**STAT 46700/ CS 59000 Topics in Data Science Spring 2025**

**Midterm Test – Take home Due: March 12, 2025**

**Time: 11:59 PM(CST)**

Please work ***individually*** and provide detail solution with R code embedded with your answers.

I affirm that I didn’t give or receive any unauthorized help on this exam and that all work is my own [***Vaishak Balachandra***]

**PUID: *0037831852***

**Q.N. 1)** In order to investigate how well one can predict and estimate deep abdominal adipose tissue (AT) from the knowledge of the waist circumference Despres et al . (1991) collected a data from men between the ages of 18 and 42 years who were free from metabolic disease that would require treatment.The variable waist measurement will be used as a predictor variable. The dataset (**MIdtermQ1data)** is provided with this assignment in the Brightspace. Waist circumference (cm), X, and Deep Abdominal AT, Y, of a sample of men are analyzed.

1. Import the data and determine how many men are included in this study.
2. Fit a simple linear regression model. Please be sure to state the equation of the model and display the fitted line with the scattered plot.
3. What is the value of the coefficient of determination? Please provide its interpretation.
4. What is the predicted value of AT for a waist circumference of 105cm?
5. Provide a 90% confidence interval and prediction interval of your estimated value in (d).

> # a

> Q1 <- read.table("C:/Users/PNW\_checkout/Downloads/sem 2/0. Coursework/0. Coursework/Data science/Vaishak\_Balachandra\_Midterm/MidterrmQ1.txt", header = T)

> head(Q1)

ID X Y

1 1 74.75 25.72

2 2 103.00 129.00

3 3 108.00 217.00

4 4 72.60 25.89

5 5 80.00 74.02

6 6 100.00 140.00

> dim(Q1)

[1] 109 3

> names(Q1)

[1] "ID" "X" "Y"

> attach(Q1)

> cat("109 mens are included in the given dataset")

109 mens are included in the given dataset

>

>

> # b

> plot(X,Y, main = "Scatterplot of Y against X", pch = 17, col = "maroon", col.main = "orange", col.lab = "darkgreen", xlab = "Waist circumference (cm)", ylab = "Deep Abdominal AT")

> model1 <- lm(Y~X)

> model1

Call:

lm(formula = Y ~ X)

Coefficients:

(Intercept) X

-215.981 3.459

> cat("Linear Fitted Model Equation:

+ Y = -215.981 + 3.459\*X

+ AT = -215.981 + 3.459\*Waist")

Linear Fitted Model Equation:

Y = -215.981 + 3.459\*X

AT = -215.981 + 3.459\*Waist

> abline(model1, lwd = 2, col = "purple")

>

>

> # c

> summary(model1)

Call:

lm(formula = Y ~ X)

Residuals:

Min 1Q Median 3Q Max

-107.288 -19.143 -2.939 16.376 90.342

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -215.9815 21.7963 -9.909 <2e-16 \*\*\*

X 3.4589 0.2347 14.740 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.06 on 107 degrees of freedom

Multiple R-squared: 0.67, Adjusted R-squared: 0.667

F-statistic: 217.3 on 1 and 107 DF, p-value: < 2.2e-16

> cat("Coefficient of Determination (R squared value) = 0.67 = 67%")

Coefficient of Determination (R squared value) = 0.67 = 67%

> cat("Inference: It means only 67% of the variability in Y(AT) is defined by X(Waist Circumference)")

Inference: It means only 67% of the variability in Y(AT) is defined by X(Waist Circumference)

>

>

>

> # d

> predict(model1, data.frame(X = 105))

1

147.1987

> cat("Here, For the waist circumference equals to 105cm, the AT is found to be 147.1987")

Here, For the waist circumference equals to 105cm, the AT is found to be 147.1987

> # e

> predict(model1, data.frame(X = 105), interval = "confidence", level = 0.9)

fit lwr upr

1 147.1987 139.8762 154.5213

> cat("Confidence Interval: [139.8762, 154.5213]")

Confidence Interval: [139.8762, 154.5213]

> predict(model1, data.frame(X = 105), interval = "pred", level = 0.9)

fit lwr upr

1 147.1987 91.85026 202.5472

> cat("Confidence Interval: [91.85026, 202.5472]")

Confidence Interval: [91.85026, 202.5472]

A graph of a graph

AI-generated content may be incorrect.

###################################################################################################

**Q.N.2)** The dataset (**Passenger**) provided with this assignment is sample data obtained from the accident data from data.gov which contains passenger’s age and the speed of the vehicle(mph) at the time of impact and the fate of the passengers (1 -survived, 0- did not survive) after the crash.

a) Fit a logistic regression model. Please be sure to write the equation of the fitted model.

b) Find the probability that a 35-year-old passenger will survive if there was a crash of a car speeding at 80mph?

> # 2

> # a

> install.packages("readxl")

> library(readxl)

> Q2 <- read\_excel("Passangers.xlsx")

> head(Q2)

# A tibble: 6 × 4

ID Age Speed Survived

*<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 1 22 65 0

2 2 38 50 1

3 3 26 45 1

4 4 35 55 1

5 5 35 85 0

6 6 26 117 0

> dim(Q2)

[1] 20 4

> names(Q2)

[1] "ID" "Age" "Speed" "Survived"

> attach(Q2)

> model2 <- glm(Survived~Age+Speed, family = "binomial")

> model2

Call: glm(formula = Survived ~ Age + Speed, family = "binomial")

Coefficients:

(Intercept) Age Speed

7.56052 0.05207 -0.14679

Degrees of Freedom: 19 Total (i.e. Null); 17 Residual

Null Deviance: 27.73

Residual Deviance: 13.32 AIC: 19.32

> cat("Logistic Fitted Model Equation:

+ Survived = [1 + exp(-7.56052 - 0.05207\*Age + 0.14679\*Speed)]^(-1)")

Logistic Fitted Model Equation:

Survived = [1 + exp(-7.56052 - 0.05207\*Age + 0.14679\*Speed)]^(-1)

>

>

> # b

> predict(model2, data.frame(Age= 35, Speed = 80), type = "response")

1

0.08625105

> cat("Probability that a 35-year-old passenger will survive if there was a crash of a car speeding at 80mph is: 0.08625105 = 8.625105%")

Probability that a 35-year-old passenger will survive if there was a crash of a car speeding at 80mph is: 0.08625105 = 8.625105%

###################################################################################################

**Q.N. 3)**  For risk management purposes, a credit card companies wants to predict the likelihood of their customers missing a payment in a given month. A publicly available data involving a cross-sectional sample of 30,000 customers from a major credit company in Taiwan are provided with this assignment (Credit\_data). (If the last digit of your PUID is less than 5 please use the first 15000 data otherwise use the last 15000 data).

1. Import the data in R and print the variable names.
2. Fit a simple logistic regression model using MISSED\_PAYMENT (1-Yes, 0-No) as the binary response and BILL\_AMT1 as a predictor variable. Please write the model equation and display the fitted model on the scatterplot.
3. Create the confusion matrix to assess the classification accuracy (assume that probabilities exceeding 0.5 to a predicted missed payment in your model)
4. Fit a multiple logistic regression model using all possible predictors and assess the classification accuracy.

> # 3

>

> # a

> Q3\_initial <- read.csv("Credit\_data.csv")

> head(Q3\_initial)

ID LIMIT\_BAL SEX EDUCATION MARRIAGE AGE PAY\_0 PAY\_2 PAY\_3 PAY\_4 PAY\_5 PAY\_6 BILL\_AMT1

1 1 20000 2 2 1 24 2 2 -1 -1 -2 -2 3913

2 2 120000 2 2 2 26 -1 2 0 0 0 2 2682

3 3 90000 2 2 2 34 0 0 0 0 0 0 29239

4 4 50000 2 2 1 37 0 0 0 0 0 0 46990

5 5 50000 1 2 1 57 -1 0 -1 0 0 0 8617

6 6 50000 1 1 2 37 0 0 0 0 0 0 64400

BILL\_AMT2 BILL\_AMT3 BILL\_AMT4 BILL\_AMT5 BILL\_AMT6 PAY\_AMT1 PAY\_AMT2 PAY\_AMT3 PAY\_AMT4

1 3102 689 0 0 0 0 689 0 0

2 1725 2682 3272 3455 3261 0 1000 1000 1000

3 14027 13559 14331 14948 15549 1518 1500 1000 1000

4 48233 49291 28314 28959 29547 2000 2019 1200 1100

5 5670 35835 20940 19146 19131 2000 36681 10000 9000

6 57069 57608 19394 19619 20024 2500 1815 657 1000

PAY\_AMT5 PAY\_AMT6 MISSED\_PAYMENT

1 0 0 1

2 0 2000 1

3 1000 5000 0

4 1069 1000 0

5 689 679 0

6 1000 800 0

> dim(Q3\_initial)

[1] 30000 25

> names(Q3\_initial)

[1] "ID" "LIMIT\_BAL" "SEX" "EDUCATION" "MARRIAGE"

[6] "AGE" "PAY\_0" "PAY\_2" "PAY\_3" "PAY\_4"

[11] "PAY\_5" "PAY\_6" "BILL\_AMT1" "BILL\_AMT2" "BILL\_AMT3"

[16] "BILL\_AMT4" "BILL\_AMT5" "BILL\_AMT6" "PAY\_AMT1" "PAY\_AMT2"

[21] "PAY\_AMT3" "PAY\_AMT4" "PAY\_AMT5" "PAY\_AMT6" "MISSED\_PAYMENT"

> attach(Q3\_initial)

>

> # Since, my PUID ends with 2 <5, I'm choosing the first 15000 rows

> Q3 <- Q3\_initial[1:15000, ]

> dim(Q3)

[1] 15000 25

>

> # b

> plot(Q3$BILL\_AMT1, Q3$MISSED\_PAYMENT, pch = "x", col = "maroon",main = "Scatterplot of Missed Payment against Bill Amnt 1", col.main = "orange", col.lab = "red", cex = 0.75, xlab = "Bill Amount 1", ylab = "Missed Payment")

> model3 <- glm(MISSED\_PAYMENT ~ BILL\_AMT1, data = Q3, family = "binomial")

> summary(model3)

Call:

glm(formula = MISSED\_PAYMENT ~ BILL\_AMT1, family = "binomial",

data = Q3)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.202e+00 2.372e-02 -50.671 <2e-16 \*\*\*

BILL\_AMT1 -6.967e-07 2.824e-07 -2.468 0.0136 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 16000 on 14999 degrees of freedom

Residual deviance: 15994 on 14998 degrees of freedom

AIC: 15998

Number of Fisher Scoring iterations: 4

> cat("Logistic Fitted Model Equation: MISSED\_PAYMENT = [1 + exp(-(1.202 + 0.0000006967\*BILL\_AMT1))]^(-1)")

Logistic Fitted Model Equation: MISSED\_PAYMENT = [1 + exp(-(1.202 + 0.0000006967\*BILL\_AMT1))]^(-1)

> curve(predict(model3, newdata = data.frame(BILL\_AMT1 = x), type = "response"), col = "blue", add = TRUE)

>

> # c

> # summary(model3)

> p = predict(model3, data = Q3, type = 'response')

> pp =ifelse(p > 0.5, 1, 0)

> # install.packages("caret")

> library(caret)

> confusionMatrix(data = factor(pp), reference = factor(Q3$MISSED\_PAYMENT), positive = "1")

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 11623 3377

1 0 0

Accuracy : 0.7749

95% CI : (0.7681, 0.7815)

No Information Rate : 0.7749

P-Value [Acc > NIR] : 0.5046

Kappa : 0

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.0000

Specificity : 1.0000

Pos Pred Value : NaN

Neg Pred Value : 0.7749

Prevalence : 0.2251

Detection Rate : 0.0000

Detection Prevalence : 0.0000

Balanced Accuracy : 0.5000

'Positive' Class : 1

> cat("Accuracy = 0.7749 = 77.49%")

Accuracy = 0.7749 = 77.49%

>

> # d

> names(Q3)

[1] "ID" "LIMIT\_BAL" "SEX" "EDUCATION" "MARRIAGE"

[6] "AGE" "PAY\_0" "PAY\_2" "PAY\_3" "PAY\_4"

[11] "PAY\_5" "PAY\_6" "BILL\_AMT1" "BILL\_AMT2" "BILL\_AMT3"

[16] "BILL\_AMT4" "BILL\_AMT5" "BILL\_AMT6" "PAY\_AMT1" "PAY\_AMT2"

[21] "PAY\_AMT3" "PAY\_AMT4" "PAY\_AMT5" "PAY\_AMT6" "MISSED\_PAYMENT"

> model3a <- glm(MISSED\_PAYMENT~., family = "binomial", data = Q3)

> model3a

Call: glm(formula = MISSED\_PAYMENT ~ ., family = "binomial", data = Q3)

Coefficients:

(Intercept) ID LIMIT\_BAL SEX EDUCATION MARRIAGE AGE

-5.317e-01 1.814e-06 -2.799e-07 -8.916e-02 -1.207e-01 -1.895e-01 3.475e-03

PAY\_0 PAY\_2 PAY\_3 PAY\_4 PAY\_5 PAY\_6 BILL\_AMT1

5.461e-01 4.594e-02 8.254e-02 4.619e-03 8.237e-02 -2.019e-02 -7.360e-06

BILL\_AMT2 BILL\_AMT3 BILL\_AMT4 BILL\_AMT5 BILL\_AMT6 PAY\_AMT1 PAY\_AMT2

4.647e-06 -5.036e-07 3.016e-07 1.488e-06 5.894e-07 -1.873e-05 -6.773e-06

PAY\_AMT3 PAY\_AMT4 PAY\_AMT5 PAY\_AMT6

-6.278e-06 -2.961e-06 -1.959e-06 -1.471e-06

Degrees of Freedom: 14999 Total (i.e. Null); 14975 Residual

Null Deviance: 16000

Residual Deviance: 14330 AIC: 14380

> p1 = predict(model3a, data = Q3, type = 'response')

> pp1 =ifelse(p1 > 0.5, 1, 0)

> library(caret)

> confusionMatrix(data = factor(pp1), reference = factor(Q3$MISSED\_PAYMENT), positive = "1")

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 11339 2694

1 284 683

Accuracy : 0.8015

95% CI : (0.795, 0.8078)

No Information Rate : 0.7749

P-Value [Acc > NIR] : 1.352e-15

Kappa : 0.2381

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.20225

Specificity : 0.97557

Pos Pred Value : 0.70631

Neg Pred Value : 0.80802

Prevalence : 0.22513

Detection Rate : 0.04553

Detection Prevalence : 0.06447

Balanced Accuracy : 0.58891

'Positive' Class : 1

> cat("Accuracy = 0.8015 = 80.15%")

Accuracy = 0.8015 = 80.15%

|  |
| --- |
| > cat("The curve in the plot appears like a stright line, as the effect size of BILL\_AMT1 on the probability of missed payment is very small (i.e., coefficient is 0.0000006967), meaning the relationship is weak.")  The curve in the plot appears like a stright line, as the effect size of BILL\_AMT1 on the probability of missed payment is very small (i.e., coefficient is 0.0000006967), meaning the relationship is weak. |
|  |
| |  | | --- | |  | |

A graph with numbers and lines

AI-generated content may be incorrect.

###################################################################################################

**Q.N. 4)** Consider the data related to the entering high-school students who make program choices among general programs, vocational programs and academic programs. Their choices might be modeled using their reading, writing, math, science scores and social economic status. The data sets are available with this assignment (**program\_data**).

a) Create a KNN-based classifier for program choice using the variables ses,prog,read, write, math, science, socst of students. (Make sure that you have normalize the data and use 80% data for training and 20% for testing. Please be sure to use set.seed() function for reproducibility of your results.

b) Evaluate the classifier’s accuracy in predicting which academic program the student will be joining.

> # 4

> # a

> library(readxl)

> Q4 <- read\_excel("program\_data.xlsx")

> head(Q4)

# A tibble: 6 × 13

id female ses schtyp prog read write math science socst honors awards cid

*<dbl>* *<chr>* *<chr>* *<chr>* *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<chr>* *<dbl>* *<dbl>*

1 45 female low public vocation 34 35 41 29 26 not enrolled 0 1

2 108 male middle public general 34 33 41 36 36 not enrolled 0 1

3 15 male high public vocation 39 39 44 26 42 not enrolled 0 1

4 67 male low public vocation 37 37 42 33 32 not enrolled 0 1

5 153 male middle public vocation 39 31 40 39 51 not enrolled 0 1

6 51 female high public general 42 36 42 31 39 not enrolled 0 1

> names(Q4)

[1] "id" "female" "ses" "schtyp" "prog" "read" "write" "math" "science"

[10] "socst" "honors" "awards" "cid"

> dim(Q4)

[1] 150 13

> set.seed(2467)

> rnum <- sample(1:nrow(Q4))

> Q4 <- Q4[rnum,]

> Q4$ses <- as.numeric(as.factor(Q4$ses))

> prog\_factor <- as.factor(Q4$prog)

> # min-max normalization

> normalize <- function(x){

+ return ((x-min(x))/(max(x)-min(x)))

+ }

> Q4\_norm <- as.data.frame(lapply(Q4[,c("ses", "read", "write", "math", "science", "socst")], normalize))

> head(Q4\_norm)

ses read write math science socst

1 0.0 0.5945946 0.3235294 0.5151515 0.2173913 0.7777778

2 1.0 0.7297297 0.2352941 0.7272727 0.5869565 0.4444444

3 1.0 0.7837838 0.7647059 0.8181818 0.6956522 0.6666667

4 1.0 0.2972973 0.6764706 0.6363636 0.5869565 0.3333333

5 0.5 0.2972973 0.6764706 0.1818182 0.4565217 0.2222222

6 0.0 0.5135135 0.6176471 0.3030303 0.4782609 0.7777778

> # Splitting the dataset

> set.seed(2467)

> train\_index <- sample(1:nrow(Q4\_norm), 0.8 \* nrow(Q4\_norm))

> train\_data <- Q4\_norm[train\_index,]

> test\_data <- Q4\_norm[-train\_index,]

> train\_labels <- prog\_factor[train\_index]

> test\_labels <- prog\_factor[-train\_index]

> # To find optimum k value

> set.seed(2467)

> k\_values <- 10:20

> accuracies <- numeric(length(k\_values))

> for(i in 1:length(k\_values)) {

+ knn\_pred <- knn(train = train\_data, test = test\_data, cl = train\_labels, k = k\_values[i])

+ accuracies[i] <- sum(knn\_pred == test\_labels) / length(test\_labels)

+ }

> plot(k\_values, accuracies, type="b", xlab="k", ylab="Accuracy", main="Accuracy for Different k Values")

> cat("Best k:", 15)

Best k: 15

>

>

> # b

> library(class)

> knn\_pred <- knn(train = train\_data, test = test\_data, cl = train\_labels, k = 15)

> confusion\_matrix <- table(Predicted = knn\_pred, Actual = test\_labels)

> print(confusion\_matrix)

Actual

Predicted academic general vocation

academic 13 2 1

general 0 2 1

vocation 3 3 5

> accuracy <- sum(knn\_pred == test\_labels) / length(test\_labels)

> cat("Accuracy: 66.67%")

Accuracy: 66.67%

A graph with numbers and lines

AI-generated content may be incorrect.