```
# solutions for HW 5
# R code
# math 150
# Fall 2010
# Jo Hardin


sepsis <- read.table("sepsis.csv",header=T,sep=",")
attach(sepsis)


# 1
sepsis.bu <- sepsis[(race==1)&(treat==0),]


sepsis.bu.log <- glm(fate ~ apache, family="binomial", data=sepsis.bu)


summary(sepsis.bu.log)$coef
#                Estimate Std. Error   z value   Pr(>|z|)
# (Intercept) -0.80765133  0.6658397 -1.212982 0.2251368
# apache       0.06147948  0.0351579  1.748668 0.0803485
```

The probability of death for a given APACHE score (x) in black untreated patients is:

p(x) = exp (-0.807 + 0.061 x) / (1 + exp (-0.807 + 0.061 x))

```
plot(apache,fate,pch=19,xlim=c(1,45),ylab="probability of dying", xlab="APACHE
score")
lines(c(1:45), exp (-0.807 + 0.061*c(1:45)) / (1 + exp (-0.807 + 0.061*c(1:45))),
        lty=3)
```

2 The odds ratio associated with a unit rise in APACHE score in untreated black patients is e^beta for the appropriate model (above sepsis.bu.log).

beta = 0.0614; e^beta = 1.06341

A 95% CI for beta is: 0.06148 +- 1.96 * 0.035158 = (-0.00742968, 0.1303897)

A 95% CI for e^beta is: (e^-0.00742968, e^0.1303897) = (0.9925979, 1.139272)

We are 95% confident that your odds of death changes by a factor of between 0.9925979 and 1.139272 for a one unit increase in APACHE score. Note that the number 1 is in

the interval, so I am unable to claim whether the odds of death go up, down, or remain the same for a one unit increase in APACHE score. Also note, however, that a

smaller level of confidence would have given me a significant result. So I interpret these results to be suggestive if not significant.

3

Yes, it is possible to predict the probability of deat for a treated black patient with an APACHE score of 50:

pi(50) = exp (-2.43 + 0.121*50) / (1 + exp (-2.43 + 0.121*50)) = 0.974

A black treated patient with an APACHE score of 50 has a probability of dying equal to 0.974. Notice, however that 50 is outside the range of observations used to fit

the model. Typically we do not like to apply our models to values outside the range (this is called extrapolation). However, we may sometimes apply it to x-values that

are only slightly larger than what we observed.

5

The APACHE score which gives median survival is the x value that gives e^0 / (1+ e^0) = 1/2 as a probability of survival. This happens when:

x = -b0/b1 = 0.80765133 / 0.06147948 = 13.137

A patient with an APACHE score of 13.137 has an equal probability of survival and death.

Note, you can calculate the APACHE score for any probability of survival. For example, let's say we wanted the 0.9 survival value.

0.9 = e^(b0+b1 x) / [1+ e^(b0+b1 x)]
0.9 * [1+ e^(b0+b1 x)] = e^(b0+b1 x)
0.9 = 0.1 * e^(b0+b1 x)
9 = e^(b0+b1 x)
ln(9) = b0 + b1 x
x = (ln(9) - b0) / b1

6

Our model is that E[Y] = exp (alpha + beta * x) / (1 + exp (alpha + beta * x))

E[Y] is the average Y value in the population (expected value just means that "the average in the population"). Because the response variable is binary, we can think of

an average of 0s and 1s as a proportion of 1s (or proportion of successes).

Because a proportion is just an average, and our particular proportion is a function of the explanatory variables (the logit of the linear function), logistic regression

is a very appropriate description.

In linear regression we are modeling the expected value (i.e., long run _average_) for a response variable at every possible X. Our model is reasonably sophisticated

because we aren't simply finding Y-bar at each X and connecting the dots. Instead, we believe that the true mean values fall on the line, so we use the entire set of

data to model the relationship as a linear function.

For logistic regression, we are modeling the expected proportion / probability (which is an average!) of success at every possible X. Again, we don't find p-hat at every

value and connet the dots. Instead, we believe that the true proportions lie on a

line related by the logit function of the true value to a linear function of the explanatory variable.

7

(a) The response variable (y) is never normally distributed around the line.

(b) The errors will not have constant variance across all values of x

(c) Often, the best line will go outside the bounds of (0,1) which isn't meaningful if we think of the predicted response as probability of success.

(d) The idea of the line being the average response at a particular value of the explanatory variable (x) is NOT necessarily violated (see # 6 above).