Business Analytics

Lecture 7

Simple Linear Regression

Dr Yufei Huang

UCL
SCHOOL OF
MANAGEMENT

Review

- Session 6: Hypothesis Testing
  - Null Hypothesis vs. Alternative Hypothesis
  - Set confidence level and significance level
  - Computing the p-value, t-stats
  - Rejecting H_0 to accept H_A requires strong statistical evidence

- Session 7 and 8: simple statistical tool for studying relationships:
  - Regression analysis

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

Introduction

- 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.

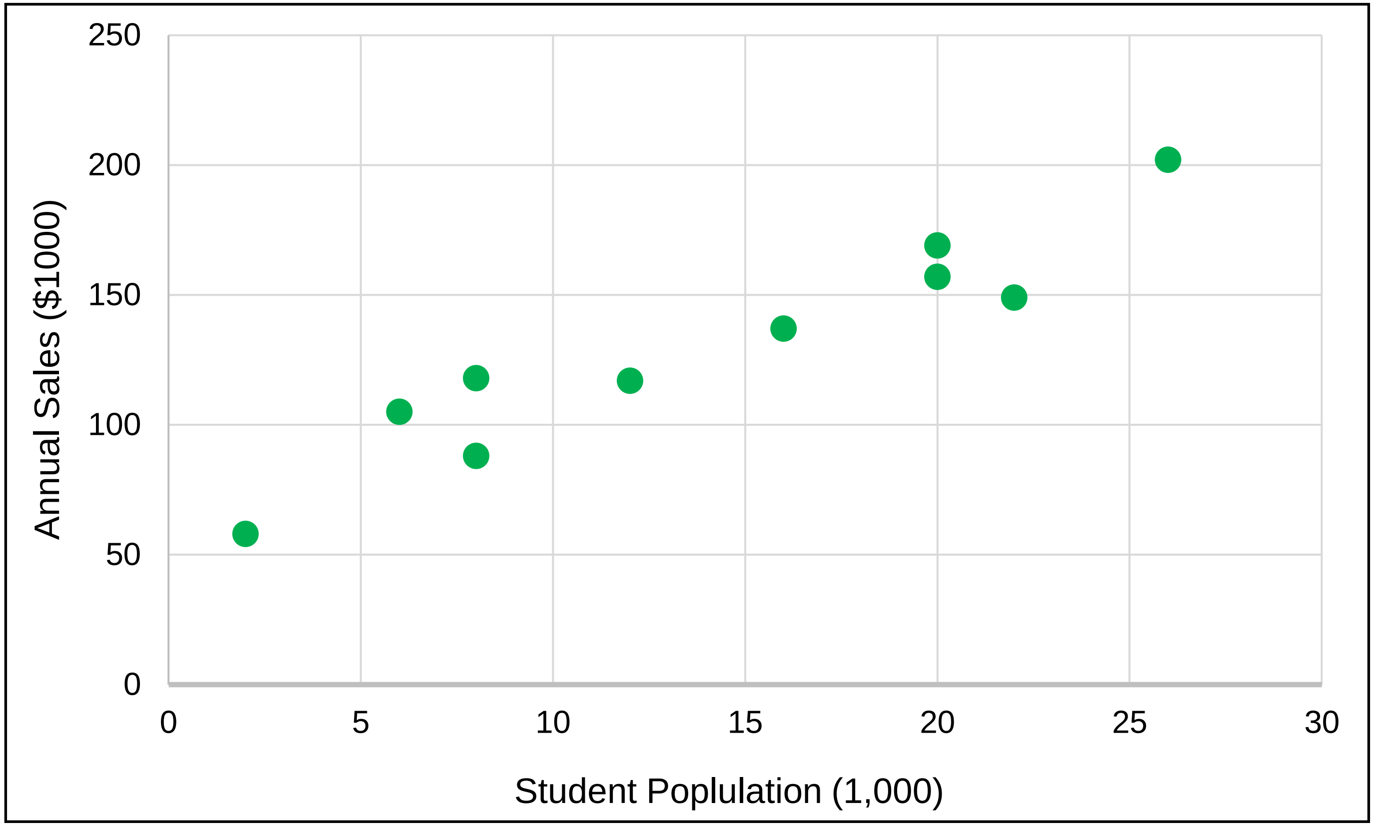
- The model we will use to depict the relationship between X and Y will be a straight-line relationship.

- 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

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.

UCL
SCHOOL OF
MANAGEMENT

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:

$$Y = b_0 + b_1 X + \varepsilon$$

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 $b_0 + b_1 X$ to a set of n points.

The estimated regression line:

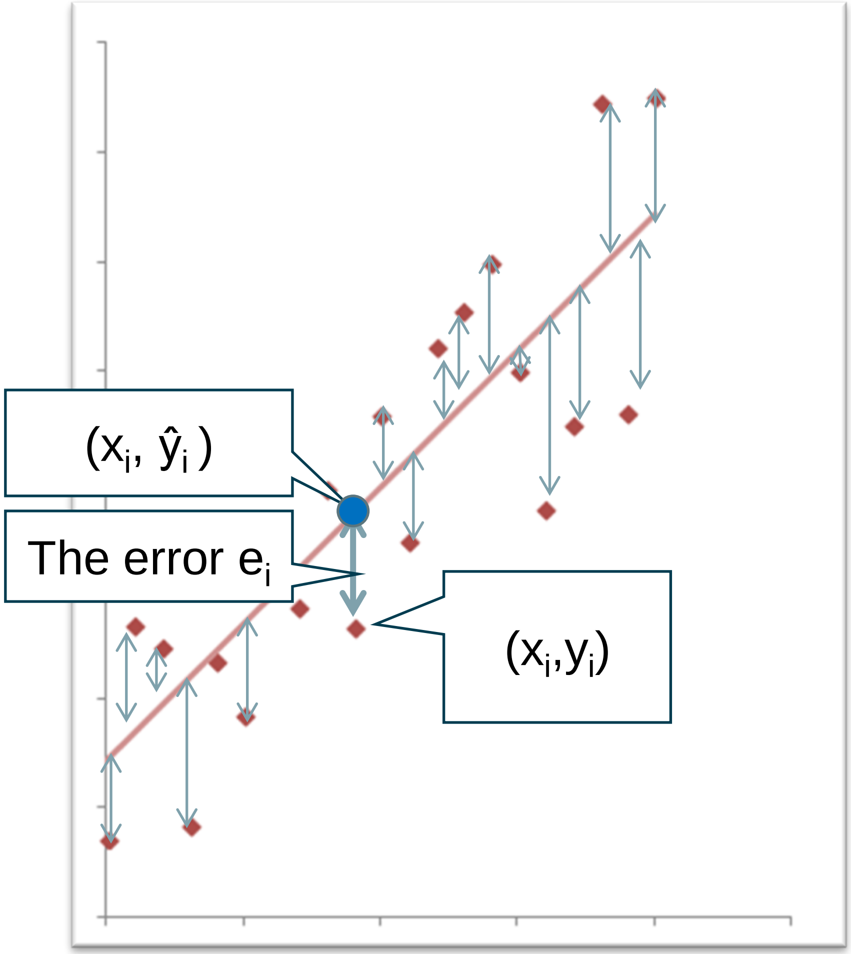
$$\hat{Y} = b_0 + b_1 X$$

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

UCL SCHOOL OF MANAGEMENT

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

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

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

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

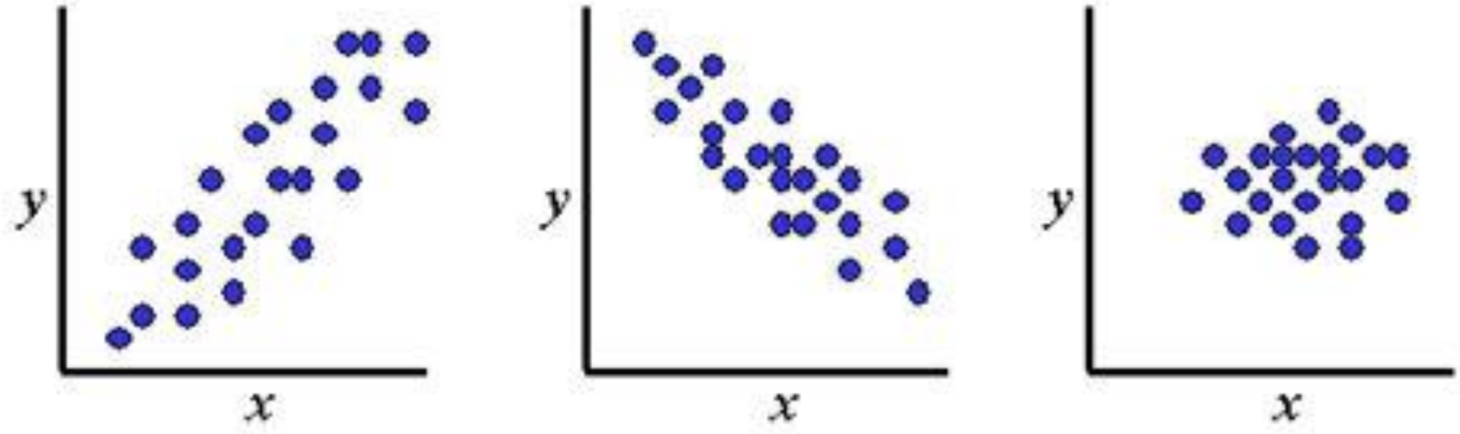
$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right).$$

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.

UCL SCHOOL OF MANAGEMENT

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

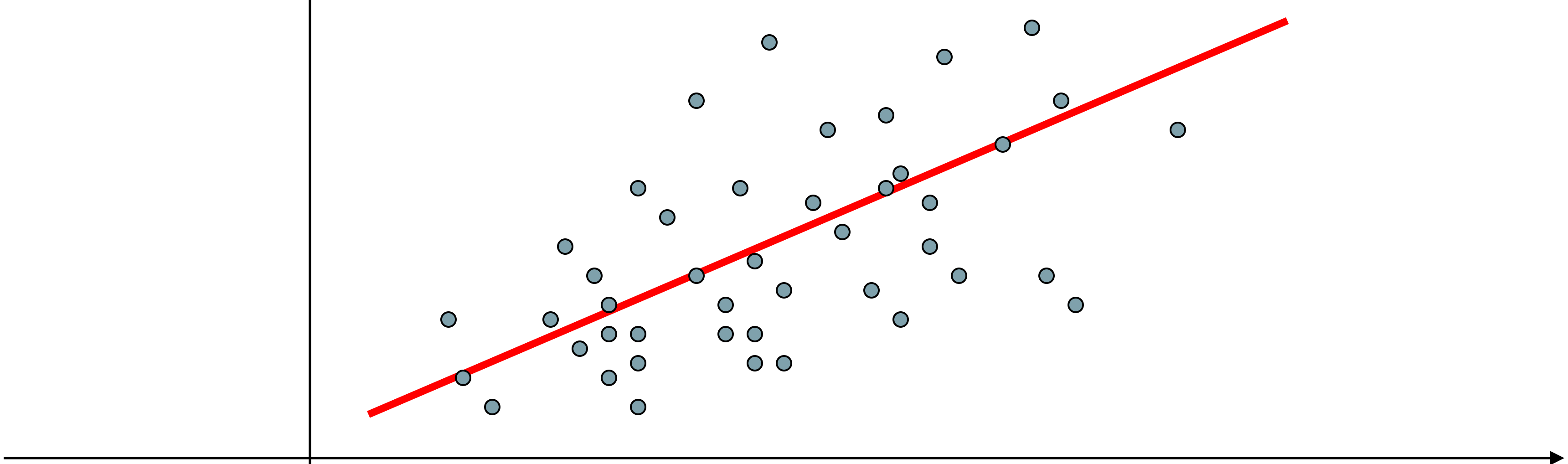
The coefficient is deemed significant at 95% confidence level:

- 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?

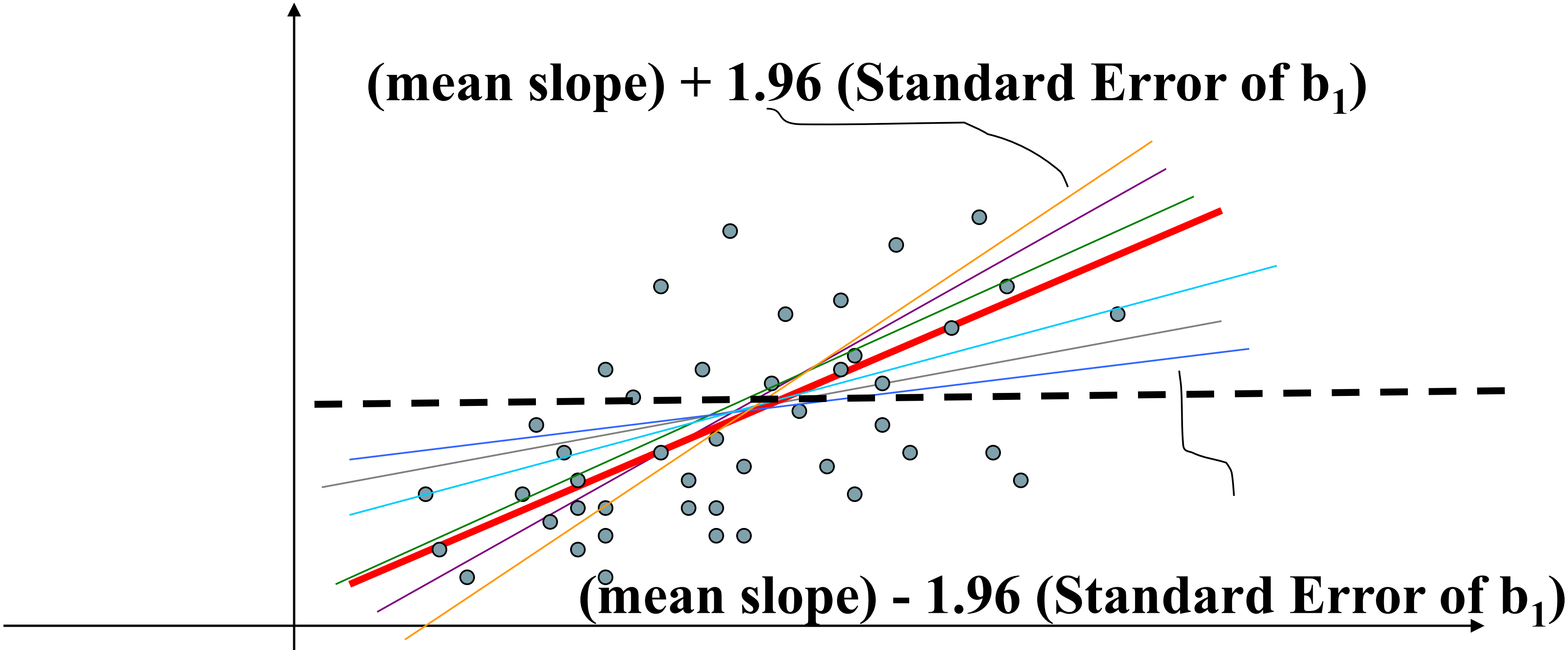
b_1 is the slope of the line.



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)

Uncertainty in Forecast

- Prediction Interval
  - With a 95% confidence level, the individual 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:

$$110 \pm 2.306 \times 13.829$$

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. ($0 \le R^2 \le 1$)

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

Summary

- Regression is useful in testing the relationship between two variables and in forecasting. Excel can generate the regression results.

- How to interpret them:

1. Write the equation of the estimated line
   - Sales = b_0 + b_1 *(student population) $+ \ \varepsilon$

2. Is the coefficient, b_1, significant? 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. What is the point forecast for the mean and the 95% prediction interval?

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

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

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

UCL SCHOOL OF MANAGEMENT

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.

Hints. 1. For scatter charts in excel, go to INSERT -> Charts -> Scatter
2. For regression in excel, go to DATA -> Data Analysis -> Regression
3. Tick "Line Fit Plots" for the fitted line in regression .

Mini Case: 2016 Rio Olympic Games

- Download Mini Case: 2016 Rio Olympic Games 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. Tick “Line Fit Plots” for the fitted line in regression .

Reference

Chapter 10 of:

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