

# Lecture 7: High-Dimensional Linear Regression

Francis J. DiTraglia

April 6, 2014

## 1 Review of Matrix Decompositions

### 1.1 The QR Decomposition

Any  $n \times k$  matrix  $A$  with full column rank can be decomposed as  $A = QR$ , where  $R$  is an  $k \times k$  upper triangular matrix and  $Q$  is an  $n \times k$  matrix with orthonormal columns. The columns of  $A$  are *orthogonalized* in  $Q$  via the Gram-Schmidt process. Since  $Q$  has orthogonal columns, we have  $Q'Q = I_k$ . It is *not* in general true that  $QQ' = I$ , however. In the special case where  $A$  is square,  $Q^{-1} = Q'$ .

**Note:** The way we have defined things here is here is sometimes called the “thin” or “economical” form of the QR decomposition, e.g. `qr_econ` in Armadillo. In our “thin” version,  $Q$  is an  $n \times k$  matrix with orthogonal columns. In the “thick” version,  $Q$  is an  $n \times n$  *orthogonal* matrix. Let  $A = QR$  be the “thick” version and  $A = Q_1R_1$  be the “thin” version. The connection between the two is as follows:

$$A = QR = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1R_1$$

**Least-Squares via the QR Decomposition** We can calculate the least squares estimator of  $\beta$  as follows

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y = [(QR)'(QR)]^{-1}(QR)'y \\ &= [R'Q'QR]^{-1}R'Q'y = (R'R)^{-1}R'Qy \\ &= R^{-1}(R')^{-1}R'Q'y = R^{-1}Q'y\end{aligned}$$

In other words,  $\hat{\beta}$  is the solution to  $R\beta = Q'y$ . While it may not be immediately apparent, this is a much easier system to solve than the normal equations  $(X'X)\beta = X'y$ . Because  $R$  is *upper triangular* we can solve  $R\beta = Q'y$  extremely quickly. The product  $Q'y$  is a vector, call it  $v$ , so the system is simply

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1,n-1} & r_{1k} \\ 0 & r_{22} & r_{23} & \cdots & r_{2,n-1} & r_{2k} \\ 0 & 0 & r_{33} & \cdots & r_{3,n-1} & r_{3k} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & r_{k-1,k-1} & r_{k-1,k} \\ 0 & 0 & \cdots & 0 & 0 & r_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{k-1} \\ \beta_k \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{k-1} \\ v_k \end{bmatrix}$$

Hence,  $\beta_k = v_k/r_k$  which we can substitute into  $\beta_{k-1}r_{k-1,k-1} + \beta_k r_{k-1,k} = v_{k-1}$  to solve for  $\beta_{k-1}$ , and so on. This is called **back substitution**. We can use the same idea when a matrix is *lower triangular* only in reverse: this is called **forward substitution**.

To calculate the variance matrix  $\sigma^2(X'X)^{-1}$  for the least-squares estimator, simply note from the derivation above that  $(X'X)^{-1} = R^{-1}(R^{-1})'$ . Inverting  $R$ , however, is easy: we simply apply back-substitution *repeatedly*. Let  $A$  be the inverse of  $R$ ,  $\mathbf{a}_j$  be the  $j$ th column of  $A$ , and  $\mathbf{e}_j$  be the  $j$ th element of the  $k \times k$  identity matrix, i.e. the  $j$ th standard basis vector. Inverting  $R$  is equivalent to solving  $R\mathbf{a}_1 = \mathbf{e}_1$ , followed by  $R\mathbf{a}_2 = \mathbf{e}_2$ , and so on all the way up to  $R\mathbf{a}_k = \mathbf{e}_k$ . In Armadillo, if you enclose a matrix in `trimatu()` or `trimatl()`, and then request the inverse, the library will carry out backward

or forward substitution, respectively.

**Othogonal Projection Matrices and the QR Decomposition** Consider a projection matrix  $P_X = X(X'X)^{-1}X'$ . Provided that  $X$  has full column rank, we have begin

$$P_X = QR(R'R)^{-1}R'Q' = QRR^{-1}(R')^{-1}R'Q' = QQ'$$

Recall that, in general, it is *not* true that  $QQ' = I$  even though  $Q'Q = I$ . It's important to keep this in mind when using the QR decomposition for more complicated matrix calculations, such as linear GMM.

## 1.2 The Singular Value Decomposition

The Singular Value Decomposition (SVD) is probably the most elegant result in linear algebra. It's also an invaluable computational and theoretical tool in statistics and econometrics. I can only give a brief overview here, but I'd encourage you to learn more when you have time. Some excellent references are Strang (1993) and Kalman (2002).

## 2 Gauss-Markov, meet James-Stein

Consider the linear regression model  $\mathbf{y} = X\beta + \epsilon$  In Econ 705 you learned that ordinary least squares (OLS) is the minimum variance unbiased linear estimator of  $\beta$  under the assumptions  $E[\epsilon|X] = \mathbf{0}$  and  $Var(\epsilon|X) = \sigma^2 I$ . When the second assumption fails, you learned that generalized least squares (GLS) provides a lower variance estimator than OLS. All of this is fine, as far as it goes, but there's an obvious objection: why are we restricting ourselves to unbiased estimators? Generically, we know that there is a bias-variance tradeoff. So what happens if we allow ourselves to consider biased estimators? Does some form of the Gauss-Markov Theorem still hold?

## A Fundamental Decomposition

### 3 Ridge Regression

#### 3.1 Ridge as Bayesian Linear Regression

#### 3.2 Another Way to Express Ridge Regression

Data-dependent mapping.

#### 3.3 Ridge Regression via OLS

From the first half of the semester, you may recall that Bayesian linear regression can be thought of as “plain-vanilla” OLS using a design matrix that has been *augmented* with “fake” observations that represent the prior. This turns out to be a very helpful way of looking at Ridge Regression. Define

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_p \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix}$$

The objective function for Ridge Regression is *identical* to the OLS objective function for the augmented dataset, namely

$$\arg \min_{\beta} (\tilde{\mathbf{y}} - \tilde{X}\beta)' (\tilde{\mathbf{y}} - \tilde{X}\beta)$$

Which we can show as follows:

$$\begin{aligned} (\tilde{\mathbf{y}} - \tilde{X}\beta)' (\tilde{\mathbf{y}} - \tilde{X}\beta) &= \begin{bmatrix} (\mathbf{y} - X\beta)' & (-\sqrt{\lambda}\beta)' \end{bmatrix} \begin{bmatrix} (\mathbf{y} - X\beta) \\ -\sqrt{\lambda}\beta \end{bmatrix} \\ &= (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda\beta'\beta \\ &= \text{RSS}(\beta) + \lambda \|\beta\|_2^2 \end{aligned}$$

**Ridge is Always Unique** We know that the OLS estimator is only unique provided that the design matrix has full column rank. In contrast there is *always* a unique solution to the Ridge Regression problem, even when there are more regressors than observations. This follows *immediately* from the preceding: the columns of  $\sqrt{\lambda}I_p$  are linearly independent, so the columns of the augmented data matrix  $\tilde{X}$  are *also* linearly independent, *regardless* of whether the same holds for the columns of  $X$ .

## Calculations for Ridge Regression

### Calculations When $p \gg n$