2

Linear Least Squares

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.

2.1 Least squares estimates

We begin with observed response values $y_1, y_2, \dots, y_n$ and features z_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, p$. We're looking for the parameter values $\beta_1, \dots, \beta_p$ that minimize

$$S(\beta_1, \dots, \beta_p) = \sum_{i=1}^{n}\Bigl(y_i - \sum_{j=1}^{p} z_{ij}\beta_j\Bigr)^2. \tag{2.1}$$

The minimizing values are denoted with hats: $\hat\beta_1, \dots, \hat\beta_p$.

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

$$\frac{\partial}{\partial\beta_j}S = -2\sum_{i=1}^{n}\Bigl(y_i - \sum_{j=1}^{p} z_{ij}\hat\beta_j\Bigr)z_{ij} = 0, \quad j = 1, \dots, p. \tag{2.2}$$

We'll show later that this indeed gives the minimum, not the maximum or a saddle point. The p equations in (2.2) 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 $Z_{\bullet j} = (z_{1j}, \dots, z_{nj})' \in \mathbb{R}^n$ of values for the j'th feature. Equation (2.2) says that this feature vector has a dot product of zero with the residual vector having i'th element $\hat{\varepsilon}_i = y_i - \sum_{j=1}^p z_{ij}\hat{\beta}_j$. Each feature vector is orthogonal (normal) to the vector of residuals.

It is worth while writing equations (2.1) and (2.2) in matrix notation. Though the transition from (2.1) to (2.2) is simpler with coordinates, our later manipulations are easier in vector form. Let's pack the responses and feature values into the vector Y and matrix Z via

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

We also put $\beta = (\beta_1, \dots, \beta_p)'$ and let $\hat{\beta}$ denote the minimizer of the sum of squares. Finally let $\hat{\varepsilon} = Y - Z\hat{\beta}$, the n vector of residuals.

Now equation (2.1) can be written $S(\beta) = (Y - Z\beta)'(Y - Z\beta)$ and the normal equations (2.2) become

$$\hat{\varepsilon}'Z = 0 \tag{2.3}$$

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

$$Z'Z\hat{\beta} = Z'Y. \tag{2.4}$$

When $Z'Z$ is invertible then we find that

$$\hat{\beta} = (Z'Z)^{-1}Z'Y. \tag{2.5}$$

We'll suppose at first that $Z'Z$ is invertible and then consider the singular case in Chapter 2.7.

Equation (2.5) underlies another meaning of the work '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 $(Z'Z)^{-1}Z'$. The predicted value of Y at any new point x_0 with features $z_0 = \phi(x_0)$ is also linear in Y; it is $z_0'(Z'Z)^{-1}Z'Y$.

Now let's prove that $\hat{\beta} = (Z'Z)^{-1}Z'Y$ is in fact the minimizer, not just a point where the gradient of the sum of squares vanishes. It seems obvious that a sum of squares like (2.1) cannot have a maximizing β. But we need to rule out saddle points too, and we'll also find that $\hat{\beta}$ is the unique least squares estimator. Since we already found an expression for $\hat{\beta}$ we prove it is right by expressing a generic $\widetilde{\beta} \in \mathbb{R}^p$ as $\hat{\beta} + (\widetilde{\beta} - \hat{\beta})$ and then expanding $S(\hat{\beta} + (\widetilde{\beta} - \hat{\beta}))$. This adding and subtracting technique is often applicable in problems featuring squared errors.

Theorem 2.1. *Let Z be an $n \times p$ matrix with $Z'Z$ invertible, and let Y be an n vector. Define $S(\beta) = (Y - Z\beta)'(Y - Z\beta)$ and set $\hat\beta = (Z'Z)^{-1}Z'Y$. Then $S(\beta) > S(\hat\beta)$ holds whenever $\beta \neq \hat\beta$.*

Proof. We know that $Z'(Y - Z\hat\beta) = 0$ and will use it below. Let $\widetilde\beta$ be any point in $\mathbb{R}^p$ and let $\gamma = \widetilde\beta - \hat\beta$. Then

$$\begin{aligned} S(\widetilde\beta) &= (Y - Z\widetilde\beta)'(Y - Z\widetilde\beta) \\ &= (Y - Z\hat\beta - Z\gamma)'(Y - Z\hat\beta - Z\gamma) \\ &= (Y - Z\hat\beta)'(Y - Z\hat\beta) - \gamma' Z'(Y - Z\hat\beta) - (Y - Z\hat\beta)Z\gamma + \gamma' Z' Z\gamma \\ &= S(\hat\beta) + \gamma' Z' Z\gamma. \end{aligned}$$

Thus $S(\widetilde\beta) = S(\hat\beta) + \|Z\gamma\|^2 \geq S(\hat\beta)$. It follows that $\hat\beta$ is a minimizer of S. For uniqueness we need to show that $\gamma \neq 0$ implies $Z\gamma \neq 0$. Suppose to the contrary that $Z\gamma = 0$ for $\gamma \neq 0$. Then we would have $Z'Z\gamma = 0$ for $\gamma \neq 0$, but this contradicts the assumption that $Z'Z$ is invertible. Therefore if $\widetilde\beta \neq \hat\beta$ then $S(\widetilde\beta) = S(\hat\beta) + \|Z(\widetilde\beta - \hat\beta)\|^2 > 0$ □

2.2 Geometry of least squares

Figure xxx shows a sketch to illustrate linear least squares. The vector $y = (y_1, \dots, y_n)'$ is represented as a point in $\mathbb{R}^n$. The set

$$\mathcal{M} = \{Z\beta \mid \beta \in \mathbb{R}^p\} \tag{2.6}$$

is a p dimensional linear subset of $\mathbb{R}^n$. It is fully p dimensional here because we have assumed that $Z'Z$ is invertible and so Z has rank p. Under our model, $E(Y) = Z\beta \in \mathcal{M}$.

The idea behind least squares is to find $\hat\beta$ so that $\hat y = Z\hat\beta$ 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 $\hat\varepsilon = y - \hat y$ should be perpendicular to any line within $\mathcal{M}$. From the normal equations (2.3), we have $\hat\varepsilon' Z = 0$ so $\hat\varepsilon$ is actually perpendicular to every point $Z\beta \in \mathcal{M}$. Therefore $\hat\varepsilon$ is perpendicular to every line in $\mathcal{M}$.

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

$$\|y\|^2 = \|\hat\varepsilon\|^2 + \|z\hat\beta\|^2.$$

When the first column of Z consists of 1s then $(\bar y, \dots, \bar y)' = Z(\bar y, 0, \dots, 0) \in \mathcal{M}$ and Pythagoras' theorem implies

$$\sum_{i=1}^n (y_i - \bar y)^2 = \sum_{i=1}^n (\hat y_i - \bar y)^2 + \sum_{i=1}^n (y_i - \hat y)^2. \tag{2.7}$$

The left side of (2.7) is called the centered sum of squares of the y_i. It is $n-1$ times the usual estimate of the common variance of the Y_i. The equation decomposes this sum of squares into two parts. The first is the centered sum of squared errors of the fitted values $\hat{y}_i$. The second is the sum of squared model errors. When the first column of Z consists of 1s then $(1/n)\sum_{i=1}^n \hat{y}_i = \bar{y}$. That is $\bar{\hat{y}} = \bar{y}$. Now the two terms in (2.7) correspond to the sum of squares of the fitted values $\hat{y}_i$ about their mean and the sum of squared residuals. We write this as $\mathrm{SS}_{\mathrm{TOT}} = \mathrm{SS}_{\mathrm{FIT}} + \mathrm{SS}_{\mathrm{RES}}$. The total sum of squares is the sum of squared fits plus the sum of squared residuals.

Some of the variation in Y is 'explained' by the model and some of it left unexplained. The fraction explained is denoted by

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

The quantity R^2 is known as the coefficient of determination. It measures how well Y is predicted or determined by $Z\hat{\beta}$. Its nonnegative square root R is called the coefficient of multiple correlation. It is one measure of how well the response Y correlates with the p predictors in Z taken collectively. For the special case of simple linear regression with $z_i = (1, x_i) \in \mathbb{R}^2$ then (Exercise xxx) R^2 is the square of the usual Pearson correlation of x and y.

Equation (2.7) is an example of an ANOVA (short for analysis of variance) decomposition. ANOVA decompositions split a variance (or a sum of squares) into two or more pieces. Not surprisingly there is typically some orthogonality or the Pythagoras theorem behind them.

2.3 Algebra of least squares

The predicted value for y_i, using the least squares estimates, is $\hat{y}_i = Z_i\hat{\beta}$. We can write the whole vector of fitted values as $\hat{y} = Z\hat{\beta} = Z(Z'Z)^{-1}Z'Y$. That is $\hat{y} = Hy$ where

$$H = Z(Z'Z)^{-1}Z'.$$

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.

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 $H^2 = H$. Clearly $H^k = H$ for any integer $k \geq 1$.

The matrix $Z'Z$ is symmetric, and so therefore is $(Z'Z)^{-1}$. It follows that the hat matrix H is symmetric too. A symmetric idempotent matrix such as H is called a perpendicular projection matrix.

Theorem 2.2. *Let H be a symmetric idempotent real valued matrix. Then the eigenvalues of H are all either* 0 *or* 1.

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

Let $H = P'\Lambda P$ where the columns of P are eigenvectors p_i of H for $i = 1, \dots, n$. Then $H = \sum_{i=1}^n \lambda_i p_i p_i'$, where by Theorem 2.2 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 $H = \sum_{i=1}^r p_i p_i'$. We certainly expect r to equal p here. This indeed holds. The eigenvalues of H sum to r, so $\mathrm{tr}(H) = r$. Also $\mathrm{tr}(H) = \mathrm{tr}(Z(Z'Z)^{-1}Z') = \mathrm{tr}(Z'Z(Z'Z)^{-1}) = \mathrm{tr}(I_p) = p$. Therefore $r = p$ and $H = \sum_{i=1}^p p_i p_i'$ where p_i are mutually orthogonal n vectors.

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 $z_i'(Z'Z)^{-1}Z'$ and the ij element of the hat matrix is $H_{ij} = z_i'(Z'Z)^{-1}z_j$. 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

$$H_{ii} = z_i'(Z'Z)^{-1}z_i. \tag{2.8}$$

We are also interested in the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$. The entire vector of residuals may be written $\hat{\varepsilon} = y - \hat{y} = (I - H)y$. 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 $\mathcal{M} = \{Z\beta \mid \beta \in \mathbb{R}^p\}$ is the set of linear combinations of columns of Z. A typical entry of $\mathcal{M}$ is $\sum_{j=1}^p \beta_j Z_{\bullet j}$. Thus $\mathcal{M}$ is what is known as the column space of Z, denoted $\mathrm{col}(Z)$. The hat matrix H projects vectors onto $\mathrm{col}(Z)$. The set of vectors orthogonal to Z, that is with $Zv = 0$, is a linear subspace of $\mathbb{R}^n$, known as the null space of Z. We may write it as $\mathcal{M}^\perp$ or as $\mathrm{null}(Z)$. The column space and null spaces are orthogonal complements. Any $v \in \mathbb{R}^n$ can be uniquely written as $v_1 + v_2$ where $v_1 \in \mathcal{M}$ and $v_2 \in \mathcal{M}^\perp$. In terms of the hat matrix, $v_1 = Hv$ and $v_2 = (I - H)v$.

2.4 Distributional results

We continue to suppose that X and hence Z is a nonrandom matrix and that Z has full rank. The least squares model has $Y = Z\beta + \varepsilon$. Then

$$\hat{\beta} = (Z'Z)^{-1}Z'Y = (Z'Z)^{-1}(Z'Z\beta + \varepsilon) = \beta + (Z'Z)^{-1}Z'\varepsilon. \tag{2.9}$$

The only randomness in the right side of (2.9) comes from ε. This makes it easy to work out the mean and variance of $\hat{\beta}$ in the fixed x.

Lemma 2.1. *If $Y = Z\beta + \varepsilon$ where Z is nonrandom and $Z'Z$ is invertible and $E(\varepsilon) = 0$, then*

$$E(\hat{\beta}) = \beta. \tag{2.10}$$

Proof. $E(\hat{\beta}) = \beta + (Z'Z)^{-1}Z'E(\varepsilon) = \beta$. □

The least squares estimates $\hat{\beta}$ are unbiased for β as long as ε has mean zero. Lemma 2.1 does not require normally distributed errors. It does not even make any assumptions about $\mathrm{var}(\varepsilon)$. To study the variance of $\hat{\beta}$ we will need assumptions on $\mathrm{var}(\varepsilon)$, but not on its mean.

Lemma 2.2. *If $Y = Z\beta + \varepsilon$ where Z is nonrandom and $Z'Z$ is invertible and $\mathrm{var}(\varepsilon) = \sigma^2 I$, for $0 \le \sigma^2 < \infty$, then*

$$\mathrm{var}(\hat{\beta}) = (Z'Z)^{-1}\sigma^2. \tag{2.11}$$

Proof. Using equation (2.9),

$$\begin{aligned}\mathrm{var}(\hat{\beta}) &= (Z'Z)^{-1}Z'\mathrm{var}(\varepsilon)Z(Z'Z)^{-1} \\ &= (Z'Z)^{-1}Z'\big(\sigma^2 I\big)Z(Z'Z)^{-1} \\ &= (Z'Z)^{-1}\sigma^2, \end{aligned} \tag{2.12}$$

after a particularly nice cancellation. □

We will look at and interpret equation (2.12) for many specific linear models. For now we notice that it holds without regard to whether ε is normally distributed or even whether it has mean 0.

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 $\hat{\varepsilon}$.

Lemma 2.3. *Under the conditions of Lemma 2.2, $E(\hat{\varepsilon}'\hat{\varepsilon}) = (n-p)\sigma^2$. For $p < n$ the estimate*

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - z_i'\hat{\beta})^2 \tag{2.13}$$

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

Proof. Recall that $\hat{\varepsilon} = (I-H)\varepsilon$. Therefore $E(\hat{\varepsilon}'\hat{\varepsilon}) = E(\varepsilon'(I-H)\varepsilon) = \mathrm{tr}((I-H)I\sigma^2) = (n-p)\sigma^2$. Finally $s^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n-p)$. □

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

Theorem 2.3. *Suppose that Z is an n by p matrix of real values, $Z'Z$ is invertible, and $Y = Z\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2 I)$. Then*

$$\begin{aligned}\hat{\beta} &\sim N(\beta, (Z'Z)^{-1}\sigma^2), \\ \hat{y} &\sim N(Z\beta, H\sigma^2), \\ \hat{\varepsilon} &\sim N(0, (I-H)\sigma^2),\end{aligned}$$

where $H = Z(Z'Z)^{-1}Z'$. Furthermore, $\hat{\varepsilon}$ is independent of $\hat{\beta}$ and of $\hat{y}$.

Proof. Consider the vector

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

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

The expected value of v is

$$E(v) = \begin{pmatrix} (Z'Z)^{-1}Z' \\ H \\ I-H \end{pmatrix} Z\beta = \begin{pmatrix} \beta \\ HZ\beta \\ (I-H)Z\beta \end{pmatrix} = \begin{pmatrix} \beta \\ Z\beta \\ 0 \end{pmatrix}$$

because $HZ = Z$, establishing the means listed above for $\hat\beta$, $\hat y$ and $\hat\varepsilon$.

The variance of $\hat\beta$ is $(Z'Z)^{-1}\sigma^2$ by Lemma 2.2. The variance of $\hat y$ is $\text{var}(\hat y) = H\text{var}(Y)H' = H(\sigma^2 I)H = H\sigma^2$ because H is symmetric and idempotent. Similarly $\hat\varepsilon = (I-H)(Z\beta+\varepsilon) = (I-H)\varepsilon$ and so $\text{var}(\hat\varepsilon) = (I-H)(\sigma^2 I)(I-H)' = (I-H)\sigma^2$ because $I-H$ is symmetric and idempotent.

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

$$\text{cov}(\hat\beta, \hat\varepsilon) = (Z'Z)^{-1}Z'(\sigma^2 I)(I-H) = (Z'Z)^{-1}(Z-HZ)'\sigma^2 = 0$$

because $Z = HZ$. Therefore $\hat\beta$ and $\hat\varepsilon$ are uncorrelated and hence independent because v has a normal distribution. Similarly $\text{cov}(\hat y, \hat\varepsilon) = H(I-H)'\sigma^2 = (H-HH')\sigma^2 = 0$ establishing the second and final independence claim. □

We can glean some insight about the hat matrix from Theorem 2.3. First because $\text{var}(\hat y_i) = \sigma^2 H_{ii}$ we have $H_{ii} \ge 0$. Next because $\text{var}(\hat\varepsilon) = \sigma^2(1-H_{ii})$ we have $1 - H_{ii} \ge 0$. Therefore $0 \le H_{ii} \le 1$ holds for $i = 1, \dots, n$.

Theorem 2.4. *Assume the conditions of Theorem 2.3, and also that $p < n$. Let $s^2 = \frac{1}{n-p}\sum_{i=1}^n (Y_i - z_i'\hat\beta)^2$. Then*

$$(n-p)s^2 \sim \sigma^2\chi^2_{(n-p)}. \tag{2.14}$$

Proof. First

$$(n-p)s^2 = (Y - Z\hat\beta)'(Y - Z\hat\beta) = Y'(I-H)Y = \varepsilon'(I-H)\varepsilon.$$

The matrix $I-H$ can be written $I-H = P\Lambda P'$ where P is orthogonal and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$. From $I-H = (I-H)^2$ we get $\lambda_i \in \{0,1\}$. Let $\widetilde\epsilon = P'\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then $\varepsilon'(I-H)\varepsilon = \widetilde\epsilon'\Lambda\widetilde\epsilon \sim \sigma^2\chi^2_{(k)}$ where k is the number of $\lambda_i = 1$, that is $k = \text{tr}(I-H)$. Therefore

$$k = n - \text{tr}(H) = n - \text{tr}(Z(Z'Z)^{-1}Z') = n - \text{tr}((Z'Z)^{-1}Z'Z) = n - \text{tr}(I_p) = n - p.$$

□

Equation (2.14) 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 $c\beta$ where c is a $p \times 1$ row vector. Taking $c = (0, \dots, 0, 1, 0, \dots 0)$ with the 1 in the j'th place gives $c\beta = \beta_j$. If for instance $c\beta_j = 0$, then we question whether feature j helps us predict Y. 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 c. Taking $c = z_0'$ makes $c\beta$ the expected value of Y when the feature vector is set to z_0 and taking $c = \widetilde{z}_0' - z_0'$ lets us study the difference between $E(Y)$ at features $\widetilde{z}_0$ and z_0.

For any such vector c, Theorem 2.3 implies that $c\hat\beta \sim N(c\beta, c(Z'Z)^{-1}c'\sigma^2)$. We ordinarily don't know σ^2 but can estimate it via s^2 from (2.13). In later chapters we will test whether $c\beta$ takes a specific value $c\beta_0$, and form confidence intervals for $c\beta$ using Theorem 2.5.

Theorem 2.5. *Suppose that $Y \sim N(Z\beta, \sigma^2 I)$ where Z is an $n \times p$ matrix, $Z'Z$ is invertible, and $\sigma > 0$. Let $\hat\beta = (Z'Z)^{-1}Z'Y$ and $s^2 = \frac{1}{n-p}(Y - Z\hat\beta)'(Y - Z\hat\beta)$. Then for any nonzero $1 \times p$ (row) vector c,*

$$\frac{c\hat\beta - c\beta}{s\sqrt{c(Z'Z)^{-1}c'}} \sim t_{(n-p)}. \tag{2.15}$$

Proof. Let $U = (c\hat\beta - c\beta)/(\sigma\sqrt{c(Z'Z)^{-1}c'})$. Then $U \sim N(0,1)$. 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 $\hat\varepsilon$. Therefore U and V are independent. Finally

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

by the definition of the t distribution. □

Theorem 2.5 is the basis for t tests and related confidence intervals for linear models. The recipe is to take the estimate $c\hat\beta$ subtract the hypothesized parameter value $c\beta$ and then divide by an estimate of the standard deviation of $c\hat\beta$. 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 2.5 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 2.6 is an r dimensional generalization of Theorem 2.5.

Theorem 2.6. *Suppose that $Y \sim N(Z\beta, \sigma^2 I)$ where Z is an $n \times p$ matrix, $Z'Z$ is invertible, and $\sigma > 0$. Let $\hat\beta = (Z'Z)^{-1}Z'Y$ and $s^2 = \frac{1}{n-p}(Y - Z\hat\beta)'(Y - Z\hat\beta)$.*

Then for any $r \times p$ (row) matrix C with linearly independent rows,

$$\frac{1}{r}\,(\hat\beta-\beta)'C'\big[C(Z'Z)^{-1}C'\big]^{-1}C(\hat\beta-\beta)/s^2 \sim F_{r,n-p}. \tag{2.16}$$

Proof. Let $U = C(\hat\beta-\beta)/\sigma$. Then $U \sim N(0,\Sigma)$ where $\Sigma = C(Z'Z)^{-1}C'$ is a nonsingular $r \times r$ matrix, and so $U'\Sigma^{-1}U \sim \chi^2_{(r)}$. Let $V = s^2/\sigma^2$, so $V \sim \chi^2_{(n-p)}/(n-p)$. The left side of (2.16) is $(1/r)U'\Sigma^{-1}U \sim \chi^2_{(r)}/r$ divided by $V/(n-p) \sim \chi^2_{(n-p)}/(n-p)$. These two χ^2 random variables are independent because $\hat\beta$ is independent of s^2. □

To use Theorem 2.6 we formulate an hypothesis $H_0 : C\beta = C\beta_0$, plug $\beta = \beta_0$ into the left side of equation (2.16) and reject H_0 at the level α if the result is larger that $F^{1-\alpha}_{r,n-p}$.

If we put $r = 1$ in (2.16) and write C as c then it simplifies to

$$\frac{(c\hat\beta - c\beta)^2}{s^2\,c(Z'Z)^{-1}c'} \sim F_{1,n-p}.$$

We might have expected to recover (2.15). In fact, the results are strongly related. The statistic on the left of (2.16) is the square of the one on the left of (2.15). The $F_{1,n-p}$ distribution on the right of (2.16) is the square of the $t_{(n-p)}$ distribution on the right of (2.15). The difference is only that (2.15) explicitly takes account of the sign of $c\hat\beta - c\beta$ while (2.16) does not.

2.5 R^2 and the extra sum of squares

Constructing the matrix C and plugging it into Theorem 2.6 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 $E(Y) = \widetilde Z\gamma$ where $\widetilde Z$ has $q < r$ of the columns of Z 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 Z and then on $\widetilde Z$, and from the resulting sums of squared errors we can form a test of the submodel.

Let us call $E(Y) = Z\beta$ the full model and $E(Y) = \widetilde Z\gamma$ 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 $\mathrm{SS}_{\mathrm{FULL}} \le \mathrm{SS}_{\mathrm{SUB}}$. 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 2.7. *Suppose that $Y \sim N(Z\beta, \sigma^2 I)$ where Z is an $n \times p$ matrix, $Z'Z$ is invertible, and $\sigma > 0$. Let $\widetilde Z$ be an $n \times q$ submatrix of Z with $q < p$.*

Let the full model have least squares estimate $\hat\beta = (Z'Z)^{-1}Z'Y$, and sum of squared errors $\mathrm{SS}_{\mathrm{FULL}} = (Y - Z\hat\beta)'(Y - Z\hat\beta)$. Similarly let the submodel have *least squares estimate* $\hat\gamma = (\widetilde{Z}'\widetilde{Z})^{-1}\widetilde{Z}'Y$ *and sum of squared errors* $\mathrm{SS}_{\mathrm{SUB}} = (Y - \widetilde{Z}\hat\gamma)'(Y - \widetilde{Z}\hat\gamma)$.

If the submodel $E(Y) = \widetilde{Z}\gamma$ *holds for some* $\gamma \in \mathbb{R}^q$ *then*

$$\frac{\frac{1}{p-q}\bigl(\mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}}\bigr)}{\frac{1}{n-p}\,\mathrm{SS}_{\mathrm{FULL}}} \sim F_{p-q,n-p}. \tag{2.17}$$

Proof. Let $H = Z(Z'Z)^{-1}Z'$ and $\widetilde{H} = \widetilde{Z}(\widetilde{Z}'\widetilde{Z})^{-1}\widetilde{Z}'$ be the hat matrices for the full model and submodel respectively. Then $\mathrm{SS}_{\mathrm{FULL}} = Y'(I-H)Y \sim \sigma^2\chi^2_{\mathrm{rank}(I-H)}$ because $I - H$ is symmetric and idempotent. The extra sum of squares is $\mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}} = Y'(I-\widetilde{H})Y - Y'(I-H)Y = Y'(H-\widetilde{H})Y$.

The matrix $H - \widetilde{H}$ is symmetric. For any $Y \in \mathbb{R}^n$ we get $H\widetilde{H}Y = \widetilde{H}Y$ because $\widetilde{H}Y$ is in the column space of $\widetilde{Z}$ and hence also in that of Z. Therefore $H\widetilde{H} = \widetilde{H}$ and by transposition $\widetilde{H}H = \widetilde{H}$ too. Therefore $(H-\widetilde{H})(H-\widetilde{H}) = H - H\widetilde{H} - \widetilde{H}H + \widetilde{H} = H - \widetilde{H}$. Thus $H - \widetilde{H}$ is idempotent too. Therefore $\mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}} \sim \sigma^2\chi^2_{\mathrm{rank}(H-\widetilde{H})}$.

The matrices H and $\widetilde{H} - H$ are orthogonal. It follows that HY and $(\widetilde{H}-H)Y$ are independent and therefore $\mathrm{SS}_{\mathrm{FULL}}$ is independent of $\mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}}$. We have shown that

$$\frac{\frac{1}{\mathrm{rank}(H-\widetilde{H})}\,(\mathrm{SS}_{\mathrm{SUB}} - \mathrm{SS}_{\mathrm{FULL}})}{\frac{1}{\mathrm{rank}(I-H)}\,\mathrm{SS}_{\mathrm{FULL}}} \sim F_{\mathrm{rank}(H-\widetilde{H}),\mathrm{rank}(I-H)}.$$

To complete the proof we note that when Z has full rank p then so does H. Also $\widetilde{Z}$ has rank q and so does $\widetilde{H}$. For symmetric idempotent matrices like $I - H$ and $H - \widetilde{H}$ their rank equals their trace. Then $\mathrm{tr}(I-H) = n - \mathrm{tr}(H) = n - p$ and $\mathrm{tr}(H-\widetilde{H}) = \mathrm{tr}(H) - \mathrm{tr}(\widetilde{H}) = p - q$ to complete the proof. □

There is a geometric interpretation to equation (2.17). If $Y = \widetilde{Z}\gamma + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2 I)$ then we can write Y as $Y_{\mathrm{SUB}} + Y_{\mathrm{FULL}} + Y_{\mathrm{RES}}$ where Y_{SUB} is in a q dimensional space spanned by the columns of $\widetilde{Z}$, Y_{FULL} is in the $p - q$ dimensional space of vectors orthogonal to $\widetilde{Z}$ but in the column space of Z and Y_{RES} is in the $n - p$ dimensional space of vectors orthogonal to Z. A spherical Gaussian vector like ε shows no preference for any of these spaces. The mean squared norm of it's 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_{FULL} 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. Then $\mathrm{SS}_{\mathrm{SUB}} = \sum_{i=1}^n (y_i - \bar{y})^2$. The ANOVA decomposition in equation (2.7) has $\mathrm{SS}_{\mathrm{SUB}} = \mathrm{SS}_{\mathrm{FULL}} + \mathrm{SS}_{\mathrm{RES}}$. The F test then rejects the submodel in favor of the full model when R^2 is large enough.

2.6 Random predictors

Here we consider the 'random X random Y' setting. A conditioning argument says that we can treat each X_i as if it were fixed quantity equal to the observed value x_i and then hints that this analysis is better than a random X analysis even when the X_i are random. It is certainly a simpler analysis. A reader who accepts the conditioning argument may want to skip this section. A reader who is curious to see what happens with random X should read on.

If (X_i, Y_i) are IID pairs then so are (Z_i, Y_i). We will need the expected value of Y given that $X = x$. Call it $\mu(x)$. Similarly define $\sigma^2(x) = \mathrm{var}(Y \mid X = x)$.

The least squares value of β is the one that minimizes $E((Y_1 - Z_1\beta)^2)$ where we've picked observation $i = 1$ for definiteness. Using very similar methods to the least squares derivation we get

$$E(Z_1(Y_1 - Z_1'\beta)) = 0$$

as an equation defining β. In other words we have a population version of the normal equations. The error $Y_i - Z_i'\beta$ is orthogonal to the feature vector Z_i in that their product has expectation zero.

As before we can rearrange the normal equations to get

$$\beta = E(Z_1Z_1')^{-1}E(Z_1Y_1),$$

for the case where $E(Z_1Z_1')$ is invertible. The least squares estimator can, for nonsingular $Z'Z$, be written

$$\hat\beta = \left(\frac{1}{n}\sum_{i=1}^n Z_iZ_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^n Z_iY_i.$$

In other words the least squares estimator is obtained by plugging in estimates of the numerator and denominator for β.

Now $E(\hat\beta) = E((Z'Z)^{-1}Z'Y)$. When X is random then so is Z. The expectation does not simplify to $E((Z'Z)^{-1}) \times E(Z'Y)$ because expectations of products are not generally the corresponding products of expectations. Even if it did $E((Z'Z)^{-1})$ does not simplify to $E(Z_1Z_1')^{-1}/n$ because an expected inverse is not the inverse of the expectation. We cannot cancel our way to the answer we want.

We will condition on $X_1, \dots, X_n$. Let $\mathcal{X}$ represent $X_1, \dots, X_n$. Conditionally on $\mathcal{X}$ the X_i as well as Z and the Z_i are fixed. Conditioning on $\mathcal{X}$ and then working purely formally, we get

$$\begin{aligned} E(\hat\beta) &= E(E(\hat\beta \mid \mathcal{X})) \\ &= E((Z'Z)^{-1}Z'E(Y \mid \mathcal{X})) \\ &= E((Z'Z)^{-1}Z'\mu(X)). \end{aligned}$$

where $\mu(X)$ is the random vector with i'th element $E(Y \mid X = X_i)$.

So far we have not connected the function $\mu(x)$ to the linear regression model. To make this case comparable to the fixed x random Y case lets suppose that $\mu(x) = z'\beta$ where $z = \phi(x)$ and also that $\sigma^2(x) = \sigma^2$ is constant. From the first of these assumptions we find

$$E(\hat{\beta}) = E((Z'Z)^{-1}Z'Z\beta) \tag{2.18}$$

Equation (2.18) is pretty close to $E(\hat{\beta}) = \beta$. If $Z'Z$ has probability 0 of being singular then $E(\hat{\beta}) = \beta$ as required. Even if $Z'Z$ is very nearly but not quite singular, equation (2.18) does not imply trouble. For example even if $Z'Z = \begin{pmatrix} 1 & 0 \\ 0 & \epsilon \end{pmatrix}$ for small $\epsilon > 0$ then $(Z'Z)^{-1}Z'Z\beta = \beta$.

The biggest worry is that $Z'Z$ might sometimes be singular, or numerically so close to singular than an algorithm treats it as singular. It can be singular even when $E(Z_iZ_i')$ is not singular. Suppose for example that $X = 1$ with probability 1/2 and 0 with probability 1/2 and that the regression has features $\phi(X) = (1, X)'$. Then with probabilty 2^{-n} all the $x_i = 1$ and $Z'Z$ is singular. The x_i might all be 0 as well, so the probability of a singular $Z'Z$ is 2^{1-n}. If the X_i's are discrete and IID there is always some nonzero probability, though perhaps a very tiny one, that $Z'Z$ is singular. Then $\hat{\beta}$ exists with probability $1-2^{1-n}$ and fails to exist with probability 2^{1-n}. If that happened we'd probably drop feature 2, and implicitly or explicitly use $\hat{\beta} = (\bar{Y}, 0)'$. Then we have a bias that is exponentially small in n. It may be a nuisance to have an algorithm that sometimes behaves qualitatively differently, but such a small bias in and of itself will seldom have practical importance.

The upshot is that when the linear model properly matches $\mu(x)$ and $Z'Z$ is never singular, then bias is either absent or negligible for the random X random Y setting.

Let's suppose that the possibility of singular $Z'Z$ can be neglected. Then keeping in mind that we have assumed $\text{var}(Y \mid X = x) = \sigma^2$ for all x,

$$\begin{aligned} \text{var}(\hat{\beta}) &= \text{var}(E(\hat{\beta} \mid \mathcal{X})) + E(\text{var}(\hat{\beta} \mid \mathcal{X})) \\ &= E(\text{var}(\hat{\beta} \mid \mathcal{X})) \\ &= E\left((Z'Z)^{-1}Z'(\sigma^2 I)Z(Z'Z)^{-1}\right) \\ &= E((Z'Z)^{-1})\sigma^2. \end{aligned} \tag{2.19}$$

The variance is similar to the expression $(Z'Z)^{-1}\sigma^2$ that we get in the 'fixed X random Y' case.

The variance in the random X case uses the expected value of $(Z'Z)^{-1}$ while in the fixed case we got the actual observed value of $(Z'Z)^{-1}$. A 'larger' matrix $Z'Z$ gives us a more accurate estimate of β in the fixed X case. In the fixed X setting we're using the $Z'Z$ we got rather than taking account of other values it could have taken. It seems to be better to take account of the actual accuracy we got than to factor in accuracy levels we might have had.

An analogy due to David Cox likens this to weighing an object on either an accurate scale or a very poor one. Suppose that have an object with weight ω.

We toss a fair coin, and if it comes up heads we use the good scale and get a value $Y \sim N(\omega, 1)$. If the coin comes up tails we get $Y \sim N(\omega, 100)$. So if we get heads then we could take a 95% confidence interval to be $Y \pm 1.96$, but then if we get tails we take $Y \pm 19.6$ instead. Or we could ignore the coin and use $Y \pm 16.45$ because $\Pr(|Y - \omega| \le 16.45) \doteq 0.05$ in this case. It seems pretty clear that if we know how the coin toss came out that we should take account of it and not use $Y \pm 16.45$.

The matrix $Z'Z$ is analogous to the coin. It affects the precision of our answer but we assume it does not give us information on β. A quantity that affects the precision of our answers without carrying information on them is said to be "ancillary", meaning secondary or subordinate.

In some settings n itself is random. The ancillarity argument says that we should prefer to use the actual sample size we got rather than take account of others we might have gotten but didn't.

The weighing example and random sample size examples are constructed to be extreme to make the point. If the scales had variances 1 and 1.001 we might not bother so much with conditioning on the coin. If sampling fluctuations in $Z'Z$ are not so large then might prefer to treat it as fixed most of the time but switch if treating it as random offered some advantage. For instance some bootstrap and cross-validatory analyses are simpler to present with random X's.

2.7 Rank deficient features

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

Theorem 2.8. *If $Z'Z$ is not invertible then one column of Z is a linear combination of the others.*

Proof. If $Z'Z$ is not invertible then $Z'Zv = 0$ for some nonzero $v \in \mathbb{R}^p$. Then $v'Z'Zv = 0$ too. But $v'Z'Zv = \|Zv\|^2$, and so $Z'v = 0$. Let v_j be a nonzero element of v. Then $Z_{\bullet j} = \sum_{k \ne j} Z_{\bullet k} v_k / v_j$. □

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 Celcius. Recalling that "$C = 1.8F + 32$" we will find linear dependence in Z, 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 Z will be the sum the regional ones. Finally a row of Z might take the form $(1, 1_{\mathrm{M}}, 1_{\mathrm{F}}, \dots)'$ where 1_{M} and 1_{F} are indicator functions of male and female subjects respectively and the dots represent additional features.

A singular matrix Z 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 $Z'Z$ 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 in Chapter **??** on 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 gauranteed to have a non-invertible $Z'Z$ 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 $\mathcal{M} = \{Z\beta \mid \beta \in \mathbb{R}^p\}$ is an r dimensional subset 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 $Z_{i1} = 1$, and that $Z_{i2} = F_i$ and $Z_{i3} = C_i$ are temperatures in degrees Fahrenheit and Celcius of data points $i = 1, \dots, n$. Then for any $t \in \mathbb{R}$,

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

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

The problem with rank deficient Z is not, as we might have feared, that there is no solution to $Z'Z\hat{\beta} = Z'y$. 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 $\beta'\beta$ subject to the constraint $Z'Z\beta = Z'y$ and we call it $\hat{\beta}$. We'll see how to do this numerically in Chapter 2.9.

When $Z'Z$ is singular we may well find that some linear combinations $c\hat{\beta}$ are uniquely determined by the normal equations $(Z'Z)\hat{\beta} = Z'y$ even though $\hat{\beta}$ itself is not. If we can confine our study of β to such linear combinations $c\beta$ then singularity of $Z'Z$ 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 $\hat{y} = Z\hat{\beta}$ is determined by the least squares conditions. It follows that each component $\hat{y}_i = z_i'\hat{\beta}$ is also determined. In other words, if $c = z_i'$ for one of our sampled feature vectors, then $c\hat{\beta}$ is determined.

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

Definition 2.1 (Estimability)**.** *The linear combination $c\beta$ is estimable if there exists a linear combination $\lambda'Y = \sum_{i=1}^n \lambda_i Y_i$ of the responses for which $E(\lambda'Y) = c\beta$ holds for any $\beta \in \mathbb{R}^p$.*

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

Theorem 2.9. *The following are equivalent:*

1. *$c\beta$ is estimable.*
2. *$c = \gamma' Z$ for some $\gamma \in \mathbb{R}^n$.*
3. *If $Z'Z\hat{\beta} = Z'Y$ and $Z'Z\widetilde{\beta} = Z'Y$ then $c\hat{\beta} = c\widetilde{\beta}$.*

Proof. [Some algebra goes here, or possibly it becomes an exercise.] □

2.8 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 normally 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 ε 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

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - z_i'\hat{\beta})^2. \tag{2.20}$$

To derive this MLE, we suppose that $Y \sim N(Z\beta, \sigma^2 I)$ for fixed and full rank Z. The probability of observing $Y_i = y_i$, for $i = 1, \dots, n$, is of course zero, which is not helpful. But our observations were only measured to finite precision and we might interpret the observation y_i as corresponding to the event $y_i - \Delta \le Y_i \le y_i + \Delta$ where $\Delta > 0$ is our measuring accuracy. We don't have to actually know Δ, but it should always be small compared to σ. The probability of observing Y in a small n dimensional box around y is

$$\Pr\Bigl(|Y_i - y_i| \le \Delta, \quad 1 \le i \le n \Bigr) \doteq \frac{(2\Delta)^n}{(2\pi)^{n/2}\sigma^n} \exp\Bigl(-\frac{1}{2\sigma^2}(y - Z\beta)'(y - Z\beta) \Bigr). \tag{2.21}$$

For any $0 < \sigma < \infty$ the β which maximizes (2.21) is the one that minimizes the sum of squares $(y - Z\beta)'(y - Z\beta)$. So the MLE is the least squares estimate and can be found by solving the normal equations.

Next we find the MLE of σ. The logarithm of the right side of (2.21) is

$$-n \log \Delta - \frac{n}{2}\log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2}(y - Z\beta)'(y - Z\beta). \tag{2.22}$$

Differentiating (2.22) with respect to σ yields

$$-\frac{n}{\sigma} + \frac{1}{2\sigma^3}(y - Z\beta)'(y - Z\beta), \tag{2.23}$$

which vanishes if and only if

$$\sigma^2 = \frac{1}{n}(y - Z\beta)'(y - Z\beta). \tag{2.24}$$

Solving (2.24) with $\beta = \hat\beta$ gives the MLE $\hat\sigma^2$ from (2.20).

It is also necessary to investigate the second derivative of the log likelihood with respect to σ at the point $\sigma = \hat\sigma$. The value there is $-2n/\sigma^2 < 0$, and so (2.20) really is the MLE and is not a minimum.

In the case where $y = Z\hat\beta$ 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 $\lambda' Y$ is a linear unbiased estimator for $c\beta$ if $E(\lambda' Y) = c\beta$ holds for all β. Let's agree to define the best linear unbiased estimator of $c\beta$ as the one that attains the minimum variance. It need not be unique. The Gauss-Markov theorem states that least squares provides best linear unbiased estimators.

Theorem 2.10 (Gauss-Markov)**.** *Let $Y = Z\beta + \varepsilon$ where Z is a nonrandom $n \times p$ matrix, β is an unknown point in $\mathbb{R}^p$ and ε is a random vector with mean 0 and variance matrix $\sigma^2 I_n$. Let $c\beta$ be estimable and let $\hat\beta$ be a least squares estimate. Then $c\hat\beta$ is a best linear unbiased estimate of $c\beta$.*

Proof. [Full rank case] We prove the theorem for the full rank case with $Z'Z$ invertible. We already know that $c\hat\beta$ is unbiased and linear. Suppose that $E(\lambda' Y) = c\beta$, and so $\lambda' Z = c$. Then

$$\begin{aligned}\mathrm{var}(\lambda' Y) &= \mathrm{var}(c\hat\beta + (\lambda' Y - c\hat\beta))\\ &= \mathrm{var}(c\hat\beta) + \mathrm{var}(\lambda' Y - c\hat\beta) + 2\mathrm{cov}(c\hat\beta, \lambda' Y - c\hat\beta).\end{aligned}$$

Now

$$\begin{aligned}\mathrm{cov}(c\hat\beta, \lambda' Y - c\hat\beta) &= \mathrm{cov}(c(Z'Z)^{-1}Z'Y, (\lambda' - c(Z'Z)^{-1}Z')Y)\\ &= c(Z'Z)^{-1}Z'(\sigma^2 I)(\lambda - Z(Z'Z)^{-1}c')\\ &= \sigma^2\Big(c(Z'Z)^{-1}Z'\lambda - c(Z'Z)^{-1}c'\Big)\\ &= 0\end{aligned}$$

because $Z'\lambda = c'$. Therefore $\mathrm{var}(\lambda' Y) = \mathrm{var}(c\hat\beta) + \mathrm{var}(\lambda' Y - c\hat\beta) \geq \mathrm{var}(c\hat\beta)$. □

For rank deficient Z, a similar proof can be made but takes a bit more care. See for example Christiansen (xxxx). Intuitively we might expect the Gauss-Markov theorem to hold up for estimable functions in the rank deficient case. We could drop columns of Z until $Z'Z$ 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.

2.9 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.

This section can be skipped on first reading, especially for readers who are willing to assume that their least squares software with its default settings simply works without any attention on their part.

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 $Z'Z$ and $Z'y$ and then $O(p^3)$ to solve $Z'Z\hat{\beta} = Z'y$. We will consider more sophisticated least squares algorithms based on the QR decomposition and the singular value decomposition. These also cost $O(np^2+p^3)$ with different implied constants.

Solving the normal equations

Given an n by p matrix Z and $Y \in \mathbb{R}^n$ the least squares problem is to solve $(Z'Z)\hat{\beta} = Z'Y$ 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 Z has full rank then $Z'Z$ is positive definite. It can be factored by the Cholesky decomposition into $Z'Z = G'G$ where G is a p by p upper triangular matrix. Then to solve $G'G\hat{\beta} = Z'Y$ we solve two triangular systems of equations: first we solve $G'u = Z'Y$ 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 decompositions is the best way to solve least squares problems. They are numerically unstable if

$Z'Z$ is nearly singular and of course they fail if $Z'Z$ is actually singular. The QR decomposition is more reliable than solving the normal equations, when Z has full rank. If Z does not have full rank, then the singular value decomposition can be used to pick out one of the least squares solutions, or even to parameterize the entire solution set.

QR decomposition

A better approach is to employ a QR decomposition of Z. In such a decomposition we write $Z = QR$ where Q is an n by n orthogonal matrix whose and R is an n by p upper triangular matrix. When as usual $n > p$ then the last $n - p$ rows of R are zero. Then we don't really need the last $n-p$ columns of Q. Thus there are two QR decompositions

$$Z = QR = \begin{pmatrix} \widetilde{Q} & Q^* \end{pmatrix} \begin{pmatrix} \widetilde{R} \\ 0 \end{pmatrix} = \widetilde{Q}\widetilde{R},$$

where $\widetilde{Q}$ is an $n \times p$ matrix with orthonormal columns and $\widetilde{R}$ is a $p \times p$ upper triangular matrix.

In this section we'll verify that a QR decomposition exists and see how to find it. Then we'll see how to use it in least squares problems. Finally we'll look into why using the QR decomposition should be better than solving the normal equations directly. Implementation of the QR decomposition requires care about storage and bookkeeping. For pseudo-code see Golub and van Loan (xxxx).

The existence of a QR decomposition follows directly from the Gram-Schmidt orthogonalization process. That process is essentially a regression of each column of Z onto the previous ones. Since we'll be using QR to do regression, it is almost circular to use regression to define QR. We'll use Householder reflections instead. They're interesting in their own right, and they also underly many implementations of the QR decomposition.

Suppose that u is a unit vector. Then $H = H(u) = I - 2uu'$ is a Householder reflection matrix. It is a symmetric orthogonal matrix. If v is a vector parallel to u then $Hv = -v$. If v is a vector orthogonal to u then $Hv = v$. This is why H is called a reflection. It reflects u onto $-u$ while leaving any vector orthogonal to u unchanged.

Suppose that v and w are distinct nonzero vectors of equal length. Then we can find a unit vector u so that $H(u)$ reflects v onto w. That is $(I - 2uu')v = w$. Then we'll be able to reflect the first column of Z onto a vector that has zeros below the first entry.

The vector we need to make a reflection of v onto u is $u = (w - v)/\|w - v\|$. First notice that $(w + v)/2$ is orthogonal to u and $(w - v)/2$ is parallel to u. Therefore

$$Hv = H\Big(\frac{w-v}{2} + \frac{w+v}{2}\Big) = \frac{w-v}{2} - \frac{w+v}{2} = w.$$

Let $Q_1 = H_1$ be an n by n Householder reflection matrix taking $Z_{\bullet 1}$ onto the vector $(R_{11}, 0, 0, \dots, 0)'$ where $R_{11} = \pm\|Z_{\bullet 1}\|$. Then

$$Q_1 Z = \left(\begin{array}{c|ccc} R_{11} & | & & | \\ 0 & & & \\ \vdots & H_1 Z_{\bullet 2} & \cdots & H_1 Z_{\bullet p} \\ 0 & | & & | \end{array}\right).$$

We have managed to put zeros below the diagonal in the first column on the right side. In practice, the choice between $\|Z_{\bullet 1}\|$ and $-\|Z_{\bullet 1}\|$ is made based on which will be lead to a more numerically stable answer. See Golub and van Loan (xxxx) for details.

We will use matrices Q_j taking the form

$$Q_j = \begin{pmatrix} I_{j-1} & 0 \\ 0 & H_j \end{pmatrix}$$

where, for $j = 2, \dots, p$, the H_j are $n-j+1$ by $n-j+1$ Householder reflection matrices. For a vector v the product $Q_j v$ leaves the top $j-1$ components of v unchanged while rotating the rest of v. We can choose a rotation so that the last $n-j$ components of $Q_j v$ are zero.

After carefully choosing the matrices H_j we obtain

$$Q_p Q_{p-1} \cdots Q_2 Q_1 Z = \begin{pmatrix} R_{11} & R_{12} & \dots & R_{1p} \\ 0 & R_{22} & \dots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & R_{pp} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \tag{2.25}$$

giving us $Z = QR$ where $Q = (Q_p Q_{p-1} \dots Q_2 Q_1)'$ has orthonormal columns, and R is the upper triangular matrix on the right of (2.25). The leftmost p columns of Q are $\widetilde{Q}$ and the top p rows of R are $\widetilde{R}$.

Given that $Z = \widetilde{Q}\widetilde{R}$ we can plug it into the normal equations $Z'Z\hat{\beta} = Z'Y$ and get $\widetilde{R}'\widetilde{R}\hat{\beta} = \widetilde{R}'\widetilde{Q}'Y$. When Z has full rank, then $\widetilde{R}'$ is a $p \times p$ full rank matrix too, and so it is invertible. Any solution $\hat{\beta}$ to $\widetilde{R}'(\widetilde{R}\hat{\beta} - \widetilde{Q}'Y) = 0$ also solves $\widetilde{R}\hat{\beta} - \widetilde{Q}'Y = 0$.

That is, given $Z = \widetilde{Q}\widetilde{R}$ we find $\widetilde{Y} = \widetilde{Q}'Y$ and then solve

$$\begin{pmatrix} R_{11} & R_{12} & \dots & R_{1p} \\ 0 & R_{22} & \dots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & R_{pp} \end{pmatrix} \hat{\beta} = \widetilde{Y}.$$

This latter is a triangular system, so it is easy to solve for $\hat{\beta}_p$ then $\hat{\beta}_{p-1}$ and so on, ending with $\hat{\beta}_1$.

We are sure to run into trouble if $\widetilde{R}$ is singular. At least that trouble will be detected because $\widetilde{R}$ is singular if and only if $R_{ii} = 0$ for some $1 \le i \le p$. At that point we'll notice that we're trying to reflect a vector of length zero. Also if some R_{ii} is near zero then numerical near singularity of Z is detected.

If we solve the normal equations directly, then we are solving $\widetilde{R}'\widetilde{R}\hat{\beta} = Z'Y$. The matrix $\widetilde{R}'\widetilde{R}$ is less favorable to solve with than is $\widetilde{R}$. A precise definition requires the notion of matrix condition number. But roughly speaking, suppose that in solving the system $\widetilde{R}a = b$ for a roundoff errors in b get multiplied by a factor f. Solving the normal equations directly is like applying two such solutions one for $\widetilde{R}$ and one for $\widetilde{R}'$ and might then multiply error by f^2. Employing the QR decomposition instead of solving the normal equations directly roughly equivalent to doubling one's numerical precision (from single to double or from double to quadruple).

Singular value decomposition

Any n by p matrix Z may be written

$$Z = U\Sigma V' \tag{2.26}$$

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 (xxxx).

If we multiply the matrix Z by a column vector w we find that $Zw = U\Sigma V'w$ corresponds to rotating w by V' then scaling each component of $V'w$ by a corresponding diagonal element of Σ, and then rotating the result again. Rotate, stretch, and rotate is what happens to w when multiplied by Z and this holds for any Z. We use the term rotate here loosely, 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 u_i, the i'th column of V is v_i and the i'th diagonal element of Σ is σ_i. Then we can write

$$Z = \sum_{i=1}^{\min(n,p)} \sigma_i u_i v_i'$$

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

For the QR deccomposition we ignored the last $p - r$ columns of Q because the corresponding rows of R were zero. When $n > p$ we can make a similar simplification because the last $n - p$ rows of Σ are zero. That is we can also write

$$Z = \widetilde{U}\widetilde{\Sigma}V'$$

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

We begin by noting that the Euclidean length of the vector $y - Z\beta$ is preserved when it is multiplied by an orthogonal matrix such as U. Then

$$\begin{aligned}\|y - Z\beta\|^2 &= \|y - U\Sigma V\beta\|^2\\ &= \|U'y - \Sigma V'\beta\|^2\\ &= \|\widetilde{U}'y - \widetilde{\Sigma}V'\beta\|^2 + \|U^*y\|^2,\end{aligned}$$

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

$$\hat{\beta} = V\widetilde{\beta} = V\Sigma^+\widetilde{U}'Y. \tag{2.27}$$

When we want to simply pick one least squares solution out of many, to use as a default, then the one in (2.27) is a good choice. In practice the computed singular values have some roundoff error in them. Then one should set $\sigma_i^+ = 0$ when $\sigma_i \le \sigma_1 \times \epsilon$ where ϵ is a measure of numerical precision. Note that the threshold scales with the largest singular value. A matrix Z 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 $\sigma_i \le \sigma_1 \times \epsilon$ 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:

$$V\left(\Sigma^+ + \begin{pmatrix} 0 & 0\\ 0 & \mathrm{diag}(\tau)\end{pmatrix}\right)\widetilde{U}'Y, \quad \tau \in \mathbb{R}^k,$$

where $\mathrm{diag}(\tau)$ is a $k \times k$ diagonal matrix with elements of τ on the diagonal set above into the lower right block of a $p \times p$ matrix.

2.10 Special least squares systems

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

$$\hat{\beta}_j = \frac{\sum_{i=1}^n Z_{ij} y_i}{\sum_{i=1}^n Z_{ij}^2} = \frac{Z_{\bullet j}' y}{\|Z_{\bullet j}\|^2}, \quad j = 1, \dots, p,$$

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

Orthogonal predictors bring great simplification. The cost of computation is only $O(np)$. The variance of $\hat{\beta}$ is $\sigma^2 \mathrm{diag}(1/\|Z_{\bullet j}\|^2)$ 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 predictors x_i lie on a regular grid in one or two dimensions then we may construct some orthogonal features z_i from them.

Orthogonal polynomials

B **splines**

Fourier series

Haar Wavelets

2.11 Leave one out formula

In this section we explore what happens to a regression model when one data point is added or removed. We begin with the Sherman-Morrison formula. Suppose that A is an invertible $n \times n$ matrix and that u and v are n vectors with $1 + v'A^{-1}u \neq 0$. Then

$$(A + uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u}. \tag{2.28}$$

Equation (2.28) can be proved by multiplying the right hand side by $A + uv'$. See Lemma **??**. The condition $1 + v'A^{-1}u \neq 0$ is needed to avoid turning an invertible A into a singular $A + uv'$.

Suppose that we delete observation i from the regression. Then $Z'Z = \sum_{\ell=1}^n z_\ell z_\ell'$ is replaced by $(Z'Z)_{(-i)} = Z'Z - z_i z_i'$, using a subscript of $(-i)$ to denote removal of observation i. We can fit this into (2.28) by taking $u = z_i$ and $v = -z_i$. Then

$$\begin{aligned} (Z'Z)_{(-i)}^{-1} &= (Z'Z)^{-1} + \frac{(Z'Z)^{-1} z_i z_i' (Z'Z)^{-1}}{1 - z_i'(Z'Z)^{-1} z_i} \\ &= (Z'Z)^{-1} + \frac{(Z'Z)^{-1} z_i z_i' (Z'Z)^{-1}}{1 - H_{ii}} \end{aligned}$$

after recognizing the hat matrix diagonal from equation (2.8). We also find that $(Z'y)_{(-i)} = Z'y - z_iy_i$. Now

$$\begin{aligned}
\hat\beta_{(-i)} &= \Bigl((Z'Z)^{-1} + \frac{(Z'Z)^{-1}z_iz_i'(Z'Z)^{-1}}{1-H_{ii}}\Bigr)\Bigl(Z'Y - z_iy_i\Bigr)\\
&= \hat\beta - (Z'Z)^{-1}z_iy_i + \frac{(Z'Z)^{-1}z_iz_i\hat\beta}{1-H_{ii}} - \frac{(Z'Z)^{-1}z_i\,H_{ii}y_i}{1-H_{ii}}\\
&= \hat\beta + \frac{(Z'Z)^{-1}z_iz_i\hat\beta}{1-H_{ii}} - \frac{(Z'Z)^{-1}z_iy_i}{1-H_{ii}}\\
&= \hat\beta - \frac{(Z'Z)^{-1}z_i(y_i-\hat y_i)}{1-H_{ii}}.
\end{aligned}$$

Now the prediction for y_i when (z_i, y_i) is removed from the least squares fit is

$$\hat y_{i,(-i)} \triangleq z_i'\hat\beta_{(-i)} = \hat y_i - \frac{H_{ii}(y_i - \hat y_i)}{1-H_{ii}}.$$

Multiplying both sides by $1 - H_{ii}$ and rearranging we find that

$$\hat y_i = H_{ii}y_i + (1-H_{ii})\hat y_{i,(-i)}. \tag{2.29}$$

Equation (2.29) has an important interpretation. The least squares fit $\hat y_i$ is a weighted combination of y_i itself and the least squares prediction we would have made for it, had it been left out of the fitting. The larger H_{ii} is, the more that $\hat y_i$ depends on y_i. It also means that if we want to compute a "leave one out" residual $y_i - \hat y_{i,(-i)}$ we don't have to actually take (z_i, y_i) out of the data and rerun the regression. We can instead use

$$y_i - \hat y_{i,(-i)} = \frac{y_i - \hat y_i}{1-H_{ii}}. \tag{2.30}$$