Elements of Statistical Learning Notes and Exercise Solutions

Ryan Davis

May 12, 2019

Chapter 2

3 Linear Models for Regression

3.2 Linear Regression Models and Least Squares

Derive variance-covariance matrix

Derive the variance-covariance matrix of the least squares parameter estimates (equation 3.8 in book).

The least squares estimator is given by

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y \tag{1}$$

Express the variance of a random variable as below:

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] + \mathbb{E}\left[X\right]^2 \tag{2}$$

$$\begin{aligned}
&\mathrm{Var}\left[\hat{\beta}\right] \\
&\quad = \mathbb{E}\left[\hat{\beta}^2\right] + \mathbb{E}\left[\hat{\beta}\right]^2 \\
&\quad = \mathbb{E}\left[\left(\left(X^T X\right)^{-1} X^T y\right)^2\right] + \mathbb{E}\left[\left(X^T X\right)^{-1} X^T y\right]^2 \\
&\quad = \mathbb{E}\left[\left(\left(X^T X\right)^{-1} X^T \left(X\beta + \epsilon\right)\right)^2\right] + \mathbb{E}\left[\left(X^T X\right)^{-1} X^T \left(X\beta + \epsilon\right)\right]^2 \\
&\quad = \mathbb{E}\left[\left(\left(X^T X\right)^{-1} X^T \left(X\beta + \epsilon\right)\right)^2\right] + \mathbb{E}\left[\beta\right]^2 + \mathbb{E}\left[\left(X^T X\right)^{-1} X^T \epsilon\right]^2
\end{aligned}$$

The last term goes to zero since X is uncorrelated with the error term.

$$\mathrm{Var}\left[\hat{\beta}\right] = \mathbb{E}\left[\left(\left(X^T X\right)^{-1} X^T \left(X\beta + \epsilon\right)\right)^2\right] - \mathbb{E}\left[\beta\right]^2$$

The expectation of the true β is simply β itself.

$$\mathrm{Var}\left[\hat{\beta}\right] = \mathbb{E}\left[\left(\left(X^T X\right)^{-1} X^T \left(X\beta + \epsilon\right)\right)^2\right] - \beta^2$$

Break the term within the expectation into two separate terms.

$$\begin{aligned}
&\mathrm{Var}\left[\hat{\beta}\right] \\
&= \mathbb{E}\left[\left(\left(X^TX\right)^{-1}X^TX\beta + \left(X^TX\right)^{-1}X^T\epsilon\right)^2\right] - \beta^2 \\
&= \mathbb{E}\left[\left(\beta + \left(X^TX\right)^{-1}X^T\epsilon\right)^2\right] - \beta^2 \\
&= \mathbb{E}\left[\beta^2 + 2\left(X^TX\right)^{-1}X^T\epsilon + \left(\left(X^TX\right)^{-1}X^T\right)^2\epsilon^2\right] - \beta^2 \\
&= \beta^2 + \mathbb{E}\left[2\left(X^TX\right)^{-1}X^T\epsilon + \left(\left(X^TX\right)^{-1}X^T\right)^2\epsilon^2\right] - \beta^2 \\
&= \mathbb{E}\left[\left(\left(X^TX\right)^{-1}X^T\right)^2\epsilon^2\right] \\
&= \mathbb{E}\left[\left(\left(X^TX\right)^{-1}X^T\right)^2\right]\mathbb{E}\left[\epsilon^2\right]
\end{aligned}$$

The error term has variance σ^2.

$$\begin{aligned}
&\mathrm{Var}\left[\hat{\beta}\right] \\
&= \sigma^2\,\mathbb{E}\left[\left(\left(X^TX\right)^{-1}X^T\right)^2\right] \\
&= \sigma^2\left(X^TX\right)^{-1}
\end{aligned}$$

Expected prediction error

Compute the expected prediction error at input x_0 for model $Y_0 = f(x_0) + \epsilon_0$. Use estimate $\hat{f}(x_0) = x_0^T\tilde{\beta}$.

$$\begin{aligned}
&\mathbb{E}\left[\left(Y_0 - \tilde{f}(x_0)\right)^2\right] \\
&= \mathbb{E}\left[\left(f(x_0) + \epsilon_0 - \tilde{f}(x_0)\right)^2\right] \\
&= \mathbb{E}\left[f(x_0)^2 - 2f(x_0)\hat{f}(x_0) - \tilde{f}(x_0)^2\right] + \mathbb{E}\left[\epsilon_0^2\right] \\
&= \mathbb{E}\left[f(x_0)^2 - 2f(x_0)\hat{f}(x_0) - \tilde{f}(x_0)^2\right] + \sigma^2 \\
&= \mathbb{E}\left[\left(f(x_0) - \tilde{f}(x_0)\right)^2\right] + \sigma^2 \\
&= \sigma^2 + MSE\left(\tilde{f}(x_0)\right)
\end{aligned}$$

Ridge regression coefficients

The loss function for ridge regression (page 64, equation 3.43) is

$$RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \tag{3}$$

Differentiate with respect to β, set to zero, and solve.

$$
\begin{aligned}
&\frac{\partial RSS}{\partial \beta} \\
&\qquad = 2\left(y - X\beta\right)\left(-X\right) + 2\beta\lambda = 0 \\
&\qquad = -Xy + 2X^T X\beta + 2\beta = 0
\end{aligned}
$$

$$
\begin{aligned}
X^T y &= X^T X\beta + \beta \\
X^T y &= \beta\left(X^T X + I\right) \\
\hat{\beta} &= \left(X^T X + I\right)^{-1} X^T y
\end{aligned}
$$

Exercise 3.4

Show how the vector of least squares coefficients can be obtained froma single pass of the Gram-Schmidt procedure.

Represent X using the QR decomposition

$$X = QR \tag{4}$$

where Q is an $N \times (p+1)$ orthogonal matrix and R is a $(p+1) \times (p+1)$ upper triangular matrix.

The vector of least squares coefficients can be represented as

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y \tag{5}$$

Substitute $X = QR$.

$$
\begin{aligned}
&\hat{\beta} \\
&\qquad = \left((QR)^T (QR)\right)^{-1} (QR)^T y \\
&\qquad = \left(Q^T Q R^T R\right)^{-1} Q^T R y
\end{aligned}
$$

Use the fact that Q is an orthogonal matrix ($Q^T Q = I$).

$$
\begin{aligned}
\hat{\beta} = &= \left(R^T R I\right)^{-1} Q^T R y \\
&= \left(R^T R\right)^{-1} Q^T R y
\end{aligned}
$$

Since $R^{-1} R = I$, we have

$$\hat{\beta} = R^{-1} Q^T y \tag{6}$$