MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Hat Matrix: Properties and Interpretation

Week 5, Lecture 1

1 Hat Matrix

1.1 From Observed to Fitted Values

The OLS estimator was found to be given by the $(p^* \times 1)$ vector,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

The predicted values $\widehat{\mathbf{y}}$ can then be written as,

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \Big(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Big)\mathbf{y} =: \mathbf{H}\mathbf{y},$$

where $\mathbf{H} := \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is an $n \times n$ matrix, which "*puts the hat on* $\mathbf{y}$" and is therefore referred to as the **hat matrix**. This shows that the fitted values are, in fact, a **linear function** of the observed values, such that for any $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$, we have letting $\widehat{\mathbf{y}} =: f(\mathbf{y})$,

$$f(\mathbf{y} + \mathbf{y}') = \mathbf{H}(\mathbf{y} + \mathbf{y}') = \mathbf{H}\mathbf{y} + \mathbf{H}\mathbf{y}' = f(\mathbf{y}) + f(\mathbf{y}'),$$

and for any $a \in \mathbb{R}$, $f(a\mathbf{y}) = af(\mathbf{y})$.

Similarly, the **residuals** can also be expressed as a function of $\mathbf{H}$,

$$\widehat{\mathbf{e}} := \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y},$$

with $\mathbf{I}$ denoting the $n \times n$ identity matrix, and where again the residuals can also be seen to be a linear function of the observed values, $\mathbf{y}$. In summary, we therefore have

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y} \qquad \text{and} \qquad \widehat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

Crucially, it can be shown that both $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are **orthogonal projections**. This can be represented visually, by contrasting two different viewpoints on multiple linear regression:

i. Thus far, we have mainly be concerned with what may be called the **variable space**, $\mathbb{R}^{p^*}$, in which each subject is a *point*, and the variables are dimensions.

ii. We need to accomplish a change of perspective by considering the **subject space**, $\mathbb{R}^n$. In this Euclidean space, each subject is a *dimension*, whereas the $\mathbf{y}$, $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{e}}$ are treated as *vectors*.

1.2 Hat Matrix as Orthogonal Projection

The matrix of a projection, which is also symmetric is an **orthogonal projection**. We can show that both $\mathbf{H}$ and $\mathbf{I}-\mathbf{H}$ are orthogonal projections. These two conditions can be re-stated as follows:

1. A **square** matrix $\mathbf{A}$ is a *projection* if it is **idempotent**,
2. A **projection A** is *orthogonal* if it is also **symmetric**.

A projection, which is not orthogonal is called an *oblique* projection. Specifically, $\mathbf{H}$ projects $\mathbf{y}$ onto the **column space** of $\mathbf{X}$, whereas $\mathbf{I}-\mathbf{H}$ projects $\mathbf{y}$ onto the **orthogonal complement** of the image of $\mathbf{H}$. The column space of matrix is defined as the **range** or the **image** of the corresponding **linear transformation**. Formally, as $\mathbf{H}$ is an $n \times n$ matrix, its column space is defined as

$$\operatorname{col}(\mathbf{H}) := \left\{ \sum_{i=1}^{n} c_j \mathbf{H}_{\cdot,j} : c_j \in \mathbb{R},\ \forall\, j = 1, \ldots, n \right\},$$

where $\mathbf{H}_{\cdot,j}$ denotes the j^{th} column of $\mathbf{H}$. This is the **span** of the linearly independent columns of $\mathbf{H}$.

1.3 Idempotency of the Hat Matrix

$\mathbf{H}$ is an $n \times n$ **square** matrix, and moreover, it is **idempotent**, which can be verified as follows,

$$\begin{aligned}\mathbf{HH} &= \Big(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Big)\Big(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Big) \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{H}.\end{aligned}$$

Similarly, $\mathbf{I}-\mathbf{H}$ can also be shown to be idempotent,

$$(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \mathbf{I} - 2\mathbf{H} + \mathbf{HH} = (\mathbf{I}-\mathbf{H}).$$

Every *square* and *idempotent* matrix is a **projection matrix**. The meaning of a projection can be understood with the following 2×2 example of a projection matrix, $\mathbf{P}$ which sends any 2-dimensional vector, $\mathbf{x}$, to a one-dimensional subspace,

$$\mathbf{Px} = \begin{vmatrix} 1 & \alpha \\ 0 & 0 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} x_1 + \alpha x_2 \\ 0 \end{vmatrix}.$$

The idempotency of $\mathbf{P}$ implies that once a vector has been projected to a subspace, it "*remains*" there, even if we re-apply the same projection.

1.4 Symmetry of the Hat Matrix

For any *square* and *invertible* matrices, the inverse and transpose operator commute,

$$(\mathbf{X}^T)^{-1} = (\mathbf{X}^{-1})^T.$$

Moreover, the transpose unary operator is an **involution**, since $(\mathbf{X}^T)^T = \mathbf{X}$. Thus, it follows that $(\mathbf{X}^T\mathbf{X})^{-1}$ is **self-transpose** (i.e. **symmetric**), since

$$[(\mathbf{X}^T\mathbf{X})^{-1}]^T = [(\mathbf{X}^T\mathbf{X})^T]^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}.$$

One can apply these two rules in order to show that $\mathbf{H}$ is *self-transpose* or *symmetric*,

$$\mathbf{H}^T = \big[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\big]^T = \mathbf{X}\big[(\mathbf{X}^T\mathbf{X})^{-1}\big]^T\mathbf{X}^T = \mathbf{H},$$

This symmetry is inherited by $\mathbf{I} - \mathbf{H}$, in the following manner,

$$(\mathbf{I} - \mathbf{H})^T = \mathbf{I}^T - \mathbf{H}^T = (\mathbf{I} - \mathbf{H}).$$

The orthogonality of a matrix is best understood through an example. Consider the following 2×2 matrix, $\mathbf{P}$, and its complement $\mathbf{I} - \mathbf{P}$, when applied to any vector $\mathbf{x} \in \mathbb{R}^2$,

$$\mathbf{P}\mathbf{x} = \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} x_1 \\ 0 \end{vmatrix}, \qquad \text{and} \qquad (\mathbf{I} - \mathbf{P})\mathbf{x} = \begin{vmatrix} 0 & 0 \\ 0 & 1 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} 0 \\ x_2 \end{vmatrix}.$$

Clearly, one can immediately see that the two resulting vectors are perpendicular, $\langle \mathbf{P}\mathbf{x}, (\mathbf{I} - \mathbf{P})\mathbf{x} \rangle = 0$; and that therefore they span two orthogonal **linear subspaces** of $\mathbb{R}^2$.

1.5 Orthogonal Projections and Orthogonal Matrices

One should be careful not to confuse the matrix of an orthogonal projection with an **orthogonal matrix**. Recall that the latter satisfies,

$$\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}.$$

Specifically, one may observe that an orthogonal projection $\mathbf{H}$ projects vectors, which are orthogonal to vectors in its null space, whereas an orthogonal matrix has column vectors, which are orthogonal. In general, an orthogonal matrix does not induce an orthogonal projection. In fact, it can be shown that the sole matrix, which is both an orthogonal projection and an orthogonal matrix is the *identity matrix*.

2 Orthogonal Decomposition

2.1 Range and Kernel of the Hat Matrix

By combining our definitions of the **fitted values** and the **residuals**, we have

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y} \qquad \text{and} \qquad \widehat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

These equations correspond to an **orthogonal decomposition** of the observed values, such that

$$\mathbf{y} = \widehat{\mathbf{y}} + \widehat{\mathbf{e}} = \mathbf{H}\mathbf{y} + (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

Observe that the **column space** or **range** of $\mathbf{H}$, denoted $\operatorname{col}(\mathbf{H})$, is identical to the column space of $\mathbf{X}$. Recall that $\mathbf{X}$ is a matrix with **real entries**, and therefore it is known that the rank of $\mathbf{X}$ is equal to the rank of its **Gram** matrix, defined as $\mathbf{X}^T\mathbf{X}$, such that

$$\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{X}^T\mathbf{X}) = p^*.$$

Moreover, we can use some basic operations on matrix ranks, such that for any *square* matrix $\mathbf{A}$ of order $k \times k$; if $\mathbf{B}$ is an $n \times k$ matrix of rank k, then

$$\operatorname{rank}(\mathbf{B}\mathbf{A}) = \operatorname{rank}(\mathbf{A}), \qquad \text{and} \qquad \operatorname{rank}(\mathbf{A}\mathbf{B}^T) = \operatorname{rank}(\mathbf{A}).$$

By assumption, we have $\operatorname{rank}(\mathbf{X}) = p^*$. Thus, it suffices to apply these rules in order to obtain

$$\operatorname{rank}(\mathbf{H}) = \operatorname{rank}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \operatorname{rank}((\mathbf{X}^T\mathbf{X})^{-1}) = \operatorname{rank}(\mathbf{X}^T\mathbf{X}) = p^*,$$

and to note that only full rank matrices are *invertible*, which implies that matrix inversion preserves rank. Thus, the column space of $\mathbf{H}$ is equal to the column space of $\mathbf{X}$, such that

$$\operatorname{col}(\mathbf{H}) = \operatorname{col}(\mathbf{X}),$$

where the column space of $\mathbf{X}$ is the set of all vectors that can be obtained as linear combinations of the columns of $\mathbf{X}$. This is commonly referred to as the span of the columns of $\mathbf{X}$.

The orthogonal complement of this vector subspace is the **kernel** or **null space** of $\mathbf{H}$, denoted $\ker(\mathbf{H})$. In summary, the space of the n columns of $\mathbf{H}$ can be divided into the following two orthogonal vector subspaces,

i. $\mathbf{y} \in \mathbb{R}^n$,

ii. $\widehat{\mathbf{y}} \in \mathbb{X} := \operatorname{col}(\mathbf{H}) = \operatorname{span}(\mathbf{x}_1, \ldots, \mathbf{x}_p)$,

iii. $\widehat{\mathbf{e}} \in \mathbb{X}^\perp := \ker(\mathbf{H}) = \{\mathbf{v} \in \mathbb{R}^n : \langle \mathbf{v}, \mathbf{x}_j \rangle = 0,\ \forall\, j = 1, \ldots, p^*\}$;

where the vector structure on $\mathbb{R}^n$ can be recovered by taking the *direct sum* of the vector structures on $\mathbb{X}$ and $\mathbb{X}^\perp$. Here, $\mathbb{X}$ and $\mathbb{X}^\perp$ are referred to as vector subspaces or *linear subspaces* of $\mathbb{R}^n$. More specifically, the linear subspace $\mathbb{X}^\perp$ is the **orthogonal complement** of $\mathbb{X}$. Naturally, the fact that $\widehat{\mathbf{y}} \in \mathbb{X}$ and $\widehat{\mathbf{e}} \in \mathbb{X}^\perp$ implies that $\widehat{\mathbf{y}} \perp \widehat{\mathbf{e}}$, which can be checked algebraically, as demonstrated in the next section.

2.2 Residuals and Fitted Values are Orthogonal

Firstly, observe that **left-multiplication** of $\mathbf{I} - \mathbf{H}$ by $\mathbf{H}$ gives an $n \times n$ matrix of zeros,

$$\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{H} - \mathbf{H}\mathbf{H} = \mathbf{H} - \mathbf{H} = \mathbf{0}.$$

Geometrically, the residuals and the fitted values are two orthogonal vectors. Therefore, one can verify that the Euclidean inner product or *dot product* between these two vectors in $\mathbb{R}^n$ is zero.

$$\langle \widehat{\mathbf{y}}, \widehat{\mathbf{e}} \rangle = \widehat{\mathbf{y}}^T \widehat{\mathbf{e}} = (\mathbf{H}\mathbf{y})^T (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}^T \mathbf{H}(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}^T \mathbf{0} \mathbf{y} = 0.$$

Therefore, $\widehat{\mathbf{y}}$ and $\mathbf{e}$ are orthogonal in $\mathbb{R}^n$.

2.3 Residuals and Fitted Values are Uncorrelated

Moreover, one can also show that $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{e}}$ are **elementwise uncorrelated**. We will need to use the fact $\mathbb{C}\mathrm{ov}[\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}] = \mathbf{A}\,\mathbb{C}\mathrm{ov}[\mathbf{x}, \mathbf{y}]\mathbf{B}^T$, for any two random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Thus,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[\widehat{\mathbf{y}}, \widehat{\mathbf{e}}|\mathbf{X}] &= \mathbb{C}\mathrm{ov}[\mathbf{H}\mathbf{y}, (\mathbf{I} - \mathbf{H})\mathbf{y}|\mathbf{X}] \\
&= \mathbf{H}\,\mathbb{C}\mathrm{ov}[\mathbf{y}, \mathbf{y}|\mathbf{X}](\mathbf{I} - \mathbf{H})^T \\
&= \sigma^2 \mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}.
\end{aligned}$$

Here, we are treating $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{e}}$ as random vectors, whereas in the previous section, we only considered two vectors of realizations from these random processes.

3 Geometric Interpretation of Regression

3.1 Degrees of Freedom of $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{e}}$

Geometrically, the degrees of freedom associated to $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{e}}$ can be seen to simply be the dimensions of the respective vector subspaces in which these two vectors have been projected. In particular, for any decomposition of a vector space $\mathbb{V}$ into its **range** and **kernel**, by a simple application of the **rank-nullity** theorem,

$$\begin{array}{ccccc} \dim(\mathbb{R}^n) & = & \dim(\mathbb{X}) & + & \dim(\mathbb{X}^{\perp}) \\ n & = & \operatorname{rank}(\mathbf{H}) & + & \operatorname{nul}(\mathbf{H}), \end{array}$$

and where $\operatorname{nul}(\mathbf{H})$ indicates the **nullity** of $\mathbf{H}$, which is defined as the dimension of the null space of $\mathbf{H}$. When considering a mean function with p^* different parameters, we have the following decomposition of the degrees of freedom,

$$\begin{array}{ccccc} \operatorname{df}(\mathbf{y}) & = & \operatorname{df}(\widehat{\mathbf{y}}) & + & \operatorname{df}(\widehat{\mathbf{e}}) \\ n & = & p^* & + & (n-p^*). \end{array} \tag{1}$$

Since $\mathbf{H}$ was noted to be an orthogonal projection, one can also observe that the eigenvalues of $\mathbf{H}$ determine the dimensions of these vector subspaces. Indeed, the eigenvalues of the matrix of an orthogonal projection can only be 0 or 1. Thus, the number of zeros in the spectrum of $\mathbf{H}$ is equal to the nullity of $\mathbf{H}$, whereas the number of ones in its spectrum is equal to its rank.

3.2 Variance Partitioning Through Pythagoras' Theorem

The vectors $\mathbf{y}$, $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{e}}$ determine three points in $\mathbb{R}^n$, which forms a triangle. Since $\langle \widehat{\mathbf{y}}, \widehat{\mathbf{e}} \rangle = 0$, it follows that this triangle is a **right triangle**, or (right-angled) triangle. Two aspects of regression analysis can be understood using basic trigonometry. Firstly, the decomposition of the total sum of squares (TSS or SYY) into estimated sum of squares (SSreg or ESS) and residual sum of squares (RSS) can be shown to constitute a special case of Pythagoras' theorem. For convenience, *we center the y_i's at $\bar{y}$.* Then, the **total sum of squares** is given by

$$\text{SYY} := \sum_{i=1}^{n} y_i^2 = ||\mathbf{y}||^2,$$

where $||\cdot||$ is the *Euclidean norm.* Using the fact that $y_i = \widehat{y}_i + \widehat{e}_i$, for every $i = 1, \dots, n$, we have

$$\begin{aligned} \sum_{i=1}^{n} (\widehat{y}_i + \widehat{e}_i)^2 &= \sum_{i=1}^{n} \widehat{y}_i^2 + 2 \sum_{i=1}^{n} \widehat{y}_i \widehat{e}_i + \sum_{i=1}^{n} \widehat{e}_i^2 \\ &= ||\widehat{\mathbf{y}}||^2 + 2 \langle \widehat{\mathbf{y}}, \widehat{\mathbf{e}} \rangle + ||\widehat{\mathbf{e}}||^2. \end{aligned}$$

Moreover, we have already shown that $\widehat{\mathbf{y}} \perp \widehat{\mathbf{e}}$. Therefore, $||\mathbf{y}||$ is the **length of the hypotenuse**, whose square is equal to the sum of the squares of the lengths of the sides such that

$$||\mathbf{y}||^2 \quad = \quad ||\widehat{\mathbf{y}}||^2 \quad + \quad ||\widehat{\mathbf{e}}||^2.$$

We will see how to compute the degrees of freedom for these three quantities in the next class. These degrees of freedom may be different from the ones of the random vectors in equation (1), depending on whether or not we have centered our data.

3.3 Coefficient of Determination using Trigonometry

Secondly, we can also apply another consequence of Pythagoras' theorem in order to measure the degree of linear dependency between $\mathbf{y}$ and $\widehat{\mathbf{y}}$. Recall that for any right triangle,

$$\cos\theta = \frac{|\text{adjacent}|}{|\text{hypotenuse}|},$$

where the angle θ is measured in Radians. In regression, we are interested in the angle at $\mathbf{0}$ in $\mathbb{R}^n$, with sides given by the vectors $\mathbf{y}$ and $\widehat{\mathbf{y}}$. Thus, similarly we have

$$(\cos\theta)^2 = \frac{||\widehat{\mathbf{y}}||^2}{||\mathbf{y}||^2} = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{SYY} - \text{RSS}}{\text{SYY}} = R^2.$$

Therefore, maximizing the fit of a regression model is equivalent to decreasing the angle between the vectors $\mathbf{y}$ and $\widehat{\mathbf{y}}$ in $\mathbb{R}^n$. A perfect correlation arises when this angle is zero, such that $\cos(0) = 1$; whereas full statistical independence (i.e. absence of correlation) is equivalent to a *right angle*, $\cos(\pi/2) = 0$. Finally, a negative correlation is obtained when the angle between $\mathbf{y}$ and $\widehat{\mathbf{y}}$ is *straight*, whereby $\cos(\pi) = -1$.

3.4 Collinear Predictors

Moreover, this geometrical perspective on multiple regression also sheds light on the use of the term **collinear** in statistics. Several points are said to be collinear if they are lying on a single line. In the context of multiple regression, each column of $\mathbf{X}$ is a vector in $\mathbb{R}^n$. If the correlation between two such column vectors, say $\mathbf{X}_{\cdot,j}$ and $\mathbf{X}_{\cdot,k}$, is perfect; then they are lying on the same line.