# THE METHOD OF LEAST SQUARES

# The goal:

to measure (determine) an unknown quantity $x$ (the value of a RV $X$) Realisation: $n$ results: $y_1, y_2, \ldots, y_j, \ldots, y_n$, (the measured values of $Y_1, Y_2, \ldots, Y_j, \ldots, Y_n$) every result is encumbered with an error $\varepsilon_j$:

$$y_j = x + \varepsilon_j; \quad j = 1, 2, \ldots, n$$

**the fundamental assumption: the errors are normally distributed with the expected value equal to zero and with some standard deviation $\sigma$**

$$\varepsilon_j \ \rightarrow \ N(0, \sigma); \quad E(\varepsilon_j) = 0; \ \ E(\varepsilon_j^2) = \sigma^2$$

if so, the probability of having a result in the range $(y_j, y_j + dy_j)$ equals to:

$$dP_j \equiv dP\left(Y_j \in (y_j, y_j + dy_j)\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_j - x)^2}{2\sigma^2}\right] dy_j$$

# Realisation:

$$dP_j \equiv dP\left(Y_j \in (y_j, y_j + dy_j)\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_j - x)^2}{2\sigma^2}\right] dy_j$$

The likelihood function $L$ is then:

$$L = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_j - x)^2}{2\sigma^2}\right]$$

and its logarithm $l$ is:

$$l = -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - x)^2 + \quad \text{a constant}$$

the demand $l = maximum \rightarrow \sum_{j=1}^{n} (y_j - x)^2 = minimum$

$$\boxed{\sum_{j=1}^{n} \varepsilon_j^2 = minimum}$$

The sum of the squares of the errors should be as small as possible if the determination of $\hat{x}$ is to be the most plausible.

# Realisation, cntd.

The ML estimator is the arithmetic mean:

$$\hat{x} = \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j, \quad \sigma^2(\hat{x}) = \frac{\sigma^2}{n}$$

or, if the errors connected with individual measurements are different:

$$\hat{x} = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} \quad \left[ w_j = \frac{1}{\sigma_j^2} \right], \quad \sigma^2(\hat{x}) = \left( \sum_{j=1}^{n} \frac{1}{\sigma_j^2} \right)^{-1}$$

Now, if $\hat{x}$ was the best estimator of $X$ (ML estimator) of RV $X$ then the quantities

$$\hat{\varepsilon_j} = y_j - \hat{x}$$

are the best estimators of the quantities $\varepsilon_j$ !!

# We may construct a statistic:

$$M = \sum_{j=1}^{n} \left( \frac{\hat{\varepsilon}_j}{\sigma_j} \right)^2 = \sum_{j=1}^{n} \frac{(y_j - \hat{x})^2}{\sigma_j^2} = \sum_{j=1}^{n} (y_j - \hat{x})^2 w_j$$

If the $\varepsilon_j$ have a normal distribution the RV $M$ should have a $\chi^2$ distribution with $n-1$ degrees of freedom. This hypothesis may be verified (tested). A positive result of the test supports the data treatment. A negative one calls for an extra analysis of the determination procedure.

# An example (from: *S. Brandt, Data analysis*

## measuring the mass of the $K^0$ meson: (all data in MeV/c$^2$)

$$
\begin{aligned}
m_1 &= 498.1; \quad \sigma_1 = 0.5, \\
m_2 &= 497.44; \quad \sigma_2 = 0.33, \\
m_3 &= 498.9; \quad \sigma_3 = 0.4, \\
m_4 &= 497.44; \quad \sigma_4 = 0.4,
\end{aligned}
$$

$$
\hat{x} = \frac{\sum_{j=1}^{4} y_j \times \frac{1}{\sigma_j^2}}{\sum_{j=1}^{4} \frac{1}{\sigma_j^2}} = 497.9 \qquad VAR(\hat{x}) = \left[ \sum_{j=1}^{4} \frac{1}{\sigma_j^2} \right]^{-1} = 0.20
$$

$$
M = \sum_{j=1}^{4} (y_j - \hat{x})^2 \frac{1}{\sigma_j^2} = 7.2 \quad \chi^2_{0.95;3} = 7.82
$$

there is no reason for discrediting the above scheme of establishing the value of $x$.

# LINEAR REGRESSION

**LINEAR REGRESSION** is a powerfull tool for studying fundamental relationships between two (or more) RVs $Y$ and $X$. The method is based on the method of least squares. Let's discuss the simplest case possible: we have a set of **bivariate data**, i.e. a set of $(x_i, y_i)$ values and we presume a **linear** relationship between the RV $Y$ (*dependent variable, response variable*) and the *explanatory (or regressor or predictor) variable* $X$. Thus we should be able to write:

$$\hat{y}_i = B_0 + B_1 \times x_i$$

**Note:** $Y_i$ are RVs, and $y_i$ are their *measured values*; $\hat{y}_i$ are the *fitted values*, i.e. the values resulting from the above relationship. We assume this relationship be true and we are interested in the numerical coefficients in the proposed dependence. We shall find them via an adequate treatement of the measurement data.

# LINEAR REGRESSION, cntd.

As for $x_i$ — these are the values of a random variable too, but of a rather different nature. For the sake of simplicity we should think about $X_i$ (or the values $x_i$) as of RV that take on values practically free of any errors (uncertainties). We shall return to this (unrealistic) assumption later on. [1] The errors $\varepsilon_i$ are to be considered as differences between the measured ($y_i$) and the "fitted" quantities:

$$\varepsilon_i = y_i - \hat{y}_i \equiv y_i - (B_0 + B_1 \times x_i)$$

As in the former case, we shall try to minimise the sum of the error squares (SSE): $Q = \sum \varepsilon_i^2$; it is not hard to show that this sum may be decomposed into 3 summands:

$$Q = S_{yy}(1 - r^2) + \left(B_1\sqrt{S_{xx}} - r\sqrt{S_{yy}}\right)^2 + n\left(\bar{y} - B_0 + B_1\bar{x}\right)^2 .$$

[1] One can imagine a situation when the values of the predictor variable $x_i$ had been "carefully prepared" prior to measurement, i.e. any errors connected with them are negligible. On the other hand, the $y_i$ values must be measured "on-line" and their errors should not be disregarded.

# LINEAR REGRESSION, cntd.

$$Q = S_{yy}(1-r^2) + \left(B_1\sqrt{S_{xx}} - r\sqrt{S_{yy}}\right)^2 + n\left(\bar{y} - B_0 + B_1\bar{x}\right)^2 .$$

The symbols used are:

$$\begin{aligned}
S_{xx} &= \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2 \\
S_{yy} &= \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n\bar{y}^2 \\
S_{xy} &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y}
\end{aligned}$$

The $\bar{x}$ and $\bar{y}$ are the usual arithmetic means; finally

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

is the sample estimator of the correlation coefficient.

In order to minimise $Q$ we are free to adjust properly the values of $B_0$ and $B_1$. It is obvious that $Q$ will be the smallest if the following equations are satisfied:

# LINEAR REGRESSION, cntd.

$$Q = S_{yy}(1-r^2) + \left(B_1\sqrt{S_{xx}} - r\sqrt{S_{yy}}\right)^2 + n\left(\bar{y} - B_0 + B_1\bar{x}\right)^2 .$$

$$\begin{aligned} B_1\sqrt{S_{xx}} - r\sqrt{S_{yy}} &= 0 \\ \bar{y} - B_0 + B_1\bar{x} &= 0. \end{aligned}$$

We shall denote the solutions for the values of $B_0$ (intercept) and $B_1$ (slope) coefficients which minimise the sum of squares in a special way:

$$\hat{\beta}_1 = \frac{r\sqrt{S_{xx}}}{\sqrt{S_{yy}}} = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

With the relation: $y = \hat{\beta}_0 + \hat{\beta}_1 x$ the SSE has minimum: $Q = S_{yy}(1-r^2)$. (N.B. this may be used to show that $|r|$ must be $\leq 1$.) For $r > 0$ the slope of the straight line is positive, and for $r < 0$ – negative.

# LINEAR REGRESSION, cntd.

The $r$ quantity (the sample correlation coefficient) gives us a measure of the adequacy of the assumed model (linear dependence). It can be easily shown that the *total sum of squares, SST* $= \sum_i (y_i - \bar{y})^2$ can be decomposed into a sum of the *regression sum of squares, SSR* and the introduced already *error sum of squares, SSE*:

$$\begin{aligned} \sum_i (y_i - \bar{y})^2 &= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \\ \text{or} \qquad SST &= SSR + SSE \end{aligned}$$

$SST$ is a quantity that constitutes a measure of total variability of the 'true' observations; $SSR$ is a measure of the variability of the fitted values, and $SSE$ is a measure of 'false' ('erroneous') variability. We have:

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

but: $SSR = SST - SSE = SST - SST(1 - r^2) = r^2 SST$ . Thus the above unity is a sum of two terms: the first of them is *the square of the sample correlation coefficient,* $r^2$ and it's sometimes called the *coefficient of determination*. The closer is $r^2$ to 1 the better is the (linear) model.

# LINEAR REGRESSION, cntd.

Up to now nothing has been said about the random nature of the fitted coefficients, $B_0, B_1$. We tacitly assume them to be some real numbers – coefficients in an equation. But in practice we calculate them from formulae that contain values of some RVs. Conclusion: $B_0, B_1$ should be also perceived as RVs, in that sense that their determination will be accomplished also with some margins of errors. The "linear relationship" should be written in the form:

$$Y_i = B_0 + B_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

or perhaps[2]

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

where $\varepsilon_i$ are errors, i.e. all possible factors other than the $X$ variable that can produce changes in $Y_i$. These errors are normally distributed with $E(\varepsilon_i) = 0$ and $VAR(\varepsilon_i)$ equal $\sigma^2$. From the above relation we have: $E(Y_i) = \beta_0 + \beta_1 x_i$ and $VAR(Y_i) = \sigma^2$ (remember: any errors on $x_i$ are to be neglected). The simple *linear regression model* has *three* unknown parameters: $\beta_0$, $\beta_1$ and $\sigma^2$.

[2] This change of notation reflects the change of our atitude to the fitted coefficients; we should think about them as about RVs.

# LINEAR REGRESSION, cntd.

The method of least squares allows us to find the numerical values of the beta coefficients – theses are the ML estimators and they should be perceived as the expected values:

$$\begin{aligned} E(\beta_1) &= \hat{\beta}_1 = \frac{r\sqrt{S_{xx}}}{\sqrt{S_{yy}}} = \frac{S_{xy}}{S_{xx}} \\ E(\beta_0) &= \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

As for the variances we have:

$$\begin{aligned} VAR(\beta_1) &= \frac{\sigma^2}{S_{xx}} \\ VAR(\beta_0) &= \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \end{aligned}$$

# verification

$$\begin{aligned} E(\beta_1) &= E\left\{\frac{S_{xy}}{S_{xx}}\right\} = E\left\{\sum_i \frac{(x_i-\bar{x})(Y_i-\bar{Y})}{S_{xx}}\right\} \stackrel{?}{=} E\left\{\sum_i \frac{(x_i-\bar{x})Y_i}{S_{xx}}\right\} \\ &= \sum_i \frac{(x_i-\bar{x})(\beta_0+\beta_1 x_i)}{S_{xx}} = \beta_0 \sum_i \frac{(x_i-\bar{x})}{S_{xx}} + \frac{\beta_1}{S_{xx}} \sum_i (x_i - \bar{x}) x_i \\ &\stackrel{?}{=} 0 + \frac{\beta_1}{S_{xx}} \sum_i (x_i-\bar{x})(x_i-\bar{x}) = \frac{\beta_1}{S_{xx}} S_{xx} = \beta_1 \end{aligned}$$

$$\begin{aligned} VAR(\beta_1) &= VAR\left\{\frac{S_{xY}}{S_{xx}}\right\} = VAR\left\{\sum_i \frac{(x_i-\bar{x})(Y_i-\bar{Y})}{S_{xx}}\right\} \\ &\stackrel{?}{=} \sum_i \frac{(x_i-\bar{x})^2}{S_{xx}^2} \times VAR(Y_i) = \frac{S_{xx}}{S_{xx}^2} \times \sigma^2 = \frac{\sigma^2}{S_{xx}} \end{aligned}$$

(some manipulations — $\stackrel{?}{=}$ — should be carefully justified; for $\beta_0$ the verification can be done in a similar manner)
The third parameter of the simple linear regression model is $\sigma^2$. It may be shown that the statistic

# verification, cntd.

$$s^2 = \frac{SSE}{n-2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$$

is an unbiased estimator of $\sigma^2$.
(The $n-2$ in the denominator reflects the fact that the data are used for determining *two* coefficients). The RV $(n-2)s^2/\sigma^2$ has a chi-square distribution with $n-2$ degrees of freedom. Replacing the values of $\sigma^2$ in the formulae for the variances of the beta coefficients by the sample estimator $s$ we conclude that these coefficients can be regarded as two standardised random variables:

$$\frac{\beta_1 - \hat{\beta}_1}{s\sqrt{S_{xx}}} \quad \text{and} \frac{\beta_0 - \hat{\beta}_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}.$$

# verification, cntd.

The ˆ-values are the ML estimators and the denominators are the estimated standard errors of our coefficients. Both standardised variables have a Student's distribution with the $n-2$ degrees of freedom. Their confidence intervals can de determined in the usual way.