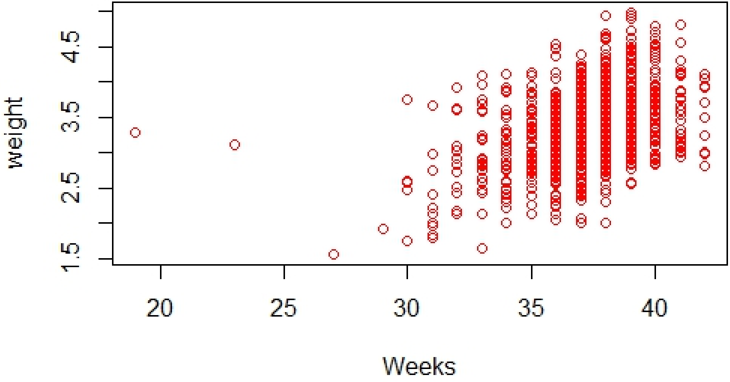
Elements in regression

Researchers are interested in the relation of weight of newborn and weeks of pregnancy. There are 1153 babies.

y = response variable
x = explanatory variable

use x to predict y



- linear trend but noisy

What is a regression?

RECALL: regression line y_hat = a + bx

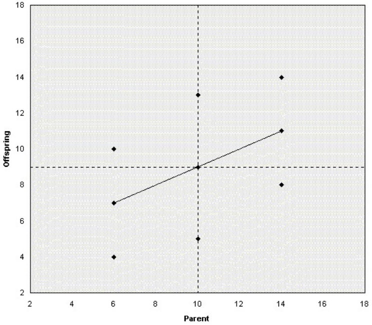
residual = y - y_hat

- **Regression** is a statistical methodology that use the statistical relation between predictor variable(s)(input, independent variable(s), etc.) and a response variable (output, dependent variable, etc.), so that a response can be predicted from the others.

- **Two distinct goals**
  1. Construct and Tests about statistical relation between predictor variables and response variables
  2. Prediction

Where to start? ...a little history

1. First used by Sir Francis Galton, 19th century.
   - Sweet pea experiment in 1875: size of mother pea and size of daughter pea.

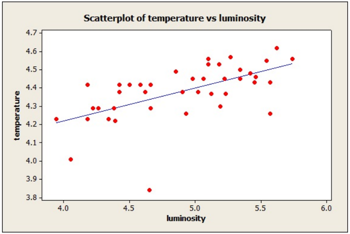


- Median weights of daughter seeds from a particular size of mother seed: a straight line with positive slope ("r") less than 1.0 approximately.
- Regression to the mean
  "Extremely large or small mother seeds typically generated substantially less extreme daughter seeds."

2. Later, Karl Pearson extended to statistics
   - median to mean (1896)
   - mechanical calculating machine (no later than 1910)

Regression: Examples

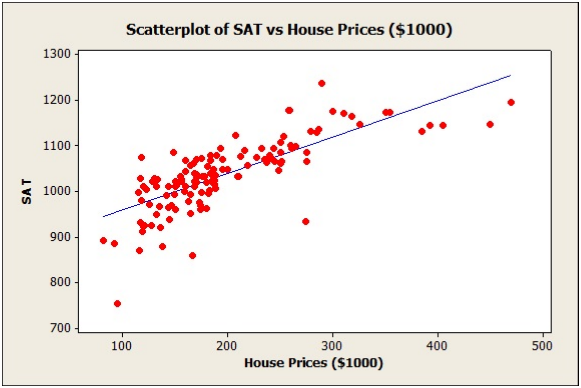
- Science and Engineering
  Luminosity and temperature of stars.



- Epidemiology and Biology
  smoking behavior, heart disease.
- Finance, economics and business.
  trend analysis

Regression: a interesting example

- House price and SAT score in Boston area



Does X cause Y??

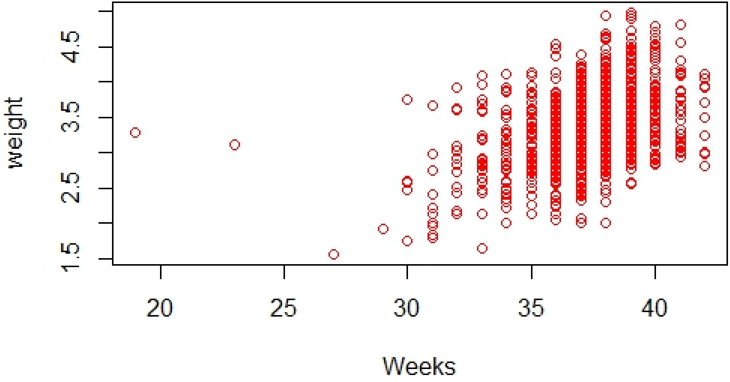
No! “**Correlation does not imply causation.**”

Goals

- **Given some data to fit regression model**
  - How do you validate model assumptions and fit a regression (or other) model to the data?
  - How to interpret the results and examine the confidence in the values of the model?
  - How to predict value for a new observation by the model?
- **Given a problem to predict some variable by some others.**
  - What kind of model should you use?
  - Which variables to be include?
  - Which transformations of variables and interaction terms should you use?

Relations between Two Variables: Regression

Recall the example of newborn baby weight and his/her mother pregnancy weeks.



Sample v.s. Population

Not only **Regression**, **Statistical Methods** are usually used to make generalization about population based on information of sample.

- A **sample** is the collection of units (people, animals, cities, fields, whatever you study) that is actually measured or surveyed in a study.
- The **population** is the large group of units we are interested in, from which the sample was selected.
- The sample, a subset of the population, is used to estimate characteristics of the population.

Example

Example: Pregnancy weeks and baby weight

- **population:**
- **Sample:**

Notations for Sample and Population

Different notations are used for sample and population characteristics. For instance,

- The mean of a sample is often denoted as $\bar{y}$.
- The mean of a population is often denoted as μ.
- An alternative notation for population mean is

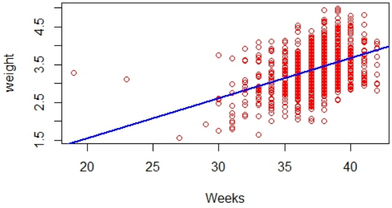
$$\mu = E(Y),$$

  called the expected value of Y, or the expectation of Y.

NOTE: In practice, we do not know the exact value of μ but only the value of $\bar{y}$. Therefore, we often use the sample characteristic to estimate the feature of the population.

Regression Notation

Based on a sample of 1153 babies,



- blue line estimated from the 1153 observations.

$$y = -0.56 + 0.11x$$

- Is it for the sample (x, y)?
- Is it for the population variable (X, Y)?

Regression Notation

A component of the simple regression model is that the mean value of the Y-variable is a straight line function of an X-variable. The two coefficients of a straight line are the intercept and the slope.

$$E(Y) = \beta_0 + \beta_1 X$$

- Intercept:

$$\begin{aligned} \text{population} &: \quad \beta_0 \\ \text{sample} &: \quad b_0 \text{ or } \hat{\beta}_0 \end{aligned}$$

- Slope:

$$\begin{aligned} \text{population} &: \quad \beta_1 \\ \text{sample} &: \quad b_1 \text{ or } \hat{\beta}_1 \end{aligned}$$

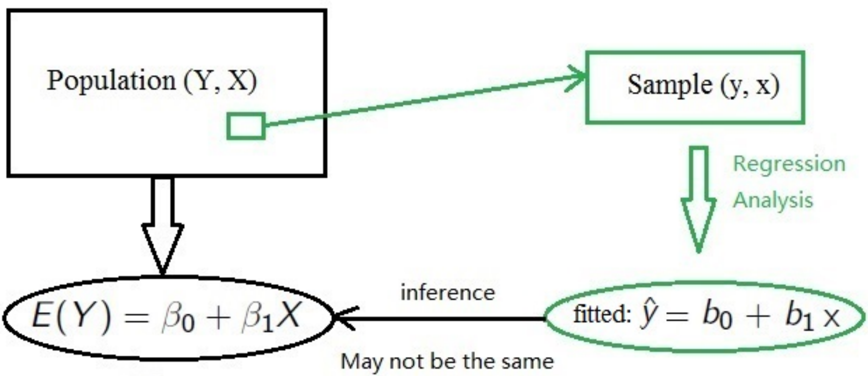
Regression Notation

In the example of newborn babies. Based on a sample of 1153 women, a regression line is

predicted weight= -0.56 + 0.11 weeks

Then the sample slope is $b_1 = 0.367$ and the sample intercept is $b_0 = -7.15$
We do not know the values of β_0 and β_1, the intercept and slope for the larger population of all individuals in the population (all babies born). It would be wrong, for example, to write $\beta_1 = 0.367$

Regression Notation



Model for Simple Linear Regression

- A **regression equation** describes how the **mean** value of a Y-variable relates to specific values of the X-variable(s) used to predict Y.
- The **simple (linear) regression equation** is that the mean of Y is a straight line function of X:

$$E(y_i) = \beta_0 + \beta_1 \cdot x_i,$$

where $E(y_i)$ is used to represent the mean value (expected value), and the subscript i denotes the *ith* unit in the population.

Model for Simple Linear Regression

The overall **simple (linear) regression model**:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon,$$

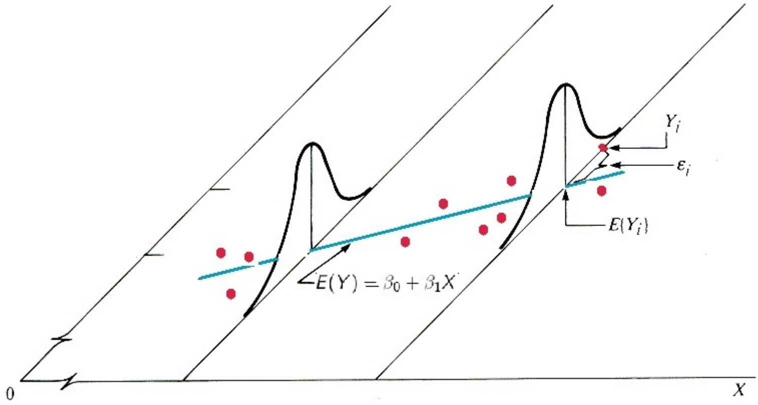
- y: **response/dependent variable**
- x: **predictor/independent variable**
- ϵ: **random error** of Y from the line $\beta_0 + \beta_1 X$.

Model for Simple Linear Regression

Assumptions of Errors

- All the errors ϵ are independent with mean 0, i.e. $E(\epsilon) = 0$.
- All the errors ϵ have the same degree of variation from the regression line for all x, i.e. $\text{var}(\epsilon) = \sigma^2$.
- For the purpose of statistical inference, we assume that the errors have a normal distribution, i.e. $\epsilon \sim N(0, \sigma^2)$.

Model for Simple Linear Regression



Sample Estimates of Model

Assume simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \ i = 1, \ldots, n. \tag{1}$$

Fitted model:

$$\hat{y}_i = b_0 + b_1 x_i;\ \ i = 1, ..., n. \tag{2}$$

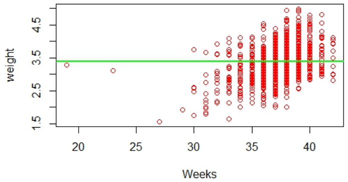
where Predicted/Fitted values: $\hat{y}_i = b_0 + b_1 x_i$

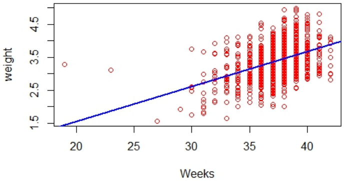
Our **goal** is to estimate β_0, β_1 by b_0, b_1 based on a sample of observations (x_i, y_i) of size n.

How to decide the best estimates based on the sample?

How to get b_0 and b_1?

Which one is better Guess 1 and Guess 2?





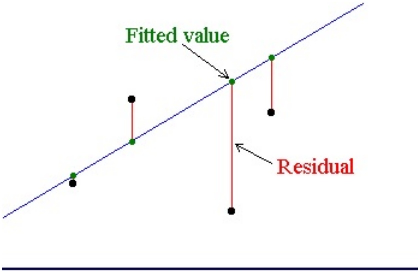
green line:$y = 3.38 + 0 \cdot x$, sum of squared lengths$= 312.0133$

blue line:$y = -0.56 + 0.11 \cdot x$, sum of squared lengths$= 254.2687$

How to get b_0 and b_1?

Our criterion is **least sum of squared errors**.

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



Predicted/Fitted values

$$\hat{y}_i = b_0 + b_1 x_i$$

Observed errors (residuals)

$$e_i = y_i - \hat{y}_i$$

Find b_0, b_1 such that SSE is minimized!

How to get b_0 and b_1?

Based on the rule of least sums of squared errors, the estimated β coefficients in the simple linear regression model have the following expressions:

$$
\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
b_0 &= \bar{y} - b_1\bar{x}
\end{aligned}
$$

b_0 and b_1 are called the **least square** estimates of β_0 and β_1.

The old example again...

The simple regression equation relating weight of a new born baby(y) and his/her mother's pregnancy weeks(x) is

$$\text{average weight} = \beta_0 + \beta_1 \cdot \text{weeks}.$$

Based on sample, sample intercept is $b_0 = -0.56$ and the sample slope is $b_1 = 0.11$. Interpret the parameters:

- b_0: the average height at weeks=0 is - 0.56
- b_1: For one unit increase in weeks, the average weight increases by 0.11 kg.

The old example again...

Note: It would be wrong, for example, to write the regression equation

$$\text{average weight} = -0.56 + 0.11 \cdot \text{week}.$$

But we can write

$$\widehat{\text{weight}} = -0.56 + 0.11 \cdot \text{week}.$$

How good the fitted model is?

- Sum of Squared Errors

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

- Mean Squared Error

$$MSE = \frac{SSE}{n-2}$$

MSE is the sample variance of the errors and estimates σ^2.
Important Note: The divisor $n-2$ only applies to simple regression. The general rule is that the divisor is $n-p$, where $p=$number of parameters in the regression equation.

$$E\{MSE\} = \sigma^2$$

How good the fitted model is?

- Standard Deviation of Errors

$$s = \sqrt{MSE}$$

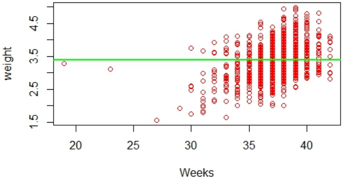
which is the sample standard deviation of the errors (residuals) from the regression line. $s = \sqrt{MSE}$ can be interpreted (roughly) as the average absolute size of deviations of individuals from the sample regression line.

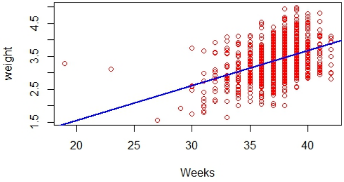
How good the fitted model is?

- *R*-square
$$R^2 = \frac{SST - SSE}{SST}$$
where SST is the **Sums of Squares Total** (or total sum of squares)
$$SST(\text{ or } SSTO) = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$
- $\sqrt{R^2} = r(\text{correlation})$
- **Interpretation:** R^2 is interpreted as the fraction of variation in y that is explained by the fitted regression equation. It is often converted to a percentage.

How good the fitted model is?

NOTE: residual std dev = sqrt(sum(y - y_hat)^2 / n - 2)





green line: $\bar{y} = 3.385$, $SSTO = 312.0133$

blue line: $y = -0.56 + 0.11x$, $SSE = 254.2687$

MSE=______________, s=______________, R^2=______________