Regression Analysis

Lecture 3

Chapter 2: Regression Analysis and Time Series Analysis

Institute of Engineering
Asst. Prof. Anita Prajapati, Ph.D.

7 June 2023

Spreadsheet Modeling & Decision Analysis:

Chapter 9 : Regression Analysis

A Practical Introduction to Management Science, 3e

by Cliff Ragsdale

Introduction to Regression Analysis (RA)

Regression Analysis is used to estimate a function $f(\ )$ that describes the relationship between a continuous dependent variable and one or more independent variables.

$$Y = f(X_1, X_2, X_3, \ldots, X_n) + \varepsilon$$

Note:

- $f(\ )$ describes *systematic* variation in the relationship.
- ε represents the *unsystematic* variation (or random error) in the relationship.

Spreadsheet Modeling and Decision Analysis, 3e, by Cliff Ragsdale. © 2001 South-Western/Thomson Learning.

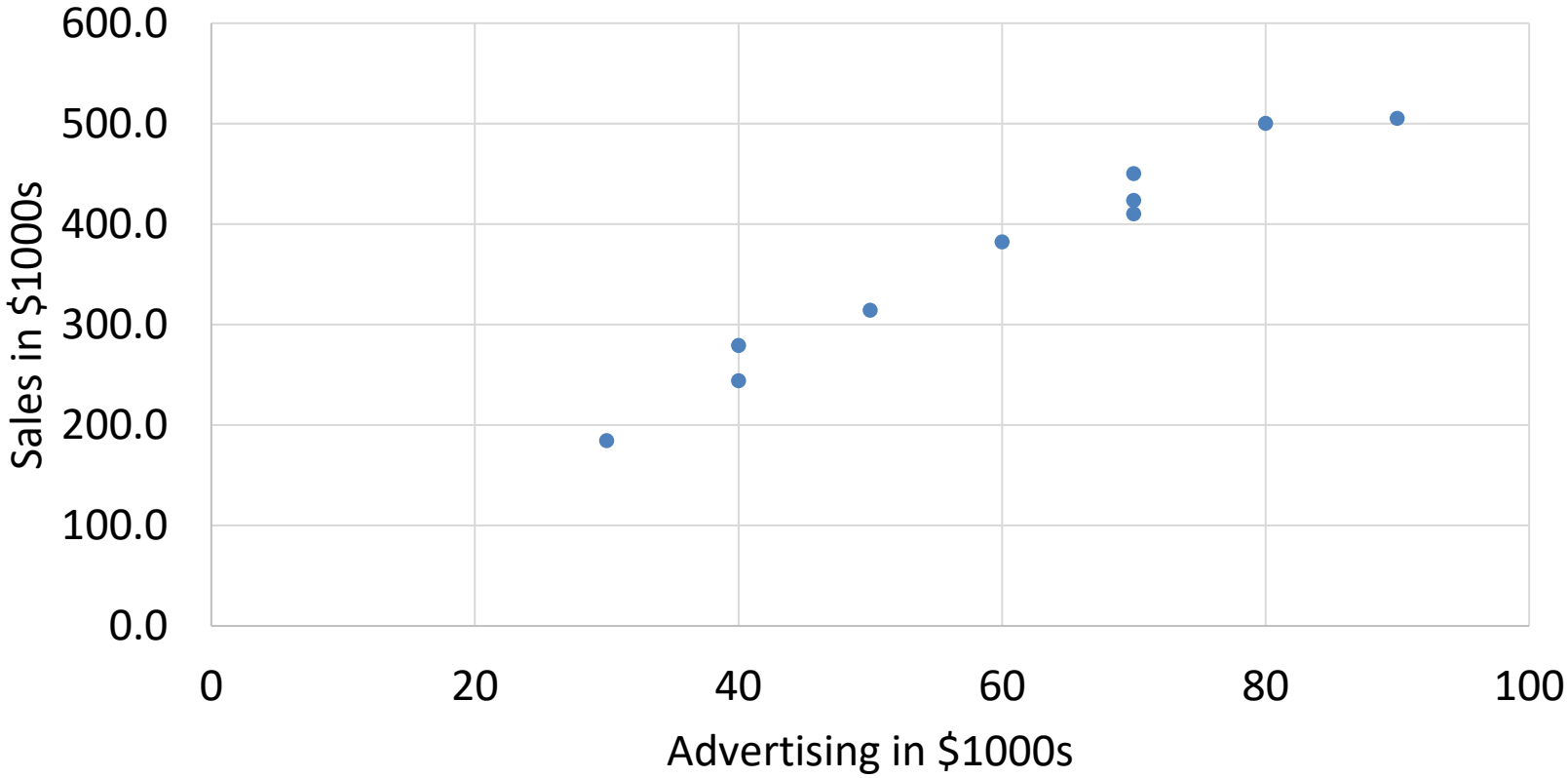
An Example

- Consider the relationship between advertising (X_1) and sales (Y) for a company.
- There probably *is* a relationship...
  - ...as advertising increases, sales should increase.
- But how would we measure and quantify this relationship?

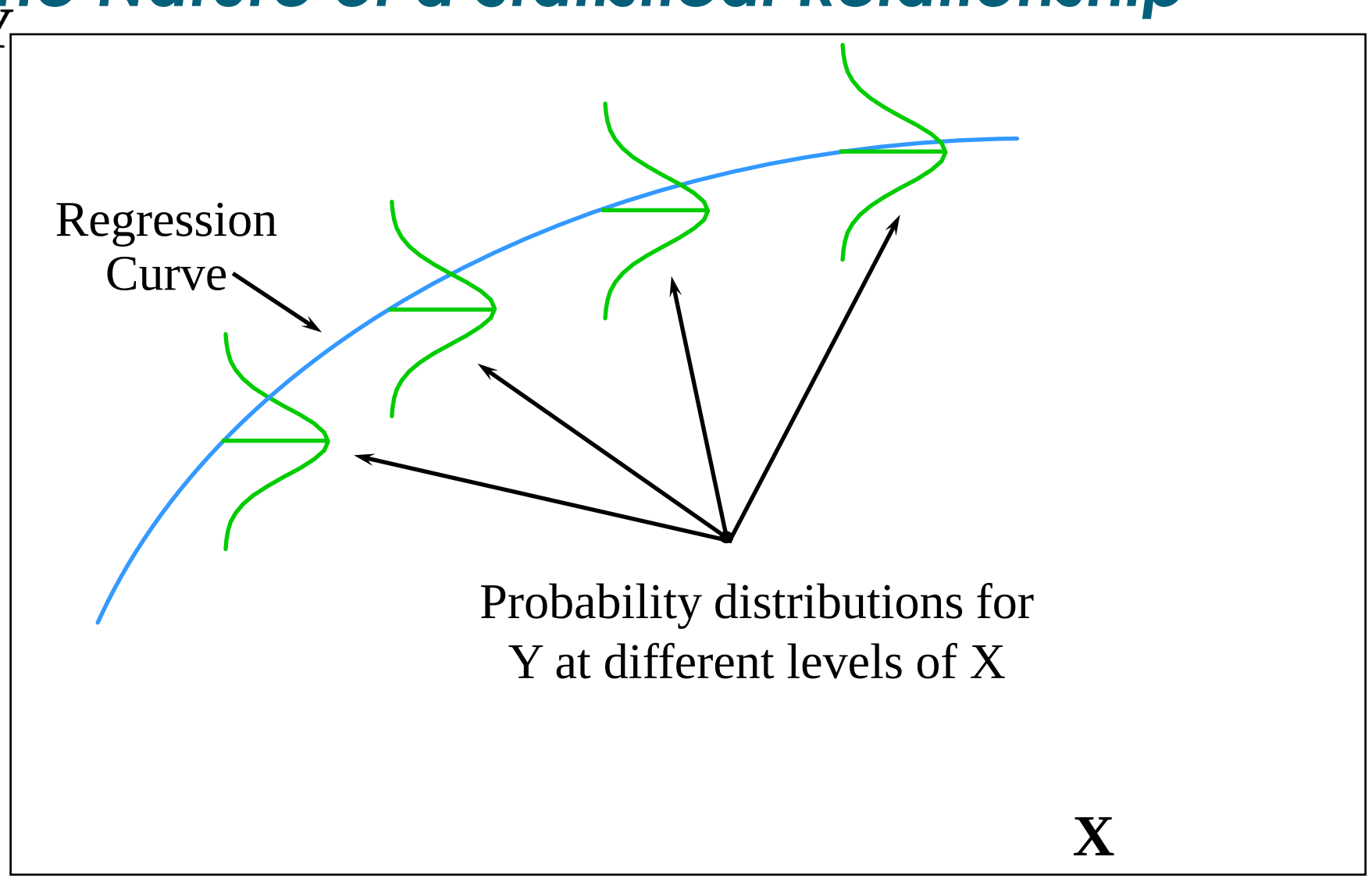
See file Fig 9-1.xls

Spreadsheet Modeling and Decision Analysis, 3e, by Cliff Ragsdale. © 2001 South-Western/Thomson Learning.

A Scatter Plot of the Data



The Nature of a Statistical Relationship



Spreadsheet Modeling and Decision Analysis, 3e, by Cliff Ragsdale. © 2001 South-Western/Thomson Learning.

A Simple Linear Regression Model

- The scatter plot shows a linear relation between advertising and sales.
- So the following regression model is suggested by the data,

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \varepsilon_i$$

This refers to the true relationship between the entire population of advertising and sales values.

- The estimated regression function (based on our sample) will be represented as,

$$\hat{Y}_i = b_0 + b_1 X_{1_i}$$

$\hat{Y}_i$ is the estimated (of fitted) value of Y at a given level of X

Determining the Best Fit

- Numerical values must be assigned to b_0 and b_1
- The method of "least squares" selects the values that minimize:

$$\mathrm{ESS} = \sum_{i=1}^{n} (\mathrm{Y}_i - \hat{\mathrm{Y}}_i)^2 = \sum_{i=1}^{n} (\mathrm{Y}_i - (b_0 + b_1 \mathrm{X}_{1_i}))^2$$

- If ESS=0 our estimated function fits the data perfectly.
- We could solve this problem using Solver...

Using Solver…

See file Fig9-4.xls

The Estimated Regression Function

The estimated regression function is:

$$\hat{Y}_i = 36.342 + 5.550X_{1_i}$$

Using the Regression Tool

- Excel also has a built-in tool for performing regression that:
  - is easier to use
  - provides a lot more information about the problem

See file Fig9-1.xls

The TREND() Function

TREND(Y-range, X-range, X-value for prediction)

where:

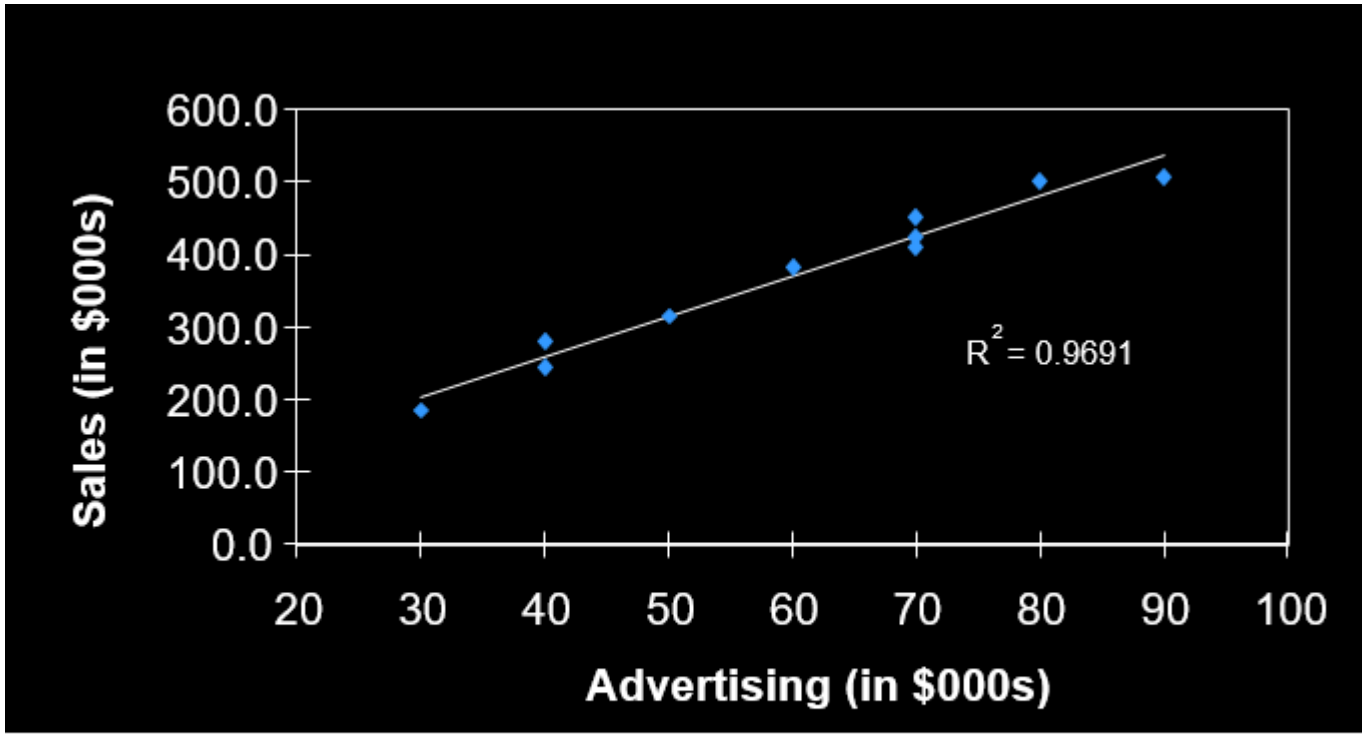
Y-range is the spreadsheet range containing the dependent Y variable,

X-range is the spreadsheet range containing the independent X variable(s),

X-value for prediction is a cell (or cells) containing the values for the independent X variable(s) for which we want an estimated value of Y.

Note: The TREND( ) function is dynamically updated whenever any inputs to the function change. However, it does not provide the statistical information provided by the regression tool. It is best to use these two different approaches to doing regression in conjunction with one another.

Evaluating the "Fit"



The R^2 Statistic

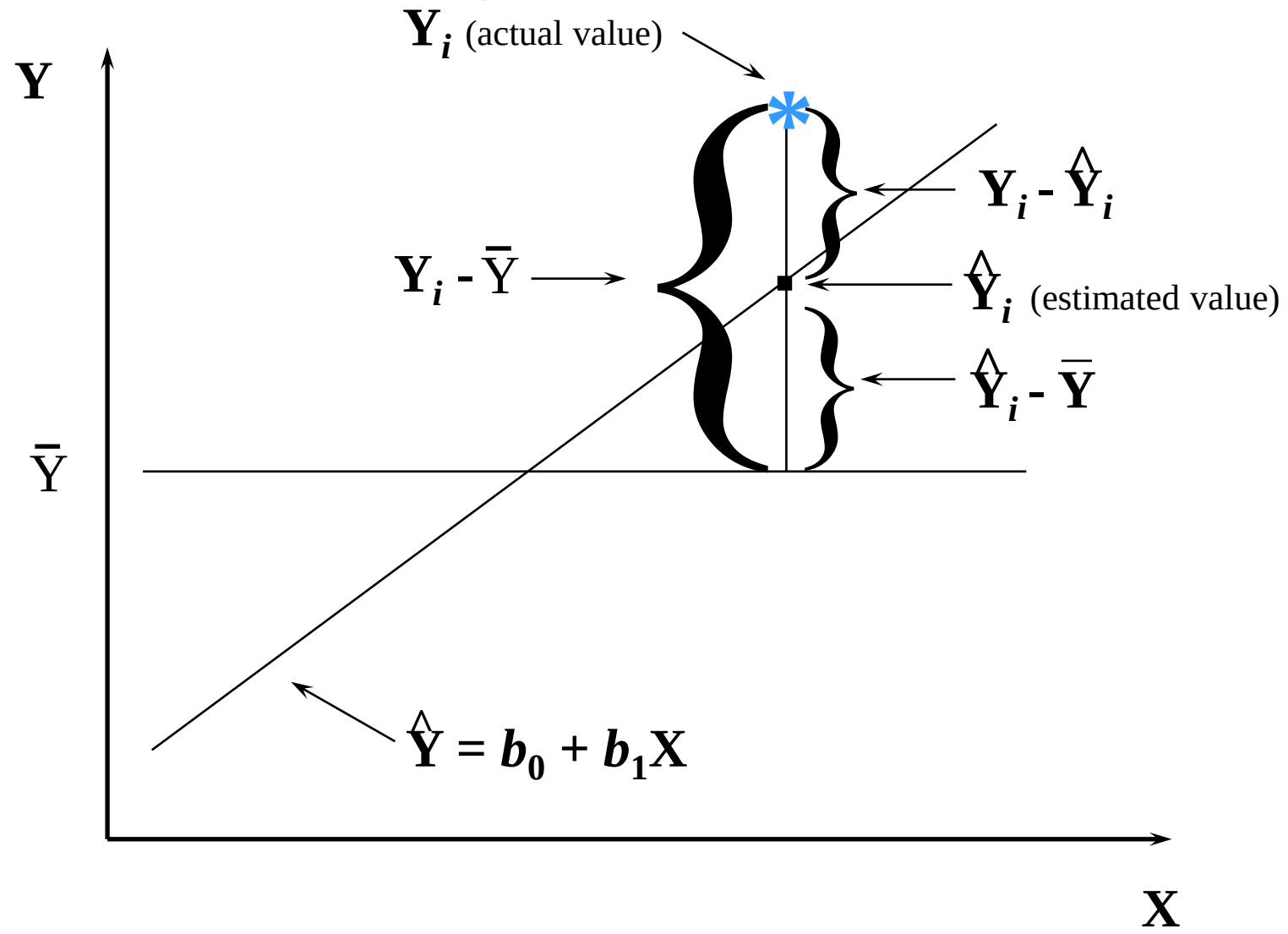
- The R^2 statistic indicates how well an estimated regression function fits the data.

$$0 <= R^2 <= 1$$

- It measures the proportion of the total variation in Y around its mean that is accounted for by the estimated regression equation.

- To understand this better, consider the following graph...

Error Decomposition



Spreadsheet Modeling and Decision Analysis, 3e, by Cliff Ragsdale. © 2001 South-Western/Thomson Learning.

Partition of the Total Sum of Squares

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

or,

TSS = ESS + RSS

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

Making Predictions

- Suppose we want to estimate the average levels of sales expected if $65,000 is spent on advertising.

$$\hat{Y}_i = 36.342 + 5.550X_{1_i}$$

- Estimated Sales = 36.342 + 5.550 * 65
- = 397.092

- So when $65,000 is spent on advertising, we expect the average sales level to be $397,092

The Standard Error

- The standard error measures the scatter in the actual data around the estimate regression line.

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n - k - 1}}$$

where k = the number of independent variables

For our example, $S_e = 20.421$

- This is helpful in making predictions...

An Approximate Prediction Interval

An approximate 95% prediction interval for a new value of Y when $X_1=X_{1h}$ is given by

$$\hat{Y}_h \pm 2S_e$$

where:

$$\hat{Y}_h = b_0 + b_1 X_{1_h}$$

Example: If $65,000 is spent on advertising:

95% lower prediction interval = 397.092 - 2*20.421 = 356.250

95% upper prediction interval = 397.092 + 2*20.421 = 437.934

If we spend $65,000 on advertising we are approximately 95% confident actual sales will be between $356,250 and $437,934.

An Exact Prediction Interval

A (1-α)% prediction interval for a new value of Y when $X_1=X_{1_h}$ is given by

$$\hat{Y}_h \pm t_{(1-\alpha/2,n-2)} S_p$$

where

$$\hat{Y}_h = b_0 + b_1 X_{1_h}$$

$$S_p = S_e \sqrt{1 + \frac{1}{n} + \frac{(X_{1_h} - \overline{X})^2}{\sum_{i=1}^{n} (X_{1_i} - \overline{X})^2}}$$

Example

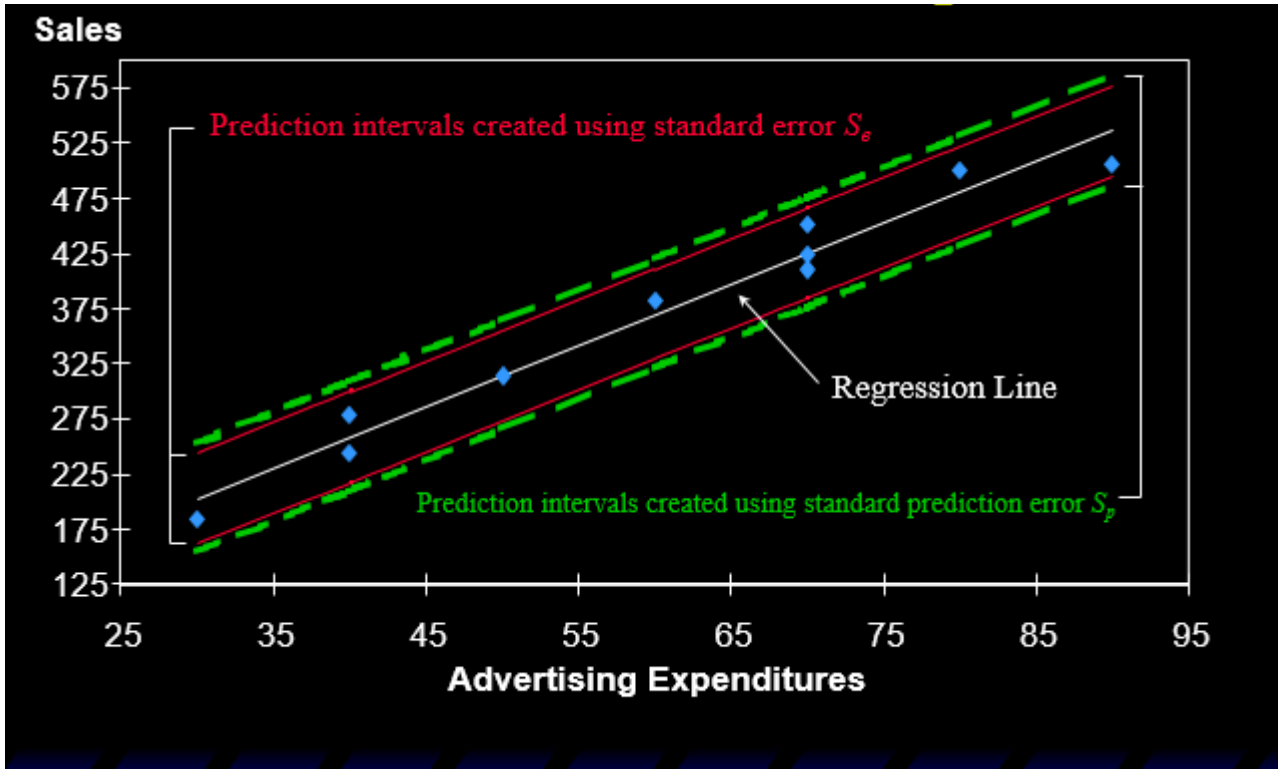
- If $65,000 is spent on advertising:
  - 95% lower prediction interval = 397.092 - 2.306*21.489 = 347.556
  - 95% upper prediction interval = 397.092 + 2.306*21.489 = 446.666

- If we spend $65,000 on advertising we are 95% confident actual sales will be between $347,556 and $446,666.

- This interval is only about $20,000 wider than the approximate one calculated earlier but was much more difficult to create.

- The greater accuracy is not always worth the trouble.

Comparison of Prediction Interval Techniques



Confidence Intervals for the Mean

A (1-α)% confidence interval for the true mean value of Y when $X_1=X_{1_h}$ is given by

$$\hat{Y}_h \pm t_{(1-\alpha/2,n-2)} S_a$$

where:

$$\hat{Y}_h = b_0 + b_1 X_{1_h}$$

$$S_a = S_e \sqrt{\frac{1}{n} + \frac{(X_{1_h} - \overline{X})^2}{\sum_{i=1}^{n} (X_{1_i} - \overline{X})^2}}$$

A Note About Extrapolation

Predictions made using an estimated regression function may have little or no validity for values of the independent variables that are substantially different from those represented in the sample.

Regression Analysis (cont…)

Lecture 4

Chapter 2: Regression Analysis and Time Series Analysis

Institute of Engineering
Asst. Prof. Anita Prajapati, Ph.D.

9 June 2023

Multiple Regression Analysis

- Most regression problems involve more than one independent variable

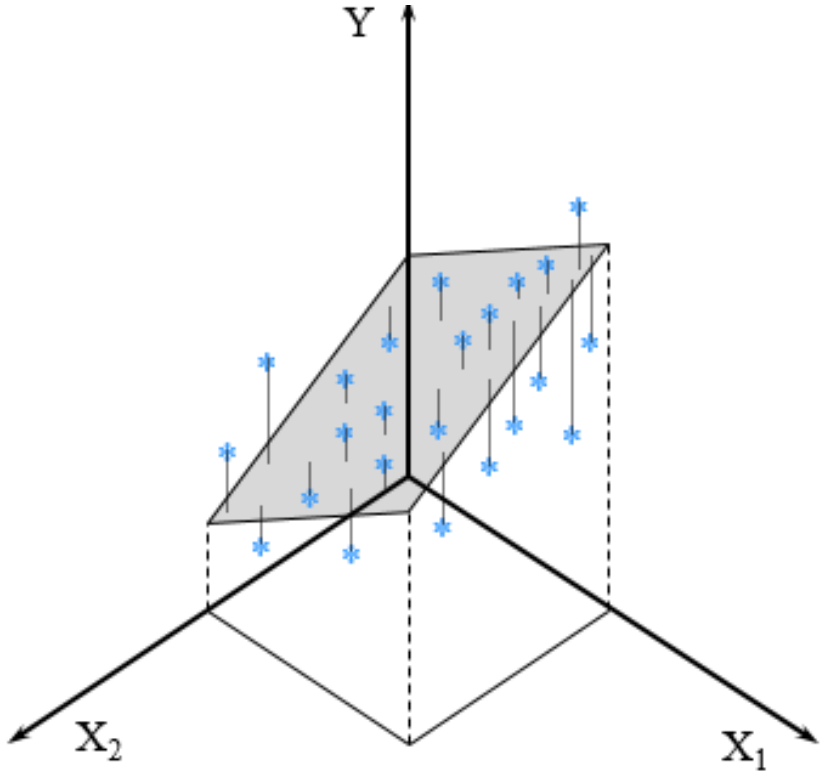
- If each independent variables varies in a linear manner with Y, the estimated regression function in this case is:

$$\hat{\mathrm{Y}}_i = b_0 + b_1 \mathrm{X}_{1_i} + b_2 \mathrm{X}_{2_i} + \cdots + b_k \mathrm{X}_{k_i}$$

- The optimal values for the b_i can again be found by minimizing the ESS.

- The resulting function fits a hyperplane to our sample data.

Example Regression Surface for Two Independent Variables



Multiple Regression Example: Real Estate Appraisal

- A real estate appraiser wants to develop a model to help predict the fair market values of residential properties.

- Three independent variables will be used to estimate the selling price of a house:
  - total square footage
  - number of bedrooms
  - size of the garage

See data in file Fig9-17.xls

Selecting the Model

- We want to identify the simplest model that adequately accounts for the systematic variation in the Y variable.

- Arbitrarily using all the independent variables may result in overfitting.

- A sample reflects characteristics:
  - representative of the population
  - specific to the sample

- We want to avoid fitting sample specific characteristics -- or overfitting the data.

Models with One Independent Variable

With simplicity in mind, suppose we fit three simple linear regression functions:

$$\hat{Y}_i = b_0 + b_1 X_{1_i}$$
$$\hat{Y}_i = b_0 + b_2 X_{2_i}$$
$$\hat{Y}_i = b_0 + b_3 X_{3_i}$$

- Key regression results are:

Variables in the Model	R^2	Adjusted R^2	S_e	Parameter Estimates
X_1	0.870	0.855	10.299	b_0=9.503, b_1=56.394
X_2	0.759	0.731	14.030	b_0=78.290, b_2=28.382
X_3	0.793	0.770	12.982	b_0=16.250, b_3=27.607

The model using X_1 accounts for 87% of the variation in Y, leaving 13% unaccounted for.

Important Software Note

When using more than one independent variable, all variables for the X-range must be in one continuous block of cells (that is, in adjacent columns).

Models with Two Independent Variables

Now suppose we fit the following models with two independent variables:

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i}$$
$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_3 X_{3_i}$$

- Key regression results are:

Variables in the Model	R^2	Adjusted R^2	S_e	Parameter Estimates
X_1	0.870	0.855	10.299	b_0=9.503, b_1=56.394
X_1 & X_2	0.939	0.924	7.471	b_0=27.684, b_1=38.576 b_2=12.875
X_1 & X_3	0.877	0.847	10.609	b_0=8.311, b_1=44.313 b_3=6.743

The model using X_1 and X_2 accounts for 93.9% of the variation in Y, leaving 6.1% unaccounted for.

The Adjusted R^2 Statistic

- As additional independent variables are added to a model:
  - The R^2 statistic can only increase.
  - The Adjusted-R^2 statistic can increase *or* decrease.

$$R_a^2 = 1 - \left(\frac{ESS}{TSS}\right)\left(\frac{n-1}{n-k-1}\right)$$

- The R^2 statistic can be artificially inflated by adding ***any*** independent variable to the model.

- We can compare adjusted-R^2 values as a heuristic to tell whether adding an additional independent variable really helps to improve a regression model.

A Comment On Multicollinearity

- It should not be surprising that adding X_3 (# of bedrooms) to the model with X_1 (total square footage) did not significantly improve the model.
- Both variables represent the same (or very similar) things -- the size of the house.
- These variables are highly correlated (or collinear).
- Multicollinearity should be avoided.

Model with Three Independent Variables

- Now suppose we fit the following model with three independent variables:

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i} + b_3 X_{3_i}$$

  - Key regression results are:

Variables in the Model	R^2	Adjusted R^2	S_e	Parameter Estimates
X_1	0.870	0.855	10.299	b_0=9.503, b_1=56.394
X_1 & X_2	0.939	0.924	7.471	b_0=27.684, b_1=38.576, b_2=12.875
X_1, X_2 & X_3	0.943	0.918	7.762	b_0=26.440, b_1=30.803, b_2=12.567, b_3=4.576

- The model using X_1 and X_2 appears to be best:
  - Highest adjusted-R^2
  - Lowest S_e (most precise prediction intervals)

Making Predictions

- Let's estimate the average selling price of a house with 2,100 square feet and a 2-car garage:

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i}$$

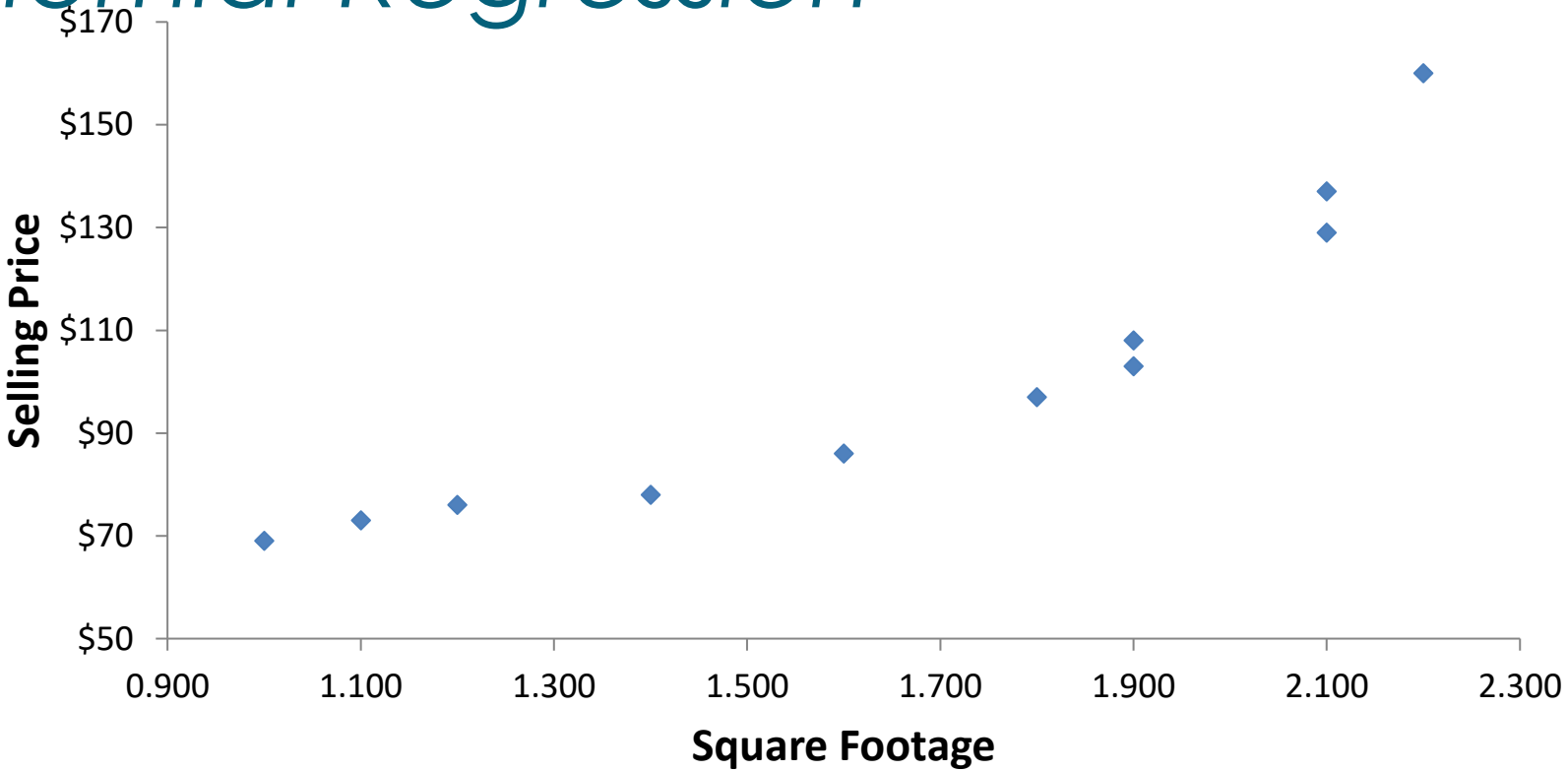
$$\hat{Y}_i = 27.684 + 38.576 * 2.1 + 12.875 * 2 = 134.444$$

- The estimated average selling price is $134,444
- A 95% prediction interval for the actual selling price is approximately:

$$\hat{Y}_h \pm 2S_e$$

- 95% lower prediction interval = 134.444 - 2*7.471 = $119,502
- 95% lower prediction interval = 134.444 + 2*7.471 = $149,386

Polynomial Regression



This graph suggests a quadratic relationship between square footage (X) and selling price (Y).

The Regression Model

An appropriate regression function in this case might be,

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{1_i}^2$$

or equivalently,

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i}$$

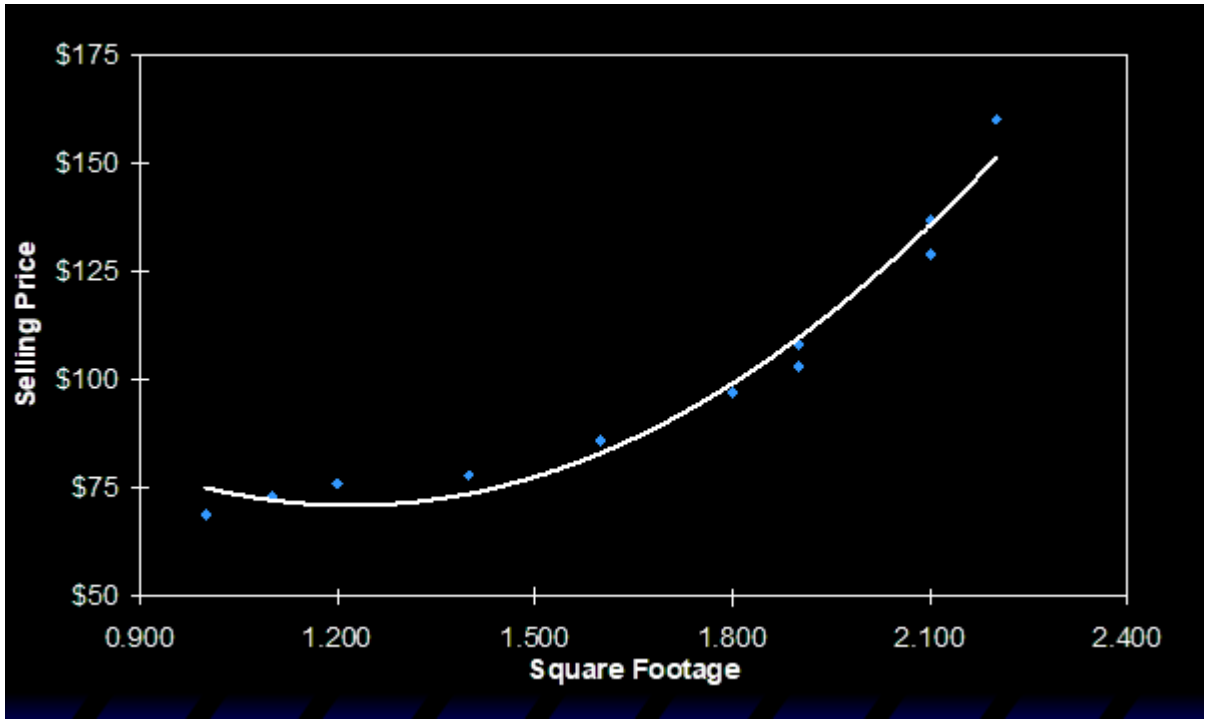
where,

$$X_{2_i} = X_{1_i}^2$$

Implementing the Model

See file Fig9-25.xls

Graph of Estimated Quadratic Regression Function



Fitting a Third Order Polynomial Model

We could also fit a third order polynomial model,

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{1_i}^2 + b_3 X_{1_i}^3$$

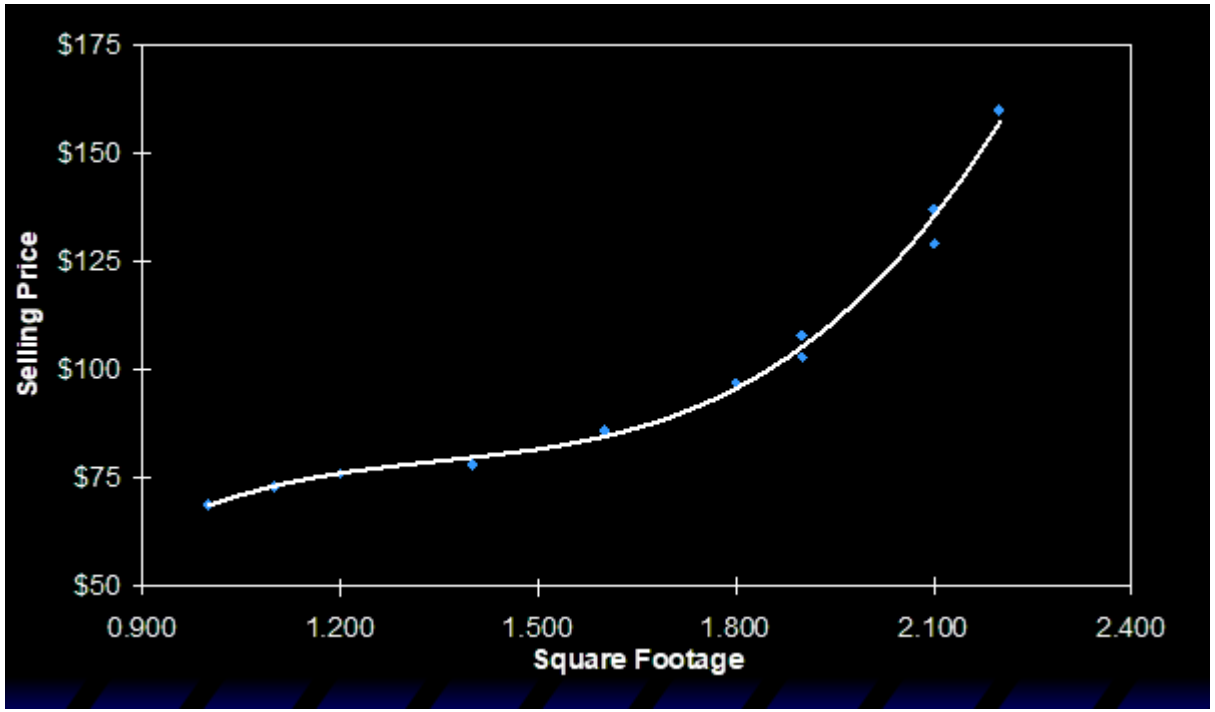
or equivalently,

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i} + b_3 X_{3_i}$$

where,

$$X_{3_i} = X_{1_i}^3$$

Graph of Estimated Third Order Polynomial Regression Function



Overfitting

When fitting polynomial models, care must be taken to avoid overfitting. The adjusted-R^2 statistic can be used for this purpose here also.

Assignments

Chapter 9

Question no: 6,8,9,10,12,13,17,20

THANK YOU!!!