REGRESSION AND CORRELATION

PMAS ARID AGRICULTURE UNIVERSITY
1994
RAWALPINDI

Learning Objectives

After completing this chapter, students will be able to:

1. **Identify variables and use them in a regression model**
2. **Develop simple linear regression equations from sample data and interpret the slope and intercept**
3. **Compute the coefficient of determination and the coefficient of correlation and interpret their meanings**
4. **Interpret the F-test in a linear regression model**
5. **List the assumptions used in regression and use residual plots to identify problems**

Triple A Construction

- Triple A Construction renovates old homes.
- They have found that the dollar volume of renovation work is dependent on the area payroll.

TRIPLE A'S SALES ($100,000's)	LOCAL PAYROLL ($100,000,000's)
6	3
8	4
9	6
5	4
4.5	2
9.5	5

Table 1

Triple A Construction

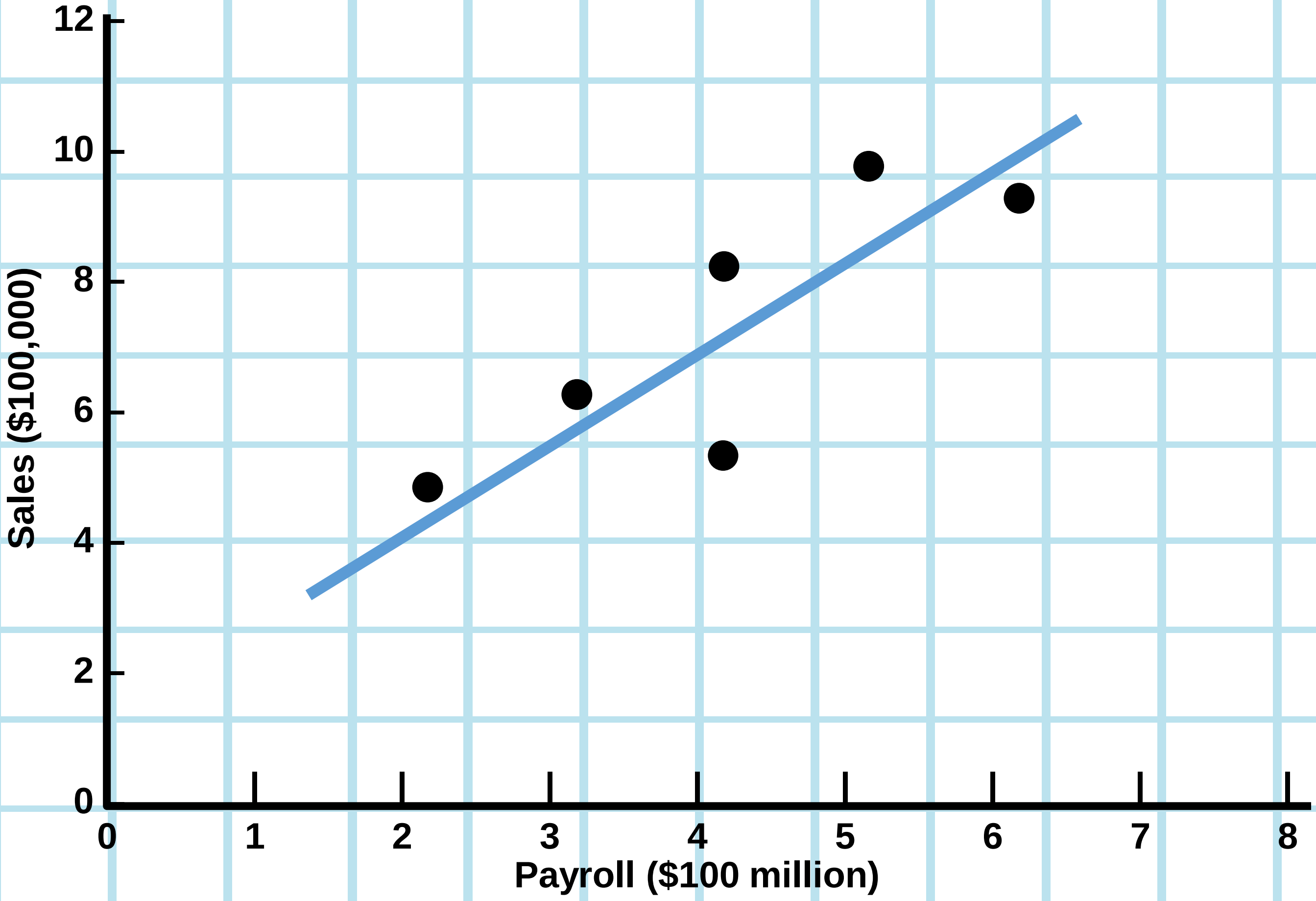

Figure 1

Measuring the Fit of the Regression Model

- Regression models can be developed for any variables *X* and *Y*
- How do we know the model is actually helpful in predicting *Y* based on *X*?
  - We could just take the average error, but the positive and negative errors would cancel each other out
- Three measures of variability are
  - *SST* – Total variability about the mean
  - *SSE* – Variability about the regression line
  - *SSR* – Total variability that is explained by the model

- **Sum of the squares total**

$$SST = \sum (Y - \bar{Y})^2$$

- **Sum of the squared error**

$$SSE = \sum e^2 = \sum (Y - \hat{Y})^2$$

- **Sum of squares due to regression**

$$SSR = \sum (\hat{Y} - \bar{Y})^2$$

- **An important relationship**

$$SST = SSR + SSE$$

Y	X	$(Y - \bar{Y})^2$	$\hat{Y}$	$(Y - \hat{Y})^2$	$(\hat{Y} - \bar{Y})^2$
6	3	$(6 - 7)^2 = 1$	2 + 1.25(3) = 5.75	0.0625	1.563
8	4	$(8 - 7)^2 = 1$	2 + 1.25(4) = 7.00	1	0
9	6	$(9 - 7)^2 = 4$	2 + 1.25(6) = 9.50	0.25	6.25
5	4	$(5 - 7)^2 = 4$	2 + 1.25(4) = 7.00	4	0
4.5	2	$(4.5 - 7)^2 = 6.25$	2 + 1.25(2) = 4.50	0	6.25
9.5	5	$(9.5 - 7)^2 = 6.25$	2 + 1.25(5) = 8.25	1.5625	1.563
		$\sum(Y - \bar{Y})^2 = 22.5$	$\sum(Y - \hat{Y})^2$	= 6.875	$\sum(\hat{Y} - \bar{Y})^2 = 15.625$
$\bar{Y} = 7$		$SST = 22.5$	SSE	= 6.875	$SSR = 15.625$

Table 3

For Triple A Construction

SST = 22.5
SSE = 6.875
SSR = 15.625

- Sum of the squares total

$$SST = \sum (Y - \overline{Y})^2$$

- Sum of the squared error

$$SSE = \sum e^2 = \sum (Y - \hat{Y})^2$$

- Sum of squares due to regression

$$SSR = \sum (\hat{Y} - \overline{Y})^2$$

- An important relationship

$$SST = SSR + SSE$$

  - SSR – explained variability
  - SSE – unexplained variability

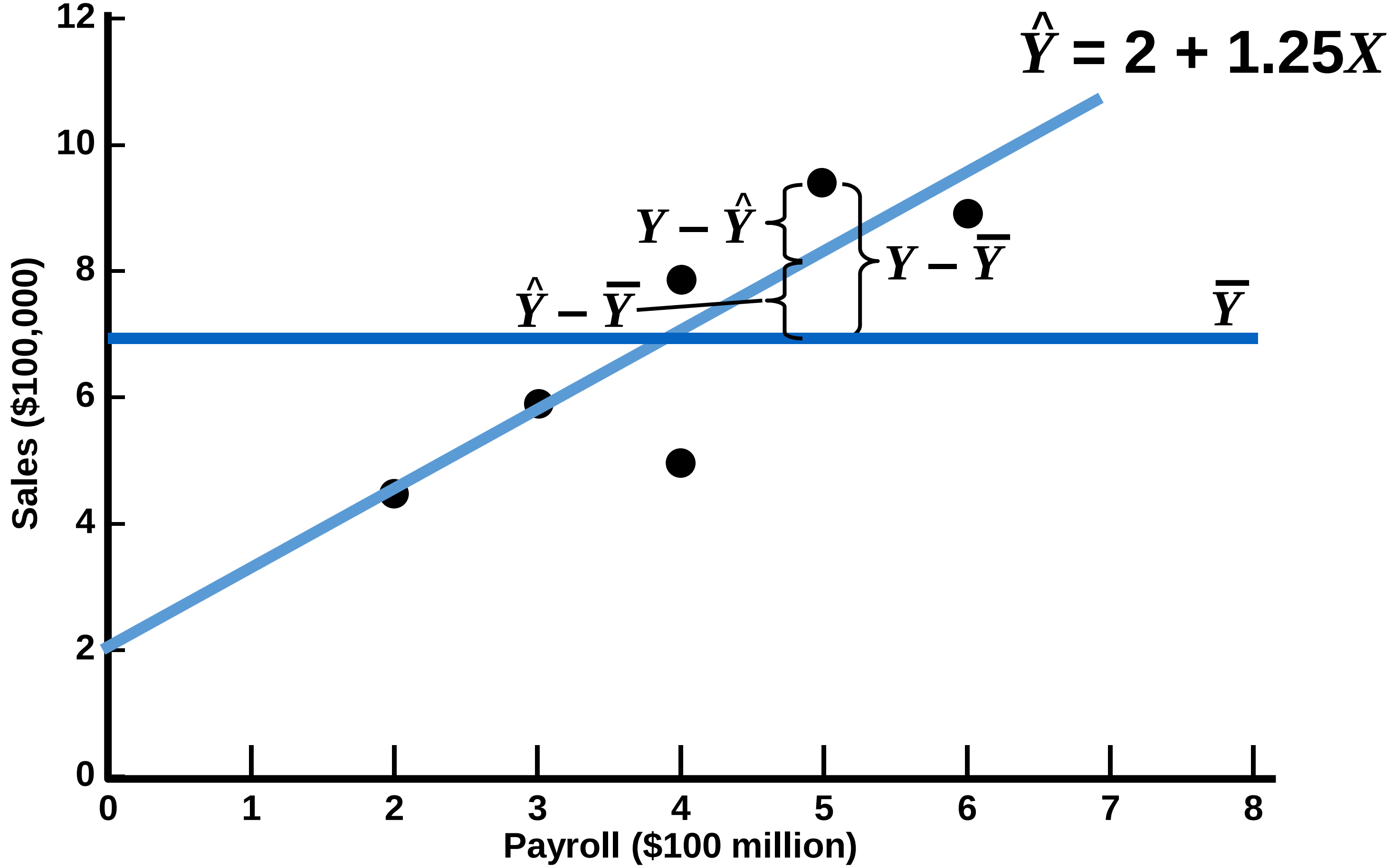
$\hat{Y} = 2 + 1.25X$
$Y - \hat{Y}$
$Y - \bar{Y}$
$\hat{Y} - \bar{Y}$
$\bar{Y}$
12
10
8
6
4
2
0
Sales ($100,000)
0
1
2
3
4
5
6
7
8
Payroll ($100 million)

Figure 4.2

Coefficient of Determination

- The proportion of the variability in Y explained by regression equation is called the *coefficient of determination*
- The coefficient of determination is r^2

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- **For Triple A Construction**

$$r^2 = \frac{15.625}{22.5} = 0.6944$$

- **About 69% of the variability in Y is explained by the equation based on payroll (X)**

Assumptions of the Regression Model

- If we make certain assumptions about the errors in a regression model, we can perform statistical tests to determine if the model is useful

  1. Errors are independent
  2. Errors are normally distributed
  3. Errors have a mean of zero
  4. Errors have a constant variance

- A plot of the residuals (errors) will often highlight any glaring violations of the assumption

REGRESSION LINE OF X ON Y

The line which expresses the trend of two observed values is called a regression line. For example if the sample data is given then the value of $\boldsymbol{y}$ corresponding to the given value of $\boldsymbol{x}$ can be estimated by the method of least squares. Now because the value of $\boldsymbol{y}$ is estimated from given value of $\boldsymbol{x}$ therefore the resulting line is called regression line of $\boldsymbol{y}$ on $\boldsymbol{x}$ which means that $\boldsymbol{y}$ is dependent on $\boldsymbol{x}$. The general equation of $\boldsymbol{y}$ on $\boldsymbol{x}$ is

$$\boldsymbol{Y = a + bx}$$

where $\boldsymbol{b} = \frac{\boldsymbol{n \sum xy - (\sum y)(\sum x)}}{\boldsymbol{n \sum x^2 - (\sum x)^2}}$ and $\mathbf{a} = \overline{\boldsymbol{y}} - \boldsymbol{b}\overline{\boldsymbol{x}} = \frac{\boldsymbol{\sum y}}{\boldsymbol{n}} - \boldsymbol{b}\frac{\boldsymbol{\sum x}}{\boldsymbol{n}}$

Question

Problem: The following table shows the chart of price and demand for an item at different periods of time.

I. Forecast demand for the price of $ 25

- Solution:

Price (x)	*Demand* (y)	xy	x^2	y^2
18	45	810	324	2025
22	40	880	484	1600
25	34	850	625	1156
30	26	780	900	676
18	42	756	324	1764
15	44	660	225	1936
12	45	540	144	2025
140	276	5276	3026	11182

Regression Line of Y on X	
$Y = a + bx$ **Where,** $b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$ $b = \frac{7(5276) - (140)(276)}{7(3026) - (140)^2}$ $b = \frac{36932 - 38640}{21182 - 19600}$ $b = \frac{-1708}{1582}$ $b = -1.08$	**And** $a = \bar{y} - b\bar{x}$ $a = \frac{276}{7} - (-1.08)\left(\frac{140}{7}\right)$ $a = 39.43 + 21.60 = 61.03$ **Therefore,** $y = 61.03 - 1.08x$ *By putting the value of* $x = 25$ *we get* $Y = 61.03 - 1.08(25)$ $Y = 61.03 - 27 = 34.03$