Elements of Statistical Learning, Solutions
Marios

Exercises for Section 2

1. Suppose each of K-classes has an associated target t_k, which is a vector of all zeros, except a one in the kth position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target, $\min_k \|t_k - \hat{y}\|$, if the elements of $\hat{y}$ sum to one.

   Solution. Let $k^* = \arg\max_k \hat{y}_k$ and suppose that there is $k' \leq k^*$ such that $\|t_{k'} - \hat{y}\| < \|t_{k^*} - \hat{y}\|$.

   - ℓ_1 norm. It holds that $\|t_k - \hat{y}\|_1 = \sum_i |t_{k,i} - \hat{y}_i| = \sum_{i \neq k} |\hat{y}_i| + |1 - \hat{y}_k|$. Hence, we get

$$\sum_{i \neq k'} |\hat{y}_i| + |1 - \hat{y}_{k'}| < \sum_{i \neq k^*} |\hat{y}_i| + |1 - \hat{y}_{k^*}| \Rightarrow |\hat{y}_{k^*}| - |1 - \hat{y}_{k^*}| < |\hat{y}_{k'}| - |1 - \hat{y}_{k'}|. \tag{1}$$

     But the function $f(y) = |y| - |1 - y|$ is increasing in $[0, 1]$ hence Equation (1) implies that $\hat{y}_{k^*} < \hat{y}_{k'}$, reaching a contradiction.

   - ℓ_2 norm. Similarly, we get that $\hat{y}_{k^*}(1 - \hat{y}_{k^*}) < \hat{y}_{k'}(1 - \hat{y}_{k'})$ and since the function $f(y) = y(1 - y)$ is increasing in $[0, 1]$, we get that $\hat{y}_{k^*} < \hat{y}_{k'}$, reaching a contradiction.

2. Show how to compute the Bayes decision boundary for the simulation example in Figure 2.5.

   Solution. If we know the exact probability distribution $\Pr[G, X]$, $X \in \mathbb{R}^p$, $G \in \mathcal{G} = \{B, O\}$, then we can probably also derive $f(X) = \Pr[B|X] = \Pr[B, X]/\Pr[X]$, namely the probability that X maps to blue in reality. This assume that we also know $\Pr[X]$ which is not necessary. Of course, $\Pr[O|X] = 1 - \Pr[B|X]$. So now, all we have to do is to check for each $x \in \mathbb{R}^p$, whether $f(x) > 1/2$. For the case where $x \in \mathbb{R}$, this is trivial. We simply solve the equation $f(x) = 1/2$. This also hold in general. So the points (in $\mathbb{R}$), the line (in $\mathbb{R}^2$), and the $(p-1)$-dimensional hyperplane (in $\mathbb{R}^p$), is the solution to the equation $f(x) = \Pr[B|X] = 1/2$. See Figure 2 for another example.

3. Derive equation 2.24. Consider N data points uniformly sampled in a p-dimensional unit ball centered at the origin. Show that the median distance from the origin to the closest data point is given by the expression

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}.$$

   Solution. We start with the cumulative distribution function (CDF) of the distance of a random point from the origin. The volume of a p-dimensional ball of radius d is $V_p(d) = c_p d^p$, where c_p is a value that does not depend on d. Therefore,

$$F_D(d) = \Pr[D \leq d] = \frac{V_p(d)}{V_p(1)} = d^p. \qquad \text{(First trick to remember)}$$

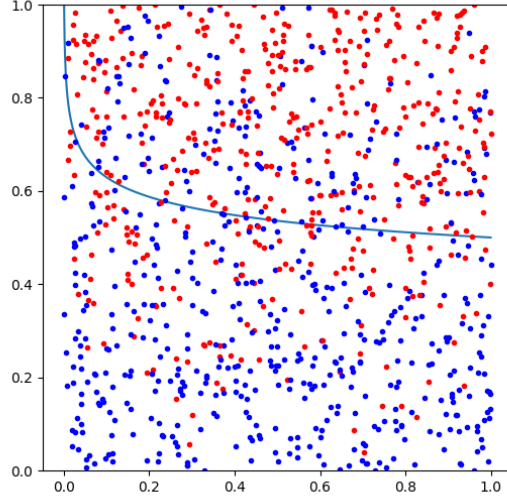   Now it is useful to compute the CDF of the distance of the closest point $C = \min_{i \in [N]} D_i$. We have

Figure 1: In this example we have computed the Bayes decision boundary when $X \sim U(0,1)^2$ and $\Pr[Y = \text{red}|X] = X_1^{1/10} X_2$. Therefore, the line is the solution to the equation $X_1^{1/10} X_2 = 1/2$.

that

$$
\begin{aligned}
F_C(d) &= \Pr[C \leq d] \\
&= 1 - \Pr[C \geq d] \\
&= 1 - \Pr\left[\min_{i \in [N]} D_i \geq d\right] \\
&= 1 - \Pr[\forall i \in [N], D_i \geq d] \\
&= 1 - \prod_{i \in [N]} \Pr[D_i \geq d] \\
&= 1 - \Pr[D \geq d]^N \\
&= 1 - (1 - \Pr[D \leq d])^N \\
&= 1 - (1 - d^p)^N.
\end{aligned}
\tag{2}
$$

By definition, the median m is defined as $F_C(m) = 1/2$. Hence, we get that $(1 - m^p)^N = 1/2$ and solving for m, we get

$$
m = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}.
$$

4. Consider inputs drawn from a spherical multinormal distribution $X \sim N(0, \mathbf{I}_p)$. The squared distance from any sample point to the origin has a χ_p^2 distribution with mean p. Consider a prediction point x_0 drawn from this distribution, and let $a = x_0/\|x_0\|$ be an associated unit vector. Let $z_i = a^T x_i$ be the projection of each of the training points on this direction.

   Show that the z_i are distributed according to $N(0,1)$ with expected squared distance from the origin 1, while the target point has expected squared distance p from the origin.

   Solution. We use the fact that for any $a \in \mathbb{R}^p$, if $x \sim N(0, \mathbf{I}_p)$, then $a^T x \sim N\left(\sum_j a_j \mu_j, \sum_j a_j^2 \sigma_j^2\right)$, where $\mu_j = E(x_j)$ and $\sigma_j = V(x_j)$ and $j \in [p]$. Since $\sigma_j = 1$ and $\mu_j = 0$, we get that $a^T x \sim$

$N\left(0, \sum_j a_j^2\right)$. Given that $\|a\|$ is a unit vector, we get that $a^T x \sim N(0,1)$. Hence $|z| = |a^T x| \sim \chi_1^2$ and $E[|z|] = 1$.

5. Suppose that we know that the true relationship between Y and X is linear,

$$Y = X^T \beta + \varepsilon, \tag{2.26}$$

where $\varepsilon \sim N(0, \sigma^2)$ and we fit the model by least squares to the training data. For an arbitrary test point x_0, we have $\hat{y}_0 = x_0^T \hat{\beta}$, which can be written as $\hat{y}_0 = x_0^T \beta + \sum_{i-1}^N \ell_i(x_0)\varepsilon_i$, where $\ell_i(x_0)$ is the ith element of $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0$. Show that

$$\text{EPE}(x_0) = \sigma^2 + E_{\mathcal{T}} x_0^T (\mathbf{X}^T\mathbf{X})^{-1} x_0 \sigma^2 + 0^2, \tag{2.27}$$

where you can use the fact that for any $\mathbf{X}$,

$$\text{Cov}[\hat{\beta}] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2. \tag{3.8}$$

Additionally, suppose N is large and $\mathcal{T}$ were selected at random. Assuming $E(X) = 0$, then $\mathbf{X}^T\mathbf{X} \to N\,\text{Cov}(X)$. Show that

$$\begin{aligned}
E_{x_0} \text{EPE}(x_0) &\approx E_{x_0} x_0^T \, \text{Cov}[X]^{-1} x_0 \sigma^2 / N + \sigma^2 \\
&= \text{trace}[\text{Cov}[X]^{-1} \, \text{Cov}[x_0]]\sigma^2 / N + \sigma^2 \\
&= \sigma^2(p/N) + \sigma^2.
\end{aligned} \tag{2.28}$$

Make use of the cyclic property of the trace operator ($\text{trace}(AB) = \text{trace}(BA)$), and its linearity (which allows us to interchange the order of trace and expectation).

Solution. In the first question, the test point x_0 is arbitrary and not sampled from the distribution. Thus the randomness is only over:

- the samples $\mathcal{T}$,
- the error ε.

In the second part, we also sample x_0 and hence we consider the expectation of $\text{EPE}(x_0)$.

(a) We start by showing that the expected prediction error equals the sum of the variance of the system, the variance of the model and the squared bias of the model:

$$\begin{aligned}
\text{EPE}(x_0) &= E_{\mathcal{T},\varepsilon}[(y_0 - \hat{y}_0)^2] \\
&= E_{\mathcal{T},\varepsilon}[y_0^2 - 2y_0\hat{y}_0 + \hat{y}_0^2] \\
&= E_{\mathcal{T},\varepsilon}[y_0^2] - 2E_{\mathcal{T},\varepsilon}[y_0\hat{y}_0] + E_{\mathcal{T},\varepsilon}[\hat{y}_0^2] \\
&= E_{\mathcal{T},\varepsilon}[y_0^2] - 2x_0^T \beta E_{\mathcal{T},\varepsilon}[\hat{y}_0] + E_{\mathcal{T},\varepsilon}[\hat{y}_0^2] \\
&= E_{\mathcal{T},\varepsilon}[y_0^2]\boxed{-E_{\mathcal{T},\varepsilon}[y_0]^2 + E_{\mathcal{T},\varepsilon}[y_0]^2} - 2x_0^T \beta E_{\mathcal{T},\varepsilon}[\hat{y}_0] + E_{\mathcal{T},\varepsilon}[\hat{y}_0^2] - \boxed{E_{\mathcal{T},\varepsilon}[\hat{y}_0]^2 + E_{\mathcal{T},\varepsilon}[\hat{y}_0]^2} \\
&= \text{Var}[y_0] + \text{Var}[\hat{y}_0] + \left(E_{\mathcal{T},\varepsilon}[\hat{y}_0] - x_0^T \beta\right)^2,
\end{aligned}$$

where the third line follows from the fact that

$$E_{\mathcal{T},\varepsilon}[y_0\hat{y}_0] = E_{\mathcal{T},\varepsilon}[(x_0^T \beta + \varepsilon)\hat{y}_0] = x_0^T \beta E_{\mathcal{T},\varepsilon}[\hat{y}_0] + E_{\mathcal{T},\varepsilon}[\varepsilon\hat{y}_0]$$

and $E_{\mathcal{T},\varepsilon}[\varepsilon\hat{y}_0] = E_{\mathcal{T},\varepsilon}[\varepsilon]E_{\mathcal{T},\varepsilon}[\hat{y}_0] = 0$, since ε is independent from $\hat{y}_0$. Now, we have that $\text{Var}[y_0] =$

σ^2. Moreover,

$$
\begin{aligned}
E_{\mathcal{T},\varepsilon}[\hat{y}_0] &= E_{\mathcal{T},\varepsilon}\left[x_0^T \beta + \sum_{i-1}^{N} \ell_i(x_0)\varepsilon_i\right] \\
&= x_0^T \beta + \sum_{i-1}^{N} E_{\mathcal{T},\varepsilon}[\ell_i(x_0)\varepsilon_i] \\
&= x_0^T \beta + \sum_{i-1}^{N} E_{\mathcal{T},\varepsilon}[\ell_i(x_0)] E_{\mathcal{T},\varepsilon}[\varepsilon_i] \\
&= x_0^T \beta,
\end{aligned}
\tag{3}
$$

since ε_i is independent of x_0 and $\mathbf{X}$. Hence, $\mathrm{Bias}(\hat{y}_0) = \left(E_{\mathcal{T},\varepsilon}[\hat{y}_0] - x_0^T \beta\right) = 0$. Last, we want to calculate the variance of our prediction $\mathrm{Var}[\hat{y}_0]$. We have

$$
\begin{aligned}
\mathrm{Var}[\hat{y}_0] &= \mathrm{Var}[x_0^T \hat{\beta}] \\
&= x_0^T \mathrm{Cov}[\hat{\beta}] x_0 \\
&= x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 x_0,
\end{aligned}
\tag{4}
$$

where the last line comes from eq. (3.8).

(b) Now we have that

$$
\begin{aligned}
E_{x_0} \mathrm{EPE}(x_0) &\approx E_{x_0}[x_0^T \mathrm{Cov}[X]^{-1} x_0]\sigma^2/N + \sigma^2 \\
&= E_{x_0}[\mathrm{trace}[x_0^T \mathrm{Cov}[X]^{-1} x_0]]\sigma^2/N + \sigma^2 \\
&= E_{x_0}[\mathrm{trace}[\mathrm{Cov}[X]^{-1} x_0 x_0^T]]\sigma^2/N + \sigma^2 \\
&= \mathrm{trace}[E_{x_0}[\mathrm{Cov}[X]^{-1} x_0 x_0^T]]\sigma^2/N + \sigma^2 \\
&= \mathrm{trace}[\mathrm{Cov}[X]^{-1} E_{x_0}[x_0 x_0^T]]\sigma^2/N + \sigma^2 \\
&= \mathrm{trace}[\mathrm{Cov}[X]^{-1} \mathrm{Cov}[x_0]]\sigma^2/N + \sigma^2 \\
&= \mathrm{trace}[\mathbf{I}_p]\sigma^2/N + \sigma^2 \\
&= p\sigma^2/N + \sigma^2
\end{aligned}
\tag{5}
$$

We see that the function $f(p) = E_{x_0} \mathrm{EPE}(x_0) = \sigma^2(p/N) + \sigma^2$ increases linearly with p, with slope $f'(p) = \sigma^2/N$. Hence as long as we have sufficiently many samples N this increase becomes negligible. In other words, even if we have a lot of dimensions, the expected EPE remains constant. Of course, the reason is that we imposed heavy restrictions on the class of models being fitted.

6. Consider a regression problem with inputs x_i and outputs y_i, and a parameterized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with *tied* or *identical* values of x, then the fit can be obtained from a reduced weighted least squares problem.

Solution. The weighted least squares problem is defined as the problem of minimizing the value

$$
\mathrm{WRSS}_{(x_i, y_i, w_i)}(\beta) = \sum_{i=1}^{N} w_i(y_i - \hat{f}_\beta(x_i))^2.
$$

It generasizes RSS since by setting $w_i = 1$ we get the RSS. The idea behind WRSS is that some pairs (x_i, y_i) may have errors and some may be more accurate. By giving them weights, we reward the more accurate ones and we penalize the less accurate.

A reduced least squares problem is one that uses fewer observations than available; $N' < N$.

Suppose that we have N observations (x_i, y_i) with some of them sharing the same x_i. Suppose we have N' distinct x_is and for each distinct x_i, we have N_i observations, so $N = \sum_{i=1}^{N'} N_i$. We wish to compute the value $\arg\min_\beta \text{RSS}(\beta)$. We will show that

$$\arg\min_\beta \text{RSS}_{(x_i, y_i)}(\beta) = \arg\min_\beta \text{WRSS}_{(x_i, \bar{y}_i, N_i)}(\beta).$$

We have

$$
\begin{aligned}
\arg\min_\beta \text{RSS}_{(x_i, y_i)}(\beta) &= \arg\min_\beta \sum_{i=1}^{N} (y_i - \hat{f}_\beta(x_i))^2 \\
&= \arg\min_\beta \sum_{i=1}^{N'} \sum_{j=1}^{N_i} (y_i - \hat{f}_\beta(x_i))^2 \\
&= \arg\min_\beta \sum_{i=1}^{N'} \sum_{j=1}^{N_i} (y_i^2 - 2 y_i \hat{f}_\beta(x_i) + \hat{f}_\beta^2(x_i)) \\
&= \arg\min_\beta \sum_{i=1}^{N'} \sum_{j=1}^{N_i} (-2 y_i \hat{f}_\beta(x_i) + \hat{f}_\beta^2(x_i)) \\
&= \arg\min_\beta \sum_{i=1}^{N'} (-2 N_i \bar{y}_i \hat{f}_\beta(x_i) + N_i \hat{f}_\beta^2(x_i)) \\
&= \arg\min_\beta \sum_{i=1}^{N'} N_i (-2 \bar{y}_i \hat{f}_\beta(x_i) + \hat{f}_\beta^2(x_i)) \\
&= \arg\min_\beta \sum_{i=1}^{N'} N_i (\bar{y}_i^2 - 2 \bar{y}_i \hat{f}_\beta(x_i) + \hat{f}_\beta^2(x_i)) \\
&= \arg\min_\beta \sum_{i=1}^{N'} N_i (\bar{y}_i - \hat{f}_\beta(x_i))^2 \\
&= \arg\min_\beta \text{WRSS}_{(x_i, \bar{y}_i, N_i)}(\beta),
\end{aligned}
\tag{6}
$$

where the fourth and the seventh lines come from the fact that for any function f and $c \in \mathbb{R}$, it holds that $\arg\min_\beta (f(\beta) + c) = \arg\min_\beta f(\beta)$.

7. Suppose we have a sample of N pairs x_i, y_i drawn i.i.d from the distribution characterized as follows:

$x_i \sim h(x)$, the design density

$y_i = f(x_i) + \varepsilon_i$, f is the regression function

$\varepsilon_i \sim (0, \sigma^2)$ (mean zero, variance σ^2)

We construct an estimator for f *linear* in the y_i,

$$\hat{f}(x_0) = \sum_{i=1}^{N} \ell_i(x_0; \mathcal{X}) y_i,$$

where the weights $\ell_i(x_0; \mathcal{X})$ do not depend on the y_i, but do depend on the entire training sequence of x_i, denoted here by $\mathcal{X}$.

(a) Show that the linear regression and k-nearest-neighbor regression are members of this class of estimators. Describe explicitly the weights $\ell_i(x_0; \mathcal{X})$ in each of these cases.

(b) Decompose the conditional mean-squared error

$$E_{\mathcal{Y}|\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$$

into a conditional squared bias and a conditional variance component. Like $\mathcal{X}$, $\mathcal{Y}$ represents the entire training sequence of y_i.

(c) Decompose the mean-squared error $E_{\mathcal{X},\mathcal{Y}}(f(x_0) - \hat{f}(x_0))^2$ into a squared bias and a variance component.

(d) Establish a relationship between the squared biases and variances in the above two cases.

Solution. (a) For linear regression, we have that $\hat{f}(x_0) = x^T(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}y = \sum_{i=1}^{N} \ell_i(x_0;\mathcal{X})y_i$, where $\ell_i(x_0;\mathcal{X})$ is the ith element of the vector $x^T(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}$. For the k-nearest neighbor, we have that $\ell_i(x_0;\mathcal{X}) = \frac{1}{k}I(x_i \in N_k(x_0, \mathcal{X}))$, where $N_k(x_0, \mathcal{X})$ is the set of the k closest points to x_0.

(b) We have that

$$
\begin{aligned}
E_{\mathcal{Y}/\mathcal{X}}[(f(x_0) - \hat{f}(x_0))^2] &= E_{\mathcal{Y}/\mathcal{X}}[f^2(x_0) - 2f(x_0)\hat{f}(x_0) + \hat{f}^2(x_0)] \\
&= \mathrm{Var}_{\mathcal{Y}/\mathcal{X}}(f(x_0)) + \mathrm{Var}_{\mathcal{Y}/\mathcal{X}}(\hat{f}(x_0)) + E_{\mathcal{Y}/\mathcal{X}}[(f(x_0) - \hat{f}(x_0))]^2 \\
&= \mathrm{Var}_{\mathcal{Y}/\mathcal{X}}(f(x_0)) + \mathrm{Var}_{\mathcal{Y}/\mathcal{X}}(\hat{f}(x_0)) + \mathrm{Bias}_{\mathcal{Y}/\mathcal{X}}[\hat{f}(x_0)]^2 \\
&= \mathrm{Var}_{\mathcal{Y}/\mathcal{X}}(\hat{f}(x_0)) + \mathrm{Bias}_{\mathcal{Y}/\mathcal{X}}[\hat{f}(x_0)]^2
\end{aligned}
\tag{7}
$$

since $f(x_0)$ does not have any randomness.

(c) Similarly,

$$E_{\mathcal{Y},\mathcal{X}}[(f(x_0) - \hat{f}(x_0))^2] = \mathrm{Var}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)) + \mathrm{Bias}_{\mathcal{Y},\mathcal{X}}[\hat{f}(x_0)]^2. \tag{8}$$

(d) We have that

$$\mathrm{Var}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)) = E_{\mathcal{X}\sim h}[\mathrm{Var}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0))]$$

$$\mathrm{Bias}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)) = E_{\mathcal{X}\sim h}[\mathrm{Bias}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0))]$$

8. Compare the classification performance of linear regression and k-nearest neighbor classification on the `zipcode` data. In particular, consider only the 2's and 3's and $k = 1, 3, 5, 7, 15$. Show both the training and test error for each choice. The `zipcode` data are available from the book website `https://hastie.su.domains/ElemStatLearn/`.

Solution. We use the `sklearn` library of Python.

```
Linear Regression
Training Error
2.48%
Testing Error
15.17%
k-nearest neighbors classifier
k = 1
Training Error
0.00%
Testing Error
2.47%
k = 3
Training Error
0.50%
Testing Error
3.02%
k = 5
Training Error
0.58%
Testing Error
3.02%
```

```
k = 7
Training Error
0.65%
Testing Error
3.30%
k = 15
Training Error
0.94%
Testing Error
3.85%
```

9. Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population as the training data. If $R_{\text{tr}}(\beta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2$ and $R_{\text{te}}(\beta) = \frac{1}{M} \sum_{i=1}^{M} (\tilde{y}_i - \beta^T \tilde{x}_i)^2$, prove that

$$E\left[R_{\text{tr}}(\hat{\beta})\right] \leq E\left[R_{\text{te}}(\hat{\beta})\right],$$

where the expectations are over all that is random in each expression (meaning the population).

Solution. We will prove a more general result. Let S be the training set and T be the testing set. Moreover, let $f(S, \beta) = E_i[f(S_i, \beta)]$ be the function we want to minimize (in our case it will be the RSS or its normalized version, the mean squared error MSE). Observe that for any β,

$$\begin{aligned}
E_S[f(S, \beta)] &= E_S E_i[f(S_i, \beta)] \\
&= E_i E_S[f(S_i, \beta)] \\
&= E_i E_S[f(S_1, \beta)] \\
&= E_S[f(S_1, \beta)],
\end{aligned} \tag{9}$$

since all S_i are i.i.d.. Let $\beta_S = \arg\min_\beta f(S, \beta)$ and observe that for any β, it holds that $f(S, \beta_S) \leq f(S, \beta)$. Let T' be the set T, truncated or augmented by sampling more data to match the size of S. We have

$$\begin{aligned}
E_{S,T}[f(T, \beta_S)] &= E_S E_{T/S}[f(T, \beta_S)] \\
&\geq E_S E_{T/S}[f(T, \beta_{T'})] \\
&= E_S E_T[f(T, \beta_{T'})] \\
&= E_T[f(T, \beta_{T'})] \\
&= E_T[f(T_1, \beta_{T'})] \\
&= E_{T'}[f(T_1, \beta_{T'})] \\
&= E_S[f(S_1, \beta_S)] \\
&= E_S[f(S, \beta_S)]
\end{aligned} \tag{10}$$

where the second line comes from the above inequality and the 5th and the 8th lines come from Equation (9).

For our case, S is the training set and T is the testing set. Moreover, we have that

$$\arg\min_\beta (\text{RSS}(\beta)) = \arg\min_\beta (\frac{1}{N} \text{RSS}(\beta)) = \arg\min_\beta (\text{MSE}(\beta)).$$

Hence it is enough to consider the function $f(S, \beta) = \text{MSE}(\beta)$, $\text{tr} = S$, $\text{te} = T$ and $\hat{\beta} = \beta_S$.

We see that this inequality is illustrated in Exercise 8.

Exercises for Section 3

1. Show that the F statistic (3.13) for dropping a single coefficient from a model is equal to the square of the corresponding z-score (3.12).

Solution. The z-score for dropping the variable $j \in [p]$ is defined as

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}},$$

where v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. Moreover, the F statistic for dropping variable j is

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)} = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1/(N - p - 1)} = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\hat{\sigma}^2}$$

Hence, it suffices to show that $\frac{\hat{\beta}_j^2}{u_j} = \mathrm{RSS}_0 - \mathrm{RSS}_1$.

Letting $V = (X^T X)^{-1}$, we have that $X^T y = V^{-1}\hat{\beta}$ and

$$\begin{aligned}
\mathrm{RSS} &= y^T y - y^T X \hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X \hat{\beta} \\
&= y^T y - \hat{\beta}^T V^{-1}\hat{\beta}
\end{aligned} \tag{11}$$

and $\mathrm{RSS}_0 - \mathrm{RSS}_1 = \hat{\beta}_0^T V_0^{-1}\hat{\beta}_0 - \hat{\beta}_1^T V_1^{-1}\hat{\beta}_1$.