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.

Making Predictions

Regression equations are often useful for ***predicting* the value of one variable**, given some specific value of the other variable. When making predictions, we should consider the following:

1. Bad Model: If the regression equation **does not appear to be useful for making predictions, *don't* use the regression equation** for making predictions. For **bad models**, the best predicted value of a variable is simply its **sample mean**. However, the sample mean is not a ***good* predicted value** because it is the predicted value for *any* value of the other variable.

2. Good Model: Use the regression equation for predictions only if the graph of the **regression line on the scatterplot confirms** that the regression line fits the points reasonably well.

Making Predictions

3. Correlation: Use the **regression equation for predictions** only if the **linear correlation coefficient *r*** indicates that there is a **linear correlation between the two variables**.

4. Scope: Use the regression line for predictions only if the data **do not go much beyond the scope of the available sample data**. (Predicting too far beyond the scope of the available sample data is called ***extrapolation***, and it could result **in bad predictions**.)

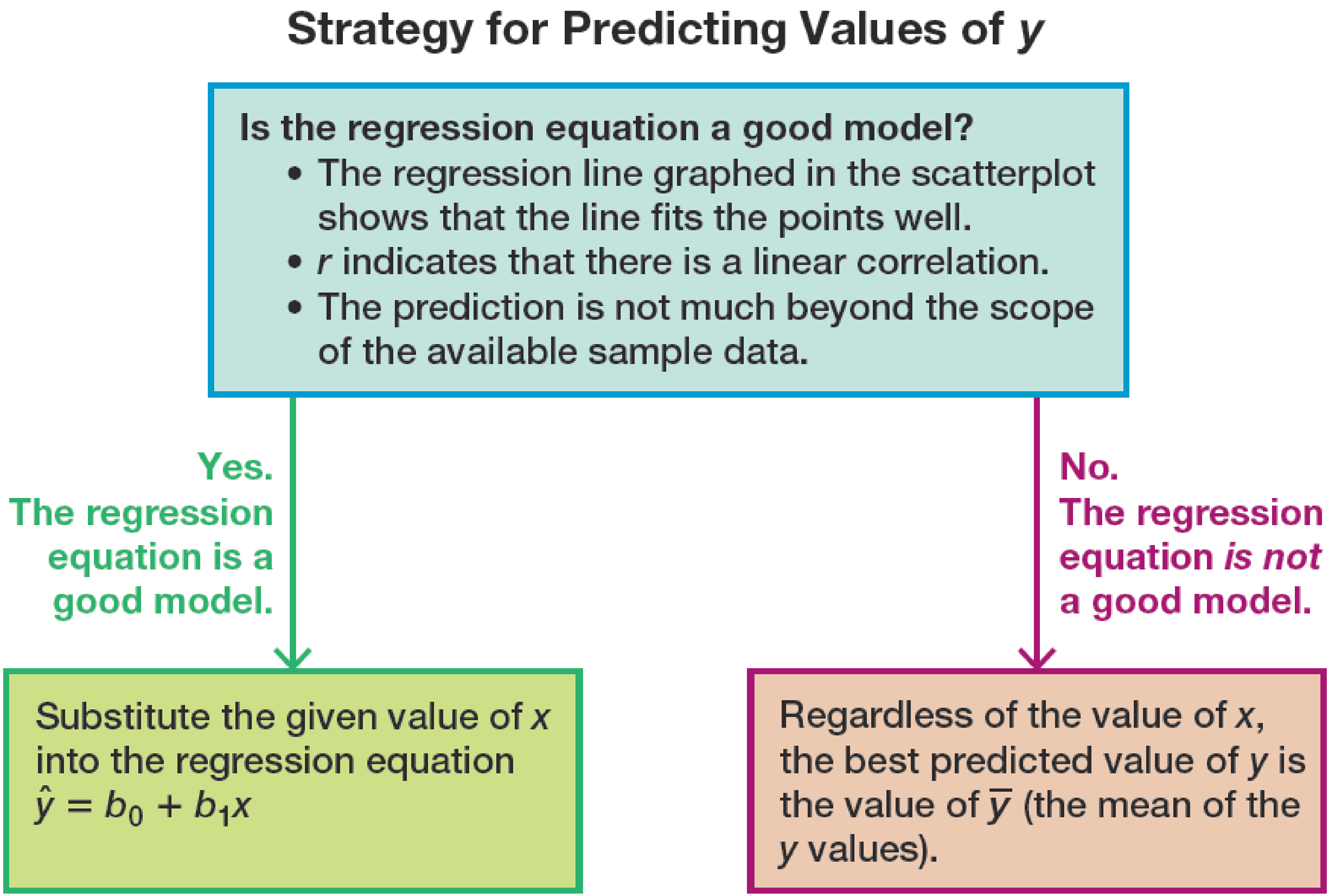


Figure 1: Recommended Strategy for Predicting Values of *y*

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

Example: Making Predictions

a. Use the **jackpot/tickets** data from **Table1** on to predict the number of lottery tickets sold when the jackpot is **$625 million**. How close is the predicted value to the actual value of **90 million tickets** that were actually sold when the Powerball lottery had a jackpot of **$625 million**?

b. Predict the **IQ score** of an adult who is **exactly 175 cm tall**.

Solution:

a. **Good Model: Use the Regression Equation for Predictions.** The regression line fits the points well, as shown in previously. Also, there is a linear correlation between Powerball jackpot amounts and numbers of tickets sold.

Because the regression equation

$\hat{y}$ **=-10.9** $+(0.1742)x$ is a good model, substitute x = 625 into the regression equation to get a predicted value **of 97.9 million tickets sold**.

The actual number of tickets sold **was 90 million**, so the predicted value of **97.9 million tickets** is pretty good.

b. Bad Model: Use $\overline{y}$ for predictions. There is no correlation between **height and IQ score**, so we know that a **regression equation is not a good model**.

Therefore, the best predicted IQ score value is the **mean IQ score**, which is **100**.

Marginal change in a variable

DEFINITION In working with two variables related by a **regression equation**, the **marginal change** in a variable is the amount that it changes when the other variable changes by **exactly one unit**. The **slope** b_1 in the **regression equation** represents the **marginal change in y that occurs when x changes by one unit**.

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

- The **slope of 0.174** tells us that if we increase the **jackpot *x* by 1 (million dollars)**, the predicted number of tickets sold will **increase by 0.174 million (or 174,000 tickets).** That is, for every **additional 1 million dollars** added to the jackpot amount, we expect the ticket sales to **increase by 174,000 tickets**.

- This realization has led **lottery officials to adjust their rules** to make winning more difficult **so that jackpots will grow considerably larger** and drive greater lottery ticket sales.

Outliers and Influential Points

A **correlation/regression** analysis of bivariate (paired) data should include an **investigation of outliers** and **influential points**, defined as follows.

DEFINITIONS

In a scatterplot, an **outlier** is a point lying **far away from the other data points**. Paired sample data may include one or more **influential points**, which are points that **strongly affect the graph of the regression line**.

Influential Point

Example: Influential Point

- Consider the nine pairs of jackpot/ticket data Problem. The scatterplot located to the left below on coming slide shows the regression line. If we include the additional pair of ***x*** **= 980 and** ***y*** **= 12**, we get the regression line shown on the coming slide.

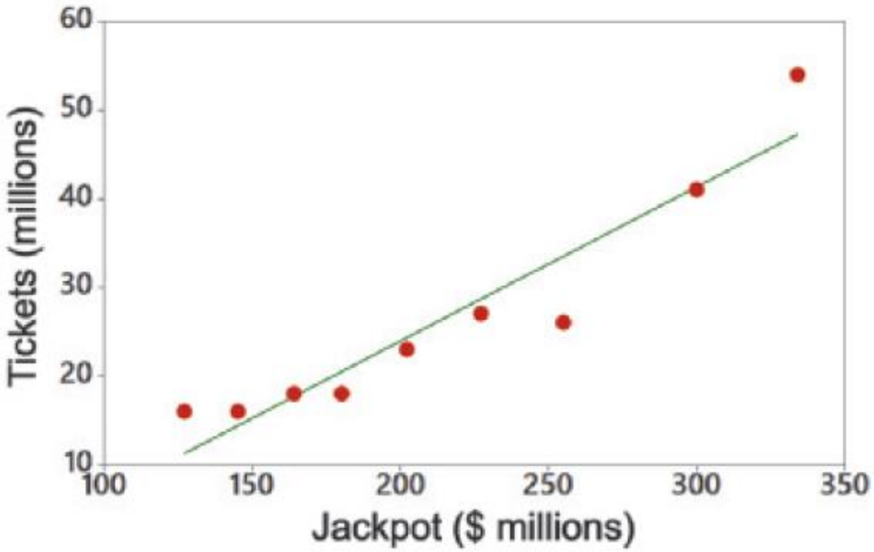
- The additional point **(980, 12) is an influential point** because the **graph of the regression line did change considerably** in the right graph on the coming slide.

Influential Point

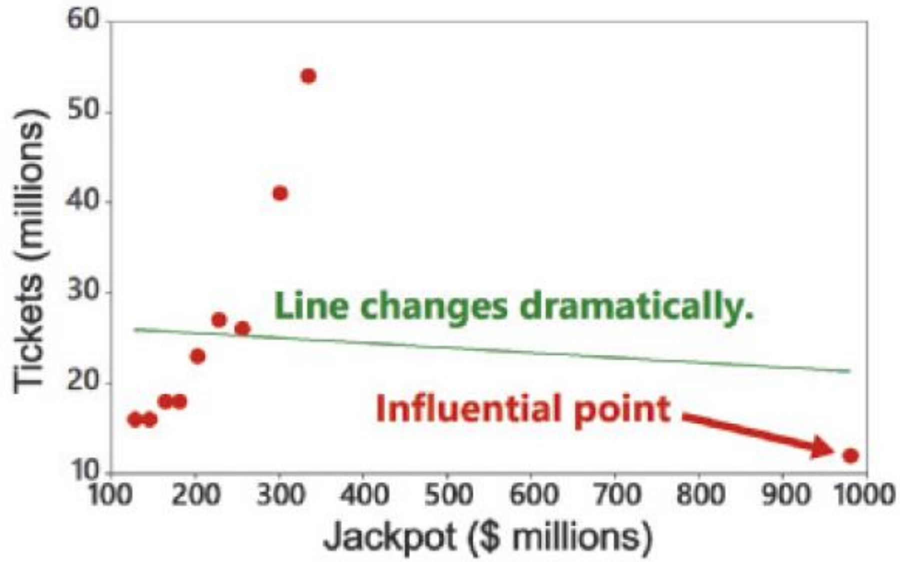
- **Example: Influential Point Cont.**

- Compare the two graphs to see clearly that the addition of this one pair of values has a very dramatic effect on the regression line, so that additional point is an **influential point.** The **additional point** is also an **outlier** because it is far from the other points.

Original Jackpot , Ticket Data from Table 1



Jackpot , Ticket Data with Additional Point: (980, 12)



Residuals and the Least-Squares Property

We stated that the regression equation represents the straight line that **"best" fits the data**. The criterion to determine the line that is better than all others is based on the **vertical distances** between **the original data points** and **the regression line**. Such distances are called ***residuals.***

DEFINITION

For a pair of **sample *x* and *y* values**, the **residual** is the difference between the ***observed* sample value of *y*** and the ***y* value that is *predicted*** by using the regression equation. That is,

Residual = **observed *y* - predicted *y* = *y* -** $\hat{y}$

Consider the sample point with coordinates of **(8, 4)** plotted in Figure 1. We get the following:

Observed value: For $x = 8$, the corresponding ***observed* value is** $y = 4$.

Predicted value: If we substitute $x = 8$ into the regression equation of $\hat{y} = 1 + x$, we get the ***predicted* value** $\hat{y} = 9$.

Residual: The difference between the **observed value** and **predicted value** is the residual, so the residual is $y - \hat{y} = 4 - 9 = -5$.

In Figure 1, the **residuals** are represented by the **dashed lines**. The paired data are plotted as red points in Figure 1.

x	8	12	20	24
y	4	24	8	32

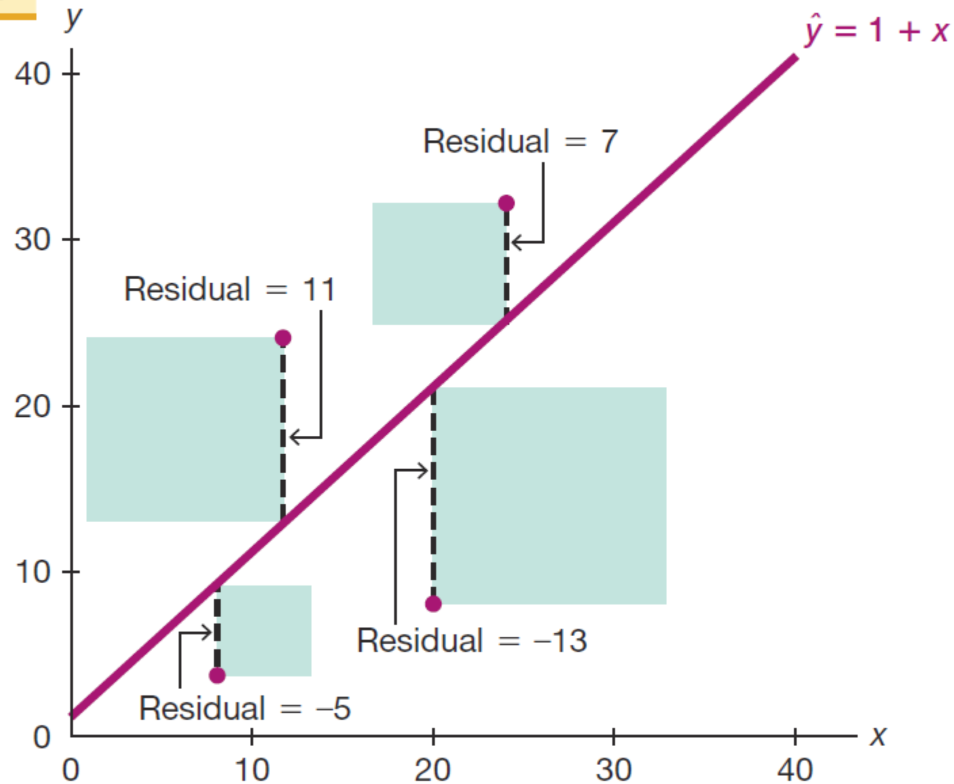


FIGURE 1 Residuals and Squares of Residuals

Least-squares property

The **regression equation** represents the line that "best" fits the points according to the following least-squares property.

DEFINITION

A straight line satisfies the **least-squares property** if the **sum of the squares of the residuals is the smallest sum possible**.

We see that the residuals are **-5, 11, -13, and 7**, so the sum of their squares is

$$(-5)^2 + (11)^2 + (-13)^2 + (7)^2 = 364$$

Least-squares property

- The sum of the shaded square areas is **364**, which is the **smallest sum possible**.

- **Use any other straight line**, and the shaded squares will combine to produce an **area larger than the combined shaded area of 364**.

Residual Plots

- We noted that we should always begin with a **scatterplot**, and we should verify that the pattern of points is **approximately a straight-line pattern**. We should also consider **outliers**.

- A ***residual plot*** can be another helpful tool for **analyzing correlation and regression results** and for checking the requirements **necessary for making inferences** about **correlation and regression**.

Residual Plot

A **residual plot** is a scatterplot of the (x, y) values after each of the y-coordinate values has been replaced by the residual value $\boldsymbol{y - \hat{y}}$ (where $\hat{y}$ denotes the predicted value of y). That is, a residual plot is a graph of the points $\boldsymbol{(x, y - \hat{y})}$

Usefulness of a Residual Plot

- A **residual plot** helps us determine whether **the regression line is a good model of the sample data**.

- A **residual plot** helps us to check the requirement that for different values of x, the corresponding y values all have the same **standard deviation**.

Criteria for Residual Plot

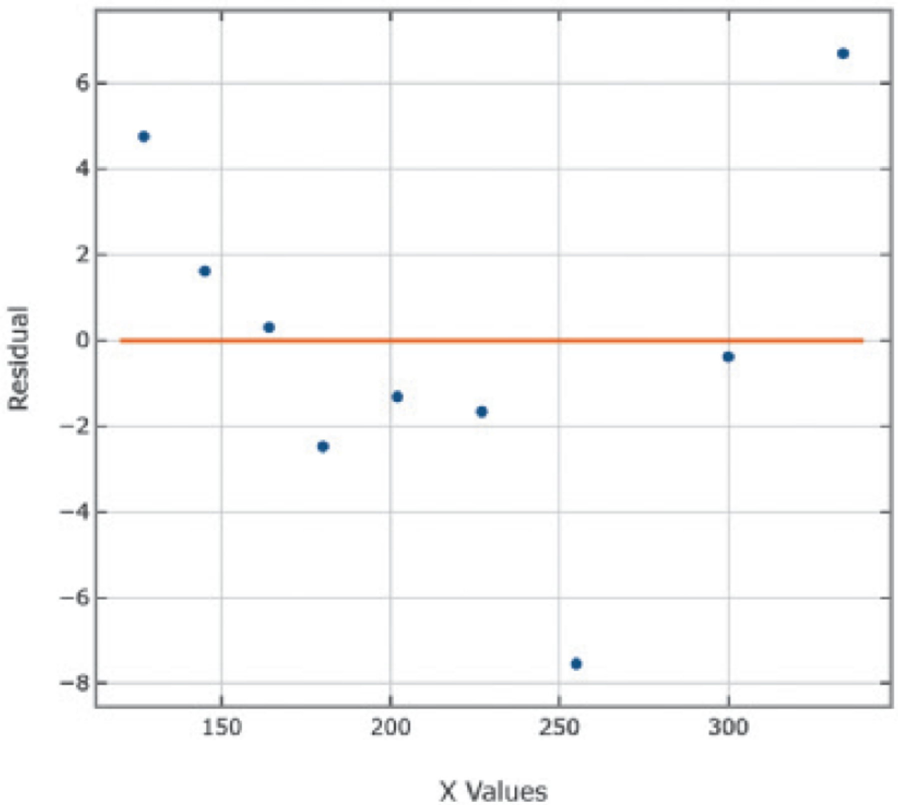
- The **residual plot** should **not have any obvious pattern** (not even a straight-line pattern). (This lack of a **pattern confirms that a scatterplot of the sample data is a straight-line pattern** instead of some other pattern.)

- The **residual plot should** not **become much wider (or thinner)** when viewed from **left to right**. (This confirms the requirement that for the different fixed values of *x*, the distributions of the corresponding *y* values all have the same standard deviation.)

Example: Residual Plot

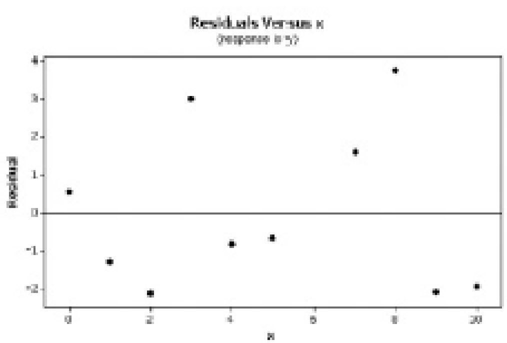
The jackpot/ ticket data from Table 1 are used to obtain the accompanying tool generated residual plot, which is a plot of the **$(x, y - \hat{y})$** values.

See that this residual plot satisfies the preceding two general criteria for residual plots.

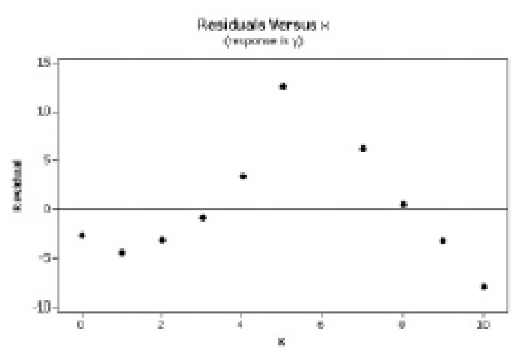


- The leftmost residual plot suggests that the **regression equation is a good model**.
- The middle residual plot shows a **distinct pattern**, suggesting that the sample data **do not follow a straight-line pattern as required**.
- The **rightmost residual** plot becomes **thicker**, which suggests that the requirement of equal standard deviations is violated.

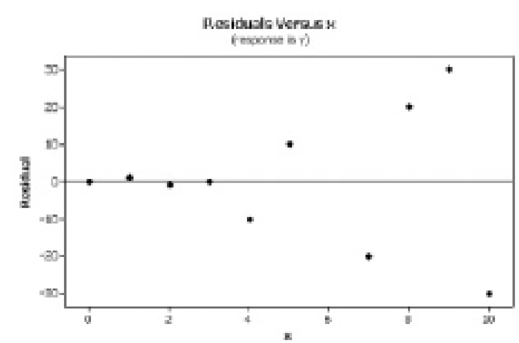
Residual Plot Suggesting That the Regression Equation Is a Good Model



Residual Plot with an Obvious Pattern, Suggesting That the Regression Equation Is Not a Good Model



Residual Plot That Becomes Wider, Suggesting That the Regression Equation Is Not a Good Model



Prediction Interval

DEFINITIONS

A **prediction interval** is a **range of values used to estimate a *variable*** (such as a predicted value of y in a regression equation).

A **confidence interval** is a **range of values used to estimate a population *parameter*** (such as ρ or μ or σ).

Prediction Intervals

Objective

Find a **prediction interval**, which is an **interval estimate of a predicted value of *y***.

Requirement

For each fixed value of x, the corresponding sample values of y are **normally distributed** about the regression line, and those **normal distributions have the same variance**.

Prediction Intervals

Given a **fixed and known value x_0**, the prediction interval for an individual *y* value is

$$\hat{y} - E < y < \hat{y} + E$$

where the margin of error is

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

and **x_0 is a given value of *x***, $t_{\alpha/2}$ has *n* - 2 degrees of freedom, and s_e is the **standard error** of estimate

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$ (This formulae is good for manual calculations or writing computer programs.)

Prediction Interval

In previous example, we showed that when using the 9 pairs of jackpot/tickets data from Table 1, the regression equation is **$\hat{y}$ =-10.9 $+$(0.1742)x,** and for a jackpot of ***x* = 625 million dollars**, the predicted value of ***y* is 97.9 million tickets** (which is found by substituting *x* = 625 in the regression equation). For *x* = 625, the "best" predicted value of *y* is 97.9, but we have **no sense of the accuracy of that estimate**, so we **need an interval estimate**

Prediction Interval

EXAMPLE 1 Powerball Jackpots and Ticket Sales: Finding a Prediction Interval

For the paired jackpot / tickets data in Table 1, we found that there is sufficient evidence to support the claim of a linear correlation between those two variables, and we found that the regression equation is $\hat{y}$ **=-10.9** $+(\mathbf{0.1742})x$. We also found that if the jackpot amount is ***x* = 625 million dollars**, the predicted number of tickets sold is 97.9 million (or 98.0 million if using calculations with more decimal places).

Use the jackpot amount of 625 million dollars to construct a 95% prediction interval for the number of tickets.

Prediction Interval

The 95% prediction interval is

$$\hat{y} - E < y < \hat{y} + E$$

where the margin of error is

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

Where

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

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	$\sum x^2 =$ 455364	$\sum y^2 =$ 7691	$\sum xy =$ 58285

Prediction Interval

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

$$r = \frac{9(58{,}2852) - (1934)(239)}{\sqrt{9(455{,}364) - (1943)^2}\sqrt{9(7651) - (239)^2}}$$

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

Prediction Interval

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

$$s_y = \sqrt{\frac{1}{n(n-1)}\{n\sum y^2 - (\sum y)^2\}}$$

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

$$s_y = 70.50611$$

$$s_x = \sqrt{\frac{1}{n(n-1)}\{n\sum x^2 - (\sum x)^2\}}$$

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

$$= 12.96255$$

Prediction Interval

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

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

$$= 0.1742$$

$$\overline{\boldsymbol{x}} = \frac{1934}{9} = 214.8889$$

$$\overline{\boldsymbol{y}} = \frac{239}{9} = 26.5556$$

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

$$\mathrm{b_0} = 26.5556 - (0.1742)(214.8889)$$

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

Prediction Interval

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

$$\Rightarrow s_e = \sqrt{\frac{7691 - (-10.8716)(239) - (0.1742)(58285)}{9-2}}$$

$$= 4.4088$$

$$t_{\left(\frac{\alpha}{2}, n-2\right)} = t_{(0.0250,\, 7)} = 2.365$$

x_0 = 625 (given)

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \overline{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

$$= (2.365)(4.4088)(\sqrt{1 + \frac{1}{9} + \frac{9(625 - 214.8889)^2}{9(455364) - (1934)^2}})$$

$$= (2.365)(4.4088)(\sqrt{1 + 0.1111 + 4.2292})$$

$$= (2.365)(4.4088)(\ (2.3109)$$

E = 24.0953

$\hat{y}$ = 97.9 (predicted value of y found by substituting x = 625 into the regression equation)

The 95% prediction interval is

$$\hat{y} - E < y < \hat{y} + E$$

$\Rightarrow$**73.7 million tickets < *y* < 122 million tickets**

(which does contain the value of **90 million tickets** that were actually sold in this particular lottery).

This means that if we select some particular lottery with a jackpot of 625 million dollars (x = 625), we have 95% confidence that the limits of

73.7 million tickets < *y* < 122 million tickets

contain the actual ticket sales in millions.

That is a wide range of values. The **prediction interval** would be **much narrower** and our estimated number of tickets would be much better if **the margin of error *E*** was not so large (due to the **small sample size** and the large difference between **the outlier jackpot of x = 625 million dollars and $\overline{x}$ = 214.8889 million dollars**).