Exam 1 B - MATH 4322

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?

b. Is this a regression or classification problem?

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

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

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

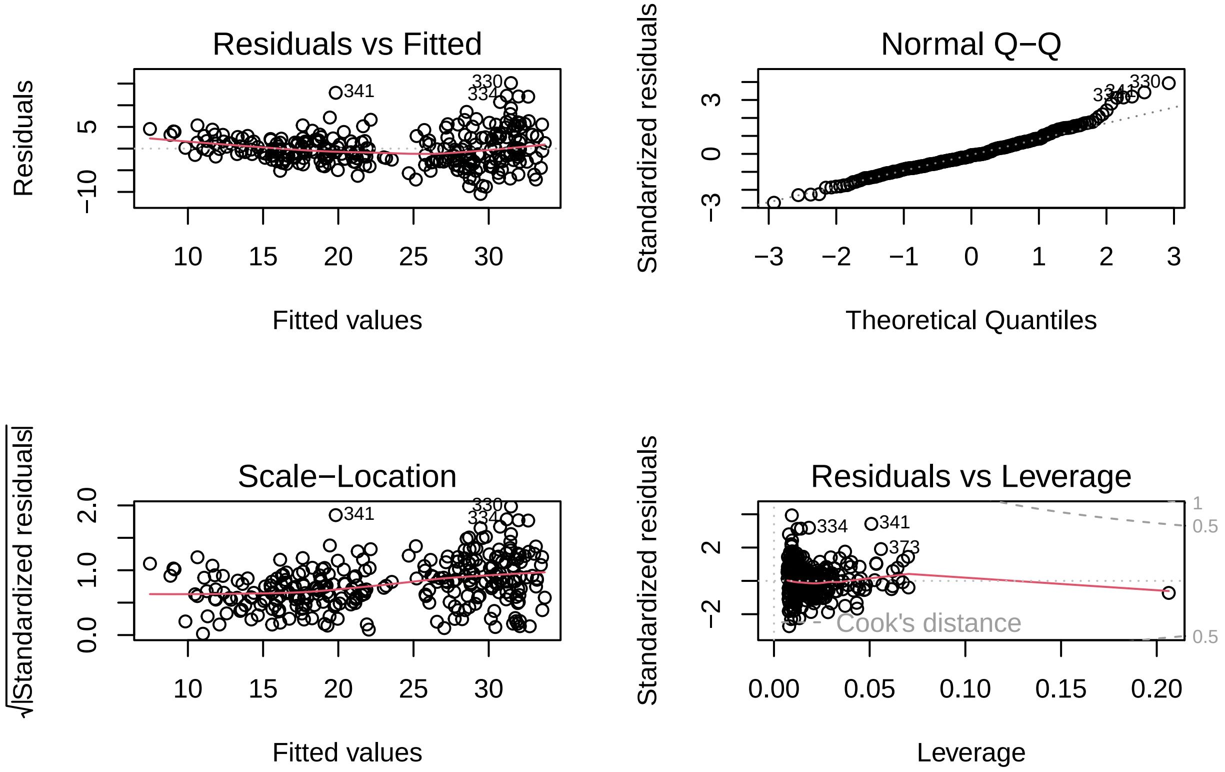
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	45.0220	1.4919	30.1767	0.0000
cylinders6	-4.3528	0.9674	-4.4994	0.0000
cylinders8	-1.6160	1.7456	-0.9258	0.3554
displacement	-0.0027	0.0094	-0.2831	0.7773
horsepower	-0.0480	0.0145	-3.3028	0.0011
weight	-4.8304	0.7566	-6.3846	0.0000

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

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

h. What are the assumptions of this model?

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



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?

b. Is this a regression or classification problem?

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

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

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.5662	0.2034	-2.78	0.0054
Recency	-0.1366	0.0211	-6.46	0.0000
Monetary	0.2545	0.0721	3.53	0.0004

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

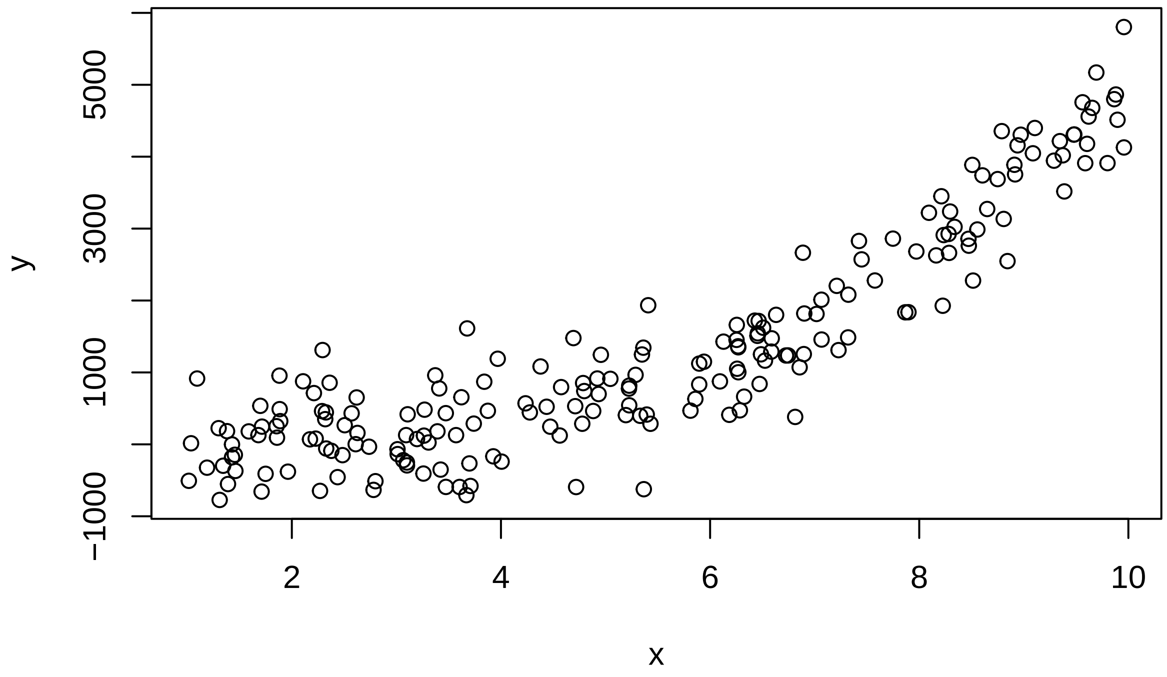
g. The following is the output from `R`. Determine R^2 and give an interpretation.

```
    Null deviance:   614.5   on   560   degrees of freedom
Residual deviance:  531.13   on   558   degrees of freedom
```

Problem 3

(8 possible points)

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



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	0.78	226.35	-292.13
Degree 2	0.89	6.63	-435.70
Degree 3	0.90	4.75	-434.30
Degree 4	0.90	5.00	-430.80

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	19	0	4
No	0	21	1
Conditional	0	0	23

i. 0.0735

ii. 0.8261

iii. 0.9545

iv. 0.9265

v. 1

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 14.89. What is the C_p statistic?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
horsepower	1	11277.08	11277.08	662.91	0.0000
weight	1	2328.67	2328.67	136.89	0.0000
Residuals	285	4848.28	17.01		

a. 473.12

b. -127.66

c. 4

d. 45.5

e. 41.5

Problem 6

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

a. 4

b. 8

c. 16

d. 32

e. 100

Problem 7

(4 points) Which stepwise selection begins with the full least squares model containing all the p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

a. forward

b. backward

c. best subset

d. none of these

Problem 8

(4 points) The following is a 95% confidence 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	24.78	24.28	25.27

a. For one automobile that weighs 2845, we predict the `mpg` to be between 24.28 and 25.27 with 95% confidence.

b. On average for all automobiles that weigh 2845, we we predict the `mpg` to be between 24.28 and 25.27 with 95% confidence.

c. For one automobile that regardless of the weight, we predict the `mpg` to be between 24.28 and 25.27 with 95% confidence.

d. On average for all automobiles regardless of the weight, we we predict the `mpg` to be between 24.28 and 25.27 with 95% confidence.

e. For an automobile that weights 2845 pounds, the `mpg` will be 24.78.