CLIMATE CHANGE

Anuranjan

24 September 2016

The file climate_change.csv contains climate data from May 1983 to December 2008.The available variables include:

Year: the observation year.
Month: the observation month.
Temp: the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.
CO2, N2O, CH4, CFC.11, CFC.12: atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division. CO2, N2O and CH4 are expressed in ppmv (parts per million by volume – i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere) CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).
Aerosols: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun’s energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA.
TSI: the total solar irradiance (TSI) in W/m2 (the rate at which the sun’s energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project website.
MEI: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

Reading CSV files.

```
climate = read.csv("climate_change.csv")
str(climate)
```

```
## 'data.frame':    308 obs. of  11 variables:
##  $ Year    : int  1983 1983 1983 1983 1983 1983 1983 1983 1984 1984 ...
##  $ Month   : int  5 6 7 8 9 10 11 12 1 2 ...
##  $ MEI     : num  2.556 2.167 1.741 1.13 0.428 ...
##  $ CO2     : num  346 346 344 342 340 ...
##  $ CH4     : num  1639 1634 1633 1631 1648 ...
##  $ N2O     : num  304 304 304 304 304 ...
##  $ CFC.11  : num  191 192 193 194 194 ...
##  $ CFC.12  : num  350 352 354 356 357 ...
##  $ TSI     : num  1366 1366 1366 1366 1366 ...
##  $ Aerosols: num  0.0863 0.0794 0.0731 0.0673 0.0619 0.0569 0.0524 0.0486 0.
0451 0.0416 ...
##  $ Temp    : num  0.109 0.118 0.137 0.176 0.149 0.093 0.232 0.078 0.089 0.01
3 ...
```

```
summary(climate)
```

```
##       Year          Month             MEI               CO2
##  Min.   :1983   Min.   : 1.000   Min.   :-1.6350   Min.   :340.2
##  1st Qu.:1989   1st Qu.: 4.000   1st Qu.:-0.3987   1st Qu.:353.0
##  Median :1996   Median : 7.000   Median : 0.2375   Median :361.7
##  Mean   :1996   Mean   : 6.552   Mean   : 0.2756   Mean   :363.2
##  3rd Qu.:2002   3rd Qu.:10.000   3rd Qu.: 0.8305   3rd Qu.:373.5
##  Max.   :2008   Max.   :12.000   Max.   : 3.0010   Max.   :388.5
##       CH4            N2O            CFC.11          CFC.12
##  Min.   :1630   Min.   :303.7   Min.   :191.3   Min.   :350.1
##  1st Qu.:1722   1st Qu.:308.1   1st Qu.:246.3   1st Qu.:472.4
##  Median :1764   Median :311.5   Median :258.3   Median :528.4
##  Mean   :1750   Mean   :312.4   Mean   :252.0   Mean   :497.5
##  3rd Qu.:1787   3rd Qu.:317.0   3rd Qu.:267.0   3rd Qu.:540.5
##  Max.   :1814   Max.   :322.2   Max.   :271.5   Max.   :543.8
##       TSI          Aerosols            Temp
##  Min.   :1365   Min.   :0.00160   Min.   :-0.2820
##  1st Qu.:1366   1st Qu.:0.00280   1st Qu.: 0.1217
##  Median :1366   Median :0.00575   Median : 0.2480
##  Mean   :1366   Mean   :0.01666   Mean   : 0.2568
##  3rd Qu.:1366   3rd Qu.:0.01260   3rd Qu.: 0.4073
##  Max.   :1367   Max.   :0.14940   Max.   : 0.7390
```

Creating Training and Test data set.

```
train = subset(climate, Year <= 2006)
test = subset(climate, Year > 2006)
```

Creating Linear Model.

Temp is dependent variable and (MEI,CO2,..) are dependent variables. we can include many by using '+'.

```
climatelm=lm(Temp ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 + TSI + Aerosols, d
ata=train)
summary(climatelm)
```

```
##
## Call:
## lm(formula = Temp ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 +
##     TSI + Aerosols, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25888 -0.05913 -0.00082  0.05649  0.32433
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.246e+02  1.989e+01  -6.265 1.43e-09 ***
## MEI          6.421e-02  6.470e-03   9.923  < 2e-16 ***
## CO2          6.457e-03  2.285e-03   2.826  0.00505 **
## CH4          1.240e-04  5.158e-04   0.240  0.81015
## N2O         -1.653e-02  8.565e-03  -1.930  0.05467 .
## CFC.11      -6.631e-03  1.626e-03  -4.078 5.96e-05 ***
## CFC.12       3.808e-03  1.014e-03   3.757  0.00021 ***
## TSI          9.314e-02  1.475e-02   6.313 1.10e-09 ***
## Aerosols    -1.538e+00  2.133e-01  -7.210 5.41e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09171 on 275 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7436
## F-statistic: 103.6 on 8 and 275 DF,  p-value: < 2.2e-16
```

- Details:
  Estimate : gives estimates of the beta values for our model. Dependent variable is directly proportional to this value.i.e.if positive, Dependent variable value will increase if it increases and decreas if it decreases. **Std.Error**:The standard error column gives a measure of how much the coefficient is likely to vary from the estimate value.
  t value:The t value is the estimate divided by the standard error. It will be negative if the estimate is negative and positive if the estimate is positive. The larger the **absolute value** of the t value, the more likely the coefficient is to be significant. So we want independent variables with a large absolute value in this column.
  Pr(>|t|):The last column of numbers gives a measure of how plausible it is that the coefficient is

actually 0, given the data we used to build the model.The less plausible it is, or the smaller the probability number in this column, the less likely it is that our coefficient estimate is actually 0. This number will be large if the absolute value of the t value is small, and it will be small if the absolute value of the t value is large. We want independent variables with small values in this column.
Stars at end of each row: More the stars, more better.
Multiple R-squared :R-squared value (1 - SSE/SST) **Adjusted R-squared**:This number adjusts the R-squared value to account for the number of independent variables used relative to the number of data points.

- NOTE:Multiple R-squared will always increase if you add more independent variables.But Adjusted R-squared will decrease if you add an independent variable that doesn't help the model.This is a good way to determine if an additional variable should even be included in the model.

So, all of the variables have at least one star except for CH4 and N2O. So MEI, CO2, CFC.11, CFC.12, TSI, and Aerosols are all significant.

Checking corelation.

```
cor(climate)
```

```
##                  Year        Month         MEI         CO2         CH4
## Year      1.00000000 -0.025789103 -0.14534485  0.98537870  0.91056328
## Month    -0.02578910  1.000000000 -0.01634543 -0.09628668  0.01755804
## MEI      -0.14534485 -0.016345434  1.00000000 -0.15291104 -0.10555472
## CO2       0.98537870 -0.096286676 -0.15291104  1.00000000  0.87225311
## CH4       0.91056328  0.017558035 -0.10555472  0.87225311  1.00000000
## N2O       0.99484971  0.012395210 -0.16237531  0.98113544  0.89440921
## CFC.11    0.46096457 -0.014913724  0.08817074  0.40128447  0.71350408
## CFC.12    0.87006746 -0.001084139 -0.03983567  0.82321031  0.95823718
## TSI       0.02235316 -0.032754296 -0.07682560  0.01786672  0.14633495
## Aerosols -0.36188438  0.014845187  0.35235073 -0.36926514 -0.29038142
## Temp      0.75573115 -0.098015821  0.13529168  0.74850465  0.69969658
##                  N2O      CFC.11       CFC.12         TSI    Aerosols
## Year      0.99484971  0.46096457  0.870067456  0.02235316 -0.36188438
## Month     0.01239521 -0.01491372 -0.001084139 -0.03275430  0.01484519
## MEI      -0.16237531  0.08817074 -0.039835666 -0.07682560  0.35235073
## CO2       0.98113544  0.40128447  0.823210310  0.01786672 -0.36926514
## CH4       0.89440921  0.71350408  0.958237181  0.14633495 -0.29038142
## N2O       1.00000000  0.41215475  0.839295454  0.03989183 -0.35349882
## CFC.11    0.41215475  1.00000000  0.831381310  0.28462884 -0.03230227
## CFC.12    0.83929545  0.83138131  1.000000000  0.18927009 -0.24378508
## TSI       0.03989183  0.28462884  0.189270090  1.00000000  0.08323812
## Aerosols -0.35349882 -0.03230227 -0.243785082  0.08323812  1.00000000
## Temp      0.74324183  0.38011134  0.688944109  0.18218561 -0.39206945
##                  Temp
## Year      0.75573115
## Month    -0.09801582
## MEI       0.13529168
## CO2       0.74850465
## CH4       0.69969658
## N2O       0.74324183
## CFC.11    0.38011134
## CFC.12    0.68894411
## TSI       0.18218561
## Aerosols -0.39206945
## Temp      1.00000000
```

- A high correlation between an independent variable and the dependent variable is a good thing since we're trying to predict the dependent variable using the independent variables.
  Due to the possibility of multicollinearity, we should remove the insignificant variables one at a time.

Since N2O seems to be highly correlated with many, building another model keeping in view this.

```
climatelm2=lm(Temp ~ MEI + TSI + Aerosols + N2O, data = train)
summary(climatelm2)
```

```
##
## Call:
## lm(formula = Temp ~ MEI + TSI + Aerosols + N2O, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27916 -0.05975 -0.00595  0.05672  0.34195
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.162e+02  2.022e+01  -5.747 2.37e-08 ***
## MEI          6.419e-02  6.652e-03   9.649  < 2e-16 ***
## TSI          7.949e-02  1.487e-02   5.344 1.89e-07 ***
## Aerosols    -1.702e+00  2.180e-01  -7.806 1.19e-13 ***
## N2O          2.532e-02  1.311e-03  19.307  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09547 on 279 degrees of freedom
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7222
## F-statistic: 184.9 on 4 and 279 DF,  p-value: < 2.2e-16
```

- We have observed that, for this problem, when we remove many variables the sign of N2O flips. The model has not lost a lot of explanatory power (the model R2 is 0.7261 compared to 0.7509 previously) despite removing many variables.In this particular problem many of the variables (CO2, CH4, N2O, CFC.11 and CFC.12) are highly correlated, since they are all driven by human industrial development.

R provides a function, step, that will automate the procedure of trying different combinations of variables to find a good compromise of model simplicity and R2. This trade-off is formalized by the Akaike information criterion (AIC) - it can be informally thought of as the quality of the model with a penalty for the number of variables in the model.

```
StepModel = step(climatelm)
```

```
## Start:  AIC=-1348.16
## Temp ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 + TSI + Aerosols
##
##            Df Sum of Sq    RSS     AIC
## - CH4       1   0.00049 2.3135 -1350.1
## <none>                  2.3130 -1348.2
## - N2O       1   0.03132 2.3443 -1346.3
## - CO2       1   0.06719 2.3802 -1342.0
## - CFC.12    1   0.11874 2.4318 -1335.9
## - CFC.11    1   0.13986 2.4529 -1333.5
## - TSI       1   0.33516 2.6482 -1311.7
## - Aerosols  1   0.43727 2.7503 -1301.0
## - MEI       1   0.82823 3.1412 -1263.2
##
## Step:  AIC=-1350.1
## Temp ~ MEI + CO2 + N2O + CFC.11 + CFC.12 + TSI + Aerosols
##
##            Df Sum of Sq    RSS     AIC
## <none>                  2.3135 -1350.1
## - N2O       1   0.03133 2.3448 -1348.3
## - CO2       1   0.06672 2.3802 -1344.0
## - CFC.12    1   0.13023 2.4437 -1336.5
## - CFC.11    1   0.13938 2.4529 -1335.5
## - TSI       1   0.33500 2.6485 -1313.7
## - Aerosols  1   0.43987 2.7534 -1302.7
## - MEI       1   0.83118 3.1447 -1264.9
```

```
summary(StepModel)
```

```
##
## Call:
## lm(formula = Temp ~ MEI + CO2 + N2O + CFC.11 + CFC.12 + TSI +
##     Aerosols, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25770 -0.05994 -0.00104  0.05588  0.32203
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.245e+02  1.985e+01  -6.273 1.37e-09 ***
## MEI          6.407e-02  6.434e-03   9.958  < 2e-16 ***
## CO2          6.402e-03  2.269e-03   2.821 0.005129 **
## N2O         -1.602e-02  8.287e-03  -1.933 0.054234 .
## CFC.11      -6.609e-03  1.621e-03  -4.078 5.95e-05 ***
## CFC.12       3.868e-03  9.812e-04   3.942 0.000103 ***
## TSI          9.312e-02  1.473e-02   6.322 1.04e-09 ***
## Aerosols    -1.540e+00  2.126e-01  -7.244 4.36e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09155 on 276 degrees of freedom
## Multiple R-squared:  0.7508, Adjusted R-squared:  0.7445
## F-statistic: 118.8 on 7 and 276 DF,  p-value: < 2.2e-16
```

It is interesting to note that the step function does not address the collinearity of the variables, except that adding highly correlated variables will not improve the R2 significantly. The consequence of this is that the step function will not necessarily produce a very interpretable model - just a model that has balanced quality and simplicity for a particular weighting of quality and simplicity (AIC).

Our residuals, or error terms, are stored in the vector StepModel1$residuals.

```
head(StepModel$residuals)
```

```
##           1           2           3           4           5
## -0.050950385 -0.027312217  0.001313177  0.068606372  0.108093940
##           6
##  0.086756035
```

Sum of Squared Error

```
SSE_model=sum(StepModel$residuals^2)
SSE_model
```

```
## [1] 2.313506
```

Using the model produced from the step function, calculating temperature predictions for the testing data set, using the predict function.

```
tempPredict = predict(StepModel, newdata = test)
SSE_prediction = sum((tempPredict - test$Temp)^2)
SSE_prediction
```

```
## [1] 0.2176444
```

```
SST = sum( (mean(train$Temp) - test$Temp)^2)
R2 = 1 - SSE_prediction/SST
R2
```

```
## [1] 0.6286051
```