Bayesian linear regressions

Jacob M. Montgomery

2017

Bayesian linear regression

The setup

$$y_i = \mathbf{x}_i' \beta + \epsilon_i$$

The setup

$$y_i = \mathbf{x}_i'\beta + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

The setup

$$y_i = \mathbf{x}_i' \beta + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

$$p(\mathbf{y}|\mathbf{x}, \beta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\right)$$

How to estimate this if the world were easy?

1. We have the likelihood
2. We just need the priors
3. And then we can can calculate the posterior.

How this is actually going to work

1. We have the likelihood
2. We add priors.
3. We can calculate the posterior for σ and can calculate the posterior β while holding the other constant.
4. So we first sample one, and then the other (composition or Gibbs sampling)

Decomposing the likelihood

- Note that it is possible to re-write the likelihood using the same "complete the squares" trick we have used all semester

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) + (\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta})$$

- Remember that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

- Note that $(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$ is just the sum of squared residuals.

- Note that $(\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta})$ is just the sum of squared residuals.
- Using the standard definition of $s^2 = \frac{SSE}{n-k}$ we can re-write this first term as $(n-k)s^2$
- Letting $\nu = n-k$ (the degrees of freedom), we get

$$(\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta}) = \nu s^2$$

Still decomposing the likelihood

- So now we can re-write the entirly likelihood as:

$$p(\mathbf{y}|\mathbf{x}, \beta, \sigma^2) \propto (\sigma^2)^{-\nu/2} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) (\sigma^2)^{\frac{-(n-\nu)}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta})\right)$$

- Let's do some chalkboard work on that

Still decomposing the likelihood

- So now we can re-write the entirly likelihood as:

$$p(\mathbf{y}|\mathbf{x}, \beta, \sigma^2) \propto (\sigma^2)^{-\nu/2} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) (\sigma^2)^{\frac{-(n-\nu)}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta})\right)$$

- Let's do some chalkboard work on that
- The key point here is that we can divide the likelihood into things one part that is related to σ and one part that has both σ and β
- AND we can recognize the kernal of some other distributions in each.

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\nu/2} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) (\sigma^2)^{\frac{-n-\nu}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right)$$

$$p(\mathbf{y}|\mathbf{x}, \beta, \sigma^2) \propto (\sigma^2)^{-\nu/2} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) (\sigma^2)^{\frac{-n-\nu}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta})\right)$$

- The easy one to see is that (thinking now of β as the random variable) this similar to the normal distribution

$$(\sigma^2)^{\frac{-n-\nu}{2}} \exp(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta}))$$

- Recalling, of course, that the multivariate normal distribution is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

where p is the number of independent parameters within the covariance matrix $\boldsymbol{\Sigma}$

$$p(\mathbf{y}|\mathbf{x},\beta,\sigma^2) \propto (\sigma^2)^{-\nu/2} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) (\sigma^2)^{\frac{-n-\nu}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta-\hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta-\hat{\beta})\right)$$

- The easy one to see is that (thinking now of β as the random variable) this similar to the normal distribution

$$(\sigma^2)^{\frac{-n-\nu}{2}} \exp(-\frac{1}{2\sigma^2}(\beta-\hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta-\hat{\beta}))$$

- Recalling, of course, that the multivariate normal distribution is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

where p is the number of independent parameters within the covariance matrix $\boldsymbol{\Sigma}$

- Let's chalk that a bit

- The harder one to see is this:

$$(\sigma^2)^{-\nu/2} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right)$$

- This has similiarties to the kernal of an inverse gamma distribution that takes the form:

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(\frac{b}{x})$$

So now we need priors

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)$$

So now we need priors

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)$$

$$\pi(\sigma^2) \propto (\sigma^2)^{\nu_0/2-1} \exp\left(-\frac{\nu_0 s_0^2}{2\sigma^2}\right)$$

So now we need priors

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)$$

$$\pi(\sigma^2) \propto (\sigma^2)^{\nu_0/2-1} \exp\left(-\frac{\nu_0 s_0^2}{2\sigma^2}\right)$$

- This is the same as the Inverse Gamma where $a_0 = \nu_0/2$ and $b_0 = 1/2\nu_0 s_0^2$ where we can interpret ν_0 as being prior the degrees of freedom and s_0 being the prior variance.

So now we need priors

$$\pi(\beta, \sigma^2) = \pi(\sigma^2)\pi(\beta|\sigma^2)$$

$$\pi(\sigma^2) \propto (\sigma^2)^{\nu_0/2-1} \exp\left(-\frac{\nu_0 s_0^2}{2\sigma^2}\right)$$

- This is the same as the Inverse Gamma where $a_0 = \nu_0/2$ and $b_0 = 1/2\nu_0 s_0^2$ where we can interpret ν_0 as being prior the degrees of freedom and s_0 being the prior variance.

- This can then be re-writen as

$$\pi(\sigma^2) \propto (\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2}\right)$$

$$\pi(\boldsymbol{\beta}|\sigma^2) \sim N(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{\Lambda}_0^{-1})$$

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto (\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'(\boldsymbol{\Lambda}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right)$$

- Compare to:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

- Let's assemble this whole mess on the chalkboard

$$p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2) \pi(\beta | \sigma^2) \pi(\sigma^2)$$

- Let's assemble this whole mess on the chalkboard

$$p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2) \pi(\beta | \sigma^2) \pi(\sigma^2)$$

- Now we need to re-arrange some terms

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\beta - \beta_0)' \mathbf{\Lambda}_0 (\beta - \beta_0)$$

- Let's assemble this whole mess on the chalkboard

$$p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2) \pi(\beta | \sigma^2) \pi(\sigma^2)$$

- Now we need to re-arrange some terms

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\beta - \beta_0)' \boldsymbol{\Lambda}_0 (\beta - \beta_0)$$

- Now we are going to introduce, essentially, the answer as: $\boldsymbol{\mu} = (\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda}_0)^{-1} (\mathbf{X}'\mathbf{X}\hat{\beta} + \boldsymbol{\Lambda}_0 \beta_0)$

- Let's assemble this whole mess on the chalkboard

$$p(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \beta, \sigma^2)\pi(\beta|\sigma^2)\pi(\sigma^2)$$

- Now we need to re-arrange some terms

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\beta - \beta_0)'\mathbf{\Lambda}_0(\beta - \beta_0)$$

- Now we are going to introduce, essentially, the answer as: $\boldsymbol{\mu} = (\mathbf{X}'\mathbf{X} + \mathbf{\Lambda}_0)^{-1}(\mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{\Lambda}_0\beta_0)$

- It turns out that we can re-write the top one in terms of μ as follows:

$$(\beta - \boldsymbol{\mu})^T(\mathbf{X}'\mathbf{X} + \mathbf{\Lambda}_0)(\beta - \boldsymbol{\mu}) + \mathbf{y}'\mathbf{y} - \boldsymbol{\mu}'(\mathbf{X}'\mathbf{X} + \mathbf{\Lambda}_0)\boldsymbol{\mu} + \beta_0'\mathbf{\Lambda}_0\beta_0$$

- Let's assemble this whole mess on the chalkboard

$$p(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \beta, \sigma^2)\pi(\beta|\sigma^2)\pi(\sigma^2)$$

- Now we need to re-arrange some terms

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\beta - \beta_0)'\mathbf{\Lambda}_0(\beta - \beta_0)$$

- Now we are going to introduce, essentially, the answer as: $\boldsymbol{\mu} = (\mathbf{X}'\mathbf{X} + \mathbf{\Lambda}_0)^{-1}(\mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{\Lambda}_0\beta_0)$

- It turns out that we can re-write the top one in terms of μ as follows:

$$(\beta - \boldsymbol{\mu})^T(\mathbf{X}'\mathbf{X} + \mathbf{\Lambda}_0)(\beta - \boldsymbol{\mu}) + \mathbf{y}'\mathbf{y} - \boldsymbol{\mu}'(\mathbf{X}'\mathbf{X} + \mathbf{\Lambda}_0)\boldsymbol{\mu} + \beta_0'\mathbf{\Lambda}_0\beta_0$$

- This is really just tedious algebra with some collecting of terms at the end. Let's do just a bit of this so you get the sense of it.

- So now we:
  - Take all of the things related to β and gather them in one exponent. The rest goes in the other.
  - We divide up the $(\sigma^2)^{whatever}$ into two parts, so that the the "normal" part is to the power $k/2$.

$$(\sigma^2)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu})'(\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda}_0)(\boldsymbol{\beta} - \boldsymbol{\mu})\right) \times$$

$$(\sigma^2)^{\frac{n+2a_0}{2}-1} \exp\left(\frac{2b_0 + \mathbf{y}'\mathbf{y} - \boldsymbol{\mu}'(\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda}_0)\boldsymbol{\mu} + \boldsymbol{\beta}_0\boldsymbol{\Lambda}_0\boldsymbol{\beta}_0}{2\sigma^2}\right)$$

- Thus,

$$p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto N(\boldsymbol{\mu}, \sigma^2 \tilde{\boldsymbol{\Sigma}}^{-1}) InvGamma(\tilde{a}, \tilde{b})$$

- Where

$$\tilde{\boldsymbol{\Sigma}} = (\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda}_0)$$

$$\boldsymbol{\mu} = (\mathbf{X}'\mathbf{X} + \boldsymbol{\Lambda}_0)^{-1}(\boldsymbol{\Lambda}_0 \beta_0 + \mathbf{X}'\mathbf{y})$$

$$\tilde{a} = a_0 + \frac{n}{2}$$

$$\tilde{b} = b_0 + \frac{1}{2}(\mathbf{y}'\mathbf{y} + \beta_0 \boldsymbol{\Lambda}_0 \beta_0 - \boldsymbol{\mu} \tilde{\boldsymbol{\Sigma}} \boldsymbol{\mu})$$

How to sample

1. Take m draws from the inverse gamma.
2. For each value of σ^2 drawn, then draw from the appropriate normal.
3. Profit.

How to sample

1. Take m draws from the inverse gamma.
2. For each value of σ^2 drawn, then draw from the appropriate normal.
3. Profit.

We could also marginalize out σ^2 by doing the indefinite integral. This would give us a multivariate t distribution.

Deep thoughts

- So what happpens if we exclude the priors? That is, what are the expected values for β and σ^2 if we ignored prior information?

Deep thoughts

- So what happpens if we exclude the priors? That is, what are the expected values for β and σ^2 if we ignored prior information?
- Thus, what are the priors doing?
- And what would this be in particular if we used this as a prior:

$$N(\mathbf{0}, \lambda \mathbf{I})$$

Ridge regression

- Hoerl and Kennard (1970) came up with the idea that we want to do a regression with *lots* of parameters.
- In that case, $X'X$ may not be invertable.

Ridge regression

- Hoerl and Kennard (1970) came up with the idea that we want to do a regression with *lots* of parameters.
- In that case, $X'X$ may not be invertable.
- So what if we set up this equation instead?

$$\sum (y_i - \mathbf{x}_i \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2$$

- This is the same as doing the normal optimization subject to, for some $c > 0$, $\sum_{j=1}^{p} \beta_j^2 < c$

Ridge regression

- Hoerl and Kennard (1970) came up with the idea that we want to do a regression with *lots* of parameters.
- In that case, $X'X$ may not be invertable.
- So what if we set up this equation instead?

$$\sum (y_i - \mathbf{x}_i\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2$$

- This is the same as doing the normal optimization subject to, for some $c > 0$, $\sum_{j=1}^{p} \beta_j^2 < c$
- Turns out that if we do this we get

$$\hat{\beta}_{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

- It also can be shown that this has better MSE for some value of λ than OLS (although we don't know for what value)

Geometry of ridge