Chapter 3

Abhimanyu Talwar

November 2, 2018

ESL Problem 3.3

(a) Let the least squares estimate of parameter $a^T\beta$ be represented as h^Ty where $h^T = a^T(X^TX)^{-1}X^T$. We have:

$$
\begin{aligned}
Var(c^Ty) &= Var(c^Ty - h^Ty + h^Ty) && (1)\\
&= Var(h^Ty) + Var(c^Ty - h^Ty) + 2Cov(c^Ty - h^Ty, h^Ty) && (2)
\end{aligned}
$$

Now I will show that $Cov(c^Ty - h^Ty, h^Ty) = 0$. That result, combined with the fact that variance of a random variable is non-negative, will prove that $Var(c^Ty) \geq Var(h^Ty)$. Using the fact that $y = X\beta + \epsilon$ where ϵ is an $N \times 1$ vector of random errors. We have:

$$
\begin{aligned}
Cov(c^Ty - h^Ty, h^Ty) &= Cov((c-h)^Ty, h^Ty) && (3)\\
&= Cov((c-h)^T(X\beta + \epsilon), h^T(X\beta + \epsilon) && (4)
\end{aligned}
$$

Now I will prove some intermediate results:

$$
\begin{aligned}
h^TX\beta &= a^T(X^TX)^{-1}X^TX\beta && (5)\\
&= a^T\beta && (6)
\end{aligned}
$$

And further:

$$
\begin{aligned}
\mathbb{E}\left[c^Ty\right] &= a^T\beta && (7)\\
\implies \mathbb{E}\left[c^T(X\beta + \epsilon)\right] &= a^T\beta && (8)\\
\implies c^TX\beta &= a^T\beta && (9)
\end{aligned}
$$

Substituting results from Eq. 6 and Eq. 9 in Eq. 4, we get:

$$
\begin{aligned}
Cov(c^Ty - h^Ty, h^Ty) &= Cov((c-h)^T\epsilon, a^T\beta + h^T\epsilon) && (10)\\
&= \mathbb{E}\left[((c-h)^T\epsilon)(h^T\epsilon)\right] && (11)\\
&= \mathbb{E}\left[\epsilon^T(c-h)h^T\epsilon\right] && (12)\\
& && (13)
\end{aligned}
$$

Now I will use the Linear Algebra identity that $\mathbb{E}\left[Z^TAZ\right] = trace(A\Sigma_Z) + \mathbb{E}\left[Z^T\right]A\mathbb{E}\left[Z\right]$ for an $N \times 1$ random vector Z and a non-random matrix A. Also, using the fact that the expectation of ϵ is a zero vector, we get:

$$
\begin{aligned}
Cov(c^Ty - h^Ty, h^Ty) &= trace((c-h)h^T\sigma^2 I) && (14)\\
&= \sigma^2(c-h)^Th && (15)\\
&= \sigma^2(c^Th - h^Th) && (16)\\
&= \sigma^2(c^TX(X^TX)^{-1}a - a^T(X^TX)^{-1}X^TX(X^TX)^{-1}a) && (17)\\
&= \sigma^2(a^T(X^TX)^{-1}a - a^T(X^TX)^{-1}a) && (18)\\
&= 0 && (19)
\end{aligned}
$$

Using Eq. 19 and Eq. 2, we can finally conclude that:

$$
\begin{aligned}
Var(c^T y) &= Var(h^T y) + Var(c^T y - h^T y) && (20)\\
&>= Var(h^T y) && (21)
\end{aligned}
$$

(b) Let $\hat{\beta}$ represent the OLS estimate of β, and let $\hat{V} = Var(\hat{\beta})$ represent its variance-covariance matrix. Let $\tilde{\beta}$ be some other unbiased linear estimate of β and let $\tilde{V}$ be its variance-covariance matrix. I will show that $\tilde{V} - \hat{V}$ is positive semi-definite by showing that for any vector $z \neq 0$, the scalar quantity $z^T(\tilde{V} - \hat{V})z \geq 0$.

$$
\begin{aligned}
z^T(\tilde{V} - \hat{V})z &= z^T\tilde{V}z - z^T\hat{V}z && (22)\\
&= Var(z^T\tilde{\beta}) - Var(z^T\hat{\beta}) && (23)
\end{aligned}
$$

Now we know from Part (a) that the parameter $a^T\hat{\beta}$ is the Best Linear Unbiased Estimator, therefore it follows that $Var(z^T\hat{\beta}) \leq Var(z^T\tilde{\beta})$. Using this result in Eq. 23, we deduce that $z^T(\tilde{V} - \hat{V})z \geq 0$ for any $z \neq 0$. This proves that $(\tilde{V} - \hat{V})$ is Positive Definite.

ESL Problem 3.11

Before any calculations, I will first define various matrices and vectors along with their dimensions:

1. **X**: The design matrix of dimensions $N \times (p+1)$ where N is the number of samples and p is the number of features.
2. $\mathbf{X}_i$: This represents the i^{th} row of X. This has dimensions $1 \times (p+1)$.
3. **Y**: The response variable matrix of dimensions $N \times K$. Here K is the number of outputs for the multivariate regression problem.
4. $\mathbf{Y}_i$: this represents the transpose of the i^{th} row of Y. This is a column vector of dimensions $K \times 1$.
5. **B**: This represents the $(p+1) \times K$ matrix of regression coefficient estimates.
6. Σ: The $K \times K$ dimensional covariance matrix for errors.

Now the squared error is defined as:

$$
\begin{aligned}
RSS(B, \Sigma) &= \sum_{i=1}^{N} (Y_i - B^T X_i^T)^T \Sigma^{-1} (Y_i - B^T X_i^T) && (24)\\
&= \sum_{i=1}^{N} \left(Y_i^T \Sigma^{-1} Y_i - Y_i^T \Sigma^{-1} B^T X_i^T - X_i B \Sigma^{-1} Y_i + X_i B \Sigma^{-1} B^T X_i^T \right) && (25)
\end{aligned}
$$

Now I want to calculate the gradient of the expression in Eq. 25 with respect to the matrix B. Let me calculate this for one particular i first, and then I will sum over all values of i. There are four sub-expressions in Eq. 25:

1. **Sub-expression 1**: $Z_1 = Y_i^T \Sigma^{-1} Y_i$
   The gradient is a $(p+1) \times K$ matrix filled with 0s because this sub-expression is not influenced by the matrix B.
2. **Sub-expression 2**: $Z_2 = Y_i^T \Sigma^{-1} B^T X_i^T$
$$\frac{\partial Z_2}{\partial B} = X_i^T Y_i^T \Sigma^{-1} \tag{26}$$
3. **Sub-expression 3**: $Z_3 = X_i B \Sigma^{-1} Y_i$
$$\frac{\partial Z_3}{\partial B} = X_i^T Y_i^T \Sigma^{-1} \tag{27}$$
4. **Sub-expression 4**: $Z_4 = X_i B \Sigma^{-1} B^T X_i^T$
$$\frac{\partial Z_4}{\partial B} = 2 X_i^T X_i B \Sigma^{-1} \tag{28}$$

Setting the gradients of $RSS(B, \Sigma)$ with respect to B equal to 0, and using these four sub-expressions, we get:

$$\sum_{i=1}^{N} \left(-X_i^T Y_i^T \Sigma^{-1} - X_i^T Y_i^T \Sigma^{-1} + 2X_i^T X_i B \Sigma^{-1}\right) = 0 \tag{29}$$

$$\implies \sum_{i=1}^{N} \left(-2X_i^T Y_i^T \Sigma^{-1} + 2X_i^T X_i B \Sigma^{-1}\right) = 0 \tag{30}$$

$$\implies \left(\sum_{i=1}^{N} \left(-2X_i^T Y_i^T + 2X_i^T X_i B\right)\right) \Sigma^{-1} = 0 \tag{31}$$

Now in Eq. 31, since we assumed that the sample covariance matrix is invertible, we can say that Σ^{-1} is Positive Definite, and so its null-space only consists of the zero vector. So for some matrix A, if $A\Sigma^{-1} = 0$, it implies that each row of A is a zero-vector, and so the matrix A is filled with zeroes. Therefore we have:

$$\sum_{i=1}^{N} \left(-2X_i^T Y_i^T + 2X_i^T X_i B\right) = 0 \tag{32}$$

$$\implies -X^T Y + X^T X B = 0 \tag{33}$$

$$\implies B = (X^T X)^{-1} X^T Y \tag{34}$$

Now if covariance matrices are different for each observation, I will not be able to get rid of the covariance matrices like I did above after Eq. 31. The solution coefficients in that case will depend on the individual covariance matrices, and I'm unsure if a closed form solution exists in that case.

ESL Problem 3.12

Let X be the $N \times p$ dimensional design matrix and let y be the $N \times 1$ dimensional response vector. Let I_p denote the $p \times p$ Identity matrix and let O_p denote a column vector of zeroes of dimensions $p \times 1$. Define the $(N+p) \times p$ dimensional augmented matrix X^* as follows:

$$X^* = \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix} \tag{35}$$

Define the augmented response variable y^* of dimensions $(N+p) \times 1$ as:

$$y^* = \begin{bmatrix} y \\ O_p \end{bmatrix} \tag{36}$$

Now our estimated coefficients β_{Ridge} are basically solutions to the following Ordinary Least Squares (OLS) minimization problem.

$$\beta_{Ridge} = argmin_{\beta} |X^* \beta - y^*|^2 \tag{37}$$

We already know that a closed form solution exists for the OLS problem which is given by:

$$\beta_{Ridge} = (X^{*T} X^*)^{-1} X^{*T} y^* \tag{38}$$

$$= \left(\begin{bmatrix} X^T & \sqrt{\lambda} I_p \end{bmatrix} \begin{bmatrix} X^T \\ \sqrt{\lambda} I_p \end{bmatrix}\right)^{-1} \begin{bmatrix} X^T & \sqrt{\lambda} I_p \end{bmatrix} \begin{bmatrix} y \\ O_p \end{bmatrix} \tag{39}$$

$$= (X^T X + \lambda I_p)^{-1} X^T y \tag{40}$$

Hence, Eq. 40 proves that the coefficient estimates for Ridge Regression may be derived by OLS for augmented data.

ESL Problem 3.28

Let X_j be the feature that we duplicate and let X_{-j} denote all other features except X_j. Let β_j denote the coefficient of X_j in the original Lasso problem, and let β_{-j} denote all the other coefficients. Then the **original Lasso problem** be written as the following optimization problem:

$$minimize_{\beta} \, \|Y - X_{-j}\beta_{-j} - X_j \beta_j\|_2^2 \tag{41}$$

$$s.t. \, \|\beta_{-j}\|_1 + |\beta_j| \leq t \tag{42}$$

Let X_j^* denote the duplicated feature and let $\tilde{\beta}_j$ and β_j^* denote the coefficients of the original feature X_j and the duplicated feature X_j^* in the new Lasso problem. Let $\tilde{\beta}_{-j}$ denote the coefficients of other feature vectors in the new Lasso problem. Then the **updated Lasso problem** can be written as:

$$minimize_\beta \left\| Y - X_{-j}\tilde{\beta}_{-j} - X_j\tilde{\beta}_j - X_j^*\beta_j^* \right\|_2^2 \tag{43}$$

$$s.t. \left\|\tilde{\beta}_{-j}\right\|_1 + |\tilde{\beta}_j| + |\beta_j^*| \leq t \tag{44}$$

Now say for a particular solution to the updated Lasso problem, our coefficients are: $\tilde{\beta}_{-j}$, $\tilde{\beta}_j$ and β_j^*. Now if we choose $\beta_{-j} = \tilde{\beta}_{-j}$, and $\beta_j = \tilde{\beta}_j + \beta_j^*$, then I claim that this set of β_{-j} and β_j is also a solution to the original Lasso problem. Using **Triangle Inequality** (i.e. $|a+b| \leq |a| + |b|$) in Eq. 44 we get:

$$\left\|\tilde{\beta}_{-j}\right\|_1 + |\tilde{\beta}_j| + |\beta_j^*| \leq t \tag{45}$$

$$\implies \left\|\tilde{\beta}_{-j}\right\|_1 + |\tilde{\beta}_j + \beta_j^*| \leq t \tag{46}$$

But we already know that for the given value of t, the optimal coefficient of X_j for the original Lasso problem is $\beta_j = a$. Therefore, this new coefficient $\tilde{\beta}_j + \beta_j^*$ also has to equal a. Further, we also know from constraint in Eq. 44 that the absolute value of each individual coefficient can never exceed t. **Therefore we conclude the solution set for coefficients of X_j and X_j^* is characterized by the following line segment**:

$$\boxed{\tilde{\beta}_j + \beta_j^* = a} \tag{47}$$

$$\text{subject to } |\tilde{\beta}_j| \leq t, |\beta_j^*| \leq t \tag{48}$$

ESL Problem 3.29
Let $X = (x_1, x_2, \cdots, x_N)$ denote a column vector containing N sample values of the single dimensional feature in this problem. Let $y = (y_1, y_2, \cdots, y_N)$ be a column vector of size N containing the response variable.

I will first prove the general case where we have M copies of X in our training set, and then I will use it for $M = 2$ to derive expressions of coefficients for the case where we have one exact copy of X.

Let $X^* = [X_1, X_2, \cdots, X_m]$ denote our design matrix of dimensions $(N \times M)$ in which each X_i is an exact copy of X. The Ridge coefficients are given by:

$$\beta_M = (X^{*T}X + \lambda I)^{-1}X^{*T}y \tag{49}$$

The matrix inverse can be written as:

$$\left(X^{*T}X + \lambda I\right)^{-1} = \left(\left(\sum_{i=1}^N x_i^2\right)\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} + \lambda I\right)^{-1} \tag{50}$$

$$= \frac{1}{a}\left(P + \frac{\lambda}{a}I\right)^{-1} \tag{51}$$

Here P denotes the $(M \times M)$ matrix each entry of which equals 1. The quantity $a = \sum_{i=1}^N x_i^2$. Now to find the inverse of the matrix in Eq. 51, I have taken help from this Stackexchange link, however the calculations are my own. Let the inverse be of the form $\left(kP + \frac{a}{\lambda}I\right)$ where k is a quantity which I will solve for. I will use the fact that $P^2 = MP$. Then we have:

$$\left(P + \frac{\lambda}{a}I\right)\left(kP + \frac{a}{\lambda}I\right) = I \tag{52}$$

$$\implies kP^2 + \frac{\lambda}{a}kP + I + \frac{a}{\lambda}P = I \tag{53}$$

$$\implies kMP + \left(\frac{\lambda}{a}k + \frac{a}{\lambda}\right)P = 0 \tag{54}$$

$$\implies k = \frac{-a^2}{\lambda(aM + \lambda)} \tag{55}$$

Substituting the value of k from Eq. 55 in our expression for inverse, i.e. $\left(kP + \frac{a}{\lambda}I\right)$, and plugging it in Eq. 51, we finally get:

$$\left(X^{*T}X + \lambda I\right)^{-1} = \frac{-a}{\lambda(aM + \lambda)}P + \frac{1}{\lambda}I \tag{56}$$

$$\tag{57}$$

Now I will substitute the result derived in Eq. 57 in our original equation for solutions of Ridge coefficients, i.e. Eq. 49. But before that, note that $X^{*T}y$ can be written as a column vector bW of size M, where W denotes a column vector of size M with each entry 1 and $b = \sum_{i=1}^{N} x_i y_i$. So finally we have from Eq. 49:

$$\beta_M = \left(\frac{-a}{\lambda(aM + \lambda)}P + \frac{1}{\lambda}I\right) bW \tag{58}$$

$$= \frac{-ab}{\lambda(aM + \lambda)}PW + \frac{b}{\lambda}W \tag{59}$$

$$= \frac{-abM}{\lambda(aM + \lambda)}W + \frac{b}{\lambda}W \tag{60}$$

$$= \frac{b}{aM + \lambda}W \tag{61}$$

$$\tag{62}$$

Since W is simply a column vector of M ones, Eq. 62 proves that each coefficient is simply equal to $\frac{b}{aM+\lambda}$.

Now coming back to the case where $M = 2$, that is we have one exact copy of X, the value of each of the two Ridge coefficients is given by:

$$\boxed{\beta = \frac{b}{2a + \lambda}} \tag{63}$$

As defined earlier, $a = \sum_{i=1}^{N} x_i^2$ and $b = \sum_{i=1}^{N} x_i y_i$.