# Chapter 1 Linear Regression with One Predictor Variable

## 1.3 Simple Linear Regressfion Model with Distribution of Error Unspecified

**Definition.** ***Simple Linear Model*** *A model that is linear in simple (1 predictor variable) and linear in parameters and linear in predictor variable. A model linear in parameter and predictor variable is called* ***first-order model***

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

*where*

1. $Y_i$ *is value of response variable and* $X_i$ *is the value of predictor variable in the i-th trial*
2. $\epsilon_i$ *is a random error term with mean* $\mathbb{E}[\epsilon_i] = 0$ *and* $\sigma^2[\epsilon_i] = \epsilon^2$*. Also* $\epsilon_i$ *and* $\epsilon_j$ *are uncorrelated* ($\sigma^2[\epsilon_1], \epsilon_j = 0$ *for all* $i \neq j$)
3. ***regression coefficients*** $\beta_1$ *is slope of the regression line, indicating change in mean of probability distribution of* $Y$ *per unit increase in* $X$*.* $\beta_0$ *is the* $Y$ *intercept of the regression line*

***Properties***

1. $Y_i$ ***is a random variable****, a summation of a constant* $\beta_0 + \beta_1 X_i$ *and the random error* $\epsilon_i$*.*
2. ***Distribution of*** $Y_i$
   (a) *By* $\mathbb{E}[\epsilon_i] = 0$
   $$\mathbb{E}[Y_i] = \mathbb{E}[\beta_0 + \beta_1 X_i + \epsilon_i] = \beta_0 + \beta_1 X_i$$
   (b) *By* $\sigma^2\{\epsilon_i\} = \sigma^2$
   $$\sigma^2\{Y_i\} = \sigma^2\{\epsilon_i\} = \sigma^2$$
3. *The* ***regression function*** *relates mean of probability distribution of* $Y$ *for given* $S$ *to level of* $X$
   $$\mathbb{E}[Y] = \beta_0 + \beta_1 X$$
4. $Y_i$ ***and*** $Y_j$ ***are uncorrelated*** *since errors are uncorrelated*

*In summary this regression model implie reseponse* $Y_i$ *come from probability distribution whose means are* $\mathbb{E}\{Y\} = \beta_0 + \beta_1 X_i$ *whose variances are* $\sigma^2$ *(same for all levels of* $X$*), furthermore, two responses* $Y_i$ *and* $Y_j$ *are uncorrelated*

### 1.6 Estimation of Regression Function

**Definition.** ***Method of Least Squares*** *is used to estimate regression paramters $\beta_0$ and $\beta_1$. The MLS considers sum of n squared deviations of $Y_i$ from its expected value*

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 y_i)^2$$

*estimators of $\beta_0$ and $\beta_1$ are values $b_0$ and $b_1$ that minimize $Q$ given sample observations $(x_1, y_1), \cdots$. By taking partials of $RSS$ and set it to zero, we derive a pair of* ***Normal Equation***

$$\sum y_i = nb_0 + b_1 \sum x_i$$

$$\sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2$$

*Solving it we get LS estimator*

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

***Properties of LS regression coefficient*** *mostly from derivation of LS estimators*

$$\sum_{i=1}^{n} e_i = 0 \text{(from } \frac{\partial RSS}{\beta_0}\text{)} \qquad \sum_{i=1}^{n} e_i^2 \quad \text{minimized(from LSE)}$$

$$\sum y_i = \sum \hat{y}_i \text{(since } \sum_{i=1}^{n} e_i = 0\text{)} \quad \sum x_i e_i = 0 \text{(from } \frac{\partial RSS}{\beta_1}\text{)} \quad \sum \hat{y}_i e_i = 0$$

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$$

**Definition.** ***Gauss Markov Theorem*** *Under conditions of regression model, the least squares estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all unbiased linear estimators*

*Proof.* 1. **For** $\hat{\beta}_1$

$$\hat{\beta}_1 = \sum c_i y_i \quad \text{where } c_i \text{ is arbitrary}$$

Now we prove its **unbiased**

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\{\sum c_i y_i\} = \sum c_i \mathbb{E}[Y_i] = \sum c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum c_i + \beta_1 \sum c_i x_i = \beta_1$$

given the restriction that

$$\sum c_i = 0 \quad \sum c_i x_i = 1$$

which holds for both $\hat{\beta}_1$ and $\hat{\beta}_0$ ...

□

**Definition.** ***Point estimation of mean response*** *Given regression function*

$$\mathbb{E}[Y] = \beta_0 + \beta_1 X$$

*so we have a estimated regression line*

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

*where $\hat{y}_i$ is value of estimated regression level at level $x$, it is a point estimate of the mean response when the level of the predictor variable is $X$. By the previous theorem, $\hat{y}$ is an unbiased estimator of $\mathbb{E}[Y]$ with minimum variance in the class of unbiased linear estimators.*

**Definition.** ***residual*** *the $i$-th residual is the difference between the observed value $y_i$ and the corresponding fitted value $\hat{y}_i$.*

$$e_i = y_i - \hat{y}_i$$

*In the case of simple linear model, we have*

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

*Informally, it is the vertical distance of $y_i$ from the fitted value $\hat{y}_i$ on the estimated regression line, which is known.*

### 1.7 Estimation of Error terms variance $\sigma^2$

**Definition.** ***Point Estimator of $\sigma^2$***

1. ***Single Population*** *We use **mean square** $s^2$ to estimate population variance $\sigma^2$*

$$s^2 = \frac{1}{n-1} \sum (y_i - \overline{y})^2$$

   *The lost degree of freedom comes with using $\overline{y}$ to estimate mean*

2. ***Regression Model*** *Note $y_i$ comes with difference probability distribution based on levels of $x_i$. So to calcualte sum of squared deviation, we have to calculate deviation around its own estimated mean $\hat{y}_i$.*

   (a) ***Residual/Error sum of square (RSS, SSE)***

$$RSS/SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

   (b) ***Residual/Error Mean Square (MSE)***

$$s^2 = \frac{RSS}{n-2} = \frac{\sum e_i^2}{n-2}$$

   *The loss of 2 degree of freedom comes from using $\hat{\beta}_0$ and $\hat{\beta}_1$ to estimate regression coefficient to get the $\hat{y}_i$*

   (c) ***MSE is an unbiased estimator of $\sigma^2$***

## 1.8 Normal Error Regression Model

*Note.* **Motivation** Least squared method provides unbiased point estimator for $\beta_0$ and $\beta_1$ regardless of the distribution of $\epsilon_i$ (and hence of $Y_i$). However, need to make assumption about form of distribution of $\epsilon_i$ to set up **interval estimate** and make tests.

**Definition.** ***Normal Error Regression Model***

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

*Additionally,* $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ *for* $i = 1, \cdots, n$.
***Properties***

1. *Note independence of* $\epsilon_i$ *comes from the the uncorrelatedness of* $\epsilon_i$ *and properties of normal distribution*

2. $Y_i$ *are independent normal random variable*

3. $\epsilon_i$ *being normal is somewhat justified as it represent all factors which tend to comply with CLT and cause error distribution approach normality as number of factor effects becomes large*

**Definition.** ***Parameter ($\beta_0$, $\beta_1$, $\sigma^2$) estimation by Method of maximum likelihood***
*Turns out* $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *are same as least squared estimator. The estimator for* $\sigma^2$ *is different however*

$$\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

*estimator* $\hat{\sigma}^2$ *is biased with following relationship to mean squre error*

$$s^2 = MSE = \frac{n}{n-2} \hat{\sigma}^2$$