Linear regression

Patrick Breheny

February 7

Introduction

- Our topic for today is linear regression
- Many of the results will be similar to what we have seen for the iid normal case – albeit with the multivariate normal distribution replacing the normal – although we will encounter a few interesting differences
- Like the iid normal case, linear regression for a normally distributed outcome does not have a fully conjugate prior, but is semi-conjugate with a multivariate normal prior on $\boldsymbol{\beta}$ and a gamma/scaled χ^2 prior on τ

Posterior for $\boldsymbol{\beta}|\mathbf{y}, \tau$

- Suppose

$$\mathbf{y}|\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I})$$
$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Omega}_0^{-1})$$

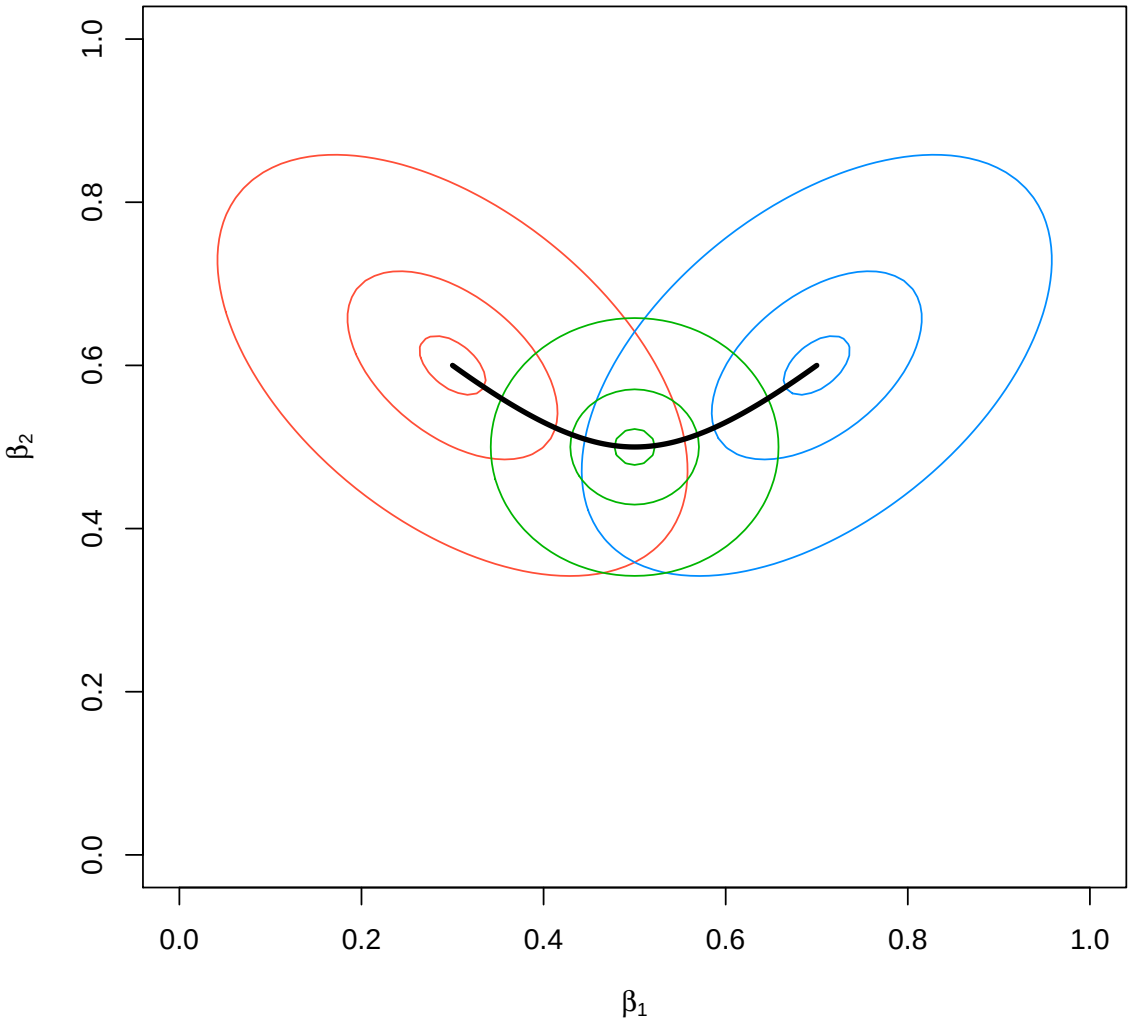
- Then, conditioning on τ, we have

$$\boldsymbol{\beta}|\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}_n, \boldsymbol{\Omega}_n^{-1}),$$

where

$$\begin{aligned}\boldsymbol{\Omega}_n &= \boldsymbol{\Omega}_0 + \tau\mathbf{X}'\mathbf{X} \\ \boldsymbol{\mu}_n &= \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\Omega}_0\boldsymbol{\mu}_0 + \tau\mathbf{X}'\mathbf{y}) \\ &= \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\Omega}_0\boldsymbol{\mu}_0 + \tau\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\end{aligned}$$

Geometry



Connection w/ ridge regression

- A common skeptical prior is $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\Omega}_0 = \omega_0 \mathbf{I}$; *i.e.*, the elements of $\boldsymbol{\beta}$ follow independent normal distributions centered at zero
- In this case,

$$\boldsymbol{\mu}_n = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y},$$

  where $\lambda = \omega_0 / \tau$; this is the ridge regression estimator
- Note that a "skeptical" prior on the intercept wouldn't really make any sense; instead, one would presumably use an uninformative prior on β_0 and have the rest follow $\mathrm{N}(0, \omega^{-1})$ distributions, as is common practice in ridge regression
- Note also that $\boldsymbol{\beta}$ is not scale-invariant, and assuming a common ω_0 for all β_j may not be appropriate

Updating τ/σ^2

- The preceding remarks have focused on $\boldsymbol{\beta}$, holding τ fixed
- For updating τ, we have the exact same result as for linear regression:

$$\begin{aligned}\mathbf{y}|\tau &\sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}) \\ \tau &\sim \text{Scaled-}\chi^2(n_0, \mathrm{RSS}_0)\end{aligned}$$

implies

$$\tau|\mathbf{y} \sim \text{Scaled-}\chi^2(n_0 + n, \mathrm{RSS}_0 + \mathrm{RSS}),$$

where again, in a Gibbs sampler, RSS is calculated conditional on fixing $\boldsymbol{\beta}$ at its most recent value

Alcohol metabolism data

- Let's apply these methods to a study involving 18 women and 14 men of why women exhibit a lower tolerance for alcohol and develop alcohol-related liver disease more readily than men
- The data set contains the following variables:
  - `Metabol`: First-pass metabolism of alcohol in the stomach (mmol/liter-hour); this is the outcome variable
  - `Gastric`: Gastric alcohol dehydrogenase activity in the stomach (μmol/min/g of tissue)
  - `Sex`: Sex of the subject
  - `Alcohol`: Whether the subject is alcoholic or not

Model

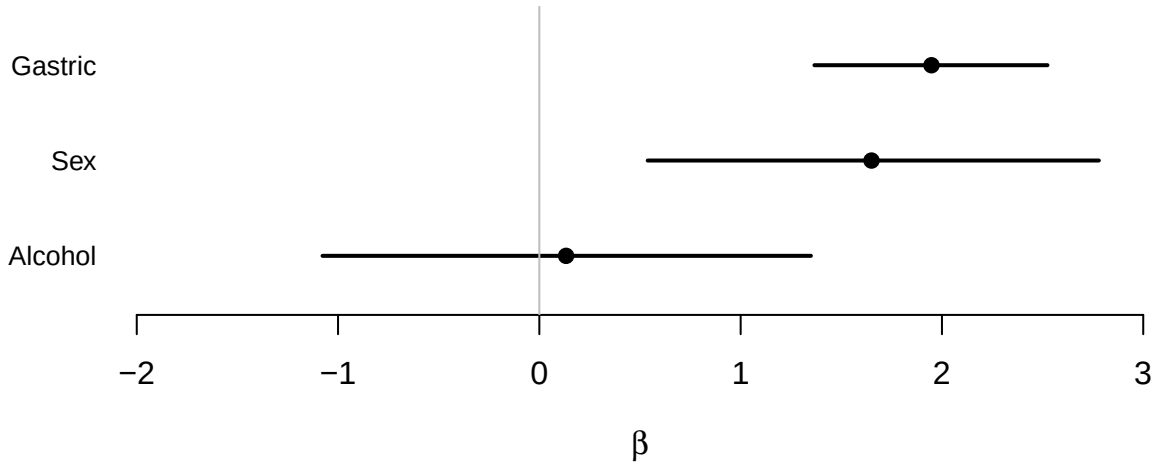
- There are interesting questions of interactions in this data set, particularly between sex and dehydrogenase activity, but we will focus only on the main effects here
- We will fit the following model:

$$\texttt{Metabol} = \beta_0 + \beta_1 \texttt{Gastric} + \beta_2 \texttt{Sex} + \beta_3 \texttt{Alcohol}$$

- We will explore the use of both uninformative/reference priors for $\boldsymbol{\beta}$ and the use of the skeptical (ridge) prior discussed earlier, with $\beta_j \sim \mathrm{N}(0, \mathrm{Var}(\mathbf{x}_j)^{-1})$ for $j \neq 0$
- The standard deviation, σ, will have an uninformative prior in both models

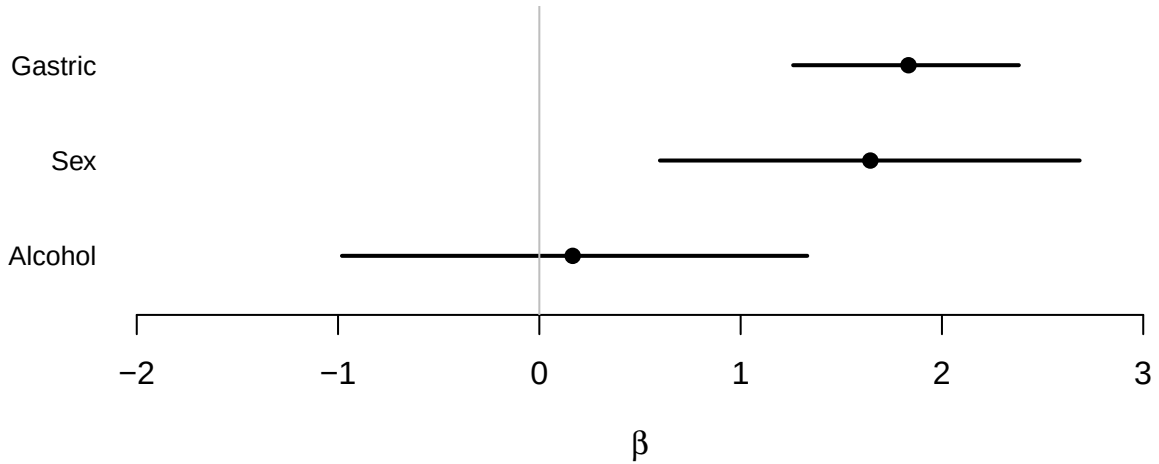
Results: Reference prior

	mean	sd	2.5%	25%	50%	75%	97.5%
β_0	-2.02	0.66	-3.33	-2.47	-2.02	-1.59	-0.73
β_1	1.95	0.30	1.36	1.75	1.95	2.15	2.54
β_2	1.65	0.58	0.54	1.27	1.65	2.03	2.81
β_3	0.13	0.62	-1.11	-0.27	0.12	0.53	1.35
σ	1.39	0.19	1.08	1.26	1.37	1.51	1.81



Results: Skeptical prior

	mean	sd	2.5%	25%	50%	75%	97.5%
β_0	-1.84	0.65	-3.12	-2.28	-1.84	-1.42	-0.54
β_1	1.83	0.29	1.26	1.65	1.84	2.03	2.38
β_2	1.65	0.54	0.58	1.30	1.65	2.01	2.72
β_3	0.17	0.59	-1.00	-0.22	0.17	0.56	1.35
σ	1.39	0.19	1.07	1.25	1.37	1.50	1.81



Remarks

- Note that the skeptical prior tends to shrink the posterior mean back towards 0, but also reduces the posterior variance
- This trade (sacrificing bias to reduce variance) is also the central idea in ridge regression
- Note also that this is only a tendency, and depends the correlation between the explanatory variables – in this example, the posterior mean of β_3 was larger in the skeptical model
- In this particular example, the posterior for σ was virtually identical in both models, although this is not always the case

Identifiability

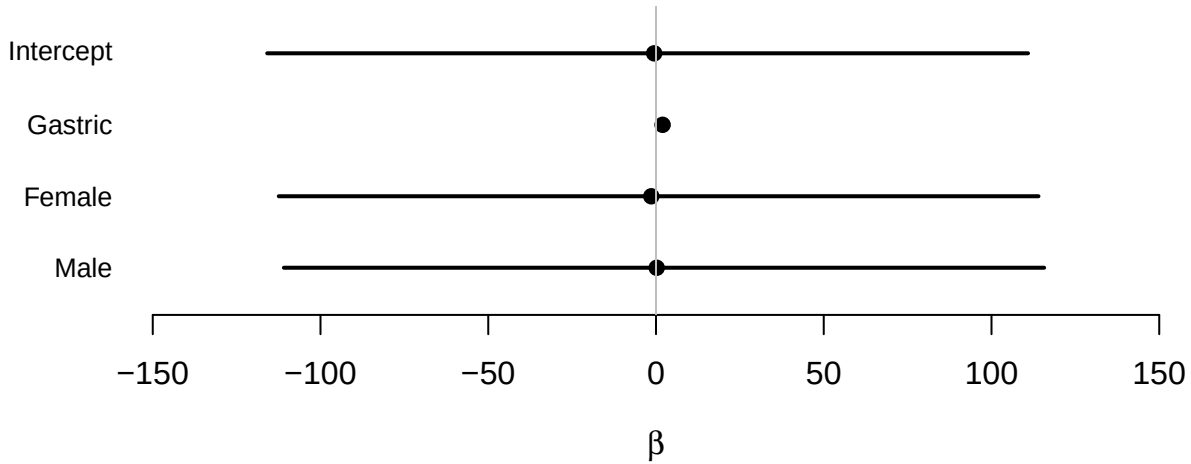
- Recall that the least-squares solution is unique only if the design matrix $\mathbf{X}$ is full-rank
- Thus, for example, we cannot fit the following model with ordinary linear regression and obtain a unique solution $\widehat{\boldsymbol{\beta}}$:

$$\texttt{Metabol} = \beta_0 + \beta_1 \texttt{Gastric} + \beta_2 \texttt{Female} + \beta_3 \texttt{Male},$$

  where `Male` and `Female` are indicator functions, as the columns of $\mathbf{X}$ are linearly dependent
- What happens with a Bayesian regression model?

Uninformative prior, non-identifiable model

	mean	sd	2.5%	25%	50%	75%	97.5%
β_0	-0.42	57.58	-113.37	-39.58	-0.29	38.31	112.66
β_1	1.96	0.28	1.41	1.79	1.96	2.15	2.50
β_2	-1.53	57.58	-114.87	-40.08	-1.62	37.74	111.21
β_3	0.09	57.59	-113.08	-38.57	0.06	39.38	113.01
σ	1.36	0.18	1.06	1.23	1.34	1.48	1.78

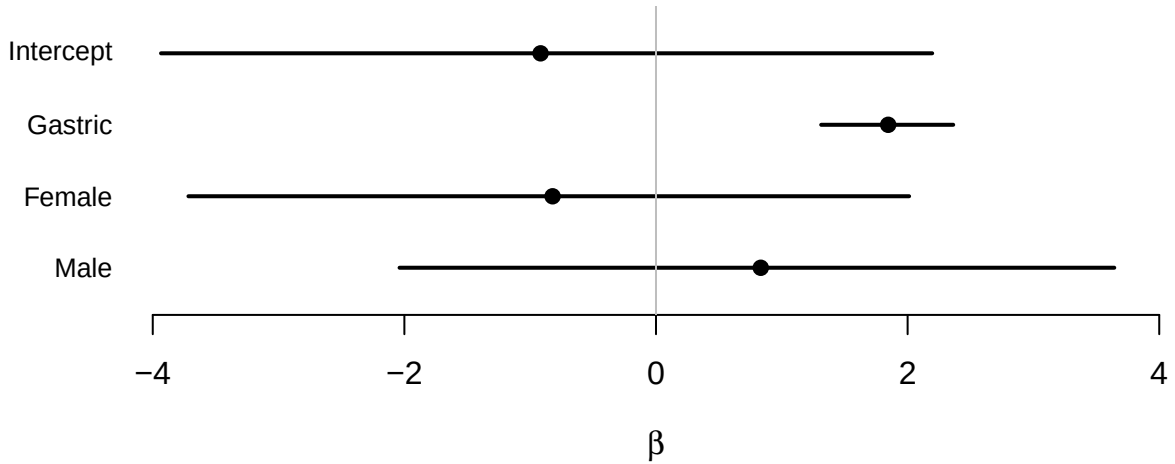


Remarks

- Note that the distributions for β_0, β_2, and β_3 (the three columns that were linearly dependent) become extremely wide, to the point where we cannot say that we know anything about these three parameters
- The posteriors for σ and β_1, on the other hand, are unaffected
- There is a certain similarity here to our "bad" Gibbs sampler from the normal distribution lecture, in that our sampler is meandering about aimlessly
- In that example, however, the posterior was perfectly well-defined – it was our sampler that failed us; here, this is actually what the posterior looks like – we just have a questionable model

Skeptical prior, non-identifiable model

	mean	sd	2.5%	25%	50%	75%	97.5%
β_0	-0.94	1.53	-3.88	-1.99	-0.96	0.09	2.04
β_1	1.85	0.27	1.31	1.67	1.85	2.03	2.37
β_2	-0.80	1.45	-3.61	-1.79	-0.78	0.18	1.98
β_3	0.85	1.44	-1.99	-0.12	0.86	1.83	3.62
σ	1.37	0.19	1.06	1.24	1.35	1.48	1.80



Remarks

- Note that "identifiability" in Bayesian regression is less black-and-white than it is in maximum likelihood estimation: having a somewhat informative prior can render a non-identifiable likelihood identifiable
- This, again, is similar to what is accomplished by ridge regression, which lends stability to models that would otherwise be overwhelmed by multicollinearity
- The posterior for β_0, β_2, and β_3 is still pretty wide, however – this model is better, but still not very good

Identifiability restriction

- Nevertheless, the unidentifiable model is on to something: the coefficients in our model would be more easily interpretable if the intercept represented an overall average, and we had parameters for males and females that represented their departure from this overall average
- What we require, however, is a restriction that keeps the model identifiable:

$$\beta_3 = \beta_0 + \delta/2$$
$$\beta_2 = \beta_0 - \delta/2$$

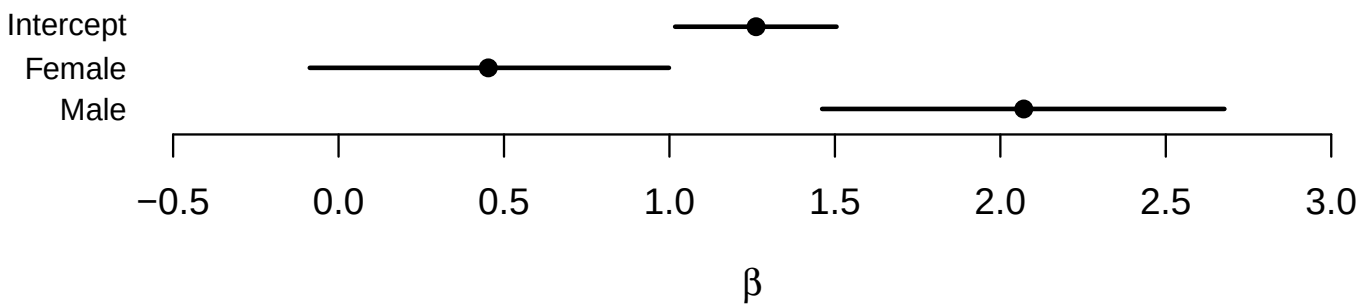
- With this parameterization in place, the likelihood for (β_0, δ) is identifiable, and (β_2, β_3) are determined by simple transformation, giving them a reasonable posterior as well

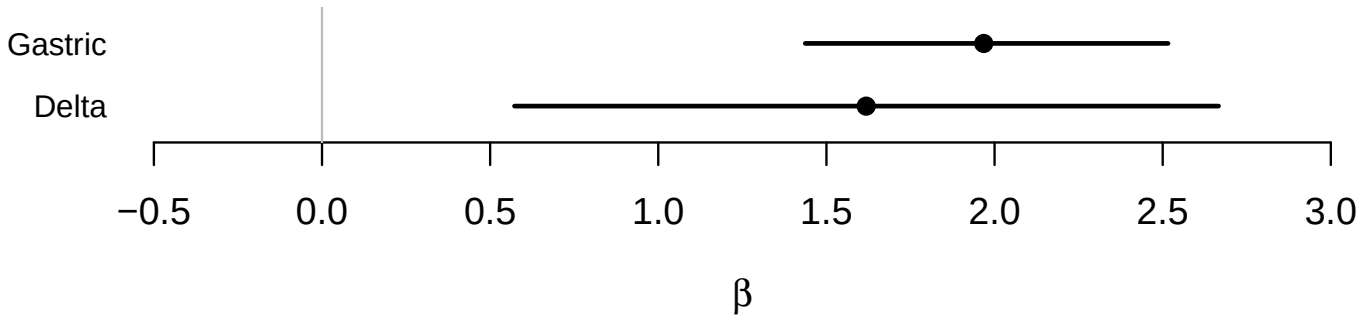
Centering

- While we're at it, we might want to subtract off the mean from `Gastric`, to increase the interpretability of β_0 as the population average metabolism (average of males and females, with average dehydrogenase levels)
- Note that this does not change the meaning of β_1, which still represents the change in metabolism caused by a one-unit change in dehydrogenase levels, but will affect the posterior of β_0
- This is referred to as *centering* a variable, and has many advantages; we will discuss it further in a few weeks

Uninformative prior, reparameterized model

	mean	sd	2.5%	25%	50%	75%	97.5%
β_0	1.26	0.12	1.03	1.18	1.26	1.34	1.51
β_1	1.97	0.28	1.41	1.78	1.97	2.15	2.50
β_2	0.45	0.27	-0.08	0.27	0.45	0.64	1.00
β_3	2.07	0.30	1.47	1.87	2.07	2.27	2.67
δ	1.62	0.53	0.59	1.26	1.62	1.97	2.65
σ	1.37	0.19	1.07	1.24	1.35	1.48	1.80





Remarks

- Note that we could have obtained these results by fitting the original model and deriving the male and female parameters as functions of the original coefficients, as is typically done in maximum likelihood estimation
- The difference, however, is that MCMC methods allow such constraints to be built directly into the model without any added complications; this is not the case for maximum likelihood
- As we will see, such intentionally overparameterized models (with suitable identifiability restrictions) will prove quite helpful in hierarchical modeling