Stat 8053, Fall 2013: Logistic Models (Faraway, Chap. 2)

The data for this example is in the package `alr4`. If you downloaded before August 30, you need to download it again to reproduce this handout.

```
install.packages("alr4")
```

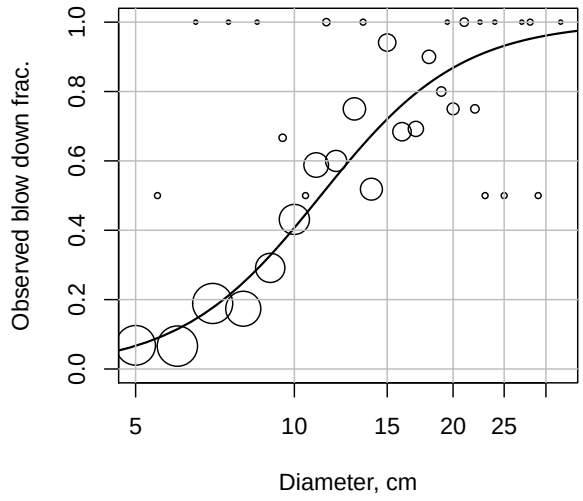
In July 1999 a very large storm devastated the tree population of the Boundary Waters Canoe Area Wilderness (BWCAW) in Northern Minnesota. This first example concern one species of trees called black spruce.

```
library(alr4) # also loads 'car' and 'effects' packages
some(BlowBS) # from the 'car' package required by 'alr4'
```

```
      d died  m
5   7.0   17 90
9   9.0   14 48
10  9.5    2  3
15 12.0   15 25
18 14.0   14 27
19 15.0   16 17
21 17.0    9 13
25 20.0    6  8
28 22.5    1  1
30 24.0    1  1
```

Here `d` is the diameter of the tree in cm to the nearest 0.5 cm, `died` is the response y, the number of trees that died, and `m` is the number of trees sampled.

```
plot( I(died/m) ~ d, BlowBS, xlab="Diameter, cm", ylab="Observed blow down frac.",
           cex=.4*sqrt(m), log="x", ylim=c(0,1))
g1 <- glm(cbind(died, m-died) ~ log(d), data=BlowBS, family=binomial)
bs <- round(coef(g1), 4)
dnew <- seq(3, 40, length=100)
lines(dnew, predict(g1, newdata=data.frame(d=dnew), type="response"), lwd=1.5)
grid(col="gray", lty="solid")
```




```
summary(g1)

Call:
glm(formula = cbind(died, m - died) ~ log(d), family = binomial,
    data = BlowBS)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.898  -0.810   0.353   1.135   2.330

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.892      0.633   -12.5   <2e-16
log(d)         3.264      0.276    11.8   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 250.856  on 34  degrees of freedom
Residual deviance:  49.891  on 33  degrees of freedom
```

```
AIC: 117.5

Number of Fisher Scoring iterations: 4
```

```
Anova(g1)
```

```
Analysis of Deviance Table (Type II tests)

Response: cbind(died, m - died)
       LR Chisq Df Pr(>Chisq)
log(d)      201  1     <2e-16
```

Coefficients

Let $\mathbf{x}$ be the observed predictors and $\hat{\boldsymbol{\beta}}$ the vector of estimates including the intercept, so the fitted log odds are $\mathbf{x}'\hat{\boldsymbol{\beta}}$. Suppose $\mathbf{x}_1$ differs from $\mathbf{x}$ by increasing x_1 to $x_1 + \delta$. Then

$$\begin{aligned}
\log\left(\frac{p(\mathbf{x}_1)}{1-p(\mathbf{x}_1)}\right) &= \mathbf{x}_1'\hat{\boldsymbol{\beta}} \\
\frac{p(\mathbf{x}_1)}{1-p(\mathbf{x}_1)} &= \exp(\mathbf{x}_1'\hat{\boldsymbol{\beta}}) \\
&= \exp(\hat{\beta}_0 + \hat{\beta}_1(x_1+\delta) + \cdots \hat{\beta}_p x_p) \\
&= \exp(\hat{\beta}_1\delta)\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots \hat{\beta}_p x_p) \\
&= \exp(\hat{\beta}_1\delta)\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right)
\end{aligned}$$

Example: If `d` is increased by 10%, then $\log(1.1\texttt{d}) = \log(\texttt{d}) + \log(1.1) = \log(\texttt{d}) + \delta = \approx \log(\texttt{d}) + 0.095$, and so the odds of blowdown are multiplied by $\exp(\hat{\beta}_1 \times 0.095) = 1.36$. Since $\log(1.25) \approx 0.223$, a 25% increase in `d` corresponds to multiplying the odds of blowdown by $\exp(\log(1.25)\hat{\beta}_1) = 2.07$. To get a confidence interval, exponentiate the end-points of an interval for β_1:

```
exp(log(1.1) * confint(g1)[2, ])
```

```
2.5 % 97.5 %
1.298  1.440
```

Per Tree

```
str(Blowdown)

'data.frame':        3666 obs. of  4 variables:
 $ d  : num  9 14 18 23 9 16 10 5 6 9 ...
 $ s  : num  0.0218 0.0218 0.0218 0.0218 0.0218 ...
 $ y  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ spp: Factor w/ 9 levels "aspen","balsam fir",..: 2 2 2 2 2 2 7 2 2 2 ...
```

This data file uses each tree as a unit of analysis, rather than grouped by diameter:

```
summary(g2 <- glm(y ~ log(d), data=Blowdown, family=binomial, subset=spp=="black spruce"))

Call:
glm(formula = y ~ log(d), family = binomial, data = Blowdown,
    subset = spp == "black spruce")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.507  -0.757  -0.494   0.810   2.327

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.892      0.633   -12.5   <2e-16
log(d)         3.264      0.276    11.8   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 856.21  on 658  degrees of freedom
Residual deviance: 655.24  on 657  degrees of freedom
AIC: 659.2

Number of Fisher Scoring iterations: 4

compareCoefs(g1, g2) # from 'car'

Call:
1:"glm(formula = cbind(died, m - died) ~ log(d), family = binomial, data = BlowBS)"
```

```
2:"glm(formula = y ~ log(d), family = binomial, data = Blowdown, subset = spp == \"black spruce\")"
            Est. 1   SE 1 Est. 2   SE 2
(Intercept) -7.892  0.633 -7.892  0.633
log(d)       3.264  0.276  3.264  0.276
```

```
Anova(g2) # also from 'car'
```

```
Analysis of Deviance Table (Type II tests)

Response: y
       LR Chisq Df Pr(>Chisq)
log(d)      201  1     <2e-16
```

A bigger regression

```
summary(g3 <- glm(y ~ log(d) + s + spp, data=Blowdown, family=binomial))
```

```
Call:
glm(formula = y ~ log(d) + s + spp, family = binomial, data = Blowdown)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.751  -0.681  -0.224   0.671   3.022

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.997195   0.374841  -16.00  < 2e-16
log(d)           1.581342   0.111460   14.19  < 2e-16
s                4.628886   0.212845   21.75  < 2e-16
sppbalsam fir   -2.242787   0.493577   -4.54  5.5e-06
sppblack spruce  0.000228   0.178933    0.00     1.00
sppcedar         0.167226   0.151751    1.10     0.27
sppjackpine     -2.076512   0.216234   -9.60  < 2e-16
spppaper birch   1.039965   0.178763    5.82  6.0e-09
sppred pine     -1.723568   0.186462   -9.24  < 2e-16
sppred maple    -1.795674   0.301934   -5.95  2.7e-09
sppblack ash     0.003138   0.413172    0.01     0.99

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 5057.9  on 3665  degrees of freedom
Residual deviance: 3259.3  on 3655  degrees of freedom
AIC: 3281

Number of Fisher Scoring iterations: 5
```

This model has an intercept for each of the 9 species, given in logit scale by

```
round(ints <- coef(g3)[1] + c(0, coef(g3)[4:11]), 3)
```

```
                  sppbalsam fir sppblack spruce        sppcedar     sppjackpine  spppaper birch
         -5.997          -8.240          -5.997          -5.830          -8.074          -4.957
    sppred pine    sppred maple    sppblack ash
         -7.721          -7.793          -5.994
```

but a common slope for `d` and for `s`.

Effects

A **typical fitted value** for each species can be obtained by getting fitted values for each species with `s` and `d` fixed at a typical value, by default their means:

```
(new <- with(Blowdown, data.frame(spp=levels(spp), d=exp(mean(log(d))), s=mean(s))))
```

```
           spp     d      s
1        aspen 13.83 0.4116
2   balsam fir 13.83 0.4116
3 black spruce 13.83 0.4116
4        cedar 13.83 0.4116
5     jackpine 13.83 0.4116
6  paper birch 13.83 0.4116
7     red pine 13.83 0.4116
8    red maple 13.83 0.4116
9    black ash 13.83 0.4116
```

```
predict(g3, newdata=new)
```

```
       1        2        3        4        5        6        7        8        9
 0.06160 -2.18119  0.06182  0.22882 -2.01492  1.10156 -1.66197 -1.73408  0.06473
```

```
predict(g3, newdata=new, type="response")

     1      2      3      4      5      6      7      8      9
0.5154 0.1015 0.5155 0.5570 0.1176 0.7506 0.1595 0.1501 0.5162

or <- order(ints)
Blowdown$spp <- with(Blowdown, factor(spp, levels=levels(spp)[or]))
g3 <- update(g3)
(e3 <- Effect("spp", g3)) # from the 'effects' package

 spp effect
spp
  balsam fir     jackpine    red maple     red pine        aspen black spruce    black ash
      0.1015       0.1176       0.1501       0.1595       0.5154       0.5155       0.5162
       cedar  paper birch
      0.5570       0.7506

plot(e3, rotx=45, grid=TRUE)
```