Notes 3

Maryclare Griffin

2/14/2023

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

Recall from the previous notes, the linear regression model for a dependent variable or response Y and independent variables, predictors, or covariates X_1, … X_{p-1} is defined as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

where:

- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
- $X_{i1}, \ldots, X_{i,p-1}$ are known constants
- ϵ_i is a random error term with mean $E\left\{\epsilon_i\right\} = 0$ and variance $\sigma^2\left\{\epsilon_i\right\} = \sigma^2$; ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma\left\{\epsilon_i, \epsilon_j\right\} = 0$ for all i, j; $i \neq j$)
- $i = 1, \ldots, n$

  Note: Sometimes, the linear regression model will have an additional predictor X_{i0} and be written as

  $$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i.$$

  This is just an issue of notation, it is **not** a different model than the one we use in this class. For this class, we will always set $X_{i0} = 1$. This is consistent with how regression models are usually used in practice.

The parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are called **regression coefficients**. They have the following meaning:

- β_0 is the value of the mean response $E\left\{Y\right\}$ when all of the predictors are exactly equal to zero, $X_{i1} = \cdots = X_{i,p-1} = 0$. Alternatively, it may be called the **intercept** of the regression plane.

- For $k > 0$, β_k is the change in the mean response $E\left\{Y\right\}$ per unit increase in predictor X_k,when the rest of the predictors are held constant. Alternatively, it may be called a **partial regression coefficient**.

  Note: When we are using the simple linear regression model with $p = 2$, we can interpret β_0 and β_1 as the **intercept** and **slope** of the regression line.

  Note: Just because a regression coefficient has an interpretation doesn't mean that that interpretation makes sense! Whether or not it makes sense is determined by the context of the problem and the observed data. For instance, the intercept β_0 only has an interpretation that makes sense if $X_{i1} = \cdots = X_{i,p-1} = 0$ is a plausible set of predictor values, i.e. if all $X_{i1} = \cdots = X_{i,p-1} = 0$ is in the **scope** of the model.

  Example 1: Consider an electrical distributor, who is studying the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week and the time required to prepare the bids. Suppose we magically know that the following regression model holds:

  $$Y_i = 9.5 + 2.1 X_i + \epsilon_i,$$

where X is the number of bids prepared in a week and Y is the number of hours required to prepare the bids. Remember, in the real world, we **never** know the true regression model. We just see the data, $Y_1, \dots, Y_n$ and $X_1, \dots, X_n$.

The intercept $\beta_0 = 9.5$ can be interpreted as the number of hours required to prepare 0 bids. The slope $\beta_1 = 2.1$ can be interpretated as the mean change in the number of hours associated with one additional bid. The regression function $E\left\{Y\right\} = 9.5 + 2.1X$ is depicted in Figure 1.

```r
curve(9.5 + 2.1*x, from = 0, to = 50,
      xlab = "Number of Bids Prepared",
      ylab = "Hours", main = "E{Y}")
```

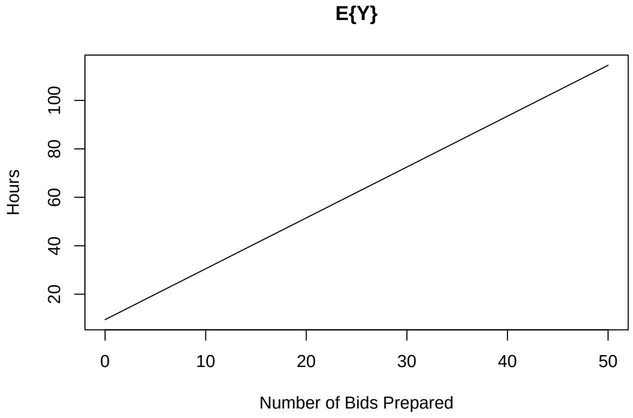

Figure 1: Example 1

Suppose that in the i-th week, $X_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$. In that case, the corresponding error ϵ_i satisfies:

$$\begin{aligned}108 &= 9.5 + 2.1\left(45\right) + \epsilon_i \\ &= 9.5 + 94.5 + \epsilon_i \\ &= 104 + \epsilon_i.\end{aligned}$$

Accordingly, the corresponding error $\epsilon_i = 4$.

Example 2: Consider an manufacturer marketing a new product, who is studying the relationship between the test market sales in tens of thousands of dollars (Y), point-of-sale expenditures in thousands of dollars (X_1), and TV expenditures in thousands of dollars (X_2). Suppose we magically know that the following regression model holds:

$$Y_i = 10 + 2X_{i1} + 5X_{i2} + \epsilon_i.$$

Again, remember that we **never** know the true regression model in the real world. We just see the data, $Y_1, \dots, Y_n$, $X_{11}, \dots, X_{n1}$, and $X_{12}, \dots, X_{n2}$.

The parameter $\beta_0 = 10$ is the intercept of the regression plane, and can be interpreted as indicating that the mean test market sales will be \$100,000 when no point of sale expenditures or TV expenditures are made.

We can interpret the parameter $\beta_1 = 2$ as indicating that a a \$1,000 increase in point-of-sale expenditures is associated with an increase of a \$20,000 in mean test market sales when TV expenditures are held constant.

We can interpret the parameter $\beta_2 = 5$ as indicating that a \$1,000 increase in TV expenditures is associated with an increase of a \$50,000 in mean test market sales when point-of-sale expenditures are held constant.

Often, interpreting β_0 doesn't make sense because it is not plausible to imagine that the predictors $X_{i1}, \ldots X_{i,p-1}$ could be equal to 0. For this reason, the linear regression model is sometimes rewritten or reparameterized. Define the average value of predictor X_{ik} in the sample to be $\bar{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ik}$. We can add and subtract $\beta_k \bar{X}_{ik}$ from the linear model equation,

$$Y_i = \beta_0 + \left(\sum_{k=1}^{p-1} \beta_k \bar{X}_k\right) + \beta_1 \left(X_{i1} - \bar{X}_1\right) + \beta_2 \left(X_{i2} - \bar{X}_2\right) + \cdots + \beta_{p-1} \left(X_{i,p-1} - \bar{X}_{p-1}\right) + \epsilon_i.$$

Then we can define $\beta_0^* = \beta_0 + \left(\sum_{k=1}^{p-1} \beta_k \bar{X}_k\right)$, and rewrite the linear model equation as:

$$Y_i = \beta_0^* + \beta_1 \left(X_{i1} - \bar{X}_1\right) + \beta_2 \left(X_{i2} - \bar{X}_2\right) + \cdots + \beta_{p-1} \left(X_{i,p-1} - \bar{X}_{p-1}\right) + \epsilon_i.$$

The regression coefficients $\beta_0^*, \beta_1, \ldots, \beta_{p-1}$ have the following meaning:

- β_0^* is the mean response when each predictor is equal to its sample mean predictors are set to their mean values, $\bar{X}_1, \ldots, \bar{X}_{p-1}$.
- For $k > 0$, each β_k is still the change in the mean response $E\left\{Y\right\}$ per unit increase in predictor X_k,when the rest of the predictors are held constant.

Note: This is the same linear model, just a different way of representing it. We will use both representations interchangeably.