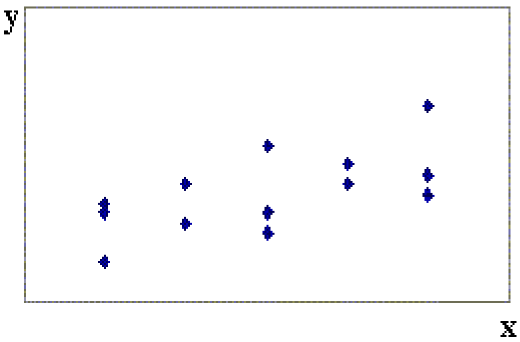
Simple Linear Regression

Sometimes an experiment is set up where the experimenter has control over the values of one or more variables X and measures the resulting values of another variable Y, producing a field of observations.

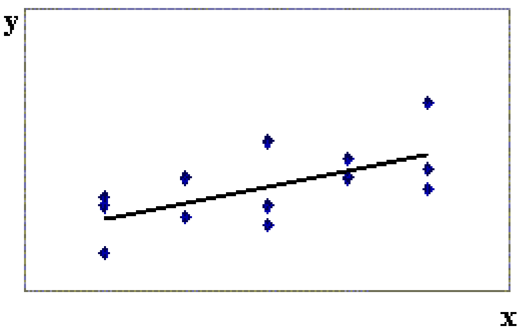


The question then arises: What is the best line (or curve) to draw through this field of points?

Values of X are controlled by the experimenter, so the non-random variable x is called the **controlled** variable or the **independent** variable or the **regressor.**

Values of Y are random, but are influenced by the value of x. Thus Y is called the **dependent** variable or the **response** variable.

We want a "line of best fit" so that, given a value of x, we can predict the value of Y for that value of x.



The **simple linear regression model** is that the **predicted value** of y is

$$\hat{y} = \beta_0 + \beta_1 x$$

and that the **observed value** of Y is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where ε_i is the **error**.

It is assumed that the errors are normally distributed as $\varepsilon_i \sim \text{N}(0, \sigma^2)$, with a constant variance σ^2. The point estimates of the errors ε_i are the **residuals** $e_i = y_i - \hat{y}_i$.

With the assumptions

1) $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

2) $x = x_0 \Rightarrow Y \sim \text{N}(\beta_0 + \beta_1 x_0, \sigma^2)$ **[$\Rightarrow$ (3) V[*Y*] is ind't of *x*]**

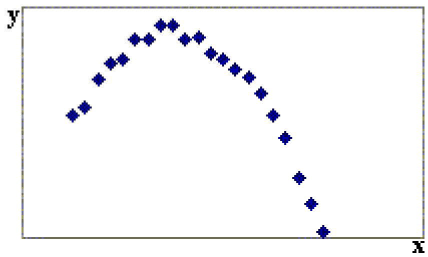
in place, it then follows that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of the coefficients β_0 and β_1.

$$\text{E}\left[\hat{\beta}_0 + \hat{\beta}_1 x\right] = \beta_0 + \beta_1 x \qquad (\text{note lower case } x)$$

Methods for dealing with non-linear regression are available in the course text, but are beyond the scope of this course.

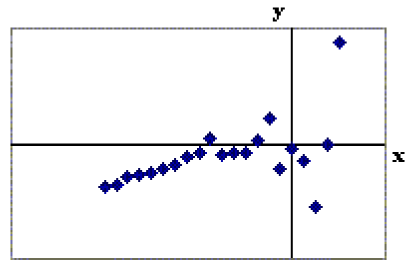
Examples illustrating violations of the assumptions in the simple linear regression model:

1.



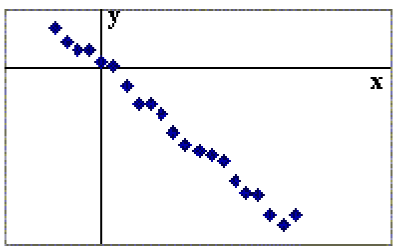
(1) violated – not linear

2.



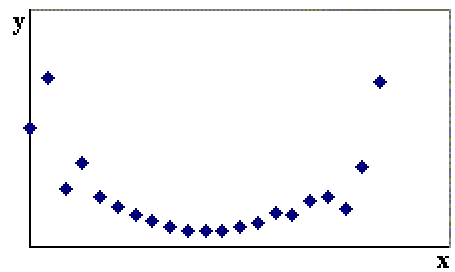
(3) violated – variance not constant

3.



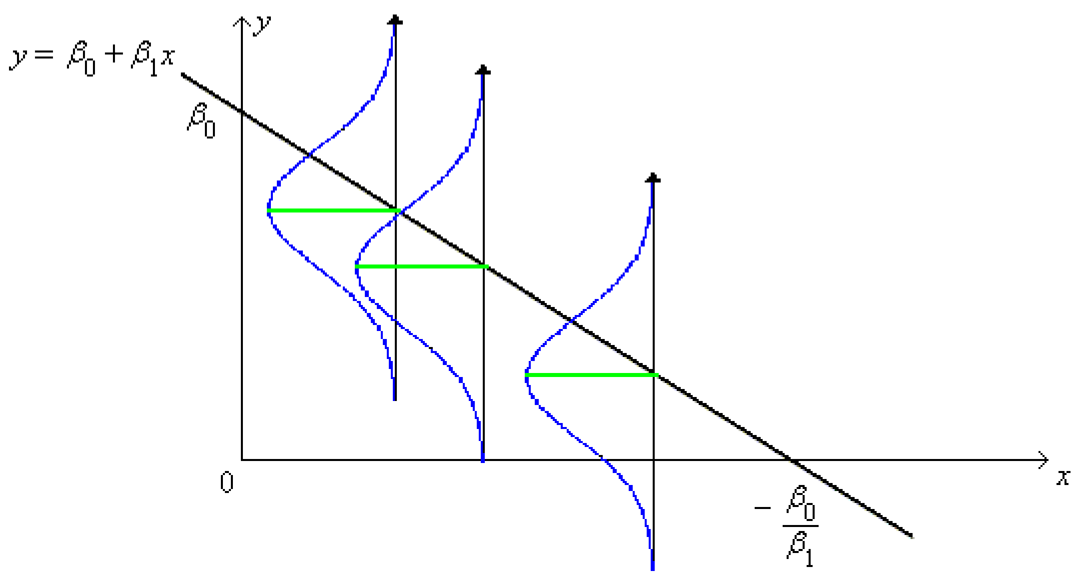
OK

4.



(1) & (3) violated

If the assumptions are true, then the probability distribution of $Y \mid x$ is $N(\beta_0 + \beta_1 x, \sigma^2)$.



Example 12.01

Given that $Y_i = 10 - 0.5\,x_i + \varepsilon_i$, where $\varepsilon_i \sim \text{N}(0, 2)$, find the probability that the observed value of y at $x = 8$ will exceed the observed value of y at $x = 7$.

$Y_i \sim \text{N}(10 - 0.5\,x_i, 2)$

Let Y_7 = the observed value of y at $x = 7$
and Y_8 = the observed value of y at $x = 8$,
then

$Y_7 \sim \text{N}(\mathbf{6.5, 2})$ and $Y_8 \sim \text{N}(\mathbf{6, 2})$

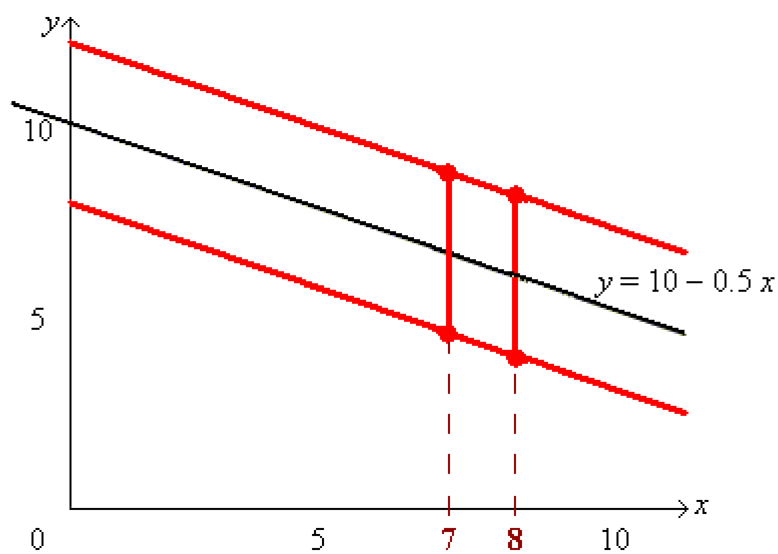
$\Rightarrow$ $Y_8 - Y_7 \sim \text{N}(\mathbf{6 - 6.5, 2 + 2})$

$\mu = \mathbf{-0.5}$ $\sigma = \sqrt{\mathbf{4}} = \mathbf{2}$

$$\text{P}\left[Y_8 - Y_7 > 0\right] = \text{P}\left[Z > \frac{0 - (-0.5)}{2}\right] = 1 - \Phi(0.25) \approx .4013$$

Despite $\beta_1 < 0$, P[$Y_8 > Y_7$] > 40% !

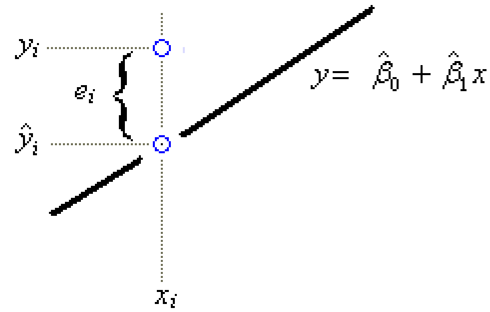
For any x_i in the range of the regression model, more than 95% of all Y_i will lie within 2σ ($= 2\sqrt{2}$) either side of the regression line.



Derivation of the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ of the regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$:

We need to minimize the errors.

Each error is estimated by the observed residual $e_i = y_i - \hat{y}_i$.



Minimize errors.

$\sum |e_i|$? NO

Use the SSE (sum of squares due to errors)

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = f\left(\hat{\beta}_o, \hat{\beta}_1\right)$$

Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that $\dfrac{\partial S}{\partial \hat{\beta}_0} = \dfrac{\partial S}{\partial \hat{\beta}_1} = 0$.

[Note: $\hat{\beta}_0, \hat{\beta}_1$ are <u>variables</u>, while x, y are constants.]

$$\frac{\partial S}{\partial \hat{\beta}_0} = 2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(0-1-0) = 0 \quad \Rightarrow \quad \hat{\beta}_0 \sum 1 + \hat{\beta}_1 \sum x = \sum y \qquad \textbf{(1)}$$

and

$$\frac{\partial S}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(0-0-x_i) = 0 \quad \Rightarrow \quad \hat{\beta}_0 \sum x + \hat{\beta}_1 \sum x^2 = \sum xy \qquad \textbf{(2)}$$

or, equivalently, $$\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} \qquad \textbf{(3)}$$

$$\Rightarrow \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} = \frac{1}{n\, S_{xx}} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} \qquad \textbf{(4)}$$

The solution to the linear system of two **normal equations (1)** and **(2)** is:
from the lower row of matrix equation **(4)**:

$$\boxed{\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}}, \quad \text{(where } \boxed{n S_{xy} = n\sum xy - \sum x \sum y}$$

$$\text{and} \quad n S_{xx} = n\sum x^2 - \left(\sum x\right)^2 \text{)}$$

or, equivalently, $\hat{\beta}_1 = \dfrac{\text{sample covariance of } (x, y)}{\text{sample variance of } x}$;

and, from equation **(1)**: $\boxed{\hat{\beta}_0 = \frac{1}{n}\left(\sum y - \hat{\beta}_1 \sum x\right)}$.

A form that is less susceptible to round-off errors (but less convenient for manual computations) is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

The regression line of Y on x is $\boxed{y - \bar{y} = \hat{\beta}_1 (x - \bar{x})}$.

Equation **(1)** guarantees that all simple linear regression lines pass through the centroid $(\bar{x}, \bar{y})$ of the data.

It turns out that the simple linear regression method remains valid even if the values of the regressor x are also random.

However, note that interchanging x with y, (so that Y is the regressor and X is the response), results in a *different* regression line (unless X and Y are perfectly correlated).
.

Example 12.02
(the same data set as Example 11.06: paired two sample t test)

Nine volunteers are tested before and after a training programme. Find the line of best fit for the posterior (after training) scores as a function of the prior (before training) scores.

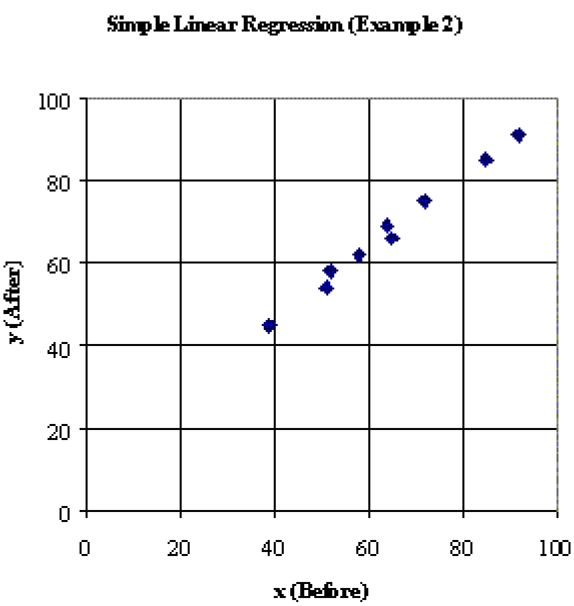
Volunteer:	1	2	3	4	5	6	7	8	9
After training:	75	66	69	45	54	85	58	91	62
Before training:	72	65	64	39	51	85	52	92	58

Let Y = score after training and X = score before training.

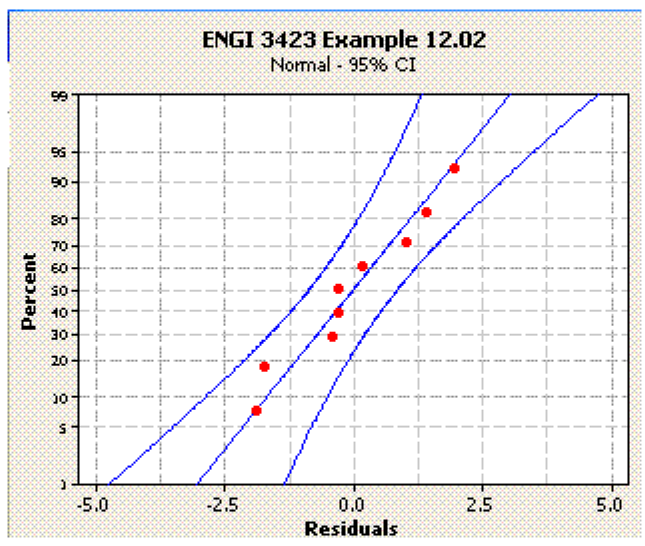
In order to use the simple linear regression model, the assumptions

$$Y_i \;=\; \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$x = x_0 \;\Rightarrow\; Y \sim \mathrm{N}(\beta_0 + \beta_1 x_0\,,\, \sigma^2)$$

must hold.
From a plot of the data
(in `http://www.engr.mun.ca/~ggeorge/3423/demos/regress2.xls`),
and `http://www.engr.mun.ca/~ggeorge/3423/demos/ex1202.mpj`),
one can see that the assumptions are reasonable.



Scatter plot



Normal probability plot of residuals

Calculations:

i	$\boldsymbol{x_i}$	$\boldsymbol{y_i}$	$\boldsymbol{x_i^2}$	$\boldsymbol{x_i \cdot y_i}$	$\boldsymbol{y_i^2}$
1	72	75	5184	5400	5625
2	65	66	4225	4290	4356
3	64	69	4096	4416	4761
4	39	45	1521	1755	2025
5	51	54	2601	2754	2916
6	85	85	7225	7225	7225
7	52	58	2704	3016	3364
8	92	91	8464	8372	8281
9	58	62	3364	3596	3844
Sum:	**578**	**605**	**39384**	**40824**	**42397**

$$nS_{xy} = n\sum xy - \sum x \sum y = 9 \times 40824 - 578 \times 605 = \mathbf{17726}$$

[Note: $\boldsymbol{n\,S_{xy}} = \boldsymbol{n\,(n-1)}$ * sample covariance of ($\boldsymbol{X, Y}$)]

$$nS_{xx} = n\sum x^2 - \left(\sum x\right)^2 = 9 \times 39384 - 578^{\,2} = \mathbf{20372}$$

[Note: $\boldsymbol{n\,S_{xx}} = \boldsymbol{n\,(n-1)}$ * sample variance of $\boldsymbol{X}$]

$$\Rightarrow \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{17726}{20372} = \underline{\underline{\mathbf{0.870116}}}$$

and $$\hat{\beta}_0 = \frac{1}{n}\left(\sum y - \hat{\beta}_1 \sum x\right) = \frac{1}{9}\left(605 - 0.807116 \times 578\right) = \underline{\underline{\mathbf{11.34145}}}$$

Each predicted value $\hat{y}_i$ of Y is then estimated using $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \approx 11.34 + 0.87\,x$ and the point estimates of the unknown errors ε_i are the observed residuals $e_i = y_i - \hat{y}_i$. [Use **un-rounded** values 11.34**...** and 0.87**...** to find residuals.]

A measure of the degree to which the regression line fails to explain the variation in Y is the sum of squares due to error,

$$SSE = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

which is given in the adjoining table.

x_i	y_i	$\hat{y}_i$	e_i	e_i^2
72	75	73.98979	1.0102	1.0205
65	66	67.89898	−1.8990	3.6061
64	69	67.02886	1.9711	3.8854
39	45	45.27597	−0.2760	0.0762
51	54	55.71736	−1.7174	2.9493
85	85	85.30130	−0.3013	0.0908
52	58	56.58747	1.4125	1.9952
92	91	91.39211	−0.3921	0.1537
58	62	61.80817	0.1918	0.0368
			SSE =	**13.8141**

An Alternative Formula for *SSE*:

$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x \quad \Rightarrow$$

$$SSE = \sum_{i=1}^{n}\left(y_i - \left(\bar y - \hat\beta_1 \bar x\right) - \hat\beta_1 x_i\right)^2 = \sum_{i=1}^{n}\left(\left(y_i - \bar y\right) - \hat\beta_1\left(x_i - \bar x\right)\right)^2$$

$$= \sum_{i=1}^{n}\left(y_i - \bar y\right)^2 - 2\hat\beta_1 \sum_{i=1}^{n}\left(x_i - \bar x\right)\left(y_i - \bar y\right) + \hat\beta_1^{\,2} \sum_{i=1}^{n}\left(x_i - \bar x\right)^2$$

$$= S_{yy} - 2\hat\beta_1 S_{xy} + \hat\beta_1^{\,2} S_{xx}$$

But $\hat\beta_1 = \dfrac{S_{xy}}{S_{xx}}$

$$\Rightarrow \quad SSE = S_{yy} - \hat\beta_1 S_{xy} \quad \text{or} \quad SSE = \frac{S_{xx}S_{yy} - S_{xy}^{\;2}}{S_{xx}} \quad \text{or}$$

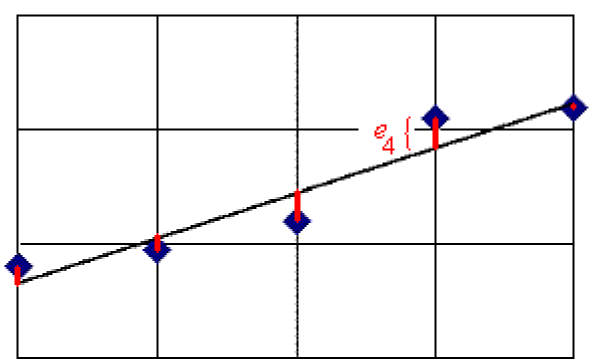
$$\boxed{SSE = \frac{\left(nS_{xx}\right)\left(nS_{yy}\right) - \left(nS_{xy}\right)^2}{n \times \left(nS_{xx}\right)}}$$

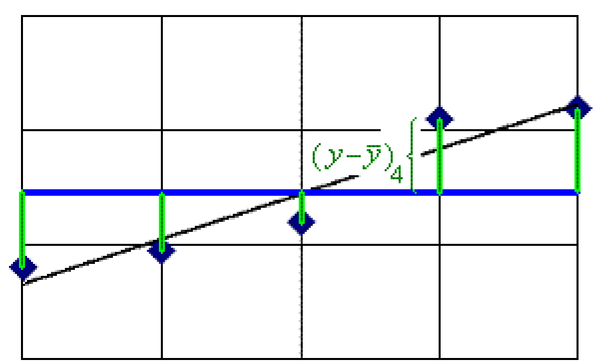
In this example,

$$SSE = \frac{20\,372 \times 15\,548 - 17\,726^2}{9 \times 20\,372} = 13.814\ldots$$

However, this formula is *very* sensitive to round-off errors:
If all terms are rounded off prematurely to three significant figures, then

$$SSE = \frac{20\,400 \times 15\,500 - 17\,700^2}{9 \times 20\,400} = 15.85 \quad (2\ \text{d.p.})$$





$$SSE = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n}\left(y_i - \hat y_i\right)^2 \qquad SST = \sum_{i=1}^{n}\left(y_i - \bar y\right)^2$$

The total variation in Y is the SST (sum of squares - total):

$$\boxed{SST = \frac{n\, S_{yy}}{n} = \sum \left(y_i - \bar{y}\right)^2}$$ (which is $(n-1) \times$ the sample variance of y).

In this example, $SST = 15\,548 / 9 =$ **<u>1 727.555...</u>**

The total variation (SST) can be partitioned into the variation that can be explained by the regression line $\left(SSR = \sum \left(\hat{y}_i - \bar{y} \right)^2 \right)$ and the variation that remains unexplained by the regression line (SSE).

$$\begin{array}{ccccc} SST & = & SSR & + & SSE \\ \uparrow & & \uparrow & & \\ S_{yy} & & \hat{\beta}_1 S_{xy} & & \end{array}$$

The proportion of the variation in Y that is explained by the regression line is known as the **coefficient of determination**

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

In this example, $r^2 = 1 - (13.81\ldots / 1727.555\ldots) = .992004\ldots$

Therefore the regression model in this example explains 99.2% of the total variation in y.

Note:

$$SSR = \hat{\beta}_1 \cdot S_{xy} = \frac{{S_{xy}}^2}{S_{xx}}$$

and $SST = S_{yy}$

$\Rightarrow$

$$\boxed{r^2 = \frac{{S_{xy}}^2}{S_{xx}\, S_{yy}}}$$

The coefficient of determination is just the square of the sample correlation coefficient r. Thus $r = \sqrt{r^2} \approx .996$. It is no surprise that the two sets of test scores in this example are very strongly correlated. Most of the points on the graph are very close to the regression line $y = 0.87\,x + 11.34$.

A point estimate of the unknown population variance σ^2 of the errors ε is the sample variance or **mean square error** $s^2 = MSE = SSE$ / (number of degrees of freedom).

But the calculation of s^2 includes two parameters that are estimated from the data: $\hat{\beta}_0$ and $\hat{\beta}_1$. Therefore two degrees of freedom are lost and $\boxed{MSE = \dfrac{SSE}{n-2}}$. In this example, $MSE \approx 1.973$.

A concise method of displaying some of this information is the **ANOVA table** (used in Chapters 10 and 11 of Devore for analysis of variance). The f value in the top right corner of the table is the square of a t value that can be used in an **hypothesis test** on the value of the slope coefficient β_1.

Sequence of manual calculations:
$\{ n, \sum x, \sum y, \sum x^2, \sum xy, \sum y^2 \} \rightarrow \{ n\,S_{xx}, n\,S_{xy}, n\,S_{yy} \} \rightarrow$
$\{ \hat{\beta}_1, \hat{\beta}_0, SSR, SST \} \rightarrow \{ R^2, SSE \} \rightarrow \{ MSR, MSE \} \rightarrow f \rightarrow t$

Source	Degrees of Freedom	Sums of Squares	Mean Squares	f
Regression	1	SSR = 1713.741...	MSR = SSR / 1 = 1713.741...	= MSR/MSE = 868.4...
Error	$n-2$ = 7	SSE = 13.81...	MSE = SSE / (n–2) = 1.973...	
Total	$n-1$ = 8	SST = 1727.555...		

To test $\mathcal{H}_o : \beta_1 = 0$ (no useful linear association) against $\mathcal{H}_a : \beta_1 \neq 0$ (a useful linear association exists), we compare $|t| = \sqrt{f}$ to $t_{\alpha/2,\,(n-2)}$.

In this example, $|t| = \sqrt{868.4...} = 29.4... \gg t_{.0005,\,7}$ (the p-value is $< 10^{-7}$)
so we reject $\mathcal{H}_o$ in favour of $\mathcal{H}_a$ at any reasonable level of significance α.

The standard error s_b of $\hat{\beta}_1$ is $s / \sqrt{S_{xx}}$ so the t value is also equal to $\boxed{\dfrac{\hat{\beta}_1 - 0}{\sqrt{\dfrac{n\,MSE}{n\,S_{xx}}}}}$.

Yet another alternative test of the significance of the linear association is an hypothesis test on the population correlation coefficient ρ, ($\mathcal{H}_o$: $\rho = 0$ vs. $\mathcal{H}_a$: $\rho \neq 0$), using the test statistic $\boxed{t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$, which is entirely equivalent to the other two t statistics above.

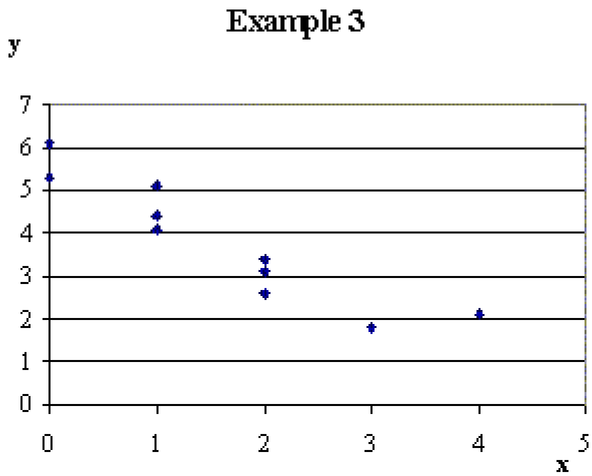
<u>Example 12.03</u>

(a) Find the line of best fit to the data

x	0	0	1	1	1	2	2	2	3	4
y	6.1	5.3	4.1	5.1	4.4	3.4	2.6	3.1	1.8	2.1

(b) Estimate the value of y when $x = 2$.
(c) Why can't the regression line be used to estimate y when $x = 10$?
(d) Find the sample correlation coefficient.
(e) Does a useful linear relationship between Y and x exist?

(a) A plot of these data follows.



The Excel spreadsheet file for these data can be found at

“`http://www.engr.mun.ca/~ggeorge/3423/demos/regress3.xls`”.

The summary statistics are

$\Sigma x = 16$ $\Sigma y = 38$ $n = 10$

$\Sigma x^2 = 40$ $\Sigma xy = 45.6$ $\Sigma y^2 = 163.06$

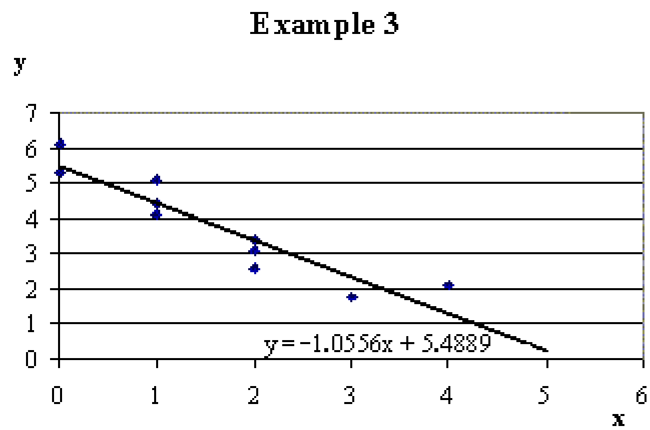
From which

$$n\,S_{xy} = n\,\Sigma\,xy - \Sigma\,x\,\Sigma\,y = -152$$

$$n\,S_{xx} = n\,\Sigma\,x^2 - (\Sigma\,x)^2 = 144 \qquad n\,S_{yy} = n\,\Sigma\,y^2 - (\Sigma\,y)^2 = 186.6$$

$$\Rightarrow \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-152}{144} = -1.0\dot{5}$$

and $$\hat{\beta}_0 = \frac{\sum y - \hat{\beta}_1 \sum x}{n} = 5.4\dot{8}$$



So the regression line is

$y = 5.489 - 1.056\,x$ **(3 d.p.)**

(b) $x = 2 \Rightarrow$ $y = \mathbf{5.488... - 1.055... \times 2} = \underline{\underline{\mathbf{3.38}}}$ **(2 d.p.)**

(c) $x = 10 \Rightarrow$ $y = \mathbf{5.488... - 1.055... \times 10} = \underline{\underline{\mathbf{-5.07}}}$ **< 0 ! (2 d.p.)**

Problem: ***x* = 10 is outside the sample range for *x*.**

$\Rightarrow$ **SLR model may be invalid. In one word: EXTRAPOLATION.**

(d) $$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{-152}{\sqrt{144 \times 186.6}} = -.92727... \approx \underline{\underline{\mathbf{-.93}}}$$

(e) $$SSR = \frac{(n\,S_{xy})^2}{n\,(n\,S_{xx})} = \frac{(-152)^2}{10 \times 144} = 16.0\dot{4}$$

$$SST = S_{yy} = (\,186.6 / 10\,) = 18.66$$

and $SSE = SST - SSR = 18.66 - 16.04... = 2.615...$

The ANOVA table is then:

Source	d.f.	SS	MS	f
R	**1**	16.04444...	**16.04444...**	**49.07...**
E	**8**	**2.61555...**	**0.3269...**	
T	**9**	18.66000		

from which $t = -\sqrt{f} \approx$ $\mathbf{-7.005}$ (3 d.p.) But $t_{.0005,8} = 5.041...$ $\mathbf{|t| > t_{.0005,\,8}}$

Therefore reject $\mathcal{H}_o : \beta_1 = 0$ in favour of $\mathcal{H}_a : \beta_1 \neq 0$ at any reasonable level of significance α. **[p-value = .00011...]**

OR $\quad t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{-.92727...\times\sqrt{8}}{\sqrt{1-.85983...}} \approx -7.005$

$\Rightarrow$ reject $\mathcal{H}_o$: $\rho = 0$ in favour of $\mathcal{H}_a$: $\rho \neq 0$ (a significant linear association exists).

[Also, from the ANOVA table, $r^2 = \dfrac{SSR}{SST} = \dfrac{16.04\ldots}{18.66} \approx .8598$

Therefore the regression line explains ~86% of the variation in *y*.
$\mathbf{r = -\sqrt{r^2} = -.927}$ **, as before.]**

<u>Confidence and Prediction Intervals</u>

The simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ leads to a line of best fit in the least squares sense, which provides an expected value of Y given a value for x :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \mathrm{E}[\,Y \,|\, x\,] = \mu_{Y\bullet x}\,.$$

The uncertainty in this expected value has two components:

- the square of the standard error of the scatter of the observed points about the regression line ($= \sigma^2 / n$), and
- the uncertainty in the position of the regression line itself, which increases with the distance of the chosen x from the centroid of the data but decreases with increasing spread of the full set of x values: $\sigma^2\left(\dfrac{(x-\bar{x})^2}{S_{xx}}\right)$.

The unknown variance σ^2 of individual points about the true regression line is estimated by the mean square error $s^2 = MSE$.

Thus a 100(1−α)% **confidence interval** for the expected value of Y at $x = x_o$ has endpoints at

$$\boxed{\left(\hat{\beta}_0 + \hat{\beta}_1 x_o\right) \pm t_{\alpha/2,(n-2)}\, s \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}}}$$

The **prediction error** for a single point is the residual $E = Y - \hat{y}$, which can be treated as the difference of two independent random variables. The variance of the prediction error is then

$$\mathrm{V}[E] = \mathrm{V}[Y] + \mathrm{V}[\hat{y}] = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{n(x_o - \bar{x})^2}{n\,S_{xx}}\right)$$

Thus a 100(1−α)% **prediction interval** for a single future observation of Y at $x = x_o$ has endpoints at

$$\boxed{\left(\hat{\beta}_0 + \hat{\beta}_1 x_o\right) \pm t_{\alpha/2,(n-2)}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}}}$$

The prediction interval is always wider than the confidence interval.

<u>Example 12.03</u> (continued)

(f) Find the 95% confidence interval for the expected value of Y at $x = 2$ and $x = 5$.
(g) Find the 95% prediction interval for a future value of Y at $x = 2$ and at $x = 5$.

(f) $\alpha = 5\% \Rightarrow \alpha/2 = .025$

Using the various values from parts (a) and (e):

$n = 10$ $\quad t_{.025,\,8} = 2.306...$ $\quad s = 0.57179...$ $\quad \bar{x} = 1.6$

$S_{xx} = 14.4$ $\quad \hat{\beta}_0 = 5.4888...$ $\quad \hat{\beta}_1 = -1.0555...$

$x_o = 2 \Rightarrow$ the 95% CI for $\mu_{Y|2}$ is

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_o\right) \pm t_{\alpha/2,(n-2)}\, s \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}} = 3.3777... \pm 1.3185... \times \sqrt{0.1111...}$$

$= 3.3777... \pm 0.4395... \Rightarrow$ <u>$2.94 \le \mathrm{E}[\,Y \mid 2\,] < 3.82$</u> (to 3 s.f.)

$x_o = 5 \Rightarrow$ the 95% CI for $\mu_{Y|5}$ is

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_o\right) \pm t_{\alpha/2,(n-2)}\, s \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}} = 0.2111\ldots \pm 1.3185\ldots \times \sqrt{0.902777\ldots}$$

$= 0.2111\ldots \pm 1.2528\ldots \Rightarrow \underline{-1.04 \le \mathrm{E}[\,Y \mid 5\,] \le 1.46}$ (to 3 s.f.)

(g) $x_o = 2 \Rightarrow$ the 95% PI for Y is

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_o\right) \pm t_{\alpha/2,(n-2)}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}} = 3.3777\ldots \pm 1.3185\ldots \times \sqrt{1.1111\ldots}$$

$= 3.3777\ldots \pm 1.3898\ldots \Rightarrow \underline{1.99 \le Y < 4.77}$ (to 3 s.f.) at $x = 2$

$x_o = 5 \Rightarrow$ the 95% PI for Y is

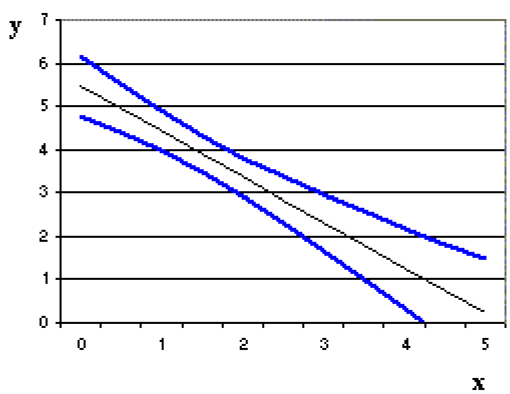
$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_o\right) \pm t_{\alpha/2,(n-2)}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}} = 0.2111\ldots \pm 1.3185\ldots \times \sqrt{1.902777\ldots}$$

$= 0.2111\ldots \pm 1.8188\ldots \Rightarrow \underline{-1.61 < Y < 2.03}$ (to 3 s.f.) at $x = 5$

Note how the confidence and prediction intervals both become wider the further away from the centroid the value of x_o is. The two intervals at $x = 5$ are wide enough to cross the x-axis, which is an illustration of the dangers of **extrapolation** beyond the range of x for which data exist.

Sketch of confidence and prediction intervals for Example 3 (f) and (g):

(f) 95% Confidence Intervals



(g) 95% Prediction Intervals