# Advanced Bayesian GLMs: Quantifying Uncertainty & Model Checking

Data 102 Fall 2022

Lecture 11

# Weekly Outline

- So far: regression, GLMs, Bayesian vs frequentist
- **Today: Advanced modeling (Bayesian perspective)**
    - Quantifying uncertainty
    - Model checking
- Next time: Frequentist perspective on uncertainty & model checking

# Recap: Generalized Linear Models (GLMs)

- Model that describes the relationship between predictors (x) and targets (y)
- How the model makes predictions:
  - Take each predictor (x), multiply it by a coefficient (beta), and add them all up
  - Apply the (inverse) link function to get the average prediction
  - Assume the targets (y) have a particular likelihood model centered around that average
- Examples:

| Model | Link function | Likelihood model |
|---|---|---|
| Linear regression: | identity (inverse) link, | Gaussian likelihood |
| Logistic regression: | sigmoid (inverse) link, | Bernoulli likelihood |
| Poisson regression: | exponential (inverse) link, | Poisson likelihood |
| Neg. binomial regression: | exponential (inverse) link, | negative binomial likelihood |
| **Molecule concentration (HW3)** | **???** | **???** |
| **Your problem here** | **<choose one>** | **<choose one>** |

# GLMs, step-by-step

1. Formulate your prediction problem (define what x, y mean)
    a. Could involve feature engineering: more on this next week
2. Gather training data (x, y pairs)
3. Choose an inverse link function and likelihood that make sense for your data
4. Fit the model using training data (in this class, using PyMC3 or statsmodels)
5. *Check that the model is actually a good fit for the data*
6. Generate predictions for new x where y is unknown
7. *Report uncertainty for the new predictions*

# Model Checking: Is my model a good fit for my data?

- Question
- Step 1: Is the model a good fit for my *training* data?
  - Focus of today & next time
- Step 2: Is the model going to be a good fit when I see *new* data?
  - Use held-out test set to answer
  - More on this next week

# Bayesian Model Checking: Posterior Predictive Checks

- Regression:
  - Consider training set $x_1, \ldots, x_n, y_1, \ldots, y_n$
  - For each $x_i$, draw PPC samples from the posterior predictive distribution for $y_i$
- Non-regression
  - Consider training set $x_1, \ldots, x_n$
  - Draw PPC samples from the posterior predictive distribution for $x_{n+1}, \ldots$
- Check whether those new PPC samples are “reasonable” given the data
  - If they do, our model is a good fit for the training data
  - If they don’t, our model is not a good fit for the training data