Stat 220 1a 9/5/12

Bayesian data analysis

central component is modeling

generative model is a story. observed data are realization from a probability dist.

God-like figure draws θ from an urn. within θ is the essence of how to draw y.

"God created the world in 7 days and we haven't seen much of him since."

vector of hyperparameters ϕ is specified or itself modeled.

inference, model checking, model improvement.

"People don't go around introducing you to their ex-wives." (why model improvement doesn't go into papers)

Bayes inference represented by matrix of posterior simulations

postprocessing

inference for qoi's - not necessarily θ

model checking

decision analysis

"The confidence interval can exclude 0, in which case I can submit it to a really good journal, or it can include 0, in which case I can look really hard and throw out some bad data points." (example of decision analysis)

PERC metabolism model

Goal: how much PERC metabolized at low doses

want population distribution (since some people are more susceptible than others)

Experimental data: expose 6 volunteers to PERC for 4 hrs, then measure concentrations in blood and air for 2 wks

4 compartment model of metabolism: well perfused tissues, poorly perfused, liver, fat

15 parameters per person.

plot looks like mixture of Expo decays. looks like Expo on log scale.

PERC released more slowly out of fat than out of well-perfused tissues.

Modeling

Sometimes model comes first, based on substantive considerations

Sometimes model chosen based on data collection (e.g. traditional statistics of surveys and experiments - "design dictates analysis")

Other times data comes first (e.g. protein folding. data-fitting.)

usually a mix.

Stat 220 1a

Four ideas that did not work.

- Fitting 4-compartment model directly to data. nonlinear Least squares.

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(y_{ij}-E(y_{ij}\mid x_i,\theta_j)\right)^2\frac{1}{\sigma_{ij}^{2}}$$

↑ ↑

individuals measurements

does not allow for time variation of parameters. (discussion of how we always know model is wrong)

fitting separately to each person is unstable b/c ≈30 data pts. for 15 params.

pooling data and estimating params for the "standard man" defeats goal of learning about population distribution.

"we always have prior information, of course. The only debate is how to use the prior information."

- Assisted model fit.

Set some parameters to fixed values from pharma literature, estimate the others.

couldn't fit the data well

↳ statistics tools for blind people.

using joystick to adjust parameters, with resistance indicating strength of prior.

" it would be cool even for people who aren't blind. I would use it. Just because blind people can do it.. I'm not too proud."

↓

difficult to get fixed (prior) values for PERC-specific parameters b/c not much prior lit.

"I'll get rid of the well-perfused tissues, get rid of poorly perfused tissues. All you are is just liver and fat. Probably true for some of you."

↳ what makes a good statistical analysis: other people can repeat what you did for other applications, and it becomes default. can do it with the click of a button.

"Statistics is said to be the science of defaults. one of our challenges is to defaultize things."

↓

1-2 compartment model

- 1 compartment $y = Ae^{-at}$

2 compartments $y = Ae^{-at} + Be^{-bt}$.

doesn't fit data well. most PERC leaves right away. but some stays in body after a week or more. w/ 2 compartments, can do slow vs. fast, but then fit is bad in the middle.

not realistic for low-dose extrapolation, which is our goal.

problem with poor-fitting model is inability to extrapolate.

Stat 220 1a

↳ thinking more about data collection.

"We don't think about x. We're all like y y y y y -- what about x? It's because it's not modeled: y is the data, not x. [someone leaves] Don't take it personally!"

↓

problem: least squares gives bad fit for low doses because we have more measurements from the beginning, where there are still high doses. these measurements are weighted too much.

however, we can't just downweight the undesirable points b/c that's ad hoc from a Bayesian perspective.

"how could that happen? it's <u>least squares</u>!" [said in forlorn, pleading voice].

"forget physicist-bashing. some<u>one</u> needs to design the new financial instruments tomorrow."

"that's not in the likelihood, that's not in the prior. where is it? it's <u>nowhere</u>!"

(why we can't assign weights as we please)

"we want to downweight these. but we can't, [jumping up and down] cause it's cheating!"

how to solve? good question.

↳ Survey sampling and causal inference are the same thing.

missing data: respondent vs. nonrespondent, treatment vs. control.

"in my head I have these simulations of Don Rubin and Jennifer Hill running on a loop."

"What Would Jennifer Do? What would Don Say?"

↓

here, can segment x axis, analogous to ~~stratification~~ subclassification in causal inference.

- Simulation from prior distribution.

get prior info on parameters from pharma lit, try to fit data within prior constraints.

but not much prior info for some params.

Bayesian model with hierarchical prior distributions

example: prior dist. for a rate parameter in the metabolism, θ_j for person j.

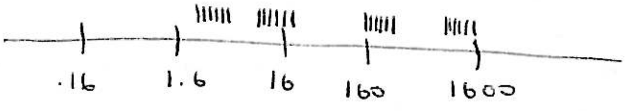
$\log \theta_j \sim N(\mu, \tau^2)$ pop. dist.

$\mu \sim N(\log 16, (\log 10)^2)$

$\tau \approx \log 2$

large uncertainty, small variation.

can learn about μ using data from several people. <u>use people to find out where μ is</u>. <u>pooling</u>.



only works with hierarchical prior because for any <u>given</u> person, they could be anywhere.

Stat 220 1a

example: parametrizing so independence seems reasonable.

A, B, C, D

$\theta_1 = \frac{A}{A+B+C+D}$, $\theta_2 = \frac{B}{A+B+C+D}$, ...

% of body mass in 4 compartments.

$\theta_1+\theta_2+\theta_3+\theta_4 = 1$.

trick:

$\theta_1 = \frac{e^{\varphi_1}}{e^{\varphi_1}+e^{\varphi_2}+e^{\varphi_3}+e^{\varphi_4}}$, $\theta_2 = \frac{e^{\varphi_2}}{\cdots}$

then put indep. Normal priors on $\varphi_1 \to \varphi_4$. ends up being 7 parameters (8 is overspecified).

induces ~~p~~ correlated priors on θ's.

called soft max.

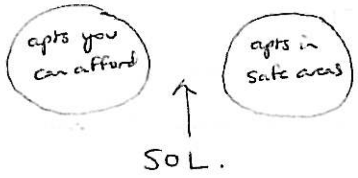
course website: http://stat.columbia.edu/~gelman/bda.course/

"This guy comes to me and says 'I have prior information and data, and I'd like to combine them, and I heard Bayes is a good way to do that.' Well, that's as good as it gets! Normally you want to do Bayes but they won't let you because they're like [in stupid voice] 'Ugh, it's subjective, I'm not allowed to, it's subjective.' But here this guy is saying 'I have prior information and data and I want to combine them'! I'm like, 'I can do that! I was trained to do that!'"

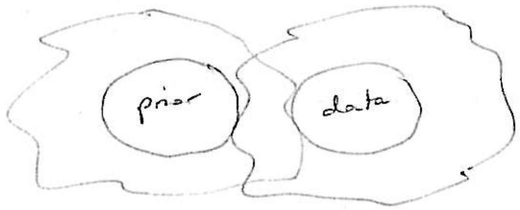
1b

9/7/12

want something that is consistent with model and data. but what if it's like finding an apt in NYC?



in our case, prior and data are probabilistic, so the boundaries are fuzzier.



"In statistics it's enough for our results to be cool. In psychology they're supposed to be correct. In economics they're supposed to be correct and consistent with your ideology."

back to example from last time: fit the model w/ Gibbs and Metropolis, re-ran model under hypothetical low-dose exposures.

scatterplot of inferences for 6 individuals: each point is draw from posterior dist.

uncertainty: within-graph variability.

variation: between-graph variability.

scale of uncertainty and variation are the same.

if variation were much higher than uncertainty (people far apart from each other), we would want to know why and would want to put it in the model. try to explain unexplained variation.

Stat 220 16

used model to predict future data

Sample theta from posterior dist, run model, add noise, rinse and repeat to get theoretical quantiles. plotted against actual data.

prediction is poor from 0-15 min. due to assumption of instantaneous mixing between compartments. but otherwise data fall within 10-90th quantiles.

putting it all together

(a) physiological model - a model where the parameters have <u>names</u>.

this is the best way to get prior info, combine info across diff. people.

$y = Ae^{-ax} + Be^{-bx}$ is a "phenomenological model"

A and B don't have meaning. difficult to choose prior for them

(b) hierarchical model - allow people to vary

without common population, not enough data to estimate params separately for each individual

but don't want complete pooling b/c there is a lot of variation even among healthy young male Dutch volunteers

(c) prior info

(d) data

need prior info on physiological params, data to learn about PERC in particular

(e) Bayesian inference

find params consistent with both prior and data, if possible

automatically includes uncertainty and variability.

(f) computation

(g) model checking - check reasonableness, consistency with prior, fit to current data, fit to new data.

"we need all g of these things. any f of them would not be enough."

with savvy parametrization, won't be much variation across people (example: use % of body mass in compartments instead of raw mass)

Bayesian inference automatically separates uncertainty and variation

"Sometimes classical statistics gives up [says there are probabilities we can't estimate].

<u>Bayes never gives up</u>. [...] so we're under more responsibility to check our models."

"you have to want to falsify the model. if you love somebody, set them free."

Stat 220 16

Ch. 1

Bayesian data analysis

"inference" too narrow - doesn't include model-checking

"statistics" too broad - includes design and data collection

wanted to call it "statistics using conditional probability." but that "wouldn't put the butts in the seats."

Bayesian: condition on data and prior information

What is Bayes?

data + regularization (rule out bad stuff)

data + prior info (add good stuff)

logical probabilistic reasoning ("not compelling, more an aspiration")

1. Probability and inference

Different approaches to statistics:

Traditional likelihood

Pure nonparametric (machine learning - no model)

Robust (econometrics - how well does least squares work under general smoothness conditions. no generative model for data)

Bayes (very model-dependent, must be willing to throw out model)

"A chicken is an egg's way of making another egg"
↓ ↓ ↓
inference model model

Goal is to use inferences from model to figure out what's wrong with the model and get a better model.

.1. Overview.

Three steps to Bayesian data analysis

Set up probability model

Inference

Model checking

"inference is the glamor boy."

Then go back and improve the model

.2. General notation for statistical inference

x (unmodeled data, like sample size) and y (modeled data)

Rubin's philosophy: all statistics is inference about missing data.

params are missing data - some observable, some inherently unobservable

In a world of prediction, what is the role of parameters?

Some say: parameters don't exist, only past data and future data

parameters allow <u>conditional independence</u>, making the model simpler

parameters are those things that persist

"reality is that which, when you stop believing in it, doesn't go away." - Philip K. Dick

Rubin's two questions:

1. What would you do if you had all the data?
2. What were you doing before you had any data? (i.e., what's your prior?)

Stat 220 1b

1.3. Bayesian inference

$N(\theta|\mu,\sigma^2)$ is the Normal PDF.

"Xiao-Li thinks our notation is better."

1.4. Example about spelling suggestions for "Kofee"

prior probabilities of coffee, Kofi, kofee depend on reference set you choose, how much information you put in. more information = better prior.

likelihood involves research too! to model the likelihood, may want to conduct experiments about who makes what typos, etc.

point: both prior and likelihood involve modeling choices.

1.5. Probability as a measure of uncertainty

frequency reference sets = Bayes probability.

1.6 and 1.7. Examples to support his argument that probability is empirical, like height or weight.

"No, it's inside the exp, you can't touch that."

Modeling using conditional probabilities

early 20th century: have data, find what dist. the data look like, and then learn about reality.

example: heights look like mixture of two Normals, % of boy births looks like $\frac{\mathrm{Bin}(n,\frac{1}{2})}{n}$.

late 20th century: regression modeling, conditional distributions.

21st century: hierarchical nonparametric modeling?

2a

9/12/12

Ch. 2.

2. Single-parameter models.

The basics:

- Data model (distinct from likelihood b/c different data models can have same likelihood)
- Prior density
- Posterior density

Why does likelihood come before prior?

suggested answer: usually you're trying to explain some data. Gelman: "usually you're trying to explain some data, you statistician."

- having likelihood tells you where you need to care about your prior.
- but prior beliefs do determine design and data collection

2.1. Estimating a probability from Binomial data

Example of bovine spongiform - Uniform prior seems inappropriate. if we observe 0 out of 75, is the posterior mean $\frac{1}{77}$?

really, we'd want to give an interval estimate $[0, \frac{3}{75}]$.

Dependence of θ and n: might need large n to estimate small θ.

2.2. Posterior as compromise between data and prior information

When is posterior variance higher than prior variance?

- bad luck. 70.1, 70.2, 69.9, 73.5, last observation brings the variance up.
- bad model. if prior is bad, you'll learn that from the data.

but even with bad model, posterior variance can still get smaller.

Stat 220 2a

Research problem: when do models have "warning lights"?

"There are two types of models. Good models, if they don't fit, you get a large standard error. Bad models, if the model doesn't fit, it goes... 'no problem.' [laughter] 'I have a great compromise for you.' The model isn't able to tell you it's bad." Need a name for this property.

Ex. 1 t can be viewed as mixture of Normals with unequal variances. Seeing an extreme observation is evidence of a high-variance Normal component, so t prior will inflate posterior variance when we have an extreme observation. but how much of this do we want? need to decide df of t.

"As you know from teaching introductory statistics, 30 is infinity."

Ex. 2. Acceptance/confidence region obtained by inverting χ^2 goodness-of-fit test. Consists of all quadratics not rejected by χ^2 test.

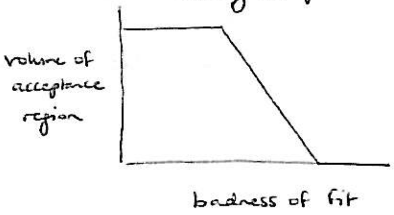
confidence region when model fits well.

everything doesn't fit. no curve is accepted by χ^2 test.

"It rejects everything, okay? The χ^2 acceptance region is the goddamn empty set."

"This leads to a scary situation where right before the model rejects everything, it gives you a very narrow confidence band.

Scary Graph



"So that's a bad thing."

Problem here is using the χ^2 test to get uncertainty statements and also to check fit of model.

↳ "Your computations are your inference." – Rubin.

If model is good but computations are wrong, the model you actually fit is the one you computed.

Conversely, "computation could save your butt." If model has bad mode but computer doesn't find it, it's as if it isn't there.

2.3 Summarizing posterior inference

Central vs. shortest intervals

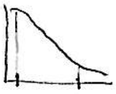
- same for symmetric distribution. for one-sided dist., central interval is silly because it excludes 0, the most likely value. in general, shortest interval is drawn toward the region with highest posterior density.
- central interval invariant to transformations.
- shortest interval harder to compute. empirical shortest interval is a little bit biased and variable "because it's the winner in a noisy competition." noisier than central interval.

Stat 220 2a

Goals of posterior intervals.

- sometimes want nested intervals: 50% inside 80% inside 95%. won't always happen with multimodal distributions, so has to be requirement of procedure.
- two purposes of interval estimation

accept what's inside } not the same thing.
reject what's outside



accept what's inside because it is a positive summary of where params are likely to be, but don't reject 0 just b/c it's outside the interval.

Finding shortest posterior intervals: combination of bootstrapping and smoothing works best. but computational tradeoffs - get more simulations if you can!

2.4. Informative prior distributions

Interpretations

- population - "the urn full of thetas"
- state of knowledge - represents our assumptions
- software defaults (statistician in a box)

"conceptually there is a true prior." true prior is *not* true subjective belief, but rather the dist. of all possible Θs for which you would apply a given statistical procedure. it's "behavioral rather than cognitive."

"we can imagine there's some sort of thetabase."

Bin model: is Beta(α, β) prior equivalent to $\alpha+\beta$ data points, or $\alpha+\beta-2$?

Probability of a girl birth

P(girl birth) = 0.485 in general population. compensates for the fact that at every age, boys die more than girls. 55% boys at embryo stage, then equal sex ratio at age 20 "which I'm told is convenient", "and then eventually you've got grandma by herself..."

variation in sex ratio completely explainable by Binomial variation.

Constructing a prior distribution

believe p should be between 0.47 and 0.50

"but that's a soft upper bound. we prefer soft power. we're like Bill Clinton; we're not like George W. Bush."

Beta(α, β) prior with mean 0.485 and SD 0.01 turns out to correspond to $\alpha+\beta = 2500$ data points, whereas $n = 980$, so prior is very strong.

Stat 220 2b 9/14/12

P(girl) for beautiful vs. ugly parents

Data: difference in P(girl) estimated from 3000 respondents

.08 ± .03 (selected comparison)

.047 ± .043 (linear regression)

Prior: $N(0, 0.003^2)$.

two reasons for prior zero mean:

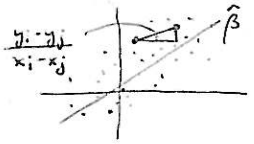
1. don't want to bias results
2. no prior reason to think beautiful parents have more girls

prior variance loosely based on previous research ("we need a probabilistic model of the scientific process")

Equivalent sample size

consider survey with n parents, compare sex ratio of prettiest $\frac{n}{3}$ to ugliest $\frac{n}{3}$.

$$\hat{\beta} = \frac{\sum_{i,j} (x_i - x_j)^2 \frac{y_i - y_j}{x_i - x_j}}{\sum_{i,j} (x_i - x_j)^2}$$



Further-apart comparisons get more weight.

Shouldn't compare top half and bottom half b/c points in middle give noise, no leverage.

more statistically efficient to do upper 3rd / lower 3rd. get rid of noise.

SE is $\sqrt{\frac{.5^2}{\frac{n}{3}} + \frac{.5^2}{\frac{n}{3}}} = .5\sqrt{\frac{6}{n}}$

equivalent info: $0.003 = .5\sqrt{\frac{6}{n}} \Rightarrow n = 166000$.

So survey with 166000 people would be equally weighted with prior.

2.6. Normal mean, known variance

$\theta \sim N(\mu_0, \tau_0^2)$, $\bar{y} | \theta \sim N(\theta, \sigma^2/n)$

$$\theta | y \sim N\left(\frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \right)$$

2.7. Other standard single-param. models

Poisson with exposure almost never do $y_i \sim \text{Pois}(\theta)$, instead do $y_i \sim \text{Pois}(\overset{\text{exposure}}{x_i}\,\overset{\text{rate}}{\theta})$

2.8. Estimating kidney cancer rates

low-population counties shrunk more to prior mean.

Parable of the two exams:

Option 1: 100 questions

Option 2: 1 question. Score is 0 or 100.

when hiring, go with option 1. that's like the Bayes estimate.

low-population counties "don't get a chance to show their stuff." we don't "hire" (believe) them.

Estimating prior dist. from data

prior is population distribution - represents true kidney cancer death rates. data more variable because noisy.

For any given county, prior is supposed to represent my knowledge of the rate in that county.

match mean of rates to mean of black curve

variance of rates to variance of black curve + average Poisson sampling variance.

Stat 220 2b

Connecting cancer rate example to Bayesian themes

- Model $y_j \sim \text{Pois}(10 n_j \theta_j)$

  assumes independence between individuals

  independence between counties, conditioned on θs

  "you can't die twice of something. it's double jeopardy."

  likelihood based on Poisson approximation to Binomial

  Poisson actually has fewer assumptions because θ's for Binomial have to be the same for people.

  for Poisson, can have different probabilities and interpret θ as average.

- Informative Gamma prior: equivalent sample size (in each county) of 20, prior mean of $\frac{20}{430000}$ ← dead people ← total

  county with 430 000 people has posterior mean halfway between prior and data

- One prior, many different datasets (each county as its own dataset)
- Data-based prior distribution

Comparing Bayesian inferences to other estimates

- Do parameter estimates look reasonable?
- Do predictions look reasonable?
- Artifacts.

  raw rates lead to artifacts - highest rates all in low-pop. counties.

  posterior means have artifacts too - highest rates all in high-pop. counties.

Cross-validation

- Take 1st half of 80s. Use 3 estimators (raw rates, Bayes' estimate, prior mean) to predict 2nd half of 80s.

  Bayes does best.

- What if data not broken down by year?

  $n = 200{,}000$, $y = 30$. artificially break up into y_1, y_2 with $y_1 \sim \text{Bin}(30, \frac{1}{2})$

  if Poisson model is true, this is exactly what it assumes! so can prove theorem that Bayes does best under mean squared error, assuming model is true and using this cross-validation.

  ⇒ Bayesian inference as implicit cross-validation.

  raw rate works best if y_1, y_2 always half and half, but that's not what Poisson model assumes.

2.9. Noninformative prior distributions

"Another way of saying a prior is proper is that it's a generative model. You can use it to generate data."

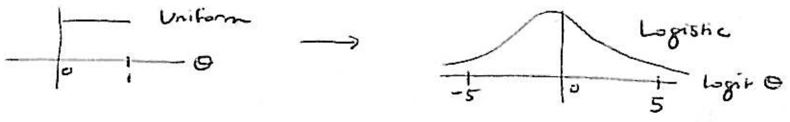
An improper prior is not a generative model, either because it

1. doesn't integrate to 1, or because it
2. depends on the data. need data to get the prior, so can't generate data with it.

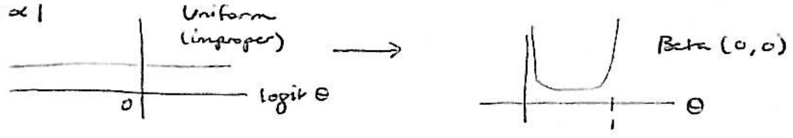
In Bayes, you can set prior based on the functional form of the data, but not the realization of the data themselves.

"Noninformative" depends on scale:

(a) $p(\theta) = 1$



(b) $p(\text{logit}\,\theta) \propto 1$



Stat 220 2b

2.10. Weakly informative priors

WIP: prior that contains some info (is proper) but less than you really have.

recall the idea of the unknown true prior $p_{true}(\theta) = N(\theta | \mu_0, \sigma_0^2)$

assumed prior (subjective prior) is $p_{subj}(\theta) = N(\theta | \mu_1, \sigma_1^2)$

WIP inflates variance by factor of k^2, $k > 1$

need to formalize the idea of using less info than is available.

3a 9/19/12

Last problem on hw 2: "You weren't supposed to be particularly Bayesian here, except in the sense that anything reasonable is Bayesian."

Deciding whether to keep covariates in a regression model: ("Jennifer wouldn't let me put this in the book")

	expected sign	unexpected
Stat. Significant	✓	!
not	keep	○

— Sign of some interaction in the underlying process.
"women are making more than men. that's wrong! oh wait."

Ch 3.

3. Introduction to multiparameter models

3.1. Averaging over nuisance parameters

"Suppose there's someone you want to get to know better, but you have to talk to all her friends too. They're like the nuisance parameters."

$p(\theta_1|y)\, p(\theta_2|\theta_1, y)$ "First I have to learn θ_1. Then I can go in for the kill."

Def. of nuisance depends on context. Example: purchaser/user of a scale cares about mean, manufacturer cares about variance.

3.2. Normal data with noninformative prior

$y_1, \ldots, y_n \sim N(\mu, \sigma^2)$

prior $p(\mu, \sigma^2) \propto \sigma^{-2}$

equivalent to $p(\mu, \sigma) \propto \sigma^{-1}$, $p(\mu, \log \sigma) \propto 1$

integrate out μ from joint posterior density: $\sigma^2 | y \sim \text{Inv-}\chi^2(n-1, s^2)$

$\mu | (\sigma^2, y) \sim N(\bar{y}, \frac{\sigma^2}{n})$.

3.3. Normal data with conjugate prior

$y_1, \ldots, y_n \sim N(\mu, \sigma^2)$

Conjugate family

$\sigma^2 \sim \text{Inv-}\chi^2$

$\mu | \sigma^2 \sim$ Normal with variance proportional to σ^2

μ and σ dependent in prior - says that the bigger the variance, the less you know about the mean

"can you forget the last five minutes? thanks. including the part where I said to forget the last five minutes."

use conjugate priors for understanding (prior as extra data) and computation reasons.

Stat 220 3a

3.4. Multinomial model

$y_1, \to y_K \sim \text{Mult}(n, \theta_1, \to \theta_K)$

unknown probabilities $\theta_1, \to \theta_K$ constrained to unit simplex, so can't be independent.

noninformative priors:

$\theta_1, \to \theta_K \sim \text{Dir}(1, \to 1)$ uniform on θ_j's

$\sim \text{Dir}(0, \to 0)$ uniform on $\log \theta_j$'s

$\sim \text{Dir}(\frac{1}{K}, \to \frac{1}{K})$

Example of joint prior on regression coefficients

$y = \beta_{j0} + \beta_{j1} x_1 + \beta_{j2} x_2 + \beta_{j3} x_3 + \text{error}$ each group has 4 coefficients

"why is it Normal? because that's the only continuous multivariate distribution we have. oh, we have the multivariate t. as if that's a different distribution."

3.6. Multivariate Normal with unknown mean and variance

good prior of cov matrix is hard b/c of dimensionality, positive-definiteness constraint.

"this is a paper we have, that's making the rounds of getting rejected."

3b 9/21/12

"It's like the joint distribution is a movie, and all you care about is the star, like Robert Downey Jr. or whatever."

"It's like you have a big network of variables, and you grab the one you care about, and you shake it and shake it."

On birthday frequencies: "Well of course there's a reason – it's not like the baby doesn't want to come out."

3.7 Example: analysis of bioassay experiment

$y_i | \theta_i \sim \text{Bin}(n_i, \theta_i)$, $i = 1, \to 4$

assumptions: independence within and between groups, conditional on θ_i.

same θ_i for each rat within a group.

if rats are fighting for food, underdispersion. if contagion, overdispersion.

"In cage 1, they all die, and then in cage 2 they all hear about it, and they're like, 'Don't eat that shit, man.'"

$\theta_i = \text{logit}^{-1}(\alpha + \beta x_i)$, $i = 1, \to 4$, $p(\alpha, \beta) \propto 1$.

assumptions: monotonicity of dose-response.

in particular, follows logistic curve

θ_i deterministic function of α, β

θ_i between 0 and 1

no group-level error

α and β "could be anything"

information in logistic regression is (how much news) × (probability of news). at the extremes, the probability is too small, and right in the middle, the news is nothing. So the most information is in the intermediate points.

↳ predicting presidential elections: only 1980, 1988, 1992, 1996, 2004, and 2008 count.

the others are "ties" or "obvious"

↲

posterior dist. for β is asymmetric: β could be really large, but we're sure of the sign (positive).

GLM vs. Bayes: GLM overfits but gets a really good fit in this case "by luck".

Stat 220 3b

α and β are positively correlated a posteriori because the data happen to be such that the curve pivots around a negative value of x. Since α is the intercept $(x=0)$ and higher β means higher intercept, we get positive corr. but the pivot pt. is close to 0, so corr. is weak.

Should parametrize so intercept is meaningful : " if your height is 0, you're not going to make any money."

↳ against artificial histogram smoothing :

"I disagree with those people. Those straw people."

"The best histograms have the seeds of their own destruction."

Difficulties with ratios as qoi's when denom. can be positive or negative.

LD50 = $\frac{-\alpha}{\beta}$. moving from $\beta > 0$ to $\beta < 0$ doesn't actually correspond to a logical sequence of models.

this is not a continuous logical progression.

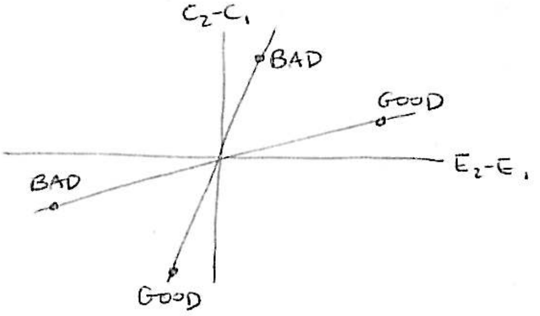
"they're two completely different models that happen to be connected by a common parametrization."

Incremental cost-effectiveness ratio

old treatment : cost C_1, efficacy E_1.

new treatment : cost C_2, efficacy E_2.

$\frac{C_2 - C_1}{E_2 - E_1}$ incremental cost-effectiveness ratio



Same cost-effectiveness ratio in quadrants I and III, with completely different interpretations. meaningless qoi.

"we economized. people died. and we didn't even save money!"

Instrumental variables

IV estimate: $\frac{\text{coef of regression of } y \text{ on } I}{\text{coef of regression of } z \text{ on } I}$

need to restrict denom to be positive and admit you're proceeding under that assumption

Feller - Creasy problem

$x_1, \to x_n \sim N(\theta_x, \sigma^2)$, $y_1, \to y_n \sim N(\theta_y, \sigma^2)$, hard to get interval estimates for $\frac{\theta_y}{\theta_x}$ with good coverage.

but what does $\frac{\theta_y}{\theta_x}$ mean if θ_x can be positive or negative?

folk theorem of computational stat : when you have computational problems, often there's a problem with your model.

Pinocchio principle : a model created solely for computational reasons can take on a life of its own.

Stat 220 4a 9/26/12

work with log density instead of actual density

"raise your hand if you haven't heard this principle before. hey, you heard a new principle!"

computing posterior density on a grid

Compute unnormalized log density on grid — log.post

rescale and exponentiate — a <- exp(log.post - max(log.post))

normalize to sum to 1 — a / sum(a)

Chicken brain data

"I called them and they told me to fuck off, basically."

Sham treatment can be thrown out after it's determined that sham has no effect

Est. 1 $\bar{y}_1^{\text{exposed}} - \bar{y}_0^{\text{exposed}} - (\bar{y}_1^{\text{sham}} - \bar{y}_0^{\text{sham}})$

Est. 2 $\bar{y}_1^{\text{exposed}} - \bar{y}_0^{\text{exposed}}$

Est. 3 $\bar{y}_1^{e} - \bar{y}_0^{e} - \lambda(\bar{y}_1^{s} - \bar{y}_0^{s})$

Ch. 4.

4. Large-sample inference and frequency properties

Normal approximations to posterior dist.

Large-sample theory

Counterexamples to the theorems - "it's not a theorem until you have counterexamples"

Frequency evaluations of Bayesian inferences

4.1 Normal approximation

$$\log p(\theta|y) = \log p(\hat{\theta}|y) - \frac{1}{2}(\theta-\hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta}) + \ldots$$

Normal centered at $\hat{\theta}$ with inverse variance

4.2 Large-sample theory

data $y_1, \to y_n \overset{\text{i.i.d.}}{\sim} f(y)$, modeled as $p(y|\theta)$, which may not contain f

as $n \to \infty$:

- if θ discrete and finite, $p(\theta|y) \to$ point mass at true θ or model closest to f in K-L distance
- if θ continuous on compact set, $p(\theta|y) \to$ point mass at true θ or closest model
- under some conditions, $p(\theta|y)$ approaches Normal dist.

4.3 Counterexamples to the theorems

unidentified parameters ($y = \theta_1 + \theta_2$)

model changing with sample size

unbounded likelihood, e.g. mixture model

Solution is to bound the variance ratio or the actual variance

there's a sense in which the mixture model includes a class of models we're not interested in, as with the cost-effectiveness ratio

improper posteriors

Constrained priors

boundary estimates

tails - the further out in the tails, the more data you need to fit it well.

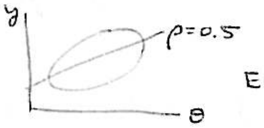
Stat 220 4a

4.4 Frequency evaluations of Bayesian inferences

example where unbiasedness doesn't make sense:

θ = height of a woman (inches)

y = height of her adult daughter



$E(y|\theta) = 64 + 0.5(\theta-64)$

$\hat{\theta}(y) = y$ is biased. bias $= -0.5(\theta-64)$

$\hat{\theta}(y) = 64 + 0.5(y-64)$ is really biased. bias $= -0.75(\theta-64)$ "this is big-ass bias."

unbiased estimate is $\hat{\theta}(y) = 64 + 2(y-64)$: need to anti-shrink!

classical statistician says θ is not a parameter; it's a "predictive quantity" b/c it has a distribution.

in predictive inference, don't condition on z. bias of prediction is $E(\hat{z}|\theta) - E(z|\theta)$.

4b

Ch 5.

9/28/12

5. Hierarchical models

rat tumor example: want more precise estimate of P(tumor) at dose of 0.

5.1 Constructing a parametrized prior dist.

$y_j \sim \text{Bin}(n_j, \theta_j)$, $j = 1, -, 71$, $\theta_1, \rightarrow \theta_{71} \sim \text{Beta}(\alpha, \beta)$

5.2 Exchangeability

$\theta_1, \rightarrow \theta_J$ exchangeable if $p(\theta_1, \rightarrow \theta_J)$ symmetric.

nonexchangeable models:

- time effects

$\hookrightarrow$ $\theta_j = \text{logit}^{-1}(\alpha + \beta x_j + \epsilon_j)$, x_j = date of study j.

no good to write $\alpha + \beta j + \epsilon_j$; index shouldn't convey information.

"having the index convey information is like having the lamp hang from the wire ... you don't want to use an electrical connection as a mechanical connection."

$\swarrow$

- knowing some experiments are in different settings
- Markov chain*

"exchangeability is a function not just of reality, but of the information you have"

Say we know the 70 previous experiments are different from the new one.

exchangeability violated. use above model, with x_j the indicator of the new experiment.

old: $\alpha + \epsilon_1, \alpha + \epsilon_2, -, \alpha + \epsilon_{71}$.

new: $\alpha + \beta + \epsilon_{71}$.

not much info about β in the likelihood.

if prior on β is noninformative, then not using the others to help with 71. if more informative, then will pool.

- group-level predictors - many nonexchangeable models fall into this category.

Stat 220 4b

5.3 Fully Bayesian analysis of conjugate hierarchical models

$p(\phi, \theta | y) \propto p(\phi)\, p(\theta | \phi)\, p(y | \theta, \phi)$

marginal posterior of hyperparams

$$p(\phi | y) = \int p(\phi, \theta | y)\, d\theta$$

$$\propto p(\phi) \int p(\theta | \phi)\, p(y | \theta, \phi)\, d\theta$$

if we can do this integral, then just need to compute $p(\phi | y)$ on grid of ϕ,

sample ϕ^s from grid and then sample θ^s from $p(\theta^s | \phi^s, y)$.

Rat tumor model: algebra

Model: $y_j \sim Bin(n_j, \theta_j)$

$\theta_j \sim Beta(\alpha, \beta)$

assumptions: independence conditional on θ, exchangeability

θ could have come from a Beta! (functional form assumption - not bimodal)

is exchangeability problematic? it's just inference conditional on no information distinguishing the θ_j. if you have info, exchangeability just ignores the info.

prior on (α, β).

Uniform prior on $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ doesn't work (improper posterior)

Unif over large box doesn't work either - mass pulled toward edges of box

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

go from joint posterior $p(\theta, \alpha, \beta | y)$ to marginal posterior $p(\alpha, \beta | y)$ so can compute on 2D grid

need to work hard for good parametrization so that we can come up with a good prior

↳ back to time effects - should measure x_j in years / decades relative to average time of experiment.

↰ otherwise intercept "will be the logit probability of death in the year that Jesus was born"

partial pooling plot: posterior mean of θ as function of observed rate

no pooling is 45° line, complete pooling is horizontal line, our model is a compromise.

5.4 Exchangeable parameters from Normal model

Model: $\bar{y}_j \sim N(\theta_j, \sigma_j^2)$

$\theta_j \sim N(\mu, \tau^2)$

Second statement is more restrictive because assumes θ_j can be modeled as Normal - no reason that this should be true.

get conditional posterior $\theta | \mu, \tau, y$, average over marginal posterior of μ, τ.

5.5 Example: parallel experiments in 8 schools

8 schools, 60 students in each, randomized to exam coaching or no coaching

Separate estimates unsatisfactory b/c if experiments had been replicated, would not expect school A to get 28 again.

pooled estimate also feels unsatisfactory b/c would want to send kid to school A over school C

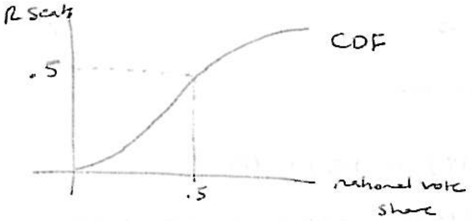
Case for partial pooling.

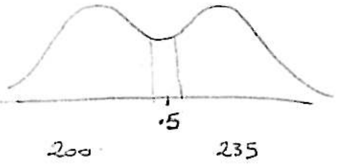
Stat 220 5a

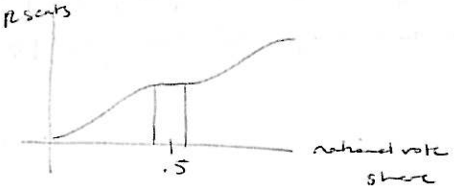
10/3/12

2J parameters, half from $N(1,1)$ and half from $N(-1,1)$. if it were a mixture of i.i.d. components,

the # in each mixture should be a r.v., not fixed.







large swing in votes
around .5 doesn't
affect # of seats

bootstrapping doesn't make sense because # in each mixture component is pretty much fixed.

instead, get standard errors by moving things side to side a little bit.

"The Alexes said you guys had a lot of bad models. They blamed themselves. And I blame them too."

" It's not like the door is open. It's like where is the door? I can't even see the wall. Maybe this describes most of my research."

Sweet spot for statistics:

- lot of data - don't need fancy methods.
- no data - can't do anything.
- in between - sweet spot.

Hierarchical models continued.

8 schools example:

- data consistent with $\tau = 0$
- Rubin's approach: draw from posterior of τ and then average over the different draws.

what if model were applied to 8 unrelated objects?

- partial pooling wouldn't happen because τ would be huge.
- if exchangeability is inappropriate, don't do it.
- can fit mixture model (7 schools + 1 speed-of-light observation, haha)

dependent prior on μ, τ would shrink μ toward 0 for low τ

5.6 Hierarchical modeling applied to a meta-analysis

transformation to log odds, then apply 8 schools model

3 levels of inference:

- avg effect
- effect in a single study (existing or new)
- prediction for a new person (in existing or new study)

Stat 220 5a

$$\theta \qquad\qquad \frac{1}{\sqrt{n}}$$

Why are doses more effective in experiments than real world?

1. researchers chose subpopulation that is most likely to benefit.
   Less expensive to multiply θ by 2 than n by 4.
2. actual dose is higher in experiment because of better compliance, controlled conditions, etc.
3. statistical significance filter.

5.7 WIP for hierarchical variance parameter

$\log \tau \sim \text{Unif}(-\infty, \infty)$ leads to improper posterior

"you use this sucker - I should say 'this sucka' - you're gonna die."

$$\left.\begin{array}{l}\tau \sim \text{Unif}(0, \infty)\\ \tau \sim \text{IGam}(10^{-3}, 10^{-3})\end{array}\right\}\text{problems}$$

$\tau \sim \text{Cauchy}^{+}(0, A)$

IGam(1,1) prior cuts off (dead zone near 0), and IGam$(10^{-3}, 10^{-3})$ is much too strong.

Unif prior with 3 schools doesn't cut off long tail. ("we want to cut off the tail. we're not dogs here.")

won't shrink a lot, and moreover, posterior predictive has huge variance

implications:

1. to make inference from 3 schools, you really have to use prior info.
2. prior info, even a little bit, can help.

5b

Ch. 6 10/5/12

6. Model checking

Why wouldn't we put all our prior info into the model?

1. Sometimes hard to express prior info in probabilistic model
2. could add too many parameters, make model more complicated
3. concerns about human psychology - don't want to fool ourselves.
   example: prior for ESP experiments.
4. statistics as science of defaults. might want something that everyone can use.
5. ☆ not wanting to cheat. inference as part of a larger process.
   there is a whole class of problems where you don't want to put down what you believe the most.
   - census data - need to preserve anonymity.
   - sports tournament - can't award victory based on prior belief
   - assigning grades - presumably we want to award grades based on actual understanding of material, and final exam is a noisy measure, so should we add the pretest score into the model?

   What these examples have in common is that the inference is embedded within a larger societal or scientific process, which necessitates different rules. we care about fairness, replicability of results, etc.

→

Stat 220 5b

"A Bayesian wants everyone else to be non-Bayesian" – if every instructor used Bayesian inference to assign grades, my likelihood function as someone who wants to hire is tremendously complicated.

if you have to save the world or humanity will be extinguished, use your prior. but normally it's embedded in a process, entangled with social goals, so there may be reasons to omit prior info.

"if I'm doing an experiment to save the world, I better use my prior."

model checking is:

comparing estimates and predictions to substantive knowledge.

comparing predictions to observed data (your original data can test your model)

graphical and numerical tests

6.1 Model checking in applied Bayesian statistics

With great power comes great responsibility

example: with only 3 schools, should you just give up? can't estimate group-level variance

Sensitivity analysis

not just sensitivity to the prior, but sensitivity to the likelihood (which enters n times!)

Gelman on sensitivity analysis: "I just never get around to it."

↳ paper on the boxer, wrestler, and coin flip.

two r.v.s: indicator of heads in a coin toss and indicator of who wins in a boxer vs. wrestler fight to the death.

different procedures do change your inference.

↓ if you were fully committed to sensitivity analysis, you would consider models

$p_1(\theta)$, $p_2(\theta)$, $p_3(\theta)$, ... which could be expressed as $p(\theta | \varphi)$,

but then you put a prior on $p(\varphi)$, and you're back at a Bayesian model again!

in 8 schools, for example, Rubin put prior on τ and averaged over it.

there's no way around this because it doesn't make sense to compare across values of φ without a probability dist. on φ – otherwise how would you know if the sensitivity is worrisome or not?

All models are false: "in the grand scheme of things, the kitten is already dead."

Real Bayes vs. Super-Bayes

real Bayes: build the model. fit the model. check the model. expand the model. rinse and repeat.

Super Bayes: all the models are already there. Start with mixture of all possible models.

"but you don't have to be Georg Cantor to know there's always some model that's not in your super-model."

how far can we take super-Bayes?

6.2 Do the inferences from the model make sense?

8-1b liver from class 1a

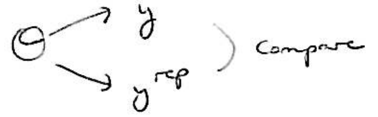
when you find a problem with the model, don't just tweak prior, change the whole model (incorporate interactions, trends, etc.)

Choosing a discrepancy statistic: check what you care about

Stat 220 5b

6.3 Posterior predictive checking

Compare observed data to replications simulated from model



Replications and p-values

classical p-value $\text{p-value}(y|\theta) = p(T(y^{rep}) \geq T(y) \mid y, \theta)$ if test stat is pivotal, doesn't depend on θ

Bayesian posterior p-value

$$p(T(y^{rep}) \geq T(y) | y) = \int \text{pvalue}(y|\theta)\, p(\theta|y)\, d\theta$$

Compute by simulation

↳ idea: generate 19 replicated datasets and put real dataset in random place. can you identify the real one? what about 99 fake datasets? here's a way to get a p-value based on the entire posterior distribution!

↓

picture of 20 replications under normal model - many of the datasets don't look normal at all.

6.4 Graphical posterior predictive checks

Tukey: picture forces us to "notice what we never expected to see"

exploratory and confirmatory data analysis are the same thing - they're both about checking model for problems.

EDA and p-values belong in the same chapter. they're not inference.

"inference is not the inverse of a hypothesis test."

delineation of inference vs. checking: you have a model. you believe it. you have a willing suspension of disbelief. you do inference. then you do model-checking / hypothesis tests.

EDA is a type of model check - same as a hypothesis test. difference is that hypothesis test uses p-value as summary, and with EDA it's graphical, checked against a vague sense of expectation (Tukey quote).

"all graphs are model checks"

residual plot is posterior predictive check where you don't bother to do 20 reps because you already know what reps are supposed to look like. res plot is a test stat with known symmetry properties.

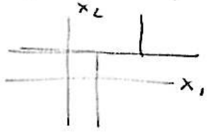
6a

10/10/12

Bayesian additive regression trees (BART)

nonparametric fitting $E(y|x) = \sum_{k=1}^{K} g_k(x)$

divides the space into regions, can get step functions (very flexible) which are smoothed via uncertainties / errors



Jennifer likes BART b/c can model all the x's and y's in their high-dimensional glory instead of collapsing into a pscore

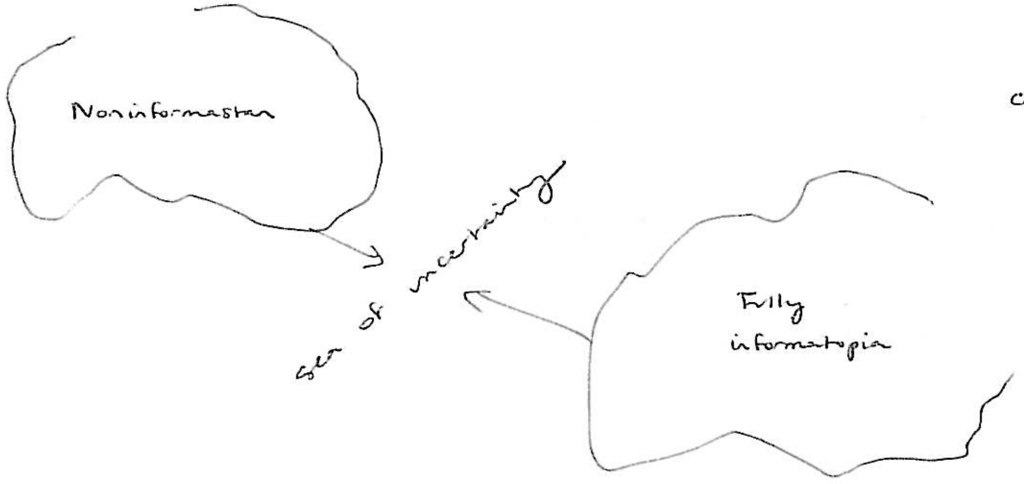
"I hope she's plugging my work in her classes."

on priors: "they don't have to be weakly informative. They can just be shitty."

Stat 220 6a

Why to relax a fully informative prior?

1. could be wrong, even after adjusting for the fact you could be wrong
2. don't want to cheat
3. better generalization to other contexts.



Can start noninformative and then tighten up, or start uptight and then relax.

inferring p for rare disease, "noninformative" prior is actually too informative b/c pulls things to the right.

Dog models

$$y_{jt} = \begin{cases} 1 & \text{if } j \text{ gets shock at time } t \\ 0 & \text{if } j \text{ avoids shock at time } t \end{cases} \qquad t = 0, 1, 2, \ldots$$

Model 1: $P(y_{jt} = 1) = p^t$. $p \in [0,1]$.

Model 2: $P(y_{jt} = 1) = \text{logit}^{-1}(\alpha + \beta t)$. β should be negative, or positive.

Model 3: $P(y_{jt} = 1) = A^{(\# \text{ prev. shocks})} B^{(\# \text{ prev. avoidances})}$

A and B between 0 and 1. need to constrain when fitting, or could get estimates >1.

$A \approx .9$, $B \approx .8$. dogs tend to learn more from avoidance than shock, plus there's the evidential impact of learning the dog is smart if it avoids right away.

dog-specific effects missing from all three - would prefer hierarchical model

Model checking continued.

6.6 Connections to classical testing

p-values and u-values

p-value $P(T(y^{rep}) \geq T(y) \mid y, \theta)$ is a random variable, function of y (which has prior predictive dist.)

u-value some function $u(y)$ that has Uniform distribution, averaging over prior predictive dist.

posterior predictive p-value typically concentrated around 0.5, so is conservative, won't reject as often.

Example where ppp is stuck near 0.5 and this makes sense

Data $y \sim N(\theta, 1)$

Prior $\theta \sim N(0, A^2)$, $A = 100$

Test stat $T(y) = y$ (sample mean)

Posterior predictive $y^{rep} \mid y \sim N(0.9999y, 1.9999)$

$ppp = \Phi\left(\frac{-y}{14000}\right)$

will essentially never reject. if prior were strong, it could reject.

Stat 220 6a

"You can't stand on the beach of the sea of uncertainty with the waves lapping at your ankles. You have to jump into the sea and swim and stick your head underwater and blow some bubbles."

6.7 Model checking for the 8 schools.

Two kinds of replications

new data from same schools : $\theta \to y^{rep\,(i)}$

new data from new schools : $(\mu,\tau) \to \theta^{rep} \to y^{rep\,(i)}$

$$\begin{tikzcd}
(\mu,\tau) \arrow[r] \arrow[ddr] & \theta \arrow[r] \arrow[dr] & y & \\
 & & y^{rep\,(i)} & \\
 & \theta^{rep} \arrow[r] & y^{rep\,(ii)} & \\
(\mu,\tau)^{rep\,(3)} \arrow[r] & \theta^{rep\,(3)} \arrow[r] & y^{rep\,(3)} & \arrow[uuu]
\end{tikzcd}$$

as you move up the ladder, p-values become more concentrated around 0.5. when you get more similar to the actual data, the p-values get less extreme.

can imagine y^{rep} coming out of the page - everything stitched together like macrame.

↳ digression to diss on graphical models

Some people think estimation is too boring, so they try to "learn" connections.

"learn - that's a trendy word."

↓

$$\begin{tikzcd}
(\mu,\tau) \arrow[r] & \theta \arrow[r] & y \\
(\mu,\tau)^{rep} \arrow[ur] & &
\end{tikzcd}$$

is <u>not</u> a posterior predictive check. arrow is pointing the wrong way.

for inference within a model, is enough to have a joint distribution on everything. but for <u>checking</u> the model, need to take into account the data-generating process, so the arrows matter.

" Inference is normal science. Model-checking is revolutionary science."

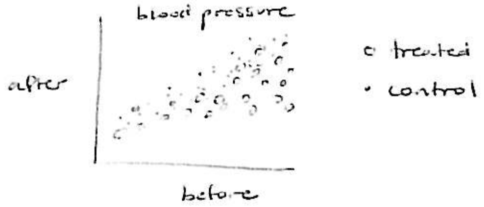
6b

10/12/12

Correlation between pre and post is higher among control than among treated.

$y = a + bx + \theta T + \text{error}$ is wrong - constant treatment effect never happens.

- blood pressure drug



if you were low to begin with, drug doesn't help much

traditional method is interaction term, but that's descriptive, doesn't have a story

Correlation pre/post higher among control.

- "subtractive" treatment effect - variance component in original population gets reduced

before : $y_i^1 = \alpha_i + \varepsilon_i^1$

after : $y_i^2 = \begin{cases} \alpha_i + \varepsilon_i^2 & \text{if control} \\ \lambda \alpha_i + \varepsilon_i^2 & \text{if treated} \end{cases}$

↑ less than 1, to capture variance reduction.

Stat 220 6b

- "additive treatment effect"
  - teaching German
  - control: no one learns German
  - treatment: some kids learn a little, some learn a lot. creates more variance.
  - still, correlation higher among control.

Ch. 7.

7. Evaluating, comparing, and expanding models

7.1 Evaluating the predictive accuracy of a model

log predictive density as a measure of fit
- harder to interpret than mean squared error
- more to compare models than evaluate single model

log data density rather than log posterior density?
- "the answer is that this is a fruitful ambiguity"
- 8 schools example - ambiguous what is prior and what is likelihood
- no-pooling model does no prediction for new schools

Observed fit, cross-validation, and external validation

difference between observed fit and LOOCV is measure of overfitting
- it's the absolute difference that matters. can't say "10% decrease" b/c you could add anything to the log density

can also do 5-fold cross-validation - divide dataset into five pieces many times and take average.

7.2 Information criteria and effective number of parameters

estimates of out-of-sample predictive accuracy
- within-sample predictive accuracy (not good because of overfitting)
- adding an adjustment
- cross-validation
- simulation

Akaike information criterion (AIC)

$\widehat{elpd}_{AIC} = \log p(y|\hat{\theta}_{mle}) - k$ elpd = expected log predictive density

goal: $elpd = E(\log p(\tilde{y}|\hat{\theta}_{mle}))$

effective # of parameters

fitting a function with 30 parameters given 30 data points
- $y_t \sim \text{Pois}(N_t \theta_t)$, $t = 35, \to 64$
  - ↑ prob. of death at given age
- uniform prior: $p(\theta) \propto 1$
- constraint of increasing convexity decreases effective # of parameters
- problem: posterior draws too high. turns out this prior is informative! uniform on θ's is uniform on second differences, so prior is concentrated on quadratic curves.

Stat 220 6b

Deviance information criterion (DIC)

posterior mean

$\widehat{elpd}_{DIC} = \log p(y|\hat{\theta}_{Bayes}) - p_{DIC}$ goal: $elpd = E(\log p(\tilde{y}|\hat{\theta}_{Bayes}))$

↑ effective # of params, based on χ^2_k approx. to -2 log lik.

$p_{DIC} = 2[(\text{log pred density given } \hat{\theta}_{Bayes}) - \text{avg}(\text{log pred density given } \theta)]$

$p_{DIC,alt} = \text{Var}_{post}(\log p(y|\theta))$

↳ noisy function of data. variance sensitive to outliers.

Watanabe-Akaike information criterion (WAIC)

$\widehat{elppd}_{WAIC} = \log p_{post}(y) - p_{WAIC}$ elppd = expected log posterior predictive density

$p_{WAIC} = \sum_{i=1}^{n} \text{Var}_{s=1}^{S} \log(p(y_i|\theta^s))$ Summing stabilizes the noisiness of the variance operation

↑ Simulation draws

need to partition data - not easy for network data, time series, etc., where data points are dependent

"Bayesian" information criterion (BIC)

AIC: $\widehat{elpd}_{AIC} = \log p(y|\hat{\theta}_{mle}) - k$

BIC: subtract $\frac{k}{2}\log n$ instead of k.

not an estimate of out-of-sample predictive accuracy.

favors smaller models.

Cross-validation

for many partitions of the data into y_{train} and $y_{holdout}$:

fit model to training set, get posterior sims

compute log posterior predictive density of $y_{holdout}$

average over simulations to get $\widehat{elppd}_{xval}$

imperfect because still depends on observed data only, so it's still a random variable. the number that gets churned out is for this dataset only.

7a

10/17/12

Ch. 8.

8. Modeling accounting for data collection.

8.1 Bayesian inference requires a model for data collection

Full model is p(data | parameters)

Need a model for the data collection process:

	Observed data	Complete data
sampling	n units in sample	N units in population
experiment	outcomes under observed treatment	other potential outcomes
rounding	rounded observations	precise values
unintentional missingness	observed values	observed and missing values

Stat 220 7a

8.2 Formal models for data collection

data y, inclusion indicators I. I can allow for partial information

ex: rounded data. $y = 10, 10, 12, 11, 9$.

likelihood $\prod_{i=1}^{n}\left[\Phi\left(\frac{y_i+.5-\theta}{\sigma}\right) - \Phi\left(\frac{y_i-.5-\theta}{\sigma}\right)\right]$

latent variable formulation:

$$p(\theta,\sigma,z\mid y) \propto p(\theta,\sigma)\underbrace{p(z\mid\theta,\sigma)}_{\substack{\text{data model}\\ \prod_{i=1}^{n} N(z_i\mid\theta,\sigma^2)}}\underbrace{p(y\mid z,\theta,\sigma)}_{\substack{\text{measurement model}\\ \prod_{i=1}^{n} 1\{y_i=\text{round}(z_i)\}}}$$

8.3 Ignorability.

"why is this chapter different from all other chapters?"

in chapters 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, ...

we write $p(\theta|y) \propto p(\theta)\, p(y|\theta)$

in chapter 8

$$p(\underset{\substack{\uparrow\\ \text{params we}\\ \text{care about}}}{\theta},\underset{\substack{\uparrow\\ \text{params having}\\ \text{to do with}\\ \text{data collection}}}{\psi}\mid y, I) \propto p(\theta,\psi)\underbrace{p(y|\theta)}_{\text{God is here}}\underbrace{p(I|y,\psi)}_{\text{humanity is here.}}$$

"selection model" - select what gets observed.

I depends on y because, for example, old people round ages more.

amount of rounding can itself be a r.v.!

distinct parameters: $p(\theta,\psi) = p(\theta)\,p(\psi)$

means selection process is independent of the truth.

violated if we purposely get bigger sample sizes for rare things

missing at random: $p(I|y,\psi) = p(I|\psi)$

means no selection.

if these are satisfied, then the RHS factors as $p(\theta)\,p(y|\theta)\,p(\psi)\,p(I|\psi)$

So we can just analyze the data ignoring the data collection process.

this is called ignorability: distinct parameters + MAR.

when we just look at $p(\theta)\,p(y|\theta)$, we're assuming ignorability.

now add covariates.

$$p(\theta,\psi|y,I,x) \propto p(\theta,\psi|x)\, p(y|\theta,x)\, p(I|y,\psi,x)$$

distinct params: $p(\theta,\psi|x) = p(\theta|x)\,p(\psi|x)$.

knowing x could be informative about θ and ψ, but conditioned on x, get no additional info about θ from ψ.

MAR: $p(I|\psi,y,x) = p(I|\psi,y^{obs},x)$

missingness can depend on info you have.

throw as many x as possible into the model so ignorability is reasonable assumption.

double-blindness is a form of conditional independence - puts a wall between information and decision.

this is a thread connecting classical stat to what we do!

Stat 220 7a

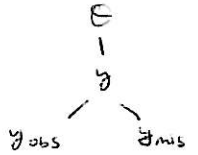
8.4 Sample surveys

Simple random sampling

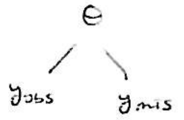
finite-population inference

$$\bar{y} = \frac{n}{N}\bar{y}_{obs} + \frac{N-n}{N}\bar{y}_{mis}$$

to do inference, assume



is the same as



then can just use y_{obs} to infer θ, get inference for $\bar{y}_{mis}$ based on posterior predictive.

asymptotically, $\bar{y} \mid y_{obs} \sim t_{n-1}\left(\bar{y}_{obs}, \left(\frac{1}{n} - \frac{1}{N}\right) s^2_{obs}\right)$.

Stratified sampling

$$\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j} \leftarrow \text{stratum sizes}$$

example: Bayes estimates of public opinion on school vouchers, by religion/race, income, and state.

"whites, .., blacks, Hispanics, and most of the people in this room."

7b

10/19/12

8.5 Designed experiments

"an experiment means doing something you wouldn't otherwise do"

we want $p(I \mid y, \varphi, x) = p(I \mid \varphi, x)$, so include enough x's to make this true

for randomized block experiment, x needs to include block membership

need hierarchical model of y given x

↳ designs that cheat.

lots of ignorable designs! consider the following sequential designs.

① $n = 20$

② 2 hours

③ $n \sim \text{Pois}(20)$

④ $n_1 = 10$

⑤ $|n_1 - n_2| > 4$

⑥ $\left| \dfrac{\bar{y}^1 - \bar{y}^2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right| > 2$

⑦ $\left| \qquad \right| > 3$

all are ignorable. all have the same likelihood function!

for ⑥ and ⑦, if θ's close together, takes forever to reach stopping rule, get precise estimate near 0.

if θ's far apart, get there right away.

if θ^{study} interval is [0.0001, 0.0003], $\theta^{tomorrow}$ probably includes 0

"maybe we think there's nothing going on because it's education research."

" it's so great when you guys laugh - it's like blood to a vampire!"

Stat 220 7b

Example: matched pairs design

item	gp	treat	y(1)	y(2)
1	1	1	✓	
2	1	2		✓
3	2	1	✓	
4	2	2		✓
5	3	1	✓	
6	3	2		✓
⋮	⋮	⋮	⋮	⋮

estimator ~~estimand~~ is $\bar{z} = \bar{y}_1 - \bar{y}_2$ where $z_j = y_j^1 - y_j^2$, j indexing the groups.

right	wrong
$\frac{s_z^2}{n}$	$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$
reg z on 1	reg y on 1, T
$z = a + \text{error}$	$y = a + bT + \text{error}$

} doesn't account for pairing.

also right: reg y on T and group indicators. gives same estimate.

Bayesian (also right): put prior on n group coefficients $\theta_1, \ldots, \theta_n \sim N(\mu, \tau^2)$. then do this regression.

"this is a textbook way too. it's just my textbook."

will only make a difference when τ is small (not much group variation)

by partially pooling, get more precise estimate, because you're partially pooling toward the "wrong" analysis that gives more degrees of freedom.

in practice, the real benefits are for more complicated designs, since this generalizes to 3 people/group, etc.

Example: experiment on 50 cows.

outcome: amount of milk fat produced.

treatments: 4 levels of feed additive.

background variables: lactation #, age, initial cow weight.

repeated randomization until balance.

how to analyze?

conditioned on lactation #, age, and cow weight, it's ignorable. those are the only covariates used in the assignment. "he wasn't face to face with the cows, as it were."

So just throw those covariates into the model. regress treatment on the covariates. then it's irrelevant how many times he randomized.

8.6 Sensitivity and the role of randomization

why randomize?

with no item-level background variables, randomization is the only ignorable design.

with background variables:

design ABABABBABABA uses background variable of location.

advantages of randomization are for model-checking (replicating future data) and robustness.

Stat 220 7b

8.7 Observational studies

Balance (on observed and unobserved variables)

Lack of complete overlap in high dimensions – hard to detect and deal with.

Matching and poststratification.

matching and regression together: matching deals with lack of overlap, and regression helps further with balance.

propensity score matching.

poststratification (example: mother's education)

8.8 Censoring and truncation

N observations from $f(y|\theta)$, only observed when < 200. observe $n = 91$ cases where $y_i < 200$.

Scenario 1: N unknown. truncated-data likelihood

$$p(\theta|y) \propto p(\theta) \frac{\prod_{i=1}^{91} f(y_i|\theta)}{F(200|\theta)^{91}}$$

Scenario 2: N known. censored-data likelihood

$$p(\theta|y,N) \propto p(\theta) \left(1 - F(200|\theta)\right)^{N-91} \prod_{i=1}^{91} f(y_i|\theta)$$

for a Bayesian, if N is unknown, we should be able to use the censored-data model + a prior on N, average over N to get $p(\theta|y)$:

$$p(\theta|y) \propto \sum_{N=91}^{\infty} p(N)\, p(\theta) \left(1 - F(200|\theta)\right)^{N-91} \prod_{i=1}^{91} f(y_i|\theta)$$

but it turns out this reduces to the truncated-data model <u>if and only if</u> $p(N) \propto \frac{1}{N}$.

?!

Summary of Chapter 8

The method of data collection dictates the <u>minimal</u> level of modeling required for a valid Bayesian analysis.

Condition on all information used in the design

in a survey: stratum and cluster indicators, any variable that determines probability of sampling or probability of non-response

in an experiment: pairs, blocks, any variable used in treatment assignment.

Mr. P: <u>regression analysis</u>, <u>multilevel</u> b/c many unknown parameters, and <u>poststratification</u>.

8a

10/24/12

"Survey weights are like McDonald's chicken nuggets – you don't know what goes into them."

open problem – poststratification on many variables.

Stat 220 8a

Latin square homework problem

"before you do anything, you gotta take the ... you gotta take the ..."

because the effects are probably multiplicative, and in any case this is more interpretable.

"you take the log so fast that you don't even see the actual data. plus you take the log because you can, because they're all positive."

$y_i = \mu + \alpha_{row[i]} + \beta_{column[i]} + \gamma_{treatment[i]} + \epsilon_i$

probability model:

$y_i \sim N(\mu + \alpha_{row[i]} + \beta_{column[i]} + \gamma_{treatment[i]}, \sigma_y^2)$ $i = 1, 2, -, 25$

$\alpha_j \sim N(0, \sigma_\alpha^2)$

$\beta_k \sim N(0, \sigma_\beta^2)$ $j, k, l = 1, -, 5.$

$\gamma_l \sim N(0, \sigma_\gamma^2)$

$p(\mu, \sigma_\alpha, \sigma_\beta, \sigma_\gamma, \sigma_y) \propto 1.$

Petri dish homework problem

6 cultures per dish, 5 dishes. individual-level analysis or dish-level analysis?

$$\bar{y}_A \pm \frac{s_A}{\sqrt{30}}, \quad \bar{y}_B \pm \frac{s_B}{\sqrt{30}} \qquad \text{vs.} \qquad \bar{\bar{y}}_A \pm \frac{s_{\bar{y}_A}}{\sqrt{5}}, \quad \bar{\bar{y}}_B \pm \frac{s_{\bar{y}_B}}{\sqrt{5}}$$

right thing is to do dish-level analysis because if she's right and there are no dish effects, then $s_{\bar{y}_A}^2$ will be 6 times less variable anyway. but if she's wrong, she should do the dish analysis!

unlike the sham data from the chicken brains, here the loss from the dish analysis is small.

even better would be to do partial pooling.

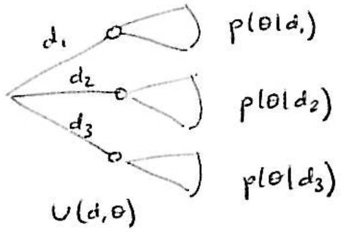
Ch 9

9. Decision analysis.

data, model > inference | decision-analysis > decision.

9.1 Bayesian decision theory in different contexts

decision trees



JITT question about microlifes: 20 cents per microlife is literal. people will accept some amount of money to increase risk, but it doesn't make sense to talk about people accepting 200k to die.

Stat 220 8a

Decision analysis vs. "statistical decision theory" - evil twin, says things like "use posterior mode for 0-1 loss" but isn't connected to real-world decision problem.

Dave Krantz: goal-based framework.

ex: effects of incentives for telephone surveys.

6 factors: incentive or not, value, form, timing, mode, burden.

fit model, see if inferences make sense, pipe them into decision analysis.

ex: radon measurements - remediate, measure, or do nothing

8b

10/26/12

Ch. 10.

10. Overview of computation

10.1 Crude estimation by ignoring some information

Simpler model

Set hyperparameters to fixed values

note: don't always want to use no pooling or complete pooling as simple model.
↓
computationally unstable since you're estimating a ton of parameters.

quick imputation of missing data.

network of models.

fit multiple models to:
(1) compare
(2) check
(3) understand what we're doing - "scaffolding our understanding"

"exploratory model analysis"

open question: how to do this systematically?

develop a "syntax of models" - models connected to other models in a network, and edge exists if they differ by only one thing. define "operations" you can perform on a model: adding a predictor, adding a hyperparameter, adding a more general distribution (t instead of Normal), etc.

↳ fractal nature of scientific revolution

true at every time scale

idea is that you throw in some data and the computer program builds the model by starting with simple things and adding snap-ons. then you need a model-checking module.

Stat 220 8b

10.2 Direct simulation

grid sampling

rejection sampling

approximation $M \cdot g(\theta)$ that dominates $p(\theta|y)$

can compute and draw from.

for $s = 1, \dots, S$:

draw θ from density proportional to g

accept with probability $\frac{p(\theta|y)}{M \cdot g(\theta)}$

usually doesn't work because acceptance rates absurdly low.

10.3 Numerical integration

goal: posterior expectation $E(h(\theta)|y) = \int h(\theta)\, p(\theta|y)\, d\theta$

estimate by $\frac{1}{S}\sum_{s=1}^{S} h(\theta^s)$ if can draw directly.

Laplace's method:

approximate $h(\theta)p(\theta|y)$ by $e^{\text{quadratic in } \theta}$

fit to mode and curvature at the mode.

using unnormalized density:

$$E(h(\theta)|y) = \frac{\int h(\theta)\, q(\theta|y)\, d\theta}{\int q(\theta|y)\, d\theta}$$

apply Laplace to numerator and denominator.

10.4 Importance sampling

goal: $E(h(\theta)|y) = \frac{\int h(\theta)\, q(\theta|y)\, d\theta}{\int q(\theta|y)\, d\theta}$

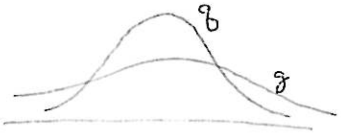
but can't draw from q, only from g.

compute importance weights $w(\theta^s) = \frac{q(\theta^s|y)}{g(\theta^s)}$

estimate of $E(h(\theta)|y)$: $\frac{\frac{1}{S}\sum_{s=1}^{S} h(\theta^s)w(\theta^s)}{\frac{1}{S}\sum_{s=1}^{S} w(\theta^s)}$

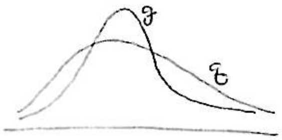
want g to have heavier tails than q.

scenario 1



okay

scenario 2



bad. could get weight of a million.

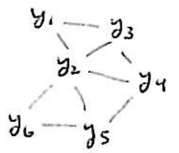
0.5 Computing normalizing factors

$p(y|\theta) = \frac{1}{z(\theta)} q(y|\theta)$. $z(\theta)$ is the normalizing factor.

typically want to calculate $z(\theta)$ offline - get $z(\theta)$ as a function of θ.

Stat 220 8b

models in BDA : "Normal, Poisson, Binomial, Gamma, that's it."



$p(y|\theta) \propto e^{-\theta \sum c_{ij} (y_i - y_j)^2}$ y_i's bounded

c_{ij} =1 if i and j are neighbors, 0 otherwise

if $\theta > 0$, says neighbors have to be similar

spatial equivalent of AR model in time series

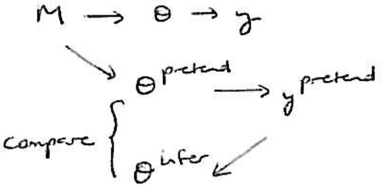
finding normalizing factor is hard.

10.6 Use of posterior simulations in Bayesian data analysis

estimating rare event probabilities by combining simulations and analytic probabilities.

10.8 Practical issues

Fake-data debugging



debugging : bridge between

simple models that can be fit successfully $\longrightarrow$ $\longleftarrow$ complex models that don't fit

- work with smaller datasets
- strip down the model
- fixed parameter values, then strong priors, then weak priors

week 9 - Hurricane Sandy. ☹

10a 11/7/12

read Ch 20 of Gelman and Hill.

JITT

	$p = .6$	$p = .75$
SE	$\frac{.5}{\sqrt{n}}$	$\frac{.5}{\sqrt{n}}$

" the SE is .75 times (1-.75) over... oh, give me a break, this works just fine."

set diff > 2.8 × SE

$.15 > \frac{2.8 \times .5 \times \sqrt{2}}{\sqrt{n}}$ solve for n.

Ch. 13.

13. Approximations to the posterior distribution

role of approximations has changed - used to be the easy thing to do, but now MCMC is sometimes easier than the approximation. approximation is more <u>scalable</u>.

def. of approximation : doesn't converge to the right stationary distribution.

multimodal posteriors often arise in discrete data (genetics - one gene vs. the other)

Stat 220 10a

13.2 Boundary-avoiding priors for modal summaries

a prior that's bad for the posterior mode might be fine for full Bayes.

conversely, if we know we're going to use PM, we might want to choose a different prior.

ex: Gamma(2, 0.1) prior for τ in 8-schools example

13.3 Normal and related mixture approximations

fit to mode and curvature at the mode

mixture of Normals or multivariate t

importance sampling (with replacement)

> "Rubin calls this sampling importance resampling, which is silly because once you say resampling, you must have sampled already. It's like you say e4, you don't say e2 to e4 because of course it was at e2, it's not like it was at e3!
>
> I guess e4 is old-fashioned nowadays. I guess I should do something like g3."

13.7 Variational Bayes

goal: approximate $p(\theta|y)$ by $g(\theta)$

typically assume $g(\theta) = \prod_{j=1}^{J} g_j(\theta_j)$

approximating each marginal distribution: $p(\theta_j|y) = \int \cdots \int p(\theta|y)\, d\theta_{-j}$

converges to a distributional estimate (many times is just a lower bound)

VB for 8 schools

$$\log p(\theta, \mu, \tau | y) = \text{const} - \frac{1}{2}\sum_{j=1}^{8} \frac{(y_j-\theta_j)^2}{\sigma_j^2} - \frac{1}{2}\sum_{j=1}^{8} \frac{(\theta_j-\mu)^2}{\tau^2} - 8\log\tau$$

look at θ_1: $\text{const} - \frac{1}{2}\frac{(y_1-\theta_1)^2}{\sigma_1^2} - \frac{1}{2}\frac{(\theta_1-\mu)^2}{\tau^2}$

θ_1 is the star! only care about terms involving θ_1

μ, τ: average over, using $g(\mu)$, $g(\tau)$

do the same with all other parameters.

VB underestimates variance.

research project: VB + particle filtering.

10b

11/9/12

JITT: if we want to find the slope of a dose-response curve and we know it's linear, just measure at endpoints!

three sources of error: measurement error, model error, natural variability.

xkcd on Bayesians vs. frequentists: "I don't like this."

frequentism is a conservative viewpoint.

- Bayesianism is conservative in that it respects prior beliefs.
- frequentism is conservative in the sense of distrusting any overall/overarching philosophy. seeing a situation like this, they would say the p-value is not appropriate.
- Bayesians have this idea that their methods should work everywhere. frequentists are more willing to say their method doesn't work in this situation.

but isn't that an unprincipled way to incorporate prior information? sure, but then the fair question is: how much do you lose from a discrete approach to prior information as opposed to a prior distribution?

Stat 220 10b

- p-value is not the only aspect of a statistical test.
- what do frequentists give up on? frequentists might say "I can't give you a probability."
  - Bayesians will give you something, but it could be bad. so it would be easy to make a cartoon where the Bayesian looks silly:
  - freq: "I can't give you a probability."
  - Bayesian: "I used a flat prior and the probability is $\frac{1}{36}$."
- "But if someone's going to be unfairly attacked, I'd rather it not be me."

Causality and statistical learning.

Forward vs. reverse causal questions

- Rubin hates reverse causal questions, but it's easy to think in terms of reverse causality.
- now, some things are easy to think about but are just wrong. like anthropomorphizing objects.
  - or folk physics: people think a baseball trajectory looks like  .
- but in this case it seems there's something to these questions - it's a fruitful inquiry.
- one way to think about reverse causal questions is as a way to generate a lot of forward causal questions,
  - which are easier to handle from a statistical perspective.
- does it make sense to formalize reverse causal inference? is it possible?

"the way I think about causal inference now is in terms of model checking"

- if attractiveness is related to earnings and that's a surprise, then it's a departure from our mental framework.

Different perspectives on causal inference

- humans: wired to ask reverse questions
- macroeconomists: state-space models
- applied micro: forward causal inference. think in terms of interactions.
- statisticians: fitting models
- computer scientists: modeling everyday reasoning
  - example: traveling salesman. really hard optimization problem, and yet people figure out how to get places!
  - similarly, causality is hard, but everyone does it every day, so it's possible for computers to learn it too. an optimistic view!
  - statement like "rain causes mud. mud does not cause rain" sounds silly, but it's about thinking of the relationships between variables in terms of how people think about them.

Example: Deterrent effect of the death penalty.

- law school prof with no stat background trusts "sophisticated econometricians"
- ridiculous regression models with additivity assumptions
- economists don't want assumptions, don't want bias, so end up with low power and high variance,
  - so they throw a ton of data at it, including inappropriate pooling across time, which creates bias!

Example: U-bend of happiness

- graph from The Economist shows middle-aged people are sad, but GSS happiness is exactly the other way around.

Stat 220 10b

Difficulties with the research program of learning causal structure

"I don't like the idea of state variables causing each other."

"learning causal structure" is all about discovering zeros, but there are no true zeros.

a lot of social science proceeds by trying to discover stylized facts and reason from them,

but the statement that something has no effect isn't generally reasonable.

11a 11/14/12

Ch. 16

16. Generalized linear models

What are gains from Bayesian treatment of GLMs?

hierarchical models and regularization.

Tinkertoy-style model expansion in Stan.

hierarchical models allow for:

spatial and time series. more data on existing people.

partial pooling within an analysis - infill.

combining different data sources - annexation.

} metaphor of expanding city.

16.1 Standard GLM likelihoods.

almost always want overdispersed model, except "50000 people need your help" from Gelman and Hill.

16.2 Setting up and interpreting GLMs

offsets - like log population size in Poisson regression. coefficient should be 1 when you do the regression.

identification - if have 3 ethnic groups and 5 grade levels, set one of each to 0.

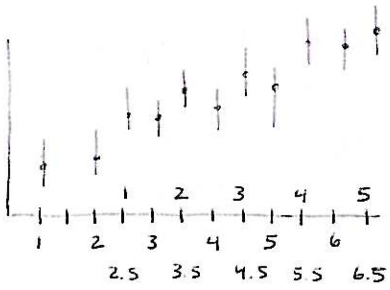
if have interactions, set an entire row and column to 0.

extended example: ETS, predicting Calc II grades from AP scores.

$y_i = a + b_{j[i]} + \text{error}$ $j[i]$ = group of obs. i index variable

same as

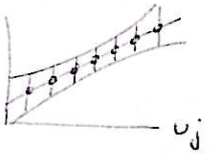
$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_5 x_{i5} + \text{error}$. indicator variables



$b_j \sim N(\beta(v_j - \bar{v}), \sigma_b^2)$ not exchangeable

$v_j = \{1, 2, 2.5, 3, \ldots, 6, 6.5\}$

if $\sigma_b = 0$, then the b_j's are all exactly on the line.



SEs are just from ordinary linear regression.

average within a group

$avg_j = a + \beta v_j + err_j$

errors confounded with a, so SEs too wide.

instead of a, β, error, look at $a + \overline{err}$, β + slope(err), residual. these are more stable.

parameters don't mean what you think.

with few groups, you have basically no info on group-level variance σ_b^2

big question is appropriate postprocessing.

Stat 220 11a

takeaway: multilevel modeling lets you estimate individual coefficients and group-level effects all at once.

"the computation is going to be hard for GLMs no matter what, so you're already dead. okay, not dead, but that bus is already gone."

16.3 Computation

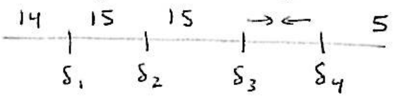
Separation - will always happen with enough predictors. only happens for discrete, not continuous.

11b

11/16/12

JITT:

ordered logit



using mode for (β, δ), $\hat{\delta}_3 = \hat{\delta}_4$,
so predictive probability is 0.

pilot study is to figure out your design, not to measure effect.

Sampling words at random:

words x_i

actual probability $g(x_i)$

target probability $p(x_i) \propto 1$

use importance resampling: $w_i = \frac{p_i}{g_i}$, get 2000 and resample 1000.

Ch 20

20. Nonparametric regression

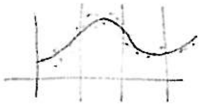
is more important than density estimation.

20.1 Splines and other basis functions

"because cubicity is closed under addition - doesn't that sound good?"

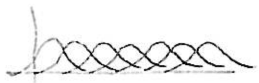
cubics allow you to fit a sequence of points piecewise and have continuous 1st derivative.

divide x-axis into pieces, fit a cubic within each piece



can express spline as sum of basis functions. then model is $y_i = \sum_{j=1}^{J} b_j \, B_j(x_i) + \text{error}$.

Gaussian kernels:



by adding scaled versions of these, can get increasing functions
or other shapes.

20.2 Prior distributions for basis-function models

instead of estimating the knots, think of there being a large but fixed number of knots, put a prior on how many you need, and only keep how many you need.

Similar to variable selection in regression.

inimodal symmetric prior means scale mixture of Normals.

Stat 220 11b

20.3 Multivariate regression surfaces

additive model (not good for hw problem because that would assume constant difference between men and women)

multivariate kernels for multivariate response

"tensor products, which are some sort of product ... of tensors, I believe."

20.4 Gaussian processes

$y_i = g(x_i) + \text{error}$

g defined on discrete points $g_1, g_2, \to g_{73}$.

model $$\begin{pmatrix} g_1 \\ \vdots \\ g_{73} \end{pmatrix} \sim N\left(\left(\quad \right), \left(\qquad \right) \right)$$

mean process.
can be regression.
can have hyperparameters.

covariance.
high correlations for nearby values, low correlations for faraway values.

want local smoothness but globally unconstrained. unlike AR process.

covariance function:

could be $c(x, x') = \phi_1 \exp(-\phi_2 \| x - x' \|^2)$

$$c(x, x') = \phi_1 \exp\left(-\sum_{j=1}^{d} \alpha_j (x_j - x_j')^2\right)$$

Gaussian process prior for coefficients of basis expansion

12a

11/28/12

Bayes helps if you're allowed to use prior information or if you have a lot of parameters.

"one of the most influential Bayesian books was written by LJ Savage, and it's full of bad ideas.
there isn't a single good idea in the whole book."

"Sergiy seems like a more serious lecturer, which is good, you get the soft jokes now and then next semester
it's time to learn some shit."

Ch. 21

21. Finite mixture models

Example: identifying a three-component mixture

"you're not going to go to heaven because you only needed two." (mixture components)

motivation: want to learn about

- partisan bias - are Dems getting more seats than they deserve?
  if they got 50% of the vote, how many seats would they get?
- electoral responsiveness - how responsive is Congress to the voters?
  if there were a 1% shift in the vote, how would that affect the # of seats assigned to each party?

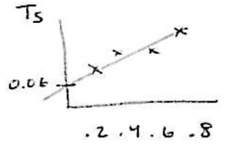
Stat 220 12a

hypothetical election: shift things and add noise

$$y_i = \alpha_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, 0.06^2)$$

$$y_i^{rep} = \alpha_i + \varepsilon_i^{rep}$$

0.06 is unexplained variation in vote share, obtained as



$$T_S = \sqrt{\tfrac{1}{2} \operatorname{avg} \left((y_{it} - \bar{y}_t) - (y_{i,t-1} - \bar{y}_{t-1}) \right)^2}$$

i = district. t = time.

21.4 More general formulation

in schizophrenic reaction time example, mixture components actually mean something, as distinguished from just fitting clusters because you can.

meaning of a cluster depends on context.

might make sense to arbitrarily break into clusters to explain variation in outcome

example: clustering the 50 states according to welfare policy.

21.5 Label switching and posterior computation

for identifiability, impose constraints like $\mu_1 < \mu_2$ or $\sigma_1 < \sigma_2$

21.6 Unspecified # of mixture components

average over unknown K

choose a prior

marginal likelihood of K or transdimensional MCMC (propose merging clusters)

or choose a large (but not huge) K and maybe some of the clusters will be empty

prior on membership probabilities

$$(\pi_1, \ldots, \pi_K) \sim \text{Dir}\left(\tfrac{1}{K}, \ldots, \tfrac{1}{K}\right).$$

"I've never done it. I just read about it in my own book."

12b

11/30/12

question on how to compare models. writes in giant letters "Out of Sample Prediction Error".

("So there's that.")

"the likelihood function never tells you the whole story, because you don't know where the data come from."

what can you do with 21 data points?

1. see whether the data are consistent with your prior
2. do univariate analyses. in this case, found that none of the 16 predictors was highly correlated with outcome. That's learning something!

"better to have analyzed and lost than never to have analyzed the data at all."

"a Bayesian version will usually make things better."

Stat 220 12b

"hierarchical models: just like what you did before, but the standard errors are a little bit fatter."

"political scientists have no pride. they'll use anything that works. or doesn't work - it's not like they would know."

Ch. 22

22 Infinite-parameter models

mixture models with potentially infinite number of components

need new classes of probability models: distribution on infinite # of r.v.s that sum to 1.

22.1 Bayesian histograms, aka better histograms

histogram without specifying # of bars or knot locations

more bars where there's more variability - solves the hist(rcauchy(1000)) problem.

Is hierarchical Bayes nonparametric?

Bayesian models are always parametric at the top level, but not at intermediate stages in the sense that the curve may not be restricted to any subspace of the space of possible curves

line between parametric and nonparametric is blurry

research principle: "you take what someone is doing and pretend they're being Bayesian."

Bayesian data analysis

The three steps!

Building confidence in classes of models.

> "I've chunked the 8 schools. As we develop, we chunk more and more things. We'll be able to chunk DPs and CRPs."

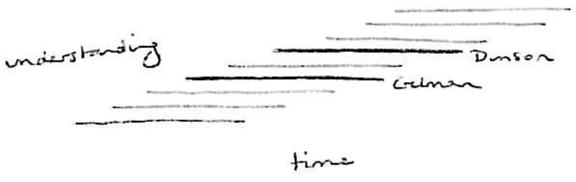
What is the role of theory?

> "the full name of theoretical statistics is the theory of applied statistics."
>
> "statistics is applied statistics."
>
> "the gambler's ruin problem, that's a theory about what happens when you're a gambler."
>
> paper with Jennifer on multiple comparisons
>
> Tukey: the model is irrelevant. you construct the model, it leads to a method, the method has statistical properties, and it's the properties that matter, not the model.
>
> "like he advocated plotting a rootogram instead of a histogram. you take the square root of the counts, because if the counts follow a Poisson and you... [gasps, covers mouth] oops, I said Poisson!"

Open questions

- systematic model choice
- grammar of models
- Computation - reparametrization.
- model understanding via graphs.
- automated model checking (write down principles)

Where will statistics be in 20 years?



you don't perceive the change, but it's there. :)