STA9690 - Midterm

Christopher Lang

November 2, 2017

Question 1

Part A

Prove that $E[y] = X\beta^*$ and $Cov[y] = \sigma^2 I$

We know the following:

- $E[\epsilon] = 0$
- $Cov[\epsilon] = \sigma^2 I$

Using the distributive property of the **expectation operator** we define the following:

$$y = X\beta^* + \epsilon$$

$$E[y] = E[X\beta^*] + E[\epsilon]$$
$$E[y] = E[X\beta^*] + 0$$
$$E[y] = E[X\beta^*]$$

X is a fixed quantity, while β^* is a known quantity **that does not vary** (though we don't know what the value is, we know it is a population parameter). If so, they are both constant values (constant matrix, constant vector, respectively), and by the expectation property of $E[a] = a$ we have the following:

$$E[y] = X\beta^*$$

Similarly, to find $Cov[y]$ we note that the covariance of a single random variable is the same as the variance of that random variable by the property of covariance:

$$Cov[Y] = \sigma_Y^2$$

Since we are only looking at a single random variable y, the **covariance operator** is equivalent to the **variance operator**. Therefore, all of the properties of the variance operator is applicable:

$$Cov[Y] = Cov[X\beta^* + \epsilon]$$

Since both X and β^* are constants, they are dropped by the variance operator:

$$Cov[Y] = Cov[\epsilon]$$

And by the known definition of $Cov[\epsilon] = \sigma^2 I$

$$Cov[Y] = \sigma^2 I$$

Part B

We know the following:

- $v \in \mathbb{R}^p$ is a vector such that $Xv = 0$
- $X \in \mathbb{R}^{n \times p}$ is a $n \times p$ real matrix of predictor variables

Let the minimizer of $\hat{\beta} + c \cdot v$ be a solution to the least squares problem. Then:

$$y - X\beta = y - X(\hat{\beta} + c \cdot v) = y - X\hat{\beta} - Xc \cdot v = y - X\hat{\beta} - c \cdot Xv$$

Since we know that $Xv = 0$ then we know that:

$$y - X\hat{\beta} - c \cdot Xv = y - X\hat{\beta} - c \cdot 0$$

So regardless of the value of c, the term $c \cdot 0 = 0$. So:

$$y - X\hat{\beta} - c \cdot Xv = y - X\hat{\beta}$$

Hence:

$$(\hat{\beta} + c \cdot v) = argmin_{\beta} ||y - X\beta||_2^2$$

In summary, the minimizer $\hat{\beta} + c \cdot v$ is equivalent to the minimizer $\hat{\beta}$ if $Xv = 0$

Part C

If the matrix X has completely linear independent columns, than the matrix is full column rank and has a unique solution for v. That solution is $v = 0$

Part D

If $p > n$ then the matrix $Xv = 0$ is an underdetermined homogeneous linear system, with a trivial solution of $v = 0$ and **infinitely many non-trivial solutions**. Therefore:

- There will be infinitely many vectors $v \neq 0$ for the homogeneous system $Xv = 0$
- For the minimizer $\hat{\beta} + c \cdot v$ in part b, vector v has infinitely many non-trivial solutions, which also means that there will be infinitely many regression coefficient estimates that are minimizers to the least squares problems, regardless of the value of c

If there are infinitely many regression estimates that is not unique, we can expect that some of them to have have their sign flipped since we can change the value of c

For example, solutions $\hat{\beta}_i$, $\hat{\beta}_i - 6v$, and $\hat{\beta}_i + 6v$ are all valid solutions under part b, yet their signs (and values) are changing

This is not useful in regression. The coefficient's value and sign are often used for insight generation:

- We often want to know the direction of contribution each variables has (positive/negative sign)
- We often want to know the magnitude of the contribution each variables has (the total value)

If these are not unique, then the model is not usable for most use cases

Part E

In ridge regression, the objective function of a standard linear regression has an extra term during minimization:

$$argmin||y - X\beta||^2 + \lambda||\beta||^2$$

Given that we know the following:

- The response $\widetilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{R}^{n \times p}$
- The predictor vector $\widetilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \in \mathbb{R} \in \mathbb{R}^{(n+p) \times p}$
- The zero vector $0 \in \mathbb{R}^p$
- The identity matrix $I \in \mathbb{R}^{p \times p}$

Then we get the following least squares objective function:

$$||\widetilde{y} - \widetilde{X}\beta||^2 = ||y_1 - X_1\beta||^2 + ||y_2 - X_2\beta||^2$$
$$||y - X\beta||^2 + ||0 - \sqrt{\lambda} I\beta||^2$$
$$||y - X\beta||^2 + \lambda(I\beta)^2$$
$$||y - X\beta||^2 + \lambda||\beta||^2$$

The result $||y - X\beta||^2 + \lambda||\beta||^2$ is equal to the original minimizer $argmin||y - X\beta||^2 + \lambda||\beta||^2$

Since this objective function minimizes the least squares, the $\widetilde{\beta}^{ridge}$ minimizers will be the vector of linear regression coefficients

Part F

For matrix $\widetilde{X}$ to be full column rank, all of its columns **must** be linearly independent from another, in all possible combinations

In $\widetilde{X}_1$ the value of $\sqrt{\lambda} I$ is guaranteed to be a full column rank because:

- By definition, the identity matrix I will have full column rank
- Multiplying an identity matrix with a scalar does not change this property

Hence the matrix $\widetilde{X}$ is also guaranteed to be of full column rank due to $\sqrt{\lambda} I$, regardless of matrix X

If $\widetilde{X}$ is full column rank it will have a unique solution. Therefore, $\widetilde{\beta}^{ridge}$ will find unique coefficients

Part G

Let:

$$||\widetilde{y} - \widetilde{X}\beta||^2 = ||y_1 - X_1\beta||^2 + \lambda||\beta||^2 = (y_1 - X_1\beta)^T(y_1 - X_1\beta) + \lambda\beta^T\beta$$
$$(y_1^T - (X_1\beta)^T)(y_1 - X_1\beta) + \lambda\beta^T\beta$$
$$y_1^T y_1 - y_1^T X_1\beta - (X_1\beta)^T y + (X_1\beta)^T X_1\beta + \lambda\beta^T\beta$$
$$y_1^T y_1 - y_1^T X_1\beta - \beta^T X_1^T y_1 + \beta^T X_1^T X_1\beta + \lambda\beta^T\beta$$
$$y_1^T y_1 - 2\beta^T X_1^T y_1 + \beta^T X_1^T X_1\beta + \lambda\beta^T\beta$$

Now let $e = y_1^T y_1 - 2\beta^T X_1^T y_1 + \beta^T X_1^T X_1 \beta + \lambda \beta^T \beta$

To find minimum, we find the partial derivative with respective to β and set it equal to zero. Then rearrange so that β is by itself on one side:

$$\frac{\partial e}{\partial \beta} = \frac{\partial e}{\partial \beta}(y_1^T y_1 - 2\beta^T X_1^T y_1 + \beta^T X_1^T X_1 \beta + \lambda \beta^T \beta) = 0$$

$$\frac{\partial e}{\partial \beta} = (0 - 2X_1^T y_1 + 2X_1^T X_1 \beta + 2\lambda\beta) = 0$$

$$-X_1^T y_1 + X_1^T X_1 \beta + \lambda\beta = 0$$

$$X_1^T X_1 \beta + \lambda\beta = X_1^T y_1$$

$$\beta(X_1^T X_1 + \lambda I) = X_1^T y_1$$

$$\beta = (X_1^T X_1 + \lambda I)^{-1} X_1^T y_1$$

We restate that $X_1 = X$, $y_1 = y$, and $\beta = \widetilde{\beta}^{ridge}$ to get the final answer:

$$\widetilde{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

Part H

Let $X = UDV^T$

Substitute it into our original formula:

$$\widetilde{\beta}^{ridge} = ((UDV^T)^T(UDV^T) + \lambda I)^{-1}(UDV^T)^T y$$

Due to the tranpose property $(AB)^T = B^T A^T$:

$$\widetilde{\beta}^{ridge} = (VD^T U^T UDV^T + \lambda I)^{-1}(VD^T U^T)y$$

Since we know that matrix U is an orthonormal matrix, we know that $U^T U = I$, where I is the identity matrix:

$$\widetilde{\beta}^{ridge} = (VD^T IDV^T + \lambda I)^{-1}(VD^T U^T)y$$

Which we can drop in this context:

$$\widetilde{\beta}^{ridge} = (VD^T DV^T + \lambda I)^{-1}(VD^T U^T)y$$

Since we know matrix D is a diagonal matrix, $D = D^T$, hence:

$$\widetilde{\beta}^{ridge} = (VD^2 V^T + \lambda I)^{-1}(VD^T U^T)y$$

Since V is an orthonormal matrix, $V = V^T = V^{-1}$, hence:

$$\widetilde{\beta}^{ridge} = ((VD^2)V^{-1} + \lambda I)^{-1}(VD^T U^T)y$$

$$\widetilde{\beta}^{ridge} = (V^{-1}(VD^2) + \lambda I)^{-1}(VD^T U^T)y$$

$$\widetilde{\beta}^{ridge} = (V^T VD^2 + \lambda I)^{-1}(VD^T U^T)y$$

Again, since V is orthonormal, $V^T V = I$ and hence:

$$\widetilde{\beta}^{ridge} = (D^2 I + \lambda I)^{-1}(VD^T U^T)y$$

$$\widetilde{\beta}^{ridge} = (D^2 + \lambda I)^{-1}(VD^T U^T)y$$

Part I

We can determine whether or not the $\widetilde{x}^T\widetilde{\beta}^{ridge}$ is a biased estimator by determining whether or not its expected value is equal to $\widetilde{x}^T\beta^*$. In other words:

If $\widetilde{x}^T\widetilde{\beta}^{ridge} \neq \widetilde{x}^T\beta^*$, then the estimator $\widetilde{x}^T\widetilde{\beta}^{ridge}$ is biased, not biased otherwise

$$E[\widetilde{x}^T\widetilde{\beta}^{ridge}] = E[\widetilde{x}^T(D^2+\lambda I)^{-1}(VD^TU^T)y]$$

Since X is fixed, then the matrices D, V, and U are also fixed. Similarly, matrix I and scalar λ are also fixed (conditional on specified λ). Therefore, following the properties of the **expectation operator**:

$$E[\widetilde{x}^T\widetilde{\beta}^{ridge}] = \widetilde{x}^T(D^2+\lambda I)^{-1}(VD^TU^T)E[y]$$

Where $E[y] = X\beta^*$ as defined in regression, as well as $X = UDV^T$:

$$E[\widetilde{x}^T\widetilde{\beta}^{ridge}] = \widetilde{x}^T(D^2+\lambda I)^{-1}(VD^TU^T)X\beta^*$$

$$E[\widetilde{x}^T\widetilde{\beta}^{ridge}] = \widetilde{x}^T(D^2+\lambda I)^{-1}(VD^TU^T)UDV^T\beta^*$$

$$E[\widetilde{x}^T\widetilde{\beta}^{ridge}] = \widetilde{x}^T(D^2+\lambda I)^{-1}(VD^2)V^T\beta^*$$

$$E[\widetilde{x}^T\widetilde{\beta}^{ridge}] = \widetilde{x}^T(D^2+\lambda I)^{-1}D^2\beta^*$$

Compare this to $E[\widetilde{x}^T\beta^*] = \widetilde{x}^T\beta^*$, the $\widetilde{x}^T\widetilde{\beta}^{ridge}$ is a **biased estimator**