df— title : ‘w271: Homework 2 (Due: 4pm Monday Week 3)’ author: “Professor Jeffrey Yau” output: pdf_document: toc: true number_sections: true fontsize: 11pt geometry: margin=1in —

Due: Before the Live Session of Week 3

Instructions (Please Read it Carefully!):

- **Page limit of the pdf report: None, but please be reasonable**
- Page setup:
- Use the following font size, margin, and linespace:
    - fontsize=11pt
    - margin=1in
    - line_spacing=single
- Submission:
    - Homework needs to be completed individually; this is not a group project.
    - Each student submits his/her homework to the course github repo by the deadline; submission and revision made after the deadline will not be graded
    - Submit 2 files:
        1. A pdf file that details your answers. Include all the R codes used to produce the answers. *Please do not suppress the codes in your pdf file.*
        2. R markdown file used to produce the pdf file
    - Use the following file-naming convensation; fail to do so will receive 10% reduction in the grade:
        - StudentFirstNameLastName_HWNumber.fileExtension
        - For example, if the student’s name is Kyle Cartman for homework 1, name your files as
            - KyleCartman_HW1.Rmd
            - KyleCartman_HW1.pdf
    - Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files.
    - For statistical methods that we cover in this course, use only the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you have to (1) provide an explanation of why such libraries and functions are used instead and (2) reference to the library documentation. **Lacking the explanation and reference to the documentation will result in a score of zero for the corresponding question.** For data wrangling and data visualization, you are free to use other libraries, such as dplyr, ggplot2, etc.
- For mathematical formulae, type them in your R markdown file. **Do not write them on a piece of paper, snap a photo, and either insert the image file or sumbit the image file separately. Doing so will receive a 0 for that whole question.**
- Students are expected to act with regards to UC Berkeley Academic Integrity.

In the live session of week 2, we discussed data analysis, EDA, and binary logistic regression. This homework is designed to review and practice these concepts and techniques. It also covers variable transformation and associated concepts covered in week 3.

For this homework, you will use the dataset *“data_wk02.csv”*, which contains a small sample of graduate school admission data from a university. The variables are specificed below:

1. admit - the depenent variable that takes two values: $0, 1$ where 1 denotes *admitted* and 0 denotes *not admitted*.

2. gre - GRE score
3. gpa - College GPA
4. rank - rank in college major

Suppose you are hired by the University’s Admission Committee and are charged to analyze this data to quantify the effect of GRE, GPA, and college rank on admission probability. We will conduct this analysis by answering the follwing questions:

Question 1: Examine the data and conduct EDA.

```
df <- read.table('data_wk02.csv', header = TRUE, sep = ',')
str(df)
```

```
## 'data.frame':    400 obs. of  5 variables:
##  $ X    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ admit: int  0 1 1 1 0 1 1 0 1 0 ...
##  $ gre  : int  380 660 800 640 520 760 560 400 540 700 ...
##  $ gpa  : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
##  $ rank : int  3 3 1 4 4 2 1 2 3 2 ...
```

```
table(df$admit)
```

```
## 
##   0   1 
## 273 127
```

```
prop.table(table(df$admit))
```

```
## 
##      0      1 
## 0.6825 0.3175
```

Question 2: Estimate a binary logistic regression using the following set of explanatory variables: gre, gpa, $rank$, gre^2, gpa^2, and $gre \times gpa$, where $gre \times gpa$ denotes the interaction between gre and gpa variables.

```
mod.fit <- glm(formula = admit ~ gre + gpa + rank + gre^2 + gpa^2 + gre * gpa, family =
 binomial (link = logit), data = df)
mod.fit
```

```
##
## Call:  glm(formula = admit ~ gre + gpa + rank + gre^2 + gpa^2 + gre *
##     gpa, family = binomial(link = logit), data = df)
##
## Coefficients:
## (Intercept)          gre          gpa         rank      gre:gpa
##  -13.196328     0.018507     3.661045    -0.565757    -0.004762
##
## Degrees of Freedom: 399 Total (i.e. Null);  395 Residual
## Null Deviance:       500
## Residual Deviance: 456.6     AIC: 466.6
```

Question 3: Test the hypothesis that GRE has no effect on admission using the likelihood ratio test.

```
mod.fit.h0 <- glm(formula = admit ~  gpa + rank  + gpa^2, family = binomial (link = logit), data = df)
anova(mod.fit.h0, mod.fit, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: admit ~ gpa + rank + gpa^2
## Model 2: admit ~ gre + gpa + rank + gre^2 + gpa^2 + gre * gpa
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       397     463.93
## 2       395     456.60  2   7.3359  0.02553 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 4: What is the estimated effect of college GPA on admission?

```
mod.fit$coefficients
```

```
##   (Intercept)           gre           gpa          rank       gre:gpa
## -13.196328476   0.018507380   3.661044911  -0.565756999  -0.004761884
```

Question 5: Construct the confidence interval for the admission probability for the students with $GPA = 3.3$, $GRE = 720$, and $rank = 1$.

```
ci.logit.profile <- confint(mod.fit, level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
ci.logit.profile
```

```
##                     2.5 %        97.5 %
## (Intercept) -2.545946e+01 -1.6941232424
## gre         -4.000396e-04  0.0384623331
## gpa          2.583081e-01  7.2536228613
## rank        -8.213621e-01 -0.3206443439
## gre:gpa     -1.054386e-02  0.0007580543
```