Hw6_q4

10/7/2019

1.

The survival rate based on the table is 0.7 for females and 0.43 for males. Here we use the logistic regression model's coefficients (which are both significant $p < 0.05$) to say that males are significantly more likely to be dead than females.

```
prop.table(table(donner$Outcome, donner$Sex), 2)
```

```
##
##        Female      Male
##   0 0.2857143 0.5714286
##   1 0.7142857 0.4285714
```

```
model <- glm(Outcome ~ Sex, data=donner, family=binomial(link="logit"))
summary(model)
```

```
##
## Call:
## glm(formula = Outcome ~ Sex, family = binomial(link = "logit"),
##     data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5829  -1.0579   0.8203   1.3018   1.3018
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.9163     0.3742   2.449  0.01433 *
## SexMale      -1.2040     0.4614  -2.609  0.00907 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 125.61  on 90  degrees of freedom
## Residual deviance: 118.36  on 89  degrees of freedom
## AIC: 122.36
##
## Number of Fisher Scoring iterations: 4
```

2. Fit logistic regression.

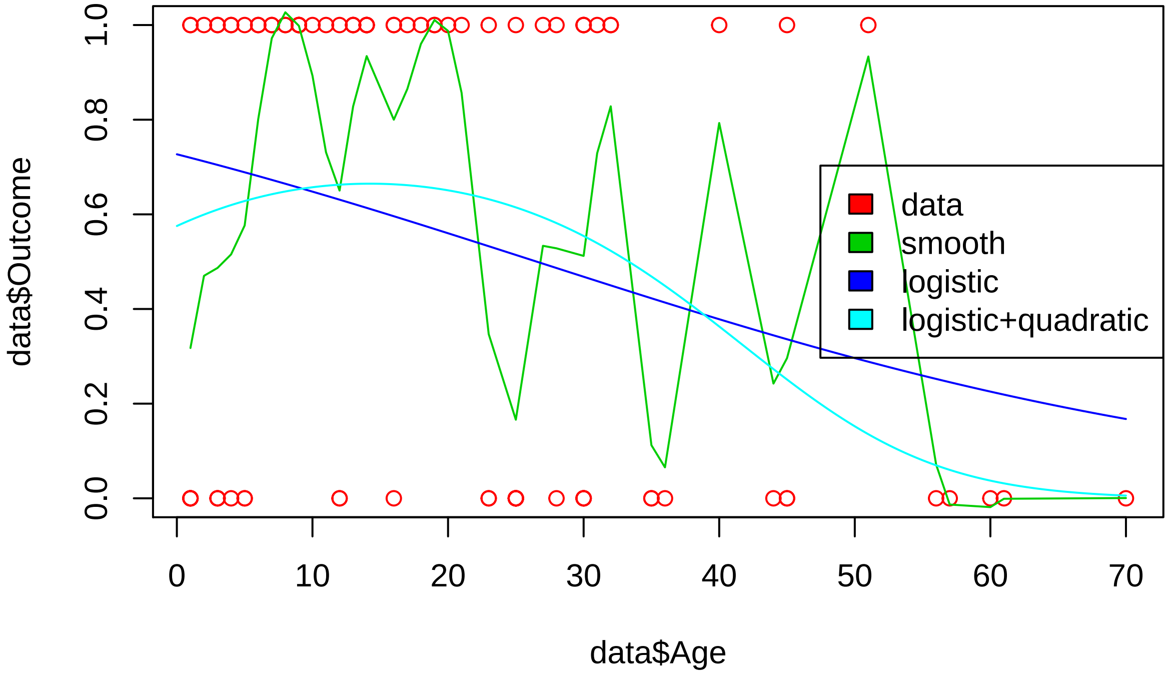
The fitted coefficient for age is -0.037 ± 0.17. This means the probability of living decreases by about 1-exp(-0.037)= 4 percent per additional year lived.

```
##
## Call:
## glm(formula = Outcome ~ Age, family = binomial(link = "logit"),
##     data = donner)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.5946  -1.2017   0.8436   0.9882   1.5765
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.97917    0.37460   2.614  0.00895 **
## Age         -0.03689    0.01493  -2.471  0.01346 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120.86  on 87  degrees of freedom
## Residual deviance: 114.02  on 86  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 118.02
##
## Number of Fisher Scoring iterations: 4
```

3.

The logistic curve from the above model and the smoothing spline curve are plottd below against the data. It looks like the smoothing spline is overfitting and the logistic curve with only an age term is just a diagonal line which is not a great fit. However the logistic curve with a quadratic term is better than the others since it shows some curving towards where the data are concentrated in outcome.



4.

Here are teh interpretation of the coefficient:

- For males, coefficient is not significant, suggesting there is no significant difference between sex in terms of survival rate.

- For StatusHired people the coefficient suggests that they are significantly more likely to die than Family by a factor of exp(-1.625e+00)=0.197 times ($p < 0.05$).
- For StatusSingle coefficient is not significant. The result suggests that they are not significantly more likely to die than Family members.
- The coefficient upon age is exp(1.675e-01)=1.18 indicating age 1 is more likely to survive than newborn. Each additional year adds a factor of exp(-3.889e-03)=0.9961186 to the survival.

```
model<- glm(Outcome ~ Age + I(Age^2) + Sex +Status,
              data=donner,
              family=binomial(link="logit"))
summary(model)
```

```
##
## Call:
## glm(formula = Outcome ~ Age + I(Age^2) + Sex + Status, family = binomial(link = "logit"),
##     data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0431  -1.0391   0.5120   0.8664   2.0797
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.986e-01  6.172e-01   0.322   0.7476
## Age           1.675e-01  7.107e-02   2.357   0.0184 *
## I(Age^2)     -3.889e-03  1.525e-03  -2.550   0.0108 *
## SexMale      -6.637e-01  5.588e-01  -1.188   0.2349
## StatusHired  -1.625e+00  7.481e-01  -2.173   0.0298 *
## StatusSingle -1.852e+01  1.760e+03  -0.011   0.9916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120.855  on 87  degrees of freedom
## Residual deviance:  92.363  on 82  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 104.36
##
## Number of Fisher Scoring iterations: 16
```

The test of the deviance of each term suggests that out of the terms selected, only sex is not significant at $p < 0.05$. The other terms add significantly to the model ($p < 0.01$).

```
drop1(model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## Outcome ~ Age + I(Age^2) + Sex + Status
##          Df Deviance    AIC     LRT Pr(>Chi)
## <none>        92.363 104.36
## Age       1   99.278 109.28  6.9153 0.008546 **
## I(Age^2)  1  102.968 112.97 10.6049 0.001128 **
## Sex       1   93.798 103.80  1.4350 0.230942
## Status    2  103.940 111.94 11.5769 0.003063 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5

We set out to understand the survival rates of a historic group of migrants from the 19th century, the Donner party, which ventured into extreme weather. The group members had three types of status: family members, hired workers, amd single individuals. Out of these, we found that more than half of males and about 30% of females died along the way. To understand what factors affected survival, we first used a logistic model to predict their outcome based upon sex and age individually. However, neither of these variables individually provided a satisfactory fit to the data. Instead we found that adding a term quadratic in age helped provide a better fit perhaps because age is very important to survival. Therefore in our final logistic regression model, we used age and its quadratic term as well as sex and status. We found that when sex and status remain constant, those with hired worker status had a decreased survival probability relative to family members by a factor of 0.2 times ($p < 0.05$). Additionally if we fix the other variables, then for each increase in age in years, we get a corresponding decrease in survival that depends upon the age of comparison. Whereas a 1-year old is about 0.18 times more likely to survive than a newborn, from each year forward the survival rate is decreased by about one percent ($p < 0.05$).