STATISTICS

INFORMED DECISIONS USING DATA

Fifth Edition

Chapter 4

Describing the Relation between Two Variables

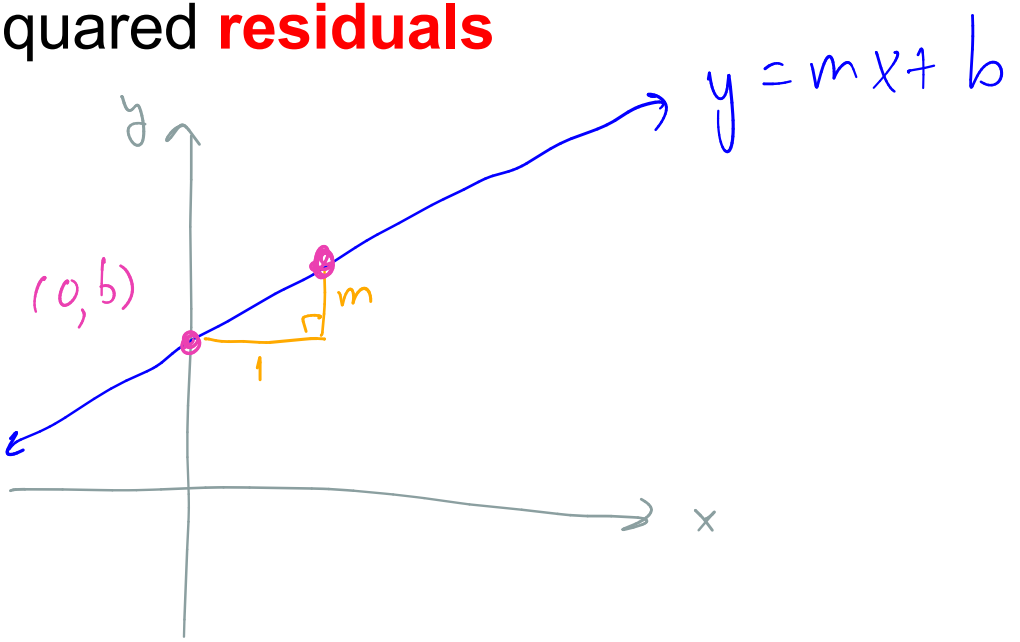
Pearson

4.2 Least-squares Regression

Learning Objectives

$\hat{y} = mx + b$
m slope
b y int

1. Find the **least-squares regression line** and use the line to make predictions
2. Interpret the **slope** and the ***y*-intercept** of the least-squares regression line
3. Compute the sum of squared **residuals**



4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (1 of 7)

(x, y) one data point

The difference between the observed value of *y* and the predicted value of *y* is the error, or residual.

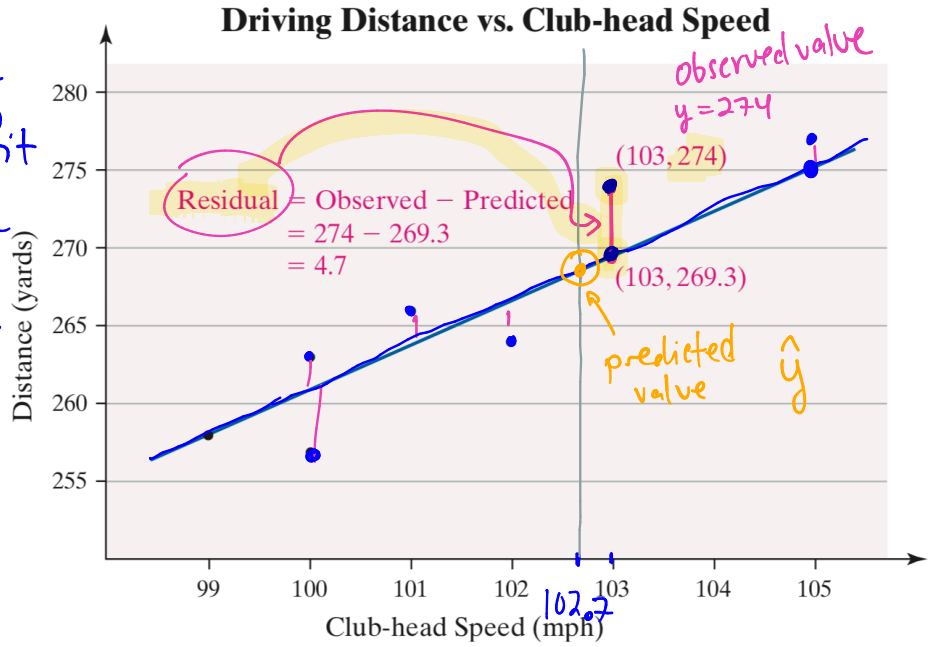
$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}$$

$$= 274 - 269.3 = 4.7$$

$\hat{y}$ "y hat"

- regression line
- line of best fit
- predicted line
- linear model

$\hat{y} = a + bx$



Notation:

- Observed: y
- Predicted: $\hat{y}$
- Residual: R

- y observed (data)
- $\hat{y}$ predicted (regression line $\hat{y} = a + bx$)
- R residual

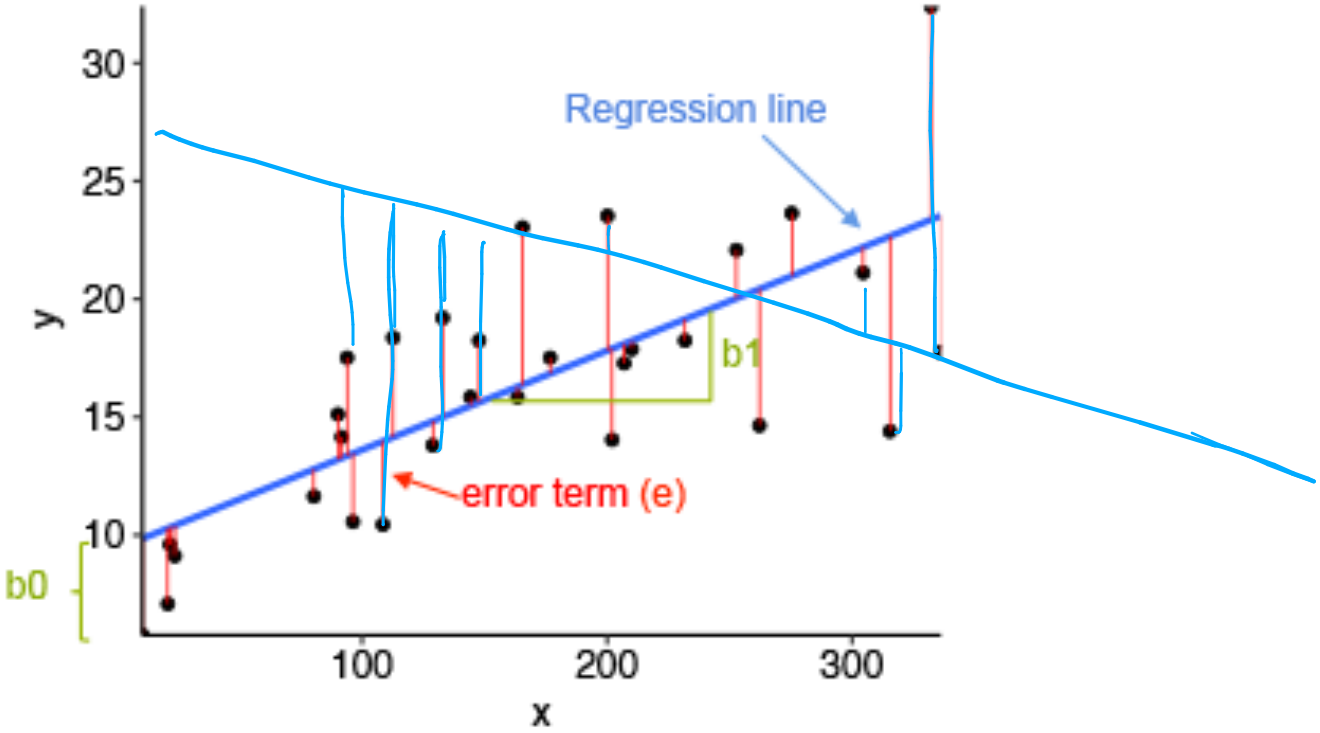
Pearson

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (2 of 7)

Least-Squares Regression Criterion

How to find regression line? Line of best fit, what does “best” mean here?



4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (2 of 7)

Least-Squares Regression Criterion

How to find regression line? Line of best fit, what does “best” mean here?

Key: goal is to minimize the total (sum) square of the ”errors” or residuals. Residuals are defined as the difference between the observed y-values and the predicted y-values.

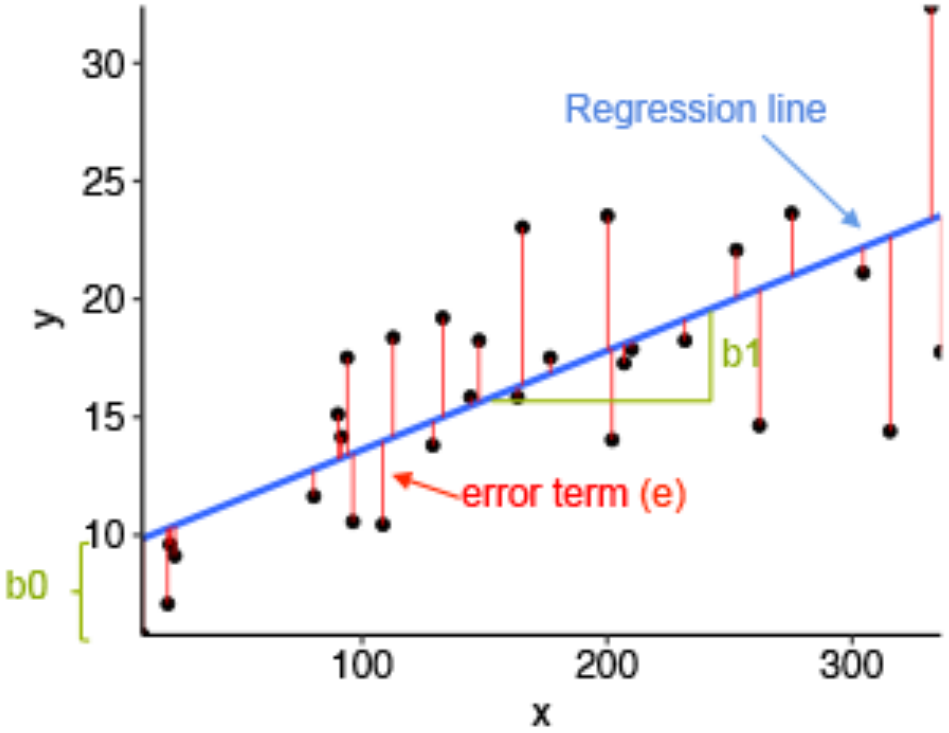
This minimizes the sum of the squares of the vertical distance between the observed values of y and those predicted by the line $\hat{y}$ (”y hat”).

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (2 of 7)

Least-Squares Regression Criterion

This is summarized: the line of best fit minimizes $\sum residuals^2$



4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (3 of 7)

The Least-Squares Regression Line

The **equation of the (least-squares) regression line** is given by

$$\hat{y} = b_1 x + b_0$$

Where

$b_1 = r \cdot \frac{s_y}{s_x}$ is the slope of the least-squares regression line

Where

$b_0 = \bar{y} - b_1 \bar{x}$ is the ***y*-intercept** of the least-squares regression line

Note: the least-squares regression line always contains the point $(\bar{x}, \bar{y})$.

Calculator:

$\hat{y} = a + bx$

$a = b_0$ y-int

$b = b_1$ slope

8: LinReg(ax+b)

Note 4: LinReg(ax+b)

4.2 Least-squares Regression
4.2.3 Compute the Sum of Squared Residuals

To illustrate the fact that the sum of squared residuals for a least-squares regression line is less than the sum of squared residuals for any other line, use the “regression by eye” applet.

>> Go to our website >> Handouts >> Textbook Additional Resources >> Additional Resources from Textbook >> Applets for the Student Activity Notebook >> “Regression by eye”

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (4 of 7)

The Least-Squares Regression Line

Note: $\bar{x}$ is the sample mean and s_x is the sample standard deviation of the explanatory variable x ;
$\bar{y}$ is the sample mean and s_y is the sample standard deviation of the response variable y.

Calculator:

$$\hat{y} = a + bx$$

- $a = b_0$
- $b = b_1$
- **>> Stat >> Calc >> 8.LinReg(a+bx)**
  - Note: turn "DiagnosticsON" in Catalog

Step 0 enter data into lists

also r

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (4 of 7)

The Least-Squares Regression Line

Calculator:

$$\hat{y} = a + bx$$

- $a = b_0$
- $b = b_1$

Rounding Rules

- We agree to round the y-int (a) and slope (b) to **FOUR decimal places**.

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (4 of 7)

$\hat{y} = a + bx$

Interpretation of predicted value: $\hat{y}$

The predicted value of y, $\hat{y}$, has an interesting interpretation. It is an estimate of the mean value of the response variable for any value of the explanatory variable. For example, suppose a least-squares regression equation is obtained that relates students' grade point average (GPA) to the number of hours studied each week. If the equation results in a predicted GPA of 3.14 when a student studies 20 hours each week, we would say the mean GPA of *all* students who study 20 hours each week is 3.14.

x = # hours studied

y = GPA

average gpa of all students who studied 20 hrs is 3.14.

$x = 20$ hours $\longrightarrow \hat{y} = 3.14$

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (5 of 7)

EXAMPLE Finding the Least-squares Regression Line

Using the drilling data

(a) Find the least-squares regression line.

(b) Predict the drilling time if drilling starts at 130 feet.

(c) Is the observed drilling time at 130 feet above, or below, average.

(d) Draw the least-squares regression line on the scatter diagram of the data.

Depth at Which Drilling Begins, x (in feet)	Time to Drill 5 Feet, y (in minutes)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (5 of 7)

Use Calc

a) $\hat{y} = a + bx$

$\hat{y} = 5.5273 + 0.0116x$ Regression Line

b) Predict $\hat{y}$

$x = 130$ feet

$\hat{y} = 5.5273 + 0.0116(130)$

$\hat{y} = 7.0353$ min

interpretation: average/mean time to drill 5 feet at 130 feet.

c) Observed $y = 6.93$ min

Observed value of 6.93 min is below the average of 7.0353 min.

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (6 of 7)

(a) We agree to round the estimates of the slope and intercept to four decimal places.

$$\hat{y} = 0.0116x + 5.5273$$

(b)
$$\begin{aligned}\hat{y} &= 0.0116x + 5.5273 \\ &= 0.0116(\mathbf{130}) + 5.5273 \\ &= 7.035\end{aligned}$$

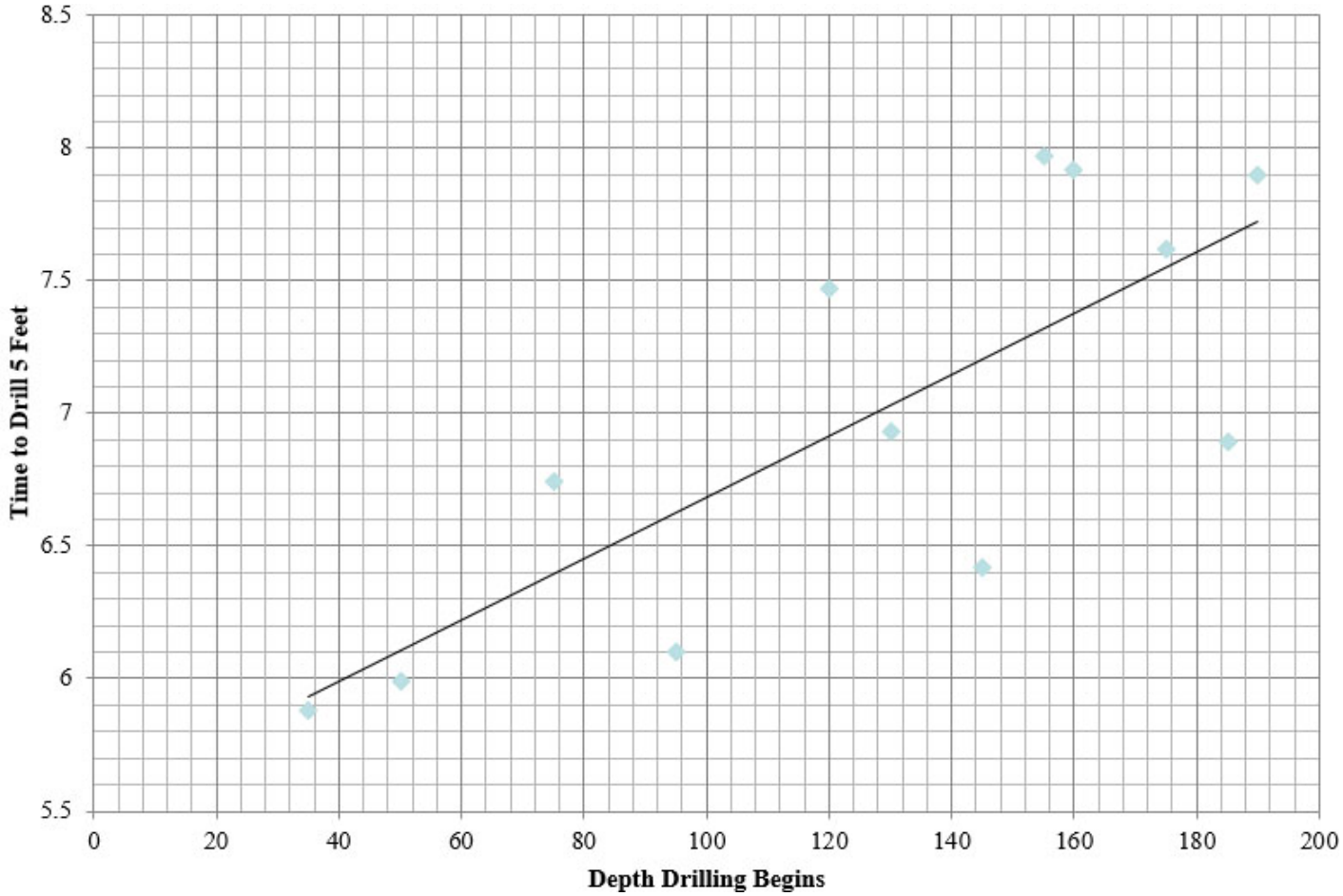
(c) The observed drilling time is 6.93 seconds. The predicted drilling time is 7.035 seconds. The residual is 6.93-7.035.

The drilling time of 6.93 seconds is below average (since residual is negative).

4.2 Least-squares Regression

4.2.1 Find the Least-Squares Regression Line and Use the Line to Make Predictions (7 of 7)

(d)



4.2 Least-squares Regression

4.2.2 Interpret the Slope and the *y*-Intercept of the Least-Squares Regression Line (1 of 3)

Interpretation of Slope:

b

The slope of the regression line is 0.0116.

For each additional foot of depth we start drilling, the time to drill five feet increases by 0.0116 minutes, on average.

$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{0.0116}{1}$ $\frac{\text{drill time (min)}}{\text{depth (feet)}}$

(Recall $4 = \frac{4}{1} = \frac{8}{2} = \frac{12}{3} = \frac{40}{10}$)

as increase x by 1, we increase y by 0.0116

as increase drill depth by 1 feet, we increase drill time by 0.0116 min.

as increase drill depth by 100 feet, we increase drill time by 1.16 min

4.2 Least-squares Regression

4.2.2 Interpret the Slope and the *y*-Intercept of the Least-Squares Regression Line (1 of 3)

Interpretation of Slope:

The slope of the regression line is 0.0116.

For each additional foot of depth we start drilling, the time to drill five feet increases by 0.0116 minutes, on average.

Recall: slope = rise/run = 4/1 = 8/2 = 12/3 = etc… (key is ratio is unchanged while rise and run can vary)

So: we can take slope b = 0.0116 = 1.16/100.

Interpret: For an additional 100 feet of depth, the time to drill 5 feet increases by 1.16 minutes, on average.

4.2 Least-squares Regression

4.2.2 Interpret the Slope and the *y*-Intercept of the Least-Squares Regression Line (2 of 3)

Interpretation of the *y*-Intercept:

The *y*-intercept of the regression line is 5.5273.

To interpret the *y*-intercept, we must first ask two questions:

1. Is 0 a reasonable value for the explanatory variable?
2. Do any observations near $x = 0$ exist in the data set?

4.2 Least-squares Regression

4.2.2 Interpret the Slope and the *y*-Intercept of the Least-Squares Regression Line (2 of 3)

Interpretation of the *y*-Intercept:

The *y*-intercept of the regression line is 5.5273. To interpret the *y*-intercept, we must first ask two questions:

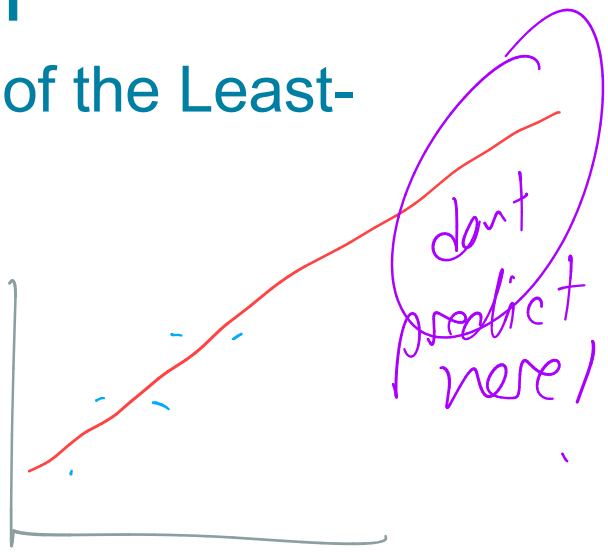
1. Is 0 a reasonable value for the explanatory variable?
2. Do any observations near $x = 0$ exist in the data set?

A value of 0 is reasonable for the drilling data (this indicates that drilling begins at the surface of Earth. The smallest observation in the data set is $x = 35$ feet, which is reasonably close to 0 for this data set. So, interpretation of the *y*-intercept is reasonable.

The time to drill five feet when we begin drilling at the surface of Earth is 5.5273 minutes.

4.2 Least-squares Regression

4.2.2 Interpret the Slope and the *y*-Intercept of the Least-Squares Regression Line (3 of 3)



If the least-squares regression line is used to make predictions based on values of the explanatory variable that are much larger or much smaller than the observed values, we say the researcher is working **outside the scope of the model**.

Never use a least-squares regression line to make predictions outside the scope of the model because we can't be sure the linear relation continues to exist.