

# 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  $k$ th 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  $\|t_{k'} - \hat{y}\| < \|t_{k^*} - \hat{y}\|$ .

- $\ell_1$  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

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

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.

- $\ell_2$  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,

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

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

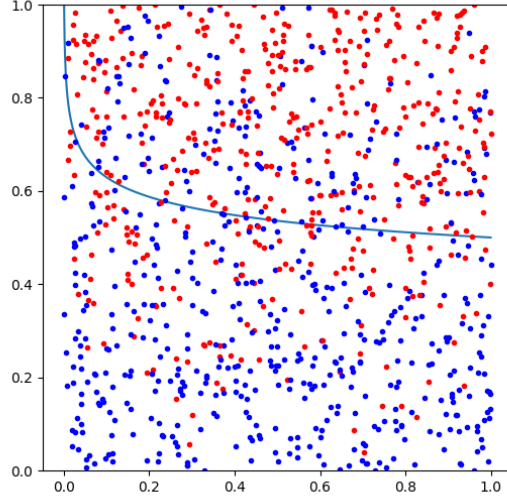


Figure 1: In this example we have computed the Bayes decision boundary when  $X \sim U(0,1)^2$  and  $\Pr[Y = \text{red}|X] = X_1^{1/10}X_2$ . Therefore, the line is the solution to the equation  $X_1^{1/10}X_2 = 1/2$ .

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  $\chi_p^2$  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  $a^T x \sim$

$N(0, \sum_j a_j^2)$ . 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,

$$Y = X^T \beta + \varepsilon, \quad (2.26)$$

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  $i$ th element of  $\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} x_0$ . Show that

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

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

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

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

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

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  $\varepsilon$ .

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:

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

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  $\varepsilon$  is independent from  $\hat{y}_0$ . Now, we have that  $\text{Var}[y_0] =$

$\sigma^2$ . 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  $\varepsilon_i$  is independent of  $x_0$  and  $\mathbf{X}$ . Hence,  $\text{Bias}(\hat{y}_0) = (E_{\mathcal{T},\varepsilon}[\hat{y}_0] - x_0^T \beta) = 0$ . Last, we want to calculate the variance of our prediction  $\text{Var}[\hat{y}_0]$ . We have

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

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

(b) Now we have that

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

We see that the function  $f(p) = E_{x_0} \text{EPE}(x_0) = \sigma^2(p/N) + \sigma^2$  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

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

It generalizes 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  $x_i$ s 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

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

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:

$$\begin{aligned} x_i &\sim h(x), \text{ the design density} \\ y_i &= f(x_i) + \varepsilon_i, f \text{ is the regression function} \\ \varepsilon_i &\sim (0, \sigma^2) \text{ (mean zero, variance } \sigma^2) \end{aligned}$$

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  $i$ th 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

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

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

- (c) Similarly,

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

- (d) We have that

$$\begin{aligned} \text{Var}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)) &= E_{\mathcal{X} \sim h}[\text{Var}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0))] \\ \text{Bias}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)) &= E_{\mathcal{X} \sim h}[\text{Bias}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0))] \end{aligned}$$

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  $(x_1, y_1), \dots, (x_N, y_N)$  drawn at random from a population. Let  $\hat{\beta}$  be the least squares estimate. Suppose we have some test data  $(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_M, \tilde{y}_M)$  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

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

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  $\beta$ ,

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

since all  $S_i$  are i.i.d.. Let  $\beta_S = \arg \min_{\beta} f(S, \beta)$  and observe that for any  $\beta$ , 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

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

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

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

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

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

$$\begin{aligned} \text{RSS} &= y^T y - y^T X \hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X \hat{\beta} \\ &= y^T y - \hat{\beta}^T V^{-1} \hat{\beta} \end{aligned} \quad (11)$$

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

2. Given data on two variables  $X$  and  $Y$ , consider fitting a cubic polynomial regression model  $f(X) = \sum_{j=0}^3 \beta_j X^j$ . In addition to plotting the fitted curve, you would like a 95% confidence band about the curve. Consider the following two approaches:

- (a) At each point  $x_0$ , form a 95% confidence interval for the linear function  $a^T \beta = \sum_{j=0}^3 \beta_j x_0^j$ .
- (b) For a 95% confidence set for  $\beta$  as in (3.15), which in turn generates confidence intervals for all  $f(x_0)$ .

How do these approaches differ? Which band is likely to be wider? Conduct a small simulation experiment to compare the two methods.

**Solution.** We consider design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix} \quad (12)$$

response matrix

$$\mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_N \end{bmatrix} \quad (13)$$

and that the true model is  $y = x^T \beta + \varepsilon$ , where  $x^T = [1, x, x^2, x^3]$  and  $\varepsilon \sim N(0, \sigma^2)$ .

By applying ordinary least squares we get

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$$

with  $\hat{\beta} \sim N(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$ .

- (a) For any  $a^T$ , it holds that  $a^T \hat{\beta}$  is a univariate normal distribution:

$$a^T \hat{\beta} \sim N(a^T \beta, a^T (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 a).$$

Therefore, the confidence interval for  $a^T \beta$  is

$$C_{a^T \beta} = a^T \hat{\beta} \pm 2 \cdot \hat{\sigma} \sqrt{a^T (\mathbf{X}^T \mathbf{X})^{-1} a}$$



(b) For a confidence interval of the whole  $\beta$ , we apply (3.15) and we get that

$$C_\beta = \left\{ \beta \mid (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \beta) \leq \hat{\sigma}^2 \chi_4^2(1-0.05) \right\}$$

and therefore

$$C_{a^T \beta} = \left\{ a^T \beta \mid (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \beta) \leq \hat{\sigma}^2 \chi_4^2(1-0.05) \right\}$$

Since  $\chi_4^2(1-0.05) > 2$ , we expect the second band to be wider.

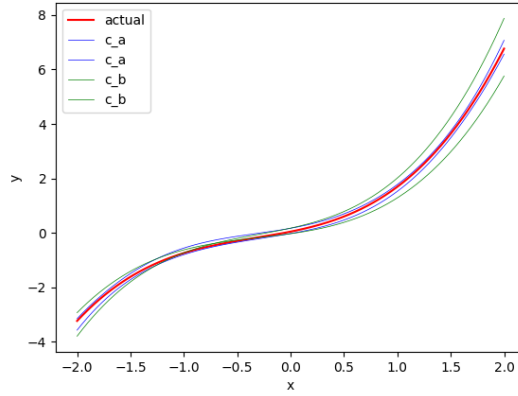


Figure 2: The actual curve together with the two confidence bands.

### 3. Gauss-Markov theorem:

- (a) Prove the Gauss-Markov theorem: the least squares estimate of a parameter  $a^T \beta$  has variance no bigger than that of any other linear unbiased estimate of  $a^T \beta$ .
- (b) The matrix inequality  $\mathbf{B} \preceq \mathbf{A}$  holds if  $\mathbf{A} - \mathbf{B}$  is positive semidefinite. Show that if  $\hat{\mathbf{V}}$  is the variance-covariance matrix of the least squares estimate  $\hat{\beta}$  and  $\tilde{\mathbf{V}}$  is the variance-covariance matrix of any other linear unbiased estimate, then  $\hat{\mathbf{V}} \preceq \tilde{\mathbf{V}}$ .

**Solution.** (a) Let  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  be the least squares estimate and consider the estimator of  $a^T \beta$  to be

$$\hat{\theta} = a^T \hat{\beta} = \mathbf{c}_0^T \mathbf{y}.$$

Now consider any other unbiased linear estimator  $\tilde{\theta} = \mathbf{c}^T \mathbf{y}$  of  $a^T \beta$ ; i.e.,  $E[\mathbf{c}^T \mathbf{y}] = a^T \beta$ . We write  $\mathbf{c}^T = \mathbf{c}_0^T + d^T$  for some  $d$  and we have:

$$\begin{aligned} E[\tilde{\theta}] &= E[(\mathbf{c}_0^T + d^T) \mathbf{y}] \\ &= E[a^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} + d^T \mathbf{y}] \\ &= a^T \beta + d^T E[\mathbf{y}] \\ &= a^T \beta + d^T \mathbf{X} \beta \end{aligned} \tag{14}$$

From which, we conclude that

$$d^T \mathbf{X} = 0$$

We now compute the variance of  $\tilde{\theta}$ :

$$\begin{aligned}
\text{Var}[\tilde{\theta}] &= \text{Var}[\mathbf{c}^T \mathbf{y}] \\
&= \mathbf{c}^T \text{Var}[\mathbf{y}] \mathbf{c} \\
&= \sigma^2 \mathbf{c}^T \mathbf{c} \\
&= \sigma^2 (\mathbf{c}_0^T + d^T) (\mathbf{c}_0 + d) \\
&= \sigma^2 (a^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + d^T) (\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} a + d) \\
&= \sigma^2 a^T (\mathbf{X}^T \mathbf{X})^{-1} a + \sigma^2 d^T d \\
&= a^T \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} a + \sigma^2 d^T d \\
&= a^T \text{Var}[\hat{\beta}] a + \sigma^2 d^T d \\
&= \text{Var}[\hat{\theta}] + \sigma^2 \|d\|^2 \\
&\geq \text{Var}[\hat{\theta}].
\end{aligned} \tag{15}$$

- (b) We can show that this extends to the whole variance-covariance matrix. Letting the above  $a = I$  the identity matrix and  $d = \mathbf{D}$  any  $(p+1) \times (p+1)$  matrix, we get that

$$\text{Var}[\tilde{\beta}] = \text{Var}[\hat{\beta}] + \sigma^2 \mathbf{D}^T \mathbf{D} \tag{16}$$

Therefore,  $\text{Var}[\tilde{\beta}] - \text{Var}[\hat{\beta}] = \sigma^2 \mathbf{D}^T \mathbf{D}$  is a Gram matrix and therefore positive-semidefinite.

**Note.** Another way of stating the Gauss-Markov theorem is that the least squares estimator  $\hat{\beta}$  is *BLUE*: best linear unbiased estimator.

4. Show how the vector of least squares coefficients can be obtained from a single pass of the Gram-Schmidt procedure (Algorithm 3.1). Represent your solution in terms of the QR decomposition of  $\mathbf{X}$ .

**Solution.** After having computed the residual vectors  $\mathbf{z}_j$  using Gram-Schmidt, it is straightforward to compute the least squares coefficients, by computing

$$\hat{\beta}_j = \frac{\langle \mathbf{z}_j, \mathbf{y} \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle}.$$

In other words,  $\hat{\beta} = (\mathbf{D}\mathbf{R})^{-1} \mathbf{Z}^T \mathbf{y}$  and

$$\begin{aligned}
\hat{\beta} &= (\mathbf{D}\mathbf{R})^{-1} \mathbf{Z}^T \mathbf{y} \\
&= \mathbf{R}^{-1} \mathbf{D}^{-1} \mathbf{Z}^T \mathbf{y} \\
&= \mathbf{R}^{-1} \mathbf{Q}^T \mathbf{y}
\end{aligned} \tag{17}$$

5. Consider the ridge regression problem (3.41). Show that this problem is equivalent to the problem

$$\hat{\beta}^c = \arg \min_{\beta^c} \left\{ \sum_{i=1}^N \left[ y_i - \beta_0^c - \sum_{j=1}^p (x_{ij} - \bar{x}_j) \beta_j^c \right]^2 + \lambda \sum_{j=1}^p \beta_j^{c2} \right\}. \tag{3.85}$$

Give the correspondence between  $\beta^c$  and the original  $\beta$  in (3.41). Characterize the solution to this modified criterion. Show that a similar result holds for the lasso.

Recall that (3.41) is

$$\hat{\beta}^{\text{ridge}} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left[ y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right]^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}. \tag{3.41}$$

**Solution.** Considering for a second the case where  $p = 1$ , observe that by centering the  $x_i$ , we do not modify  $\beta_1$  since  $\beta_1$  estimates the slope and in both cases, the slope remains the same. On the other hand, affects the intercept of the line, and hence  $\beta_0$ . Moreover, assuming that the model is linear, the training data  $(y_i, x_i - \bar{x})$  would give the same model as the training data  $(y_i + \bar{y}, x_i)$ , since both data fall on the same line.

In the case of Ridge Regression, we do not attempt to constrain the intercept, hence  $\beta_0$  is free to be picked as the intercept of the line.

In the case where  $p = 2$ , we shift the dependent variables towards some line, again without affecting the normal vector of the plane, only its intercept. Similarly, this shift is equivalent to shifting all the  $y_i$  by a constant amount.

This idea generalizes to  $p$  dimensions.