VE406 Homework5

Yu Xinmiao 518021910792

11/24/2020

Q1

(a)

After load in the semi.df, we observe the `str(semi.df)` do not have large data values, so randomly choose three small coefficient as the initial starting point for nls.

```
semi.df = read.table(file = semi.txt, header = TRUE)
# str(semi.df)
parameter_each = capture.output({
  semi.NL = nls(
    Reactivity~b1*(1-exp(-exp(b2+b3*Thickness))),
    start = list(b1=1, b2=1, b3=2), data = semi.df, trace = TRUE
    )
})

parameter_each
```

```
##  [1] "2.732035 :  1 1 2"
##  [2] "2.448122 :   0.8145931   6.2907163 -31.5691395"
##  [3] "2.087135 :   0.8540583   3.5183279 -17.8746785"
##  [4] "1.830549 :   0.9203079   1.9259211 -10.3513852"
##  [5] "1.669153 :  1.047910  1.304834 -8.497452"
##  [6] "1.517795 :  1.2387815  0.8973009 -7.9415093"
##  [7] "1.484195 :  1.7109713  0.3397431 -7.9448983"
##  [8] "0.7722244 :   2.4644855   0.1831752 -10.0357352"
##  [9] "0.638288 :   1.9932920   0.6200694 -11.1894357"
## [10] "0.2758684 :   1.9018436   0.9892926 -12.7388601"
## [11] "0.1367902 :   1.981707   1.219382 -14.124957"
## [12] "0.1366601 :   1.949793   1.265391 -14.335590"
## [13] "0.1366565 :   1.949105   1.268657 -14.357340"
## [14] "0.1366565 :   1.948384   1.269827 -14.362532"
## [15] "0.1366565 :   1.948360   1.269891 -14.362957"
## [16] "0.1366565 :   1.948346   1.269914 -14.363057"
```

From the output, we found the coefficients approaches around $b_1 = 2, b_2 = 1, b_3 = -14$. So with these three initial values, we fit the model again.

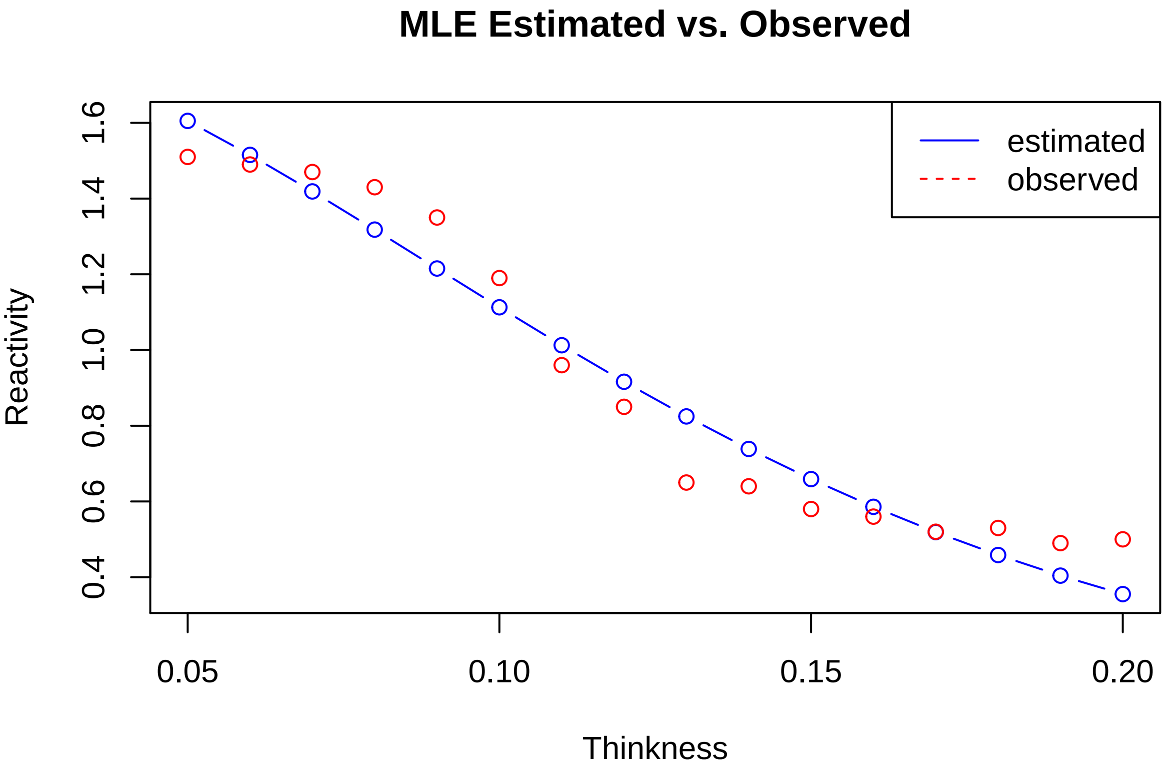
```
semi.NL = nls(
  Reactivity~b1*(1-exp(-exp(b2+b3*Thickness))),
  start = list(b1=2, b2=1, b3=-14), data = semi.df)
summary(semi.NL)
```

```
##
## Formula: Reactivity ~ b1 * (1 - exp(-exp(b2 + b3 * Thickness)))
##
## Parameters:
##    Estimate Std. Error t value Pr(>|t|)
## b1   1.9484     0.4725   4.124 0.001199 **
## b2   1.2699     0.6809   1.865 0.084902 .
## b3 -14.3630     2.7038  -5.312 0.000141 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1025 on 13 degrees of freedom
##
## Number of iterations to convergence: 6
## Achieved convergence tolerance: 8.139e-06
```

And the maximum estimates of the parameters $\beta_1, \beta_2, \beta_3$ are shown in the summary.

(b)

```
fvs = fitted.values(semi.NL)
obs = semi.df$Reactivity
plot(semi.df$Thickness, fvs, type="b", col="blue", lty=1, xlab="Thinkness", ylab="Reactivity", main="ML
points(semi.df$Thickness, obs, col="red", lty=2)
legend("topright", lty = 1:2, col = c("blue", "red"), legend = c("estimated", "observed"))
```



(c)

The estimated $\hat{\sigma}^2$ is

$$\frac{1}{n-3}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2,$$

with the data introduced before, calculated that

```
sum((obs - fvs) * (obs - fvs)) / (length(obs) - 3)
```

```
## [1] 0.01051204
```

So $\hat{\sigma}^2 = 0.0105$.

(d)

When β_1, β_3 are known to be positive, as the trend of $e^{\beta_2+\beta_3 x_i}$ will not be influenced by β_2, the whole equation will decrease while x is increasing. So the upper bound is

$$\beta_1(1 - exp(-e^{-\beta_2}))$$

Estimate through the previous estimate, which gives result as 0.4771.

```
1.9484*(1-exp(-exp(-1.2699)))
```

```
## [1] 0.4770964
```

Q2

(a)

```
newcar.df = read.table(file = newcar.csv, header = TRUE, sep = ",")
newcar.LG = glm(NEWCAR ~ INCOME + CAR.AGE, family = binomial, data = newcar.df)
summary(newcar.LG)
```

```
##
## Call:
## glm(formula = NEWCAR ~ INCOME + CAR.AGE, family = binomial, data = newcar.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6189  -0.8949  -0.5880   0.9653   2.0846
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.73931    2.10195  -2.255   0.0242 *
## INCOME       0.06773    0.02806   2.414   0.0158 *
## CAR.AGE      0.59863    0.39007   1.535   0.1249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 44.987  on 32  degrees of freedom
## Residual deviance: 36.690  on 30  degrees of freedom
## AIC: 42.69
```

```
##
## Number of Fisher Scoring iterations: 4
```

The result of Wald Test is shown in the summary. Base on the p-value of testing the null hypothesis H_0 : the coefficients of `CAR.AGE` is zero, which is $0.1249 > 0.05$, there is no evidence to reject the hypothesis. So we can conclude that `CAR.AGE` could be dropped from the model.

(b)

```
newcar.noCarAge.LG = glm(NEWCAR ~ INCOME, family = binomial, data = newcar.df)
LR.test = 2 * (logLik(newcar.LG)[1]-logLik(newcar.noCarAge.LG)[1])
1-pchisq(LR.test, 1)
```

```
## [1] 0.1058638
```

The likelihood ratio test with null hypothesis H_0 : the reduced model is valid, such that the coefficients of `CAR.AGE` forced to zero is indeed zero. The p-value is 0.106, which is not small enough. There is no evidence to reject the hypothesis. So `CAR.AGE` is not needed. ## (c)

```
predf = data.frame(INCOME = 50, CAR.AGE = 3)
mhat = predict(newcar.LG, newdata = predf, type = "response")
mhat
```

```
##         1
## 0.6090245
```

which gives the estimated probability is 0.609.

Q3

(a)

```
leukemia.df = data.frame(
  County = c("Marin", "Contra Costa", "Alameda", "San Francisco", "San Mateo" ),
  Count = c(22,146,226,47,52),
  Population = c(247289,948816,1443741,776733,70716)
  )
leukemia.PS = glm(Count ~ Population, family = poisson, data = leukemia.df)
summary(leukemia.PS)
```

```
##
## Call:
## glm(formula = Count ~ Population, family = poisson, data = leukemia.df)
##
## Deviance Residuals:
##       1        2        3        4        5
## -3.0650   3.2450  -0.0446  -4.5831   3.5002
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.320e+00  1.141e-01   29.10   <2e-16 ***
## Population  1.457e-06  1.013e-07   14.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 287.348  on 4  degrees of freedom
## Residual deviance:  53.182  on 3  degrees of freedom
## AIC: 87.685
##
## Number of Fisher Scoring iterations: 4
```

So the maximum likelihood estimate of θ is $1.457e-6$ ## (b)

```
leukemia.county = glm(Count ~ Population + County, family = poisson, data = leukemia.df)
1-pchisq( leukemia.PS$deviance - leukemia.county$deviance , 1)
```

```
## [1] 3.039791e-13
```

As the large sample inference is valid, as the sample size becomes large, the difference in the deviances follows a chi-squared distribution, with the null hypothesis that the simpler model is correctly specified.

The number of degree of freedom is the difference in the number of paramters of the two model, which is 1.

The p-value is $3.039791e-13$, so we need to introduce the categorical variable `County` into the model.