Midterm - 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) $y = 2.456 + 1.475x$

(B) $y = 6.4 + 1.475x$

(C) $y = -6.4 + 1.475x$

(D) $y = -6.4 - 1.475x$

(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 $n = 200$ 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) 712
(B) 792
(C) 0.8
(D) 859
(E) 688

Problem 3

Peter observes the following coffee prices in his company cafeteria:

- 1 bagel for 1.00 (USD)
- 2 bagel for 1.50 (USD)

The cafeteria announces that they will begin to sell any amount of bagels for a price that is the value predicted by a simple linear regression using least squares of the current prices.

With the new pricing model, how much Peter would save if he bought 10 bagels instead of 5 bagels twice?

(A) It would cost him more so he would not save any money.
(B) It would cost the same.
(C) He would save 0.5 (USD)
(D) He would save 1 (USD)
(E) He would save 1.5 (USD)

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
10	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) -2.00
(B) -0.75
(C) -0.5
(D) 0.5
(E) 1.75

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 0.8. Calculate the F-statistics to test the significant of the model.

(A) 45
(B) 85
(C) 125
(D) 155
(E) 195

Problem 6

The following two models were fit to 20 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 100 months. Her software calculates the F statistic for the regression as $F = 30$, 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 $x = 1$ and $x = 2$ are 0.3 and 0.4, respectively.

Calculate $\hat{\beta}_1$ 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 10 drivers and a prior rate changes of 7%.

(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 $\times$ Gender: Female	1	0.012	< 0.0001
Age $\times$ Gender: Male	0	0.000	

Calculate the predicted number deaths for a population of 200,000 females 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 15% prior rate increase and has 150,000 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 (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	6	7	8	9	10
y_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, $\widehat{y}_{19}$.

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

Problem 13 (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 $w = 0.8$ are:

Time (t)	97	98	99	100
Observation (y_t)	96.9	98.1	99.0	100.2
Estimates $(\widehat{s^{(}}1)_t)$	93.1	94.1	95.1	
Estimates $(\widehat{s^{(}}2)_t)$	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 $w = 0.8$.

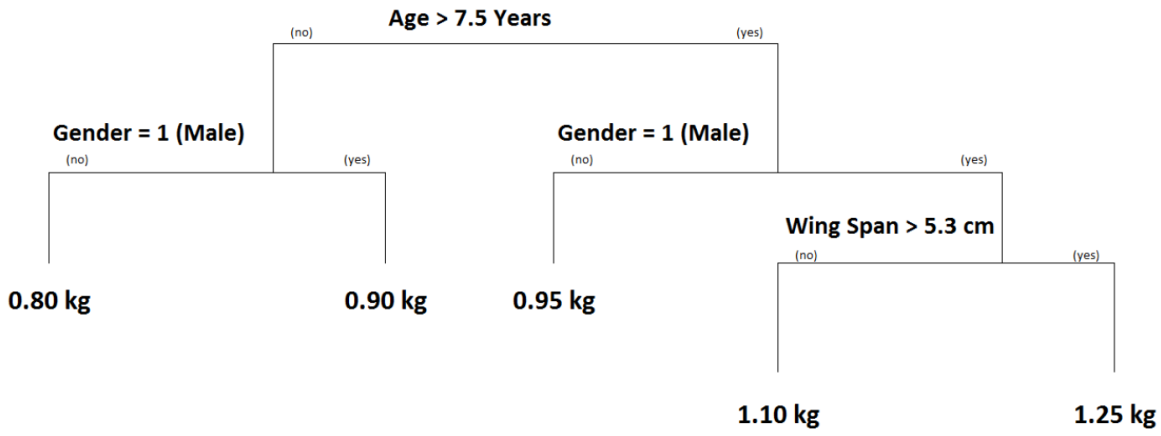
Let G be the predicted value of y_{102} using double exponential smoothing with $w = 0.8$.

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

(A) 0.0
(B) 2.1
(C) 4.2
(D) 6.3
(E) 8.4

Problem 14 (Sample - Question 51)

You are given the following regression tree predicting the weight of ducks in kilograms (kg):



You predict the weight of the following three ducks:

X: Wing Span = 5.5 cm, Male, Age = 7 years Y: Wing Span = 5.8 cm, Female, Age = 5 years Z: Wing Span = 5.7 cm, Male, Age = 8 years

Determine the order of the predicted weights of the three ducks.

(A) X < Y < Z
(B) X < Z < Y
(C) Y < X < Z
(D) Y < Z < X
(E) Z < X < Y

Problem 15 (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 20 observations with 12 lapses and one node has 80 observations with 8 lapses.
- Split 2: One node has 10 observations with 8 lapses and one node has 90 observations with 15 lapses.

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 split is better based on total entropy
