Student
number

Semester 2 Assignment, 2024

School of Mathematics and Statistics

MAST90104 A First Course in Statistical Learning

Submission deadline: 6:00 pm on Friday 18 October 2024

This assignment consists of 6 pages (including this page) with 2 questions and 30 total marks

Instructions to Students

Writing

- Please submit a scanned or other electronic copy of your work via the Learning Management System. Your answer to all questions should be a single PDF file.
- You can type or handwrite your answer. If you handwrite your answer, write on A4 paper. Write on one side of each sheet only. Follow the instructions below for scanning and submitting of your assignment.
- If you use R, please include the R output in your answer and submit your R code as an additional file.
- Avoid only showing the R code and/or output as your answer without any explanation. You may lose marks if your answer is unclear.
- Page 1 should only have your student number, the subject code and the subject name. Each question should be on a new page. The question number must be written at the top of each page.

Scanning and Submitting

- Put the pages in question order and all the same way up. Use a scanning app to scan all pages to PDF. Scan directly from above. Crop pages to A4.
- Your scanned assignment must be a single PDF file. Check that all pages are present and readable before submitting.

Blank page

Question 1 (15 marks)

The data *winequality-red.csv* includes data from the paper *Modeling wine preferences by data mining from physicochemical properties* by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis (2009). The data consists of 1599 observations of red variants of the Portuguese "Vinho Verd" wine. The variables are:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol : percent alcohol content of the wine
- quality : Output variable (score between 0 and 10)

The quality score in this data set ranges from 3 to 8, but we will recode the levels as *bad*, *average* and *good*:

```
wine$quality = factor(wine$quality)
levels(wine$quality ) <- c("bad","bad","average","average" , "good","good")
```

(a) Fit a multinomial logit model to predict the wine quality by category, considering all available predictors. Refine the model using stepwise selection with AIC as selection criteria.

(b) Repeat the analysis the analysis in part (a) with an ordinal model, with "good" being the highest level of the output. Comment on any differences.

(c) We have a new observation with these attributes:

```
newobs
# fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
#           7.9              0.4         0.2            1.7       0.1
# free.sulfur.dioxide total.sulfur.dioxide  density   pH  sulphates alcohol
#                  10                   36    0.997  3.3        0.9      10
```

Compute the probabilities that this wine variant is a "bad", "average" and "good" wine, according to the refined multinomial and ordinal models. You should NOT use the function `predict()` for this question.

(d) Under the ordinal model, what is the odds ratio of being classified as "bad" or "average" of a wine variant with chlorides level 0.08 compared to a variant with chlorides level 0.2, given that the other attributes of the two variants are the same?

Question 2 (15 marks) *Use only content provided in this question to answer all parts:* The data `moons` contains the diameter, mass, distance from the sun, and number of moons for 13 planets, gas giants, and dwarf planets in our solar system. The variables are

- `Name`: a character variable with the name of the planet, gas giant, or dwarf planet
- `Distance`: distance from sun, relative to earth’s
- `Diameter`: diameter of the planet, relative to earth’s
- `Mass`: mass, relative to earth’s
- `Moons`: number of moons

The first 3 rows of the data are shown below:

Name	Distance	Diameter	Mass	Moons
Mercury	0.39	0.382	0.0600	0
Venus	0.72	0.949	0.8200	0
Earth	1.00	1.000	1.0000	1

We want to see if the number of moons of a planet is related to its size. We will use Poisson regression (with log link) to model the number of moons of a planet.

(a) We first fit a model that assumes the number of moons does not depend on any other variable in the data (Model 0). Next, we fit a model for the number of moons with predictors `Diameter` and `Mass`. We call this Model 1. The following R output is provided.

```
model0 <- glm(Moons~1,family = poisson(link = ’log’),data = moons)
model1 <- glm(Moons~ Mass + Diameter,family = poisson(link = ’log’),data = moons)
summary(model1)

# Call:
# glm(formula = Moons ~ Mass + Diameter, family = poisson(link = "log"),
#       data = moons)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -2.45936  -2.05425  -0.91278   0.27366   4.32665
#
# Coefficients:
#                Estimate  Std. Error z value  Pr(>|z|)
# (Intercept)  0.71322809  0.19912109  3.5819 0.0003411 ***
# Mass        -0.00401033  0.00094609 -4.2388 2.247e-05 ***
# Diameter     0.41803308  0.03270228 12.7830 < 2.2e-16 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
#     Null deviance: 388.2529  on 12  degrees of freedom
# Residual deviance:  44.8173  on 10  degrees of freedom
# AIC: 84.8331
> c(qchisq(0.9,10), qchisq(0.9,2), qnorm(0.95))
[1] 15.987179  4.605170  1.644854
```

Compared to Model 0, is Model 1 a better model for predicting the number of moons? Conduct an appropriate hypothesis test at 10% significance level and state clearly the test you use, the test statistic value, its null distribution, and the conclusion. *Hint: the Null deviance in the R output corresponds to the deviance of Model 0. The residual deviance corresponds to the deviance of Model 1.*

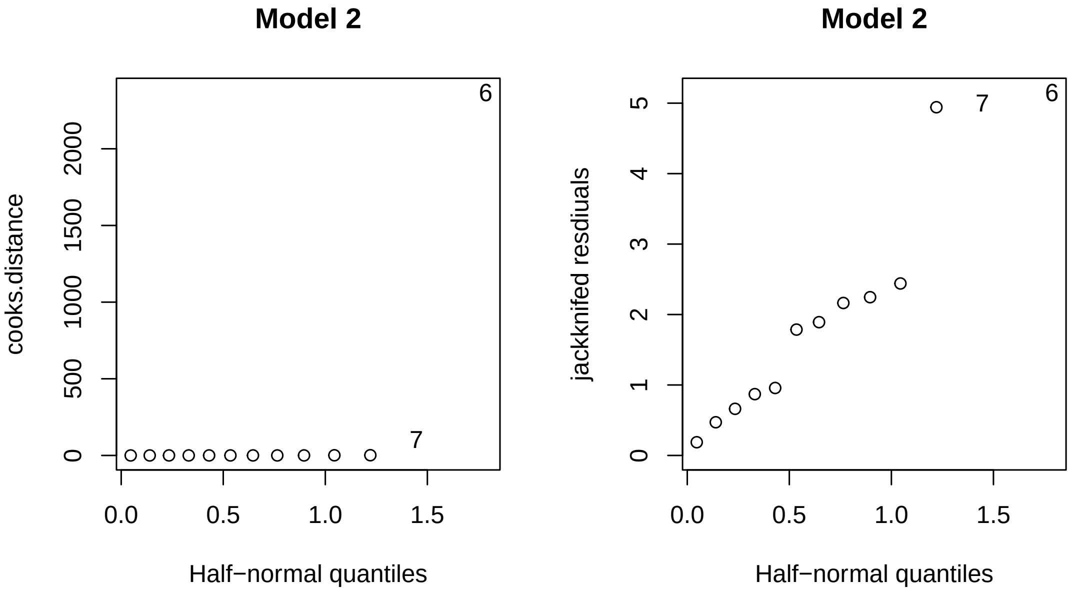
(b) We suspect that distance from sun is also useful in predicting the number of moons of a planet. We then add `Distance` to our model, and call this new model Model 2. The model's summary is shown below (with a `MISSING` output):

```
model2 <- glm(Moons˜ Mass + Diameter + Distance, family = poisson(link = ’log’),
data = moons)
summary(model2)

# Call:
# glm(formula = Moons ˜ Mass + Diameter + Distance, family = poisson(link = "log"),
#     data = moons)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -2.34540  -1.73936  -0.87962   0.27045   4.40580
#
# Coefficients:
#               Estimate  Std. Error z value  Pr(>|z|)
# (Intercept)  0.31230542  0.33261090  0.9390   0.34776
# Mass        -0.00392976  0.00094607 -4.1538 3.271e-05 ***
# Diameter     0.44500885  0.03824737 11.6350 < 2.2e-16 ***
# Distance     0.01410618  0.00826520  1.7067   0.08788 .
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
#     Null deviance: 388.2529  on 12  degrees of freedom
# Residual deviance:  41.8974  on  9  degrees of freedom
# AIC: <MISSING>
```

Compute the missing output, i.e., model 2's AIC.

(c) The figure in the next page shows the half-normal plots of Cook's distance (left panel) and jackknife residuals (right panel) from Model 2. Can you identify any influential observations or outliers from the two plots?



End of Assignment — Total Available Marks = 30