ES/STT 7140: Homework 5

Brandon M. Greenwell

Monday, March 26, 2018 (due Monday, April 2, 2018)

Question 1 Early diagnosis of heart attack is important if the best care is to be given to patients. One suggested diagnostic aid is the level of the enzyme creatinine kinase (CK) in the blood stream. A study was conducted (Smith, 1967) in which the level of CK was measured for 360 patients suspected of suffering from a heart attack. Whether or not each patient had really suffered a heart attack was established later, after more prolonged medical investigation. The original paper classified patients according to ranges of CK level, but in the this problem only midpoints of the range are given.

It would be good to be able to base diagnostic criteria on data like these, so that CK level can be used to estimate the probability that a patient has had a heart attack. We can go some way towards such a goal, by constructing a model which tries to explain the proportion of patients suffering a heart attack, from the CK levels. In the following code chunk, the data are loaded into an R data frame called `heart`. It contains the variables `ha`, `ok`, and `ck`, giving numbers of patients who subsequently turned out to have had, or not to have had, heart attacks, at each CK level.

```
heart <- data.frame(
  ck = 0:11*40 + 20,
  ha = c(2, 13, 30, 30, 21, 19, 18, 13, 19, 15, 7, 8),
  ok = c(88, 26, 8, 5, 0, 1, 1, 1, 1, 0, 0, 0)
)
```

a) It makes sense to plot the observed proportions against CK level first. Using the code below, obtain a plot of the observed proportion of heart attacks against CK level and comment.

```
plot(
  x = heart$ck,                          # x-axis: CK level
  y = heart$ha/(heart$ha + heart$ok),    # y-axis: proportion of heart attacks
  xlab = "Creatinine kinase level",      # x-axis label
  ylab = "Proportion heart attack",      # y-axis label
  las = 1,                               # horizontal y-axis labels
  pch = 19,                              # point shape
  col = "dodgerblue2",                   # point color
  cex = 1.2                              # point size
)
```

b) Use the code chunk below to fit a simple logistic regression model of the form

$$logit\,(p) = \beta_0 + \beta_1 CK,$$

where p is the probability of a heart attack as a function of CK. Interpret the slope parameter in the context of this problem (e.g., what does $e^{\widehat{\beta}_1}$ mean here?).

```
fit1 <- glm(cbind(ha, ok) ~ ck, family = binomial(link = "logit"),
            data = heart)
```

c) What is the estimated proportion of heart attacks for a CK level of 100?

d) What is the estimated CK level that corresponds to a proportion of heart attacks of 0.8?

e) Standard residual plots (not discussed in class for logistic regression models) would suggest fitting a more complex model with a cubic term. (For the curious minded, you can obtain one such plot using `plot(fit1, which = 1)`, which suggests the possibility of adding a quadratic and cubic term to the

model.) Use the code chunk below to fit such a model and test the null hypothesis of whether or not the more complicated model should be used in favor of the simpler model fit in part a). **Hint:** use the `anova()` function as we did in class. Also, be sure to write out specifically the null and alternative hypotheses (i.e., H_0 and H_1) being tested here.

```
fit2 <- glm(cbind(ha ,ok) ~ ck + I(ck^2) + I(ck^3),
            family = binomial(link = "logit"), data = heart)
```