MODEL SELECTION FRAMEWORK

Q1) Prove Theorem 1

For each model, least square estimators for the regression coefficient are given as follows:

$$\hat{\gamma}_1(F) = \hat{\gamma}_1(S) - A\hat{\gamma}_2(F)$$
$$\hat{\gamma}_2(F) = (X_{2,1\perp}^T X_{2,1\perp})^{-1} X_{2,1\perp}^T y$$
$$\hat{\gamma}_1(S) = (X_1^T X_1)^{-1} X_1^T y$$

where, $A = (X_1^T X_1)^{-1} X_1^T X_2$ and $X_{2,1\perp} = X_2 - \pi_{X_1} X_2 = X_2 - X_1 A$

(proof)

(Full Model): $y = X_1\gamma_1 + X_2\gamma_2 + \epsilon$ $\quad \epsilon = (\epsilon_1, \cdots \epsilon_p)^T, \epsilon_i \sim (0, \sigma^2)$
(Sub/Reduced Model): $y = X_1\gamma_1 + \epsilon$

$$\hat{\gamma}_1(S) = (X_1^T X_1)^{-1} X_1^T y$$
$$\hat{\gamma}_2(F) = (X_{2,1\perp}^T X_{2,1\perp})^{-1} X_{2,1\perp}^T y$$

$$\pi_X y = \pi_{X_1} y + \pi_{X_{2,1\perp}} y = X_1(X_1^T X_1)^{-1} X_1^T y + X_{2,1\perp}(X_{2,1\perp}^T X_{2,1\perp})^{-1} X_{2,1\perp}^T y$$
$$= X_1\hat{\gamma}_1 + X_2\hat{\gamma}_2 = X_1\hat{\gamma}_1 + \pi_{X_1} X_2\hat{\gamma}_2 + X_{2,1\perp}\hat{\gamma}_2$$

$$\therefore X_1\hat{\gamma}_1(S) = X_1\hat{\gamma}_1 + \pi_{X_1} X_2\hat{\gamma}_2$$
$$= X_1\hat{\gamma}_1 + X_1(X_1^T X_1)^{-1} X_1^T X_2\hat{\gamma}_2(F)$$

$$\therefore \hat{\gamma}_1(F) = \hat{\gamma}_1(S) - A\hat{\gamma}_2(F)$$

Q2) Prove the following:

$$MSE[\hat{\gamma}_1(F), \gamma_1] = Var(\hat{\gamma}_1(S)) + A\,Var(\hat{\gamma}_2(F))\,A^T$$
$$MSE[\hat{\gamma}_1(S), \gamma_1] = Var(\hat{\gamma}_1(S)) + A\gamma_2\gamma_2^T A^T$$

Here, the MSE is matrix valued and not scalar valued.

$$MSE[\hat{\gamma}_1(F), \gamma_1] = E[(\hat{\gamma}_1(F) - \gamma_1)(\hat{\gamma}_1(F) - \gamma_1)^T]$$
$$= Var(\hat{\gamma}_1(F)) \quad (\because \hat{\gamma}_1(F) \text{ is an unbiased estimator})$$
$$= Var(\hat{\gamma}_1(S) - A\hat{\gamma}_2(F))$$
$$= Var(\hat{\gamma}_1(S)) + Var(A\hat{\gamma}_2(F)) - 2Cov(\hat{\gamma}_1(S), A\hat{\gamma}_2(F))$$

Now, $\hat{\gamma}_1(S) \in C(X_1)$ and $\hat{\gamma}_2(F) \in C(X_{2,1\perp})$ in which $C(X_1) \perp\!\!\!\perp C(X_{2,1\perp})$
$\therefore Cov(\hat{\gamma}_1(S), \hat{\gamma}_2(F)) = 0$

$$= Var(\hat{\gamma}_1(S)) + A\,Var(\hat{\gamma}_2(F))\,A^T \quad \square$$

$$MSE[\hat{\gamma}_1(S), \gamma_1] = E[(\hat{\gamma}_1(S) - \gamma_1)(\gamma_1(S) - \gamma_1)^T]$$

$$= E[(\hat{\gamma}_1(S) - E[\hat{\gamma}_1(S)] + E[\hat{\gamma}_1(S)] - \gamma_1)(\hat{\gamma}_1(S) - E[\hat{\gamma}_1(S)] + E[\hat{\gamma}_1(S)] - \gamma_1)^T]$$
$$= E[(\hat{\gamma}_1(S) - E[\hat{\gamma}_1(S)])(\hat{\gamma}_1(S) - E[\hat{\gamma}_1(S)])^T] + E[(E[\hat{\gamma}_1(S)] - \gamma_1)(E[\hat{\gamma}_1(S)] - \gamma_1)^T]$$
$$= Var(\hat{\gamma}_1(S)) + E[(\gamma_1 + A\gamma_2 - \gamma_1)(\gamma_1 + A\gamma_2 - \gamma_1)^T]$$
$$= Var(\hat{\gamma}_1(S)) + A\gamma_2\gamma_2^T A^T$$

Approach #2.

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T]$$
$$= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^T]$$
$$= E[(\hat{\theta} - E[\hat{\theta}])(\hat{\theta} - E[\hat{\theta}])^T] + E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] + E[(E[\hat{\theta}] - \theta)(\hat{\theta} - E[\hat{\theta}])^T] + E[(E[\hat{\theta}] - \theta)(E[\hat{\theta}] - \theta)^T]$$
$$= Var(\hat{\theta}) + \underbrace{E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)]}_{=0 \text{ if } \hat{\theta} \text{ is unbiased.}} + E[(E[\hat{\theta}] - \theta)(E[\hat{\theta}] - \theta)^T]$$

Q3) Prove Theorem 2

Consider two predictions of y_z at $z = (z_1^T\ z_2^T)^T$ which is decomposed in the same way as X.

$$\hat{y}_z(F) = z_1^T\hat{\gamma}_1(F) + z_2^T\hat{\gamma}_2(F)$$
$$\hat{y}_z(S) = z_1^T\hat{\gamma}_1(S)$$

Prove the following:

$$MSPE(\hat{y}_z(F), y_z) = \sigma^2 + z_1^T Var(\hat{\gamma}_1(S))\,z_1 + (z_2 - A^T z_1)^T Var(\hat{\gamma}_2(F))(z_2 - A^T z_1)$$

$$MSPE(\hat{y}_z(S), y_z) = \sigma^2 + z_1^T Var(\hat{\gamma}_1(S))\,z_1 + (z_2 - A^T z_1)^T\gamma_2\gamma_2^T(z_2 - A^T z_1)$$

(proof)

$$MSPE(\hat{y}_z(F), y_z) = E[(\hat{y}_z - y_z)^2] = E[(\hat{y}_z - E[\hat{y}_z] + E[y_z] - y_z)^2] \quad * E[\hat{y}_z] = E[y_z]$$
$$= E[(\hat{y}_z - E[\hat{y}_z])^2] + E[(y_z - E[y_z])^2]$$
$$= Var(\hat{y}_z) + Var(\hat{y}_z)$$
$$= Var(z_1^T\hat{\gamma}_1(F) + z_2^T\hat{\gamma}_2(F)) + Var(\hat{y}_z)$$
$$= Var(z_1^T\hat{\gamma}_1(S) - z_1^T A\gamma_2(F) + z_2^T\hat{\gamma}_2(F)) + Var(\hat{y}_z)$$
$$= z_1^T Var(\hat{\gamma}_1(S))\,z_1 + (z_2^T - z_1^T A)\,Var(\hat{\gamma}_2(F))(z_2 - A^T z_1)$$
$$= z_1^T Var(\hat{\gamma}_1(S))\,z_1 + (z_2 - A^T z_1)^T Var(\hat{\gamma}_2(F))(z_2 - A^T z_1)$$

$$MSPE(\hat{y}_z(S), y_z) = E[(\hat{y}_z(S) - y_z)^2]$$
$$= E[(\hat{y}_z(S) - E[\hat{y}_z(S)] + E[\hat{y}_z(S)] - E[y_z] + E[y_z] - y_z)^2]$$
$$* \rightarrow = Var(\hat{y}_z(S)) + E[(E[\hat{y}_z(S)] - E[\hat{y}_z])^2] + Var(y_z)$$
$$= z_1^T Var(\hat{\gamma}_1(S))\,z_1 + E[(z_1^T(\gamma_1 + A\gamma_2) - z_1^T\gamma_1 - z_2^T\gamma_2)^2] + \sigma^2$$
$$= \sigma^2 + z_1^T Var(\hat{\gamma}_1(S))\,z_1 + E[(z_1^T A - z_2^T)\gamma_2\gamma_2^T(A^T z_1 - z_2)]$$
$$= \sigma^2 + z_1^T Var(\hat{\gamma}_1(S))\,z_1 + (z_2 - A^T z_1)^T\gamma_2\gamma_2^T(z_2 - A^T z_1)$$

* cross-product goes to 0

MODEL SELECTION CRITERION

Q4)

$$R^2 = \frac{SSR(F)}{SST(F)} = \frac{SSR(F)}{SST}, \quad R^2(S) = \frac{SSR(S)}{SST}$$

Show that $R^2(S) \le R^2(F)$. Hint: $R(\gamma_2|\gamma_1) = \|\Pi_X y\|^2 - \|\Pi_{X_1} y\|^2$

$$SSR(F) = y^T(\Pi_X - \Pi_1)y$$
$$SSR(S) = y^T(\Pi_{X_1} - \Pi_1)y$$
$$\therefore SSR(F) - SSR(S) = y^T(\Pi_X - \Pi_{X_1})y = y^T(\Pi_{X_2|1})y \ge 0$$

Q5) Show that under H_0 (i.e., $H_0: \beta_1 = \cdots = \beta_p = 0$),

$$R^2(S) = \frac{SSR(S)}{SST} \sim Beta\left(\frac{q}{2}, \frac{n-q-1}{2}\right)$$

$$R^2(S) = \frac{SSR(S)}{SSR(S)+SSE(S)}$$

under $\beta_1 = \cdots = \beta_p = 0$, $y = \beta_0 \mathbb{1} + \varepsilon$

Q6) what is the expectation of $R^2(S)$?

$$E[R^2(S)] = \frac{q}{q+(n-q-1)} = \frac{q}{n-1}$$

Q7) What is the "adjusted R^2 (R^2_a)" for the sub-model?

While the regular R^2 measures the proportion of total variance explained by your model, it has a fundamental mathematical quirk: it mathematically cannot decrease when you add more predictors, even if those predictors are completely unrelated to the response variable.

The adjusted R^2 (R^2_{adj}) corrects for this by penalizing the addition of extraneous predictors. Instead of using the raw sum of squares, it uses "Mean Squares", which divides the Sums of Squares by their respective degrees of freedom to provide unbiased estimators of the population variances.

For a sub-model S with n observations and q predictors (plus one intercept), the formula is:

$$R^2_{adj}(S) = 1 - \frac{MSE(S)}{MST} = 1 - \frac{SSE(S)/(n-q-1)}{SST/(n-1)}$$

Derivation from Regular $R^2(S)$:

$$R^2(S) = 1 - \frac{SSE(S)}{SSR(S)} \Leftrightarrow \frac{SSE(S)}{SSR(S)} = 1 - R^2(S)$$
$$\Leftrightarrow \frac{SSE(S)/(n-q-1)}{SSR(S)/(n-1)} = (1-R^2(S))\frac{n-1}{n-q-1}$$
$$\Leftrightarrow 1 - \frac{SSE(S)/(n-q-1)}{SSR(S)/(n-1)} = 1 - (1-R^2(S))\frac{n-1}{n-q-1}$$

Q8) What is another representation of R^2_a (adjusted R^2)?

$$R^2_{adj} = 1 - \frac{SSE(S)/(n-q-1)}{SST/(n-1)}$$

* note: $R^2_{adj} < 0$ is possible.

Q9) what is the mathematical representation of:

1. In-Sample MSE
2. Out-of-Sample MSE

in-sample MSE: $MSE(s) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i(s))^2 = \frac{1}{n}SSE(s)$

out-of-sample MSE: $MSE_o(s) = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{y}_i(s))^2$, $y^* = X\beta + \epsilon^*$, $\epsilon_1^*, \cdots, \epsilon_n^* \sim N(0,1)$

Q10) what are the two approximations of out-of-sample MSE?

1. K-fold cross validation
   - split the data in K groups
   - Test the ith group with the model fitted by the other K-1 groups.
   - Repeat it K times
   - out-of-sample MSE ($MSE_o(s)$) is approximated by:

$$CV(s) = \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in G_j}(y_i - \hat{y}_{i,(j)}(s))^2$$

G_j : j^{th} test group in the sample

$\hat{y}_{i,(j)}(s)$: predicted value of y_i on the jth step.
(⇔ predicted value of y_i by fitting the model with samples that exclude the jth group)

$$\hat{y}_{i,(j)}(s) = z_i^T\hat{\beta}_{(j)}(s)$$

2. Leave-one-out cross-validation (⇔ n-fold cross validation)

LOOCV score: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,(i)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\epsilon}_i(s)}{1-h_{ii}}\right)^2$, h_{ii} : (i,i) element of H

H: hat matrix.

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i(s)}{1-h_{ii}}\right)^2$$

Q11) Prove Theorem 6

Let the sub-model have q number of predictors ($1 \le q \le p$).

Derive the following:

$$E[MSE_o(s)] = \frac{1}{n}\sum_{i=1}^{n}[\sigma^2 + Bias^2[\hat{y}_i(s)] + Var(\hat{y}_i(s))$$

$$E[MSE(s)] = E[MSE_o(s)] - \frac{2}{n}\sigma^2 q$$

$$E[MSE_o(s)] = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_z - \hat{y}_z(s))^2\right] = \frac{1}{n}\sum_{i=1}^{n}E[(y_z - \hat{y}_z(s))^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}E\left[(y_i - E[y_i] + E[y_i] - E[\hat{y}_i(s)] + E[\hat{y}_i(s)] - \hat{y}_i(s))^2\right]$$

$$= E[MSE_o(s)] = \frac{1}{n}\sum_{i=1}^{n}Var(y_i) + E[(E[y_i] - E[\hat{y}_i])^2] + Var(\hat{y}_i(s))$$

cross product terms = 0

$$= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 + Bias(\hat{y}_z(s))^2 + Var(\hat{y}_i(s)) \quad \square$$

$$E[MSE(s)] = E[MSE_o(s)] - \frac{2}{n}\sigma^2 q$$

$$E[MSE(s)] = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i(s))^2\right]$$

$$= \frac{1}{n}\sum_{i}^{n}E[(y_i - E[y_i] + E[y_i] - E[\hat{y}_i(s)] + E[\hat{y}_i(s)] - \hat{y}_i(s))^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}E[(y_i - E[y_i])^2] + E[E[y_i] - E[\hat{y}_i(s)])^2] + E[(E[\hat{y}_i(s)] - \hat{y}_i(s))^2]$$
$$+ 2\frac{1}{n}\sum_{i=1}^{n}E[(y_i - E[y_i])(E[\hat{y}_i(s)] - \hat{y}_i(s))]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 + bias(\hat{y}_i(s))^2 + Var(\hat{y}_i(s)) - \frac{2}{n}\sum_{i=1}^{n}Cov(y_i, \hat{y}_i(s))$$

$$= E[MSE_o(s)] - \frac{2}{n}trace(Cov(y, \hat{y}(s)))$$
$$= E[MSE_o(s)] - \frac{2}{n}trace(Var(y)\cdot X_1(X_1^TX_1)^{-1}X_1^T)$$
$$= E[MSE_o(s)] - \frac{2}{n}trace(\sigma^2 X_1(X_1^TX_1)^{-1}X_1^T)$$
$$= E[MSE_o(s)] - \frac{2\sigma^2}{n}trace(I_q)$$
$$= E[MSE_o(s)] - \frac{2\sigma^2 q}{n} \quad \square$$

Q12) what is the mathematical representation of "Mallow's C_p"?

Mallow's C_p statistic exploits this relationship in order to approximate the out-of-sample MSE

$$C_p(S) = MSE(S) + \frac{2}{n}\hat{\sigma}^2 q$$

* $MSE(S) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i(S))^2$
* $\hat{\sigma}^2$ is estimated only once with the full model.

Q13) What is the mathematical representation of the "Kullback-Leibler divergence" and explain its meaning.

KL-divergence: $E_{\theta_0}[-\log\{f(x;\theta)/f(x;\theta_0)\}]$

The KL-divergence measures the difference between two probability distributions: $f(x;\theta)$ and $f(x;\theta_0)$.

$$E_{\theta_0}[-\log\{f(x;\theta)/f(x;\theta_0)\}]$$
$$= E_{\theta_0}[-\log f(x;\theta)] + E_{\theta_0}[\log f(x;\theta_0)]$$
$$= E_{\theta_0}[-\log f(x;\theta)] + C$$

Since we only care about comparing different θ values (different models), Statisticians often drop the constant C and casually refer to the remainding operational part as the "KL-divergence" (or relative KL divergence for model selection processes.

2. Why multiply by 2?

· In statistics, Deviance is defined as $-2\log\mathcal{L}(\theta)$ (where $\mathcal{L}$ is the likelihood)
· Multiplying the log-likelihood by 2 ensures that the difference between two models follow's Wilk's theorem, meaning the likelihood ratio test statistic will have an asymptotic Chi-Square (χ^2) distribution.

By defining $K(\theta)$ with the -2 multiplier, your notes are framing the expected cross-entropy on the "deviance-scale".

- Gemini 3.1 pro

Q14) Find $\hat{K}(\theta)$: estimate of the KL-divergence.

estimation of the KL-divergence:

$$\hat{K}(\theta) = -\frac{2}{n}\sum_{i=1}^{n}\log f_\theta(y_i) = -\frac{2}{n}\log\prod_{i=1}^{n} f_\theta(y_i)$$

$$= -\frac{2}{n}\log f_\theta(y), \quad y = [y_1 \cdots y_n]^T \sim N(X\beta, \sigma^2 I_n)$$

$$= -\frac{2}{n}\log\left\{\det(2\pi\Sigma)^{-1/2}\exp\left\{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right\}\right\}$$
$$= -\frac{2}{n}\log\left\{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right\}\right\}$$

$$= \log(2\pi\sigma^2) + \frac{1}{n\sigma^2}\|y - X\beta\|^2$$

Q16) Using the profile method, find $\hat{\theta}(S) = \underset{\theta\in\Omega_S}{\operatorname{argmin}}\ \hat{K}(\theta)$

Ω_S: parameter space of the sub-model/ reduced model.

① Fix $\sigma^2 > 0$

$$\hat{\gamma}_i(\sigma^2) = \underset{\gamma_i}{\operatorname{argmin}}\ \hat{K}(\theta) = \underset{\gamma_i}{\operatorname{argmin}}\ \|y - X_i\gamma_i\|^2 = (X_i^TX_i)^{-1}X_i^Ty = \hat{\gamma}_i(S)$$

② $\hat{\sigma}^2$: $\underset{\sigma^2}{\operatorname{argmin}}\ \hat{K}(\hat{\gamma}_i(\sigma^2)^T, 0^T, \sigma^2)$

$$= \underset{\sigma^2>0}{\operatorname{argmin}}\left\{\log(2\pi\sigma^2) + \frac{1}{n\sigma^2}\|y - X_i\hat{\gamma}_i(S)\|^2\right\}$$

$$= \underset{\sigma^2>0}{\operatorname{argmin}}\left\{\log(2\pi\sigma^2) + \frac{1}{n\sigma^2}SSE(S)\right\}$$

미분해서 0이 되는 지점 찾기 $\frac{\partial L}{\partial\sigma^2} = \frac{1}{\sigma^2} - \frac{1}{n(\sigma^2)^2}SSE(S) = 0 \Leftrightarrow \sigma^2 = \frac{1}{n}SSE(S)$

$$= \frac{1}{n}SSE(S)$$

$\therefore \hat{\theta}(S) = (\hat{\gamma}_i(S), 0^T, \frac{1}{n}SSE)$

Q16) Show that $\forall\gamma_i \in \mathbb{R}^q$, $\sigma^2 \in \mathbb{R}$

$$\hat{K}(\gamma_i^T, 0^T, \sigma^2) \geq \hat{K}(\hat{\gamma}_i(S), 0^T, \frac{1}{n}SSE(S))$$

(proof)

$$\hat{K}(\gamma_i(\sigma^2)^T, 0^T, \sigma^2) \geq \hat{K}(\hat{\gamma}_i(S), 0^T, \sigma^2)$$
$$\geq \hat{K}(\hat{\gamma}_i(S), 0^T, \hat{\sigma}^2)$$
$$= \hat{K}(\hat{\gamma}_i(S), 0^T, \frac{1}{n}SSE(S))$$

Q17) What is the mathematical representation of:

1. Akaike Information Criterion
2. Bayesian Information Criterion

$$AIC(S) = -\frac{2}{n}\sum_{i=1}^{n}\log f(y_i;\hat{\theta}(S)) + \frac{2|S|}{n},$$

$|S|$: size of the submodel; Number of free parameters in θ_i

In the MLRM, $AIC(S) = \log SSE(S) + \frac{2q}{n}$, q: number of predictors in the model.

$$AIC(S) = \hat{K}(\hat{\theta}(S)) + \frac{2|S|}{n} = \log(2\pi\hat{\sigma}^2) + \frac{1}{n\hat{\sigma}^2}\|y - X_i\hat{\gamma}_i\|^2 + \frac{2q}{n}$$

$$= \log(2\pi\frac{1}{n}SSE(S)) + \frac{1}{n\cdot\frac{1}{n}SSE(S)}\|y - X_i\hat{\gamma}_i\|^2 + \frac{2q}{n}$$

$$= \log(\frac{2\pi}{n}) + \log(SSE(S)) + 1 + \frac{2q}{n}$$

$$\Rightarrow \log(SSE(S)) + \frac{2q}{n}$$ (∵ drop factors that are uncorrelated to the model)

$\therefore AIC(S) = \log\{SSE(S)\} + \frac{2q}{n}$

$$BIC(S) = -\frac{2}{n}\sum_{i=1}^{n} \log f(y_i|\hat{\theta}(S)) + \frac{|S| \cdot \log n}{n}$$

- In the MLRM, $BIC(S) = \log SSE(S) + \frac{q \log n}{n}$

(proof)

$$BIC(S) = \underbrace{-\frac{2}{n}\sum_{i=1}^{n} \log f(y_i|\hat{\theta}(S))}_{*} + \frac{|S| \cdot \log n}{n}$$

$* : \hat{k}(\theta(S)) = \log(2\pi\hat{\sigma}^2) + \frac{1}{n\hat{\sigma}^2}\|y - x_i\hat{\gamma}_i\|^2$

$= \log(\frac{2\pi}{n}) + \log(SSE(S)) + 1$

$\therefore BIC(S) = \log(\frac{2\pi}{n}) + \log(SSE(S)) + 1 + \frac{q \cdot \log n}{n}$

$\Rightarrow \log(SSE(S)) + \frac{q \cdot \log n}{n}$ (after dropping constants)

$\therefore BIC(S) = \log(SSE(S)) + \frac{q \cdot \log n}{n}$

* Mallow's C_p, AIC and BIC are estimates of the KL-divergence. Higher the value of these statistics, stronger is the implication that the models fitted are inadequate.

Q(8) Discuss the model complexity and the goodness of fit in:

1. Mallow's C_p
2. AIC(S)
3. BIC(S)

	Goodness of fit	Model Complexity	
$\frac{nC_p(S)}{\hat{\sigma}^2}$	$SSE(S)/\hat{\sigma}^2$	q	Future Prediction
$AIC(S)$	$\log SSE(S)$	$2q/n$	Parameter Estimation
$BIC(S)$	$\log SSE(S)$	$q\log n/n$	Parameter Estimation

* typically $\log n > 2$. Hence BIC is penalizing more for the increase in model complexity than AIC.

STEPWISE VARIABLE SELECTION.

Q19)

- Describe the "forward selection method."
- When does the procedure stop?
- What kind of statistic is $F_{K,j}$ and what test is used?
- Show that $F_{K,j}$ is an increasing function of

$$R^2(\beta_0, \beta_\ell : \ell \in S_{K-1}, \beta_j) = \frac{SSR(\beta_0, \beta_\ell : \ell \in S_{K-1}, \beta_j)}{SST}$$

The forward selection method:

→ (Step 0) : Set the index subset $S_0 = \{\emptyset\}$

(Step K) : Find an index $j = j_K$ that maximizes the statistic:

$$F_{Kj} = \frac{R(\beta_j \mid \beta_0, \beta_\ell : \ell \in S_{K-1})}{SSE(\beta_0, \beta_\ell : \ell \in S_{K-1}, \beta_j)/(n-K-1)} \quad (j \notin S_{K-1})$$

For a predetermined $F_{in} > 0$, Stop the procedure at this step and the subset $S^* = S_{K-1}$ is selected as our model if :

$$F_{K,jK} < F_{in}$$

Otherwise, update the index subset S_K and proceed to the next step:

$$S_K = S_{K-1} \cup \{j_K\}$$

(a) On the step $K \in \mathbb{Z}^+$, the statistic $F_{K,j}$ is the F-test statistic for testing $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ under the following model:

$$y_i = \left\{\beta_0 + \sum_{\ell \in S_{K-1}} \beta_\ell x_{i\ell}\right\} + \beta_j x_{ij} + \varepsilon_i \quad (i = 1, 2, \cdots, n)$$

(b) The statistic $F_{K,j}$ is an increasing function of

$$R^2(\beta_0, \beta_\ell : \ell \in S_{K-1}, \beta_j) = \frac{SSR(\beta_0, \beta_\ell : \ell \in S_{K-1}, \beta_j)}{SST}$$

which is the coefficient of determination for the model in (a)

$$\therefore F_{K,j} = \frac{R(\beta_j \mid \beta_0, \beta_\ell : \ell \in S_{K-1})}{SSE(\beta_0, \beta_\ell : \ell \in S_{K-1}, \beta_j)/(n-K-1)}$$

$(n-K-1)$로 나누어주는 이유 : $F_{K,j} \sim F(1, n-K-1)$

Let $R(\beta_j \mid \beta_0, \beta_\ell : \ell \in S_{K-1}) = \underbrace{SSE(R)}_{\text{SSE of reduced model}} - \underbrace{SSE(F)}_{\text{SSE of full model}} = 1 - SSR(R) - (1 - SSR(F))$

$$SSE(\beta_0, \beta_\ell : \ell \in S_{K-1}, \beta_j) = SSE(F) = SST - SSR(F)$$

$$F_{K,j} = \frac{SSR(F) - SSR(R)}{SST - SSR(F)} \cdot (n-p-1) = \frac{\{SSR(F) - SSR(R)\}/SST}{\{SST - SSR(F)\}/SST} \cdot (n-p-1)$$

$$= \frac{R^2(F) - R^2(R)}{1 - R^2(F)} \cdot (n-p-1)$$

$\therefore F_{K,j}$ is an increasing function of $R^2(F)$

Q20) Describe the backward selection method.

(Step 0) : Set the index subset $S_0 = \{1, 2, \cdots, k\}$

(Step K) : Find an index that minimizes the statistic:

$$F_{K,j} = \frac{R(\beta_j \mid \beta_0, \beta_\ell : \ell \in S_{K-1} \setminus \{j\})}{SSE(\beta_0, \beta_\ell : \ell \in S_{K-1})/(n-K)} \quad (j \in S_{K-1})$$

For a predetermined $F_{out} > 0$, stop the procedure at this step and the subset $S^* = S_{K-1}$ is selected as our final model if

$$F_{K,jK} \geq F_{out}$$

Otherwise, update the index subset S_K and proceed to the next step.

$$S_K = S_{K-1} \setminus \{j_K\}$$

* $F_{K,j} = \frac{SSE(R) - SSE(F)}{SSE(F)/(n-K)} \sim F(1, n-K)$

Q21) Describe the stepwise variable selection method.
Which should be stricter: F_{in} or F_{out}? Explain why.

(Step 0) : Set the index subset $S_0 = \{\emptyset\}$

(Step K) : Find an index $j = j_K$ that maximizes the statistic:

$$F_{K,j} = \frac{R(\beta_j \mid \beta_0, \beta_\ell : \ell \in S_{K-1})}{SSE(\beta_0, \beta_\ell : \ell \in S_{K-1}, \beta_j)/(n-K-1)}, \quad j \notin S_K$$

For a predetermined $F_{in} > 0$, Stop the procedure at this step and the subset $S^* = S_{K-1}$ is selected as our final model if

$$F_{K,jK} < F_{in}$$

Otherwise update the index subset S_K and proceed to the next step:

$$S_K = S_{K-1} \cup \{j_K\}$$

(Step K Continued)

We also compute a criterion for removing predictors:

$$F^*_{K,j} = \frac{R(\beta_j \mid \beta_0, \beta_\ell : \ell \in S_K \setminus \{j\})}{SSE(\beta_0, \beta_\ell : \ell \in S_K)/(n-K-1)} \quad (j \in S_K)$$

Find a set $D_K = \{\ell \in S_K : F^*_{K,\ell} < F_{out}\}$

Now, update the index subset S_K and proceed to the next step

$$S_K = S_K \setminus D_K$$

* Note : $F_{in} \geq F_{out}$. Suppose that $F_{in} \leq F_{out}$. The for all indices satisfying $F_{in} \leq F_{K,jK} \leq F_{out}$, the index will infinitely come in and go out of our index set S_K and the algorithm will never end.

LASSO REGRESSION.

Q22) what is the mathematical representation of a "LASSO estimator"?

Consider the MLRM of centered version:

$$y_c = X_c\gamma + \varepsilon, \text{ where } \varepsilon_1, \cdots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$$

LASSO estimator is defined as the minimizer of

$$\hat{\gamma}^L(k) = \underset{\gamma \in \mathbb{R}^p}{\operatorname{argmin}} \left(\|y_c - X_c\gamma\|^2 - k\|\gamma\|_1 \right), \quad (k \geq 0)$$

Here, $\|\cdot\|_1$ denotes the L_1 norm defined as:

$$\|\gamma\|_1 = \sum_{j=1}^{p} |\gamma_j|, \quad \gamma = [\gamma_1 \cdots \gamma_p]^T \in \mathbb{R}^p$$

* The LASSO regression is a variant of penalized least square estimator.

Q23) Prove Theorem 8

Consider the MLRM of centered version. Let $\hat{\gamma}^L(k)$ denote the LASSO estimator for some $k \geq 0$. In this case

$$\hat{\gamma}^L(k) = \underset{\gamma \in \mathbb{R}^p : \|\gamma\|_1 \leq d}{\operatorname{argmin}} \|y_c - X_c\gamma\|^2 = \underset{\gamma \in \mathbb{R}^p : \|\gamma\|_1 \leq d}{\operatorname{argmin}} (\gamma - \hat{\gamma})^T X_c^T X_c (\gamma - \hat{\gamma})$$

Where $\hat{\gamma}$ is the OLS estimator for γ and $d = \|\hat{\gamma}^L(k)\|$

(proof)

Let $\mathcal{U}(\gamma, \lambda) = \|y_c - X_c\gamma\|^2 + \lambda(\|\gamma\|_1 - d)$, ($\gamma \in \mathbb{R}^p, \lambda \in \mathbb{R}$, λ is the Lagrange multiplier)

① Fix $\lambda > 0$ and find $\hat{\gamma}(\lambda) = \underset{\gamma}{\operatorname{argmin}}\ \mathcal{U}(\gamma, \lambda)$ } Strategy

② Find $\lambda^* > 0$ s.t. $\|\hat{\gamma}(\lambda^*)\|_1 = d$

If ① and ② are satisfied,

$$\|y_c - X_c\gamma\|^2 \geq \|y_c - X_c\gamma\|^2 + \lambda^*(\|\gamma\|_1 - d)$$

$$= \mathcal{U}(\gamma, \lambda^*)$$
$$\geq \mathcal{U}(\hat{\gamma}(\lambda^*), \lambda^*) \quad \text{by ①}$$
$$= \|y_c - X_c\gamma\|^2 + \lambda^*(\|\hat{\gamma}(\lambda^*)\|_1 - d)$$

$\|\hat{\gamma}(\lambda^*)\|_1 - d = 0$ by ②

Gemini 3.1. Pro explanation.

The objective function to minimize is:

$$L(\beta) = \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$$

$y \in \mathbb{R}^{n \times 1}$
$X \in \mathbb{R}^{n \times p}$
$\lambda > 0$
$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the L_1 norm.

2. The orthonormal Design Assumption.

Because the absolute value function $|\beta_j|$ is not differentiable at zero, we cannot find a closed form analytical solution for the general X matrix.

However, to derive the exact mathematical behavior of the estimator, we make a standard simplifying assumption: Assume the columns of X are orthonormal. This means $X^TX = I$.

Under this assumption:

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

If we expand the RSS portion of the LASSO objective:

$$\frac{1}{2}\|y - X\beta\|^2 = \frac{1}{2}y^Ty - y^TX\beta + \frac{1}{2}\beta^TX^TX\beta$$

$$= \frac{1}{2}y^Ty - (\hat{\beta}^{OLS})^T\beta + \frac{1}{2}\|\beta\|_2^2$$

By completing the square (and ignoring the terms that don't depend on β), minimizing the original objective is mathematically equivalent to minimizing:

$$f(\beta) = \sum_{j=1}^{p} \left(\frac{1}{2}(\beta_j - \hat{\beta}_j^{OLS})^2 + \lambda|\beta_j| \right)$$

3. Subdifferential Calculus.

Because the objective function perfectly decouples, we can optimize for each parameter β_j independently. we need to minimize:

$$f(\beta_j) = \frac{1}{2}(\beta_j - \hat{\beta}_j^{OLS})^2 + \lambda|\beta_j|$$

Since $|\beta_j|$ has a sharp corner at zero, it is not differentiable there. We must use its subdifferential, denoted as $\partial|\beta_j|$, which is defined as:

- $\{1\}$ if $\beta_j > 0$
- $\{-1\}$ if $\beta_j < 0$
- $[-1, 1]$ if $\beta_j = 0$

Taking the subgradient of $f(\beta_j)$ and setting it to zero gives:

$$\beta_j - \hat{\beta}_j^{OLS} + \lambda s = 0, \text{ where } s \in \partial|\beta_j|$$

4. Solving for the 3 cases:

We solve this equation by looking at the three possible regions for β_j:

Case 1: $\beta_j > 0$

Here, $S = 1$. The equation becomes:

$$\beta_j - \hat{\beta}_j^{OLS} + \lambda = 0 \Rightarrow \hat{\beta}_j^{Lasso} = \hat{\beta}_j^{OLS} - \lambda$$

* Condition: for β_j to actually be positive, we must have $\hat{\beta}_j^{OLS} > \lambda$.

Case 2: $\beta_j < 0$

Here, $S = -1$. The equation becomes:

$$\beta_j - \hat{\beta}_j^{OLS} - \lambda = 0 \Rightarrow \hat{\beta}_j^{LASSO} = \hat{\beta}_j^{OLS} + \lambda$$

* condition: For β_j to actually be negative, we must have $\hat{\beta}_j^{OLS} < -\lambda$.

Case 3: $\beta_j = 0$

Here, $S = [-1, 1]$. The equation becomes:

$$0 - \hat{\beta}_j^{OLS} + \lambda S = 0 \Rightarrow \hat{\beta}_j^{OLS} = \lambda \cdot S$$

Since S can be any value between -1 and 1, this implies that $\hat{\beta}_j^{LASSO} = 0$ whenever $\hat{\beta}_j^{OLS}$ falls in the interval $[-\lambda, \lambda]$

5. The final Estimator (Soft-thresholding):

Combine these three cases yields the exact analytical formula for the LASSO estimator (under orthonormal design). It is known as the Soft-thresholding Operator, typically denoted as S_λ:

$$\hat{\beta}_j^{LASSO} = \begin{cases} \hat{\beta}_j^{OLS} - \lambda & \text{if } \hat{\beta}_j^{OLS} > \lambda \\ 0 & \text{if } |\hat{\beta}_j^{OLS}| \leq \lambda \\ \hat{\beta}_j^{OLS} + \lambda & \text{if } \hat{\beta}_j^{OLS} < -\lambda \end{cases}$$

Alternatively, it can be written in a single, compact mathematical expression:

$$\hat{\beta}_j^{LASSO} = \text{Sign}(\hat{\beta}_j^{OLS}) \max\{0, |\hat{\beta}_j^{OLS}| - \lambda\}$$

The interpretation: The math explicitly proves how LASSO achieves feature selection. If the original OLS estimate is smaller than the penalty magnitude λ, the subgradient completely absorbs it, and the coefficient is forced to exactly 0. If it is larger than λ, the coefficient is kept but shrunk towards zero by exactly λ.