Homework 08

Spencer Pease

12/09/2019

(Q1) (OR) Response: 5-Year mortality; Predictor: High Serum Creatinine Level

In this question we look at a logistic regression model with an indicator of death within five years as the response (*death*), and an indicator of high serum creatinine level as the predictor (crt_H):

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 \cdot crt_H$$
$$Pr(death_i = 1) = p_i$$

(Q1.a)

This model is saturated, since the predictor variable (*high crt*) has the same number of groupings (*low*, *high*) 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.

(Q1.b)

For this model, we have:

- Intercept β_0 = -1.86
- Slope β_1 = 0.86

Here, the exponentiated intercept is interpreted as the odds of death within five years for the subset of the population that does not have a high serum creatinine levels.

The exponentiated slope is interpreted as odds ratio of death within five years between the subset of the population that has high creatinine levels and the subset that doesn't have high creatinine levels.

(Q1.c,d)

- Estimated **odds** of dying within 5 years for subjects with **low** creatinine levels: 0.16
- Estimated **probability** of dying within 5 years for subjects with **low** creatinine levels: 0.14
- Estimated **odds** of dying within 5 years for subjects with **high** creatinine levels: 0.37
- Estimated **probability** of dying within 5 years for subjects with **high** creatinine levels: 0.27

(Q1.e)

Table 1: Statistical inference of the relationship between binary indicator of death and binary indicator of high crt

Estimate	2.5%	97.5%	P-val
2.37	1.55	3.61	6.8e-5

From logistic regression analysis, we estimate that for two groups that differ by binary indicator of serum creatinine level ($high > 1.2\frac{mg}{dl}$, $low \leq 1.2\frac{mg}{dl}$), the odds of death are 2.37 times larger for the group with high creatinine levels. A 95% confidence interval suggests that this observation is not unusual if the true odd ratio were anywhere from 1.55 to 3.61. Because this interval does not include 1, in addition to the fact that the two-sided P-value for this estimate is less than .05, suggests that we can reject the null hypothesis that there is no association between five-year mortality and binary indicator of serum creatinine level.

(Q1.f,g)

Table 2: Comparison of different representations of the same underlying model

parameter	death~crtH	death~crtL	survival~crtH
Intercept	-1.86	-0.99	1.86
Slope	0.86	-0.86	-0.86

If our original model instead used an indicator of **low** serum creatinine level as the predictor, the statistical evidence for an association would not change, since changing the crt indicator will only change the sign of the slope, not the magnitude.

If our original model instead used an indicator of **surviving at least 5 years** as the response variable, the statistical evidence for an association would not change, only the starting point of the intercept and the direction of the slope.

(Q2) (OR) Response: High Serum Creatinine Level; Predictor: 5-Year mortality

In this question we look at a logistic regression model with an indicator of high serum creatinine level as the response (crt_H), and an indicator of death within 5 years as the predictor ($death$):

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 \cdot death$$

$$Pr(crt_{Hi} = 1) = p_i$$

(Q2.a)

For this model, we have:

- Intercept β_0 = -1.42
- Slope β_1 = 0.86

Here, the exponentiated intercept is interpreted as the odds of having a high creatinine level for the subset of the population that has survived at least five years

The exponentiated slope is interpreted as odds ratio of having a high creatinine level between the subset of the population that died within five years and the subset that survived at least five years.

(Q2.b,c)

- Estimated **odds** of high creatinine level for subjects who **die** within 5 years: 0.57
- Estimated **probability** of high serum creatinine for subjects who **die** within 5 years: 0.36
- Estimated **odds** of high creatinine level for subjects who **survive** at least 5 years: 0.24
- Estimated **probability** of high serum creatinine for subjects who **survive** at least 5 years: 0.19

(Q2.d)

Table 3: Statistical inference of the relationship between binary indicator of high crt and binary indicator of death

Estimate	2.5%	97.5%	P-val
2.37	1.55	3.61	6.8e-5

From logistics regression analysis, we estimate that for two groups that differ by binary indicator of five-year all-cause mortality, the odds of having high serum creatinine level ($crt > 1.2\frac{mg}{dl}$) are 2.37 times larger in the group that died within five years than the group that survived at least five years. A 95% confidence interval suggests this observation is not unusual if the true odds ratio is anywhere from 1.55 to 3.61. Since this interval does not include 1 and our two-sided P-value is much less than .05, we have sufficient confidence to reject the null hypothesis that there is no association between binary indicator of serum creatinine level and five-year mortality.

(Q2.e)

Both this retrospective analysis and the prospective analysis from *(Q1)* produce the same estimate of the odds ratio, as well as the confidence interval and P-value of that estimate. This is because both models are testing for association between the same two indicators, with only the response and predictor labels swapped.

(Q3) (RD) Response: 5-Year mortality; Predictor: High Serum Creatinine Level

This question asks us to assess the association between 5-year mortality and serum creatinine level via a risk difference (RD) contrast. Since our response is binary ($death$), we can use a linear model like:

$$E[death_i|crt_{Hi}] = \beta_0 + \beta_1 \cdot death_i$$

(Q3.a,b)

For this model, we have:

- Intercept $\beta_0 = 0.14$
- Slope $\beta_1 = 0.13$

Here, the intercept is interpreted as the estimated mean proportion of death within five years for the subset of the population that does not have high serum creatinine levels.

The slope is interpreted as the estimated difference in population means of death within five years between the two population subsets defined by high serum creatinine level indicator.

(Q3.c)

Table 4: Statistical inference of the difference in mean estimates of proportion of death within 5 years between indicators of high creatinine

Estimate	2.5%	97.5%	P-val
0.13	0.06	0.21	3.65e-4

From a linear regression analysis of five-year all-cause mortality indicator and high serum creatinine level ($crt > 1.2\frac{mg}{dl}$) indicator, using Huber-White estimates of the standard error, we find that the estimated risk difference in probability of death between the population of individuals with high creatinine level and those without is 0. A 95% confidence interval suggests this estimate is not unusual if the true risk difference is anywhere from 0 to 0. Because this interval does not include 0, and since the P-value for this estimate is much less than .05, we can reject the null hypothesis that there is no association between high serum creatinine level indicator and five-year mortality with confidence.

(Q4)

(Q4.a)

(Q4.a.i)

- Intercept β_0 = -3.6
- Slope $\beta_1 = 1.79$

Here, the exponentiated intercept is interpreted as the odds of death within five years for the subset of the population that has a mean serum creatinine level of 0. This is not useful scientifically, as we never observed an individual with level of less than .05.

The exponentiated slope is interpreted as the ratio of the odds of death between two subsets of the population differing by one unit of mean creatinine level. Scientifically, this is a useful value, because it allows us to test for difference between two subsets of the population.

(Q4.a.ii)

Table 5: Statistical inference of the OR relationship between creatinine level and binary indicator of death

Estimate	2.5%	97.5%	P-val
5.99	3.12	11.5	0.01e-5

From logistic regression analysis, we estimate that for two groups differing in mean serum creatinine level by one unit ($1\frac{mg}{dl}$), the odds of death within five years increases by 5.99 times. A 95% confidence interval suggests this observation is not unusual if the true odds ratio is anywhere from 3.12 to 11.5. Since this interval does not include 1, and the two-sided P-value for this estimate is less than .05, we can, with confidence, reject the null hypothesis that there is no association between serum creatinine level and five-year mortality.

(Q4.b)

(Q4.b.i)

- Intercept β_0 = -0.13
- Slope β_1 = 0.27

Here, the intercept is interpreted as the estimated mean proportion of death within five years for the subset of the population with a mean serum creatinine level of $0\frac{mg}{dl}$. Again, this is not scientifically useful, as there are no observed individuals with a measured creatinine level of less than .05.

The slope is interpreted as the estimated difference in mean probability of death within five years between the two population subsets that differ by $1\frac{mg}{dl}$ of mean serum creatinine level. This does have scientific value, as it lets us estimate differences across groups within our population.

(Q4.b.ii)

(Q4.b.iii)

Table 6: Statistical inference of the RD relationship between creatinine level and binary indicator of death

Estimate	2.5%	97.5%	P-val
0.27	0.18	0.36	3.83e-9

From a linear regression analysis of five-year all-cause mortality and and serum creatinine level, using Huber-White estimates of the standard error, we find the estimated risk difference in probability of death between between two groups differing by one unit of mean serum level ($1\frac{mg}{dl}$) increases by 0.27. A 95% confidence interval suggest this estimate is not unusual if the true risk difference is in the range 0.18 to 0.36. 1 not being included in this interval and a two-sided P-value less than .05 suggest we can with high confidence reject the null hypothesis that there is no association between serum creatinine level and five-year mortality.

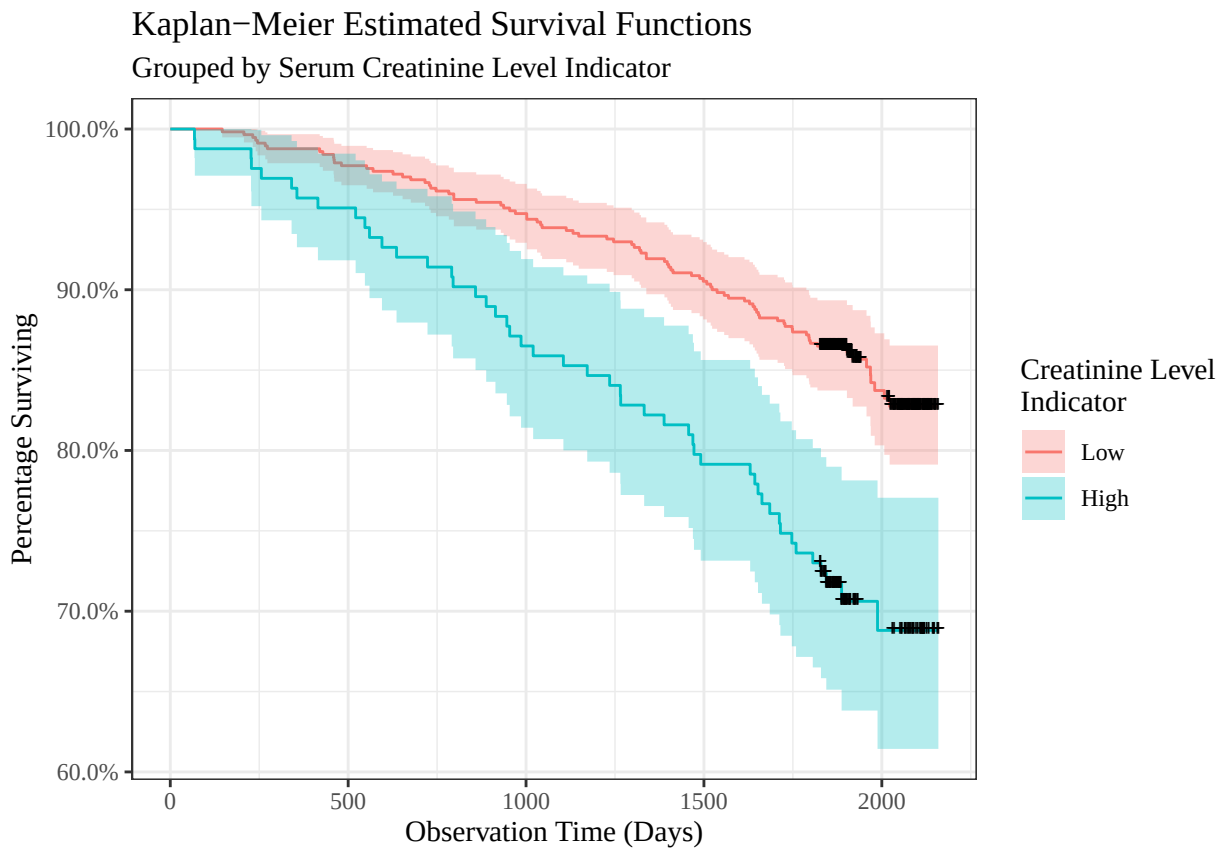
(Q4.c)

Table 7: Comparison of probability of death between OR and RD models with continuous crt variable

crt	OR	RD
0.8	0.10	0.09
1.8	0.41	0.37
2.8	0.80	0.64
3.8	0.96	0.91

(Q5) Survival Analysis

(Q5.a)



(Q5.b)

The high creatinine level indicator curve decays faster than the low indicator, though both decay at independent steady rates until about 1500 days, at which point both decays increase.

(Q5.c)

Table 8: Association of time-to-death and serum creatinine level

crt Indicator	N obs	Observed	Expected
Low	85	85	105.66
High	48	48	27.34

From the Kaplan-Meier survival estimates of time-to-death and high serum creatinine level indicator, the P-value of $9e-6$ and χ^2 value of 19.66 suggest that we can reject the null hypothesis that there is no association between survival time and high serum creatinine level indicator.