STAT 605 Project Proposal

Samach Sathitvudh, Sheng Huang, Kyungjin Cho, Shravan Kaul, Midhun Satheesh

Variables description

We obtain the data from Kaggle containing daily observations climate records over global land areas. We have observed the sample data and realize that there are several useful columns that we are going to use in our analysis.

```
library(data.table)
dt = fread("2019.csv.gz")
head(dt)
```

Columns used and description

- V1 = station id where the data is recorded
- V2 = date id in the form of YYYYMMDD
- V3 = elements of record, which the followings aim to be used:
  - TMAX = Maximum temperature (tenth of degrees C)
  - TMIN = Minimum temperature (tenth of degrees C)
  - TOBS = Temperature at a time of observation
  - PRCP = Precipitation of rain (tenths of mm)
  - SNOW = Precipitation of snowfall (mm)
- V4 = the value recorded for the element

(However, for other elements, they might be used in case of necessary to make a model more reliable)

Statistical methods

According to the topic we are interested in, we consider applying the logistic regression technique to model the probability of global warming as $\mathbb{P}(y_i = 1) = \frac{1}{1+exp(-x_i^T \beta)}$

Computational steps

1. Seperately download the climate data file each year.
2. Assign each job to calculate the average records (temperature, precipitation, etc.) for each year.
3. Concatenate all of the results into one file prepared for the next analysis.
4. Perform logistic regression in R, based on the calculated average values.
5. Perform data visualization and interpret the results from the model and graphics.

Github repository link

https://github.com/kcho36/Stat605-Final-Project