Power Plant Capacity

Sam Reeves

Here we have a dataset containing info concerning about 30k power plants globally. We are going to try to use all the variables given to predict the capacity in megawatts of a powerplant on Earth. That is, we will make a generalized linear model that describes the dependent variable capacity.

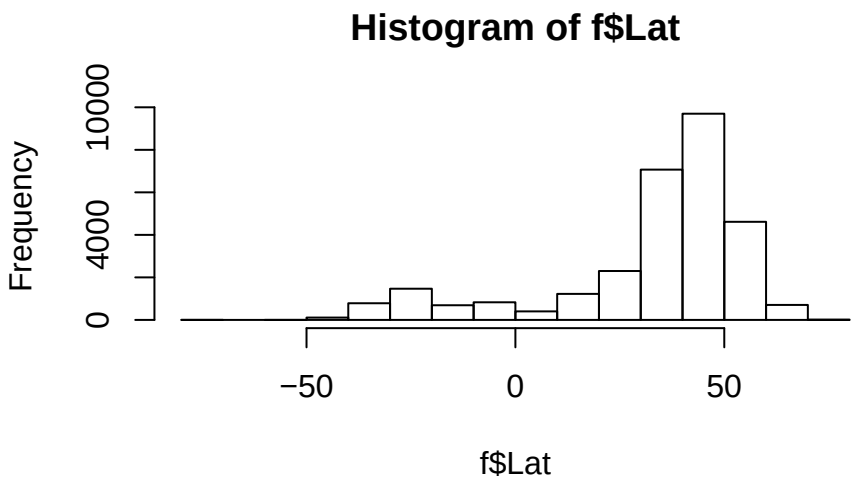
```
library(dplyr)
library(geosphere)

f <- read.csv("power.csv") %>%
  tibble() %>%
  select(c(Country, Capacity..MW., Latitude, Longitude, Primary.Fuel, Owner, Source)) %>%
  rename(Cap = Capacity..MW.,
         Fuel = Primary.Fuel,
         Lat = Latitude,
         Long = Longitude) %>%
  arrange(Lat)

head(f)
```

```
## # A tibble: 6 x 7
##   Country       Cap   Lat  Long Fuel  Owner               Source
##   <fct>       <dbl> <dbl> <dbl> <fct> <fct>               <fct>
## 1 Antarctica    6.6 -77.8 167.  Oil   ""                  Antarctica Australia
## 2 Antarctica    1   -77.8 167.  Wind  "Meridian Energy"   Meridian Energy
## 3 Argentina   750   -53.8 -67.7 Hydro "EPEC"              Ministerio de Energía~
## 4 Argentina     2.4 -46.8 -68.0 Wind  "MUNICIPAL"         Ministerio de Energía~
## 5 Argentina    45.6 -46.8 -67.9 Gas   "CT PATAGONICAS SA" Ministerio de Energía~
## 6 New Zealand  83   -45.9 170.  Hydro "Trustpower"        Trustpower
```

Dichotomous term

Since there is no dichotomous term given automatically by the dataset, I suppose we will have to make one out of latitude. There are substantially more power plants in the Northern hemisphere than the southern, and I believe that adding this feature should fall in line with other global economic studies. Frequently, development economists discuss trade and infrastructure in terms of the “global north” or “global south”.

```
f <- mutate(f, North = ifelse(Lat > 0, 1, 0))
hist(f$Lat)
```


Histogram of f$Lat
Frequency
0
4000
10000
−50
0
50
f$Lat

Quadratic term

There isn't an obvious choice for a quadratic term either. So, I want to combine this dataset with another that contains information on many of the world's cities. We will relate each power plant with the nearest city with a population over 1,000,000.

To compute the distance, we can use a number of functions from the geosphere package. It's not clear which is fastest, so we will use the one which relies on the law of cosines... There is a lot of data.

```
g <- read.csv("worldcities.csv") %>%
  tibble() %>%
  filter(population > 8000000) %>%
  rename(pop = population,
         type = capital) %>%
  select(c(city, lat, lng, pop, type)) %>%
  arrange(lat)

head(g)
```

```
## # A tibble: 6 x 5
##   city              lat   lng      pop type
##   <fct>           <dbl> <dbl>    <dbl> <fct>
## 1 Buenos Aires   -34.6  -58.4 16157000 primary
## 2 São Paulo      -23.6  -46.6 22046000 admin
## 3 Rio de Janeiro -22.9  -43.2 12272000 admin
## 4 Lima           -12.0  -77.0  9848000 primary
## 5 Luanda          -8.84  13.2  8417000 primary
## 6 Jakarta         -6.21 107.  34540000 primary
```

```
findNear <- function(f, g) {
  nearest <- tibble()
  for (i in 1:nrow(f)) {
    cities <- tibble()
    p1 <- c(f[i,]$Long, f[i,]$Lat)

    for (j in 1:nrow(g)) {
      p2 <- c(g[j,]$lng, g[j,]$lat)
      dist <- distCosine(p1, p2)
      cities <- rbind(cities, cbind(g[j,], dist))
    }
    nearest <- rbind(nearest, cities[which.min(cities$dist),])
  }
  return(nearest)
}
```

Dichotomous vs. quantitative interaction term

Interpret all coefficients

Conduct residual analysis

Was the linear model appropriate? Why or why not?