HW 6 P1 & P2

Amy Ly

2/9/2023

Problem 1 Context

An experiment was performed to test the effect of a toxic substance. 1500 insects were randomly assigned to 6 groups and each group were exposed to a fixed dose of the toxic substance. A day later, the number of dead insects (out of 250) was recorded.

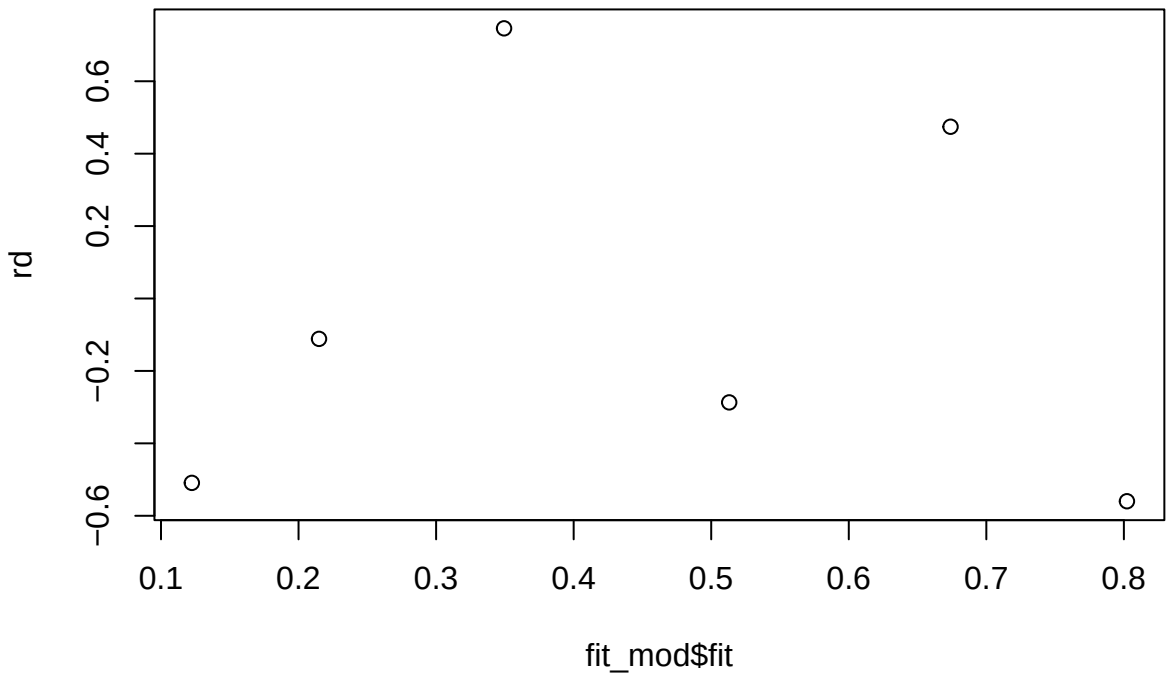
X denotes the dose level (on a log scale) received by the insects in each group.

Y denotes the number of insects that died (out of 250) in each group.

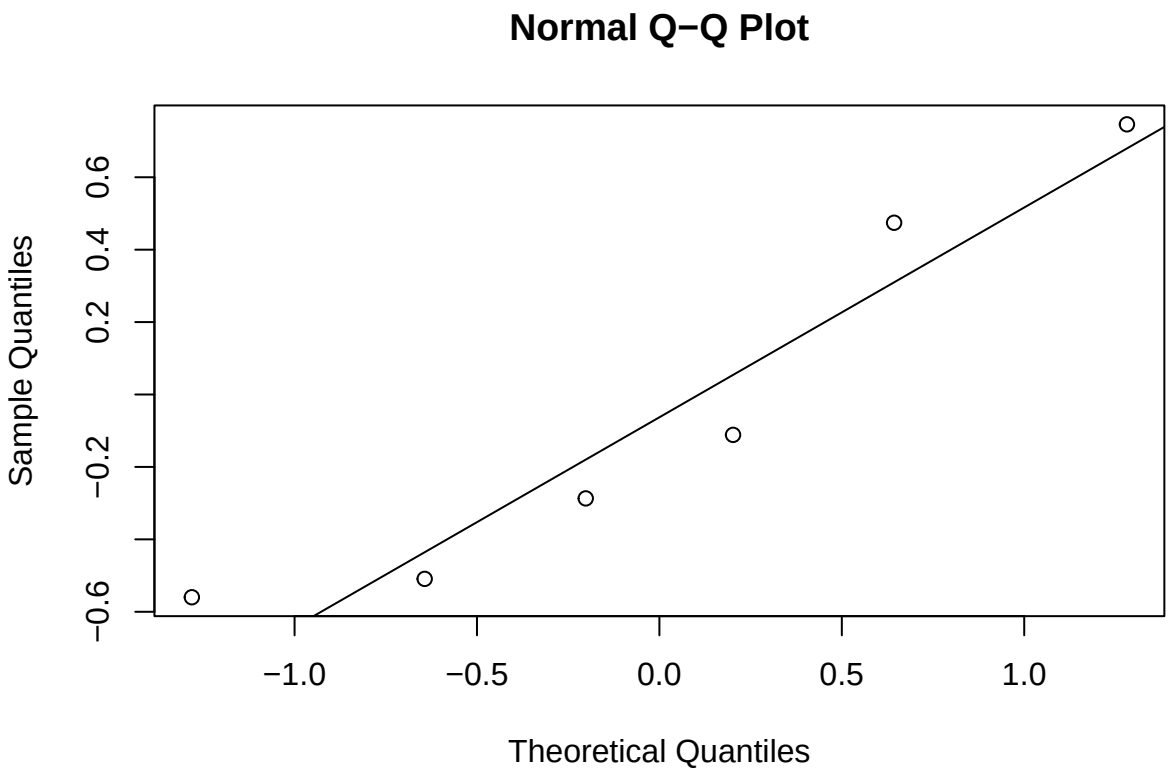
Part A

Fit a logistic regression model with X as the explanatory variable and famile =binomial.

```
df <- data.frame(X = c(1, 2, 3, 4, 5, 6),
                 dead = c(28, 53, 93, 126, 172, 197),
                 tot = c(250, 250, 250, 250, 250, 250)) %>%
  mutate(alive = tot-dead)

fit_mod <-  glm(cbind(dead, alive) ~ X, family=binomial, data=df)

rd <- resid(fit_mod, "deviance")
plot(fit_mod$fit,rd)
```




```
qqnorm(rd)
qqline(rd)
```



Part B

Conduct the deviance goodness-of-fit test to assess if the fitted model is adequate.

```
df2 <- df %>%
  mutate(X = as.factor(X))

sat <-  glm(cbind(dead, alive) ~ X, family=binomial, data=df2)

anova(fit_mod, sat, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(dead, alive) ~ X
## Model 2: cbind(dead, alive) ~ X
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         4     1.4491
## 2         0     0.0000  4   1.4491   0.8356
```

```
p.val <- anova(fit_mod, sat, test="Chisq")$Pr[2]

test.stat <- anova(fit_mod, sat, test="Chisq")$Deviance[2]
```

Based on the goodness of fit test, there is little evidence (D =1.449093 and p-value = 0.8356191) that the saturated model (the fitted model from part a) is a better fit for the data.

Part c

Is there evidence for overdispersion?

```
summary(fit_mod)
```

```
##
## Call:
## glm(formula = cbind(dead, alive) ~ X, family = binomial, data = df)
##
## Deviance Residuals:
##       1        2        3        4        5        6
## -0.5092  -0.1115   0.7461  -0.2869   0.4744  -0.5599
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.64367    0.15610  -16.93   <2e-16 ***
## X            0.67399    0.03911   17.23   <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: 383.0695  on 5  degrees of freedom
## Residual deviance:   1.4491  on 4  degrees of freedom
## AIC: 39.358
##
## Number of Fisher Scoring iterations: 3
```

Usually, to assess if there is overdispersion with count data, we would divide the residual deviance by the degrees of freedom. Here, that ratio would be $1.4491/4 = 0.362275$, which is much less than 1.

In this scenario, we do not have the issue of overdispersion. If the ratio was less than 0.01, then I would be concerned about underdispersion which may indicate that the model may be mispecified or that the link function is incorrect.

Part d

Fit an appropriate model based on your answer in part c. Conduct a drop-in-deviance test to assess the potential effect of dose on mortality rate.

Note that the drop-in deviance test (similar to the Extra sum of squares F-test in ordinary regression) compares the residual deviance between 2 nested models.

```
fit_mod2 <-  glm(cbind(dead, alive) ~ 1, family=binomial(link="log"), data=df)

anova(fit_mod2, fit_mod, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(dead, alive) ~ 1
## Model 2: cbind(dead, alive) ~ X
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
```

```
## 1         5     383.07
## 2         4       1.45  1   381.62 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is more evidence that the appropriate model is the more complex model, which is the fitted model from part a.

Part e

Using the model fitted in the previous part, interpret the estimated slope coefficient and give the 95% confidence interval.

```
a <- fit_mod$coef[1]
b <- fit_mod$coef[2]
V <- vcov(fit_mod)

CI <- c(fit_mod$coef[2]-1.96*sqrt(V[2,2]),
  fit_mod$coef[2]+1.96*sqrt(V[2,2]))
```

For every unit increase in log dose, there is a $\beta = 0.6739928$ rate of change of the log odds ratio for death of insects. The 95% confidence interval is [0.597339, 0.7506465]

Problem 2 Context

When farming practices are ecological, the number of butterflies observed on fields is expected to increase. To investigate this, several species of butterflies are observed and counted at a number of farms around Uppsala and Scania (in Sweden).

Covariates in the dataset are region (U for Uppsala, S for Scania) and years (number of years in ecological farming). There were 3 different species counted (Large Skipper, Pearly Heath, and Ringlet).

Problem

Fit a Poisson regression model for the counts of Large Skipper depending on the region and the number of years in ecological farming. Examine the residuals, residual deviance, and deviance goodness-of-fit test. Weigh the evidence for or against overdispersion. Build a final model and summarize your findings.

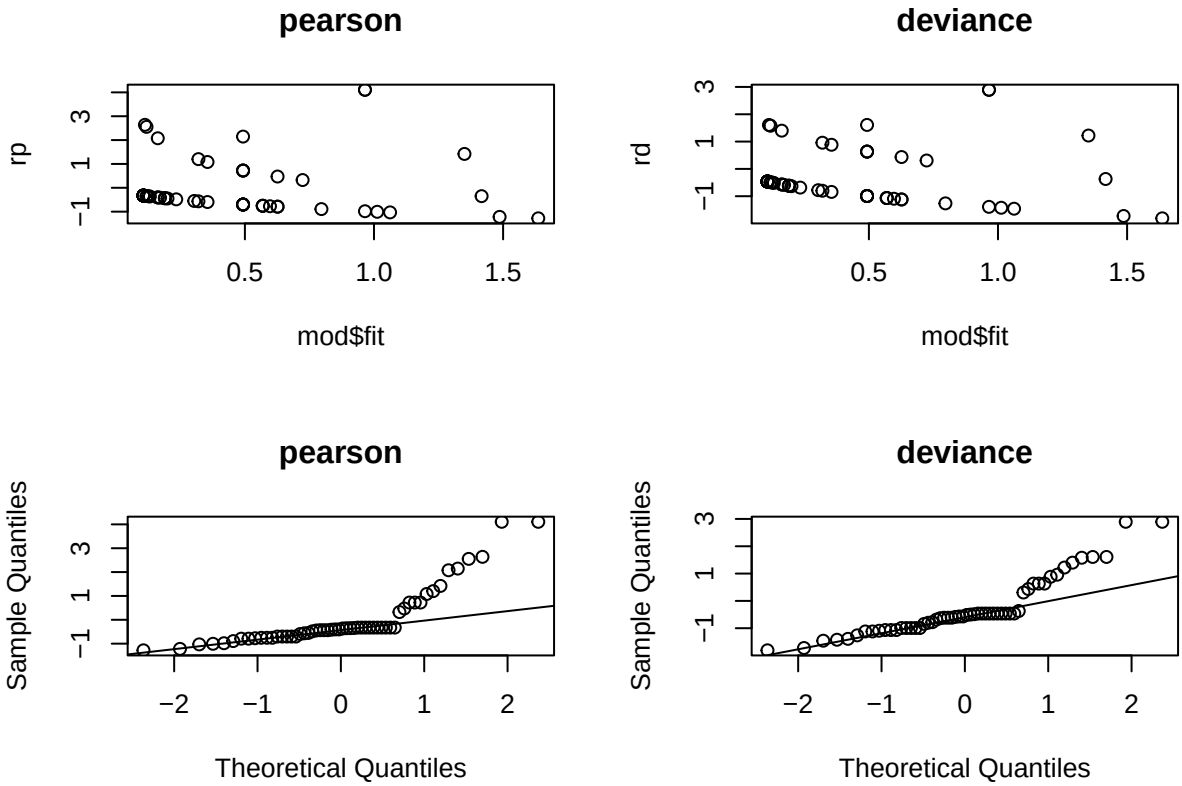
```
butterfly <- read.csv("butterflies.csv")

butterfly <- butterfly%>%
  select(Region, X.years, LargeSkipper)

#w where LargeSkipper species and S region are the reference categories
mod <- glm(LargeSkipper~Region*X.years, data=butterfly, family=poisson(link="log"))
mod$aic
```

```
## [1] 107.5876
```

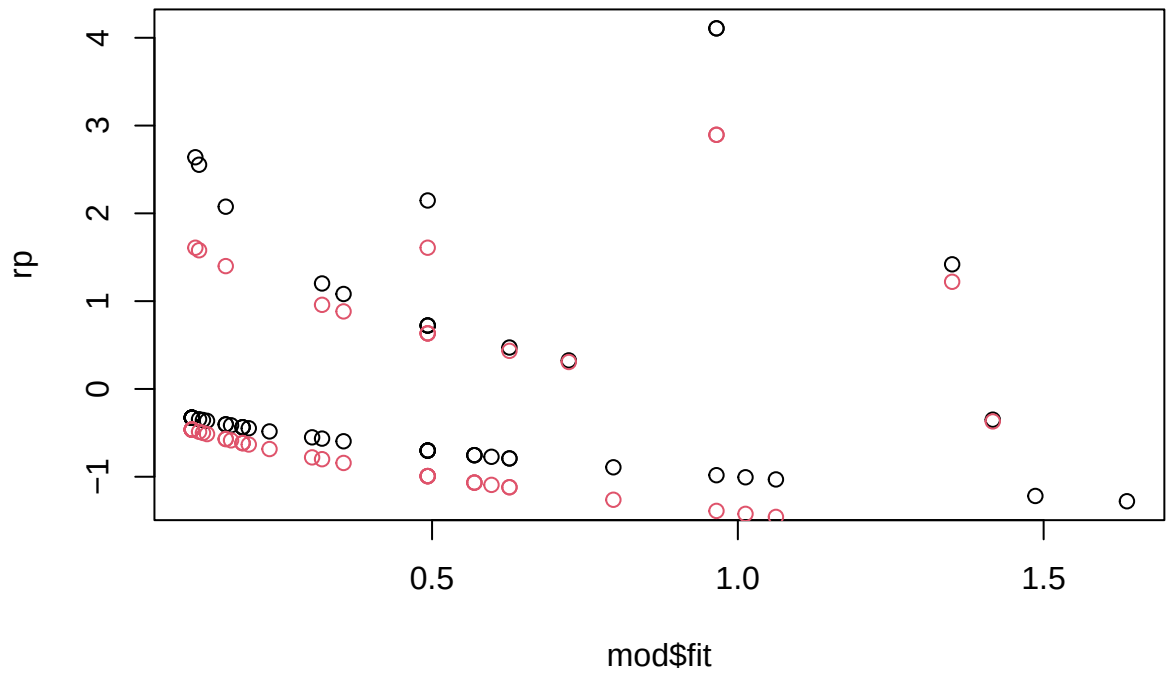
```
summary(mod)
```

```
##
## Call:
## glm(formula = LargeSkipper ~ Region * X.years, family = poisson(link = "log"),
##     data = butterfly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8090  -0.9928  -0.5422  -0.2006   2.8949
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.707518   0.353682  -2.000   0.0455 *
## RegionU         -1.525202   0.804930  -1.895   0.0581 .
## X.years          0.047996   0.024559   1.954   0.0507 .
## RegionU:X.years  0.004086   0.057194   0.071   0.9430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 81.450  on 55  degrees of freedom
## Residual deviance: 65.021  on 52  degrees of freedom
## AIC: 107.59
##
## Number of Fisher Scoring iterations: 6
```

```
rp <-resid(mod, "pearson")
rd <- resid(mod, "deviance")

par(mfrow=c(2,2))

plot(mod$fit,rp, main="pearson")
plot(mod$fit,rd,main="deviance")

qqnorm(rp, main="pearson")
qqline(rp)

qqnorm(rd,main="deviance")
qqline(rd)
```




```
par(mfrow=c(1,1))
plot(mod$fit,rp, col=1)
points(mod$fit,rd, col=2)
```




```
summary(mod)
```

```
##
## Call:
## glm(formula = LargeSkipper ~ Region * X.years, family = poisson(link = "log"),
##     data = butterfly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8090  -0.9928  -0.5422  -0.2006   2.8949
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.707518   0.353682  -2.000   0.0455 *
## RegionU         -1.525202   0.804930  -1.895   0.0581 .
## X.years          0.047996   0.024559   1.954   0.0507 .
## RegionU:X.years  0.004086   0.057194   0.071   0.9430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 81.450  on 55  degrees of freedom
## Residual deviance: 65.021  on 52  degrees of freedom
## AIC: 107.59
```

```
##
## Number of Fisher Scoring iterations: 6
```

When we check the ratio ($64.021/52 = 1.23$), we see that it is slightly greater than 1 and potentially indicates overdispersion.

When we look at the Q-Q plots, we can see that there are several values that do not fall along the straight line near the right tail end. These may be outliers that are driving overdispersion. This is backed up by the fact that there are several points at the right end of the residuals versus fitted plots do not seem to follow one of the two bands that we expected to see (one for each region).

I will explore other models:

```
mod2 <- glm(LargeSkipper~Region + X.years, data=butterfly, family=poisson(link="log"))
anova(mod2, mod, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: LargeSkipper ~ Region + X.years
## Model 2: LargeSkipper ~ Region * X.years
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1        53     65.026
## 2        52     65.021  1 0.0051016   0.9431
```

```
mod3 <- glm(LargeSkipper~ X.years, data=butterfly, family=poisson(link="log"))
anova(mod3, mod, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: LargeSkipper ~ X.years
## Model 2: LargeSkipper ~ Region * X.years
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        54     76.337
## 2        52     65.021  2   11.316 0.003489 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod4 <- glm(LargeSkipper~Region, data=butterfly, family=poisson(link="log"))
anova(mod4, mod, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: LargeSkipper ~ Region
## Model 2: LargeSkipper ~ Region * X.years
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        54     69.687
## 2        52     65.021  2   4.6659  0.09701 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the drop in deviance tests where we have either Region or X.Years as a term, there is strong evidence that the more complex model (the fitted model that has all of the interaction terms) is a better fit.

However, when we compare mod2 (When we have Region + X.Years as terms) to mod, we see that there is little evidence for the complex model.

Another way to fix the overdispersion is to use the quasipoisson family:

```
mod_quasi <- glm(LargeSkipper~Region +X.years, data=butterfly, family = quasipoisson)

summary(mod_quasi)
```

```
##
## Call:
## glm(formula = LargeSkipper ~ Region + X.years, family = quasipoisson,
##     data = butterfly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8183  -0.9886  -0.5465  -0.2068   2.8921
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.71607    0.40843  -1.753   0.0853 .
## RegionU     -1.48039    0.60898  -2.431   0.0185 *
## X.years      0.04875    0.02713   1.797   0.0781 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.496457)
##
##     Null deviance: 81.450  on 55  degrees of freedom
## Residual deviance: 65.026  on 53  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

After fitting the model with a quasipoisson family, we can see that fewer of the coefficients are statistically significant.

Therefore, my final model is:

`glm(formula = LargeSkipper ~ Region + X.years, family = quasipoisson,  data = butterfly)`

The ratio of residual deviance to degrees of freedom is slightly reduced and we see from the model that the Region that the Largeskippers come from has the greatest influence. Compared to the Scania Region, there is a -1.48 multiplicative change in expected number of Largeskippers.