Exploration

Jiayi Lily Ma (jm4303)

5/11/2019

```
#detach("package:dplyr", unload=TRUE)
library(dyn)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(here)
```

```
## here() starts at /Users/linatian/Desktop/msd final project/Sales
```

```
source("oosf_exploration.R")

#original data from paper
dat <- read.csv("data/merged.csv")

min.model <- lm(dat$sales~1,data = dat)
full.model <- formula(lm(sales~suvs+insurance, data=dat))
step(min.model,scope = full.model,direction = c("forward"))
```

```
## Start:  AIC=1668.48
## dat$sales ~ 1
##
##             Df  Sum of Sq        RSS    AIC
## + insurance  1 1674106800 6497005629 1649.6
## + suvs       1 1081567972 7089544457 1657.6
## <none>                    8171112429 1668.5
##
## Step:  AIC=1649.62
## dat$sales ~ insurance
##
##        Df Sum of Sq        RSS    AIC
## + suvs  1 193845195 6303160433 1648.9
## <none>              6497005629 1649.6
##
## Step:  AIC=1648.86
## dat$sales ~ insurance + suvs
```

```
##
## Call:
## lm(formula = dat$sales ~ insurance + suvs, data = dat)
##
## Coefficients:
## (Intercept)    insurance         suvs
##       80246        43133        31072
```

Insurance predictor doesn't seem to be very useful on paper's data as AIC decreased by less than 1.

```
#we try the same thing on our data: 2004-2008
merged0408 <- read.csv("data/merged_04_08.csv")
min.model <- lm(merged0408$sales~1,data = merged0408)
full.model <- formula(lm(sales~suvs+insurance, data=merged0408))
step(min.model,scope = full.model,direction = c("forward"))
```

```
## Start:  AIC=1083.5
## merged0408$sales ~ 1
##
##             Df Sum of Sq        RSS    AIC
## + suvs       1 687090658 3351919224 1074.3
## + insurance  1 630296149 3408713732 1075.3
## <none>                   4039009881 1083.5
##
## Step:  AIC=1074.31
## merged0408$sales ~ suvs
##
##             Df Sum of Sq        RSS    AIC
## <none>                   3351919224 1074.3
## + insurance  1  11884583 3340034641 1076.1
```

```
##
## Call:
## lm(formula = merged0408$sales ~ suvs, data = merged0408)
##
## Coefficients:
## (Intercept)         suvs
##     46384.5        379.8
```

In this case, insurance isn't useful at all.

Since insurance doesn't seem to be a useful predictor, we'll try a model without it.

We try:

$$y = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-1}^2 + \beta_3 y_{t-12} + \beta_4 suvs$$

```
getModelandEvaluation2 <- function(data, print=1, dateBegin=NULL, dateEnd=NULL) {
  zooObj <- zoo(data[,-1], as.Date(data[,1]))
  y <- log(zooObj$sales)
  x <- zooObj[, "suvs"]

  reg0 <- dyn$lm(y~lag(y,-1) + I(lag(y, -1)^2) + lag(y, -12))
  reg1 <- dyn$lm(y~lag(y,-1)+I(lag(y, -1)^2)+lag(y,-12)+suvs,data=data)

  if(print == 1){
    print(summary(reg0))
    print(summary(reg1))
  }

  z <- OutOfSampleForecast12_newModel(y,x,17)
  print(MaeReport(z))

  if(!is.null(dateBegin) && !is.null(dateEnd)){
    print("dates not null")
    print(MaeReport(z,dateBegin, dateEnd))
```

```
  }
}
```

Paper's data, new model

```
getModelandEvaluation2(dat, dateBegin = "2007-12-01", dateEnd = "2009-06-30")
```

```
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12)))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.215883 -0.037117  0.005776  0.037219  0.232368
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -41.69352   44.76012  -0.931    0.355
## lag(y, -1)        8.26571    8.05212   1.027    0.308
## I(lag(y, -1)^2)  -0.34386    0.36324  -0.947    0.347
## lag(y, -12)       0.30772    0.07398   4.160  8.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0799 on 75 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7218, Adjusted R-squared:  0.7107
## F-statistic: 64.88 on 3 and 75 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12) +
##     suvs), data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168576 -0.037237  0.003466  0.040400  0.179618
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -47.76507   38.94430  -1.226    0.224
## lag(y, -1)        9.37266    7.00597   1.338    0.185
## I(lag(y, -1)^2)  -0.39723    0.31607  -1.257    0.213
## lag(y, -12)       0.35177    0.06493   5.418 7.21e-07 ***
## suvs              0.78229    0.15593   5.017 3.50e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06949 on 74 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7924, Adjusted R-squared:  0.7812
```

```
## F-statistic: 70.63 on 4 and 74 DF,  p-value: < 2.2e-16
```


```
##   mae.base mae.trends  mae.delta
## 0.06730051 0.06149628 0.08624353
## [1] "dates not null"
```


```
##   mae.base mae.trends  mae.delta
## 0.09767965 0.07496843 0.23250718
```

First, we observe that y_{t-12} is the only significant predictor in the base model, compared to the paper's base model $y = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12}$ where all the predictors except the intercept are significant.

The second observation is that in the model with suv Trend data, only y_{t-12} and *suvs* are significant.

Using this new model on the paper's data, we see that the overall MAE decreased by about 8.6% with the addition of Trend data; this is about 2% less than the model that the paper use. However, the MAE for the Recession years between "2007-12-01" and "2009-06-30" decreased by about 23%, which is 2% higher than the decrease of 21% the paper's model.

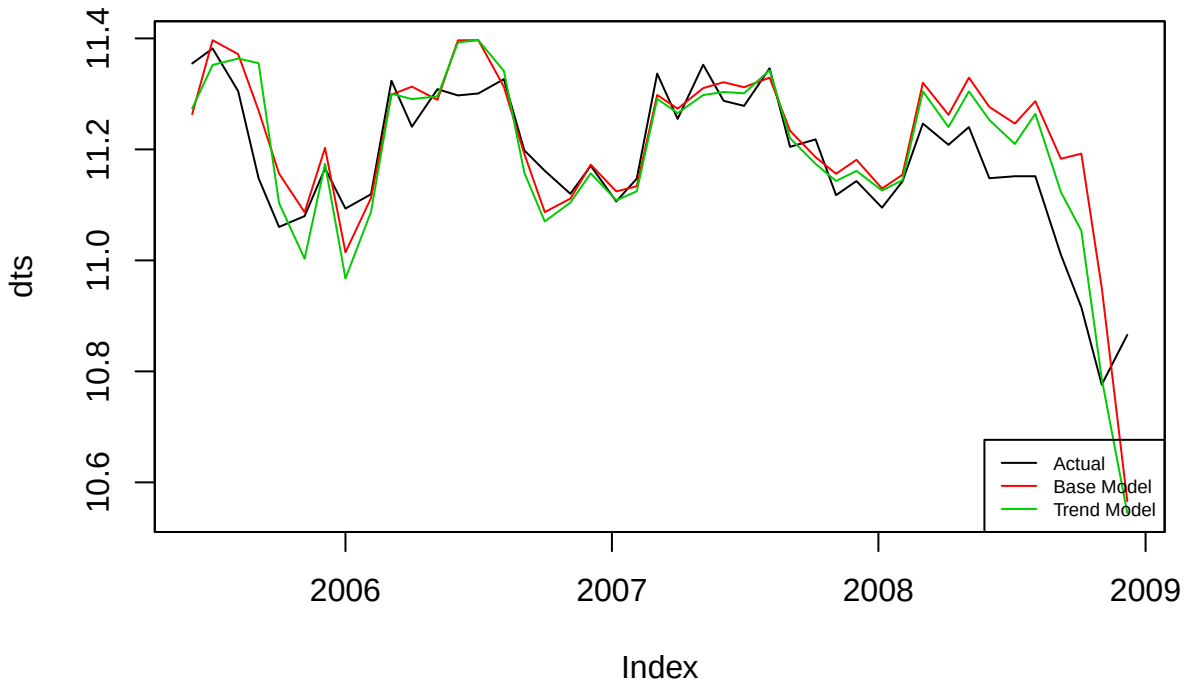
Our data: 2004-2008, new model

Note that this is the time period in which we were able to replicate the paper's results most closely.

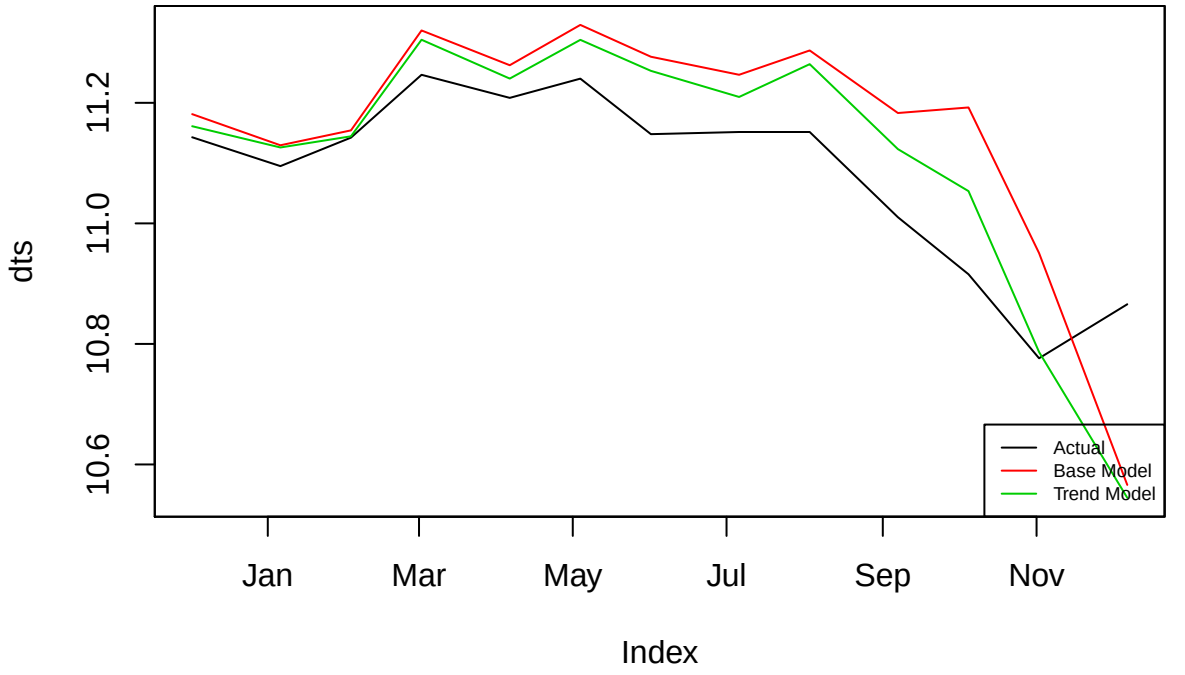
```
getModelandEvaluation2(merged0408,  dateBegin = "2007-12-02", dateEnd = "2008-12-07")
```

```
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16412 -0.03946  0.01090  0.04613  0.11954
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -109.3014    56.6102  -1.931   0.0600 .
## lag(y, -1)        19.8919    10.1408   1.962   0.0562 .
## I(lag(y, -1)^2)   -0.8659     0.4555  -1.901   0.0639 .
## lag(y, -12)        0.5643     0.1232   4.579 3.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07263 on 44 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7005, Adjusted R-squared:  0.6801
## F-statistic:  34.3 on 3 and 44 DF,  p-value: 1.375e-11
##
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12) +
##     suvs), data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.127084 -0.038662  0.009153  0.033421  0.093417
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -92.970976  44.446845  -2.092   0.0424 *
## lag(y, -1)       16.844466   7.963541   2.115   0.0402 *
## I(lag(y, -1)^2)  -0.737636   0.357600  -2.063   0.0452 *
## lag(y, -12)       0.679259   0.098880   6.870 1.99e-08 ***
## suvs              0.006417   0.001198   5.359 3.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05689 on 43 degrees of freedom
```

```
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.8204, Adjusted R-squared:  0.8037
## F-statistic: 49.11 on 4 and 43 DF,  p-value: 1.725e-15
```



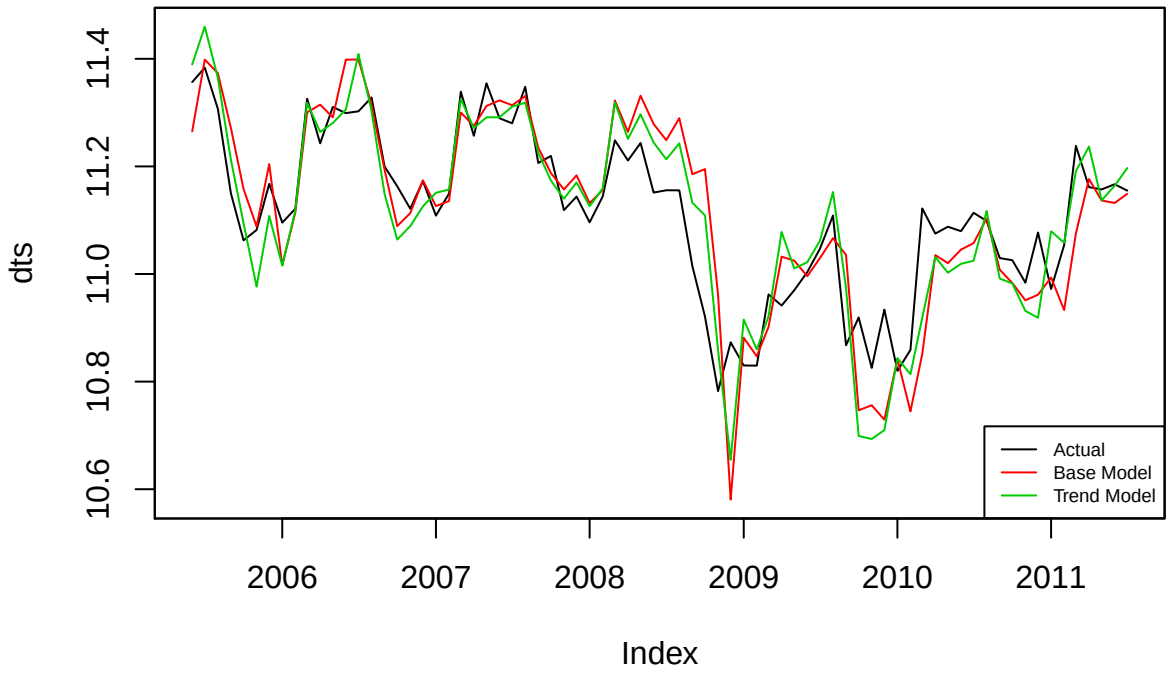

```
##   mae.base mae.trends  mae.delta
## 0.06596471 0.05681596 0.13869151
## [1] "dates not null"
```



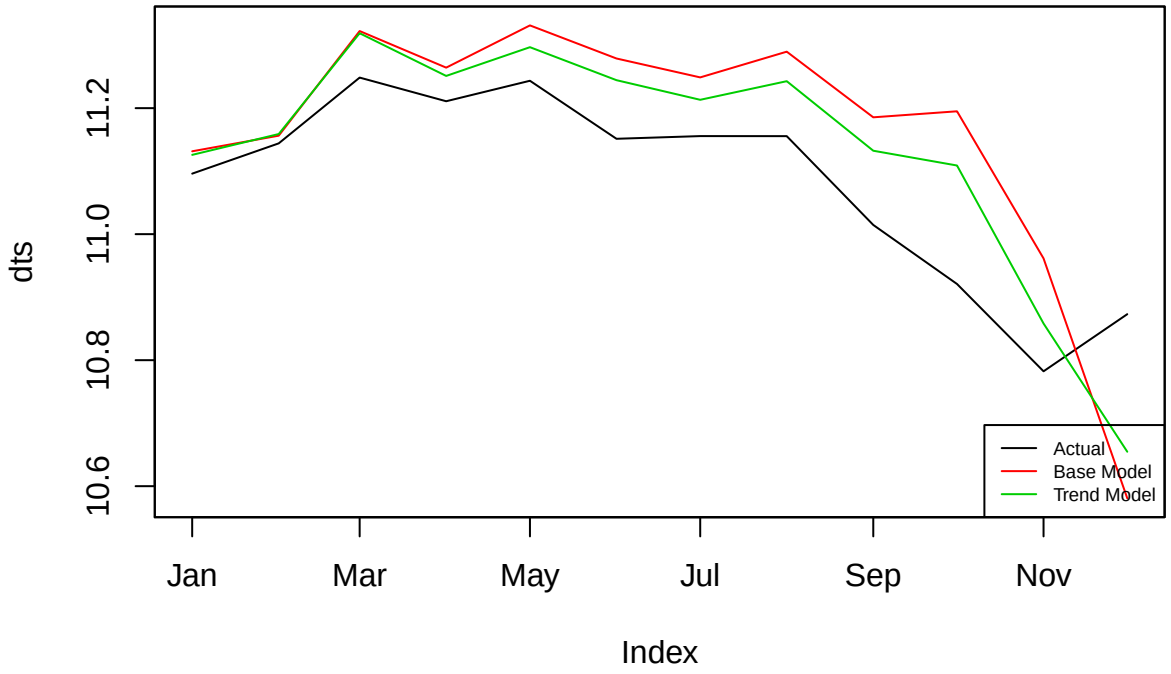

```
##   mae.base mae.trends  mae.delta
## 0.12187361 0.08184502 0.32844344
```

```r
getModelandEvaluation2(dat, dateBegin = "2007-12-02", dateEnd = "2008-12-07")
```

```
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12)))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.215883 -0.037117  0.005776  0.037219  0.232368
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -41.69352   44.76012  -0.931    0.355
## lag(y, -1)        8.26571    8.05212   1.027    0.308
## I(lag(y, -1)^2)  -0.34386    0.36324  -0.947    0.347
## lag(y, -12)       0.30772    0.07398   4.160  8.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0799 on 75 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7218, Adjusted R-squared:  0.7107
## F-statistic: 64.88 on 3 and 75 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12) +
##     suvs), data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168576 -0.037237  0.003466  0.040400  0.179618
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -47.76507   38.94430  -1.226    0.224
## lag(y, -1)        9.37266    7.00597   1.338    0.185
## I(lag(y, -1)^2)  -0.39723    0.31607  -1.257    0.213
## lag(y, -12)       0.35177    0.06493   5.418 7.21e-07 ***
## suvs              0.78229    0.15593   5.017 3.50e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06949 on 74 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7924, Adjusted R-squared:  0.7812
## F-statistic: 70.63 on 4 and 74 DF,  p-value: < 2.2e-16
```




```
##   mae.base mae.trends  mae.delta
## 0.06730051 0.06149628 0.08624353
## [1] "dates not null"
```




```
##   mae.base mae.trends  mae.delta
## 0.12780393 0.08720023 0.31770305
```

This model improved overall MAE more than the previous model on our data (~13% vs ~10%).

Also, we see that in the same time period (in the midst of Recession), this new model gives a better MAE improvement than the original model, for both the paper's data and our data (~32% vs ~28%).

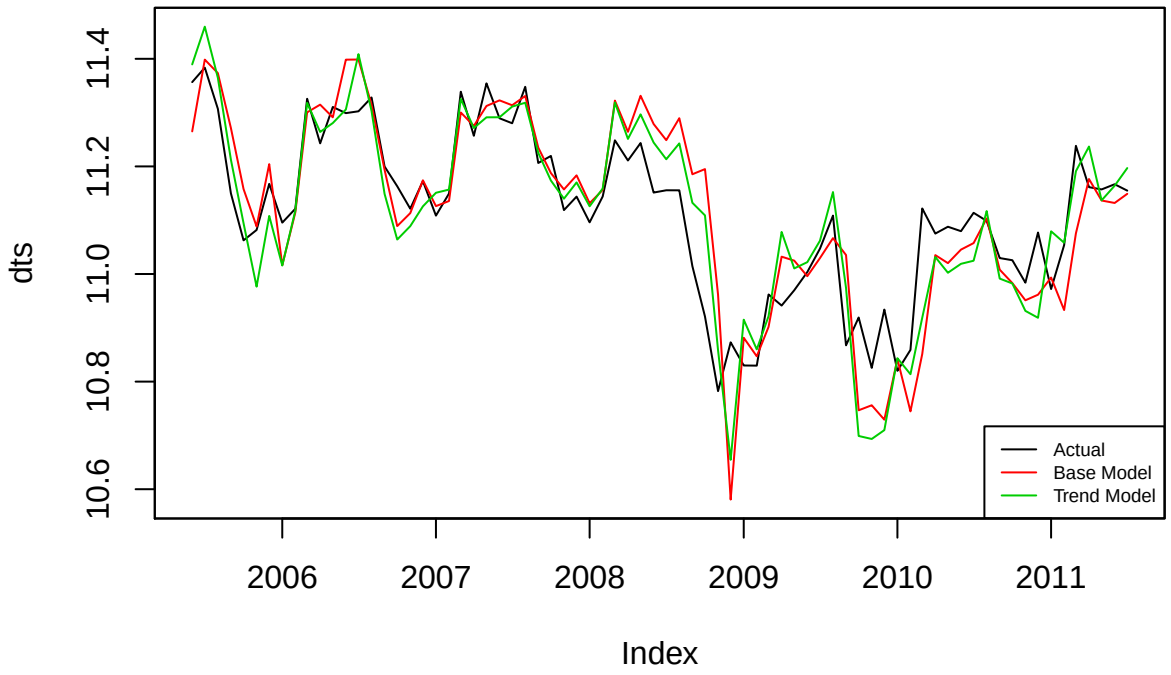
Since our investigation with the paper's model led us to suspect that Trend data helps improve MAE only in

period of major economic turbulence, we check this speculation here with the new model by looking at the paper's data during non-recession years. We repeat that with our data too.

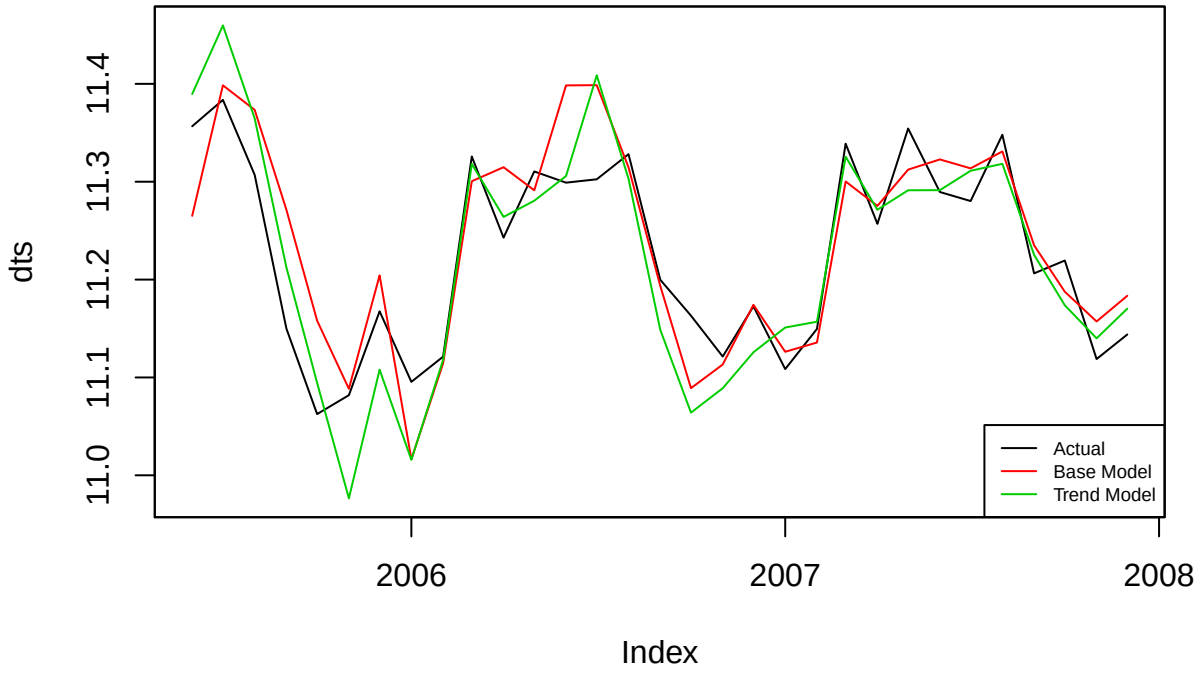
Non-recession years

```
getModelandEvaluation2(dat,  dateBegin = "2004-01-01", dateEnd = "2007-12-01")
```

```
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12)))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.215883 -0.037117  0.005776  0.037219  0.232368
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -41.69352   44.76012  -0.931    0.355
## lag(y, -1)        8.26571    8.05212   1.027    0.308
## I(lag(y, -1)^2)  -0.34386    0.36324  -0.947    0.347
## lag(y, -12)       0.30772    0.07398   4.160  8.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0799 on 75 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7218, Adjusted R-squared:  0.7107
## F-statistic: 64.88 on 3 and 75 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12) +
##     suvs), data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168576 -0.037237  0.003466  0.040400  0.179618
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -47.76507   38.94430  -1.226    0.224
## lag(y, -1)        9.37266    7.00597   1.338    0.185
## I(lag(y, -1)^2)  -0.39723    0.31607  -1.257    0.213
## lag(y, -12)       0.35177    0.06493   5.418 7.21e-07 ***
## suvs              0.78229    0.15593   5.017 3.50e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06949 on 74 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7924, Adjusted R-squared:  0.7812
## F-statistic: 70.63 on 4 and 74 DF,  p-value: < 2.2e-16
```




```
##   mae.base mae.trends  mae.delta
## 0.06730051 0.06149628 0.08624353
## [1] "dates not null"
```




```
##   mae.base mae.trends  mae.delta
## 0.04150881 0.04033344 0.02831624
```

```
getModelandEvaluation2(merged0408,  dateBegin = "2004-01-04", dateEnd = "2007-12-02")
```

```
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12)))
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16412 -0.03946  0.01090  0.04613  0.11954
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -109.3014    56.6102  -1.931   0.0600 .
## lag(y, -1)        19.8919    10.1408   1.962   0.0562 .
## I(lag(y, -1)^2)   -0.8659     0.4555  -1.901   0.0639 .
## lag(y, -12)        0.5643     0.1232   4.579 3.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07263 on 44 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7005, Adjusted R-squared:  0.6801
## F-statistic:  34.3 on 3 and 44 DF,  p-value: 1.375e-11
##
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12) +
##     suvs), data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.127084 -0.038662  0.009153  0.033421  0.093417
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -92.970976  44.446845  -2.092   0.0424 *
## lag(y, -1)       16.844466   7.963541   2.115   0.0402 *
## I(lag(y, -1)^2)  -0.737636   0.357600  -2.063   0.0452 *
## lag(y, -12)       0.679259   0.098880   6.870 1.99e-08 ***
## suvs              0.006417   0.001198   5.359 3.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05689 on 43 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.8204, Adjusted R-squared:  0.8037
## F-statistic: 49.11 on 4 and 43 DF,  p-value: 1.725e-15
```


```
##   mae.base mae.trends  mae.delta
## 0.06596471 0.05681596 0.13869151
## [1] "dates not null"
```


```
##    mae.base  mae.trends   mae.delta
##  0.04162837  0.04507837 -0.08287616
```

On the paper's data, we see that Trend data helps very little during non-recession years; MAE improved by about 2.8% (compared to the ~10% overall and ~21% during Recession.) On our data, Trend data didn't decrease MAE but rather increased it.

Here we rescale the Trend data using $log(x/100)$ to see if the same holds:

```r
rescaled_merged0408 <- merged0408
rescaled_merged0408$suvs <- log(merged0408$suvs/100)
rescaled_merged0408$insurance <- log(merged0408$insurance/100)

getModelandEvaluation2(rescaled_merged0408,  dateBegin = "2004-01-04", dateEnd = "2007-12-02")
```

```
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16412 -0.03946  0.01090  0.04613  0.11954
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -109.3014    56.6102  -1.931   0.0600 .
## lag(y, -1)        19.8919    10.1408   1.962   0.0562 .
## I(lag(y, -1)^2)   -0.8659     0.4555  -1.901   0.0639 .
## lag(y, -12)        0.5643     0.1232   4.579 3.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07263 on 44 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7005, Adjusted R-squared:  0.6801
## F-statistic:  34.3 on 3 and 44 DF,  p-value: 1.375e-11
##
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + I(lag(y, -1)^2) + lag(y, -12) +
##     suvs), data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.123103 -0.037730  0.008395  0.029270  0.093148
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -84.48354   43.02092  -1.964   0.0560 .
## lag(y, -1)       15.43969    7.70662   2.003   0.0515 .
## I(lag(y, -1)^2)  -0.67578    0.34601  -1.953   0.0573 .
## lag(y, -12)       0.68811    0.09558   7.199 6.63e-09 ***
## suvs              0.44867    0.07701   5.826 6.55e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05493 on 43 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.8326, Adjusted R-squared:  0.8171
## F-statistic: 53.48 on 4 and 43 DF,  p-value: 3.855e-16
```


```
##   mae.base mae.trends  mae.delta
## 0.06596471 0.05538415 0.16039727
## [1] "dates not null"
```


```
##    mae.base  mae.trends   mae.delta
##  0.04162837  0.04529120 -0.08798859
```

Conclusion: This new model seems to be able to improve MAE more than the paper's model for economic turbulent period.

Future exploration:

Looking at another Google Trend category–Currencies

Since we suspect Trend data helps only in time period of major economic changes, we want to check if 'currencies' is a useful predictor since it's a major indicator of economic activities.

```
merged0408_c <- read.csv("data/merged_04_08_c.csv")

#we only look at the Recession period from 2007/06/03 - 2008/12/07 (since our data only goes up to 2008
row <- which(merged0408_c == "2007/06/03")
small_merged <- merged0408_c[row:nrow(merged0408_c),]

min.model <- lm(sales~1,data = small_merged)
full.model <- formula(lm(sales~suvs+insurance+currencies, data=small_merged))
step(min.model,scope = full.model,direction = c("forward"))
```

```
## Start:  AIC=349.31
## sales ~ 1
##
##              Df  Sum of Sq        RSS    AIC
## + insurance   1 1203416841  445971872 326.46
## + suvs        1  967083916  682304797 334.53
## + currencies  1  529966068 1119422644 343.94
## <none>                     1649388713 349.31
##
## Step:  AIC=326.46
## sales ~ insurance
##
##              Df Sum of Sq       RSS    AIC
## + currencies  1 100481625 345490246 323.60
## <none>                    445971872 326.46
## + suvs        1  19345145 426626727 327.61
##
## Step:  AIC=323.6
## sales ~ insurance + currencies
##
##        Df Sum of Sq       RSS    AIC
## <none>              345490246 323.60
## + suvs  1   4985851 340504395 325.33
```

```
##
## Call:
## lm(formula = sales ~ insurance + currencies, data = small_merged)
##
## Coefficients:
## (Intercept)    insurance   currencies
##       29067         1063         -328
```

```
#all 2004-2008 years
min.model <- lm(sales~1,data = merged0408_c)
full.model <- formula(lm(sales~suvs+insurance+currencies, data=merged0408_c))
step(min.model,scope = full.model,direction = c("forward"))
```

```
## Start:  AIC=1083.5
```

```
## sales ~ 1
##
##              Df Sum of Sq        RSS    AIC
## + suvs        1 687090658 3351919224 1074.3
## + insurance   1 630296149 3408713732 1075.3
## <none>                    4039009881 1083.5
## + currencies  1    151856 4038858025 1085.5
##
## Step:  AIC=1074.31
## sales ~ suvs
##
##              Df Sum of Sq        RSS    AIC
## + currencies  1 645598684 2706320540 1063.5
## <none>                    3351919224 1074.3
## + insurance   1  11884583 3340034641 1076.1
##
## Step:  AIC=1063.47
## sales ~ suvs + currencies
##
##             Df Sum of Sq        RSS    AIC
## + insurance  1  88992924 2617327616 1063.5
## <none>                   2706320540 1063.5
##
## Step:  AIC=1063.46
## sales ~ suvs + currencies + insurance
```

```
##
## Call:
## lm(formula = sales ~ suvs + currencies + insurance, data = merged0408_c)
##
## Coefficients:
## (Intercept)         suvs   currencies    insurance
##     53762.7        485.1       -436.9        241.5
```

'currencies' and 'insurance' seem to be a useful predictor in Recession years where as 'suvs' is not; this makes sense because during Recession, people don't have extra money to buy cars. But during normal times, three predictors seem to be useful in predicting sales.