Final Exam - CSC/DSC 265/465 - May 10, 2018

NAME: --

PLEASE NOTE that all students will do a total of 8 questions.

Undergraduate students do questions 1-8.

Graduate students do questions 3-10.

The exam will last three hours. You are allowed up to five aid sheets on standard 8.5×11 inch paper (both sides) and a calculator. Answer the questions in the space provided. Use the back of the sheet if needed (please indicate if you have done this). Critical values for the t distributions are given in tabular form on the first page of this exam sheet. No other critical values will be needed. All answers must be justified with sufficient detail.

Table 1: Critical values for the t distribution with ν degrees of freedom.

df $= \nu$	$t_{\nu,0.05}$	$t_{\nu,0.025}$
20	1.725	2.086
21	1.721	2.080
22	1.717	2.074
23	1.714	2.069
24	1.711	2.064
25	1.708	2.060
26	1.706	2.056
27	1.703	2.052
28	1.701	2.048
29	1.699	2.045
30	1.697	2.042
35	1.690	2.030
40	1.684	2.021
45	1.679	2.014
50	1.676	2.009
55	1.673	2.004
60	1.671	2.000
65	1.669	1.997
70	1.667	1.994
75	1.665	1.992
80	1.664	1.990
85	1.663	1.988
90	1.662	1.987
95	1.661	1.985
100	1.660	1.984

Q1: [Undergraduate Students Only] A certain classification problem involves 2 classes $j = 1, 2$, and a random observation of the form $X \in \{1, 2, 3, 4\}$. Suppose the prior probabilities π_j of class j are given by $\pi_1 = 1 - \pi_2 = 3/4$. The following table gives the conditional distribution $f(x \mid j)$ of X:

$x =$	1	2	3	4
$f(x \mid j = 1)$	1/2	1/4	1/4	0
$f(x \mid j = 2)$	0	1/3	1/3	1/3

(a) What is the posterior probability of class $j = 1$ given $X = 2$?
(b) Give the prediction made by a Bayes classifier for each outcome $X = 1, 2, 3, 4$. Justify your answers numerically.

SOLUTION:

(a) We have

$$\begin{aligned}
P(j = 1 \mid X = 2) &= \frac{P(X = 2 \mid j = 1)P(j = 1)}{P(X = 2)} \\
&= \frac{f(2 \mid j = 1)\pi_1}{f(2 \mid j = 1)\pi_1 + f(2 \mid j = 2)\pi_2} \\
&= \frac{(1/4) \times (3/4)}{(1/4) \times (3/4) + (1/3) \times (1/4)} \\
&= \frac{(3/4)}{(3/4) + (1/3)} \\
&= \frac{9}{13}.
\end{aligned}$$

(b) The Bayes classifier is given by

$$\hat{j} = \text{argmax}_j h_j(x)$$

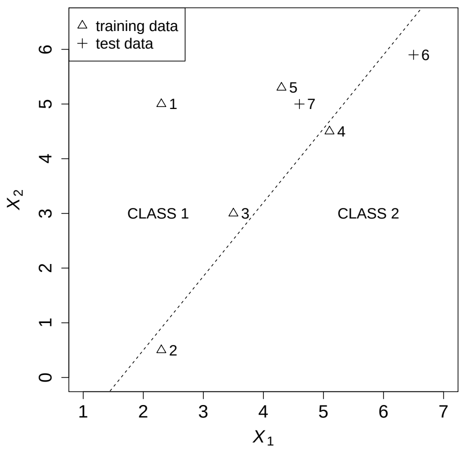
where

$$h_j(x) = f(x \mid j)\pi_j.$$

These values, along with $\hat{j}$, are given in the following table:

$x =$	1	2	3	4
$h_1(x)$	3/8	3/16	3/16	0
$h_2(x)$	0	1/12	1/12	1/12
$\hat{j}$	1	1	1	2

Q2: [Undergraduate Students Only] To build a KNN classifier, the data in the following plot is used, partitioned into training and test data (see the appropriate symbols in the plot legend). As it happens, there are two classes, indicated in the plot by a class boundary (the dashed line). The pairwise distances are also given. By evaluating the classifier with the test data, estimate the classification errors for neighborhood sizes $K = 1$ and $K = 3$. When evaluating a prediction, specify the neighborhood exactly. Note that the KNN classifier itself is built using only the training data.



Pairwise Distance =

	1	2	3	4	5	6
2	4.500					
3	2.332	2.773				
4	2.844	4.883	2.193			
5	2.022	5.200	2.435	1.131		
6	4.295	6.841	4.173	1.980	2.280	
7	2.300	5.054	2.283	0.707	0.424	2.102

SOLUTION: The correct classes for test observations $i = 6, 7$ are $y_i = 2, 1$.

For $K = 1$, observation $i = 6$, the neighborhood is $N = \{4\}$, so $\hat{y}_6 = 2$. For $i = 7$, $N = \{5\}$, $\hat{y}_7 = 1$. This means $CE = 0.0$.

For $K = 3$, observation $i = 6$, the neighborhood is $N = \{3, 4, 5\}$, so $\hat{y}_6 = 1$ (2/3 in N are class 1). For $i = 7$, $N = \{3, 4, 5\}$, $\hat{y}_7 = 1$ (2/3 in N are class 2). This means $CE = 1/2$.

Q3: We are given 2 classes, $j = 1, 2$. The distribution of a single dimensional observation is given by $X \sim N(\mu_j, \sigma_j^2)$, given classes $j = 1, 2$. Available estimates of μ_j are given by $\bar{X}_1 = 102.5$, $\bar{X}_2 = 143.8$. We assume $\sigma_1^2 = \sigma_2^2$, and a pooled estimate of the common variance is given by $s^2_{pooled} = 5.03$. We accept as prior class probabilities $\pi_1 = 0.7, \pi_2 = 0.3$. Suppose an LDA classifier is constructed. Determine the region for X which predicts class $j = 1$.

SOLUTION: For LDA, the classifier is given by

$$\hat{y} = \text{argmax}_j h_j(x)$$

where

$$h_j(x) = x\mu_j/\sigma^2 - \frac{1}{2}\mu_j^2/\sigma^2 + \log(\pi_j).$$

The classification boundary x_b is the solution to $h_1(x_b) = h_2(x_b)$. There is only one, since the $h_j(x)$ are linear. This gives, after substituting the estimates,

$$x_b \times (102.5/5.03) - \frac{1}{2} \times 102.5^2/5.03 + \log(0.7) = x_b \times (143.8/5.03) - \frac{1}{2} \times 143.8^2/5.03 + \log(0.3)$$

or,

$$\begin{aligned} x_b \times \frac{102.5 - 143.8}{5.03} &= -\frac{1}{2} \times \frac{102.5^2 - 143.8^2}{5.03} + \log(0.7/0.3), \\ x_b &= 123.2532, \end{aligned}$$

so that class $y = 1$ is predicted when $X < x_b = 123.2532$.

Q4: Suppose we have $n = 5$ observations of a feature vector. The distances between observations i and j, denoted d_{ij}, are given in the following distance matrix:

	1	2	3	4	5
1	0.000	8.853	9.022	9.540	10.982
2	8.853	0.000	7.803	9.537	10.753
3	9.022	7.803	0.000	9.377	10.562
4	9.540	9.537	9.377	0.000	9.957
5	10.982	10.753	10.562	9.957	0.000

Using the compact agglomeration method, for which cluster distance is defined by

$$D(A, B) = \max_{i \in A, j \in B} d_{ij}$$

for any two clusters A, B, construct a hierarchical cluster for this data. Justify each step precisely. Sketch a dendogram, indicating precisely the height of each node.

SOLUTION: The compact distance between two clusters A and B is

$$D(A, B) \quad = \quad \max_{i \in A, j \in B} d_{ij}.$$

To construct the clustering, we use the following steps:

1. Start with clusters {1},{2},{3},{4},{5}.
2. First join the two nearest observations, which are 2 and 3 ($d_{2,3} = 7.803$). This gives clusters $\{1\}$, $\{4\}$, $\{5\}$ and $\{2, 3\}$ joined at distance 7.80.
3. The cluster distances are now

$$\begin{aligned}
D(\{1\}, \{4\}) &= d_{1,4} = 9.540, \\
D(\{1\}, \{5\}) &= d_{1,5} = 10.982, \\
D(\{4\}, \{5\}) &= d_{4,5} = 9.957, \\
D(\{1\}, \{2, 3\}) &= \max\{d_{1,2}, d_{1,3}\} = \max\{8.853, 9.022\} = 9.022, \\
D(\{4\}, \{2, 3\}) &= \max\{d_{4,2}, d_{4,3}\} = \max\{9.540, 9.377\} = 9.540, \\
D(\{5\}, \{2, 3\}) &= \max\{d_{5,2}, d_{5,3}\} = \max\{10.753, 10.562\} = 10.753.
\end{aligned}$$

   The smallest cluster distance is $D(\{1\}, \{2, 3\}) = 9.022$, so combine clusters $\{1\}$ and $\{2, 3\}$. This gives clusters $\{1, 2, 3\}$, $\{4\}$ and $\{5\}$, joined at distance 9.022.
4. The cluster distances are now

$$\begin{aligned}
D(\{1, 2, 3\}, \{4\}) &= \max\{d_{1,4}, d_{2,4}, d_{3,4}\} = \max\{9.540, 9.537, 9.377\} = 9.540, \\
D(\{1, 2, 3\}, \{5\}) &= \max\{d_{1,5}, d_{2,5}, d_{3,5}\} = \max\{10.982, 10.753, 10.562\} = 10.982, \\
D(\{4\}, \{5\}) &= \max\{d_{4,5}\} = \max\{9.96\} = 9.96.
\end{aligned}$$

   The smallest cluster distance is $D(\{1, 2, 3\}, \{4\}) = 9.540$, so combine clusters $\{1, 2, 3\}$ and $\{4\}$. This gives clusters $\{1, 2, 3, 4\}$ and $\{5\}$, joined at distance 9.540.
5. Join the remaining two clusters, at cluster distance

$$D(\{1, 2, 3, 4\}, \{5\}) = \max\{d_{1,5}, d_{2,5}, d_{3,5}, d_{4,5}\} = \max\{10.982, 10.753, 10.562, 9.957\} = 10.982.$$

This gives the dendogram shown in Figure 1.

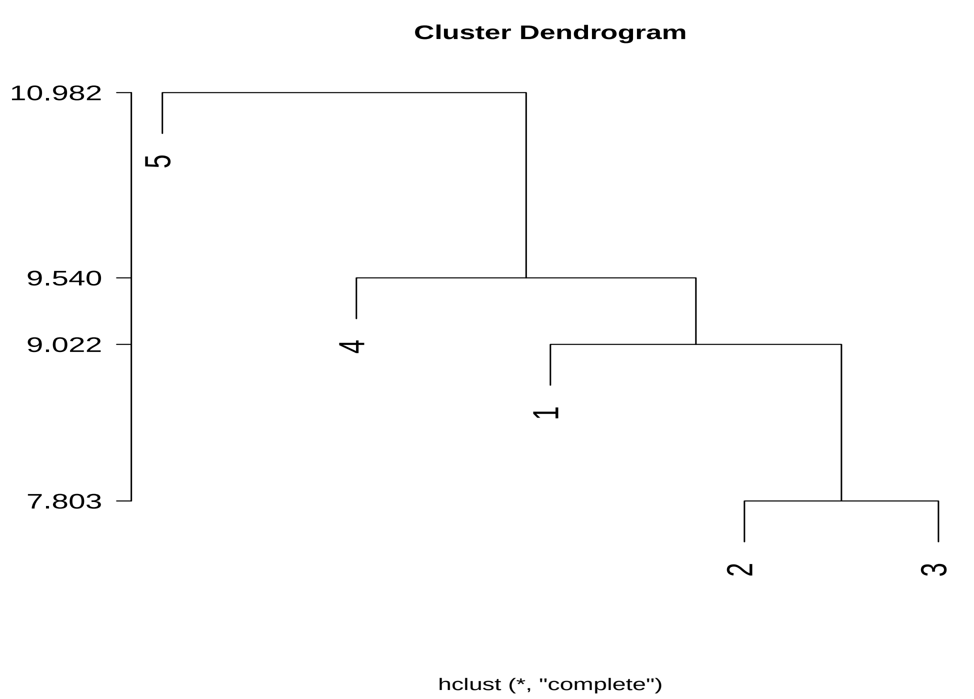


Figure 1: Dendogram for Question **Q4**.

Q5: Suppose in an unsupervised learning application we are given observations $\dot{x}_1, \dots, \dot{x}_n$. Recall the *within cluster sum of squares*, for K clusters $A_1, \dots, A_K$ where d is a distance function and $g(A_i)$ is a cluster centroid:

$$SS_{within} = \sum_{i=1}^{K} \sum_{j \in A_i} d(\dot{x}_j, g(A_i))^2.$$

A K-means clustering algorithm was applied to the data, allowing the number of clusters K to vary from 1 to 6. The following table gives the separate sum of squares within each cluster:

	1	2	3	4	5	6
1	38608.0	-	-	-	-	-
2	258.3	4911.1	-	-	-	-
3	501.9	258.3	218.6	-	-	-
4	191.6	112.6	94.8	258.3	-	-
5	53.5	42.1	94.8	191.6	112.6	-
6	42.1	77.3	40.9	53.5	112.6	38.8

Let R^2 be the proportion of total variation explained by the clustering. If we accept as the number of clusters the smallest value of K for which $R^2 \geq 95\%$, what is this number?

SOLUTION: The total sum of squares SS_{total} is simply the SS for the $K = 1$ model, so

$$SS_{total} = 38608.0.$$

Otherwise, SS_{within} is the sum of the individual cluster sums of squares. Then

$$R^2 = 1 - \frac{SS_{within}}{SS_{total}}.$$

This gives, for $K = 1, 2, 3$:

$$\begin{aligned} R^2[1] &= 1 - \frac{SS_{total}}{SS_{total}} = 0, \\ R^2[2] &= 1 - \frac{258.3 + 4911.1}{38608.0} = 0.866, \\ R^2[3] &= 1 - \frac{501.9 + 258.3 + 218.6}{38608.0} = 0.975. \end{aligned}$$

The smallest number of clusters that yield at least 95% variation explained is $K = 3$.

Q6: We wish to fit a model of the form

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent error terms, and $x_i \in [10, 20]$ is a predictor variable. We consider the following six models

M1 $g(x) = \beta_1 x$, where β_1 is to be estimated.

M2 $g(x) = \beta_0 + \beta_1 x$, where β_0, β_1 are to be estimated.

M3 $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, where $\beta_0, \beta_1, \beta_2$ are to be estimated.

M4 $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, where $\beta_0, \beta_1, \beta_2, \beta_3$ are to be estimated.

M5 $g(x)$ is a continuous piecewise linear spline with 1 knot at $\xi = 13$.

M6 $g(x)$ is a natural cubic spline with 2 knots at $\xi = 15, 17$ (note that $g(x)$ is continuous, and possesses continuous derivatives, at each knot).

The relevant SSE values are given in the following table. The sample size is $n = 181$. Which model is preferred based on the BIC score (use form $BIC = n \log(SSE/n) + C$)?

Model	SSE
M1	607.807
M2	32.163
M3	14.116
M4	8.707
M5	6.263
M6	9.523

SOLUTION: The equation is

$$BIC \quad = \quad n \log(SSE/n) + \log(n)k,$$

where k is the number of parameters. Other than σ^2, the number of parameters is

M1 β_1, $k = 1$.

M2 β_0, β_1, $k = 2$.

M3 $\beta_0, \beta_1, \beta_2$, $k = 3$.

M4 $\beta_0, \beta_1, \beta_2, \beta_3$, $k = 4$.

M5 4 parameters with one constraint, so $k = 4 - 1 = 3$.

M6 2+4+2 parameters with 4 constraints, so $k = 8 - 4 = 4$.

The number of parameters does not include σ^2, but if this was included the model selection procedure would be unchanged, since we would simply add 1 to each k.

We can construct table:

		Without σ^2		With σ^2	
Model	SSE	k	BIC	k	BIC
M1	607.807	1	224.455	2	229.653
M2	32.163	2	-302.313	3	-297.115
M3	14.116	3	-446.171	4	-440.973
M4	8.707	4	-528.429	5	-523.231
M5	6.263	3	-593.260	4	-588.062
M6	9.523	4	-512.217	5	-507.018

So, model **M5** has the lowest BIC, and is therefore the preferred model.

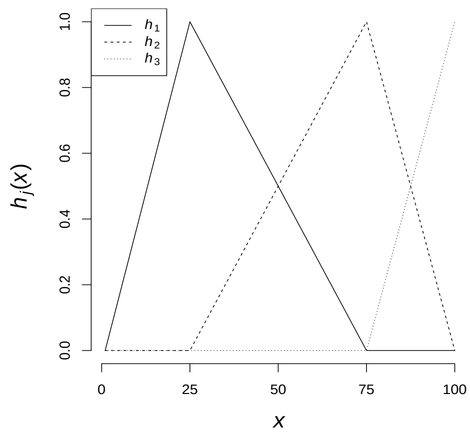
Q7: We wish to fit a model of the form

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent error terms, and $x_i \in [0, 100]$ is a predictor variable. We will assume that $g(x)$ is a continuous linear spline with two knots at $\xi = 25, 75$. One way to do this is to use the basis functions

$$b_1(x) = x; \;\; b_2(x) = (x - 25)I\{x > 25\}; \;\; b_3(x) = (x - 75)I\{x > 75\},$$

then set $g(x) = \beta_0 + \sum_{j=1}^{3} \beta_j b_j(x)$. Suppose we then consider alternative basis functions $h_j(x)$, $j = 1, 2, 3$ shown in the following graph:



Each $h_j(x)$ is a continuous piecewise linear spline with $h_j(0) = 0$. The maximum of each $h_j(x)$ on the range $x \in [0, 100]$ is 1, and the discontinuites in slope occur at the knots $\xi = 25, 75$. Note that in the plot the functions overlap at various places on the horizontal axis. We then set $g(x) = \beta_0^* + \sum_{j=1}^{3} \beta_j^* h_j(x)$.

(a) Write explicitly each basis function $h_j(x)$, $j = 1, 2, 3$ as a linear combination of the functions $b_1(x), b_2(x), b_3(x)$.
(b) Suppose we use multiple linear regression to estimate the coefficients β_j using basis functions b_1, b_2, b_3. Suppose then that we use multiple linear regression to estimate the coefficients β_j^* using basis functions h_1, h_2, h_3. Show that the fitted values will be identical.

SOLUTION:

(a) If we write

$$h(x) = \alpha_1 b_1(x) + \alpha_2 b_2(x) + \alpha_3 b_3(x)$$

then $h(0) = 0$, since $b_j(0) = 0$ for $j = 1, 2, 3$. Clearly, $h(x)$ is also a linear spline with knots $\xi = 25, 75$. In addition, the slope of $h(x)$ is α_1 for $x < 25$, $\alpha_1 + \alpha_2$ for $x \in (25, 75)$, and $\alpha_1 + \alpha_2 + \alpha_3$ for $x > 75$.

Then note that the slope of $h_1(x)$ is 1/25 for $x < 25$, -1/50 for $x \in (25, 75)$, and 0 for $x > 75$. Therefore, if

$$h_1(x) = \alpha_1 b_1(x) + \alpha_2 b_2(x) + \alpha_3 b_3(x)$$

then we must have $\alpha_1 = 1/25$, $\alpha_2 = -1/50 - \alpha_1 = -3/50$, $\alpha_3 = 0 - \alpha_1 - \alpha_2 = 1/50$.

Next, the slope of $h_2(x)$ is 0 for $x < 25$, 1/50 for $x \in (25, 75)$, and -1/25 for $x > 75$. Therefore, $\alpha_1 = 0$, $\alpha_2 = 1/50$, $\alpha_3 = -1/25 - \alpha_2 = -3/50$.

Finally, $h_3(x) = (1/25) \cdot b_3(x)$. To summarize:

$$\begin{aligned} h_1(x) &= (1/25) \cdot b_1(x) - (3/50) \cdot b_2(x) + (1/50) \cdot b_3(x) \\ h_2(x) &= 0 \cdot b_1(x) + (1/50) \cdot b_2(x) - (3/50) \cdot b_3(x) \\ h_3(x) &= 0 \cdot b_1(x) + 0 \cdot b_2(x) + (1/25) \cdot b_3(x). \end{aligned}$$

(b) The easiest approach is to note that the two sets of basis functions are related by a linear transformation:

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = \begin{bmatrix} 1/25 & -3/50 & 1/50 \\ 0 & 1/50 & -3/50 \\ 0 & 0 & 1/25 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

From Part (a) we have shown that any function $h_j(x)$ is a linear combination of the basis functions b_1, b_2, b_3. Since the linear transformation is clearly invertible, any function $b_j(x)$ is a linear combination of the basis functions h_1, h_2, h_3. Therefore, each set of basis functions span the same function space. This in turn implies that the least squares estimate of $g(x)$ will be the same using either set of basis functions.

Q8: We are given paired observations (x_i, y_i), $i = 1, \ldots, n$. We wish to fit a model of the form

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent error terms, and x_i is a predictor variable. We decide to use locally weighted linear regression based on kernel density:

$$\phi(x) = \begin{cases} 1 + x & ; \;\; x \in [-1, 0) \\ 1 - x & ; \;\; x \in [0, 1] \\ 0 & ; \;\; \text{elsewhere} \end{cases} .$$

The following table gives a partial listing of the data (sorted in increasing order of x_i):

i	x_i	y_i
⋮	⋮	
11	3.5	18.56
12	3.9	21.34
13	4.6	23.45
14	5.3	22.72
15	5.9	28.51
16	6.2	27.67
⋮	⋮	

Suppose we wish to calculate estimate $\hat{g}(4.7)$ of $g(4.7)$. Write the weighted sum of squares which must be minimized in order to do this. Give the numerical values of any weights used.

SOLUTION: To evaluate $\hat{g}(x)$ at $x = 4.7$ the weighted sum of squares is

$$SSE_x = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2,$$

where

$$w_i = \phi(x_i - 4.7), \;\; i = 1, \ldots, n.$$

Then w_i is nonzero only if $|x_i - 4.7| < 1$. Noting that the data are sorted in increasing order of x_i, this occurs only for $i = 12, 13, 14$. We then have

$$\begin{aligned} w_{12} &= \phi(3.9 - 4.7) = 1 + 3.9 - 4.7 = 0.2, \\ w_{13} &= \phi(4.6 - 4.7) = 1 + 4.6 - 4.7 = 0.9, \\ w_{14} &= \phi(5.3 - 4.7) = 1 - 5.3 - 4.7 = 0.4. \end{aligned}$$

Q9: [Graduate Students Only] A logistic regression model is used to model $P(Y = 1)$ for some binary response variable Y. It depends on two predictors, a quantitative predictor `x` and the indicator variable `i.class`. The following logistic regression model is used:

$$P(Y = 1) = \frac{e^{\eta}}{1 + e^{\eta}}, \text{ where } \eta = \beta_0 + \beta_1 \texttt{x} + \beta_2 \texttt{i.class} + \beta_3 \texttt{x} \times \texttt{i.class}.$$

Using data with sample size $n = 94$, the following coefficient estimates were obtained. The estimated covariance matrix for the estimated coefficients in vector form $[\hat{\beta}_0, \ldots \hat{\beta}_3]^T$ is given immediately following.

```
>
> ### coefficient estimates
>
> summary(fit)$coef
              Estimate Std. Error   z value    Pr(>|z|)
(Intercept) -1.0488571  0.7736844 -1.355665 0.175205679
x            0.7183279  0.2540876  2.827087 0.004697354
i.class      1.3787316  0.9861359  1.398115 0.162078447
x:i.class   -0.9788835  0.2823500 -3.466915 0.000526468
>
> ### estimated covariance matrix
>
> summary(fit)$cov.scaled
            (Intercept)           x    i.class   x:i.class
(Intercept)   0.5985876 -0.15894404 -0.5985876  0.15894404
x            -0.1589440  0.06456053  0.1589440 -0.06456053
i.class      -0.5985876  0.15894404  0.9724639 -0.22173994
x:i.class     0.1589440 -0.06456053 -0.2217399  0.07972153
```

(a) Carry out a hypothesis test for null hypothesis H_o and alternative hypothesis H_a given by:

$$\begin{aligned} H_o &: P(Y = 1) \text{ is not an increasing function of } \texttt{x} \text{ for fixed } \texttt{i.class} = 0, \text{ against} \\ H_a &: P(Y = 1) \text{ is an increasing function of } \texttt{x} \text{ for fixed } \texttt{i.class} = 0. \end{aligned}$$

Use a t-statistic based on the appropriate degrees of freedom. Use significance level $\alpha = 0.05$.

(b) Carry out a hypothesis test for null hypothesis H_o and alternative hypothesis H_a given by:

$$\begin{aligned} H_o &: P(Y = 1) \text{ is not a decreasing function of } \texttt{x} \text{ for fixed } \texttt{i.class} = 1, \text{ against} \\ H_a &: P(Y = 1) \text{ is a decreasing function of } \texttt{x i.class} = 1 \text{for fixed } \texttt{i.class} = 1. \end{aligned}$$

Use a t-statistic based on the appropriate degrees of freedom. Use significance level $\alpha = 0.05$.

SOLUTION:

(a) The required hypothesis test is

$$\begin{aligned} H_o &: \beta_1 \leq 0, \text{ against} \\ H_a &: \beta_1 > 0. \end{aligned}$$

From the coefficient table we have estimate and standard deviation $\hat{\beta}_1 = 0.7183279$, $S = 0.2540876$, giving t-statistic

$$T = \frac{\hat{\beta}_1}{S} = \frac{0.7183279}{0.2540876} = 2.827088.$$

There are $p = 4$ coefficients, so the appropriate degrees of freedom is $n - 4 = 90$. We reject H_o if $T > t_{90,0.05} = 1.662$. Therefore, we reject H_o, and conclude that $P(Y = 1)$ is an increasing function of `x` for fixed `i.class` $=$ 0.

(b) When `i.class` $= 1$, the slope of η is $\beta_1 + \beta_3$. The required hypothesis test is therefore

$$\begin{aligned} H_o &: \beta_1 + \beta_3 \geq 0, \text{ against} \\ H_a &: \beta_1 + \beta_3 < 0. \end{aligned}$$

The estimate of $\beta_1 + \beta_3$ is

$$\hat{\beta}_1 + \hat{\beta}_3 = 0.7183279 - 0.9788835 = -0.2605556.$$

To calculate the standard error of $\hat{\beta}_1 + \hat{\beta}_3$ we need the standard errors S_1, S_3 of $\hat{\beta}_1$ and $\hat{\beta}_3$, and the estimated covariance S_{13}. From the estimated covariance matrix we have

$$\begin{aligned} S_1^2 &= 0.06456053 \\ S_3^2 &= 0.07972153 \\ S_{13} &= -0.06456053 \end{aligned}$$

The standard error S_+ of $\hat{\beta}_1 + \hat{\beta}_3$ is then given by

$$S_+^2 = S_1^2 + S_3^2 + 2S_{13} = 0.06456053 + 0.07972153 - 2 \times 0.06456053 = 0.015161.$$

The t-statistic is then

$$T = \frac{\hat{\beta}_1}{S} = \frac{\hat{\beta}_1 + \hat{\beta}_3}{S_+} = \frac{-0.2605556}{0.015161^{1/2}} = \frac{-0.2605556}{0.12313} = -2.116102.$$

There are $p = 4$ coefficients, so the appropriate degrees of freedom is $n - 4 = 90$. We reject H_o if $T < t_{90,0.05} = 1.662$. Therefore, we reject H_o, and conclude that $P(Y = 1)$ is a deacreasing function of `x` for fixed `i.class` $= 1$.

Q10: [Graduate Students Only] Suppose we are given an $n \times 3$ matrix $\mathbf{X}$, with columns defining 3 standardized predictors x_1, x_2, x_3. The three principal components are then calculated, and given in form

$$PC_j = a_{1j}x_1 + a_{2j}x_2 + a_{3j}x_3, \;\; j = 1, 2, 3.$$

Suppose the matrix of variable loadings a_{ij} is given, in part, by

$$A = \begin{bmatrix} 1/2 & a_{12} & a_{13} \\ 1/2 & a_{22} & a_{23} \\ a_{31} & 1/\sqrt{2} & a_{33} \end{bmatrix}$$

Determine all values of the variable loadings a_{ij} left unspecified. For convenience, you may assume $a_{31} > 0$ and $a_{13} > 0$.

SOLUTION: **1st PC:** The sum of squares of each column of A equals 1. Therefore

$$a_{31}^2 = 1 - (1/2)^2 - (1/2)^2 = 1/2.$$

Since $a_{31} > 0$ we must have $a_{31} = 1/\sqrt{2}$.

2nd PC: The columns of A are orthogonal. This means

$$a_{12}/2 + a_{22}/2 + 1/2 = 0.$$

In addition,

$$a_{12}^2 + a_{22}^2 + 1/2 = 1.$$

Substition gives

$$(1 + a_{22})^2 + a_{22}^2 + 1/2 = 1,$$

or equivalently,

$$2a_{22}^2 + 2a_{22} + 1/2 = 2(a_{22} + 1/2)^2 = 0.$$

The unique solution is $a_{22} = -1/2$. Then substituting gives $a_{12} = -1/2$.

3rd PC: The columns of A are mutually orthogonal. This means

$$\begin{aligned} a_{13}/2 + a_{23}/2 + a_{33}/\sqrt{2} &= 0 \\ -a_{13}/2 + -a_{23}/2 + a_{33}/\sqrt{2} &= 0. \end{aligned}$$

Adding the equations gives $2a_{33}/\sqrt{2} = 0$, or $a_{33} = 0$. This then implies $a_{13} = -a_{23}$. If $a_{13} > 0$, and the sum of squares of each column equals, we must then have $a_{13} = -a_{23} = 1/\sqrt{2}$.

To summarize, we then have

$$A = \begin{bmatrix} 1/2 & -1/2 & 1/\sqrt{2} \\ 1/2 & -1/2 & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix}$$