hw6

Mohammad Mottahedi

February 21, 2016

Problem 1

9pts Consider the saturated model for a 2 by 2 table under multinomial sampling. There are two variables, X and Y , each taking values 0 or 1. The saturated model has three degrees of freedom. Inside it is the independence model with its two degees of freedom. Suppose we represent these degrees of freedom using three parameters: α, the log odds that Y = 1 ignoring X, γ, the log odds that X = 1 ignoring γ , and δ, the log odds ratio.

(a)

Explain how any independence model (a multinomial distribution (p11, p12, p21, p22) such that p11p22/p12p21= 1) can be represented by some choice of α, γ but fixing $\delta = 0$

$\alpha = log(p_{+1}/(1 - p_{+1}))$

$p_{+1} = e^{\alpha}/(1 + e^{\alpha})$

$p_{+2} = 1/(1 + e^{\alpha})$

$\gamma = log(p_{1+}/(1 - p_{1+}))$

$p_{1+} = e^{\gamma}/(1 + e^{\gamma})$

$p_{2+} = 1/(1 + e^{\gamma})$

$p11 = e^{\alpha}/(1 + e^{\alpha})xe^{\gamma}/(1 + e^{\gamma})$

$p12 = 1/(1 + e^{\gamma})xe^{\gamma}/(1 + e^{\gamma})$

$p21 = 1/(1 + e^{\gamma})xe^{\alpha}/(1 + e^{\alpha})$

$p22 = 1/(1 + e^{\gamma})x1/(1 + e^{\alpha})$

$p_{11}p_{22}/p_{12}p_{21}$

The two parts in above equation represent the conditional probabilities in each cell in a 2 by 2 table under multinomial sampling.

(b)

Show that any multinomial distribution (p11, p12, p21, p22) can be represented by some choice of α, γ, δ (now δ must be allowed to be nonzero).

saturated model: $log(\mu_{ij}) = \lambda_i^{\alpha} + \lambda_j^{\gamma} + \lambda_j^{\delta}$

independece model: $log(\mu_{ij}) = \lambda_i^{\alpha} + \lambda_j^{\gamma}$

(c)

Explain how to get the functional form of logistic regression for Y predicted by X from this set-up (hint: compute P r(Y = 1|X = 1) as a function of α, γ, δ). Which of our parameters α, γ, δ correspond to the β_0 and β_1 in binary logistic regression with a single categorical predictor?

$\alpha = \beta_0$, $\delta = \beta_1$

$log(\pi/(1-\pi)) = \beta_0 + beta_1 X$

$Pr(Y = 1|X = 0) = \beta_0$

$Pr(Y = 1|X = 0) = 1/(1+e^{\gamma})xe^{\alpha}/(1+e^{\alpha})/1/(1+e^{\gamma})$

$Pr(Y = 1|X = 0) = e^{\alpha}/(1+e^{\alpha}) {-\!\!-\!\!>} \alpha = \beta_0$

$Pr(Y = 1|X = 1) = \beta_0 + \beta_1 = \alpha + \beta_1$

$OddRatio = \frac{p11*p22}{p21*p12} = P(Y = 1|X = 1)/P(Y = 0|X = 1)] * [P(Y = 0|X = 0)]/P(Y = 1|X = 0)]$

$Pr(Y = 1|X = 0) - e^{\alpha+\beta_1}/(1+e^{\alpha+\beta_1})$

$Pr(Y = 0|X = 1) - 1/(1+e^{\alpha+\beta_1})$

$Pr(Y = 1|X = 0) - e^{\alpha}/(1+e^{\alpha})$

$Pr(Y = 0|X = 0)/Pr(Y = 1|X = 0) = e^{-\alpha} {-\!\!-\!\!>} OddsRatio = e^{\alpha+\beta_1} * e^{-\alpha} = e^{\beta_1}$

$p(x) = e^{\alpha+\delta x}/(1+e^{\alpha+\delta x}), \beta_0 = \alpha, \beta_1 = \delta$

Problem 2

9pts Let Y1, Y2, . . . , YN be independent random variables. For each of the following models, state whether it is or is not a generalized linear model. If it is a GLM, then what is the link function and the systematic component. If it is not a GLM, explain why is it not a GLM. The basic intro to GLMs was given in Lesson 5.

check part a

(a)

No, the link function is not linear because of the log term.

(b)

No, the relationship between transformed response and independent variables are not linear. $log(Y_i) = \alpha + log(\beta_1)x_{1i} + \beta_2 log(x_{2i})$

(c)

No, the systematic component are not linear in independent variables. $log\frac{\pi_i}{1-\pi_i} = \alpha + \beta e^{x_i}$

Problem 3

Children of ages 0-15 were classified according to whether they carried Streptococcus pyogenes and according to the size of their tonsils; see homework 4. Let π_i be the probability of a child being a carrier of Streptococcus and Xi tonsil size.

(a)

Fit the logistic regression model. Write the model in terms of π_i , β's, and X0s. Make sure you are modeling the ”event” of ”being a carrier”. How many dummy variables do you have? How many parameters does your saturated model have? What did you choose as you baseline(reference) level for the explanatory variable.

```r
tonsil.size <- as.factor(c("normal", "Senlarged", "Venlarged"))
response <- cbind(c(19, 29, 24), c(497, 560, 269))

carrier.logistic <- glm(response ~ tonsil.size, family = binomial(link=logit))
```

equation:

$\hat{pi}_i = \frac{exp(\beta_0+\beta_1X_1+\beta_2X_2)}{1+exp(\beta_0+\beta_1X_1+\beta_2X_2)}$

Current model has 2 dummy variables.

The model has 3 parameters $\beta_0, \beta_1, \beta_2$.

The baseline in chosen to be normal tonsil size.

(b)

Report the estimated model, and more specifically report the estimated coefficients and standard errors. Interpret exp (β 1) and report 95% confidence interval for exp (β 1). How about β 2?

```r
carrier.logistic
```

```
##
## Call:  glm(formula = response ~ tonsil.size, family = binomial(link = logit))
##
## Coefficients:
##          (Intercept)  tonsil.sizeSenlarged  tonsil.sizeVenlarged
##              -3.2642                0.3035                0.8475
##
## Degrees of Freedom: 2 Total (i.e. Null);  0 Residual
## Null Deviance:       7.321
## Residual Deviance: 1.146e-13     AIC: 20.85
```

```r
1-pchisq(carrier.logistic$null.deviance,carrier.logistic$df.null)
```

```
## [1] 0.02572057
```

$\hat{\pi}_i = \frac{exp(-3.2642+0.3035X_1+0.8475X_2)}{1+exp(-3.2642+0.3035X_1+0.8475X_2)}$

$exp(\beta_1) = \frac{oddsofslightlyenlargedtonsil}{oddsofnormaltonsil}$

$exp(\beta_2) = \frac{oddsofveryenlargedtonsil}{oddsofnormaltonsil}$

Cofidence interval:

$\beta_1 : (-0.28744, 0.89444)$

$\beta_2 : (0.227552, 0.227552)$

(c)

Does the model appear to fit the data? Perform an overall goodness-of-fit test and comment on the results. Is this test trustworthy?

```
summary(carrier.logistic)
```

```
## 
## Call:
## glm(formula = response ~ tonsil.size, family = binomial(link = logit))
## 
## Deviance Residuals: 
## [1]  0  0  0
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)           -3.2642     0.2338 -13.964  < 2e-16 ***
## tonsil.sizeSenlarged   0.3035     0.3015   1.007  0.31412    
## tonsil.sizeVenlarged   0.8475     0.3163   2.680  0.00737 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 7.3209e+00  on 2  degrees of freedom
## Residual deviance: 1.1458e-13  on 0  degrees of freedom
## AIC: 20.851
## 
## Number of Fisher Scoring iterations: 3
```

```
1-pchisq(carrier.logistic$null.deviance,carrier.logistic$df.null)
```

```
## [1] 0.02572057
```

```
#anova(carrier.logistic)
```

To check the overal fit of the model we have to look at null model deviance and residual deviance (saturated model). The residual deviance for the staturated model is almost zero, this means the saturated model is a good fit(as expected). The null deviance is equal to 7.32 with p-value 0.0257 which is significant which indicated the null model is not a good fit. The goodness of fit statistics is valid when the cell count are sufficiently large and expceted values are ≥ 5 and $n_i - \hat{\mu}_i \geq 5$ at leat 80% of the times.

Also,the hypothesis testing for indidual parametres whos that the we can't reject the null hypothesis that β_1 is equal to zero (p-value $= 0.314$).

(d)

How do results from this model compare to your conclusions from homework 4?

Both showed that the being a carrier and tosil size have a strong relationship. also they both confirm that the risk being a carrier is much higher when tonsil is very large.

(e)

What is the predicted probability of a child being a carrier if she has ”very enlarged” tonsils? P(A child being a carrier | having very enlarged tonsils)

$P(carrier|verylargetonsil) = \frac{exp(-3.2642+0.8475))}{1+exp(-3.2642+0.8475)} = 0.0819081$

Problem 4

$\frac{spouse}{others} = \frac{\frac{spouse}{alone}}{\frac{others}{alone}} = \frac{2.02}{1.71} = 1.1812865$

$\frac{25k+}{10k-25k} = \frac{\frac{25k+}{<10k}}{\frac{10k-25k}{<10k}} = \frac{0.41}{0.72} = 0.5694444$

Problem 5

```r
setwd("~/Dropbox/spring2016/discretDataAnalysis/hw_solutions/hw6")
montana <- read.csv("montana.csv", header = T, na.strings = ".")
#montana <- na.omit(montana)
attach(montana)
require(ResourceSelection)
```

```
## Loading required package: ResourceSelection
```

```
## ResourceSelection 0.2-6   2016-02-15
```

a

```r
inc_pol <- xtabs(data=montana, ~Sex+Stat)
inc_pol
```

```
##    Stat
## Sex  0  1
##   0 58 41
##   1 60 22
```

```r
stat <- as.factor(c(0,1))
response <- cbind(c(58, 60), c(41,22))
stat.sex <- glm(response ~ stat, family = binomial(link=logit))
summary(stat.sex)
```

```
##
## Call:
## glm(formula = response ~ stat, family = binomial(link = logit))
##
## Deviance Residuals:
## [1]  0  0
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3469     0.2040   1.700   0.0891 .
## stat1         0.6564     0.3221   2.038   0.0416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.2552  on 1  degrees of freedom
## Residual deviance: 0.0000  on 0  degrees of freedom
## AIC: 13.647
##
## Number of Fisher Scoring iterations: 3
```

the null model deviance is large (p-value=.039), the null model is not a good fit. The current model residual deviance is zero and the model fits well.

The Women are more likely to have postiive outlook.

b

```r
attach(montana)
```

```
## The following objects are masked from montana (pos = 4):
##
##     Age, Area, Fin, Inc, Pol, Sex, Stat
```

```r
montana$Age = as.factor(montana$Age) #2 factors
montana$Sex = as.factor(montana$Sex)
montana$Inc = as.factor(montana$Inc)
montana$Pol = as.factor(montana$Pol)
montana$Area = as.factor(montana$Area) # 3 factors
montana$Fin = as.factor(montana$Fin) #3 factors
montana$Stat = as.factor(montana$Stat)
```

```r
stat.logistic <- glm(Stat ~ ., data=montana, family=binomial(link=logit))
summary(stat.logistic)
```

```
##
## Call:
## glm(formula = Stat ~ ., family = binomial(link = logit), data = montana)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8745  -0.8653  -0.5599   1.0215   2.1188
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.31447    0.59427  -0.529   0.5967
## Age2        -1.08428    0.47706  -2.273   0.0230 *
## Age3        -0.71978    0.47204  -1.525   0.1273
## Sex1        -0.81694    0.39572  -2.064   0.0390 *
## Inc2         0.07748    0.48402   0.160   0.8728
## Inc3        -0.09860    0.52731  -0.187   0.8517
## Pol2        -0.34867    0.53569  -0.651   0.5151
## Pol3         0.50937    0.42507   1.198   0.2308
## Area2       -0.86827    0.49285  -1.762   0.0781 .
## Area3        0.48308    0.42938   1.125   0.2606
## Fin2         0.51792    0.47059   1.101   0.2711
## Fin3         0.98811    0.46494   2.125   0.0336 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 211.01  on 162  degrees of freedom
## Residual deviance: 182.39  on 151  degrees of freedom
##   (46 observations deleted due to missingness)
## AIC: 206.39
##
## Number of Fisher Scoring iterations: 4
```

```
#Hoslem.test

hoslem.test(stat.logistic$y, stat.logistic$fitted.values)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  stat.logistic$y, stat.logistic$fitted.values
## X-squared = 11.553, df = 8, p-value = 0.1723
```

The pearson χ^2 value is 11.5 with df=8 (p-value > 0.05) we can't reject the null hypothesis and the model fits well. pia # c

```
mont <- na.omit(montana)
stat.logistic2 <- glm(Stat ~ ., data=mont, family=binomial(link=logit))
step(stat.logistic2, direction = "backward")
```

```
## Start:  AIC=206.39
## Stat ~ Age + Sex + Inc + Pol + Area + Fin
##
##        Df Deviance    AIC
## - Inc   2   182.55 202.55
## - Pol   2   185.37 205.37
## <none>      182.39 206.39
## - Fin   2   187.05 207.05
## - Age   2   188.01 208.01
## - Sex   1   186.82 208.82
## - Area  2   191.17 211.17
##
```

```
## Step:  AIC=202.55
## Stat ~ Age + Sex + Pol + Area + Fin
##
##        Df Deviance    AIC
## - Pol   2   185.43 201.43
## <none>       182.55 202.55
## - Fin   2   187.29 203.29
## - Age   2   188.74 204.74
## - Sex   1   186.88 204.88
## - Area  2   191.58 207.58
##
## Step:  AIC=201.43
## Stat ~ Age + Sex + Area + Fin
##
##        Df Deviance    AIC
## <none>       185.43 201.43
## - Fin   2   192.47 204.47
## - Age   2   192.47 204.47
## - Sex   1   190.74 204.74
## - Area  2   193.71 205.71
```

```
##
## Call:  glm(formula = Stat ~ Age + Sex + Area + Fin, family = binomial(link = logit),
##     data = mont)
##
## Coefficients:
## (Intercept)         Age2         Age3         Sex1        Area2
##     -0.2221      -1.1684      -0.7221      -0.8568      -0.7361
##       Area3         Fin2         Fin3
##      0.5487       0.4376       1.1481
##
## Degrees of Freedom: 162 Total (i.e. Null);  155 Residual
## Null Deviance:       211
## Residual Deviance: 185.4     AIC: 201.4
```

The variable income is not in the step 1 while it is inside step 0.

variables age, area, sex, and fin are in the final model. the residual deviance in this smaller model is not much larger than saturated model in part (b).

6

a

full model -> df = 0

$log(\pi/(1-\pi)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_3 + \beta_6 X_1 X_4 + \beta_7 X_2 X_3 + \beta_8 X_2 X_4$

b

$log(\pi/(1-\pi)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

df = 6

c

$log(\pi/(1-\pi)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

df = 4