Exam 1 B Solutions - MATH 4322

Fall 2023

Name: **PSID**:

Instructions

- Allow one sheet of notes front and back to be turned in for extra credit.
- Allow calculator.
- Total possible points 100.
- For multiple choice circle your answer on this test paper.
- For short answer questions answer fully on this test paper, partial credit will be given.
- Once completed turn in to TA or instructor.
- Data sets are coming from

UCI Machine Learning Repository and `R`.

Problem 1

(28 possible points) Among the certified rice grown in TURKEY, the Osmancik species, which has a large planting area since 1997 and the Cammeo species grown since 2014 have been selected for the study. We have four features to predict the type of rice.

The features are as follows:

- *Area*: Returns the number of pixels within the boundaries of the rice grain
- *Perimeter*: Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain
- *Major_Axis*: The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives
- *Convex*: Returns the pixel count of the smallest convex shell of the region formed by the rice grain
- The response variable - *Class*: Cammeo ($Y = 0$) and Osmancik ($Y = 1$)

The name of the data set is called **rice**.

a. Is this a inference or prediction statistical learning problem?

This is a prediction learning problem

b. Is this a regression or classification problem?

This is a classification problem

c. Give the model formula for our problem. Use the variable names in the formula.

$$P(X) = \frac{exp(\beta_0 + \beta_1 \times Area + \beta_2 \times Perimeter + \beta_3 \times Major\_Axis + \beta_4 \times Convex)}{1 + exp(\beta_0 + \beta_1 \times Area + \beta_2 \times Perimeter + \beta_3 \times Major\_Axis + \beta_4 \times Convex)}$$

or

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 \times Area + \beta_2 \times Perimeter + \beta_3 \times Major\_Axis + \beta_4 \times Convex$$

d. Give the R code to get the model for predicting the probability of **Osmancik**, given the features. That is, $P(\text{Class} = 1|X)$.

rice.glm = glm(Class ~ Area + Perimeter + Major_Axis + Convex, family = binomial,data = rice)

e. The following is the output from the data. Write out the equation with the estimates.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	45.3667	6.5070	6.97	0.0000
Area	0.0049	0.0020	2.51	0.0121
Perimeter	0.0701	0.0462	1.52	0.1293
Major_Axis	-0.3545	0.0542	-6.54	0.0000
Convex	-0.0055	0.0023	-2.46	0.0139

$$\hat{P(Class} = 1|X) = \frac{exp(45.3667 + 0.0049 \times Area + 0.0701 \times Perimeter - 0.3545 \times Major\_Axis - 0.0055 \times Convex)}{1 + exp(45.3667 + 0.0049 \times Area + 0.0701 \times Perimeter - 0.3545 \times Major\_Axis - 0.0055 \times Convex)}$$

or

$$log\left(\frac{p(X)}{1-p(x)}\right) = 45.3667 + 0.0049 \times Area + 0.0701 \times Perimeter - 0.3545 \times Major\_Axis - 0.0055 \times Convex$$

f. Give the predicted probability of the type of rice being *Osmanick*, given Area = 15000, Perimeter = 425, Major_Axis = 175, and Convex = 15000. Given these values, which type of rice are we predicting?

$$45.3667 + 0.0049 \times 15000 + 0.0701 \times 425 - 0.3545 \times 175 - 0.0055 \times 15000 = 4.1217$$

$$P(Class = 1|X) = \frac{exp(4.1217)}{1 + exp(4.1217)} = 0.9840419$$

We predict this type of rice to be Osmanick.

g. The following is the output from `R`. Determine R^2 and give an interpretation.

Null deviance:	2087.44	on	1523	degrees of freedom
Residual deviance:	516.84	on	1519	degrees of freedom

$$R^2 = 1 - \frac{516.84}{2087.44} = 0.7524049$$

This tells us how good of a fit we have to determine which type of rice.

Problem 2

(36 points) An MRI was conducted to several patients to determine what features are related to atrophy.

Response - atrophy: a measure of loss of neurons estimated by the degree of ventricular enlargement relative to the predicted ventricular size; a continuous value with 0 indicating no atrophy and 100 indicating the most severe degree of atrophy

Predictors

- *age*: participant age at time of MRI, in years.
- *sex*: participant sex (male or female). female = 0 and male = 1.
- *weight*: participant's weight at time of MRI, in pounds.
- Data is called **mri**.

a. Is this a inference or prediction statistical learning problem?

This is an inference learning problem

b. Is this a regression or classification problem?

This is a regression problem

c. Give the model formula for our problem. Use the variable names in the formula.

$$atrophy = \beta_0 + \beta_1 \times age + \beta_2 \times sex + \beta_3 \times weight + \epsilon, \epsilon \sim N(0,1)$$

or

$$atrophy = \begin{cases} \beta_0 + \beta_1 \times age + \beta_3 \times weight + \epsilon, & \text{if female} \\ \beta_0 + \beta_1 \times age + \beta_2 + \beta_3 \times weight + \epsilon, & \text{if male} \end{cases}, \epsilon \sim N(0,1)$$

d. Give the R code to get the model with `atrophy` as the response variable and the other 3 features as the input variables.

mri.lm = lm(atrophy ~ age + sex+weight,data = mri)

e. The following is the output from the data. Write out the equation with the estimates.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15.3709	9.9163	-1.5501	0.1219
age	0.6140	0.1129	5.4364	0.0000
sexMale	4.8190	1.3443	3.5846	0.0004
weight	0.0165	0.0221	0.7467	0.4557

$$\hat{atrophy} = \begin{cases} -15.3709 + 0.614 \times age + 0.0165 \times weight, & \text{if female} \\ -10.5519 + 0.614 \times age + 0.0165 \times weight, & \text{if male} \end{cases}$$

f. Give the interpretation of the coefficient for the variable `age`.
For fixed values of sex and weight, the atrophy will increase on average by 0.614 for an increase of one year.

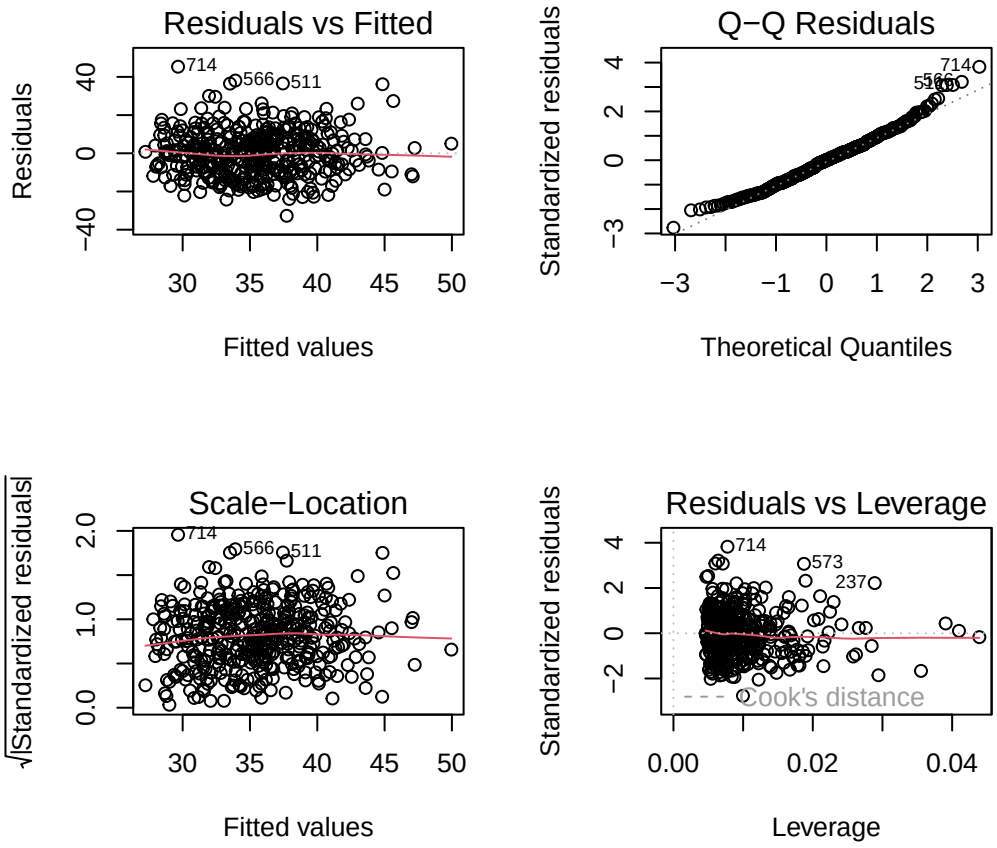
g. Are there any variables that are not needed in this model? Justify your answer.

Yes, weight is not needed. Since the p-value is larger than 0.05, we determine that weight is not significant in predicting atrophy, given age and sex in the model.

h. What are the assumptions of this model?

Linear, Independent samples, Equal variance of the residuals for each value of X, Normal distribution of the residuals. Also assume no extreme values.

i. The plot below are the diagnostics plots. Are any of the assumptions violated with this model?

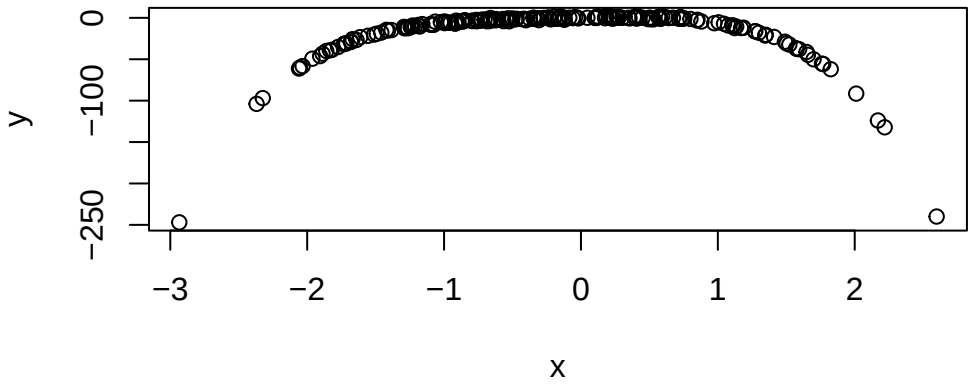


There does not appear to be an violations of the data.

Problem 3

(12 possible points)

a. Using the following plot below do we have a linear relationship?



This is NOT a linear relationship

b. We will use a 10-fold cross validation for a regression model with degrees 1, 2, 3 and 4 respectively. What is the calculation for the MSE of the cross validations?

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

c. What is the correct function and library to use for a 10-fold cross validation?

Library uses boot, **Function is cv.glm**

d. The following is the cross validation output for a regression model with degree 1, 2, 3 and 4 respectively, based on the data represented from the plot above. According to these values, write out the formula for the best model.

	Degree 1	Degree 2	Degree 3	Degree 4
MSE	1070.80	205.01	305.51	1.00

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$$

Problem 4

(8 points) The famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for flowers from each of 3 species of iris. We would like a model to predict the species *Iris setosa*, *versicolor*, and *virginica*.

a. Circle the best model to use for this example.

i. Simple Linear Regression

ii. Logistic Regression

iii. Multiple Linear Regression

iv. Linear Discriminat Analysis (LDA)

v. Polynomial Regression

b. The following is the confusion matrix based on the model. What is the error rate?

	setosa	versicolor	virginica
setosa	39	0	0
versicolor	0	39	1
virginica	0	0	41

i. 0.0083

ii. 1

iii. 0.975

iv. 0.9917

v. 1

Problem 5

(4 points) The following is a 95% prediction interval for `atrophy` from problem 2, with only `age` and `sex` as the predictors. We want to predict `atrophy` for a male that is 70 years old. Which statement is correct?

	fit	lwr	upr
1	35.43	11.98	58.88

a. For one male that is 70 years old, we predict the `atrophy` to be between 11.98 and 58.88 with 95% confidence.

b. On average for all males that 70 years old, we we predict the `atrophy` to be between 11.98 and 58.88 with 95% confidence.

c. For one male that regardless of the age, we predict the `atrophy` to be between 11.98 and 58.88 with 95% confidence.

d. On average for all males regardless of the age, we we predict the `atrophy` to be between 11.98 and 58.88 with 95% confidence.

e. For a male that is 70 years old, the `atrophy` will be 35.43.

Problem 6

(4 points) The following is the ANOVA table from problem 2, where $n = 412$ using `age` and `sex` to predict `atrophy`. What is the AIC statistic?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	3783.93	3783.93	26.77	0.0000
sex	1	2878.19	2878.19	20.36	0.0000
Residuals	409	57822.81	141.38		

a. 919.6082

b. 806.884

c. 3783.9318

d. 2878.1892

e. 2042.9755

Problem 7

(4 points) What value can we calculate to determine if there is **multicolinearity**, a situation in which two or more predictor variables are closely related to one another?

a. VIF

b. C_p

c. BIC

d. Adjusted R^2

e. AIC

Problem 8

(4 points) Which stepwise selection begins with a model containing all of the predictors, and then iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.

a. forward

b. backward

c. best subset

d. none of these