R Notebook

Hide

```
rm(list=ls(all=TRUE))

library(MASS)
summary(bacteria)
```

```
 y       ap      hilo       week            ID            trt
 n: 43   a:124   hi:122   Min.   : 0.000   X03    : 5   placebo:96
 y:177   p: 96   lo: 98   1st Qu.: 2.000   X04    : 5   drug   :62
                          Median : 4.000   X05    : 5   drug+  :62
                          Mean   : 4.455   X07    : 5
                          3rd Qu.: 6.000   X08    : 5
                          Max.   :11.000   X09    : 5
                                           (Other):190
```

Hide

```
?bacteria
show(bacteria)
```

	y <fctr>	ap <fctr>	hilo <fctr>	week <int>	ID <fctr>	trt <fctr>
1	y	p	hi	0	X01	placebo
2	y	p	hi	2	X01	placebo
3	y	p	hi	4	X01	placebo
4	y	p	hi	11	X01	placebo
5	y	a	hi	0	X02	drug+
6	y	a	hi	2	X02	drug+
7	n	a	hi	6	X02	drug+
8	y	a	hi	11	X02	drug+
9	y	a	lo	0	X03	drug
10	y	a	lo	2	X03	drug

1-10 of 220 rows   Previous **1** 2 3 4 5 6 … 22 Next

We first bring up the appropriate library and a summary of the dataset that we will be using in this analysis. In this case, we pose our research question as the following: Does the use of drugs eradicate bacteria that causes otitis media? Our null hypothesis, or the status quo, is that: The use of drugs DO NOT have an effect on the bacteria that causes otirtis media. On the other hand, the alternative hypothesis is that: The use of drugs DO have an effect on the bacteria that causes otirtis media. This is the idea that we are trying to achieve.

Hide

```
wk11 <- subset(bacteria, week == 11)
bacteria$yn <- 1
bacteria$yn[which(bacteria$y == "y")] <- 1
bacteria$yn[which(bacteria$y == "n")] <- 0
show(wk11)
```

	y <fctr>	ap <fctr>	hilo <fctr>	week <int>	ID <fctr>	trt <fctr>	yn <dbl>
4	y	p	hi	11	X01	placebo	1
8	y	a	hi	11	X02	drug+	1
13	y	a	lo	11	X03	drug	1
18	y	p	lo	11	X04	placebo	1
23	y	p	lo	11	X05	placebo	1
27	y	a	lo	11	X06	drug	1
32	y	a	hi	11	X07	drug+	1
37	y	p	hi	11	X08	placebo	1
42	y	p	lo	11	X09	placebo	1
49	y	p	hi	11	X11	placebo	1

1-10 of 44 rows   Previous **1** 2 3 4 5 Next

In the above portion, we filter out the appropriate subset from week 11 so that we can determine whether the drug had any effect on the bacteria at the end of the study. We then create a new response variable, "yn", that identifies the presence or absence of bacteria with 1 and 0, respectively, as opposed to the factors y and n.

Hide

```
bacteria.model <- glm(yn ~ trt, data = wk11, family = binomial)
show(bacteria.model)
```

```
Call:  glm(formula = yn ~ trt, family = binomial, data = wk11)

Coefficients:
(Intercept)        trtdrug        trtdrug+
     1.3863        -0.6931         -0.6931

Degrees of Freedom: 43 Total (i.e. Null);   41 Residual
Null Deviance:        51.56
Residual Deviance: 50.57      AIC: 56.57
```

Hide

```
summary(bacteria.model)
```

```
Call:
glm(formula = yn ~ trt, family = binomial, data = wk11)

Deviance Residuals:
    Min        1Q    Median        3Q        Max
-1.7941   -1.4823    0.6681    0.9005    0.9005

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.3863     0.5590   2.480   0.0131 *
trtdrug       -0.6931     0.8292  -0.836   0.4032
trtdrug+      -0.6931     0.8292  -0.836   0.4032
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51.564  on 43  degrees of freedom
Residual deviance: 50.569  on 41  degrees of freedom
AIC: 56.569

Number of Fisher Scoring iterations: 4
```

We use the general linear model equation that plots the presence or absence of bacteria at week 11 (yn) against the treatment type (trt). The reason that we are only comparing it to trt is due to the fact that "trt" factors in both "ap" and "hilo", and adding those to our linear model would only complicate the model. The predicted values of the Intercept show a significant value, however, our null deviance and residual deviance show little difference and thus signifies that our model is NOT a good fit. With that said, we can reject our hypothesis and state that drug use has no effect on the presence of H. influenzae.

```
pred.est <- predict.glm(bacteria.model, wk11)
wk11$pred.value <- round(exp(pred.est) / (1 + exp(pred.est)))
accuracy <- table(wk11$pred.value, wk11$yn)
accuracy
```

```
    0  1
  1 12 32
```

```
sum(diag(accuracy))/sum(accuracy)
```

```
[1] 0.2727273
```

We get an accuracy of 0.273, which signifies that the results are accurate only 27% of the time, which is highly unreliable.

```
pR2(bacteria.model)
```

```
         llh       llhNull          G2      McFadden          r2ML          r2CU
  -25.28438851 -25.78191521   0.99505339    0.01929751    0.02236105    0.03239681
```

Finally, McFaddens R2 value is nowhere near 1, further cementing the idea that this model is not a good fit.