⚠ Lagunita is retiring and will shut down at 12 noon Pacific Time on March 31, 2020. A few courses may be open for self-enrollment for a limited time. We will continue to offer courses on other online learning platforms; visit http://online.stanford.edu.

Course > EDA: Examining Relationships > Case Q→Q: Linear Relationships > Linear Relationships: Regression

🔖 Bookmark this page

Linear Relationships: Regression

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.

Linear Regression: Summarizing the Pattern of the Data with a Line

So far we've used the scatterplot to describe the relationship between two quantitative variables, and in the special case of a linear relationship, we have supplemented the scatterplot with the correlation (r). The correlation, however, doesn't fully characterize the linear relationship between two quantitative variables—it only measures the strength and direction. We often want to describe more precisely how one variable changes with the other (by "more precisely," we mean more than just the direction), or **predict** the value of the response variable for a given value of the explanatory variable. In order to be able to do that, we need to summarize the linear relationship with a line that best fits the linear pattern of the data. In the remainder of this section, we will introduce a way to find such a line, learn how to interpret it, and use it (cautiously) to make predictions.

Again, let's start with a motivating example:

Earlier, we examined the linear relationship between the age of a driver and the maximum distance at which a highway sign was legible, using both a scatterplot and the correlation coefficient. Suppose a government agency wanted to predict the maximum distance at which the sign would be legible for 60-year-old drivers, and thus make sure that the sign could be used safely and effectively.

How would we make this prediction?

Making Predictions

⬛⬛⬛⬛⬛⬛⬛⬛⬛⬛⬛	Start of transcript. Skip to the end.
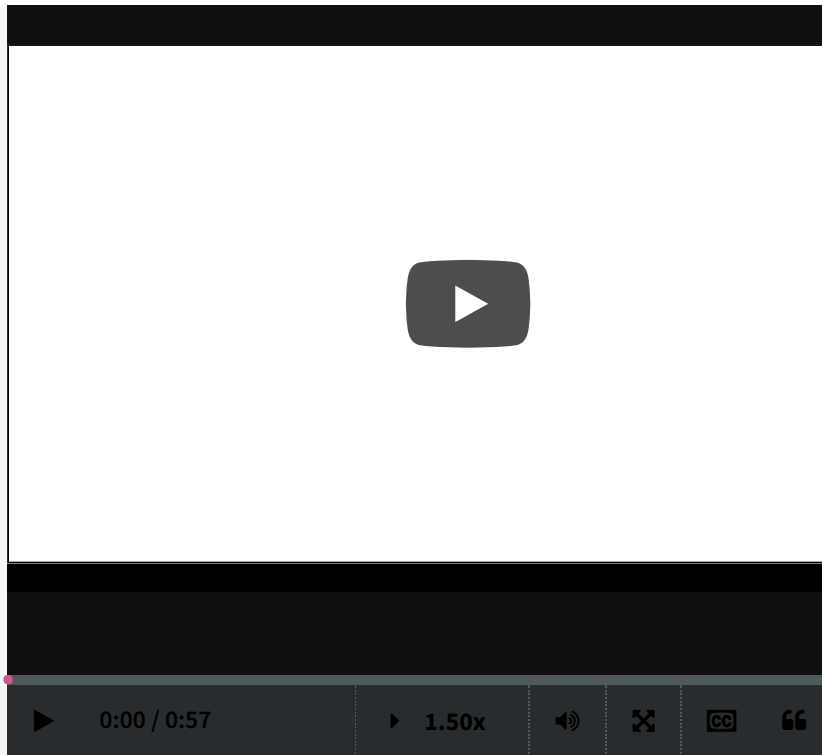
Here again is the scatterplot of the drivers' ages and maximum distances at which

a sign is legible. The age for which an agency wishes to predict the legibility

distance, 60, is marked in red. How can we make this prediction? It would be useful if we

could find a line (such as the one that is presented on the scatterplot)

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

How and why did we pick this particular line (the one shown in red in the above video) to describe the dependence of the maximum distance at which a sign is legible upon the age of a driver? What line exactly did we choose? We will return to this example once we can answer that question with a bit more precision.

Open Learning Initiative