Module 5 – Linear Regression

$(0, 5)$

$(1, -3)$

$(2, 4)$

$\vdots$

$f(\underbrace{0.5}) \simeq$

Lesson goals

1. Knowing how to compute the slope and intercept of a best-fit straight line with linear regression.
2. Knowing how to compute and understand the meaning of the coefficient of determination and the standard error of the estimate.
3. Understanding how to use transformations to linearize nonlinear equations so that they can be fit with linear regression.
4. Knowing how to implement linear regression with MATLAB.

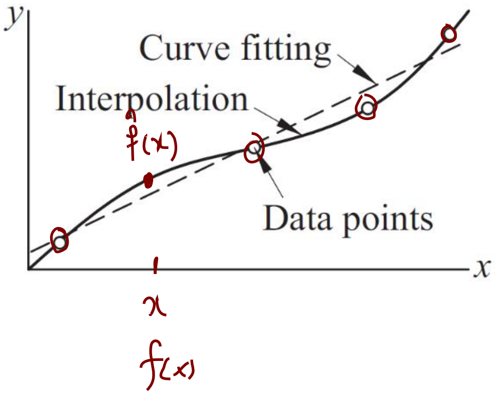
Introduction

Interpolation vs curve fitting: Discrete data sets, or tables of the form

x_0	x_1	x_2	…	x_n
y_0	y_1	y_2	…	y_n

are commonly involved in technical calculations. The source of the data may be experimental observations or numerical computations.

There is a distinction between interpolation and curve fitting. In interpolation we construct a curve through the data points. In doing so, we make the implicit assumption that the data points are accurate and distinct. In contrast, curve fitting is applied to data that contain scatter (noise), usually caused by measurement errors. Here we want to find a smooth curve that approximates the data in some sense. Thus, the curve does not necessarily hit the data points. The difference between interpolation and curve fitting is illustrated in the following figure.



Curve fitting

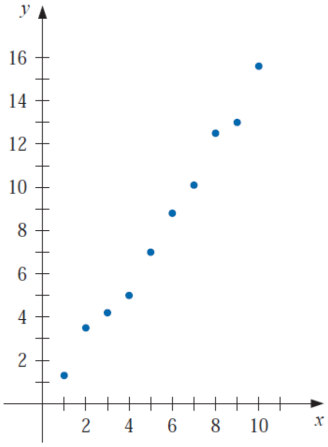
Curve fitting examines the relationship between one or more predictors (independent variables) and a response variable (dependent variable), with the goal of defining a "best fit" model of the relationship.

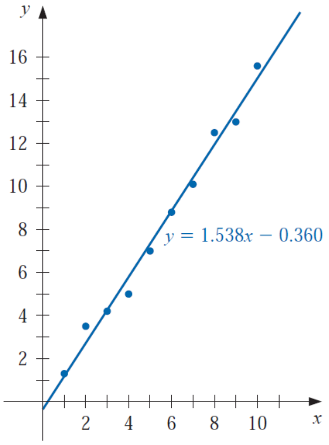
The best curve-fitting strategy is to derive an approximating function that fits the shape or general trend of the data without necessarily matching the individual points. One approach to do this is to visually inspect the plotted data and then sketch a “best”

$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$

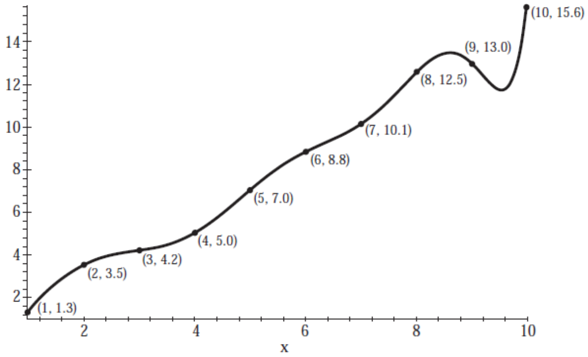
$(1, 3), (2, 5.5), (3, 6), (4, 8.8), (5, 11)$

line through the points. Consider the following scatter plot.





A better approach would be to find the "best" (in some sense) approximating line



This polynomial is clearly a poor predictor of information between a number of the data points

Linear regression

The simplest case it to fit a straight line to a set of paired observations:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

The mathematical expression for the line of best fit (regression line) is $\boxed{\hat{y} = a_0 + a_1 x.}$ ✓
Thus, the actual y can be written as:

$$y = \hat{y} + e = a_0 + a_1 x + e$$

where a_0 and a_1 are coefficients representing the y-intercept and the slope, respectively, e is the *error* or *residual* between the model and the observations, and $\hat{y}$ is the predicted value for actual y. Note that

$$e = y - \hat{y} = y - (a_0 + a_1 x)$$

(x_1, y_1) ✓

$e_1 = y_1 - \hat{y}_1$

(x_2, y_2)

$e_2 = y_2 - \hat{y}_2$

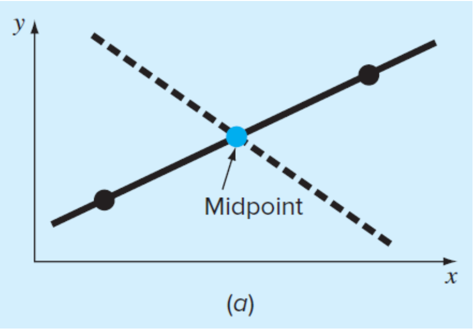
$\vdots$

e_n

One strategy for fitting a "best" line through the data would be to minimize the sum of the residual errors for all the available data, as in

$$F(a_0, a_1) = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)]$$

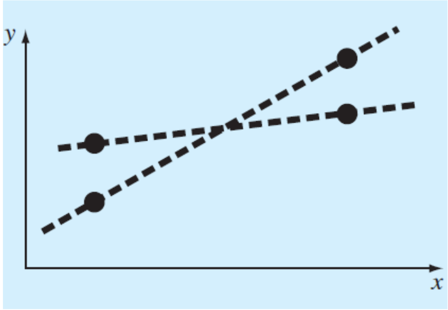
Inadequate criterion as positive and negative errors cancel



One way to remove the effect of the signs might be to minimize the sum of the absolute values of the discrepancies (**absolute deviations**), as in

$$F(a_0, a_1) = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - (a_0 + a_1 x_i)|$$

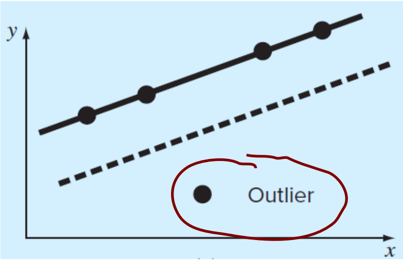
Inadequate criterion as it does not yield a unique best fit. Also, the function is not differentiable at zero.



Another strategy for fitting a best line is the *minimax* criterion, that is, to minimize the following maximization problem.

$$\underbrace{\max\{|y_i - (a_0 + a_1 x_i)| \ : \ for\ 1 \le i \le n\}}$$

This strategy is ill-suited for regression because it gives undue influence to an outlier–that is, a single point with a large error.



Linear Least-squares regression. The least squares approach to this problem involves determining the best approximating line when the error involved is the sum of the squares of the differences between the y-values on the approximating line and the given y-values. Hence, constants a_0 and a_1 must be found that minimize the least squares error:

min. $S_r(a_0, a_1)$

$a_0, a_1 \in \mathbb{R}$

$$\boxed{S_r} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [\underbrace{y_i - (a_0 + a_1 x_i)}]^2 = S_r(a_0, a_1)$$

The least squares method is the most convenient procedure for determining best linear approximations and puts substantially more weight on a point that is out of line with the rest of the data but will not permit that point to completely dominate the approximation.

$$(u^2)' = 2uu'$$

For a minimum to occur, we need both

$$\frac{\partial S_r}{\partial a_0} = \sum 2[y_i - (a_0 + a_1 x_i)](-1) = 0$$

$$\frac{\partial S_r}{\partial a_0} = 0, \qquad \frac{\partial S_r}{\partial a_1} = 0.$$

That is

$$\frac{\partial S_r}{\partial a_1} = \sum 2[y_i - (a_0 + a_1 x_i)](-x_i) = 0$$

$$\begin{cases} \dfrac{\partial S_r}{\partial a_0} = -2\displaystyle\sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)] = 0 \\ \dfrac{\partial S_r}{\partial a_1} = -2\displaystyle\sum_{i=1}^{n} x_i \underbrace{[y_i - (a_0 + a_1 x_i)]} = 0 \end{cases}$$

$$\sum y_i - \underbrace{\sum a_0} - \sum a_1 x_i = 0 \Rightarrow \begin{cases} n\,a_0 + a_1 \boxed{\sum x_i} = \sum y_i \\ a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

$$\sum x_i y_i - a_0 \sum x_i - a_1 \sum x_i^2 = 0 \mapsto$$

or

$$\begin{cases} na_0 \quad + \quad a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

The solution to this system of equations is:

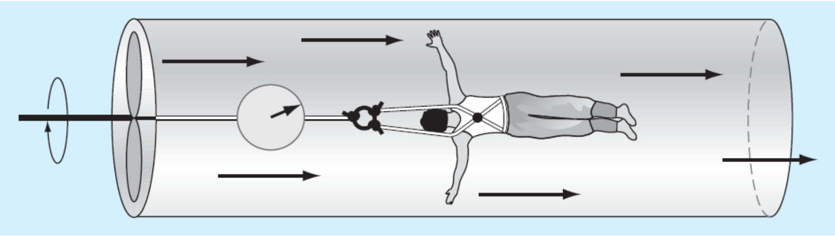
$$\begin{cases} a_1 = \dfrac{n \sum_{i=1}^{n} \overset{\swarrow}{x_i y_i} - (\overset{\swarrow}{\sum_{i=1}^{n} x_i})(\overset{\swarrow}{\sum_{i=1}^{n} y_i})}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad \checkmark \quad \text{Slope} \\ a_0 = \boxed{\bar{y}} - a_1 \boxed{\bar{x}} \quad \checkmark \quad y\text{- intercept} \end{cases}$$

Where $\underset{\substack{\downarrow \\ \text{mean of } y_i\text{'s}}}{\bar{y}} = (\sum_{i=1}^{n} y_i)/n$ and $\underset{\substack{\downarrow \\ \text{mean of } x_i\text{'s}}}{\bar{x}} = (\sum_{i=1}^{n} x_i)/n$.

Example. An individual is suspended in a wind tunnel (any volunteers?) and the force measured for various levels of wind velocity. The results are given in the following table. Fit a straight line to the values in table.

Experimental data for force (N) and velocity (m/s) from a wind tunnel experiment.

$x_i \rightarrow$	v, m/s	10	20	30	40	50	60	70	80
$y_i \rightarrow$	F, N	25	70	380	550	610	1220	830	1450



$n = 8$

Solution. Complete the following table and find a_0 and a_1.

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	10	25	100	250	625
2	20	70	400	1400	4900
3	30	380	900	11400	144400
4	40	550	1600	22000	302500
5	50	610	2500	30500	372100
6	60	1,220	3600	73200	1488400
7	70	830	4900	58100	688900
8	80	1,450	6400	116000	2102500
Σ	=360	= 5135	= 20400	= 312850	= 5104325

$$\text{Slope}: \quad a_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{(8)(312850) - (360)(5135)}{(8)(20400) - (360)^2} = 19.47$$

$$a_0 = \bar{y} - a_0\bar{x} = \frac{5135}{8} - (19.47)\frac{360}{8} = -234.28$$

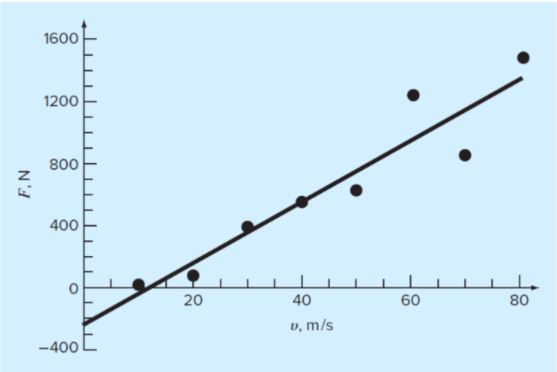
$$\boxed{\hat{y} = -234.28 + 19.47x}$$

What is the estimated value for the force if $v = 25$ m/s

$$\hat{y}(25) = -234.28 + (19.47)(25)$$

$$= \underline{\underline{252.47}}$$

Least-squares fit of a straight line to the data from Table



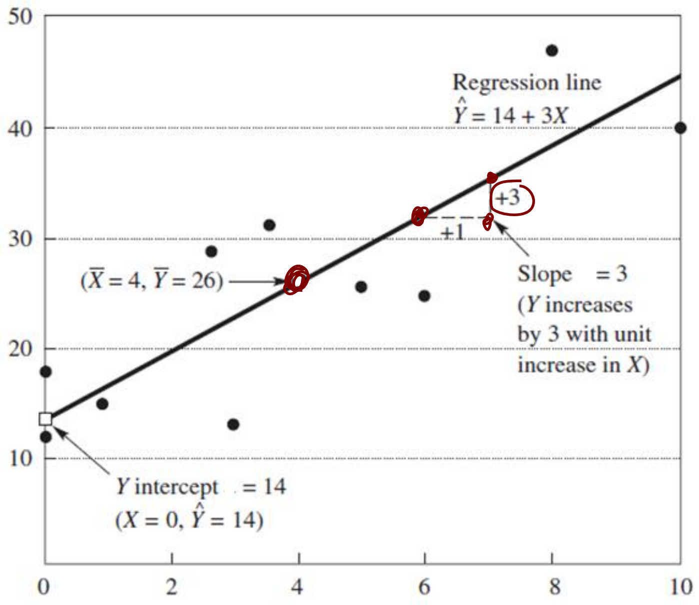
Quantification of error of linear regression and interpretation

Let

$$\hat{y} = a_0 + a_1 x$$

be the equation of the regression line. Note that

- ***y*- intercept** a_0 represents the value of y when $x = 0$
- **Slope** a_1 means that for each one unit increase in x, the average value of y is estimated by a_1 units, i.e., ***increased*** by a_1 units if a_1 is "+", or ***decreased*** by a_1 units if a_1 is "-").



Types of errors.

- The difference between the points (y) and the regression line ($\hat{y}$) is the *error* or *disturbance term* (e).

$$e = y - \hat{y}$$

- To measure predictive ability, we use *error sum of squares* (or *residual sum of squares*), denoted by SS_{error}, which is

$$SSE = SS_{\text{error}} = \sum e^2 = \sum (y - \hat{y})^2$$

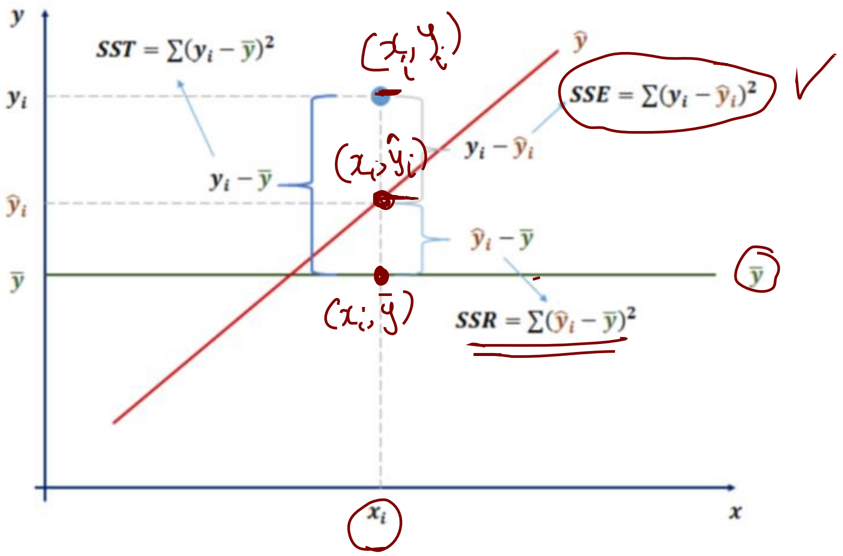
- Total sum of squares:

$$\boxed{SS_{\text{total}}} = \sum (y - \bar{y})^2$$

- Regression sum of squares (or explained sum of squares)

$$\boxed{SSR} = SS_{\text{reg}} = \underbrace{SS_{\text{total}} - \cancel{SS}_{\text{error}}}$$

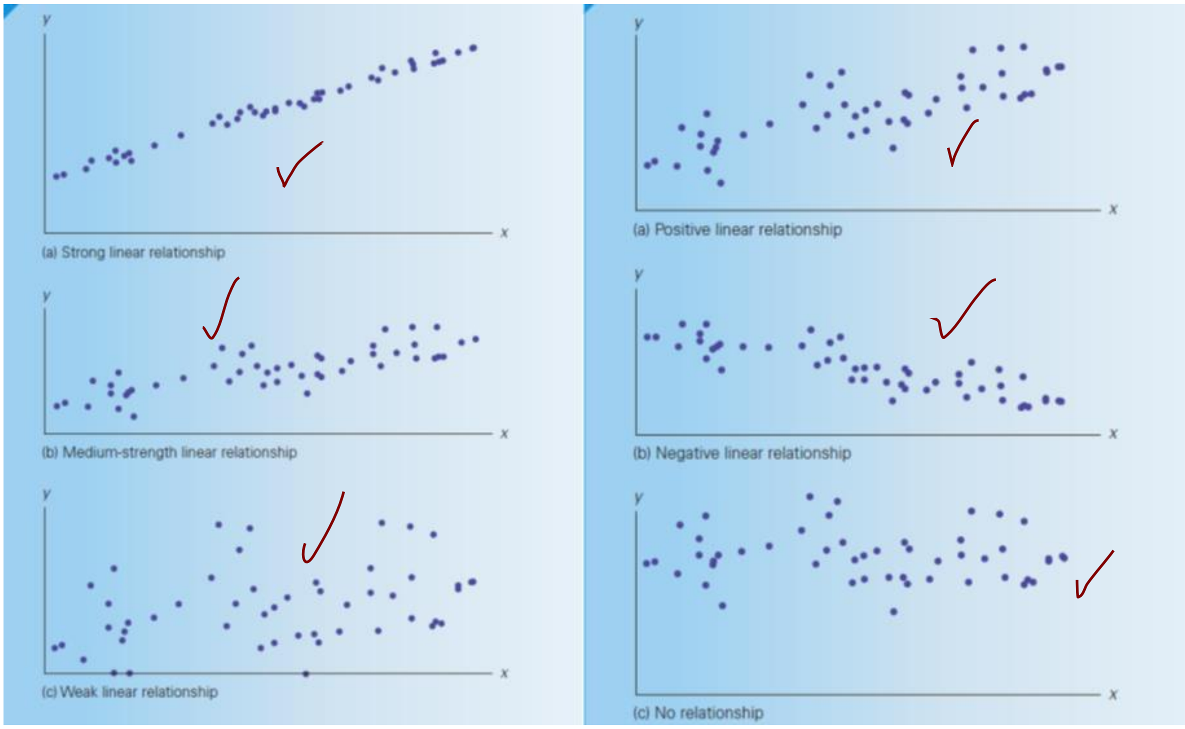
$$\boxed{\frac{SSR}{SST}} = R^2$$



Linear Correlation Coefficient.

The linear Correlation Coefficient (or Pearson's Correlation Coefficient) r, measures the direction and strength of the linear relationship between the variables x and y in a sample. It ranges between -1.0 and + 1.0

- The sign (either – or +) indicates the direction of the relationship
- Values close to zero indicate little or no correlation
- Values closer to -1 or +1, indicate stronger correlations



Linear Correlation Coefficient r can be computed using the following formula:

$$a_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\boxed{r} = \frac{n\sum(x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum \boxed{y_i^2} - (\sum y_i)^2}} \quad ✓ \qquad -1 \le r \le 1$$

The linear correlation coefficient r and the slope have the *same sign*.

- If the slope is (+), the correlation coefficient r is (+)
- If the slope is (–), the correlation coefficient r is (–).

For the previous example, we have.

$$r = \frac{(8)(312850) - (360)(5135)}{\sqrt{(8)(20400) - (360)^2}\sqrt{(8)(5104325) - (5135)^2}} = \boxed{0.94}$$

Very Strong Positive linear relationship

$$r^2 = (0.94)^2 = 0.8836$$

Coefficient of determination.

The coefficient of determination is used to check the “goodness” of the regression line. It is defined as

$$\text{Coefficient of determination} = r^2 = R^2$$

It measures the amount of variation in the dependent variable that is explained by the variation in the independent variable. In other words, it is the amount of the variation in y that is explained by the regression

$r = -0.75$

$r^2 = 0.5625$

$$r^2 = \frac{SS_{\text{total}} - SS_{\text{error}}}{SS_{\text{total}}} = \frac{SS_{\text{reg}}}{SS_{\text{total}}}$$

Thus, $1 - r^2$ is the *coefficient of nondetermination* and provides us with the amount of variation in y left unexplained.

For example, if $r = 0.8711$, then

$$\text{Coefficient of determination} = r^2 = 0.7588$$

This tells us that **75.88%** of the variation in y is **explained** by the variation in the variable x. The remaining **24.12%** (100% – 75.88% = 24.12%) is **unexplained** due to other factors.

Standard error of estimate vs standard deviation.

We define

$y_i - \bar{y}$ = deviation of y_i

$$s_{y/x} = \sqrt{\frac{S_r}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

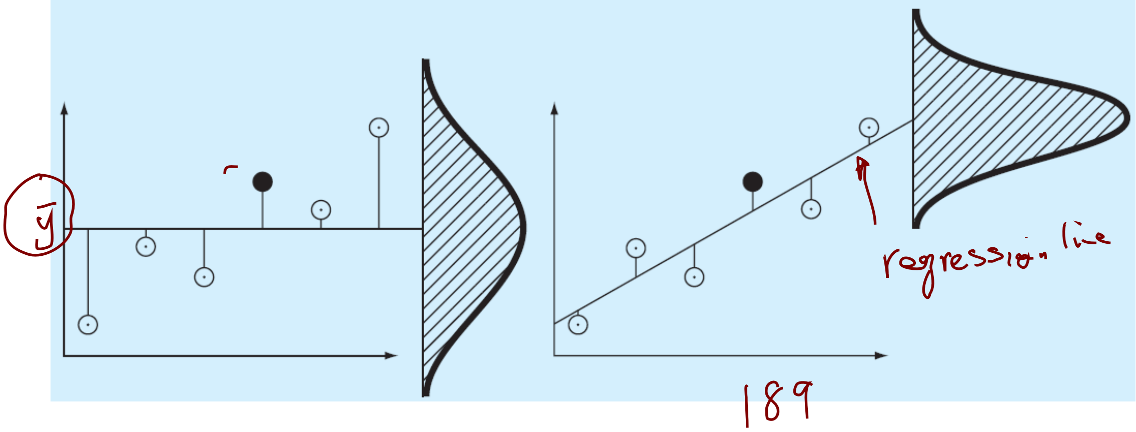
as the *standard error of estimate* that quantifies the spread around the regression line while

$y_1, y_2, \dots, y_n$

$\bar{y}$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

quantifies the spread around the mean.



508

Example. Compute the total standard deviation, the standard error of the estimate, and the correlation coefficient for the fit in previous example.

Solution. Complete the following table.

$\hat{y} = -234.28 + 19.47x$

i	x_i	y_i	$\hat{y}_i = a_0 + a_1x_i$	$(y_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$ $(y_i - a_0 - a_1x_i)^2$
1	10	25	−39.58	380540.93	4170.58
2	20	70	155.12	327046.73	7245.41
3	30	380	349.82	68581.13	910.83
4	40	550	544.52	8441.93	30.03
5	50	610	739.22	1016.33	16697.81
6	60	1,220	933.92	334222.73	81841.77
7	70	830	1128.62	35389.13	89173.90
8	80	1,450	1323.32	653057.93	16047.82
Σ	360	5,135		$\Sigma = 1808296.84$	

Correlation coefficient and coefficient of determination have been computed in page 10.

Standard deviation of y-values: $\bar{y} = \dfrac{\sum y_i}{8} = \dfrac{5135}{8} = 641.88$

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{8-1}} = \underbrace{508.26}$$ ✓

Standard error of estimate:

$$S_{y/x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{8-2}} = \sqrt{\frac{216118.16}{6}} = 189.79$$ ✓

Linearization of Nonlinear Models.

This is not always the case that the relationship between dependent and independent variables is linear. The first step in any regression analysis should be to plot and visually inspect the data to ascertain whether a linear model applies. It might be the case polynomial or other nonlinear models are suitable. Sometimes, transformations can be used to express the data in a form that is compatible with linear regression.

- *Exponential model:*

$$\boxed{y = \overset{\downarrow}{\alpha_1} e^{\overset{\downarrow}{\beta_1} x}}$$

where α_1 and β_1 are constants. This model is used in many fields of engineering and science to characterize quantities that increase (positive β_1) or decrease (negative β_1) at a rate that is directly proportional to their own magnitude (population growth model).

- *Power equation:*

$$\boxed{y = \alpha_2 x^{\beta_2}} \quad \checkmark$$

where α_2 and β_2 are constant coefficients. This model has wide applicability in all fields of engineering and science. It is very frequently used to fit experimental data when the underlying model is not known.

✓ $y = \alpha_1 e^{\beta_1 x}$

$\underbrace{\ln y} = \ln(\alpha_1 e^{\beta_1 x})$

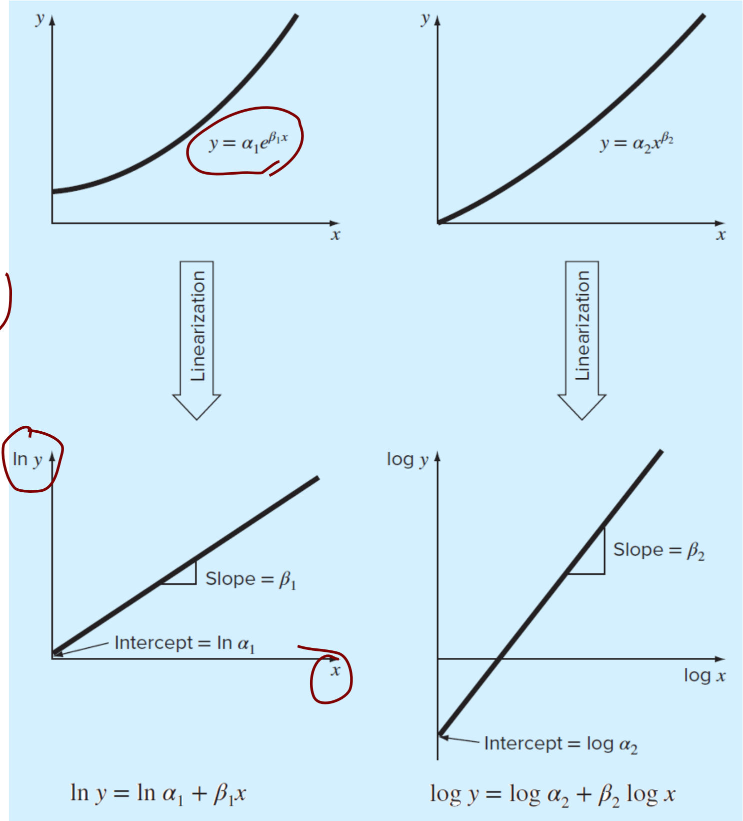
$= \ln(\alpha_1) + \underbrace{\ln(e^{\beta_1 x})}$

$= \underbrace{\ln(\alpha_1)}_{A} + \beta_1 x$

$\boxed{Y} = \boxed{A} + \underset{B}{\boxed{\beta_1}} x$

$\alpha_1 = e^{A}$

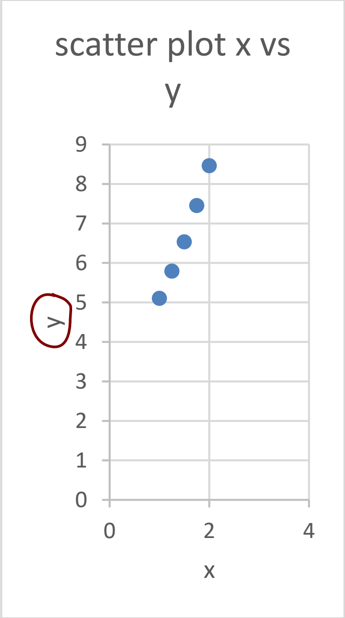
✓ x_i	y_i	✓ $\ln(y_1)$

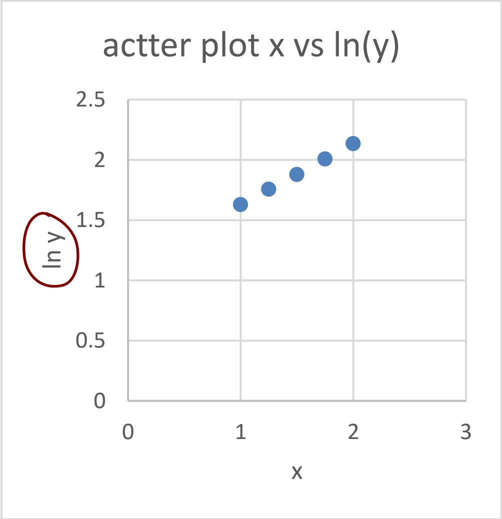


Example. Consider the collection of data in the following table.

i	x_i	y_i
1	1.00	5.10
2	1.25	5.79
3	1.50	6.53
4	1.75	7.45
5	2.00	8.46

The scatter plot shows that the data are exponentially related.





(x_i, y_i)

$e_i = y_i - \hat{y}_i$

$E = \sum e_i^2$

Direct approach:

The relationship requires the approximating function to be of the form

$$\boxed{y = be^{ax}} \quad ✓ \; ✓$$

Thus, for m points (x_i, y_i), we need the function

$$E = \sum_{i=1}^{m} (y_i - be^{ax_i})^2, = E(a,b) \quad ✓$$

to be minimized with respect to a and b. This means that

$$\begin{cases} 0 = \dfrac{\partial E}{\partial a} = 2 \displaystyle\sum_{i=1}^{m} (y_i - be^{ax_i})(-bx_i e^{ax_i}) \\ 0 = \dfrac{\partial E}{\partial \overset{b}{a}} = 2 \displaystyle\sum_{i=1}^{m} (y_i - be^{ax_i})(-bx_i e^{ax_i}) \end{cases}$$

No exact solution to this system in a and b can generally be found.

Alternative approach:

The method that is commonly used when the data are suspected to be exponentially related is to consider the logarithm of the approximating equation:

$$\ln y = \ln b + ax$$

In this case, a linear problem now appears, and solutions for *ln b* and *a* can be obtained by applying linear least squares approach.

For the above table, we need to complete the following table and use the formulas to find a and $\ln b$.

$y = ae^{bx}$

$\begin{cases} Y = \ln(y) \\ X = x \end{cases}$

$\hat{Y} = A + BX$

	X_i		Y_i	X_i^2	X_iY_i
i	x_i	y_i	$\ln y_i$	x_i^2	$x_i \ln y_i$
1	1.00	5.10	1.63	1	1.63
2	1.25	5.79	1.76	1.56	2.2
3	1.50	6.53	1.88	2.25	2.82
4	1.75	7.45	2.01	3.06	3.52
5	2.00	8.46	2.14	4	4.28
	7.5	X	9.42	11.87	14.45

$$\text{Slope: } a_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{(5)(14.45) - (7.5)(9.42)}{(5)(11.87) - (7.5)^2} = 0.52$$

$$\text{y-intercept: } a_0 = \bar{Y} - a_1\bar{X} = \frac{9.42}{5} - (0.52)\frac{7.5}{5} = 1.1$$

$$\hat{Y} = A + BX = 1.1 + 0.52X$$

$$\begin{cases} b = B \longrightarrow b = 0.52 \\ A = \ln(a) \longrightarrow 1.1 = \ln(a) \longrightarrow a = e^{1.1} \simeq 3 \end{cases}$$

$$\boxed{\hat{y} = ae^{bx} = 3e^{0.52x}}$$

References

1. Chapra, Steven C. (2018). *Numerical Methods with* MATLAB *for Engineers and Scientists*, 4th Ed. McGraw Hill.
2. Burden, Richard L., Faires, J. Douglas (2011). *Numerical Analysis*, 9th Ed. Brooks/Cole Cengage Learning