STAT 331 Final Project

Fall 2016

Tim Huang, 20520988
Ayush Venkatesh, 20578894
Ian Waudby-Smith, 20526823

Summary

The objective of this report is to investigate a dataset containing strike data belonging to the Organization for Economic Co-operation and Development (OECD) during the postwar period 1951-1985. More specifically, we are interested in the relationship between several macroeconomic variables and strike activity. We fit a series of regression models using techniques including, but not limited to multiplicative error models, variable transformation, forward variable selection and stepwise variable selection. We arrive at two strong candidate multiplicative error models, the first of which is quite simple and interpretable, while the other is more complex. Depending on the measurement used to compare predictive performance, each model can seem to outperform the other. We ultimately choose the simple model because in addition to being more interpretable, it appears to be more robust and does not have substantially less predictive ability than the more complex model.

Model Selection

We note that the dataset includes a column for the numeric value, *Year*. If a model is fit using *Year*, the intercept of that model will have an interpretation with *Year = 0*. As we are primarily interested in analyzing strike activity during the period of 1951 - 1985 (or other years not veering too far from this range), we let the column $Year1951 = Year - 1951$. This way, when interpreting the intercept of our model, we can think of this being when $Year = 1951$.

We denote model 1 by *M1* and let

$$M1 : Strike \sim Country + Year1951 + Unemp + Infl + Demo + Centr + Dens$$

We create the following diagnostic plots to assess the model fit:

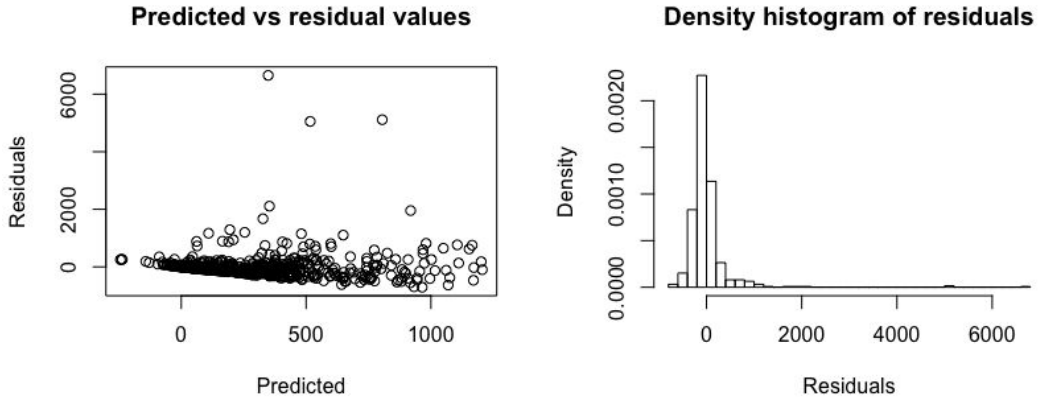

Figure 1: Diagnostic plots for model 1

As we can see in figure 1, some regression assumptions are being severely violated. In both plots we see the existence of large outliers. In investigating these outliers, we discover that many of them correspond to “general” or “mass” strikes in history - a strike in which a substantial percentage of the total labour force in a region or

country participates. As such, we add a categorical covariate, *GenStrike* to indicate whether a country experienced a general/mass strike in a given year.

Additionally, in the first plot in figure 1, we can see that the variance in the errors seems to increase with the predicted value. As such, a multiplicative error model is worth investigating.

We denote model 2 by *M2* and let

$$M2 : \log(Strike + 1) \sim Country + Year1951 + Unemp + Infl + Demo + Centr + Dens + GenStrike$$

Note: $\log(Strike + 1)$ is used instead of $\log(Strike)$ because there exist data points with Strike = 0. This allows us to avoid encountering -Inf values in R. Additionally, data points with 0 strike activity will have values of 0 in the multiplicative error model.

Since we have changed the scale of our response variate, it is reasonable to check whether *GenStrike* is still an "important" covariate (since the general/mass strike outliers were found on the original scale). We consider *M2* and $M2^{(-GenStrike)}$ which corresponds to M2 without *GenStrike* as a covariate. The result of an anova test (p-value = 0.0006239) between *M2* and $M2^{(-GenStrike)}$ suggests that *GenStrike* is a potentially "important" covariate.

It also seems to be the case that the outliers previously found in *M1* are not causing the same issues in model fit when the response is taken to be on the log scale. Therefore, although these data points are considered somewhat "different" (i.e. general strikes), we do not have the same inclination to remove them as it seems plausible that the log-scale model, *M2* may be able to account for them.

Consider the following diagnostic plots for model 2:

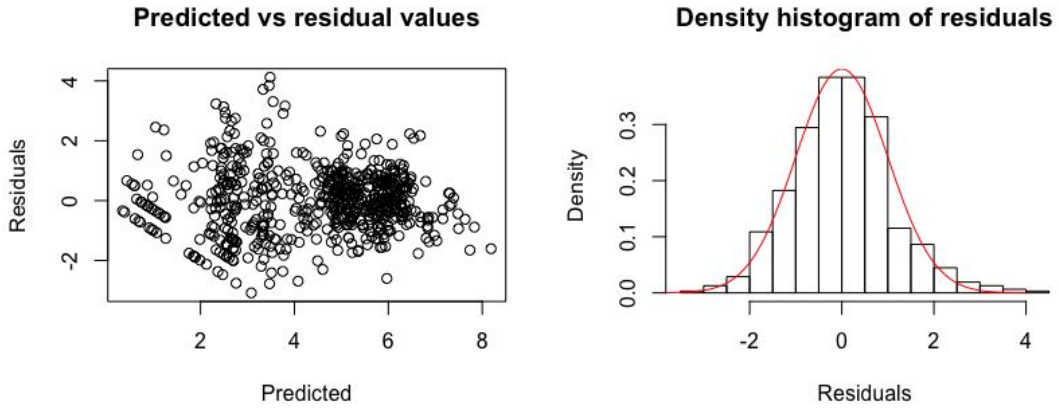

Figure 2: Diagnostic plots for model 2

There is still some evidence that assumptions are being violated, i.e. there is a noticeable pattern in the "predicted vs residuals" plot and the histogram of residuals is slightly right-skewed. However, model 2 is an improvement from model 1 and provides a good starting point for investigating new models.

We observe that the covariate *Centr* is the same for each *Country* regardless of *Year*, and thus it is a linear combination of *Country*. Therefore, we remove *Centr* from the model.

We also observe that in *M2*, the adjusted R-squared is 0.6889, and in a model where we include *Country* as the only covariate, the adjusted R-squared is 0.652. This suggests that the covariate *Country* is responsible for "explaining" a large portion of the variability in the response. Now we would like to see which covariates we can have in addition to *Country* that are significant but not redundant. As such, it is natural to perform forward selection starting with the following model:

$$\log(Strike + 1) \sim Country$$

Forward selection with *M2* as the upper model yields the following model:

$$\log(Strike + 1) \sim Country + Dens + Infl + Year1951 + GenStrike$$

However, we choose to use a variant of this model that does not include *GenStrike*. The reason for this is that *GenStrike* only makes up 5 of 625 observations and so the corresponding β-coefficient may be poorly estimated. Consider the resulting model which we will denote as *M3*:

$$M3 : \log(Strike + 1) \sim Country + Dens + Infl + Year1951$$

Stepwise selection

From our previous work in assessing whether regression assumptions are being met, we note that log(Strike + 1) is a more appropriate response variable than Strike. It is now worth considering whether a model on this scale including interaction terms can more effectively explain the data. Given the large number of possible interaction models, we perform stepwise selection as follows:

Consider model 0 (*M0*):

$$M0 : \log(Strike + 1) \sim 1$$

Now consider the following model (*Mmix*):

$$Mmix : \log(Strike + 1) \sim (Centr + Year1951 + Unemp + Infl + Demo + Dens)^2 + Country + GenStrike$$

Note that interaction effects with the covariate *Country* were not included as this would require estimating a large number of parameters for each interaction and could possibly lead to overfitting. Additionally, interaction effects with the covariate *GenStrike* were not included since general/mass strikes only account for 5 of the 625 data points so including all interaction effects with *GenStrike* could lead to poorly estimated coefficients. However, we

chose to leave *GenStrike* in *Mmix* in case it has an adequate amount of predictive power (as decided by stepwise selection).

We perform stepwise selection (in both directions) with *M3* as the starting model, *M0* as the lower bound and *Mmix* as the upper bound. Stepwise selection in both directions was chosen since we have already seen that *M3* seems to be a satisfactory, yet simple model. As such it is desirable to use it as a "starting point" for further variable selection to obtain a model that can potentially make better predictions. The algorithm yields the following model which we will refer to as *Mstep*:

$$Mstep : \log(Strike + 1) \sim Country + Dens + GenStrike + Infl + Year1951 + Infl : Year1951 + Dens : Infl$$

Given the output to our two variable selection algorithms, we decide to investigate M3 and Mstep in depth. For the remainder of this article, we refer to M3 and Mstep as M(a) and M(b) respectively for clarity.

Model Diagnostics

Residual Plots

The four types of residuals we will analyze are standardized residuals, studentized residuals, PRESS, and DFFITS. For all the residual vs predicted value plots in this section, we will show standardized residuals as black, studentized residuals as orange, PRESS as red, and DFFITS as blue. Also, the histogram of the studentized residuals are compared with the standard normal density function shown as a red curve. The studentized residuals were chosen to compare with the standard normal distribution because if the model were true, the residuals should follow $N(0, \sigma^2(1-H))$ distribution where H is the hat matrix. The studentized residual is the standardized residual divided by $\sqrt{(1-h_i)}$ which should bring the studentized residual closer to the standard normal distribution.

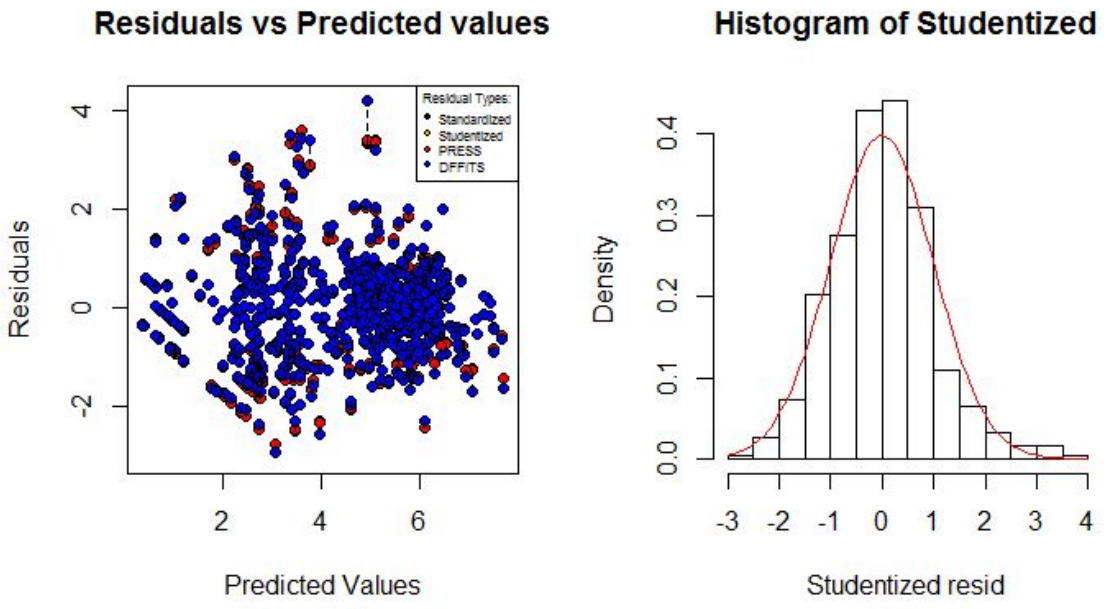

Figure 3: Residual Plot and Histogram of Studentized Residuals for M(a)

As seen in figure 3, there seems to be little difference between standardized, studentized, and PRESS residuals. Since the hat values (or leverage) of each observation is very low (close to zero), the value $\sqrt{(1-h_i)}$ would be close to 1. It can be shown that $\hat{\sigma}$ of the residuals for M(a) is 1.1089 which is also fairly close to 1. This means that the denominator of the standardized, studentized, and PRESS residuals are all close to one which allows the three values to be similar for each observation. This is why it is hard to distinguish the location of the 3 residuals on the plot.

The DFFITS are fairly similar to the other residual types in this model. Also, the distribution of the studentized residuals appears to be very similar to a standard normal distribution based on the histogram.

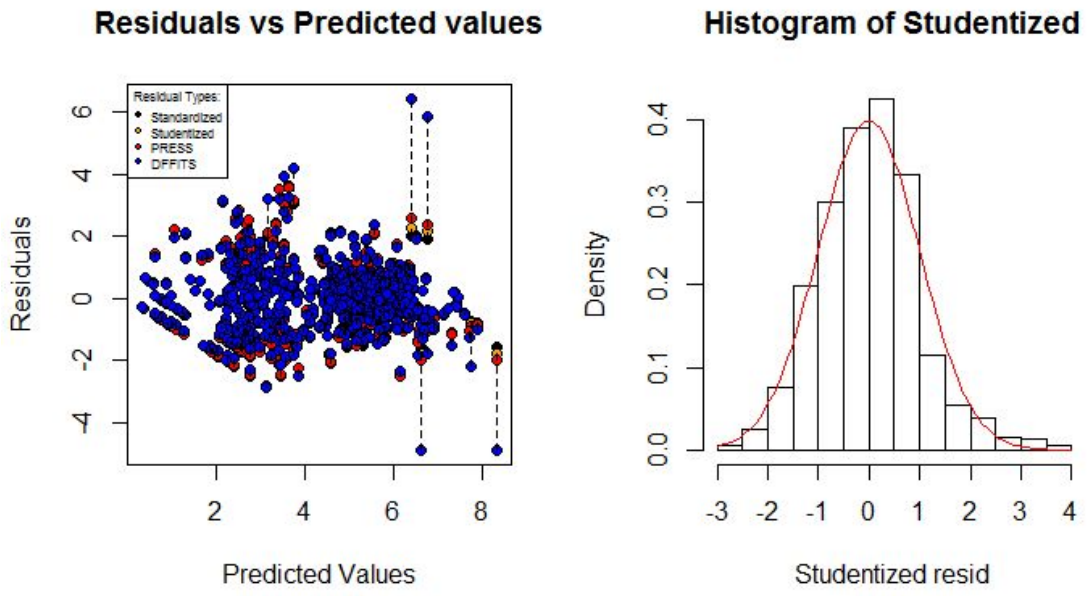

Figure 4: Residual Plot and Histogram of Studentized Residuals for M(b)

The residual plot for the other model M(b) is slightly different than the one for M(a). As shown in figure 4, there are four observations that have DFFITS values very different than the other residual values. This would indicate that there are four observations that have a large influence on the estimated β-coefficients of the model. A notable difference between this model and M(a) is the "General Strike" covariate. Since the number of general strike observations are very limited, the impact of these observations on the corresponding covariate can result in large DFFITS values.

The $\hat{\sigma}$ of the residuals for M(b) is 1.0966 which is close to 1. So based on the same reasoning for M(a), the majority of the observations with low hat values would have very similar standardized, studentized, and PRESS residuals. Similar to M(a), the histogram of the studentized residuals for M(b) appear to closely follow the standard normal distribution.

Influence and Leverage Measures

The following plots show the Leverage of the data points vs Cook's D Statistic (or Cook's influence). Observations with high influence are shaded green and observations with high leverage are shaded orange. Data points with high leverage are defined to be those with leverage more than two times the mean leverage. Data points with high influence are defined to be those with Cook's influence greater than 80% of the maximum Cook's influence among the observations.

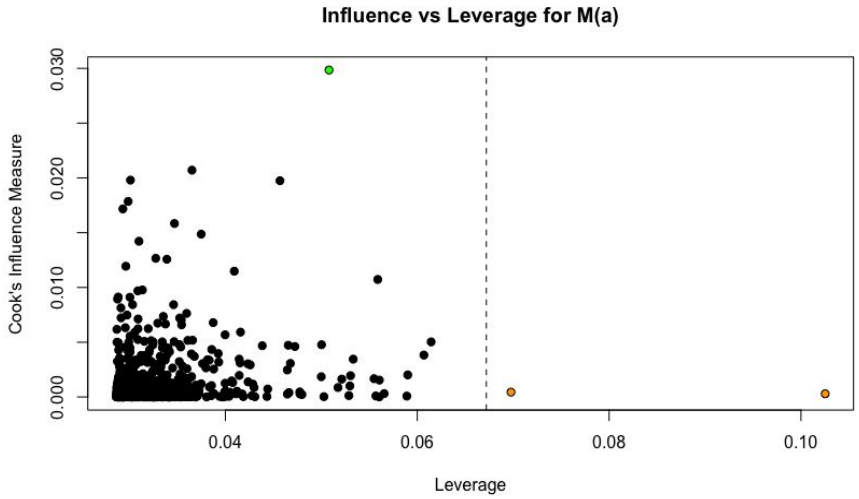

Figure 5: Cook's Influence vs Leverage for M(a)

As seen in figure 5, only the data point with maximum influence is identified as high influence. This means that the influence of the observation with the second most influence is not that similar to the observation with the most influence.

Also, it is important to note that there are only two observations that have high leverage and both of them have very low influence. This is a good sign as it means that the two observations that have the most potential of changing the coefficient parameters of the model (if they had high influence) are not very influential. In other words, if those two observations were removed, there would be very little change in the estimators of the model. If the high leverage observations were also to have high influence, then removing the observations would have more impact on the model than removing other observations with the same influence but lower leverage. As it is the case in this model, it is important for high leverage observations to have low influence.

The figure below demonstrates the Cook's influence vs leverage for M(b) using the same colouring pattern as in figure 5:

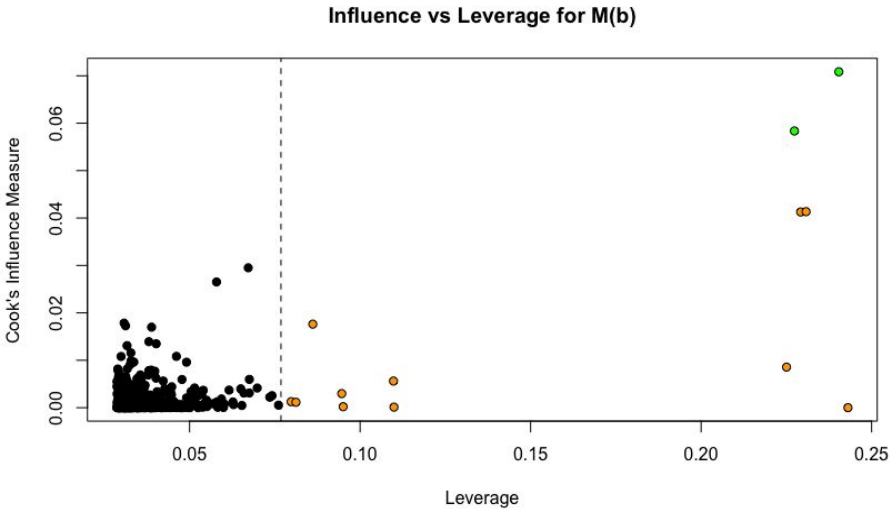

Figure 6: Cook's Influence vs Leverage for M(b)

Unlike figure 5, the plot in figure 6 contains many observations with leverage more than two times the average leverage. In fact, six of the high leverage observations have significantly more leverage than any of the other observations.

Through further investigation, we found that all five observations that were flagged as general/mass strikes were among those six observations. Based on the way that this model was created, it would make sense that the five general strike observations have high leverage. This is because with only five observations, every single observation flagged as a general strike would have more potential to affect the corresponding β-coefficient than observations not flagged as general strikes.

There is still a possibility that this model can provide fair predictability for observations of year and country that do not have a general strike. The observation within the top six leverage values that isn't a general strike observation is the one that has a Cook's influence close to zero. Thus, similar to the high leverage observations in figure 5, this observation has very little effect on the estimators of the model (see appendix for details).

Cross Validation

To assess the predictive power in M(a) and M(b) we perform cross-validation with a training set of 580 data points (92.8% of the data) and a testing set of 45 data points (7.2% of the data). This particular training set size was chosen since smaller sizes tended to result in rank-deficient fits. We plot the histogram of the Λ^{test} statistic values as well as the box plots for the sum of square errors as follows:

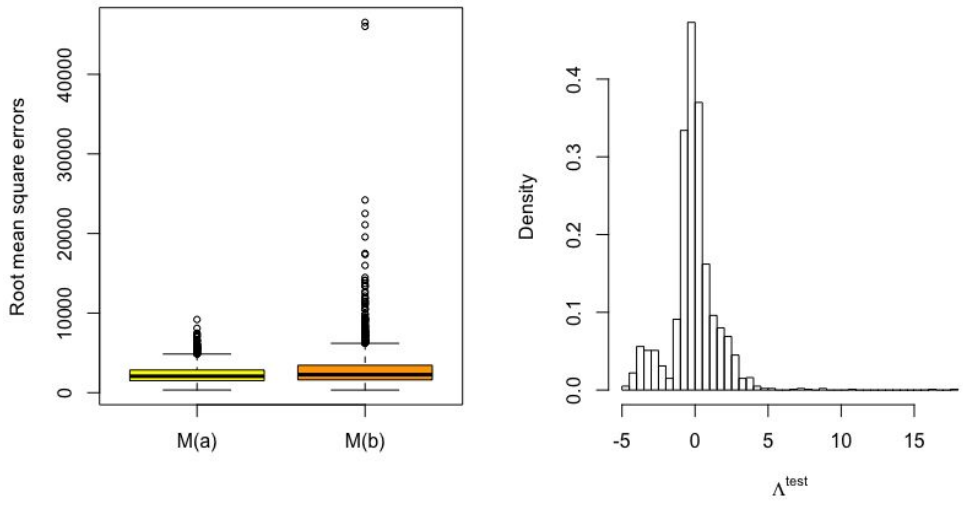

Figure 7: Cross validation; left: box plot for root mean errors, right: likelihood ratio test statistic

In figure 7, note that higher values of Λ^{test} favour M(a). The boxplots represent the distribution of root mean square errors (RMSE), computed as follows: $RMSE = \sqrt{SSE/n_{test}}$ where SSE is the sum of square errors either from M(a) or M(b) and n_{test} is the number of test data points used. RMSE is the measure of choice for comparing prediction errors as it is in units of strike-hours and thus is measuring the error of the predictions our model is ultimately meant to make (that is, on the original scale). The mean RMSE values are computed to be 2324.92 and 2923.77 for M(a) and M(b) respectively. The mean Λ^{test} value is computed to be -0.16. As such, we see that the RMSE values would suggest that M(a) outperforms M(b) on the original scale, and the likelihood ratio statistic suggests that M(b) outperforms M(a) on the log scale.

Final model selection

Both M(a) and M(b) prove to be strong candidates for modelling the data, but they are not without their own shortcomings. M(b) is favoured in making predictions on the log scale as suggested by the likelihood ratio statistic. However, M(a) is a considerably simpler and more interpretable model and seems to make slightly better predictions on the original scale of strike-hours.

Additionally, we see large difference between studentized and standardized residuals for some data points with respect to M(b). M(b) also seems to have a higher number of high-leverage and high-influence data points than M(a). This suggests that M(a) is, in a sense, a more “robust” model.

In summary, M(a) is the simpler, more interpretable model as it is only a multiplicative error model, whereas M(b) contains multiple interaction terms. It also outperforms M(b) on the original scale of strike-hours. As such,

we choose M(a) as the final model. See the appendix for a table (Table 1) containing parameter estimates and 95% confidence intervals.

Discussion

The country in which the strikes take place appears to be one of the most important factors for estimating strike activity, both in terms of its statistical significance (p-values) but also in how drastically different the strike activity can be estimated to be in two countries with all covariates equal. For instance, holding all other covariates equal, the expected strike activity (shifted up by 1) for Australia is $e^{4.086} = 59.5$ times the expected strike activity + 1 for Switzerland (note that the coefficient for Switzerland is approximately -4.0856 when looking at log(Strike + 1) so this translates to a factor of $e^{-4.086}$ on the original scale). This is not extremely surprising as different countries have unique legislation and culture that may heavily influence how much strike activity there is in a given year.

In M(a), trade union density and inflation appear to be macroeconomic variables with statistically significant relationships with strike activity. Each unit increase in inflation increases the expected strike activity + 1 by 5.8% (since $e^{0.0563} = 1.0579$). Each unit increase in trade union density increases expected strike activity + 1 by 3.6% (since $e^{0.03554} = 1.0362$).

The only parameters left in the final model with high p-values are certain categories of our categorical variable (Country) which is overall a significant covariate. The reason for not including more covariates with high p-values is mainly due to the fact that we allowed for forward selection to select many of the important covariates. However, those that were selected have very small p-values and are thus important parameters for explaining strike activity on the log scale.

When we first found that in certain years, some countries experienced a mass strike, we decided to treat this event as a categorical covariate to see if its inclusion in a model would have added predictive power or would explain some of the variation in the response. Since mass strikes were such rare events in the 1951 - 1985 period, there were not enough data points to accurately measure its effect on the response. As such, it may have been reasonable to remove these points and treat them as outliers.

There are certainly some regression assumptions being violated in the final model, M(a). Firstly, as seen in figure 3, the studentized residuals do seem to be slightly right-skewed, so the assumption of normally distributed errors may be violated. Additionally, there seems to be a noticeable pattern in the residual plot. Namely, if we look at figure 3, we can see a negative linear trend in the residual points in the region $predicted \in (0, 4)$. We also notice that the variance in the residuals seems to be slightly higher in the $predicted \in (0, 4)$ range than in the $predicted \in (5, 7)$ where there seems to be a cluster of points.

None of the violations described here are enormous and are considerably less severe than the assumption violations seen in the simple additive error model, *M1* (as seen in figure 1). Given the inherent complexity in modelling economic variables, the model proposed here is a good starting point for further investigations into the macroeconomic variables associated with strike activity despite the assumption violations found.

Appendix

Table 1: Parameter estimates and their confidence intervals for M(a)

	Estimate	**95% CI**	
		2.5 %	**97.5 %**
(Intercept)	3.9249	3.14	4.71
CountryAustria	-3.6562	-4.19	-3.12
CountryBelgium	-0.4956	-1.04	0.05
CountryCanada	1.4157	0.80	2.03
CountryDenmark	-2.5876	-3.14	-2.04
CountryFinland	-0.7002	-1.22	-0.18
CountryFrance	0.5473	-0.15	1.25
CountryGermany	-2.3716	-2.93	-1.81
CountryIreland	0.2957	-0.23	0.82
CountryItaly	1.2702	0.72	1.82
CountryJapan	-0.4788	-1.06	0.11
CountryNetherlands	-2.5773	-3.13	-2.02
CountryNewZealand	-0.4490	-0.99	0.09
CountryNorway	-2.5324	-3.06	-2.01
CountrySweden	-3.7781	-4.38	-3.18
CountrySwitzerland	-4.0860	-4.66	-3.51
CountryUnitedKingdom	-0.1413	-0.67	0.39
CountryUnitedStates	1.2378	0.59	1.88
Year1951	-0.0235	-0.03	-0.01
Infl	0.0563	0.03	0.08
Dens	0.0355	0.02	0.05

R Code:

```
OECD.df$Year1951 = OECD.df$Year - 1951
OECD.df$logStrikeShifted <- log(OECD.df$Strike + 1)

# Adding General/Mass Strikes:

OECD.df$GenStrike <- NA
OECD.df[OECD.df$Country == "Finland" & OECD.df$Year == 1956,]$GenStrike <- TRUE
OECD.df[OECD.df$Country == "Belgium" & OECD.df$Year == 1960,]$GenStrike <- TRUE
OECD.df[OECD.df$Country == "France" & OECD.df$Year == 1968,]$GenStrike <- TRUE
OECD.df[OECD.df$Country == "Canada" & OECD.df$Year == 1972,]$GenStrike <- TRUE
OECD.df[OECD.df$Country == "Ireland" & OECD.df$Year == 1974,]$GenStrike <- TRUE
OECD.df[is.na(OECD.df$GenStrike),]$GenStrike <- FALSE

M.mult.err <- lm(logStrikeShifted ~ Country + Year1951 + Unemp + Infl +
                   Demo + Centr + Dens, data=OECD.df)

M <- lm(Strike ~ I(Year - 1951) + Unemp + Infl +
          Demo + Country + Dens, data = OECD.df)

M2 <- lm(logStrikeShifted ~ Country + I(Year - 1951) + Unemp + Infl + Demo +
Centr + Dens + GenStrike,
         data=OECD.df)

Ma <- lm(logStrikeShifted ~ Country + I(Year - 1951) + Infl + Dens, data=OECD.df)

# Model 1 diagnostics

plot(predict(M, OECD.df),residuals(M))

par(mfrow=c(1, 2))
plot(predict(M), residuals(M), xlab = "Predicted",
     ylab="Residuals", main="Predicted vs residual values")
hist(residuals(M), breaks=30, xlab ="Residuals",
     ylab="Density", prob=TRUE, main="Density histogram of residuals")

# Model 2 diagnostics

par(mfrow=c(1, 2))
```

```
plot(predict(M2), residuals(M2), xlab = "Predicted",
     ylab="Residuals", main="Predicted vs residual values")
hist(residuals(M2), breaks=25, xlab ="Residuals",
     ylab="Density", prob=TRUE, main="Density histogram of residuals")
curve(dnorm, from=-4, to=4, add = TRUE, col="red")
```

Automated Selection:

```
M0 = lm(logStrike ~ Country, data = OECD.df)
Mmix = lm(logStrike ~ (Year1951 + Unemp + Infl + Demo + Dens)^2 + Country +
GenStrike, data = OECD.df)
M.no.interact = lm(logStrike ~ Year1951 + Unemp + Infl + Demo + Dens + Country,
data = OECD.df)

Ma = step(object = M0, scope = list(lower = M0, upper = M.no.interact),
            direction = "forward", trace = FALSE)
formula(Ma) #logStrike ~ Country + Dens + Infl + Year1951

Mb = step(object = Ma, scope = list(lower = M0, upper = Mmix),
             direction = "both", trace = FALSE)
formula(Mb) #logStrike ~ Country + Dens + Infl + Year1951 + GenStrike +
Infl:Year1951 + Dens:Infl
```

Residual Plots, Leverage, and Influence:

```
#hat value h is optional, can input to avoid repeated calculation
get.res.df <- function(M, h=numeric(0)) {

  n = nobs(M)
  sigma.hat = summary(M)$sigma
  if(length(h) == 0) h = hatvalues(M)

  # original residuals
  res = resid(M)

  # standardized residuals
  stan = res/sigma.hat

  # studentized residuals
```

```
  stud = stan/sqrt(1-h)

  # PRESS residuals
  press = res/(1-h)

  # DFFITS residuals
  dfts = dffits(M)

  # data frame to return
  res.df = data.frame(orig=res,    stan=stan, stud=stud,
                      press=press, dfts=dfts )

  return(res.df)
}

leverage.plots <- function(M) {
  # residuals vs predicted values
  y.hat = predict(M)
  sigma.hat = summary(M)$sigma

  # Computing leverage
  h = hatvalues(M)

  # Data frame of residuals
  res.df = get.res.df(M, h)

  p = length(coef(M))
  n = nobs(M)
  hbar = p/n #average leverage

  #standardize the resids so they are same at average leverage
  res.df$stud = res.df$stud*sqrt(1-hbar)
  res.df$press = res.df$press*(1-hbar)/sigma.hat
  res.df$dfts = res.df$dfts*(1-hbar)/sqrt(hbar)

  #plot predicted vs residuals
  par(mfrow = c(1,2))
  plot(y.hat, numeric(n), type = "n", ylim = range(res.df),
       xlab = "Predicted Values", ylab = "Residuals",
       main = "Residuals vs Predicted values")
  points(y.hat, res.df$stan, bg = "black", pch = 21)
  points(y.hat, res.df$stud, bg = "orange", pch = 21)
  points(y.hat, res.df$press, bg = "red", pch = 21)
```

```
  points(y.hat, res.df$dfts, bg = "blue", pch = 21)
  segments(x0 = y.hat, lty = 2,
                y0 = pmin(res.df$stan, res.df$stud, res.df$press, res.df$dfts),
                y1 = pmax(res.df$stan, res.df$stud, res.df$press, res.df$dfts))

  #histogram of standardized resid
  hist(res.df$stud, freq = FALSE, xlab = "Studentized resid", main = "Histogram
of Studentized")
  curve(dnorm(x,0,1), add = TRUE, col = "red")

  #plot leverage vs residuals
  par(mfrow = c(1,1))
  plot(h, numeric(n), type = "n", ylim = range(res.df),
       xlab = "Leverages", ylab = "Residuals",
       main = "Residuals vs Leverages")
  points(h, res.df$stan, bg = "black", pch = 21)
  points(h, res.df$stud, bg = "orange", pch = 21)
  points(h, res.df$press, bg = "red", pch = 21)
  points(h, res.df$dfts, bg = "blue", pch = 21)
  abline(v = hbar, col="black", lty = 2)
  segments(x0 = h, lty = 2,
           y0 = pmin(res.df$stan, res.df$stud, res.df$press, res.df$dfts),
           y1 = pmax(res.df$stan, res.df$stud, res.df$press, res.df$dfts))

  #cook's D statistics vs leverage
  D = cooks.distance(M)

  highD = max(D)*0.8
  infl.index = D > highD #most influential point
  lev.index = h > 2*hbar  #leverage more than 2x the average

  col = rep("black", len = n)
  col[lev.index] = "orange"
  col[infl.index] = "green"

  par(mfrow = c(1,1))
  plot(h, D, xlab = "Leverage", ylab = "Cook's Influence Measure",
       main = "Influence vs Leverage", pch = 21, bg = col)
  abline(v = 2*hbar, col="black", lty = 2)
}

leverage.plots(Ma)
leverage.plots(Mb)
```

```
#example of how to find which observations has high leverage
h = hatvalues(Ma)
OECD.df[(h > 0.065),] #only 2 observations with leverage more than 0.065 from
cook's influence plot

#effect of removing low influence and high leverage in M(b), refer to figure 6
h = hatvalues(Mb)
D = cooks.distance(Mb)
OECD.df[(h > 0.24),] #refer to figure 6 for h > 0.24, outputs index 36 and 176
D[c(36,176)] #index 36 has influence close to 0
temp = OECD.df[-36,]
Mtemp = lm(formula(Mb), data = temp)
coefs = cbind(coef(Mb),coef(Mtemp))
(coefs[,1] - coefs[,2])/coefs[,1]
```

```
> cbind(coef(Mstep),coef(Mtemp))
                          [,1]         [,2]
(Intercept)          70.257799287 70.236459840
CountryAustria       -3.570231372 -3.569822048
CountryBelgium       -0.570587358 -0.570517195
CountryCanada         1.215704416  1.215637112
CountryDenmark       -2.565654194 -2.565626176
CountryFinland       -0.814200119 -0.814311027
CountryFrance         0.364938512  0.364853413
CountryGermany       -2.447445150 -2.447466935
CountryIreland        0.160559525  0.160525110
CountryItaly          1.127181944  1.127124965
CountryJapan         -0.552520365 -0.552625733
CountryNetherlands   -2.640215587 -2.640253662
CountryNewZealand    -0.522465143 -0.522458549
CountryNorway        -2.512287352 -2.512257874
CountrySweden        -3.718169790 -3.718121653
CountrySwitzerland   -4.177022398 -4.177057453
CountryUnitedKingdom -0.210352670 -0.210369140
CountryUnitedStates   1.078320061  1.078244649
Dens                  0.024631645  0.024615046
Infl                 -3.335759755 -3.329646641
Year                 -0.033652123 -0.033640952
GenStrikeTRUE         1.697408939  1.697304233
Infl:Year             0.001699459  0.001696319
Dens:Infl             0.001041983  0.001044033
```

```
> (coefs[,1] - coefs[,2])/coefs[,1]
        (Intercept)       CountryAustria       CountryBelgium        CountryCanada       CountryDenmark
       3.037307e-04         1.146491e-04         1.229665e-04         5.536208e-05         1.092013e-05
     CountryFinland        CountryFrance       CountryGermany       CountryIreland         CountryItaly
      -1.362168e-04         2.331867e-04        -8.901241e-06         2.143473e-04         5.055013e-05
       CountryJapan   CountryNetherlands    CountryNewZealand       CountryNorway        CountrySweden
      -1.907037e-04        -1.442097e-05         1.262105e-05         1.173328e-05         1.294658e-05
  CountrySwitzerland CountryUnitedKingdom  CountryUnitedStates                 Dens                 Infl
      -8.392249e-06        -7.829244e-05         6.993481e-05         6.738991e-04         1.832600e-03
                Year        GenStrikeTRUE            Infl:Year            Dens:Infl
       3.319675e-04         6.168577e-05         1.847541e-03        -1.967513e-03
```

Cross Validation:

```
strike=read.csv("strikes_clean.csv")
strike$logStrikeShifted <- log(strike$Strike + 1)

#The two models
Mb<-lm(logStrikeShifted ~ Country + Dens + GenStrike + Infl + Year1951 +
Infl:Year1951 + Dens:Infl, data=strike)
Ma<-lm(logStrikeShifted ~ Country + Dens + Infl + Year1951 ,data=strike)

nreps <- 2e3    # number of replications
ntot <- nrow(strike) # total number of observations
ntrain <- 580 # size of training set
ntest <- ntot-ntrain # size of test set
sse_Ma <- rep(NA, nreps) # sum-of-square errors for each CV replication
sse_Mb <- rep(NA, nreps)
Lambda <- rep(NA, nreps) # likelihood ratio statistic for each replication
sd.test.Ma <- rep(NA, nreps)
sd.test.Mb <- rep(NA, nreps)
for(ii in 1:nreps) {

  trainSet=sample(ntot,ntrain)
  Ma_cv=update(Ma,subset=trainSet)
  Mb_cv=update(Mb,subset=trainSet)

  #Convert back to regular scale. These 3 variables below are for SSE
  regularStrike=strike$Strike[-trainSet]
  sigma.hat <- sd(strike$logStrikeShifted[-trainSet])
  regular_Ma_Strike=exp(predict(Ma_cv,newdata=strike[-trainSet,]) +
                          (sigma.hat^2)/2) - 1
  regular_Mb_Strike=exp(predict(Mb_cv,newdata=strike[-trainSet,]) +
                           (sigma.hat^2)/2) - 1

  #Calculate residuals on the log scale

resid_Mb=strike$logStrikeShifted[-trainSet]-(predict(Mb_cv,newdata=strike[-trainS
et,]))

resid_Ma=strike$logStrikeShifted[-trainSet]-(predict(Ma_cv,newdata=strike[-trainS
et,]))

  #The sum of squared errors:
  sse_Mb[ii]=sum((regularStrike-regular_Mb_Strike)^2)
```

```
  sse_Ma[ii]=sum((regularStrike-regular_Ma_Strike)^2)

  #Calculating MLE of sigma for both models:
  sigma_Mb=sqrt(sum(resid(Mb_cv)^2)/ntrain)
  sigma_Ma=sqrt(sum(resid(Ma_cv)^2)/ntrain)

  sd.test.Mb[ii] <- sqrt(sse_Mb[ii]/ntest)
  sd.test.Ma[ii] <- sqrt(sse_Ma[ii]/ntest)

  #Calculating lambda:
  Lambda[ii]=sum(dnorm(resid_Ma,mean=0,sd=sigma_Ma,log=TRUE))
  Lambda[ii]=Lambda[ii]-sum(dnorm(resid_Mb,mean=0,sd=sigma_Mb,log = TRUE))
}

#BoxPlot
boxplot(x = list(sd.test.Ma, sd.test.Mb), names = c("M(a)","M(b)"), cex = .7,
        ylab = "Root mean square errors", col = c("yellow", "orange"))

#Histogram of lambda
hist(Lambda, breaks = 50, freq = FALSE,
     xlab = expression(Lambda^{test}),
     main = "", cex = .7)

median(Lambda)

#Sum of squared errors
c(SSE1 = mean(sse_Ma), SSE2 = mean(sse_Mb))
```

```
summary(Ma)
Call:
lm(formula = logStrike ~ Country + Dens + Infl + Year1951, data = OECD.df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0806 -0.6889 -0.0128  0.5893  3.9971

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.924883   0.398492   9.849  < 2e-16 ***
CountryAustria      -3.656242   0.271923 -13.446  < 2e-16 ***
CountryBelgium      -0.495552   0.277018  -1.789 0.074136 .
CountryCanada        1.415676   0.313124   4.521 7.40e-06 ***
CountryDenmark      -2.587607   0.280552  -9.223  < 2e-16 ***
CountryFinland      -0.700208   0.265237  -2.640 0.008506 **
CountryFrance        0.547321   0.356611   1.535 0.125361
CountryGermany      -2.371556   0.284291  -8.342 4.97e-16 ***
CountryIreland       0.295660   0.267730   1.104 0.269893
CountryItaly         1.270223   0.278592   4.559 6.21e-06 ***
CountryJapan        -0.478753   0.298393  -1.604 0.109140
CountryNetherlands  -2.577267   0.283765  -9.082  < 2e-16 ***
CountryNewZealand   -0.449020   0.275161  -1.632 0.103233
CountryNorway       -2.532430   0.267995  -9.450  < 2e-16 ***
CountrySweden       -3.778149   0.305177 -12.380  < 2e-16 ***
CountrySwitzerland  -4.085999   0.291072 -14.038  < 2e-16 ***
CountryUnitedKingdom -0.141312   0.269097  -0.525 0.599684
CountryUnitedStates  1.237793   0.327548   3.779 0.000173 ***
Dens                 0.035540   0.007108   5.000 7.52e-07 ***
Infl                 0.056286   0.011628   4.840 1.65e-06 ***
Year1951            -0.023465   0.005057  -4.640 4.27e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.109 on 604 degrees of freedom
Multiple R-squared:  0.694,     Adjusted R-squared:  0.6839
F-statistic:  68.5 on 20 and 604 DF,  p-value: < 2.2e-16
```