18.650. Fundamentals of Statistics
Fall 2023. Recitation sheet 3

1 Regression

Problem 1

Let (X, Z, Y) be a tuple of random variable following the model

$$Y = \beta_0^* + \beta_1^* X + \beta_2^* XZ + \varepsilon\,,$$

where ε, X, Z are independent, $\varepsilon, X \sim \mathcal{N}(0,1)$ and $Z \sim \mathcal{N}(0,2)$.

Assume that we observe n i.i.d copies $(X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n)$ of (X, Z, Y).

1. What is the regression function $f(x,z)$ of Y onto X, Z?
   Solution: The regression function is $f(x,z) = \beta_0^* + \beta_1^* x + \beta_2^* xz$.
2. Define $\vec{Y} = (Y_1, \ldots, Y_n)^\top$. Show that

   $$\vec{Y} | \{(X_1, Z_1 \ldots, X_n, Z_n)\} \sim \mathcal{N}_n(\mathbb{X}\beta^*, I_n)\,,$$

   for some design matrix $\mathbb{X}$ and some vector β^* to be made explicit.
   Solution: Let:

   $$\mathbb{X} = \begin{pmatrix} 1 & X_1 & X_1 Z_1 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n Z_n \end{pmatrix}$$

   Additionally, let $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ and β^* and $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)$. Then we have:

   $$\vec{Y} = \mathbb{X}\beta^* + \vec{\epsilon}$$

   Conditioned on the random variables, $\mathbb{X}\beta^*$ is simply a constant and $\vec{\epsilon}$ is a $\mathcal{N}_n(0, I_n)$ random variable as the $\vec{\epsilon}$ are iid. Hence, we get the desired.
3. Show that

   $$D = \lim_{n\to\infty} \frac{1}{n} \mathbb{X}^\top \mathbb{X}$$

   is a deterministic diagonal matrix to be made explicit. `[Hint: check the limit of each entry in the matrix]`.

Solution: Expanding out the matrix products we get:

$$\frac{1}{n}(\mathbb{X}^\top \mathbb{X})_{i,j} = \begin{cases} \frac{1}{n}\sum_{i=1}^n X_i & i,j = 1,2 \\ \frac{1}{n}\sum_{i=1}^n X_i Z_i & i,j = 1,3 \\ \frac{1}{n}\sum_{i=1}^n X_i^2 Z_i & i,j = 2,3 \\ \frac{1}{n}\sum_{i=1}^n 1 & i,j = 1,1 \\ \frac{1}{n}\sum_{i=1}^n X_i^2 & i,j = 2,2 \\ \frac{1}{n}\sum_{i=1}^n X_i^2 Z_i^2 & i,j = 3,3 \end{cases}$$

By the LLN, each summation will converge to the expectation of the random variable the sum is over. As each diagonal element contains a non-squared copy of X_i, Z_i and $\mathbb{E}X_i = \mathbb{E}Z_i = 0$, the expectation of those elements will converge to 0. The diagonal elements are easily evaluted using the given variances as:

$$\frac{1}{n}(\mathbb{X}^\top \mathbb{X}) \to \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

In the rest of this problem, we assume that n is large enough so that we can take

$$D = \frac{1}{n}\mathbb{X}^\top \mathbb{X}$$

4. Let $\hat{\beta}$ be the least squares estimator. What is the asymptotic distribution of $\hat{\beta}$?
Solution: Asymptotically,

$$\hat{\beta} \sim \mathcal{N}(\beta^*, (\mathbb{X}^\top \mathbb{X})^{-1}) = \mathcal{N}\left(\beta^*, \frac{1}{n}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}\right)$$

Problem 2

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random pairs, where each $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ ($p \geq 1$). Assume that there exists some $\beta \in \mathbb{R}^p$ such that, for all $i = 1, \ldots, n$, $Y_i = X_i^\top \beta + \varepsilon_i$, where ε_i is a real valued random variable that satisfies $\mathsf{cov}(X_i, \varepsilon_i) = 0$.

1. Write the above identities (for $i = 1, \ldots, n$) as one matrix identity.
Solution: We can write out the above identities as

$$Y = X\beta + \epsilon,$$

where

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, X = \begin{pmatrix} —X_1^T— \\ \dots \\ —X_n^T— \end{pmatrix} \in \mathbb{R}^{n \times p}, \text{ and } \epsilon = \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n.$$

2. Now assume that with probability one, the family of vectors $(X_1, \dots, X_n)$ has rank p. Compute the least square estimator of β.
Solution: The LSE minimizes $f(\beta) = ||Y - X\beta||^2$ with respect to $\beta \in \mathbb{R}^p$. We can take the gradient and set it equal to zero:

$$\begin{aligned} \nabla f(\beta) &= -2X^T(Y - X\beta) = 0 \\ \implies X_T Y &= X^T X\beta \\ \implies \hat{\beta}^{LSE} &= (X^T X)^{-1}(X^T Y), \end{aligned}$$

since $X^T X$ is invertible (rank of $X^T X$ is the rank of X, which is p).

3. Let f be the pdf of X_1 and assume that $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$ is independent of X_1, where $\sigma^2 > 0$.
   a) Compute the MLE $(\hat{\beta}, \hat{\sigma}^2)$ of (β, σ^2).
   Solution: The joint pdf of (X_1, Y_1) is $g(x, y) = f(x)\phi_{Y_1|X_1=x}(y)$, where ϕ is the conditional pdf of Y. Since $Y_1 = X_1^T\beta + \epsilon$, the conditional distribution of Y_1 given $X_1 = x$ is

$$\phi_{Y_1|X_1=x}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - x^T\beta)^2},$$

   so our likelihood function is

$$\mathcal{L}_n((X_1, Y_1), \dots, (X_n, Y_n)) = [\prod_{i=1}^n f(x_i)](2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i^T\beta)^2}.$$

   Thus our $\hat{\beta}$ MLE needs to maximize just the last term or equivalently minimize the exponent of the last term, which is simply $||Y - X^T\beta||^2$, and so $\hat{\beta} = \hat{\beta}^{LSE}$. Then, $\widehat{\sigma^2}$ minimizes

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{||Y - X\hat{\beta}||^2}{2\sigma^2},$$

   so $\widehat{\sigma^2} = \frac{||Y - X\hat{\beta}||^2}{n}$.
   b) Conditional on $X_1, \dots, X_n$, what is the distribution of $\hat{\beta}$?

Solution:

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1}(X^T Y) \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= \beta + (X^T X)^{-1} X^T \epsilon,\end{aligned}$$

and since $\epsilon \sim \mathcal{N}_p(0, \sigma^2 I_p)$ is independent of X, the conditional distribution of $\hat{\beta}$ is $\mathcal{N}_p(\beta, \sigma^2((X^T X)^{-1}X^T)((X^T X)^{-1}X^T)^T) = \mathcal{N}_p(\beta, \sigma^2(X^T X)^{-1})$.

4. Let $u \in \mathbb{R}^p$ be a given nonzero vector. Consider the following hypotheses:

$$H_0 : \text{“}u^\top \beta \geq 0\text{”} \text{ v.s. } H_1 : \text{“}u^\top \beta < 0\text{”}.$$

a) We assume that σ^2 is known. Propose a test that has non asymptotic level α ($\alpha \in (0,1)$), conditional on $X_1, \ldots, X_n$.
Solution: Using the previous part of the problem, conditional on X we have that $u^T \hat{\beta} \sim \mathcal{N}(u^T \beta, \sigma^2 u^T (X^T X)^{-1} u)$, so we have that

$$\frac{u^T \hat{\beta} - u^T \beta}{\sqrt{\sigma^2 u^T (X^T X)^{-1} u}} \sim \mathcal{N}(0,1).$$

Thus, we can take the test statistic

$$T = \frac{u^T \hat{\beta}}{\sqrt{\sigma^2 u^T (X^T X)^{-1} u}}$$

and rejection region

$$R = \{T < -\Phi^{-1}(1-\alpha)\}.$$

b) How would you test whether $\beta_1 \geq \beta_2$?
Solution: Take $u = \begin{pmatrix} 1 \\ -1 \\ 0 \\ \ldots \\ 0 \end{pmatrix}$.

2 Survival Analysis

Problem 3 6 people who tested positive for COVID on the same day (call it day #0) were enrolled in a study to see how long it would take for them to test negative. Each day, they took a COVID test and reported whether they had tested negative. The results were as follows:

#days since pos. test	tested neg.
5	1
7	1
8	0 (censored)
10	0 (censored)
11	1
12	1

The censored data is from people who dropped out of the study after the indicated number of days in which they continued to test positive.

1. Compute a simple estimator for the survival curve $S(t) = P(T > t)$ for each time between 0 and 12 days.
   Solution: A simple estimator which is better than throwing out all censored datapoints (though not as efficient as Kaplan-Meier) is the following:

$$\hat{S}(t) = \frac{\sum_{i=1}^{n} \mathbb{1}(\tilde{T}_i > t, C_i > t)}{\sum_{i=1}^{n} \mathbb{1}(C_i > t)}$$

   We count the proportion of observations greater than t among those which were censored at a time greater than t. This gives the survival curve shown in Figure 1.

2. Calculate the Kaplan-Meier estimator for the survival curve $S(t) = P(T > t)$ for each time between 0 and 12 days. **Solution:** Remember that

$$\hat{S}(t) = \prod_{s=0}^{t} \left(1 - \frac{\#\{i : \tilde{T}_i = s, C_i > s\}}{\#\{i : \tilde{T}_i \geq s\}}\right).$$

   This gives the survival curve shown in Figure 2.

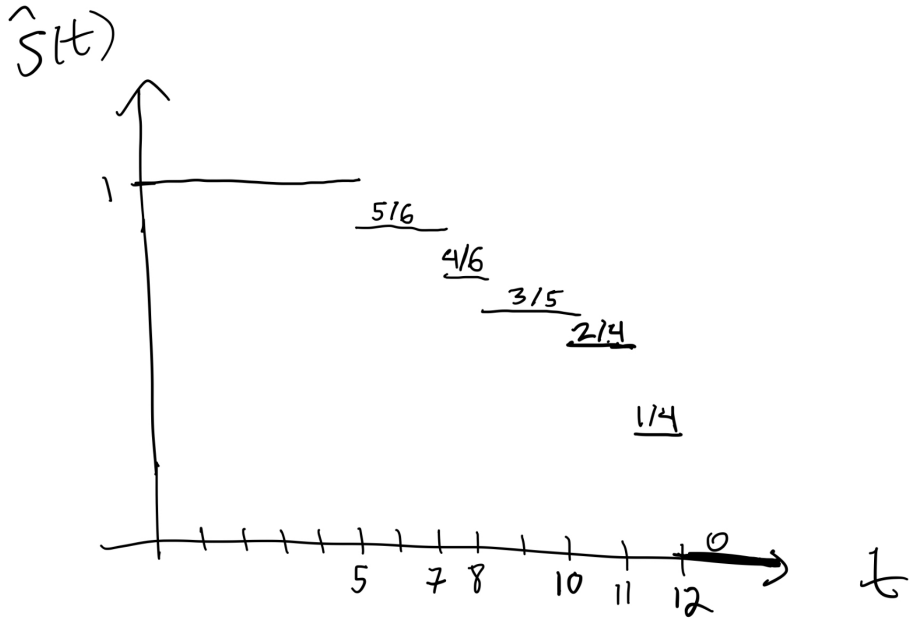

Figure 1: Answer to Problem 3, part 1

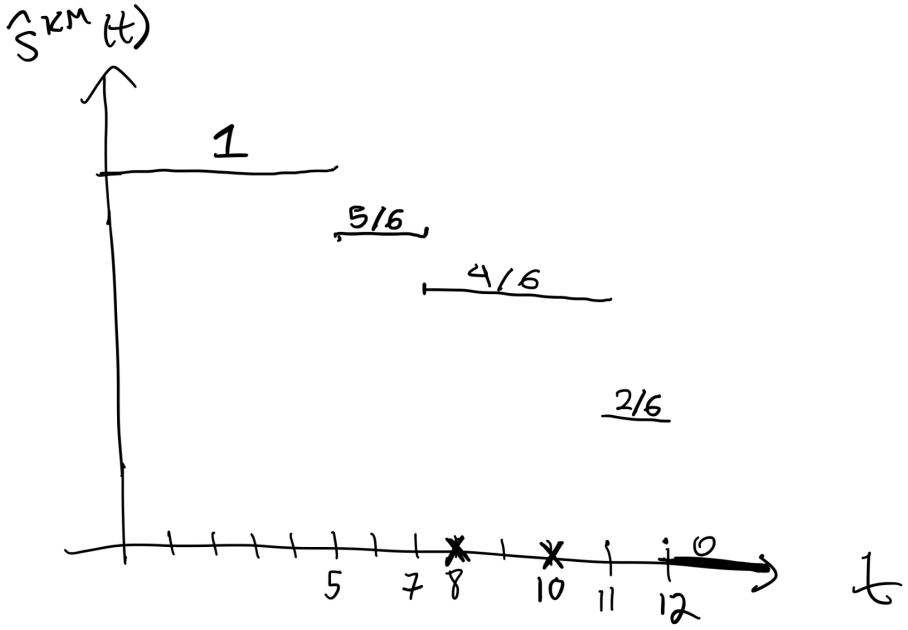

Figure 2: Answer to Problem 3, part 2

3 Causal Inference

Problem 4

Suppose that $U = Unif[0,1]$. Let $X \in \{0,1\}$ be a binary treatment and let (C_0, C_1) denote the corresponding potential outcomes. Consider an experiment where patients are assigned to the treatment group $X = 1$ if $U \geq 1/2$ and to the control otherwise. The potential outcomes are given by

$$C_1 = Ber(U)$$
$$C_0 = Ber(U).$$

Compute the average treatment effect θ and the association α. Comment on your result.

Solution: We have

$$\mathbb{E}C_1 = \mathbb{E}\mathbb{P}(C_1 = 1|U) = \mathbb{E}U = \frac{1}{2}$$
$$\mathbb{E}C_0 = \mathbb{E}\mathbb{P}(C_0 = 1|U) = \mathbb{E}U = \frac{1}{2}.$$

This gives an average treatment effect of $\theta = 0$. (The calculation was not necessary since clearly C_1 and C_0 have the same distribution, but it's good practice.) For the association we have

$$\begin{aligned}\mathbb{E}[Y|X=1] &= \mathbb{E}[C_1|X=1] \\ &= \mathbb{E}[C_1|U \geq 1/2] \\ &= 2\int_{1/2}^{1} \mathbb{P}(C_1 = 1|U=u)du \\ &= 2\int_{1/2}^{1} udu \\ &= \frac{3}{4}.\end{aligned}$$

Similarly we have

$$\begin{aligned}\mathbb{E}[Y|X=0] &= \mathbb{E}[C_0|X=0] \\ &= \mathbb{E}[C_0|U < 1/2] \\ &= 2\int_{0}^{1/2} \mathbb{P}(C_0 = 1|U=u)du \\ &= 2\int_{0}^{1/2} udu = \frac{1}{4}.\end{aligned}$$

Therefore, $\alpha = \frac{3}{4} - \frac{1}{4} = \frac{1}{2} \neq \theta$. Despite the fact that the treatments were selected randomly we see that association does not equal average causal effect. This is because the distribution of the outcome was not independent from how the treatments were randomly selected.

Problem 5 Suppose you are given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from an observational study, where $X_i \in \{0, 1\}$ and $Y_i \in \{0, 1\}$. Although it is not possible to estimate the causal effect θ, it is possible to put bounds on θ. Find upper and lower bounds on θ that can be consistently estimated from the data. Show that the bounds have width 1.
Hint: note that $\mathbb{E}[C_1] = \mathbb{E}[C_1|X = 1]\mathbb{P}(X = 1) + \mathbb{E}[C_1|X = 0]\mathbb{P}(X = 0)$. Which of these quantities can you estimate from the data? **Solution:**

$$\begin{aligned}\mathbb{E}[C_1] - \mathbb{E}[C_0] =& \mathbb{E}[C_1|X = 1]P(X = 1) + \mathbb{E}[C_1|X = 0]P(X = 0) \\ &- \mathbb{E}[C_0|X = 1]P(X = 1) - \mathbb{E}[C_0|X = 0]P(X = 0) \\ =& \mathbb{E}[C_1|X = 1]P(X = 1) - \mathbb{E}[C_0|X = 0]P(X = 0) \\ &+ \mathbb{E}[C_1|X = 0]P(X = 0) - \mathbb{E}[C_0|X = 1]P(X = 1)\end{aligned} \tag{1}$$

Note: $\mathbb{E}[C_1|X = 1] = \mathbb{E}[C_X|X = 1] = \mathbb{E}[Y|X = 1]$ and $\mathbb{E}[C_0|X = 0] = \mathbb{E}[C_X|X = 0] = \mathbb{E}[Y|X = 1]$. Therefore, $\mathbb{E}[C_1|X = 1]P(X = 1) = \mathbb{E}[Y|X = 1]P(X = 1) = P(Y = 1, X = 1)$, since C_1 is Bernoulli. Similarly, $\mathbb{E}[C_0|X = 0]P(X = 0) = \mathbb{E}[Y|X = 0]P(X = 0) = P(Y = 0, X = 0)$. Both of these quantities can be estimated from the data. For $\mathbb{E}[C_0|X = 1]$ and $\mathbb{E}[C_1|X = 0]$, all we know is that they are between 0 and 1. Therefore we get

$$\begin{aligned}-P(X = 1) + [P(Y = 1, X = 1) - P(Y = 0, X = 0)] \\ \leq \mathbb{E}[C_1] - \mathbb{E}[C_0] \\ \leq P(X = 0) + [P(Y = 1, X = 1) - P(Y = 0, X = 0)]\end{aligned} \tag{2}$$

All quantities in this inequality can be estimated from the data, and the upper bound minus the lower bound is $P(X = 0) + P(X = 1) = 1$.

4 Nonparametric Curve Estimation

Problem 6 Let n, m and K be three positive integers such that $n = Km$. Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be independent such that $x_i = (i - 1)/n$ and

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

for some ε_i that are i.i.d $N(0, 1)$. Here f is the unknown regression function of interest.

1. Recall the definition of the regressogram $\hat{f}$ with m bins.
   Solution: Set $B_i = [\frac{i}{m}, \frac{i+1}{m})$. For $x \in B_l$, the regressogram is given by:

$$\hat{f}(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in B_l)}{\sum_{i=1}^n \mathbb{1}(X_i \in B_l)}$$

2. How many x_is fall into the first bin $B_1 = [0, \frac{1}{m})$?
   Solution: We have $x_1, \ldots, x_K \in B_1$ so the answer is K. This result also holds for each bin.
3. For $x \in B_1$, what is the distribution of $\hat{f}(x)$? Compute the bias $b(x)$ and the variance $v(x)$ of $\hat{f}(x)$.
   Solution: By the previous part, the regressogram simplifies to:

$$\hat{f}(x) = \frac{1}{K}\sum_{i=1}^K Y_i = \frac{1}{K}\sum_{i=1}^K f(x_i) + \frac{1}{K}\sum_{i=1}^K \epsilon_i$$

   The first summation is a constant, and the second is an average of K standard normals. Thus $v(x) = \frac{1}{K}$,

$$b(x) = \left| f(x) - \frac{1}{K}\sum_{i=1}^K f(x_i) \right|$$

   and

$$\hat{f}(x) \sim \mathcal{N}\left(\frac{1}{K}\sum_{i=1}^K f(x_i), \frac{1}{K}\right).$$

   Assume now f is 1-Lipschitz and $K \geq 3$.
4. Show that the bias is upper bounded as

$$|b(x)| \leq \frac{K}{n}$$

   `[Hint: if` x `and` x_i `are in the same bin then` $|x - x_i| \leq 1/m$`.]`
   Solution: By the 1-Lipshitz property:

$$|f(x) - f(x_i)| \leq |x - x_i| \leq \frac{1}{m}$$

   whenever $x, x_i \in B_1$. By the triangle inequality then:

$$b(x) \leq \frac{1}{K}\sum_{i=1}^K |f(x) - f(x_i)| \leq \frac{1}{m} = \frac{K}{n}.$$

5. Give a choice of K such that

$$\mathsf{MISE}(\hat{f}) \leq \frac{2}{n^{2/3}}$$

`[Tip: Don't bother with rounding; when optimizing over` K`, simply assume that the optimizer is an integer. ]`

Solution: Note that the logic from the previous parts applies to all bins not just B_1. Taking $K = n^{-2/3}$, we get:

$$\mathsf{MISE}(\hat{f}) = \int_0^1 (b(x)^2 + v(x))dx \leq \int_0^1 \left(\frac{K^2}{n^2} + \frac{1}{K}\right) dx = \frac{2}{n^{2/3}}$$

5 Test practice: Bayes + linear regression

Problem 7 Bayesian estimation and linear regression Let $X_1, \ldots, X_n$ be n deterministic vectors in $\mathbb{R}^p$ and let $\mathbb{X}$ be the $n \times p$ matrix whose rows are $X_1^\top, \ldots, X_n^\top$. Suppose

$$Y_i = X_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

where σ^2 is a given positive number. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

1. What is the distribution of $\boldsymbol{Y}$?
2. Assume that the prior distribution of $\boldsymbol{\beta}$ is $\mathcal{N}_p(0, \tau^2 I_p)$, where τ^2 is a fixed positive number.
   a) Prove that the posterior distribution of $\boldsymbol{\beta}$ (i.e., the distribution of $\boldsymbol{\beta}$ conditional on $\boldsymbol{Y}$) has a density proportional to $\exp\left(-\frac{1}{2\sigma^2}\left(\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2\right)\right)$, for some λ to be determined.
   b) Conclude that the posterior distribution of $\boldsymbol{\beta}$ is Gaussian and determine its parameters.
   c) What is the posterior mean of $\boldsymbol{\beta}$?
3. Consider a frequentist approach: Assume that the linear regression of Y_i on X_i is $Y_i = X_i^\top \boldsymbol{\beta} + \varepsilon_i$, for some unknown vector $\boldsymbol{\beta} \in \mathbb{R}^p$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, for some $\sigma^2 > 0$. The *Ridge* estimator of $\boldsymbol{\beta}$ is defined as the minimizer of a penalized version of the sum of squared errors, namely, it solves

$$\min_{\boldsymbol{b} \in \mathbb{R}^p} \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|^2,$$

where λ is a *tuning parameter*, i.e., a given number chosen by the statistician.

a) Compute the Ridge estimator.
b) Using the previous questions, prove that there exists a value of τ^2 such that $\hat{\boldsymbol{\beta}}^{(R)}$ is equal to the Bayesian estimator of $\boldsymbol{\beta}$ using the prior distribution $\mathcal{N}_p(0, \tau^2 I_p)$.
c) What is the distribution of $\hat{\boldsymbol{\beta}}^{(R)}$?

6 Classification

Problem 8 Consider a classification problem where the goal is to predict whether or not a patient will live through a certain disease, using some features.

Let $Y = 1$ indicate that the patient survived and $Y = 0$ indicate that the patient did not survive. Our main predictor is the number X, the systolic blood pressure of the patient. Assume that the regression function of Y onto X is given by

$$f(x) = \frac{8000}{x^2}.$$

Furthermore, assume that $X \sim \mathsf{Unif}[100, 150]$.

1. What is the distribution of Y given $X = 120$?
   Solution: Note that Y is binary (i.e. 0 versus 1), and so it follows a Bernoulli distribution with parameter
   $$\mathbb{E}[Y|X = 120] = f(120) = \frac{5}{9}.$$

2. Compute the Bayes classifier $h^*(x)$.
   Solution: Note that $f(x) = \frac{1}{2}$ when $x = \sqrt{16000} \approx 126.491$. The Bayes classifier is 1 when $f(x) > \frac{1}{2}$ and 0 otherwise giving:
   $$h^*(x) = \begin{cases} 1 & x \leq \sqrt{16000} \\ 0 & x > \sqrt{16000} \end{cases}$$

3. Compute the true error rate of the Bayes classification rule.

Solution: We have:

$$\begin{aligned}
\mathcal{L}(h^*) &= \mathbb{P}(h^*(X) \neq Y) \\
&= \frac{1}{50}\int_{100}^{150} P(h^*(x) \neq Y | X = x)dx \\
&= \frac{1}{50}\int_{100}^{\sqrt{16000}} \left(1 - \frac{8000}{x^2}\right)dx + \frac{1}{50}\int_{\sqrt{16000}}^{150} \frac{8000}{x^2}dx \\
&= \frac{1}{50}\left((60\sqrt{10} - 180) + \left(20\sqrt{10} - \frac{160}{3}\right)\right) \\
&= \frac{24\sqrt{10} - 70}{15} \approx .393
\end{aligned}$$

4. Compute the density of $X|Y = 1$.
Solution: By Bayes rule:

$$p(x|Y = 1) = \frac{\mathbb{P}(Y = 1|X = x)p(x)}{\int_{100}^{150} \mathbb{P}(Y = 1|X = x)p(x)dx} = \frac{\frac{8000}{x^2}}{\int_{100}^{150} \frac{8000}{x^2}dx} = \frac{300}{x^2}$$

where $p(x)$ denotes the density of X.

Problem 9

Consider the classification problem where we sample random variables (X, Y) such that Y takes values in $\{0, 1\}$ and and for each $y \in \mathcal{Y}$ the random variable $X|Y = y$ has a normal distribution. The parameters for these normal distribution may change with y.

1. Recall the definition of regression function, Bayes classifier, decision boundary.
Solution: The regression function is given by:

$$r(x) = \mathbb{E}[Y|X = x].$$

The Bayes classifier is given by:

$$h^*(x) = \mathbb{1}\left(r(x) > \frac{1}{2}\right).$$

The decision boundary is the set of points where $r(x) = \frac{1}{2}$.

2. Suppose $X|Y = 0 \sim \mathcal{N}(0, 1)$, and $X|Y = 1 \sim \mathcal{N}(1, 4)$. Without using theorem 22.7 from AoS, determine the Bayes classifier when $P(Y = 0) = \frac{1}{3}$.
Solution: The Bayes classifier equals 1 if:

$$P(Y = 1|X = x) > P(Y = 0|X = x)$$

By Bayes' rule this is equivalent to:

$$f(x|Y=1)P(Y=1) > f(x|Y=0)P(Y=0)$$

where $f(x|Y=k)$ is the distribution of $x|Y=k$. This is equivalent to:

$$\frac{1}{2\sqrt{2\pi}}\exp(-(x-1)^2/8)\cdot\frac{2}{3} > \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)\cdot\frac{1}{3}.$$

The constants cancel so it suffices to compare $-\frac{(x-1)^2}{8}$ to $-\frac{x^2}{2}$. From this we get:

$$h^*(x) = \begin{cases} 0 & -1 \le x \le \frac{1}{3} \\ 1 & \text{else} \end{cases}$$

3. Suppose X is a 2 dimensional Gaussian and let $h(x)$ be the optimal classification rule. Construct examples where the decision boundary is (a) a line, (b) a circle, (c) an ellipse, (d) a parabola. You may use theorem 22.7 here.
Solution: Let $X|Y=k \sim N(\mu_k, \Sigma_k)$ and suppose $P(Y=0) = P(Y=1)$. The decision boundary is a line if $\Sigma_0 = \Sigma_1$. For (b) we can set $\mu_0 = \mu_1$, $\Sigma_0 = I_2$, and $\Sigma_1 = 2I_2$ where I_2 is the 2×2 identity matrix. For (c) we can use the same setup just we replace $\Sigma_1 = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ where $a, b > 1$ and $a \neq b$. For (d), we again use the same setup just we set $a = 1$ and $b > 1$. For all of these, one just needs to match (22.10) in AoS to the proper quadratic equation.
4. Give an examples of a region that cannot be the decision boundary.
Solution: Gaussians give quadratic decision boundaries, so cubic curves like $y = x^3$ cannot be the decision boundary.