STAT 512 – Assignment 5

Vignesh M. Pagadala
Vignesh.Pagadala@ColoState.Edu

1. For this problem use the data described in Ott and Longnecker Example 12.22 (p 664 in the 7 th edition). The data are available from Canvas as "CKheart.csv". Read the description of the data in the book. You can use the output in the book to check your own R calculations.

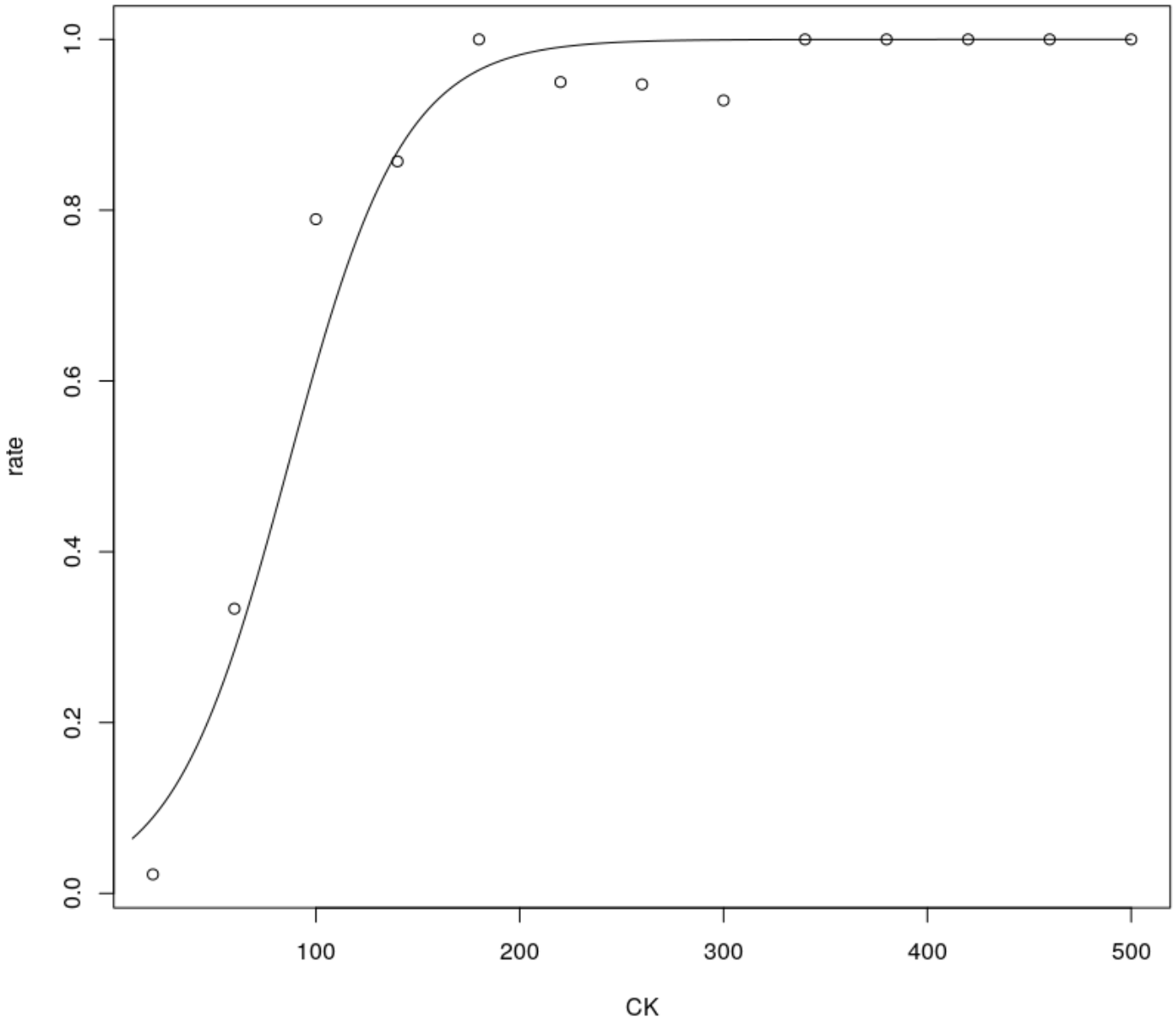
A. Use glm() to fit a logistic regression model that estimates the probability of a heart attack as a function of CK value. Include the Coefficients table in your assignment.

ANSWER:

```
Call:
glm(formula = cbind(withHA, withoutHA) ~ CK, family = binomial(link = "logit"),
    data = InData)

Deviance Residuals:
      Min        1Q    Median        3Q       Max
-2.79579  -1.34637   0.00587   0.07173   2.26860

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.028360   0.366977  -8.252   <2e-16 ***
CK           0.035104   0.004081   8.602   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 311.29  on 12  degrees of freedom
Residual deviance:  28.14  on 11  degrees of freedom
AIC: 51.596

Number of Fisher Scoring iterations: 6
```

B. Construct a plot of the data with the fitted logistic regression curve overlaid. Include the plot in your assignment.

ANSWER:



C. Give an estimate of the odds ratio corresponding to CK and an approximate 95% confidence interval.

ANSWER:

```
(Intercept)          CK
 0.04839496  1.03572783

            2.5 %     97.5 %
(Intercept) 0.02234051 0.09487507
CK          1.02809855 1.04473381
```

D. Give a one-sentence description of the odds of heart attack among those with a given level of CK, compared to the odds of a heart attack among those with a level of CK ten points higher. (4 pts)

ANSWER:
The individual with a CK value which is 10 points higher is 1.4 times more likely to suffer a heart attack than the individual with a lower CK value.

E. Calculate McFadden's pseudo R 2 for the model.

ANSWER:
```
'log Lik.' 0.8560926 (df=2)
```

F. Give an estimate of the CK level at which doctors would be 90% sure that a subject has had a heart attack.

ANSWER:
```
              Dose       SE
p = 0.10:  23.67610 8.196009
p = 0.15:  36.85461 7.092263
p = 0.20:  46.77663 6.379136
p = 0.25:  54.97168 5.896211
p = 0.30:  62.13074 5.574251
p = 0.35:  68.63302 5.377604
p = 0.40:  74.71699 5.285684
p = 0.45:  80.55086 5.285667
p = 0.50:  86.26725 5.369401
p = 0.55:  91.98365 5.532264
p = 0.60:  97.81752 5.773095
p = 0.65: 103.90149 6.094916
p = 0.70: 110.40377 6.506617
p = 0.75: 117.56283 7.026380
p = 0.80: 125.75788 7.689113
p = 0.85: 135.67990 8.565022
p = 0.90: 148.85841 9.817534
```

2. An observational study was done to investigate risk factors associated with low infant birth weight. Data from 189 (singleton) pregnancies were collected at Baystate Medical Center, Springfield, MA during 1986. The response variable was low (1 if birth weight was less than 2.5 kg, 0 otherwise). The predictor variables included: age (mother's age in years), mwt (mother's weight in pounds prior to pregnancy), race (mother's race, 1= white, 2=black, 3=other) and smoke (1=mother smoked during pregnancy, 0 otherwise). The data is available from Canvas as "birthweight.csv". Important note: Be sure to define race and smoke as factors!

A. To examine the relationship between low vs race: calculate the proportion of births resulting in low birthweight for each race category and present the p-value from a chi-square test. (4 pts)

ANSWER:

```
     low
race         0         1
   1 0.7604167 0.2395833
   2 0.5769231 0.4230769
   3 0.6268657 0.3731343
```

B. To examine the relationship between low vs smoke: calculate the proportion of births resulting in low birthweight for each smoke category and present the p-value from a chi-square test. (4 pts)

ANSWER:

```
      low
smoke         0         1
     0 0.7478261 0.2521739
     1 0.5945946 0.4054054

         Pearson's Chi-squared test with Yates' continuity correction

data:  Table2
X-squared = 4.2359, df = 1, p-value = 0.03958
```

C. Run a logistic regression with smoke as the only predictor variable. Calculate the emmeans using type = "response" for each smoke group (copy/paste the results to your assignment). Note: these should match your simple proportions from part B. (4 pts)

ANSWER:

```
 smoke  prob     SE  df asymp.LCL asymp.UCL
 0     0.252 0.0405 Inf     0.181     0.339
 1     0.405 0.0571 Inf     0.300     0.520

Confidence level used: 0.95
```

```
Intervals are back-transformed from the logit scale
```

D. Now consider all 4 predictors (age, mwt, race, smoke). Using best subsets selection with AIC criteria, which variables are included in the final model? Include the Coefficients table and Type3 Anova table in your assignment. (4 pts)
NOTE: Use the selected model from the previous question for all further questions!

ANSWER:

Model selection table

```
     (Intrc)      age      mwt race smoke df   logLik   AIC delta weight
15 -0.10920          -0.01326    +     +  5 -107.507 225.0  0.00  0.475
16  0.33250 -0.02248 -0.01253    +     +  6 -107.289 226.6  1.56  0.218
13 -1.84100                      +     +  4 -109.987 228.0  2.96  0.108
14 -1.00800 -0.03488             +     +  5 -109.431 228.9  3.85  0.069
11  0.62200          -0.01332          +  3 -112.170 230.3  5.33  0.033
12  1.36800 -0.03899 -0.01214          +  4 -111.440 230.9  5.86  0.025
7   0.80580          -0.01522    +        4 -111.630 231.3  6.24  0.021
8   1.30700 -0.02552 -0.01435    +        5 -111.330 232.7  7.65  0.010
3   0.99830          -0.01406             2 -114.345 232.7  7.68  0.010
4   1.74900 -0.03979 -0.01278             3 -113.562 233.1  8.11  0.008
10  0.06091 -0.04978                   +  3 -113.638 233.3  8.26  0.008
9  -1.08700                            +  2 -114.902 233.8  8.79  0.006
5  -1.15500                      +        3 -114.831 235.7 10.65  0.002
2   0.38460 -0.05115                    2 -115.956 235.9 10.90  0.002
6  -0.20800 -0.03951             +        4 -114.064 236.1 11.11  0.002
1  -0.79000                             1 -117.336 236.7 11.66  0.001
Models ranked by AIC(x)
```

Call:

```
glm(formula = low ~ mwt + race + smoke, family = binomial, data = BirthData)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.5278  -0.9053  -0.5863   1.2878   2.0364

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.10922    0.88211  -0.124  0.90146
mwt         -0.01326    0.00631  -2.101  0.03562 *
race2        1.29009    0.51087   2.525  0.01156 *
race3        0.97052    0.41224   2.354  0.01856 *
smoke1       1.06001    0.37832   2.802  0.00508 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 215.01  on 184  degrees of freedom
AIC: 225.01

Number of Fisher Scoring iterations: 4
```

Analysis of Deviance Table (Type III tests)

```
Response: low
       LR Chisq Df Pr(>Chisq)
mwt     4.9601  1   0.025939 *
race    9.3260  2   0.009438 **
smoke   8.2444  1   0.004088 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

E. Based on the model selected above, give the estimated odds ratio and corresponding 95% CI for Smokers vs Non-Smokers (smoke 1 vs 0).

ANSWER:

```
(Intercept)         mwt       race2       race3      smoke1
  0.8965324   0.9868281   3.6331297   2.6393030   2.8863880

Waiting for profiling to be done...
                    2.5 %     97.5 %
(Intercept) 0.1658131  5.3953378
mwt         0.9738926  0.9984753
race2       1.3380529 10.0823263
race3       1.1927957  6.0578281
smoke1      1.3945161  6.1980517
```

F. Calculate the emmeans using type = “response” for each smoke group (copy/paste the results to your assignment). Note that these values are different from what you found in part C because of the additional variables included in the model.

ANSWER:

```
$emmeans
 smoke  prob     SE  df asymp.LCL asymp.UCL
 0     0.254 0.0467 Inf     0.174     0.356
 1     0.496 0.0710 Inf     0.360     0.632

Results are averaged over the levels of: race
Confidence level used: 0.95
Intervals are back-transformed from the logit scale

$contrasts
 contrast odds.ratio    SE  df z.ratio p.value
 0 / 1         0.346 0.131 Inf -2.802  0.0051

Results are averaged over the levels of: race
Tests are performed on the log odds ratio scale
```

G. Give the p-value corresponding to the Hosmer-Lemeshow test. Use hoslem.test() from the ResourceSelection package with g = 10 groups. Based on this test, is there evidence of lack of fit?

ANSWER:

Hosmer and Lemeshow goodness of fit (GOF) test

```
data:  Model2$y, fitted(Model2)
X-squared = 7.3472, df = 8, p-value = 0.4997
```