14.3: Estimating the Regression Model with the Least-Square Line

We now return to the case where we know the data and can see the linear correlation in a scatterplot, but we do not know the values of the parameters of the underlying model. The three parameters that are unknown to us are the y-intercept β_0, the slope (β_1) and the standard deviation of the residual error (σ):

Slope parameter: b_1 will be an estimator for β_1

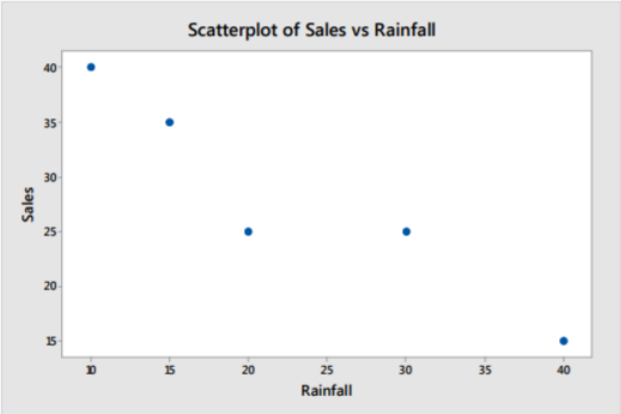
Y-intercept parameter: b_0 will be an estimator for β_0

Standard deviation: s_e will be an estimator for σ

$$\text{Regression line: } \hat{Y} = b_0 + b_1 X$$

Example

Take the example comparing rainfall to sales of sunglasses in which the scatterplot shows a negative correlation. However, there are many lines we could draw. How do we find the line of best fit?



Solution

Minimizing Sum of Squared Residual Errors (SSE)

We are going to define the "best line" as the line that minimizes the Sum of Squared Residual Errors (SSE).

Suppose we try to fit this data with a line that goes through the first and last point. We can then calculate the equation of this line using algebra:

$$\hat{Y} = \frac{145}{3} - \frac{5}{6}X \approx 48.3 - 0.833X$$

The SSE for this line is 47.917:

Rainfall	Sales	Predicted Sales	Residual	Squared Residuals
10	40	40	0	0
15	35	35.833	-0.833	0.694
20	25	31.667	-6.667	44.444
30	25	23.333	1.667	2.778
40	15	15	0	0
		Sum of Squared Residuals =		47.917

Although this line is a good fit, it is not the best line. The slope(b_1) and intercept(b_o) for the line that minimizes SSE is be calculated using the least squares principle formulas:

Least squares principle formulas

$$SSX = \Sigma X^2 - \frac{1}{n}(\Sigma X)^2$$

$$SSY = \Sigma Y^2 - \frac{1}{n}(\Sigma Y)^2$$

$$SSXY = \Sigma XY - \frac{1}{n}(\Sigma X \cdot \Sigma y)$$

$$b_1 = \frac{SSXY}{SSX}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

In the Rainfall example where X=Rainfall and Y=Sales of Sunglasses:

	X	Y	X^2	Y^2	XY
	10	40	100	1600	400
	15	35	225	1225	525
	20	25	400	625	500
	30	25	900	625	750
	40	15	1600	225	600
Σ	115	140	3225	4300	2775

- SSX = 580
- SSY = 380
- SSXY = -445

- b_1 = -.767
- b_0 = 45.647
- $\hat{Y}$ = 45.647 - .767X

The Sum of Squared Residual Errors (SSE) for this line is 38.578, making it the "best line". (Compare to the value above, in which we picked the line that perfectly fit the two most extreme points).

Rainfall	Sales	Predicted Sales	Residual	Squared Residuals
10	40	37.977	2.023	4.092529
15	35	34.142	0.858	0.736
20	25	30.307	-5.307	28.164
30	25	22.637	2.363	5.584
40	15	14.967	0.033	0.001089
		Sum of Squared Residuals =		38.578

In practice, we will use technology such as Minitab to calculate this line. Here is the example using the Regression Fitted Line Plot option in Minitab, which determines and graphs the regression equation. The point (20,25) has the highest residual error, but the overall Sum of Squared Residual Errors (SSE) is minimized.