Normal model

Javier Esteban Aragoneses, Mauricio Marcos Fajgenbaun, Danyu Zhang, Daniel Alonso

March 5th, 2021

Introduction

The Los Angeles ozone pollution dataset contains 13 meteorological variables and 366 observations. It's a time series where each observation represents daily measures obtained in 1976. In this research we will only consider one of them, which is the concentration of ozone (represented in the dataset by the 4th column: "*Daily maximum one-hour-average ozone reading*"). We will use a prior distribution and the dataset itself to then perform the imputation of missing values in this particular column of the dataset. Our objective is to characterize the average of the ozone level in New York.

Model

To begin with, our data is a time series. This means, that is not a set of i.i.d. random variables (as in a time series, values depend on their past, until certain extent). Nevertheless, we will work with the data as if it was independent (thus, assuming i.i.d.).

Then, we will try to characterize the data as a normal distribution. So we have to check (through a qq plot and the Shapiro-Wilk test) to asses normality.

Testing normality

We perform a normality test and a generate a QQ-plot in order to assess whether our data comes from a normal distribution or not.

- **Shapiro-Wilk normality test (Original data)**: We can see from the p-value returned (*5.988324e-14*) that we reject that the data comes from a normal distribution.
- **QQ-plot (Original data)**: The QQ plot also coincides with these results where we can see that our data most certainly does not follow a normal distribution.
- **Shapiro-Wilk normality test (Transformed data)**: We perform a *log* transformation to normalize the data and then perform the Shapiro-Wilk normality test on the transformed dataset. After performing the transformation we can see that the situation improves by several orders of magnitude (p-value: *9.681899e-07*), however, it's still not enough evidence to conclude it comes from a normal distribution.
- **QQ-plot (Transformed data)**: Creating a QQ-plot for the transformed dataset yields similar results where we clearly notice that the data most likely does not come from a normal distribution, even if there's a significant improvement over the original dataset in attempting to achieve such normality.

Analysis

As these test show that the data does not behave normal, we proceed to fix the assimetry problem by taking logarithms. We can see that this improves our normal assumption, although it does not behave strictly normal. We decided not to change more the data, as it will complicate our final interpretation.

In general, a random variable is said to be normally distributed with mean θ and variance $\sigma^2 > 0$ if the density is given by:

$$p(y|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y-\theta}{\sigma})^2}, \ -\infty < y < \infty$$

As we do not know neither the mean nor the variance of the population, we will try to do inference for the mean not conditioning on a known variance, through a Bayesian approach. To do that, as for any joint prior distribution for the mean and variation, we will do posterior inferences using Baye´s rule. We will derive the full posterior distribution of this model, and using this full posterior derive the marginal posteriors for both the expected value μ.

We will suppose our prior distribution and sampling model are as follows:

$$\begin{aligned} 1/\sigma^2 &\sim gamma(\nu_0/2, \nu_0\sigma_0^2/2) \\ \theta|\sigma^2 &\sim normal(\mu_0, \sigma^2/\kappa_0) \\ Y_1, \ldots, Y_n|\theta, \sigma^2 &\sim \text{i.i.d. normal}(\theta, \sigma^2) \end{aligned}$$

It is important to recall that 1/sigmasq can be interpreted as the precision. We can also interpret the prior parameters (σ_0^2, ν_0) as the sample variance and sample size of prior observations. μ_0 and κ_0 can be interpreted as the mean and sample size from a set of prior observations.

In this case, we know from an expert that in nature the values of ozone are between 0.001 and 0.125. This is our prior information. So we have to translate this into prior parameters. The average between these two values is of 0.0625. However, as New York is a very populated and busy city, we will take 0.07 as the prior mean (a bit more than the nature). The prior standard deviation is harder to grasp. We will say it is 0.05, as this way all the data would be considered inside two standard deviations. Taking into account that in a normal distribution 95% of the data in included when taking almost two standard deviations, it seems like a good prior measure. However, we are really not sure about this, as none of us has really prior knowledge about the production and concentration of ozone in a city. So we choose $\kappa_0 = \nu_0 = 1$ so that our prior distributions are only weakly centered around these estimates from our expert and us.

The sample mean of data set is ≈ 0.012 and the sample variance is $\approx 6.56 * 10^{-5}$. From this values and the prior parameters already stablished we can compute μ_n:

$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n} = \frac{1 * 0.07 + 366 * 0.012}{\kappa_n} = \frac{4.462}{367} = 0.0122$$

We can also calculate our posterior variance as follows:

$$\sigma_n^2 = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2] = 0.000212$$

This is important, because now we can also determine our joint posterior distribution (that is completely determined by the values that we calculated). This can be expressed as follows:

$$\begin{aligned} \{\theta|y_1, \ldots, y_n, \sigma^2\} &\sim \text{ normal}(\mu_n, \sigma^2/\nu_n) \\ \{1/\sigma^2|y_1, \ldots, y_n\} &\sim \text{ gamma}(\nu_n/2, \nu_n * \sigma_n^2/2) \end{aligned}$$

As our marginal posterior is not defined in a close form, we have to use Monte Carlo method to approximate (simulations of random samples for the inverse-gamma, to calculate the variance and simulations of normal distributions to elicit the mean distribution).

Data imputation

To simulate the values for the normal in R, we require concrete values for the population mean and variance. However, we have posterior probability distributions.

The algorithm we have works as follows:

- Create a vector for *means*, *variances* and *vals* (the values that will be used to impute the dataset)
- A sample of 10 values is chosen randomly from the simulation of the s^2 and θ
- The mean of the 10 value sample is calculated and appended to the *means* and *variances* vectors
- A single normally distributed value is obtained using **rnorm** utilizing the mean and variance of the iteration (as this mean and variance will change per iteration given that we choose a different 10 value sample in each iteration)

- We then apply the exponential function to it (as the data used had a previous log transformation) and add it to the vector of values to impute
- The NAs are then replaced by the values of this vector

Conclusions

The means calculated through the bayesian approach lie between 0.01 and 0.013, with a variance between 0.00016 and 0.00024. Given these results, we have to say that our prior were probably not so good. A good approach would be to get another sample, and with this update our priors (using this last results). This way, we can learn about the population mean and variance (through updating our priors). The histogram with the posterior distribution of the mean and variance are shown in the annex, as follows.

Annex: Code

```
library(dplyr)
library(mlbench)
library(UsingR)
library(car)

data(Ozone)

# prior
mu0=0.07
k0=1
s20=0.05
nu0=1

# Normality tests (no transformation)
shapiro.test(Ozone$V4)$p.value
car::qqPlot(Ozone$V4)

# Normality test (after log transformation)
lozone = log(Ozone$V4)
shapiro.test(lozone)$p.value
car::qqPlot(lozone)

# data
n=length(lozone);ybar=mean(na.omit(lozone));s2=var(na.omit(lozone))
# posterior inference
kn=k0+n; nun=nu0+n
mun=(k0*mu0+n*ybar)/kn
s2n=(nu0*s20+(n-1)*s2+k0*n*(ybar-mu0)^2/(kn))/(nun)
s2.postsample=1/rgamma(10000,nun/2,s2n*nun/2)
theta.postsample=rnorm(10000,mun,sqrt(s2.postsample/kn))

means <- c(); vars <- c(); vals <- c()
for (i in 1:5) {
    means <- c(means, mean(sample(theta.postsample, 10)))
    vars <- c(vars, mean(sample(s2.postsample, 10)))
    gen <- rnorm(1, mean=means[i], sd=sqrt(vars[i]))
    vals <- c(vals, exp(abs(gen)))

}
Ozone$V4[is.na(Ozone$V4)] <- vals
```