516 Case1 Climate Change

Kun Qian

1/28/2021

a. Build a linear regression model to predict Temp, using all of other vairables as independent variables.

```
# load data sets
train <- read.csv("ClimateChangeTrain.csv")
test <- read.csv("ClimateChangeTest.csv")

# Take a look at the training data
head(train)
```

```
##   Year Month   MEI    CO2     CH4     N2O  CFC.11  CFC.12      TSI Aerosols
## 1 1983     5 2.556 345.96 1638.59 303.677 191.324 350.113 1366.102   0.0863
## 2 1983     6 2.167 345.52 1633.71 303.746 192.057 351.848 1366.121   0.0794
## 3 1983     7 1.741 344.15 1633.22 303.795 192.818 353.725 1366.285   0.0731
## 4 1983     8 1.130 342.25 1631.35 303.839 193.602 355.633 1366.420   0.0673
## 5 1983     9 0.428 340.17 1648.40 303.901 194.392 357.465 1366.234   0.0619
## 6 1983    10 0.002 340.30 1663.79 303.970 195.171 359.174 1366.059   0.0569
##    Temp
## 1 0.109
## 2 0.118
## 3 0.137
## 4 0.176
## 5 0.149
## 6 0.093
```

```
#head(test)

# build the first linear model to predict Temp
lm1 <- lm(Temp~.-Year-Month, data=train)
```

i. What is the linear regression equation produced by the model?

```
summary(lm1)
```

```
##
## Call:
## lm(formula = Temp ~ . - Year - Month, 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
```

The linear regression equation is built with the intercept plus the sum of the product of the variables and their corresponding coefficients.

ii. Evaluate the quality of the model.

The R-squared of the model is 0.75. Significant variables are:

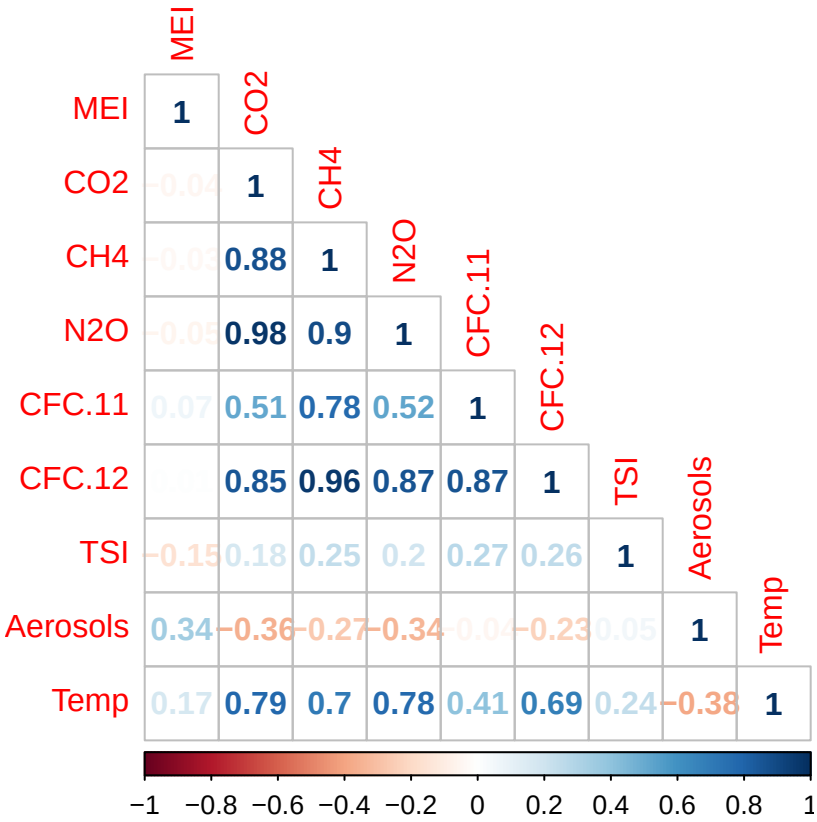
- MEI
- CO2
- CFC.11
- CFC.12
- TSI
- Aerosols

iii. What is the simplest explanation for the contradiction?

N2O and CFC.11 might be highly correlated with other variables, i.e. there might be multicollinearity. This could cause errors in sign and magnitude of the coefficients.

iv. Compute the correlations between independent variables in the training set. Which independent variables is N2O highly correlated with? Which independent variable is CFC.11 highly correlated with?

```
library(corrplot)
corrplot(cor(train[,-c(1,2)]), method='number', type="lower", )
```



- Among all independent variables, N2O is highly correlated with CFC.12.
- Among all independent variables, CFC.11 is highly correlated with CFC.12.

b Build a new linear regression model with only MEI, TSI, Aerosols, and N2O as independent variables.

```
# build the linear model
lm2 <- lm(Temp~MEI+TSI+Aerosols+N2O, data = train)
```

i. How does the coefficient for N2O in this model compare to the coefficient in the previous model?

```
summary(lm2)
```

```
##
## 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
```

The coefficient of N2O in this model is now positive and statistically significant. The sign has flipped and the contradiction no longer exists.

ii. How does the quality of this model compare to the previous one?

The R-squared is 0.73, slightly dropped compare to the previous model, but still quite good. All variables are significant in this model. Since it has less variables with similar performance, it seems to be a simple and strong model.

c. Using the simplified model I created in part(b), calculate predictions for the testing dataset. What is the R-squared on the test set? What does this tell me about the model?

```r
# Make prediction on the test set data using the second linear model
pred <- predict(lm2, test)

# Calculate R-squared
SSE = sum((test$Temp - pred)^2)
SST = sum((test$Temp - mean(train$Temp))^2)
1 - SSE/SST
```

```
## [1] 0.4967795
```

The R-squared drops from 0.73 to 0.50. This tells us that the model perform well on the training data, but perform relatively poorly on unseen testing data.