Lecture 19

We've seen how to explore the relationship between two quantitative variables graphically with a scatterplot. When the relationship has a straight-line pattern, the correlation describes it numerically. We can analyze the data further by finding an equation for the straight line that best describes that pattern. This equation predicts the value of the response variable from the value of the explanatory variable.

Linear Regression

The **regression line** predicts the value for the response variable y as a straight-line function of the value x of the explanatory variable. Let $\hat{y}$ denote the **predicted value** of y. The equation for the regression line has the form

$$\hat{y} = a + bx \tag{1}$$

where a denotes the **y-intercept** and b denotes the **slope**.

Example:
Anthropologists often try to reconstruct information using partial human remains at burial sites. For instance, after finding a femur (thighbone), they may want to predict how tall an individual was. An equation they use to do this is the regression line,

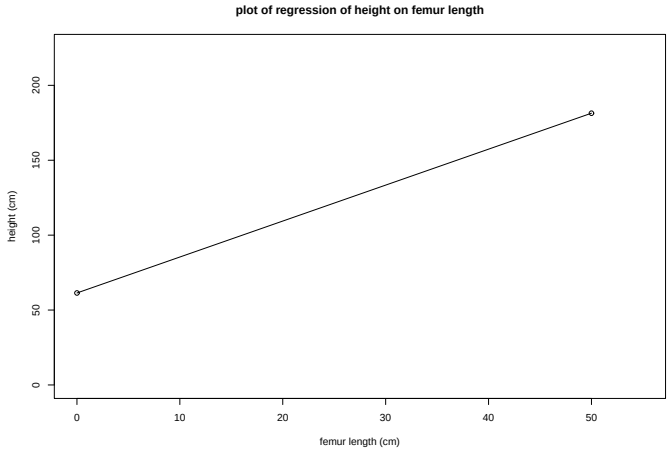
$$\hat{y} = 61.4 + 2.4x$$

where $\hat{y}$ is the predicted height and x is the length of the femur, both in centimeters.

At $x = 0$, $\hat{y} = 61.4 + 2.4(0) = 61.4$, the y-intercept. In practice there would not be a femur of length 0 but this value represents the baseline or starting point for the equation.

At $x = 0$, $\hat{y} = 61.4 + 2.4(50) = 181.4$. When the femur is 50 cm, the predicted height of the person is 181.4 cm.

We can plot the line $\hat{y} = 61.4+2.4x$ by connecting the two coordinates (0,61.4) and (50,181.4).



Interpreting the y-intercept and the slope

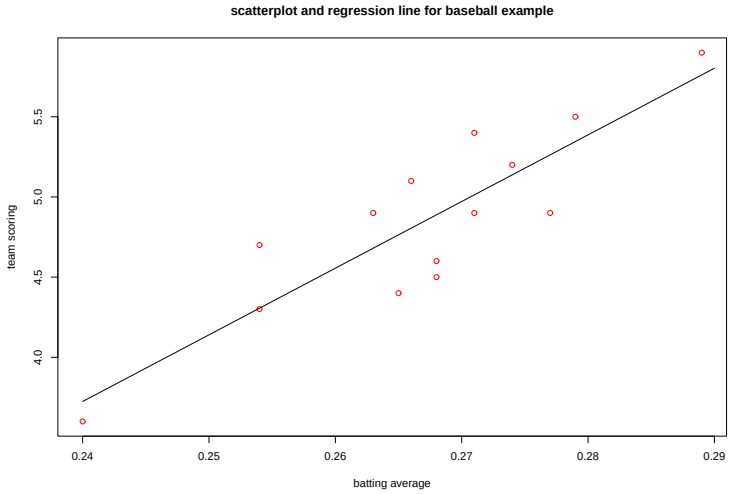
The y-intercept a is the predicted value of y when $x = 0$. This fact helps us plot the line, but it may not have any interpretative value if no overvations had x values near 0. It does not make sense for femur length to be 0 cm, so the y-intercept for the equation $\hat{y} = 61.4 + 2.4x$ is not a relevant predicted height.

The slope b in the equation $\hat{y} = a + bx$ equals the amount that $\hat{y}$ changes when x increases by one unit. For two x values that differ by 1.0, the $\hat{y}$ values differ by b. For the line $\hat{y} = 61.4 + 2.4x$, we've seen that $\hat{y} = 181.4$ for $x = 50$. If x increases by 1.0 to $x = 51$, we get $\hat{y} = 61.4 + 2.4(51) = 183.8$. The increase in $\hat{y}$ from 181.4 to 183.8 is 2.4 which is the slope. Therefore, for each 1 cm increase in femur length, height is predicted to increase by 2.4cm.

Example:
In baseball, two summaries of a team's offensive ability are the teams batting average (the proportion of times the team's players get a hit, out of the times they get a hit or instead are "out") and team scoring (the team's mean number of runs scored per game. We have the data below for the 2003 season for American League teams.

Team	**Batting Average**	**Team Scoring**
Boston	0.289	5.9
Toronto	0.279	5.5
Minnesota	0.277	4.9
Kansas City	0.274	5.2
Seattle	0.271	4.9
New York	0.271	5.4
Anaheim	0.268	4.5
Baltimore	0.268	4.6
Texas	0.266	5.1
Tampa Bay	0.265	4.4
Chicago	0.263	4.9
Oakland	0.254	4.7
Cleveland	0.254	4.3
Detroit	0.240	3.6

Scoring runs is a result of hitting, so team scoring is the response variable y and team batting average is the explanatory variable x. The scatterplot below shows a straight-line trend summarized by a strong positive correlation, $r = 0.875$.



This line represents the equation $\hat{y} = -6.1 + 41.2x$ so the y-intercept $a = -6.3$ and the slope $b = 41.2$. We can predict that an American League team with a team batting average of 0.275 will score an average of $\hat{y} = -6.1 + 41.2(0.275) = 5.34$ runs per game.

Since the slope $b = 41.2$ is positive, the association is positive: The predicted team scoring increases as team batting average increases. The slope refers to the change in $\hat{y}$ for a 1-unit change in x. However, x is a proportion. The team batting averages fall between about 0.24 and 0.29, a range of 0.05. An increase of 0.05 in x corresponds to an increase of (0.05)41.2=2.1 in predicted team scoring. Therefore, the mean number of scores per game is predicted to be about 2 higher for the best hitting teams than for the worst hitting teams.

Residuals

The regression equation $\hat{y} = -6.1 + 41.2x$ predicts team scoring for a given level of x=team batting average. We can compare the predicted values to the actual team scoring to check the accuracy of those prediction.

For example, New York had values

$$y = 5.4 \; and \; x = 0.271 \tag{2}$$

The prediction for y=mean number of runs per game at $x = 0.271$ is

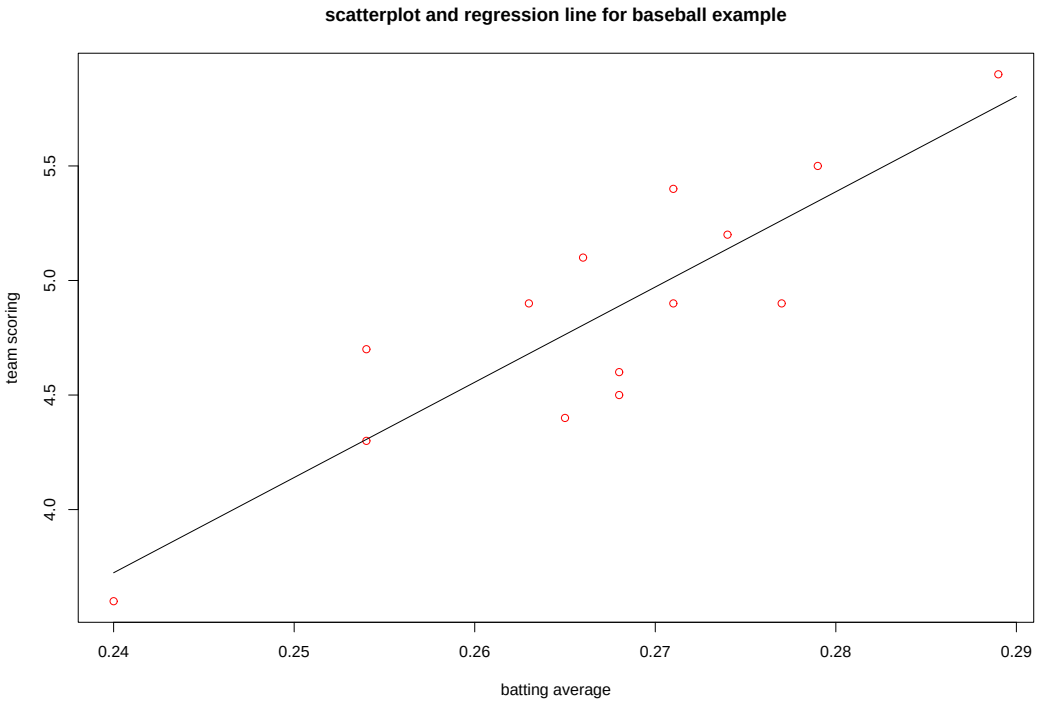
$$\hat{y} = -6.1 + 41.2(0.271) = 5.1 \tag{3}$$

The **residual** or prediction error is the difference between the actual y value and the predicted $\hat{y}$ value at a certain value of x. Here it is

$$residual \; = \; y - \hat{y} \tag{4}$$

$$5.4 - 5.1 = 0.3 \tag{5}$$

For New York, the regression equation under predicts y by 0.3 runs per game.



Example:
Find the residuals for Tampa Bay and mark them on the scatterplot.

The method of least-squares yields the regression line

We choose the optimal line to fit through the data by making the residuals as small as possible. The summary measure used to evaluate predictions is

$$residual\ sum\ of\ squares = \sum_{i=1}^{n}(y - \hat{y})^2 \tag{6}$$

This squares each vertical distance between a point and the line and then adds them up. The line that minimizes the residual sum of squares more than any other line is the regression line selected through the **least squares method**.

The method of least squares provides formulas for the y-intercept and slope, based on statistics for the sample data. Let $\bar{x}$ denote the mean of x, $\bar{y}$ denote the mean of y, σ_x denote the standard deviation of x, σ_y denote the mean of y. The slope b is directly related to the correlation r.

$$b = r\frac{\sigma_y}{\sigma_x} \tag{7}$$

$$a = \bar{y} - b\bar{x} \tag{8}$$

Example:
In the baseball example, $\bar{x} = 0.267$, $\bar{y} = 4.865$, $\sigma_x = 0.0121$, $\sigma_y = 0.575$ and $r = 0.875$. So,

$$b = r\frac{\sigma_y}{\sigma_x} = (0.875)\frac{0.568}{0.0121} = 41.2$$
$$a = \bar{y} - b\bar{x} = 4.865 - (41.2)0.267 = -6.1$$

which are the results in the given regression line $\hat{y} = -6.1 + 41.2x$.