STP 530
Lecture 1: Introduction to Regression Analysis, Simple Linear Regression

Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

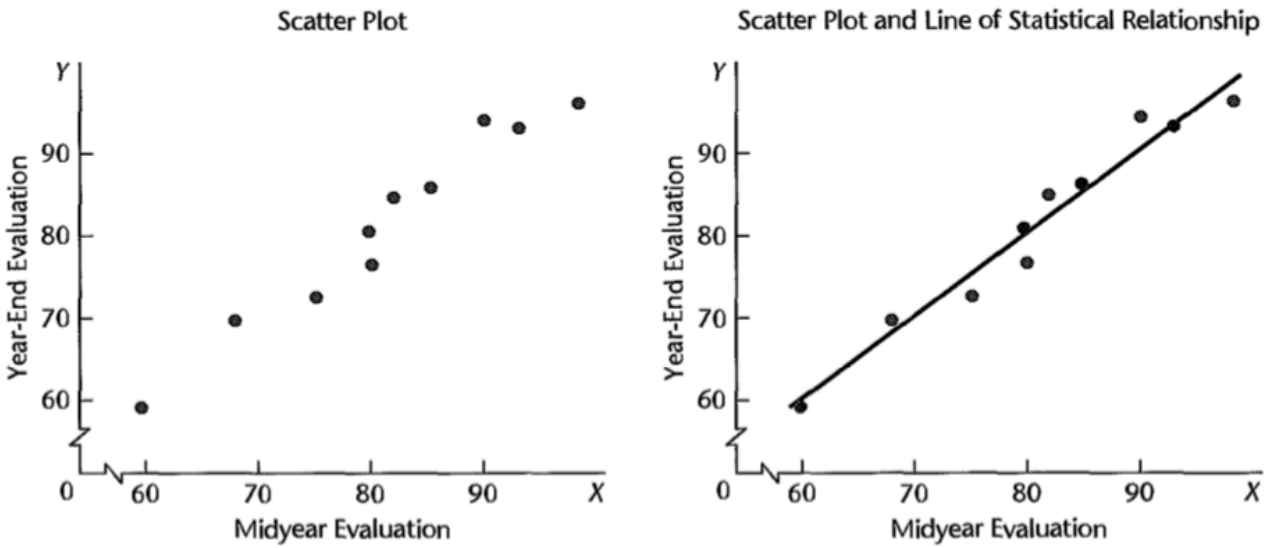
ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

August 22, 2023

What is Regression Analysis?

What is regression analysis?

Regression analysis is a statistical methodology that studies the relation among variables.



$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{OR} \quad E\{Y\} = \beta_0 + \beta_1 X$$

A regression model describes the general trend existing among the variables (the signal), separated from the "random" perturbation (the noise).

Relation between variables

Two types of relations between variables:

- Functional relation
- Statistical relation

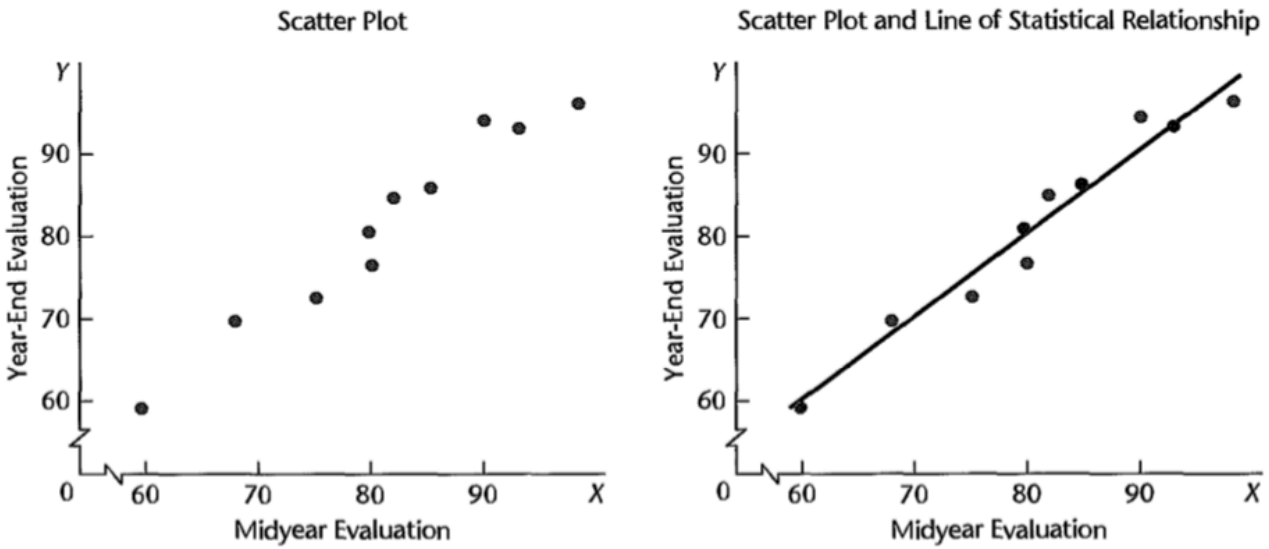
A **functional** relation is expressed by a mathematical formula. No deviation from the function is considered.

Example: Temperature in Celsius (X) and temperature in Fahrenheit (Y).

$$Y = 32 + 1.8X$$

Relation between variables

A **statistical** relation considers and models random deviation from the mathematical formula. In general, observations do not fall directly on the curve of relationship.



A **regression model** is a formal means of expressing two essential ingredients of a statistical relation:

1. A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion.
2. A scattering of points around the curve of statistical relationship.

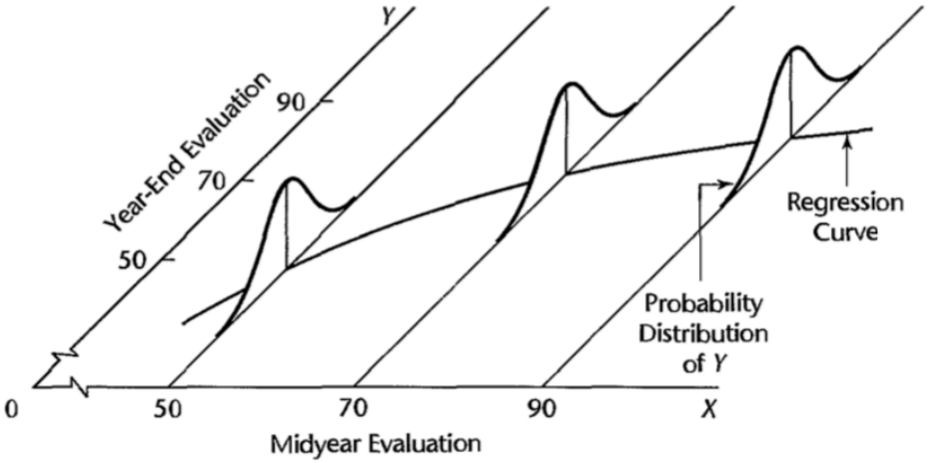
These two characteristics are embodied in a regression model by postulating that:

1. There is a probability distribution of Y for each level of X.

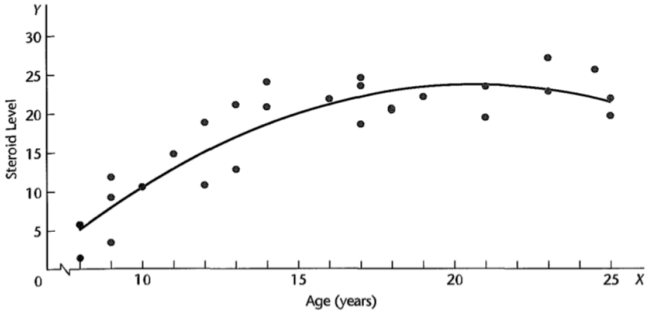
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

2. The means of these probability distributions vary in some systematic fashion with X.

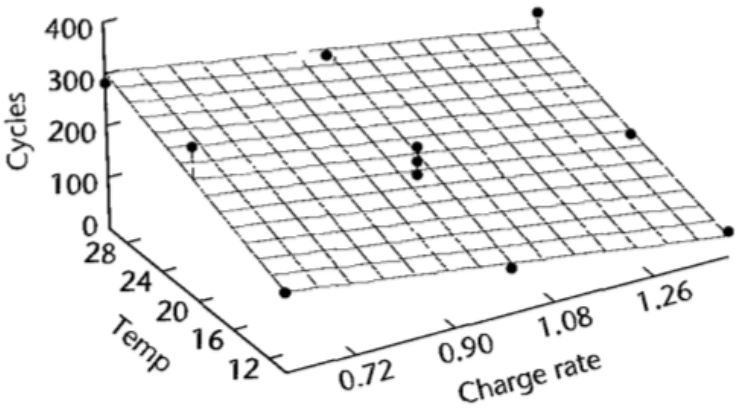
$$E\{Y\} = \beta_0 + \beta_1 X$$



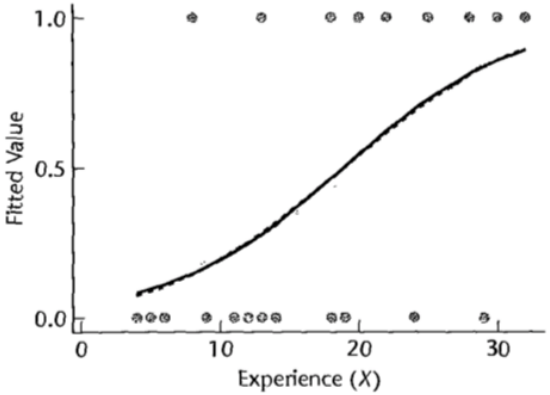
More examples of regression models



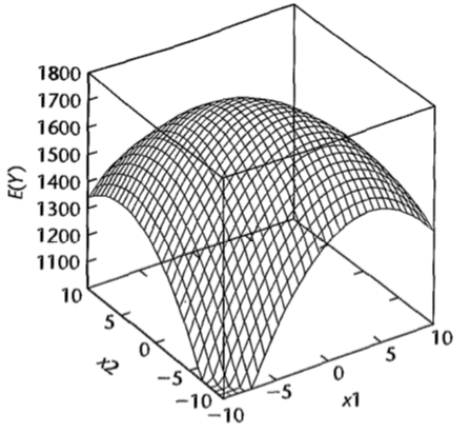
$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$$



$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



$$ln\frac{E\{Y\}}{1 - E\{Y\}} = \beta_0 + \beta_1 X$$



$$E\{Y\} = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_1 X_2$$

Why call it "regression" at all???

According to the Oxford Dictionary:

regression

NOUN

[mass noun]

1 A return to a former or less developed state.
'it is easy to blame unrest on economic regression'
+ More example sentences + Synonyms

1.1 A return to an earlier stage of life or a supposed previous life, especially through hypnosis.
'the details recalled by Carol during the regression'
[as modifier] 'regression therapy'
[count noun] 'a past-life regression'
+ More example sentences + Synonyms

1.2 A lessening of the severity of a disease or its symptoms.
'there was 46.7 per cent complete regression in the placebo group'
+ More example sentences

2 *Statistics*
A measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).

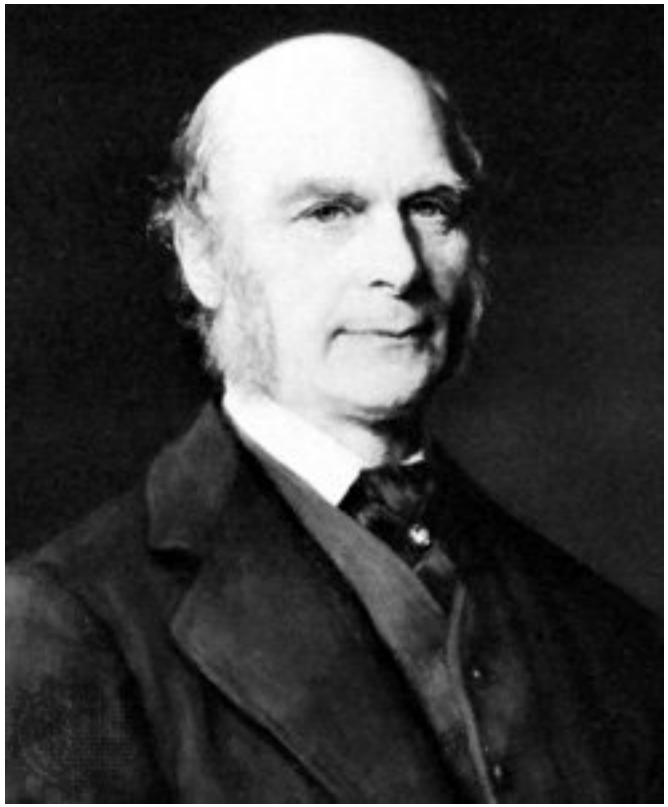
Why call it "regression"?

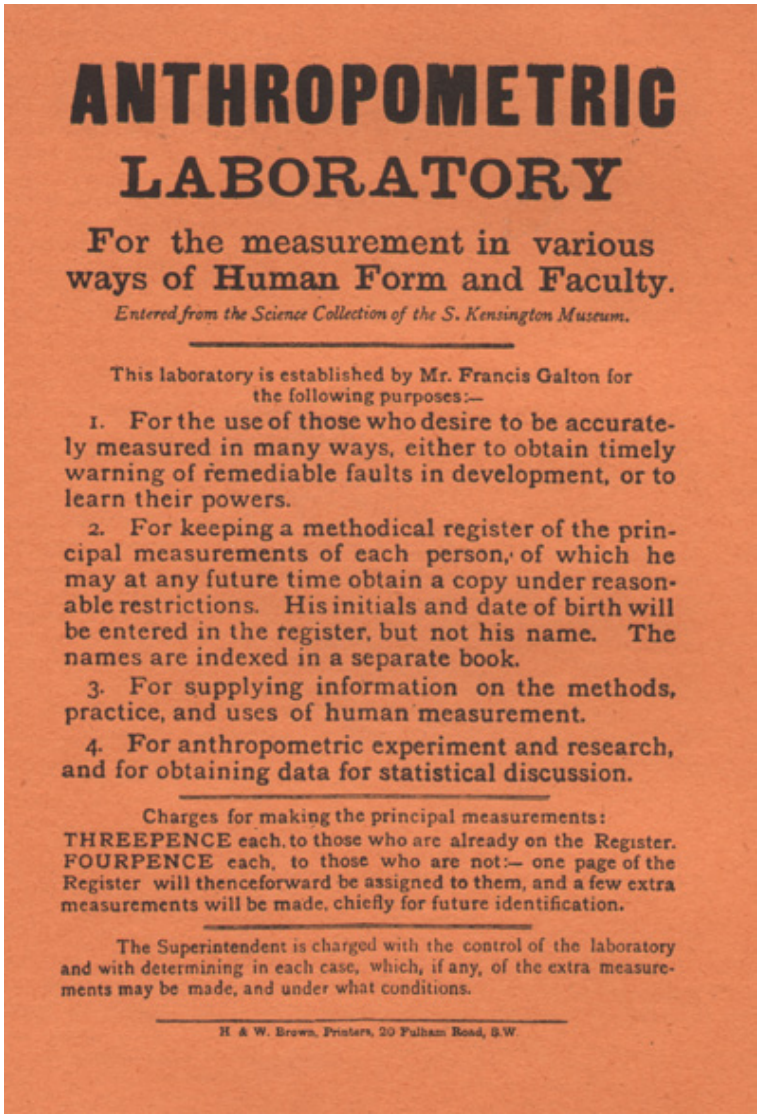
Textbook Section 1.2, page 5:

"Regression analysis was first developed by Sir Francis Galton in the latter part of the 19th century. Galton had studied the relation between heights of parents and children and noted that the heights of children of both tall and short parents appeared to "revert" or "regress" to the mean of the group. He considered this tendency to be a regression to "mediocrity." Galton developed a mathematical description of this regression tendency, the precursor of today's regression models. The term regression persists to this day to describe statistical relations between variables."

Why call it "regression"?

Sir Francis Galton
(English Statistician, 1822-1911)





ANTHROPOMETRIC
LABORATORY

For the measurement in various ways of Human Form and Faculty.

Entered from the Science Collection of the S. Kensington Museum.

This laboratory is established by Mr. Francis Galton for the following purposes:—

1. For the use of those who desire to be accurately measured in many ways, either to obtain timely warning of remediable faults in development, or to learn their powers.
2. For keeping a methodical register of the principal measurements of each person, of which he may at any future time obtain a copy under reasonable restrictions. His initials and date of birth will be entered in the register, but not his name. The names are indexed in a separate book.
3. For supplying information on the methods, practice, and uses of human measurement.
4. For anthropometric experiment and research, and for obtaining data for statistical discussion.

Charges for making the principal measurements:
THREEPENCE each, to those who are already on the Register.
FOURPENCE each, to those who are not:— one page of the Register will thenceforward be assigned to them, and a few extra measurements will be made, chiefly for future identification.

The Superintendent is charged with the control of the laboratory and with determining in each case, which, if any, of the extra measurements may be made, and under what conditions.

H. & W. Brown, Printers, 20 Fulham Road, S.W.

Why call it "regression"?

Galton's 1st anthropometric lab at the International Health Exhibition, London, 1885.

At least 17,000 individuals were testing during 1880s-1890s. About 7,500 of the data records have survived to the present day.

Galton's 2nd anthropometric lab, South Kensington Museum, London, 1891-95.

Two Purposes of Regression Analysis

Example: Analyzing the bike sharing dataset

Data source: The UC Irvine Machine Learning Repository

`https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset`

Pictures above: Bike sharing in New York City

Example: Analyzing the bike sharing dataset

Bike Sharing Dataset

Donated on 12/19/2013

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

Dataset Characteristics	Subject Area	Associated Tasks
Univariate	Social	Regression
Attribute Type	**# Instances**	**# Attributes**
Integer, Real	17389	16

Information

Additional Information

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed ...

SHOW MORE

Has Missing Values?

No

Example: Analyzing the bike sharing dataset

Attribute Information

Additional Information

Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

- instant: record index
- dteday : date
- season : season (1:winter, 2:spring, 3:summer, 4:fall)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
+ weathersit :
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

Example: Analyzing the bike sharing dataset

Regression purpose 1: Making predictions

The managers of the bike-sharing system may wish to predict ridership, say for the following question:

> Tomorrow, Sunday, is expected to be sunny and cool, say 62 degrees Fahrenheit. We may wish to predict the number of riders, so that we can get some idea as to how many bikes will need repair. We may try to predict ridership, given the weather conditions, day of the week, time of year and so on.

Some bike-sharing services actually have trucks to move large numbers of bikes to locations that are expected to have high demand. Prediction would be even more useful here.

Example: Analyzing the bike sharing dataset

Regression purpose 2: Describing patterns

We may be interested in determining what factors affect ridership, such as:

- How much effect does wind speed have in influencing whether people wish to borrow a bike?
- Does the day of the week make a difference in bike sharing activities?
- How does bike sharing vary by seasons?
- Does weather alter bike sharing activities to a greater extent during holidays vs non-holidays?

Example: Predicting Body Fat

Special note about the *prediction* goal:

We can use regression to:

- Predict the future (e.g., the bike sharing example)
- "Predict" the past or the present

In the *latter* scenario, we are interested in the *specific values* of the response variable given the predictor values of *a specific case*.

> For example, consider the bodyfat data set, available in the R package, **mfp**. Direct measurement of body fat is expensive and unwieldy, as it involves underwater weighing. Thus it would be highly desirable to "predict" that quantity from easily measurable variables such as height, age, weight, abdomen circumference, etc.

Two purposes of regression analysis

Two distinctive purposes of regression analysis:

1. Description
   - What are the most important factors that impact the outcome?
   - In what functional form is X associated with Y?
   - How close is the estimated equation to the truth?
   - Which factor is more important/influential than the others?
   - Does the effect of one factor on Y vary by the level of another factor?
2. Prediction
   - What is the equation that provides the most accurate prediction of Y?
   - What is the margin of error for each prediction?
   - Is the equation robust enough to generate accurate prediction for new data?

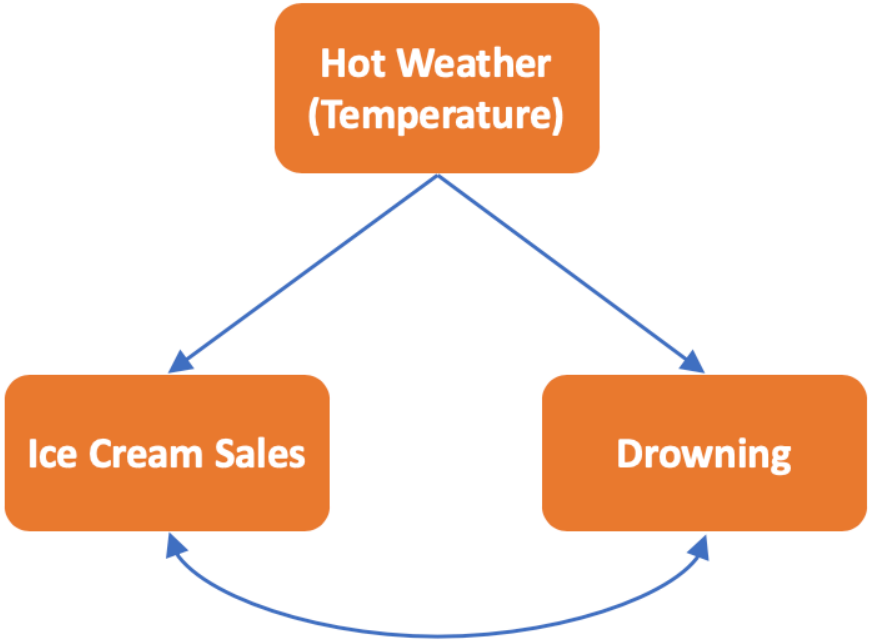
Key Tasks of Constructing Regression Models

Key tasks of constructing regression models

1. Select predictor variables
   - To what extent a predictor variable contributes to reducing the unexplained variation in Y?
   - Can observations of the variable be obtained reliably, accurately, quickly, and/or economically?
   - Is the predictor variable an important causal agent (theoretically) in the process under analysis?
   - Can the variable be controlled?

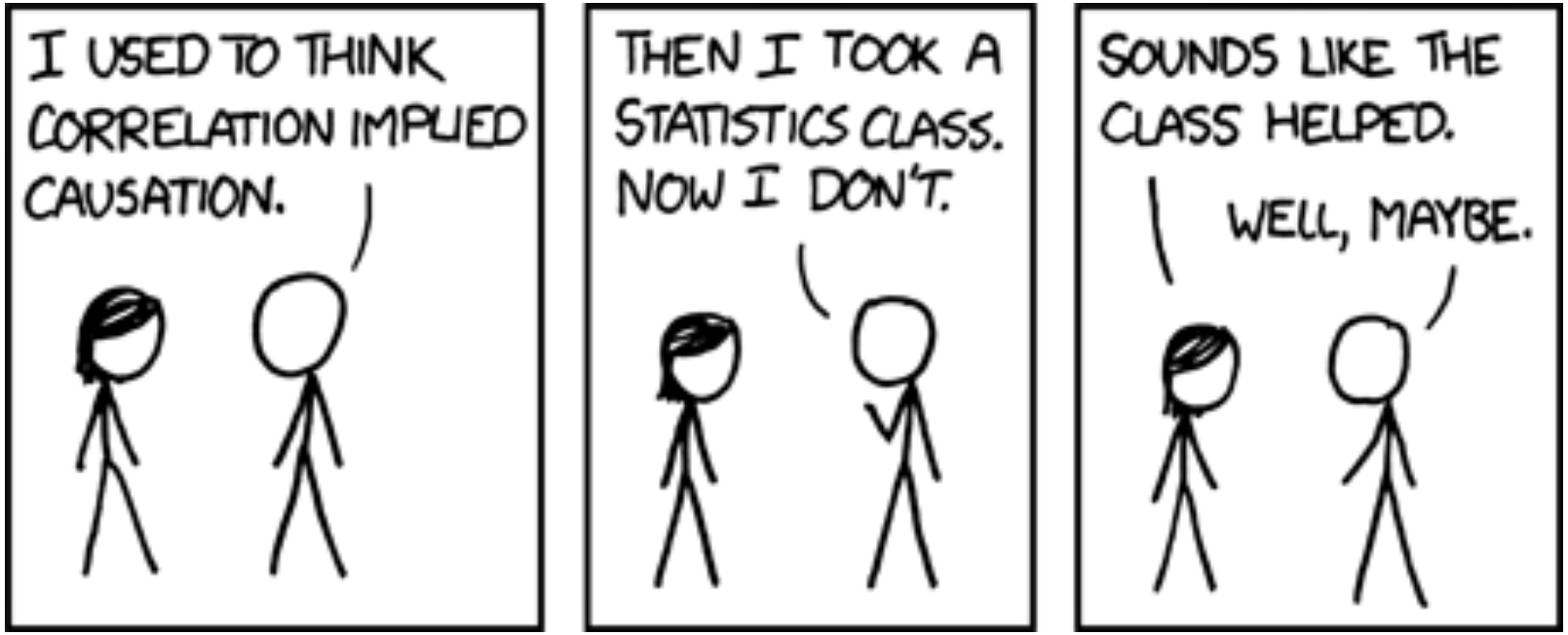
Regression doesn't imply causation

Regression (Association) doesn't imply causation.



THEN WHAT??

Regression doesn't imply causation. Correct, BUT...



We do care about causation!

Regression doesn't imply causation. Correct, BUT...

How do we reveal causation?

Three necessary conditions for causality by British philosopher John Stuart Mill (1806-1873):

1. A relationship between the potential cause and effect;
2. The cause to temporally precede the effect;
3. No plausible alternative explanation exists.

Key tasks of constructing regression models

2. Functional form of regression relation

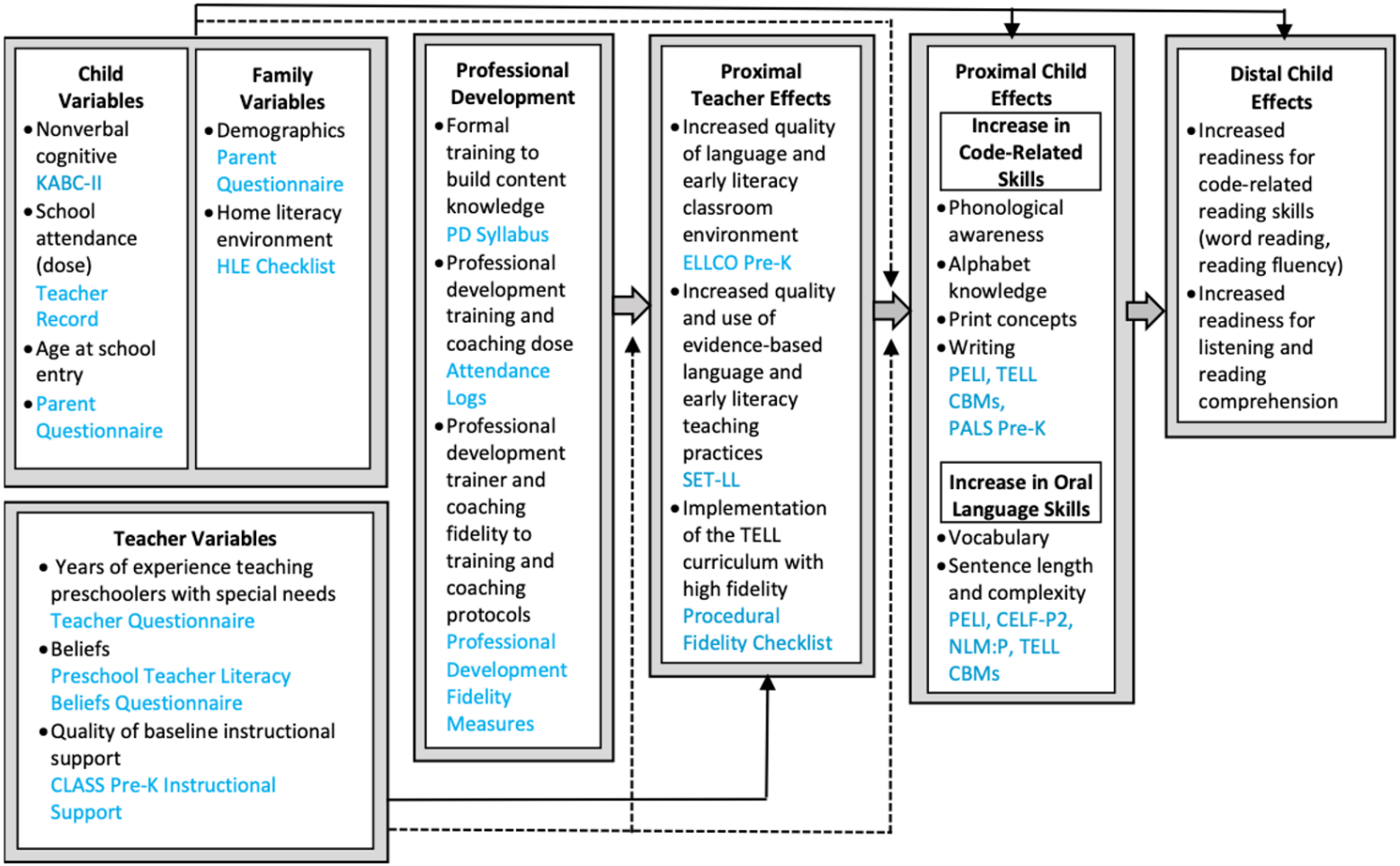
Two ways to determine the functional form:

- **Guided by theory**: Theory hypothesizes a causal process. The regression analysis is used to validate the theory.
- **Data-driven**: Model building $\rightarrow$ Statistical learning. The promises and caveats of data science.

Theory-driven regression analysis

Example: **TELL curriculum development grant (Gray, Wilcox, Zheng)**

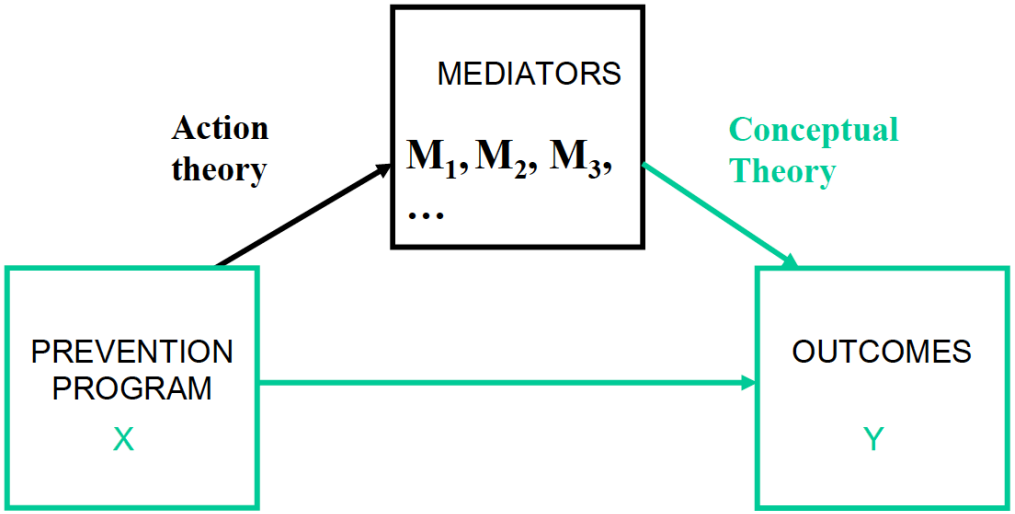
Figure 1: Proposed Theory of Change



Theory-driven regression analysis

Example: **mediation analysis**
(Excerpts from online tutorial by Dr. David MacKinnon, ASU Psychology. For more detail see supplemental resource on Canvas course site.)

Intervention Mediation Model

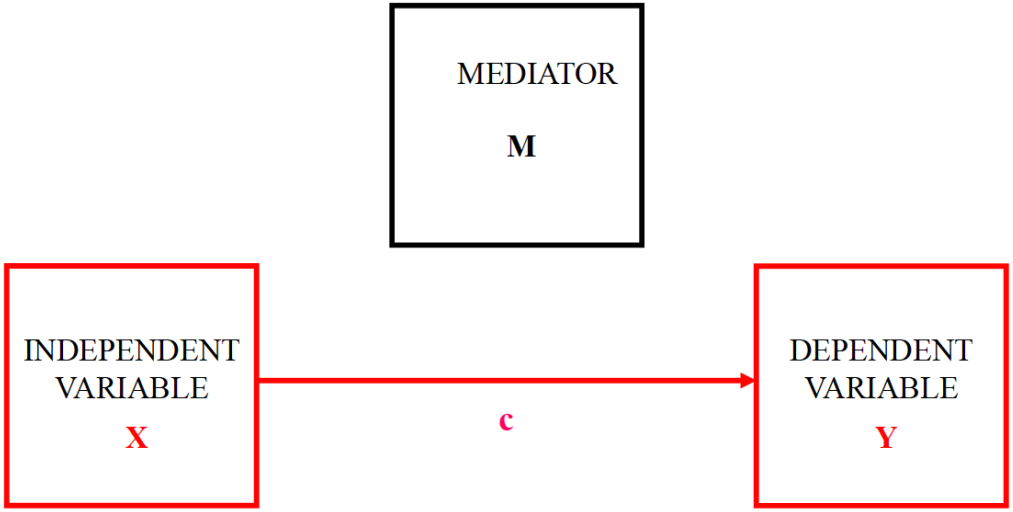


If the mediators selected are causally related to Y, then changing the mediators will change Y.

Theory-driven regression analysis

Example: **mediation analysis**

Regression Equation 1



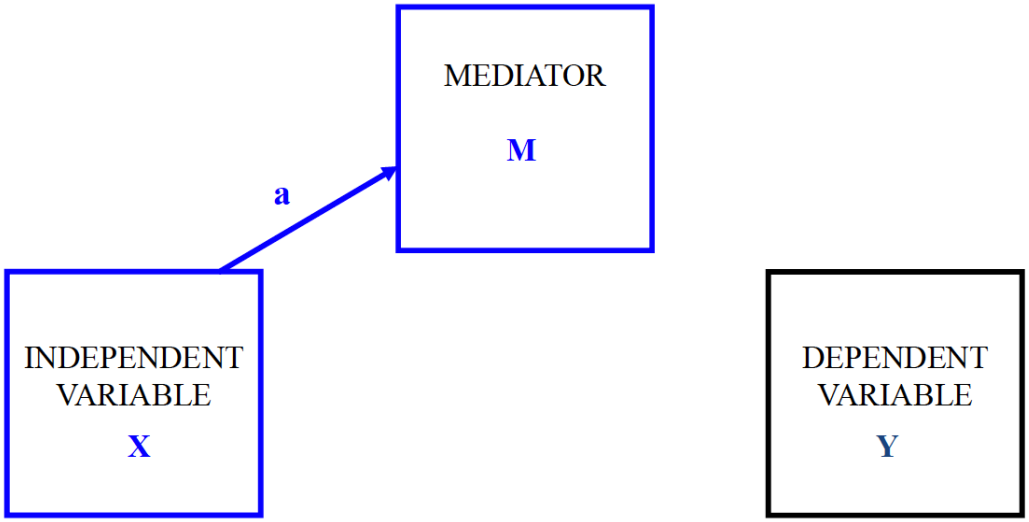
1. The independent variable is related to the dependent variable:

$$Y = i_1 + \hat{c}X + e_1$$

Theory-driven regression analysis

Example: **mediation analysis**

Regression Equation 2



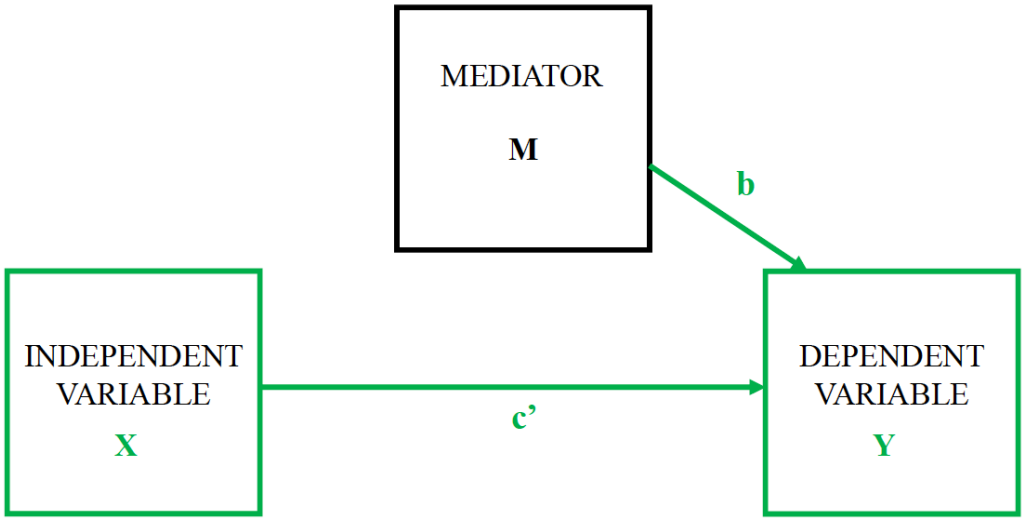
2. The independent variable is related to the potential mediator:

$$M = i_2 + \hat{a}X + e_2$$

Theory-driven regression analysis

Example: **mediation analysis**

Regression Equation 3



3. The mediator is related to the dependent variable controlling for exposure to the independent variable:

$$Y = i_3 + \hat{c}' X + \hat{b} M + e_3$$

Theory-driven regression analysis

Example: **mediation analysis**

Mediated Effect Measures

Mediated effect = *ab* Product of Coefficients

Mediated effect = *c*-*c'* Difference in Coefficients

Mediated effect = *ab* = *c*-*c'*
(see MacKinnon et al., 1995 for a proof)

Direct effect = *c'* & Total effect = *ab* + *c'* = *c*

Recap: Two purposes of regression analysis

Two distinctive (and often contradictory) purposes of regression analysis:

1. Description **(theory-driven or data-driven)**
   - What are the most important factors that impact the outcome?
   - In what functional form is X associated with Y?
   - How close is the estimated equation to the truth?
   - Which factor is more important/influential than the others?
   - Does the effect of one factor on Y vary by the level of another factor?
2. Prediction **(mostly data-driven)**
   - What is the equation that provides the most accurate prediction of Y?
   - What is the margin of error for each prediction?
   - Is the equation robust enough to generate accurate prediction for new data?

Simple Linear Regression Model

Formal statement of model

A one-predictor linear regression model (a.k.a., simple linear regression model) is stated as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

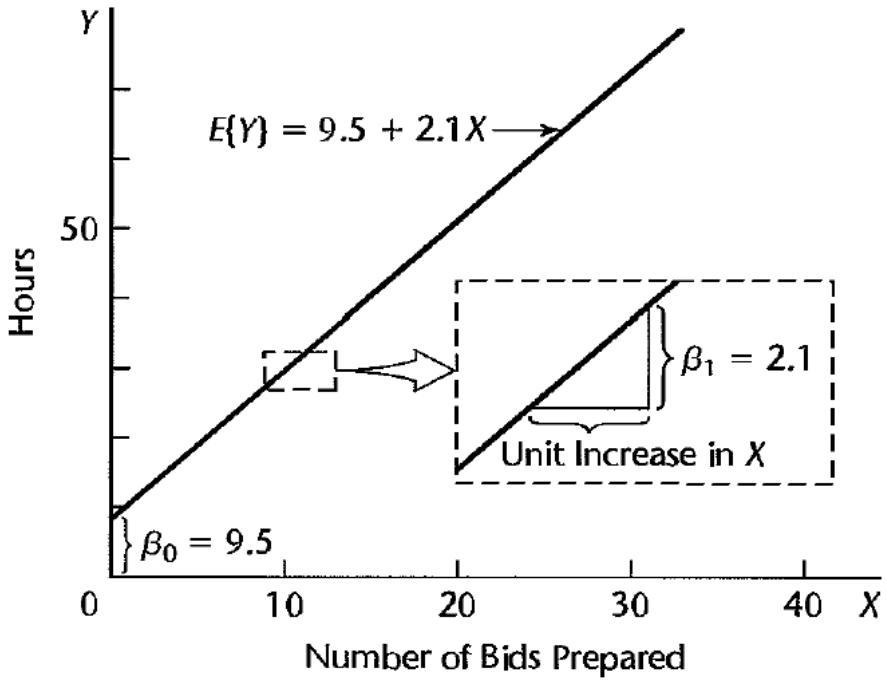
where:

- Y_i is the value of the response variable in the ith trial
- β_0 and β_1 are parameters
- X_i is the value of the predictor variable in the ith trial
- ε_i is a random error term with mean $E\{\varepsilon_i\} = 0$ and variance $\sigma^2\{\varepsilon_i\} = \sigma^2$; ε_i and ε_j are uncorrelated, i.e. $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$.

Important features of model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- β_0 is the intercept: The **expected** Y value when $X = 0$.
- β_1 is the slope: The change in the **expected** Y value per 1-unit increase in X.

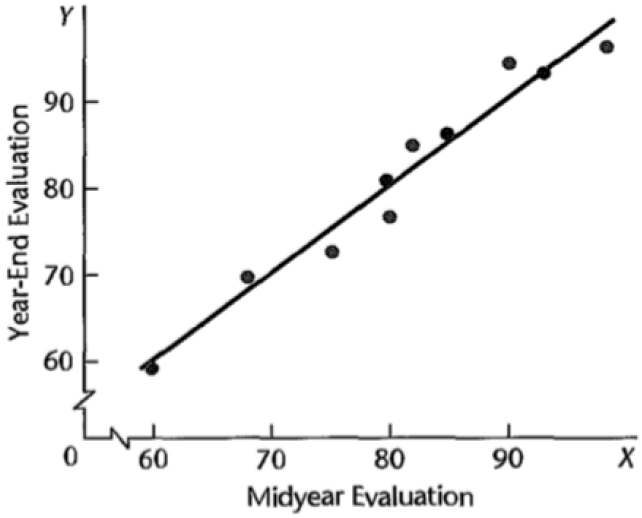


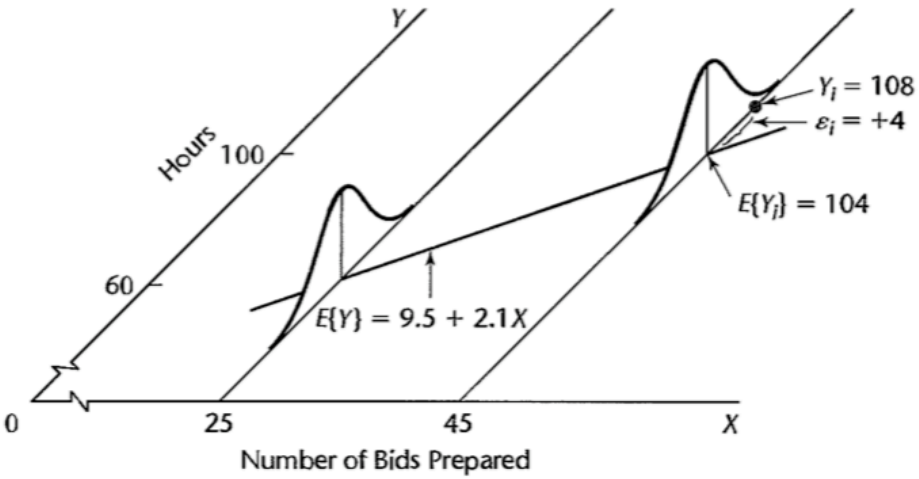
Important features of model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- The **mean** of the probability distribution of Y_i given X_i:
  $E\{Y_i\} = E\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \beta_0 + \beta_1 X_i + E\{\varepsilon_i\} = \beta_0 + \beta_1 X_i$

- The **variance** of the probability distribution of Y_i given X_i:
  $\sigma^2\{Y_i\} = \sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sigma^2\{\varepsilon_i\} = \sigma^2$

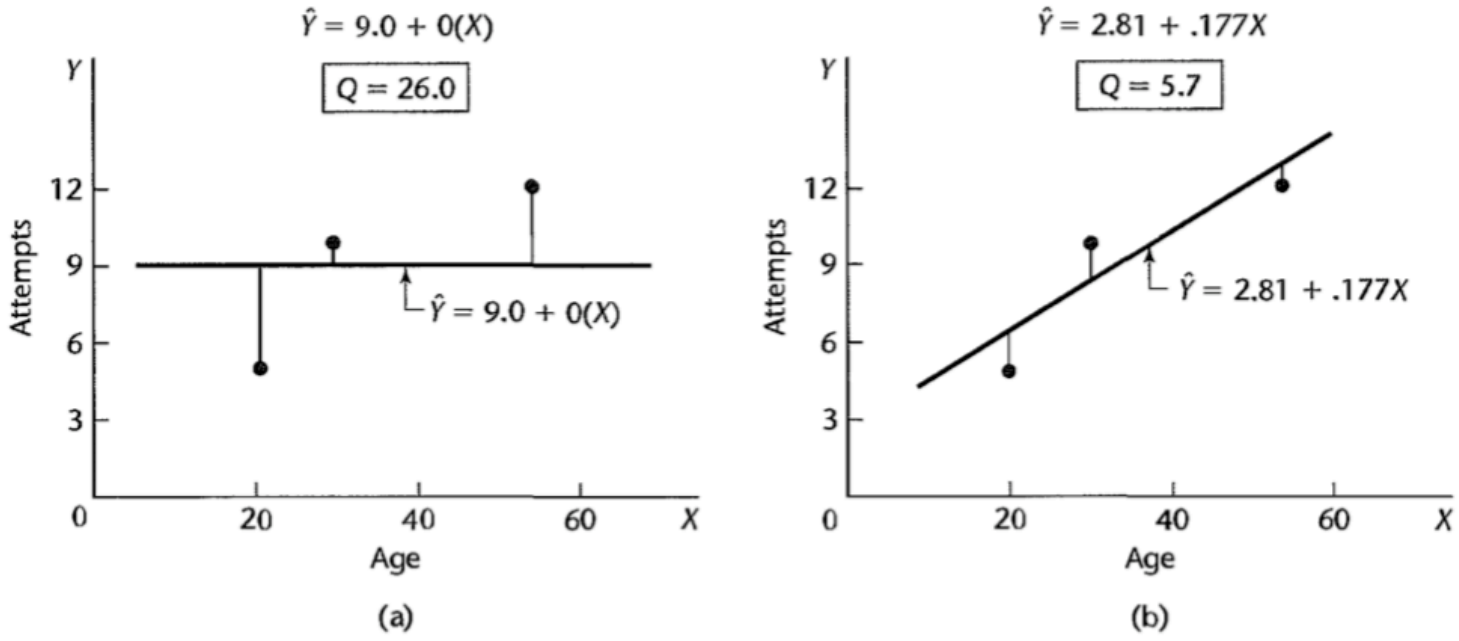




Estimation of Regression Function

Least square estimation

FIGURE 1.9 Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.



For a given sample of observation pairs $(X_i, Y_i), i = 1, \cdots, n$, the **method of least squares (LS)** minimizes Q, the **sum of squared errors**:

$$Q = \text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Least square estimation

The problem: Minimize $Q = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$

The estimation problem can be solved in two ways:

1. **Numerical search procedures**: e.g., grid search with "all" possible values of β's and choose the values when Q is minimized.
2. **Analytical procedures**: e.g., solve the partial derivative equation system.

Least square estimation: The partial derivative approach

The β_0 and β_1 values that minimize Q satisfy that:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum X_i(Y_i - \beta_0 - \beta_1 X_i) = 0$$

Solve the above equations for β_0 and β_1. We get:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{Cov\{X, Y\}}{Var\{X\}}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

Note on notation: Greek letters for parameters (theoretically posited). Roman letters for statistics (empirically obtained).

Important distinctions:

- The hypothesized model: $E\{Y\} = \beta_0 + \beta_1 X$
- The estimated prediction equation: $\hat{Y} = b_0 + b_1 X$

- The error term (hypothesized, unknown): $\varepsilon_i = Y_i - E\{Y_i\}$
- The residual (calculated, known, an estimate of ε_i): $e_i = Y_i - \hat{Y}_i$

Least square estimation: properties

A few useful properties to remember:

1. From the least square solution $b_0 = \bar{Y} - b_1\bar{X}$, we have $\bar{Y} = b_0 + b_1\bar{X}$, which means the point $(\bar{X}, \bar{Y})$ is always on the fitted regression line.
2. It will be shown in Chapter 2 that

$$E\{b_0\} = \beta_0$$

$$E\{b_1\} = \beta_1$$

   which means b_0 and b_1 are **unbiased estimators** for β_0 and β_1.
3. $\hat{Y}$ is an unbiased estimator of $E\{Y\}$ because

$$E\{\hat{Y}\} = E\{b_0 + b_1X\} = E\{b_0\} + E\{b_1\}X = \beta_0 + \beta_1X$$

Least square estimation: properties

From the partial derivative equations

$$\frac{\partial Q}{\partial \beta_0} = -2\sum(Y_i - \beta_0 - \beta_1 X_i) = 0; \quad \frac{\partial Q}{\partial \beta_1} = -2\sum X_i(Y_i - \beta_0 - \beta_1 X_i) = 0$$

We know that the solutions b_0 and b_1 satisfy the following **normal equations**:

$$\sum(Y_i - b_0 - b_1 X_i) = 0; \qquad \sum X_i(Y_i - b_0 - b_1 X_i) = 0$$

So the least square solution has the following properties:

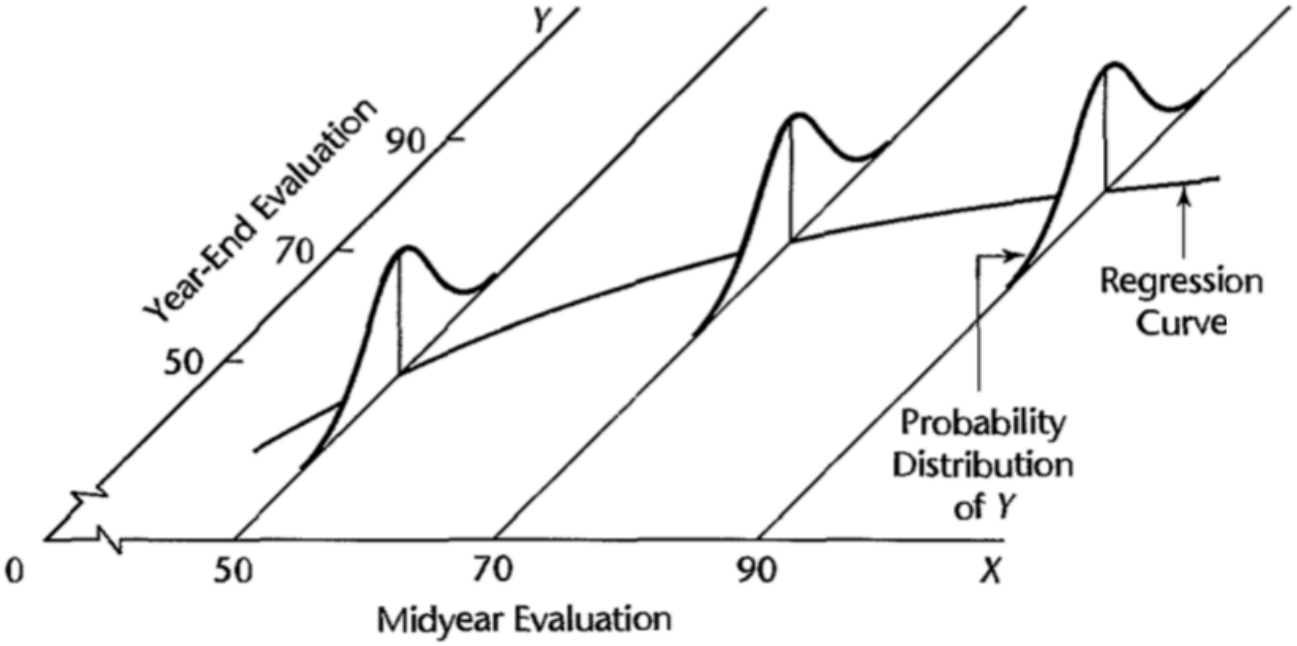
4. $\sum e_i = 0$, because $\sum e_i = \sum(Y_i - b_0 - b_1 X_i) = 0$
5. $\sum Y_i = \sum \hat{Y}_i$, because $\sum Y_i = \sum(\hat{Y}_i + e_i) = \sum \hat{Y}_i + \sum e_i = \sum \hat{Y}_i$
6. $\sum X_i e_i = 0$ (a.k.a. $\boldsymbol{X} \perp \boldsymbol{e}$), because $\sum X_i e_i = \sum X_i(Y_i - b_0 - b_1 X_i) = 0$
7. $\sum \hat{Y}_i e_i = 0$ (a.k.a. $\hat{\boldsymbol{Y}} \perp \boldsymbol{e}$), because
   $\sum \hat{Y}_i e_i = \sum(b_0 + b_1 X_i)e_i = b_0 \sum e_i + b_1 \sum X_i e_i = 0$

Estimation of Error Term Variance σ^2

Estimation of error term variance σ^2

First note:

- σ^2 is the variance of the error term ε_i (hypothesized, unknown)
- s^2 denotes the estimate of σ^2 (calculated, known), but it's not the variance of e_i (will be explained later).



Estimation of error term variance σ^2

The error term variance σ^2 is estimated by:

$$s^2 = MSE = \frac{SSE}{n-p} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p}$$

where p denotes the total number of unknown parameters to be estimated in the model (including the intercept), which equals the number of predictors in the model plus 1.

The value $s = \sqrt{s^2}$ (a.k.a., residual standard error in R outputs) is useful to quantify the amount of random deviation the model curve to observed data.

R code: `s = sqrt((sum((y - yhat) ^ 2))/(n - p))`