SOCIETY OF ACTUARIES

EXAM SRM - STATISTICS FOR RISK MODELING

EXAM SRM SAMPLE QUESTIONS AND SOLUTIONS

These questions and solutions are representative of the types of questions that might be asked of candidates sitting for Exam SRM. These questions are intended to represent the depth of understanding required of candidates. The distribution of questions by topic is not intended to represent the distribution of questions on future exams.

April 2019 update: Question 2, answer III was changed. While the statement was true, it was not directly supported by the required readings.
July 2019 update: Questions 29-32 were added.
September 2019 update: Questions 33-44 were added.
December 2019 update: Question 40 item I was modified.
January 2020 update: Question 41 item I was modified.
November 2020 update: Questions 6 and 14 were modified and Questions 45-48 were added.
February 2021 update: Questions 49-53 were added.
July 2021 update: Questions 54-55 were added.
October 2021 update: Questions 56-57 added.
February 2022 update: Questions 58-60 were added. Question 59 corrected in July 2022.
June 2022 update: Questions 61-63 were added.
September 2022 update: Question 64 was added.
August 2023 update: Questions 66-67 were added, Questions 17, 28, 47, and 65 were deleted (no longer on the syllabus)
June 2024 update: Questions 68-70 were added.
July 2024 update: Questions 71-72 were added

QUESTIONS

1. You are given the following four pairs of observations:

$$x_1 = (-1, 0), \quad x_2 = (1, 1), \quad x_3 = (2, -1), \text{ and } x_4 = (5, 10).$$

A hierarchical clustering algorithm is used with complete linkage and Euclidean distance.

Calculate the intercluster dissimilarity between $\{x_1, x_2\}$ and $\{x_4\}$.

(A) 2.2
(B) 3.2
(C) 9.9
(D) 10.8
(E) 11.7

2. Determine which of the following statements is/are true.

I. The number of clusters must be pre-specified for both *K*-means and hierarchical clustering.

II. The *K*-means clustering algorithm is less sensitive to the presence of outliers than the hierarchical clustering algorithm.

III. The *K*-means clustering algorithm requires random assignments while the hierarchical clustering algorithm does not.

(A) I only
(B) II only
(C) III only
(D) I, II and II
(E) The correct answer is not given by (A), (B), (C), or (D)

3. You are given:

i) The random walk model

$$y_t = y_0 + c_1 + c_2 + \cdots + c_t$$

where $c_t, t = 0,1,2,\ldots,T$ denote observations from a white noise process.

ii) The following nine observed values of c_t:

t	11	12	13	14	15	16	17	18	19
c_t	2	3	5	3	4	2	4	1	2

iii) The average value of $c_1, c_2, \ldots, c_{10}$ is 2.

iv) The 9 step ahead forecast of y_{19}, $\hat{y}_{19}$, is estimated based on the observed value of y_{10}.

Calculate the forecast error, $y_{19} - \hat{y}_{19}$.

(A) 1

(B) 2

(C) 3

(D) 8

(E) 18

4. You are given:

i) The random walk model

$$y_t = y_0 + c_1 + c_2 + \cdots + c_t$$

where $c_t, t = 0,1,2,\ldots T$ denote observations from a white noise process.

ii) The following ten observed values of y_t.

t	1	2	3	4	5	6	7	8	9	10
y_t	2	5	10	13	18	20	24	25	27	30

iii) $y_0 = 0$

Calculate the standard error of the 9 step-ahead forecast, $\hat{y}_{19}$.

(A) 4/3

(B) 4

(C) 9

(D) 12

(E) 16

5. Consider the following statements:

I. Principal Component Analysis (PCA) provide low-dimensional linear surfaces that are closest to the observations.

II. The first principal component is the line in p-dimensional space that is closest to the observations.

III. PCA finds a low dimension representation of a dataset that contains as much variation as possible.

IV. PCA serves as a tool for data visualization.

Determine which of the statements are correct.

(A) Statements I, II, and III only
(B) Statements I, II, and IV only
(C) Statements I, III, and IV only
(D) Statements II, III, and IV only
(E) Statements I, II, III, and IV are all correct

6. Consider the following statements:

I. The proportion of variance explained by an additional principal component never decreases as more principal components are added.

II. The cumulative proportion of variance explained never decreases as more principal components are added.

III. Using all possible principal components provides the best understanding of the data.

IV. A scree plot provides a method for determining the number of principal components to use.

Determine which of the statements are correct.

(A) Statements I and II only
(B) Statements I and III only
(C) Statements I and IV only
(D) Statements II and III only
(E) Statements II and IV only

7. Determine which of the following pairs of distribution and link function is the most appropriate to model if a person is hospitalized or not.

(A) Normal distribution, identity link function
(B) Normal distribution, logit link function
(C) Binomial distribution, linear link function
(D) Binomial distribution, logit link function
(E) It cannot be determined from the information given.

8. Determine which of the following statements describe the advantages of using an alternative fitting procedure, such as subset selection and shrinkage, instead of least squares.

I. Doing so will likely result in a simpler model
II. Doing so will likely improve prediction accuracy
III. The results are likely to be easier to interpret

(A) I only
(B) II only
(C) III only
(D) I, II, and III
(E) The correct answer is not given by (A), (B), (C), or (D)

9. A classification tree is being constructed to predict if an insurance policy will lapse. A random sample of 100 policies contains 30 that lapsed. You are considering two splits:

Split 1: One node has 20 observations with 12 lapses and one node has 80 observations with 18 lapses.

Split 2: One node has 10 observations with 8 lapses and one node has 90 observations with 22 lapses.

The total Gini index after a split is the weighted average of the Gini index at each node, with the weights proportional to the number of observations in each node.

The total entropy after a split is the weighted average of the entropy at each node, with the weights proportional to the number of observations in each node.

Determine which of the following statements is/are true?

I. Split 1 is preferred based on the total Gini index.

II. Split 1 is preferred based on the total entropy.

III. Split 1 is preferred based on having fewer classification errors.

(A) I only

(B) II only

(C) III only

(D) I, II, and III

(E) The correct answer is not given by (A), (B), (C), or (D).

10. Determine which of the following statements about random forests is/are true?

I. If the number of predictors used at each split is equal to the total number of available predictors, the result is the same as using bagging.

II. When building a specific tree, the same subset of predictor variables is used at each split.

III. Random forests are an improvement over bagging because the trees are decorrelated.

(A) None

(B) I and II only

(C) I and III only

(D) II and III only

(E) The correct answer is not given by (A), (B), (C), or (D).

11. You are given the following results from a regression model.

Observation number (i)	y_i	$\hat{f}(x_i)$
1	2	4
2	5	3
3	6	9
4	8	3
5	4	6

Calculate the sum of squared errors (SSE).

(A) –35
(B) –5
(C) 5
(D) 35
(E) 46

12. Determine which of the following statements is true

(A) Linear regression is a flexible approach
(B) Lasso is more flexible than a linear regression approach
(C) Bagging is a low flexibility approach
(D) There are methods that have high flexibility and are also easy to interpret
(E) None of (A), (B), (C), or (D) are true

13. Determine which of the following statements is/are true for a simple linear relationship, $y = \beta_0 + \beta_1 x + \varepsilon$.

I. If $\varepsilon = 0$, the 95% confidence interval is equal to the 95% prediction interval.

II. The prediction interval is always at least as wide as the confidence interval.

III. The prediction interval quantifies the possible range for $E(y \mid x)$.

(A) I only

(B) II only

(C) III only

(D) I, II, and III

(E) The correct answer is not given by (A), (B), (C), or (D).

14. From an investigation of the residuals of fitting a linear regression by ordinary least squares it is clear that the spread of the residuals increases as the predicted values increase. Observed values of the dependent variable range from 0 to 100.

Determine which of the following statements is/are true with regard to transforming the dependent variable to make the variance of the residuals more constant.

I. Taking the logarithm of one plus the value of the dependent variable may make the variance of the residuals more constant.

II. A square root transformation may make the variance of the residuals more constant.

III. A logit transformation may make the variance of the residuals more constant.

(A) None

(B) I and II only

(C) I and III only

(D) II and III only

(E) The correct answer is not given by (A), (B), (C), or (D).

15. You are performing a *K*-means clustering algorithm on a set of data. The data has been initialized randomly with 3 clusters as follows:

Cluster	Data Point
A	(2, –1)
A	(–1, 2)
A	(–2, 1)
A	(1, 2)
B	(4, 0)
B	(4, –1)
B	(0, –2)
B	(0, –5)
C	(–1, 0)
C	(3, 8)
C	(–2, 0)
C	(0, 0)

A single iteration of the algorithm is performed using the Euclidian distance between points and the cluster containing the fewest number of data points is identified.

Calculate the number of data points in this cluster.

(A) 0
(B) 1
(C) 2
(D) 3
(E) 4

16. Determine which of the following statements is applicable to *K*-means clustering and is not applicable to hierarchical clustering.

(A) If two different people are given the same data and perform one iteration of the algorithm, their results at that point will be the same.

(B) At each iteration of the algorithm, the number of clusters will be greater than the number of clusters in the previous iteration of the algorithm.

(C) The algorithm needs to be run only once, regardless of how many clusters are ultimately decided to use.

(D) The algorithm must be initialized with an assignment of the data points to a cluster.

(E) None of (A), (B), (C), or (D) meet the meet the stated criterion.

17. DELETED

18. For a simple linear regression model the sum of squares of the residuals is $\sum_{i=1}^{25} e_i^2 = 230$ and the R^2 statistic is 0.64.

Calculate the total sum of squares (TSS) for this model.

(A) 605.94

(B) 638.89

(C) 690.77

(D) 701.59

(E) 750.87

19. The regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$ is being investigated.

The following maximized log-likelihoods are obtained:

- Using only the intercept term: –1126.91
- Using only the intercept term, X_1, and X_2: –1122.41
- Using all four terms: –1121.91

The null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is being tested at the 5% significance level using the likelihood ratio test.

Determine which of the following is true.

(A) The test statistic is equal to 1 and the hypothesis cannot be rejected.
(B) The test statistic is equal to 9 and the hypothesis cannot be rejected
(C) The test statistic is equal to 10 and the hypothesis cannot be rejected.
(D) The test statistic is equal to 9 and the hypothesis should be rejected.
(E) The test statistic is equal to 10 and the hypothesis should be rejected.

20. An analyst is modeling the probability of a certain phenomenon occurring. The analyst has observed that the simple linear model currently in use results in predicted values less than zero and greater than one.

Determine which of the following is the most appropriate way to address this issue.

(A) Limit the data to observations that are expected to result in predicted values between 0 and 1.
(B) Consider predicted values below 0 as 0 and values above 1 as 1.
(C) Use a logit function to transform the linear model into only predicting values between 0 and 1.
(D) Use the canonical link function for the Poisson distribution to transform the linear model into only predicting values between 0 and 1.
(E) None of the above.

21. A random walk is expressed as

$$y_t = y_{t-1} + c_t \text{ for } t = 1,2, \ldots$$

where

$$E(c_t) = \mu_c \text{ and } Var(c_t) = \sigma_c^2, \quad t = 1, 2, \ldots$$

Determine which statements is/are true with respect to a random walk model.

I. If $\mu_c \neq 0$, then the random walk is nonstationary in the mean.

II. If $\sigma_c^2 = 0$, then the random walk is nonstationary in the variance.

III. If $\sigma_c^2 > 0$, then the random walk is nonstationary in the variance.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

22. A stationary autoregressive model of order one can be written as

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t, \quad t = 1, 2, \ldots .$$

Determine which of the following statements about this model is false

(A) The parameter β_0 must not equal 1.
(B) The absolute value of the parameter β_1 must be less than 1.
(C) If the parameter $\beta_1 = 0$, then the model reduces to a white noise process.
(D) If the parameter $\beta_1 = 1$, then the model is a random walk.
(E) Only the immediate past value, y_{t-1}, is used as a predictor for y_t.

23. Toby observes the following coffee prices in his company cafeteria:

- 12 ounces for 1.00
- 16 ounces for 1.20
- 20 ounces for 1.40

The cafeteria announces that they will begin to sell any amount of coffee for a price that is the value predicted by a simple linear regression using least squares of the current prices on size.

Toby and his co-worker Karen want to determine how much they would save each day, using the new pricing, if, instead of each buying a 24-ounce coffee, they bought a 48-ounce coffee and shared it.

Calculate the amount they would save.

(A) It would cost them 0.40 more.

(B) It would cost the same.

(C) They would save 0.40.

(D) They would save 0.80.

(E) They would save 1.20.

24. Sarah performs a regression of the return on a mutual fund (y) on four predictors plus an intercept. She uses monthly returns over 105 months.

Her software calculates the F statistic for the regression as $F = 20.0$, but then it quits working before it calculates the value of R^2. While she waits on hold with the help desk, she tries to calculate R^2 from the F-statistic.

Determine which of the following statements about the attempted calculation is true.

(A) There is insufficient information, but it could be calculated if she had the value of the residual sum of squares (RSS).

(B) There is insufficient information, but it could be calculated if she had the value of the total sum of squares (TSS) and RSS.

(C) $R^2 = 0.44$

(D) $R^2 = 0.56$

(E) $R^2 = 0.80$

25. Determine which of the following statements concerning decision tree pruning is/are true.

I. The recursive binary splitting method can lead to overfitting the data.

II. A tree with more splits tends to have lower variance.

III. When using the cost complexity pruning method, $\alpha = 0$ results in a very large tree.

(A) None

(B) I and II only

(C) I and III only

(D) II and III only

(E) The correct answer is not given by (A), (B), (C), or (D).

26. Each picture below represents a two-dimensional space where observations are classified into two categories. The categories are representing by light and dark shading. A classification tree is to be constructed for each space.

Determine which space can be modeled with no error by a classification tree.

I.

II.

III.

(A) I only

(B) II only

(C) III only

(D) I, II, and III

(E) The correct answer is not given by (A), (B), (C), or (D).

27. Trevor is modeling monthly incurred dental claims. Trevor has 48 monthly claims observations and three potential predictors:

- Number of weekdays in the month
- Number of weekend days in the month
- Average number of insured members during the month

Trevor obtained the following results from a linear regression:

	Coefficient	**Standard Error**	***t* Stat**	***p*-value**
Intercept	–45,765,767.76	20,441,816.55	–2.24	0.0303
Number of weekdays	513,280.76	233,143.23	2.20	0.0330
Number of weekend days	280,148.46	483,001.55	0.58	0.5649
Average number of members	38.64	6.42	6.01	0.0000

Determine which of the following variables should be dropped, using a 5% significance level.

I. Intercept
II. Number of weekdays
III. Number of weekend days
IV. Number of members.

(A) I only
(B) II only
(C) III only
(D) IV only
(E) None should be dropped from the model

28. DELETED

29. Determine which of the following considerations may make decision trees preferable to other statistical learning methods.

I. Decision trees are easily interpretable.
II. Decision trees can be displayed graphically.
III. Decision trees are easier to explain than linear regression methods.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

30. Principal component analysis is applied to a large data set with four variables. Loadings for the first four principal components are estimated.

Determine which of the following statements is/are true with respect the loadings.

I. The loadings are unique.
II. For a given principal component, the sum of the squares of the loadings across the four variables is one.
III. Together, the four principal components explain 100% of the variance.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

31. Determine which of the following indicates that a nonstationary time series can be represented as a random walk

I. A control chart of the series detects a linear trend in time and increasing variability.
II. The differenced series follows a white noise model.
III. The standard deviation of the original series is greater than the standard deviation of the differenced series.

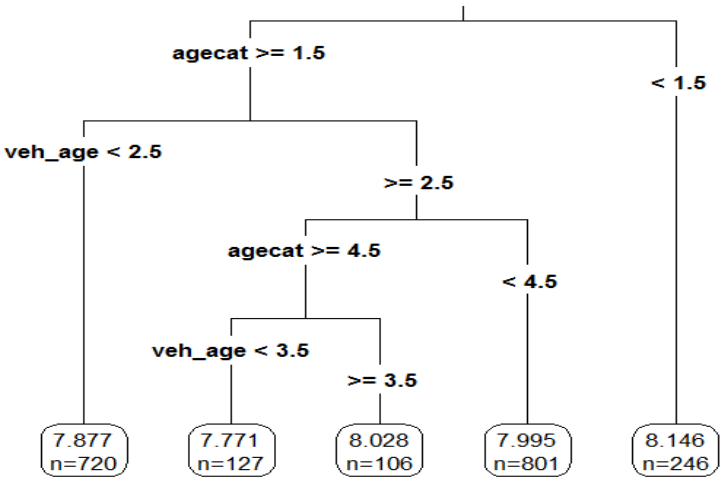
(A) I only
(B) II only
(C) III only
(D) I, II and III
(E) The correct answer is not given by (A), (B), (C), or (D).

32. You are given a set of n observations, each with p features.
Determine which of the following statements is/are true with respect to clustering methods.

I. The n observations can be clustered on the basis of the p features to identify subgroups among the observations.
II. The p features can be clustered on the basis of the n observations to identify subgroups among the features.
III. Clustering is an unsupervised learning method and is often performed as part of an exploratory data analysis.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

33. The regression tree shown below was produced from a dataset of auto claim payments. Age Category (agecat: 1, 2, 3, 4, 5, 6) and Vehicle Age (veh_age: 1, 2, 3, 4, …) are both predictor variables, and log of claim amount (LCA) is the dependent variable.



Consider three autos I, II, III:

I: An Auto in Age Category 1 and Vehicle Age 4
II: An Auto in Age Category 5 and Vehicle Age 5
III: An Auto in Age Category 5 and Vehicle Age 3

Rank the estimated LCA of Autos I, II, and III.

(A) LCA(I) < LCA(II) < LCA(III)
(B) LCA(I) < LCA(III) < LCA(II)
(C) LCA(II) < LCA(I) < LCA(III)
(D) LCA(II) < LCA(III) < LCA(I)
(E) LCA(III) < LCA(II) < LCA(I)

34. Determine which of the following statements is/are true about clustering methods:

I. If K is held constant, K-means clustering will always produce the same cluster assignments.

II. Given a linkage and a dissimilarity measure, hierarchical clustering will always produce the same cluster assignments for a specific number of clusters.

III. Given identical data sets, cutting a dendrogram to obtain five clusters produces the same cluster assignments as K-means clustering with $K = 5$.

(A) I only

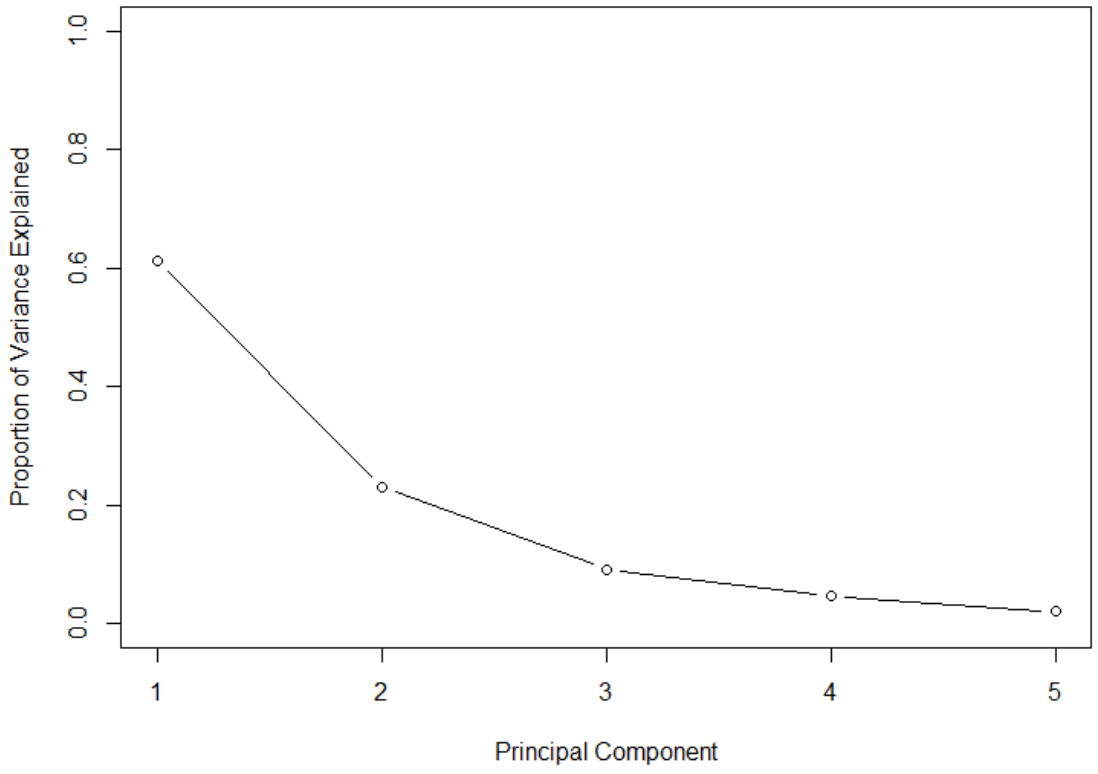
(B) II only

(C) III only

(D) I, II and III

(E) The correct answer is not given by (A), (B), (C), or (D).

35. Using the following scree plot, determine the minimum number of principal components that are needed to explain at least 80% of the variance of the original dataset.



(A) One

(B) Two

(C) Three

(D) Four

(E) It cannot be determined from the information given.

36. Determine which of the following statements about hierarchical clustering is/are true.

I. The method may not assign extreme outliers to any cluster.
II. The resulting dendrogram can be used to obtain different numbers of clusters.
III. The method is not robust to small changes in the data.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

37. Analysts W, X, Y, and Z are each performing Principal Components Analysis on the same data set with three variables. They use different programs with their default settings and discover that they have different factor loadings for the first principal component. Their loadings are:

	Variable 1	Variable 2	Variable 3
W	–0.549	–0.594	0.587
X	–0.549	0.594	0.587
Y	0.549	–0.594	–0.587
Z	0.140	–0.570	–0.809

Determine which of the following is/are plausible explanations for the different loadings.

I. Loadings are unique up to a sign flip and hence X's and Y's programs could make different arbitrary sign choices.

II. Z's program defaults to not scaling the variables while Y's program defaults to scaling them.

III. Loadings are unique up to a sign flip and hence W's and X's programs could make different arbitrary sign choices.

(A) None

(B) I and II only

(C) I and III only

(D) II and III only

(E) The correct answer is not given by (A), (B), (C), or (D).

38. You are given two models:

Model L: $y_t = \beta_0 + \beta_1 t + \varepsilon_t$

where $\{\varepsilon_t\}$ is a white noise process, for t = 0,1,2, …

Model M: $y_t = y_0 + \mu_c t + u_t$

$c_t = y_t - y_{t-1}$

$u_t = \sum_{j=1}^{t} \varepsilon_j$

where $\{\varepsilon_t\}$ is a white noise process, for t = 0,1,2, …

Determine which of the following statements is/are true.

I. Model L is a linear trend in time model where the error component is not a random walk.

II. Model M is a random walk model where the error component of the model is also a random walk.

III. The comparison between Model L and Model M is not clear when the parameter $\mu_c = 0$.

(A) I only

(B) II only

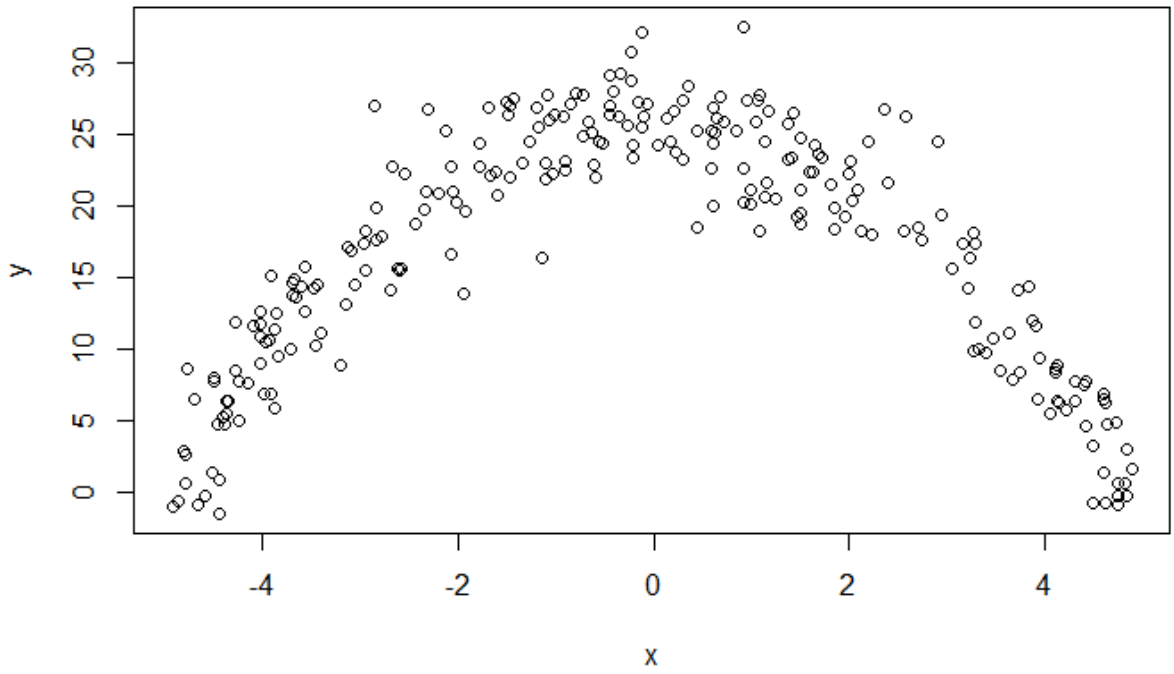
(C) III only

(D) I, II and III

(E) The correct answer is not given by (A), (B), (C), or (D).

39. You are given a dataset with two variables, which is graphed below. You want to predict y using x.

Determine which statement regarding using a generalized linear model (GLM) or a random forest is true.



(A) A random forest is appropriate because the dataset contains only quantitative variables.

(B) A random forest is appropriate because the data does not follow a straight line.

(C) A GLM is not appropriate because the variance of y given x is not constant.

(D) A random forest is appropriate because there is a clear relationship between y and x.

(E) A GLM is appropriate because it can accommodate polynomial relationships.

40. Determine which of the following statements about clustering is/are true.

I. Cutting a dendrogram at a lower height will not decrease the number of clusters.

II. *K*-means clustering requires plotting the data before determining the number of clusters.

III. For a given number of clusters, hierarchical clustering can sometimes yield less accurate results than *K*-means clustering.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

41. For a random forest, let p be the total number of features and m be the number of features selected at each split.
Determine which of the following statements is/are true.

I. When $m = p$, random forest and bagging are the same procedure.

II. $\frac{p-m}{p}$ is the probability a split will not consider the strongest predictor.

III. The typical choice of m is $\frac{p}{2}$.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

42. Determine which of the following statements is NOT true about the linear probability, logistic, and probit regression models for binary dependent variables.

(A) The three major drawbacks of the linear probability model are poor fitted values, heteroscedasticity, and meaningless residual analysis.

(B) The logistic and probit regression models aim to circumvent the drawbacks of linear probability models.

(C) The logit function is given by $\pi(z) = e^z / (1 + e^z)$.

(D) The probit function is given by $\pi(z) = \Phi(z)$ where Φ is the standard normal distribution function.

(E) The logit and probit functions are substantially different.

43. Determine which of the following statements about clustering methods is NOT true.

(A) Clustering is used to discover structure within a data set.

(B) Clustering is used to find homogeneous subgroups among the observations within a data set.

(C) Clustering is an unsupervised learning method.

(D) Clustering is used to reduce the dimensionality of a dataset while retaining explanation for a good fraction of the variance.

(E) In *K*-means clustering, it is necessary to pre-specify the number of clusters.

44. Two actuaries are analyzing dental claims for a group of $n = 100$ participants. The predictor variable is sex, with 0 and 1 as possible values.

Actuary 1 uses the following regression model:

$$Y = \beta + \varepsilon .$$

Actuary 2 uses the following regression model:

$$Y = \beta_0 + \beta_1 \times \text{Sex} + \varepsilon .$$

The residual sum of squares for the regression of Actuary 2 is 250,000 and the total sum of squares is 490,000.

Calculate the F-statistic to test whether the model of Actuary 2 is a significant improvement over the model of Actuary 1.

(A) 92

(B) 93

(C) 94

(D) 95

(E) 96

45. The actuarial student committee of a large firm has collected data on exam scores. A generalized linear model where the target is the exam score on a 0-10 scale is constructed using a log link, resulting in the following estimated coefficients

Predictor Variables	Coefficient
Intercept	– 0.1
Study Time (in units of 100 hours)	0.5
Attempt (1 for first attempt, else 0)	0.5
Master's degree (1 for Yes, 0 for No)	– 0.1
Interaction of Attempt and Master's degree	0.2

The company is about to offer a job to an applicant who has a Master's degree and for whom the exam would be a first attempt. It would like to offer half of the study time that will result in an expected exam score of 6.0.

Calculate the amount of study time that the company should offer.

(A) 123 hours
(B) 126 hours
(C) 129 hours
(D) 132 hours
(E) 135 hours

46. A time series was observed at times 0, 1, …, 100. The last four observations along with estimates based on exponential and double exponential smoothing with w = 0.8 are:

Time (t)	97	98	99	100
Observation (y_t)	96.9	98.1	99.0	100.2
Estimates ($\hat{s}_t^{(1)}$)	93.1	94.1	95.1	
Estimates ($\hat{s}_t^{(2)}$)	88.9	89.9		

All forecasts should be rounded to one decimal place and the trend should be rounded to three decimal places.

Let F be the predicted value of y_{102} using exponential smoothing with w = 0.8.

Let G be the predicted value of y_{102} using double exponential smoothing with w = 0.8.

Calculate the absolute difference between F and G, $|F - G|$.

(A) 0.0

(B) 2.1

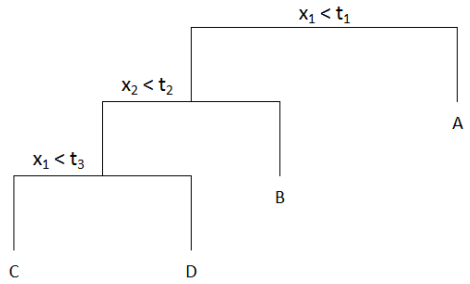
(C) 4.2

(D) 6.3

(E) 8.4

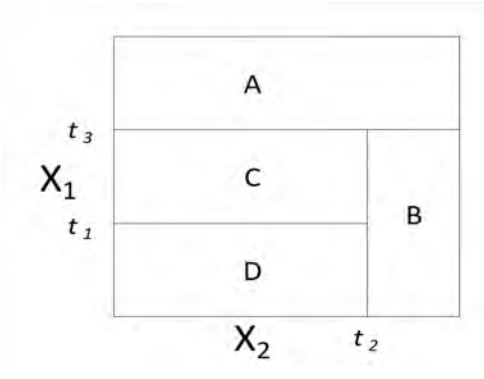
47. DELETED

48. The following tree was constructed using recursive binary splitting with the left branch indicating that the inequality is true.

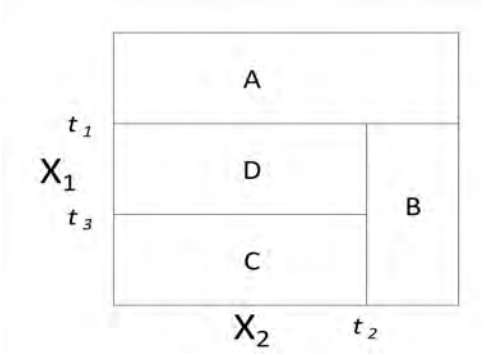


Determine which of the following plots represents this tree.

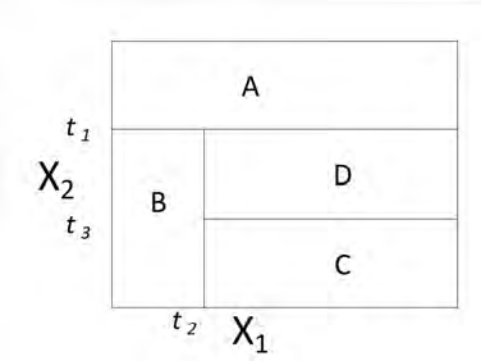
(A)



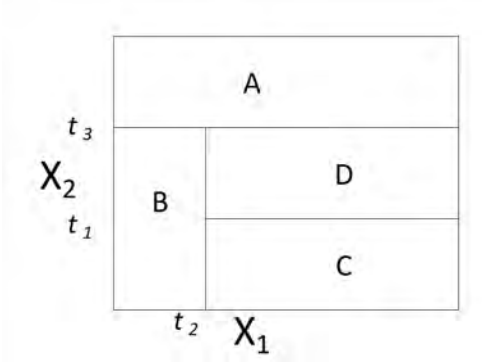
(B)



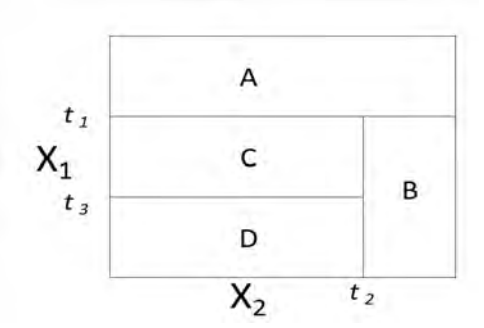
(C)



(D)



(E)



49. Trish runs a regression on a data set of n observations. She then calculates a 95% confidence interval (t,u) on y for a given set of predictors. She also calculates a 95% prediction interval (v,w) on y for the same set of predictors.

Determine which of the following must be true.

I. $\lim_{n\to\infty}(u-t)=0$

II. $\lim_{n\to\infty}(w-v)=0$

III. $w-v>u-t$

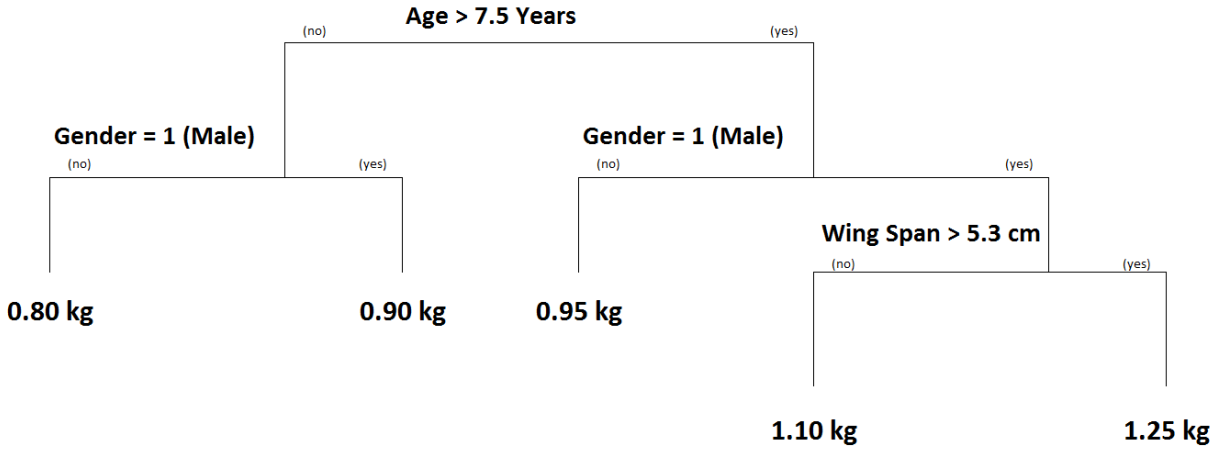
(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

50. Determine which of the following statements regarding statistical learning methods is/are true.

I. Methods that are highly interpretable are more likely to be highly flexible.
II. When inference is the goal, there are clear advantages to using a lasso method versus a bagging method.
III. Using a more flexible method will produce a more accurate prediction against unseen data.

(A) I only
(B) II only
(C) III only
(D) I, II and III
(E) The correct answer is not given by (A), (B), (C), or (D).

51. You are given the following regression tree predicting the weight of ducks in kilograms (kg):



You predict the weight of the following three ducks:

X: Wing Span = 5.5 cm, Male, Age = 7 years

Y: Wing Span = 5.8 cm, Female, Age = 5 years

Z: Wing Span = 5.7 cm, Male, Age = 8 years

Determine the order of the predicted weights of the three ducks.

(A) X < Y < Z

(B) X < Z < Y

(C) Y < X < Z

(D) Y < Z < X

(E) Z < X < Y

52. Determine which of the following statements is/are true about Pearson residuals.

I. They can be used to calculate a goodness-of-fit statistic.
II. They can be used to detect if additional variables of interest can be used to improve the model specification.
III. They can be used to identify unusual observations.

(A) I only
(B) II only
(C) III only
(D) I, II, and III
(E) The correct answer is not given by (A), (B), (C), or (D).

53. Determine which of the following statements is NOT true about the equation

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

(A) β_0 is the expected value of Y.
(B) β_1 is the average increase in Y associated with a one-unit increase in X.
(C) The error term, ε, is typically assumed to be independent of X.
(D) The equation defines the population regression line.
(E) The method of least squares is commonly used to estimate the coefficients β_0 and β_1.

54. For a regression model of executive compensation, you are given:

i) The following statistics:

Executive Compensation				
Coefficients:	Estimate	Std. Error	*t*-statistic	*p*-value
(INTERCEPT)	–28,595.5	220.5	–129.7	<0.001
AGEMINUS35	7,366.3	12.5	588.1	<0.001
TOPSCHOOL	50.0	119.7	0.4	0.676
LARGECITY	147.9	119.7	1.2	0.217
MBA	2,490.9	119.7	20.8	<0.001
YEARSEXP	15,286.6	7.2	2132.8	<0.001

ii) The acceptable significance level is $\alpha = 0.10$.

Determine which variable or variables should be removed first prior to rerunning the model.

(A) (INTERCEPT)
(B) AGEMINUS35, MBA, and YEARSEXP
(C) TOPSCHOOL
(D) TOPSCHOOL and LARGECITY
(E) YEARSEXP

55. You are given the following eight observations from a time series that follows a random walk model:

Time (t)	0	1	2	3	4	5	6	7
Observation (y_t)	3	5	7	8	12	15	21	22

You plan to fit this model to the first five observations and then evaluate it against the last three observations using one-step forecast residuals. The estimated mean of the white noise process is 2.25.

Let F be the mean error (ME) of the three predicted observations.

Let G be the mean square error (MSE) of the three predicted observations.

Calculate the absolute difference between F and G, $|F - G|$.

(A) 3.48
(B) 4.31
(C) 5.54
(D) 6.47
(E) 7.63

56. Determine which of the following statements about prediction is true.

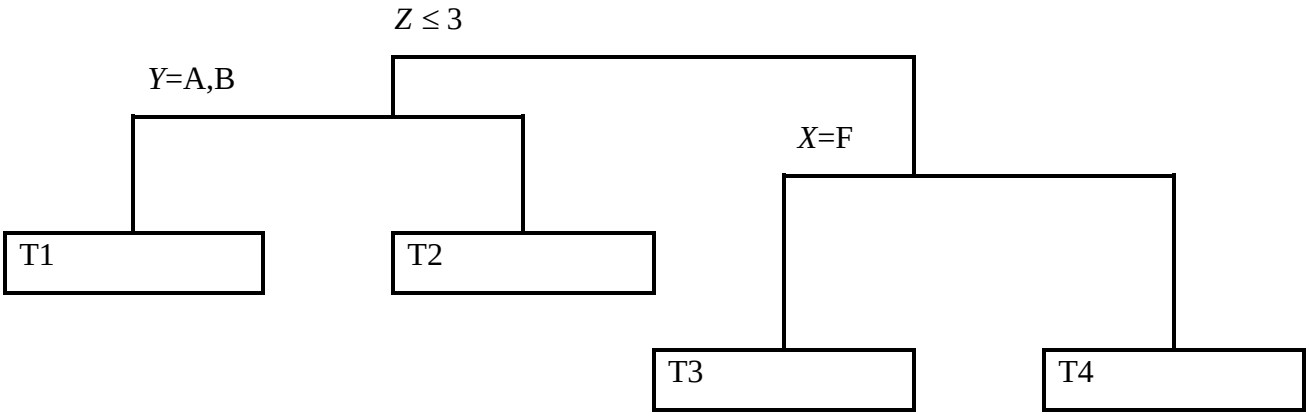
(A) Each of several candidate regression models must produce the same prediction.
(B) When making predictions, it is assumed that the new observation follows the same model as the one used in the sample.
(C) A point prediction is more reliable than an interval prediction.
(D) A wider prediction interval is more informative than a narrower prediction interval.
(E) A prediction interval should not contain the single point prediction.

57. You are given:

i) The following observed values of the response variable, *R*, and predictor variables *X*, *Y*, *Z*:

R	4.75	4.67	4.67	4.56	4.53	3.91	3.90	3.90	3.89
X	M	F	M	F	M	F	F	M	M
Y	A	A	D	D	B	C	B	D	B
Z	2	4	1	3	2	2	5	5	1

ii) The following plot of the corresponding regression tree:



Calculate the Mean Response (MR) for each of the end nodes.

(A) MR(T1) = 4.39, MR(T2) = 4.38, MR(T3) = 4.29, MR(T4) = 3.90

(B) MR(T1) = 4.26, MR(T2) = 4.38, MR(T3) = 4.62, MR(T4) = 3.90

(C) MR(T1) = 4.26, MR(T2) = 4.39, MR(T3) = 3.90, MR(T4) = 4.29

(D) MR(T1) = 4.64, MR(T2) = 4.29, MR(T3) = 4.38, MR(T4) = 3.90

(E) MR(T1) = 4.64, MR(T2) = 4.38, MR(T3) = 4.39, MR(T4) = 3.90

58. You are given the following six observed values of the autoregressive model of order one time series

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t \text{ with } Var(\varepsilon_t) = \sigma^2.$$

t	1	2	3	4	5	6
y_t	31	35	37	41	45	51

The approximation to the conditional least squares method is used to estimate β_0 and β_1.

Calculate the mean squared error s^2 that estimates σ^2.

(A) 13

(B) 21

(C) 22

(D) 26

(E) 35

59. You apply 2-means clustering to a set of five observations with two features. You are given the following initial cluster assignments:

Observation	$X1$	$X2$	Initial cluster
1	1	3	1
2	0	4	1
3	6	2	1
4	5	2	2
5	1	6	2

Calculate the total within-cluster variation of the initial cluster assignments, based on Euclidean distance measure.

(A) 32.0
(B) 70.3
(C) 77.3
(D) 118.3
(E) 141.0

60. Determine which of the following statements about selecting the optimal number of clusters in *K*-means clustering is/are true.

I. *K* should be set equal to *n*, the number of observations.
II. Choose *K* such that the total within-cluster variation is minimized.
III. The determination of *K* is subjective and there does not exist one method to determine the optimal number of clusters.

(A) I only
(B) II only
(C) III only
(D) I, II and III
(E) The correct answer is not given by (A), (B), (C), or (D).

61. A linear model has been fit to a dataset containing six predictor variables, F, G, H, I, J, and K. Determine which of the following statements regarding using Akaike information criterion (AIC) or Bayesian information criterion (BIC) to select an optimal set of predictor variables for this linear model is/are true.

I. AIC and BIC provide a direct estimate of the test error.
II. When choosing between the subsets {F, G, H} and {I, J, K}, AIC and BIC will always select the same subset.
III. For large sample sizes ($n > 7$), the number of variables selected by BIC will be less than or equal to the number selected by AIC.

(A) None
(B) I and II only
(C) I and III only
(D) II and III only
(E) The correct answer is not given by (A), (B), (C), or (D).

62. In a simple linear regression model based on over 100 observations, you are given the following estimates.

i) The estimated slope is –1.03.

ii) The standard error of the estimated slope is 0.06.

Calculate the 95% confidence interval for the slope.

(A) (–1.15, –0.91)

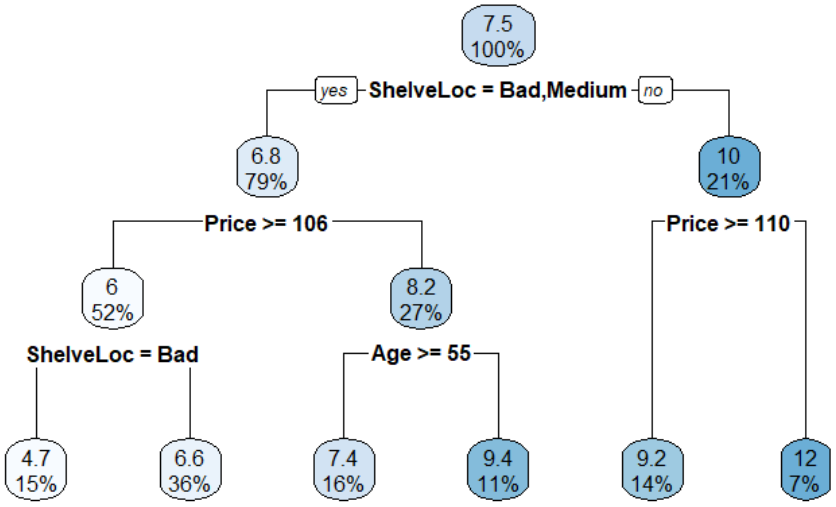
(B) (–1.13, –0.93)

(C) (–1.11, –0.95)

(D) (–1.09, –0.97)

(E) (–1.07, –0.99)

63. You have constructed the following regression tree predicting unit sales (in thousands) of car seats. The variable ShelveLoc has possible values Good, Medium, and Bad.



Variable	Observed Value
ShelveLoc	Good
Price	120
Age	57
Advertising	12

Determine the predicted unit sales (in thousands) for the above observation based on the regression tree.

(A) 4.7

(B) 6.6

(C) 7.4

(D) 9.2

(E) 9.4

64. You are given a stationary AR(1) model, $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t, \quad t = 2,\ldots,T.$

Determine which or the following is always true.

(A) $\beta_0 \neq 0$

(B) $\beta_0 = 1$

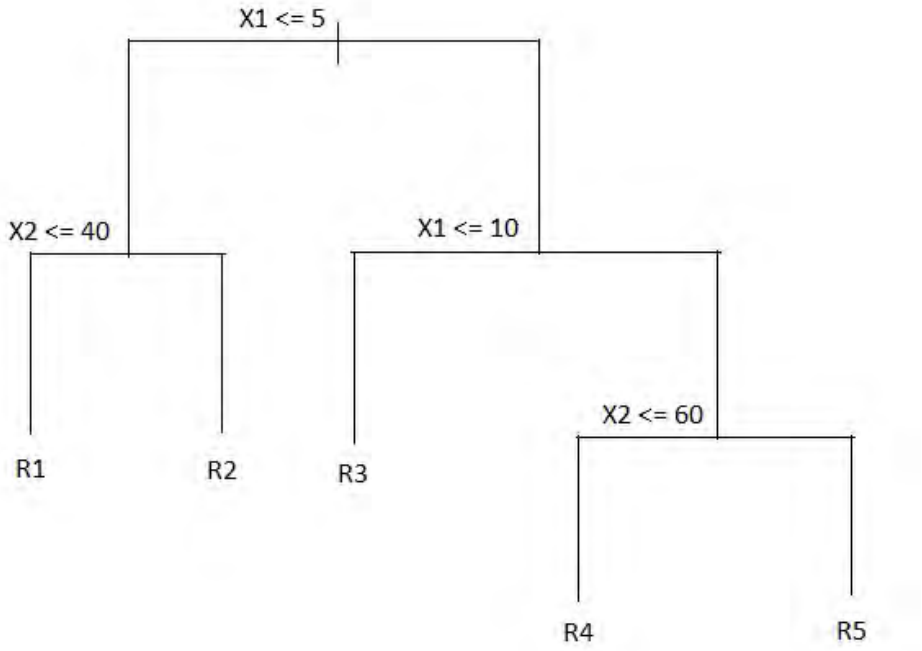
(C) $\beta_1 = 0$

(D) $\beta_1 = 1$

(E) $|\beta_1| < 1$

65. DELETED

66. You are given the following regression tree (where the inequality represents the values in the left hand branch of a split.



Determine the set of regions that represents this regression tree.

(A) Region 1: $\{X|\ X_1 \leq 5, X_2 \leq 40\}$
Region 2: $\{X|\ X_1 \leq 5, X_2 > 40\}$
Region 3: $\{X|\ X_1 > 5\}$
Region 4: $\{X|\ X_1 > 10, X_2 \leq 60\}$
Region 5: $\{X|\ X_1 > 10, X_2 > 60\}$

(B) Region 1: $\{X|\ X_1 \leq 5, X_2 \leq 40\}$
Region 2: $\{X|\ X_1 \leq 5, X_2 > 40\}$
Region 3: $\{X|\ X_1 > 5, X_2 \leq 60\}$
Region 4: $\{X|\ X_1 > 10, X_2 \leq 60\}$
Region 5: $\{X|\ X_1 > 10, X_2 > 60\}$

(C) Region 1: $\{X|\ X_1 \leq 5, X_2 > 40\}$
Region 2: $\{X|\ X_1 \leq 5, X_2 \leq 40\}$
Region 3: $\{X|\ 5 < X_1 \leq 10\}$
Region 4: $\{X|\ X_1 > 10, X_2 \leq 60\}$
Region 5: $\{X|\ X_1 > 10, X_2 > 60\}$

(D) Region 1: $\{X \mid X_1 \leq 5, X_2 \leq 40\}$
Region 2: $\{X \mid X_1 \leq 5, X_2 > 40\}$
Region 3: $\{X \mid 5 < X_1 \leq 10\}$
Region 4: $\{X \mid X_1 > 10, X_2 \leq 60\}$
Region 5: $\{X \mid X_1 > 10, X_2 > 60\}$

(E) Region 1: $\{X \mid X_1 \leq 5, X_2 \leq 40$
Region 2: $\{X \mid X_1 \leq 5, X_2 > 40\}$
Region 3: $\{X \mid X_1 > 5, X_2 \leq 60\}$
Region 4: $\{X \mid X_1 > 10, X_2 > 60\}$
Region 5: $\{X \mid X_1 > 10, X_2 \leq 60\}$

67. A and B each attempt a task 10 times. A is successful 8 out of 10 times while B is successful 2 out of 10 times. They fit a logistic regression model to predict an individual's success with linear term $\beta_0 + \beta_1 X$, where X is an indicator variable for individual A.

The full model estimates that A will be successful 80% of the time and B will be successful 20% of the time. Under the null hypothesis that $\beta_1 = 0$, the estimate is that both A and B will be successful 50% of the time. The likelihood ratio test is applied to the logistic regression model.

Determine the lowest available *p*-value at which the null hypothesis is rejected.

(A) 0.100
(B) 0.050
(C) 0.025
(D) 0.010
(E) 0.005

68. You are fitting a boosted regression tree with shrinkage parameter $\lambda = 0.08$. There is a single numeric predictor variable, x, and the target variable is y. Each tree is to have a single split. The first tree, built using all the training data has the following split:

- If x < 17.5, then the predicted value of y is 125.4.
- If x ≥ 17.5, then the predicted value of y is 350.5.

The following are three observations from the training set:

Observation	x	y
1	15.2	139
2	12.9	156
3	23.6	289

Calculate the updated residuals for these three observations (r_1, r_2, r_3).

(A) (–406.1, –389.1, –31.1)

(B) (–211.4, –194.4, 163.6)

(C) (10.0, 10.0, 28.0)

(D) (111.0, 128.0, 279.0)

(E) (129.0, 146.0, 261.0)

69. You are fitting a boosted regression tree with shrinkage parameter $\lambda = 0.08$. There is a single numeric predictor variable, x, and the target variable is y. Each tree is to have a single split. The first tree, built using all the training data has the following split:

- If x < 17.5, then the predicted value of y is 125.4.
- If x ≥ 17.5, then the predicted value of y is 350.5.

The following are three observations from the training set:

Observation	x	y
1	15.2	139
2	12.9	156
3	23.6	289

Calculate the updated predictions for these three observations $(\hat{f}(x_1), \hat{f}(x_2), \hat{f}(x_3))$.

(A) (–406.1, –389.1, –31.1)

(B) (–211.4, –194.4, 163.6)

(C) (10.0, 10.0, 28.0)

(D) (111.0, 128.0, 279.0)

(E) (129.0, 146.0, 261.0)

70. A dataset has 100 observations with six predictor variables, X1, X2, X3, X4, X5, and X6 and a response variable, Y. Three ordinary regression models have been run, with the following results:

Model Number	Variables Used	Residual Sum of Squares	Residual Standard Error
I	X1 only	183,663.30	43.2911
II	X1 and X2 only	8,826.47	9.5391
III	X1, X2, X3, X4, X5, and X6	8,319.59	9.4582

The Akaike Information Criterion (AIC) is to be used to select the best model from these three choices.

Determine which of the following statements is true.

(A) Select Model I with an AIC value of 1,838.42

(B) Select Model I with an AIC value of 1,874.12

(C) Select Model II with an AIC value of 91.84

(D) Select Model II with an AIC value of 91.90

(E) Select Model III with an AIC value of 93.93

71. Determine which of the following statements regarding the effect of omitting important variables from the regression model specification is NOT true.

(A) It will result in underfitting.
(B) It will increase the width of prediction intervals.
(C) The coefficient estimates will remain unbiased.
(D) The estimate of the variance, s^2, is inflated.
(E) The total sum of squares remains unchanged.

72. An actuary is investigating the probability of a policyholder incurring positive medical expenditures. Using a sample of data from the company's claims files, a logistic regression model is fit and produces the following table of estimates:

Predictor Variable	Coefficient Estimate	Standard error	*p*-value
Age	0.257	0.102	0.0117
Sex	–0.955	0.542	0.0781
Education	0.103	0.069	0.1355
Log(Income)	–0.516	0.157	0.0010

Determine which of the following predictor variables are significant, using a 5% level of significance.

(A) Log(Income) only
(B) Education only
(C) Age and Log(Income) only
(D) Sex and Education only
(E) All four are significant