Multiparameter Models - Normal Data, Poisson Regression

Statistics 220

Spring 2005

Normal Inference Models - Semiconjugate Prior

Another popular prior is

$$\begin{aligned} \mu|\sigma^2 &\sim N(\mu_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inv}-\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

In this case, μ and σ^2 are independent apriori. This prior is useful when the prior information on μ isn't thought of in terms of a number of prior measurements.

Note that this isn't a conjugate prior. The posterior is not the product of normal and $\text{Inv}-\chi^2$ densities. In fact the posterior is not particularly nice, in that parts of it do not reduce to standard densities.

$p(\mu|\sigma^2, y)$:

Given that σ^2 is fixed, this is a case we have already seen

$$\mu|\sigma^2, y \sim N(\mu_n, \tau_n^2)$$

where

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

This gives an idea where the term semi-conjugate comes from. If we consider the posterior distribution of one parameter conditional on the other parameters, the posterior is of the same form as the prior.

$p(\sigma^2|\mu, y)$:

Similarly

$$\sigma^2|\mu, y \sim \text{Inv}{-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + nv}{\nu_0 + n}\right)$$

where

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

$p(\sigma^2|y)$:

Here is where the nice distributional results breakdown.

$$\begin{aligned}
\sigma^2|y &\propto \int \text{Inv}{-}\chi^2(\sigma^2|\nu_0, \sigma_0^2) N(\mu|\mu_0, \tau_0^2) \prod_{i=1}^{n} N(y_i|\mu, \sigma^2) d\mu \\
&\propto \text{Inv}{-}\chi^2\left(\sigma^2|\nu_0 + n, \frac{\nu_0\sigma_0^2 + (n-1)s^2}{\nu_0 + n}\right) \int N(\mu|\mu_0, \tau_0^2) N\left(\bar{y}|\mu, \frac{\sigma^2}{n}\right) d\mu
\end{aligned}$$

Since the part inside the integral is proportional to a normal density, the density $p(\sigma^2|y)$ can be calculated in closed form.

Unfortunately this isn't a standard density. However we can get a handle on it based on the fact

$$p(\sigma^2|y) = \frac{p(\mu, \sigma^2|y)}{p(\mu|\sigma^2, y)}$$

This comes directly from

$$p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)p(\sigma^2|y)$$

So

$$p(\sigma^2|y) \propto \frac{N(\mu|\mu_0, \tau_0^2)\text{Inv}{-}\chi^2(\sigma^2|\nu_0, \sigma_0^2)\prod_{i=1}^n N(y_i|\mu, \sigma^2)}{N(\mu|\mu_n, \tau_n^2)}$$

While it appears that this depends on μ, it actually doesn't so we can pick any value of μ to make computation as easy as possible. A good choice is to evaluate this at $\mu = \mu_n$, giving

$$p(\sigma^2|y) \propto \tau_n N(\mu_n|\mu_0, \tau_0^2)\text{Inv}{-}\chi^2(\sigma^2|\nu_0, \sigma_0^2)\prod_{i=1}^n N(y_i|\mu_n, \sigma^2)$$

$p(\mu|y)$:

This is even uglier. While it appears that a closed form solution to the integral

$$p(\mu|y) \propto \int p(\mu, \sigma^2|y) d\sigma^2$$

is possible (the integrand is proportion to an inverse gamma density), this is usually handled by simulation. On approach is the two stage simulation approach mentioned before

1. Simulate $\sigma_1^2, \ldots, \sigma_m^2 \overset{iid}{\sim} \sigma^2|y$

2. Simulate $\mu_i \sim \mu|\sigma_i^2, y = N(\mu_i, \tau_i^2)$

Step 1 could be done by an acceptance-rejection method or by the grid simulation approach discussed in the text.

An alternative approach would be to use a Gibbs sampler for both μ and σ^2.

This approach will be taken for the example.

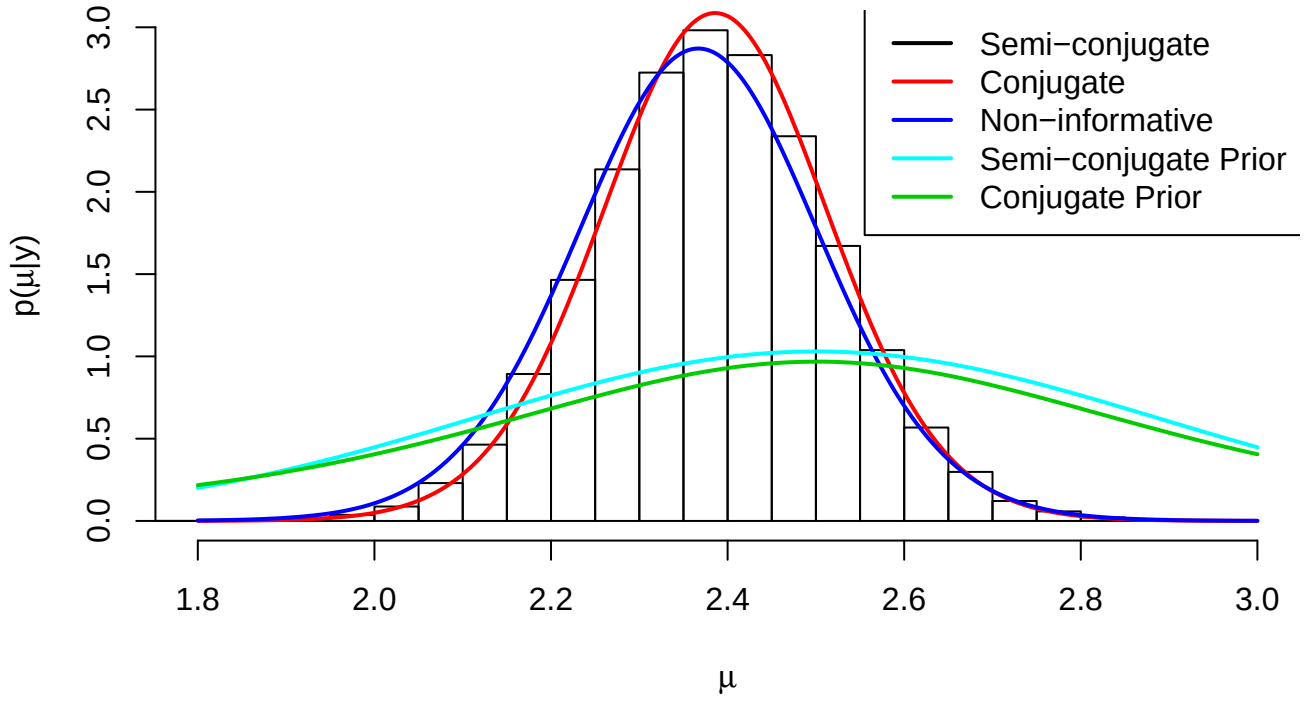
The following prior was chosen, trying to match the conjugate prior used last class

$$\mu|\sigma^2 \sim N\left(2.5, 0.15 = \frac{0.75}{5}\right) \qquad \sigma^2 \sim \text{Inv}{-}\chi^2(4, 0.75)$$

In this case μ_0 was set to μ_0 and τ_0^2 was set to $\frac{\sigma_0^2}{\kappa_0}$ as used in the conjugate prior. The prior on σ^2 was the same in each case.

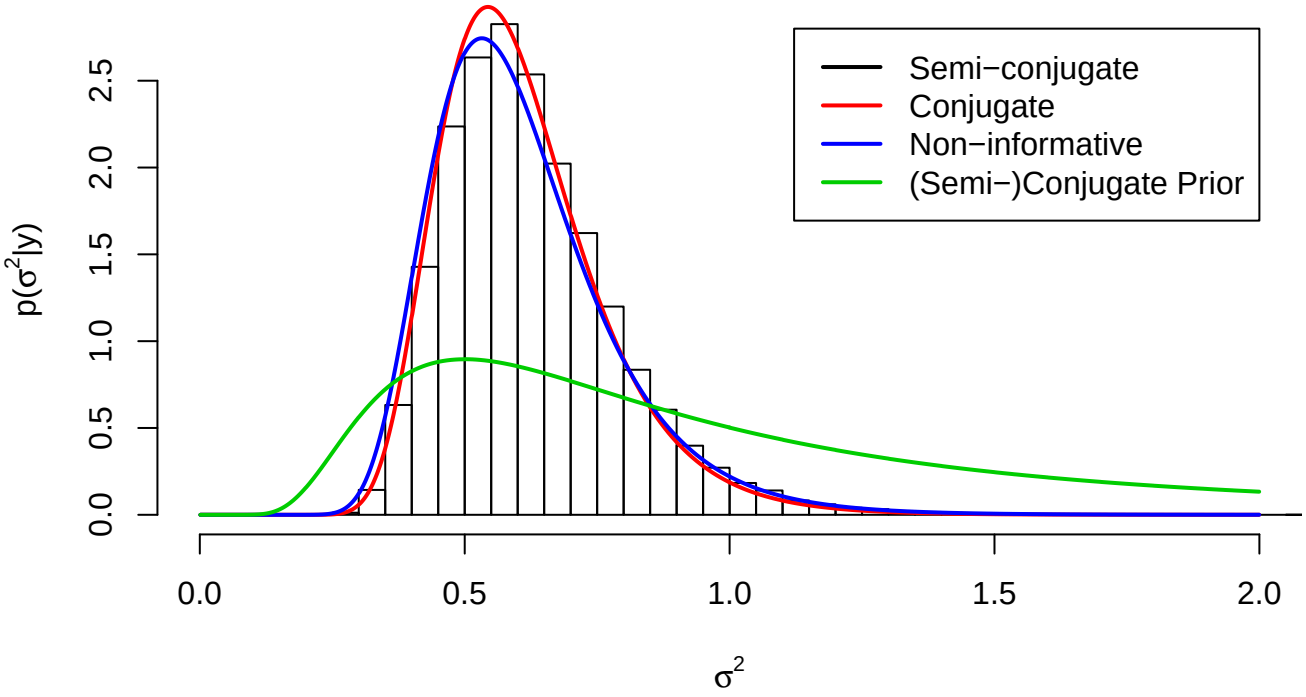
The results are based on 500,000 imputations.

$\mu|y$:



Prior	$E[\mu\|y]$	$\mathrm{SD}(\mu\|y)$	95% Cred. Int.
Non-informative	2.367	0.143	(2.085, 2.649)
Conjugate	2.386	0.132	(2.125, 2.647)
Semi-conjugate	2.383	0.135	(2.119, 2.650)

$\sigma^2|y$



Prior	$E[\sigma^2\|y]$	$\mathrm{SD}(\sigma^2\|y)$	95% Cred. Int.
Non-informative	0.6114	0.1729	(0.3610, 1.0287)
Conjugate	0.6120	0.1580	(0.3768, 0.9887)
Semi-conjugate	0.6270	0.1641	(0.3835 1.0189)

In this case, the two informative priors give similar answers, though the semi-conjugate prior seems to give slightly larger answers for σ^2. This isn't particularly surprising as the form of the $N-\text{Inv}-\chi^2$ distribution should lead to small values of μ pulling down σ^2. In this case, the data suggests that μ should be a bit lower than the prior specified.

Though one surprising result is that the posterior correlation between μ and σ^2 seems larger in the semi-conjugate case ($r = 0.0253$) than in the conjugate case ($r = 0.0014$).

However there is a suggestion that there might be a problem with the simulation calculating these (particularly in the conjugate case) so this might be taken with a grain of salt).

Multivariate Normal Models

y is a vector of length d with mean vector μ (also of length d and $d \times d$ variance matrix Σ, $(y|\mu, \Sigma \sim N_d(\mu, \Sigma))$. The density of a single observation is

$$p(y|\mu, \Sigma) \propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)$$

where $|\Sigma|$ is the determinant of the matrix Σ.

The likelihood of n iid observations is

$$\begin{aligned} p(y_1, \ldots, y_n|\mu, \Sigma) &\propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right) \\ &= |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S_0)\right) \end{aligned}$$

where $\mathrm{tr}(A)$ is the trace of the matrix A (the sum of the diagonal entries) and

$$S_0 = \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^T$$

So the density and likelihood look like what we get in the univariate case, but with matrix and vectors instead.

Note that most of the inference in this model is a direct analogue to the univariate case. However we need a multivariate analogue to the χ^2 and $\mathrm{Inv}{-}\chi^2$ distributions.

Wishart and Inverse Wishart Distributions

- Wishart distribution ($\text{Wishart}_\nu(\Lambda)$)

  Multivariate analogue of a scaled χ^2 distribution

  If $z_1, \ldots, z_\nu \overset{iid}{\sim} N_d(0, \Lambda)$ then

$$\Sigma = \sum_{i=1}^{\nu} z_i z_i^T \sim \text{Wishart}_\nu(\Lambda)$$

  like $z_1, \ldots, z_\nu \overset{iid}{\sim} N(0, \tau^2)$ then

$$S = \sum_{i=1}^{\nu} z_i^2 \sim \tau^2 \chi^2_\nu$$

- Inverse Wishart distribution ($\mathrm{Inv}{-}\mathrm{Wishart}_\nu(\Lambda^{-1})$)

  Multivariate analogue of a scaled $\mathrm{Inv}{-}\chi^2$ distribution

  If $\Sigma \sim \mathrm{Wishart}_\nu(\Lambda)$ then

$$\Sigma^{-1} \sim \mathrm{Inv}{-}\mathrm{Wishart}_\nu(\Lambda^{-1})$$

Common Multivariate Normal Models

- Unknown mean but known variance

$$\mu|\Sigma \sim N(\mu_0, \Lambda_0)$$

$$\mu|\Sigma, y \sim N(\mu_n, \Lambda_n)$$

where

$$\begin{aligned} \mu_n &= (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}) \\ \Lambda_n^{-1} &= \Lambda_0^{-1} + n\Sigma^{-1} \end{aligned}$$

Like the univariate case, the posterior mean is a weighted average of the prior mean and the sample average and the posterior precision matrix is the prior 'precision matrix + data precision matrix.

- Unknown mean and variance - conjugate prior

$$\begin{aligned} \Sigma &\sim \text{Inv}-\text{Wishart}_{\nu_0}(\Lambda_0^{-1}) \\ \mu|\Sigma &\sim N(\mu_0, \Sigma/\kappa_0) \end{aligned}$$

The posterior distribution satisfies

$$\begin{aligned} \Sigma|y &\sim \text{Inv}-\text{Wishart}_{\nu_n}(\Lambda_n^{-1}) \\ \mu|\Sigma, y &\sim N(\mu_n, \Sigma/\kappa_n) \end{aligned}$$

where

$$
\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T \\
S &= \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T
\end{aligned}
$$

In addition, it is possible to integrate out the variance matrix showing that

$$\mu|y \sim t_{\nu_n - d + 1}(\mu_n, \Lambda_n/(\kappa_n(\nu_n - d + 1)))$$

(i.e. multivariate t with $\nu_n - d + 1$ degrees of freedom)

- Unknown mean and variance - non-informative prior

$$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$$

which is the Jeffreys' prior and is the limit of the conjugate prior as $\kappa_0 \to 0$, $\nu_0 \to -1$, and $|\Sigma_0| \to 0$.

The posterior in this case satisfies

$$\begin{aligned} \Sigma|y &\sim \text{Inv}-\text{Wishart}_{n-1}(S) \\ \mu|\Sigma, y &\sim N(\bar{y}, \Sigma/n) \end{aligned}$$

Similarly to the univariate case,

$$\mu|y \sim t_{n-d}(\bar{y}, S/(n(n-d)))$$

Poisson Regression

Example: Geriatric study

A researcher in geriatrics designed a 6 month prospective study on $n = 100$ subjects to investigate the effects of two interventions on the frequency of falls. We will examine the effect of the intervention along with one of the covariates (Strength index) believed to be associated with the number of falls.

Data model: (y_i = number of falls during study, z_i = Intervention, x_i = Balance index)

$$\begin{aligned} y_i|\lambda_i &\overset{ind}{\sim} Pois(\lambda_i) \\ \log \lambda_i &= a + bx_i + cz_i \end{aligned}$$

Prior:

Assume a, b, and c are independent with

$$\begin{aligned} a &\sim N(0, 100) \\ b &\sim N(0, 100) \\ c &\sim N(0, 100) \end{aligned}$$

This is intended to be a fairly non-informative prior and clearly isn't a conjugate. The posterior distribution is of the form

$$p(a, b, c|y) \propto e^{-a^2/200} e^{-b^2/200} e^{-c^2/200} \prod_{i=1}^{n} e^{(a+bx_i+cz_i)y_i} e^{-e^{a+bx_i+cz_i}}$$

Given the form of this posterior, it will need to be examined by simulation. 5000 samples will be generated by the Gibbs sampler.

Questions of interest:

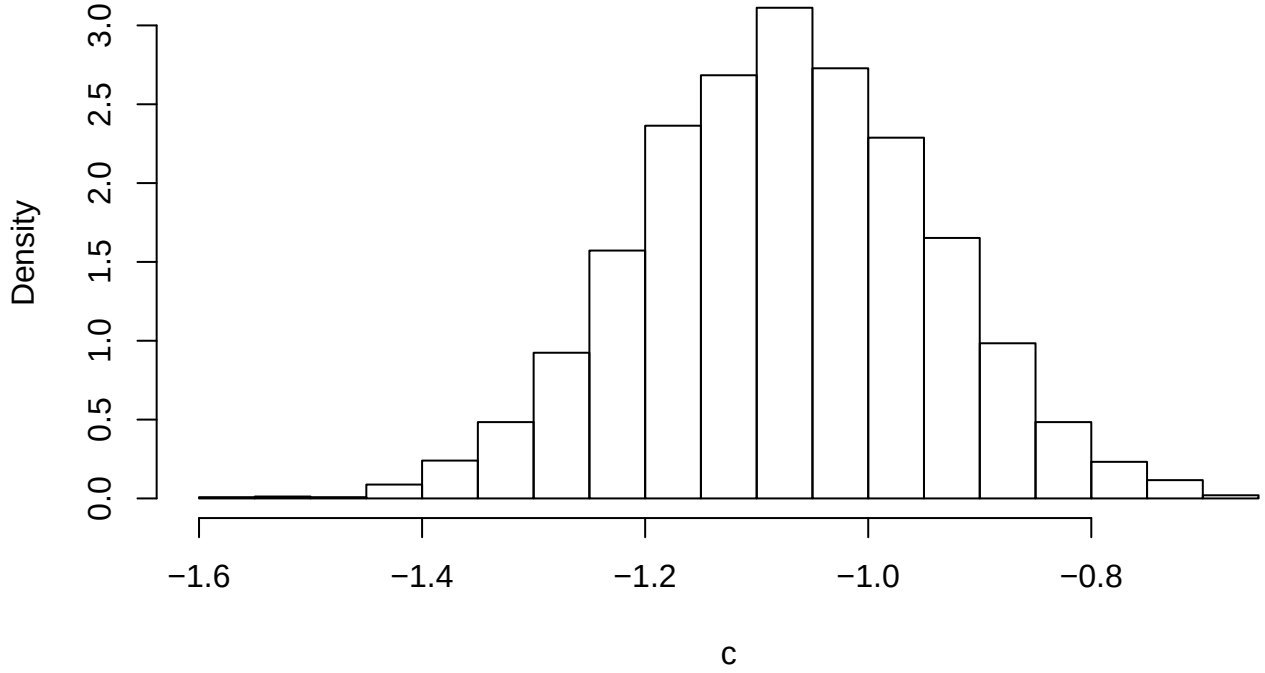
- $p(c|y)$
- $p(b|y)$
- $p(b, c|y)$
- $p(a, b|y)$
- $P[c < 0|y]$ ($c < 0$ indicates intervention works)
- $p(e^c|y)$ (e^c gives the rate of change in the expected number of falls)

$$\lambda = e^{cz} e^{a+bx}$$

- λ when $x = 40$ under no intervention and intervention

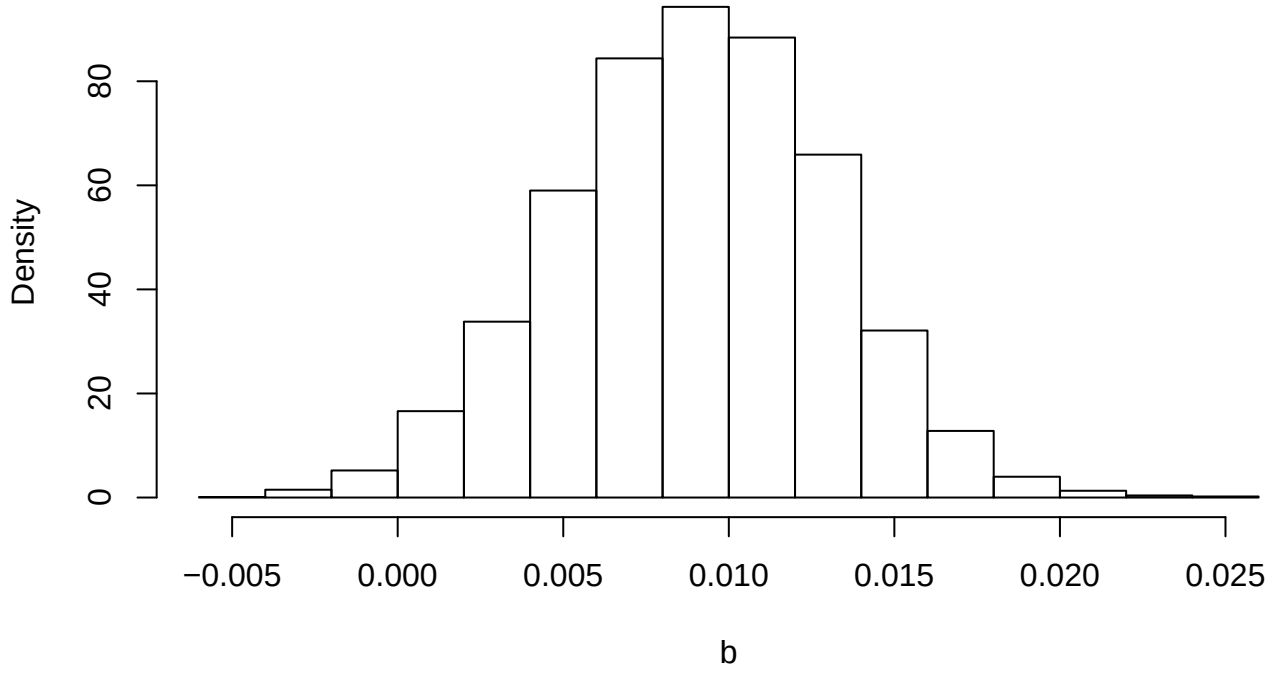
Answers:

- $p(c|y)$



$$E[c|y] = -1.074; \qquad \mathrm{SD}(c|y) = 0.131$$

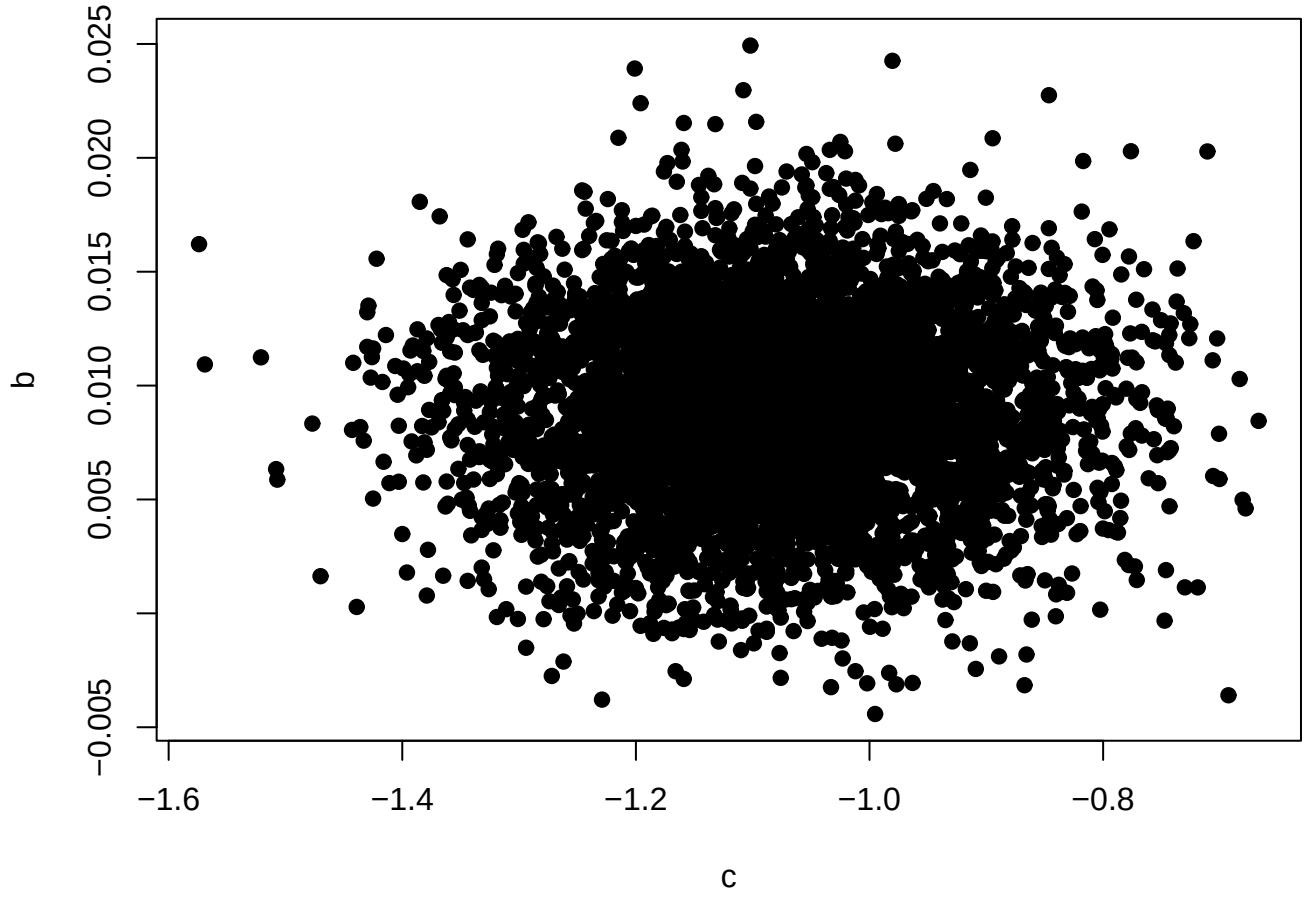
- $p(b|y)$



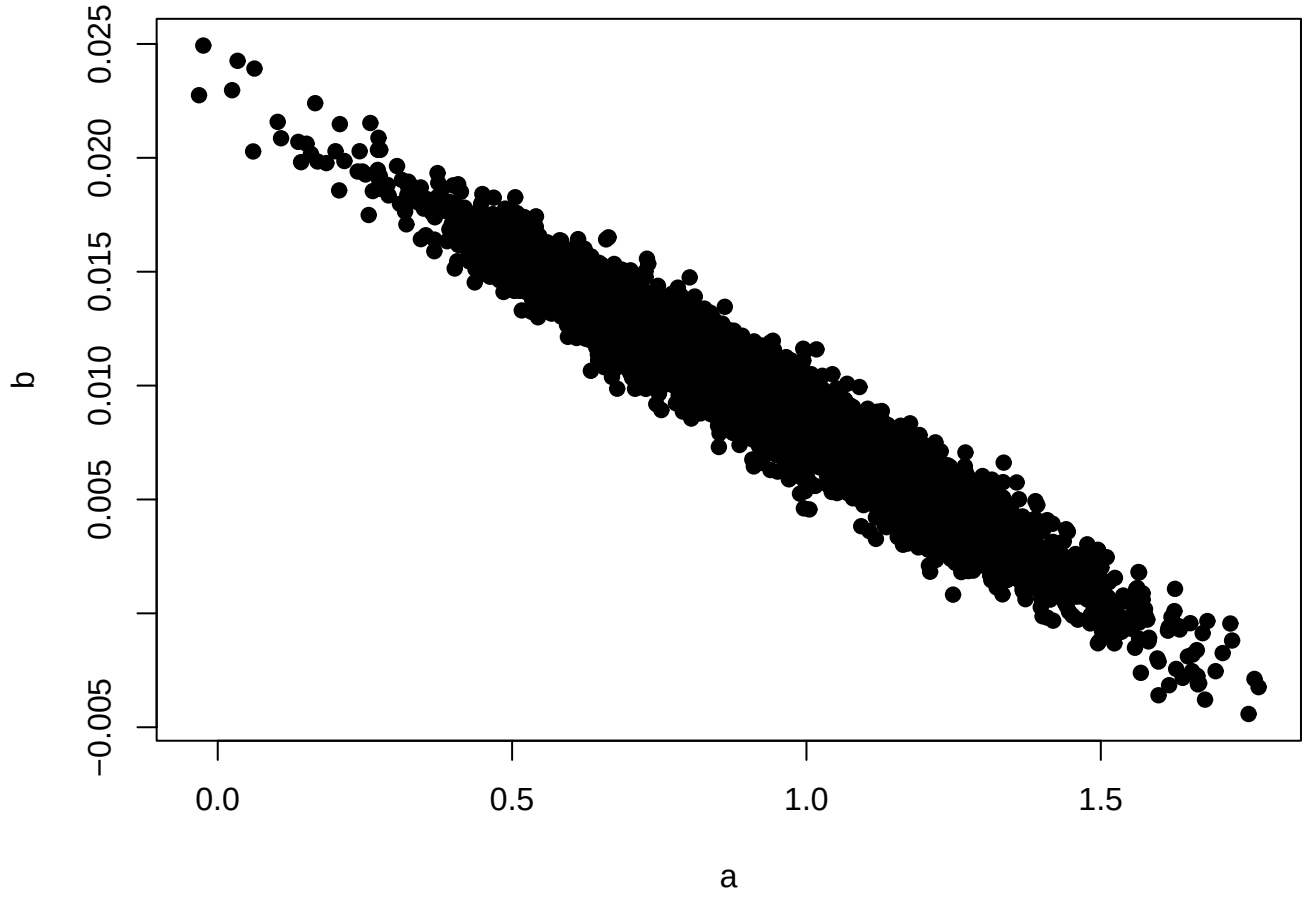
$$E[b|y] = 0.00898; \qquad \mathrm{SD}(b|y) = 0.00406$$

Note that this result is a bit surprising, since an increased strength index is expected to lead to fewer falls.

- $p(b, c|y)$



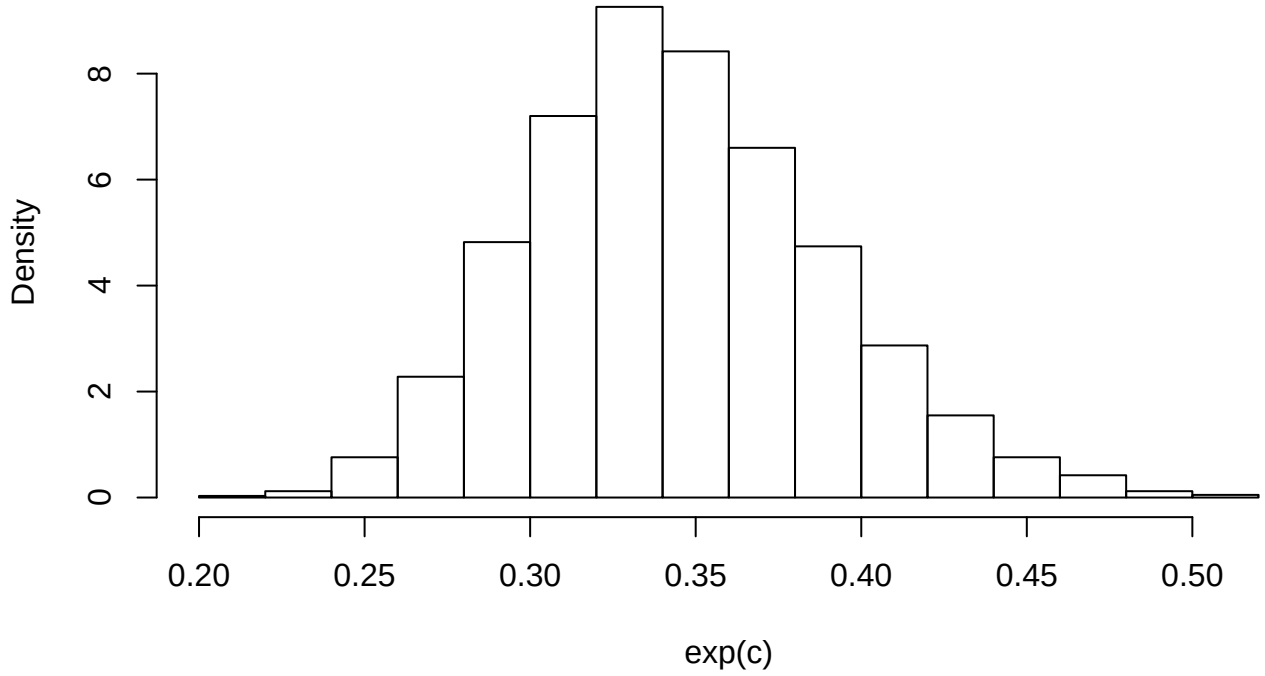
- $p(a, b|y)$



- $P[c < 0|y]$

$$P[c < 0|y] \approx \frac{1}{m} \sum_{i=1}^{m} I(c_i < 0) = 1$$

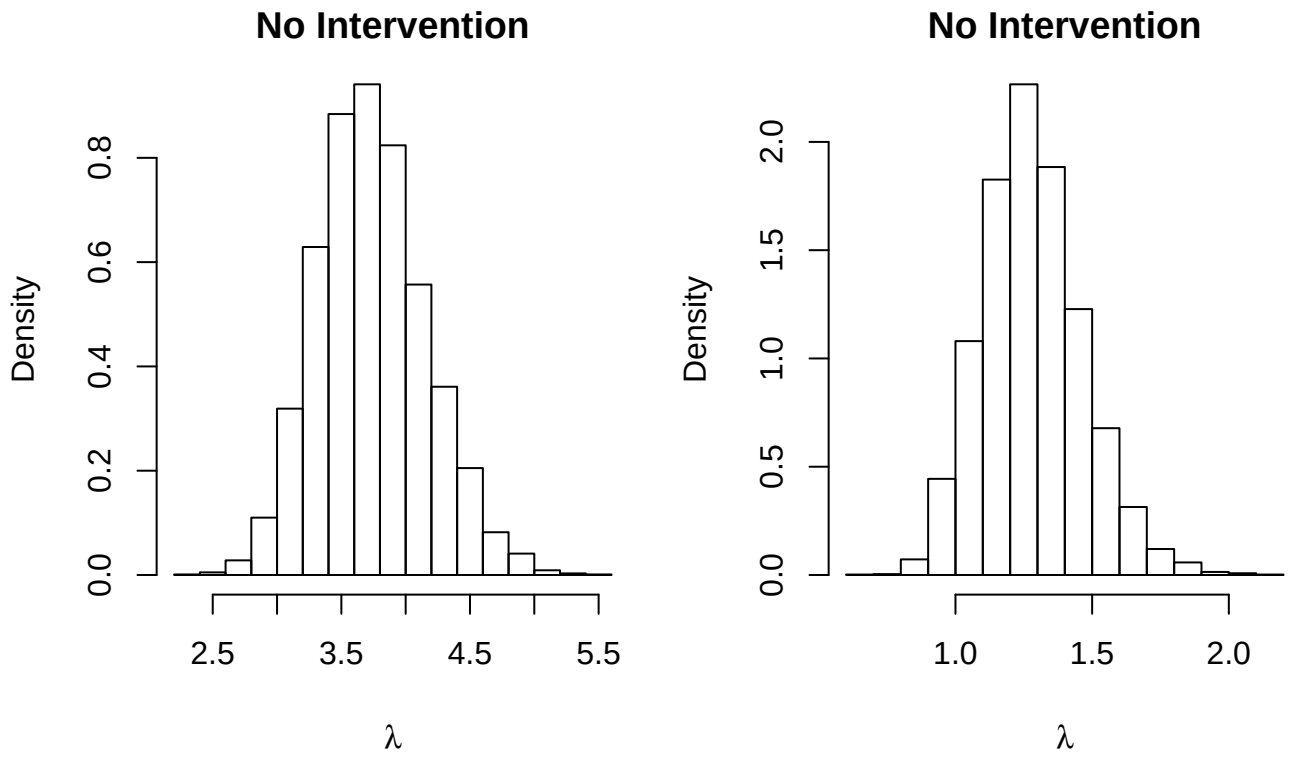
- $p(e^c|y)$



$$E[e^c|y] = 0.345; \qquad \mathrm{SD}(e^c|y) = 0.0451$$

This implies that the effect of the intervention should lead to people having less than half a many falls. The best guess is about a third as many.

- λ when $x = 40$ under no intervention and intervention



Intervention	$E[\lambda\|y]$	$\mathrm{SD}(\lambda\|y)$
No	3.735	0.421
Yes	1.281	0.184