Least squares regression:

\[
\min_\beta \sum_n (y_n - \beta x_n)^2
\]

"objective" or "error" function

In the space of measurements:

\[
\hat{\beta} = \arg \min_\beta \sum_n (y_n - \beta x_n)^2
\]
\[
\min_\beta \sum_n (y_n - \beta x_n)^2
\]

... or, with linear algebra!
\[
\min_\beta \|\tilde{y} - \beta \tilde{x}\|^2
\]

Geometry:

Note: this is a 2-D cartoon of the N-D vectors, not the two-dimensional \((x,y)\) measurement space of previous plots!

Note: partition of sum of squared data values:
\[
\|\tilde{y}\|^2 = \|\beta_{opt} \tilde{x}\|^2 + \|\tilde{y} - \beta_{opt} \tilde{x}\|^2
\] explained residual
Multiple regression:
\[
\min_{\beta} \| \mathbf{y} - \sum_{k} \beta_k \mathbf{x}_k \|^2 = \min_{\beta} \| \mathbf{y} - \mathbf{X} \mathbf{\beta} \|^2
\]

2D example:

(\text{observation}) \quad (\text{regressor 1}) \quad (\text{regressor 2})

Solution via the “Orthogonality Principle”:

Construct matrix $X$, containing columns $\mathbf{x}_1$ and $\mathbf{x}_2$

Orthogonality: $X^T(\mathbf{y} - \mathbf{X} \mathbf{\beta}) = \mathbf{0}$

Alternatively, can solve using SVD...

\[
\min_{\beta} \| \mathbf{y} - \mathbf{X} \mathbf{\beta} \|^2 = \min_{\beta} \| \mathbf{y} - \mathbf{U} \mathbf{S} \mathbf{V}^T \mathbf{\beta} \|^2 \\
= \min_{\beta} \| \mathbf{U}^T \mathbf{y} - \mathbf{S} \mathbf{V}^T \mathbf{\beta} \|^2 \\
= \min_{\mathbf{\beta}^*} \| \mathbf{y}^* - \mathbf{S} \mathbf{\beta}^* \|^2 \\
\text{where} \quad \mathbf{y}^* = \mathbf{U}^T \mathbf{y}, \quad \mathbf{\beta}^* = \mathbf{V}^T \mathbf{\beta}
\]

Solution: $\beta_{\text{opt},k} = y_k^* / s_k$, for each $k$

or $\mathbf{\beta}_{\text{opt}} = \mathbf{S} \mathbf{\beta}^* \Rightarrow \mathbf{\beta}_{\text{opt}} = \mathbf{V} \mathbf{S} \mathbf{U}^T \mathbf{y}$

[on board: transformations, elliptical geometry]
Fitting a parametric model (general)

Experimental Data $\bar{x}_n \rightarrow \bar{y}_n$  

Model $f_\beta(\bar{x})$

To fit model $f_\beta(\bar{x})$ to data $\{\bar{x}_n, \bar{y}_n\}$,

**optimize parameters $\beta$ to minimize an error function:**

$$\min_{\beta} \sum_n E(\bar{y}_n, f_\beta(\bar{x}_n))$$

Ingredients: data, model, error function, optimization method

---

**Optimization**

Heuristics, exhaustive search, (pain & suffering)

Iterative descent, (possibly) nonunique

Iterative descent, guaranteed

Closed-form guaranteed

---

**Interpretation warning:** fitting a line does not guarantee data actually lie along a line

These 4 data sets give the same regression fit, and same error:

[Anscombe, 1973]
Polynomial regression

Observation \( \bar{y} \)

constant \( \bar{x}_0 \)

linear \( \bar{x}_1 \)

quadratic \( \bar{x}_2 \)

\[-\beta_0 \quad -\beta_1 \quad -\beta_2\]

(to be continued, when we get to “statistics”...)

Polynomial regression - how many terms?

Weighted Least Squares

\[
\min_{\beta} \sum_{n} [w_n(y_n - \beta x_n)]^2
\]

\[
= \min_{\beta} \|W(\bar{y} - \beta \bar{x})\|^2
\]

Solution via simple extensions of basic regression solution

(i.e., let \( \bar{y}^* = W\bar{y} \) and \( \bar{x}^* = W\bar{x} \) then solve for \( \beta \))
Outliers

![Outliers graph](image1)

Outliers

![Outliers graph](image2)

Outliers

![Outliers graph](image3)
“Trimming”… discard points with large error (note: a special case of weighted least squares)

Trimming can be done iteratively (discard outlier, re-fit, repeat), a so-called “greedy” method. When should you stop?

More generally, use a “robust” error metric. For example:

\[ f(d) = d^2 \]
\[ f(d) = \log(e^2 + d^2) \]

“Lorentzian”

Note: generally can’t obtain solution directly (i.e., requires an iterative optimization procedure, such as gradient descent). In some cases, can use iteratively re-weighted least squares (IRLS)...
Constrained Least Squares

Linear constraint:
\[ \arg\min_{\beta} \left\| \tilde{y} - X\beta \right\|^2, \text{ where } \hat{c}^T\beta = 1 \]

Quadratic constraint:
\[ \arg\min_{\beta} \left\| X\beta \right\|^2, \text{ where } \|\beta\|^2 = 1 \]

Can be solved exactly using linear algebra (SVD)... [on board, with geometry]
Standard Least Squares regression

Error is \textit{vertical} distance (in the “dependent variable”) from the fitted line...

\[
\arg \min_{\beta} ||\hat{y} - \beta \hat{x}||^2
\]

Total Least Squares Regression (a.k.a “orthogonal regression”)

Error is squared distance from the fitted line...

expressed as: \[\min_\hat{u} ||D\hat{u}||^2, \text{ where } ||\hat{u}||^2 = 1\]

Note: “data” matrix \(D\) now includes both \(x\) and \(y\) coordinates

Variance of data \(D\), projected onto axis \(\hat{u}\):
\[
||USV^T \hat{u}||^2 = ||SV^T \hat{u}||^2 = ||S\hat{u}^*||^2 = ||\hat{u}^*||^2,
\]
where \(D = USV^T\), \(\hat{u}^* = V^T \hat{u}\), \(\hat{u}^* = S\hat{u}^*\)

Set of \(\hat{u}\)'s of length 1 (i.e., unit vectors)
Set of \(\hat{u}^*\)'s of length 1 (i.e., unit vectors)
First two components of \(\hat{u}^*\) (rest are zero!), for three example \(S\)'s.
Total Least Squares Regression
(a.k.a “orthogonal regression”)

Error is squared distance from the fitted line...

expressed as: \( \min_{\hat{u}} \|D\hat{u}\|^2 \), where \( \|\hat{u}\|^2 = 1 \)

Note: “data” matrix \( D \) now includes both \( x \) and \( y \) coordinates

Olympic gold medalists
(Rio, 2016)

3D geometry:
Javelin, Discus, Shotput…

Principal Component Analysis (PCA)

The shape of a data cloud can be summarized with an ellipse (ellipsoid), centered around the mean, using a simple procedure:
(1) Subtract mean of all data points, to re-center around origin
(2) Assemble centered data vectors in rows of a matrix, \( D \)
(3) Compute the SVD:

\[
D = U S V^T
\]

or just use the smaller matrix

\[
C = D^T D = V S^T S V^T = V \Lambda V^T
\]

(4) Columns of \( V \) are the principal components (axes) of the ellipsoid, diagonal elements \( s_k \) or \( \sqrt{\lambda_k} \) are the corresponding principle radii, and their product is the volume.
Population activity time courses — how it unfolds over time, how different types of information can be encoded in network models, and dimensionality reduction, this review focuses on studies that incorporate network models to study working memory, decision making, motor control, and more.

Fig 1. Relating biological and model networks using population analyses: Because a model network typically does not attempt to replicate the precise anatomical connectivity of a biological network, there is not a one-to-one correspondence of each biological neuron with a model neuron. Dimensionality reduction can be used to obtain a concise summary of the population activity from each network. This provides common ground for incisive comparisons between biological and model networks. Discrepancies in the population activity structure between biological and model networks can then help to refine model networks.

[Williamson, Doiron, Smith, Yu 2019]

Example: PCA for dimensionality reduction and visualization

Eigenvectors/eigenvalues

- An eigenvector of a matrix is a vector that is rescaled by the matrix (i.e., the direction is unchanged)
- The corresponding scale factor is called the eigenvalue

- For matrix $C = D^T D = VAV^T$ the columns of $V$ (denoted $\hat{v}_k$) are eigenvectors, with corresponding eigenvalues $\lambda_k$:

$$C\hat{v}_k = VAV^T\hat{v}_k$$
$$= V\lambda_k \hat{v}_k$$
$$= \lambda_k \hat{v}_k$$