University of Ferrara

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

Stefano Bonnini

Multiple Regression Analysis

Summary

- Brief notes on probability and inference
- Simple linear regression analysis
- Multiple linear regression analysis

Brief Notes on Probability and Inference

Games of chance

Game where a randomizing device (dice, playing cards, roulette wheels, lottery, …) influences the outcome

→ Each of the possible outcomes has a given probability of occurrence

→ **Probability distribution**: each event ***E*** is given a probability ***P(E)***

↓

Probability function for discrete numerical variables → P(x): probability of number x

$\Sigma_{x \in A}$P(x): probability of the set A

Probability density function for continuous numerical variables → *f(x)*: density of x

$\int_{x \in A} f(x)dx$: probability of A

Brief Notes on Probability and Inference

Inferential Statistics and Probability Theory

Inferential Statistics is based on

(a) **Sample** data processing for obtaining information

(b) Probabilistic **assumptions**

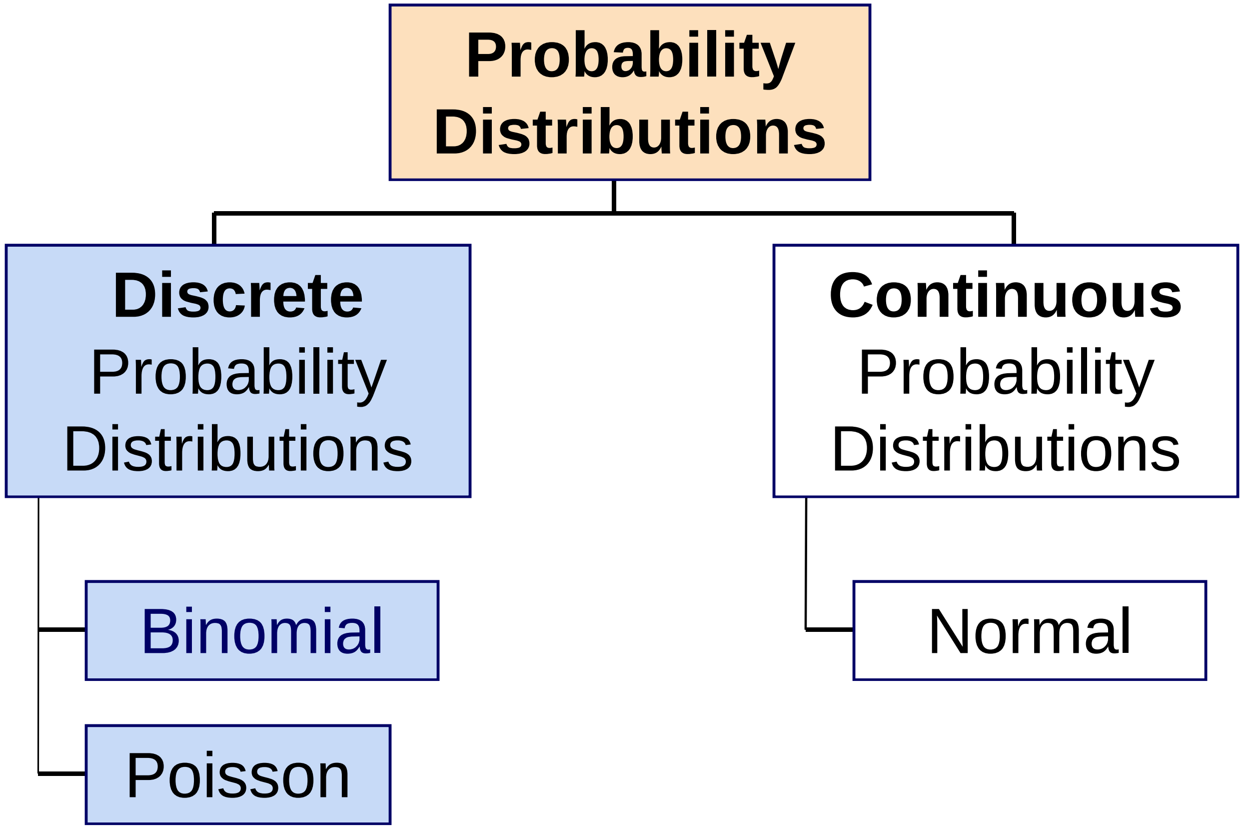
Observed data are determinations of random variables X_1 ,…,X_n characterized by unknown probability distributions (e.g. Binomial, Poisson, Normal, …)

Information provided by sample statistics leads to plausible (but uncertain) results, also computing the risk of making wrong decisions

Parametric methods assume that distribution functions are known except for some unknown parameters

Nonparametric methods are based on less stringent assumptions and give more importance on (a)

Probability Distributions



Brief Notes on Probability and Inference

- A **probability distribution for a discrete random variable** is a mutually exclusive listing of all possible numerical outcomes for that variable and a probability of occurrence associated with each outcome.

Number of Classes Taken	Probability
2	0.2
3	0.4
4	0.24
5	0.16

$$\mu = \mathrm{E(X)} = \sum_{i=1}^{N} X_i P(X_i) = 2\cdot 0.2 + 3\cdot 0.4 + 4\cdot 0.24 + 5\cdot 0.16 = 3.36$$

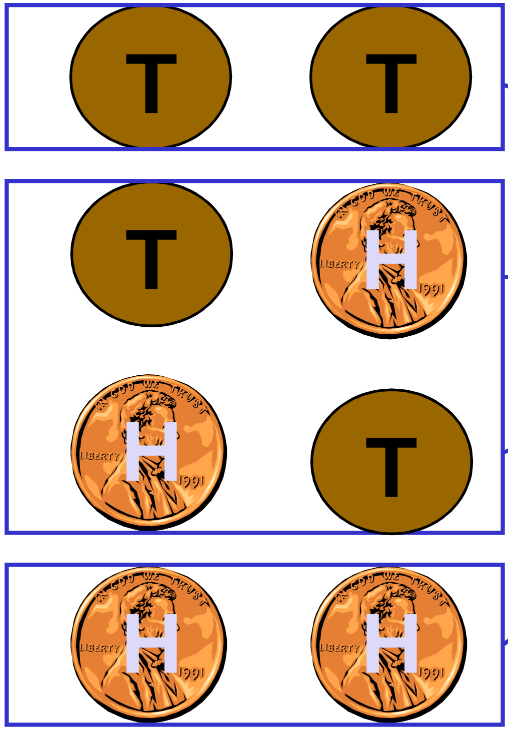
$$\sigma^2 = \sum_{\mathrm{i=1}}^{\mathrm{N}} [\mathrm{X_i} - \mathrm{E(X)}]^2 \mathrm{P(X_i)} = (2-3.36)^2 \cdot 0.2 + (3-3.36)^2 \cdot 0.4 + (4-3.36)^2 \cdot 0.24 + (5-3.36)^2 \cdot 0.16 = 0.9504$$

$$\sigma = \sqrt{\sigma^2} = 0.9749$$

Brief Notes on Probability and Inference

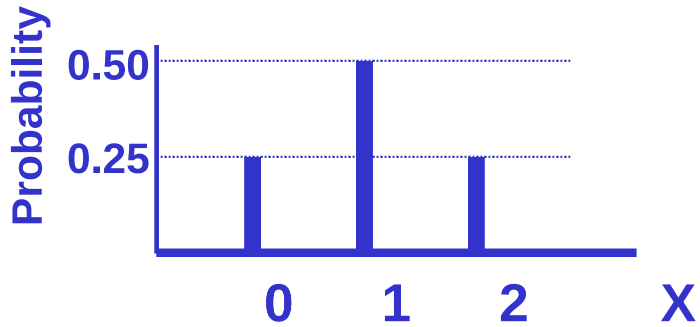
Experiment: Toss 2 Coins. Let X = # heads.

4 possible outcomes



Probability Distribution

X Value	Probability
0	1/4 = 0.25
1	2/4 = 0.50
2	1/4 = 0.25



Brief Notes on Probability and Inference

- Expected Value (or mean) of a discrete random variable (Weighted Average)

$$\mu = E(X) = \sum_{i=1}^{N} X_i P(X_i)$$

- Example: Toss 2 coins,
  X = # of heads,
  compute expected value of X:

E(X) = ((0)(0.25) + (1)(0.50) + (2)(0.25))
= 1.0

X	P(X)
0	0.25
1	0.50
2	0.25

Brief Notes on Probability and Inference

- Variance of a discrete random variable

$$\sigma^2 = \sum_{i=1}^{N} [X_i - E(X)]^2 P(X_i)$$

- Standard Deviation of a discrete random variable

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} [X_i - E(X)]^2 P(X_i)}$$

Brief Notes on Probability and Inference

(continued)

- Example: Toss 2 coins, X = # heads, compute standard deviation (recall E(X) = 1)

$$\sigma = \sqrt{\sum [X_i - E(X)]^2 P(X_i)}$$

$$\sigma = \sqrt{(0-1)^2(0.25) + (1-1)^2(0.50) + (2-1)^2(0.25)} = \sqrt{0.50} = 0.707$$

Possible number of heads = 0, 1, or 2

Brief Notes on Probability and Inference

- A continuous random variable is a variable that can assume any value on a continuum (can assume an uncountable number of values)
  - thickness of an item
  - time required to complete a task
  - temperature of a solution
  - height, in centimeters

$$\mu = \mathrm{E}(\mathrm{X}) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

$$\sigma^2 = \mathrm{E}(\mathrm{X} - \mu)^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)dx$$

f(x): probability density function

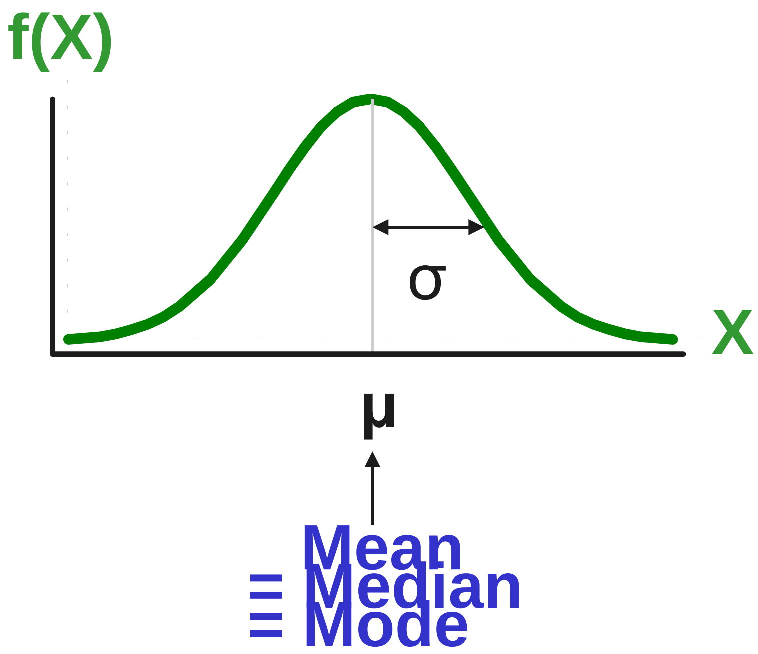
Brief Notes on Probability and Inference

- Bell Shaped
- Symmetrical
- Mean, Median and Mode are Equal

Location is determined by the mean, μ

Spread is determined by the standard deviation, σ

The random variable has an infinite theoretical range: $+\infty$ to $-\infty$



Brief Notes on Probability and Inference

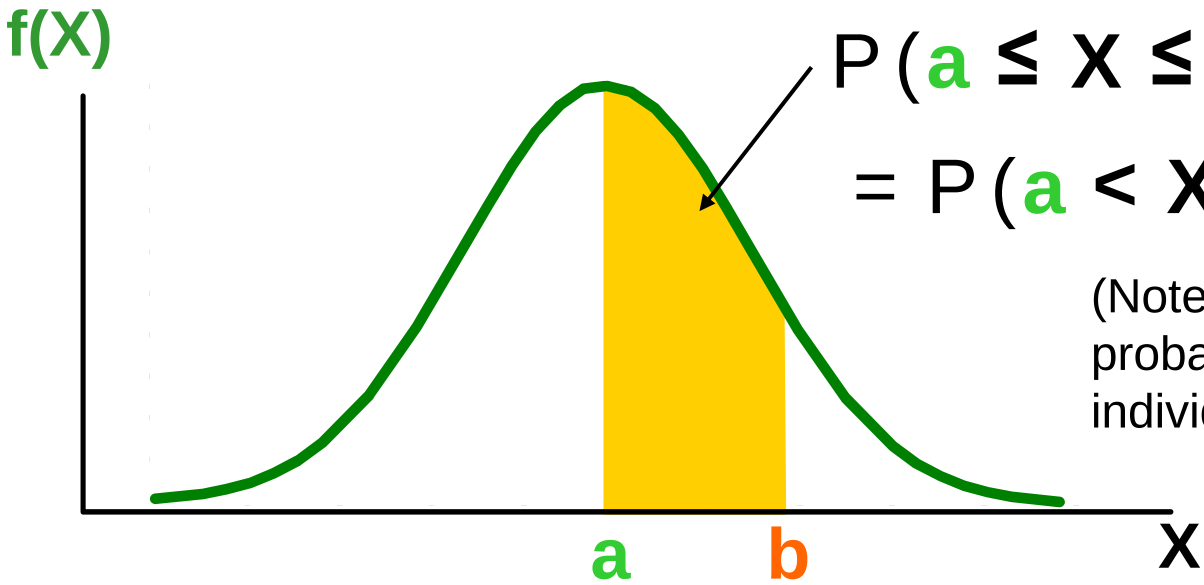
- The formula for the normal probability density function is

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2}$$

Where
e = the mathematical constant approximated by 2.71828
π = the mathematical constant approximated by 3.14159
μ = the population mean
σ = the population standard deviation
X = any value of the continuous variable

Brief Notes on Probability and Inference

Probability is measured by the area under the curve



$$P(a \leq X \leq b) = P(a < X < b)$$

(Note that the probability of any individual value is zero)

Brief Notes on Probability and Inference

- Not all continuous distributions are normal
- It is important to evaluate the plausibility of the assumption of normality.
- Normally distributed data should approximate the theoretical normal distribution:
  - The normal distribution is bell shaped (symmetrical) where the mean is equal to the median.
  - The empirical rule applies to the normal distribution.
  - The interquartile range of a normal distribution is 1.33 standard deviations.

Brief Notes on Probability and Inference

(continued)

Comparing data characteristics to theoretical properties

- Construct charts or graphs
  - For small- or moderate-sized data sets, construct a boxplot to check for symmetry
  - For large data sets, does the histogram or polygon appear bell-shaped?
- Compute descriptive summary measures
  - Do the mean, median and mode have similar values?
  - Is the interquartile range approximately 1.33 σ?
  - Is the range approximately 6 σ?

Brief Notes on Probability and Inference

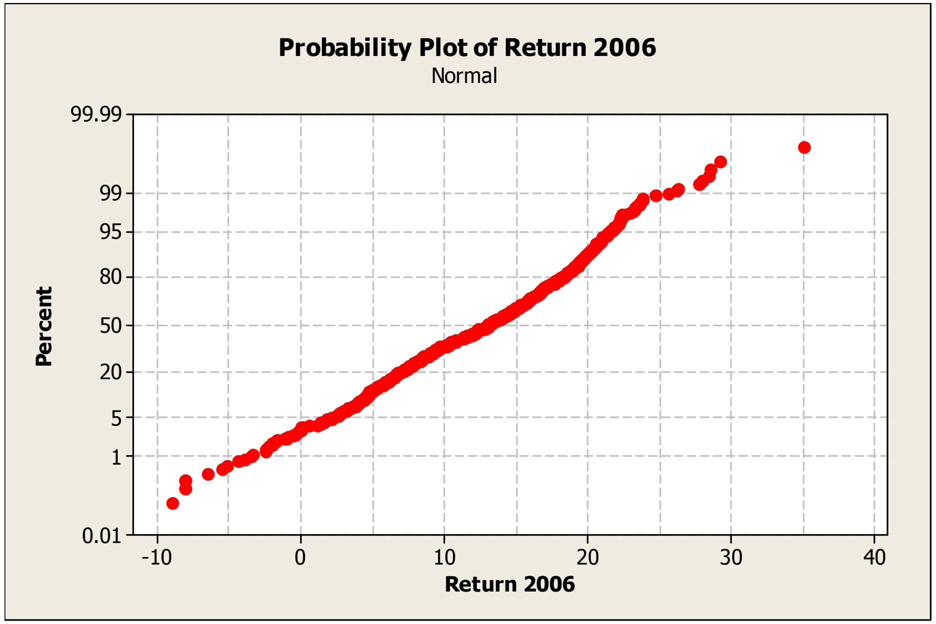
(continued)

Comparing data characteristics to theoretical properties

- Observe the distribution of the data set
  - Do approximately 2/3 of the observations lie within mean ±1 standard deviation?
  - Do approximately 80% of the observations lie within mean ±1.28 standard deviations?
  - Do approximately 95% of the observations lie within mean ±2 standard deviations?
- Evaluate normal probability plot
  - Is the normal probability plot approximately linear (i.e. a straight line) with positive slope?

Brief Notes on Probability and Inference

(continued)



Plot is approximately a straight line except for a few outliers at the low end and the high end.

Brief Notes on Probability and Inference

TEST OF HYPOTHESIS

A **test of hypothesis** is an inferential procedure based on sample data to test some assertions related to one or more populations

NULL HYPOTHESIS H_0

The **null hypothesis** usually corresponds to the status quo or the hypothesis of no effect, no difference, etc.

ALTERNATIVE HYPOTHESIS H_1

The **alternative hypothesis** represents the assertion that needs to be proved by the empirical evidence through sample data.

Summary

- Brief notes on probability and inference
- Simple linear regression analysis
- Multiple linear regression analysis

Simple Linear Regression Analysis

- A scatter plot can be used to show the relationship between two variables
- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - No causal effect is implied with correlation

Simple Linear Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable Y based on the value of at least one independent variable
  - Explain the impact on the dependent variable of changes in independent (explanatory) variables $X_1,\ldots,X_k$

Dependent variable: the variable we wish to predict or explain

Independent or explanatory variable(s): the variable(s) used to predict or explain the dependent variable

Simple Linear Regression Analysis

- Relationship between *Y* and $X_1,\ldots,X_k$ is described by a linear function
- Only **one** independent variable, X $\Rightarrow$ Simple Linear Regression Model
- ***k≥2*** independent variables, $X_1,\ldots,X_k \Rightarrow$ Multiple Linear Regression Model

Simple Linear Regression Analysis

Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

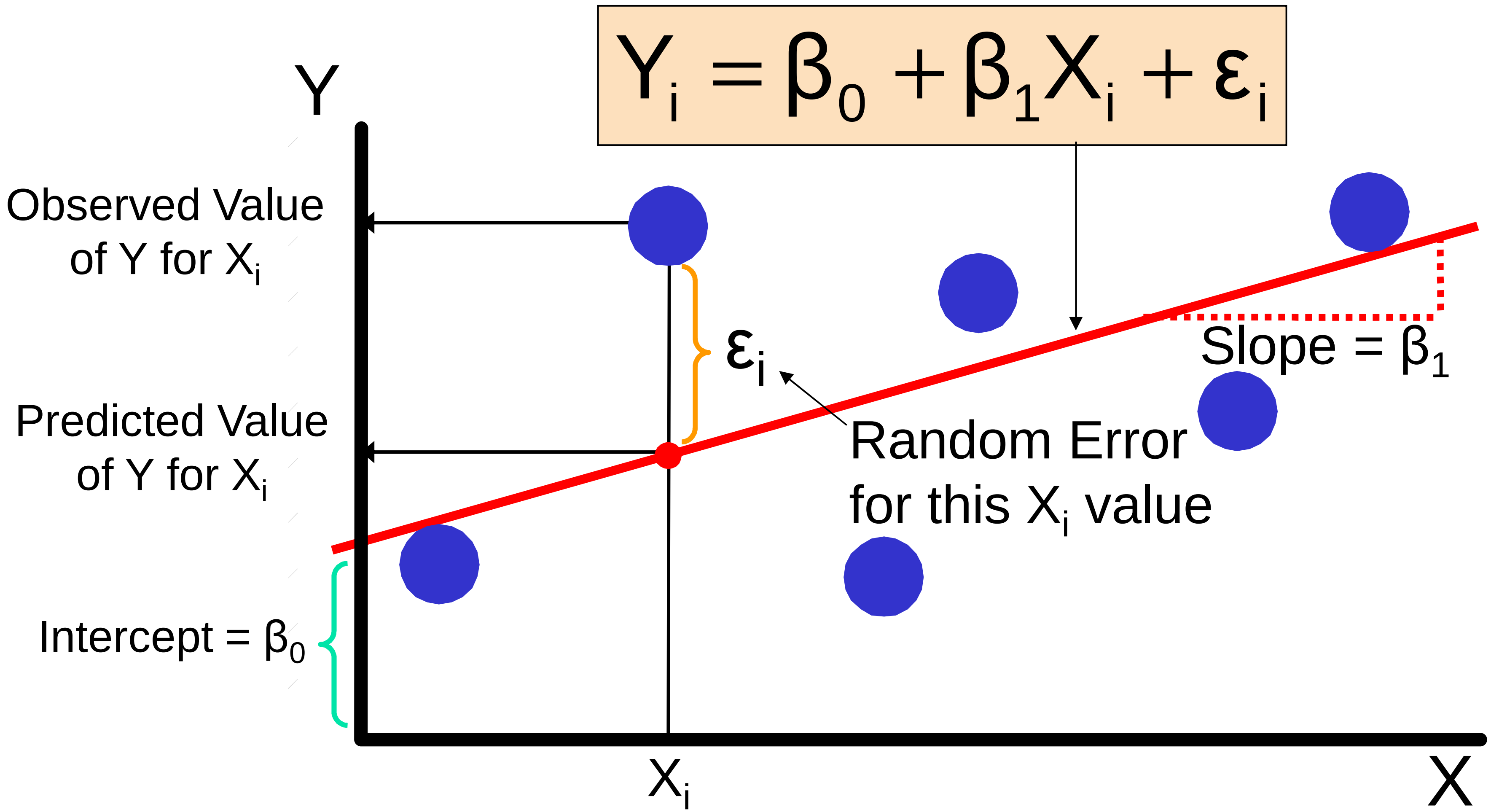
Random Error term

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

Linear component

Random Error component

Simple Linear Regression Analysis



Simple Linear Regression Analysis

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

Simple Linear Regression Analysis

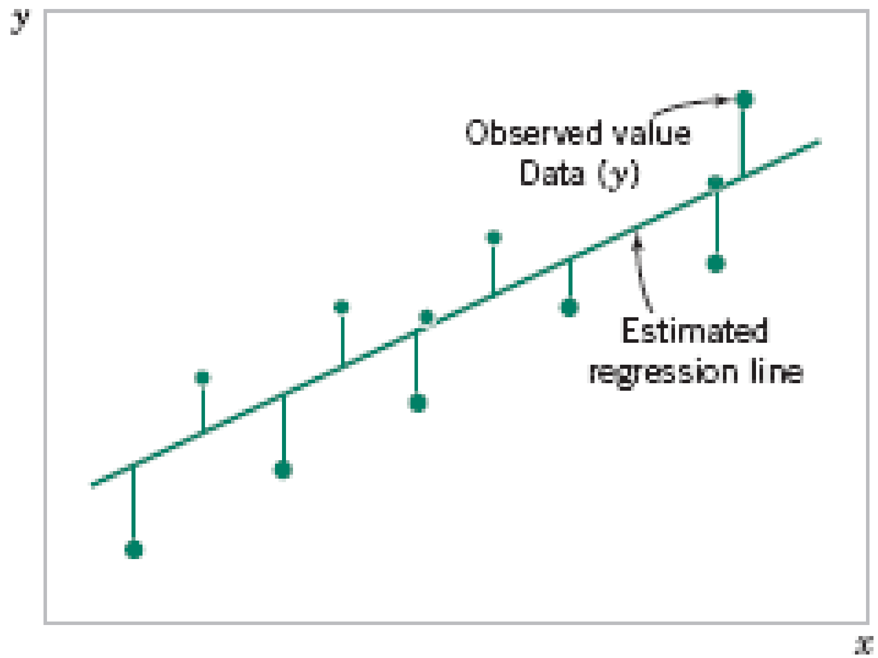
b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Simple Linear Regression Analysis

- Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

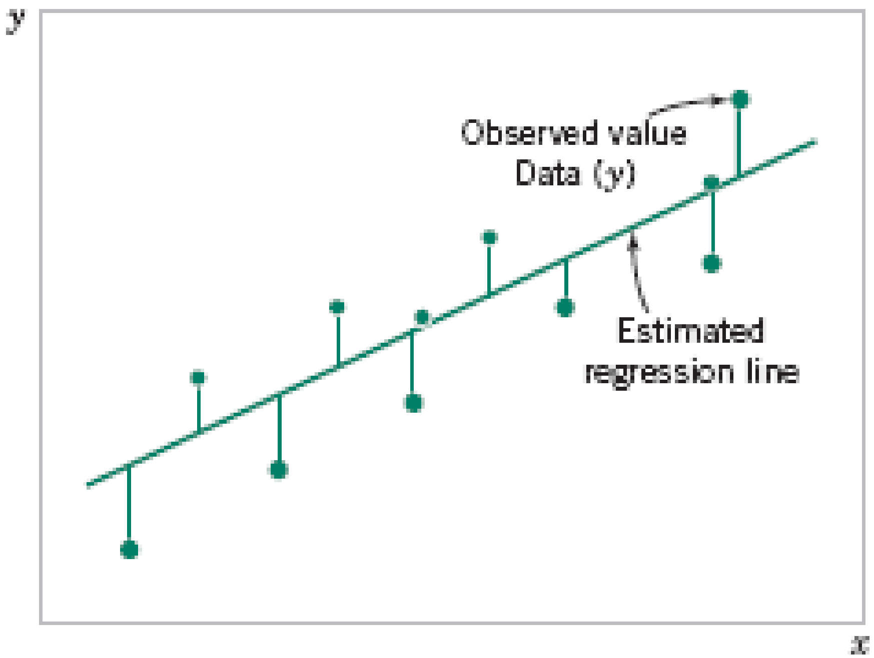
Deviations of the data from the estimated regression model.



Simple Linear Regression Analysis

• The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations.

Deviations of the data from the estimated regression model.



Simple Linear Regression Analysis

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

The least squares estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Simple Linear Regression Analysis

Simplifying these two equations yields

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad (11\text{-}6)$$

Equations 11-6 are called the **least squares normal equations.** The solution to the normal equations results in the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Simple Linear Regression Analysis

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} y_i x_i - \frac{\left(\displaystyle\sum_{i=1}^{n} y_i\right)\left(\displaystyle\sum_{i=1}^{n} x_i\right)}{n}}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11-8}$$

where $\bar{y} = (1/n) \sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$.

Simple Linear Regression Analysis

- $b_0 = \hat{\beta}_0$ is the estimated mean value of Y when the value of X is zero

- $b_1 = \hat{\beta}_1$ is the estimated change in the mean value of Y as a result of a one-unit change in X

Simple Linear Regression Analysis

Example:

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

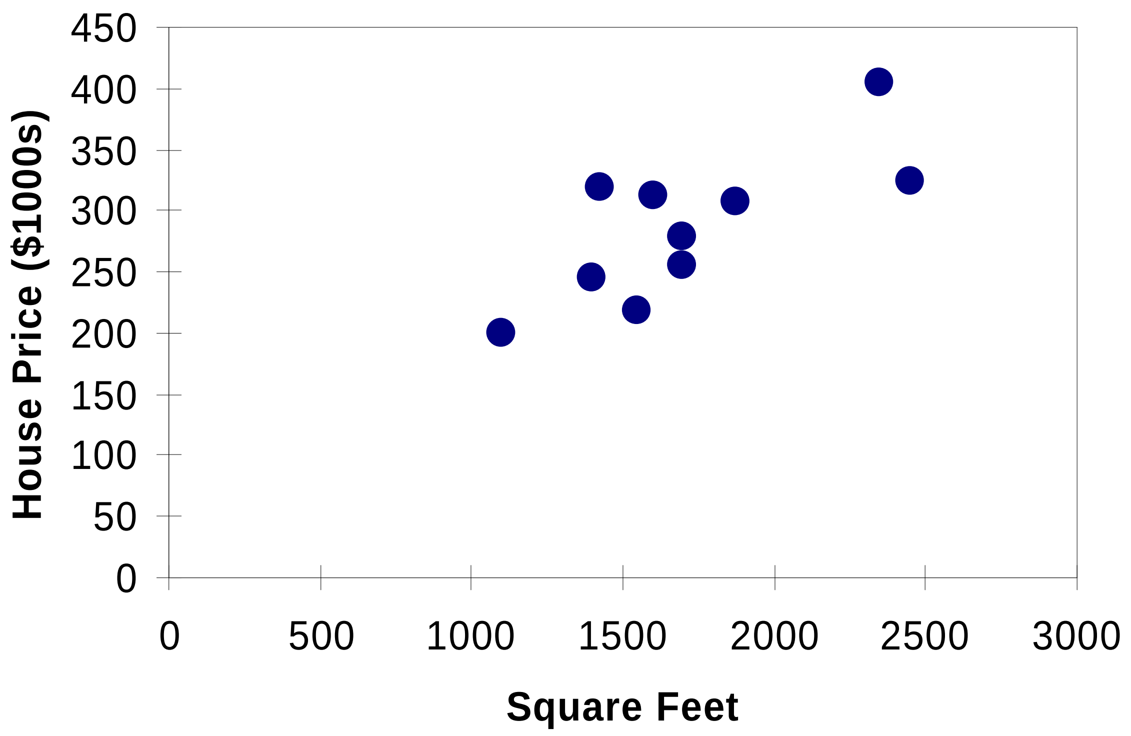
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Simple Linear Regression Analysis

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Simple Linear Regression Analysis

House price model: Scatter Plot



Simple Linear Regression Analysis

	Y	X	$(Y-\bar{Y})$	$(X-\bar{X})$	$(Y-\bar{Y})^2$	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$
	245	1400	-41.5	-315	1722.25	99225	13072.5
	312	1600	25.5	-115	650.25	13225	-2932.5
	279	1700	-7.5	-15	56.25	225	112.5
	308	1875	21.5	160	462.25	25600	3440
	199	1100	-87.5	-615	7656.25	378225	53812.5
	219	1550	-67.5	-165	4556.25	27225	11137.5
	405	2350	118.5	635	14042.25	403225	75247.5
	324	2450	37.5	735	1406.25	540225	27562.5
	319	1425	32.5	-290	1056.25	84100	-9425
	255	1700	-31.5	-15	992.25	225	472.5
sum	2865	17150	0	0	32600.5	1571500	172500
mean	286.5	1715			3260.05	157150	17250

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{172500}{1571500} = 0.109768$$

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1\bar{x} = 286.5 - 0.109768 \cdot 1715 = 98.24833$$

Simple Linear Regression Analysis

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

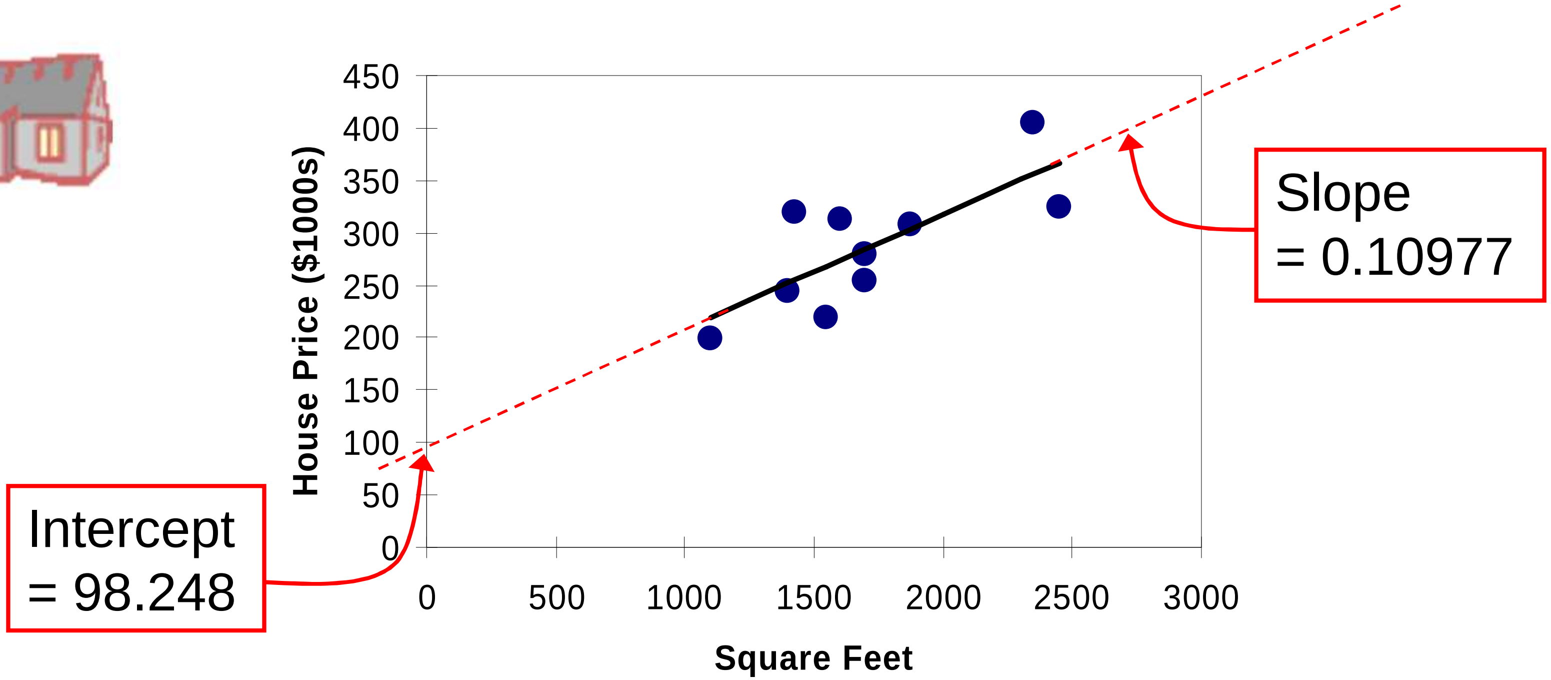
The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

House price model: Scatter Plot and Prediction Line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

Simple Linear Regression Analysis

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Simple Linear Regression Analysis

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \bar{Y})^2$	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

where:

$\bar{Y}$ = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

Simple Linear Regression Analysis

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression } \textit{sum} \text{ of squares}}{\textit{total} \text{ sum of squares}}$$

note: $$0 \leq r^2 \leq 1$$

Simple Linear Regression Analysis

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Simple Linear Regression Analysis

Assumptions of the model:

- Linearity
  - The relationship between X and Y is linear
- Independence of Errors
  - Error values are statistically independent
- Normality of Error
  - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance

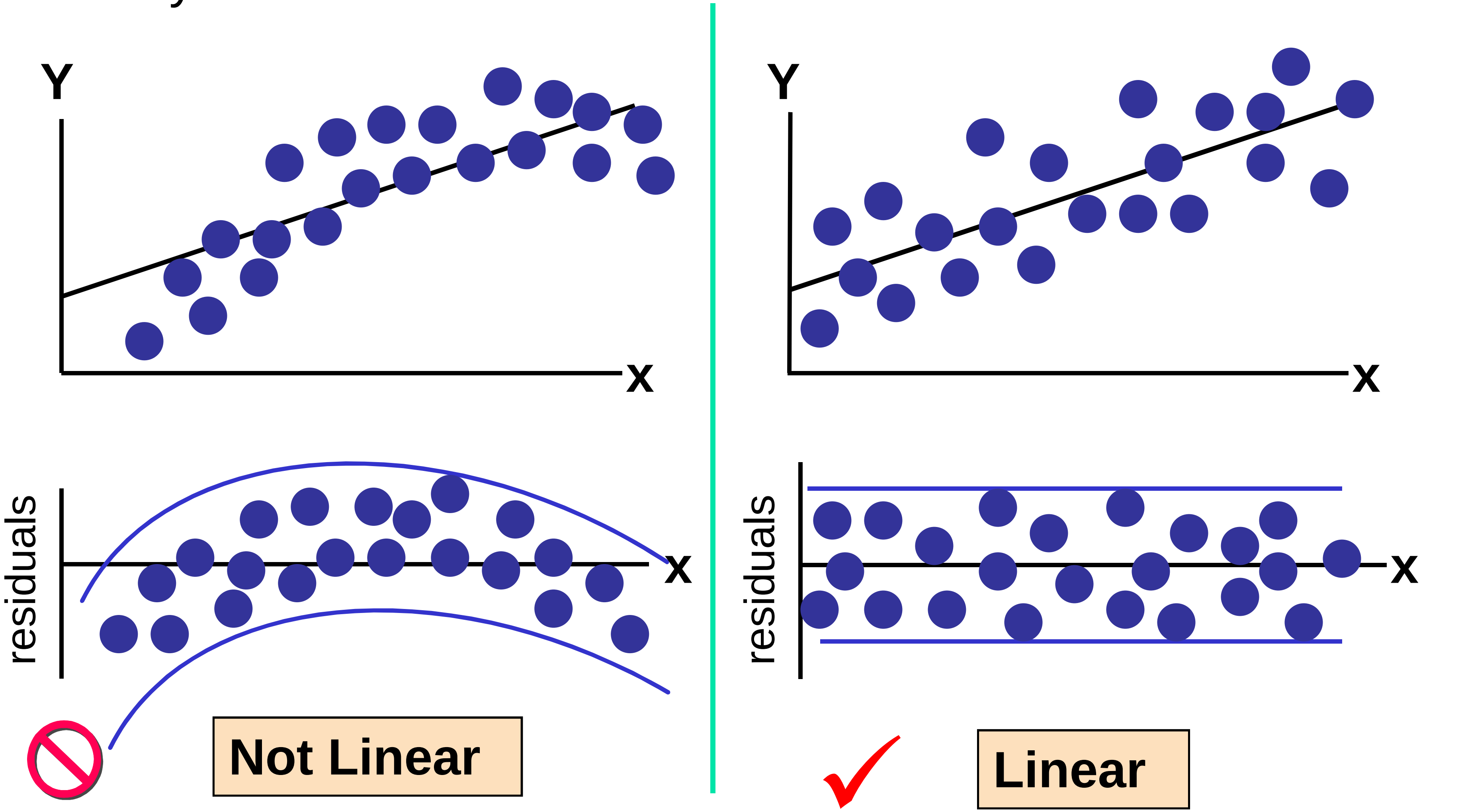
Simple Linear Regression Analysis

$$e_i = Y_i - \hat{Y}_i$$

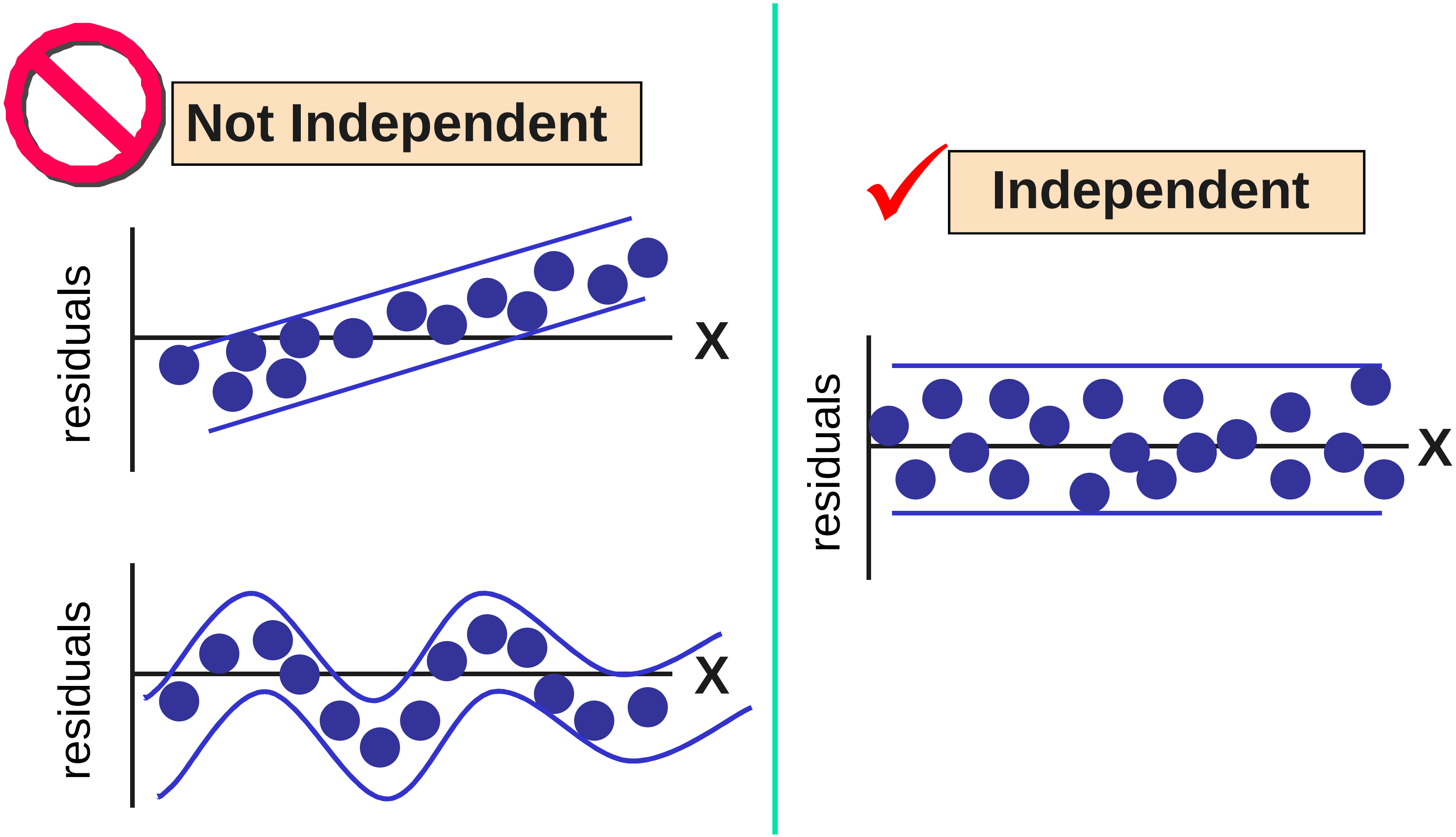
- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

Simple Linear Regression Analysis

Analysis of residuals



Simple Linear Regression Analysis



Simple Linear Regression Analysis

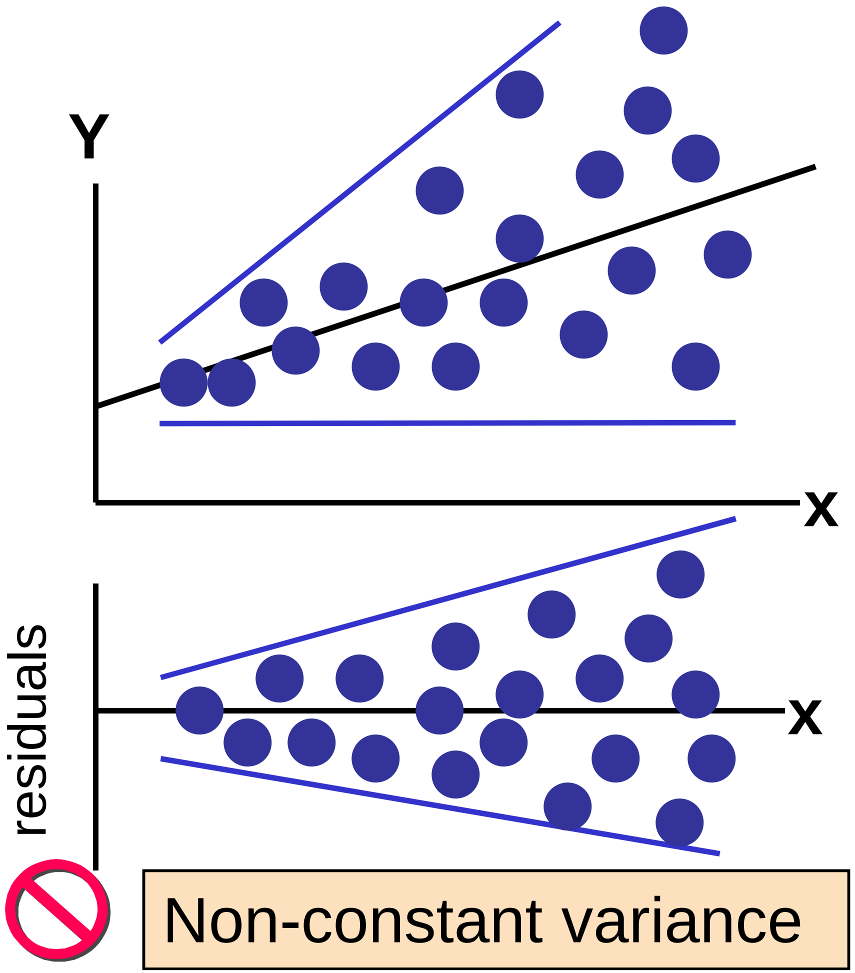
Checking for normality:

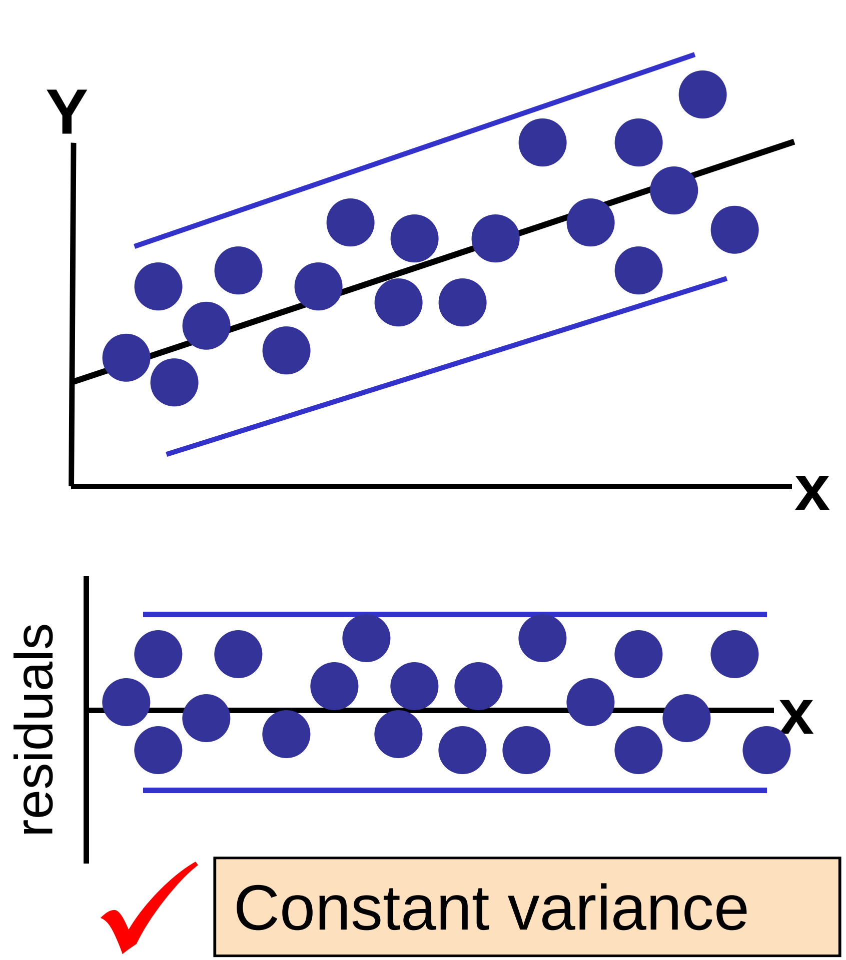
- Examine the Histogram of the Residuals

- Construct a Normal Probability Plot of the Residuals

Simple Linear Regression Analysis

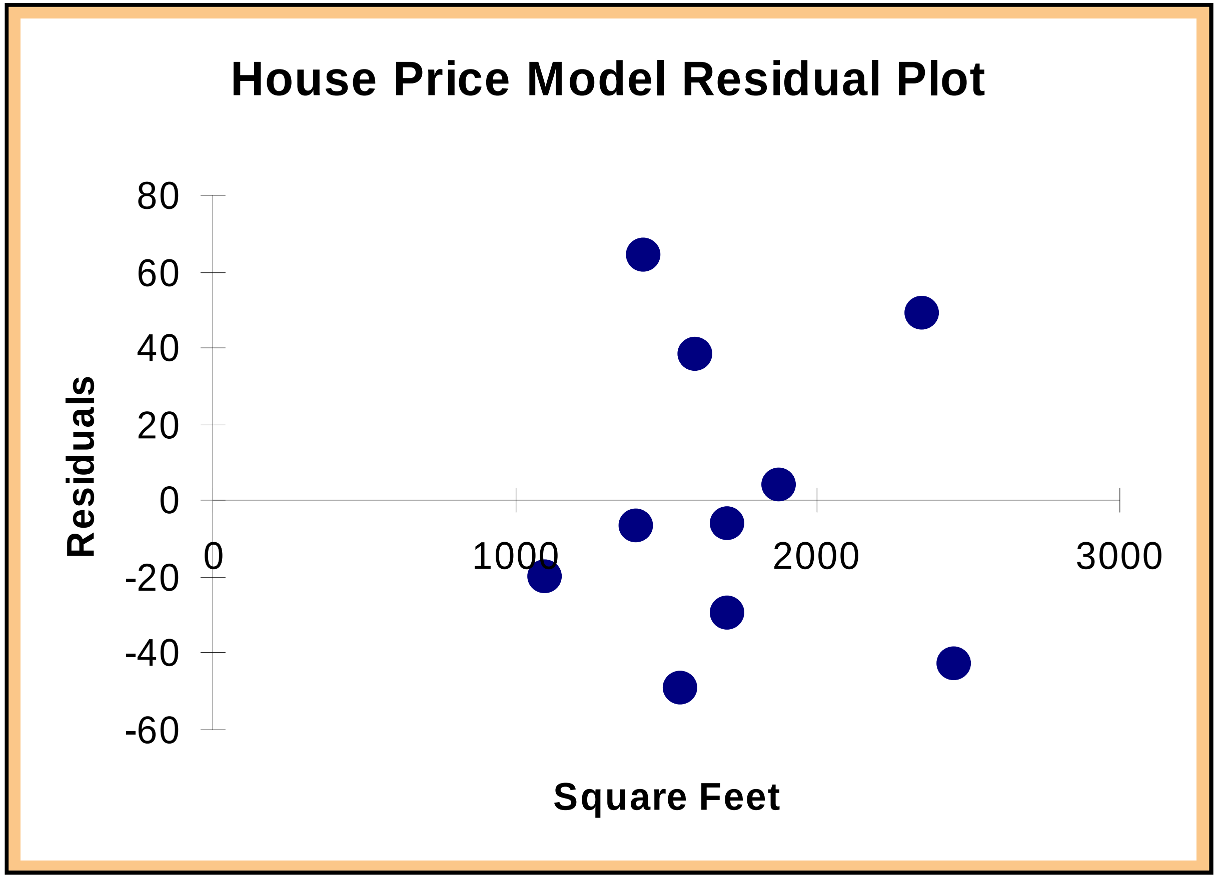
Checking for homoschedasticity





Simple Linear Regression Analysis

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Does not appear to violate any regression assumptions

Simple Linear Regression Analysis

- The standard error of the regression slope coefficient (b_1) is estimated by

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

where:

S_{b_1} = Estimate of the standard error of the slope

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

Simple Linear Regression Analysis

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

Simple Linear Regression Analysis

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Software output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

b_1 (0.10977), S_{b_1} (0.03297)

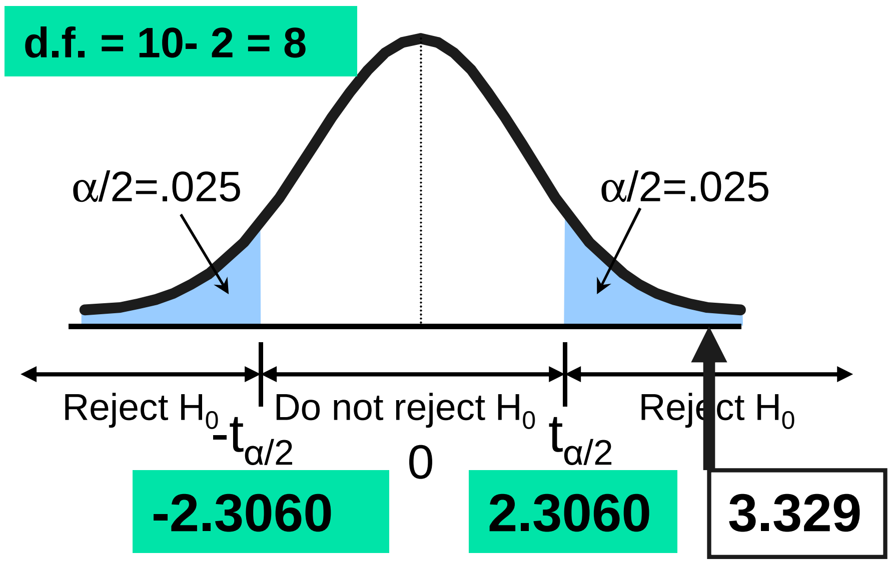
$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Simple Linear Regression Analysis

Test Statistic: $\mathbf{t_{STAT} = 3.329}$

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Summary

- Brief notes on probability and inference
- Simple linear regression analysis
- Multiple linear regression analysis

Multiple Linear Regression Analysis

Idea: Examine the linear relationship between
1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

Y-intercept — Population slopes — Random Error

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Multiple Linear Regression Analysis

Matrix representation:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$$

nx1 $n \times (k+1)$ $(k+1) \times 1$ *nx1*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Multiple Linear Regression Analysis

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:

Estimated (or predicted) value of Y

Estimated intercept

Estimated slope coefficients

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

Multiple Linear Regression Analysis

- The least squares function is given by

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

- The least squares estimates must satisfy

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k$$

Multiple Linear Regression Analysis

- The **least squares normal Equations** are

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i
\end{aligned}
$$

- The solution to the normal Equations are the **least squares estimators** of the regression coefficients.

Multiple Linear Regression Analysis

Matrix Approach

We wish to find the vector of least squares estimators that minimizes:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The resulting least squares estimate is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (12\text{-}13)$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

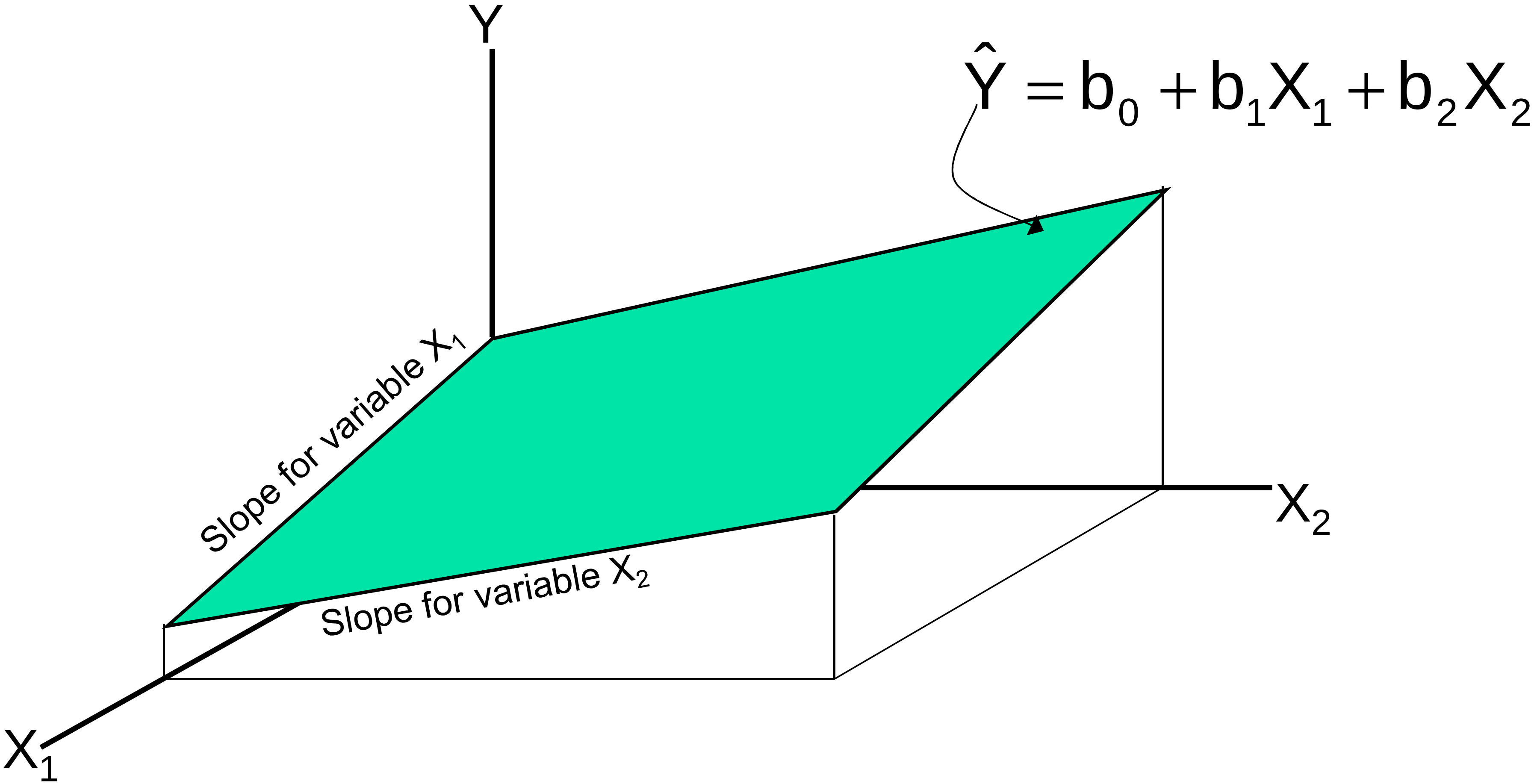
$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} =$$

$$= \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{y}$$

Multiple Linear Regression Analysis

(continued)

Two variable model



Multiple Linear Regression Analysis

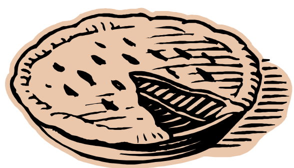
- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in $), Advertising ($100's)
- Data are collected for 15 weeks

Multiple Linear Regression Analysis

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression equation:

$$\widehat{Sales} = b_0 + b_1\,(Price) + b_2\,(Advertising)$$



Multiple Linear Regression Analysis

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

b_1 = -24.975: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising

b_2 = 74.131: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price

$$\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Multiple Linear Regression Analysis

Predict sales for a week in which the selling price is $5.50 and advertising is $350:

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

$$= 306.526 - 24.975(5.50) + 74.131(3.5)$$

$$= 428.62$$

Predicted sales is 428.62 pies

Note that Advertising is in $100's, so $350 means that $X_2 = 3.5$

Multiple Linear Regression Analysis

- Reports the proportion of total variation in Y explained by all X variables taken together

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Multiple Linear Regression Analysis

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Multiple Linear Regression Analysis

(continued)

- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used and sample size

$$r_{adj}^2 = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right]$$

(where n = sample size, k = number of independent variables)

- Penalize excessive use of unimportant independent variables
- Smaller than r^2
- Useful in comparing among models

Multiple Linear Regression Analysis

- F Test for Overall Significance of the Model
- Shows if there is a linear relationship between all of the X variables considered together and Y
- Use F-test statistic
- Hypotheses:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (no linear relationship)

H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

Multiple Linear Regression Analysis

- Test statistic:

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\dfrac{SSR}{k}}{\dfrac{SSE}{n-k-1}}$$

where F_{STAT} has numerator d.f. = k and
denominator d.f. = (n – k - 1)

Multiple Linear Regression Analysis

(continued)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Multiple Linear Regression Analysis

Errors (residuals) from the regression model:

$$e_i = (Y_i - \hat{Y}_i)$$

Assumptions:

- Independence of errors
  - Error values are statistically independent
- Normality of errors
  - Error values are normally distributed for any given set of X values
- Equal Variance (also called Homoscedasticity)
  - The probability distribution of the errors has constant variance

Multiple Linear Regression Analysis

- These residual plots are used in multiple regression:
  - Residuals vs. $\hat{Y}_i$
  - Residuals vs. X_{1i}
  - Residuals vs. X_{2i}
  - Residuals vs. time (if time series data)

Use the residual plots to check for violations of regression assumptions

Multiple Linear Regression Analysis

- Use t tests of individual variable slopes
- Shows if there is a linear relationship between the variable X_j and Y holding constant the effects of other X variables
- Hypotheses:
  - H_0: $\beta_j = 0$ (no linear relationship)
  - H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Multiple Linear Regression Analysis

(continued)

H_0: $\beta_j = 0$ (no linear relationship)

H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Test Statistic:

$$t_{STAT} = \frac{b_j - 0}{S_{b_j}}$$

(df = n – k – 1)

Multiple Linear Regression Analysis

(continued)

t Stat for Price is t_{STAT} = -2.306, with p-value .0398

t Stat for Advertising is t_{STAT} = 2.855, with p-value .0145

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Multiple Linear Regression Analysis

Multicollinearity (also collinearity) occurs when two or more explanatory variables of the multiple regression model are highly correlated

In the presence of multicollinearity the coefficients estimates can change with high variability as a consequence of small changes in the data (low efficiency).

Perfect multicollinearity $\Rightarrow$ ***X*** matrix is singular and cannot be inverted $\Rightarrow$ least square estimates cannot be computed

Multiple Linear Regression Analysis

One way to detect multicollinearity is by computing the variance inflation factors

$$VIF(\beta_j) = \frac{1}{(1 - R_j^2)} \qquad j = 1, 2, \ldots, k \qquad (12\text{-}50)$$

R_j^2: coefficient of determination of the regression of X_j on all the other explanatory variables

A VIF greater than or equal to 5 indicates a multicollinearity problem

In the presence of multicollinearity one or more explanatory variables should be removed by the model

Example: VIF(Price)=VIF(Advertising)= $1/(1-R_1^2)= 1/(1-0.0009264^2) \cong 1$

Price and *Advertising* are almost uncorrelated $\Rightarrow$ absence of collinearity

Multiple Linear Regression Analysis

Regression analysis procedure

- Specification of the multiple regression model
- Test the significance of the multiple regression model
- Test the significance of the regression coefficents
- Discuss adjusted r^2
- Use residual plots to check model assumptions

R exercises

Problem 1 - Passito

- Perform a multiple regression analysis for predicting LIKE_PAS as function of LIKE_AROMA, LIKE_SWEET, LIKE_ALCOHOL and LIKE_TASTE
- Predict the value of LIKE_PAS when LIKE_AROMA=LIKE_ALCOHOL=5

  LIKE_TASTE=LIKE_SWEET=6

R exercises

Problem 2 - Hotel

- Perform a multiple regression analysis for predicting *Price* as function of *Cleanliness* and *Courtesy*
- Predict the value of *Price* when *Cleanliness*=80 and *Courtesy*=40

R exercises

Problem 3 - Mall

- Perform a multiple regression analysis for predicting *Product_assortment* as function of *Temp_Level*, *Brightness*, *Salesman* and *Music_volume*
- Predict the value of *Product_assortment* when *Temp_Level*=-50, *Brightness=20*, *Salesman=30* and *Music_volume=-70*

R exercises

Problem 4 - Students

- Perform a multiple regression analysis for predicting *Econometrics* as function of *Statistics* and *Mathematics*
- Predict the value of *Econometrics* when *Statistics=8* and *Mathematics=7*