Computer Age Statistical Inference: Exercises

Bradley Efron and Trevor Hastie
Stanford University

Many of these exercises use data used in the book. These datasets can be found on the book webpage `https://web.stanford.edu/~hastie/CASI`.

Chapter 1 Exercises

1. (a) Fit a cubic regression, as a function of age, to the `kidney` data of Figures 1.1 and 1.2, calculating estimates and standard errors at ages 20, 30, 40, 50, 60, 70, 80.
   (b) How do the results compare with those in Table 1.1?

2. The `lowess` curve in Figure 1.2 has a flat spot between ages 25 and 35. Discuss how one might use bootstrap replications like those in Figure 1.3 to suggest whether the flat spot is genuine or just a statistical artifact.

3. Suppose that there were no differences between AML and ALL patients for any gene, so that t in (1.6) exactly followed a student-t distribution with 70 degrees of freedom in all 7128 cases. *About* how big might you expect the largest observed t value to be? Hint: $1/7128 = 0.00014$.

4. (a) Perform 1000 nonparametric bootstrap replications of $\overline{\text{ALL}}$ (1.5). You can use program `bcanon` from the CRAN library "bootstrap" or type in the little program *Algorithm 10.1* on page 178.
   (b) Do the same for $\overline{\text{AML}}$.
   (c) Plot histograms of the results, and suggest an inference.

Chapter 2 Exercises

1. A coin with probability of heads θ is independently flipped n times, after which θ is estimated by
$$\hat{\theta} = \frac{s+1}{n+2},$$
with s equal the number of heads observed.
   (a) What are the bias and variance of $\hat{\theta}$?

(b) How would you apply the plug-in principle to get a practical estimate of $\text{se}(\hat{\theta})$?

2. Supplement Table 2.1 with entries for trimmed means, trim proportions 0.1, 0.2, 0.3, 0.4.

3. Page 14 presents two definitions of frequentism, one in terms of probabilistic accuracy and one in terms of an infinite sequence of future trials. Give a heuristic argument relating the two.

4. Suppose that in (2.15) we plugged in $\hat{\sigma}$ to get an approximate 95% normal theory hypothesis test for $H_0 : \theta = 0$. How would it compare with the student-t hypothesis test?

5. Recompute the Neyman–Pearson alpha-beta curve in Figure 2.2, now with $n = 20$. In qualitative terms, how does it compare with the $n = 10$ curve?

Chapter 3 Exercises

1. Suppose the parameter μ in the Poisson density (3.3) is known to have prior density $e^{-\mu}$. What is the posterior density of μ given x?

2. In Figure 3.1, suppose the doctor had said "$1/2, 1/2$" instead of "$1/3, 2/3$". What would be the answer to the physicist's question?

3. Let X be binomial,

$$\Pr_\pi\{X = x\} = \binom{n}{x}\pi^x(1-\pi)^{n-x} \qquad \text{for } x = 0, 1, \ldots, n.$$

What is the Fisher information $\mathcal{I}_\pi$ (3.16)? How does $\mathcal{I}_\pi$ relate to the estimate $\hat{\pi} = x/n$?

4. (a) Run the following simulation 200 times:
   - $x_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1)$ for $i = 1, 2, \ldots, 500$
   - $\mu_i = 3i/500$
   - $i_{\max}$ = index of largest x_i
   - $d = x_{i_{\max}} - \mu_{i_{\max}}$

   (b) Plot the histogram of the 200 d values.

   (c) What is the relation to Figure 3.4?

5. Give a brief nontechnical explanation of why $x_{610} = 5.29$ was likely to be an overestimate of θ_{610} in Figure 3.4.

6. Given prior density $g(\mu)$ and observation $x \sim \text{Poi}(\mu)$, you compute $g(\mu \mid x)$, the posterior density of μ given x. Later you are told that x could only be observed if it were greater than 0. (Table 6.2 presents an example of this situation.) Does this change the posterior density of μ given x?

Chapter 4 Exercises

1. (a) Verify formula (4.10).
   (b) Why isn't the formula for $\hat{\sigma}$ the one generally used in practice?

2. Draw a schematic graph of $\dot{l}_x(\theta)$ versus θ. Use it to justify (4.25).

3. You observe $x_1 \sim \text{Bin}(20, \theta)$ and, independently, $x_2 \sim \text{Poi}(10 \cdot \theta)$. Numerically compute the Cramér–Rao lower bound (4.33). Hint: Fisher information adds for independent observations.

4. A coin with unknown probability of heads θ is flipped n_1 times, yielding x_1 heads; then it is flipped another x_1 times, yielding x_2 heads.
   (a) What is an intuitively plausible estimate of θ?
   (b) What Fisherian principle have you invoked?

5. Recreate a version of Figure 4.3 based on 1000 permutations.

6. A one-parameter family of densities $f_\theta(x)$ gives an observed value x. Statistician A computes the MLE $\hat{\theta}$. Statistician B uses a flat prior density $g(\theta) = 1$ to compute $\bar{\theta}$, the Bayes posterior expectation of θ given x. Describe the relationship between the two methods.

Chapter 5 Exercises

1. Suppose $X \sim \text{Poi}(\mu)$ where μ has a $\text{Gam}(\nu, 1)$ prior (as in Table 5.1).
   (a) What is the marginal density of X?
   (b) What is the conditional density of μ given $X = x$?

2. X is said to have an "F distribution with degrees of freedom ν_1 and ν_2", denoted $F_{\nu_1,\nu_2}(x)$, if
$$X \sim \frac{\nu_2}{\nu_1} \frac{\text{Gam}(\nu_1, \sigma)}{\text{Gam}(\nu_2, \sigma)},$$
   the two gamma variates being independent. How does the F distribution relate to the beta distribution?

3. Draw a sample of 1000 bivariate normal vectors $x = (x_1, x_2)'$, with
$$x \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right).$$
   (a) Regress x_2 on x_1, and numerically check (5.18).
   (b) Do the same regressing x_1 on x_2.

4. Suppose $x \sim \mathcal{N}_p(\mu, \mathbf{\Sigma})$ as in (5.14), with $\mathbf{\Sigma}$ a *known* $p \times p$ matrix. Use (5.26) to directly calculate the information matrix $\mathcal{I}_\mu$. How does this relate to (5.27)?

5. Draw the equivalent of Figure 5.5 for $x \sim \text{Mult}_3(5, \boldsymbol{\pi})$.

6. If $x \sim \text{Mult}_L(n, \boldsymbol{\pi})$, use the Poisson trick (5.44) to approximate the mean and variance of x_1/x_2. (Here we are assuming that $n\pi_2$ is large enough to ignore the possibility $x_2 = 0$.) Hint: In notation (5.41),

$$\frac{S_1}{S_2} \doteq \frac{\mu_1}{\mu_2}\left(1 + \frac{S_1 - \mu_1}{\mu_1} - \frac{S_2 - \mu_2}{\mu_2}\right).$$

7. Show explicitly how the binomial density $\text{bi}(12, 0.3)$ is an exponential tilt of $\text{bi}(12, 0.6)$.

Chapter 6 Exercises

1. Suppose that instead of the Poisson model (6.1), we assume a binomial model

$$\Pr\{x_k = x\} = \binom{n}{x}\theta_k^x(1-\theta_k)^{n-x},$$

   n some fixed and known integer such as $n = 10$. What is the equivalent of Robbins' formula (6.5)?

2. Define $V\{\theta \mid x\}$ as the variance of θ given x. In the Poisson situation (6.1), show that

$$V\{\theta \mid x\} = E\{\theta \mid x\} \cdot \left(E\{\theta \mid x+1\} - E\{\theta \mid x\}\right),$$

   where $E\{\theta \mid x\}$ is as given in (6.5).

3. Instead of (6.8), assume $g(\theta) = (1/\sigma)e^{-\theta/\sigma}$ for $\theta > 0$.

   (a) Numerically find the maximum likelihood estimate $\hat{\sigma}$ for the Poisson model (6.1) fit to the count data in Table 6.1.

   (b) Calculate the estimates of $\widehat{E}\{\theta \mid x\}$, as in the third row of Table 6.1.

4. Suppose the `butterfly` data consisted of only the first 12 counts in Table 6.2. Recalculate Table 6.3.

5. Explain carefully why equation (6.27) is valid.

6. Let $E_1(t)$ be the number of species seen exactly once in the initial trapping period and then seen at least once in the new trapping period.

   (a) Derive the equivalent of formula (6.15).

   (b) What is the equivalent of (6.19)?

7. The `nodes` data of Section 6.3 consists of 844 pairs (n_i, x_i).
   (a) Plot x_i versus n_i.
   (b) Perform a cubic regression of x_i versus n_i and add it to the plot.
   (c) What would you expect the plot to look like if the values of n_i were assigned randomly before surgery?

Chapter 7 Exercises

1. Suppose $\mu \sim \mathcal{N}(M, A)$ and $x \mid \mu \sim \mathcal{N}(\mu, D)$, $D > 0$ known.
   (a) What is the marginal distribution of x?
   (b) What is the posterior distribution of μ given x?
2. In Table 7.1, suppose the MLE batting averages were based on 180 at-bats for each player, rather than 90. What would the JS column look like?
3. In Table 7.1, calculate the JS column based on (7.20).
4. Perform a simulation with $B = 1000$ binomial (n, P) replicates to check the accuracy of (7.21)–(7.22), using $n = 90$ and $P = 0.265$.
5. Your brother-in-law's favorite player, number 4 in Table 7.1, is batting .311 after 90 at-bats, but JS predicts only .272. He says that this is due to the lousy 17 other players, who didn't have anything to do with number 4's results and are averaging only .250. How would you answer him?
6. Verify (7.39).
7. (a) How were the columns sd(0) and sd(0.1) calculated in Table 7.3?
   (b) Calculate $\hat{\beta}(0.2)$ and sd(0.2).
8. Derive (7.43).
9. Carry out the differentiation following (7.41) to derive (7.36).
10. Derive (7.43).

Chapter 8 Exercises

1. In Figure 8.2, the numbers of mice dying in the 11 groups were $0, 0, 0, 3, 6, 6, 5, 9, 9, 10, 10$. Use the R package `glm` to calculate the red logistic regression curve. What were the regression curve values at $x = 0, 1, 2, \ldots, 10$?
2. Verify formula (8.9) for the binomial density.

3. Calculate the "deviance residuals"

$$R_{ij} = \text{sign}(p_{ij} - \hat{\pi}_{ij})\sqrt{D(p_{ij}, \hat{\pi}_{ij})}$$

($D(p_{ij}, \hat{\pi}_{ij})$ as in (8.14)), in Table 8.2. If model (8.16) fit perfectly we would expect the R_{ij}'s to follow an approximate $\mathcal{N}(0, 1)$ distribution. How well do you think the model worked?

4. Verify the Poisson deviance formula in Table 8.4.

5. The expectation of the sufficient statistic $z = \boldsymbol{X'y}$ in (8.25) is $\boldsymbol{X'\mu}(\alpha)$ according to (8.27). Use this to give an intuitive interpretation of the MLE equation (8.28).

6. (a) Fit the Poisson regression model (8.39) to the `galaxy` data, Table 8.5.
   (b) Plot the Poisson deviance residuals.
   (c) Where does the fit seem poor?
   (d) How might you add to model (8.39) to get a better fit?

7. Verify formula (8.60).

Chapter 9 Exercises

1. Formula (9.4), with $i = 1$, gives

$$S_j = \prod_{k=1}^{j}(1 - h_k)$$

as the probability of surviving past age j. How does this relate to formula (9.1)?

2. What does formula (9.17) reduce to if there is *no* censoring? (And why does this make sense?)

3. Redraw Figure 9.2, changing the "knot" location from 11 to 12.

4. Compute the equivalent of Table 9.4 for months 7 through 12.

5. Why does the hypergeometric distribution enter into formula (9.24)?

6. Derive formula (9.34).

7. Using the `bivariate normal` data for Figure 9.3, recreate Table 9.8.

8. Give a more careful version of argument (9.56)–(9.58).

Chapter 10 Exercises

1. (a) Use the jackknife to assess the standard error of $\hat{\theta} = \text{correlation}(x_i, y_i)$ for the `kidney function` data.

   (b) Examine the differences $\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}$. Do any of the observations (x_i, y_i) make particularly big contributions to $\widehat{\text{se}}_{\text{jack}}$?

2. Why is a nonparametric bootstrap sample the same as an i.i.d. sample of size n drawn from the empirical probability distribution $\widehat{F}$?

3. Use the R program `boot` in package `boot` to recreate Figure 10.2.

4. Verify formula (10.38) for the number of distinct bootstrap samples.

5. A normal theory least squares model (7.28)–(7.30) yields $\hat{\beta}$ (7.32). Describe the parametric bootstrap estimates for the standard errors of the components of $\hat{\beta}$.

6. Type in algorithm (10.1), page 178. Use it to assess the standard error of $\hat{\theta} = \text{correlation}(x_i, y_i)$ for the `kidney function` data.

7. Verify formula (10.70).

8. Suppose $n = 3$, $\boldsymbol{x} = (x_1, x_2, x_3) = (10, 2, 6)$, and $\hat{\theta} = \text{mean}(\boldsymbol{x})$. Fill in the bootstrap and jackknife values for all of the points in Figure 10.3.

9. A survey in a small town showed incomes $x_1, x_2, \ldots, x_m$ for men and $y_1, y_2, \ldots, y_n$ for women. As an estimate of the differences,

$$\hat{\theta} = \text{median}\{x_1, x_2, \ldots, x_m\} - \text{median}\{y_1, y_2, \ldots, y_n\}$$

   was computed.

   (a) How would you use nonparametric bootstrapping to assess the accuracy of $\hat{\theta}$?

   (b) Do you think your method makes full use of the bootstrap replications?

Chapter 11 Exercises

1. We observe $y \sim \lambda G_{10}$ to be $y = 20$. Here λ is an unknown parameter while G_{10} represents a gamma random variable with 10 degrees of freedom ($y \sim G(10, \lambda)$ in the notation of Table 5.1). Apply the Neyman construction as in Figure 11.1 to find the confidence limit endpoints $\hat{\lambda}(0.025)$ and $\hat{\lambda}(0.975)$.

2. (a) Say why the standard method intervals are *not* transformation invariant.

   (b) Give a clear example of your explanation.

3. Suppose $\widehat{G}$ in (11.33) was perfectly normal, say $\widehat{G} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$. What does $\hat{\theta}_{\text{BC}}(\alpha)$ reduce to in this case, and why does this make intuitive sense?

4. Show the following:

   (a) If a is zero in (11.39) then the BCa endpoints are the same as the BC endpoints (11.33).

   (b) If both z_0 and a are zero then the BC endpoints reduce to the percentile endpoints (11.18).

   (c) If z_0 and a are zero and $\widehat{G}$ is normal, $\widehat{G} \sim \mathcal{N}(\hat{\theta}, \hat{\sigma}^2)$, then the percentile endpoints reduce to the standard endpoints $\hat{\theta} \pm z^{(\alpha)}\hat{\sigma}$.

5. Suppose $\hat{\theta} \sim$ Poisson(θ) is observed to equal 16. Without employing simulation, compute the 95% central BCa interval for θ. (You can use the good approximation $z_0 = a = 1/(6\hat{\theta}^{1/2})$.)

6. Use the R program `bcajack` (available with its help file from `efron.web.stanford.edu` under "Talks") to find BCa confidence limits for the student score eigenratio statistic as in Figure 10.2.

7. Write a simulation program to find the bootstrap-t distribution for the `student score` data as in Figure 11.5. Does (11.54) seem about right?

8. One can approximate $d\alpha/d\theta$ numerically by $[\alpha(\theta+\epsilon) - \alpha(\theta-\epsilon)]/(2\epsilon)$, ϵ some small value such as 0.1. Numerically approximate the Poisson confidence density for $x = 10$ in Figure 11.6. *Note*: It may be more convenient to first approximate $d\theta/d\alpha$.

Chapter 12 Exercises

1. (a) Fit a linear model to the `supernova` data and recreate Figure 12.1.

   (b) Remove the five predictors with the smallest (in absolute value) regression coefficients, and refit the `supernova` data using just the five remaining predictors.

   (c) Compare the two fits in terms of squared error predictive power.

2. (a) Compute the cross-validated error (12.21) for the two models in problem 12.1.

   (b) Recompute the cross-validated error following the "remove five smallest" rule at each cross-validation step.

3. Give a clear explanation of why Figure 12.3 is counterintuitive, and what the last sentence of Section 12.2 means.

4. Equation (12.45) says that positive correlation between y_i and $\hat{\mu}_i$ should increase our estimate of prediction error. Why is this intuitively correct?

5. Show that the SURE formula (12.58) agrees with Mallow's estimate (12.51) in the linear case (12.47).

6. Use the bootstrap to compute the degrees of freedom as in Figure 12.4, but now for `lowess`$(\boldsymbol{x}, \boldsymbol{y}, 1/6)$.

7. Verify (12.70).

8. Give a careful argument relating Figure 12.5 to Figure 12.6.

Chapter 13 Exercises

1. A one-parameter family $f_\mu(x)$ gives MLE $\hat{\mu}$ having standard deviation $\sigma_\mu = [\mathrm{var}_\mu(\hat{\mu})]^{1/2}$. Jeffreys' prior is $g^{\mathrm{jeff}}(\mu) = 1/\sigma_\mu$. Suppose we transform coordinates to parameter λ according to the transformation
$$\frac{d\lambda}{d\mu} = \frac{1}{\sigma_\mu}.$$
What does $g^{\mathrm{jeff}}(\mu)$ transform into on the λ scale?

2. Verify (13.21) by the direct application of Bayes rule.

3. Compute the ratio of posterior densities, Jeffreys' prior over flat prior, as in the top two panels of Figure 13.2. Display the ratios as a grid of values.

4. Calculate the parametric bootstrap confidence density (11.68) for the `vasoconstriction` data, model (13.24)–(13.25).

5. One thousand independent flips of a coin yielded 563 heads and 437 tails. What is B_{BIC} for testing
$$H_0 : \text{the coin is fair?}$$
(You can use normal approximations.) Compare Jeffreys' and Fisher's assessments of H_0.

6. Suppose the i.i.d. data $x_1, x_2, \ldots, x_n$ were grouped into pairs,
$$X_1 = (x_1, x_2), X_2 = (x_3, x_4), \ldots, X_{n/2} = (x_{n-1}, x_n),$$
with the X_i's considered to be the individual data points (rather than the x_i's). How would this affect the BIC criterion (13.40)?

7. Carry out 1000 Gibbs sampling steps (13.70)–(13.71), as in the blue histogram of Figure 13.5.

8. Give an interpretation of the results in Figure 13.7.

Chapter 15 Exercises

1. Show that Holm's procedure (15.10) is more generous than Bonferroni in declaring rejections.

2. Redraw Figure 15.3 for $q = 0.2$.

3. (a) Let $S_0(z) = 1 - F_0(z)$ and $\widehat{S}(z) = \#\{z_i \geq z\}/N$. Show that (15.14) is equivalent to $\widehat{S}(z_{(i)}) \geq S_0(z_{(i)})/q$.
   (b) Give an intuitive explanation of what this says about the Benjamini–Hochberg rejection region if, say, $q = 0.1$.

4. For an observed data set of z-values $z_1, z_2, \ldots, z_N$, a case z_i of particular interest just barely made it into the Benjamini–Hochberg $\mathcal{D}_q$ rejection region. Later you find out that 25 of the very negative other z-values were actually positive, and exceed z_i. Is H_{0i} still rejected?

5. In the two-groups model (15.19), we define the "true discovery rate" as $\text{tdr}(z_0) = \Pr\{\text{case } i \text{ is non-null} \mid z_i = z_0\}$. What is the expected value of $\text{tdr}(z)$?

6. Suppose we believe $g(\mu) \sim \mathcal{N}(0.10, 0.63^2)$ for the `police` data, as suggested in *Effect size considerations*, page 288. What would this say about bias in the 2006 New York City police force?

7. The histogram in Figure 15.9 uses 49 bins, equally spaced between -4 and 4.4.
   (a) Compute the histogram counts y_i, $i = 1, 2, \ldots, 49$.
   (b) Fit a Poisson regression $\texttt{glm}(y \sim \texttt{poly}(x, 6), \texttt{Poisson})$, with x the vector of bin centers.
   (c) Compute the Poisson deviance residuals (8.41). Do you think the fit is satisfactory?
   (d) Plot the equivalent of Figure 15.6.
   (e) Apply `locfdr` and comment.

8. In searching for interesting voxels in the DTI study, what would be a simple way to compensate for the wave effect seen in Figure 15.10?

Chapter 16 Exercises

1. In forward-stepwise regression, we include the variable at each step that improves the residual-sum-of squares the most. You notice that in a software package you were using, the variable is chosen that has the maximum absolute correlation with the current residual. Are these two approaches equivalent? Explain.

2. Describe in some detail an efficient approach for computing the forward-stepwise regression model path.

3. In (16.5) on page 309, we show that the coefficient profile for the lasso path is piecewise linear. Can you use this relationship to discover at what value of $\lambda < \lambda_1$ the active set $\mathcal{A}$ changes? Explain.

4. Run a simulation to compare the df of best-subset regression and lasso. Use $p = 30$ variables and $n = 200$ observations to build an $\boldsymbol{X}$ matrix, generated from a multivariate Gaussian distribution with non-trivial covariance (of your choice). Now pose a response model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$ and specify β in advance. In your simulations hold $\boldsymbol{X}$ and β fixed, and generate new $\boldsymbol{\varepsilon}$ at each run. Make a plot similar to the right plot in Figure 16.8

5. Derive the coordinate-descent update (16.17).

Chapter 17 Exercises

1. Explain in detail why OOB error for random forests is almost identical to LOO (leave one out) error when the number B of trees is large.

2. Fit a sequence of random forests to the `spam` data, varying the parameter m from 1 to 57 (about 10 values) and using a large number $B = 5000$ trees each time. Plot the OOB error as a function of m, as well as the test error. Construct a plot that shows how the variable importance measures change with m. Make some overall conclusions from what you have learned.

3. Consider algorithm 17.3, step 2(a). Suppose we have found a tree $g(x; \gamma)$ with K terminal nodes, as in step 2(b) of algorithm 17.4. Assume the loss L corresponds to an exponential family model (e.g binomial, Poisson etc). Show how we can replace (and improve) the constants in the terminal nodes by a new set of constants by fitting a GLM with an offset and a K level factor variable as the only predictor.

4. Consider the gradient boosting algorithm 17.4. Instead of approximating the gradient with a tree, approximate it by univariate linear regression on the predictor most correlated with it, followed by shrinkage by ϵ. Outline this algorithm, and show that at each step the fit is a linear model. What will happen as the number of steps gets large?

5. Implement your algorithm in 4, and apply it to the `spam` data (use the "started" log transformation for each variable, and the binomial deviance as a loss function). Make a plot of the evolving coefficients as a function of the step number, using $\epsilon = .01$. Compare the coefficient profile to that obtained using the logistic-regression lasso.

6. When we fit a lasso to the `spam` data, we worry about the skewness of the input features, and apply a (started) log transformation. Should we do the same for a random forest? Explain.

Chapter 18 Exercises

Some of the exercises for this section require the installation of software for fitting deep neural networks. The authors have used both the `h2o.ai` software, as well as the `keras` package in `R`.

1. Fit the model as shown in Figure 18.3 to the `MNIST digit` data, using for example the `keras` package in `R`. Compute the confusion matrix on the test data for your fitted model.

2. Derive and verify equations 18.10–18.15 for a neural network with $K = 3$: an input layer, a single hidden layer, and an output layer. Assume a single output and squared-error loss.

3. Verify the assertion below (18.21) that $\hat{\mathbf{A}} = \mathbf{V}$.

4. Using a simulated data set, verify empirically that making multiple copies of the training data with added noise gives a solution close to ridge regression. Use a data matrix of size 100×5 (with non-trivial correlations), and compare with the traditional version of ridge linear regression (squared-error loss).

5. Fit a deep CNN to the CIFAR 10 image dataset using the `keras` package or similar software (data can be found at `https://www.cs.toronto.edu/~kriz/cifar.html`.) Try and get your test error below 10%, and report the confusion matrix for your network.

6. Suppose we solve a least-squares problem with $p > n$ by gradient descent: $\beta \leftarrow \beta - \epsilon \frac{\partial L(\beta)}{\partial \beta}$. We start at $\beta = 0$ and use a small stepsize ϵ.

   (a) Show that if the points are in general position, the residuals will converge to zero (modulo ϵ).

   (b) Show that the converged β corresponds to the *minimum-ℓ_2-norm* solution.

Chapter 19 Exercises

1. Consider the `leukemia` data in Figure 19.2. Suppose we code the binary response `y` as +1 and -1, and fit a ridge-regression using the 3571 gene-expression variables:

$$\min_{\alpha,\beta} \|\boldsymbol{y} - \mathbf{1}\alpha - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|^2.$$

   (a) Show that as $\lambda \downarrow 0$, $\boldsymbol{X}\hat{\beta}_\lambda \to \boldsymbol{y}$.

   (b) Show that the limiting $\hat{\beta}_0$ has minimum ℓ_2 norm among all least-squares solutions that fit the response vector $\boldsymbol{y}$ exactly.

   (c) The solution could be represented as in the left plot of Figure 19.2, except *all* the training points would be on the margin.

(d) Which solution would have the widest margin? The SVM or minimum-norm regression?

(e) Using the `leukemia` data, reproduce the left panel using this minimum-norm regression approach.

2. Show that solving criterion (19.5) is equivalent to solving (19.6).

3. Consider fitting a logistic regression problem with $p \gg n$, such as for the `leukemia` data, using a ridge penalty on β. Using the QR decomposition of $\boldsymbol{X}^T$, show that the problem can be solved instead by fitting a ridged logistic regression with n variables. Discuss what would need to be done if 10-fold cross-validation were to be used to select λ.

4. Consider fitting a kernel logistic regression to the `leukemia` data, along the lines of (19.17) using a radial kernel.

   (a) Show that the problem can be reduced to fitting a standard ridged logistic regression.

   (b) Use the package `glmnet` fit the ridge path for these data, and show the test error as a function of λ.

   (c) Repeat the above with two other values for the kernel parameter γ (see 19.11), as well as for the linear ridge regression model. Superimpose on your plot the SVM test error as a horizontal line.

5. Consider the `spam` dataset, and recode each predictor x as $x' = I(x > 0)$. Some of the recoded predictors are all 1, and these can be removed for the purposes of this exercise. Fit the SVM path using (19.6), as well as the ridge path using (19.7), and show their performance on the test data as a function of $\|\hat{\beta}\|^2$. Summarize what you see. Is plotting against $\|\hat{\beta}\|^2$ the correct basis for comparison? Propose a better way to compare their performance.

Chapter 20 Exercises

1. 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) Form a 95% confidence set for β as in (20.12), which in turn generates confidence intervals for $f(x_0)$.

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

2. Consider the `Cholersterol` data in Figure 20.1.

   (a) Use the `gam` function in package `mgcv` to fit a smooth curve. The function fits a smoothing spline, and selects the amount of smoothing automatically. Save the smoothing parameter `‘‘sp’’` from the fit. Plot the data and the fitted curve.

   (b) Run a bootstrap experiment, with $B = 300$. For each bootstrap sample, refit the GAM model with automatic selection of the smoothing parameter. For the same sample, refit the GAM with the saved value of the smoothing parameter. For each pair of fits, superimpose their curves on the original plot (in two different colors).

   Summarise your conclusions.

3. In medical applications we are often interested in the relative improvement of one treatment versus another. Suppose a proportion r_A of n_A randomly selected mice responded to treatment A. Likewise r_B of n_B randomly chosen (and different) mice responded to treatment B. We define the *log-odds ratio* as

$$L = \log \frac{r_A}{1 - r_A} - \log \frac{r_B}{1 - r_B}.$$

   Use the delta method to derive an expression for the variance of L.

4. Using the `prostate` data, write an R program to estimate the marginal density for the gene expression data. (Hint: model the log-density, using the “Poisson trick” on page 249, using a basis of natural splines). Use your fitted density to produce a version of the black conditional-mean curve in Figure 20.7.

Chapter 21 Exercises

1. This problem applies the *g-modeling* approach of Sections 21.2–21.4. Since the publication of the book, an R package `deconvolvR` has been produced by the first author and B. Narasimhan. This package is available on CRAN, and should be installed.

   (a) Estimate the prior distribution for the effect sizes for the `prostate` data, as described in Section 21.4. Use a basis of natural splines to represent the log-density for the prior. Assume a common conditional variance σ_0^2 (21.49), but give some justification for the value you used.

   (b) Use your fitted posterior distribution to estimate $\Pr(|\mu| < 0.2)$.

   (c) Compute the conditional mean curve $E(\mu|X = x)$ and compare to the one produced in Figure 20.7.

2. Use the package `deconvolveR` to fit the prior density for the `butterfly` data in Chapter 6. Help can be obtained by following the *Vignette* for the package, and the referenced paper. Describe the bias in the prior density estimate, and apply the

bias correction suggested in the vignette. Use your model to produce a version of Figure 6.2.

3. Describe how the estimation for the GLMM in Section 21.5 proceeds. Your description can consist of a series of steps, with a high level description of each step, but sufficient detail such that it could actually be implemented by a statistically savvy data scientist.