Lecture 05
Geometry of Least Squares

16 September 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Goals for today

1. Geometry of least squares
2. Projection matrix P and annihilator matrix M
3. Application to

Geometry of Least Squares

Last time, we established that the least squares solution to the model:

$$y = X\beta + \epsilon$$

Yields the solution:

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

As long as the matrix $X^t X$ is invertable.

Define the column space of the matrix X as:

$$\mathcal{R}(X) = \{\theta : \theta = X\beta,\ \beta \in \mathbb{R}^p\} \subset \mathbb{R}^n$$

This is the space spanned by the p columns on X sitting in n-dimensional space.

Define the column space of the matrix X as:

$$\mathcal{R}(X) = \{\theta : \theta = X\beta,\ \beta \in \mathbb{R}^p\} \subset \mathbb{R}^n$$

This is the space spanned by the p columns on X sitting in n-dimensional space.

Notice that the least squares problem can be re-written as:

$$\widehat{\theta} = \arg\min_{\theta} \left\{ ||y - \theta||_2^2, \quad \text{s.t} \quad \theta = X\beta \right\}$$

Where $\widehat{\beta} = X\widehat{\theta}$.

Theorem 3.2 (p.g. 37, Rao & Toutenburg) The minimum, $\widehat{\theta}$ is attained when $(y - \widehat{\theta}) \perp \mathcal{R}(X)$. In other words, $(y - \widehat{\theta})$ is perpendicular to all vectors in $\mathcal{R}$.

Proof: Pick a $\widehat{\theta}$ in $\mathcal{R}$ such that $(y - \widehat{\theta}) \perp \mathcal{R}(X)$.

Proof: Pick a $\widehat{\theta}$ in $\mathcal{R}$ such that $(y - \widehat{\theta}) \perp \mathcal{R}(X)$. This implies that $X^t(y - \widehat{\theta}) = 0$. Then for all $\theta \in \mathcal{R}$:

$$||y - \theta||^2 \quad = \quad (y - \widehat{\theta} + \widehat{\theta} - \theta)^t(y - \widehat{\theta} + \widehat{\theta} - \theta)$$

Proof: Pick a $\widehat{\theta}$ in $\mathcal{R}$ such that $(y - \widehat{\theta}) \perp \mathcal{R}(X)$. This implies that $X^t(y - \widehat{\theta}) = 0$. Then for all $\theta \in \mathcal{R}$:

$$\begin{aligned}
||y - \theta||^2 &= (y - \widehat{\theta} + \widehat{\theta} - \theta)^t(y - \widehat{\theta} + \widehat{\theta} - \theta) \\
&= (y - \widehat{\theta})^t(y - \widehat{\theta}) + (\widehat{\theta} - \theta)^t(\widehat{\theta} - \theta) + 2(y - \widehat{\theta})^t(\widehat{\theta} - \theta)
\end{aligned}$$

Proof: Pick a $\widehat{\theta}$ in $\mathcal{R}$ such that $(y-\widehat{\theta}) \perp \mathcal{R}(X)$. This implies that $X^t(y-\widehat{\theta}) = 0$. Then for all $\theta \in \mathcal{R}$:

$$\begin{aligned} ||y-\theta||^2 &= (y-\widehat{\theta}+\widehat{\theta}-\theta)^t(y-\widehat{\theta}+\widehat{\theta}-\theta) \\ &= (y-\widehat{\theta})^t(y-\widehat{\theta}) + (\widehat{\theta}-\theta)^t(\widehat{\theta}-\theta) + 2(y-\widehat{\theta})^t(\widehat{\theta}-\theta) \\ &= (y-\widehat{\theta})^t(y-\widehat{\theta}) + (\widehat{\theta}-\theta)^t(\widehat{\theta}-\theta) \end{aligned}$$

Proof: Pick a $\widehat{\theta}$ in $\mathcal{R}$ such that $(y - \widehat{\theta}) \perp \mathcal{R}(X)$. This implies that $X^t(y - \widehat{\theta}) = 0$. Then for all $\theta \in \mathcal{R}$:

$$
\begin{aligned}
||y - \theta||^2 &= (y - \widehat{\theta} + \widehat{\theta} - \theta)^t (y - \widehat{\theta} + \widehat{\theta} - \theta) \\
&= (y - \widehat{\theta})^t (y - \widehat{\theta}) + (\widehat{\theta} - \theta)^t (\widehat{\theta} - \theta) + 2(y - \widehat{\theta})^t (\widehat{\theta} - \theta) \\
&= (y - \widehat{\theta})^t (y - \widehat{\theta}) + (\widehat{\theta} - \theta)^t (\widehat{\theta} - \theta) \\
&= ||y - \widehat{\theta}||_2^2 + ||\widehat{\theta} - \theta||_2^2
\end{aligned}
$$

Proof: Pick a $\widehat{\theta}$ in $\mathcal{R}$ such that $(y-\widehat{\theta}) \perp \mathcal{R}(X)$. This implies that $X^t(y-\widehat{\theta}) = 0$. Then for all $\theta \in \mathcal{R}$:

$$\begin{aligned}
||y-\theta||^2 &= (y-\widehat{\theta}+\widehat{\theta}-\theta)^t(y-\widehat{\theta}+\widehat{\theta}-\theta) \\
&= (y-\widehat{\theta})^t(y-\widehat{\theta}) + (\widehat{\theta}-\theta)^t(\widehat{\theta}-\theta) + 2(y-\widehat{\theta})^t(\widehat{\theta}-\theta) \\
&= (y-\widehat{\theta})^t(y-\widehat{\theta}) + (\widehat{\theta}-\theta)^t(\widehat{\theta}-\theta) \\
&= ||y-\widehat{\theta}||_2^2 + ||\widehat{\theta}-\theta||_2^2 \\
&\geq ||y-\widehat{\theta}||_2^2
\end{aligned}$$

Proof: Pick a $\widehat{\theta}$ in $\mathcal{R}$ such that $(y - \widehat{\theta}) \perp \mathcal{R}(X)$. This implies that $X^t(y - \widehat{\theta}) = 0$. Then for all $\theta \in \mathcal{R}$:

$$
\begin{aligned}
||y - \theta||^2 &= (y - \widehat{\theta} + \widehat{\theta} - \theta)^t (y - \widehat{\theta} + \widehat{\theta} - \theta) \\
&= (y - \widehat{\theta})^t (y - \widehat{\theta}) + (\widehat{\theta} - \theta)^t (\widehat{\theta} - \theta) + 2(y - \widehat{\theta})^t (\widehat{\theta} - \theta) \\
&= (y - \widehat{\theta})^t (y - \widehat{\theta}) + (\widehat{\theta} - \theta)^t (\widehat{\theta} - \theta) \\
&= ||y - \widehat{\theta}||_2^2 + ||\widehat{\theta} - \theta||_2^2 \\
&\geq ||y - \widehat{\theta}||_2^2
\end{aligned}
$$

So, if such a $\widehat{\theta}$ exists it attains the minimum. To see that it does, write $\widehat{\theta} = X\widehat{\beta}$.

Proof: Pick a $\widehat{\theta}$ in $\mathcal{R}$ such that $(y - \widehat{\theta}) \perp \mathcal{R}(X)$. This implies that $X^t(y - \widehat{\theta}) = 0$. Then for all $\theta \in \mathcal{R}$:

$$\begin{array}{rcl} ||y - \theta||^2 & = & (y - \widehat{\theta} + \widehat{\theta} - \theta)^t(y - \widehat{\theta} + \widehat{\theta} - \theta) \\ & = & (y - \widehat{\theta})^t(y - \widehat{\theta}) + (\widehat{\theta} - \theta)^t(\widehat{\theta} - \theta) + 2(y - \widehat{\theta})^t(\widehat{\theta} - \theta) \\ & = & (y - \widehat{\theta})^t(y - \widehat{\theta}) + (\widehat{\theta} - \theta)^t(\widehat{\theta} - \theta) \\ & = & ||y - \widehat{\theta}||_2^2 + ||\widehat{\theta} - \theta||_2^2 \\ & \geq & ||y - \widehat{\theta}||_2^2 \end{array}$$

So, if such a $\widehat{\theta}$ exists it attains the minimum. To see that it does, write $\widehat{\theta} = X\widehat{\beta}$. Then:

$$\begin{aligned} X^t(y - \widehat{\theta}) \quad &= X^t(y - X\widehat{\beta}) \\ &= X^t y - X^t X\widehat{\beta} \end{aligned}$$

To see that such a $\widehat{\theta}$ does exist, write $\widehat{\theta} = X\widehat{\beta}$.

To see that such a $\widehat{\theta}$ does exist, write $\widehat{\theta} = X\widehat{\beta}$. Then:

$$X^t(y - \widehat{\theta}) = X^t(y - X\widehat{\beta})$$

To see that such a $\widehat{\theta}$ does exist, write $\widehat{\theta} = X\widehat{\beta}$. Then:

$$\begin{aligned} X^t(y - \widehat{\theta}) &= X^t(y - X\widehat{\beta}) \\ &= X^t y - X^t X\widehat{\beta} \end{aligned}$$

To see that such a $\widehat{\theta}$ does exist, write $\widehat{\theta} = X\widehat{\beta}$. Then:

$$
\begin{aligned}
X^t(y - \widehat{\theta}) &= X^t(y - X\widehat{\beta}) \\
&= X^t y - X^t X\widehat{\beta} \\
&= X^t y - X^t X(X^t X)^{-1} X^t y
\end{aligned}
$$

To see that such a $\widehat{\theta}$ does exist, write $\widehat{\theta} = X\widehat{\beta}$. Then:

$$\begin{aligned}
X^t(y - \widehat{\theta}) &= X^t(y - X\widehat{\beta}) \\
&= X^t y - X^t X\widehat{\beta} \\
&= X^t y - X^t X(X^t X)^{-1} X^t y \\
&= X^t y - X^t y \\
&= 0
\end{aligned}$$

And therefore our proposed $\widehat{\theta} \in \mathcal{R}(X)$.

From this geometric interpretation of the least squares estimator, we introduce an important matrix P_X called the *projection matrix*.

$$P_X = X(X^tX)^{-1}X^t$$

I'll often drop the subscript as it should be understood that the projection is on the data matrix X.

Notice that $PX = X$:

$$
\begin{aligned}
PX &= X(X^t X)^{-1} X^t X \\
&= X
\end{aligned}
$$

Notice that $PX = X$:

$$
\begin{aligned}
PX &= X(X^tX)^{-1}X^tX \\
&= X
\end{aligned}
$$

And Py gives the fitted values $\widehat{y}$:

$$
\begin{aligned}
Py &= X(X^tX)^{-1}X^tXy \\
&= X\widehat{\beta} \\
&= \widehat{\theta} \\
&= \widehat{y}
\end{aligned}
$$

Do you see why the projection matrix is called the projection matrix?

Notice that $PX = X$:

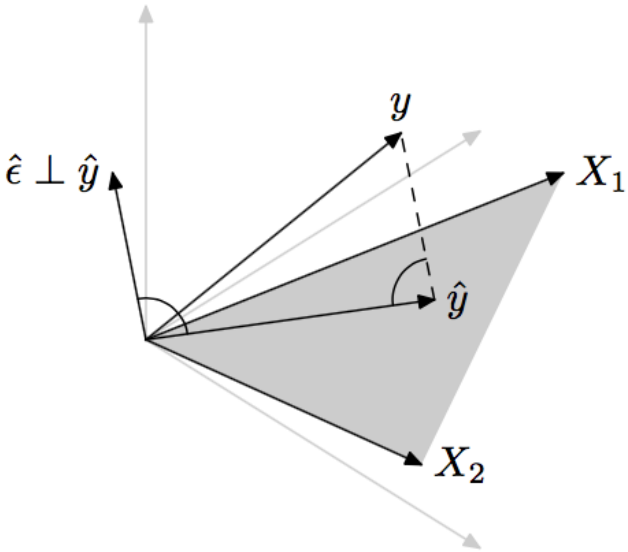
$$\begin{aligned} PX &= X(X^t X)^{-1} X^t X \\ &= X \end{aligned}$$

And Py gives the fitted values $\widehat{y}$:

$$\begin{aligned} Py &= X(X^t X)^{-1} X^t X y \\ &= X\widehat{\beta} \\ &= \widehat{\theta} \\ &= \widehat{y} \end{aligned}$$

Do you see why the projection matrix is called the projection matrix?

The projection matrix is sometimes called the *hat matrix*. Any thoughts as to why?



y
$\hat{\epsilon} \perp \hat{y}$
X_1
$\hat{y}$
X_2

A closely related matrix to P is the *annihilator matrix* M:

$$M = I_n - P$$

A closely related matrix to P is the *annihilator matrix* M:

$$M = I_n - P$$

It gets its name because $MX = 0$.

The matrix $P = X(X^tX)^{-1}X^t$ is clearly symettric. It is also idempotent:

$$\begin{aligned} P^2 &= X(X^tX)^{-1}X^tX(X^tX)^{-1}X^t \\ &= X(X^tX)^{-1}(X^tX)(X^tX)^{-1}X^t \\ &= X(X^tX)^{-1}X^t \\ &= P \end{aligned}$$

M is also symettric

$$\begin{aligned} M^t &= (I_n - P)^t \\ &= (I_n - P^t) \\ &= M \end{aligned}$$

M is also symettric

$$\begin{aligned} M^t &= (I_n - P)^t \\ &= (I_n - P^t) \\ &= M \end{aligned}$$

And idempotent:

$$\begin{aligned} M^2 &= (I_n - P)^2 \\ &= (I_n - P)(I_n - P) \end{aligned}$$

M is also symettric

$$\begin{aligned} M^t &= (I_n - P)^t \\ &= (I_n - P^t) \\ &= M \end{aligned}$$

And idempotent:

$$\begin{aligned} M^2 &= (I_n - P)^2 \\ &= (I_n - P)(I_n - P) \\ &= I_n - 2 * P + P^2 \end{aligned}$$

M is also symettric

$$\begin{aligned} M^t &= (I_n - P)^t \\ &= (I_n - P^t) \\ &= M \end{aligned}$$

And idempotent:

$$\begin{aligned} M^2 &= (I_n - P)^2 \\ &= (I_n - P)(I_n - P) \\ &= I_n - 2 * P + P^2 \\ &= I_n - 2 * P + P \end{aligned}$$

M is also symettric

$$\begin{aligned} M^t &= (I_n - P)^t \\ &= (I_n - P^t) \\ &= M \end{aligned}$$

And idempotent:

$$\begin{aligned} M^2 &= (I_n - P)^2 \\ &= (I_n - P)(I_n - P) \\ &= I_n - 2 * P + P^2 \\ &= I_n - 2 * P + P \\ &= I_n - P \\ &= M \end{aligned}$$

M is also symettric

$$
\begin{aligned}
M^t &= (I_n - P)^t \\
&= (I_n - P^t) \\
&= M
\end{aligned}
$$

And idempotent:

$$
\begin{aligned}
M^2 &= (I_n - P)^2 \\
&= (I_n - P)(I_n - P) \\
&= I_n - 2 * P + P^2 \\
&= I_n - 2 * P + P \\
&= I_n - P \\
&= M
\end{aligned}
$$

These properties both make sense given the geometric interpretation of P and M as projections; into the column space of X and the compliment of the columns space of X.

These properties are quite useful. Notice how we can easily rewrite the following for the residual vector $r = y - X\widehat{\beta}$:

$$r = y - X\widehat{\beta}$$

These properties are quite useful. Notice how we can easily rewrite the following for the residual vector $r = y - X\widehat{\beta}$:

$$
\begin{aligned}
r &= y - X\widehat{\beta} \\
&= y - Py
\end{aligned}
$$

These properties are quite useful. Notice how we can easily rewrite the following for the residual vector $r = y - X\widehat{\beta}$:

$$\begin{aligned} r &= y - X\widehat{\beta} \\ &= y - Py \\ &= (I_n - P)y \end{aligned}$$

These properties are quite useful. Notice how we can easily rewrite the following for the residual vector $r = y - X\widehat{\beta}$:

$$\begin{aligned} r &= y - X\widehat{\beta} \\ &= y - Py \\ &= (I_n - P)y \\ &= My \end{aligned}$$

These properties are quite useful. Notice how we can easily rewrite the following for the residual vector $r = y - X\widehat{\beta}$:

$$
\begin{aligned}
r &= y - X\widehat{\beta} \\
&= y - Py \\
&= (I_n - P)y \\
&= My \\
&= M(X\beta - \epsilon)
\end{aligned}
$$

These properties are quite useful. Notice how we can easily rewrite the following for the residual vector $r = y - X\widehat{\beta}$:

$$
\begin{aligned}
r &= y - X\widehat{\beta} \\
&= y - Py \\
&= (I_n - P)y \\
&= My \\
&= M(X\beta - \epsilon) \\
&= M\epsilon
\end{aligned}
$$

These properties are quite useful. Notice how we can easily rewrite the following for the residual vector $r = y - X\widehat{\beta}$:

$$\begin{aligned} r &= y - X\widehat{\beta} \\ &= y - Py \\ &= (I_n - P)y \\ &= My \\ &= M(X\beta - \epsilon) \\ &= M\epsilon \end{aligned}$$

The matricies P and M not only help make the derivation easier, they also give geometric insight into what we are doing.

One particularly useful formula will be writing the squared residuals as:

$$\begin{aligned} ||r||_2^2 &= ||M\epsilon||_2^2 \\ &= \epsilon^t M^t M\epsilon \\ &= \epsilon^t M\epsilon \end{aligned}$$

One particularly useful formula will be writing the squared residuals as:

$$\begin{aligned} ||r||_2^2 &= ||M\epsilon||_2^2 \\ &= \epsilon^t M^t M \epsilon \\ &= \epsilon^t M \epsilon \end{aligned}$$

So the matrix M translates the sum of squared residuals into the sum of the square errors, which are estimated by the residuals.

Applications