Hands on Assignment #3

1) Use worksheet "baseball stats" to perform a multiple regression analysis on the dataset found in BB2011 tab, using Wins as the dependent variable, and League, ERA, Runs Scored, Hits Allowed, Walks Allowed, Saves, and Errors as candidates for the independent variables. Perform the analysis at the 5% significance level.

a) Create a full write up, where you write your statistical analysis step by step. In your write up, make sure you address the following points.

- Methodology and steps that you took to get to your final regression equation.

Answer: To Predict the number of wins (dependent variable), we used League, ERA, Runs Scored, Hits Allowed, Walks Allowed, Saves, and Errors as independent variable.

I. We performed multiple regression with all independent variable and found the below results from the regression output:

Independent Variable	Sign of co-efficient of independent variables	Interpretation
League	Negative	Chances of winning will be less if number of leagues will increase.
E.R.A.	Negative	Number of wins will be high if E.R.A. score is low.
Runs Scored	Positive	High run scored will contribute to increase in number of wins
Hits Allowed	Negative	Increase in number of hits can decrease the number of wins
Walks Allowed	Negative	Increase in number of walks allowed will decrease the number of wins
Saves	Positive	Increase in saving point will increase the number of wins
Errors	Negative	More error can cause less chance of wins.

II. The purpose of the regression is to predict the number of wins of baseball team. A positive co-efficient will increase the number of wins while negative co-efficient will decrease the number of wins.

III. To identify the independent variable which support to predict the number of wins, we checked the p-value of all predictors. We excluded the predictors whose p-value is greater than 0.05. Below is the table for all predictors with p-value:

Independent Variable	P-value
League	0.40675
E.R.A.	0.03774

Runs Scored	0.00000
Hits Allowed	0.30121
Walks Allowed	0.09817
Saves	0.00706
Errors	0.45996

Highlighted values show p-value >0.05
So, League, Hits Allowed, Walks Allowed, errors have insignificant p-value which exceeded the level of significance 0.05. Hence, excluded from the multiple regression analysis.

IV. We continued to perform multiple regression analysis with predictors E.R.A., Runs Scored and Saves.

- Final regression equation output.
  Answer: Below is the final regression output after removing predictors which did not fit in the model:

Regression Statistics	
Multiple R	0.944566583
R Square	0.892206029
Adjusted R Square	0.879768263
Standard Error	3.95819414
Observations	30

ANOVA

	df	***SS***	***MS***	***F***	***Significance F***
Regression	3	3371.617	1123.872	71.734	0.000
Residual	26	407.350	15.667		
Total	29	3778.967			

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***	***Lower 95.0%***	***Upper 95.0%***
Intercept	62.610	12.755	4.908	0.000	36.391	88.829	36.391	88.829
E.R.A.	-15.426	2.159	-7.146	0.000	-19.864	-10.989	-19.864	-10.989
Runs Scored	0.094	0.009	10.288	0.000	0.075	0.113	0.075	0.113
Saves	0.338	0.128	2.637	0.014	0.074	0.601	0.074	0.601

- What is the final regression equation?
  Answer: Final regression equation: **62.610 -15.426(E.R.A.) + 0.094(Runs Scored) + 0.338(Saves)**

- Interpret all the coefficients in the equation.

Independent Variable	Sign of co-efficient of independent variables	Interpretation
E.R.A.	Negative	Number of wins will be high if E.R.A. score is low. Each extra increase in E.R.A. will decrease the number of wins by 15.426. However, co-efficient can be rounded off as 15 as 15.426 has not logical meaning of wins.
Runs Scored	Positive	High run scored will contribute to increase in number of wins. Each unit of increase in runs scored will increase the number of wins by 0.094.
Saves	Positive	Increase in saving point will increase the number of wins. Each unit of increase in Saves will increase the number of wins by 0.338.

- Speak to whether the signs on the coefficients make sense.

  Answer: Yes, sign of the co-efficient make sense. In real world, ERA score should be low in baseball team. In our regression analysis, it has negative sign which has the logical meaning for predicting number of wins. If team has high run scored and saves, their chance of winning is high. In our analysis, Runs Scored and Saves have positive sign which also have logical significance to predict the number of wins.

- Interpret R squared.

  Answer: The co-efficient of determination R^2 is 0.8922 which is high. The chosen independent variables help to predict the number of wins. Also, total variation explained by the model is 89.22% while 10.78% is unexplained. The model is good and can be used to forecast the number of wins.

- Include a full residual analysis.

  Table1: Residual Analysis

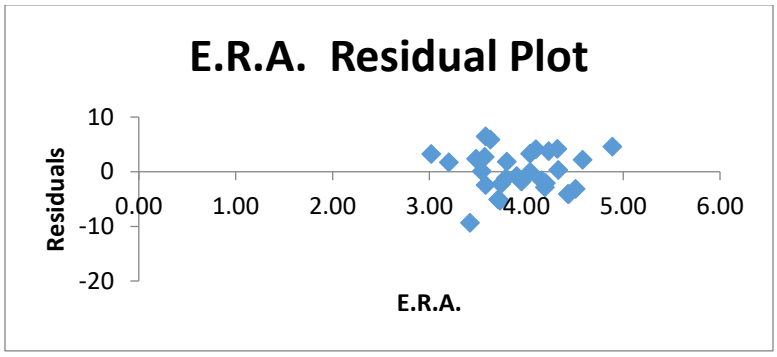
RESIDUAL OUTPUT			
Observation	***Predicted Wins***	***Residuals***	***Standard Residuals***
1	64.4241	4.5759	1.2209
2	92.0902	-2.0902	-0.5577
3	74.9227	4.0773	1.0879
4	76.2574	3.7426	0.9986

5	91.7073	3.2927	0.8786
6	75.1198	-4.1198	-1.0992
7	83.3047	2.6953	0.7192
8	60.8549	2.1451	0.5723
9	102.3067	-5.3067	-1.4159
10	79.0807	-5.0807	-1.3556
11	67.7984	-0.7984	-0.2130
12	84.5386	6.4614	1.7240
13	97.2141	-1.2141	-0.3239
14	76.8392	4.1608	1.1102
15	92.1822	1.8178	0.4850
16	86.6460	2.3540	0.6281
17	70.6989	0.3011	0.0803
18	80.5841	-1.5841	-0.4227
19	77.0948	-4.0948	-1.0926
20	59.1941	-3.1941	-0.8523
21	81.9472	0.0528	0.0141
22	73.8396	-1.8396	-0.4908
23	90.1494	5.8506	1.5610
24	79.8777	-2.8777	-0.7678
25	98.8087	3.1913	0.8515
26	72.0574	-0.0574	-0.0153
27	92.2998	-2.2998	-0.6136
28	80.3643	-9.3643	-2.4986
29	84.3030	1.6970	0.4528
30	82.4945	-2.4945	-0.6656

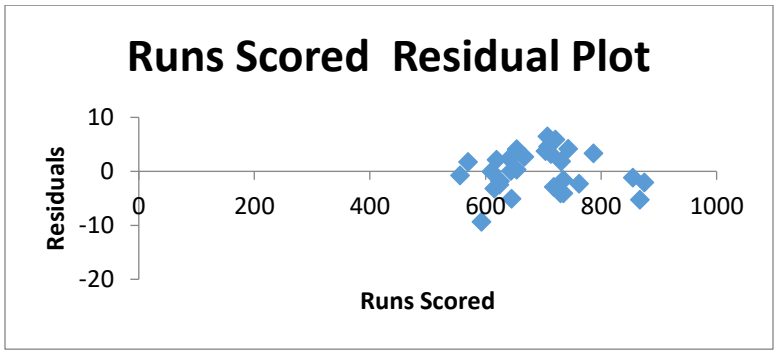
Residual plot graphs

For regression, four criteria should be met: linearity, independence of errors, normality of error and equal variance. To check the non-linearity, outliers, unequal variance and error, we generate the residual plot graphs for each predictor:

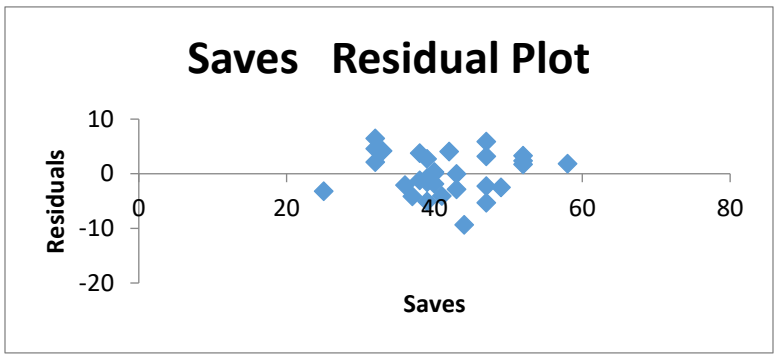
- Residual plot for E.R.A.



- Residual Plot for Runs Scored



- Residual Plot for Saves

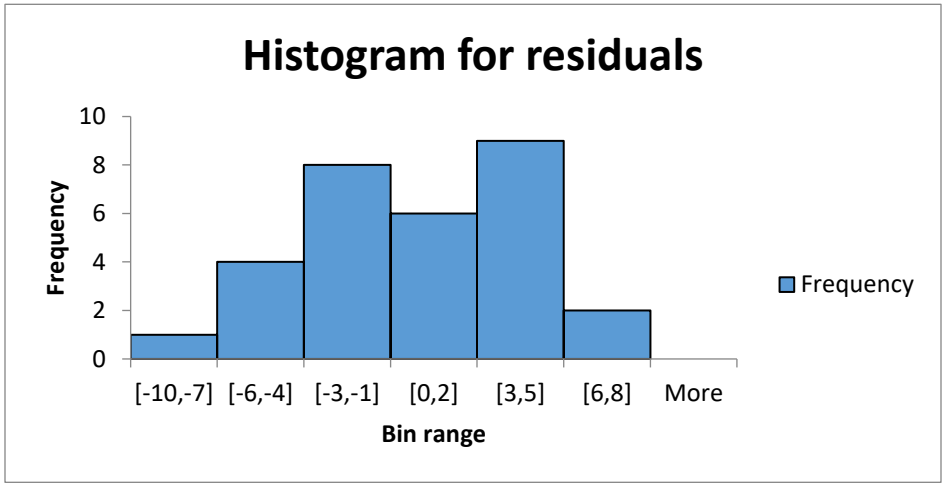


Above residual plot graphs show no pattern of distribution and no heteroscedasticity (unequal variance). We can conclude that model is great fit.

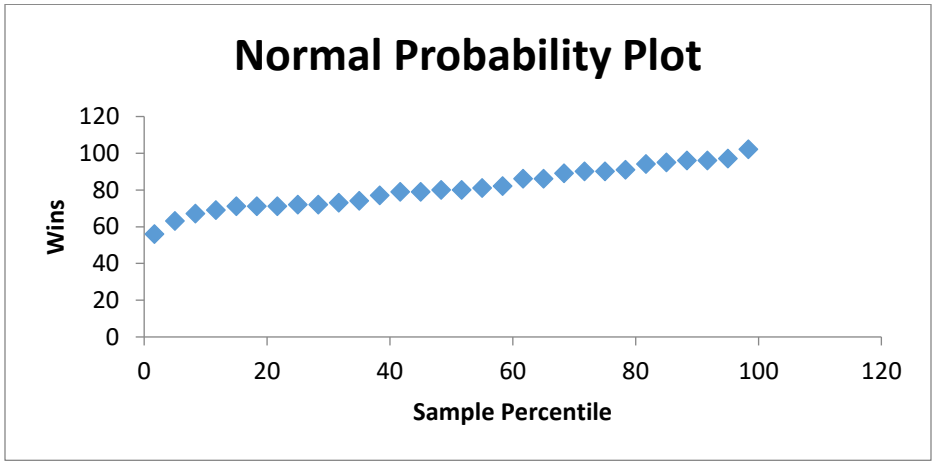
HISTOGRAM AND NORMAL PROBABILITY PLOT:

Histogram and normal probability plot were drawn to check the residuals and errors are normally distributed. Histogram shows that residuals were normally distributed, no skewness is observed. Normal Probability plot shows that data is normally distributed. Hence, the model is best fit.

Histogram for Residual:



Normal probability plot:



From the residual analysis, we conclude that model is acceptable and can be used to forecast the wins.

- What is the residual of the Tampa Bay observation?
  Answer: The residual of the Tampa Bay observation is 6.46. (Highlighted in yellow color in Table1).

b) Now, use tab BB2012 to make predictions of wins in 2012, using the model you created with the 2011 stats.

Table2

Team	WINS	E.R.A.	Runs Scored	Hits Allowed	Walks Allowed	Saves	Errors	Wins predicted
HOU	55	4.56	583	1493	540	31	118	58
SFG	94	3.68	718	1361	489	53	112	91
WSN	98	3.33	731	1296	497	51	96	97

- How many games are the Giants (SFG) expected to win in 2012?
  Answer: From table 2: For SFG, E.R.A.=3.68, Runs Scored=718, Saves=53
  Using final regression equation: 62.610 -15.426(E.R.A.) + 0.094(Runs Scored) + 0.338(Saves);
  Number of games Giants (SFG) expected to win=91

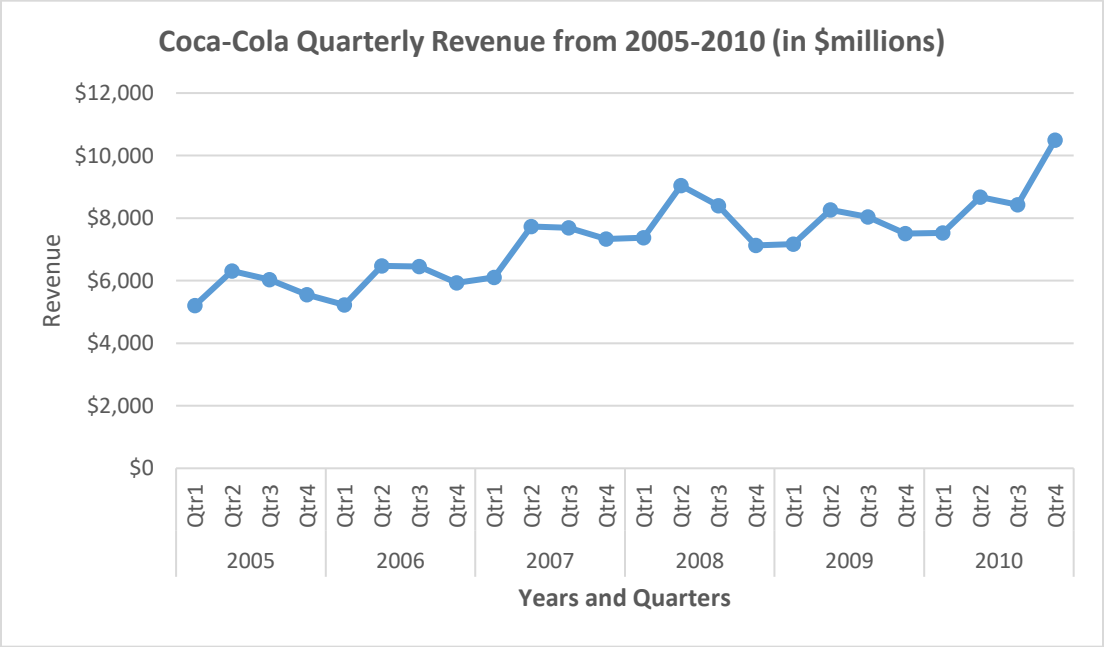
- Which team is predicted by the model to have the worst record in 2012?
  Answer: Team HOU have the worst record in 2012 with number of wins 58 (Table2).

- Which team is predicted by the model to have the best record in 2012?
  Answer: Team WSN have the best record in 2012 with the number of wins 97 (Table2).

2) Use the CocaCola data set to analyze quarterly data on Coca Cola's revenues.

   a. Plot the time series as a line graph. Make sure the graph is polished enough to be included in a formal report - i.e. it has a good title, axis titles/formatting, etc. (extra credit: format the x axis to have years and quarters)

   Answer: Below is the graph for Coca Cola's revenue for years 2005 to 2010



   b. Perform a regression using seasonal binaries. **Use 0.1 level of significance** for eliminating p-values. Include the final output and write the final regression equation.
   Answer: Below is the regression output at 0.1 level of significance:

Regression Statistics	
Multiple R	0.8926
R Square	0.7968
Adjusted R Square	0.7663
Standard Error	625.866
Observations	24

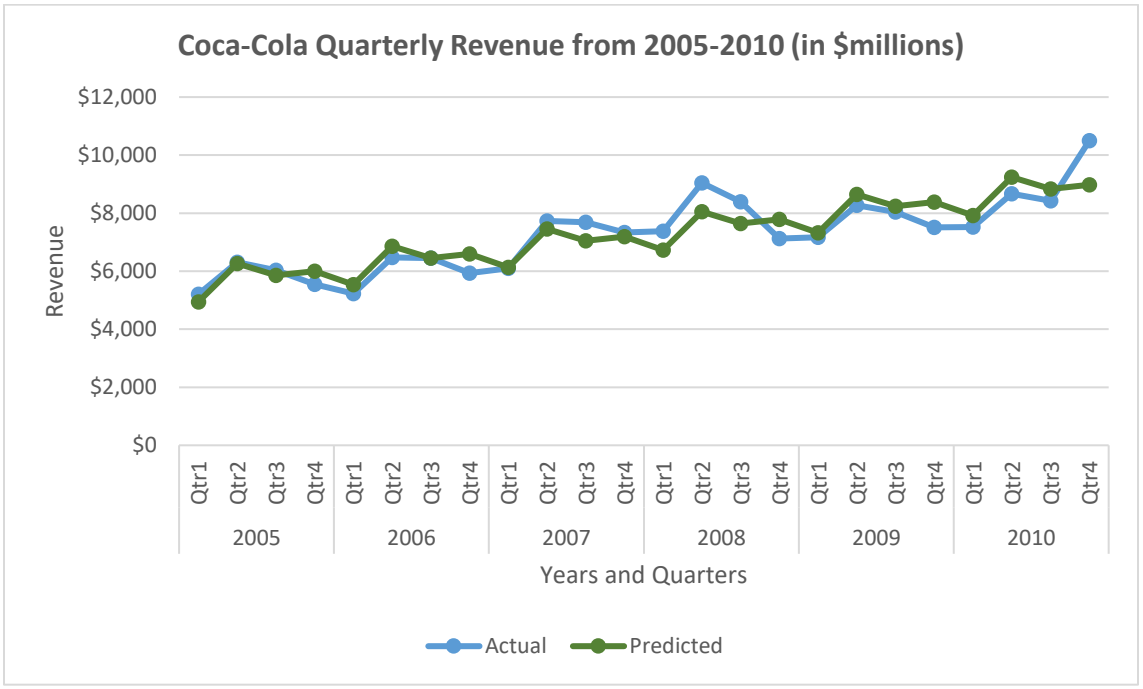
ANOVA	df	SS	MS	F	Significance F
Regression	3	30713251.20	10237750.40	26.1362	4.01613E-07
Residual	20	7834153.30	391707.66		
Total	23	38547404.50			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 90.0%	Upper 90.0%
Intercept	5405.86	310.181	17.428	1.5E-13	4758.836	6052.887	4870.888	5940.836
Time index	148.874	18.676	7.971	1.2E-07	109.916	187.833	116.663	181.086
Q1	-608.81	316.397	-1.924	0.069	-1268.806	51.178	-1154.509	-63.119
Q2	558.645	314.184	1.778	0.091	-96.732	1214.022	16.766	1100.524

Final Regression equation:
Quarterly Revenue=**5405.86 + 148.874(Time Index) - 608.814 (Q1) + 558.645 (Q2)**

c. Plot the fitted values of your time series on the same graph as the actuals. Does your regression look good?

Answer: Below is the graph for actual and predicted revenues for Coca-Cola company.



Yes, Regression looks good. The line graph of fitted value is very close to the actual revenue in each year and quarter. The regression output shows R^2 is 0.7968 which is high and shows 79.68% of total variation in revenue is explained by model while 20.32% is unexplained. Hence, model is acceptable and can be used for forecasting the revenue of Coca-Cola company.

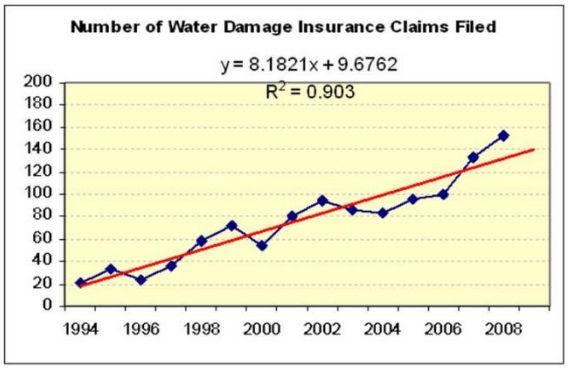
d. Interpret the results. Make sure to talk about seasonality.
Answer: From the graph and regression model, it suggests that in quarter2 and quarter4, people consume more cold drinks than other quarters. So, these quarters seem to have high level of Coca-Cola revenue. Seasonality is a repetitive cyclical pattern within a year which occurs in quarter2 and quarter3. Final regression equation: Quarterly Revenue=**5405.86 + 148.874(Time Index) - 608.814 (Q1) + 558.645 (Q2)**
Since co-efficient of Quarter 1 is -608.814, we can conclude that in quarter1, Coca-Cola's revenue will decrease by $608.814 million. Similarly, for Q2, the co-efficient is 558.645 which illustrate the revenue will increase by $558.645 million in quarter 2.

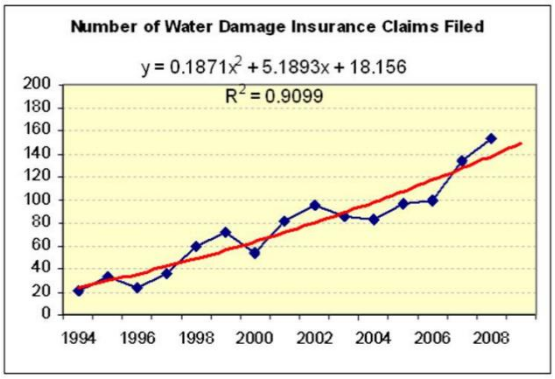
e. Use the regression equation to make a prediction for each quarter in 2011.
Using Regression question **5405.86 + 148.874(Time Index) - 608.814 (Q1) + 558.645 (Q2)**
For 2011, following are the quarterly revenue in 2011:

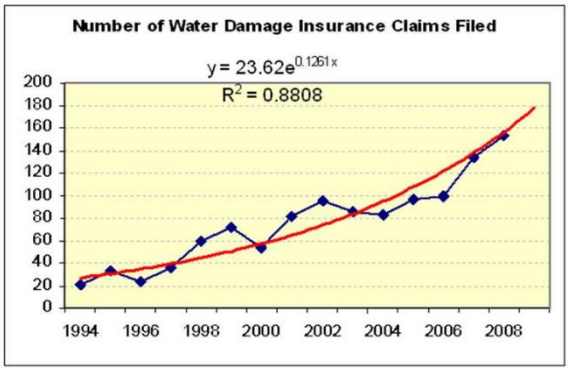
Q1 = $8518.896
Q2 = $9835.229
Q3 = $9425.458
Q4 = $9574.332

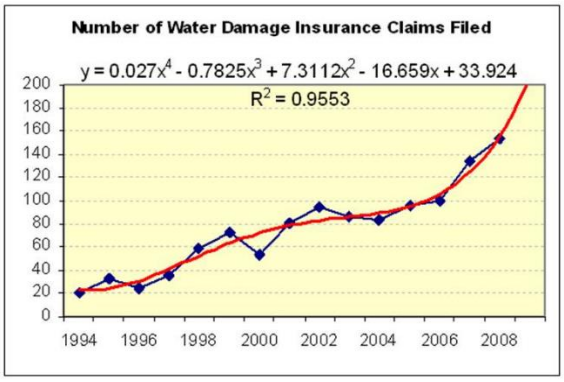
3) Bob analyzed water damage claims filed at a small Louisiana home insurance company over the last 15 years. He fitted several different trend models, shown below. Which trend model seems most reasonable (or more than one) for making forecasts for the next three years? What about the principle of Occam's Razor?

Answer: The first model which is linear model seems most reasonable as it has high R^2 with value 0.903. Occam's razor principle suggests choosing the simplest model. In this case, the linear model is simplest among all trend model and familiar to everyone.


Number of Water Damage Insurance Claims Filed
y = 8.1821x + 9.6762
R² = 0.903
200
180
160
140
120
100
80
60
40
20
0
1994
1996
1998
2000
2002
2004
2006
2008


Number of Water Damage Insurance Claims Filed
y = 0.1871x² + 5.1893x + 18.156
R² = 0.9099
200
180
160
140
120
100
80
60
40
20
0
1994
1996
1998
2000
2002
2004
2006
2008


Number of Water Damage Insurance Claims Filed
y = 23.62e^0.1261x
R² = 0.8808
200
180
160
140
120
100
80
60
40
20
0
1994
1996
1998
2000
2002
2004
2006
2008


Number of Water Damage Insurance Claims Filed
y = 0.027x⁴ - 0.7825x³ + 7.3112x² - 16.659x + 33.924
R² = 0.9553
200
180
160
140
120
100
80
60
40
20
0
1994
1996
1998
2000
2002
2004
2006
2008