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.

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 increases as more principal components are added.

II. The cumulative proportion of variance explained increases 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. Because the logarithm of zero is negative infinity, a logarithm transformation cannot be used.

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) 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).

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. The regression model is $y = \beta_0 + \beta_1 x + \varepsilon$. There are six observations.

The summary statistics are:

$\sum y_i = 8.5$, $\sum x_i = 6$, $\sum x_i^2 = 16$, $\sum x_i y_i = 15.5$, $\sum y_i^2 = 17.25$.

Calculate the least squares estimate of β_1.

(A) 0.1

(B) 0.3

(C) 0.5

(D) 0.7

(E) 0.9

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. Dental claims experience was collected on 6480 policies. There were a total of 9720 claims on these policies. The following table shows the number of dental policies having varying numbers of claims.

Number of Claims	Number of Policies
0	1282
1	2218
2	1856
3	801
4	235
5	81
6 or more	7
Total	**6480**

Calculate the chi-squared statistic to test if a Poisson model with no predictors provides an adequate fit to the data.

(A) 80

(B) 83

(C) 86

(D) 89

(E) 92

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. Sarah is applying principal component analysis 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. We can cluster the n observations on the basis of the p features in order to identify subgroups among the observations.

II. We can cluster p features on the basis of the n observations in order to discover subgroups among the features.

III. Clustering is an unsupervised learning method and is often performed as part of an exploratory data analysis.

(A) None

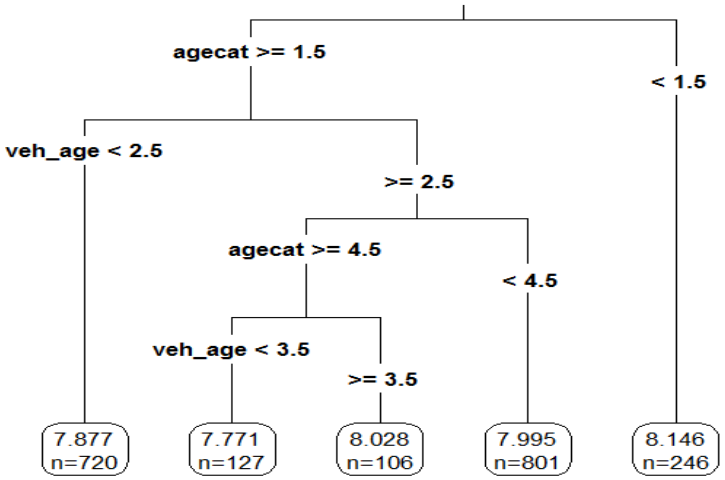
(B) I and II only

(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

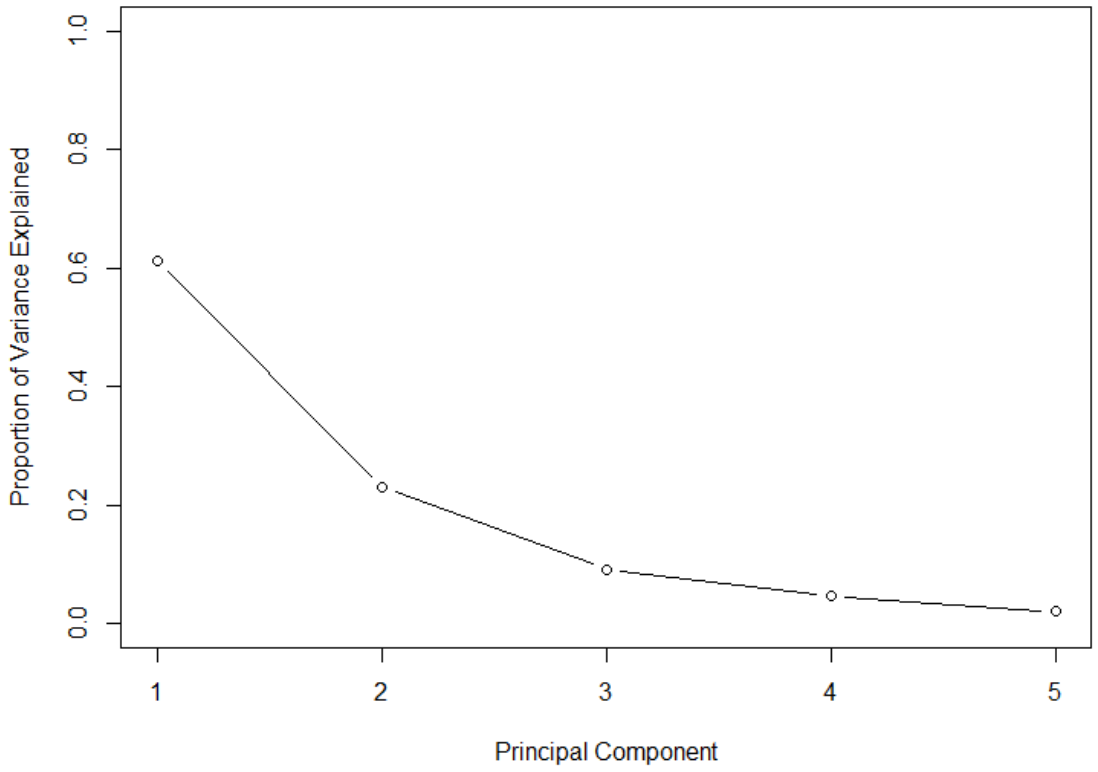
(B) II only

(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 $t = 0,1,2, \ldots$

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 $t = 0,1,2, \ldots$

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

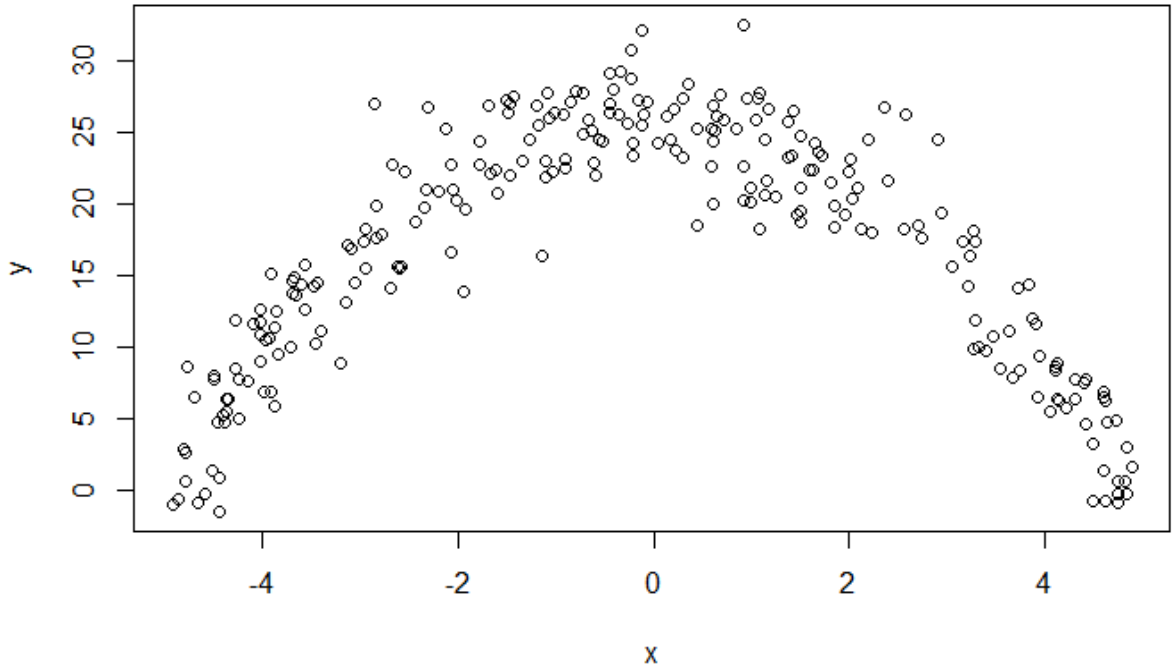
(C) III only

(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 is NOT true about clustering methods.

(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 gender, with 0 and 1 as possible values.

Actuary 1 uses the following regression model:

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

Actuary 2 uses the following regression model:

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

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

SOLUTIONS

1. **Key: E**

First, calculate the distance between pairs of elements in each set. There are two pairs here:

$$x_1, x_4 : \sqrt{(-1-5)^2 + (0-10)^2} = \sqrt{136} = 11.66$$
$$x_2, x_4 : \sqrt{(1-5)^2 + (1-10)^2} = \sqrt{97} = 9.85.$$

For complete linkage, the dissimilarity measure used is the maximum, which is 11.66.

2. **Key: C**

I is false because the number of clusters is pre-specified in the *K*-means algorithm but not for the hierarchical algorithm. II is also false because both algorithms force each observation to a cluster so that both may be heavily distorted by the presence of outliers. III is true.

3. **Key: D**

$$y_{10} = y_0 + c_1 + \cdots + c_{10} = y_0 + 20$$
$$y_{19} = y_{10} + c_{11} + \cdots + c_{19} = y_0 + 20 + c_{11} + \cdots + c_{19} = y_0 + 20 + 26 = y_0 + 46$$
$$\hat{y}_{19} = y_{10} + \hat{c}_{11} + \cdots + \hat{c}_{19} = y_0 + 20 + 9(2) = y_0 + 38$$
$$y_{19} - \hat{y}_{19} = (y_0 + 46) - (y_0 + 38) = 8.$$

4. **Key: B**

$c_t = y_t - y_{t-1}$ and hence $c_1, c_2, \ldots c_{10} = 2, 3, 5, 3, 5, 2, 4, 1, 2, 3.$

The mean of the *c* values is 3, the variance is (1+0+4+0+4+1+1+4+1+0)/9 = 16/9. The standard deviation is 4/3. The standard error of the forecast is $(4/3)\sqrt{9} = 4.$

5. **Key: E**

Statement I is correct – Principal components provide low-dimensional linear surfaces that are closest to the observations.

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

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

Statement IV is correct – PCA serves as a tool for data visualization.

6. **Key: E**

Statement I is incorrect – The proportion of variance explained by an additional principal component decreases as more principal components are added.

Statement II is correct – The cumulative proportion of variance explained increases as more principal components are added.

Statement III is incorrect – We want to use the least number of principal components required to get the best understanding of the data.

Statement IV is correct – Typically, the number of principal components are chosen based on a scree plot.

7. The intent is to model a binary outcome, thus a classification model is desired. In GLM, this is equivalent to binomial distribution. The link function should be one that restricts values to the range zero to one. Of linear and logit, only logit has this property.

8. Key: D

Alternative fitting procedures will tend to remove the irrelevant variables from the predictors, thus resulting in a simpler and easier to interpret model. Accuracy will likely be improved due to reduction in variance.

9. **Key: E**

The total Gini index for Split 1 is 2[20(12/20)(8/20) + 80(18/80)(62/80)]/100 = 0.375 and for Split 2 is 2[10(8/10)(2/10) + 90(22/90)(68/90)]/100 = 0.3644. Smaller is better, so Split 2 is preferred. The factor of 2 is due to summing two identical terms (which occurs when there are only two classes).

The total entropy for Split 1 is –[20(12/20)ln(12/20) +20(8/20)ln(12/20) + 80(18/80)ln(18/80) + 80(62/80)ln(62/80)]/100 = 0.5611 and for Split 2 is –[10(8/10)ln(8/10) +10(2/10)ln(2/10) + 90(22/90)ln(22/90) + 90(68/90)ln(68/90)]/100 = 0.5506. Smaller is better, so Split 2 is preferred.

For Split 1, there are 8 + 18 = 26 errors and for Split 2 there are 2 + 22 = 24 errors. With fewer errors, Split 2 is preferred.

10. **Key: C**

II is false because with random forest a new subset of predictors is selected for each split.

11. **Key: E**

Solution: SSE is sum of the squared differences between the observed and predicted values. That is, $[(2-4)^2+(5-3)^2+(6-9)^2+(8-3)^2+(4-6)^2]=46.$

12. **Key: E**

A is false, linear regression is considered inflexible because the number of possible models is restricted to a certain form.

B is false, the lasso determines the subset of variables to use while linear regression allows the analyst discretion regarding adding or moving variables.

C is false, bagging provides additional flexibility.

D is false, there is a tradeoff between being flexible and easy to interpret.

13. **Key E**

I is true. The prediction interval includes the irreducible error, but in this case it is zero.

II is true. Because it includes the irreducible error, the prediction interval is at least as wide as the confidence interval.

III. is false. It is the confidence interval that quantifies this range.

14. **Key: B**

Changing the transformation to $log(1 + y)$ can solve the problem of y being zero. Hence I is false. Power transformations with the power less than one, such as the square root transformation, may make the variance constant. Hence II is true. A logit transformation requires that the variable take on values between 0 and 1 and hence cannot be used here.

15. **Key: D**

The cluster centers are A: (0, 1), B: (2, –2), and C: (0, 2). The new assignments are:

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

Cluster C has the fewest points with three.

16. **Key: D**

(A) For *K*-means the initial cluster assignments are random. Thus different people can have different clusters, so the statement is not true for *K*-means clustering. It is true for hierarchical clustering.

(B) For *K*-means the number of clusters is set in advance and does not change as the algorithm is run. For hierarchical clustering the number of clusters is determined after the algorithm is completed.

(C) For *K*-means the algorithm needs to be re-run if the number of clusters is changed. This is not the case for hierarchical clustering.

(D) This is true for *K*-means clustering. Agglomerative hierarchical clustering starts with each data point being its own cluster.

17. **Key: D**

$\bar{x} = 6/6 = 1, \quad \bar{y} = 8.5/6 = 1.41667$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{15.5 - 6(1)(1.41667)}{16 - 6(1^2)} = 0.7.$$

18. **Key: B**

$TSS = RSS/(1-R^2) = 638.89$

19. **Key: E**

The only two models that need to be considered are the full model with all four coefficients and the null model with only the intercept term. The test statistic is twice the difference of the log-likelihoods, which is 10.

The number of degrees of freedom is the difference in the number of coefficients in the two models, which is three.

At 5% significance with three degrees of freedom, the test statistic of 10 exceeds the 7.81 threshold, so the null hypothesis should be rejected.

20. **Key: C**

(A) is not appropriate because removing data will likely bias the model estimates.

(B) is not appropriate because altering data will likely bias the model estimates.

(C) is correct.

(D) is not appropriate because the canonical link function is the logarithm, which will not restrict values to the range zero to one.

21. **Key: C**

I is true because the mean $E(y_t) = y_0 + t\mu_c$ depends on *t*.

II is false because the variance $Var(y_t) = t\sigma_c^2 = 0$ does not depend in *t*.

III is true because the variance depends on *t*.

22. **Key: A**

The intercept term may be any value, hence (A) is false.

23. **Key: C**

The regression line is $y = 0.40 + 0.05x$. This can be obtained by either using the formula for the regression coefficients or by observing that the three points lie on a straight line and hence that line must be the solution. A 24-ounce cup costs 1.60. Two of them cost 3.20. A 48-ounce cup costs $2.80. So the savings is 0.40.

24. **Key: C**

Even though the formula for R^2 involves RSS and TSS, she just needs their ratio, which can be obtained from *F*.

$$F = \frac{(TSS - RSS)/4}{RSS/(105-4-1)} = 20$$

$$\frac{TSS - RSS}{RSS} = 20(4)/100 = 0.80$$

$$\frac{TSS}{RSS} = 1.80$$

$$\frac{RSS}{TSS} = \frac{1}{1.80}$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{1}{1.80} = \frac{0.80}{1.80} = 0.44$$

25. **Key: C**

I is true because the method optimizes with respect to the training set, but may perform poorly on the test set.

II is false because additional splits tends to increase variance due to adding to the complexity of the model.

III is true because in this case only the training error is measured.

26. **Key: A**

Each step must divide the one the existing regions into two parts, using either a vertical or a horizontal line. Only I can be created this way.

27. **Key: C**

Only variables with a *p*-value greater than 0.05 should be considered. Because only one of the variables (Number of weekend days) meets this criterion, it should be dropped.

28. **Key: C**

The average number of claims on a policy is 9720/6480 = 1.5. Using this as the parameter of a Poisson distribution yields the following table of actual (given) and expected (calculated) policies with given numbers of claims:

Number of Claims	Actual Number of Policies	Poisson Probability	Expected Number of Policies	Chi-square
0	1282	0.2231	1445.69	18.53
1	2218	0.3347	2168.86	1.11
2	1856	0.2510	1626.48	32.39
3	801	0.1255	813.24	0.18
4	235	0.0471	305.21	16.15
5	81	0.0141	91.37	1.18
6 or more	7	0.0045	29.16	16.84
Total	**6480**	**1.0000**	**6480.01**	**86.38**

29. **Key: E**

All three statements are true. See Section 8.1 of *An Introduction to Statistical Learning*. The statement that trees are easier to explain than linear regression methods may not be obvious. For those familiar with regression but just learning about trees, the reverse may be the case. However, for those not familiar with regression, relating the dependent variable to the independent variables, especially if the dependent variable has been transformed, can be difficult to explain.

30. **Key: D**

I is false because the loadings are unique only up to a sign flip.

II is true. Principal components are designed to maximize variance. If there are no constraints on the magnitude of the loadings, the variance can be made arbitrarily large. The PCA algorithm's constraint is that the sum of the squares of the loadings equals 1.

III is true because four components can capture all the variation in four variables, provided there are at least four data points (note that the problem states that the data set is large).

31. **Key: D**

See Page 242 of *Regression Modeling with Actuarial and Financial Applications*.

I is true because a random walk is characterized by a linear trend and increasing variability.

II is true because differencing removes the linear trend and stabilizes the variance.

III is true as both the linear trend and the increasing variability contribute to a higher standard deviation.

32. **Key: E**

I and II are both true because the roles of rows and columns can be reversed in the clustering algorithm. (See Section 10.3 of *An Introduction to Statistical Learning*.)

III is true. Clustering is unsupervised learning because there is no dependent (target) variable. It can be used in exploratory data analysis to learn about relationships between observations or features.

33. **Key: E**

LCA(I) = 8.146

LCA(II) = 8.028

LCA(III) = 7.771

34. **Key: B**

I is false. *K*-means clustering is subject to the random initial assignment of clusters.

II is true. Hierarchical clustering is deterministic, not requiring a random initial assignment.

III is false. The two methods differ in their approaches and hence may not yield the same clusters.

35. **Key: B**

The first PC explains about 62% of the variance. The second PC explains about 23% of the variance. Combined they explain about 85% of the variance, and hence two PCs are sufficient.

36. **Key: D**

I is false. All observations are assigned to a cluster.

II is true. By cutting the dendrogram at different heights, any number of clusters can be obtained.

III is true. Clustering methods have high variance, that is, having a different random sample from the population can lead to different clustering.

37. **Key: B**

I is true. Uniqueness up to a sign flip means all three signs must be flipped. This is true for X and Y.

II is true. The presence of absence of scaling can change the loadings.

III is false. Uniqueness up to a sign flip means all three signs must be flipped. For W and X only the second loading is flipped.

38. **Key: D**

I is true. See formula (7.8) in the Frees text.

II. is true. See formula (7.9) in the Frees text.

III is true. The only difference is the error terms, which are difficult to compare. See page 243 in the Frees text.

39. **Key: E**

(A) is false. Trees work better with qualitative data.

(B) is false. While trees accommodate nonlinear relations, as seen in (E) a linear model can work very well here.

(C) is false. The variance is constant, so that is not an issue here.

(D) is false. There is a clear relationship as noted in answer (E).

(E) is true. The points appear to lie on a quadratic curve so a model such as $y = \beta_0 + \beta_1 x + \beta_2 x^2$ can work well here. Recall that linear models must be linear in the coefficients, not necessarily linear in the variables.

40. **Key: C**

I is true. At the lowest height, each observation is its own cluster. The number of clusters decreases as the height increases.

II is false. There is no need to plot the data to perform *K*-means clustering.

III is true. *K*-means does a fresh analysis for each value of *K* while for hierarchical clustering, reduction in the number of clusters is tied to clusters already made. This can miss cases where the clusters are not nested.

41. **Key: B**

I is true. Random forests differ from bagging by setting $m < p$.

II is true. $p - m$ represents the splits not chosen.

III is false. Typical choices are the square root of p or $p/3$.

42. **Key: E**

The logit and probit models are similar (see page 307 of Frees, which also discusses items A-D).

43. **Key: D**

Item D is a statement about principal components analysis, not clustering.

44. **Key: C**

The model of Actuary 1 is the null model and hence values from it are not needed. The solution is $F = \frac{(TSS - RSS)/1}{RSS/(n-2)} = \frac{490,000 - 250,000}{250,000/98} = 94.08.$