Homework

Computational Statistics and Data Analysis
Summer Semester, 2020

Ryan Hutchins
Ruprecht Karls Universität Heidelberg

8 Mai, 2020

You may find the code for this assignment here.

1 Problem 1: Numerical Optimization

(a) Appropriate H_0 and H_1 for these tests are: $H_0 : F_X = F_Y$ and $F_X \neq F_Y$. The null hypothesis states that the social distancing measures have no effect on the new infection rate of the disease. Based on these test results, it does appear that the social distancing measures make a substantial difference. The numbers that I ran told me to **reject** the null hypothesis.

I compute a t statistics value of: 4.0238208193893055
The endpoints of the 95% interval for the null hypothesis were: (-2.1199052992210112, 2.1199052992210112)

This led to a rejection of the null hypothesis.

(b) The null hypothesis, H_0 is that social distancing measures have no effect, and therefore that we can expect the distribution before and after the measures were taken to be the same, up to random variation, so $F_X = F_Y$. My results show that that is false. When I construct the Empirical Distribution Function of t_{N-1}, I get that my bootstrap values are outside the 95% interval for the EDF in all but around 1% of cases. Thus, t_{N-1} is significantly different from 0, if my α value is 0.05.

2 Problem 2: Likelihood Ratio Tests

The null hypothesis in this case is that $\beta_1 = 0$, where β_1 is the linear coefficient in the model:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot \cos\left(\frac{\pi}{6}x_i\right) + \epsilon_i \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, iid.

(a) My test results seem to suggest that there is no linear trend in the data.

(b) My estimates for the full and reduced parameter set β values are:

$$\beta_0 = 10.7023703, \beta_1 = .002823587, \beta_2 = -7.78643352 \qquad \text{full set}$$
$$\beta_0 = 11.21202778, \beta_2 = -7.78360994 \qquad \text{reduced set}$$

In order to test whether they are significantly different from 0, I need to construct a t-distributed statistic. This means I need a statistic that is the ratio of $z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_{k-1}$. From script equation (2.7), I know that:

$$\hat{\beta} \sim \mathcal{N}(\vec{\beta}, (X^T X)^{-1}\sigma^2) \tag{2}$$

so I must get a location-scale transformation on $\hat{\beta}$ to make it standard normal. To do this, I would need:

$$\hat{\beta}^* = \frac{\hat{\beta} - \vec{\beta}}{(X^T X)^{-1/2}\sigma} \tag{3}$$

where $\vec{\beta}$ is the true mean of β the parameters. Given the location-scale transformation applied, we should have $\beta \sim \mathcal{N}(0,1)$, although I am unclear on how to take the square root of a matrix. Since the sample variance is always χ^2_{N-1} distributed, where N is the number of data records, we could then take the statistic:

$$t = \frac{\hat{\beta}^*}{\sqrt{S_N^2/N}} \tag{4}$$

That should give us a t-distributed statistic.
The null hypothesis would be that the true mean is zero, i.e., that $\vec{\beta} = 0$, so that $\hat{\beta}^* = \frac{\hat{\beta}}{\sqrt{S_N^2/N}}$. This test can also be done for any given component of β, in which case we reduce the dimensionality of the problem. In our case, we were interested in whether or not β_1 was significantly different from 0, and it does not appear to be.

(c) The model assumes that the noise in the observations is independent, and I do not believe that that is actually satisfied by our weather data. I think the average temperature of the earth in a given month influences that in the next given heat retention by, among other things, water vapor.
I believe that the t distribution has a sample size that is already approaching sufficiency after N = 5, and we have N = 16, so I suppose it does fit the model in that regard.