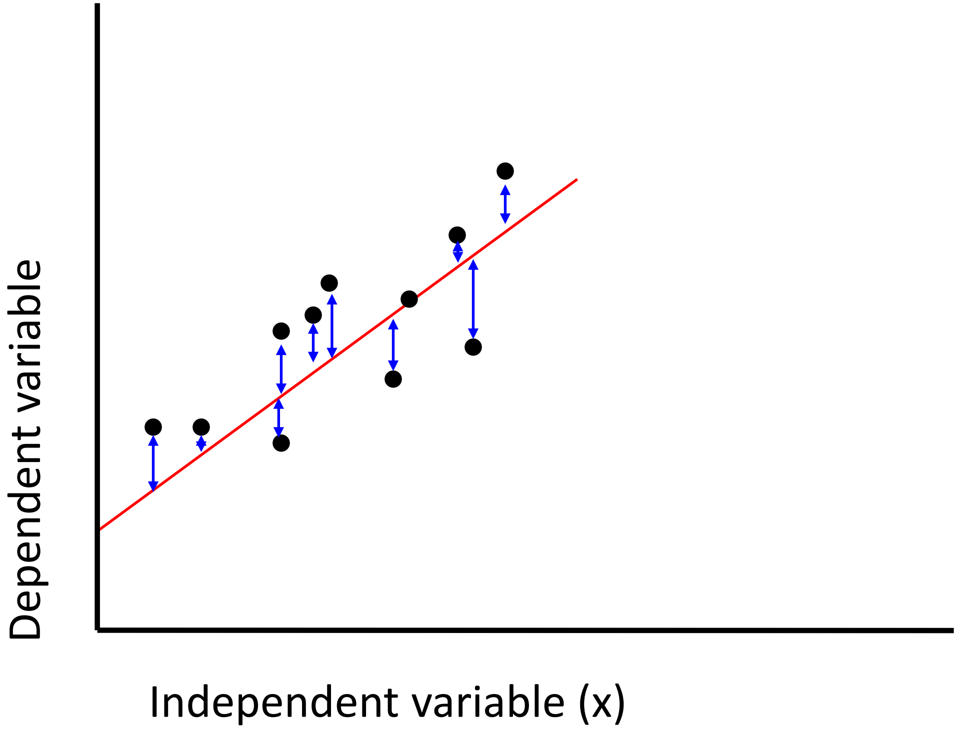
Capturing Relationship between Two Quantities

Measures of Accuracy

Assessing the Model

- The least squares method will produce a regression line whether or not there is a linear relationship between x and y.
- Consequently, it is important to assess how well the linear model fits the data.
- Several methods are used to assess the model:
  - Testing and/or estimating the coefficients.
  - Using descriptive measurements.

Sum of Squares of Errors (SSE)



A least squares regression selects the line with the lowest total sum of squared prediction errors.

This value is called the Sum of Squares of Error, or SSE.

Sum of squares for errors

– This is the sum of differences between the observation points and the regression line.

– It can serve as a measure of how well the line fits the data.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

$$SSE = (n-1)s_Y^2 - \frac{\text{cov}(X,Y)}{s_x^2}$$

– This statistic plays a role in every statistical technique we employ to assess the model

Standard error of estimate

– The mean error is equal to zero.

– If σ_ε is small the errors tend to be close to zero (close to the mean error). Then, the model fits the data well.

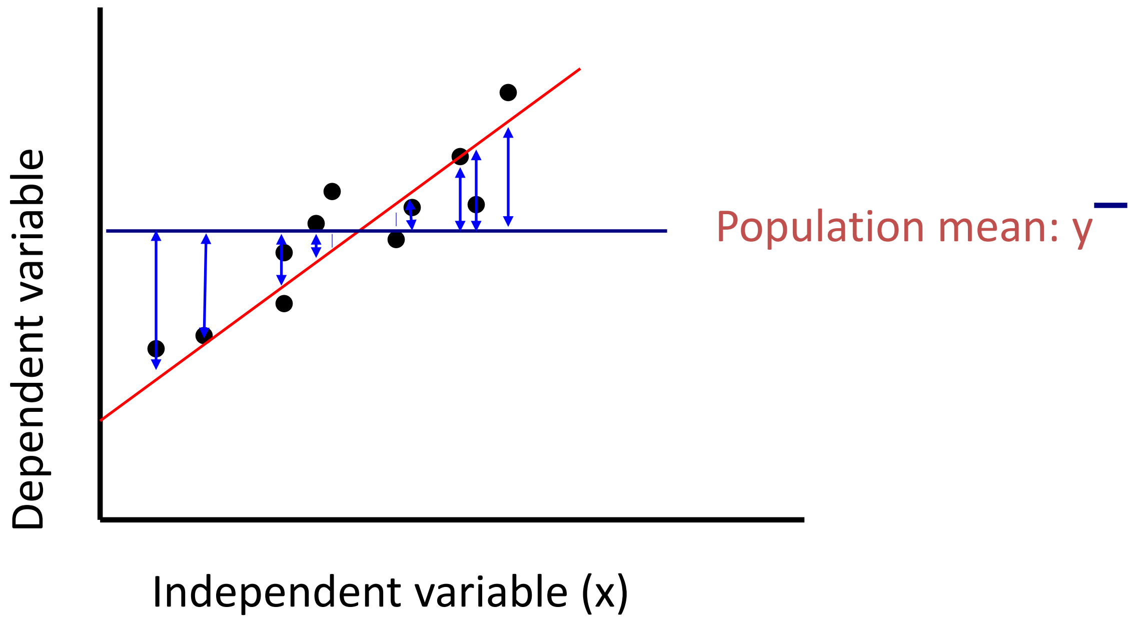
– Therefore, we can, use σ_ε as a measure of the suitability of using a linear model.

– An unbiased estimator of σ_ε^2 is given by s_ε^2

Standard Error of Estimate

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

Sum of Squares Regression (SSR)



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

Total Sum of Squares (SST)

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

$$SSR = \sum (\hat{y} - \overline{y})^2 \quad \text{(measure of explained variation)}$$

$$SSE = \sum (y - \hat{y})^2 \quad \text{(measure of unexplained variation)}$$

$$SST = SSR + SSE = \sum (y - \overline{y})^2 \quad \text{(measure of total variation in y)}$$

What about variance of our model?

Total variance of y: $$s_y^2 = \frac{\sum(y - \bar{y})^2}{n - 1}$$

Variance of predicted y values (ŷ):

$$s_{\hat{y}}^2 = \frac{\sum(\hat{y} - \bar{y})^2}{n - 1}$$

This is the variance explained by our regression model

Error variance:

$$s_{error}^2 = \frac{\sum(y - \hat{y})^2}{n - 2}$$

This is the variance of the error between our predicted y values and the actual y values, and thus is the variance in y that is NOT explained by the regression model

Coefficient of determination

The proportion of total variation (SST) that is explained by the regression (SSR)

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

The value of R^2 can range between 0 and 1, and the higher its value the more accurate the regression model is

Coefficient of determination

– When we want to measure the strength of the linear relationship, we use the coefficient of determination.

$$R^2 = \frac{[\operatorname{cov}(X, Y)]^2}{s_x^2 s_y^2} \quad or \quad R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

Coefficient of Determination

- R^2 measures the proportion of the variation in y that is explained by the variation in x.

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - SSE}{\sum (y_i - \bar{y})^2} = \frac{SSR}{\sum (y_i - \bar{y})^2}$$

- R^2 takes on any value between zero and one.

  $R^2 = 1$: Perfect match between the line and the data points.

  $R^2 = 0$: There are no linear relationship between x and y.