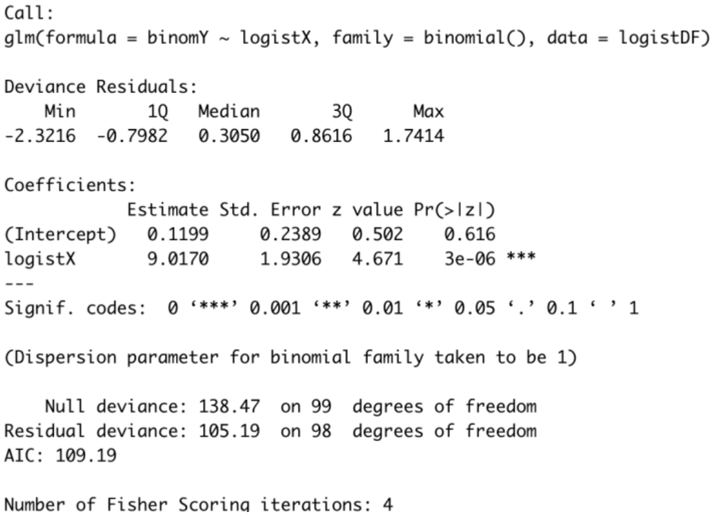
IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Wanyue Xiao
Date: September 27, 2020
Chapter Number: #10
Title of Chapter: **Logistic Regression**

Logistics Regression

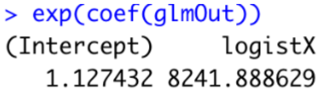
Maximum Likelihood Estimation: a method that used to determine values for the parameters of a model so that the parameter values can maximize the likelihood that the process described by the model. With maximum likelihood estimation, only one estimate will be produced for each population parameter while no posterior distribution will be produced.
Inverse Logic: (aka Logistic Curve) a function that uses natural logarithm to generate a S-shaped curve so that the model can predict binary result.

- The residual is always 0. Therefore, a very strongly positive or negative median indicates the existence of outliers.
- The **Log-odds**, which is also called the logarithm of the odds of the Y variable, is being represented by the coefficient of the linear regression.
  Here is the example of the glm() function’s result:



```
Call:
glm(formula = binomY ~ logistX, family = binomial(), data = logistDF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3216  -0.7982   0.3050   0.8616   1.7414

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1199     0.2389   0.502    0.616
logistX       9.0170     1.9306   4.671    3e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.47  on 99  degrees of freedom
Residual deviance: 105.19  on 98  degrees of freedom
AIC: 109.19

Number of Fisher Scoring iterations: 4
```

  Transform the coefficient by using exp() function.



```
> exp(coef(glmOut))
(Intercept)     logistX
   1.127432 8241.888629
```

- **Interpretation:** assumed that the transformed coefficient of X is 8241.8. One-unit change in X will have an 8241.8:1 change in the odds of Y. Namely, if the X variable increase by 1, it is 8241.8 times more likely that the Y result will be classified as the other class.
- One can use ***exp(confint(glmOut))*** to check the confidence interval of the coefficient of X variable. If the interval straddles 1:1, which is the null hypothesis, one can conclude that the coefficient is not significant. Otherwise, one can conclude that 95% confidence interval for the X variable runs from a low of 254.5:1 up to a. high of 525128.2:1. There is an amount of uncertainty around the 8241.8:1.
- Specifically, the **Null Deviance** indicates the number of errors in the model. This number is perceived as a baseline that peopled can used to compare other models. 1 df is given to

the calculation of the proportion of the two levels (or classes) in the Y variable. Since the population is 100, the final result of df is 99.

- The **residual** represents the number of errors reduced by introducing the X variable. Again, 1 df is given to the introduction of X variable. The difference between the null variance and residual is distributed as chi-square. According to the textbook, one can use the ***anova(glmOut, test='Chisq')*** command to get the exact same result. If the p-value is lower than 0.05, then one can reject the null hypothesis that "the introduction of X into the model caused 0 reduction of model error".
- The **AIC** can be used to compare the performance of nonnested models (models that do not have exact same predictors). Lower AIC indicates model with better performance.

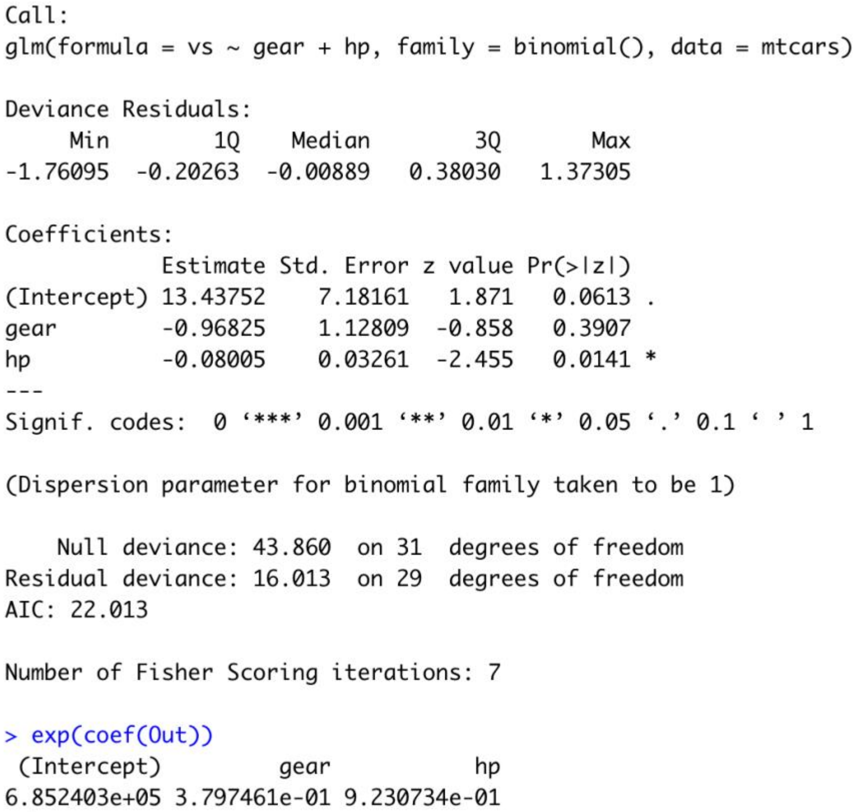
Multinomial Logistic Regression

There are several types of logistics regression. One is Binomial Logistic Regression whose result is binary (dichotomous). Another one is multinomial logistic model whose results can be classified into more than 2 groups.

- To classify the results into different groups, one can use one of the groups as baseline while compare the other groups to this baseline. Namely, the multinomial model consist of k-1 binomial models.

Exercise Review

1.



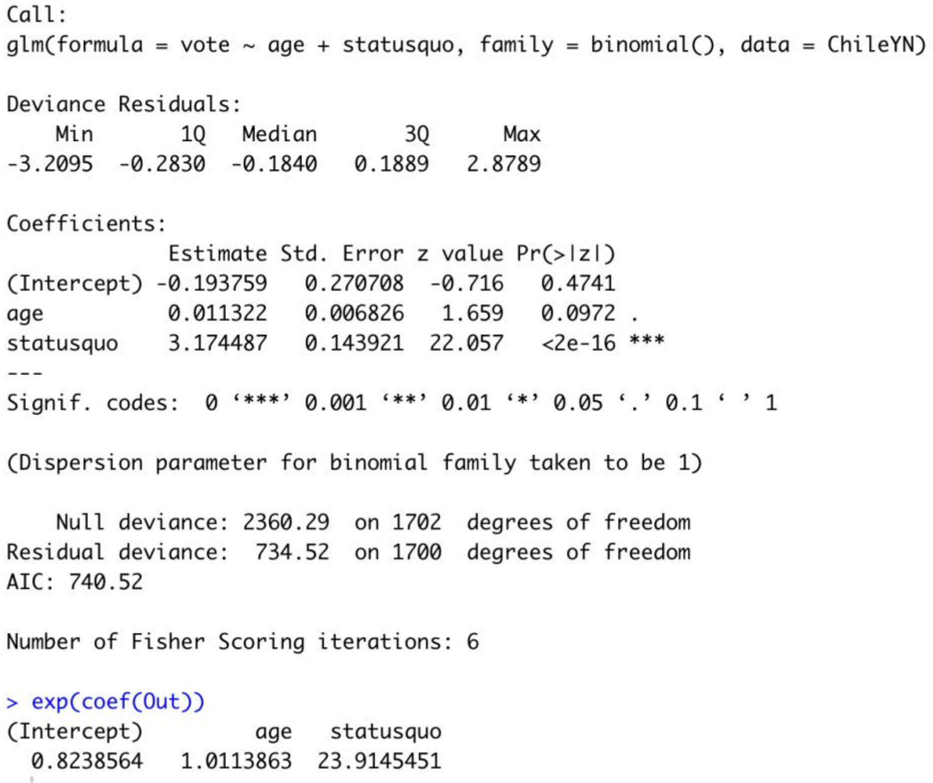
```
Call:
glm(formula = vs ~ gear + hp, family = binomial(), data = mtcars)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.76095  -0.20263  -0.00889   0.38030   1.37305

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 13.43752    7.18161   1.871   0.0613 .
gear        -0.96825    1.12809  -0.858   0.3907
hp          -0.08005    0.03261  -2.455   0.0141 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.860  on 31  degrees of freedom
Residual deviance: 16.013  on 29  degrees of freedom
AIC: 22.013

Number of Fisher Scoring iterations: 7

> exp(coef(Out))
 (Intercept)         gear           hp
6.852403e+05 3.797461e-01 9.230734e-01
```

The p-value of Intercept and gear are higher than 0.05, indicating that both of the two results are not significant. The p-value of hp is lower than 0.05, indicating that we can reject the null hypothesis that the coefficient of hp is not 0.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

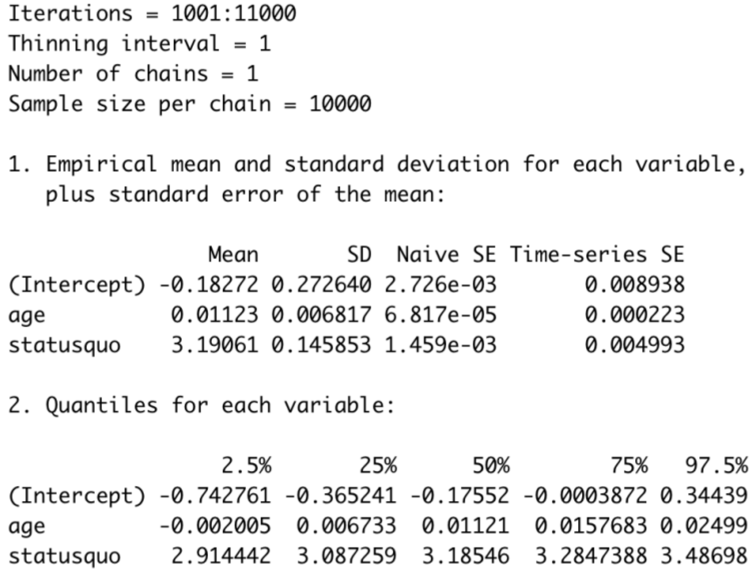
5.
The package is not available for the current version of RStudio.

6.



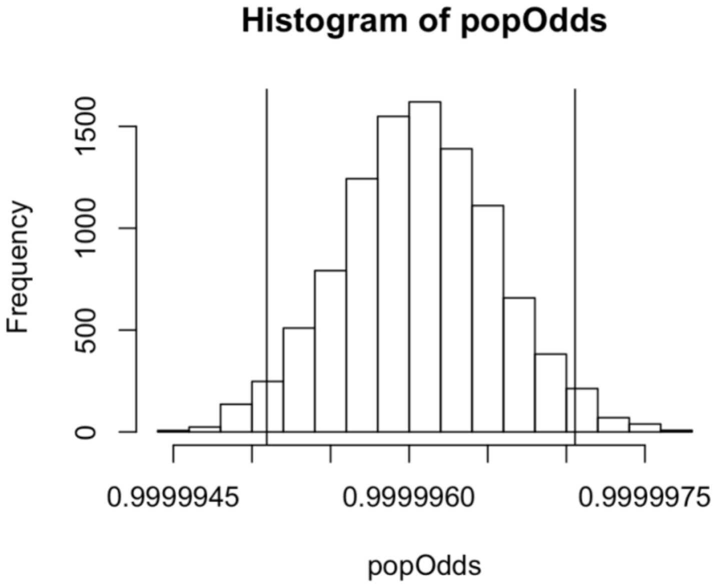
```
Call:
glm(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2095  -0.2830  -0.1840   0.1889   2.8789

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.193759   0.270708  -0.716   0.4741
age          0.011322   0.006826   1.659   0.0972 .
statusquo    3.174487   0.143921  22.057   <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: 2360.29  on 1702  degrees of freedom
Residual deviance:  734.52  on 1700  degrees of freedom
AIC: 740.52

Number of Fisher Scoring iterations: 6

> exp(coef(Out))
(Intercept)         age   statusquo
  0.8238564   1.0113863  23.9145451
```

The p-values of Intercept and age are higher than 0.05, indicating that both of those two results are not significant. Namely, one fail to reject the null hypothesis. The p-value of statusquo is lower than 0.05, indicating that we can reject the null hypothesis that the coefficient of statusquo is not 0. The AIC value of this model is far lower than the model created in the chapter, therefore, one can conclude that this model has a better performance.



```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean       SD  Naive SE Time-series SE
(Intercept) -0.18272 0.272640 2.726e-03       0.008938
age          0.01123 0.006817 6.817e-05       0.000223
statusquo    3.19061 0.145853 1.459e-03       0.004993

2. Quantiles for each variable:

                 2.5%       25%      50%        75%   97.5%
(Intercept) -0.742761 -0.365241 -0.17552 -0.0003872 0.34439
age         -0.002005  0.006733  0.01121  0.0157683 0.02499
statusquo    2.914442  3.087259  3.18546  3.2847388 3.48698
```

7.



If we use the population as the predictor, the result seems to be 1:1, indicating that one fail to reject the null hypothesis.

R Code Fragment and Explanation

Generalized linear models are fit using the **glm()** function.

glm(*formula***, family=***familytype***(link=***linkfunction***), data=)**

Family	**Default Link Function**
binomial	(link = "logit")
gaussian	(link = "identity")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")

Code Example:

```
fit <- glm(F~x1+x2+x3,data=mydata,family=binomial())
summary(fit)             # display results
confint(fit)             # get the 95% CI for the coefficients
exp(coef(fit))           # get the exponentiated coefficients
exp(confint(fit))        # get the 95% CI for exponentiated coefficients
predict(fit, type="response")          # get the predicted values
residuals(fit, type="deviance")        # get the residuals
```

Question for Class

1. What is the difference between logit and probit?