

1 The Role of the Prior: A Thought Experiment

1.1 Preamble

Koop chapter 2 provides a complete Bayesian analysis of the simple regression model with one regressor and no intercept, given in his equation (2.1)

$$y_i = \beta x_i + \varepsilon_i \quad (1)$$

with a sample of N observations and under the same assumptions. This analysis includes posterior for parameters, model comparison and prediction.

Here we focus on interpretation of the posterior. Natural conjugate priors are priors that return posteriors with the same distribution as the likelihood. An obvious advantage is tractability. A more subtle result is that then “the prior can be interpreted as arising from a fictitious data set from the same process that generated the actual data (Koop page 18).”

This handout focuses on this interpretation, for a model with normal likelihood (which can be re-expressed as normal-gamma) with normal prior for β and gamma prior for h leading to a normal-gamma posterior.

1.2 Interpretation of Prior Information as Data

Consider the expression of the log-likelihood, $\log(p(y_i|\beta, \sigma^2))$ under the assumption of normality and split the sample into two sub-samples, such that $N_1 + N_2 = N$:

$$\begin{aligned} \log(p(y_i|\beta, \sigma^2)) &= -\frac{N_1}{2} \log(2\pi) - \frac{N_1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum^{N_1} (y_i - \beta x_i)^2 \\ &\quad - \frac{N_2}{2} \log(2\pi) - \frac{N_2}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum^{N_2} (y_i - \beta x_i)^2 \end{aligned}$$

The usual maximum-likelihood estimator for β is obtained by taking the derivative of the log-likelihood with respect to β

$$\frac{\partial L}{\partial \beta} = \frac{1}{\sigma^2} \left[\left(\sum^{N_1} y_i x_i - \beta \sum^{N_1} x_i^2 \right) + \left(\sum^{N_2} y_i x_i - \beta \sum^{N_2} x_i^2 \right) \right] = 0$$

Hence

$$\left(\sum^{N_1} y_i x_i + \sum^{N_2} y_i x_i \right) = \beta \left(\sum^{N_1} x_i^2 + \sum^{N_2} x_i^2 \right)$$

For reasons that will become clear momentarily, define:

$$\begin{aligned}\bar{V} &= \left(\sum^{N_1} x_i^2 + \sum^{N_2} x_i^2 \right)^{-1} \\ \hat{\beta}_1 &= \frac{\sum^{N_1} y_i x_i}{\sum^{N_1} x_i^2},\end{aligned}$$

and similarly for $\hat{\beta}_2$. Then, the maximum-likelihood estimator of β can be expressed as

$$\bar{\beta} = \bar{V} \left(\hat{\beta}_1 \sum^{N_1} x_i^2 + \hat{\beta}_2 \sum^{N_2} x_i^2 \right) \quad (2)$$

or in other words, the maximum-likelihood estimator $\bar{\beta}$ is a weighted average of the subsample estimates with weights

$$\frac{\sum^{N_1} x_i^2}{\sum^{N_1} x_i^2 + \sum^{N_2} x_i^2}; \text{ and } \frac{\sum^{N_2} x_i^2}{\sum^{N_1} x_i^2 + \sum^{N_2} x_i^2}$$

that add up to one. The notation \bar{V} is chosen to correspond to the notation used in expression (2.9) in Koop. The previous derivations are perhaps useful in gauging the weight of the prior on the posterior mean. Specifically, notice that expression (2.10) in Koop gives the posterior mean as

$$\bar{\beta} = \bar{V} \left(\underline{\beta} \underline{V}^{-1} + \hat{\beta} \sum x_i^2 \right) \quad (3)$$

where $\hat{\beta}$ would be the usual OLS or MLE estimator, $\bar{\beta}$ is the posterior mean, the prior is given by $\beta|h \sim N(\underline{\beta}, h^{-1}\underline{V})$, and

$$\bar{V} = \left(\underline{V}^{-1} + \sum x_i^2 \right)^{-1} \quad (4)$$

so that the correspondence between (2) and (3) is clear.

More generally, suppose that in a classical context we want to estimate (1) by maximum likelihood and we want to impose a constraint of the type $r = \beta R$ (e.g., $r = 0$, and $R = 1$, would be the usual null hypothesis of significance for β) in a “flexible” manner. For example, one way to specify the uncertainty about this constraint would be

$$r = \beta R + \underline{V}^{1/2} \varepsilon_{N+1} \quad \varepsilon_{N+1} \sim N(0, \sigma^2) \quad (5)$$

Thus, as $\underline{V} \rightarrow 0$, the constraint becomes binding, and as $\underline{V} \rightarrow \infty$ we express our uncertainty about the validity of the constraint. Using the Koop notation and setting $r = \underline{\beta}$; $R = 1$ and taking σ^2 as known (for simplicity), notice that (5) is just another way of specifying a prior on β .

Taking logs of the product of the likelihood times this prior is equivalent to maximizing the log-likelihood over the sample N , augmented by the prior, that is

$$\begin{aligned} \log(p(y_i|\beta, \sigma^2)p(\beta|\sigma^2)) &= -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \beta x_i)^2 \\ &\quad - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log \sigma^2 \underline{V} - \frac{1}{2\sigma^2 \underline{V}} (\underline{\beta} - \beta)^2 \end{aligned}$$

Maximization with respect to β involves the first order condition

$$\frac{\partial L}{\partial \beta} = \frac{1}{\sigma^2} \left[\left(\sum y_i x_i - \beta \sum x_i^2 \right) + \left(\frac{\beta}{\underline{V}} - \frac{\underline{\beta}}{\underline{V}} \right) \right] = 0$$

Using definition (2.9) in Koop, which appears in expression (4) as well, we have that the solution to this maximization problem is

$$\bar{\beta} = \bar{V} \left(\hat{\beta} \sum x_i^2 + \underline{\beta} \underline{V}^{-1} \right)$$

which is just expression (3).

1.3 An alternative derivation in terms of OLS

Consider now the corresponding derivations in the context of OLS. For OLS of y_i on x_i with the sample split into two (with $N = N_1 + N_2$ observations), OLS yields

$$\begin{aligned} \hat{\beta} &= \left(\sum_{i=1}^N x_i^2 \right)^{-1} \sum_{i=1}^N x_i y_i \\ &= \left(\sum_{i=1}^{N_1} x_i^2 + \sum_{i=1}^{N_2} x_i^2 \right)^{-1} \left(\sum_{i=1}^{N_1} x_i y_i + \sum_{i=1}^{N_2} x_i y_i \right) \\ &= \left(\sum_{i=1}^{N_1} x_i^2 + \sum_{i=1}^{N_2} x_i^2 \right)^{-1} \left(\sum_{i=1}^{N_1} x_i^2 \left[\sum_{i=1}^{N_1} x_i^2 \right]^{-1} \sum_{i=1}^{N_1} x_i y_i + \sum_{i=1}^{N_2} x_i^2 \left[\sum_{i=1}^{N_2} x_i^2 \right]^{-1} \sum_{i=1}^{N_2} x_i y_i \right) \end{aligned}$$

$$= \bar{V} \left(\sum^{N_1} x_i^2 \hat{\beta}_1 + \sum^{N_2} x_i^2 \hat{\beta}_2 \right)$$

which is (2) given earlier.

For regression with stochastic constraints, stack the model as

$$\begin{aligned} y_1 &= x_1 \beta + \varepsilon_1 \\ y_2 &= x_2 \beta + \varepsilon_2 \\ &\vdots \\ y_N &= x_N \beta + \varepsilon_N \\ r &= R\beta + \underline{V}^{1/2} \varepsilon_{N+1} \end{aligned}$$

and OLS with these $N + 1$ observations will yield $\hat{\beta}$ equal estimator in (3).

1.4 Summary

What have we learned from this exercise?

1. The assumption on the prior variance term \underline{V} is an assumption about the relative certainty/weight that we have on the prior information.
2. Thinking of the prior as augmenting the sample of data with additional information, notice that \underline{V} plays the same role as $(\sum^{N_1} x_i^2)^{-1}$ in expression (2).
3. Another interpretation of the prior in a classical context is that of imposing a “flexible” constraint on the possible range of values of β , as expression (5) shows.
4. Notice that when the prior is given a lot of weight, then $\underline{V} \rightarrow 0$ and from expression (3), it is clear that $\bar{\beta} \rightarrow \beta$. Conversely, a completely agnostic prior means a lot of uncertainty in your beliefs, or $\underline{V} \rightarrow \infty$. In this case, $\bar{\beta} \rightarrow \hat{\beta}$.
5. Hence, the quantity $\bar{V} \underline{V}^{-1}$ is an expression of the relative weight of the prior since $\bar{V} (\sum x_i^2 + \underline{V}^{-1})$ adds up to one.