Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

BU7142 Foundations of Business Analytics

Lecture 6

Regressions

Dr Yufei Huang

Key Issue

- Find a group
https://docs.google.com/spreadsheets/d/13NOMmmCuqb_dIKWFnXtdLt99KxbFEjZ9U6kOv3DPjiM/edit#gid=0

- Group Assignment Deadline: 23.59, 23 Oct (Sunday)

- Timed Individual Assignment (24h window, 9am, 05 Oct. to 9am 06 Oct.)

- Please fill in Teaching Evaluation Form when you receive the email. Many thanks!

Content

- **Simple linear regression**
  - One independent variable that can influence the dependent variable

- **Multiple Variable Regression Model**
  - There may be more than one independent variables that can influence the dependent variable

- **Non linear regression**
  - The relationship between independent variable and dependent variable may not be linear

Regression

- Regression refers to the statistical technique of modeling the relationship between variables.

- In simple linear regression, we model the relationship between two variables.

- One of the variables, denoted by Y, is called the dependent variable and the other, denoted by X, is called the independent variable.

- The model we will use to depict the relationship between X and Y will be a straight-line relationship.

- A graphical sketch of the pairs (X, Y) is called a scatter plot.

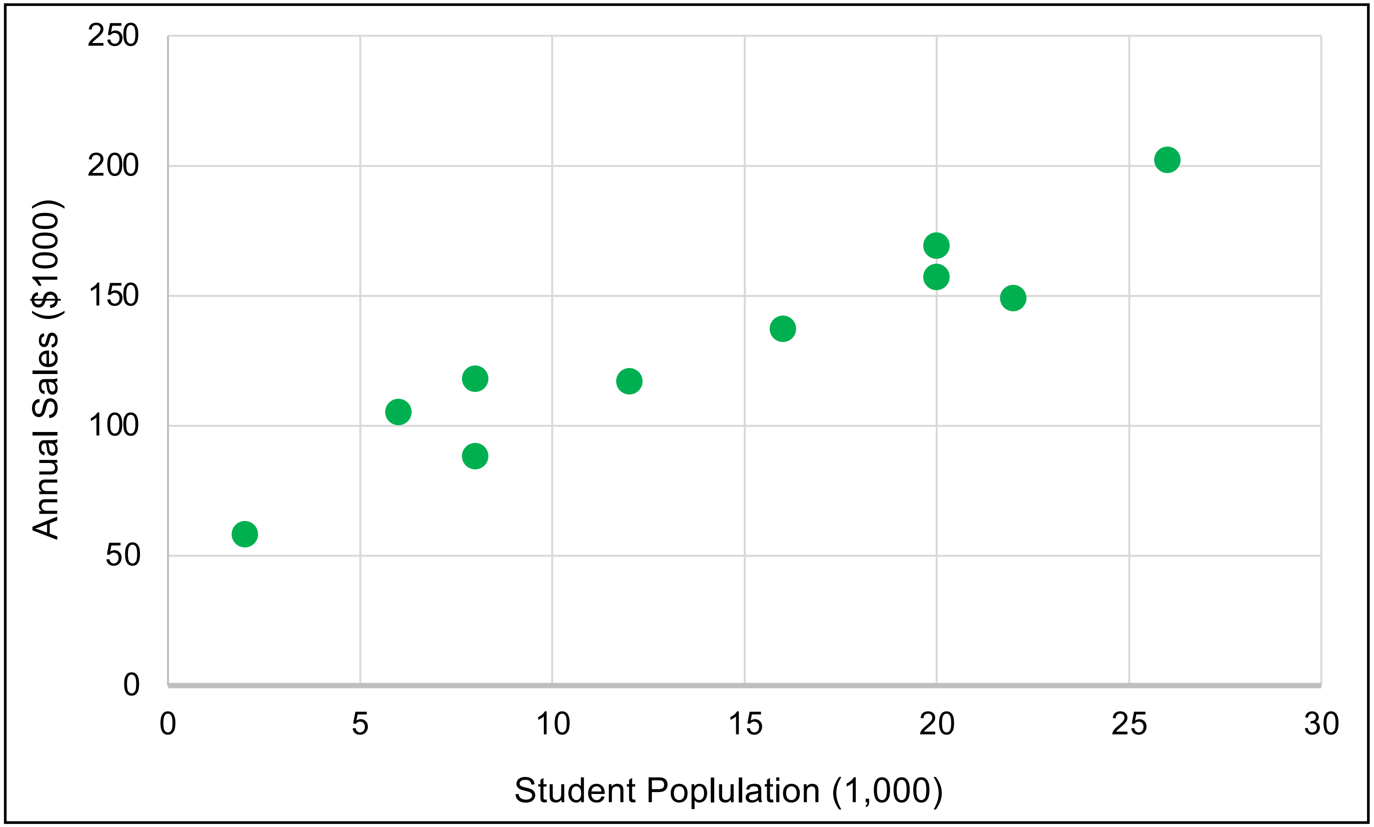
The Goal

- The basic idea in simple linear regression is to
  - (i) **establish** a relationship between a dependent variable Y and an independent variable X
  - (ii) **quantify** the magnitude of the impact of X on Y
  - (iii) **find** the 95% prediction interval for forecasting

Example: Armand's Pizza

Restaurant i	Student Population ('000) X_i	Annual Sales ($ '000) Y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Armand’s Pizza : Scatter Plot



Any relationship between Student Population and Annual Sales?
We need a statistical model to answer this question.

Model Building

A statistical model separates the systematic component of a relationship from the random component.

Data

↓

Statistical model

↓

Systematic component
+
Random errors

In regression, the systematic component is the overall linear relationship, and the random component is the variation around the line.

The Simple Linear Regression Model

The population simple linear regression model:

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{Nonrandom or Systematic Component}} + \underbrace{\varepsilon}_{\text{Random Component}}$$

where

- Y is the dependent variable, the variable we wish to explain or predict
- X is the independent variable, also called the predictor variable
- ε is the error term, the only random component in the model, and thus, the only source of randomness in Y

- β_0 is the intercept of the systematic component of the regression relationship
- β_1 is the slope of the systematic component

Assumptions of the Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- β_0 Y-intercept of the line
- β_1 the slope of the line
- ε the error

1. The error ε is a random variable with mean 0.
2. The variance of ε, denoted as σ2, is the same for all values of X.
3. The values of ε are independent.
4. The error term ε is Normally distributed.

How to Estimate?

Estimation of a simple linear regression relationship involves finding estimated or predicted values of the intercept and slope of the linear regression line.

The estimated regression equation:

$$Y = b_0 + b_1X + \varepsilon$$

where

- b_0 estimates the intercept of the population regression line, β_0 ;
- b_1 estimates the slope of the population regression line, β_1;
- ε stands for the observed errors - the residuals from fitting the estimated regression line $b_0 + b_1X$ to a set of n points.

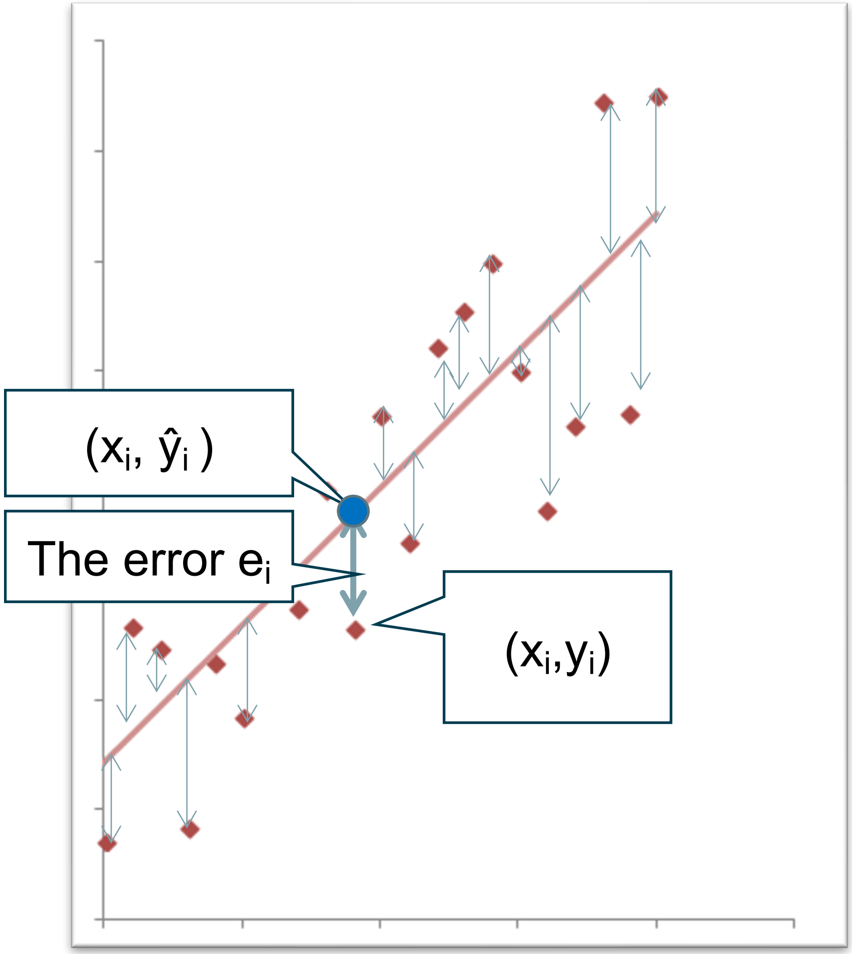
The estimated regression line:

$$\hat{Y} = b_0 + b_1 X$$

where $\hat{Y}$ (Y-hat) is the value of Y lying on the fitted regression line for a given value of X.

The method of least squares

- To find coefficients b_0, b_1,
- we denote each data point by (x_i, y_i).
- The line gives us an approximated value: $\hat{y}_i = b_0 + b_1 x_i$.
- The approximation error of each point is $e_i = |y_i - \hat{y}_i|$.
- The Sum of Squares for Errors in regression is:



$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

To find b_0, b_1, which minimise SSE

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Theorem. The following b_0 and b_1 minimise SSE :
(Least Squares Estimator)

$$\boxed{\begin{aligned} b_1 &= \frac{SS_{xy}}{SS_x}, \\ b_0 &= \bar{y} - b_1 \bar{x}, \end{aligned}}$$

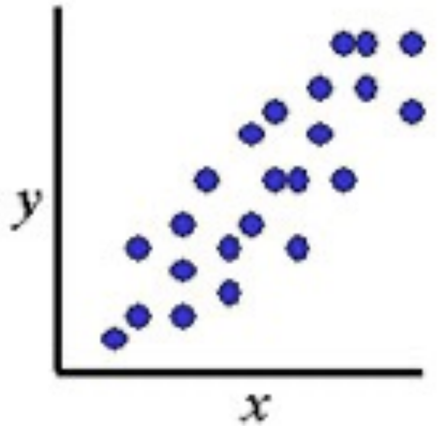
where $\bar{x} = \text{mean}(X), \bar{y} = \text{mean}(Y)$

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

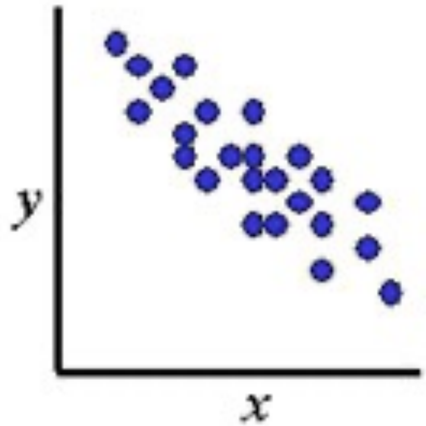
$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right).$$

What is b_1's sign in the following relationships?

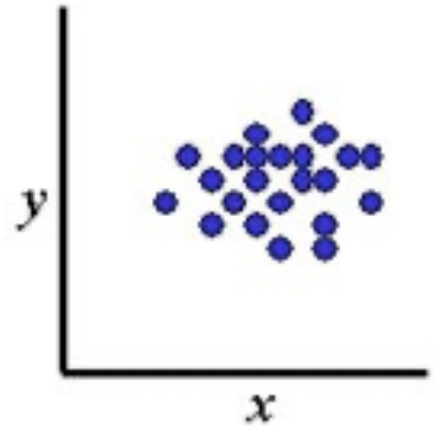
Positive b_1
As x increases,
y increases



Negative b_1
As x increases,
y decreases



b_1=0
No relation
between x and y



- It is important to check whether b_1 is significantly different that 0.
- How? Hypothesis testing.

Hypothesis testing for a linear relationship

Hypotheses:

H_0: $b_1=0$

H_1: $b_1 \neq 0$.

The test statistic for the existence of a linear relationship between X and Y can be calculated in Excel.

Armand’s Pizza: Excel Output

Regression Statistics	
Multiple R	0.950122955
R Square	0.90273363
Adjusted R Square	0.890575334
Standard Error	13.82931669
Observations	10

Standard error for Y

Sample size

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14200	14200	74.24837	2.54887E-05
Residual	8	1530	191.25		
Total	9	15730			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	60	9.22603481	6.503336	0.000187	38.72471182	81.27528818
X Variable	5	0.580265238	8.616749	2.55E-05	3.661905096	6.338094904

Estimated b1

Standard error for b1

Test statistic based on confidence level defined

Confidence Interval for b1

Regression Results

$$Y = 60 + 5^{*}X$$

Interpretation of coefficients:

- b_0 =60, is the Y-intercept of the line
- b_1 =5, is the slope of the line
- b_1 = 5 means that for a unit increase in X-value, the value of Y increases by 5 units

Forecasting: fit a line using the Least Squares Method:

- Y = 60 + 5X
- Forecast sales for X = 10: $y = 60 + 5 * 10 = 110$

Significant Relationship

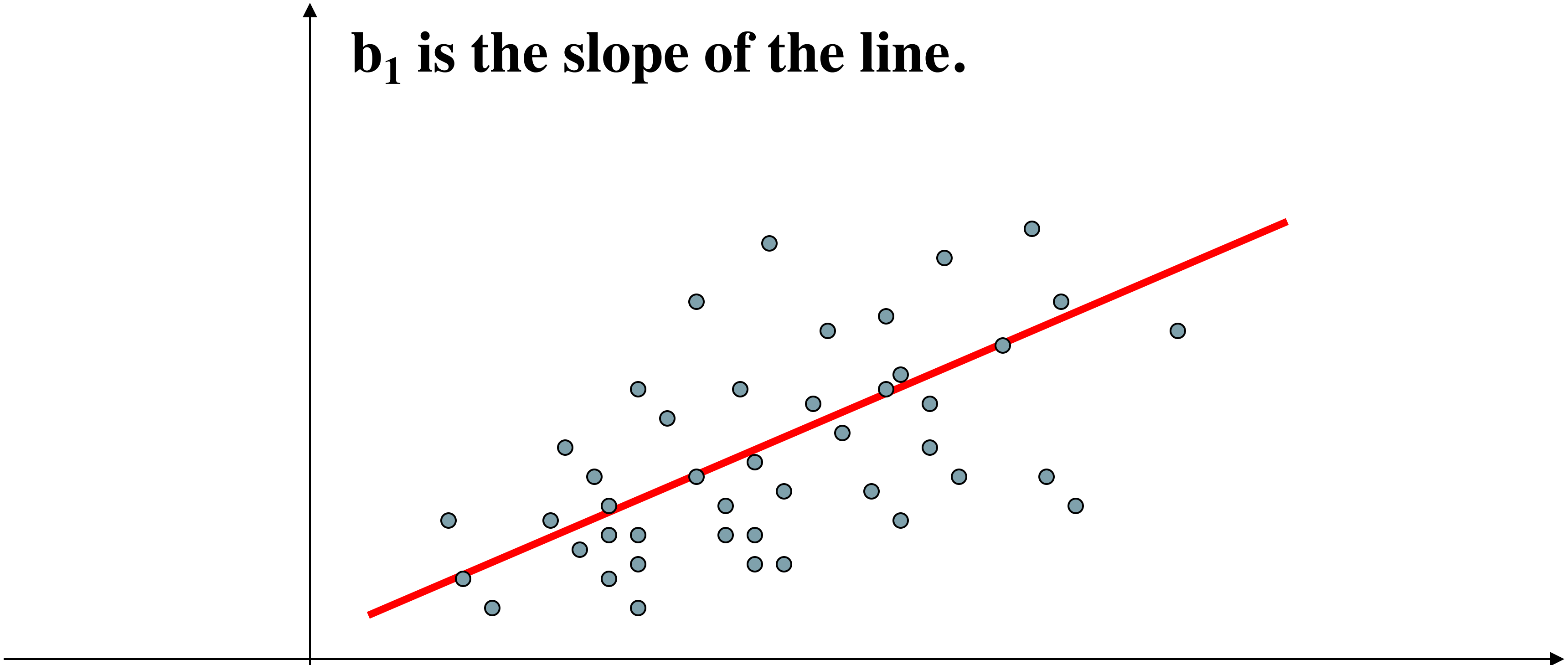
The coefficient is deemed significant at 95% confidence level:

- If the p-value associated with a coefficient is less than 0.05 (the significance level)
- If the t-stat associated with a coefficient is larger than 1.96 (normal distribution) or t(n-2,0.025) (for t distribution)
- If 0 is outside the 95% confidence interval

Then we can reject the null hypothesis ($b_1=0$), namely there is a relationship between X and Y

Is there a relationship?

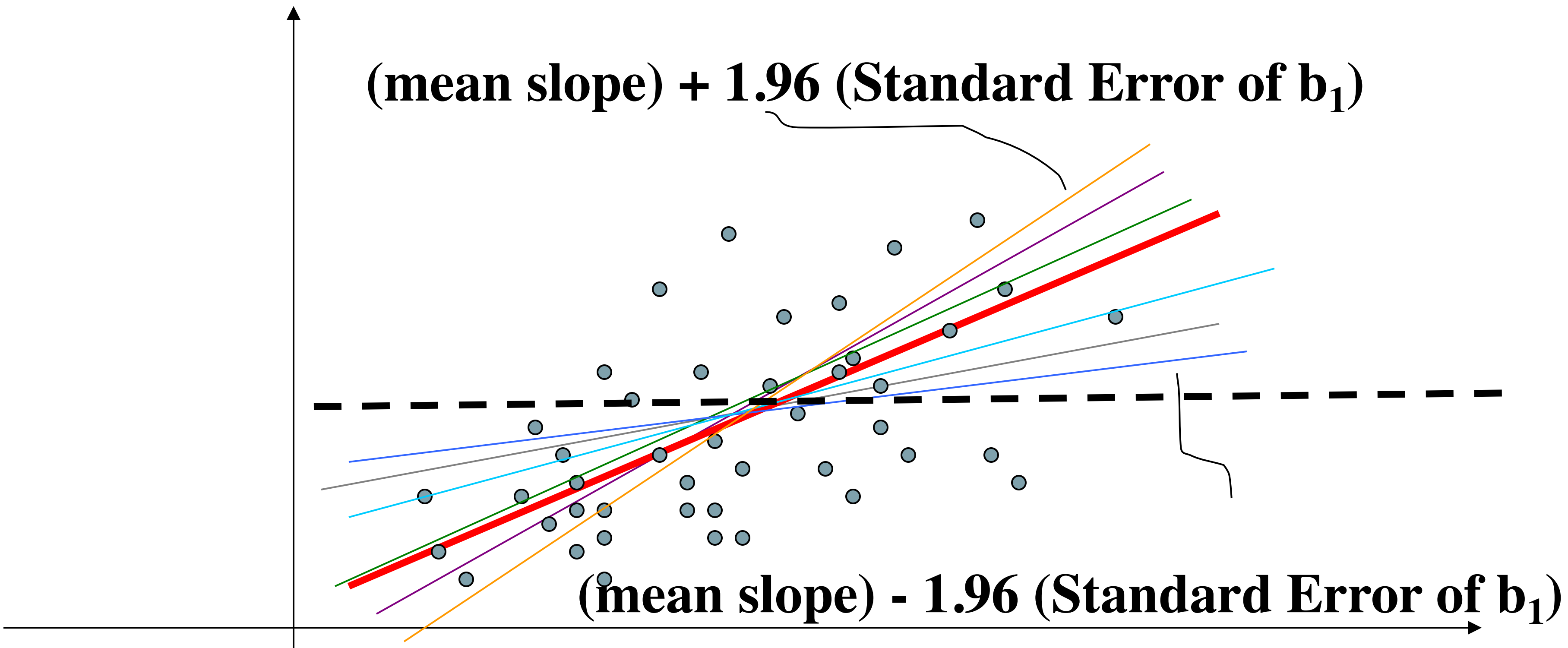
b_1 is the slope of the line.



Make sure within 95% confidence interval, the line doesn't go flat!

(mean slope) ± 1.96 (Standard Error of b_1)

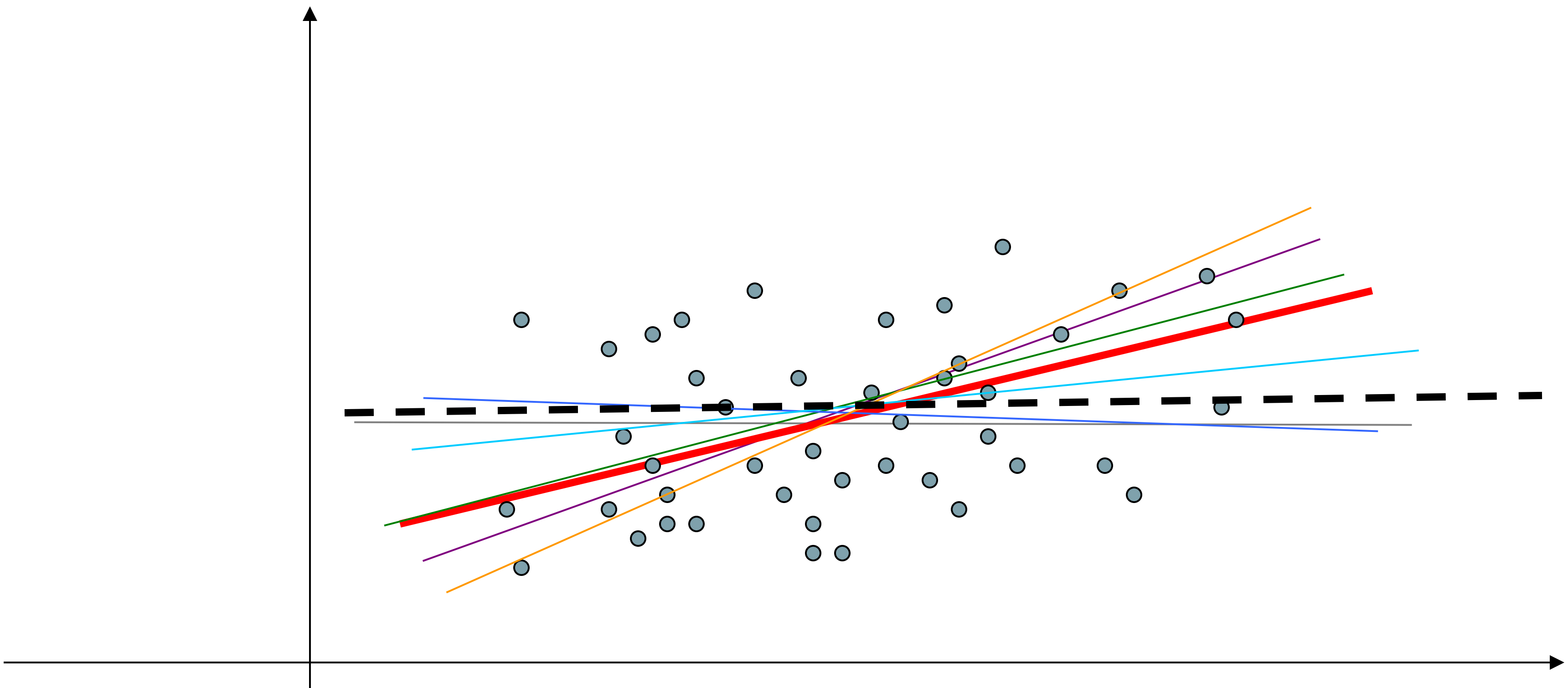
Is there a relationship?



Make sure within 95% confidence interval, the line doesn't go flat!

(mean slope) ± 1.96 (Standard Error of b_1)

Question: Is there a relationship here?



If the line could go flat, we don't claim a relationship!

(mean slope) ± 1.96 (Standard Error of b_1)

Uncertainty in Forecast

- Prediction Interval
  - With a 95% confidence level, the <u>individual</u> value of y for a given value of x will lie in the interval:

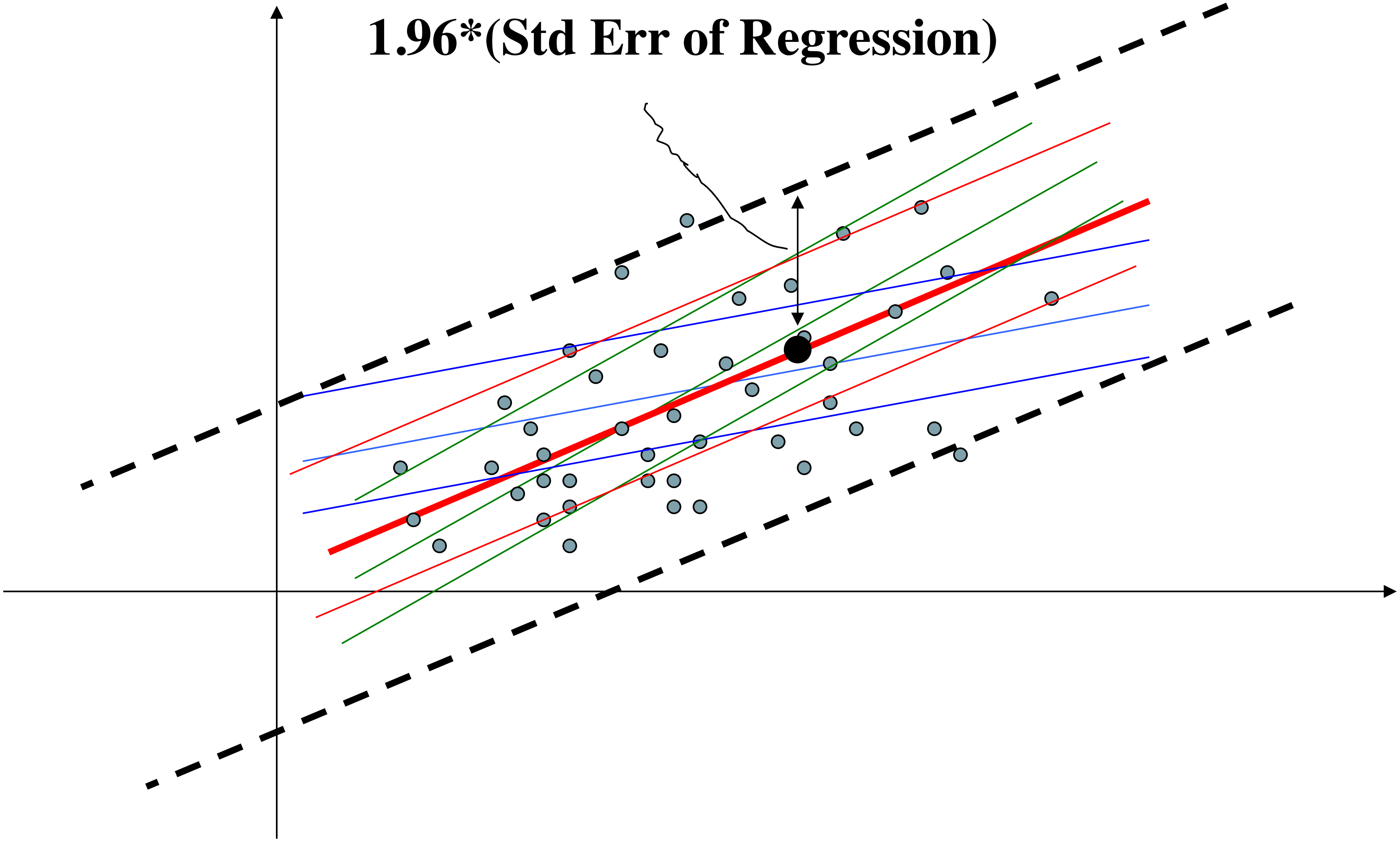
$$\boxed{\hat{y} \pm 1.96 \times \text{standard error of the estimate}}$$

When t-distribution is used (i.e., for small sample size), 1.96 needs to be replaced by $t_{(n-2,0.025)}$

  - For x = 10, the 95% prediction interval is:

$$110 \pm 2.306 \times 13.829$$

95% Prediction Interval



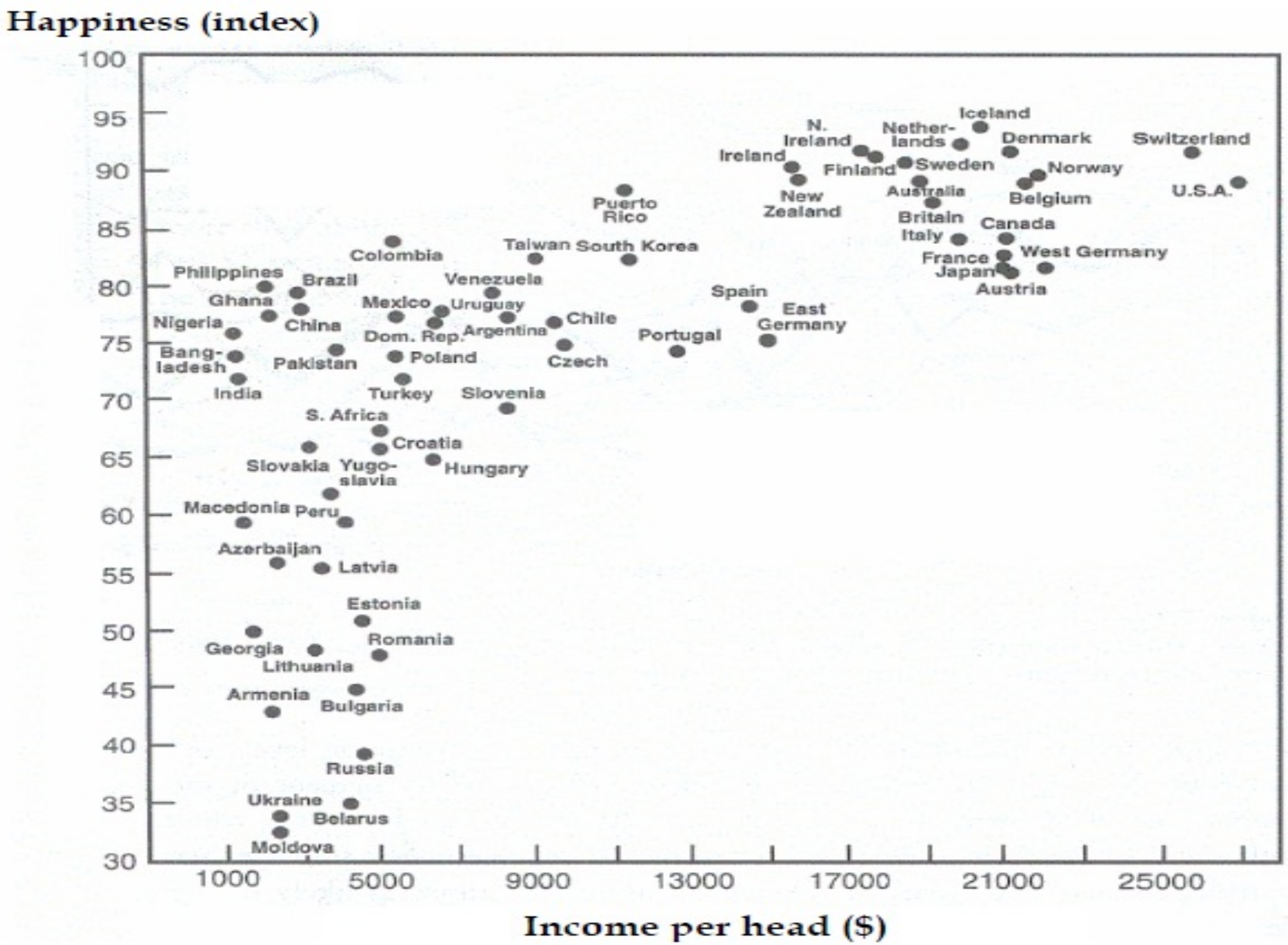
How Good Is the Fit?

- R^2 measures how well the regression line fits the data. In the pizza example, $R^2 = 0.90$. This means that 90% of the variation in sales is due to the variation in student population. The other 10% of the variation remains unexplained. ($0 \leq R^2 \leq 1$)

- R^2 is one of several statistics that should be used in evaluating the quality of the regression model.

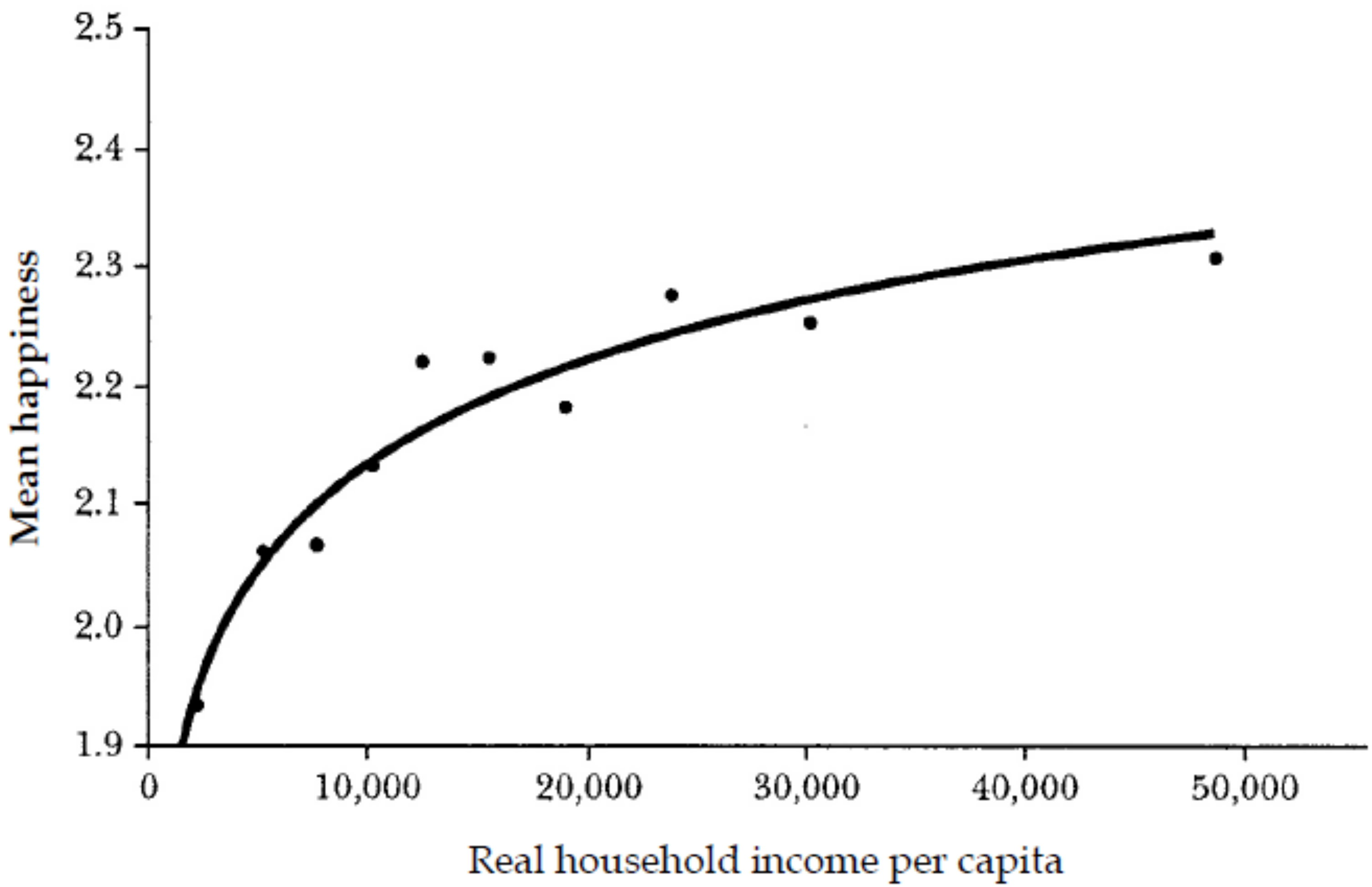
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Country Comparison of Income and Happiness



Source: Inglehart and Klingemann (2000, Fig. 7.2 and Table 7.1).

Mean Happiness and Household Income



Source: diTella and MacCullouch (2006).

Nonlinear Regression

- A nonlinear relationship may be a better model than a linear relationship.
- A widely used regression for nonlinear relationship is multiplicative regression

$$\text{The multiplicative model}:$$
$$Y = \beta_0 X^{\beta_1} \varepsilon$$

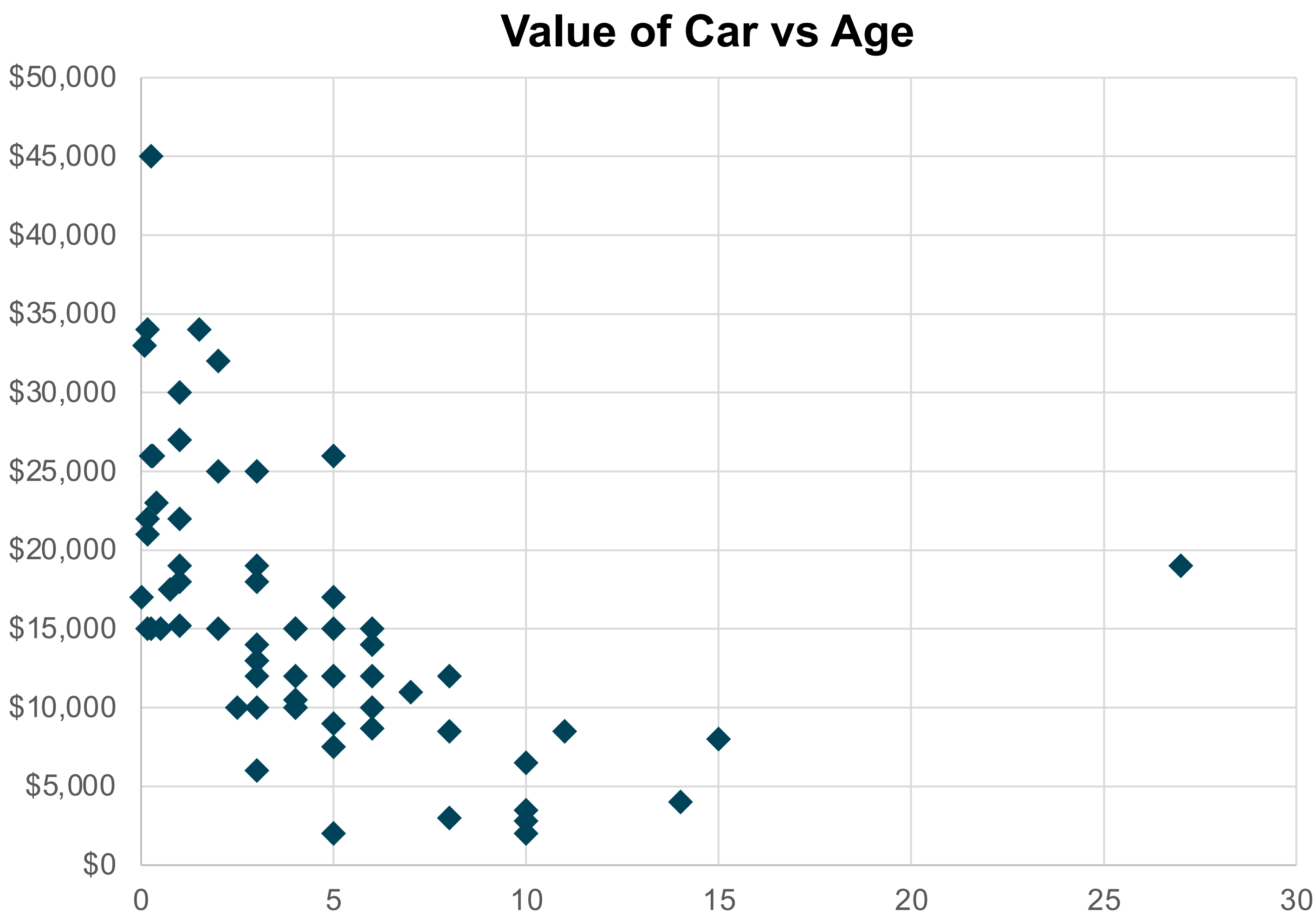
$$\text{The logarithmic transformation}:$$
$$\ln Y = \ln \beta_0 + \beta_1 \ln X + \ln \varepsilon$$

Interpreting Multiplicative Models

- $y = b_0 + b_1 \, LN(x_1) + \varepsilon$ Model (1)
  - If x_1 increases by 1%, then y increases by approximately 0.01 b_1 units.
- $LN(y) = b_0 + b_1 x_1 + \varepsilon$ Model (2)
  - If x_1 increases by 1 unit, then y increases by approximately 100 b_1%.
- $LN(y) = b_0 + b_1 \, LN(x_1) + \varepsilon$ Model (3)
  - If x_1 increases by 1%, then y increases by approximately b_1%.

- Interpretation of the coefficient b_1 is of managerial use. For example, if y is sales or demand and x_1 is price then in Model 3, the coefficient b_1 measures the elasticity of sales with respect to price. That is, in Model 3, a 1% change in price leads to approximately b_1% change in sales.

Example: Value of Second-hand Cars



Example: Value of Second-hand Cars

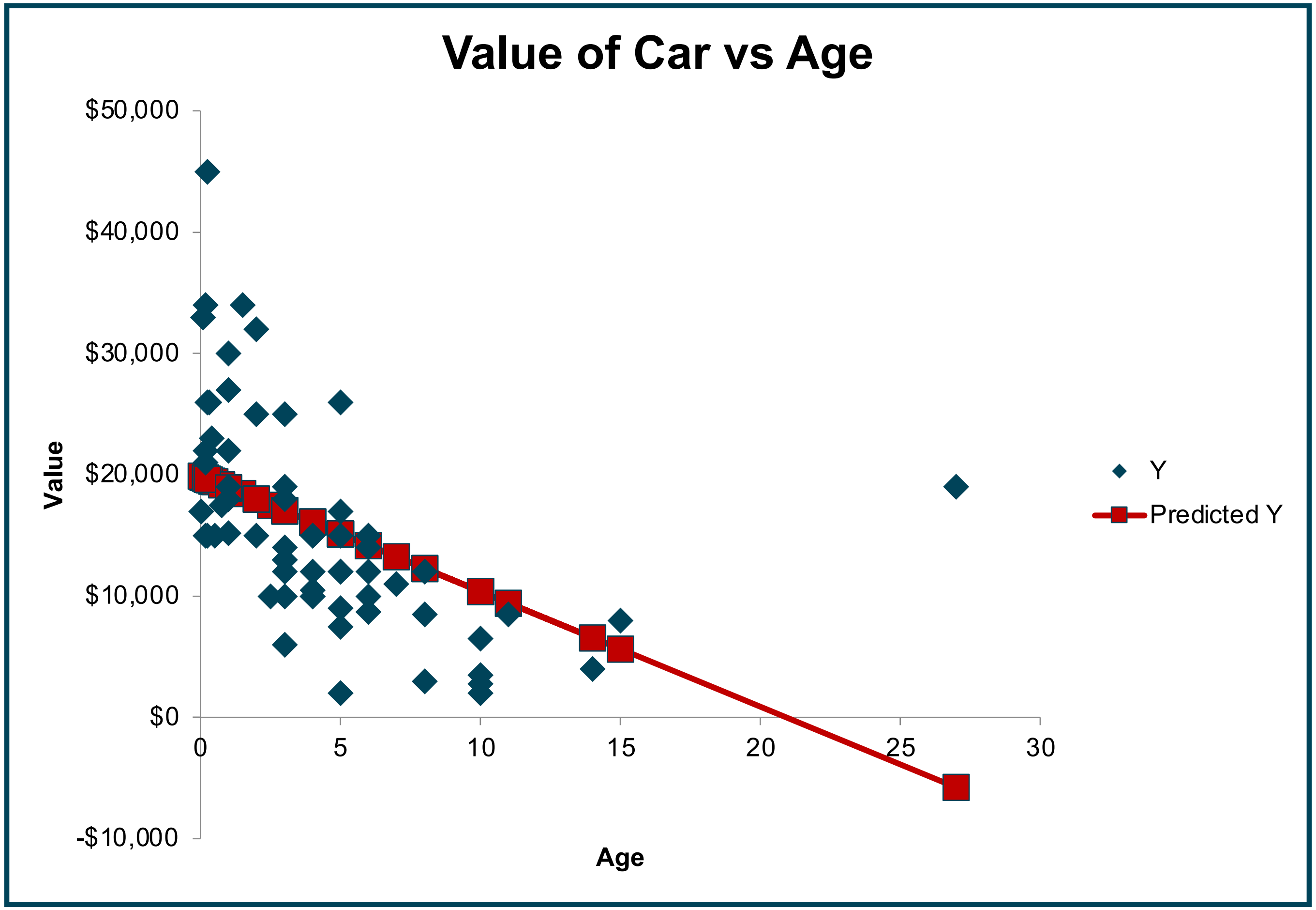
A simple linear regression model: Value = b_0 + b_1 * Age + ε

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.48506759
R Square	0.23529056
Adjusted R Square	0.22295654
Standard Error	7803.4037
Observations	64

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95.0%*	*Upper 95.0%*
Intercept	19889.8746	1359.561	14.62962	8.95E-22	17172.15	22607.6
Age	-950.6942	217.6662	-4.36767	4.86E-05	-1385.8	-515.58

Example: Value of Second-hand Cars



Example: Value of Second-hand Cars

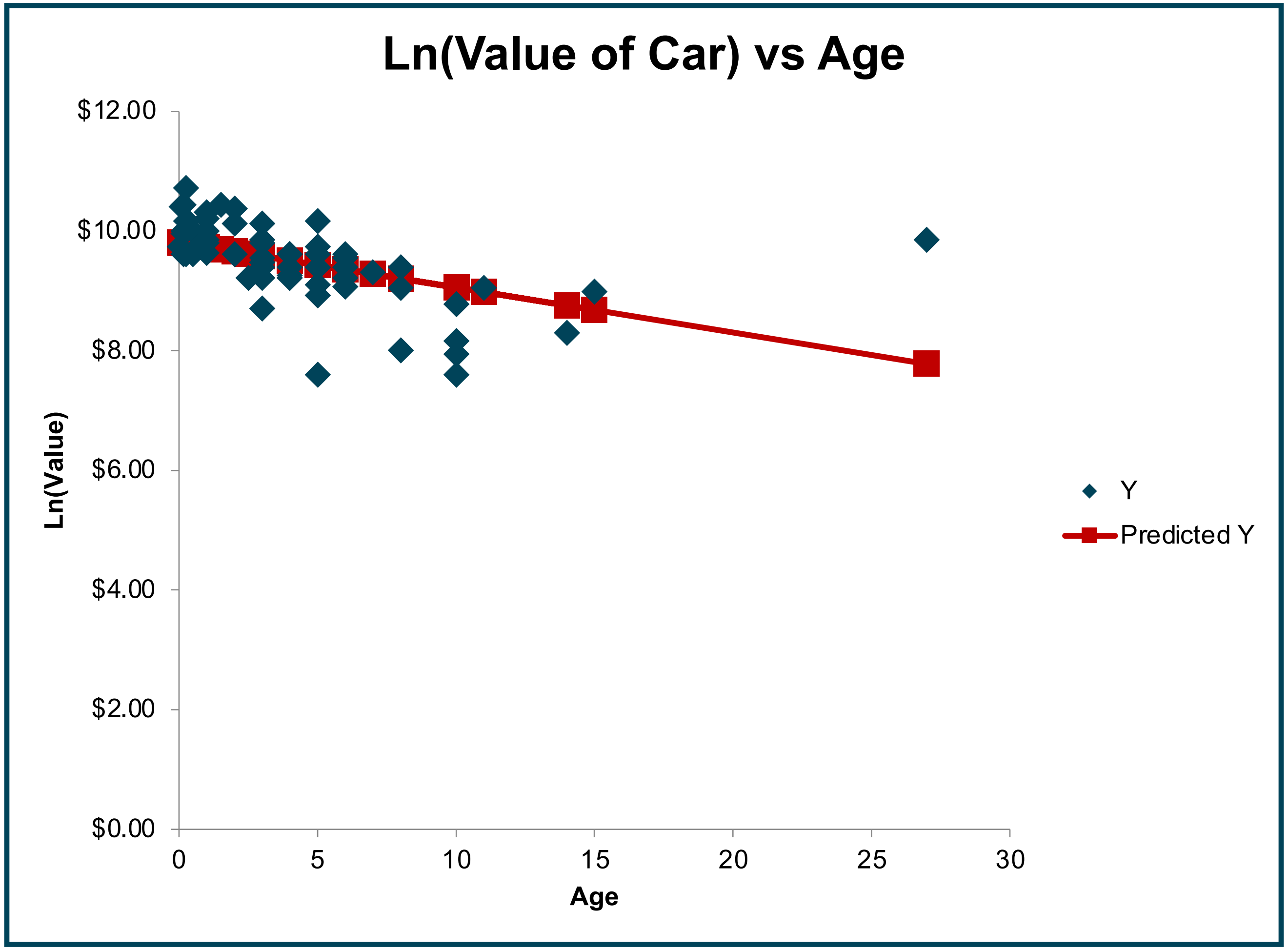
Nonlinear regression model: Ln(Value) = $b_0 + b_1 * \text{Age} + \varepsilon$

SUMMARY OUTPUT

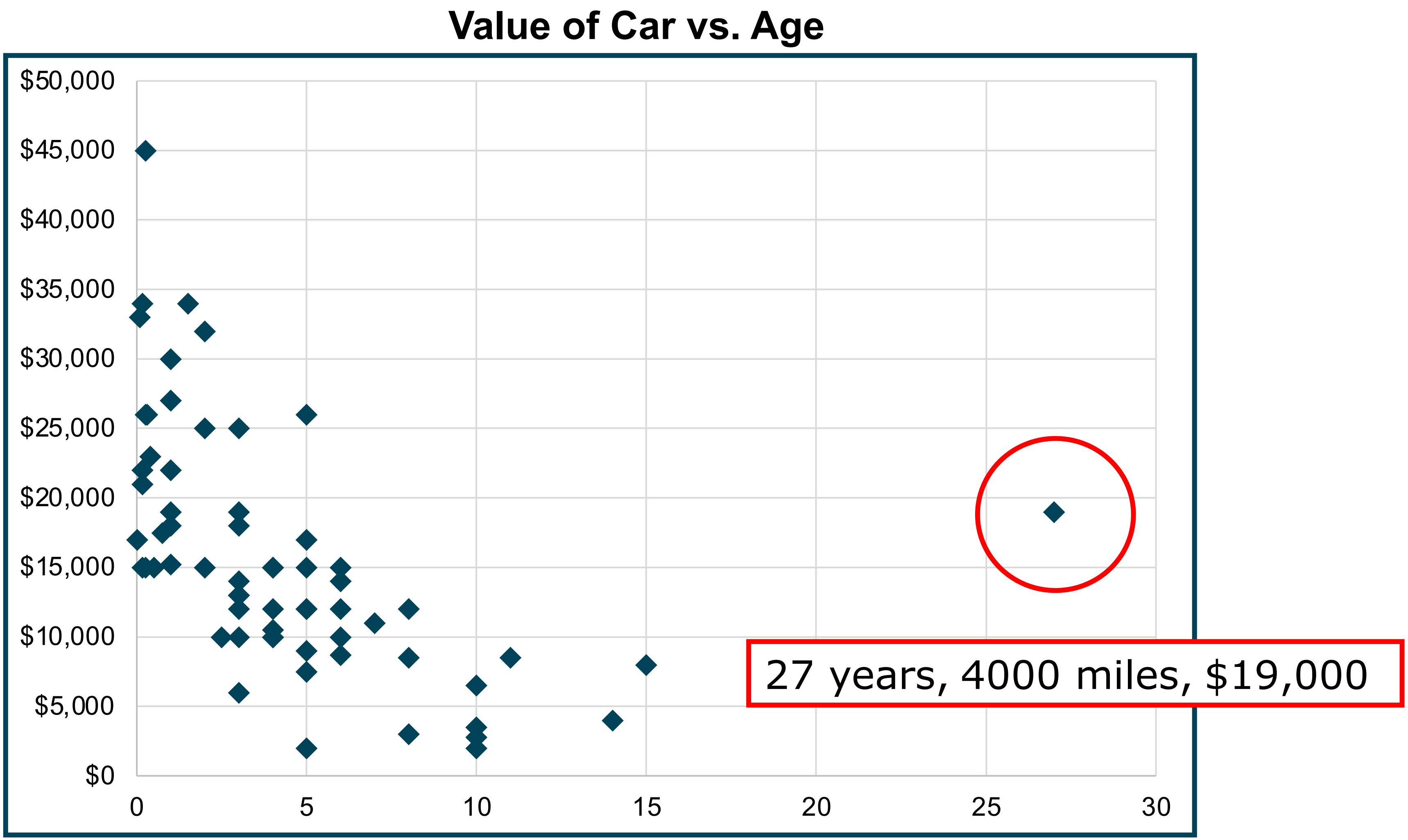
Regression Statistics	
Multiple R	0.508079632
R Square	0.258144913
Adjusted R Square	0.246179508
Standard Error	0.580212169
Observations	64

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95.0%*	*Upper 95.0%*
Intercept	9.809117654	0.101088	97.03498	1.96E-69	9.607045	10.01119
Age	-0.07517299	0.016184	-4.64481	1.82E-05	-0.10752	-0.04282

Example: Value of Second-hand Cars



Is Age enough to study the value of cars?



Mileage is also important!!
So we need to use multidimensional regression!

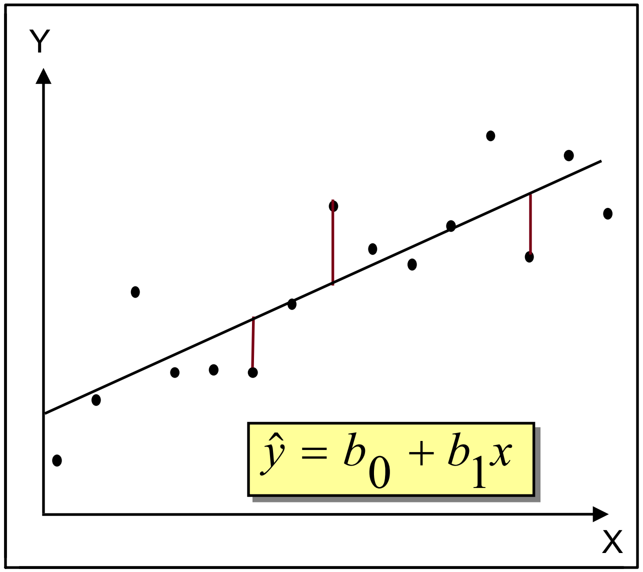
Multidimensional Regression

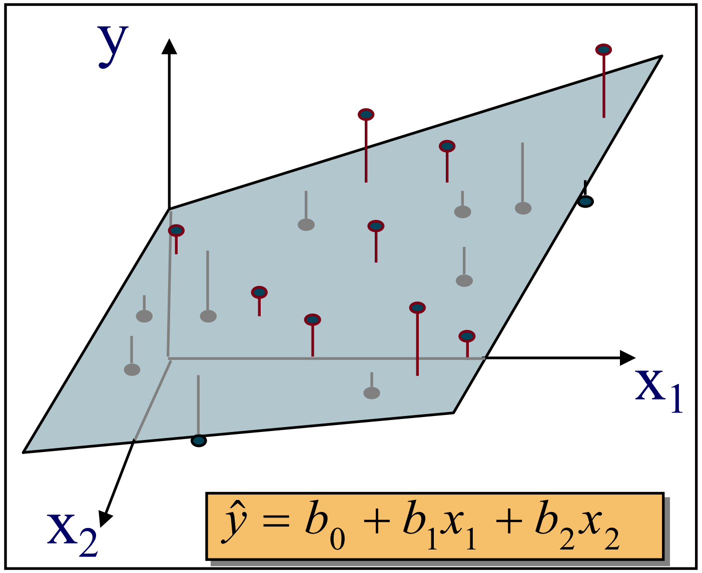
- Where $X_1,\ldots,X_p$ are p independent variables and $b_0,\ldots,b_p$ are the coefficients obtained by the Least Squares Method.

$$Y = b_0 + b_1 X_1 + \ldots + b_p X_p + \varepsilon$$

- Interpretation of b_i: The magnitude of b_i represents an estimate of the change in Y corresponding to a one unit change in X_i when all other independent variables are held constant.

Simple and Multiple Least-Squares Regression

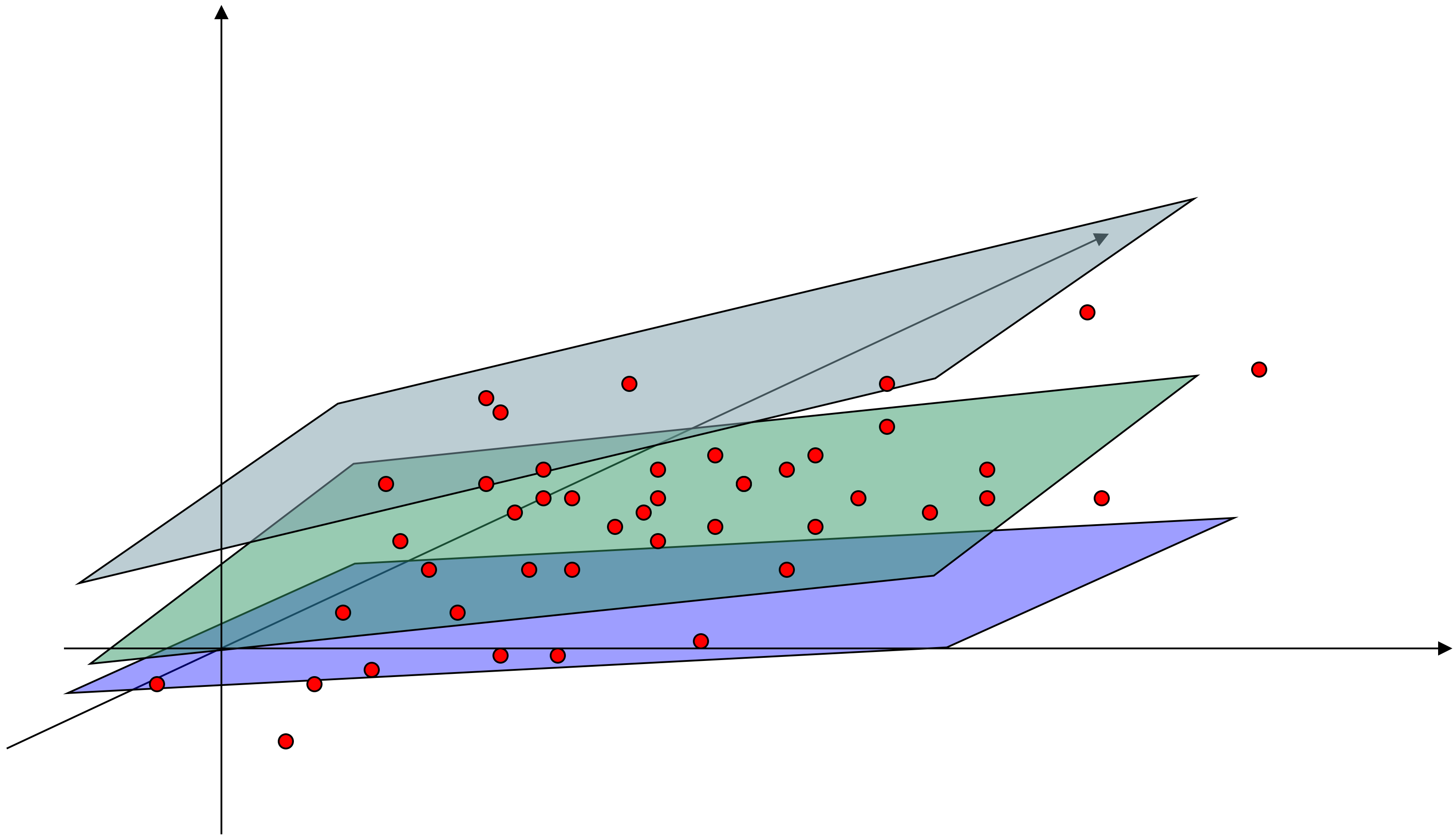




In a **simple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression line.

In a **multiple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression plane.

3 Dimensional Interpretation



Example: Value of Second-hand Cars

Multiple variable model: Ln(Value) = $b_0 + b_1 * \text{Age} + b_2 * \text{Mileage} + \varepsilon$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.791538
R Square	0.626532
Adjusted R Square	0.614288
Standard Error	0.415035
Observations	64

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	17.62747	8.813733	51.16708	8.99E-14
Residual	61	10.50749	0.172254		
Total	63	28.13496			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	10.08315	0.080479	125.2898	2.73E-75	9.922227	10.24408
Age	-0.01226	0.014135	-0.86709	0.389289	-0.04052	0.016009
Mileage	-1.2E-05	1.6E-06	-7.75695	1.15E-10	-1.6E-05	-9.2E-06

Significance Test

Rigorously test: "Do all the variables X_i that we have included in the model have an impact on Y?"

- For overall model, null Hypothesis:
  - b_1=0 AND b_2=0 AND b_3=0 …
  - If "significance F" < 0.05, then model is statistically significant.

- For individual coefficients, check the p-value, t-stats, or CI (similar to simple linear regression)

Goodness of Fit

- R^2, represents the variability in y that is explained by the estimated regression equation.

- Adjusted R^2 modifies R^2 for the number of independent variables to avoid unnecessary inclusion of additional independent variables.

Multicollinearity

- Occurs if two or more independent variables have high correlation
- Causes regression coefficients to have the “wrong” sign and the associated t-values to be low
- Can be detected by computing a correlation matrix of the independent variables
- Can be avoided by dropping one of the variables that has a high correlation with another variable.

Excel Example: Armand’s Pizza

- Download data file from Moodle: Armand’s Pizza.xlsx
- Draw scatter plot
- Run regression and interpret the results
- Plot predicted value and draw regression line.

Mini Case: 2016 Rio Olympic Game

- Download Mini Case: 2016 Rio Olympic Game and the related data file from.

Hints. 1. For scatter charts in excel, go to INSERT->Charts->Scatter

2. For regression in excel, go to DATA->Data Analysis- >Regression

3. For multiple regression in excel, include more than one column in the Input X Range. Note, however, that the regressors need to be in contiguous columns. If this is not the case in the original data, then columns need to be copied to get the regressors in contiguous columns.

Seminar-Regressions

Exercise 1

As the CEO of a company selling smart phones, you want to know whether advertising expenditure (unit: £) can influence sales (unit: £) or not. After analysing the data using simple linear regression, the result is as follows:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.83
R Square	0.70
Adjusted R Square	0.67
Standard Error	0.81
Observations	15

	Coeffts	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	22.94	0.59	39.13	0.000000	21.67	24.21
Advert	2.16	0.39	5.47	0.000108	1.31	3.01

(a) Based on the Excel output, write down your linear regression model.

(b) What is R square? What does it mean?

(c) Is the Advert coefficient significantly different from 0?

(d) Predict sales when spending 1M on advertising, and give 95% confidence interval for your predicted sales

Exercise 2

A supermarket manager is interested in the relationship between weekly sales level (y), shelf space (X_1), and the height of the shelf (X_2). The manager sampled 12 products, and ran a regression of the weekly sales with the independent variables X_1 and X_2. The results are presented below.

1. Write down the regression model
2. What proportion of the variation in weekly sales can be explained by the shelf space and the height of the shelf?
3. What's the relationship between sales and shelf space?
4. What's the relationship between sales and height of shelf?
5. Is the model overall significant?

3	*Regression Statistics*						
4	Multiple R	0.896668366					
5	R Square	0.804014159					
6	Adjusted R Squ	0.760461749					
7	Standard Error	25.57011034					
8	Observations	12					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	24140.52511	12070.26256	18.46084232	0.000653147	
13	Residual	9	5884.474886	653.8305429			
14	Total	11	30025				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	55.1826484	42.30630913	1.304359788	0.224478996	-40.52087163	150.8861684
18	x1	7.913242009	1.338400746	5.912460849	0.000225563	4.885569182	10.94091484
19	x2	0.641552511	0.273199908	2.348289631	0.043426589	0.023531384	1.259573639

Regression Statistics	
Multiple R	0.896668366
R Square	0.804014159
Adjusted R Squ	0.760461749
Standard Error	25.57011034
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	24140.52511	12070.26256	18.46084232	0.000653147
Residual	9	5884.474886	653.8305429		
Total	11	30025			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	55.1826484	42.30630913	1.304359788	0.224478996	-40.52087163	150.8861684
x1	7.913242009	1.338400746	5.912460849	0.000225563	4.885569182	10.94091484
x2	0.641552511	0.273199908	2.348289631	0.043426589	0.023531384	1.259573639

Exercise 3

The marketing manager at the Dean Dome is interested in estimating the demand function for concert t-shirts. He obtains the following data regarding the sales (in hundreds) of t-shirts at various prices:

Sales	47	42	47	31	36	58	38	35	38	32	61	46
Price	11	15	13	19	16	12	16	18	18	20	13	14

R	R Square	Adj. R Sqr	St. Err of Est	df	F	p-value
0.846	0.715	0.686	5.384	10	25.084	0.0005

Variable	Coeff.	Std. Err	t-value	p-value
Constant	85.709	8.750	9.796	0.0000
Price	-2.797	0.559	-5.008	0.0005

Exercise 3

1. Write out the regression equation relating sales as a function of price.
2. Is the coefficient for the Price (the ‘Beta’) significantly different from zero?
3. What does the Price ‘Beta’ tell us?
4. Construct a 95% prediction interval for the change in expected sales’ i.e. ‘Beta’’ when the price is increased by £1.

Thank You!

Any Questions ?