Chapter 4

# Over-parametrized models

## 1 Rank, subspaces, and bases

Once more suppose $X = (x_1, \dots, x_p)$ is an $n \times p$ matrix of rank $m$, with $m < p$. That is, the space $\mathcal{X}$ spanned by all the columns of $X$ can also be spanned by some subset of $m$ linearly independent columns. The space $\mathcal{X}$ is ***over-parametrized***; we don't need all $p$ parameters to specify vectors in $\mathcal{X}$, because there is a set of $m$ linearly independent columns that spans $\mathcal{X}$. For each $z$ in $\mathcal{X}$ there are many different $b$ in $\mathbb{R}^p$ for which $z = Xb$. The non-uniqueness of $b$ leads to several difficulties when the columns of $X$ are used as the predictors in a least squares problem.

The $p \times n$ matrix $X^T = (w_1, \dots, w_n)$ also has rank $m$ (Axler, 2015, pages 111–112). The subspace $\mathcal{W}$ of $\mathbb{R}^p$ spanned by all the columns of $X^T$ can also be spanned by some subset of $m$ linearly independent columns, which (without loss of generality) we may suppose correspond to the first $m$ rows of $X$. Put another way,

$$X^T = \begin{array}{cc} & \begin{array}{cc} m & p-m \end{array} \\ p & \begin{bmatrix} W_1 & W_2 \end{bmatrix} \end{array}$$

where the linearly independent columns of $W_1$ form a basis for $\mathcal{W}$ and $W_2 = W_1 A$ for some $m \times (p-m)$ matrix $A$.

A vector $z$ in $\mathbb{R}^n$ belongs to $\mathcal{X}$ if and only if it can be written as $Xb$ for some $b$ in $\mathbb{R}^p$. If we partition $z$ into a vector $z_1$ of length $m$ and a vector $z_2$ of length $n - m$ then

$$Xb = \binom{z_1}{z_2} \text{ if and only if } z_1 = W_1^T b \text{ and } z_2 = W_2^T b = A^T z_1.$$

 

That is, if $z \in \mathcal{X}$ then $z_2 = A^T z_1$ and $Xb = z$ if and only if $W_1^T b = z_1$. Write $b_z$ for the orthogonal projection of $b$ onto $\mathcal{W}$, so that $w = b - b_z \in \mathcal{W}^\perp$. The vector $b_z$ is the unique member of $\mathcal{W}$ for which $W_1^T b_z = z_1$; it is the same for every solution of $W_1^T b = z_1$. The $w$ could be anything in $\mathcal{W}^\perp$. In summary: if $z \in \mathcal{X}$ there is a unique $b_z$ in $\mathcal{W}$ for which $W_1^T b_z = z_1$ and

<4.1> $$\mathcal{B}_z = \{b \in \mathbb{R}^p : Xb = z\} = \{b_z + w : w \in \mathcal{W}^\perp\}.$$

The solution set could also be characterized using the svd. The singular value decomposition of $X$ is given by an $n \times n$ orthogonal matrix $U$ and a $p \times p$ orthogonal matrix $V$, and nonzero singular values $\lambda_1, \dots, \lambda_m$. If we partition $U$ and $V$ as

$$U = \begin{array}{cc} & \begin{array}{cc} m & n-m \end{array} \\ n & \begin{bmatrix} U_1 & U_2 \end{bmatrix} \end{array} \quad \text{AND} \quad V = \begin{array}{cc} & \begin{array}{cc} m & p-m \end{array} \\ p & \begin{bmatrix} V_1 & V_2 \end{bmatrix} \end{array}$$

then $X = U_1 \Lambda_1 V_1^T$, where $\Lambda_1 = \text{diag}(\lambda_1, \dots, \lambda_m)$, a nonsingular $m \times m$ matrix with inverse $\Lambda_1^{-1} = \text{diag}(\lambda_1^{-1}, \dots, \lambda_m^{-1})$. The columns of $U_1$ provide an orthonormal basis (onb) for $\mathcal{X}$; the columns of $U_2$ provide an onb for $\mathcal{X}^\perp$; the columns of $V_1$ provide an onb for $\mathcal{W}$; and the columns of $V_2$ provide an onb for $\mathcal{W}^\perp$.

## 2 A bad thing about non-uniqueness

In statistical applications, we often think of $y$ as a random vector whose expected value $\mu = \mathbb{E}y$ is modelled as an unknown element of the subspace $\mathcal{X}$ of $\mathbb{R}^n$ spanned by the columns of a given matrix $X$. By assumption, the unknown $\mu$ can be written as $X\beta$ for some unknown $\beta$ in $\mathbb{R}^p$. The fitted vector $\widehat{y}$ is then thought of as an estimator for the unknown $\mu$ and the $\widehat{b}$ for which $\widehat{y} = X\widehat{b}$ is thought of as an estimator for $\beta$. Non-uniqueness of $\widehat{b}$ (or of $\beta$ itself) clearly causes some embarrassment. How can we interpret quantities that are not uniquely determined?

Statisticians use two general strategies for avoiding this embarrassment.

(i) Restrict attention to linear functions $L^T\beta$ (with $L \in \mathbb{R}^p$) of the unknown $\beta$ that are uniquely determined by $\mu$. That is, only interpret the linear combinations for which there exists some vector $\ell$ in $\mathbb{R}^n$ such that $L^T\beta = \ell^T\mu$ whenever $\mu = X\beta$. Such linear combinations are said to be ***estimable***.

(ii) Impose a set of $p - m$ linearly independent linear constraints on $b$, say $D_1^T b = 0$ for fixed $p \times (p-m)$ matrix $D_1$, so that to every $z \in \mathcal{X}$ there exists a unique $b$ in $\mathbb{R}^p$ for which both $z = Xb$ and $D_1^T b = 0$.

# 3 Estimable functions

If $L \in \mathcal{W}$ then it can be written as a linear combination of the columns of $X^T$, that is, $L = X^T \ell$ for some $\ell$ in $\mathbb{R}^n$. If $Xb = z$ then

$$L^T b = \ell^T X b = \ell^T z.$$

In particular, if $X\beta = \mu \in \mathcal{X}$ then $L^T\beta = \ell^T \mu$. That is, $L^T\beta$ is an estimable function.

Conversely, suppose that $L = L_1 + L_2$ with $L_1 \in \mathcal{W}$ and $L_2 \in \mathcal{W}^\perp$ and $b = b_Z + w \in \mathcal{B}_z$, as in <4.1>. Then

$$L^T b = L_1^T b_z + L_2^T w.$$

If $L_2 \neq 0$ then we can generate many different $L^T b$ values by varying $w$. For example, try $w = 0$ then $w = L_2$.

In short, the estimable functions $L^T\beta$ of the unknown parameters are precisely those for which $L \in \mathcal{W}$, that is, $L = X^T\ell$ for some $\ell$ is $\mathbb{R}^n$. In that case, $L^T\widehat{b} = \ell^T X \widehat{b} = \ell^T \widehat{y}$ for every solution $\widehat{b}$ of the equation $X\widehat{b} = \widehat{y}$. Under the linear model where $y = \mu + \xi$ with $\mathbb{E}y = \mu \in \mathcal{X}$ and $\text{var}(\xi) = \sigma^2 I_n$ we have

$$\mathbb{E}L^T\widehat{b} = L^T\beta \quad \text{AND} \quad \text{var}\left(L^T\widehat{b}\right) = \sigma^2 \|\ell\|^2 .$$

**Remark.** There is a role for estimability if we start with an $X$ of full rank but lose some of the data. When the corresponding rows are removed from $X$ we might be left with a reduced model matrix that is not of full rank.

# 4 Linear constraints

Suppose we constrain the parametrization by adding $p - m$ more rows to the $X$ matrix, in such a way that the augmented matrix

$$\widetilde{X} = \begin{array}{c} \\ n \\ p-m \end{array} \begin{array}{c} p \\ \begin{bmatrix} X \\ D_1^T \end{bmatrix} \end{array} = \begin{array}{c} \\ m \\ n-m \\ p-m \end{array} \begin{array}{c} p \\ \begin{bmatrix} W_1^T \\ W_2^T \\ D_1^T \end{bmatrix} \end{array}$$

has rank $p$. The columns of $\widetilde{X}^T$ span $\mathbb{R}^p$. The columns of the $p \times p$ matrix $M^T = [W_1, D_1]$ span the same space. Thus $M$ has rank $p$; it is nonsingular. For each $z$ in $\mathcal{X}$ there is now a unique $b$ in $\mathbb{R}^p$ for which $Xb = z$

and $D_1^T b = 0$:

$$\widetilde{X}b = \begin{array}{r} \\ n \\ p-m \end{array} \begin{array}{c} 1 \\ \left[\begin{array}{c} z \\ 0 \end{array}\right] \end{array} \text{ iff } \begin{pmatrix} W_1^T \\ D_1^T \end{pmatrix} b = \begin{pmatrix} z_1 \\ 0 \end{pmatrix} \text{ iff } b = M^{-1} \begin{pmatrix} z_1 \\ 0 \end{pmatrix}.$$

Here is another way to derive the solution. It corresponds to the way **R** actually handles the over-parametrization problem for factors. The linearly independent columns of $D_1$ span a $(p-m)$-dimensional subspace $\mathcal{D}$ of $\mathbb{R}^p$. Find a $p \times m$ matrix $D_2$ whose columns span $\mathcal{D}^\perp$, the $m$-dimensional subspace of all vectors in $\mathbb{R}^p$ that are orthogonal to $\mathcal{D}$. The requirement $D_1^T b = 0$ means that $b$ should be orthogonal to $\mathcal{D}$. That is, $d = D_2 a$ for some $a$ in $\mathbb{R}^m$. The second requirement then becomes $XD_2 a = z$. We now have a new parametrization for the model with the $n \times m$ model matrix $XD_2$ and parameters $a \in \mathbb{R}^m$. The equation $XD_2 a = z$ has a unique solution for each $z$ in $\mathcal{X}$.

<4.2> **Example.** Suppose observations $y_1, y_2, \ldots, y_9$ are each identified as coming from one of three groups by means of a factor variable $G$ with levels “A”, “B”, and “C”:

$$G = [A, A, A, B, B, B, C, C, C].$$

The **R** command `lm(y ~ G)` would, conceptually, create a $9 \times 4$ model matrix $X$ with columns $(\mathbb{1}, G_A, G_B, G_C)$, where $G_A$ has ones in the first three positions and zeros in the remaining six positions. More succinctly, $G_A$ is the indicator function that takes the value 1 when the item comes from group A and zero otherwise. And so on. The least squares problem seeks $b_0, b_A, b_B, b_C$ to minimize

$$\|y - b_0\mathbb{1} - b_A G_A - b_B G_B - b_C G_C\|^2.$$

The matrix $X$ has rank 3, because $\mathbb{1} = G_A + G_B + G_C$. The minimizing $b_i$'s are not unique.

We could make the solution unique by eliminating the intercept term, that is, by putting $b_0 = 0$. We could also set one of the other $b_i$'s to zero (treatment contrasts). We could also constrain $b_A + b_B + b_C = 0$ (sum contrasts or Helmert contrasts).

The last three of these alternatives correspond to working with a model matrix of the form $(\mathbb{1}, \mathbb{G}D_2)$, where $\mathbb{G} = (G_A, G_B, G_C)$ and $D_2$ is one of the following three types of matrix:

```
> contr.sum(3)
  [,1] [,2]
1    1    0
2    0    1
3   -1   -1
> contr.treatment(3)
  2 3
1 0 0
2 1 0
3 0 1
> contr.helmert(3)
  [,1] [,2]
1   -1   -1
2    1   -1
3    0    2
```

For details consult Chambers and Hastie (1992, Chapter 2) and Venables and Ripley (2002, Section 6.2). I'll also be creating a new handout to show how the interpretation of the summary output is affected by the different choices of constraint.

□

## References

Axler, S. J. (2015). *Linear Algebra Done Right* (Third ed.). Undergraduate Texts in Mathematics. Springer.

Chambers, J. M. and T. J. Hastie (Eds.) (1992). *Statistical Models in S.* Wadsworth.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4th ed.). Springer-Verlag.