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 transpose 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 $X = UDV^T$:

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

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

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

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

We find that $E[\widetilde{x}^T\widetilde{\beta}^{ridge}] \neq E[\widetilde{x}^T\beta^*]$ as:

$$\widetilde{x}^T(D^2+\lambda I)^{-1}D^2\beta^* \neq \widetilde{x}^T\beta^*$$

Therefore, we conclude that $\widetilde{x}^T\widetilde{\beta}^{ridge}$ is a **biased estimator**

Question 2

Part A

We can reformulate the least squares problem into an algebraic form:

$$argmin_\beta ||y - X\beta||_2^2 + \lambda||\beta||_2^2 = argmin_\beta \sum_{i=1}^{n}(y_i - \beta_i)^2 + \lambda\sum_{i=1}^{n}\beta_i^2$$

To find the solution that minimizes the algebraic form, we take the partial derivative of the objective function and set it zero:

$$\frac{\partial}{\partial\beta}\sum_{i=1}^{n}(y_i - \beta_i)^2 + \lambda\sum_{i=1}^{n}\beta_i^2 = 0$$

$$\sum_{i=1}^{n}2(y_i - \beta_i)(0-1) + \lambda\sum_{i=1}^{n}2\beta_i = 0$$

$$\sum_{i=1}^{n}2(\beta_i - y_i) + \lambda\sum_{i=1}^{n}2\beta_i = 0$$

$$\sum_{i=1}^{n}\beta_i - \sum_{i=1}^{n}y_i + \lambda\sum_{i=1}^{n}\beta_i = 0$$

$$\sum_{i=1}^{n}\beta_i + \lambda\sum_{i=1}^{n}\beta_i = \sum_{i=1}^{n}y_i$$

$$\sum_{i=1}^{n}\beta_i(1+\lambda) = \sum_{i=1}^{n}y_i$$

$$\sum_{i=1}^{n}\beta_i = \sum_{i=1}^{n}\frac{y_i}{(1+\lambda)}$$

Now let $\beta = \widetilde{\beta}^{ridge}$, so that:

$$\sum_{i=1}^{n}\widetilde{\beta}_i^{ridge} = \sum_{i=1}^{n}\frac{y_i}{(1+\lambda)}$$

Therefore, for any given $i = 1,\ 2,\ ...,\ n$:

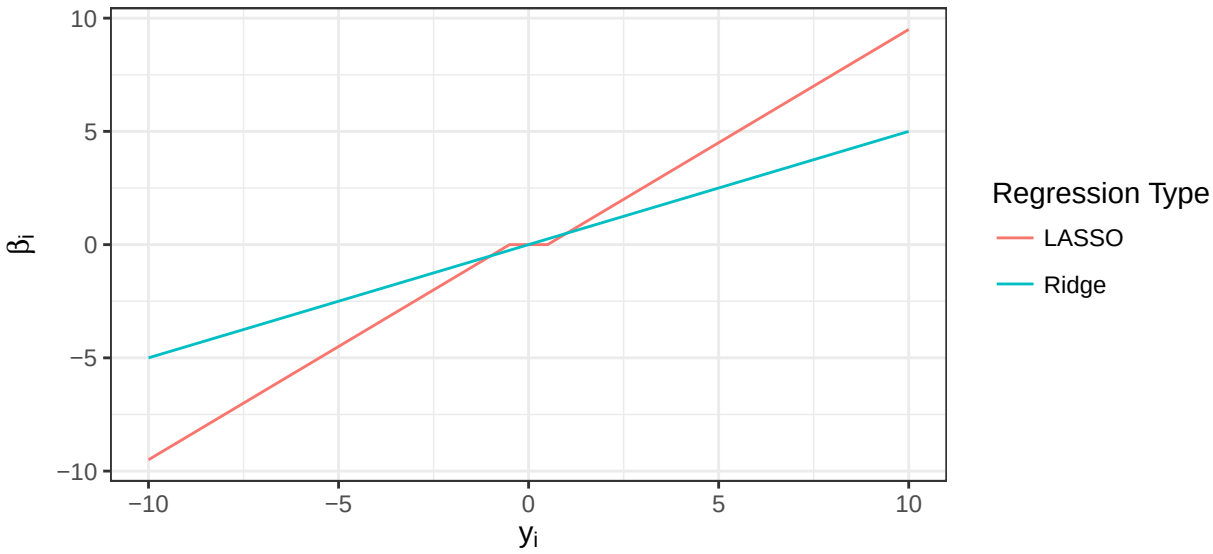
$$\widetilde{\beta}_i^{ridge} = \frac{y_i}{(1+\lambda)}$$

Part B

Let:

- $\widetilde{\beta}_i^{ridge} = \frac{y_i}{(1+\lambda)}$
- $\widetilde{\beta}_i^{lasso} = sign(y_i) \cdot max(|y_i| - \frac{\lambda}{2}, 0)$

```
library(tidyverse)

beta_ridge <- function(y, lambda=1) {
  y / (1 + lambda)
}

beta_lasso <- function(y, lambda=1) {
  sign(y) * pmax(abs(y) - (lambda / 2), 0)
}

y_values <- seq(-10, 10, by = 0.01)

df <- (
    data_frame(y = y_values, b_ridge = beta_ridge(y_values),
               b_lasso = beta_lasso(y_values)) %>%
      gather(regression_type, beta, b_ridge, b_lasso)
)

ggplot(df) + aes(y, beta, color = regression_type) + geom_line() + theme_bw() +
  labs(x = expression(""*y[i]*""), y = expression(""*beta[i]*"")) +
  scale_color_discrete(guide = guide_legend("Regression Type"),
                       label = c('b_lasso' = 'LASSO', 'b_ridge' = 'Ridge'))
```



Observation into the difference between LASSO and Ridge is:

- The LASSO coefficients shows a steeper slope in regards to the coefficients vs. Ridge. Hence LASSO allows for a greater coefficient size compared to Ridge. This is due to Ridge's greater penalty imposed on the objective function, limiting the size of the coefficients
- LASSO also shows a range of y_i values where the coefficient becomes zero (between $-\frac{1}{2}$ and $\frac{1}{2}$). This is LASSO's stronger shrinkage capabilities, forcing more coefficients to be zero vs. Ridge

Part C

We first start with the general least square minimization problem:

$$\widetilde{\beta} = argmin_{\beta}||y - X\beta||_2^2$$

We know that the predictor matrix X is orthogonal, hence it will have the following properties:

- $X^TX = XX^T = I$
- $X^T = X^{-1}$
- That $X^T(X^T)^T = X^TX = XX^T = (X^T)^TX^T$, hence X^T is also orthogonal
- $||Xv||_2 = ||v||_2$, therefore $||Xv||_2^2 = ||v||_2^2$, where $v \in \mathbb{R}^n$
- $||y||_2^2 = ||Xy||_2^2 = ||X^Ty||_2^2$, where $y \in \mathbb{R}^n$

Assuming predictor matrix X is orthogonal, we find that:

$$||y - X\beta||_2^2 = ||y||_2^2 - 2y^TX\beta + ||X\beta||_2^2 = ||X^Ty||_2^2 - 2(XX^T)y^TX\beta + ||\beta||_2^2 = ||X^Ty||_2^2 - 2Xy^T\beta + ||\beta||_2^2$$

Therefore:

$$||y - X\beta||_2^2 = ||X^Ty - \beta||_2^2$$

Hence, we have the following least squares:

- $\widetilde{\beta}^{ridge} = argmin_{\beta}||X^Ty - \beta||_2^2 + \lambda||\beta||_2^2$
- $\widetilde{\beta}^{lasso} = argmin_{\beta}||X^Ty - \beta||_2^2 + \lambda||\beta||_1$

So we find that if predictor matrix X is orthogonal, the least square problems (and their solutions) will have the same formula as if X was not orthogonal, with y replaced by X^Ty

Part D

Assume that predictor matrix $X \in \mathbb{R}^{n \times p}$ is orthogonal, and $y \in \mathbb{R}^n$:

We know that the solution to the ordinary least square problem is:

$$\widetilde{\beta}^{OLS} = (X^TX)^{-1}X^Ty$$
$$\widetilde{\beta}^{OLS} = (I)^{-1}X^Ty = X^Ty$$
$$\widetilde{\beta}^{OLS} = X^Ty$$

From part G in question 1:

$$\widetilde{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^Ty$$
$$\widetilde{\beta}^{ridge} = (I + \lambda I)^{-1}X^Ty$$
$$\widetilde{\beta}^{ridge} = \frac{X^Ty}{1+\lambda} = \frac{\widetilde{\beta}^{OLS}}{1+\lambda}$$

$$\widetilde{\beta}^{ridge} = \frac{\widetilde{\beta}^{OLS}}{1+\lambda}$$

Let $\lambda \in \mathbb{R} \geq 0$

For ridge regression:

- If $lambda = 0$, the solution to ridge regression is equivalent to ordinary least squares
- If $lambda > 0$, the ridge regression coefficients will become smaller than ordinary least squares coefficient as λ increases in value

Hence the λ parameter controls the ridge regression's coefficient penalization's strength. The greater the value, the greater the penalty

For lasso regression:

We take the answer from part B, where the solution to the lasso regression is

$$\widetilde{\beta}^{lasso} = sign(y_i) \cdot max(|y_i| - \frac{\lambda}{2}, 0)$$

And replace y_i with $\widetilde{\beta}^{OLS}$:

$$\widetilde{\beta}^{lasso} = sign(\widetilde{\beta}^{OLS}) \cdot max(|\widetilde{\beta}^{OLS}| - \frac{\lambda}{2}, 0)$$

- If $\widetilde{\beta}^{lasso} < 0$, then the lasso coefficient is always negative if $|\widetilde{\beta}^{OLS}| < \frac{lambda}{2}$, zero otherwise. Hence there is a range of $\widetilde{\beta}^{OLS}$ values where the lasso coefficient is forced to zero
- If $\widetilde{\beta}^{lasso} = 0$, then the solution to the lasso regression is always zero
- If $\widetilde{\beta}^{lasso} > 0$, then the lasso coefficient is always positive if $|\widetilde{\beta}^{OLS}| > \frac{lambda}{2}$, zero otherwise. Similar to when $\widetilde{\beta}^{lasso} < 0$, there is a certain range of OLS coefficient values where lasso will be force to zero

The λ parameter in lasso controls the coefficient penalization by extending or contracting the range where lasso coefficients are forced to zero

Question 3

Part A

The answers to this question has been placed in the script `takeHome_christopherlang.R` script included in the submittal

Under each `#TODO` comment, there is a `||-ANSWER-||` tag that contains the filled in code

Part B

The plots are located in the `takeHome_christopherlang.R` script, under *//–ANSWER–// for part B* section

In regards to the selected λ, answers are below

- For the ridge regression:
  - The *usual rule* selected $\lambda = 4$
  - The *one standard error rule* selected $\lambda = 15$
- For the lasso regression:
  - The *usual rule* selected $\lambda = 0.057$
  - The *one standard error rule* selected $\lambda = 0.120$

We can get both the *usual rule* and *one standard error rule* mean cross validation errors, for both methods. The code used to get these values remain under the same section

For ridge regression, we find that:

- The mean CV error for *usual rule* is 271.1551
- The mean CV error for *one standard error rule* is 282.0665

For lasso regression, we find that:

- The mean CV error for *usual rule* is 247.844
- The mean CV error for *one standard error rule* is 255.2767

Between the two methods, lasso regression had lower CV error overall, according to both rules

Part C

The code used is located in the `takeHome_christopherlang.R` script, under *//–ANSWER–// for part C* section

When looking at just the lasso images, the main difference between the usual rule and standard error rule is that the usual rule is slightly more noisy than standard error, but more saturated in coloration (specifically the darker areas). Which image is found to look better is a bit subjective, but we found that the usual rule image is better. Though it is more noisy (though only slightly), it seems closer to the original `bstar` image due to the closer coloration it has

Between the ridge and lasso images, the difference is much more stark however. The ridge images is simply far too noisy, which causes it to lose details in the shape, as well as reducing the color's saturation compared to the lasso images

Ultimately, the lasso images are much closer to the original `bstar` image than ridge, primarily due to the closer color match it has to the original. As stated before, of the four images, the lasso usual rule image is found to be the closest to the original `bstar` image

Part D

The code is located in the `takeHome_christopherlang.R` script, under *//–ANSWER–// for part D* section

We compute the total sum of squared error for both ridge and lasso, and for both usual and standard error rule, against the original `bstar` coefficients:

- Ridge regression, usual rule $SSE = 226.3899$
- Ridge regression, standard error rule $SSE = 245.3781$
- Lasso regression, usual rule $SSE = 148.6028$
- Lasso regression, standard error rule $SSE = 159.603$

Of the four, the lasso regression with usual rule λ selection is found to the have the lowest total sum of squared error, and therefore has the closest prediction of the original `bstar` coefficients