Exam 1 C - MATH 4322 Solutions

Instructor - Cathy Poliak

Fall 2022

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

Problem 1

(36 possible points) We want to understand how the input variables relate to miles per gallon, `mpg`. The input variables are:

- `cylinders` - as qualitative 4, 6 or 8
- `displacement` - cubic inches
- `horsepower` - gross horsepower
- `weight` - per 1000 pounds

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

(3 points)

- Inference

b. Is this a regression or classification problem?

(3 points)

- Regression

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

(4 points)

- acknowledged that cylinders is categorical
- Need error term
- linear regression model
- the distribution of the error term

$$\text{mpg} = \begin{cases} \beta_0 + \beta_3 \times \text{displaement} + \beta_4 \times \text{horsepower} + \beta_5 \times \text{weight} + \epsilon & \text{if cylinders is 4} \\ \beta_0 + \beta_1 + \beta_3 \times \text{displaement} + \beta_4 \times \text{horsepower} + \beta_5 \times \text{weight} + \epsilon & \text{if cylinders is 6} \\ \beta_0 + \beta_2 + \beta_3 \times \text{displaement} + \beta_4 \times \text{horsepower} + \beta_5 \times \text{weight} + \epsilon & \text{if cylinders is 8} \end{cases}$$

$$\epsilon \sim N(0, \sigma^2)$$

d. Give the R code to get the model for predicting the `mpg` based on the 4 input variables.

(4 points)

- lm
- mpg first
- additive of the predictors
- summary function

```
mpg.fit = lm(mpg ~ cylinders + displacement + horsepower + weight, data = auto_mpg)
summary(mpg.fit)
```

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	46.3833	1.5647	29.6434	0.0000
cylinders6	-3.5210	1.0506	-3.3514	0.0009
cylinders8	0.6573	1.9370	0.3394	0.7346
displacement	0.0007	0.0099	0.0750	0.9403
horsepower	-0.0893	0.0158	-5.6474	0.0000
weight	-4.4521	0.7921	-5.6210	0.0000

(5 points)

- Recognize how it is set up with cylinders
- Watch for +/-
- Do not include epsilon here
- Recognize that this is an estimate
- Include all predictors

$$\hat{mpg} = \begin{cases} 46.3833 + (7 \times 10^{-4}) \times \text{displacement} + (-0.0893) \times \text{horsepower} + (-4.4521) \times \text{weight} & \text{if cylinders is 4} \\ 42.8623 + (7 \times 10^{-4}) \times \text{displacement} + (-0.0893) \times \text{horsepower} + (-4.4521) \times \text{weight} & \text{if cylinders is 6} \\ 47.0406 + (7 \times 10^{-4}) \times \text{displacement} + (-0.0893) \times \text{horsepower} + (-4.4521) \times \text{weight} & \text{if cylinders is 8} \end{cases}$$

f. Give the interpretation of the coefficient for the variable `horespower`.

(4 points)

- $\beta_4 = -0.0893$
- For each additional horsepower, the `mpg` will decrease by 0.0893 (2 points)
- Holding the other predictors at a fixed value.

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

(5 points)

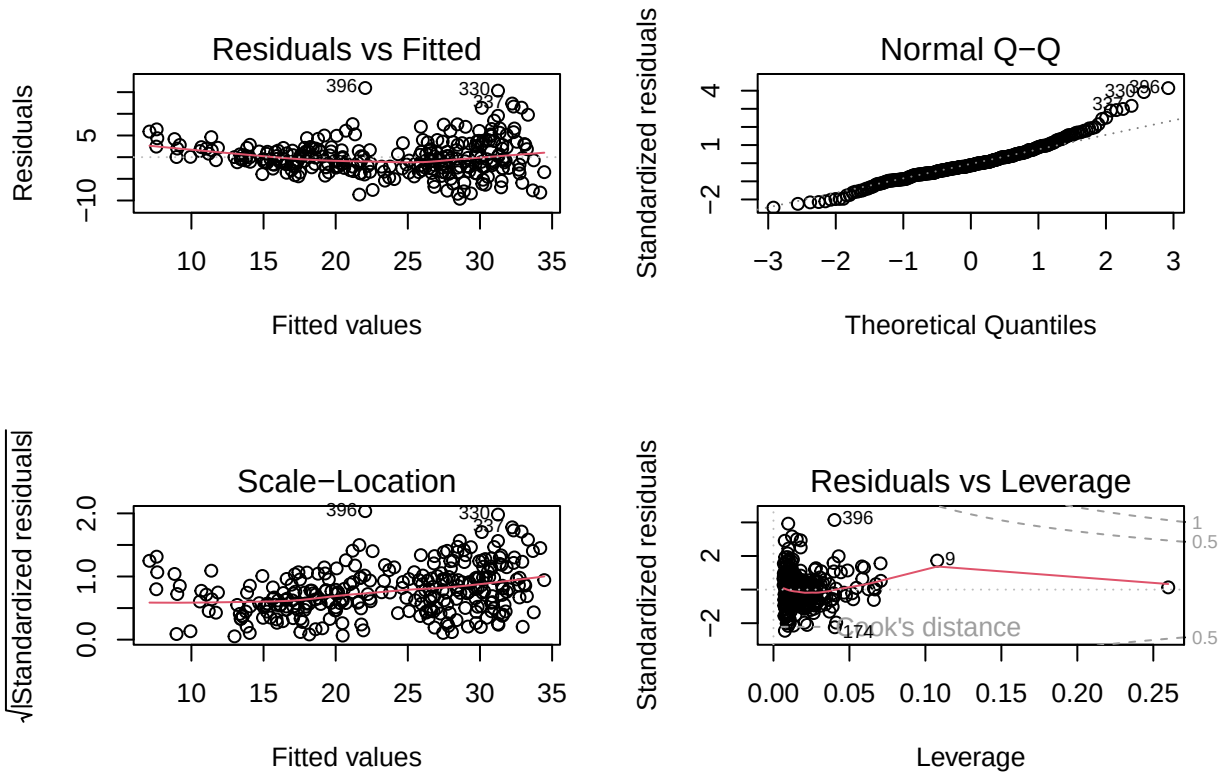
- Yes
- Cylinders and Displacement
- We use the t-test for each individual coefficient if the p-value > 0.1 we determine that that variable is not significant.
- $H : \beta_2 = 0$ (`cylinders8`), given that the other variables are in the model, p-value $= 0.7346$, thus we fail to reject H_0
- $H : \beta_3 = 0$ (`displaclement`), given that the other variables are in the model, p-value $= 0.9403$, thus we fail to reject H_0.

h. What are the assumptions of this model?

(4 points)

- L - Linear
- I - Independent sample
- N - Normal distribution for error term
- E - Equal variance for error term

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



(4 points)

- This may not be linear by the Residuals vs Fitted plot
- Approximately normal by Normal Q-Q plot
- Equal variance by the Scale - Location plot
- Some observations have hig leverage by the Residuals vs Leverage plot

Problem 2

(32 possible points) We want to predict whether a person will donate blood or not. The variables are:

- `Monetary` - total blood donated in c.c per 1000.
- `Recency` - months since last donation.
- `Donate` - a binary variable representing whether he/she donated blood (1 stand for donating blood; 0 stands for not donating blood).

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

(4 points)

- Prediction

b. Is this a regression or classification problem?

(4 points)

- Classification

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

(5 points)

- Cannot be linear
- Recognizes logistic
- Includes both predictors
- $p(X) = P(\text{Donate} = 1|\text{Monetary and Recency})$
- Either of these two formulas will be accepted

$$p(X) = \frac{exp\left(\beta_0 + \beta_1 \times \text{Monetary} + \beta_2 \times \text{Recency}\right)}{1 + exp\left(\beta_0 + \beta_1 \times \text{Monetary} + \beta_2 \times \text{Recency}\right)}$$

or

$$\log \quad \frac{p(X)}{1 - p(X)} \quad = \beta_0 + \beta_1 \times \text{Monetary} + \beta_2 \times \text{Recency}$$

d. Give the R code to get the model for predicting the probability of donating blood based on the 2 input variables.

(5 points)

- glm
- Donate first in the model
- additive of the predictor
- include family = "binomial"
- summary function

```
blood.fit = glm(Donate ~ Monetary + Recency, data = blood,
                family = "binomial")
summary(blood.fit)
```

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.4931	0.2011	-2.45	0.0142
Recency	-0.1199	0.0191	-6.29	0.0000
Monetary	0.1880	0.0692	2.72	0.0066

(5 points)

- Not linear
- Watch for +/-
- Include all predictors (interchanging the terms is fine)
- Can do log or exponential (will need exponential for part f)
- Equation

$$p(X) = \frac{exp\left(-0.4931 + (-0.1199) \times \text{Recency} + (0.188) \times \text{Monetary}\right)}{1 + exp\left(-0.4931 + (-0.1199) \times \text{Recency} + (0.188) \times \text{Monetary}\right)}$$

f. Give the predicted probability of donating blood for a donor that has donated 1400 c.c. of blood and last donation was 4 months ago.

(5 points)

- Recognize the units of Monetary it is per 1000 c.c so they need to use 1.4.
- Input 4 into Recency
- Use the $p(X)$ formula

$$p(X) = \frac{exp\left(-0.4931 + (-0.1199) \times 4 + (0.188) \times 1.4\right)}{1 + exp\left(-0.4931 + (-0.1199) \times 4 + (0.188) \times 1.4\right)} = 0.3297$$

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

Null deviance:	619.14	on	560	degrees of freedom
Residual deviance:	551.4	on	558	degrees of freedom

(4 points)

- Correct formula (2 points)
- Interpretation (2 points)

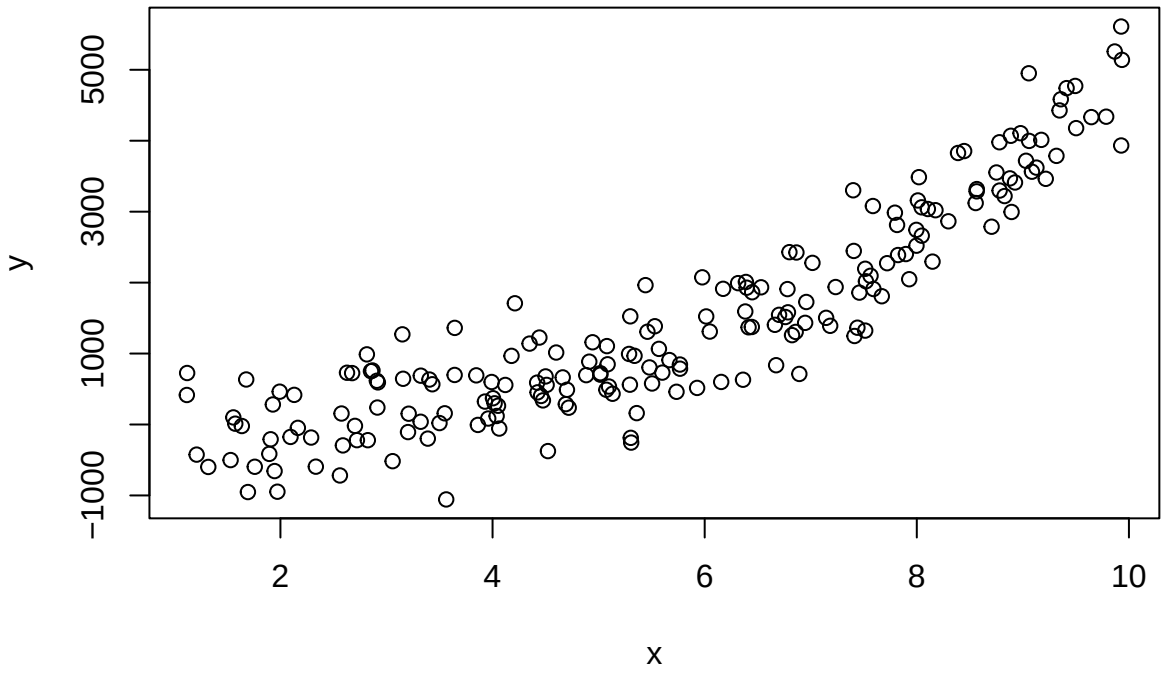
$$R^2 = 1 - \frac{551.4}{619.14} = 0.1094$$

This is not a good fit to predict if they will donate blood.

Problem 3

(8 possible points)

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



(4 points)

- This is not linear

b. The following is an 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 statistics, write out the formula for the best model.

	Adj.R2	Cp	BIC
Degree 1	0.80	152.03	-310.46
Degree 2	0.88	8.12	-413.85
Degree 3	0.89	3.16	-415.55
Degree 4	0.89	5.00	-410.42

(4 points)

- Correct number of terms
- Include the error term
- Degree 3 is the best
- Formula

$$y = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \beta_3 \times X^3 + \epsilon$$

Problem 4

(8 points) A graduate program is making decisions to admit students into the program with the variables `GPA`, and the score on the `GRE`. The response variable is `Decision`, there are three decisions that are made; *yes*, *no*, and *conditional*.

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?

	Yes	No	Conditional
Yes	24	0	2
No	0	19	1
Conditional	0	1	21

i. **0.0588**

ii. 0.9231

iii. 0.95

iv. 0.9412

v. 0.9545

Problem 5

(4 points) The following is the ANOVA table from problem 1, where $n = 288$ and the MSE for the full model from problem 1 is 15.45. What is the C_p statistic?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
horsepower	1	11180.36	11180.36	642.83	0.0000
weight	1	1522.33	1522.33	87.53	0.0000
Residuals	285	4956.83	17.39		

a. 439.57

b. -185.48

c. 4

d. 40.8

e. **38.8**

Problem 6

(4 points) Suppose we have $p = 3$ predictors. How many possible additive models contain subsets of the 3 predictors?

a. 4

b. **8**

c. 16

d. 36

e. 100

Problem 7

(4 points) Which stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the significant predictors are in the model.

a. **forward**

b. backward

c. best subset

d. none of these

Problem 8

(4 points) The following is a 95% prediction interval for the `mpg` from problem 1, with only `weight` as the predictor. We wanted to predict where `weight` is 2845 pounds. Which statement is correct?

	fit	lwr	upr
1	24.48	16.03	32.93

a. **For one automobile that weighs 2845, we predict the `mpg` to be between 16.03 and 32.93 with 95% confidence.**

b. On average for all automobiles that weigh 2845, we we predict the `mpg` to be between 16.03 and 32.93 with 95% confidence.

c. For one automobile that regardless of the weight, we predict the `mpg` to be between 16.03 and 32.93 with 95% confidence.

d. On average for all automobiles regardless of the weight, we we predict the `mpg` to be between 16.03 and 32.93 with 95% confidence.

e. For an automobile that weights 2845 pounds, the `mpg` will be 24.48.