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 or stays the same as more principal components are added.

Statement II is correct – The cumulative proportion of variance explained increases or stays the same 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 is 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(8/20) + 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**

Adding a constant to the dependent variable avoids the problem of the logarithm of zero being negative infinity. In general, a log transformation may make the variance constant. Hence I is true. 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.

(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. DELETED

18. **Key: B**

$$TSS = (\text{Residual Sum of Squares}) / (1 - R^2) = 230 / (1 - 0.64) = 638.89$$

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 + 0.05 *x*. 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. DELETED

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 $F = \frac{(TSS - RSS)/1}{RSS/(n-2)} = \frac{490,000 - 250,000}{250,000/98} = 94.08$.

45. **Key: C**

Let T be the study time offered. The equation to solve is

$6.0 = \exp[–0.1 + 0.5(2T) + 0.5 – 0.1 + 0.2] = \exp(0.5 + T)$

$1.79176 = 0.5 + T$

$T = 1.29176$ or 129 hours.

46. **Key: D**

The smoothed forecast at 100 is 0.2(100.2) + 0.8(95.1) = 96.1. This is also the forecast at 102.

The smoothed values are at 99: 0.2(95.1) + 0.8(89.9) = 90.9. At 100: 0.2(96.1) + 0.8(90.9) = 91.9. The trend is 0.2(96.1 – 91.9)/.8 = 1.05. The intercept (value at 100) is 2(96.1) – 91.9 = 100.3. The forecast at time 102 is 100.3 + 2(1.05) = 102.4.

The difference is 6.3.

47. DELETED

48. **Key: B**

The first split is $X_1 < t_1$. This requires a horizontal line at t_1 on the vertical axis. Graphs B, C, and E have such a line.

The second split is the case where the first split is true. That means all further action is below the line just described. All three graphs do that. The second split is $x_2 < t_2$. This requires a vertical line at t_2 on the horizontal axis with the line only going up to t_1. Again, all three graphs have this.

The third split is when the second split is true. That means all further action is to the left of the line just described. That rules out graph C. The only difference between graphs B and E is which part relates to node C and which to node D. The third split indicates that node C is the case when $x_1 < t_3$. Only graph B has this region marked as C.

49. **Key: C**

I is true. A confidence interval reflects the error in estimating the expected value. With an infinite amount of data, the error goes to zero.

II is false. A prediction interval also reflects the uncertainty in the predicted observation. That uncertainty is independent of the sample size and thus the interval cannot go to zero.

III is true. The additional uncertainty in making predictions about a future value as compared to estimating its expected value leads to a wider interval.

50. **Key: B**

I is false. Highly flexible models are harder to interpret. For example, a ninth degree polynomial is harder to interpret than a straight line.

II is true. Inference is easier when using simple and relatively inflexible methods. Lasso is simpler and less flexible than bagging.

III is false. Flexible methods tend to overfit the training set and be less accurate when applied to unseen data.

51. **Key: C**

For duck X: Age = 7 => go left. Gender = Male => go right. Prediction = 0.90 kg.

For duck Y: Age = 5 => go left. Gender = Female => go left. Prediction = 0.80 kg.

For duck Z: Age = 8 => go right. Gender = Male => go right. Wing Span = 5.7 => go right. Prediction = 1.25.

Y < X < Z.

52. **Key: D**

I is true. See Page 348 of Frees.

II is true. See Page 347 of Frees.

III is true. See Page 347 of Frees.

53. **Key: A**

(A) is false. β_0 is the expected value of Y when $X = 0$.

The other four statements are true (see Page 63 of James, et al.)

54. **Key: C**

The variables TOPSCHOOL and LARGECITY both lack significance and are candidates for removal. However, only one variable should be removed at a time, and it should be the one with the highest *p*-value, which is TOPSCHOOL. After removing TOPSCHOOL and rerunning the model it is possible that LARGECITY will become significant.

55. **Key: B**

The three one-step predicted values are 12+2.25 = 14.25; 15+2.25 = 17.25; and 21+2.25 = 23.25. The errors are 15 – 14.25 = 0.75; 21 – 17.25 = 3.75; and 22 – 23.25 = –1.25.

$F = (0.75 + 3.75 - 1.25)/3 = 1.083$

$G = (0.75^2 + 3.75^2 + 1.25^2)/3 = 5.396$

The absolute difference is 4.313.

56. **Key: B**

A is false because different models are likely to produce different results.

B is true because using the fitted model implies that this model continues to apply.

C is false as there is no easy way to compare reliability of the two approaches.

D is false in that a narrower interval provides more useful information about the true value.

E is false in that the interval contains the most likely values with the point prediction being the single most likely point.

57. **Key: A**

T1 has observations with $Z \leq 3$ and Y = A or B, which are observations 1, 5, and 9. The values are 4.75, 4.53, and 3.89, which average to 4.39.

T2 has observations with $Z \leq 3$ and Y = C or D, which are observations 3, 4, and 6. The values are 4.67, 4.56, and 3.91, which average to 4.38.

T3 has observations with $Z > 3$ and X = F, which are observations 2 and 7. The values are 4.67 and 3.90, which average to 4.29.

T4 has observations with $Z > 3$ and X = M, which is observation 8. The value is 3.90.

58. **Key: C**

The estimate of β_1 is

$$b_1 = r_1 = \frac{\begin{array}{c}(31-40)(35-40)+(35-40)(37-40)+(37-40)(41-40)\\ +(41-40)(45-40)+(45-40)(51-40)\end{array}}{(31-40)^2+(35-40)^2+(37-40)^2+(41-40)^2+(45-40)^2+(51-40)^2} = \frac{117}{262} = 0.4466.$$

The estimator of β_0 is $b_0 = \bar{y}(1-r_1) = 22.136.$

The residuals are then

$e_2 = 35-(22.136+0.4466\times 31) = -0.9806$

$e_3 = 37-(22.136+0.4466\times 35) = -0.7670$

$e_4 = 41-(22.136+0.4466\times 37) = 2.3398$

$e_5 = 45-(22.136+0.4466\times 41) = 4.5534$

$e_6 = 51-(22.136+0.4466\times 45) = 8.7670.$

The average residual is $\bar{e} = 2.78252$ and then the mean square error is

$$s^2 = \frac{\begin{array}{c}(-0.9806-2.78252)^2+(-0.7670-2.78252)^2+(2.3398-2.78252)^2\\ +(4.5534-2.78252)^2+(8.7670-2.78252)^2\end{array}}{6-3} = 21.969.$$

59. **Key: C**

The means for cluster 1 are (1 + 0 + 6)/3 = 2.3333 for *X*1 and (3 + 4 + 2)/3 = 3 for *X*2 and the variation is

$(1-2.3333)^2+(3-3)^2+(0-2.3333)^2+(4-3)^2+(6-2.3333)^2+(2-3)^2 = 22.6667.$

The means for cluster 2 are (5 + 1)/2 = 3 for *X*1 and (2 + 6)/2 = 4 for *X*2 and the variation is $(5-3)^2+(2-4)^2+(1-3)^2+(6-4)^2 = 16.$

The total within-cluster variation is (per equation (10.12) in the first edition of *ISLR*) 2(22.6667 + 16) = 77.33.

60. **Key: C**

I is false. Setting $K = n$ will almost certainly overfit as there are unlikely to be that many true clusters.

II is false. Within-cluster variation is minimized when $K = n$, which as noted above is unlikely to be optimal.

III is true. There is no exact method for determining the optimal value of K.

61. **Key: D**

I is false, AIC and BIC make an indirect estimate by adjusting the training error.

II is true, for a fixed value of the number of predictors, the two provide the same ranking.

III is true as for $n > 7$, BIC provides a greater penalty for additional variables, and hence will select a number less than equal to that selected by AIC.

62. **Key: A**

Using the rule of thumb t-value of 2 produces and interval of –1.03 ± 2(0.06) for an interval of (–1.15, –0.91). Note that if the more precise value of 1.96 is used, the same answer (to two decimal places) is obtained.

63. **Key: D**

For the first split, ShelveLoc = Good, the answer is no, and the branch to the right is taken. For the second split, Price = 120 ≥ 110, the answer is yes, and the branch to the left is taken. The node has a predicted value of 9.2.

64. **Key: E**

From the definition of a stationary AR(1) process on Page 254 of Frees, the parameter β_0 can be any fixed constant, ruling out answers (A) and (B). To be stationary it is necessary that $-1 < \beta_1 < 1$, which makes answer (E) correct.

65. **DELETED**

66. **Key:D**

To be in Region 1, X_1 must be ≤ 5 and X_2 must be ≤ 40. This eliminates answer C

To be in Region 2, X_1 must be ≤ 5 and X_2 must be > 40. This also eliminates answer C.

To be in Region 3, X_1 must be > 5 and be ≤ 10. This eliminates answers A, B, and E.

That leaves only answer D as possible. To confirm:

To be in Region 4, X_1 must be .> 10 and X_2 must be ≤ 60.

To be in Region 5, X_1 must be .> 10 and X_2 must be > 60.

Both of these are satisfied by answer D.

67. **Key: D**

In the full model, there are 16 outcomes with a probability of 0.8 and 4 outcomes with a probability of 0.2, so the loglikelihood is 16 ln 0.8 + 4 ln 0.2 = –10.01. In the reduced model, the probability of each of the 20 outcomes is 0.5, so the loglikelihood is 20 ln 0.5 = –13.86.

The likelihood ratio test statistic (equal to twice the difference in loglikelihoods) has a chi-square distribution with degrees of freedom equal to the difference in the number of explanatory variables, in this case 1. The test statistic is 2(–10.01 – (–13.86)) = 7.71. The critical values for one degree of freedom are 6.635 and 7.879 for p = 0.010 and 0.005 respectively, so the null hypothesis can be rejected at 0.010 but not 0.005.

68. **Key: E**

The initial residuals are the observations (139, 156, 289). They are updated by subtracting λ times the prediction from the tree. For the three observations the predictions are 125.4, 125.4, and 350.5. The updated residuals are 139 – 0.08(125.4) = 129.0, 156 – 0.08(125.4) = 146.0, and 289 – 0.08(350.5) = 261.0.

69. **Key: C**

The initial fitted values are all zero They are updated by adding λ times the prediction from the tree. For the three observations the predictions are 125.4, 125.4, and 350.5. The updated predictions are 0 + 0.08(125.4) = 10.0, 0 + 0.08(125.4) = 10.0, and 0 + 0.08(350.5) = 28.0.

70. **Key: C**

The AIC formula uses the estimated variance from the full model, $\hat{\sigma}^2 = 9.4582^2 = 89.45755.$ For each model, the AIC is:

Model I: [183,663.30 + 2(1)(89.45755]/100 = 1,836.42

Model II: [8,826.47 + 2(2)(89.45755]/100 = 91.84

Model III: [8,319.59 + 2(6)(89.45755]/100 = 93.93

The model with the smallest AIC should be selected, which is Model II with AIC = 91.84.

Note that the residual standard error did not have to be given. It can be calculated from the residual sum of squares, the sample size, and the number of predictors.

71. **Key: C**

Per Frees (page 197), the estimates will be biased.

72. **Key: C**

Any variable with a *p*-value less than 0.05 is significant. Hence, Age and Log(Income) are the only significant variables.