Linear Regression
Introduction

Course: Statistics 1

Lecturer: Dr. Courtney Pindling

Introduction

Regression:

The statistical technique for finding the best-fitting straight line for two sets of data

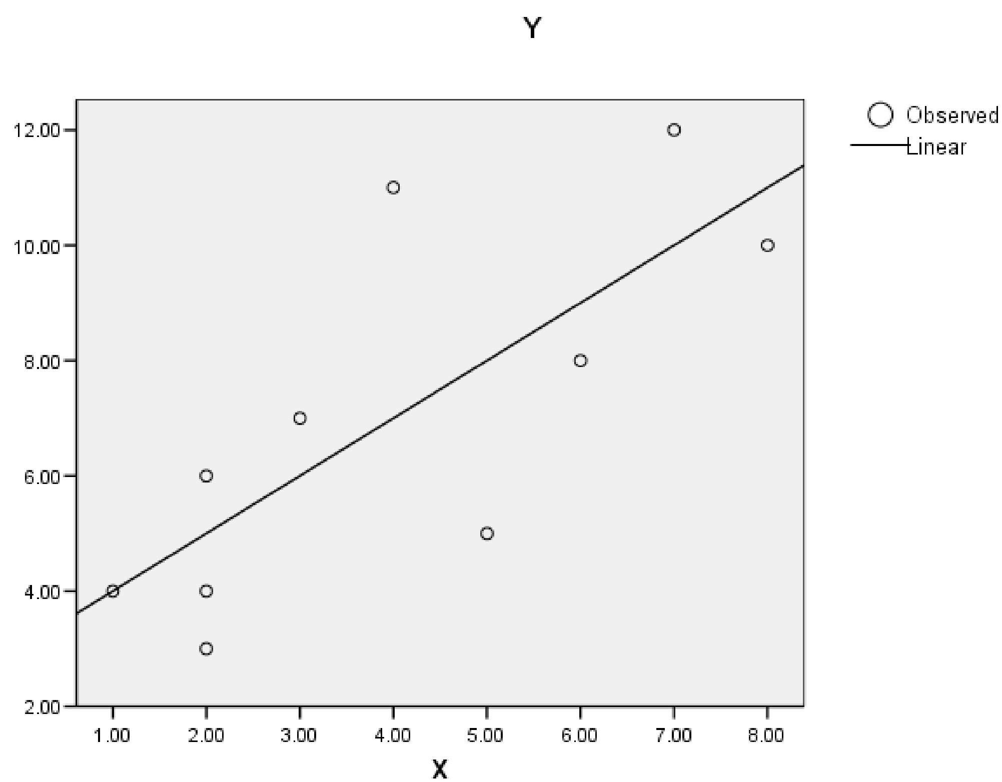
Regression Line:

The best-fitted straight line from a regression technique

Regression Plot

Regression Line:

1. Make **relationship** between X and Y **easier to see**
2. The line identifies the center or ***central tendency*** of the relationship
3. The line can be use for **prediction**



Regression Model

- Linear Relationship between X and Y
- X is the independent variable
- Y is the dependent variable
- ***$Y = B_1 X + B_0$***
- ***B_1*** is the slope of the line
- ***B_0*** is the y-intercept or value of Y when X = 0
- The residual is the vertical distance between each data point and the regression line

X ***Independent***	**Y** ***Dependent***
4	11
3	7
1	4
7	12
2	6

Regression Table

X	Y	X - M_X	Y - M_Y	$(X-MX)^2$ (5)	$(Y-MY)^2$ (6)	(5)(6)
4	11	0.6	3	0.36	9	1.8
3	7	-0.4	-1	0.16	1	0.4
1	4	-2.4	-4	5.76	16	9.6
7	12	3.6	4	12.96	16	14.4
2	6	-1.4	-2	1.96	4	2.8
M_X	M_Y			SS_X	SS_Y	SP
3.4	8			21.2	46	29

Regression Equation

- **From Regression Table:**

  SP = 29, SS_X = 21.2 , M_X = 3.4 , M_Y = 8

- $Y = B_1 X + B_0$
- B_1 is the slope of the line

  $B_1 = SP/SS_X = 29/21.2 = 1.368$

- B_0 is the y-intercept

  $B_0 = M_Y - B_1 M_X = 8 - 1.368(3.4) = 3.35$

- $Y = 1.368X + 3.35$

Residual Sum of Square: SS_{Res}

X	Y	X - M_X	Y - M_Y	$(X-M_X)^2$ (5)	$(Y-M_Y)^2$ (6)	(5)(6)	Predicted Y *1.368X+3.35*	Residual Res	Res^2
4	11	0.6	3	0.36	9	1.8	8.82	2.18	4.75
3	7	-0.4	-1	0.16	1	0.4	7.45	-0.45	0.21
1	4	-2.4	-4	5.76	16	9.6	4.72	-0.72	0.51
7	12	3.6	4	12.96	16	14.4	12.92	-0.92	0.85
2	6	-1.4	-2	1.96	4	2.8	6.08	-0.08	0.007
M_X	M_Y			SS_X	SS_Y	SP			SS_{Res}
3.4	8			21.2	46	29			**6.33**

Predicted Y: for x = 5, $Y = 1.368(\mathbf{5}) + 3.35 = \mathbf{10.19}$

Standard Error of Estimate

- **Standard error of the estimate**: gives a measure of the standard distance between a regression line and the actual data values
- $SS_{Residual}$ = **Sum** (Y – Predicted Y)^2 = *6.33*
- Variance = *SS/df*
- $df = n - 2 = 5 - 2 = 3$

$$\text{std error of est} = \sqrt{\frac{SS_{Res}}{df}} = \sqrt{\frac{6.33}{3}} = 1.45$$

Standard Error and Correlation

- *Std error of the estimate* is **directly related** to the *magnitude of the correlation* between X and Y
- When the *correlation is near 1.00* (or -1.00) the data values will be *clustered close to regression line*
- As the correlation *nears zero*, the line will provide *less accurate predictions*
- r^2 measures the portion of the variability in the Y scores that is *predicted by the regression equation*
- ($1 - r^2$) measures the unpredicted portion

Pearson Correlation

- Predicted variability = SS_{Reg} = $\boldsymbol{r^2 SS_Y}$
- Unpredicted variability = SS_{Res} = $\boldsymbol{(1 - r^2) SS_Y}$
- Pearson Correlation

$$\text{std error of est} = \sqrt{\frac{SS_{Res}}{df}} = \sqrt{\frac{(1-r^2)SS_Y}{n-2}}$$

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{29}{\sqrt{21.2(46)}} = 0.929 \text{ and } r^2 = 0.863$$

$$SS_{Reg} = r^2 SS_Y = (0.929)^2(46) = 39.70$$

$$SS_{Res} = (1-r^2)SS_Y = [1-(0.929)^2](46) = 6.30$$

SPSS Output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.929[a]	.862	.817	1.45261

a. Predictors: (Constant), X

Cautions for Predictions

- The predicted value is not perfect unless
  $r = +1.00$ or -1.00
- The regression equation should not be used to make prediction for X values that fall outside the range of values covered by the original data set
- Inclusion of extreme values may bias the regression equation prediction capability