Practice questions

Statistical Thinking

14/11/2021

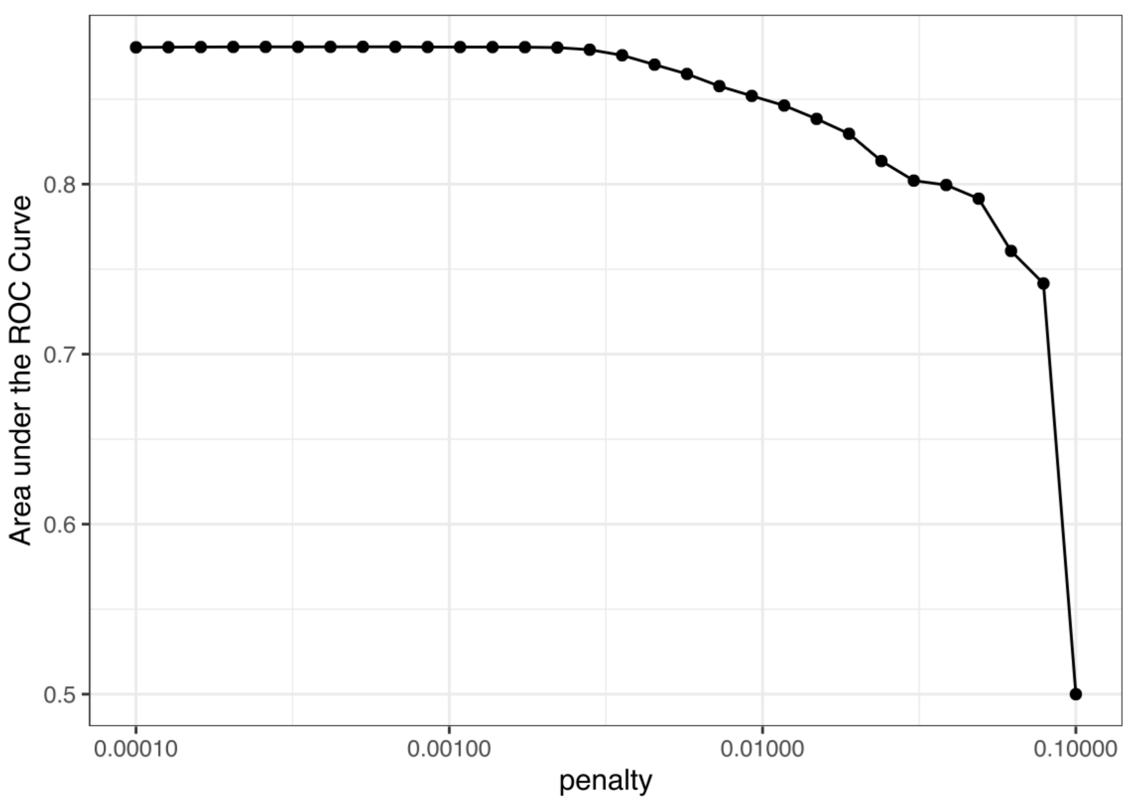
Question 1

After fittling a linear regression using the formula `y ~ x`, you compute the reisdual and plot the residuals on the y-axis and the covariate `x` on the x-axis. This plot shows a pronouned U shape. Sketch a dataset that would lead to this diagnostic plot.

Question 2

You are interesed in producing a confidence interval for the kurtosis[1] of a sample. Your friend Margaret gives you a procedure for computing $L(y)$ and $U(y)$ but she can't remember what the type-1 error is. Describe in words, using bullet point, the procedure to compute the type-1 error of the proposed confidence interval.

Question 3



[1] a measure of how heavy the tails of a distribution are

The above graph shows the AUC (bigger is better) for various values of a tuning parameter. If you know that a larger penalty produces a less complex model, choose and justify an appropriate penalty parameter for this problem.

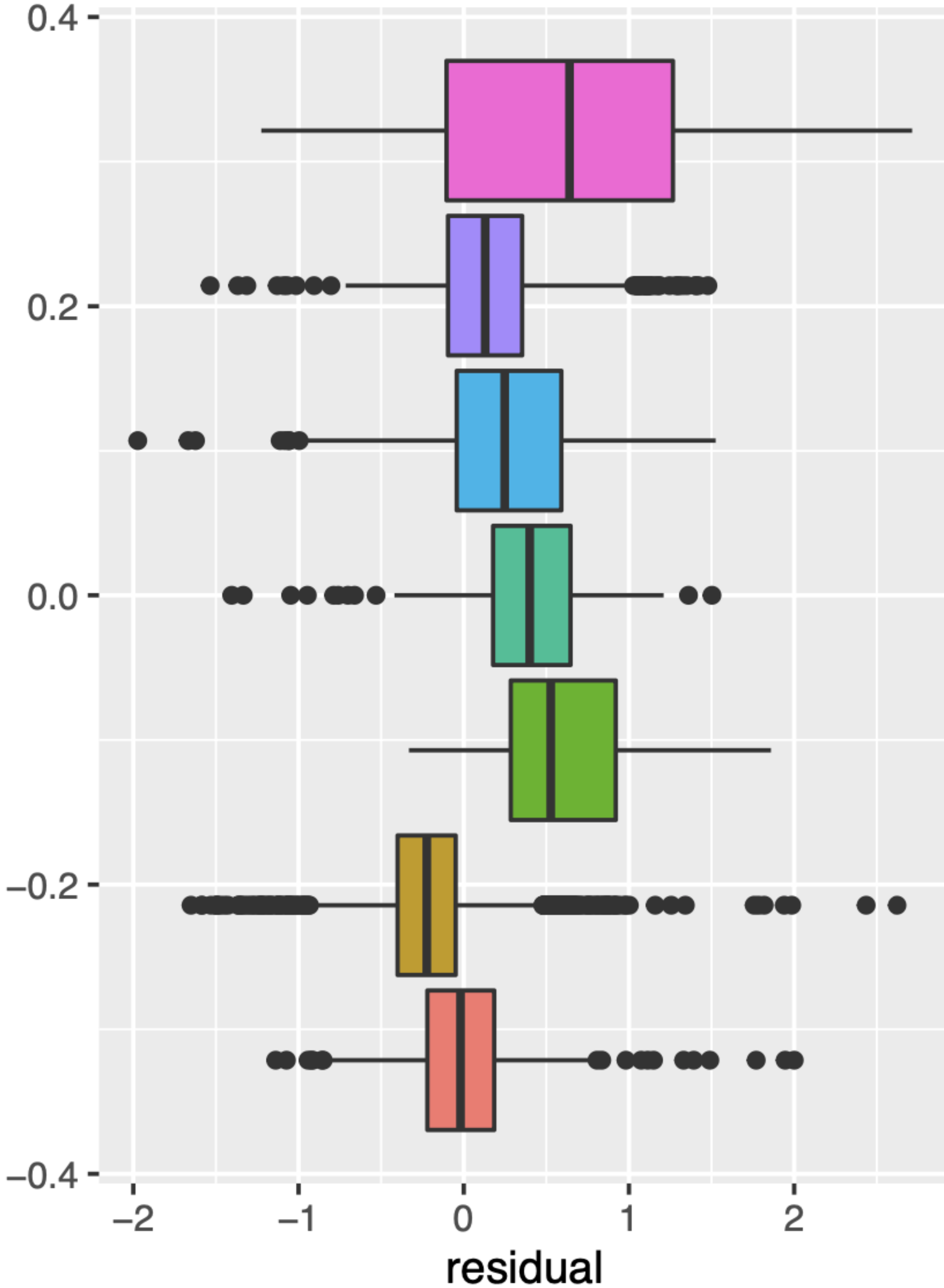
Question 4

Consider a regression problem where you are trying to estimate the causal effect of X on Y in the presence of other variables A, B, C, and D. The appropriate regression for estimating this causal effect is `Y ~ X + A + B`. Draw a DAG that includes at least one each of forks, pipes, and colliders that is consistent with the stated regression estimating the causal effect of X on Y.

Question 5

Draw a data set where points in (x_1, x_2)-space are labelled as either $y = 0$ or $y = 1$, where logistic regression using the formula `y ~ x1 + x2` would fail to yield a good classifier.

Question 6



The above boxplot shows the residuals for the linear regression `y ~ x + w` grouped by a categorical variable

z. How could you improve this regression?

Question 7

Consider the data in the image below.

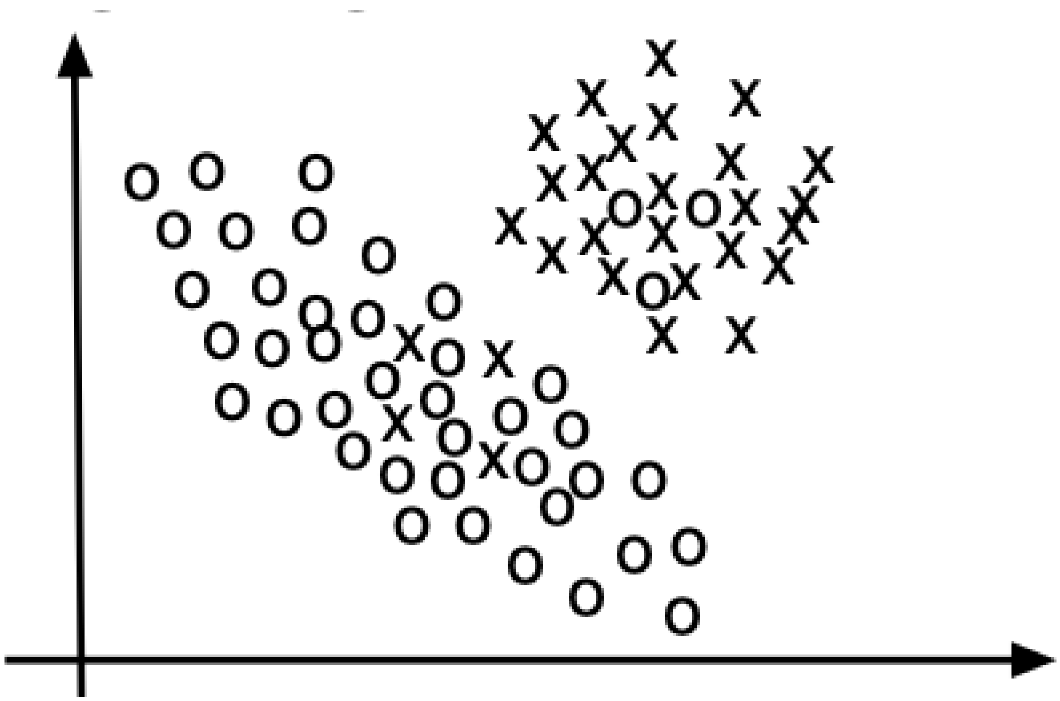


Figure 1: A data set

Which would you expect to have the smallest training error:

- 1-nearest neighbour classification
- 5-nearest neighbour classification
- logistic regression

Question 7

Let p denote the probability that a tossed coin will return a *Head* outcome.

Suppose your friend Ella gives you a coin, and tells you that she is 70% sure it is a 'fair' coin. But, she thinks it is also possible that a $Head$ outcome could occur with probability $p = 0.6$, but that no other values of the $Head$ outcome probability are possible.

You decide to give the coin from Ella to your friend, Wei. You tell Wei that you do not know if it is a 'fair' coin, but you neglect to tell him anything about your belief (or Ella's) regarding the likely values of p, nor do you tell him anything about your previous coin toss.

Knowing only that the coin may not be 'fair', Wei does not consider any individual value of $p \in (0, 1)$ more likely than any other possible value. Wanting to update his belief regarding the value of p, Wei decides to run his own experiment comprised of ten independent tosses the coin. He observes 3 *Head* outcomes.

Part A (TRUE/FALSE)

Ella is acting like a Bayesian by suggesting probabilities for certain possible values of the unknown parameter, p.

Part B (TRUE/FALSE)

Given n independent tosses of the coin, the likelihood function associated with the unknown parameter p is equal to the probability (mass) function associated with the distribution for the number of Head outcomes, viewed as a function of the unknown parameter p.

Part C (Multichoice)

Given n independent tosses of the coin, the probability distribution associated with a Head outcome, for a given value of p, is...

a. $Normal(\mu, \sigma^2)$

b. t_ν

c. $Binomial(n, p)$

d. $Beta(\alpha, \beta)$

e. $Gamma(\alpha, \beta)$

f. None of the above

Part D (Multichoice)

Ella's prior distribuiton for the possible values of p can be represented as...

a. $Beta(\alpha = 1,\ \beta = 1)$

b. $\Pr(p = 0.5) = 0.7$ and $\Pr(p = 0.6) = 0.3$

c. $\Pr(p = 0.7) = 0.3$ and $\Pr(p = 0.6) = 0.4$

d. $Beta(\alpha = 0.55,\ \beta = 0.45)$

e. None of the above.

Part E (Short answer)

Report both Wei's *prior distribution* and his *posterior distribution*, and briefly explain how the two distributions are related to each other.

Part F (Short answer)

Using only the outcomes from Wei's ten coin tosses, briefly explain how a Frequentist would attempt to determine the value of p.