Project 2:
Non-Linear Models for Cross Section and Panel Data

Due: Sunday November 8th, 2020 at 22:00

Part 1: Marginal Effects and 'Munkit'

Suppose that we observe n independent observations $\{(Y_i, X_i)\}_1^n$ sampled from the censored regression model

$$Y = \max\{0, Y^*\},$$

where the latent outcome Y^* is given by

$$Y^* = \beta_0 X + \sigma_0 \varepsilon,$$

$\beta_0 \in \mathbf{R}$ and $\sigma_0 \in \mathbf{R}_{++}$ are unknown parameters (to be estimated), ε and X are independent, and ε is distributed according to a *known* differentiable CDF $G : \mathbf{R} \to [0,1]$ with a continuous and positive PDF g [i.e., $g(t) > 0$ for all $t \in \mathbf{R}$].

(1) Derive the conditional probabilities of Y^* being censored and uncensored (from below, at zero), respectively, conditional on $X = x$, and relate these probabilities to the probabilities $\mathrm{P}(Y = 0 | X = x)$ and $\mathrm{P}(Y > 0 | X = x)$ pertaining to the outcome Y.

(2) Derive the CDF $F_{Y|X}(\cdot | x)$ of Y conditional on $X = x$ and comment on the nature of $F_{Y|X}(\cdot | x)$.

(3) Derive the likelihood contribution function of the ith observation and define the maximum likelihood estimator of $\boldsymbol{\theta}_0 := (\beta_0, \sigma_0)$ based on $\{(Y_i, X_i)\}_1^n$.

(4) Show that $\mathrm{E}[Y | X = x] = \beta_0 x [1 - G(-\beta_0 x/\sigma_0)] + \sigma_0 \int_{-\beta_0 x/\sigma_0}^{\infty} t g(t)\, \mathrm{d}t$.

(5) Derive an expression for the marginal effect $\mathrm{ME}(x) := (\mathrm{d}/\mathrm{d}x)\, \mathrm{E}[Y | X = x]$ of X on the conditional mean of Y at x and comment on its dependence on x.

(6) Evaluate the claim: "Censoring leads to a reduction of the marginal effect of X relative to its marginal effect on the latent outcome."

(7) Suppose that you have already established consistency of $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ as $n \to \infty$. Suggest a consistent estimator $\widehat{\mathrm{ME}}(x)$ of the marginal effect $\mathrm{ME}(x)$ and argue its consistency at some point x.

(8) Suppose now that you have already established that $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathrm{N}(\mathbf{0}, \mathbf{V}_0)$ as $n \to \infty$ for some 2×2 variance matrix $\mathbf{V}_0$. What is the asymptotic distribution of the estimator $\widehat{\mathrm{ME}}(x)$ from your answer to the previous question?

(9) Discuss the components necessary to construct a 95% confidence interval for $\mathrm{ME}(x)$ and argue in one sense it is valid.

Hints

- Simplify your answers as much as possible.
- Note that G need <u>not</u> be the CDF of the standard normal distribution.
- *Leibniz integral rule* states that for an integral of the form

$$\int_{a(x)}^{b(x)} h(x,t)\,\mathrm{d}t,$$

where $a(\cdot)$ and $b(\cdot)$ are differentiable functions of x, the derivative of the integral is expressible as

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_{a(x)}^{b(x)} h(x,t)\,\mathrm{d}t = h(x,b(x))\,b'(x) - h(x,a(x))\,a'(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} h(x,t)\,\mathrm{d}t.$$

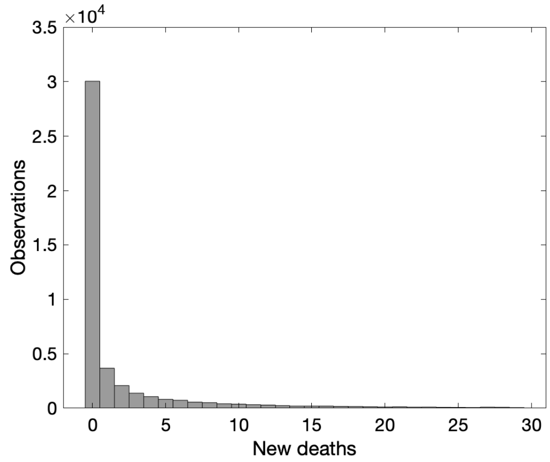
- You may want to consult Cameron & Trivedi (2005) Theorems A.3, A.11 and/or A.12.

Part 2: COVID-19 and Temperature

The accompanying dataset, `covid.csv`, contains a panel datasets of $N = 209$ countries over $T = 299$ days covering the global COVID-19 pandemic. At the bottom of this project description, you will find a more detailed description of the variables but the key variables that we will focus on are:

- `new_deaths`: number of new deaths reported due to COVID-19 for the country on that day. Additionally, the dataset contains the variable `new_deaths_per_million`, which is `new_deaths` divided by the country's population in millions.
- `temperature`: a measure of the temperature in degrees celsius. The measures are daily averages of raw measurements from weather stations in the country.

Figure 1: New deaths due to COVID-19



The overarching research question of this empirical project is: *What is the expected development in COVID-19 deaths due to the seasonal changes in temperature?* An accurate answer to this question may help us in forecasting what lies ahead in terms of the severity of the pandemic over the winter season.

In particular, you should focus on two non-linear models that explicitly deal with the following two features of the outcome variable:

- The number of deaths per day cannot take negative values,
- The number of deaths is an integer, i.e. $y_{it} \in \{0, 1, 2, ...\}$.

These two features are clear from Figure 1 clearly shows a substantial mass point at zero.

In what follows, let y_{it} the number of deaths (in levels or per million capita), let z_{it} denote temperature, and $\mathbf{x}_{it}$ be a vector of additional regressors. You should everywhere focus on *static models* (i.e. no lagged outcome, y_{it-1}, or first-differences), emphasizing $N \to \infty$ asymptotics in your analysis. The focus will be on estimating models of $\mathrm{E}(y_{it}|z_{it}, \mathbf{x}_{it})$, and in particular the marginal effect with respect to z_{it}: $\frac{\mathrm{d}}{\mathrm{d}z_{it}}\mathrm{E}(y_{it}|z_{it}, \mathbf{x}_{it})$.

(1) Pick an estimation sample and a set of additional covariates, $\mathbf{x}_{it}$, and justify your decision. You should keep this fixed throughout the rest of the questions.

- *Suggested approach: start by estimating univariate linear models of $E(y_{it}|z_{it})$ for different samples, e.g. Denmark, Europe, and the whole world. Then proceed to explore sources of variation in temperature: pooled, between, and within countries. Finally, choose the set of covariates you find plausibly exogenous and important, balancing data availability. You should not report all you have done in this step but find a simple way of summarizing the most important insights.*

(2) Estimate models of $\mathrm{E}(y_{it}|z_{it}, \mathbf{x}_{it})$ using respectively a Tobit model, and a Poisson regression model (see Cameron & Trivedi, 2005, ch. 5.2.1 and 20.2.1). Focus your comparison on the marginal effect of z_{it}.

(3) Assess the fit of the two models first in terms of $\mathrm{E}(y_{it}|z_{it}, \mathbf{x}_{it})$, and then in terms of other features of the distribution. Which model is the most suitable for understanding the development in Denmark?

- *Hint: with maximum likelihood, you have made assumptions on the full distribution of y given $(z_{it}, \mathbf{x}_{it})$. Hence, you can compute* $\Pr(y_{it} = 0|z_{it}, \mathbf{x}_{it})$ *or even the full distribution for any observation i, t.*

(4) Assess the robustness of your estimated marginal effects from the Tobit model with respect to the assumed distribution for the error term.

- *Hint: estimate two or three alternative models. You may use your derivations from Part 1 (justify your choice of G) or any other estimator you deem relevant.*

(5) Is the effect of temperature on COVID-19 deaths constant across countries and over time? Are some countries likely to see sharper increases in fatalities over the coming months?

Hints

- Regarding numerical optimization:
  - State what starting values you used for the optimizer.
  - Report the termination flag from `fminunc` and check that it has indeed converged. If gradients are not flat, attempt to use `fminsearch` and if that also cannot improve further, it may be that your criterion function has noise in it.
- The date variable is a `datetime` object in Matlab. This means that you can extract e.g. the month with the function `month(T.date)`.
- The function `dummyvar(cat)` takes a categorical vector (i.e. a vector taking only integer values) and creates a matrix of dummy vectors, one for each of the possible values.
- The function `icdf('normal', 0.5, 0, 1)` returns $\Phi^{-1}(0.5)$, i.e. it evaluates the inverse of the normal cdf at 0.5. Type `help icdf` to see the many other built-in distributions. Similarly, `cdf('normal, 1.96, 0, 1)` is the cdf itself, $\Phi(1.96) = 97.50\%$.

- Your writing is constrained in terms of the amount of text you may write. Therefore, make sure that you use tables effectively. There is no need to reiterate every single coefficient from a table: that is the purpose of the table. Instead, your text should focus on the key highlights, differences, similarities, conclusions based on your tests, etc.

- Long tables: if you have a long list of coefficients in your model, you are allowed to only report the most important ones. For instance, if you a model with country fixed effects, it is not necessary to report every single estimate: simply include a row with "Country FE: Yes/No" at the bottom of your table (assuming that you have coefficients in rows and different models in columns). But report the number of fixed effects included and the reference category.

Formal Requirements

- You must hand in a report containing your answers to all parts of the assignment.

- The report must be written in English and uploaded to Peergrade via Absalon as one single PDF file.

- You must obey the following page constraints:

  - Part 1: No constraints,

  - Part 2: Maximum five normal pages of text plus two pages with output in the form of tables and/or figures. [1]

- You are allowed to work in groups of up to three people (not necessarily in the same exercise class as yours). List all group members on the front page of your report.

- The assessment criteria are posted on the course page on Absalon.

Data description

The dataset is constructed from three primary sources:

- OWID Covid dataset: `https://ourworldindata.org/coronavirus-source-data`. Covid dataset with deaths, tests, cases, and time-constant country information.

[1] One normal page of text is defined as: Font size = 12p, 1.5 line spacing, and margins of 2.5 cm.

- Apple mobility data: `https://covid19.apple.com/mobility`: Daily data on mobility from Apple's devices (by car, walking, or public transportation).
- Google location data: `https://www.google.com/covid19/mobility/`: Daily data on where people are located (home, work, shopping, parks).
- Climate data, from the US' NOAA: `ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/`: Daily data from weather stations across the world.

The most important variables contained in the data are:

- `new_deaths`: the number of new deaths within that date (midnight to midnight) that were due to COVID-19.
- `temperature`: The average of all the day's measurements from weather stations in the country at hand on that day. Specifically, it is the average of the variable `TAVG` in the NOAA dataset, which is the average temperature during the day in celsius. It should be noted that some countries have seemingly very large values.
- `date`: the calendar date for the observation.
- `country`: the name of the country in question.
- `continent`: name of the continent (Africa, Asia, Europe, North America, Oceania, or South America).

Some additional variables below.

- Apple data: variables with the prefix `mobility_*`. These are mobility indices that measure the search intensity on Apple's maps for trips of different types. The indices are relative to the activity in the same country on January 13, 2020.
  - `mobility_driving`: driving trips,
  - `mobility_walking`: walking trips.
  - `mobility_transit`: public transportation trips.
- Google data: variables that begin with `location_*`. These variables are *relative* measures of mobility patterns, relative to the median for the same weekday for January 3 to February 6, 2020.
  - `location_retail_and_recreation`
  - `location_grocery_and_pharmacy`

  - `location_parks`
  - `location_transit_stations`
  - `location_workplaces`
  - `location_residential`

- `cases`, `deaths`, `tests`: variables describing respectively the number of new confirmed COVID-19 cases, deaths, and conducted tests. Each variable is available as `new_X` (variable in level), `new_X_per_million` (variable normalized by population), `new_X_smoothed` (variable smoothed by a running average), and `new_X_smoothed_per_million` (same, but normalized).

Remaining variables should be self-explanatory by the variable name. Note that many variables are country-specific but do not vary over time (e.g. `diabetes_prevalence` or `hospital_beds_per_thousand`). You can find a complete description here: `github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv`