STAT GR5703—Spring 2020

Group: Zining Fan, Mutian Wang, Siyuan Wang

UNI: zf2234, mw3386, sw3418

Graded Homework 3- Exercise 4

1.

$$\begin{aligned}\sigma^2\Gamma &= \sum_i E[(\hat{Y} - E[Y_i])^2] \\ &= \sum_i var(\hat{Y} - E[Y_i]) + E[\hat{Y} - E[Y_i]]^2 \\ &= \sum_i var(\delta_i) + E[\hat{Y} - Y_i]^2 \\ &= \sum_i var(\delta_i) + E[(\hat{Y} - Y_i)^2] - var(\hat{Y} - Y_i) \\ &= \sum_i E[(\hat{Y} - Y_i)^2] - var(\hat{\epsilon_i}) + var(\delta_i) \\ &= E[RSS(\hat{\mathbf{Y}}) - \sum_i var(\hat{\epsilon_i}) + \sum_i var(\delta_i)]\end{aligned}$$

2. Note that $var(Y) = \sigma^2\mathbf{I_n}$ since ϵ_i is i.i.d.

$$\begin{aligned}\sum var(\hat{\epsilon_i}) &= tr(var(\mathbf{Y} - \hat{\mathbf{Y}})) \\ &= tr(var(\mathbf{Y} - \mathbf{SY})) \\ &= tr((\mathbf{I} - \mathbf{S})var(\mathbf{Y})(\mathbf{I} - \mathbf{S})^T) \\ &= \sigma^2 tr((\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T)\end{aligned}$$

$$\begin{aligned}\sum var(\delta_i) &= tr(var(\hat{\mathbf{Y}} - E[\mathbf{Y}])) \\ &= tr(var(\hat{\mathbf{Y}})) \\ &= tr(var(\mathbf{SY})) \\ &= \sigma^2 tr(\mathbf{SS}^T)\end{aligned}$$

From above, we know

$$\begin{aligned}-\sum_i var(\hat{\epsilon_i}) + \sum_i var(\delta_i) &= -\sigma^2 tr((\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T) + \sigma^2 tr(\mathbf{SS}^T) \\ &= \sigma^2[-tr(\mathbf{I}) + \mathbf{S}^T + \mathbf{S} - \mathbf{SS}^T + \mathbf{SS}^T] \\ &= \sigma^2(2tr(\mathbf{S}) - tr(\mathbf{I})) \\ &= \sigma^2(2tr(\mathbf{S}) - n)\end{aligned}$$

Now we calculate $E[C]$.

$$\begin{aligned}E[C] &= \frac{1}{\sigma^2}E[RSS(\hat{\mathbf{Y}}) + \sigma^2(2tr(\mathbf{S}) - n)] \\ &= \frac{1}{\sigma^2}E[RSS(\hat{\mathbf{Y}}) - \sum_i var(\hat{\epsilon_i}) + \sum_i var(\delta_i)] \\ &= \Gamma\end{aligned}$$

Therefore, C is an unbiased estimator of Γ.

3. Let $\mathbf{S}$ in point 2 be the projection matrix, i.e. $\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Therefore, $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y} = \mathbf{X}\hat{\beta}$.

   Since $\mathbf{S}$ is a square matrix and $\mathbf{X}$ is a $n \times p$ matrix, so $tr(\mathbf{S}) = tr(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = tr(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) = tr(\mathbf{I_p}) = p$.

   Based on the conclusion of point 2, we know $C_p = \frac{1}{\sigma^2}RSS(\hat{\mathbf{Y}}) + 2tr(\mathbf{S}) - n = \frac{1}{\sigma^2}RSS(\hat{\mathbf{Y}}) + 2p - n$ is an unbiased estimator of the prediction error.

   Remember that $AIC = n \log RSS(\hat{\mathbf{Y}}) + 2p$, so we can see both AIC and C_p are made up with two components. One is about the goodness of fit, i.e. RSS, and the other one is about the model complexity, i.e. the number of features. More specifically, $C_p = \frac{1}{\sigma^2}\exp(\frac{AIC-2p}{n}) + 2p - n$. In a nutshell, they are very similar.

4.

$$\begin{aligned}
&P\{AIC(\hat{\beta}_{q+1}) < AIC(\hat{\beta}_q)\}\\
=&P\{\log\frac{SS(\hat{\beta}_q)}{SS(\hat{\beta}_{q+1})} > \frac{2}{n}\}\\
=&P\{1 + \frac{1}{n-q-1}F_{1,n-q-1} > \exp(\frac{2}{n})\}\\
=&P\{F_{1,n-q-1} > (n-q-1)[\exp(\frac{2}{n}) - 1]\} \xrightarrow{n\to\infty} P\{\chi_1^2 > 0\} = 1
\end{aligned}$$

   When n goes to infinity, $F_{1,n-q-1}$ asymptotically has a χ_1^2 distribution; the right side will go to 0, which can be checked by L'Hospital's rule. Thus the probability goes to 1. That is, when the sample size goes to infinity, AIC will tend to overfit.

5.

$$\begin{aligned}
&P\{BIC(\hat{\beta}_{q+1}) < BIC(\hat{\beta}_q)\}\\
=&P\{\log\frac{SS(\hat{\beta}_q)}{SS(\hat{\beta}_{q+1})} > \frac{\log n}{n}\}\\
=&P\{1 + \frac{1}{n-q-1}F_{1,n-q-1} > \exp(\frac{\log n}{n})\}\\
=&\{F_{1,n-q-1} > (n-q-1)[\exp(\frac{\log n}{n}) - 1]\} \xrightarrow{n\to\infty} P\{\chi_1^2 > \infty\} = 0
\end{aligned}$$

   When n goes to infinity, $F_{1,n-q-1}$ asymptotically has a χ_1^2 distribution; the right side will go to infinity, which can be checked by L'Hospital's rule. Thus the probability goes to 0. That is, when the sample size goes to infinity, BIC won't tend to overfit.