Climate Change

There have been many studies documenting that the average global temperature has been increasing over the last century. The consequences of a continued rise in global temperature will be dire. Rising sea levels and an increased frequency of extreme weather events will affect billions of people. In this problem, we will study the relationship between average global temperature and several other factors. The file *ClimateChange.csv*, which is available in the Online Companion, contains climate data from May 1983 to December 2008. The variables are described in Table 22.2.

a) Start by splitting the dataset into a *training* set, consisting of observations up to and including 2006, and a *testing* set, consisting of observations after 2006. You will use the training set to build your model, and the testing set to evaluate the predictive ability of your model.

Then, build a linear regression model to predict `Temp`, using all of the other variables as independent variables (except `Year` and `Month`). You should use the training set that you just created to build the model.

i) What is the linear regression equation produced by your model?

ii) Evaluate the quality of the model. What is the R^2 value? (Note that this is called "Multiple R-squared" in some software output.) Which independent variables are significant?

iii) Current scientific opinion is that N2O and CFC-11 are greenhouse gases: gases that are able to trap heat from the sun and contribute to the heating of the Earth. However, you should see that the regression coefficients of both the N2O and CFC-11 variables in this model are negative, indicating that increasing atmospheric concentrations of either of these two compounds is

Table 22.2: Variables in the dataset *ClimateChange.csv*.

Variable	Description
`Year`	The observation year.
`Month`	The observation month.
`Temp`	The difference in degrees Celsius between the average global temperature in the observation month and a reference value.
`CFC.11`	The atmospheric concentration of trichlorofluoromethane (commonly referred to as CFC-11), expressed in ppbv (parts per billion by volume – i.e., 397 ppbv of CFC-11 means that CFC-11 constitutes 397 billionths of the total volume of the atmosphere).
`CFC.12`	The atmospheric concentration of dichlorodifluoromethane (commonly referred to as CFC-12), expressed in ppbv.
`CO2`	The atmospheric concentration of carbon dioxide (CO2), expressed in ppmv (parts per million by volume).
`N2O`	The atmospheric concentration of nitrous oxide (N2O), expressed in ppmv.
`CH4`	The atmospheric concentration of methane (CH4), expressed in ppmv.
`Aerosols`	The mean stratospheric aerosol optical depth at 550 nm.
`TSI`	The total solar irradiance (TSI) in W/m^2, which is the rate at which the sun's energy is deposited per unit area.
`MEI`	Multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of a weather effect in the Pacific Ocean.

associated with lower global temperatures. What is the simplest explanation for this contradiction?

iv) Compute the correlations between all independent variables in the training set. Which independent variables is `N2O` highly correlated with (absolute correlation greater than 0.7)? Which

independent variables is `CFC.11` highly correlated with (absolute correlation greater than 0.7)?

b) Now build a new linear regression model, this time only using `MEI`, `TSI`, `Aerosols`, and `N2O` as the independent variables. Remember to use the training set to build the model.

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

ii) How does the quality of this model compare to the previous one? Consider the R^2 value and the significance of the independent variables when answering this question.

c) We have developed an understanding of how well we can fit a linear regression model to the training data, but now we want to see if the model quality holds when applied to unseen data. Using the simplified model you created in part (b), calculate predictions for the testing dataset. What is the R^2 on the test set? What does this tell you about the model?