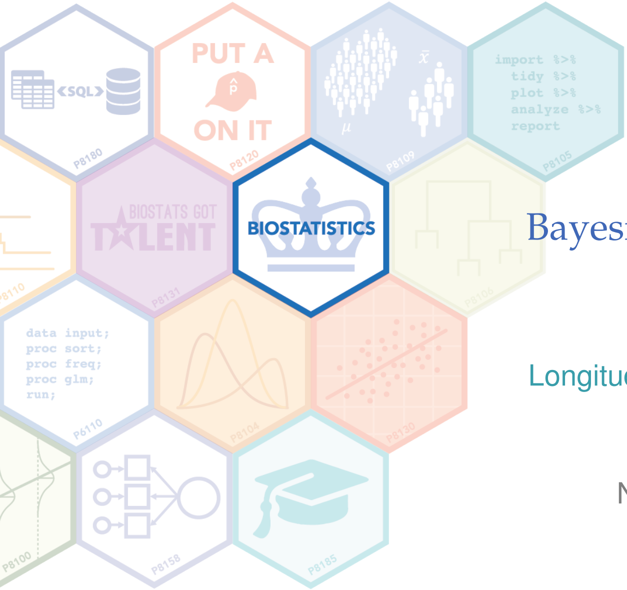
COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH



# Bayesian Regression

Julia Wrobel

Longitudinal Data Analysis

November 26, 2018

# Today’s Lecture

- Bayesian methods
  - One-sample Normal-Normal
  - Regression using Normal-Normal model
  - Regression with unknown variance
  - Regression implementation
  - Random slope model with pigs data

# A simple motivating example

Imagine we are interested in the daily number of steps taken by Mailman students. We use accelerometers to collect a random sample of step counts from $n = 10$ students.

- Let $y_i$ be the step count for the $i^{th}$ student
- Assume $y_i \sim N(\mu, \sigma_y^2)$

We want to learn about $\mu$, the average daily number of steps taken by Mailman students.

# A simple motivating example

Before we start, what's your best guess about $\mu$?

# The frequentist approach

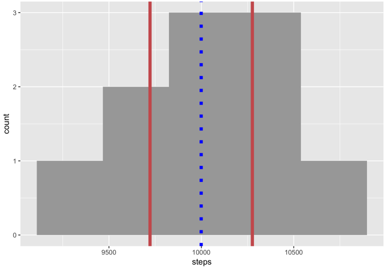
We want to learn about $\mu$, the average daily number of steps taken by Mailman students.

- Parameter $\mu$ is fixed and unknown
- The data $y$ is random
- The sample mean $\hat{\mu} = \bar{y}$ is a statistic, and a frequentist estimator of the population mean $\mu$
- Base inference on p-values and confidence intervals

Follow along with examples in the file **bayesian_code.R.**

# Frequentist inference about $\mu$

```
> head(steps)
[1]  9732  9270 10446 10699
>
> mean(steps)
[1] 9999.2
>
> sd(steps)
[1] 446.6863
```



We are 95% confident that the true number of steps taken by Mailman students lis between 9722 and 10276 steps.

# The Bayesian approach

- Bayesians treat the parameter $\mu$ as random
- Express uncertainty about $\mu$ using probability distributions
- The distribution before observing the data is called the **prior distribution**
  - Allows incorporation of prior knowledge
- The distribution after observing the data is called the **posterior distribution**
- Inference is conditional on the dataset we observe

# The Bayesian approach

What do I think know?

- $y_i|\mu \sim \mathsf{N}\left[\mu, \sigma_y^2\right]$
- $\mu \sim \mathsf{N}\left[\mu_0, \sigma_\mu^2\right]$

What do I want to learn?

- $\mu|y_i \sim ???$

This is a **one sample Normal-Normal** model

# The Bayesian approach

Relate prior, likelihood, and posterior through Bayes' formula:

$$\begin{aligned} p(\mu|y_i) &= \frac{p(y_i|\mu)p(\mu)}{p(y_i)} \\ &= \frac{p(y_i|\mu)p(\mu)}{\int_\mu p(y_i|\mu)p(\mu)d\mu} \\ &\propto p(y_i|\mu)p(\mu) \end{aligned}$$

For the Normal likelihood with a Normal prior for $\mu$, the posterior is also Normal:

$$\mu|y_i \sim \mathsf{N}\left[\frac{\sigma_\mu^2}{\frac{\sigma_y^2}{n}+\sigma_\mu^2}\bar{y}+\frac{\frac{\sigma_y^2}{n}}{\frac{\sigma_y^2}{n}+\sigma_\mu^2}\mu_0, \frac{\frac{\sigma_y^2}{n}\sigma_\mu^2}{\frac{\sigma_y^2}{n}+\sigma_\mu^2}\right]$$

# Deriving posterior distribution for $\mu$

We could do this calculation using sums. Instead we'll use multivariate distributions.

$p(\mu|y_i) \propto p(y_i|\mu)p(\mu)$

# Distribution for $\mu$

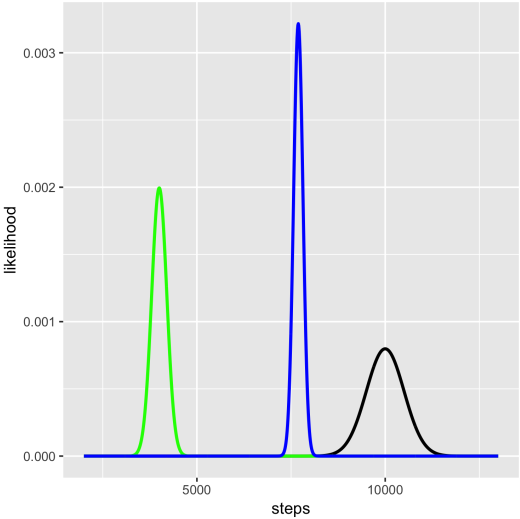
# Distribution for $\mu$

# Choosing the prior distribution

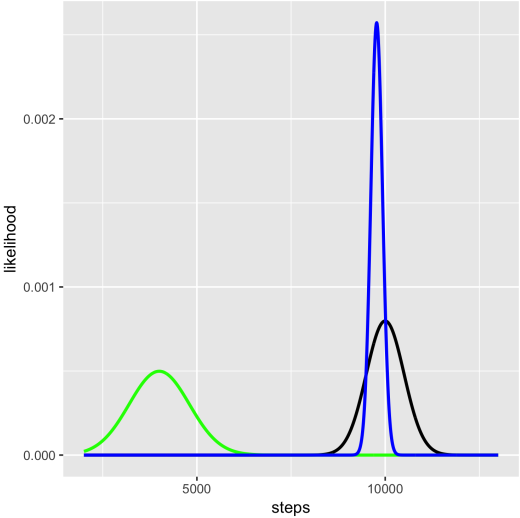
A study in *Medicine & Science in Sports & Exercise* reports that Americans take an average 4000 steps per day with a standard deviation of 200 steps. We come up with three different priors:

- Informative prior
  - Maybe we really believe the previous study
  - $\mu \sim N(4000, 200^2)$
- Weakly informative prior
  - Or maybe we believe it a little bit
  - $\mu \sim N(4000, 800^2)$
- Uninformative or diffuse prior
  - Or maybe we don't have any scientific information
  - $\mu \sim N(0, 4000^2)$

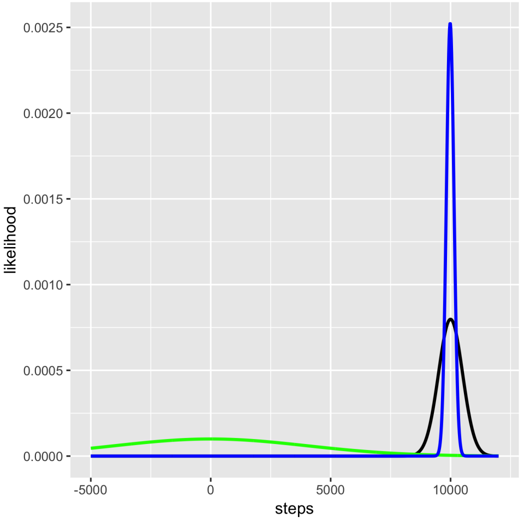
# Effect of informative prior



# Effect of weakly informative prior



# Effect of uninformative prior



# Bayesian inference about $\mu$

We decide to go with the uninformative prior.

- posterior mean: $\mu|y = 9984$ steps
- posterior 95% **credible interval** is just the (2.5%, 97.5%) quantile of the posterior distribution

```
> qnorm(c(0.025, 0.927), mu_post, sigma_post)
[1]  9673.945 10213.288
```

- Posterior probability that $\mu < 10000$ given $y$

```
> pnorm(10000, mu_post, sigma_post)
[1] 0.5413359
```

Bayesian inference allows us to make probability statements about $\mu$ given the data.

# Linear regression

We are often interested in estimating the parameters in the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2)$$

Specifically, we want to estimate $\boldsymbol{\beta}$ and $\sigma^2$

# Least squares / frequentist methods

- Estimate $\boldsymbol{\beta}$ using least squares (or maximum likelihood)

$$\hat{\boldsymbol{\beta}}_{OLS} = argmin_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- Compute residuals using $\hat{\boldsymbol{\beta}}$, and from this estimate $\hat{\sigma}^2$
- Base inference for $\boldsymbol{\beta}$ on $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$
- For *iid* observations and normal errors the likelihood is

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

# Steps for a Bayesian linear regression

The Bayesian approach is a little different

- Assign prior distributions to parameters of interest
  - Normal prior for $\beta$
  - Inverse gamma for $\sigma^2$
- Choose hyper-parameters
  - Prior mean and variance for $\beta$
  - Shape and scale for $\sigma^2$
- Obtain joint posterior distribution, and base inference on this

# Bayesian linear regression (known variance)

We want a Bayesian framework for the regression model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\epsilon} \sim \mathrm{N}\left[0, \sigma_{\boldsymbol{y}}^2 \mathbf{I}_n\right]$.

- Need to make distributional assumptions about $\boldsymbol{\beta}$
  - Normal priors seemed to work well in the past ...
- Try $\boldsymbol{\beta} \sim \mathrm{N}\left[0, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}_p\right]$ where $p$ includes the intercept

# Bayesian regression (known variance)

We want to obtain the posterior

$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X})p(\boldsymbol{\beta})$

# Bayesian regression (known variance)

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N(\boldsymbol{\beta}_{post}, \Sigma_{post})$$

$$\Sigma_{post} = \left(\frac{1}{\sigma_y^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\boldsymbol{I}_p\right)^{-1}$$

$$\begin{aligned}\boldsymbol{\beta}_{post} &= \Sigma_{post}\left(\frac{1}{\sigma_y^2}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{\sigma_\beta^2}\boldsymbol{\beta}_0\right) \\ &= \left(\boldsymbol{X}^T\boldsymbol{X} + \frac{\sigma_y^2}{\sigma_\beta^2}\boldsymbol{I}_p\right)^{-1}(\boldsymbol{X}^T\boldsymbol{y})\end{aligned}$$

# So, about the variances

- Throughout all of this we have implicitly conditioned on the variances $\sigma_{\boldsymbol{y}}^2$ and $\sigma_{\boldsymbol{\beta}}^2$
- How do we “choose” the variance terms??

# Prior variance

- The prior variance $\sigma^2_\beta$ is often pre-selected to indicate the amount of prior knowledge
- Typically, this is chosen very large to indicate a lack of knowledge
  - Synonyms includes a “diffuse” prior or an “uninformative” prior; $\sigma^2_\beta$ is called a “hyper-parameter”
  - Basically it means that your prior will be dominated by the data and likelihood

# Outcome variance

- The outcome variance $\sigma_y^2$ is a quantity of interest
- Ideally, we'd like to estimate this using the observed data
- Since we're already being Bayesians, why not assign a prior distribution and try to find a posterior?
- The inverse-gamma distribution works pretty well ...

# Inverse-gamma distribution

We'll use an inverse-gamma distribution for the variance component $\sigma_y^2$:

$$p(\sigma_y^2|A,B) = \frac{B^A}{\Gamma(A)}(\sigma_y^2)^{-A-1}\exp\left(-\frac{B}{\sigma_y^2}\right)$$

# Outcome variance posterior

We want to find

$$p(\sigma_{\boldsymbol{y}}^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma_{\boldsymbol{y}}^2, \boldsymbol{X})p(\sigma_{\boldsymbol{y}}^2)$$

The posterior outcome variance is also Inverse Gamma!

# Some Bayesian language

- The parameters we're interested in are $\boldsymbol{\beta}, \sigma^2_{\boldsymbol{y}}$
- The priors we've used are called **conjugate priors**
- The **hyperparameters** we've chosen are $\sigma^2_{\boldsymbol{\beta}}, A, B$
- The distributions we've calculated are

  $$p(\sigma^2_{\boldsymbol{y}} | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) \text{ and } p(\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X}, \sigma^2_{\boldsymbol{y}}).$$

  These are called **full conditionals**
- The posterior distribution of interest is $p(\boldsymbol{\beta}, \sigma^2_{\boldsymbol{y}} | \boldsymbol{y}, \boldsymbol{X})$, which is called the **joint posterior**

# An interlude on conjugate priors

- A prior is **conjugate** if the posterior is a member of the same parametric family
- Some examples are
  - **beta-binomial** model: If the response is binomial and we use a beta prior, the posterior is beta
  - **gamma-Poisson**: Poisson response + gamma prior = gamma posterior
  - **normal-normal** model: normal response + normal prior = normal posterior
- Advantage of a conjugate prior is that the posterior is available in closed form

# Back to the joint posterior distribution

Our goal is to learn about the joint posterior distribution:

$$p(\boldsymbol{\beta}, \sigma_{\boldsymbol{y}}^2 | \boldsymbol{y}, \boldsymbol{X}, \text{hyperparameters})$$

We can calculate this from the full conditionals using the law of conditional probability:

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma_{\boldsymbol{y}}^2 | \boldsymbol{y}, \boldsymbol{X}, h) &= p(\boldsymbol{\beta} | \sigma_{\boldsymbol{y}}^2, \boldsymbol{y}, \boldsymbol{X}, h) p(\sigma_{\boldsymbol{y}}^2 | \boldsymbol{y}, \boldsymbol{X}, h) \\ &= p(\boldsymbol{\beta} | \sigma_{\boldsymbol{y}}^2, \boldsymbol{y}, \boldsymbol{X}, h) \int_{\beta} p(\sigma_{\boldsymbol{y}}^2 | \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}, h) d\boldsymbol{\beta} \end{aligned}$$

...

# Joint posterior distribution

Our goal is to learn about the joint posterior distribution

$$p(\boldsymbol{\beta}, \sigma_{\boldsymbol{y}}^2 | \boldsymbol{y}, \boldsymbol{X}, \text{hyperparameters})$$

but (as we saw above) this is hard.

- Often the joint posterior is analytically intractable
- For the conjugate priors we use, though, the full conditionals were “easy”
- If we can’t write down the joint posterior, maybe we can sample from it

# MCMC let's you sample from complicated posterior distributions

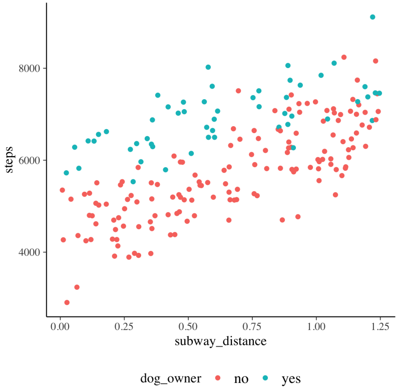
- Markov chain Monte Carlo (MCMC) methods are a collection of tools used to sample from a target distribution
- Once you have a sample from the posterior distribution you can use summary statistics (mean, variance, quantiles) to do posterior inference
- There are lots of ways to do MCMC
- The code provided runs MCMC but I'm not really going to explain how it works here

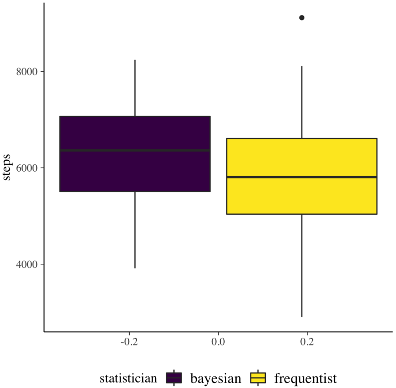
# Bayesian regression example

We now want to consider the effect of covariates on the activity levels of Mailman students. For $N = 500$ students we collect **step count** and the following variables:

- **age** in years
- **dog_ownership**: whether or not student owns a dog ( $0 = no$, $1 = yes$)
- **subway_distance**: distance in miles from student's home to nearest subway
- **statistician**: type of statistician ($0 = bayesian$, $1 = frequentist$)

# Some EDA


steps
8000
6000
4000
0.00
0.25
0.50
0.75
1.00
1.25
subway_distance
dog_owner
● no
● yes


steps
8000
6000
4000
-0.2
0.0
0.2
statistician
bayesian
frequentist

# Model specification

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

Define the likelihood:

- $\boldsymbol{\epsilon} \sim N(0, \sigma_{\boldsymbol{y}}^2 \boldsymbol{I}_n)$
- implies $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma_{\boldsymbol{y}}^2 \boldsymbol{I}_n)$

Define the priors:

- $\boldsymbol{\beta} \sim N(0, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{I}_b)$
- $\sigma_{\boldsymbol{y}}^2 | \lambda \sim exponential(\lambda)$

Define the hyper parameters:

- Let $\sigma_{\boldsymbol{\beta}}^2 = 100$
- Let $\lambda = 1$

# Coding it up

We will use the `R` package `rstanarm`

```
blr_mod = stan_glm(steps ~ subway_distance +
                         age +
                         dog_owner +
                         statistician,
                   data = steps_df,
                   prior_intercept = normal(0, 10),
                   prior = normal(0, 10),
                   prior_aux = exponential(rate = 1))
```

# Regression coefficients

Coefficients from Bayesian regression are similar to least squares in this setting

```
|term            | coef_blr| coef_lm| true_value|
|:---------------|--------:|-------:|----------:|
|(Intercept)     |   5875.7|  5877.8|       6000|
|subway_distance |   1885.0|  1884.2|       1800|
|age             |    -45.8|   -46.0|        -50|
|dog_owner       |   1341.0|  1340.7|       1300|
|statistician    |   -583.0|  -582.8|       -500|
```

# Bayesian regression diagnostics

The package `rstanarm` uses MCMC to draw samples from the posterior distribution. There are standard checks to ensure that your results come from independent draws of the posterior distribution.

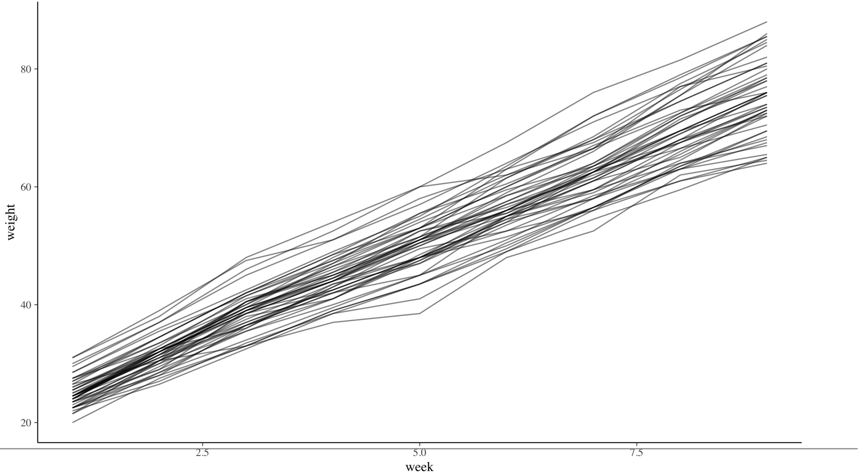
These are beyond the scope of this lecture but I encourage you to check them out!

# Bayesian hierarchical models

- Provide a way to do multilevel modeling, the Bayesian way
- Basically, just nest your priors
- A natural framework for doing Bayesian random effects models
- We'll go back to the pigs data for an example

# Bayesian random effects with pigs data

48 piglets were weighed each week for 9 weeks. They got bigger!



# Bayesian random effects model specification

The general (frequentist) form of the random effects model is:

$$y_{ij} = X_{ij}\boldsymbol{\beta} + Z_{ij}\boldsymbol{b}_i + \epsilon_{ij}$$

A (**frequentist**) random slope model for the pigs data is then:

$$y_{ij} = \beta_0 + \beta_1 week_{ij} + b_{0i} + b_{1i} week_{ij} + \epsilon_{ij}$$

For the **Bayesian** model only specify the subject-level effects:

$$y_{ij} = b_{0i} + b_{1i} week_{ij} + \epsilon_{ij}$$

The population level effects will get defined in the prior specification step

# Bayesian random effects model specification

$$y_{ij} = b_{0i} + b_{1i} week_{ij} + \epsilon_{ij}$$

Define the likelihood:

- $\epsilon_{ij} \sim N(0, \sigma_{\boldsymbol{y}}^2); \qquad y_{ij} \sim N(b_{0i} + b_{1i} week_{ij}, \sigma_{\boldsymbol{y}}^2)$

Define the priors:

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{b_0} \\ \mu_{b_1} \end{pmatrix}, \quad \begin{pmatrix} \sigma_{b_0}^2 & \rho\sigma_{b_0}\sigma_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix}\right]$$

Define priors on the hyperparameters

- $\mu_{b_0} \sim N(m_0, \tau_0^2)$
- $\mu_{b_1} \sim N(m_1, \tau_1^2)$

# Bayesian random effects model specification

The parameters are at the subject level

- these give subject specific effects and fitted values
- still have interpretation of deviation from population mean, as in frequentist approach

The hyperparameters are at the population level

- these “borrow strength” across pigs
- we’ll let the data “choose” the hyperparameter values

# Coding it up

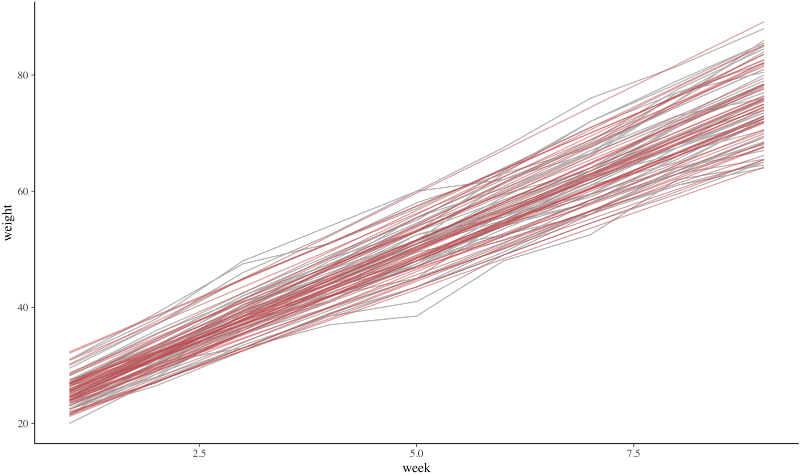
We can use `rstanarm::stan_lmer()` to specify the model.

```
bayesian_lmer = stan_lmer(weight ~ week +
                            (1 + week|id),
                          data = pigs_long)
```

We’ll use the default priors chosen by the software. After running this model we can examine these priors using:

```
> prior_summary(bayesian_lmer)
Priors for model ’bayesian_lmer’
------
...
```

# Fitted values for the data



# Some ways to estimate Bayesian models using MCMC

- WINBUGS: Ask a real Bayesian about this
- STAN: short person time, long computer time
- DIY: long person time, short(er) computer time

# Why you should consider Bayesian methods

- Can obtain joint distributions for parameters of interest – more fully account for sources of uncertainty
- Credible intervals are based on posterior probabilities
- More natural in some cases – sometimes full conditionals are an easier route than maximum likelihood

# Why you should avoid Bayesian methods

- Less familiar to collaborators
- More computationally demanding
- Can be sensitive to prior specification (although there are sensitivity issues with frequentist methods as well, as we’ve seen)

# References and Acknowledgements

Slides inspired by:

- Jeff Goldsmith’s 2014 course on Linear Regression Models
- NC State Applied Bayesian Analysis course
  - https://www4.stat.ncsu.edu/ reich/ABA/

A full tutorial on `rstanarm`:

- http://www.tqmp.org/RegularArticles/vol14-2/p099/p099.pdf

Thanks! Slides are on my website (juliawrobel.com)