

Linear regression example

Consider the experimental data in Table 11.1, which were obtained from 33 samples of chemically treated waste in a study conducted at Virginia Tech. Readings on x , the percent reduction in total solids, and y , the percent reduction in chemical oxygen demand, were recorded.

Table 11.1: Measures of Reduction in Solids and Oxygen Demand

Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)	Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

The data are plotted in a scatter diagram in Figure 11.3. From an inspection of this scatter diagram, it can be seen that the points closely follow a straight line, indicating that the assumption of linearity between the two variables appears to be reasonable.

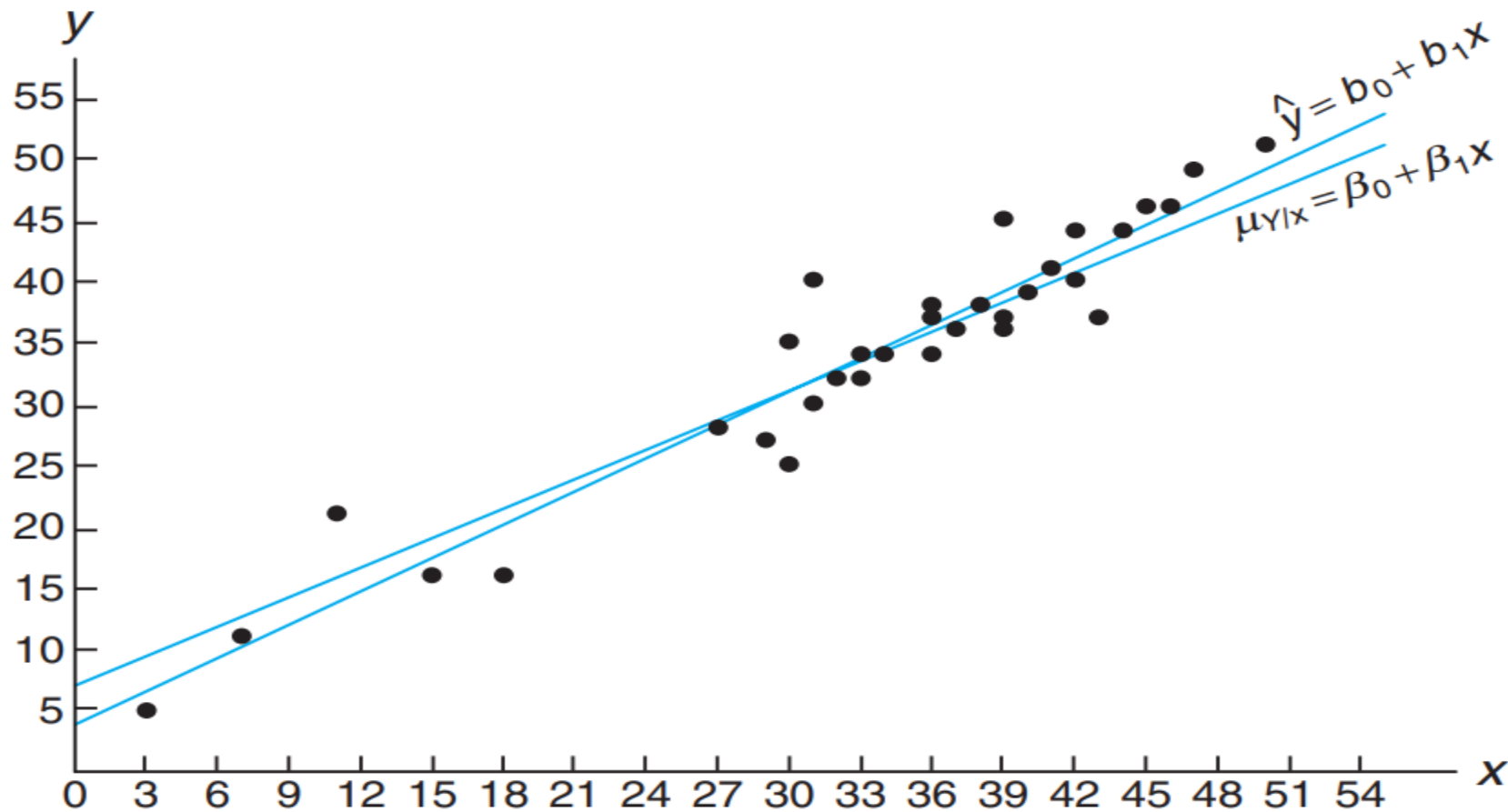
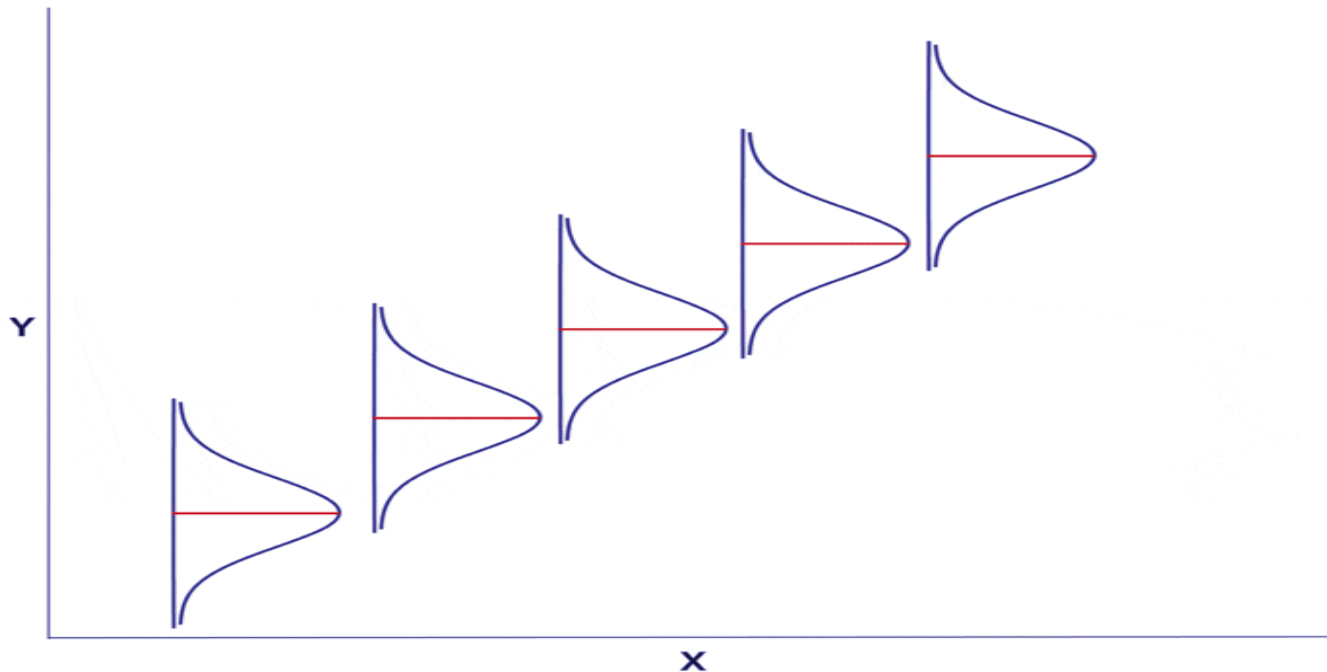


Figure 11.3: Scatter diagram with regression lines.

Model Assumptions

- The residual errors are random and are normally distributed.
- The standard deviation of the residual error does not depend on X
- A linear relationship exists between X and Y
- The samples are randomly selected

<https://youtu.be/h8cTBrYHWqA>



We shall find b_0 and b_1 , the estimates of β_0 and β_1 , so that the sum of the squares of the residuals is a minimum. The residual sum of squares is often called the sum of squares of the errors about the regression line and is denoted by SSE. This minimization procedure for estimating the parameters is called the method of least squares. Hence, we shall find a and b so as to minimize

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

Given the sample $\{(x_i, y_i); i = 1, 2, \dots, n\}$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and}$$

$$b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} = \bar{y} - b_1 \bar{x}.$$

Estimate the regression line for the pollution data of Table 11.1.

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41,355, \quad \sum_{i=1}^{33} x_i^2 = 41,086$$

Therefore,

$$b_1 = \frac{(33)(41,355) - (1104)(1124)}{(33)(41,086) - (1104)^2} = 0.903643 \text{ and}$$
$$b_0 = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Thus, the estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

Hypothesis Test for Simple Linear Regression

We will now describe a hypothesis test to determine if the regression model is meaningful; in other words, does the value of X in any way help predict the expected value of Y ?

An unbiased estimate of σ^2 is

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}.$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Hypothesis Testing on the Slope

To test the null hypothesis H_0 that $\beta_1 = \beta_{10}$ against a suitable alternative, we again use the t -distribution with $n - 2$ degrees of freedom to establish a critical region and then base our decision on the value of

$$t = \frac{b_1 - \beta_{10}}{s/\sqrt{S_{xx}}}.$$


The method is illustrated by the following example.

Example 11.3: Using the estimated value $b_1 = 0.903643$ of Example 11.1, test the hypothesis that $\beta_1 = 1.0$ against the alternative that $\beta_1 < 1.0$.

Solution: The hypotheses are $H_0: \beta_1 = 1.0$ and $H_1: \beta_1 < 1.0$. So

$$t = \frac{0.903643 - 1.0}{3.2295/\sqrt{4152.18}} = -1.92,$$

with $n - 2 = 31$ degrees of freedom ($P \approx 0.03$).

Decision: The t -value is significant at the 0.03 level, suggesting strong evidence that $\beta_1 < 1.0$. 

Test Hypotheses using ANOVA

- Ho: X and Y are not correlated
- Ha: X and Y are correlated

Or

- Ho: β_1 (slope) = 0
- Ha: β_1 (slope) \neq 0

Test Statistic

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Error}}} \quad df_{\text{num}} = 1$$
$$df_{\text{den}} = n - 2$$

Sum of Squares

$$SS_{\text{Total}} = \sum (Y - \bar{Y})^2$$

$$SS_{\text{Error}} = \sum (Y - \hat{Y})^2$$

$$SS_{\text{Regression}} = SS_{\text{Total}} - SS_{\text{Error}}$$

In simple linear regression, this is equivalent to saying, "Are X and Y correlated?"

In reviewing the model, $Y = \beta_0 + \beta_1 X + \varepsilon$, as long as the slope (β_1) has any non-zero value, X will add value in helping predict the expected value of Y . However, if there is no correlation between X and Y , the value of the slope (β_1) will be zero.

The model we can use is very similar to One Factor ANOVA.

The Results of the test can be summarized in a special ANOVA table:

Source of Variation	Sum of Squares (SS)	Degrees of freedom (df)	Mean Square (MS)	F
Factor (due to X)	$SS_{\text{Regression}}$	1	$MS_{\text{Factor}} = SS_{\text{Factor}} / 1$	$F = MS_{\text{Factor}} / MS_{\text{Error}}$
Error (Residual)	SS_{Error}	$n - 2$	$MS_{\text{Error}} = SS_{\text{Error}} / n - 2$	
Total	SS_{Total}	$n - 1$		

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3390.6	3390.6	325.08	0.000
Residual Error	31	323.3	10.4		
Total	32	3713.9			

Decision: Since the value of F statistics is large $F = 325.08$, Then The F-value is significant at the 0.0000 level, suggesting strong evidence that $\beta_1 \neq 0$.

So, we can't accept H_0