1.

$$\bar{x} = \frac{26}{5} = 5.2, \qquad \bar{y} = \frac{68}{5} = 13.6$$

$$b_1 = \frac{156.4}{64.8} = 2.4136$$

$$b_0 = 13.6 - 2.4136(5.2) = 1.0494$$

The answer is A

2.

$$\widehat{b_1} = \frac{183 - 5\left(\frac{15}{5}\right)\left(\frac{47}{5}\right)}{55 - 5\left(\left(\frac{15}{5}\right)^2\right)} = 4.2$$

The answer is B.

3.

$$SSR = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = (-0.75^2) * 1000 = 562.5$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 640$$

$$SSE = 640 - 562.5 = 77.5$$

$$F\ statistic = \frac{562.5}{\frac{77.5}{20 - 1 - 1}} = 130.6 \approx 131$$

The answer is A

4.

Problem 4

$R^2 = 1 - \frac{SSError}{SSTotal}$ or $\frac{SSReg}{SST}$

$SST = \sum y^2 - n(\bar{y})^2 = 81,004 - 8(100) = 1,004$

$SSR = \sum \hat{y}^2 - n(\bar{y})^2 = 80,525 - 8(100) = 525$

$\frac{525}{1,004} = .5229$ D

5. E

6.

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 0 & 0 \\ 1 & 0 & 0 & 2 & 1 \end{bmatrix} * \begin{bmatrix} 2 \\ 3 \\ 4 \\ 6 \\ 10 \end{bmatrix} = \begin{bmatrix} 25 \\ 13 \\ 24 \end{bmatrix}$$

$$(X'X) * (X'Y) = \begin{bmatrix} 2.333 & -1.333 & -1.333 \\ -1.333 & 0.933 & 0.733 \\ -1.333 & 0.733 & 0.933 \end{bmatrix} * \begin{bmatrix} 25 \\ 13 \\ 24 \end{bmatrix} = \begin{bmatrix} 9.004 \\ -3.604 \\ -1.404 \end{bmatrix}$$

$\hat{b}_1 = -3.6$

The answer is B

7.

Using models 1 and 5

$$\frac{\frac{(4508761-2250956)}{2}}{\frac{2250952}{15-3-1}}=5.52$$

The answer is D.

8.

Solution. The fact that house prices are non-negative makes the normal distribution inappropriate, leaving only Options (C) and (D). The link function should be one that ensures that the response mean is non-negative. Among the identity and log links, only the log link has this property. **(Answer: (D))** □

9.

$$\frac{1}{\mu^2}=0.00279-0.001-0.00007(25)=0.00004$$

$$\mu=158.11\approx 160$$

The answer is B.

10.

$$0.3321-2.37(0.518)+4.04(0.26)=0.154985$$

$$\mu=\frac{1}{1+e^{-0.154985}}=0.538\approx 0.54$$

The answer is C.

11.

(Odds ratio calculation)

Solution. The estimated odds ratio equals $e^{4\hat{\beta}_1}=e^{4(1.0286)}=\boxed{61.2155}$. **(Answer: (E))**

12.

Solution. Inputting $\{(y_i, \hat{y}_i)\}_{i=1}^{5}$ into a financial calculator yields $R^2 = r^2 = 0.978494^2 = 0.957451$. Then by (4.3.1), the adjusted R^2 equals

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{5-1}{5-2-1}(1-0.957451) = \boxed{0.9149}. \quad \textbf{(Answer: (A))}$$

13.

Solution. I and III are correct, but not II. Marital status takes only a distinct set of levels such as married, single, widowed. Modeling marital status is a classification problem. **(Answer: (E))** □

14.

Solution. Inputting $\{(x_i, y_i)\}_{i=1}^{8}$ into a financial calculator yields $r = 0.927820$. Thus $R^2 = r^2 = 0.860849$. By (4.3.1), the adjusted R^2 equals

$$R_a^2 = 1 - \frac{n-1}{n-k-1}(1-R^2) = 1 - \frac{8-1}{8-1-1}(1-0.860849) = \boxed{0.8377}. \quad \textbf{(Answer: (A))}$$

□

15.

Solution. The sample mean is

$$\bar{y} = \frac{1 + 1.5 + 1.6 + 1.4 + 1.5 + 1.7}{6} = 1.45.$$

The sum of squares of the whole series is

$$\sum_{t=1}^{6}(y_t - \bar{y})^2 = (-0.45)^2 + 0.05^2 + 0.15^2 + (-0.05)^2 + 0.05^2 + 0.25^2 = 0.295,$$

and the sum of lag-3 cross products is

$$\sum_{t=4}^{6}(y_{t-3} - \bar{y})(y_t - \bar{y}) = (-0.45)(-0.05) + (0.05)(0.05) + 0.15(0.25) = 0.0625.$$

By (6.2.1), the sample autocorrelation at lag 3 is $r_3 = 0.0625/0.295 = \boxed{0.2119}$. **(Answer: (D))** □

16.

$$\bar{c} = 3$$

$$s_c^2 = \frac{1}{9}\sum_{i=1}^{10}(c_i - 3)^2 = \frac{16}{9}$$

$$s_c = \sqrt{\frac{16}{9}} = \frac{4}{3}$$

$$\text{standard error} = s_c\sqrt{l} = \frac{4}{3} * \sqrt{9} = 4$$

The answer is B.

17.

$$\bar{y} = 40$$

$$b_1 = r_1 = \frac{117}{262} = 0.4466$$

$$b_0 = \bar{y}(1 - r_1) = 22.1374$$

$$\bar{e} = 2.7825$$

$$s^2 = \frac{\sum_{t=2}^{6}(e_t - \bar{e})^2}{6 - 3} = \frac{65.9121}{3} = 21.97 \approx 22$$

The answer is C.

18.

The fitted values for the four points of test data are 5.5, 7.2, 12.0, and 10.4 respectively. The mean squared error is

$$\frac{(4 - 5.5)^2 + (10 - 7.2)^2 + (11 - 12.0)^2 + (13 - 10.4)^2}{4} = \boxed{4.4625} \quad \textbf{(A)}$$

19.

The only observations that share the region of $x_1 = 18, x_2 = 11$, are:

(17,4,27)

(20,3,18)

(20,10,24)

Taking the average of the y values:

$$\frac{27 + 18 + 24}{3} = 23$$

The answer is D.

20.

$$\mu = e^{2.05+1.32+0.405} = 2.1$$

The answer is A.

21.

t	y	$\hat{y}$	e	$\left\|\frac{e}{y}\right\|$
17	8	9	-1	0.125
18	12	15	-3	0.25
19	14	18	-4	0.285714
20	22	20	2	0.090909

$$100 * \frac{1}{4} \sum \left|\frac{e}{y}\right| = 18.79$$

The answer is E.

22.

The three clusters have center points of A: (0, 0.75), B: (2, –2), and C: (0, 1.5)

	x	u	DistA	DistB	DistC	New Cluster	Has Changed
A	2	-2	**3.400368**	**0**	**4.031129**	B	1
A	-1	2	**1.600781**	**5**	**1.118034**	C	1
A	-2	1	**2.015564**	**5**	**2.061553**	A	0
A	1	2	**1.600781**	**4.123106**	**1.118034**	C	1
B	4	0	**4.069705**	**2.828427**	**4.272002**	B	0
B	4	-1	**4.366062**	**2.236068**	**4.716991**	B	0
B	0	-2	**2.75**	**2**	**3.5**	B	0
B	0	-5	**5.75**	**3.605551**	**6.5**	B	0
C	-1	0	**1.25**	**3.605551**	**1.802776**	A	1
C	3	8	**7.846177**	**10.04988**	**7.158911**	C	0
C	-2	-2	**3.400368**	**4**	**4.031129**	A	1
C	0	0	**0.75**	**2.828427**	**1.5**	A	1

There are 6 data points that moved clusters.

The answer is C.

23.

[**Section 17.1**] Let ϕ be the loading we are solving for. The other loading is $\sqrt{1-\phi^2}$. We are given

$$4.3077 = 4\phi + 3\sqrt{1-\phi^2}$$

Let's solve for ϕ.

$$(4.3077 - 4\phi)^2 = 9(1-\phi^2)$$
$$18.5563 - 34.4616\phi + 16\phi^2 = 9 - 9\phi^2$$
$$25\phi^2 - 34.4616\phi + 9.5563 = 0$$
$$\phi = 0.9938, \boxed{0.3846} \qquad \textbf{(D)}$$

24.

Solution. We should look for the K observations which are closest to $(X_1, X_2) = (0, 5)$. The closest five observations, in that order, are:

X_1	X_2	Y	Distance to $(X_1, X_2) = (0, 5)$
0	5	Yes	0.0
1	6	No	1.4
2	5	No	2.0
2	3	Yes	2.8
2	7	Yes	2.8

- When $K = 1$, the closest observation has $Y =$ "Yes", so the predicted response is "Yes."
- When $K = 3$, there are two observations with $Y =$ "No" and one observation with $Y =$ "Yes", so the predicted response is "No."
- When $K = 5$, there are three observations with $Y =$ "Yes" and two observations with $Y =$ "No", so the predicted response is "Yes." **(Answer: (A))** □

25.

B

26.

Complete linkage uses the maximum inter-cluster dissimilarity, which happens between (1,1) and (6,6).

$\sqrt{(1-6)^2 + (1-6)^2} = 7.07$

The answer is B.

28.

Solution. I. True. Clustering is an unsupervised learning technique.

II. False. The number of clusters remains the same for the K-means clustering algorithm, but not for the hierarchical clustering algorithm.

III. False. Only the K-means clustering algorithm requires randomization of clusters at the outset. **(Answer: (A))** □

29.

Solution. I. True. Both quantitative and qualitative predictors can be partitioned to form splits.

II. False. A disadvantage of the classification error rate is that it is not sensitive enough to node impurity as it only looks at the maximum of the class proportions.

III. False. A pure node is characterized by a small value (close to zero) of the Gini index or cross-entropy. **(Answer: (A))** □

30.

Since there are 50 observations, the maximum value of K is 50.

The answer is D.

31.

Solution. We first calculate the Euclidean distance between all possible pairs of observations from the two clusters.

$$\begin{aligned} d_{13} &= \sqrt{(-1-6)^2+[1-(-2)]^2} = 7.6158, \\ d_{14} &= \sqrt{(-1-10)^2+(1-5)^2} = 11.7047, \\ d_{23} &= \sqrt{(2-6)^2+[-1-(-2)]^2} = 4.1231, \\ d_{24} &= \sqrt{(2-10)^2+(-1-5)^2} = 10. \end{aligned}$$

With single linkage, we take the minimum pairwise dissimilarities, which is $\boxed{4.1231}$, as the inter-cluster dissimilarity. **(Answer: (A))** □

32.

Solution. I. False. The KNN classier predicts a given test observation $\mathbf{x}_0$ by identifying the K points in the *training* data that are closest to $\mathbf{x}_0$ and using the most commonly-occurring class among these K points as the predicted class.

II. True. As K increases, the classifier becomes less sensitive and flexible, so the training error rate tends to increase.

III. False. As K increases, the test error rate tends to follow a U-shape rather than a monotonically increasing shape. **(Answer: (B))** □

33.

Solution. I. True. The predicted salary for players with less than 5 years of experience (i.e., those sent to the left branch) is 5.1, compared to 6 or 6.7 for those with 5 or more years of experience (i.e., those sent to the right branch).

II. True. All players with less than 5 years of experience share the same predicted salary, regardless of the number of hits they made.

III. True. Players with 5 or more years of experience are further partitioned by the number of hits they made. In other words, Hits has an effect on salary only when Years exceeds 5. By definition, there is an interaction between Years and Hits. **(Answer: (D))** □

34.

Solution. I. False. Neither a random forest nor boosting is a special case of each other. Only bagging is a special case of a random forest (with $m = p$).

II. True, by the design of random forests and boosting.

III. True, because boosting does not involve generating bootstrapped samples, on which out-of-bag estimation relies. **(Answer: (D))** □

35.

Solution. Since the data is centered and scaled, $\frac{1}{4}\sum_{i=1}^{4} x_{ij}^2 = 1$ for $j = 1, 2, 3$, and so $\sum_{j=1}^{3}\sum_{i=1}^{4} x_{ij}^2 = 12$. Then the cumulative PVE by the first two PCs is

$$\begin{aligned}\frac{\sum_{m=1}^{2}\sum_{i=1}^{4} z_{im}^2}{\sum_{j=1}^{3}\sum_{i=1}^{4} x_{ij}^2} &= \frac{(-1.0903)^2 + \cdots + 1.4819^2 + 0.8970^2 + \cdots + 0.9056^2}{12} \\ &= \frac{11.97705}{12} \\ &= \boxed{99.81\%}. \quad \textbf{(Answer: (E))}\end{aligned}$$

□