Homework 4

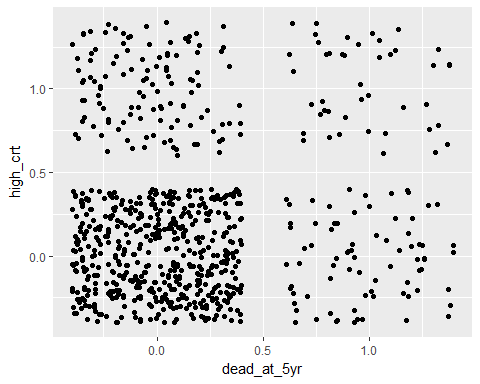
Piotr Mankowski

February 2, 2018

## Question 1

**Prospective logistic regression analysis of the association between serum creatinine level and 5-year all-cause mortality. Mortality is the response variable, and serum creatinine level is the predictor.**

### Response vs. Predictor



##   
## Call:  
## glm(formula = dead\_at\_5yr ~ high\_crt, family = "binomial", data = mri)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7933 -0.5387 -0.5387 -0.5387 2.0009   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.8567 0.1225 -15.152 < 2e-16 \*\*\*  
## high\_crt 0.8618 0.2148 4.012 6.03e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 656.75 on 732 degrees of freedom  
## Residual deviance: 641.50 on 731 degrees of freedom  
## (2 observations deleted due to missingness)  
## AIC: 645.5  
##   
## Number of Fisher Scoring iterations: 4

1. *Is this a saturated model? Explain your answer.*

In a saturated regression model, the number of groups of the predictor variable equal the number of parameters in the model. Logstic regression fits a single parameter, . Since the predictor variable is binomial (high and low crt), it has two groups. This is more than the number of parameters, so this is not a saturated regression model.

1. *Interpretation of the slope and intercept.*

## [1] "Exponentiated Slope: 2.367 Exponentiated Intercept: 0.156"

The intercept gives us the log odds when the predictor is 0, while the transformed intercept () represents the odds of the response when the predictor is 0. In this case, the intercepts gives us the odds of being dead at 5 years given you have low creatinine levels.

The slope gives us the estimated difference in log odds of being dead at 5 years for two groups differing by 1 in the predictor variable. Since our predictor is binary, the slope gives us the difference in log odds between the two mortality groups. When exponentiated, the intercept () gives us the odds ratio between the two groups.

In this case, the model estimates that the high-creatinine group (high\_crt == 1) has a higher odds of being dead at 5 years than the low-creatinine group.

1. *Estimated odds and probability of dying within 5 years for subjects with low creatinine levels.*

To get estimated odds of dying within 5 years (‘dead\_at\_5yr == 0’) for subjects with low creatinine (‘high\_crt == 0’), we can use the value for the exponentiated intercept, since it represents the odds when the predictor (‘high\_crt’) is 0. We can get the estimated probability by tranforming the odds, since and

## [1] "Odds: 0.156 Probability: 0.135"

1. *Comparing sample odds/probability with estimates for low creatinine*

## [1] "Sample odds: 0.156 Sample proportion: 0.135"

The sample odds and proportion for those who died within 5 years in teh low serum creatinine group are virtually equal to the estimates we got in the logistic regression. This simmilarity is not suprising, since the model was trained on this data, and would be expected to make the estimates based on the actual proportions in each of the four predictor/response groups (since each variable is binary, we have 4 quadrants in the data, and our logistic regression is simmilar to a test.)

1. *Estimated odds and probability of dying within 5 years for subjects with high creatinine levels.*

To get the odds and probability in the high-creatinine group, we can use the intercept and multiply it by the slope. Since the slope represents the odds ratio between the groups (), we can get the odds for the high creatinine group:

## [1] "Odds: 0.370 Probability: 0.270"

1. *Comparing sample odds/probability with estimates for high creatinene*

sample\_p\_1 <- mri[high\_crt == 1, mean(dead\_at\_5yr)]  
sample\_o\_1 <- sample\_p\_1/(1 - sample\_p\_1)  
  
sprintf("Sample odds: %.3f Sample proportion: %.3f", sample\_o\_1, sample\_p\_1)

## [1] "Sample odds: 0.370 Sample proportion: 0.270"

Again, the sample odds and proportion match the estimates from the logistic regression model. This simmilarity is still not suprising, since the model was trained on this data, and would be expected to make the estimates based on the actual proportions in the sample.

1. *Full inference between 5-year all-cause mortality and high creatinine levels.*

We performed logistic regression analysis to investigate the association between 5-year all-cause mortality and high serum creatinine. From the analysis, we estimate that the odds of dying are 137% higher in the high-creatinine group than the low-creatinine group. A 95% CI suggests that the results would not be suprising if the real odds of dying within 5 years were between 94.6% and 179% higher in the high-creatinine group. A two-sided p-value of < 0.0005 suggests we should reject the null hypothesis that creatinine level has association with 5-year all-cause mortality.

1. *Effect of using low creatinine level as predictor*

In this scenario, the predictor groups would swap, with individuals labeled by a 1 in the high-creatinine group recieving a 0 in the low-creatinine group, and vice-versa. Nothing else about the analysis would change; however, swapping the 1’s and 0’s means that the intercept would represent the odds of dying for high-creatinine group, and the slope would adjust, with the new slope , since the odds would decrease for the low-creatinine group. The statistical evidence would not change in any way; the labeling is arbitrary.

1. *Effect of using 5-year survival as response variable*

As in part h, swapping the arbitrary labeling of the response variable would affect the interprestation of the resulting estimates, but not change the statsitical evidence for the association between mortality and serum creatinine. The intercept would represent the odds of surviving 5 years () in the low-creatinine (high\_crt == 0) group, and the slope would represent the odds ratio of 5-year survival between the two groups. Since the probability of survival () is related to the probability of death () by , . Therefore, the new intercept would = the inverse of the old intercept, and the new slope the inverse of the slope from part b.

## Question 2

**Retrospective logistic regression analysis of the association between serum creatinine level and 5-year all-cause mortality. Mortality is the predictor variable, and serum creatinine level is the response**

mod2 <- regress("odds", high\_crt~dead\_at\_5yr, data=mri)  
mod2.pointest <- mod2$transformed[2,1]  
mod2.intercept <- mod2$transformed[1,1]  
mod2.95ci <- c(mod2$transformed[2,2], mod2$transformed[2,3])  
  
sprintf("Slope: %.2f | Intercept: %.3f | 95%% CI: (%.2f, %.2f)", mod2.pointest, mod2.intercept, mod2.95ci[1], mod2.95ci[2])

## [1] "Slope: 2.37 | Intercept: 0.241 | 95% CI: (1.55, 3.61)"

1. *Interpretation of slope and intercept* After running the analysis and transforming the output from log odds to odds, we find the estimated slope and intercept to be 2.37 and 0.241 respectively.

In this scenario, the intercept represents the odds of the response (high serum creatinine levels, ‘high\_crt == 0’) when 5-year all-cause mortality is false (‘dead\_at\_5yr == 0’); in other words, it represents the odds of having had high creatinine levels for those who survived at least 5 years after the levels were measured.

The slope of 2.37 represents the odds ratio of having had high creatinine between those who died within 5 years and those who survived at least 5 years. In other words, the odds of having had high creatinine for those who died within 5 years (‘dead\_at\_5yr == 1’) are higher than for those surviving at least 5 years.

1. *Probability and odds of high creatinine in those dead at 5 years* To get the odds for high creatinine in this group where ‘dead\_at\_5yr == 1’, we can use the intercept (odds for ‘dead\_at\_5yr == 0’) and multiply it by the slope (odds ratio between the two groups).

## [1] "Estimated odds: 0.571 | Estimated probability: 0.364"

1. *Probability and odds of high creatinene in those surviving 5 years* Since the predictor variable value in this case == 0, we can use the intercept as an estimate for the odds of this group. Then we just have to tranfrom it into probability.

## [1] "Estimated odds: 0.241 | Estimated probability: 0.194"

1. *Full inference on retrospective association between 5-year all-cause mortality and high serum creatinine levels*

## [1] "Slope: 2.37 | Intercept: 0.241 | 95% CI: (1.55, 3.61)"

From logistic regression analysis, we estimate that the odds of having had high serum creatinine levels are 137% higher in individuals who die within 5 years than in individuals who survive at least 5 years. A 95% CI suggests that this observation would not be unusual if the true % difference between the two groups is between 55% and 261% larger in those who experience 5-year all-cause mortality. A two-sided p-value of 1\*10^-4 suggests that, at , we can reject the null hypothesis of the odds being of high creatinine levels being the same in the two mortality groups.

1. *Comparison of 1g and 2d*

## Question 3

*Investigating the association between odds of death within 5 years and non-dichotomozed serum creatinine levels*

In this case, the response variable - 5-year all-cause mortality (‘dead\_at\_5yr’) - is binary, while the predictor variable - serum creatinine level, (‘crt’) - is continuous. We can use a logistic regression to perform inference.

mod3 <- glm(dead\_at\_5yr~crt, family="binomial", data=mri)  
mod3.pointest <- exp(summary(mod3)$coefficients['crt', "Estimate"])  
mod3.intercept <- exp(summary(mod3)$coefficients["(Intercept)", "Estimate"])  
mod3.95ci <- exp(confint.default(mod3))  
mod3.pval <- summary(mod3)$coefficients[2,4]  
  
sprintf("Slope: %.3f | Intercept: %.4f | 95%% CI: (%.4f, %.2f) | p-value: 1.4e-07", mod3.pointest, mod3.intercept, mod3.95ci[2,1], mod3.95ci[2,2], mod3.pval)

## [1] "Slope: 5.986 | Intercept: 0.0272 | 95% CI: (3.0751, 11.65) | p-value: 1.4e-07"

1. *Interpretation of the slope and intercept* The exponentiated intercept of 0.0272 represents the odds of death at 5 years for those with a serum creatinine level of 0. In reality, creatinine levels of 0 might not even be biologically viable; in our dataset, the range of creatinine levels is (0.5, 4.0). In this case, the intercept does not seem to have a meaningful interpretation.

The exponentiated slope of 5.99 represents the odds ratio of death within 5 years between groups with a 1-unit (in this case, mg/dl) difference in creatinine. The slope suggests that the odds of death within 5 years are 499% higher in groups with higher creatinine levels by 1 unit.

1. *Full Inference* From our logistic regression analysis, we estimate that for two groups that differ by one mg/dl in serum creatinine levels, the odds of death within 5 years are 499% higher in the higher-creatinine group.

A 95% CI suggests that this observation is not unusual if a group if the true difference in odds of 5-year all-cause mortality is between 208% and 1065% higher in the high-creatinine group.

A two-sided p value of 1.4e-07 suggest that we can reject, at , the null hypothesis that there is no association of 5-year all-cause mortality and serum creatinine.