Forecasting revenue with time series models

Geoffery Mullings

Housekeeping: Let's grab the data...

**require**(tseries)

## Loading required package: tseries

**require**(zoo)

## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

**require**(dyn)

## Loading required package: dyn

file <- 'http://aux.zicklin.baruch.cuny.edu/manzan/eco9723/files/EPS\_REV\_SPRING2015.csv'  
data <- **read.csv**(file, header=TRUE)  
mytick <- "F"  
index <- **which**(data[, "tic"] == mytick)  
mydata <- data[index,]  
**head**(mydata)

## datadate tic datacqtr epsfxq revtq  
## 2584 3/31/90 F 1990Q1 1.09 23638.1  
## 2585 6/30/90 F 1990Q2 1.66 26834.2  
## 2586 9/30/90 F 1990Q3 0.22 22986.9  
## 2587 12/31/90 F 1990Q4 -1.12 24190.8  
## 2588 3/31/91 F 1991Q1 -1.86 21341.0  
## 2589 6/30/91 F 1991Q2 -0.68 23846.7

**tail**(mydata)

## datadate tic datacqtr epsfxq revtq  
## 2677 6/30/13 F 2013Q2 0.30 37923  
## 2678 9/30/13 F 2013Q3 0.31 35775  
## 2679 12/31/13 F 2013Q4 0.74 37036  
## 2680 3/31/14 F 2014Q1 0.24 35876  
## 2681 6/30/14 F 2014Q2 0.32 37411  
## 2682 9/30/14 F 2014Q3 0.21 34920

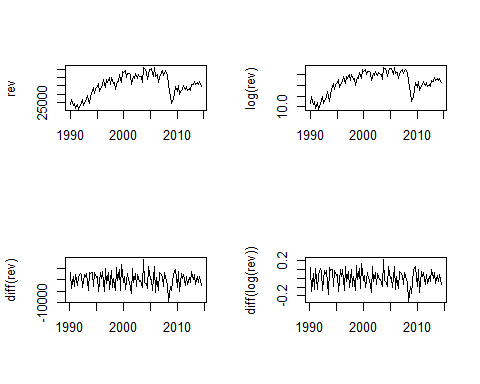
Now to create the trend, log of revenue, and seasonal dummy variables that will be needed.

startdate <- mydata[1, "datacqtr"]  
rev <- **zooreg**(mydata[, "revtq"], start=**as.yearqtr**(startdate), frequency=4)  
lrev <- **log**(rev)  
trend <- **zooreg**(1:**length**(rev), start=**as.yearqtr**(startdate), frequency=4)  
trendsq <- trend^2  
Q1 <- **zooreg**(**as.numeric**(**cycle**(rev) ==1), start=**start**(rev), frequency=4)  
Q2 <- **zooreg**(**as.numeric**(**cycle**(rev) ==2), start=**start**(rev), frequency=4)  
Q3 <- **zooreg**(**as.numeric**(**cycle**(rev) ==3), start=**start**(rev), frequency=4)  
Q4 <- **zooreg**(**as.numeric**(**cycle**(rev) ==4), start=**start**(rev), frequency=4)

Section 1: To Log or Not To log?

In general using log data seems to be best for consistency, so we will be using it regardless. But the following is to explore whether log data was needed

**par**(mfrow=**c**(2,2))  
**plot**(rev,xlab="")  
**plot**(**log**(rev), xlab="")  
**plot**(**diff**(rev), xlab="")  
**plot**(**diff**(**log**(rev)), xlab="")

 The shape of the data hardly seems to change between log and standard. Still, to be consistent, we will stick with log data.

Section 2: Nonstationary or Stationary?

If the t statistic is greater than -2.89 without a trend, or greater than -3.45 with a trend then we accept the null hypothesis that the revenue data is non-stationary.

dlrev <- **diff**(**log**(rev))  
adffit <- dyn$**lm**(dlrev ~ **lag**(rev, -1) + **lag**(dlrev, -1:-4) + trend + Q2 + Q3 + Q4)  
**summary**(adffit)

##   
## Call:  
## lm(formula = dyn(dlrev ~ lag(rev, -1) + lag(dlrev, -1:-4) + trend +   
## Q2 + Q3 + Q4))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.173413 -0.030714 0.000898 0.037447 0.161238   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.306e-02 4.415e-02 0.976 0.33211   
## lag(rev, -1) -2.584e-06 1.090e-06 -2.372 0.01999 \*   
## lag(dlrev, -1:-4)1 2.313e-01 1.042e-01 2.220 0.02914 \*   
## lag(dlrev, -1:-4)2 -1.753e-01 1.072e-01 -1.635 0.10570   
## lag(dlrev, -1:-4)3 1.545e-02 1.057e-01 0.146 0.88414   
## lag(dlrev, -1:-4)4 -4.689e-02 1.030e-01 -0.455 0.65006   
## trend -1.059e-04 2.436e-04 -0.435 0.66477   
## Q2 1.302e-01 4.111e-02 3.168 0.00214 \*\*   
## Q3 -6.797e-02 2.180e-02 -3.118 0.00250 \*\*   
## Q4 1.834e-01 3.971e-02 4.617 1.39e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05802 on 84 degrees of freedom  
## (9 observations deleted due to missingness)  
## Multiple R-squared: 0.6773, Adjusted R-squared: 0.6427   
## F-statistic: 19.59 on 9 and 84 DF, p-value: < 2.2e-16

At -2.372 and even at -.435 with a trend, we do not reject the null hypothesis. The data seems non-stationary, invalidating the use of a trend-stationary model.

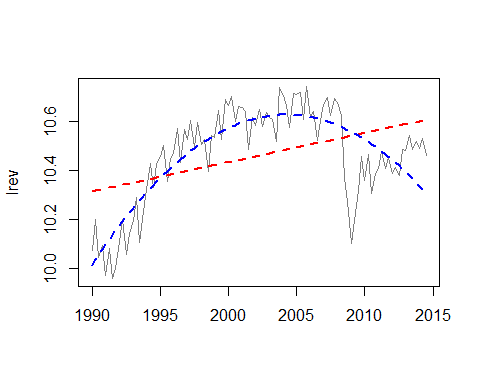
Section 3: Linear or Quadratic Trend?

Let's run that ADF test again with a quadratic trend, and plot revenue with linear and quadratic trend lines.

adffitquad <- dyn$**lm**(dlrev ~ **lag**(rev, -1) + **lag**(dlrev, -1:-4) + trend + trendsq + Q2 + Q3 + Q4)  
**summary**(adffitquad)

##   
## Call:  
## lm(formula = dyn(dlrev ~ lag(rev, -1) + lag(dlrev, -1:-4) + trend +   
## trendsq + Q2 + Q3 + Q4))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.173044 -0.030200 0.000475 0.034490 0.146644   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.550e-02 5.009e-02 1.507 0.13556   
## lag(rev, -1) -5.150e-06 2.191e-06 -2.351 0.02111 \*   
## lag(dlrev, -1:-4)1 2.818e-01 1.103e-01 2.556 0.01242 \*   
## lag(dlrev, -1:-4)2 -1.190e-01 1.146e-01 -1.038 0.30213   
## lag(dlrev, -1:-4)3 5.075e-02 1.084e-01 0.468 0.64085   
## lag(dlrev, -1:-4)4 3.912e-03 1.092e-01 0.036 0.97151   
## trend 2.733e-03 2.120e-03 1.289 0.20099   
## trendsq -2.481e-05 1.841e-05 -1.348 0.18140   
## Q2 1.251e-01 4.109e-02 3.045 0.00312 \*\*   
## Q3 -6.518e-02 2.180e-02 -2.990 0.00367 \*\*   
## Q4 1.726e-01 4.033e-02 4.279 4.99e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05774 on 83 degrees of freedom  
## (9 observations deleted due to missingness)  
## Multiple R-squared: 0.6842, Adjusted R-squared: 0.6462   
## F-statistic: 17.98 on 10 and 83 DF, p-value: < 2.2e-16

fitlin <- dyn$**lm**(lrev ~ trend)  
fitquad <- dyn$**lm**(lrev ~ trend + trendsq)  
**plot**(lrev, xlab="", col="gray50")  
**lines**( **fitted**(fitlin) ,col=2,lwd=2,lty=2)  
**lines**( **fitted**(fitquad) ,col=4,lwd=2,lty=2)



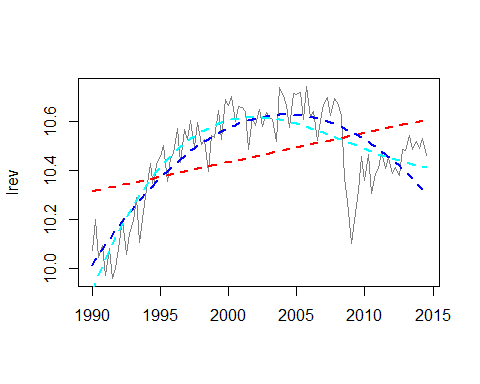
At -2.351 and even at -1.348 with a quadratic trend, we still do not reject the null hypothesis. The data still seems non-stationary, invalidating the use of a trend-stationary model. According to the

Would a cubic trend help?

trendcub <- trend^3  
adffitcub <- dyn$**lm**(dlrev ~ **lag**(rev, -1) + **lag**(dlrev, -1:-4) + trend + trendsq + trendcub + Q2 + Q3 + Q4)  
**summary**(adffitcub)

##   
## Call:  
## lm(formula = dyn(dlrev ~ lag(rev, -1) + lag(dlrev, -1:-4) + trend +   
## trendsq + trendcub + Q2 + Q3 + Q4))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.165459 -0.032868 0.000257 0.031050 0.138515   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.509e-02 5.172e-02 0.872 0.38580   
## lag(rev, -1) -6.712e-06 2.301e-06 -2.917 0.00455 \*\*   
## lag(dlrev, -1:-4)1 2.824e-01 1.085e-01 2.603 0.01095 \*   
## lag(dlrev, -1:-4)2 -1.112e-01 1.128e-01 -0.985 0.32729   
## lag(dlrev, -1:-4)3 4.405e-02 1.067e-01 0.413 0.68074   
## lag(dlrev, -1:-4)4 -5.832e-03 1.076e-01 -0.054 0.95689   
## trend 9.596e-03 4.108e-03 2.336 0.02194 \*   
## trendsq -1.629e-04 7.347e-05 -2.217 0.02939 \*   
## trendcub 8.079e-07 4.166e-07 1.939 0.05592 .   
## Q2 1.227e-01 4.044e-02 3.034 0.00323 \*\*   
## Q3 -6.530e-02 2.144e-02 -3.045 0.00312 \*\*   
## Q4 1.674e-01 3.976e-02 4.211 6.49e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0568 on 82 degrees of freedom  
## (9 observations deleted due to missingness)  
## Multiple R-squared: 0.6981, Adjusted R-squared: 0.6576   
## F-statistic: 17.23 on 11 and 82 DF, p-value: < 2.2e-16

fitlin <- dyn$**lm**(lrev ~ trend)  
fitquad <- dyn$**lm**(lrev ~ trend + trendsq)  
fitcub <- dyn$**lm**(lrev ~ trend + trendsq + trendcub)  
**plot**(lrev, xlab="", col="gray50")  
**lines**( **fitted**(fitlin) ,col=2,lwd=2,lty=2)  
**lines**( **fitted**(fitquad) ,col=4,lwd=2,lty=2)  
**lines**( **fitted**(fitcub) ,col=5,lwd=2,lty=2)



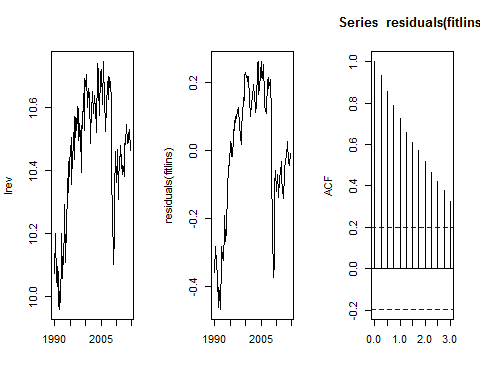
Judging by the R squared coefficient of 0.6576, highest of all three trends tested, a cubic trend may be most appropriate.

Section 4: Seasonal or Not Seasonal?

fitlins <- dyn$**lm**(lrev ~ trendcub + Q2 + Q3 +Q4)  
**summary**(fitlins)

##   
## Call:  
## lm(formula = dyn(lrev ~ trendcub + Q2 + Q3 + Q4))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.46629 -0.10402 0.02911 0.15761 0.26208   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.043e+01 4.306e-02 242.154 <2e-16 \*\*\*  
## trendcub 9.876e-08 7.115e-08 1.388 0.168   
## Q2 5.195e-02 5.591e-02 0.929 0.355   
## Q3 -5.665e-02 5.593e-02 -1.013 0.314   
## Q4 3.634e-02 5.649e-02 0.643 0.522   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1977 on 94 degrees of freedom  
## Multiple R-squared: 0.06249, Adjusted R-squared: 0.0226   
## F-statistic: 1.566 on 4 and 94 DF, p-value: 0.1897

**par**(mfrow=**c**(1,3))  
**plot**(lrev, xlab="")  
**plot**(**residuals**(fitlins), xlab="")  
**acf**(**residuals**(fitlins), lag=12, xlab="")



The dummy variable coefficients do not seem statistically significant and do not suggest the need for seasonal adjustment. But there is a need for an auto-regressive function now as too many of the residuals are auto-correlated (systematically under-predicting the variable).

Section 4: Auto-Regressive?

fitlins <- dyn$**lm**(lrev ~ trendcub + Q2 + Q3 + Q4)  
resid <- **residuals**(fitlins)  
fitresid <- **ar**(resid, aic=TRUE, order.max=8, demean=FALSE, method="ols")  
ord <- 1:fitresid$order  
ord

## [1] 1 2 3

fitlinsar <- dyn$**lm**(lrev ~ **lag**(lrev, -ord) + trend + trendcub + Q2 + Q3 + Q4)  
**summary**(fitlinsar)

##   
## Call:  
## lm(formula = dyn(lrev ~ lag(lrev, -ord) + trend + trendcub +   
## Q2 + Q3 + Q4))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.168293 -0.028189 0.006048 0.036121 0.149399   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.196e+00 5.902e-01 2.027 0.045721 \*   
## lag(lrev, -ord)1 1.162e+00 1.066e-01 10.902 < 2e-16 \*\*\*  
## lag(lrev, -ord)2 -4.092e-01 1.564e-01 -2.616 0.010476 \*   
## lag(lrev, -ord)3 1.243e-01 1.087e-01 1.143 0.256207   
## trend 1.156e-03 1.032e-03 1.120 0.265827   
## trendcub -1.054e-07 9.219e-08 -1.143 0.256051   
## Q2 1.255e-01 3.117e-02 4.025 0.000121 \*\*\*  
## Q3 -5.815e-02 2.011e-02 -2.891 0.004844 \*\*   
## Q4 1.808e-01 3.389e-02 5.333 7.54e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05883 on 87 degrees of freedom  
## (6 observations deleted due to missingness)  
## Multiple R-squared: 0.9143, Adjusted R-squared: 0.9064   
## F-statistic: 116 on 8 and 87 DF, p-value: < 2.2e-16

It seems that the third lag is not significant at 10% (it is below 1.6626). We'll drop it now to improve the model.

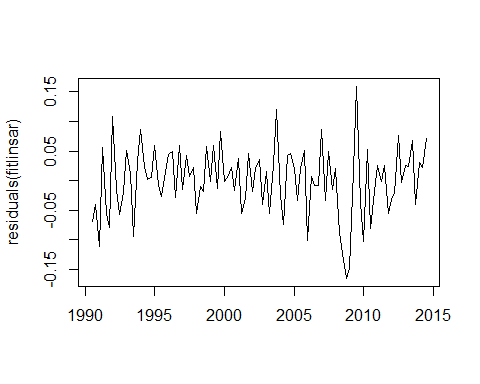
ord <- ord[-3]  
ord

## [1] 1 2

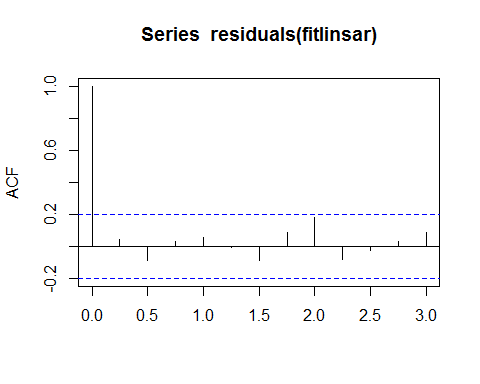
fitlinsar <- dyn$**lm**(lrev ~ **lag**(lrev, -ord) + trend + trendcub + Q2 + Q3 + Q4)  
**summary**(fitlinsar)

##   
## Call:  
## lm(formula = dyn(lrev ~ lag(lrev, -ord) + trend + trendcub +   
## Q2 + Q3 + Q4))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.16477 -0.03250 0.00461 0.03707 0.15898   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.465e+00 5.473e-01 2.677 0.00885 \*\*   
## lag(lrev, -ord)1 1.116e+00 1.025e-01 10.885 < 2e-16 \*\*\*  
## lag(lrev, -ord)2 -2.648e-01 1.048e-01 -2.527 0.01326 \*   
## trend 1.745e-03 9.604e-04 1.817 0.07262 .   
## trendcub -1.530e-07 8.692e-08 -1.760 0.08178 .   
## Q2 9.733e-02 2.065e-02 4.713 8.97e-06 \*\*\*  
## Q3 -7.234e-02 1.762e-02 -4.105 8.96e-05 \*\*\*  
## Q4 1.542e-01 2.655e-02 5.807 9.68e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0591 on 89 degrees of freedom  
## (4 observations deleted due to missingness)  
## Multiple R-squared: 0.9158, Adjusted R-squared: 0.9092   
## F-statistic: 138.3 on 7 and 89 DF, p-value: < 2.2e-16

**plot**(**residuals**(fitlinsar), xlab="")



**acf**(**residuals**(fitlinsar), lag=12, xlab="")



The autocorrelation for all lags are now 0 at a 95% confidence interval, so the variable should be adequately predicted now by the model.

Section 5: Revenue Forecasts

file2 <- 'http://aux.zicklin.baruch.cuny.edu/manzan/eco9723/files/myforecast.txt'  
**source**(file2)  
myf <- **myforecast**(lrev, ord=**c**(1,2), n.ahead=5, trend="cubic", seasonal="no")  
myf

## 2014 Q4 2015 Q1 2015 Q2 2015 Q3 2015 Q4   
## 10.49207 10.48203 10.48875 10.48842 10.49083

Compared to analyst estimates: Avg. Estimate 34.18B (15Q1) 36.52B (15Q2) 143.51B (2015) / Low Estimate 31.13B (15Q1) 33.61B (15Q2) 134.30B (2015) / High Estimate 37.54B (15Q1) 38.82B (15Q2) 150.49B (2015)

My model's forecasts are far off, most likely for a few reasons: 1) Cubic trending alone isn't the greatest predictor of revenue at the Ford company since the 1990's, it just roughly accounts for the downturn in 2008, 2) Vehicle sales are probably a function of multiple variables, including oil prices, average income, and perhaps even the percentage of the population in urban areas, among others.