I. Simultaneous Equations

A. Equations in Matrix Form

1. Consider the following equations:

\[ \begin{align*}
2x_1 + x_2 + x_3 &= 0 \\
x_1 + x_2 + x_3 &= 1 \\
x_1 + 2x_2 + x_3 &= 3
\end{align*} \]

2. This can be written as:

\[
\begin{pmatrix}
2 & 1 & 1 \\
1 & 1 & 1 \\
1 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\ x_2 \\ x_3
\end{pmatrix}
=
\begin{pmatrix}
0 \\ 1 \\ 3
\end{pmatrix}
\]

or \( Ax = b \).

3. How to solve for \( x \) depends on the number of unknowns relative to the number of linearly independent equations:

a. Fully determined (same number of equations as unknowns)
   - Solve by elimination and back-substitution (Gauss)
   - Or using inverse, noting that \( x = A^{-1}b \)
   
   → Brief excursion on relationship between inverse, determinants, Cramer’s Rule (to which we will refer later in finding eigenvalues):
   
   i. Define the adjugate matrix \( \text{adj}A \) as the matrix of all cofactors of \( A \) (see Linear Algebra I), arranged in “transpose form”:

   \[
   \text{adj}A =
   \begin{pmatrix}
   A_{11} & A_{21} & \cdots & A_{p1} \\
   A_{12} & A_{22} & \cdots & A_{p2} \\
   \vdots & \vdots & \ddots & \vdots \\
   A_{1p} & A_{2p} & \cdots & A_{pp}
   \end{pmatrix}
   \]

   ii. It can be shown that \( A^{-1} = \frac{\text{adj}A}{\det A} \).

   iii. From (ii) it follows that \( x = A^{-1}b = \frac{\text{adj}A}{\det A} \).

   iv. From (iii) it follows that the \( j \)th element of \( x \) can be obtained by:

   \[
   x_j = \frac{\det B_j}{\det A},
   \]

   where \( B_j \) is the matrix \( A \) with \( b \) substituted for the \( j \)th column. This is Cramer’s rule.

b. Underdetermined (more unknowns than equations)

   - Generally yields a family of solutions. E.g. \( x + y = 1 \) has solutions on the line \( y = 1 - x \).

c. Overdetermined (more equations than unknowns). For example:

\[
\begin{align*}
1 \cdot x_1 + 0 \cdot x_2 &= b_1 \\
5 \cdot x_1 + 4 \cdot x_2 &= b_2 \text{ or } \begin{pmatrix} 1 & 0 \\ 5 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \text{ or } x_1 \cdot \begin{pmatrix} 1 \\ 5 \end{pmatrix} + x_2 \cdot \begin{pmatrix} 0 \\ 4 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\
2 \cdot x_1 + 4 \cdot x_2 &= b_3
\end{align*}
\]
If it is possible to express $\vec{b}$ as a linear combination of the columns of $A$ (weighted by $x_1$ and $x_2$), then there is an exact solution. In other words, there is an exact solution if $\vec{b}$ lies in the subspace spanned by the columns of $A$. This is the consistent case.

Otherwise the equations are inconsistent and there is no exact solution. This is where we resort to least squares estimation.

Previous approach to Least Squares: Define squared residual, use calculus to minimize.

B. Least Squares Revisited

1. Consider 3 equations in one unknown:

$$
\begin{align*}
2x &= b_1 \\
3x &= b_2 \\
4x &= b_3
\end{align*}
$$

or $A\vec{x} = \vec{b}$, where $A = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$, $\vec{x} = (x)$, $\vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$. Since $A$ has only one column, let’s write it as $\vec{a}$.

2. Expression for sum of squared residuals: $E^2 = (a_1 x - b_1)^2 + \ldots + (a_k x - b_k)^2$. In this case, $k = 3$.

3. Vector equivalent of this expression: $E^2 = \|\vec{a}\vec{x} - \vec{b}\|^2$

4. An equivalent form: $E^2 = (\vec{a}\vec{x} - \vec{b})^T (\vec{a}\vec{x} - \vec{b}) = \vec{a}^T \vec{a}\vec{x}^2 - 2\vec{a}^T \vec{b}\vec{x} + \vec{b}^T \vec{b}$

5. Take first derivative of this with respect to $\vec{x}$ and set to zero:

$$
\frac{dE^2}{d\vec{x}} = 2\vec{a}^T \vec{a}\vec{x} - 2\vec{a}^T \vec{b} = 0
$$

The solution is $\vec{x} = \vec{a}^T \vec{b} / \|\vec{a}\|^2$.

6. Geometrically, it can be seen that $\vec{a}\vec{x}$ goes in the direction of $\vec{a}$, and extends to the point where the distance to $\vec{b}$ is minimized. Thus, the error vector $\vec{a}\vec{x} - \vec{b}$ is perpendicular to $\vec{a}$. Put another way, if we project $\vec{b}$ onto $\vec{a}$, the resulting vector is equal to $\vec{a}\vec{x}$.

7. The same reasoning holds for larger problems. For example, consider:

$$
\begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22} \\
a_{31} & a_{32}
\end{pmatrix}
\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}
$$

8. The column space of $A$ is spanned by $\vec{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}$ and $\vec{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix}$.

9. The vectors $\vec{a}_1$ and $\vec{a}_2$ define a plane, and in general $\vec{b}$ will not be in this plane.

10. If we project $\vec{b}$ onto this plane, the projection $\vec{p}$ will be equal to $A\vec{x}$.

11. But what is $\vec{x}$? By a general extension of the error minimization procedure given above, $\vec{x}$—the least squares solution—satisfies

$$
A^T A\vec{x} = A^T \vec{b}
$$

This compact formula represents the set of normal equations, which, in less compact form, look like this (for the two-dimensional case):

$$
\sum_{i=1}^{n} a_{i1}^2 |\vec{x}_1| + \sum_{i=1}^{n} a_{i1} a_{i2} |\vec{x}_2| = \sum_{i=1}^{n} a_{i1} b_i
$$
\[
\sum_{i=1}^{n} a_{i1}x_1 + \sum_{i=1}^{n} a_{i2}x_2 = \sum_{i=1}^{b} a_{i2}b_i,
\]

where \( n \) is the number of rows of \( A \), i.e. the number of equations in \( Ax = \bar{b} \), which in this case is 3. Note that the number of normal equations is equal to the number of unknowns, which in this case is 2.

12. Solving for \( \bar{x} \), we have:

\[
\bar{x} = (A^T A)^{-1} A^T \bar{b},
\]

and the projection of \( \bar{b} \) onto the column space is

\[
\bar{p} = A\bar{x} = A(A^T A)^{-1} A^T \bar{b}
\]

II. Multiple Regression

A. Uses (not necessarily good ones)

1. Description of structural relationship between one “dependent” variable and several “independent” variables
2. Prediction
3. Estimation of missing values

B. Model

\[
Y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p + \epsilon
\]

- \( Y \): “response” (observed)
- \( b_0 \): intercept (unknown)
- \( x_1, \ldots, x_p \): set of “predictor” variables (observed)
- \( \epsilon \): residual (error)
- \( b_i \): partial regression coefficient (coeff. of regression of \( Y \) on \( x_i \) with all other \( x \) held fixed at mean) (unknowns)

C. Data

1. Assume more equations than unknowns (\( n \) equations, \( n > p \))
2. General equation holds for each observation

\[
b_0 + b_1x_{11} + b_2x_{12} + \ldots + b_px_{1p} = Y_1
\]

\[
b_0 + b_1x_{21} + b_2x_{22} + \ldots + b_px_{2p} = Y_2
\]

\[\ldots\]

\[
b_0 + b_1x_{n1} + b_2x_{n2} + \ldots + b_px_{np} = Y_n
\]

3. These equations in matrix form:

\[
\begin{pmatrix}
1 & x_{11} & x_{12} & \ldots & x_{1p} \\
1 & x_{21} & x_{22} & \ldots & x_{2p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \ldots & x_{np}
\end{pmatrix}
\begin{pmatrix}
b_0 \\
b_1 \\
\vdots \\
b_p
\end{pmatrix}
= 
\begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{pmatrix}
\]

or

\[
X\bar{b} = \bar{Y},
\]
where \( \vec{b} \) are the unknowns for which we must solve.

D. Least Squares Solution

Use normal equations as already established for overdetermined case of simultaneous equations:

\[
\hat{\vec{b}} = (X^T X)^{-1} X^T \vec{Y}.
\]

E. Goodness of Fit etc.

1. ANOVA for overall regression
   a. Total sum of squares (i.e. squared deviations from mean) in \( Y \): \( SS_T \).
   b. Regression sum of squares equal to sum of squares in \( \hat{\vec{Y}} \), the predicted values: \( SS_R \).
   c. Residual sum of squares: \( SS_D = SS_T - SS_R \).
   d. Regression degrees of freedom equals \( p \); thus regression mean square: \( MS_R = SS_R/p \).
   f. Residual degrees of freedom equals \( (n - p - 1) \); thus residual mean square: \( MS_D = SS_D/(n - p - 1) \).
   g. F-test: \( F = MS_R/MS_D \) with \( p \) d.f. in the numerator and \( (n - p - 1) \) d.f. in the denominator.
   h. Coefficient of multiple determination: \( R^2 = SS_R/SS_T \).

2. Test for improvement in fit with added terms
   a. Let \( R^2_{k_1} \) and \( R^2_{k_2} \) be the coefficients of determination for multiple regressions with \( k_1 \) and \( k_2 \) predictor variables, where \( k_2 > k_1 \) and the \( k_1 \) predictors are a subset of the \( k_2 \) predictors (i.e. the models are nested).
   b. Then evaluate \( F = (R^2_{k_2} - R^2_{k_1})/(k_2 - k_1) \) with \( (k_2 - k_1) \) d.f. in the numerator and \( (n - k_2 - 1) \) d.f. in the denominator.

F. Sepkoski example (1976, Paleobiology 4:298-303)

- Does habitable area explain extra variance in fossil species richness once the effect of rock amount is accounted for?

G. Other variations on multiple regression

1. Polynomial regression: \( Y = b_0 + b_1 X + b_2 X^2 + \cdots + b_p X^p \)


H. Problems with Multiple Regression

1. Overestimation of \( R^2 \)

   - Matrix of normal equations in effect treats sample (co)variances as if they were parametric values.
   - ANOVA test of significance implicitly corrects by taking degrees of freedom into account.

2. Model specification

   - How many predictor variables to include? which ones?
   - Stepwise multiple regression may help, but no panacea
   - Is predictor-response model reasonable?
   - Is assumption of additivity (no interaction) reasonable?

3. Multiple collinearity among predictors

   - Estimates of \( \vec{b} \) very sensitive to sampling error in \( Y \) if \( x \)’s highly intercorrelated.
   - Multicollinearity also leads to inflated statistical significance.
   - “Solutions”
• Regression of \( Y \) on principal components (which won’t make sense until we cover principal component analysis...)
• Ridge regression: Add a small constant to diagonal terms of \( X^TX \), which makes it more stable when inverted.