Handout 04: Normal Equations

2 Ordinary least squares

Many of the advantages of linear models concern the beneficial properties of the standard estimators used to compute the unknown parameters β_j from observed data. As a next step we would like to explore the definition of these estimators. To this aim, it will be useful to provide a compact matrix-based description of a linear model. Throughout this text, unless otherwise noted, we use a notation where n is the sample size, p is the number of variables, i is an index over the samples, and j is the index over the variables. With this notation a complete general description of a linear model can be given by

$$y_i = \beta_1 \cdot x_{i,1} + \cdots + \beta_p \cdot x_{i,p} + \epsilon_i, \quad \forall\, i = 1, \ldots, n. \tag{2.1}$$

Or simply

$$y_i = \sum_j \beta_j \cdot x_{i,j} + \epsilon_i, \quad \forall\, i = 1, \ldots, n. \tag{2.2}$$

Notice that we do not need to include an explicit intercept term β_0. If one is required this can be included by setting $x_{i,1}$ equal to one for every single observation i. Using matrix notation, we can write the linear model equation simultaneously for all observations as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ x_{1,2} & \ddots & & x_{p,2} \\ \vdots & & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{2.3}$$

which can be compactly written in terms of a vector y of the responses, a matrix X of the predictor variables, a vector β of the unknown parameters, and a vector ϵ of the errors

$$y = X\beta + \epsilon. \tag{2.4}$$

Beyond compactness, this notation is also useful as many of the computational properties of linear models can be reduced to linear algebraic properties of the matrix X.

It is desirable for an estimate $\widehat{\beta}$ of the unknown vector β to be able to explain as much variation as possible of the responses y. One way of viewing a linear model is as a decomposition of y into a fixed, deterministic signal $X\beta$ and a stochastic random noise term ϵ. Prediction and inference both benefit from making the signal term as

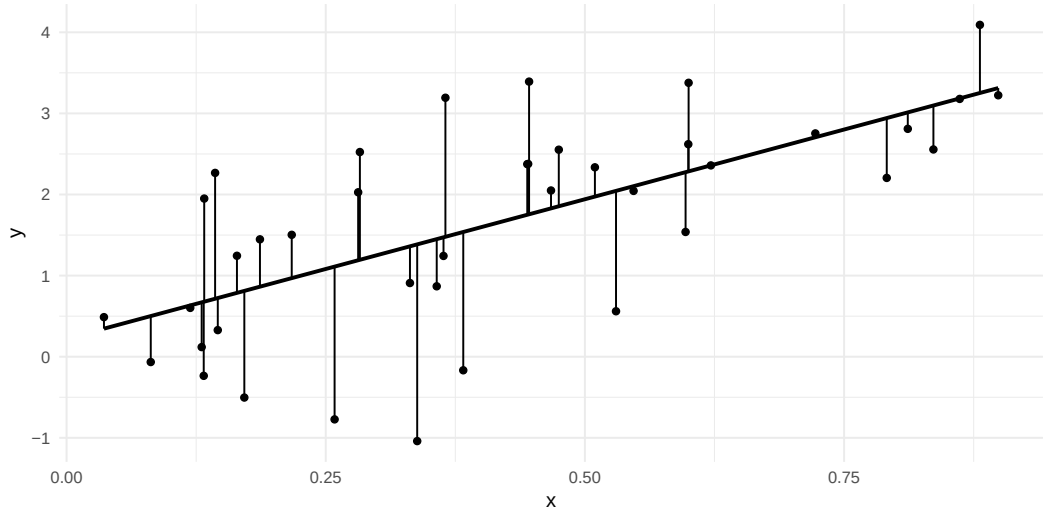

Figure 1: Visualization of residuals from the linear model $y = \beta_0 + \beta_1 x$.

dominant as possible. A good method for measuring this is to construct the vector of residuals

$$r = y - X\widehat{\beta}. \tag{2.5}$$

We can compare estimators by comparing the size of their residual vectors. A graphical representation of residuals from a linear model is given in Equation 1.

There are many choices for measuring the size of a regression vector, several of which lead to important, and distinct, estimators. The hinged loss, which penalizes positive residuals more than negative ones (or vice versa), leads to quantile regression. Metrics that penalize all residuals past some large threshold equally lead to robust regression techniques. Metrics that give each sample a weight w_i depending on the specific values of the data x_i can result in kernel regression (Section **??**), local regression (Section **??**) and are an important intermediate step in solving the generalized linear model problems that arise in Chapters **??**, **??**, and **??**.

In this chapter, we will focus on the most popular choice of metric to measure the size of the regression vector: the sum of squared residuals. Minimizing the sum of squared residuals leads to the *ordinary least squares* (OLS) estimator. Why is this such a popular choice? For one thing, it allows us to write the metric in terms of an inner product or vector norm

$$\sum_i r_i^2 = r^t r = ||r||_2^2, \tag{2.6}$$

a form that is easy to work with both computationally and theoretically. The choice of the sum of squared residuals is also motivated by the maximum likelihood estimator when the ϵ_i's are independent and identically distributed random variables with a normal distribution having zero mean.

3 The normal equations

We now have a formal specification of the ordinary least squares estimator. Computing the estimator given a set of observed data requires solving an optimization problem. This particular optimization problem is unconstrained and has a continuous gradient, so an obvious first step would be to find the gradient of the least squares objective function with respect to the vector b

$$\nabla_b \left[||y - Xb||_2^2\right] = \nabla_b \left[y^t y + b^t X^t X b - 2y^t X b\right] \tag{3.1}$$
$$= 2X^t X b - 2X^t y. \tag{3.2}$$

A necessary condition for minimizing the objective function is to have the gradient equal to the zero vector, $\vec{0}$. If the Hessian matrix is positive definite at this solution, only then are we guaranteed to have a local minimum. The Hessian matrix here is constant everywhere, in that it does not depend on the value of b. Specifically, it is given by

$$H\left(||y - Xb||_2^2\right) = X^t X. \tag{3.3}$$

For a matrix M to be positive definite, we need to have $z^t M z$ be strictly positive for any vector z not equal to the zero vector. Notice that in our case this matrix product can be written as a vector norm

$$z^t H\left(||y - Xb||_2^2\right) z = z^t X^t X z \tag{3.4}$$
$$= ||Xz||_2^2. \tag{3.5}$$

The squared ℓ_2-norm is never negative and is only zero at the zero vector.

We see then that the Hessian is positive definite everywhere if and only if a non-zero vector z does not exist such that Xz is the zero vector. This, in turn is true if and only if X is not full rank. In this case, there are many possible values of b that all attain the minimum least squares solution. Such a result should not surprise us. If we have such a z, then there are many parameter vectors b that result in the exact same estimates for y as there are for the true β

$$X(\beta + a \cdot z) = X\beta + aXz \tag{3.6}$$
$$= X\beta + \vec{0} \tag{3.7}$$
$$= X\beta. \tag{3.8}$$

In such cases it is possible to place constraints on the problem to formulate a related problem with a unique solution. For instance, Section **??** illustrates how to find the unique OLS solution of minimal norm. Although minimum-norm least squares solutions are widely used in many science and engineering applications, it is more common in statistics to constrain solutions to rank-deficient problems in other ways. In particular, R's `lm` and `glm` solvers reformulate rank-deficient problems into full-rank ones by selecting a subset of columns using a heuristic procedure based on the

model matrix column order. Other common subset selection approaches include the lasso (see Chapter **??**), which solves a penalized version of OLS.

Satisfied that we attain a local minimum wherever the gradient is zero, we return to Equation 3.2. Setting this equal to zero we get what are known as the *normal equations*, a linear system of equations of p variables over p unknowns expressed in matrix form as

$$X^t X b = X^t y. \tag{3.9}$$

Solving systems like the normal equations for b in a numerically stable and efficient manner is an important problem encountered repeatedly in this text. Linear systems of equations are generically solved by Gaussian elimination, but we will see that other more efficient and/or numerically stable methods based on the Cholesky, QR, or SVD decompositions can be used depending on context.

References

LAB QUESTIONS

1. Here is a thing!