228.371 Computer Lab: Logistic Regression

Semester One 2015 - Week 8

Instructions: Read each section and try the commands. Then try the Stream worksheet questions suggested to test your knowledge. The worksheet is "adaptive" which means if you get an answer wrong, you can try again. This quiz is to help you monitor your progress, it does not count toward your mark.

Note that because of fonts, especially for symbols like quotation marks, cutting and pasting commands from this document occasionally will not work - you may have to retype.

1 Simple Logistic Regression

Consider the data in `SurfaceToAir.csv`, which records the velocity of the target and whether the surface to air missle system being tested hit (1) or missed (0). If you call data dataset `surface` in `R`, you can fit a logistic regression with the command

```
surface <- read.table ("Data/SurfaceToAir.txt", header=TRUE)
m1 <- glm(y ~ TargetSpeed, data=surface, family="binomial")
summary(m1)
```

```
Call:
glm(formula = y ~ TargetSpeed, family = "binomial", data = surface)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0620  -0.4868   0.3915   0.5476   2.1682

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.070884   2.108996   2.879  0.00399 **
TargetSpeed -0.017705   0.006076  -2.914  0.00357 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.617  on 24  degrees of freedom
Residual deviance: 20.364  on 23  degrees of freedom
AIC: 24.364
```

```
Number of Fisher Scoring iterations: 4
```

You can find the probability of hitting the target at one of the target speeds used in the experiment by looking at the fitted values:

```
m1$fitted[surface$TargetSpeed==220]
```

```
        2        22 
0.8980598 0.8980598 
```

(the value appears twice because the speed appears twice in the data set). We can also compute this value by hand (there are differences in the final digits due to rounding error:

```
exp(6.070994-0.017705*220)/(1+exp(6.070994-0.017705*220))
```

```
[1] 0.8980638
```

The expression for $\mathbb{E}(\text{logit}\,(p_i))$ when $x_i = 220$ is 6.070994-0.017705*220. This is called the "linear predictor" and can be extracted with:

```
m1$linear[surface$TargetSpeed==220]
```

```
      2      22 
2.17585 2.17585 
```

As with linear regression, there is `predict` function that can be used to get either the linear predictor (using `type="link"`) or the probability (using `type="response"`) for values that are not in the data set. For example:

```
newdata=data.frame(TargetSpeed=c(205, 215, 225))
predict.glm (m1, newdata, type="response")
```

```
        1         2         3 
0.9199318 0.9058832 0.8896653 
```

```
predict.glm (m1, newdata, type="link")
```

```
       1        2        3 
2.441421 2.264374 2.087327 
```

Now answer worksheet questions 1-3.

2 ANOVA

The file `geeseag.txt` contains an experiment conducted at an oil drilling station to determine the effect of helicopter flights to service the station on the local geese. Different altitutes and latteral distances from the flock were tried, and whether or not there was a large flight response (1=large response). The glm model can be fit with

```
geese <- read.table ("Data/geeseag.txt", header=TRUE)
m2 <- glm (formula = cbind(Response, 5 - Response) ~ factor(LAT) *
              factor(ALT), family = "binomial", data = geese)
```

This can be compared to sub-models with

```
anova(m2, test="Chisq")
```

If a term is unnecesary, the reduction in residual deviance will be distributed chi-squared with the difference in degrees of freedom. For instance, for the interaction term, the reduction in deviance is 31.187-0 and the degrees of freedom are 27-0. In this case the p-value is 0.26, suggesting that the interaction term is unnecessary. The residuals from the additive model can be examined with:

```
m3 <- glm (formula = cbind(Response, 5 - Response) ~ factor(LAT) *
           factor(ALT), family = "binomial", data = geese)
plot (m3)
```

Now answer worksheet questions 4-6.

Read in the data `groundwater.csv`. The response is 1 if a detectable level of methyl tert-butyl (a potential contaminant from gasoline) is present, 0 otherwise. The potential predictors are well class (public or private), pH, Depth, Disolved Oxygen, Distance from the nearest potential source, and Aquifer type (bedrock or unconsolidated). Use the step function to select a model from these predictors:

```
groundwater <- read.csv ("Data/groundwater.csv", header=TRUE)
m4 <- glm(Response~1,data=groundwater,family="binomial")
step(m4, scope=~pH+DissOxy+WellClass+Aquifier+Depth+Distance)
```

Now answer worksheet questions 7-8.