Business Analytics

Lecture 8

Nonlinear and Multiple Regression

Dr Yufei Huang

UCL
SCHOOL OF
MANAGEMENT

Review: Simple Linear Regression

1. Write the equation of the estimated line
   - Dependent Variable Y = b_0 + b_1 *(Independent Variable X) + ε

2. Is the coefficient, b_1, significant? For 95% confidence level, check,
   - p-value < 0.05?
   - t-stats > Z-value from normal distribution (or t-value from t-distribution)
   - does the 95% interval for the coefficient contain 0?

3. Point forecast for the mean and the 95% prediction interval?

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

4. How good is the fit? Look at the R^2.

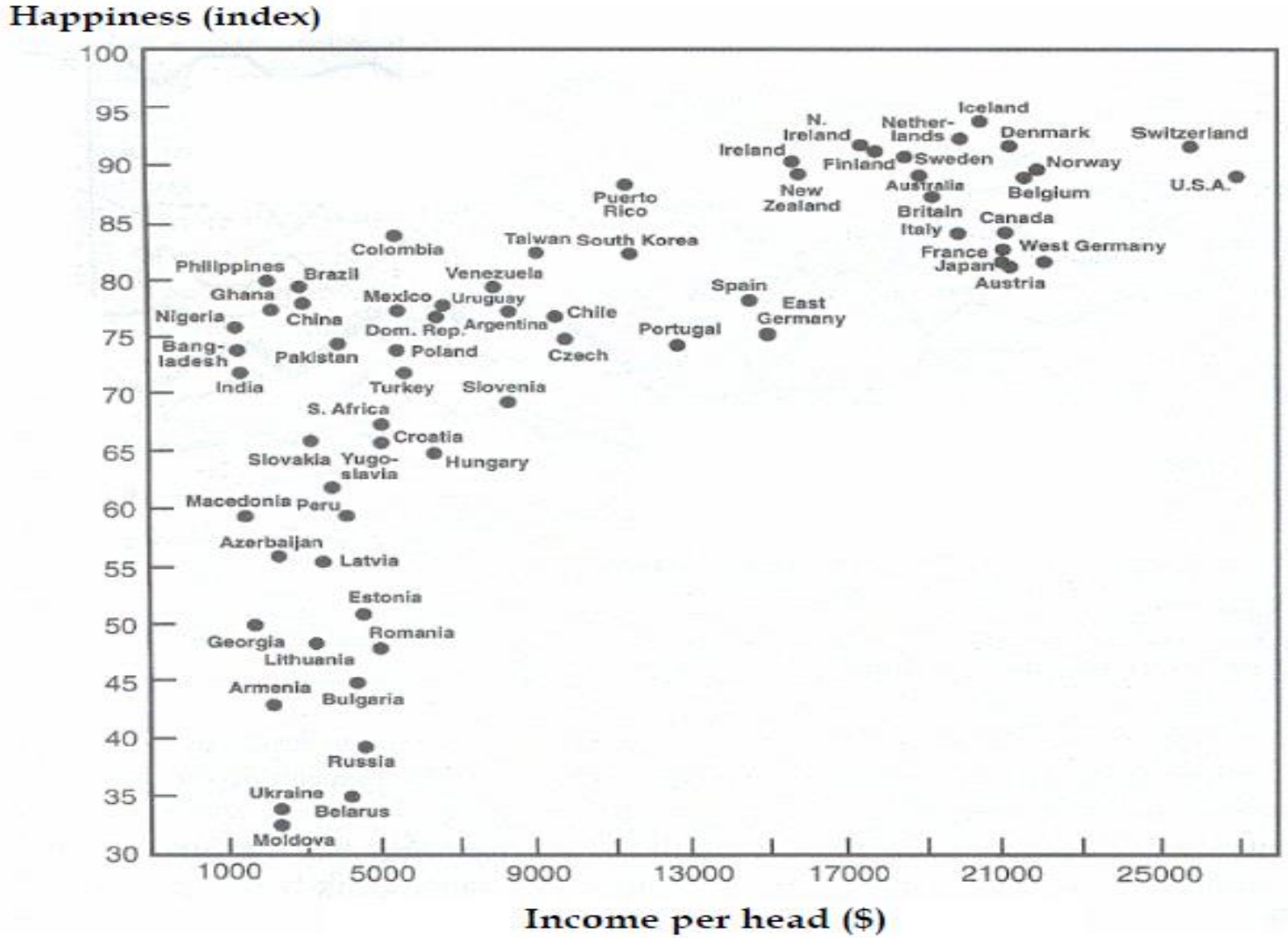
5. Degree of freedom
   - n-1 for hypothesis testing
   - n-2 for simple linear regression

Preview

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

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

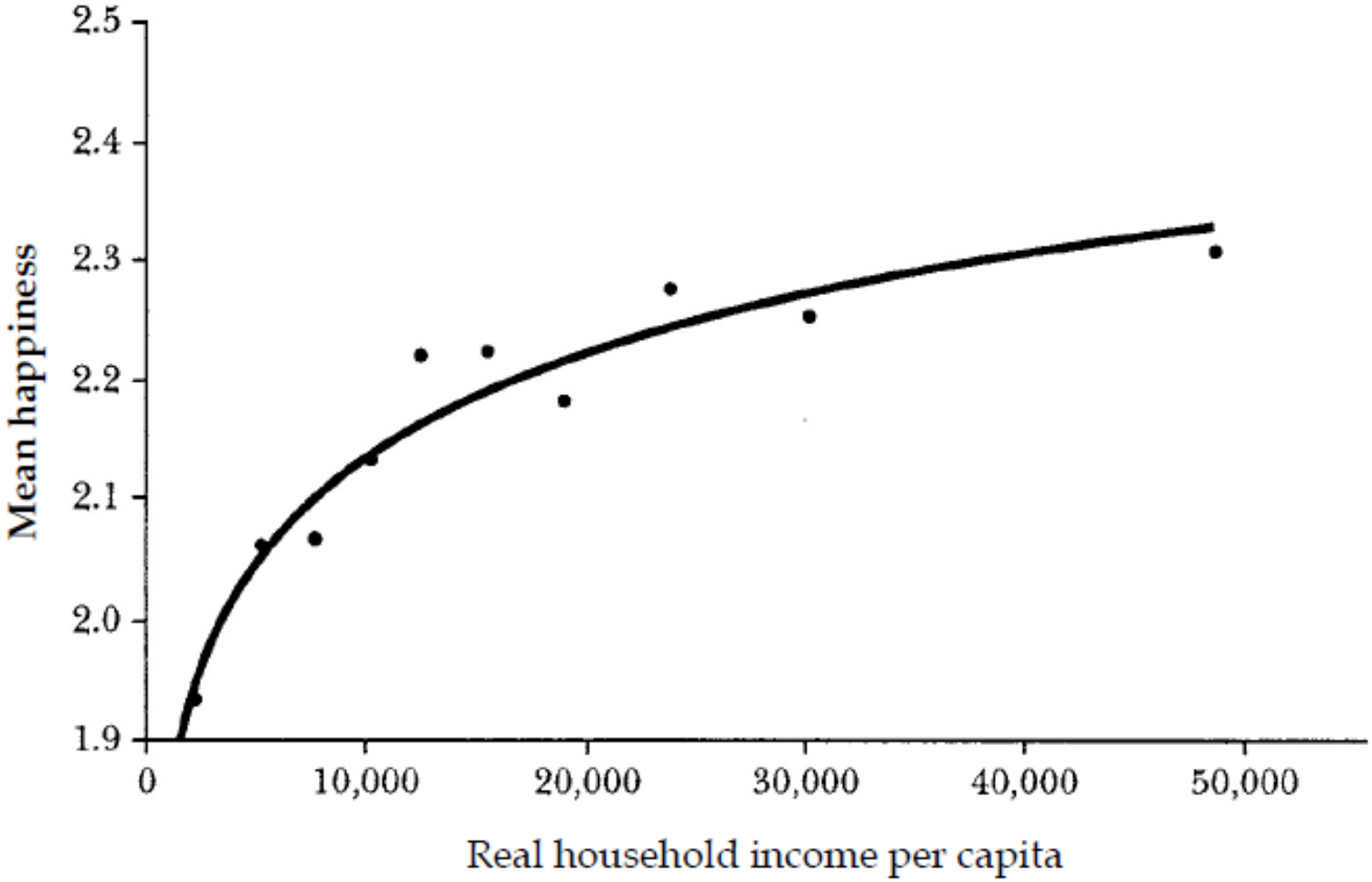
Country Comparison of Income and Happiness



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

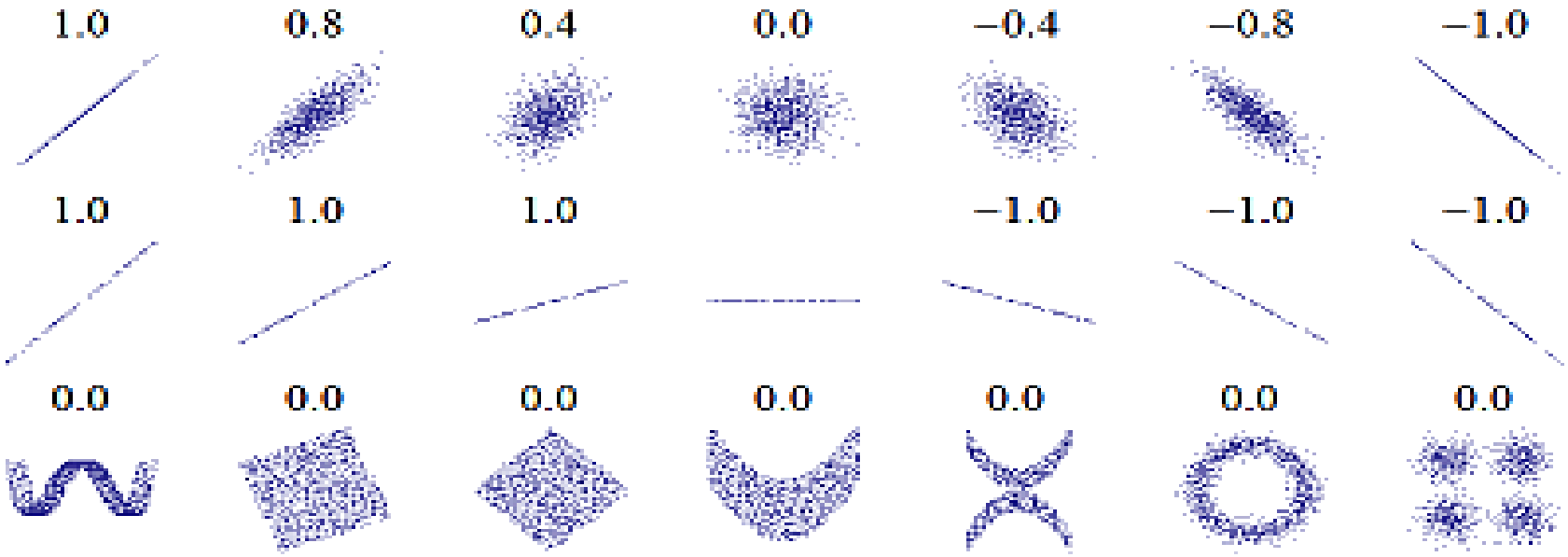
UCL SCHOOL OF MANAGEMENT

Mean Happiness and Household Income



Source: diTella and MacCullouch (2006).

In real life, data might not fit a linear model



Nonlinear Regression

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

The multiplicative model:

$$Y = \beta_0 X^{\beta_1} \varepsilon$$

The logarithmic transformation:

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

Interpreting Multiplicative Models

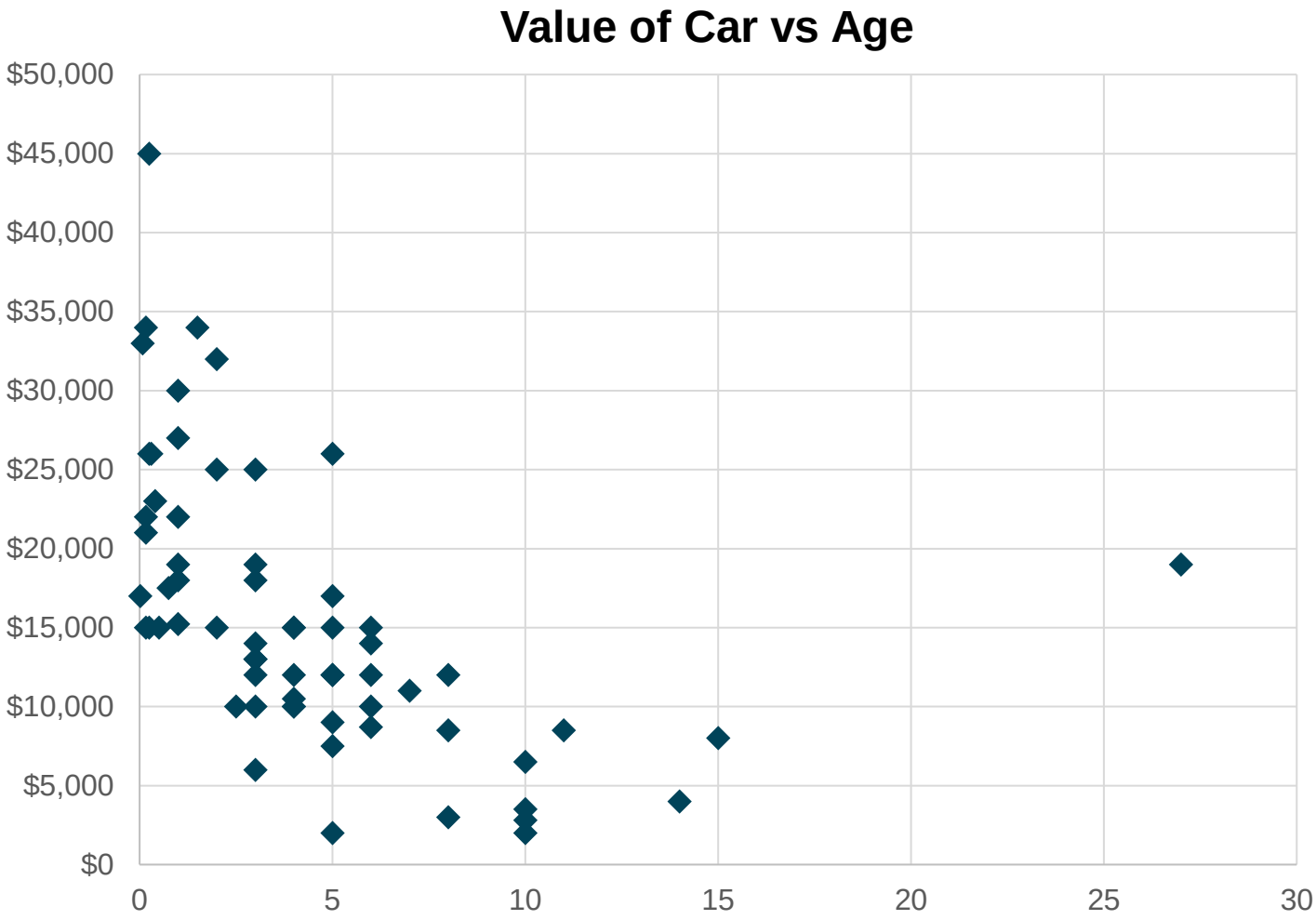
- $y = b_0 + b_1 LN(x_1) + \varepsilon$ Model (1)
  - If x_1 increases by 1%, then y increases by approximately 0.01 b_1 units.

- $LN(y) = b_0 + b_1 x_1 + \varepsilon$ 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



UCL SCHOOL OF MANAGEMENT

Example: Value of Second-hand Cars

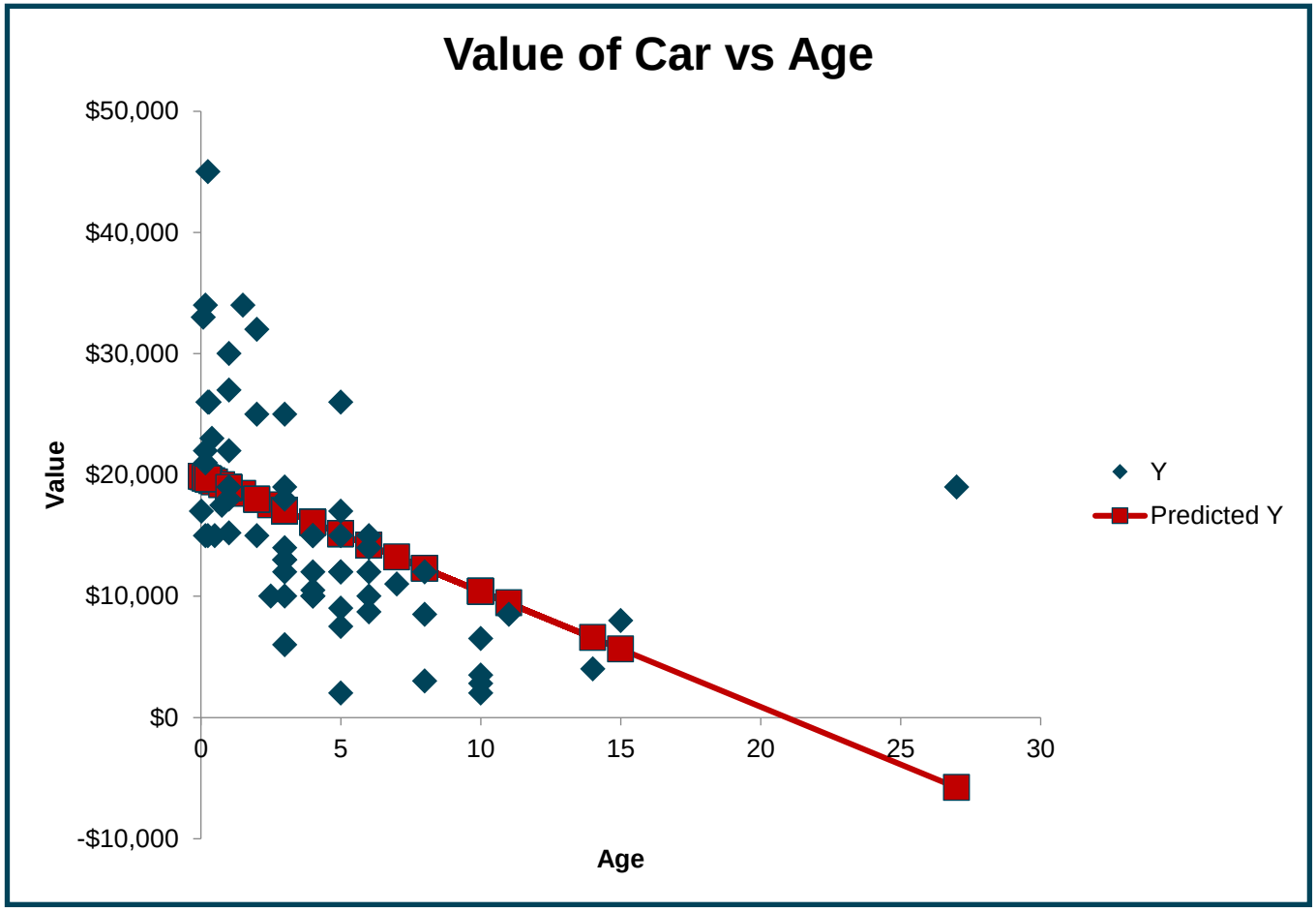
A simple linear regression model: Value = b_0 + b_1 * Age + ε

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

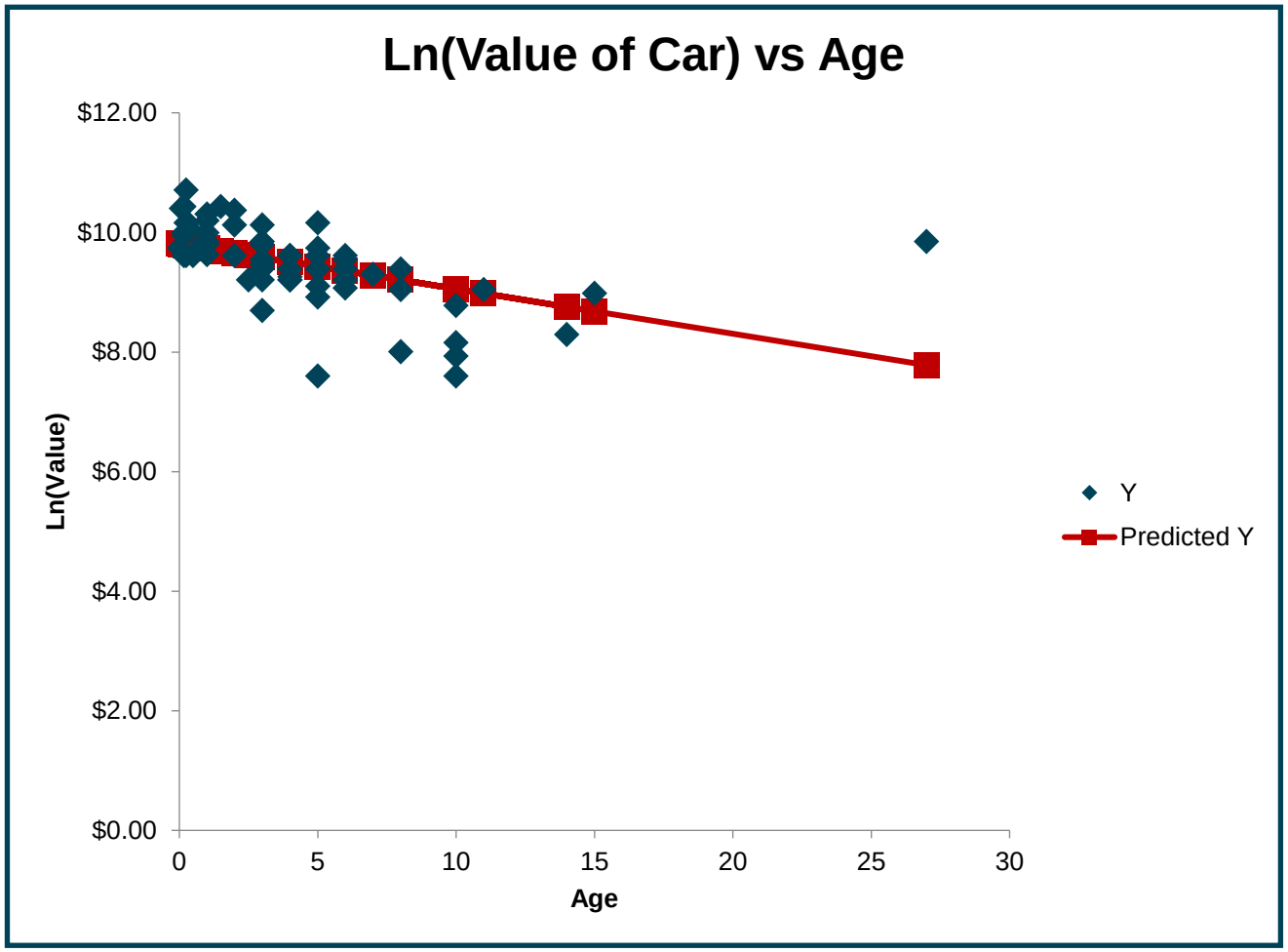
Nonlinear regression model: Ln(Value) = b_0 + b_1 * Age + ε

SUMMARY OUTPUT

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

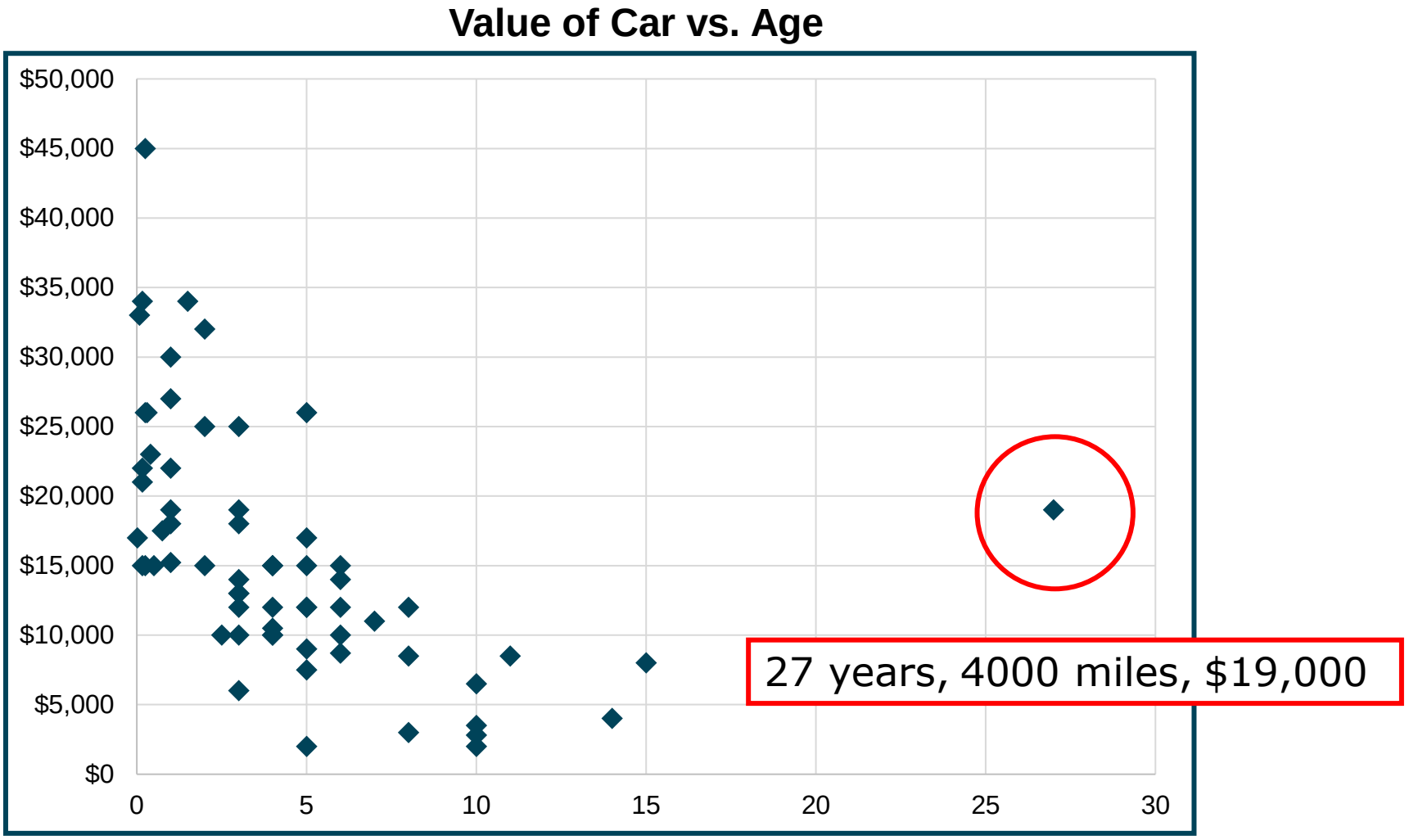
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95.0%*	*Upper 95.0%*
Intercept	9.809117654	0.101088	97.03498	1.96E-69	9.607045	10.01119
Age	-0.07517299	0.016184	-4.64481	1.82E-05	-0.10752	-0.04282

Example: Value of Second-hand Cars



UCL SCHOOL OF MANAGEMENT

Is Age enough to study the value of cars?



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

UCL SCHOOL OF MANAGEMENT

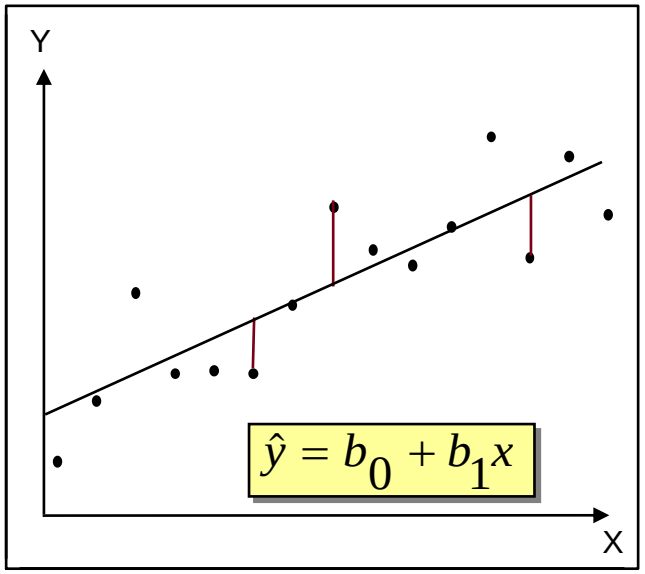
Multidimensional Regression

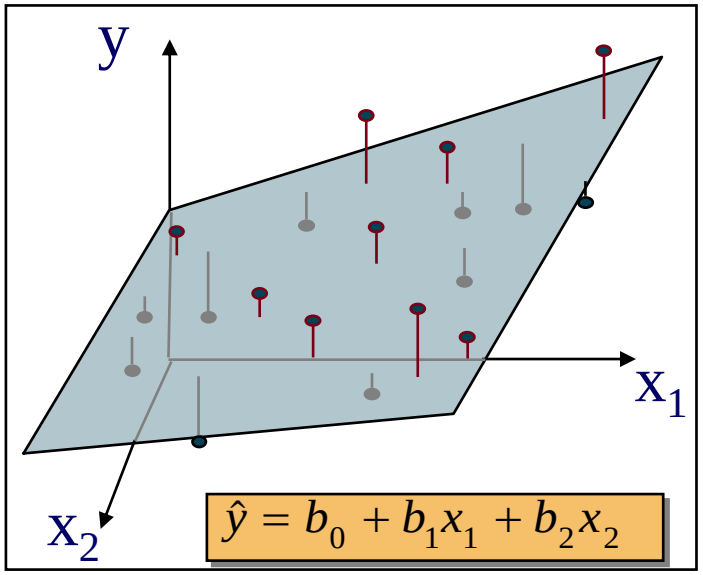
- Where $X_1,\ldots,X_p$ are p independent variables and $b_0,\ldots,b_p$ are the coefficients obtained by the Least Squares Method.

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

- 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

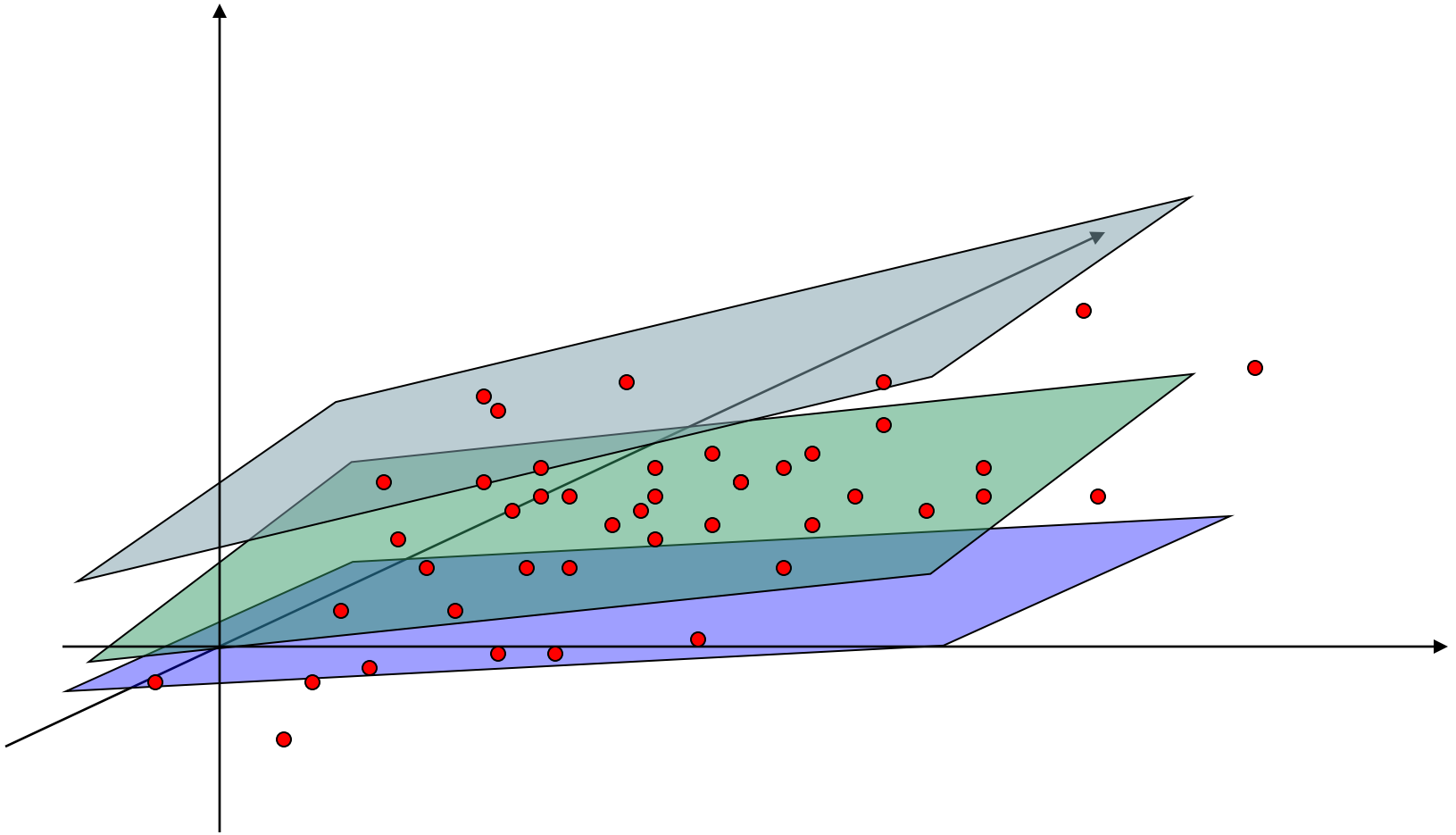




In a **simple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression line.

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 * Age + b_2 * Mileage + ε

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.

Summary

- Nonlinear Models
  - Linear regression can be applied to capture nonlinear relationships using the multiplicative models.
  - “Ln” denotes the percentage change

- Multidimensional Regression
  - You can add more independent variables to improve the understanding of different factors that influence the outcome.
  - Significance for overall model: “significance of F”
  - Goodness of fit: R^2 , more variable usually leads to higher R^2
  - But watch out for multicollinearity

- Next Lecture:
  - Revision

Mini Case: 2016 Rio Olympic Game-Revisited

- Download Mini Case: 2016 Rio Olympic Game-Revisited and the related data file from Moodle, and follow the instructions.

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.

Reference

Chapter 11 of:

Aczel, A., & J. Sounderpandian. 2008. Complete Business Statistics. McGraw-Hill/Irwin, Seventh Edition.