	7.576	7.205	6.890	6.500	6.245	6.023	5.685						
7	8.105	7.855	7.445	7.163	6.861	6.553	6.287	5.946							
8	8.053	7.778	7.430	7.111	6.892	6.578	6.248								
9	8.102	7.756	7.596	7.252	6.924	6.573									
10	8.100	7.874	7.478	7.178	6.862										
11	8.063	7.792	7.598	7.277											
12	8.150	7.743	7.545												
13	8.090	7.921													
14	8.166														

Observing the data in triangle form, there appears to be a linear relationship between the Accident Year and Log(Incremental Paid) for each given Development Year, and likewise, a linear relationship between the Development Year and Log(Incremental Paid) for each given Accident Year. Hence, a natural first choice is to select the explanatory variables of Accident Year (x1) and Development year (x2).

A regression analysis with those variables is shown in sheet Stable_Residuals. The R^2 is extremely good at 99.79%. The intercept and explanatory variable coefficients are significant, as their values are significantly different from 0 (they are more than 2 standard errors away from 0), which is confirmed by their extremely low P-Values and high t-statistics. The F test shows that the null hypothesis that all coefficients are 0 is rejected up to an extremely high level of confidence.

Regression Statistics	
Multiple R	0.998971
R Square	0.997943

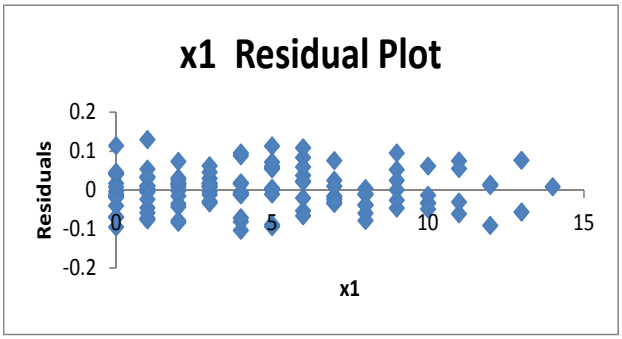
Adjusted R Square	0.997908
Standard Error	0.051208
Observations	120

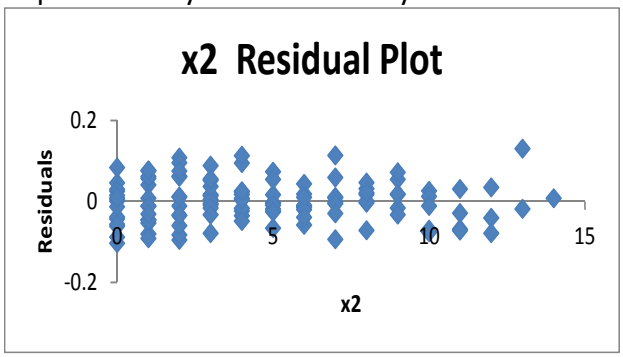
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	148.8677	74.43383	28385.35	6.6E-158
Residual	117	0.306805	0.002622		
Total	119	149.1745			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	8.004079	0.012877	621.575	1E-207
X1	0.010949	0.001484	7.375702	2.56E-11
X2	-0.30069	0.001484	-202.558	8.3E-151

Review of the residual plots generally suggest homoscedasticity of the residuals for both x1 and x2 as they are centred around 0 and have generally constant variance. AY 3 shows a smaller variance and AY 8 shows a negative bias, but these appear to be driven by randomness and the sample size is too small to draw raise any concerns. I would conclude that the residuals are IID. This, combined with the high R2 and high significance of the coefficients, while only using 2 explanatory variables suggests that this is an extremely well fitting model so I do not deem it necessary to perform any additional analysis.





Unstable Scenario

In order to demonstrate regression techniques on data that is not homogeneous, I created a sheet Unstable_Generator that generates a random Log(Incremental Paid), Accident Year/Development Year triangle with changes to trends (Accident Year and Calendar year effects) as well as changes in development patterns and benefit level (as might be seen following a product reform). In my reserving role for a Home & Auto insurer in Ontario Canada, this topic has been particularly relevant in recent years as Ontario Auto Reform lead to significant benefit level changes and development patterns in Ontario Accident Benefits auto insurance.

I started with a similar approach to the Stable data. I applied a reduction to the level of the benefits after Accident year 6. This can be thought of as an adjustment to the intercept (hence my labelling as alpha1) but is more appropriately described as a coefficient for a dichotomous explanatory variable. I overlaid a change in Calendar Year trend after the 7th diagonal, which is a dichotomous interaction between the

explanatory variables x1 and x2. Also, the development pattern changes after Accident Year 6, as an additional decay coefficient is applied to the Development Year variable (a dichotomous interaction between Accident Year and Development Year).

The effect of each coefficient can be checked and directly observed in the resulting triangle by setting the standard error (sigma) to 0, and all other coefficients to 0.

The selected coefficients are:

sigma	0.1	standard error of the regression
alpha	9	intercept of the regression equation
alpha1	-1.5	adjustment to intercept after AY 6 (reform to benefit levels)
beta1	0.05	accident period trend (inflation)
beta2	-0.3	development period decay
beta3	0.15	additional calendar period trend after 7th diagonal (i.e. x1+x2 > 6)
beta4	-0.3	additional development period decay after AY 6 (change in development pattern)

The Unstable_Fixed sheet is a static (fixed) version of the Unstable_Generator sheet, the only difference being that the random numbers are fixed. The regression will be performed on this dataset. Effectively a dataset has been generated, and as before, I will pretend that I do not know the process used to generate the dataset and use techniques from this course to fit a regression model.

The Paid and Log(Paid) triangles are as follows:

Incremental Paid	Development Year														
Accident Yr	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	6,280	5,544	4,267	2,855	2,238	1,933	1,290	1,069	991	928	894	730	527	517	419
1	7,568	6,463	4,404	3,277	3,393	2,241	1,784	1,433	1,137	1,092	863	873	572	513	
2	9,333	6,668	4,689	3,916	2,728	2,382	2,068	1,734	1,532	1,333	1,001	816	873		
3	9,610	6,551	5,627	3,862	3,508	3,460	2,620	2,600	1,656	1,670	1,338	1,080			
4	8,343	7,758	5,540	4,537	4,566	2,920	3,021	2,554	2,118	2,221	2,064				
5	9,981	7,411	6,681	5,737	4,605	4,388	3,693	2,981	2,776	2,357					
6	13,064	10,335	8,377	6,567	6,324	5,706	3,656	3,871	3,238						
7	3,392	1,700	1,228	731	429	296	180	154							
8	3,886	1,801	1,583	982	667	394	259								
9	3,926	3,206	1,789	1,227	709	549									
10	6,401	3,964	1,878	1,528	747										
11	6,658	4,053	3,317	1,638											
12	9,291	4,817	3,877												
13	9,153	7,910													
14	14,811														
Log(Incremental Pai	Development Year														
Accident Yr	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	8.745	8.620	8.359	7.957	7.714	7.567	7.163	6.974	6.898	6.833	6.796	6.593	6.267	6.248	6.038
1	8.932	8.774	8.390	8.095	8.130	7.714	7.487	7.267	7.036	6.996	6.761	6.772	6.350	6.241	
2	9.141	8.805	8.453	8.273	7.912	7.776	7.634	7.458	7.335	7.195	6.909	6.704	6.772		
3	9.171	8.787	8.635	8.259	8.163	8.149	7.871	7.863	7.412	7.421	7.199	6.985			
4	9.029	8.957	8.620	8.420	8.426	7.979	8.013	7.845	7.658	7.706	7.632				
5	9.208	8.911	8.807	8.655	8.435	8.387	8.214	8.000	7.929	7.765					
6	9.478	9.243	9.033	8.790	8.752	8.649	8.204	8.261	8.083						
7	8.129	7.438	7.113	6.594	6.061	5.690	5.192	5.034							
8	8.265	7.496	7.367	6.890	6.502	5.976	5.555								
9	8.275	8.073	7.490	7.112	6.564	6.307									
10	8.764	8.285	7.538	7.332	6.616										
11	8.804	8.307	8.107	7.401											
12	9.137	8.480	8.263												
13	9.122	8.976													
14	9.603														

We can clearly see the product reform effect that I introduced, and although close inspection reveals the increased CY trend and change in development pattern post-reform, these are less apparent.

The approach is much the same as before (for the Stable data), but it is clear that there is some heterogeneity in the data, so it is likely that further explanatory variables will be needed. I will not show every table and every residual plot as they are in the excel file, but I will show the initial run, the final run, and interesting plots along the way.

Unstable_Residuals1

A regression analysis with 2 explanatory variables (x1=Accident Year, x2=Development Year) is shown in sheet Unstable_Residuals1. The R^2 is fairly poor at 48.65%. The intercept and explanatory variable coefficients are significant, as their values are significantly different from 0 (they are more than 2 standard errors away from 0), which is confirmed by their extremely low P-Values and high t-statistics. The F test shows that the null hypothesis that all coefficients are 0 is rejected up to an extremely high level of confidence. I conclude from this that the explanatory variables chosen are useful (since they are significant), but not sufficient (since the R^2 is low) without the addition of further variables in order to get a good fit.

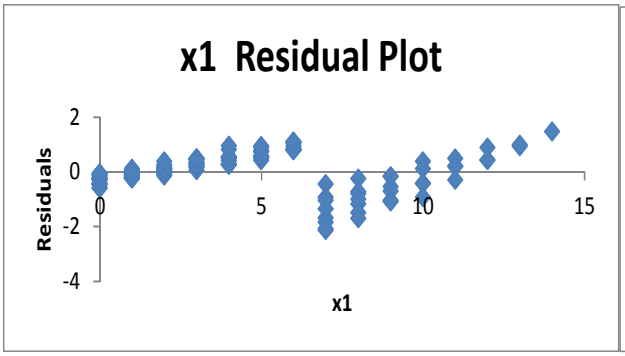
Regression Statistics	
Multiple R	0.697478
R Square	0.486475
Adjusted R Square	0.477697
Standard Error	0.700319
Observations	120

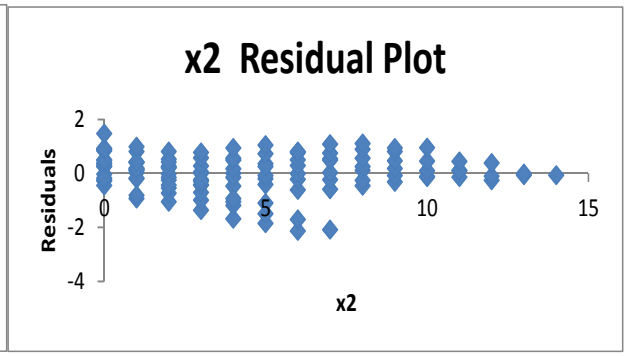
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	54.35964	27.17982	55.41851	1.17E-17
Residual	117	57.38225	0.490447		
Total	119	111.7419			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	9.027107	0.176107	51.25933	5.19E-82
x1	-0.06349	0.020301	-3.12748	0.002225
x2	-0.20849	0.020301	-10.2696	5E-18

Review of the residual plots shows clear heteroscedasticity of the residuals for both x1 and x2. There is a clear trend in the x1 residuals, and a drastic change at AY 7, which would be consistent with our knowledge that there has been a product reform. This suggests that I should add a dichotomous (dummy) variable for AY>6 in the next run. The x2 residual plot also shows heteroscedasticity as the residuals for DY 8+ seem to have positive bias, and the residuals for DY 0-7 show a decreasing bias.

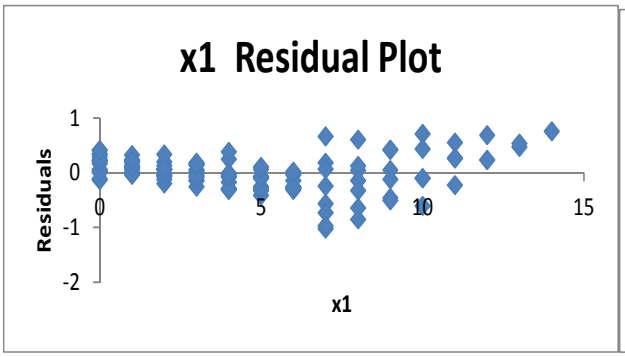


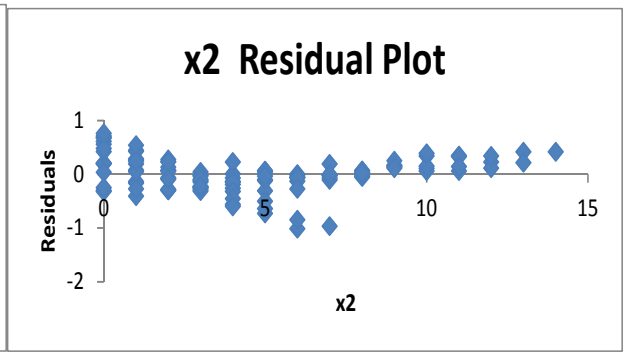


Unstable Residuals2

As previously mentioned, I add the dichotomous variable x3 for AY>6 (a dummy variable). This improved the R^2 significantly to 88.96%, and the coefficient is significant.

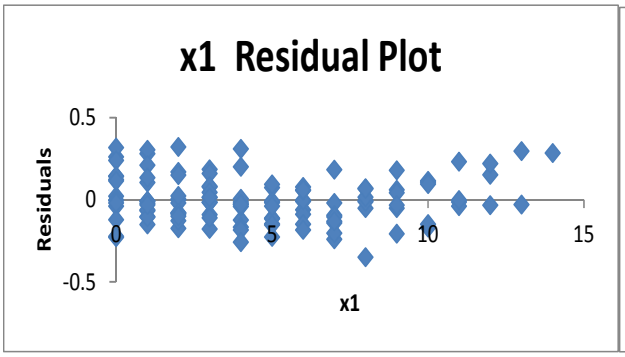
However, review of the residual plots still shows clear heteroscedasticity of the residuals for both x1 and x2. x1 residuals have low variance for AY0-6 and higher variance for AY7+, as well as an increasing trend (positive bias) in the residuals after AY7. The x2 residuals also show some heteroscedasticity, similar to the previous run – there appears to be 2 distinct development patterns in the first 7 development years. Reviewing the residual triangle suggests that the development pattern changes for AY 7+ - this suggests that there may be some interaction between x2 (DY) and x3 (AY>6), so I'll add this variable next.

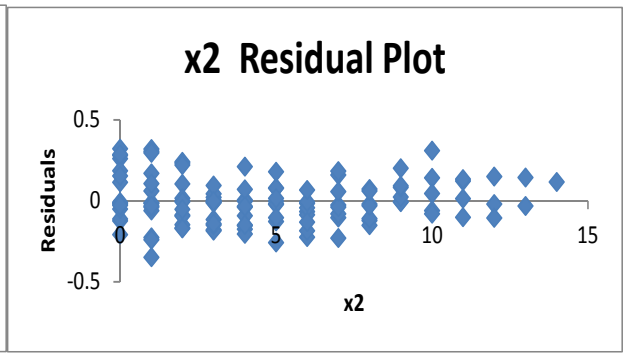




Unstable Residuals3

I add the interaction variable x4=x2*x3. This variable is significant and improves the R^2 to 97.88%. This is an extremely good fit. Review of the residual plots shows no clear indication of heteroscedasticity – we can reasonably conclude that there is homoscedasticity here. On this information alone, it might be reasonable to stop here.

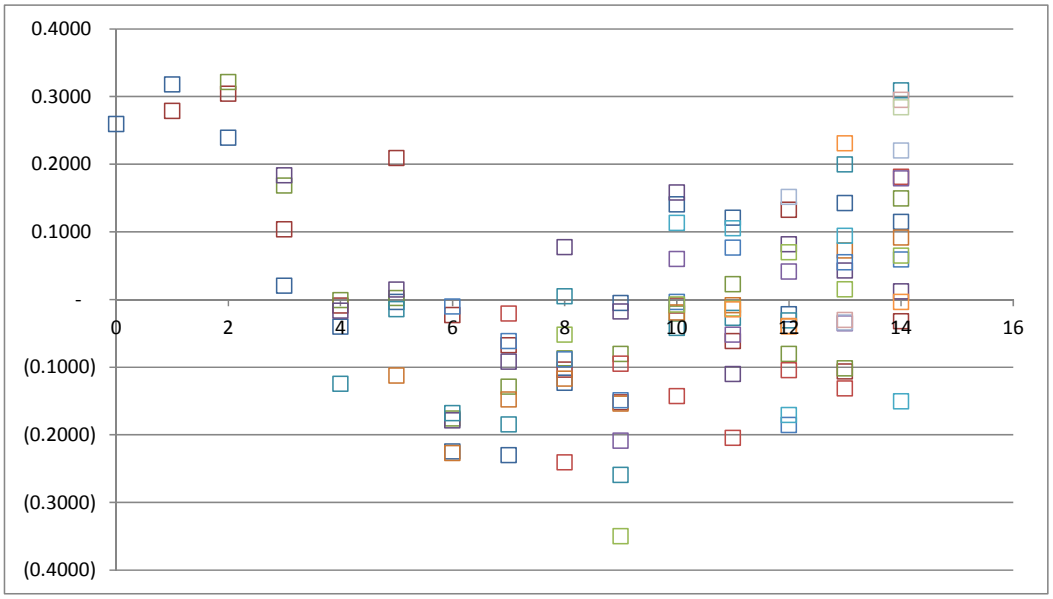




Review of the residual triangle suggests that an improvement can be made. The earlier diagonals and late diagonals generally have positive residuals and the middle diagonals generally have negative residuals, as can be seen by the triangle itself, and the "V shape" in the residual plot below (both in sheet "Unstable_Residuals3"):

Residuals	*DY*														
AY	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.2593	0.3176	0.2390	0.0202	(0.0401)	(0.0037)	(0.2248)	(0.2300)	(0.1230)	(0.0051)	0.1404	0.1209	(0.0218)	0.1422	0.1147
1	0.2787	0.3040	0.1035	(0.0089)	0.2090	(0.0231)	(0.0678)	(0.1039)	(0.1521)	(0.0092)	(0.0616)	0.1326	(0.1065)	(0.0324)	
2	0.3214	0.1682	(0.0009)	0.0021	(0.1762)	(0.1289)	(0.0872)	(0.0805)	(0.0208)	0.0227	(0.0805)	(0.1017)	0.1493		
3	0.1836	(0.0165)	0.0145	(0.1789)	(0.0920)	0.0774	(0.0176)	0.1580	(0.1103)	0.0814	0.0429	0.0118			
4	(0.1248)	(0.0144)	(0.1681)	(0.1847)	0.0047	(0.2592)	(0.0423)	(0.0271)	(0.0312)	0.1995	0.3091				
5	(0.1126)	(0.2272)	(0.1478)	(0.1171)	(0.1539)	(0.0192)	(0.0085)	(0.0396)	0.0722	0.0916					
6	(0.0104)	(0.0617)	(0.0886)	(0.1491)	(0.0037)	0.0765	(0.1856)	0.0548	0.0592						
7	(0.0208)	(0.2407)	(0.0950)	(0.1430)	(0.2046)	(0.1044)	(0.1317)	0.1814							
8	(0.0520)	(0.3499)	(0.0076)	(0.0137)	0.0697	0.0149	0.0647								
9	(0.2087)	0.0597	(0.0522)	0.0413	(0.0355)	0.1789									
10	0.1132	0.1051	(0.1708)	0.0940	(0.1507)										
11	(0.0146)	(0.0398)	0.2307	(0.0039)											
12	0.1516	(0.0341)	0.2200												
13	(0.0303)	0.2948													
14	0.2839														

[The conditional formatting (red = high, green = low) really helps to observe this effect.]



[Clear V shape in the residuals by Calendar Year]

This suggests that we might add a dichotomous variable around the middle diagonal, as we suspect that there has been a calendar year effect caused by the reform.

Unstable_Residuals4

I add x5, a Calendar Year dichotomous variable defined as (x1+x2>6). [Note that adding Calendar Year as Accident Year + Development Year does not work, as we run into the problem of collinearity.] This variable is significant, and slightly improves the R^2 to 98.03%. We still have evidence of the "V-shape" in the residual triangle (across calendar years [diagonals]), which leads me to believe that there is an interaction between the x1 and x5 variables, as well as the x2 and x5 variables that I need to consider.

Unstable_Residuals5

To begin with, I add x6=x1*x5, an interaction variable. This variable is significant but only slightly improves the R^2 to 98.26%. We still have the "V-shape" in the calendar year residuals, so I'll consider interaction between x2 and x5 variables.

Unstable_Residuals6

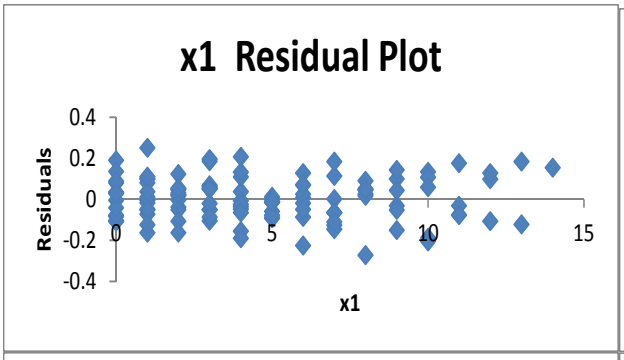
I now add x7=x2*x5, an interaction variable. This variable is significant, and improves the R^2 to 98.9%. The residual triangle now exhibits randomness only, and there is no "V-shape" to speak of. I shall stop here - the R^2 is extremely good, all variables are significant, all residual plots satisfactorily exhibit homoscedasticity, and the residual triangle accordingly shows pure randomness only. I don't believe this model can be improved – adding variables will likely only introduce overfitting. I will support this conclusion by removing the variability.

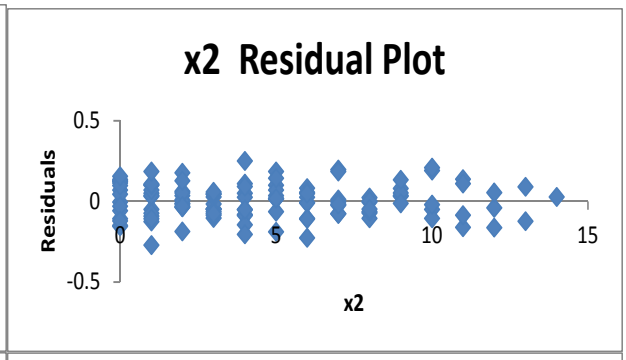
Regression Statistics	
Multiple R	0.994485
R Square	0.989001
Adjusted R Square	0.988313
Standard Error	0.104756
Observations	120

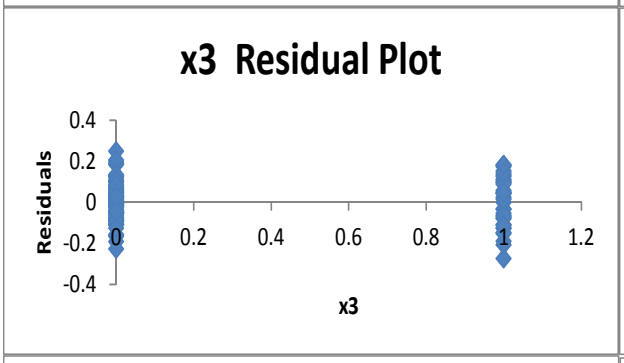
ANOVA

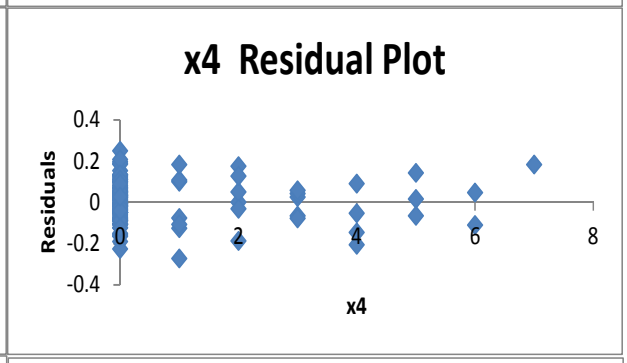
	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	110.5128	15.78755	1438.662	1.5E-106
Residual	112	1.229063	0.010974		
Total	119	111.7419			

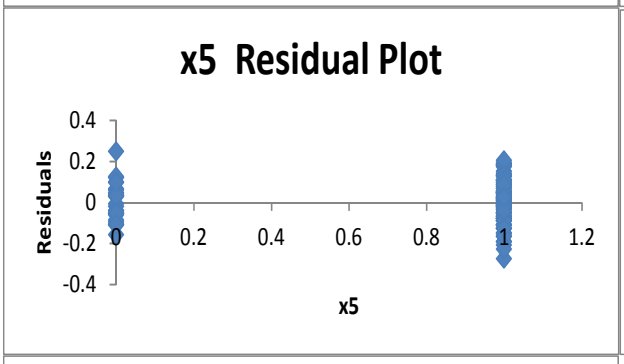
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	8.855935	0.049821	177.754	3.7E-139
x1	0.082709	0.013198	6.266781	7.06E-09
x2	-0.2643	0.013198	-20.0259	8.25E-39
x3	-1.51218	0.054551	-27.7206	5.74E-52
x4	-0.30338	0.009989	-30.3704	6.85E-56
x5	-0.75986	0.07558	-10.0536	2.47E-17
x6	0.122119	0.014882	8.206101	4.3E-13
x7	0.115489	0.01429	8.081634	8.2E-13

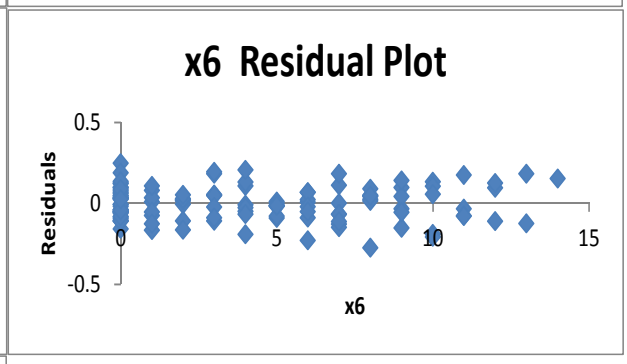


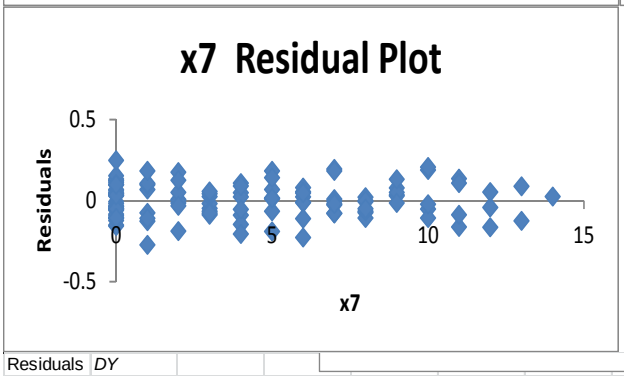




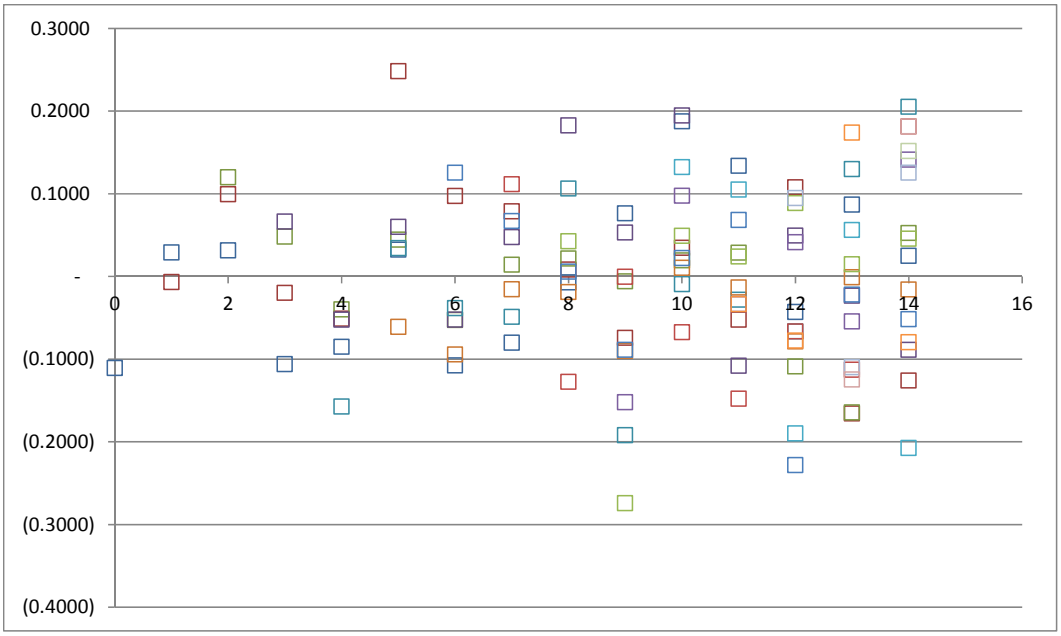








Residuals	DY														
AY	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	(0.1107)	0.0288	0.0314	(0.1062)	(0.0852)	0.0324	(0.1075)	(0.0800)	(0.0073)	0.0764	0.1876	0.1338	(0.0431)	0.0867	0.0249
1	(0.0070)	0.0995	(0.0198)	(0.0509)	0.2482	0.0973	0.0787	0.0083	(0.0741)	0.0345	(0.0522)	0.1078	(0.1656)	(0.1257)	
2	0.1200	0.0480	(0.0398)	0.0444	(0.0526)	0.0140	0.0215	(0.0061)	0.0193	0.0286	(0.1089)	(0.1644)	0.0524		
3	0.0665	(0.0524)	0.0599	(0.0523)	0.0474	0.1825	0.0532	0.1946	(0.1079)	0.0494	(0.0233)	(0.0886)			
4	(0.1575)	0.0341	(0.0384)	(0.0489)	0.1063	(0.1919)	(0.0092)	(0.0283)	(0.0667)	0.1298	0.2052				
5	(0.0611)	(0.0944)	(0.0155)	(0.0191)	(0.0901)	0.0104	(0.0133)	(0.0786)	(0.0010)	(0.0159)					
6	0.1255	0.0671	0.0059	(0.0888)	0.0223	0.0682	(0.2281)	(0.0220)	(0.0518)						
7	0.1115	(0.1273)	(0.0005)	(0.0674)	(0.1479)	(0.0666)	(0.1128)	0.1814							
8	0.0425	(0.2743)	0.0491	0.0241	0.0886	0.0149	0.0458								
9	(0.1520)	0.0975	(0.0333)	0.0413	(0.0544)	0.1411									
10	0.1321	0.1051	(0.1897)	0.0562	(0.2074)										
11	(0.0335)	(0.0776)	0.1740	(0.0795)											
12	0.0949	(0.1096)	0.1255												
13	(0.1248)	0.1814													
14	0.1517														



Unstable_Fixed no Variability

This sheet is simply a copy of the Unstable_Fixed sheet, but with 0 standard error.

Unstable_Residuals no Variab

This sheet is a copy of Unstable_Residuals, but uses the data with no variability. The regression is refreshed and the R^2 is 100%, while each variable is significant as it has 0 P-value. The residuals are all 0 (excel shows some tiny values, but this is simply the effect of floating point error I believe). Therefore, the Unstable_Residuals6 model is the optimal model in that it includes all necessary variables, without including any extra variables that would only cause overfitting (capture the effect of randomness only). Of course, in practice this confirmation would not be possible (if it were, the regression analysis would be completely unnecessary as we would understand the process perfectly).