Biostat 200C Homework 1

Due Apr 14 @ 11:59PM

Contents

To submit homework, please submit Rmd and html files to bruinlearn by the deadline.

Q1. Reivew of linear models

The swiss data — use Fertility as the response to practice

- An initial data analysis that explores the numerical and graphical characteristics of the data.
- Variable selection to choose the best model.
- An exploration of transformations to improve the fit of the model.
- Diagnostics to check the assumptions of your model.
- Some predictions of future observations for interesting values of the predictors.
- An interpretation of the meaning of the model by writing a scientific abstract. (<150 words)
  - BACKGROUND: brief intro of the study background, what are the existing findings

- OBJECTIVE: state the overall purpose of your research, e.g., what kind of knowledge gap you are trying to fill in
- METHODS: study design (how these data were collected), outcome definitions, statistical procedures used
- RESULTS: summary of major findings to address the question raised in objective
- CONCLUSIONS:

Q2. Concavity of logistic regression log-likelihood

Q2.1

Write down the log-likelihood function of logistic regression for binomial responses.

Q2.2

Derive the gradient vector and Hessian matrix of the log-likelhood function with respect to the regression coefficients $\boldsymbol{\beta}$.

Q2.3

Show that the log-likelihood function of logistic regression is a concave function in regression coefficients $\boldsymbol{\beta}$. (Hint: show that the negative Hessian is a positive semidefinite matrix.)

Q3.

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima`.

Q3.1

Create a factor version of the test results and use this to produce an interleaved histogram to show how the distribution of insulin differs between those testing positive and negative. Do you notice anything unbelievable about the plot?

Q3.2

Replace the zero values of `insulin` with the missing value code `NA`. Recreatethe interleaved histogram plot and comment on the distribution.

Q3.3

Replace the incredible zeroes in other variables with the missing value code. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting? Why is this less than the number of observations in the data frame.

Q3.4

Refit the model but now without the insulin and triceps predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question.

Q3.5

Use AIC to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?

Q3.6

Create a variable that indicates whether the case contains a missing value. Use this variable as a predictor of the test result. Is missingness associated with the test result? Refit the selected model, but now using as much of the data as reasonable. Explain why it is appropriate to do this.

```
library(faraway)
```

```
## Warning in check_dep_version(): ABI version mismatch:
## lme4 was built with Matrix ABI version 1
## Current Matrix ABI version is 0
## Please re-install lme4 from source or restore original 'Matrix' package
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
pima <- pima %>%
  mutate(
    glucose2  = ifelse(glucose == 0, NA, glucose),
    diastolic2 = ifelse(diastolic == 0, NA, diastolic),
    triceps2 = ifelse(triceps == 0, NA, triceps),
    insulin2 = ifelse(insulin == 0, NA, insulin),
    bmi2 = ifelse(bmi == 0, NA, bmi),
    diabetes2 = ifelse(diabetes == 0, NA, diabetes),
    age2 = ifelse(age == 0, NA, age))

pima$missingNA = ifelse(apply(is.na(dplyr::select(pima, contains("2"))), 1, sum) > 0, 1, 0)

missing.glm <- glm(test ~ missingNA, family = binomial(), data = pima)

library(gtsummary)
```

```
missing.glm %>%
  tbl_regression() %>%
  bold_labels() %>%
  bold_p(t = 0.05)
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	log(OR)	95% CI	p-value
missingNA	0.16	-0.14, 0.45	0.3

From above regression, we found missingness was not associate with outcome. This means that the distribution of outcome when removing data with missing is still a representative of the original distribution. This justifies the use of "complete case" analysis.

Q3.7

Using the last fitted model of the previous question, what is the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.

Q3.8

Do women who test positive have higher diastolic blood pressures? Is the dias- tolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.