Generalized Linear Models (GLMs)

GENERAL ASSEMBLY

Learning Objectives

- Understand OLS from a new perspective
- Understand logistic regression from a new perspective
- Define the three components of a generalized linear model:
  - Systematic component
  - Random component
  - Link component
- Two new models: Poisson regression, gamma regression
- Interpreting their coefficients

Prelude I: “Linearity”

Aside: Linearity

The term **linear** is often misunderstood. Is the following model linear?

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Aside: Linearity

The term **linear** is often misunderstood. Is the following model linear?

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Yes!

Aside: Linearity

The term **linear** is often misunderstood. Is the following model linear?

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Yes!

Why? Because it can be written:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= \mathbf{x}_i^T \beta \end{aligned}$$

$x_2 = x_1^2$

Aside: Linearity

The term **linear** is often misunderstood. Is the following model linear?

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Yes!

Why? Because it can be written:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$\boxed{= \mathbf{x}_i^T \beta}$$

$x_2 = x_1^2$

If a model can be written like this, it is linear!

Aside: Linearity

So then, what ***isn't*** a linear model? Anything that can't be written as in the previous slide. For example:

$$Y_i = \frac{\beta_1 x_1}{\beta_2 x_2 + \beta_3 x_3}$$

Prelude II: Prediction vs. Inference

Prediction vs. Inference

Statistical modeling (aka *machine learning*) has two primary goals:

Prediction

"I care most about estimating

$$\hat{y}$$

as accurately as possible."

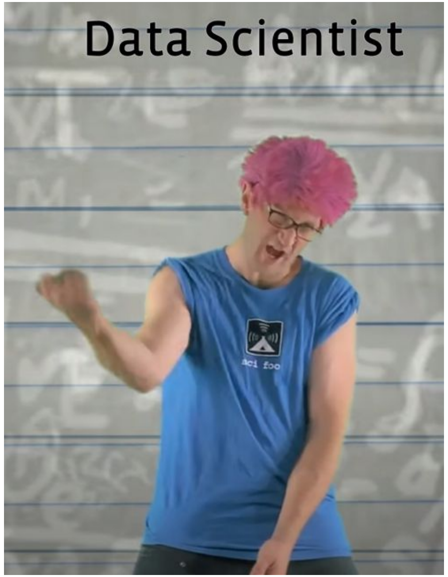
Inference

"I care most about estimating

$$\hat{\beta}$$

as accurately as possible."

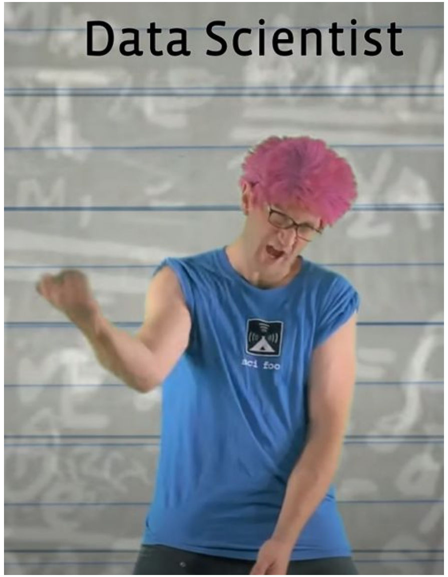
Prediction



Inference

Source: https://www.youtube.com/watch?v=uHGlCi9jOWY

Prediction



Data science, we predict the outcome
And if you see it’s workin’
No need to question “how come”

Source: https://www.youtube.com/watch?v=uHGlCi9jOWY

Inference

Statisticians - we want it precise!
How is it "science"
If you can't explain your model designs?

Source: https://www.youtube.com/watch?v=uHGlCi9jOWY

Inference

Statisticians - we reveal how nature functions
All I need is a model with articulated assumptions!

Source: https://www.youtube.com/watch?v=uHGlCi9jOWY

But really....

Source: https://www.youtube.com/watch?v=uHGlCi9jOWY

They're both important. Our priorities just shift from project to project.

This lesson will focus itself on **inference**. These models won't be amazing at prediction.

A closer look at OLS

A closer look at OLS

Our parametrization of OLS has been:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \\
&= \mathbf{x}_i^T \beta + \varepsilon_i
\end{aligned}
$$

where

$$\varepsilon_i \sim N(0, \sigma)$$

A closer look at OLS

Our parametrization of OLS has been:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \\
&= \mathbf{x}_i^T \beta + \varepsilon_i
\end{aligned}
$$

$$\mathbf{x}_i^T = \begin{bmatrix} 1 & x_{1i} & \cdots & x_{pi} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

where

$$\varepsilon_i \sim N(0, \sigma)$$

A closer look at OLS

Our parametrization of OLS has been:

$$
\begin{aligned}
Y_i &= \overset{L}{\boxed{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i}} \\
&= \mathbf{x}_i^T \beta + \varepsilon_i
\end{aligned}
$$

where

$$\varepsilon_i \sim N(0, \sigma)$$

A closer look at OLS

Our parametrization of OLS has been:

$$\begin{aligned} Y_i &= \overset{\mathbf{L}}{\boxed{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i}} \\ &= \mathbf{x}_i^T \beta + \varepsilon_i \end{aligned}$$

where

$$\varepsilon_i \overset{\mathbf{I}}{\boxed{\sim}} N(0, \sigma)$$

A closer look at OLS

Our parametrization of OLS has been:

$$\begin{aligned} Y_i &= \overset{\textbf{L}}{\boxed{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i}} \\ &= \mathbf{x}_i^T \beta + \varepsilon_i \end{aligned}$$

where

$$\varepsilon_i \overset{\textbf{I}}{\boxed{\sim}} \overset{\textbf{N}}{\boxed{N}}(0, \sigma)$$

A closer look at OLS

Our parametrization of OLS has been:

$$
\begin{aligned}
Y_i &= \overset{\textbf{L}}{\boxed{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i}} \\
&= \mathbf{x}_i^T \beta + \varepsilon_i
\end{aligned}
$$

where

$$
\varepsilon_i \overset{\textbf{I}}{\boxed{\sim}} \overset{\textbf{N}}{\boxed{N}}(0, \overset{\textbf{E}}{\boxed{\sigma}})
$$

A closer look at OLS

It turns out our LINE assumptions don't just come from nowhere. They're a consequence of how we parametrize OLS! What's more, is that we can condense this:

A closer look at OLS

It turns out our LINE assumptions don't just come from nowhere. They're a consequence of how we parametrize OLS! What's more, is that we can condense this:

$$Y_i = \underbrace{\mathbf{x}_i^T}_{\text{const}} \beta + \underbrace{\varepsilon_i}_{\text{rv}}$$

A closer look at OLS

It turns out our LINE assumptions don't just come from nowhere. They're a consequence of how we parametrize OLS! What's more, is that we can condense this:

$$Y_i = \underbrace{\mathbf{x}_i^T \beta}_{\text{const}} + \underbrace{\varepsilon_i}_{\text{rv}}$$

$$\implies \quad Y_i \sim N(\mathbf{x}_i^T \beta, \sigma)$$

A closer look at OLS

$$Y_i \sim N(\mathbf{x}_i^T \beta, \sigma)$$

Finally, I'm going to split this formula up just a little. Let's store this away for later:

Blue Box #1

$$\boxed{\begin{aligned} Y_i &\sim N(\mu_i, \sigma) \\ \mu_i &= \mathbf{x}_i^T \beta \end{aligned}}$$

A closer look at logistic regression

A closer look at logistic regression

$$\log \frac{p_i}{1-p_i} = \operatorname{logit} p_i = \mathbf{x}_i^T \beta$$

Why is there no epsilon error term shown here?

A closer look at logistic regression

$$\log \frac{p_i}{1-p_i} = \operatorname{logit} p_i = \mathbf{x}_i^T \beta$$

Y_i is not depicted here. What distribution do we implicitly assume it follows?

A closer look at logistic regression

Blue Box #2

$$\begin{aligned} Y_i &\sim \mathrm{Ber}\left(p_i\right) \\ \operatorname{logit} p_i &= \mathbf{x}_i^T \beta \end{aligned}$$

The Generalized Linear Model

What do these blue boxes have in common?

Blue Box #1

$$\boxed{\begin{aligned} Y_i &\sim N(\mu_i, \sigma) \\ \mu_i &= \mathbf{x}_i^T \beta \end{aligned}}$$

Blue Box #2

$$\boxed{\begin{aligned} Y_i &\sim \mathrm{Ber}\,(p_i) \\ \mathrm{logit}\, p_i &= \mathbf{x}_i^T \beta \end{aligned}}$$

What do these blue boxes have in common?

Blue Box #1

$$Y_i \sim N(\mu_i, \sigma)$$
$$\mu_i = \mathbf{x}_i^T \beta$$

Blue Box #2

$$Y_i \sim \mathrm{Ber}(p_i)$$
$$\mathrm{logit}\, p_i = \mathbf{x}_i^T \beta$$

Linear regression and logistic regression couldn't be more different. One's regression, the other's classification! However, they're both linear models. Actually, they're both **generalized linear models (GLMs)**.

Anatomy of a GLM

GLMs have three components:

- **Systematic component** (or **linear component**) - This is the linear part of the model, the part we're already familiar with.

Anatomy of a GLM

GLMs have three components:

- **Systematic component** (or **linear component**) - This is the linear part of the model, the part we're already familiar with.
- **Random component** - The distributional assumption of our target variable. Where does the randomness in our model come from? What is the distribution of Y?

Anatomy of a GLM

GLMs have three components:

- **Systematic component** (or **linear component**) - This is the linear part of the model, the part we're already familiar with.
- **Random component** - The distributional assumption of our target variable. Where does the randomness in our model come from? What is the distribution of Y?
- **Link function** - The function that maps values of μ_i to the real numbers. This ensures we don't predict anything impossible (like probabilities outside 0 and 1).

Anatomy of a GLM

Blue Box #2

$$
\begin{aligned}
Y_i &\sim \mathrm{Ber}(p_i) \\
\operatorname{logit} p_i &= \mathbf{x}_i^T \beta
\end{aligned}
$$

What is the random component?

What is the link function?

Anatomy of a GLM

We assume that Y_i follows a Bernoulli distribution. That is, it is randomly either 0 or 1 with some probability p_i

Blue Box #2

$$
\begin{aligned}
Y_i &\sim \boxed{\mathrm{Ber}}(p_i) \\
\operatorname{logit} p_i &= \mathbf{x}_i^T \beta
\end{aligned}
$$

Anatomy of a GLM

Blue Box #2

$$
\begin{aligned}
Y_i &\sim \mathrm{Ber}\left(p_i\right) \\
\boxed{\mathrm{logit}}\, p_i &= \mathbf{x}_i^T \beta
\end{aligned}
$$

We usually use the logit link function for logistic regression. It takes numbers between 0 and 1 and maps them onto the whole real line.

Anatomy of a GLM

Blue Box #1

$$\boxed{\begin{aligned} Y_i &\sim N(\mu_i, \sigma) \\ \mu_i &= \mathbf{x}_i^T \beta \end{aligned}}$$

What is the random component?

What is the link function?

Anatomy of a GLM

We make strong normality assumptions when conducting OLS!

Blue Box #1

$$Y_i \sim \boxed{N}(\mu_i\,,\sigma)$$
$$\mu_i = \mathbf{x}_i^T\beta$$

Anatomy of a GLM

Blue Box #1

$$Y_i \sim N(\mu_i, \sigma)$$
$$\boxed{\mu_i} = \mathbf{x}_i^T \beta$$

Since the normal distribution can already take on any real number, we use the "identity link". That is, the link that maps a number to itself. Another way of saying "the function that does nothing".

The GLM

$$\boxed{\begin{aligned} Y_i &\sim F(\mu_i) \\ g(\mu_i) &= \mathbf{x}_i^T \beta \end{aligned}}$$

All **GLMs** have these three components.

We can invent **new** GLMs simply by swapping out some of these components!

For example...

Probit regression

$$\boxed{\begin{aligned} Y_i &\sim \mathrm{Ber}\left(p_i\right) \\ \Phi^{-1}\left(p_i\right) &= \mathbf{x}_i^T \beta \end{aligned}}$$

Instead of the logit link function, we can use the **probit** link function. This is a slight variation on logistic regression.

The **probit function** is the inverse standard normal CDF. It takes in a percentile and outputs where on the normal distribution it lies.

Probit regression

$$Y_i \sim \mathrm{Ber}(p_i)$$
$$\Phi^{-1}(p_i) = \mathbf{x}_i^T \beta$$

The probit function in practice looks *really* similar to the logit function. So why use this?

It's for **interpretability**. The logit link function allows us to interpret our results in terms of **log odds**. Lots of fields, especially medical ones, like log odds.

Some fields, such as **economics** and **education**, prefer working in **percentiles**.

Some New Models

What distribution might these variables follow?

- How many students will enroll in DSI next cohort?
- How many puppies will be in the next litter?
- How many lightbulbs will we go through in the bathroom next year?
- How many sales will my website generate tomorrow?

What distribution might these variables follow?

- How many students will enroll in DSI next cohort?
- How many puppies will be in the next litter?
- How many lightbulbs will we go through in the bathroom next year?
- How many sales will my website generate tomorrow?

https://www.nndb.com/people/857/000093578/

Poisson!

These are all count data. Count data are typically modeled well with the Poisson distribution

What’s a good link function for this situation?

$$\boxed{\begin{aligned} Y_i &\sim \mathrm{Poi}\,(\mu_i) \\ g(\mu_i) &= \mathbf{x}_i^T\beta \end{aligned}}$$

What's a good link function for this situation?

Blue Box #3: Poisson Regression

$$\boxed{\begin{aligned} Y_i &\sim \mathrm{Poi}\left(\mu_i\right) \\ \log \mu_i &= \mathbf{x}_i^T \beta \end{aligned}}$$

The **log link** is an excellent choice for Poisson regression, since it maps the possible values of μ_i (positive numbers) onto the real line.

Another popular choice is the **square root link**. Even though it doesn't map to the full real line, it has some nice theoretical properties that some people enjoy. It's not as interpretable though, so we'll stay away from it.

Let’s try it out!

Interpreting Poisson coefficients

Remember how we did this for OLS:

$$
\begin{aligned}
\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1(x+1) \\
&= (\hat{\beta}_0 + \hat{\beta}_1 x) + \hat{\beta}_1 \\
&= \hat{y} + \hat{\beta}_1
\end{aligned}
$$

Interpreting Poisson coefficients

Remember how we did this for OLS:

For a one-unit increase in x...

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \boxed{(x+1)}$$
$$= (\hat{\beta}_0 + \hat{\beta}_1 x) + \hat{\beta}_1$$
$$= \boxed{\hat{y} + \hat{\beta}_1}$$

We expect *y* to increase by β_1

Interpreting Poisson coefficients

$$
\begin{aligned}
\log \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 (x + 1) \\
\hat{y} &= \exp\left(\hat{\beta}_0 + \hat{\beta}_1 (x + 1)\right) \\
&= \exp\left((\hat{\beta}_0 + \hat{\beta}_1 x) + \hat{\beta}_1\right) \\
&= \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \exp\left(\hat{\beta}_1\right) \\
&= \hat{y} \cdot \exp\left(\hat{\beta}_1\right)
\end{aligned}
$$

Interpreting Poisson coefficients

For a one-unit increase in x...

$$\begin{aligned}
\log \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 \boxed{(x+1)} \\
\hat{y} &= \exp\left(\hat{\beta}_0 + \hat{\beta}_1 (x+1)\right) \\
&= \exp\left((\hat{\beta}_0 + \hat{\beta}_1 x) + \hat{\beta}_1\right) \\
&= \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \exp\left(\hat{\beta}_1\right) \\
&= \hat{y} \cdot \boxed{\exp\left(\hat{\beta}_1\right)}
\end{aligned}$$

We expect *y* to increase by a *factor* of exp(β_1)

What does this mean...?

$$\hat{\beta}_1 > 0 \implies e^{\hat{\beta}_1} > 1$$

$$\hat{\beta}_1 = 0 \implies e^{\hat{\beta}_1} = 1$$

$$\hat{\beta}_1 < 0 \implies e^{\hat{\beta}_1} < 1$$

What does this mean...?

$$\hat{\beta}_1 > 0 \Longrightarrow e^{\hat{\beta}_1} > 1$$

$$\hat{\beta}_1 = 0 \Longrightarrow e^{\hat{\beta}_1} = 1$$

$$\hat{\beta}_1 < 0 \Longrightarrow e^{\hat{\beta}_1} < 1$$

We're multiplying by a number greater than 1. Larger *x* means things are increasing by a multiplicative factor

What does this mean...?

$$\hat{\beta}_1 > 0 \implies e^{\hat{\beta}_1} > 1$$

$$\hat{\beta}_1 = 0 \implies e^{\hat{\beta}_1} = 1$$

$$\hat{\beta}_1 < 0 \implies e^{\hat{\beta}_1} < 1$$

We're multiplying by a number less than 1. Larger *x* means things are decreasing by a multiplicative factor

What does this mean...?

$$\hat{\beta}_1 > 0 \implies e^{\hat{\beta}_1} > 1$$

$$\hat{\beta}_1 = 0 \implies e^{\hat{\beta}_1} = 1$$

$$\hat{\beta}_1 < 0 \implies e^{\hat{\beta}_1} < 1$$

We're multiplying by 1. Things are not changing with respect to *x*.

Let’s try it out!

What distribution might these variables follow?

- How long until this light bulb goes out?
- How long will this rain last?
- How long until the next economic downturn?
- How long until this user unsubscribes from our service?

What distribution might these variables follow?

- How long until this light bulb goes out?
- How long will this rain last?
- How long until the next economic downturn?
- How long until this user unsubscribes from our service?

Γ

Gamma!

These are all waiting-time variables. The gamma (a more general exponential distribution) is often used to measure such values.

Gamma regression

Blue Box #4: Gamma Regression

$$Y_i \sim \text{Gamma}\left(\mu_i\,, \nu\right)$$
$$\log \mu_i = \mathbf{x}_i^T \beta$$

The **log link** is an excellent choice for gamma regression since it is interpretable.

Another popular choice is the **negative inverse link** which usually gets better predictive results, but is not interpretable.

Gamma regression

Blue Box #4: Gamma Regression

$$\boxed{\begin{aligned} Y_i &\sim \mathrm{Gamma}\boxed{(\mu_i\,,\nu)} \\ \log \mu_i &= \mathbf{x}_i^T \beta \end{aligned}}$$

Side note: This is an all-new parametrization of the gamma distribution that we've never seen before. This parametrization is *only* used in the context of GLMs since it gets them to work. Just know that μ_i is the mean and that ν is the **dispersion parameter**, sort of like a measurement of variance.

Let’s try it out!