

**Exercise A.** Let  $y = X\beta + e$ , where  $y = (y_1, \dots, y_N)$ ,  $X$  is an  $N \times P$  matrix with  $i$ th row  $x_i$  and  $e$  a vector of model residuals. Further let  $W$  be the  $N \times N$  diagonal vector of weights  $w_i$ .

We may then rewrite the equation  $\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^P} \sum_{i=1}^N \frac{w_i}{2} (y_i - x_i^T \beta)^2$  in vector form as

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^P} (y - X\beta)^T W (y - X\beta)$$

. We note that this is equivalent to

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^P} \|y - X\beta\|_W$$

. Let  $r = y - X\beta$ . We want to show that a vector  $\beta \in \mathbb{R}^P$  minimizes the weighted norm  $\|r\|_W = \|y - X\beta\|_W$  iff  $X^T W y = X^T W X \beta$

We note, however, that

$$\begin{aligned} X^T W y &= X^T W X \beta \Rightarrow \\ X^T W y - X^T W X \beta &= 0 \\ X^T W (y - X\beta) &= 0 \\ X^T W r &= 0 \text{ (note that, because } W \text{ is diagonal, } W = W^T) \end{aligned}$$

Thus, it is equivalent to prove that  $\beta \in \mathbb{R}^P$  minimizes the weighted norm  $\|r\|_W = \|y - X\beta\|_W$  iff  $r$  is orthogonal to  $\text{range}(X)$  under the inner product induced by the weighted norm (the weighted inner product). This, however, follows immediately from the properties of orthogonality.

This orthogonal approach gives good geometric intuition into the problem.

Another method that can be used here is a differentiation argument, as follows. We wish to minimize

$$(y - X\beta)^T W (y - X\beta)$$

. We wish to find  $\hat{\beta}$  s.t.  $(y - X\beta)^T W (y - X\beta)$  is minimized. Thus, in order for this quantity

to be minimized, it must satisfy:

$$\begin{aligned}
\frac{d}{d\beta} \left( (y - X\hat{\beta})^T W (y - X\hat{\beta}) \right) &= 0 \\
\frac{d}{d\beta} \left( (y^T - (X\hat{\beta})^T) W (y - X\hat{\beta}) \right) &= 0 \\
\frac{d}{d\beta} \left( (y^T W - \hat{\beta}^T X^T W) (y - X\hat{\beta}) \right) &= 0 \\
\frac{d}{d\beta} \left( y^T W y - y^T W X \hat{\beta} - \hat{\beta}^T X^T W y + \hat{\beta}^T X^T W X \hat{\beta} \right) &= 0 \\
\frac{d}{d\beta} \left( y^T W y - (\hat{\beta}^T X^T W y)^T - \hat{\beta}^T X^T W y + \hat{\beta}^T X^T W X \hat{\beta} \right) &= 0 \\
\frac{d}{d\beta} \left( y^T W y - 2\hat{\beta}^T X^T W y + \hat{\beta}^T X^T W X \hat{\beta} \right) &= 0 \text{ (note that } \hat{\beta}^T X^T W y \text{ is a scalar)} \\
0 - 2X^T W y + 2X^T W X \hat{\beta} &= 0 \\
-X^T W y + X^T W X \hat{\beta} &= 0
\end{aligned}$$

Thus  $\hat{\beta}$  must satisfy  $X^T W X \hat{\beta} = X^T W y$ .  $\square$

**Exercise B.** Numerically speaking, the inversion method is certainly not a very stable way to actually solve the linear system. Furthermore, when the problem is formulated in terms of computing an explicit inverse (rather than a matrix solve method), it is certainly not a very fast way to solve the linear system.

In numerical linear algebra, there are three fairly standard factorizations used to formulate the least squares problem without any explicit inversions: Cholesky factorization, QR factorization, and singular value decomposition. Each is valuable in different scenarios. Solving least squares via cholesky factorization utilizes the normal equations and is one of the fastest methods to solve for  $\beta$ . However, this method is unstable. On the other hand, solving the least squares problem via the SVD is stable, but is one of the slowest methods. It is the best algorithm for cases when the matrix  $X$  is close to being rank-deficient (the smallest singular value is very small). Generally, solving WLS via a QR factorization is the most appropriate method. This method is faster than the SVD and far more stable than the Cholesky method. We will implement WLS via QR factorization as “our method”.

Essentially, the idea of the QR factorization method is to reduce the problem to a triangular solve. Thus we take  $QR = W^{\frac{1}{2}}X$ , where  $QR$  is the reduced QR factorization of  $W^{\frac{1}{2}}X$  (where  $R$  is a square, triangular matrix and  $Q$  has orthonormal columns) and substitute this to obtain:

$$\begin{aligned}
X^T W X \beta &= X^T W y \\
X^T W^{\frac{1}{2}} W^{\frac{1}{2}} X \beta &= X^T W^{\frac{1}{2}} W^{\frac{1}{2}} y \\
R^T Q^T Q R \beta &= R^T Q^T W^{\frac{1}{2}} y \\
R^T R \beta &= R^T Q^T W^{\frac{1}{2}} y \\
R \beta &= Q^T W^{\frac{1}{2}} y
\end{aligned}$$

Thus, the basic algorithm gives us:

1. Take  $W^5X$  (note that this is a simple broadcast)
2. Compute the reduced  $QR$  factorization of  $W^5X$ .
3. Compute  $Q^TW^5y$
4. Solve the upper triangular system  $R\beta = Q^TW^5y$ .

**Exercise C.** Please see the attached code

**Exercise D.** Please see the attached code