Regression models on strikes in OECD countries

Chen Yuan 20561973
Yuke Wu 20566786
Bochao Zhang 20589851

Summary

The goal of this project is to obtain the relation between strikes activity and some macroeconomic factors, such as unemployment rate and inflation. In order to achieve this goal, after excluding the dependent variable, we selected two candidate models by applying automated model selection and log transformation. We then provide the models' diagnostics information including residual plots, leverage and Cook's distance, so that we can choose the more suitable one.

We favored one of the models based on the model selection and model diagnostics conducted.The final model indicates the necessity of log transformation and the interaction between covariates including the interaction between unemployment rate & inflation and inflation & trade union density. Moreover, discovered by observing specific graphs, the model may violate heteroscedasticity assumption and normality assumption.

Model Selection

First, we would like to take a look at the general information of the data that we are going to analyze.

The summary of the strikes data is as follows

```
      Country         Year          Strike           Unemp
 Australia: 35   Min.   :1951   Min.   :   0.0   Min.   : 0.000
 Austria  : 35   1st Qu.:1959   1st Qu.:  22.0   1st Qu.: 1.200
 Canada   : 35   Median :1968   Median : 129.0   Median : 2.500
 Denmark  : 35   Mean   :1968   Mean   : 302.3   Mean   : 3.555
 Finland  : 35   3rd Qu.:1977   3rd Qu.: 362.0   3rd Qu.: 5.500
 France   : 35   Max.   :1985   Max.   :7000.0   Max.   :17.000
 (Other)  :415
      Infl            Demo           Centr            Dens
 Min.   :-2.900   Min.   : 8.16   Min.   :0.000   Min.   :15.10
 1st Qu.: 2.700   1st Qu.:32.20   1st Qu.:0.250   1st Qu.:33.90
 Median : 4.800   Median :42.50   Median :0.375   Median :43.50
 Mean   : 5.957   Mean   :40.85   Mean   :0.456   Mean   :45.69
 3rd Qu.: 8.200   3rd Qu.:49.70   3rd Qu.:0.750   3rd Qu.:57.20
 Max.   :27.500   Max.   :78.70   Max.   :1.000   Max.   :91.50
```

We then drew a table to obtain an overview of the correlation of all covariates excluding the country name.

Correlation table

	Year	Strike	Unemp	Infl	Demo	Centr	Dens
Year	1.0000000	0.0128132	0.3435994	0.4229497	0.0350355	-0.0112541	0.1446531
Strike	0.0128132	1.0000000	0.1716496	0.1753685	0.0168376	-0.2423249	-0.1046925
Unemp	0.3435994	0.1716496	1.0000000	0.1753565	-0.0395134	-0.2187568	-0.0081223
Infl	0.4229497	0.1753685	0.1753565	1.0000000	-0.0115602	0.0065534	0.2223047

	Year	Strike	Unemp	Infl	Demo	Centr	Dens
Demo	0.0350355	0.0168376	-0.0395134	-0.0115602	1.0000000	-0.1153820	0.0318138
Centr	-0.0112541	-0.2423249	-0.2187568	0.0065534	-0.1153820	1.0000000	0.7182962
Dens	0.1446531	-0.1046925	-0.0081223	0.2223047	0.0318138	0.7182962	1.0000000

Before conducting the automated model selection, we have to find out if there are any linearly dependent covariates.NA values in a model suggest that the parameter is linearly dependent with other parameters.

```
Model :
Strike ~ ((Country + Year + Unemp + Infl + Demo + Centr + Dens) -
    Country)^2 + Country

Complete :
                    (Intercept) Year Unemp Infl Demo Centr Dens
CountryUnitedStates    1           0    0     0    0 -8/3     0
                    CountryAustria CountryBelgium CountryCanada
CountryUnitedStates  5/3              1             -1
                    CountryDenmark CountryFinland CountryFrance
CountryUnitedStates  1/3              1             -1
                    CountryGermany CountryIreland CountryItaly
CountryUnitedStates -1/3            1/3           -1/3
                    CountryJapan CountryNetherlands CountryNewZealand
CountryUnitedStates -2/3          1                  0
                    CountryNorway CountrySweden CountrySwitzerland
CountryUnitedStates  4/3           4/3           1/3
                    CountryUnitedKingdom Year:Unemp Year:Infl Year:Demo
CountryUnitedStates    0                    0          0         0
                    Year:Centr Year:Dens Unemp:Infl Unemp:Demo Unemp:Centr
CountryUnitedStates    0          0         0          0          0
                    Unemp:Dens Infl:Demo Infl:Centr Infl:Dens Demo:Centr
CountryUnitedStates    0          0         0          0         0
                    Demo:Dens Centr:Dens
CountryUnitedStates    0         0
```

Since union centralization is a linearly dependent parameter as showed above, we have to exclude it from the final data set used to conduct automated model selection. We also made some changes to the raw data so that the model is found more easily and fits better. We deducted the year values from raw data by 1900, and made inflation values non-negative by adding the minimum inflation value to all values.

We first tried automated model selection directly. We set the minimal model to be the one containing the intercept only, and the maximal model to be the one consisting country name and the interactions between any two covariates except country name, i.e.

Minimal model: $Strike \sim 1$

Maximal model: $Strike \sim (Year+Inflation+Unemploymentrate+Democracyindex+Tradeuniondensity)^2+Country$

```
      Country          Year           Strike           Unemp      
 Australia: 35   Min.   :51.00   Min.   :   0.0   Min.   : 0.000  
 Austria  : 35   1st Qu.:59.00   1st Qu.:  22.0   1st Qu.: 1.200  
 Canada   : 35   Median :68.00   Median : 129.0   Median : 2.500  
 Denmark  : 35   Mean   :67.88   Mean   : 302.3   Mean   : 3.555  
 Finland  : 35   3rd Qu.:77.00   3rd Qu.: 362.0   3rd Qu.: 5.500
 France   : 35   Max.   :85.00   Max.   :7000.0   Max.   :17.000
 (Other)  :415
      Infl            Demo            Dens
 Min.   : 0.000   Min.   : 8.16   Min.   :15.10
 1st Qu.: 5.600   1st Qu.:32.20   1st Qu.:33.90
 Median : 7.700   Median :42.50   Median :43.50
 Mean   : 8.857   Mean   :40.85   Mean   :45.69
 3rd Qu.:11.100   3rd Qu.:49.70   3rd Qu.:57.20
 Max.   :30.400   Max.   :78.70   Max.   :91.50
```

The results are listed below:

Using forward selection:

```
lm(formula = Strike ~ Country + Infl + Unemp, data = strikes1)
```

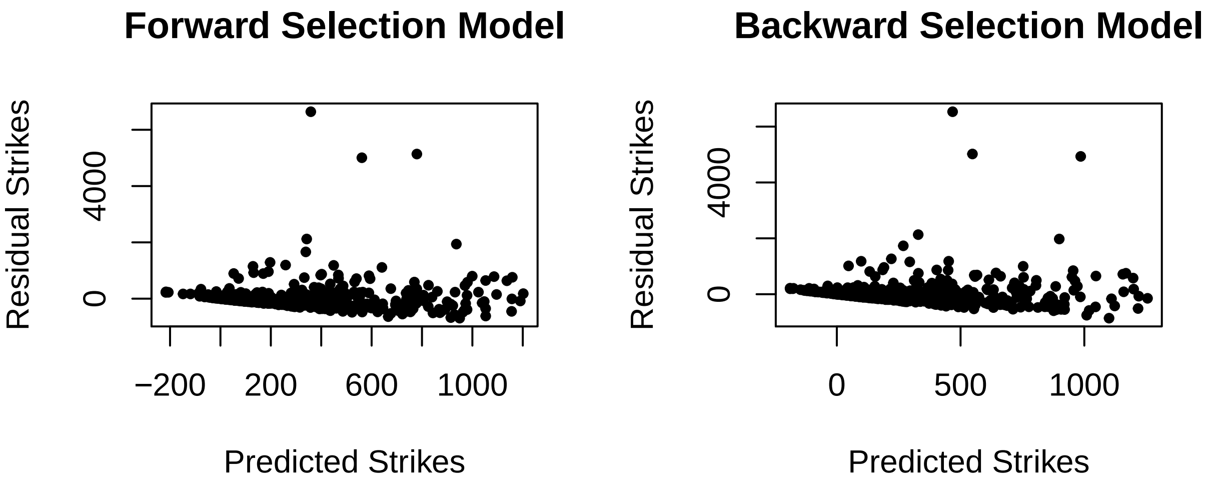
Using backward selection:

```
lm(formula = Strike ~ Year + Unemp + Infl + Demo + Dens + Country +
    Year:Demo + Unemp:Demo + Demo:Dens, data = strikes1)
```

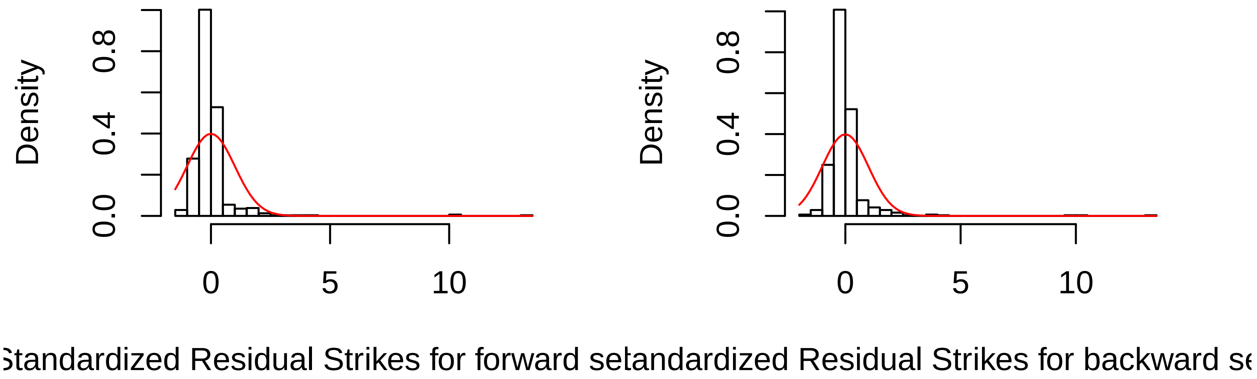
Using stepwise selection:

```
lm(formula = Strike ~ Country + Unemp + Infl, data = strikes1)
```

To analyze the suitability of the two models, we plotted a graph of residual against predicted strikes and a histogram of the residuals against their theoretical normal distribution.



For both graphs above, the dots seem to be a little linear rather than randomly spread around 0, which means these may not be good models for the given data since homoscedasticity assumption is violated.



From the histograms, the residuals are slightly biased towards positive values and have long tails, which indicates further modeling is required. Therefore, we tried log transformation.

We first took log of only one side, the strike. The starting models then became:

Minimal model: $log(strike + 1) \sim 1$

Maximal model: $log(strike + 1) \sim (Year + Inflation + Unemploymentrate + Democracyindex + Tradeuniondensity)^2 + Country$

We then conducted automated model selection, which gives

forward selection:

```
lm(formula = log(Strike + 1) ~ Country + Dens + Infl + Year +
    Infl:Year + Dens:Infl, data = strikes1)
```

backward selection:

```
lm(formula = log(Strike + 1) ~ Year + Unemp + Infl + Dens + Country +
    Year:Unemp + Year:Infl + Unemp:Infl + Infl:Dens, data = strikes1)
```

stepwise selection:

```
lm(formula = log(Strike + 1) ~ Country + Year + Unemp + Infl +
    Dens + Year:Unemp + Year:Infl + Unemp:Infl + Infl:Dens, data = strikes1)
```

We then tried log transformation on both sides. We gave new names to strike and covariates except country after they got log transformed, e.g. log_Infl is defined as log(Infl + 1). The new starting models are:

Minimal model: $log(Strike + 1) \sim 1$

Maximal model: $log(Strike + 1) \sim (log(Year + 1) + log(Infl + 1) + log(Unemp + 1) + log(Demo + 1) + log(Dens + 1))^2 + Country$

And the models given by automated model selection are:

forward selection:

```
lm(formula = (log_Strike_) ~ Country + log_Dens_ + log_Year_ +
    log_Infl_ + log_Unemp_ + log_Year_:log_Infl_ + log_Dens_:log_Infl_ +
    log_Year_:log_Unemp_ + log_Infl_:log_Unemp_, data = strikes_log)
```

backward selection:

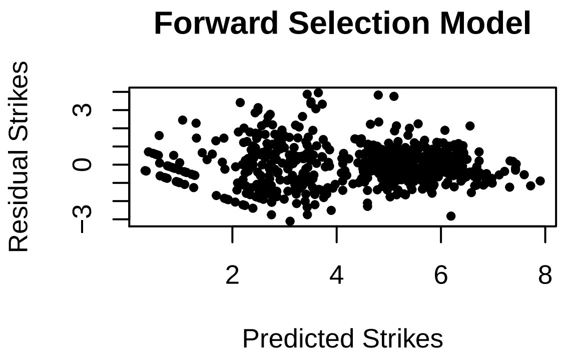
```
lm(formula = log_Strike_ ~ log_Year_ + log_Unemp_ + log_Infl_ +
    log_Dens_ + Country + log_Year_:log_Unemp_ + log_Year_:log_Infl_ +
    log_Unemp_:log_Infl_ + log_Infl_:log_Dens_, data = strikes_log)
```

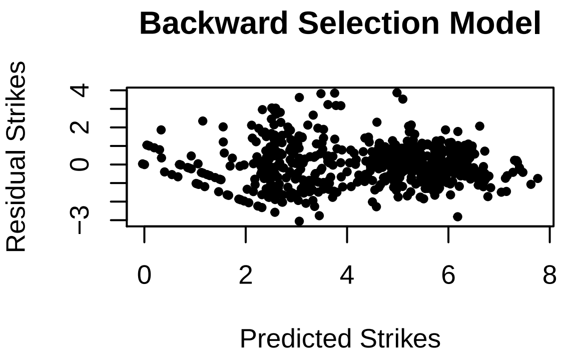
stepwise selection:

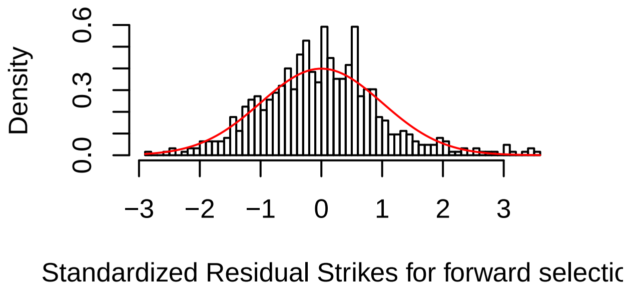
```
lm(formula = log_Strike_ ~ Country + log_Year_ + log_Unemp_ +
    log_Infl_ + log_Dens_ + log_Year_:log_Infl_ + log_Infl_:log_Dens_ +
    log_Year_:log_Unemp_ + log_Unemp_:log_Infl_, data = strikes_log)
```

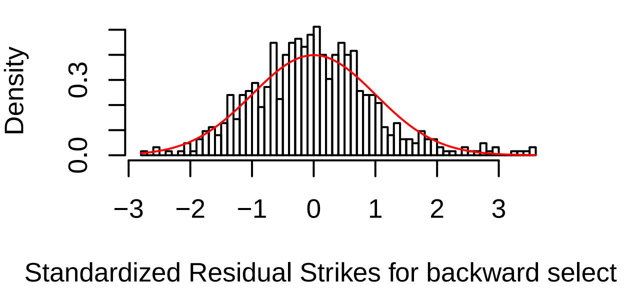
Again, we plot the graph of residual against predicted strikes and the histogram of the residuals against their theoretical normal distribution.

First, the log-one-side models.



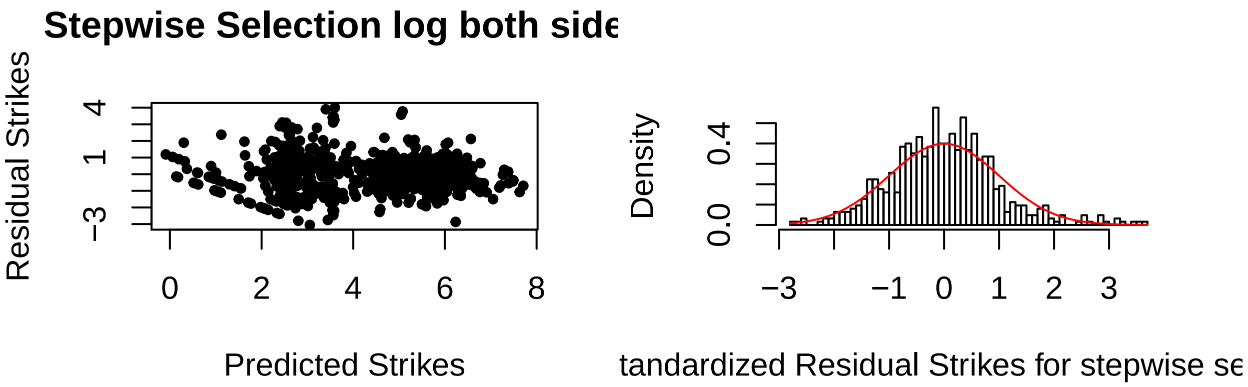






These two models seem to fit better than the previous two. The dots are randomly spread around 0. For the histograms, the standardized residuals seem very normal and the tail is short and acceptable, especially the one for the backward selection model. Therefore, we chose the backward selection model as our first candidate model.

We then take a look at the two graphs for the log-two-sides model.



As observed, the dots in the plot are randomly spread, and the histogram seems quite random, ignoring the little acceptable tail. This model became our second candidate model.

Model Diagnostics

We now have our two candidate models, which are

Model 1: $log(Strike+1) \sim Year+Unemp+Infl+Dens+Country+Year:Unemp+Year:Infl+Unemp:Infl+Infl:Dens$

Model 2: $log(Strike+1) \sim log(Year+1)+log(Unemp+1)+log(Infl+1)+log(Dens+1)+Country+log(Year+1):log(Unemp+1)+log(Year+1):log(Infl+1)+log(Unemp+1):log(Infl+1)+log(Infl+1):log(Dens+1)$

However, we still need to conduct some analysis before we are able to pick the better model from the two candidates.

First we calculate the AIC statistic for both models.

```
    AIC1     AIC2
1913.706 1917.065
```

Since AIC of Model 2 is larger than that of Model 1, AIC picks Model 1 over Model 2.

For PRESS Statistic,

```
  PRESS1   PRESS2
781.6154 786.9358
```

Again, PRESS favors Model 1 since it has a smaller PRESS statistic.

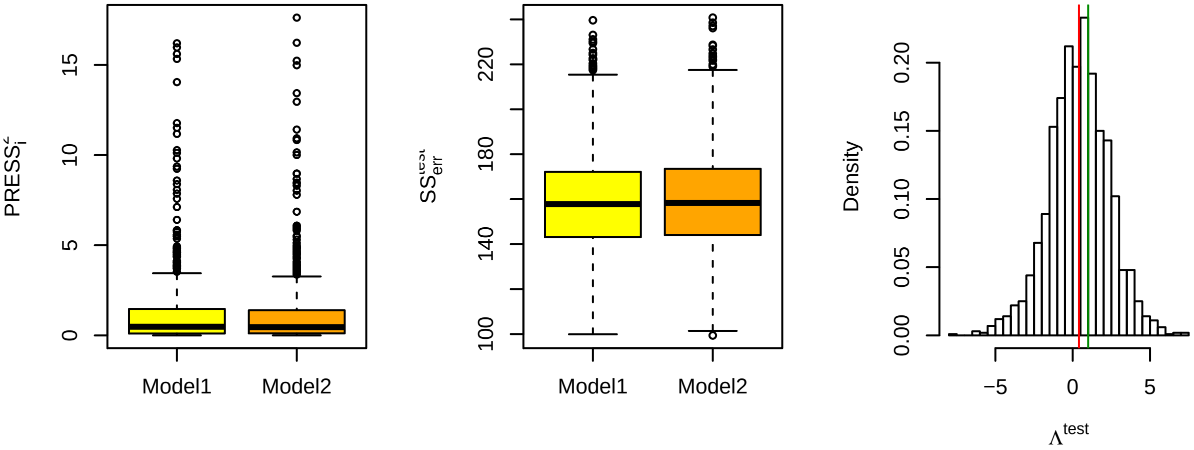
We then conduct cross-validation for two candidate models.

```
    SSE1     SSE2
158.5695 159.5523
```

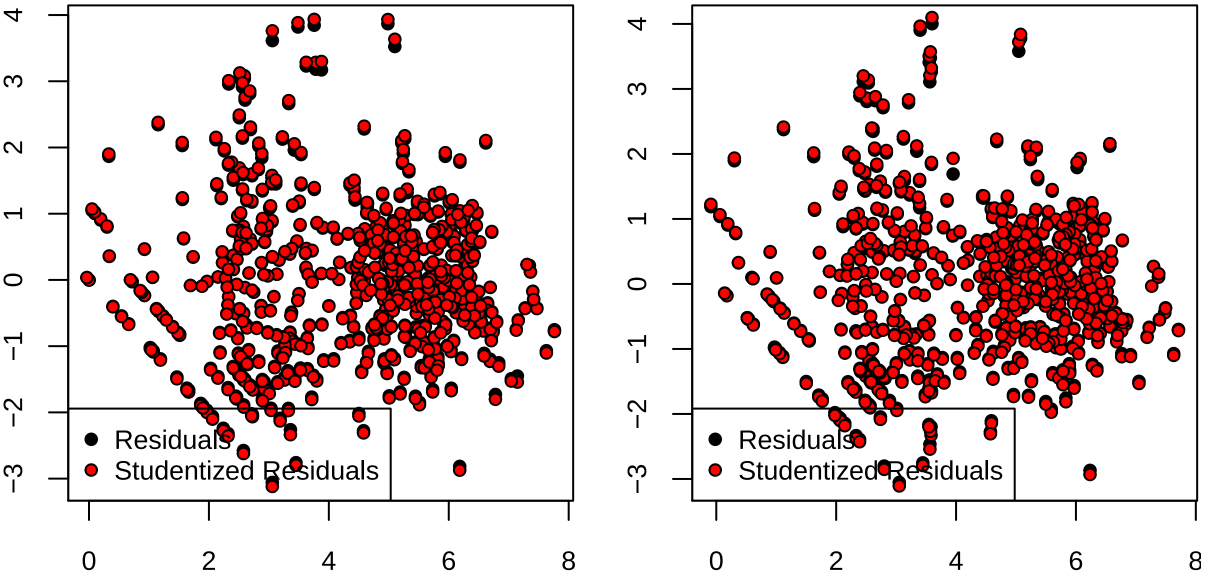
Cross-validation also favors Model 1, which has a smaller sum of square errors.

From the VIF values(see Appendix 1), we can tell that all interaction terms in both models are highly correlated, and the interation terms in Model 2 may be more correlated than those in Model 1.

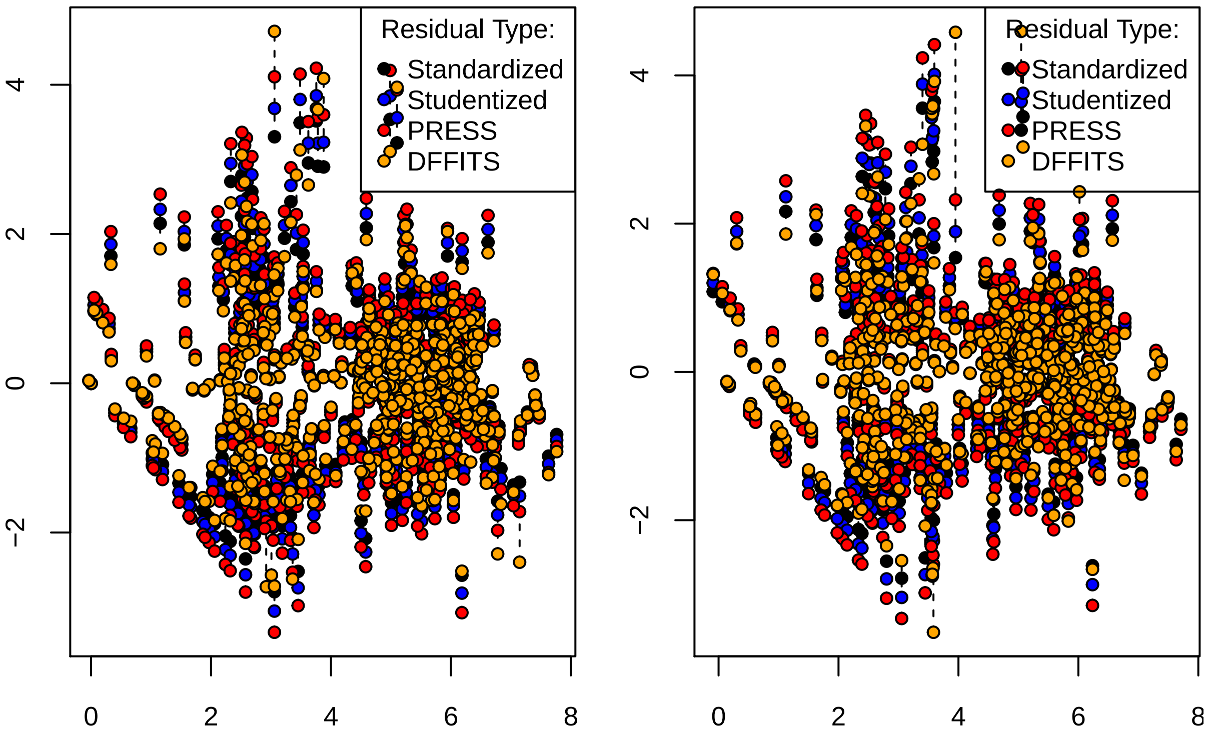
To view in graphs, we plotted some graphs, including PRESS statistic, SSE values and Lambda.



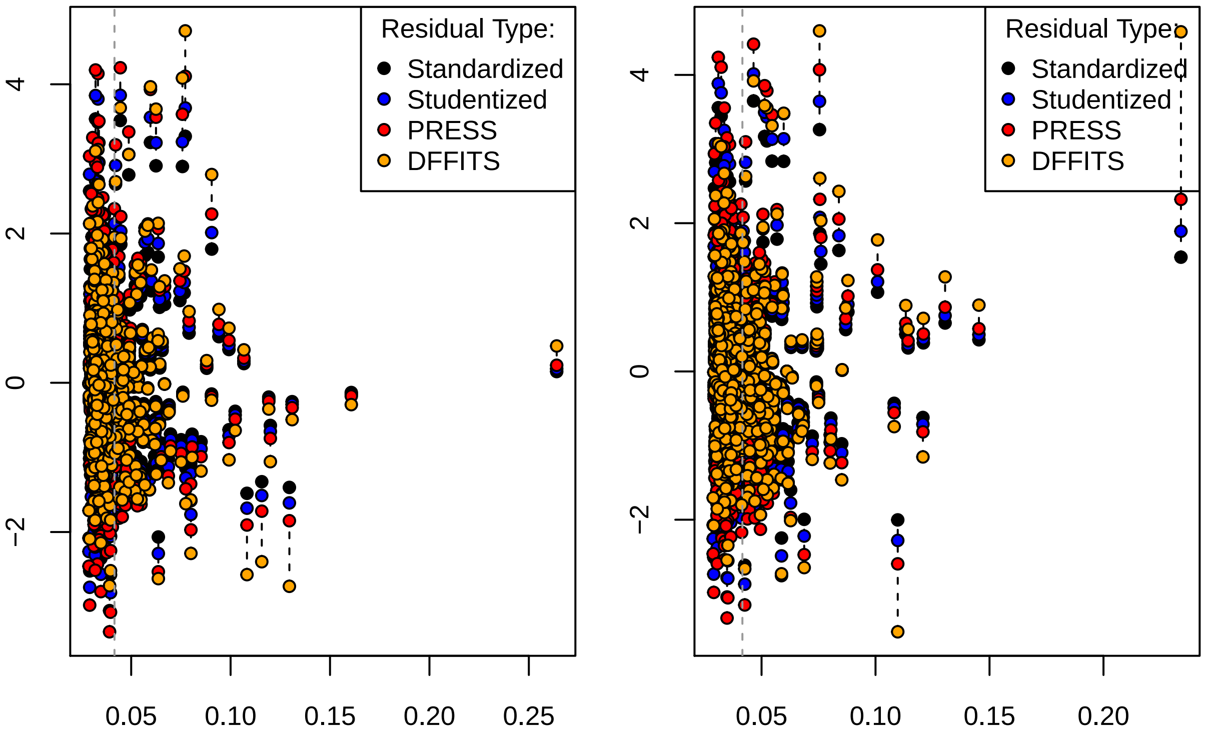
The box plot seems similar for both models, but Λ<1 means Model 2 is better. Λ has a potential overfitting problem as it tends to favor the model with more covariates, so the preference may not be very accurate.

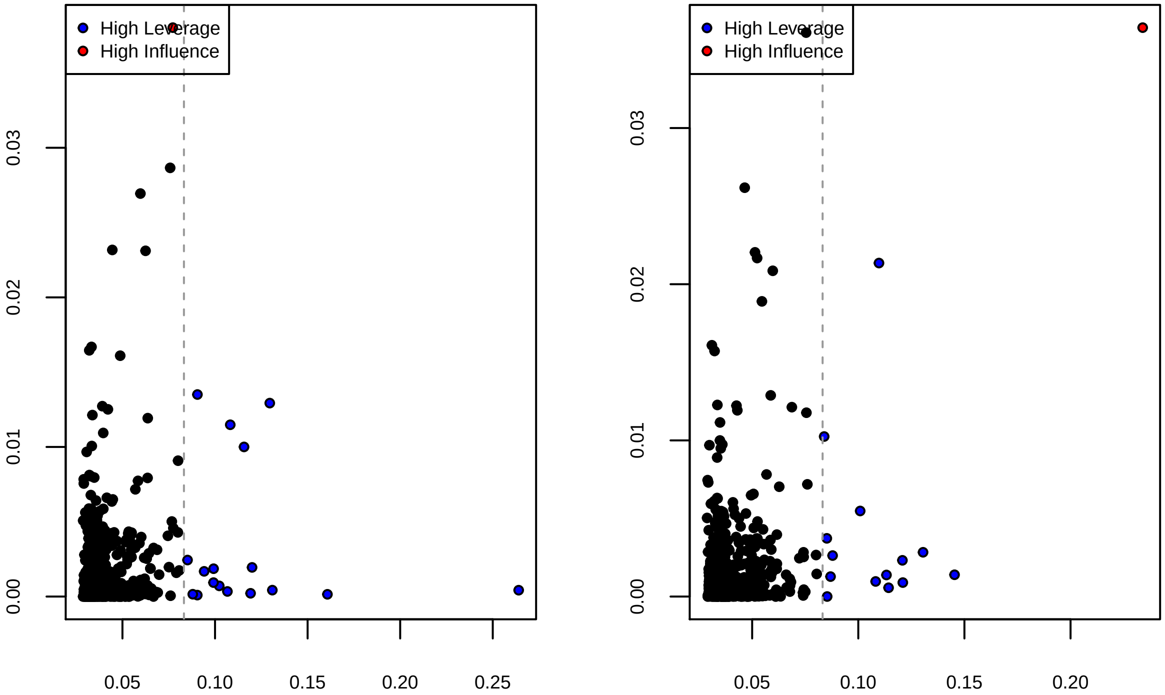


The graphs above include predict strikes, residuals and studentized residuals. They seem similar. Both have minuscule difference between residuals and studentized residuals. This may not be a good justification tool.



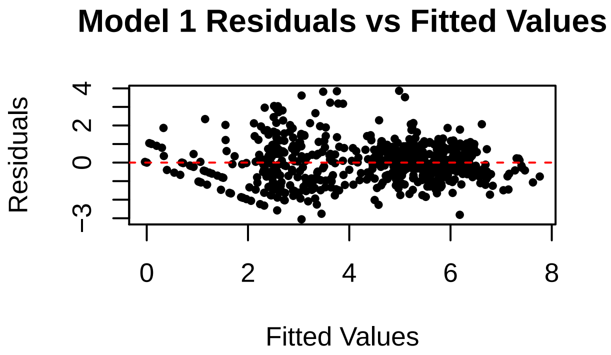
After the residuals got amplified, both models have significant DIFFTS values, but DFFITS of Model 1 is obviously larger than DFFITS of Model 2 from the graph.

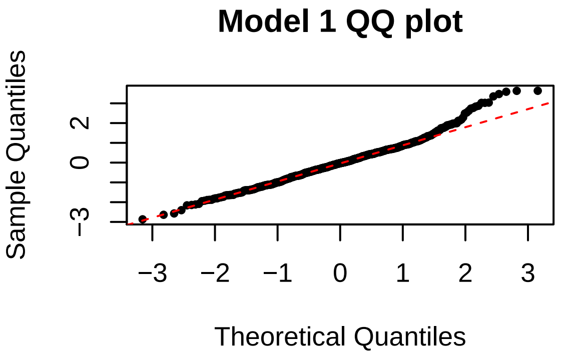


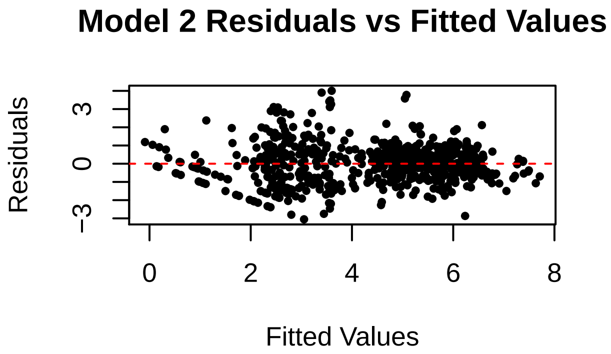


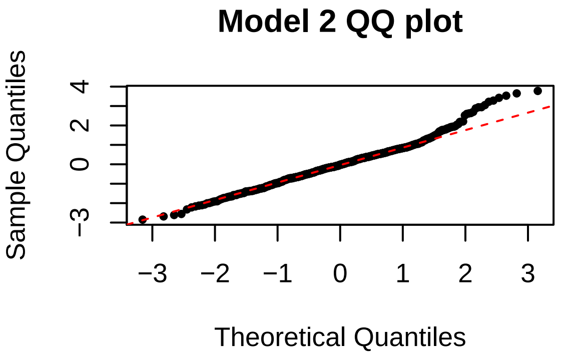
From the graph Cook's Distance against leverage, Model 1 has 17 observations with higher leverage, and a low-leverage high-influence observation. Model 2 has 16 observations with higher leverage, and a high-leverage high-influence observation. This may be because there are some extreme observations affected by external factors or the models are not suitable.

In this particular case, it is a little difficult to tell which one is better since they have similar extreme observations. Model 2 may be better since its extreme values of leverage and Cook's distance is smaller.









Both models have large variance of residuals, which may be an indication of the violantion of homoscedascity assumption, and heavy tails in the QQ plots, which means the models violates normality assumption.

Conclusion

Based on all the analysis above, we decide to retain Model 1 as the indication of strike activity and given covariates. Below is some important information of Model 1.

It seems important factors of this model includes country, year and the interaction between year and unemployment rate.

Some factors with high p-value retained in the final model. Most of them are countries. This may be because strike activities are influenced by business culture in the specific country.

The regression assumptions that Model 1 violates are homoscedascity assumption and normality assumption. From Figure above, we can see some residuals with relatively large variance which indicates the violation of homoscedascity assupmtion. This may be suggesting that there are some external factors which are not included in the data provided. From the QQ plot above, the heavy tail shows that normality assumption does not holds in this particular case.

Appendix 1

VIF for model 1

```
                Year                Unemp                 Infl 
            9.256713            51.906368            61.883595 
                Dens       CountryAustria       CountryBelgium 
           17.037805             2.079157             1.794197 
       CountryCanada       CountryDenmark       CountryFinland 
            2.942756             2.249411             1.954042 
       CountryFrance       CountryGermany       CountryIreland 
            3.605903             2.245776             2.319699 
        CountryItaly         CountryJapan   CountryNetherlands 
            2.454683             2.504054             2.222727 
   CountryNewZealand        CountryNorway        CountrySweden 
            2.192717             2.004662             2.602772 
  CountrySwitzerland CountryUnitedKingdom  CountryUnitedStates 
            2.507121             1.985421             3.192356 
          Year:Unemp            Year:Infl           Unemp:Infl 
           65.731905            78.573025            12.946452 
           Infl:Dens 
           26.171779 
```

VIF for model 2

```
      CountryAustria       CountryBelgium        CountryCanada 
            2.049772             1.797388             3.297460 
      CountryDenmark       CountryFinland        CountryFrance 
            2.072357             1.991730             5.208391 
      CountryGermany       CountryIreland         CountryItaly 
            2.225544             2.184248             2.406625 
        CountryJapan   CountryNetherlands    CountryNewZealand 
            2.608417             2.229244             2.361540 
       CountryNorway        CountrySweden   CountrySwitzerland 
            2.035251             2.294570             2.763644 
CountryUnitedKingdom  CountryUnitedStates            log_Year_ 
            1.973630             3.886297            41.056813 
          log_Unemp_            log_Infl_            log_Dens_ 
          937.212156          1059.206220            47.255899 
 log_Year_:log_Infl_  log_Infl_:log_Dens_ log_Year_:log_Unemp_ 
         1218.570268           215.407322          1110.388717 
log_Unemp_:log_Infl_ 
           48.016725 
```

Model1

```

Call:
lm(formula = log(Strike + 1) ~ Year + Unemp + Infl + Dens + Country + 
    Year:Unemp + Year:Infl + Unemp:Infl + Infl:Dens, data = strikes1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0573 -0.7025 -0.0270  0.6199  3.8686 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           8.6265076  1.0193725   8.463  < 2e-16 ***
Year                 -0.0724423  0.0132518  -5.467 6.74e-08 ***
Unemp                -0.2815140  0.1040378  -2.706 0.007006 ** 
Infl                 -0.1545549  0.0745092  -2.074 0.038477 *  
Dens                  0.0187178  0.0115601   1.619 0.105936    
CountryAustria       -3.4439655  0.2744992 -12.546  < 2e-16 ***
CountryBelgium       -0.4701281  0.2742668  -1.714 0.087023 .  
CountryCanada         1.2205081  0.3265686   3.737 0.000204 ***
CountryDenmark       -2.4501125  0.2855169  -8.581  < 2e-16 ***
CountryFinland       -0.8194241  0.2661119  -3.079 0.002170 ** 
CountryFrance         0.3404365  0.3614968   0.942 0.346705    
CountryGermany       -2.4133212  0.2852861  -8.459  < 2e-16 ***
CountryIreland        0.3476904  0.2899434   1.199 0.230937    
CountryItaly          1.2544640  0.2982601   4.206 3.00e-05 ***
CountryJapan         -0.5472830  0.3012446  -1.817 0.069757 .  
CountryNetherlands   -2.7392185  0.2838183  -9.651  < 2e-16 ***
CountryNewZealand    -0.6152693  0.2818958  -2.183 0.029452 *  
CountryNorway        -2.4983870  0.2695367  -9.269  < 2e-16 ***
CountrySweden        -3.6733157  0.3071252 -11.960  < 2e-16 ***
CountrySwitzerland   -4.0707646  0.3014290 -13.505  < 2e-16 ***
CountryUnitedKingdom -0.2773643  0.2682401  -1.034 0.301546    
CountryUnitedStates   1.0484533  0.3401363   3.082 0.002147 ** 
Year:Unemp            0.0054371  0.0015034   3.617 0.000324 ***
Year:Infl             0.0029656  0.0010596   2.799 0.005294 ** 
Unemp:Infl           -0.0114277  0.0040851  -2.797 0.005317 ** 
Infl:Dens             0.0012156  0.0007493   1.622 0.105258    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.094 on 599 degrees of freedom
Multiple R-squared:  0.7045,	Adjusted R-squared:  0.6922 
F-statistic: 57.13 on 25 and 599 DF,  p-value: < 2.2e-16
```

Model2

```
Call:
lm(formula = log_Strike_ ~ Country + log_Year_ + log_Unemp_ + 
    log_Infl_ + log_Dens_ + log_Year_:log_Infl_ + log_Infl_:log_Dens_ + 
    log_Year_:log_Unemp_ + log_Unemp_:log_Infl_, data = strikes_log)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0529 -0.6900 -0.0244  0.6126  4.0029 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           41.6397     8.5167   4.889 1.30e-06 ***
```

```
CountryAustria        -3.3887     0.2733 -12.400  < 2e-16 ***
CountryBelgium        -0.4996     0.2752  -1.815 0.070013 .
CountryCanada          1.1062     0.3466   3.191 0.001490 **
CountryDenmark        -2.4844     0.2748  -9.041  < 2e-16 ***
CountryFinland        -0.7140     0.2694  -2.650 0.008253 **
CountryFrance          0.5791     0.4356   1.329 0.184233
CountryGermany        -2.4479     0.2848  -8.596  < 2e-16 ***
CountryIreland         0.1089     0.2821   0.386 0.699523
CountryItaly           1.0320     0.2961   3.485 0.000528 ***
CountryJapan          -0.5087     0.3083  -1.650 0.099458 .
CountryNetherlands    -2.7282     0.2850  -9.573  < 2e-16 ***
CountryNewZealand     -0.5091     0.2933  -1.736 0.083125 .
CountryNorway         -2.4141     0.2723  -8.865  < 2e-16 ***
CountrySweden         -3.4555     0.2891 -11.951  < 2e-16 ***
CountrySwitzerland    -3.9115     0.3173 -12.326  < 2e-16 ***
CountryUnitedKingdom  -0.2790     0.2682  -1.040 0.298585
CountryUnitedStates    1.0272     0.3763   2.730 0.006525 **
log_Year_             -8.6831     1.8995  -4.571 5.89e-06 ***
log_Unemp_            -4.2227     2.0228  -2.088 0.037264 *
log_Infl_            -11.9446     3.1293  -3.817 0.000149 ***
log_Dens_             -0.5921     0.8575  -0.690 0.490199
log_Year_:log_Infl_    2.3457     0.7333   3.199 0.001453 **
log_Infl_:log_Dens_    0.8340     0.3156   2.643 0.008435 **
log_Year_:log_Unemp_   1.2064     0.5076   2.377 0.017784 *
log_Unemp_:log_Infl_  -0.3249     0.1784  -1.821 0.069047 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.097 on 599 degrees of freedom
Multiple R-squared:  0.7029,    Adjusted R-squared:  0.6905
F-statistic: 56.69 on 25 and 599 DF,  p-value: < 2.2e-16
```

Appendix 2

Rcode for Analysis

```r
# read data
strikes <- read.csv("strikes_clean.csv")

#correlation tables
require(knitr)
corr <- cor(strikes[-1])
kable(corr)

#automated model selection
M0 <- lm(Strike ~ 1, data=strikes)
Mfull <- lm(Strike ~ (.-Country)^2 + Country, data=strikes)
alias(Mfull)

# doing automated model selection directly
strikes1 <- strikes[-7] #remove Centr
```

```r
strikes1$Year <- strikes1$Year - 1900 #deduct year by 1900
strikes1$Infl <- strikes1$Infl + abs(min(strikes1$Infl)) #make inflation positive
summary(strikes1)
M0_red <- lm(Strike ~ 1, data=strikes1) #reduced M0
Mfull_red <- lm(Strike ~ (.-Country)^2 + Country, data=strikes1) #reduced Mfull

Mfwd <- step(object = M0_red, scope = list(lower = M0_red, upper = Mfull_red),
             direction = "forward", trace = FALSE)
Mfwd$call

Mback <- step(object = Mfull_red, scope = list(lower = M0_red, upper = Mfull_red),
              direction = "backward", trace = FALSE)
Mback$call

Mstart <- lm(Strike ~ ., data=strikes1)
Mstep <- step(object = Mstart, scope = list(lower = M0_red, upper = Mfull_red),
              direction = "both", trace = FALSE)
Mstep$call

par(mfrow=c(1,2))
plot(predict(Mfwd), resid(Mfwd), pch=21, bg="black", cex = .6, xlab="Predicted Strikes",
     ylab = "Residual Strikes", main="Forward Selection Model")
plot(predict(Mback), resid(Mback), pch=21, bg="black", cex=.6, xlab="Predicted Strikes",
     ylab = "Residual Strikes", main="Backward Selection Model")

par(mfrow=c(1,2))
sigma.hat_fwd <- sqrt(sum(resid(Mfwd)^2)/Mfwd$df)
sigma.hat_back <- sqrt(sum(resid(Mback)^2)/Mback$df)
hist(resid(Mfwd)/sigma.hat_fwd, breaks = 50, freq = FALSE,
     xlab = "Standardized Residual Strikes for forward selection", main = "")
curve(dnorm(x), col = "red", add = TRUE)
hist(resid(Mback)/sigma.hat_back, breaks = 50, freq = FALSE,
     xlab = "Standardized Residual Strikes for backward selection", main = "")
curve(dnorm(x), col = "red", add = TRUE)


# doing log transfomation on left side only
M0_log <- lm(log(Strike + 1) ~ 1, data=strikes1)
Mfull_log <- lm(log(Strike + 1) ~ (.-Country)^2 + Country, data=strikes1)

Mfwd_log <- step(object = M0_log, scope= list(lower = M0_log, upper = Mfull_log),
                 direction = "forward", trace = FALSE)
Mfwd_log$call

Mback_log <- step(object = Mfull_log, scope = list(lower = M0_log, upper = Mfull_log),
                  direction = "backward", trace = FALSE)
Mback_log$call

Mstep_start <- lm(log(Strike + 1) ~., data=strikes1)
Mstep_log <- step(object=Mstep_start, scope = list(lower = M0_log, upper = Mfull_log),
                  direction = "both", trace = FALSE)
Mstep_log$call
```

```
#doing log transcormation on both sides
strikes_log <- strikes1
strikes_log[-1] <- log(strikes_log[-1] + 1)
names(strikes_log)[-1] <- paste0("log_", names(strikes_log)[-1], "_")
M0_log_d <- lm((log_Strike_) ~ 1, data = strikes_log)
Mfull_log_d <- lm(log_Strike_ ~ (.-Country)^2 + Country, data=strikes_log)

Mfwd_log_d <- step(object=M0_log_d, scope=list(lower = M0_log_d, upper=Mfull_log_d),
                   direction="forward", trace=FALSE)
Mfwd_log_d$call

Mback_log_d <- step(object=Mfull_log_d, scope=list(lower=M0_log_d, upper=Mfull_log_d),
                    direction="backward", trace=FALSE)
Mback_log_d$call

Mstart_log_d <- lm(log_Strike_ ~., data=strikes_log)
Mstep_log_d <- step(object=Mstart_log_d, scope=list(lower=M0_log_d, upper=Mfull_log_d),
                    direction="both", trace=FALSE)
Mstep_log_d$call

par(mfrow=c(2,2))
plot(predict(Mfwd_log), resid(Mfwd_log), pch=21, bg="black", cex = .6,
     xlab="Predicted Strikes",
     ylab = "Residual Strikes", main="Forward Selection Model")
plot(predict(Mback_log), resid(Mback_log), pch=21, bg="black", cex=.6,
     xlab="Predicted Strikes",
     ylab = "Residual Strikes", main="Backward Selection Model")
sigma.hat_fwdlog <- sqrt(sum(resid(Mfwd_log)^2)/Mfwd_log$df)
sigma.hat_backlog <- sqrt(sum(resid(Mback_log)^2)/Mback_log$df)
hist(resid(Mfwd_log)/sigma.hat_fwdlog, breaks = 50, freq = FALSE,
     xlab = "Standardized Residual Strikes for forward selection", main = "")
curve(dnorm(x), col = "red", add = TRUE)
hist(resid(Mback_log)/sigma.hat_backlog, breaks = 50, freq = FALSE,
     xlab = "Standardized Residual Strikes for backward selection", main = "")
curve(dnorm(x), col = "red", add = TRUE)

par(mfrow=c(1,2))
plot(predict(Mstep_log_d), resid(Mstep_log_d), pch=21, bg="black", cex=.6,
     xlab="Predicted Strikes", ylab = "Residual Strikes",
     main="Stepwise Selection log both sides")
sigma.hat_step_d <- sqrt(sum(resid(Mstep_log_d)^2)/Mstep_log_d$df)
hist(resid(Mstep_log_d)/sigma.hat_step_d, breaks=50, freq=FALSE,
     xlab="Standardized Residual Strikes for stepwise selection", main="")
curve(dnorm(x),col="red", add=TRUE)

# Model Diagnostics
M1 <- Mback_log
M2 <- Mstep_log_d
Mnames <- expression(Model1, Model2)
c(AIC1 = AIC(M1), AIC2 = AIC(M2))

press1 <- resid(M1)/(1-hatvalues(M1)) # M1
press2 <- resid(M2)/(1-hatvalues(M2)) # M2
```

```r
c(PRESS1 = sum(press1^2), PRESS2 = sum(press2^2)) # favors M1

nreps <- 2e3 # number of replications
ntot <- nrow(strikes1) # total number of observations
ntrain <- 500 # size of training set
ntest <- ntot-ntrain # size of test set
sse1 <- rep(NA, nreps) # sum-of-square errors for each CV replication
sse2 <- rep(NA, nreps)
Lambda <- rep(NA, nreps) # likelihod ratio statistic for each replication
system.time({
for(ii in 1:nreps) {
  if(ii%%400 == 0) message("ii = ", ii)
  # randomly select training observations
  train.ind <- sample(ntot, ntrain) # training observations
  # this is the faster R way
  M1.cv <- update(M1, subset = train.ind)
  M2.cv <- update(M2, subset = train.ind)
  # testing residuals for both models
  # that is, testing data - predictions with training parameters
  M1.res <- log(strikes1$Strike+1)[-train.ind] -
    predict(M1.cv, newdata = strikes1[-train.ind,])
  M2.res <- strikes_log$log_Strike_[-train.ind] -
    predict(M2.cv, newdata = strikes_log[-train.ind,])
  # total sum of square errors
  sse1[ii] <- sum((M1.res)^2)
  sse2[ii] <- sum((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))
  }
})

c(SSE1 = mean(sse1), SSE2 = mean(sse2))

# VIF
X_log <- cor(model.matrix(M1))
X_log <- X_log[-1,-1]
VIF_log <- diag(solve(X_log))

X_log_d <- cor(model.matrix(M2))
X_log_d <- X_log_d[-1,-1]
VIF_log_d <- diag(solve(X_log_d))

par(mfrow = c(1,3))
boxplot(x = list(press1^2, press2^2), names = Mnames,
        ylab = expression(PRESS[i]^2), col = c("yellow", "orange"))
# plot cross-validation SSE and Lambda
boxplot(x = list(sse1, sse2), names = Mnames, cex = .7,
ylab = expression(SS[err]^{test}), col = c("yellow", "orange"))
hist(Lambda, breaks = 50, freq = FALSE, xlab = expression(Lambda^{test}),
     main = "", cex = .7)
```

```r
abline(v = mean(Lambda), col = "red") # average value
abline(v = 1, col = "cadetblue") # to compare with 1

MDstep_log <- Mstep_log
MDstep_log_d <- Mstep_log_d

par(mfrow=c(1,2))
# Studentized residual plots
# In _log model
res_log <- resid(MDstep_log)
H_log <- model.matrix(MDstep_log)
H_log <- H_log %*% solve(crossprod(H_log),t(H_log))
h_log <- diag(H_log)
res.stu_log <- resid(MDstep_log)/sqrt(1-h_log)

cex <- .8
par(mar = c(2,3,.1,.1))
plot(predict(MDstep_log), res_log, pch = 21, bg="black", cex=cex, cex.axis=cex,
     xlab="Predicted Strikes", ylab="Residual Strikes")
points(predict(MDstep_log), res.stu_log, pch=21, bg="red", cex=cex)
legend(x="bottomleft", c("Residuals", "Studentized Residuals"), pch=21,
       pt.bg=c("black","red"), pt.cex=cex, cex=cex)

# In _log_d model
res_log_d <- resid(MDstep_log_d)
H_log_d <- model.matrix(MDstep_log_d)
H_log_d <- H_log_d %*% solve(crossprod(H_log_d),t(H_log_d))
h_log_d <- diag(H_log_d)
res.stu_log_d <- resid(MDstep_log_d)/sqrt(1-h_log_d)

cex <- .8
par(mar = c(2,3,.1,.1))
plot(predict(Mstep_log_d), res_log_d, pch=21, bg="black", cex=cex, cex.axis=cex,
     xlab="Predicted Strikes", ylab="Residual Strikes")
points(predict(Mstep_log_d), res.stu_log_d, pch=21, bg="red", cex=cex)
legend(x="bottomleft", c("Residuals", "Studentized Residuals"), pch=21,
       pt.bg=c("black","red"), pt.cex=cex, cex=cex)

# Standardize
# log model
# Press residual
press_log <- res_log/(1-h_log)
press_log_d <- res_log_d/(1-h_log_d)

# Dffits residuals
dfts_log <- dffits(MDstep_log)
dfts_log_d <- dffits(MDstep_log_d)
p_log <- length(coef(MDstep_log))
n_log <- nobs(MDstep_log)
hbar_log <- p_log/n_log
stud.res_log <- res.stu_log*sqrt(1-hbar_log)
press_log <- press_log*(1-hbar_log)*summary(MDstep_log)$sigma
dfts_log <- dfts_log*(1-hbar_log)/sqrt(hbar_log)
```

```
y.hat_log <- predict(MDstep_log)
stan.res_log <- res_log/summary(MDstep_log)$sigma
par(mfrow=c(1,2), mar=c(3,3,.1,.1))
plot(y.hat_log,rep(0, length(y.hat_log)),type= "n",
     ylim =range(stan.res_log,stud.res_log, press_log, dfts_log),
     cex.axis=cex, xlab="Predicted Values", ylab="Residuals")

segments(x0 = y.hat_log,
         y0 = pmin(stan.res_log, stud.res_log, press_log, dfts_log),
         y1 = pmax(stan.res_log, stud.res_log, press_log, dfts_log),
         lty = 2)
points(y.hat_log, stan.res_log, pch = 21, bg = "black", cex = cex)
points(y.hat_log, stud.res_log, pch = 21, bg = "blue", cex = cex)
points(y.hat_log, press_log, pch = 21, bg = "red", cex = cex)
points(y.hat_log, dfts_log, pch = 21, bg = "orange", cex = cex)
legend("topright", legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
pch = 21, pt.bg = c("black", "blue", "red", "orange"), title = "Residual Type:",
cex = cex, pt.cex = cex)

# double log model
p_log_d <- length(coef(MDstep_log_d))
n_log_d <- nobs(MDstep_log_d)
hbar_log_d <- p_log_d/n_log_d
stud.res_log_d <- res.stu_log_d*sqrt(1-hbar_log_d)
press_log_d <- press_log_d*(1-hbar_log_d)*summary(MDstep_log_d)$sigma
dfts_log_d <- dfts_log_d*(1-hbar_log_d)/sqrt(hbar_log_d)

y.hat_log_d <- predict(MDstep_log_d)
stan.res_log_d <- res_log_d/summary(MDstep_log_d)$sigma
plot(y.hat_log_d,rep(0, length(y.hat_log_d)),type= "n",
     ylim =range(stan.res_log_d,stud.res_log_d, press_log_d, dfts_log_d),
     cex.axis=cex,
     xlab="Predicted Values", ylab="Residuals")

segments(x0 = y.hat_log_d,
         y0 = pmin(stan.res_log_d, stud.res_log_d, press_log_d, dfts_log_d),
         y1 = pmax(stan.res_log_d, stud.res_log_d, press_log_d, dfts_log_d),
         lty = 2)
points(y.hat_log_d, stan.res_log_d, pch = 21, bg = "black", cex = cex)
points(y.hat_log_d, stud.res_log_d, pch = 21, bg = "blue", cex = cex)
points(y.hat_log_d, press_log_d, pch = 21, bg = "red", cex = cex)
points(y.hat_log_d, dfts_log_d, pch = 21, bg = "orange", cex = cex)
legend("topright", legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
pch = 21, pt.bg = c("black", "blue", "red", "orange"), title = "Residual Type:",
cex = cex, pt.cex = cex)

par(mfrow=c(1,2), mar=c(3,3,.1,.1))
# Against leverage
# log model
plot(h_log, rep(0, length(y.hat_log)), type = "n", cex.axis = cex,
     ylim = range(stan.res_log, stud.res_log, press_log, dfts_log),
     xlab = "Leverages", ylab = "Residuals")
segments(x0 = h_log,
         y0 = pmin(stan.res_log, stud.res_log, press_log, dfts_log),
         y1 = pmax(stan.res_log, stud.res_log, press_log, dfts_log),
         lty = 2)
points(h_log, stan.res_log, pch = 21, bg = "black", cex = cex)
points(h_log, stud.res_log, pch = 21, bg = "blue", cex = cex)
points(h_log, press_log, pch = 21, bg = "red", cex = cex)
points(h_log, dfts_log, pch = 21, bg = "orange", cex = cex)
abline(v = hbar_log, col = "grey60", lty = 2)
legend("topright", legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
       pch = 21, pt.bg = c("black", "blue", "red", "orange"),
       title = "Residual Type:",
       cex = cex, pt.cex = cex)

# double log model
plot(h_log_d, rep(0, length(y.hat_log_d)), type = "n", cex.axis = cex,
     ylim = range(stan.res_log_d, stud.res_log_d, press_log_d, dfts_log_d),
     xlab = "Leverages", ylab = "Residuals")
segments(x0 = h_log_d,
         y0 = pmin(stan.res_log_d, stud.res_log_d, press_log_d, dfts_log_d),
         y1 = pmax(stan.res_log_d, stud.res_log_d, press_log_d, dfts_log_d),
         lty = 2)
points(h_log_d, stan.res_log_d, pch = 21, bg = "black", cex = cex)
points(h_log_d, stud.res_log_d, pch = 21, bg = "blue", cex = cex)
points(h_log_d, press_log_d, pch = 21, bg = "red", cex = cex)
points(h_log_d, dfts_log_d, pch = 21, bg = "orange", cex = cex)
abline(v = hbar_log_d, col = "grey60", lty = 2)
legend("topright", legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
       pch = 21, pt.bg = c("black", "blue", "red", "orange"),
       title = "Residual Type:",
       cex = cex, pt.cex = cex)

par(mfrow=c(1,2))
# cook's distance vs. leverage
# log model
D_log <- cooks.distance(MDstep_log)
infl.ind_log <- which.max(D_log)
lev.ind_log <- h_log > 2*hbar_log
clrs_log <- rep("black", len = n_log)
clrs_log[lev.ind_log] <- "blue"
clrs_log[infl.ind_log] <- "red"
par(mar = c(3,3,1,1))
cex <- .6
plot(h_log, D_log, xlab = "Leverage", ylab = "Cook's Influence Measure",
     pch = 21, bg = clrs_log, cex = cex, cex.axis = cex)
p_log <- length(coef(Mstep_log))
n_log <- nrow(strikes1)
hbar_log <- p_log/n_log
abline(v = 2*hbar_log, col = "grey60", lty = 2)
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21,
       pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)

# double log model
D_log_d <- cooks.distance(MDstep_log_d)
```

```
infl.ind_log_d <- which.max(D_log_d)
lev.ind_log_d <- h_log_d > 2*hbar_log_d
clrs_log_d <- rep("black", len = n_log_d)
clrs_log_d[lev.ind_log_d] <- "blue"
clrs_log_d[infl.ind_log_d] <- "red"
par(mar = c(3,3,1,1))
cex <- .6
plot(h_log_d, D_log_d, xlab = "Leverage", ylab = "Cook's Influence Measure",
     pch = 21, bg = clrs_log_d, cex = cex, cex.axis = cex)
p_log_d <- length(coef(Mstep_log_d))
n_log_d <- nrow(strikes1)
hbar_log_d <- p_log_d/n_log_d
abline(v = 2*hbar_log_d, col = "grey60", lty = 2)
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21,
       pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)

# Residual plots & QQ plots
# log model
par(mfrow=c(2,2))
plot(predict(MDstep_log),residuals(MDstep_log),xlab="Fitted Values",
     ylab="Residuals", main="Model 1 Residuals vs Fitted Values", pch=16, cex=0.7)
abline(h=0, col="red",lty=2)

qqnorm(rstudent(MDstep_log), main="Model 1 QQ plot", pch=16, cex=.7)
qqline(y=rstudent(MDstep_log), col="red",lty=2)

# double log model
plot(predict(MDstep_log_d),residuals(MDstep_log_d),xlab="Fitted Values",
     ylab="Residuals", main="Model 2 Residuals vs Fitted Values", pch=16, cex=0.7)
abline(h=0, col="red",lty=2)

qqnorm(rstudent(MDstep_log_d), main="Model 2 QQ plot", pch=16, cex=.7)
qqline(y=rstudent(MDstep_log_d), col="red",lty=2)
```