All of Statistics - Chapter 13 Solutions

May 26, 2021

1.

It is easier to work in the multivariate setting for this proof. In light of this, let X_i be a random p dimensional vector. Define X_{-0} as the $n \times p$ matrix whose rows are $X_i^\intercal$. Augment this matrix to obtain $X = (e \mid X_{-0})$ where e is the vector of ones, corresponding to a design matrix with a bias column. Let Y be the vector whose coordinates are Y_i.

Using the fact that $\sum_i \hat{\epsilon}_i^2 = \|Y - X\hat{\beta}\|^2$ and matrix calculus, it is straightforward to show that the RSS is minimized when $\hat{\beta}$ is chosen to satisfy the linear system

$$X^\intercal X \hat{\beta} = X^\intercal Y.$$

Note that

$$X^\intercal Y = \begin{pmatrix} e^\intercal Y \\ X_{-0}^\intercal Y \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ X_{-0}^\intercal Y \end{pmatrix}$$

and

$$X^\intercal X = \begin{pmatrix} n & e^\intercal X_{-0} \\ X_{-0}^\intercal e & X_{-0}^\intercal X_{-0} \end{pmatrix}.$$

Let $\hat{\beta} = (\hat{\beta}_0 \mid \hat{\beta}_{-0})$ where $\hat{\beta}_0$ is a scalar. The first row of the linear system yields

$$\hat{\beta}_0 = \overline{Y} - \frac{1}{n} e^\intercal X_{-0} \hat{\beta}_{-0}.$$

Since $e^\intercal X_{-0} = n\overline{X}$ when $p = 1$, the above is equivalent to Eq. (13.6). Substituting the above into the second row of the linear system yields

$$\left(X_{-0}^\intercal X_{-0} - \frac{1}{n} X_{-0}^\intercal e e^\intercal X_{-0} \right) \hat{\beta}_{-0} = X_{-0}^\intercal Y - X_{-0}^\intercal e \overline{Y}.$$

If $p = 1$, the above simplifies to

$$\left(\sum_i X_i^2 - n\overline{X}^2 \right) \hat{\beta}_1 = \sum_i X_i Y_i - n\overline{X}\,\overline{Y}$$

which, with some work, can be shown to be equivalent to Eq. (13.5).

Next, denoting by $\hat{\epsilon}$ the vector with coordinates $\hat{\epsilon}_i$, we have

$$\hat{\epsilon} = Y - X\hat{\beta} = MY$$

where $M = I - X(X^\intercal X)^{-1}X^\intercal$. Denoting by ϵ the vector with coordinates ϵ_i and β the vector of true coefficients,

$$\hat{\epsilon} = MY = M(X\beta + \epsilon) = M\epsilon.$$

Using the fact that M is both symmetric and idempotent,

$$\mathrm{RSS} = \sum_i \hat{\epsilon}_i^2 = \hat{\epsilon}^\intercal \hat{\epsilon} = \epsilon^\intercal M^\intercal M \epsilon = \epsilon^\intercal M \epsilon.$$

For brevity, we abuse notation by writing $\mathbb{E}f$ to mean $\mathbb{E}[f \mid X]$. Then,

$$\mathbb{E}\left[\mathrm{RSS}\right] = \mathbb{E}\left[\epsilon^\intercal M \epsilon\right] = \operatorname{tr}(M\mathbb{E}\left[\epsilon\epsilon^\intercal\right]).$$

Assuming that ϵ_i and ϵ_j are independent whenever $i \neq j$ yields $\mathbb{E}[\epsilon\epsilon^\intercal] = \sigma^2 I$ and hence

$$\mathbb{E}\left[\mathrm{RSS}\right] = \sigma^2 \operatorname{tr}(M).$$

Moreover,

$$\operatorname{tr}(M) = \operatorname{tr}(I_{n\times n}) - \operatorname{tr}(X^\intercal X(X^\intercal X)^{-1}) = \operatorname{tr}(I_{n\times n}) - \operatorname{tr}(I_{(p+1)\times(p+1)}) = n - (p+1)\,,$$

establishing that (13.7) is an unbiased estimator of the noise variance.

2.

We continue to use the notation established in the answer to the first exercise. First, note that

$$\mathbb{E}Y = \mathbb{E}\left[X\beta + \epsilon\right] = X\beta$$

and

$$\mathbb{E}\left[YY^\intercal\right] = \mathbb{E}\left[\left(X\beta + \epsilon\right)\left(X\beta + \epsilon\right)^\intercal\right] = \mathbb{E}\left[X\beta\beta^\intercal X^\intercal + 2X\beta\epsilon^\intercal + \epsilon\epsilon^\intercal\right] = X\beta\beta^\intercal X^\intercal + \sigma^2 I.$$

Therefore,

$$\mathbb{E}\hat{\beta} = (X^\intercal X)^{-1} X^\intercal \mathbb{E}\left[Y\right] = \beta$$

and

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}\hat{\beta}^\intercal\right] &= \mathbb{E}\left[(X^\intercal X)^{-1} X^\intercal Y Y^\intercal X (X^\intercal X)^{-1}\right] \\ &= (X^\intercal X)^{-1} X^\intercal \mathbb{E}\left[YY^\intercal\right] X (X^\intercal X)^{-1} = \beta\beta^\intercal + \sigma^2 (X^\intercal X)^{-1}.\end{aligned}$$

Combining the above yields

$$\mathbb{V}(\hat{\beta}\hat{\beta}^\intercal) = \mathbb{E}\left[\hat{\beta}\hat{\beta}^\intercal\right] - \mathbb{E}\left[\hat{\beta}\right]\mathbb{E}\left[\hat{\beta}\right]^\intercal = \sigma^2 (X^\intercal X)^{-1}.$$

In the univariate case, the form

$$X^\intercal X = \begin{pmatrix} n & n\overline{X} \\ n\overline{X} & \sum_i X_i^2 \end{pmatrix}$$

can be used to derive a closed form expression for the inverse which in turn yields (13.11) as desired.

3.

A univariate regression through the origin is a special case of the multivariate regression seen in Exercise 1. It has least squares coefficient

$$\frac{\sum_i X_i Y_i}{\sum_i X_i^2}.$$

This is well-defined whenever at least one of the X_i is nonzero.

The standard error of this coefficient is also a special case of the standard error for the multivariate case seen in Exercise 2. It is

$$\frac{\sigma^2}{\sum_i X_i^2}.$$

Since the least squares estimate is an MLE, it is consistent whenever it is well-defined.

4.

Using the fact that Y_i and Y_i^* are IID,

$$\begin{aligned} \mathbb{E}\left[\hat{R}_{\text{tr}}(S)\right] - R(S) &= \sum_i \mathbb{E}\left[\left(\hat{Y}_i(S) - Y_i\right)^2 - \left(\hat{Y}_i(S) - Y_i^*\right)^2\right] \\ &= \sum_i \mathbb{E}\left[\hat{Y}_i(S)^2 - 2\hat{Y}_i(S)Y_i + Y_i^2 - \hat{Y}_i(S)^2 + 2\hat{Y}_i(S)Y_i^* - \left(Y_i^*\right)^2\right] \\ &= \sum_i -2\mathbb{E}\left[\hat{Y}_i(S)Y_i\right] + \mathbb{E}\left[Y_i^2\right] + 2\mathbb{E}\left[\hat{Y}_i(S)Y_i^*\right] - \mathbb{E}\left[\left(Y_i^*\right)^2\right] \\ &= -2\sum_i \mathbb{E}\left[\hat{Y}_i(S)Y_i\right] - \mathbb{E}\left[\hat{Y}_i(S)\right]\mathbb{E}\left[Y_i\right] \\ &= -2\sum_i \operatorname{Cov}(\hat{Y}_i(S), Y_i). \end{aligned}$$

5.

Let $\hat{\delta} = \hat{\beta}_1 - 17\hat{\beta}_0$. By Theorem 13.8,

$$\mathbb{V}(\hat{\delta}) = \mathbb{V}(\hat{\beta}_1) + 17^2\mathbb{V}(\hat{\beta}_0) - 17\,\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{\sigma^2}{n s_X^2}\left(1 + 17\overline{X} + \frac{17^2}{n}\sum_i X_i^2\right).$$

Replacing σ by $\hat{\sigma}$ and taking square roots yields $\hat{\mathrm{se}}(\hat{\delta})$. The Wald statistic is $W = \hat{\delta}/\hat{\mathrm{se}}(\hat{\delta})$.

6.

TODO (Computer experiment).

7.

TODO (Computer experiment).

8.

Maximizing AIC is equivalent to minimizing $-2\sigma^2\mathrm{AIC}$. This is equivalent to minimizing Mallow's C_p statistic since

$$\begin{aligned}
-2\sigma^2\mathrm{AIC} &= -2\sigma^2\ell_S + 2\,|S|\,\sigma^2 \\
&= -2\sigma^2\left\{\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_i\left(\hat{Y}_i(S) - Y_i\right)^2\right\} + 2\,|S|\,\sigma^2 \\
&= \text{const.} + \sum_i\left(\hat{Y}_i(S) - Y_i\right)^2 + 2\,|S|\,\sigma^2 \\
&= \text{const.} + C_p + 2\,|S|\,\sigma^2.
\end{aligned}$$

9.

Choosing the model with the highest AIC is equivalent to choosing the model with the lowest Mallow's C_p statistic. The two models have Mallow's statistics $C_p^0 = \sum_i X_i^2$ and $C_p^1 = [\sum_i (X_i - \hat{\theta})^2] + 2$ with $\hat{\theta} = \overline{X}$. Note that

$$C_p^0 - C_p^1 = \sum_i X_i^2 - \sum_i\left(X_i - \hat{\theta}\right)^2 + 2 = n\hat{\theta}^2 - 2.$$

Therefore, $\mathcal{M}_0$ is picked if and only if $\hat{\theta}^2 < 2/n$.

a)

First, note that $\hat{\theta} \sim N(\theta, 1/n)$. If $\theta = 0$, then

$$\mathbb{P}(J_n = 0) = \mathbb{P}(|\hat{\theta}| < \sqrt{2}n^{-1/2}) = \mathbb{P}(|Z| < \sqrt{2}) = 2\Phi(\sqrt{2}) - 1 \approx 0.8427.$$

If $\theta \neq 0$, then

$$\begin{aligned}
&\mathbb{P}(J_n = 0) = \mathbb{P}(|\hat{\theta}| < \sqrt{2}n^{-1/2}) = \mathbb{P}(|Zn^{-1/2} + \theta| < \sqrt{2}n^{-1/2}) \\
&= \mathbb{P}(-\sqrt{2} - \theta\sqrt{n} < Z < \sqrt{2} - \theta\sqrt{n}) = \Phi(\sqrt{2} - \theta\sqrt{n}) - \Phi(-\sqrt{2} - \theta\sqrt{n}) \to 0.
\end{aligned}$$

b)

Let $\mu = \hat{\theta} I_{\{J_n=1\}}$ so that

$$\hat{f}_n(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2} \right).$$

Let $Z \sim N(0,1)$. The KL distance between ϕ_0 and $\hat{f}_n$ is

$$\begin{aligned}
D(\phi_0, \hat{f}_n) &= \int \phi_0(z) \left(\log \phi_0(z) - \log \hat{f}_n(z) \right) dz \\
&= \mathbb{E}\left[\log \phi_0(Z) - \log \hat{f}_n(Z) \right] \\
&= \frac{1}{2} \mathbb{E}\left[-Z^2 + (Z-\mu)^2 \right] \\
&= \frac{1}{2} \mathbb{E}\left[-2\mu Z + \mu^2 \right] = \frac{1}{2}\mu^2.
\end{aligned}$$

If $\theta = 0$, this quantity converges to zero in probability since

$$\mathbb{P}(\mu^2 > \epsilon) = \mathbb{P}(\hat{\theta}^2 I_{\{J_n=1\}} > \epsilon) \leq \mathbb{P}(\hat{\theta}^2 > \epsilon) = \mathbb{P}(|Z| > \sqrt{n\epsilon}).$$

Next, the KL distance between $\phi_{\hat{\theta}}$ and $\hat{f}_n$ is

$$\begin{aligned}
D(\phi_{\hat{\theta}}, \hat{f}_n) &= \int \phi_{\hat{\theta}}(x) \left(\log \phi_{\hat{\theta}}(x) - \log \hat{f}_n(x) \right) dx \\
&= \int \phi_0(z) \left(\log \phi_0(z) - \log \hat{f}_n(z + \hat{\theta}) \right) dz \\
&= \mathbb{E}\left[\log \phi_0(Z) - \log \hat{f}_n(Z + \hat{\theta}) \right] \\
&= \frac{1}{2} \mathbb{E}\left[-Z^2 + \left(Z + \hat{\theta} - \mu \right)^2 \right] \\
&= \frac{1}{2} \mathbb{E}\left[2\left(\hat{\theta} - \mu \right) Z + \hat{\theta}^2 - 2\hat{\theta}\mu + \mu^2 \right] \\
&= \frac{1}{2} \left(\hat{\theta}^2 - 2\hat{\theta}\mu + \mu^2 \right).
\end{aligned}$$

By the LLN, $\hat{\theta}$ converges to θ in probability. Suppose that $\theta \neq 0$. Our findings in Part (a) imply that $I_{\{J_n=1\}}$ converges to one in probability. Therefore, by Theorem 5.5, μ converges to θ in probability and hence $D(\phi_{\hat{\theta}}, \hat{f}_n)$ converges to zero in probability.

c)

Noting that the only difference between the AIC and BIC criteria is replacing the penalty of 2 by $\log n$, we can conclude that if $\theta = 0$, then

$$\mathbb{P}(J_n = 0) = 2\Phi(\sqrt{\log n}) - 1 \to 1.$$

Recall that even in the limit, the corresponding quantity for AIC was not one. Similarly, if $\theta \neq 0$, then

$$\mathbb{P}(J_n = 0) = \Phi(\sqrt{\log n} - \theta\sqrt{n}) - \Phi(-\sqrt{\log n} - \theta\sqrt{n}) \to 0.$$

The limiting KL distances are also as before.

10.

a)

Suppose $\epsilon \sim N(0, \sigma^2)$. Since ϵ is independent of $\hat{\theta}$ (recall that X_* correspond to a sample that hasn't been trained on),

$$\frac{Y_* - \hat{Y}_*}{s} = -\frac{\hat{\theta} - \theta}{s} + \frac{\epsilon}{s} \approx N\left(0, 1 + \frac{\sigma^2}{s^2}\right).$$

b)

Similarly to Part (a),

$$\begin{aligned} \frac{Y_* - \hat{Y}_*}{\xi_n} = -\frac{\hat{\theta} - \theta}{\xi_n} + \frac{\epsilon}{\xi_n} &= -\frac{\hat{\theta} - \theta}{s} \frac{s}{\sqrt{s^2 + \sigma^2}} + \frac{\epsilon}{\sqrt{s^2 + \sigma^2}} \\ &\approx N\left(0, \frac{s^2}{s^2 + \sigma^2}\right) + N\left(0, \frac{\sigma^2}{s^2 + \sigma^2}\right) = N(0, 1). \end{aligned}$$

11.

TODO (Computer experiment).