PoliSci 4782 Political Analysis II

Probability Distributions and Generalized Linear Models

Seamus Wagner

The Ohio State University

Probability as a Model

$$Prob(Y|M) \equiv Prob(Data|Model)$$
$$where \quad M = (f, g, \boldsymbol{X}, \beta, \eta)$$

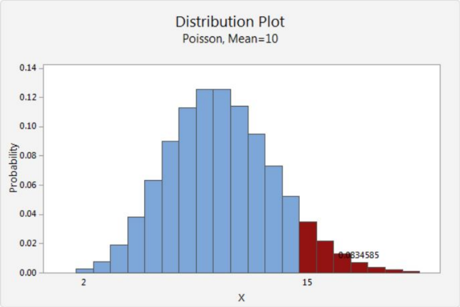
1 assume a model formula (f, g) with unknown parameters (β, η);
2 feed data ($\boldsymbol{X}$ and y) into the model to estimate those parameters to get the complete model formula;
3 use the model to predict Y

Probability

$Pr(\cdot)$ is defined by three axioms:

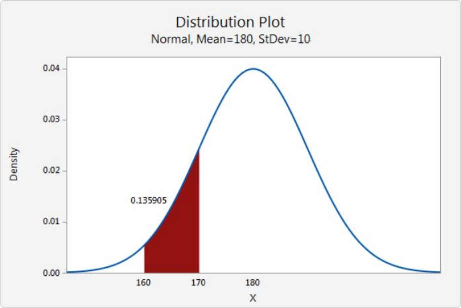
1. $Pr(y) \geq 0$ for any event y
2. $Pr(\phi) = 1$, where $\phi \ni \{Y | y_1, y_2, \cdots y_n\}$
3. if $y_1, y_2, \cdots y_n$ are mutually exclusive,
   $Pr(y_1 \cup y_2 \cup \cdots y_n) = Pr(y_1) + Pr(y_2) + \cdots Pr(y_n)$

Probability Mass Function (for discrete variables)



- $\sum_{all\ y} Pr(Y) = 1$
- $Pr(a \leq Y \leq b) = \sum_{a \leq Y \leq b} Pr(y) = Pr(Y \leq b) - Pr(Y \leq a)$

Probability Density Function (for continuous variables)



- $\int_{-\infty}^{\infty} f(y)dy = 1$
- $Pr(a \leq Y \leq b) = \int_a^b f(y)dy = \int_{-\infty}^b f(y)dy - \int_{-\infty}^a f(y)dy$

Linear Models Based on the Normal Distribution

$$E(Y) = \mu = \beta \boldsymbol{X}, \ \textit{where} \ Y \sim \mathcal{N}(\mu, \sigma^2)$$

- Simple linear regression is based on the assumption that $Y \sim N(\mu, \sigma^2)$
  - or at least the normal is a good approximation of the actual probability density of Y
- We can think of some aspects of Y to examine the validity of this assumption:
  - value assignment of Y (continuous and $-\infty \leq y \leq \infty$)
  - distribution of y in our sample (unimodal, symmetric/not terribly skewed, not terribly peaky or flat, etc.)

Linear Regression

$$M = (f, g, \boldsymbol{X}, \beta, \sigma^2)$$

- f is $\mathcal{N}(\mu, \sigma^2)$
- $g = \beta\boldsymbol{X}$, without any additional transformation (the simplest form)
- effect parameters β and ancillary parameter σ^2 need to be estimated

Estimation of β in OLS

The methodological idea is to minimize the sum of mean square errors to find the best β:

$$min \sum_{i=1}^{n} {\epsilon_i}^2 \ or \ min \sum_{i=1}^{n} (y_i - \boldsymbol{X_i}\beta)^2$$

Mathematically, it leads to the following estimator:

$$\hat{\beta} = (\boldsymbol{X^T X})\boldsymbol{X}^T \boldsymbol{Y}$$

where $\boldsymbol{X^T}$ *is the transpose of the original matrix* $\boldsymbol{X}$ *in linear algebra.*

Estimation of σ^2 in OLS

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

where n is the number of observations; k is the number of effect parameters (including the intercept); $\hat{\epsilon}$ is regression residual.

Estimation Uncertainty of β in OLS

The estimation uncertainty of $\hat{\beta}$ is indicated by its standard error:

$$se(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where x_i is the value for that explanatory variable X in observation i and $\bar{x}$ is the mean of X in the sample.

Deciding Statistical Significance: CI

- Confidence intervals are computed by

$$\hat{\beta} \pm [q_t \times \mathrm{se}(\hat{\beta})]$$

where the value for q_t is the appropriate t quantile for a given confidence level (conventionally 95%).

- As a rule of thumb, in a sample of $N \geq 50$, we can assume $q_t \approx 2$ at 95% confidence level. So the confidence interval is:

$$(\hat{\beta} - 2 \times se(\hat{\beta}), \hat{\beta} + 2 \times se(\hat{\beta}))$$

- If 0 is included in the CI, $\hat{\beta}$ is statistically insignificant at the given level; otherwise, significant.

Deciding Statistical Significance: T-Score Test

- Another way to test significance is a two-tailed t-score test (we get a non-zero estimate simply by chance).
- It is expressed as

$$H_0\colon \beta = 0,$$

$$H_1\colon \beta \neq 0,$$

  where H_0 is the null hypothesis (we get a our non-zero estimated value simply by chance) and H_1 is the alternative hypothesis.
- Like all other statistical tests, we gain confidence in H_1 by falsifying H_0.

Deciding Statistical Significance: T-Score Test

- In light of our null hypothesis, we calculate a t statistic in which β^* is set equal to 0:

$$t(df = n - k) = \frac{\hat{\beta} - \beta^*}{\mathrm{se}(\hat{\beta})}.$$

- k is the number of effect parameters (including the intercept) and df stands for degree of freedom.

- The t with the chosen df will correspond to a probability, which tells us how likely we observe something *as extreme or more extreme* as what we find in our data if the true β were 0.

- If this probability is sufficiently low (conventionally lower than 5%), we are confident in rejecting H_0 and believing $\hat{\beta}$.

Preparing for Transition to Generalized Linear Model

- Generalized linear model (GLM) is an extension of linear model (LM).
- In LM, we have

$$
\begin{aligned}
&Y \sim \mathcal{N}(\mu, \sigma^2) && \text{probability density of } Y \\
&g(\mu) = \mu = \boldsymbol{X}\boldsymbol{\beta} && \text{link function} \\
&E(Y) = \mu = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) && \text{mean function}
\end{aligned}
$$

- In other words, we do not need a special link function $g(\cdot)$ to travel between μ and $\boldsymbol{X}\boldsymbol{\beta}$ in LM.

Transition to GLM

- The probability distribution of Y becomes much flexible:
  - binary outcomes: whether to vote
  - count outcomes: number of protests
  - ordered categorical outcomes: high, medium, or low income
  - unordered categorical outcomes: White, African American, etc.
  - duration outcomes: incumbency of the president

- In GLM, a context-specific $g(\cdot)$ needs to be invented to seamlessly connect μ to $\boldsymbol{X\beta}$.

$$
\begin{aligned}
Y &\sim f(\mu, \eta) && \text{probability density of } Y \\
g(\mu) &= \boldsymbol{X\beta} && \text{link function} \\
\mu &= g^{-1}(\boldsymbol{X\beta}) && \text{mean function}
\end{aligned}
$$

Common GLMs in a Nutshell

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\beta = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\beta = \mu$	$\mu = \mathbf{X}\beta$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\beta = -\mu^{-1}$	$\mu = -(\mathbf{X}\beta)^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\beta = \mu^{-2}$	$\mu = (\mathbf{X}\beta)^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\beta = \ln(\mu)$	$\mu = \exp(\mathbf{X}\beta)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)} = \frac{1}{1+\exp(-\mathbf{X}\beta)}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$	outcome of single K-way occurrence			
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model

What Comes Next?

- Linear regression analysis in `R` (this week)
- Theoretical foundations for model estimation (next week)
- Maximum likelihood estimation as the mainstream method for estimating GLM (next week)