F24 PSTAT 127 Homework 5 Questions - there are two main questions. Answer all parts

Due uploaded to Gradescope by 11:59 pm on Wed Nov 13**, as a pdf file**
I also will accept assignments uploaded on Thurs Nov 14 without penalty - one-day overlap with next assignment

1. The dataset `wbca` (in the R library `faraway`) comes from a study of breast cancer in Wisconsin. The data represent 681 cases of potentially cancerous tumors of which 238 are actually malignant. At the time of this study, assessment of whether a tumor is malignant usually involved an invasive surgical procedure. The purpose of this study was to determine whether a *new procedure (i.e., new at the time of this study)* called fine needle aspiration *(which draws only a small sample of tissue)* could be effective in determining tumor status.

   (a) Fit a logistic regression with `CLASS` as the response and the other nine variables as predictors using glm in R. Report the residual deviance and associated degrees of freedom.

   (b) Use AIC as the criterion to determine the best subset of variables if you only consider models obtained by dropping one explanatory variable out at a time. (Read the help file for the `step` function.)

   Side-note: this is not necessarily the optimum model if we had considered all possible subsets of the 9 variables, but that would be a lot of models to consider and beyond the scope of this particular homework!
   Here I just ask you to drop out one variable at a time sequentially.

   (c) Use the reduced model in step (1b) to estimate the probability of malignancy for a new patient with predictor variables 1, 1, 3, 2, 1, 1, 4, 1, 1 (same order as in the data frame). Hint: use function “predict.glm” with the appropriate choice of ”type”.

   (d) Suppose that a cancer is classified as benign if $p > .5$ and malignant if $p < .5$. Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.

Important for 1 b)

We discovered that some Rstudio versions have a conflict with the "step" wrapper command, leading to an error in 1(b)

If you receive an error running "step" (error about a class name being too long),

then replace "step" by "stats::step" to specify that R should use the "step" function that belongs in the "stats" library.

The step command works fine for some of us without specifying stats library --- but not for all.

Ask if questions about this - and also let me know if any other commands don't run as expected in your R version

2. A study was run to investigate whether a statistical model could be used to estimate the probability of a household purchasing a new car within a 12-month period, based both on the income of the household and on the age of the oldest car belonging to the household at the start of that 12 month period.

   Data were collected from a random sample of n households. Each household was asked the age of their oldest automobile (variable labelled ”age” measured in years), and their income (variable labelled ”income”). One year later, a follow-up visit asked if the household had bought a new car in that 12 month period (variable ”purchase” - coded as ”1” if they had bought a new car, and ”0” otherwise”).

   Two models were fitted using R.

   On canvas I provide the data set car.txt from books written by ”Kutner, Nachtsheim, Neter” and co-authors. Run the code I type below yourself (or equivalent code - for example there are alternatives to attaching the data frame by including ”data=car.dat” within glm command). Include all your code and results within your uploaded answer pdf file, together with the answers to each of the questions I ask.

```
car.dat <- read.table("...  enter your path here ..../car.txt", header=T)

attach(car.dat)

> fit1 <- glm(purchase ~ income + age, family=binomial(link=logit))
> summary(fit1)   ## edited

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6189  -0.8949  -0.5880   0.9653   2.0846

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.73931    2.10195  -2.255   0.0242 *
income       0.06773    0.02806   2.414   0.0158 *
age          0.59863    0.39007   1.535   0.1249
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.987  on 32  degrees of freedom
Residual deviance: 36.690  on 30  degrees of freedom
AIC: 42.69

Number of Fisher Scoring iterations: 4
```

   (a) What is the value of n (i.e., how many households are in our random sample)?

   (b) Write down model ”fit1” and the assumptions of this model. Define all notation that you use.

```
> fit2 <- update(fit1, .~.-age)
> summary(fit2)

Call:
glm(formula = purchase ~ income, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5883  -0.8430  -0.7121   0.9262   1.7688

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.98079    0.85720  -2.311   0.0208 *
income       0.04342    0.02011   2.159   0.0308 *
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.987  on 32  degrees of freedom
Residual deviance: 39.305  on 31  degrees of freedom
AIC: 43.305

Number of Fisher Scoring iterations: 4

# ---------------------------------------------------------------------------------------------------
> round( vcov(fit2), digits=5 )        # vcov provides the estimated variance-covariance matrix
            (Intercept)   income                # of the beta parameter estimators in fit2 results
(Intercept)     0.73478 -0.0154
income         -0.01540  0.0004
```

(c) Load R library "MASS" and read the help file for "confint.glm". The last example in this help file has a glm for the budworm data that you can run to practice to learn the R syntax.

```
library(MASS)
help("confint.glm")
```

Now use the "confint.glm" command to obtain a 99% confidence interval estimate for parameter β_{income}, assuming model "fit2" holds. Include your code and confidence interval results within your answer file.

(d) i. Explain why I receive the warning message below when I run the following anova command in R.

```
> anova(fit2, fit1, test="F")
Analysis of Deviance Table

Model 1: purchase ~ income
Model 2: purchase ~ income + age
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        31     39.305
2        30     36.690  1   2.6149 2.6149 0.1059

Warning message:
using F test with a 'binomial' family is inappropriate
```

ii. How should I modify this anova command in order to obtain the p-value for the nested model hypothesis test that we studied in class, of
H_0 : model “fit2”, versus H_A: model “fit1”?

Now write and run this modified command. Include the code and R results within your answer file.

Using significance level $\alpha = 0.01$, state the decision of this hypothesis test, i.e., state if the p-value you obtained leads you to reject the null hypothesis in favor of H_A, or not, at this significance level.

As discussed in class, the p-value obtained is only approximate, but answer as though this test is exact for this homework exercise.