Stat 440/540: Regression Analysis

Instructor: Yan Lu

Albuquerque, UNM

Chapter 1 Linear Regression with One Predictor Variable

Statisticians typically analyze data by creating probability models for the data.

- A model for the data is simply a statement of the assumptions
- Assumptions about the data: typically, the observations are independent, have equal variances, and that either the observations are normally distributed or involve large sample sizes
- Point estimation
- Interval estimation: confidence intervals, prediction intervals
- Tests of a null hypothesis
- Validity of the model: diagnostics

Inference on single parameters: four things

1. the parameter of interest, Par
2. the estimate of the parameter, Est
3. the standard error of the estimate, SE(Est)
4. the appropriate reference distribution

$$\frac{\text{Est - Par}}{\text{SE(Est)}}$$

has a distribution that is some member of the family of t distributions, say $t(\text{df})$, where df specifies the degrees of freedom.

Linear Regression with One Predictor Variable

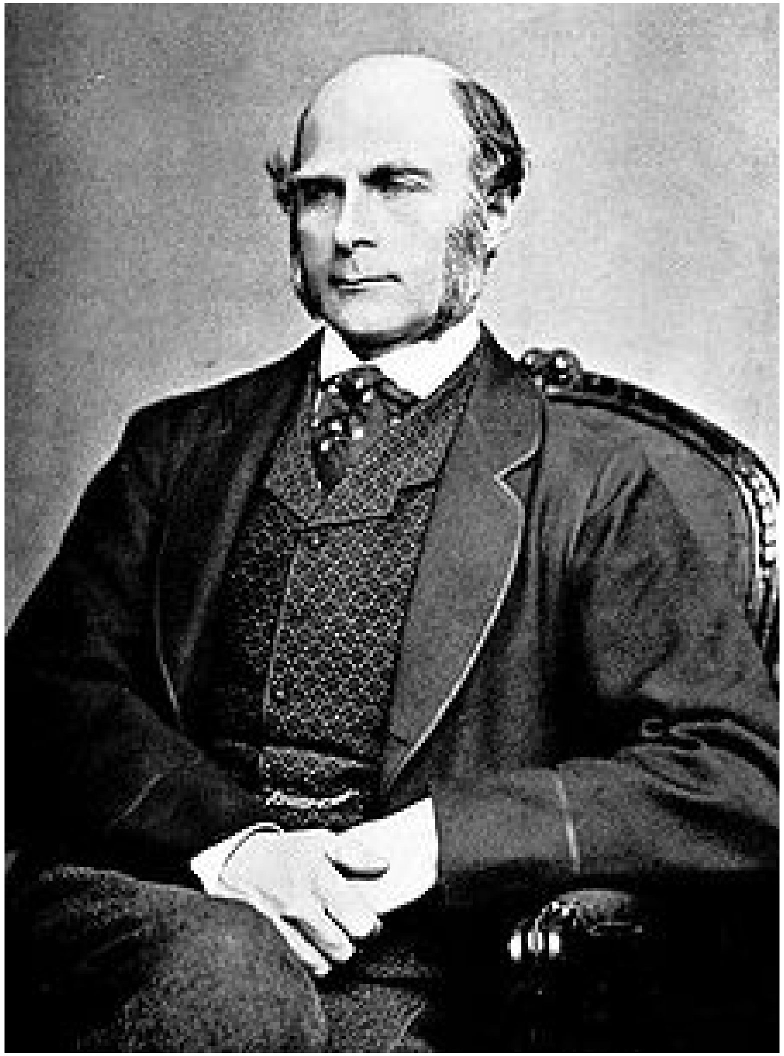
Example: Price Analysis for Diamond Rings

Variables

- Response variable (dependent variable) Y : price in dollars;
- Predictor variable (independent variable) X: weight of diamond in carats;
- Want to discover the relationship between price for diamond rings and weight of diamond in carats

Regression analysis

- A statistical methodology that utilizes the relation between response variable and predictor variable, so that a response variable can be predicted from the predictor variables
- The term “regression” was coined by Francis Galton (1822-1911, England) to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).



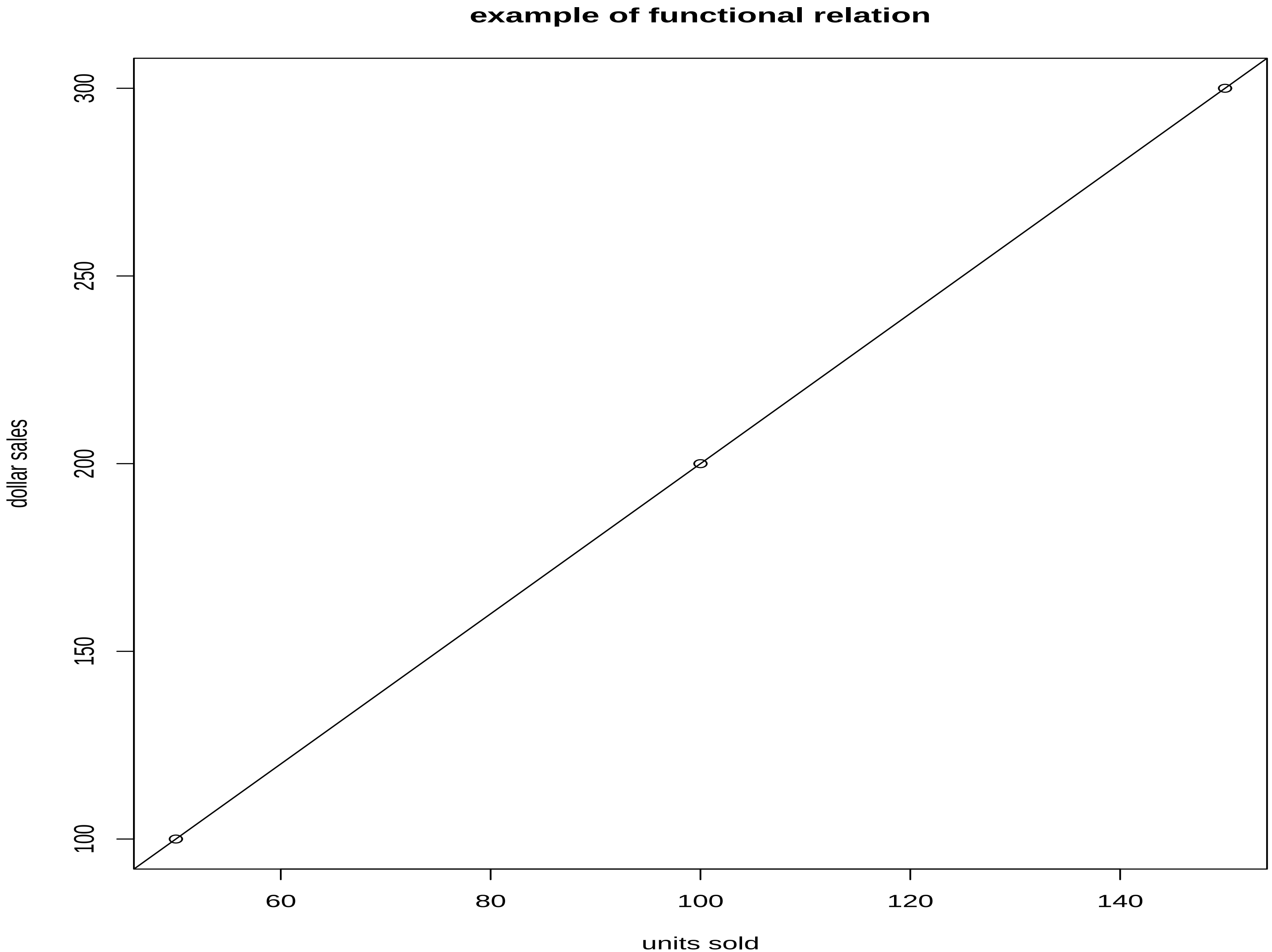
- Galton’s work was later extended by Yule, Pearson and Fisher to a more general statistical context.

Relations between variables

Functional relation between two variables: expressed by a mathematical formula

Example: consider a product's sale

- y: Dollar sales
- x: Number of units sold
- Selling price: \$2 per unit
- The relation between dollar sales and number of units sold is expressed by the equation $y = 2x$

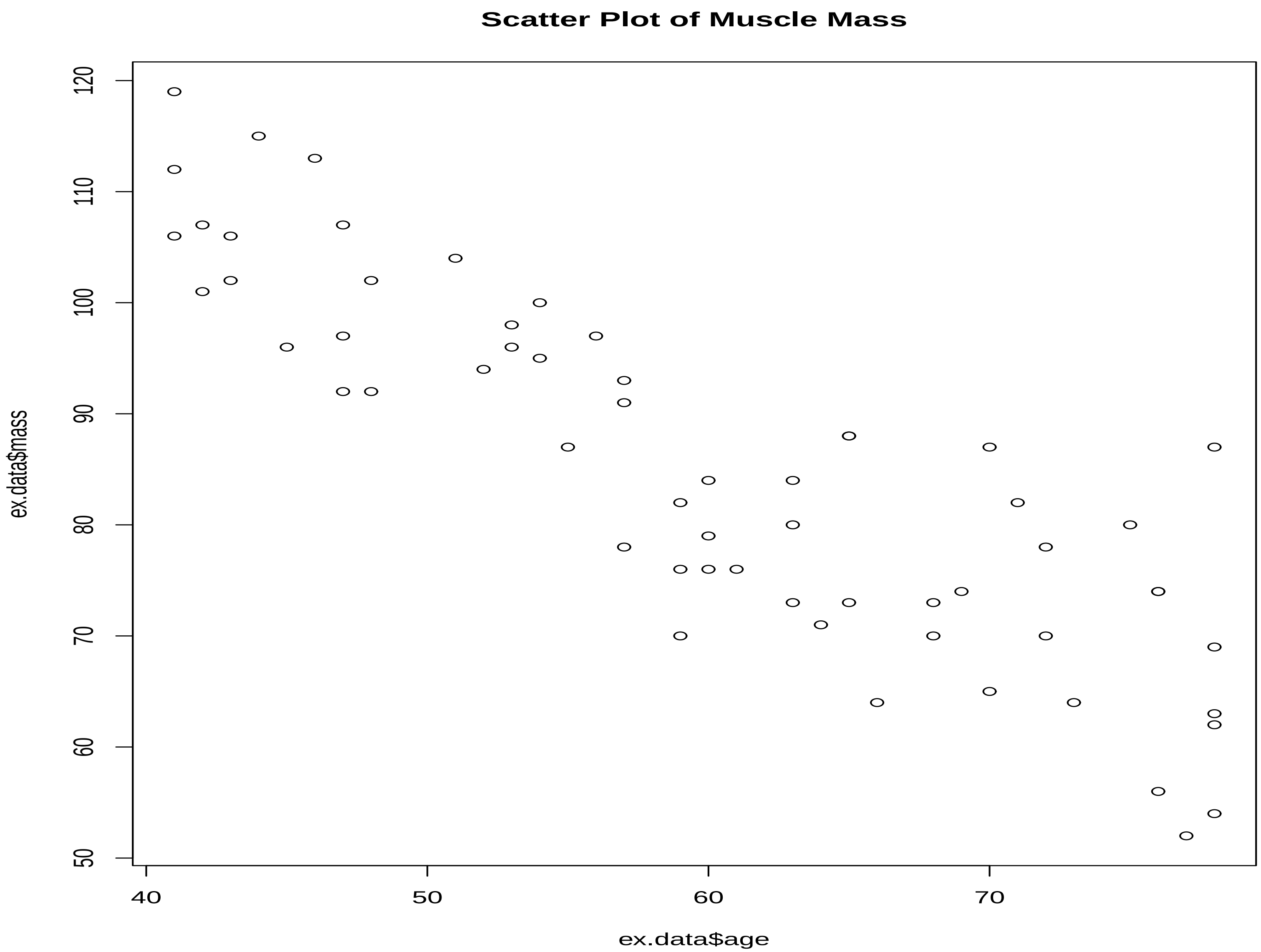

example of functional relation
dollar sales
300
250
200
150
100
60
80
100
120
140
units sold

Statistical relation between two variables

- Not a perfect relation
- In general, the observations for a statistical relation do not fall directly on the curve of relationship
- Statistical relation could be very useful, even though they do not have the exactitude of a functional relation

Example: A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79 with a total number of 60 women.

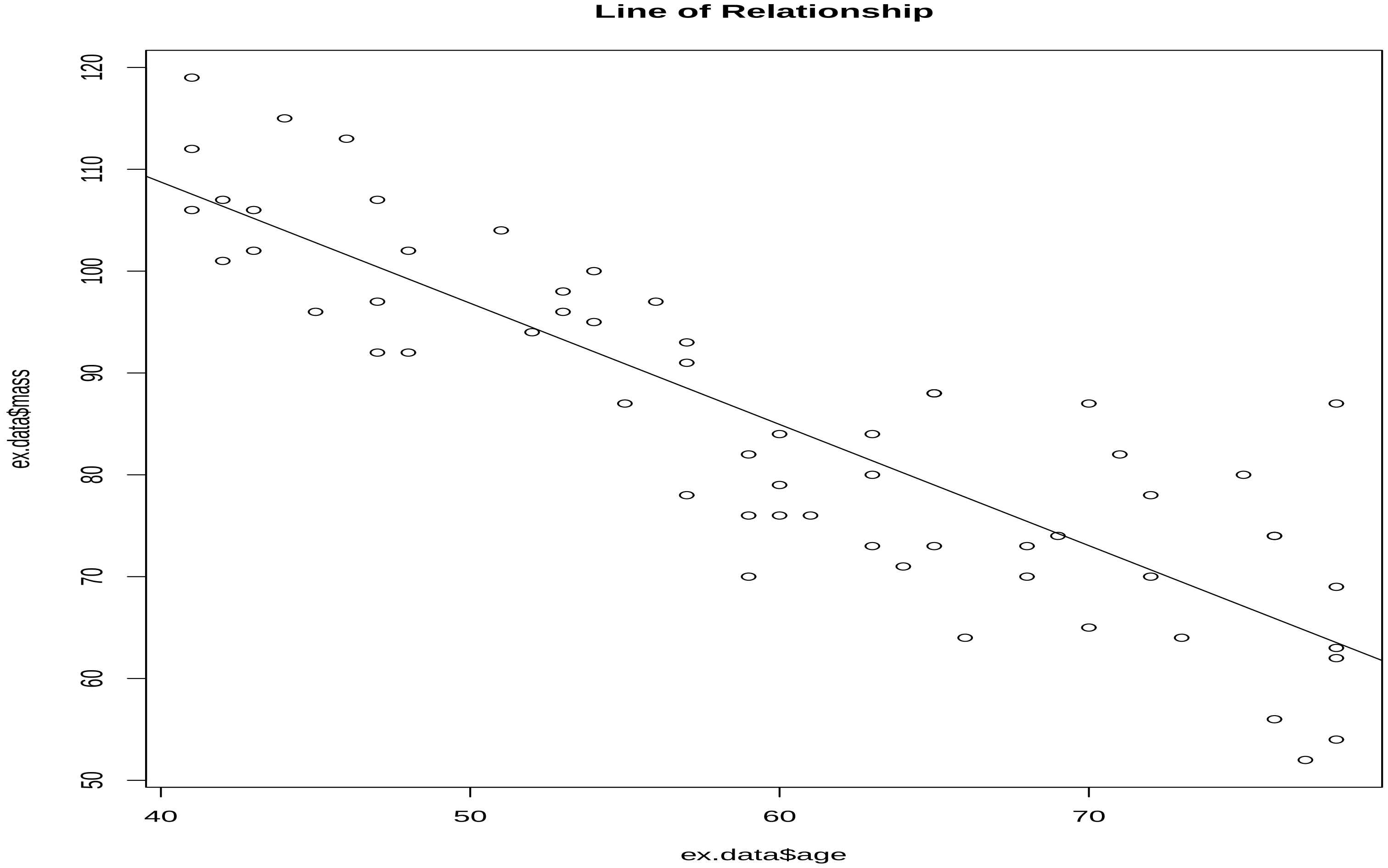
- Scatterplot of data



Scatter Plot of Muscle Mass
ex.data$mass
ex.data$age
50
60
70
80
90
100
110
120
40
50
60
70

- In statistical terminology, each point in the scatter plot represents a trial or a case
- Plot suggests a negative relation between age and muscle mass in women
- Clearly, the relation is not a perfect one
- Variation in muscle mass is not accounted for by age

Plot a line of relationship that describes the statistical relation between muscle mass and age



- The line indicates the general tendency by which muscle mass vary with age

- Most of the points do not fall directly on the line of statistical relationship

- The scattering of points around the line represents variation in muscle mass that is not associated with age and that is usually considered to be of a random nature

Simple linear regression and multiple linear regression

- One response variable for both simple linear and multiple linear regression
- Simple linear regression—one predictor variable
- Multiple linear regression—more than one predictor variable

Regression models

A regression model is a formal means of expressing

- A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion

- A scattering of points around the curve of statistical relationship by assuming that

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

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

Regression and Causality

Example:

- Subjects: a sample of young children aged $5 - 10$
- Predictor variable X: size of vocabulary
- Response variable Y: writing speed
- Data shows a positive regression relation

Question: Can we draw the conclusion from the data that an increase in vocabulary causes a faster writing speed?

- **No**, the positive relation discovered from the data doesn't imply that an increase in vocabulary causes a faster writing speed

- Other explanatory variables, such as age of the child and amount of education, affect both the vocabulary X and the writing speed Y

- The existence of a statistical relation between the response variable Y and the predictor variable X does not imply in any way that Y depends causally on X

- Regression analysis by itself provides no information about causal patterns and must be supplemented by additional analyses to obtain insights about causal relations

Use of R software

- Download a copy from www.r-project.org
- Practice using R handout1

Review: Regression analysis

- Discover the relationship between response variable and predictor variable

- Statistical relationship, not a perfect exact relationship: the value of the variable to be predicted do not fall exactly on a curve

- Assume a distribution of possible y values for each x value, usually assume a normal distribution

Simple linear regression model—one predictor variable

Data: $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$ could be arise in two ways

- Experimental setting, control x values and observe y values

  Example: An insurance company wishes to study the relation between productivity of its analysts in processing claims and length of training.

  —Nine analysts are to be used in the study

  —Three of them are selected at random and trained for two weeks, three for three weeks, and three for 5 weeks

  —Observe the productivity of the analysts during the next 10 weeks

- Observational setting, the value of x and y come as pairs from nonexperimental studies, we do not set the x value first

  Example: height, weight

- Notes:

  —Regression analyses are frequently based on observational data

  —Major limitation of observational data is that they often do not provide adequate information about cause-and-effect relationships

  —When control over the predictor variable(s) is exercised through random assignments, the resulting experimental data provide much stronger information about the cause-and-effect relationships than do observational data

Simple linear regression model

Y = (systematic part) + (random part)

= $(\beta_0 + \beta_1 x) + (\epsilon)$

Formal Statement of simple linear regression model with distribution of error terms unspecified

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

- y_i: the value of the response variable in the ith trial
- x_i: a known constant, the value of the predictor variable in the ith trial
- β_0 and β_1: parameters
- ϵ_i: random error term

—$E(\epsilon_i) = 0$

—$V(\epsilon_i) = \sigma^2$

—ϵ_i and ϵ_j are uncorrelated, $cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$

Comments:

- Simple: there is only one predictor variable
- Linear:

—-Linear in the parameters: no parameter appears as an exponent or is multiplied or divided by another parameter

Example of nonlinear regression model:

$$y = e^{\beta_0 + \beta_1 x} + \epsilon$$

$$y = \text{sin}\beta_0\beta_1 x + e^{\beta_1 x} + \epsilon$$

—-Linear in the predictor variable: the predictor variable appears only in the first power

—-A model that is linear in the parameters and in the predictor variable is also called a first-order model

Important features of model:

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the sum of a constant term and the random term. Hence, y_i is a random variable
- $E(y_i) = E(\beta_0 + \beta_1 x_i) + E(\epsilon_i) = \beta_0 + \beta_1 x_i, V(y_i) = \sigma^2$. Regression model (1) implies that the response y_i come from probability distributions whose means are $\beta_0 + \beta_1 x_i$ and whose variances are σ^2. Two responses y_i and y_j are uncorrelated
- Regression line: $E(y) = \beta_0 + \beta_1 x$
- β_1: slope of the regression line gives the change in mean value of y for a unit change in x

- β_0: y-intercept, β_0 is the mean of y when $x = 0$. So only if $x = 0$ is in the domain(scope), β_0 is interpretable
- ϵ_i: $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$, deviation between the observed y and the expected mean of y at the given x

Example: A consultant for an electrical distributor is studying the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week and the time required to prepare the bids. Suppose regression model is applicable and is as follows

$$y_i = 9.5 + 2.1x_i + \epsilon_i$$

- y_i: number of hours required to prepare the bids
- x_i: number of bids prepared in a week

- Regression function: $E(y) = 9.5 + 2.1x$

  if $\quad x_i = 25, E(y_i) = 9.5 + 2.1 \times 25 = 62$

  if $\quad x_i = 45, E(y_i) = 9.5 + 2.1 \times 45 = 104$

- ϵ_i: the deviation of y_i from its mean value $E(y_i)$

  —If $x_i = 45$ and $y_i = 108$

  —Then the error term $\epsilon_i = y_i - E(y_i) = 108 - 104 = 4$

- $\beta_1 = 2.1$ indicates that the preparation of one additional bid in a week leads to an increase in the mean of the probability distribution of y of 2.1 hours

- $\beta_0 = 9.5$ indicates the value of the regression line at $x = 0$. But, linear regression model was formulated to apply to weeks where the number of bids prepared ranged from 20 to 80, so β_0 does not have any intrinsic meaning of its own here

  Question: how to estimate the regression function parameters β_0, β_1 and σ^2 using information from the available data?

Least square estimators:

- Consider the deviation of y_i from its expected value $[y_i - (\beta_0 + \beta_1 x_i)]$
- Measure:

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Objective: to find estimates b_0 and b_1 for β_0 and β_1 respectively, for which Q is minimum

Steps to find LS estimators

- Take partial derivatives from Q and set to 0

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

- Normal equations

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

- Solve normal equations to find least square estimators of β_0 and β_1

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

$$b_0 = \bar{y} - b_1 \bar{x}$$

Computational formula

$$b_1 = \frac{\sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n}$$

- Regression line $E(y) = \beta_0 + \beta_1 x$ is estimated by

$$\hat{y} = b_0 + b_1 x$$

Properties of least squares estimators

- Unbiased: $E(b_0) = \beta_0$ and $E(b_1) = \beta_1$

- b_1 is a linear combination of the y_i and hence a linear estimator. So is b_0.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} k_i y_i$$

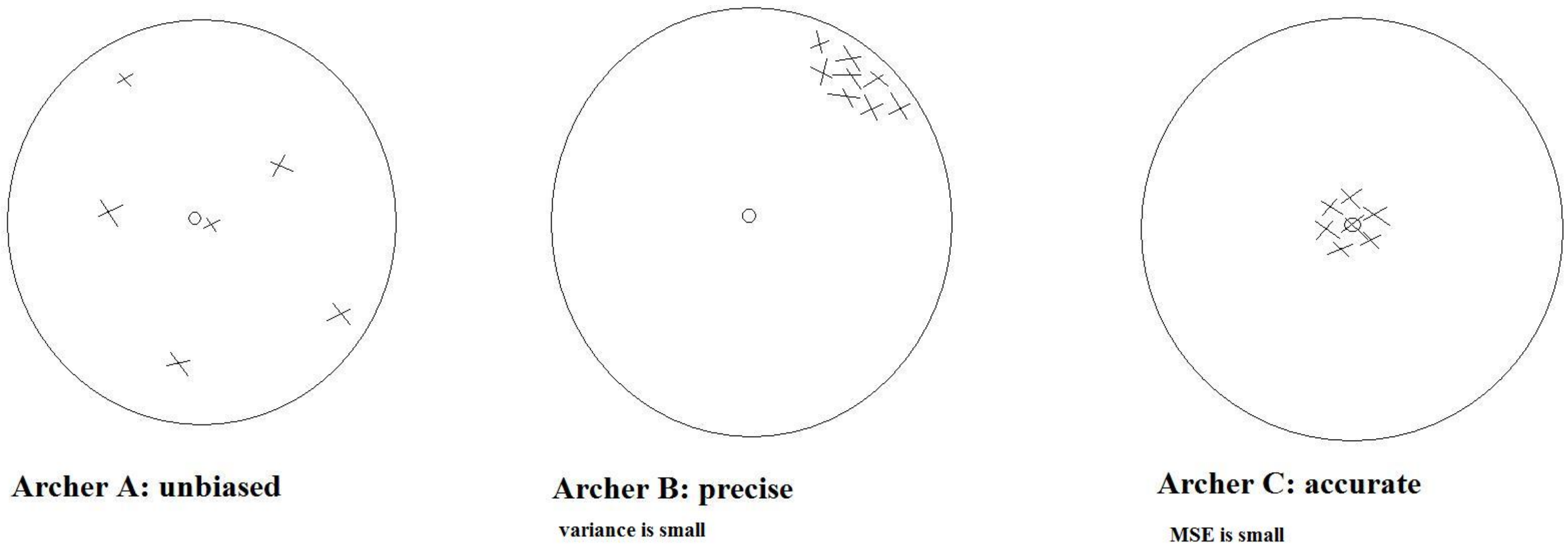
where

$$k_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- b_0 and b_1 have the smallest possible variance than any other estimators belonging to the class of unbiased estimators that are linear functions of the observed y values

- $\hat{y} = b_0 + b_1 x$ is an unbiased estimate of the regression line $E(y) = \beta_0 + \beta_1 x$

Figure 1: Unbiased, precise and accurate archers



Notations: consider model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

- Predicted (fitted or mean) value of y_i at x_i: $\hat{y}_i = b_0 + b_1 x_i$

  —the fitted value $\hat{y}_i$ is not the same as y_i

  —y_i is the observed value and $\hat{y}_i$ is the predicted value

- Residual $e_i = y_i - \hat{y}_i$: vertical deviation between y_i and the estimated regression function

- Error term $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$: vertical deviation between y_i and the true regression line

- Residual e_i is a prediction of ϵ_i

  —- $e_i \neq \epsilon_i$

Properties of residuals and fitted values

- $\sum_{i=1}^{n} e_i = 0$
- $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$ = smallest value of Q
- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$, so $\bar{y} = \bar{\hat{y}}$
- $\sum_{i=1}^{n} x_i e_i = 0$, x_i's and e_i's are uncorrelated
- $\sum_{i=1}^{n} \hat{y}_i e_i = 0$, $\hat{y}_i$'s and e_i's are uncorrelated
- the fitted regression line passes through $(\bar{x}, \bar{y})$

$$\hat{y} = b_0 + b_1 x = \bar{y} - b_1 \bar{x} + b_1 x$$

Estimation of σ^2

To estimate σ^2, we find the deviation of each y value from its mean and square the deviation sum, then divided by a function of n to get an average deviation

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2}{n-2} \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n} e_i^2\end{aligned}$$

where e_i is the ith residual $y_i - \hat{y}_i$

Notes:

- Sum of squares due to error or error sum of squares or residual sum of squares: $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

- Mean square error or error mean square: $MSE = SSE/(n-2) = \hat{\sigma}^2$

Computational formulas for SSE

$$SSE = \sum_{i=1}^{n} y_i^2 - b_0 \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i y_i$$

or

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

-

$$E(MSE) = \sigma^2$$

Normal Error in Simple Linear Regression Model

- To do statistical inference, testing hypothesis and to construct confidence interval, we need to make an assumption about the distribution of ϵ in the regression model
- Common assumption: ϵ is a normal distribution
- Why assume normal distribution of errors?

  —Sometimes the errors have approximately normal distributions

  —We get nice methods for statistical inferences

  —If the errors are only approximately normal, the methods developed assuming normality still perform approximately as we would expect

Review:

- under normal distributions, independence and uncorrelated are the same

  Uncorrelated $\Leftrightarrow$ Independence

- This is not true in general, in general

  Uncorrelated $\not\Rightarrow$ independence

  Independence $\Rightarrow$ uncorrelated

Normal error regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i: observed response in the ith trial
- x_i: a known constant, the level of the predictor variable in the ith trial
- β_0 and β_1: parameters
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2, \cdots, n$
- $E(y_i) = \beta_0 + \beta_1 x_i$, var$(y_i) = \sigma^2$

Estimation of parameters by method of maximum likelihood —-The functional form of the probability distribution of the error terms is specified

- The density of an observation y_i for the normal error regression model is

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \text{exp}[-\frac{1}{2}(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma})^2]$$

- The likelihood function for n observations $y_1, y_2, \cdots, y_n$ is the product of the individual densities

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \text{exp}[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \text{exp}[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2] \end{aligned}$$

- The values of β_0, β_1 and σ^2 that maximize this likelihood function are the maximum likelihood estimators and are denoted by $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$

Parameter	MLE	LS	Relation
β_0	$\hat{\beta}_0$	b_0	$\hat{\beta}_0 = b_0$
β_1	$\hat{\beta}_1$	b_1	$\hat{\beta}_1 = b_1$
σ^2	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$	MLE: biased v.s. LS: unbiased

Example: (page 15) In a small scale study of persistence, an experimenter gave three subjects a very difficult task. Data on the age of the subject x and on the number of attempts to accomplish the task before giving up y is as follows

Subject i	1	2	3
Age x_i	20	55	30
# of attempts y_i	5	12	10

$n = 3$

$(x_1, y_1) = (20, 5)$

$(x_2, y_2) = (55, 12)$

$(x_3, y_3) = (30, 10)$

$\bar{x} = 35$

$\bar{y} = 9$

$$b_1 = \frac{\sum_{i=1}^{3}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{3}(x_i - \bar{x})^2} = .177$$

$$b_0 = \bar{y} - b_1\bar{x} = 2.81$$

$$\hat{y} = 2.81 + .177x$$

Predicted value

$\hat{y}_1 = 2.81 + .177 \times 20 = 6.35$

$\hat{y}_2 = 2.81 + .177 \times 55 = 12.538462$

$\hat{y}_3 = 2.81 + .177 \times 30 = 8.12$

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

$$e_1 = 5 - 6.35 = -1.35$$

$$e_2 = 12 - 12.5384 = -.5384$$

$$e_3 = 10 - 8.12 = 1.8846154$$

Estimate σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{3}(y_i - \hat{y}_i)^2}{n-2} = \sum_{i=1}^{3} e_i^2 = 5.66415$$