

Chapter 1 – Linear Regression with One Predictor Variable

- statistical relation vs. deterministic functional relation
- scatter plot may provide evidence of a relation; maybe linear or curvilinear
- two essential ingredients of a statistical relation that lead to a regression model:
 - 1) tendency of Y to vary with X in a systematic fashion
 - 2) a scattering of points around the curve of statistical relationship
- general regression model postulates:
 - 1) probability distribution (possibly unspecified) of Y for each level of X
 - 2) means of probability distribution vary in some systematic fashion with X
- model construction:
 - selection of predictor variables
 - selecting functional form
 - determining scope of the model
- uses of regression analysis:
 - description
 - control
 - prediction
- a regression relationship does not imply Y depends causally on X
- **simple linear model (1.1) with distribution of error terms unspecified**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{for } i = 1, 2, \dots, n$$

where Y_i is value of response variable in i th trial

β_0 and β_1 are parameters (constants, typically unknown)

X_i is a known constant, value of predictor variable in i th trial

ϵ_i is random error term in i th trial with $E\{\epsilon_i\} = 0$, $\sigma^2\{\epsilon_i\} = \sigma^2$,
and $\text{cov}\{\epsilon_i, \epsilon_j\} = \sigma\{\epsilon_i, \epsilon_j\} = 0$ for $i \neq j$.

- simple refers to the fact that there is only one predictor variable
- linear refers to the fact that Y is a linear function of β_0 and β_1 and also to the fact that Y is a linear function of X
- the **regression function** for model (1.1) is $E\{Y\} = \beta_0 + \beta_1 X$
- $\sigma^2\{Y_i\} = \sigma^2$ and $\sigma\{Y_i, Y_j\} = 0$ for $i \neq j$

Alternative versions of model (1.1): $Y_i = \beta_0 X_0 + \beta_1 X_i + \epsilon_i$ where $X_0 \equiv 1$

or

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i \text{ where } \beta_0^* = \beta_0 + \beta_1 \bar{X}$$

-
- observational data vs. experimental data

observational data: predictor variable not controlled; obtained from nonexperimental study

experimental data: control exercised over predictor variable; control through random assignments.

- completely randomized design – randomized assignments of 'treatments' to 'experimental units'
- see flowchart for regression analysis strategy in text

-
- **least squares method** for estimating β_0 and β_1 :

minimizing Q (the sum of squared deviations)

$$\text{where } Q = Q(\beta_0, \beta_1) = \sum (Y_i - E\{Y_i\})^2 = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

normal equations (with values b_0 and b_1 that minimize Q):

$$(i) \sum Y_i = nb_0 + b_1 \sum X_i \quad \text{and} \quad (ii) \sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

-
- solving normal equations provides the estimates:

$$b_1 = \frac{SS_{xy}}{SS_x} \quad \text{and} \quad b_0 = \bar{Y} - b_1 \bar{X}$$

$$\text{where } SS_{xy} \equiv \sum (X_i - \bar{X})(Y_i - \bar{Y}) \quad \text{and} \quad SS_x = \sum (X_i - \bar{X})^2$$

-
- estimated (or fitted) regression line: $\hat{Y} = b_0 + b_1 X$

-
- **ith residual**: $e_i = Y_i - \hat{Y}_i$, that is, response value minus fitted value

- facts: $\sum e_i = 0$, $\sum Y_i = \sum \hat{Y}_i$, $\sum X_i e_i = 0$, and $\sum \hat{Y}_i e_i = 0$

- the fitted regression line always goes through the point (\bar{X}, \bar{Y})

- **error sum of squares**: $SSE = \sum e_i^2$; **error mean square**: $MSE = SSE/(n - 2)$

- under model (1.1), $E\{MSE\} = \sigma^2$, i.e., MSE is unbiased estimator of σ^2

- **Gauss-Markov Theorem**: under conditions of regression model(1.1), b_0 and b_1 are BLUE (best linear unbiased estimators).

- **normal error regression model (1.24)** : $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

where ϵ_i are i.i.d. $N(0, \sigma^2)$ for $i = 1, \dots, n$

– then Y_i are independent random variables

with distribution $N(\beta_0 + \beta_1 X_i, \sigma^2)$

– maximum likelihood estimates of β_0 and β_1 :

$$\hat{\beta}_0 = b_0 \quad \text{and} \quad \hat{\beta}_1 = b_1 \quad (\text{same as least squares est.})$$

– under normal error model (1.24), b_0 and b_1 are also

MVU (minimum variance unbiased), consistent, and sufficient. (definitions in App. A)