STA 138 Take Home Midterm

Madeline Ye

11/14/2016

9.7  
A. Loglinear Model Selection  
##First data frame where S = Seat belt, yes or no  
##E = whether the victim was ejected yes or no  
##I = degree of injury, yes =fatal and no = nonfatal

##   
## Call:  
## glm(formula = count ~ S + E + I + S:E + S:I + E:I, family = poisson,   
## data = seat.data)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7   
## 0.20704 -0.10095 -0.01071 0.01731 -1.59987 0.30951 0.31400   
## 8   
## -0.21583   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.92251 0.03110 222.56 <2e-16 \*\*\*  
## SYes -0.75682 0.05394 -14.03 <2e-16 \*\*\*  
## EYes -0.72784 0.05345 -13.62 <2e-16 \*\*\*  
## IYes 5.04362 0.03120 161.65 <2e-16 \*\*\*  
## SYes:EYes -2.39964 0.03334 -71.97 <2e-16 \*\*\*  
## SYes:IYes 1.71732 0.05402 31.79 <2e-16 \*\*\*  
## EYes:IYes -2.79779 0.05526 -50.63 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1.6249e+06 on 7 degrees of freedom  
## Residual deviance: 2.8540e+00 on 1 degrees of freedom  
## AIC: 93.853  
##   
## Number of Fisher Scoring iterations: 3

## 1 2 3 4 5   
## 1098.13193 4630.86807 411117.86807 157335.13193 20.86807   
## 6 7 8   
## 490.13193 476.13193 1014.86807

## Analysis of Deviance Table  
##   
## Model 1: count ~ S + E + I + S:I + E:I  
## Model 2: count ~ S + E + I + S:E + S:I + E:I  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 2 7134.0   
## 2 1 2.9 1 7131.1 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## models gsq p.val df AIC  
## 1 Sat. model 0.000000 1.00000000 8 92.99875  
## 2 SE,SI,EI 2.854022 0.09114565 7 93.85277  
## 3 SI, EI 7133.977725 0.00000000 6 7222.97648  
## 4 SE, I 11444.375350 0.00000000 5 3654.72175  
## 5 S,E,I 11444.375350 0.00000000 4 11529.37410

From the results we take a look at the p-values in addition to the AIC values. With these in mind we find that the second model (SE, SI, EI) is the best fit. Although the saturated model has the smallest AIC value, I chose the second model because it has the smallest p-value and the difference between between the AIC values of the first and second model is minimal. We conclude that there is homogeneity amongst the three variables, meaning there is an conditional association for each pair of variables given the third variable.

B. Logistic Model Selection

##   
## Call: glm(formula = count ~ S + E + I + S:E + S:I + E:I, family = poisson,   
## data = seat.data)  
##   
## Coefficients:  
## (Intercept) SYes EYes IYes SYes:EYes   
## 6.9225 -0.7568 -0.7278 5.0436 -2.3996   
## SYes:IYes EYes:IYes   
## 1.7173 -2.7978   
##   
## Degrees of Freedom: 7 Total (i.e. Null); 1 Residual  
## Null Deviance: 1625000   
## Residual Deviance: 2.854 AIC: 93.85

The loglinear model (SE, SI, EI) is equivalent to the logit model with seat belt ad ejection as the explanatory variables and injury as the response variable. From the table we can pull the value of wearing a seatbelt and still being injured while the ejection variable is ommitted. the value is 1.7173 and we exponentiate this value. This tells us that that those not wearing a sealt belt are 5.57 times more likely to be in a fatal accident than those that do wear a seat belt.

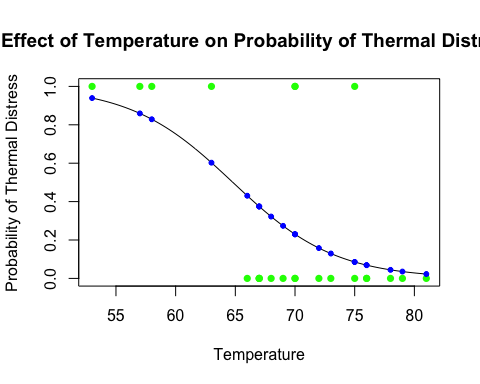
C. Index of Dissimilarity

## [1] 0.003287491

The index of dissimilarity measures how far the model fit falls from the data. The index falls between 0 and 1, with larger values representing a poorer fit. Our index value is 0.003287491 is close to 0 and far from 1 so we can conclude that our model selection does not fall far from fitting the data.

5.6  
A. Logistic Regression

##   
## Call:  
## glm(formula = distress ~ temp, family = binomial(link = logit),   
## data = rocket.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0611 -0.7613 -0.3783 0.4524 2.2175   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 15.0429 7.3786 2.039 0.0415 \*  
## temp -0.2322 0.1082 -2.145 0.0320 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28.267 on 22 degrees of freedom  
## Residual deviance: 20.315 on 21 degrees of freedom  
## AIC: 24.315  
##   
## Number of Fisher Scoring iterations: 5

 As we can observe from the plot, the higher the temperature, the less likely thermal distress is to occur at a 5% level of significance according to the model we have chosen. The relationship implies that the probability of thermal distress increases as temperature decreases.

1. Probability

The estimated probability of thermal distress at 31 degrees fahrenheit is 0.9996088. This is a large probability that is close to 1.

C. Confidence Interval

## Waiting for profiling to be done...

## OR 2.5 % 97.5 %  
## (Intercept) 3.412315e+06 27.9546841 8.214986e+14  
## temp 7.928171e-01 0.5972188 9.409919e-01

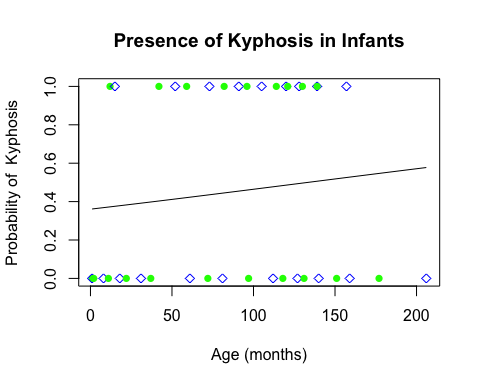
The confidence intervals shows that with each degree of temperature increase, the probability of thermal distress increases by the factor of 0.7928171. The confidence interval does not contain 1 which means that temperature has a significant effect on thermal distress.

5.8  
A. Logistic Regression Model

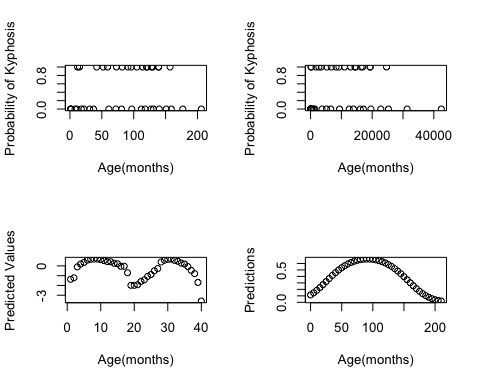
##   
## Call:  
## glm(formula = kyphosis ~ age, family = binomial(link = logit),   
## data = kyp.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3126 -1.0907 -0.9482 1.2170 1.4052   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -0.572693 0.602395 -0.951 0.342  
## age 0.004296 0.005849 0.734 0.463  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 55.051 on 39 degrees of freedom  
## Residual deviance: 54.504 on 38 degrees of freedom  
## AIC: 58.504  
##   
## Number of Fisher Scoring iterations: 4

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: kyphosis  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)  
## NULL 39 55.051   
## age 1 0.54689 38 54.504 0.4596

By the summary and a Chi square test, we see that age does not have a significant effect on the presence of kyphosis in infants.

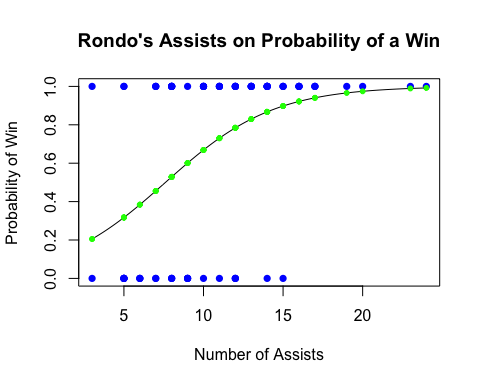
B. Plot  
 From the plot of the model we used in part A we find that it may not be the best fit. We try an exponential fit instead.

##   
## Call:  
## glm(formula = kyphosis ~ age + age.sq, family = binomial(link = logit),   
## data = kypdata.2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.482 -1.009 -0.507 1.012 1.788   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.0462547 0.9943478 -2.058 0.0396 \*  
## age 0.0600398 0.0267808 2.242 0.0250 \*  
## age.sq -0.0003279 0.0001564 -2.097 0.0360 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 55.051 on 39 degrees of freedom  
## Residual deviance: 48.228 on 37 degrees of freedom  
## AIC: 54.228  
##   
## Number of Fisher Scoring iterations: 4

 Once we add the age squared variable to the model, the same conclusion is reached that age does not have a significant effect on whether kyphosis is present or not. From the plots we see that new plot with squaring the slope follows more of a normal distribution. This suggest that predictions may be more accurate with the altered model.

5.3

##   
## Call:  
## glm(formula = wins ~ assists, family = binomial(link = logit),   
## data = rondo)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1351 -0.8737 0.4647 0.7928 1.7797   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.23527 0.83183 -2.687 0.007206 \*\*   
## assists 0.29378 0.08325 3.529 0.000417 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 97.073 on 76 degrees of freedom  
## Residual deviance: 79.619 on 75 degrees of freedom  
## AIC: 83.619  
##   
## Number of Fisher Scoring iterations: 5



## (Intercept)   
## 0.205227

## (Intercept)   
## 0.9919608

## Waiting for profiling to be done...

## OR 2.5 % 97.5 %  
## (Intercept) 0.1069636 0.01851579 0.4995626  
## assists 1.3414829 1.15506467 1.6057242

When x = 3, the probability of a win is 0.205227 and for x=24, the probability rises to 0.9919608. From the confidence interval and the plot we see a positive correlation between the the number of assists Rajon Rondo has in a game and the probability of winning said game. Since the confidence interval does not contain 1 we can say there is evidence that the number of assists Rondo executes has significant effect on whether a game is won or not. This is expected though as assists lead to points and the team with the most points win. This could be said for any player in the NBA. A comparison of different players with high assists numbers may be useful in this case.