Several Simple Models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

September 2, 2020

Outline

Previously:

- Binomial model (spinning coin)
- Poisson model (kidney cancer)

Now:

- Normal models (known or unknown variance)
- Some commentary on priors
- Linear regression as a normal model

Normal model with known variance

The normal model, known variance

We'll start with a normal model, and as an example case we'll use a data set for basketball scores: final scores y_i from all NCAA men's tournament games from about 1939-1995.

- Often normal models get used out of convenience or out of tradition
- When justified, usually justified by the central limit theorem: sum or average of many IID components gives rise to normal distribution

A visual inspection of the data distribution shows a normal distribution really does fit here, but it's reasonably well justified from first principles

The normal model, known variance

As usual, our starting point is specifying a model and priors for our parameters:

$$y_i \sim \mathrm{Normal}(\theta, \sigma)$$
$$\theta \sim \mathrm{Normal}(\mu_0, \tau_0)$$

Take $\sigma = 14$. Here, we are choosing a normal prior for convenience (it's conjugate to the normal likelihood)

Can we choose a value for μ_0?

Checking our prior: prior predictive simulations

We've set a prior – we should check that it's at least slightly reasonable.

Simple thing to do: draw some samples, make sure they're not off-the-wall ridiculous

- We're not looking for the prior predictions to be a perfect model for the data
- But, if our predictive draws have games with negative score, or teams scoring 500 points, maybe something is off

Calculating the posterior

Assume we start with one observation y. Since we are using a conjugate prior, the posterior is analytically expressible:

$$\begin{aligned} p(\theta|y) &\propto \exp\left(-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\tau_0^2}\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}\right)\theta^2-\left(\frac{2y}{\sigma^2}+\frac{2\mu_0}{\tau_0^2}\right)\theta+\frac{y^2}{\sigma^2}+\frac{\mu_0^2}{\tau_0^2}\right) \end{aligned}$$

Then some magic happens...

Calculating the posterior

$$\theta | y \sim \mathrm{Normal}(\mu_1, \tau_1)$$

where

$$\mu_1 = \frac{\frac{1}{\sigma^2} y + \frac{1}{\tau_0^2} \mu_0}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}$$

$$\frac{1}{\tau_1} = \frac{1}{\sigma_2} + \frac{1}{\tau_0^2}$$

The inverse variances $1/\sigma^2, 1/\tau^2$ are called the *precisions* of these distributions

(Where's the magic? Complete the square (exercise 2.14(a)) in the book)

The posterior as a compromise

Three ways of writing the posterior mean of θ:

$$\mu_1 = \frac{\frac{1}{\sigma^2} y + \frac{1}{\tau_0^2}\mu_0}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}$$

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

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

- Weighted average of μ_0 and y
- Prior mean μ_0 adjusted toward the data
- Data "shrunk" toward the prior mean

Generalizing to many observations

We don't have to iterate this process a thousand times to incorporate our thousand games (although the ability to incorporate observations one by one can be considered a feature of the Bayesian approach); the posterior depends on $y_1, y_2, \ldots$ only through the sample mean $\bar{y}$[1]

$$\theta | y_1, y_2, \ldots \sim \mathrm{Normal}(\mu_n, \tau_n)$$

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

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

[1] $\bar{y}$ is called a *sufficient statistic* in this model

Posterior predictions

The posterior predictive distribution is (unsurprisingly) also normal (details in section 2.5 of BDA) with

$$E(y|y_{\mathrm{obs}}) = \mu_n$$

$$\mathrm{var}(y|y_{\mathrm{obs}}) = \sigma^2 + \tau_n^2$$

Intuitively:

- mean prediction is posterior mean of θ
- uncertainty of prediction is the uncertainty in θ (epistemic uncertainty, τ_n^2) plus the uncertainty of individual observations (aleatoric uncertainty, σ^2)

Posterior predictions

Let's do some posterior predictions...

Priors

Informative vs. uninformative priors

Most often, priors are categorized as *informative* or *uninformative priors* depending on whether they incorporate outside scientific information

- informative priors: bring in knowledge about the application domain, or results of previous study, as a starting point for estimation and inference
- uninformative priors: avoid using external knowledge, "let the data speak for itself"

In reality, informative vs. uninformative is not always a sharp binary distinction.

Proper and improper prior distributions

The prior precision $1/\tau_0^2$ is the weight given to the prior mean in the posterior distribution; if this precision is very small, it is as if $\tau^2 = \infty$.

In other words, if $n/\sigma^2 \gg 1/\tau_0^2$, then the posterior distribution is approximately

$$p(\theta|y) \sim \mathrm{Normal}(\bar{y}, \sigma/\sqrt{n})$$

We could imagine that we had assigned a prior that is constant/uniform. Problem: has an infinite integral!

Improper prior distributions can produce proper posteriors

This example shows that even with an improper uniform prior on θ, the posterior distribution is proper – i.e. $p(\theta|y)$ has a finite integral for any possible data y (as long as there is at least one observation).

- This must be checked any time you use an improper prior
- Most reasonable interpretation of the posterior: as an approximation, valid when the likelihood dominates the prior density
- This is generally dependent on both sufficient amount of data and sufficiently localized likelihood

Uninformative priors

Some issues about uninformative priors:

- uninformative doesn't always mean "flat" / uniform
    - a prior that is flat in one parameterization may be non-flat if you change variables
    - flat priors can be improper
    - flat priors can be practically nonsensical
    - from Aki Vehtari – how about a flat prior on "the amount of money in my wallet"?

Weakly informative priors

A compromise between the informative and uninformative priors is so-called "weakly informative" priors, which generally attempt to include enough outside knowledge to ensure that the prior is proper and sensible, but the information in the prior is intentionally weaker than the availabele outside informations.

- Example: in the coin spinning problem, take $\mathrm{Beta}(3, 3)$ in place of uniform or $\mathrm{Beta}(1, 1)$.
- Example (from the book): in estimating the proportion of female births, choose a prior with the probability mass concentrated between, say, 0.4 and 0.6 (e.g. $\mathrm{Normal}(0.5, 0.1)$)

Normal model, unknown variance

Introduction to multi-parameter models

The known-variance assumption isn't necessarily particularly realistic. So instead, we can allow σ^2 to be an unknown parameter in our model.

New model (simple, improper priors):

$$
\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu, \sigma^2) \\
p(\mu, \sigma^2) &\propto (\sigma^2)^{-1}
\end{aligned}
$$

The improper prior is derived from applying a uniform prior to $\mu, \log \sigma^2$.

The joint posterior

As before we can get the joint posterior by simply multiplying this prior by the likelihood $N(\mu, \sigma^2)$ for our data $y_1, \ldots, y_n$.

$$
\begin{aligned}
p(\mu, \sigma^2|y) &\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \\
&= \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)]\right)
\end{aligned}
$$

where $\bar{y}$ is the sample mean, s^2 the sample variance.

Interpreting the posterior

The posterior here shows that μ, σ are not independent.

In order to interpret this, it's easier to think about the distribution of μ conditional on σ^2. This is simple:

$$\mu|\sigma^2 \sim \mathrm{Normal}(\bar{y}, \sigma^2/n)$$

which is the same as the known variance case. To factor the joint posterior, then, we need $p(\sigma^2|y)$.

Marginal distribution of σ^2

The marginal posterior distribution of σ^2 is obtained by averaging the joint posterior over μ:

$$p(\sigma^2|y) \propto \int_{-\infty}^{\infty} \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)]\right) d\mu$$

Looks gnarly, but it's not – the exponential factors into the part dependent on s and the part dependent on μ. The part dependent on μ is a simple Gaussian integral, proportional to σ^{-1}. So, we get...

Marginal distribution of σ^2

The marginal posterior distribution of σ^2 is

$$p(\sigma^2|y) \propto \sigma^{n-3} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)$$

which is a standard density on σ^2:

$$\sigma^2|y \sim \mathrm{Inv}{-}\chi^2(n-1, s^2)$$

Sampling from the joint posterior

It is easy to sample from the joint posterior distribution:

- draw a value of σ^2 from the posterior for σ^2
- draw a value of μ from the conditional posterior given your value of σ

Sampling from the joint posterior

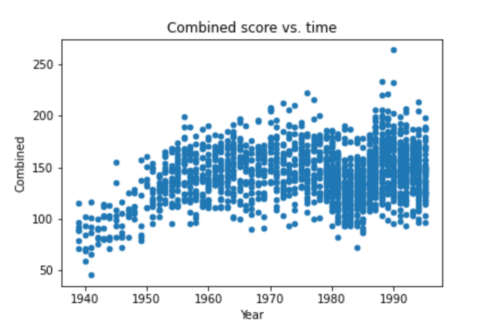
For the following plots:

- Restricted to year ≥ 1988
- Subsampled: $n = 10, 40, 100$

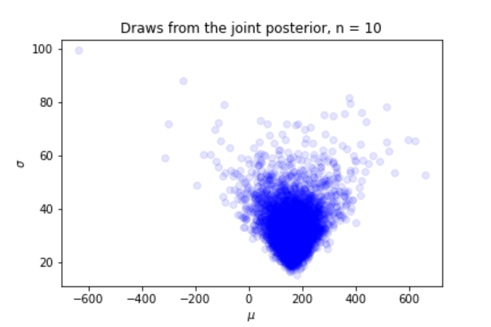
Why limit to data after 1988?

Time series

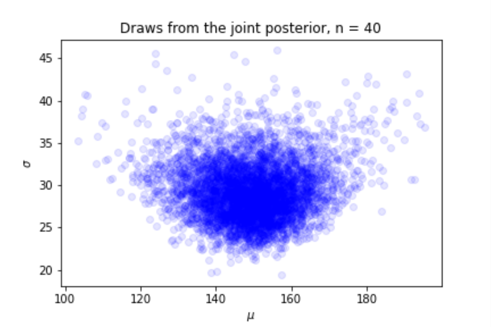
Plotting all of the combined scores against time:



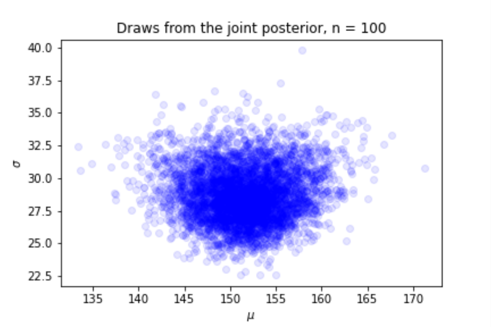
Sampling from the joint posterior



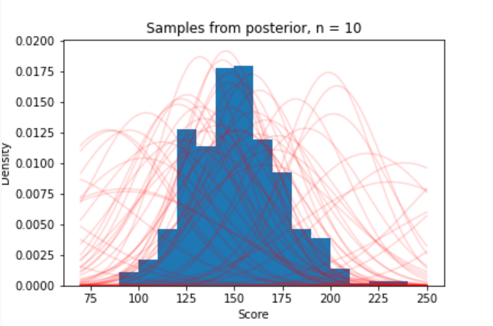
Sampling from the joint posterior



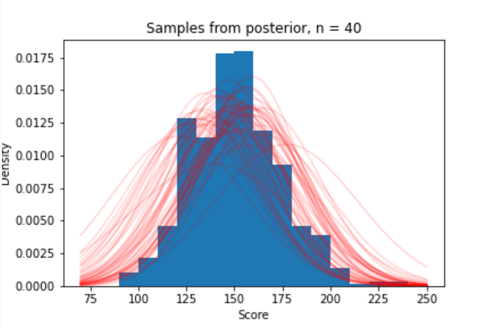
Sampling from the joint posterior



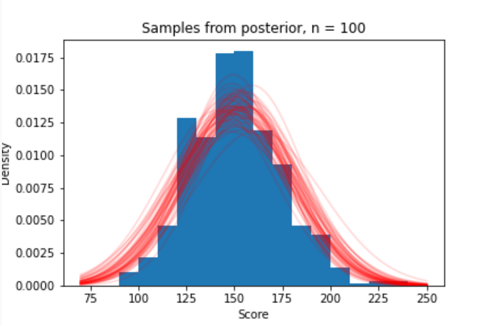
Visualizing uncertainty



Visualizing uncertainty



Visualizing uncertainty



Regression

Regression models

We all know about simple linear regression:

- Given data (x_i, y_i)
- Model the data as
$$y = \alpha + \beta x + \varepsilon$$
where $\varepsilon \sim \mathrm{Normal}(0, \sigma)$
- Fit values of α, β by some optimization procedure (classically, minimize sum of squared errors)

Standardize your data

It's a good idea with this model to standardize the data (subtract the mean, divide by the standard deviation).

Why do this?

- computationally simpler – when fitting MAP estimate, the optimizer doesn't have to guess at the correct scale
- sometimes easier to interpret parameters

Regression models

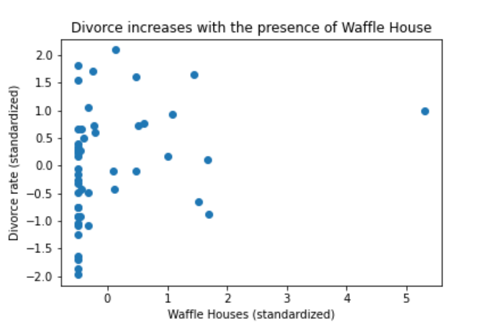
Our form of the regression model:

$$
\begin{aligned}
y_i|\beta, \sigma, x &\sim \mathrm{Normal}(\beta x, \sigma^2) \\
p(\beta, \sigma^2|X) &\propto \sigma^{-2}
\end{aligned}
$$

where we have taken the same improper prior as we did before for the normal model (because this regression really is a normal model!).

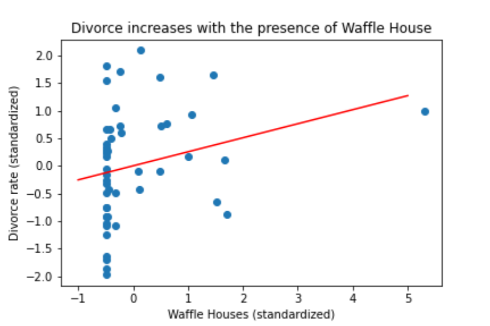
If you look in section 14.2 of the book, you will see this described for vector X.

Example



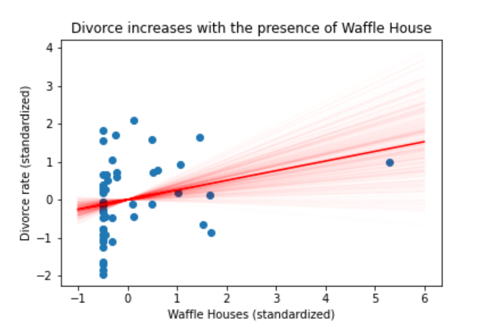
Note: some of these plots were made using a slightly different procedure / prior than described on these slides. Sorry! The main points stand.

Example



$\beta \approx 0.25$

Sampling from the posterior



Priors and regularization

Uniform prior

Previously, we took

$$
\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta x_i \\
\beta &\sim \mathrm{Uniform}(-\infty, \infty)
\end{aligned}
$$

Uniform($-\infty, \infty$) is an improper prior, but the likelihood will dominate, so we can get away with it (or we can cut off the uniform distribution at some sufficiently large limits)

Sampling from the prior

With standardized data, what values of β should be possible for a linear regression?

Sampling from the prior

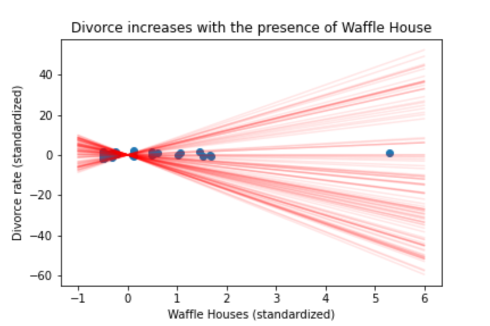
With standardized data, what values of β should be possible for a linear regression?

$\beta \in [-1, 1]$

Sampling from the prior

With standardized data, what values of β should be possible for a linear regression?

$\beta \in [-1, 1]$



Uniform prior

With a uniform prior on β, the joint likelihood is

$$p(\mathrm{y}|\beta) = \frac{1}{Z(\sigma, \beta)} \exp\left(-\frac{1}{\sigma^2}\sum_i (y_i - \beta x_i)^2\right)$$

Uniform prior

With a uniform prior on β, the joint likelihood is

$$p(\mathrm{y}|\beta) = \frac{1}{Z(\sigma, \beta)} \exp\left(-\frac{1}{\sigma^2}\sum_i (y_i - \beta x_i)^2\right)$$

This is a justification for the least squares procedure for fitting linear regressions – it's just the maximum likelihood estimate.

(This also shows why we assume constant variance for least squares... without it the MLE is trickier.)

Normal prior

With our Normal model:

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \beta x_i \\ \beta &\sim \text{Normal}(0, 1/\tau) \end{aligned}$$

τ is a parameter sometimes called the *precision* of a normal distribution (it measures inverse dispersion, instead of measuring dispersion as the standard deviation does)

Normal prior

Now the posterior looks a bit different:

$$
\begin{aligned}
p(\beta|\mathrm{y}) &\propto \exp\left(-\frac{1}{\sigma^2}\sum_i (y_i - \beta x_i)^2\right)\exp(-\frac{\tau^2}{\beta^2}) \\
&= \exp\left(-\frac{1}{\sigma^2}\left[\sum_i (y_i - \beta x_i)^2 + \sigma^2\tau^2\beta^2\right]\right)
\end{aligned}
$$

So the MAP estimate is the β which minimizes:

$$
L(\beta) = \sum_i (y_i - \beta x_i)^2 + \lambda\beta^2
$$

where we have absorbed the constants multiplying β^2 into a single parameter.

Recognize this?

Ridge regression

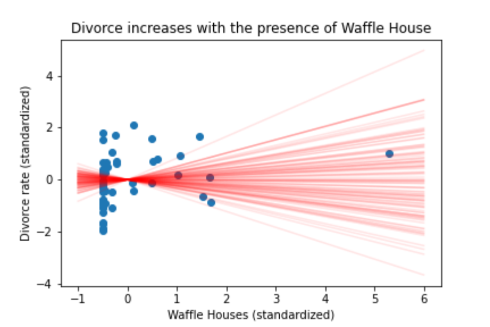
The model minimizing the loss function

$$L(\beta) = \sum_i (y_i - \beta x_i)^2 + \lambda \beta^2$$

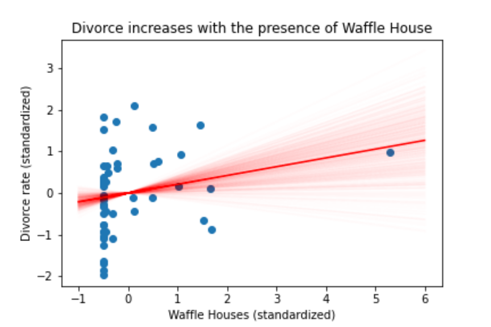
is known as "ridge regression" (or Tikhonov regularization) and is a common way to reduce variance and overfitting in linear models, especially those with several highly correlated predictors (multicollinearity).

What we've seen is that this regularization can be interpreted as placing a Normal prior on the model coefficients. The precision we specify for the prior controls the amount of regularization.

Sampling from the prior



Sampling from the posterior



Other regularizations

Do you know any similar approaches to regularizing linear models?

Other regularizations

Do you know any similar approaches to regularizing linear models?

Lasso regression: minimizing the loss function

$$L(\beta) = \sum_i (y_i - \beta x_i)^2 + \lambda|\beta|$$

Can this be written in terms of a prior on β? What prior do you need?

Priors and shrinkage

The effect of both of these types of regularization is to "shrink" model coefficients toward 0.

Notice we saw this effect in the kidney cancer Poisson model:

- estimated death rates (posterior means) shrank toward the national average
- estimates shrink more when there is less data available

Preview

Next week:

- Other regression models
- Multiple regression
- directed acyclic graphs and causal inference