

# Data Analytics for Pharmaceutical Sciences

## Part IV: Generalized Linear Models

---

Stéphane Guerrier, Data Analytics Lab, University of Geneva 🇨🇭

Dominique-L. Couturier, Cancer Research UK, University of Cambridge 🇬🇧

Yuming Zhang, Data Analytics Lab, University of Geneva 🇨🇭



License: CC BY NC SA 4.0

# Motivation example: Bronchitis

---

Problem

Data

Graph

A scientist is interested in studying the effects of smoking on being diagnosed with bronchitis. He collected 212 people for an experiment where he collected their daily number of smoked cigarettes. Then the scientist recorded the **presence (encoded as 1)** or **absence (encoded as 0)** of bronchitis of each person as the outcome of the experiment.

Based on these samples, the scientist wants to verify that smoking every day significantly contributes to the presence of bronchitis.

# Non-normal conditional distributions

In practice, we often encounter data which is actually not suitable for a linear model. For example,

1. A researcher is interested in studying the **proportion** of individuals with a specific genetics mutation in different groups.
2. A hospital wants to investigate the **number** of children diagnosed with an infectious disease per day.
3. A lab wants to analyze the **probability** of myocardial infarction per treatment group.

In these cases, linear models are not suitable for two reasons:

1. **Non-linearity**: a linear model may give predicted values outside of the possible range (e.g. negative values for probability/proportion, decimal values for number of people).
2. **Heteroscedasticity**: the (conditional) variance of response is not constant and depends on its (conditional) mean.

# Solution: Generalized linear models

- Recall that in linear regression, we consider

$$Y_i \stackrel{iid}{\sim} \mathcal{N} \left( \beta_0 + \sum_{j=1}^p \beta_j X_{ij}, \sigma^2 \right).$$

- In **Generalized Linear Models (GLM)**, we consider

$$Y_i \sim \text{distribution} \left( f \left( \beta_0 + \sum_{j=1}^p \beta_j X_{ij} \right), \phi \right),$$

where **distribution** (classically) belongs to the **exponential family**, such as Bernoulli distribution or Poisson distribution. The function  $f(\cdot)$  🙌 is monotonically increasing, and  $\phi$  is a nuisance parameter related to the variance of  $Y_i$  (which is typically also a function of  $\mu_i$ ).

🙌 The function  $f(\cdot)$  is actually the inverse of a function called the link function.

# Logistic regression

- **Logistic regression** is a classical example of GLM, and is most commonly used to model a binary response, where  $Y_i = 0$  or  $1$ .
- It assumes that the response  $Y_i$  follows a Bernoulli distribution with parameter  $\mu_i$  such that  $\Pr(Y_i = 1) = \mu_i$  and  $\Pr(Y_i = 0) = 1 - \mu_i$ .
- The parameter  $\mu_i$ , i.e. the conditional mean of  $Y_i$ , is then modelled as a function of the covariates  $\mu_i = f\left(\beta_0 + \sum_{j=1}^p \beta_j X_{ij}\right)$ . In other words,

$$Y_i \sim \text{Bernoulli} \left( f \left( \beta_0 + \sum_{j=1}^p \beta_j X_{ij} \right), \phi \right), \quad \text{where } \phi = 1.$$

- The function  $f(\cdot)$  allows to transform the whole real line to  $(0, 1)$ , which makes unrestricted linear modeling possible. Some common choices of  $f(\cdot)$  include:
  1. **Logit link**:  $f(z) = e^z / (1 + e^z) = 1 / (1 + e^{-z})$ .
  2. **Probit link**:  $f(z) = \Phi(z)$ , where  $\Phi(\cdot)$  is the CDF of a  $\mathcal{N}(0, 1)$ .

# Logistic regression

Compared to linear regression, logistic regression makes **more relaxed assumptions**:

1. The conditional mean  $\mu_i$  is assumed to be a general (not necessarily linear) function of the covariates.
2. The errors are assumed to be independent. However, they do not need to be normally distributed and actually have different variance that depends on  $\mu_i$ .


The parameters of the model (i.e.  $\beta_0, \beta_1, \dots, \beta_p$ ) are estimated by **maximum likelihood estimation**. In practice, the Maximum Likelihood Estimator (MLE) has no closed-form solution and needs to be computed using **iterative methods** such as the Newton-Raphson algorithm and the Iteratively ReWeighted Least Squares (IRWLS) algorithm. The details of these computational methods are beyond the scope of this class.


# Example: Bronchitis

In the bronchitis example, using the logit link function, we can formulate a logistic regression model as follows:

$$\Pr(\text{Bronchitis}_i = 1 | \text{Cigarettes}_i) = \frac{\exp(\beta_0 + \beta_1 \text{Cigarettes}_i)}{1 + \exp(\beta_0 + \beta_1 \text{Cigarettes}_i)}.$$

- **Bronchitis<sub>i</sub>**: indicator of the presence of bronchitis for the  $i$ -th person (i.e. **Bronchitis<sub>i</sub>** = 1 if presence and **Bronchitis<sub>i</sub>** = 0 if absence).
- **Cigarettes<sub>i</sub>**: the daily number of smoked cigarettes of the  $i$ -th person.

We also name  $\exp(\beta_0 + \beta_1 \text{Cigarettes}_i)$  as the **odds** , the linear combination of covariates  $\beta_0 + \beta_1 \text{Cigarettes}_i$  as the **log odds**, and the coefficients to be **log-odds ratios**.

 The odd related to a probability  $\pi$  equals  $\pi/(1 - \pi)$  (the probability of success over the one of failure).

# Example: Bronchitis

---

R Code

Output

In R, we use the function `glm( ... , family=binomial(link="logit"))`.

Here is the code for our example:

```
# Import data (if you haven't already)
dat = read.csv("data/bronchitis.csv")

# Fit a logistic regression
mod1 = glm(bron ~ cigs, data = dat, family=binomial(link="logit"))
summary(mod1)
```



# Interpretation of estimated coefficients

- $\hat{\beta}_0 = -2.2840$  represents the **log odds** that a person is diagnosed with bronchitis when no cigarettes are consumed. Alternatively, the probability that a non-smoking person diagnosed with bronchitis is:

$$\frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)} \approx 9.25\%.$$

- $\hat{\beta}_1 = 0.2094$  means that the **log odds** of having bronchitis is estimated to **increase by** 0.2094 by consuming one more cigarette. Therefore, the probability that a person **consuming 20 cigarettes a day** to be diagnosed with bronchitis is

$$\frac{\exp(\hat{\beta}_0 + 20\hat{\beta}_1)}{1 + \exp(\hat{\beta}_0 + 20\hat{\beta}_1)} \approx 87.03\%.$$

# Interpretation of coefficients

In general, the logistic model stipulates that the effect of a covariate on the chance of "success" is **linear on the log odds scale, or multiplicative on the odds scale**.

- If  $\beta_j > 0$  then  $\exp(\beta_j) > 1$ , and therefore the odds (probability) increases.
- If  $\beta_j < 0$  then  $\exp(\beta_j) < 1$ , and therefore the odds (probability) decreases.
- If  $\beta_j = 0$  then  $\exp(\beta_j) = 1$ , which implies that the log odds is a constant function, and the response  $Y_i$  is independent of the covariate  $X_{ij}$ .
- The **magnitude** of  $\beta_j$  (i.e.  $|\beta_j|$ ) determines how fast the log odds increases or decreases as a function of the covariate  $X_{ij}$ .

# Interpretation of coefficient p-values

- Like for linear models, for each coefficient  $\beta_j$ , there is a corresponding p-value associated to the Wald test of  $H_0 : \beta_j = 0$  and  $H_a : \beta_j \neq 0$ .
- A covariate with a small p-value (typically smaller than 5%) is considered to be a significant (meaningful) addition to the model, as changes in the values of such covariate can lead to changes in the response variable.
- On the other hand, a large p-value (typically larger than 5%) suggests that the corresponding covariate is not (significantly) associated with changes in the response or that we don't have enough evidence (data) to show its effect.
- ⚠ The Wald test for each coefficient  $\beta_j$  of a GLM is analogous to a t-test for a coefficient in a linear regression. But there is a slight difference. Unlike in a linear regression, there is generally no exact distribution theory for a GLM and we rely instead on large-sample approximations based on the asymptotic normality of the MLE. 🧐

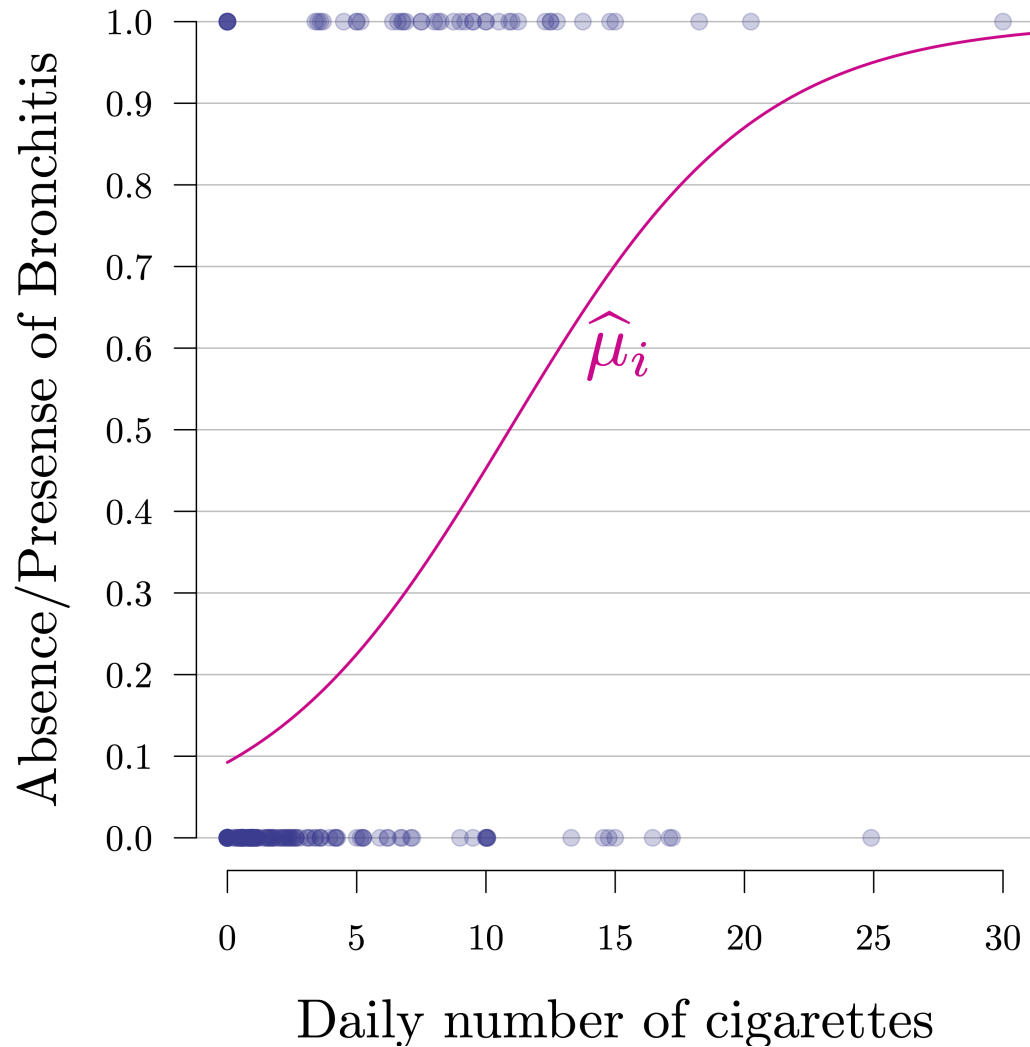
# Interpretation of coefficient p-values

- In this example, the coefficient p-value associated to the `Cigarettes` covariate is  $2.59 \times 10^{-6}\%$ . This suggests that whether people consume cigarettes every day makes a significant **difference** in the presence of bronchitis, compared to those who do not smoke, at the significance level of 5%.
- To further verify that smoking contributes to the presence of bronchitis, we can compute the p-value of a one-sided test using the following relations (like in linear regression):

	$H_a : \beta_j > 0$	$H_a : \beta_j < 0$
$\hat{\beta}_j > 0$	$p/2$	$1 - p/2$
$\hat{\beta}_j < 0$	$1 - p/2$	$p/2$

In this case,  $\hat{\beta}_1 = 0.2094$  and  $p = 2.59 \times 10^{-6}\%$ . So the p-value for the test  $H_0 : \beta_1 = 0$  and  $H_a = \beta_1 > 0$  is  $2.59 \times 10^{-6}\%/2 \approx 1.3 \times 10^{-6}\% < \alpha$ . So we can conclude that smoking **significantly contributes** to the presence of bronchitis.

# Model fit



# Example: Bronchitis

The scientist points out that the **environmental pollution** also plays an important role in the development of bronchitis. So he wants to consider the level of environmental pollution a person is exposed to as an additional covariate, which can be accessed as follows:

```
head(dat$poll)
```

```
#> [1] 67.1 66.9 66.7 65.8 64.4 64.4
```

The scientist wants to know whether **both smoking and environmental pollution** contribute to the development of bronchitis.

# Let's update our model

Taking into account the new covariate `pollution`, we can update our model as follows:

$$\Pr(\text{Bronchitis}_i = 1 | \text{Cigarettes}_i, \text{Pollution}_i) = \frac{\exp(\beta_0 + \beta_1 \text{Cigarettes}_i + \beta_2 \text{Pollution}_i)}{1 + \exp(\beta_0 + \beta_1 \text{Cigarettes}_i + \beta_2 \text{Pollution}_i)}.$$

- **Bronchitis<sub>i</sub>**: indicator of the presence of bronchitis for the  $i$ -th person (i.e. **Bronchitis<sub>i</sub>** = 1 if presence and **Bronchitis<sub>i</sub>** = 0 if absence).
- **Cigarettes<sub>i</sub>**: the daily number of smoked cigarettes of the  $i$ -th person.
- **Pollution<sub>i</sub>**: the level of pollution that the  $i$ -th person is exposed to.

The goal of the scientist is now to assess if both  $\beta_1$  and  $\beta_2$  are significantly larger than 0.

# Example: Bronchitis

R Code

Output

AIC

Here is the code to fit our second model:

```
# Import data (if you haven't already)
dat = read.csv("data/bronchitis.csv")
head(dat)
```

```
#>   bron  cigs poll
#> 1     0 5.15 67.1
#> 2     1 0.00 66.9
#> 3     0 2.50 66.7
#> 4     0 1.75 65.8
#> 5     0 6.75 64.4
#> 6     0 0.00 64.4
```

```
# Fit a logistic regression
mod2 = glm(bron ~ cigs + poll, data = dat, family=binomial(link="logit"))
summary(mod2)
```



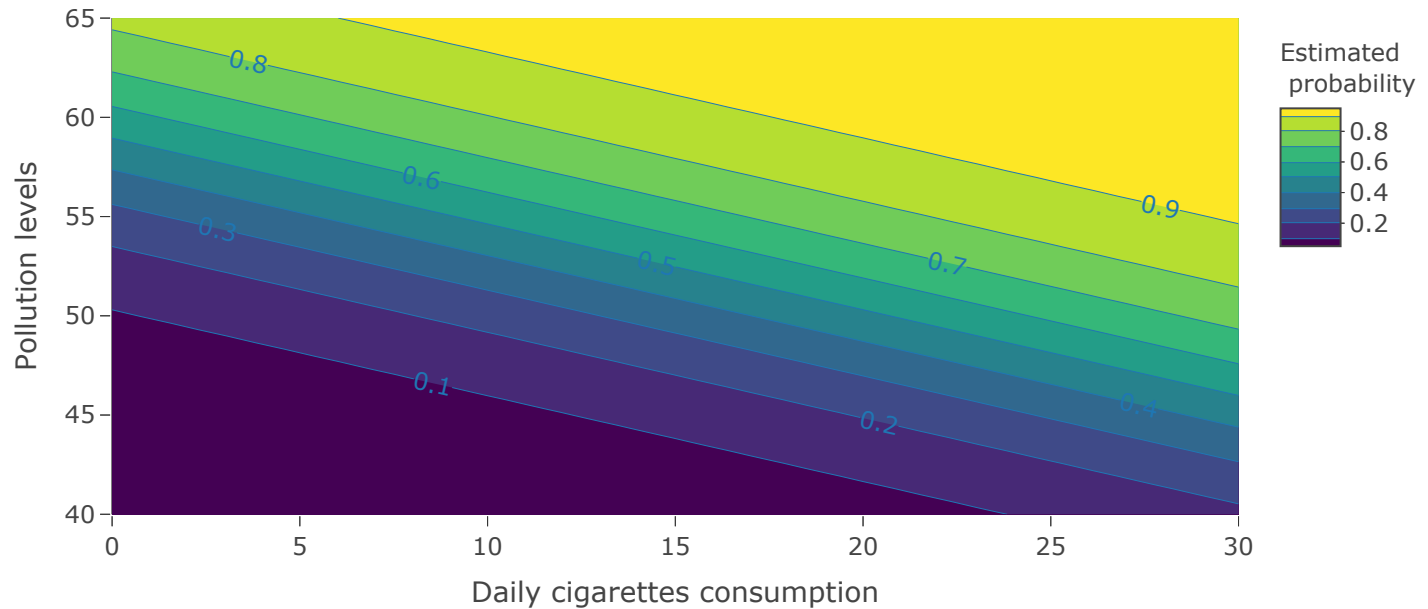
# Interpretation of estimated coefficients

- $\hat{\beta}_0 = -10.0849$  represents the **log odds** that a person is diagnosed with bronchitis when the person consumes no cigarettes and is not exposed to any pollution.
- $\hat{\beta}_1 = 0.2117$  means that the **log odds** of having bronchitis is estimated to **increase by** 0.2117 by consuming one more cigarette.
- $\hat{\beta}_2 = 0.1318$  means that the **log odds** of having bronchitis is estimated to **increase by** 0.1318 by being exposed to one more unit of pollution.
- Therefore, we can compute the following probabilities:

$$\frac{\exp(\hat{\beta}_0 + 20\hat{\beta}_1 + 65\hat{\beta}_2)}{1 + \exp(\hat{\beta}_0 + 20\hat{\beta}_1 + 65\hat{\beta}_2)} \approx 93.80\%,$$

$$\frac{\exp(\hat{\beta}_0 + 10\hat{\beta}_1 + 50\hat{\beta}_2)}{1 + \exp(\hat{\beta}_0 + 10\hat{\beta}_1 + 50\hat{\beta}_2)} \approx 20.13\%.$$

# Predicted probabilities



# Interpretation of coefficient p-values

We now consider whether both  $\beta_1$  and  $\beta_2$  are significantly greater than 0.

- Since  $\hat{\beta}_1 = 0.2117$  and the p-value of the two sided test is  $2.83 \times 10^{-6}\%$ , we obtain the p-value for the one sided test  $H_0 : \beta_1 = 0$  and  $H_a : \beta_1 > 0$  as  $2.83 \times 10^{-6}\%/2 \approx 1.42 \times 10^{-6}\% < \alpha$ , and therefore,  $\beta_1$  is significantly greater than 0 at the significance level of 5%.
- Since  $\hat{\beta}_2 = 0.1318$  and the p-value of the two sided test is 0.7113%, we obtain the p-value for the one sided test  $H_0 : \beta_2 = 0$  and  $H_a : \beta_2 > 0$  as  $0.7113\%/2 \approx 0.3557\% < \alpha$ , and therefore,  $\beta_2$  is also significantly greater than 0 at the significance level of 5%.

Therefore, we conclude that both smoking cigarettes and environmental pollution significantly contribute to the development of bronchitis.

# Poisson regression

- **Poisson regression** is another example of GLM that is used to model count data  $Y_i \in \mathbb{N}$ , such as a number of events occurring during a time interval or in a given area.
- It assumes that, during a time interval or in a given area, **the events occur independently at the same rate given the covariates**.
- The count response  $Y_i$  can be modeled with a Poisson distribution with rate parameter  $\mu_i$  such that  $\mu_i = \exp\left(\beta_0 + \sum_{j=1}^p \beta_j X_{ij}\right)$ . Then

$$Y_i \sim \text{Poisson}\left(\exp\left(\beta_0 + \sum_{j=1}^p \beta_j X_{ij}\right), \phi\right), \quad \text{where } \phi = 1.$$

- In R, we can use the function `glm(..., family=poisson)`.
- The Poisson regression has a restrictive assumption that the mean and variance of  $Y_i$  are the same given the covariates. In practice, however, data are often overdispersed, exhibiting larger variance than the mean. This can be handled, for example, by the **negative binomial regression**.

# Concluding remarks

- GLM extend the normal linear regression model in two ways:
  1. It allows the response variable to **come from an exponential family distribution**. Exponential family includes some of the most commonly used statistical models, including the normal, Bernoulli, and Poisson distributions.
  2. It allows the conditional mean of the response variable to vary linearly with the covariates **through a monotonic transformation**.
- In GLM, the parameters are estimated by the **method of maximum likelihood**. In general, there are no closed-form solutions to the MLE so **iterative methods** are used to compute the estimators.
- Like in linear regression, GLM **only consider independent data**, which sometimes may not be the case in practice. To consider dependent data, we can, for example, consider **generalized linear mixed models**.
- All regression models **should not be used to extrapolate**.

# Exercise: Intensive care admission

---

Problem

Import

We are interested in studying the predictability of the need to be admitted to an Intensive Care Unit (ICU) for COVID 19 patients. For this purpose, data were collected on patients admitted to an emergency care facility in Belgium and were used to assess their relationship to an ICU admission during the course of their illness. The dataset is quite large and we will consider here a small subset with the following variables: `ic` admission to ICU (0 for no, 1 for yes), `sex` gender of the patient (0 for female, 1 for male), `age` age of the patient, `ldh` lactic acid dehydrogenase of the patient and `spo2` oxygen saturation of the patient. Can you conclude that these variables are significant to predict admissions to an ICU?