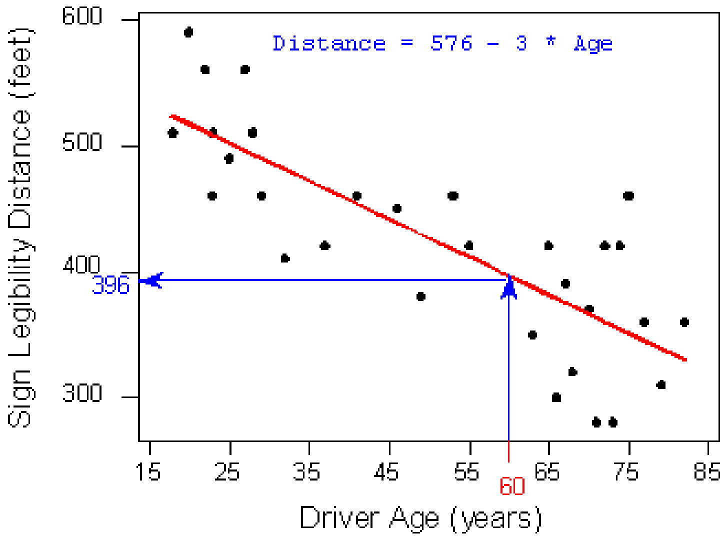
# Linear Relationships: Prediction

**Learning Objective: In the special case of linear relationship, use the least squares regression line as a summary of the overall pattern, and use it to make predictions.**

Let's go back now to our motivating example, in which we wanted to predict the maximum distance at which a sign is legible for a 60-year-old. Now that we have found the least squares regression line, this prediction becomes quite easy:



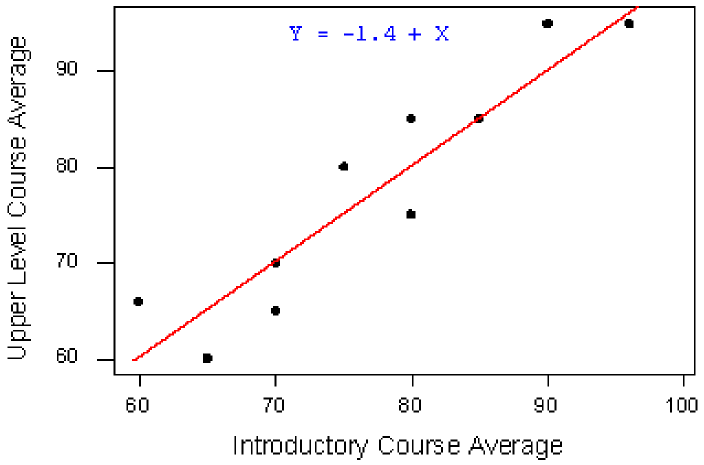
Practically, what the figure tells us is that in order to find the predicted legibility distance for a 60-year-old, we plug Age = 60 into the regression line equation, to find that:

Predicted distance = 576 + (- 3 * 60) = 396

396 feet is our best prediction for the maximum distance at which a sign is legible for a 60-year-old.

## Scenario: Progress Until Graduation

**Background:** A statistics department is interested in tracking the progress of its students from entry until graduation. As part of the study, the department tabulates the performance of 10 students in an introductory course and in an upper-level course required for graduation. The scatterplot below includes the least squares line (the line that best explains the upper-level course average based on the lower-level course average), and its equation:



## Did I Get This

1/1 point (graded)

What is the slope of the regression line?

○ 1.4

○ −1.4

● 1 ✓

○ −1

**Answer**

Correct:

The equation of the line has the following form: Y = intercept + (slope * X). The number before X is 1 (1 * X = X).

Submit

## Did I Get This

1/1 point (graded)

Which of the following is the correct interpretation of the slope of the regression line?

○ A student who performs 1 point better than another on the upper-level course is likely to perform 1 point worse in the introductory course.

○ A student who performs 1 point worse than another on the upper-level course is likely to perform 1 point better in the introductory course.

○ A student who performs 1 point better than another in the introductory course is likely to perform 10 − 1.4 = 8.6 points better in the upper-level course.

● A student who performs 1 point better than another on the introductory course is likely to perform 1 point better in the upper-level course. ✓

**Answer**

Correct:

The slope is the change we would expect in the response variable for an increase of 1 unit in the explanatory variable. A positive slope of 1.0 for the regression line means that every change of 1 unit in the explanatory variable leads us to expect a change of 1 unit in the response variable.

Submit

## Did I Get This (1/1 point)

A student scored an average of 90 in the introductory course. What score can we expect for this student in the upper-level course?

**Your Answer:**

88.6

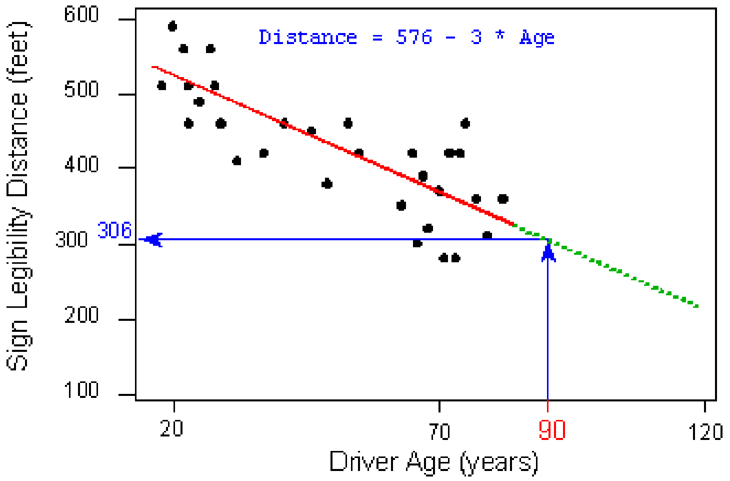
**Our Answer:**

Using the regression line, the correct answer is -1.4 + 90 = 88.6

Resubmit Reset

## Comment About Predictions

Suppose a government agency wanted to design a sign appropriate for an even wider range of drivers than were present in the original study. They want to predict the maximum distance at which the sign would be legible for a 90-year-old. Using the least squares regression line again as our summary of the linear dependence of the distances upon the drivers' ages, the agency predicts that 90-year-old drivers can see the sign at no more than 576 + (- 3 * 90) = 306 feet:



(The green segment of the line is the region of ages beyond 82, the age of the oldest individual in the study.)

**Question:** Is our prediction for 90-year-old drivers reliable?

**Answer:** Our original age data ranged from 18 (youngest driver) to 82 (oldest driver), and our regression line is therefore a summary of the linear relationship **in that age range only.** When we plug the value 90 into the regression line equation, we are assuming that the same linear relationship extends beyond the range of our age data (18-82) into the green segment. **There is no justification for such an assumption.** It might be the case that the vision of drivers older than 82 falls off more rapidly than it does for younger drivers. (i.e., the slope changes from -3 to something more negative). Our prediction for age = 90 is therefore **not reliable.**

## In General

Prediction for ranges of the explanatory variable that are not in the data is called **extrapolation**. Since there is no way of knowing whether a relationship holds beyond the range of the explanatory variable in the data, extrapolation is not reliable, and should be avoided. In our example, like most others, extrapolation can lead to very poor or illogical predictions.