MONT 105N – Analyzing Environmental Data
Chapter 12 Project
April 15 (and 17?), 2019

Introduction

For this final chapter project, we will draw on many of the topics studied in this and previous chapters to try to understand some data on severe weather and climate events collected and analyzed by scientists at NOAA, the National Oceanographic and Atmospheric Administration. As hinted at in the Preface of our text, whether we ultimately decide to try to arrest or reverse the ongoing changes in the earth's atmosphere caused by fossil fuel burning and the effects that we have studied in previous Chapter Projects depends on the costs to individuals and to society that are involved. Are we willing to bear those costs to maintain our current economic and lifestyle choices?

Background

The first step here, of course, is to understand exactly what those costs have been. One predicted effect of higher global average temperatures is an increase in the frequency and severity of extreme weather events, including both droughts and flooding in different locations, severe storms in both summer and winter, hurricanes in the tropics, and so forth. The data we will look at, from

`https://www.ncdc.noaa.gov/billions`

includes information about the number and severity of "billion-dollar events" in the U.S. since 1980, that is, weather events that have caused more than $1 billion in estimated damages.

There are several things to keep in mind in looking at this data:

- The types of events covered here include droughts, floods, freezes, severe storms, severe winter storms, tropical cyclones (hurricanes and weaker tropical storms), and wildfires caused by dry conditions. It is predicted that climate change should produce more extremes of all sorts, not just "global warming."
- There has been inflation in the dollar over this whole period, so to make fair comparisons, all dollar amounts have been indexed to the current value of the dollar. This means, in particular, that some events in the past would not have been this costly in dollars in use at that time. But taking inflation into account, the current value of the associated damages is over the $1 billion limit.
- The cost figures come from estimates made by FEMA (the Federal Emergency Management Agency), from insurance claims information, and from other sources. This means that there is often a time lag between when an event happens and when the best estimate for the total cost becomes available.
- In addition, the cost figures probably *understate* the true cost in some cases. For instance, if a storm hits an area where many of the residents have fewer financial resources (the 2017

impact of hurricane Maria on the island of Puerto Rico is an example), many of them may not have insurance coverage, and their losses may not be estimated accurately.

- For the reason immediately above, all of the loss amounts are reported with *confidence intervals* estimated by applying methods related to those we have discussed to similar events, in addition to single dollar amounts. The "believable" values of the true cost lie somewhere in those intervals.

The Form of the Data

I have condensed the data from the NOAA website above into the file `Chapter12ProjectData.txt` that you will download from our course homepage. When you open this file in an editor, you will see that there are 7 rows of numbers. The first are the number of years since 1980–0 (representing 1980) to 38 (representing 2018, the last year for which complete data is available). The other six rows are, in order:

- the total counts of billion-dollar events for each year
- the total cost of the billion-dollar events for each year (in billions of inflation-adjusted dollars)
- the total counts of billion-dollar severe storms (excluding tropical cyclones) for each year
- the total cost of the billion-dollar severe storms for each year (in billions of inflation-adjusted dollars)
- the total counts of billion-dollar tropical cyclones for each year
- the total cost of the billion-dollar tropical cyclones for each year (in billions of inflation-adjusted dollars)

A Worked Example in `R`

The file above is organized so you can easily copy and paste one of the rows of data into an `R` list data structure and work with it. (There are also much more powerful `R` facilities for reading in information from data files, but those are somewhat intricate, so we won't go into them!) To start

- On the beginning of an `R` input line type in `Years <- c(`
- Copy the list of years from the data file and paste it into the `R` window immediately after the `c(` above
- Type in a `)` to close off the list.
- Now on the next input line enter `TotalCost <- c(`
- Copy the list of total cost vales from the data file and paste it into the `R` window immediately after this `c(`, then type in `)` to close off the list.

In R, the basic way to do a simple linear regression (the linear trendline calculation we used extensively last semester) is this. After the `Years, TotalCost` data have been entered, type in these commands:

```
CostModel = lm(TotalCosts ~ Years)
summary(CostModel)
```

The first line computes the linear model best fitting `Total Costs` as a function of `Years`. The second command generates a list of *regression statistics* going much farther than anything we discussed last semester. The part after the heading `Residuals:` gives the 5-number summary for the regression residuals–the differences between the data values and the predicted values from the model.

The next part after the heading `Coefficients:` is completely new. The idea here is that when you do a regression, you are *also* computing all the raw material for *several statistical hypothesis tests*. Let's look at this in detail:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7141    17.5240  -0.041  0.96771
Years         2.3172     0.7935   2.920  0.00593 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 55.77 on 37 degrees of freedom
Multiple R-squared:  0.1873,Adjusted R-squared:  0.1653
F-statistic: 8.527 on 1 and 37 DF,  p-value: 0.005928
```

What does this all mean? Well:

- the values in the `Estimate` column are just the coefficients in the linear model, the intercept and the slope of the line. So the linear model itself is

  $$\text{Total Costs} = 2.3172 \cdot \text{Years} - .7141$$

  (units of billions of dollars).

- the values in the `Std. Error` column are standard errors (SD's) of the estimators for the two coefficients in the linear model

- the values in the `t value` column are the values of a test statistic (with a $t(37)$-distribution–37 degrees of freedom–we can see this indicated below as well).

- The last column on the right gives a p-value for a test of the null hypothesis that the coefficient is $= 0$ vs. an alternative hypothesis that the coefficient is $\neq 0$.

- Here we see that there is *pretty strong evidence* ($p = .00593$) to reject the null hypothesis that the slope is zero, but no evidence at all ($p = .96771$) to reject the null hypothesis that the intercept is zero(!)

- The last part gives even more detail about the regression, including the fact that the linear model doesn't fit the data very well ($R^2 \doteq .1873$). But we can safely ignore that because we are *asking a different question now*–not whether a linear model comes close to the data points, but whether the best-fitting linear model gives evidence to say there is an upward trend in the data.

In practical terms, we can say that *there is strong evidence for the statement that the total cost per year of these natural disasters is increasing with time*, because the slope 2.3172 of the linear model is *positive* (the units would be billion dollars per year).

Another important question to ask here is whether the *number of these large-cost events per year* has been increasing. In this part you will carry out some computations along the lines of one discussion of an article cited in the text. The idea is that since we are dealing with the numbers of events in intervals of time, it would be reasonable to model them using the Poisson distributions from Chapter 10. However, we allow the parameter λ in the Poisson probability mass function to depend on time, and we estimate how $\lambda(t)$ is changing using a log-linear model of the form

$$\ln(\lambda(t)) = \lambda_0 + \lambda_1 t + \text{random variation.} \tag{1}$$

Once we estimate λ_0 and λ_1 from the data, from (1), we get an exponential model for $\lambda(t)$ as a function of t as we discussed in Chapter 5. To do the estimation of λ_0 and λ_1 you will need to use the `R` command `glm` ("generalized linear model") with the options set for *Poisson regression.* The details of exactly how this works are somewhat beyond the scope of this discussion, so we won't try to describe them.

Start off by reading the total counts data into an `R` list called `Events` as we did above for the `Years` and `TotalCosts`. The Poisson fitting is done via the following command:

```
CountModel = glm(Events ~ Years,family=poisson)
```

You can see a summary of the results using

```
summary(Countmodel)
```

Lab Questions

(A) What do the results of the Poisson regression mean here? Do we have evidence for saying the number of large-cost weather events per year has increased over this period? Roughly how fast is the average number per year increasing? Note that the left side of (1) uses the *natural logarithm.* The reason is that that's the way `R` is doing this computation.

(B) Now look at the other data from the `Chapter12ProjectData.txt` file.

   (1) What do basic linear regressions of the severe storm costs and tropical cyclone costs against `Years` tell us?

   (2) What do Poisson regressions of the severe storm counts and the tropical cyclone counts against `Years` tell us?

Assignment

Write a report summarizing your findings, and addressing the question whether this data indicates that the costs and number of severe weather events each year are increasing. Something to think about: if these trends continue, what real world actions might be necessary?