36-617: Applied Linear Models

Bayes, Shrinkage, and Multi-Level Models

Brian Junker

132E Baker Hall

brian@stat.cmu.edu

Announcements

- Project 02 Schedule:
  - **Fri Nov 19:** Draft Technical Appendix with **HW 10**.
  - **Mon Nov 29 (or earlier):** Full IDMRAD paper first draft.
  - **Fri Dec 3:** Peer reviews due.
  - **Fri Dec 10 (or earlier):** Full IDMRAD paper final draft!
- Regular classes Nov 22, 29 & Dec 1
- No more graded hw's or quizzes
  - I may give one more ___ungraded___ hw, if I have time to set it up
  - End of semester feedback for me Nov 29
- *BJ's remaining office hours will be in the MSP space in FMS – __not__ in his office.*

Outline

- ## Today:
  - ❑ Shrinkage
  - ❑ Review of MLE
  - ❑ Crash course in Bayes
  - ❑ Normal-Normal Model & Shrinkage
  - ❑ MLM's and Shrinkage
- ## Mon:
  - ❑ Multilevel glm's (multilevel binomial, poisson, etc.)
- ## After Thanksgiving:
 - ❑ Some ideas from nonparametric regression

An MLM phenomenon: Shrinkage

The fitted multilevel model underpredicts high obs's and overpredicts low ones.

The distribution assumptions underlying lmer() "smooth out" extreme observations!

Multi-level models provide more smoothing/shrinkage to groups with smaller sample sizes (since there is less evidence that their values should be different from "grand mean".)

We'll talk about **_why_** today…
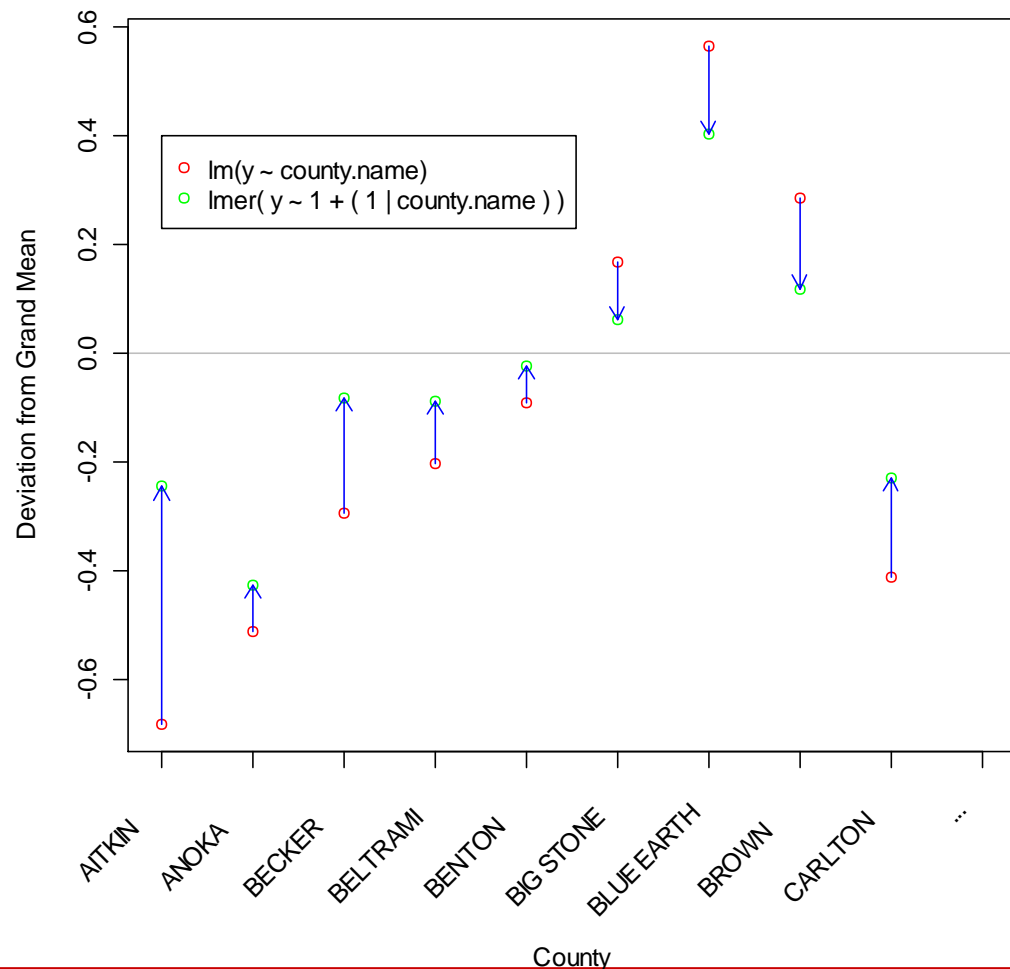
Methods of Estimation – How can we systematically construct "good" estimators?

- Several methods have proven useful:

  - *Method of Moments (MoM)*: The kth moment of X is $E[X^{\underline{k}}]$. MoM estimators combine unbiased estimates of moments of X.

  - *Least Squares (LS)*: Obtained by minimizing squared error $\sum_{i=1}^{n}(Y_i - E[Y_i])^2$. Ordinary linear regression!

  - *Maximum likelihood (ML)*: The likelihood is the probability of the data we observed. ML estimators (MLE's) choose parameter values that maximize the likelihood.

  - *Bayesian Estimation (Bayes)*: Treat the parameters as random variables, and use Bayes' rule to pick the parameter value most likely, given the data (the "reverse" of ML!)

Maximum Likelihood Estimators (MLE's)

- Let $X_1, ..., X_n$ be an iid sample from $f_X(x;\theta)$, $x_1, ..., x_n$ are the observed values

- The *likelihood* of the sample is the joint density

$$L(\theta) = f(x_1, ..., x_n; \theta) = f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta)$$

$$= \prod_{i=1}^{n} f(x_i; \theta)$$

- The *maximum likelihood estimate* $\hat{\theta}_{MLE}$ maximizes $L(\theta)$:

$$L(\hat{\theta}_{MLE}) \geq L(\theta) \quad \forall\ \theta$$

- Strategy: It's usually (but not always) easier to work with the *log likelihood*

$$LL(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)\ .$$

Florida, 2020 Pre-Election Poll

- Donald Trump (R) running for election to the presidency against Joe Biden (D)
- In a Suffolk University Poll (October 1-4, 2020):
  - 451 of 500 voters expressed a preference for Trump or Biden.
  - Of those 451: 226 prefer Donald Trump.
- In most polling, weights are attached to each response, to adjust the "representativeness" of the response for things like
  - who is likely to be home when survey worker calls
  - who refuses to answer
  - etc
- We will ignore weights etc and treat the 451 as a simple random sample.

Possible models for the data

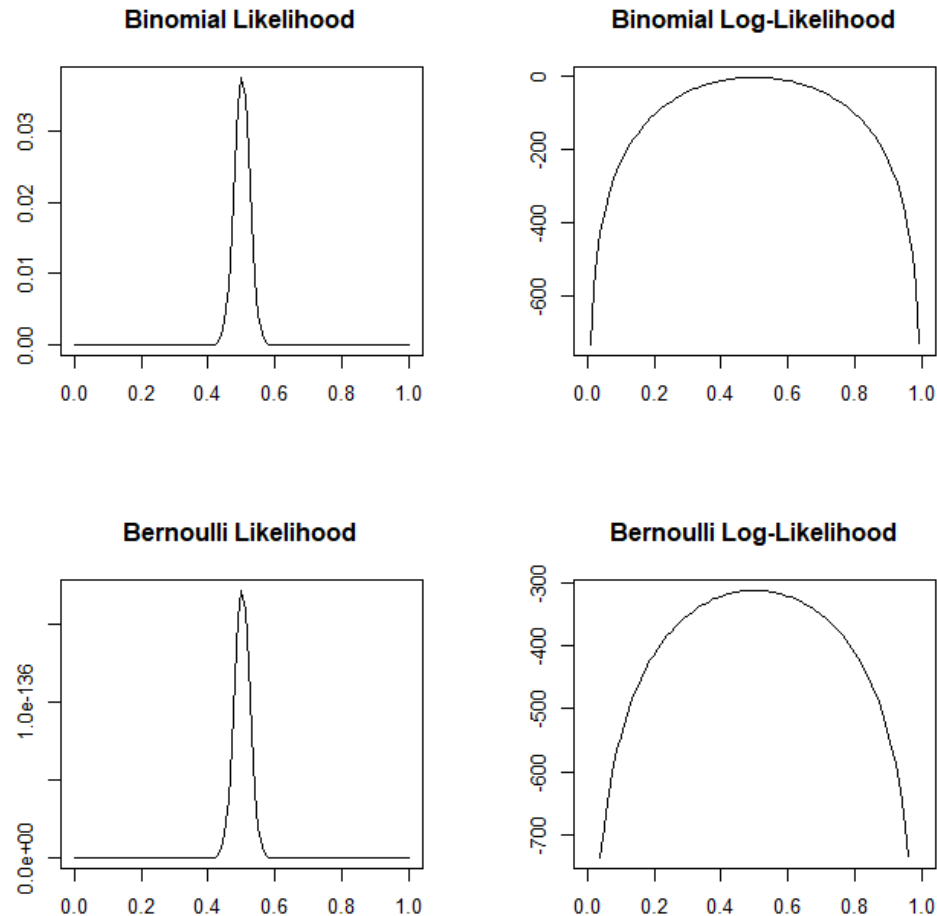- 451 individual Bernoulli coin flips, $x_i = 1$ for Trump, $x_i = 0$ for Biden

$$L_{ber}(p) = \prod_{i=1}^{451} p^{x_i}(1-p)^{1-x_i} = p^{226}(1-p)^{225}$$

- 451 trials, 226 "successes" (Trump voters)

$$L_{bin}(p) = \binom{451}{226} p^{226}(1-p)^{225}$$

- What matters for MLE and SE is *shape*, not *size*!

Binomial and Bernoulli Likelihoods

Finding the MLE...

- **If we use the Bernoulli likelihood,**

$$
\begin{aligned}
LL_{ber}(p) &= \log L_{ber}(p) \\
= \log p^k (1-p)^{n-k} &= k \log p + (n-k)\log(1-p)
\end{aligned}
$$

- **If we use the Binomial likelihood**

$$
\begin{aligned}
LL_{bin}(p) &= \log L_{bin}(p) \\
= \log \binom{n}{k} p^k (1-p)^{n-k} &\propto k \log p + (n-k)\log(1-p)
\end{aligned}
$$

- **Either way we want to maximize**

$$
k \log p + (n-k)\log(1-p)
$$

with k = 226, n=451

MLE: Point Estimate

■ Differentiating and setting to zero...

$$0 \quad = \quad LL'(p) \quad = \quad \frac{d}{dp} \left[k \log p + (n - k) \log(1 - p) \right]$$

$$= \quad \frac{k}{p} - \frac{n - k}{1 - p} \quad = \quad \frac{k - pn}{p(1 - p)}$$

■ so, clearly,

$$\hat{p} = \frac{k}{n} = \frac{226}{451} = 0.501$$

Bayes' Rule (a.k.a. Bayes' Theorem)

- A very simple idea with very powerful consequences

- We often start with information like P[A|B] and what we really want is P[B|A]. Bayes' Theorem lets us "turn the conditioning around":

$$\mathbf{P[B|A]} = \frac{P[A\&B]}{P[A]} = \frac{\mathbf{P[A|B]}P[B]}{P[A]}$$

- See https://arbital.com/p/bayes_rule/ for lots of examples and proselytizing.

Conditional probability & conditional density

- P[A|B] = P[A&B]/P[B]

- P[B] = P[B|A]P[A] + P[B|A^c]P[A^c]

- P[A & B] = P[B|A]P[A]

- Bayes' Theorem:

$$P[B|A] = \frac{P[A\&B]}{P[A]} = \frac{P[A|B]P[B]}{P[A]}$$
$$= \frac{P[A|B]P[B]}{P[A|B]P[B] + P[A|B^c]P[B^c]}$$

- f(x|y) = f(x,y)/f(y)

- $f(y) = \int f(y|x)f(x)dx$

- f(x,y) = f(y|x) f(x)

- Bayes' Theorem:

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{f(x|y)f(y)}{f(x)}$$
$$= \frac{f(x|y)f(y)}{\int f(x|y^*)f(y^*)dy^*}$$

Bayes' Theorem for Data

- Bayes' Theorem

$$f(y|x) \quad = \quad \frac{f(x,y)}{f(x)} \quad = \quad \frac{f(x|y)f(y)}{f(x)}$$

$$= \quad \frac{f(x|y)f(y)}{\int f(x|y^*)f(y^*)dy^*}$$

Dummy variable of integration

- Let x = data, y = θ (parameter!); then

$$f(\theta|\text{data}) \quad = \quad \frac{f(\text{data},\theta)}{f(\text{data})} \quad = \quad \frac{f(\text{data}|\theta)f(\theta)}{f(\text{data})}$$

$$= \quad \frac{f(\text{data}|\theta)f(\theta)}{\int f(\text{data}|\theta^*)f(\theta^*)d\theta^*}$$

Bayes' Theorem for Data

- ■ We call
  - ❑ f(θ) the *prior distribution*
  - ❑ f(data|θ) = L(θ) the *likelihood*
  - ❑ f(θ|data) the *posterior distribution*
- ■ So Bayes' Theorem says

$$f(\theta|\text{data}) \quad = \quad \frac{f(\text{data}|\theta)f(\theta)}{f(\text{data})} \quad \propto \quad f(\text{data}|\theta)f(\theta)$$

- ■ Slogan: (posterior) $\propto$ (likelihood)$\times$(prior)

Back to 2016 Florida pre-election poll

- The *likelihood* is the same as before:

$$L(p) \propto p^k (1-p)^{n-k}$$

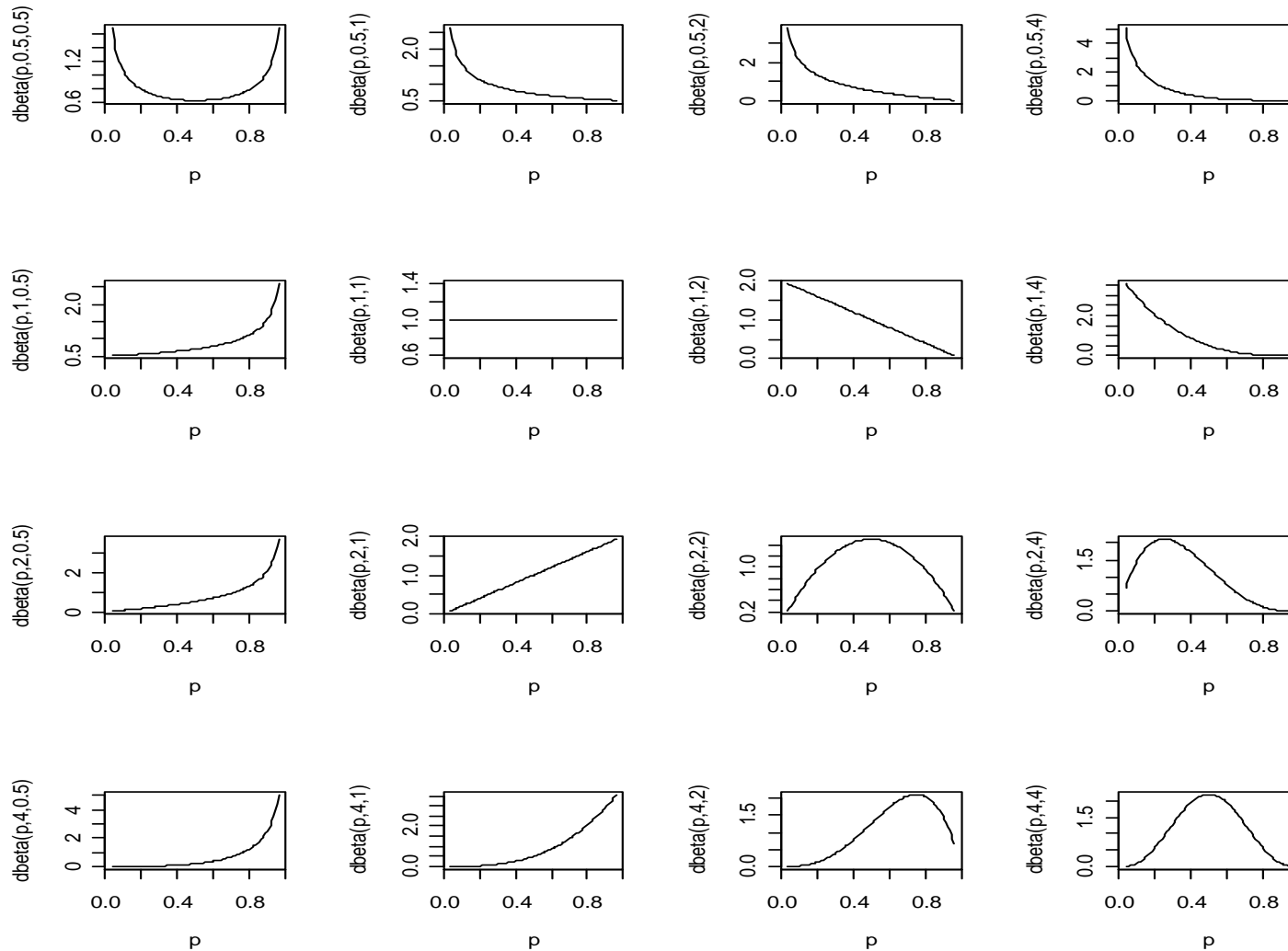- We need a *prior distribution*. One good choice is a *beta distribution*, with
  - Density $f(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$

  - Mean $E[p] = \frac{\alpha}{\alpha+\beta}$

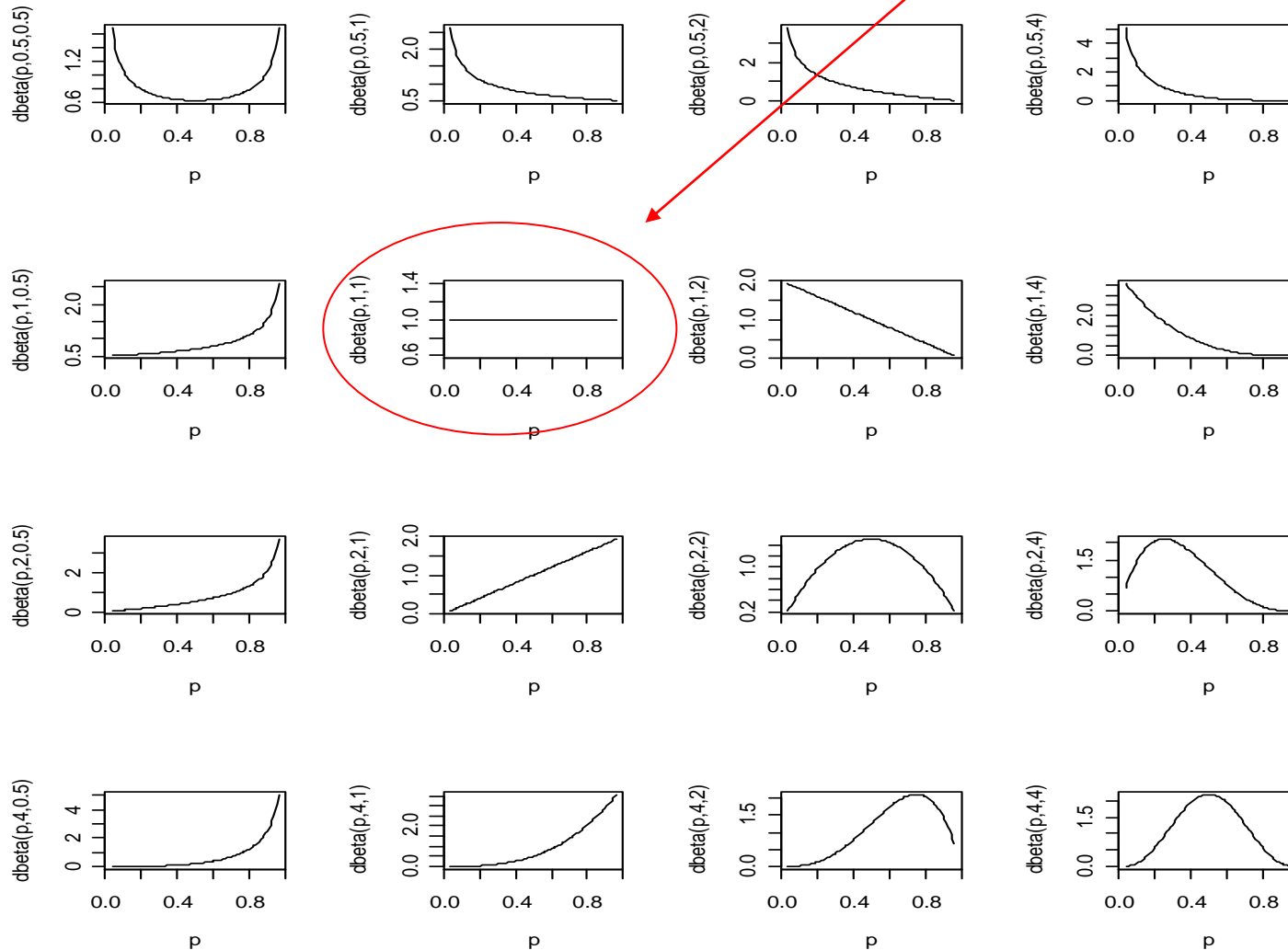 - Variance $\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

- Some graphs of beta densities appear on the next slide

Some Beta Densities

Some Beta Densities

uniform distribution!

Choosing prior parameters…

- The *likelihood* is the same as before:
$$L(p) \propto p^k (1-p)^{n-k} = p^{226}(1-p)^{225}$$

- The *prior distribution* is a beta distribution
$$f(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

  - $\alpha = 1$, $\beta = 1$ gives a uniform distribution – no preference for one p over another!

  - Suppose that in a previous poll, 942 prefer Trump and 1008 prefer Biden. Could set α=942, β=1008
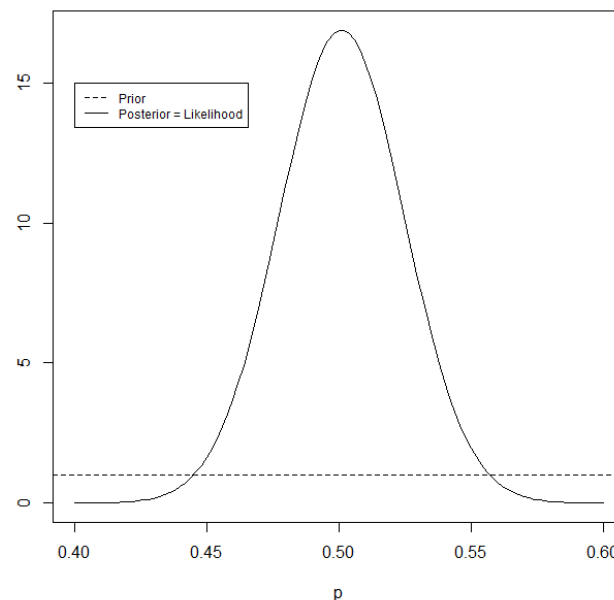
If α=1 and β=1...

- (posterior) $\propto$ (likelihood)$\times$(prior):

$$f(p|\text{data}) \quad \propto \quad L(p) \times 1 \;=\; p^{226}(1-p)^{225}$$

- Since f(p|data)=L(p), posterior mode = MLE = 226/451 = 0.5011086

- Since f(p|data) is a beta with α=227, β=226, E[p|data] = 227/453 = 0.5011038

If α=942, β=1008...
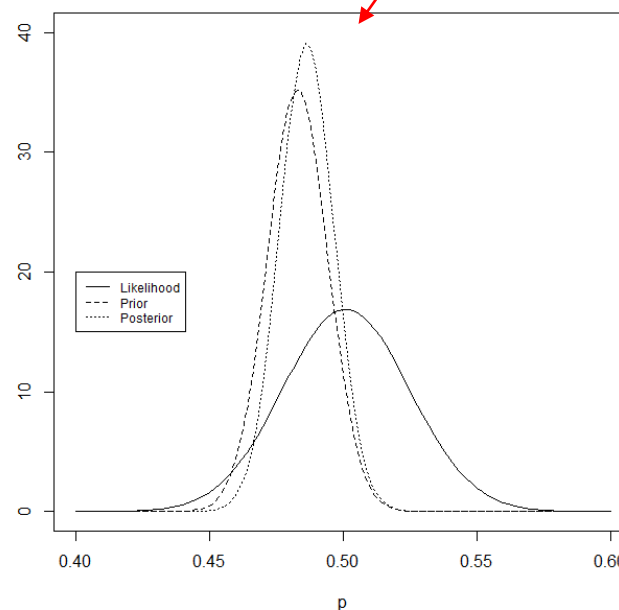
- (posterior) $\propto$ (likelihood) $\times$ (prior):

$$f(p|\text{data}) \quad \propto \quad L(p) \times p^{941}(1-p)^{1007}$$
$$= \quad p^{1167}(1-p)^{1232}$$

- Since f(p|data) = beta(p,1168,1233), E[p|data] = 1168/2401 = 0.486 vs MLE=0.501

Normal Model: Estimate μ, with σ^2 Known, One Observation y $\sim$ N(μ,σ^2)

- For our prior distribution, we'll assume $\mu \sim$ N(μ_0,$\tau_0{}^2$):

$$
\begin{aligned}
f(y|\mu) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \\
f(\mu) &= \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2}
\end{aligned}
$$

$$
f(\mu|y) \;\propto\; f(y|\mu)f(\mu) \;\propto\; \exp\left\{-\frac{1}{2}\left[\frac{(y-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2}\right]\right\}
$$

- Posterior must be normal for μ (quadratic in μ!); to identify it, complete the square…

- The exponent of $f(\mu|y)$ looks like -1/2 times

$$
\begin{aligned}
\frac{(y-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2} \quad &= \quad \frac{\tau_0^2+\sigma^2}{\tau_0^2\sigma^2}\left[\mu^2 - \frac{2y\mu\tau_0^2 + 2\mu\mu_0\sigma^2}{\tau_0^2+\sigma^2} + \frac{y^2\tau_0^2 + \mu_0^2\sigma^2}{\tau_0^2+\sigma^2}\right] \\
&= \quad \frac{\tau_0^2+\sigma^2}{\tau_0^2\sigma^2}\left[\left(\mu - \frac{y\tau_0^2 + \mu_0\sigma^2}{\tau_0^2+\sigma^2}\right)^2 + \mathsf{junk}(y,\sigma^2,\mu_0,\tau_0^2)\right] \\
&= \quad \frac{1}{\tau_1^2}(\mu-\mu_1)^2 + (\mathsf{known\ junk})
\end{aligned}
$$

so that $\mu|y \sim N(\mu_1, \tau_1{}^2)$, where

$$
\begin{aligned}
\tau_1^2 \quad &= \quad \frac{\tau_0^2\sigma^2}{\tau_0^2+\sigma^2} = \frac{1}{1/\sigma^2 + 1/\tau_0^2} \\
\mu_1 \quad &= \quad \frac{y\tau_0^2 + \mu_0\sigma^2}{\tau_0^2+\sigma^2} = \left(\frac{\tau_0^2}{\tau_0^2+\sigma^2}\right)y + \left(\frac{\sigma^2}{\tau_0^2+\sigma^2}\right)\mu_0
\end{aligned}
$$

- The exponent of $f(\mu|y)$ looks like -1/2 times

$$\frac{(y-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2} = \frac{\tau_0^2+\sigma^2}{\tau_0^2\sigma^2}\left[\mu^2 - \frac{2y\mu\tau_0^2 + 2\mu\mu_0\sigma^2}{\tau_0^2+\sigma^2} + \frac{y^2\tau_0^2 + \mu_0^2\sigma^2}{\tau_0^2+\sigma^2}\right]$$

$$= \frac{\tau_0^2+\sigma^2}{\tau_0^2\sigma^2}\left[\left(\mu - \frac{y\tau_0^2+\mu_0\sigma^2}{\tau_0^2+\sigma^2}\right)^2 + \text{junk}(y,\sigma^2,\mu_0,\tau_0^2)\right]$$

$$= \frac{1}{\tau_1^2}(\mu-\mu_1)^2 + (\text{known junk})$$

so that $\mu|y \sim N(\mu_1, \tau_1^2)$, where

$$\tau_1^2 = \frac{\tau_0^2\sigma^2}{\tau_0^2+\sigma^2} = \frac{1}{1/\sigma^2 + 1/\tau_0^2}$$

$$\mu_1 = \frac{y\tau_0^2+\mu_0\sigma^2}{\tau_0^2+\sigma^2} = \left(\frac{\tau_0^2}{\tau_0^2+\sigma^2}\right)y + \left(\frac{\sigma^2}{\tau_0^2+\sigma^2}\right)\mu_0$$

n Observations $y_i \sim N(\mu, \sigma^2)$

- Since

$$p(y_1, \ldots, y_n|\mu) \quad = \quad N(y_1, \ldots, y_n|\mu, \sigma^2) \quad = \quad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2}$$

$$\propto \quad N(\bar{y}|\mu, \sigma^2/n) \quad \equiv \quad \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{1}{2\sigma^2/n}(\bar{y} - \mu)^2}$$

we can apply the results for one observation

- $p(y_1, \ldots, y_n|\mu) \propto N(\bar{y}|\mu, \sigma_n^2)$, $\sigma_n^2 = \sigma^2/n$

- $p(\mu) = N(\mu|\mu_o, \tau_o^2)$

- $p(\mu|\text{data}) = N(\mu|\mu_n, \tau_n^2)$ where

$$\tau_n^2 \quad = \quad \frac{1}{1/\sigma_n^2 + 1/\tau_0^2} = \frac{1}{n/\sigma^2 + 1/\tau_0^2}$$

$$\mu_n \quad = \quad \frac{\bar{y}/\sigma_n^2 + \mu_0/\tau_0^2}{1/\sigma_n^2 + 1/\tau_0^2} = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}\right)\bar{y} + \left(\frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n}\right)\mu_0$$
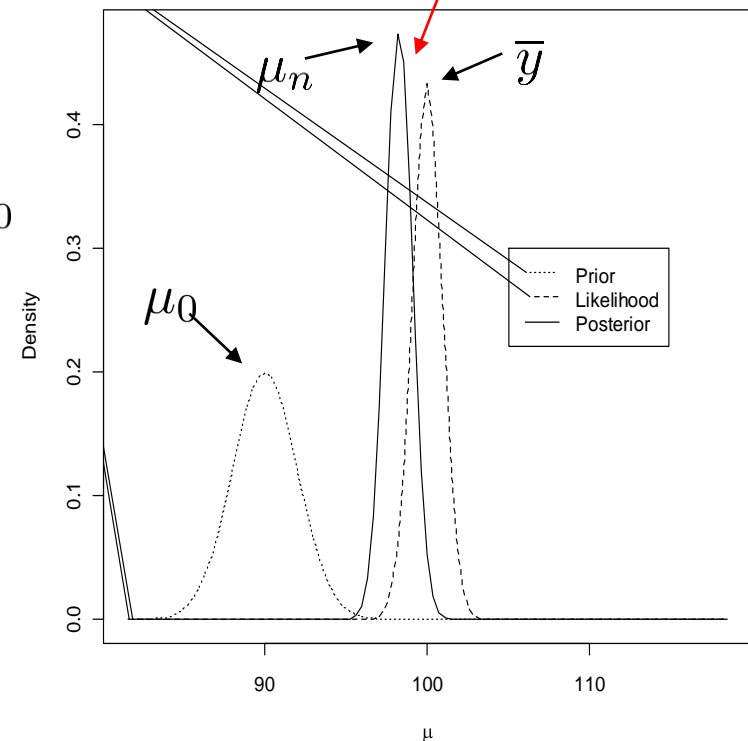
Normal Mean, Example

> "shrinkage": posterior between prior & likelihood

- Suppose we know σ=12, we look at n=169 IQ scores, and we find $\bar{y}$ = 100.

- We use as prior N(μ_0, $\tau_0{}^2$) with μ_0=90, $\tau_0{}^2$ = 4

- Shrinkage determined by

$$\mu_n = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n} \right) \bar{y} + \left(\frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n} \right) \mu_0$$

- $\dfrac{\tau_0^2}{\tau_0^2 + \sigma^2/n}$ is the *reliability*

- n larger $\Rightarrow$

  reliability larger $\Rightarrow$

  less shrinkage

Minnesota Radon Example

- **Emphasize Distribution Structure**

$$\text{Level 2:} \quad \mu_j \quad \overset{iid}{\sim} \quad N(\mu_0, \tau_0^2)$$

$$\text{Level 1:} \quad y_{j[i]} \quad \overset{indep}{\sim} \quad N(\mu_j, \sigma^2)$$

- **Emphasize Bayesian point of view**

$$\text{Prior:} \quad \mu_j \quad \overset{iid}{\sim} \quad N(\mu_0, \tau_0^2)$$

$$\text{Likelihood:} \quad y_{j[i]} \quad \overset{indep}{\sim} \quad N(\mu_j, \sigma^2)$$

- **Emphasize two-stage (multistage) sampling**

$$\mu_0$$

Mean radon across MN

$$\mu_1 \quad \mu_2 \quad \cdots \quad \mu_J$$

County-level differences
from grand mean

$$y_1 y_2 \cdots y_{n_1} \quad y_{n_1+1} \cdots y_{n_2} \quad y_{n_{J-1}+1} \cdots y_n$$

individual house levels

Minnesota Radon Example

$$\mu_0$$

Mean radon across MN

$$\mu_1 \qquad \mu_2 \qquad \cdots \qquad \mu_I$$

County-level differences
from grand mean

$$y_{11} y_{12} \cdots y_{1n_1} \quad y_{21} y_{22} \cdots y_{2n_2} \qquad y_{I1} y_{I2} \cdots y_{In_I}$$

individual house levels
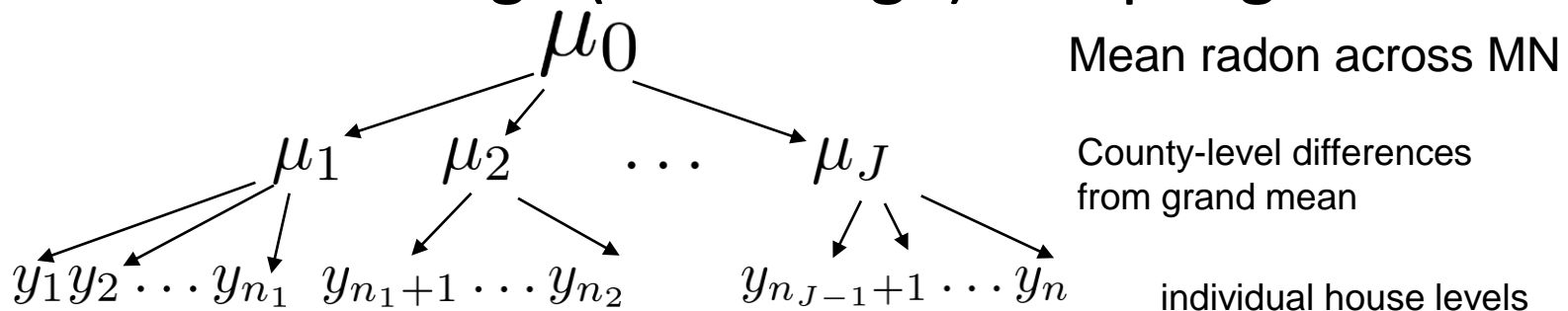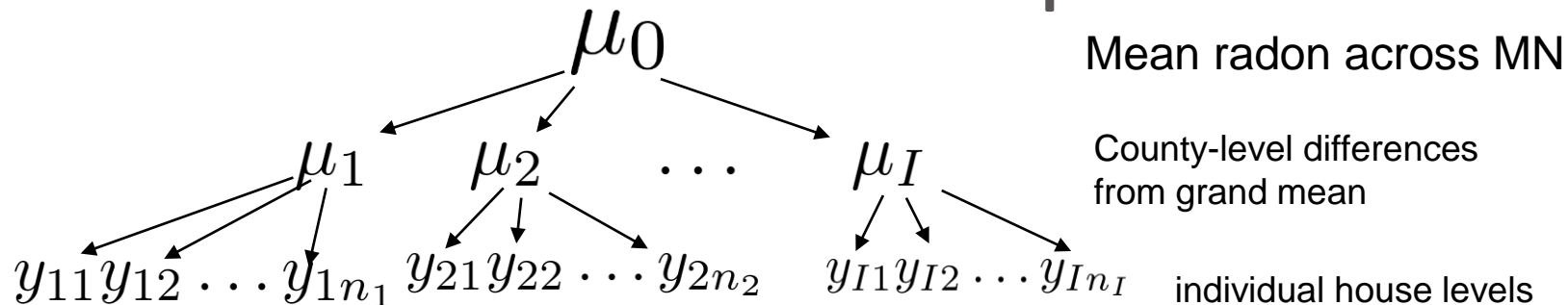
- In each county i with n_i houses, the posterior mean radon level $E[\mu_i | y_{i1}, \ldots, y_{in_i}]$ will be

$$\mu_i^{post} = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n_i} \right) \bar{y}_i + \left(\frac{\sigma^2/n_i}{\tau_0^2 + \sigma^2/n_i} \right) \mu_0$$

- When n_i large, $\mu_i{}^{post} \approx \bar{y}_i$
- When n_i small, $\mu_i{}^{post} \approx \mu_o$

Minnesota Radon Example

- **In the figure, the grand mean is μ_0**

- **In each county i with n_i houses, posterior mean is**

$$\mu_i^{post} = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n_i} \right) \bar{y}_i$$
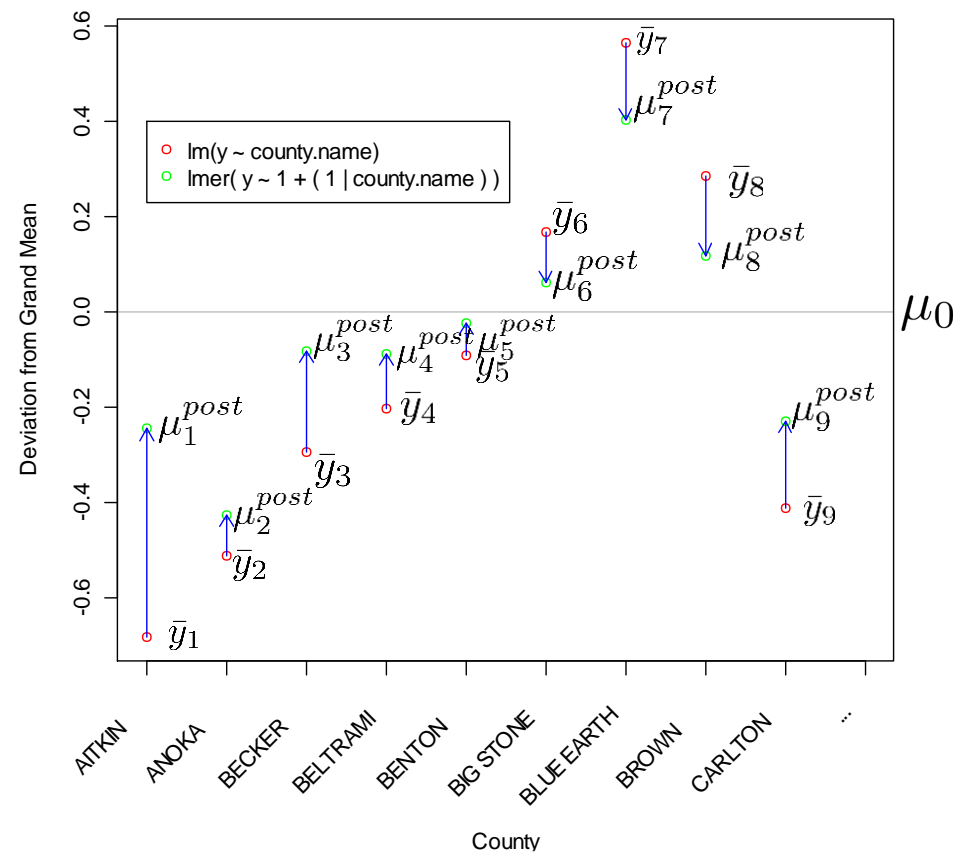$$+ \left(\frac{\sigma^2/n_i}{\tau_0^2 + \sigma^2/n_i} \right) \mu_0$$

- ❑ When n_i large, $\mu_i^{post} \approx \bar{y}_i$
- ❑ When n_i small, $\mu_i^{post} \approx \mu_0$

MLM's and Shrinkage

- The random effect "estimates" that `lmer` produces with `ranef()` are a form of posterior means $E[\eta|\text{data}]$ for each η.

- The posterior means $E[\eta|\text{data}]$ are always shrunk toward the prior mean 0, so that the random effects α are always shrunk toward the corresponding level-2 fixed effects β.

- The Bayesian pov not only provides insights, but also

    - Novel ways to expand the multi-level model framework

    - Simulation-based methods of estimation (MCMC with `jags`, Hamiltonian MC with `stan`, etc.)

(all for a different course focusing on MLM's & Bayes!)

Summary

- Today:
  - Shrinkage
  - Review of MLE
  - Crash course in Bayes
  - Normal-Normal Model & Shrinkage
  - MLM's and Shrinkage
- Mon:
  - Multilevel glm's (multilevel binomial, poisson, etc.)
- After Thanksgiving:
  - Maybe a little practical Bayes / MCMC