Homework8

Jyoti Chaudhary

April 16, 2018

Problem 1.a

Fit a logistic regression model to the response variable y.

```
data1 <- read.csv(paste(getwd(),"/Data_HW_8_1.csv",sep = ""),header=T)

mdl1= glm ( y ~ x, data = data1, family = binomial)

summary(mdl1) ##### PARAMETER ESTIMATES
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial, data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0620  -0.4868   0.3915   0.5476   2.1682
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.070884   2.108996   2.879  0.00399 **
## x           -0.017705   0.006076  -2.914  0.00357 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.617  on 24  degrees of freedom
## Residual deviance: 20.364  on 23  degrees of freedom
## AIC: 24.364
##
## Number of Fisher Scoring iterations: 4
```

```
# The coefficients obtained are:

mdl1$coefficients
```

```
## (Intercept)           x
##   6.0708839  -0.0177047
```

The fitted model is:

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{-(6.07 - 0.0177x)}} = \frac{1}{1 + e^{-6.07 + 0.0177x}}$$

(1.b)

Does the model deviance indicate that the logistic regression model from part a is adequate?

```
anova(mdl1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    24     34.617
## x     1   14.254        23     20.364 0.0001597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Deviance = 20.364 with df = 23. As D/df = 0.8 which is close to 1.00, this indicates that the model is adequate and good fit for the data. Also, the p-value (=0.0001597) is very low which further indicates that model is adequate.

(1.c)

Provide an interpretation of the parameter beta1 in this model.

```
OR=exp(coef(mdl1)[2])
OR
```

```
##         x
## 0.9824511
```

This odds ratio implies that for every unit increase in speed, the odds of hitting the target decrease by 1.75%.

(1.d)

Expand the linear predictor to include a quadratic term in target speed. Is there any evidence that this quadratic term is required in the model?

```
mdl2= glm ( y ~ x + x^2, data = data1, family = binomial)

anova(mdl1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                   24     34.617
## x     1   14.254         23     20.364 0.0001597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is no difference in the deviance of the two models and hence there is no need for the quadratic term.

(2.a)

Fit a logistic regression model to the data.

```
data2 <- read.csv(paste(getwd(),"/Data_HW_8_2.csv",sep = ""),header=T)

mdl3= glm ( y ~ x1 + x2, data = data2, family = binomial)

summary(mdl3) ##### PARAMETER ESTIMATES
```

```
##
## Call:
## glm(formula = y ~ x1 + x2, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5635  -0.8045  -0.1397   0.9535   1.7915
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.047e+00  4.674e+00  -1.508    0.132
## x1           7.382e-05  6.371e-05   1.159    0.247
## x2           9.879e-01  5.274e-01   1.873    0.061 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 21.082  on 17  degrees of freedom
## AIC: 27.082
##
## Number of Fisher Scoring iterations: 5
```

```
# The coefficients obtained are:

mdl3$coefficients
```

```
##   (Intercept)            x1            x2 
## -7.047061e+00  7.381679e-05  9.878861e-01
```

The fitted model is:

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{-(-7.047 + 7.382e{-}05x1 + 0.9879x2)}} = \frac{1}{1 + e^{7.047 - 7.382e{-}05x1 - 0.9879x2}}$$

```

# (2.b)

Does the model deviance indicate that the logistic regression model from part a is adequate?

```
anova(mdl3, test="Chisq")
```

```
## Analysis of Deviance Table
## 
## Model: binomial, link: logit
## 
## Response: y
## 
## Terms added sequentially (first to last)
## 
## 
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)  
## NULL                    19     27.726           
## x1    1   0.7349        18     26.991  0.39129  
## x2    1   5.9094        17     21.081  0.01506 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Deviance = 21.08 with df = 17. As D/df = 1.24 which is greater than 1.00, this indicates that the model fits the data, however model fit can be improved by further analysis.Also, the p-value (=0.01506) is low which further indicates that model is adequate.

(2.c)

Interpret the model coefficients beta1 and beta2.

```
OR=exp(coef(mdl3))
OR
```
```

```
##  (Intercept)           x1           x2 
## 0.0008699617 1.0000738195 2.6855513881
```

The above odds ratio implies that:

- For a unit increase in the income, the odds of purchasing a new vehicle increase by 0.0074%.
- For a unit increase in the age of family's oldest vehicle, the odds of purchasing a new vehicle increase by 268.5% which implies the odds are more than approx 2.5 times.

(2.d)

What is the estimated probability that a family with an income of $45,000 and a car that is 5 years old will purchase a new vehicle in the next 6 months?

```
pred2=subset(data.frame(data2,predict(mdl3, se.fit=TRUE, type='response')),select = -c(residual.
scale))

nwdt=with(pred2, data.frame(x1=45000,x2=5))  # NEW DATA POINT
nwdt
```

```
##      x1 x2
## 1 45000  5
```

```
pct=0.95

nwdt2=subset(cbind(nwdt,predict(mdl3,newdata=nwdt, type="link", se=TRUE)),select = -c(residual.s
cale))
nwdt2
```

```
##      x1 x2      fit    se.fit
## 1 45000  5 1.214124 0.8630815
```

```
nwdt3=within(nwdt2,{PredictedProb <- plogis(fit)
LL <- plogis(fit - (qnorm((1+pct)/2) * se.fit))
UL <- plogis(fit + (qnorm((1+pct)/2) * se.fit))})
nwdt3
```

```
##      x1 x2      fit    se.fit        UL        LL PredictedProb
## 1 45000  5 1.214124 0.8630815 0.9481291 0.3828464     0.7710279
```

We can see from the above result that the predicted probability is 0.77 of purchasing a new vehicle.

(2.e)

Expand the linear predictor to include an interaction term. Is there any evidence that this term is required in the model?

```r
mdl4= glm ( y ~ x1*x2, data = data2, family = binomial)

summary(mdl4) ##### PARAMETER ESTIMATES
```

```
##
## Call:
## glm(formula = y ~ x1 * x2, family = binomial, data = data2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.63981  -0.62754  -0.05642   0.66213   1.85666
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.144e-01  6.394e+00   0.049    0.961
## x1          -1.411e-04  1.412e-04  -0.999    0.318
## x2          -2.462e+00  2.081e+00  -1.183    0.237
## x1:x2        1.014e-04  6.297e-05   1.610    0.107
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 16.551  on 16  degrees of freedom
## AIC: 24.551
##
## Number of Fisher Scoring iterations: 6
```

```r
anova(mdl4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                     19     27.726
## x1     1   0.7349        18     26.991  0.39129
## x2     1   5.9094        17     21.081  0.01506 *
## x1:x2  1   4.5307        16     16.551  0.03329 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The deviance has significantly reduced due to interaction term with a p-value of 0.033. As D/df = 1.03, this indicates that the model is adequate and good fit for the data. Hence its evident that the interaction term is required in the model.

(2.f)

If income goes up by $1000 in model of part (a) while age remain fixed, how much the odds of buying change.

```
OR=exp(coef(mdl3)[2])
OR
```

```
##       x1
## 1.000074
```

From the odds ratio we can determine that if the income goes up by $1000, the odds of purchasing a new vehicle increase by 7.4%.

(2.g)

Find approximate 95% confidence intervals on the model parameters for the logistic regression model from part a.

```
confint(mdl3, level=0.95)
```

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

```
##                     2.5 %       97.5 %
## (Intercept) -1.805544e+01 1.0275430082
## x1          -4.361540e-05 0.0002184223
## x2           1.544228e-01 2.2872127855
```