University of Manitoba
Department of Statistics

STAT 4600: Computational Statistics

Assignment 4 – Part I

Due: Tuesday, February 26

Question 1:

The `faithful` data set available in R gives 272 waiting time before eruptions and the duration of these eruptions of the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. These data, taken between November 1 and November 15, 1985, are often examined in research papers because of their interesting bimodal (i.e. two modes) structure. We note that, for example,

```
> faithful[1,]
  eruptions waiting
1       3.6      79
```

meaning that the first eruption lasted 3.6 minutes and occurred after a 79 minute wait. Now, define the variables `X = faithful[,'waiting']` and `Y = faithful[,'eruptions']`. Our goal, here, is to study a regression of Y on X, considering the prediction of eruption durations based on waiting times before eruptions occur. In particular, we write

regression

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1}$$

for $i = 1, 2, \ldots, 272$ and where the random errors are assumed to be i.i.d. with mean zero and common variance σ^2.

(A) If an eruption just started after a 61 minute wait, what is its expected duration?

To answer this question, construct a 90% confidence interval for the expected duration of eruptions happening after a 61 minute wait based on least-squares estimation. For this, use the approach seen in class for bootstrapping parameter estimates in a regression context.

(B) Do the same using the bootstrapping approach described on page 3.

Note: The function `lm()` can be used to easily calculate regression estimates and residuals.

Question 2: A Poisson regression model for English Premiership Soccer

To model the scores of soccer games from the English Premier League, assume that in every league game, the number of goals X scored by team i when playing team j at home satisfies

$$X \sim \text{Poisson}(\exp(\mu + \Delta + \alpha_i + \beta_j)),$$

and that the number of goals scored by team j on the road playing team i satisfies

$$Y \sim \text{Poisson}(\exp(\mu + \alpha_j + \beta_i)).$$

Assume also that X and Y are independent (which is unrealistic, but nevertheless will allow us to do some interesting investigation). The parameters of the model have the following interpretation:

- μ : overall league scoring parameter,
- Δ : scoring effect for playing at home (same for all teams in the league),
- α_i : offensive effect for team i (a large positive value indicates a strong offense),
- β_j : defensive effect for team j (a large negative value indicates a strong defense).

It is assumed, as in traditional ANOVA, that

$$\sum_i \alpha_i = 0 = \sum_j \beta_j, \tag{2}$$

implying that for an average team, we have that $\alpha_i = \beta_i = 0$. Note that this implies that the expected number of goals for an average team playing another average team at home is $e^{\mu+\Delta}$, while it is e^{μ} when playing that same team on the road. Also, $\Delta > 0$ implies that teams tend to score more goals playing at home then they do playing on the road.

The paper by Lee (1997), that is available with the assignment on UMLearn, may be useful to have more insight into what we are doing here. Note that Lee is not using the constraint (2) but is instead fixing some of the parameters to zero.

Now, in our context, It is possible to show that the maximum likelihood estimates of the model parameters obtained from the Premiership data are $\hat{\mu} = 0.0891$ and $\hat{\Delta} = 0.2838$, the team specific parameters being given below in Table 1. (Maximum Likelihood estimation will be discussed later.) You may make use of the `soccer.estimates` function available in the `soccer_estimates.R` file available on UMLearn to obtain all parameter estimates given the `game.results` for a complete season.

(A) Create an R function `new.season` that simulates, based on the above Poisson model, a new season from a `calendar` of games, a data frame containing the dates of all the games and which teams play home and away for each game, and the parameters `mu`, `Delta`, `alpha` and `beta` of the model.

Your function should return a data frame containing the calendar and scores to all the games in the same format as the `soccer` object we have used before.

Note: It may be a good idea to use the function created in Assignment 2 to make sure your results are consistent with those obtained in the original data, although you should not expect to reproduce those exactly.

(B) Using the parametric bootstrap approach, create an R function that returns simultaneous confidence intervals for all parameters of the model. Specifically, your function should take the following arguments:

- the `game.results` of a season of interest,
- the number `N` of replicates of the full season that should be used for bootstrapping,
- the confidence level `C` that is desired,

and return a matrix (with appropriate row and column names) containing the wanted confidence intervals.

Note: You can use the `soccer.estimates` and `new.season` functions for this.

Table 1: Estimates of all team specific parameters for the EPL data.

Team	$\hat{\alpha}_i$	$\hat{\beta}_i$
AFC Bournemouth	0.1035	0.2757
Arsenal	0.4179	-0.1241
Burnley	-0.2532	0.0618
Chelsea	0.5060	-0.4046
Crystal Palace	0.0038	0.2089
Everton	0.2006	-0.1391
Hull City	-0.2812	0.4355
Leicester City	-0.0371	0.2069
Liverpool	0.4288	-0.1698
Manchester City	0.4512	-0.2421
Manchester United	0.0474	-0.5647
Middlesbrough	-0.6235	0.0130
Southampton	-0.2100	-0.0726
Stoke City	-0.2021	0.0818
Sunderland	-0.5363	0.2792
Swansea City	-0.0949	0.3096
Tottenham Hotspur	0.5107	-0.6426
Watford	-0.2150	0.2755
West Bromwich Albion	-0.1593	-0.0099
West Ham United	-0.0573	0.2217

Another approach for bootstrapping regression parameter estimates:

An alternate approach to bootstrapping $\hat{\beta}_0$ and $\hat{\beta}_1$ in model (1) is the following.

1. Estimate the model parameters by calculating $\hat{\beta}_0$ and $\hat{\beta}_1$ from the original sample and calculate the fitted values and residuals

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad \text{and} \qquad e_i = Y_i - \hat{Y}_i.$$

   Define $R_e = \{e_1, e_2, \ldots, e_n\}$ as the sample of observed residuals.

2. Then, for $k = 1, 2, \ldots, N$:

   - generate a bootstrap sample $R^*_{e,k} = \{e^*_1, e^*_2, \ldots, e^*_n\}$ of residuals by sampling from R_e with replacement and calculate new observations

$$Y^*_i = \hat{Y}_i + e^*_i,$$

   - calculate the associated bootstrap replicates $\hat{\beta}^*_{0,k}$ and $\hat{\beta}^*_{1,k}$.

Note: It is especially important to proceed this way when the design is fixed (which isn't the case here). Also, the function `lm()` can be used to easily calculate regression estimates and residuals.