Introduction to Bayesian Econometrics

Andrea Carriero

January 2018

The classical linear regression model (CLRM)

- Consider the following linear regression and the task of estimating β

$$Y = X\beta + \varepsilon;\ \varepsilon \sim N(0, \sigma^2 I_T)$$

- In the standard approach we write down the likelihood function

$$p(Y|\beta, \sigma^2) = (2\pi)^{-\frac{T}{2}} \left|\sigma^2 I_T\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(Y - X\beta)'\left(\sigma^2 I_T\right)^{-1}(Y - X\beta)\right]$$

- Then we obtain data and maximize $p(Y|\beta, \sigma^2)$, which gives the standard OLS estimator

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y$$

- Incorporates information from the data only. Bayesian analysis allows to combine our beliefs about β with information from the data

More on the CLRM

- More specifically, Maximum Likelihood esimation gives:

$$\hat{\beta} \sim N(\beta, \sigma^2 \left(X'X\right)^{-1}),$$

but since σ^2 is usually unknown it is estimated with

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{T - k}.$$

- Noting that

$$(T - k)\hat{\sigma}^2/\sigma^2 = \frac{\varepsilon'}{\sigma}(I - P_X)\frac{\varepsilon}{\sigma} \sim \chi^2_{T-k},$$

we have

$$\frac{\frac{\hat{\beta}-\beta}{\sqrt{\sigma^2\left(X'X\right)^{-1}}}}{\sqrt{\frac{(T-k)\hat{\sigma}^2/\sigma^2}{(T-k)}}} \sim t_{T-K} \rightarrow \hat{\beta} \sim t_{T-K}(\beta, \hat{\sigma}^2 \left(X'X\right)^{-1}),$$

which is approximately normal in reasonably large samples.

Updating a linear projection

- Start with:

$$y = X\beta + \varepsilon; \ \varepsilon \sim N(0, \sigma^2 I_T)$$

  and get

$$\hat{\beta} = (X'X)^{-1} X'Y$$

- Add data:

$$\begin{bmatrix} Y \\ Y_1 \end{bmatrix} = \begin{bmatrix} X \\ X_1 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ \varepsilon_1 \end{bmatrix}; \ \varepsilon \sim N(0, \sigma^2 I_{T+T_1})$$

  and get

$$\begin{aligned} \hat{\beta} &= \left(\begin{bmatrix} X' & X_1' \end{bmatrix} \begin{bmatrix} X \\ X_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} X' & X_1' \end{bmatrix} \begin{bmatrix} Y \\ Y_1 \end{bmatrix} \\ &= \left(X'X + X_1'X_1 \right)^{-1} (X'Y + X_1'Y_1) \end{aligned}$$

The Bayesian approach to the CLRM

Bayesian approach

1. The researcher starts with a **prior** belief about the coefficient β. The prior belief is in the form of a distribution $p(\beta)$

$$\beta \sim N(\beta_0, \Sigma_0)$$

2. Collect data and write down the **likelihood** function as before $p(Y|\beta)$.
3. Update your prior belief on the basis of the information in the data. Combine the prior distribution $p(\beta)$ and the likelihood function $p(Y|\beta)$ to obtain the **posterior** distribution $p(\beta|Y)$

Key identities

- These three steps come from Bayes Theorem:

$$p(\beta|Y) = \frac{p(Y|\beta) \times p(\beta)}{p(Y)}$$

- Useful identities:

$$\underset{joint}{p(Y,\beta)} = \underset{data\ density}{p(Y)} \times \underset{posterior}{p(\beta|Y)} = \underset{likelihood}{p(Y|\beta)} \times \underset{prior}{p(\beta)}$$

- $p(Y)$ is the data density (also known as marginal likelihood).It is the constant of integration of the posterior:

$$\int p(\beta|Y)d\beta = \frac{1}{p(Y)} \int \underbrace{p(Y|\beta) \times p(\beta)}_{posterior\ \ kernel} d\beta = 1,$$

therefore it is not needed if we are only interested in the posterior kernel. The posterior kernel is sufficient to compute e.g. mean and variance of $p(\beta|Y)$.

Prior distribution of coefficients

We assume for the moment that σ^2 is known, k is the number of regressors.

1. Set prior distribution for $\beta \sim N(\beta_0, \Sigma_0)$

$$\begin{aligned} p(\beta) &= (2\pi)^{-\frac{k}{2}} \left|\Sigma_0\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)\right] \\ &\propto \exp\left[-0.5(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)\right] \end{aligned}$$

2. Obtain data and form the likelihood function:

$$\begin{aligned} p(Y|\beta) &= (2\pi)^{-\frac{T}{2}} \left|\sigma^2 I_T\right|^{-\frac{1}{2}} \exp[-\frac{1}{2}(Y-X\beta)'(\sigma^2 I_T)^{-1}(Y-X\beta)] \\ &\propto \exp[-0.5(Y-X\beta)'(Y-X\beta)/\sigma^2] \end{aligned}$$

Posterior distribution of coefficients

3. Obtain the posterior kernel

$$
\begin{aligned}
p(\beta|Y) &\propto p(Y|\beta) \times p(\beta) \\
&\propto \exp[-0.5\,(\beta-\beta_0)'\,\Sigma_0^{-1}\,(\beta-\beta_0)] \\
&\quad \times \exp[-0.5\,(Y-X\beta)'\,(Y-X\beta)\,/\sigma^2] \\
&\propto \exp[-0.5\{(\beta-\beta_0)'\,\Sigma_0^{-1}\,(\beta-\beta_0) + (Y-X\beta)'\,(Y-X\beta)\,/\sigma^2\}] \\
&\propto \exp[-0.5\,(\beta-\beta_1)'\,\Sigma_1^{-1}\,(\beta-\beta_1)]
\end{aligned}
$$

where the last step uses:

$$
\Sigma_1 = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1} \tag{1}
$$

$$
\beta_1 = \Sigma_1\left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2}X'Y\right) \tag{2}
$$

This is the kernel of a normal distribution. Therefore we can write:

$$
\beta|\sigma^2, Y \sim N\,(\beta_1, \Sigma_1)
$$

Posterior distribution of coefficients - details

we have:

$$p(\beta|Y) \propto \exp[-0.5\{(\beta-\beta_0)' \Sigma_0^{-1} (\beta-\beta_0) + (Y - X\beta)' (Y - X\beta) / \sigma^2\}]$$

completing the squares gives:

$$\begin{aligned} k &= \beta'\Sigma_0^{-1}\beta - \beta'\Sigma_0^{-1}\beta_0 - \beta_0'\Sigma_0^{-1}\beta + \beta_0'\Sigma_0^{-1}\beta_0 + \\ &+ Y'(\sigma^2)^{-1}Y - Y'(\sigma^2)^{-1}X\beta - \beta'X'(\sigma^2)^{-1}Y + \beta'X'(\sigma^2)^{-1}X\beta \end{aligned}$$

regrouping gives:

$$\begin{aligned} k = &\beta'\underbrace{(\Sigma_0^{-1} + X'(\sigma^2)^{-1}X)}_{\Sigma_1^{-1}}\beta - \beta'\underbrace{(\Sigma_0^{-1}\beta_0 + X'(\sigma^2)^{-1}Y)}_{\Sigma_1^{-1}\beta_1} + \\ &- \underbrace{(\beta_0'\Sigma_0^{-1} + Y'(\sigma^2)^{-1}X)}_{\beta_1'\Sigma_1^{-1}}\beta + \beta_0'\Sigma_0^{-1}\beta_0 + Y'(\sigma^2)^{-1}Y \end{aligned}$$

where the elements in braces follow from the definitions (1) and (2).

Posterior distribution of coefficients - details

$$k = \beta'\Sigma_1^{-1}\beta - \beta'\Sigma_1^{-1}\beta_1 - \beta_1'\Sigma_1^{-1}\beta + \beta_0'\Sigma_0^{-1}\beta_0 + Y'(\sigma^2)^{-1}Y \tag{3}$$

The last two terms will remain as they are. Rewrite the first term as:

$$\begin{aligned}
\beta'\Sigma_1^{-1}\beta &= (\underbrace{\beta - \beta_1} + \underbrace{\beta_1})'\Sigma_1^{-1}(\beta - \beta_1 + \beta_1) \\
&= \underbrace{(\beta - \beta_1)'}\Sigma_1^{-1}\underbrace{(\beta - \beta_1)} + (\underbrace{\beta_1})'\Sigma_1^{-1}(\underbrace{\beta_1}) \\
&\quad + \underbrace{(\beta - \beta_1)'}\Sigma_1^{-1}(\underbrace{\beta_1}) + (\underbrace{\beta_1})'\Sigma_1^{-1}\underbrace{(\beta - \beta_1)}.
\end{aligned}$$

Simplifying the $\beta_1'\Sigma_1^{-1}\beta_1$ appearing in the last three terms (+,-,-) gives:

$$\beta'\Sigma_1^{-1}\beta = (\beta - \beta_1)'\Sigma_1^{-1}(\beta - \beta_1) - \beta_1'\Sigma_1^{-1}\beta_1 + \underline{\beta'\Sigma_1^{-1}\beta_1} + \underline{\beta_1{}'\Sigma_1^{-1}\beta}$$

The terms underlined simplify with those in (3), which then becomes:

$$\begin{aligned}
k &= (\beta - \beta_1)'\Sigma_1^{-1}(\beta - \beta_1) - \beta_1'\Sigma_1^{-1}\beta_1 + \beta_0'\Sigma_0^{-1}\beta_0 + Y'(\sigma^2)^{-1}Y \\
&\propto (\beta - \beta_1)'\Sigma_1^{-1}(\beta - \beta_1)
\end{aligned}$$

Comparison with OLS

- Note that, as $X'X\hat{\beta} = X'Y$, we have:

$$\beta_1 = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} X'X\right)^{-1} \left(\Sigma_0^{-1}\beta_0 + \frac{1}{\sigma^2} X'X\hat{\beta}\right)$$

- Without the priors, these moments are simply the OLS estimates
- Without the data, these moments are simply the priors
- The mean is a weighted average of the prior and OLS.
- The weights are inversely proportional to the precision of prior and data information
- Σ_0^{-1} and $\Sigma_0^{-1}\beta_0$ are the prior moments, can be interpreted as dummy observations/pre-sample observations.
- Setting $\Sigma_0^{-1} = \frac{\lambda}{\sigma^2} I$ and $\beta_0 = 0$ gives the Ridge regression:

$$\beta_1 = \left(\lambda I + X'X\right)^{-1} X'Y.$$

The Likelihood Principle

- Consider tossing a drawing pin (Lindley and Phillips 1976). Luigi says he tossed it 12 times and obtained :

$$\{U, U, U, D, U, D, U, U, U, U, U, D\}$$

- You -as a statistician- are asked to give a 5% rejection region for the null that U and D are equally likely
- Obtaining 9 U's out of 12 suggests that the chance of its falling uppermost (U) exceeds 50%.The results that would even more strongly support this conclusion are:

$$(10, 2),\ (11, 1),\ \textit{and}\ (12, 0),$$

  so that, under the null hypothesis $\theta = 1/2$, the chance of the observed result, or more extreme, is:

$$\left\{ \binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0} \right\} \left(\theta = \frac{1}{2} \right)^{12} = \mathbf{7.5\%} > 5\%$$

- Hence, you do NOT reject the null that U and D are equally likely (50%).

The Likelihood Principle

- However, now Luigi tells you: *"but I didn't set to throw the pin* 12 *times. My plan was to throw the pin until 3 Ds appeared"*
- Does this change your inference? Yes it does
- Under the new scenario, the more exreme events would be:

$$(10, 3),\ (11, 3),\ (12, 3),\ ...,$$

  while the events $(10, 2)$, $(11, 1)$, and $(12, 0)$ actually can NOT take place under this design.
- So the chance of the observed result under the null hypothesis becomes:

$$\left\{ 1 - \binom{10}{2} \left(\frac{1}{2}\right)^{11} - \binom{9}{2} \left(\frac{1}{2}\right)^{10} - ... - \binom{2}{2} \left(\frac{1}{2}\right)^{3} \right\} = \mathbf{3.25}\% < 5\%$$

- Why is this happening? Because the two setups imply a different **stopping rule** (stop at 12 draws, or stop at 3 D draws). This, more generally, alters the **sample space**.

The Likelihood Principle

- Things are even more problematic. Think if Luigi says *"I just kept drawing the pin until lunch was served"*. How would you tackle this?
- Confidence intervals similarly demand consideration of the sample space. Indeed, so does every statistical technique, with the exception of maximum likelihood.

Lindley and Phillips (1976): Many people's intuition says this specification is irrelevant. Their argument might more formally be expressed by saying that the evidence is of 12 honestly reported tosses, 9 of which were U; 3, D. Furthermore, these were in a particular order, that reported above. **Of what relevance are things that might have happened [e.g. no lunch], but did not?**

Indeed, this helps us understand **The LIKELIHOOD PRINCIPLE**:

1. All the information about θ obtainable from an experiment is contained in the likelihood function for θ given the data.
2. Two likelihood functions for θ(from the same or different experiments) contain the same information about θ if they are proportional to one another.

The Likelihood Principle

- By using only the likelihood, and nothing else from the experiment, the answer to the problem is the same regardless of the stopping rule.
- Indeed, let $x_1 = \#U$ in experiment 1 $x_2 = \#U$ in experiment 2
- In experiment 1 (E1) we have a binomial density:

$$f_\theta^1(x_1) = \binom{12}{x_1}\theta^{x_1}(1-\theta)^{12-x_1} \Longrightarrow \ell_\theta^1(9) = \binom{12}{9}\theta^9(1-\theta)^3$$

- In experiment 2 (E2) we have a negative binomial density:

$$f_\theta^2(x_2) = \binom{x_2+3-1}{x_2}\theta^{x_2}(1-\theta)^3 \Longrightarrow \ell_\theta^2(9) = \binom{11}{9}\theta^9(1-\theta)^3$$

- In this situation, the Likelihood Principle says that:
  1. for experiment E1 alone the information about θ is contained solely in $\ell_\theta^1(9)$;
  2. for experiment E2 alone the information about θ is contained solely in $\ell_\theta^2(9)$;
  3. since $\ell_\theta^1(9)$ and $\ell_\theta^2(9)$ are proportional as functions of θ, the information about θ in the two experiments is identical

Error variance

- We assume for the moment that β is known. A typical prior for the variance σ^2 is an inverse Gamma prior.
- Suppose we have ν_0 $i.i.d.$ observations from a normal distribution:

$$v_t \sim N(0, 1/s_0^2).$$

- Then $s_0 v_t \sim N(0,1)$ and the sum of squares of these is

$$\sum_{t=1}^{\nu_0} (s_0 v_t)^2 \sim \chi^2(\nu_0).$$

- Defining $h = \sum_{t=1}^{\nu_0} v_t^2$ we can write $s_0^2 h \sim \chi^2(\nu_0)$ with pdf:

$$f_{s_0^2 h}\left(s_0^2 h\right) = [2^{\frac{\nu_0}{2}} \Gamma(\nu_0/2)]^{-1} (s_0^2 h)^{\frac{\nu_0 - 2}{2}} \exp(-s_0^2 h/2).$$

Error variance - gamma

- If $s_0^2 h \sim \chi^2(\nu_0)$ then h has the so-called gamma distribution (and vice-versa):

$$\begin{aligned} h &= \sum_{t=1}^{\nu_0} v_t^2 \sim \Gamma\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right); \\ f_h(h) &= [\Gamma(\nu_0/2)]^{-1}(s_0^2/2)^{\frac{\nu_0}{2}} h^{\frac{\nu_0-2}{2}} \exp(-s_0^2 h/2) \end{aligned}$$

- The pdf $f_h(h)$ above can be obtained using the change of variable theorem.
- This theorem states that if x is a random variable (and we know its pdf $f_x(\cdot)$), and $z = r(x)$ is an invertible function of it (and therefore $x = r^{-1}(z)$), then the pdf of z can be derived as follows:

$$f_z(z) = \left|\frac{d}{dz} r^{-1}(z)\right| \times f_x(r^{-1}(z))$$

- In this case $x = s_0^2 h \sim f_x$ and $z = h = x/s_0^2 \sim f_z$. So $r^{-1}(z) = x = s_0^2 \times h$, and:

$$f_h(h) = \left|s_0^2\right| \times f_x(s_0^2 h)$$

Error variance - change of variable

Indeed we have:

$$f_{s_0^2 h}\left(s_0^2 h\right) = [2^{\frac{\nu_0}{2}}\Gamma(\nu_0/2)]^{-1}(s_0^2 h)^{\frac{\nu_0-2}{2}}\exp(-s_0^2 h/2) \sim \chi^2(\nu_0).$$

and

$$\begin{aligned}
f_h(h) &= \left|s_0^2\right| \times f_x(s_0^2 h) \\
&= \left|s_0^2\right| [2^{\frac{\nu_0}{2}}\Gamma(\nu_0/2)]^{-1}(s_0^2 h)^{\frac{\nu_0-2}{2}}\exp(-s_0^2 h/2) \\
&= \left|s_0^2\right| [2^{\frac{\nu_0}{2}}\Gamma(\nu_0/2)]^{-1} s_0^{2\left(\frac{\nu_0}{2}-1\right)} h^{\frac{\nu_0-2}{2}}\exp(-s_0^2 h/2) \\
&= [2^{\frac{\nu_0}{2}}\Gamma(\nu_0/2)]^{-1} s_0^{2\frac{\nu_0}{2}} h^{\frac{\nu_0-2}{2}}\exp(-s_0^2 h/2) \\
&= [\Gamma(\nu_0/2)]^{-1}(s_0^2/2)^{\frac{\nu_0}{2}} h^{\frac{\nu_0-2}{2}}\exp(-s_0^2 h/2) \sim \Gamma\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right)
\end{aligned}$$

Error variance - inverse gamma

- Using a second change of variable $\sigma^2 = h^{-1}$ yields:

$$\begin{aligned}\sigma^2 &\sim \Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right); \\ f_{\sigma^2}(\sigma^2) &\propto [\Gamma(\nu_0/2)]^{-1}(s_0^2/2)^{\frac{\nu_0}{2}}(\sigma^2)^{-\frac{\nu_0+2}{2}}\exp(-s_0^2/2\sigma^2),\end{aligned}$$

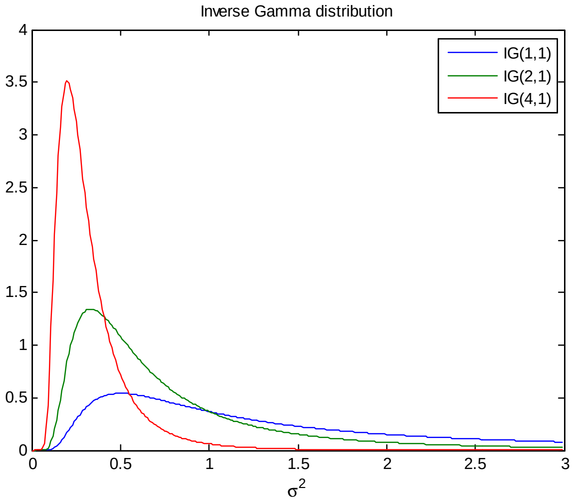
  In this case $x = h \sim f_x$ and $z = h^{-1} \sim f_z$. So $r^{-1}(z) = x = h$, and $f_{\sigma^2}(h^{-1}) = 1 \times f_h(h)$, that is we simply use $h = \frac{1}{\sigma^2}$ in $f_h\,(h)$.
- Then σ^2 has an inverse gamma distribution with mean $\frac{s_0^2}{2}/(\frac{\nu_0}{2} - 1)$ and variance $(\frac{s_0^2}{2})^2/((\frac{\nu_0}{2} - 1)^2(\frac{\nu_0}{2} - 2))$.
- Instead h is the precision, and has a gamma distribution with mean $\frac{\nu_0}{2}/\frac{s_0^2}{2} = \nu_0/s_0^2$ and variance $\frac{\nu_0}{2}/\left(\frac{s_0^2}{2}\right)^2 = 2\nu_0/s_0^4$.

Error variance - drawing from

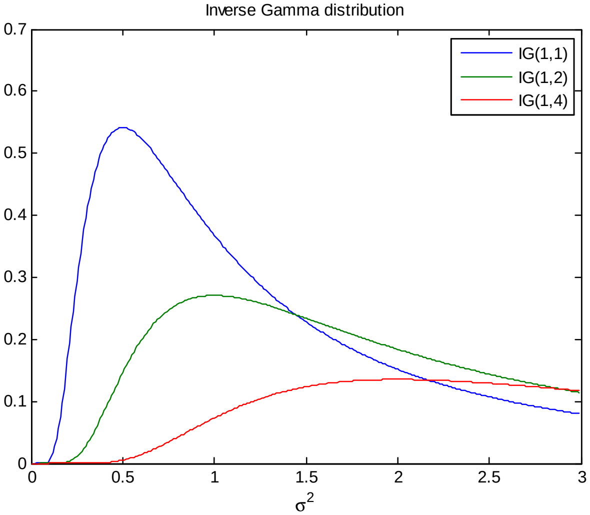
To draw σ^2 we can:

- Draw a vector of dimension ν_0 from a Gaussian distribution, i.e. $\underset{\nu_0 \times 1}{a} = s_0 v$ where $v \sim N(0, I_{\nu_0})$.
- The quantity $a'a = \frac{1}{s_0^2} v'v$ is a random draw of the precision h from $\Gamma\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right)$.
- The inverse $(a'a)^{-1} = s_0^2 / v'v$ is a draw of σ^2 from $\Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{s_0^2}{2}\right)$.

The prior distribution for different degrees of freedom



The prior distribution for different scale matrices



Conditional Posterior of error variance

1. Set prior distribution $\sigma^2 \sim \Gamma^{-1}\left(v_0/2, s_0^2/2\right)$

$$p(\sigma^2) = [\Gamma(\nu_0/2)]^{-1}(s_0^2/2)^{\frac{\nu_0}{2}}(\sigma^2)^{-\frac{\nu_0+2}{2}} \exp\left(-s_0^2/2\sigma^2\right)$$

2. Obtain data and form the likelihood function

$$p(Y|\beta,\sigma^2) = (2\pi)^{-\frac{T}{2}}\left|\sigma^2 I_T\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(Y - X\beta)'\left(\sigma^2 I_T\right)^{-1}(Y - X\beta)\right]$$

3. Obtain the conditional posterior kernel

$$p\left(\sigma^2|Y,\beta\right) \propto (\sigma^2)^{-\frac{T+\nu_0+2}{2}} \exp\left[-\{s_0^2 + (Y - X\beta)'(Y - X\beta)\}/2\sigma^2\right]$$

which is the kernel of an inverse gamma

$$\Gamma^{-1}(v_1/2, s_1^2/2)$$

with

$$v_1 = T + v_0,\ \ s_1^2 = s_0^2 + (Y - X\beta)'(Y - X\beta).$$