Key: Overdispersion and Quasibinomial Logistic Regression

Logistic regression – Stat 230

In this class, you learn about extra-binomial variation (a case of overdispersion) and how we can use quasibinomial logistic regression to adjust our inferences to handle this issue. This topic is not discussed in your book.

Data overview

Orobanche is a genus of parasitic plants without chlorophyll that grow on the roots of flowering plants. In an experiment, seeds from two varieties, Orobanche aegyptiaca 75 and Orobanche aegyptiaca 73, where brushed with extract from either a cucumber root or a bean root. Researchers recorded the number of seeds that eventually germinated.

```r
orobanche <- read.csv("http://aloy.rbind.io/data/orobanche.csv")
```

The variables in the data set are:

- `y`: count of seeds germinated
- `n`: number of seeds
- `variety`: either oa75 or oa73
- `root`: either cucumber or bean

Notice that both explanatory variables are categorical (i.e., factors).

Task 1. Does there appear to be a relationship between the proportion of seeds that germinate and variety? What about root? To answer this, I recommend looking at both a summary table and boxplots of the estimated proportion by each variable.

To calculate a summary of counts for each combination of levels, use the commands `ftable()` and `xtabs()`

```r
ftable(xtabs(cbind(y, n) ~ variety + root, data = orobanche))
```

To create a `prop` column in the data set, you can run the below chunk:

```r
orobanche$prop <- orobanche$y/ orobanche$n
```

Additive model

To begin, let’s fit an additive logistic regression model where `variety` and `root` are both included.

```
oro_glm1 <- glm(y/n ~ variety + root, data = orobanche, family = binomial, weights = n)
```

Task 2. Perform a deviance goodness-of-fit test. What do you conclude?

The deviance goodness-of-fit statistic is in the summary of the model:

```
summary(oro_glm1)
```

```

Call:
glm(formula = y/n ~ variety + root, family = binomial, data = orobanche,
    weights = n)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3919  -0.9949  -0.3744   0.9831   2.4766

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.7005     0.1507  -4.648 3.36e-06 ***
varietyoa75    0.2705     0.1547   1.748   0.0804 .
rootcucumber   1.0647     0.1442   7.383 1.55e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 39.686  on 18  degrees of freedom
AIC: 122.28

Number of Fisher Scoring iterations: 4
```

We find the statistic is $D^2 = 39.686$. Then we used a chi-squared distribution with df $= 18$. To find the p-value we run:

```
1 - pchisq(39.686, df = 18)
```

```
[1] 0.00230269
```

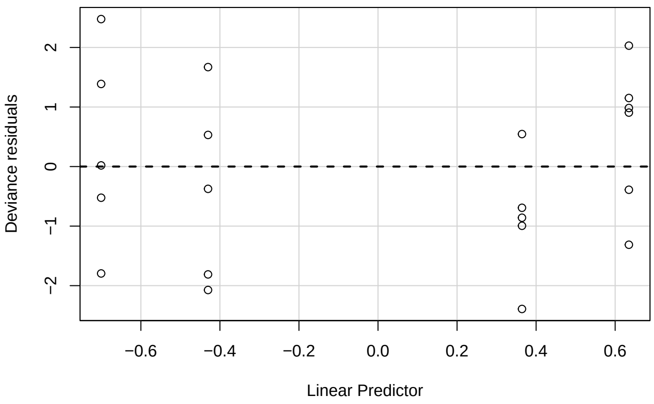
We have strong evidence that the model is not adequate ($D^2 = 39.686$, df $= 18$, p-value $= 0.002$).

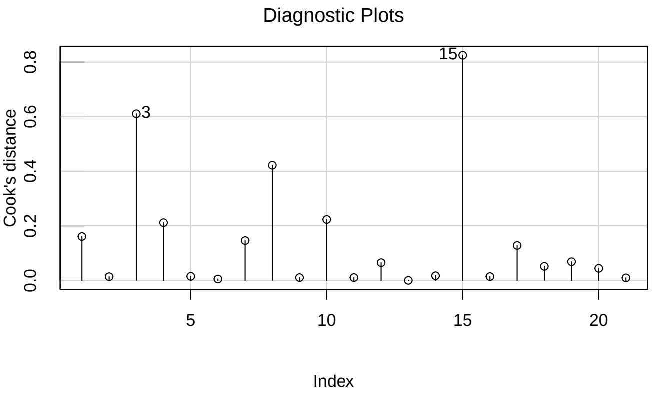
Task 3. Are there any concerning outliers? Check residual plots to investigate this.

There don’t appear to be any problematic outliers from the residual plot. Observation #15 is borderline influential based on Cook’s distance, but we’ll retain it for now.

```
residualPlot(oro_glm1, type = "deviance", smooth = FALSE)

influenceIndexPlot(oro_glm1, vars = "cook")
```





Adding an interaction term

You should have found evidence of lack of fit and that there were no concerning outliers, so we may be missing a term. The only other thing we can add here is an interaction. The below code chunk updates our first model to include an interaction term.

```
oro_glm2 <- update(oro_glm1, . ~ . + variety:root)
```

Task 4. Perform a deviance goodness-of-fit test. What do you conclude?

The deviance goodness-of-fit statistic is in the summary of the model:

```
summary(oro_glm2)
```

```
Call:
glm(formula = y/n ~ variety + root + variety:root, family = binomial,
    data = orobanche, weights = n)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01617  -1.24398   0.05995   0.84695   2.12123

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.4122     0.1842  -2.238   0.0252 *
varietyoa75               -0.1459     0.2232  -0.654   0.5132
rootcucumber               0.5401     0.2498   2.162   0.0306 *
varietyoa75:rootcucumber   0.7781     0.3064   2.539   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
AIC: 117.87

Number of Fisher Scoring iterations: 4
```

We find the statistic is $D^2 = 33.278$. Then we used a chi-squared distribution with df $= 17$. To find the p-value we run:

```
1 - pchisq(33.278, df = 18)
```

```
[1] 0.01544297
```

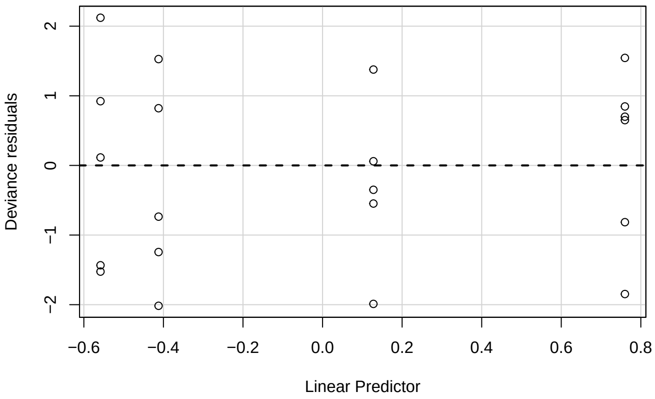
We have strong evidence that the model is not adequate ($D^2 = 39.686$, df $= 18$, p-value $= 0.015$).

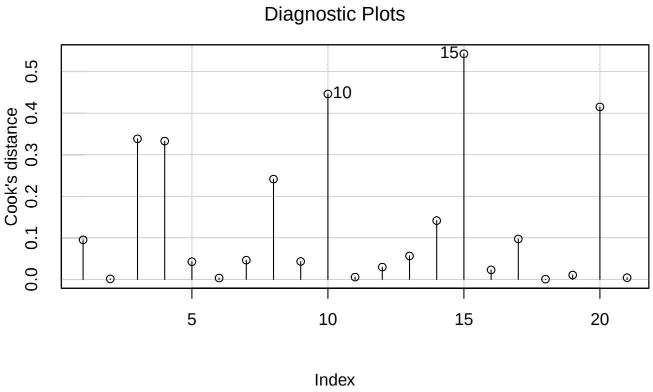
Task 5. Check the residual plots again to see if outliers are a problem.

We don't have any concerns about outliers here.

```
residualPlot(oro_glm2, type = "deviance", smooth = FALSE)

influenceIndexPlot(oro_glm2, vars = "cook")
```





Investigating over-dispersion

Since there aren't any more terms we can add and outliers aren't a problem, we suspect over-dispersion.

Task 6. Do the *ad hoc* calculation of the estimate of the dispersion parameter. Is it "much" larger than 1?

The *ad hoc* dispersion estimate is about two times larger than 1 (what we would expect).

$$\widehat{\psi} = \frac{\text{residual deviance}}{df} = \frac{33.278}{17} \approx 1.958$$

Task 7. The below code to fit the interaction model via quasilikelihood (i.e., this fits the quasibinomial model). How do the standard errors and test statistics change? Do any of the variables change in significance?

The standard errors are larger, which reduce the magnitude of the test statistics. This increases the p-values, and we see that `root` is no longer appears to be associated with the log odds (after controlling for the other variables), and now there is weak evidence that the interaction term is needed.

```r
oro_glm3 <- glm(y/n ~ variety * root, data = orobanche, family = quasibinomial,
                weights = n)
summary(oro_glm3)
```

```
Call:
glm(formula = y/n ~ variety * root, family = quasibinomial, data = orobanche,
    weights = n)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01617  -1.24398   0.05995   0.84695   2.12123

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.4122     0.2513  -1.640   0.1193
varietyoa75               -0.1459     0.3045  -0.479   0.6379
rootcucumber               0.5401     0.3409   1.584   0.1315
varietyoa75:rootcucumber   0.7781     0.4181   1.861   0.0801 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.861832)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Task 8. In the `summary()` output for `oro_glm3` there is a dispersion parameter given. Multiply the square root of this with the standard error of the interaction term in `oro_glm2`. Compare this to the standard error of the interaction term in `oro_glm3`. What do you notice?

$$SE_{\text{quasi}}(\widehat{\beta_3}) = \sqrt{\widehat{\psi}} SE(\widehat{\beta_3}) = \sqrt(1.861832) * 0.3064 \approx 0.4181$$

This is the SE given for the interaction term in `oro_glm3`. Thus, we can see that $\sqrt{\widehat{\psi}}$ is how much we need to increase the variability of our estimates.

Comparing models

Let's work with the quasibinomial version now.

Task 9. Determine if we can remove the `variety:root` interaction from the model. To do this, fit the reduced quasibinomial logistic regression model and then use the `anova()` command to compare the models, specifying `test = "F"`. (Remember that in the quasibinomial case, the drop in deviance test now uses the F distribution!)

There is weak evidence of an interaction effect between variety and root ($F = 6.4081$, df $= 1, 18$, p-value $= 0.081$).

```
oro_glm4 <- glm(y/n ~ variety + root, data = orobanche, family = quasibinomial,
                weights = n)

anova(oro_glm4, oro_glm3, test = "F")
```

```
Analysis of Deviance Table

Model 1: y/n ~ variety + root
Model 2: y/n ~ variety * root
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
1        18     39.686
2        17     33.278  1   6.4081 3.4418 0.08099 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```