Advanced Statistics

Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics,** 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

- **Elementary Statistics**, 14th Edition, Mario F. Triola
- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

These notes contain material from the above resources.

Explained and Unexplained Variation

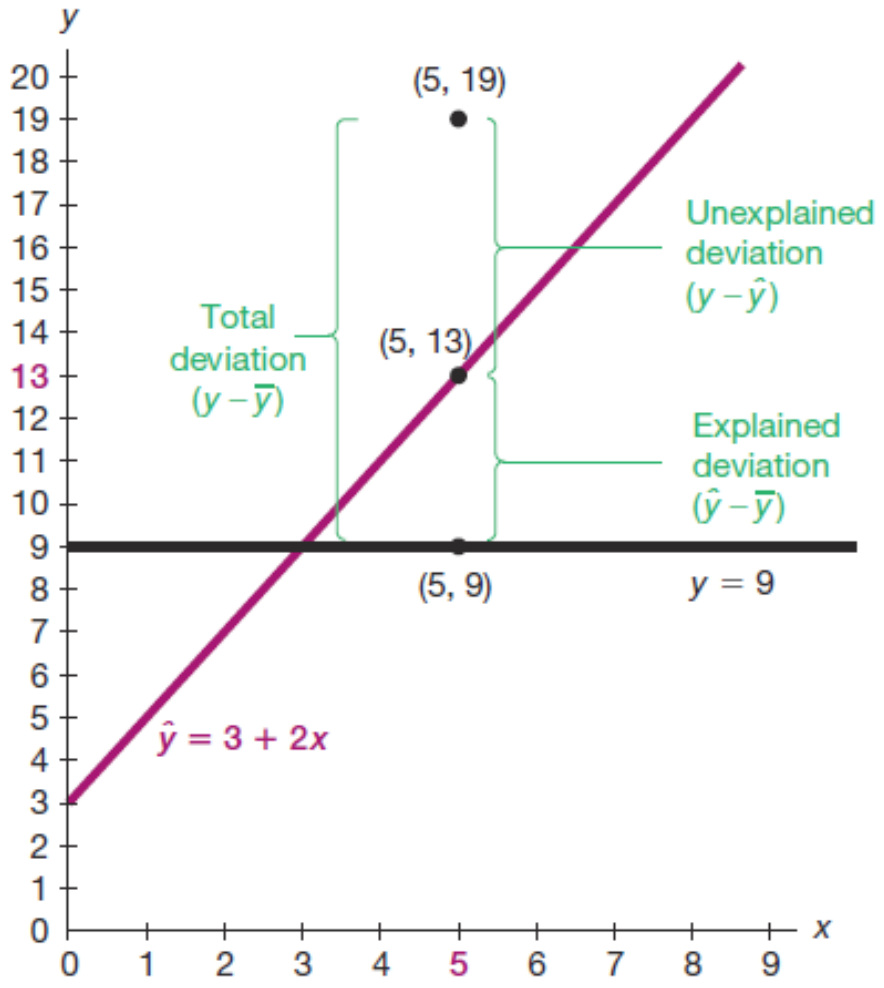

FIGURE 1 Total, Explained, and Unexplained Deviation

Assume that we have a sample of paired data having the following properties shown in Figure 1:

- There is sufficient evidence to support the claim of a **linear correlation between *x* and *y***.
- The equation of the regression line is $\hat{y} = 3 + 2x$.
- The mean of the *y* values is given by $\bar{y} = 9$.
- One of the pairs of sample data is ***x* = 5 and *y* = 19**.
- The point **(5, 13)** is one of the points on the regression line, because substituting ***x* = 5** into the regression equation of $\hat{y} = 3 + 2x$ yields $\hat{y} = 13$.

- Figure 1 shows that the point **(5, 13) lies on the regression** line, but the **point (5, 19) from the original data set does not lie on the regression line**.

- If we completely **ignore correlation and regression** concepts and want to **predict a value of *y* given a value of *x*** and a collection of paired (*x*, *y*) data, our best guess would be the **mean *y* = 9**.

- But in this case there is a **linear correlation between *x* and *y***, so a better way to predict the value of *y* when *x* = 5 is to substitute ***x* = 5 into the regression equation** to get $\hat{y}$ **= 13**.

- We can ***explain* the discrepancy** between $\bar{y}$ **= 9** and $\hat{y}$ **= 13** by noting that there is a **linear relationship best described** by the **regression line**.

- Consequently, when ***x* = 5**, the **predicted value of *y* is 13**, not the **mean value of 9**. For *x* = 5, the **predicted value of *y* is 13**, but the **observed sample value of *y*** is actually **19**.

- The discrepancy between $\hat{y}$ **= 13** and ***y* = 19** cannot be explained by the **regression line**, and it is called a ***residual* or *unexplained deviation***, which can be expressed in the general format of ***y* -** $\hat{y}$.

We have defined the standard deviation, we again consider a *deviation* to be a **difference between a value and the mean**. (In this case, the mean is $\bar{y} = 9$.) Examine Figure 1 carefully and note these specific deviations from $\bar{y} = 9$:

***Total deviation* (from $\bar{y} = 9$)** of the point (5, 19) $= y - \bar{y}$

$= 19 - 9$

$= 10$

***Explained deviation* (from $\bar{y} = 9$)** of the point (5, 19) $= \hat{y} - \bar{y}$

$= 13 - 9$

$= 4$

***Unexplained deviation* (from $\bar{y} = 9$)** of the point (5, 19) $= y - \hat{y}$

$= 19 - 13$

$= 6$

Total Deviation, Explained Deviation, and Unexplained Deviation

Assume that we have a collection of paired data containing the **sample point (*x*, *y*),** that $\hat{y}$ is the predicted value of *y* (obtained by using the regression equation), and that the **mean of the sample *y*** values is $\overline{y}$.

- The **total deviation** of (*x*, *y*) is the vertical distance ***y* - $\overline{y}$**, which is the distance between the **point (*x*, *y*) and the horizontal line passing through the sample mean $\overline{y}$**.
- The **explained deviation** is the vertical distance $\hat{y}$ - $\overline{y}$, which is the distance between the **predicted *y* value and the horizontal line passing through the sample mean $\overline{y}$**.
- The **unexplained deviation** is the vertical distance ***y* - $\hat{y}$**, which is the vertical distance between the point (*x*, *y*) and the regression line. (The distance ***y* - $\hat{y}$** is also called a *residual)*

In Figure 1 we can see the following relationship for an individual point (*x*, *y*):

(total deviation) = (explained deviation) + (unexplained deviation)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

The expression above involves deviations away from the mean, and it applies to any one particular point (*x*, *y*). If we sum the squares of deviations using all points (*x*, *y*), we get **amounts of *variation***.

(total variation) = (explained variation) + (unexplained variation)

$$\sum(y - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y - \hat{y})^2$$

Coefficient of Determination

The **coefficient of determination** is **the proportion of the variation in *y* that is explained by the regression line**. It is computed as

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

$$= \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

We can compute r^2 by using the above formula, or we can **simply square the linear correlation coefficient *r***.

Example: Jackpot/Tickets Data: Finding the Coefficient of Determination

If we use the nine pairs of jackpot/tickets data from Table 1, we find that the **linear correlation coefficient is $r = 0.947$**. Find the **coefficient of determination**. Also, find **the percentage of the total variation in *y* (tickets)** that can be explained by the linear correlation between the jackpot amount and number of tickets sold.

Solution

With r = 0.947 the coefficient of determination is r^2 = 0.897.

INTERPRETATION

Because r^2 is the **proportion of total variation** that can be explained, we conclude that **89.7% of the total variation** in tickets sold can be explained by the amount of the jackpot, and the other **10.3% cannot be explained by the jackpot**. The other **10.3%** might be explained by some other factors and/or random variation

A multiple regression equation

The following *multiple regression equation* describes linear relationships involving **more than two variables**.

A **multiple regression equation** expresses a linear relationship between a **response variable *y*** and **two or more predictor variables (x_1, x_2, . . . , x_k)**. The general form of a multiple regression equation obtained from sample data is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Finding a Multiple Regression Equation

Objective

Use sample matched data from **three or more variables** to find a multiple regression equation that is useful for predicting values of the **response variable *y***.

Notation

$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$ (multiple regression equation found from *sample* data)

$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_kx_k$
(multiple regression equation for the *population* of data)

$\hat{y}$ = **predicted value of *y*** (computed using the multiple regression equation)

k = **number of *predictor* variables** (also called *independent variables* or *x* variables)

n = **sample size** (number of values for any one of the variables)

Requirements

For any specific set of ***x* values**, the regression equation is associated with a **random error often denoted by e**. We assume that such errors are **normally distributed** with a **mean** of **0** and a **standard deviation** of σ and that the random errors are independent.

Data Set 1: Body Data

Body and exam measurements are from 300 subjects (first five rows shown here). **AGE** is in years, for **GENDER** 1 = male and 0 = female, **PULSE** is pulse rate (beats per minute), **SYSTOLIC** is systolic blood pressure (mm Hg), **DIASTOLIC** is diastolic blood pressure (mm Hg), **HDL** is HDL cholesterol (mg/dL), **LDL** is LDL cholesterol (mg/dL), **WHITE** is white blood cell count

AGE	GENDER (1 = M)	PULSE	SYSTOLIC	DIASTOLIC	HDL	LDL	WHITE	RED	PLATE	WEIGHT	HEIGHT	WAIST	ARM CIRC	BMI
43	0	80	100	70	73	68	8.7	4.80	319	98.6	172.0	120.4	40.7	33.3
57	1	84	112	70	35	116	4.9	4.73	187	96.9	186.0	107.8	37.0	28.0
38	0	94	134	94	36	223	6.9	4.47	297	108.2	154.4	120.3	44.3	45.4
80	1	74	126	64	37	83	7.5	4.32	170	73.1	160.5	97.2	30.3	28.4
34	1	50	114	68	50	104	6.1	4.95	140	83.1	179.0	95.1	34.0	25.9

EXAMPLE Predicting Weight

Data Set 1 “Body Data” in previous slide includes heights (cm), waist circumferences (cm), and weights (kg) from a sample of 153 males. Find the multiple regression equation in which the **response variable (*y*)** is the **weight of a male** and the **predictor variables** are **height (x_1)** and **waist circumference (x_2).**

Using Technology with the sample data in Data Set 1. The coefficients b_0, b_1, and b_2 are used in the multiple regression equation:

$$\hat{y} = -149 + 0.769x_1 + 1.01x_2$$

or

Weight = -149 + 0.769Height + 1.01Waist

The obvious advantage of the second format above is that it is easier to keep track of the roles that the variables play.

Finding a Multiple Regression Equation

A researcher wants to determine how **employee salaries** at a company are related to the **length of employment, previous experience, and education**. The researcher selects eight employees from the company and obtains the data shown in the table.

Employee	Salary (in dollars), y	Employment (in years), x_1	Experience (in years), x_2	Education (in years), x_3
A	57,310	10	2	16
B	57,380	5	6	16
C	54,135	3	1	12
D	56,985	6	5	14
E	58,715	8	8	16
F	60,620	20	0	12
G	59,200	8	4	18
H	60,320	14	6	17

Use Technology to find a multiple regression equation that models the data.

The multiple regression equation is

$\hat{y} = 49764 + 364x_1 + 228x_2 + 267x_2$

Or

Salary = $49764 + 364\boldsymbol{Employment} + 228\boldsymbol{Expereince} + 267\boldsymbol{Education}$

EXAMPLE Predicting *y*-Values Using Multiple Regression Equations

Use the regression equation found in the previous example to predict an employee's salary for these conditions.

1. 12 years of current employment, **5 years of previous experience**, and **16 years of education**

2. 4 years of current employment, **2 years of previous experience**, and **12 years of education**

3. 8 years of current employment, **7 years of previous experience**, and **17 years of education**

Solution

To predict each employee's salary, substitute the values for x_1, x_2, and x_3 into the regression equation. Then calculate $\hat{y}$.

1. $\hat{y} = 49764 + 364x_1 + 228x_2 + 267x_3$

= 49,764 + 3641122 + 228152 + 2671162

= **59,544**

The employee's predicted salary is **$59,544**.

2. $\hat{y} = 49764 + 364x_1 + 228x_2 + 267x_3$

= 49,764 + 364142 + 228122 + 2671122

= **54,880**

The employee's predicted salary is **$54,880**.

3. $\hat{y} = 49764 + 364x_1 + 228x_2 + 267x_3$

= 49,764 + 364182 + 228172 + 2671172

= **58,811**

The employee's predicted salary is **$58,811**.

Example A statistics professor wants to determine how **students' final grades** are related to the **midterm exam grades** and **number of classes missed**. The professor selects 10 students and obtains the data shown in the table.

Student	Final grade, y	Midterm exam, x_1	Classes missed, x_2
1	81	75	1
2	90	80	0
3	86	91	2
4	76	80	3
5	51	62	6
6	75	90	4
7	44	60	7
8	81	82	2
9	94	88	0
10	93	96	1

Use technology to find a multiple regression equation that models the data.
Calculate the regression line

$\hat{y} = 46.385 + 0.540x_1 - 4.897x_2$

Or

$Final\ grade = 46.385 + 0.54\text{MidtermExam} - 4.897\ \text{ClassMissed}$

Example: Use the regression equation found in previous example to predict a **student's final grade** for these conditions.

1. A student has a **midterm exam score of 89** and misses **1 class**.

2. A student has a **midterm exam score of 78** and **misses 3 classes**.

3. A student has a **midterm exam score of 83** and **misses 2 classes**.

a. Substitute the midterm score for x_1 into the regression equation.

b. Substitute the corresponding number of missed classes for x_2 into the regression equation.

c. Calculate $\hat{y}$.

d. What is each student's final grade?

Solution

$\hat{y} = 46.385 + 0.540x_1 - 4.897x_2$

or

$Final\ grade = 46.385 + 0.540\ \mathrm{MidtermExam} - 4.897\ \mathrm{ClassMissed}$

(1) $\hat{y} = 46.385 + 0.540 \times (89) - 4.897 \times (1)$

(2) $\hat{y} = 46.385 + 0.540 \times (78) - 4.897 \times (3)$

(3) $\hat{y} = 46.385 + 0.540 \times (83) - 4.897 \times (2)$

c. (1) $\hat{y} = 89.548$ **(2)** $\hat{y} = 73.814$ **(3)** $\hat{y} = 81.411$

d. (1) 90 **(2)** 74 **(3)** 81

R^2

R^2 denotes the **multiple coefficient of determination**, which is a measure of **how well the multiple regression equation** fits the sample data.

- **A perfect fit** would result in **$R^2 = 1$**.
- **A very good** fit results in a **value near 1**.
- **A very poor fit** results in a value of **R^2 close to 0.**

Adjusted R^2

- However, the **multiple coefficient of determination R^2** has a **serious flaw: As more variables are included, R^2 increases.** (R^2 could remain the same, but it usually increases.)

- The **largest R^2** is obtained by simply **including *all* of the available variables**, but the **best multiple regression equation does not necessarily use all of the available variables**.

- Because of that flaw, it is better to use **the *adjusted coefficient of determination*,** which is **R^2 adjusted** for the number of variables and the sample size.

Adjusted R^2

DEFINITION

The **adjusted coefficient of determination** is the **multiple coefficient of determination R^2** modified to account for the number of variables and the sample size. It is calculated by using Formula

$$\text{Adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

n = sample size

k = number of predictor (x) variables

Finding the Best Multiple Regression Equation

- When trying to find the **best multiple regression equation**, we should **not necessarily include all of the available predictor variables**.

- Finding **the best multiple regression equation requires abundant use of judgment and common sense**, and there is **no exact and automatic procedure that can be used to find the best multiple regression equation**.

- ***Determination of the best multiple regression equation is often quite difficult*** *and is beyond the scope of this section,* but the following guidelines are helpful.

Guidelines for Finding the Best Multiple Regression Equation

1. Use **common sense** and **practical considerations** to **include or exclude variables**.

For example, when trying to find a good multiple regression equation for predicting the height of a daughter, we should exclude the height of the physician who delivered the daughter, because that height is obviously irrelevant.

2. Consider **the P-value**. Select an equation having overall significance, as **determined by a low P-value**.

Guidelines for Finding the Best Multiple Regression Equation

3. Consider equations with high values of **adjusted R^2**, and try to include **only a few variables**. Instead of including **almost every available variable**, try to include **relatively few predictor (x) variables**. Use these guidelines:

- Select an equation having a value of adjusted R^2 with this property: If **an additional predictor variable is included**, the value of **adjusted R^2 does not increase very much**.

- For a particular number of predictor (x) variables, select the equation with the **largest value of adjusted R^2**.

Guidelines for Finding the Best Multiple Regression Equation

- In **excluding predictor (x) variables** that don't have much of an effect on the **response (y) variable,** it might be helpful to find the **linear correlation coefficient r for each pair of variables** being considered.

- If **two predictor values have a very high linear correlation coefficient (called multicollinearity)**, there is **no need to include them both**, and we should exclude the variable with the lower value of **adjusted R^2**.

Data Set : Foot and Height

Foot and height measurements are from 40 subjects (first five rows shown here). **SEX** is gender of subject, **AGE** is age in years, **FOOT LENGTH** is length of foot (cm), **SHOE PRINT** is length of shoe (cm), **SHOE SIZE** is reported shoe size, and **HEIGHT** is height (cm) of the subject.

SEX	AGE	FOOT LENGTH	SHOE PRINT	SHOE SIZE	HEIGHT
M	67	27.8	31.3	11.0	180.3
M	47	25.7	29.7	9.0	175.3
M	41	26.7	31.3	11.0	184.8
M	42	25.9	31.8	10.0	177.8
M	48	26.4	31.4	10.0	182.3

EXAMPLE Predicting Height from Footprint Evidence

Data Set "Foot and Height" includes the **age, foot length, shoe print length, shoe size**, and **height** for each of 40 different subjects. Using those sample data, find the **regression equation** that is best for **predicting height**. Is the **"best" regression equation** a *good* equation for predicting height?

Solution

- Using the response variable of height and possible predictor variables of **age, foot length, shoe print length, and shoe size**, there are 15 different possible combinations of predictor variables.

- Table 1 includes key results from five of those combinations.

- **Blind and thoughtless application of regression methods** would suggest that the **best regression equation** uses all **four of the predictor variables**, because that combination yields the highest **adjusted R^2 value of 0.7585**. However, given the objective of using evidence to estimate **the height of a suspect**, we use *critical thinking* as follows.

1. **Delete** the **variable of age**, because **criminals rarely leave evidence identifying their ages.**

2. **Delete** the **variable of shoe size**, because it is really a **rounded form of foot length**.

3. For the remaining variables of **foot length and shoe print** length, use only **foot length** because its **adjusted R^2 value of 0.7014** is greater than 0.6520 for **shoe print length**, and it is not very much less than the adjusted R^2 value of **0.7484 for both foot length and shoe print length**. In this case, it is better to use **one predictor variable instead of two**.

4. Although it appears that the use of the **single variable of foot length is best**, we also note that **criminals usually wear shoes**, so **shoe print lengths** are more likely to be found than **foot lengths**.

TABLE 1 Select Key Results from Data Set 9 "Foot and Height"

Predictor Variables	Adjusted R^2	*P*-Value	
Age	0.1772	0.004	← **Not best:** Adjusted R^2 is far less than 0.7014 for Foot Length.
Foot Length	**0.7014**	**0.000**	← **Best:** High adjusted R^2 and lowest *P*-value.
Shoe Print Length	0.6520	0.000	← **Not best:** Adjusted R^2 is less than 0.7014 for Foot Length.
Foot Length/Shoe Print Length	0.7484	0.000	← **Not best:** The adjusted R^2 value is not very much higher than 0.7014 for the single variable of Foot Length.
Age/Foot Length/ Shoe Print Length/ Shoe Size	0.7585	0.000	← **Not best:** There are other cases using fewer variables with adjusted R^2 that are not too much smaller.

INTERPRETATION

Blind use of regression methods suggests that when estimating the height of a subject, we should use all of the available data by including **all four predictor variables of age, foot length, shoe print length, and shoe size**, but **practical considerations suggest that it is best to use the single predictor variable of foot length**. So the best regression equation appears to be:

Height = 64.1 + 4.29 (Foot Length).

INTERPRETATION Cont.

- However, given that **criminals usually wear shoes**, it is best to use the **single predictor variable of shoe print length**, so the best practical regression equation appears to be this: **Height = 80.9 + 3.22 (Shoe Print Length)**.
- The ***P*-value of 0.000** suggests that the **regression equation yields a good model for estimating height.** Because the results of this example are based on sample data from only 40 subjects, estimates of heights will not be very accurate. As is usually the case, better results could be obtained by using larger samples.