Stat4ACSAI / Homework 04

Marco Stefanucci

Due 5 days before the oral presentation

General Instructions

I expect you to upload your solutions (one per team) on Moodle as a **single running** `Notebook` (`.ipynb`) **named with your surnames** + pictures converted and collected in a single `pdf`-file of your handwritten exercises. Alternatively, a `zip`-file with all the material inside will be fine too.

Your responses must be supported by textual explanations, the code **you** wrote to produce the results and **extensive comments** on both the code and the results.

Please Notice

- Remember our **policy on collaboration**: *collaboration on homework assignments with fellow students is **encouraged**. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had **discussions** (no more) concerning your solution. You may **not**, however, share written work or code after discussing a problem with others. The solutions should be written by **you and your group** only.*

Exercise 01 | Everybody in the kitchen!

A way to check a model is to inspect numerical measures that detect observations that highly influence the model fit. The **leverage** is a measure of an observation's potential influence on the fit. For now, we mention that the leverage measures how far the explanatory variable values fall from their means, and it takes values between 0 and 1. Observations for which explanatory variables are far from their means have greater potential influence on the least squares estimates. For an observation to actually be influential, it must have both a relatively large leverage and a relatively large residual. **Cook's distance** is a diagnostic that uses both, based on the change in $\{\widehat{\beta}_j\}$ when the observation is removed from the data set. For the residual e_i and leverage h_i for observation i, Cook's distance is

$$D_i = \frac{e_i^2 h_i}{(p+1)s^2(1-h_i)^2}$$

where s^2 is an estimate of the conditional variance σ^2. This diagnostic is nonnegative, with a relatively large D_i occurring when both e_i and h_i are relatively large.

In the file *Cook.txt* are reported some data regarding several seats of the same company where Y is (the logarithm of) the sales volume in 2021 and X is (the logarithm of) the sales volume in 2020. The interest is in how the sales volume on 2021 depend on the previous year's volume.

1) Compute the OLS solution of the linear regression problem. Individuate those observations with large leverage and those with large Cook's distance.

2) Remove the highly influential observations and re–run the OLS procedure. Does the fit look better? Comment on the estimated coefficients.

The problem of outliers can be partially mitigated by changing the loss function. While in ordinary least squares the loss is

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\rho(e_i)$$

where $\rho(e_i) = e_i^2$ is the square loss, other choices are possible. The loss $\rho(e_i) = |e_i| = |y_i - \mathbf{x}_i^T\boldsymbol{\beta}|$ for example lead to the so called **median regression**, an instance of the broader methodology of quantile regression. This loss should be more resistant to outliers with respect to the square loss.

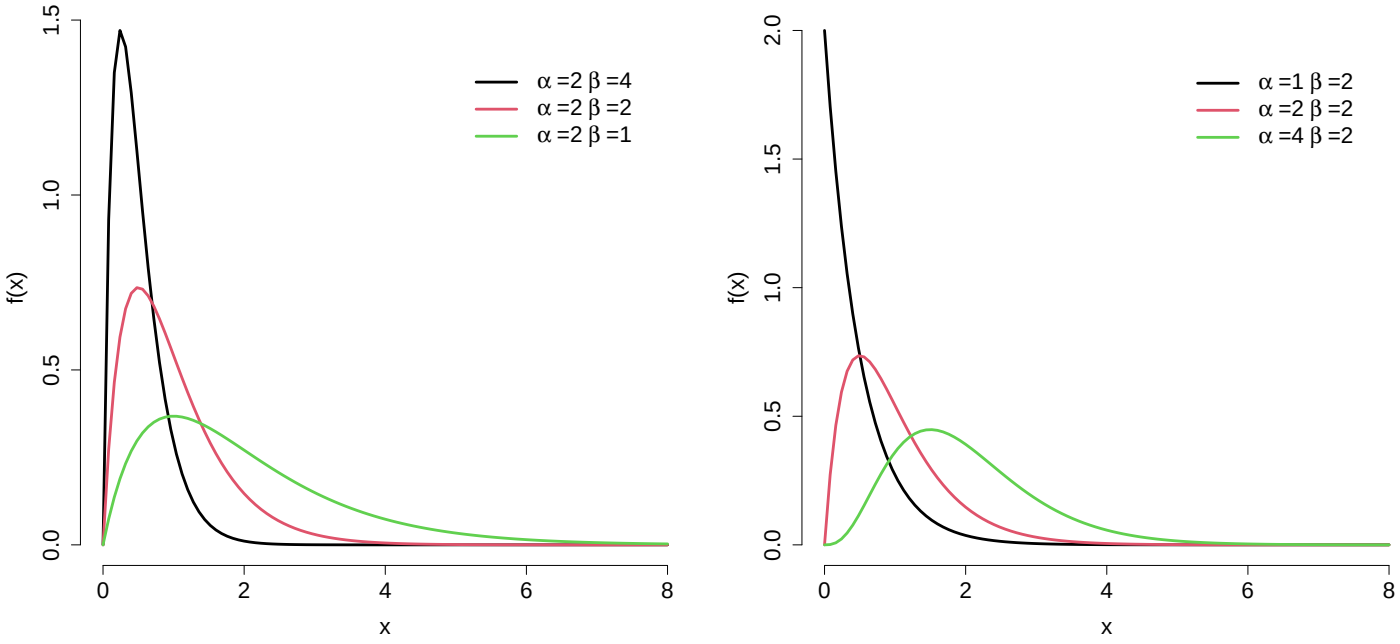
3) Write *your own* code to obtain solution of the median regression problem in `Python` and apply it to the *Cook* data. Compare the results to the previous ones.

Exercise 02 | Yet another model on... Covid data!

The **Gamma random variable** is a positively skewed random variable useful to model only–positive continuous data. The pdf is

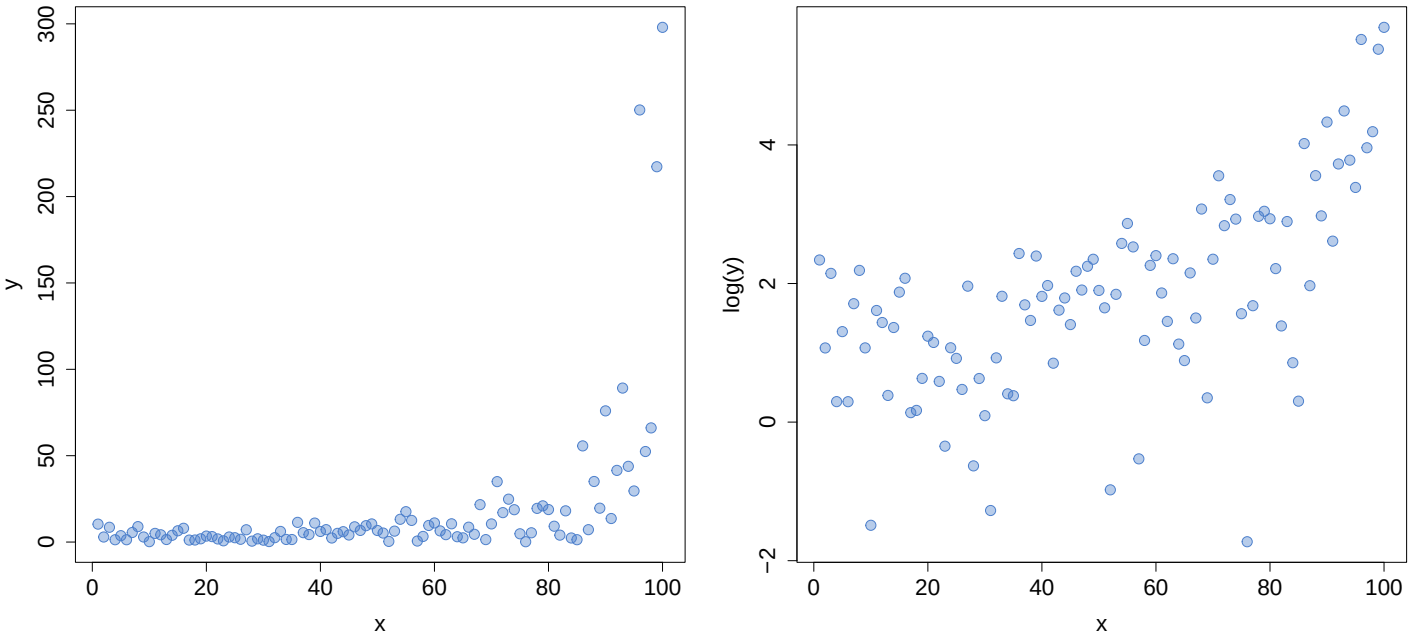
$$f(x) = \frac{\delta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-\delta x}, \quad x > 0$$

where $\Gamma(\alpha)$ is the special mathematical function Gamma. Some examples of Gamma random variables with different choices of the parameters are illustrated in the figure.



1) Prove that the Gamma random variable *with fixed parameter* α belongs to the exponential family. Determine the canonical parameter θ, the dispersion parameter ϕ and the function $b(\theta)$.

In the next figure percentages ($\times 10000$) of Covid-19 cases in a country in the first 100 days of pandemic are plotted. The plot shows that y = percentage of cases was approximately an exponential function of x = day number during the early days of the pandemic. The $\{\log(y_i)\}$ values, also shown in the figure, more closely follow a linear pattern. A simple idea is to fit a normal linear model to those log-transformed values, namely $\mathbb{E}[\log(Y_i)] = \beta_0 + \beta_1 x_i$. By contrast, rather than modeling *the expected value of the transformation*, we could model *a transformation of the expected value*, $\log[\mathbb{E}(Y_i)] = \beta_0 + \beta_1 x_i$ thus using a GLM with log link function and normal distribution for Y. However, since data are positive and the variability can be related to the mean, a last option is to build a **Gamma GLM**, using the *canonical link* and assuming $Y_i \sim Ga(\alpha, \delta_i)$.



2) Derive *analitically* the canonical link, variance function, pseudo-responses and all things you need in order to build the Fisher scoring algorithm for the Gamma GLM. Implement the algorithm from scratch in `Python` (set $\alpha = 1$) and apply it to the Covid data, available in the file *Covid.txt*.

3) Compare the model to the linear model applied to transformed data and to the GLM with log link and normal response. Explain the theoretical differences about the assumption on the mean and the variance. Which one would you pick?