Homework #5

Dustin Leatherman

5/12/2019

10

$$\omega = \frac{\pi}{1-\pi} = e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}$$

Let $A = X_i$

$$\omega_A = e^{\beta_0 + \beta_1 A}$$

Let $B = X_i$

$$\omega_B = e^{\beta_0 + \beta_1 B}$$

$$\frac{\omega_A}{\omega_B} = \frac{e^{\beta_0 + \beta_1 A}}{e^{\beta_0 + \beta_1 B}} = e^{e^{\beta_0 + \beta_1 A} - (e^{\beta_0 + \beta_1 B})} = e^{\beta_1 (A-B)}$$

11

The data in Display 20.14 are the launch temperatures (degrees Fahrenheit) and an indicator of O-ring failures for 24 space shuttle launches prior to the space shuttle Challenger disaster of January 28, 1986. (See the description in Section 4.1.1 for more background information.)

(a) Fit the logistic regression of Failure (1 for failure) on Temperature. Report the estimated coefficients and their standard errors.

(b) Test whether the coefficient of Temperature is 0, using Wald's test. Report a one-sided p-value (the alternative hypothesis is that the coefficient is negative; odds of failure decrease with increasing temperature).

(c) Test whether the coefficient of Temperature is 0, using the drop-in-deviance test.

(d) Give a 95% confidence interval for the coefficient of Temperature.

(e) What is the estimated logit of failure probability at 31°F (the launch temperature on January 28, 1986)? What is the estimated probability of failure?

(f) Why must the answer to part (e) be treated cautiously? (Answer: It is a prediction outside the range of the available explanatory variable values.)

a

```
failure.lr <- glm(Failure ~ Temperature, family = "binomial", data = ex2011)
summary(failure.lr)
```

```
##
## Call:
## glm(formula = Failure ~ Temperature, family = "binomial", data = ex2011)
##
## Deviance Residuals:
##     Min       1Q    Median       3Q       Max
## -1.2125   -0.8253   -0.4706    0.5907    2.0512
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 10.87535     5.70291    1.907   0.0565 .
## Temperature -0.17132     0.08344   -2.053   0.0400 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.975  on 23  degrees of freedom
## Residual deviance: 23.030  on 22  degrees of freedom
## AIC: 27.03
##
## Number of Fisher Scoring iterations: 4
```

b

There is suggestive evidence that the temperature coefficient is non-zero (Wald Lower Tail T-Test. p-value = 0.02).

```
est <- summary(failure.lr)$coef[2,]
pnorm(est[3])
```

```
##    z value
## 0.02002412
```

c

There is convincing evidence that there is an association between O-ring failure and Temperature (Drop-in-Deviance Test. p-value = 0.0148)

```
failure.reduced <- glm(Failure ~ 1, family = "binomial", data = ex2011)
dind(failure.lr, failure.reduced)
```

```
## [1] 0.01476632
```

d

```
confint(failure.lr)[2,]
```

```
## Waiting for profiling to be done...
```

```
##      2.5 %     97.5 %
## -0.3779407 -0.0305637
```

e

The estimated logit of failure probablitiy is 1.7164. The estimated failure probability is 0.9962.

```
eta <- predict(failure.lr, newdata = data.frame(Temperature=c(31)))
est.logit <- log(eta)
est.prob <- exp(eta) / (1 + exp(eta))

est.logit
```

```
##        1
## 1.716392
```

```
est.prob
```

```
##        1
## 0.9961828
```

f

Since 31 is not within the range of temperature data provided, this result is an extrapolation and cannot be tested against without further data collection.

12

Duchenne Muscular Dystrophy (DMD) is a genetically transmitted disease, passed from a mother to her chil

(a) Make a scatterplot of H versus log(CK); use one plotting symbol to represent the controls on the pl

(b) Fit the logistic regression of carrier on CK and CK-squared. Does the CK-squared term significantl

(c) Fit the logistic regression of carrier on log(CK) and H. Report the coefficients and standard erro

(d) Carry out a drop-in-deviance test for the hypothesis that neither log(CK) nor H are useful predict

(e) Typical values of CK and H are 80 and 85. Suppose that a suspected carrier has values of 300 and 1

```r
qplot(H, log(CK), color = Group, data = ex2012)
```
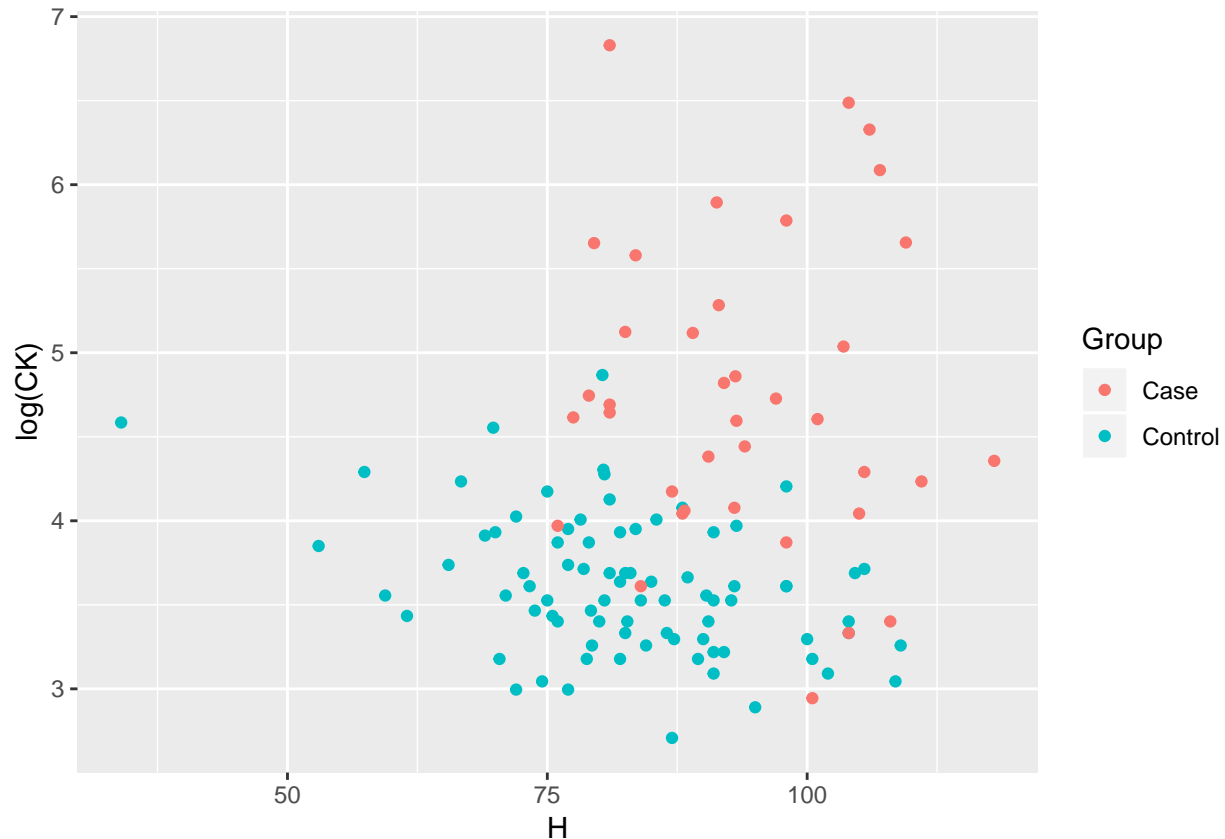

It appears that there is a decent amount of separation between Carriers and Controls so there appears to be usefulness for using these enzymes as predictors.

b

There is no evidence that the squared CK term is non-zero (Two-tailed Wald Test for single logistic regression coefficient. p-value = 0.122).

There is no evidence that the squared log(CK) term is non-zero (Two-tailed Wald Test for single logistic regression coefficient. p-value = 0.1737).

The log transformation is more appropriate for this data since it is right-skewed. A log transformation doesn't make it entirely normal but it does help.

```r
mdc.lr.ck <- glm(Group ~ poly(CK, 2), family = "binomial", data = ex2012)
summary(mdc.lr.ck)
```

```
##
## Call:
## glm(formula = Group ~ poly(CK, 2), family = "binomial", data = ex2012)
```
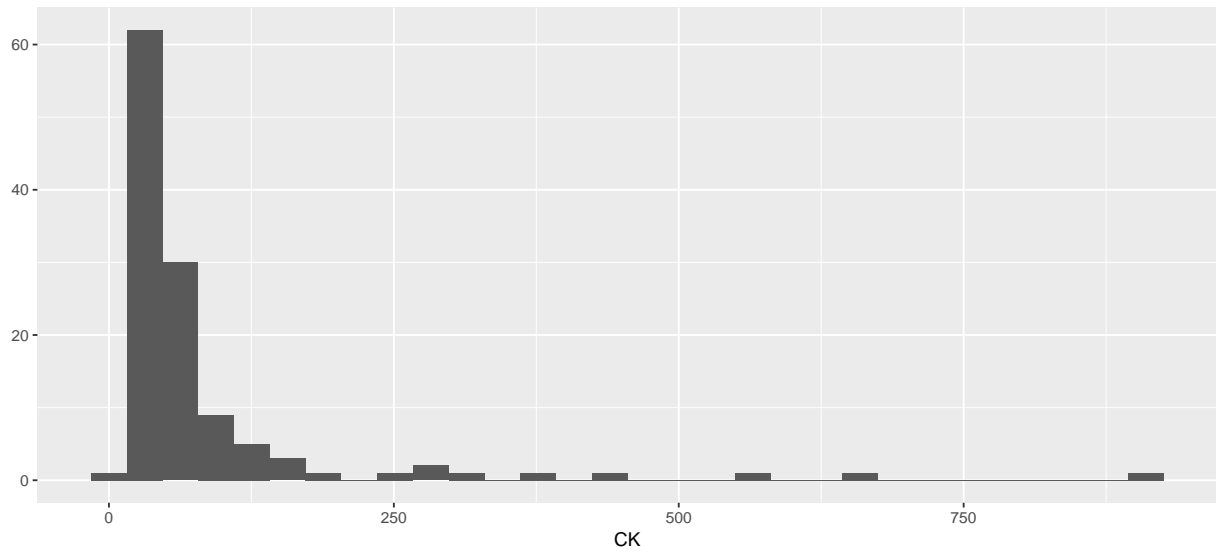
```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.50536  -0.03915   0.37969   0.51841   2.27337
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.4887     0.5342   0.915    0.360
## poly(CK, 2)1 -31.4849    23.4325  -1.344    0.179
## poly(CK, 2)2  19.0390    12.3095   1.547    0.122
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.84  on 119  degrees of freedom
## Residual deviance:  85.47  on 117  degrees of freedom
## AIC: 91.47
##
## Number of Fisher Scoring iterations: 9
```

```
mdc.lr.logck <- glm(Group ~ poly(log(CK), 2), family = "binomial", data = ex2012)
summary(mdc.lr.logck)
```

```
##
## Call:
## glm(formula = Group ~ poly(log(CK), 2), family = "binomial",
##     data = ex2012)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.39368  -0.03111   0.38041   0.50222   2.28558
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.3558     0.5363   0.663  0.50706
## poly(log(CK), 2)1 -39.5356    12.0472  -3.282  0.00103 **
## poly(log(CK), 2)2 -14.1441    10.3972  -1.360  0.17371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.840  on 119  degrees of freedom
## Residual deviance:  85.017  on 117  degrees of freedom
## AIC: 91.017
##
## Number of Fisher Scoring iterations: 7
```
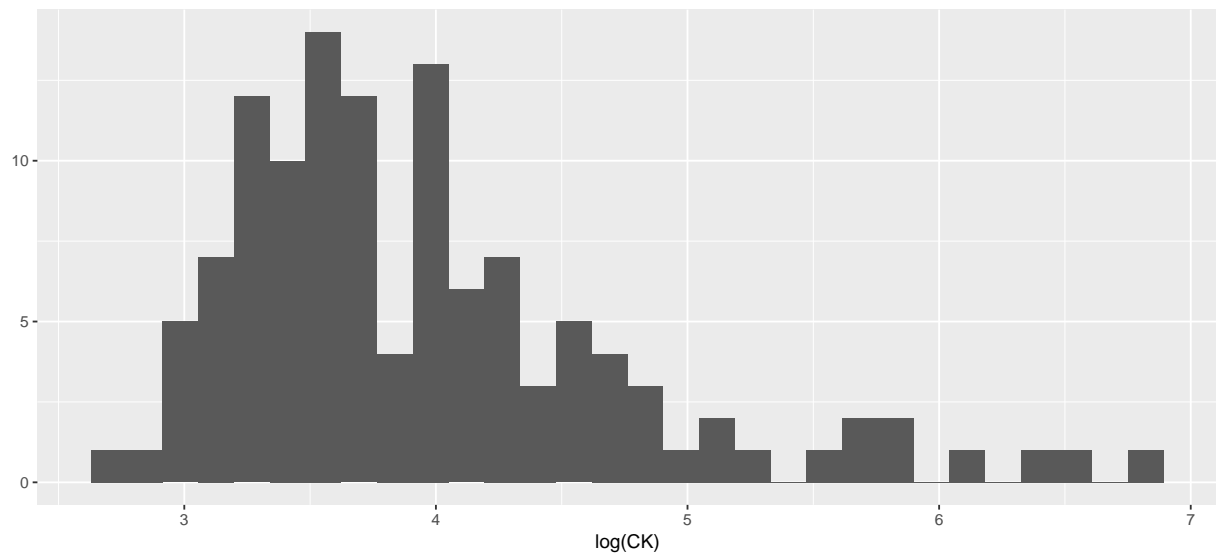
```
qplot(x = CK, geom="histogram", data = ex2012)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
qplot(x = log(CK), geom="histogram", data = ex2012)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


c

```
mdc.lr <- glm(Group ~ log(CK) + H, family = "binomial", data = ex2012)
summary(mdc.lr)
```

```
##
## Call:
## glm(formula = Group ~ log(CK) + H, family = "binomial", data = ex2012)
##
```

```
## Deviance Residuals:
##      Min       1Q    Median       3Q      Max
## -2.60371  -0.09903   0.16696   0.38782   1.89706
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 28.91340    5.80017   4.985 6.20e-07 ***
## log(CK)     -4.02043    0.82910  -4.849 1.24e-06 ***
## H           -0.13652    0.03654  -3.736 0.000187 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.840  on 119  degrees of freedom
## Residual deviance:  61.992  on 117  degrees of freedom
## AIC: 67.992
##
## Number of Fisher Scoring iterations: 7
```

d

There is convincing evidence that at least one of the predictors, log(CK) and H, are significant for determining whether a woman is a muscular dystrophy carrier (Drop-in-Deviance Test. p-value = 0).

```
mdc.reduced <- glm(Group ~ 1, family = "binomial", data = ex2012)
dind(mdc.lr, mdc.reduced)
```

```
## [1] 0
```

e

Typical values of CK and H are 80 and 85. Suppose that a suspected carrier has values of 300 and 100. What are the odds that she is a carrier relative to the odds that a woman with typical values (80 and 85) is a carrier?

A woman with CK = 300 and H = 100 is 177.8424 times more likely to be a carrier.

$$e^{\beta_0 + \beta_1(300-80) + \beta_2(100-85)} = e^{28.9134 + log(220)\beta_1 + 15\beta_2} = 177.8424$$

16

Hermon Bumpus analyzed various characteristics of some house sparrows that were found on the ground after a severe winter storm in 1898. Some of the sparrows survived and some perished. The data on male sparrows in Display 20.17 are survival status (1 = survived, 2 = perished), age (1 = adult, 2 = juvenile), the length from tip of beak to tip of tail (in mm), the alar extent (length from tip to tip of the extended wings, in mm), the weight in grams, the length of the head in mm, the length of the humerus (arm bone, in inches), the length of the femur (thigh bones, in inches), the length of the tibio-tarsus (leg bone, in inches), the breadth of the skull in inches, and the length of the sternum in inches. (A subset of this data was discussed in Exercise 2.21.)

Analyze the data to see whether the probability of survival is associated with physical characteristics of the birds. This would be consistent, according to Bumpus, with the theory of natural selection: those that survived did so because of some superior physical traits. Realize that (i) the sampling is from a population of grounded sparrows, and (ii) physical measurements and survival are both random. (Thus, either could be treated as the response variable.)

Null Hypothesis: Physical characteristics are not associated with the survival of grounded sparrows. Alternative Hypothesis: Physical characteristics are associated with the survival of grounded sparrows.

Comparing a logistic regression intercept-only model with a full logistic regression model using the drop-in-deviance test will tell if any of the physical characteristics are useful in predicting survival. There is convincing evidence that there is an association between physical characteristics and survival with grounded sparrows (Drop-in-Deviance Test. p-value = 1.099047e-07).

```
birds <- ex2016
# get a 0-based variable
birds$AGScale <- birds$AG - 1
birds.full <- glm(Status ~ ., family = "binomial", data = birds %>% select(-AG))
birds.reduced <- glm(Status ~ 1, family = "binomial", data = birds)
summary(birds.full)
```

```
##
## Call:
## glm(formula = Status ~ ., family = "binomial", data = birds %>%
##     select(-AG))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2252  -0.5232   0.1397   0.5131   2.0134
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 27.56606   26.42412   1.043 0.296848
## TL          -0.73634    0.18965  -3.883 0.000103 ***
## AE           0.08275    0.12622   0.656 0.512060
## WT          -0.88860    0.34182  -2.600 0.009333 **
## BH           0.58293    0.59735   0.976 0.329131
## HL          56.03494   31.05541   1.804 0.071176 .
## FL          -6.64680   31.73442  -0.209 0.834096
## TT           5.05213   14.05263   0.360 0.719210
## SK          21.53121   27.28482   0.789 0.430037
## KL          23.56111   12.03826   1.957 0.050326 .
## AGScale      0.10631    0.68253   0.156 0.876225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 118.01  on 86  degrees of freedom
## Residual deviance:  65.92  on 76  degrees of freedom
## AIC: 87.92
##
## Number of Fisher Scoring iterations: 6
```

```
dind(birds.full, birds.reduced)
```

```
## [1] 1.099047e-07
```

19

Following a successful test of an anti-ballistic missile (ABM) in July 1999, many prominent U.S. politicians called for the early deployment of a full ABM system. The scientific community was less enthusiastic about the efficacy of such a system. G. N. Lewis, T. A. Postol, and J. Pike ("Why National Missile Defense Won't Work," Scientific American, 281(2): 36–41, August 1999) traced the history of ABM testing, reporting the results shown in Display 20.20. Do these data suggest any improvement in ABM test success probability over time?

Null Hypothesis: Time is not associated the results of a missle test Alternative Hypothesis: Time is associated with the results of a missle test

As done in the previous problem, comparing an intercept-only logistic regression model with a logistic regression model containing time using the drop-in-deviance statistic is telling in the efficacy of missle testing.

There is no evidence that there is an association between Time and the Results of a Missle Test (Drop-in-Deviance. p-value = 0.6909)

```
missle.full <- glm(Result ~ Months, family = "binomial", data = ex2019)
missle.reduced <- glm(Result ~ 1, family = "binomial", data = ex2019)
summary(missle.full)
```

```
##
## Call:
## glm(formula = Result ~ Months, family = "binomial", data = ex2019)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7478  -0.6215  -0.5797  -0.5507   2.0013
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.131263   1.161775  -0.974    0.330
## Months      -0.003706   0.009231  -0.402    0.688
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.844  on 16  degrees of freedom
## Residual deviance: 15.686  on 15  degrees of freedom
## AIC: 19.686
##
## Number of Fisher Scoring iterations: 4
```

```
summary(missle.reduced)
```

```
##
## Call:
```

```
## glm(formula = Result ~ 1, family = "binomial", data = ex2019)
##
## Deviance Residuals:
##     Min        1Q    Median        3Q       Max
## -0.6231   -0.6231   -0.6231   -0.6231    1.8626
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5404     0.6362  -2.421   0.0155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.844  on 16  degrees of freedom
## Residual deviance: 15.844  on 16  degrees of freedom
## AIC: 17.844
##
## Number of Fisher Scoring iterations: 4
```

```
dind(missle.full, missle.reduced)
```

```
## [1] 0.6909111
```

20

Reconsider the NLSY79 data in Exercise 12.23. The variable Esteem1 is the respondent's degree of agreement in 2006 with the statement "I feel that I'm a person of worth, at least on equal basis with others," with values of 1, 2, 3, and 4 signifying strong agreement, agreement, disagreement, or strong disagreement, respectively. Construct a new binary response variable from this, which takes on the value 1 for strong agreement and 0 for agreement, disagreement, or strong disagreement.

Explore the dependence of the probability that a person has positive self-esteem (as indicated by a response of strong agreement on Esteem1) on these explanatory variables: log annual income in 2005 (Income2005), intelligence as measured by the AFQT score in 1981 (AFQT), years of education (Educ), and gender (Gender).

There is convincing evidence that log income, AFQT score, and Years of Education are associated with person with positive self esteem (Two-tailed Wald Test for single logistic regression coefficient. p-value = 0.0004, 0.0000365, 0.000297 respectively). There is suggestive but weak evidence that being Male is associated with positive self-esteem.

```
esteem <- ex1223
esteem$high <- ifelse(esteem$Esteem1 == 1, 1, 0)
esteem.full <- glm(high ~ log(Income2005) + AFQT + Educ + Gender, family = "binomial", data = esteem)
summary(esteem.full)
```

```
##
## Call:
## glm(formula = high ~ log(Income2005) + AFQT + Educ + Gender,
##     family = "binomial", data = esteem)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7891  -1.2240   0.8174   1.0410   1.6339
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.773457   0.485833  -5.709 1.14e-08 ***
## log(Income2005)  0.165970   0.047114   3.523 0.000427 ***
## AFQT             0.007603   0.001842   4.128 3.65e-05 ***
## Educ             0.076135   0.021045   3.618 0.000297 ***
## Gendermale      -0.146214   0.087097  -1.679 0.093201 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3508.3  on 2583  degrees of freedom
## Residual deviance: 3395.5  on 2579  degrees of freedom
## AIC: 3405.5
##
## Number of Fisher Scoring iterations: 4
```