STAT 153 - Introduction to Time Series Lecture Three

Fall 2022, UC Berkeley

Aditya Guntuboyina

September 2, 2022

1 Last Class

In the last class, we studied the analysis of the simple trend model:

$$Y_i = \beta_0 + \beta_1 t_i + Z_i \tag{1}$$

where $Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. The goal is to infer the three parameters β_0, β_1, σ given observed data $(t_1, y_1), \dots, (t_n, y_n)$. Of these three parameters, the main parameter is β_0, β_1 while σ can be considered as a nuisance parameter.

A key role in this inference is played by the least squares criterion:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 t_i)^2 .$$

The point estimates for β_0, β_1 are given by the minimizers of $S(\beta_0, \beta_1)$:

$$(\hat{\beta}_0, \hat{\beta}_1) := \underset{\beta_0, \beta_1}{\text{argmin}} \ S(\beta_0, \beta_1).$$

In the above argmin refers to minimizer. We saw that

$$\hat{\beta} := \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \left(X^T X\right)^{-1} X^T Y$$

where

$$Y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & t_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & t_n \end{pmatrix}.$$

We also studied the Bayesian analysis of this model which allows for full uncertainty quantification. We worked with the prior:

$$\beta_0, \beta_1, \log \sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-C, C)$$

for a large constant C. Under this prior, the joint posterior for (β_0, β_1) came out to be:

$$f_{\beta_0, \beta_1 | \text{data}}(\beta_0, \beta_1) \propto \left(\frac{1}{S(\beta_0, \beta_1)}\right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \tag{2}$$

2 More on the Bayesian Posterior

In most regression problems, the least squares criterion $S(\beta_0, \beta_1)$ will take large values (for example, in the US population dataset, the smallest possible value of $S(\beta_0, \beta_1)$ is of the order of billions). This would mean that $\left(\frac{1}{S(\beta_0,\beta_1)}\right)^{n/2}$ would be very small for all values of β_0, β_1 (of course, the normalizing constant in front of (2) would then have to be quite large). In order to not deal with such small values, it makes sense to rewrite the posterior density as:

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0, \beta_1) \propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)}\right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \tag{3}$$

Note that (2) and (3) represent exactly the same density because the term $(S(\hat{\beta}_0, \hat{\beta}_1))^{n/2}$ does not depend on β_0, β_1 and is thus a constant.

Generally, the density (3) will be quite sharply concentrated around the least squares estimator $(\hat{\beta}_0, \hat{\beta}_1)$ especially when n is large. This is because, when (β_0, β_1) is such that $S(\beta_0, \beta_1)$ is large compared to $S(\hat{\beta}_0, \hat{\beta}_1)$, the quantity

$$\left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)}\right)^{n/2}$$

would be quite negligible because of the large power $n/2$. As a result, the posterior density $f_{\beta_0,\beta_1|\text{data}}(\beta_0, \beta_1)$ will be concentrated around those values of (β_0, β_1) for which $S(\beta_0, \beta_1)$ is quite close to $S(\hat{\beta}_0, \hat{\beta}_1)$. For example, suppose $n = 762$ (as in the US population dataset), and that (β_0, β_1) is such that $S(\beta_0, \beta_1) = (1.1)S(\hat{\beta}_0, \hat{\beta}_1)$. Then

$$\left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)}\right)^{n/2} = \left(\frac{1}{1.1}\right)^{381} \approx 1.7 \times 10^{-16}.$$

Such (β_0, β_1) will thus get negligible posterior probability. Even for (β_0, β_1) such that $S(\beta_0, \beta_1) = (1.1)S(\hat{\beta}_0, \hat{\beta}_1)$, we have

$$\left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)}\right)^{n/2} = \left(\frac{1}{1.01}\right)^{381} \approx 0.02$$

and so such (β_0, β_1) will also get fairly small posterior probability.

To sum up, when n is large, the posterior probability will be concentrated around those (β_0, β_1) for which $S(\beta_0, \beta_1)$ is very close to $S(\hat{\beta}_0, \hat{\beta}_1)$. Generally, this would imply that (β_0, β_1) would itself have to be close to $(\hat{\beta}_0, \hat{\beta}_1)$. For this reason, the indicator term in (3) has no effect when C is large. From now on, we shall drop this indicator term and refer to the Bayesian posterior as simply

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0, \beta_1) \propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)}\right)^{n/2}. \tag{4}$$

A more precise understanding of the posterior density can be obtained by noting its connection to the multivariate t-density. Before looking at this connection, let us briefly recall t-densities.

3 t-densities

We first look at the univariate case.

3.1 Univariate t-density

The t-density is obtained by changing the scale of a normally distributed random variable through an independent chi-squared distributed random variable. More precisely, suppose X has the $N(\mu, \sigma^2)$ distribution. First write

$$X = \mu + (X - \mu).$$

Now consider an independent random variable V such that

$$V \sim \chi^2_v.$$

Recall that χ^2_v is the same as the Gamma$(v/2, 1/2)$ distribution so that

$$f_V(x) \propto x^{\frac{v}{2}-1} e^{-x/2} I\{x > 0\}.$$

We now change the scale of X using V to create a new random variable T by

$$T := \mu + \frac{X - \mu}{\sqrt{\frac{V}{v}}}. \tag{5}$$

The distribution of T will be denoted by $t_v(\mu.\sigma^2)$ (here v is known as the degrees of freedom). The density of T can be derived as follows:

$$f_T(y) = \int_0^\infty f_{T|V=x}(y) f_V(x) dx.$$

Observe now that

$$T \mid V = x \quad = \mu + \frac{X - \mu}{\sqrt{\frac{x}{v}}} \sim N\left(\mu, \sigma^2 \frac{v}{x}\right)$$

so that

$$f_{T|V=x}(y) = \frac{\sqrt{x}}{\sqrt{2\pi}\sigma\sqrt{v}} \exp\left(-\frac{x}{2\sigma^2 v}(y - \mu)^2\right).$$

As a result

$$\begin{aligned} f_T(y) &= \int_0^\infty f_{T|V=x}(y) f_V(x) dx \\ &\propto \int_0^\infty \frac{\sqrt{x}}{\sqrt{2\pi}\sigma\sqrt{v}} \exp\left(-\frac{x}{2\sigma^2 v}(y - \mu)^2\right) x^{\frac{v}{2}-1} e^{-x/2} dx \\ &\propto \int_0^\infty x^{\frac{v}{2}-\frac{1}{2}} \exp\left(-\frac{x}{2}\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)\right) dx. \end{aligned}$$

The change of variable

$$t = x\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)$$

now leads to

$$f_T(y) \propto \frac{1}{\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)^{\frac{v+1}{2}}} \int_0^\infty t^{\frac{v}{2}-1} e^{-t/2} dt \propto \frac{1}{\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)^{\frac{v+1}{2}}}.$$

Therefore the density corresponding to the $t_v(\mu, \sigma^2)$ distribution is proportional to

$$y \mapsto \frac{1}{\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)^{\frac{v+1}{2}}}.$$

It is useful to note that when the degrees of freedom v is large, the distribution $t_v(\mu, \sigma^2)$ is very close to the normal distribution $N(\mu, \sigma^2)$. There are many ways of seeing this. One way is to note that the mean and variance of $V \sim \chi^2_v$ are given by v and $2v$ respectively. This implies that

$$\mathbb{E}\left(\frac{V}{v}\right) = 1 \quad \text{and} \quad \text{var}\left(\frac{V}{v}\right) = \frac{2v}{v^2} = \frac{2}{v}.$$

Thus when v is large, the random variable $\frac{V}{v}$ has mean 1 and very small variance so that $\frac{V}{v}$ will be very close to 1 with very high probability. As a result, the scale change by $\sqrt{V/v}$ in (5) has little effect so that T will have the same distribution as $X \sim N(\mu, \sigma^2)$.

3.2 Multivariate t-density

The multivariate t-density is obtained by changing the scale of a **multivariate** normal density. Let X have the p-variate normal distribution $N_p(\mu, \Sigma)$. This means that X is a $p \times 1$ random vector, μ is a $p \times 1$ vector, Σ is a $p \times p$ (positive-definite) matrix and the density of X is equal to

$$x \mapsto \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

Just as in the one-dimensional case, we define

$$T := \mu + \frac{X - \mu}{\sqrt{\frac{V}{v}}}.$$

Here $V \sim \chi^2_v$ and is independent of X (exactly as in the one-dimensional case). Note that X and T are both $p \times 1$ random vectors while V is a scalar. In other words, T is given by

$$\begin{pmatrix} T_1 \\ \cdot \\ \cdot \\ \cdot \\ T_d \end{pmatrix} = \begin{pmatrix} \mu_1 + \frac{X_1 - \mu_1}{\sqrt{\frac{V}{v}}} \\ \cdot \\ \cdot \\ \cdot \\ \mu_d + \frac{X_d - \mu_d}{\sqrt{\frac{V}{v}}} \end{pmatrix}. \tag{6}$$

Note specifically that the scale change on each component through the same random variable V.

The distribution of this random vector T will be denoted by $t_{v,p}(\mu, \Sigma)$. Its density can be derived just as in the univariate case in the following way:

$$f_T(y) = \int_0^\infty f_{T|V=x}(y) f_V(x) dx.$$

Observe that

$$T \mid V = x \sim N\left(\mu, \frac{v}{x}\Sigma\right)$$

so that

$$f_{T|V=x}(y) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\frac{v}{x}\Sigma)}} \exp\left[-\frac{1}{2}(y-\mu)^T \left(\frac{v}{x}\Sigma\right)^{-1}(y-\mu)\right]$$
$$= \frac{x^{p/2}}{(2\pi)^{p/2}v^{p/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{x}{2v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)$$

where we used $\det(\frac{v}{x}\Sigma) = (v/x)^p\det(\Sigma)$. As a result

$$f_T(y) = \int_0^\infty f_{T|V=x}(y)f_V(x)dx$$
$$\propto \int_0^\infty \frac{x^{p/2}}{(2\pi)^{p/2}v^{p/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{x}{2v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) x^{\frac{v}{2}-1}e^{-x/2}dx$$
$$\propto \int_0^\infty x^{\frac{p+v}{2}-1} \exp\left(-\frac{x}{2}\left[1+\frac{1}{v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]\right) dx.$$

The change of variable

$$t = x\left[1+\frac{1}{v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]$$

leads to

$$f_T(y) \propto \frac{1}{\left[1+\frac{1}{v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]^{\frac{v+p}{2}}} \int_0^\infty t^{\frac{v+p}{2}-1}e^{-t/2}dt$$
$$\propto \frac{1}{\left[1+\frac{1}{v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]^{\frac{v+p}{2}}}.$$

Therefore the density corresponding to $t_{v,p}(\mu,\Sigma)$ distribution is proportional to

$$y \mapsto \frac{1}{\left[1+\frac{1}{v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]^{\frac{v+p}{2}}}. \tag{7}$$

Note that, in the notation $t_{v,p}(\mu,\Sigma)$, v denotes degrees of freedom, p denotes dimension, μ and Σ denote the mean vector and covariance matrix of the corresponding normal random vector X.

As in the univariate case, when v is large, $t_{v,p}(\mu,\Sigma)$ is very close to $N_p(\mu,\Sigma)$.

The following fact will be useful in the sequel.

Fact 3.1. *If $T \sim t_{v,p}(\mu,\Sigma)$ has components $T_1,\dots,T_p$, then, for each $j = 1,\dots,p$,*

$$T_j \sim t_v(\mu_j, \Sigma_{jj})$$

where μ_j is the j^{th} component of μ and Σ_{jj} is the $(j,j)^{th}$ entry of Σ.

This fact follows directly from (6) (and the univariate definition of the t-density) because

$$T_j = \mu_j + \frac{X_j - \mu_j}{\sqrt{\frac{V}{v}}}$$

and $X_j \sim N(\mu_j, \Sigma_{jj})$.

4 Back to the Bayesian Posterior in Linear Regression

We now show that the Bayesian posterior (4) in the linear trend model (1) equals $t_{v,p}(\mu, \Sigma)$ for some $v, p = 2, \mu$ and Σ. For this, first note that (below $\beta := \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}$ and $\hat{\beta} := \begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{pmatrix}$)

$$S(\beta) = \|Y - X\beta\|^2 = \|Y - X\hat{\beta}\|^2 + \|X\beta - X\hat{\beta}\|^2 = S(\hat{\beta}) + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}).$$

We can thus write

$$f_{\beta|\text{data}}(\beta) \propto \left(\frac{S(\hat{\beta})}{S(\beta)}\right)^{n/2} = \left(\frac{S(\hat{\beta})}{S(\hat{\beta}) + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})}\right)^{n/2} = \left(\frac{1}{1 + (\beta - \hat{\beta})^T \left(\frac{X^T X}{S(\hat{\beta})}\right)(\beta - \hat{\beta})}\right)^{n/2}.$$

Comparing the above to (7) with $p = 2$, we can easily deduce that this is the multivariate t-density $t_{v,2}(\mu, \Sigma)$ with $v = n - 1$, $\mu = \hat{\beta}$ and

$$\Sigma^{-1} = \frac{n-2}{S(\hat{\beta})}(X^T X) \quad \text{so that } \Sigma = \frac{S(\hat{\beta})}{n-2}(X^T X)^{-1}.$$

It is customary in linear regression to use the notation

$$\hat{\sigma} := \sqrt{\frac{S(\hat{\beta})}{n-2}}$$

so that

$$\Sigma = \hat{\sigma}^2 (X^T X)^{-1}.$$

We have thus proved that

$$\beta \mid \text{data} \sim t_{n-2,2}(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1}).$$

With this posterior density, one can do uncertainty quantification about the parameters β_0 and β_1. One can generate multiple samples from $t_{n-2,2}(\hat{\beta}, \hat{\sigma}^2(X^T X)^{-1})$ and plot the resulting lines to visualize the uncertainty in β_0 and β_1. One can also use Fact 3.1 to deduce that

$$\beta_0 \mid \text{data} \sim t_{n-2}(\hat{\beta}_0, \hat{\sigma}^2 (X^T X)^{11}) \quad \text{and} \quad \beta_1 \mid \text{data} \sim t_{n-2}(\hat{\beta}_1, \hat{\sigma}^2 (X^T X)^{22})$$

where $(X^T X)^{11}$ and $(X^T X)^{22}$ are the first and second diagonal entries of $(X^T X)^{-1}$ respectively. These univariate t-densities describe the marginal uncertainty in the intercept and slope parameters. When n is large, these will be close to the normal distributions $N(\hat{\beta}_0, \hat{\sigma}^2 (X^T X)^{11})$ and $N(\hat{\beta}_1, \hat{\sigma}^2 (X^T X)^{22})$ respectively. The quantities $\hat{\sigma}\sqrt{(X^T X)^{11}}$ and $\hat{\sigma}\sqrt{(X^T X)^{22}}$ are known as the standard errors of the intercept and the slope respectively.

5 More General Trend Models

This linear regression based methodology can be used to fit more complicated trend models as well. For example, one can fit the quadratic trend model

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + Z_i \tag{8}$$

with $Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ by the same methodology with

$$X = \begin{pmatrix} 1 & t_1 & t_1^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & t_n & t_n^2 \end{pmatrix}.$$

The posterior density of $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$ will then be given by $t_{n-3,3}(\hat{\beta}, \hat{\sigma}^2(X^TX)^{-1})$ (note that the dimension now is 3 and the degrees of freedom is $n-3$).

A more general polynomial trend model (of degree k) can be fit analogously (the dimension of β will then be $k+1$ and the degrees of freedom of the posterior t-density will be $n-k-1$).

To capture seasonal trend with known period s (for example $s = 12$ in monthly data), one can use a model of the form

$$Y_i = \beta_0 + \sum_{f=1}^{r} \left(\beta_{f1} \cos \frac{2\pi f t_i}{s} + \beta_{f2} \sin \frac{2\pi f t_i}{s} \right) + Z_i.$$

This is also a linear regression model with

$$\beta = (\beta_0, \beta_{11}, \beta_{12}, \beta_{21}\beta_{22}, \dots, \beta_{r1}, \beta_{r2})^T$$

and the i^{th} row of the $n \times (2r+1)$ matrix X is given by

$$\left(1, \cos \frac{2\pi t_i}{s}, \sin \frac{2\pi t_i}{s}, \cos \frac{2\pi(2)t_i}{s}, \sin \frac{2\pi(2)t_i}{s}, \dots, \cos \frac{2\pi(r)t_i}{s}, \sin \frac{2\pi(r)t_i}{s} \right).$$

Here the posterior t-density of β will have dimension $2r+1$ and degrees of freedom $n-(2r+1)$.

Time series datasets often have both trend and seasonality. These effects can be estimated by models of the form:

$$Y_i = \sum_{j=0}^{k} \beta_j^{(1)} t_i^j + \sum_{f=1}^{r} \left(\beta_{f1} \cos \frac{2\pi f t_i}{s} + \beta_{f2} \sin \frac{2\pi f t_i}{s} \right) + Z_i.$$

Inference for this model can also be done through linear regression. The degrees of freedom for the posterior t-density of the coefficients will now be $n - (2r + k + 1)$. Our methodology will work as long as $n > 2r + k + 1$.