STAT 331 Final Project, Fall 2016

Daniel Matheson, 20270871

Summary

The objective of this report is to analyze strike activity in 18 countries belonging to the Organization for Economic Co-operation and Development in the years 1951-1985 to determine which variables are significant in predicting strike activity[1]*: the countries themselves, the year, unemployment rates, inflation rates, democracy index*[2]*, union centralization*[3]*, union density*[4]*.*

We created an algorithmic model, and a qualitative model based on macroeconomic intuition. The latter was the most accurate and it showed that the number of strike days is predicted by the country, the year, inflation, and the combination of union density working with union centralization (i.e as a product)

[1] Defined as the number of days in the year lost per 1,000 workers due to strikes

[2] *The Democracy Index* is defined as the proportion of left-party parliamentary representation

[3] The measure of *Union Centralization* refers to "the authority that union confederations have over their members". The higher this value, the more powerful the union. This measure is aggregated over all years in a given country. More information here: https://lanekenworthy.files.wordpress.com/2014/07/2003ijs.pdf

[4] *Trade Union Density* is the fraction of wage earners in the country who belong to a union

1. Model Selection

We wish to pick *two* linear regression models which we believe represent the data well, in order to compare them in the next section.
In order to accomplish this, we will use *Automated Model Selection*, using all three versions of this process (Forward, Backward, and Stepwise) to select one model, and qualitative analysis to determine the second.
Before we begin, it is useful to look at the plots of paired (non-categorical) data:

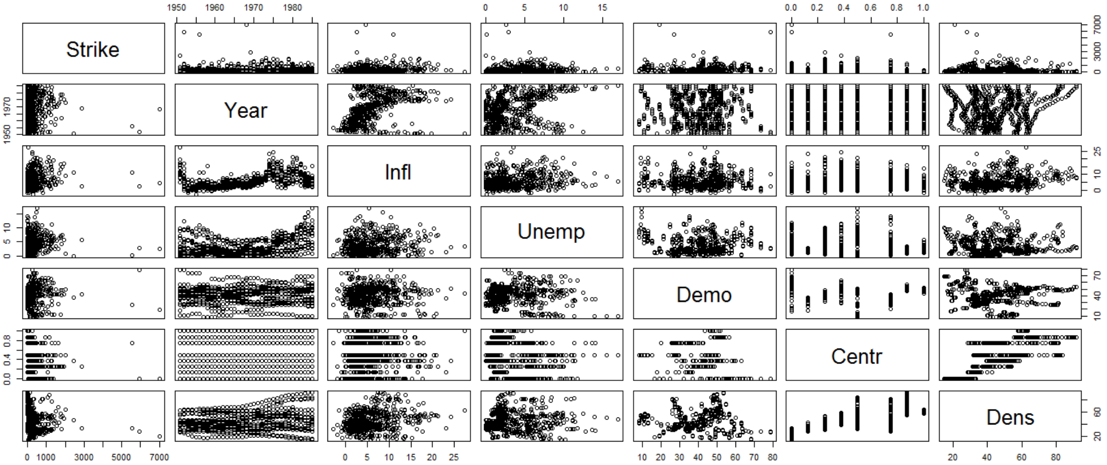


Figure 1: Paired Data

From these plots it is already clear that there may be some outliers (seen in the leftmost plots). Due to these extremely large values of strike days, it is hard to gauge visually if the other co-variates have any effect.
As explained in Appendix A, the data was found to have some abnormalities. The standardized (and studentized) residuals were found to be consistently negative or near zero for all residuals corresponding to the observations with less than the median number of strikes. In addition, the residuals were found to have non-constant variance, and did not follow a normal distribution.
We propose a transformation of the data to account for these problems; specifically, a transformation on the number of strike days. The transformation we decided on is:

$$\text{Strikes} \longleftarrow \log(\text{Strikes} + 1)$$

where “Strikes” is the number of strike days in each observation.
We now have a different, and more clear picture of the interactions between the variables in our pairs plot, and we can now spot some clear trends.
Additionally, the points that represented the extremely large number of strike days have been brought in closer to the rest of the points.

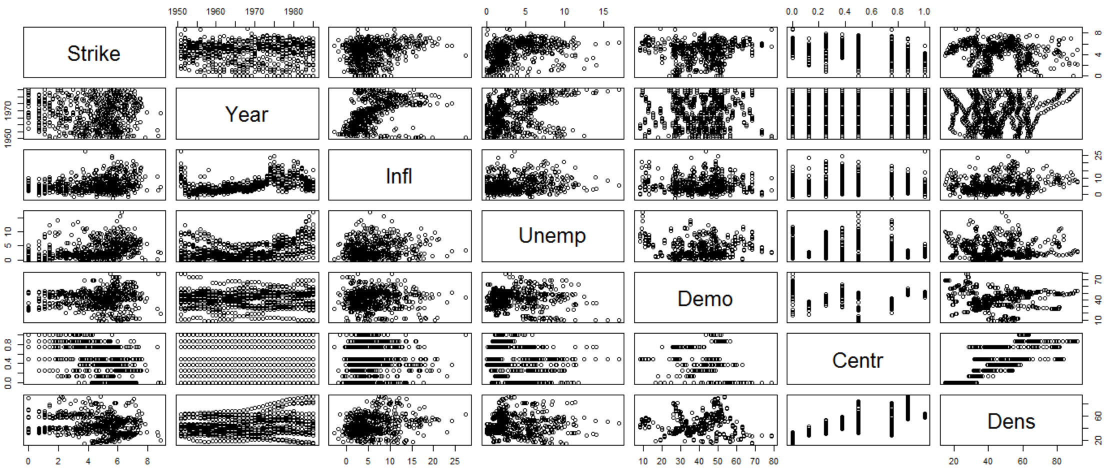


Figure 2: Paired Data after taking log of Strike

We will also perform another transformation on the data, to check if there is a relationship between neighbouring countries. That is, we will create a new co-variate **Region** which can take on values: *"North America", "Europe", "Scandanavia", "Australia and New Zealand", and "Japan".* The reason for separating Scandanavia (Denmark, Finland, Netherlands, Norway, and Sweden) is that these countries appeared to have different trends than the others. See Appendix B for details on the Scanadavian separation.

1.1 Quantitative Model Selection

We now choose our models, starting with the "quantitative model":
We use automated model selection in the same way done in Appendix A for the pre-transformation data, for the now-transformed data.[5]
Once we got our Forward, Backward and Stepwise models; we saw that the Forward model had far too many coefficients and would most likely over-fit the data. To choose between the remaining two models, we took the co-variates which were common to both, and made a new "Test" model with those co-variates. Using ANOVA tests, with some intuitive reasoning, we decided to select the Backward model.
Therefore, our first candidate model is the Backward Selection Model:

$$\begin{aligned}\text{Strike} \sim\ &\text{Country} + \text{Year} + \text{Inflation} + \text{Union Density} + \text{Unemployment}\\ &+ \text{Country:Year} + \text{Country:Unemployment} + \text{Year:(Union Density)}\\ &+ \text{Unemployment:(Union Density)}\end{aligned}$$

Which fits the data fairly well, with relatively Normal and constant-variance residuals, as can be seen in Figure 3 (top of next page). Note: the other residuals (press and studentized) had very similar histograms.
See Appendix C for details of the narrowing down of automated models.

[5] i.e we added Region to the maximal and stepwise starting models, and Strike is now log(Strike)

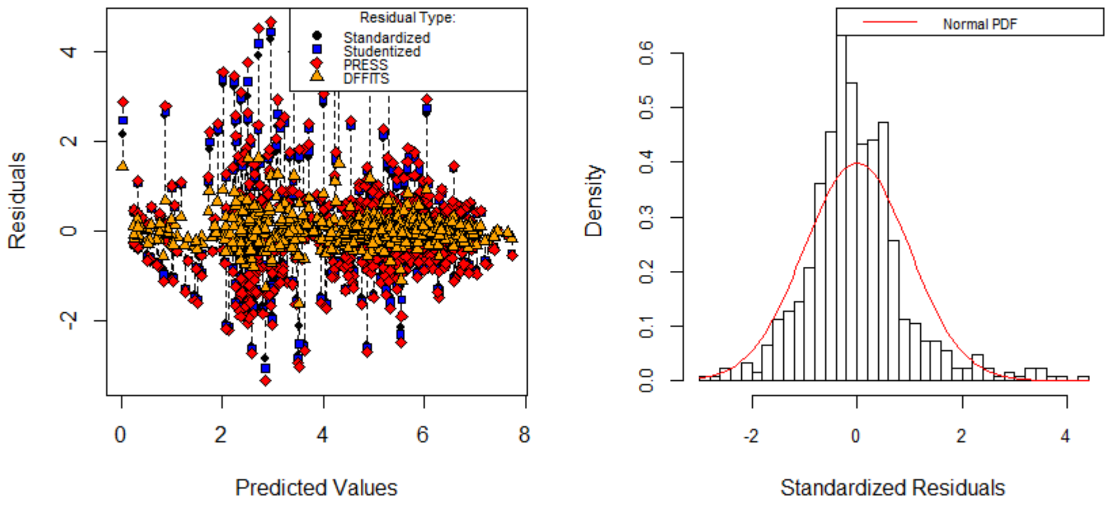


Figure 3: Backward Model Residuals

1.2 Qualitative Model Selection

In order to construct a qualitative model, we first and foremost consulted the paired data plots. Here we are looking for any linear relationships between any two variables.
We observed the following linear relationships with (log) strike days (see Figure 4, next page): Year, Inflation, Unemployment, Union Centralization, and Union Density (vaguely). We will also assume a relationship between strike days and country/region.

We now look at the paired data for any interaction relationships, as well as using our own subjective judgement. Some such judgements are the following:

- Union Centralization and Union Density are clearly interacting in a linear manner when plotted (see Figure 5 on the next page, bottom-right plot); but also, in an intuitive sense it should be clear that the more authority a union exercises - as measured by centralization - the more its density is important. And vice-versa: the higher the fraction of the population in the unions - as measured by union density - the more authority the unions will have.
- The Year:Country co-variate could be significant, if only for the fact that certain large values of strike days will be taken into account, such as the three largest ones from 1952:Canada, 1956:Finland, and 1968:France - of course, this may not be a good thing and could contribute to "making the model fit the data". However, because of the common sense assumption that each country may have its own pattern of strike days throughout the years, this could be another reason for a good fit by this co-variate.
- Year:Inflation, Year:Unemployment, Year:Democracy Index could also be considerations due to the expected fluctuations in these macroeconomic factors which normally follow trends of some sort throughout the years.

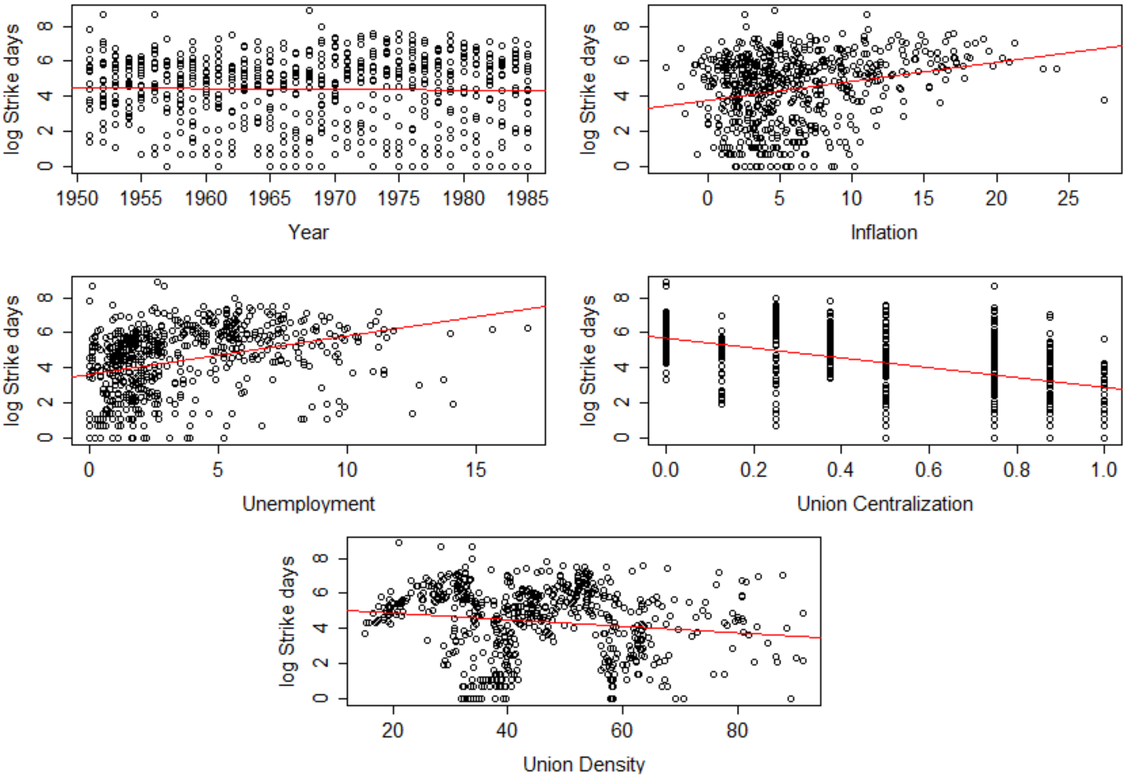


Figure 4: Linear Relationships between log strike days and the mentioned co-variates

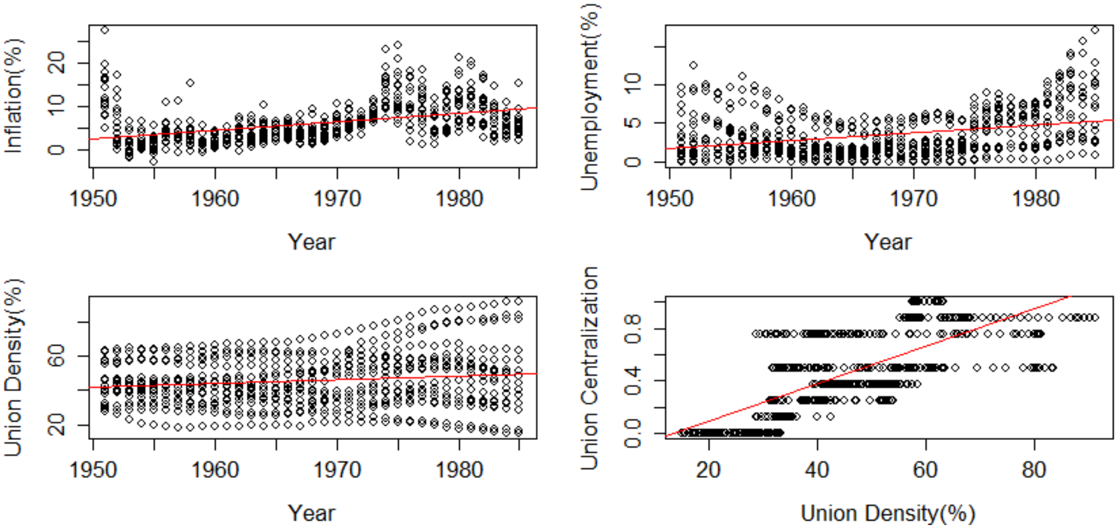


Figure 5: Linear Relationships between the above co-variates

From analyzing Figure 5, there appear to be linear relationships between the following co-variates (although they may be fairly weak) relationships:

Year:Inflation   Year:Unemployment
Year:(Union Density)   (Union Centralization):(Union Density)

We proceed as follows: We add in the co-variates which we believe to be the most important, and test for significance at each stage; as well as testing for the use of countries vs. the use of regions. Additionally, some co-variates (notably, the interactions) were tested for their significance diligently through use of many, many ANOVA tests. This was a *very* lengthy process and we will not elaborate on all the details, but will list the order in which we ranked the co-variates in terms of importance (in Appendix D).

The rankings were based on either the intuitive macroeconomic link between the co-variate and number of strike days, or between two co-variates; or the observed strength of the linear relationship in the plots in Figures 4 & 5.

In the end we chose the model with countries, rather than regions; because the model with regions[6] had standard squared errors that were far larger (∼1665 for regions, as opposed to ∼750 for countries).

Therefore, the final qualitative model that we selected is:

$$\begin{aligned}\text{Strike} \sim \text{Country} + \text{Inflation} + \text{Year}\\ + (\text{Union Density}):(\text{Union Centralization}) - 1\end{aligned}$$

This model is nice and simple, and very easily interpretable[7]. There were many other interaction terms that were considered, but they did not change the fit of the model very much, and only complicated things, and in some cases making interpretation very confusing. See Appendix D for more details.

One remarkable thing was that replacing Year with only Country:Year did not change the sum of squared residuals at all, but Year is easier to interpret so we will keep it. Removing Unemployment from the equation did not have much effect, and this was not surprising due to the Phillips curve[8] explained in detail in Appendix C.

This ends the process of Model Selection.

2. Model Diagnostics

We've selected our two candidate models, and now we will examine them in more detail. We will compare some fundamentals of both models to begin, and then we will proceed to more in-depth analysis.

We will begin with the following metrics:

```
                            Quantitative Model  Qualitative Model
Sum-of-squared Press               706.7084943        804.0835476
Sum-of-squared DFFITS               67.5450811         22.0237561
Akaike Information Criterion      1843.6544546       1932.1018628
R^2                                  0.7615597          0.9487787
```

[6] The model with regions was Strike ∼ Region + Year + Infl + Dens:Centr -1

[7] The intercept was removed for more easily interpretable estimates

[8] http://www.econlib.org/library/Enc/PhillipsCurve.html

We can see that our Qualitative model is preferred with respect to the sum-of-squared DFFITS residuals and Coefficient of Determination (R^2), but for the sum-of-squared Press residuals and Akaike Information Criterion the Quantitative model is preferred.

We will deem the results inconclusive, for now. With that said, we should note the large difference in the sum-of-squared DFFITS; which to us indicates that perhaps the Quantitative model is being affected by high leverage observations in a way that is not affecting the Qualitative model.

Therefore we will begin by investigating the high-leverage observations and their influence on the fit of both models.

Figure 6 shows the Leverage plotted against the Cook's Distance for both models, and some striking differences can be seen.

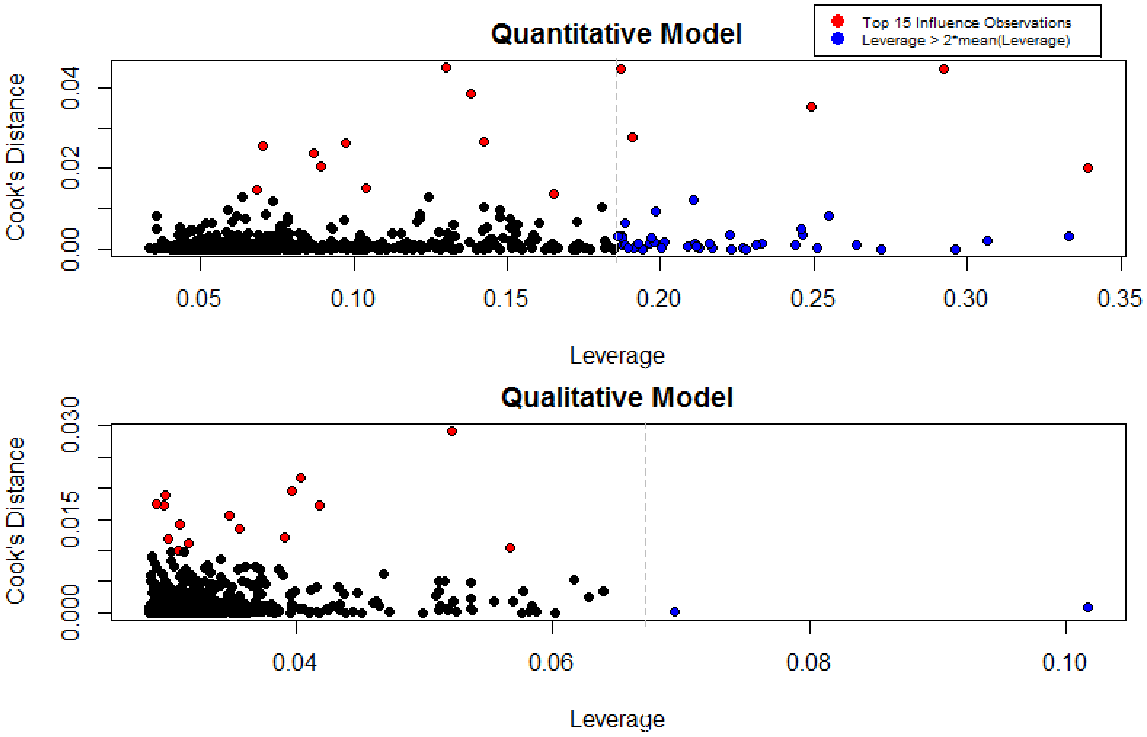


Figure 6: Leverage vs. Cook's Distance

For the Quantitative model there are 5 high-leverage observations which are in the top 15 influence observations. As opposed to *zero* such observations in the Qualitative model. It is also worth noting that in the Qualitative model there are far less high-leverage observations, and that the top influence observations are not as far from the others as they are in the Quantitative model. So, as we suspected; the Quantitative model is being highly influenced by observations with high leverage.

Let us analyze the Quantitative Model first:

- The 5 observations which have high leverage and influence are all to be found in the 98.2th percentile of DFFITS residuals and 93.1th percentile of PRESS residuals; meaning that when we take only *one* of these observations out, it changes the estimator $\hat{\beta}$ far more than any other observations.

- None of the high leverage&influence observations appear to actually be of particular interest when singled out in the paired plots; whether on the log(strikes) scale or the original. In particular, they were not those three points which were extremely high in strike days (>5000).
- One common thread between the five high leverage&influence observations is that four of them have very high Union Centralization (0.75, 0.875, 0.875, 1). We guess that the fifth observation is influential and high leverage due to its high Unemployment rate of 8.2% (91.9th percentile).

Details of this analysis are in Appendix E.

And now for the Qualitative Model:

- The qualitative model has the advantage of being more easily interpretable, by an extremely large margin.
- The Qualitative model has R^2 value far closer to 1 (0.9488 compared to 0.7616); indicating a good fit.
- The fit of the Qualitative method produces residuals which closer to normal and closer to constant-variance:

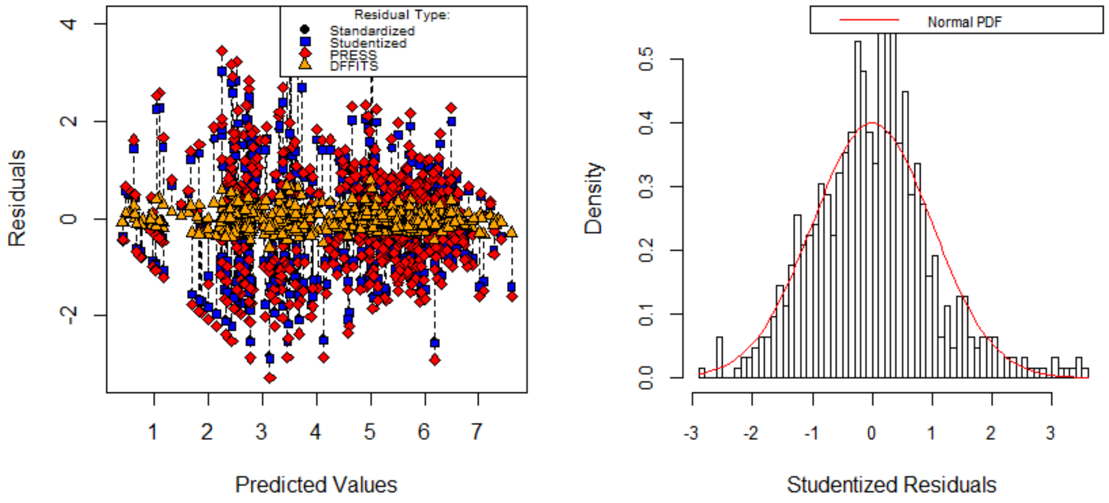


Figure 7: Fit of the Qualitative Model

We are tempted to already pick the Qualitative model, but one last test remains:
We will perform cross-validation with 600 observations for training. Note that cross-validation will be run with 8,000 replications for extremely good accuracy, and the sum-of-squared test residuals are on the original scale of strike days.

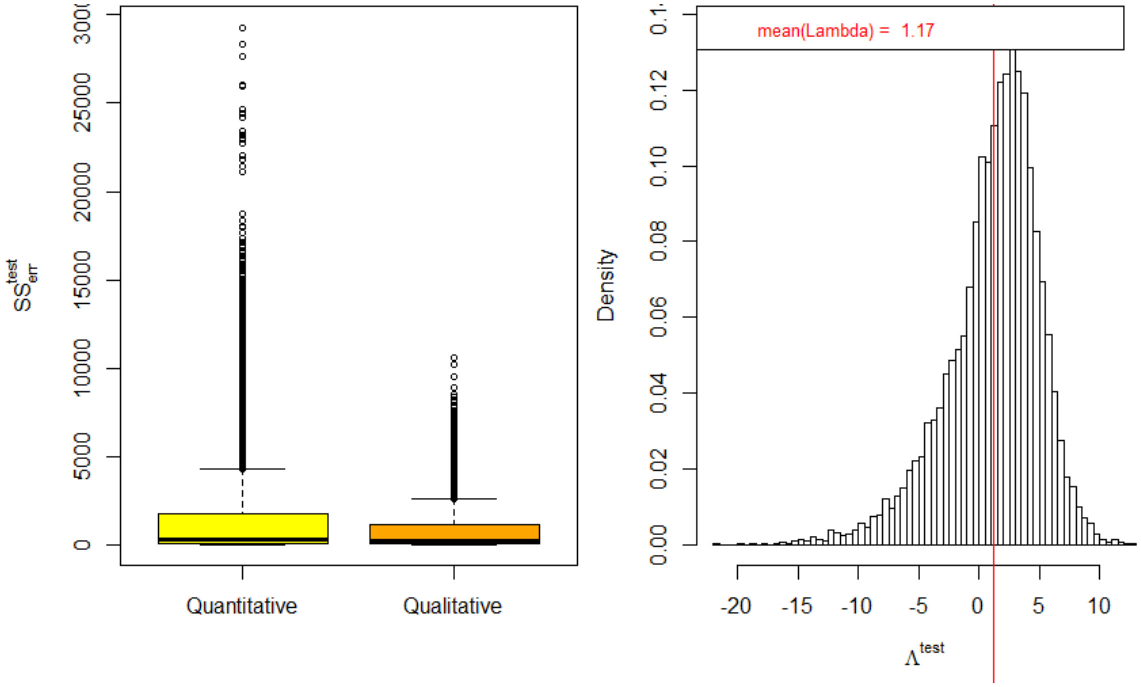


Figure 8: Cross-Validation with n=600 training sample size; slightly prefers Quantitative Model

Therefore, **we pick the Qualitative model as our final choice**. For the following reasons:

• The sum-of-squared PRESS residuals and Akaike Information Criterion do not show an extremely large difference between the two models. The sum-of-squared DFFITS residuals prefer the Qualitative model by a very large margin, however: with the Qualitative model's sum-of-squared DFFITS being about $\frac{1}{3}$ of the Quantitative model's.

• *ALL* of the co-variates in the Qualitative Model are significant (see Appendix F).

• Although the Cross-validation technically preferred the Quantitative model in terms of mean Lambda value, the box-plot tells a different story. The Quantitative model showed some very large sum-of-square test errors (shown in Figure 8). Clearly it is over-fitting the data. Also, in the histogram of Lambda values, there are far more extreme values on the negative side of the axis - which again points to the Qualitative model being far more accurate.

• The Quantitative model has far more high influence observations, and more high influence observations which also have high leverage (the Qualitative model, again, has zero). This is of course represented in the sum-of-squared DFFITS residuals, but the extent to which it affects the fit of the model is not entirely clear from that alone. The cross-validation box-plot in Figure 8 provides more evidence of the over-fitting that the sum-of-squared DFFITS residuals were pointing to.

• As mentioned before, Figure 7 shows that the residuals from the Qualitative model are closer to regression assumptions, as compared to those of the Quantitative model in Figure 3 (Page 4).

Final Model:

log(Strike + 1) = β_0 Country + β_1 Inflation + β_2 Year + β_3 (Union Density)×(Union Centralization) + ϵ

$$\implies \text{Strike} \sim \prod_{i=1}^{18} \left[\left(e^{\text{II(Country=country}_i)} \right)^{\beta_{0i}} \right] \cdot \left(e^{\text{Inflation}} \right)^{\beta_1} \cdot \left(e^{\text{Year}} \right)^{\beta_2}$$

$$\cdot \left(e^{\text{(Union Density)} \times \text{(Union Centralization)}} \right)^{\beta_3} \Big) - 1$$

Where β_{0i} corresponds to the given country$_i$ and II is the indicator function;

$$\text{i.e II(x)} = \begin{cases} 0 \text{ if x is false} \\ 1 \text{ if x is true} \end{cases}$$

2.1 Final Model Parameter Estimates & Confidence Intervals

```
                      Estimate   95% Confidence Interval
CountryAustralia       52.253     [ 32.445 , 72.06  ]
CountryAustria         47.040     [ 27.432 , 66.647 ]
CountryBelgium         50.872     [ 31.173 , 70.572 ]
CountryCanada          53.726     [ 33.787 , 73.665 ]
CountryDenmark         49.535     [ 29.799 , 69.271 ]
CountryFinland         50.652     [ 30.962 , 70.342 ]
CountryFrance          52.507     [ 32.586 , 72.428 ]
CountryGermany         49.844     [ 29.949 , 69.739 ]
CountryIreland         52.185     [ 32.416 , 71.954 ]
CountryItaly           53.537     [ 33.687 , 73.388 ]
CountryJapan           51.793     [ 31.887 , 71.698 ]
CountryNetherlands     48.763     [  28.99 , 68.536 ]
CountryNewZealand      51.613     [ 31.793 , 71.432 ]
CountryNorway          48.505     [ 28.862 , 68.149 ]
CountrySweden          47.200     [ 27.624 , 66.775 ]
CountrySwitzerland     47.662     [ 27.816 , 67.508 ]
CountryUnitedKingdom   51.991     [ 32.174 , 71.807 ]
CountryUnitedStates    53.419     [ 33.477 , 73.361 ]
Infl                    0.063     [  0.041 , 0.086  ]
Year                   -0.024     [ -0.034 , -0.014 ]
Centr:Dens              0.045     [  0.024 , 0.065  ]
```

Then if we are given a country, the year, the inflation rate, union density and union centralization, we could predict strike days as follows:

$$\text{Strike} \sim \exp\Big(\beta_{0i} + 0.063 \cdot \text{Inflation} - 0.024 \cdot \text{Year}$$

$$+0.045 \cdot (\text{Union Density} \times \text{Union Centralization}) \Big) - 1$$

Using the appropriate β_{0i} depending on the Country in question. Let us use an example of the USA in 1970, with 6.4% inflation 0.375 union centralization and 48.2% union density. Then we expect 1564[9] strike days. A confidence interval would be more complex.

[9] $\exp\Big(53.419 + 0.063 \cdot 6.4 - 0.024 \cdot 1970 + 0.045 \cdot 0.375 \cdot 48.2\Big) - 1$

3. Closing Remarks

From our final model we can make some remarks with regards to the factors influencing strike activity in OECD countries during the postwar period;

• Inflation was seen to be an important factor in nearly every model considered, and it has a large significance in our final model. This is not unexpected whatsoever; as inflation pushes prices up while wages stay stagnant - which could prompt workers to strike for higher wages which match inflationary pressures.

• Union Density:Union Centralization is also an important factor, which as explained previously makes perfect sense. The more people that are in unions, the higher the authority of the unions. Similarly; the more authority a union has, the more relevant it becomes how many people are in the union. This combination of predictors had a positive correlation with the number of strike days: indicating that the higher the union density and union centralization (which also have a positive correlation of 0.718), the higher the expected number of strike days. Rightly so; organized workers with a voice will have incentive to "push back" to benefit themselves.

• Having the Year as a predictor shouldn't be shocking. Certain political movements throughout the years can effect all sorts of macroeconomic trends, and there are certain years where there were periods of high strike days across the board. This can be seen in the pairs plots as a "wave" of sorts in the Strike-Years plot; the number of strikes appears to move in a cyclical manner with respect to the year.

• Lastly, the Country predictor is obvious. Every country has its own internal economic structure and political movements which can affect strikes.

We had considered removing some of the extremely large outlying values of strike days; there were three such points at 5918, 5568, and 7000 strike days. The fourth largest observation had only 2875 strike days, and the 99th percentile of strike days is only 1906.

One of these observations (5568 strike days) was a major political revolt in Finland[10].

The observation with 5,918 strike days comes from Canada in 1952 where there was a large strike of 1,200 employees from a large company called Dupuis Freres in Quebec[11].

Lastly, the observation with 7,000 strike days came from France in 1968; where there was not only a general strike as in Finland, but a student revolt[12]. There were actually *10 million* workers who joined the strike, so the data may even be under-representing the scale of it.

With all that said, we did not remove these observations, although we would have certainly removed the one from France if it were accurately reported as *millions* of strike days. The data set may have just set a maximum of 7,000.

We are content with our model's fit, and its alignment with regression assumptions. The one issue we foresee is perhaps its simplicity. While making interpretation easy, it may be its downfall. There were decent candidate co-variates to add to the model but we were concerned with their relevance as they did not show promise as significant predictors.

It should be noted, again, that our model contains *zero* insignificant predictors. In fact, the highest p-value is of the order 10^{-5}; see Appendix F.

[10] https://en.wikipedia.org/wiki/General_strike_of_1956

[11] http://www.virtualmuseum.ca/edu/ViewLoitDa.do;jsessionid=280BB4770A45F6525272514C8398D2A0?method=preview&lang=EN&id=25204

[12] https://www.wsws.org/en/articles/2008/05/may1-m28.html

Appendix A: Data Transformations

$$\begin{aligned}\textbf{Minimal Model}:&\ \text{Strike} \sim 1\\ \textbf{Maximal Model}^{\dagger}:&\ \text{Strike} \sim \Big(\text{Country} + \text{Year} + \text{Unemployment} + \text{Inflation}\\ &\qquad + \text{Democracy Index} + \text{Union Centralization}\\ &\qquad + \text{Union Density}\Big)^2\\ \textbf{Stepwise Starting Model}:&\ \text{Strike} \sim \Big(\text{Country} + \text{Year} + \text{Unemployment} + \text{Inflation}\\ &\qquad + \text{Democracy Index} + \text{Union Centralization}\\ &\qquad + \text{Union Density}\Big)\end{aligned}$$

†: The square term indicates that all first-order interaction effects are included. These terms are later denoted as A:B for A interacting with B.

The results of this process are:

$$\begin{aligned}\textbf{Forward Model}:&\ \text{Strike} \sim \text{Country} + \text{Inflation} + \text{Unemployment}\\ \textbf{Backward Model}:&\ \text{Strike} \sim \text{Country} + \text{Year} + \text{Unemployment} + \text{Inflation}\\ &\quad + \text{Democracy Index} + \text{Union Density} + \text{Year:(Democracy Index)}\\ &\quad + \text{Unemployment:(Democracy Index)}\\ &\quad + \text{(Democracy Index):(Union Density)}\\ \textbf{Stepwise Model}:&\ \text{Strike} \sim \text{Country} + \text{Unemployment} + \text{Inflation}\end{aligned}$$

Note that the Forward and Stepwise models are the same, and that they are both contained within the Backward model; that is, they are reduced forms of the Backward model.
Therefore, we can use the *ANOVA* function in **R** to calculate the significance of the additional coefficients found in the Backward model.
It has an F-statistic p-value of 0.1483, which is not significant. Therefore we will use the Forward (or Stepwise) model for this brief analysis.
Taking a look at the plot of the residuals for the Forward model, we see a striking problem. The residuals do not have constant variance (fan shape indicates increasing variance as the predicted values increase), and from the histograms we can see that they are also not normally distributed.
Furthermore, we can see that the residuals corresponding to the observations with less than the median number of strikes (302.3 - in red) show a downward negative trend as the predicted value increases, and they are nearly all negative. We speculate that this is due to the extremely large outliers (in blue) which have 10σ residuals which is a very large unbalance.
Therefore we must perform a transformation of the data. After some trial and error, the most sensible transformation seems to be the following:

$$\text{Strikes} \longleftarrow \log(\text{Strikes} + 1)$$

We include the “+1” in order to avoid the issue of infinite numbers, since some values of Strikes (strike days) are zero, and these values still now map to zero.

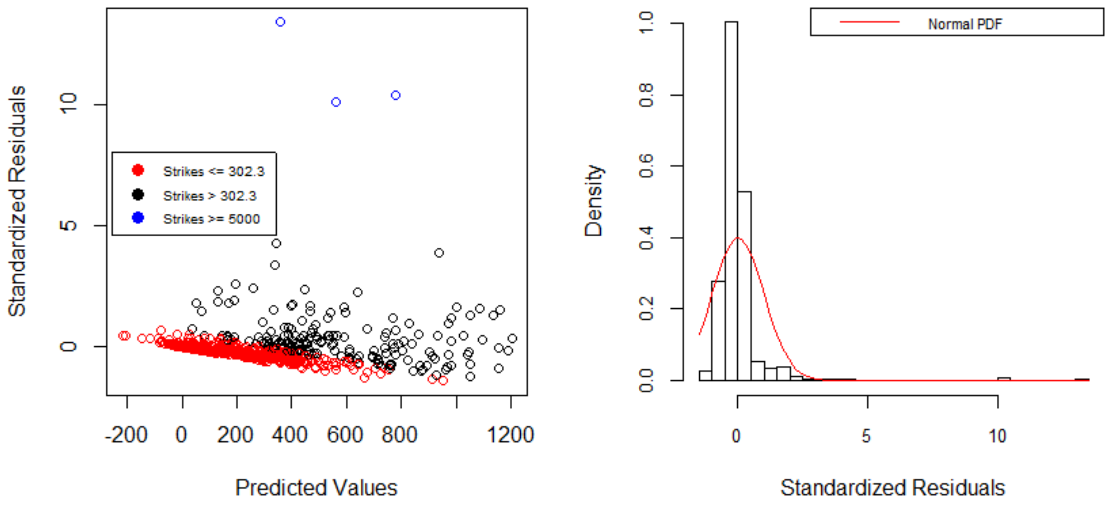


Figure 9: Standardized Residuals from the Forward/Stepwise models

As we can see; after the transformation, the residuals appear to be much closer to normal, as well as having constant variance.

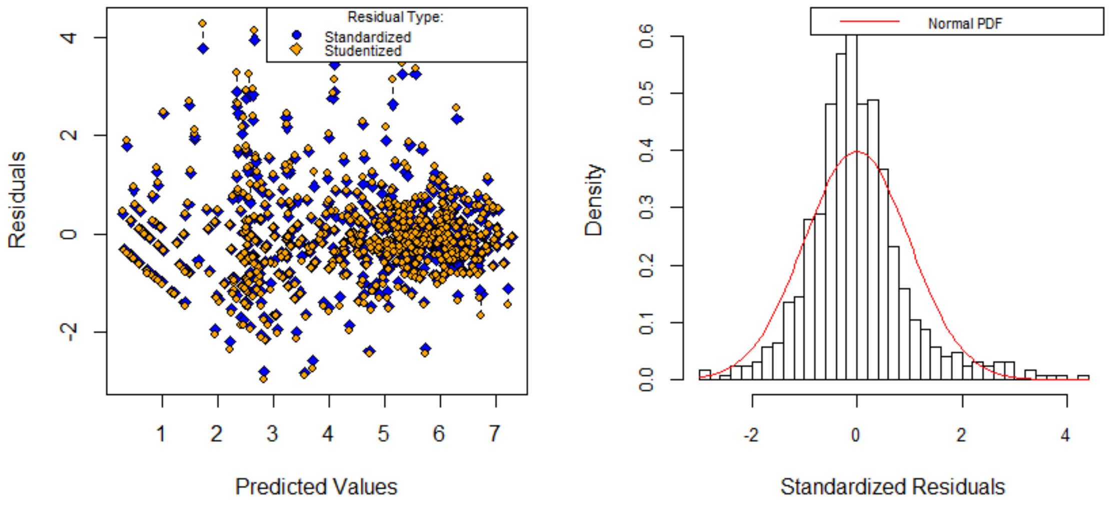
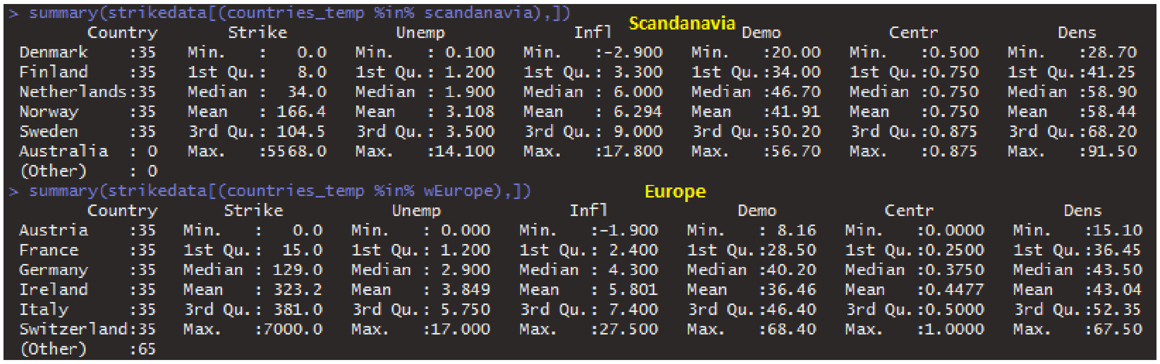


Figure 10: Residuals after the transformation of Strikes

Click to Return

Appendix B: Reasoning in separating Scandanavia from Europe



```
> summary(strikedata[(countries_temp %in% scandanavia),])
        Country        Strike           Unemp             Infl             Demo            Centr             Dens      
 Denmark    :35   Min.   :   0.0   Min.   : 0.100   Min.   :-2.900   Min.   :20.00   Min.   :0.500   Min.   :28.70  
 Finland    :35   1st Qu.:   8.0   1st Qu.: 1.200   1st Qu.: 3.300   1st Qu.:34.00   1st Qu.:0.750   1st Qu.:41.25  
 Netherlands:35   Median :  34.0   Median : 1.900   Median : 6.000   Median :46.70   Median :0.750   Median :58.90  
 Norway     :35   Mean   : 166.4   Mean   : 3.108   Mean   : 6.294   Mean   :41.91   Mean   :0.750   Mean   :58.44  
 Sweden     :35   3rd Qu.: 104.5   3rd Qu.: 3.500   3rd Qu.: 9.000   3rd Qu.:50.20   3rd Qu.:0.875   3rd Qu.:68.20  
 Australia  : 0   Max.   :5568.0   Max.   :14.100   Max.   :17.800   Max.   :56.70   Max.   :0.875   Max.   :91.50  
 (Other)    : 0                                                                                                     
> summary(strikedata[(countries_temp %in% wEurope),])
        Country        Strike           Unemp             Infl             Demo            Centr             Dens      
 Austria    :35   Min.   :   0.0   Min.   : 0.000   Min.   :-1.900   Min.   : 8.16   Min.   :0.0000   Min.   :15.10  
 France     :35   1st Qu.:  15.0   1st Qu.: 1.200   1st Qu.: 2.400   1st Qu.:28.50   1st Qu.:0.2500   1st Qu.:36.45  
 Germany    :35   Median : 129.0   Median : 2.900   Median : 4.300   Median :40.20   Median :0.3750   Median :43.50  
 Ireland    :35   Mean   : 323.2   Mean   : 3.849   Mean   : 5.801   Mean   :36.46   Mean   :0.4477   Mean   :43.04  
 Italy      :35   3rd Qu.: 381.0   3rd Qu.: 5.750   3rd Qu.: 7.400   3rd Qu.:46.40   3rd Qu.:0.5000   3rd Qu.:52.35  
 Switzerland:35   Max.   :7000.0   Max.   :17.000   Max.   :27.500   Max.   :68.40   Max.   :1.0000   Max.   :67.50  
 (Other)    :65                                                                                                      
```

Figure 11: Scandanavian Data vs. European Data

We can see from the data summaries that Scandanavia differs significantly from Europe in several categories:
† Scandanavia has a lower number of strikes, by a significant margin[13]
† Scandanavia also has lower unemployment, lower inflation, consistent union centralization, and high trade union density.
† It is also generally known that Scandanavia has higher standards of living, education, and work than the rest of Europe.

Click to Return

[13] one of the potential outliers at 5,568 strike days is from Finland in 1956 - referred to as the "General Strike of 1956", the last of the three general strikes after the first two in 1905 and 1917. The central federation of the trade unions started the action (Union Centralization was 0.75 in this year - quite high), and wages were increased 6 to 10 % as a result https://en.wikipedia.org/wiki/General_strike_of_1956

Appendix C: Quantitative Model selection details

The results of Automated Model Selection were:

$$\begin{aligned}
\textbf{Forward Model} : \text{Strike} &\sim \text{Country} + \text{Year} + \text{Inflation} + \text{Union Density}\\
&+ \text{Country:(Union Density)} + \text{Country:(Inflation)} + \text{Country:Year}\\
\textbf{Backward Model} : \text{Strike} &\sim \text{Country} + \text{Year} + \text{Inflation} + \text{Union Density} + \text{Unemployment}\\
&+ \text{Country:Year} + \text{Country:Unemployment} + \text{Year:(Union Density)}\\
&+ \text{Unemployment:(Union Density)}\\
\textbf{Stepwise Model} : \text{Strike} &\sim \text{Country} + \text{Year} + \text{Inflation} + \text{Union Density} + \text{Unemployment}\\
&+ \text{Country:Year} + \text{Country:(Union Density)}\\
&+ \text{Unemployment:Inflation}
\end{aligned}$$

Note that there are common elements in all three models (in blue): Country, Year, Inflation, Union Density, Country:Year.
We will define a new model named "Test Model" with these co-variates, and compare it (as a reduced model) to the three automated models with ANOVA.
All three models were seen to have significant additional terms when compared to the Test Model. However, the Forward model has too many coefficients (72 for 625 data points), and the Backward model has more significance towards co-variates that both the Backward and Stepwise model have in common. The co-variates Year:(Union Density) and Unemployment:(Union Density) are also significant in the Backward model (p-values 2.8325×10^{-2} and 7.68×10^{-4} respectively), which are not present in the Stepwise model.
Additionally, the Stepwise model contains the Unemployment:Inflation term which may be deceptive as according to the Phillips Curve[14] found in macroeconomics, Unemployment and Inflation follow an inverse relationship. Therefore, if we multiply them together we may introduce some peculiarities into the model.
As explained in the Model Selection section, we tested the "Test Model" against the three automated models, using ANOVA, since Test Model is a reduced form of all three.
The results are as follows:

Model	F-statistic p-value	# of β in model
Forward	2.31×10^{-3}	72
Backward	3.511×10^{-4}	58
Stepwise	1.779×10^{-3}	57
Test	N/A	38

The reason for choosing the Backward Model over the Stepwise Model was the following:

Model	Unemployment(p-val)	Union Density(p-val)	Residual Std. Error	R^2
Backward	0.014427	0.032383	1.01	0.7616
Stepwise	0.871147	0.503114	1.015	0.7588

[14] http://www.econlib.org/library/Enc/PhillipsCurve.html

The p-values for Unemployment and Union Density are very large for the stepwise model. It essentially does not give any significance to Unemployment or Union Density.
Note that these models had extremely similar residual standard errors and R^2 values so this was not a factor in the decision.
Additionally the terms that are not found in either model varied in their significance in such a way that prefers the Backward Model:

Model	Co-variate	p-value
Backward	Year:Unemployment	2.8325×10^{-2}
	Unemployment:(Union Density)	7.86×10^{-4}
Forward	Unemployment:Inflation	7.7121×10^{-2}

The interaction of Unemployment and Inflation is being assumed to be a poor co-variate due to the Phillips Curve mentioned previously.

Click to Return

Appendix D: Qualitative Model selection details

The order of most important, to least important co-variate is, in our opinion:

1. Country(or Region)
2. Year
3. Inflation
4. Unemployment
5. Country:Year
6. Union Density
7. (Union Centralization):(Union Density)
8. Year:Inflation
9. Year:Unemployment
10. Year:(Union Density)

As mentioned, when going through the value of these co-variates in the model, we tried many combinations and particularly tried to find the significance and change in residuals from including or excluding certain interaction terms.
Here are the results:

```
  Added -->   Unemployment Union Density Year:Inflation Year:Unemployment
SSQ Resid      749.378       741.347        745.360           749.319
Change           0.000        -8.031         -4.018            -0.060
R^2              0.949         0.949          0.949             0.949
Change           0.000         0.001          0.000             0.000
Sigma hat        1.115         1.109          1.112             1.115
Change           0.000        -0.006         -0.003             0.000
p-value          0.327         0.006          0.040             0.315
```

There were not many significant changes, and although some had low p-values, they did not affect the fit enough to justify their addition. i.e: Year:Inflation complicates the model too much for little change in fit, and Union Density being included changed the p-value of (Union Density):(Union Centralization) to a non-significant value from its previously very significant value of 2.07×10^{-5}.
Code used to get the table above:

```
################################################################################
################################## APPENDIX D ##################################

### Models which all add in one co-variate. In this order: Unemp, Dens,
  # Year:Infl, Year:Unemp
Mqualtest1 <- lm(Strike ~ Country + Infl + Dens:Centr + Year - 1 + Unemp,
                 data = strikedata)
Mqualtest2 <- lm(Strike ~ Country + Infl + Dens:Centr + Year - 1 + Dens,
                 data = strikedata)
Mqualtest3 <- lm(Strike ~ Country + Infl + Dens:Centr + Year - 1 + Year:Infl,
                 data = strikedata)
Mqualtest4 <- lm(Strike ~ Country + Infl + Dens:Centr + Year - 1 + Year:Unemp,
                 data = strikedata)

# Get sum-of-squared residuals for all 4 models
test.resid.sq <- c(sum(resid(Mqualtest1)^2),sum(resid(Mqualtest2)^2),
                   sum(resid(Mqualtest3)^2),sum(resid(Mqualtest4)^2))
names(test.resid.sq) <- c("Unemployment", "Union Density", "Year:Inflation",
                          "Year:Unemployment")
# Compare to Mqual's sum-of-squared residuals
test.resid.sq.change <- test.resid.sq - sum(resid(Mqual)^2)

# Get R^2 values for all 4 models
test.rsq <- c(summary(Mqualtest1)$r.squared,summary(Mqualtest2)$r.squared,
              summary(Mqualtest3)$r.squared,summary(Mqualtest4)$r.squared)
# Compare to Mqual's R^2 value
test.rsq.change <- test.rsq - summary(Mqual)$r.squared

# Get sigma hat for all 4 models
test.sigma <- c(summary(Mqualtest1)$sigma,summary(Mqualtest2)$sigma,
                summary(Mqualtest3)$sigma,summary(Mqualtest4)$sigma)
# Compare to Mqual's sigma hat
test.sigma.change <- test.sigma - summary(Mqual)$sigma

# Get p-values for the significance of each additional co-variate in their
# respective models.
test.pval <- c(summary(Mqualtest1)$coefficients[21,4],
               summary(Mqualtest2)$coefficients[21,4],
               summary(Mqualtest3)$coefficients[22,4],
               summary(Mqualtest4)$coefficients[22,4])

### Final Output ###
test.matrix <- rbind(round(test.resid.sq,3), round(test.resid.sq.change,3),
                     round(test.rsq,3), round(test.rsq.change, 3),
                     round(test.sigma,3), round(test.sigma.change,3),
                     round(test.pval,3))
row.names(test.matrix) <- c("SSQ Resid", "Change", "R^2", "Change","Sigma hat",
                            "Change", "p-value")
```

Click to Return

Appendix E: Quantitative Model analysis

```
  Observation -->           1      2      3      4     5
DFFITS Percentile       0.9936 0.9824 0.9984 0.9968 0.992
Press Percentile        0.9728 0.9312 0.9920 0.9792 0.976
Union Centralization    1.0000 0.7500 0.2500 0.8750 0.875
Unemployment Percentile 0.6512 0.8640 0.9088 0.5952 0.560
```

Code used for this table:

```
##############################################################################
################################ APPENDIX E ##################################

Mquant.h <- hatvalues(Mquant)                      # Hat values (leverages)
Mquant.h.bar <- mean(Mquant.h)                     # Average leverage
Mquant.highlev <- (Mquant.h >= 2*Mquant.h.bar)     # High leverage indices
Mquant.cookD <- cooks.distance(Mquant)             # Cook's distance
Mquant.high.cookD <- (Mquant.cookD >=             # top 15 influence observations
                        quantile(Mquant.cookD, probs = (610/625)))
Mqual.high.cookD <- (Mqual.cookD >=               # top 15 influence observations
                       quantile(Mqual.cookD, probs = (610/625)))
# Finding indices which are both high leverage and high influence:
Mquant.remove <- (Mquant.highlev & Mquant.high.cookD)
Mquant.remove.index <- as.numeric(
  row.names(strikedata[(Mquant.highlev & Mquant.high.cookD),]))

# Observations from strikedata which correspond to the indices above
 # i.e high leverage and influence
E.obs <- strikedata[Mquant.remove.index,]
# The magnitude of the DFFITS residuals of the observations found in E.obs
E.DFFITS <- abs(Mquant.DFFITS[Mquant.remove.index])
# The magnitude of the press residuals of the observations found in E.obs
E.press <- abs(Mquant.press[Mquant.remove.index])
# The unemployment figures for the observations in E.obs
E.unemp <- E.obs[,4]
# The union centralization figures for the observations in E.obs
E.centr <- E.obs[,7]

### Percentiles of DFFITS:
E.DFFITS.CDF <- ecdf(abs(Mquant.DFFITS)) # Empirical CDF function
E.DFFITS.p <- c(E.DFFITS.CDF(E.DFFITS))

### Percentiles of PRESS:
E.press.CDF <- ecdf(abs(Mquant.press))   # Empirical CDF function
E.press.p <- c(E.press.CDF(E.press))

### Percentiles of Unemployment
E.unemp.CDF <- ecdf(strikedata$Unemp)    # Empirical CDF function
E.unemp.p <- c(E.unemp.CDF(E.unemp))

#### FINAL OUTPUT ####
E.matrix <- rbind(E.DFFITS.p, E.press.p, E.centr, E.unemp.p)
row.names(E.matrix) <- c("DFFITS Percentile","Press Percentile",
                         "Union Centralization", "Unemployment Percentile")
colnames(E.matrix) <- c(1,2,3,4,5)
```

Additionally there was the claim that the five points in question were of no particular interest in the pairs plots. Here is that pairs plot:

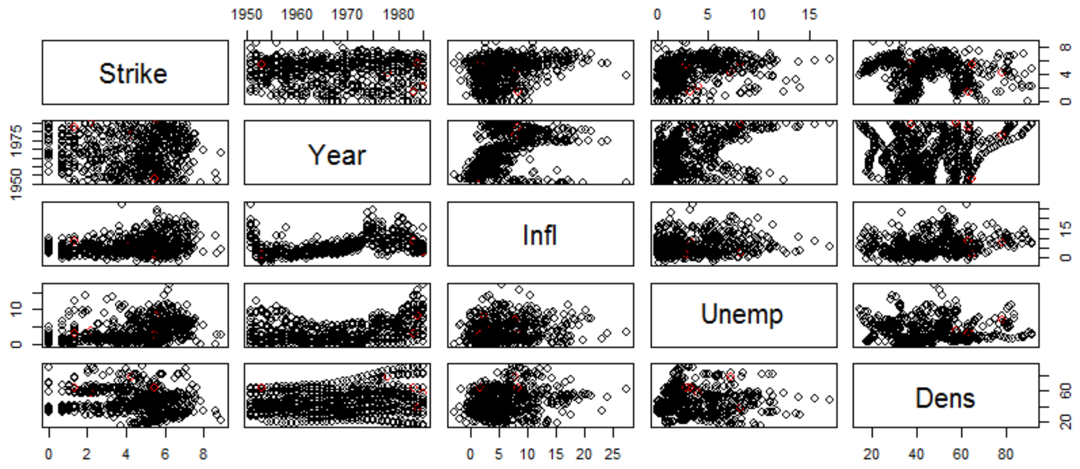


Figure 12: Paired Plots relevant to the Quantitative model. The high leverage&influence points are marked as red

Code for figure 12:

```
###### FIGURE 12: Pairs Plot with Quantitative Model influencial observations
##### labelled to show that they don't appear to be "special" in any way.
clrs.quant.outliers <- rep("black", nrow(strikedata))
clrs.quant.outliers[Mquant.remove.index] <- "red"
pairs(~ Strike + Year + Infl + Unemp + Dens + Year:Dens + Unemp:Dens,
      col = clrs.quant.outliers, data = strikedata, cex = 1.3)
```

Click to Return

Appendix F: Summary of Final Model

```
Call:
lm(formula = Strike ~ Country + Infl + Centr:Dens + Year - 1,
    data = strikedata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1399 -0.7170 -0.0149  0.6002  3.9485

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
CountryAustralia     52.252658  10.106016   5.170 3.18e-07 ***
CountryAustria       47.039777  10.003846   4.702 3.19e-06 ***
CountryBelgium       50.872474  10.050998   5.061 5.53e-07 ***
CountryCanada        53.726113  10.172969   5.281 1.79e-07 ***
CountryDenmark       49.535172  10.069321   4.919 1.12e-06 ***
CountryFinland       50.651874  10.045854   5.042 6.10e-07 ***
CountryFrance        52.507097  10.163767   5.166 3.25e-07 ***
CountryGermany       49.844360  10.150444   4.911 1.17e-06 ***
CountryIreland       52.185295  10.086126   5.174 3.12e-07 ***
CountryItaly         53.537469  10.127600   5.286 1.75e-07 ***
CountryJapan         51.792762  10.155755   5.100 4.56e-07 ***
CountryNetherlands   48.762974  10.088405   4.834 1.70e-06 ***
CountryNewZealand    51.612627  10.111903   5.104 4.46e-07 ***
CountryNorway        48.505387  10.022237   4.840 1.65e-06 ***
CountrySweden        47.199679   9.987535   4.726 2.85e-06 ***
CountrySwitzerland   47.661935  10.125619   4.707 3.12e-06 ***
CountryUnitedKingdom 51.990580  10.110576   5.142 3.67e-07 ***
CountryUnitedStates  53.418636  10.174551   5.250 2.11e-07 ***
Infl                  0.063099   0.011474   5.499 5.63e-08 ***
Year                 -0.024276   0.005179  -4.688 3.42e-06 ***
Centr:Dens            0.044904   0.010464   4.291 2.07e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.115 on 604 degrees of freedom
Multiple R-squared:  0.9488,Adjusted R-squared:  0.947
F-statistic: 532.8 on 21 and 604 DF,  p-value: < 2.2e-16
```

Click to Return

Appendix G: Source Code[15]

```
################################################################################
##                            STAT 331 Final Project                          ##
##                               Daniel Matheson                              ##
##                                  20270871                                  ##
################################################################################

options(scipen=3) # sets the threshold for scientific notation
strikedata <- read.csv("strikes_clean.csv") # import data
## strikedata:
# Country: 18 countries
# Year: every year from 1951 to 1985
# Strike: number of days in the year lost per 1,000 workers due to strikes
# Unemp: Unemployment rate (%)
# Infl: Inflation rate (%)
# Demo: A democracy index, defined as proportion of left-party parliamentary
# representation
# Centr: Measure of union centralization, which refers to "the authority that
# union condeferations have over their members". The higher this value,
# the more powerful the union. The measure is aggregated over all years in a
# given country.
# Dens: Trade union density, which is the fraction of wage earners in the
# country who belong to a trade union.

################################################################################
############################ VARIABLE MANIPULATION #############################
# Creating the Region co-variate:

countries_temp <- strikedata$Country
wEurope <- c("Austria","Belgium","France","Germany","Ireland","Italy",
             "UnitedKingdom","Switzerland")
strikedata$Region[ (countries_temp %in% wEurope)] <- "Europe"

north_america <- c("Canada","UnitedStates")
strikedata$Region[(countries_temp %in% north_america)] <- "NorthAmerica"

ausnz <- c("Australia", "NewZealand")
strikedata$Region[(countries_temp %in% ausnz)] <- "AusNZ"

strikedata$Region[(countries_temp == "Japan")] <- "Japan"

scandanavia <- c("Denmark","Finland","Netherlands","Norway","Sweden")
strikedata$Region[(countries_temp %in% scandanavia)] <- "Scandanavia"

# Taking log(Strike + 1)
strikedata$Strike <- log(strikedata$Strike + 1)

################################################################################
########################  AUTOMATED MODEL SELECTION ############################
M0 <- lm(Strike ~ 1, data = strikedata)          # Minimal Model
Mfull <- lm(Strike ~ (.)^2, data = strikedata)   # Maximal Model

```

[15] Excludes any code already listed in Appendices D and E

```r
## Foward Model
Mfwd <- step(object = M0, scope = list(lower = M0, upper = Mfull),
             direction = "forward", trace = F)
## Backward Model
Mback <- step(object = Mfull, scope = list(lower = M0, upper = Mfull),
              direction = "backward", trace = F)
## Stepwise Model
Mstart <- lm(Strike ~ ., data = strikedata)      # Stepwise starting model
Mstep <- step(object = Mstart, scope = list(lower = M0, upper = Mfull),
              direction = "both", trace = F)

models <- c(Mfwd$call, Mback$call, Mstep$call)    # Summarize Results
names(models) <- c("FWD", "BACK","STEP")

## Model which has the co-variates common to Mfwd, Mback, Mstep
## Used to perform three ANOVA tests (We will omit these as they are long)
Mtest1 <- lm(Strike ~ Country + Year + Infl + Dens + Country:Year,
             data = strikedata)

#####################################################################################
############################ MODEL DEFINITIONS ####################################
Mquant <- lm(Strike ~ Country + Year + Infl + Unemp + Dens + Country:Year
             + Country:Unemp + Year:Dens + Unemp:Dens, data = strikedata)
Mqual <- lm(Strike ~ Country + Infl + Dens:Centr + Year - 1, data = strikedata)
Mqualregion<- lm(Strike ~ Region + Year + Infl + Dens:Centr + Year -1,
                 data = strikedata) # Not used

#####################################################################################
######################### 2. MODEL DIAGNOSTICS ####################################

### PRESS residuals
Mquant.h <- hatvalues(Mquant)                        # Hat values
Mqual.h <- hatvalues(Mqual)

Mquant.press <- resid(Mquant)/(1-Mquant.h)           # PRESS residuals
Mqual.press <- resid(Mqual)/(1-Mqual.h)

Mquant.press.sq <- sum(Mquant.press^2)               # Sum-of-squared Press
Mqual.press.sq <- sum(Mqual.press^2)                   # residuals

press.list <- c(Mquant.press.sq, Mqual.press.sq)       # Combined for nice output

### DFFITS residuals
Mquant.DFFITS <- dffits(Mquant)                      # DFFITS residuals
Mqual.DFFITS <- dffits(Mqual)

Mquant.DFFITS.sq <- sum(Mquant.DFFITS^2)             # Sum-of-squared DFFITS
Mqual.DFFITS.sq <- sum(Mqual.DFFITS^2)                # residuals

DFFITS.list <- c(Mquant.DFFITS.sq, Mqual.DFFITS.sq)  # Combined for nice output

#### Akaike Information Criterion (AIC)
Mquant.AIC <- AIC(Mquant)                            # AIC
Mqual.AIC <- AIC(Mqual)
AIC.list <- c(Mquant.AIC, Mqual.AIC)                 # Combined for nice output
```

```

#### R^2
Mquant.R2 <- summary(Mquant)$r.squared          # R squared values
Mqual.R2 <- summary(Mqual)$r.squared
R2.list <- c(Mquant.R2, Mqual.R2)              # Combined for nice output

#### Final Output as a Matrix (Bottom of Page 6) ###
diagnost.matrix <- rbind(press.list,DFFITS.list,AIC.list, R2.list)
row.names(diagnost.matrix) <- c("Sum-of-squared Press", "Sum-of-squared DFFITS",
                                "Akaike Information Criterion", "R^2")
colnames(diagnost.matrix) <- c("Quantitative Model", "Qualitative Model")

##############################################################################
##################### MODEL SELECTION: Cross Validation #######################
M1 <- Mquant     # models
M2 <- Mqual
nreps <- 8000                           # number of replications
ntot <- nrow(strikedata)                # total number of observations
ntrain <- 600                           # size of training set
ntest <- ntot-ntrain                    # size of test set
sse1 <- rep(NA, nreps)                  # sum-of-square errors for
sse2 <- rep(NA, nreps)                     #  each CV replication
Lambda <- rep(NA, nreps)       # likelihod ratio statistic for each replication
system.time({     # measures time taken to perform analysis
  for(ii in 1:nreps) {
    if(ii%%400 == 0) message("ii = ", ii)  # progress indicator
    # randomly select training observations
    train.ind <- sample(ntot, ntrain) # training observations
    # predict from training observations
    M1.cv <- update(M1, subset = train.ind)
    M2.cv <- update(M2, subset = train.ind)
    # test models with trained estimates
    M1.res <- strikedata$Strike[-train.ind] -
      predict(M1.cv, newdata = strikedata[-train.ind,])
    M2.res <- strikedata$Strike[-train.ind] -
      predict(M2.cv, newdata = strikedata[-train.ind,])
    # total sum of square errors (applying exp to return to original scale)
    sse1[ii] <- sum((exp(M1.res))^2)
    sse2[ii] <- sum(exp((M2.res))^2)
    # testing likelihood ratio
    M1.sigma <- sqrt(sum(resid(M1.cv)^2)/ntrain) # MLE of sigma
    M2.sigma <- sqrt(sum(resid(M2.cv)^2)/ntrain)
    Lambda[ii] <- sum(dnorm(M1.res, mean = 0, sd = M1.sigma, log = TRUE))
    Lambda[ii] <- Lambda[ii] - sum(dnorm(M2.res, mean = 0, sd = M2.sigma,
                                         log = TRUE))
  }})

# SSE.list <- c(SSE1 = mean(sse1), SSE2 = mean(sse2)) # not used right now

####### FIGURE 8: Cross validation with Histogram #######

par(mfrow = c(1,2), mar = c(4.5, 4.5, .1, .1))     # set up graph frame
# boxplots:
boxplot(x = list(sse1, sse2), names = c("Quantitative","Qualitative"), cex = .7,
        ylab = expression(SS[err]^{test}), col = c("yellow", "orange"))
```

```
# histogram:
hist(Lambda, breaks = 50, freq = FALSE, xlab = expression(Lambda^{test}),
     main = "", cex = .7)
abline(v = mean(Lambda), col = "red") # overlays average Lambda value
legend("topleft", legend = c(paste("mean(Lambda) = ",
                                   as.character(round(mean(Lambda),2)))),
       cex = 0.8, text.col = "red")  # legend indicating Lambda

##########################################################################
####### 2.1: Paramater Estimates and Confidence Intervals Of Final Model #######

Mqual.coefdata <- summary(Mqual)$coefficients # Pulls estimates, sigmas, F-stats
                                              # and p-values from summary
final.estimates <- Mqual.coefdata[,1]      # estimates
final.sigmas <- Mqual.coefdata[,2]         # sigmas
final.CIs <- paste("[", round(final.estimates - 1.96 * abs(final.sigmas), 3),
                   ",", round(final.estimates + 1.96 * abs(final.sigmas), 3),
                   "]")                    # creates 95% confidence intervals
                                            # for all betas
final.estimates <- round(final.estimates, 3)  # rounds estimates for output
final.matrix <- data.frame(as.numeric(final.estimates),final.CIs,
                           stringsAsFactors = FALSE)  # arrange for output
colnames(final.matrix) <- c("Estimate", "95% Confidence Interval")
rownames(final.matrix) <- rownames(Mqualtest.coefdata)

##########################################################################
################################### FIGURES ########################################


   #### FIGURE 1 & 2: Pairs Plots
# Figure 1 before log(Strike), Figure 2 after
pairs(~ Strike + Year + Infl + Unemp + Demo + Centr + Dens, data = strikedata)

   #### FIGURE 3 & 7: Residual Fit and Histogram
M <- Mquant # Mquant for Figure 3, Mqual for Figure 7.

M.stand.res <- resid(M)/summary(M)$sigma       # Standardized Residuals
M.stud.res <- M.stand.res/sqrt(1-hatvalues(M))  # Studentized Residuals
M.press <- resid(M)/(1-hatvalues(M))            # Press Residuals
M.dffits <- dffits(M)                           # DFFITS residuals

cex = 0.9 # dot size
par(mfrow = c(1,2), mar = c(4,4,2,2))           # sets up graph frame
plot(x = predict(M), y = rep(0,length(predict(M))), type = "n", # empty plot
     ylim = range(M.stand.res,M.stud.res,M.press,M.dffits),      # for all the
     xlab = "Predicted Values", ylab = "Residuals")              # points
segments(x0 = predict(M),
         y0 = pmin(M.stand.res, M.stud.res, M.press, M.dffits), # lines between
         y1 = pmax(M.stand.res, M.stud.res, M.press, M.dffits),  # residuals
         lty = 2)
points(predict(M), M.stand.res, pch = 21, bg = "black", cex = cex) # draws
points(predict(M), M.stud.res, pch = 22, bg = "blue", cex = cex)    # the
points(predict(M), M.press, pch = 23, bg = "red", cex = cex)        # points
points(predict(M), M.dffits , pch = 24, bg = "orange", cex = cex)
legend("topright", legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
       pch = c(21,22,23,24), pt.bg = c("black", "blue", "red", "orange"),
       title = "Residual Type:", cex = 0.7, pt.cex = 1)            # legend
hist.resid <- M.stand.res # Change residual to be used for Histogram
hist(hist.resid, breaks = 50, freq = F, cex.axis = 0.8,
     xlab = "Standardized Residuals", main = "")  # Histogram
curve(dnorm(x),col = "red", add = T)               # Overlay Normal curve
legend("topright", legend = c("Normal PDF"), lty = 1, # Legend for Normal curve
       col = "red", cex = 0.7)


   #### FIGURE 4: Linear Relationships with log(Strike)
M.strike.year <- lm(Strike ~ Year, data = strikedata)   # Linear regressions to
M.strike.infl <- lm(Strike ~ Infl, data = strikedata)   # plot the linear
M.strike.unemp <- lm(Strike ~ Unemp, data = strikedata) # relationships on top
M.strike.centr <- lm(Strike ~ Centr, data = strikedata) # of the scatter plots
M.strike.dens <- lm(Strike ~ Dens, data = strikedata)

cex4 = 1.4
par(mfrow = c(3,2), mar = c(4,4,2,2))                    # set up the graph frame
plot(strikedata$Year, strikedata$Strike, xlab = "Year",    # plots vvvv
     ylab = "log Strike days", cex.axis = cex4, cex.lab = cex4)
abline(M.strike.year , col = "red") # overlays the linear regression line
plot(strikedata$Infl, strikedata$Strike, xlab = "Inflation",
     ylab = "log Strike days", cex.axis = cex4, cex.lab = cex4)
abline(M.strike.infl, col = "red")  # overlays the linear regression line
plot(strikedata$Unemp, strikedata$Strike, xlab = "Unemployment",
     ylab = "log Strike days", cex.axis = cex4, cex.lab = cex4)
abline(M.strike.unemp, col = "red") # overlays the linear regression line
plot(strikedata$Centr, strikedata$Strike, xlab = "Union Centralization",
     ylab = "log Strike days", cex.axis = cex4, cex.lab = cex4)
abline(M.strike.centr, col = "red") # overlays the linear regression line
plot(strikedata$Dens, strikedata$Strike, xlab = "Union Density",
     ylab = "log Strike days", cex.axis = cex4, cex.lab = cex4)
abline(M.strike.dens, col = "red")  # overlays the linear regression line

    #### FIGURE 5: Linear Relationships between co-variates
M.year.infl <- lm(Infl ~ Year, data = strikedata)   # similar as above, linear
M.year.unemp <- lm(Unemp ~ Year, data = strikedata) # regressions to plot
M.year.dens <- lm(Dens ~ Year, data = strikedata)   # over the scatter plots
M.centr.dens <- lm(Centr ~ Dens, data = strikedata)

cex5 = 1.3
par(mfrow = c(2,2), mar = c(4,4,2,2))          # sets up the graph frame
plot(strikedata$Year, strikedata$Infl, xlab = "Year", ylab = "Inflation(%)",
     cex.axis = cex5, cex.lab = cex5)
abline(M.year.infl , col = "red") # overlays the linear regression line
plot(strikedata$Year, strikedata$Unemp, xlab = "Year", ylab = "Unemployment(%)",
     cex.axis = cex5, cex.lab = cex5)
abline(M.year.unemp, col = "red") # overlays the linear regression line
plot(strikedata$Year, strikedata$Dens, xlab = "Year", ylab = "Union Density(%)",
     cex.axis = cex5, cex.lab = cex5)
abline(M.year.dens, col = "red")  # overlays the linear regression line
plot(strikedata$Dens, strikedata$Centr, xlab = "Union Density(%)",
     ylab = "Union Centralization", cex.axis = cex5, cex.lab = cex5)
abline(M.centr.dens, col = "red") # overlays the linear regression line
```

```


    #### FIGURE 6: Leverage vs. Influence graphs
Mquant.h <- hatvalues(Mquant)                        # Hat values (leverages)
Mqual.h <- hatvalues(Mqual)

Mquant.h.bar <- mean(Mquant.h)                       # Average leverages
Mqual.h.bar <- mean(Mqual.h)

Mquant.highlev <- (Mquant.h >= 2*Mquant.h.bar)    # indices that are more than
Mqual.highlev <- (Mqual.h >= 2*Mqual.h.bar)        # twice the average leverage

Mquant.cookD <- cooks.distance(Mquant)              # cook's distance or cook's
Mqual.cookD <- cooks.distance(Mqual)                 # influence measure

Mquant.high.cookD <- (Mquant.cookD >=         # indices of top 15 influence obs
                        quantile(Mquant.cookD, probs = (610/625)))
Mqual.high.cookD <- (Mqual.cookD >=           # indices of top 15 influence obs
                       quantile(Mqual.cookD, probs = (610/625)))

### Figure starts here:
clrs <- rep("black", len = nrow(strikedata))    # empty colors vector
clrs[Mquant.highlev] <- "blue"                  # high leverage in blue
clrs[Mquant.high.cookD] <- "red"                # high influence in red
par(mfrow = c(2,1), mar = c(4,4,2,4), xpd = T)  # set up the graph frame
                                          # xpd = T allows legend outside of plot
plot(Mquant.h, Mquant.cookD, xlab = "Leverage", ylab = "Cook's Distance",
     main = "Quantitative Model", pch = 21, bg = clrs) # plot
abline(v = 2*Mquant.h.bar, col = "grey", lty = 2)   # add 2*mean(leverage) line
legend(x = 0.25, y = 0.061, legend = c("Top 15 Influence Observations",
                                       "Leverage > 2*mean(Leverage)"),
       cex = 0.6, pt.cex = 1.2,                 # legend
       pch = 19, col =c("red", "blue"))
clrs2 <- rep("black", len = nrow(strikedata))   # empty colors vector
clrs2[Mqual.highlev] <- "blue"                  # high leverage in blue
clrs2[Mqual.high.cookD] <- "red"                # high influence in red
plot(Mqual.h, Mqual.cookD, xlab = "Leverage", ylab = "Cook's Distance",
     main = "Qualitative Model", pch = 21, bg = clrs2) # plot
abline(v = 2*Mqual.h.bar, col = "grey", lty = 2) # add 2*mean(leverage) line
legend(x = 0.14, y = 0.06, legend = c("Top 15 Influence Observations",
                                      "Leverage > 2*mean(Leverage)"),
       cex = 0.6, pt.cex = 1.2,                 # legend
       pch = 19, col =c("red", "blue"))

################################################################################
################################ APPENDIX A ####################################

#####  IDENTIFYING OUTLIERS

strike_outliers <- strikedata[(strikedata$Strike > 5000),] # finds indices of
outliers_indices <- as.numeric(row.names(strike_outliers))   # outliers

##### FIGURE 9
M <- Mfwd

```

```r
H <- hatvalues(M)                         # Hat values/leverages
M.res <- resid(M)                         # Residuals
M.y.hat <- predict(M)                     # Predicted Values
M.sigma.hat <- summary(M)$sigma           # Sigma hat
M.stand.res <- M.res/M.sigma.hat          # Standardized Residuals
M.stud.res <- M.stand.res/sqrt(1-H)       # Studentized Residuals

color.index <- (strikedata$Strike <= 302.3)       # Setting colors for obs.
pt.col <- rep("black", length(strikedata$Strike)) # with strikes <= mean
pt.col[color.index] <- "red"
pt.col[outliers_indices] <- "blue"                # Blue for 3 large outliers

par(mfrow = c(1,2), mar = c(4,4,2,2))             # Set up graph frame
plot(x = M.y.hat, y = M.stand.res, col = pt.col, xlab = "Predicted Values",
     ylab = "Standardized Residuals")             # plot
legend(x = -150, y = 8, legend = c("Strikes <= 302.3", "Strikes > 302.3",
                                   "Strikes >= 5000"),
       cex = 0.6, y.intersp = 2, xjust = 0.1, pt.cex = 1.2, pch = 19,
       col = c("red", "black","blue"))            # legend
hist(M.stand.res, breaks = 50, freq = FALSE, cex.axis = 0.8,
     xlab = "Standardized Residuals", main = "")  # histogram of residuals
curve(dnorm(x),col = "red", add = TRUE)           # normal curve overlay
legend("topright", legend = c("Normal PDF"), lty = 1, # legend
       col = "red", cex = 0.7)

##### FIGURE 10
cex = 0.9        # point size
par(mfrow = c(1,2), mar = c(4,4,2,2))    # set up graph frame
plot(x = M.y.hat, y= rep(0,length(M.y.hat)), type = "n", # empty plot to put the
     ylim = range(M.stand.res, M.stud.res),               # points onto
     xlab = "Predicted Values", ylab = "Residuals")
segments(x0 = M.y.hat,                                    # lines between the
         y0 = pmin(M.stand.res, M.stud.res),               # residuals
         y1 = pmax(M.stand.res, M.stud.res),
         lty = 2)
points(M.y.hat, M.stand.res, pch = 23, bg = "blue", cex = cex)   # points
points(M.y.hat, M.stud.res, pch = 21, bg = "orange", cex = cex)
legend("topright", legend = c("Standardized", "Studentized"),    # legend
        pch = c(21,23), pt.bg = c("blue", "orange"),
        title = "Residual Type:", cex = 0.7, pt.cex = 1)
hist(M.stand.res, breaks = 50, freq = FALSE, cex.axis = 0.8,
     xlab = "Standardized Residuals", main = "")  # histogram of residuals
curve(dnorm(x),col = "red", add = TRUE)           # normal curve overlay
legend("topright", legend = c("Normal PDF"), lty = 1, # legend
       col = "red", cex = 0.7)
```