CHAPTER 8

LEAST-SQUARE REGRESSION

Least-Squares Regression

Regression Line

- If correlation measures the **direction** and **strength** of the **linear** relationship between **two quantitative** variables, and a scatterplot shows a linear relationship, then we want to summarize this linear relationship by drawing a line on the scatter plot – this line is known as a **regression line**.

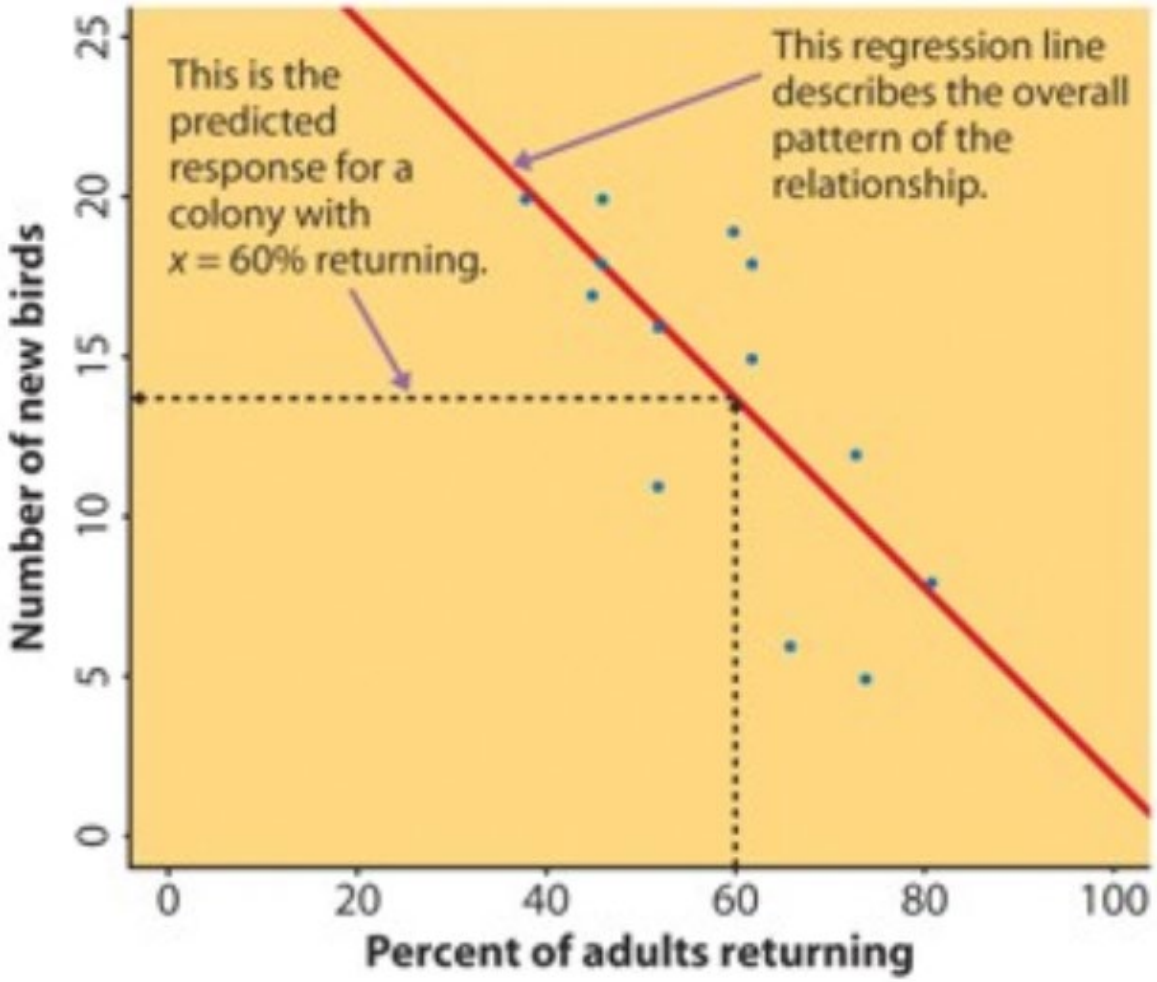
- **A regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes.

- We can use a regression line to predict the value of y for a given value of x.

Example

Predict the number of new adult birds that join the colony based on the percent of adult birds that return to the colony from the previous year.

- If 60% of adults return, How many new birds are predicted?

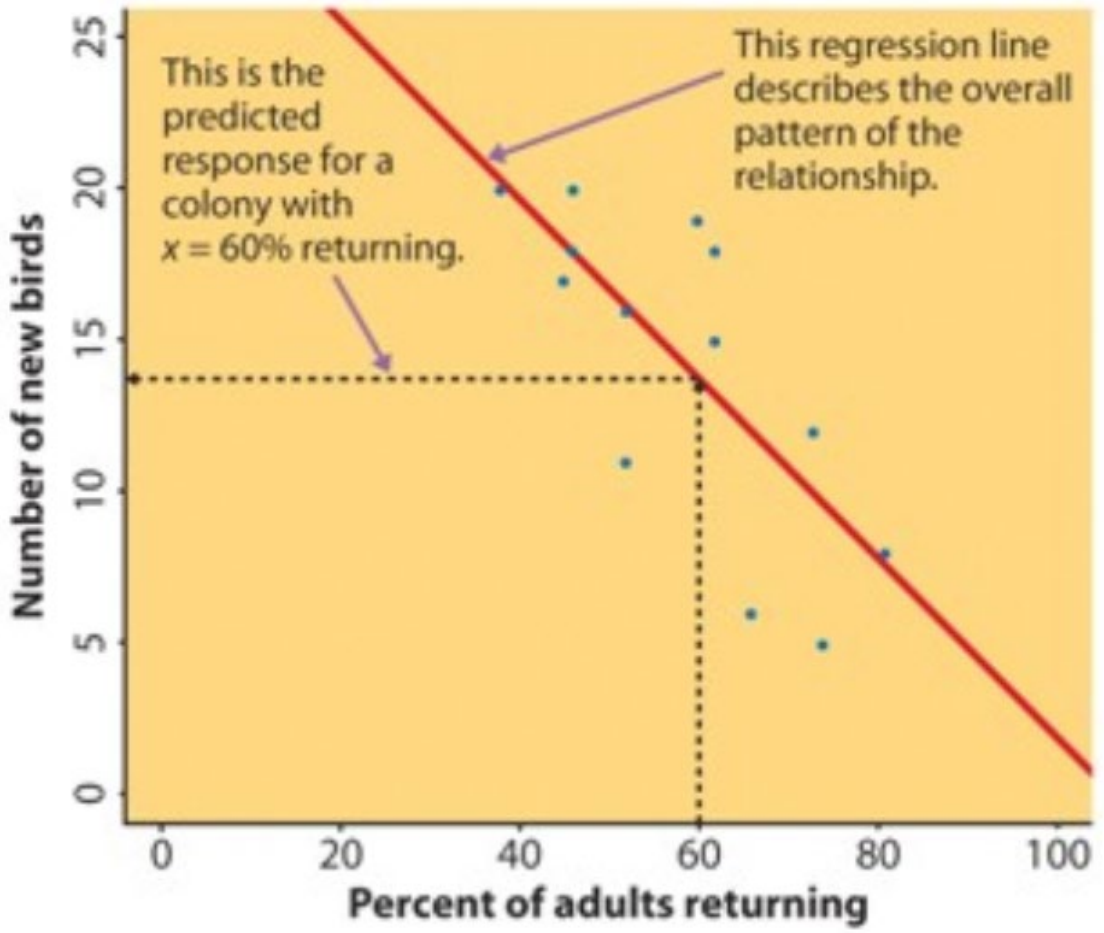


Example

If 10% of adults return, How many new birds are predicted?

Extrapolation: Is the use of a regression line for prediction far outside the range of values of the explanatory variable x used to obtain the line.

- Such predictions are often not accurate and should be avoided.



Regression Line

Equation of a straight line: $y = mx + b$

Equation of a regression line: $\hat{y} = b_0 + b_1 x$

- x is the value of the explanatory variable.
- $\hat{y}$ is the predicted value of the response variable for a given value of x.
- b_1 is the **slope,** the amount by which y changes for each one-unit increase in x.
- b_0 is the **intercept,** the value of y when $x = 0$.

Least Square Regression Line

Since we are trying to predict y, we want the regression line to be as close as possible to the data points in the vertical (y) direction.

Least-Squares Regression Line (LSRL):

The line that minimizes the sum of the squares of the vertical distances of the data points from the line.

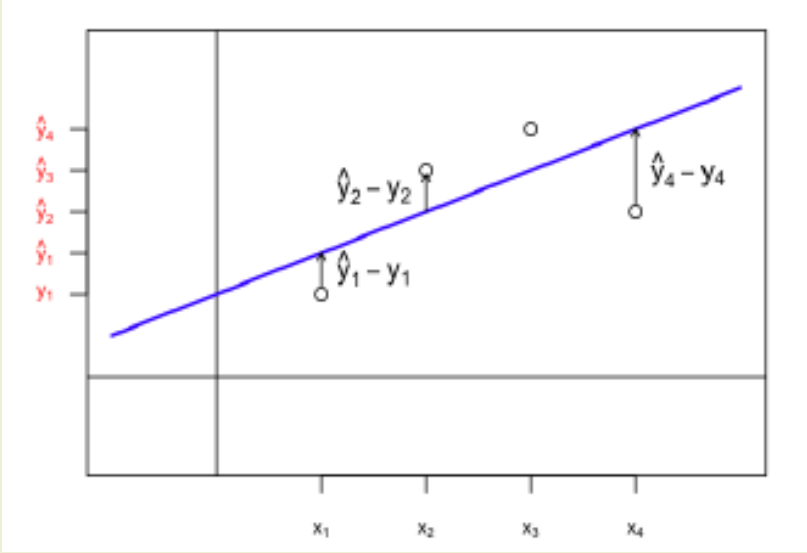
The least-squares regression line:

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

Where,

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$



Facts about Least –Square Regression

Fact 1: The LSRL always passes through (x-bar, y-bar)

Fact 2: The distinction between explanatory and response variables is essential.

Example: The relationship between body weight and backpack weight for a group of hikers.
Write the least squares regression line.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

Correlation and Regression

There is a close connection between correlation and regression

The **square of the correlation, r^2,** is the fraction of the variation in values of y that is explained by the least-squares regression of y on x.

- r^2 is called the **coefficient of determination.**

The coefficient of determination measure how well your regression equation truly represent your set of data.

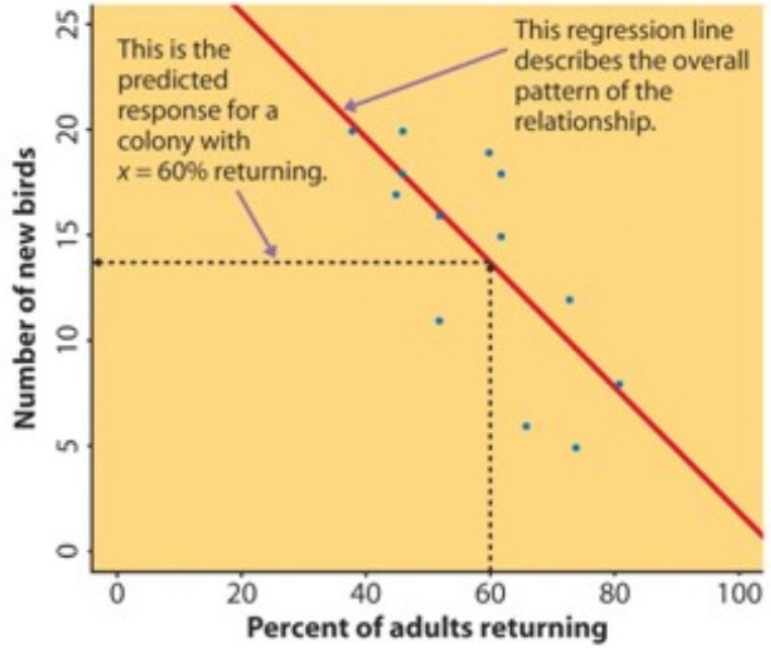
$R^2 = 1$, it is perfect fit, $R^2 = 0.5$, it is 50% fit,

Example

For the returning birds example, The LSRL is :

$$\hat{y} = 31.9343 - 0.3040x$$

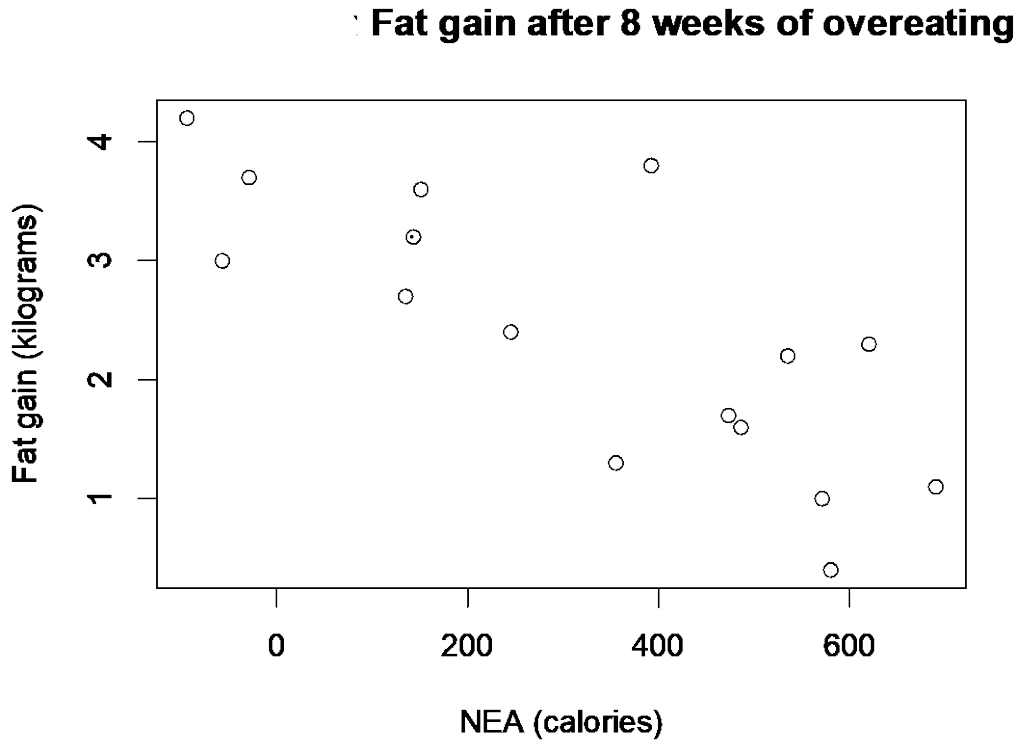
Suppose we know that an individual colony has 60% returning. What would we predict the number of new birds to be for just that colony?



Example

Does fidgeting keep you slim? Researchers deliberately overfed 16 healthy young adults for 8 weeks

- **Response:** fat gain (in kilograms)
- **Explanatory variable:** "non-exercise activity" (NEA), an increase in energy use (in calories) from activity other than deliberate exercise – fidgeting, daily living, and the like.



NEA increase (cal)	−94	−57	−29	135	143	151	245	355
Fat gain (kg)	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA increase (cal)	392	473	486	535	571	580	620	690
Fat gain (kg)	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Example continued…

- Verify from the data in that the mean and standard deviation of the 16 increases in NEA are as listed below.

$$\bar{x} = 324.8 \qquad \bar{y} = 2.388 \qquad r = -0.7786$$

$$s_x = 257.66 \qquad s_y = 1.1389$$

- Find the slope and intercept for the least-squares line.
- Write out the equation of the least-squares line.

Example Cont.

Cautions about Correlation and Regression

Residuals

In LSRL we want the vertical distances between the data points and the LSRL to have the smallest possible sum of squares.

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line:

Residual = Observed *y* – Predicted *y*

Residual Plots

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

- Ideally there should be a "random" scatter around zero.
  - The line residual = 0 corresponds to the fitted line: The residual plot magnifies the deviations from the line to make patterns easier to see.
- Residual ***patterns*** suggest deviations from a linear relationship.

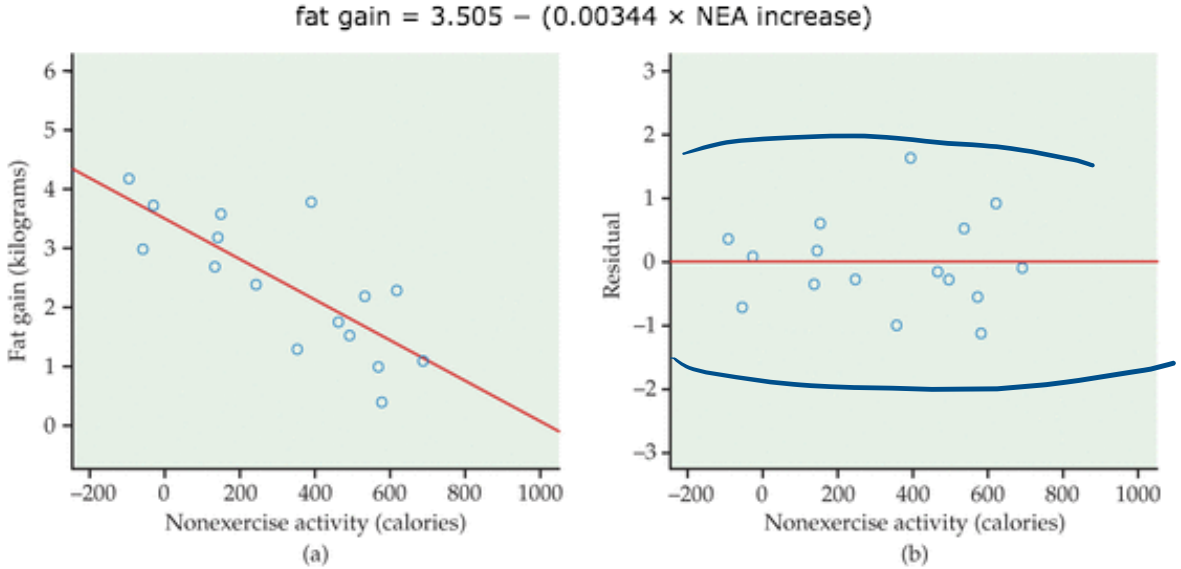


FIGURE 2.23 (a) Scatterplot of fat gain versus increase in nonexercise activity, with the least-squares regression line, for Example 2.25. (b) Residual plot for the regression displayed in (a). The line at $y = 0$ marks the mean of the residuals.

Outliers and Influential Points

- When looking at scatterplots and residuals, look for striking individual points as well as for an overall pattern.
- An ***outlier*** is an observation that lies outside the overall pattern of the other observations.
- Points that are outliers in the **y direction** of a scatterplot have large regression residuals.
- An observation is ***influential*** for a statistical calculation if removing it would markedly change the result of the calculation.

Cautions About Correlation and Regression

- Both describe linear relationships.
- Both are affected by outliers.
- Always plot the data before interpreting.
- Beware of ***extrapolation.***
  - Use caution in predicting *y* when *x* is outside the range of observed *x*'s.
- Beware of ***lurking variables.***
  - These have an important effect on the relationship among the variables in a study but are not included in the study.

Example

The data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class are:

Number of absences	6	2	15	9	12	5	8
Final grade	82	86	43	74	58	90	78

Find the equation of the least-square regression line.