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**

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

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}:
$$\text{SS}_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$
- Sum of Squares SS_{xy}:
$$\text{SS}_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

The **least-squares estimates** are:

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

yielding **fitted-values**

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

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

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

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

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

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	PCV x	Viscosity y	Unit	PCV x	Viscosity y	Unit	PCV x	Viscosity y	Unit	PCV x	Viscosity y
1	40.00	3.71	9	46.75	4.14	17	51.25	4.68	25	49.50	5.12
2	40.00	3.78	10	48.00	4.20	18	50.25	4.73	26	56.00	5.15
3	42.50	3.85	11	46.00	4.20	19	49.00	4.87	27	50.00	5.17
4	42.00	3.88	12	47.00	4.27	20	50.00	4.94	28	47.00	5.18
5	45.00	3.98	13	43.25	4.27	21	50.00	4.95	29	53.25	5.38
6	42.00	4.03	14	45.00	4.37	22	49.00	4.96	30	57.00	5.77
7	42.50	4.05	15	50.00	4.41	23	50.50	5.02	31	54.00	5.90
8	47.00	4.14	16	45.00	4.64	24	51.25	5.02	32	54.00	5.90

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

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

Thus

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

The estimate of the residual error variance is

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