Problem 1

(32 possible points) We want to predict whether income exceeds \$50K per year based on census data. The variables are: **Age**, **Education** (in years), **Gender** (0 for Female and 1 for Male), **Hours** (hours per week), and **Income** (0 for $\leq 50K$ and 1 for $> 50K$).

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

This is a prediction learning problem

only want to see whether income exceeds 50k, not checking which variables cause it to exceed 50k, so this is prediction.

b. Is this a regression or classification problem?

This is a classification problem

the question is "does income exceed 50k, yes or no?" we're not predicting what the income is, just if it is exceeding 50k or not since we're predicting categories, this is classification.

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

Answer

we're predicting two categories (yes or no) so we use logistic regression

$$p(\text{Income} > 50K|X) = \frac{exp(\beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{Education} + \beta_3 \times \text{Gender} + \beta_4 \times \text{Hours})}{1 + exp(\beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{Education} + \beta_3 \times \text{Gender} + \beta_4 \times \text{Hours})}$$

The probability that the income exceeds 50K given a predictor X

d. Give the R code to predict the probability of income being greater than \$50K.

glm.income = glm(Income ~ Age + Education + Gender + Hours, family = 'binomial')

From lecture 9 & 10:

In R we use the function: `glm(model, family="binomial")` function.

```
fit.bc = glm(Class ~ Cell.shape, family = "binomial", data = bc)
#              ^          ^                    ^
#            y var      x var             logistic reg.
# if we don't do "family = "binomial"" it only does
# regular regression
```

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

Predictor	Estimate	Std. Error	t value	P value
(Intercept)	-9.54	1.425	-6.69	0.0000
Age	0.04	0.013	2.87	0.0041
Education	0.45	0.083	5.44	0.0000
Gender Male	1.50	0.469	3.21	0.0013
Hours	0.02	0.014	1.52	0.1285

Answer

$$\hat{p(X)} = \begin{cases} \frac{exp(-8.04+0.04\text{Age}+0.45\text{Education}+0.02\text{Hours})}{1+exp(-8.04+0.04\text{Age}+0.45\text{Education}+0.02\text{Hours})} & \text{if Male} \\ \frac{exp(-9.54+0.04\text{Age}+0.45\text{Education}+0.02\text{Hours})}{1+exp(-9.54+0.04\text{Age}+0.45\text{Education}+0.02\text{Hours})} & \text{if Female} \end{cases}$$

Remember our original model:

$$p(\text{Income} > 50K|X) = \frac{exp(\beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{Education} + \beta_3 \times \text{Gender} + \beta_4 \times \text{Hours})}{1 + exp(\beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{Education} + \beta_3 \times \text{Gender} + \beta_4 \times \text{Hours})}$$

In this case our beta0 is -9.54, our gender variable is categorical so we have to split our model into two, if the gender is male and if the gender is female.

If the gender is male then "Gender" = 1, so in the "if Male" part we have a -9.54 and a 1.50 (for the male) and -9.54 + 1.50 = -8.04.

In the "if Female" part since we have encoded it as '0' which that 1.50 won't show up in our model (since 1.50 is only present if the gender is male).

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

As the years of education increase, the probability of making more than $50,000 per year increases as well.

template: "as [quantitative variable] increases the probability of [categorical output succeeding] increases"

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

Yes, the number of hours is not needed in the model

hours is not needed in the model because its p-value is greater than 0.05!

fancier explanation:

To test $H_0 : \beta_4 = 0$, given that the other terms are in the model we get a p-value of 0.1285, thus we would fail to reject the null hypothesis and state that there is no evidence that the number of hours are significant in predicting income, given that the other variables are in the model.

h. The following is the confusion matrix based on the removal of the variable. What is the error rate for this model?

		Predicted > $50K	
		No	Yes
Actual	No	178	14
> $50K	Yes	37	21

missclassified

missclassified

Answer

$$\text{Error Rate} = \frac{14+37}{178+14+37+21} = 0.204$$

error rate: $\frac{\text{false\_positive+false\_negative}}{\text{sample\_size}}$

Problem 2

(36 possible points) We want to be able to see the affect of student performance in secondary education by some predictors. The following are the variables used.

- age - student's age (numeric: from 15 to 22)
- internet - Internet access at home (binary: yes or no)
- absences - number of school absences (numeric: from 0 to 93)
- score - final grade (numeric: from 0 to 20, output target or response variable)

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

This is an inference statistical learning problem

We want to see how the predictors affect the response, not just prediction. This is a 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.

Answer

$$score = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{internet} + \beta_3 \times \text{absences} + \epsilon$$

The score is a numeric output, so we can use (multiple) linear regression.

d. The following is an output for predicting the final grade. Write out the equation with the estimates.

Predictor	Estimate	Std. Error	t value	P value
(Intercept)	17.81	2.559	6.96	0.0000
age	-0.37	0.155	-2.38	0.0179
internetyes	0.61	0.437	1.39	0.1664
absences	-0.09	0.042	-2.17	0.0311

Answer

$$\hat{score} = \begin{cases} 18.42 - 0.37\text{age} - 0.09\text{absences} & \text{If they have internent} \\ 17.81 - 0.37\text{age} - 0.09\text{absences} & \text{If they do not have internet} \end{cases}$$

internet is a categorical predictor, so we split our model into two different parts, one for if there is internet and one for no internet.

if there is internet we add 0.61 to the model, and 17.81 is beta0, so 17.81 + 0.61 = 18.42

if there is no internet then beta2 becomes 0 and the rest of the model stays the same

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

For each year increase in age, the final score will decrease on average by 0.37 points

template: "for increase in [quantitative predictor] the [quantitative response] increases/decreases on average by [amount]."

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

Yes, if they have internet or not   its p-value is greater than 0.05!

fancy explanation:

When testing $H_0 : \beta_2 = 0$ given that the other terms are in the model, we get a p-value = 0.1664. Thus we fail to reject the null hypothesis and state that there is no evidence that having access to the internet is significant in predicting the final score, given that the other variables are in the model.

g. What is the predicted value of the final score, where the student is 17 years old, does not have internet, and has 2 absences?

predicted score = 17.81 -0.37(17) -0.09(2) = 11.34

remember our model:

$$\hat{score} = \begin{cases} 18.42 - 0.37\text{age} - 0.09\text{absences} & \text{If they have internent} \\ 17.81 - 0.37\text{age} - 0.09\text{absences} & \text{If they do not have internet} \end{cases}$$

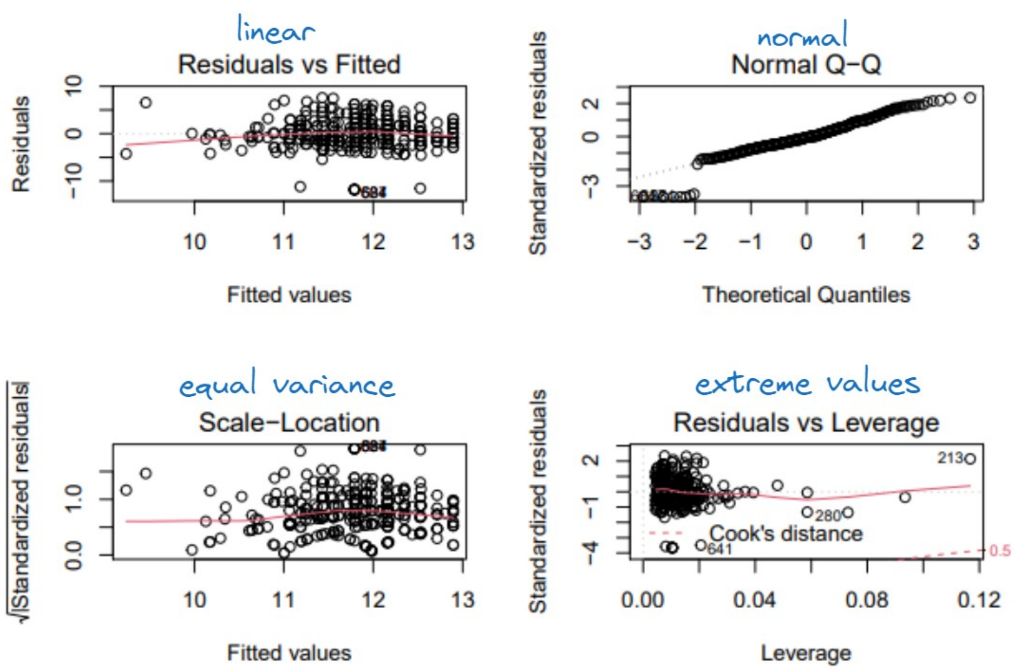
since we are predicting for someone that does not have internet
we use the second part: 17.81 - 0.37*(17) - 0.09*(2)

h. What are the assumptions of this model?

Linear, Independent random sample, Normal distribution, and Equal variance among the residuals for each value of x.

(remember these assumptions about linear models ^)

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



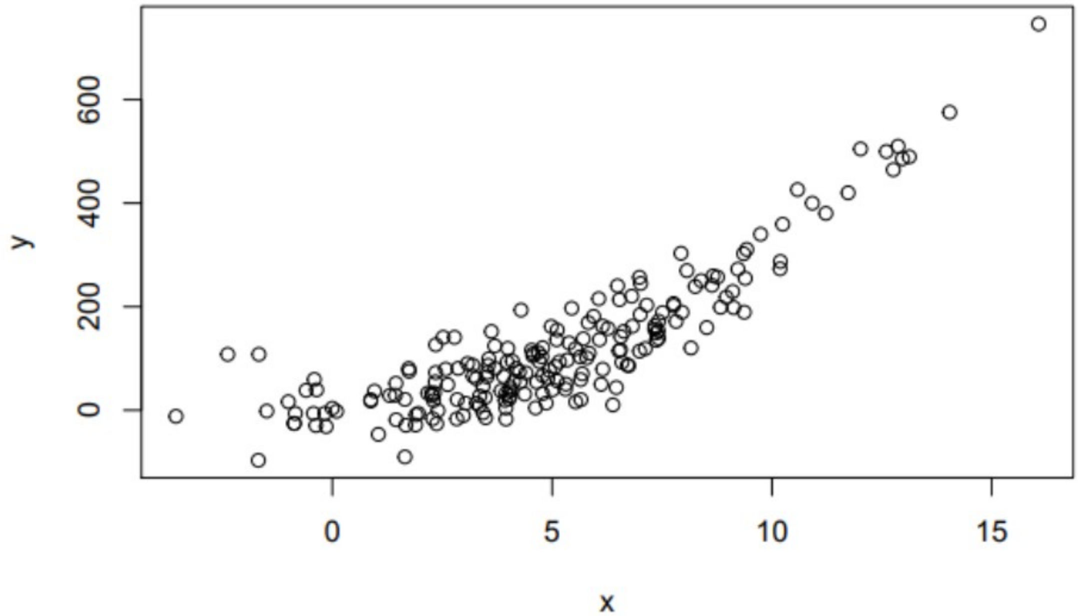
The only violation that appears is that there may be extreme values.

eyeball the plots ig lol :shrug_emoji:

Problem 3

(8 possible points) a. Using the following plot below do we have a linear relationship?

use your eyes
("or get a meter stick and draw a straight line" - Kevin Z)



No, this does not apear to be linear

b. The following is an output for a regression model with degree 1, 2, 3 and 4 respectively. Give the formula for the best model.

```
##          Adj.R2       Cp       BIC
## Degree 1 0.7347 187.6954 -255.8245
## Degree 2 0.8642   1.5024 -385.3815
## Degree 3 0.8638   3.0339 -380.5630
## Degree 4 0.8631   5.0000 -375.2995
```

Since the adjusted R^2 is large, and C_p and BIC are small for Degree 2, the following is the best equation

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Adjusted R^2 must be as close to 1 as possible,
Cp and BIC (and AIC if its given) must all be
as small as possible

look at all three values to determine the best models

Problem 4

(8 possible points) Suppose we have a data set with five predictors, X_1 =GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get the estimated model:

$$\hat{Salary} = 5.5 + 20X_1 + 0.05X_2 - 6.25X_3 + 0.03X4 + 30X_5$$

True or False: For a fixed value of IQ and GPA, females earn more on average than males provided the GPA for females is high. Justify your answer.

TRUE

Suppose IQ = 150 and GPA = 3.5
For Males: Salary = 5.5 + 20(3.5) +0.05(150) +0.03 (3.5)(150) = 98.75
For Females: Salary = 5.5 + 20(3.5) + 0.05(150) -6.25 +.03(3.5)(150) + 30(3.5) = 197.5

In this case the female had a higher salary. Actually because of the coefficient of the interaction term being so large, the female will always have a higher salary.

Remember that interaction means that the value for those predictors are being multiplied basically

Problem 5

(4 points) Given the following ANOVA table, determine the AIC of this model. There are 200 observations.

only one row before the total, so only one predictor

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## x           1 2480746 2480746  552.22 < 2.2e-16 ***
## Residuals 198  889471    4492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSE_p

a. 1950.43

b. 1684.01

c. 1889.15

d. 4492.28

e. 67.0244

formula for AIC:

$$2(p+1) + n\ ln(\frac{SSE_p}{n})$$

where p is the number of predictors

$$2 \cdot (1+1) + 200 \cdot \ln\left(\frac{889471}{200}\right) = 1684.012963$$

Problem 6

(4 points) Given the confusion matrix below, determine the sensitivity rate.

		Predicted > $50K: No	Predicted > $50K: Yes
Actual > $50K	No	178	14
	Yes	37	21

true positives

a. 0.36

b. 0.6

c. 0.93

d. 0.83

e. 0.08

Sensitivity: When its actually positive, how often does it predict positive? Also called the true positive rate.

$$\frac{\text{true positive}}{\text{total positives}}$$

$$\frac{21}{21+37} = 0.3620689655$$

Problem 7

(4 points) Given the training data set, testing data set and MSE which statement is true?

a. The data sets most of the time will have the same vale of MSE.
(no they could be different, which is why we do resampling)

b. If the testing data set has a larger MSE, this is called overfitting the data.

c. The training data set will have a MSE of zero (0).

d. The training data set most of the time will have the largest MSE.

e. The testing data set most of the time will have the largest MSE.

Problem 8

(4 points) Given a 95% confidence interval for the students final score in problem 2 below, which statement is correct?

[10.613, 12.131]

a. For one student, we predict the score to be between 10.613 and 12.131 with 95% confidence.

b. We predict the average score of the students to be between 10.613 and 12.131 with 95% confidence.

c. For one student, there is a 95% chance that the score is between 10.613 and 12.131.

d. There is a 95% chance that the average score of the students is between 10.613 and 12.131.

e. None of these are correct.

Remember:

- prediction interval: The prediction interval predicts the response for *one* observation.

- confidence interval: The confidence interval predicts the *average* across all observations