Department of Statistics

COURSE STATS 330

Model answers for Assignment 5 2008

1. *Ignoring the variable age, group the data into groups corresponding to each age-group/sex/class combination. Compute the logits for each combination as we did for the plum tree data. Make a data frame containing the logits, and the categorical variables. You should have one line in the data frame for each combination of the factor levels. Show the R code used.[8 marks]*

Hints: the function **`tapply`** will be useful to calculate the r and n values, and hence the logits. This will result in a 3-dimensional array which you can turn into a vector with the function **`as.vector`**. The function **`expand.grid`** is useful for generating all possible combinations of factor levels.

The following code will do the trick

```
# read data frame

titanic.df = read.table(file.choose(), header=T)

# compute r and n, and the n logits

r = tapply(titanic.df$survived, list(titanic.df$age.group,
titanic.df$pclass,titanic.df$sex), sum)

n = tapply(titanic.df$survived, list(titanic.df$age.group,
titanic.df$pclass,titanic.df$sex), length)


logits = as.vector(log((r+0.5)/(n-r+0.5)))

# make data frame

titanic.logits.df = data.frame(logits,
expand.grid(age.group=levels(titanic.df$age.group),pclass =
levels(titanic.df$pclass),
sex = levels(titanic.df$sex)))
```

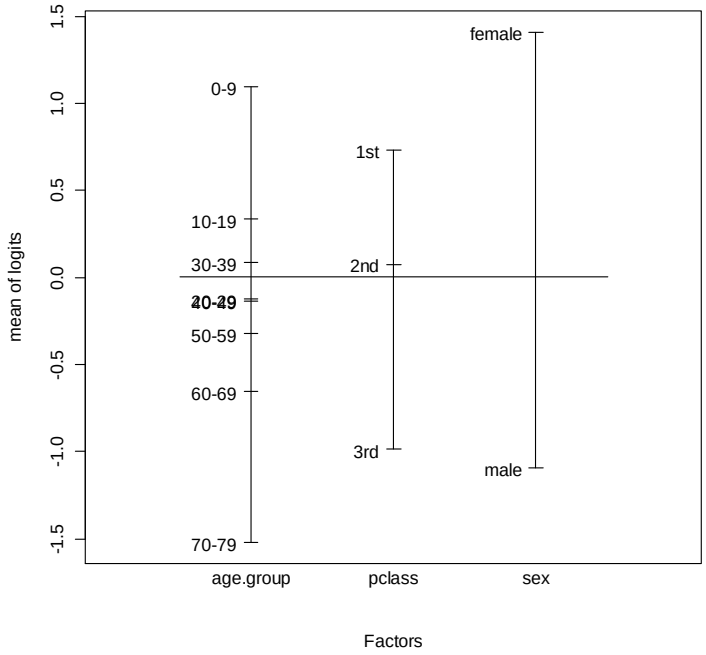
2. *Draw suitable plots and use them to assess graphically how the factors `class, age and sex` are related to the logits. Report your conclusions.[8 marks]*

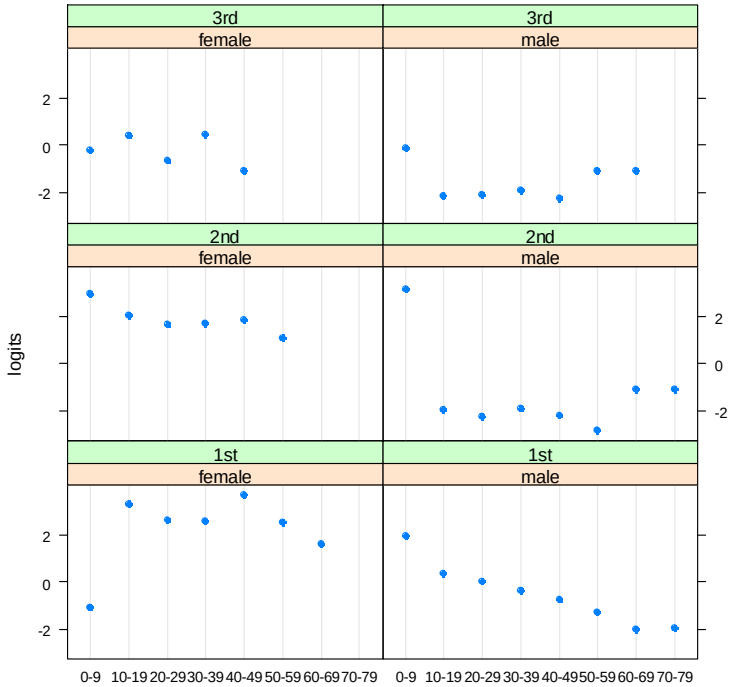
We will draw a plot with the function `plot.design` to assess how the individual factors affect the logits, and then a trellis plot to assess how the three factors together affect the logits. The R code is

```
plot.design( titanic.logits.df, y=logits)

library(lattice)
dotplot(logits~age.group|sex*pclass, data=titanic.logits.df )
```

This produces the plots





We can draw the following conclusions:

- Survival was higher for females, upper class passengers, and younger passengers.
- The relationship between survival and age is reasonably linear, with survival decreasing with age for all class/sex groups except for second and third class males.
- Children under ten survived better than other ages except for first class, but we can discount this as there were only 4 first-class children in the data set.

3. *Fit a suitable model to the data, using the ages rather than the age groups (i.e. using the original data frame). Are there significant interactions between the explanatory variables? [8 marks]*

We fit the model and do an anova:

```
> full.model = glm(survived~sex*pclass*age, data=titanic.df,
family=binomial)
> anova(full.model, test="Chisq")
Analysis of Deviance Table

               Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          632     869.54
sex             1   238.51       631     631.03 8.314e-54
pclass          2    55.78       629     575.25 7.706e-13
age             1    35.70       628     539.55 2.299e-09
sex:pclass      2    15.92       626     523.63 3.489e-04
sex:age         1    10.80       625     512.83 1.017e-03
pclass:age      2     7.69       623     505.14      0.02
sex:pclass:age  2     1.43       621     503.71      0.49
```

Looks like all interactions are required except the 3-factor. A stepwise regression (not shown) confirms this. However, the 0-9 group did not follow the pattern of the other ages. We could eliminate these and refit the model:

```
delete = titanic.df$age<10
full.model.10 = glm(survived~sex*pclass*age, data=titanic.df,
family=binomial, subset = -delete)
anova(full.model.10, test="Chisq")

               Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          631     867.92
sex             1   237.28       630     630.63 1.539e-53
pclass          2    55.55       628     575.08 8.661e-13
age             1    35.62       627     539.46 2.393e-09
sex:pclass      2    15.88       625     523.58 3.558e-04
sex:age         1    10.84       624     512.74 9.943e-04
pclass:age      2     7.72       622     505.02      0.02
sex:pclass:age  2     1.43       620     503.60      0.49
```

but this shows very little difference.

Our model is thus `survived~sex*pclass*age – sex:pclass:age`
Residual plots show no problems, and HL stat is OK. Call this the “final model”

4. *Estimate the probability that a female 1st class passenger aged 50 will survive. [8 marks]*

```
> new.data = data.frame(pclass="1st", sex="female", age=50)
> predict(final.model, new.data, type="response", se.fit=TRUE)
$fit
        1
0.9608722

$se.fit
         1
0.02153902

> 0.9608722+ 1.96*0.02153902*c(-1,1)
[1] 0.9186557 1.0030887
```

Thus, the confidence interval is (0.919, 1.000). This is slightly unpleasant since the upper CL is greater than 1.

Alternatively, we could get a CI for the log-odds and then apply the function exp(x)/(1+exp(x)) to each endpoint. This works as the function exp(x)/(1+exp(x)) is increasing.

```
> predict(final.model, new.data, se.fit=TRUE)
$fit
       1
3.201007

$se.fit
[1] 0.5728941

$residual.scale
[1] 1

> 3.201007 + 1.96*0.5728941*c(-1,1)
[1] 2.078135 4.323879

logist = function(x) exp(x)/(1+exp(x))

> logist(c(2.078135, 4.323879))
[1] 0.8887598 0.9869248
```

5. *What is the relationship between the three factors age-group, sex and class? Are any independent of the others? Fit a contingency table model for this part. (see Lecture 28) [8 marks]*

First, we make a suitable data frame, containing one line for each factor level combination, and a variable for the counts:

```
counts = table(titanic.df$age.group, titanic.df$pclass, titanic.df$sex)
```

```
poisson.df = data.frame(counts=as.vector(counts),
expand.grid(age.group=levels(titanic.df$age.group),
 pclass = levels(titanic.df$pclass), sex = levels(titanic.df$sex)))
```

Then, fit the Poisson regression and check for interactions:

```
poisson.glm=glm(counts ~ pclass*age.group*sex, family=poisson, data =
poisson.df)
anova(poisson.glm, test="Chisq")
```

```
                     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                    47     601.00
pclass                2     1.89        45     599.11      0.39
age.group             7   393.48        38     205.63 5.961e-81
sex                   1    32.80        37     172.84 1.023e-08
pclass:age.group     14   135.94        23      36.89 4.539e-22
pclass:sex            2    11.89        21      25.01 2.620e-03
age.group:sex         7    14.48        14      10.52      0.04
pclass:age.group:sex 14    10.52         0  5.021e-10      0.72
```

seems like the 3-factor interaction is zero, and possibly the age.group:sex as well. Let's compare the model without these interactions to the full model:

```
> anova(sub.df, poisson.glm, test="Chisq")
Analysis of Deviance Table


Model 1: counts ~ pclass * age.group + pclass * sex
Model 2: counts ~ pclass * age.group * sex
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        21    25.0052
2         0  5.021e-10 21  25.0052    0.2469
```

Seems like the model counts ~ pclass * age.group + pclass * sex is OK ie that age group and sex are independent given class.