

# Notes 8

Maryclare Griffin

3/23/2023

These notes are based on Chapters 1 and 6 of KNNL.

So far, we have not assumed a specific distribution for the errors. Going forward, we will assume that the errors are normally distributed. This will allow us to describe the distribution of the estimated regression coefficients and fitted values, in addition to their means and variances.

The **normal error linear regression model** for a dependent variable or response  $Y$  and independent variables, predictors, or covariates  $X_1, \dots, X_{p-1}$  is defined as:

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- The elements of  $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$  are parameters
- The elements of the  $n \times p$  matrix  $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \dots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{n,p-1} \end{pmatrix}$  are known constants
- $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$  is a random error term elements that are  $\epsilon_i$  that are independent and normally distributed with mean  $E\{\epsilon_i\} = 0$  and variance  $\sigma^2\{\epsilon_i\} = \sigma^2$ .

Going forward, we will always be using the normal error linear regression model unless otherwise specified.

Setting aside everything we've done the past couple weeks, let's revisit the question of how we might estimate  $\boldsymbol{\beta}$  under this model. Now that we have assumed a distribution for the errors, **maximum likelihood** is an option! First, we need to determine the distribution of the observed data under the normal error regression model.

**Note:** If  $Z$  is a normal random variable and  $a$  and  $c$  are fixed constants, then the transformed variable  $V = a + cZ$  is normally distributed, with mean  $E\{V\} = a + cE\{Z\}$  and variance  $\sigma^2\{V\} = c^2\sigma^2\{Z\}$ .

Accordingly, each  $Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i$  is a normal random variable with mean  $E\{Y_i\} = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}$  and variance  $\sigma^2\{Y_i\} = \sigma^2$ , and that each  $Y_i$  is independent of  $Y_j$  for all  $j \neq i$ .

**Note:** If  $Z$  is a normal random variable with mean  $\mu$  and variance  $\sigma^2$ , then the density function

of  $Z$  is

$$f(Z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Z - \mu)^2\right\}.$$

The **likelihood** of the data  $\mathbf{Y}$  given the parameters  $\boldsymbol{\beta}$ , the predictors  $\mathbf{X}$ , and the noise variance  $\sigma^2$  under the normal error linear regression model is:

$$\begin{aligned} L &= \prod_{i=1}^n f(Y_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^n \exp\left\{-\frac{1}{2\sigma^2}\left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n \left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2\right\}. \end{aligned}$$

What values of  $\sigma^2$  and  $\boldsymbol{\beta}$  maximize the likelihood  $L$ ?

To answer that, we need to take derivatives with respect to  $\sigma^2$  and elements of  $\boldsymbol{\beta}$  and find the values that set them equal to 0. However, we can tell this is going to be a bit of a pain just by examining how  $L$  depends on  $\boldsymbol{\beta}$ ; the term  $\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n \left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2\right\}$  is going to have some complicated derivatives.

Fortunately, finding the values of  $\sigma^2$  and  $\boldsymbol{\beta}$  that maximize the likelihood  $L$  is equivalent to finding the values of  $\sigma^2$  and  $\boldsymbol{\beta}$  that maximize the log-likelihood  $\ell = \log(L)$ ! This is a consequence of the fact that  $\log(x)$  is a monotonic, increasing function of  $x$ .

$$\begin{aligned} \ell &= \log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n \left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2\right\}\right) \\ &= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n \left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2 \\ &= -n \log(\sqrt{2\pi}) - \left(\frac{n}{2}\right) \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n \left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2 \end{aligned}$$

Aha! This depends on  $\boldsymbol{\beta}$  in a much nicer way! Now to find the values of  $\sigma^2$  and  $\boldsymbol{\beta}$  maximize the likelihood  $\log(L)$ , we need to take derivatives, not only with respect to  $\beta_0, \beta_1, \dots, \beta_{p-1}$  but also  $\sigma^2$ .

Before we take a lot of derivatives, maybe there's a shortcut. Let's see if we can use the derivative with respect to  $\sigma^2$  to find an expression for the value  $\hat{\sigma}^2$  that maximizes  $\ell$  for fixed values of  $\boldsymbol{\beta}$ . If we can, we can plug it into  $\ell$  and then only have to worry about maximizing over  $\boldsymbol{\beta}$  as opposed to maximizing over  $\boldsymbol{\beta}$  and  $\sigma^2$ .

The derivative of  $\ell$  with respect to  $\sigma^2$  is

$$-\left(\frac{n}{2}\right) \left(\frac{1}{\sigma^2}\right) - \left(-\frac{1}{\sigma^4}\right) \frac{1}{2} \sum_{i=1}^n \left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2.$$

Letting  $\hat{\sigma}^2$  refer to the value of  $\sigma^2$  that sets this derivative equal to 0 and simplifying, we have

$$-n \left(\frac{1}{\hat{\sigma}^2}\right) + \left(\frac{1}{\hat{\sigma}^4}\right) \sum_{i=1}^n \left(Y_i - \left(\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}\right)\right)^2 = 0.$$

Rearranging and multiplying through by  $\hat{\sigma}^4$  yields:

$$n\hat{\sigma}^2 = \sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2.$$

Great! Let's plug this expression for  $\hat{\sigma}^2$  into  $\ell$  for  $\sigma^2$ :

$$\ell = -n \log(\sqrt{2\pi}) - \left(\frac{n}{2}\right) \log \left( \frac{1}{n} \sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2 \right) - \frac{1}{2 \left( \frac{1}{n} \sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2 \right)} \sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2.$$

We can simplify this a lot! The sums of squares  $\sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2$  in the last term cancel, yielding

$$\ell = -n \log(\sqrt{2\pi}) - \left(\frac{n}{2}\right) \log \left( \sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2 \right) - \left(\frac{n}{2}\right) \log(n) - \frac{n}{2}.$$

The only part of this equation that depends on  $\beta$  is the middle term,  $-\left(\frac{n}{2}\right) \log \left( \sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2 \right)$ !

Thus, maximizing  $\ell$  with respect to  $\beta$  is equivalent to maximizing

$$-\left(\frac{n}{2}\right) \log \left( \sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2 \right).$$

Multiplying by a positive constant  $\left(\frac{n}{2}\right)$  and taking the logarithm does not change the values of  $\beta$  that maximize the function, so maximizing  $\ell$  with respect to  $\beta$  is equivalent to maximizing

$$-\sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2.$$

Remember,  $\sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2$  is the least squares objective  $Q$  that we minimized to obtain the least squares estimates  $\mathbf{b}$  of  $\beta$ ! Substituting  $Q$  in for  $\sum_{i=1}^n \left( Y_i - \left( \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} \right) \right)^2$  and remembering that maximizing the negative of a function is the same as minimizing it, we can see that maximizing  $\ell$  with respect to  $\beta$  is equivalent to minimizing  $Q$ !

Thus, the maximum likelihood estimate  $\hat{\beta}$  of  $\beta$  is equal to the least squares estimate  $\mathbf{b}$  of  $\beta$ , i.e.  $\hat{\beta} = \mathbf{b}$ !!!

This is very convenient, it means that when we assume the normal errors linear regression model, our maximum likelihood estimate is the same as the least squares estimate! As a result, all of the derivations we did in the past few weeks to obtain the least squares estimator  $\mathbf{b}$  and derive its properties apply to the maximum likelihood estimator  $\hat{\beta}$  as well. This is a nice coincidence that is a direct result of assuming that the normal errors. If we had assumed another distribution for the errors, it might not have been true.