

# ACSC71-326 Advanced Regression

## Final Exam – SAMPLE

### Total Marks: 45

#### Question 1: (9 Total Marks)

Provide a Short Answer (3 marks each)

- Briefly describe what is meant by the “bias variance trade-off” and give two examples of how it operates in modelling techniques we have studied this term.
- When employing a Poisson regression model to assess contingency table data, how do we test whether the two discrete variables involved are independent or not?
- Linear discriminant analysis (LDA) finds class boundaries in the predictor space that are linear in the predictors to determine which category a data point belongs in. Quadratic discriminant analysis (QDA) does essentially the same thing, except the boundaries are quadratic. What is the key difference in the model assumptions for QDA that creates these quadratic boundaries?

#### Question 2: (18 Total Marks)

The file `baby.dat` contains data on 247 premature births. For each, the birthweight (in grams), the gestational age (in weeks), the one and five minute Apgar scores (on a scale of 1 to 9) and the pH level of the venous blood are recorded. In addition, an indicator of whether the baby survived or not is in the first column, a value of 1 indicating survival and 0 indicating death.

- A pregnant woman is experiencing complications and her doctors are considering inducing labour at either 30 or 31 weeks. Assuming the other predictors would be the same regardless of when the induction is performed, the doctors want an estimate of how much the extra week of gestational age would change the baby’s survival chances. Build an appropriate model and give an estimate (and confidence interval) as requested. [9 marks]
- At the time of the one minute Apgar score calculation, the only variables available are the one minute Apgar value itself, the birthweight and the gestational age. Use these variables to develop a model to predict the five minute Apgar score. [9 marks]

#### Question 3: (18 Total Marks)

The file `crckt.dat` contains data for ball-by-ball outcomes of 217 men’s international 50-over matches, 151 men’s international 20-over matches, 72 women’s international 50-over matches and 70 women’s international 20-over matches played during 2015 and 2016. Each row of the data corresponds to a single ball of a match. For each ball, the information recorded is:

- `BallsRem` – the number of balls still remaining in the match, including the current one (e.g., for the first ball of a 50-over match, this value would be 300, as there are 6 balls/over).
  - `Runs` – the total number of runs scored on the ball (including extras for illegal deliveries).
  - `Wckts` – the number of wickets down at the time of the ball.
  - `WLastBall` – an indicator of whether a wicket fell on the previous ball (1 = Yes, 0 = No).
  - `Year` – the calendar year of the match.
  - `GameType` – an indicator of the type of match (1 = 50-over, 2 = 20-over)
  - `Gender` – an indicator of the gender for the match (1 = men, 2 = women)
- There is debate about whether scoring rates for 20-over matches are the same as those in the final 20-overs of a 50-over match (i.e., balls 120 down to 1). Use an appropriate technique to model the relationship between expected runs scored and balls remaining which allows comparison across match types and appropriately adjusts for any other important factors. Further, produce a plot (or small collection of plots) to illustrate the similarity or dissimilarity in the expected runs scored versus balls remaining relationship across the two match types. [9 marks]
  - When a wicket falls, a new batsman must start their innings. It is often claimed this is the most difficult time for a batter to score. Investigate whether there is a difference in the runs scored at any given stage of a match depending on whether a wicket has fallen on the previous delivery or not. [9 marks]