Chapter 12

Conditions for regression inference (LINER) Suppose we have n observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

- **Linear:** The relationship between x and y is linear in the population. For any fixed value of x, the mean response μ_y falls on the population regression line $\mu_y = \alpha + \beta x$. The slope β and intercept α are usually unknown parameters.
- **Independent:** Individual observations are independent of each other.
- **Normal:** For any fixed value of x, the response y varies according to a Normal distribution.
- **Equal variance:** The standard deviation of y (call it σ) is the same for all values of x. The common standard deviation σ is usually an unknown parameter.
- **Random:** The data come from a well-designed random sample or randomized experiment.

Exponential model A relationship of the form $y = ab^x$. If the relationship between two variables follows an exponential model, and we plot the logarithm (base 10 or base e) of y against x, we should observe a straight-line pattern in the transformed data.

Population regression line (true regression line) The regression line $\mu_y = \alpha + \beta x$ based on the entire population of data.

Power model A relationship of the form $y = ax^p$. When experience or theory suggests that the relationship between two variables is described by a power model, you can transform the data to achieve linearity in two ways: (1) raise the values of the explanatory variable x to the p power and plot the points $\left(x^p, y\right)$, or (2) take the pth root of the values of the response variable y and plot the points $\left(x, \sqrt[p]{y}\right)$. If you don't know what power to use, taking the logarithms of both variables should produce a linear pattern.

Sample regression line (estimated regression line) The least-squares regression line $\hat{y} = a + bx$ computed from the sample data.

Standard error of the slope Used to estimate the spread of the sampling distribution of b.

$$\mathrm{SE}_b = \frac{s}{\sqrt{n-1} \cdot s_x}$$

***t* interval for the slope β** When the conditions for regression inference are met, a level C confidence interval for the slope β of the population regression line is

$$b \pm t^* \mathrm{SE}_\mathrm{b}$$

In this formula, the standard error of the slope is

$$\mathrm{SE}_b = \frac{s}{\sqrt{n-1} \cdot s_x}$$

and t^* is the critical value for the t distribution with df = $n - 2$ having area C between $-t^*$ and t^*.

***t* test for the slope** Suppose the conditions for inference are met. To test the hypothesis $H_0 : \beta = \text{hypothesized value}$, compute the test statistic

$$t = \frac{b - \beta_0}{\mathrm{SE}_b}$$

Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with df = $n - 2$.

Transforming Applying a function such as the logarithm or square root to a quantitative variable is called transforming the data.