# STAT 5170: Applied Time Series

*Course notes for part A of learning unit 4*

Our discussion thus far has focused primarily on time series models, while only touching on issues of inference. Our attention now turns to the latter.

**Section 4A.1: Bayesian inference.**

When faced with the task of data analysis, we might have a model for the phenomenon that produced the data in mind (*e.g.*, an $AR(1)$ model), but we would be substantially *uncertain* about the values its parameters (*e.g.*, the value of $\phi_1$ in the $AR(1)$ model). The general goal of data analysis is to *reduce uncertainty* about the model parameters. Once that is accomplished, our aim is to *summarize uncertainty* by producing statistical reports in the form of point and interval estimates, predictions, or other relevant evidence summaries about the model's characteristics.

The approach just describes outlines the *Bayesian* framework for statistical inference. The components of this approach are as follows:

1. The state of uncertainty about the model parameters, $\boldsymbol{\theta}$, as the analyst faces the task of data analysis is described probabilistically, in the form of probability distribution called the *prior distribution*. In multiparameter models, the notation $\boldsymbol{\theta}$ is to represent all model parameters (*e.g.*, $\boldsymbol{\theta} = (\phi_1, \sigma_w^2)$ in the $AR(1)$ model). The parameter space is denoted $\Theta$, which is the space of all sensible parameter values. The notation $p(\boldsymbol{\theta})$ identifies the probability mass or density function of the prior distribution. In the examples we will discuss here, the parameter space is continuous and $p(\boldsymbol{\theta})$ is a density function.

2. The model for the data at hand, $\boldsymbol{x}$ (*e.g.*, a single time series $\boldsymbol{x} = x_1, \ldots, x_n$) is described probabilistically in the form of probability distribution called the *data-generating distribution*. The notation $p(\boldsymbol{x}|\boldsymbol{\theta})$ identifies the probability mass or density function of the data-generating distribution. Oftentimes, out of consistency with broader discussions of statistical inference, $p(\boldsymbol{x}|\boldsymbol{\theta})$ is called the *likelihood function*.

3. The reduced state of uncertainty about the model parameters that arises from examining the data at hand is deduced by applying Bayes's Theorem to the prior and data-generating distributions. The result is called the *posterior distribution*, whose probability mass or density function is

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

   This formula is no more than a restatement of Bayes's Theorem. The act of working out the posterior distribution is sometimes called "updating" prior knowledge.

The quantity in the numerator of Bayes's Theorem, $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, is the probability mass or density function that defines the joint distribution of the parameters and data. The

quantity in the denominator, $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, is the probability mass or density function that defines the marginal distribution of the data; it is sometimes instead called the *integrated likelihood*.

Using such generic terminology as joint and marginal distributions may take some of the magic out of such notions as an "initial state of uncertainty" and "updating uncertainty," and recast the core of Bayesian methodology as no more than developing a joint probability model for the parameters, $\boldsymbol{\theta}$, and data, $\boldsymbol{x}$, whose mass or density function is $p(\boldsymbol{x}, \boldsymbol{\theta})$. From this perspective, the prior and data-generating distributions are just devices that help us to formulate the joint model, and Bayes's Theorem is a device for summarizing what the data tell us about the parameter.

A huge advantage of Bayesian methods over other methods of inference is its flexibility in both conceptualizing and interpreting statistical reports that arise as summaries of the posterior distribution. For example, if the report of uncertainty about a univariate parameter, $\theta$, is given as 2.5% and 97.5% percentiles of the posterior distribution, then it makes sense to interpret that report as the posterior *probability* that the parameter falls between those values. There is no conceptual fuss that would recast the reported interval as a statement of "confidence," or as a probabilistic report about hypothetical replications of data-generation that are never to be carried out. Bayesian methods are also valuable for their capacity to capture and summarize extraordinarily complicated models and associated aspects of uncertainty.

Nevertheless, the development of Bayesian methods has faced a number of challenges:

1. The first is the modeling challenge of specifying a prior distribution. As illustrated in the examples, below, choosing a prior distribution is rife with subtle complications around issues of parameterization (specifically, the issue of reparameterization by transformation) and around what information is used or not used to model prior uncertainty. On the latter point, there has come to be a split between methods that either (*i.*) aim to *elicit* prior uncertainty from a variety of existing sources apart from the data at hand, or (*ii.*) aim to formulate a prior distribution that describes a state of *ignorance* about prior uncertainty. The second approach is connected to what is sometimes called the *objective Bayes* framework, which has come into prominence in recent decades. Because it simplifies the modeling step of selecting a prior distribution (though not without ambiguities and, ultimately, implications of logical inconsistency), the objective Bayes approach is convenient for setting up examples in a classroom setting such as ours.

2. The second challenge is the practical difficulty of producing a posterior distribution. Among the largest hurdles to implementing a Bayesian data-analysis can be formulating a strategy for dealing with the denominator of Bayes's Theorem, $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, the integrated likelihood function. As you know from previous calculus classes, integrals can be immensely challenging to work out analytically, even with such clever tools for change-of-variables and integration-by-parts. In common Bayesian formulations, it is not unusual that the integrated likelihood function is *impossible* to work out analytically, or simply not worth the effort. A satisfactory solution has been

to look toward *numerical integration* methods, particularly those that involve simulation. Subsequent discussion and examples will highlight a small handful of these methods, just a fraction of the array of computational strategies that have become available for implementing Bayesian methods in practice.

A final comment before diving into examples of Bayesian data-analysis is that certain specialized summaries of the posterior distribution have also been developed for purposes of *predicting* unobserved values of the data, such as values that are yet to be generated. In time series analysis, this offers another approach to *forecasting* future values of a time series, beyond those of the data at hand. However, the general prediction framework is also useful for the separate purpose of *model checking*, in which diagnostics are examined for evaluating whether the form of the data-generating distribution (or other aspects of the joint Bayesian model) captures features that may appear to be present in the data.

Within the prediction setup, the aim is to summarize uncertainty about an unobserved data value $\boldsymbol{x}^*$, such as a value that is yet to be generated. As part of the overall Bayesian setup, an extension of the data-generating distribution is specified to describe how the unobserved data value *would* be generated. It is described probabilistically in the form of probability distribution with mass or density function $p(\boldsymbol{x}^*|\boldsymbol{\theta})$, which could be extended to $p(\boldsymbol{x}^*|\boldsymbol{x}, \boldsymbol{\theta})$ in order to take into account dependencies of the unobserved data value, $\boldsymbol{x}^*$, with the data at hand, $\boldsymbol{x}$. For example, when the aim is to forecast a future value given by a time series model, it typically makes sense, due to autocorrelation, that $\boldsymbol{x}^*$ would be dependent on $\boldsymbol{x}$. Subsequently, the reduced state of uncertainty about the unobserved data value is described by the *posterior predictive distribution*, whose probability mass or density function is given by the formula

$$p(\boldsymbol{x}^*|\boldsymbol{x}) = \int p(\boldsymbol{x}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta} \quad \text{or} \quad p(\boldsymbol{x}^*|\boldsymbol{x}) = \int p(\boldsymbol{x}^*|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta},$$

depending on whether $\boldsymbol{x}^*$ and $\boldsymbol{x}$ are dependent. Observe that the posterior predictive distribution is defined by integration with respect to the posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{x})$, which itself is formulated through integration. Such double-integration might be thought to intensify the computational burden of Bayesian data-analysis. We will see, however, that prediction fits neatly within computational strategies that make use of simulation.

**Section 4A.2: Elementary techniques.**

The sequence of examples that follow introduce the techniques that we will use for Bayesian analysis of time series, starting will elementary contexts that you are likely to have encountered in previous courses, and then specializing to contexts for analysis of time-series data.

**Example: Binomial-beta models**

As a sort of "warm-up" to deeper explorations, we consider the canonical setup for incidence-counts within a sequence of independent trials. Suppose $X|\theta \sim \text{binomial}(n, \theta)$ is a binomial random variable counting the total number of incidences of some interesting outcome in a sequence of $n$ independent trials. This defines the data-generating distribution, with

probability mass function

$$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

A very early approach to Bayes methodology under this formulation would specify the uniform distribution, $\theta \sim \text{uniform}(0,1)$, as the prior distribution, as a description of "ignorance" about the parameter. Greater flexibility in specifying the prior distribution is made possible by embedding the uniform distribution distribution within the class of beta distributions, $\theta \sim \text{beta}(\alpha, \beta)$. The associated prior density is

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

The flexibility to specify the prior parameters $\alpha$ and $\beta$ offers the opportunity to incorporate any available background knowledge (and uncertainty) about the data-generating process.

A convenient aspect of specifying a beta prior distribution under a binomial data-generating distribution is that it yields a posterior distribution that is also beta. Specifically, $\theta|x \sim \text{beta}(\alpha + x, \beta + n - x)$, for which the associated posterior density is

$$p(\theta|x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.$$

This posterior density is quickly deduced by examining the *forms* of the relevant probability mass and densities functions as Bayes's Theorem is applied:

$$p(x|\theta) \propto \theta^x(1-\theta)^{n-x} \text{ and } p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

implies

$$p(\theta|x) \propto p(x|\theta)p(\theta) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.$$

Proportional equivalences (*i.e.*, $a \propto b$ rather than $a = b$) are all that are needed in this deduction because the form of the mass or density function implies the constant factor that induces it sum or integrate to one. Note that $p(\theta|x) \propto p(x|\theta)p(\theta)$ is a restatement of Bayes's formula in terms of proportional equivalence, where only the numerator in the usual formula is needed for deduction. Numerical work with these distributions must account for both the numerator and denominator.

Example posterior summaries in the binomial-beta setup that would be reported as data-analysis results could include the posterior mean and posterior standard deviation,

$$E[\theta|x] = \frac{\alpha+x}{\alpha+\beta+n} \quad \text{and} \quad SD[\theta|x] = \sqrt{\frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}},$$

and an array of posterior quantiles, which provide interval estimates. Recall that, for an absolutely continuous random variable $X$ with cumulative distribution function $F(x)$, the $100p\%$ quantile is the value $x_p$ such that $F(x_p) = p$.

For example, if $x = 3$ incidences are observed in $n = 10$ trials, and the prior distribution is specified as $\theta \sim \text{uniform}(0, 1) = \text{beta}(1, 1)$, the posterior distribution is $\theta|x \sim \text{beta}(4, 8)$. The posterior mean and standard deviation are $E[\theta|x] = 0.3333$ and $SD[\theta|x] = 0.1307$, and a potentially helpful set of quantiles is

| $100p\%$ | 2.5% | 25% | 50% | 75% | 97.5% |
|---|---|---|---|---|---|
| quantile | 0.1093 | 0.2364 | 0.3238 | 0.4205 | 0.6097 |

These values imply that, given $x = 3$, there is a 95% probability that $\theta$ is between 0.1093 and 0.6097.

Some additional comments about the prior distribution are in order:

- A situation in which the data-generating distribution is associated with a particular class of prior distributions each of which yields a posterior distribution in the same class is called a *conjugate* setup. That particular class of prior distributions is called a *conjugate* class, relative to the form of data-generating distribution. In the present example, we see that the beta distributions form a conjugate class to the binomial data-generating distributions. Many other examples of conjugate setups are available, and they are important for identifying situations in which complicated numerical algorithms are not needed to produce simple summaries of uncertainty from the posterior distribution.

- The proposition to specify a prior distribution that would describe a state of "ignorance" about the parameter has produced a number of guiding principles and a variety of widely-applicable potential solutions. For the binomial-beta model, $\theta \sim \text{beta}(1, 1)$, $\theta \sim \text{beta}(1/2, 1/2)$, and $\theta \sim \text{beta}(0, 0)$ have all been proposed as settings of the prior that would be specified by "default" when the analyst chooses not to incorporate background information into the analysis. The latter setting, $\theta \sim \text{beta}(0, 0)$ is interesting for defining a density function that fails to integrate to one; yet, provided $x$ is neither 0 nor $n$, the corresponding posterior density does integrate to one. Such situations are said to involve an *improper prior* that nevertheless yields a *proper posterior*. The "correct" default setting for a prior distribution is elusive; the choices made largely reflect the preferences of individual data-analysts.

**Example: Bayesian analysis of normal random samples**

Suppose the data-generating distribution is defined from a random sample, $\boldsymbol{x} = (x_1, \ldots, x_n)$, of normal measurement. That is, given a common mean, $\mu$, and variance, $\sigma^2$, the $x_i$ are independent and identically distributed such that each $x_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$. This setup involves two parameters, $\mu$ and $\sigma^2$. The task of specifying a prior distribution on $\boldsymbol{\theta} = (\mu, \sigma^2)$

may be handled sequentially, by conditioning, starting by specifying a conditional prior distribution for $\mu$ given $\sigma^2$, and then specifying a marginal prior distribution for $\sigma^2$.

As for a conditional prior distribution for $\mu$, given $\sigma^2$, it turns out that when the conditional distribution is normal, the corresponding conditional posterior distribution is also normal. Specifically,

$$\mu|\sigma^2 \sim N(m_0, c_0\sigma^2) \text{ implies } \mu|\boldsymbol{x}, \sigma^2 \sim N(m, c\sigma^2),$$

where

$$m = \frac{m_0 + c_0 n\bar{x}}{1 + c_0 n} \quad \text{and} \quad c = \frac{c_0}{1 + c_0 n}.$$

As for a marginal prior distribution $\sigma^2$, consider specifying this distribution in such a way that, for specified constants $\lambda_0$ and $\kappa_0$, the transformation $\lambda_0/\sigma^2$ has a $\chi^2_{\kappa_0}$ distribution. Associated terminology would state that $\sigma^2$ has a *scaled inverse-$\chi^2$ distribution* with scale parameter $\lambda_0$ and $\kappa_0$ degrees of freedom, or, in shorthand notation, $\sigma^2 \sim \text{ScInv-}\chi^2(\lambda_0, \kappa_0)$. A density for this distribution is

$$p(\sigma^2) \quad = \quad \frac{(\lambda_0/2)^{\kappa_0/2}}{\Gamma(\kappa_0/2)} (\sigma^2)^{-(\kappa_0/2+1)} e^{-\lambda_0/(2\sigma^2)}.$$

It can be shown that

$$\sigma^2 \sim \text{ScInv-}\chi^2(\lambda_0, \kappa_0) \text{ implies } \sigma^2|\boldsymbol{x} \sim \text{ScInv-}\chi^2(\lambda, \kappa),$$

where

$$\lambda = \lambda_0 + (n-1)s^2 \quad \text{and} \quad \kappa = \kappa_0 + n - 1,$$

having written

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

for the sample variance.

Because the conditional prior and posterior distributions for $\mu$, given $\sigma^2$ are in the same family of distributions (the normal distributions), and also the marginal prior and posterior distributions for $\sigma^2$ are in the same family of distributions (the scaled inverse-$\chi^2$ distributions), we are in a conjugate setup. The above distributional results may be deduced from $p(\mu, \sigma^2|\boldsymbol{x}) \propto p(\boldsymbol{x}|\mu, \sigma^2)p(\mu, \sigma^2)$, after expanding $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$ and $p(\mu, \sigma^2|\boldsymbol{x}) = p(\mu|\boldsymbol{x}, \sigma^2)p(\sigma^2|\boldsymbol{x})$, and noting that, because $\sum_{i=1}^{n}(x_i - \mu)^2 = (n-1)s^2 + n(\bar{x} - \mu)^2$, the data-generating density factors according to

$$\begin{aligned} p(\boldsymbol{x}|\mu, \sigma^2) \quad &= \quad (2\pi\sigma^2)^{-n/2} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/(2\sigma^2)} \\ &\propto \quad \left\{(2\pi\sigma^2)^{-1/2} e^{-n(\bar{x}-\mu)^2/(2\sigma^2)}\right\} \times \left\{(\sigma^2)^{-(n-1)/2} e^{-(n-1)s^2/(2\sigma^2)}\right\}. \end{aligned}$$

This factorization connects the density to the two forms that define the normal and scaled inverse-$\chi^2$ distributions; the above formulas for $m$ and $c$ are found by completing the square.

In many circumstances, a statistical report under this model would focus on inferences about the mean, $\mu$, which presents a complication since we have only this deduced a posterior distribution for $\mu$ conditional on $\sigma^2$. The desired inference about $\mu$ is to be unconditional; hence, it would arise from the marginal posterior distribution of $\mu$, whose density is

$$p(\mu|\boldsymbol{x}) = \int p(\mu|\boldsymbol{x}, \sigma^2)p(\sigma^2|\boldsymbol{x})d\sigma^2.$$

Carrying out this integration will show that the marginal posterior distribution of $\mu$ is determined by t distribution. Specifically, given $\boldsymbol{x}$, the shift-and-scale transformation

$$t = \frac{\mu - m}{\sqrt{\lambda c/\kappa}}$$

has a t distribution with $\kappa$ degrees of freedom; *i.e.*, $t|\boldsymbol{x} \sim t_\kappa$. Thus, if $t_{\kappa,p}$ is the right-tail cutoff of a t distribution with $\kappa$ degrees of freedom, with probability $p$ in the right tail (*i.e.*, $P[T > t_{\kappa,p}] = p$, where $T \sim t_\kappa$) then it would make sense to report

$$m \pm t_{\kappa,\alpha/2}\sqrt{\lambda c/\kappa}$$

as the endpoints of a $(1-\alpha)100\%$ posterior interval for $\mu$.

Two additional comments are as follows.

- The formulas presented above for the posterior parameters $m$, $c$, $\lambda$, and $\kappa$ are functions of the prior parameters $m_0$, $c_0$, $\lambda_0$, and $\kappa_0$, and the summary statistics $\bar{x}$ and $s^2$. Examining the formulas for $m$ and $c$, asymptotically as $c_0$ increases without bound, it is seen that

  $$m \to \bar{x} \text{ and } c \to 1/n, \quad \text{as} \quad c_0 \to \infty.$$

  Moreover, the formulas for $\lambda$ and $\kappa$ are such that

  $$\lambda = (n-1)s^2 \text{ and } \kappa = n - 1, \quad \text{if} \quad \lambda_0 = \kappa_0 = 0.$$

  If follows that when $\lambda_0 = \kappa_0 = 0$ and $c_0$ is very large (*i.e.* we consider the limit as $c_0 \to \infty$), it is approximately the case that the conditional posterior distribution for $\mu$ and marginal posterior distribution for $\sigma^2$ are

  $$\mu|\boldsymbol{x}, \sigma^2 \sim N(\bar{x}, \sigma^2/n) \text{ and } \sigma^2|\boldsymbol{x} \sim \text{ScInv-}\chi^2((n-1)s^2, n-1),$$

  the relevant shift-and-scale transformation of $\mu$ is, given $\boldsymbol{x}$,

  $$\frac{\mu - \bar{x}}{s/\sqrt{n}} \sim t_{n-1}$$

and

$$\bar{x} \pm t_{n-1,\alpha/2} s/\sqrt{n}$$

are the endpoints of a $(1-\alpha)100\%$ posterior interval for $\mu$. This is remarkable for producing the same distributional and inference formulas that are associated with the classical statistical methods you are likely to have learned in your earlier classes. Accordingly, many regard the Bayesian setup implied by the settings $\lambda_0 = \kappa_0 = 0$ and $c_0 \to \infty$ to involve a prior distribution that describes ignorance. Careful attention to the density formulas given above will suggest that this prior could be defined by

$$p(\mu|\sigma^2) \propto 1 \text{ and } p(\sigma^2) \propto 1/\sigma^2.$$

each of which must be understood in terms of proportional equivalences since the density formula is improper.

- In not too much longer we will see that, in practice, it is not really necessary to know the marginal posterior distribution of $\mu$ in order to produce a statistical report such as a $(1-\alpha)100\%$ posterior interval for that parameter. This is because of the availability of numerical simulation algorithms that would allow us to generate a simulated sample from $p(\mu|\boldsymbol{x})$ using just $p(\mu|\boldsymbol{x}, \sigma^2)$ and $p(\sigma^2|\boldsymbol{x})$. The steps in this algorithm are as follows

  STEP 1: Simulate a value of $\sigma^2$ from a ScInv-$\chi^2(\lambda, \kappa)$ distribution;

  STEP 2: Given the value $\sigma^2$ simulated in STEP 1, simulate a value of $\mu$ from a $\mu|\boldsymbol{x}, \sigma^2 \sim N(m, c\sigma^2)$ distribution;

  STEP 3: Repeat STEPS 1 and 2 many times, each time storing the simulated values of $\sigma^2$ and $\mu$.

  STEP 4: The distribution of simulated, stored values of $\mu$ are representative of the marginal posterior distribution of $\mu$. Their $(\alpha/2)100\%$ and $(1-\alpha/2)100\%$ quantiles provide the endpoints to an approximate $(1-\alpha)100\%$ posterior interval for $\mu$.

  This is a simple example of a *Gibbs sampling algorithm*, where individual parameters are simulated conditionally on the other parameters, cycling through all of them at each iteration. This and additional numerical simulation strategies will be discussed further in what follows. In complex Bayesian formulations, the use of simulation algorithms for practical implementation may be more a necessity than a convenience. In general, even in simple problems, such algorithms add a great deal of flexibility to what is possible for inference *via* the Bayesian approach, as the next example demonstrates.

□

**Section 4A.3: Specialized techniques.**

The relationships between prior and posterior distributions highlighted in the last example broadly extend to more general contexts, though the mathematical expressions may start to

become complicated. This leads us to consider Bayesian analysis of multiple linear regression models. In what follows, we start by describing the basic setup of linear regression, much of which is likely to look familiar from your previous classes, then embed the setup into a Bayesian formulation that extends that of the previous example. A subsequent example showcases the flexibility of Bayesian numerical techniques.

In describing the regression setup, we apply notation from time series to help keep our discussion in context. In multiple regression, the data arise as $(q+1)$-tuples of measurements,

$$(x_1, z_{11}, \ldots, z_{1q}), \ldots, (x_n, z_{n1}, \ldots, z_{nq})$$

collected at $n$ time points. The variable $x_t$ is the *response variable* and the $z_{t1}, \ldots, z_{tq}$ are *regressor variables*. Models for regression data describe the dependency of the response variable on the regressor variables. The *multiple linear regression* model is given by

$$x_t = \beta_1 z_{t1} + \cdots + \beta_q z_{tq} + w_t,$$

where $\beta_1, \ldots, \beta_q$ are *regression coefficients* and $(w_t)$ is Gaussian white noise with *residual variance* $\sigma_w^2$. Sometimes $(w_t)$ is called the *residual error* series.

It is convenient to describe the components of the regression model using matrix-vector notation: The $t$'th *regression vector* collects the $q$ regressor measurements at time $t$ into a column vector.

$$\boldsymbol{z}_t = [z_{t1}, \ldots, z_{tq}]^T.$$

which allows the data to be succinctly written $(x_1, \boldsymbol{z}_1), \ldots, (x_n, \boldsymbol{z}_n)$. The *regression matrix* is the $n \times q$ such that the $t$'th row is the transpose of the $t$'th regression vector:

$$\boldsymbol{Z} = [\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T]^T = \begin{bmatrix} z_{11} & \cdots & z_{1q} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nq} \end{bmatrix}.$$

The *response vector* is the $n \times 1$ column vector with the $t$'th response measurement filling its $t$'th entry,

$$\boldsymbol{x} = [x_1, \ldots, x_n]^T.$$

Similarly, the *residual error* and *regression coefficient vectors* are the respective $n \times 1$ and $q \times 1$ vectors

$$\boldsymbol{w} = [w_1, \ldots, w_n]^T \quad \text{and} \quad \boldsymbol{\beta} = [\beta_1, \ldots, \beta_q]^T.$$

Using this notation, the regression model is

$$\boldsymbol{x} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{w}.$$

In a Bayesian formulation, the regression model is expressed to reflect the Gaussian distributional assumption on the residual errors and to define the data-generating distribution

$$\boldsymbol{x} | \boldsymbol{\beta}, \sigma_w^2 \sim N(\boldsymbol{Z}\boldsymbol{\beta}, \sigma_w^2 \boldsymbol{I}),$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Extending the distributional relationships we discussed for Bayesian analysis of normal random samples, it is possible to specify a prior distribution on $\boldsymbol{\beta}$ and $\sigma_w^2$ such that the posterior distribution would reflect formulas from classical least-squares regression. The associated prior density is improper, and is specified through a conditional-marginal factorization according to

$$p(\boldsymbol{\beta}|\sigma_w^2) \propto 1 \quad \text{and} \quad p(\sigma_w^2) \propto 1/\sigma_w^2.$$

Observe the resemblance of these prior densities to the ignorance-prior discussed in the last example. Incidentally, a rich conjugate setup, which allows specification of informative prior distributions, is also available for use in linear regression, derived from the normal and scaled inverse-$\chi^2$ distributions; however, to simplify matters, let us restrict our focus to the non-informative setting above.

The posterior distribution can be described by some of the calculations that arise in least-squares analysis. The vector of *fitted regression coefficients*, $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \ldots, \hat{\beta}_q]^T$, is calculated as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{x}.$$

The *sample residual variance* is

$$s_w^2 = \frac{1}{n-q}\sum_{t=1}^{n}(x_t - \boldsymbol{z}_t^T\hat{\boldsymbol{\beta}})^2.$$

With these statistics defined, the conditional and marginal posterior distributions can be expressed as

$$\boldsymbol{\beta}|\boldsymbol{x}, \sigma_w^2 \sim N(\hat{\boldsymbol{\beta}}, \sigma_w^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}) \quad \text{and} \quad \sigma_w^2|\boldsymbol{x} \sim \text{ScInv-}\chi^2((n-q)s_w^2, n-q).$$

It may furthermore be deduced that, in the marginal posterior distribution of $\boldsymbol{\beta}$, each individual fitted regression coefficient may be understood in terms of a t distribution: given $\boldsymbol{x}$, each

$$\frac{\beta_i - \hat{\beta}_i}{s_w^2\sqrt{[(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]_{ii}}} \sim t_{n-q}$$

where $[(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]_{ii}$ is the $i$'th diagonal entry of $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$. Reflecting a classical interval-estimation formula from least-squares regression, the marginal posterior distribution suggests that

$$\hat{\beta}_i \pm t_{n-q,\alpha/2}s_w^2\sqrt{[(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]_{ii}}$$

are the endpoints of a $(1-\alpha)100\%$ posterior interval for $\mu$.

A more general description of the marginal posterior distribution is that, given $\boldsymbol{x}$, $\boldsymbol{\beta}$ has a multivariate t distribution with location parameter $\hat{\boldsymbol{\beta}}$ and covariance parameter $s_w^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$.

We leave this type of multivariate distribution without a precise definition, instead noting that the possibility of numerical simulation makes it unnecessary to have an explicit distributional expression for the marginal posterior distribution.

Numerical calculation within the Bayesian regression setup is demonstrated in the following example.

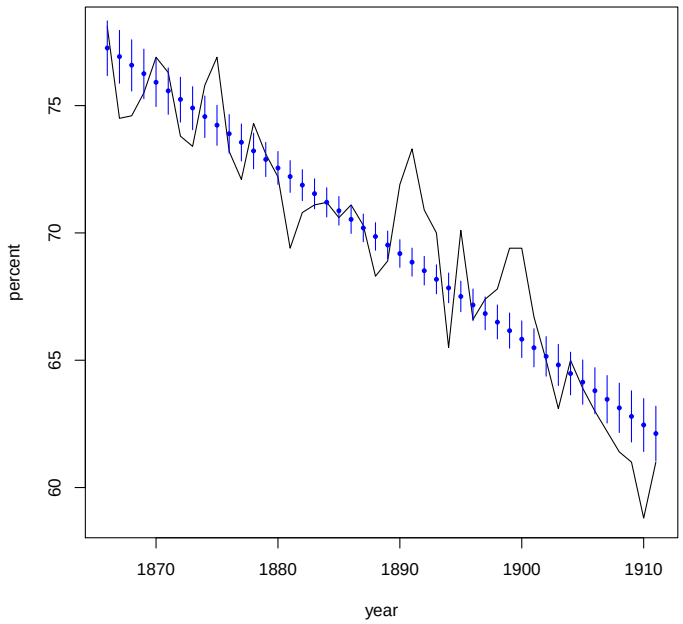
**Example: Bayesian regression**

In time series analysis, a simple regression model might arise when attempting to "de-trend" a time series $\boldsymbol{x} = (x_1, \ldots, x_n)$. In this setup, the response variable is the time series measurements $x_t$ are the response variable, and time points define a single regressor variable, $z_t = t$. The model is

$$x_t = \beta_1 + \beta_2 t + w_t,$$

where $\beta_1$ and $\beta_2$ are regression parameters and $(w_t)$ is Gaussian white noise with variance parameter $\sigma_w^2$. As we know, in time series analysis the $x_t$ are likely to be dependent; this implies that the residual error series $(w_t)$ would also be dependent. An advantage of the Bayesian approach is that it offers a satisfying way to *check* the assumption that $(w_t)$ is white noise. This is explored in comments that appear below, after the main example.

The "de-trending" technique is now illustrated on the "Marriages in Church of England" data from learning unit 1. Recall that the data exhibit a negative trend over time, as is seen in the following time plot:




The line the would be drawn through the larger dots at the center of the vertical lines data has the formula

$$\hat{x}_t = 704.78 - 0.3363t,$$

where $t$ is time. The two coefficients are determined by least-squares analysis; that is, $\hat{\beta}_1 = 704.78$ and $\hat{\beta}_2 = -0.3363$. Also shown, intersecting vertically with this line are 2.5% and 97.5% quantiles of the posterior time-specific mean values, $\mu_t = \boldsymbol{z}_t^T \boldsymbol{\beta}$, at each time point $t$. These offer a sense of the magnitude of posterior uncertainty reflected in the marginal posterior distribution of $\sigma_w^2$. □

How are the quantiles of $\mu_t$ calculated? Numerical simulation offers a relatively straightforward approach.

Following the Gibbs sampling strategy highlighted in the last example, an iterative algorithm is applied wherein at each iteration the following steps are taken:

STEP 1: Simulate a value $\sigma_w^2$ from a ScInv-$\chi^2((n-q)s_w^2, n-q)$ distribution;

STEP 2: Given the value $\sigma_w^2$ simulated in STEP 1, simulate a value of $\boldsymbol{\beta}$ from a $N(\hat{\boldsymbol{\beta}}, \sigma_w^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1})$ distribution;

STEP 3: Using the simulated $\boldsymbol{\beta}$ from STEP 2, calculate the time-specific mean values, $\mu_t = \boldsymbol{z}_t^T\boldsymbol{\beta}$ across $t = 1, \ldots, n$, and store them for later.

After many iterations, the posterior quantiles displayed in the plot above are produced, at time $t$, as the corresponding quantiles of the stored simulated $\mu_t$-values.

As a quick comment, simulating $\boldsymbol{\beta}$ in STEP 2 requires an amount of careful thought about technical aspects. It is typical that a statistical software package would include built-in functionality for simulating independent normal random variables, but unlikely that it would provide built-in functionality for simulating from a general multivariate-normal distribution. If the software is reasonably sophisticated in its handling of matrix manipulations, then a potential approach to simulating $\boldsymbol{\beta}$ is as follows. As an initial step, use the software's functionality to obtain the *matrix square-root* of $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$: For a given symmetric, positive-definite matrix, $\boldsymbol{A}$, its matrix square-root is the matrix $\boldsymbol{A}^{1/2}$ such that $\boldsymbol{A} = \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}$. The matrix square-root of $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$ would be denoted $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1/2}$. Now, to generate $\boldsymbol{\beta}$, first simulate a $q \times 1$ vector, $\boldsymbol{u} = [u_1, \ldots, u_q]^T$ of independent and identically distributed standard-normal random variables, $u_i \sim N(0,1)$. The desired simulated vector is

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} + \sigma_w(\boldsymbol{Z}^T\boldsymbol{Z})^{-1/2}.$$

Other strategies for simulating $\boldsymbol{\beta}$ are available. Among the more widely discussed are those that make use of something called a QR decomposition of $\boldsymbol{Z}^T\boldsymbol{Z}$, and related approaches that work with that matrix's eigenvalue-eigenvector decomposition.

Bayesian numerical computation makes possible the following approach to checking the assumption of independent errors. Consider adding to each iteration the step of simulating a brand new set of time-series data from the data-generating distribution

STEP 4: Using the simulated $\sigma_w^2$ and $\boldsymbol{\beta}$ from previous steps, simulate a value $\boldsymbol{x}^*$ from $N(\boldsymbol{Z}\boldsymbol{\beta}, \sigma_w^2\boldsymbol{I})$ distribution.

The simulated data, $\boldsymbol{x}^* = [x_1^*, \ldots, x_n^*]^T$, is often called a *prediction*, although in the present context our goal is not to forecast new measurements; we will discuss forecasting later. Instead, $\boldsymbol{x}^*$ is a simulated value reflective of all the assumptions of the data-generating model (such as independent errors).

An particular use for $\boldsymbol{x}^*$ is to compare its patterns with those of the data at hand, $\boldsymbol{x}$, focusing on some characteristic that is of specific interest. In the present analysis, the characteristic of interest is autocorrelation. Accordingly, we add the following step at each iteration

STEP 5: Using the previously-simulated $\boldsymbol{\beta}$ and $\boldsymbol{x}^*$, calculate the observed and predicted residual errors,

$$w_t = x_t - \boldsymbol{z}_t^T\boldsymbol{\beta} \text{ and } w_t^* = x_t^* - \boldsymbol{z}_t^T\boldsymbol{\beta}.$$

From each series of residual errors, calculate the predicted autocorrelations

$$\tilde{\rho}(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}\frac{w_{t+h}w_t}{\sigma_w^2} \quad \text{and} \quad \tilde{\rho}^*(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}\frac{w_{t+h}^*w_t^*}{\sigma_w^2}$$

Compare the magnitudes of autocorrelations at each lag-value $h$, noting whether $|\tilde{\rho}(h)| > |\tilde{\rho}^*(h)|$. Alternatively, the same comparison could be made using a statistic that aggregates autocorrelations, for which a variety of choices are available. One option is the maximum absolute autocorrelation up to lag $H$, whose predicted values are

$$S(\boldsymbol{x}, \beta) = \max_{h=1,\ldots,H}|\tilde{\rho}(h)| \quad \text{and} \quad S(\boldsymbol{x}^*, \beta) = \max_{h=1,\ldots,n-1}|\tilde{\rho}^*(h)|.$$

Another is the sum of absolute autocorrelation up to lag $H$, whose predicted values are

$$S(\boldsymbol{x}, \beta) = \sum_{h=1}^{H}|\tilde{\rho}(h)| \quad \text{and} \quad S(\boldsymbol{x}^*, \beta) = \sum_{h=1}^{H}|\tilde{\rho}^*(h)|.$$

A third choice is the weighted sum of absolute autocorrelation up to lag $H$, whose predicted values are

$$S(\boldsymbol{x}, \beta) = \sum_{h=1}^{H}a_h|\tilde{\rho}(h)| \quad \text{and} \quad S(\boldsymbol{x}^*, \beta) = \sum_{h=1}^{H}a_h|\tilde{\rho}^*(h)|,$$

where $a_1, \ldots a_H$ is a sequence of positive weights, which is typically decreasing; *i.e.*, $a_1 > a_2 > \cdots > a_H$. Still other choices are possible by replacing absolute deviation ($|\rho|$) with the squaring operation ($\rho^2$). Moreover, it may be helpful to examine any of these aggregated statistics at multiple values of the maximum lag, $H$. For any definition that is adopted, one is to note whether $S(\boldsymbol{x}, \beta) > S(\boldsymbol{x}^*, \beta)$.

The underlying idea, here, is that too-frequent incidences of the magnitude of autocorrelation in $\boldsymbol{x}$ exceeding that in $\boldsymbol{x}^*$ would indicate a deficiency of the postulated model with respect to the assumption of independent errors. It is assessed by calculating a *posterior predictive p-value*, which is the relative frequency of observing the problematic incidence across all iterations of the simulation.

These relative frequencies for the example analysis of the "Marriages in Church of England" data are tabulated as follows.

| stat. | $\lvert\tilde{\rho}(1)\rvert$ | $\lvert\tilde{\rho}(2)\rvert$ | $\lvert\tilde{\rho}(3)\rvert$ | $\lvert\tilde{\rho}(4)\rvert$ | $\lvert\tilde{\rho}(5)\rvert$ | $\max_h \lvert\tilde{\rho}(h)\rvert$ | $\sum_{h=1}^{H} \lvert\tilde{\rho}(h)\rvert$ |
|---|---|---|---|---|---|---|---|
| ppp-val | 0.0118 | 0.4131 | 0.7488 | 0.2050 | 0.2216 | 0.0563 | 0.1052 |

Noting that several posterior predictive p-values take values close to zero, it is evident from these results that the model is deficient for not accounting for patterns of autocorrelation in the residual errors. This is particularly evident in the p-value associated with the lag-1 autocorrelation, which is defined by the statistic $|\tilde{\rho}(1)|$. The posterior predictive p-values arising from the aggregated statistics are also small, with that defined from the maximum absolute autocorrelation, $\max_h |\tilde{\rho}(h)|$, notably so. However, neither of these values are not extraordinarily small, and may simply reflect the unaccounted for lag-1 autocorrelation that has already been detected, rather than some deficiency reflective of combined autocorrelation across multiple lags. In other words, the posterior predictive p-values associated with the aggregated statistics do not add any new insights in this case.

In general, the posterior predictive p-value is described as follows. It is always defined relative to a statistic, $S(\boldsymbol{x}, \boldsymbol{\theta})$, that would summarize the data-characteristic of interest. As indicated in the notation, the summary statistic may depend on the parameter as well as the data. For example, the version of autocovariance used in the example depends on the parameter, as is made clear by expanding

$$\tilde{\rho}(h) \quad = \quad \frac{1}{n-h}\sum_{t=1}^{n-h} w_{t+h}w_t = \frac{1}{n-h}\sum_{t=1}^{n-h}(x_{t+h} - \boldsymbol{z}_{t+h}^T\boldsymbol{\beta})(x_t - \boldsymbol{z}_t^T\boldsymbol{\beta}),$$

in which the rightmost expression give the explicit dependency on $\boldsymbol{\beta}$. In some situations, it would make sense to define the statistic independently of the parameter, $S(\boldsymbol{x}, \boldsymbol{\theta}) = S(\boldsymbol{x})$. For instance, in the example, it would have been perfectly fine to have replaced $\tilde{\rho}(h)$ with sample autocovariance,

$$\hat{\rho}(h) \quad = \quad \frac{1}{n-h}\sum_{t=1}^{n-h}(w_{t+h} - \bar{w})(w_t - \bar{w}),$$

but it made sense to fill in $\boldsymbol{z}_t^T\boldsymbol{\beta}$ for $\bar{w}$, since that is the quantity the latter is meant to estimate. Once a suitable statistic, $S(\boldsymbol{x}, \boldsymbol{\theta})$, is defined, the posterior predictive p-value is the quantity

$$P[S(\boldsymbol{x}, \boldsymbol{\theta}) > S(\boldsymbol{x}^*, \boldsymbol{\theta})|\boldsymbol{x}] = \int\int I[S(\boldsymbol{x}, \boldsymbol{\theta}) > S(\boldsymbol{x}^*, \boldsymbol{\theta})]p(\boldsymbol{x}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{x}^*d\theta,$$

where $I(\cdot)$ takes the value one if the condition is true and zero otherwise.