University of Ferrara

DIPARTIMENTO
DI ECONOMIA
E MANAGEMENT

Stefano Bonnini & Valentina Mini

Simple Linear Regression Analysis

Lecture 3
December 14, 2018

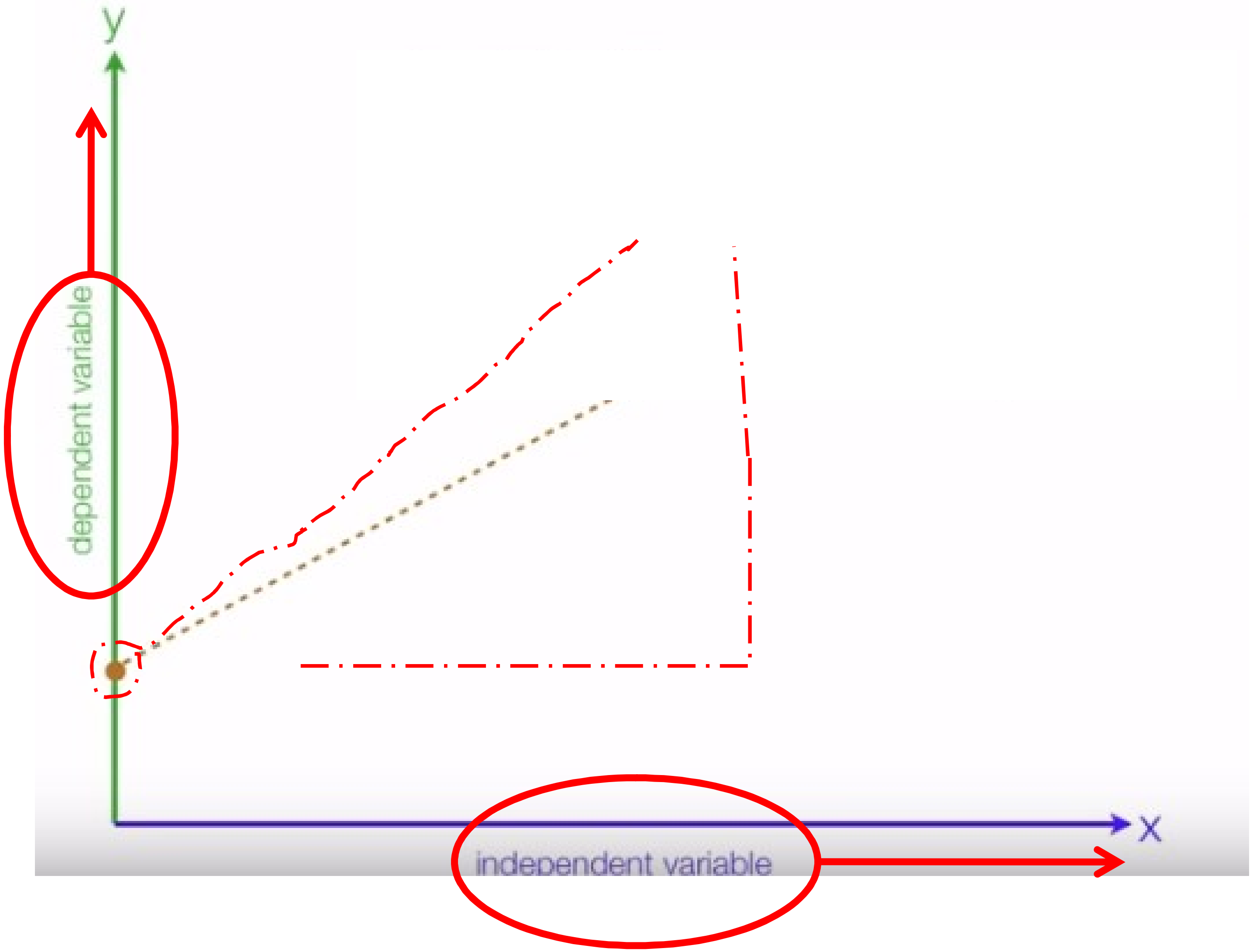
structure of the lecture

1) Linear Regression Model: theoretical approach

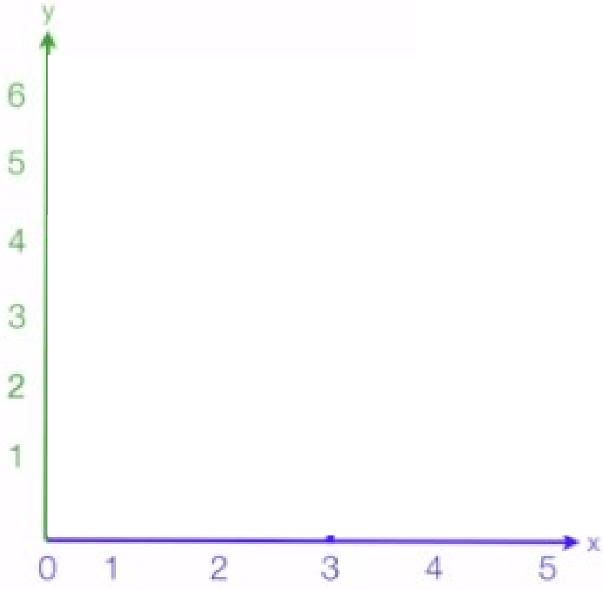
2) Linear Regression Model: a step-by-step simulation analysis

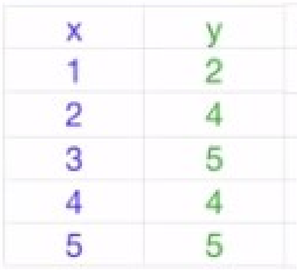
3) LRM in R: practice and exercises

Introduction to Linear Regression Model



Introduction to Linear Regression Model





x	y
1	2
2	4
3	5
4	4
5	5

contents

- Linear regression model: main concepts
- **Regression coefficients:** b0 and b1 – The Least Squares Method
- **Interpretation** of the coefficients
- How well the model is fitting data? The **coefficient of determination r^2**
- The estimates' **standard error**
- **4 basic assumptions** for the linear Regression Model
- **The significance: is the model statistically significant**?
- Inference
- *Exercises using R*

Main targets

- Use of one explanatory variable (x) to estimate a dependent variable (y)

    y = dependent variable,

    x = independent or explanatory variable

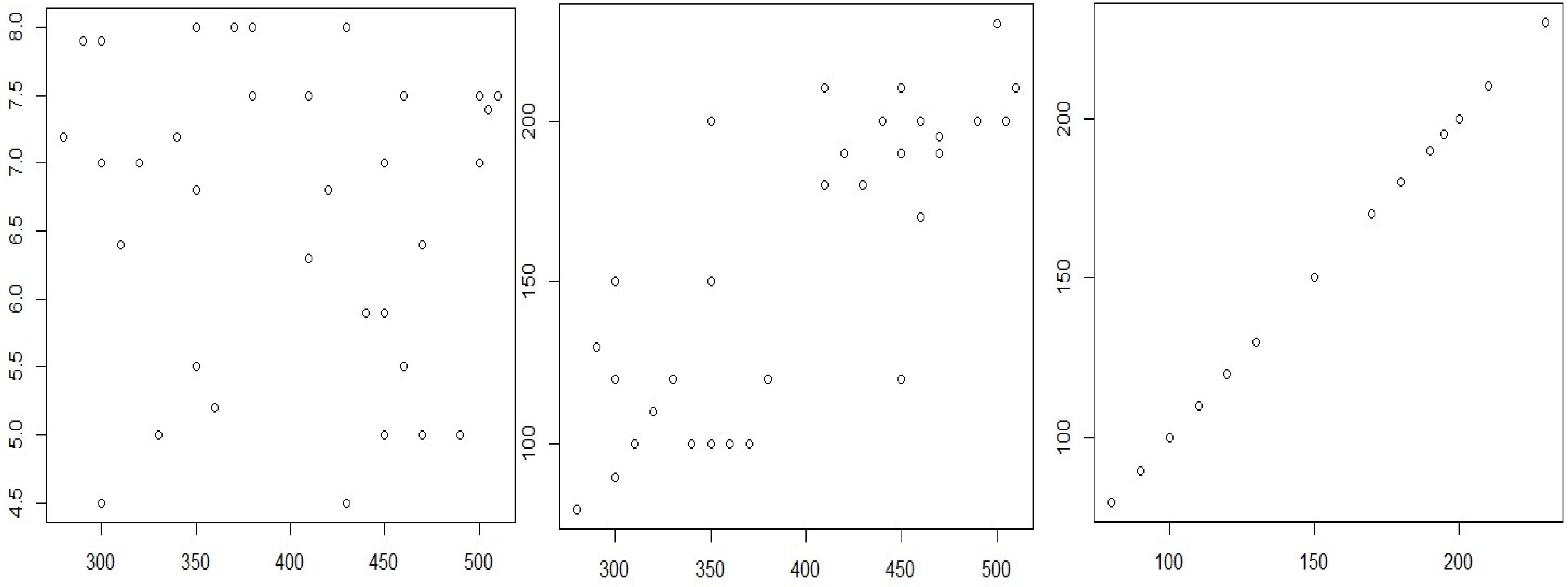
- Estimate and find the meaning of the regression coefficients b_0 e b_1
- Do prevision of y values, based on x (n.b. range)
- Do evaluation of regression's assumptions respect
- Do inference (on coefficients and Y values).

Simple Linear Regression Analysis: checking the relationship between two variables

- 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

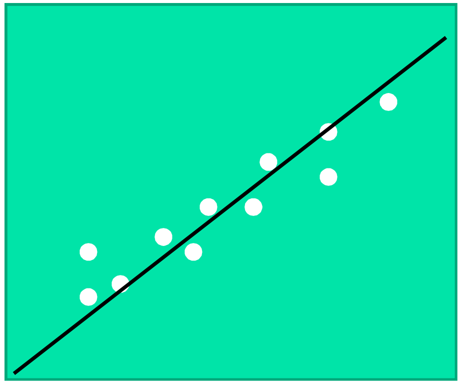
Simple Linear Regression Analysis: checking the relationship between two variables

- Examples of scatter plots

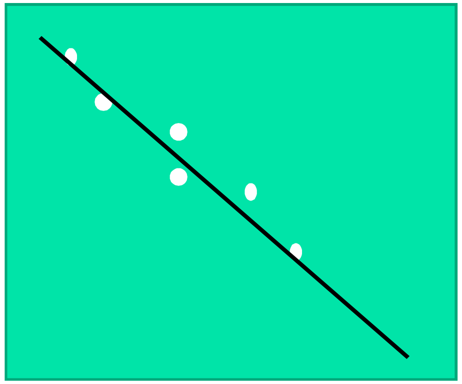


Simple Linear Regression Analysis: checking the relationship between two variables

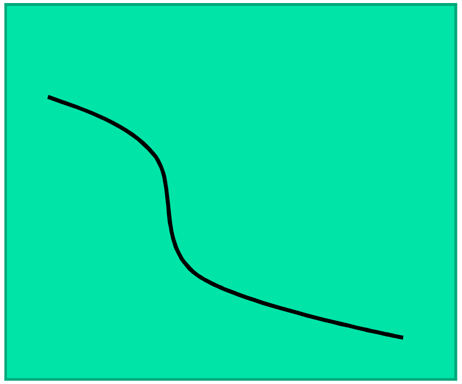
Using the scatter plot we do individuate the possible relationship between two observed variables



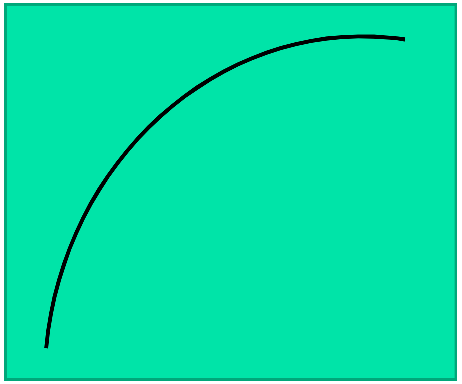
Linear positive relation



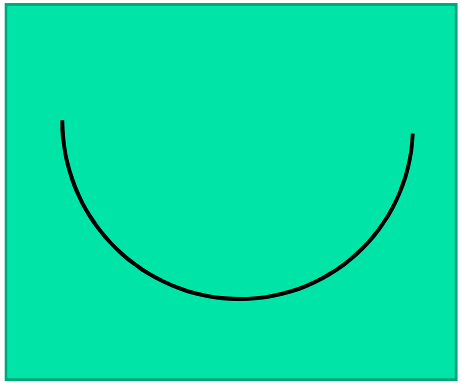
Linear negative relation



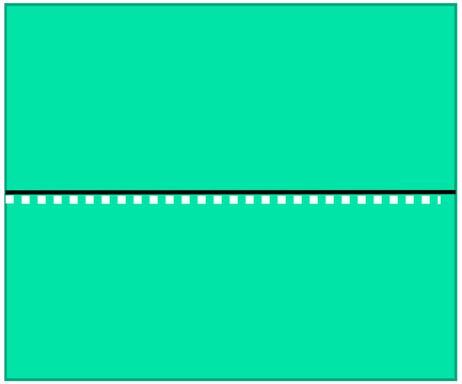
Non-linear relation



Exponential relation



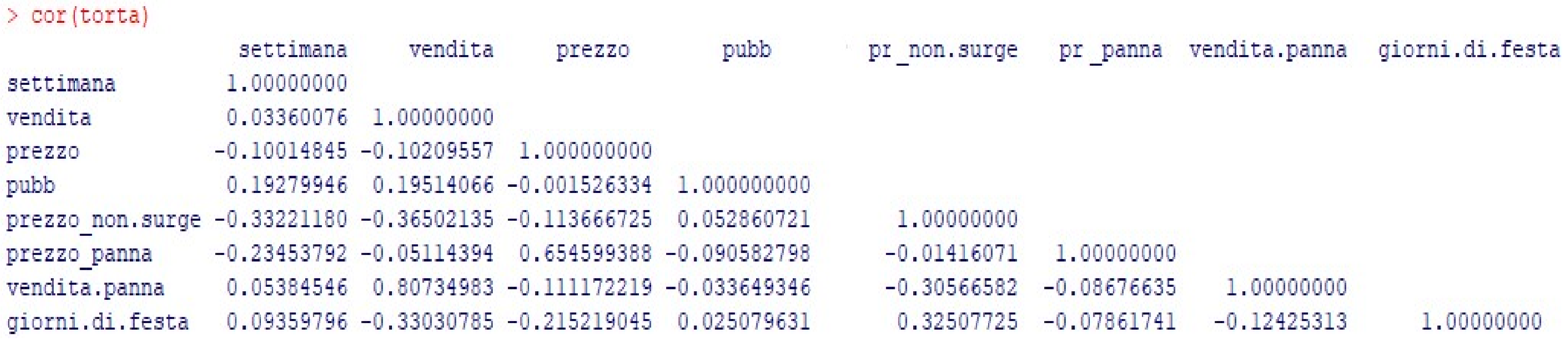
U relation



Absence of relation

Simple Linear Regression Analysis:
checking the relationship between two variables

Example of a correlation matrix



```
> cor(torta)
                   settimana     vendita       prezzo         pubb     pr_non.surge      pr_panna  vendita.panna   giorni.di.festa
settimana         1.00000000
vendita           0.03360076  1.00000000
prezzo           -0.10014845 -0.10209557  1.000000000
pubb              0.19279946  0.19514066 -0.001526334  1.000000000
prezzo_non.surge -0.33221180 -0.36502135 -0.113666725  0.052860721       1.00000000
prezzo_panna     -0.23453792 -0.05114394  0.654599388 -0.090582798      -0.01416071    1.00000000
vendita.panna     0.05384546  0.80734983 -0.111172219 -0.033649346      -0.30566582   -0.08676635     1.00000000
giorni.di.festa   0.09359796 -0.33030785 -0.215219045  0.025079631       0.32507725   -0.07861741    -0.12425313        1.00000000
```

•*Correlation is only concerned with strength of the relationship*

•*No causal effect is implied with correlation*

Simple Linear Regression Analysis: aim

- Regression analysis is used to:
  - Predict the value of a dependent variable *Y* based on the value of one independent variable
  - Explain the impact on the dependent variable of changes in independent (explanatory) variable *X*

Dependent variable: the variable we wish to predict or explain

Independent or explanatory variable: the variable used to predict or explain the dependent variable

Simple Linear Regression Analysis: only one explanatory variable (x)

- Relationship between *Y* and *X* is described by a linear function

- Only **one** independent variable, X ⇒ Simple Linear Regression Model

- ***X*≥2** independent variables, $X_1,\ldots,X_k$ ⇒ Multiple Linear Regression Model

Simple Linear Regression Analysis: the model

Simple Linear Regression Model (SLM)

Population Y intercept

Population Slope Coefficient

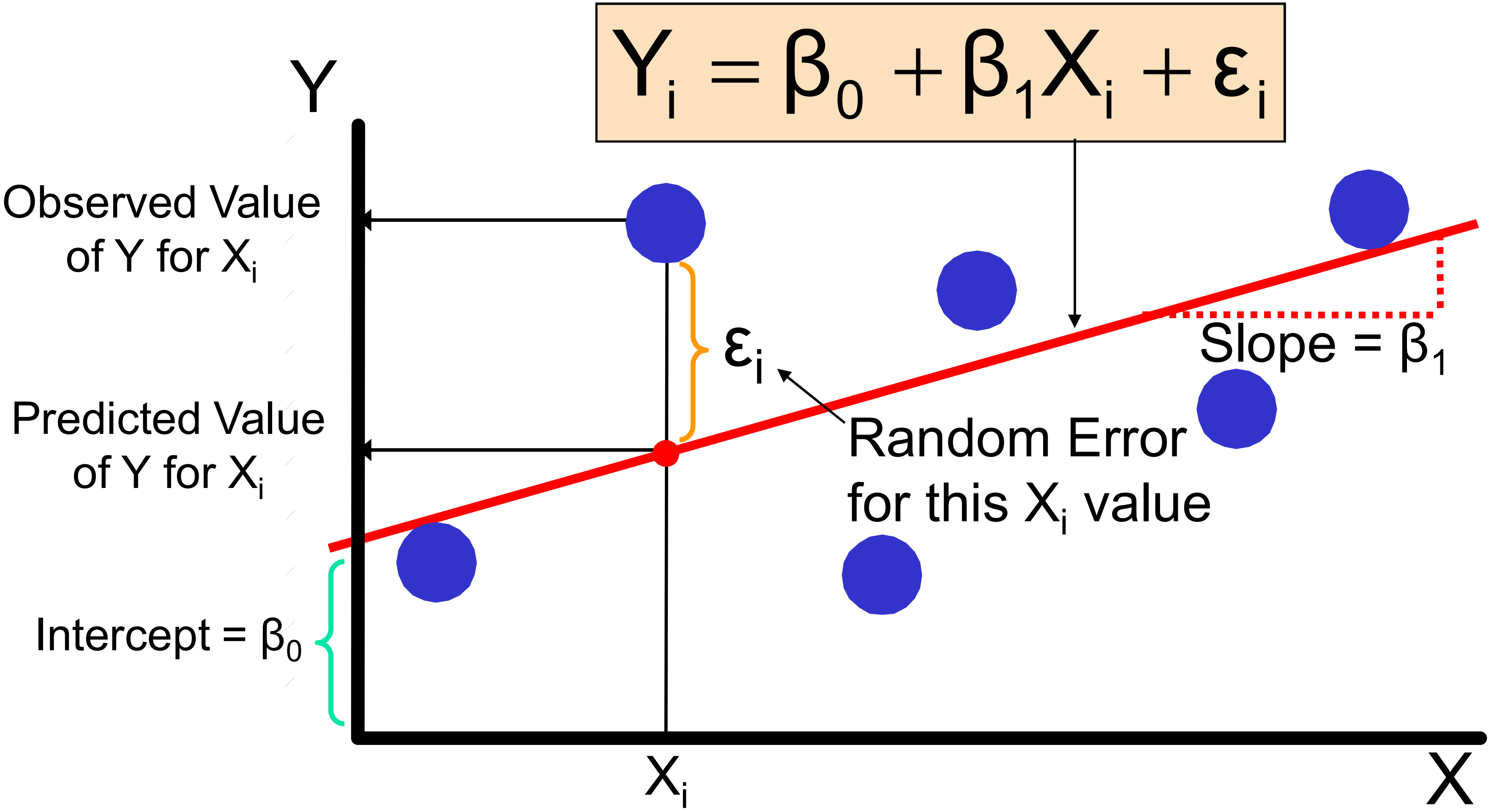
Independent Variable

Random Error term

Dependent Variable

$$\boxed{Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}}$$

Simple Linear Regression Analysis: graphical representation



Simple Linear Regression Analysis: the equation

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

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

Value of X for observation i

Simple Linear Regression Analysis: the least squares method (OLS)

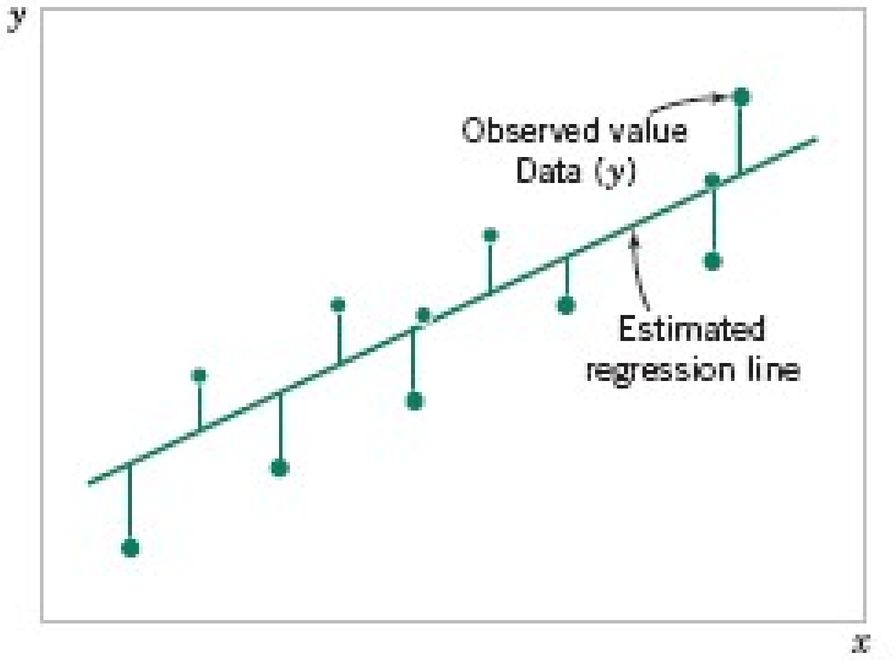
b_0 and b_1 are obtained by finding the values that minimize the sum of the squared differences between Y and $\hat{Y}$:

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

Simple Linear Regression Analysis: the least squares method (OLS)

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

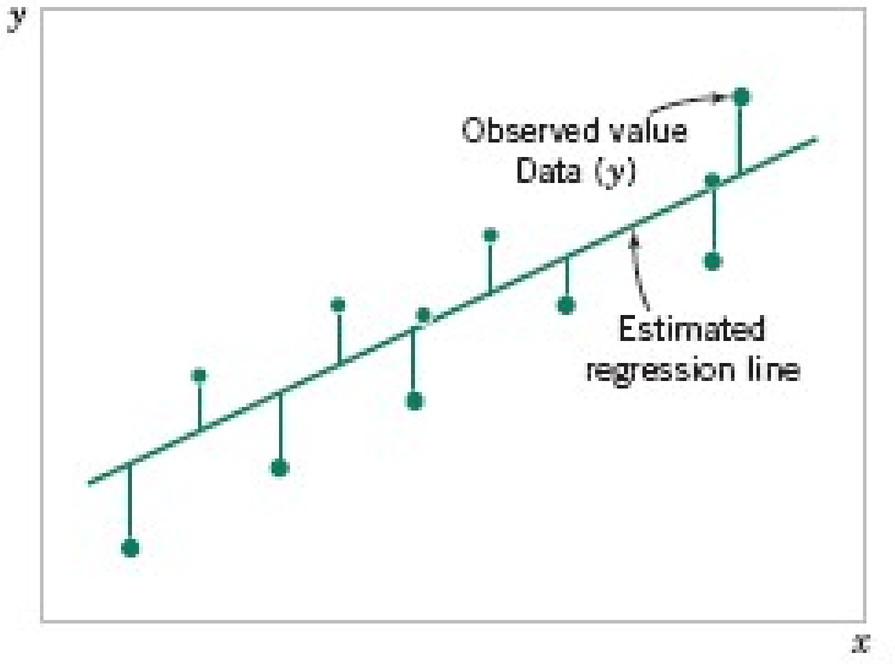
Deviations of the data from the estimated regression model.



Simple Linear Regression Analysis: the least squares method (OLS)

- 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: the least squares method (OLS)

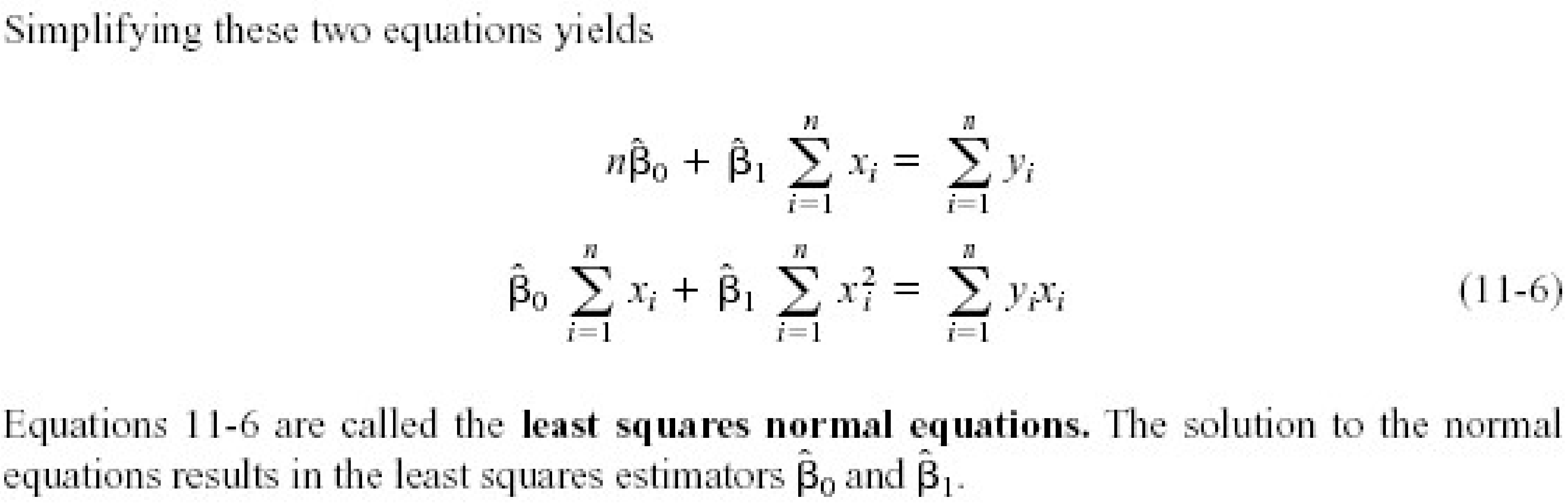
$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

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: the least squares method (OLS)



Simplifying these two equations yields

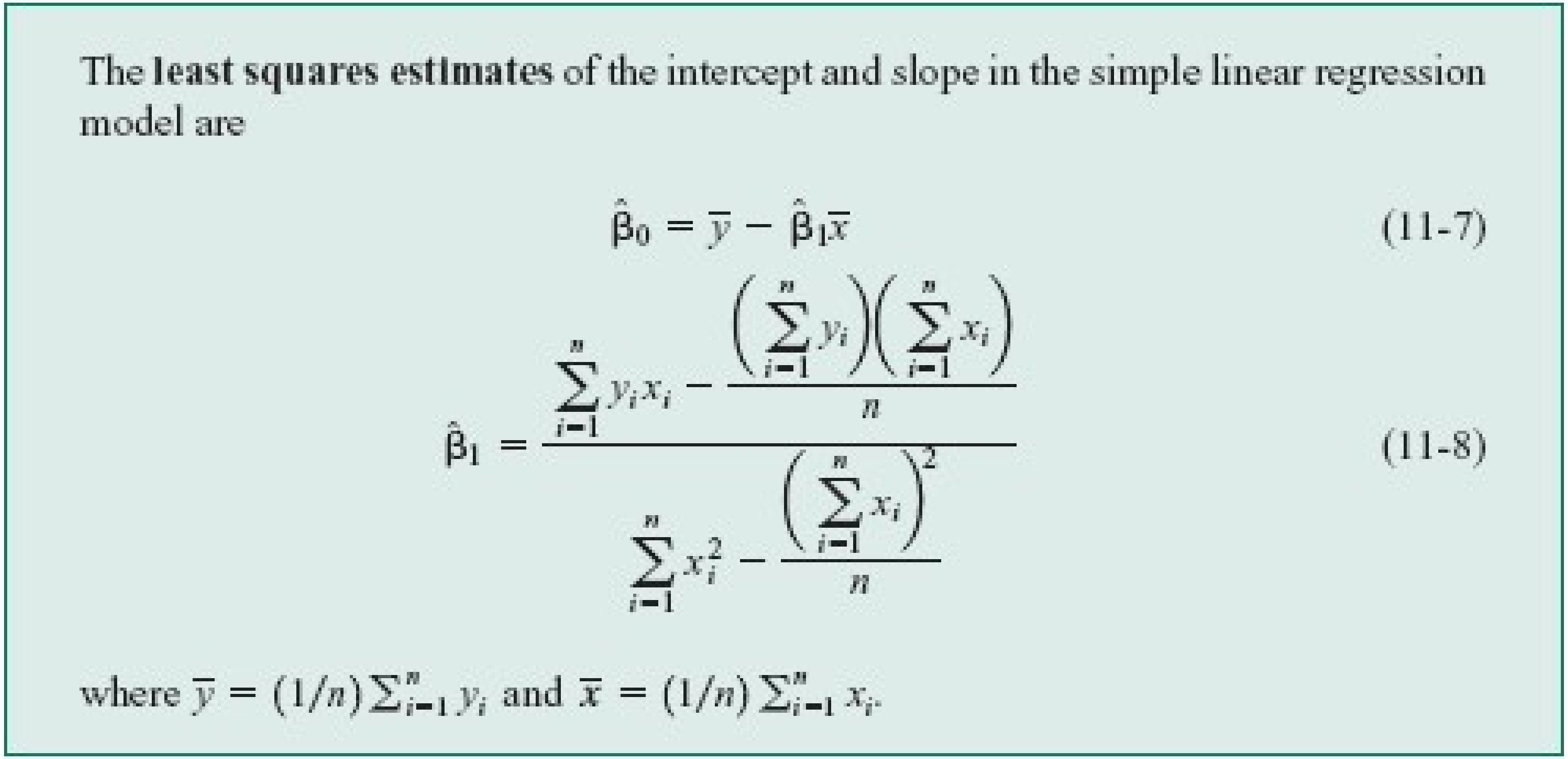
$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

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

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:
least squares estimates

Definition



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

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

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

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

Simple Linear Regression Analysis: least squares estimates

- $b_0 = \hat{\beta}_0$ 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: an example

Ex:

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

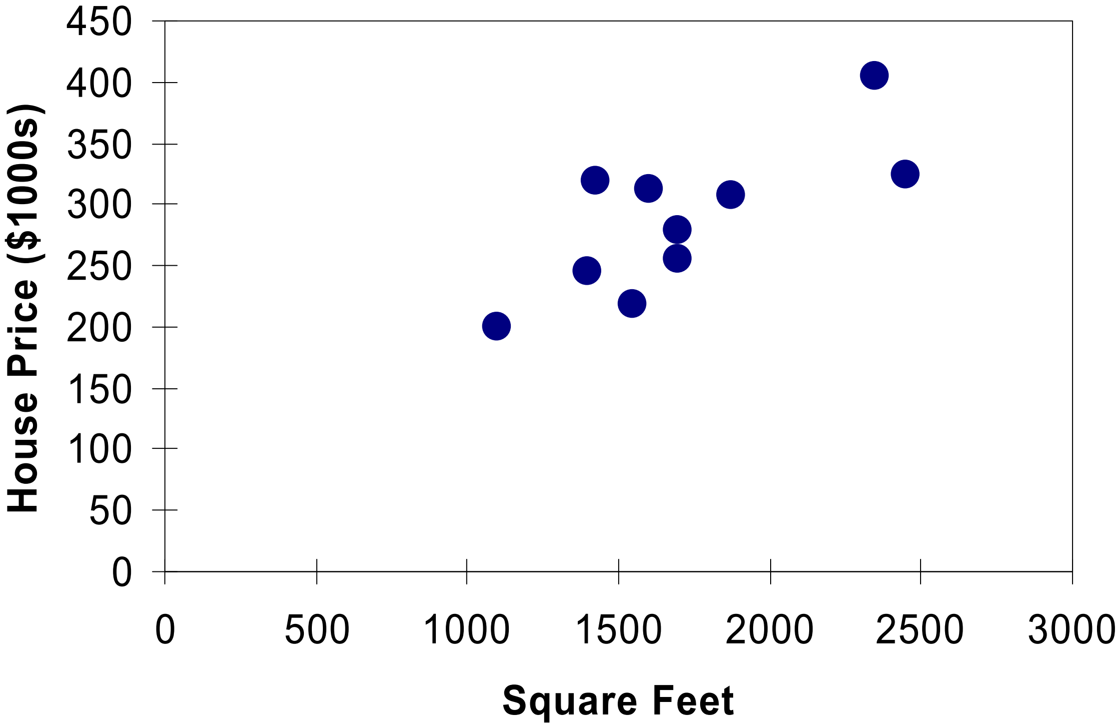
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Simple Linear Regression Analysis: an example

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: an example

House price model: Scatter Plot



Simple Linear Regression Analysis: an example

	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:
an example

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:

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

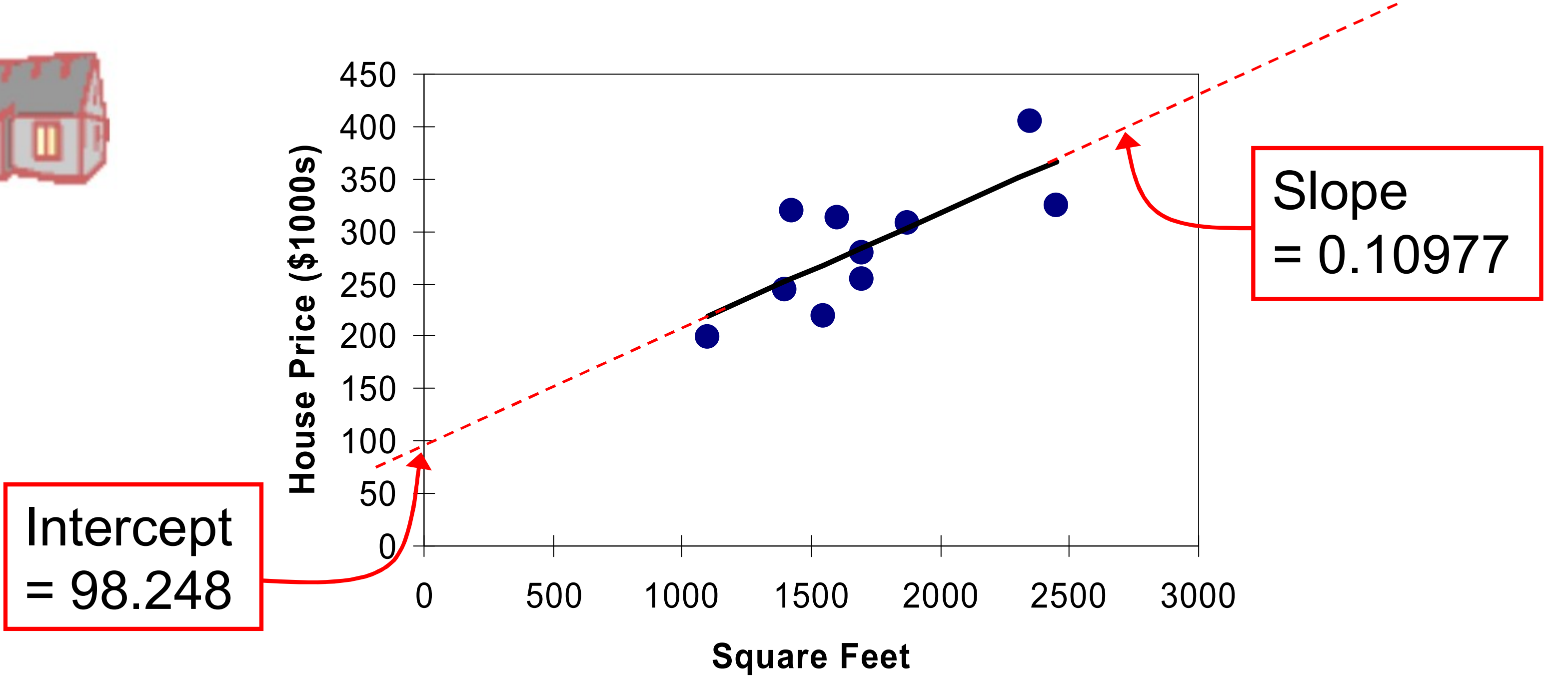
ANOVA

	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**

Simple Linear Regression Analysis: an example

House price model: Scatter Plot and Prediction Line



Intercept
= 98.248

Slope
= 0.10977

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

Simple Linear Regression Analysis: an example

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: the total variation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares

Regression Sum of Squares

Error Sum of Squares

$$SST = \sum (Y_i - \overline{Y})^2$$

$$SSR = \sum (\hat{Y}_i - \overline{Y})^2$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

where:

$\overline{Y}$ = 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

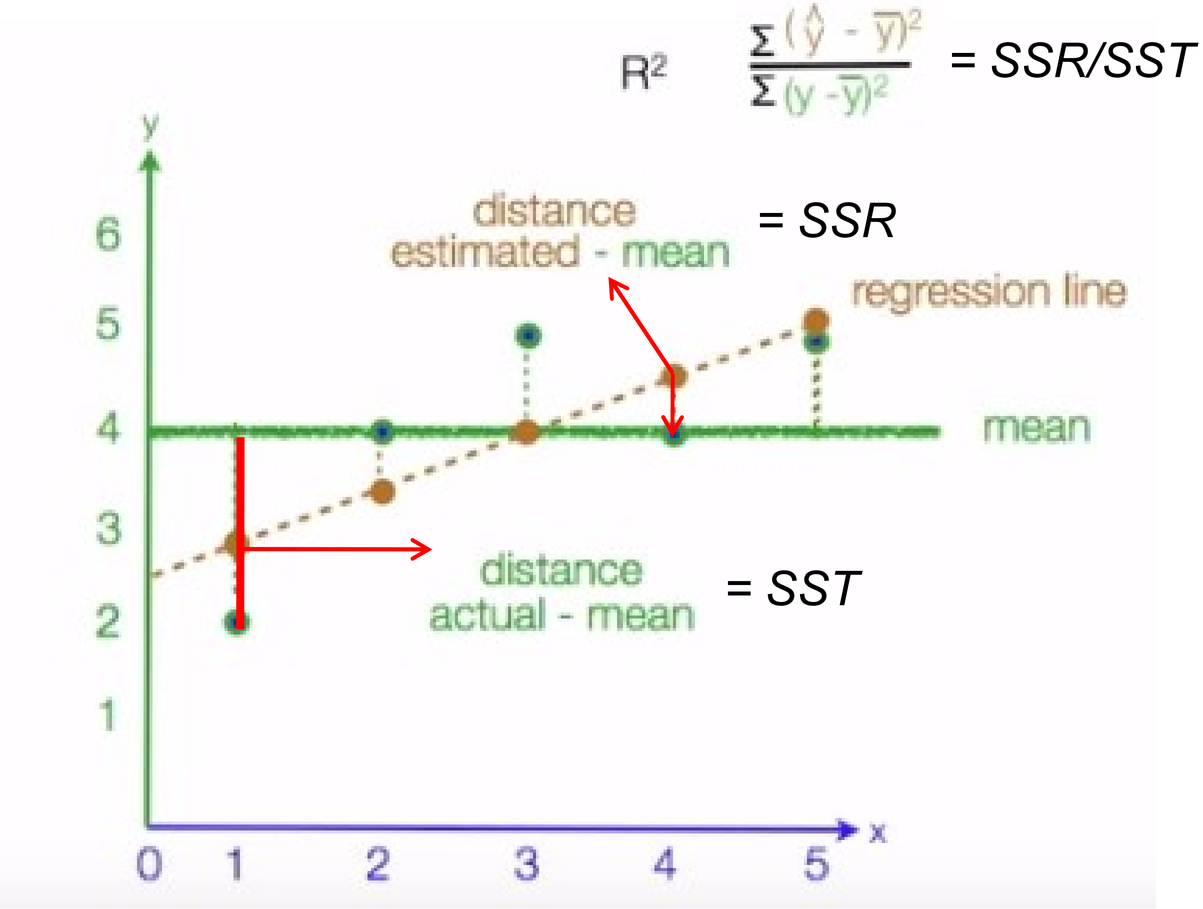
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as r^2

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

note: $\boxed{0 \leq r^2 \leq 1}$

Simple Linear Regression Analysis: the coefficient of determination

How well the regressed values estimated the real/actual values



$$R^2 \quad \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = SSR/SST$$

= *SSR*

= *SST*

Simple Linear Regression Analysis: the coefficient of determination

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:
central assumptions

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

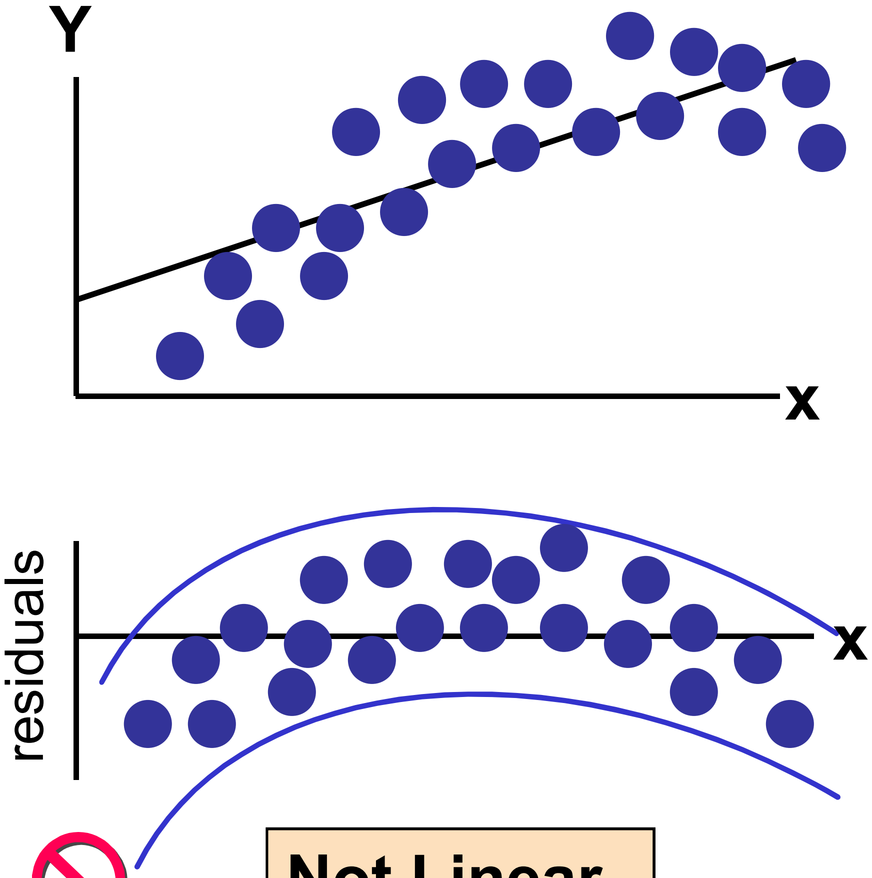
Simple Linear Regression Analysis: central assumptions

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

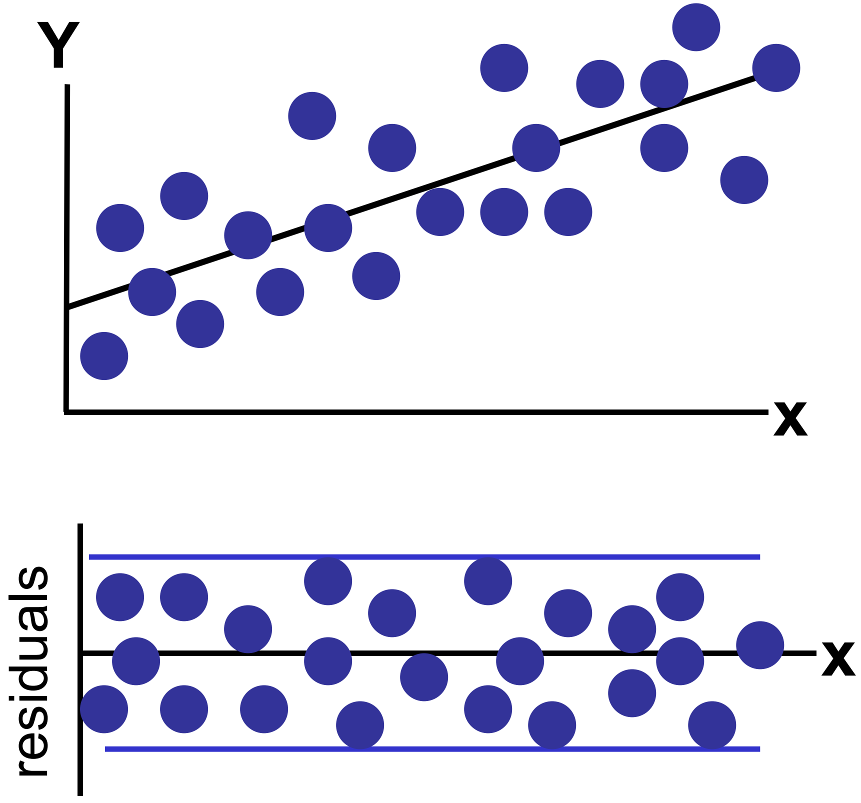
- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant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

Simple Linear Regression Analysis: central assumptions

Analysis of residuals



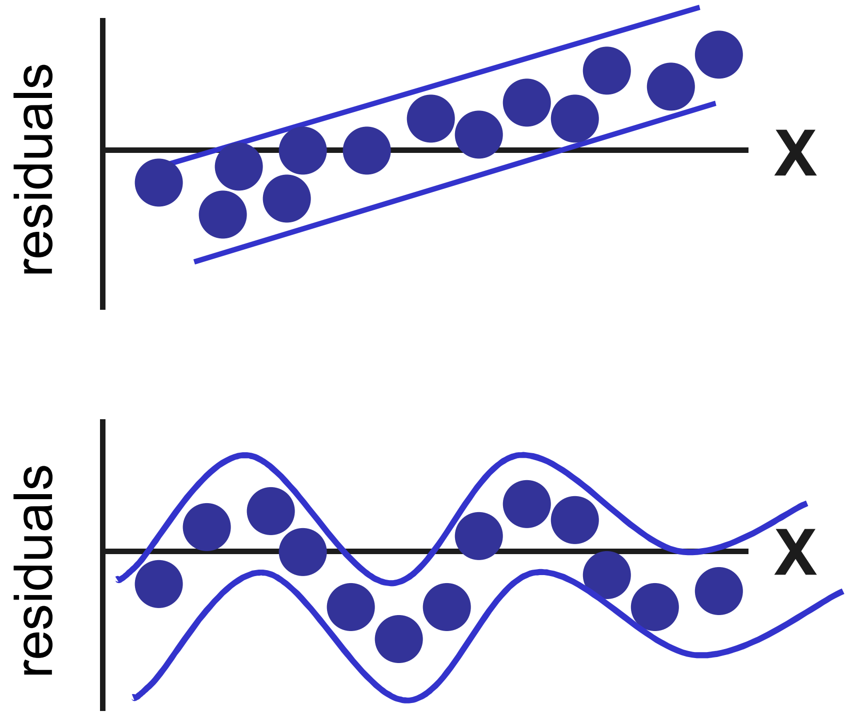
🚫 **Not Linear**



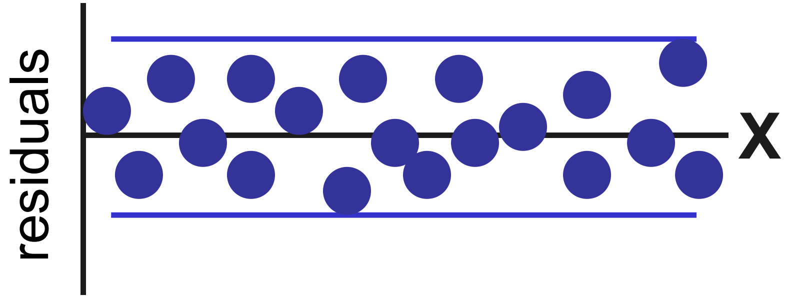
✓ **Linear**

Simple Linear Regression Analysis: central assumptions

🚫 **Not Independent**



✓ **Independent**



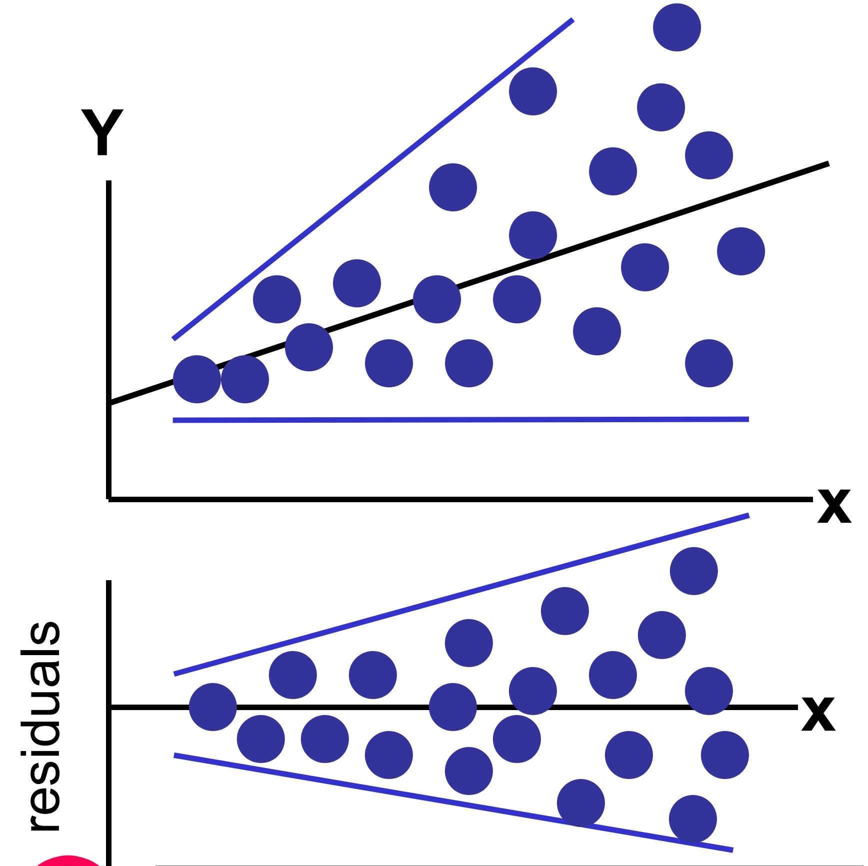
Simple Linear Regression Analysis: central assumptions

Checking for normality:

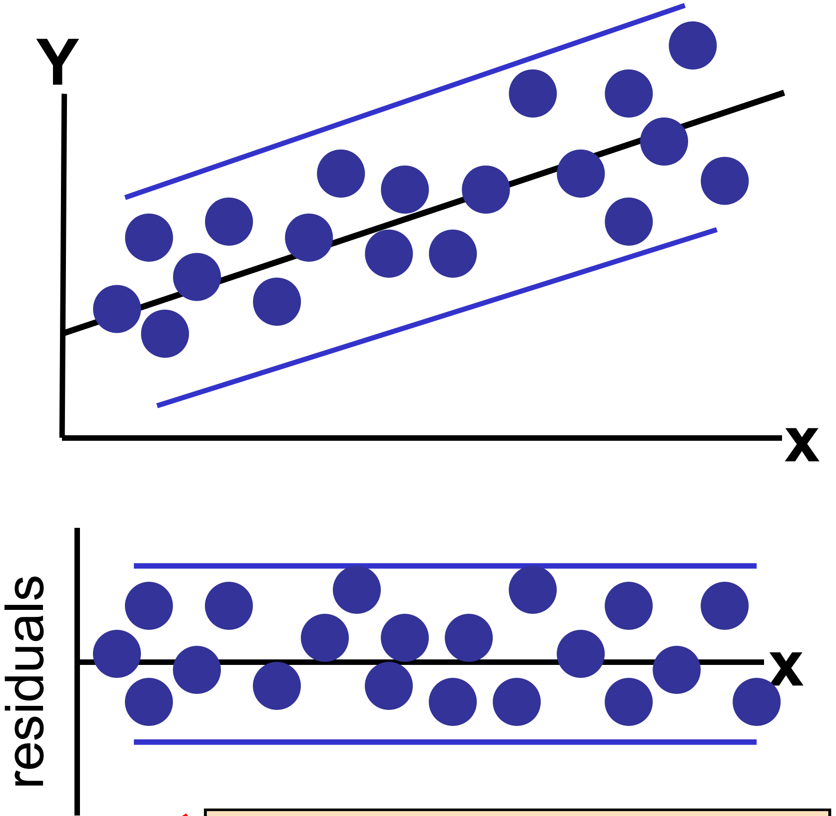
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

Simple Linear Regression Analysis: central assumptions

Checking for homoschedasticity



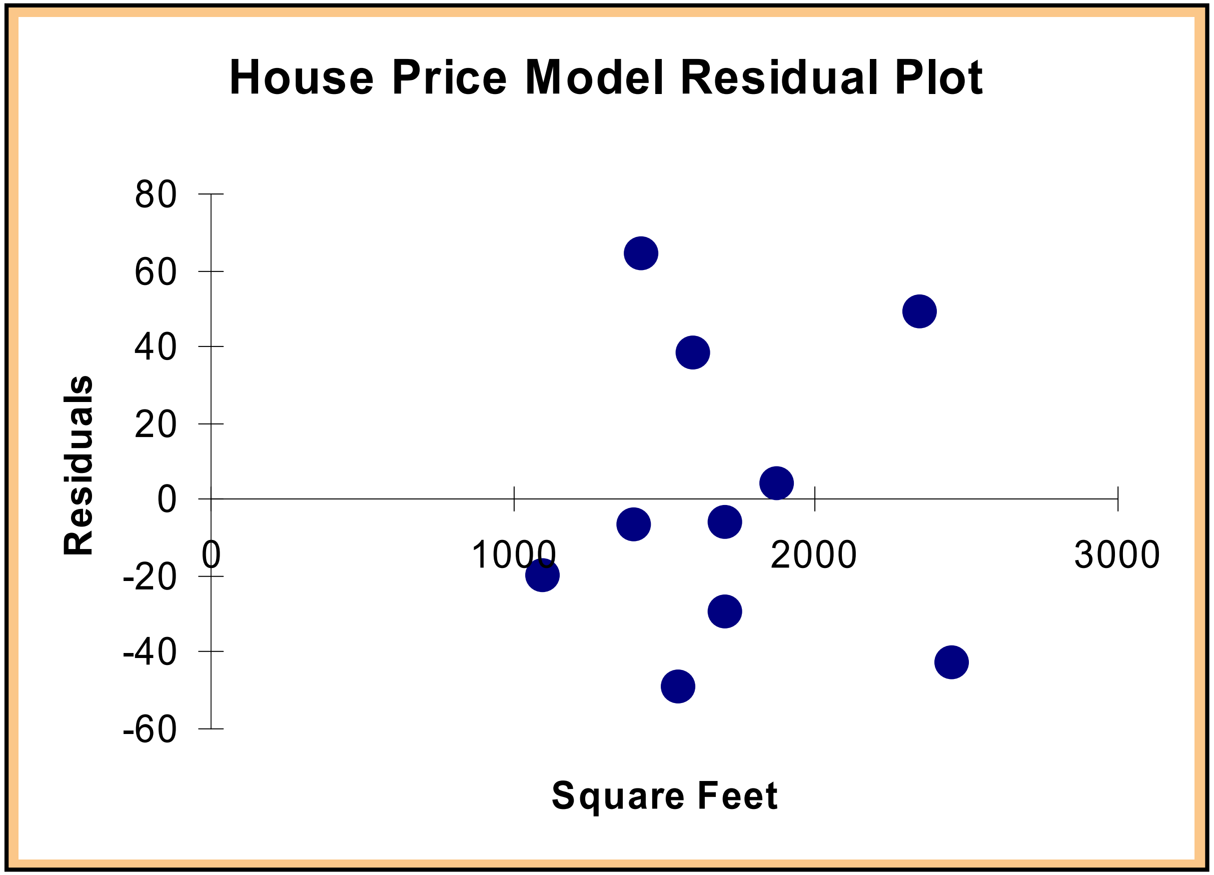
🚫 Non-constant variance



✓ Constant variance

Simple Linear Regression Analysis: central assumptions

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:
standard error of the regression slope coefficient

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

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

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: inference

- 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: inference

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

S_{b_1}

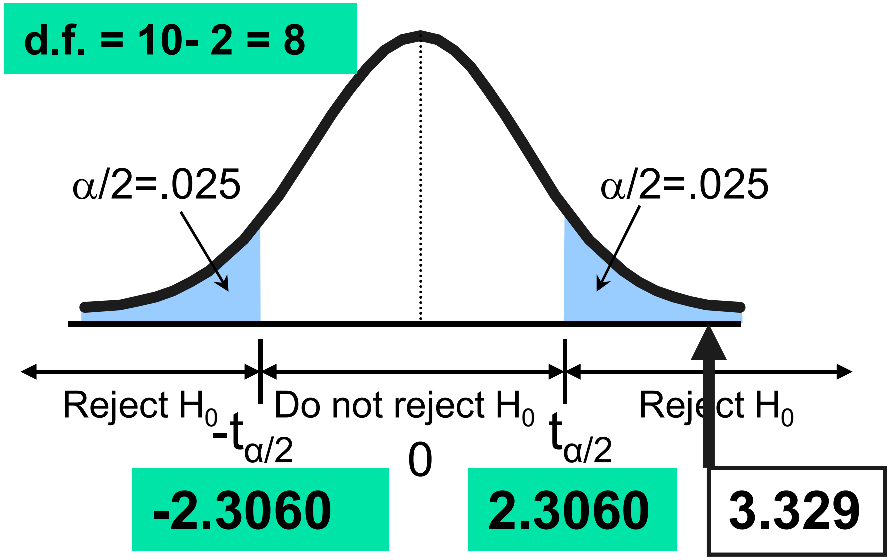
$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Simple Linear Regression Analysis: inference

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

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



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Simple Linear Regression:
a step-by-step Analysis

Example based on a case study

The Mini Market Company is a chain of small convenience retail shops that stocks a range of everyday items such as groceries, snack foods, confectionery, soft drinks ect.

The director is considering the possibility to open a new shop in Ferrara City Center; however before to construct the business plan, he wants understand the causal relationship of the shop size on sails volume.

For this reason the Director is asking you a technical advise.

Data sample:

14 shops,

Shop’s size (100 m^2) and

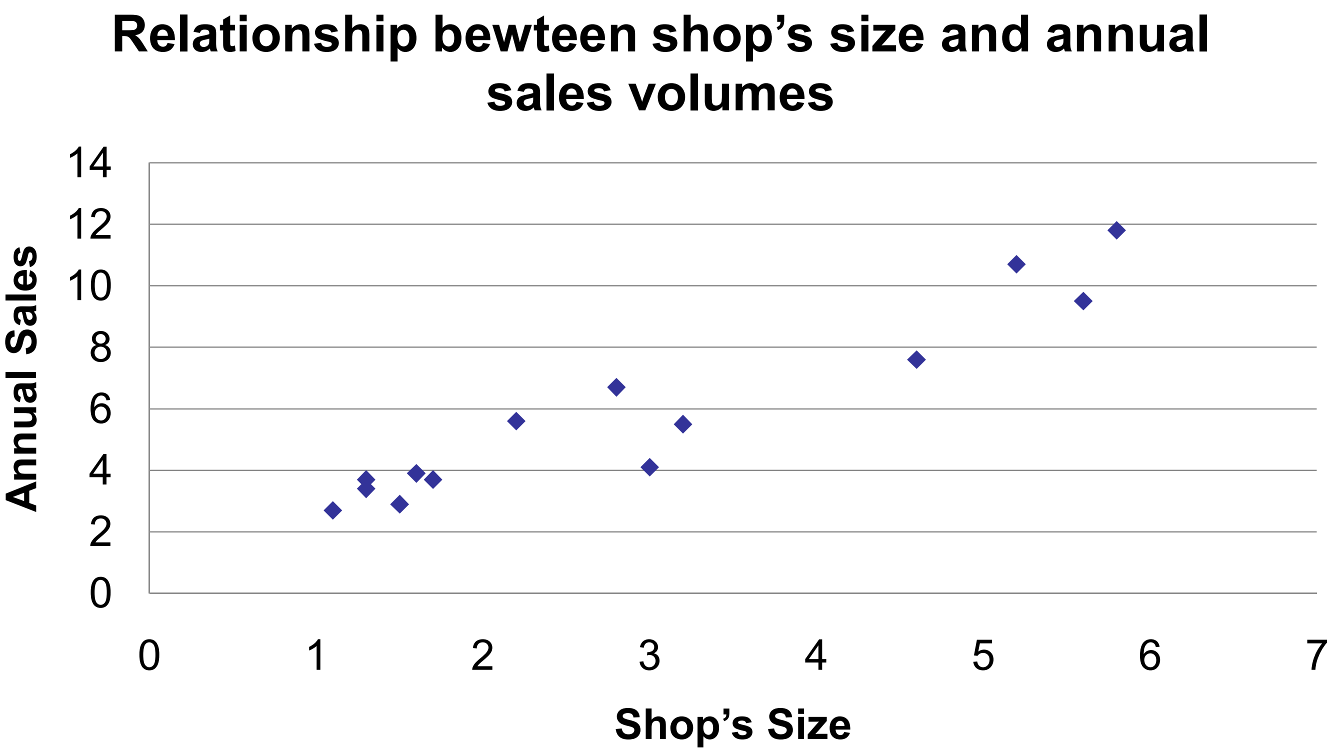
Annual sales volume (1’000 €)

The database of sampled data

Shop's ID	Shop's size (100 M²)	Annual Sales Volume (1000 €)
1	1,7	3,7
2	1,6	3,9
3	2,8	6,7
4	5,6	9,5
5	1,3	3,4
6	2,2	5,6
7	1,3	3,7
8	1,1	2,7
9	3,2	5,5
10	1,5	2,9
11	5,2	10,7
12	4,6	7,6
13	5,8	11,8
14	3	4,1

Central question:
in the explorative phase, what we can say about the relationship between this two variables?

Step 1:
Graphical representation of the correlation between 2 variables



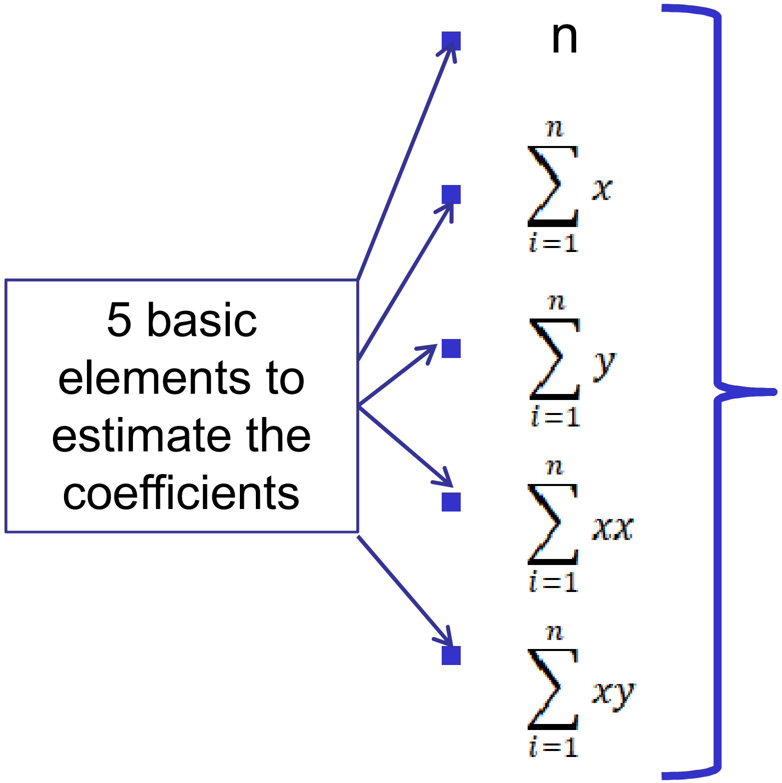
Relationship bewteen shop’s size and annual sales volumes

Dependent Variable (y)

Exlanatory variable

The scatter-plot must form a linear pattern.

Step 2:
Estimating regression coefficients – b_1 and b_0



5 basic elements to estimate the coefficients

- n
- $\sum_{i=1}^{n} x$
- $\sum_{i=1}^{n} y$
- $\sum_{i=1}^{n} xx$
- $\sum_{i=1}^{n} xy$

$$b_1 = ssxy/ssx$$

SSXY=

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

SSX=

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

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

$$\bar{Y} = \frac{\sum_{1=i}^{n} y_i}{n}$$

$$\bar{x} = \frac{\sum_{1=i}^{n} x_i}{n}$$

$$\hat{y}_i - b_0 + b_1 x_1$$

Step 2:
Estimating regression coefficients – b_1 and b_0

ID negozio	M² (100) X	Sales Volume (1’000) y	X²	X*Y
1	1,7	3,7	2,89	6,29
2	1,6	3,9	2,56	6,24
3	2,8	6,7	7,84	18,76
4	5,6	9,5	31,36	53,2
5	1,3	3,4	1,69	4,42
6	2,2	5,6	4,84	12,32
7	1,3	3,7	1,69	4,81
8	1,1	2,7	1,21	2,97
9	3,2	5,5	10,24	17,6
10	1,5	2,9	2,25	4,35
11	5,2	10,7	27,04	55,64
12	4,6	7,6	21,16	34,96
13	5,8	11,8	33,64	68,44
14	3	4,1	9	12,3
14	**40,9**	**81,8**	**157,41**	**302,3**
n	$\sum_{i=1}^{n} x$	$\sum_{i=1}^{n} y$	$\sum_{i=1}^{n} xx$	$\sum_{i=1}^{n} xy$

Step 2:
Estimating regression coefficients – b_1 and b_0

b_1= SSXY/SSX=

SSXY = 302.3-(40.9*81.8)/14 = 63.3271

SSX = 157- (40.9*40.9)/14 = 37.92358

= 63.3271/37.9235= 1.6699

b_0= (81.8/14) – 1.6699*(40.9/14)

= 5.843-4.8785 =

= 0.9645

Estimated Model

Y = 0.9645+1.6699 Xi

ID	M² (100) X	€ annaul sales (1'000) y	Estimated Model
1	1,7	3,7	3.8
2	1,6	3,9	3.64
3	2,8	6,7	5.64
4	5,6	9,5	10.31
5	1,3	3,4	3.13
6	2,2	5,6	4.64
7	1,3	3,7	3.13
8	1,1	2,7	…
9	3,2	5,5	…
10	1,5	2,9	…
11	5,2	10,7	…
12	4,6	7,6	…
13	5,8	11,8	…
14	3	4,1	5.97
14	**40,9**	**81,8**	

Step 4:
Interpreting the estimated regression coefficients

b_1 - This is the SLOPE of the regression line.
Thus this is the amount that the Y variable (dependent) will change for each 1 unit change in the X variable.
So for each increase of 100 m² in the Shop's Size (X), we estimate that the annual sales (Y) will increase by 1'996,6 Euros.

----

b_0 - This is the intercept of the regression line with the y-axis.
In other words it is the value of Y if the value of X = 0.
Theoretically, in pour case when the shop's size =0, the annual sales will be 964,5€

Question: Does this interpretation make sense?

Attention to the X-values range!

If the X value is outside the range,
we are not able to give a practical interpretation of b_0

Step 5:
Making predictions using our estimated model

Before making predictions, check the data!
Be sure that
the range of sampled X (Xmin, Xmax) includes
the value you are using for your prediction

Considering our Mini Market case:

-How much will be the Annual Shops Sales if the Shop's Size is 200 squared meters?

→*Sales (1'000)= 0.9645+1.6699*200*

-How much will be the Annual Shops Sales if the Shop's Size is 100 squared meters?

→ *We cannot do the prediction because the value X=100 is outside the range of sampled X (so the relationship between the two variables could be different)*

Step 6:
Assessing the Model's goodness of fit

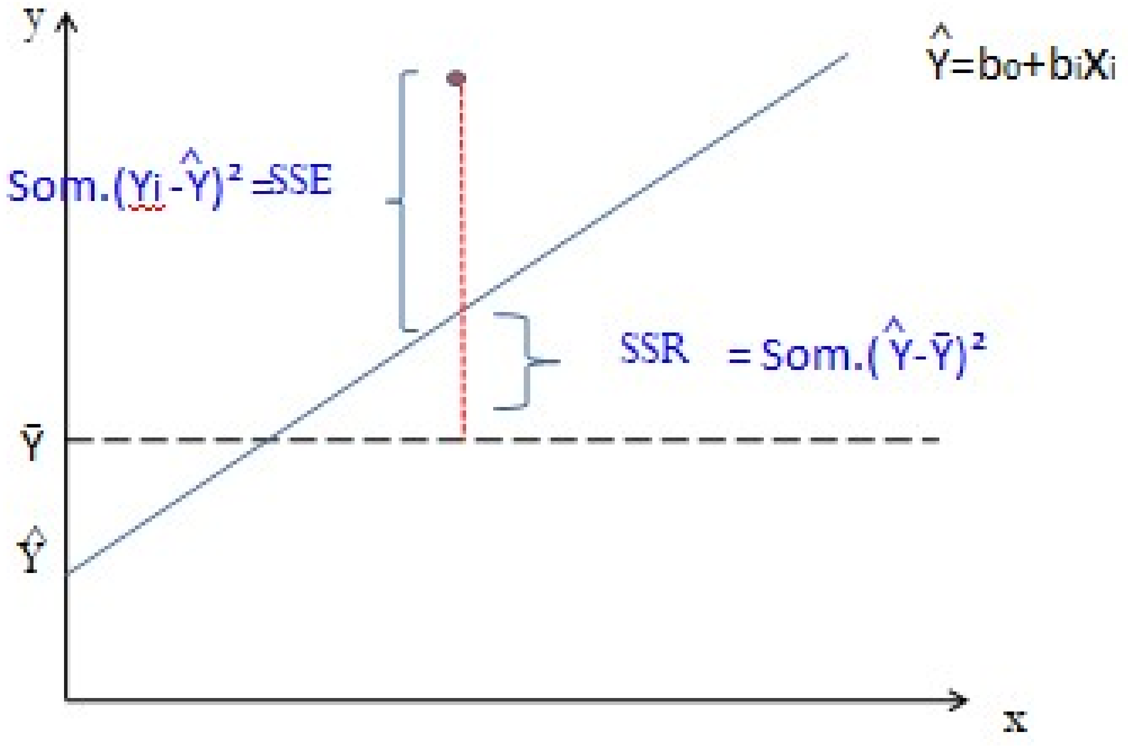
R^2 = coefficient of determination.
It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

$$R^2 = \frac{\text{Regression Variability}}{\text{Total Variability}}$$

This implies that *R2%* of the variability of the dependent variable has been accounted for, and the remaining *(1-R2)%* of the variability is still unaccounted for.

Step 6:
Assessing the Model's goodness of fit

Graphical representation



SST = SSE+SSR

Step 6:
Assessing the Model’s goodness of fit

$R^2 = SSR/SST$

$SSR = SUM\ (\hat{Y}-\bar{Y})^2$

$SST << SSR + SSE = SUM(y_i-\bar{Y})^2$

$\hat{Y}$	$\hat{Y}-\bar{Y}$	$(\hat{Y}-\bar{Y})^2$ SSR	$y_i-\bar{Y}$	$(y_i-\bar{Y})^2$ SST
3,8	3,8-5,84= -2,04	4,16	-2,14	4,58
3,64	-2,2	4,84	-1,94	3,76
5,64	-0,2	0,04	0,86	0,79
10,31	4,47	19,98	3,66	13,39
3,13	-2,71	7,34	-2,44	5,95
4,64	-1,2	1,44	-0,24	0,06
3,13	-2,71	7,34	-2,14	4,58
2,8	-3,04	9,24	-3,14	9,86
6,3	0,46	0,21	-0,34	0,11
3,47	-2,37	5,62	-2,94	8,64
9,65	3,81	14,52	4,86	23,62
8,65	2,81	7,89	1,76	3,1
10,65	4,81	23,13	5,96	35,52
5,97	0,13	0,01	-1,74	3,03
		105,76		116,99

$R^2 = 105.72/116.99 = 0.903669 =$ *0.904 → 90.4% of var accounted*

Step 7:
Standard Error of the Estimates

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

In our example:

SSE= sum(Yi-Y)²

n=14 → (n-2=12)

→ Syx= 0.966

INTERPRETATION

Standard error=0.966, Thus equals to 966 Euros.

→ The mean deviation of the estimated sales value and the real one is equals to 966 Euros.

Step 8:
Graphical analysis of the assumptions

Using graphical representations, we need to check the 4 main assumptions of the Linear Regression Model

Linear Regression Mode using R