Lecture 4: Maximum Likelihood, Logit and Poisson Regression

Antoine Chapel

Introduction

- Yesterday was mostly dedicated to Optimization, like the first half of math+econ+code
- Today, we move to econometrics, and in particular the sort that requires numerical methods
- We will apply the function seen earlier to do gradient descent to solve estimation problems

Outline:

1. Reminder on maximum likelihood
2. The Poisson Regression
3. The logit model

Maximum Likelihood

Suppose you have the following data:

$$y = \begin{bmatrix} 81 & 75 & 100 & 82 & 64 \end{bmatrix}$$

- What is the chance/the likelihood that this sample was generated by draws from a normal distribution $\mathcal{N}(0,1)$? Very low
- By a normal distribution $\mathcal{N}(82,1)$? Much higher.
- Maximum likelihood formalizes that idea

Maximum Likelihood

- normal density:

$$f(y_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \tag{1}$$

- The likelihood function, given a sample $y = \begin{pmatrix} y_1, & \dots, & y_n \end{pmatrix}$:

$$\begin{aligned}
\mathcal{L}(\mu,\sigma^2|y) &= \prod_{i=1}^{n} f(y_i|\mu,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}
\end{aligned}$$

Maximum Likelihood

Taking the $\log$-likelihood makes thing much easier to derivate.

$$\begin{aligned}\log \mathcal{L}(\mu, \sigma^2|y) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\mu + \mu^2)\end{aligned}$$

Now, the Maximum Likelihood Estimator (MLE) is found by taking the FOC of the ($\log$-)likelihood with respect to the parameters.

$$\frac{\partial \log \mathcal{L}}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(2\mu - 2y_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0$$

$$\Leftrightarrow \hat{\mu} = \frac{1}{n}\sum_{i}^{n} y_i$$

Maximum Likelihood

For variance:

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\mu, \sigma^2 | y)}{\partial \sigma^2} &= -\frac{n}{2}\frac{2\sigma}{\sigma^2} - \frac{-\left(\sum_{i=1}^{n}(y_i - \mu)^2\right)4\sigma}{4\sigma^4} = 0 \\
&= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{\sigma^3} = 0 \\
&\Leftrightarrow \hat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu})^2}{n}
\end{aligned}
$$

As you know this estimator is not unbiased. MLEs are in general consistent and asymptotically normal.

Maximum Likelihood: OLS

You may also find the OLS estimator of β through maximum likelihood, assuming $y_i \sim \mathcal{N}(x_i'\beta, \sigma^2)$. Solution on the next slide

$$\begin{aligned}\log \mathcal{L}(\beta, \sigma^2 | y) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X'\beta)'(y - X'\beta)\end{aligned}$$

Notice how you are maximizing "-(squared residuals)"

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta} &= -\frac{1}{2\sigma^2}(-2X'y + 2X'X\beta) = 0 \\ \Leftrightarrow \hat{\beta} &= (X'X)^{-1}(X'y)\end{aligned}$$

Maximum Likelihood Estimators

- MLE requires you to make an assumption on the whole distribution of data: can be quite strong
- In comparison, Method of Moments estimators only enforce a single condition on moments

Maximum Likelihood: Poisson Regression

- Poisson regression will appear several times in the masterclass. This presentation gives you a basic understanding of what it is, and what it does.
- Poisson distribution: $f(y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$, with $E[y] = \lambda$
- It describes the distribution of discrete positive events (typically, number of visits to the doctor per year)
- Ordinary Linear Regression relies on the assumption that the y is normally distributed, with $E[y|X] = X'\beta$
- Poisson regression instead uses the assumption that y_i is Poisson distributed, with the specification that $E[y|X] = e^{X'\beta}$. This is part of the family of generalized linear models (GLM)

$$f(y_i|x_i,\beta) = \frac{e^{-\exp(x_i'\beta)}\exp(x_i'\beta)^{y_i}}{y_i!}$$

Maximum Likelihood: Poisson regression

$$
\begin{aligned}
\mathcal{L}(\beta|y) &= \prod_{i=1}^{n} \frac{e^{-\exp(x_i'\beta)} \exp(x_i'\beta)^{y_i}}{y_i!} \\
\Leftrightarrow \log \mathcal{L}(\beta|y) &= \frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{e^{-\exp(x_i'\beta)} \exp(x_i'\beta)^{y_i}}{y_i!} \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} (-\exp(x_i'\beta) + y_i x_i'\beta - \log y_i!)
\end{aligned}
$$

$$
\frac{\partial \log \mathcal{L}}{\partial \beta} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \exp(x_i'\beta))x_i' = 0
$$

Clearly, no way to isolate β in here and find a closed form expression for an estimator $\hat{\beta}$! So let us take this to Python.

Logit

- Logit is the basis of discrete choice models
- It is here again a member of the GLM family, but the specification is a bit more subtle
- In this preparation session, in IO and in m+e+c, it will be used for demand estimation.

Logit

An individual i choosing alternative j obtains utility:

$$U_{ij} = V_j + \epsilon_{ij}$$

- Where V_j is the ”representative utility” that every agent gets from alternative j. Typically, it would be parametrized as $V_j = x'_j\beta$
- ϵ_{ij} denotes an unobserved heterogeneity: that is, a specific taste from individual i for a given alternative, unobserved characteristics.
- We make the assumption that ϵ_{ij} is distributed according to a Gumbel distribution (also called Type 1 extreme value): $F(\epsilon) = e^{-\exp(-\epsilon)}$

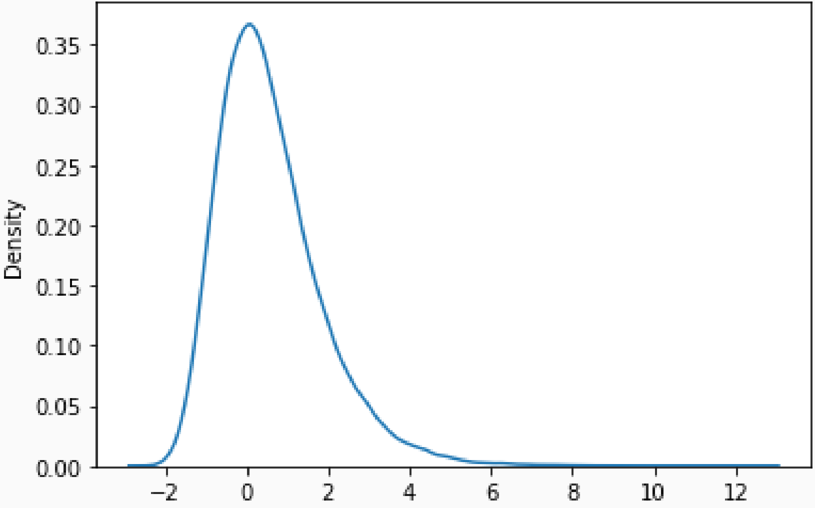

Figure 1: The Gumbel distribution

Logit

The probability that agent i chooses alternative j, given shock ϵ_{ij}, is:

$$\begin{aligned}
P_{j|\epsilon} &= P(U_{ik} < U_{ij} | \epsilon_{ij} \ \forall k \neq j) \\
&= P(V_k + \epsilon_{ik} < V_j + \epsilon_{ij} \ \forall k \neq j) \\
&= P(\epsilon_{ik} < \epsilon_{ij} + V_j - V_k \ \forall k \neq j) \\
&= F(\epsilon_{ij} + V_j - V_k \ \forall k \neq j) \\
&= \prod_{k \neq j} e^{-e^{-\epsilon_{ij} + V_j - V_k}}
\end{aligned}$$

$$\begin{aligned}
P_j &= \int_{-\infty}^{+\infty} \prod_{k \neq j} e^{-e^{-\epsilon_{ij} + V_j - V_k}} f(\epsilon_{ij}) d\epsilon_{ij} \\
&= \frac{e^{V_j}}{\sum_k e^{V_k}}
\end{aligned}$$

For reference, you can find the full derivation of this result on the notebook.

Logit

- P_j denotes the probability that any individual would choose alternative j, given representative utility V_j.
- The simplest Industrial Organization method assumes that $V_j = X_j\beta + \xi_j$, where ξ_j is normally distributed unobserved heterogeneity of alternative j. P_j is then interpreted as alternative j's market share, that you observe from the data.

$$s_j = \frac{e^{V_j}}{\sum_k e^{V_k}} \quad s_0 = \frac{e^{V_0}}{\sum_k e^{V_k}}$$

$$\frac{s_j}{s_0} = \frac{e^{V_j}}{e^{V_0}} = \frac{e^{V_j}}{e^0} = e^{V_j}$$

$$\log(s_j) - \log(s_0) = V_j$$

$$\log(s_j) - \log(s_0) = X_j'\theta + \xi_j$$