6: The k-Variable Linear Model 1
ECON 837

Brian Krauth (adapted from notes by Simon Woodcock), Spring 2010

Matrix Formulation

Now we turn our attention to the generalization and matrix formulation of the linear model. Suppose that:

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \varepsilon_1 \\ y_2 &= \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \varepsilon_n \end{aligned}$$

where (usually) $x_{11} = x_{12} = \cdots = x_{1n} = 1$, i.e., our model has an intercept. Good practice requires including an intercept unless there is a compelling reason not to. Then we can write

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where, following convention, $\mathbf{y}$ is $n \times 1$, $\mathbf{X}$ is $n \times k$, β is $k \times 1$, and ε is $n \times 1$. Now,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ 1 & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

and a typical row is

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + \varepsilon_i.$$

As we did for simple regression, we will add assumptions one at a time, and derive the least squares estimator and its properties.

Least Squares

Assumption (Linearity) We assume the regression function is linear, that is $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$. Note this is equivalent to $E[\varepsilon|\mathbf{X}] = \mathbf{0}$, and implies $E[\varepsilon] = \mathbf{0}$ by the law of iterated expectations. If $\mathbf{X}$ is nonstochastic, then we can write this assumption more simply as $E[\mathbf{y}] = \mathbf{X}\beta$ and $E[\varepsilon] = 0$.

For simplicity of exposition, we'll assume the nonstochastic case. Keep in mind that we can allow stochastic $\mathbf{X}$ without any great complication as long as $E[\varepsilon|\mathbf{X}] = \mathbf{0}$.

Just like the simple regression case, the least squares estimator is the parameter vector that minimizes the sum of squares function. That is,

$$\begin{aligned} \hat{\beta} &= \arg\min_{\mathbf{b}} S(\mathbf{b}) \\ \text{where } S(\mathbf{b}) &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}. \end{aligned}$$

The first order condition for a minimum yields the **least squares normal equations** that define $\hat{\beta}$:

$$\mathbf{X}'\mathbf{X}\hat{\beta} - \mathbf{X}'\mathbf{y} = \mathbf{0}.$$

These equations **always** have at least one solution. Whether there is a unique solution depends on $\mathbf{X}'\mathbf{X}$.

Proposition 1 *If $\mathbf{X}'\mathbf{X}$ is invertible, the solution to the normal equations is unique and*

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

Proof. This is obvious. ■

Proposition 2 *$\hat{\beta}$ minimizes $S(\mathbf{b})$.*

Proof. Let $\mathbf{b}$ be any other k-vector. Since $\mathbf{b} = \mathbf{b} + \hat{\beta} - \hat{\beta}$, we have

$$\begin{aligned}
(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) &= \left(\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\left(\hat{\beta} - \mathbf{b}\right)\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\left(\hat{\beta} - \mathbf{b}\right)\right) \\
&= \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) + 2\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\mathbf{X}\left(\hat{\beta} - \mathbf{b}\right) + \left(\hat{\beta} - \mathbf{b}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta} - \mathbf{b}\right) \\
&= \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) + \left(\hat{\beta} - \mathbf{b}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta} - \mathbf{b}\right) \\
&\geq \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)
\end{aligned}$$

since $\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\mathbf{X} = \mathbf{0}$ by the least squares normal equations and $\left(\hat{\beta} - \mathbf{b}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta} - \mathbf{b}\right) \geq 0$ since this is a quadratic form and $\mathbf{X}'\mathbf{X}$ is psd.

Less elegantly, we could just check the second order conditions:

$$\frac{\partial^2 S}{\partial \mathbf{b}\partial \mathbf{b}'} = 2\mathbf{X}'\mathbf{X} \text{ psd.}$$

■

Definition 3 *$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$ is the $n \times 1$ vector of least squares residuals.*

Note that as a simple matter of algebra, $\mathbf{X}'\mathbf{e} = \mathbf{0}$ by the least squares normal equations. Assuming the first column in $\mathbf{X}$ is an intercept, this means that the residuals average out to exactly zero.

Statistically, $E[\mathbf{e}] = E\left[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = \mathbf{X}\beta - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \mathbf{0}$,.

We can write:

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon
\end{aligned}$$

This expresses the least squares estimator as its estimand β plus some data-dependent error.

Proposition 4 *The least squares estimator $\hat{\beta}$ is unbiased.*

Proof. $E\left[\hat{\beta}\right] = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon)\right] = \beta$. ■

To derive more properties of the least squares estimator we need to make additional assumptions. Before doing so, let's have a look at the geometry of least squares.

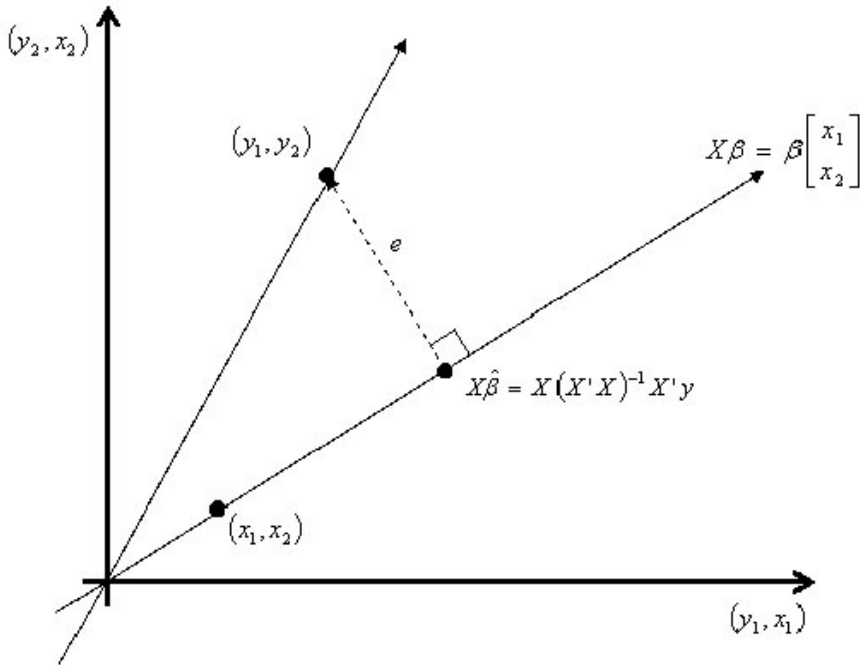

Figure 1: The Geometry of Least Squares

Geometry of Least Squares

We'll keep things simple and consider the case of a single regressor, no intercept, and two observations. That is, we'll consider $\mathbf{y} = \mathbf{x}\beta + \varepsilon$ where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \ \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The least squares estimator $\hat{\beta}$ **projects** $\mathbf{y}$ onto the **column space** of $\mathbf{x}$. What does this look like? See Figure 1.

We'll discuss the figure informally, then introduce some formal concepts.

In Figure 1, we've plotted the data $\mathbf{y}$ and $\mathbf{x}$ as two coordinate points: one for $\mathbf{y}$ and one for $\mathbf{x}$. These points can be represented in two dimensions because we have only two observations. The line $\mathbf{x}\beta$ is the called the **space spanned by** $\mathbf{x}$ (or the column space of $\mathbf{x}$). Since $\mathbf{x}$ is a vector, it is one-dimensional and represents all real re-scalings of $\mathbf{x}$. I've plotted a similar line through $\mathbf{y}$, which represents the space spanned by $\mathbf{y}$. The goal of least squares is to choose a particular number $\hat{\beta}$ that minimizes the (squared) distance between the space spanned by $\mathbf{x}$ and the point $\mathbf{y}$. This distance is given by the length of the residual vector $\mathbf{e}$, and as is minimized when $\mathbf{e}$ is orthogonal (perpendicular) to the space spanned by $\mathbf{x}$. Generalizing the geometry of least squares to more than one regressor and more than two observations is straightforward, but more difficult to represent graphically.

Definition 5 *The* ***space spanned by the matrix*** $\mathbf{X}$ *is the vector space that consists of*

all linear combinations of the column vectors of $\mathbf{X}$. *Sometimes this is called the column space of* $\mathbf{X}$ *or the range space of* $\mathbf{X}$, $R(\mathbf{X})$. *We can regard* $R(\mathbf{X})$ *as an operator* $R : \mathbb{R}^k \to \mathbb{R}^n = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{X}\lambda = \mathbf{x}\ \forall \lambda \in \mathbb{R}^k\right\}$.

Definition 6 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ *is the orthogonal projection of* $\mathbf{y}$ *onto the space spanned by* $\mathbf{X}$ *(that is, onto* $R(\mathbf{X})$*).*

So, as in the case of one regressor and two observations, as β varies, $\mathbf{X}\beta$ is the space spanned by the columns of $\mathbf{X}$. That is, as β varies, $\mathbf{X}\beta$ represents the set of all linear combinations of the column vectors of $\mathbf{X}$. The least squares estimator $\hat{\beta}$ minimizes the distance between $R(\mathbf{X})$ and $\mathbf{y}$ because $\mathbf{X}\hat{\beta}$ is the orthogonal projection of $\mathbf{y}$ onto $R(\mathbf{X})$. That is, $\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Definition 7 *The linear subspace of* $\mathbb{R}^n$ *orthogonal to the range space of* $\mathbf{X}$ *is the* ***null space of*** $\mathbf{X}'$, $N(\mathbf{X}') = \{\mathbf{a} \in \mathbb{R}^n : \mathbf{X}'\mathbf{a} = \mathbf{0}\}$.

Proposition 8 $\mathbf{e}$ *is orthogonal to* $\mathbf{X}$, *i.e.,* $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

Proof. We already know this from the least squares normal equations, but let's show it directly.

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathbf{X}\hat{\beta} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{y} \\ \mathbf{X}'\mathbf{e} &= (\mathbf{X}' - \mathbf{X}')\mathbf{y} = \mathbf{0}\end{aligned}$$

■

Thus the equation $\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{e}$ gives $\mathbf{y}$ as the sum of a vector in $R(\mathbf{X})$ and a vector in $N(\mathbf{X}')$.

Common Projection Matrices

1. The matrix that projects onto the space orthogonal to the space spanned by the columns of $\mathbf{X}$ (i.e., onto $N(\mathbf{X}')$) is

   $$\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

   Note that $\mathbf{e} = \mathbf{M}\mathbf{y}$. If $\mathbf{X}$ has full column rank then $\mathbf{M}$ has rank $n - k$ (that is, it projects onto an $n - k$ dimensional subspace of $\mathbb{R}^n$). Sometimes we informally call $\mathbf{M}$ the **residual maker**.

2. The matrix that projects to the space spanned by the columns of $\mathbf{X}$ (i.e., onto $R(\mathbf{X})$) is

   $$\mathbf{I}_n - \mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

   Note that $\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = \mathbf{y} - \mathbf{M}\mathbf{y} = (\mathbf{I}_n - \mathbf{M})\mathbf{y}$. If $\mathbf{X}$ has full column rank, then $\mathbf{I}_n - \mathbf{M}$ has rank k (that is, it projects onto a k dimensional subspace of $\mathbb{R}^n$).

3. The matrix that "sweeps out" means is

$$\begin{aligned}\mathbf{J} &= \mathbf{I}_n - \mathbf{i}\left(\mathbf{i}'\mathbf{i}\right)^{-1}\mathbf{i}' = \mathbf{I}_n - \mathbf{i}\mathbf{i}'n^{-1} \\ &= \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & & \vdots \\ \vdots & & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \cdots & -\frac{1}{n} & 1-\frac{1}{n} \end{bmatrix}\end{aligned}$$

where $\mathbf{i}$ is an n-vector of ones. That is, $\mathbf{J}\mathbf{x} = \mathbf{x} - \bar{x}\mathbf{i}$ for any vector $\mathbf{x}$.

Properties of Projection Matrices

1. Projection matrices are **idempotent**. In other words, $\mathbf{M}\mathbf{M} = \mathbf{M}$ and $\left(\mathbf{I}_n - \mathbf{M}\right)\left(\mathbf{I}_n - \mathbf{M}\right) = \mathbf{I}_n - \mathbf{M}$. Let's check the latter.

Proof. $\left(\mathbf{I}_n - \mathbf{M}\right)\left(\mathbf{I}_n - \mathbf{M}\right) = \left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \mathbf{I}_n - \mathbf{M}$.
The intuition here is pretty simple. Consider $\mathbf{M}$. It projects from $\mathbb{R}^n$ onto an $n-k$ dimensional subspace of $\mathbb{R}^n$. Repeating the projection doesn't change the space onto which you project.
Note that $\mathbf{M}\left(\mathbf{I}_n - \mathbf{M}\right) = \mathbf{0}$ since $\mathbf{M}$ and $\mathbf{I}_n - \mathbf{M}$ project onto orthogonal subspaces. ■

2. Idempotent matrices have eigenvalues equal to zero or one. Note this implies that idempotent matrices are psd.

Proof. Consider the characteristic equation of $\mathbf{M}$:

$$\mathbf{M}\mathbf{c} = \lambda\mathbf{c}.$$

Now consider the characteristic equation of $\mathbf{M}\mathbf{M}$:

$$\mathbf{M}\mathbf{M}\mathbf{c} = \lambda^2\mathbf{c}.$$

But since $\mathbf{M}$ is idempotent, we know that $\mathbf{M}\mathbf{M}\mathbf{c} = \mathbf{M}\mathbf{c}$, so $\lambda^2\mathbf{c} = \lambda\mathbf{c}$, which implies $\lambda = 0$ or $\lambda = 1$. ■

3. For idempotent matrices, trace = rank.

The projection matrices we will see are symmetric, but note that there are idempotent projection matrices that are not symmetric, e.g., $\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}$.

Sampling Variance of $\hat{\beta}$ and the GMT

Now we add an assumption to our k-variable linear model so that we can obtain the sampling variance of $\hat{\beta}$.

Assumption (Spherical Errors) $Var[\mathbf{y}] = Var[\varepsilon] = \sigma^2\mathbf{I}_n$.

Proposition 9 $Var\left[\hat{\beta}\right] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Proof.

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta+\varepsilon) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
Var\left[\hat{\beta}\right] &= E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'\right] \\
&= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\varepsilon\varepsilon']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

■

A basic optimality property of the least squares estimator is given by the Gauss-Markov Theorem (GMT). It's not a very strong property, but it's all that is available without further assumptions.

Theorem 10 (Gauss-Markov) *The least squares estimator is BLUE.*

Proof. Consider estimating any linear combination of the coefficients, $\mathbf{c}'\beta$. A possible estimator is $\mathbf{c}'\hat{\beta}$ with variance $\sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$.
Consider any alternative linear unbiased estimator $\mathbf{b} = \mathbf{a}'\mathbf{y}$. Then $E[\mathbf{b}] = \mathbf{a}'E[\mathbf{y}] = \mathbf{a}'\mathbf{X}\beta$. Since $\mathbf{b}$ is an unbiased estimator of $\mathbf{c}'\beta$, $\mathbf{a}'\mathbf{X} = \mathbf{c}'$. Thus,

$$\begin{aligned}
\mathbf{b} &= \mathbf{a}'\mathbf{y} = \mathbf{a}'(\mathbf{X}\beta+\varepsilon) = \mathbf{a}'\mathbf{X}\beta + \mathbf{a}'\varepsilon = \mathbf{c}'\beta + \mathbf{a}'\varepsilon \\
Var[\mathbf{b}] &= \sigma^2\mathbf{a}'\mathbf{a}.
\end{aligned}$$

Now, $Var\left[\mathbf{c}'\hat{\beta}\right] = \sigma^2\mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{a}$ since $\mathbf{c}' = \mathbf{a}'\mathbf{X}$. Therefore,

$$Var[\mathbf{b}] - Var\left[\mathbf{c}'\hat{\beta}\right] = \sigma^2\mathbf{a}'\mathbf{M}\mathbf{a}$$

which is psd because $\mathbf{M}$ is idempotent. ■

Estimating σ^2

Without further assumptions, we can't hope for too much. However, we can derive an unbiased estimator of σ^2.

Proposition 11 $s^2 = \mathbf{e}'\mathbf{e}/(n-k)$ *is an unbiased estimator of* σ^2.

Proof.

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{X}\beta + \mathbf{M}\varepsilon = \mathbf{M}\varepsilon$$

because $\mathbf{MX} = \mathbf{0}$ (it projects onto the space orthogonal to the columns of $\mathbf{X}$). Therefore

$$\begin{aligned}
\mathbf{e}'\mathbf{e} &= \varepsilon'\mathbf{M}\varepsilon \\
E[\mathbf{e}'\mathbf{e}] &= E[\varepsilon'\mathbf{M}\varepsilon] \\
&= E[tr(\varepsilon'\mathbf{M}\varepsilon)] \qquad (\varepsilon' M \varepsilon \text{ is a scalar}) \qquad (1)\\
&= E[tr(\mathbf{M}\varepsilon\varepsilon')] \qquad (\text{cyclic permutation}) \qquad (2)\\
&= tr(E[\mathbf{M}\varepsilon\varepsilon']) \qquad (\text{trace is a linear operator}) \qquad (3)\\
&= tr(\mathbf{M}E[\varepsilon\varepsilon']) \\
&= \sigma^2 tr(\mathbf{M}) \\
&= \sigma^2(n-k) \qquad (\text{trace} = \text{rank}) \qquad (4)
\end{aligned}$$

■

Note this proof uses the fact that $\mathbf{M}$ is idempotent and that $tr(\mathbf{ABC}) = tr(\mathbf{CAB}) = tr(\mathbf{BCA})$ when the matrices are conformable.

Fit: Does the Regression Model Explain the Data?

Recall the idempotent matrix $\mathbf{J} = \mathbf{I}_n - \mathbf{ii}'/n$ that sweeps out means. Note that $\mathbf{JM} = \mathbf{M}$ when $\mathbf{X}$ contains a constant term (show this yourself). Intuition: $\mathbf{M}$ projects onto an $n-k$ dimensional subspace of $\mathbb{R}^n$, namely $N(\mathbf{X}')$; and $\mathbf{J}$ projects onto an $n-1$ dimensional subspace of $\mathbb{R}^n$, namely $N(\mathbf{i}')$. Since $\mathbf{X}$ contains $\mathbf{i}$, we have $R(\mathbf{i}) \subset R(\mathbf{X})$, and hence $N(\mathbf{X}') \subset N(\mathbf{i}')$. That is, $\mathbf{M}$ projects onto a subspace of the space onto which $\mathbf{J}$ projects.

Definition 12 *The* ***squared correlation coefficient*** *in the k-variable case is*

$$R^2 = \frac{\textit{sum of squares explained by } \mathbf{X}}{\textit{total sum of squares}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{J}\mathbf{y}}.$$

Since $\mathbf{J}$ is idempotent and symmetric, $\mathbf{y}'\mathbf{J}\mathbf{y} = (\mathbf{J}\mathbf{y})'\mathbf{J}\mathbf{y} = \sum_i (y_i - \bar{y})^2$. Furthermore,

$$\begin{aligned}
\mathbf{y}'\mathbf{J}\mathbf{y} &= (\mathbf{J}\mathbf{y})'\mathbf{J}\mathbf{y} = (\mathbf{J}\hat{\mathbf{y}} + \mathbf{J}\mathbf{e})'(\mathbf{J}\hat{\mathbf{y}} + \mathbf{J}\mathbf{e}) \\
&= \hat{\mathbf{y}}'\mathbf{J}\hat{\mathbf{y}} + 2\hat{\mathbf{y}}'\mathbf{J}\mathbf{e} + \mathbf{e}'\mathbf{J}\mathbf{e} \\
&= \hat{\mathbf{y}}'\mathbf{J}\hat{\mathbf{y}} + 2\mathbf{y}'(\mathbf{I}_n - \mathbf{M})\mathbf{J}\mathbf{M}\mathbf{y} + \mathbf{e}'\mathbf{J}\mathbf{e} \\
&= \hat{\mathbf{y}}'\mathbf{J}\hat{\mathbf{y}} + 2\mathbf{y}'(\mathbf{I}_n - \mathbf{M})\mathbf{M}\mathbf{y} + \mathbf{e}'\mathbf{J}\mathbf{e} \\
&= \hat{\mathbf{y}}'\mathbf{J}\hat{\mathbf{y}} + \mathbf{e}'\mathbf{J}\mathbf{e} \\
&= \hat{\mathbf{y}}'\mathbf{J}\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}
\end{aligned}$$

because $(\mathbf{I}_n - \mathbf{M})\mathbf{M} = \mathbf{0}$ and $\mathbf{e}'\mathbf{J}\mathbf{e} = \mathbf{e}'\mathbf{e}$ (why?). This expression says

total sum of squares = sum of squares explained by $\mathbf{X}$ + unexplained sum of squares

Dividing both sides by $\mathbf{y}'\mathbf{J}\mathbf{y}$ gives

$$1 = \frac{\hat{\mathbf{y}}'\mathbf{J}\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{J}\mathbf{y}} + \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{J}\mathbf{y}}$$

that is, the proportion of variation due to $\mathbf{X}$ and the proportion of variation due to error. Since R^2 gives the proportion of variation explained by $\mathbf{X}$, it is $R^2 = 1 - \mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{J}\mathbf{y}$. Note also that is the squared correlation coefficient between $\mathbf{y}$ and $\hat{\mathbf{y}}$.

Sometimes we prefer an alternate measure of model fit that "penalizes" models with many regressors. This is the **adjusted squared correlation coefficient**:

$$\bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-k)}{\mathbf{y}'\mathbf{J}\mathbf{y}/(n-1)}.$$

We might prefer this because adding an additional regressor never decreases R^2. This is not the case for $\bar{R}^2$. What is the numerator in the fraction? What is the denominator?

A Note on Reporting Regression Output

Some good practice:

1. Always report the characteristics of the sample (sample means & standard deviations, observation counts, anything unusual or surprising, population from which the sample was collected, how it was collected (unless well-known), and how the estimation sample was selected).
2. Always report $\hat{\beta}$ and standard errors (**not** t-statistics). The usual format is

$$\begin{matrix} \hat{\beta} \\ \Big(\text{s.e. of } \hat{\beta}\Big) \end{matrix}$$

3. Specify s^2 or σ^2_{ML}.
4. Report n and R^2 (and/or $\bar{R}^2$).
5. Plots are important. It is good practice to plot the density of $\mathbf{y}$ and $\mathbf{e}$, as well as plots of predicted vs. actual values, or actual values over time in a time series setting.