Homework 2

Aji John

Winter Quarter 2018

This template is meant to be a helpful starting point for writing up homeworks for students who want to use Rmarkdown. Modify/distribute it as you see fit. At a minimum, you'll want to remove these comments when you turn in your assignments. For additional information about using Rmarkdown, see the help files and links posted to the course Canvas website. Remember, raw R code/output is not acceptable in your homework.

Question 1

In this question, you will perform statistical analyses evaluating an association between serum creatinine levels (measured from blood) and 5 year all-cause mortality by comparing mean creatinine levels across groups defined by vital status at 5 years using a t-test that presumes homoscedasticity (i.e., equal variances across groups). As this problem is directed toward illustrating correspondences between the t-test and linear regression, you do not need to provide full statistical inference for this problem. Instead, just answer the following questions.

(a)

What are the sample size, sample mean and sample standard deviation of serum creatinine levels among subjects who survived at least 5 years?

n	mean	sd
612	1.033987	0.2455265

- Two samples skipped as 'NA' featured in them.
- Units for mean is mg/dl

(b)

What are the sample size, sample mean and sample standard deviation of creatinine levels among subjects who died within 5 years?

n	mean	sd
121	1.215702	0.4745187

- Units for mean is mg/dl

(c)

What are the point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for the mean creatinine level in a population of similar subjects who would survive at least 5 years?

Point estimate is 1.033 mg/dl SE is 0.009974404 95% CI is 1.01 to 1.05

(d)

What are the point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for the mean creatinine level in a population of similar subjects who would die within 5 years?

Point estimate is 1.2157 mg/dl SE is 0.01918 95% CI is 1.13 to 1.30

(e)

What are the point estimate and the 95% confidence interval for the difference in creatinine means between a population of similar subjects that survives at least 5 years and a population of similar subjects that dies within 5 years? What is the P value for testing the hypothesis that the two populations have the same mean creatinine level? What conclusions do you reach about a statistically significant association between serum creatinine and 5 year all-cause mortality?

The best estimate of the difference, i.e. the point estimate is -0.18175 units (Survive Sample, Observed to Die Sample.)

95% CI , difference of mean 'crt' between the groups is (-0.26927031 , -0.09416079).

p-value is 7.011e-05, and is found to be significant, and we reject our null hypothesis that the two populations have the same mean creatinine levels.

(f)

Although we did not consider age at the time of enrollment in the questions above, could the association analysis for creatinine level and 5 year all-cause mortality conducted using the t-test potentially be confounded by the age of the subjects at the time of the MRI? Briefly explain why or why not this is plausible? Provide any descriptive statistics (e.g. an appropriate table, plot, etc.) giving evidence for or against the association results and conclusions above with the t-test potentially being confounded by age.

The age distribution is notable in the study, we can see from the above figure that median is not the same for the two groups, and it is possibly adding the variation.

Question 2

Perform statistical analyses evaluating an association between serum creatinine and 5 year all-cause mortality by comparing mean creatinine levels across groups defined by vital status at 5 years using linear regression that presumes homoscedasticity. As this problem is directed toward illustrating correspondences between the t test and linear regression, you do not need to provide full statistical inference for this problem. Instead, just answer the following questions.

(a)

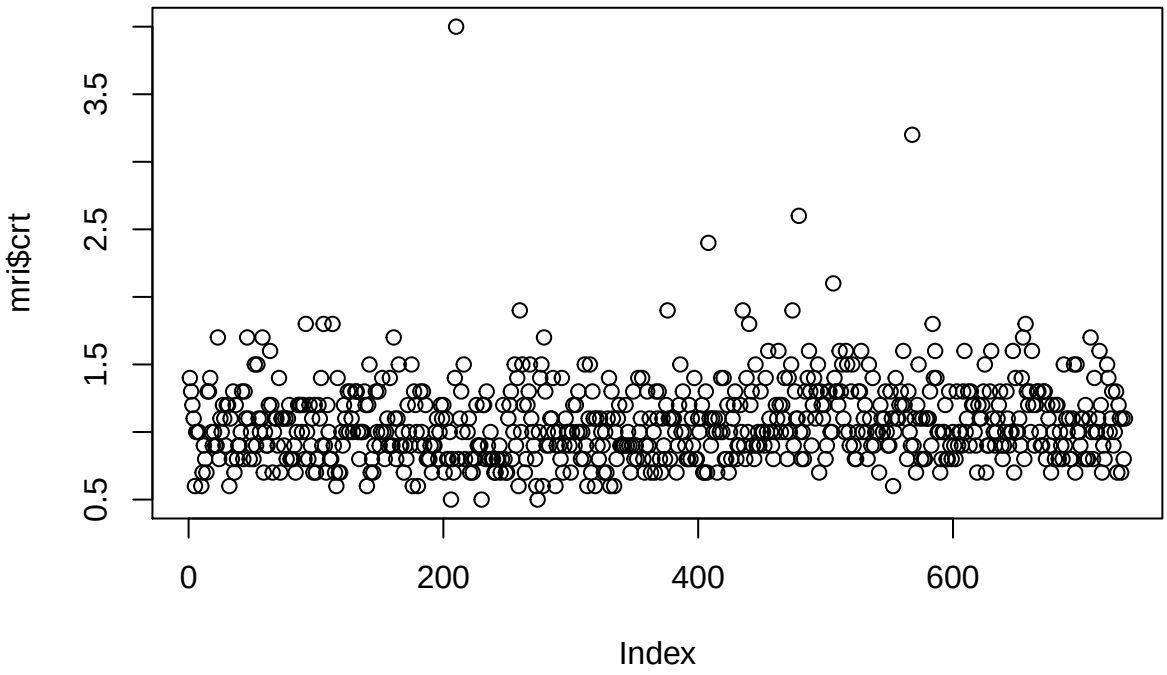
Fit a regression model where the response variable is creatinine level and the predictor variable is an indicator variable for a subject dying within 5 years (i.e., a value of 1 if subject died within 5 years, and a value of 0 if the subject survived at least 5 years). Provide an interpretation of the intercept and slope of this regression model.

Our model came out to be 'crt = 0.18 * death + 1.03

(b)

Is the regression model you fit a saturated model? Briefly explain why or why not.

```
## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not
## a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not
## a graphical parameter

## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter
```



(c)

Using the parameter estimates from the regression model, what is the estimate of the mean creatinine level in a population of similar subjects who would die within 5 years? How does this compare to the corresponding estimate from problem 1? If there are any differences, explain the source of the differences.

```
##
## Call:
## lm(formula = crt ~ deathin5, data = mri)
##
## Coefficients:
##  (Intercept)  deathin5TRUE
##       1.0340        0.1817
```