Section 1.8. Normal Error Regression Model

Note. In the simple linear regression model of Section 1.3, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, no particular distribution was imposed on the error terms ε_i. In this section we assume the error terms are normally distributed with distribution $N(0, \sigma^2)$. In Section 1.3 we *did* assume the mean of each ε_i was 0 and the variance of each ε_i was the constant σ^2. The new condition here is the normality of the distribution.

Definition. The *normal error regression model* is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1.24}$$

where:

Y_i is the value of the response variable in the ith trial,

X_i is a known constant, the "level" of the predictor variable in the ith trial,

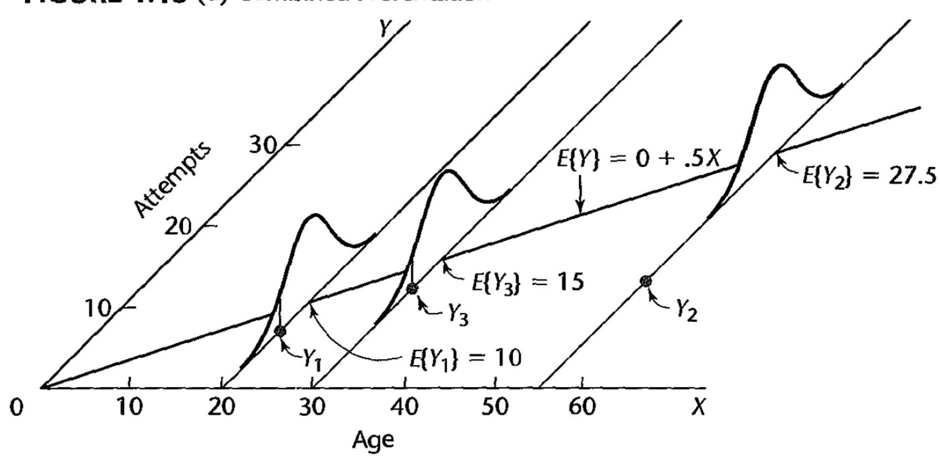
β_0 and β_1 are parameters, and

ε_i are independent with distribution $N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$.

Note 1.7.A. The Y_i are independent normal random variables with mean $E\{Y_i\} = \beta_0 + \beta_1 X_i$ and variance σ^2. This is illustrated in Figure 1.6 of Section 1.3. Kutner et al. justify the normality assumption in two ways. First, a larger number of small (mutually independent) random effects on response Y_i combine to give, by the Central Limit Theorem (for details on the Central Limit Theorem, see my online notes for Mathematical Statistics 1 [STAT 4047/5047] on Section 5.3. Central Limit Theorem), an approximately normal distribution. Second, unless the departures from normality are large (such as a strongly skewed distribution), the actual confidence coefficients will be close to the levels for exact normality.

Note. In the normal error regression model, for a fixed i, Y_i has a normal distribution with mean $\beta_0 + \beta_1 X_i$ and variance σ^2. So when we sample Y_i, the observed value has associated value when substituted into the normal distribution. Kutner et al. call this the "density of the probability distribution" at Y_i (notice this value is not a probability since each value Y_i will appear with probability 0 since the normal distribution is a continuous distribution). We next associate a likelihood value based on the observed data. We then maximize this likelihood to estimate the values of β_0 and β_1. Figure 1.15(d) illustrates this idea when there are three observations Y_1, Y_2, Y_3 (the fit of the line to this data is poor and the likelihood associated with these observations is low for the chosen values of β_0 and β_1 [which are 0 and 0.5, respectively, in the figure]).

FIGURE 1.15 (d) Combined Presentation



Definition. In the normal error regression model, the *density* of an observation Y_i is

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{Y_i - \beta_0 = \beta_1 X_i}{\sigma}\right)^2\right).$$

The *likelihood function* for n observations $Y_1, Y_2, \ldots, Y_n$ is the product of the individual densities:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right)$$
$$= \frac{1}{(2\pi\sigma^2)^{-1/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} Y_i = \beta_0 = \beta_1 X_i)^2\right).$$

Note/Definition. We seek the values of $\beta_0, \beta_1, \sigma^2$ that maximize the likelihood. These are called *maximum likelihood estimators* and are denoted $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, respectively. We next claim that $\hat{\beta}_0 = b_0$ and $\hat{\beta}_1 = b_1$ where b_0 and b_1 are the least squares values that minimize quantity Q of Theorem 1.6.A. We do so by finding a critical point of the likelihood function, but we do not (like Kutner et al.) confirm that the critical point gives a maximum.

Theorem 1.8.A. In the normal error regression model, the likelihood function $L(\beta_0, \beta_1, \sigma^2)$ is minimized at $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ where $\hat{\beta}_0 = b_0$, $\hat{\beta}_1 = b_1$, and $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$ where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1$.

Note. Kutner et al. state that $\hat{\sigma}^2$ is a biased estimator of parameter σ^2 (that is, $E\{\sigma^2\} \neq \hat{\sigma}^2$), and that the unbiased estimator

$$MSE = s^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$$

is used in place of $\hat{\sigma}^2$. Of course, for n large, $n/(n-2) \approx 1$.

Revised: 9/30/2022