# Regression (Page 125-128, Chapter 5)

**TODAY YOU WILL BE ABLE TO…**

* Quantify the linear relationship between an explanatory variable (x) and a response variable (y).
* Use a regression line to predict values of (y) for values of (x).

**Backpack Weight**

X: Body Weight (lb.)

Y: Backpack Weight (lb.)

Body weight “explains” backpack weight.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X: | 120 | 187 | 109 | 103 | 131 | 165 | 158 | 116 |
| Y: | 26 | 30 | 26 | 24 | 29 | 35 | 31 | 28 |

**THE REGRESSION LINE (also called the LEAST-SQUARES 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.

***Exercise 1:*** Using the scatterplot, predict the backpack weight of a 150-pound person.

Backpack is \_\_\_\_\_\_\_\_\_\_\_lbs.

**NOTATION**

***x*** is the value of the explanatory variable

pronounced ***“y-hat,”*** is the predicted value of the response variable for a given value of *x*

***b*** is the **slope,** the amount by which changes for each one-unit increase in *x*

***a*** is the **intercept,** the value of *y* when *x = 0*

the regression equation

Many lines could be drawn “by sight” through the set of points on a scatterplot; however, there is one line that is the best-fit through the data. That is the line defined as How this line is determined will be discussed later. For now, it is important to be able to use the line to make predictions.

**USE THE REGRESSION LINE TO PREDICT THE RESPONSE**

***Exercise 2:*** The regression line for predicting backpack weight from body weight is

If x=150…

**What does the slope tell you?**

On average, backpack weight increases by 0.0908 pounds for every 1 pound of body weight. The slope does **not** tell you how strong a relationship is.

**The intercept is not statistically meaningful.**

The intercept is necessary to make accurate predictions, however, it has no statistical meaning.

For example, when body weight is 0, There is no such thing as a zero-pound person. You could count an embryo as a person; however, such a person will not be carrying a 16-pound backpack.

Using the regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line is called **extrapolation**.

***Exercise 3:*** We expect a car’s highway gas mileage to be related to its city gas mileage. Data for all 1040 vehicles in the government’s 2010 *Fuel Economy Guide* give the regression line:

Highway MPG = 6.554 + (1.016 × City MPG)

1. What is the slope of this line?
2. What does the slope tell you?
3. What is the intercept? What does it mean?
4. Find the predicted highway mileage for a car that get 16 miles per gallon in the city.
5. Find the predicted highway mileage for a car that get 28 miles per gallon in the city.

# Regression (Page 128-134, Chapter 5)

**TODAY YOU WILL BE ABLE TO…**

* Describe how the line of best fit is determined
* Calculate the slope and intercept of the regression line

**THE LEAST-SQUARES REGRESSION LINE**

Just as each data value x is “some” distance away from its mean, each data point (x, y) on the scatterplot is “some” distance away from the regression line.



Just as the mean is the “center” of the values of a single-variable set x, the best-fitting line is the “center” of the values of a two-variable set (x, y).

Just as the sum of the differences between each point x and the mean in a single-variable set is \_\_\_\_\_\_\_\_, the sum of the vertical distances between each point (x, y) and the regression line in a two-variable set is also \_\_\_\_\_\_\_\_.

The best-fitting line has the smallest “sum of squared vertical distances” of any line that could be drawn through the data. Hence the term, least-squares regression line.

You can calculate the slope and intercept of the least-squares regression line from what you know about each variable, x and y.

**Necessary Calculations**

* The correlation, r, between x and y
* The mean of x, , and the mean of y,
* The standard deviation of x, sx, and the standard deviation of y, sy

is the least-squares regression line with

, “y-hat,” is used to represent the predicted value of y because of the scatter of points about the line. The predicted response will usually not be the same as the actual observed response y.

**Exercise 4:** Scientists have examined data on mean sea surface temperatures, in Celsius, and mean coral growth, in millimeters per year, over a several year period at locations in the red sea.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sea Surface Temperature | 29.68 | 29.87 | 30.16 | 30.22 | 30.48 | 30.65 | 30.90 |
| Growth | 2.63 | 2.58 | 2.60 | 2.48 | 2.26 | 2.38 | 2.26 |

Calculate…

**FACTS ABOUT LEAST-SQUARES REGRESSION**

1. **The distinction between explanatory and response variables is essential.**

Least-squares regression makes the sum of squared distances from the line small only in the y-direction. If you reverse the roles of the variables, you get a different regression line.

1. **The slope b and correlation r always have the same sign.**

Positive slope = positive relationship

Negative slope = negative relationship

1. **The least-squares regression line always passes through the point   
   (, ).**

This is simply a consequence of calculating the least-squares regression line.

1. **The square of the correlation, r2, is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.**

The correlation r describes the strength of a linear relationship. In the regression setting, the explanatory variable x “explains” the response variable y. The square of the correlation, r2, is the fraction or portion of the variability in y that is explained by the explanatory variable.

1. **r2 is called the coefficient of determination.**

***Exercise 5:*** Sea surfacetemperatures explainwhat fraction of the variability in coral reef growth?

**Scatterplot with least-squares regression line for Coral Growth vs. Sea Surface Temperature**



# Regression (Page 135-146, Chapter 5)

**TODAY YOU WILL BE ABLE TO…**

* Interpret residuals.
* Recognize influential observations.
* Describe cautions about correlation and regression.

**INTERPRETING RESIDUALS**

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

Values of the response, y, that are below the least-squares regression line, i.e., their value is smaller than the corresponding predicted , have negative residuals. Values of y that are above the line have positive residuals.

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable and is used to assess the fit of a regression line. A point on the zero line of this plot indicates that the observed value lies directly on the least-squares line.

When looking at a residual plot, look for a “random” scatter of points around the zero line, i.e., there is no pattern to the values of the residuals.



**Residual Plot for Coral Growth vs. Sea Surface Temperature**

**INFLUENTIAL OBSERVATIONS**

An outlier is an observation that lies far away from the other observations. In regression the observations are represented by coordinate pairs, (x, y). A point (xi, yi) is an outlier if it lies far away from the other observations on the scatterplot.

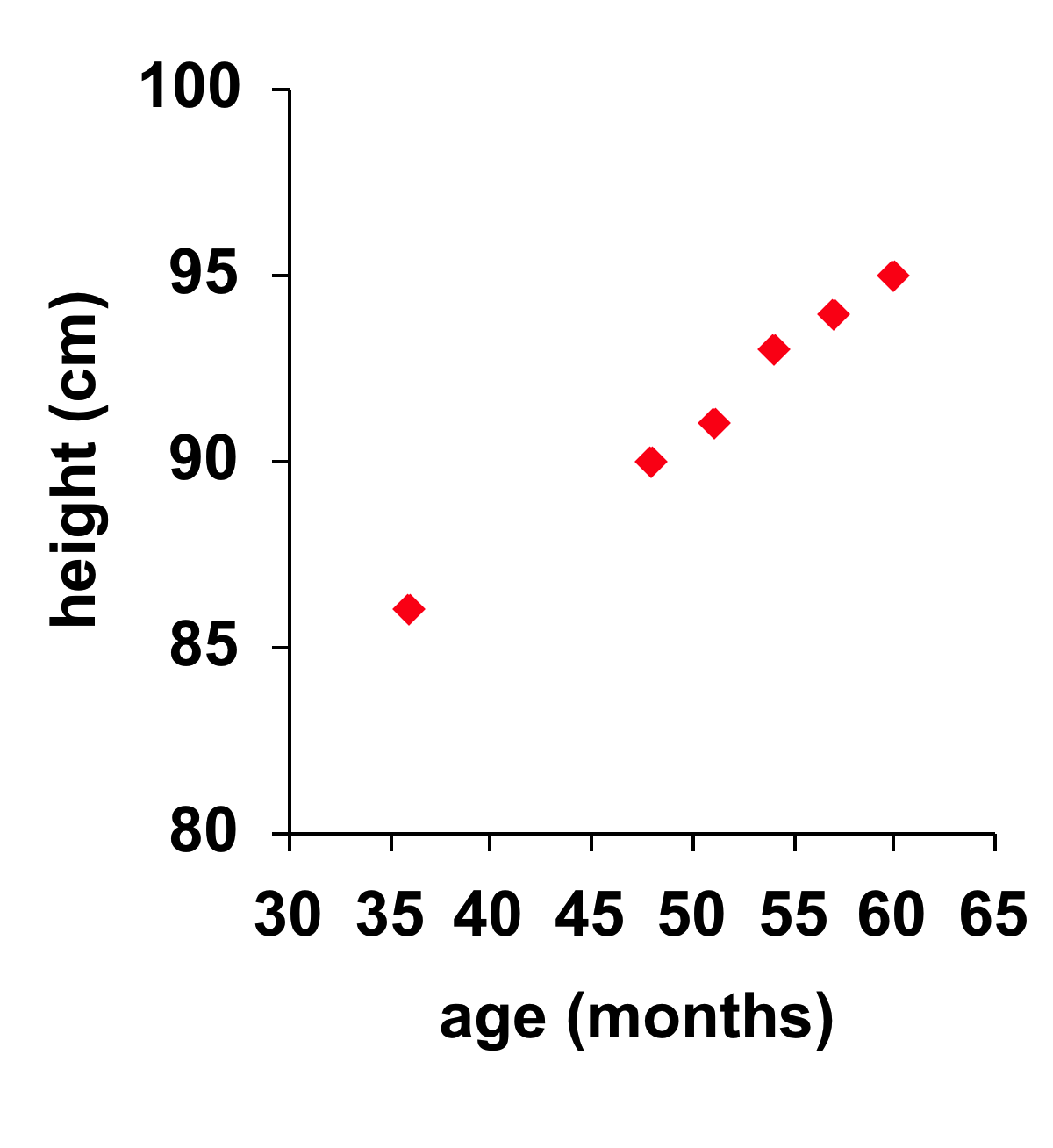
* The x-value of an outlying point in regression may not be an outlier with respect to the other explanatory values.
* The y-value of an outlying point in regression may not be an outlier with respect to the other response values.

**Directions of outliers**

* Outliers in the y direction have large residuals.
* Outliers in the x direction are often **influential** for the least-squares regression line, meaning that the removal of such points would markedly change the equation of the line.

**GUIDELINES ABOUT CORRELATION AND REGRESSION**

* Always plot the data before interpreting.
* Beware of extrapolation (predicting outside of the range of x)
  + Sarah’s height was plotted against her age. The regression line is
  + Can you predict her height at age 42 months?
  + Can you predict her height at age 30 years (360 months)?



* Beware of lurking variables.
  + These have an important effect on the interpretation of relationships among the variables in a study, but are not included in the study.
* Remember that correlation does not imply causation!
  + Even very strong correlations may not correspond to a real causal relationship (changes in x actually causing changes in y).
  + Correlation may be caused by one or more lurking variables
  + Only a properly conducted experiment may establish causation.

**Social Relationships and Health**

*House, J., Landis, K., and Umberson, D. “Social Relationships and Health,” Science, Vol. 241 (1988), pp 540-545.*

A strong correlation was found between lack of social relationships and illness. Does lack of social relationships cause people to become ill?

* Or, are unhealthy people less likely to establish and maintain social relationships? (reversed relationship)
* Or, is there some other factor that predisposes people both to have lower social activity and become ill?