Section 1.6. Estimation of Regression Function

Note. In this section we start our computation exploration of linear models. We use the "least squares" approach to derive estimations of the optimal values of β_0 and β_1 in the simple linear regression model.

Note. As an introductory example, consider a small-scale experiment in which $n = 3$ subjects are asked to perform a difficult task. The subjects attempt the task until they succeed. We take as the predictor variable (or the "explanatory variable") as age and represent it by X_i. We take the response variable as the number of attempts and represent it by Y_i. In general, we associate with trial i the ordered pair (X_i, Y_i). Consider the data:

Subject i**:**	**1**	**2**	**3**
Age X_i:	20	55	30
Number of attempts Y_i:	5	12	10

Here, we have $n = 3$ and the data points $(X_1, Y_1) = (20, 5)$, $(X_2, Y_2) = (55, 12)$, and $(X_3, Y_3) = (30, 10)$.

Note 1.6.A. We employ the least squares method to find "good" estimators of the regression parameters β_0 and β_1. This is also addressed in my online notes for Calculus 3 (MATH 2110) on Section 14.7. Extreme Values and Saddle Points (see Page 828 Number 65). The simple linear regression model predicts a value of $\beta_0 + \beta_1 X_i$ for the response variable when the predictor variable is X_i, whereas the observed value of the response variable is Y_i. So the deviation of the predicted

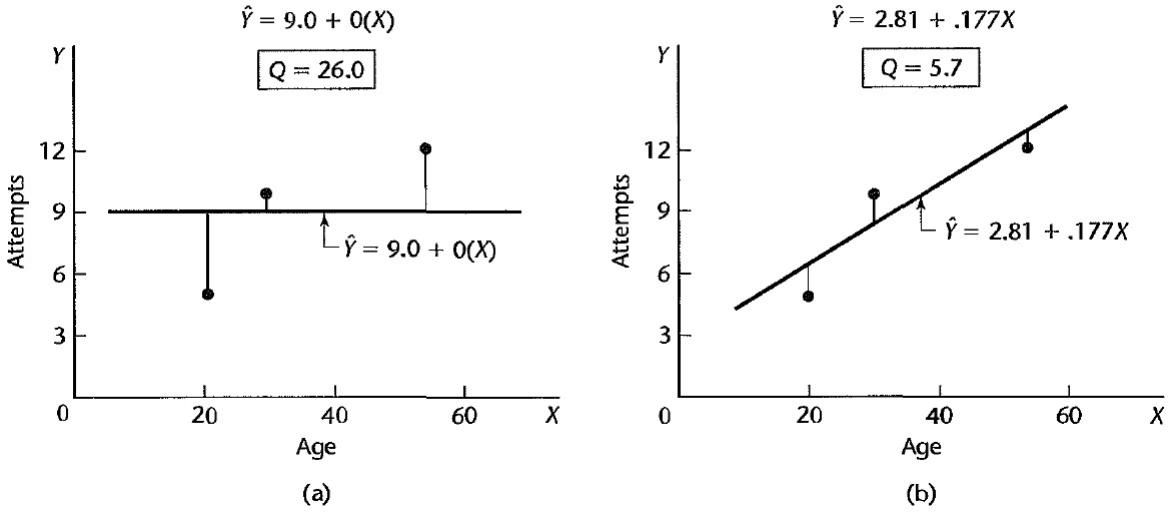
value from the observed value is $Y_i - (\beta_0 + \beta_1 X_i) = Y_i - \beta_0 - \beta_1 X_i$. We square this deviation for each i and then sum the squares of the deviation to produce quantity Q. That is, we want to find values of β_0 and β_1 that will minimize Q, where

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

We will denote the values of β_0 and β_1 which minimize Q as b_0 and b_1, respectively.

Note. Figure 1.9 gives two plots of the data from the table above. Two lines are plotted, the line $\hat{Y} = 9.0 + 0(X)$ in Figure 1.9(a), and $\hat{Y} = 2.81 + 0.177X$ in Figure 1.9(b). Since we have the raw data and the lines, then we can calculate the value of the sum of the squares of the deviations (i.e., the value of Q) in both cases. For the line $\hat{Y} = 9.0 + 0(X)$ we have $Q = 26.0$, and for the line $\hat{Y} = 2.81 + 0.177X$ we have $Q = 5.7$. So the second line is a better fit in terms of making Q small. In fact, the second lint is the least squares regression line.

FIGURE 1.9 Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.



Theorem 1.6.A. For data points (X_i, Y_i) where $i = 1, 2, \ldots, n$, the values of β_0 and β_1 which minimize

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

are

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

respectively.

Note. The next theorem will be proved in the next chapter in the special case when the error terms are normally distributed (see Section 2.1. Inferences Concerning β_1).

Theorem 1.11. The Gauss-Markov Theorem.

Consider the data points (X_i, Y_i) for $i = 1, 2, \ldots, n$ and the simple linear regression model $Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$ given in (1.1) (see Section 1.3. Simple Linear Regression Model with Distribution of Error Terms Unspecified for the detailed assumptions of the model). 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).

Note/Definition. The equation $y = b_1x + b_0$ is the *regression equation* or *regression line*. To simplify by-hand computations, we will establish in Section 2.1. Inferences Concerning β_1 the relationship

$$b_1 = \frac{\sum(X_i - \overline{X})Y_i}{\sum(X_i - \overline{X})^2} = \sum k_iY_i \text{ where } k_i = \frac{X_i - \overline{X}}{\sum(X_i - \overline{X})^2}.$$

Example 1.6.A. Kutner et al. give an example based on 25 data points (the data is available on the CD-ROM that comes with the book). Here, we give an example from William Navidi's *Statistics for Engineers and Scientists*, 3rd Edition, McGraw-Hill (2011). This is the book used in ETSU's Foundations of Probability and Statistics-Calculus (MATH 2050), and I have online notes in preparation for this class. Navidi's Exercise 7.2.9 is as follows.

Exercise 7.2.9. The article "Testing the Influence of Climate, Human Impact and Fire on the Holocene Population Expansion of *Fagus sylvatica* in the Southern Prealps (Italy)" (V. Valsecchi, W. Flnsinger, et al., *The Holocene* 2008: 603–614) presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

Depth	284.5	407.5	512.0	551.0	578.5	697.0	746.5
Age	1255	3390	5560	6670	7160	9820	11030

First, we use the predictor variable X_i of Depth and the response variable Y_i of Age. Notice that the averages (to two decimal places) are

$$\overline{X} = \sum X_i/n = (284.5 + 407.5 + 512.0 + 551.0 + 578.5 + 697.0 + 746.5)/7 = 539.57$$

$$\overline{Y} = \sum Y_i/n = (1255 + 3390 + 5560 + 6670 + 7160 + 9820 + 11030)/7 = 6412.14$$

i	X_i	Y_i	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})(Y_i - \overline{Y})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$
1	284.5	1255	−255.1	−5157.1	1315576.2	65076.0	26595680.4
2	407.5	3390	−132.1	−3022.1	399219.4	17450.4	9133088.4
3	512.0	5560	−27.6	−852.1	23518.0	761.8	726074.4
4	551.0	6670	11.4	257.9	2940.1	130.0	66512.4
5	578.5	7160	38.9	747.9	29093.3	1513.2	559354.4
6	697.0	9820	157.4	3407.9	536403.5	24774.8	11613782.4
7	746.5	11030	206.9	4617.9	955443.5	42807.6	21325000.4
sum	3777	44885	−0.2	0.3	3262194	152513.8	70019492.8

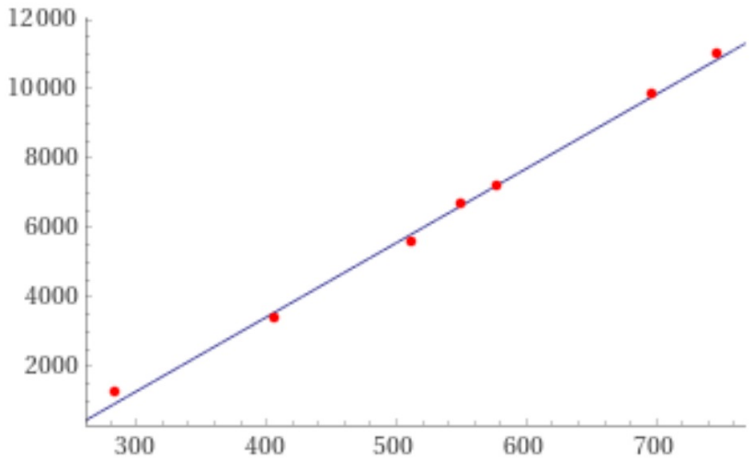
The sums of the $X_i - \overline{X}$ and the $Y_i - \overline{Y}$ should be 0, reflecting round off error. Therefore, we have the least squares estimators of β_1 and β_0 (to two decimal places) as

$$b_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{3262194}{152513.8} = 21.39$$

and

$$b_0 = \frac{1}{n}\left(\sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i\right) = \overline{Y} - b_1\overline{X} = 6412.14 - (21.39)(539.57) = -5129.26,$$

respectively. The regression equation is $y = 21.39x - 5129.94$. A plot of the data points and the regression line is given below. This was generated with the online software Wolfram Alpha.



A plot of the Depth/Age data points and the regression line.

Note. In Note 1.3.A, we have the regression model $E\{Y\} = \beta_0 + \beta_1 X$. With the sample estimators b_0 and b_1, we estimate the regression model as $\hat{Y} = b_0 + b_1 X$, where $\hat{Y}$ (pronounced "Y hat") is the estimate of the response variable Y for a given value of the predictor variable X.

Definition. In the estimate of the regression model, $\hat{Y} = b_0 + b_1 X$, a value of the predictor variable is the *level* of X, a value of the response variable Y is a *response*, and $E\{Y\}$ is the *mean response*. That is, the mean response $E\{Y\}$ is the mean of the probability distribution of Y corresponding to the level of the predictor variable X. When $\hat{Y}_i = b_0 + b_1 X_i$ (where $i = 1, 2, \ldots, n$), $\hat{Y}_i$ is the *fitted value* for the ith case ("trial"), whereas Y_i is the *observed value*.

Example 1.6.B. The estimated regression function of Example 1.6.A is $\hat{Y} = -5129.26 + 21.39X$. When the depth $X = 600$ cm, the point estimate of the age $\hat{Y} = -5129.26 + 21.39(600) = 7704.74$ years. That is, if several samples are taken at a depth of 600 cm then the average age of these samples is expected to be 7704.74 years. However, since there are error terms in the model reflecting the variability in the ages of samples at a given depth.

Note 1.6.B. In Note 1.3.C, we introduced the alternative form of simple linear regression model as $Y_i = \beta_0^* + \beta_1(X_i - \overline{X}) + \varepsilon_i$ where $\beta_0^* = \beta_0 + \beta_1\overline{X}$. By Theorem 1.6.A, $b_0 = \overline{Y} - b_1\overline{X}$, so that the least squares estimator of

$$b_0^* = b_0 + b_1\overline{X} = (\overline{Y} - b_1\overline{X}) + b_1\overline{X} = \overline{Y}.$$

The estimated regression model is the

$$\hat{Y} = b_0^* + b_1(X - \overline{X}) = \overline{Y} + b_1(X - \overline{X}).$$

In Example 1.6.A, we have $\overline{Y} = 6412.14$ and $\overline{X} = 539.57$ so that the estimated regression function in the alternative form is: $\hat{Y} = 6412.14 + 21.39(X - 539.57)$ (which simplifies to $\hat{Y} = -5129.26 + 21.39X$, the original estimated regression function).

Definition. The ith *residual* in a set of data is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$. This residual is denoted by e_i, so that $e_i = Y_i - \hat{Y}_i$.

Example 1.6.C. Of course the residual e_i indicates how far the data point (X_i, Y_i) is from point $(X_i, \hat{Y}_i)$, and hence reflects the vertical distance from a data point to the estimated regression function. For Example 1.6.A, we have the following residuals and residuals squared:

i	X_i	Y_i	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$	$e_1^2 = (Y_i - \hat{Y}_i)^2$
1	284.5	1255	956.20	298.8	89281.44
2	407.5	3390	3587.17	−197.17	38876.01
3	512.0	5560	5822.42	−262.42	68864.26
4	551.0	6670	6656.63	13.37	178.76
5	578.5	7160	7244.86	−84.86	7201.22
6	697.0	9820	9779.57	40.43	1634.58
7	746.5	11030	10838.38	191.62	36718.22
sum	3777	44885	44885.23	−0.23	242754.49

The sum of of the residuals, −0.23, should be 0, reflecting round off error.

Note. Notice that the model error term $\varepsilon_i = Y_i - E\{Y_i\}$ is unknown, whereas the residual $e_i = Y_1 - \hat{Y}_i$ is known from the data. We'll use residuals again in Chapter 3, "Diagnostics and Remedial Measure" (notice sections 3.2, 3.3, and 3.4). We now prove several properties of residuals and the estimated regression function. You may have noticed some of these properties in Examples 1.6.A and 1.6.C.

Theorem 1.6.B. For data points (X_i, Y_i) where $i = 1, 2, \ldots, n$, estimated regression model $\hat{Y} = b_0 + b_1 X$, and residuals $e_i = Y_i - \hat{Y}_i$, we have the following properties.

1. The sum of the residuals is zero: $\sum_{i=1}^{n} e_i = 0$.

2. The sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$, is a minimum.

3. The sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_i$: $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$.

4. The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial. That is, $\sum_{i=1}^{n} X_i e_i = 0$.

5. The sum of the weighted residuals is zero when the ith trial is weighted by the fitted value of the response variable for the ith trial. That is, $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$.

6. The regression line always goes through the point $(\overline{X}, \overline{Y})$.

Revised: 10/1/2022