Week 10 - AYU - Individual

Problem 1

An automobile insurance company wants to use gender ($x_1 = 0$, if female and $x_1 = 1, if male$) and traffic penalty point $(x)2)$ to predict the number of claims (y). The observed values of these variables for a sample of six motorists are given by:

Motorist	1	2	3	4	5	6
x_1	0	0	0	1	1	1
x_2	0	1	2	0	1	2
y	1	0	2	1	3	5

You are using the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, 2, ..., 6$$

You have determine

$$(X'X)^{-1} = \frac{1}{12}\begin{bmatrix} 7 & -4 & -3 \\ -4 & 8 & 0 \\ -3 & 0 & 3 \end{bmatrix}$$

Determine $\hat{\beta}_2$.

(A) -0.25
(B) 0.25
(C) 1.25
(D) 2.00
(E) 4.25

Problem 2

You are performing a multiple regression of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

You have obtain the following data

y	x_1	x_2
1	-1	-1
2	1	-1
3	-1	1
4	1	1

Determine $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$

(A) 3.5
(B) 4.0
(C) 4.5
(D) 5.0
(E) 5.5

Problem 3

The following two models were fit to 18 observations:

Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$

The result of the regression are:

Model Number	Error Sum of Squares	Regression Sum of Squares
1	102	23
2	78	47

Calculate the value of the F-statistics used to test the hypothesis that $\beta_3 = \beta_4 = \beta_5 = 0$

(A) Less than 1.30
(B) At least 1.30, but less than 1.40
(C) At least 1.40, but less than 1.50
(D) At least 1.50, but less than 1.60
(E) At least 1.60

Problem 4

An automobile insurance company wants to use gender ($x_1 = 0$, if female and $x_1 = 1, if male$) and traffic penalty point $(x)2)$ to predict the number of claims (y). The observed values of these variables for a sample of six motorists are given by:

Motorist	1	2	3	4	5	6
x_1	0	0	0	1	1	1
x_2	0	1	2	0	1	2
y	1	0	2	1	3	5

You are using the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, 2, ..., 6$$

You have determine

$$(X'X)^{-1} = \frac{1}{12} \begin{bmatrix} 7 & -4 & -3 \\ -4 & 8 & 0 \\ -3 & 0 & 3 \end{bmatrix}$$

Determine $\hat{\beta}_2$.

(A) -0.25
(B) 0.25

(C) 1.25
(D) 2.00
(E) 4.25

Problem 5

You are performing a multiple regression of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

You have obtain the following data

y	x_1	x_2
1	-1	-1
2	1	-1
3	-1	1
4	1	1

Determine $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$

(A) 3.5
(B) 4.0
(C) 4.5
(D) 5.0
(E) 5.5

Problem 6

The following two models were fit to 18 observations:

Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$

The result of the regression are:

Model Number	Error Sum of Squares	Regression Sum of Squares
1	102	23
2	78	47

Calculate the value of the F-statistics used to test the hypothesis that $\beta_3 = \beta_4 = \beta_5 = 0$

(A) Less than 1.30
(B) At least 1.30, but less than 1.40
(C) At least 1.40, but less than 1.50
(D) At least 1.50, but less than 1.60
(E) At least 1.60

Problem 7

A professor ran an experiment in three sections of a psychology courses to show that the more digits in a number, the more difficult it is to remember. The following variables were used in a multiple regression:

\$x_1 = \$ number of digits in the number $x_2 = 1$ if student was in section 1, 0 otherwise $x_3 = 1$ if student was in section 2, 0 otherwise \$y = \$ percentage of students correctly remembering the number

You are given

- A total of 42 students participated in the study
- The regression equation
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \epsilon$$
was to fit the data and resulted in $R^2 = 0.940$
- A second regression equation $y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_1^2 + \epsilon$ was to fit to the data and resulted $R^2 = 0.915$

Determine the F statistic used to test whether class section is significant variable.

(A) 5.4
(B) 7.3
(C) 7.7
(D) 7.9
(E) 8.3

Problem 8

You are determining the relationship of salary (y) to experience (x_1) for both men ($x_2 = 1$) and woman ($x2 = 0$). You fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

You are given

Sources	Sum of Squares	Degree of Freedom
Regression	330.0117	3
Error	12.8156	7

You also fit the below model to the same observations.

$$y = \gamma_0 + \gamma_1 x_1 + \epsilon$$

You are given

Sources	Sum of Squares	Degree of Freedom
Regression	315.0992	1
Error	27.7281	9

Determine the F-ratio to test whether the linear rellationship between salary and experience is identical for men and women.

(A) 0.6
(B) 2.0
(C) 3.5
(D) 4.1
(E) 6.2

Problem 9

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.

Problem 10

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).

Problem 11

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.

Problem 12 You are given:

i) The random walk model

$$y_t = y_0 + c_1 + c_2 + c_3 + ... + c_t,$$

where $c_i, (i = 1, 2, ..., t)$ 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 $c_1, c_2, ..., c_{10}$ is 2.

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

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

(A) 1
(B) 2
(C) 3
(D) 8
(E) 18

Problem 13 You are given:

i) The random walk model

$$y_t = y_0 + c_1 + c_2 + c_3 + ... + c_t,$$

where $c_i, (i = 1, 2, ..., t)$ denote observations from a white noise process.

ii) The following ten observed values of c_t:

t	1	2	3	4	5	6	7	8	9	10
c_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, 19 $\hat{y}$.

(A) 4/3
(B) 4
(C) 9
(D) 12
(E) 16

Problem 14 A random walk is expressed as

$$y_y = y_{t-1} + c_t$$

where

$E(c_t) = \mu_t$ and $Var(c_t) = \sigma_c^2$

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).

Problem 15 A stationary autoregressive model of order one can be written as

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

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 .

Problem 16 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).

Problem 17 You are given two models:

Model L:

$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

where ϵ_t is a white noise process, for $t = 0, 1, 2, \ldots$

Model M:

$$\begin{aligned} y_t &= y_0 + \mu_c t + u_t \\ c_t &= y_t - y_{t-1} \\ u_t &= \sum_{j=1}^{t} \epsilon_j \end{aligned}$$

where ϵ_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).

Problem 18

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)

Problem 19

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).

Problem 20

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).

Problem 21

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.

I.

II.

III.

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

(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).

Problem 22

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).

Problem 23

You are given the following four pairs of observations:

$x_1 = (-1, 0),\ x_2 = (1, 1),\ x_3 = (2, -1),\ x_4 = (5, 10)$

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

Problem 24

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)

Problem 25 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)

Cluster	Data Point
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

Problem 26 (SRM - Sample Question 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.

Problem 27 (SRM - Sample Question 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).

Problem 28 (SRM - Sample Question 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

Problem 29 (SRM - Sample Question 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

Problem 30 (SRM - Sample Question 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)