3.0 Linear Regression with Matrices

- Answer Questions
- Assumptions
- Maximum Likelihood Estimators
- A Second Proof of $\hat{\boldsymbol{\beta}}$
- Residuals
- Prediction and Confidence Intervals
- Bases and Projections
- More on the Hat Matrix

3.1 Assumptions

Multiple linear regression assumes that:

- The vectors of explanatory variables $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, are not random.
- Given $\boldsymbol{x}_i$, the response Y_i is normally distributed.
- $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$.
- The variance of Y_i, given $\boldsymbol{x}_i$, is σ^2 (usually unknown, but the same for all i).
- The $Y_1, \ldots, Y_n$ are independent given the $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$.

The statistical model that satisfies these assumptions is called the **General Linear Model**.

In matrix notation, we write

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{X}$ is $n \times (p+1)$, $\boldsymbol{\beta}$ is $(p+1) \times 1$ and $\boldsymbol{\epsilon}$ is $n \times 1$.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

3.2 MLEs

In the general linear model, the distribution of the observations Y_i has density

$$f(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2\right].$$

To see this, realize that the independence of the Y_i values implies their joint density is just the product of each density; i.e., $f(\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i)$, where each term in the product is normal with mean $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ and common variance σ^2.

We want to find the maximum likelihood estimates of $\beta_0, \ldots, \beta_p$. Inspection of the joint density shows it is sufficient to find the estimates that minimize the quadratic form

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

The values that minimize Q are sometimes called the **Least Squares Estimates**, which here, but not in general, agree with the maximum likelihood estimates.

To solve, find the partial derivatives $\partial Q/\partial \beta_j$ for $j = 0, \ldots, p$. Set these partials to zero, giving $p + 1$ equations in $p + 1$ unknowns. Solving these gives the estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$.

The **Design Matrix** is the $n \times (p+1)$ matrix $\boldsymbol{X}$ whose ith row is $(1, x_{i1}, \ldots, x_{ip})$ for $i = 1, \ldots n$. The name comes from the fact that in some multiple regressions, the experimenter gets to select the values of the explanatory variables by design. When this is possible, one can learn more efficiently about the relationship between the response variable and the explanatory variables.

For now, assume that the set of $p+1$ equations has a unique solution. This implies the design matrix is full rank; i.e., has rank $p+1$.

The rank of an $n \times m$ matrix is

- the number of linearly independent rows, or
- the number of linearly independent columns, or
- the dimensions of the largest square submatrix that is invertible (nonsingular).

Some facts:

- $\text{Rank}(\boldsymbol{AB}) \leq \min\{\text{Rank}(\boldsymbol{A}), \text{Rank}(\boldsymbol{B})\}$.
- If $\boldsymbol{A}$ is $n \times n$ and has determinant equal to 0, then $\text{Rank}(\boldsymbol{A}) < n$.
- If $\boldsymbol{A}$ is $m \times n$, then $\boldsymbol{A}^\top$ is the $n \times m$ **transpose** that turns the rows into columns.

Let $\boldsymbol{X}$ denote the design matrix, $\boldsymbol{\beta}$ the true but unknown coefficients, $\boldsymbol{y}$ the response vector, and $\hat{\boldsymbol{\beta}}$ the maximum likelihood estimates of the coefficients. We shall show that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Define the inverse $\boldsymbol{A}^{-1}$ as the square matrix such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, the identity matrix that is symmetric with ones on the diagonal and zeroes elsewhere.

Proof: Write $Q = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Then

$$\frac{\partial Q}{\partial \beta_k} = -2\sum_{i=1}^{n} x_{ik}y_i + 2\sum_{j=0}^{p}\beta_j\left(\sum_{i=1}^{n} x_{ik}x_{ij}\right).$$

Set the right-hand sides of the $p+1$ equations to zero, and then rewrite the equations as $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{y}$. Because we assumed that $\boldsymbol{X}^\top\boldsymbol{X}$ had rank $p+1$, then from linear algebra we know it is invertible, and thus

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \qquad \blacksquare$$

Let $S^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$. Then one can show that the maximum likelihood estimate of σ^2 is $\hat{\sigma}^2 = S^2/n$, but this is biased. An unbiased estimate is $\hat{\sigma}^2 = S^2/(n-p-1)$.

Theorem: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Proof: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbb{E}[\boldsymbol{y}] = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\beta}$.

Theorem: $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sigma^2$.

Proof: Recall, from linear algebra, that if $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, then $\boldsymbol{A}^\top = \boldsymbol{C}^\top \boldsymbol{B}^\top$.

$$\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{\beta}}] &= \mathrm{Var}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}] \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathrm{Var}[\boldsymbol{y}] [(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top]^\top \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \sigma^2 \boldsymbol{I} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sigma^2.
\end{aligned}$$

This uses the linear algebra fact that $\boldsymbol{X}^\top \boldsymbol{X}$ is symmetric, so its inverse is symmetric, so the transpose of the inverse is itself.

3.3 A Proof of $\hat{\boldsymbol{\beta}}$

Some facts:

- If $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, then $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{A}$.
- If $\boldsymbol{A}$ is symmetric and $\boldsymbol{y} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, then $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}^\top \boldsymbol{A}$.
- If $\boldsymbol{A}$ is symmetric, the derivative of the transpose is the transpose of the derivative.
- $(\boldsymbol{A}\boldsymbol{B})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$.

The rules for differentiating vectors and matrices look very much like the usual calculus for univariate quantities.

Theorem: If $\boldsymbol{X}$ has rank $p+1$, then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$.

Proof: Recall that $Q = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Thus $Q = \boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}$. Taking the derivative of Q with respect to $\boldsymbol{\beta}$ and using the facts shows that the derivative of the first term is zero, the second term goes to $\boldsymbol{y}^\top\boldsymbol{X}$, the third term goes to $(\boldsymbol{X}^\top\boldsymbol{y})^\top = \boldsymbol{y}^\top\boldsymbol{X}$, and the last term has derivative $\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}$.

Set the derivative to zero and solve. One gets $0 = -2\boldsymbol{y}^\top\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}$ so $\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\boldsymbol{X}) = \boldsymbol{y}^\top\boldsymbol{X}$. Take the transpose of both sides to get $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{y}$. Since $\boldsymbol{X}$ has rank $p+1$, then $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible, and so $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. ■

3.4 Residuals

What else can we do with matrix algebra in regression?

- Derive properties of regression diagnostics, such as residuals.
- Calculate prediction intervals for outcomes at a new value $\boldsymbol{x}$.
- Interpret the MLEs geometrically.

The ith residual $\hat{\epsilon}_i$ is the difference between the actual value of y_i in the training sample and the predicted value $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$. In matrix notation, this is

$$\begin{aligned} \hat{\boldsymbol{\epsilon}} &= \boldsymbol{y} - \hat{\boldsymbol{y}} \\ &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{y}. \end{aligned}$$

Theorem: $\mathbb{E}[\hat{\boldsymbol{\epsilon}}] = \mathbf{0}$.

Proof: $\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{y}$ so

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\epsilon}}] &= \mathbb{E}[(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{y}] \\
&= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\mathbb{E}[\boldsymbol{y}] \\
&= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top * \boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} \\
&= \mathbf{0}. \qquad \blacksquare
\end{aligned}$$

Of course, this is completely unsurprising.

Consider a random vector $\boldsymbol{y} \in \mathbb{R}^n$ with mean $\boldsymbol{\mu}$. Its covariance matrix, $\mathrm{Cov}(\boldsymbol{y})$, is $\boldsymbol{\Sigma}$ where

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^\top].$$

This definition does not require multivariate normality.

Here $\boldsymbol{\Sigma}$ is $n \times n$ and symmetric; i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top$. Its diagonal entries are the the variances of the i random variable y_i. The (i, j)th entry is the covariance between the random variables y_i and y_j.

Theorem: $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{y}\ \boldsymbol{y}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$.

The proof is an easy exercise. The logic is exactly like the corresponding univariate case, with $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

If $\boldsymbol{A}$ is not random, it is pretty clear that $\mathbb{E}[\boldsymbol{Ay}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{y}] = \boldsymbol{A\mu}$. For the covariance of a matrix product, it is a little more interesting.

Theorem: $\mathrm{Cov}(\boldsymbol{Ay}) = \boldsymbol{A\Sigma A}^\top$.

Proof: Use the definition.

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{Ay}) &= \mathbb{E}[\boldsymbol{Ayy}^\top\boldsymbol{A}^\top] - \mathbb{E}[\boldsymbol{Ay}](\mathbb{E}[\boldsymbol{Ay}])^\top \\
&= \boldsymbol{A}\mathbb{E}[\boldsymbol{yy}^\top]\boldsymbol{A}^\top - \boldsymbol{A}\mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^\top\boldsymbol{A}^\top \\
&= \boldsymbol{A}\left(\mathbb{E}[\boldsymbol{yy}^\top] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^\top\right)\boldsymbol{A}^\top \\
&= \boldsymbol{A}\mathrm{Cov}(\boldsymbol{y})\boldsymbol{A}^\top. \qquad \blacksquare
\end{aligned}$$

Note that in the second line we used the fact that $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top\boldsymbol{A}^\top$.

Some facts: The inverse of a symmetric matrix is symmetric. And $\boldsymbol{X}^\top \boldsymbol{X}$ is symmetric, since its transpose equals itself.

Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. This called the **hat matrix** since $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$.

Theorem: $\text{Cov}(\boldsymbol{r}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$.

Proof: Since $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$, then

$$\begin{aligned}
\text{Cov}(\boldsymbol{r}) &= (\boldsymbol{I} - \boldsymbol{H})\text{Cov}(\boldsymbol{y})(\boldsymbol{I} - \boldsymbol{H})^\top \\
&= (\boldsymbol{I} - \boldsymbol{H})\sigma^2\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{H})^\top \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I}^\top - \boldsymbol{H}^\top) \\
&= \sigma^2(\boldsymbol{I}\boldsymbol{I}^\top - \boldsymbol{I}\boldsymbol{H}^\top - \boldsymbol{H}\boldsymbol{I}^\top + \boldsymbol{H}\boldsymbol{H}^\top).
\end{aligned}$$

Note that $\boldsymbol{I}^\top = \boldsymbol{I}$ and that $\boldsymbol{H}\boldsymbol{H}^\top = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top[\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top]^\top$ which simplifies to $\boldsymbol{H}$. ■

One implication of the theorem is that the variance of the ith residual is $(1 - h_{ii})\sigma^2$, which would by estimated by $(1 - h_{ii})\hat{\sigma}^2$, where $\hat{\sigma}^2$ is the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The **standardized residual** is $(y_i - \hat{y}_i)/\sqrt{(1 - h_{ii})}\hat{\sigma}$. This follows a t distribution with $n - p - 1$ degrees of freedom.

For example, one could use the standardized residual to decide whether an observation might be an outlier.

3.5 Confidence and Prediction Intervals

Suppose we want to predict the response Y for a new vector of covariates $\boldsymbol{x}_{\text{new}}$. The point estimate is easy: $\boldsymbol{x}_{\text{new}}^{\top}\hat{\boldsymbol{\beta}}$.

The variance depends upon whether we want to estimate the average response for an observation with the value $\boldsymbol{x}_{\text{new}}$ (i.e., the location of response on the estimated hyperflat) or the value of an individual with $\boldsymbol{x}_{\text{new}}$.

This is the distinction we saw previously for simple linear regression between setting a confidence interval on the response $\hat{\beta}_0 + \hat{\beta}_1 x$ and a prediction interval on an individual that has explanatory variable x.

The variance for the prediction interval should be larger, since the individual is unlikely to have the average value.

For the average response at $\boldsymbol{x}_{\text{new}}$ we can calculate

$$\begin{aligned}
\text{Var}(\boldsymbol{x}_{\text{new}}^\top \hat{\beta}) &= \boldsymbol{x}_{\text{new}}^\top \text{Var}(\hat{\beta})(\boldsymbol{x}_{\text{new}}^\top)^\top \\
&= \boldsymbol{x}_{\text{new}}^\top \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\boldsymbol{x}_{\text{new}}^\top)^\top \\
&= \sigma^2 \boldsymbol{x}_{\text{new}}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{\text{new}}.
\end{aligned}$$

For σ^2, we use the usual estimate

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The two-sided $(1-\alpha)100\%$ confidence interval is just

$$U, L = \boldsymbol{x}_{\textbf{new}}^\top \hat{\boldsymbol{\beta}} \pm \hat{\sigma} \sqrt{\boldsymbol{x}_{\textbf{new}}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{\textbf{new}}} \times t_{n-p-1,\alpha/2}.$$

For predictions on individuals, there are two sources that contribute to the variance:

- the variance due to estimating the average response
- the variance of an individual's value around the average response.

The prediction interval must combine both:

$$\begin{aligned}
\mathrm{Var}(\hat{y}_{\mathrm{new,ind}}) &= \mathrm{Var}(\hat{y}_{\mathrm{new}} + \epsilon_{\mathrm{new}}) \\
&= \mathrm{Var}(\hat{y}_{\mathrm{new}}) + \mathrm{Var}(\epsilon_{\mathrm{new}}) \\
&= \sigma^2 \boldsymbol{x}_{\mathrm{new}}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{\mathrm{new}} + \sigma^2 \\
&= \sigma^2 (\boldsymbol{x}_{\mathrm{new}}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{\mathrm{new}} + 1).
\end{aligned}$$

This uses the fact that $\hat{y}_{\mathrm{new}}$ and ϵ_{new} are independent.

The $\hat{y}_{\text{new}}$ and ϵ_{new} are independent since

- ϵ_{new} was not used in fitting the model **and**
- the general linear model assumes that the errors (ϵs) are independent.

Therefore the two-sided $(1-\alpha)100\%$ prediction interval is just

$$U, L = \boldsymbol{x}_{\mathbf{new}}^{\top}\hat{\boldsymbol{\beta}} \pm \hat{\sigma}\sqrt{1 + \boldsymbol{x}_{\mathbf{new}}^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{x}_{\mathbf{new}}} \times t_{n-p-1,\alpha/2}.$$

For the case of simple linear regression, with one explanatory variable, it is easy to show that the confidence interval formula and the prediction interval formula agree with the formulae given in Lecture 2. It just takes some algebra.

3.6 Bases and Projections

A **basis** of a vector space $\mathbb{R}^p$ is a set of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ such that any vector $\boldsymbol{v} \in \mathbb{R}^p$ can be expressed as $\boldsymbol{v} = \sum_{i=1}^{p} a_i \boldsymbol{x}_i$ but none of the $\boldsymbol{x}_i$ is a linear combination of the other basis vectors.

For example, a standard choice of basis is $(1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $(0, 0, 1, \ldots, 0)$ up to $(0, 0, 0, \ldots, 1)$. In fact, this is an **orthonormal** basis since each member is orthogonal to the others (i.e., had dot product equal to zero) and is normalized (has unit length).

A subspace of $\mathbb{R}^p$ is $\mathbb{R}^q$ for $q < p$. Any vector in $\mathbb{R}^q$ can be written as a linear combination of q basis vectors.

A **projection matrix** $\boldsymbol{P}$ is a square matrix such that $\boldsymbol{P}^2 = \boldsymbol{P}$. For example, the projection matrix

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

projects $\mathbb{R}^2$ onto the subspace $\mathbb{R}^1$ corresponding to the line $x_1 = x_2$, since $\boldsymbol{P}(x_1, x_2)^\top = (x_2, x_2)^\top$.

Note: The line $x_1 = x_2$ is a proper subspace of $\mathbb{R}^2$. It contains $\mathbf{0}$, and $(1, 1)^\top$ is a basis element.

Note: It makes sense that $\boldsymbol{P}^2 = \boldsymbol{P}$ since if one applies the same projection twice, nothing changes.

A projection matrix is symmetric if and only if the vector space projection is orthogonal. In an **orthogonal projection** of $\mathbb{R}^p$ onto $\mathbb{R}^q$ for $q < p$, any vector $\boldsymbol{v}$ can be written as $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ where $\boldsymbol{v}_1 = \boldsymbol{P}\boldsymbol{v}$ and $\boldsymbol{v}_2 = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{v}$.

This implies that the dot product $\boldsymbol{v}_1^\top \boldsymbol{v}_2 = 0$.

Consider the projection matrix

$$\tilde{\boldsymbol{P}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

This is symmetric and projects $(x_1, x_2)^\top$ onto $(x_1, 0)^\top$. And it is an orthogonal projection since

$$(x_1, x_2)^\top = \tilde{\boldsymbol{P}}(x_1, x_2)^\top + (\boldsymbol{I} - \tilde{\boldsymbol{P}})(x_1, x_2)^\top = (x_1, 0)^\top + (0, x_2)^\top$$

with $< (x_1, 0), (0, x_2) >= 0$.

In contrast, $\boldsymbol{P}$ is not an orthogonal projection since

$$\boldsymbol{P}(x_1, x_2)^\top + (\boldsymbol{I} - \boldsymbol{P})(x_1, x_2)^\top = (x_2, x_2)^\top + (x_1 - x_2, 0)^\top$$

and $< (X_2, x_2), (x_1 - x_2, 0 > \neq 0$.

The orthogonal projection of the hat matrix minimizes the sum of the squared vertical distances onto the subspace. Recall that in multiple linear regression we assume the explanatory variables are measured without error, and thus we want to minimize the sum of the squared vertical distances.

The hat matrix is a projection.

$$\boldsymbol{H}^2 = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \times \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{H}.$$

But to what subspace does it project? To the column space of $\boldsymbol{X}$.

The **column space** $\mathcal{C}(\boldsymbol{A})$ is the space spanned by the columns of $\boldsymbol{A}$. It is formed by taking all possible linear combinations of the columns of $\boldsymbol{A}$. One views the columns of $\boldsymbol{A}$ as the basis elements of the space. If $\boldsymbol{A}$ has rank n, then the column space has dimension n.

The **null space** of an $m \times n$ matrix $\boldsymbol{A}$ is the space of vectors in $\mathbb{R}^n$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$. If $\boldsymbol{A}$ is invertible, then the null space is just $\boldsymbol{0}$. For

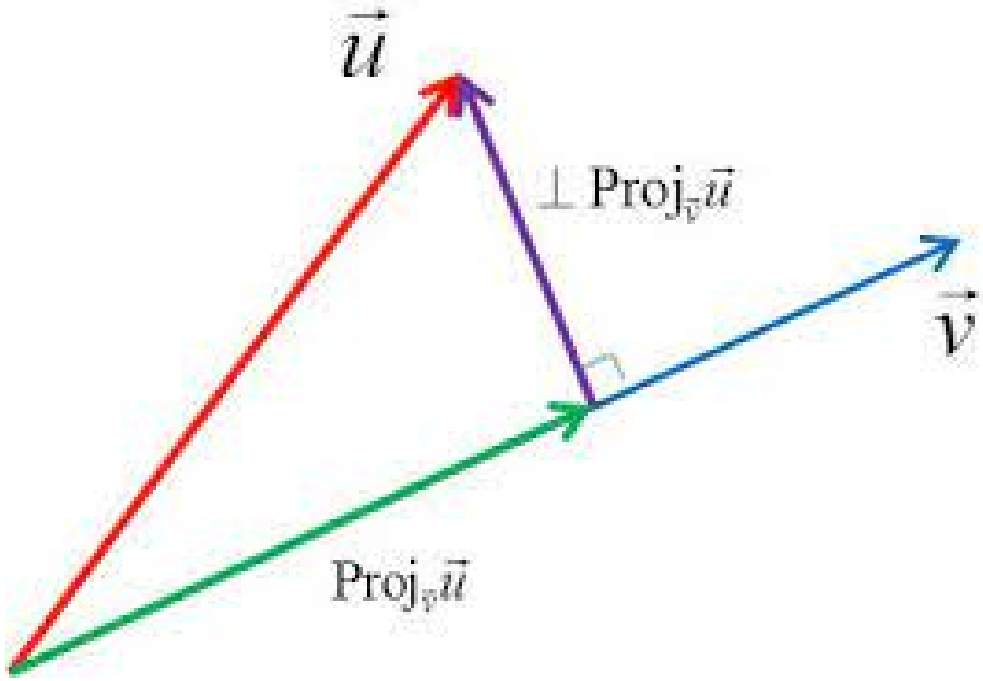
$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix},$$

the null space is $\{(x_1, 0)^\top\}$.

The **row space** $\mathcal{R}(\boldsymbol{A})$ is the space formed by taking all possible linear combinations of the rows of $\boldsymbol{A}$. If $\boldsymbol{A}$ has rank m, then the row space has dimension m.

Some properties of orthogonal projections $\boldsymbol{H}$:

- $\boldsymbol{v}_1 = \boldsymbol{H}\boldsymbol{v}$, $\boldsymbol{v}_2 = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{v}$ and $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$.
- $\boldsymbol{v}_1$ is orthogonal to $\boldsymbol{v}_2$, so their dot product is zero. We write this as $\boldsymbol{v}_1 \perp \boldsymbol{v}_2$.
- $\boldsymbol{H}$ and $\boldsymbol{I} - \boldsymbol{H}$ are symmetric, and their squares equal themselves.
- In the hat matrix, if $\boldsymbol{X}$ is $n \times (p+1)$ of rank $p+1$, then $\boldsymbol{H}$ has rank $p+1$.
- In the hat matrix, the eigenvalues of $\boldsymbol{H}$ consist of $p+1$ ones and $n-p-1$ zeroes.
- $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$, $(\boldsymbol{I}$ - $\boldsymbol{H})$ $\boldsymbol{X} = \boldsymbol{0}$.
- For the hat matrix, the trace of $\boldsymbol{H}$ equals the rank of $\boldsymbol{X}$.



This image uses the notation $\perp \boldsymbol{v}$ to denote the component of $\boldsymbol{u}$ that is orthogonal to the orthogonal projection of $\boldsymbol{u}$ onto the subspace containing $\boldsymbol{v}$. In this class, we shall usually denote this by $\boldsymbol{v}^{\perp}$.

More results from linear algebra:

Theorem: A vector that is orthogonal to the column space of an $n \times m$ matrix $\boldsymbol{X}$ lies in the null space of $\boldsymbol{X}^\top$.

Proof: Let $\boldsymbol{u}$ be orthogonal to the column space of an $n \times m$ matrix $\boldsymbol{X}$. So for any vector $\boldsymbol{v}$ in the column space of $\boldsymbol{X}$, $\boldsymbol{v}^\top \boldsymbol{u} = 0$. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be the columns of $\boldsymbol{X}$. Each $\boldsymbol{x}_i$ is in the column space of $\boldsymbol{X}$, so $\boldsymbol{x}_i^\top \boldsymbol{u} = 0 \; \forall i$. Since the rows of $\boldsymbol{X}^\top$ are the columns of $\boldsymbol{X}$, then $\boldsymbol{u}$ lies in the null space of $\boldsymbol{X}^\top$. ■

Theorem: A vector in the column space of an $n \times m$ matrix $\boldsymbol{X}$ is orthogonal to a vector in the null space of $\boldsymbol{X}^\top$.

Proof: Let $\boldsymbol{v}$ be in the column space of $\boldsymbol{X}$ and $\boldsymbol{u}$ be in the null space of $\boldsymbol{X}^\top$. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be the columns of $\boldsymbol{X}$. Since $\boldsymbol{v}$ the columns of $\boldsymbol{X}$ are a basis for the column space, then there exists $\boldsymbol{a}$ such that

$$\boldsymbol{v} = \sum_{i=1}^{m} a_i \boldsymbol{x}_i = \boldsymbol{X}\boldsymbol{a}.$$

So the inner product $\boldsymbol{u}^\top \boldsymbol{v} = \boldsymbol{u}\boldsymbol{X}\boldsymbol{a} = (\boldsymbol{X}^\top \boldsymbol{u})\boldsymbol{a}$ which is zero since we previously showed that $\boldsymbol{u}$ is in the null space of $\boldsymbol{X}^\top$. ■.

Note: Thus $\hat{\boldsymbol{\beta}}$ is obtained by projecting $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.

3.7 More on the Hat Matrix

The hat matrix projects $\boldsymbol{y} \in \mathbb{R}^n$ onto the column space of $\boldsymbol{X}$, which has dimension $p+1$. That imposes $p+1$ linear constraints on what the response can be, accounting for the loss of $p+1$ degrees of freedom.

The diagonal entry of the hat matrix lies in $[0,1]$ since $\boldsymbol{H}^2 = \boldsymbol{H}$. A diagonal element of $\boldsymbol{H}$ is h_{ii}. The corresponding diagonal element of $\boldsymbol{H}^2$ is $\boldsymbol{h}_i^\top \boldsymbol{h}_i = \sum h_{ij}^2 = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2$. Thus $h_{ii}^2 \leq h_{ii}$ so $0 \leq h_{ii} \leq 1$.

An **eigenvalue** of $\boldsymbol{H}$ satisfies $\boldsymbol{H}\boldsymbol{x} = \lambda \boldsymbol{x}$. Since $\boldsymbol{H}^2\boldsymbol{x} = \boldsymbol{H}\boldsymbol{x} = \lambda \boldsymbol{H}\boldsymbol{x} = \lambda^2 \boldsymbol{x}$, it follows that $\lambda = \lambda^2$. An eigenvalue of a projection matrix is zero or one.