The Coefficient of Determination

The **coefficient of determination** is the proportion of variation in the observed values of the response variable explained by the regression.

Sums of Squares

- **Total sum of squares (SST)**: the total variation in the observed values of the response variable.

$$SST = \sum (y_i - \bar{y})^2$$

- Note the connection to standard deviation!

Sums of Squares

- **Regression sum of squares (SSR)**: the variation in the observed values of the response variables explained by the regression.

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

- **Error sum of squares (SSE)**: the variation in the observed values of the regression variable not explained by the regression.

$$SSE = \sum(y_i - \hat{y}_i)^2$$

Coefficient of Determination

The **coefficient of determination,** r^2, is the proportion of variation in the observed values of the response variable explained by the regression.

$$r^2 = \frac{SSR}{SST}$$

This value will always be between 0 and 1.

Regression Identity

The **regression identity** states

$$SST = SSR + SSE.$$

This allows us to find r^2 in other ways:

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

Regression Identity

We can also interpret r^2 as the percent reduction in total squared error when we use a regression instead of $\bar{y}$ to make predictions.

Values of r^2 near 1 suggest that the independent variable is quite useful in predicting values of the dependent variable.
(Values near 0 suggest the opposite.)

Correlation

The **correlation** between two variables describes the strength of their linear relationship. It always takes values between -1 and 1.

Correlation

We denote the correlation (or correlation coefficient) by r:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y} \right)$$

where s_x and s_y are the respective standard deviations for x and y.

Correlation

The correlation coefficient r is $\sqrt{r^2}$, but we need some additional information to use this!

With information about the slope, we can determine whether r is positive or negative.

The sign of the correlation will match the sign of the slope!

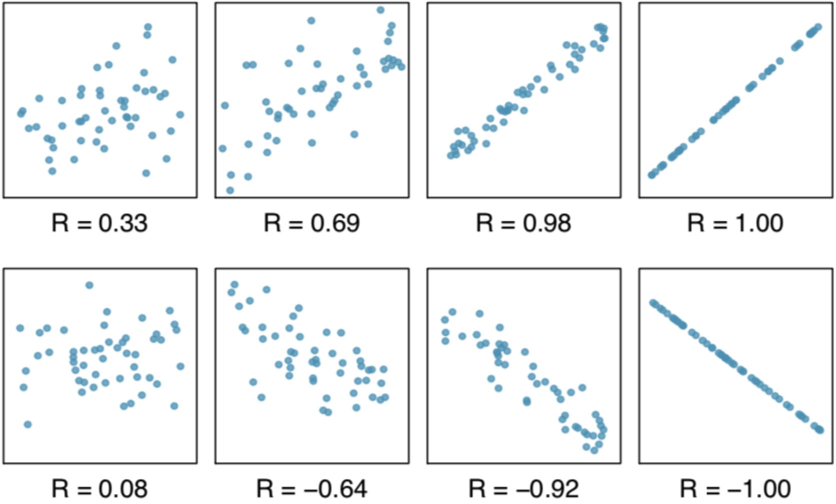
- If $r < 0$, there is a downward trend and $b_1 < 0$.
- If $r > 0$, there is an upward trend and $b_1 > 0$.
- If $r \approx 0$, there is no relationship and $b_1 \approx 0$.

Correlation

Correlations

- Close to -1 suggest strong, negative linear relationships.
- Close to +1 suggest strong, positive linear relationships.
- Close to 0 have little-to-no linear relationship.

Correlation



Correlations

Correlations only represent *linear* trends!