Module 5: Non-linear Models and Maximum Likelihood Estimation

Econometrics II

Sannah Tijani (stijani@wu.ac.at)
Department of Economics, WU Vienna

Max Heinze (mheinze@wu.ac.at)
Department of Economics, WU Vienna

December 18, 2025

Limited Dependent Variables

Linear Probability Model

Modeling Probabilities

Interpretation

Limited Dependent Variables

So far, we have mostly focused on **continuous** and **unconstrained** dependent variables, i.e. $Y \in \mathbb{R}$. However, many interesting variables are **limited** in some form.

- **Probabilities** range from 0 to 1
- **GDP** is a positive variable
- **Political orientation** is a categorical variable

Until now, we have treated these variables as approximately continuous, but this may cause severe issues.

A solution is to use specialised **limited dependent variable** (LDV) models.

Example of LDVs

We can distinguish between **classification** and **regression**.

Classification: we speak of a classification model when the outcome is

- **Binary**: pass vs. fail, good vs. bad, sick vs. not, employed vs. unemployed
- **Categorical**: nationality, dog breeds, means of transport; or **ordinal**: good – okay – bad

Regression: we speak of a regression model when the outcome is

- **Censored, truncated, or positive**: wages, wealth, time
- **Count data**: number of votes, days since an accident

Regression is generally used in a broader sense and may encompass **classification**.

Limited Dependent Variables

Linear Probability Model

Modeling Probabilities

Interpretation

Inference

Linear Probability Model

The **Linear Probability Model** is an **OLS regression** applied to a binary dependent variable $\boldsymbol{y} \in \{0, 1\}$ as in:

$$Y = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

The model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$$

The expected value of the **dependent variable** is equal to the probability p that $\boldsymbol{y} = 1$.

Conditional on the regressor $\boldsymbol{X}$, we have:

$$\mathbb{E}[y \mid \boldsymbol{X}] = \mathbb{P}(y = 1 \mid \boldsymbol{X}) = \boldsymbol{X}\beta.$$

Understanding the LPM

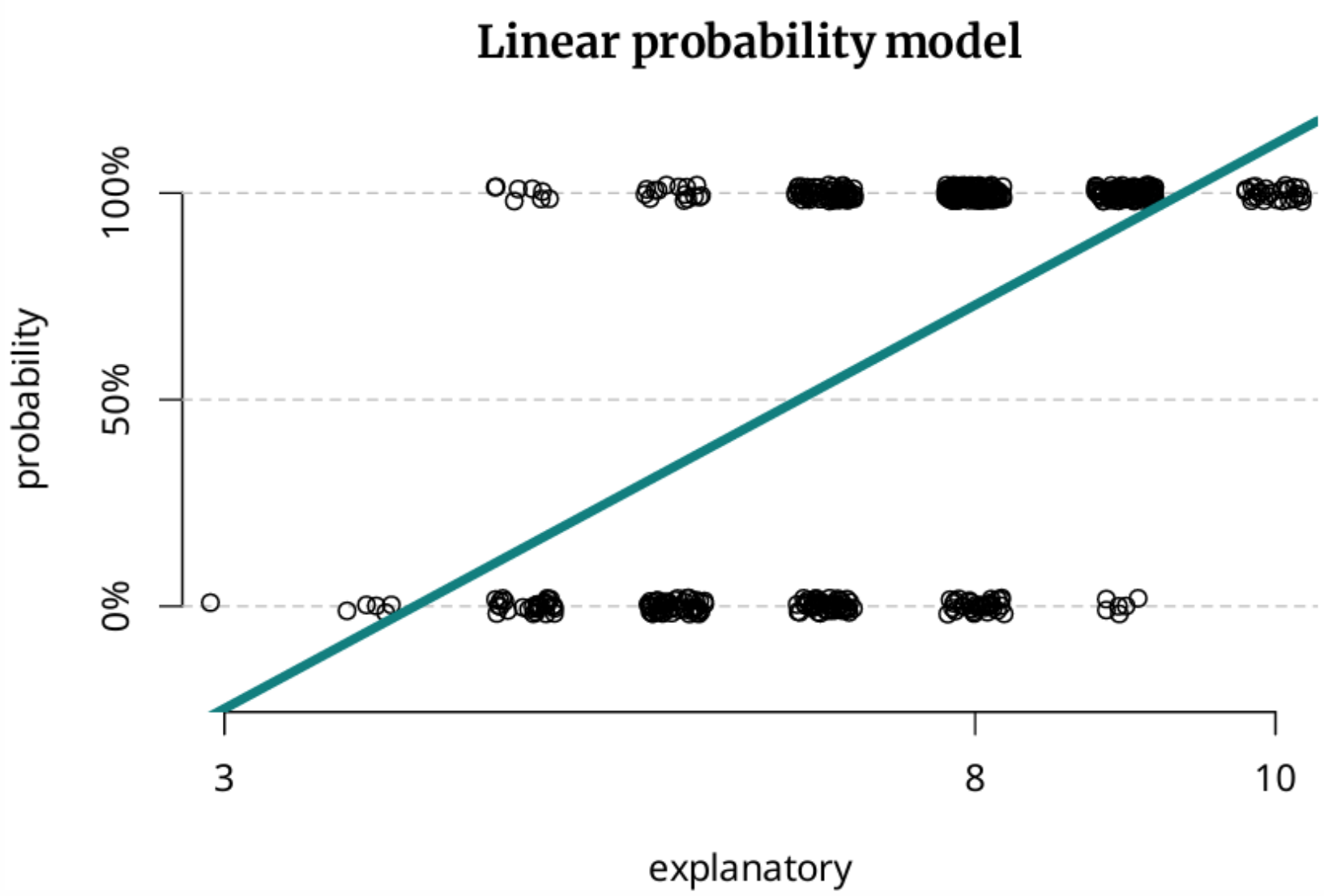
The **LPM** implies that β_j gives us the **expected absolute change in probability** when $\boldsymbol{x}_j$ increases by 1. This linearity assumption can be a major limitation, as probabilities are naturally nonlinear.

$$\mathbb{P}(y \mid \boldsymbol{X}) = \beta_0 + \boldsymbol{x}_1\beta_1 + \cdots + \boldsymbol{x}_k\beta_k$$

This linearity assumption can be a **major limitation** because:

- probabilities are naturally **nonlinear**
- **constant marginal effects** are unrealistic
- the linear prediction can exceed **0 or 1**

Problems with the LPM



Limited Dependent Variables

Linear Probability Model

Modeling Probabilities

Interpretation

Inference

Count Data

Modeling Probabilities

- When dealing with **probabilities**, the linearity assumption for f may be too strong. We need another approach.
- We consider a function G that satisfies $0 < G(z) < 1$.
- We can use G to adapt our model to

$$P(\boldsymbol{y} \mid \boldsymbol{X}) = G(\boldsymbol{X}\beta).$$

- In this way, we model a latent variable $\boldsymbol{z} = \boldsymbol{X}\beta$ using a linear model and link it to the dependent variable $\boldsymbol{y}$ via the **non-linear** function G, giving us $\boldsymbol{y} = G(\boldsymbol{z})$.
- The inverse function $G^{-1}(z)$ is called the **link function**.
- We can use different functional forms for G.

Cumulative Distribution Function

The probability function $G(\boldsymbol{X}\beta)$ comes directly from the CDF of the error term in the latent variable model.

Start from $\boldsymbol{y}^* = \boldsymbol{X}\beta + u$, where:

- $\boldsymbol{y}^*$ is an unobserved continuous propensity
- u is a random error term

We only observe:

$$y = \begin{cases} 1 & \text{if } y^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that the expected value of y depends on the distribution of u:

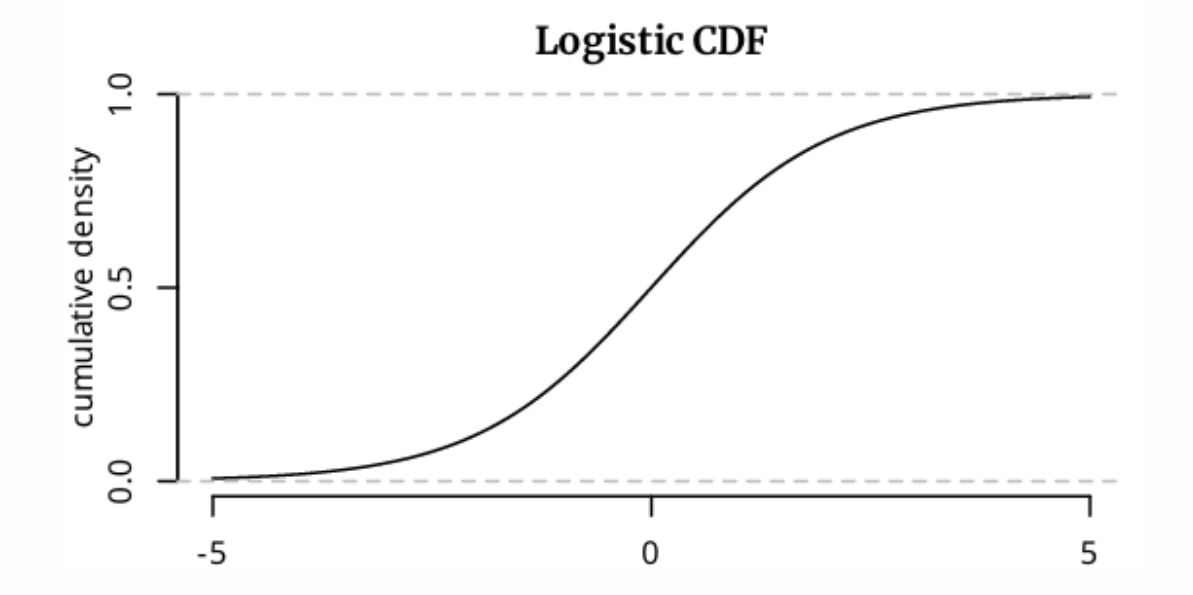
$$\mathbb{P}(\boldsymbol{y} = 1 \mid \boldsymbol{X}) = \mathbb{P}(\boldsymbol{y}^* > 0|\boldsymbol{X}) = \mathbb{P}(u > -\boldsymbol{X}\beta) = \mathbb{P}(u < \boldsymbol{X}\beta) = F_u(\boldsymbol{X})\beta)$$

The Logit Model

For the **logit model**, we use the cumulative distribution function (CDF) of a logistic variable — the logistic function — for G.

The link function is the **log-odds**: $\log \frac{p}{1-p}$.

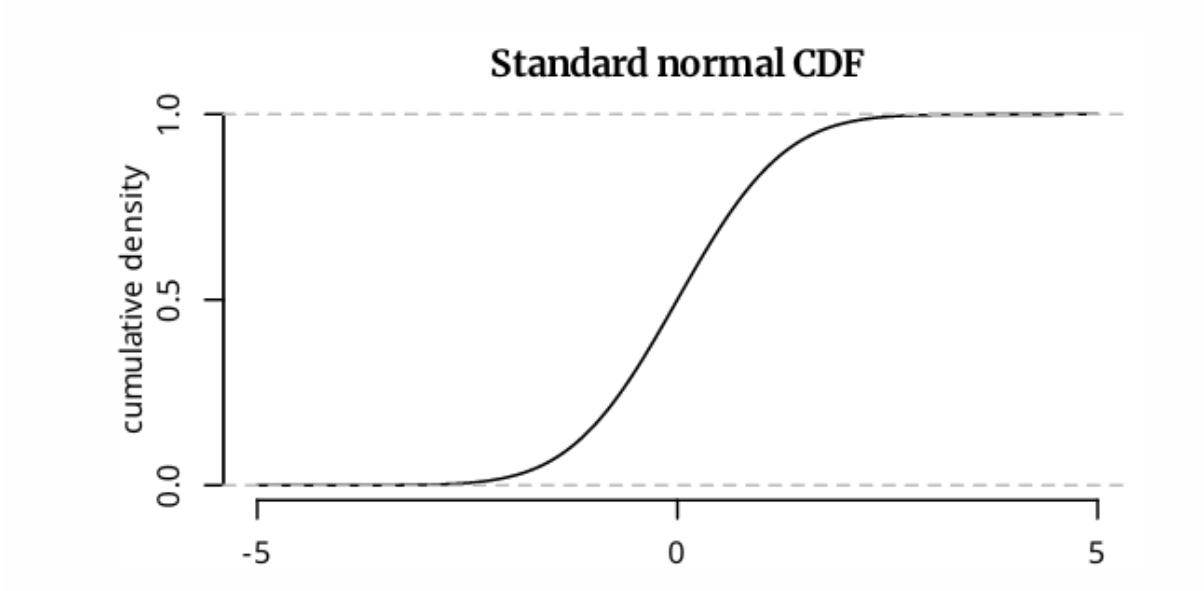
$$G(\boldsymbol{z} = \frac{e^z}{e^z + 1})$$



The Probit Model

For the **probit model**, we use the CDF of a standard normal distribution, which gives us the probability that the standard normal variable Z is smaller than z.

$$G(z) = \Phi(z) = \mathbb{P}(Z \leq z), \quad \text{where } Z \sim \mathcal{N}(0, 1).$$



Limited Dependent Variables

Linear Probability Model

Modeling Probabilities

Interpretation

Inference

Count Data

General Models

Interpretation

The interpretation of non-linear models such as logit and probit is not as straightforward as in linear models due to their non-linearity.We can intrepret:

- sign of the coefficients, the direction of the expected change
- significance of coefficients

If $\beta_j > 0$ we expect the probability to increase with $\boldsymbol{x_j}$.

However, we cannot interpret the magnitude of coefficients as magnitude of the effect of $\boldsymbol{X}$ on $\boldsymbol{y}$. Instead it captures the effects of $\boldsymbol{X}$ on the latent $\boldsymbol{z}$, which we rarely care about.

Partial effects

The problem with interpreting coefficients is that the partial effects of x_j are affected by all
other variables. Assume x_1 is a dummy. Then:

$$\mathbb{P}(y \mid x_1 = 1, x_2, \ldots) = G(\beta_0 + \beta_1 + x_2\beta_2 + \ldots)$$

$$\mathbb{P}(y \mid x_1 = 0, x_2, \ldots) = G(\beta_0 + x_2\beta_2 + \ldots)$$

The change depends on the level of x_2 and other variables.

The same holds for continuous variables, where the partial effect is given by:

$$\frac{\partial \mathbb{P}(y \mid x_j = x_{j,\cdot})}{\partial x_j} = G(z)$$

where $g(z) = G'(z)$, i.e., the first derivative of the link function.

Partial effects (2)

We can use **summary measures** to help interpret partial effects in non-linear models.

- **Partial effect at the average**: the partial effect at the average evaluates the effect when the **explanatory variables are at their mean**.

$$g(\bar{X}\hat{\beta})\hat{\beta}_j$$

- **Average partial effect**: we calculate the **partial effect for each observation** and then take the average.

$$\frac{1}{N}\sum_{j=1}^{N} G(X\hat{\beta})\hat{\beta}_j$$

Linear Probability Model

Modeling Probabilities

Interpretation

Inference

Count Data

General Models

Maximum Likelihood Estimation

Testing

To test the significance of single coefficients, we can use t values.

For multiple coefficients we can use the **likelihood ratio** test

$$LR = 2(\log \mathcal{L}_u - \log \mathcal{L}_r)$$

We compare the likelihood of the **unrestricted** ($\mathcal{L}_u$) and **restricted** ($\mathcal{L}_r$) models, where the models are required to be nested (the complex model nests the simpler one).

Likelihood: The likelihood function is the joint probability of the observed data, viewed as a function of the parameters.

Comparing Models

We can compare model specifications using

- R^2, the proportion of **explained variance**, for non-linear models there are various pseudo R^2 measures,
- the **likelihood**, $\mathcal{L}$, of a given model,
- **information criteria** (IC), which rewards model fit but penalizes model complexity

Many measures of model fit always increase with complexity — IC prefer **parsimony**.

Akaike information criterion

$$\mathrm{AIC} = 2K - 2\log\hat{\mathcal{L}}$$

Bayesian (or Schwarz) information criterion:

$$\mathrm{BIC} = K\log N - 2\log\hat{\mathcal{L}}$$

Other Probability Models

Probabilities are not the only limited dependent variables, and there is a range of other **specialised models**. This includes the:

- **Poisson model** for count variables,
  e.g. $Y \in \{0, 1, 2, \ldots\}$ for votes
- **Tobit model** for censored variables,
  e.g. $Y > 0$ for forest loss
- **Heckit model** for non-random samples,
  which uses the Heckman correction by modeling the sampling probability
- **Multinomial probit/logit model** for categorical variables,
  e.g. $Y \in \{\text{agree}, \text{disagree}, \text{unsure}\}$

Modeling Probabilities

Interpretation

Inference

Count Data

General Models

Maximum Likelihood Estimation

Linear Model and Maximum Likelihood

Count Data

Count data take **non-negative integer values** $(0, 1, 2, \ldots)$ and often include a substantial number of zero outcomes ("zero-inflated").

To build a model for this kind of data, we could:

- think of a **latent normal variable** behind Y, or
- use a **discrete probability distribution**.

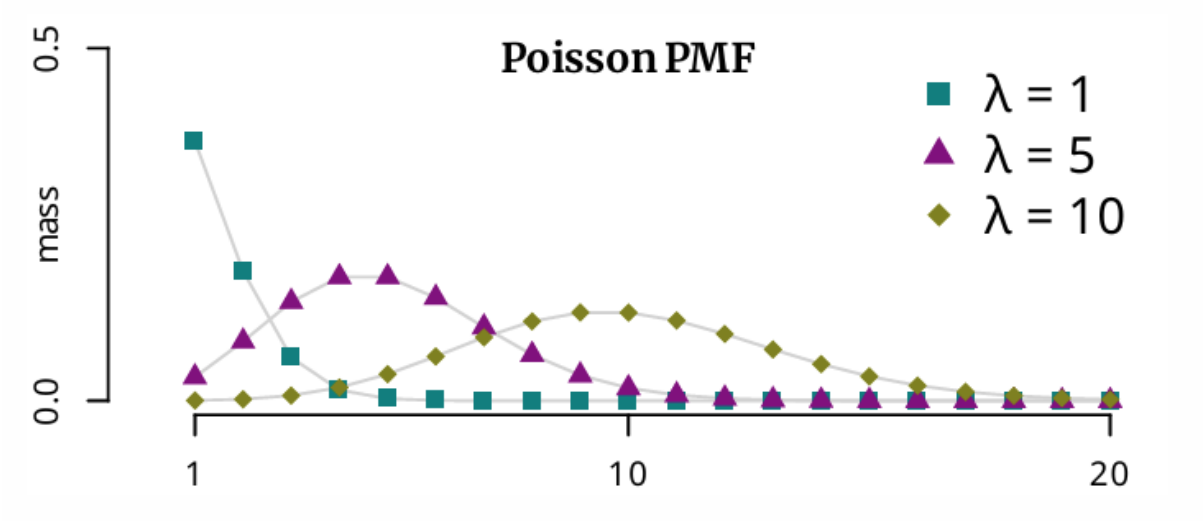
The Poisson distribution is one example; we can use it to express the probability that a given number of events occurs in a fixed interval.

Poisson Distribution

The probability mass function (PMF) of the Poisson distribution is

$$\mathbb{P}(Y = y_i \mid \lambda) = \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!}, \qquad y_i = 0, 1, 2, \dots$$

where the parameter λ is also the expectation $\mathbb{E}[Y]$ and the variance $\mathbb{V}(Y)$.



Poisson Distribution(2)

We generally expect that the expectation, i.e. the mean $\lambda = \mathbb{E}[y]$, depends on other variables. Consider a **Poisson model** with a dependent mean; let

$$\lambda = \mathbb{E}[y \mid X; \beta] = \exp(X\beta),$$

where we use the exponential function to ensure that $\mathbb{E}[y \mid X] > 0$. We obtain

$$\mathbb{P}(Y = y_i \mid x_i; \beta) = \frac{\exp(x_i\beta)^{\,y_i}\ \exp(-\exp(x_i\beta))}{y_i!},$$

describing the probability of each observation.

Interpretation

Inference

Count Data

General Models

Maximum Likelihood Estimation

Linear Model and Maximum Likelihood

Shrinkage

General Models

- We need a better way to **estimate** more **general models**, such as probit and logit, which are **non-linear in parameters**.

$$\boldsymbol{y} = G(\boldsymbol{X}, \boldsymbol{\beta}) + \boldsymbol{u}$$

- If we apply the OLS method, we should minimise (′).
- This will be problematic because we need to consider (K) ((^K)) partial derivatives, and there is no closed-form solution.
- Non-linear least squares estimation is a conceptually straightforward approach. First, we approximate with a linear model, and then refine the estimates iteratively. However, estimates are generally not unique and inefficient — **OLS is not BLUE**.

Inference

Count Data

General Models

Maximum Likelihood Estimation

Linear Model and Maximum Likelihood

Shrinkage

Summary

Maximum Likelihood Estimation

- **Maximum Likelihood** estimation is a method for estimating parameters.
- It maximizes a likelihood function, **the joint probability distribution of the data** as a function of the parameters:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \mathbb{P}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$$

- Intuitively, we set $\boldsymbol{\beta}_{ML}$ such that the observed data is **most probable** within our model
- The resulting estimator is **consistent, asymptotically normal, and asymptotically efficient** in most cases
- The likelihood $\mathcal{L}(\theta \mid X)$ itself is not a probability - θ can vary, not X

Maximum Likelihood Estimation (2)

- To make the computation easier we usually work with **log-likelihood**

$$\ell(\boldsymbol{\beta}) = \log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log \mathbb{P}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$$

- $\boldsymbol{\beta}_{ML}$ is then the estimate that maximises the **log-likelihood function**
- The equation $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta} = 0$ generally has no closed form solution, and iterative optimization algorithms are used such as **Gradient Descent** and **Newton's method**

Maximum Likelihood for binary outcomes

- A **distributional assumption** lies at the center of Maximum Likelihood estimation.
- For **binary outcomes**, where $Y \in (0, 1)$, we can use the **Bernoulli distribution** with proability mass function:

$$f(y_i \mid p) = p^{y_i}(1-p)^{1-y_i}$$

- Given the independence of observations, the **joint probability** is

$$f(y_1, y_2, \ldots, y_n \mid p) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i}$$

Maximum Likelihood for binary outcomes (2)

- With a **Bernoulli** outcome, we can use the following **likelihood**

$$\mathcal{L}(p) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i}$$

- To find p_{ML}, we maximise the likelihood by solving $\frac{\partial \mathcal{L}}{\partial p} = 0$
- The product form of the likelihood is difficult to differentiate — we would prefer a sum.
- Using properties of the logarithm, we instead maximise the log-likelihood.
- We therefore solve:

$$\frac{\partial \ell}{\partial p} = \frac{\partial \sum_i \log\left[p^{y_i}(1-p)^{1-y_i}\right]}{\partial p} = 0$$

Maximum Likelihood for binary outcomes (3)

To obtain the ML estimate, we first reformulate the log-likelihood as

$$\begin{aligned}\ell(p) &= \sum_{i=1}^{N} \log\left[p^{y_i}(1-p)^{1-y_i}\right] \\ &= \sum_{i=1}^{N} y_i \log p + (1-y_i)\log(1-p) \\ &= N\bar{y}\log p + N(1-\bar{y})\log(1-p).\end{aligned}$$

Where the last step relates the summation to the sample mean:

$$\sum_{i=1}^{N} y_i = N\bar{y}.$$

Maximum Likelihood for binary outcomes (4)

We know that

$$\ell(p) = N\bar{y}\log p + N(1-\bar{y})\log(1-p),$$

which we need to differentiate with respect to p and solve for p_{ML}.

$$\begin{aligned}
\frac{\partial \ell(p)}{\partial p} &= \frac{N\bar{y}}{p} - \frac{N(1-\bar{y})}{1-p} = 0 \\
\frac{N\bar{y}}{p} &= \frac{N(1-\bar{y})}{1-p} \\
\bar{y}(1-p) &= p(1-\bar{y}) \\
p_{ML} &= \bar{y}.
\end{aligned}$$

The maximum likelihood estimate is the **average number of occurrences in the sample**.

Maximum Likelihood for logit models

With logit models, we have a Bernoulli outcome Y, and model the probability p using the logistic function. We have the following PMF:

$$\begin{aligned}\mathbb{P}(Y = y_i \mid x_i) &= p^{y_i}(1-p)^{1-y_i} \\ &= \left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)^{y_i}\left(1-\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)^{1-y_i}\end{aligned}$$

We then set β_{ML} by (numerically) maximising the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{N}\left[-\log\left(1+e^{x_i\beta}\right) + y_i x_i \beta\right].$$

Maximum Likelihood for poisson models

With Poisson models, we have a Poisson outcome Y, and model the mean λ using an exponential function. We have the following PMF:

$$\mathbb{P}(Y = y_i \mid x_i) = \frac{\exp(x_i\beta)^{y_i} \exp^{-\exp(x_i\beta)}}{y_i!}.$$

We then set β_{ML} by (numerically) maximising the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{N} [y_i x_i \beta - \exp(x_i\beta)].$$

Count Data

General Models

Maximum Likelihood Estimation

Linear Model and Maximum Likelihood

Shrinkage

Summary

Exercises

Linear Model and Maximum Likelihood

- Consider the **standard linear model** with normally distributed errors, given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \qquad u \sim \mathcal{N}(0, \sigma^2).$$

- This implies that $y \sim \mathcal{N}(X\beta, \sigma^2)$
- So far, we have used **ordinary least squares** to estimate the parameters — now we can also use **maximum likelihood estimation**.
- The Normal distribution, denoted by $\mathcal{N}(\mu, \sigma^2)$, has the probability density function

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Linear Model and Maximum Likelihood (2)

We can obtain the likelihood function from the PDF:

$$\mathcal{L}(\beta, \sigma^2) = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N} \exp\left\{ -\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) \right\}.$$

To obtain estimates, we work with the log-likelihood:

$$\ell(\beta, \sigma^2) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

We will focus on β_{ML} — notice how the last term measures the **squared deviations**.

Linear Model and Maximum Likelihood (3)

To find β_{ML}, we need to maximise the log-likelihood:

$$\ell(\beta, \sigma^2) = \frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

When taking the derivative, the first two terms drop out, and we obtain:

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta} = -2\sigma^{-2}\left(-2X'y + 2X'X\beta\right).$$

- We obtain β_{ML} from $\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta} = 0$
- And check whether $\ell(\beta, \sigma^2)$is maximal by checking the second derivative.

For the linear model with Normal errors, the **OLS and ML estimates of β coincide**.

General Models

Maximum Likelihood Estimation

Linear Model and Maximum Likelihood

Shrinkage

Summary

Exercises

References

Shrinkage estimators - LASSO

Let's discard the **constraint of unbiased estimators**.

- Theoretically, there is an **unlimited number of regressors**; most are irrelevant.
- We only want to **keep important regressors**, and pull coefficients towards the mean

How can we achieve this in the linear model?

$$\hat{\beta} = \arg\min_{\beta} \left\{ (y - X\beta)'(y - X\beta) \right\}$$

$$\hat{\beta} = \arg\min_{\beta} \left\{ (y - X\beta)'(y - X\beta) + \lambda|\beta| \right\}.$$

We can introduce various **penalty terms** to punish larger coefficient values.

Maximum Likelihood Estimation

Linear Model and Maximum Likelihood

Shrinkage

Summary

Exercises

References

Summary

- ML estimators are based on the **probability distribution** of Y.
- We learn about the:
  - **parameters** of this underlying distribution,
  - **conditional** on the data we observe and the chosen distribution.

To find an ML estimator, we:

(1) Model the **probability** of each observation.

(2) Derive the **joint probability** of all observations.

(3) Consider the joint probability as a function of its parameters θ, conditional on the data $\mathcal{D}$. This gives us the **likelihood function** $\mathcal{L}(\theta \mid \mathcal{D})$.

(4) Maximise the **log-likelihood** $\ell(\theta \mid \mathcal{D})$ with respect to θ.

Linear Model and Maximum Likelihood

Shrinkage

Summary

Exercises

References

Exercises

- Derive the log-likelihood for the logit model (result slide 11)
- Derive the log-likelihood for the poisson model (result slide 12)

Shrinkage

Summary

Exercises

References