edX   **MITx:** 15.071x The Analytics Edge   **Help**

Unit 2: Linear Regression > Assignment 2 > Climate Change

Climate Change

🔖 Bookmark this page

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 attempt to study the relationship between average global temperature and several other factors.

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 (CO_2), nitrous oxide (N_2O), methane (CH_4), trichlorofluoromethane (CCl_3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl_2F_2; 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/m^2 (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.

Problem 1.1 - Creating Our First Model

2.0 points possible (graded)
We are interested in how changes in these variables affect future temperatures, as well as how well these variables explain temperature changes so far. To do this, first read the dataset climate_change.csv into R.

Then, split the data into a *training set*, consisting of all the observations up to and including 2006, and a *testing set* consisting of the remaining years (hint: use subset). A training set refers to the data that will be used to build the model (this is the data we give to the lm() function), and a testing set refers to the data we will use to test our predictive ability.

Next, build a linear regression model to predict the dependent variable Temp, using MEI, CO2, CH4, N2O, CFC.11, CFC.12, TSI, and Aerosols as independent variables (*Year* and *Month* should NOT be used in the model). Use the training set to build the model.

Enter the model R^2 (the "Multiple R-squared" value):

Submit   You have used 0 of 5 attempts

Problem 1.2 - Creating Our First Model

1 point possible (graded)

Which variables are significant in the model? We will consider a variable signficant only if the p-value is below 0.05. (Select all that apply.)

- ☐ MEI

- ☐ CO2

- ☐ CH4

- ☐ N2O

- ☐ CFC.11

- ☐ CFC.12

- ☐ TSI

- ☐ Aerosols

Submit   You have used 0 of 2 attempts

Problem 2.1 - Understanding the Model

1 point possible (graded)

Current scientific opinion is that nitrous oxide 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, the regression coefficients of both the N2O and CFC-11 variables are **negative**, indicating that increasing atmospheric concentrations of either of these two compounds is associated with lower global temperatures.

Which of the following is the *simplest* correct explanation for this contradiction?

○ Climate scientists are wrong that N2O and CFC-11 are greenhouse gases - this regression analysis constitutes part of a disproof.

○ There is not enough data, so the regression coefficients being estimated are not accurate.

○ All of the gas concentration variables reflect human development - N2O and CFC.11 are correlated with other variables in the data set.

Submit	You have used 0 of 1 attempt

Problem 2.2 - Understanding the Model

2 points possible (graded)

Compute the correlations between all the variables in the training set. Which of the following independent variables is N2O highly correlated with (absolute correlation greater than 0.7)? Select all that apply.

☐ MEI

☐ CO2

☐ CH4

☐ CFC.11

☐ CFC.12

☐ Aerosols

☐ TSI

Which of the following independent variables is CFC.11 highly correlated with? Select all that apply.

☐ MEI

☐ CO2

☐ CH4

☐ N2O

☐ CFC.12

☐ Aerosols

☐ TSI

Submit   You have used 0 of 2 attempts

Problem 3 - Simplifying the Model

2.0 points possible (graded)
Given that the correlations are so high, let us focus on the N2O variable and build a model with only MEI, TSI, Aerosols and N2O as independent variables. Remember to use the training set to build the model.

Enter the coefficient of N2O in this reduced model:

(How does this compare to the coefficient in the previous model with all of the variables?)

Enter the model R^2:

[   ]

Submit   You have used 0 of 5 attempts

Problem 4 - Automatically Building the Model

4.0 points possible (graded)
We have many variables in this problem, and as we have seen above, dropping some from the model does not decrease model quality. 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 R^2. 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.

The step function has one argument - the name of the initial model. It returns a simplified model. Use the step function in R to derive a new model, with the full model as the initial model (HINT: If your initial full model was called "climateLM", you could create a new model with the step function by typing step(climateLM). Be sure to save your new model to a variable name so that you can look at the summary. For more information about the step function, type ?step in your R console.)

Enter the R^2 value of the model produced by the step function:

[   ]

Which of the following variable(s) were eliminated from the full model by the step function? Select all that apply.

☐ MEI

☐ CO2

☐ CH4

☐ N2O

☐ CFC.11

☐ CFC.12

☐ TSI

☐ Aerosols

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 R^2 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).

| Submit | You have used 0 of 4 attempts

Problem 5 - Testing on Unseen Data

2.0 points possible (graded)
We have developed an understanding of how well we can fit a linear regression to the training data, but does the model quality hold when applied to unseen data?

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

Enter the testing set R^2:

| Submit | You have used 0 of 5 attempts

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Discussion

Show Discussion

Topic: Unit 2 / Unit 2, Homework: Climate Change

POWERED BY
OPENedX