Solution to Homework for Logistic Regression

1. Logistic regression for the combined data on incubation temperature and number of male and female turtles from eggs collected in Illinois is presented in Examples 5.5-5.8. The original data is given below.

<table>
<thead>
<tr>
<th>Temp</th>
<th>male</th>
<th>female</th>
<th>% male</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>27.2</td>
<td>0</td>
<td>8</td>
<td>11%</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>20%</td>
<td>80%</td>
</tr>
<tr>
<td>28.4</td>
<td>5</td>
<td>3</td>
<td>63%</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>27.7</td>
<td>4</td>
<td>2</td>
<td>67%</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>28.3</td>
<td>6</td>
<td>3</td>
<td>67%</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>88%</td>
<td></td>
</tr>
<tr>
<td>29.9</td>
<td>8</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

(a) Use the complete data, 3 observed proportions for each temperature, to fit a logistic regression model. You can fit the model in either of two ways:
- Use the proportion male, as the response and use the number of turtles as weights.
- Use cbind to create a two column response containing the number of males and the number of females.

How does this fit compare to that of the combined data? Look at the residual deviance as well as the fitted equation.

The logistic regression fit for the original data produces the same intercept and slope coefficients as those for the combined data.

\[ \hat{\pi}_i' = -61.3183 + 2.2110 \text{Temp} \]

The residual deviance is now 24.94 on 13 degrees of freedom. This has a P-value of 0.0235. This is a better fit than before but there is still a significant lack of fit for the model.

(b) Is temperature significant in the logistic regression model using the complete data? Justify your answer statistically.

The change in deviance is the same as in the combined data analysis; 44.56555 on 1 df. This is highly significant with a P-value virtually zero. Temperature is a statistically significant predictor of the proportion of male turtles in the logistic regression model.

(c) What is the incubation temperature that would give a predicted proportion of males of 50%?

For a 50:50 split, an odds of 1, we want the incubation temperature that would have a predicted log-odds of 0; temperature=61.3183/2.211=27.7 degrees.
Below is S+ code and output to perform the analysis. The data has been read in from a data frame and is stored in turtle.IL

```R
> turtle.IL
  Temp  Male  Female Total
1 27.2   1     9    10
2 27.7   7     3    10
3 28.3  13     0    13
4 28.4   7     3    10
5 29.9  10     1    11
6 27.2   0     8     8
7 27.7   4     2     6
8 28.3   6     3     9
9 28.4   5     3     8
10 29.9  8     0     8
11 27.2   1     8     9
12 27.7   6     2     8
13 28.3   7     1     8
14 28.4   7     2     9
15 29.9  9     0     9
> attach(turtle.IL)

#---------------------------------------
# fit glm using proportion male and the number of turtles as weights
#---------------------------------------

> PMale.IL <- Male/Total
> turtle.IL.logit1 <- glm(PMale~Temp,family=binomial,weights=Total)
> summary(turtle.IL.logit1)

Deviance Residuals:
  Min      1Q  Median      3Q     Max
-2.07212 -1.02915 -0.27137  0.80867  2.55496

Coefficients:
                        Value Std. Error t value
(Intercept)          -61.3183   12.0156  -5.1032
Temp                   2.2110    0.4306   5.1347

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 74.50804 on 14 degrees of freedom
Residual Deviance: 24.94249 on 13 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:
                          (Intercept)         Temp
(Intercept)            -0.9998299
Temp                   -0.9998299

#---------------------------------------
# fit glm using two column response matrix of number of males, females
#---------------------------------------
```

2
> Response.IL <- cbind(Male,Female)
> turtle.IL.logit2 <- glm(Response~Temp,family=binomial)
> summary(turtle.IL.logit2)

Deviance Residuals:
    Min 1Q Median 3Q Max
-2.072127 -1.029156 -0.2713629 0.808678 2.554963

Coefficients:
             Value Std. Error t value
(Intercept) -61.317982  12.0002551 -5.109723
    Temp   2.211019   0.4300549  5.141249

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 74.50804 on 14 degrees of freedom
Residual Deviance: 24.94249 on 13 degrees of freedom
Number of Fisher Scoring Iterations: 4

Correlation of Coefficients:
       (Intercept)     Temp
(Intercept)   1.0000000
    Temp    -0.9998297

#--------------------------------------------------------
# plot the data and the fitted logistic regression curve
#--------------------------------------------------------

> t.IL <- c(270:300/10)
> yhat.IL <- coefficients(turtle.IL.logit2)[c(1)]+
coefficients(turtle.IL.logit2)[c(2)]*t.IL
> pihat.IL <- exp(yhat.IL)/(1+exp(yhat.IL))

> plot(Temp,PMale,xlab="Incubation Temperature",ylab="Proportion Male",
+ main="Proportion of Male Turtles (IL) as a Function of Temperature")
> lines(t.IL,pihat.IL)
Lines out of bounds X= 27.1 Y= 0.1979173
2. There is also data on the relationship between the number of male turtles and incubation temperature for turtle eggs from New Mexico. The turtles are the same species as those from Illinois.

<table>
<thead>
<tr>
<th>Temp</th>
<th>male</th>
<th>female</th>
<th>% male</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.20</td>
<td>0</td>
<td>5</td>
<td>0%</td>
</tr>
<tr>
<td>27.20</td>
<td>0</td>
<td>3</td>
<td>0%</td>
</tr>
<tr>
<td>27.20</td>
<td>0</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>28.30</td>
<td>6</td>
<td>1</td>
<td>86%</td>
</tr>
<tr>
<td>28.30</td>
<td>2</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>28.30</td>
<td>0</td>
<td>3</td>
<td>0%</td>
</tr>
<tr>
<td>29.90</td>
<td>4</td>
<td>1</td>
<td>80%</td>
</tr>
<tr>
<td>29.90</td>
<td>1</td>
<td>1</td>
<td>50%</td>
</tr>
<tr>
<td>29.90</td>
<td>3</td>
<td>0</td>
<td>100%</td>
</tr>
</tbody>
</table>

(a) Use logistic regression to analyze these data. You can either use the proportion male as the response with the number of turtles as weights or you can use cbind to create a two column response containing the number of males and number of females. You do NOT have to use both ways, only one. Turn in the summary of the logistic regression fit. Give the equation, comment on the residual deviance and what it indicates, and test to see if temperature is significant.

The New Mexico data has fewer eggs and only three temperatures. The logistic regression gives the equations:

\[ \hat{\pi}_i' = -41.7419 + 1.4696 \text{Temp} \]

This is quite a bit different from the Illinois data. The residual deviance is 17.84 on 7 degrees of freedom. This corresponds to a P-value of 0.0127. Again there is significant lack of fit in the model. The change in deviance is 30.84318 minus 17.84488 or 12.99831 on 1 df. This is a highly significant result with a P-value that is virtually zero. Again, temperature is a significant predictor of the proportion male.
(b) What is the temperature at which you would get a 50:50 split of males to females?

The temperature for a 50:50 split (odds=1) is found by looking for the temperature that gives you a log-odds of 0; temperature=41.7419/1.4696=28.4 degrees. This is over half a degree higher that the temperature for Illinois turtles.

c) Turn in a plot of the data with the logistic regression curve superimposed. Make sure your plot has appropriate labels and a title.

(d) How do the New Mexico turtles compare to the Illinois turtles in terms of the effect of temperature on the sex of the turtles?

As mentioned above, temperature is significant for both the Illinois and New Mexico turtles. The equations are quite a bit different, resulting in over a half a degree difference in the temperature to obtain a 50:50 split. The two curves are presented in the figure below.

e) How would you analyze the Illinois and New Mexico data together? You do not have to do this analysis, simply tell me what variables you would include in your model and what procedure in you would use to fit the model.

Temperature would be one variable. A second variable would be an indicator variable that indicated whether the proportion male was for Illinois turtles or New Mexico turtles. A third variable would be a cross product of the temperature and the indicator. These three variables would allow you to fit separate slopes and separate intercepts for the two groups of turtles and test to see if there were significant differences in slopes or intercepts or both. This also allows you to combine the data to get more degrees of freedom for the residual deviance.
Below is R code and output to perform the analysis. The data has been read in from a data frame and is stored in turtle.NM

```r
> turtle.NM
   Temp Male Female Total
1   27.2   0     5     5
2   27.2   0     3     3
3   27.2   0     2     2
4   28.3   6     1     7
5   28.3   2     0     2
6   28.3   0     3     3
7   29.9   4     1     5
8   29.9   1     1     2
9   29.9   3     0     3
> attach(turtle.NM)
> Response.NM <- cbind(Male,Female)
> turtle.NM.logit <- glm(Response.NM~Temp,family=binomial)
> summary(turtle.NM.logit)

Deviance Residuals:
       Min      1Q  Median      3Q     Max
-1.928218 -1.254542 -0.793442  0.794589  2.183381

Coefficients:
                               Estimate Std. Error t value
(Intercept)               -41.741852  14.5174040 -2.8753
Temp                         1.469578   0.5124967  2.8675

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 30.84318 on 8 degrees of freedom
Residual Deviance: 17.84488 on 7 degrees of freedom
```
3. A study was conducted to see the effect of coupons on purchasing habits of potential customers. In the study, 1000 homes were selected and a coupon and advertising material for a particular product was sent to each home. The advertising material was the same but the amount of the discount on the coupon varied from 5% to 30%. The number of coupons redeemed was counted. Below are the data.

<table>
<thead>
<tr>
<th>Price Reduction (X_i)</th>
<th>Number of Coupons (n_i)</th>
<th>Number Redeemed (Y_i)</th>
<th>Proportion Redeemed (p_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>200</td>
<td>32</td>
<td>0.160</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
<td>51</td>
<td>0.255</td>
</tr>
<tr>
<td>15</td>
<td>200</td>
<td>70</td>
<td>0.350</td>
</tr>
<tr>
<td>20</td>
<td>200</td>
<td>103</td>
<td>0.515</td>
</tr>
<tr>
<td>30</td>
<td>200</td>
<td>148</td>
<td>0.740</td>
</tr>
</tbody>
</table>

(a) Fit a simple linear regression to the observed proportions. Is there a significant linear relationship between the proportion redeemed and the price reduction? Use this regression to estimate the proportion redeemed. According to this regression at what price reduction will you get a 25% redemption rate?

Simple linear regression of observed proportions redeemed on the price reduction.

The simple linear regression equation is

\[ \hat{p}_i = 0.0243 + 0.0237 \times \text{PriceReduce} \]

The slope coefficient for Price Reduction has a t-value of 20.26 with an associated P-value of 0.0003. Such a small P-value indicates that such a large slope coefficient could not have happened by chance alone and therefore there is a significant linear relationship between the proportion redeemed and price reduction. A note of caution, we are violating the assumptions for the t-test. However, the result is so extreme the violation is comparatively unimportant. If we want a 25% redemption rate we need to offer a 9.5% price reduction.
Coefficients:

Value Std. Error t value Pr(>|t|)
(Intercept) 0.0243 0.0213 1.1431 0.3359
PriceReduce 0.0237 0.0012 20.2585 0.0003

Residual standard error: 0.02253 on 3 degrees of freedom
Multiple R-Squared: 0.9927
F-statistic: 410.4 on 1 and 3 degrees of freedom, the p-value is 0.0002629

> predict(Redeemed.slr)
     1     2     3     4     5
0.14297 0.26162 0.38027 0.49892 0.73622

> Price.slr <- (0.25 - coefficients(Redeemed.slr)[c(1)])
                    + coefficients(Redeemed.slr)[c(2)]
> Price.slr
9.510251

(b) Compute the logits for the observed proportions at each price reduction level.

\[
\begin{array}{cccc}
\text{Price Reduction} & \text{Number of Coupons} & \text{Number Redeemed} & \text{Proportion Redeemed} & \text{Logits} \\
X_i & n_i & Y_i & p_i & \\
5 & 200 & 32 & 0.160 & -1.65823 \\
10 & 200 & 51 & 0.255 & -1.07212 \\
15 & 200 & 70 & 0.350 & -0.61904 \\
20 & 200 & 103 & 0.515 & 0.06002 \\
30 & 200 & 148 & 0.740 & 1.04597 \\
\end{array}
\]

> logits <- log(PRedeemed/(1-PRedeemed))
> logits
[1] -1.65822808 -1.07212067 -0.61903921 0.06001801 1.04596856

(c) Fit a simple linear regression of the logit transformed proportions on the price reduction. Is there a significant linear relationship between the logit and the price reduction? Use this regression to estimate the proportion redeemed. According to this regression at what price reduction will you get a 25% redemption rate?

The simple linear regression of the logit of the proportion redeemed on the price reduction is:

\[
\hat{\pi}_i' = -2.18602 + 0.10858 \text{PriceReduce}
\]

See the table below for the estimates of the proportions redeemed.

\[
\begin{array}{cccccc}
\text{Price Reduction} & \text{Number of Coupons} & \text{Number Redeemed} & \text{Proportion Redeemed} & \text{Predicted Proportions} \\
X_i & n_i & Y_i & p_i & \hat{p}_i & \\
5 & 200 & 32 & 0.160 & 0.162 \\
10 & 200 & 51 & 0.255 & 0.250 \\
15 & 200 & 70 & 0.350 & 0.364 \\
20 & 200 & 103 & 0.515 & 0.496 \\
30 & 200 & 148 & 0.740 & 0.745 \\
\end{array}
\]
The t-value is 34.47 with a P-value of essentially zero. Such a small P-value indicates that such a large slope coefficient could not have happened by chance alone and therefore there is a significant linear relationship between the logit and price reduction. A note of caution, we are violating the assumptions for the t-test. However, the result is so extreme the violation is comparatively unimportant. If we want a 25% redemption rate we need to offer a 10% price reduction.

```r
> Redeemed.logitslr <- glm(logits~PriceReduce)
> summary(Redeemed.logitslr)
```

Deviance Residuals:
```
1    2    3    4    5
-0.01512903 0.02806075 -0.06177541 0.07436419 -0.02552055
```

Coefficients: Value Std. Error t value
```
(Intercept) -2.1860167 0.057222197 -38.20225
PriceReduce 0.1085835 0.003149978 34.47120
```

(Dispersion Parameter for Gaussian family taken to be 0.0036713 )

Null Deviance: 4.373455 on 4 degrees of freedom

Residual Deviance: 0.0110138 on 3 degrees of freedom

Number of Fisher Scoring Iterations: 1

```r
> pihat.logitslr <- exp(predict(Redeemed.logitslr)) + 1 + exp(predict(Redeemed.logitslr)))
> pihat.logitslr
```

```
1    2    3    4    5
0.1620438 0.2497059 0.3641808 0.4964135 0.74488
```

```r
> Price.logitslr <- (log(1/3)-coefficients(Redeemed.logitslr)[c(1)]) + coefficients(Redeemed.logitslr)[c(2)]
> Price.logitslr
```

```
10.01445
```

(d) Use the general linear model (glm) function in S+ to fit a logistic regression of the proportion redeemed on the price reduction. Use this regression to estimate the proportion redeemed. According to this regression at what price reduction will you get a 25% redemption rate?

The logistic regression of the of the proportion redeemed on the price reduction is:

\[ \hat{\pi} = -2.18551 + 0.10872 \cdot \text{PriceReduce} \]

See the table below for the estimates of the proportions redeemed.

<table>
<thead>
<tr>
<th>Price Reduction</th>
<th>Number of Coupons</th>
<th>Number Redeemed</th>
<th>Proportion Redeemed</th>
<th>Predicted Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_i )</td>
<td>( n_i )</td>
<td>( Y_i )</td>
<td>( p_i )</td>
<td>( \hat{p}_i )</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>32</td>
<td>0.160</td>
<td>0.162</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
<td>51</td>
<td>0.255</td>
<td>0.250</td>
</tr>
<tr>
<td>15</td>
<td>200</td>
<td>70</td>
<td>0.350</td>
<td>0.365</td>
</tr>
<tr>
<td>20</td>
<td>200</td>
<td>103</td>
<td>0.515</td>
<td>0.497</td>
</tr>
<tr>
<td>30</td>
<td>200</td>
<td>148</td>
<td>0.740</td>
<td>0.746</td>
</tr>
</tbody>
</table>
The residual deviance is 0.5102 on 3 df. Such a small residual deviance indicates that the logistic regression model is a very good fit. The change in deviance is 180.54 on 1 df. The associated P-value would be virtually zero. This indicates that the predictor variable, Price Reduction is statistically significant.

If we want a 25% redemption rate we need to offer a 10% price reduction.

```r
> Redeemed.logistic <- glm(PRedeemed~PriceReduce,family=binomial, weights=Coupons)
> summary(Redeemed.logistic)

Deviance Residuals:
  1       2       3       4       5
-0.08476077 0.1611105 -0.4352312 0.5029435 -0.1862615

Coefficients:
                 Value Std. Error t value
(Intercept) -2.1855052 0.16466609 -13.27235
PriceReduce  0.1087191 0.00884287 12.29455

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 180.9475 on 4 degrees of freedom
Residual Deviance: 0.5102127 on 3 degrees of freedom
Number of Fisher Scoring Iterations: 4
```

```r
> pihat.logistic <- exp(predict(Redeemed.logistic))/(1+exp(predict(Redeemed.logistic)))
> pihat.logistic

1       2       3       4       5
0.1622054 0.2500559 0.3647703 0.4972192 0.7457491

> Price.logistic <- (log(1/3)-coefficients(Redeemed.logistic)[c(1)])/
coefficients(Redeemed.logistic)[c(2)]

> Price.logistic

9.997258
```

(c) Compare the three regression equations and price reductions to get a 25% redemption rate.

Comparison of equations, predicted proportions and predicted price reduction necessary for a 25% redemption.

The two equations using the logit transform (logitslr and logistic) are virtually the same in terms of the estimates. The standard errors are somewhat different. The predicted proportions are very similar, especially for the logitlr and the logistic regression. The slr model indicated a 9.5% price reduction while the other two models say to offer a 10% price reduction to get a 25% redemption rate.
<table>
<thead>
<tr>
<th>Price Reduction</th>
<th>Proportion Redeemed</th>
<th>SLR Predicted Proportion ($p_i^{(slr)}$)</th>
<th>LogitSLR Predicted Proportion ($p_i^{(logitslr)}$)</th>
<th>Logistic Predicted Proportion ($p_i^{(logistic)}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.160</td>
<td>0.14297</td>
<td>0.16204</td>
<td>0.16221</td>
</tr>
<tr>
<td>10</td>
<td>0.255</td>
<td>0.26162</td>
<td>0.24971</td>
<td>0.25006</td>
</tr>
<tr>
<td>15</td>
<td>0.350</td>
<td>0.38027</td>
<td>0.36418</td>
<td>0.36477</td>
</tr>
<tr>
<td>20</td>
<td>0.515</td>
<td>0.49892</td>
<td>0.49641</td>
<td>0.49722</td>
</tr>
<tr>
<td>30</td>
<td>0.740</td>
<td>0.73622</td>
<td>0.74488</td>
<td>0.74575</td>
</tr>
</tbody>
</table>

(f) Create plots that show the data and the fits.

Each of the models is so close that the resolution of the plots does not highlight the difference. One plot, the logistic regression fit is given below.
4. Kyphosis is a spinal deformity found in young children who have corrective spinal surgery. The incidence of spinal deformities following corrective spinal surgery (Kyp=1 if deformity is present, Kyp=0 if there is no deformity present) is thought to be related to the Age (in months) at the time of surgery, Start (the starting vertebra for the surgery) and Num (the number of vertebrae involved in the surgery).

| Age | Start | Num | Kyp | Age | Start | Num | Kyp | Age | Start | Num | Kyp |
|-----|-------|-----|-----|-----|-------|-----|-----|-----|-------|-----|-----|-----|
| 71  | 5     | 3   | 0   | 100 | 14    | 3   | 0   | 140 | 15    | 4   | 0   |
| 158 | 14    | 3   | 0   | 4   | 16    | 3   | 0   | 72  | 15    | 5   | 0   |
| 128 | 5     | 4   | 1   | 151 | 16    | 2   | 0   | 2   | 13    | 3   | 0   |
| 2   | 1     | 5   | 0   | 31  | 11    | 3   | 0   | 120 | 8     | 5   | 1   |
| 1   | 15    | 4   | 0   | 125 | 11    | 2   | 0   | 51  | 9     | 7   | 0   |
| 1   | 16    | 2   | 0   | 130 | 13    | 5   | 0   | 102 | 13    | 3   | 0   |
| 61  | 17    | 2   | 0   | 112 | 16    | 3   | 0   | 130 | 1     | 4   | 1   |
| 37  | 16    | 3   | 0   | 140 | 11    | 5   | 0   | 114 | 8     | 7   | 1   |
| 113 | 16    | 2   | 0   | 93  | 16    | 3   | 0   | 81  | 1     | 4   | 0   |
| 59  | 12    | 6   | 1   | 1   | 9     | 3   | 0   | 118 | 16    | 3   | 0   |
| 82  | 14    | 5   | 1   | 52  | 6     | 5   | 1   | 118 | 16    | 4   | 0   |
| 148 | 16    | 3   | 0   | 20  | 9     | 6   | 0   | 17  | 10    | 4   | 0   |
| 18  | 2     | 5   | 0   | 91  | 12    | 5   | 1   | 195 | 17    | 2   | 0   |
| 1   | 12    | 4   | 0   | 73  | 1     | 5   | 1   | 159 | 13    | 4   | 0   |
| 168 | 18    | 3   | 0   | 35  | 13    | 3   | 0   | 18  | 11    | 4   | 0   |
| 1   | 16    | 3   | 0   | 143 | 3     | 9   | 0   | 15  | 16    | 5   | 0   |
| 78  | 15    | 6   | 0   | 61  | 1     | 4   | 0   | 158 | 14    | 5   | 0   |
| 175 | 13    | 5   | 0   | 97  | 16    | 3   | 0   | 127 | 12    | 4   | 0   |
| 80  | 16    | 5   | 0   | 139 | 10    | 3   | 1   | 87  | 16    | 4   | 0   |
| 27  | 9     | 4   | 0   | 136 | 15    | 4   | 0   | 206 | 10    | 4   | 0   |
| 22  | 16    | 2   | 0   | 131 | 13    | 5   | 0   | 11  | 15    | 3   | 0   |
| 105 | 5     | 6   | 1   | 121 | 3     | 3   | 1   | 178 | 15    | 4   | 0   |
| 96  | 12    | 3   | 1   | 177 | 14    | 2   | 0   | 157 | 13    | 3   | 1   |
| 131 | 3     | 2   | 0   | 68  | 10    | 5   | 0   | 26  | 13    | 7   | 0   |
| 15  | 2     | 7   | 1   | 9   | 17    | 2   | 0   | 120 | 13    | 2   | 0   |
| 9   | 13    | 5   | 0   | 139 | 6     | 10  | 1   | 42  | 6     | 7   | 1   |
| 8   | 6     | 3   | 0   | 2   | 17    | 2   | 0   | 36  | 13    | 4   | 0   |

(a) Plot the binary response for the incidence of Kyphosis versus the age of the child. Fit a logistic regression of incidence of Kyphosis on Age. Examine the fit, significance of Age and look at the residuals.

The logistic regression of kyphosis on Age is

\[ \hat{\pi}' = -1.8091 + 0.00544 \times \text{Age} \]

The residual deviance is 81.93 on 79 df which indicates no significant lack of fit. However the change in deviance is only 1.302 on 1 df which has a P-value of 0.25. So Age is not a significant predictor of Kyphosis. Looking at the graph, although there is a slight increasing trend in Kyphosis with Age, what is more apparent is that Kyphosis is not as common for very young or older children. A plot of residuals is difficult to interpret because of the binary nature of the response and the poor fit of the model. Below is the S+ code and output for this analysis.

> kyphosis <- read.table("stat415_textdata/kyphosis.dat",head=T)
> kyphosis
attach(kyphosis)
plot(Age,Kyp,xlab="Age in months",ylab="Incidence of kyphosis",
  main="Incidence of kyphosis versus age")
kyp.linA <- glm(Kyp~Age,family=binomial)
summary(kyp.linA)

Call: glm(formula = Kyp ~ Age, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
  -0.9022496 -0.7396895 -0.6028076 -0.5522001  1.944772

Coefficients:
            Value     Std. Error   t value
(Intercept) -1.809076886  0.525924169 -3.439806
  Age         0.005439848  0.004797095  1.133988

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 83.23447 on 80 degrees of freedom
Residual Deviance: 81.93249 on 79 degrees of freedom
Number of Fisher Scoring Iterations: 3

(b) Fit a quadratic logistic regression model in Age. You will need to create a new variable AgeSq = Age*Age. Examine the fit, significance of Age and AgeSq and look at the residuals.

The quadratic logistic regression of Kyphosis on Age is

\[ \hat{\pi}' = -3.7703 + 0.07003 \times Age - 0.00365 \times Age^2 \]

The residual deviance is 72.74 on 78 df which indicates no significant lack of fit. However the change in deviance from the model with only Age in it is 9.19 on 1 df, this is highly significant. Note also that the z-test statistic for AgeSq is \(-2.47\)
which also indicates the significance of the addition of AgeSq. What is somewhat unusual is that the z-test for Age (indicating the significance of Age once AgeSq is in the model) is 2.60 which is also significant. So together both Age and AgeSq are significant predictors of Kyphosis. Below is the S+ code and output for this quadratic logistic regression.

```R
> AgeSq <- Age*Age
> kyp.quadA <- glm(Kyp ~ Age + AgeSq, family=binomial)
> summary(kyp.quadA)
```

Call: `glm(formula = Kyp ~ Age + AgeSq, family = binomial)`

Deviance Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.007884</td>
<td>-0.8411795</td>
<td>-0.4154928</td>
<td>-0.2209433</td>
<td>2.392006</td>
</tr>
</tbody>
</table>

Coefficients:

<table>
<thead>
<tr>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.7702868712</td>
<td>1.1501219836</td>
</tr>
<tr>
<td>Age</td>
<td>0.0700349930</td>
<td>0.0269651682</td>
</tr>
<tr>
<td>AgeSq</td>
<td>-0.0003651679</td>
<td>0.0001477186</td>
</tr>
</tbody>
</table>

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 83.23447 on 80 degrees of freedom
Residual Deviance: 72.73858 on 78 degrees of freedom
Number of Fisher Scoring Iterations: 5

(c) Repeat part (a) with the explanatory variable Number.

The logistic regression of Kyphosis on the Number of vertebrae is

$$\hat{\pi}_i = -3.6510 + 0.5317 \times \text{Num}$$
Both the change in deviance of 9.88 on 1 df and the z-test statistic of 2.88, indicate that Number is a significant predictor of Kyphosis. Again, a plot of residuals is not very informative but it does not indicate a quadratic term is necessary. The S+ code and partial output appears below:

```r
> kyp.linN <- glm(Kyp~Num,family=binomial)
> summary(kyp.linN)
```

Call: glm(formula = Kyp ~ Num, family = binomial)
Deviance Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.68</td>
<td>-0.63</td>
<td>-0.49</td>
<td>-0.38</td>
<td>2.09</td>
</tr>
</tbody>
</table>

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.651</td>
<td>0.898</td>
<td>-4.07</td>
</tr>
<tr>
<td>Num</td>
<td>0.532</td>
<td>0.185</td>
<td>2.88</td>
</tr>
</tbody>
</table>

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 83.23447 on 80 degrees of freedom
Residual Deviance: 73.35706 on 79 degrees of freedom

Number of Fisher Scoring Iterations: 4

(d) Fit a full quadratic logistic regression model in Age and Num, that is include variables Age, Num, Age*Num, AgeSq and NumSq. Examine the fit and significance of each of the variables.

It appears from the z-tests (S+ calls these t values) that Age and AgeSq are statistically significant while Num is close to being significant. From the sequential anova() analysis Num and AgeSq appear to be significant. This suggests that a model with Age AgeSq and Num might do a good job of fitting the Kyphosis data. Below is the S+ code and partial output.
> AgeNum <- Age*Num
> NumSq <- Num*Num
> kyp.quadAN <- glm(Kyp ~ Age + Num + AgeNum + AgeSq + NumSq, family = binomial)
> summary(kyp.quadAN)

Call: glm(formula = Kyp ~ Age + Num + AgeNum + AgeSq + NumSq, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max
-1.838459 -0.632569 3 -0.1766221 -0.007887338 1.992313

Coefficients:

Value Std. Error t value
(Intercept) -1.710103e+01 7.2369589224 -2.363013
Age 1.944470e-01 0.0908284362 2.140817
Num 2.201546e+00 1.1665625196 1.887208
AgeNum -1.383399e-02 0.0077913916 -1.775548
AgeSq -6.214753e-04 0.0002834587 -2.192472
NumSq -3.784011e-03 0.0805485732 -0.046978

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 83.23447 on 80 degrees of freedom

Residual Deviance: 58.67218 on 75 degrees of freedom

Number of Fisher Scoring Iterations: 7

> anova(kyp.quadAN, test = "Chisq")

Analysis of Deviance Table

Binomial model

Response: Kyp

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(Chi)
NULL 80 83.23447
Age 1 1.30198 79 81.93249 0.2538510
Num 1 10.30593 78 71.62656 0.0013260
AgeNum 1 1.59143 77 70.03513 0.2071226
AgeSq 1 11.36074 76 58.67439 0.0007501
NumSq 1 0.00221 75 58.67218 0.9625337

(e) Give a final model that includes only those explanatory variables from the full quadratic logistic regression model that are significant.

The final logistic regression model includes Age, AgeSq and Num

\[ \hat{\pi}' = -6.5269 + 0.07586 \times \text{Age} - 0.000384 \times \text{AgeSq} + 0.5440 \times \text{Num} \]

The residual deviance is 63.86 on 77 df which indicates a very good fit. Also the difference between this residual deviance and that of the full model is 63.86 minus 58.67 or 5.19 on 2 df. This change has a P-value of 0.075 which indicates that dropping the other two variables, NumSq and Age*Num has not decreased
the residual deviance significantly. From the z-test values, all three explanatory variables are statistically significant. Also from the sequential anova() although Age does not appear significant when first entered, we know from the previous analysis that together with AgeSq it is. Below is the S+ code and partial output.

```splus
> kyp.final <- glm(Kyp~Age+AgeSq+Num,family=binomial)
> summary(kyp.final)

Call: glm(formula = Kyp ~ Age + AgeSq + Num, family = binomial)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.78199 -0.719609 -0.3527064 -0.1141014 2.253575

Coefficients:
            Value      Std. Error t value
(Intercept) -6.5269011  1.7771671396 -3.672643
Age          0.0758597  0.0312331308  2.428820
AgeSq       -0.0003843  0.0001693231 -2.269544
Num         0.5440377  0.2055715192  2.646464

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 83.23447 on 80 degrees of freedom
Residual Deviance: 63.86261 on 77 degrees of freedom
Number of Fisher Scoring Iterations: 5

> anova(kyp.final,test="Chisq")

Analysis of Deviance Table

Binomial model
Response: Kyp
Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>80</td>
<td>83.23447</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>1.301985</td>
<td>81.93249</td>
</tr>
<tr>
<td>AgeSq</td>
<td>1</td>
<td>9.193908</td>
<td>72.73858</td>
</tr>
<tr>
<td>Num</td>
<td>1</td>
<td>8.875971</td>
<td>63.86261</td>
</tr>
</tbody>
</table>
```