STAT/Q SCI 403: Introduction to Resampling Methods **Spring 2025**

Lecture 3: MLE and Regression

Instructor: Course notes courtesy of Yen-Chi Chen

3.1 Parameters and Distributions

Some distributions are indexed by their underlying parameters. Thus, as long as we know the parameter, we know the entire distribution. For instance, for Normal distributions $N(\mu, \sigma^2)$, if we know μ and σ^2, the entire distribution is determined. For another example, for Exponential distributions $\mathsf{Exp}(\lambda)$, as long as we know the value of λ, we know the entire distribution. Because these distributions are determined by their parameters, they are sometimes called *parametric distributions*.

Because parameters in the parametric distributions determine the entire distribution, finding these parameters is very important in practice. There are many approaches of finding parameters; here we will introduce the most famous and perhaps most important one–the maximum likelihood estimator (MLE).

3.2 MLE: Maximum Likelihood Estimator

Assume that our random sample $X_1, \cdots, X_n \sim F$, where $F = F_\theta$ is a distribution depending on a parameter θ. For instance, if F is a Normal distribution, then $\theta = (\mu, \sigma^2)$, the mean and the variance; if F is an Exponential distribution, then $\theta = \lambda$, the rate; if F is a Bernoulli distribution, then $\theta = p$, the probability of generating 1.

The idea of MLE is to use the PDF or PMF to find the most likely parameter. For simplicity, here we use the PDF as an illustration. Because the CDF $F = F_\theta$, the PDF (or PMF) $p = p_\theta$ will also be determined by the parameter θ. By the independence property, the joint PDF of the random sample $X_1, \cdots, X_n$

$$p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} p_\theta(x_i).$$

Because $p_\theta(x)$ also changes when θ changes, we rewrite it as $p(x; \theta) = p_\theta(x)$. Thus, the joint PDF can be rewritten as

$$p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i; \theta).$$

Having observed $x_1 = X_1, \cdots, x_n = X_n$, how can we tell which parameter is most likely? Here is a simple proposal that the MLE uses. Based on the joint PDF and $x_1 = X_1, \cdots, x_n = X_n$, we can rewrite the joint PDF as a *function of parameter* θ:

$$L(\theta | X_1, \cdots, X_n) = \prod_{i=1}^{n} p(X_i; \theta).$$

The function L is called the likelihood function. And the MLE finds the maximizer of the likelihood function. Namely,

$$\widehat{\theta}_{MLE} = \widehat{\theta}_n = \underset{\theta}{\mathsf{argmax}}\ L(\theta | X_1, \cdots, X_n).$$

In many cases, maximizing the likelihood function might not be easy so people consider maximizing the *log-likelihood* function:

$$\widehat{\theta}_{MLE} = \widehat{\theta}_n = \underset{\theta}{\mathsf{argmax}}\ \ell(\theta|X_1, \cdots, X_n) = \underset{\theta}{\mathsf{argmax}}\ \sum_{i=1}^n \log p(X_i;\theta) = \underset{\theta}{\mathsf{argmax}}\ \sum_{i=1}^n \ell(\theta|X_i),$$

where $\ell(\theta|X_i) = \log p(X_i;\theta)$ but now we fix each X_i and view it as a function of θ. It is easy to see that maximizing the likelihood function is the same as maximizing the log-likelihood function.

When the log-likelihood function is differentiable with respect to θ, we can use our knowledge from calculus to find $\widehat{\theta}_n$. Let $s(\theta|X_i) = \frac{\partial}{\partial\theta}\ell(\theta|X_i)$ be the derivative of $\ell(\theta|X_i)$ with respect to θ (here for simplicity we assume θ is one-dimensional). The function $s(\theta|X_i)$ is called the *score function*. Then the MLE is from solving the following likelihood equation:

$$s(\widehat{\theta}_n|X_1, \cdots, X_n) = \sum_{i=1}^n s(\widehat{\theta}_n|X_i) = 0. \tag{3.1}$$

Note that we generally need to verify that the solution $\widehat{\theta}_n$ is the maximum by checking the second derivative but here we ignore it for simplicity.

◆ : Equation (**??**) is related to the *generalized estimating equations*[1], a common approach to obtain estimators.

◆ : The idea of obtaining an estimator by maximizing certain criteria (or minimizing some criteria) is called an *M-estimator*[2]. In linear regression, we have learned that the estimators of the slope/intercept is from minimizing the sum of squares of errors (least square estimator). Thus, the least square method is another M-estimator.

Example. (Normal distribution) Here is an example of finding the MLE of the Normal distribution. We assume $X_1, \cdots, X_n \sim N(\mu, 1)$. The goal is to find an estimator of the mean parameter μ. Because the density of such a Normal is $p_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$, the log-likelihood function $\ell(\mu|X_i)$ is

$$\ell(\mu|X_i) = \log p_\mu(X_i) = -\frac{(X_i-\mu)^2}{2} - \frac{1}{2}\log(2\pi),$$

which further implies that the score function is

$$s(\mu|X_i) = \frac{\partial}{\partial\mu}\ell(\mu|X_i) = \mu - X_i.$$

Thus, the MLE $\widehat{\mu}_n$ satisfies

$$0 = \sum_{i=1}^n s(\widehat{\mu}_n|X_i) = \sum_{i=1}^n (\widehat{\mu}_n - X_i) = n\widehat{\mu}_n - \sum_{i=1}^n X_i \Longrightarrow \widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

Thus, the MLE of the mean parameter is just the sample mean.

Example. (Exponential distribution) Assume $X_1, \cdots, X_n \sim \mathsf{Exp}(\lambda)$. Now we find an estimator of λ using the MLE. By definition of the exponential distribution, the density is $p_\lambda(x) = \lambda e^{-\lambda x}$. Thus, the log-likelihood function and the score function are

$$\ell(\lambda|X_i) = \log p_\lambda(X_i) = \log\lambda - \lambda X_i, \quad s(\lambda|X_i) = \frac{1}{\lambda} - X_i.$$

[1] A bit advanced materials can be found in https://en.wikipedia.org/wiki/Generalized_estimating_equation
[2] https://en.wikipedia.org/wiki/M-estimator

As a result, the MLE comes from solving

$$0 = \sum_{i=1}^{n} s(\widehat{\lambda}_n | X_i) = \sum_{i=1}^{n} \left(\frac{1}{\widehat{\lambda}_n} - X_i \right) = \frac{n}{\widehat{\lambda}_n} - \sum_{i=1}^{n} X_i \Longrightarrow \widehat{\lambda}_n = \frac{n}{\sum_{i=1}^{n} X_i}.$$

Namely, the MLE is the inverse of the sample average.

Example. (Uniform distribution) Here is a case where we cannot use the score function to obtain the MLE but still we can directly find the MLE. Assume $X_1, \cdots, X_n \sim \mathsf{Uni}[0, \theta]$. Namely, the random sample is from an uniform distribution over the interval $[0, \theta]$, where the upper limit parameter θ is the parameter of interest. Then the density function is $p_\theta(x) = \frac{1}{\theta} 1(0 \le x \le \theta)$. Here we cannot use the log-likelihood function (think about why) so we use the original likelihood function:

$$L(\theta | X_1, \cdots, X_n) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\theta} 1(0 \le X_i \le \theta) = \frac{1}{\theta^n} 1(0 \le X_{(1)} \le X_{(n)} \le \theta),$$

where $X_{(1)} = \min\{X_1, \cdots, X_n\}$ and $X_{(n)} = \max X_1, \cdots, X_n$ are the minimum and maximum value of the sample (here we use the notations from *order statistic*). Observing from $L(\theta | X_1, \cdots, X_n) = \frac{1}{\theta^n} 1(0 \le X_{(1)} \le X_{(n)} \le \theta)$, we see that a smaller θ, a higher value of the likelihood function. However, there is a restriction– the value of θ cannot be below $X_{(n)}$ otherwise the indicator function outputs 0. Thus, the maximum value of $L(\theta | X_1, \cdots, X_n)$ occurs when $\theta = X_{(n)}$, so the MLE is $\widehat{\theta}_n = X_{(n)}$, the maximum value of the sample.

Example. (Bernoulli distribution) Finally, we provide an example of finding the MLE of a discrete random variable. Suppose $X_1, \cdots, X_n \sim \mathsf{Ber}(p)$. Then the PMF is $X = 1$ with a probability of p and $X = 0$ with a probability of $1 - p$. We can rewrite the PMF in the following succinct form:

$$P(x) = p^X (1 - p)^{1 - X}.$$

You can verify that $P(1) = P(X = 1) = p$ and $P(0) = P(X = 0) = 1 - p$. The likelihood function will be

$$L(p | X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i) = \prod_{i=1}^{n} p^{X_i} (1 - p)^{1 - X_i}$$

We can then compute the log-likelihood function and the score function:

$$\ell(p | X_1, \cdots, X_n) = \sum_{i=1}^{n} \left(X_i \log p + (1 - X_i) \log(1 - p) \right), \quad s(p | X_1, \cdots, X_n) = \sum_{i=1}^{n} \left(\frac{X_i}{p} - \frac{1 - X_i}{1 - p} \right).$$

Therefore, the MLE can be obtained by solving

$$0 = s(\widehat{p}_n | X_1, \cdots, X_n) = \sum_{i=1}^{n} \left(\frac{X_i}{\widehat{p}_n} - \frac{1 - X_i}{1 - \widehat{p}_n} \right).$$

Multiplying both sides by $\widehat{p}_n (1 - \widehat{p}_n)$,

$$0 = \sum_{i=1}^{n} X_i \cdot (1 - \widehat{p}_n) - (1 - X_i) \cdot \widehat{p}_n = \sum_{i=1}^{n} (X_i - \widehat{p}_n) \Longrightarrow \widehat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Again, the MLE is the sample mean.

♦ : In many problems (such as the *mixture models*[3]), we do not have a closed form of the MLE. The only way to compute the MLE is via computational methods such as the *EM algorithm (Expectation-Maximization)*[4],

[3] https://en.wikipedia.org/wiki/Mixture_model

[4] https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

which is like a gradient ascent approach. However, the EM algorithm will stuck at the local maximum, so we have to rerun the algorithm many times to get the real MLE (the MLE is the parameters of 'global' maximum). In machine learning/data science, how to numerically find the MLE (or approximate the MLE) is an important topic. A common solution is to propose other computationally feasible estimators that are similar to the MLE and switch our target to these new estimators.

3.3 Theory of MLE

The MLE has many appealing properties. Here we will focus on one of its most desirable properties: *asymptotic normality* and *asymptotic variance*.

What is asymptotic normality? It means that the estimator $\widehat{\theta}_n$ and its target parameter θ has the following elegant relation:

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{D} N(0, I^{-1}(\theta)), \tag{3.2}$$

where $\sigma^2(\theta)$ is called the *asymptotic variance*; it is a quantity depending only on θ (and the form of the density function). Simply put, the asymptotic normality refers to the case where we have the convergence in distribution to a Normal limit centered at the target parameter. Moreover, this asymptotic variance has an elegant form:

$$I(\theta) = \mathbb{E}\left(\left(\frac{\partial}{\partial\theta}\log p(X;\theta)\right)^2\right) = \mathbb{E}\left(s^2(\theta|X)\right). \tag{3.3}$$

The asymptotic variance $I(\theta)$ is also called the *Fisher information*. This quantity plays a key role in both statistical theory and information theory.

Here is a simplified derivation of equation (**??**) and (**??**). Let $X_1, \cdots, X_n \sim p_{\theta_0}$, where θ_0 is the parameter generating the random sample. For simplicity, we assume $\theta_0 \in \mathbb{R}$ and θ_0 satisfies

$$\theta_0 = \underset{\theta}{\mathsf{argmax}}\ \mathbb{E}(\ell(\theta|X)).$$

Let $\widehat{\theta}_n$ be the MLE. Recall that the MLE solves the equation

$$\sum_{i=1}^{n} s(\widehat{\theta}_n|X_i) = 0.$$

Because θ_0 is the maximizer of $\mathbb{E}(\ell(\theta|X))$, it also satisfies $\mathbb{E}(s(\theta|X)) = 0$. Now consider the following expansion:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} s(\theta_0|X_i) - \underbrace{\mathbb{E}(s(\theta_0|X))}_{=0} &= \frac{1}{n}\sum_{i=1}^{n} s(\theta_0|X_i) - \underbrace{\sum_{i=1}^{n} s(\widehat{\theta}_n|X_i)}_{=0} \\
&= f_n(\theta_0) - f_n(\widehat{\theta}_n) \\
&= (\theta_0 - \widehat{\theta}_n) f_n'(\theta^*), \quad \theta^* \in [\theta_0, \widehat{\theta}_n] \quad \text{by mean value theorem} \\
&\approx (\theta_0 - \widehat{\theta}_n) f_n'(\theta_0) \\
&= (\theta_0 - \widehat{\theta}_n)\frac{1}{n}\sum_{i=1}^{n} s'(\theta_0|X_i) \\
&\approx (\theta_0 - \widehat{\theta}_n)\mathbb{E}(s'(\theta_0|X_i)).
\end{aligned} \tag{3.4}$$

By Central Limit Theorem, the right hand side of equation (**??**)

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} s(\theta_0|X_i) - \mathbb{E}(s(\theta_0|X))\right) \xrightarrow{D} N(0, \sigma^2(\theta)), \tag{3.5}$$

where $\sigma^2(\theta) = \mathsf{Var}(s(\theta_0|X_i))$.

Therefore, rearranging the quantities in equation (**??**) and (**??**),

$$\begin{aligned}
\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) &= \sqrt{n}\cdot\frac{-1}{\mathbb{E}(s'(\theta_0|X_i))}\cdot\left(\frac{1}{n}\sum_{i=1}^{n} s(\theta_0|X_i) - \mathbb{E}(s(\theta_0|X))\right) \\
&= \frac{-1}{\mathbb{E}(s'(\theta_0|X_i))}\cdot\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} s(\theta_0|X_i) - \mathbb{E}(s(\theta_0|X))\right) \\
&\xrightarrow{D} N\left(0, \frac{\sigma^2(\theta)}{\left(\mathbb{E}(s'(\theta_0|X_i)\right)^2}\right).
\end{aligned} \tag{3.6}$$

Now we have already shown the asymptotic normality. The next step is to show the asymptotic variance $\frac{\sigma^2(\theta)}{(\mathbb{E}(s'(\theta_0|X_i))^2} = I(\theta_0)$.

First, we expand $\sigma^2(\theta)$:

$$\sigma^2(\theta) = \mathsf{Var}(s(\theta_0|X_i)) = \underbrace{\mathbb{E}(s^2(\theta_0|X_i))}_{=I(\theta)} - \left(\underbrace{\mathbb{E}(s(\theta_0|X_i))}_{=0}\right)^2. \tag{3.7}$$

Second, we expand $\mathbb{E}(s'(\theta_0|X_i))$. But first we focus on $s'(\theta_0|X_i)$:

$$\begin{aligned}
s'(\theta_0|X_i) = \frac{\partial^2}{\partial\theta_0^2}\log p(X_i;\theta_0) &= \frac{\partial}{\partial\theta_0}\frac{\frac{\partial}{\partial\theta_0}p(X_i;\theta_0)}{p(X_i;\theta_0)} \\
&= \frac{\frac{\partial^2}{\partial\theta_0^2}p(X_i;\theta_0)}{p(X_i;\theta_0)} - \left(\frac{\frac{\partial}{\partial\theta_0}p(X_i;\theta_0)}{p(X_i;\theta_0)}\right)^2 \\
&= \frac{\frac{\partial^2}{\partial\theta_0^2}p(X_i;\theta_0)}{p(X_i;\theta_0)} - \left(\frac{\partial}{\partial\theta_0}\log p(X_i;\theta_0)\right)^2 \\
&= \frac{\frac{\partial^2}{\partial\theta_0^2}p(X_i;\theta_0)}{p(X_i;\theta_0)} - s^2(\theta_0|X_i).
\end{aligned}$$

For the first quantity,

$$\mathbb{E}\left(\frac{\frac{\partial^2}{\partial\theta_0^2}p(X_i;\theta_0)}{p(X_i;\theta_0)}\right) = \int\frac{\frac{\partial^2}{\partial\theta_0^2}p(x;\theta_0)}{p(x;\theta_0)}p(x;\theta_0)dx = \int\frac{\partial^2}{\partial\theta_0^2}p(x;\theta_0)dx = \frac{\partial^2}{\partial\theta_0^2}\underbrace{\int p(x;\theta_0)dx}_{=1} = 0.$$

Note that because we exchange the positions of the derivative and the integration, we assume that the parameter is independent of the support of the density function.

Thus,

$$\mathbb{E}(s'(\theta_0|X_i)) = \mathbb{E}\left(\frac{\frac{\partial^2}{\partial\theta_0^2}p(X_i;\theta_0)}{p(X_i;\theta_0)}\right) - \mathbb{E}\left(s^2(\theta_0|X_i)\right) = -\mathbb{E}\left(s^2(\theta_0|X_i)\right) = -I(\theta_0). \tag{3.8}$$

Plugging equations (**??**) and (**??**) into equation (**??**), we conclude

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{D} N\left(0, \frac{I(\theta_0)}{I^2(\theta_0)}\right) \overset{D}{=} N\left(0, I^{-1}(\theta_0)\right), \tag{3.9}$$

which is the desired result.

As a byproduct, we also showed that

$$I(\theta_0) = \mathbb{E}\left(s^2(\theta_0|X)\right) = -\mathbb{E}\left(s'(\theta_0|X)\right).$$

Equation (**??**) provides several useful properties of MLE:

- The MLE $\widehat{\theta}_n$ is an unbiased estimator of θ_0.
- The mean square error (MSE) of $\widehat{\theta}_n$ is

$$\mathsf{MSE}\left(\widehat{\theta}_n, \theta_0\right) = \underbrace{\mathbf{bias}^2(\widehat{\theta})}_{=0} + \mathsf{Var}(\widehat{\theta}_n) \approx \frac{1}{n \cdot I(\theta_0)}.$$

- The estimator error of $\widehat{\theta}_n$ is asymptotically Normal.
- If we know $I(\theta_0)$ or have an estimate $\widehat{I}(\theta_0)$ of it , we can construct a $1-\alpha$ confidence interval using

$$\left[\widehat{\theta}_n - \frac{z_{1-\alpha/2}}{\sqrt{n\widehat{I}(\theta_0)}},\ \widehat{\theta}_n - \frac{z_{1-\alpha/2}}{\sqrt{n\widehat{I}(\theta_0)}}\right].$$

- If we want to test $H_0 : \theta_0 = \theta^*$ versus $H_a : \theta_0 \neq \theta^*$ under significance level α, we can first compute $I(\theta^*)$ and then reject the null hypothesis if

$$\sqrt{nI(\theta^*)} \cdot |\widehat{\theta}_n - \theta^*| \geq z_{1-\alpha/2}.$$

Example. (Normal distribution) In the example where $X_1, \cdots, X_n \sim N(\mu, 1)$, we have seen that $s(\mu|X_i) = \mu - X_i$. Thus, $I(\mu) = \mathbb{E}(s^2(\mu|X_i)) = \mathbb{E}((\mu - X_i)^2) = 1$ because the variance is 1. Moreover, $\mathbb{E}(s'(\mu|X_i)) = \mathbb{E}(1) = 1 = I(\mu)$ agrees with the fact that $\mathbb{E}(s^2(\mu|X_i)) = \mathbb{E}(s'(\mu|X_i))$. Another way to check this is directly compute the variance of the MLE $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$, which equals to $\frac{1}{n} = \frac{1}{nI(\mu)}$. Again, we obtain $I(\mu) = 1$.

Example. (Exponential distribution) For the example where $X_1, \cdots, X_n \sim \mathsf{Exp}(\lambda)$, the Fisher information is more involved. Because the MLE $\widehat{\lambda}_n = \frac{n}{\sum_{i=1}^n X_i}$, we cannot directly compute its variance. However, using the Fisher information, we can still obtain its asymptotic variance. Recall that $s(\lambda|X_i) = \frac{1}{\lambda} - X_i$. Thus,

$$\mathbb{E}(s^2(\lambda|X_i)) = \mathbb{E}\left(\frac{1}{\lambda^2} - \frac{2X_i}{\lambda} + X_i^2\right).$$

For an exponential distribution $\mathsf{Exp}(\lambda)$, its mean is λ^{-1} and variance is λ^{-2}, which implies the second moment $\mathbb{E}(X_i^2) = \mathsf{Var}(X_i) + \mathbb{E}^2(X_i) = \frac{2}{\lambda^2}$. Putting it altogether,

$$I(\lambda) = \mathbb{E}(s^2(\lambda|X_i)) = \frac{1}{\lambda^2} - \frac{2}{\lambda^2} + \frac{2}{\lambda^2} = \frac{1}{\lambda^2}.$$

If we use $\mathbb{E}(s'(\lambda|X_i))$, we obtain

$$\mathbb{E}(s'(\lambda|X_i)) = \frac{1}{\lambda^2} = I(\lambda),$$

which agrees with the previous result.

♦ : Note that we may obtain an approximation of the variance of the MLE $\widehat{\lambda}$ by the *delta method*[5].

♠ : Not all MLEs have the above elegant properties. There are MLEs who do not satisfy the assumption of the theory. For instance, in the example of $X_1, \cdots, X_n \sim \mathsf{Uni}$ over $[0, \theta]$, the MLE $\widehat{\theta}_n = X_{(n)}$ does not satisfy the assumption. An assumption is that the support of the density does not depend on the parameter but here the support is $[0, \theta]$, which depends on the parameter.

3.4 Simple Linear Regression

Under certain conditions, the least square estimator (LSE) in linear regression can be framed as an MLE of the regression slope and intercept. Let's recall the least square estimator of the linear regression first. We observe IID bivariate random variables $(X_1, Y_1), \cdots, (X_n, Y_n)$ such that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where the ϵ_i is a mean 0 noise independent of X_i. We also assume that the marginal distribution of the covariate $X_1, \cdots, X_n \sim p_X$.

The LSE finds $\widehat{\beta}_{0,LSE}, \widehat{\beta}_{1,LSE}$ by

$$(\widehat{\beta}_{0,LSE}, \widehat{\beta}_{1,LSE}) = \underset{\beta_0, \beta_1}{\mathsf{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 . \tag{3.10}$$

In other words, we are choose $\widehat{\beta}_{0,LSE}, \widehat{\beta}_{1,LSE}$ such that the sum of squares of errors is minimized.

Now, here is the key assumption the links the MLE and the LSE: We assume that

$$\epsilon_1, \cdots, \epsilon_n \sim N(0, \sigma^2). \tag{3.11}$$

Namely, the noises are IID from a mean 0 Normal distribution.

Under equation (**??**), what is the MLE of β_0 and β_1? Here is a derivation of that. Let $p_\epsilon(e)$ be the distribution of the ϵ_i . The joint density of (X_i, Y_i) is

$$p_{XY}(x, y) = p(y|x) \cdot p_X(x) = p_\epsilon(y - \beta_0 - \beta_1 x) \cdot p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x)^2} \cdot p_X(x).$$

Thus, the log-likelihood function of β_0, β_1 is

$$\begin{aligned} \ell(\beta_0, \beta_1 | X_1, Y_1, \cdots, X_n, Y_n) &= \sum_{i=1}^{n} \log p_{XY}(X_i, Y_i) \\ &= \sum_{i=1}^{n} \left(\underbrace{-\frac{1}{2} \log(2\pi\sigma^2)}_{\perp \beta_0, \beta_1} - \frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 X_i)^2 + \underbrace{\log p_X(X_i)}_{\perp \beta_0, \beta_1} \right). \end{aligned}$$

Therefore, only the center term is related to the parameter of interest β_0, β_1. This means that the MLE is also the maximizer of

$$\ell^*(\beta_0, \beta_1 | X_1, Y_1, \cdots, X_n, Y_n) = -\sum_{i=1}^{n} \frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 X_i)^2 .$$

[5] https://en.wikipedia.org/wiki/Delta_method and http://www.stat.cmu.edu/~larry/=stat705/Lecture4.pdf

Because multiplying $\ell^*(\beta_0, \beta_1 | X_1, Y_1, \cdots, X_n, Y_n)$ by any positive number will not affect the maximizer, we multiply it by $2\sigma^2$, which leads to the following criterion:

$$\ell^\dagger(\beta_0, \beta_1 | X_1, Y_1, \cdots, X_n, Y_n) = -\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 .$$

Accordingly, the MLE finds $\widehat{\beta}_{0,MLE}, \widehat{\beta}_{1,MLE}$ by maximizing $-\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$, which is equivalent to minimizing $\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$, the same criterion as equation (**??**). So the LSE and the MLE coincides in this case and the theory of MLE applied to LSE (asymptotic normality, Fisher information, ...etc).

♠ : Think about the closed form of $\widehat{\beta}_{0,LSE}, \widehat{\beta}_{1,LSE}$ given IID bivariate random variables $(X_1, Y_1), \cdots, (X_n, Y_n)$. And also think about if they converges to the true parameter β_0, β_1 and if we can derive the asymptotic normality of them (you may need to use the *Slutsky's theorem*[6]).

3.5 Logistic Regression

Now we consider a special case in the regression problem: binary response case. We observe IID bivariate random variables

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

such that Y_i takes value 0 and 1. Here Y_i's are called the response variable and X_i's are called the covariate. Because Y_i takes value between 0 and 1, we can view it as a Bernoulli random variable with parameter q but this parameter q depends on the value of the corresponding covariate X_i. Namely, if we observe $X_i = x$, then the probability $P(Y_i = 1 | X_i = x) = q(x)$ for some function $q(x)$. Here are some examples why this model is reasonable.

Example. In graduate school admission, we are wondering how a student's GPA affects the chance that this applicant received the admission. In this case, each observations is a student and the response variable Y represents whether the student received admission ($Y = 1$) or not ($Y = 0$). GPA is the covariate X. Thus, we can model the probability

$$P(\mathsf{admitted} | \mathsf{GPA} = x) = P(Y = 1 | X = x) = q(x).$$

Example. In medical research, people are often wondering if the heretability of the type-2 diabetes is related to some mutation from of a gene. Researchers record if the subject has the type-2 diabetes (response) and measure the mutation signature of genes (covariate X). Thus, the response variable $Y = 1$ if this subject has the type-2 diabetes. A statistical model to associate the covariate X and the response Y is through

$$P(\text{subject has type-2 diabetes} | \text{mutation signature} = x) = P(Y = 1 | X = x) = q(x).$$

Thus, the function $q(x)$ now plays a key role in determining how the response Y and the covariate X are associated. The logistic regression provides a simple and elegant way to characterize the function $q(x)$ in a 'linear' way. Because $q(x)$ represents a *probability*, it ranges within $[0, 1]$ so naively using a linear regression will not work. However, consider the following quantity:

$$O(x) = \frac{q(x)}{1 - q(x)} = \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)} \in [0, \infty).$$

The quantity $O(x)$ is called the *odds* that measures the contrast between the event $Y = 1$ versus $Y = 0$. When the odds is greater than 1, we have a higher change of getting $Y = 1$ than $Y = 0$. The odds

[6] https://en.wikipedia.org/wiki/Slutsky's_theorem

has an interesting asymmetric form– if $P(Y = 1|X = x) = 2P(Y = 0|X = x)$, then $O(x) = 2$ but if $P(Y = 0|X = x) = 2P(Y = 1|X = x)$, then $O(x) = \frac{1}{2}$. To symmetrize the odds, a straight-forward approach is to take (natural) logarithm of it:

$$\log O(x) = \log \frac{q(x)}{1 - q(x)}.$$

This quantity is called *log odds*. The log odds has several beautiful properties, for instance when the two probabilities are the same ($P(Y = 1|X = x) = P(Y = 0|X = x)$), $\log O(x) = 0$, and

$$\begin{aligned} P(Y = 1|X = x) = 2P(Y = 0|X = x) &\Rightarrow \log O(x) = \log 2 \\ P(Y = 0|X = x) = 2P(Y = 1|X = x) &\Rightarrow \log O(x) = -\log 2. \end{aligned}$$

The logistic regression is to impose a linear model to the log odds. Namely, the logistic regression models

$$\log O(x) = \log \frac{q(x)}{1 - q(x)} = \beta_0 + \beta_1 x$$

leading to

$$P(Y = 1|X = x) = q(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

Thus, the quantity $q(x) = q(x; \beta_0, \beta_1)$ depends on the two parameter β_0, β_1. Here β_0 behaves like the intercept and β_1 behaves like the slope (they are the intercept and slope in terms of the log odds).

When we observe data, how can we estimate these two parameters? In general, people will use the MLE to estimate them and here is the likelihood function of logistic regression. Recall that we observe IID bivariate random sample:

$$(X_1, Y_1), \cdots, (X_n, Y_n).$$

Let $p_X(x)$ denotes the probability density of X; note that we will not use it in estimating β_0, β_1. For a given pair X_i, Y_i, recalled that the random variable Y_i given X_i is just a Bernoulli random variable with parameter $q(x = X_i)$. Thus, the PMF of Y_i given X_i is

$$\begin{aligned} L(\beta_0, \beta_1|X_i, Y_i) = P(Y = Y_i|X_i) &= q(X_i)^{Y_i}(1 - q(X_i))^{1-Y_i} \\ &= \left(\frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}}\right)^{Y_i} \left(\frac{1}{1 + e^{\beta_0+\beta_1 X_i}}\right)^{1-Y_i} \\ &= \frac{e^{\beta_0 Y_i+\beta_1 X_i Y_i}}{1 + e^{\beta_0+\beta_1 X_i}}. \end{aligned}$$

Note that here we construct the likelihood function using only the conditional PMF because similarly to the linear regression, the distribution of the covariate X does not depends on the parameter β_0, β_1. Thus, the log-likelihood function is

$$\begin{aligned} \ell(\beta_0, \beta_1|X_1, Y_1, \cdots, X_n, Y_n) &= \sum_{i=1}^{n} \log L(\beta_0, \beta_1|X_i, Y_i) \\ &= \sum_{i=1}^{n} \log \left(\frac{e^{\beta_0 Y_i+\beta_1 X_i Y_i}}{1 + e^{\beta_0+\beta_1 X_i}}\right) \\ &= \sum_{i=1}^{n} \beta_0 Y_i + \beta_1 X_i Y_i - \log\left(1 + e^{\beta_0+\beta_1 X_i}\right). \end{aligned}$$

We can then take derivative to find the maximizer.

However, the derivative of the log-likelihood function $\ell(\beta_0, \beta_1 | X_1, Y_1, \cdots, X_n, Y_n)$ does not have a closed-form solution so we cannot write down a simple expression of the estimator. Despite this disadvantage, such a log-likelihood function can be optimized by gradient ascent approach such as the Newton-Raphson[7].

[7] some references can be found: https://www.cs.princeton.edu/~bee/courses/lec/lec_jan24.pdf