More PyMC3 / Regression Models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

February 17, 2021

Outline

Last time:

- Posterior predictive checking
- Specifying and sampling from a model in PyMC3

Today:

- Simple linear regression; the role of priors
- Prior predictive checking
- More PyMC3

Regression

Linear regression

We all know simple linear regression:

- Have a predictor variable x and a response variable y
- Pose a model equation of the form

$$\hat{y} = a + bx$$

- Seek a, b so that the mean squared error

$$\frac{1}{N}\sum_i (y_i - \hat{y}_i)^2$$

  is minimized

How is this reproduced in our Bayesian modeling framework?

Linear regression

This is easily reframed as a Bayesian model:

$$\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= \underbrace{a + bx_i} \\
\sigma_i &\sim \text{(some prior)} \\
a &\sim \text{(some prior)} \\
b &\sim \text{(some prior)}
\end{aligned}$$

want something defined on $[0, \infty)$

Half-Cauchy

$$p(\sigma) = \frac{1}{1+\sigma^2}$$

$$\sigma \sim \text{Exponential}(1)$$

$$p(\log \sigma^2) \propto 1$$

$$p(\sigma^2) \propto (\sigma^2)^{-1}$$

Prior predictive simulations

Another way to make sense of this idea is to look at prior predictive simulations:

- Last time, we used posterior predictive simulations to assess model fit
- Prior predictive simulations can be used to

Example: CO2 vs. global temperature anomaly

Carbon dioxide and temperature

Our toy model for today: global average temperature as a function of atmospheric CO2, measured between 1959 and 2016

- $c =$ CO2 concentration in units of 100 ppm
- $T =$ global average temperature (Celsius) relative to 20th century average

Linear model:

$$\begin{aligned} T_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\ \mu_i &= a + bc_i \end{aligned}$$

Carbon dioxide and temperature

To specify the model, we need to put some priors on a, b, σ.

Let's explore three options.

- Vague prior: normal with a wide variance
- Weakly informative: normal with a narrower variance
- Slightly more informative: normal on a, log-normal on b

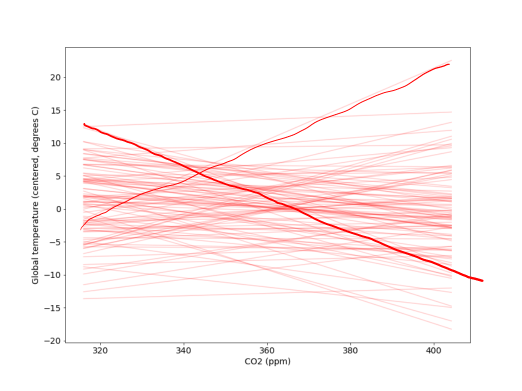
Why might we prefer some of these options over the others?

Prior predictive simulations

Prior #1:

a $\alpha \sim \mathrm{Normal}(0, 5)$

b $\beta \sim \mathrm{Normal}(0, 10)$



Prior predictive simulations

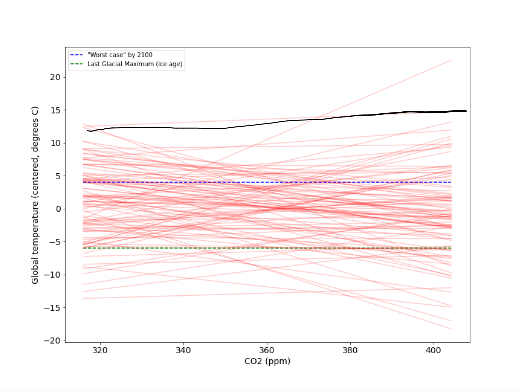
Prior #1:

$$\alpha \sim \mathrm{Normal}(0, 5)$$

$$\beta \sim \mathrm{Normal}(0, 10)$$

should be fairly close to 0 (when fitting against centered x values)

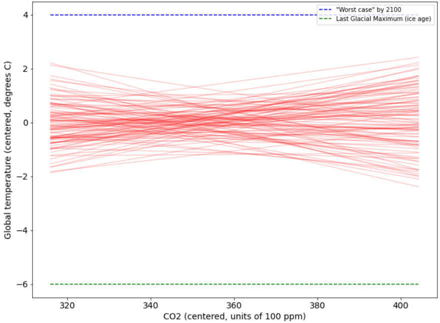
guess slope should be between (-2, +2)



Prior predictive simulations

Prior #1:

$$\alpha \sim \mathrm{Normal}(0, 0.5)$$
$$\beta \sim \mathrm{Normal}(0, 1)$$



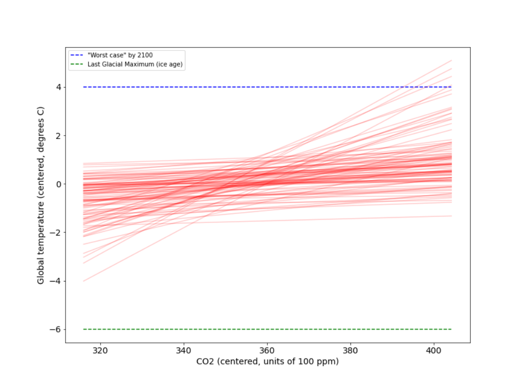
Prior predictive simulations

Prior #1:

$$\alpha \sim \text{Normal}(0, 0.5)$$

$$\beta \sim \text{LogNormal}(0, 1)$$

$\log(\beta) \sim \text{Normal}(0, 1)$



Connection to OLS

Simple case: no intercept, only slope.

Joint likelihood:

$$p(y_i|b,\sigma) \underset{\text{const}}{\propto} \overbrace{\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - bx_i)^2}{\sigma^2}\right)}^{\text{function of } y, b, \sigma}$$

- Conditional on σ, likelihood is maximized by minimizing the mean squared error

each $y_i \sim N(bx_i, \sigma)$

$$p(y_i) = \frac{1}{\text{const}} \exp\left(-\frac{1}{2}\frac{(y_i - bx_i)^2}{\sigma^2}\right)$$

What does the prior on b do?

Let's suppose we put a normal prior on the slope parameter $\overset{b}{\cancel{\beta}}$:

$$
\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= b x_i \\
\sigma_i &\sim \text{(some prior)} \\
b &\sim \mathrm{Normal}(0, \tau^{\cancel{2}})
\end{aligned}
$$

$$p(b \mid y) = \underbrace{p(y \mid b)}_{\text{last slide}} \times p(b)$$

What does the prior on b do?

Putting this prior in, we get a posterior distribution (again conditional on σ):

$$p(b|y,\sigma) \propto \exp\left(-\frac{1}{2}\left(\overbrace{\sum_{i=1}^{N}\frac{(y_i - bx_i)^2}{\sigma^2} + \frac{1}{\tau^2}b^2}^{\text{minimize this}}\right)\right)$$

Suppose we seek the mode (maximum) of the posterior distribution.

- What quantity would we minimize?
- Does this look familiar?

Ridge regression

Ridge regression: a "penalized" version of OLS that minimizes the loss function

$$\mathcal{L}(b) = \underline{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} + \underbrace{\alpha}_{\text{not the same } \alpha} b^2$$

- a form of regularization – reducing model variability to combat overfitting
- although we don't generally use the MAP estimate in Bayesian inference, a weakly informative prior on the coefficients still performs this function

Logistic regression

Logistic regression

A quick sidebar to translate another familiar model into the Bayesian framework.

Section 3.7 from BDA: logistic regression for a bioassay experiment.

- Toxicity testing of chemicals on animals
- 20 animals split into four groups, each given a different dose x_i
- Number of deaths recorded

The model

Within each group, we model each animal as independent

This leads to the model:

$$\begin{aligned} y_i|\theta_i &\sim \mathrm{Binomial}(n_i, \theta_i) \\ \mathrm{logit}(\theta_i) &= \alpha + \beta x_i \\ \alpha &\sim \mathrm{Normal}(0, 1) \\ \beta &\sim \mathrm{Normal}(0, 5) \end{aligned}$$

Let's run it in PyMC3...

Intro to multilevel models

How much traffic is bicycle traffic?

In the book there is a data set with observations of bicycle traffic on a number of streets.

- Data collected by standing at the roadside for some amount of time
- Count number of bicycles and number of non-bicycle vehicles
- Parameter of interest: proportion of traffic that is bicycles

Why multilevel modeling?

Two competing approaches to describing this:

- There is a single rate of bicycle traffic θ, and each observation is a separate observation of a $\mathrm{Binomial}$ with the same parameter
- Every street is different; there are 10 independent θ_i; each observation y_i is an observation of a $\mathrm{Binomial}$ with that θ_i

The hierarchical model attempts to combine the merits of these two ideas.

Why multilevel modeling?

Imagine you're collecting data for this.

- You go out to Mountain Ave. and count cars and bicycles for an hour.
- The following can both be true:
  - Mountain Ave. is not representative of the rate of bicycle traffic in Tucson – if we treated it that way, we'd probably overestimate
  - Mountain Ave. is in part a reflection of the fact that Tucson is, overall, fairly bicycle-friendly

Variability at two levels

There is variability at two levels in this problem:

- Different cities have, overall, more or less bicycle traffic
- Within a given city, different streets have more or less bicycle traffic

So, observations of one street should inform our estimates of other streets, *without* making the same estimate for every street

Diagram of the multilevel model

Exercise 3.8

The model:

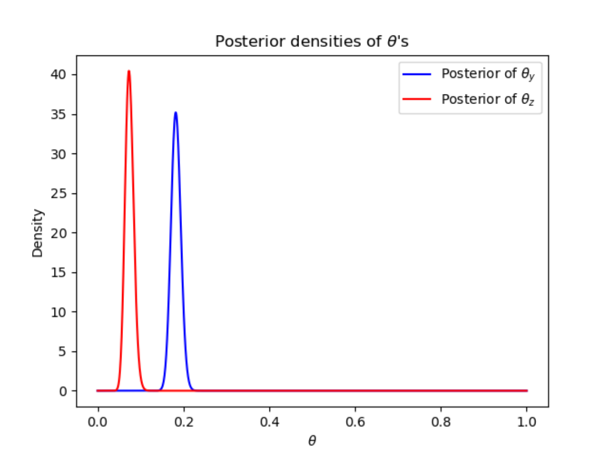
$$y_j \sim \mathrm{Binomial}(\theta, n_j)$$
$$\theta \sim \mathrm{Beta}(\alpha_0, \beta_0)$$

for fixed α_0, β_0.

- Choosing $\alpha_0 = 1, \beta_0 = 1$ gives a completely noninformative (flat) prior
- Weakly informative prior also reasonable, e.g. $\alpha_0 = 1, \beta_0 = 3$ for prior mean of 25% bicycle traffic

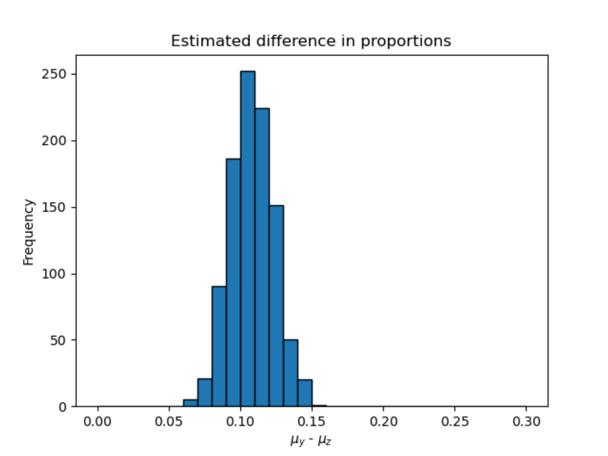
Exercise 3.8

Posteriors for the streets with and without bike lanes:



Exercise 3.8

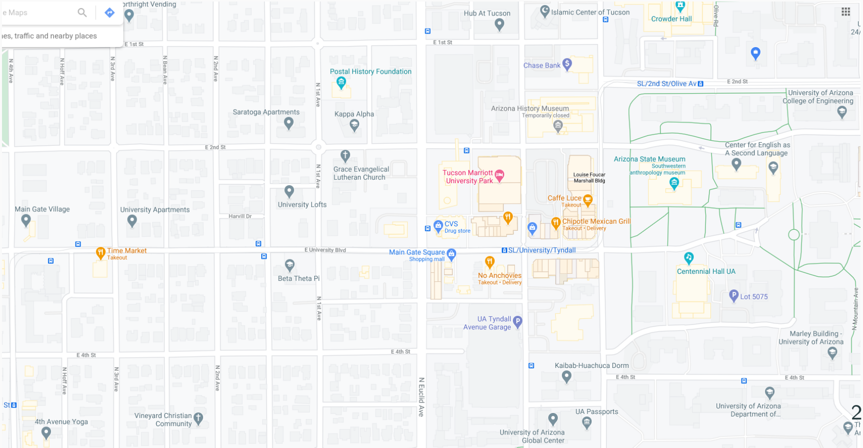
Then, by sampling from the posterior, we can estimate the distribution of quantities of interest, such as $\theta_y - \theta_z$:



A hierarchical model

Why a hierarchical model?

The model we wrote in the previous section treats all streets as the same; each street's observation is an observation of the same underlying proportion.



Why a hierarchical model?

As an alternative, we could treat each street as an independent entity:

Why a hierarchical model?

As an alternative, we could treat each street as an independent entity:

$$
\begin{aligned}
y_j &\sim \mathrm{Binomial}(\theta_j, n_j) \\
\theta_j &\sim \mathrm{Beta}(\alpha_0, \beta_0)
\end{aligned}
$$

for fixed α_0, β_0.

- Exactly like the previous model, except we now have 10 independent θ_js for the 10 streets
- Same considerations for choice of prior

Call this the separate-effects model.

Why a hierarchical model?

Choosing between the two models: classically, do an analysis of variance

- Compare variance within groups (streets) to variance between streets
- Test against the null hypothesis that all streets are the same
- If we reject the null, take the separate-effects model
- If we don't take the pooled model

Problem: false dichotomy!

Why a hierarchical model?

In reality, it is most plausible that both of the following are true:

- The streets are not identical; some of the streets are more popular with cyclists
- Observations of one street can inform our knowledge of the others

So: neither side of this dichotomy is preferable.

Analogy: cafes

Imagine you're walking into a cafe; how long will it take you to get your coffee?

- Varies among cafes / franchises
- Not completely independent

The Bayesian solution

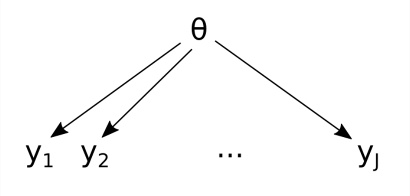
With a Bayesian approach, we can find a compromise.

- We have a θ for each street
- However, instead of being fully independent, each θ is drawn from a common probability distribution
- This probability distribution, a *hyperprior*, depends on *hyperparameters* which we estimate from the data

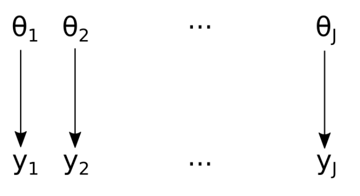
(note: slightly different sense of the term *hyperparameter* from its common use in ML)

Examining this graphically

Pooled model:

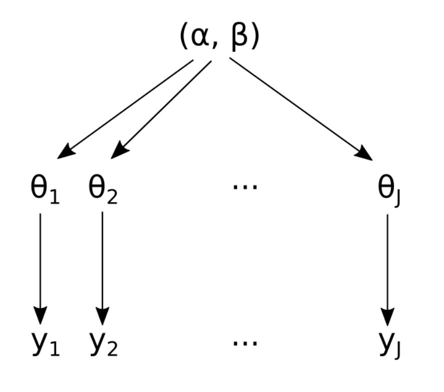


Separate model:



Examining this graphically

Hierarchical model combines the features of these two:



Setting up the model

This is conceptually only a slight difference from our previous model:

$$
\begin{aligned}
y_j &\sim \mathrm{Binomial}(\theta_j, n_j) \\
\theta_j &\sim \mathrm{Beta}(\alpha, \beta) \\
p(\alpha, \beta) &\propto ???
\end{aligned}
$$

Choosing a hyperprior

We need a prior distribution for α, β; this can be a tricky part of this sort of modeling, because the interpretation of these parameters is not so simple compared to θ_j.

Starting with the idea that α and β can represent "pseudocounts", parameterize in terms of

- $\mu = \frac{\alpha}{\alpha+\beta}$ – prior expectation
- $\eta = \alpha + \beta$ – prior "sample size" (think of this like a precision)

Choosing a hyperprior

Prior for μ:

- Need $0 < \mu < 1$, so we'll choose a Beta
- Informative version: cars outnumber bikes, so try $\mathrm{Beta}(1, 3)$

Prior for η:

- $\eta > 0$, so choose something with range $(0, \infty)$
- Fairly spread out, but put more prior mass near 0; try a half-Cauchy

Setting up the model

This is conceptually only a slight difference from our previous model:

$$
\begin{aligned}
y_j &\sim \mathrm{Binomial}(\theta_j, n_j) \\
\theta_j &\sim \mathrm{Beta}(\alpha, \beta) \\
\mu &:= \frac{\alpha}{\alpha + \beta} \\
\eta &:= \alpha + \beta \\
p(\mu) &\sim \mathrm{Beta}(1, 3) \\
p(\eta) &\sim \mathrm{HalfCauchy}(1)
\end{aligned}
$$

Note: BDA uses a vaguely similar but fairly opaque approach to reach a prior that is qualitatively fairly similar

Doing inference

How can we hope to get a handle on the posterior distribution for this model?

Inference the hard way

As usual, we can make inferences by sampling from the posterior distribution. This can be done the hard way (directly), or the easy way (MCMC).

Hard way:

1. Calculate the posterior density $p(\alpha, \beta | y)$ on a grid of α and β values.
2. Sum over the β values to get an estimate of the marginal posterior $p(\alpha | y)$; use this to draw samples of α.
3. For each sampled value of α, use the conditional posterior $p(\beta | \alpha, y)$ (which is a slice of)
4. For each sampled pair (α_i, β_i), draw values of θ_j from the beta distribution $\mathrm{Beta}(\alpha_i + y_j, \beta_i + y_j - n_j)$

Inference the easy way

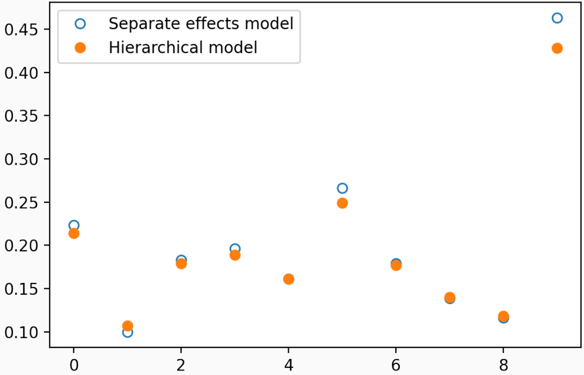
The easier approach: use MCMC to sample from the posterior.

Let's see this in action...

Comparison

What is the difference in the results?

Let's compare point estimates:



Shrinkage and regularization

The shrinkage effect we see is a form of regularization:

- Most extreme observations “shrunk” toward an overall average
- Amount of shrinkage tuned to relative sample size

Difference: we learned the strength of regularization from the data

Underfitting and overfitting

Another way to think about this, in terms of underfitting and overfitting:

- The pooled model: strong underfitting
- The separate-effects model: strong overfitting
- Hierarchical model: adaptive regularization

With enough observations the seperate effects model will estimate each street similarly to the hierarchical model.

Summary

Hierarchical models:

- Have several “levels” of parameters stacked
- Perform adaptive regularization – learn priors from the data

Next time: more models

Appendix: hyperprior calculation

Choosing a hyperprior

BDA suggests the following as a prior for a similar example:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

Choosing a hyperprior

BDA suggests the following as a prior for a similar example:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

In a beta distribution, interpretation of parameters as "pseudocounts":

Choosing a hyperprior

BDA suggests the following as a prior for a similar example:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

In a beta distribution, interpretation of parameters as "pseudocounts":

- If we start with $\mathrm{Beta}(\alpha, \beta)$ and make binomial observations, we update to the posterior $\mathrm{Beta}(\alpha + n_s, \beta + n_f)$, with n_s successes and n_f failures
- So, we can think of α and β as "counts" of imaginary observations

Choosing a hyperprior

Goal: prior is noninformative on the mean value of θ_j and the spread, or scale, of that mean

- Mean is $\frac{\alpha}{\alpha+\beta}$
- Scale parameters (standard errors) for means are distributed like $n^{-1/2}$ where n is the sample size

So: set up a prior distribution that is uniform on $\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right)$

Choosing a hyperprior

Define:

$$\begin{aligned} w &= \frac{\alpha}{\alpha + \beta} \\ z &= (\alpha + \beta)^{-1/2} \\ p(w, z) &\propto 1 \end{aligned}$$

To get to

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

we have to do some calculus (on the following slides!)

Reminder

As a reminder, our prior distribution was uniform on

$$\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right)$$

Define $w = \frac{\alpha}{\alpha+\beta}, z = (\alpha+\beta)^{-1/2}$, and set $p(w,z) \propto 1$.

Changing variables for probability densities comes from changing variables for integrals, because the PDF is defined by the property that

$$\mathrm{Pr(x_1, \ldots, x_n \in A) = \int_A p(x_1, \ldots, x_n) dx_1 \ldots dx_n}$$

To perform the change of variables, we need to multiply by the absolute determinant of the Jacobian matrix

$$J = \begin{pmatrix} \frac{\partial w}{\partial \alpha} & \frac{\partial w}{\partial \beta} \\ \frac{\partial z}{\partial \alpha} & \frac{\partial z}{\partial \alpha} \end{pmatrix}$$

To perform the change of variables, we need to multiply by the absolute determinant of the Jacobian matrix

$$J = \begin{pmatrix} \frac{\partial w}{\partial \alpha} & \frac{\partial w}{\partial \beta} \\ \frac{\partial z}{\partial \alpha} & \frac{\partial z}{\partial \alpha} \end{pmatrix}$$

$$J = \begin{pmatrix} \frac{\beta}{(\alpha+\beta)^2} & \frac{-\alpha}{(\alpha+\beta)^2} \\ -\frac{1}{2}(\alpha+\beta)^{-3/2} & -\frac{1}{2}(\alpha+\beta)^{-3/2} \end{pmatrix}$$

so $|\det J| = \frac{1}{2}(\alpha+\beta)^{-5/2}$ (and we can drop the 1/2 because it's a constant)