Homework 3

Aji John

Winter Quarter 2018

Question 1

Perform a statistical analysis evaluating an association between serum creatinine level and 5 year all-cause mortality by comparing geometric mean serum creatinine levels between groups defined by vital status at 5 years. In your analysis, allow for heteroscedasticity. Provide full statistical inference about an association between creatinine and 5 year all-cause mortality.

Model we got is E[log(crt)|deathin5]=0.005763+0.132212×deathin5.

From linear regression analysis on log transformed creatinine level(CRT) using Huber-White estimates of the standard error, we estimate that between two groups defined by vital status at 5 years, the geometric mean CRT is 14.13% higher in the group which survives in 5 years. A 95% CI suggests that this observation is not unusual(?) if the true relationship between geometric means were such that the group which survives in 5 years geometric mean CRT were between 7.38% and 21.31% higher from the group that dies in 5 years. Because the two-sided P value is $P < .0005$, we reject the null hypothesis that there is no linear trend in the geometric mean CRT across the two groups.

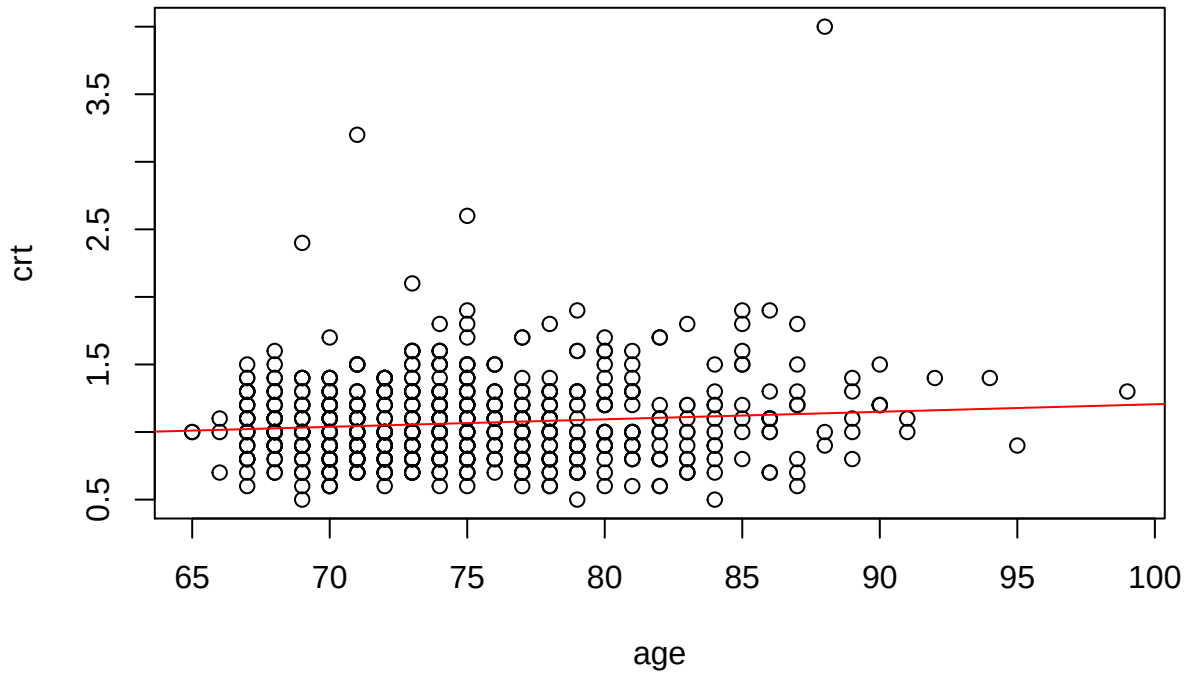
- Sample size 735
- Two samples skipped as 'NA' featured in them.
- Units for mean is mg/dl

Question 2

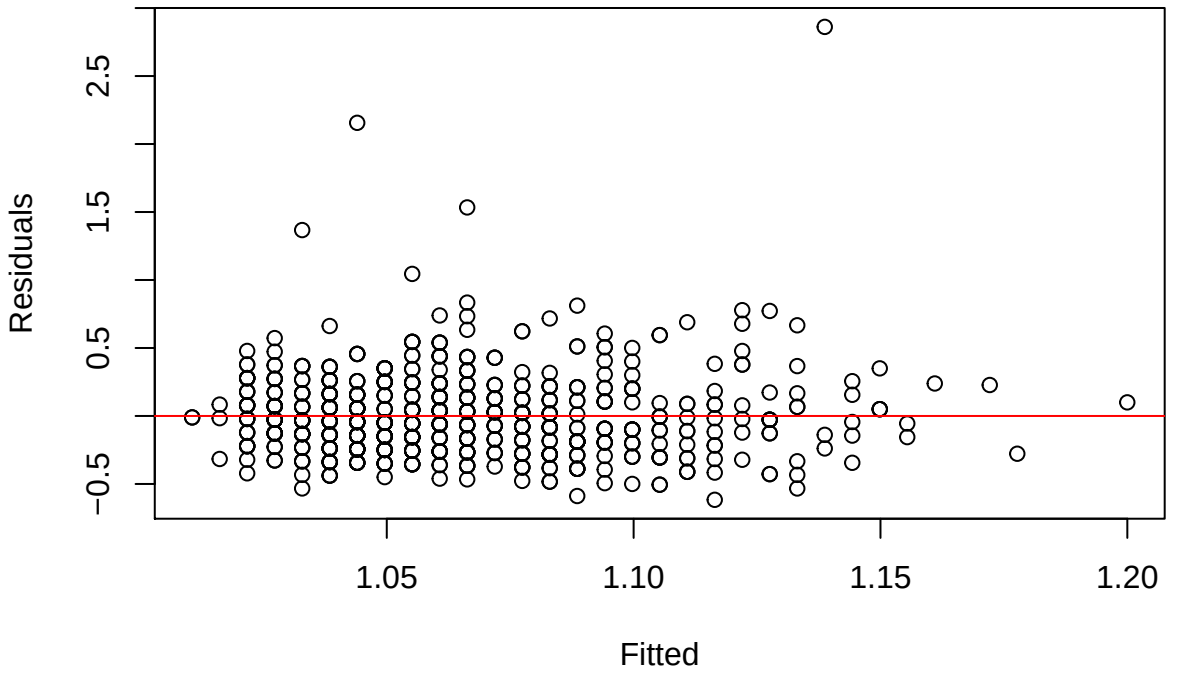
Perform a regression analysis evaluating an association between serum creatinine level and age by comparing mean serum creatinine levels across groups defined by age as a continuous variable. In your analysis, allow for heteroscedasticity. (Provide formal inference where asked to.)

(a)

Provide a brief description of the statistical methods for the model you fit to address the question of an association between creatinine and age.



We first do the some exploratory plots to so some look into relationship. From the above figure it looks like there is weak linear relationship. Lets look closely by plotting the residuals.



By looking at the residulas, it confirms the same. Lets now look at the linear analysis output.

From linear regression analysis using Huber-White estimates of the standard error, we estimate that for each year difference in age between two populations, the difference in mean CRT is 0.005 points higher in the older population. A 95% CI suggests that this observation is not unusual if the true difference in mean CRT were between 0.000494 and 0.01065 points higher per year difference in age. Because the two sided P value is $P < .0005$(p-value: 0.03154), we reject the null hypothesis that there is no linear trend in the average CRT across age groups.

- Two samples skipped as 'NA' featured in them.

- Units for mean is mg/dl

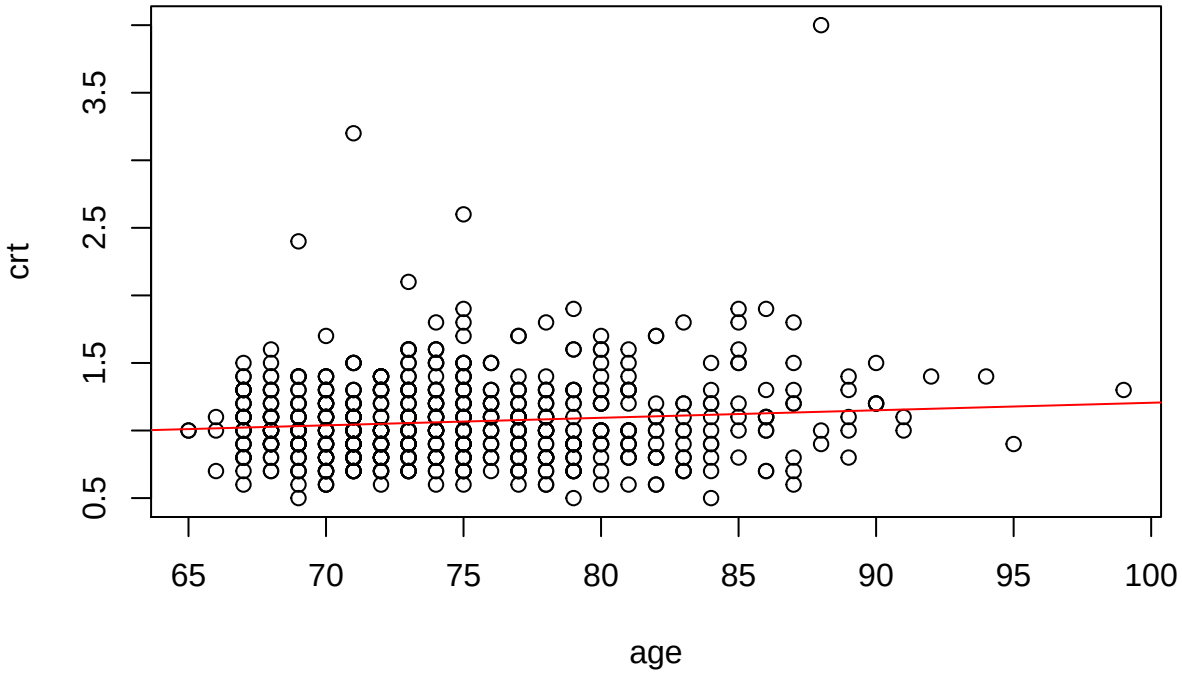
(b)

Is this a saturated model? Explain your answer.

This is not a saturated model, as number of parameters does not equal to number of coefficients, and predictor is a continous variable which can take range of values from 65 to 99.

(c)

Provide a scatterplot illustrating the relationship between serum creatinine and age and include in the plot the regression line from your regression analysis.



(d)

Based on your regression model, what is the estimated mean creatinine level among a population of 70 year old subjects?

E[CRTi|Agei]= 0.6485+ 0.005571×Agei

E[CRTi|Agei=70]=0.6485+ 0.005571×70=1.03847. Hence, estimated mean is 1.038 mg/dl.

(e)

Based on your regression model, what is the estimated mean creatinine level among a population of 85 year old subjects? How does the difference between your answer to this problem and your answer to part d relate to the slope?

E[CRTi|Agei=85]=0.6485+ 0.005571×85=1.122035. Hence, estimated mean is 1.122 mg/dl. The number of units between 85 and 70 is 15 years, and as slope is per unit(1 year) change, if you multiply 15 by slope, you get the difference in these two age groups i.e. 15*0.005571 = 1.122035 -1.03847 mg/dl.

(f)

Based on your regression model, what is the estimated mean creatinine level among a population of 101 year old subjects? Do you think this estimate is a reliable estimate for the mean creatinine of a population of 101 year old subjects? Briefly explain why or why not?

E[CRTi|Agei=85]=0.6485+ 0.005571×101=1.211171. Hence, estimated mean is 1.211 mg/dl. It is not reliable as our age range of subjects stops at 99.

(g)

What is the interpretation of the intercept in your model? Does it have a relevant scientific interpretation?

It is the mean CRT value when age is 0, it is not relevant as no subjects of age 0 is applicable.

(h)

What is the interpretation of the slope?

For each year difference in age between two populations, the difference in mean CRT is 0.005571 points (higher in the older population).

(i)

Provide full statistical inference about an association between serum creatinine and age based on your regression model.

From linear regression analysis using Huber-White estimates of the standard error, we estimate that for each year difference in age between two populations, the difference in mean CRT is 0.005571 points higher in the older population. A 95% CI suggests that this observation is not unusual if the true difference in mean CRT were between 0.000494 and 0.01065 points higher per year difference in age. Because the two sided P value is $P < .0005$(p-value: 0.03154), we reject the null hypothesis that there is no linear trend in the average CRT across age groups.

(j)

Suppose we wanted an estimate and a 95% CI for the difference in mean creatinine across groups that differ by 10 years in age. What would you report?

CI would be 0.00494 mg/dl to 0.106 mg/dl. From linear regression analysis using Huber-White estimates of the standard error, we estimate that for each 10 year difference in age between two populations, the difference in mean CRT is 0.05571 points higher in the older population.

Question 3

Now perform a regression analysis evaluating an association between serum creatinine levels and age by comparing the geometric mean of serum creatinine levels across groups defined by age as a continuous variable. In your analysis, allow for heteroscedasticity. (Provide formal inference where asked to.)

(a)

Provide a description of the statistical methods for the model you fit to address the question of an association between serum creatinine and age.

We used the linear regression method to look into association between log transformed CRT and age.

(b)

Based on your regression model, what is the estimated geometric mean serum creatinine level among a population of 70 year old subjects, 80 year old subjects, and 90 year old subjects.

Our model is E[log(CRT) | Agei] = -0.2459+0.003667 $\times$ Agei, so it gives the estimates of geometric mean.

70 yr olds - 0.01079 (Would we use the same units ??) - 1% higher 80 yr olds - 0.04746 - 4.7% higher 90 yr olds - 0.08413 - 8.4% higher

(c)

What is the interpretation of the intercept? Does it have a relevant scientific interpretation?

It gives the geometric mean CRT value when age is 0, it is not relevant as no subjects of age 0 is applicable.

(d)

What is the interpretation of the slope?

Ratio of geometric means of CRT between groups differing in age 1 year, older subjects having a .0367% higher CRT levels.

exponent(beta1) = e^0.003667 = 1.003674

(e)

Provide full statistical inference about an association between serum creatinine and age based on your regression model.

From linear regression analysis on log transformed CRT using Huber-White estimates of the standard error, we estimate that for every 1 year difference in age between two groups of children, the geometric mean CRT is .36% higher in the older population. A 95% CI suggests that this observation is not unusual if the true relationship between geometric means were such that the older group's geometric mean CRT were between 0.01% lower and .7% higher for each 1 year difference in age Because the two-sided P value is 0.05395, we fail to reject the null hypothesis that there is no linear trend in the average log CRT across height groups(alpha 0.05).

(f)

Provide an estimate and 95% confidence interval (CI) for the percent change in geometric mean serum creatinine between two groups that differ by 10 years in age.

exp(10*0.003667) is 1.037 i.e 3.7% higher when the groups differ by 10 years. The 95% CI is (0.9994 1.0768) i.e -.06% lower to 7.6% higher.

(g)

Compare your estimates of geometric mean serum creatinine level in question (b) to estimates of (arithmetic) mean serum creatinine levels for 70, and 80, and 90 year old subject from a linear regression model with serum creatinine levels as the response and age as the predictor, e.g., the regression model for problem 2 above. Briefly discuss any similarities or differences.

Put it in table (if needed)

Geometric mean 70 - 1.010848 mg/dl 80 - 1.048604 mg/dl 90 - 1.08777 mg/dl

Regression mean

70 - 1.03847 mg/dl 80 - 1.09418 mg/dl 90 - 1.14989 mg/dl