The method of least squares

1. Connection with maximum likelihood
2. Linear LS problem
3. LS fit of a polynomial
4. Testing goodness-of-fit with LS
5. LS with binned data
6. Combining measurements with LS
Suppose we have Gaussian r.v.s \(y_i, i = 1, \ldots, N\)

\[E[y_i] = \lambda_i = \lambda(x_i; \tilde{\theta}),\]

where \(x_1, \ldots, x_N\) and \(V[y_i] = \sigma_i^2\) are known.

**Goal:** estimate parameters \(\tilde{\theta}\),

i.e. fit the curve through the points.

The joint pdf for independent Gaussian \(y_i\) is

\[g(\vec{y}; \vec{\lambda}, \vec{\sigma}^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - \lambda_i)^2}{2\sigma_i^2}\right)\]

i.e. the log-likelihood function is (drop terms not depending on \(\tilde{\theta}\)),

\[
\log L(\tilde{\theta}) = -\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \tilde{\theta}))^2}{\sigma_i^2}
\]

\(\rightarrow\) maximizing \(\log L(\tilde{\theta})\) same as minimizing

\[
\chi^2(\tilde{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \tilde{\theta}))^2}{\sigma_i^2}
\]
Definition of least squares (LS) estimators

If the \( y_i \) follow a multivariate Gaussian, covariance matrix \( V \),

\[
g(\bar{y}; \hat{\lambda}, V) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp \left[ -\frac{1}{2}(\bar{y} - \hat{\lambda})^T V^{-1}(\bar{y} - \hat{\lambda}) \right]
\]

then the log-likelihood is

\[
\log L(\hat{\theta}) = -\frac{1}{2} \sum_{i,j=1}^{N} (y_i - \lambda(x_i; \hat{\theta}))(V^{-1})_{ij}(y_j - \lambda(x_j; \hat{\theta})),
\]

i.e. we should minimize

\[
\chi^2(\hat{\theta}) = \sum_{i,j=1}^{N} (y_i - \lambda(x_i; \hat{\theta}))(V^{-1})_{ij}(y_j - \lambda(x_j; \hat{\theta}))
\]

Its minimum defines the least squares (LS) estimators \( \hat{\theta} \),
even when \( y_i \) not Gaussian. (In fact, \( y_i \) often Gaussian because
central limit theorem leads to Gaussian measurement errors.)

C.F. Gauss, Theoria Motus Corporum Coelestium in Sectionibus
Conicis Solem Ambientium, Hamburghi Sumtibus Friderici Perthes et
H. Besser Liber II, Sectio II (1809);
C.F. Gauss, Theoria Combinationis Observationum Erroribus
Minimis Obnoxiae, pars prior (15.2.1821) et pars posterior
(2.2.1823), Commentationes Societatis Regiae Scientiarum
Gottingensis Recensiones Vol. V (MDCCCXXIII).

G. Cowan – Computing and Statistical Data Analysis – University of London Postgraduate Lectures
Linear least squares fit

LS has particularly simple properties if $\lambda(x; \tilde{\theta})$ linear in $\tilde{\theta}$:

$$\lambda(x; \tilde{\theta}) = \sum_{j=1}^{m} a_j(x) \theta_j$$

where $a_j(x)$ are any linearly independent functions of $x$.

$\rightarrow \tilde{\theta}$ have zero bias, minimum variance (Gauss–Markov theorem)

Matrix notation: let $A_{i,j} = a_j(x_i)$,

$$\chi^2(\tilde{\theta}) = (\bar{y} - \bar{X})^T V^{-1} (\bar{y} - \bar{X})$$

$$= (\bar{y} - A\tilde{\theta})^T V^{-1} (\bar{y} - A\tilde{\theta})$$

Set derivatives with respect to $\theta_i$ to zero,

$$\nabla \chi^2 = -2(\bar{y}^T V^{-1} \bar{y} - \bar{y}^T V^{-1} A \tilde{\theta}) = 0$$

Solve to get the LS estimators,

$$\tilde{\theta} = (A^T V^{-1} A)^{-1} A^T V^{-1} \bar{y} \equiv B \bar{y}$$

N.B. estimators $\tilde{\theta}_i$ are linear functions of the measurements $y_i$. 

---

G. Cowan – Computing and Statistical Data Analysis – University of London Postgraduate Lectures
Variance of LS estimators

Error propagation (exact for linear problem) for \( U_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j] \):

\[
U = BVB^T = (A^T V^{-1} A)^{-1}
\]

Equivalently, use

\[
(U^{-1})_{ij} = \frac{1}{2} \left[ \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right]_{\tilde{\theta} = \tilde{\theta}}
\]

\( \rightarrow \) coincides with RCF bound if \( y_i \) are Gaussian.

For \( \lambda(x; \tilde{\theta}) \) linear in the parameters, \( \chi^2(\tilde{\theta}) \) is quadratic,

\[
\chi^2(\tilde{\theta}) = \chi^2(\tilde{\theta}) + \frac{1}{2} \sum_{i,j=1}^{m} \left[ \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right]_{\tilde{\theta} = \tilde{\theta}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)
\]

\( \rightarrow \) variances from tangent planes to (hyper)ellipse,

\[
\chi^2(\tilde{\theta}) = \chi^2(\tilde{\theta}) + 1 = \chi^2_{\text{min}} + 1
\]

If \( \lambda(x; \tilde{\theta}) \) not linear in \( \tilde{\theta} \), then expressions above not exact (but may still be good approximations).

Still interpret region \( \chi^2(\tilde{\theta}) \leq \chi^2_{\text{min}} + 1 \) as ‘confidence region’, having given probability of containing true \( \tilde{\theta} \) (more later).

N.B. formulae above don’t depend on \( y_i \) being Gaussian, but in any case need \( V_{ij} = \text{cov}[y_i, y_j] \).

G. Cowan – Computing and Statistical Data Analysis – University of London Postgraduate Lectures
LS fit of a polynomial

Fit a polynomial: \( \lambda(x; \theta_0, \ldots, \theta_m) = \sum_{j=0}^{m} \theta_j x^j \)

\( a_j(x) = x^j \)

Examples:
0th order (1 parameter)
1st order (2 parameters)
4th order (5 parameters)

1-parameter fit (i.e. horizontal line):

\( \hat{\theta}_0 = 2.66 \pm 0.13 \)

\( \chi^2_{\text{min}} = 45.5 \)

\( \sigma_{\hat{\theta}_0 \text{ from } \chi^2(\hat{\theta}_0 \pm \sigma_{\hat{\theta}_0}) = \chi^2_{\text{min}} + 1.} \)
2-parameter case (line with nonzero slope):

\[
\hat{\theta}_0 = 0.93 \pm 0.30, \\
\hat{\theta}_1 = 0.68 \pm 0.10 \\
cov[\hat{\theta}_0, \hat{\theta}_1] = -0.028 \\
r = -0.90 \\
\chi^2 = 3.99
\]

Tangent lines \(\rightarrow \sigma_{\hat{\theta}_0}, \sigma_{\hat{\theta}_1}\).

Angle of ellipse \(\rightarrow\) correlation (same as for ML)

Could transform \((\hat{\theta}_0, \hat{\theta}_1) \rightarrow (\hat{\eta}_0, \hat{\eta}_1)\) such that \(\text{cov}[\hat{\eta}_0, \hat{\eta}_1] = 0\),

easier to work with uncorrelated estimators, but interpretation

of new parameters may not be obvious, cf. SDA Section 1.7.

5-parameter case:

curve goes through all points,

\[\chi^2_{\text{min}} = 0,\]

(number of parameters = number of data points)

Value of \(\chi^2_{\text{min}}\) reflects agreement between data and hypothesis,

\(\rightarrow\) use as goodness-of-fit test statistic
Testing goodness-of-fit with LS

If: the $y_i$, $i = 1, \ldots, N$, are Gaussian ($V_{ij}$ known),
the hypothesis $\lambda(x; \overrightarrow{\theta})$ is linear in $\theta_i$, $i = 1, \ldots, m$, and
the form of the hypothesis $\lambda(x; \overrightarrow{\theta})$ is correct,
then $\chi^2_{\text{min}}$ follows chi-square pdf for $N - m$ degrees of freedom.

From this compute $P$-value,

$$P = \int_{\chi^2_{\text{min}}}^{\infty} f(z; n_d) dz$$

Consider e.g. 2-parameter fit:

$$\chi^2_{\text{min}} = 3.99, \ N - m = 3 \rightarrow P = 0.263$$

i.e. repeat experiment many times, 26.3% will have higher $\chi^2_{\text{min}}$:

For the horizontal line fit, we had

$$\chi^2_{\text{min}} = 45.5, \ N - m = 4 \rightarrow P = 3.1 \times 10^{-9}$$
Goodness-of-fit vs. smallness-of-errors

Small statistical error does not mean a good fit (nor vice versa).

Curvature of $\chi^2$ near its minimum $\rightarrow$ statistical errors ($\sigma_{\hat{\theta}}$)

Value of $\chi^2_{\min}$ $\rightarrow$ goodness-of-fit

Horizontal line fit, move the data points, keep errors on points same:

$$\hat{\theta}_0 = 2.84 \pm 0.13$$

$$\chi^2_{\min} = 4.48$$

Variance same as before,
now $\chi^2_{\min}$ ‘good’.

$\rightarrow \chi^2(\theta_0)$ shifted down, same curvature as before.

Variance of estimator (statistical error) tells us:

if experiment repeated many times, how wide is the distribution
of the estimates $\hat{\theta}$. (Doesn’t tell us whether hypothesis correct.)

$P$-value tells us:

if hypothesis is correct and experiment repeated many times,
what fraction will give equal or worse agreement between data
and hypothesis according to the statistic $\chi^2_{\min}$.

Low $P$-value $\rightarrow$ hypothesis may be wrong $\rightarrow$ systematic error.
Histogram:

$N$ bins, $n$ entries.

Hypothesized pdf:

$f(x; \theta)$

We have

$y_i =$ number of entries in bin $i$,

$$
\lambda_i(\bar{\theta}) = n \int_{x_i^{\text{min}}}^{x_i^{\text{max}}} f(x; \bar{\theta}) dx = n p_i(\bar{\theta})
$$

LS fit: minimize

$$
\chi^2(\bar{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\bar{\theta}))^2}{\sigma_i^2}
$$

where $\sigma_i^2 = V[y_i]$, here not known a priori.

Treat the $y_i$ as Poisson r.v.s, in place of true variance take either

$\sigma_i^2 = \lambda_i(\bar{\theta})$ \hspace{1cm} (LS method)

$\sigma_i^2 = y_i$ \hspace{1cm} (Modified LS method)

MLS sometimes easier computationally, but $\chi^2_{\text{min}}$ no longer follows chi-square pdf (or is undefined) if some bins have few (or no) entries.
Normalization with binned LS

Do not ‘fit the normalization’:

\[ \lambda_i(\bar{\theta}, \nu) = \nu \int_{x_i^{\text{min}}}^{x_i^{\text{max}}} f(x; \bar{\theta}) \, dx = \nu p_i(\bar{\theta}) \]

i.e. introduce adjustable \( \nu \), fit along with \( \bar{\theta} \).

\( \hat{\nu} \) is a bad estimator for \( n \) (which we know, anyway!)

\[ \hat{\nu}_{\text{LS}} = n + \frac{\chi^2_{\text{min}}}{2} \]

\[ \hat{\nu}_{\text{MLS}} = n - \chi^2_{\text{min}} \]

Example with \( n = 400 \) entries, \( N = 20 \) bins:

![Graphs showing data and fit for LS and MLS methods with chi-squared values and uncertainties](image)

Expect \( \chi^2_{\text{min}} \) around \( N - m \),

\[ \rightarrow \text{relative error in } \hat{\nu} \text{ large when } N \text{ large, } n \text{ small} \]

Either get \( n \) directly from data for LS (or better, use ML).
Combining measurements with LS

Use LS to obtain weighted average of $N$ measurements of $\lambda$:

$$y_i = \text{result of measurement } i, \ i = 1, \ldots, N;$$
$$\sigma_i^2 = V[y_i], \text{ assume known;}$$
$$\lambda = \text{true value (plays role of } \theta).$$

For uncorrelated $y_i$, minimize

$$\chi^2(\lambda) = \frac{\sum_{i=1}^{N} (y_i - \lambda)^2}{\sigma_i^2} ,$$

Set $\frac{\partial \chi^2}{\partial \lambda} = 0$ and solve,

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} y_i / \sigma_i^2}{\sum_{j=1}^{N} 1 / \sigma_j^2}$$

$$V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^{N} 1 / \sigma_i^2}$$

If $\text{cov}[y_i, y_j] = V_{ij}$, minimize

$$\chi^2(\lambda) = \sum_{i,j=1}^{N} (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda),$$

$$\hat{\lambda} = \sum_{i=1}^{N} w_i y_i, \quad w_i = \frac{\sum_{j=1}^{N} (V^{-1})_{ij}}{\sum_{k,l=1}^{N} (V^{-1})_{kl}}$$

$$V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j$$

LS $\hat{\lambda}$ has zero bias, minimum variance (Gauss–Markov theorem).
Example of averaging two correlated measurements

Suppose we have \( y_1, y_2 \), and 
\[
V = \begin{pmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{pmatrix}
\]

\[ \hat{\lambda} = w y_1 + (1 - w) y_2, \quad w = \frac{\sigma_2^2 - \rho \sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2 - 2 \rho \sigma_1 \sigma_2} \]

\[
V[\hat{\lambda}] = \frac{(1 - \rho^2) \sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2 \rho \sigma_1 \sigma_2} = \sigma^2
\]

The increase in inverse variance due to 2nd measurement is

\[
\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} = \frac{1}{1 - \rho^2} \left( \frac{\rho}{\sigma_1} - \frac{1}{\sigma_2} \right)^2 > 0
\]

\[ \rightarrow \text{2nd measurement can only help.} \]

If \( \rho > \sigma_1/\sigma_2 \), \( \rightarrow w < 0 \),

\[ \rightarrow \text{weighted average is not between } y_1 \text{ and } y_2 \text{ (!?)} \]

Cannot happen if correlation due to common data, but possible for shared random effect; very unreliable if e.g. \( \rho, \sigma_1, \sigma_2 \) incorrect.

See example in SDA Section 7.6.1 with two measurements at same temperature using two rulers, different thermal expansion coefficients: average is outside the two measurements; used to improve estimate of temperature.
The method of least squares

1. **Connection with maximum likelihood:** ML and LS same for Gaussian $y_i$.

2. **Linear LS problem:** if $\lambda(x; \bar{\theta})$ linear in the parameters, LS can be solved by matrix inversion; estimators are linear functions of the measurements $y_i$.

3. **LS fit of a polynomial:** an example of the linear problem. $\chi^2_{\text{min}}$ gets smaller when using more parameters, goes to zero for $N = m$.

4. **Testing goodness-of-fit with LS:** use $\chi^2_{\text{min}}$ as goodness-of-fit statistic, follows chi-square pdf for $N - m$ degrees of freedom.

5. **LS with binned data:** treat $y_i$ as Poisson,

   \[
   \text{LS: } \sigma^2_i = \lambda_i(\bar{\theta}),
   \]

   \[
   \text{MLS: } \sigma^2_i = y_i
   \]

   Do not fit the normalization (get $n$ from the data).

6. **Combining measurements with LS:** LS gives zero bias, minimum variance. Additional measurements can only help. For large correlations, weights can be negative.

---

G. Cowan – Computing and Statistical Data Analysis – University of London Postgraduate Lectures