Chapter 2. Inferences in Regression and Correlation Analysis

Note. In this chapter we consider "inferences" of the regression parameters β_0, β_1, and $E\{Y\}$. These inferences are largely based on confidence intervals for these parameters. In Section 2.6. Confidence Band for Regression Line we present a "confidence band" that contains the regression line. **Throughout Chapter 2 (excluding Section 2.11) and in the remainder of Part I unless stated otherwise, we assume that we are addressing the normal error regression model (1.24).**

Section 2.1. Inferences Concerning β_1

Note. Recall that β_1 is the slope of the regression line. In this section we consider the expected value of the estimator b_1 (and in the process give a proof of the Gauss-Markov Theorem, Theorem 1.11), the variance of b_1, confidence intervals for β_1, and discuss hypothesis tests in this setting.

Note. Recall from Theorem 1.6.A that "point estimator" b_1 of β_1 based on date points (X_i, Y_i) for $i = 1, 2, \ldots, n$ in the simple linear regression model is

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Definition. The *sampling distribution* of b_1 is the collection of different values of b_1 that result from repeated sampling when the levels (i.e. values) of the predictor variable X are held constant from sample to sample.

Note. We wish to find the mean and variance of the sampling distribution of b_1 (that is, $E\{b_1\}$ and $\sigma^2\{b_1\}$). First, we need a preliminary lemma.

Lemma 2.1.A. Statistic b_1 is a linear combination of the observations Y_i:

$$b_1 = \sum_{i=1}^{n} k_i Y_i \text{ where } k_i = \frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Note 2.1.A. Two properties of the k_i of Lemma 2.1.A that we need are:

$$\begin{aligned}
\sum_{i=1}^{n} k_i &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right) \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \overline{X})^2} \sum_{i=1}^{n}(X_i - \overline{X}) = \frac{0}{\sum_{i=1}^{n}(x_i - \overline{X})^2} = 0. \qquad (2.5) \\
\sum_{i=1}^{n} k_i^2 &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right)^2 \\
&= \frac{1}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2} \sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2}. \qquad (2.7)
\end{aligned}$$

Note 2.1.B. In the normal error regression model, by Note 1.7.A, the Y_i are independent normal random variables. As shown in Mathematical Statistics 1 (MATH 4047/5047), a linear combination of independent normally distributed random variables is itself normally distributed (see my online notes for Mathematical Statistics

1 on Section 3.4. The Normal Distribution; notice Theorem 3.4.2). So by Lemma 2.1.A, we see that b_1 is normally distributed(!).

Note. We are ready to prove the Gauss-Markov Theorem (Theorem 1.11; see Section 1.6. Estimation of Regression Function) in the special case that that the error terms are normally distributed. As a point of history, Carl Friedrich Gauss (April 30, 1777–February 23, 1855) first proved the result under the assumption of normally distributed error terms in 1821. His results were written in Latin, but were translated into French and published (by Joseph Bertrand) in 1855. Andrei Markov (June 14, 1856–July 20, 1922) dropped the normality condition and the version of the result as we stated in it Section 1.6. This appeared in a chapter on the method of least squares in a book he published in 1912.

Carl F. Gauss

Andrei Markov

The above images are from the MacTutor History of Mathematics Archive. These historical comments are based on R. L. Plackett's "A Historical Note on the Method of Least Squares," *Biometrika*, **36**(3,4), 458–460 (1949); a copy is available online from JSTOR. These websites were accessed 10/1/2022.

Theorem 1.11. The Gauss-Markov Theorem for the Normal Error Regression Model.

Consider the data points (X_i, Y_i) for $i = 1, 2, \ldots, n$ and the normal error linear regression model $Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$ given in (1.1) with the added hypothesis that each error term has a $N(0, \sigma^2)$ distribution. The least squares estimators

$$b_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \text{ and } b_0 = \frac{1}{n}\left(\sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i\right) = \overline{Y} - b_1 \overline{X}$$

are unbiased (that is, $E\{b_0\} = \beta_0$ and $E\{b_1\} = \beta_1$) and have minimum variance among all unbiased linear estimators (i.e., linear combinations of the Y_i).

Revised: 10/1/2022