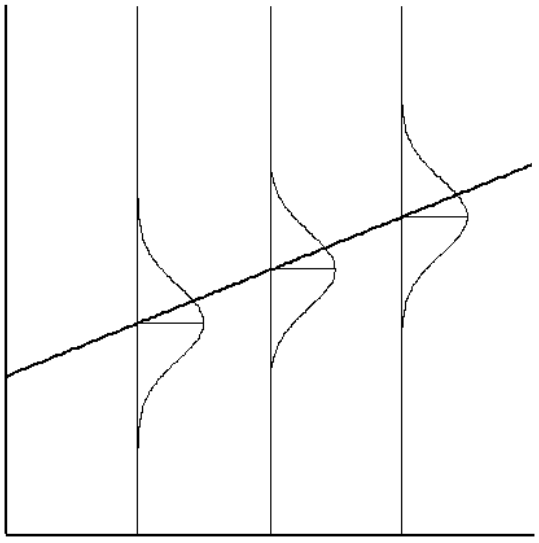
Examples for 09/03/2015

The (normal) simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

where ε_i's are independent Normal$(0, \sigma^2)$ (iid Normal$(0, \sigma^2)$).

β_0, β_1, and σ^2 are unknown model parameters.



Least-Squares Approach: minimize $\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$

$$SXX = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$SXY = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

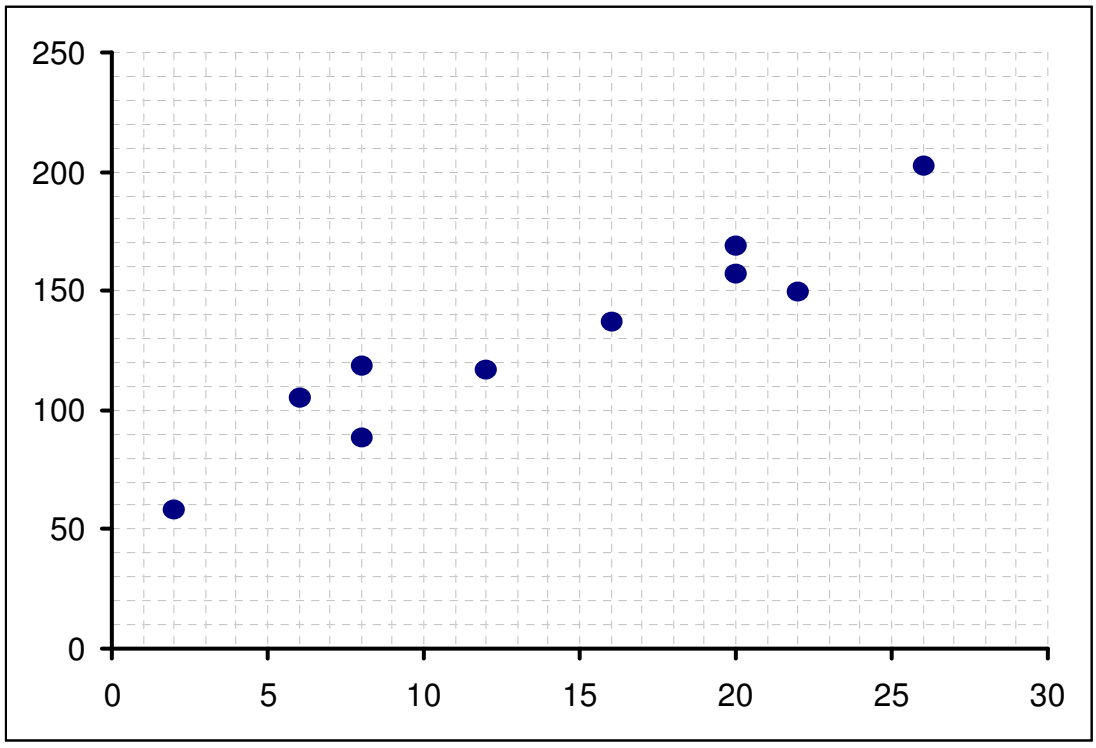
$$SYY = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

Slope $\hat{\beta}_1 = \dfrac{SXY}{SXX}$

Y-intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

1. The owner of *Momma Leona's Pizza* restaurant chain believes that if a restaurant is located near a college campus, then there is a linear relationship between sales and the size of the student population. Suppose data were collected from a sample of 10 *Momma Leona's Pizza* restaurants located near college campuses. For the ith restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars). The values of x_i and y_i for the 10 restaurants in the sample are summarized in the following table:

Restaurant	Student Population (1000s)	Quarterly Sales ($1000s)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202



a) Find the equation of the least-squares regression line.

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})\cdot(y-\bar{y})$	$(y-\bar{y})^2$
	2	58	−12	−72	144	864	5184
	6	105	−8	−25	64	200	625
	8	88	−6	−42	36	252	1764
	8	118	−6	−12	36	72	144
	12	117	−2	−13	4	26	169
	16	137	2	7	4	14	49
	20	157	6	27	36	162	729
	20	169	6	39	36	234	1521
	22	149	8	19	64	152	361
	26	202	12	72	144	864	5184
Total:	140	1300	0	0	568	2840	15730
	Σx	Σy			*SXX*	*SXY*	*SYY*

$$\bar{x} = \frac{\sum x}{n} = \frac{140}{10} = 14 \qquad \bar{y} = \frac{\sum y}{n} = \frac{1300}{10} = 130$$

$$\hat{\beta}_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{2840}{568} = \mathbf{5}.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 130 - 5 \cdot 14 = \mathbf{60}.$$

The least-squares regression line: $\hat{y} = 60 + 5 \cdot x.$

OR

	x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})\cdot y$
	2	58	−12	144	−696
	6	105	−8	64	−840
	8	88	−6	36	−528
	8	118	−6	36	−708
	12	117	−2	4	−234
	16	137	2	4	274
	20	157	6	36	942
	20	169	6	36	1014
	22	149	8	64	1192
	26	202	12	144	2424
Total:	140	1300	0	568	2840
	Σx	Σy		*SXX*	*SXY*

$$\bar{x} = \frac{\sum x}{n} = \frac{140}{10} = 14 \qquad \bar{y} = \frac{\sum y}{n} = \frac{1300}{10} = 130$$

$$\hat{\beta}_1 = \frac{\sum (x-\bar{x})\, y}{\sum (x-\bar{x})^2} = \frac{2840}{568} = \mathbf{5}.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 130 - 5 \cdot 14 = \mathbf{60}.$$

The least-squares regression line: $\hat{y} = 60 + 5 \cdot x.$

OR

	x	y	x^2	xy	y^2
	2	58	4	116	3364
	6	105	36	630	11025
	8	88	64	704	7744
	8	118	64	944	13924
	12	117	144	1404	13689
	16	137	256	2192	18769
	20	157	400	3140	24649
	20	169	400	3380	28561
	22	149	484	3278	22201
	26	202	676	5252	40804
Total:	140	1300	2528	21040	184730
	Σx	Σy	Σx^2	Σxy	Σy^2

$$\bar{x} = \frac{\sum x}{n} = \frac{140}{10} = 14 \qquad \bar{y} = \frac{\sum y}{n} = \frac{1300}{10} = 130$$

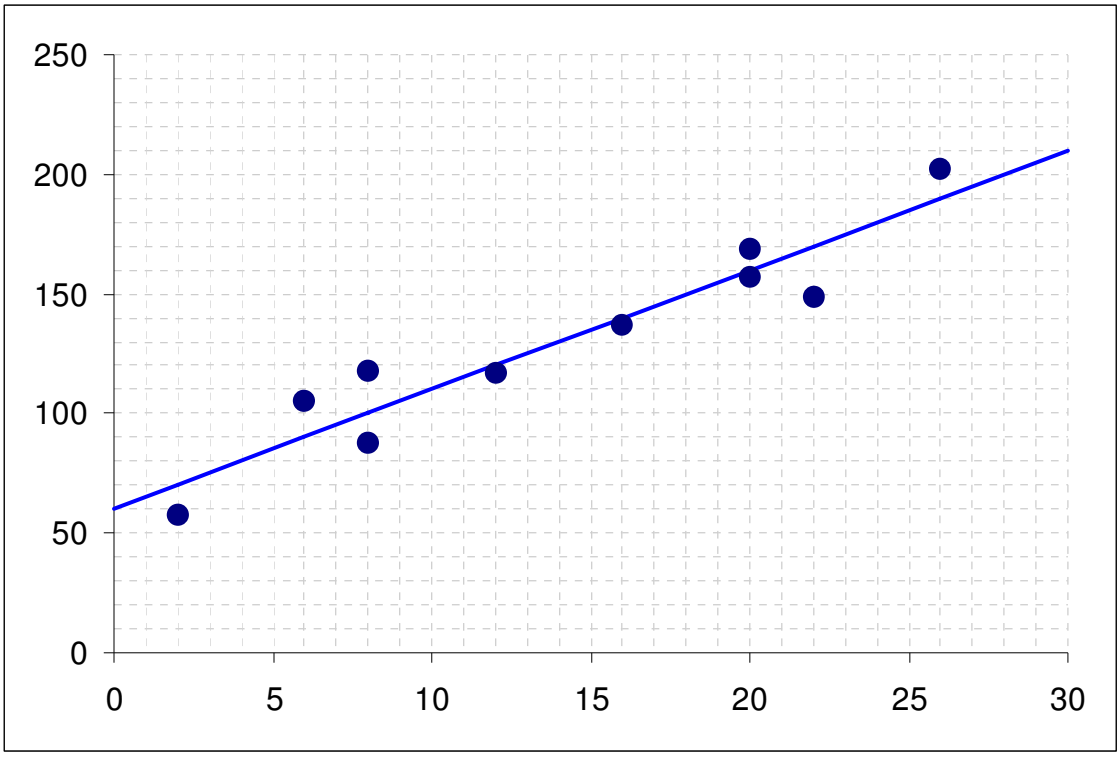
$$\hat{\beta}_1 = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{\sum x^2 - \frac{1}{n}(\sum x)^2} = \frac{21040 - \frac{1}{10}\cdot 140 \cdot 1300}{2528 - \frac{1}{10}\cdot 140^2} = \frac{2840}{568} = \mathbf{5}.$$

OR

$$\hat{\beta}_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{10\cdot 21040 - 140\cdot 1300}{10\cdot 2528 - 140^2} = \frac{28400}{5680} = \mathbf{5}.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 130 - 5 \cdot 14 = \mathbf{60}.$$

The least-squares regression line: $\hat{y} = 60 + 5 \cdot x.$



b) Suppose that Anytown College has 10 thousand students. Use the regression line to predict the level of quarterly sales if a *Momma Leona's Pizza* restaurant is opened near the Anytown College campus.

$$x = 10. \quad \Rightarrow \quad \hat{y} = 60 + 5 \cdot x = 60 + 5 \cdot 10 = \mathbf{110}.$$

c) Suppose that University of Illinois at Urbana-Champaign has 38 thousand students. Use the regression line to predict the level of quarterly sales if a *Momma Leona's Pizza* restaurant is opened near the UIUC campus.

$$x = 38. \quad \Rightarrow \quad \hat{y} = 60 + 5 \cdot x = 60 + 5 \cdot 38 = \mathbf{250}.$$

d) In which prediction (part (b) or part (c)) would we have more confidence?

We have more confidence in the **part (b)** prediction since $x = 10$ is within the range of the data that we have, and $x = 38$ is not. There is no guarantee that the relationship would remain the same for larger x.

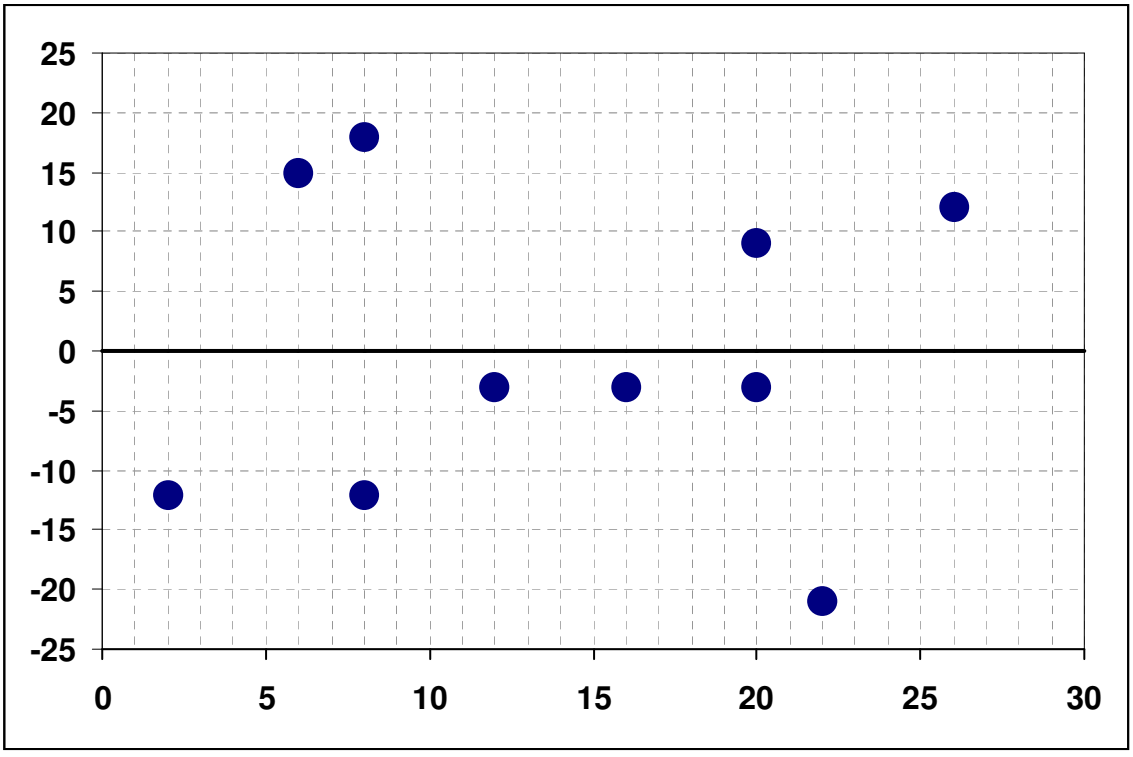
The least-squares regression line: $\hat{y} = 60 + 5 \cdot x.$

DATA = PREDICTION OF MODEL + RESIDUAL

$y = \hat{y} + e$

$e = y - \hat{y} = y - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x)$

i	x	y	$\hat{y}$	$e = y - \hat{y}$	e^2
1	2	58	70	– 12	144
2	6	105	90	15	225
3	8	88	100	– 12	144
4	8	118	100	18	324
5	12	117	120	– 3	9
6	16	137	140	– 3	9
7	20	157	160	– 3	9
8	20	169	160	9	81
9	22	149	170	– 21	441
10	26	202	190	12	144
			Total	0	1530



$RSS = \sum e_i^2 = \sum \left(y_i - \hat{y}_i \right)^2 = 1530$ (another common notation is *SSE*)

$$\underbrace{\sum \left(y_i - \bar{y} \right)^2}_{\text{Total variation}} = \underbrace{\sum \left(y_i - \hat{y}_i \right)^2}_{\text{Unexplained variation}} + \underbrace{\sum \left(\hat{y}_i - \bar{y} \right)^2}_{\text{Explained variation}}$$

coefficient of determination:

$$R^2 = \frac{\Sigma \left(\hat{y}_i - \bar{y} \right)^2}{\Sigma \left(y_i - \bar{y} \right)^2} = 1 - \frac{\Sigma \left(y_i - \hat{y}_i \right)^2}{\Sigma \left(y_i - \bar{y} \right)^2} = 1 - \frac{\Sigma e_i^2}{\Sigma \left(y_i - \bar{y} \right)^2}$$

$0 \le R^2 \le 1$

Coefficient of determination is interpreted as the proportion of observed *y* variation that can be explained by the simple linear regression model (attributed to an approximate linear relationship between *y* and *x*).

$R^2 = 1 - \frac{1530}{15730} \approx \mathbf{0.9027}.$ **90.27%**

To estimate σ^2:

Maximum Likelihood Estimator:

$\hat{\sigma}^2 = \frac{1}{n} \sum \left(y_i - \hat{y}_i \right)^2 .$

Simple linear regression sample variance:

$s_e^2 = \frac{1}{n-2} \sum \left(y_i - \hat{y}_i \right)^2 .$

$s_e^2 = \frac{1}{n-2} \sum \left(y_i - \hat{y}_i \right)^2 = \frac{1}{n-2} \sum e_i^2 = \frac{1}{8} \cdot 1530 = \mathbf{191.25}.$

$s_e = \sqrt{191.25} \approx \mathbf{13.83}.$

OR

$\hat{\sigma}^2 = \frac{1}{n} \sum \left(y_i - \hat{y}_i \right)^2 = \frac{1}{n} \sum e_i^2 = \frac{1}{10} \cdot 1530 = \mathbf{153}.$

$\hat{\sigma} = \sqrt{153} \approx \mathbf{12.37}.$