VE414 Lecture 21

Jing Liu

UM-SJTU Joint Institute

November 19, 2019

- So far we have largely used data to only estimate unobservable,

$$Y$$

- Linear regression model is a way to study the relationship of an observable

$$Y$$

in terms of a set of other observable variables

$$X_1, X_2, \ldots X_k$$

specifically, it is a type of smoothly changing model for

$$f_{Y|\{X_1, X_2, \ldots\}}$$

in which the conditional expectation $\mathbb{E}\left[Y \mid \{X_1, \ldots X_k\}\right]$ has a form that is linear in a set of unobservable β_i, which are often known as the parameters

$$\mathbb{E}\left[Y \mid \{X_1, \ldots X_k\}\right] = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}$$

- In addition to being linear,

$$\mathbb{E}\left[Y \mid \{X_1, \ldots, X_k\}\right] = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$$

the variability around the mean, i.e. the error,

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i$$

is often assumed to be normal

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{Normal}\left(0, \sigma^2\right)$$

- Under the above specification, we have the following density function

$$\begin{aligned} f_{\{Y_1, Y_2, \cdots, Y_n\} \mid \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2\}} &= \prod_{i=1}^{n} f_{Y_i \mid \{\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2\}} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right) \end{aligned}$$

- We can put the density function into a vector form,

$$\begin{aligned} f_{\{Y_1,Y_2,\cdots,Y_n\}|\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n,\boldsymbol{\beta},\sigma^2\}} &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\mathsf{RSS}\right) \end{aligned}$$

where residual sum of squares is given by

$$\mathsf{RSS} = \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)$$

- Thus our model in vector form is $\mathbf{Y} \mid \{\mathbf{X}, \boldsymbol{\beta}, \sigma^2\} \sim \mathrm{Normal}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$.

Q: What would frequentists do next?

- Frequentists would maximise the likelihood by treating the density function as a function of the unknown parameters, which is equivalent to minimise

$$\mathsf{RSS}\left(\mathbf{b}\right) = \left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)$$

- Recall to minimise a function,

$$\mathsf{RSS}(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

we set the gradient to zero,

$$\nabla \mathsf{RSS} = 0 - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

Setting this to zero, we have

$$\hat{\boldsymbol{\beta}}_{\mathsf{MLE}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- Hence, the fitted value is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{P}\mathbf{y}$$

and the residual can be found using

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P})\,\mathbf{y}$$

- With more linear algebra, we have

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}\right) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{e}$$

which means it is unbiased as expected,

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbb{E}\left[\boldsymbol{\varepsilon} \mid \mathbf{X}\right] = \boldsymbol{\beta}$$

- The variance is given by

$$\begin{aligned}\mathrm{Var}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{Var}\left[\boldsymbol{\varepsilon} \mid \mathbf{X}\right]\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\end{aligned}$$

- With the normal assumption, we see

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$$

- To estimate σ^2, frequentists typically use the following

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\hat{\mathbf{e}}^{\mathrm{T}}\hat{\mathbf{e}} \qquad \text{where} \quad \hat{\mathbf{e}} = (\mathbf{I}-\mathbf{P})\,\mathbf{y}$$

which is unbiased as well as being consistent.

- It can be shown the residual

$$\hat{\mathbf{e}} = (\mathbf{I}-\mathbf{P})\,\mathbf{y} = (\mathbf{I}-\mathbf{P})\,(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})$$

is an unbiased and consistent estimator of the error $\mathbf{e}$, and the variance is

$$\begin{aligned}
\operatorname{Var}[\hat{\mathbf{e}} \mid \mathbf{X}] &= \operatorname{Var}[(\mathbf{I}-\mathbf{P})\,(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \mid \mathbf{X}] \\
&= (\mathbf{I}-\mathbf{P})\operatorname{Var}[\boldsymbol{\varepsilon} \mid \mathbf{X}]\,(\mathbf{I}-\mathbf{P})^{\mathrm{T}} \\
&= (\mathbf{I}-\mathbf{P})\,\sigma^2\mathbf{I}\,(\mathbf{I}-\mathbf{P})^{\mathrm{T}} = \sigma^2\,(\mathbf{I}-\mathbf{P})
\end{aligned}$$

- Thus with the normal assumption, we have

$$\hat{\mathbf{e}} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\,(\mathbf{I}-\mathbf{P})\right)$$

Q: How would Bayesian approach the same problem?

$$f_{\mathbf{Y}|\{\mathbf{X},\boldsymbol{\beta},\sigma^2\}} = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\mathsf{RSS}\left(\boldsymbol{\beta}\right)\right)$$

where

$$\mathsf{RSS}\left(\boldsymbol{\beta}\right) = \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

- Using a normal prior for $\boldsymbol{\beta} \sim \mathrm{Normal}\left(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0\right)$, we have

$$f_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) = \frac{1}{\sqrt{(2\pi)^k \det\left(\boldsymbol{\Sigma}_0\right)}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right)$$
$$\propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right)$$

Q: What is the conditional posterior of $\boldsymbol{\beta}$?

$$f_{\boldsymbol{\beta}|\{\sigma^2,\mathbf{Y},\mathbf{X}\}}$$

- Using the precision parameter in the likelihood instead of σ^2, that is

$$\tau = \frac{1}{\sigma^2}$$

and using a gamma prior for $\tau \sim \mathrm{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$,

$$\begin{aligned} f_\tau &= \frac{\left(\nu_0\sigma_0^2/2\right)^{\nu_0/2}}{\Gamma\left(\nu_0/2\right)}\tau^{\nu_0/2-1}\exp\left(-\frac{\nu_0\sigma_0^2}{2}\tau\right) \\ &\propto \tau^{\nu_0/2-1}\exp\left(-\frac{\nu_0\sigma_0^2}{2}\tau\right) \end{aligned}$$

Q: What is the conditional posterior of $\boldsymbol{\tau}$?

$$f_{\sigma^2|\{\boldsymbol{\beta},\mathbf{Y},\mathbf{X}\}}$$

Q: How can we sample from the Joint posterior?

$$f_{\{\boldsymbol{\beta},\sigma^2\}|\{\mathbf{Y},\mathbf{X}\}}$$

- Since both conditionals are readily available, and both are pretty standard,

$$\boldsymbol{\beta} \mid \{\sigma^2, \mathbf{Y}, \mathbf{X}\} \sim \text{Normal}\,(\mathbf{m}, \mathbf{V})$$
$$\sigma^2 \mid \{\boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}\} \sim \text{Inverse-Gamma}\,(\alpha, \beta)$$

where

$$\begin{aligned}
\mathbf{m} &= \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{X}/\sigma^2\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^{\mathrm{T}}\mathbf{y}/\sigma^2\right) \\
\mathbf{V} &= \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{X}/\sigma^2\right)^{-1} \\
\alpha &= \frac{\nu_0 + n}{2}; \qquad \beta = \frac{\nu_0\sigma_0^2 + \mathsf{RSS}\,(\boldsymbol{\beta})}{2}
\end{aligned}$$

and positivity is satisfied, using Gibbs sampling is then straightforward

$$\left(\boldsymbol{\beta}, \sigma^2\right) \in \mathbb{R}^k \times (0, \infty)$$

- If other priors are used, we will have a different joint and a different sampling scheme, but the essences of Bayesian linear regression are the same.