BRN Research Methods Workshop 3

3. OLS vs MLE- A Gateway to Generalized Linear Model

Inhwan Ko (Univ. of Washington, Seattle)

July 29th, 2021

Contents

Review Questions

- What is the Law of Large Numbers?
- What is the Central Limit Theorem?
- What is p-value?
- What is confidence interval?
- What is the standard error of the regression (σ^2?)
- What is the square root of the mean of the squared errors (RMSE)?

Refresher

Good Beta, Bad Beta

A good beta estimator should have three characteristics:

1. No bias
2. Efficiency
3. Consistency

Let's take a look at each.

Unbiased Estimator

Bias can be formally written as:

$$\mathbb{E}(\hat{\beta} - \beta)$$

That is: how much is the estimate $\hat{\beta}$ expected to be different from the true parameter β?

The more unbiased an estimator is, the closer it is to the true parameter.

Efficient Estimator

If we were to take multiple times of sample and derive a beta estimate,

how often do we get the estimate that is close to the true parameter?

This is the question of efficiency: remember, "how often". So we need to consider not only the bias but also the variance of estimator.

We calculate efficiency with mean squared error:

$$\text{MSE} = \frac{1}{n}\text{RSS} = \mathbb{E}[(\beta - \hat{\beta})^2] = Var(\hat{\beta}) + \text{Bias}$$

If there is no bias, then MSE reduces to var($\hat{\beta}$).

Consistent Estimator

An estimator is consistent when bias converges to zero as $N \to \infty$.

Although this sounds unquestionably clear, there are some cases where bias does not coverge to zero even though the sample size approaches infinity.

Yet, not a big concern relative to the two earlier concepts.

Unbiasedness and Efficiency

It is happy to have an unbiased and efficient estimator all the time. But it is not possible always. What do we mean by having (un)biased and (in)efficient estimator?

Let's say there is a sniper trying to practice in a shooting range. The sniper is the best on earth- as long as the rifle has no problem, the sniper will always send the bullets to the center.

But let's say the rifle may have two problems: ACOG(advanced combat optical gunsight) and gun barrel.

ACOG: relates to unbiasedness
Gun barrel: relates to efficiency

If a sniper shoots with good ACOG (unbiased) but bad barrel (inefficient), although the sniper will correctly aim at the bull's eye, the shot group will scatter. Nevertheless, on average a sniper has aimed at the bull's eye.

If a sniper shoots with bad ACOG (biased) but good barrel (efficient), the sniper's aim will be incorrect but the shot group will relatively less scatter.

Gauss-Markov Assumptions for Linear Regression

Even if our sample is representative of the population and the model is correctly specified (meaning we don't have any pre-treatment bias), there are still a few problems for conducting linear regression safely.

They relate to the key assumptions of linear regression to have "Best Linear Unbiased and Efficient" estimate, or BLUE. Those assumptions are:

1. No perfect collinearity
2. Exogenous covariates
3. Mean zero
4. Homoskedasticity
5. No error correlation
6. Non-normal disturbances

Let's take a look at each.

No Perfect Collinearity

Perfect collinearity occurs when $X'X$ is singular.

If this happens, $|X'X| = 0$ (determinant is zero), and β cannot be defined (since $\beta = (X'X)^{-1}X'Y$).

But no worries, usually R or other stat tools will automatically drop a few covariates to way around this issue.

Endogeneity & Exogeneity

Our linear regression in matrix form, $Y = X\beta + \epsilon$ implies that Y is endogenous to X. This means that Y is determined by X as its systematic component.

In other words, X is exogenous to Y as it is assumed to cause Y, not caused by Y. That is, all variance of X must be independent from Y while all variance of Y must be solely dependent upon X.

This second GM assumption is violated when the relationship between X and Y switches- especially when we cannot identify correctly whether X or Y is the cause of the other. $\mathbb{E}(x_i, e_i) \neq 0$

Violating this will give us biased and inconsistent estimator.

Mean Zero

Mean zero means $\mathbb{E}(\epsilon) = 0$. This is because we assume that:

$$\mathbb{E}(\mathrm{y}) = X\beta$$

Violating this will give us biased and inconsistent estimator.

Homoskedasticity

Homoskedasticity means the diagonal elements of Σ are not the same. Therefore:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

This means that how far each individual's observed outcome, y_i, is from the expected outcome predicted from the model, $\hat{y}$, is different across individuals.

If this happens, our linear regression will ignore why each individual has different distance between the predicted line and its observed value and instead **produce an inefficient (but unbiased) estimator**. This situation is called "heteroskedasticity".

No error correlation

Errors should not have correlation with each other. Formally, $\mathbb{E}(\varepsilon_i, \varepsilon_j) = 0$. If this happens along with heteroskedasticity, let's see what happens to our Σ.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Violating this too will give us unbiased yet inefficient estimator.

Non-Normal Disturbances

Relatively simple, this formally writes: $\epsilon \sim N(0, \sigma^2)$. If this occurs, unless $N \to \infty$ our standard error will be biased.

Regression?

Recall that "regression" implies there is an assumption that something many "regress" towards a certain value.

And those most likely regress toward the "mean" value.
Also recall that the regression coefficients are "conditional means" ($\mathbb{E}(Y|X,$ while other variables held constant)).

If we do not know those conditional means ex ante, we need to estimate, and the logic of estimation was to calculate the value that minimizes the error.

In order for such a beta coefficient to be BLUE, the model needs to satisfy the GM assumptions. However, what if such assumptions are not met?

Maximum Likelihood Estimation (MLE)

Recall we used a partial derivation to find the beta coefficient, β, that minimizes the residual sum of squares, $\sum_{i=1}^{n} \varepsilon_i^2$.

Instead, in MLE, we *maximize* something.

And we maximize the **likelihood**.

Likelihood of what?

Maximum Likelihood Estimation (MLE)

Remember that we understand our model to consist of two components:

1. Stochastic component
2. Systematic component

In OLS, we minimized our stochatic component to find the coefficient.
In MLE, we maximize the *likelihood* of our systematic component.

To write both components into functional forms:

1. Stochastic component: $\text{y} \sim f(\mu, \alpha)$
2. Systematic component: $\mu \sim g(X, \beta)$

In MLE, let's denote our estimates (β, σ^2) as θ. We need to maximize the likelihood of θ given that we have observations, y.

So we maximize $P(\theta|\text{y})$.

Bayesian Approach

Using *Bayes Rule*, we know that:

$$P(\theta|\text{y}) = \frac{P(\theta)P(\text{y}|\theta)}{P(\text{y})}$$

In Bayesian analysis, we (subjectively) guess what the $P(\theta)$ was before we collected the data, y. Then we calculate $P(\text{y}|\theta)$ to calculate $P(\theta|\text{y})$.

Likelihood Approach

However, in MLE, we consider $P(\theta)$ as something we cannot know. Thus:

$$P(\theta|\mathrm{y}) = \frac{P(\theta)}{P(\mathrm{y})} P(\mathrm{y}|\theta)$$
$$\mathcal{L}(\theta|\mathrm{y}) = k(\mathrm{y}) \times P(\mathrm{y}|\theta)$$
$$\mathcal{L}(\theta|\mathrm{y}) \propto P(\mathrm{y}|\theta)$$

Note that $P(\theta|\mathrm{y})$ is now denoted as $\mathcal{L}(\theta|\mathrm{y})$. This indicates the likelihood of θ given our observation, y.

Although we don't know the $k(\mathrm{y})$, we know that the likelihood is proportional to the probability of having our observation given our estimates, $P(\mathrm{y}|\theta)$. This is something we can calculate from our data.

Hence, we maximize the likelihood by maximizing its proportionate, $P(\mathrm{y}|\theta)$.

Flexibility of MLE

In OLS, we assumed that our observation and error terms asymptotically follow the Normal distribution.

In MLE, we do not have to. Instead, we can assume them to follow various probabilistic distributions we know.

- If our y follows the Normal: $P(\mathrm{y}_i|\theta_i) = f\mathcal{N}(\mathrm{y}_i|\mu_i, \sigma^2)$.
- If our y follows the Bernoulli: $P(\mathrm{y}_i|\theta) = f_{Bern}(\pi_i)$.
- If our y follows the Poisson: $P(\mathrm{y}_i|\theta) = f_{Pois}(\lambda_i)$.
- And more. . . .

Bernoulli Distribution

Let's say our observation, y, follows the Bernoulli distribution. In other words, it is a random variable, either 0 or 1, drawn from the population that follows the Bernoulli distribution.

We know that our first observation, y_1, is a function of this Bernoulli distribution:

$$P(y_1|\pi_1) = f_{Bern}(\pi_1)$$

So is our second observation, y_2.

$$P(y_2|\pi_2) = f_{Bern}(\pi_2)$$

Bernoulli Distribution

Their joint probability will be:

$$P(y_1, y_2|\pi_1, \pi_2) = f_{Bern}(\pi_1) \times f_{Bern}(\pi_2)$$

And the joint probability of all observations will be:

$$\begin{aligned} P(y_1, y_2, \ldots, y_i|\pi_1, \pi_2, \ldots, \pi_i) &= f_{Bern}(\pi_1) \times f_{Bern}(\pi_2) \times \ldots \ f_{Bern}(\pi_i) \\ &= \prod_{i=1}^{n} f_{Bern}(\pi_i) \\ &= \prod_{i=1}^{n} \pi^{y_i}(1-\pi_i)^{1-y_i} \end{aligned}$$

Where π_i is the probability that $y_i = 1$.

Bernoulli Distribution

Since it's hard to calculate the product of multiple exponents, we take the log:

$$\log[P(y_1, y_2, \ldots, y_i|\pi_1, \pi_2, \ldots, \pi_i)] = log[\prod_{i=1}^{n} \pi^{y_i}(1-\pi_i)^{1-y_i}]$$

$$= \sum_{i}^{n}[y_i \log \pi_i] - \sum_{i}^{n}[y_i \log(1-\pi_i)] + \sum_{i}^{n} \log(1-\pi_i)$$

$$= \sum_{i}^{n}[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)]$$

Therefore, we find π which maximizes this log-likelihood function.

Bernoulli Experiment

If we assume that $y_i (i = 1, 2, ..., 8)$ is defined as:

```
y <- c(1,0,0,1,0,0,0,0)
y
```

```
## [1] 1 0 0 1 0 0 0 0
```

which follows a Bernoulli distribution with its parameter π,

$$y_i \sim f_{Bern}(\pi_i)$$

Bernoulli Experiment

We can write down its likelihood function:

$$\begin{aligned} L(\pi_i|y_i) &\propto P(y_i|\pi_i) \\ &= f_{Bern}(\pi_1) \times f_{Bern}(\pi_2) \times \cdots \times f_{Bern}(\pi_8) \\ &= \prod_{i=1}^{8} f_{Bern}(\pi_i) \\ &= \prod_{i=1}^{8} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \end{aligned}$$

Therefore, plugging in our data will change it into:

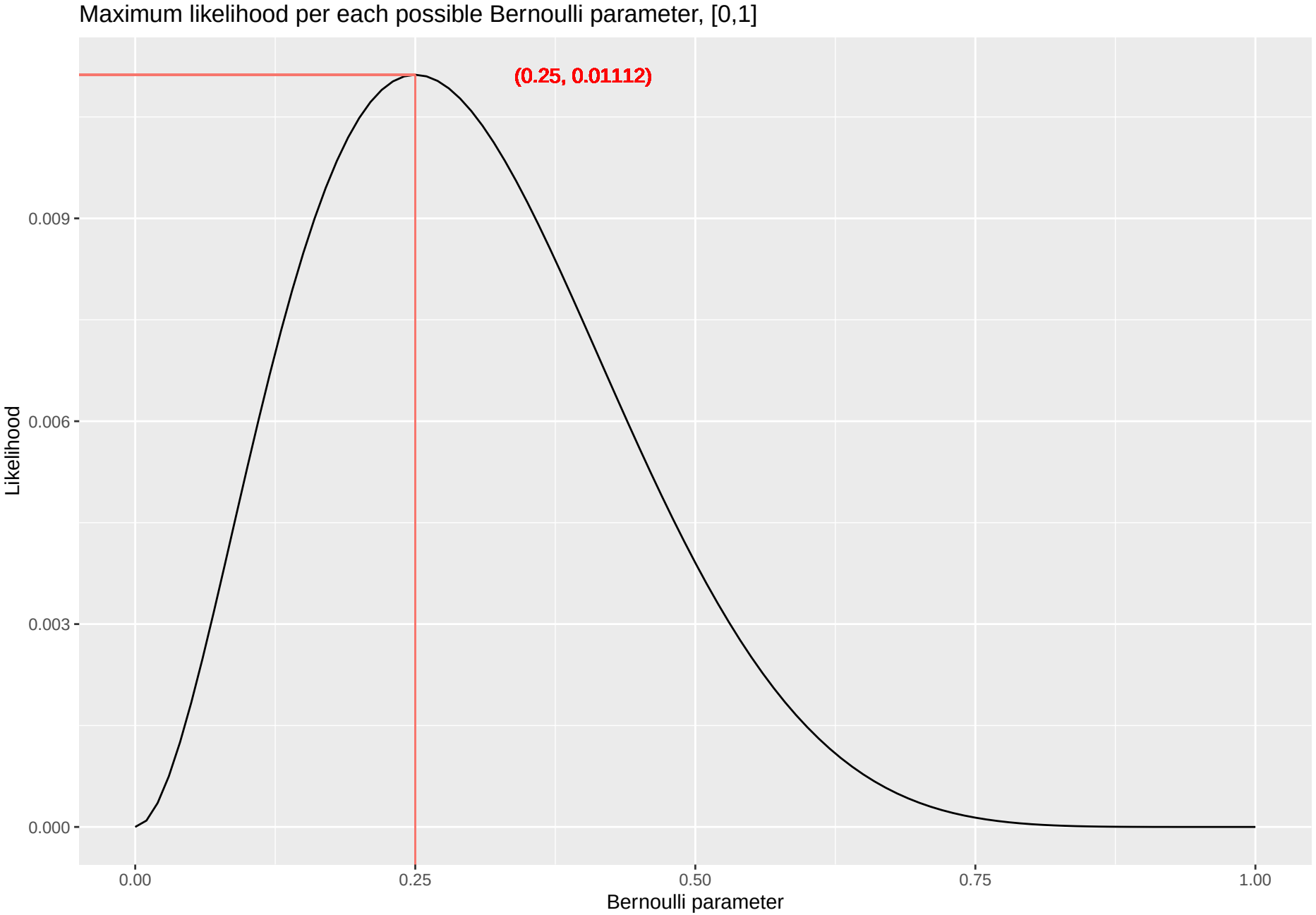
$$L(\pi_i|y_i) \propto \pi^2 \times (1-\pi)^6$$

Bernoulli Experiment

```r
pi <- seq(from=0, to=1, by=0.01)

L <- function(pi){
  val= (pi^2)*((1-pi)^6)
  return(val)
} # transform the likelihood equation into an R function

likelihood <- data.frame(pi=pi, y=L(pi)) %>%
  arrange(desc(y)) # arrange by making y decrease to find the ML and MLE

head(likelihood) # pi=0.25 is the MLE
```

```
##     pi          y
## 1 0.25 0.01112366
## 2 0.26 0.01110036
## 3 0.24 0.01109952
## 4 0.27 0.01103227
## 5 0.23 0.01102554
## 6 0.28 0.01092222
```

Bernoulli Experiment



Link Function

Link function translates our expected value of Y into the systematic component of our model, which is:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = X_i \beta$$

For instance, in logit regression, our link function is an inverse logit function

$$\pi_i = \text{logit}^{-1}(X_i\beta) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} = \frac{1}{1+\exp(-X_i\beta)}$$

Why do we use an inverse logit function?

From MLE to Link Function: A Case of Logit

$$L(\pi|y) \propto \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$L(\beta|y) \propto \prod_{i=1}^{n} (\frac{1}{1+\exp(-X_i\beta)})^{y_i}(1-(\frac{1}{1+\exp(-X_i\beta)}))^{1-y_i}$$

$$L(\beta|y) \propto \prod_{i=1}^{n} (1+\exp(-X_i\beta))^{-y_i}(1+\exp(-X_i\beta))^{-(1-y_i)}$$

$$\log L(\beta|y) \propto \sum_{i=1}^{n} [-y_i \log(1+\exp(-X_i\beta)) - (1-y_i)\log[(1+\exp(-X_i\beta))]$$

From MLE to Link Function: A Case of Logit

Now note that:

$$\pi = \text{logit}^{-1}(X_i\beta) \;\rightarrow\; \text{logit}(\pi) = X_i\beta$$

Therefore:

$$\log(\frac{\pi}{1-\pi}) = X_i\beta$$

Now, we need to estimate $log(\frac{\pi_i}{1-\pi_i})$. In fact, this is a logged **odds ratio.**

From MLE to Link Function: A Case of Logit

Odds ratio, $\frac{\pi}{1-\pi}$ indicates the probability of an event A occurs rather than does not occur, if the probability of an event A is π.

For instance, imagine you bet in a horse racing and your horse wins under 75% probability. Its odds ratio of winning is $0.75/1 - 0.75 = 0.75/0.25 = 3$. This indicates that the horse wins the race 3 times more likely than does not win. In our daily language, the horse wins 3 out of 4 games.

The logged odds ratio of π is the logit function of π which is then linked to the linear form of $X_i\beta$. Therefore:

$$log(\frac{\pi_i}{1-\pi_i}) = X_i\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i$$

That is, we want our model's systematic component to calculate the logged odds ratio of the probability that our dependent variable y is 1.

Interpretation of Odds Ratio

Remember, odds ratio $\neq$ OLS coefficients. Also, we want to know how much 1 unit increase in x changes the odds ratio, which is:

$$\frac{\text{odds}(x+1)}{\text{odds}(x)}$$

To interpret our model results using any regressions whose link function relates to odds ratio, we exponentiate the model result:

$$e^{\hat{\beta}}$$

And this equals $\frac{\text{odds}(x+1)}{\text{odds}(x)}$. I will show why.

Interpretation of Odds Ratio

For example, if our true model is $y_i = \beta_0 + \beta_1 x_i + e_i$, π_i will look like:

$$\text{odds}(x) = \frac{\pi_i}{1 - \pi_i} = e^{\beta_0 + \beta_1 x_i}$$

Let's denote this as odds(x). If we let x_i to increase by 1 unit:

$$\begin{aligned} \text{odds}(x+1) &= e^{\beta_0 + \beta_1 (x_i + 1)} \\ &= e^{\beta_0 + \beta_1 x_i} \times e^{\beta_1} \\ &= \text{odds}(x) \times e^{\beta_1} \end{aligned}$$

Interpretation of Odds Ratio

Therefore, we can calculate below:

$$\frac{\text{odds}(x+1)}{\text{odds}(x)} = \frac{\text{odds}(x) \times e^{\beta_1}}{\text{odds}(x)} = e^{\beta_1}$$

The same logic applies if we add more beta coefficients. Therefore, we need to exponentiate our beta from the model results using any regressions whose link function relates to odds ratio.

Binary and count models (i.e. logit, probit, poisson, negative binomial) all use odds ratio.

95% Confidence Interval of Odds Ratio

We judge whether our beta coefficient is significant at 95% confidence level by checking if its odds ratio has 95% confidence interval over/below zero.

We calculate the 95% confidence interval of odds ratio as follows:

1. Calculate a 95% confidence interval of $\hat{\beta}$.

$$\hat{\beta} - 1.96 \times \sigma\hat{\beta},\ \hat{\beta} + 1.96 \times \sigma\hat{\beta}$$

2. Exponentiate both ends of the 95% confidence interval of $\hat{\beta}$.

$$e^{\hat{\beta} - 1.96 \times \sigma\hat{\beta}},\ e^{\hat{\beta} + 1.96 \times \sigma\hat{\beta}}$$

How do we interpret if the 95% confidence interval is over zero? Below zero?

What if it includes zero?

Binomial Distribution

Bernoulli distribution is a part of a binomial distribution. For instance:

$$\frac{m!}{y!(m-y)!}\pi^y(1-\pi)^{m-y}$$

is a probability density function of a binomial distribution, where $y = 0, 1, 2..., m$. And Bernoulli distribution is a binomial distribution where $y = 0, 1$ and therefore $m = 1$. It reduces into:

$$\frac{1!}{y!(1-y)!}\pi^y(1-\pi)^{m-y} = \pi^y(1-\pi)^{(1-y)}$$

Poisson Distribution

Now, depending on our assumption on the distribution of our dependent variable, we can conduct MLE in various ways. If our dependent variable is a count variable (i.e. the number of civil war occurence in a country), we assume that it follows the Poisson distribution, therefore:

$$\begin{aligned} P(y_1, y_2, \ldots, y_i | \lambda_1, \lambda_2, \ldots, \lambda_i) &= f_{Pois}(\lambda_1) \times f_{Pois}(\lambda_2) \times \ldots \ f_{Pois}(\lambda_i) \\ &= \prod_{i=1}^{n} f_{Pois}(\lambda_i) \\ &= \prod_{i=1}^{n} \frac{\lambda^n e^{-\lambda}}{n!} \end{aligned}$$

And we take the same step as we did in the case of logit. I will show you (outside of the slide) that the poisson and binomial are actually related with each other.

More Appropriate Model?- A Goodness of Fit

After we specify multiple models, we need to decide which one fits our data the most.

This is called "a goodness of fit" test. We can take a look at three different statistics.

1. Likelihood-Ratio Test

$$LR = (-2 \times logL_o) - (-2 \times logL_1)$$

Where L_o is the maximum likelihood of the reference model and L_1 is that of the tested model. LR follows a chi-square distribution, so we use this value to test if there is a statistically significant improvement with reference to that distribution.

More Appropriate Model?- A Goodness of Fit

2. Akaike Information Criteria

It is known that the maximum likelihood increases simply because parameters to be estimated increase in number. Therefore, we penalize it by modifying the likelihood as below, which gives us Akaike Information Criteria (AIC):

$$AIC = 2k - 2logL$$

Where k is the number of parameters in our model.

More Appropriate Model?- A Goodness of Fit

3. Error-based criterion: SER, RSME, MAE

Or simply, we can use RSME as a reference. After we conduct MLE, we get the parameter estimate, therefore we can also get $\hat{y}$. This means we can calculate below:

$$\mathbb{E}((y-\hat{y})^2) = \mathbb{E}(\epsilon'\epsilon) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \sigma^2$$

As we all know, this is the standard error of the regression. Its root is the RMSE (Square root of the mean of the squared errors). Or, we can try:

$$\mathbb{E}(|y-\hat{y}|) = \mathbb{E}(|\epsilon'\epsilon|) = \frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|$$

This is called a mean absolute error.