Section 2: Least Squares

Least squares regression:

$$\min_\beta \sum_n (y_n - \beta x_n)^2$$

“Objective” or “error” function

In the space of measurements:

$$\hat{\beta} = \arg \min_\beta \sum_n (y_n - \beta x_n)^2$$
can solve this with calculus… *(on board)*

... or, with linear algebra!

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

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:

\[
\| \mathbf{y} \|^2 = \| \mathbf{\beta_{opt} x} \|^2 + \| \mathbf{\hat{y} - \beta_{opt} x} \|^2
\]

Multiple regression:

\[
\min_{\beta} \| \mathbf{y} - \sum_k \beta_k \mathbf{x}_k \|^2 = \min_{\beta} \| \mathbf{y} - \mathbf{X}\mathbf{\beta} \|^2
\]
Solution via the “Orthogonality Principle”:

Construct matrix \( X \), containing columns \( \tilde{x}_1 \) and \( \tilde{x}_2 \)

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

Alternatively, can solve using SVD...

\[
\begin{align*}
\min_\beta ||\tilde{y} - X \tilde{\beta}||^2 &= \min_\beta ||\tilde{y} - USV^T \tilde{\beta}||^2 \\
&= \min_\beta ||USV^T \tilde{\beta}||^2 \\
&= \min_\beta ||\tilde{y}^* - S \tilde{\beta}^*||^2 \\
\text{where} \quad \tilde{y}^* &= U^T \tilde{y}, \quad \tilde{\beta}^* = V^T \tilde{\beta}
\end{align*}
\]

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

or \( \tilde{\beta}_{\text{opt}} = S^# \tilde{y}^* \Rightarrow \beta_{\text{opt}} = V S^# U^T \tilde{y} \)

[on board: transformations, elliptical geometry]

Fitting a parametric model (general)

Experimental Data \( \tilde{x}_n \rightarrow \tilde{y}_n \)

Model \( f_\beta (\tilde{x}) \)

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

optimize parameters \( \beta \) to minimize an error function:

\[
\min_\beta \sum_n E (\tilde{y}_n, f_\beta (\tilde{x}_n))
\]

Ingredients: data, model, error function, optimization algorithm
Optimization problems

- Heuristics, exhaustive search, (pain & suffering)
- Iterative descent, (possibly) nonunique
- Iterative descent, guaranteed
- Closed-form guaranteed

Be careful with interpretation: 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

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \]
Polynomial regression - how many terms?

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

Weighted Least Squares

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

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

Solution via simple extensions of basic regression solution
(i.e., let \( \tilde{y}^* = W\tilde{y} \) and \( \tilde{x}^* = W\tilde{x} \) and solve for \( \beta \))

Outliers
Outliers

Solution 1: “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 do you stop?
Solution 2: Use a “robust” error metric. For example:

\[ f(d) = d^2 \]

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

“Lorentzian”

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

Iteratively Re-weighted Least Squares (IRLS)

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

\[ w_n^{(i+1)} = \frac{f(y_n - \beta^{(i)} x_n)}{|y_n - \beta^{(i)} x_n|} \]

Constrained Least Squares

Linear constraint:

\[ \min_{\vec{\beta}} \| \vec{y} - X \vec{\beta} \|^2, \text{ where } \vec{c} \cdot \vec{\beta} = \alpha \]

Quadratic constraint:

\[ \min_{\vec{\beta}} \| \vec{y} - X \vec{\beta} \|^2, \text{ where } \| \vec{\beta} \|^2 = 1 \]

Both can be solved exactly using linear algebra (SVD)...
Standard Least Squares regression

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

arg min \( \beta \) \( ||\vec{y} - \beta \vec{x}||^2 \)
Total Least Squares Regression
(a.k.a “orthogonal regression”)

Error is squared distance from the fitted line...

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

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

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

Set of \( \hat{\mathbf{u}} \)'s of length 1 (i.e., unit vectors)
Set of \( \hat{\mathbf{u}}^* \)'s of length 1 (i.e., unit vectors)
First two components of \( \hat{\mathbf{u}}^{**} \) (rest are zero!), for three example \( S \)'s.

Eigenvalues/eigenvectors

Define symmetric matrix:
\[
C = D^T D = (USV^T)^T(USV^T) = V S^T U^T U S V^T = V (S^T S) V^T
\]

• An eigenvector is a vector that is rescaled by a matrix (i.e., direction is unchanged)
• The corresponding scale factor is called the eigenvalue
• The columns of \( V \) (denoted \( \mathbf{v}_k \)) are eigenvectors of \( C \) with corresponding eigenvalues \( s_k^2 \):
\[
C \mathbf{v}_k = V (S^T S) V^T \mathbf{v}_k = V (S^T S) \hat{\mathbf{v}}_k = s_k^2 V \hat{\mathbf{v}}_k = s_k^2 \mathbf{v}_k
\]

“rotate, stretch, rotate back”
• The matrix \( C \) “summarizes” the shape of the data with an ellipsoid: principal axes are columns of \( V \), dimensions are diagonal elements of \( S \)
Principal Component Analysis (PCA)

The shape of a data cloud can be summarized with an ellipse (ellipsoid) 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 of $D$:
   \[ D = U S V^T \]
   or equivalently compute eigenvectors of $C = D^T D$:
   \[ C = V \Lambda V^T \]
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

Example: PCA for dimensionality reduction and visualization

[Image: Russo et. al., 2018]