Bayesian Econometrics

Lecture 2: Bayesian Estimation of Linear Regression Models

Prof. Dr. Kai Carstensen

Kiel University

Winter Term 2024/25

Outline of this lecture

In this lecture we study estimation and inference for a normally distributed random sample. We proceed step by step, increasing generality and complexity.

1. Estimating the mean of a normal distribution
2. Estimating mean and variance of a normal distribution
3. Regression models with one explanatory variable
4. Regression models with many explanatory variables
5. Model comparison
6. Prediction

This lecture follows Chapters 2 and 3 of Koop (2003). Please read them carefully.

1. Estimating the mean of a normal distribution

Likelihood

Let $y = (y_1, \ldots, y_N)'$ be a random sample from a normal distribution with unknown mean μ and **known** precision $h = \sigma^{-2}$ (inverse of the variance). We want to estimate μ.

The distribution is

$$p(y_i|\mu, h) = (2\pi)^{-\frac{1}{2}} h^{\frac{1}{2}} \exp\left[-\frac{1}{2}h(y_i - \mu)^2\right].$$

This gives rise to the likelihood

$$p(y|\mu, h) = \prod_{i=1}^{N} p(y_i|\mu, h) = (2\pi)^{-\frac{N}{2}} h^{\frac{N}{2}} \exp\left[-\frac{1}{2}h\sum_{i=1}^{N}(y_i - \mu)^2\right].$$

Prior

Suppose prior beliefs concerning μ are represented by a normal distribution with mean $\underline{\mu}$ and precision $\underline{\kappa}$. This yields prior moments

$$\mathrm{E}(\mu) = \underline{\mu} \qquad \text{and} \qquad \mathrm{Var}(\mu) = \underline{\kappa}^{-1}.$$

The prior pdf is

$$p(\mu|\underline{\mu}, \underline{\kappa}) = (2\pi)^{-\frac{1}{2}} \underline{\kappa}^{\frac{1}{2}} \exp\left[-\frac{1}{2}\underline{\kappa}(\mu - \underline{\mu})^2\right].$$

Not surprisingly, the prior has the same functional form as the likelihood. As we will see, the posterior is of the same form.

Hence, this is a natural conjugate prior. The interpretation is nice: $\underline{\mu}$ is the prior mean and $\underline{\kappa}$ represents the precision of this prior knowledge.

Posterior

The posterior is proportional to

$$p(\mu|y) \propto p(y|\mu, h)p(\mu|\underline{\mu}, \underline{\kappa})$$

Neglecting constants, we obtain

$$\begin{aligned} p(\mu|y) &\propto h^{\frac{N}{2}} \exp\left[-\frac{1}{2}h\sum_{i=1}^{N}(y_i-\mu)^2\right] \exp\left[-\frac{1}{2}\underline{\kappa}(\mu-\underline{\mu})^2\right] \\ &\propto h^{\frac{N}{2}} \exp\left[-\frac{1}{2}h\sum_{i=1}^{N}(y_i-\mu)^2 - \frac{1}{2}\underline{\kappa}(\mu-\underline{\mu})^2\right] \end{aligned}$$

This looks similar to a normal distribution of μ and in fact it is one. However, finding the parameters of this posterior is a bit tedious and left to the Tutorial.

It turns out that the posterior distribution of μ is normal with mean

$$\mathsf{E}(\mu|y) = \frac{hN\bar{y} + \underline{\kappa}\underline{\mu}}{hN + \underline{\kappa}}$$

and variance

$$\mathsf{Var}(\mu|y) = (hN + \underline{\kappa})^{-1} < \mathsf{Var}(\mu) = \underline{\kappa}^{-1}.$$

Hence, beliefs w.r.t. μ are updated: the expectation shifts towards the observed mean and the variance shrinks (beliefs become more precise as more information arrives).

To facilitate interpretation let us think of a fictitious "prior sample".

To this end, compare the kernels of the likelihood and the prior distribution:

$$\begin{aligned}
\exp\left[-\tfrac{1}{2}\underline{\kappa}(\mu-\underline{\mu})^2\right] &= \exp\left[-\tfrac{\underline{\kappa}}{2}\left(\underline{\mu}^2 - 2\underline{\mu}\mu + \mu^2\right)\right] \\
\exp\left[-\tfrac{1}{2}h\textstyle\sum_{i=1}^{N}(y_i-\mu)^2\right] &= \exp\left[-\tfrac{h}{2}\left(N\overline{y^2} - 2\mu N\bar{y} + N\mu^2\right)\right]
\end{aligned}$$

Now define $\underline{\kappa} = \underline{h}P$, where P is the fictitious "prior sample" size.

$$\begin{aligned}
\exp\left[-\tfrac{1}{2}\underline{\kappa}(\mu-\underline{\mu})^2\right] &= \exp\left[-\tfrac{\underline{h}}{2}\left(P\underline{\mu}^2 - 2P\underline{\mu}\mu + P\mu^2\right)\right] \\
\exp\left[-\tfrac{1}{2}h\textstyle\sum_{i=1}^{N}(y_i-\mu)^2\right] &= \exp\left[-\tfrac{h}{2}\left(N\overline{y^2} - 2N\bar{y}\mu + N\mu^2\right)\right]
\end{aligned}$$

This loosely suggests that P is the "prior sample" size, $\underline{\mu}$ is the "prior sample" mean and $\underline{h}$ is the "prior sample" precision.

Then the posterior expectation (Bayesian point estimator) can be written as

$$\mathrm{E}(\mu|y) = \frac{hN\bar{y} + \underline{h}P\underline{\mu}}{hN + \underline{h}P}.$$

Hence, the estimator is a weighted average of the sample mean $\bar{y}$ and the fictitious "prior sample" mean $\underline{\mu}$, where the weights depend on the respective sample sizes and precisions.

If we additionally assume that the precisions in both samples are identical, $h = \underline{h}$, we obtain

$$\mathrm{E}(\mu|y) = \frac{N\bar{y} + P\underline{\mu}}{N + P}.$$

Then the estimator is is identical to the mean of the combined samples.

2. Estimating mean and variance of a normal distribution

Likelihood

Let $y = (y_1, \ldots, y_N)'$ be a random sample from a normal distribution with unknown mean μ and precision h. We want to estimate both μ and h.

The likelihood again is

$$p(y|\mu, h) = \prod_{i=1}^{N} p(y_i|\mu, h) = (2\pi)^{-\frac{N}{2}} h^{\frac{N}{2}} \exp\left[-\frac{1}{2} h \sum_{i=1}^{N} (y_i - \mu)^2\right].$$

Prior

Suppose prior beliefs concerning μ are, conditional on h, represented by a normal distribution with mean $\underline{\mu}$ and precision $h\underline{\kappa}$, , $\mu|h \sim \mathcal{N}(\underline{\mu}, (h\underline{\kappa})^{-1})$, with density

$$p(\mu|\underline{\mu}, h\underline{\kappa}) = (2\pi)^{-\frac{1}{2}}(h\underline{\kappa})^{\frac{1}{2}} \exp\left[-\frac{1}{2}h\underline{\kappa}(\mu - \underline{\mu})^2\right],$$

and prior beliefs concerning h are represented by a gamma distribution with parameters $\underline{s}^{-2}$ and $\underline{\nu}$, $h \sim \text{Gamma}(\underline{s}^{-2}, \underline{\nu})$, with density

$$p(h|\underline{s}^{-2}, \underline{\nu}) = \left(\frac{\underline{s}^2\underline{\nu}}{2}\right)^{\frac{\underline{\nu}}{2}} \Gamma\left(\frac{\underline{\nu}}{2}\right)^{-1} h^{\frac{\underline{\nu}-2}{2}} \exp\left[-\frac{1}{2}h\underline{\nu}\underline{s}^2\right].$$

Here we use the Gamma function $\Gamma(\cdot)$.

The joint prior distribution of $\theta = (\mu, h)'$ is called normal-gamma,

$$NG(\underline{\mu}, \underline{\kappa}^{-1}, \underline{s}^{-2}, \underline{\nu}).$$

It is obtained as

$$\begin{aligned} p(\theta) &= p(\mu|h)p(h) \\ &= (2\pi)^{-\frac{1}{2}}(h\underline{\kappa})^{\frac{1}{2}} \exp\left[-\frac{1}{2}h\underline{\kappa}(\mu-\underline{\mu})^2\right] \times \left(\frac{\underline{s}^2\underline{\nu}}{2}\right)^{\frac{\underline{\nu}}{2}} \Gamma\left(\frac{\underline{\nu}}{2}\right)^{-1} h^{\frac{\underline{\nu}-2}{2}} \exp\left[-\frac{1}{2}h\underline{\nu}\underline{s}^2\right] \\ &= \underbrace{(2\pi)^{-\frac{1}{2}}\left(\frac{\underline{s}^2\underline{\nu}}{2}\right)^{\frac{\underline{\nu}}{2}} \Gamma\left(\frac{\underline{\nu}}{2}\right)^{-1} \underline{\kappa}^{\frac{1}{2}}}_{\text{integrating constant}} \underbrace{h^{\frac{\underline{\nu}-1}{2}} \exp\left[-\frac{h}{2}\left\{\underline{\kappa}(\mu-\underline{\mu})^2 + \underline{\nu}\underline{s}^2\right\}\right]}_{\text{kernel of the normal-gamma density}} \end{aligned}$$

and thus proportional to

$$p(\theta) \propto h^{\frac{\underline{\nu}-1}{2}} \exp\left[-\frac{h}{2}\left\{\underline{\kappa}(\mu-\underline{\mu})^2 + \underline{\nu}\underline{s}^2\right\}\right].$$

Prior moments

It is instructive to consider the moments implied by this prior because it is easier to think of prior moments than about prior parameters.

The marginal prior distribution of h is, by the definition of the normal-gamma distribution, just a gamma distribution with parameters $\underline{s}^{-2}$ and $\underline{\nu}$. From the properties of the gamma, we thus have

$$\mathrm{E}(h) = \underline{s}^{-2}$$

and variance

$$\mathrm{Var}(h) = \frac{2}{\underline{s}^4 \underline{\nu}} = \frac{2\,\mathrm{E}(h)^2}{\underline{\nu}}.$$

Hence, once you can come up with the first two moments of your prior beliefs regarding h (and you are willing to assume it is represented by a normal-gamma distribution), you effectively fix the parameters $\underline{s}$ and $\underline{\nu}$.

The marginal prior distribution of μ is more difficult to find (we do not show it in this lecture). It turns out that it is the generalized t distribution $\mu \sim t(\underline{\mu}, \underline{s}^2/\underline{\kappa}, \underline{\nu})$. It has mean

$$\mathsf{E}(\mu) = \underline{\mu} \qquad \text{if} \quad \underline{\nu} > 1$$

and variance

$$\mathsf{Var}(\mu) = \frac{\underline{\nu}}{\underline{\nu} - 2} \frac{\underline{s}^2}{\underline{\kappa}} \qquad \text{if} \quad \underline{\nu} > 2.$$

Hence, once you can come up with the first two moments of your prior beliefs regarding h to fix $\underline{s}$ and $\underline{\nu}$ and regarding μ, you effectively fix the parameters $\underline{\mu}$ and $\underline{\kappa}$.

Posterior

Finding the posterior is even more tedious than in part I (see Tutorial). It turns out that the joint posterior distribution of μ and h is normal-gamma, $\theta \sim NG(\bar{\mu}, \bar{\kappa}^{-1}, \bar{s}^{-2}, \bar{\nu})$, with parameters

$$\bar{\mu} = \frac{\underline{\kappa}\underline{\mu} + N\bar{y}}{\underline{\kappa} + N}$$

$$\bar{\kappa} = \underline{\kappa} + N$$

$$\bar{s}^2 = \frac{\underline{\nu}\underline{s}^2 + \nu s^2 + \frac{1}{\frac{1}{\underline{\kappa}} + \frac{1}{N}} \left(\underline{\mu} - \bar{y}\right)^2}{\underline{\nu} + N}$$

$$\bar{\nu} = \underline{\nu} + N$$

using the sample statistics

$$\nu = N - 1 \qquad \text{and} \qquad s^2 = \frac{1}{\nu} \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

The marginal posterior distribution of h is, by the definition of the normal-gamma distribution, just a gamma distribution with parameters $\bar{s}^{-2}$ and $\bar{\nu}$. From the properties of the gamma, we thus have

$$\mathrm{E}(h|y) = \bar{s}^{-2}$$

and variance

$$\mathrm{Var}(h|y) = \frac{2}{\bar{s}^4 \bar{\nu}}.$$

In addition, interval estimators can be constructed from the quantiles of the gamma distribution.

The marginal posterior distribution of μ is again a generalized t distribution

$$\mu|y \sim t(\bar{\mu}, \bar{s}^2/\bar{\kappa}, \bar{\nu})$$

with pdf

$$f(\mu|\bar{\mu}, \bar{s}^2/\bar{\kappa}, \bar{\nu}) = \frac{\Gamma\left(\frac{\bar{\nu}+1}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)} \left(\frac{\bar{\kappa}}{\bar{s}^2\bar{\nu}}\right)^{\frac{1}{2}} \left[1 + \frac{\bar{\kappa}}{\bar{s}^2\bar{\nu}}(\mu - \bar{\mu})^2\right]^{-\frac{\bar{\nu}+1}{2}}.$$

It has mean

$$\mathrm{E}(\mu|y) = \bar{\mu} \qquad \text{if} \quad \bar{\nu} > 1$$

and variance

$$\mathrm{Var}(\mu|y) = \frac{\bar{\nu}}{\bar{\nu} - 2}\frac{\bar{s}^2}{\bar{\kappa}} \qquad \text{if} \quad \bar{\nu} > 2.$$

Interval estimators can be constructed from the quantiles of this t distribution.

Improper prior

Let us think about a noninformative normal-gamma prior. What is noninformativeness here?

The conditional normal prior distribution

$$\mu|h \sim \mathcal{N}(\underline{\mu}, (h\underline{\kappa})^{-1})$$

becomes, given h, flatter (i.e., more dispersed) as prior precision $\underline{\kappa}$ tends towards zero. But

$$p(\mu|\underline{\mu}, h\underline{\kappa}) = (2\pi)^{-\frac{1}{2}}(h\underline{\kappa})^{\frac{1}{2}} \exp\left[-\frac{1}{2}h\underline{\kappa}(\mu-\underline{\mu})^2\right] \to 0 \qquad \text{as} \qquad \underline{\kappa} \to 0.$$

The marginal gamma prior distribution

$$h \sim \text{Gamma}(\underline{s}^{-2}, \underline{\nu})$$

becomes, given a prior mean $\underline{s}^{-2}$, flatter (i.e., more dispersed) as its variance

$$\text{Var}(h) = \frac{2}{\underline{s}^4 \underline{\nu}}$$

increases. This is achieved if $\underline{\nu}$ tends towards zero. But again

$$p(h|\underline{s}^{-2}, \underline{\nu}) = \left(\frac{\underline{s}^2 \underline{\nu}}{2}\right)^{\frac{\underline{\nu}}{2}} \Gamma\left(\frac{\underline{\nu}}{2}\right)^{-1} h^{\frac{\underline{\nu}-2}{2}} \exp\left[-\frac{1}{2} h \underline{\nu} \underline{s}^2\right] \to 0 \qquad \text{as} \qquad \underline{\nu} \to 0$$

because then $\Gamma\left(\frac{\underline{\nu}}{2}\right) \to \infty$.

Altogether for the conjugate normal-gamma prior we have:

$$p(\theta) = p(\mu|h)p(h) \to 0 \qquad \text{as} \qquad \underline{\kappa} \to 0 \ \text{ and } \ \underline{\nu} \to 0.$$

This is a degenerate distribution difficult to work with: if you multiply this prior with the likelihood, you obtain a value of zero.

In a slight abuse of notation, one may go around this problem by neglecting the integration constants. Just consider

$$p(\theta) \propto h^{\frac{\underline{\nu}-1}{2}} \exp\left[-\frac{h}{2}\left\{\underline{\kappa}(\mu-\underline{\mu})^2+\underline{\nu s}^2\right\}\right].$$

Now letting $\underline{\kappa}$ and $\underline{\nu}$ tend towards zero yields

$$p(\theta) \propto h^{-\frac{1}{2}}, \qquad -\infty < \mu < \infty, 0 < h < \infty.$$

This is called an improper prior because it does not integrate to 1 no matter which (finite) integration constant you choose.

Nevertheless, it works when you use it in the way we studied above.

In fact, based on $\underline{\kappa} = 0$ and $\underline{\nu} = 0$ and thus

$$p(\theta) \propto h^{-\frac{1}{2}}$$

you obtain the following posterior normal-gamma parameters, which only depend on the data:

$$\lim_{\underline{\kappa},\underline{\nu}\to 0} \bar{\mu} = \frac{\underline{\kappa}\underline{\mu} + N\bar{y}}{\underline{\kappa} + N} = \bar{y}$$

$$\lim_{\underline{\kappa},\underline{\nu}\to 0} \bar{\kappa} = \underline{\kappa} + N = N$$

$$\lim_{\underline{\kappa},\underline{\nu}\to 0} \bar{s}^2 = \frac{\underline{\nu}\underline{s}^2 + \nu s^2 + \frac{\underline{\kappa} N}{\underline{\kappa}+N}\left(\underline{\mu} - \bar{y}\right)^2}{\underline{\nu} + N} = \frac{\nu}{N} s^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \hat{\sigma}^2_{ML}$$

$$\lim_{\underline{\kappa},\underline{\nu}\to 0} \bar{\nu} = \underline{\nu} + N = N$$

3. Regression models with one explanatory variable

Assumptions

We are interested in the scalar parameter β of the relationship

$$y_i = \beta x_i + \varepsilon_i$$

Assumptions:

- Random sample: x_i, y_i, $i = 1, \ldots, N$ are a random sample.
- Normality: ε_i is iid normal with mean zero and precision $h = \sigma^{-2}$.
- Regressor exogeneity: The x_i are independent of ε_i with pdf $p(x_i|\lambda)$, where λ is a vector of parameters that does not include, or depend on, β and h. This allows to condition on x_i without loss of information regarding β.

Conditional likelihood

Conditioning on x_i yields the moments

$$\mathrm{E}(y_i|\beta, h, x_i) = \beta x_i \qquad \text{and} \qquad \mathrm{Var}(y_i|\beta, h, x_i) = \sigma^2 = h^{-1}.$$

Note that for the simplicity of notation, we will henceforth not include x_i in the conditioning set.

Due to normality of ε_i, the conditional distribution of y_i is also normal

$$p(y_i|\beta, h) = (2\pi)^{-\frac{1}{2}} h^{\frac{1}{2}} \exp\left[-\frac{h}{2}(y_i - \beta x_i)^2\right].$$

Due to random sampling, the likelihood is (leaving out constants)

$$p(y|\beta, h) = \prod_{i=1}^{N} p(y_i|\beta, h) \propto h^{\frac{N}{2}} \exp\left[-\frac{h}{2}\sum_{i=1}^{N}(y_i - \beta x_i)^2\right].$$

For further use let us define the "frequentist" statistics (which are purely data-dependent)

Degrees of freedom: $\nu = N - 1$

OLS estimator: $\hat{\beta} = \left(\sum_{i=1}^{N} x_i^2\right)^{-1} \sum_{i=1}^{N} x_i y_i$

Variance estimator: $s^2 = \frac{1}{\nu} \sum_{i=1}^{N} (y_i - \hat{\beta} x_i)^2$

OLS precision: $\kappa = \sum_{i=1}^{N} x_i^2$

Use this to rewrite (it is easiest to show this backwards)

$$\sum_{i=1}^{N} (y_i - \beta x_i)^2 = \nu s^2 + (\beta - \hat{\beta})^2 \sum_{i=1}^{N} x_i^2 = \nu s^2 + (\beta - \hat{\beta})^2 \kappa$$

and substitute it into the likelihood.

Substitution yields

$$p(y|\beta,h) \propto h^{\frac{N}{2}} \exp\left[-\frac{h}{2}\left\{(\beta-\hat{\beta})^2\kappa+\nu s^2\right\}\right]$$

$$\propto h^{\frac{N}{2}} \exp\left[-\frac{h}{2}\kappa(\beta-\hat{\beta})^2\right] \exp\left[-\frac{\nu s^2}{2}h\right]$$

Now "allocate" h to the two exponential functions

$$p(y|\beta,h) \propto \underbrace{\left\{h^{\frac{1}{2}} \exp\left[-\frac{h\kappa}{2}(\beta-\hat{\beta})^2\right]\right\}}_{\text{looks similar to a normal density}} \underbrace{\left\{h^{\frac{\nu}{2}} \exp\left[-\frac{\nu s^2}{2}h\right]\right\}}_{\text{looks similar to a Gamma density}}.$$

This looks like a normal-gamma distribution and suggests that the normal-gamma is a natural conjugate prior (in fact, it is).

Normal-Gamma prior

Conceptually, the normal regression model is very similar to a previous case in which we estimated mean and variance (or precision) of a normal distribution using a random sample (in fact, that was a special case with $x_i = 1$).

We know already that the normal-gamma distribution is a natural conjugate prior in that case and it turns out to be the same here. Hence, let us assume

$$\beta | h \sim \mathcal{N}(\underline{\beta}, h^{-1}\underline{V})$$

and

$$h \sim G(\underline{s}^{-2}, \underline{\nu}).$$

Then the joint prior is

$$\beta, h \sim NG(\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}).$$

Normal-Gamma posterior

Finding the posterior proceeds along the same lines as for the previous case of a normal distribution with unknown mean and precision. It turns out that the joint posterior distribution of β and h is normal-gamma, $\beta, h \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, \bar{\nu})$, with parameters

$$\bar{\beta} = \frac{\underline{V}^{-1}\underline{\beta} + \kappa\hat{\beta}}{\underline{V}^{-1} + \kappa}$$

$$\bar{V} = \left(\underline{V}^{-1} + \kappa\right)^{-1}$$

$$\bar{s}^2 = \frac{\underline{\nu}\underline{s}^2 + \nu s^2 + \frac{1}{\underline{V}+\kappa^{-1}}\left(\hat{\beta} - \underline{\beta}\right)^2}{\underline{\nu} + N}$$

$$\bar{\nu} = \underline{\nu} + N$$

The expressions are easier to interpret if we define precisions $\underline{\kappa} = \underline{V}^{-1}$ and $\bar{\kappa} = \bar{V}^{-1}$, and replace accordingly:

$$\bar{\beta} = \frac{\underline{\kappa}\underline{\beta} + \kappa\hat{\beta}}{\underline{\kappa} + \kappa}$$

$$\bar{\kappa} = \underline{\kappa} + \kappa$$

$$\bar{s}^2 = \frac{\underline{\nu}\underline{s}^2 + \nu s^2 + \frac{1}{\underline{\kappa}^{-1}+\kappa^{-1}}\left(\hat{\beta} - \underline{\beta}\right)^2}{\underline{\nu} + N}$$

$$\bar{\nu} = \underline{\nu} + N$$

Estimator of β

We already know that the marginal distribution of the first normal-gamma variable is a generalized t distribution. Specifically, from

$$\beta, h \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, \bar{\nu})$$

follows

$$\beta \sim t(\bar{\beta}, \bar{s}^2 \bar{V}, \bar{\nu}).$$

From the properties of the t distribution we obtain

$$\mathrm{E}(\beta|y) = \bar{\beta} = \frac{\underline{\kappa}\underline{\beta} + \kappa\hat{\beta}}{\underline{\kappa} + \kappa} \qquad \text{and} \qquad \mathrm{Var}(\beta|y) = \frac{\bar{\nu}}{\bar{\nu} - 2}\bar{s}^2 \bar{V} = \frac{\bar{\nu}}{\bar{\nu} - 2}\frac{\bar{s}^2}{\bar{\kappa}}.$$

Interval estimators can be constructed from the quantiles of the $t(\bar{\beta}, \bar{s}^2 \bar{V}, \bar{\nu})$ distribution.

4. Regression models with many explanatory variable

Assumptions

We are interested in the k-dimensional parameter vector β of the relationship

$$y_i = x_{1i}\beta_1 + \cdots + x_{ki}\beta_k + \varepsilon_i = x_i\beta + \varepsilon_i$$

Assumptions:

- Dimension: x_i is a $1 \times k$ row vector.
- Intercept: $x_{1i} = 1$, hence β_1 is the intercept.
- Random sample: x_i, y_i, $i = 1, \ldots, N$ are a random sample.
- Normality: ε_i is iid normal with mean zero and precision h.
- Regressor exogeneity: The x_i are independent of ε_i.

Stacking all observations into matrices yields

$$y = X\beta + \varepsilon,$$

where y and ε are $N \times 1$, X is $N \times k$, and β is $k \times 1$.

Conditional likelihood

Conditioning on x_i yields the moments

$$\mathsf{E}(y_i|\beta, h, x_i) = x_i\beta \qquad \text{and} \qquad \mathsf{Var}(y_i|\beta, h, x_i) = h^{-2}.$$

Note that for the simplicity of notation, we will henceforth not include x_i in the conditioning set.

Due to normality of ε_i and random sampling, the likelihood is

$$p(y|\beta, h) = \prod_{i=1}^{N} p(y_i|\beta, h) \propto h^{\frac{N}{2}} \exp\left[-\frac{h}{2}\sum_{i=1}^{N}(y_i - x_i\beta)^2\right].$$

It is more convenient to use matrix notation instead of sums. To this end, recall that

$$\sum_{i=1}^{N}(y_i - x_i\beta)^2 = (y - X\beta)'(y - X\beta).$$

This yields the likelihood function

$$p(y|\beta, h) \propto h^{\frac{N}{2}} \exp\left[-\frac{h}{2}(y - X\beta)'(y - X\beta)\right].$$

For further use let us define the "frequentist" statistics (which are purely data-dependent)

Degrees of freedom: $\nu = N - k$

OLS estimator: $\hat{\beta} = (X'X)^{-1} X'y$

Variance estimator: $s^2 = \frac{1}{\nu}(y - X\hat{\beta})'(y - X\hat{\beta}) = \frac{1}{\nu}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{\nu}\varepsilon' M_X \varepsilon$

OLS precision: $\boldsymbol{\kappa} = X'X$

From OLS algebra, recall $M_X = I_N - X(X'X)^{-1}X'$, $M_X = M_X M_X$, and $\hat{\varepsilon} = M_X \varepsilon$.

Use this to rewrite (it is easiest to show this backwards)

$$(y - X\beta)'(y - X\beta) = \nu s^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) = \nu s^2 + (\beta - \hat{\beta})'\boldsymbol{\kappa}(\beta - \hat{\beta}).$$

Now substitute this into the likelihood

$$p(y|\beta,h) \propto h^{\frac{N}{2}} \exp\left[-\frac{h}{2}(\beta-\hat{\beta})'\boldsymbol{\kappa}(\beta-\hat{\beta})\right] \exp\left[-\frac{h}{2}\nu s^2\right].$$

and rearrange the parameter h to the two exponential functions

$$p(y|\beta,h) \propto \underbrace{\left\{h^{\frac{k}{2}} \exp\left[-\frac{1}{2}(\beta-\hat{\beta})'h\boldsymbol{\kappa}(\beta-\hat{\beta})\right]\right\}}_{\text{looks like a multivariate normal density with mean vector } \hat{\beta} \text{ and precision matrix } h\boldsymbol{\kappa} \text{ or, equivalently, variance matrix } \sigma^2(X'X)^{-1}} \underbrace{\left\{h^{\frac{\nu}{2}} \exp\left[-\frac{\nu s^2}{2}h\right]\right\}}_{\text{looks like a Gamma density}}.$$

This once again looks like a normal-gamma distribution and suggests that the normal-gamma is a natural conjugate prior (in fact, it is).

Normal-Gamma prior

For the case of many regressors, the normal-gamma distribution is again a natural conjugate prior. This time, we have to use the multivariate normal distribution

$$\beta|h \sim \mathcal{N}(\underline{\beta}, h^{-1}\underline{V}),$$

where $\underline{V}$ is a positive definite $k \times k$ matrix with inverse $\underline{\boldsymbol{\kappa}}$, and

$$h \sim G(\underline{s}^{-2}, \underline{\nu}).$$

Then the joint prior is the multivariate normal-gamma distribution

$$\beta, h \sim NG(\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}).$$

Normal-Gamma posterior

Finding the posterior is quite tedious and left to the tutorial. It turns out that the joint posterior distribution of β and h is normal-gamma, $\beta, h \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, \bar{\nu})$, with parameters

$$\bar{\beta} = (\underline{V}^{-1} + \boldsymbol{\kappa})^{-1}(\underline{V}^{-1}\underline{\beta} + \boldsymbol{\kappa}\hat{\beta})$$

$$\bar{V} = (\underline{V}^{-1} + \boldsymbol{\kappa})^{-1}$$

$$\bar{s}^2 = \frac{\underline{\nu s}^2 + \nu s^2 + \left(\hat{\beta} - \underline{\beta}\right)' \left[\underline{V} + \boldsymbol{\kappa}^{-1}\right]^{-1} \left(\hat{\beta} - \underline{\beta}\right)}{\underline{\nu} + N}$$

$$\bar{\nu} = \underline{\nu} + N$$

The parameters are again easier to interpret if we define the posterior precision $\bar{\boldsymbol{\kappa}} = \bar{V}^{-1}$ and replace accordingly:

$$\bar{\beta} = (\underline{\boldsymbol{\kappa}} + \boldsymbol{\kappa})^{-1}(\underline{\boldsymbol{\kappa}}\underline{\beta} + \boldsymbol{\kappa}\hat{\beta})$$

$$\bar{\boldsymbol{\kappa}} = \underline{\boldsymbol{\kappa}} + \boldsymbol{\kappa}$$

$$\bar{s}^2 = \frac{\underline{\nu s}^2 + \nu s^2 + \left(\hat{\beta} - \underline{\beta}\right)' \left[\underline{\boldsymbol{\kappa}}^{-1} + \boldsymbol{\kappa}^{-1}\right]^{-1} \left(\hat{\beta} - \underline{\beta}\right)}{\underline{\nu} + N}$$

$$\bar{\nu} = \underline{\nu} + N$$

Marginal posterior for β

The marginal posterior distribution of β is the general multivariate t distribution

$$\beta|y \sim t(\bar{\beta}, \bar{s}^2\bar{V}, \bar{\nu})$$

with pdf

$$f(\beta|\bar{\beta}, \bar{s}^2\bar{V}, \bar{\nu}) = \frac{\Gamma\left(\frac{\bar{\nu}+k}{2}\right)}{(\bar{\nu}\pi)^{\frac{k}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)}|\bar{s}^2\bar{V}|^{-\frac{1}{2}}\left[1+(\beta-\bar{\beta})'(\bar{s}^2\bar{V})^{-1}(\beta-\bar{\beta})/\bar{\nu}\right]^{-\frac{\bar{\nu}+k}{2}}.$$

It can be used to find posterior expectation (if $\bar{\nu} > 1$)

$$\mathrm{E}(\beta|y) = \bar{\beta},$$

variance matrix (if $\bar{\nu} > 2$)

$$\mathrm{Var}(\beta|y) = \frac{\bar{\nu}}{\bar{\nu}-2}\bar{s}^2\bar{V},$$

and posterior intervals.

How to compute quantiles from the multivariate t distribution

Consider the k-dimensional random vector $u \sim t(m, S, \nu)$.

Then element i has marginal (univariate) distribution $u_i \sim t(m_i, S_{ii}, \nu)$.

The standardized element $\tilde{u}_i = (u_i - m_i)/\sqrt{S_{ii}} \sim t(0, 1, \nu)$. This is the "textbook" t distribution, quantiles of which are tabulated in many books or available in software packages such as Matlab.

Hence, a symmetric $100(1-\alpha)\%$ interval for $\tilde{u}_i$ is constructed as

$$p(t_{\alpha/2} \leq \tilde{u}_i \leq t_{1-\alpha/2}) = p(m_i + t_{\alpha/2}\sqrt{S_{ii}} \leq u_i \leq m_i + t_{1-\alpha/2}\sqrt{S_{ii}}) = 1 - \alpha.$$

Using the symmetry of the t distribution yields

$$p(m_i - t_{1-\alpha/2}\sqrt{S_{ii}} \leq u_i \leq m_i + t_{1-\alpha/2}\sqrt{S_{ii}}) = 1 - \alpha.$$

5. Model comparison

Two regression models

We have two regression models M_1 and M_2 which differ in their explanatory variable:

$$M_j: \qquad y_i = \beta_j x_{ji} + \varepsilon_{ji}, \quad \varepsilon_{ji} \sim \mathcal{N}(0, h_j^{-1}), \quad j = 1, 2.$$

In addition, we are given two normal-gamma priors

$$\beta_j, h_j | M_j \sim NG(\underline{\beta}_j, \underline{V}_j, \underline{s}_j^{-2}, \underline{\nu}_j), \quad j = 1, 2.$$

For simplicity, x_{1i} and x_{2i} are scalars. As before, they are exogenous.

Which model should we favor a posteriori?

The models and priors imply the two posterior distributions

$$\beta_j, h_j | y, M_j \sim NG(\bar{\beta}_j, \bar{V}_j, \bar{s}_j^{-2}, \bar{\nu}_j), \quad j = 1, 2.$$

We are interested in the posterior odds ratio

$$PO_{12} = \frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)p(M_1)}{p(y|M_2)p(M_2)}.$$

Hence, we need to calculate the marginal likelihood

$$p(y|M_j) = \int \int p(y|\beta_j, h_j)\, p(\beta_j, h_j)\, d\beta_j\, dh_j.$$

While we could do this numerically for any set of parameters, an analytic solution is feasible that leads to some insights.

Marginal likelihood and posterior odds ratio

It is straightforward (at least for those who love to integrate) but tedious to show that

$$p(y|M_j) = c_j \left(\frac{\bar{V}_j}{\underline{V}_j}\right)^{\frac{1}{2}} (\bar{\nu}_j \bar{s}_j^2)^{-\frac{\bar{\nu}_j}{2}}$$

where c_j is a complicated function of the prior and posterior parameters (see textbook, p. 25). This yields the posterior odds ratio

$$PO_{12} = \frac{c_1 \left(\frac{\bar{V}_1}{\underline{V}_1}\right)^{\frac{1}{2}} (\bar{\nu}_1 \bar{s}_1^2)^{-\frac{\bar{\nu}_1}{2}} p(M_1)}{c_2 \left(\frac{\bar{V}_2}{\underline{V}_2}\right)^{\frac{1}{2}} (\bar{\nu}_2 \bar{s}_2^2)^{-\frac{\bar{\nu}_2}{2}} p(M_2)}.$$

What does this tell us?

$$PO_{12} = \frac{c_1 \left(\frac{\bar{V}_1}{\underline{V}_1}\right)^{\frac{1}{2}} (\bar{\nu}_1 \bar{s}_1^2)^{-\frac{\bar{\nu}_1}{2}} p(M_1)}{c_2 \left(\frac{\bar{V}_2}{\underline{V}_2}\right)^{\frac{1}{2}} (\bar{\nu}_2 \bar{s}_2^2)^{-\frac{\bar{\nu}_2}{2}} p(M_2)}.$$

Support for M_1 is high if

- the prior odds ratio $p(M_1)/p(M_2)$ is high,
- the fit of M_1 exceeds that of M_2: $\bar{\nu}_1 \bar{s}_1^2$ contains $\nu_1 s_1^2$, the sum of squared errors,
- coherency between prior and data of M_1 exceeds that of M_2: $\bar{\nu}_1 \bar{s}_1^2$ contains $(\hat{\beta}_j - \underline{\beta}_j)^2$,
- the prior information relative to posterior information of M_1 is higher: this is represented by the relative variance $\frac{\bar{V}_1}{\underline{V}_1}$, or equivalently, the inverse of the relative precision.

Regression models with multiple regressors

If we allow for multiple regressors k_1 and k_2, there is an additional reward for parsimony (or penalty for the number of parameters).

This reward can completely dominate the posterior odds ratio when noninformative priors for β are used. Therefore, comparison of models with different equality restrictions on the β's should be based on informative priors.

For details, see textbook p. 41-43.

Inequality restrictions

Often, we are interested in whether an effect is positive (or negative). Then we may compare

$$M_1 : \beta_i > 0$$

with

$$M_2 : \beta_i \text{ unrestricted.}$$

The posterior odds ration simply is

$$PO_{12} = \frac{p(M_1|y)}{p(M_2|y)} = \frac{p(\beta_i > 0|y)}{p(\beta_i \text{ unrestricted}|y)} = p(\beta_i > 0|y).$$

The probability can be computed from the marginal t distribution of β_i.

6. Prediction

Setup

Recall that we assumed (a) the regressors are exogenous and (b) the disturbances are iid.

This simple setup does not allow for typical time series forecasting (autoregressive models etc).

Nevertheless, we may ask what our model, based on a prior and N observations (y, X), predicts given T out-of-sample observations X^*.

Model:

$$y^* = X^*\beta + \varepsilon^*,$$

where $\varepsilon^* \sim \mathcal{N}(0, h^{-1}I_T)$ is **independent** of ε. Also note that y^* and ε^* are $T \times 1$ and X^* is $T \times k$.

Posterior predictive density

The predictive density is

$$p(y^*|y) = \int \int p(y^*|y, \beta, h)\, p(\beta, h|y)\, d\beta\, dh,$$

where integration is with respect to h and to all elements of β.

It can be shown that for the normal-gamma prior and posterior, this leads to a t distribution:

$$y^*|y \sim t(X^*\bar{\beta}, \bar{s}^2[I_T + X^*\bar{V}X^{*\prime}], \bar{\nu}).$$

This allows to compute point and interval predictions.

Bayesian model averaging

Frequently, it is not obvious which model to use: more than one model fits the data well.

In such a situation, we may compute an "average" predictive density. For two models, we obtain

$$p(y^*|y) = p(y^*|y, M_1)p(M_1|y) + p(y^*|y, M_2)p(M_2|y)$$

with point predictor

$$\mathrm{E}(y^*|y) = \mathrm{E}(y^*|y, M_1)p(M_1|y) + \mathrm{E}(y^*|y, M_2)p(M_2|y).$$

7. Empirical Example

Economic question

What is the quantitative effect of house configuration on house prices?

Data: $N = 546$ houses sold in Windsor (Canada) in 1987 downloadable from the companion website of the textbook.

Literature: Anglin and Gencay (1996)

Variables:

- y: sales price in Canadian dollars
- x_1: intercept
- x_2: lot size in square feet
- x_3: number of bedrooms
- x_4: number of bathrooms
- x_5: number of storeys

Priors

Prior for h

Recall: $h = 1/\sigma^2$ (inverse of error variance)

House prices are between \$50,000 and \$150,000.

Hopefully, errors are between -\$10,000 and \$10,000.

This suggests that $[-\$10,000; \$10,000]$ is a $\pm 2\sigma$ interval $\Rightarrow \underline{\sigma} = 5000$.

Therefore, we choose

$$\underline{s}^{-2} = \mathrm{E}(h) = \frac{1}{\underline{\sigma}^2} = 4.0 \times 10^{-8}.$$

Since we are very unsure, we choose a small degree of freedom (fictitious prior sample size) of

$$\underline{\nu} = 5.$$

Priors

Prior for β

These are crude guesses (not really "prior knowledge" in the correct sense):

$$\underline{\beta} = \begin{pmatrix} 0 \\ 10 \\ 5000 \\ 10000 \\ 10000 \end{pmatrix}$$

This means:

- $\underline{\beta}_2 = 10$: dollar price of 1 additional square foot lot size
- $\underline{\beta}_3 = 5000$: dollar price of 1 additional bedroom
- $\underline{\beta}_4 = 10000$: dollar price of 1 additional bathroom
- $\underline{\beta}_5 = 10000$: dollar price of 1 additional storey

Priors

Prior for Var(β)

Think again in terms of approximate 95% intervals ($\pm 2\sigma$) based on the (not fully correct) normal distribution:

- $\underline{\beta}_1 \in (-20,000; +20,000)$: intercept
- $\underline{\beta}_2 \in (0; 20)$: dollar price of 1 additional square foot lot size
- $\underline{\beta}_3 \in (0; 10,000)$: dollar price of 1 additional bedroom
- $\underline{\beta}_4 \in (0; 20,000)$: dollar price of 1 additional bathroom
- $\underline{\beta}_5 \in (0; 20,000)$: dollar price of 1 additional storey

Assuming correlations are zero, this leads to the variance matrix

$$\text{Var}(\beta) = \begin{pmatrix} 10,000^2 & 0 & 0 & 0 & 0 \\ 0 & 5^2 & 0 & 0 & 0 \\ 0 & 0 & 2500 & 0 & 0 \\ 0 & 0 & 0 & 5000^2 & 0 \\ 0 & 0 & 0 & 0 & 5000^2 \end{pmatrix}$$

To obtain the prior scale matrix $\underline{V}$, recall that, for the normal-gamma distribution,

$$\text{Var}(\beta) = \frac{\underline{\nu}}{\underline{\nu}-2}\underline{s}^2\underline{V} \qquad \Rightarrow \qquad \underline{V} = \frac{\underline{\nu}-2}{\underline{\nu}}\underline{s}^{-2}\,\text{Var}(\beta)$$

Given our choices $\underline{\nu} = 5$ and $\underline{s}^2 = 5000^2$, we obtain

$$\underline{V} = \begin{pmatrix} 2.40 & 0 & 0 & 0 & 0 \\ 0 & 6.0 \times 10^{-7} & 0 & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 \\ 0 & 0 & 0 & 0.60 & 0 \\ 0 & 0 & 0 & 0 & 0.60 \end{pmatrix}$$

Results

Prior and posterior means for β

MATLAB Command Window Page 1

Prior and posterior means for beta (std.dev. in brackets)

Prior Informative		Posterior basend on noninformative prior		Posterior basend on informative prior	
0.00	(10000.00)	-4009.55	(3593.16)	-4035.05	(3530.16)
10.00	(5.00)	5.43	(0.37)	5.43	(0.37)
5000.00	(2500.00)	2824.61	(1211.45)	2886.81	(1184.93)
10000.00	(5000.00)	17105.17	(1729.65)	16965.24	(1708.02)
10000.00	(5000.00)	7634.90	(1005.19)	7641.23	(997.02)

Prior and posterior properties for h

MATLAB Command Window Page 1

Prior and posterior means for h (std.dev. in brackets

Prior	Posterior basend on	
Informative	noninformative prior	informative prior
4.00e-08 (2.53e-08) --	3.03e-09 (1.83e-10) --	3.05e-09 (1.84e-10)

Model comparison involving β

MATLAB Command Window Page 1

Informative prior

p(beta_j>0|y)	95% HPDI	99% HPDI	PO for beta_j=0
0.13	[-10956.68, 2886.57]	[-13143.17, 5073.07]	4.14e+00
1.00	[4.71, 6.15]	[4.49, 6.38]	2.25e-39
0.99	[563.52, 5210.11]	[-170.40, 5944.02]	3.94e-01
1.00	[13616.30, 20314.17]	[12558.39, 21372.08]	1.72e-19
1.00	[5686.37, 9596.10]	[5068.85, 10213.62]	1.22e-11

Nonnformative prior

p(beta_j>0|y)	95% HPDI	99% HPDI	PO for beta_j=0
0.13	[-11054.72, 3035.62]	[-13280.35, 5261.25]	NaN
1.00	[4.71, 6.15]	[4.48, 6.38]	NaN
0.99	[449.30, 5199.93]	[-301.09, 5950.32]	NaN
1.00	[13713.83, 20496.52]	[12642.47, 21567.88]	NaN
1.00	[5664.00, 9605.79]	[5041.38, 10228.42]	NaN