Data Simulation

Xiangyi Xu

Thursday, April 09, 2015

Data simulation is based on article: Pastor-Valero, M. (2013). Fruit and vegetable intake and vitamins C and E are associated with a reduced prevalence of cataract in a Spanish Mediterranean population. BMC Ophthalmology, 13:52.

Background Cataract is among the major causes of vision impairment and blindness worldwide. Epidemiological studies support the role of antioxidants in the etioloty of cataract, but the evidence of one specific antioxidant over another is inconsistent. Few studies have examined the association of cataract with fruit and vegetable intake with inconclusive results. In this study, the relationship between cataract and fruit and vegetable intake and dietary and blood levels of carotenoids, vitamins C and E were examined in a Spanish Mediterranean population.

Data Simulation The dataset is simulated from the above study. Among 433 elderly with cataract or cataract extraction, 54% are women and 46% are men, their average age iss around 72. The study shows that increasing quartiles `(13~83, 83~107, 107~143, 143~408 mg/d)` of dietary intakes from `107mg/d` of vitamin C indicating a significant decreasing association with prevalence of cataract or cataract extraction. In the simulated dataset, with enough data, it will show that both age and intake of vitamin C matter much in having cataract or not. Noise such as zodiac signs are added in the study too. Hint: log of VitaminC for the analysis of data

Features Feature1: age
Feature2: obesity- Y, N
Feature3: sex- M, F
Feature4: marital_status- Y, N
Feature5: diabetes- Y, N
Feature6: smoking- Never, Past, Current
Feature7: VitaminC
Feature8: zodiac signs
Outcome: cataract Y, N

Manipulate function is used to generate coefficients. library(manipulate) manipulate(with(df,{ score <- 0.73 + 2^a*(age - 72) - 2^b*log(VitaminC) prob <- logistic(score) hist(prob, breaks=50) }), a=slider(-9, 9, step=0.1, initial = 0), b=slider(-9, 9, step=0.1, initial = 0))

```
generate_dataset <- function(N=100){

        age <- runif(N, min=72-10, max=72+10)

        obesity <- sample(c("Y", "N"), N, replace=TRUE, prob=c(.36, .64))

        sex <- sample(c("M", "F"), N, replace=TRUE, prob=c(.46, .54))

        marital_status <- sample(c("Y", "N"), N, replace=TRUE,
```

```
                                prob=c(.70, .30))

        diabetes <- sample(c("Y", "N"), N, replace=TRUE, prob=c(.26, .74))

        smoking <- sample(c("Never", "Past", "Current"), N, replace=TRUE,
                          prob=c(.52, .31, .17))

        VitaminC <- rnorm(N, mean=107, sd=15)



        zodiac <- c("Aries", "Taurus", "Gemini", "Cancer", "Leo", "Virgo",
                    "Libra","Scorpio", "Sagittarius", "Capricorn","Aquarius",
                      "Pisces")

        sign <- sample(zodiac, N, replace=TRUE)

        simulate_cataract <- function(age, VitaminC){
                logistic <- function(t) 1 / (1 + exp(-t))
                score <- 0.73 +  2^(-1.2)*(age - 72) - 2^(-1.2)*log(VitaminC)

                prob <- logistic(score)

                result <- ifelse(runif(length(prob)) < prob, "Y", "N")
        }

        cataract <- simulate_cataract(age, VitaminC)

        data.frame(age, sex, obesity, marital_status, smoking, VitaminC,
                         sign, cataract)


}

df <- generate_dataset(2e5)

head(df)
```

generate dataset

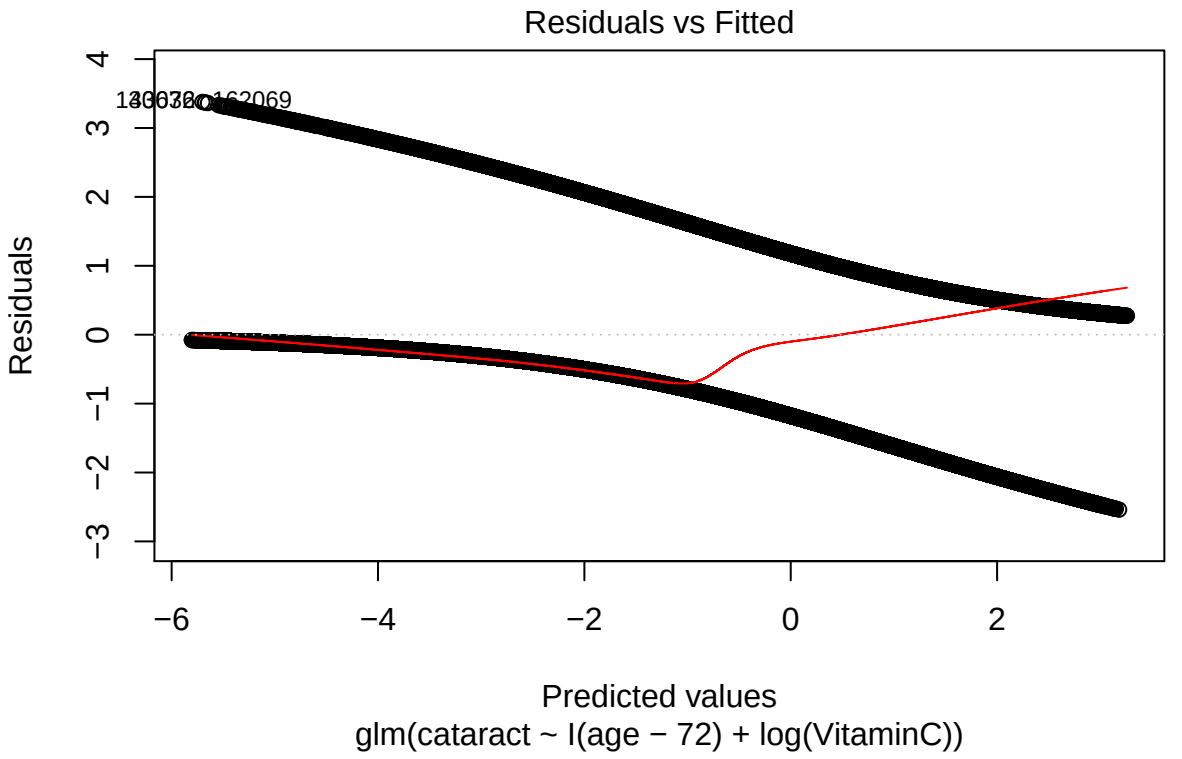
```
##        age sex obesity marital_status smoking  VitaminC     sign cataract
## 1 63.20792   F       N              Y    Past 106.16689 Aquarius        N
## 2 75.27488   F       N              Y Current 100.28545    Virgo        Y
## 3 63.84553   F       N              N    Past  88.76279   Gemini        N
## 4 64.77161   F       Y              Y   Never 113.11708      Leo        N
## 5 78.87637   M       N              Y    Past 114.32050   Taurus        Y
## 6 78.70342   F       N              N   Never 119.27459    Libra        Y
```

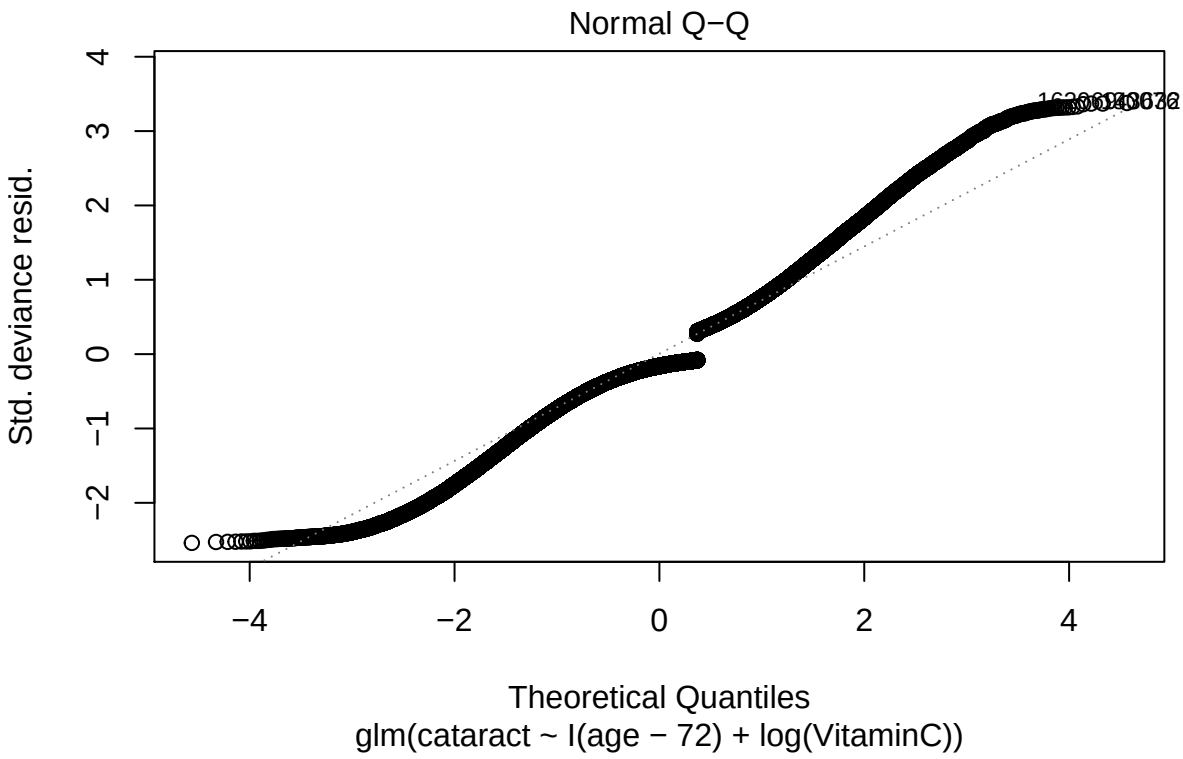
```
fit <- glm(cataract ~ I(age-72) + log(VitaminC), data=df, family="binomial")
summary(fit)
```

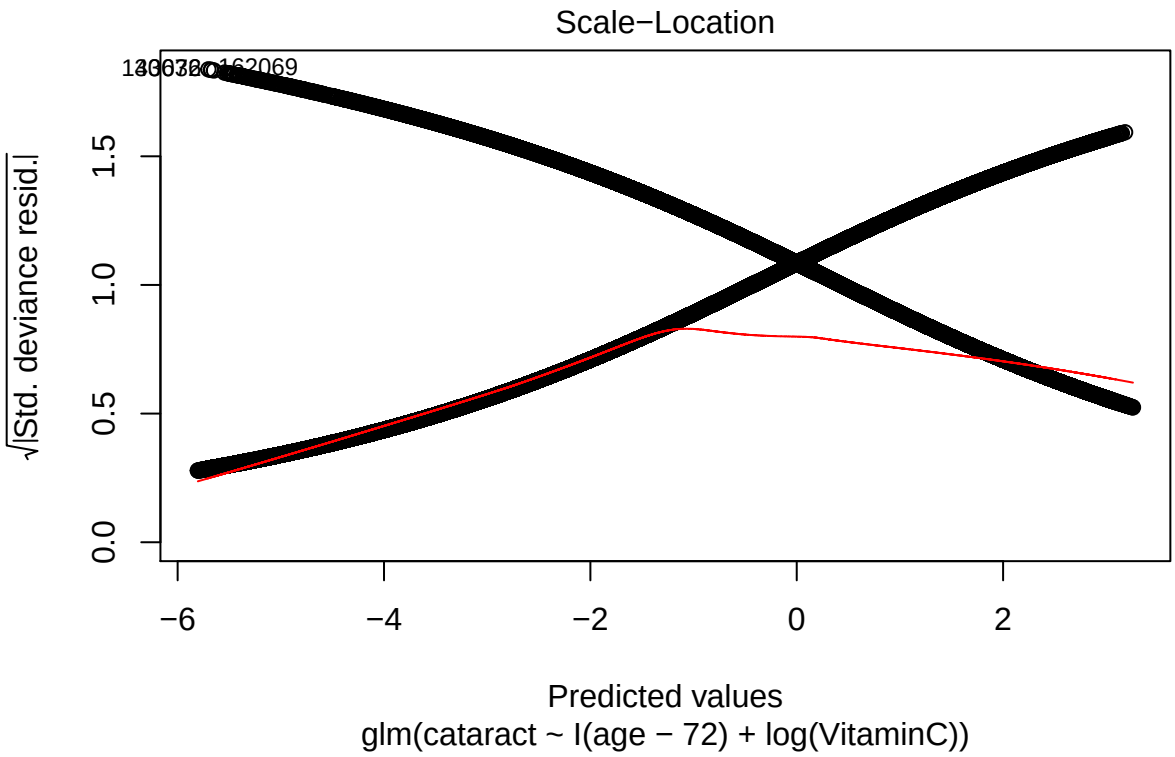
```
##
## Call:
```

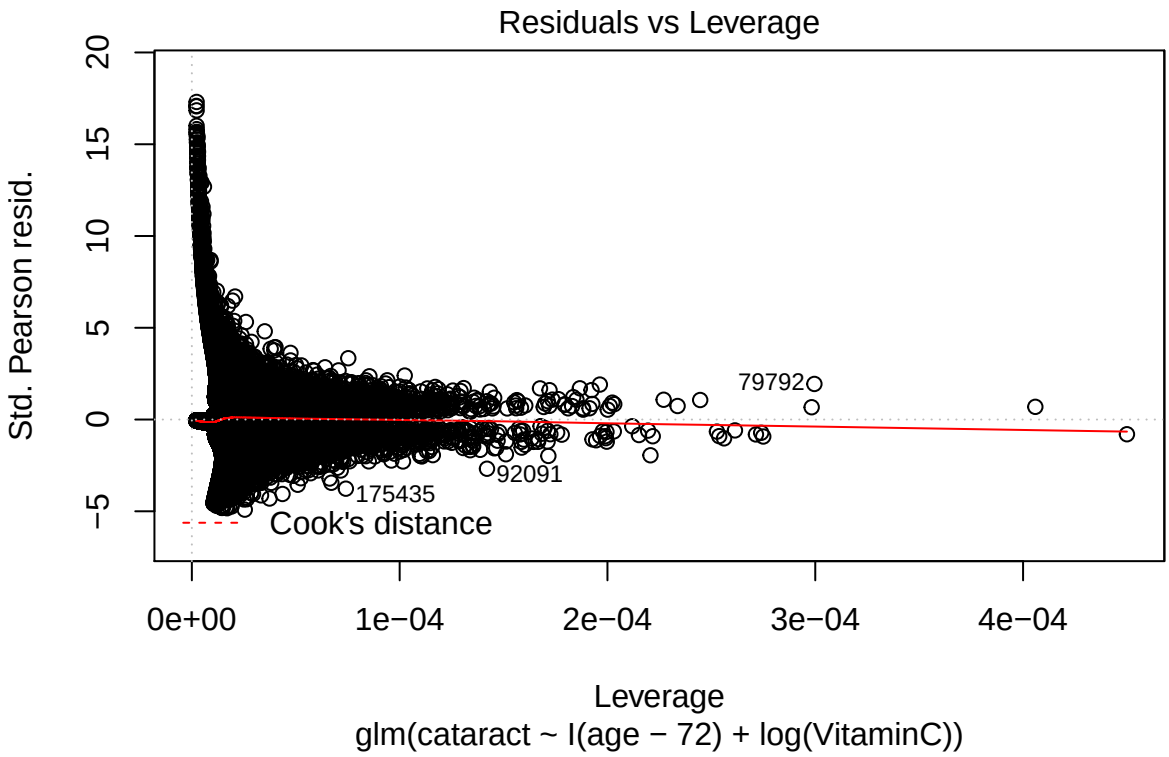
```
## glm(formula = cataract ~ I(age - 72) + log(VitaminC), family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5388  -0.4805  -0.1592   0.4921   3.3779
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.04004    0.21895    4.75 2.03e-06 ***
## I(age - 72)    0.43345    0.00188  230.56  < 2e-16 ***
## log(VitaminC) -0.50185    0.04696  -10.69  < 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: 260466  on 199999  degrees of freedom
## Residual deviance: 139628  on 199997  degrees of freedom
## AIC: 139634
##
## Number of Fisher Scoring iterations: 6
```

```
plot(fit)
```


Residuals vs Fitted
Residuals
Predicted values
glm(cataract ~ I(age − 72) + log(VitaminC))


Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(cataract ~ I(age − 72) + log(VitaminC))


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
−6
−4
−2
0
2
Predicted values
glm(cataract ~ I(age − 72) + log(VitaminC))




```
fit2 <- glm(cataract ~ age + log(VitaminC) + obesity + marital_status, data=df, family="binomial")
summary(fit2)
```

```
##
## Call:
## glm(formula = cataract ~ age + log(VitaminC) + obesity + marital_status,
##     family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5442  -0.4808  -0.1591   0.4922   3.3752
##
## Coefficients:
##                   Estimate Std. Error  z value Pr(>|z|)
## (Intercept)     -30.159665   0.255677 -117.960   <2e-16 ***
## age               0.433455   0.001880  230.559   <2e-16 ***
## log(VitaminC)    -0.501854   0.046964  -10.686   <2e-16 ***
## obesityY          0.005115   0.014075    0.363    0.716
## marital_statusY  -0.015597   0.014727   -1.059    0.290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 260466  on 199999  degrees of freedom
```

```
## Residual deviance: 139627  on 199995  degrees of freedom
## AIC: 139637
##
## Number of Fisher Scoring iterations: 6
```