Simple Linear Regression

Dr. Zahoor Tanoli
COMSATS Attock

Correlation vs. Regression

- 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

Introduction to Regression Analysis

- Statistical process for estimating the relationships among variables
- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to predict or explain

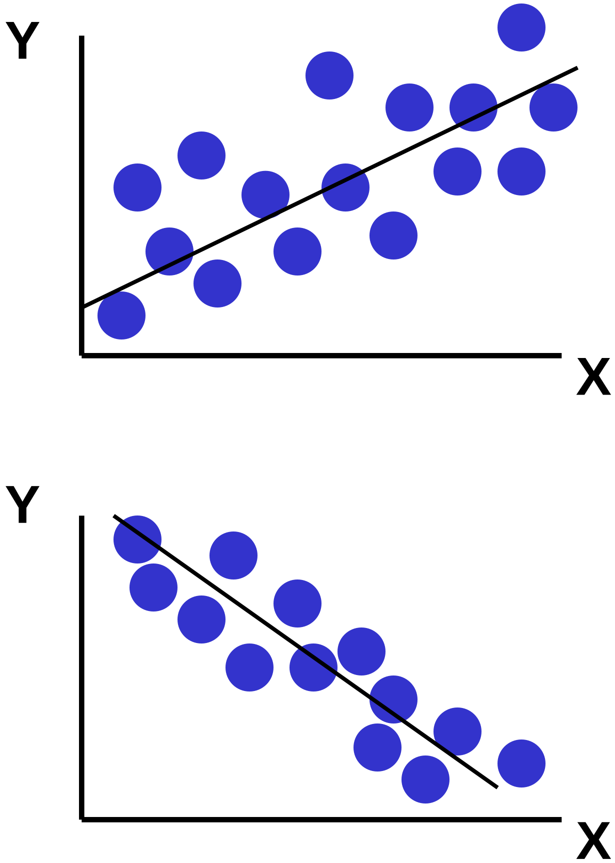
Independent variable: the variable used to predict or explain the dependent variable

Simple Linear Regression Model

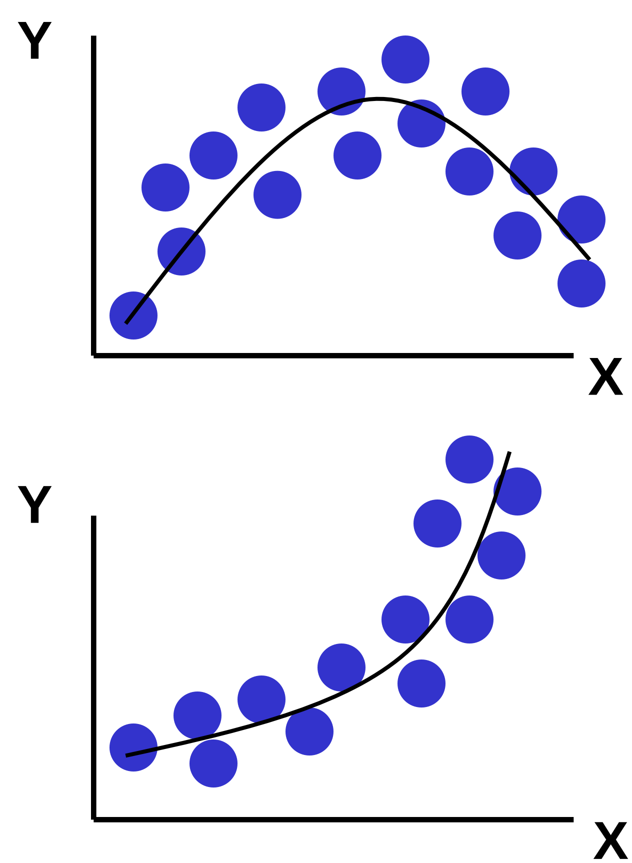
- Only **one** independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be related to changes in X

Types of Relationships

Linear relationships



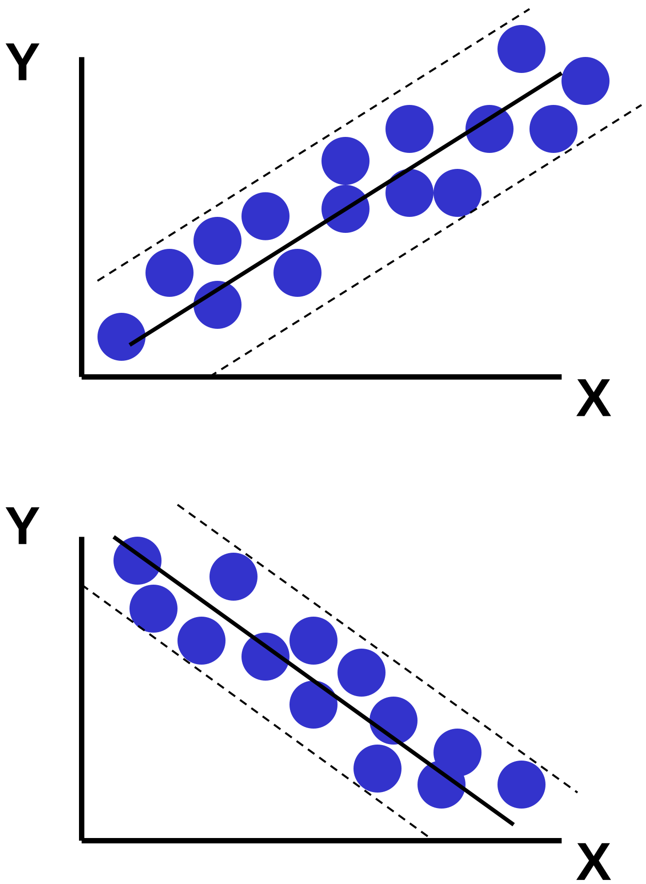
Curvilinear relationships



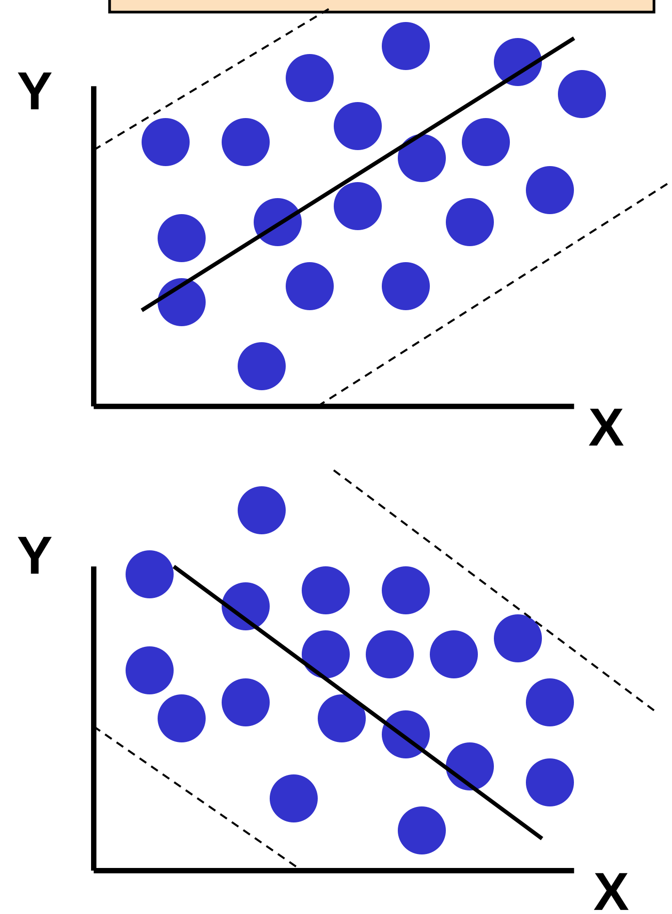
Types of Relationships

(continued)

Strong relationships



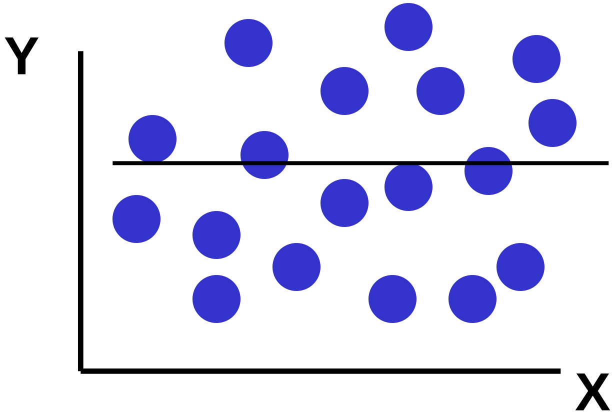
Weak relationships

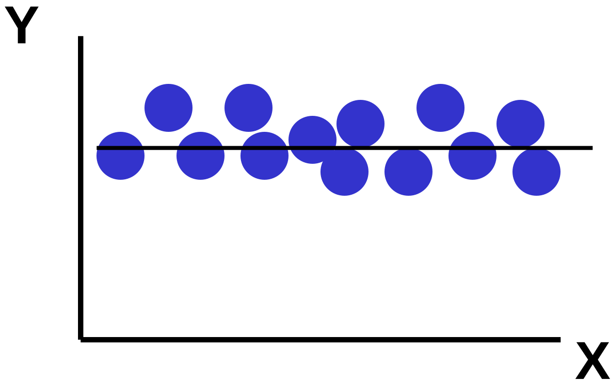


Types of Relationships

(continued)

No relationship





Simple Linear Regression Model

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

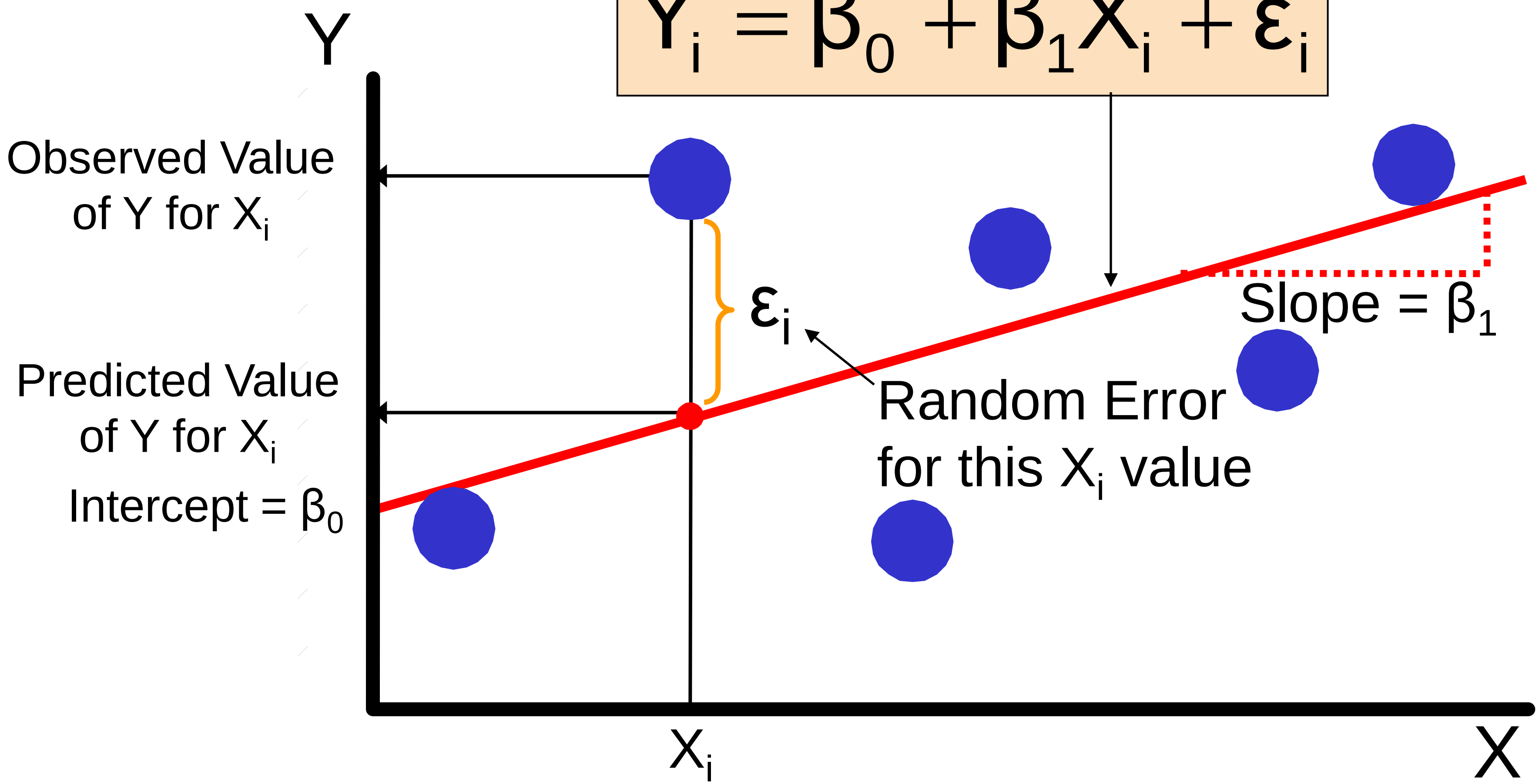
Dependent Variable

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

Simple Linear Regression Model

(continued)

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



Simple Linear Regression Equation (Prediction Line)

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

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

The Least Squares Method

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

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

Interpretation of the Slope and the Intercept

- b_0 is the estimated mean value of Y when the value of X is zero

- b_1 is the estimated change in the mean value of Y as a result of a one-unit increase in X

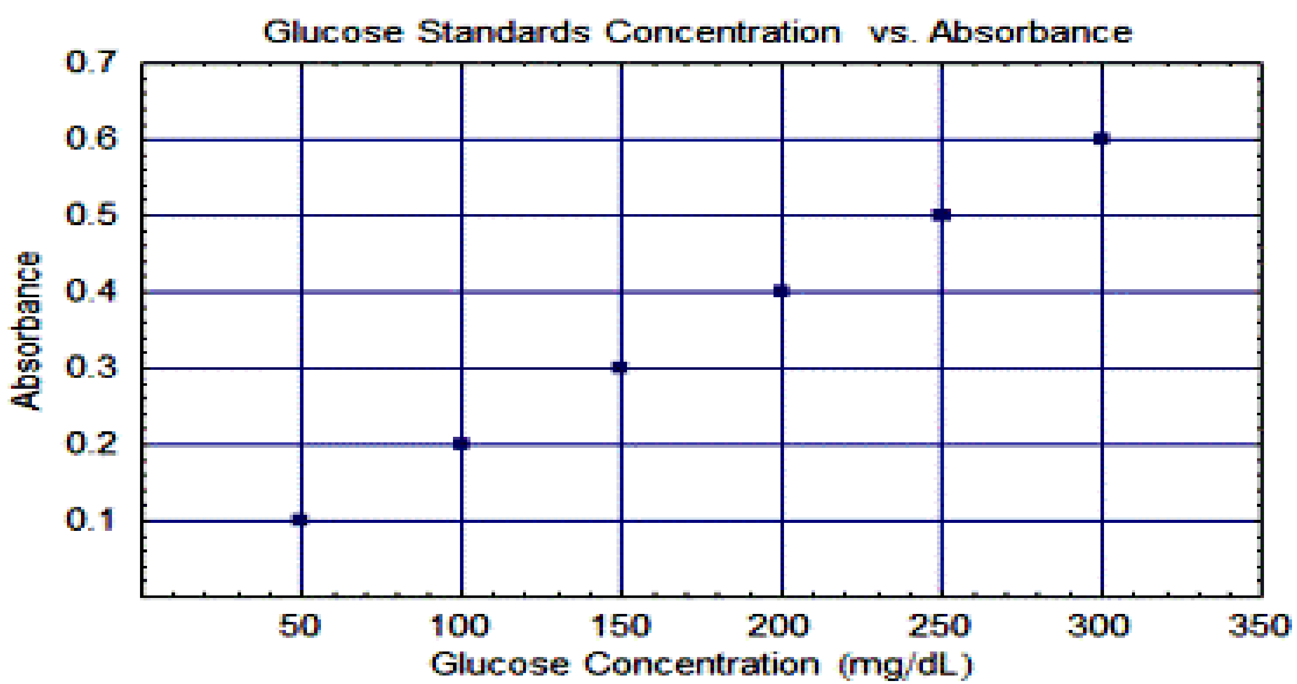
Calculating the Y-Intercept

Glucose (mg/dL)	Absorbance
50	.10
100	.20
150	.30
200	.40
250	.50
300	.60

To find the y-intercept, calculate $\bar{x}$ and $\bar{y}$ the average of the x- and y-values respectively

$$\bar{x} = 175$$

$$\bar{y} = 0.35$$



Calculating the Y-Intercept

- Then substitute these two values for x and y in the $\bar{y}$ = b$\bar{x}$ + a equation. Finally, solve for the unknown quantity a. Remember from the previous page that:

Formula

- Regression Equation(y) = a + bx
- Slope(b) = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$
- Intercept(a) = $(\Sigma Y - b(\Sigma X)) / N$
- Where
  - x and y are the variables.
  - b = The slope of the regression line
  - a = The intercept point of the regression line and the y axis.
  - N = Number of values or elements
  - X = First Score
  - Y = Second Score
  - ΣXY = Sum of the product of first and Second Scores
  - ΣX = Sum of First Scores
  - ΣY = Sum of Second Scores
  - ΣX^2 = Sum of square First Scores

Regression Example

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

To find regression equation, we will first find slope, intercept and use it to form regression equation

Step 1 and 2

- Count the number of values. N = 5
- Find XY, X^2

X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 = 186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Step 3

- Find ΣX, ΣY, ΣXY, ΣX^2.
  - $\Sigma X = 311$
  - $\Sigma Y = 18.6$
  - $\Sigma XY = 1159.7$
  - $\Sigma X^2 = 19359$

Step 4

- Substitute in slope formula given
  - $\text{Slope}(b) = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$
  - $((5)*(1159.7)-(311)*(18.6))/((5)*(19359)-(311)^2)$
  - $(5798.5 - 5784.6)/(96795 - 96721) = 13.9/74 = 0.19$

Step 5

- Substitute in intercept formula
  - Intercept(a) = $(\Sigma Y - b(\Sigma X)) / N$
  - (18.6 - 0.19(311))/5
  - (18.6 - 59.09)/5
  - -40.49/5 = **-8.098**

Step 6

- Then substitute these values in regression equation formula
- Regression Equation(y) = a + bx
  - **-8.098 + 0.19x**
- Suppose if we want to know the approximate y value for the variable x = 64. Then we can substitute the value
  - Regression Equation(y) = a + bx
  - -8.098 + 0.19(64).
  - -8.098 + 12.16 = **4.06**

Simple Linear Regression Example

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

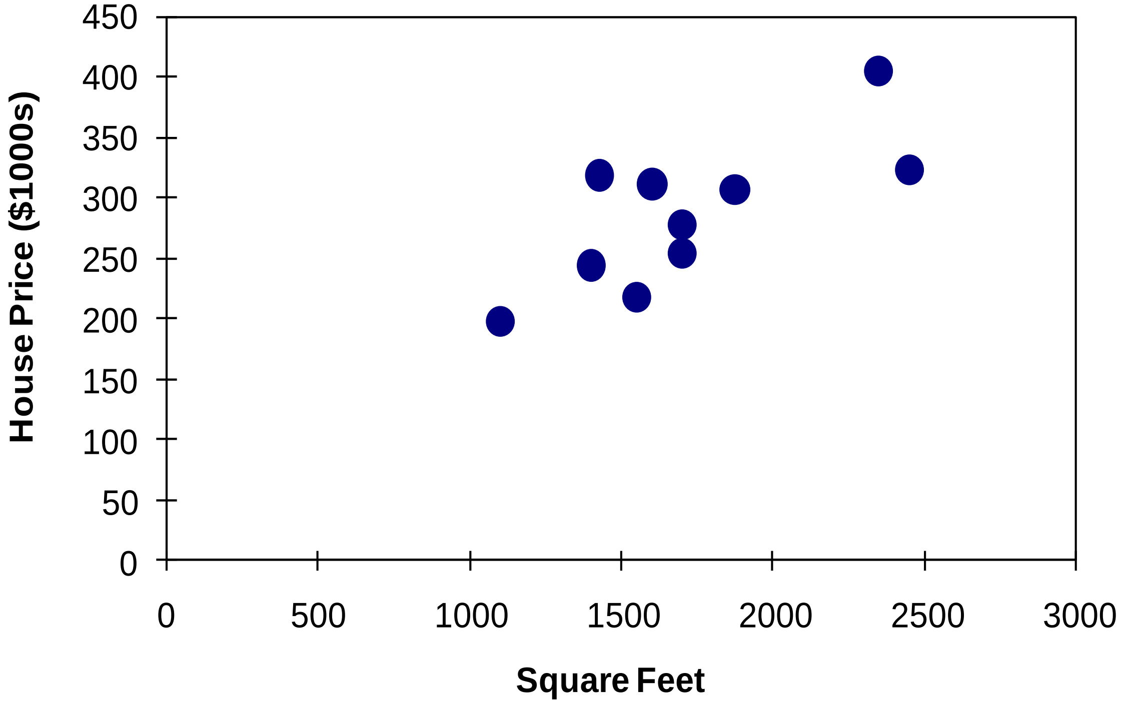
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Simple Linear Regression Example: Data

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 Example: Scatter Plot

House price model: Scatter Plot



SPSS Output

Coefficients[a]

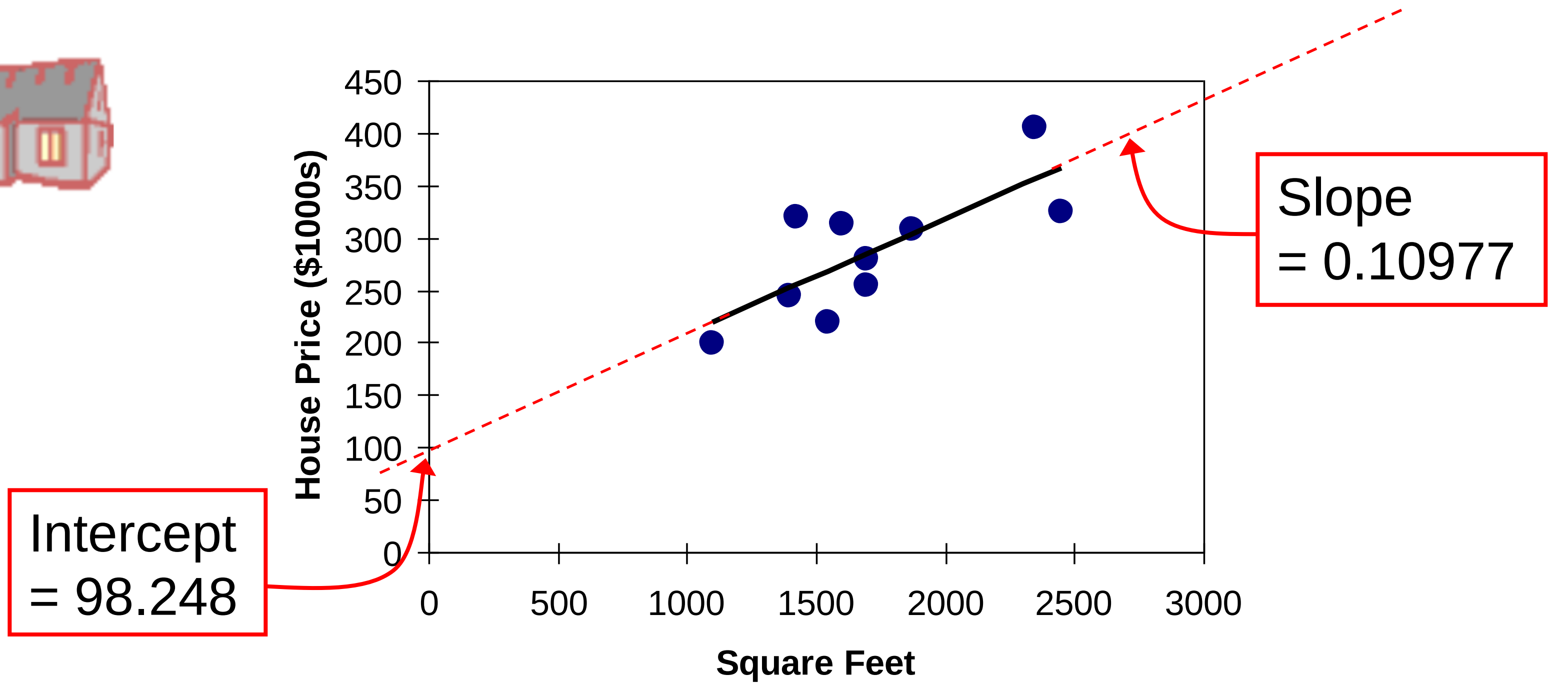
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.248	58.033		1.693	.129
	Area in Square Feet	.110	.033	.762	3.329	.010

a. Dependent Variable: House Price

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

Simple Linear Regression Example: Graphical Representation

House price model: Scatter Plot and Prediction Line



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

Simple Linear Regression Example: Interpretation of b_o

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

- b_0 is the estimated mean value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
- Because a house cannot have a square footage of 0, b_0 has no practical application

Example

A statistics professor wants to use the number of hours a student studies for a statistics final exam (X) to predict the final exam score (Y). A regression model was fit based on data collected for a class during the previous semester, with the following results:

$$\hat{Y}_i = 35.0 + 3X_i$$

What is the interpretation of the Y intercept, b_0, and the slope, b_1?

SOLUTION The Y intercept $b_0 = 35.0$ indicates that when the student does not study for the final exam, the predicted final exam score is 35.0. The slope $b_1 = 3$ indicates that for each increase of one hour in studying time, the mean change in the final exam score is predicted to be $+3.0$. In other words, the final exam score is predicted to increase by 3 points for each one-hour increase in studying time.

Simple Linear Regression Example: Interpreting b_1

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

- b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X
  - Here, $\boxed{b_1 = 0.10977}$ tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Simple Linear Regression Example: Making Predictions

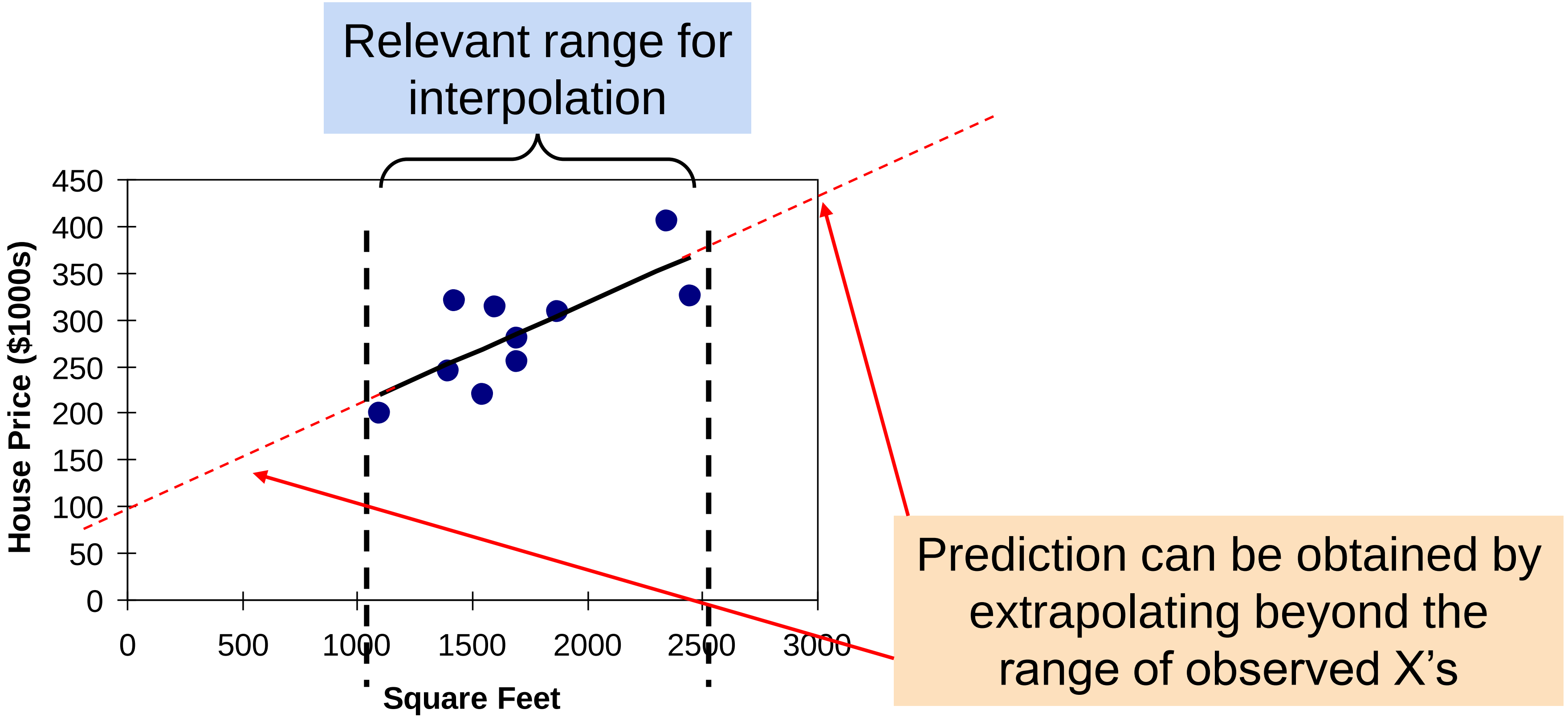
Predict the price for a house with 2000 square feet:

$$
\begin{aligned}
\widehat{\text{house price}} &= 98.24833 + 0.10977\,(\text{sq.ft.}) \\
&= 98.24833 + 0.10977(2000) \\
&= 317.78
\end{aligned}
$$

The predicted price for a house with 2000 square feet is 317.78($1,000s) = $317,780

Simple Linear Regression Example: Making Predictions

- What will happened when we try to extrapolate the results?



Store	Square Feet (X)	Annual Sales (Y)	X^2	Y^2	XY
1	1.7	3.7	2.89	13.69	6.29
2	1.6	3.9	2.56	15.21	6.24
3	2.8	6.7	7.84	44.89	18.76
4	5.6	9.5	31.36	90.25	53.20
5	1.3	3.4	1.69	11.56	4.42
6	2.2	5.6	4.84	31.36	12.32
7	1.3	3.7	1.69	13.69	4.81
8	1.1	2.7	1.21	7.29	2.97
9	3.2	5.5	10.24	30.25	17.60
10	1.5	2.9	2.25	8.41	4.35
11	5.2	10.7	27.04	114.49	55.64
12	4.6	7.6	21.16	57.76	34.96
13	5.8	11.8	33.64	139.24	68.44
14	3.0	4.1	9.00	16.81	12.30
Totals	40.9	81.8	157.41	594.90	302.30

COMPUTATIONAL FORMULA FOR THE SLOPE, b_1

$$b_1 = \frac{SSXY}{SSX}$$

where

$$SSXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SSX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

COMPUTATIONAL FORMULA FOR THE Y INTERCEPT, b_0

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

where

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$SSXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SSXY = 302.3 - \frac{(40.9)(81.8)}{14}$$

$$= 302.3 - 238.97285$$

$$= 63.32715$$

$$SSX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

$$= 157.41 - \frac{(40.9)^2}{14}$$

$$= 157.41 - 119.48642$$

$$= 37.92358$$

$$b_1 = \frac{SSXY}{SSX} = \frac{63.32715}{37.92358} = 1.67$$

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{81.8}{14} = 5.842857$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{40.9}{14} = 2.92143$$

$$\begin{aligned} b_0 &= \overline{Y} - b_1 \overline{X} \\ &= 5.842857 - (1.6699)(2.92143) \\ &= 0.9645 \end{aligned}$$

Check the results with SPSS, Rapidminer, etc.