# Studio 12 Linear regression (Climate change)
# 18.05, Spring 2025

## Overview of the studio

In this studio we will practice applying and interpreting multivariate linear regression in the context of exploring detection and attribution of climate change. This studio has been prepared by Paolo Giani, Mohammad Reza Karimi, and Jonathan Bloom.

## R introduced in this studio

We will explore data manipulation and linear models in R.

New functions: `lm()`

## Download the zip file

- You should have downloaded the studio12 zip file from our Canvas site.
- Unzip it in your 18.05 studio folder.
- You should see the following R files

  studio12.r

  and the following other files

  studio12-instructions.pdf (this file), climate.csv

**There are no test files or sample codes in this studio.**

## Prepping R Studio

- In R studio, open studio12.r
- Using the Session menu, set the working directory to source file location. (This is a good habit to develop!)
- Answer the questions in the detailed instructions just below. Your answers should be put in studio12.r
- Solution code will be posted on Saturday at 4 am

## Detailed instructions for the studio

Climate change detection and attribution studies are fundamental to our understanding of anthropogenic climate change. *Detection* refers to the process of demonstrating that an observed change in climate is statistically significant and cannot be explained by natural internal variability alone. *Attribution* goes a step further by quantifying the contributions of different factors (both natural and anthropogenic) to the observed changes.

In this studio, we consider the following main forcing factors:

- **Greenhouse Gases (GHG):** Primarily $CO_2$, $CH_4$, $N_2O$, and halocarbons, which trap outgoing longwave radiation and warm the atmosphere.
- **Aerosols (AER):** Airborne particles that generally have a cooling effect by reflecting sunlight back to space, though some aerosols (like black carbon) can absorb radiation and contribute to warming.
- **Natural Forcing (NAT):** Includes solar variability and volcanic eruptions, which can cause both warming and cooling effects.
- **Anthropogenic Forcing (ANT):** The combined effect of human-induced factors, typically the sum of greenhouse gases and aerosol effects.

In this studio, we work with the Earth's yearly average temperature data (average of the whole surface of the Earth) for studying climate patterns. To measure climate change more precisely, we establish a baseline period—commonly the pre-industrial era from 1850 to 1870—and calculate what's known as the "temperature anomaly" by subtracting this baseline from current measurements.

Physical-based climate simulations play a crucial role in this analysis. These sophisticated models compute how Earth's temperature responds to various forcing factors mentioned above. For example, recent models like CMIP6 incorporate high-resolution atmospheric dynamics, improved ocean circulation patterns, and more realistic cloud formation processes to provide increasingly accurate projections.

A common approach in detection and attribution studies is to use multiple linear regression. The observed temperature anomaly is modeled as a linear combination of simulated responses to different forcing factors:

$$T_{\text{obs}} = \beta_0 + \beta_{\text{ghg}} \cdot T_{\text{ghg}} + \beta_{\text{aer}} \cdot T_{\text{aer}} + \beta_{\text{nat}} \cdot T_{\text{nat}} + \varepsilon \tag{1}$$

Here, $T_{\text{obs}}$ is the observed temperature anomaly, $T_{\text{ghg}}$, $T_{\text{aer}}$, and $T_{\text{nat}}$ are the (simulated) temperature responses to greenhouse gases, aerosols, and natural forcings, $\beta_0$ is the intercept, $\beta_{\text{ghg}}$, $\beta_{\text{aer}}$, and $\beta_{\text{nat}}$ are regression coefficients, and $\varepsilon$ is the residual error.

The regression coefficients ($\beta$ values) have a physical interpretation: If the coefficient is statistically significantly different from 0, we can "detect" the influence of that forcing on the observed warming pattern.

Once we have estimated the regression coefficients, we can calculate the *attributable warming*—the portion of observed warming that can be attributed to each forcing factor is $\hat{\beta}_{\text{factor}} \cdot T_{\text{factor}}$. This allows us to quantify how much of the observed warming is due to greenhouse gases, aerosols, and natural factors.

## Problem 1

The data will be provided in a CSV file named `climate.csv` with the following columns:

- `year`: The year of observation
- `obs`: Observed global mean temperature anomaly (reference period: 1850–1870)
- `ghg`: Simulated temperature response to greenhouse gas forcing
- `aer`: Simulated temperature response to aerosol forcing
- `nat`: Simulated temperature response to natural forcings

All the measurements are in degrees Celcius.

First, make sure you can read the data and run the line

```
climateData = read.csv("climate.csv")
```

Be sure to set your working directory to the folder containing all files for this studio.

**Problem 1a.** Here you will complete the code for the function

```
studio12_problem_1a()
```

This function plots the observations along with the values for ghg, nat, and aer. Make educated guesses about the reason for the trends you see in the plot, and briefly write your findings in the cat statement at the end of the function.

The following quote from NASA's website might help you understand the trend for aerosols:

> Aerosols come in many forms. They can be natural, like wildfire smoke, volcanic gases, or salty sea spray. Human activities can also generate aerosols, such as particles of air pollution or soot. The role of aerosols in climate science is complex. In general, light-colored particles in the atmosphere will reflect incoming sunlight and cause cooling. Dark-colored particles absorb sunlight and make the atmosphere warmer. Because different types of particles have different effects, aerosols are a hot topic in climate research.

**Problem 1b.** Here you will finish the code for the function

```
studio12_problem_1b()
```

This function performs the first linear regression analysis, where the (adjusted) observation (`obs`) is the dependent variable and `ghg`, `nat`, and `aer` are the independent ones.

1. Start by printing the summary of the model.
2. Use the first 'cat' statement to print the $R^2$ value of the model. Then, explain in words what that value means.
3. Use the second 'cat' statement to print the linear model with the correct coefficients. That is, print $\text{obs} = a + b \times \text{ghg} + c \times \text{nat} + d \times \text{aer}$, where $a, b, c, d$ are the coefficients of your model.
4. Use the third 'cat' statement to print which independent variable is the least significant. (The intercept is not a variable).

5. Use the fourth 'cat' statement to briefly explain why it is reasonable to expect that the variable you chose in the previous item should be the least significant.

**Problem 1c.** Here you will finish the code for the function

`studio12_problem_1c()`

This function performs the second linear regression analysis, where the (adjusted) observation (`obs`) is the dependent variable and `nat` and `ant` are the independent ones.

1. Start by printing the summary of the model.
2. Use the first 'cat' statement to print the $R^2$ value of the model. Then, explain in words what that value means.
3. Use the second 'cat' statement to print the linear model with the correct coefficients. That is, print $\text{obs} = a + b \times \text{nat} + c \times \text{ant}$, where $a, b, c$ are the coefficients of your model.
4. Use the third 'cat' statement to describe what you see in terms of significance of the independent variables. (The intercept is not a variable).

**Problem 1d.** Here you will finish the code for the function

`studio12_problem_1d()`

This function should plot the fitted values for both of the models in problems 1b and 1c, along with the observations and forcing effects.

**Problem 1e.** Here you will finish the code for the function

`studio12_problem_1e()`

Here you compute the "attributable warming", that is, how much of the observed warming is due to ghg, nat and aer? Compute the attribution (in degrees Celcius) of each factor for the year 2014 and use the cat statement to print these values.

### Further References

If you want to know more about detection and attribution of climate change, we encourage you to take a look at chapter 10 from the latest IPCC report (AR6): `https://www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5_Chapter10_FINAL.pdf`. Box 10.1 in this pdf is basically what you have done in this studio. You can see how this is just a very simple example compared to the much wider literature of detection and attribution that is reviewed in that IPCC chapter.

You can also take a look at `https://rls.sites.oasis.unc.edu/postscript/rs/Hammerling_34_Final.pdf`.

## Before uploading your code

1. Make sure all your code is in studio12.r. Also make sure it is all inside the functions for the problems.
2. Clean the environment and plots window.

3. Source the file.
4. Call each of the problem functions with the same parameters as the instructions file.
5. Make sure it runs without error and outputs just the answers asked for in the questions.

## Upload your code

Upload your code to Gradescope.

Leave the file name as studio12.r.

You can upload more than once. We will grade the last file you upload.

## Due date

**Due date:** The goal is to upload your work by the end of class. If you need extra time, you can upload your work any time before 6 PM ET on the day of the studio (Friday).

**Solutions uploaded:** Solution code will be posted on Canvas at 4 AM the day after the studio.