Topics in Computational Economics

Lecture 7

John Stachurski

NYU 2016

Today's Lecture

- Orthogonal Projections
- Orthonormal Bases
- Overdetermined Systems
- Least Squares and Regression

Comments on Homework Set 4

Overall, many improvements

Further comments:

Write readable code, not clever code

> *Debugging is twice as hard as writing the code in the first place. Therefore, if you write the code as cleverly as possible, you are, by definition, not smart enough to debug it.*
>
> – *Brian W. Kernighan*

Be clever with your algorithms, not your code

Write readable code, not clever code

> *There is no faster way for a trading firm to destroy itself than to deploy a piece of trading software that makes a bad decision over and over in a tight loop. Part of Jane Street's reaction to these technological risks was to put a very strong focus on building software that was easily understood—software that was readable.*
>
> *– Yaron Minsky, Jane Street*

Always avoid numbers in the middle of your algos

```
for n in range(len(N_int)): # For each sample size
    alpha_estimates = []

    for i in range(4000): # Run 1000 estimations
        xlag, x = ar1_gen(N_int[n], a, 1, 1)
```

But the good use of PEP8 makes me happy...

- https://www.python.org/dev/peps/pep-0008/

Always break your code up into smallish functions

- Each function has one simple, well defined task
- Add some documentation to say what each function does
- Test each function before you move on

Use descriptive function and variable names

- `simulation_size` is often better than `N`

Read other people's solutions:

- See http://nbviewer.jupyter.org/github/jstac/quantecon_nyu_2016_homework/tree/master/hw_set_4/

Orthogonal Projections

Many problems in linear algebra and analysis are related to orthogonal projection

- Least squares and linear regression
- Conditional expectation
- Gram–Schmidt orthogonalization
- QR decomposition
- Orthogonal polynomials
- Wavelets as basis functions
- Galerkin projection

Let's start with the basic concepts in $\mathbb{R}^n$

If

$$\mathbf{x}, \mathbf{z} \in \mathbb{R}^n \quad \text{and} \quad \langle \mathbf{x}, \mathbf{z} \rangle = 0$$

then $\mathbf{x}$ and $\mathbf{z}$ said to be **orthogonal** and we write $\mathbf{x} \perp \mathbf{z}$

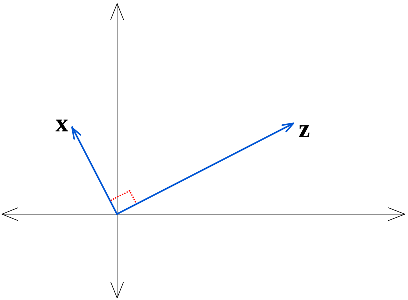
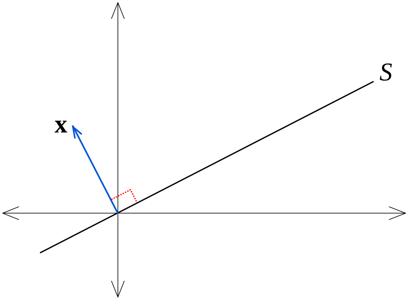


Figure: $\mathbf{x} \perp \mathbf{z}$

Let S be a linear subspace of $\mathbb{R}^n$

We call $\mathbf{x} \in \mathbb{R}^n$ **orthogonal to S** if $\mathbf{x} \perp \mathbf{z}$ for all $\mathbf{z} \in S$

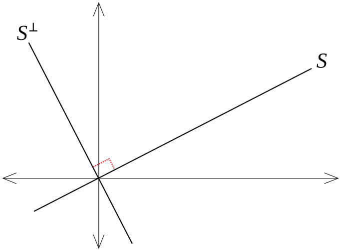
Write $\mathbf{x} \perp S$



Orthogonal Complement

The **orthogonal complement** of S is

$$S^{\perp} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \perp S\}$$



Fact. $S^\perp$ is always a linear subspace of $\mathbb{R}^n$

Proof: Fix $\mathbf{x}, \mathbf{y} \in S^\perp$ and $\alpha, \beta \in \mathbb{R}$

We claim that $\alpha\mathbf{x} + \beta\mathbf{y} \in S^\perp$

True, because if $\mathbf{z} \in S$, then

$$\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle = \alpha \times 0 + \beta \times 0 = 0$$

$$\therefore \quad \alpha\mathbf{x} + \beta\mathbf{y} \in S^\perp$$

Fact. For any nonempty $S \subset \mathbb{R}^n$, we have $S \cap S^\perp = \{\mathbf{0}\}$

A set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subset \mathbb{R}^n$ is called an **orthogonal set** if

$$\mathbf{x}_i \perp \mathbf{x}_j \quad \text{whenever} \quad i \neq j$$

Pythagorean Law. If $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is an orthogonal set, then

$$\|\mathbf{x}_1 + \cdots + \mathbf{x}_k\|^2 = \|\mathbf{x}_1\|^2 + \cdots + \|\mathbf{x}_k\|^2$$

Proof for the case of $k = 2$:

$$\begin{aligned}
\|\mathbf{x}_1 + \mathbf{x}_2\|^2 &= \langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_1 + \mathbf{x}_2 \rangle \\
&= \langle \mathbf{x}_1, \mathbf{x}_1 \rangle + 2 \langle \mathbf{x}_2, \mathbf{x}_1 \rangle + \langle \mathbf{x}_2, \mathbf{x}_2 \rangle \\
&= \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2
\end{aligned}$$

Linear Independence vs Orthogonality

Fact. If

- $X \subset \mathbb{R}^n$ is an orthogonal set and
- $\mathbf{0} \notin X$,

then X is linearly independent

Ex. Prove it

The converse is not true

But a partial converse holds — See Gram–Schmidt theorem below

Orthogonal Projections

Problem:

Given $\mathbf{y} \in \mathbb{R}^n$ and subspace S, find closest element of S to $\mathbf{y}$

Formally: Solve for

$$\hat{\mathbf{y}} := \operatorname*{argmin}_{\mathbf{z} \in S} \|\mathbf{y} - \mathbf{z}\|$$

Let $\mathbf{y} \in \mathbb{R}^n$ and let S be a linear subspace of $\mathbb{R}^n$

Theorem. (OPT I) There exists a unique solution to

$$\hat{\mathbf{y}} := \operatorname*{argmin}_{\mathbf{z} \in S} \|\mathbf{y} - \mathbf{z}\|$$

Moreover, the solution $\hat{\mathbf{y}}$ is the unique vector in $\mathbb{R}^n$ such that

1. $\hat{\mathbf{y}} \in S$
2. $\mathbf{y} - \hat{\mathbf{y}} \perp S$

The vector $\hat{\mathbf{y}}$ is called the **orthogonal projection of $\mathbf{y}$ onto S**

Proof, sufficiency: Let $\mathbf{y} \in \mathbb{R}^n$ and let S be a linear subspace of $\mathbb{R}^n$

Let $\hat{\mathbf{y}}$ be a vector in $\mathbb{R}^n$ such that

1. $\hat{\mathbf{y}} \in S$
2. $\mathbf{y} - \hat{\mathbf{y}} \perp S$

Let $\mathbf{z}$ be any other point in S

We have

$$\|\mathbf{y} - \mathbf{z}\|^2 = \|(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \mathbf{z})\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{z}\|^2$$

Hence

$$\|\mathbf{y} - \mathbf{z}\| \geqslant \|\mathbf{y} - \hat{\mathbf{y}}\|$$

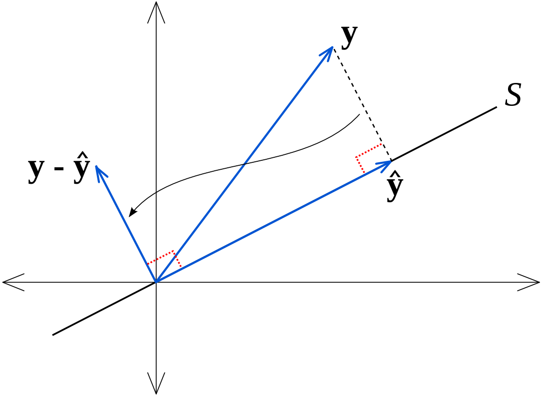


Figure: Orthogonal projection

Holding S fixed, we have a functional relationship

$$\mathbf{y} \;\mapsto\; \text{its orthogonal projection } \hat{\mathbf{y}} \in S$$

A well-defined function from $\mathbb{R}^n$ to $\mathbb{R}^n$

Typically denoted by $\mathbf{P}$

- Write $\mathbf{P} = \operatorname{proj} S$
- $\mathbf{P}\mathbf{y}$ represents the projection $\hat{\mathbf{y}}$

$\mathbf{P}$ is called the **orthogonal projection mapping onto S**

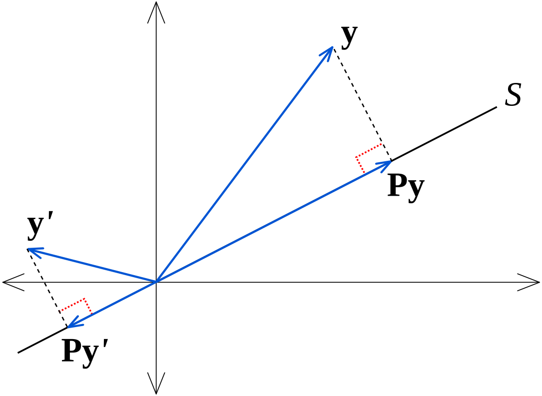


Figure: Orthogonal projection under $\mathbf{P}$

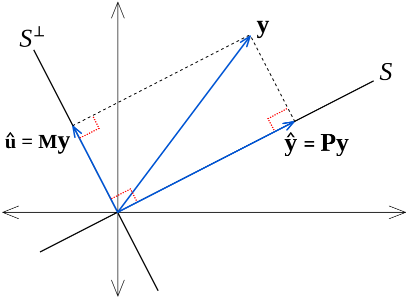
Let S be any linear subspace, and let $\mathbf{P} = \operatorname{proj} S$

Theorem. (OPT II) For any $\mathbf{y} \in \mathbb{R}^n$, we have

1. $\mathbf{P}\mathbf{y} \in S$,
2. $\mathbf{y} - \mathbf{P}\mathbf{y} \perp S$,
3. $\|\mathbf{y}\|^2 = \|\mathbf{P}\mathbf{y}\|^2 + \|\mathbf{y} - \mathbf{P}\mathbf{y}\|^2$,
4. $\|\mathbf{P}\mathbf{y}\| \leqslant \|\mathbf{y}\|$, and
5. $\mathbf{P}\mathbf{y} = \mathbf{y}$ if and only if $\mathbf{y} \in S$.

Proof of 3:

- Observe $\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{y} - \mathbf{P}\mathbf{y}$, apply ?



Theorem. (OPT III) If S is a linear subspace of $\mathbb{R}^n$, $\mathbf{P} = \operatorname{proj} S$ and $\mathbf{M} = \operatorname{proj} S^\perp$, then

$$\mathbf{Py} \perp \mathbf{My} \quad \text{and} \quad \mathbf{y} = \mathbf{Py} + \mathbf{My} \quad \text{for all } \mathbf{y} \in \mathbb{R}^n$$

Orthonormal Bases

An orthogonal set $O \subset \mathbb{R}^n$ is called an **orthonormal set** if $\|\mathbf{u}\| = 1$ for all $\mathbf{u} \in O$

If S is a linear subspace of $\mathbb{R}^n$,

- O is orthonormal
- $O \subset S$
- $\operatorname{span} O = S$

then O is called an **orthonormal basis** of S

It is, necessarily, a basis of S

Example. The canonical basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ forms an orthonormal basis of $\mathbb{R}^n$

Fact. If $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is an orthonormal basis of linear subspace S, then

$$\mathbf{x} = \sum_{i=1}^{k} \langle \mathbf{x}, \mathbf{u}_i \rangle \, \mathbf{u}_i \quad \text{for all} \quad \mathbf{x} \in S$$

Proof: Since $\mathbf{x} \in \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, we can find scalars $\alpha_1, \ldots, \alpha_k$ s.t.

$$\mathbf{x} = \sum_{j=1}^{k} \alpha_j \mathbf{u}_j \tag{1}$$

Taking the inner product with respect to $\mathbf{u}_i$ gives

$$\langle \mathbf{x}, \mathbf{u}_i \rangle = \sum_{j=1}^{k} \alpha_j \langle \mathbf{u}_j, \mathbf{u}_i \rangle = \alpha_i$$

Combining this result with (1) verifies the claim

Projection onto an Orthonormal Basis

Theorem. If $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is an orthonormal basis for S, then

$$\mathbf{P}\mathbf{y} = \sum_{i=1}^{k} \langle \mathbf{y}, \mathbf{u}_i \rangle \, \mathbf{u}_i, \quad \forall \, \mathbf{y} \in \mathbb{R}^n$$

Proof: Fix $\mathbf{y} \in \mathbb{R}^n$ and let $\mathbf{P}\mathbf{y}$ be as defined above. Clearly, $\mathbf{P}\mathbf{y} \in S$

We claim that $\mathbf{y} - \mathbf{P}\mathbf{y} \perp S$ also holds

Sufficies to show that $\mathbf{y} - \mathbf{P}\mathbf{y} \perp$ any basis element (why?)

We have

$$\left\langle \mathbf{y} - \sum_{i=1}^{k} \langle \mathbf{y}, \mathbf{u}_i \rangle \, \mathbf{u}_i, \mathbf{u}_j \right\rangle = \langle \mathbf{y}, \mathbf{u}_j \rangle - \sum_{i=1}^{k} \langle \mathbf{y}, \mathbf{u}_i \rangle \langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$$

Projection Using Matrix Algebra

Ex. If S is any linear subspace of $\mathbb{R}^n$ and $\mathbf{P} = \operatorname{proj} S$, then $\mathbf{P}$ is a linear function from $\mathbb{R}^n$ to $\mathbb{R}^n$

Therefore $\mathbf{P} = \operatorname{proj} S$ has a unique representation as a matrix

In what follows we use $\mathbf{P}$ for both the function and the matrix

But what does the matrix look like?

Theorem. If $\mathbf{P} = \operatorname{proj} S$ and the columns of $\mathbf{X} \in \mathcal{M}(n \times k)$ form a basis of S, then

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Proof: Given arbitrary $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, our claim is that

1. $\mathbf{P}\mathbf{y} \in S$, and
2. $\mathbf{y} - \mathbf{P}\mathbf{y} \perp S$

Here 1 is true because

$$\mathbf{P}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}\mathbf{a} \quad \text{when} \quad \mathbf{a} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

On the other hand, 2 is equivalent to the statement

$$\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \perp \mathbf{X}\mathbf{b} \quad \text{for all} \quad \mathbf{b} \in \mathbb{R}^K$$

This is true: If $\mathbf{b} \in \mathbb{R}^K$, then

$$(\mathbf{X}\mathbf{b})'[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = \mathbf{b}'[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y}] = 0$$

Example. Let $\mathbf{X} \in \mathcal{M}(n \times k)$ with linearly independent columns

Let

$$S := \operatorname{span} \mathbf{X} := \operatorname{span}\{\operatorname{col}_1 \mathbf{X}, \ldots, \operatorname{col}_k \mathbf{X}\}$$

Note that the columns of $\mathbf{X}$ form a basis of S

Now define

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad \text{and} \quad \mathbf{M} = \mathbf{I} - \mathbf{P}$$

- $\mathbf{P} = \operatorname{proj} S$ is called the **projection matrix**
- $\mathbf{M} = \operatorname{proj}(S^{\perp})$ is called the **annihilator**

Ex. Show that $\mathbf{P}$ and $\mathbf{M}$ are both idempotent and symmetric

Example. Suppose that $\mathbf{U} \in \mathcal{M}(n \times k)$ has orthonormal columns

Let $\mathbf{u}_i := \operatorname{col} \mathbf{U}_i$ for each i, let $S := \operatorname{span} \mathbf{U}$ and let $\mathbf{y} \in \mathbb{R}^n$

We know that the projection of $\mathbf{y}$ onto S is

$$\mathbf{P}\mathbf{y} = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{y}$$

Since $\mathbf{U}$ has orthonormal columns, we have $\mathbf{U}'\mathbf{U} = \mathbf{I}$

Hence

$$\mathbf{P}\mathbf{y} = \mathbf{U}\mathbf{U}'\mathbf{y} = \sum_{i=1}^{k} \langle \mathbf{u}_i, \mathbf{y} \rangle \, \mathbf{u}_i$$

We have recovered our earlier result about projecting onto the span of an orthonormal basis

Overdetermined Systems of Equations

Consider system of equations $\mathbf{X}\mathbf{b} = \mathbf{y}$

Given $\mathbf{X}$ and $\mathbf{y}$, we seek $\mathbf{b} \in \mathbb{R}^k$ satisfying this eq

Assume:

- $\mathbf{X} \in \mathcal{M}(n \times k)$ with linearly independent columns
- $\mathbf{b} \in \mathbb{R}^k$ and $\mathbf{y} \in \mathbb{R}^n$

Suppose that $n > k$ (more eqs than unknowns)

In this case, the system said to be **overdetermined**

(May not be able find a $\mathbf{b}$ that satisfies all n equations)

A geometric view: Observe that

$$\mathrm{span}(\mathbf{X}) = \{\text{all } \mathbf{Xb} \text{ with } \mathbf{b} \in \mathbb{R}^K\}$$

$$\therefore \quad \exists\, \mathbf{b} \text{ such that } \mathbf{Xb} = \mathbf{y} \iff \mathbf{y} \in \mathrm{span}(\mathbf{X})$$

When $k < n$ there is usually no such $\mathbf{b}$ because

- $\mathbf{y}$ is an arbitrary point in $\mathbb{R}^n$
- $\mathrm{span}(\mathbf{X})$ has dimension k
- k-dim subspace has measure zero in $\mathbb{R}^n$ whenever $k < n$

Hence:

1. Accept that an exact solution may not exist
2. Look instead for an approximate solution

How to define an approximate solution?

Find $\mathbf{b} \in \mathbb{R}^k$ such that $\mathbf{X}\mathbf{b}$ is as close to $\mathbf{y}$ as possible

That is, choose

$$\hat{\beta} := \operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^k} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|$$

The vector $\hat{\beta}$ is called the **least squares** solution to the overdetermined system

- But does it exist?
- Is it unique?

Thm. The unique minimizer of $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|$ over $\mathbf{b} \in \mathbb{R}^K$ is

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

Proof: Note that

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

Since $\mathbf{P}\mathbf{y}$ is the orthogonal projection onto $\mathrm{span}(\mathbf{X})$ we have

$$\|\mathbf{y} - \mathbf{P}\mathbf{y}\| \leqslant \|\mathbf{y} - \mathbf{z}\| \text{ for any } \mathbf{z} \in \mathrm{span}(\mathbf{X})$$

In other words,

$$\|\mathbf{y} - \mathbf{X}\hat{\beta}\| \leqslant \|\mathbf{y} - \mathbf{X}\mathbf{b}\| \text{ for any } \mathbf{b} \in \mathbb{R}^K$$

Linear Regression

Given pairs $(\mathbf{x}, y) \in \mathbb{R}^{K+1}$, choose $f \colon \mathbb{R}^K \to \mathbb{R}$ to minimize risk

$$R(f) := \mathbb{E}\left[(y - f(\mathbf{x}))^2\right]$$

If $\mathbb{E}$ unknown but sample available, replace risk with empirical risk:

$$\min_{f \in \mathscr{F}} \frac{1}{N} \sum_{n=1}^{N} (y_n - f(\mathbf{x}_n))^2$$

Letting $\mathscr{F} =$ linear functions and dropping $1/N$, the problem is

$$\min_{\mathbf{b} \in \mathbb{R}^K} \sum_{n=1}^{N} (y_n - \mathbf{b}' \mathbf{x}_n)^2$$

Switching to matrix notation,

$$\mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{x}_n := \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nK} \end{pmatrix} = n\text{-th obs on all regressors}$$

and

$$\mathbf{X} := \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{pmatrix} =: \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix}$$

We assume throughout that $N > K$ and $\mathbf{X}$ is full column rank

Ex. Verify that

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = \sum_{n=1}^{N} (y_n - \mathbf{b}'\mathbf{x}_n)^2$$

Since increasing transforms don't affect minimizers we have

$$\operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^K} \sum_{n=1}^{N} (y_n - \mathbf{b}'\mathbf{x}_n)^2 = \operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|$$

By the theory of overdetermined systems, the solution is

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

Notation

Let $\mathbf{P}$ and $\mathbf{M}$ be the projection and annihilator associated with $\mathbf{X}$:

$$\mathbf{P} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad \text{and} \quad \mathbf{M} := \mathbf{I} - \mathbf{P}$$

The **vector of fitted values** is

$$\hat{\mathbf{y}} := \mathbf{X}\hat{\beta} = \mathbf{P}\mathbf{y}$$

The **vector of residuals** is

$$\hat{\mathbf{u}} := \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{P}\mathbf{y} = \mathbf{M}\mathbf{y}$$

Applying the OPT we obtain

$$\hat{\mathbf{u}} \perp \hat{\mathbf{y}} \quad \text{and} \quad \mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{u}}$$

More standard definitions:

- **Total sum of squares** $:= \|\mathbf{y}\|^2$
- **Sum of squared residuals** $:= \|\hat{\mathbf{u}}\|^2$
- **Explained sum of squares** $:= \|\hat{\mathbf{y}}\|^2$

Fact. $\mathrm{TSS} = \mathrm{ESS} + \mathrm{SSR}$

Proof: $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{u}}$ and $\hat{\mathbf{u}} \perp \hat{\mathbf{y}}$

Centered R^2 is defined as

$$R^2 = \frac{\sum_{n=1}^{N} (\hat{y}_n - \bar{y})^2}{\sum_{n=1}^{N} (y_n - \bar{y})^2}$$

Gram-Schmidt Orthogonalization

Theorem. For each linearly independent set $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subset \mathbb{R}^n$, there exists an orthonormal set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ with

$$\operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_i\} = \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_i\} \quad \text{for} \quad i = 1, \ldots, k$$

Construction uses the **Gram-Schmidt orthogonalization** procedure: For $i = 1, \ldots, k$, set

1. $S_i := \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_i\}$ and $\mathbf{M}_i := \operatorname{proj} S_i^\perp$
2. $\mathbf{v}_i := \mathbf{M}_{i-1}\mathbf{x}_i$ where $\mathbf{M}_0$ is the identity mapping
3. $\mathbf{u}_i := \mathbf{v}_i / \|\mathbf{v}_i\|$

Sequence $\mathbf{u}_1, \ldots, \mathbf{u}_k$ has the stated properties (proof omitted)

QR Decomposition

Theorem. If $\mathbf{X} \in \mathcal{M}(n \times k)$ with linearly independent columns, then $\exists$ a factorization $\mathbf{X} = \mathbf{QR}$ where

1. $\mathbf{R}$ is $k \times k$, upper triangular and nonsingular
2. $\mathbf{Q}$ is $n \times k$, with orthonormal columns

Proof sketch: Let $\mathbf{x}_j := \mathrm{col}_j(\mathbf{X})$, let $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ be orthonormal with same span as $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and let $\mathbf{Q}$ be formed from cols $\mathbf{u}_i$

Since $\mathbf{x}_j \in \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$, we have

$$\mathbf{x}_j = \sum_{i=1}^{j} \langle \mathbf{u}_i, \mathbf{x} \rangle \, \mathbf{u}_i \quad \text{for } j = 1, \ldots, k$$

Some rearranging gives $\mathbf{X} = \mathbf{QR}$

Linear Regression via QR Decomposition

For $\mathbf{X}$ and $\mathbf{y}$ as above we have

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

Using the QR decomposition $\mathbf{X} = \mathbf{QR}$ gives

$$\begin{aligned}
\hat{\beta} &= (\mathbf{R}'\mathbf{Q}'\mathbf{QR})^{-1}\mathbf{R}'\mathbf{Q}'\mathbf{y} \\
&= (\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{Q}'\mathbf{y} \\
&= \mathbf{R}^{-1}(\mathbf{R}')^{-1}\mathbf{R}'\mathbf{Q}'\mathbf{y} = \mathbf{R}^{-1}\mathbf{Q}'\mathbf{y}
\end{aligned}$$

Numerical routines use $\mathbf{R}\hat{\beta} = \mathbf{Q}'\mathbf{y}$ and back substitution