Matrix Algebra Overview

Paul E. Johnson\textsuperscript{1, 2}

\textsuperscript{1}Department of Political Science

\textsuperscript{2}Center for Research Methods and Data Analysis, University of Kansas

2015
Outline

1. Objectives
2. Vector
3. Matrix
4. Decompositions
5. Regression
Outline

1 Objectives

2 Vector

3 Matrix

4 Decompositions

5 Regression
Matrix View of Multiple Regression

Assume:

\[ Y = X\beta + e \]

With 2 predictors, that’s short for:

\[
\begin{bmatrix}
    y_1 \\
    y_2 \\
    \vdots \\
    y_N
\end{bmatrix}
= 
\begin{bmatrix}
    1 & X_{11} & X_{12} \\
    1 & X_{12} & X_{22} \\
    1 & \ldots & \ldots \\
    1 & X_{1N} & X_{2N}
\end{bmatrix}
\begin{bmatrix}
    \beta_0 \\
    \beta_1 \\
    \beta_2
\end{bmatrix}
+ 
\begin{bmatrix}
    e_1 \\
    e_2 \\
    \vdots \\
    e_n
\end{bmatrix}
\]
Become comfortable with $X\beta$

- $X\beta$ is the “linear predictor”, the substantively important variables are combined together into a single number.
- Whether in OLS or Generalized Linear Models, the key aspect of this is the same.
  
  _All of the predictive elements boil down to a single number_

- Which may then be transformed . . .
- In matrices:

  $$y = X\beta + \varepsilon$$

  rather than

  $$y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \ldots + \varepsilon_i$$

- $y$ is an $(N \times 1)$ column, $X$ is an $(N \times p)$ (where $p = 7 + 1$) rectangular matrix, $\varepsilon$ is $(N \times 1)$. 

Why Matrices?

- Notational compactness -> clarity of thought!
- It's not necessarily faster in the computer, but might be with clever algorithms (BLAS).
- You can't really read the literature unless you invest some effort to learn this
## OLS Regression in Matrices

### Objectives

**OLS Regression in Matrices**

<table>
<thead>
<tr>
<th>dep. var</th>
<th>Slopes</th>
<th>indep var</th>
<th>estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\begin{bmatrix} y_1 \ y_2 \ \vdots \ y_N \end{bmatrix}$</td>
<td>$\begin{bmatrix} \beta_0 \ \beta_1 \ \vdots \ \beta_p \end{bmatrix}$</td>
<td>$\begin{bmatrix} 1 &amp; x_{11} &amp; x_{21} \ 1 &amp; x_{12} &amp; x_{22} \ \vdots &amp; \vdots &amp; \vdots \ 1 &amp; x_{1N} &amp; x_{2N} \end{bmatrix}$</td>
<td>$\begin{bmatrix} \hat{\beta}_0 \ \hat{\beta}_1 \ \vdots \ \hat{\beta}_p \end{bmatrix}$</td>
</tr>
</tbody>
</table>

### Predicted values

$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = X\hat{\beta}$

### Residuals

$\hat{\varepsilon} = y - \hat{y} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_N - \hat{y}_N \end{bmatrix}$
One thing I do which is not standard

- I refer to the elements inside the matrix $X$ by style “variable name” subscript $i$.

$$X = \begin{bmatrix}
\text{x1}_1 & \text{x2}_1 \\
\text{x1}_2 & \text{x2}_2 \\
\vdots & \vdots \\
\text{x1}_N & \text{x2}_N
\end{bmatrix} \quad \text{My Way}$$

I’m thinking of the matrix as a collection of columns, so inside the matrix, I just refer to them by the same notation we would use to refer to individual columns.

- Often, the first column is a column of 1’s, the regression intercept
One thing I do which is not standard ...

$$X = \begin{bmatrix}
1 & x_{11} & x_{21} \\
1 & x_{12} & x_{22} \\
\vdots & \vdots & \vdots \\
1 & x_{1N} & x_{2N}
\end{bmatrix} \quad \text{My Way}$$

- The more popular notation in stats is to refer to $X$ like this

$$X = \begin{bmatrix}
x_{11} & x_{12} & x_{13} \\
x_{21} & x_{22} & x_{23} \\
\vdots & \vdots & \vdots \\
x_{N1} & x_{N2} & x_{N3}
\end{bmatrix} \quad \text{Usual Way}$$

where $i$ is the row and $j$ is the column.

- The usual way bothers me because it destroys the emotional connection I have to my columns!
One thing I do which is not standard ... 

- Example, combine variables *Religion*, *Income* into a matrix, and add an intercept term too

\[
\begin{bmatrix}
\text{Religion}_1 \\
\text{Religion}_2 \\
\text{Religion}_3 \\
\text{Religion}_4 \\
\text{Religion}_5 \\
\text{Religion}_6
\end{bmatrix},
\begin{bmatrix}
\text{Income}_1 \\
\text{Income}_2 \\
\text{Income}_3 \\
\text{Income}_4 \\
\text{Income}_5 \\
\text{Income}_6
\end{bmatrix}
\rightarrow
\begin{bmatrix}
1 & \text{Religion}_1 & \text{Income}_1 \\
1 & \text{Religion}_2 & \text{Income}_2 \\
1 & \text{Religion}_3 & \text{Income}_3 \\
1 & \text{Religion}_4 & \text{Income}_4 \\
1 & \text{Religion}_5 & \text{Income}_5 \\
1 & \text{Religion}_6 & \text{Income}_6
\end{bmatrix}
\]

- To mathematicians, $X_1$, is a row \{1, *Religion*$_1$, *Income*$_1$\}.

- But if I want to scan down my *Religion* column, I’m forced to have a first-subscript varying faster (my mind is thinking $x_{2i}$, but the usual matrix notation makes me think $x_{i2}$. I don’t like the fact that the column number becomes the 2nd subscript, when to me it is the important thing.
One thing I do which is not standard ...

\[
\begin{bmatrix}
x_{12} \\
x_{22} \\
x_{32} \\
x_{42}
\end{bmatrix}
\]

- Honestly, though, it is just an emotional hangup for me.
- Some people get past it by referring to column two with “missing index” notation. If you mean to refer to all of the rows, but the second column, put a period in the row index position

\[X_{.2}\]
Outline

1. Objectives
2. Vector
3. Matrix
4. Decompositions
5. Regression
Terminology: vector versus “row vector.”

- By custom, a vector is a column vector.

\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \]

In social science (statistics in general), it is very conventional to refer to the sample size as \( N \) or \( n \), so we’ll do that.

- If we need a row vector, we apply the transpose that:

\[ y^T = [y_1, y_2, \ldots, y_N] \]

transpose: turns a column into a row

- A row vector “saves space” on the printed page, thus we a lot of mental gymnastics where people want to write about a column but they are instead forced to write about \((some \ column)^T\)
Multiply 2 vectors (Inner Product)

- The product as a row times a column

\[ x^T \cdot y \]

\[
\begin{bmatrix}
  a & b & c & d & e \\
\end{bmatrix}
\begin{bmatrix}
  f \\
  g \\
  h \\
  i \\
  j \\
\end{bmatrix} = af + bg + ch + di + ej
\]

- \([1 \times 5] \cdot [5 \times 1]\) yields a \([1 \times 1]\) result, just a single number (a “\textbf{scalar}”)
- Vector Multiplication is NOT DEFINED if the inner 5’s in \([1 \times 5] \cdot [5 \times 1]\) don’t match (Vectors must “conform”)
- Linear Algebra: This is a “dot product” (AKA “inner product”)
Multiply 2 vectors (Inner Product) ...

- Symbolized in many books as

\[ < x, y > \text{ or sometimes } (x, y) \]

- One of the most common uses will calculate “sum of squares”

\[
\begin{bmatrix}
a \\
b \\
c \\
d \\
e
\end{bmatrix}
\cdot
\begin{bmatrix}
a \\
b \\
c \\
d \\
e
\end{bmatrix} = a^2 + b^2 + c^2 + d^2 + e^2
\]

- Example: Sum of squared errors the scalar way:
Multiply 2 vectors (Inner Product) ...

\[ \sum_{i}^{N} (y_i - \hat{y}_i)^2 \]

\[ = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y})^2 + \ldots \]

Example: Sum of squared errors the vector way:

\[ \text{residuals} : y - \hat{y} \]

\[ (y - \hat{y})^T = (y_1 - \hat{y}_1, y_2 - \hat{y}_2, y_3 - \hat{y}_3 \ldots, y_N - \hat{y}_N) \]

\[ (y - \hat{y})^T (y - \hat{y}) = (y_1 - \hat{y}_1, y_2 - \hat{y}_2, y_3 - \hat{y}_3 \ldots, y_N - \hat{y}_N) \]

\[
\begin{bmatrix}
  y_1 - \hat{y}_1 \\
  y_2 - \hat{y}_2 \\
  \vdots \\
  y_N - \hat{y}_N
\end{bmatrix}
\]
Multiply 2 vectors (Inner Product) ...

- Recall the pythagorean theorem, the length of a hypotenuse is square root of the sum of squared sides?

  \[ a^2 + b^2 = c^2 \implies c = \sqrt{a^2 + b^2} \]

- You see here that the sum of squares is the squared “length” of the residual vector.

- In linear algebra, the length of a vector is called its “norm”, represented as \( \| x \| \). Sometimes they put a subscript 2 on that to remember it is a distance calculated by a sum of squares, \( \| x \|_2 \).

- Think of the sum of squares as a squared norm of the residual.

  \[ (y - \hat{y})^T (y - \hat{y}) = \| y - \hat{y} \|_2^2 \]
Outline

1. Objectives
2. Vector
3. Matrix
4. Decompositions
5. Regression
Multiply a matrix times a vector

\[
\begin{bmatrix}
  a & b & c & d & e \\
  r & s & t & u & v
\end{bmatrix}
\cdot
\begin{bmatrix}
  f \\
  g \\
  h \\
  i \\
  j
\end{bmatrix}
= \begin{bmatrix}
  af + bg + ch + di + ej \\
  rf + sg + th + ui + vj
\end{bmatrix}
\]

- Treat matrix as two rows, then conduct multiplication separately for each one.
- \([2 \times 5] \cdot [5 \times 1]\) yields a \([2 \times 1]\) result
- Example: \(\hat{y} = Xb\)
Multiply a matrix times a vector ...

\[
\begin{bmatrix}
\hat{y}_1 \\
\hat{y}_2 \\
\vdots \\
\hat{y}_N \\
\end{bmatrix} = \begin{bmatrix}
1 & x_{11} & x_{21} \\
1 & x_{12} & x_{22} \\
1 & \ldots & \ldots \\
1 & x_{1N} & x_{2N} \\
\end{bmatrix} \begin{bmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\vdots \\
\end{bmatrix} = \begin{bmatrix}
\hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{21} \\
\hat{\beta}_0 + \hat{\beta}_1 x_{12} + \hat{\beta}_2 x_{22} \\
\vdots \\
\hat{\beta}_0 + \hat{\beta}_1 x_{1N} + \hat{\beta}_2 x_{2N} \\
\end{bmatrix}
\]

- Example with numbers:

\[
\begin{bmatrix}
1 & 19 & 1 & 0.1 & 1 & 0 & 22 & 3 \\
1 & 22 & 2 & 1.1 & 0 & 1 & 42 & 1 \\
1 & 17 \\
\vdots & \vdots \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5 \\
\beta_6 \\
\beta_7 \\
\end{bmatrix}
\]

\( \text{(1)} \)

Result.
Multiply a matrix times a vector ...

Row 1: \[ 1 \cdot \beta_0 + \beta_1 \cdot 19 + \beta_2 \cdot 1 + \beta_3 \cdot 0.1 + \beta_4 \cdot 1 + \beta_5 \cdot 0 + \beta_6 \cdot 22 + \beta_7 \cdot 3 \] (2)

Row 2: \[ \beta_0 + \beta_1 \cdot 22 + \beta_2 \cdot 2 + \beta_3 \cdot 1.1 + \beta_4 \cdot 0 + \beta_5 \cdot 1 + \beta_6 \cdot 42 + \beta_7 \cdot 1 \] (3)
Multiply a matrix times a matrix

\[
\begin{bmatrix}
a & b & c & d & e \\
r & s & t & u & v
\end{bmatrix}
\cdot
\begin{bmatrix}
f & k \\
g & l \\
h & m \\
i & n \\
j & o
\end{bmatrix}
\]

= \[
\begin{bmatrix}
af + bg + ch + di + ej & ak + bl + cm + dn + eo \\
rf + sg + th + ui + vj & rk + sl + tm + un + vo
\end{bmatrix}
\]

- Break into sequences of vector multiplications, row 1 \cdot column 1, row2 \cdot column 1, row 1 \cdot column 2, row 2 \cdot column 2.
- \([2 \times 5] \cdot [5 \times 2]\) yields a \([2 \times 2]\) result
Transpose

- $X^T$ means “$X$ transpose”, or “$X$ turned on its side”

\[
X^T = \begin{bmatrix}
1 & 1 & 1 & \ldots & 1 \\
x_1^T & x_2^T & x_3^T & \cdots & x_N^T
\end{bmatrix}
\]

- The left column becomes the first row

\[
X = \begin{bmatrix}
1 & 3 & 33 \\
1 & 2 & 62 \\
1 & 5 & 65 \\
1 & 1 & 45 \\
1 & 5 & 66
\end{bmatrix}
\]

$X$ is 5x3

\[
X^T = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 \\
3 & 2 & 5 & 1 & 5 \\
33 & 62 & 65 & 45 & 66
\end{bmatrix}
\]

$X^T$ is 3x5
Matrix $X^T X$ can be much smaller than $X$ or $X^T$

- Suppose $X$ is $N \times p$
- Then $X^T X$ means we are multiplying a $(p \times N)$ and an $(N \times p)$
- The result is $(p \times p)$.
- In the pencil-and-paper days of stats, hard work would go into calculating $X^T X$
- Estimates of all important quantities can be phrased $X^T X$. 
I’ve forgotten most of the detailed rules of “Linear Algebra I”, which I took in 1983. However, one sticks with me.

Rule to remember: The transpose of a product is the product of the transposes, in reverse order.

\[(XY)^T = Y^T X^T\]

This works if many matrices are multiplied together.
Variance-Covariance Matrix

- The variance of a column variable $x_1$

$$\text{Var}(x_1) = \frac{\sum_{i=1}^{N}(x_{1i} - \bar{x_1})^2}{N} \quad (4)$$

- can be re-arranged as

$$\text{Var}(x_1) = \frac{\sum_{i=1}^{N}x_{1i}^2}{N} - (\bar{x_1})^2 \quad (5)$$

$$= \text{mean of } \{x_1 \text{ squared}\} - \{\text{mean of } x_1\} \text{ squared} \quad (6)$$

$$= \frac{\sum_{i=1}^{N}x_{1i}^2}{N} - (\bar{x_1})^2$$

- In the pencil-and-paper days, it was noted that to calculate variance with formula 4 required 2 passes through the data

1. Calculate $\bar{x_1} = \frac{1}{N} \sum x_{1i}$
2. Calculate the squares $\sum(x_{1i} - \bar{x_1})$
Variance-Covariance Matrix ...

- The shorter formula (5), on the other hand, requires one pass through the rows, during which we calculate both.

\[
\begin{align*}
1 & \sum x_{1i} \\
2 & \sum x_{1i}^2
\end{align*}
\]

- The Covariance of 2 variables is similar

\[
Cov(x_1, x_2) = \frac{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{N}
\]

This can be written with the same simplifying trick as

\[
Cov(x_1, x_2) = \frac{\sum_{i=1}^{N}(x_{1i} \cdot x_{2i})}{N} - \bar{x}_1 \cdot \bar{x}_2
\]

\[= mean \ of \ \{x_1 \cdot x_2\} - \{mean \ of \ x_1\}\{mean \ of \ x_2\}\]
At least in the pencil-and-paper math era, it was possible to see that variety is summarized if we

$$\sum x_1^2, \sum x_2^2 \text{ and } \sum x_1 \cdot x_2$$

for all the pairs of variables.

The variance/covariance matrix, often just called the variance matrix

$$\text{Var}(X) = \begin{bmatrix}
\text{Var}(x_1) & \text{Cov}(x_1, x_2) & \text{Cov}(x_1, x_3) & \text{Cov}(x_1, x_4) & \text{Cov}(x_1, x_5) \\
\text{Cov}(x_2, x_1) & \text{Var}(x_2) & \text{Cov}(x_2, x_3) & \text{Cov}(x_2, x_4) & \text{Cov}(x_2, x_5) \\
\text{Cov}(x_3, x_1) & \text{Cov}(x_3, x_2) & \text{Var}(x_3) & \text{Cov}(x_3, x_4) & \text{Cov}(x_3, x_5) \\
\text{Cov}(x_4, x_1) & \text{Cov}(x_4, x_2) & \text{Cov}(x_4, x_3) & \text{Var}(x_4) & \text{Cov}(x_4, x_5) \\
\text{Cov}(x_5, x_1) & \text{Cov}(x_5, x_2) & \text{Cov}(x_5, x_3) & \text{Cov}(x_5, x_4) & \text{Var}(x_5)
\end{bmatrix}$$

Making it obvious that

- $\text{Var}(x_1) = \text{Cov}(x_1, x_1)$
- $\text{Cov}(x_2, x_1) = \text{Cov}(x_1, x_2)$, so this is symmetric (same in lower triangle as upper right triangle)
Variance-Covariance Matrix ...

- The Standard Deviations of the variables are the square root of the diagonal of $\text{Var}(x)$
  $$\sqrt{\text{diag}(\text{Var}(X))}$$
- It is very common to use the letter $\sigma$ with subscripts to fill in this kind of matrix. The matrix is often named capital sigma, $\Sigma$.
- In the “usual” matrix style
  $$\Sigma = \begin{bmatrix}
  \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\
  \sigma_{21} & \sigma_{22} & \sigma_{23} & & \\
  & \ddots & \ddots & \ddots & \\
  \sigma_{51} & & & \sigma_{55}
  \end{bmatrix}$$
- But if you stubbornly stick to my kind of notation, you have to do
Variance-Covariance Matrix …

\[
\Sigma = \begin{bmatrix}
\sigma_{x_1}^2 & \sigma_{x_1 x_2} & \sigma_{x_1 x_3} & \sigma_{x_1 x_4} \\
\sigma_{x_2 x_1} & \sigma_{x_2}^2 & \sigma_{x_2 x_3} & \\
\sigma_{x_3 x_1} & \sigma_{x_3 x_2} & \sigma_{x_3}^2 & \\
\sigma_{x_4 x_1} & \sigma_{x_4 x_2} & \sigma_{x_4 x_3} & \\
\sigma_{x_5 x_1} & \sigma_{x_5 x_2} & \sigma_{x_5 x_3} & \sigma_{x_5 x_4}^2 \\
\end{bmatrix}
\]

- Here we see the standard deviations are just the square root of the diagonal. Do you want to call that $S$ or $\sigma$?
Sum of Squares and Crossproducts

- Think of the data analysis situation in which we have columns representing variables and the rows are the observations (cases).

$$
\begin{bmatrix}
  1 & x_{11} & x_{21} \\
  1 & x_{12} & x_{22} \\
  1 & \ldots & \ldots \\
  1 & x_{1N} & x_{2N}
\end{bmatrix}
$$

- $X^TX$ is a sum of squares and cross-products matrix.
Sum of Squares and Crossproducts ...

\[
X^TX = \begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
x_1 & x_1 & x_1 & \cdots & x_1 \\
x_2 & x_2 & x_2 & \cdots & x_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_N & x_N & x_N & \cdots & x_N \\
\end{bmatrix}
\begin{bmatrix}
1 & x_1 & x_2 \\
1 & x_1 & x_2 \\
\vdots & \vdots & \vdots \\
1 & x_1 & x_2 \\
\end{bmatrix}
\]

\[
= \begin{bmatrix}
N & \sum x_1 & \sum x_1^2 & \sum x_1x_2 \\
\sum x_2 & \sum x_1x_2 & \sum x_2^2 \\
\end{bmatrix}
\]

(7)

- Each element of the matrix is a “sum of \{\text{squares, crossproducts}\}”
- Refer to a column
Sum of Squares and Crossproducts ...

\[ x_1^T x_1 = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \sum x_1^2 \]

- This matrix is *symmetric*. 
The term “sweep operator” refers to the act of
- Calculating the mean of each column of $X$,
- Subtracting that result from each row in the matrix.

The result is said to be “mean-centered”

If we start with this

\[
\begin{bmatrix}
1 & x_{11} & x_{21} & x_{31} & x_{41} & x_{51} \\
1 & x_{12} & x_{22} & x_{32} & x_{42} & x_{52} \\
1 & \cdots & \cdots & \cdot & \\
1 & x_{1N} & x_{2N} & x_{3N} & x_{5N} \\
\end{bmatrix}
\]

the means will be \(\{1, \overline{x_1}, \overline{x_2}\}\) (intercept was in column 1)
The mean centered version has a column of 0’s on the first column

\[
XC = \begin{bmatrix}
0 & xc1_1 & xc2_1 & xc3_1 & xc4_1 & xc5_1 \\
0 & xc1_2 & xc2_2 & xc3_2 & xc4_2 & xc5_2 \\
0 & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & xc1_N & xc2_N & xc3_N & xc4_N & xc5_N
\end{bmatrix}
\]

Simplifying benefits of working with mean centered data

- We lose a column, we generally throw away the first one, since it is all zeros

\[
XC = \begin{bmatrix}
xc1_1 & xc2_1 & xc3_1 & xc4_1 & xc5_1 \\
xc1_2 & xc2_2 & xc3_2 & xc4_2 & xc5_2 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
xc1_N & xc2_N & xc3_N & xc4_N & xc5_N
\end{bmatrix}
\]
The Variance and Covariance formulas simplify and we can read off their values from $XC^T XC$. Variance matrix is

$$
\text{Var}(XC) = \begin{bmatrix}
\frac{1}{N} \sum xc_1^2_i & \frac{1}{N} \sum xc_1 xc_2_i & \frac{1}{N} \sum xc_1 xc_3_i & \frac{1}{N} \sum xc_1 xc_4_i & \frac{1}{N} \sum xc_1 xc_5_i \\
\frac{1}{N} \sum xc_1 xc_2_i & \frac{1}{N} \sum xc_2^2_i & \frac{1}{N} \sum xc_2 xc_3_i & \frac{1}{N} \sum xc_2 xc_4_i & \frac{1}{N} \sum xc_2 xc_5_i \\
\frac{1}{N} \sum xc_1 xc_3_i & \frac{1}{N} \sum xc_2 xc_3_i & \frac{1}{N} \sum xc_3^2_i & \frac{1}{N} \sum xc_3 xc_4_i & \frac{1}{N} \sum xc_3 xc_5_i \\
\frac{1}{N} \sum xc_1 xc_4_i & \frac{1}{N} \sum xc_2 xc_4_i & \frac{1}{N} \sum xc_3 xc_4_i & \frac{1}{N} \sum xc_4^2_i & \frac{1}{N} \sum xc_4 xc_5_i \\
\frac{1}{N} \sum xc_1 xc_5_i & \frac{1}{N} \sum xc_2 xc_5_i & \frac{1}{N} \sum xc_3 xc_5_i & \frac{1}{N} \sum xc_4 xc_5_i & \frac{1}{N} \sum xc_5^2_i
\end{bmatrix}
$$

The variance of $x1$ would ordinarily be $\frac{1}{N} \sum x1_i^2 - (\bar{x1})^2$, but the last part is now 0 because we are mean centered.

And if you factor out the $1/N$, then you see the variance matrix for a centered variable is simply proportional to $XC^T XC$.

I hope you agree it is “easy to see” that $\frac{1}{N} XC^T XC$ is a variance matrix, but the next claim is not easy for me to prove.

- $\frac{1}{N} X^T X$ is the covariance matrix even if $X$ is not mean-centered.
- There is a proof in Greene’s *Econometric Analysis*
Some projects require input in the form of a COVARIANCE MATRIX (e.g., simulating from MVN)

Covariance matrix is difficult to comprehend.

My solution: Re-produce covariance by multiplying correlation and standard deviations together.

Correlation matrix is understandable, all between -1 and +1.

\[
\rho = \begin{bmatrix}
1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\
\rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2p} \\
\rho_{31} & \cdots & 1 & \cdots & \rho_{3p} \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\rho_{p1} & \rho_{11} & \rho_{11} & \cdots & 1
\end{bmatrix}
\]

Relationship between Covariance and Correlation...

- **Standard deviation**: The diagonal from variance matrix would be

\[
\begin{bmatrix}
\sigma_1^2 & 0 & 0 & 0 & 0 \\
0 & \sigma_2^2 & 0 & 0 & 0 \\
0 & 0 & \sigma_3^2 & 0 & 0 \\
0 & 0 & 0 & \sigma_4^2 & 0 \\
0 & 0 & 0 & 0 & \sigma_5^2
\end{bmatrix}
\]

- **Covariance!**

\[
\text{Covariance} =
\begin{bmatrix}
\sigma_1 & 0 & 0 & 0 & 0 \\
0 & \sigma_2 & 0 & 0 & 0 \\
0 & 0 & \sigma_3 & 0 & 0 \\
0 & 0 & 0 & \sigma_4 & 0 \\
0 & 0 & 0 & 0 & \sigma_5
\end{bmatrix}
\begin{bmatrix}
1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\
\rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2p} \\
\rho_{31} & \cdots & 1 & \cdots & \rho_{3p} \\
\vdots & \ddots & 11 & \cdots & \rho_{p1} \\
\rho_{p1} & \rho_{11} & \rho_{11} & \cdots & 1
\end{bmatrix}
\]

- **Can write more succinctly.** Take out the diagonal as a vector:
Relationship between Covariance and Correlation ...

\[ \sigma = \sqrt{\text{diag}(\text{Var}(X))} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \vdots \\ \sigma_N \end{bmatrix} \]

- The Covariance can be calculated as

\[ \sigma \rho \sigma^T \]
Re-scaling variables

- Given a variance matrix $\text{Var}(x)$
- Rescale that matrix by multiplying by some weight matrix $C$

\[ y = Cx \]

- The variance matrix of $Cx$ is

\[ \text{Var}(y) = C^T \text{Var}(x) C \]
Inverse Matrix

- Recall math with scalars. The reciprocal is an inverse, $k^{-1} = 1/k$, in the sense that $(k^{-1}) \cdot k = 1$
- Possible to calculate inverses for some matrices in same way
- Consider a square matrix $A$. Not all square matrices are invertible, but in order for an inverse to exist, the matrix must be square
- If $A$ is square and invertible, then

\[
A^{-1} A = I, \quad \text{where } I = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

- The derivation of Ordinary Least Squares involves $X^T X$, a product that happens to be
  - square and
Inverse Matrix ...

- usually invertible.

\[(X^T X)^{-1}(X^T X) = I, \text{ where } I = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}\] (9)

- Note we don’t require that $X$ is square, but $X^T X$ is always a square.
Inverse does not always exist

- The inverse of a matrix with redundant columns does not exist.
- Redundant: linearly dependent, one column can be reproduced by adding together the other ones.

\[
\begin{bmatrix}
  a_1 & b_1 & c_1 & d_1 \\
  a_2 & b_2 & c_2 & d_2 \\
  a_3 & b_3 & c_3 & d_3 \\
  a_4 & b_4 & c_4 & d_4 \\
\end{bmatrix}
\]
Inverse does not always exist ... 

- One vector $a$ is **linearly dependent** on $b$ if it is possible to find a coefficient $k$ so that

\[ a = kb \]

I’ll write this one out fully

\[
\begin{bmatrix}
 a_1 \\
 a_2 \\
 a_3 \\
 a_4 \\
 a_5 \\
\end{bmatrix}
 = k

\begin{bmatrix}
 b_1 \\
 b_2 \\
 b_3 \\
 b_4 \\
 b_5 \\
\end{bmatrix}
\]

- One vector $a$ is dependent on $b$, $c$, and $d$ if there are coefficients $k_1$, $k_2$, $k_3$ such that

\[ a = k_1 b + k_2 c + k_3 d \]
Inverse does not always exist ...

- The vectors are **linearly independent** if there is no such vector of $k$'s to make the equality hold.
- The **rank** of a matrix is the number of linearly independent columns.
- A matrix is said to be singular if the rank is less than the number of columns.
- There are a number of rules of thumb and diagnostic procedures that can be applied to assess the rank of a matrix and find out if it is invertible.
Two more rules I remember from 1983

1. Inverse of a product
   - The inverse of a product is the product of the inverses of the individual matrices, in reverse order.
     \[(XY)^{-1} = Y^{-1}X^{-1}\]
     \[(XYZA)^{-1} = A^{-1}Z^{-1}Y^{-1}X^{-1}\]
   - I should mention, of course, this assumes the thing being inverted is invertible (is square, etc).

2. Transpose of an inverse equals Inverse of Transpose
   \[(X^T)^{-1} = (X^{-1})^T\]

   These things come in handy in regression analysis because we often end up with these complicated looking matrix expressions for parameter variance that we want to simplify (look up Huber-White corrected variance estimator, for example)
Orthogonal Matrix, Orthonormal Matrix

- Consider a set of columns that makes up a matrix
  \[ X = \{ X_1, X_2, X_3, X_4 \} \]

- Orthogonal means that
  - the dot product of any column with any other column is 0
    \[ X_1^T \cdot X_2 = 0 \]
  - Those two columns are “unrelated”, there’s no tendency of positive values to coincide, or negative values to coincide. They “uncorrelated” in a statistical sense. If we plotted their values in a 2-D scatterplot, there would be no visible “relationship”.
Orthogonal Matrix, Orthonormal Matrix ...

Property 1: The product is a “diagonal” matrix

\[
X^T X = \begin{bmatrix}
a & 0 & 0 & 0 & 0 \\
0 & b & 0 & 0 & 0 \\
0 & 0 & c & 0 & 0 \\
0 & 0 & 0 & d & 0 \\
0 & 0 & 0 & 0 & e \\
\end{bmatrix}
\]

The product is thus diagonal, but the elements on the main diagonal are not all equal to the same number.

Property 2: The product is an identity matrix:

\[
X^T X = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\]
Orthogonal Matrix, Orthonormal Matrix …

- Property 2 holds when the columns of $X$ are “unit vectors”, meaning they have length 1.
  - Recall the norm (aka length) of $x$ is, following the Pythagorean theorem:

  $$\|x\| = \sqrt{x_1^2 + x_2^2 + \ldots}$$

  $$\|x\|^2 = x^T x$$

- Sometimes I become confused because when people say “orthogonal”, they simply mean property 1 holds, while others discussing orthogonal matrices are referring to property 2.
  - Orthogonal $\implies$ All zeros above and below the diagonal of $X^TX$
  - Orthonormal $\implies$ Columns are scaled so that their norms are all 1, ie $X_1^TX_1 = 1$, $X_2^TX_2 = 1$.

- Notice how convenient a diagonal matrix is when we need its inverse.
Orthogonal Matrix, Orthonormal Matrix ...

\[
\begin{bmatrix}
a & 0 & 0 & 0 & 0 \\
0 & b & 0 & 0 & 0 \\
0 & 0 & c & 0 & 0 \\
0 & 0 & 0 & d & 0 \\
0 & 0 & 0 & 0 & e
\end{bmatrix}
\begin{bmatrix}
1/a & 0 & 0 & 0 & 0 \\
0 & 1/b & 0 & 0 & 0 \\
0 & 0 & 1/c & 0 & 0 \\
0 & 0 & 0 & 1/d & 0 \\
0 & 0 & 0 & 0 & 1/e
\end{bmatrix} = I
\]

- Orthonormal Matrices Simplify Everything
  - If \( X \) is an orthogonal matrix of unit vectors (orthonormal), then
    \[ XX^T = X^T X = I \]
  - Hence, the inverse of \( X^T X \) or \( XX^T \) is EXTREMELY SUPER VERY EASY to calculate: Just Transpose. (With this, regression analysis is terrific, fun, numerically stable and life-extending!)
  - Another quirk: An orthogonal matrix changes the “direction” of a vector, but it does not change its magnitude.
Orthogonal Matrix, Orthonormal Matrix ...

- Let $Q$ be an orthogonal matrix and let $x$ be a vector.
- Claim: The norm of $Qx$ is equal to the norm of $x$

$$\|Qx\| = \|x\|$$

- Proof

$$\|Qx\|^2 = (Qx)^T(Qx)$$
$$= x^TQ^TQx$$
$$= x^Tx$$

- The significance of this is that the orthonormal matrix can be thought of as a “direction changer” or “reflector” of any vector $x$. 
Outline

1. Objectives

2. Vector

3. Matrix

4. Decompositions

5. Regression
Equivalent, Smaller, More Stable

What’s not to like? Faster! Smaller! More accurate!

- After you get past the first layer of matrix algebra, you become interested in doing computations on observed data
- Possibly huge data.
- Missions in the decomposition process: find a representation that
  1. has more accurate results in a digital computer (numerical linear algebra)
  2. allows symbolic manipulation and simplification so as to avoid huge storage and calculation costs
- Two of the decompositions that arise most frequently in regression contexts are the Cholesky decomposition and the QR decomposition
Cholesky decomposition: “matrix square root”

- Given a square matrix, such as

\[
\Sigma = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1N} \\
\sigma_{21} & \sigma_{22} & & \\
\vdots & & & \\
\sigma_{N1} & & & \sigma_{NN}
\end{bmatrix}
\]

- Find an upper-triangular matrix of the same size

\[
\begin{bmatrix}
c_{11} & c_{12} & \cdots & c_{1N} \\
0 & c_{22} & & c_{2N} \\
\vdots & 0 & & \\
0 & 0 & 0 & c_{NN}
\end{bmatrix}
\]
Cholesky decomposition: “matrix square root” ...

so that if you multiply that by its transpose, then you recover $\Sigma$.

$$\Sigma = \begin{bmatrix}
c_{11} & 0 & \cdots & 0 \\
c_{12} & c_{22} & \cdots & 0 \\
\vdots & 0 & \ddots & \vdots \\
c_{1N} & 0 & 0 & c_{NN}
\end{bmatrix} \begin{bmatrix}
c_{11} & c_{12} & \cdots & c_{1N} \\
0 & c_{22} & \cdots & c_{2N} \\
\vdots & 0 & \ddots & \vdots \\
0 & 0 & 0 & c_{NN}
\end{bmatrix}$$

$$\Sigma = R^T R$$

- See why they call this a matrix square root, yes?
- The Cholesky decomposition does not always exist, but it does for most of the matrices that we are interested in (positive definite). In particular, it exists for variance matrices (and anything created as a cross product, such as $X^TX$).
- There is a computational benefit here on a number of levels, but right now I’m interested in the symbolic benefit.
Cholesky decomposition: "matrix square root" ...

- We often arrive at a point where we have a calculation like this

\[ y^T \Sigma y \quad (10) \]

- Replace \( \Sigma \) by its Cholesky decomposition

\[ y^T R^T Ry \]

- Put some parentheses in to make clear what’s coming next

\[ (y^T R^T)(Ry) \]

- Because of the rule “the transpose of a product is the product of transposes in reverse order”,

\[ y^T R^T = (Ry)^T \]
Hence we can simplify the calculation of (10) if we create an updated, “re-weighted” version of \( \tilde{y} \)

\[
\tilde{y} = Ry
\]

Then (10) is simply

\[
\tilde{y}^T \tilde{y}
\]

All the algorithms and statistical results for sums of squares, for example, now are applicable to this re-weighted version of \( y \).

This is how we gracefully manage Weighted Least Squares, Generalized Least Squares, Penalized Least Squares, etc.
QR decomposition

- Cholesky gives a decomposition of a square (positive definite) matrix.
- Other decompositions can work on an $N \times p$ matrix.
- The QR decomposition says that a matrix can be reproduced as the product of 2 parts,
  1. An orthogonal matrix $Q$
  2. An upper right triangular $R$ (with rows of 0’s padding the bottom so that it is length $N$).
- Suppose $X$ is $m \times n$. If the QR decomposition exists, we can reproduce $X$ as

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$$
QR decomposition ...

- The bottom part of the stack, \( \begin{bmatrix} R \\ 0 \end{bmatrix} \), is \( m - n \) rows of 0's:

\[
\begin{bmatrix}
  r_{11} & r_{12} & r_{13} & r_{14} & \cdots & r_{1n} \\
  0 & r_{22} & r_{23} & r_{24} & \cdots & r_{2n} \\
  0 & 0 & r_{33} & r_{34} & \cdots & r_{3n} \\
  0 & 0 & 0 & \cdots & \cdots & r_{(m-1)n} \\
  0 & 0 & 0 & 0 & \cdots & r_{mm} \\
  0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

- I call that the “full sized” version of QR. Q is \( m \times m \), which is Larger Than The Original Data. It is numerically tractable, but expensive to store in a computer.
**Fear not!** But we don’t often don’t need all that information: The fact that the bottom rows of $R$ are all zeros means we don’t generally need the right side of $Q$!

\[
X = \begin{bmatrix}
q_{11} & q_{12} & q_{14} & q_{1m} \\
q_{21} & & & \\
& & \ddots & \\
q_{m1} & & q_{mm} & \\
\end{bmatrix}
\begin{bmatrix}
r_{11} & r_{12} & r_{13} & r_{14} & r_{1n} \\
0 & r_{22} & r_{23} & r_{24} & r_{2n} \\
0 & 0 & r_{33} & r_{34} & r_{3n} \\
0 & 0 & 0 & \ddots & r_{(n-1)n} \\
0 & 0 & 0 & 0 & r_{nn} \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

The last $m - n$ columns of $Q$ get zeroed out by the 0’s on the bottom of $R$. 

(12)
QR decomposition ...

- So the more “petite” version \((Q_f)\) would give

\[
X = Q_f R
\]  

(13)

\[
X = \begin{bmatrix}
q_{11} & q_{12} & q_{13} & \ldots & q_{1n} \\
q_{21} & q_{22} & q_{23} & \ldots & q_{2n} \\
q_{31} & q_{32} & \ldots & \ldots & \ldots \\
q_{41} & \ldots & \ldots & \ldots & \ldots \\
q_{m1} & \ldots & \ldots & \ldots & q_{mn}
\end{bmatrix}
\]

\[
\begin{bmatrix}
0 & 0 & r_{33} & r_{34} & \ldots & r_{3n} \\
0 & 0 & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & \ldots & r_{(n-1)n} \\
0 & 0 & 0 & 0 & 0 & r_{nn}
\end{bmatrix}
\]

We have no need to “pad” the bottom of \(R\) with 0’s and \(Q\) is now only \(n\) columns wide (still \(m\) rows deep)
Recall orthonormal matrices invert themselves: $QQ^T = I$ and $Q^T Q = I$. Literally, $Q^T = Q^{-1}$.

In regression analysis, we symbolically derive

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$  \hspace{1cm} (14)

A very accurate, reasonably fast way to calculate that is with QR. Replace $X$ by the petite $Q_f R$.

$$\hat{\beta} = ((Q_f R)^T(Q_f R))^{-1}(Q_f R)^T y$$  \hspace{1cm} (15)

If we use the rules for inverses and transposes mentioned above, we can algebraically reduce that:
QR decomposition ...

\[
\hat{\beta} = (R^T Q_f^T Q_f R)^{-1}(Q_f R)^T y 
\]
\[
(R^T R)^{-1}R^T Q_f^T y \tag{16}
\]
\[
R^{-1}R^{-1}R^T Q_f^T y \tag{17}
\]
\[
R^{-1}Q_f^T y \tag{18}
\]

More notes on this at http://pj.freefaculty.org/guides/stat/Math/Matrix-Decompositions
Outline

1. Objectives

2. Vector

3. Matrix

4. Decompositions

5. Regression
The objective function is to minimize

\[(y - \hat{y})^T (y - \hat{y})\]

or

\[\|y - \hat{y}\|_2\]

Results. symbolic solution:

\[\hat{\beta} = (X^T X)^{-1} X^T y\]

\[\text{Var}(\hat{\beta}) = \hat{\sigma}_e^2 (X^T X)^{-1}\]

(20)
What’s important in the formula?

- The matrix expression

\[ \hat{\beta} = (X^T X)^{-1} X^T y \]

\( (X^T X) \) is the “sum of squares and cross-products” matrix, similar to variance of x in one-predictor model.

- In the predicted value formula \( \hat{y} = X\hat{\beta} \), replace \( \hat{\beta} \).

\[ \hat{y} = X(X^T X)^{-1} X^T y \quad (21) \]

- The matrix \( H = X(X^T X)^{-1} X^T \) is size \( N \times N \). It serves as a weighting matrix that translates the outcome y into the predicted values.

  - \( H \) is called the “hat matrix”. See why? \( \hat{y} = H y \)