Exam 1 B Solutions - MATH 4322

Spring 2024

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

(36 possible points, 4 points each part) We want to know which features helps us predict the average amount of `Balance` individuals maintain on their credit cards.

The features are as follows:

- *Income*: Income in \$1,000's
- *Rating*: Credit rating
- *Age*: Age in years
- *Student*: A factor with levels **No** and **Yes** indicating whether the individual is a student or not
- The response variable - *Balance*: Average credit card balance in \$.

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

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

This is an inference 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.

$$\hat{Balance} = \beta_0 + \beta_1 \times \text{Income} + \beta_2 \times \text{Rating} + \beta_3 \times \text{Age} + \beta_4 \times \text{Student} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

d. Give the R code to get the model for predicting the average **Balance**, given the features.

credit.lm = lm(Balance ~ Income + Rating + Age + Student, data = Credit2)

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-533.9378	29.6293	-18.0206	0.0000
Income	-7.3586	0.3425	-21.4824	0.0000
Rating	3.9899	0.0752	53.0258	0.0000
Age	-1.2487	0.4311	-2.8968	0.0042
StudentYes	444.0252	25.7229	17.2619	0.0000

$$\hat{Balance} = \begin{cases} -533.9378 + (-7.3586) \times \text{Income} + 3.9899 \times \text{Rating} + (-1.2487) \times \text{Age}, & \text{if Student} = \text{No} \\ -89.9126 + (-7.3586) \times \text{Income} + 3.9899 \times \text{Rating} + (-1.2487) \times \text{Age}, & \text{if Student} = \text{Yes} \end{cases}$$

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

For each year an individual gets older, the average credit card balance will decrease by 1.2487 with the other variables being fixed.

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

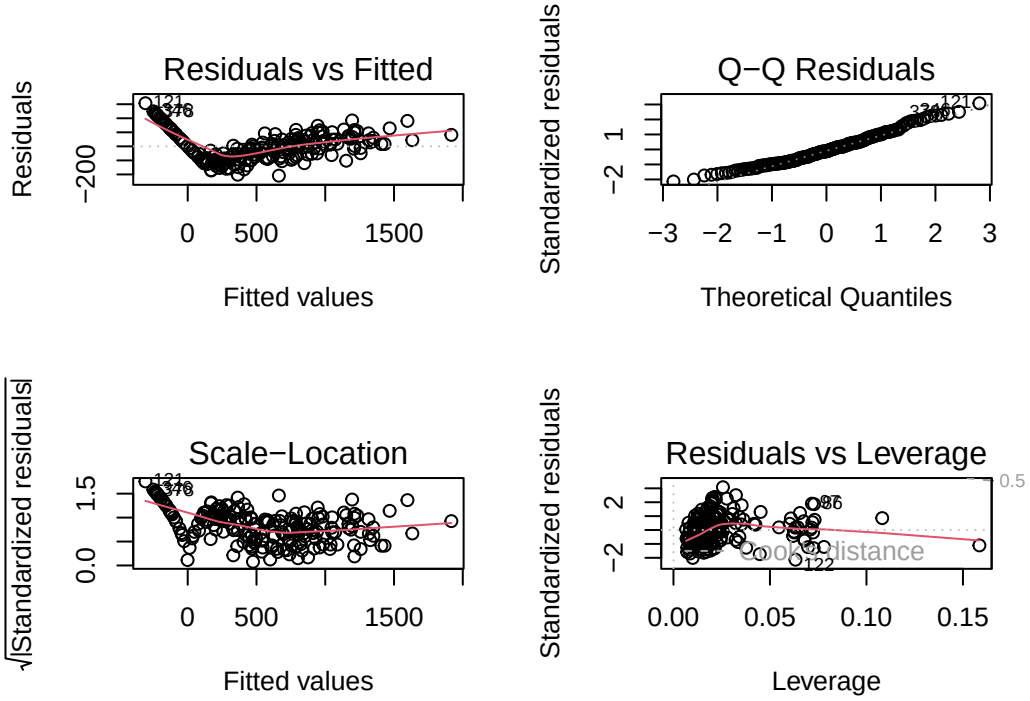
No, since the p-value for all variables are less than 0.05, we can say that all of the predictors can be used in the model, given that the other predictors are in the model.

h. What are the assumptions of this model?

Assumptions

- Linear
- Independent observations
- Normal distribution of error terms
- Equal variance of error for each value of x.
- No extreme values.

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



Yes, this is not linear.

Problem 2

(28 points, 4 points each part) The aim is to predict which customers will default on their credit card debt.

Response - default: A factor with levels **No** and **Yes** indicating whether the customer defaulted on their debt.
Predictors

- *student*: A factor with levels **No** and **Yes** indicating whether the customer is a student.
- *balance*: The average balance that the customer has remaining on their credit card after making their monthly payment.
- *income*: Income of customer.
- Data is called **Default**.

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

This is a prediction 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.

$$P(Y = Yes|X) = \frac{exp^{\beta_0+\beta_1\times\text{student}+\beta_2\times\text{balance}+\beta_3\times\text{income}}}{1 + exp^{\beta_0+\beta_1\times\text{student}+\beta_2\times\text{balance}+\beta_3\times\text{income}}}$$

d. Give the R code to get the model to predict the probability that a customer defaults given the other 3 features as the input variables.

default.glm = glm(default ~ student + balance + income, family = binomial,data = default2)

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-17.6286	4.5389	-3.88	0.0001
studentYes	0.5933	1.4664	0.40	0.6858
balance	0.0085	0.0023	3.72	0.0002
income	0.0001	0.0000	1.10	0.2722

$$P(X) = \begin{cases} \frac{exp^{-17.6286+0.0085\times\text{balance}+10^{-4}\times\text{income}}}{1+exp^{-17.6286+0.0085\times\text{balance}+10^{-4}\times\text{income}}} & \text{if Student = No} \\ \frac{exp^{-17.0353+0.0085\times\text{balance}+10^{-4}\times\text{income}}}{1+exp^{-17.0353+0.0085\times\text{balance}+10^{-4}\times\text{income}}} & \text{if Student = Yes} \end{cases}$$

f. Give the predicted probability of the customer default, given not a student, balance = 835.50 and income = 34,000. Would you predict the customer to default given these inputs?

The predicted probability of defaulting is 0.08%, which very low.

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

Null deviance:	93.53	on	399	degrees of freedom
Residual deviance:	43.61	on	396	degrees of freedom

$$R^2 = 1 - \frac{43.61}{93.53} = 0.5337$$

This shows how well the model fits the data.

Problem 3

(8 possible points) The following is using the `regsubset` function in `R` to determine the best subsets of variables to predict **Balance**.

```
         Income Rating Age StudentYes
1  ( 1 ) " "    "*"    " " " "
2  ( 1 ) "*"    "*"    " " " "
3  ( 1 ) "*"    "*"    " " "*"
4  ( 1 ) "*"    "*"    "*" "*"
```

	Adj.R2	Cp	BIC
1 Predictor	0.76	835.40	-275.42
2 Predictors	0.88	317.76	-410.29
3 Predictors	0.95	11.39	-589.53
4 Predictors	0.95	5.00	-592.66

a. How many predictors would be best to predict **Balance**, (1, 2, 3, or 4)?

4 predictors

b. Write the formula of the model with the best subset of predictors based on the output above.

Answer

$$\text{Balance} = \beta_0 + \beta_1 \times \text{Income} + \beta_2 \times \text{Rating} + \beta_3 \times \text{Student} + \beta_4 \times \text{Student} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

Problem 4

(12 points) The following is the confusion matrix based on the model from problem 2 to predict if a customer will default. The columns are the actual customers (given observations), the rows are the predicted customers (predicted values).

	No	Yes
No	390	7
Yes	0	3

a. Write the full code used to get this confusion matrix.

default.perc = predict(default.glm,default2,type = "response")
default.pred = ifelse(default.perc < 0.5, "No","Yes")
conf.tab = table(default.pred,default2$default)

b. Based on the confusion matrix, what is the error rate?

$$\text{error rate} = \frac{7}{400} = 0.0175$$

c. Based on the confusion matrix, what is the sensitivity rate?

$$\text{Sensitivity} = \frac{3}{10} = 0.3$$

Problem 5

(4 points) The following is a 95% confidence interval for `balance` from problem 1, with only `income` and `student` as the predictors. We want to predict `balance` for a student that has an income of $45,000. Which statement is correct?

	fit	lwr	upr
1	952.82	762.56	1143.08

a. For one student that has an income of $45,000, we predict the `balance` to be between 762.56 and 1143.08 with 95% confidence.

b. On average for all students that have an income of $45,000, we we predict the `balance` to be between 762.56 and 1143.08 with 95% confidence.

c. For one student that regardless of the income, we predict the `balance` to be between 762.56 and 1143.08 with 95% confidence.

d. On average for all students regardless of income, we we predict the `balance` to be between 762.56 and 1143.08 with 95% confidence.

e. For a student that has an income of $45,000, the `balance` will be 952.82.

Problem 6

(4 points) The following is the ANOVA table from problem 1, where $n = 200$ using `Income` and `Student` to predict `Balance`. What is the AIC statistic?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Income	1	9999215.61	9999215.61	63.38	0.0000
Student	1	2922953.54	2922953.54	18.53	0.0000
Residuals	197	31080112.64	157767.07		

a. 2169.94

b. 1923.9576

c. 9999216

d. 2922954

e. 2396.7523

Problem 7

(4 points) What value refers to the error that is introduced by approximating a real-life problem. This is the measurement of the accuracy of the data.

a. Variance

b. Bias

c. $\hat{\beta}_j$

d. VIF

e. AIC

Problem 8

(4 points) Why do we split the data into training and testing data?

a. To reduce the number of predictors.

b. To increase the error.

c. To avoid overfitting.

d. To make sure we are correct all of the time with the model.

e. None of these.