4/8/2010 36-402/608 ADA-II H. Seltman
Breakout #21: Logistic Regression Comments

In R, logistic regression is performed using
`result = glm(y ~ x..., data=my.dtf, family=binomial)`
where "y" is 0 or 1, and "x..." is any prediction formula. **Warning:** If you forget to specify the family, it defaults to "normal" and you get the very wrong lm() analysis.

As usual `summary(result)` has the standard errors and (Wald) Z and p-values, as well as AIC `$aic`.

The `glm()` object has a `$deviance` component that can be used for the likelihood ratio test. E.g., to compare glm objects named "full" and "reduced" use:

```
p.val = 1 - pchisq(reduced$deviance - full$deviance, reduced$df.res - full$df.res)
```

If you prefer probit regression to logistic regression use `family=binomial(link="probit")`.

For binomial (instead of Bernoulli) outcomes use `cbind(successCount,FailureCount)` instead of `y` in the formula.

Use `family=quasibinomial` to check for under/over dispersion.

The analysis shown here is the famous British moth predation data, often cited as an example of observable natural selection due to sooty pollution making moths that are resting on (normally) light colored trees easier to spot by predators. It comes from The Sleuth, chapter 21.

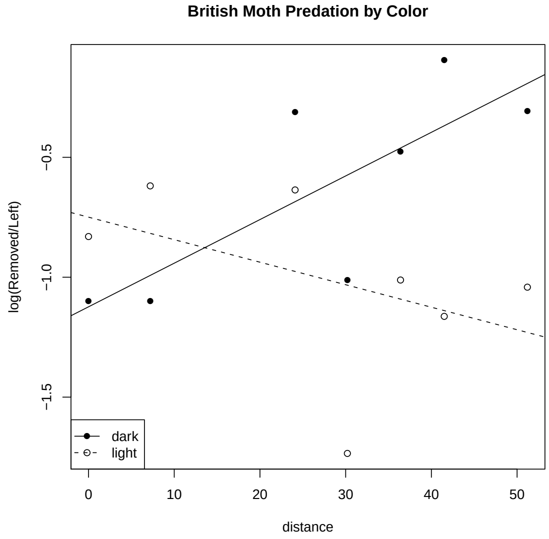
The outcome is expressed in the two data columns "placed" (total number of moths on a tree) and "removed" (number killed by predators). The explanatory variables are "distance" from the city center (in km.; used as a surrogate for level of pollution) and whether the moth is light or dark colored ("morph"). The study comprises 968 moths in 7 locations.

```
moths=read.csv("case2102.csv")
names(moths)=casefold(names(moths))
dim(moths) # 14 4
moths[1:3,]
#  morph distance placed removed
#1 light      0.0     56      17
#2  dark      0.0     56      14
#3 light      7.2     80      28
```

```
summary(moths)
#  morph      distance        placed         removed
# dark :7   Min.   : 0.00   Min.   :52.00   Min.   : 9.00
# light:7   1st Qu.:11.42   1st Qu.:57.00   1st Qu.:16.25
#           Median :30.20   Median :60.00   Median :20.00
#           Mean   :27.23   Mean   :69.14   Mean   :21.86
#           3rd Qu.:40.23   3rd Qu.:83.00   3rd Qu.:23.75
#           Max.   :51.20   Max.   :92.00   Max.   :40.00

# Create ‘‘failure’’ variable
moths$left = moths$placed - moths$removed

# EDA plot
with(moths, plot(distance, log(removed/left), ylab="log(Removed/Left)",
     pch=1+15*(morph=="dark"), main="British Moth Predation by Color"))
with(moths[moths$morph=="dark",], abline(lm(log(removed/left)~distance)))
with(moths[moths$morph=="light",], abline(lm(log(removed/left)~distance), lty=2))
legend("bottomleft", c("dark","light"), lty=1:2, pch=c(16,1))
```



Question 1: What model is suggested by the EDA?

It looks like the slope of log odds of predation vs. distance differs for light and dark moths, so we need a logistic regression model that uses distance, morphology, and their interaction.

```
noia = glm(cbind(removed,left)~morph+distance,moths,family="binomial")
summary(noia)
```

```
# Coefficients: Estimate Std. Error z value Pr(>|z|)
# (Intercept)  -0.732690   0.151221  -4.845 1.27e-06 ***
# morphlight   -0.404052   0.139377  -2.899  0.00374 **
# distance      0.005314   0.004002   1.328  0.18422
# (Dispersion parameter for binomial family taken to be 1)
#     Null deviance: 35.385  on 13  degrees of freedom
# Residual deviance: 25.161  on 11  degrees of freedom
# AIC: 93.836

exp(noia$coef)
# (Intercept)  morphlight    distance
#   0.4806144   0.6676093   1.0053278
```

Question 2: Write out the prediction model on the log odds scale. Then write out the prediction model on the odds scale. Finally, write out and simplify the prediction equation for the ratio of the odds of removal for light colored moths at distance "d" to the odds for removal for dark colored moths at distance "d". What do you conclude?

I am using "R" for removed (success in this model), and LO() as the estimated log odds, and O() for estimated odds, "L" for light, "D" for dark, and "d" for distance.

$$\mathrm{LO}(R) = b_0 + b_L L + b_d d$$

$$O(R) = \exp(b_0 + b_L L + b_d d)$$

$$O(R_{Ld})/O(R_{Dd}) = \frac{\exp(b_0 + b_L + b_d d)}{\exp(b_0 + b_d d)} = \exp(b_0 + b_L + b_d d - (b_0 + b_d d)) = \exp(b_L)$$

```
full = glm(cbind(removed,left)˜morph*distance,moths,family="binomial")
summary(full)
# Coefficients:        Estimate Std. Error z value Pr(>|z|)
# (Intercept)         -1.128987   0.197906  -5.705 1.17e-08 ***
# morphlight           0.411257   0.274490   1.498 0.134066
# distance             0.018502   0.005645   3.277 0.001048 **
# morphlight:distance -0.027789   0.008085  -3.437 0.000588 ***
# (Dispersion parameter for binomial family taken to be 1)
#     Null deviance: 35.385  on 13  degrees of freedom
# Residual deviance: 13.230  on 10  degrees of freedom
# AIC: 83.904
# Number of Fisher Scoring iterations: 4

exp(full$coef)
```

```
#  (Intercept)  morphlight   distance morphlight:distance
#    0.3233608   1.5087132  1.0186745           0.9725935
```

Question 3: Use the notation "L0" for light moths at distance "d", "L1" for light moths at distance "d+1", and similarly "D0" and "D1" for dark moths. Write out and simplify the prediction equation for: $\frac{\textbf{odds}(L1)/\textbf{odds}(D1)}{\textbf{odds}(L0)/\textbf{odds}(D0)}$ What do you conclude?

$$\text{LO}(R) = b_0 + b_L L + b_d d + b_{Ld} L d$$

$$O(R) = \exp(b_0 + b_L L + b_d d + b_{Ld} L d)$$

$$O(R_{L0})/O(R_{D0}) = \frac{\exp(b_0 + b_L + b_d d + b_{Ld} d)}{\exp(b_0 + b_d d)} = \exp(b_0 + b_L + b_d d + b_{Ld} d - (b_0 + b_d d)) = \exp(b_L + b_{Ld} d)$$

$$\begin{aligned} O(R_{L1})/O(R_{D1}) &= \frac{\exp(b_0 + b_L + b_d(d+1) + b_{Ld}(d+1))}{\exp(b_0 + b_d(d+1))} \\ &= \exp(b_0 + b_L + b_d(d+1) + b_{Ld}(d+1) - (b_0 + b_d(d+1))) \\ &= \exp(b_L + b_{Ld}(d+1)) \end{aligned}$$

$$\frac{O(R_{L1})/O(R_{D1})}{O(R_{L0})/O(R_{D0})} = \exp(b_L + b_{Ld}(d+1)) - (b_L + b_{Ld} d)) = \exp(b_{Ld})$$

```
# CI for interaction effect
tmp = summary(full)$coef[4,]
tmp
#      Estimate    Std. Error       z value      Pr(>|z|)
# -0.0277890438  0.0080854798 -3.4369072050  0.0005883972
LOCI = tmp[1] + c(-1,1)*1.96*tmp[2]
round(LOCI,3) #  -0.044 -0.012
LOCI10 = 10*tmp[1] + c(-1,1)*1.96*10*tmp[2]
round(exp(10*tmp[1]),3) # 0.757
round(exp(LOCI10),3) # 0.646 0.887
```

Question 4: Explain the meaning of 0.757 and its CI. The odds ratio for removal (predation) for light vs. dark moths drops by (1-0.757) = 24.3% for each additional kilometer distant from the city. We are 95% confident that the true drop is 11.3 to 35.4%.

Here is the likelihood ratio test for the interaction:

```
cat(noia$df.residual, full$df.residual, "\n") # 11 10
1-pchisq(noia$deviance-full$deviance,1) # 0.000552
```

Question 5: What do you conclude? (Additional Note: A similar test for the need for distance squared gave p=0.479.)

The interaction is definitely needed with p=0.000552 for the null hypothesis that the interaction coefficient is zero. (There is no evidence that we need the next most complicated model with a curvature in time.)

Here is the test for under/over-dispersion:

```
qfull = glm(cbind(removed,left)~morph*distance,moths,family="quasibinomial")
summary(qfull)
# Coefficients:       Estimate Std. Error t value Pr(>|t|)
# (Intercept)         -1.128987   0.223104  -5.060 0.000492 ***
# morphlight           0.411257   0.309439   1.329 0.213360
# distance             0.018502   0.006364   2.907 0.015637 *
# morphlight:distance -0.027789   0.009115  -3.049 0.012278 *
# (Dispersion parameter for quasibinomial family taken to be 1.270859)
#     Null deviance: 35.385  on 13  degrees of freedom
# Residual deviance: 13.230  on 10  degrees of freedom
# AIC: NA

1 - pchisq(summary(qfull)$dispersion * qfull$df.residual, qfull$df.residual)
# 0.2404243
```

Question 6: What do you conclude?

We estimate the variances to be 1.27 times bigger than expected, but this estimate is consistent with a dispersion of 1.0 (p=0.24 for the null hypothesis that the dispersion is 1), so we can use the original logistic regression results. If this p-value were small, we would use the dispersion adjusted estimates and p-values from the quasi-binomial analysis.