7 Multicollinearity. Ridge and LASSO. Model Selection for Prediction. Mallow's C_P Criterion

7.0.1 Perfect Multicollinearity

Definition 7.0.1: Perfect Multicollinearity

This defines the case where rank $(X) \neq k$ and GM1 is violated

$\Rightarrow \exists$ an exact linear dependence between the columns of X so that

$$X\alpha = 0 \quad \text{for some } \alpha \neq 0$$

Corollary 7.0.1. Models

$$Y = X\beta + \varepsilon \quad \text{and} \quad Y = X\tilde{\beta} + \varepsilon$$

where $\tilde{\beta} = \beta + \lambda\alpha$ for any λ are equivalent. Thus β is not identified as we cannot distinguish between the two.

Corollary 7.0.2. In the OLS case, rank $(X) \neq k$:

$\Rightarrow$ non invertability of $X'X$
$\Rightarrow$ infinite number of solutions to the OLS problem. $min_b||Y - Xb||^2$. Thus OLS estimator is not well defined.

Example. Dummy Variable Trap:

Consider the model $Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where X_0 is a constant vector $= \vec{1}$ and X_1, X_2 are dummies such that $X_2 = 1 - X_1$.

$\Rightarrow$ X_2 is a linear combination of X_1 and X_0 and thus model is not identified. Specifically, $X\alpha = 0, \text{for } \alpha = (1, -1, -1)'$

$\Rightarrow$ $X'X$ is not invertible.

$\Rightarrow$ OLS estimator is not well defined.

We also can have perfect multicollinearity with small sample sizes s.t. $n < k$. Thus rank $(X) \leq n < k$.

7.0.2 Imperfect Multicollinearity

Often there will exist a linear combination of X that is almost but not exactly 0

Definition 7.0.2

Imperfect multicollinearity

$$rank(X) = k, \quad X\alpha \approx 0 \quad \text{for some } \alpha \neq 0$$

But this is not a unit invariant quantity, instead we can look at the relative size of the eigenvalues of $X'X$.

$$\frac{\lambda_{max}}{\lambda_{min}}$$

Corollary 7.0.3. Multicollinearity will result in a large norm of the variance-covariance matrix of the OLS estimator $\sigma^2(X'X)^{-1}$ and thus a very large trace. Trace of the variance of OLS estimator yields:

$$tr(\sigma^2(X'X)^{-1}) = E(||\hat{\beta} - \beta||^2|X)$$

Thus large multicollinearity will mean a large expected squared distance between true and estimated value of parameters.

Note:-

Norm here is defined as $\max_{x\neq 0} \frac{||Ax||}{||x||}$, more details

Also:

$$\frac{\lambda_{max}}{\lambda_{min}}$$

represents the condition number $C = ||A||\,||A^{-1}||$ of matrix $X'X$ (if this matrix is positive definite and (given) symmetric)[a]. Intution behind condition number: For a set of simulateneous equations $Ax = b$ the condition number of A sets a bound of the sensitivity of the relative (i.e. invariant to units) solution error (in x) induced by errors in the problem vector (b). $\frac{||\Delta x||}{||x||} \leq C\frac{||\Delta b||}{||b||}$[b]. Thus if C is large, then the solution can be very sensitive to small changes in the problem. (Worst case when b points in the smallest eigenvector direction of A and Δb points in the largest eigenvector direction of A.)

Applying this to $X'X$, we can see that if the condition number is large, then the "solution error" in β can be very sensitive to small changes in the problem, i.e. small changes in $X'Y$, i.e. ε. Here, $Ax = b \equiv X'X\beta = X'Y$. This shows why it is a good condition to represent the large variance of the OLS estimator in the presence of multicollinearity.

[a] Proof in Appendix

[b] Proof in Appendix

7.1 Ridge Regression

By G-M the resulting large MSE is still best among any other conditionally unbiased estimator. Thus, solutions recommend minimising expected MSE via introducing bias and reducing variance.

Definition 7.1.1

Ridge Regression Estimator:

This solves the OLS problem but where size of $||\beta||^2 = \beta'\beta$ is penalised by λ.

$$\min_{\beta} \left(||Y - X\beta||^2 + \lambda||\beta||^2\right) \text{ for some } \lambda > 0$$

We can show this results in the estimate (for $k < n$):

$$\hat{\beta}_r = (X'X + \lambda I)^{-1}X'Y$$

Note:

$$E(\hat{\beta}_r|X) = (X'X + \lambda I)^{-1}X'X\beta$$

and thus $\hat{\beta}_r$ is biased. When λ rises the bias increases and the variance decreases, illustrating the exploitable bias-variance tradeoff.

$$Var(\hat{\beta}_r|X) = \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}$$

We can see this explicitly by looking at the condition number of $X'X + \lambda I$:[c]

$$= \frac{\lambda_{max} + \lambda}{\lambda_{min} + \lambda} < \frac{\lambda_{max}}{\lambda_{min}}$$

As $(X'X + \lambda I)\hat{\beta}_r = X'Y$, we can see that the relative solution error in $\hat{\beta}_r$ is bounded by:

$$\frac{||\Delta\hat{\beta}_r||}{||\hat{\beta}_r||} \leq \frac{\lambda_{max} + \lambda}{\lambda_{min} + \lambda} \frac{||\Delta X'Y||}{||X'Y||}$$

Thus, as λ rises, the solution error in $\hat{\beta}_r$ becomes less sensitive to changes in $X'Y$, i.e. ε. This is the bias-variance tradeoff.

[c] Proof in Appendix

Proof.

$$\begin{aligned} &\equiv \min_b(Y'Y - b'X'Y - Y'Xb + b'X'Xb + \lambda b'b) \\ &FOC : -X'Y - X'Y + 2X'Xb + 2\lambda b = 0 \\ &2(X'X + \lambda I)b = 2X'Y \\ &b = (X'X + \lambda I)^{-1}X'Y = \hat{\beta}_r \end{aligned}$$

□

7.1.1 Cross-Validation

The parameter λ is usually chosen by CV:

Overall idea is to minimise expected squared prediction error $E(y - x'\hat{\beta}_r)^2$, where (y, x) is a new observation from the joint distribution of the dependent and explanatory variables. As we do not have such a new observation we approximate it via the leave one out CV method:

1. Drop the i-th observation (y_i, x_i) from the sample
2. Estimate $\hat{\beta}_{r(-i)}$ on the remaining $n-1$ observations
3. Estimate $E(y - x'\hat{\beta}_{-i,r})^2$ via taking the average of the squared prediction errors on the dropped observation (y_i, x_i) via $\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{-i,r})^2$
4. Choose λ that minimises the above estimate.

Example. **Ridge in an Overidentified Model**: $k > n$

Clearly here the classical OLS estimator would not be defined, since $X'X$ is not invertible. Its rank is bounded by n but dimensions are $k \times k$. Thus we need to introduce some bias to make the problem well defined. We note that, assuming (X is of rank n), the column space of X will span $\mathbb{R}^n$. Thus we can find exact vectors b such that $Xb = Y$, where multiplicity due to the fact $k > n$. We select among these by minimising the L2 norm:

$$\min_b ||b||^2 \quad \text{subject to} \quad Xb = Y$$

$$\mathbb{L} = b'b + \lambda'(Y - Xb)$$

$$\text{FOC} : \begin{cases} b : 2b - X'\lambda = 0 \\ \lambda : (Y - Xb) = 0 \end{cases}$$

$$\Rightarrow b = \frac{1}{2}X'\lambda \quad \text{and} \quad Y = \frac{1}{2}XX'\lambda$$

$$\Rightarrow \frac{\lambda}{2} = (XX')^{-1}Y$$

$$\Rightarrow b = X'(XX')^{-1}Y$$

Note:-

Exact Identification $n = k$

Provided X is full rank we then simply have a unique solution to the OLS problem and take:

$$\hat{\beta} = X^{-1}Y$$

7.2 Mallow's C_P Criterion

When we allow biased estimators and only care about expected prediction error we then are no longer bound to correctly specified models. Intuitively if some regressors have non-zero, but small coefficients, we may still want to drop them to reduce the variance of prediction at the expense of introducing some, hopefully small, bias.

Proposition 7.2.1. Consider two models, unrestricted and restricted:

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

$$Y = X_1\beta_1 + \varepsilon$$

where X is $n \times k$ and X_1 is $n \times p$ and X_2 is $n \times q$, where $p + q = k$. Suppose the true model is the unrestricted one and this regression satisfies GM1-3 assumptions, and $\varepsilon | X_1, X_2 \sim N(0, \sigma^2 I_n)$.

Under the Mallow's criterion we prefer the unrestricted model if:

$$C_p = \frac{SSR_r}{\hat{\sigma}^2} - n + 2p > k$$

Intuitively we may still prefer the restricted model if β_2 could only be estimated very imprecisely.

We show the bounds of this intuition as follows:

Long model: $\hat{Y} = X\hat{\beta}$
Short model: $\tilde{Y} = X_1\tilde{\beta}_1$

Our measure of accuracy for any predictor $\check{Y}$ of Y is the expected scaled sum of squared deviations of $\check{Y}$ from the best (infeasible) predictor $X\beta$.

$$J = E\left(\frac{1}{\sigma^2}(\check{Y} - X\beta)'(\check{Y} - X\beta)\right)$$

n.b. the following expectations are <u>all condtional on X</u>

Lemma 7.2.1. For $\check{Y} = \hat{Y}$, we have $J_u = \frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{\sigma^2} = k$

Proof.

$$\begin{aligned} J_u &= E\left[\frac{1}{\sigma^2}(\hat{Y} - X\beta)'(\hat{Y} - X\beta)\right] \\ &= E\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{\sigma^2} \\ &= Etr(\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{\sigma^2}) \\ &= tr(E\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{\sigma^2}) \\ &= tr(\frac{X'X}{\sigma^2}E(\hat{\beta}-\beta)(\hat{\beta}-\beta)') \\ &= tr(\frac{X'X}{\sigma^2}Var(\hat{\beta})) \\ &= tr(\frac{X'X}{\sigma^2}\sigma^2(X'X)^{-1}) \\ &= tr(I_k) = k \end{aligned}$$

□

Lemma 7.2.2. For $\check{Y} = \tilde{Y}$, we have $J_r = p + \frac{1}{\sigma^2}\beta_2'X_2'M_2X_2\beta_2$

Proof.

$$J_r = E\frac{(\begin{pmatrix}\tilde{\beta}_1 \\ 0\end{pmatrix} - \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix})'X'X(\begin{pmatrix}\tilde{\beta}_1 \\ 0\end{pmatrix} - \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix})}{\sigma^2}$$

Omitted Variable Bias:

$$\begin{aligned} E(\tilde{\beta}_1|X) &= E((X_1'X_1)^{-1}X_1'Y|X) = E((X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)|X) \\ &= \beta_1 + \underline{(X_1'X_1)^{-1}X_1'X_2\beta_2} \end{aligned}$$

$\left(\begin{smallmatrix}\tilde{\beta}_1\\0\end{smallmatrix}\right)-\left(\begin{smallmatrix}\beta_1\\\beta_2\end{smallmatrix}\right)=\left(\begin{smallmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\0\end{smallmatrix}\right)-\left(\begin{smallmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\\beta_2\end{smallmatrix}\right)=\left(\begin{smallmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\0\end{smallmatrix}\right)+\left(\begin{smallmatrix}(X_1'X_1)^{-1}X_1'X_2\beta_2\\-\beta_2\end{smallmatrix}\right)$provided some cross product

Thus we can decompose J_r as follows:

$$J_r = E\frac{\left(\begin{smallmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\0\end{smallmatrix}\right)'X'X\left(\begin{smallmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\0\end{smallmatrix}\right)}{\sigma^2}+\frac{\left(\begin{smallmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\-\beta_2\end{smallmatrix}\right)'X'X\left(\begin{smallmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\-\beta_2\end{smallmatrix}\right)}{\sigma^2}(+\text{... cross terms }=0)^*$$

$$J_r = E\frac{(\tilde{\beta}_1-E(\tilde{\beta}_1|X))'X'X(\tilde{\beta}_1-E(\tilde{\beta}_1|X))}{\sigma^2}+E\frac{\left(\begin{smallmatrix}(X_1'X_1)^{-1}X_1'X_2\beta_2\\-\beta_2\end{smallmatrix}\right)'\left(\begin{smallmatrix}X_1'X_1 & X_1'X_2\\X_2'X_1 & X_2'X_2\end{smallmatrix}\right)\left(\begin{smallmatrix}(X_1'X_1)^{-1}X_1'X_2\beta_2\\-\beta_2\end{smallmatrix}\right)}{\sigma^2}$$

Simplifying latter term's numerator:

$$\begin{aligned}&= \beta_2'X_2'X_1(X_1'X_1)^{-1}(X_1'X_1)(X_1'X_1)^{-1}X_1'X_2\beta_2 - 2\beta_2'X_2'X_1(X_1'X_1)^{-1}X_1'X_2\beta_2 + \beta_2'X_2'X_2\beta_2)\\&= (-\beta_2'X_2'X_1(X_1'X_1)^{-1}X_1'X_2\beta_2+\beta_2'X_2'X_2\beta_2)\\&= \beta_2'X_2'(I-P_1)X_2\beta_2\\&= \beta_2'X_2'M_1X_2\beta_2\end{aligned}$$

**Evaluating cross terms is equivalent to evaluating the above except substituting one* $-\beta_2 = 0$. *Clearly then we have full cancellation and thus we do not need to consider these terms.*

Simplifying the first term:

$$\begin{aligned}&\frac{1}{\sigma^2}E[tr((\tilde{\beta}_1-E(\tilde{\beta}_1|X))'X_1'X_1(\tilde{\beta}_1-E(\tilde{\beta}_1|X))|X)]\\&= \frac{1}{\sigma^2}tr(X'XVar(\tilde{\beta}_1|X))\end{aligned}$$

$$\begin{aligned}\text{But: } Var(\tilde{\beta}_1|X) &= Var((X_1'X_1)^{-1}X_1'Y|X) = (X_1'X_1)^{-1}X_1'Var(Y|X)X_1(X_1'X_1)^{-1}\\&= \sigma^2(X_1'X_1)^{-1}\end{aligned}$$

Thus we have:

$$\begin{aligned}J_r &= tr(I_p)+\frac{1}{\sigma^2}\beta_2'X_2'M_1X_2\beta_2\\&= p+\frac{1}{\sigma^2}\beta_2'X_2'M_1X_2\beta_2\end{aligned}$$

□

Solution:-

Therefore:

$$J_r < J_u \quad \text{if and only if} \quad \frac{1}{\sigma^2}\beta_2'X_2'M_1X_2\beta_2 < q$$

This is likely to occur if

- β_2 is small (small bias when omitted)
- X_2 is highly correlated with X_1, i.e. high multicollinearity (lowers value of $X_2'M_1X_2$, since this is the SSR from regressing X_2 on X_1) ($\hat{\varepsilon} = M_1X_2, SSR = \hat{\varepsilon}'\hat{\varepsilon} = X_2'M_1'M_1X_2$)
- σ^2 is large

We can estimate the LHS via considering the restricted SSR

$$SSR_r = (Y - X_1\tilde{\beta}_1)'(Y - X_1\tilde{\beta}_1) = Y'M_1Y$$

$$\begin{aligned} E(SSR_r) = E((Y'M_1Y)|X) &= E(X_1\beta_1 + X_2\beta_2 + \varepsilon)'M_1(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= \beta_2'X_2'M_1X_2\beta_2 + E(\varepsilon'M_1\varepsilon|X) \quad \text{since} M_1X_1 = 0 \\ &= \beta_2'X_2'M_1X_2\beta_2 + trE(M_1\varepsilon'\varepsilon) = \beta_2'X_2'M_1X_2\beta_2 + \sigma^2 tr(M_1) \\ &= \beta_2'X_2'M_1X_2\beta_2 + \sigma^2(n-p) \end{aligned}$$

Therefore as we can use the MOM sample analogue of $E(SSR_r)$ and σ^2, we can estimate the LHS of the inequality above as:

$$\frac{SSR_r}{\hat{\sigma}^2} - (n-p)$$
$$\Rightarrow \hat{J}_r = p + \frac{SSR_r}{\hat{\sigma}^2} - (n-p)$$

Definition 7.2.1

Mallow's C_p Model Selection Criterion:

$$C_p = \frac{SSR_r}{\hat{\sigma}^2} - n + 2p$$

$$\text{where} \quad \hat{\sigma}^2 = \frac{SSR_u}{n-k} \quad \text{is estimated from the long regression}$$

Minimising this across different sub-models of the 'long' model yields a 'short' model most adequate for 'prediction'. We then only prefer the 'long' model to short if $C_p > J_u = k$.

As C_p only estimates J_r, choosing the minimum is not a guarantee of the best model, especially under small sample sizes and model subsets that have similar predictive power (flat minimum).

7.2.1 F-Test Interpretation

Mallows' C_p can be thought of as providing a guidance for the choice of the 'optimal' critical value of the F-test in testing the hypothesis that $\beta_2 = 0$.

$$\begin{aligned} C_p = \frac{SSR_r}{SSR_u/(n-k)} - n + 2p &= \frac{SSR_r - SSR_u}{SSR_u/(n-k)} + 2p - k \\ &= (k-p)(\text{F-stat}) + 2p - k \end{aligned}$$

Hence, $C_P > k$ (and so choose long) if and only if

$$(k-p)(\text{F-stat}) + 2p - 2k > 0$$
$$\text{F-stat} > 2$$

7.2.2 Penalised Least Squares interpretation

Proposition 7.2.2. The OLS estimator in the restricted model chosen by C_p (not considering the long regression) can also be viewed as the result of the following penalised least squares criterion:

$$\min(||Y - X\beta||^2 + \lambda||\beta||_0), \quad \text{where}\, \lambda = 2\hat{\sigma}^2$$

and the 0-norm of a vector is the number of non-zero elements in it. Note this norm is not a convex function of β, which makes minimisation difficult when k is large (as we would need to consider 2^k possible models with different combinations of included regressors and compare the penalised least squares results).

Proof.

$$\underset{M_j \in M_r}{\operatorname{argmin}}\, \underset{b \in M_j}{\operatorname{argmin}}\, C_p = \frac{RSS_r}{\hat{\sigma}^2} - n + 2p,$$

where M_r denotes the set of all possible models under differing restrictions.

$$= \underset{M_j \in M_r}{\operatorname{argmin}}\, \underset{b \in M_j}{\operatorname{argmin}}\, RSS_r - (\hat{\sigma}^2)n + 2(\hat{\sigma}^2)p$$

$$= \underset{M_j \in M_r}{\operatorname{argmin}}\, \underset{b \in M_j}{\operatorname{argmin}}\, RSS_r + 2\hat{\sigma}^2 p$$

$$= \underset{M_j \in M_r}{\operatorname{argmin}}\, \underset{b \in M_j}{\operatorname{argmin}} (Y - Xb)'(Y - Xb) + 2\hat{\sigma}^2 ||b||_0$$

□

7.3 Least Absolute Shrinkage and Selection Operator

Definition 7.3.1

LASSO solves the following problem:

$$\min_b (||Y - Xb||^2 + \lambda ||b||_1), \text{ where } ||\beta||_1 = \sum_{j=1}^{k} |\beta_j| \quad \text{for some } \lambda \geq 0$$

Similarly to $\hat{\beta}_r$, $\hat{\beta}_{LASSO}$ estimates are more stable than OLS. In addition many components of $\hat{\beta}_{LASSO}$ are exactly 0. Hence, we can think of LASSO as not only estimating the parameters but also performing model selection, making the model more parsimonious and thus better interpretable.

Example. Orthonormal design

We proceed with the following special case to build intuition:

$$X'X = I_k$$

$$\Rightarrow ||Y - Xb||^2 = ||(X\hat{\beta}_{OLS} + \hat{\varepsilon}) - Xb||^2$$
$$= (\hat{\beta}_{OLS} - b)'X'X(\hat{\beta}_{OLS} - b) + \hat{\varepsilon}'\hat{\varepsilon}$$

$$\Rightarrow \text{Objective function: } \min_b \left[\sum_{j=1}^{k} (\hat{\beta}_{OLS,j} - b_j)^2 + \lambda \sum_{j=1}^{k} |b_j| \right]$$

(OLS residuals dropped as do not depend on choice of LASSO estimator b)

We can minimise each element of the overall sum separately:

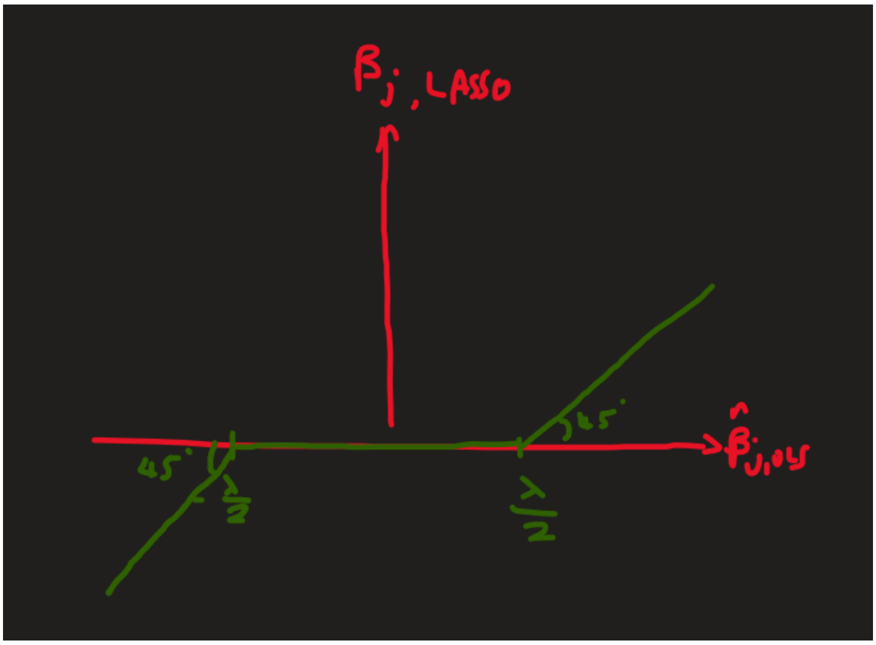
$$\min_{b_j} (b_j - \hat{\beta}_{j,OLS})^2 + \lambda |b_j| \quad \text{for each } j$$

$$\text{But:}\quad (b_j - \hat{\beta}_{j,OLS})^2 + \lambda|b_j| = \begin{cases} (b_j - \hat{\beta}_{j,OLS})^2 + \lambda b_j & \text{for } b_j \geq 0 \\ (b_j - \hat{\beta}_{j,OLS})^2 - \lambda b_j & \text{for } b_j < 0 \end{cases}$$

$$\therefore \hat{b}_{j,LASSO} = \begin{cases} \hat{\beta}_{j,OLS} - \lambda/2 & \text{for } \hat{\beta}_{j,OLS} \geq \lambda/2 \\ \hat{\beta}_{j,OLS} + \lambda/2 & \text{for } \hat{\beta}_{j,OLS} \leq -\lambda/2 \\ 0 & \text{otherwise} \end{cases}$$

7.3.1 Thresholding vs Shrinking

This estimator is an example of a soft thresholding estimator: when $|\hat{\beta}_{j,OLS}|$ is below the threshold $\lambda/2$ we set the estimator $\hat{\beta}_{j,LASSO}$ to zero, and when $|\hat{\beta}_{j,OLS}|$ is above the threshold, set estimator to $sgn\hat{\beta}_{j,OLS}(|\hat{\beta}_{j,OLS}| - \lambda/2)$.



In contrast the ridge estimator in the orthonormal design case has form:

$$\hat{\beta}_{j,ridge} = (I + \lambda I)^{-1} X'Y = \frac{\hat{\beta}_{j,OLS}}{1 + \lambda}$$

Thus it does not set any $\hat{\beta}_{j,r}$ to zero. It just shrinks $\hat{\beta}_{OLS}$.

Note:-

As in practice we do not work inside the orthonormal design case, we need to standardise the regressors to make shrinkage apply fairly. Thus we demean all regressors to avoid shrinking the constant. And we divide all variables by their standard deviation to avoid shrinking variables with larger variance more, as this is unit variant.

7.4 Appendix

7.4.1 Proof Condition Number is Ratio of Eigenvalues

Theorem 7.4.1. For any positive definite symmetric matrix A, the condition number $C ==$ $||A||\,||A^{-1}||$ is equal to the ratio of the largest eigenvalue to the smallest eigenvalue of A.

$$||A||\,||A^{-1}|| = \frac{\lambda_{max}}{\lambda_{min}}$$

Recall that for any symmetric matrix A, we have that A is diagonalisable (by spectral theorem). This means there exists an orthonormal basis of eigenvectors $(\vec{q_1}, ..., \vec{q_n}) = Q$ that spans $\mathbb{R}^n$, with corresponding eigenvalues $(\lambda_1, ..., \lambda_n)$, where wlog $|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_n|$.
Thus we can write any vector $\vec{x} \in \mathbb{R}^n$ as a linear combination of these eigenvectors: $\vec{x} = Q\vec{c}$
Thus we can rewrite:

$$||A|| = \max_{x \neq 0} \frac{||Ax||}{||x||} = \max_{x \neq 0} \frac{||AQc||}{||Qc||}$$

$$= \max_{x \neq 0} \frac{||c_1\lambda_1 q_1 + ...c_n\lambda_n q_n||}{||c_1 q_1 + ...c_n q_n||}$$

Consider:

$$\frac{||Ax||^2}{||x||^2} = \frac{(c_1\lambda_1 q_1 + ...c_n\lambda_n q_n)'(c_1\lambda_1 q_1 + ...c_n\lambda_n q_n)}{(c_1 q_1 + ...c_n q_n)'(c_1 q_1 + ...c_n q_n)}$$

As eigenvectors orthonormal $q_i'q_j = 0$ for $i \neq j$ and $q_i'q_i = 1$:

$$\therefore = \frac{c_1^2\lambda_1^2 + ...c_n^2\lambda_n^2}{c_1^2 + ...c_n^2} \leq \frac{c_1^2\lambda_1^2 + ...c_n^2\lambda_1^2}{c_1^2 + ...c_n^2} = |\lambda_1|^2$$

$$\Rightarrow \frac{||Ax||}{||x||} \leq |\lambda_1|$$

But we can achieve this bound at $x = q_1$.

Lemma 7.4.1. Thus:

$$||A|| = \max_{x \neq 0} \frac{||Ax||}{||x||} = |\lambda_1|$$

For a *positive definite* symmetric matrix we know that all $\lambda_i > 0$ Thus

$$||A|| = \max(\lambda_1, ..., \lambda_n)$$

Lemma 7.4.2. A matrix A has eigenvalue λ if and only if A^{-1} has eigenvalue λ^{-1}.

Proof. Let v be an eigenvector of A with eigenvalue λ. Then:

$$Av = \lambda v \Rightarrow A^{-1}Av = A^{-1}\lambda v \Rightarrow A^{-1}v = \frac{1}{\lambda}v$$

□

Lemma 7.4.3. A positive definite $\Rightarrow A^{-1}$ (exists and is) positive definite.

> **Proof.** A positive definite implies A invertible, as only solution to Ax=0 is x=0 (thus full rank by rank nullity theorem).
> Consider for any $x \in \mathbb{R}^n$
>
> $$x'A^{-1}x$$
>
> Define $x = Ay$ as for any x there must exist y where $y = A^{-1}x$. Thus:
>
> $$x'A^{-1}x = (Ay)'A^{-1}(Ay) = y'A'y$$
>
> As $y'A'y$ is a scalar, it is equal to its transpose. Thus:
>
> $$y'A'y = (y'A'y)' = y'Ay > 0 \quad \text{as A positive definite}$$
>
> Thus A^{-1} is positive definite. □

Thus, as A^{-1} positive definite:

$$||A^{-1}|| = \max(\lambda_1^{-1}, ..., \lambda_n^{-1})$$

As $\lambda_i > 0$ for all i,

$$||A^{-1}|| = \frac{1}{\lambda_{min}}$$

Thus:

$$C = ||A||\,||A^{-1}|| = \frac{\lambda_{max}}{\lambda_{min}}$$

Note:-

We can apply this to the regression case by considering $A = X'X$ so long as $X'X$ is positive definite. We prove that $X'X$ is positive definite if X is full rank and $n > k$: This is because for any vector $\alpha \neq 0$ we have:

$$\alpha'X'X\alpha = (X\alpha)'X\alpha = ||X\alpha||^2 \geq 0$$

To show inequality is strict, we must show that the null space of $X'X$ is trivial. We are given $null(X)$ is empty as X is full rank by rank nullity theorem, thus we simply must show $null(X'X) = null(X)$. (only true for real matrices)

$\underline{null(X) \subseteq null(X'X)}$:

$$\text{Let vector}\, v \in null(X), \text{then}\, Xv = 0$$
$$\Rightarrow X'Xv = X'0 = 0$$
$$\Rightarrow v \in null(X'X)$$

$\underline{null(X'X) \subseteq null(X)}$:

$$\text{Let vector}\, v \in null(X'X), \text{then}\, X'Xv = 0$$
$$\Rightarrow v'X'Xv = 0$$
$$\Rightarrow ||Xv||^2 = 0$$
$$\Rightarrow Xv = 0 \quad \text{provided X is real valued}$$
$$\Rightarrow v \in null(X)$$

Thus $null(X'X) = null(X) = 0$ and $X'X$ is positive definite.

7.4.2 Proof Condition Number Bounds Solution Sensitivity

Theorem 7.4.2. For a set of simulateneous equations $Ax = b$ the condition number of A sets a bound of the sensitivity of the relative (i.e. invariant to units) solution error (in x) induced by errors in the problem vector (b). $\frac{||\Delta x||}{||x||} \leq C \frac{||\Delta b||}{||b||}$

Proof. Recall the submultiplicity property of the norm:

$$||Ax|| \leq ||A|| \, ||x||, ||AB|| \leq ||A|| \, ||B||$$

Suppose there exists a problem error in b such that $\tilde{b} = b + \Delta b$.
Then the solution to the new problem is $\tilde{x} = A^{-1}\tilde{b}$. We seek a bound on the unit invariant relative error in the solution:

Consider:

$$A(x + \Delta x) = b + \Delta b$$

$$Ax = b \Rightarrow A\Delta x = \Delta b \Rightarrow \Delta x = A^{-1}\Delta b$$

$$\Rightarrow ||\Delta x|| \leq ||A^{-1}|| \, ||\Delta b||$$

But:

$$Ax = b$$

$$\Rightarrow ||b|| \leq ||A|| \, ||x||$$

Now multiplying inequalities:

$$||\Delta x|| \, ||b|| \leq ||A^{-1}|| \, ||\Delta b|| \, ||A|| \, ||x||$$

$$\Rightarrow \frac{||\Delta x||}{||x||} \leq ||A^{-1}|| \, ||A|| \frac{||\Delta b||}{||b||}$$

□

7.4.3 Proof Ridge Regression Reduces Condition Number

Theorem 7.4.3. Let A be a positive definite symmetric matrix with condition number:

$$C = ||A|| \, ||A^{-1}|| = \frac{\lambda_{max}}{\lambda_{min}}$$

Then the condition number of $(A + \mu I)$ is:

$$\frac{\lambda_{max} + \mu}{\lambda_{min} + \mu} < \frac{\lambda_{max}}{\lambda_{min}}$$

Proof. Let Q be an orthonormal basis of eigenvectors of A, with corresponding eigenvalues $\lambda_1, ..., \lambda_n$. Then:

$$(A + \mu I)Q = AQ + \mu IQ = Q\Lambda + Q\mu I = Q(\Lambda + \mu I)$$

Thus we can diagonalise (A+μI) with eigenvalues $\lambda_1 + \mu, ..., \lambda_n + \mu$ as follows.

$$(A + \mu I) = Q(\tilde{\Lambda} + \mu I)Q'$$

Where Q is the same orthonormal basis, and all eigenvalues are simply augmented by μ.

Thus:

$$\frac{\tilde{\lambda}_{max}}{\tilde{\lambda}_{min}} = \frac{\lambda_{max} + \mu}{\lambda_{min} + \mu}$$

□

Note:-

$$Q' = Q^{-1}$$

$$\because Q'Q = QQ' = I$$