Statistical Theory for the Linear Model

Brad Jones¹

¹Department of Political Science
University of California, Davis

January 26, 2010
Today: Statistical Theory
Statistical Theory for the Linear Model

- First consider the statistical theory underlying the model.
- Second, consider implications of the properties as well as diagnostic methods.
- Requires use of matrix notation and matrix algebra.
- The next several slides are from the week 1 slide set. I will not go over them in class. They cover basic manipulations. We will skip to slides for regression.
Some Matrix Basics

- What is a matrix?
- “A rectangular array of elements arranged in rows and columns.”

\[
\begin{pmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0 \\
\end{pmatrix}
\]

- All data are really matrices.
- Suppose we have “votes,” “money,” and “PID.”
- Dimension of a matrix is $r \times c : 3 \times 3$. 
Some Matrix Basics

- Dimensionality
- If we had 435 MCs and 3 variables:

\[
\begin{pmatrix}
55 & 900 & 0 \\
67 & 1112 & 1 \\
73 & 525 & 0 \\
\vdots & \vdots & \vdots \\
67 & 874 & 1 \\
\end{pmatrix}
\]

- Dimensionality: $435 \times 3$
- Applied issue: always a good idea to know the dimensions of your research problem.
Some Matrix Basics

Symbolism

We can use symbols to denote $rc$:

$$
\begin{pmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{pmatrix}
$$

Often will denote matrices with bold-faced symbols:

$$
\mathbf{A} = \begin{pmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{pmatrix}
$$
Vectors

- A matrix with a single column is a “column vector.”

\[
\begin{pmatrix}
a_{11} \\
a_{21} \\
a_{31}
\end{pmatrix}
\]

- A matrix with a single row is a “row vector.”

\[
\begin{pmatrix}
a_{11} & a_{12} & a_{13}
\end{pmatrix}
\]

- Dimensions of the column vector?

- Dimensions of the row vector?

- Answers: 3 × 1; 1 × 3

- Linear models, y is usually an \( n \times 1 \) matrix.
The transpose of matrix $A$ is another matrix, $A'$. It is obtained by interchanging corresponding columns and rows. So if $A$ is:

$$
\begin{pmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{pmatrix}
$$

Then $A'$ is:

$$
\begin{pmatrix}
  a_{11} & a_{21} & a_{31} \\
  a_{12} & a_{22} & a_{32} \\
  a_{13} & a_{23} & a_{33}
\end{pmatrix}
$$

That is, the first column of $A$ is now the first row of $A'$. 
Transposing Matrices

- The transpose of a column vector is a row vector (and vice versa).
- Note that the dimensions of the input matrix $A$ is $r \times c$.
- For the transpose matrix, $A'$, the dimension is $c \times r$.
- Transposing matrices is really important in regression-like settings.
Fun with R

Generating “data” in R as vectors:

\[ x_0 \leftarrow \text{c}(1, 1, 1, 1, 1) \]
\[ x_1 \leftarrow \text{c}(2, 4, 6, 8, 10) \]
\[ x_2 \leftarrow \text{c}(7.3, 8.6, 9.2, 3.2, 1.2) \]
\[ y \leftarrow \text{c}(32, 14, 71, 81, 101) \]
Creating a matrix $X$:

```r
xmat <- cbind(x0, x1, x2); xmat

x0  x1  x2
[1,] 1  2  7.3
[2,] 1  4  8.6
[3,] 1  6  9.2
[4,] 1  8  3.2
[5,] 1 10  1.2

xmat[1,3]
[1] 7.3

xmat[1,1]
[1] 1

xmat[,2]
[1] 2 4 6 8 10

xmat[1,]
 x0  x1  x2
1.0  2.0  7.3
```
Basic Manipulations: Multiplication by a scalar:

$$2 \times \text{xmat}$$

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4 14.6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8 17.2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>12 18.4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>16  6.4</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>20  2.4</td>
</tr>
</tbody>
</table>
Basic Manipulations: Transposing $X$ to create $X'$:

$$xprime<-t(xmat); xprime$$

$$\begin{array}{cccccc}
[,]&[,]&[,]&[,]&[,]\\
x0&1.0&1.0&1.0&1.0&1.0 \\
x1&2.0&4.0&6.0&8.0&10.0 \\
x2&7.3&8.6&9.2&3.2&1.2 \\
\end{array}$$

Transpose $Y$ to create $Y'$:

$$y$$

$$\begin{array}{cccccc}
[1,]&[2,]&[3,]&[4,]&[5,]\\
[1,] & 32 & 14 & 71 & 81 & 101 \\
\end{array}$$

$$yprime<-t(y); yprime$$

$$\begin{array}{cccccc}
[,]&[,]&[,]&[,]&[,]\\
[1,]&[2,]&[3,]&[4,]&[5,]\\
[1,] & 32 & 14 & 71 & 81 & 101 \\
\end{array}$$
Multiplying Matrices

- Multiplying matrices will be important for us in order to derive variance and covariances.

- Multiplication of matrices having the same dimensions is not problematic.

- Suppose \( A \) is:
  
  \[
  \begin{pmatrix}
  2 & 5 \\
  4 & 1
  \end{pmatrix}
  \]

- Suppose \( B \) is:
  
  \[
  \begin{pmatrix}
  4 & 6 \\
  5 & 8
  \end{pmatrix}
  \]

- The product is a matrix, \( AB \). Its elements are obtained by finding the cross-products of rows of \( A \) with columns of \( B \).
Multiplying Matrices

- For these, the product matrix is obtained by:

\[
\begin{align*}
2(4) + 5(5) & = 33 \\
2(6) + 5(8) & = 52 \\
4(4) + 1(5) & = 21 \\
4(6) + 1(8) & = 32
\end{align*}
\]

- Product matrix \(AB\):

\[
\begin{pmatrix}
33 & 52 \\
21 & 32
\end{pmatrix}
\]

- Above, we say we postmultiplied \(A\) by \(B\) or premultiplied \(B\) by \(A\).

- Why the language? Multiplication is not always defined on matrices.
Multiplying Matrices

- Scalar multiplication has the commutative property:
  \[ xy = yx \]

- However, \( AB \) may not equal \( BA \).

- The product matrix is defined when the number of columns in \( A \) equals the number of rows in \( B \).

- When this condition holds, we say the matrices are conformable for multiplication.

- If \( A \) is \( 2 \times 2 \) and \( B \) is \( 2 \times 2 \), \( AB \) is conformable (and would have dimension \( 2 \times 2 \)).

- If \( A \) is \( 2 \times 3 \) and \( B \) is \( 3 \times 1 \), \( AB \) is conformable (and would have dimension \( 2 \times 1 \)).

- Note, \( BA \) is not conformable (why?).
Regression

- The OLS model can be written as:

\[ y = X\beta + \epsilon. \]  

- What are the dimensions of each matrix?

- They are:

\[
\begin{align*}
y & : n \times 1 \\
X & : n \times k \\
\beta & : k \times 1 \\
\epsilon & : n \times 1
\end{align*}
\]

- Note that \(X\beta\) is conformable for multiplication (i.e. we have one parameter for each column vector in \(X\)).
Basic Manipulations: Multiplying Matrix $X'$ by $X$ to create $X'X$.

```r
xprime_x <- xprime %*% xmat
>
xprime_x
   x0  x1  x2
x0  5.0 30.0 29.50
x1  30.0 220.0 141.80
x2  29.5 141.8 223.57
```

$X'$ had dimension $3 \times 5$.
$X$ had dimension $5 \times 3$.
Thus, $X'X$ is $3 \times 3$ (and is obviously conformable for multiplication (i.e. $n$ columns of $X'$ equal $n$ rows of $X$; i.e. $5=5$).
Inversion

➤ One of the most important manipulations we need to perform is matrix inversion.

➤ In matrix algebra, inversion is “kind of like” division.

➤ For a scalar, $k^{-1} = 1/k$.

➤ Multiplication by an inverse has the property:

$$kk^{-1} = 1$$

➤ The question for us is this: can we find a matrix such that

$$BA = AB = I$$

➤ What is $I$?
Identity Matrix

- Suppose $A$ is:
  \[
  \begin{pmatrix}
  1 & 1 \\
  3 & 4 
  \end{pmatrix}
  \]

- Suppose $B$ is:
  \[
  \begin{pmatrix}
  4 & -1 \\
  -3 & 1 
  \end{pmatrix}
  \]

- Multiplying $A \times B$ yields:
  \[
  \begin{pmatrix}
  1 & 0 \\
  0 & 1 
  \end{pmatrix}
  \]

- Call this $I$: an “identity matrix.”

- Here, $B$ is the inverse matrix and $A$ is said to be invertible.

- BTW, verify you could obtain $I$. 
Identity Matrix

- A diagonal matrix is a matrix with all 0s on the off-diagonal.
- An identity matrix is a diagonal matrix with 1s down the main diagonal.
- In matrix algebra, the identity matrix serves the role of a “1.”
Determinants

- Determinants are a scalar: a number associated with a matrix.
- They are useful in many mathematical endeavors.
- wrt matrix algebra, the determinant is useful in informing us as to whether or not a matrix is invertible.
- A matrix with a determinant of 0 is a **singular matrix**.
- Important: a singular matrix is NOT invertible. This will cause us trouble!
- A matrix with a non-zero determinant is a nonsingular matrix.
Determinants

- If all the elements of any row of a matrix are 0, its determinant is 0 and the matrix cannot be inverted.
- If matrix is not of full rank, its determinant is 0.
- Finding a determinant is a bit complicated beyond the $2 \times 2$ setting (not mathematically complicated, just computationally tedious).
- Consider this $2 \times 2$ matrix:

\[
\begin{pmatrix}
  a_{11} & a_{21} \\
  a_{12} & a_{22}
\end{pmatrix}
\]

- In this setting, the determinant is found by cross-multiplication of elements on the main diagonal and subtracting them from cross-product elements on the off-diagonal.
- $||A|| = a_{11}a_{22} - a_{21}a_{12}$
Determinants

- Beyond the $2 \times 2$ setting, finding the determinant is time-consuming (why? LOTS of cross-product manipulations).
- Easy to do computationally. The process is sometimes called matrix expansion.
- The important idea: if a determinant is 0, “stop!”
- The matrix is not invertible.
- As noted before, if a matrix is not of full rank, it will not be invertible.
- \( \therefore \) the rank of a matrix is the dimension (or order) of the largest square submatrix whose determinant is not 0.
- For a matrix to be invertible, an $n \times n$ matrix must be of rank $n$. 
Basic Manipulations: Finding the determinant:

> Define matrix q
>
> > q1<-c(3,6)
> > q2<-c(5,2)
> > qmat<-cbind(q1,q2);qmat
>     q1  q2
> [1,] 3  5
> [2,] 6  2

> #The Determinant
>
> #Finding determinant in 2x2 case
>
> > det<-qmat[1,1]*qmat[2,2]-qmat[2,1]*qmat[1,2]; det
> [1] -24

The determinant is $-24$ (i.e. it’s not $0$; therefore, the matrix must be invertible.)
Basic Manipulations: Finding the inverse:

> qmat
  
  q1  q2
  [1,]  3  5
  [2,]  6  2

> #Solving by hand
> qmat[1,1]/det
[1] -0.125
> qmat[2,2]/det
[1] -0.08333333
> -qmat[1,2]/det
[1] 0.2083333
> -qmat[2,1]/det
[1] 0.25

This would produce the inverse matrix. I could tell R to do this for me:

> #Using R to do it for us
> qinverse<-solve(qmat); qinverse

[,1] [,2]  
q1 -0.0833333 0.2083333  
q2 0.2500000 -0.1250000

These cells are identical to what we computed above.
Recall from before, if $BA = AB = I$, then from above is we multiply matrix $Q$ by $Q^{-1}$, we should obtain an identity matrix. Let’s check:

\[
\begin{array}{cc}
\text{qmat} \ %*% \ qinverse \\
[,1] & [,2] \\
[1,] & 1 & 0 \\
[2,] & 0 & 1 \\
\end{array}
\]

Formally, this is $QQ^{-1}$ (and it’s an identity matrix).
The OLS model can be written as:

\[ y = X\beta + \epsilon. \] (3)

Dimensions:

\[ y : \quad n \times 1 \]
\[ X : \quad n \times k \]
\[ \beta : \quad k \times 1 \]
\[ \epsilon : \quad n \times 1 \] (4)

Note that \( X\beta \) is conformable for multiplication (i.e. we have one parameter for each column vector in \( X \)).
Regression Model

- The matrix $X$ is the *model matrix*.
- Assumptions:
  - Expected Value: $E(\epsilon) = 0$
  - Variance: $\text{var} = E(\epsilon \epsilon^\prime) = \sigma^2 I_n$
  - $\therefore \epsilon \sim N_n(0, \sigma^2 I_n)$
Regression Model

- Properties:
  \[
  \mu = E(y) = E(X\beta + \epsilon) = X\beta + E(\epsilon) = X\beta \\
  \text{var}(y) = E[(y - \mu)(y - \mu)'] = E[(y - X\beta)(y - X\beta)'] \\
  = E(\epsilon\epsilon') = \sigma^2 I_n
  \]

- Estimation of
  \[
  y = X\beta + \epsilon
  \]

- The dimensions of each matrix are as before.
- The residual sums of squares (in scalar form):
  \[
  \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2
  \]
- In matrix form: \( \hat{\epsilon}'\hat{\epsilon} \). (A bit simpler).
Regression Model

- This is a $1 \times 1$ matrix . . . aka a scaler.
- Multiply a column vector, $\hat{\epsilon}'$, by a row vector, $\hat{\epsilon}$, and you will get a scalar.
- Note that dimension of the first matrix is $[1 \times n]$; the second is $[n \times 1]$; the dimension of the product: $[1]$.
- Thus, $\hat{\epsilon}'\hat{\epsilon} = \text{RSS}$
- A residual:
  \[
  \epsilon = y - X\hat{\beta}. \tag{8}
  \]
- The least squares solution is the one that minimizes the sum of the squared errors.
- To solve for $\beta$, you partially differentiate the RSS
Regression Model

Let $\epsilon = [E_1, E_2, \ldots E_n]'$ denote the vector of residuals and $\beta = [B_0, +B_1 + \ldots B_k]'$ give the vector of fitted regression parameters.

Objective:

\[
\sum E_i^2 = \hat{\epsilon}'\hat{\epsilon} = (y - X\hat{\beta})'(y - X\hat{\beta}) \\
= y'y - 2\beta'X'y + \beta'X'X\beta
\] (9)

This result obtains because of properties of the transpose: $(X\beta)' = \beta'X'$; and so $\beta'X'y = y'X\beta$.

Verify that using R.

The goal is to minimize the RSS

In scalar form (which we’ve seen already):
Regression

For two variable regression, differentiate

$$\Sigma \hat{e}_i^2 = \Sigma (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2$$

partially with respect to the three unknown parameter estimates.

This gives

$$\frac{\partial \Sigma \hat{e}_i^2}{\partial \beta_0} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-1),$$

$$\frac{\partial \Sigma \hat{e}_i^2}{\partial \beta_1} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-X_{1i}),$$

$$\frac{\partial \Sigma \hat{e}_i^2}{\partial \beta_2} = 2\Sigma (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-X_{2i}),$$
Regression

When set to 0 and rearranging terms produces the normal equations:

\[
\begin{align*}
Y_i &= \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \\
\Sigma Y_i X_{1i} &= \hat{\beta}_0 \Sigma X_{1i} + \hat{\beta}_1 \Sigma X_{1i}^2 + \hat{\beta}_2 \Sigma X_{1i} X_{2i} \\
\Sigma Y_i X_{2i} &= \hat{\beta}_0 \Sigma X_{2i} + \hat{\beta}_1 \Sigma X_{2i}^2 + \hat{\beta}_2 \Sigma X_{1i} X_{2i}.
\end{align*}
\]

These equations can be rewritten yet again in terms of the parameter estimates:

\[
\begin{align*}
\hat{\beta}_1 &= \frac{\Sigma (X_1 - \bar{X}_1)(Y_i - \bar{Y})(\Sigma (X_2 - \bar{X}))^2 - \Sigma (X_2 - \bar{X}_2)(Y_i - \bar{Y})(\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2))}{\Sigma (X_1 - \bar{X}_1)^2 \Sigma (X_2 - \bar{X}_2)^2 - (\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2} \\
\hat{\beta}_2 &= \frac{\Sigma (X_2 - \bar{X}_2)(Y_i - \bar{Y})(\Sigma (X_1 - \bar{X}))^2 - \Sigma (X_1 - \bar{X}_1)(Y_i - \bar{Y})(\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2))}{\Sigma (X_1 - \bar{X}_1)^2 \Sigma (X_2 - \bar{X}_2)^2 - (\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2} \\
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2
\end{align*}
\]
The objective: find the vector partial derivative wrt to \( \beta \), that is:

\[
\frac{\partial \hat{e}' \hat{e}}{\partial \beta} = 0 - 2X'y + 2X'X\beta
\]  

(10)

Set this to 0:

\[
(X'X)\hat{\beta} = X'y
\]

(11)

This gives the matrix form of the normal equations.

The are \( k + 1 \) normal equations (why?). So long as \( X'X \) is nonsingular (or of full rank), then there is a solution in closed form:

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

(12)
Regression

What is \( X'X \)?

It is a matrix of the raw sums of squares and cross-products for the \( X \) variables (include a constant term, which is a vector of 1s).

Observations:

i. The rank of \( X'X \) equals the rank of \( X \) (why?)

ii. If \( \text{det}(X'X) = 0 \), the inverse will not exist and \( R_k^2 = 1 \).

iii. The largest submatrix that does not have a 0 determinant is \( k + 1 \).

iv. The \( k + 1 \) columns of \( X \) must be linearly independent.
If above conditions are satisfied, then

$$\beta = (X'X)^{-1}X'y$$

(13)

is the least-squares estimator.

The matrix $X'y$ contains the sums of cross products between the regressors and the response variable.

I think it’s useful to reconsider the regression model in scalar form for a minute (see next slide).
Regression

\[
(X'X)\beta := \begin{bmatrix}
\sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{2i} & \cdots & \sum_{i=1}^{n} X_{ki} \\
\sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{1i}^2 & \cdots & \sum_{i=1}^{n} X_{1i}X_{ki} \\
\sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{1i}^2 & \cdots & \sum_{i=1}^{n} X_{2i}X_{ki} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} X_{ki}^2 & \cdots & \sum_{i=1}^{n} X_{ki}^2
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_k
\end{bmatrix}
\]

\[X'y := \begin{bmatrix}
1 & 1 & \cdots & 1 \\
X_{11} & X_{12} & \cdots & X_{1n} \\
X_{21} & X_{22} & \cdots & X_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
X_{k1} & X_{k2} & \cdots & X_{kn}
\end{bmatrix}
\begin{bmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
\vdots \\
Y_k
\end{bmatrix}
\]

\[
(X'X)\hat{\beta} = X'y
\] (14)

\[
\beta = (X'X)^{-1}X'y
\] (15)
Regression

- From before, we created $\mathbf{X}'\mathbf{X}$ from our simulated data:
  \[
  \text{xprime \ % \ * \ % \ xmat}
  \]

- Returning:
  \[
  \begin{bmatrix}
  1 & X_{11} & X_{21} \\
  1 & 5 & 30 & 29.5 \\
  X_{11} & 30 & 220 & 141.8 \\
  X_{21} & 29.5 & 141.8 & 223.57
  \end{bmatrix}
  \]

- Note several things: [1, 1] is simply $n$; [2, 1] = $\sum X_{1i}$; [3, 1] = $\sum X_{2i}$; [2, 2] = $\sum X_{1i}^2$; [3, 3] = $\sum X_{1i}^2$; [3, 2] = $\sum X_{2i}X_{1i}$.

- That is, these are simply functions of the data, nothing more.

- On your own, extract $\mathbf{X}'\mathbf{y}$

- Solving for $\beta$
Regression Matrix Manipulation

- From the normal equations, it is clear we have known quantities in $(X'X)$ and $X'y$.
- We of course, do not know what $\hat{\beta}$ are.
- We solve for them.
- Let’s take advantage of the inversion result: recall $QQ^{-1} = I$ (if the inverse exists; it may not).
- So, consider this. Let us premultiply equation (11) by $(X'X)^{-1}$:

\[
(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y \\
I\hat{\beta} = (X'X)^{-1}X'y
\]  
(16)

- The existence of an inverse is important; without it, we cannot obtain the OLS estimator:

\[
\hat{\beta} = (X'X)^{-1}X'y
\]  
(17)
This is the OLS estimator: the regression parameters are a function of the data.

Continuing with our example, let’s create $(X'X)^{-1}$:
\[
\text{solve}(xprime_x)
\]

Returning:
\[
\begin{bmatrix}
x0 & x1 & x2 \\
x0 & 7.8403149 & -0.68054357 & -0.6028904 \\
x1 & -0.6805436 & 0.06676014 & 0.0474547 \\
x2 & -0.6028904 & 0.04745470 & 0.0539258
\end{bmatrix}
\]
Regression Matrix Manipulation

- Verify that \((X'X)^{-1}(X'X)\) is an identity matrix:
  
  ```r
  solve(xprime\_x)%*%xprimex
  ```

- Returning:

  \[
  \begin{bmatrix}
  x0 & x1 & x2 \\
  x0 & 1.000000e+00 & -8.701373e-15 & -4.009293e-14 \\
  x1 & 1.783513e-16 & 1.000000e+00 & 5.793976e-16 \\
  x2 & -2.123952e-16 & -2.896988e-16 & 1.000000e+00
  \end{bmatrix}
  \]

- This is an identity matrix \(I\).
We can solve for $\hat{\beta}$ by first creating $X'y$:

\[ x_{\text{prime}} \times y \]

Returning:

\[
\begin{bmatrix}
  x_0 & 299.0 \\
  x_1 & 2204.0 \\
  x_2 & 1387.6
\end{bmatrix}
\]

The least squares estimates are obtained by

$\hat{\beta} = (X'X)^{-1}X'y$:

\[ xx_{\text{inverse}} \times x_{\text{prime}_y} \]

Returning:

\[
\begin{bmatrix}
  [, 1] \\
  x_0 & 7.765380 \\
  x_1 & 9.504961 \\
  x_2 & -0.846635
\end{bmatrix}
\]
Are we correct?
\begin{tiny} \begin{verbatim}
> model1<-lm(y~x1+ x2); summary(model1)

Call:
  lm(formula = y ~ x1 + x2)

Residuals:
1           2           3           4           5
  11.40513   -24.50416   13.99390   -0.09584   -0.79903

Coefficients:
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  7.7654     60.2828 0.1290     0.909
   x1        9.5050      5.5627 1.7089     0.230
   x2       -0.8466      4.9995 -0.1692     0.881

Residual standard error: 21.53 on 2 degrees of freedom
Multiple R-Squared: 0.8197, Adjusted R-squared: 0.6395
F-statistic: 4.548 on 2 and 2 DF,  p-value: 0.1803
\end{verbatim} \end{tiny}
Theory

- \( \beta \) is a linear estimator:

\[
\beta = (X'X)^{-1}X'y = My
\]  

(using Fox’s notation).

- Expected value result:

\[
E(\hat{\beta}) = E(My) = ME(y) = (X'X)^{-1}X'(X\beta) = \beta
\]  

- Verify this result on your own.

- Another way to see this:

\[
\hat{\beta} = (X'X)^{-1}X'(X\beta + \epsilon)
\]

\[
= \beta + (X'X)^{-1}X'\epsilon
\]

- Since \( E(X'\epsilon) = 0 \), then \( \hat{\beta} \) is a linear unbiased estimator of \( \beta \).
Theory

▶ Variance result:

$$\text{var}(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

$$= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}]$$

$$= (X'X)^{-1}X'E[\epsilon\epsilon']X(X'X)^{-1}$$

$$= (X'X)^{-1}X'(\sigma^2I)X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1} \quad (21)$$

▶ Again, verify this on your own (especially the last step).

▶ It is clear that the variances and covariances of the regression parameters depend only on $X$ and the residual variance (note that $\sigma^2$ is estimated by $MSE$).
None of the above required distributional assumptions.

If $y \sim N$, however, then

$$
\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}]
$$

(22)

Now consider the Gauss-Markov Theorem again.

Review the assumptions of the model (which are neatly summarized above).
Gauss-Markov Theorem

Let \( \tilde{\beta} = Cy \).

Let \( C \) be:

\[
(X'X)^{-1}X' + D
\]

where \( D \) is \( k \times n \) (nonzero).

Then

\[
E[Cy] = E[(X'X)^{-1}X' + D)(X\beta + \epsilon)]
\]

\[
= ((X'X)^{-1}X' + D)X\beta + ((X'X)^{-1}X' + D)E(\epsilon)
\]

\[
= (X'X)^{-1}X'X\beta + DX\beta
\]

\[
= (I + DX)\beta
\]

\[
\therefore \tilde{\beta} \text{ is unbiased } \iff DX = 0.
\]
Variance result:

\[
V(\tilde{\beta}) = V(Cy) = CV(Y)C' = \sigma^2 CC' \\
= \sigma^2(X'X)^{-1} + \sigma^2 DD'
\]  

(24)

Some algebraic steps are omitted here.

The result implies the following: if $DD'$ is a positive semidefinite matrix, then the variance of $\tilde{\beta}$ must exceed the variance of $\hat{\beta}$.

The least-squares estimator is the minimum variance linear unbiased estimator for $\beta$.

This is the Gauss-Markov Theorem.
Inference and Interpretation

- Inference depends on our distributional assumptions which we have taken as the normal.
- Because $\sigma^2$ is estimated by the MSE, all the elements in $\hat{\beta}$ follow the $t$-distribution with $n - k - 1$ degrees-of-freedom.
- The square root of the variance shown earlier gives us a standard error.
- Standard $t$-test emerges:

$$t = \frac{\hat{\beta}_k - \beta^*_k}{se(\hat{\beta}_k)}$$

(25)

where the $\beta_k$ is an element of $\hat{\beta}$. 
More formally, the estimated variance is:

\[ V(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1} = \frac{\epsilon \epsilon'}{n-k-1} (X'X)^{-1} \quad (26) \]

Only difference from before is we substitute in the MSE.

The standard error is square root of the kth diagonal of \( V(\hat{\beta}) \).

Continue with illustration using simulated data.
The “hat” matrix is a useful matrix as we will see.

Predicted values:

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

That is: \( X \) times \( \hat{\beta} \).

Let \( H = X(X'X)^{-1}X' \); we call \( H \) the “hat” matrix.

It puts the “hat” on \( y \).

Thus, \( \hat{y} = Hy \).

Illustration using simulated data.
Compute the hat matrix:

\[ \text{hat} \leftarrow \text{xmat} \%\% \left( \text{xxinverse} \right) \%\% \text{xprime} \]

And now, let's get the fitted values \( \hat{y} = Hy \):

\> \text{yhat} \leftarrow \text{hat} \%\% \text{y}; \text{yhat}

\[,1\]
[1,] 20.59487
[2,] 38.50416
[3,] 57.00610
[4,] 81.09584
[5,] 101.79903

Now the residuals, \( y - \hat{y} \):

\text{residual} \leftarrow y - \text{yhat}; \text{residual}

\[,1\]
[1,] 11.40513374
[2,] -24.50416307
[3,] 13.99389560
[4,] -0.09583693
[5,] -0.79902934
Since $\sum e_i^2 = RSS$, we can compute this quantity using matrix operations:

```r
> res<-cbind(residual)
> eprime<-t(res); eprime
[1,] 11.40513 -24.50416 13.99390 -0.09583693 -0.7990293

> eprimee<-eprime %*% res; eprimee
[,1]
[1,] 927.0078
```

... or the old-fashioned way:

```r
> SSError=sum(residual^2); SSError
[1] 927.0078
```

Other quantities? How about the $R^2$:

```r
> meany<-cbind(rep(c(mean(y)), 5))
> explained<-yhat-meany; explained
[,1]
[1,] -39.205134
[2,] -21.295837
[3,] -2.793896
[4,]  21.295837
[5,]  41.999029

> SSRegress<-sum(explained^2)
> r2=SSRegress/(eprimee+SSRegress); r2
[,1]
[1,] 0.8197465
```
This is what we want. So now that we have the variance components, we can rule the world.

Mean Square Error:
> MSE<-eprimee/(2); MSE
[1,] 463.5039

Root Mean Square Error (s.e. of the estimate):
MSE^(.5)
[1,] 21.52914

Note that the variance-covariance matrix of $\beta = MSE(X'X)^{-1}$; $\therefore$

> varcov<-463.5039 * xxinverse; varcov
     x0      x1      x2
x0 3634.0165 -315.43460 -279.44206
x1 -315.4346  30.94358  21.99544
x2 -279.4421  21.99544  24.99482
>
> #Now we can obtain the standard errors:
> s.e._int<-sqrt(varcov[1,1])
> s.e._b1<-sqrt(varcov[2,2])
> s.e._b2<-sqrt(varcov[3,3])
> s.e<-cbind(s.e._int, s.e._b1, s.e._b2); s.e
     s.e._int  s.e._b1  s.e._b2
[1,]  60.2828  5.562696  4.999482
Recall:

```r
> model1<-lm(y~x1+ x2); summary(model1)

Call:
  lm(formula = y ~ x1 + x2)

Residuals:
     1       2       3       4       5
  11.40513 -24.50416  13.99390  -0.09584  -0.79903

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.7654      60.2828   0.129    0.909
x1            9.5050      5.5627   1.709    0.230
x2           -0.8466      4.9995  -0.169    0.881

Residual standard error: 21.53 on 2 degrees of freedom
Multiple R-Squared: 0.8197,    Adjusted R-squared: 0.6395
F-statistic: 4.548 on 2 and 2 DF,  p-value: 0.1803
```
We’ve basically just replicated everything here (or could replicate everything here).
The hat matrix: \( H = X(X'X)^{-1}X' \)

- Has some useful properties.
- The elements in the hat matrix are, in a sense, weights, that when applied to \( Y \), give us the predicted value.
- If we were to multiply the elements in the hat matrix for the \( i \)th observation, we would retrieve the predicted value for this observation; that is,

\[
\hat{Y}_i = h_{11} Y_1 + h_{12} Y_2 + h_{13} Y_3 + \ldots + h_{1n} Y_n.
\]
These weights give us the contribution, then, of observation $Y_i$ on the fitted values, $\hat{Y}_j$.

Since the fitted values will be sensitive to the values of other observations (why might this be the case?), inspection of the hat matrix can give us an indication of which, if any, of the observations are exerting influence on the regression function.

If $h_{11}$ is very large, this could imply that this observation is exerting a lot of influence on the regression model.

Note in matrix terms, the residuals can be written as

$$e = (I - H)Y,$$
Regression

- Subtraction of $H$ from an identity matrix makes sense because if $h_{11}$ gives us the unique weight associated with the the first observation on the overall predicted value, then $1 - h_{11}$ will give us the “leftovers” so to speak, for this observation.

- Specifically, the residual for the $i$th observation is given by

$$\hat{e}_i = 1 - h_{11}Y_1 - h_{12}Y_2 - h_{13}Y_3 - \ldots - h_{1n}Y_n;$$

- Note that when subtracting from an identity matrix, the main diagonal of $H$ will be $1 - h_{ii}$ and the off diagonals will be $0 - h_{ij}$ (thus resulting in the reversal of signs).

- Because the main diagonal element in the hat matrix gives us a measure of the weight an observation is putting on the regression line (or plane), it is natural to analyze these elements.
Regression

- The diagonal element is the hat matrix is given by
  \[ h_{ii} = X'_i (X'X)^{-1} X_i, \]
  where \( X_i \) corresponds to the \( X_i \) column vector in \( X \) and \( X'_i \) is its transpose (which is equivalent to the \( i \)th row of data pertaining to the \( i \)th case).

- The diagonal elements of \( h_{ii} \) have some useful properties. First, their values always are between 0 and 1 and they sum to \( k + 1 \); that is,
  \[ 0 \leq h_{ii} \leq 1 \quad \sum_{i=1}^{n} h_{ii} = k + 1. \]

- The diagonal elements gives us information about leverage (in terms of the independent variables) of the \( i \)th case.
Regression

- The hat diagonal is a measure of distance between $X$ values for the $i$th case and the means of the $X$ values for all the cases.
- A large leverage value of $h_{ii}$ indicates that the $i$th case is distant from the center of all $X$ observations.
- In general, large values of $h_{ii}$ indicate the observation is exerting a lot of leverage on the regression function.
- $\hat{Y}_i$ is a linear combination of the observed $Y$ values and since $h_{ii}$ is the weight associated with the $i$th observation, the larger $h_{ii}$ is, the more important the $i$th observation is in affecting the fitted value of $\hat{Y}_i$. 
Another important result to note is that the variance of the residuals can be written as

$$\text{var}(e) = \sigma^2(I - H),$$

The variance of the \(i\)th residual, that is \(e_i\) is

$$\text{var}(e_i) = \sigma^2(1 - h_{ii}).$$

Since \(\sigma^2\) is usually unknown, we substitute the \(MSE\) into the above giving us

$$\text{var}(e_i) = MSE(1 - h_{ii}).$$
Regression

- The important thing to note here is that the larger $h_{ii}$ is, the smaller will be its variance.
- Thus, the larger $h_{ii}$ is, the closer the fitted value of $\hat{Y}_i$ will tend to be to the observed value of $Y_i$.
- In the extreme case when $h_{ii} = 1$, the variance is 0 and so the regression line is forced through the data point.
- This last point is important: *high leverage observations will usually be associated with small residuals.*
- This is in contrast to the view that “outlying” observations are the observations to worry about. Outliers are usually characterized as having large residuals; they may or may not actually influence the regression function in any particular way.
The hat diagonal is a useful tool for diagnostics. Since the mean leverage is

\[ \bar{h} = \frac{k + 1}{n}, \]

then high-leverage observations will be (by this rule) \(2\bar{h}\). Extract \(H\) and inspect the diagonal elements.