This is a project to practice the concepts of Retrospective and Prospective models and Overdispersion

Hana Akbarnejad

2/24/2020

The Full code chunks and results for this assignment can be found at the end of this file

Problem 1

This problem investigates data deriven from a retrospective study that studies the relationship between alcohol consumption (alcohol concentration) and emergence of cancer. The study has been adjusted for age.

Note that this is retrospective data, so we cannot use it to calculate relative risk (RR), even with fitting prospective model. But we can treat the study as prospective (despite the fact that the data has been collected retrospectively) and treat disease status as response.

Fitting a prospective model to this data, we will have the following result:

```
# fitting a prospective model
logit_prosp = glm(response ~ cancer_data$alcohol_consump + cancer_data$age, family=binomial(link='logit
summary(logit_prosp)
```

```
##
## Call:
## glm(formula = response ~ cancer_data$alcohol_consump + cancer_data$age,
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.59974  -1.72957   0.06822   1.19015   1.50808
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -5.023449   0.418224 -12.011   <2e-16 ***
## cancer_data$alcohol_consump80+ g  1.780000   0.187086   9.514   <2e-16 ***
## cancer_data$age                   0.061579   0.007291   8.446   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 211.608  on 11  degrees of freedom
## Residual deviance:  31.932  on  9  degrees of freedom
## AIC: 78.259
##
## Number of Fisher Scoring iterations: 4
```

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 (alcohol) + \beta_2 (age)$$

We can observe that β_0 is -5.02 which is the the log odds of cancer in age 0 and alcohol consumption of less thab 80 grams (unexposed group).

the estimate for β_1 is 1.7799995, this means that the log odds ratio of developing cancer is 1.78 for the exposed group (the group with daily alcohol comsumption of equal to or more than 80 grams) compared to unexposed group (the group with daily alcohol comsumption of 0-79 grams), holding age constant.

Also, we can observe that the estimate for β_2 is 0.0615787, this means that the log odds ratio of developing cancer is 0.06 for each one unit increase in age, holding exposure (alcohol consumption status) constant.

These results show that age and alcohol comsumption are positively associated with this specific cancer.

Problem 2

Part 1

In this problem, we are going to fit a logistic regression model to study the relation between germination rates and different types of seed and root extract.

This is an example of prospective study with root and seed type as predictor and germination as response. Seed types are O. aegyptiaca 75 (coded as 0) or O. aegyptiaca 73(coded as 1) and root types are bean(coded as 0) and cucumber(coded as 1). So we have two predictors and each of them are binary. Our response is continuous.

Fitting a model to study the relation between germination rates and different types of seed and root extract and interpretation of results:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 (seed) + \beta_2 (medium)$$

```
germ_logit = glm(germ_rate ~ seed + root, family = binomial(link = 'logit'), data = germ_data)
summary(germ_logit)
```

```
##
## Call:
## glm(formula = germ_rate ~ seed + root, family = binomial(link = "logit"),
##     data = germ_data)
##
## 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)  -0.4300     0.1137  -3.781 0.000156 ***
## seed         -0.2705     0.1547  -1.748 0.080435 .
## root          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
```

The summary of logistic regression model shows that the log odds of germination is -0.43 for O. aegyptiaca 75 plants if grown in bean root extract media.

The log odds ratio of germination is -0.27 between seed type species O. aegyptiaca 75(0) versus O. aegyptiaca 73(1) seed type, holding the root media type constant. Note that the p-value for seed type coefficient is 0.08, because this value is greater than $\alpha = 0.05$, we can conclude that that seed type is an insignificant variable.

The log odds ratio for germination is 1.06 between bean root medium type compared to cucumber root extract medium, holding the seed type constant.

Part 2

In this part we are interested to check the model for possible over dispersion. To do so, we need to calculate Generalized Pearson χ^2 as follows:

$$G = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)\phi} \sim \chi^2(n-p)$$

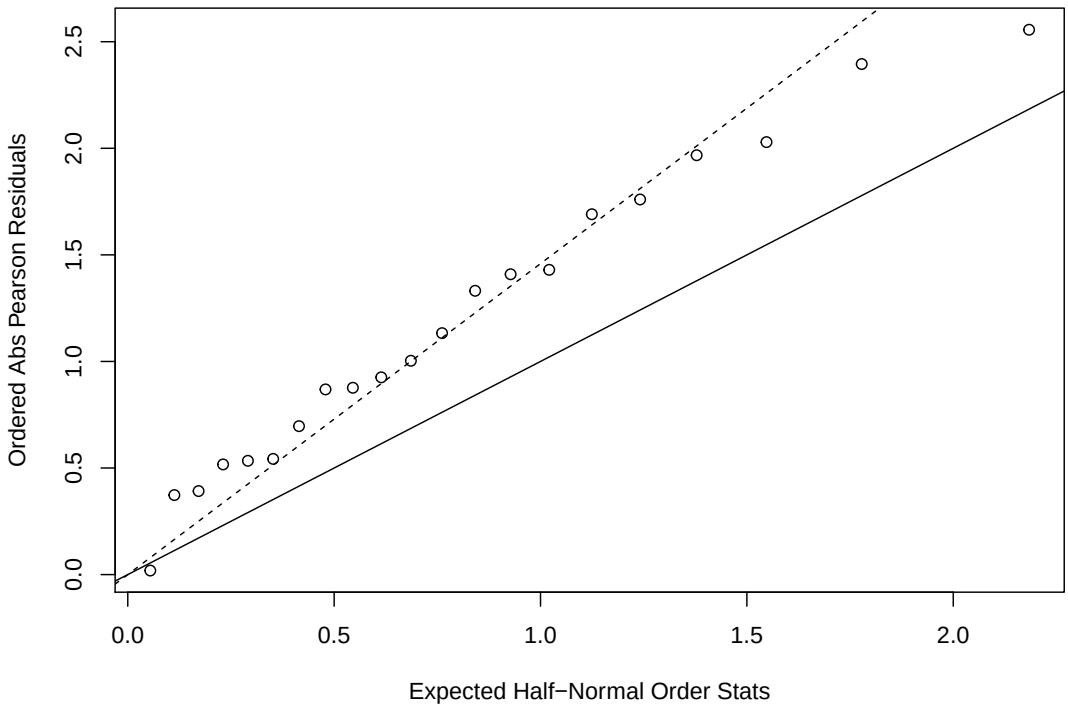
Where

$$G_0 = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

is the Original Pearson χ^2 statistics we grt from binomial distribution without dispersion. So, to estinate ϕ, we use the formula:

$\hat{\phi} = G_0/(n-p)$ or $\hat{\phi} = D_0/(n-p)$ where D_0 is the deviance of original model without overdispersion. The results are similar.

We can see that over dispersion parameter estimated from Generalized Pearson χ^2 is 2.13 which is greater than 1 and confirms that our model is over dispersed. To visualize and confirm this result, we use half-normal plot:



As the above half-normal plot shows, our model is over dispersed. So we need to update our model:

```
summary(germ_logit, dispersion = phi_hat_G)
```

```
##
## Call:
## glm(formula = germ_rate ~ seed + root, family = binomial(link = "logit"),
##     data = germ_data)
##
## 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)  -0.4300     0.1659  -2.592  0.00955 **
## seed         -0.2705     0.2257  -1.198  0.23081
## root          1.0647     0.2104   5.061 4.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 2.128368)
##
##     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 can see that after updating the model with dispersion, our standard error has been increased which is what we were expecting.

The summary of logistic regression model shows that the log odds of germination is -0.43 for O. aegyptiaca 75 plants if grown in bean root media, considering overdispersion.

The log odds ratio of germination is changed by -0.27 (reduces) when we go from seed type species O. aegyptiaca 75(0) to O. aegyptiaca 73(1), holding the root type constant. Note that the p-value for seed type coefficient is 0.23, because this value is greater than $\alpha = 0.05$, we can conclude that that seed type is still insignificant variable after updating the moel with dispersion.

The log odds ratio for germination is changed by 1.06 (increases) when we go from bean root compared to cucumber root, holding the seed type constant.

Part 3

The source of overdispersion might be Intra-class correlation, because the germination in plants from the same seed might be correlated in other characteristics and how they grow on different medium.

Also, we have clusters with different sizes that have different germination rates.