רגרסיה ומודלים סטטיסטיים- פתרון תרגיל 4

שאלה 1:

נתון שהמטריצה $X \in R^{n\times(p+1)}$ מדרגה מלאה. הניחו שהנחות המודל הלינארי מתקיימות.

חשבו את $cov(e), cov(\epsilon, e), cov(\hat{\beta}, \epsilon), Cov(Y_i, \hat{Y}_i)$ וציינו את המימד של כל אחד מהגדלים.

פתרון :

$$cov(e) = cov\left((I - P_X)Y\right) = (I - P_X)cov(Y)(I - P_X)^T = \sigma^2(I - P_X) \in R^{n\times n}$$
$$cov(\epsilon, e) = cov(\epsilon, (I - P_X)Y) = cov(\epsilon, Y)(I - P_X)^T = cov(\epsilon, X\beta + \epsilon)(I - P_X)^T$$
$$= cov(\epsilon)(I - P_X) = \sigma^2(I - P_X) \in R^{n\times n}$$
$$cov\left(\epsilon, \hat{\beta}\right) = cov(\epsilon, (X^TX)^{-1}X^TY) = cov(\epsilon, X\beta + \epsilon)X(X^TX)^{-1} = \sigma^2X(X^TX)^{-1} \in R^{n\times(p+1)}$$
$$cov\left(Y_i, \hat{Y}_i\right) = cov(Y, P_XY)_{ii} = [cov(Y)P_X^T]_{ii} = \sigma^2 P_{X_{ii}} \in R^{n\times n}$$

שאלה 2:

נתון מודל ליניארי, $Y = X\beta + \varepsilon$ כאשר $\varepsilon \sim (0, \sigma^2 I_n)$.

עבור $u, v \in \mathbb{R}^n$ כלשהם נגדיר את נורמת מהלנוביס (Mahalanobis) באופן הבא:

$$||u - v||_\Sigma = \sqrt{(u - v)^T \Sigma^{-1} (u - v)}$$

כאשר $\Sigma \in \mathbb{R}^{n \times n}$ מטריצה סימטרית חיובית.

כמו כן נגדיר את $\hat{\beta}^\Sigma$ באופן הבא:

$$\hat{\beta}^\Sigma = \underset{\beta \in \mathbb{R}^{p+1}}{argmin} ||Y - X\beta||_\Sigma^2$$

א. הראו שמתקיים

$$\hat{\beta}^\Sigma = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

והסבירו את האומד שמתקבל במקרה $\Sigma = I_n$.

<u>הדרכה:</u> אפשר (לא חייב ש) להעזר בעובדה הבאה: $A \in \mathbb{R}^{m \times m}$ מטריצה סימטרית חיובית, אזי קיימת מטריצה $B \in \mathbb{R}^{m \times m}$ סימטרית חיובית כך ש: $B^T B = A$. ניתן להפעיל מסקנה זו על Σ כך שיהיה ניתן לייצג את Σ^{-1} באופן הבא: $\Sigma^{-1} = C^T C$ עבור מטריצה C ריבועית כלשהי. מצאו מי היא C, הסבירו מדוע היא מוגדרת היטב, ואז הגדירו משתנים חדשים

$\tilde{X} = \boldsymbol{C}X \qquad \tilde{Y} = \boldsymbol{C}Y,$ ופתרו את הבעיה (שימו לב שעבור ה **C** הנכון הפתרון מיידי, חישבו מדוע).

ב. הראו כי $\hat{\beta}^\Sigma$ הינו אומד חסר הטיה לβ, וכן מצאו את $cov(\hat{\beta}^\Sigma)$.

ג. מה אומר משפט גאוס מרקוב לגבי המטריצה$cov(\hat{\beta}^\Sigma) - cov(\hat{\beta})$, כאשר $\hat{\beta}$ אומד הריבועים הפחותים?

ד. כעת הניחו מודל ליניארי כללי יותר: $Y = X\beta + \varepsilon$ כאשר $\varepsilon \sim (0, \sigma^2 \Sigma)$ עבור אותה מטריצה סימטרית חיובית $\Sigma \in \mathbb{R}^{n \times n}$ שביחס אליה מוגדר האומד $\hat{\beta}^\Sigma$ (שימו לב ש Σ מטריצה סימטרית חיובית כלשהי, זהו אכן מודל כללי יותר שכן עבור $\Sigma = I_n$ מתקבל המודל "הרגיל").

הראו שבמקרה זה $\hat{\beta}^\Sigma$ הינו האומד הליניארי חסר ההטיה הטוב ביותר מבחינת שגיאה ריבועית מבין האומדים הליניאריים חסרי ההטיה לβ. כלומר, שלכל צירוף ליניארי $\theta = a^T \beta$ האומד $\hat{\theta}^\Sigma = a^T \hat{\beta}^\Sigma$ משיג שונות מינימאלית מבין כל האומדים הליניאריים חסרי ההטיה ל θ.

<u>הדרכה:</u> השתמשו בהדרכה מסעיף א', כתבו את המודל במונחי $\tilde{X}$, $\tilde{Y}$. הראו שמתקבל המודל הליניארי "הרגיל" ומשם ההוכחה שקולה להוכחת משפט גאוס מרקוב "הרגיל".

פתרון :

52371 – סמסטר ב' תשפ"ב – בוחן 1

שאלה 1.

(א). Σ הפיכה (לא רק אי-שלילית) לפי הנחה, לכן קיימת הצגה

$$\Sigma = UDU^T, \qquad (1)$$

כאשר $U \in \mathbb{R}^{n\times n}$ אורתוגונלית, וכאשר $D = diag(d_1, \dots, d_n)$ עם $d_i > 0$.

נגדיר עכשיו

$$D^{-1/2} := diag(d_1^{-1/2}, \dots, d_n^{-1/2}), \quad C := UD^{-1/2}U^T,$$

ונשים לב שמתקיים:

$$C^TC = (UD^{-1/2}U^T)^T UD^{-1/2}U^T = UD^{-1/2}\cancel{U^TU}D^{-1/2}U^T =$$
$$= UD^{-1}U^T = (UDU^T)^{-1} = \Sigma^{-1}$$

ובמעבר הלפני-אחרון השתמשנו בעובדה ש- $U^{-1} = U^T$ (כי U אורתוגונלית).

נגדיר טרנספורמציה של הבעיה ע"י

$$\tilde{X} := CX, \quad \tilde{Y} := CY. \qquad (2)$$

אז מתקיים:

$$\begin{aligned}\|Y - Xb\|_\Sigma^2 &= (Y - Xb)^T \Sigma^{-1} (Y - Xb) = \\ &= (Y - Xb)^T C^TC (Y - Xb) \\ &= \|C(Y - Xb)\|_2^2 \\ &= \|\tilde{Y} - \tilde{X}b\|_2^2\end{aligned}$$

והינה שבביטוי האחרון מופיעה הנורמה האוקלידית (הרגילה), קיבלנו לפונקציית המטרה של הריבועים הפחותים, והפתרון:

$$\hat{\beta}^\Sigma = \underset{b}{\operatorname{argmin}} \|Y - Xb\|_\Sigma^2 = \underset{b}{\operatorname{argmin}} \|\tilde{Y} - \tilde{X}b\|_2^2 = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} =$$
$$= [(CX)^TCX]^{-1}(CX)^TCY$$

$$= (X^T C^T C X)^{-1} X^T C^T C Y$$
$$= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

כנדרש. <u>הערה</u>: שים לב שהטענה בסעיף ג' (ובעצם גם ההוכחה) תקפה לכל ערך <u>קבוע</u> Y, X (בהנחה ש- $X^T\Sigma^{-1}X$ הפיכה) — אין פה שום חישוב <u>סטטיסטי</u>.

ד.
$$E[\hat\beta^{\Sigma}] = E[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y] =$$
$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}E[Y] =$$
$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X\beta = \beta \qquad (\forall\beta\in\mathbb{R}^{p+1}).$$

$$cov(\hat\beta^{\Sigma}) = cov[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y] =$$
$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}[cov(Y)]\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$$
$$= \sigma^2(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$$
$$= \sigma^2(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-2}X(X^T\Sigma^{-1}X)^{-1}$$

ה. אם המודל נכון, משפט גאוס-מרקוב אומר שאומד הריבועים הפחותים
$$\hat\beta = (X^TX)^{-1}X^TY$$
הוא האומד הלינארי חסר-ההטייה הטוב ביותר מבחינת MSE, כלומר לכל אומד לינארי חסר-הטייה $\tilde\beta$, מתקיים שלכל וקטור קבוע $a\in\mathbb{R}^{p+1}$,

$$Var(a^T\hat\beta) \le Var(a^T\tilde\beta)$$
$$\Leftarrow \quad cov(a^T\hat\beta) \le cov(a^T\tilde\beta)$$
$$\Leftarrow \quad a^T cov(\hat\beta)a \le a^T cov(\tilde\beta)a$$
$$\Leftarrow \quad 0 \le a^T[cov(\tilde\beta) - cov(\hat\beta)]a$$

ולפי הגדרה זה אומר שהמטריצה $[cov(\tilde\beta) - cov(\hat\beta)]$ היא אי-שלילית.
מכיוון ש- $\hat\beta^{\Sigma}$ אומד לינארי חסר-הטייה, בפרט נקבל ש:

$$\mathrm{cov}(\hat\beta^{\Sigma}) - \mathrm{cov}(\hat\beta)$$

מטריצה אי-שלילית.

3. עכשיו המודל הוא:

$$Y = X\beta + \epsilon, \qquad \epsilon \sim (0, \sigma^2 \Sigma)$$

עבור מטריצה Σ שביחס אליה מוגדר $\hat\beta^{\Sigma}$.

עבור הטרנספורמציה של הבעיה ל- $\tilde X$ ו- $\tilde Y$ כפי שהוגדרה במשוואה (2), מתקיים

$$\tilde Y = CY = C(X\beta + \epsilon) = CX\beta + C\epsilon = \tilde X\beta + \tilde\epsilon$$

עבור $\tilde\epsilon := C\epsilon$. כמו כן, $E[\tilde\epsilon] = CE[\epsilon] = 0$, וגם:

$$\begin{aligned}
\mathrm{cov}(\tilde\epsilon) = \mathrm{cov}(C\epsilon) = C[\mathrm{cov}(\epsilon)]C^T &= \\
&= \sigma^2 C\Sigma C^T = \\
&= \sigma^2 UD^{-1/2}U^T(UDU^T)(UD^{-1/2}U^T)^T \\
&= \sigma^2 UD^{-1/2}U^T(UDU^T)UD^{-1/2}U^T \\
&= \sigma^2 U\underbrace{(D^{-1/2}DD^{-1/2})}_{=I_n}U^T \\
&= \sigma^2 UU^T = \sigma^2 I_n
\end{aligned}$$

במילים אחרות, עבור הטרנספורמציה של הבעיה מתקיים המודל

$$\tilde Y = \tilde X\beta + \tilde\epsilon, \qquad \tilde\epsilon \sim (0, \sigma^2 I_n) \qquad (3)$$

ואכן זהו המודל הלינארי "הרגיל", כלומר עם וקטור שגיאה בעל $\mathrm{cov} = \sigma^2 I_n$, ונשים לב ש- β ב-(3) זהו אותו וקטור β (לא ידוע) שמופיע ב-(1).

משפט גאוס-מרקוב עבור המודל (3) אומר שהאומד הלינארי חסר-ההטיה הטוב ביותר נתון ע"י

$$(\tilde X^T \tilde X)^{-1} \tilde X^T \tilde Y, \qquad (4)$$

Q.2.

As seen in class, the multivariate MSE is given by:

$$E||\hat{\theta} - \theta||^2 = tr(Var(\hat{\theta})) + ||E(\hat{\theta}) - \theta||^2$$

We already know that $\hat{\beta}_{OLS}$ is an unbiased estimator, and its variance matrix is given by $\sigma^2(X^TX)^{-1}$. From the previous calculations, we can compute the same terms for $\hat{\beta}Ridge$:

$$Var(\hat{\beta}_{Ridge}) = \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$$

Where its $bias^2$ is:

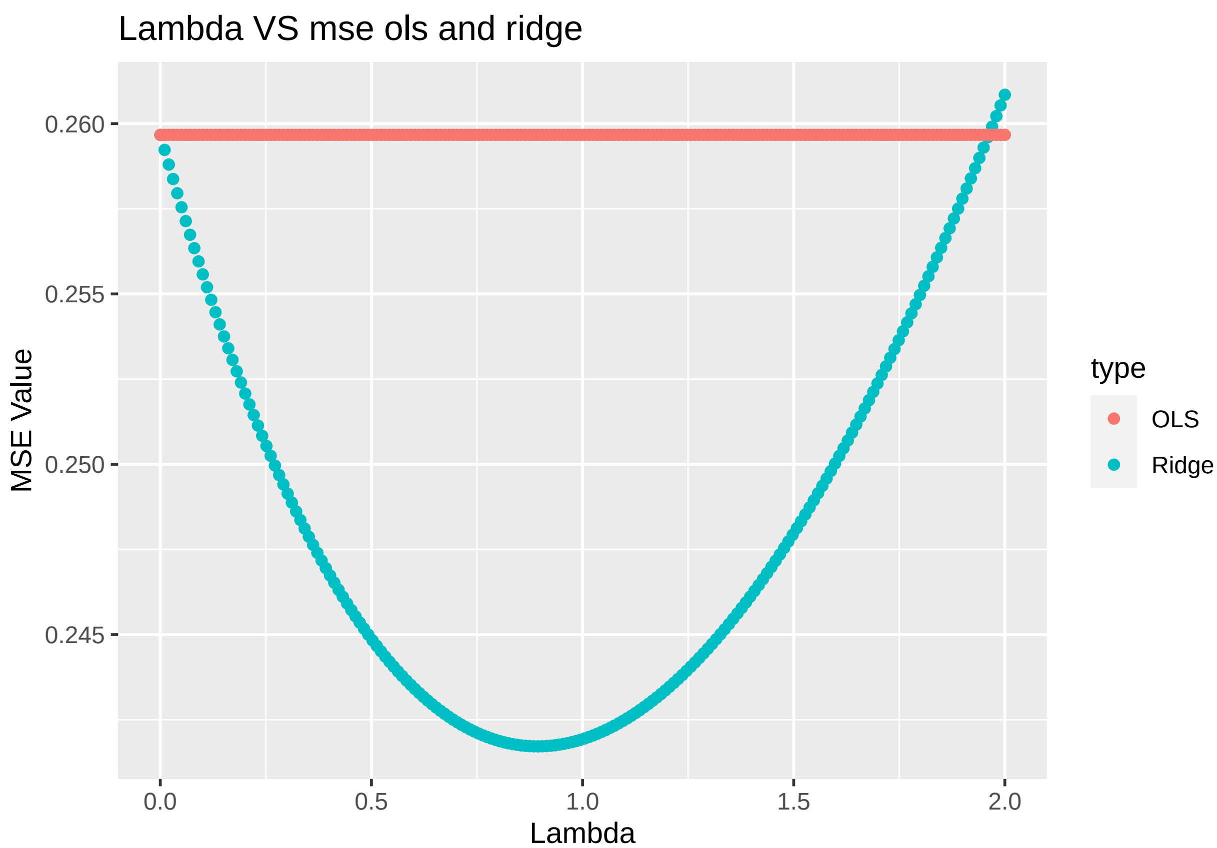
$$||(X^TX + \lambda I)^{-1}X^TX\beta - \beta||^2$$

Now we'll read and scale the data for estimations:

```
df <- read.csv("C:/Users/nivbr/Downloads/quiz2_df.csv")

#scaling
n = length(df$Y)
fac = sqrt(n / (n - 1))
for (i in c(3:9)){
  df[ ,i] = as.numeric(scale(df[,i])) * fac
}
x_matrix = cbind(df$X0,df$X1,df$X2,df$X3,df$X4,df$X5,df$X6,df$X7)
```

Estimations of for each value of λ with its corresponding MSE. Then, create the plot:

```
mse_ridge <- c()
for (i in lambda_seq){
  mse_ridge <- c(mse_ridge,ridge_aux_functions(x_matrix,df$Y,i,sigma_sq,beta)[2])
}

mse_ols <- c()
for (i in c(1:length(lambda_seq))){
  mse_ols <- c(mse_ols,ridge_aux_functions(x_matrix,df$Y,lambda_seq[i],sigma_sq,beta)[3])
}

type = c(rep("Ridge",200),rep("OLS",200))
value = c(unlist(mse_ridge),unlist(mse_ols))
data_for_ggplot <- data.frame(lambda = c(lambda_seq,lambda_seq),type = type,value=value )

ggplot(data_for_ggplot, aes(x=lambda, y = value, color=type)) +
  geom_point() + ylab("MSE Value") + xlab("Lambda") +
  labs(title="Lambda VS mse ols and ridge ")
```

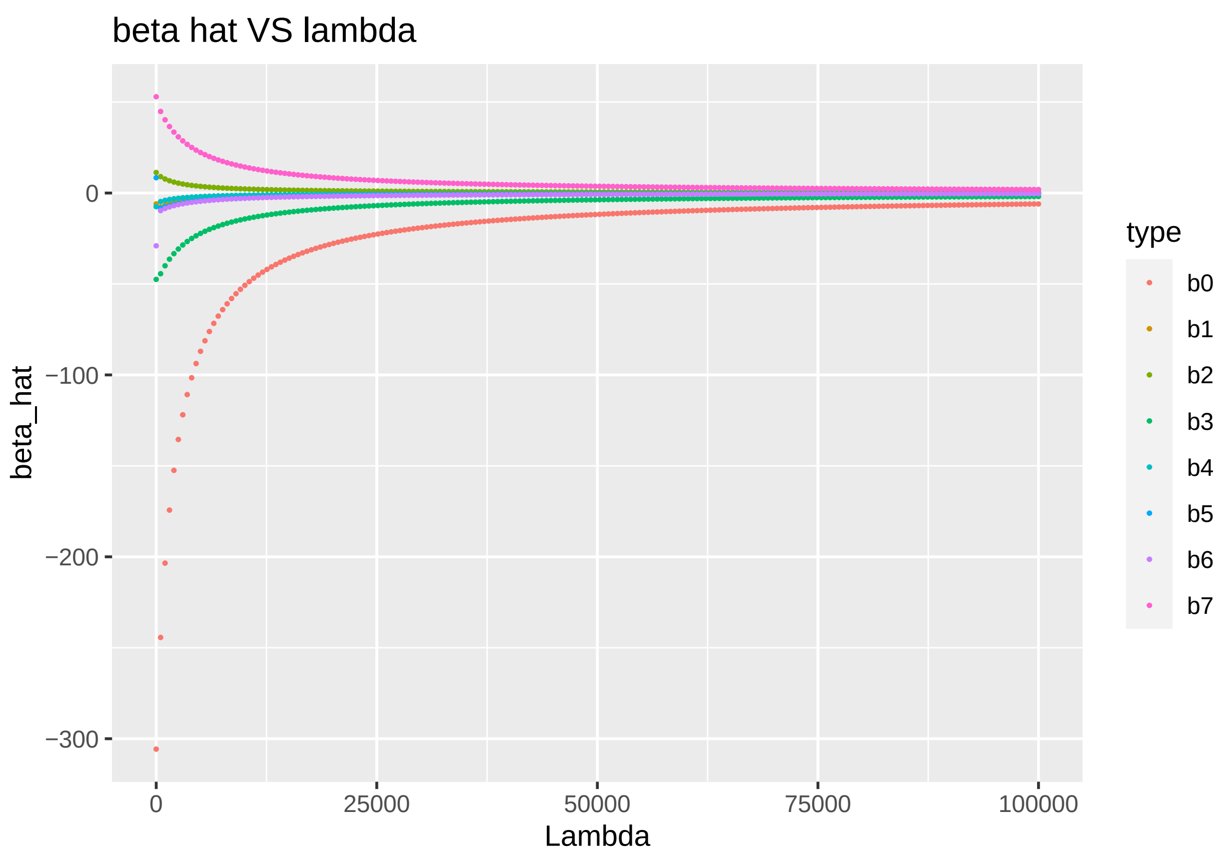


We can see in the plot that the MSE of the OLS is constant (as it does not change with different λ values). When λ is zero, the OLS MSE and the Ridge MSE are identical. The shape of the Ridge-MSE graph resembles a parabola. This means that there is a certain λ from which (around 1) the Ridge MSE starts to increase. In other words, the trade-off between the bias and the variance flips, and the added bias becomes greater than the decreasing variance. Using Ridge regression with the given λ penalties (which are between 0 and 1) does seem to produce a smaller MSE in this particular model and might be preferable to OLS regression. (This actually happens because this particular data is highly multicollinear, which increases the estimator variances).

Next we'll examine the shrinking effect of the λ-s size on the $||\hat{\beta_{Ridge}}||$. As we saw previously, this norm is a monotonically decreasing function of λ:

```
beta_hat_func <- function(X, Y, lambda){
  I <- diag(ncol(X))
  beta_hat_ols <- solve(t(X)%*%X) %*% t(X)%*%Y
  beta_hat_ridge <- solve(t(X)%*%X + lambda*I) %*% t(X)%*%Y
  return(list(lambda = lambda, beta_hat_ridge = beta_hat_ridge)[2])
}


beta_hat_ridge <- c()
for (i in c(1:length(lambda_seq_mod))){
  beta_hat_ridge <- c(beta_hat_ridge,beta_hat_func(x_matrix,df$Y,lambda_seq_mod[i]))
}


#creating dataFrame for the plot
```

```
b0 <- sapply(beta_hat_ridge, "[[", 1)
b1 <- sapply(beta_hat_ridge, "[[", 2)
b2 <- sapply(beta_hat_ridge, "[[", 3)
b3 <- sapply(beta_hat_ridge, "[[", 4)
b4 <- sapply(beta_hat_ridge, "[[", 5)
b5 <- sapply(beta_hat_ridge, "[[", 6)
b6 <- sapply(beta_hat_ridge, "[[", 7)
b7 <- sapply(beta_hat_ridge, "[[", 8)

values <- c(b0,b1,b2,b3,b4,b5,b6,b7)
type = c(rep("b0",200),rep("b1",200),rep("b2",200),rep("b3",200),rep("b4",200),rep("b5",200),rep("b6",20
lambda2 <- rep(lambda_seq_mod,8)

data_for_q12 <- data.frame(lambda = lambda2,type=type,values=values)


ggplot(data_for_q12, aes(x=lambda, y = values, color=type)) + geom_point(size=0.3) + ylab("beta_hat") +
```



as expected, we can see that as lambda increases, the beta's estimators converge to zero. the ridge regression puts a penalty lambda on beta. when the panelty increases, the absolute value of beta decreases, as the beta adds little contribution to the model.

Question 3. For Q3, we need to use the following results:

- Let C_1, C_2 be positive definite matrices with spectral decompositions $C_1 = UD_1U^T$ and $C_2 = UD_2U^T$ respectively. Then
$$C_1 + C_2 = U(D_1 + D_2)U^T,$$
and
$$C_1C_2 = U(D_1D_2)U^T.$$
- Moreover, write $D_1 = diag(d_1, ..., d_p)$ and $D_2 = diag(\tilde{d}_1, ..., \tilde{d}_p)$. Then
$$D_1 + D_2 = diag(d_1 + \tilde{d}_1, ..., d_p + \tilde{d}_p),$$
and
$$D_1D_2 = diag(d_1\tilde{d}_1, ..., d_p\tilde{d}_p).$$
- For any unitary U ($UU^T = I$), the following holds:
$$I = UIU^T$$
- As a consequence, $C_1^{-1} = UD_1^{-1}U^T$, and $D_1^{-1} = diag(d_1^{-1}, ..., d_p^{-1})$.

(1) Note that,
$$X^TX + cI = UDU^T + cI = UDU^T + cUU^T = UD^*U^T$$
where
$$D^* = \begin{bmatrix} d_1 + c & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & d_p + c \end{bmatrix}$$
hence
$$\mathbb{E}(\hat{\beta}_{ridge}) = \mathbb{E}\left((X^TX + cI)^{-1}X^TY\right) = (X^TX+cI)^{-1}X^TX\beta+(X^TX+cI)^{-1}X^T\mathbb{E}[\varepsilon]$$
and taking $E[\varepsilon] = \mathbf{0}$ the second argument vanishes.

We get
$$(X^TX + cI)^{-1}X^TX\beta = (UD^*U^T)^{-1}UDU^T\beta = UD^{*-1}DU^T\beta,$$
where
$$D^{*-1}D = \begin{bmatrix} \frac{d_1}{d_1+c} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \frac{d_p}{d_p+c} \end{bmatrix}.$$

(2)
$$\text{cov}(\hat{\beta}_{ridge}) = \text{cov}\left((X^TX + cI)^{-1}X^T(X\beta + \varepsilon)\right) =$$
$$(X^TX + cI)^{-1}X^T \text{cov}(\varepsilon)((X^TX + cI)^{-1}X^T)^T = \sigma^2(X^TX + cI)^{-1}X^TX(X^TX + cI)^{-1} =$$
$$\sigma^2UD^{*-1}U^TUDU^TUD^{*-1}U^T = \sigma^2UD^{**}U^T,$$

where

$$D^{**} = \begin{bmatrix} \frac{d_1}{(d_1+c)^2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \frac{d_p}{(d_p+c)^2}. \end{bmatrix}$$

(3)

$$var(a^T\hat{\beta}_{ridge}) = \sigma^2 a^T U D^{**} U^T a$$

and

$$var(a^T\hat{\beta}_{OLS}) = \sigma^2 a^T (X^T X)^{-1} a.$$

So

$$var(a^T\hat{\beta}_{OLS}) - var(a^T\hat{\beta}_{ridge}) = \sigma^2 a^T U(D^{-1} - D^{**})U^T a.$$

It is sufficient to prove that $U(D^{-1} - D^{**})U^T$ is positive definite. This is true if and only if the eigen-values are positive, meaning the elements on the diagonal of $(D^{-1} - D^{**})$ are positive.

The $i'th$ element in $(D^{-1} - D^{**})$ is of the form:

$$(D^{-1} - D^{**})_{ii} = \frac{1}{d_i} - \frac{d_i}{(di+c)^2}.$$

Note that for any $d_i, c > 0$:

$$\frac{d_i}{(di+c)^2} \le \frac{d_i}{(d_i)^2} = \frac{1}{d_i}.$$

and the proof is complete.

According to Gauss Markov, $\hat{\beta}_{OLS}$ should be the best unbiased linear estimator for β, and in particular $a^T\hat{\beta}_{OLS}$ should be the minimal variance estimator for $a^T\beta$, However, $\hat{\beta}_{ridge}$ is biased (for any $\beta \neq \mathbf{0}$), and so is not covered by the conditions of Gauss-Markov.

To see the bias in $\hat{\beta}_{ridge}$, check the result from part 1:

$$\mathbb{E}[\hat{\beta}_{ridge}] = UD^{*-1}DU^T\beta.$$

We can therefore compute the bias:

$$\begin{aligned} &\mathbb{E}[\hat{\beta}_{ridge} - \beta] = UD^{*-1}DU^T\beta - UIU^T\beta = \\ &= U(D^{*-1}D - I)U^T\beta. \end{aligned}$$

Therefore, assuming $\beta \neq \mathbf{0}$, the bias is $\mathbf{0}$ if and only if $(D^{*-1}D - I) \equiv \mathbf{0}$. It's easy to see that the diagonal elements in $D^{*-1}D$ are $d_i/(d_i+c) \neq 1$ and the estimator is biased.