STAT 3690 Homework 2

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Due at Mar 16 11:59 pm (Central Time)

Answers must be submitted electronically via Crowdmark. Please enclose your R source code (if applicable) as well.

1. 10 subjects with bronchus cancer were enrolled in a clinical study. For each of them, two survival times (in week) were recorded: T_1 = survival time from the first hospital admission; T_2 = survival time from the beginning of nontreatability.
   a. The gap between the two survival times is often of interet, because it may reflect the progression of disease as well as the treatment effect. Write down the distribution of time gap $T_1 - T_2$, assuming $[T_1, T_2]^\top \sim MVN_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
   b. Let $\mathbf{T}_i = [T_{i1}, T_{i2}]^\top$, where T_{i1} (resp. T_{i2}) were the observation of T_1 (resp. T_2) for subject i, $i = 1, \ldots, 10$. Suppose $\mathbf{T}_1, \ldots, \mathbf{T}_{10} \overset{\text{iid}}{\sim} MVN_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = [\mu_1, \mu_2]^\top$. Present both the Bonferroni's and Scheffe's simultaneous 95% confidence intervals for μ_1, μ_2 and $\mu_2 - \mu_1$.

Subject	T_1	T_2
1	81	74
2	461	423
3	20	16
4	450	450
5	246	87
6	166	115
7	63	50
8	64	50
9	155	113
10	151	38

2. Consider the `Wolves` dataset from the package `candisc`. The variable `sex` indicates the sex of wolves (`f=female`, `m=male`), while `location` encodes wolves' habitats (`ar=Arctic`, `rm=Rocky Mountain`). The combination of `location` and `sex` is exactly `group`. Variables `x1` to `x9` correspond to 9 different skull morphological measurements of wolves, respectively. **We will merely focus on six measurements `x4` to `x9`.**
    a. Perform an appropriate test to compare the mean skull measurements of male and female wolves. Is there any statistical evidence to claim that the the morphology of the skull differs between males and females at 5% level? (**Hereafter**, in reporting results of hypothesis testing, don't forget to include your hypotheses, the name of method, the value of test statistic, and the rejection region/p-value, before coming to the conclusion.)
    b. What are the assumptions required to perform the test in part a?
    c. Repeat parts a and b for wolves only from the Arctic.
    d. Provide plausible explanations (both statistical and subject-matter) about any discrepancy between the analysis in parts a and c.
    e. For now we are not sure whether the covariance matrix of the six measurements vary with `sex`. Please confirm it via a hypothesis test at level $\alpha = .05$.

```
##      group location sex  x1  x2  x3   x4   x5   x6   x7   x8   x9
## rmm1  rm:m       rm   m 126 104 141 81.0 31.8 65.7 50.9 44.0 18.2
## rmm2  rm:m       rm   m 128 111 151 80.4 33.8 69.8 52.7 43.2 18.5
## rmm3  rm:m       rm   m 126 108 152 85.7 34.7 69.1 49.3 45.6 17.9
## rmm4  rm:m       rm   m 125 109 141 83.1 34.0 68.0 48.2 43.8 18.4
## rmm5  rm:m       rm   m 126 107 143 81.9 34.0 66.1 49.0 42.4 17.9
## rmm6  rm:m       rm   m 128 110 143 80.6 33.0 65.0 46.4 40.2 18.2
```

3. There is a dataset presented by Dean De Cock (2011, *Journal of Statistics Education*, 19(3)). It describes the sale of individual residential property in Ames, Iowa, U.S. from 2006 to 2010, containing 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) for the assessment of home values. We will focus on the following five variables:
    - `Sale_Price`: sale price (in USD);
    - `Lot_Frontage`: linear feet of street connected to property (in feet);
    - `Lot_Area`: lot size (in square feet);
    - `First_Flr_SF`: first floor square feet;
    - `Year_Sold`: year sold.

```r
# The code below creates a clean version of the dataset.
# For more information, type ?AmesHousing::make_ames.
install.packages('AmesHousing')
library(AmesHousing)
ames_data <- make_ames()
```

- There are six subquestions.
    a. Create a pairs plot for variables `Sale_Price`, `Lot_Frontage`, `Lot_Area` and `First_Flr_SF`.
    b. Fit a linear regression model with variables `Sale_Price`, `Lot_Frontage`, `Lot_Area` and `First_Flr_SF` as outcomes and `Year_Sold` as the only explanatory variable. Carefully interpret the regression coefficient estimates.
    c. Test the multivariate regression model in part b against the empty model. Then test each of the four univariate regression models against the empty model. By comparing the multivariate result with the four univariate ones, what conclusions can you draw?
    d. Use the Cook's distance to identify the most influential observations for model in part b.
    e. Investigate the distribution of the residuals and the overall model fit for model in part b.
    f. Based on your observations above, suggest ways of improving the model fit in part b.