# Lecture 15: Model Mis-specification in Generalized Linear Models

Jacob Steinhardt

March 9, 2021

# Overview

So far, looked at issues at **training time**: what happens if data is corrupted.

Now will switch focus: to statistical inferences (e.g. uncertainty estimates or causal estimates).

- In particular, how are things inferences affected by model mis-specification?

This lecture: generalized linear models (GLMs)

- Introduce and review classical uncertainty estimates
- Show these can go very wrong (COVID-19 case study)
- Discuss how to fix

# Review: Classification and Regression

Observe data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$

- or $y^{(i)} \in \{0, 1\}$, $\mathbb{N}$, etc.

Minimize loss function $L(\beta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; \beta)$

Example:

- $\ell(x, y; \beta) = (y - \beta^\top x)^2$ (least squares regression)
- $\ell(x, y; \beta) = \log(1 + \exp((-1)^y \beta^\top x))$ (logistic regression)
- Other examples? What about count data?

# Generalized Linear Models

Observe data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

Model $y \mid x, \beta$ has two parts:

- Prediction of mean via *link function*: $\mathbb{E}[y \mid x] = g(\beta^\top x)$
- Exponential family $F(y \mid \mu)$ with mean $\mu$:
  - $y \sim N(\mu, 1)$ (regression)
  - $y \sim \mathrm{Bernoulli}(\mu)$ (classification)
  - $y \sim \mathrm{Poisson}(\mu)$ (count data)

# Generalized Linear Models

Observe data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

Model $y \mid x, \beta$ has two parts:

- Prediction of mean via *link function*: $\mathbb{E}[y \mid x] = g(\beta^\top x)$
- Exponential family $F(y \mid \mu)$ with mean $\mu$:
  - $y \sim N(\mu, 1)$ (regression)
  - $y \sim \mathrm{Bernoulli}(\mu)$ (classification)
  - $y \sim \mathrm{Poisson}(\mu)$ (count data)

Link function $g$ can be arbitrary but often canonical:

- $F = N(\mu, 1)$, $g(z) = z$
- $F = \mathrm{Bernoulli}(\mu)$, $g(z) = \frac{1}{1+\exp(-z)}$
- $F = \mathrm{Poisson}(\mu)$, $g(z) = \exp(z)$

# Example: Poisson

Poisson likelihood, exponential link:

$$
\begin{aligned}
p(y \mid x, \beta) &= \mathrm{Poisson}(y; \exp(\beta^\top x)) \\
&= \exp(-\exp(\beta^\top x)) \exp(\beta^\top x)^y / y! \\
&\propto \exp(y\beta^\top x - \exp(\beta^\top x))
\end{aligned}
$$

Log-likelihood (up to constants):

$$
L(y \mid x, \beta) = \sum_{i=1}^{n} y^{(i)} \beta^\top x^{(i)} - \exp(\beta^\top x^{(i)}).
$$

# Example: Poisson

Poisson likelihood, exponential link:

$$\begin{aligned} p(y \mid x, \beta) &= \mathrm{Poisson}(y; \exp(\beta^\top x)) \\ &= \exp(-\exp(\beta^\top x)) \exp(\beta^\top x)^y / y! \\ &\propto \exp(y\beta^\top x - \exp(\beta^\top x)) \end{aligned}$$

Log-likelihood (up to constants):

$$L(y \mid x, \beta) = \sum_{i=1}^n y^{(i)} \beta^\top x^{(i)} - \exp(\beta^\top x^{(i)}).$$

MLE ($\nabla L = 0$): predicted expectation equals empirical expectation:

$$\sum_{i=1}^n x^{(i)} y^{(i)} = \sum_{i=1}^n x^{(i)} \exp(\beta^\top x^{(i)})$$

# Linear regression on COVID-19 data

Count data: $y^{(t)}$ is number of COVID-19 cases on day $t$.

Assuming exponential growth, $\mathbb{E}[y^{(t)}] = \exp(\beta_0 + \beta_1 t)$ (Poisson with exponential link function)

Can implement using `statsmodels` package.

[Jupyter demo]

# What Went Wrong?

Recall form of log-likelihood:

$$L(y \mid x, \beta) = \sum_{i=1}^{n} y^{(i)} \beta^\top x^{(i)} - \exp(\beta^\top x^{(i)})$$

$$\nabla L(y \mid x, \beta) = \sum_{i=1}^{n} y^{(i)} x^{(i)} - \exp(\beta^\top x^{(i)}) x^{(i)}$$

$$\nabla^2 L(y \mid x, \beta) = -\sum_{i=1}^{n} \exp(\beta^\top x^{(i)}) (x^{(i)} x^{(i)})^\top$$

# What Went Wrong?

Recall form of log-likelihood:

$$L(y \mid x, \beta) = \sum_{i=1}^{n} y^{(i)} \beta^\top x^{(i)} - \exp(\beta^\top x^{(i)})$$

$$\nabla L(y \mid x, \beta) = \sum_{i=1}^{n} y^{(i)} x^{(i)} - \exp(\beta^\top x^{(i)}) x^{(i)}$$

$$\nabla^2 L(y \mid x, \beta) = -\sum_{i=1}^{n} \exp(\beta^\top x^{(i)}) (x^{(i)} x^{(i)})^\top$$

Confidence intervals based on Fisher information: $\mathrm{I}(\beta) = -\nabla^2 L$

$$\mathrm{I}(\beta) = \sum_{t=1}^{T} \exp(\beta_0 + \beta_1 t) \begin{bmatrix} 1 & t \\ t & t^2 \end{bmatrix}$$

Large whenever counts are large, independent of variation!

# Misspecification Issues

Peril of assumptions: at the mercy of your model; $\mathrm{Var}(\mathrm{Poisson}(\mu)) = \mu$

Poisson distribution too narrow, leads to overconfident posterior

Common issue (esp. with count data): **overdispersion**

# Misspecification Issues

Peril of assumptions: at the mercy of your model; $\mathrm{Var}(\mathrm{Poisson}(\mu)) = \mu$

Poisson distribution too narrow, leads to overconfident posterior

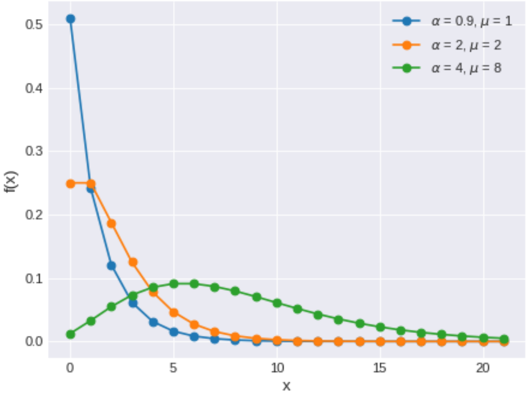
Common issue (esp. with count data): **overdispersion**

Typical fix: negative binomial distribution

$$p_{\mu,\alpha}(k) \propto \binom{k+\alpha-1}{k}\left(\frac{\mu}{\mu+\alpha}\right)^k$$

Mean $\mu$, overdispersion $\alpha$ (variance $\mu \cdot (1+\mu/\alpha)$)

# Negative binomial plots




[Credit: PyMC3 docs]

# Negative binomial regression on COVID-19 data

Instead of $F(\mu) = \mathrm{Poisson}(\mu)$, use $F_\alpha(\mu) = \mathrm{NegativeBinomial}(\mu, \alpha)$

Standard ways of fitting $\alpha$, i.e. MLE (or just set to a constant, but confidence intervals scale with $\alpha$)

[Jupyter demo]

Medium post: `https://medium.com/@jsteinhardt/the-growth-rate-of-covid-19-74944fc1d0f6`

# Discussion: modeling assumptions

What other modeling assumptions might be violated for the COVID-19 data?