COLUMBIA UNIVERSITY — Prof. Marco Avella
Statistical Inference and Modeling — TA: Ian Kinsella
Graded Homework 4 — Fall 2019

You are free to present your solutions to the exercises below in *groups of at most three*. This homework is due on Dec 5th. In bocca al lupo!

Exercise 1

We would like to model the number of cigarettes smoked per day among smokers according to some available covariates[1]. We have data on 310 individuals all living in the USA, but in different states. Due to the different state legislatures, cigarettes prices and smoking restrictions vary from one individual to the other. Our data feature the following variables:

- the response `cigs`, the number of cigarettes smoked per day;
- `cigpric`, the price (in ten cents) for a cigarette pack in the individual's state;
- `income` in US dollars;
- the dummy `restaurn` indicating whether there are restaurant smoking restrictions in the state or not (`1` = yes, `0` = no);
- the dummy `white` coding for the individual's skin color (`1` = white, `0` = non-white);
- `educ`, the number of years of education;
- and `age` given in years.

1. Comment the distribution of the response.
2. Fit a GAM to these data. Which variables are you going to keep in the parametric part ?
3. Interpret the overall effect of age on the number of cigarettes smoked per day.
4. Can we include income in a linear fashion in our model? If not, how should we transform this variable if we still want it in the parametric part ?

[1]Data from Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. MA: MIT Press

5. Does the price of a cigarette pack really have an impact on the number cigarettes smoked per day ?

Exercise 2

The aim of this exercise is to implement a test in R in order to check whether the functional form is linear or not. For this purpose we will use the functions "lm" for the linear regression estimator and "gam", in the "mgcv" library, for a nonparametric estimator.
Simulate the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

where $\beta_0 = 2$, $\beta_1 = 3$, $x_i \overset{i.i.d.}{\sim} N(0, 9)$ and $\varepsilon_i \overset{i.i.d.}{\sim} N(0, 3)$ for $i = 1, ..., 300$.

1. Estimate (1) under the null hypothesis of being a linear model, using "lm".
2. Find the distribution of the parameter β_1 estimate using the naive bootstrap.
3. Estimate (1) with "lm" and "gam". Compute the residuals linear model, as well as the fitted values for both denoted as, $\hat{y}_{i,\text{lin}}$ and $\hat{y}_{i,\text{gam}}$.
4. Compute the statistic given below:

$$T_1 = \sum_{i=1}^{300} (\hat{y}_{i,\text{gam}} - \hat{y}_{i,\text{lin}}) \hat{\varepsilon}_{i,\text{lin}} \tag{2}$$

5. In order to draw a conclusion concerning the values of the before-calculated statistics we will perform a wild bootstrap procedure. This method will allow us to compute the distribution of T_1 . For this purpose compute:

$$y_i^{*,(b)} = \hat{y}_{i,\text{lin}} + (y_i - \hat{y}_{i,\text{lin}}) \varepsilon_i^{*,(b)}, \quad i = 1, ..., 300, \quad b = 1, ..., 1000. \tag{3}$$

   where (b) denotes each bootstrap sample and $\varepsilon_i^{*,(b)} \overset{i.i.d.}{\sim} N(0, 1)$.

6. Then for each bootstrap follow the following steps:
   (a) Run $y_i^{*,(b)}$ on x_i with the functions "lm" and "gam" for each bootstrap sample. Compute the fitted values for both regressions, denoted as $\hat{y}_{i,\text{lin}}^{(b)}$ and $\hat{y}_{i,\text{gam}}^{(b)}$ respectively, as well as the residuals, denoted as $\hat{\varepsilon}_{i,\text{lin}}^{(b)}$.
   (b) Compute (2) for each bootstrap sample. That is,

$$T_1^{(b)} = \sum_{i=1}^{300} (\hat{y}_{i,\text{gam}}^{(b)} - \hat{y}_{i,\text{lin}}^{(b)}) \hat{\varepsilon}_{i,\text{lin}}^{(b)} \tag{4}$$

7. Compare T_1 with $T^b_{1,0.95}$, where $T^b_{1,0.95}$ is the 95^{th} quantiles of the bootstrapped distributions of (4). What do the tests conclude? Comment.

Exercise 3

The `falls.txt` file features data coming from a recent study aiming at reducing the number of falls among elder people. In this study, the 340 participants were randomly allocated either to a control group (169 people), where they went through usual care programs, or to an intervention group (171 people) in which they performed exercises aiming at increasing balance and lower limbs muscle strength. The researchers want to check the hypothesis that performing these exercises will help elder people fall less.
The response variable `falls` consists of the reported number of falls during the period of study. The `days` variable indicates the length of this period, as it varies between participants. The `id` variable gives a numerical code for each participant. The available covariates are:

- `falls_past12mo`: the recalled number of falls over the 12 months prior to the study;
- `gait_speed`: the walking speed in m/s, measured over a 4 m walk at the beginning of the study;
- `age`: age in years;
- `male`, a dummy coding 1 for male participants;
- `walk_aid`, a categorical variable with 3 levels (Rollator, Stick, and Nil) showing the type of walking aid, if any, required by the participant;
- and the dummy `intervention` indicating the group allocation, where "A" stands for the control group.

As a statistician, you are asked to provide a sensible model for these data and test the research hypothesis. In your analysis, you may explore the main effects of the covariates and some meaningful interactions. Answer these questions and join your outputs **where necessary**:

1. Describe your modeling approach (nature of the response, hypotheses, model specifications, etc.).
2. Explain how you have obtained your final good model(s). Validate this (these) model(s) using the results and graphs of your analysis.

The next questions should be answered according to your (one of your) final model(s).

3. How did you introduce the `falls_past12mo` variable in your model ? Justify your approach.

4. What is the role of the `days` variable in such a count model ?

5. Is the `intervention` variable significant ? Interpret its coefficient, even if it's not significantly different from zero. What would you conclude about the research hypothesis ? You may have found an unexpected effect, try nevertheless to give an explanation for it.

6. Did you find any significant interaction ? Interpret them and try to give them some meaning regarding the data.

Exercise 4 (Review question)

Let $X_1, \ldots, X_n$ be the budget shares for food of n households. We assume them to be identically distributed observations according to a distribution with density

$$f_\beta(x) = \begin{cases} \beta(1-x)^{\beta-1} & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

1. Compute the estimator of moments $\hat{\beta}_M$ of β.

2. Compute the asymptotic distribution of $\sqrt{n}(\hat{\beta}_M - \beta)$.

3. Compute the maximum likelihood estimator $\hat{\beta}_{ML}$ of β.

4. Give an approximation of $var(\hat{\beta}_{ML})$ when n is large.

5. Compare the asymptotic efficiency of $\hat{\beta}_M$ versus $\hat{\beta}_{ML}$ and sketch this curve as a function of β. What can you conclude?

6. Compute the maximum likelihood estimator $\hat{\eta}_{ML}$ of the transformed parameter $\eta = 1/\beta$.

7. Give the test statistic of the most powerful test for testing the null hypothesis $\beta = \beta_0$ against the alternative $\beta > \beta_0$.

8. For a given level α, determine an approximate critical value of the test derived at the point 7.

9. Construct a confidence interval for β based on the asymptotic distribution of $\hat{\beta}_{ML}$.