Normal Equations

This approach exploits knowledge of the orthogonality of the residual.

Recall $A \in \mathbb{R}^{n \times k}$ is full rank, $b \in \mathbb{R}^n$, $x \in \mathbb{R}^k$. The vector $x$ minimizes over all vectors in $\mathbb{R}^k$ the two norm of the residual

$$\|b - Ax\|_2$$

Recall

$$r \perp \mathcal{R}(A)$$
We have

\[ r = b - Ax \]
\[ r \perp \mathcal{R}(A) \]
\[ 0 = A^T r \]
\[ = A^T b - A^T Ax \]
\[ A^T Ax = A^T b \]
\[ x = (A^T A)^{-1} A^T b \]
\[ = A^\dagger b \]

\( A^\dagger \) is called the **Moore-Penrose inverse or pseudoinverse**. Note \( A^\dagger = A^{-1} \) if \( n = k \).
• $A^T A$ is then symmetric positive definite since $A$ is full rank.

• The inverse must exist and therefore we have a unique $x$.

• Even if $A$ is not full rank the normal equations must have a solution (the residual must be orthogonal to any $x$ that gives a minimal residual norm).

• Uniqueness is lost if $A$ is not full rank and the algorithm for solving the singular system must be carefully designed. (Recall we mentioned briefly in class the notion of a low-rank factorization that produces trapezoidal factors.)
Complexity

- Forming $A^TA$:
  
  - The matrix is symmetric so only half of its elements are needed.
  
  - $\approx \frac{k^2}{2}$ inner products are needed each of length $n$.
  
  - $nk^2 + O(k^2)$ operations
  
  - dominates the algorithm

- Forming $A^Tb$: $O(nk)$
• Cholesky factorization $A^T A = R^T R$: $\frac{k^3}{3} + O(k^2)$ (a lower order term – but the most significant – always check $k^3$ relative to $n$ for rectangular problems).

• Triangular solves: $O(k^2)$ (a lower order term)

• Total operations: $nk^2 + O(k^3)$
• This approach is the fastest that we will consider in terms of operation count.

• If $A$ is square it would not be used since $A$ would be factored via $LU$ directly rather than via $A^T A$ and the Cholesky factorization.

• The backward error involves a perturbation to $A^T A$ and to guarantee that $A^T A$ is positive definite places contraints on the conditioning of $A$ (measured via $\kappa(A) = \|A\|\|A^\dagger\|$)
• The backward error may be projected back to $A$ but the size of the perturbation is related to $\kappa(A)$

• So this approach is worthwhile only for very well-conditioned problems and has earned (a somewhat undeserved) reputation of being a method to avoid in general. It is in fact used successfully in many applications.
Linear Regression

Assume you are given a set of points \((\xi_i, \eta_i), \ i = 1, \ldots, n\)

Linear regression determines a line \(\eta = f(\xi) = \alpha \xi + \beta\) that minimizes the sum of the squares of the errors

\[
\sum_{i=1}^{n} (\eta_i - f(\xi_i))^2
\]

Standard statistics texts express the solution in terms of the means \(\bar{\xi}, \bar{\eta}\), the variance, \(\sigma_x^2\) and the covariance \(\sigma_{xy}\)
Standard Form

\[ \bar{\eta} = \frac{1}{n} \sum_{i} \eta_i \]

\[ \bar{\xi} = \frac{1}{n} \sum_{i} \xi_i \]

\[ s_{xy} = \sum_{i} \xi_i \eta_i \]

\[ s_{xx} = \sum_{i} \xi_i \xi_i \]

\[ s_x = \sum_{i} \xi_i \]

\[ s_y = \sum_{i} \eta_i \]
\[
\sigma_{xy} = \frac{1}{n-1} \sum_i (\xi_i - \bar{\xi})(\eta_i - \bar{\eta})
\]
\[
= \frac{1}{n-1}(s_{xy} - \frac{1}{n}s_{sx} s_{sy})
\]
\[
\sigma^2_x = \frac{1}{n-1} \sum_i (\xi_i - \bar{\xi})^2
\]
\[
= \frac{1}{n-1}(s_{xx} - \frac{1}{n}s_{sx} s_{sx})
\]
\[
\gamma = \frac{\sigma_{xy}}{\sigma^2_x}
\]
\[
= \frac{s_{xy} - \frac{1}{n}s_{sx} s_{sy}}{s_{xx} - \frac{1}{n}s_{sx} s_{sx}}
\]
Given all of the definitions above the statistics texts form is:

\[ \eta - \bar{\eta} = \gamma(\xi - \bar{\xi}) \]
\[ f(\xi) = \gamma(\xi - \bar{\xi}) + \bar{\eta} \]

**NOTE:** The means satisfy the regression equation, i.e.,

\[ \bar{\eta} = f(\bar{\xi}) \]
What are the least squares and normal equation forms?

\[
\begin{pmatrix}
\xi_1 & 1 \\
\xi_2 & 1 \\
\vdots & \vdots \\
\xi_n & 1
\end{pmatrix}
\begin{pmatrix}
\alpha \\
\beta
\end{pmatrix}
= 
\begin{pmatrix}
\eta_1 \\
\eta_2 \\
\vdots \\
\eta_n
\end{pmatrix}
\]

is the overdetermined set of observation equations, $Mv = y$. 
Normal Equations

\[
(M^T M) \nu = M^T y
\]

\[
\begin{pmatrix}
    s_{xx} & s_x \\
    s_x & n
\end{pmatrix}
\begin{pmatrix}
    \alpha \\
    \beta
\end{pmatrix}
= 
\begin{pmatrix}
    s_{xy} \\
    s_y
\end{pmatrix}
\]

\[
\begin{pmatrix}
    n^{-1} & 0 \\
    0 & n^{-1}
\end{pmatrix}
\begin{pmatrix}
    s_{xx} & s_x \\
    s_x & n
\end{pmatrix}
\begin{pmatrix}
    \alpha \\
    \beta
\end{pmatrix}
= 
\begin{pmatrix}
    n^{-1} & 0 \\
    0 & n^{-1}
\end{pmatrix}
\]

\[
\begin{pmatrix}
    s_{xy} \\
    s_y
\end{pmatrix}
\begin{pmatrix}
    \bar\xi \\
    1
\end{pmatrix}
\begin{pmatrix}
    \alpha \\
    \beta
\end{pmatrix}
= 
\begin{pmatrix}
    \bar\eta
\end{pmatrix}
\]
The consistency between the two forms is easily seen.

Starting from the statistics form we have:

\[ \eta - \bar{\eta} = \gamma (\xi - \bar{\xi}) \]
\[ = \gamma (\xi - \bar{\xi}) + \bar{\eta} \]
\[ = \gamma \xi + (\bar{\eta} - \gamma \bar{\xi}) \]
\[ \alpha = \gamma \]
\[ \beta = \bar{\eta} - \gamma \bar{\xi} \]
\[ = \bar{\eta} - \alpha \bar{\xi} \]
Now start from the normal equations:

\[
\begin{pmatrix}
\frac{s_{xx}}{n} & \bar{\xi}
\
\xi & 1
\end{pmatrix}
\begin{pmatrix}
\alpha \\
\beta
\end{pmatrix}
= 
\begin{pmatrix}
\frac{s_{xy}}{n} \\
\bar{\eta}
\end{pmatrix}
\]

The second equation says:

\[
\alpha \bar{\xi} + \beta = \bar{\eta}
\]

\[
\bar{\eta} - \alpha \bar{\xi} = \beta
\]

as desired.
The first equation and substitution of $\beta$ yields:

$$\frac{s_{xx}}{n} \alpha + \bar{\xi} \beta = \frac{s_{xy}}{n}$$

$$\alpha \left( \frac{s_{xx}}{n} - \bar{\xi}^2 \right) + \bar{\xi} \bar{\eta} = \frac{s_{xy}}{n}$$

$$\alpha = \frac{s_{xy}}{n} \bar{\xi} \bar{\eta}$$

$$= \frac{s_{xx}}{n} - \bar{\xi}^2$$

$$= \gamma$$