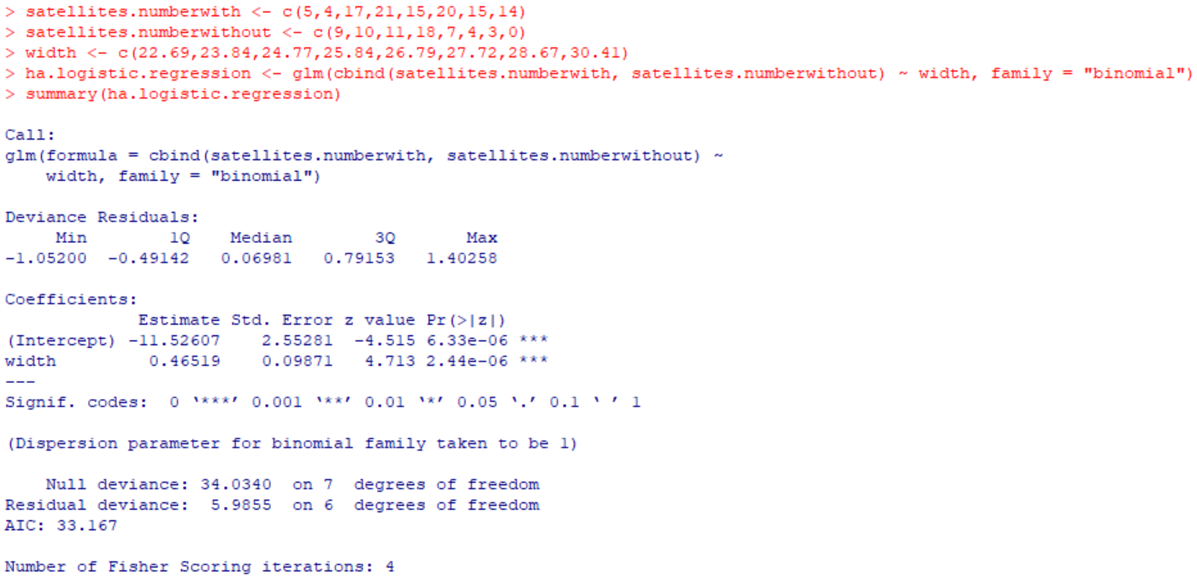
limaida_300536130_assignment4

1a. (and 1b.)



```
> satellites.numberwith <- c(5,4,17,21,15,20,15,14)
> satellites.numberwithout <- c(9,10,11,18,7,4,3,0)
> width <- c(22.69,23.84,24.77,25.84,26.79,27.72,28.67,30.41)
> ha.logistic.regression <- glm(cbind(satellites.numberwith, satellites.numberwithout) ~ width, family = "binomial")
> summary(ha.logistic.regression)

Call:
glm(formula = cbind(satellites.numberwith, satellites.numberwithout) ~
    width, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.05200  -0.49142   0.06981   0.79153   1.40258

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.52607    2.55281  -4.515 6.33e-06 ***
width         0.46519    0.09871   4.713 2.44e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.0340  on 7  degrees of freedom
Residual deviance:  5.9855  on 6  degrees of freedom
AIC: 33.167

Number of Fisher Scoring iterations: 4
```

$\hat{\beta}_0$ = -11.5261

$\hat{\beta}_1$ = 0.4652

log(p(X)/(1-p(X))) ≈ -11.5261 + 0.4652X   or, equivalently,

$\hat{p} \approx \frac{\exp(-11.5261+0.4652X)}{1+\exp(-11.5261+0.4652X)}$,

where X denotes the width.

Need a graph here?

c.

The wald test statistic is given by

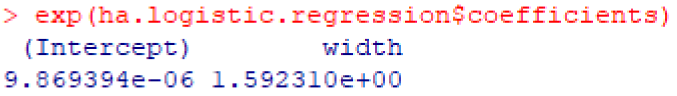
$z^* = \frac{\hat{\beta}1}{SE(\hat{\beta}1)} \approx \frac{0.4652}{0.0987} \approx 4.7$

And the p-value is given by

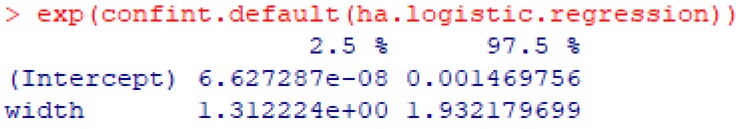
p-value = 2 x P(Z > |z*|) ≈ 2 x P(Z > 4.7) ≈ less than any reasonable significance level.

As the p-value is well below a = 0.05 (or any reasonable significance level), we reject H0 at the a = 0.05 significance level. Thus, there is strong evidence that satellites have association with width.

d.



```
> exp(ha.logistic.regression$coefficients)
 (Intercept)        width
9.869394e-06 1.592310e+00
```



```
> exp(confint.default(ha.logistic.regression))
                   2.5 %      97.5 %
(Intercept) 6.627287e-08 0.001469756
width       1.312224e+00 1.932179699
```

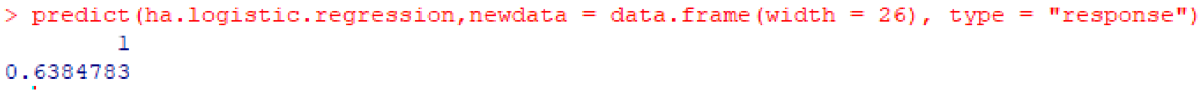
The association between width and number of satellites can be interpreted through exp(β1). We estimate that an increase in the number of 1 is associated with a multiplicative change of 1.592 (1.312, 1.932) in the odds of the incidence satellites (*i.e.*, each additional cm in width is estimated to increase the odds of a satellite by 59.2%).

To obtain the estimated odds ratio, we exponentiate the model coefficients (contained in the coefficients variable stored in a glm object) and confidence intervals for model coefficients (which can be obtained using the confint.default function). The code below shows to obtain the estimate odds ratio and corresponding confidence interval.

e.

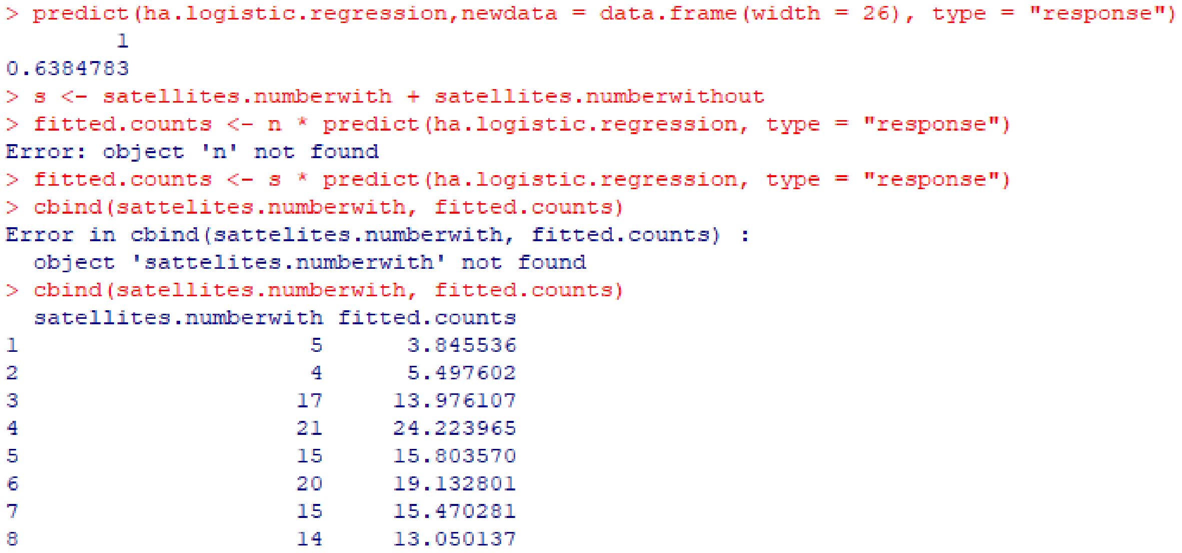
$$\hat{p}(26) \approx \frac{\exp\,(-11.5261+0.4652*26)}{1+\,\exp\,(-11.5261+0.4652*26)} \approx 0.6385$$

Thus, the predicted probability of having a satellite when crab has a width of 26cm is approx. 0.6383. To obtain this predicted probability in R, we use the predict function and pass the desired value for cigarettes in a data frame to the newdata argument, as shown in the following code.



```
> predict(ha.logistic.regression,newdata = data.frame(width = 26), type = "response")
        1
0.6384783
```

f.



```
> predict(ha.logistic.regression,newdata = data.frame(width = 26), type = "response")
        1
0.6384783
> s <- satellites.numberwith + satellites.numberwithout
> fitted.counts <- n * predict(ha.logistic.regression, type = "response")
Error: object 'n' not found
> fitted.counts <- s * predict(ha.logistic.regression, type = "response")
> cbind(sattelites.numberwith, fitted.counts)
Error in cbind(sattelites.numberwith, fitted.counts) :
  object 'sattelites.numberwith' not found
> cbind(satellites.numberwith, fitted.counts)
  satellites.numberwith fitted.counts
1                     5      3.845536
2                     4      5.497602
3                    17     13.976107
4                    21     24.223965
5                    15     15.803570
6                    20     19.132801
7                    15     15.470281
8                    14     13.050137
```

g.

We note that the observed counts and fitted counts produced in part f are quite close, suggesting that the model fits the observed data quite well. Here we formally test that, using a goodness-of-fit test. We test the hypothesis

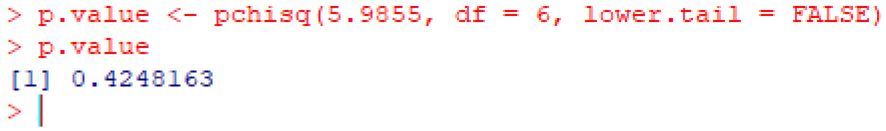
H_0: Model M provides a good fit to the data.

H_1: Model M does not provide a good fit to the data.

Where the model M denotes the logistic regression model that we fit.

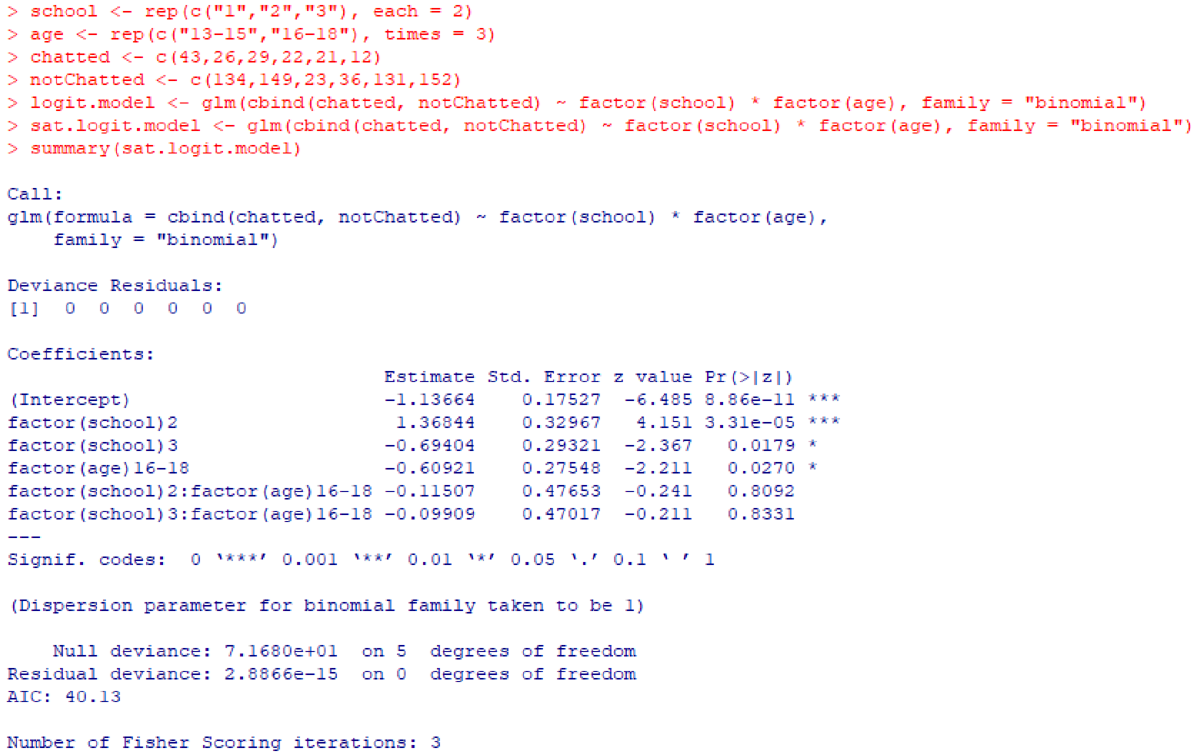
Under the null hypothesis, the deviance is given by $G_2(M)\approx 5.9855$

which has an approximately a X^{6}_{2} distribution. P-value ≈ $P(X^{6}_{2} > 5.9855) \approx 0.42$



```
> p.value <- pchisq(5.9855, df = 6, lower.tail = FALSE)
> p.value
[1] 0.4248163
> 
```

2a.



```
> school <- rep(c("1","2","3"), each = 2)
> age <- rep(c("13-15","16-18"), times = 3)
> chatted <- c(43,26,29,22,21,12)
> notChatted <- c(134,149,23,36,131,152)
> logit.model <- glm(cbind(chatted, notChatted) ~ factor(school) * factor(age), family = "binomial")
> sat.logit.model <- glm(cbind(chatted, notChatted) ~ factor(school) * factor(age), family = "binomial")
> summary(sat.logit.model)

Call:
glm(formula = cbind(chatted, notChatted) ~ factor(school) * factor(age), 
    family = "binomial")

Deviance Residuals: 
[1]  0  0  0  0  0  0

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -1.13664    0.17527  -6.485 8.86e-11 ***
factor(school)2                   1.36844    0.32967   4.151 3.31e-05 ***
factor(school)3                  -0.69404    0.29321  -2.367   0.0179 *  
factor(age)16-18                 -0.60921    0.27548  -2.211   0.0270 *  
factor(school)2:factor(age)16-18 -0.11507    0.47653  -0.241   0.8092    
factor(school)3:factor(age)16-18 -0.09909    0.47017  -0.211   0.8331    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7.1680e+01  on 5  degrees of freedom
Residual deviance: 2.8866e-15  on 0  degrees of freedom
AIC: 40.13

Number of Fisher Scoring iterations: 3
```

b.

There are 6 Logits (corresponding to the 3x2 possible school-age combinations). The model contains non-redundant parameters, so Residual df = (no. of logits)-(no. of redundant parameters) = 6-6=0.

Since the residual degrees of freedom is 0, the model is a saturated model.

c.

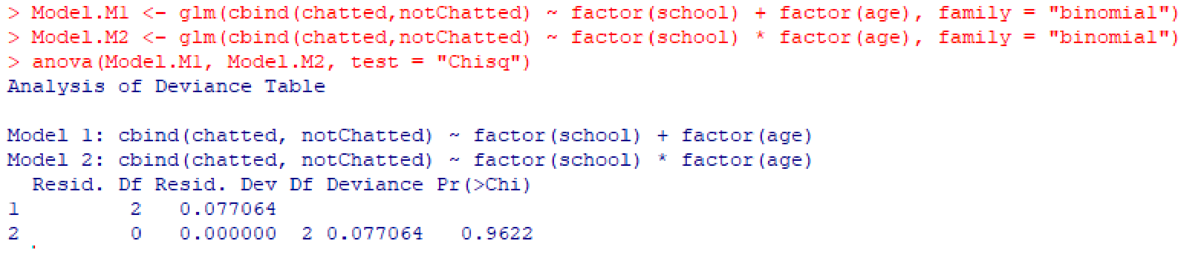
This is equivalent to a test of

$H_0 : \beta_{ij}^{AB} = 0$ for all i and j.

$H_1 : \beta_{ij}^{AB}$!= 0 for some i and j.

For the saturated model from part (a).

The following R code fits both the original saturated model (given by Model.M2) and this reduced model (given by Model.M1) and carries out a model comparison test:



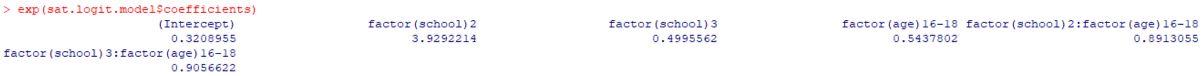
```
> Model.M1 <- glm(cbind(chatted,notChatted) ~ factor(school) + factor(age), family = "binomial")
> Model.M2 <- glm(cbind(chatted,notChatted) ~ factor(school) * factor(age), family = "binomial")
> anova(Model.M1, Model.M2, test = "Chisq")
Analysis of Deviance Table

Model 1: cbind(chatted, notChatted) ~ factor(school) + factor(age)
Model 2: cbind(chatted, notChatted) ~ factor(school) * factor(age)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2   0.077064                     
2         0   0.000000  2 0.077064   0.9622
```

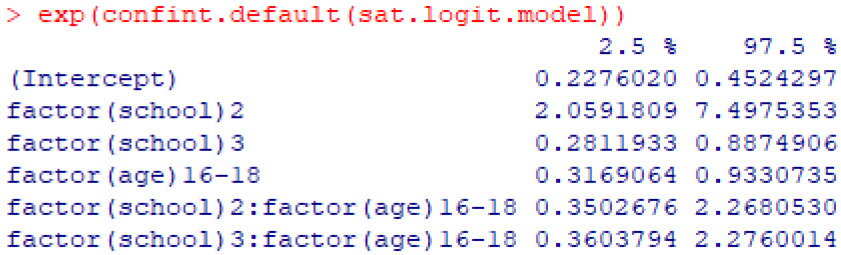
This produces a test stastic of $G^2 = G^2(M_1) – G^2(M_2) \approx 0.7706 – 0 = 0.7706$

As the p-value exceeds a significance level of a = 0.05, we have insufficient evidence to reject the null hypothesis, meaning that the interaction terms in the saturated model can be deleted, leading to the form of the reduced model that excludes these interactions.

d.



```
> exp(sat.logit.model$coefficients)
                   (Intercept)                factor(school)2                factor(school)3               factor(age)16-18 factor(school)2:factor(age)16-18
                     0.3208955                      3.9292214                      0.4995562                      0.5437802                        0.8913055
factor(school)3:factor(age)16-18
                     0.9056622
```



```
> exp(confint.default(sat.logit.model))
                                     2.5 %    97.5 %
(Intercept)                      0.2276020 0.4524297
factor(school)2                  2.0591809 7.4975353
factor(school)3                  0.2811933 0.8874906
factor(age)16-18                 0.3169064 0.9330735
factor(school)2:factor(age)16-18 0.3502676 2.2680530
factor(school)3:factor(age)16-18 0.3603794 2.2760014
```

We estimate that the odds of students from school 2 chatting to strangers is 3.9292

We estimate that the odds of students from school 3 chatting to strangers is 0.4996

We estimate that the odds of students from age 16-18 chatting to strangers is 0.5438

We estimate that the odds of students from school 2 and age 16-18 chatting to strangers is 0.8913

We estimate that the odds of students from school 3 and age 16-18 chatting to strangers is 0.90566

As the confidence interval is only completely above 1 for students from school 2, this is where it is significantly more likely to happen at a 0.05 significance level.