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

These notes contain material from the above resources.

Definitions

Given a collection of paired sample data, the **regression equation**

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the relationship **between the two variables**. The graph of the **regression equation** is called the **regression line** (or ***line of best fit***, or ***least-squares line***).

Notation for Regression Equation

	Population Parameter	Sample Statistic
y-intercept of regression equation	β_0	b_0
Slope of regression equation	β_1	b_1
Equation of the regression line	$Y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Finding the slope b_1 and y-intercept b_0 in the regression equation $\hat{y} = b_0 + b_1 x$

Slope	$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$
***y*-intercept:**	$b_0 = \bar{y} - b_1\bar{x}$ **or** $b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$

Finding the slope b_1 and *y*-intercept b_0 in the regression equation $\hat{y} = b_0 + b_1 x$

Slope: $b_1 = r\frac{s_y}{s_x}$

where ***r*** is the linear correlation coefficient, **s_y is the standard deviation** of the *y* values, and **s_x is the standard deviation** of the *x* values.

***y*-intercept: $b_0 = \bar{y} - b_1\bar{x}$**

Example: Table 1 is reproduced here. (Jackpot amounts are in millions of dollars and numbers of tickets sold are in millions.) Find the equation of the **regression line in which the explanatory variable (or *x* variable)** is the amount of the lottery jackpot and the response variable (or *y* variable) is the corresponding number of lottery tickets sold.

Table 1 Powerball Tickets Sold and Jackpot Amounts

Jackpot	334	127	300	227	202	180	164	145	255
Tickets	54	16	41	27	23	18	18	16	26

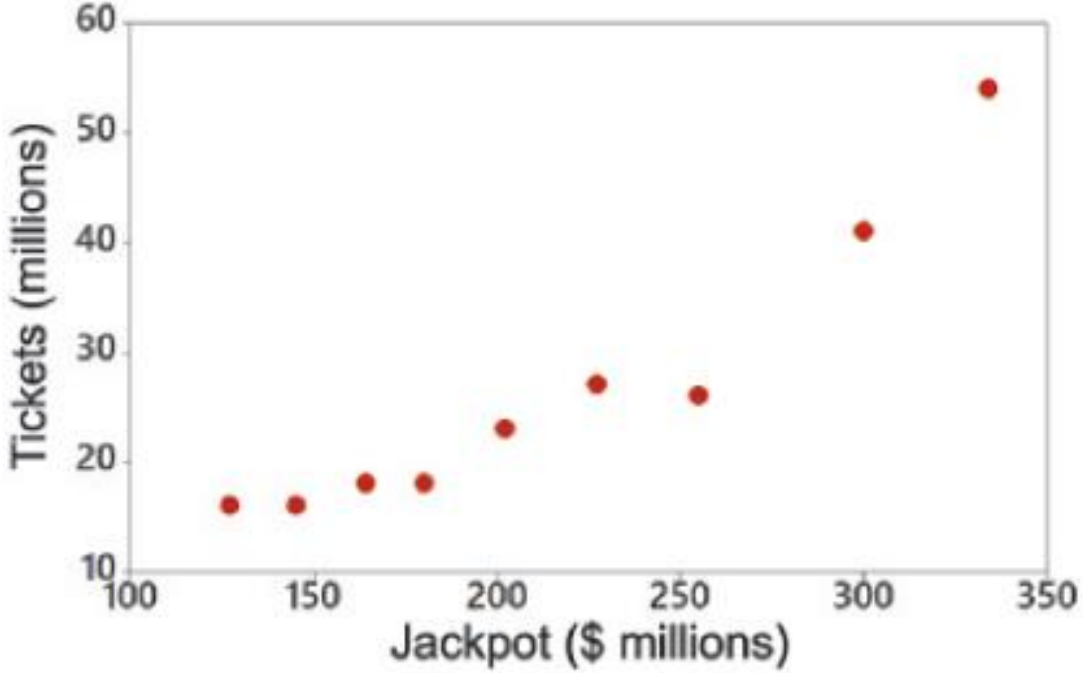

FIGURE 1 **Scatterplot from Table 1**

1. The data are a simple random sample.

2. The scatterplot in Figure 1 on previous slide shows that the pattern of points is reasonably close to a **straight-line pattern**.

3. The scatterplot also shows that there are **no outliers**.

   The requirements are satisfied.

x(Jackpot)	y(Tickets)	x^2	y^2	xy
334	54	111,556	2916	18,036
127	16	16,129	256	2032
300	41	90,000	1681	12,300
227	27	51,529	729	6129
202	23	40,804	529	4646
180	18	32,400	324	3240
164	18	26,896	324	2952
145	16	21,025	256	2320
255	26	65,025	676	6630
$\sum x$ = 1934	$\sum y$ = 239	$\sum x^2$= 455,364	$\sum y^2$= 7691	$\sum xy$ = 58,285

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{9(158{,}2852) - (1934)(239)}{\sqrt{9(455{,}364) - (1943)^2}\sqrt{9(7651) - (239)^2}}$$

$$r = \frac{62{,}339}{\sqrt{357{,}920}\sqrt{12{,}098}} = 0.947$$

$$b_1 = r\frac{s_y}{s_x}$$

$$s_x = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} y^2{}_i - (\sum_{i=1}^{n} y_i)^2\}}$$

$$s_x = \sqrt{\frac{1}{9(9-1)}\{9(7691) - (239)^2\}}$$

$$s_x = 70.50611$$

$$s_y = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2\}}$$

$$s_y = \sqrt{\frac{1}{9(9-1)}\{9(455{,}364) - (1934)^2\}}$$

$$= 12.96255$$

$$b_1 = r\frac{s_y}{s_x}$$

$$= 0.947 \times \frac{12.9625}{70.5061}$$

$$= 0.1742$$

$$\bar{x} = \frac{1934}{9} = 214.8889$$

$$\bar{y} = \frac{239}{9} = 26.5556$$

$$\mathbf{b_0 = \bar{y} - b_1 \bar{x}}$$

$$b_0 = 26.5556 - (0.1742)(214.8889)$$

$$\mathbf{b_0} = -10.8716$$

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = -10.8716 + (0.1742)x$$

Or

$$\hat{y} = -10.9 + (0.1742)x$$

where $\hat{y}$ is the predicted number of tickets sold and ***x*** is the amount of the jackpot.

Graph the regression equation $\hat{y}$ =-10.9 $+(0.1742)x$ on the scatterplot of the jackpot/tickets data from Table 1 and examine the graph to subjectively determine how well the regression line fits the data.