MAT 3375
Regression Analysis

Chapter 7
Generalized Linear Regression

P. Boily (uOttawa)

Summer – 2023

Outline

7.1 – General Framework (p.3)

7.2 – Logistic Regression (p.9)

- Maximum Likelihood Estimation (p.11)
- Significance of Predictors (p.18)
- Deviance Goodness-of-Fit Test (p.21)

7.3 – Poisson Regression (p.25)

7 – Generalized Linear Regression

In real-world situations, the linearity assumption (i.e., the assumption that the fitted response takes the form $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$) is not always met (this is definitely the case if the response must be positive, say, as in a height, or non-negative, as in a salary, or categorical, as in voting intentions).

There are various non-linear modeling approaches to handle such situations, including neural networks, support vector machines, decision trees, etc.

But the machinery of linear regression is quite useful and easy to use, and it seems a waste to let it go by the wayside.

Generalized linear models provide additional modeling flexibility without necessarily sacrificing the convenience of linear regression.

7.1 – General Framework

Consider a bivariate dataset $\{(X_i, Y_i) \mid i = 1, \ldots, n\}$ for which the response Y is **categorical** has $m = 2$ levels ($Y \in \{0, 1\}$, say). Write

$$P(Y_i = 1) = \beta_0 + \beta_1 X_i = \pi_i, \ \mathrm{E}\{Y_i\} = 1 \cdot \pi_i + 0 \cdot (1 - \pi_1) = \pi_i$$

$$\sigma^2\{Y_i\} = \mathrm{E}\left\{Y_i^2\right\} - \left(\mathrm{E}\{Y_i\}\right)^2 = 1^2 \cdot \pi_i + 0^2 \cdot (1 - \pi_i) - \pi_i^2 = \pi_i(1 - \pi_i)$$

If we use a SLR model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with **independent** errors:

1. the range of ε contains only $m = 2$ values for a fixed predictor level X,

$$\varepsilon = \begin{cases} 1 - (\beta_0 + \beta_1 X), & \text{if } Y = 1 \\ 0 - (\beta_0 + \beta_1 X), & \text{if } Y = 0 \end{cases}$$

2. the variance of the error terms is

$$\sigma^2\{\varepsilon_i\} = \sigma^2\{Y_i\} = \pi_i(1-\pi_i) = (\beta_0+\beta_1 X_i)(1-\beta_0-\beta_1 X_i)$$

3. $E\{Y_i\} = \pi_i$ is such that $\pi_i \in [0,1]$.

The first consequence violates the assumption that error terms are **normal**; the second that the variance is **constant**, and the third that the response is **linear** in the predictor.

The first two issues can be mitigated by the CLT and weighted regression if n is large enough, but the third one cannot: SLR is **not** an appropriate model in this case.

Generalized linear models (GLM) extend the ordinary least square (OLS) paradigm by accommodating response variables with **non-normal** conditional distributions.

Apart from the **error structure**, a GLM is essentially a linear model

$$Y_i \sim \mathcal{D}(\mu_i), \quad \text{where } g(\mu_i) = \eta_i = \mathbf{X}_i\boldsymbol{\beta}.$$

A GLM consists of:

- a **systematic component** $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$ (linear in $\boldsymbol{\beta}$);
- a **random component**, specified by the distribution $\mathcal{D}$, and
- a **link function** g.

The general ideas and concepts of OLS carry over to GLM:

- the **systematic component** is specified in terms of the linear predictor for the i^{th} observation $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$;
- the **link function** g associating the **systematic component** η_i to the **distribution** of the response Y_i must be **smooth** (or at least differentiable) and **monotonic** (and so **invertible**);
- the **distribution** $\mathcal{D}$ for the response Y_i is usually selected from the **exponential family** of distributions, with p.d.f. satisfying

  $$f(\mathbf{x} \mid \boldsymbol{\theta}) = h(\mathbf{x})k(\boldsymbol{\theta})\exp(\boldsymbol{\phi}(\boldsymbol{\theta}) \cdot \boldsymbol{T}(\mathbf{x})), \quad \text{for appropriate } h,\ k,\ \phi,\ \mathbf{T};$$

  this includes the normal, binomial, Poisson, Gamma distributions, etc.; these are all distributions with **conjugate priors**.

OLS is an example of GLM, with:

- **systematic component** $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$;
- **random component** $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$;
- **link function** $g(\mu) = \mu$,

which we had previously written as

$$\mathbf{Y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

The motivating example at the start of this section is also a GLM.

The main **advantages** of GLM are that:

- there is no need to transform $\mathbf{Y} \mid \mathbf{X}$ if it is not normal;
- if the link produces **additive effects**, homoscedasticity is not required;
- the choice of the link is separate from the choice of random component;
- models are still fitted *via* a **maximum likelihood** procedure;
- **inference tools** and **model checks** still apply;
- they are easily implemented in `R` *via* `glm()`, and
- various regression modeling approaches (OLS, logistic, etc.) are GLM.

7.2 – Logistic Regression

Consider n Bernoulli responses $Y_i \in \{0,1\}$; the GLM **random component** is

$$Y_i \sim \mathcal{B}(\pi_i), \quad \text{(where } \pi_i \text{ plays the role of } \mu_i\text{)},$$

with $\mathrm{E}\{Y_i\} = \pi_i = P(Y_i = 1 \mid \mathbf{X}_i)$ and $\sigma^2\{Y_i\} = \pi_i(1-\pi_i)$.

We define the **link logit response function** as

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right).$$

With the regular **systematic component** $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$, the **logistic regression** model

$$g(\pi_i) = \eta_i, \quad Y_i \sim \mathcal{B}(\pi_i), \quad \text{with independent errors}$$

can be re-written as

$$\begin{aligned} g(\pi_i) = \eta_i &\Longleftrightarrow \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{X}_i\boldsymbol{\beta} \\ &\Longleftrightarrow \frac{\pi_i}{1-\pi_i} = \exp(\mathbf{X}_i\boldsymbol{\beta}) \quad \text{(odds ratio)} \\ &\Longleftrightarrow \pi_i = (1-\pi_i)\exp(\mathbf{X}_i\boldsymbol{\beta}) \\ &\Longleftrightarrow \pi_i + \pi_i\exp(\mathbf{X}_i\boldsymbol{\beta}) = \exp(\mathbf{X}_i\boldsymbol{\beta}) \\ &\Longleftrightarrow \pi_i = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}, \quad Y_i \sim \mathcal{B}(\pi_i). \end{aligned}$$

7.2.1 – Maximum Likelihood Estimation

The question then becomes, how do we find the parameter vector $\boldsymbol{\beta}$? The p.d.f. of a Bernoulli r.v. $Y_i \sim \mathcal{B}(\pi_i)$ is

$$f(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}, \quad Y_i \in \{0,1\};$$

if we assume independence of the error terms, the joint **maximum likelihood function** is

$$L(\boldsymbol{\beta}) = f(\mathbf{Y}) = \prod_{i=1}^{n} f(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}, \quad Y_i \in \{0,1\}.$$

The **log-likelihood function** is thus

$$\begin{aligned}\ln L(\boldsymbol{\beta}) &= \sum_{i=1}^n \ln f(Y_i) = \sum_{i=1}^n \left[Y_i \ln \pi_i + (1-Y_i)\ln(1-\pi_i)\right] \\ &= \sum_{i=1}^n \left[Y_i \ln \pi_i - Y_i \ln(1-\pi_i) + \ln(1-\pi_i)\right] \\ &= \sum_{i=1}^n Y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^n \ln(1-\pi_i) \\ &= \sum_{i=1}^n Y_i \mathbf{X}_i\boldsymbol{\beta} + \sum_{i=1}^n \ln\left(1-\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right),\end{aligned}$$

or

$$\begin{aligned}\ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} Y_i\mathbf{X}_i\boldsymbol{\beta} + \sum_{i=1}^{n} \ln\left(\frac{1}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right) \\ &= \sum_{i=1}^{n} Y_i\mathbf{X}_i\boldsymbol{\beta} - \sum_{i=1}^{n} \ln\left(1+\exp(\mathbf{X}_i\boldsymbol{\beta})\right)\end{aligned}$$

The **maximum likelihood estimator vector** $\mathbf{b}$ solves

$$0 = \left.\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\mathbf{b}} = \nabla_{\boldsymbol{\beta}} \ln L(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\mathbf{b}}$$

which, unlike in the OLS case, is found using **numerical methods**.

The corresponding **fitted values** are

$$\hat{\pi}_i = \frac{\exp(\mathbf{X}_i \boldsymbol{b})}{1 + \exp(\mathbf{X}_i \boldsymbol{b})},$$

the **fitted logistic response function** at a generic predictor $\mathbf{x}$ is

$$\hat{\pi}(\mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{b})}{1 + \exp(\mathbf{x}\boldsymbol{b})},$$

and the **fitted logit response function** at $\mathbf{x}$ is

$$\hat{\pi}'(\mathbf{x}) = \mathbf{x}\boldsymbol{b} = \ln\left(\frac{\hat{\pi}(\mathbf{x})}{1 - \pi(\mathbf{x})}\right).$$

Example: we use a logistic regression model to fit the disease status ($Y = 0, 1$) of an individual in a population based on their age (X_1), employment status ($X_2 = 0, 1$), and whether they live in a city or a rural area ($X_3 = 0, 1$). The estimated coefficients are

$$\boldsymbol{b} = (b_0, b_1, b_2, b_3) = (-4.9988, 0.13086, -0.41636, 2.6858).$$

What is the probability that individual i has the disease ($Y = 1$) if their predictor vector is $\mathbf{X}_i = (44, 1, 1)$?

Solution: the fitted logistic response function is

$$\hat{\pi}(\mathbf{x}) = \frac{\exp(-4.9988 + 0.13086x_1 - 0.41636x_2 + 2.6858x_3)}{1 + \exp(-4.9988 + 0.13086x_1 - 0.41636x_2 + 2.6858x_3)},$$

so $P(Y = 1 \mid \mathbf{X} = (44, 1, 1)) = \hat{\pi}(44, 1, 1) = 0.9538443$.

Interprétation du coefficient b_k

Let $\mathbf{X}_1^k, \mathbf{X}_2^k = \mathbf{X}_1^k + \mathbf{e}_k$ be predictor vectors in the model's range, which is to say that $\mathbf{X}_2^k$ differs from $\mathbf{X}_1^k$ only by 1 unit in the kth position, i.e., $\Delta X_k = 1$.

On the one hand,

$$\Delta\hat{\pi}' = \hat{\pi}'(\mathbf{X}_2^k) - \hat{\pi}'(\mathbf{X}_1^k) = \hat{\pi}'(\mathbf{X}_1^k) + b_k - \hat{\pi}'(\mathbf{X}_1^k) = b_k;$$

On the other hand,

$$\begin{aligned}\Delta\hat{\pi}' = \hat{\pi}'(\mathbf{X}_2^k) - \hat{\pi}'(\mathbf{X}_1^k) &= \ln\left(\frac{\hat{\pi}(\mathbf{X}_2^k)}{1-\hat{\pi}(\mathbf{X}_2^k)}\right) - \ln\left(\frac{\hat{\pi}(\mathbf{X}_1^k)}{1-\hat{\pi}(\mathbf{X}_\mathbf{1}^\mathbf{k})}\right) \\ &= \ln\left(\frac{\hat{\pi}(\mathbf{X}_2^k)}{1-\hat{\pi}(\mathbf{X}_2^k)}\Big/\frac{\hat{\pi}(\mathbf{X}_1^k)}{1-\hat{\pi}(\mathbf{X}_1^k)}\right).\end{aligned}$$

Thus,

$$e^{b_k} = \frac{\text{odds ratio}(\mathbf{X_2^k})}{\text{odds ratio}(\mathbf{X_1^k})} = \frac{\hat{\pi}(\mathbf{X}_2^k)}{1-\hat{\pi}(\mathbf{X}_2^k)} \Big/ \frac{\hat{\pi}(\mathbf{X}_1^k)}{1-\hat{\pi}(\mathbf{X}_1^k)};$$

the % increase in the odds ratios before and after the change is then $e^{b_k} - 1$.

Example: suppose that $e^{b_k} = 1.26$; then if X_k increases by 1, there is a 26% increase in the odds ratio.

If $e^{b_k} = 0.25$; then if X_k increases by 1, there is a 75% decrease in the odds ratio.

7.2.2 – Significance of Predictors

As with OLS, we can provide inferential tests for the model parameters: when n is large, we use the test statistic

$$z = \frac{b_k - \beta_k}{s\{b_k\}}.$$

But what is $s\{b_k\}$ in the context of **logistic regression**?

We use **Wald's Test** to test

$$\begin{cases} H_0 : \beta_k = 0 \\ H_1 : \beta_k \neq 0 \end{cases}$$

Under H_0, the test statistic is $z^* = \frac{b_k}{s\{b_k\}}$.

The standard deviation is provided by the **approximate estimated variance-covariance matrix**

$$\mathrm{s}^2\{\mathbf{b}\} = -G^{-1} = -\left[\frac{\partial^2}{\partial\beta_i\partial\beta_j}\ln L(\beta)\right]^{-1},$$

which once again must be computed numerically. When n is large and H_0 holds, we expect $z^* \sim \mathcal{N}(0,1)$, approximately.

Decision Rule: If $|z^*| > z(1-\frac{\alpha}{2})$, we **reject** H_0 at significance level α; $b_k \pm z(1-\frac{\alpha}{2})s\{b_k\}$ gives an approximate $100(1-\alpha)\%$ C.I. for β_k.

Simultaneous inferences are also possible (see Module 3).

Example: let us assume that the approximate variance-covariance matrix of the disease status example is

$$\mathrm{s}^2\{\mathbf{b}\} = \begin{pmatrix} 0.5130 & -0.1168 & -0.3121 & -0.2743 \\ -0.1168 & 0.00029 & 0.00084 & 0.00045 \\ -0.3121 & 0.00084 & 0.4761 & 0.1734 \\ -0.2743 & 0.00045 & 0.1734 & 0.3627 \end{pmatrix}.$$

Find an approximate 95% C.I. for β_3 and for the odds ratio $\exp(\beta_3)$.

Solution: we have $\alpha = 0.05$, $b_3 = 2.6858$, $s\{b_3\} = \sqrt{0.3627} = 0.6022$, and $z(1-0.05/2) = z(0.975) = 1.96$. Thus

$$\text{CI}(\beta_3; 0.95) : 2.6858 \pm 1.96(0.6022) \equiv (1.505, 3.866) \quad \text{and}$$

$$\text{CI}(\exp(\beta_3); 0.95) : \exp(1.505, 3.866) \equiv (4.506, 47.756).$$

7.2.3 – Deviance Goodness-of-Fit Test

There is **no readily available** descriptive statistic that acts for logistic regression as R^2 or R^2_a did for OLS (**candidates exist, such as Nagelkerke's R^2, but they all have important flaws**).

Instead, we use the **deviance goodness-of-fit test**:

$$\begin{cases} H_0 : \pi = \dfrac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)} \\ H_1 : H_0 \text{ is false} \end{cases}$$

We can use this test if we have replicates for **at least one** of the predictor levels (other such tests: Pearson χ^2 and Hosmer-Lemeshow).

Let there be c different predictor levels $(\mathbf{X}_1,\ldots,\mathbf{X}_c)$ and let $Y_{i,j}$ be the ith response for group j (i.e., at $\mathbf{X}=\mathbf{X}_j$), $i=1,\ldots,n_j$, $j=1,\ldots,c$.

The **full** and **reduced** models are $\mathrm{E}\{Y_{i,j}\}=\pi_j$, $\mathrm{E}\{Y_{i,j}\}=\hat{\pi}(\mathbf{X}_j)=\hat{\pi}_j$, resp. Let p_j be the **proportion of observations with $Y_{i,j}=1$ in group j**.

The statistic for the **deviance of the fitted model** is

$$\mathrm{DEV}(p)=-2\sum_{j=1}^{c}\left[n_jp_j\ln\frac{\hat{\pi}_j}{p_j}+(n_j-n_jp_j)\ln\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right],$$

which follows a $\chi^2(c-p)$ distribution when H_0 holds.

Decision Rule: if $\mathrm{DEV}(p)>\chi^2(1-\alpha;c-p)$, **reject** H_0 at significance level α.

Example: we fit a logistic model with $p = 3$ to attempt to explain whether students will pass a difficult exam on the first try; $c = 6$ classes of students are identified. The fit statistics are shown below:

class j	n_j	$\hat{\pi}_j$	p_j	**class** j	n_j	$\hat{\pi}_j$	p_j
1	190	0.1865	0.154	4	202	0.4920	0.515
2	186	0.2645	0.296	5	175	0.7718	0.742
3	200	0.3562	0.350	6	112	0.8557	0.902

Does the corresponding logistic model provide a satisfactory fit at $\alpha = 0.05$?

Solution: we have $p = 3$ and

$$\mathrm{DEV}(3) = -2\left[n_1p_1\ln\left(\frac{\hat{\pi}_1}{p_1}\right) + (n_1 - n_1p_1)\ln\left(\frac{1-\hat{\pi}_1}{1-p_1}\right) + \cdots \right.$$
$$\left. \cdots + n_6p_6\ln\left(\frac{\hat{\pi}_6}{p_6}\right) + (n_6 - n_6p_6)\ln\left(\frac{1-\hat{\pi}_6}{1-p_6}\right)\right]$$

$$= -2\left[190(0.154)\ln\left(\frac{0.1865}{0.154}\right) + (190 - 190(0.154))\ln\left(\frac{1-0.1865}{1-0.154}\right) + \cdots\right.$$

$$\left.\cdots + 112(0.902)\ln\left(\frac{0.8557}{0.902}\right) + (112 - 112(0.902))\ln\left(\frac{1-0.8557}{1-0.902}\right)\right] = 5.786.$$

At $\alpha = 0.05$ and $c - p = 3$, the decision threshold is $\chi^2(0.95;3) = 7.81$. Since $\mathrm{DEV}(3) = 5.786 \leq 7.81$, we conclude that the logistic model provides **a satisfactory fit**.

In practice, if $p_j = 0, 1$ for some class j, then we use the convention

$$n_j p_j \ln\left(\frac{\hat{\pi}_j}{p_j}\right) = 0 \quad \text{or} \quad (n_j - n_j p_j)\ln\left(\frac{1-\hat{\pi}_j}{1-p_j}\right) = 0,$$

respectively.

7.3 – Poisson Regression

In the early stages of a rumour spreading, the rate at which new individual learn the information increases exponentially over time. If μ_i is the **expected number of people who have heard the rumor on day** x_i, a model of the form $\mu_i = \gamma \exp(\delta x_i)$ might be appropriate:

$$\underbrace{\ln(\mu_i)}_{\text{link}} = \ln\gamma + \delta x_i = \beta_0 + \beta_1 x_i = \underbrace{(1, x_i)^\top (\beta_0, \beta_1)}_{\text{systematic component}} = \mathbf{x}_i\boldsymbol{\beta}.$$

Furthermore, since we measure a count of individuals, the **Poisson** distribution could be a reasonable choice, leading to the following GLM:

$$Y_i \sim \underbrace{\text{Poisson}(\mu_i),}_{\text{random component}} \quad \ln(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}.$$

In general, if $Y_i \sim \text{Poisson}(\mu_i)$ are independent and if the link function is $g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$, the **maximum likelihood function** is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(Y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{Y_i}}{Y_i!};$$

the **log-likelihood function** is

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \ln \mu_i - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \ln(Y_i!),$$

and the $\mathbf{b}_{\text{MLE}}$ solves $\nabla_{\boldsymbol{\beta}} \ln L(\boldsymbol{\beta}) = 0$, which needs to be found numerically.

The corresponding **fitted values** are $\hat{\mu}_i = g^{-1}\left(\mathbf{x}_i\boldsymbol{b}\right)$.

Likelihood Ratio Test: let $q < p$. We test for the significance of some of the (possibly re-ordered) predictors according to

$$\begin{cases} H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \\ H_1 : \text{one of } \beta_q, \beta_{q+1}, \ldots, \beta_{p-1} \neq 0 \end{cases}$$

by computing the test statistic

$$G^2 = -2\ln\left(\frac{L(R)}{L(F)}\right) = -2\ln\left(\frac{L(\mathbf{b}_R)}{L(\mathbf{b}_F)}\right),$$

where $\mathbf{b}_F, \mathbf{b}_R$ are the MLEs of the **full** and **reduced** models, respectively, which follows a $\chi^2(p-q)$ distribution under H_0.

Decision Rule: if $G^2 > \chi^2(1-\alpha; p-q)$, we **reject** H_0 at significance level α.

Deviance Goodness-of-Fit Test: in order to test for

$$\begin{cases} H_0 : g(\mu) = \mathbf{X}\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \\ H_1 : H_0 \text{ is false} \end{cases}$$

we compute the **model deviance** test statistic

$$G^2 = \mathrm{DEV}(p) = -2\left(\sum Y_i \ln \frac{\hat{\mu}_i}{Y_i} + \sum (Y_i - \hat{\mu}_i)\right),$$

which follows a $\chi^2(n-p)$ distribution when H_0 holds.

Decision Rule: if $\mathrm{DEV}(p) > \chi^2(1-\alpha; n-p)$, we **reject** H_0 at significance level of α.