THE UNIVERSITY OF TEXAS AT AUSTIN

Department of Statistics and Data Sciences

College of Natural Sciences

Summary: Bayesian linear models

Sinead Williamson

The University of Texas Department of Statistics and Data Science

Starting point: Bayesian linear regression

Basic model:

$$
\begin{aligned}
\mathbf{y}|\beta, X \sim& \text{Normal}(X\beta, (\omega\Lambda)^{-1}) \\
\beta \sim& \text{Normal}(\mu, (\omega K)^{-1}) \\
\omega \sim& \text{Gamma}(a, b)
\end{aligned}
$$

So,

$$
\begin{aligned}
p(\beta, \omega|\mathbf{y}) \propto& N(\mathbf{y}; X\beta, (\omega\Lambda)^{-1}) N(\beta; \mu, (\omega K)^{-1}) Ga(\omega; a, b) \\
\propto& \omega^{a+p/2+n/2-1} e^{-\frac{\omega}{2}\left((y-X\beta)^T\Lambda(y-X\beta)+(\beta-\mu)^T K(\beta-\mu)+2b\right)} \\
=& \omega^{a_n-1} e^{-\frac{\omega}{2}\left((\beta-\mu_n)^T K_n(\beta-\mu_n)\right)} \omega^{p/2} e^{-\omega b} \\
=& N(\beta; \mu_n, (\omega K_n)^{-1}) Ga(\omega; a_n, b_n)
\end{aligned}
$$

where:

- $K_n = X^T\Lambda X + K$
- $\mu_n = K_n^{-1}(X^T\Lambda y + K\mu)$
- $a_n = a + n/2$
- $b_n = b + \frac{1}{2}(y^T\Lambda y + \mu^T K\mu - \mu_n^T K_n \mu_n)$

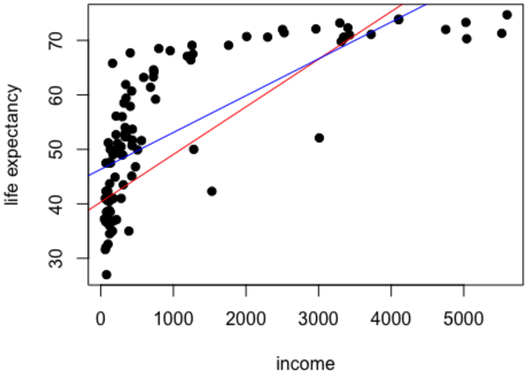
From here we can obtain the conditional and marginal posterior distributions:

$$
\begin{aligned}
p(\beta|\omega, y) =& N(\beta; \mu_n, (\omega K_n)^{-1}) \\
p(\omega|y) =& Ga(\omega; a_n, b_n) \\
p(\beta|y) \propto & \int_0^\infty p(\beta, \omega|y) d\omega \\
=& \int_0^n \omega^{a_n + p/2 - 1} \exp\left\{ -\frac{\omega}{2} \left((\beta - \mu_n)^T K_n(\beta - \mu_n) + 2b_n \right) \right\} d\omega \\
=& \Gamma(a_n + p/2) \left(\frac{(\beta - \mu_n)^T K_n(\beta - \mu_n) + 2b_n}{2} \right)^{-a_n - p/2} \\
\propto & \left(1 + \frac{1}{2a_n} \frac{\beta - \mu_n)^T K_n(\beta - \mu_n)|}{b_n/a_n} \right)^{-a_n + p/2}
\end{aligned}
$$

Specific example: countries' life expectancy

- Data: y = average lifespan, X = average income, plus intercept.

```
Lamb <- 0.1 * diag(n); K <- 0.1 * diag(p); mu <- numeric(2); a <- 1; b <- 1

K_n <- t(X)%*%Lamb %*%X + K
K_n_inv <- solve(K_n)
mu_n <- K_n_inv %*% (t(X) %*% Lamb %*% y + K %*% mu)
a_n <- a+n/2
b_n <- b + 0.5*(t(y)%*%y + t(mu)%*%K%*%mu - t(mu_n)%*%K_n%*%mu_n)

plot(X[,2],y,pch=19,xlab="income",ylab="life expectancy")
abline(mu_n[1],mu_n[2],col="red")

ls_model<-lm(life~income,data=life)
abline(ls_model,col="blue")
```



- Red = posterior mean
- Blue = LS

A heavier tailed model

Rather than have $\Lambda = \lambda I_n$, let's allow each country to have its own λ_i so that $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$:

$$
\begin{aligned}
y|\beta, \omega, \Lambda \sim& N(X\beta, (\omega\Lambda)^{-1} \\
\lambda_i \sim& \text{Gamma}(\tau, \tau) \\
\beta|\omega \sim& N(\mu, (\omega K)^{-1}) \\
\omega \sim& \text{Gamma}(a, b)
\end{aligned}
$$

The only new conditional is $p(\lambda_i|y, \beta, \omega)$

$$
\begin{aligned}
p(\lambda_i|y, \beta, \omega) \propto& p(y|\lambda_i, \beta, \omega)p(\lambda_i|\tau) \\
=& p(y_i|\lambda_i, \beta, \omega)p(\lambda_i|\tau) \\
\propto& \lambda_i^{\tau+1/2-1} \exp\left\{-\lambda_i \left(\frac{\omega(y_i - x_i^T\beta)^2}{2} + \tau\right)\right\} \\
\propto& \text{Gamma}\left(\lambda_i; \tau + 1/2, \tau + \frac{\omega(y_i - x_i^T\beta)^2}{2}\right)
\end{aligned}
$$

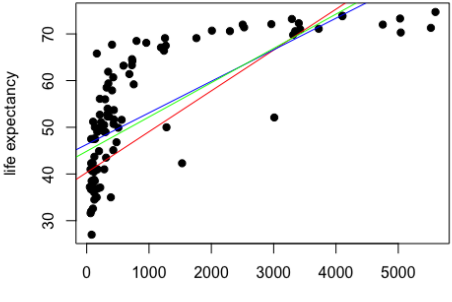
A Gibbs sampler for this model

- A Gibbs sampler generates a sequence of samples by iteratively sampling from the conditional distributions of each of the parameters.
- Asymptotically, this will generate samples from the posterior.
- In our case, we will sample from:
  - $\omega \sim \text{Gamma}(a_n, b_n)$
  - $\beta \sim \text{Normal}(\mu_n, (\omega K_n)^{-1})$
  - $\lambda_i \sim \text{Gamma}\Big(\Big(\lambda_i; \tau + 1/2, \tau + \frac{\omega(y_i - x_i^T\beta)^2}{2}\Big)$

  where
  - $K_n = X^T \Lambda X + K$
  - $\mu_n = K_n^{-1}(X^T \Lambda y + K\mu)$
  - $a_n = a + n/2$
  - $b_n = b + \frac{1}{2}(y^T \Lambda y + \mu^T K \mu - \mu_n^T K_n \mu_n)$

```
num_samples = 1000
betas <- matrix(nrow=p,ncol=num_samples)
omegas <-rep(NA,num_samples)
lambs <- matrix(nrow=num_samples,ncol=n)
omegas[1]=1
betas[,1]=0
lambs[1,]=.1
tau = 1
for (i in 2:num_samples){
  Lamb <- diag(lambs[i-1,])
  K_n <- t(X)%*%Lamb %*%X + K

  K_n_inv <- solve(K_n)
  mu_n <- K_n_inv %*% (t(X) %*% Lamb %*% y + K %*% mu)

  betas[,i] = mvrnorm(n=1,mu=mu_n,Sigma=(K_n_inv/omegas[i-1]))

  a_n <- a + n/2
  b_n <- b + 0.5*(t(y)%*%Lamb %*% y + t(mu)%*%K%*%mu - t(mu_n)%*%K_n%*%mu_n)
  omegas[i] = rgamma(n=1,shape=a_n, rate=b_n)

  lambda_rate = tau + 0.5*omegas[i]*(y - X %*% betas[,i])^2
  lambs[i,] = rgamma(n=n,shape=tau+0.5, rate = lambda_rate)
}
```

Comparison



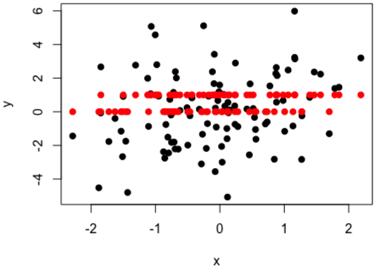
- Red = posterior mean (old model)
- Green = posterior mean (new model)
- Blue = LS

Generalized linear models

- With a Gaussian prior and a Gaussian likelihood, everything is easy!
- With a non-Gaussian likelihood, things get harder...
- Sometimes we can re-write our model in ways that give us conjugacy...
- In other cases, we will have to make approximations or resort to alternative MCMC methods.

Conjugacy in an auxiliary variable model: Probit regression

- Let's assume we have *latent* observations y_i generated according to a standard linear regression model, $y_i \sim \text{Normal}(x_i^T \beta, \sigma^2)$
- And then, our actual data z_i are set to 1 if $y_i > 0$, 0 otherwise.
- So, $\mathbf{P}(z_i = 1|\beta, x_i) = \Phi\left(\frac{x_i^T \beta}{\sigma}\right)$



Conjugacy in an auxiliary variable model: Probit regression

- Conditioned on the y_i, we just have a standard linear model.
- To sample the y_i, we need the conditional distribution.

$$p(y_i|\beta, x_i, \sigma) = Normal(y_i; x_i^T\beta, \sigma^2)$$

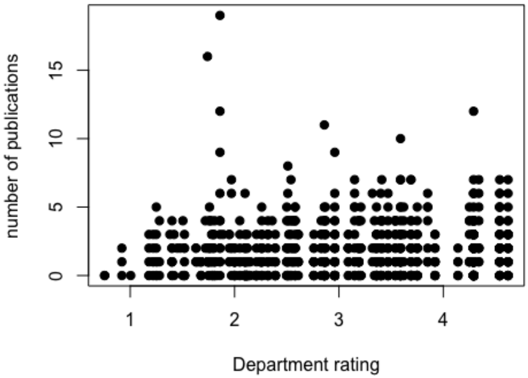
$$p(z_i|y_i) = \begin{cases} 1 & z_i = 1 \text{ and } y_i > 0 \\ 1 & z_i = 0 \text{ and } y_i < 0 \\ 0 & z_i = 1 \text{ and } y_i < 0 \\ 0 & z_i = 0 \text{ and } y_i > 0 \end{cases} \quad p(y_i|z_i\beta, x_i, \sigma) = \begin{cases} Trunc - N_{0,\infty}(y_i; x_i^T \\ Trunc - N_{-\infty,0}(y_i; x_i \\ 0 \end{cases}$$

No route to conjugacy... what next?

- In general, we won't have conjugacy.
- Let's consider a case of count data.
- A Poisson is a natural model... but we need to transform our parameter:

$$\beta \sim \text{Normal}(\mu, (\omega K)^{-1})$$
$$y_i \sim \text{Poisson}\left(\exp\{x_i^T \beta\}\right)$$

Dataset: Number of publications of bio students



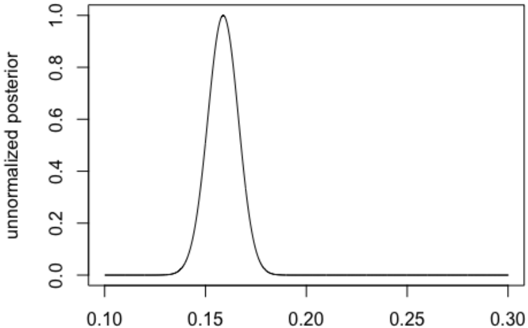
- Other predictors include gender, number of children, marital status, publications by advisor.

Looking at the posterior

- We know that

$$p(\beta|x, y) \propto Normal(\beta; \mu, (\omega K)^{-1}) \prod_{i=1}^{n} Poisson(y_i; \exp\{x_i^T \beta\}$$

- We can plot this (going to assume no intercept for now)...



Laplace's Approximation

- The Laplace transform is a way to approximate a posterior with a Gaussian.
- Let $P^*(\theta)$ be our unnormalized posterior, and let $\hat{\theta}$ be the value that maximizes the posterior.
- By a Taylor expansion,

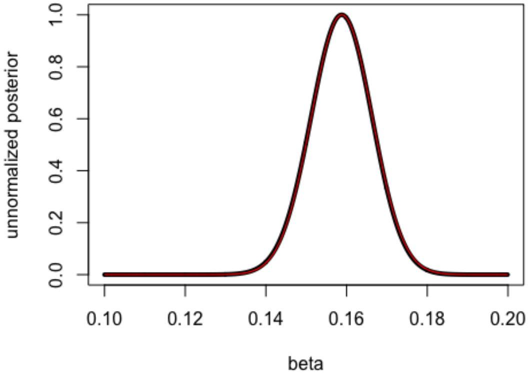
$$\begin{aligned}
\log P^*(\theta) &\approx \log P^*(\hat{\theta}) + (\theta - \hat{\theta})\frac{d}{d\theta}\log P^*(\theta)\Big|_{\theta=\hat{\theta}} + \frac{(\theta-\hat{\theta})^2}{2}\frac{d^2}{d\theta^2}\log P^*(\theta)\Big|_{\theta=\hat{\theta}} \\
&= \log P^*(\hat{\theta}) + \frac{(\theta-\hat{\theta})^2}{2}\frac{d^2}{d\theta^2}\log P^*(\theta)\Big|_{\theta=\hat{\theta}}
\end{aligned}$$

- This looks like the log pdf of a Gaussian, with precision $\frac{d^2}{d\theta^2}\log P^*(\theta)\Big|_{\theta=\hat{\theta}}$

Laplace's Approximation

- Using R's optimize with $\mu = 0, \sigma = 1$, $\hat{\beta} = 0.159$.
- We have $\log P^*(\beta) = \frac{(\beta-\mu)^2}{2\sigma^2} + \sum_{i=1}^{n} y_i x_i \beta - e^{x_i \beta}$
- First derivative: $\frac{\beta-\mu}{\sigma^2} + \sum_i y_i x_i - x_i e^{x_i \beta}$
- Second derivative: $\frac{1}{\sigma^2} - \sum_i x_i^2 e^{x_i \beta}$
- So, approximating precision is $\sum_i x_i^2 e^{0.159 x_i} - 1 = 17476.5$

Looking at the approximation

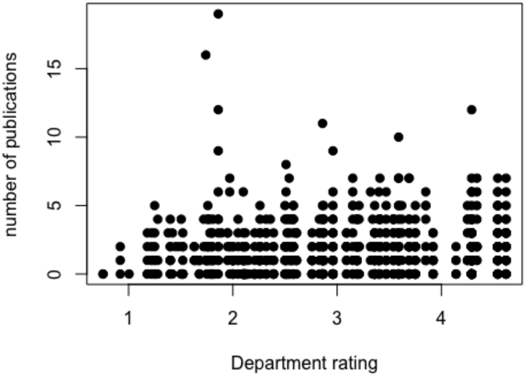


Multivariate case

- We can do the same in the multivariate case... we use the Hessian in place of the second derivative.
- Be careful with the cross terms!
- $\frac{d}{d\beta_j} \log P^*(\beta) = \frac{\beta_j - \mu_j}{\sigma^2} + \sum_i y_i x_{ij} - x_{ij} \exp\{x_i^T \beta\}$
- $\frac{d^2}{d\beta_j^2} = \frac{1}{\sigma^2} - x_{ij}^2 \exp\{x_i^T \beta\}$
- $\frac{d^2}{d\beta_j d\beta_k} = \frac{1}{\sigma^2} - x_{ij} x_{ik} \exp\{x_i^T \beta\}$

Improving our regression

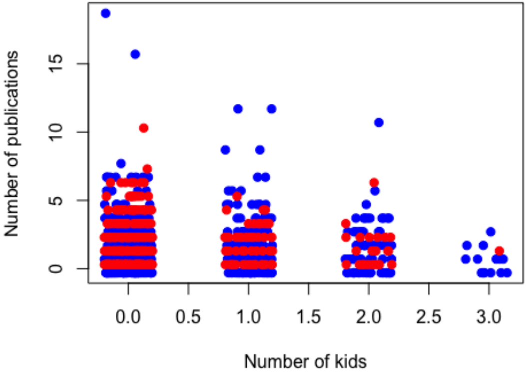
We might want to add things to our regression...



- How might we deal with heavy-tailed residuals?

Improving our regression

The effect of having kids seems to vary with gender... how could we capture this?



Projects

We need to start thinking about projects!

- Obvious suggestions: Regression / function learning in an interesting setting.
- Slightly trickier: Rates of events, causal inference, clustering, latent variable modeling.
- Appleseed will have some good examples that fall into the above.
- Kaggle is another good source of data.
- Or, you might have something from your own research.