Statistical Theory and Modeling (ST2601)
Lecture 10 - Logistic, Poisson regression and Beyond

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- **Linear Gaussian regression as a probability model**
- **Logistic regression**
- **Poisson regression**
- **Generalized linear models and beyond**

Linear Gaussian Regression

- The usual formulation for the ith observation

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$$

- The usual formulation in vector form

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$$

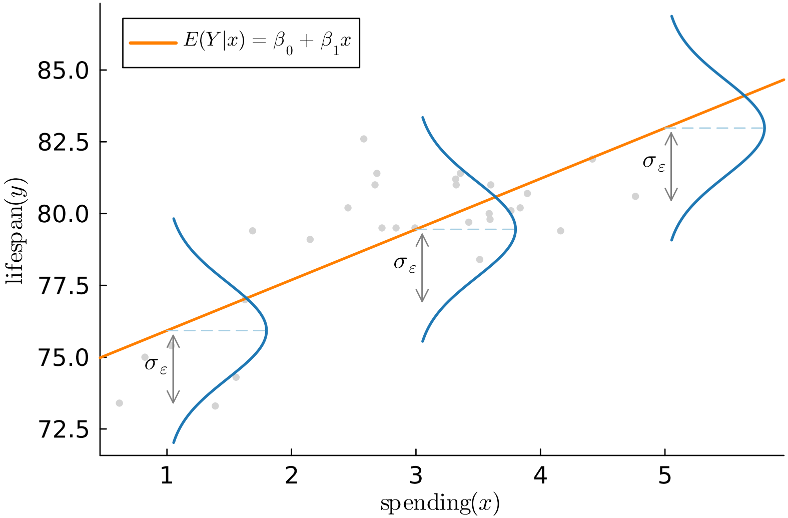
- Equivalent formulation

$$\begin{aligned} Y_i \mid \boldsymbol{x}_i &\overset{\text{ind}}{\sim} N(\mu_i, \sigma_\varepsilon^2) \\ \mu_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta} \end{aligned}$$

- Regression is a model for a **conditional distribution** $f(y|\boldsymbol{x})$.

- The ith observation has its own mean μ_i given by the **regression line**.

Regression models a conditional distribution $Y|x$



Logistic regression for binary response variable

- Assume now that the response Y_i is binary (0 or 1).
- Without covariates: model **distribution** as Bernoulli

$$Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\mu)$$

- With covariates: model **conditional distribution** as Bernoulli

$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_i)$$

- Modeling the conditional mean as $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ is no good.
- Use **logistic function** $f(z) = \frac{1}{1+e^{-z}}$ to ensure that $0 \leq \mu_i \leq 1$

$$\mu_i = \Pr(Y_i = 1 \mid \boldsymbol{x}_i) = \frac{1}{1 + e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$

- **Logistic regression**

$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_i)$$
$$\mu_i = \frac{1}{1 + e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$

Maximum likelihood for Bernoulli data

- Model: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ [Note $\mu = \mathbb{E}(Y_i) = p$]
- Bernoulli probability function

$$p(y) = \begin{cases} 1-p & \text{if } y = 0 \\ p & \text{if } y = 1 \end{cases}$$

or

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

- **Likelihood**

$$\prod_{i=1}^{n} p(y_i|p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}$$
$$= p^s(1-p)^f$$

- $s = \sum_{i=1}^n y_i$ is the number of successes
- $f = n - s$ is the number of failures.

Maximum likelihood for Bernoulli data

- **Likelihood**

$$L(p) = p^s (1-p)^f$$

- **Log-likelihood**

$$\ell(p) = s \log p + f \log(1-p)$$

- First derivative (recall: $f(x) = \log(x)$ then $f'(x) = 1/x$)

$$\ell'(p) = \frac{s}{p} - \frac{f}{1-p}$$

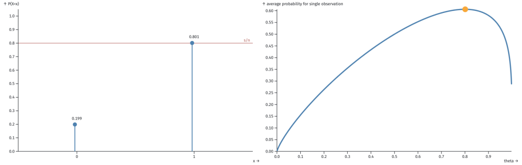
- **Maximum likelihood estimate** $\hat{p}$ is the p that solves

$$\ell'(p) = \frac{s}{p} - \frac{f}{1-p} = 0$$

which has solution

$$\hat{p} = \frac{s}{n}$$

Maximum likelihood for Bernoulli data - widget



Maximum likelihood for Logistic regression 😘

- **Logistic regression**

$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_i)$$
$$\mu_i = \frac{1}{1 + e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$

- **Data**: responses $\boldsymbol{y}$ $(n \times 1)$ and covariates $\boldsymbol{X}(n \times p)$.
- **Likelihood function** (covariates assumed fixed, non-random

$$\begin{aligned}
L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i) \\
&= \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}\right)^{y_i} \left(\frac{e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1+e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}\right)^{1-y_i}
\end{aligned}$$

- **Numerical maximization** with optim.

Poisson regression for count data

- Assume now that the response Y_i is a count $(0, 1, 2, \dots)$.
- Without covariates: **distribution** is Poisson

$$Y_i \overset{\text{iid}}{\sim} \text{Poisson}(\mu)$$

- With covariates: **conditional distribution** is Poisson

$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Poisson}(\mu_i)$$

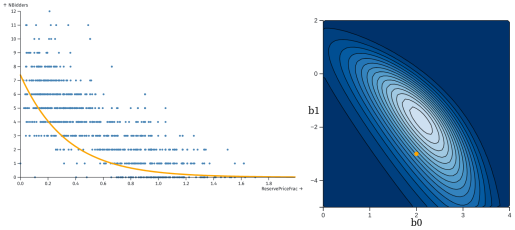
- Modeling the conditional mean as $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ is no good.
- Use **exponential function** to ensure that $\mu_i > 0$

$$\mu_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}$$

- **Poisson regression**

$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Poisson}(\mu_i)$$
$$\mu_i = \mathbb{E}(Y_i | \boldsymbol{x}_i) = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}$$

ML for Poisson regression - widget



Exponential regression

- Continuous positive data with no features.
- **Exponential distribution**

$$Y_i|\beta \overset{\text{iid}}{\sim} \text{Expon}(\beta)$$

- Continuous positive data with features $\boldsymbol{x}$.
- **Exponential regression**

$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Expon}\left(e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right)$$

$$\mu_i = \mathbb{E}(Y_i|\boldsymbol{x}_i) = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}$$

Generalized linear models (GLM)

- Continuous positive data. **Gamma regression**

$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Gamma}(\alpha, e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})$$
$$\mu_i = \mathbb{E}(Y_i|\boldsymbol{x}_i) = \alpha e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}$$

- Data as proportions: **Beta regression**.
- **Truncated data**: truncated normal regression: widget
- ... and so on ...
- **Generalized Linear Models**.
- Maximum likelihood by numerical maximization.
- **Sampling distribution** from **Observed information**

$$\hat{\boldsymbol{\beta}} \overset{\text{approx}}{\sim} N\left(\boldsymbol{\beta}, \mathcal{J}_n^{-1}(\hat{\boldsymbol{\beta}})\right) \text{ for large } n$$

GLMs are linear models

- **Logistic regression**

$$Y_i \mid \boldsymbol{x}_i \stackrel{\text{ind}}{\sim} \text{Bernoulli}(\mu_i)$$
$$\mu_i = \frac{1}{1 + e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$

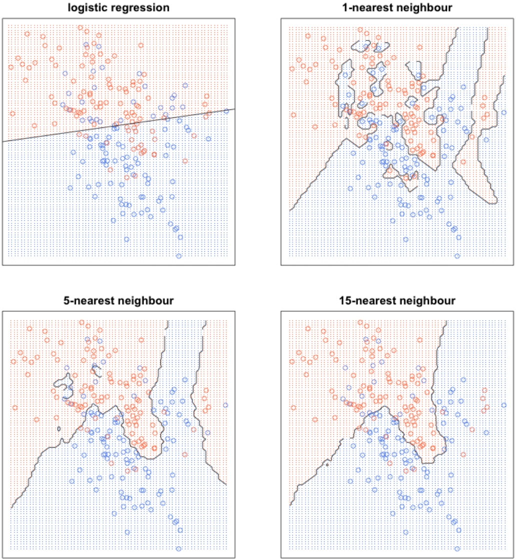
- **Decision boundary**: $\Pr(y_i = 1|\boldsymbol{x}_i) = \Pr(y_i = 0|\boldsymbol{x}_i)$

$$\mu_i = 1 - \mu_i \quad \iff \quad \frac{1}{1 + e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}} = \frac{e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$
$$1 = e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}} \quad \iff \quad 0 = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

(take log on both sides, and recall $\log 1 = 0$ and $\log e^a = a$).

- Decision **boundaries are linear in the features, x**.
- **Linear GLMs** are:
  - **highly interpretable**.
  - **robust to overfitting**.
  - **restrictive**.

Logistic regression - linear decision boundaries



Non-linear regression

- Example: **Poisson non-linear regression**

$$
\begin{aligned}
Y_i \mid \boldsymbol{x}_i &\overset{\text{ind}}{\sim} \text{Poisson}(\mu_i) \\
\mu_i &= e^{f(\boldsymbol{x}_i)}
\end{aligned}
$$

  where $f(\boldsymbol{x}_i)$ is some (non-linear) function of the covariates.

- Examples:
  - Linear: $f(x) = \beta_0 + \beta_1 x$
  - Polynomial: $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots$

- Other **non-linear models**:
  - **Splines**
  - **Regression trees**
  - **Neural networks**
  - **Gaussian processes**