ECON6300/7320/8300 Advanced Microeconometrics

Maximum Likelihood

Christiern Rose

[1] University of Queensland

Lecture 3

Likelihood principle

- **Likelihood principle:** choose as estimator of the parameter vector θ^0 that value of θ that maximizes the likelihood of observing the actual sample.
- Joint probability mass function or density $f(\mathbf{y}, \mathbf{X}|\theta)$ is a function of θ given the data $(\mathbf{y}, \mathbf{X})$.

- **Likelihood function** is denoted by $\mathrm{L}_N(\theta|\mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} f(y_i|\mathbf{x}_i, \theta)$.
- Maximizing $\mathrm{L}_N(\theta)$ is equivalent to maximizing the **log-likelihood function:** $\mathcal{L}_N(\theta) = \ln L_N(\theta) = \sum_{i=1}^{N} \ln f_i(\mathbf{y}, \mathbf{X}|\theta)$

Likelihood principle (2)

- Likelihood function $L_N(\theta) = f(\mathbf{y}, \mathbf{X}|\theta) = f(\mathbf{y}|\mathbf{X}, \theta) f(\mathbf{X}|\theta)$ requires specification of both the conditional density of $\mathbf{y}$ given $\mathbf{X}$ and the marginal density of $\mathbf{X}$.
- Under exogeneity assumption $f(\mathbf{X})$ depends on mutually exclusive sets of parameters and can be ignored.
- Objective function is the average log-likelihood function

$$Q_N(\theta) = N^{-1}\mathcal{L}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ln f(y_i|\mathbf{x}_i, \theta), \tag{1}$$

Equivalence of MLE and OLS under normality

- A linear regression model

$$y = x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k + u$$

where $u \sim \mathcal{N}(0, \sigma^2)$, least squares and MLE estimators of β are equivalent.

- To show this result note that

$$
\begin{aligned}
f(y_i) &= [2\pi\sigma^2]^{-1/2} e^{-(y_i-\mu_i)^2/2\sigma^2} \\
\mu_i &= \mathbf{x}_i'\beta \, ,
\end{aligned}
$$

Then given independence of u_i the likelihood is

$$
\begin{aligned}
L_N(\beta, \sigma) &= \prod_{i=1}^{N} f(u_i) = \prod_{i=1}^{N} [2\pi\sigma^2]^{-1/2} e^{-(y_i-\mu_i)^2/2\sigma^2} \\
\ln L_N(\beta, \sigma) &= -\frac{N}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{N} (y_i - \mu_i)^2/2\sigma^2
\end{aligned}
$$

Equivalence of MLE and OLS under normality (2)

- To maximize the log-likelihood wrt β, we need to minimize the second term

$$\sum_{i=1}^{N}(y_i - \mu_i)^2/2\sigma^2 = \sum_{i=1}^{N}(y_i - \mathbf{x}_i'\beta)^2/2\sigma^2.$$

- I.e., find the value of β which minimizes the sum of squares function $\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\beta)^2/2\sigma^2$.
- However, this is the least squares criterion function. It follows that the first-order conditions for maximizing likelihood with respect to β, or minimizing the residual sum of squares with respect to β, are the same. This means that OLS and MLE of β are equivalent when the errors are assumed to be normal distributed.

Equivalence of MLE and OLS under normality (3)

- This result does **not** depend upon linearity of μ_i. The result would hold if, for example, we had $\mu_i = \exp(\mathbf{x}_i'\beta)$.
- This result in general will not hold for other distributions. To verify or illustrate this claim, check it out for the following distributions associated with nonlinear models.

Model	**Range of y**	**Density $f(y)$**	**Common Parameterization**
Exponential	$(0, \infty)$	$\lambda e^{-\lambda y}$	$\lambda = e^{\mathbf{x}'\beta}$ or $1/\lambda = e^{\mathbf{x}'\beta}$
Poisson	$0, 1, 2, \ldots$	$e^{-\lambda}\lambda^{y}/y!$	$\lambda = e^{\mathbf{x}'\beta}$

Motivation for studying MLE (or fully parametric methods)

1. Classical method whose properties have been very well studied.
2. MLE benchmarks other widely used estimators
3. Makes strong distributional assumptions whose role we want to understand
4. Forms the basis of the Bayesian approach which is widely used

ML estimator

- MLE is defined by the solution of the first-order conditions ("score equations")

$$\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta} = \sum_{i=1}^{N} \frac{\partial \ln f(y_i|\mathbf{x}_i, \theta)}{\partial \theta} = \mathbf{0}. \tag{2}$$

- For regular problems with unique maximum of the likelihood, FOC and maximizing likelihood are equivalent. In this case the log-likelihood is **globally concave.**
- For some problems the likelihood may have more than one mode (local maximum).

ML estimator (2)

- Gradient vector $s(\theta) = \partial \mathcal{L}_N(\theta)/\partial\theta$ is called the **score**, or **efficient score** when evaluated at θ_0

- (*Check*!) Score function in the case of normal linear regression:

$$\frac{\partial \mathcal{L}_N(\beta, \sigma^2)}{\partial \beta} = \mathbf{X}'\mathbf{u}/(\sigma^2) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)/\sigma^2 = \mathbf{0}$$

or

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \mathbf{0} \Longrightarrow \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\widehat{\beta}$$

- What is the score equation for σ^2? How does its MLE compare with s^2?

MLE for Nonlinear Regression

- Discrete choice or qualitative response models are for y that takes only a finite number of discrete values.
- Here we consider binary outcome models where only two values are taken: 0 and 1.
- Particularly logit and probit models, which are **nonlinear models**.
- Extremely widely used models in social sciences

MLE for Nonlinear Regression (2)

- General properties of binary outcome models
- Probit, logit, LPM and OLS models.
- Latent variable formulations, especially random utility model.

Simple binary outcome model

- The coin toss example of introductory statistics.
- Let p denote the probability of a head ($y = 1$) on one coin toss.
- Then $\Pr[y = 1] = p$ and $\Pr[y = 0] = 1 - p$.
- For N tosses y_i is the i^{th} of N independent realizations of head or tail.
- The MLE for p is the sample mean $\bar{y}$,
  i.e. the proportion of tosses that are heads.

Bernoulli distribution

- $\Pr[y = 1] = p$ and $\Pr[y = 0] = 1 - p$.
- Compact expression for density

$$f(y) = p^y(1-p)^{1-y}.$$

- This is Bernoulli density which is the binomial with one trial per observation.
- Moments
  - $\mathrm{E}[y] = 1 \times p + 0 \times (1-p) = p$
  - $\mathrm{V}[y] = (1-p) \times p + (0-p) \times (1-p) = p(1-p)$.
- Note that p can be interpreted as $\mathrm{E}[y]$ or as $\Pr[y = 1]$.

Examples

- Economic applications are
  - labor supply: $y = 1$ if work and $y = 0$ otherwise
  - insurance status: $y = 1$ if have private health insurance, $y = 0$ otherwise
- Assumption of independent trials may be reasonable.
- Assuming a constant probability p for each trial is not reasonable. It should depend on an individual's characteristics.
- Extend the Bernoulli model so p_i may be a function of regressors $\mathbf{x}_i$.

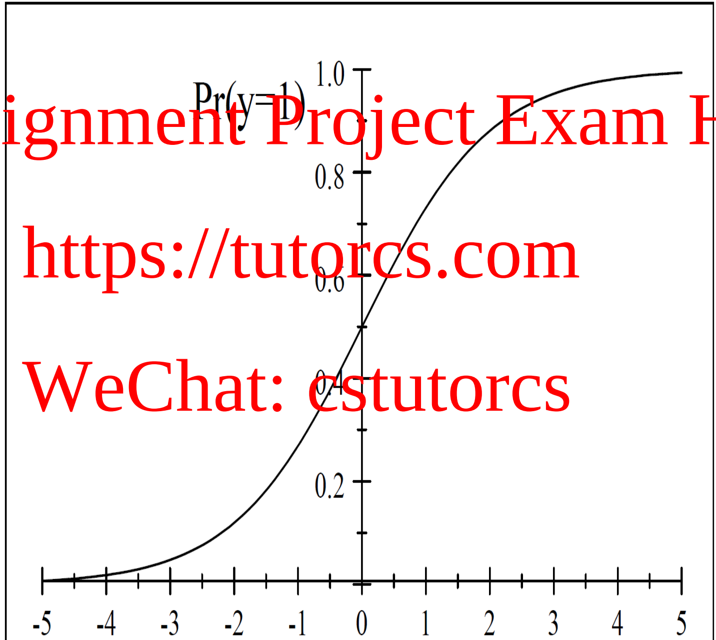
Binary outcome models

- Regression model formed by parameterizing p_i to depend on regressors $\mathbf{x}_i$ and parameters β.
- Usually specify single-index model

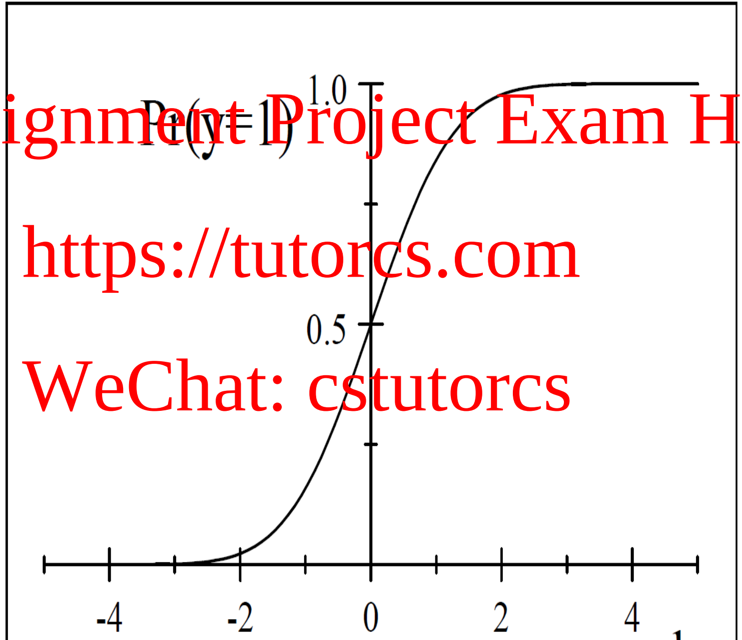
$$\mathrm{E}[y_i|\mathbf{x}_i] = p_i = F(\mathbf{x}_i'\beta).$$

- Usually choose $F(\cdot)$ to be a cumulative distribution function (cdf).
- Then $0 \leq F(\cdot) \leq 1 \Rightarrow 0 \leq p_i \leq 1$.
  - logistic cdf gives logit model.
  - standard normal cdf gives probit model.

Logistic CDF



Normal CDF



Maximum likelihood (MLE)

- Density

$$f(y) = p^y(1-p)^{1-y}, \quad p = F(\mathbf{x}'\beta)$$

$$\Rightarrow \ln f(y) = y \ln F(\mathbf{x}'\beta) + (1-y)\ln(1 - F(\mathbf{x}'\beta))$$

- Log-likelihood function is

$$\mathcal{L}_N(\beta) = \sum_{i=1}^{N} \left\{ y_i \ln F(\mathbf{x}_i'\beta) + (1-y_i)\ln(1 - F(\mathbf{x}_i'\beta)) \right\}.$$

- Let $F'(z) = \partial F(z)/\partial z$. MLE solves

$$\sum_{i=1}^{N} \left\{ \frac{y_i}{F(\mathbf{x}_i'\beta)} F'(\mathbf{x}_i'\beta)\mathbf{x}_i - \frac{1-y_i}{1-F(\mathbf{x}_i'\beta)} F'(\mathbf{x}_i'\beta)\mathbf{x}_i \right\} = \mathbf{0}.$$

Asymptotic Distribution OF MLE

- The MLE FOC simplify to

$$\sum_{i=1}^{N} \frac{y_i - F(\mathbf{x}_i'\beta)}{F(\mathbf{x}_i'\beta)(1 - F(\mathbf{x}_i'\beta))} F'(\mathbf{x}_i'\beta)\mathbf{x}_i = \mathbf{0}$$

$$\sum_{i=1}^{N} \left[\frac{y_i - F(\mathbf{x}_i'\beta)}{\sqrt{F(\mathbf{x}_i'\beta)(1 - F(\mathbf{x}_i'\beta))}} \right] \frac{F'(\mathbf{x}_i'\beta)\mathbf{x}_i}{\sqrt{F(\mathbf{x}_i'\beta)(1 - F(\mathbf{x}_i'\beta))}} = \mathbf{0}.$$

- General ML result if density correctly specified
- For binary outcome MLE

$$\widehat{\beta}_{ML} \overset{a}{\sim} \mathcal{N}\left[\beta_0, \left(-\mathrm{E}[\partial^2 \mathcal{L}_N / \partial\beta\partial\beta'] \Big|_{\beta_0} \right)^{-1} \right]$$

$$\overset{a}{\sim} \mathcal{N}\left[\beta_0, \left(\sum_{i=1}^{N} \frac{1}{F(\mathbf{x}_i'\beta_0)(1 - F(\mathbf{x}_i'\beta_0))} F'(\mathbf{x}_i'\beta_0)^2 \mathbf{x}_i\mathbf{x}_i' \right)^{-1} \right]$$

Misspecification

- For binary data the dgp density is always Bernoulli as

$$\Pr[y=1]=p$$
$$\Rightarrow \Pr[y=0]=1-\Pr[y=1]=1-F(\mathbf{x}'\beta).$$

- Therefore only possible misspecification of dgp is if $p \neq F(\mathbf{x}'\beta)$.

- Clearly inconsistent estimator if $p \neq F(\mathbf{x}'\beta)$ as then

$$\mathrm{E}[y_i - F(\mathbf{x}_i'\beta)] \neq 0$$

leading to left-hand side of the FOC not having expected value $\mathbf{0}$.

Logit model

- Logit is a widely used functional form, especially in biometrics

- Computationally convenient.
- The logit model specifies

$$p = \Lambda(\mathbf{x}'\beta) = \frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}},$$

- $\Lambda(z) = e^z/(1 + e^z) = 1/(1 + e^{-z})$ is the logistic cdf.

- The derivative $\Lambda'(z) = \Lambda(z)(1 - \Lambda(z))$ is the logistic density.
- For this reason also called logistic regression model .

Logit MLE

- The logit ML conditions simplify to

$$\sum_{i=1}^{N}(y_i - \Lambda(\mathbf{x}_i'\beta))\mathbf{x}_i = \mathbf{0}.$$

- FOC are nonlinear in parameters
- Notice that there is no "error term" in the binary model.
- The logit MLE has distribution

$$\widehat{\beta}_{Logit} \overset{a}{\sim} \mathcal{N}\left[\beta_0, \left(\sum_{i=1}^{N}\Lambda(\mathbf{x}_i'\beta_0)(1-\Lambda(\mathbf{x}_i'\beta_0))\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right].$$

Probit model

- The probit model specifies

$$p = \Phi(\mathbf{x}'\beta).$$

- $\Phi(z) = \int_{-\infty}^{z} \phi(s)ds = \int_{-\infty}^{z} (1/\sqrt{2\pi}) \exp(-s^2/2)ds$ is the c.d.f. of the standard normal.

- The derivative $\Phi'(z) = \phi(z) = (1/\sqrt{2\pi}) \exp(-z^2/2)$ is the standard normal p.d.f.
- The FOC do not simplify, unlike logit case.
- The probit MLE has distribution

$$\widehat{\beta}_{Probit} \overset{a}{\sim} \mathcal{N}\left[\beta_0, \left(\sum_{i=1}^{N} \frac{\phi(\mathbf{x}_i'\beta_0)^2}{\Phi(\mathbf{x}_i'\beta_0)(1-\Phi(\mathbf{x}_i'\beta_0))}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right].$$

Linear probability model (LPM)

- The LPM specifies

$$p = \mathbf{x}'\beta.$$

- The LPM MLE FOC are

$$\sum_{i=1}^{N} \frac{y_i - \mathbf{x}_i'\beta}{\mathbf{x}_i'\beta(1 - \mathbf{x}_i'\beta)} \mathbf{x}_i = \mathbf{0}$$

- The LPM model has the obvious weakness of permitting probabilities outside the [0, 1] interval.

- Furthermore, the MLE estimator can be numerically unstable if $\mathbf{x}_i'\beta$ close to 0 or 1.

OLS

- The LPM is more simply estimated by OLS, which also specifies $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\beta$.
- The LPM OLS FOC are

$$\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\beta)\mathbf{x}_i = \mathbf{0},$$

- Important to allow for the intrinsic heteroskedasticity of binary data

$$\widehat{\beta}_{OLS} \overset{a}{\sim} \mathcal{N}\left[\beta_0, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right]$$

where for Ω use

$$\widehat{\Omega} = \mathrm{Diag}[(y_i - \mathbf{x}_i'\widehat{\beta})^2]$$

How to interpret coefficients (1)

- Coefficients in different models are not directly comparable **due to different scaling.**
- Instead compare across models effect of a one unit change in regressors on $\mathrm{P}[y = 1|\mathbf{x}] = \mathrm{E}[y|\mathbf{x}]$.
- Now

$$\mathrm{E}[y|\mathbf{x}] = F(\mathbf{x}'\beta)$$
$$\partial \mathrm{E}[y|\mathbf{x}]/\partial \mathbf{x} = F'(\mathbf{x}'\beta) \times \beta$$

  where $F'(z) = \partial F(z)/\partial z$.
- Thus the effect depends on the functional form of F and the evaluation point $\mathbf{x}$, in addition to parameter β.

How to interpret coefficients (2)

- This suggests for slope parameters the rule of thumb

$$\begin{aligned}\widehat{\beta}_{Logit} &\simeq 4\widehat{\beta}_{OLS}\\ \widehat{\beta}_{Probit} &\simeq 2.5\widehat{\beta}_{OLS}\\ \widehat{\beta}_{Logit} &\simeq 1.6\widehat{\beta}_{Probit}.\end{aligned}$$

- This works quite well, for $0.1 \leq F(\mathbf{x}'\beta) \leq 0.9$.
- Better to compare marginal effects, not coefficients.

Marginal effects in binary response models

- Recall that:

$$E[y|\mathbf{x}] = P[y = 1|\mathbf{x}] = F(\mathbf{x}'\beta)$$

$$\partial E[y|\mathbf{x}]/\partial \mathbf{x} = F'(\mathbf{x}'\beta) \times \beta$$

- The marginal effects (ME) of a change in $\mathbf{x}$ thus depends on both $\mathbf{x}$ and β. We can estimate:
  - The ME at some value of the covariates $\mathbf{x}_*$: $F'(\mathbf{x}_*'\hat{\beta}) \times \hat{\beta}$
  - The ME at the average i.e. take $\mathbf{x}_* = N^{-1}\sum_{i=1}^{N} \mathbf{x}_i$
  - The average ME: $N^{-1}\sum_{i=1}^{N} F'(\mathbf{x}_i'\hat{\beta}) \times \hat{\beta}$

Odds ratio calculations for logit (1)

- For the logit model

$$\begin{aligned} p &= \exp(\mathbf{x}'\beta)/(1+\exp(\mathbf{x}'\beta)) \\ \Rightarrow \frac{p}{1-p} &= \exp(\mathbf{x}'\beta) \\ \Rightarrow \ln\frac{p}{1-p} &= \mathbf{x}'\beta. \end{aligned}$$

- $p/(1-p)$ is the odds ratio which measures the probability that $y=1$ relative to the probability that $y=0$.
- E.g. Pharmaceutical drug study where $y=1$ denotes survival and $y=0$ denotes death. An odds ratio of 2 means that the odds of survival are twice those of death.

Odds ratio calculations for logit (2)

- Statistical analyses and packages offer the option of printing the odds ratio $p/(1-p) = \exp(\mathbf{x}'\beta)$.
- Suppose the j^{th} regressor increases by one unit.
  Then $\mathbf{x}'\beta$ increases to $\mathbf{x}'\beta + \beta_j$.
  And $\exp(\mathbf{x}'\beta)$ increases to
  $\exp(\mathbf{x}'\beta + \beta_j) = \exp(\mathbf{x}'\beta) \times \exp(\beta_j)$
- Thus the odds ratio has increased by a multiple $\exp(\beta_j)$.
- E.g. a logit slope parameter of 0.1 means that a one unit change in the regressor increases the odds ratio by a multiple $\exp(0.1) \simeq 1.105$. The relative probability of $y = 1$ has increased by 10.5 percent.
- This interpretation is widely used in applied biostatistics.

Semi-elasticity interpretation

- For economists it is more natural to interpret β_j as a semi-elasticity for the odds ratio, since $\ln p/(1-p) = \mathbf{x}'\beta$.
- Then a logit slope parameter of 0.1 means that a one unit change in the regressor increases the odds ratio by a multiple 0.1.
- This coincides exactly with the interpretation used in biostatistics for very small of β_j, since then $\exp(\beta_j) = 1 + \beta_j$.

Which functional form?

- Which F – logit, probit or linear probability?
- Theoretically it depends on the data generating process (DGP)
- Unlike other applications of ML there is no problem in specifying the distribution – the only possible distribution for a (0, 1) variable is the Bernoulli.
- The problem lies in specifying a functional form for the parameter of this distribution.
- If the DGP has $p_i = \Lambda(\mathbf{x}_i'\beta_0)$ then a logit model should be used, and estimators based on other models such as probit are potentially inconsistent.
- Similar conclusions hold if instead for the dgp has $p_i = \Phi(\mathbf{x}_i'\beta_0)$ or $p_i = \mathbf{x}_i'\beta_0$.

Why Logit?

- Logit model is the binary model used by statisticians :
  - FOC and asymptotic distribution are relatively simple.
  - Logit model corresponds to the canonical link function for the binomial, a generalized linear model

  - Coefficients can be interpreted in terms of the log-odds ratio.
  - Easy generalization to multinomial logit.

Why Probit?

- The probit model is often used by economists .
  - It is motivated by a latent normal random variable.
  - So ties in with tobit models and multinomial probit.

- Empirically either logit and probit can be used
  - Little difference between results from probit and logit analysis, (after rescaling parameters).
  - Greatest difference is in prediction of probabilities close to 0 or 1.

Why LPM?

- The LPM should not be used as probabilities outside the [0, 1] interval and be numerically unstable.
- Nonetheless OLS can be useful for preliminary data analysis.

- Very widely used in the context of endogenous variables
- In practice standard errors of slope coefficients are often quite similar across logit, probit and OLS (even using the incorrect $s^2(\mathbf{X}'\mathbf{X})^{-1}$ in the case of OLS).

- Final results should, however, use probit or logit.

Measuring the fit of the model

- Several measures of model adequacy have been proposed.
- Many are very specific to binary outcome models.
- There is no single best measure.
- Approaches:
  - R-squared measures.
  - Compare $\widehat{y}$ with y.
  - Compare predicted $\widehat{Pr}[y = 1]$ with actual $\Pr[y = 1]$.

Pseudo-R-squared

- There are many R-squareds for binary models as R^2 in linear model has many interpretations
- McFadden proposed two. We favor McFadden (1974)

$$R^2 = 1 - \frac{\mathcal{L}_{fit}}{\mathcal{L}_0},$$

where

  - $\mathcal{L}_{fit}$ = log-likelihood in the fitted model
  - $\mathcal{L}_0$ is the log-likelihood in the intercept-only model.
- This R^2 should be only used for discrete choice models.

- In other nonlinear models instead use

$$R^2 = 1 - (\mathcal{L}_{\max} - \mathcal{L}_{fit})/(\mathcal{L}_{\max} - \mathcal{L}_0),$$

  where $\mathcal{L}_{\max}$ is the maximum possible value of the log-likelihood.

- For binary outcome models $\mathcal{L}_{\max} = 0$.
- For some other models $\mathcal{L}_{\max}$ can be unbounded restricting use of this.