Homework 06

Spencer Pease

11/20/2019

(Q1) Diabetes and Race

(Q1.a)

Table 1: Presence of diabetes by self-reported race

	Non-diabetic	Diabetic
white	516	56
black	86	18
Asian	44	3

(Q1.b)

- H_0: There is no significant association between the the occurrence of diabetes and racial groups.
- H_a: There is a significant association between the occurrence of diabetes and racial groups.

(Q1.c)

Table 2: Expected cell counts of (1a) under the null hypothesis

	Non-diabetic	Diabetic
white	511.08	60.92
black	92.92	11.08
Asian	41.99	5.01

(Q1.d)

The Pearson χ^2-Square test is appropriate for testing if the observed counts are significantly different form the expected cell counts under the null hypothesis.

The distribution of the test statistic under the null hypothesis is equal to the distribution of the sum of 2 (from the degrees of freedom) random samples from a normal distribution squared.

(Q1.e)

Using a χ^2-Square test to see if there is a significant difference between the observed cell counts of presence of diabetes among individuals in different racial groups and expected cell counts under the null hypothesis produces a test statistics of 6.188. Given the test statistic is greater that 0 and has an associated P-value of 0.045 ($< .05$), we reject the null hypothesis that there is no significant association between diabetes and different racial groups.

(Q2) Creatinine Level and 5-Year All-Cause Mortality

(Q2.a)

Both models A and B are saturated since their predictor variable has the same number of groupings (survived 5+ years, died within 5 years) as the regression model has parameters (β_0, β_1). Or, put another way, each group can be fit exactly with the model - information does not need to be borrowed from across groups.

(Q2.b)

Models A and B show the same relationship, but in "opposite" directions. In both models, the slope β_1 is the difference between population group means and therefore has the same magnitude, differing only in sign depending on which direction the difference is taken. This also means, since each model parameter represents the estimated population mean of predictor group, $\beta_0^A + \beta_1^A = \beta_0^B$ and $\beta_0^B + \beta_1^B = \beta_0^A$.

Finally, since the models are equivalent ways of fitting the same data, the test statistic, standard error, and P-value are the same for the slope.

(Q2.c,d)

(Questions Q2 c and d use model A in the calculations)

For a population of subjects who survive at least 5 years:

- Estimate of the mean creatinine: **1.034**
- 95% confidence interval: **(1.011, 1.057)**

(Q2.e,f)

(Questions Q2 e and f use model B in the calculations)

For a population of subjects who die within 5 years:

- Estimate of the mean creatinine: **1.216**
- 95% confidence interval: **(1.163, 1.268)**

(Q2.g)

In model A, the intercept represents the estimated population mean creatinine level of individuals who survived at least 5 years. The slope of model A represents the estimated difference in population means between the population of individuals who dies within 5 years and the population which survived at least 5 years.

(Q2.h)

Table 3: Inference table for the difference in mean creatinine level between a population that survived 5+ years, and a population that died within 5 years

Point est.	Std. error	P-val	2.5 %	97.5 %
-0.182	0.029	1.07e-9	-0.239	-0.124

From looking at our 95% confidence interval (which doesn't include 0) and our P-value (which is $< .05$), we can conclude that there is a statistically significant association between mean creatinine level and survivorship status at 5 years.

(Q3) Creatinine Level and Age

(Q3.a)

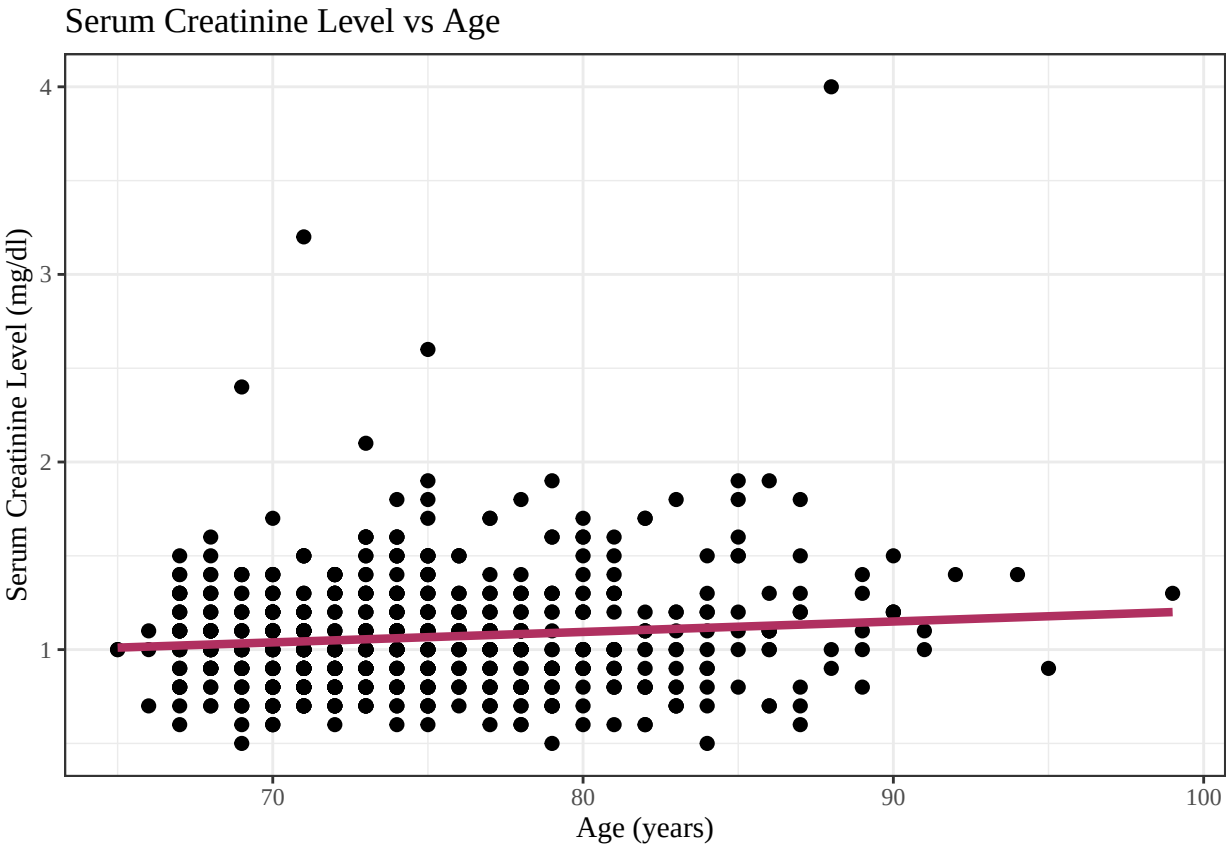
The statistical model used to determine the association between serum creatinine level and age fit the formula:

$$E(crt|age) = \beta_0 + \beta_1 \cdot age$$

With:

- $\beta_0 = 0.648$
- $\beta_1 = 0.006$

(Q3.b)



(Q3.c,d)

- The estimated mean serum creatinine level among a population of **72-year-old** subjects is: **1.05**.
- The estimated mean serum creatinine level among a population of **82-year-old** subjects is: **1.105**.

Since we fit the relationship between serum creatinine level and age to a linear model, the the difference in estimated mean between these two groups divided by the difference in age between these groups is the slope of the model.

(Q3.e)

- The estimated mean serum creatinine level among a population of **99-year-old** subjects is: **1.2**.

Since there is sparse data around the 99 year age group (only a single observation each at 99, 95, 94, and 92 years) the linear model will borrow information more heavily from other age groups, which may not be indicative of the actual mean value in the 99 year age group. Therefore, this estimate is not a reliable estimate of the mean creatinine level in a population of 99-year-old subjects.

(Q3.f)

In this model, we interpret the intercept as the mean creatinine level in a population of 0-year-old subjects. Scientifically, this is a meaningless interpretation since we fit a linear model on data for 65+ year-olds. The

intercept value is then only an artifact of the model.

(Q3.g)

The slope of this model is interpreted as the difference in mean creatinine level between two populations differing in age by a single year (one unit of the prediction variable). This has a meaningful scientific interpretation, as we can use this value to assess if there is a true difference in the mean response between two groups.

(Q3.h)

Table 4: Inference table for the association between creatinine level and age

Estimate	p-value	2.5 %	97.5 %
0.006	0.007	0.002	0.01

For the linear regression analysis of age and serum creatinine level, we estimate that for each year difference in age, there is a mean difference in creatinine level of 0.006 $\frac{mg}{dl}$. With 95% confidence we expect the true mean difference per year age to fall between 0.002 and 0.01 $\frac{mg}{dl}$. Since the two-sided P-value is 0.007, we reject a null hypothesis that there is no true linear relationship between mean creatinine level across age groups.

(Q3.i)

For differences in mean creatinine level across groups that differ by 5 years in age, the 95% confidence interval is:

Table 5: 95% CI for mean creatinine level across groups differing by 5 years in age

2.5 %	97.5 %
0.008	0.048