Poisson Regression

Jarvis T. Chen (jarvis@hsph.harvard.edu)

7 October 2025

Harvard T.H. Chan School of Public Health

Learning Objectives

- Be able to define a Poisson regression model for describing association between a count random variable and covariates, and interpret the parameters (or estimated parameters)
- Review Poisson regression as a member of the family of generalized linear models, maximum likelihood as a method for estimating parameters.
- Use an offset term in a Poisson regression model to fit a model describing the association between a rate and covariates
- Recognize that this approach can be used to take account of a variety of "exposures"
- Extend this concept to analysis of summary data and to standardization of rates (Appendix)

Readings

The following short readings complement the material covered in this class (emphasis on “complement” rather than “replicate”):

- Vittinghoff: 8.1, 8.3 (section introduction and 8.3.1) [Reading all of Chapter 8 gives a nice review of GLMs through examples].
- Wooldridge: section 17.3

Appendix also provides some more technical details about GLMs

The outcome Y is a count

Some examples:

- Number of asthma attacks experienced by a child in a year
- Number of worker's compensation claims filed in a five-year period
- Number of new HIV infections in a year in a specific geographical area
- Number of homicides in a city in a year
- Number of ...

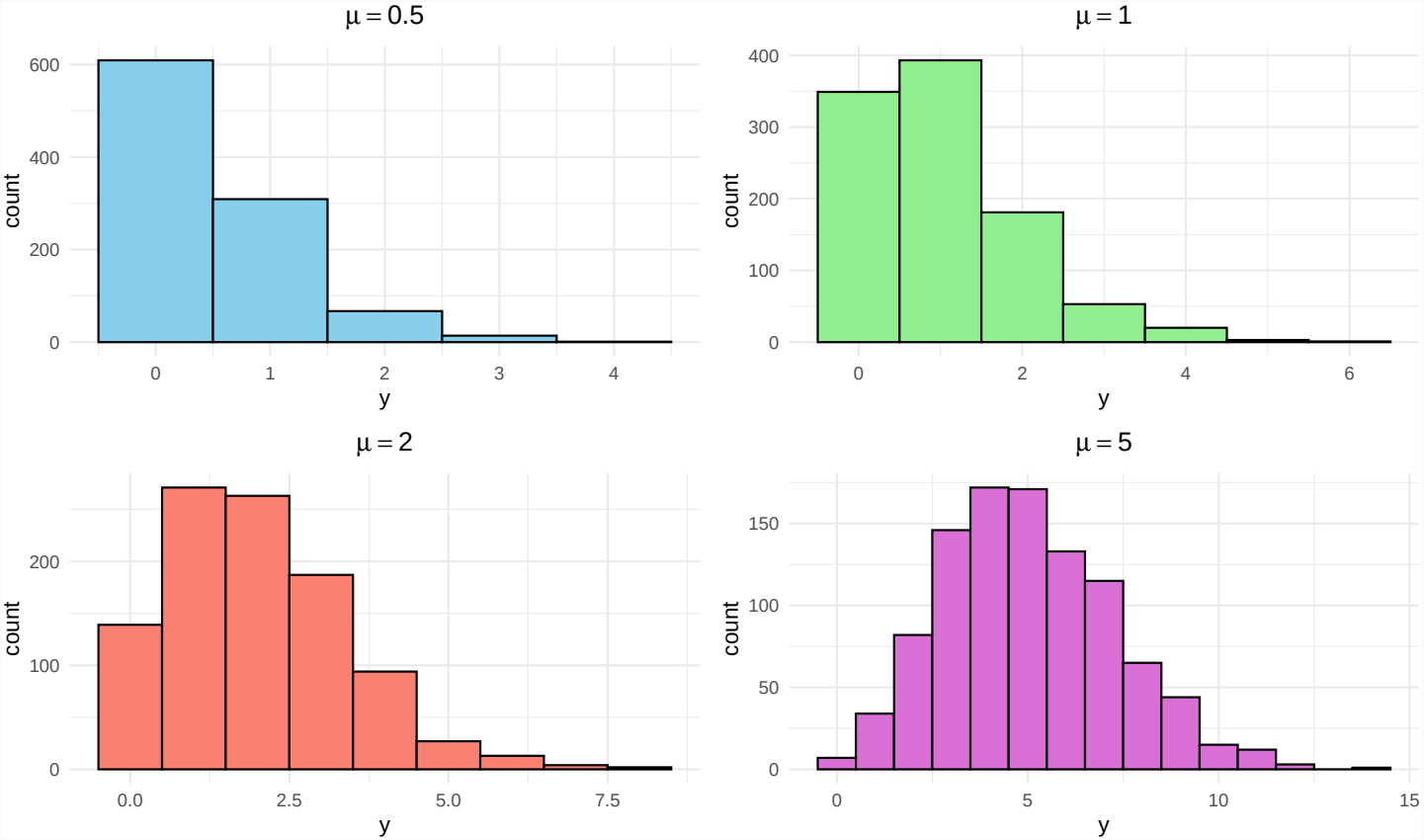
Random variable Y takes on non-negative integer values, so possible values are $y \in \{0, 1, 2, ...\}$

Random variable Y has a **Poisson distribution** with parameter $\mu > 0$ if it takes non-negative integer values $y = 0, 1, 2, ...$, with probability:

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}$$

where $y! = y \times (y-1) \times \cdots \times 1$ and $0! = 1$.

Poisson distribution



Key feature: Mean and variance of Poisson distribution are the same: $\mathbb{E}(Y) = V(Y) = \mu$

Poisson regression

- As with linear regression, we might want to define a model for the mean count in terms of one or more covariates.
- **Question**: What problem might be encountered with the following model for the mean?

$$\mu(X) = \mathbb{E}(Y|\mathbf{X}) = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{X}^\top \boldsymbol{\beta}$$

(As before, $X_0 = 1$ always)

Poisson regression

Similar to what we saw with logistic regression, we have a situation where $\mu(\mathbf{X}) > 0$ but we want to allow the linear predictor to be able to take on values from $-\infty$ to $+\infty$.
Let's use a log transformation!

$$\log[\mu(\mathbf{X})] = \log\{\mathbb{E}(Y|\mathbf{X})\} = \mathbf{X}^\top \beta$$

The full model (in generalized linear model [GLM] format) is:

Mean	$\log[\mu(X)] = X^\top \beta$	(systematic component)
Distribution	$Y\|\mathbf{X} \sim$ Poisson$[\mu(\mathbf{X})]$	(random component)
Link function	$\log$	

- "log" here refers to the **natural logarithm** (sometimes written as "$\ln$" or "$\log_e$").
- The Poisson model is sometimes referred to as a **log-linear model**.
- In generalized linear modeling terminology, we say that the **link function** for this model is "log."

Some generalized linear models

Linear regression

- Y is continuous
- Model

$$\mathbb{E}(Y|\mathbf{X}) = \mathbf{X}^\top\boldsymbol{\beta}$$
$$\mu(\mathbf{X}) = \mathbf{X}^\top\boldsymbol{\beta}$$

- Distributional assumption:

$$Y|\mathbf{X} \sim \mathcal{N}\left[\mu(\mathbf{X}), \sigma^2\right]$$

- Link function: identity

Logistic regression

- Y is binary
- Model

$$\text{logit}\left[\mathbb{E}(Y|\mathbf{X})\right] = \mathbf{X}^\top\boldsymbol{\beta}$$
$$\text{logit}\left[\mu(\mathbf{X})\right] = \mathbf{X}^\top\boldsymbol{\beta}$$

- Distributional assumption:

$$Y|\mathbf{X} \sim \text{Bernoulli}\left[\mu(\mathbf{X})\right]$$

- Link function: logit

Poisson regression

- Y is count
- Model

$$\log\left[\mathbb{E}(Y|\mathbf{X})\right] = \mathbf{X}^\top\boldsymbol{\beta}$$
$$\log\left[\mu(\mathbf{X})\right] = \mathbf{X}^\top\boldsymbol{\beta}$$

- Distributional assumption:

$$Y|\mathbf{X} \sim \text{Poisson}\left[\mu(\mathbf{X})\right]$$

- Link function: log

Note: The systematic part of the model involves some function of the mean of Y given X, $\mathbb{E}\left[Y|\mathbf{X}\right]$, being written in terms of a linear combination of the covariates (linear in the parameters $\boldsymbol{\beta}$). The random part of the model is the distributional assumption.

Maximum likelihood estimation

- Assuming Y_i's are independent, the **likelihood function** in terms of the μ_i's is:

$$L = \prod_{i=1}^{n} \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}$$

- The **log likelihood function** in terms of the μ_i's is:

$$\ell = \sum_{i=1}^{n} \{-\mu_i + y_i \log(\mu_i) - \log(y_i!)\}$$

- Now, using the linear model $\log[\mu_i] = \mathbf{x}_i^\top \boldsymbol{\beta}$ or equivalently $\mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$, we can write the log likelihood in terms of the vector of parameters, $\boldsymbol{\beta}$:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{-\exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + y_i \log(\exp(\mathbf{x}_i^\top \boldsymbol{\beta})) - \log(y_i!)\}$$

- The last term in this expression is sometimes omitted as it is a constant for a given dataset (i.e., it depends only on data values, not on parameters).
- **Maximization** of the likelihood and inferences follow in the usual way (usually done numerically as no analytical solution exists).

Poisson regression model: interpretation

- **Poisson model for mean**:

$$\log[\mu(\mathbf{X})] = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- β_0 is the **log of the mean** when all covariates, X_1 to X_p, take the value zero.
  - As with other models we have considered, this may not have a meaningful interpretation.
- For any $k = 1, \ldots, p$, β_k represents the **difference in the log of the mean** for a unit difference in X_k, holding all other covariates constant.

Poisson regression model: interpretation

- In practice, to facilitate easier interpretation, we often transform back to the original measurement (count) scale.
- So, exponentiating (i.e., "antilog"):

$$\exp\{\log[\mu(\mathbf{X})]\} = \exp\{\mathbf{X}^\top \boldsymbol{\beta}\} = \exp\{\beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}$$

  - And so:

$$\mu(\mathbf{X}) = e^{\beta_0} \times e^{\beta_1 X_1} \times \cdots \times e^{\beta_p X_p}$$

- On the original scale, this is a **multiplicative model**.
- e^{β_0} is the mean count when all covariates, X_1 to X_p, take the value zero.
- Increasing X_k by one unit **multiplies** the mean count by a factor of e^{β_k}, holding all other covariates constant.

Poisson regression model: interpretation

Suppose that our Poisson regression for the mean is

$$\log[\mu(X)] = \beta_0 + \beta_1 X$$

and we want to compare the predicted mean for $X = x$ vs. $X = x + 1$.

- The predicted mean for $X = x$:

$$\mu(x) = \exp(\beta_0 + \beta_1 x)$$

- The predicted mean for $X = x + 1$:

$$\mu(x + 1) = \exp(\beta_0 + \beta_1 (x + 1)) = \exp(\beta_0 + \beta_1 x + \beta_1)$$

- The **ratio** of the two predicted means is:

$$\frac{\mu(x + 1)}{\mu(x)} = \frac{\exp(\beta_0 + \beta_1 x + \beta_1)}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1)$$

- This shows that increasing X by 1 unit **multiplies** the mean count by a factor of $\exp(\beta_1)$

Poisson regression model: interpretation

- Approach to interpretation of parameters (and to estimates of parameters) regarding log mean count in Poisson regression follows that for standard linear regression
    - Same considerations for quantitative, ordinal, binary, and categorical variables
    - Same considerations for statistical interactions/e□ect modification
- Difference is that we have used a log link function, and so generally take exponential of parameters (antilog) to facilitate interpretation in terms of mean counts instead of log mean counts

Interpretation: example

Suppose that we are conducting a study of the n street intersections in a city in order to understand factors associated with the number of accidents at an intersection.

- **Outcome variable** (Y): The number of traffic accidents at a particular intersection in a given year.
- **Covariates of interest**:
  - X_1: Average speed in mph on the nearby streets.
  - X_2: Indicator for whether the intersection has a traffic signal ($X_2 = 1$ if yes, $X_2 = 0$ if no).
- **Possible model**:

$$\begin{aligned} &\text{Mean} && \log[\mu(\mathbf{X})] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ &\text{Distribution} && Y|\mathbf{X} \sim \text{Poisson}[\mu(\mathbf{X})] \end{aligned}$$

- **Parameter estimates**:

$$\hat{\beta}_0 = 2.8, \quad \hat{\beta}_1 = 0.012, \quad \hat{\beta}_2 = -0.20$$

Interpretation: Example

$\hat{\beta}_0 = 2.8$:

- This is the predicted value of $\log(\mu)$ when $X_1 = 0$ and $X_2 = 0$.
- Difficult to interpret, as no street will have an average speed of 0 mph ($X_1 = 0$)

$\hat{\beta}_1 = 0.012$:

- This represents the difference in log mean count for a 1 mph difference in average traffic speed.
- Equivalently, each 1 mph difference in average traffic speed is associated with a multiplicative increase of $e^{0.012} = 1.012$
- This corresponds to a 1.2% increase in the mean count of traffic accidents in a year, adjusted for traffic signals.
- For a 10 mph difference in average speed: $e^{10 \times 0.012} = 1.127$
- This corresponds to a 12.7% higher mean count of traffic accidents per 10 mph higher average traffic speed, adjusted for traffic signals.

$\hat{\beta}_2 = -0.20$:

- The multiplicative factor associated with having a traffic signal at the intersection is: $e^{-0.20} = 0.82$
- This implies an 18% reduction in the mean count of traffic accidents for intersections with the same average traffic speed.

Poisson regression for rates

- Count data often arise when we observe Y events (e.g., asthma attacks) during a time period of length t units of exposure (e.g., years), where each individual may have a different period of follow-up (i.e., different value for t).
- Let us assume the underlying true rate of events is constant over time within individuals and across individuals, with parameter λ (this can be relaxed to depend on an individual's covariate values, i.e., $\lambda(\mathbf{X})$). This is a homogeneity assumption.
- How many events would you expect to occur in time t?
- Assume that the count of events, Y, in time t follows a Poisson distribution with mean λt:

$$P(Y = y|t, \lambda, \mathbf{X}) = P(Y = y|t, \lambda) = \frac{e^{-\lambda t}(\lambda t)^y}{y!}$$

- This is just replacing the mean μ in our earlier definition by λt.
- This is called a **homogeneous Poisson process** – the "homogeneous" refers to the constancy of the rate, λ, over time within an individual.

Poisson regression for rates

- Count data, $Y = y$, in time t follows a Poisson distribution with mean $\lambda(\mathbf{X})t$:

$$P(Y = y|t, \mathbf{X}) = \frac{e^{-\lambda(\mathbf{X})t}(\lambda(\mathbf{X})t)^y}{y!}$$

- So, $Y \sim \text{Poisson}(\lambda(\mathbf{X})t)$, where the rate $\lambda(\mathbf{X})$ depends on covariates $\mathbf{X}$.
- Now we assume a model for the rate parameter $\lambda(\mathbf{X})$ in terms of a vector of covariates $\mathbf{X}$:

$$\log(\lambda(\mathbf{X})) = \mathbf{X}^\top \beta$$

- Noting that the expected count $\mu(X) = \lambda(X)t$, we have:

$$\log(\lambda(\mathbf{X})) = \log\left(\frac{\mu(\mathbf{X})}{t}\right) = \log(\mu(\mathbf{X})) - \log(t) = X^\top \beta$$

- Hence:

$$\log(\mu(\mathbf{X})) = \log(t) + \mathbf{X}^\top \beta$$

- We could also write this as:

$$\log[\mu(t, \mathbf{X})] = \log(t) + \mathbf{X}^\top \beta$$

Poisson regression for rates

Hence we can fit the model for the mean $\log(\mu(X)) = \log(t) + \mathbf{X}^\top \boldsymbol{\beta}$

- $\log(t)$ is called the **offset** in this model.
- The implied model for the rate is: $\log(\lambda(\mathbf{X})) = \log(\mu(\mathbf{X})) - \log(t) = \mathbf{X}^\top \boldsymbol{\beta}$
- Standard software for Poisson regression can be used to fit a model for the rate $\lambda(\mathbf{X})$ in terms of covariates $\mathbf{X}$, by including an offset term $\log(t)$ in the model for counts, with its coefficient forced to be one:

$$\log(\mu(\mathbf{X})) = 1 \cdot \log(t) + \mathbf{X}^\top \boldsymbol{\beta}$$

- Data provided to the software include:
  - Count outcome (y)
  - Covariate vector ($\mathbf{X}$)
  - Offset definition/variable ($\log(t)$)

Example call in R:

```
glm(formula = y ~x1 + x2, family = poisson(link = "log"), offset = log(t))
```

Poisson regression for rates

- We have focused on rates per unit of time (e.g., per year).
- The concept of using an offset is exactly the same for other types of "exposure."
- Examples:
  - In population studies, "exposure" might be the *size of the population* studied in each of multiple locations.
  - In the traffic accident example, "exposure" might be the *number of vehicles* traveling through the intersection, so the rate being modeled is the number of accidents relative to the number of vehicles traveling through the intersection.
  - In spatial applications, exposure might relate to the *size of a geographical area* studied in each of multiple places.
- See Appendix for:
  - Poisson regression when **summary data** are available (e.g., counts of deaths and person-years of follow-up summarized by age and gender).
  - Poisson regression for **standardized rates**.

Appendix (1)

A. Poisson regression applied to summary data

B. Poisson regression for standardized rates

Poisson regression applied to summary data

What if our data are already summarized into groups according to covariate patterns?

Group i	Gender	Age	Person-years of follow-up	Number of deaths due to cancer
1	F	25-44	n_1	d_1
2	M	25-44	n_2	d_2
3	F	45-64	n_3	d_3
4	M	45-64	n_4	d_4
etc.				

Poisson regression applied to summary data

A useful property of the **Poisson distribution** is that the **sum of independent Poisson variables** also has a Poisson distribution.

For example, if Y_1 and Y_2 are independent random variables with $Y_1 \sim \text{Poisson}(\mu_1)$ and $Y_2 \sim \text{Poisson}(\mu_2)$, then $Y_1 + Y_2 \sim \text{Poisson}(\mu_1 + \mu_2)$

More generally, if $Y_1, \ldots, Y_m$ are independent random variables with $Y_j \sim \text{Poisson}(\mu_j)$, then $Y_1 + \cdots + Y_m \sim \text{Poisson}\left(\sum_{j=1}^{m} \mu_j\right)$

Application to grouped data:

- Suppose we have grouped data where a group is defined by a particular covariate pattern (e.g., gender and age).
- Let Y_{ij} be the count of events experienced by the j-th individual in the i-th group, for $j = 1, \ldots, m_i$, and let the group count be $Y_i = \sum_{j=1}^{m_i} Y_{ij}$
- Assuming independence of the Y_{ij}'s, if each $Y_{ij} \sim \text{Poisson}(\mu_i)$, then the group count $Y_i \sim \text{Poisson}(m_i \mu_i)$

Poisson regression applied to summary data

- From the previous slide, we have group counts following a Poisson distribution.
- We assumed that the count for each individual in the group followed a Poisson distribution with the same mean, a homogeneity assumption.
- The observed group count in our example is the number of cancer deaths, d_i.
- How to deal with different person-years of follow-up, n_i, in each group?
- Define a model for the rate of cancer deaths per person-year:

$$\log(\lambda_i) = X_i^\top \beta$$

- So that:

$$\log(\lambda_i) = \log\left(\frac{\mu_i}{n_i}\right) = \log(\mu_i) - \log(n_i) = X_i^\top \beta$$

- Hence, fit a model with an offset term:

$$\log(\mu_i) = \log(n_i) + X_i^\top \beta$$

- Data for group i provided to the software are:
  - Group count for the outcome, d_i,
  - Covariate vector, X_i,
  - Offset definition/variable, $\log(n_i)$.

Poisson Regression Applied to Summary Data: Example

Hypoglycemia in the Diabetes Control and Complications Trial

- Randomized trial of "intensive" versus "conventional" intervention.
- Outcome (adverse event): Hypoglycemia episode.
- Group data:

Group (i)	Sample Size	Person-years (PY)	Events	$\hat{\lambda}_i$	Rate per 100 PY
Intensive (1)	363	2598.5	1723	0.6631	66.3
Conventional (0)	352	2480.2	543	0.2189	21.9

- **Fit a Poisson model** for the count of hypoglycemia episodes, with an indicator variable for intervention as the only covariate, and an **offset term** equal to person-years of follow-up.
- **Estimate**: $\hat{\beta}_1 = 1.108$ with $SE = 0.0492$
- **Exponentiating**: Incidence rate ratio (IRR) is: $\exp(1.108) = 3.03$
- **Asymptotic 95% CI** for $\hat{\beta}_1$ is $1.108 \pm 1.96 \times 0.0492 = 1.0116$ to 1.2044
- **Exponentiating confidence bounds** gives a 95% CI for the IRR: 2.75 to 3.33

(Source: Lachin, J (2000) *Biostatistical Methods.* Wiley, New York. Page 322).

Poisson regression for standardized rates

- Poisson regression can be used to model **Standardized Incidence Ratios (SIRs)** or **Standardized Mortality Ratios (SMRs)**.
- Example: Modeling the rate of cancer deaths in counties in the United States.
- We know that cancer rates vary dramatically by age, so we decide to use **indirect standardization**.
- Given a set of **age-specific reference rates** for the U.S., R_j for age groups $j = 1, \dots, J$, we can compute the **expected count** of cancer deaths in county i (assuming the U.S. rates apply in every county), based on the count of age-specific population at risk in each age group in the county, n_{ij}:

$$E_i = \sum_{j=1}^{J} (n_{ij} R_j)$$

That is, we apply the national age-specific reference rates to the age-specific population in county i to compute E_i: the **expected number of cancer deaths** in county i, adjusting for the age distribution.

Poisson regression for standardized rates

- From the previous slide, the **expected count** for county i is: $E_i = \sum_{j=1}^{J} (n_{ij} R_j)$
- The **observed count** of cancer deaths for county i is: $O_i = \sum_{j=1}^{J} y_{ij}$
- Let θ_i be the **Standardized Mortality Ratio (SMR)** for county i.
- Then considering E_i as the "exposure," we can write the model as:

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\theta_i E_i) \\
\log(\theta_i) &= \mathbf{x}_i^\top \boldsymbol{\beta} \\
\log(\mu_i) &= \log(E_i) + \mathbf{x}_i^\top \boldsymbol{\beta}
\end{aligned}
$$

- i.e. we can model the expected mean of the county of cancer deaths using a Poisson model with $\log(E_i)$ as an offset.
- This approach is very common in spatial modeling of small-area disease rates.

Appendix (2):

GLMs: Some more technical detail for those interested

(details other than what is presented in the main slides above will not appear on any exam)

More detail on GLMs

A generalized linear model has three components:

Random Component: Outcome random variable, Y_i, has a distribution in the **exponential family**, taking the form:

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

- Here, θ_i and ϕ are parameters, and $a_i(\phi)$, $b(\theta_i)$, and $c(y_i, \phi)$ are known functions. - We denote the mean of Y_i as $E(Y_i) = \mu_i$.

Systematic Component: Covariates are given by the vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$, producing a linear predictor η_i given by:

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$

Link Function: This function links the random and systematic components. It is a one-to-one continuous differentiable function, $g(\cdot)$:

$$\eta_i = g(\mu_i)$$

Some generalized linear models

Linear regression

- Y is continuous
- Model

$$\mathbb{E}(Y|\mathbf{X}) = \mathbf{X}^\top\boldsymbol{\beta}$$
$$\mu(\mathbf{X}) = \mathbf{X}^\top\boldsymbol{\beta}$$

- Distributional assumption:

$$Y|\mathbf{X} \sim \mathcal{N}\left[\mu(\mathbf{X}), \sigma^2\right]$$

- Link function: identity

Logistic regression

- Y is binary
- Model

$$\text{logit}\left[\mathbb{E}(Y|\mathbf{X})\right] = \mathbf{X}^\top\boldsymbol{\beta}$$
$$\text{logit}\left[\mu(\mathbf{X})\right] = \mathbf{X}^\top\boldsymbol{\beta}$$

- Distributional assumption:

$$Y|\mathbf{X} \sim \text{Bernoulli}\left[\mu(\mathbf{X})\right]$$

- Link function: logit

Poisson regression

- Y is count
- Model

$$\log\left[\mathbb{E}(Y|\mathbf{X})\right] = \mathbf{X}^\top\boldsymbol{\beta}$$
$$\log\left[\mu(\mathbf{X})\right] = \mathbf{X}^\top\boldsymbol{\beta}$$

- Distributional assumption:

$$Y|\mathbf{X} \sim \text{Poisson}\left[\mu(\mathbf{X})\right]$$

- Link function: log

Note: The systematic part of the model involves some function of the mean of Y given X, $\mathbb{E}\left[Y|\mathbf{X}\right]$, being written in terms of a linear combination of the covariates (linear in the parameters $\boldsymbol{\beta}$). The random part of the model is the distributional assumption.

Systematic component and link function in GLMs we have met

Systematic component: Covariates given by the vector $\mathbf{x}_i = (x_{1i}, \dots, x_{pi})^\top$ produce a **linear predictor**, η_i, given by $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$

Link function: This function links the random and systematic components. It is a one-to-one continuous differentiable function, $g(\cdot) : \eta_i = g(\mu_i)$. If the link function transforms us onto the scale of the linear predictor, then the inverse link function $g^{-1}(\cdot)$ takes us back to the scale of the mean: $\mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$.

	Systematic	Link $g(\cdot)$	Systematic + link	Inverse link
Linear	$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$	identity $\rightarrow \eta_i = \mu_i$	$\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$	$\mu_i = \eta_i \rightarrow \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$
Poisson	$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$	log $\rightarrow \eta_i = \log(\mu_i)$	$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$	$\mu_i = \exp(\eta_i) \rightarrow \mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$
Logistic	$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$	logit $\rightarrow \eta_i = \text{logit}(\mu_i)$	$\text{logit}(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$	$\mu_i = \text{expit}(\eta_i) \rightarrow \mu_i = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}$

GLMs: Random component

Random component: Outcome random variable, Y_i, has a distribution in the exponential family, taking the general form

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

where θ_i and ϕ are parameters, and $a_i(\phi)$, $b_i(\theta_i)$, and $c(y_i, \phi)$ are known functions. We denote the mean of Y_i as $\mathbb{E}(Y_i) = \mu_i$.

> We can write many important distributions in the above form, and it turns out that the normal (Gaussian), Bernoulli, and Poisson distributions all belong to the exponential family.

The parameter θ_i generally tells us something about the location of the distribution. θ_i is called the natural or canonical parameter

The parameter ϕ generally tells us something about scale/dispersion

- e.g. for standard linear regression, besides parameter $\boldsymbol{\beta}$, we have an additional variance (or dispersion) parameter, σ^2

In all of the models we've considered, the function $a_i(\phi)$ has the form $a_i(\phi) = \phi / p_i$, where p_i is a known *prior weight*, usually 1.

GLMs: Random component

It can be confusing to try to see how the pdf's of the binomial, Poisson, and normal distributions share the same general structure as the expression above.

Let's work through each distribution and try to see the connections. (**This is purely optional material and you are not required to know this for any exam!**).

Binomial (Bernoulli) distribution

The binomial pdf is

$$f_i(y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

(Remember that the Bernoulli is a special case of the Bernoulli where the number of trials is 1.)

Taking logs, we obtain

$$\log\left[f_i(y_i)\right] = y_i \log\left(\pi_i\right) + \left(n_i - y_i\right) \log(1-\pi_i) + \log \binom{n_i}{y_i}$$

Collecting terms on y_i we can write

$$\log\left[f_i(y_i)\right] = y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log \binom{n_i}{y_i}$$

Looking at the first part of this expression, we can see that the canonical parameter is

$$\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

and the canonical link is the logit transformation.

Binomial (Bernoulli) distribution

Solving for π_i, we see that

$$\pi_i = \frac{e^{\theta_i}}{1 + e^{\theta_i}}, \qquad \text{so} \qquad 1 - \pi_i = \frac{1}{1 + e^{\theta_i}}.$$

If we write the second term in the pdf as a function of θ_i, $\log(1 - \pi_i) = -\log(1 + e^{\theta_i})$, then we see that the cumulant function is

$$b(\theta_i) = n_i \log(1 + e^{\theta_i}).$$

The remaining term in the pdf is a function of y_i but not π_i, so

$$c(y_i, \phi) = \log \binom{n_i}{y_i}.$$

We set $a_i(\phi) = \phi$ and $\phi = 1$.

Binomial (Bernoulli) distribution

To verify the mean, we can follow the directions above for obtaining the mean and variance

$$\mathbb{E}(Y_i) = \mu_i = b'(\theta_i)$$
$$\text{Var}(Y_i) = \sigma_i^2 = b''(\theta_i)a_i(\phi),$$

so we differentiate $b(\theta_i)$ with respect to θ_i and obtain

$$\mathbb{E}(Y_i) = b'(\theta_i) = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} = n_i\pi_i \qquad \leftarrow \text{this is the mean of a binomial random variable}$$

Differentiating again using the quotient rule, we find that

$$\text{Var}(Y_i) = a_i(\phi)b''(\theta_i) = n_i \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = n_i\pi_i(1-\pi_i) \quad \leftarrow \text{this is the variance of a binomial r.v.}$$

So yes, we have shown that the binomial distribution is part of the exponential family!

Poisson distribution

Recall that a Poisson random variable has the probability distribution function

$$f_i(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

for $y_i \in 0, 1, 2, \ldots$. The mean and variance are

$$\mathbb{E}(Y_i) = \text{Var}(Y_i) = \mu_i.$$

Taking the log of $f_i(y_i)$, we obtain

$$\log f_i(y_i) = y_i \log(\mu_i) - \mu_i - \log(y_i!).$$

Looking at the first term in the log pdf, we can see that the canonical parameter is

$$\theta_i = \log(\mu_i)$$

and the canonical link is the log.

Poisson distribution

The second term, μ_i, can be written as

$$b(\phi_i) = e^{\theta_i}.$$

The last term is a function of y_i only, so

$$c(y_i, \phi) = -\log(y_i!).$$

Finally, note that we can take $a_i(\phi) = \phi$ and $\phi = 1$.

Let's verify the mean and variance. Differentiating the cumulant function, $b(\theta_i)$, we have

$$\mathbb{E}(Y_i) = b'(\theta_i) = e^{\theta_i} = \mu_i \qquad \leftarrow \text{this is the mean of a Poisson random variable}$$

Differentiating again, we obtain

$$\text{Var}(Y_i) = a_i(\phi)b''(\theta_i) = e^{\theta_i} = \mu_i \qquad \leftarrow \text{this is the variance of a Poisson random variable}$$

As we already knew, the mean equals the variance. So yes, the Poisson distribution also belongs to the exponential family!

Normal distribution

The normal distribution has density

$$
\begin{aligned}
f(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i - \mu_i)^2}{\sigma^2} \right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i^2 + \mu_i^2 - 2y_i\mu_i)}{\sigma^2} \right\} \\
&= \exp\left\{ \frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right\}
\end{aligned}
$$

From this, we can identify θ_i as μ_i and ϕ as σ^2 with $a_i(\phi) = \phi$.

$$
b(\theta_i) = \frac{1}{2}\theta_i^2, \qquad c(y_i, \phi) = -\frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)
$$

Verifying the mean and variance,

$$
\begin{aligned}
\mathbb{E}(Y_i) &= b'(\theta_i) = \theta_i = \mu_i && \leftarrow \text{this is the mean of a Normal random variable} \\
\text{Var} Y_i &= b"(\theta_i) a_i(\phi) = \sigma^2 && \leftarrow \text{this is the variance of a Normal random variable}
\end{aligned}
$$

So yes, the normal distribution is also part of the exponential family!

Canonical link functions

- Normal distribution: $\theta_i = \mu_i$
- Poisson distribution: $\theta_i = \log(\mu_i)$
- Bernoulli/binomial distribution: $\theta_i = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$
- We call θ_i the **natural or canonical parameter**. Statistically, it is an easier parameter to work with in the general distribution expression for the exponential family than the mean (though they are the same for the normal distribution).
- The **link function** that makes the linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ the same as the canonical parameter θ_i is called the **natural or canonical link function**.
- For the Bernoulli/binomial, this is one reason why we might prefer the **logit link function** over other possible link functions (e.g., probit).
- While using the canonical link function may be statistically convenient, it's important to check that its use with the linear predictor provides a good fit to the data

Parameter estimation for GLMs

- Several seemingly different models can be written in a standard GLM format, involving:
  - The **exponential family** of distributions.
  - A **linear predictor** of covariates.
  - A **link** function.
- It is also possible to incorporate differential weighting of observations.
- Parameter estimates are obtained using maximum likelihood estimation applied to this standard GLM, leading to asymptotic inferences based on likelihood theory.
  - In general, numerical methods are needed for estimation.
  - The same numerical algorithm can be applied to fit any of the seemingly different models.
  - The standard algorithm originally used is known as **"iteratively reweighted least squares"** (IRLS) (essentially an extension of **ordinary least squares** (OLS) and **weighted least squares** (WLS)).
  - Inferences only depend on the **first and second moments** of the distribution (relates to the method of moments estimation).
  - Extensions of GLMs: GLMs can be extended to distributions specified using just the first and second moments, known as **quasi-likelihood**.