Final - Exam

Problem 1

You are given the following summary statistics:

$$\begin{aligned}\sum x &= 40\\ \sum y &= 91\\ \sum (x_i - \bar{x})^2 &= 40\\ \sum (x_i - \bar{x})(y_i - \bar{y}) &= 59\\ \sum (y_i - \bar{y})^2 &= 214.8\end{aligned}$$

Determine the equation of the regression line, using the least squares method.

(A) $y = 9.338 + 1.475x$

(B) $y = -9.338 + 1.419x$

(C) $y = -9.338 - 1.475x$

(D) $y = 9.338 - 1.475x$

(E) The correct answer is not given by (A), (B), (C), or (D).

Problem 2

Two actuaries are analyzing dental claims for a group of $n = 20$ participants. The predictor variable is sex, with 0 and 1 as possible values.

Actuary 1 uses the following regression model:

$$Y = \beta + \epsilon$$

Actuary 2 uses the following regression model:

$$Y = \beta_0 + \beta_1 \times Sex + \epsilon$$

Given $R^2 = .8$. Calculate the F-statistic to test whether the model of Actuary 2 is a significant improvement over the model of Actuary 1.

(A) 72
(B) 79
(C) .8
(D) 85
(E) 68

Problem 3

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.

Problem 4

You are given the following data

y	x_1	x_2
2	1	1
3	1	1
4	2	2
6	3	2
8	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

$$(X'X)^{-1} = \begin{bmatrix} 0.75 & -0.25 & -0.25 \\ -0.25 & 0.1944 & -0.0278 \\ -0.25 & -0.0278 & 0.3611 \end{bmatrix}$$

Determine $\widehat{\beta}_1$.

(A) -2.003

(B) -0.753

(C) -0.533

(D) 1.083

(E) 1.753

Problem 5 You fit a multiple linear regression to a data of 20 observation and 4 predictors. You have determined that the coefficient of determination of the model is 0.9. Calculate the F-statistics to test the significant of the model.

(A) 45.75

(B) 33.75

(C) 125.75

(D) 15.75

(E) 19.75

Problem 6

The following two models were fit to 50 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	110	15
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

Sarah performs a regression of the return on a mutual fund (y) on five predictors plus an intercept. She uses monthly returns over 200 months. Her software calculates the F statistic for the regression as $F = 50$, 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) $R^2 = 0.61$

Problem 8

A statistician uses logistic regression to model a probability of success of a random variable. You are given

- There is one predictors and an intercept in the model
- The estimates of success at $x = 0$ and $x = 1$ are 0.1 and 0.4, respectively.

Calculate $\widehat{\beta}_1$ the estimated slope of the model.

Problem 9

You are given the following information for a GLM of customer retention

Response variable	Retention
Response distribution	Binomial
Link	Logit

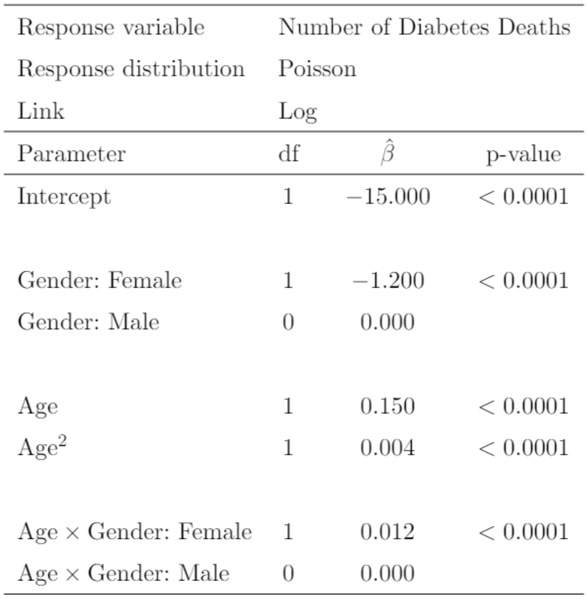
Parameter	df	$\hat{\beta}$
Intercept	1	1.530
Number of Drivers	1	
1	0	0.000
>1	1	0.735
Last Rate Change	2	
$< 0\%$	0	0.000
0%-10%	1	-0.031
$> 10\%$	1	-0.372

Calculate the probability of retention for a policy with 1 drivers and a prior rate changes of 10%.

(A) Less than 0.85
(B) At least 0.85, but less than 0.87
(C) At least 0.87, but less than 0.89
(D) At least 0.89, but less than 0.91
(E) At least 0.91

Problem 10

You are given the follow.



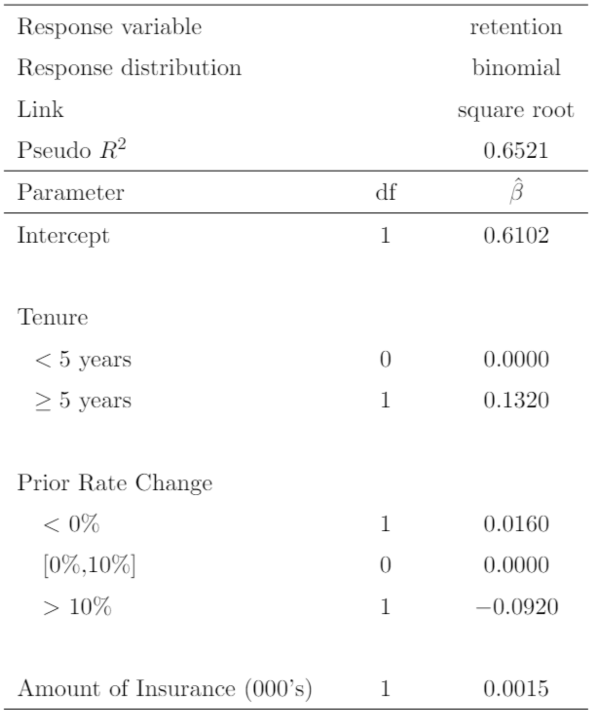
Response variable	Number of Diabetes Deaths		
Response distribution	Poisson		
Link	Log		
Parameter	df	$\hat{\beta}$	p-value
Intercept	1	-15.000	< 0.0001
Gender: Female	1	-1.200	< 0.0001
Gender: Male	0	0.000	
Age	1	0.150	< 0.0001
Age^2	1	0.004	< 0.0001
Age × Gender: Female	1	0.012	< 0.0001
Age × Gender: Male	0	0.000	

Calculate the predicted number deaths for a population of 300,000 males age 25

(A) Less than 3
(B) At least 3, but less than 5
(C) At least 5, but less than 7
(D) At least 7, but less than 9
(E) At least 9

Problem 11

You are given the following output of an GLM.



Response variable		retention
Response distribution		binomial
Link		square root
Pseudo R^2		0.6521
Parameter	df	$\hat{\beta}$
Intercept	1	0.6102
Tenure		
< 5 years	0	0.0000
≥ 5 years	1	0.1320
Prior Rate Change		
$< 0\%$	1	0.0160
[0%,10%]	0	0.0000
$> 10\%$	1	-0.0920
Amount of Insurance (000's)	1	0.0015

Calculate the probability of a policy with 2 years of tenure that experienced at a 10% prior rate increase and has 250,000 in amount of insurance will retain into the next policy term.

(A) Less than 0.6
(B) At least 0.6, but less than 0.7
(C) At least 0.7, but less than 0.8
(D) At least 0.8, but less than 0.9
(E) At least 0.9

Problem 12 (Similar to Sample - Question 4)

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
y_t	2	5	8	15	18

iii) $y_0 = 0$

Calculate the 9 step-ahead forecast, $\hat{y}_{10}$.

Problem 13 (SImilar to Sample - Question 46)

A time series was observed at times 0, 1, ..., 100. The last four observations along with estimates based on exponential and double exponential smoothing with $w = 0.7$ are:

Time (t)	97	98	99	100
Observation (y_t)	96.9	98.1	99.0	100.2
Estimates ($\hat{s}^{(}1)_t$)	93.1	94.1	95.1	
Estimates ($\hat{s}^{(}2)_t$)	88.9	89.9		

All forecasts should be rounded to one decimal place and the trend should be rounded to three decimal places.

Let F be the predicted value of y_{102} using exponential smoothing with $w = 0.8$.

Let G be the predicted value of y_{102} using double exponential smoothing with $w = 0.8$.

Calculate the absolute difference between F and G, $F - G$

(A) Less than 2
(B) At least 2, but less than 4
(C) At least 4, but less than 6
(D) At least 6, but less than 8
(E) At least 8

Problem 14 (Similar to Sample - Question 58)

You are given the following six observed values of the autoregressive model of order one time series

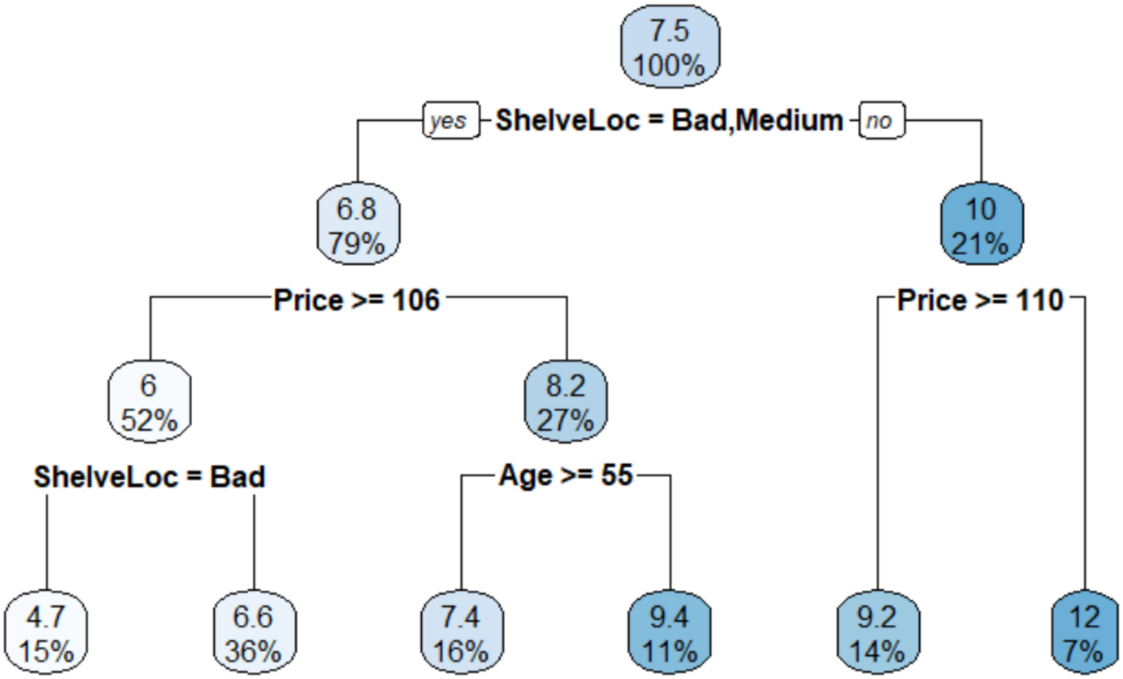
$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t, \text{ with } Var(\epsilon_t) = \sigma^2.$$

t	1	2	3	4	5	6
y_t	1	5	7	11	15	21

Calculate the conditional least squares estimators of β_0 and β_1.

Problem 15 (Similar Sample - Question 63)

You have constructed the following regression tree predicting unit sales (in thousands) of car seats. The variable ShelveLoc has possible values Good, Medium, and Bad



Variable	Observed Value
ShelveLoc	Bad
Price	100
Age	60
Advertising	12

Determine the predicted unit sales (in thousands) for the above observation based on the regression tree.

Problem 16 (Sample - Question 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 50 observations with 12 lapses and one node has 50 observations with 18 lapses.
- Split 2: One node has 40 observations with 10 lapses and one node has 60 observations with 20 lapses.

The total Gini-Index 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 split is better based on total Gini-Index.

Problem 17

X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

Using the RSS to decide the best split among

- Split 1: Region 1 $X_1 < 4$, Region 2 $X_1 \geq 4$
- Split 2: Region 1 $X_2 < 2$, Region 2 $X_2 \geq 2$

Problem 18

Given the below data of an insurance policy.

	Age	Sex	Claim
A	27	M	0
B	30	F	1
C	80	F	1
D	50	M	0
E	60	F	0
F	70	F	1

Let G be a female of 35 years old. Use 3NN to predict if G claim the policy (`Claim =1`).

Problem 19

Given the following data, use 1NN and 3NN to predict the salary for G (a female of 25 years old).

	Age	Sex	Salary (k)
A	27	M	80
B	30	F	70
C	80	F	90
D	50	M	60
E	60	F	10
F	70	F	100

Problem 20

Given the following data, use 1NN and 3NN with weighted distance to predict the salary for G (a female of 25 years old).

	Age	Sex	Salary (k)
A	27	M	80
B	30	F	70
C	80	F	90
D	50	M	60
E	60	F	10

	Age	Sex	Salary (k)
F	70	F	100

Problem 21 (SRM - Sample Question 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	(1, 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, 1)
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.

Problem 22 (SRM - Sample Question 59)

You apply 2-means clustering to a set of five observations with two features. You are given the following initial cluster assignments:

X_1	X_2	Initial cluster
1	3	A
0	4	A
6	2	B

X_1	X_2	Initial cluster
5	2	B
1	0	B

Calculate the total within-cluster variation of the initial cluster assignments, based on Euclidean distance measure.

Problem 23 (SRM - Sample Question 1)

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 single linkage and Euclidean distance. Calculate the intercluster dissimilarity between x_1, x_2 and x_4.

Problem 24 (SRM - Sample Question 1)

You are given the following four pairs of observations:

$A = (-1, 0),\ B = (1, 1),\ C = (2, -1),\ D = (0, 0)$

A hierarchical clustering algorithm is used with single linkage and Euclidean distance. Suppose we want to group the data into three clusters. Determine the points of each clusters.

Problem 25 (SRM - Sample Question 1)

You are given the following data

x	y
1	1
2	2
3	5
4	7

The principal loading matrix are given as follows.

	PC1	PC2
x	0.42	-0.91
y	0.91	0.42

Calculate the first Principal Component Score.