Group 7

Analysis of Property Data Set on Predicting Incurred Loss

Tana Wuren, Tianyue Mao, Yuting Xiong, Xinyu Fei

Introduction

1. Goal of Analysis

2. Data Description

3.Original Triangle

Methods and Results

1.Data Cleaning

2.Data Reshape

3.Model Fitting Process

4.Model Fitting Technique

5.Results

6.Model Comparison

Conclusions

1.Recommendation

2.Future Study and Unsolved Issue

Appendices

1.Details of Analysis

2.Individual Contribution

Introduction

1. Goal of Analysis:

Our goal for this project is to detect the best model that will predict the most accurate incurred loss of future years while using our current year as the explanatory variable. The models we chose are Chain Ladder, Tweedie, Poisson, and Gaussian models. We will first find the best model to predict incurred loss of future years using the loss incurred from the previous year. Then, we will calculate the ultimate loss with that model to compare the accuracy of the prediction to our original data set. Lastly, we will conduct a ratio test to determine the accuracy of our models and provide valid evidence for the client to accept.

2. Dataset Description

The dataset we obtained from the client was property, which means we will be dealing with non-life insurance. We first looked through the variables provided to see what kind of possible analysis that could be done on the dataset. Then, we will determine what factors will be used to create the loss triangles on the incurred loss and compare the output triangles with different methods. However, before we proceeded with intense coding, we made sure to clean the dataset so that our results were not skewed by abnormal changes. Ultimately, our group decided to use claim number, incurred loss, and development year to create and analyze our triangles.

3. Original Triangle

We obtained the Original Triangle using 76297 claims in the Property dataset. The actual triangle is created based on the values of evaluations in the Property dataset. Also, we calculated the ultimate loss based on the original triangle.

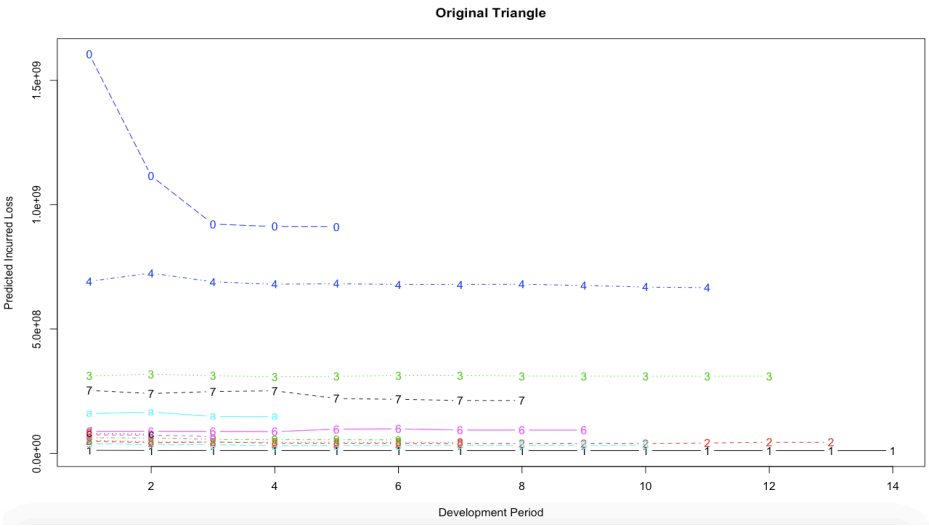
Table.1 Original loss triangle

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	11996603	10792631	10792628	10668514	10668514	10668514	10520128	10520127	10520127	10520127	10520127	10520127	10520127	10520127
2003	49187789	41852416	45431753	40424565	40205080	39662678	39589968	39624898	39849932	39852362	41352575	43876756	43876756	NA
2004	311401148	317640891	312100106	307749470	309113489	313723410	313666739	310956479	310605835	3.1E+08	3.1E+08	310373654	NA	NA
2005	691715640	724745887	689620719	679900230	682174001	678886532	679103832	679555787	675271041	6.69E+08	6.66E+08	NA	NA	NA
2006	39737708	35158035	34571737	31234870	32001853	31862913	32587614	32358857	32365698	32353732	NA	NA	NA	NA
2007	87819912	88427298	87951555	87064276	97481549	98369803	93424192	93441113	93197144	NA	NA	NA	NA	NA
2008	252762792	240607829	248388404	251312852	220642439	216780733	212651207	212319763	NA	NA	NA	NA	NA	NA
2009	51202868	47784324	44586495	43678398	44005644	43806384	43804877	NA	NA	NA	NA	NA	NA	NA
2010	62246474	61077405	56601247	55441916	55211458	54054835	NA	NA	NA	NA	NA	NA	NA	NA
2011	1.604E+09	1.116E+09	922010444	911971467	911380088	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	159381368	166069442	148024435	147959829	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	74126760	72585938	67836292	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	79286781	76617179	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	89991821	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table.2 Orignial loss triangle with ultimate loss

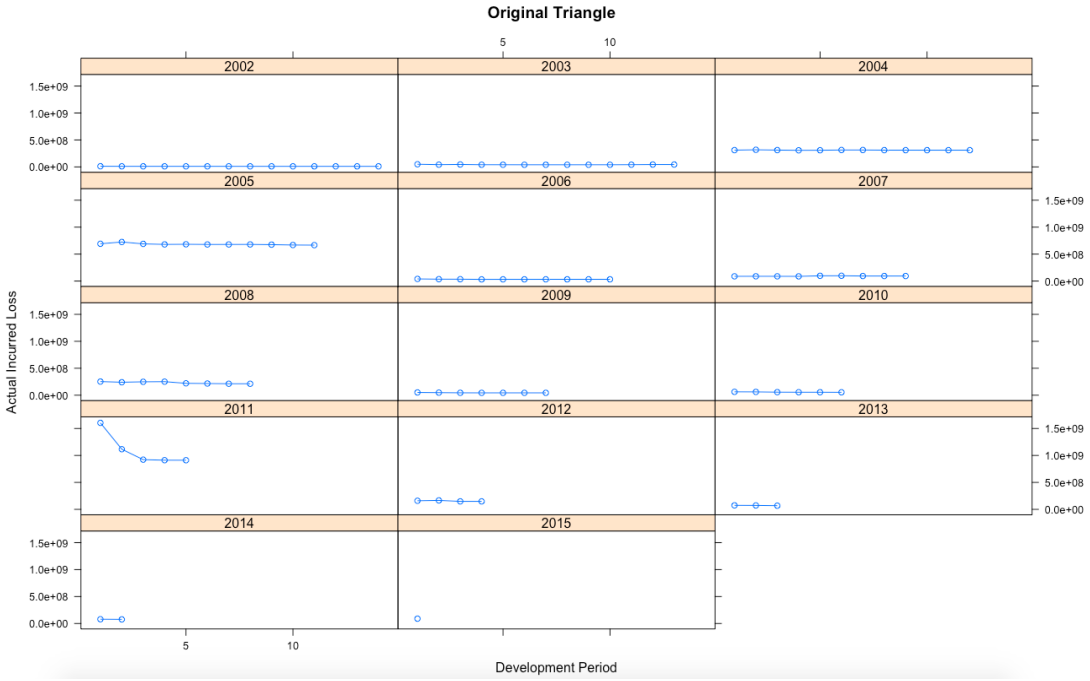
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ult
2002	11996603	10792631	10792628	10668514	10668514	10668514	10520128	10520127	10520127	10520127	10520127	10520127	10520127	10520127	10520127
2003	49187789	41852416	45431753	40424565	40205080	39662678	39589968	39624898	39849932	39852362	41352575	43876756	43876756	43876756	43876756
2004	311401148	317640891	312100106	307749470	309113489	313723410	313666739	310956479	310605835	310454590	310373654	310373654	310373654	310373654	310373654
2005	691715640	724745887	689620719	679900230	682174001	678886532	679103832	679555787	675271041	668542717	666444535	671088410	671088410	671088410	671088410
2006	39737708	35158035	34571737	31234870	32001853	31862913	32587614	32358857	32365698	32353732	32332393	32557690	32557690	32557690	32557690
2007	87819912	88427298	87951555	87064276	97481549	98369803	93424192	93441113	93197144	92596323	92535253	93180051	93180051	93180051	93180051
2008	252762792	240607829	248388404	251312852	220642439	216780733	212651207	212319763	211473823	210110499	209971924	211435036	211435036	211435036	211435036
2009	51202868	47784324	44586495	43678398	44005644	43806384	43804877	43717154	43542973	43262261	43233728	43534986	43534986	43534986	43534986
2010	62246474	61077405	56601247	55441916	55211458	54054835	53737676	53630061	53416385	53072021	53037018	53406587	53406587	53406587	53406587
2011	1604188980	1115698497	922010444	911971467	911380088	909126407	903792223	901982305	898388561	892596852	892008153	898223786	898223786	898223786	898223786
2012	159381368	166069442	148024435	147959829	146946963	146583589	145723529	145431705	144852266	143918436	143823516	144825698	144825698	144825698	144825698
2013	74126760	72585938	67836292	66983847	66525306	66360800	65971437	65839323	65577002	65154241	65111270	65564974	65564974	65564974	65564974
2014	79286781	76617179	69944342	69065407	68592616	68422999	68021535	67885317	67614843	67178945	67134638	67602442	67602442	67602442	67602442
2015	89991821	77665154	70901045	70010088	69530831	69358893	68951939	68813857	68539684	68097824	68052911	68527113	68527113	68527113	68527113

Figure.1 plot of Original triangle



Based on the triangle, we obtained the plots for comparing incurred loss over fourteen development periods. We can see that the aggregate incurred losses for claims whose accident years are in the first few years, from 2002 to 2008, which are labeled from 1 to 7 in the graph, have relatively flat incurred losses over their development periods. However, aggregate incurred losses for claims whose accident year are in the 10th development period, which is labeled as 0 in the plot, significantly decreases during its four development periods. The graphs for individual development periods also indicate similar patterns.

Figure.2 plot of Original triangle based on individual year



We also calculated the year to year ratio based on the original triangle. We can see that the yearly ratio gets close to 1 as development period goes by.

Table.3 Year to year ratio

Development Period	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6	6 to 7	7 to 8	8 to 9	9 to 10	10 to 11	11 to 12	12 to 13	13 to 14
Year to Year ratio	0.863	0.913	0.9874	0.993	0.998	0.9941	0.998	0.996	0.994	0.99934	1.007	1	1

Methods and Results

1 Data Cleaning

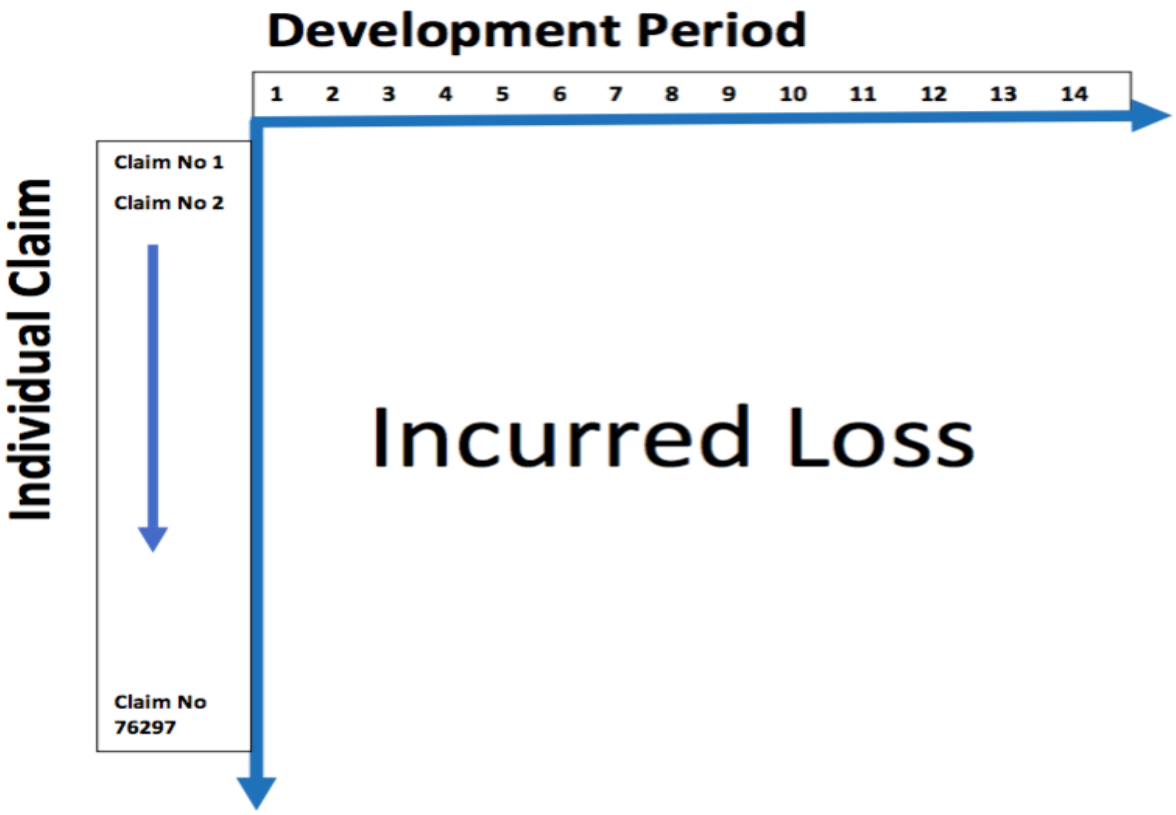
During data cleaning, many aspects of our data were changed and evaluated in order to use the most accurate data during our analysis. Four types of evaluations were removed from the property dataset which began with 166064 evaluations. Firstly, 28 evaluations whose "Loss_Incurred" was missing ('NA' values) were removed. Secondly, we noticed that some claims are evaluated multiple times in the same month within the same year. To be conservative, we kept the largest evaluations and removed the redundant evaluations (11107 observations) for such claims. Lastly, we noticed that the property dataset consists of claims that are evaluated only in December or in both June and December within the same year. In order to obtain a consistent evaluation period, we kept the evaluations in December and removed the June evaluations. Using the new evaluation period, we calculated the development period for every evaluation and removed 9805 evaluations with negative development periods. The dataset resulted is 76297 evaluations after data cleaning.

2 Reshape

To analyze the loss incurred for individual claims over fourteen development periods, we reshaped the property dataset. In our new data set, each row is a unique claim, and each column is the incurred loss for a development period. The new dataset consists of 11085 unique claims and fourteen columns of loss incurred. We also noted that "NA" values for an incurred loss

means that the claim has not been opened and thus did not have an evaluation in that development period. The graph below visualizes our new dataset after reshape:

Figure.3



3 Model Fitting Process

Baseline Model: Chain Ladder

We chose to use Chain Ladder model in our analysis because it was a model the client was interested in seeing and also because it is one of the most popular models to forecast future claims. Therefore, we decided to use Chain Ladder as our baseline case to compare to other models to see which one provides better results. We used the "ChainLadder" package in R to complete the process. Once we obtained our triangle for Chain Ladder, we compared the ratio to that of the original triangle and the ratios were very good since they were all close to 1. However, in the early years of the Chain Ladder triangle there seemed to be a bigger difference in the ratios between that of the original triangle. We suspect that the large difference in the predictions was caused by the fact that many claims have yet to close and new claims were also opening in the early developmental years. However, as development year grew, the ratio between the two triangles eventually converged to one. Our calculated ultimate losses for the Chain Ladder model were also very close to that of the original triangle. These comparisons already give an indication that the Chain Ladder model is a very good model for our goal, which is to predict the most accurate ultimate loss.

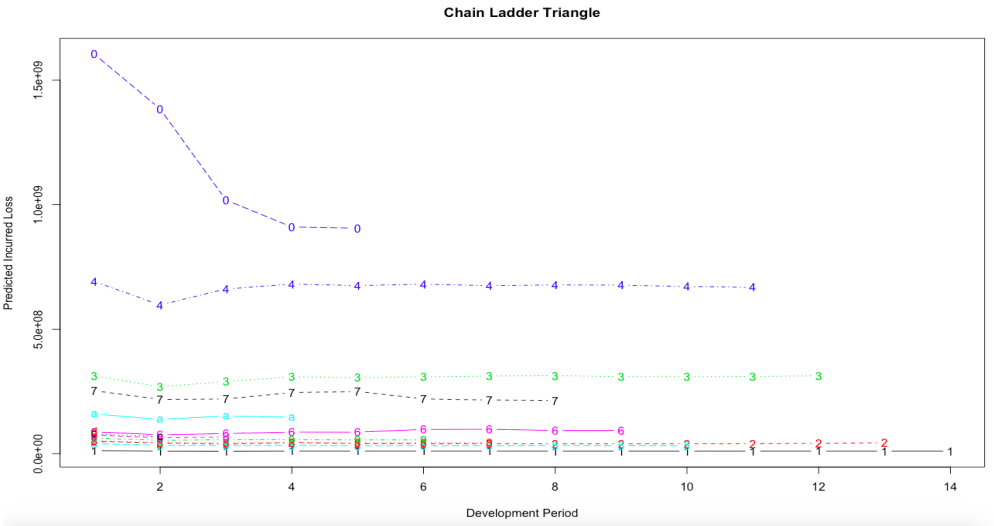
Table4. loss triangle of Chain-ladder

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	11996603	10353364	9852665.5	10657005	10595482	10642132	10605918	10499061	10478212	10452306	10513189	10593433	10520127	10520127
2003	49187789	42450272	38207354	44860848	40147837	40105660	39429962	39510686	39467021	39593029	39826078	41640726	43876756	NA
2004	311401148	268746849	289976521	308178192	305642756	308349106	311882678	313038595	309717543	308603429	310249835	312536379	NA	NA
2005	691715640	596967608	661625429	680954802	675245939	680487107	674903252	677743871	676848251	670917721	668101790	NA	NA	NA
2006	39737708	34294619	32096008	34137301	31021050	31922718	31675962	32522355	32229931	32157043	NA	NA	NA	NA
2007	87819912	75790743	80725880	86846337	86468273	97240494	97792630	93237102	93068818	NA	NA	NA	NA	NA
2008	252762792	218140505	219652517	245267104	249592477	220096829	215508800	212225356	NA	NA	NA	NA	NA	NA
2009	51202868	44189334	43622633	44026212	43379395	43896825	43549356	NA	NA	NA	NA	NA	NA	NA
2010	62246474	53720238	55757977	55889984	55062385	55074930	NA	NA	NA	NA	NA	NA	NA	NA
2011	1.604E+09	1.384E+09	1.019E+09	910424270	905728521	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	159381368	137550041	151605919	146164329	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	74126760	63973217	66264195	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	79286781	68426442	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	89991821	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table5. loss triangle of chain-ladder with ultimate loss

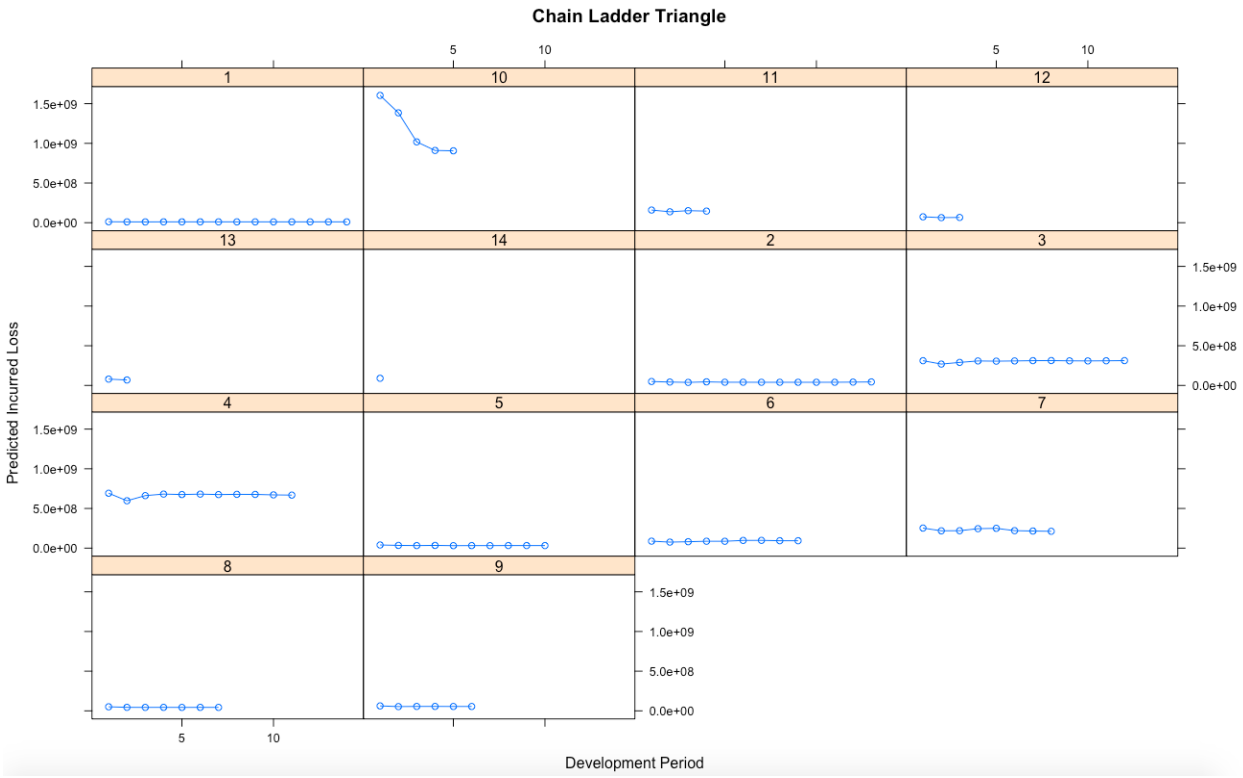
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ult
2002	11996603	10353364	9852666	10657005	10595482	10642132	10605918	10499061	10478212	10452306	10513189	10593433	10520127	10520127	10520127
2003	49187789	42450272	38207354	44860848	40147837	40105660	39429962	39510686	39467021	39593029	39826078	41640726	43876756	43876756	43876756
2004	311401148	268746849	289976521	308178192	305642756	308349106	311882678	313038595	309717543	308603429	310249835	312536379	325476767	325476767	325476767
2005	691715640	596967608	661625429	680954802	675245939	680487107	674903252	677743871	676848251	670917721	668101790	675849181	703832326	703832326	703832326
2006	39737708	34294619	32096008	34137301	31021050	31922718	31675962	32522355	32229931	32157043	32129695	32502275	33848013	33848013	33848013
2007	87819912	75790743	80725880	86846337	86468273	97240494	97792630	93237102	93068818	92457722	92379091	93450331	97319588	97319588	97319588
2008	252762792	218140505	219652517	245267104	249592477	220096829	215508800	212225356	211362685	209974863	209796290	212229113	221016337	221016337	221016337
2009	51202868	44189334	43622633	44026212	43379395	43896825	43549356	43454108	43277472	42993309	42956746	43454877	45254101	45254101	45254101
2010	62246474	53720238	55757977	55889984	55062385	55074930	54790768	54670933	54448702	54091188	54045186	54671901	56935560	56935560	56935560
2011	1604188980	1384454541	1018528715	910424270	905728521	900078263	895434254	893475822	889843950	884001170	883249373	893491642	930486147	930486147	930486147
2012	159381368	137550041	151605919	146164329	145056106	144151194	143407437	143093786	142512127	141576382	141455979	143096320	149021140	149021140	149021140
2013	74126760	63973217	66264195	65391967	64896163	64491317	64158570	64018247	63758021	63339381	63285514	64019381	66670066	66670066	66670066
2014	79286781	68426442	62292375	61472428	61006342	60625763	60312960	60181048	59936419	59542873	59492234	60182114	62673919	62673919	62673919
2015	89991821	77665154	70702886	69772232	69243217	68811253	68456217	68306494	68028837	67582154	67524679	68307704	71135945	71135945	71135945

Figure4. Plot of Chain-ladder triangle



As we can see from the Chain Ladder triangle, the trends are very similar to that of the original triangle. Both graphs have a negative trend in loss incurred for the claims in the 2011 accident year. Otherwise, both triangles have similar trends in other years. One big difference would be the lower dip in the second development year for the 2005 accident year in the chain ladder model. This difference in the 2005 prediction can be caused by the reality of higher number of new claims opening in the original triangle that chain ladder did not account for in its model. Otherwise, the trends in chain ladder are extremely similar to that of the original triangle and so far it seems like a good model for predicting loss incurred.

Figure.5 plot of chain-ladder triangle based on individual year



The number labels for this graph is that 2002 is the equivalent of 1. Number 0 represents 2011. The rest of the years exceeding 2011 are represented by a, b, c… We also made sure that the y-axis are the same as that of other triangles that we created when we completed our visual comparison.

Table.6 Year to year ratio of chain-ladder

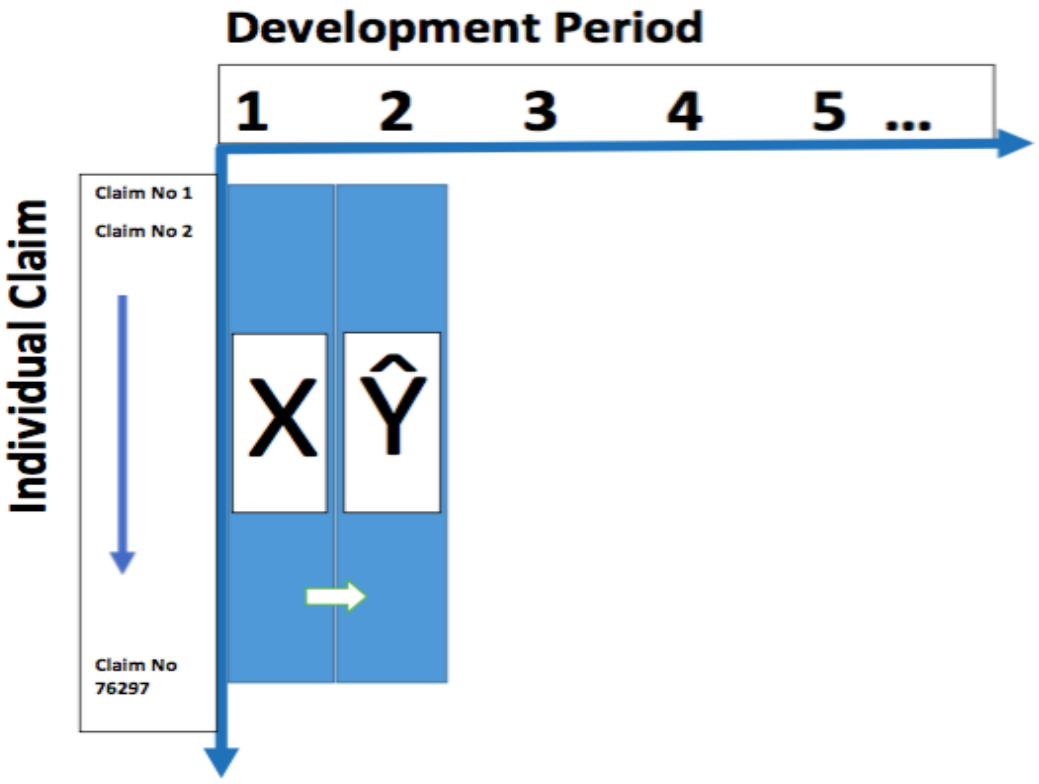
Development Period	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6	6 to 7	7 to 8	8 to 9	9 to 10	10 to 11	11 to 12	12 to 13	13 to 14
Year to Year Ratio	0.86302459	0.91035532	0.98683712	0.99241797	0.99376164	0.99484044	0.99781287	0.99593512	0.99343393	0.99914955	1.01159612	1.04140442	1

We then took a closer look at each year's incurred loss created by the chain ladder triangle. Aside from the difference in 2005 that we previously mentioned, there are only extremely minor up and downs for the trend in each of the accident years. Just to make sure our predictions are close, we created a ratio test that would measure how close chain ladder's values

are to that of the original triangle. As shown above, the values are mostly close to 1, aside from 2002 and 2003, and eventually the probability grows closer and becomes 1, which means the prediction is an exact match as the incurred loss in the original triangle. One explanation for the difference in 2002 and 2003 for the chain ladder model is that it may not have accounted for the growing number of claims opening in those two years. Furthermore, chain ladder model's later incurred losses were close to that in original triangle since the original triangle would have more closed claims and also more consistent case losses. These factors make it easier to predict the future years using chain ladder.

Exploratory Models (Tweedie, Gaussian, and Poisson Models)

Figure.6



Model Fitting Technique:

For each of the model that we chose, we will use the same model fitting technique because we aim to use loss incurred in the previous year to find out which model produced the closest results compared to the actual data. We use the incurred loss in the previous year as the

independent variable to predict the incurred loss in the current year under our selected distribution. As shown in the graph, each column in the reshape data was set as the loss incurred during an individual development period. To predict the incurred loss over fourteen development periods, we fit the model thirteen times and aggregated the estimated individual incurred loss to obtain the triangle.

Tweedie Model

The first model we chose was Tweedie because two of its properties meet the characteristics of insurance data. As a model of the generalized linear model family, Tweedie model has a Compound Poisson response distribution with both discrete and continuous property when its parameter is set between 1 and 2. The insurance data is also considered as both discrete and continuous: claim counts follows a discrete Poisson distribution and the severity of claims follow a gamma distribution. Moreover, under the Tweedie model with Compound Poisson response distribution, the independent variables are non-negative with mass at zero; the insurance data has many claims with zero values for their incurred loss. Therefore, because Tweedie model with parameter between 1 and 2 well fit two characteristics of insurance data, we decide to fit our data using a Compound Poisson response distribution.

During the model, fitting process where we fit the model thirteen times under the Compound Poisson response distribution, we first created subsets, each consisting the column of our independent variable incurred loss in the previous year, and the column of our dependent variable, the incurred loss in the current year. Within each subset, the incurred loss with zero values are eliminated in both columns, because our focus is to predict the incurred loss in the current year based on the positive incurred loss in the previous year. Also, claims with the "NA" values in both columns are removed: since the Na values indicates that such claims have not been opened previous development period, it would not be reasonable to predict the incurred loss in the next period.

Secondly, after creating thirteen subsets, we used the maximized likelihood estimation to estimate the parameter used for Tweedie model given the claims in that subsets. Using the such maximum likelihood estimation method, we can find the parameter values that maximize the

likelihood of making the observations. Limiting the parameter between 1 and 2 for the Compound Poisson response distribution, we found that the best parameters yield from each subtest range from 1.6 to 1.8. However, since the parameters re the largest maximized value for a certain subset, we chose 1.6 over 1.8 as the parameter for our Tweedie model to keep our analysis consistent over fourteen development periods.

After determining the appropriate parameter, we fit the model within each subset on individual claim level to obtain the individual predicted incurred loss over thirteen development periods. Lastly, we aggregate the individually predicted incurred loss to build the new triangle and calculate the ultimate loss.

In the graph below, we marked our Tweedie incurred loss with pink outlines and values below the triangle indicate the predicted IBNR values. The pink box to the very right of this graph shows our final ultimate loss for the Tweedie model.

Table.7 loss triangle of Tweedie with ultimate loss

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ult
2002	11996603	4231390	4593057	4087759	3332746	2714564	3688084	5287814	5704318	5884207	6511525	4042163	23746918	10520127	4660524
2003	49187789	53770947	55881193	53679141	52850728	42566271	56762641	75670959	89504241	92332585	100921736	62501275	30649971	13578250	6015303
2004	311401148	190200028	187870377	174647486	171017372	147362617	199295254	259400490	304460312	313780868	347170046	298228222	243790940	108001876	47845934
2005	691715640	185815630	179055780	170601683	167956541	432809042	471143720	520564514	541084532	546124266	574088712	460646215	376561859	166820749	73903295
2006	39737708	68109096	63136995	63165330	61434249	48018008	64531688	84873847	99880138	103602732	111233550	89253303	72961394	32322642	14319261
2007	87819912	92197396	79406200	76572359	88089846	59008295	79473332	102924181	121239421	123696650	132807478	106564126	87112374	38591671	17096504
2008	252762792	258206250	253103647	254975916	235617645	187122238	253497151	330066113	365678382	373089796	400569577	321415237	262745498	116398939	51565918
2009	51202868	297987101	295029118	284263517	276825370	218813678	296967568	362866221	402017437	410165355	440375922	353355670	288855663	127966009	56690248
2010	62246474	472353515	470301690	451751236	441117152	349417005	437489800	534571071	592248270	604251705	648757626	520560218	425539420	188518309	83515536
2011	1604188980	1030924246	868443790	914354042	904657606	898371956	1124812365	1374414101	1522705618	1553567166	1667994543	1338391362	1094087225	484691813	2.15E+08
2012	159381368	155020861	126005896	119312164	117109305	116295618	145608674	177919999	197116562	201111636	215924435	173256801	141631258	62744094	27796274
2013	74126760	87722552	84670460	84165055	82611117	82037127	102715110	125508130	139049748	141867949	152317177	122218621	99909366	44260869	19608017
2014	79286781	102523025	94416123	93852546	92119747	91479690	114537732	139954253	155054526	158197105	169849051	136286119	111409043	49355344	21864921
2015	89991821	77665265	71523965	71097032	69784369	69299500	86766883	106020907	117459964	119840593	128667405	103242150	84396777	37388634	16563547

Development Period	1 to 2	2 to 3	3 to 4	4 to 5	5 to 6	6 to 7	7 to 8	8 to 9	9 to 10	10 to 11	11 to 12	12 to 13	13 to 14
Year to Year Ratio	0.86302582	0.92092604	0.99403092	0.98153701	0.9930519	1.25205641	1.22190522	1.10789435	1.02026757	1.07365461	0.80239553	0.81746435	0.44301021

Table.8 Year to year ratio of tweedie model

Figure7. Plot of tweedie triangle

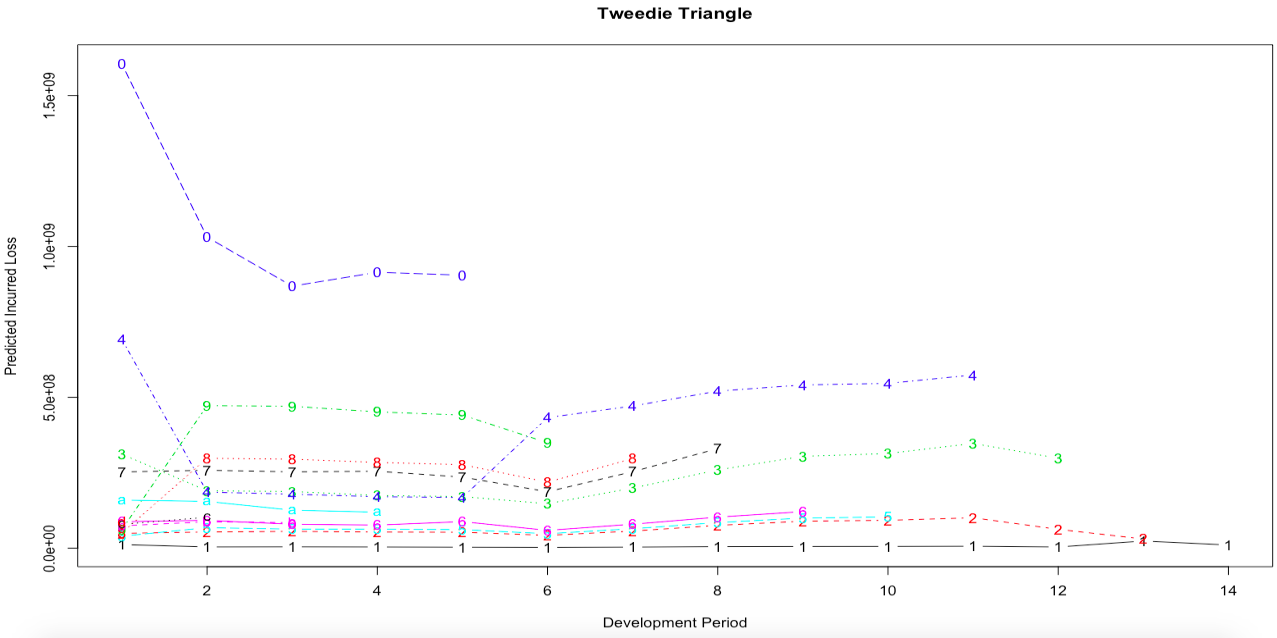
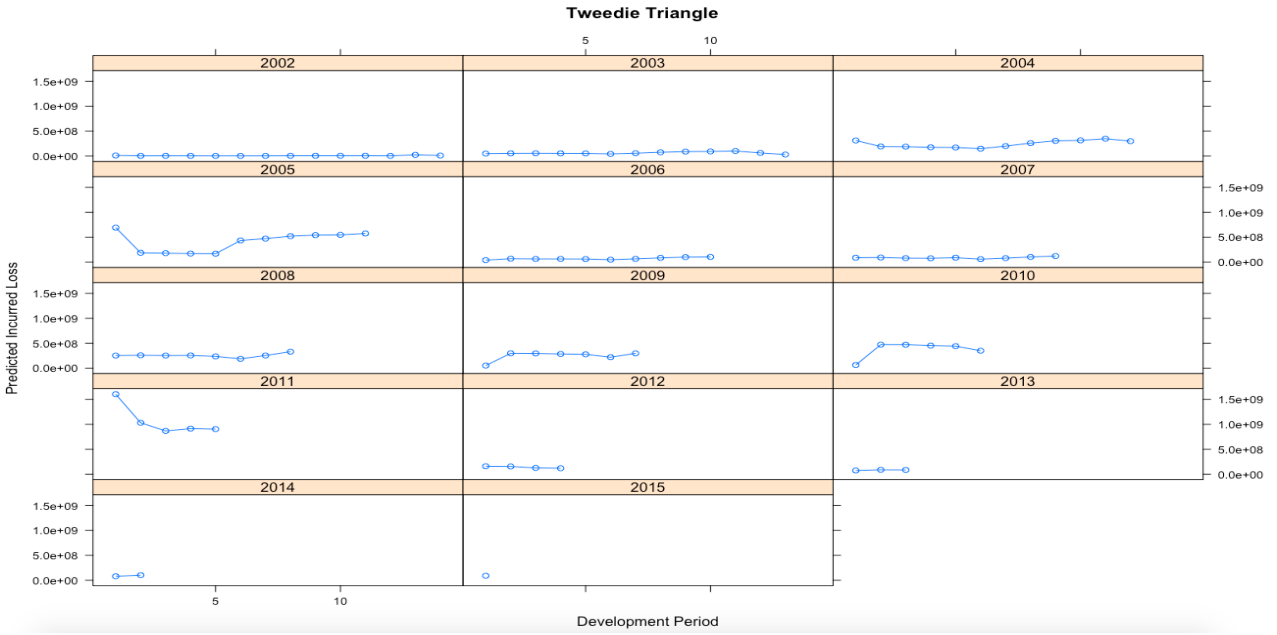


Figure.8 plot of tweedie triangle based on individual year



The graph for Tweedie has much more variation from the original triangle than chain ladder. Tweedie's graph showed a negative incurred loss that was shown in both the original and the chain ladder triangles. However, Tweedie's triangle had a huge difference in the 2004, 2005, and 2010 than the result in the original triangle. These trends were not consistent with that of common sense and even slight variation in new opening claims would not make such a big

difference in the skew of the incurred losses. Therefore, it could be said that the differences are caused by the Tweedie model or parameters in the Tweedie model that were not accurately used. Furthermore, we are using the Tweedie model along with GLM and since we only used incurred loss as a predictor, having more predictors could have potentially made the model more accurate. The individual incurred losses also showed the same trends for the three particular years that had a significant change from the original triangle. Finally, we used the ratio test to mathematically calculate the proportion differences between each of the development years from Tweedie and the original triangles. It could be said that the two triangles are similar since the ratios are generally around 1. However, the accuracy is nowhere close to that of the ones created by Chain Ladder and the final development year is only 0.4 times of the original triangle. Therefore, we can conclude that Tweedie wasn't the best model to use from the way we coded the model and the predictors included.

Gaussian Model

The second model we chose is the Gaussian model, which is also a model of the generalized linear model family. As one of the most popular methods used in the insurance industry, Gaussian is chosen because it fits the appropriate assumptions for the property dataset, where the number of event is large. In this case, we consider the incurred loss has a continuous distribution.

Same as the process of fitting Tweedie, we built the triangle of GLM with Gaussian family.

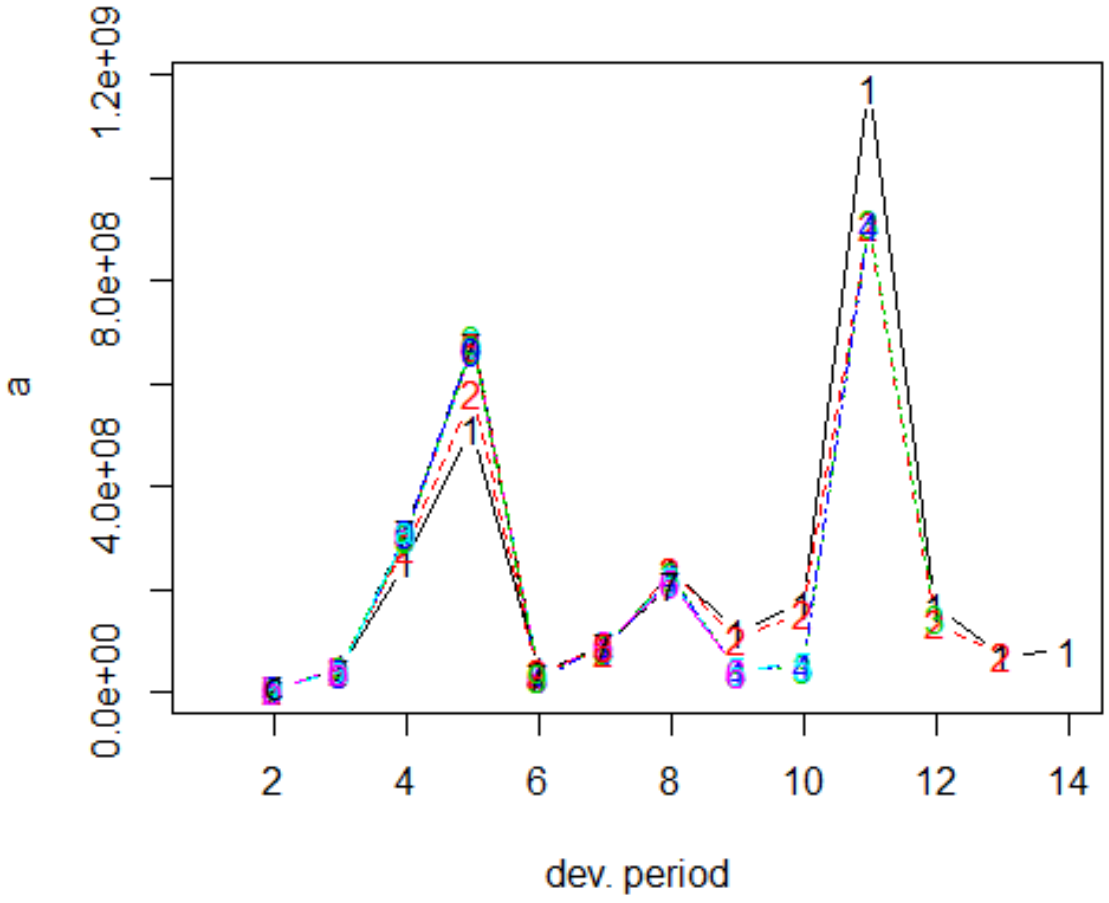
Table9. loss triangle of Gaussian

	Development year													
Accident year	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	11996603.4	9403162.94	48535383.8	257577247	514523819	44759893.1	85105181.1	240477800	118023700	174277627	1177351827	168456054	73781278.8	81625721.8
2003	49187789.3	9140764.45	44456529.4	284164215	582600108	41196848	83196830.1	239821574	104818673	155032024	911461185	139157012	71899638.3	NA
2004	311401148	10629772.3	44896853.4	303567161	687467939	33937785.5	86920807.8	230063748	42785260.6	52312764.4	915308143	146722051	NA	NA
2005	691715640	10660998.9	40269271.3	307356948	679385088	31049745.9	86191598.4	224614588	43253848.7	54771937	911339429	NA	NA	NA
2006	39737708.2	10643654.4	39875469.3	308748718	680035513	31944067.2	97179901.2	220183839	43990514.9	55230035.5	NA	NA	NA	NA
2007	87819911.6	10621129.8	39428425.8	312264653	675944002	31649208.1	97188522.3	215031490	43231997.9	NA	NA	NA	NA	NA
2008	252762792	10499336.4	39602975.2	312922510	677698534	32592193.1	93310224.4	212162139	NA	NA	NA	NA	NA	NA
2009	51202867.6	10488169.5	39238220.8	309518360	677787202	31938311.8	92856317.9	NA	NA	NA	NA	NA	NA	NA
2010	62246474.5	10463199.2	39249673.2	308271680	672023257	31716670.5	NA	NA	NA	NA	NA	NA	NA	NA
2011	1604188980	10531733.2	40019975	310989419	667150892	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	159381368	10567650.4	41831111.5	312372898	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	74126760.1	10520127	43876761	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	79286781.4	10520127	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	89991820.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

After building the predicted triangle of Gaussian, we predict the future loss and calculated the ultimate loss.

Table.10 Ultimate loss of Gaussian

Ultimate loss	Ratio
90304188	8.58394466
88001164	2.00564426
85046215	0.27401235
79891141	0.11904712
43440771	1.33427067
47374401	0.50841785
67209981	0.31787533
73988874	1.69952677
65379137	1.2241774
73091632	0.08137352
73937302	0.51052612
75557251	1.15240267
73135441	1.08184614
24248399	0.35385117



As one can see, the plot for Gaussian was nowhere close to that of the original triangle, therefore, we already know that it will most likely be a bad model for predicting the property dataset.

Poisson Model

For Poisson distribution, it has some unique properties that are different with other models. For example, Poisson distribution is a discrete distribution. While our data set is not completely discrete, it can be viewed as partial discrete for several reasons. First, the claim numbers of the insurance data set is definitely discrete. Second, the incurred loss we are cared about has a distance of one year so it can so be viewed as sort of discrete. What's more, Poisson also has other interesting properties like it is infinitely invisible. Thus, Poisson model is one that worth to try in our case.

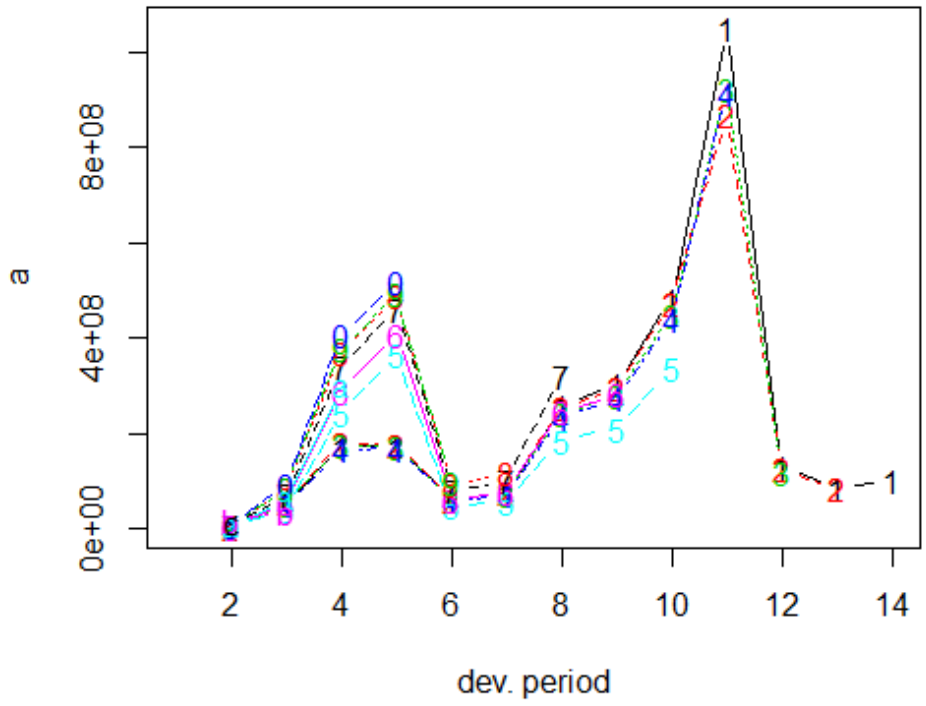
By fitting the data in GLM with Poisson family model, we can build a new predicted triangle, we need use this triangle to compare with our initial triangle, to see if they are close enough.

Table.11 predicted triangle- poisson

	Development year													
Accident year	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	11996603.4	4435145.14	51879227.7	173406488	173922162	62023151.1	76024824.3	257150625	303263303	483742787	1052460111	128728579	85079876.2	100743540
2003	49187789.3	4318911.2	55331847.7	181586339	176414458	62417651.4	77943154.6	252987627	295876701	473476936	870330257	125188506	83207794.6	NA
2004	311401148	3117891.3	53440803.2	175267532	170878143	61837426.7	74355130.6	240536582	280978907	450355359	924491843	118105686	NA	NA
2005	691715640	3374679.94	52485632.8	166213786	169458913	59703523.1	72932711.3	234842605	275368897	439445933	914939028	NA	NA	NA
2006	39737708.2	2973924.09	40843448.3	244025217	364951041	46074526.6	58289384	185027985	209949221	335330362	NA	NA	NA	NA
2007	87819911.6	3953829.76	54881224.1	289250931	408150156	61959183.3	77556995.3	247022004	282540723	NA	NA	NA	NA	NA
2008	252762792	5071691.28	72008182.2	340845927	461410075	80380900.8	100020523	319016011	NA	NA	NA	NA	NA	NA
2009	51202867.6	5890574.19	84757397.9	372265266	488458713	94064698	116389944	NA	NA	NA	NA	NA	NA	NA
2010	62246474.5	6057187.5	86232818	378673536	493887810	96873127.8	NA	NA	NA	NA	NA	NA	NA	NA
2011	1604188980	6636546.32	94772937.7	406405032	520877510	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	159381368	4337114.85	60711762.1	299722783	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	74126760.1	18105803.8	36291084.2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	79286781.4	10520127	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	89991820.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

table.12 Poisson ultimate loss

Ultimate loss	Ratio
119290968	11.33931
116666112	2.65895
109751862	0.353612
111089306	0.165536
82932636	2.547252
111682890	1.198571
146597206	0.693344
172922219	3.972029
177983475	3.332613
205825754	0.229148
148847542	1.02777
77143245	1.176592
243572734	3.603017
47243172	0.689408



Once again, Poisson model had the same issue with Gaussian in that none its values were even relevant for analysis since the predicted incurred losses were so far off from the original triangle and plot. Therefore, we can also begin ruling out Poisson as a good model for predicting the Property incurred losses.

Model Comparison

After applying different models in the case, we should check these models to see which one is best fit in the case, and which one probably not suit this case. Therefore, model comparison is necessary. To make sure the results are correct, we use different method to compare them:

- R-squared
- Ration triangle
- Sum of Square
- Residual plots

1.R-squared

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determinations for multiple regressions.

Firstly, numeric the triangles and then set them as matrix to be easier to fit in the linear regression. Secondly, fitting the linear regression of original triangle and our predicted triangles, we can get the Multiple R-squared from the output. In this case, as the R-squared closer to 1, the model will be better fit.

Original Triangle ~ Predicted Triangle

Table.1 R-Square of different models compared to original triangle.

Chain-Ladder	0.9892
Tweedie	0.7136
Gaussain	0.1037
Poisson	0.1329

From the table.1, it is easy to observe that the value of R-squared from Chain-Ladder model is closest to 1, which means the predicted triangle of Chain-Ladder is the very close to the original triangle. The Tweedie model is pretty good to fit as well. But for Poisson and Gaussian, they may have problem to use in our case. We need do more things to double check our results.

But one thing we need focus on further study is that we use the linear regression on triangle level, not individual level. There may have big differences if we compare the data on individual level, and we will go deep analysis in the future.

2. Ratio Triangle

To compare each value in the triangle, simply divide the original triangle by the predicted triangle and check the ratio will be an effective way. Let denote the incurred loss amount in original triangle, for accident year, w, and development year, d, for $1 \leq w \leq 14$ and $1 \leq d \leq 14$.

Similarly, Let denote the incurred denote the incurred loss amount we predicted by different models. The ratio is given by

$$f = \frac{C_{w,d}}{P_{w,d}}$$

Then, we built the ratio triangle and highlight the numbers that are close to 1 within 0.05 in Chain-ladder and Tweedie ratio triangle, and 1 within 0.5 in Gaussian and Poisson model. The more highlight part means the closer to original triangle, the model are better fits.

Table2. Ratio triangle of tweedie.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	1	2.55	2.35	2.61	3.2	3.93	2.85	1.99	1.84	1.79	1.62	2.6	0.44	**1**
2003	1	0.78	0.81	0.75	0.76	**0.93**	0.7	0.52	0.45	0.43	0.41	0.7	1.43	NA
2004	1	1.67	1.66	1.76	1.81	2.13	1.57	1.2	**1.02**	**0.99**	0.89	**1.04**	NA	NA
2005	1	3.9	3.85	3.99	4.06	1.57	1.44	1.31	1.25	1.22	**1.16**	NA	NA	NA
2006	1	0.52	0.55	0.49	0.52	0.66	0.5	0.38	0.32	0.31	NA	NA	NA	NA
2007	1	**0.96**	**1.11**	**1.14**	**1.11**	1.67	**1.18**	**0.91**	0.77	NA	NA	NA	NA	NA
2008	1	**0.93**	**0.98**	**0.99**	**0.94**	**1.16**	0.84	0.64	NA	NA	NA	NA	NA	NA
2009	1	0.16	0.15	0.15	0.16	0.2	0.15	NA	NA	NA	NA	NA	NA	NA
2010	1	0.13	0.12	0.12	0.13	0.15	NA	NA	NA	NA	NA	NA	NA	NA
2011	1	**1.08**	**1.06**	**1**	**1.01**	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	1	**1.07**	**1.17**	1.24	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	1	0.83	0.8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	1	0.75	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

By observed the highlight part in different ratio triangle, it is easily to conclude that the Chain-ladder fits the model best. And the results are same as the R-squared tell.

3. Sum of Squared

The sum of squares represents a measure of variation or deviation from the mean. It is calculated as a summation of the squares of the differences from the mean. The calculation of the total sum of squares considers both the sum of squares from the factors and from randomness or error.

We calculated the sum of squared of each predicted triangle with original triangle,

$$(C_{w,d} - P_{w,d})^2 \qquad 1 \leq d \leq 14$$

Then add every cell together. After calculating the sum of squared values, we have the table below.

Table.3 sum of squared of different models

Chain Ladder	1.05E+17
Tweedie	2.49E+18
Gaussian	9.73E+18
Poisson	1.16E+19

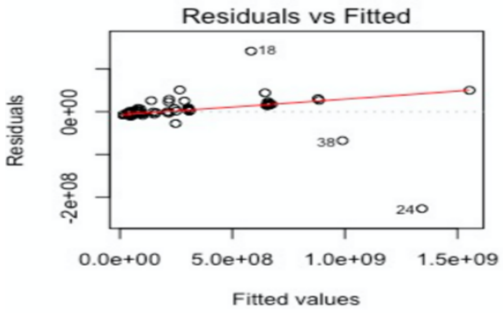
By comparing the total sum of squares, we determined the proportion of the total variation that is explained by the regression model. The smaller this value is, the better the predicted triangle equal to the original triangle. The model with the smallest sum of squared value will be more reliable to predict the future incurred loss. Same as what we analyzed before, the Chain-Ladder and Tweedie are the two best models (see Appendix).

4.Residual Plots

When conducting a residual analysis, a "residuals versus fits plot" is the most frequently created plot. It is a scatter plot of residuals on the y-axis and fitted values (estimated responses) on the x-axis. The plot is used to detect non-linearity, unequal error variances, and outliers. Trend should be roughly flat with equal vertical spread when there are no problems.

Here's what the corresponding residuals versus fits plot looks like for the data set's simple linear regression model with original triangle as the response, and the Chain-Ladder predicted triangle as the predictor:

Figure.1 Chain-ladder

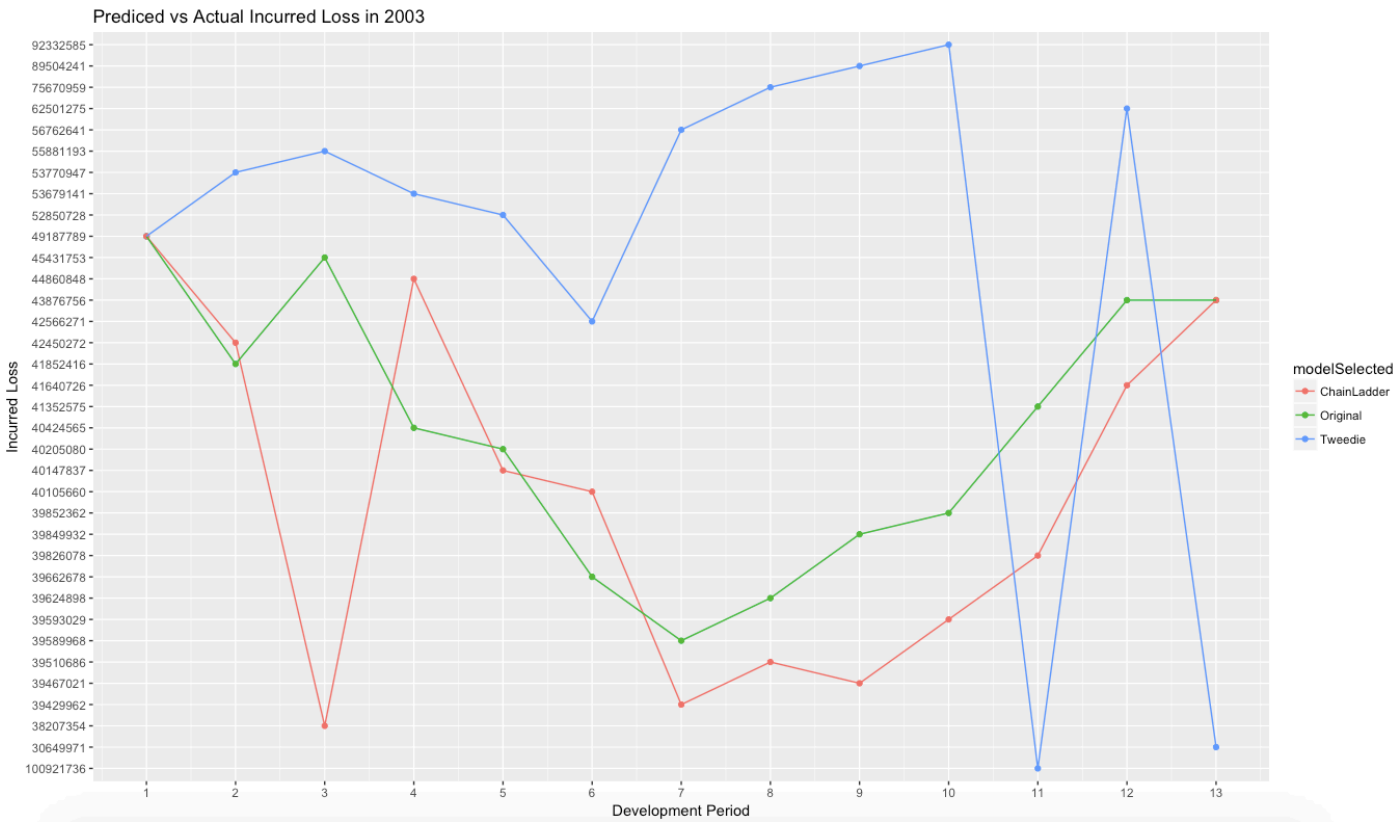


The plot suggests that there is a pretty flat linear relationship between original triangle and predicted triangle. It also suggests that there are no unusual data points in the data set, except point 18. Furthermore, it illustrates that the variation around the estimated regression line is constant which is suggesting that the assumption of equal error variances is reasonable. Compared plots of different models with each other, can see the Chain-ladder is the best model is this case (see Appendix).

To make the residual of each models more visualized, we create residual triangle for the superior two models, Chain-Ladder and Tweedie. Here is the residual triangle of Chain-ladder model, (the residual triangle of Tweedie model see in Appendix);

Table.4 Residual triangle of tweedie

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	0	4.3E+13	3.8E+13	4.3E+13	5.4E+13	6.3E+13	4.7E+13	2.7E+13	2.3E+13	2.1E+13	1.6E+13	4.2E+13	1.7E+14	0
2003	0	1.4E+14	1.1E+14	1.8E+14	1.6E+14	8.4E+12	2.9E+14	1.3E+15	2.5E+15	2.8E+15	3.5E+15	3.5E+14	1.7E+14	NA
2004	0	1.6E+16	1.5E+16	1.8E+16	1.9E+16	2.8E+16	1.3E+16	2.7E+15	3.8E+13	1.1E+13	1.4E+15	1.5E+14	NA	NA
2005	0	2.9E+17	2.6E+17	2.6E+17	2.6E+17	6.1E+16	4.3E+16	2.5E+16	1.8E+16	1.5E+16	8.5E+15	NA	NA	NA
2006	0	1.1E+15	8.2E+14	1E+15	8.7E+14	2.6E+14	1E+15	2.8E+15	4.6E+15	5.1E+15	NA	NA	NA	NA
2007	0	1.4E+13	7.3E+13	1.1E+14	8.8E+13	1.5E+15	1.9E+14	9E+13	7.9E+14	NA	NA	NA	NA	NA
2008	0	3.1E+14	2.2E+13	1.3E+13	2.2E+14	8.8E+14	1.7E+15	1.4E+16	NA	NA	NA	NA	NA	NA
2009	0	6.3E+16	6.3E+16	5.8E+16	5.4E+16	3.1E+16	6.4E+16	NA	NA	NA	NA	NA	NA	NA
2010	0	1.7E+17	1.7E+17	1.6E+17	1.5E+17	8.7E+16	NA	NA	NA	NA	NA	NA	NA	NA
2011	0	7.2E+15	2.9E+15	5.7E+12	4.5E+13	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	0	1.2E+14	4.8E+14	8.2E+14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	0	2.3E+14	2.8E+14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	0	6.7E+14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA



The graph above indicates the differences between our top three models for the year 2003. We realized that specific year had a huge difference in the accuracy of our predictions so we wanted to give a closer inspection to see how different the prediction really was. Even though there was major difference, it is still clear that Chain Ladder followed the general pattern as the original triangle and was the closer one in terms of prediction.

Conclusions

1. Recommendation

Considering the results of model comparisons using various criteria, we conclude that Chain-Ladder model is the best model for the Property Dataset. Moreover, Tweedie fits the best among the exploratory models we chose. On the other hand, we do not suggest use Poisson and Gaussian the predicted incurred loss and ultimate loss obtained by such models are less accurate.

2. Future Study and Unsolved Issues

(1) Adding Training and Testing Datasets

We analyzed the predicted incurred loss without splitting our data into training and testing datasets. Our analysis for the Original Triangle and Chain Ladder triangle is based on all individual evaluations in Property Dataset, and our analysis using Tweedie, Gaussian, and Poisson models is based on all the individual claims in the reshape dataset. However, analysis based on the whole dataset does not guarantee the overall reliability of the results. Therefore, we suggest that in order to obtain more reliability, splitting the dataset into training and testing datasets after data cleaning and reshape before further analysis to assess the overall strength of the results obtained from the models

(2) Sub -setting Dataset based on Lines of Business

We analyzed the Property dataset without sub setting the dataset according to any categorical variables However, we think that the Line Of Business might be closely related to the incurred loss. Therefore, we think that further analysis based on different lines of business is worthy investigation.

(3) Predicting Claim Counts

During our analysis, we aimed to find the most preferable model for predicting the positive incurred loss using the loss in previous year. However, we think that it is interesting to

look further into the claim counts. We suggest further investigate in claim count using discrete Poisson distribution.

(5) Obtaining R-squared on Individual Claim Level

During our model comparison, we calculated the R-squared based on the triangle level using the triangles that we obtained from our predicted incurred loss. However, it is very likely that the R squares for each model calculated on triangle level will results in inaccuracy because the aggregation causes the R square do not effectively reflect the closeness of data to the regression line. Thus, we suggest to calculate the R-Square based on the individual claim level for more accuracy.

(5) Obtaining parameter using Residual Maximum Likelihood (REML)

During our analysis using the Tweedie model, we obtained the parameter using the Maximized Likelihood Estimator (MLE) method. However, the residual maximum likelihood (REML) method for estimating the variances maybe more appropriate, because the REML method considers the fact that the means of data were estimated. As a result, the REML method is able to conduct less biased estimators for the variances compared to MLE.

(6) Exploring Categorical Variables

We conducted our analysis using the information on previous incurred loss. However, for future study, we think that many explanatory variables are worth exploring. In particular, some explanatory variables have larger impact on the claim counts (variance), and some have large impact on the claim amount (mean). Therefore, we think that further study in these two situations might be interesting.

Appendices

1.Details of Analysis

Table.1.1 Ratio triangle of chain-ladder model

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	1	1.04	1.1	1	1.01	1	0.99	1	1	1.01	1	0.99	1	1
2003	1	0.99	1.19	0.9	1	0.99	1	1	1.01	1.01	1.04	1.05	1	NA
2004	1	1.18	1.08	1	1.01	1.02	1.01	0.99	1	1.01	1	0.99	NA	NA
2005	1	1.21	1.04	1	1.01	1	1.01	1	1	1	1	NA	NA	NA
2006	1	1.03	1.08	0.91	1.03	1	1.03	0.99	1	1.01	NA	NA	NA	NA
2007	1	1.17	1.09	1	1.13	1.01	0.96	1	1	NA	NA	NA	NA	NA
2008	1	1.1	1.13	1.02	0.88	0.98	0.99	1	NA	NA	NA	NA	NA	NA
2009	1	1.08	1.02	0.99	1.01	1	1.01	NA	NA	NA	NA	NA	NA	NA
2010	1	1.14	1.02	0.99	1	0.98	NA	NA	NA	NA	NA	NA	NA	NA
2011	1	0.81	0.91	1	1.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	1	1.21	0.98	1.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	1	1.13	1.02	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	1	1.12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table.1.2 Ratio triangle of Poisson

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	1	1.15	0.22	0.04	0.02	0.24	0.12	0.04	0.09	0.06	0.01	0.06	0.14	0.13
2003	1	4.58	1.02	0.14	0.07	0.96	0.48	0.17	0.38	0.26	0.05	0.32	0.61	NA
2004	1	29.88	6.95	1.01	0.45	9.24	3.61	1.35	7.26	5.93	0.34	2.12	NA	NA
2005	1	67.98	17.13	2.21	1	21.86	7.88	3.03	15.61	12.21	0.73	NA	NA	NA
2006	1	3.3	0.87	0.1	0.05	1	0.34	0.15	0.74	0.59	NA	NA	NA	NA
2007	1	8.33	2.23	0.28	0.14	3.11	0.96	0.43	2.16	NA	NA	NA	NA	NA
2008	1	22.92	6.27	0.8	0.33	6.65	2.28	1	NA	NA	NA	NA	NA	NA
2009	1	4.56	1.14	0.14	0.06	1.37	0.47	NA	NA	NA	NA	NA	NA	NA
2010	1	5.84	1.44	0.18	0.08	1.7	NA	NA	NA	NA	NA	NA	NA	NA
2011	1	105.94	23.04	2.93	1.37	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	1	15.71	3.54	0.47	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	1	6.9	1.55	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	1	7.28	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table.1.3 Ratio triangle of Poisson

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	1	2.43	0.21	0.06	0.06	0.17	0.14	0.04	0.03	0.02	0.01	0.08	0.12	0.1
2003	1	9.69	0.82	0.22	0.23	0.64	0.51	0.16	0.13	0.08	0.05	0.35	0.53	NA
2004	1	101.88	5.84	1.76	1.81	5.07	4.22	1.29	1.11	0.69	0.34	2.63	NA	NA
2005	1	214.76	13.14	4.09	4.03	11.37	9.31	2.89	2.45	1.52	0.73	NA	NA	NA
2006	1	11.82	0.85	0.13	0.09	0.69	0.56	0.17	0.15	0.1	NA	NA	NA	NA
2007	1	22.36	1.6	0.3	0.24	1.59	1.2	0.38	0.33	NA	NA	NA	NA	NA
2008	1	47.44	3.45	0.74	0.48	2.7	2.13	0.67	NA	NA	NA	NA	NA	NA
2009	1	8.11	0.53	0.12	0.09	0.47	0.38	NA	NA	NA	NA	NA	NA	NA
2010	1	10.08	0.66	0.15	0.11	0.56	NA	NA	NA	NA	NA	NA	NA	NA
2011	1	168.11	9.73	2.24	1.75	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	1	38.29	2.44	0.49	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	1	4.01	1.87	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	1	7.28	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Figire.1.1 Tweedie

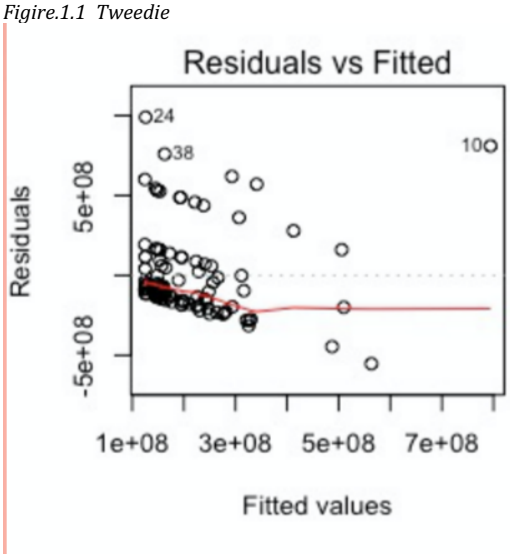


Figure.1.2 Poisson

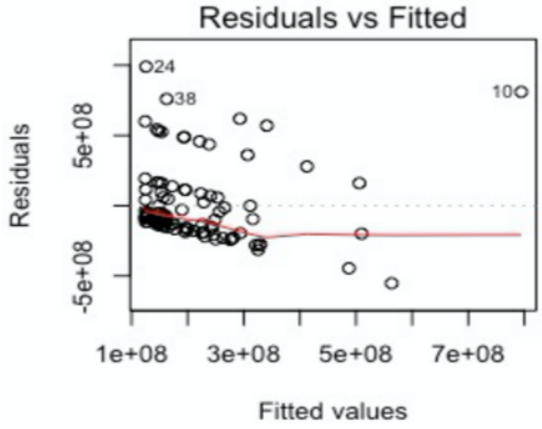


Figure.1.3Gaussian

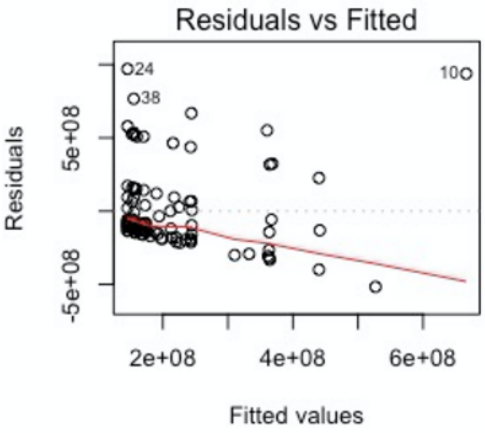


Table.6 residual triangle of chain-ladder

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2002	0	1.9E+11	8.8E+11	1.3E+08	5.33E+09	7E+08	7E+09	4E+08	1.8E+09	4.6E+09	5E+07	5E+09	0	0
2003	0	3.6E+11	5.2E+13	2E+13	3.28E+09	2E+11	3E+10	1E+10	1.5E+11	6.7E+10	2E+12	5E+12	0	NA
2004	0	2.4E+15	4.9E+14	1.8E+11	1.2E+13	3E+13	3E+12	4E+12	7.9E+11	3.4E+12	2E+10	5E+12	NA	NA
2005	0	1.6E+16	7.8E+14	1.1E+12	4.8E+13	3E+12	2E+13	3E+12	2.5E+12	5.6E+12	3E+12	NA	NA	NA
2006	0	7.5E+11	6.1E+12	8.4E+12	9.62E+11	4E+09	8E+11	3E+10	1.8E+10	3.9E+10	NA	NA	NA	NA
2007	0	1.6E+14	5.2E+13	4.7E+10	1.21E+14	1E+12	2E+13	4E+10	1.6E+10	NA	NA	NA	NA	NA
2008	0	5E+14	8.3E+14	3.7E+13	8.38E+14	1E+13	8E+12	9E+09	NA	NA	NA	NA	NA	NA
2009	0	1.3E+13	9.3E+11	1.2E+11	3.92E+11	8E+09	7E+10	NA	NA	NA	NA	NA	NA	NA
2010	0	5.4E+13	7.1E+11	2E+11	2.22E+10	1E+12	NA	NA	NA	NA	NA	NA	NA	NA
2011	0	7.2E+16	9.3E+15	2.4E+12	3.19E+13	NA	NA	NA	NA	NA	NA	NA	NA	NA
2012	0	8.1E+14	1.3E+13	3.2E+12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2013	0	7.4E+13	2.5E+12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2014	0	6.7E+13	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2015	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Packages Used:

lubridate, *fitdistrplus*, *ChainLadder*, *dplyr*, *tibble*, *reshape2*, *statmod*, *glm2*, *MASS*, *cplm*, *stringr*, *ggplot2*

2.Individual Contribution

Andy Mao: Data Reshape, Tweedie model, Chain-Ladder, Original Triangle, Final Report

Tana Wuren: Data Cleaning, Tweedie model, Chain-Ladder, Original Triangle, Final Report

Yuting Xiong: Model comparison, Poisson model, Gaussian model, Final Report

Xinyu Fei: Poisson and Gaussian model