NAIC Expense Data

Nicholas Porrone

2018-11-05

Contents

1 Introduction

1.1 Objective

This dataset is discussed in Exercise 3.5 of Edward Frees' text, *Regression Modeling with Actuarial and Financial Applications*. The data is given on the course website: http://fisher.stats.uwo.ca/faculty/aim/2018/3859A/data/NAICExpense.csv. The following report will analyize and discuss this data.

1.2 Data Summary

As referenced in the exercise, the data was apart of the database from the National Association of Insurance Commisioners. This database contained more than 3,000 other insurance companies. However, we are only looking at 384 of those companies with incured losses on file.

NOTE: AGENTWAGE contains 19 missing data points. These data points should be omitted and left out for future calculations.

	Description
EXPENSES	Total expenses incurred, in millions of dollars
LONGLOSS	Losses incurred for long tail lines, in millions of dollars
SHORTLOSS	Losses incurred for short tail lines, in millions of dollars
GPWPERSONAL	Gross premium written for personal lines, in millions of dollars
GPWCOMM	Gross premium written for commercial lines, in millions of dollars
ASSETS	Net admitted assets, in millions of dollars
CASH	Cash and invested assets, in millions of dollars
GROUP	Binary, Indicates whether the company is affiliated
STOCK	Binary, Indicates whether the company is a stock company
MUTUAL	Binary, Indicates whether the company is a mutual company
STAFFWAGE	Annual average wage of the insurer's administrative staff, in thousands of dollars
AGENTWAGE	Annual average wage of the insurance agent, in thousands of dollars

Table 1: Description of the variables used to explain the expenses of the insurance companies.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev
EXPENSES	-0.00	0.00	0.01	0.04	0.03	1.24	0.12
STAFFWAGE	51.73	80.06	84.36	87.26	93.82	137.48	11.93
AGENTWAGE	47.47	74.81	78.77	80.15	85.44	126.17	9.10
LONGLOSS	-0.07	0.00	0.00	0.03	0.01	0.85	0.09
SHORTLOSS	-0.00	0.00	0.00	0.04	0.02	1.17	0.12
GPWPERSONAL	-0.00	0.00	0.00	0.06	0.03	1.82	0.18
GPWCOMM	-0.00	0.00	0.03	0.13	0.09	4.19	0.32
ASSETS	0.00	0.01	0.06	0.36	0.19	8.71	1.03
CASH	0.00	0.01	0.05	0.33	0.18	8.82	0.98

Table 2: The values of the explantory variables.

We notice that all variables above have a right-skew because all of their medians are less than their means.

1.3 Transform the Data

We apply a log-transformation (ln(1 + x)) to the data to remove or reduce skewness.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LNEXPENSES	-0.00	0.00	0.01	0.04	0.03	0.81
LNSTAFFWAGE	3.97	4.40	4.45	4.47	4.55	4.93
LNAGENTWAGE	3.88	4.33	4.38	4.39	4.46	4.85
LNLONGLOSS	-0.07	0.00	0.00	0.02	0.01	0.62
LNSHORTLOSS	-0.00	0.00	0.00	0.03	0.02	0.78
LNGPWPERSONAL	-0.00	0.00	0.00	0.05	0.03	1.04
LNGPWCOMM	-0.00	0.00	0.03	0.10	0.09	1.65
LNASSETS	0.00	0.01	0.06	0.20	0.18	2.27
LNCASH	0.00	0.01	0.05	0.19	0.16	2.28

Table 3: The values of the transformed explantory variables.

Although we were not able to fully remove the right-skewness in the data , we were able to reduce it.

1.4 Correlations

Now we would like to take a look at the correlation between all variables within the transformed data set.

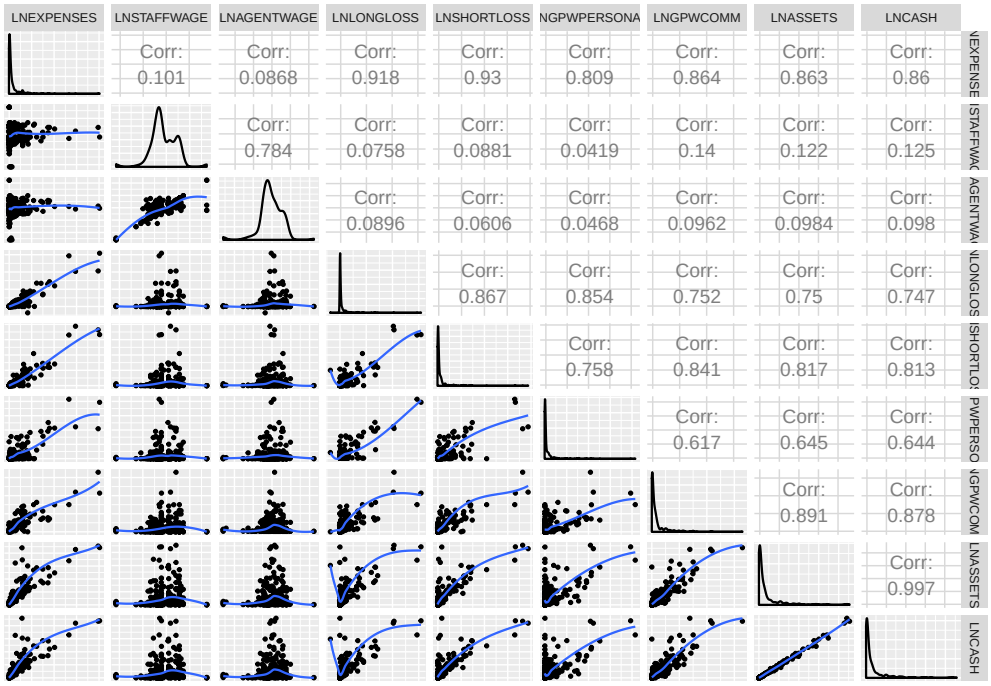
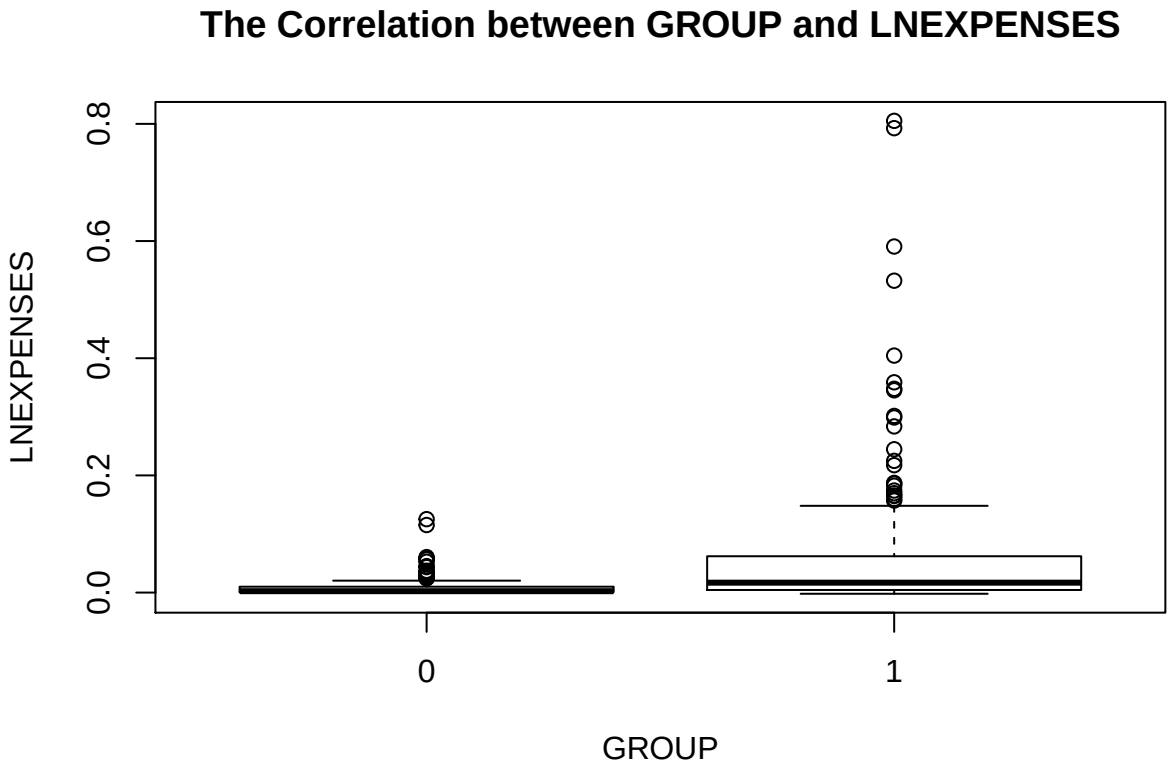


Figure 1: Correlations Scatterplot

From the scatterplot above we can see that aside from *LNAGENTWAGE* and *LNSTAFFWAGE* the variables are still heavily right-skewed. We may also take note that the 3 most highly correlated variables with *LNEXPENSES* are *LNLONGLOSS* , *LNSHORTLOSS* and *LNGPWCOMM* with correlation values .918, .93 , and .864 respectively.

1.5 Boxplot



From the box plot we can clearly see that *GROUP* 1 (The affilated companies) have a larger *LNEXPENSE*(the adjusted expenses in log dollars) than *GROUP* 0 (The unaffiliated companies). The median of the unaffilated companies expense in dollars is $2382.83 while the median of the affiliated companies is $17033.44.

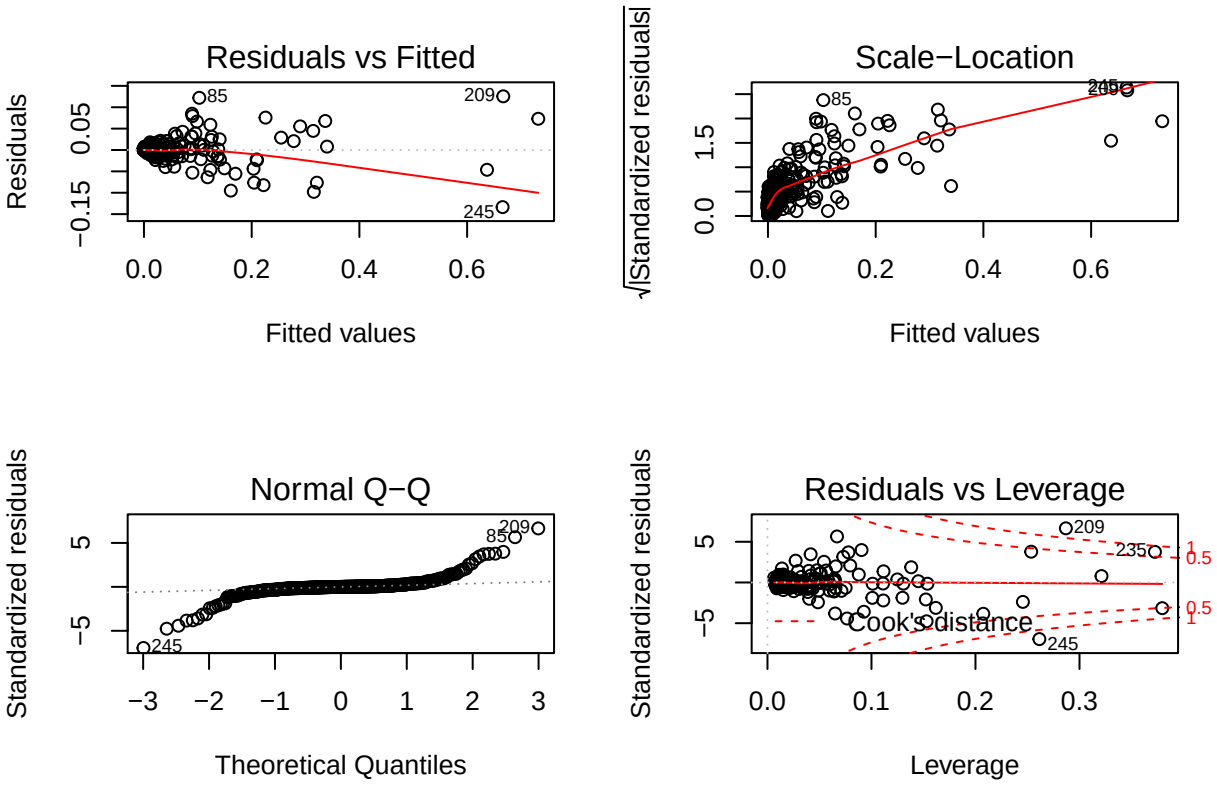
2 Linear Models and Regression Analysis

2.1 Model 1, LNEXPENSES On All Explanatory Variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0073	0.0458	-0.16	0.8741
GROUP	-0.0002	0.0028	-0.07	0.9472
MUTUAL	-0.0021	0.0043	-0.48	0.6301
STOCK	-0.0033	0.0038	-0.86	0.3910
LNSTAFFWAGE	0.0006	0.0144	0.04	0.9670
LNAGENTWAGE	0.0017	0.0166	0.10	0.9194
LNLONGLOSS	0.4576	0.0454	10.07	0.0000
LNSHORTLOSS	0.3091	0.0326	9.48	0.0000
LNGPWPERSONAL	0.0618	0.0193	3.19	0.0015
LNGPWCOMM	0.0799	0.0169	4.74	0.0000
LNASSETS	-0.0517	0.0475	-1.09	0.2772
LNCASH	0.0957	0.0469	2.04	0.0418

Table 4: A linear model fit by LNEXPENSES on all of the explanatory variables.

By this table we can see that the two premiums, two loss and asset varibles are highly significant to this model as we can be over 99% sure they are non-zero. Also note that if we were to increase each variable by 1 unit then *LNLONGLOSS* would result in the greatest increase in *LNEXPENSES*.



Through the plots above we can see that as the Fitted values increase we can see the data tends to vary more. The data "fans out".

St. Err	R^2	Adj. R^2
0.022	0.942	0.940

Table 5: Standard Error of Residuals, R^2, and Adjusted R^2 values

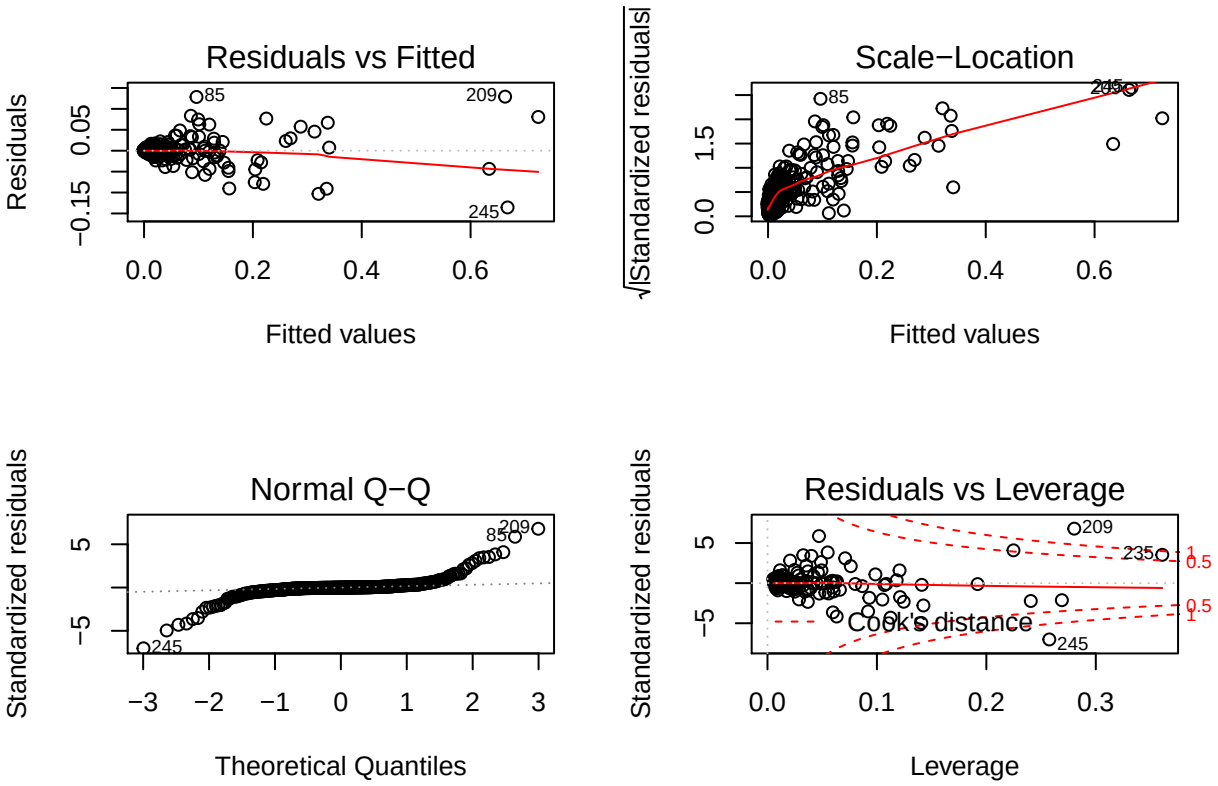
This table shows us that 94% of the change in y relies on the change x. One thing to note is, the standard deviation is comparable to the othersubsequent models. We can also note that this model seems to have a good fit for the regression of the variables.

2.2 Model 2, Drop CASH, MUTUAL, and STOCK

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0124	0.0457	-0.27	0.7869
GROUP	-0.0014	0.0026	-0.55	0.5818
LNSTAFFWAGE	0.0027	0.0143	0.19	0.8480
LNAGENTWAGE	0.0002	0.0166	0.01	0.9907
LNLONGLOSS	0.4582	0.0450	10.18	0.0000
LNSHORTLOSS	0.3132	0.0325	9.63	0.0000
LNGPWPERSONAL	0.0632	0.0191	3.30	0.0011
LNGPWCOMM	0.0689	0.0161	4.29	0.0000
LNASSETS	0.0442	0.0074	6.00	0.0000

Table 6: A linear model fit by LNEXPENSES on all of the explanatory variables except CASH,MUTUAL, and STOCK.

Similar to model 1, we see that the two premiums, two loss and asset varibles are highly significant to this model as we can be over 99% sure they are non-zero. We can also see that *GROUP* now has a bigger effect on the change of *LNEXPENSES* when we leave out *CASH* , *MUTUAL* AND *STOCK*. We can also note that for every unit increase of *LNGPWCOMM* , *LNEXPENSES* will increase by 0.0689. Given an inital expense and value of *GDPCOMM* we can describe the new dollar amount of expenses using the formula $\left[e^{\beta \ln\left(\frac{1+GPWCOMM+0.000001}{1+GPWCOMM}\right)}\right] - 1$ .If we were to suppose that *GPWCOMM* has increased by \$1 than *EXPENSES* would increase by \$0.0677. (We take the median of GPW and plug it in(0.02533)).



Just as we seen in Model 1, the plots show that as the fitted values increase we can see the data tends to vary more. The data "fans out".

St. Err	R^2	Adj. R^2
0.022	0.941	0.940

Table 7: Standard Error of Residuals, R^2, and Adjusted R^2 values

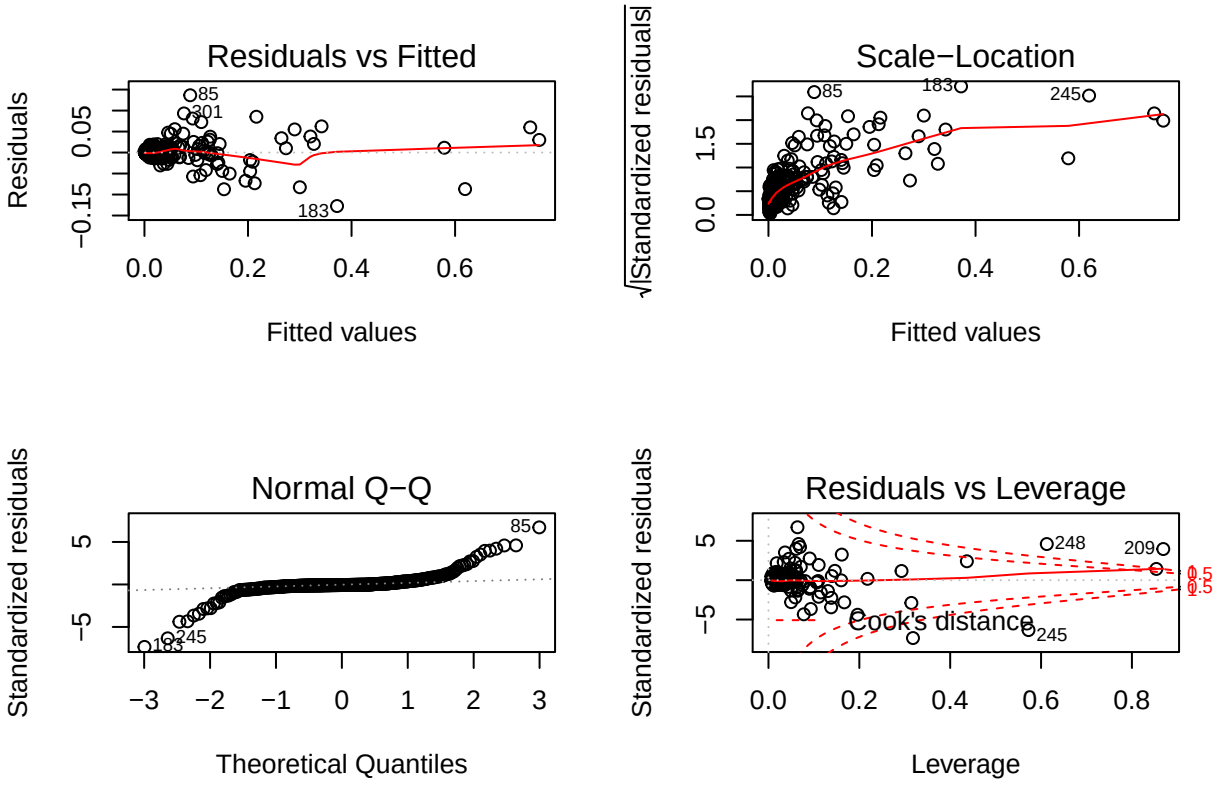
Similarly to Model 1,This table shows us that 94% of the change in y relies on the change x. One thing to note is, the standard deviation is comparable to the othersubsequent models. We can also note that this model seems to be slighly less effective than Model 1 as the R^2 value is slightly less.

2.3 Model 3, Square The Two Loss and The Two Gross Premium Variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0170	0.0429	-0.40	0.6912
GROUP	0.0006	0.0025	0.25	0.8058
LNSTAFFWAGE	0.0013	0.0134	0.10	0.9219
LNAGENTWAGE	0.0029	0.0156	0.19	0.8512
LNLONGLOSS	0.4043	0.0639	6.33	0.0000
LNSHORTLOSS	0.3912	0.0485	8.07	0.0000
LNGPWPERSONAL	0.1259	0.0362	3.47	0.0006
LNGPWCOMM	-0.0299	0.0225	-1.33	0.1842
LNASSETS	0.0512	0.0072	7.07	0.0000
LONGLOSSsq	0.5301	0.2523	2.10	0.0364
SHORTLOSSsq	-0.3690	0.0870	-4.24	0.0000
GPWPERSONALsq	-0.1518	0.0855	-1.77	0.0768
GPWCOMMsq	0.1191	0.0207	5.74	0.0000

Table 8: A linear model fit by LNEXPENSES on all of the explanatory variables and the two loss/gross variables squared.

From this table we can see that the new quadratic variables appear to be similar to the explanatory variables. This is because they have very low probabilty to be larger than the t value. More will be said in the table below which displays R values and std deviation.



St. Err	R^2	Adj. R^2
0.021	0.949	0.947

Table 9: Standard Error of Residuals, R^2, and Adjusted R^2 values

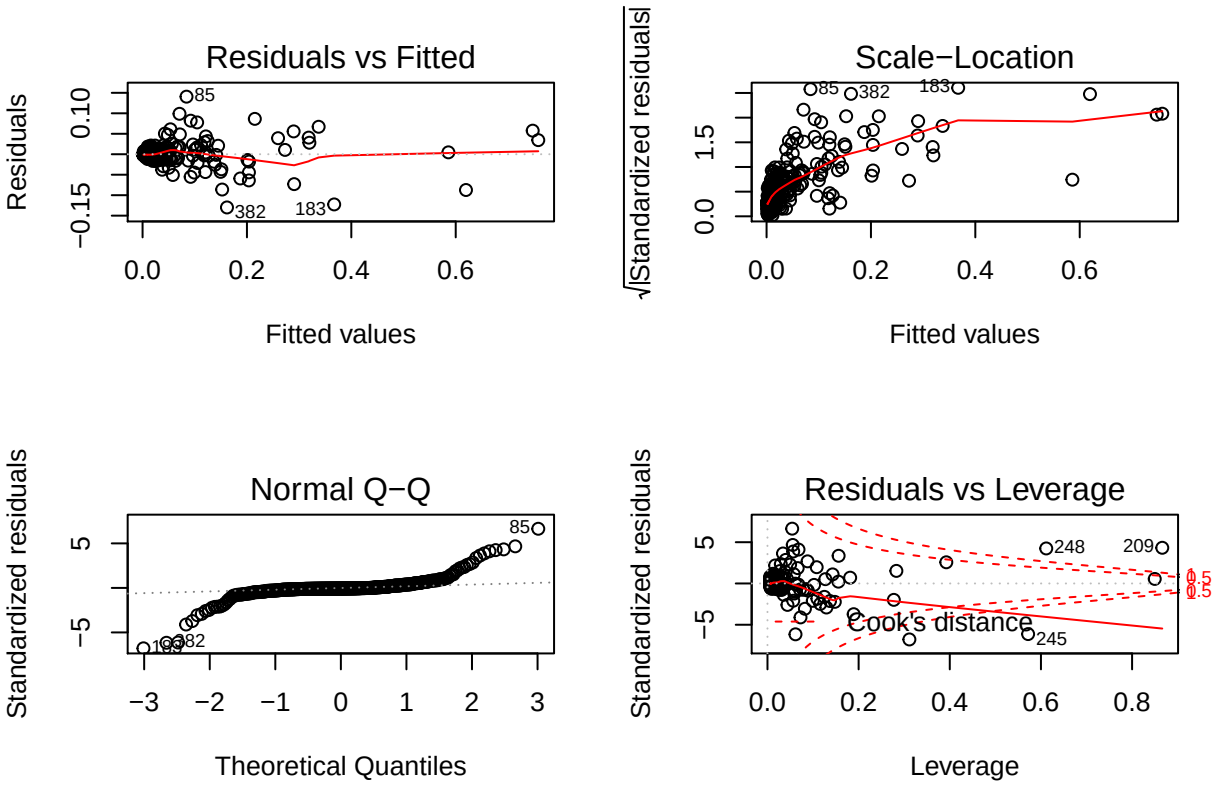
These R^2 values are the highest of all models and the standard deviation is the lowest of all models. Since the ajusted didnt decrease we can infer that these quadratic variables are not completely useless. Infact, they have made this the most accuarte model.

2.4 Model 4, Omit The Two BLS Variables From Model 3

It is important to mention that this model has 384 observations opposed to the 3 previous models of size 365. The model has 19 more observations due to the elimination of the BLS variables. This is because the BLS variables held 19 null observations.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0017	0.0018	0.94	0.3479
GROUP	0.0000	0.0025	0.01	0.9933
LNLONGLOSS	0.3721	0.0652	5.71	0.0000
LNSHORTLOSS	0.3755	0.0498	7.54	0.0000
LNGPWPERSONAL	0.1631	0.0368	4.43	0.0000
LNGPWCOMM	-0.0199	0.0222	-0.90	0.3699
LNASSETS	0.0446	0.0068	6.58	0.0000
I(LNLONGLOSS^2)	0.6571	0.2595	2.53	0.0117
I(LNSHORTLOSS^2)	-0.3307	0.0897	-3.69	0.0003
I(LNGPWPERSONAL^2)	-0.2072	0.0877	-2.36	0.0187
I(LNGPWCOMM^2)	0.1093	0.0213	5.14	0.0000

Table 10: A linear model fit by LNEXPENSES on all of the explanatory variables, the two loss/gross variables squared and drop the two BLS variables.



St. Err	R^2	Adj. R^2
0.022	0.942	0.941

Table 11: Standard Error of Residuals, R^2, and Adjusted R^2 values

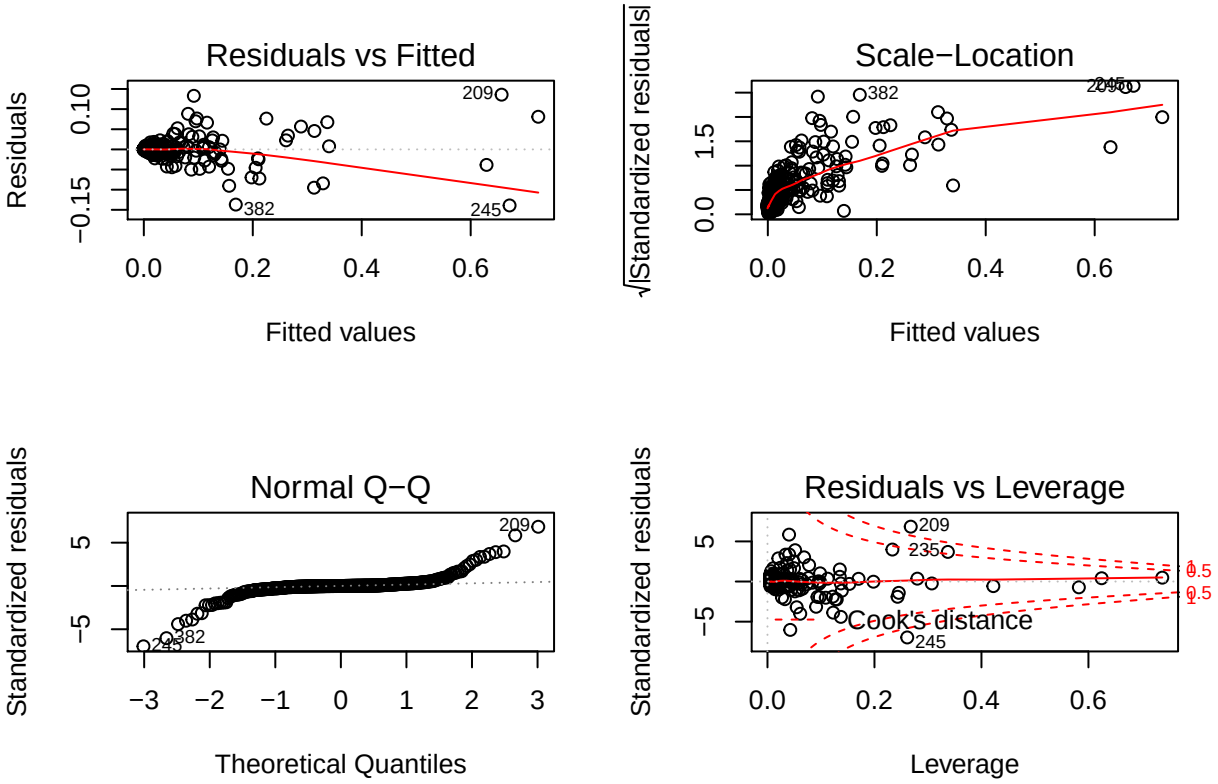
The R^2 values are the 2nd highest and the standard deviation is tied with the Model 3 for the lowest of all models. These numbers are especially impressive with the largest amount of observations. We can see eliminating the 2 BLS variables has a postive effect on the regression.

2.5 Model 5, Drop The Quadratic Terms From Model 3 and Add Interaction Terms With the GROUP and ASSET Variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0003	0.0023	0.15	0.8818
GROUP	-0.0011	0.0030	-0.35	0.7267
LNLONGLOSS	0.4257	0.2428	1.75	0.0804
LNSHORTLOSS	0.4580	0.3751	1.22	0.2229
LNGPWPERSONAL	0.0624	0.1884	0.33	0.7405
LNGPWCOMM	0.0956	0.1399	0.68	0.4949
LNASSETS	0.0289	0.0503	0.57	0.5659
LONGLOSSint	0.0292	0.2480	0.12	0.9062
SHORTLOSSint	-0.1428	0.3768	-0.38	0.7049
GPWPERSONALint	0.0140	0.1895	0.07	0.9412
GPWCOMMint	-0.0272	0.1408	-0.19	0.8469
ASSETSint	0.0099	0.0508	0.19	0.8463

Table 12: A linear model fit by LNEXPENSES on all of the explanatory variables except the two WAGES variables and include Interaction terms with the GROUP and ASSET variables

Similarly to Model 2, an inital expense and value of *GDPCOMM* we can describe the new dollar amount of expenses using the formula $\left[e^{\beta_1 \ln\left(\frac{1+GPWCOMM+0.000001}{1+GPWCOMM}\right)-\beta_2 \ln\left(\frac{1+GPWCOMM+0.000001}{1+GPWCOMM}\right)}\right] - 1$ then be sure to multiple it by 1000000. β_1 represents the GPWCOMM coefficent meanwhile β_2 represents the interaction term. If we suppose that GPWCOMM increases by \$1.00, we expect EXPENSES to increase by \$0.0932 for GROUP = 0(unafilliaed) companies. If we suppose that GPWCOMM increases by \$1.00, we expect EXPENSES to increase by \$0.0666 for GROUP = 1(affiliated) companies.



St. Err	R^2	Adj. R^2
0.023	0.935	0.933

Table 13: Standard Error of Residuals, R^2, and Adjusted R^2 values

The R^2 values are the lowest and the standard deviation is the highest of across models. Based on these values, this model is the least accurate of the 5.