SIMPLE LINEAR REGRESSION

We consider the model for response variable, Y, as a function of the predictor, X, observed to take the value x. Specifically we consider the model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 and β_1 are the **intercept** and **slope** parameters respectively, and ϵ is a random variable with expectation zero and variance σ^2. In this model

$$E[Y|X = x] = \beta_0 + \beta_1 x.$$

To estimate the parameters β_0 and β_1 from data $(x_i, y_i), i = 1, \ldots, n$, we use the **least-squares** criterion, and choose the values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize the **sum of squared errors**

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

It can be shown that the parameter estimates depend on the following sample summary statistics:

- Sample mean of x values:
$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
- Sample mean of y values:
$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$
- Sum of Squares SS_{xx}:
$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$
- Sum of Squares SS_{xy}:
$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

The **least-squares estimates** are:

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

yielding **fitted-values**

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

and **residual errors** (or **residuals**)

$$\hat{e}_i = y_i - \hat{y}_i.$$

An estimate of the **residual error variance** is given by

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n - 2}$$

EXAMPLE: BLOOD VISCOSITY AND PACKED CELL VOLUME

The following data are measurements of packed cell volume (PCV) and blood viscosity in samples taken from 32 hospital patients. We wish to model viscosity (y) as a function of PCV (x).

Reference: Begg, C. B. and Hearns, J. B. (1966) Components of Blood Viscosity. The relative contributions of haematocrit, plasma fibrinogen and other proteins, *Clinical Science*, **31**, 87-92.

Unit ID	PCV (%) x	Viscosity y	$x_i - \overline{x}$	$y_i - \overline{y}$	Fitted $\hat{y}_i$	Error $\hat{e}_i$
1	40.00	3.71	-7.938	-0.936	3.674	0.036
2	40.00	3.78	-7.938	-0.866	3.674	0.106
3	42.50	3.85	-5.438	-0.796	3.980	-0.130
4	42.00	3.88	-5.938	-0.766	3.919	-0.039
5	45.00	3.98	-2.938	-0.666	4.286	-0.306
6	42.00	4.03	-5.938	-0.616	3.919	0.111
7	42.50	4.05	-5.438	-0.596	3.980	0.070
8	47.00	4.14	-0.938	-0.506	4.531	-0.391
9	46.75	4.14	-1.188	-0.506	4.500	-0.360
10	48.00	4.20	0.062	-0.446	4.653	-0.453
11	46.00	4.20	-1.938	-0.446	4.408	-0.208
12	47.00	4.27	-0.938	-0.376	4.531	-0.261
13	43.25	4.27	-4.688	-0.376	4.072	0.198
14	45.00	4.37	-2.938	-0.276	4.286	0.084
15	50.00	4.41	2.062	-0.236	4.898	-0.488
16	45.00	4.64	-2.938	-0.006	4.286	0.354
17	51.25	4.68	3.312	0.034	5.051	-0.371
18	50.25	4.73	2.312	0.084	4.929	-0.199
19	49.00	4.87	1.062	0.224	4.776	0.094
20	50.00	4.94	2.062	0.294	4.898	0.042
21	50.00	4.95	2.062	0.304	4.898	0.052
22	49.00	4.96	1.062	0.314	4.776	0.184
23	50.50	5.02	2.562	0.374	4.959	0.061
24	51.25	5.02	3.312	0.374	5.051	-0.031
25	49.50	5.12	1.562	0.474	4.837	0.283
26	56.00	5.15	8.062	0.504	5.633	-0.483
27	50.00	5.17	2.062	0.524	4.898	0.272
28	47.00	5.18	-0.938	0.534	4.531	0.649
29	53.25	5.38	5.312	0.734	5.296	0.084
30	57.00	5.77	9.062	1.124	5.755	0.015
31	54.00	5.90	6.062	1.254	5.388	0.512
32	54.00	5.90	6.062	1.254	5.388	0.512

- Sample mean of x values: $\overline{x} = 47.938$
- Sample mean of y values: $\overline{y} = 4.646$
- Sums of Squares

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = 615.75 \qquad SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = 75.386$$

Thus

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 0.122 \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = -1.223$$

The estimate of the residual error variance is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2} = \frac{2.721}{30} = 0.091$$