DXE_EMTR 2021
First assignment (20% of total grade)

Please submit the assignment by 29 Oct in the IS MUNI system. You are allowed to work in groups of maximum size 3.

1 Regression basics

Write a short essay (no more than 1000 words) discussing the following article:

Imbens, Guido W. "Statistical Significance, p-Values, and the Reporting of Uncertainty."Journal of Economic Perspectives 35.3 (2021): 157-74.

These questions could help you to streamline the discussion:

- what makes the (ab)use of p-values problematic in some contexts?
- should p-values be banned?
- what are some possible strategies for addressing the problem of 'p-hacking' and publication bias?

Make sure to add your views/perspective, that may be specific to the field of your research expertise.

2 Identification

Assume that

- Y, X, ϵ are random variables,
- $Y = (X - \theta)^2 + \epsilon$,
- $E(\epsilon) = 0$,
- Data reveals ϕ which is the joint distribution of (Y, X),
- θ is the parameter of interest.

Here are your tasks:

- Define: the model and the structure in the framework of Lewbel, Arthur. "The identification zoo: Meanings of identification in econometrics."Journal of Economic Literature 57.4 (2019): 835-903..
- Show under what conditions is the parameter θ
  - point identified
  - set identified and find the identified set
  - not identified.
- Try to find an intuitive explanation for your answer from the previous subquestion.
- Suppose we have an additional variable Z, so that the data reveals ϕ which is the joint distribution of (Y, X, Z). Furthermore we replace the assumption $E(\epsilon) = 0$ with the following assumptions: $E(Z) = 0$, $E(Ze) = 0$ and $E(ZX) \neq 0$. Is θ point identified? If so, how would you estimate it?

3 Maximum likelihood

In many situations our dependent variables describe a number of events, so that $Y \in \{0, 1, 2, ...\}$ without a natural upper bound. This may be, for instance, a number of car crashes, a number of earthquakes or a number of visitors. Our ambition may be to find an association between the variation in the y and some explanatory variables $x_1, \cdots x_p$, these may include weather conditions, geographic location or time of day, depending on what the variable y is.

Consider the special type of regression (called Poisson regression), where we assume the following assumptions:

- Data sample consists of n i.i.d. observations $(y_i, x_{i1}, \cdots, x_{ip})$,
- $y_i \sim Pois(\lambda_i)$,
- $\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip}$,

You are asked to do the following:

- We are interested in estimating the vector of unknown parameters $(\beta_0, \beta_1, \cdots, \beta_p)$. Derive the likelihood function (conditional on the covariates $x_1, \cdots, x_p$), the score function and the Fisher information matrix for this model.
- Using a small simulation study in `R`:
  - demonstrate that the maximum likelihood estimator of β_1 for this particular model has asymptotically normal distribution. (You don't need to implement the optimization yourself, it is OK if you make use of `glm` function in `R` with option `family = poisson`).
  - Explore how the sample size affects the variance of the estimator.

4 Bootstrap

Consider the maximum likelihood estimator of the unknown parameter β_1 from the previous task.

- Construct a 95% confidence interval based on the non-parametric percentile bootstrap.
- Construct a 95% confidence interval based on the normal approximation and use bootstrap to estimate the standard errors.
- Using a simulation study in `R`, compare the coverage properties of these two confidence intervals. That is: show that the confidence intervals cover the true value in approximately 95% simulated cases.[1]

Make sure to comment your code and make your best effort to adhere to some reasonable coding standards. Your code must be easy to read. Present your results in a coherent way and whenever possible make use of visualization.

[1] Notice that this may require a lot of computing time - if a single bootstrap confidence interval is based on 100 bootstrap samples and you will run 500 simulations, you will need to estimate the Poisson regression model $100 \cdot 500 = 50000$ times. So you may need to keep the basic model specification simple.