Lecture 3

Linear Least Squares

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

The linear model is the main technique in regression problems and the primary tool for it is least squares fitting. We minimize a sum of squared errors, or equivalently the sample average of squared errors. That is a natural choice when we're interested in finding the regression function which minimizes the corresponding expected squared error.

This chapter presents the basic theory of linear least squares estimation, looking at it with calculus, linear algebra and geometry. It also develops some distribution theory for linear least squares and computational aspects of linear regression. We will draw repeatedly on the material here in later chapters that look at specific data analysis problems.

In this chapter $X_{\bullet j}$ denotes the j'th column of X.

3.1 Least squares estimates

We begin with observed response values y_1, y_2, ..., y_n and features x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We're looking for the parameter values

$\beta_1, \ldots, \beta_p$ that minimize

$$S(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2. \tag{3.1}$$

The minimizing values are denoted with hats: $\hat{\beta}_1, \ldots, \hat{\beta}_p$.

To minimize S, we set its partial derivative with respect to each β_j to 0. The solution satisfies

$$\frac{\partial}{\partial \beta_j} S = -2 \sum_{i=1}^{n} \Big(y_i - \sum_{k=1}^{p} x_{ik}\hat{\beta}_k \Big) x_{ij} = 0, \quad j = 1, \ldots, p. \tag{3.2}$$

The solution must be a minimum because S is convex.

The p equations in (3.2) are known as the ***normal equations***. This is due to normal being a synonym for perpendicular or orthogonal, and not due to any assumption about the normal distribution. Consider the vector $X_{\bullet j} = (x_{1j}, \ldots, x_{nj})^{\mathsf{T}} \in \mathbb{R}^n$ of values for the j'th feature. Equation (3.2) says that this feature vector has a dot product of zero with the residual vector having i'th element $\hat{\varepsilon}_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$. Each feature vector is orthogonal (normal) to the vector of residuals.

It is worthwhile writing equations (3.1) and (3.2) in matrix notation. Though the transition from (3.1) to (3.2) is simpler with coordinates, our later manipulations are easier in vector form. We pack the responses and feature values into the vector Y and matrix X via

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

We also put $\beta = (\beta_1, \ldots, \beta_p)^{\mathsf{T}}$ and let $\hat{\beta}$ denote the minimizer of the sum of squares. Finally let $\hat{\boldsymbol{\varepsilon}} = Y - X\hat{\beta}$, the n vector of residuals.

Now equation (3.1) can be written $S(\beta) = (Y - X\beta)^{\mathsf{T}}(Y - X\beta)$ and the normal equations (3.2) become

$$\hat{\boldsymbol{\varepsilon}}^{\mathsf{T}} X = \mathbf{0} \tag{3.3}$$

after dividing by -2. These normal equations can also be written

$$X^{\mathsf{T}} X \hat{\beta} = X^{\mathsf{T}} Y. \tag{3.4}$$

If $X^\mathsf{T}X$ is invertible, then

$$\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y. \tag{3.5}$$

We'll suppose at first that $X^\mathsf{T}X$ is invertible and then consider the singular case in Section 3.9. This should remind you of $\hat{\beta}_1 = \mathrm{SXY}/\mathrm{SXX}$ for the slope of a plain linear regression from a first course in regression.

Equation (3.5) underlies a second meaning of the word 'linear' in linear regression. The estimated coefficient $\hat{\beta}$ is a fixed linear combination of Y, meaning that we get it by multiplying Y by the matrix $(X^\mathsf{T}X)^{-1}X^\mathsf{T}$. The predicted value of Y at any new point $\boldsymbol{x}_0$ is also linear in Y; it is $\boldsymbol{x}_0^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$.

3.2 Geometry of least squares

In class I made a sketch to illustrate linear least squares. The vector $Y = (y_1, \ldots, y_n)^\mathsf{T}$ is represented as a point in $\mathbb{R}^n$. The set

$$\mathcal{M} = \{X\beta \mid \beta \in \mathbb{R}^p\} \tag{3.6}$$

is a p dimensional subspace of $\mathbb{R}^n$. It is fully p dimensional here because we have assumed that $X^\mathsf{T}X$ is invertible and so X has rank p. Under our model, $\mathbb{E}(Y) = X\beta \in \mathcal{M}$ though we do not ordinarily have $Y \in \mathcal{M}$.

The idea behind least squares is to find $\hat{\beta}$ so that $\hat{Y} = X\hat{\beta}$ is the closest point to Y from within $\mathcal{M}$. We expect to find this closest point to Y by "dropping a perpendicular line" to $\mathcal{M}$. That is, the error $\hat{\boldsymbol{\varepsilon}} = Y - \hat{Y}$ should be perpendicular to any line within $\mathcal{M}$. From the normal equations (3.3), we have $\hat{\boldsymbol{\varepsilon}}^\mathsf{T}X = 0$ so $\hat{\boldsymbol{\varepsilon}}$ is orthogonal to every point $X\beta \in \mathcal{M}$.

We can form a right angle triangle using the three points Y, $\hat{Y}$ and $X\widetilde{\beta}$ for any $\widetilde{\beta} \in \mathbb{R}^p$. Taking $\widetilde{\beta} = 0$ and using Pythagoras' theorem we get

$$\|Y\|^2 = \|\hat{\boldsymbol{\varepsilon}}\|^2 + \|X\hat{\beta}\|^2$$

for a hypotenuse Y.

When the first column of X consists of 1s then $(\bar{y}, \ldots, \bar{y})^\mathsf{T} = X(\bar{y}, 0, \ldots, 0) \in \mathcal{M}$. We can form a right angle triangle with corners Y, $\hat{Y}$ and $\bar{y}\mathbf{1}$. Using a hypotenuse $Y - \bar{y}\mathbf{1}$ (the side opposite $\hat{Y}$) we get

$$\|Y - \bar{y}\mathbf{1}\|^2 = \|\hat{Y} - \bar{y}\mathbf{1}\|^2 + \|Y - \hat{Y}\|^2.$$

Writing this out we get

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2. \tag{3.7}$$

Equation (3.7) is known as an ANOVA (analysis of variance) decomposition. We can write it as $\mathrm{SS}_{\mathrm{TOT}}=\mathrm{SS}_{\mathrm{FIT}}+\mathrm{SS}_{\mathrm{RES}}$ splitting a total sum of squares of centered y_i values as a sum of squared centered fitted values plus a sum of squared residuals. Then the ratio

$$R^2=\frac{\mathrm{SS}_{\mathrm{FIT}}}{\mathrm{SS}_{\mathrm{TOT}}}=1-\frac{\mathrm{SS}_{\mathrm{RES}}}{\mathrm{SS}_{\mathrm{TOT}}}$$

is interpreted as the fraction of variation in y_i explained by the model.

3.3 Algebra of least squares

The predicted value for y_i, using the least squares estimates, is $\hat{y}_i=\boldsymbol{x}_i^{\mathsf{T}}\hat{\beta}$. We can write the whole vector of fitted values as $\hat{Y}=X\hat{\beta}=X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y$. That is $\hat{Y}=HY$ where

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

Tukey coined the term "hat matrix" for H because it puts the hat on Y. Some simple properties of the hat matrix are important in interpreting least squares. In several places we use

$$HX=X \tag{3.8}$$

which follows immediately from the definition of H.

By writing $H^2=HH$ out fully and cancelling we find $H^2=H$. A matrix H with $H^2=H$ is called idempotent. In hindsight, it is geometrically obvious that we should have had $H^2=H$. For any $Y\in\mathbb{R}^n$ the closest point to Y inside of $\mathcal{M}$ is HY. Since HY is already in $\mathcal{M}$, $H(HY)=HY$. That is $H^2Y=HY$ for any Y and so $H^2=H$. Any column of X is in $\mathcal{M}$ so $HX_{\bullet j}=X_{\bullet j}$ providing another explanation of equation (3.8). Clearly $H^k=H$ for any integer $k\geqslant 1$.

The matrix $X^{\mathsf{T}}X$ is symmetric, and so therefore is $(X^{\mathsf{T}}X)^{-1}$. It follows that H is symmetric too. A symmetric idempotent matrix such as H is called a perpendicular projection matrix (PPM).

Theorem 1. *Let H be a symmetric idempotent real valued matrix. Then the eigenvalues of H are all either 0 or 1.*

Proof. Suppose that $\boldsymbol{v}$ is an eigenvector of H with eigenvalue λ, so $H\boldsymbol{v} = \lambda\boldsymbol{v}$. Because H is idempotent $H^2\boldsymbol{v} = H\boldsymbol{v} = \lambda\boldsymbol{v}$. But we also have $H^2\boldsymbol{v} = H(H\boldsymbol{v}) = H(\lambda\boldsymbol{v}) = \lambda^2\boldsymbol{v}$. Therefore $\lambda\boldsymbol{v} = \lambda^2\boldsymbol{v}$. The definition of eigenvector does not allow $\boldsymbol{v} = \boldsymbol{0}$ and so we must have $\lambda^2 = \lambda$, so $\lambda \in \{0, 1\}$. □

Let $H = P\Lambda P^{\mathsf{T}}$ where the columns of P are eigenvectors p_i of H for $i = 1, \ldots, n$. Then $H = \sum_{i=1}^{n} \lambda_i p_i p_i^{\mathsf{T}}$, where by Theorem 1 each λ_i is 0 or 1. With no loss of generality, we can arrange for the ones to precede the zeros. Suppose that there are r ones. Then $H = \sum_{i=1}^{r} p_i p_i^{\mathsf{T}}$. The eigenvalues of H sum to r, so $\mathrm{tr}(H) = r$. Also

$$\mathrm{tr}(H) = \mathrm{tr}(X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}) = \mathrm{tr}(X^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}) = \mathrm{tr}(I_p) = p.$$

Therefore $r = p$ and $H = \sum_{i=1}^{p} p_i p_i^{\mathsf{T}}$ where $p_i \in \mathbb{R}^n$ are mutually orthogonal.

The prediction for observation i can be written as $\hat{y}_i = H_{i\bullet}Y$ where $H_{i\bullet}$ is the i'th row of the hat matrix. The i'th row of H is simply $\boldsymbol{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$ and the ij element of the hat matrix is $H_{ij} = \boldsymbol{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\boldsymbol{x}_j$. Because $H_{ij} = H_{ji}$ the contribution of y_i to $\hat{y}_j$ equals that of y_j to $\hat{y}_i$. The diagonal elements of the hat matrix will prove to be very important. They are

$$H_{ii} = \boldsymbol{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\boldsymbol{x}_i. \tag{3.9}$$

We are also interested in the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$. The entire vector of residuals may be written $\hat{\boldsymbol{\varepsilon}} = Y - \hat{Y} = (I - H)Y$. It is easy to see that if H is a PPM, then so is $I - H$. Symmetry is trivial, and $(I - H)(I - H) = I - H - H + HH = I - H$.

The model space $\mathcal{M} = \{X\beta \mid \beta \in \mathbb{R}^p\}$ is the set of linear combinations of columns of X. A typical entry of $\mathcal{M}$ is $\sum_{j=1}^{p} \beta_j X_{\bullet j}$. Thus $\mathcal{M}$ is what is known as the column space of X, denoted $\mathrm{col}(X)$. The hat matrix H projects vectors onto $\mathrm{col}(X)$. The set of vectors $\boldsymbol{v}$ orthogonal to X, that is with $X^{\mathsf{T}}\boldsymbol{v} = \boldsymbol{0}$, is a linear subspace of $\mathbb{R}^n$, known as the null space of X^{T}. We may write it as $\mathcal{M}^{\perp}$ or as $\mathrm{null}(X^{\mathsf{T}})$. The column space and null spaces are orthogonal complements. Any $\boldsymbol{v} \in \mathbb{R}^n$ can be uniquely written as $\boldsymbol{v}_1 + \boldsymbol{v}_2$ where $\boldsymbol{v}_1 \in \mathcal{M}$ and $\boldsymbol{v}_2 \in \mathcal{M}^{\perp}$. In terms of the hat matrix, $\boldsymbol{v}_1 = H\boldsymbol{v}$ and $\boldsymbol{v}_2 = (I - H)\boldsymbol{v}$.

3.4 Distributional results

There are two main models for our $(\boldsymbol{x}_i, y_i)$ data named by David Freedman. In the ***regression model*** the $\boldsymbol{x}_i$ are fixed and then $y_i \overset{\text{ind}}{\sim} p(y_i \,|\, \boldsymbol{x}_i)$. In the

correlation model, $(\boldsymbol{x}_i, y_i) \overset{\text{iid}}{\sim} p(\boldsymbol{x}, y)$ and then X is the observed value of a random matrix $\mathcal{X} \in \mathbb{R}^{n\times p}$. Sometimes we really do choose $\boldsymbol{x}_i$ and observe y_i. Other times they are jointly observed.

Even when they are jointly observed we usually work conditionally on the observed values of $\boldsymbol{x}_i$, treating them as fixed. Consider a process that generates X and then generates Y given X. Seeing X probably doesn't tell us much or anything about β and σ. Those are in the conditional distribution of Y given X. A probability model for the data is

$$p(X, Y; \theta) = p(X; \theta_1) \times p(Y \,|\, X; \theta_2)$$

where $\theta = (\theta_1^{\mathsf{T}}, \theta_2^{\mathsf{T}})^{\mathsf{T}}$. The X values might tell us about θ_1 but unless there is a ***known connection*** relating θ_1 to $\theta_2 = (\beta^{\mathsf{T}}, \sigma^2)^{\mathsf{T}}$ learning about θ_1 gives no information about θ_2.

Let X be fixed and of full rank, and assume that $Y = X\beta + \boldsymbol{\varepsilon}$ with $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$. Then

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}(X\beta + \boldsymbol{\varepsilon}) = \beta + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{\varepsilon}. \quad (3.10)$$

The only randomness in the right side of (3.10) comes from $\boldsymbol{\varepsilon}$. This makes it easy to work out the mean and variance of $\hat{\beta}$ in the fixed X case.

Lemma 1. *If $Y = X\beta + \boldsymbol{\varepsilon}$ where X is nonrandom and $X^{\mathsf{T}}X$ is invertible and $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$, then*

$$\mathbb{E}(\hat{\beta}) = \beta. \quad (3.11)$$

Proof. $\mathbb{E}(\hat{\beta}) = \beta + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbb{E}(\boldsymbol{\varepsilon}) = \beta.$ □

The least squares estimates $\hat{\beta}$ are unbiased for β as long as $\boldsymbol{\varepsilon}$ has mean zero. Lemma 1 does not require normally distributed errors. It does not even make any assumptions about $\mathrm{var}(\boldsymbol{\varepsilon})$. To study the variance of $\hat{\beta}$ we will need assumptions on $\mathrm{var}(\boldsymbol{\varepsilon})$, but not on its mean.

Lemma 2. *If $Y = X\beta + \boldsymbol{\varepsilon}$ where X is nonrandom and $X^{\mathsf{T}}X$ is invertible and $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 I$, for $0 \leqslant \sigma^2 < \infty$, then*

$$\mathrm{var}(\hat{\beta}) = (X^{\mathsf{T}}X)^{-1}\sigma^2. \quad (3.12)$$

Proof. Using equation (3.10),

$$\begin{aligned}\mathrm{var}(\hat{\beta}) &= (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathrm{var}(\boldsymbol{\varepsilon})X(X^{\mathsf{T}}X)^{-1}\\ &= (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\big(\sigma^2 I\big)X(X^{\mathsf{T}}X)^{-1}\\ &= (X^{\mathsf{T}}X)^{-1}\sigma^2,\end{aligned}$$

after a particularly nice cancellation. □

We will look at and interpret equation (3.12) for many specific linear models. For now we notice that it holds without regard to whether $\boldsymbol{\varepsilon}$ is normally distributed or even whether it has mean $\mathbf{0}$.

Now let's make a brief remark on the correlation model. We assume that the random X always has rank p. Then under the correlation model

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}(\mathbb{E}(\hat{\beta}\,|\,X)) = \mathbb{E}(\beta) = \beta.$$

Also

$$\mathrm{var}(\hat{\beta}) = \mathbb{E}(\mathrm{var}(\hat{\beta}\,|\,X)) + \mathrm{var}(\mathbb{E}(\hat{\beta}\,|\,X)) = \mathbb{E}((X^{\mathsf{T}}X)^{-1})\sigma^2.$$

I would rather work with actual $(X^{\mathsf{T}}X)^{-1}\sigma^2$ than the expected value of $(X^{\mathsf{T}}X)^{-1}\sigma^2$ accounting for different X that I might have gotten (but didn't). So let's treat $\boldsymbol{x}_i$ as fixed even if they weren't. We might prefer to make them random later when we use the bootstrap and cross-validation.

Up to now we have studied the mean and variance of the estimate of β. Next we turn our attention to σ^2. The key to estimating σ^2 is the residual vector $\hat{\boldsymbol{\varepsilon}}$.

Lemma 3. *Under the conditions of Lemma 2,* $\mathbb{E}(\hat{\boldsymbol{\varepsilon}}^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}) = (n-p)\sigma^2$. *For* $p < n$ *the estimate*

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^{\mathsf{T}}\hat{\beta})^2 \tag{3.13}$$

satisfies $\mathbb{E}(s^2) = \sigma^2$.

Proof. Using (3.8), $\hat{\boldsymbol{\varepsilon}} = (I-H)(X\beta + \boldsymbol{\varepsilon}) = (I-H)\boldsymbol{\varepsilon}$. Therefore $\mathbb{E}(\hat{\boldsymbol{\varepsilon}}^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}) = \mathbb{E}(\boldsymbol{\varepsilon}^{\mathsf{T}}(I-H)\boldsymbol{\varepsilon}) = \mathrm{tr}((I-H)I\sigma^2) = (n-p)\sigma^2$. Finally $s^2 = \hat{\boldsymbol{\varepsilon}}^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}/(n-p)$. □

Now we add in the strong assumption that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$. Normally distributed errors make for normally distributed least squares estimates, fits and residuals.

Theorem 2. *Let* $X \in \mathbb{R}^{n\times p}$ *with* $X^{\mathsf{T}}X$ *invertible, and* $Y = X\beta + \boldsymbol{\varepsilon}$ *where* $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$. *Then*

$$\begin{aligned}
\hat{\beta} &\sim \mathcal{N}(\beta, (X^{\mathsf{T}}X)^{-1}\sigma^2),\\
\hat{Y} &\sim \mathcal{N}(X\beta, H\sigma^2), \quad \textit{and}\\
\hat{\boldsymbol{\varepsilon}} &\sim \mathcal{N}(\mathbf{0}, (I-H)\sigma^2),
\end{aligned}$$

where $H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$. *Furthermore,* $\hat{\boldsymbol{\varepsilon}}$ *is independent of* $\hat{\beta}$ *and of* $\hat{Y}$.

Proof. Consider the vector

$$\boldsymbol{v} = \begin{pmatrix} \hat{\beta} \\ \hat{Y} \\ \hat{\boldsymbol{\varepsilon}} \end{pmatrix} = \begin{pmatrix} (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}} \\ H \\ I - H \end{pmatrix} Y \in \mathbb{R}^{p+2n}.$$

The response Y has a multivariate normal distribution and so therefore does $\boldsymbol{v}$. Hence each of $\hat{\beta}$, $\hat{Y}$, and $\hat{\boldsymbol{\varepsilon}}$ is multivariate normal.

The expected value of $\boldsymbol{v}$ is

$$\mathbb{E}(\boldsymbol{v}) = \begin{pmatrix} (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}} \\ H \\ I - H \end{pmatrix} X\beta = \begin{pmatrix} \beta \\ HX\beta \\ (I-H)X\beta \end{pmatrix} = \begin{pmatrix} \beta \\ X\beta \\ 0 \end{pmatrix}$$

because $HX = X$, establishing the means listed above for $\hat{\beta}$, $\hat{Y}$ and $\hat{\boldsymbol{\varepsilon}}$.

The variance of $\hat{\beta}$ is $(X^{\mathsf{T}}X)^{-1}\sigma^2$ by Lemma 2. Next the fits have $\mathrm{var}(\hat{Y}) = H\mathrm{var}(Y)H^{\mathsf{T}} = H(\sigma^2 I)H = H\sigma^2$ because H is symmetric and idempotent. Similarly $\hat{\boldsymbol{\varepsilon}} = (I-H)(X\beta + \boldsymbol{\varepsilon}) = (I-H)\boldsymbol{\varepsilon}$ and so $\mathrm{var}(\hat{\boldsymbol{\varepsilon}}) = (I-H)(\sigma^2 I)(I-H)^{\mathsf{T}} = (I-H)\sigma^2$.

We have established the three distributions displayed but have yet to prove the claimed independence. To this end

$$\mathrm{cov}(\hat{\beta}, \hat{\boldsymbol{\varepsilon}}) = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}(\sigma^2 I)(I-H) = (X^{\mathsf{T}}X)^{-1}(X - HX)^{\mathsf{T}}\sigma^2 = 0$$

because $X = HX$. Therefore $\hat{\beta}$ and $\hat{\boldsymbol{\varepsilon}}$ are uncorrelated and hence independent because $\boldsymbol{v}$ has a normal distribution. Similarly $\mathrm{cov}(\hat{Y}, \hat{\boldsymbol{\varepsilon}}) = H(I-H)^{\mathsf{T}}\sigma^2 = (H - HH^{\mathsf{T}})\sigma^2 = 0$ establishing the second and final independence claim. □

We can glean some insight about the hat matrix from Theorem 2. First because $\mathrm{var}(\hat{y}_i) = \sigma^2 H_{ii}$ we have $H_{ii} \geqslant 0$. Next because $\mathrm{var}(\hat{\varepsilon}_i) = \sigma^2(1 - H_{ii})$ we have $1 - H_{ii} \geqslant 0$. Therefore $0 \leqslant H_{ii} \leqslant 1$ holds for $i = 1, \ldots, n$.

Theorem 3. *Assume the conditions of Theorem 2, and also that $p < n$. Let $s^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^{\mathsf{T}}\hat{\beta})^2$. Then*

$$(n-p)s^2 \sim \sigma^2 \chi^2_{(n-p)}. \tag{3.14}$$

Proof. First

$$(n-p)s^2 = (Y - X\hat{\beta})^{\mathsf{T}}(Y - X\hat{\beta}) = Y^{\mathsf{T}}(I-H)Y = \boldsymbol{\varepsilon}^{\mathsf{T}}(I-H)\boldsymbol{\varepsilon}.$$

The matrix $I - H$ can be written $I - H = P\Lambda P^\mathsf{T}$ where P is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. From $I - H = (I - H)^2$ we get $\lambda_i \in \{0, 1\}$. Let $\tilde{\boldsymbol{\varepsilon}} = P^\mathsf{T}\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I)$. Then $\boldsymbol{\varepsilon}^\mathsf{T}(I - H)\boldsymbol{\varepsilon} = \tilde{\boldsymbol{\varepsilon}}^\mathsf{T}\Lambda\tilde{\boldsymbol{\varepsilon}} \sim \sigma^2\chi^2_{(k)}$ where k is the number of $\lambda_i = 1$, that is $k = \mathrm{tr}(I - H) = n - \mathrm{tr}(H)$. Now

$$\mathrm{tr}(H) = \mathrm{tr}(X(X^\mathsf{T}X)^{-1}X^\mathsf{T}) = \mathrm{tr}((X^\mathsf{T}X)^{-1}X^\mathsf{T}X) = p$$

so $k = n - p$. □

Equation (3.14) still holds trivially, even if $\sigma = 0$. The theorem statement excludes the edge case with $n = p$ because s^2 is then not well defined (zero over zero).

Very often we focus our attention on a specific linear combination of the components of β. We write such a combination as $\boldsymbol{c}\beta$ where $\boldsymbol{c}$ is a $1 \times p$ row vector. Taking $\boldsymbol{c} = (0, \ldots, 0, 1, 0, \ldots 0)$ with the 1 in the jth place gives $\boldsymbol{c}\beta = \beta_j$. More generally, parameter combinations like $\beta_2 - \beta_1$ or $\beta_1 - 2\beta_2 + \beta_3$ can be studied by making a judicious choice of $\boldsymbol{c}$. Taking $\boldsymbol{c} = \boldsymbol{x}_0^\mathsf{T}$ makes $\boldsymbol{c}\beta = \mathbb{E}(y\,|\,\boldsymbol{x}_0)$ and taking $\boldsymbol{c} = \boldsymbol{x}_0^\mathsf{T} - \tilde{\boldsymbol{x}}_0^\mathsf{T}$ lets us study $\mathbb{E}(y\,|\,\boldsymbol{x}_0) - \mathbb{E}(y\,|\,\tilde{\boldsymbol{x}}_0)$.

For any $\boldsymbol{c} \in \mathbb{R}^{1\times p}$, Theorem 2 implies that $\boldsymbol{c}\hat{\beta} \sim \mathcal{N}(\boldsymbol{c}\beta, \boldsymbol{c}(X^\mathsf{T}X)^{-1}\boldsymbol{c}^\mathsf{T}\sigma^2)$. We ordinarily don't know σ^2 but can estimate it via s^2 from (3.13). In later lectures we will test whether $\boldsymbol{c}\beta$ takes a specific value $\boldsymbol{c}\beta_0$, and form confidence intervals for $\boldsymbol{c}\beta$ using Theorem 4.

Theorem 4. *Let $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$ for $X \in \mathbb{R}^{n\times p}$ with $X^\mathsf{T}X$ invertible, and $\sigma > 0$. Let $\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$ and $s^2 = \frac{1}{n-p}(Y - X\hat{\beta})^\mathsf{T}(Y - X\hat{\beta})$. Then for any nonzero $1 \times p$ (row) vector $\boldsymbol{c}$,*

$$\frac{\boldsymbol{c}\hat{\beta} - \boldsymbol{c}\beta}{s\sqrt{\boldsymbol{c}(X^\mathsf{T}X)^{-1}\boldsymbol{c}^\mathsf{T}}} \sim t_{(n-p)}. \tag{3.15}$$

Proof. Let $U = (\boldsymbol{c}\hat{\beta} - \boldsymbol{c}\beta)/(\sigma\sqrt{\boldsymbol{c}(X^\mathsf{T}X)^{-1}\boldsymbol{c}^\mathsf{T}})$. Then $U \sim \mathcal{N}(0, 1)$. Let $V = s^2/\sigma^2$, so $V \sim \chi^2_{(n-p)}/(n-p)$. Now U is a function of $\hat{\beta}$ and V is a function of $\hat{\boldsymbol{\varepsilon}}$. Therefore U and V are independent. Finally

$$\frac{\boldsymbol{c}\hat{\beta} - \boldsymbol{c}\beta}{s\sqrt{\boldsymbol{c}(X^\mathsf{T}X)^{-1}\boldsymbol{c}^\mathsf{T}}} = \frac{\frac{(\boldsymbol{c}\hat{\beta} - \boldsymbol{c}\beta)}{\sqrt{\boldsymbol{c}(X^\mathsf{T}X)^{-1}\boldsymbol{c}^\mathsf{T}}\sigma}}{s/\sigma} = \frac{U}{\sqrt{V}} \sim t_{(n-p)}$$

by the definition of the t distribution. □

Theorem 4 is the basis for t tests and related confidence intervals for linear models. The recipe is to take the estimate $\boldsymbol{c}\hat{\beta}$ subtract the hypothesized parameter value $\boldsymbol{c}\beta$ and then divide by an estimate of the standard

deviation of $\boldsymbol{c}\hat{\beta}$. Having gone through these steps we're left with a $t_{(n-p)}$ distributed quantity. We will use those steps repeatedly in a variety of linear model settings.

Sometimes Theorem 4 is not quite powerful enough. We may have r different linear combinations of β embodied in an $r \times p$ matrix C and we wish to test whether $C\beta = C\beta_0$. We could test r rows of C individually, but testing them all at once is different and some problems will require such a simultaneous test. Theorem 5 is an r dimensional generalization of Theorem 4.

Theorem 5. *Let $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$ for $X \in \mathbb{R}^{n\times p}$ with $X^{\mathsf{T}}X$ invertible, and $\sigma > 0$. Let $\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y$ and $s^2 = \frac{1}{n-p}(Y - X\hat{\beta})^{\mathsf{T}}(Y - X\hat{\beta})$. Then for any $C \in \mathbb{R}^{r\times p}$ with linearly independent rows,*

$$\frac{1}{r}\,(C\hat{\beta} - C\beta)^{\mathsf{T}}\big[C(X^{\mathsf{T}}X)^{-1}C^{\mathsf{T}}\big]^{-1}(C\hat{\beta} - C\beta)/s^2 \sim F_{r,n-p}. \tag{3.16}$$

Proof. Let $U = C(\hat{\beta} - \beta)/\sigma$. Then $U \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$ where $\Sigma = C(X^{\mathsf{T}}X)^{-1}C^{\mathsf{T}}$ is a nonsingular $r \times r$ matrix, and so $U^{\mathsf{T}}\Sigma^{-1}U \sim \chi^2_{(r)}$. Let $V = s^2/\sigma^2$, so $V \sim \chi^2_{(n-p)}/(n-p)$. The left side of (3.16) is $(1/r)U^{\mathsf{T}}\Sigma^{-1}U \sim \chi^2_{(r)}/r$ divided by $V \sim \chi^2_{(n-p)}/(n-p)$. These two χ^2 random variables are independent because $\hat{\beta}$ is independent of s^2. □

To use Theorem 5 we formulate an hypothesis $H_0 : C\beta = C\beta_0$, plug $C\beta = C\beta_0$ into the left side of equation (3.16) and reject H_0 at the level α if the result is larger than $F^{1-\alpha}_{r,n-p}$.

If we put $r = 1$ in (3.16) and write C as $\boldsymbol{c}$ then it simplifies to

$$\frac{(\boldsymbol{c}\hat{\beta} - \boldsymbol{c}\beta)^2}{s^2\,\boldsymbol{c}(X^{\mathsf{T}}X)^{-1}\boldsymbol{c}^{\mathsf{T}}} \sim F_{1,n-p}.$$

The statistic on the left of (3.16) is the square of the one on the left of (3.15). The $F_{1,n-p}$ distribution on the right of (3.16) is the square of the $t_{(n-p)}$ distribution on the right of (3.15). The difference is that (3.15) takes account of the sign of $\boldsymbol{c}\hat{\beta} - \boldsymbol{c}\beta$ while (3.16) does not.

3.5 R^2 and the extra sum of squares

Constructing the matrix C and plugging it into Theorem 5 is a little awkward. In practice there is a simpler and more direct way to test hypotheses. Suppose that we want to test the hypothesis that some given subset of the

components of β are all 0. Then $\mathbb{E}(Y) = \widetilde{X}\gamma$ where $\widetilde{X}$ has $q < p$ of the columns of X and $\gamma \in \mathbb{R}^q$ contains coefficients for just those q columns. All we have to do is run the regression both ways, first on X and then on $\widetilde{X}$, and from the resulting sums of squared errors we can form a test of the submodel.

Let us call $\mathbb{E}(Y) = X\beta$ the full model and $\mathbb{E}(Y) = \widetilde{X}\gamma$ the submodel. Their sums of squared residuals are $\mathrm{SS}_{\mathrm{FULL}}$ and $\mathrm{SS}_{\mathrm{SUB}}$ respectively. Because the full model can reproduce any model that the submodel can, we always have $\mathrm{SS}_{\mathrm{FULL}} \leqslant \mathrm{SS}_{\mathrm{SUB}}$. But if the sum of squares increases too much under the submodel, then that is evidence against the submodel. The extra sum of squares is $\mathrm{SS}_{\mathrm{EXTRA}} = \mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}}$. We reject the submodel when $\mathrm{SS}_{\mathrm{EXTRA}}/\mathrm{SS}_{\mathrm{FULL}}$ is too large to have arisen by chance.

Theorem 6. *Suppose that $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$ where $X \in \mathbb{R}^{n\times p}$ with $X^{\mathsf{T}}X$ invertible, and $\sigma > 0$. Let $\widetilde{X} \in \mathbb{R}^{n\times q}$ contain $q < p$ columns of X. Let the full model have least squares estimate $\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y$, and sum of squared errors $\mathrm{SS}_{\mathrm{FULL}} = (Y - X\hat{\beta})^{\mathsf{T}}(Y - X\hat{\beta})$. Similarly let the submodel have least squares estimate $\hat{\gamma} = (\widetilde{X}^{\mathsf{T}}\widetilde{X})^{-1}\widetilde{X}^{\mathsf{T}}Y$ and sum of squared errors $\mathrm{SS}_{\mathrm{SUB}} = (Y - \widetilde{X}\hat{\gamma})^{\mathsf{T}}(Y - \widetilde{X}\hat{\gamma})$.*

If the submodel $Y = \widetilde{X}\gamma + \boldsymbol{\varepsilon}$ holds for some $\gamma \in \mathbb{R}^q$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$, then

$$\frac{\frac{1}{p-q}\big(\mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}}\big)}{\frac{1}{n-p}\,\mathrm{SS}_{\mathrm{FULL}}} \sim F_{p-q,n-p}. \tag{3.17}$$

Proof. Let $H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$ and $\widetilde{H} = \widetilde{X}(\widetilde{X}^{\mathsf{T}}\widetilde{X})^{-1}\widetilde{X}^{\mathsf{T}}$ be the hat matrices for the full model and submodel respectively. We know that $\mathrm{SS}_{\mathrm{FULL}} = Y^{\mathsf{T}}(I - H)Y \sim \sigma^2\chi^2_{\mathrm{rank}(I-H)}$ holds under the submodel because we have seen that it holds under the more general full model in Theorem 3. Also $\mathrm{rank}(I - H) = n - p$. It remains to show a scaled χ^2 distribution for the numerator and also show that it is independent of the denominator.

The extra sum of squares is

$$\mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}} = Y^{\mathsf{T}}(I - \widetilde{H})Y - Y^{\mathsf{T}}(I - H)Y = Y^{\mathsf{T}}(H - \widetilde{H})Y.$$

Now $\widetilde{H}\widetilde{X} = \widetilde{X}$ by (3.8) and similarly $H\widetilde{X} = \widetilde{X}$ because each column of $\widetilde{X}$ is also a column of X. Then

$$(H - \widetilde{H})Y = (H - \widetilde{H})(\widetilde{X}\gamma + \boldsymbol{\varepsilon}) = (H - \widetilde{H})\boldsymbol{\varepsilon},$$

so

$$Y^{\mathsf{T}}(H - \widetilde{H})Y = \boldsymbol{\varepsilon}^{\mathsf{T}}(H - \widetilde{H})\boldsymbol{\varepsilon}.$$

Now $H - \widetilde{H}$ is symmetric and idempotent. Symmetry is obvious, for idempotence write

$$\begin{aligned}(H-\widetilde{H})^2 &= H^2 - H\widetilde{H} - \widetilde{H}H + \widetilde{H}^2 \\ &= H - H\widetilde{H} - \widetilde{H}H + \widetilde{H}. \end{aligned} \tag{3.18}$$

Next $H\widetilde{H} = \widetilde{H}$ because the columns of $\widetilde{H}$ are in the submodel. Also by symmetry of hat matrices

$$\widetilde{H}H = \widetilde{H}^{\mathsf{T}}H^{\mathsf{T}} = (H\widetilde{H})^{\mathsf{T}} = \widetilde{H}^{\mathsf{T}} = \widetilde{H}.$$

So (3.18) reduces to $H - \widetilde{H}$ as required.

Now $\boldsymbol{\varepsilon}^{\mathsf{T}}(H-\widetilde{H})\boldsymbol{\varepsilon} \sim \sigma^2\chi^2_{\mathrm{rank}(H-\widetilde{H})}$, where

$$\mathrm{rank}(H-\widetilde{H}) = \mathrm{tr}(H-\widetilde{H}) = \mathrm{tr}(H) - \mathrm{tr}(\widetilde{H}) = p - q,$$

so $\mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}} \sim \sigma^2\chi^2_{p-q}$.

It remains to show that the numerator and denominator in (3.17) are independent. For that we note that $(H-\widetilde{H})(I-H) = 0$. Then $(H-\widetilde{H})\boldsymbol{\varepsilon}$ and $(I-H)\boldsymbol{\varepsilon}$ are uncorrelated and jointly Gaussian, hence independent, completing the proof. □

There is a geometric interpretation to equation (3.17). If $Y = \widetilde{X}\gamma + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ then we can write Y as $Y_{\mathrm{SUB}} + (Y_{\mathrm{FULL}} - Y_{\mathrm{SUB}}) + Y_{\mathrm{RES}}$ where $Y_{\mathrm{SUB}} \widetilde{H}Y$ is in a q dimensional space spanned by the columns of $\widetilde{X}$, $Y_{\mathrm{FULL}} = HY$ and $Y_{\mathrm{FULL}} - Y_{\mathrm{SUB}}$ is in the $p-q$ dimensional space of vectors orthogonal to $\widetilde{X}$ but in the column space of X and Y_{RES} is in the $n-p$ dimensional space of vectors orthogonal to X. A spherical Gaussian vector like $\boldsymbol{\varepsilon}$ shows no preference for any of these spaces. The mean squared norm of its projection in each space is just σ^2 times the dimension of the space. The F test is measuring whether the dimensions corresponding to $Y_{\mathrm{FULL}} - Y_{\mathrm{SUB}}$ are getting more than their share of the projection.

In an important special case, the full model contains an intercept and the submodel only has the intercept, so $q = 1$. Then $\mathrm{SS}_{\mathrm{SUB}} = \sum_{i=1}^n (y_i - \bar{y})^2$. The ANOVA decomposition in equation (3.7) has $\mathrm{SS}_{\mathrm{SUB}} = \mathrm{SS}_{\mathrm{FIT}} + \mathrm{SS}_{\mathrm{RES}}$. The F test then rejects the submodel in favor of the full model when R^2 is large enough.

3.6 Maximum likelihood and Gauss Markov

This section looks at motivations or justifications for choosing least squares. The first is via maximum likelihood. If we assume that the errors are nor-

mally distributed then least squares emerges as the maximum likelihood estimator of the regression coefficient β.

Using least squares is not the same as assuming normal errors. There are other assumptions that can lead one to using least squares. Under the weaker assumption that the error vector $\boldsymbol{\varepsilon}$ has mean zero and variance matrix $\sigma^2 I$, the least squares estimators have the smallest variance among the class of linear unbiased estimators. This optimality is made precise by the Gauss-Markov theorem.

The least squares estimates $\hat{\beta}$ can also be obtained as maximum likelihood estimates of β under a normal distribution model. That is, we look for the value of β which maximizes the probability of our observed data and we will find it is the same $\hat{\beta}$ that we found by least squares and studied via projections. The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^{\mathsf{T}}\hat{\beta})^2. \tag{3.19}$$

To derive this MLE, we suppose that $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$ for fixed and full rank X. The likelihood is

$$\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\Bigl(-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\Bigr). \tag{3.20}$$

For any $0 < \sigma < \infty$ the β which maximizes (3.20) is the one that minimizes the sum of squares $(Y - X\beta)^{\mathsf{T}}(Y - X\beta)$. So the MLE is the least squares estimate $\hat{\beta}$. Next we find the MLE of σ. The logarithm of (3.20) is

$$-\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}(Y - X\beta)^{\mathsf{T}}(Y - X\beta). \tag{3.21}$$

Differentiating (3.21) with respect to σ yields

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}(Y - X\beta)^{\mathsf{T}}(Y - X\beta), \tag{3.22}$$

which vanishes if and only if

$$\sigma^2 = \frac{1}{n}(Y - X\beta)^{\mathsf{T}}(Y - X\beta). \tag{3.23}$$

Solving (3.23) with $\beta = \hat{\beta}$ gives $\hat{\sigma}^2$ from (3.19). To check that this is the maximum we write the derivative (3.22) as $n(\hat{\sigma}^2 - \sigma^2)/\sigma^3$, for $\beta = \hat{\beta}$, which is positive for $\sigma^2 < \hat{\sigma}^2$ and negative for $\sigma^2 > \hat{\sigma}^2$.

In the case where $Y = X\hat{\beta}$ exactly, the log likelihood is degenerate. The MLE formulas for $\hat{\beta}$ and $\hat{\sigma}$ are still interpretable. The first is the value that makes error zero and the second reduces to 0.

The second justification for least squares is via the Gauss-Markov theorem. We say that $\lambda^{\mathsf{T}}Y$ is a linear unbiased estimator for $\boldsymbol{c}\beta$ if $\mathbb{E}(\lambda^{\mathsf{T}}Y) = \boldsymbol{c}\beta$ holds for all β. We define the best linear unbiased estimator of $\boldsymbol{c}\beta$ to be the one that attains the minimum variance. The Gauss-Markov theorem states that least squares provides best linear unbiased estimators. We make use of the notion of 'estimable' linear combinations $\boldsymbol{c}\beta$. Those are defined in Section 3.9. When X has full rank as we consider here then every linear combination $\boldsymbol{c}\beta$ is estimable.

Theorem 7 (Gauss-Markov)**.** *Let $Y = X\beta + \boldsymbol{\varepsilon}$ where $X \in \mathbb{R}^{n\times p}$, β is an unknown point in $\mathbb{R}^p$ and $\boldsymbol{\varepsilon}$ is a random vector with mean 0 and variance matrix $\sigma^2 I_n$. Let $\boldsymbol{c}\beta$ be estimable and let $\hat{\beta}$ be a least squares estimate. Then $\boldsymbol{c}\hat{\beta}$ is a best linear unbiased estimate of $\boldsymbol{c}\beta$.*

Full rank case. We prove the theorem for the full rank case with $X^{\mathsf{T}}X$ invertible. We already know that $\boldsymbol{c}\hat{\beta}$ is unbiased and linear. Suppose that $\mathbb{E}(\lambda^{\mathsf{T}}Y) = \boldsymbol{c}\beta$, and so $\lambda^{\mathsf{T}}X = \boldsymbol{c}$. Then

$$\begin{aligned}\operatorname{var}(\lambda^{\mathsf{T}}Y) &= \operatorname{var}(\boldsymbol{c}\hat{\beta} + (\lambda^{\mathsf{T}}Y - \boldsymbol{c}\hat{\beta}))\\ &= \operatorname{var}(\boldsymbol{c}\hat{\beta}) + \operatorname{var}(\lambda^{\mathsf{T}}Y - \boldsymbol{c}\hat{\beta}) + 2\operatorname{cov}(\boldsymbol{c}\hat{\beta}, \lambda^{\mathsf{T}}Y - \boldsymbol{c}\hat{\beta}).\end{aligned}$$

Now

$$\begin{aligned}\operatorname{cov}(\boldsymbol{c}\hat{\beta}, \lambda^{\mathsf{T}}Y - \boldsymbol{c}\hat{\beta}) &= \operatorname{cov}(\boldsymbol{c}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y, (\lambda^{\mathsf{T}} - \boldsymbol{c}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}})Y)\\ &= \boldsymbol{c}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}(\sigma^2 I)(\lambda - X(X^{\mathsf{T}}X)^{-1}\boldsymbol{c}^{\mathsf{T}})\\ &= \sigma^2\Big(\boldsymbol{c}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\lambda - \boldsymbol{c}(X^{\mathsf{T}}X)^{-1}\boldsymbol{c}^{\mathsf{T}}\Big)\\ &= 0\end{aligned}$$

because $X^{\mathsf{T}}\lambda = \boldsymbol{c}^{\mathsf{T}}$. Therefore $\operatorname{var}(\lambda^{\mathsf{T}}Y) = \operatorname{var}(\boldsymbol{c}\hat{\beta}) + \operatorname{var}(\lambda^{\mathsf{T}}Y - \boldsymbol{c}\hat{\beta}) \geqslant \operatorname{var}(\boldsymbol{c}\hat{\beta})$. □

For rank deficient X, a similar proof can be made but takes a bit more care. See for example Christensen (2011). We could drop columns of X until $X^{\mathsf{T}}X$ has full rank, get a best linear unbiased estimator in the resulting full rank problem, and then because we only dropped redundant columns, reinstate the ones we dropped.

3.7 Least squares computation

It is by now rare for somebody analyzing data to have to program up a least squares algorithm. In a research setting however, it might be simpler to embed a least squares solver in one's software than to write the software in some other tool that has least squares solvers built in. Other times our problem is too big to fit our usual data analysis environment and we have to compute least squares some other way. Even outside of these special settings, we may find that our software offers us several least squares algorithms to choose from and then it pays to understand the choices.

Before considering special cases, we note that the cost to do least squares computation typically grows proportionally to $O(np^2 + p^3)$. For example when solving the normal equations directly, it takes $O(np^2)$ work simply to form $X^\mathsf{T}X$ and $X^\mathsf{T}Y$ and then $O(p^3)$ to solve $X^\mathsf{T}X\hat{\beta} = X^\mathsf{T}Y$. We will consider more sophisticated least squares algorithms based on the singular value decomposition. It also costs $O(np^2 + p^3)$ with different implied constants.

Solving the normal equations

Given an n by p matrix X and $Y \in \mathbb{R}^n$ the least squares problem is to solve $(X^\mathsf{T}X)\hat{\beta} = X^\mathsf{T}Y$ for $\hat{\beta}$. We have p equations in p unknowns. The standard way to solve p linear equations in p unknowns is to apply Gaussian elimination. Sometimes Gaussian elimination encounters numerical difficulty because it requires division by a small number. In that case pivoting methods, which rearrange the order of the p equations can improve accuracy.

When X has full rank then $X^\mathsf{T}X$ is positive definite. It can be factored by the Cholesky decomposition into $X^\mathsf{T}X = G^\mathsf{T}G$ where G is a p by p upper triangular matrix. Then to solve $G^\mathsf{T}G\hat{\beta} = X^\mathsf{T}Y$ we solve two triangular systems of equations: first we solve $G^\mathsf{T}u = X^\mathsf{T}Y$ for u and then we solve $G\hat{\beta} = u$ for $\hat{\beta}$. A triangular system is convenient because, one can start with a single equation in a single unknown and each equation thereafter brings in one new variable to solve for.

It turns out that neither Gaussian elimination nor Cholesky decomposition is the best way to solve least squares problems. They are numerically unstable if $X^\mathsf{T}X$ is nearly singular and of course they fail if $X^\mathsf{T}X$ is actually singular. The SVD considered below is more reliable and can even handle X that does not have full rank.

Singular value decomposition

Any n by p matrix X may be written

$$X = U\Sigma V^{\mathsf{T}} \tag{3.24}$$

where U is an $n \times n$ orthogonal matrix, V is a $p \times p$ orthogonal matrix, and Σ is an $n \times p$ diagonal matrix with nonnegative elements. For a proof of the existence of the SVD and for algorithms to compute it, see Golub and Van Loan (2013).

If we multiply the matrix X by a column vector $\boldsymbol{w}$ we find that $X\boldsymbol{w} = U\Sigma V^{\mathsf{T}}\boldsymbol{w}$ corresponds to rotating $\boldsymbol{w}$ by V^{T} then scaling each component of $V^{\mathsf{T}}\boldsymbol{w}$ by a corresponding diagonal element of Σ, and then rotating the result again. Rotate, stretch, and rotate is what happens to $\boldsymbol{w}$ when multiplied by X and this holds for any X. We use the term rotate here to mean a rigid motion due to the action of an orthogonal matrix. Reflections and coordinate permutations are included in this notion of rotation.

Suppose that the i'th column of U is $\boldsymbol{u}_i$, the i'th column of V is $\boldsymbol{v}_i$ and the i'th diagonal element of Σ is σ_i. Then we can write

$$X = \sum_{i=1}^{\min(n,p)} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}$$

as a sum of rank one matrices. It is customary to order the indices so that $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_{\min(n,p)} \geqslant 0$. The rank of X is the number r of nonzero σ_i.

When $n > p$, the last $n - p$ rows of Σ are zero. That is we can also write

$$X = \widetilde{U}\widetilde{\Sigma}V^{\mathsf{T}}$$

where $\widetilde{U}$ has the first p columns of $U = (\widetilde{U}\ U^*)$ and $\widetilde{\Sigma}$ has the first p rows of Σ.

The length of the vector $Y - X\beta$ is preserved when it is multiplied by an orthogonal matrix such as U^{T}. Then

$$\begin{aligned}
\|Y - X\beta\|^2 &= \|Y - U\Sigma V^{\mathsf{T}}\beta\|^2 \\
&= \|U^{\mathsf{T}}Y - \Sigma V^{\mathsf{T}}\beta\|^2 \\
&= \|\widetilde{U}^{\mathsf{T}}Y - \widetilde{\Sigma}V^{\mathsf{T}}\beta\|^2 + \|U^{*\mathsf{T}}Y\|^2,
\end{aligned}$$

after splitting the first p from the last $n - p$ components of $U^{\mathsf{T}}Y - \Sigma V^{\mathsf{T}}\beta$. Any vector β that minimizes $\|\widetilde{U}^{\mathsf{T}}Y - \widetilde{\Sigma}V^{\mathsf{T}}\beta\|$ is a least squares solution. Let $\widetilde{Y} = \widetilde{U}^{\mathsf{T}}Y \in \mathbb{R}^p$ and $\widetilde{\beta} = V^{\mathsf{T}}\beta \in \mathbb{R}^p$. In the nonsingular case each

$\sigma_i > 0$. Then the least squares solution is found by putting $\widetilde{\beta}_i = \widetilde{y}_i/\sigma_i$ for $i = 1, \dots, p$ and $\beta = V\widetilde{\beta}$.

In the singular case one or more of the σ_i are zero. Suppose that the last k of the σ_i are zero. Then setting $\widetilde{\beta}_i = \widetilde{y}_i/\sigma_i$ for $i = 1, \dots, p-k$ and taking any real values at all for $\widetilde{\beta}_i$ with $i = p-k+1, \dots, p$ gives a least squares solution.

The shortest least squares vector is found by taking $\widetilde{\beta}_i = 0$ for $i = p-k+1, \dots, p$. Writing

$$\sigma_i^+ = \begin{cases} 1/\sigma_i & \sigma_i > 0 \\ 0 & \sigma_i = 0 \end{cases}$$

and $\Sigma^+ = \mathrm{diag}(\sigma_1^+, \dots, \sigma_p^+)$ we find that the shortest least squares solution is

$$\hat{\beta} = V\widetilde{\beta} = V\Sigma^+\widetilde{U}^{\mathsf{T}}Y. \tag{3.25}$$

When we want to simply pick one least squares solution out of many, to use as a default, then the one in (3.25) is a good choice. In practice the computed singular values have some roundoff error in them. Then one should set $\sigma_i^+ = 0$ when $\sigma_i \leqslant \sigma_1 \times \epsilon$ where ϵ is a measure of numerical precision. Note that the threshold scales with the largest singular value. A matrix X with $\sigma_1 = \sigma_2 = \cdots = \sigma_p = \epsilon/2$ would lead to $\hat{\beta} = 0$ unless we use a relative error. In the unusual event that all σ_i are zero then the condition $\sigma_i \leqslant \sigma_1 \times \epsilon$ is fulfilled as we would like while the condition $\sigma_i < \sigma_1 \times \epsilon$ is not.

Now suppose that we want a parameterized form for the entire set of least squares solutions. Perhaps we're going to pick one of them by numerically optimizing some other condition instead of $\|\hat{\beta}\|$ used above. When k of the σ_i are zero, then we write the entire set as a function of $\tau \in \mathbb{R}^k$, as follows:

$$V\Sigma^+\widetilde{U}^{\mathsf{T}}Y + V_2\tau$$

where V_2 has the last k columns of V.

3.8 Special least squares systems

If the columns $X_{\bullet j}$ of X are orthogonal then least squares problems simplify considerably. The reason is that $X^{\mathsf{T}}X$ is diagonal and then so is $(X^{\mathsf{T}}X)^{-1}$ in the full rank case. The result is that

$$\hat{\beta}_j = \frac{\sum_{i=1}^n X_{ij}y_i}{\sum_{i=1}^n X_{ij}^2} = \frac{X_{\bullet j}^{\mathsf{T}}Y}{\|X_{\bullet j}\|^2}, \quad j = 1, \dots, p,$$

so the coefficients can be computed independently, one at a time. In the even more special case that each of $X_{\bullet j}$ is a unit vector then $\hat{\beta}_j = X_{\bullet j}^{\mathsf{T}} Y$.

Orthogonal predictors bring great simplification. The cost of computation is only $O(np)$. The variance of $\hat{\beta}$ is $\sigma^2 \mathrm{diag}(1/\|X_{\bullet j}\|^2)$ so the components of $\hat{\beta}$ are uncorrelated. In the Gaussian case the $\hat{\beta}_j$ are statistically independent in addition to the computational independence noted above.

In problems where scattered data are observed we won't ordinarily find orthogonal columns. In designed experiments we might arrange for orthogonal columns.

When the real world data $\boldsymbol{u}_i$ lie on a regular grid in one or two dimensions then we may construct some orthogonal features $\boldsymbol{x}_i$ from them.

3.9 Rank deficient features

Sometimes $X^{\mathsf{T}}X$ has no inverse and so $(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y$ is not available as an estimate. This will happen if one of the columns of X is a linear combination of the others. In fact that is the only way that a singular $X^{\mathsf{T}}X$ matrix can arise:

Theorem 8. *If $X^{\mathsf{T}}X$ is not invertible then one column of X is a linear combination of the others.*

Proof. If $X^{\mathsf{T}}X$ is not invertible then $X^{\mathsf{T}}X\boldsymbol{v} = 0$ for some nonzero $\boldsymbol{v} \in \mathbb{R}^p$. Then $\boldsymbol{v}^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{v} = 0$ too. But $\boldsymbol{v}^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{v} = \|X\boldsymbol{v}\|^2$, and so $X\boldsymbol{v} = 0$. Let v_j be a nonzero element of $\boldsymbol{v}$. Then $X_{\bullet j} = -\sum_{k\neq j} X_{\bullet k} v_k / v_j$. □

There is more than one way that we could find one of our features to be a linear combination of the others. Perhaps we have measured temperature in degrees Fahrenheit as well as Celsius. Recalling that "$F = 1.8C + 32$" we will find linear dependence in X, if as usual the model includes an intercept. Similarly, if a financial model makes features out of sales from every region and also the total sales over all regions, then 'total' column within X will be the sum of the regional ones. Finally a row of X might take the form $(1, 1_{\mathrm{M}}, 1_{\mathrm{F}}, \dots)^{\mathsf{T}}$ where 1_{M} and 1_{F} are indicator functions of male and female subjects respectively and the dots represent additional features.

A rank deficient matrix X is often a sign that we have made a mistake in encoding our regression. In that case we fix it by removing redundant features until $X^{\mathsf{T}}X$ is invertible, and then using methods and results for the invertible case. The rest of this section describes settings where we might actually prefer to work with a singular encoding.

One reason to work with a redundant encoding is for symmetry. We'll see examples of that for the one way analysis of variance. Another reason is that we might be planning to do a large number of regressions, say one for every gene in an organism, one for every customer of a company, or one per second in an electronic device. These regressions might be all on the same variables but some may be singular as a consequence of the data sets they're given. It may not always be the same variable that is responsible for the singularity. Also, no matter what the features are, we are guaranteed to have a non-invertible $X^\mathsf{T}X$ if $p > n$. When we cannot tweak each regression individually to remove redundant features, then we need a graceful way to do regressions with or without linearly dependent columns.

In the singular setting $\mathcal{M} = \{X\beta \mid \beta \in \mathbb{R}^p\}$ is an r dimensional subspace of $\mathbb{R}^n$ where $r < p$. We have an r dimensional space indexed by a p dimensional vector β. There is thus a $p - r$ dimensional set of labels β for each point in $\mathcal{M}$. Suppose for example that $X_{i1} = 1$, and that $X_{i2} = F_i$ and $X_{i3} = C_i$ are temperatures in degrees Fahrenheit and Celsius of data points $i = 1, \dots, n$. Then for any $t \in \mathbb{R}$,

$$\beta_0 + \beta_1 F_i + \beta_2 C_i = (\beta_0 + 32t) + (\beta_1 - t)F_i + (\beta_2 + 1.8t)C_i.$$

If y_i is near to $\beta_0 + \beta_1 F_i + \beta_2 C_i$ then it is equally near to $(\beta_0 + 32t) + (\beta_1 - t)F_i + (\beta_2 + 1.8t)C_i$ for any value of t.

The problem with rank deficient X is not, as we might have feared, that there is no solution to $X^\mathsf{T}X\hat\beta = X^\mathsf{T}Y$. It is that there is an infinite family of solutions. That family is a linear subspace of $\mathbb{R}^p$. A reasonable approach is to simply choose the shortest solution vector. That is we find the unique minimizer of $\beta^\mathsf{T}\beta$ subject to the constraint $X^\mathsf{T}X\beta = X^\mathsf{T}Y$ and we call it $\hat\beta$. We saw how to do this numerically in Section 3.7.

When $X^\mathsf{T}X$ is singular we may well find that some linear combinations $\boldsymbol{c}\hat\beta$ are uniquely determined by the normal equations $(X^\mathsf{T}X)\hat\beta = X^\mathsf{T}Y$ even though $\hat\beta$ itself is not. If we can confine our study of β to such linear combinations $\boldsymbol{c}\beta$ then singularity of $X^\mathsf{T}X$ is not so deleterious.

From the geometry, it is clear that there has to be a unique point in $\mathcal{M}$ that is closest to the vector Y. Therefore $\hat Y = X\hat\beta$ is determined by the least squares conditions. It follows that each component $\hat y_i = \boldsymbol{x}_i^\mathsf{T}\hat\beta$ is also determined. So we know that if $\boldsymbol{c} = \boldsymbol{x}_i^\mathsf{T}$ for one of our sampled feature vectors, then $\boldsymbol{c}\hat\beta$ is determined.

It then follows that if $\boldsymbol{c}$ is a linear combination of $\boldsymbol{x}_i^\mathsf{T}$, that is a linear combination of rows of the design matrix X, then $\boldsymbol{c}\hat\beta$ is determined. We can write such $\boldsymbol{c}$ as $\boldsymbol{c} = \sum_{i=1}^n \gamma_i \boldsymbol{x}_i^\mathsf{T} = \gamma^\mathsf{T}X$ for $\gamma \in \mathbb{R}^n$. In fact, the only estimable $\boldsymbol{c}\beta$ take this form.

Definition 1 (Estimability)**.** The linear combination $\boldsymbol{c}\beta$ is estimable if there exists a linear combination $\lambda^\mathsf{T} Y = \sum_{i=1}^n \lambda_i y_i$ of the responses for which $\mathbb{E}(\lambda^\mathsf{T} Y) = \boldsymbol{c}\beta$ holds for all $\beta \in \mathbb{R}^p$.

The definition of estimability above stipulates that we can find a linear combination of the y_i that is unbiased for $\boldsymbol{c}\beta$. The next theorem gives two equivalent properties that could also have been used as the definition. One is that estimable linear combinations $\boldsymbol{c}\beta$ have unique least squares estimators. The other is that estimable linear combinations are linear combinations of the rows of the design matrix X. If there is some $\boldsymbol{c}\beta$ that we really want to estimate, we can make sure it is estimable when we're choosing the $\boldsymbol{x}_i$ at which to gather data.

For this course, remember the definition of estimability and the properties below. If $\boldsymbol{c}\beta$ is estimable then $\boldsymbol{c}$ is a linear combination of the $\boldsymbol{x}_i$ and conversely. Also $\boldsymbol{c}\beta$ is estimable if and only if $\boldsymbol{c}\hat{\beta}$ is the same for any solution $\hat{\beta}$ of the normal equations.

Theorem 9. *The following are equivalent:*

1) $\boldsymbol{c}\beta$ *is estimable.*

2) $\boldsymbol{c} = \gamma^\mathsf{T} X$ *for some* $\gamma \in \mathbb{R}^n$.

3) If $X^\mathsf{T} X\hat{\beta} = X^\mathsf{T} Y$ *and* $X^\mathsf{T} X\tilde{\beta} = X^\mathsf{T} Y$ *then* $\boldsymbol{c}\hat{\beta} = \boldsymbol{c}\tilde{\beta}$.

Proof. If $\boldsymbol{c}\beta$ is estimable then for some λ, $\boldsymbol{c}\beta = \mathbb{E}(\lambda^\mathsf{T} Y) = \lambda^\mathsf{T} X\beta$ holds for all β. Taking $\gamma = \lambda$ shows that $1 \Rightarrow 2$.

If 2) holds, then $\boldsymbol{c}\beta = \gamma^\mathsf{T} X\beta = \gamma^\mathsf{T}\mathbb{E}(Y)$ so taking $\lambda = \gamma$ establishes that $\boldsymbol{c}\beta$ is estimable, so $2 \Rightarrow 1$.

Let's suppose that $\boldsymbol{c} = \gamma^\mathsf{T} X$ and that $\hat{\beta}$ and $\tilde{\beta}$ both solve the normal equations. Therefore $(X^\mathsf{T} X)\hat{\beta} = (X^\mathsf{T} X)\tilde{\beta}$ and so $(X^\mathsf{T} X)(\hat{\beta} - \tilde{\beta}) = \boldsymbol{0}$. That implies that $X(\hat{\beta} - \tilde{\beta}) = \boldsymbol{0}$ as we saw in the proof of Theorem 8. Now $\boldsymbol{c}\hat{\beta} - \boldsymbol{c}\tilde{\beta} = \gamma^\mathsf{T} X(\hat{\beta} - \tilde{\beta}) = 0$. That gives us $\boldsymbol{c}\hat{\beta} = \boldsymbol{c}\tilde{\beta}$ and so $2 \Rightarrow 3$.

Now suppose that 3) holds making the least squares estimates of $\boldsymbol{c}\beta$ unique but 2) does not hold. Since 2) does not hold, $\boldsymbol{c}$ is a linear combination of the rows of X plus something orthogonal to all of those rows. We write $\boldsymbol{c} = \gamma^\mathsf{T} X + \boldsymbol{v}^\mathsf{T}$ where $X\boldsymbol{v} = \boldsymbol{0}$, for nonzero $\boldsymbol{v} \in \mathbb{R}^p$. If $\hat{\beta}$ is a least squares estimate then so is $\tilde{\beta} = \hat{\beta} + \boldsymbol{v}$. Now $\boldsymbol{c}(\tilde{\beta} - \hat{\beta}) = \boldsymbol{c}\boldsymbol{v} = \gamma^\mathsf{T} X\boldsymbol{v} + \|\boldsymbol{v}\|^2 = \|\boldsymbol{v}\|^2 > 0$, a contradiction. Therefore $3 \Rightarrow 2$. □

Bibliography

R. Christensen. *Plane answers to complex questions: the theory of linear models*. Springer, 2011.

G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU press, 2013.