p8105_hw3_mp3745

Matthew Parker

2020-02-24

Problem 1

1)

Enter in data

```
cancer = tibble(
  age = rep(c(25, 35, 45, 55, 65, 75), 2),
  alc_cons = c(rep("0-79", 6), rep("80+", 6)),
  case = c(0, 5, 21, 34, 36, 8, 1, 4, 25, 42, 19, 5),
  ctrl = c(106, 164, 138, 139, 88, 31, 9, 26, 29, 27, 18, 0)
)
```

Model

```
# Response matrix
resp = cancer %>%
  select(case, ctrl) %>%
  as.matrix()

# Model
logit_cancer = glm(resp ~ age + alc_cons, family = binomial(link = 'logit'), data = cancer)

logit_cancer %>%
  broom::tidy() %>%
  knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-5.0234488	0.4182235	-12.011398	0
age	0.0615787	0.0072907	8.446252	0
alc_cons80+	1.7799995	0.1870860	9.514337	0

From the logit model, the log odds of esophageal cancer for someone with daily alcohol consumption of 0-79g that is age 0 is -5.0234488. The log odds ratio of esophageal cancer comparing someone with daily alcohol consumption 80+g to 0-79g is 1.7799995, holding age constant. And the log odds ratio of esophageal cancer associated with a one year increase in age is 0.0615787, holding daily alcohol consumption constant. From these results, age and is positively associated with esophageal cancer and a higher daily alcohol consumption (80+g) is also associated with esophageal cancer.

Problem 2

1)

Enter in data

```
germ = tibble(
  species = c(rep("oa_75", 11), rep("oa_73", 10)),
```

```
  root = c(rep("bean", 5), rep("cucumber", 6), rep("bean", 5), rep("cucumber", 5)),
  germinated = c(10, 23, 23, 26, 17, 5, 53, 55, 32, 46, 10, 8, 10, 8, 23, 0, 3, 22, 15, 32, 3),
  total = c(39, 62, 81, 51, 39, 6, 74, 72, 51, 79, 13, 16, 30, 28, 45, 4, 12, 41, 30, 51, 7),
  failed = total - germinated
)

# response
germ_resp = germ %>%
  select(germinated, failed) %>%
  as.matrix()
```

Fit model

```
germ_logit = glm(germ_resp ~ species + root, family = binomial(link = 'logit'), data = germ)

germ_logit %>%
  broom::tidy() %>%
  knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.7004832	0.1507214	-4.647537	0.0000034
speciesoa_75	0.2704511	0.1547056	1.748166	0.0804353
rootcucumber	1.0647498	0.1442142	7.383115	0.0000000

From the logit model, the log odds of germinating for an O. aegyptiaca 73 seed with bean root extract is -0.7004832. The log odds ratio of germinating comparing an O. aegyptiaca 75 seed with an O. aegyptiaca 73 seed is 0.2704511, holding root extract media constant. The log odds ratio of germinating comparing a seed with cucumber root extract to bean root extract is 1.0647498, holding seed type constant. From these results, root extract medium is strongly associated with germinating. O. aegyptiaca 75 may also be associated with germinating, although its p-value is greater than 0.05, indicating the term is insignificant, accounting for root extract media.

2)

Check for lack of fit

```
pval = 1 - pchisq(germ_logit$deviance, 21 - 3)
```

Since p-value = 0.0023028 < 0.05, we have sufficient evidence to reject the null hypothesis that the model fits the data well.

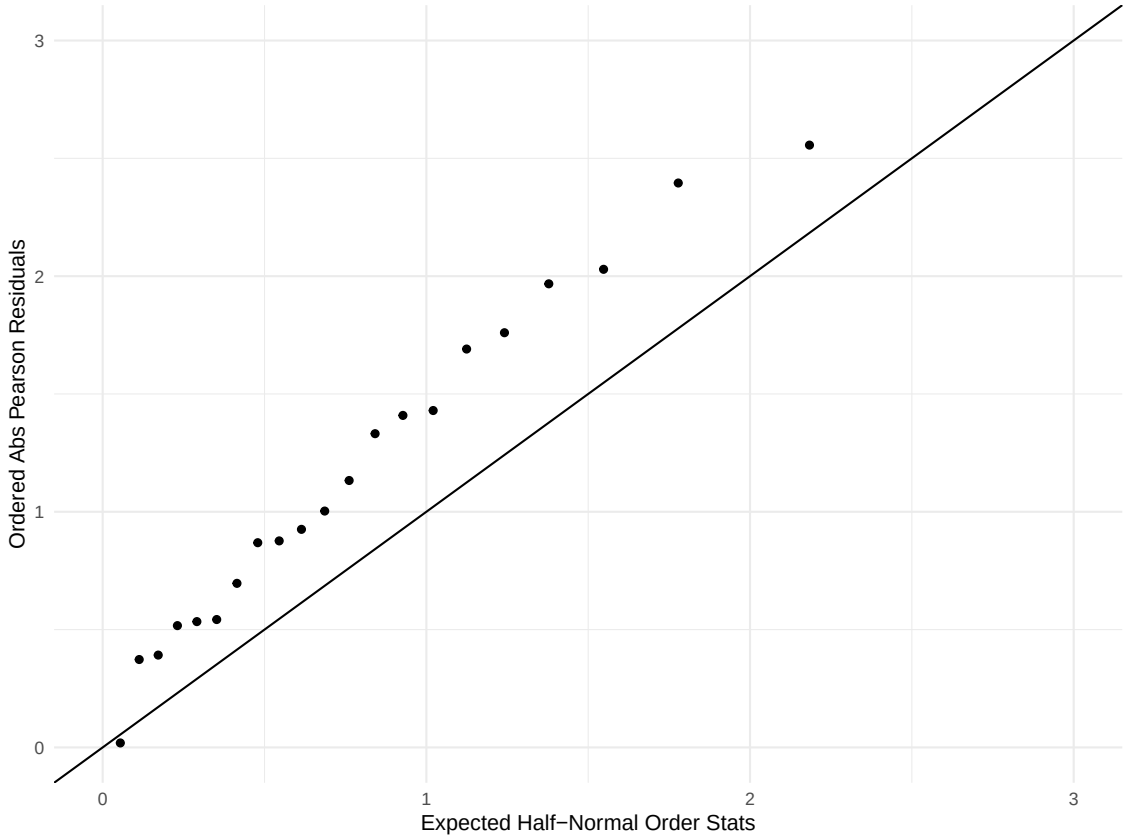
Check for overdispersion

```
res = residuals(germ_logit, type = 'pearson')

resid_points = tibble(
  x = qnorm((21 + 1:21 + 0.5)/(2 * 21 + 1.125)),
  y = sort(abs(res))
)

resid_points %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_abline(slope = 1) +
```

```
  scale_x_continuous(limits = c(0, 3)) +
  scale_y_continuous(limits = c(0, 3)) +
  labs(
    x = "Expected Half-Normal Order Stats",
    y = "Ordered Abs Pearson Residuals"
  )
```



Based on above plot, it appears they may be a liner deviation from the reference line, which indicates possible overdispersion.

Calculate estimate of dispersion parameter

```
g_stat = sum(residuals(germ_logit, type = 'pearson')^2)

phi = g_stat/(21 - 3)
```

The estimated dispersion parameter is 2.1283678

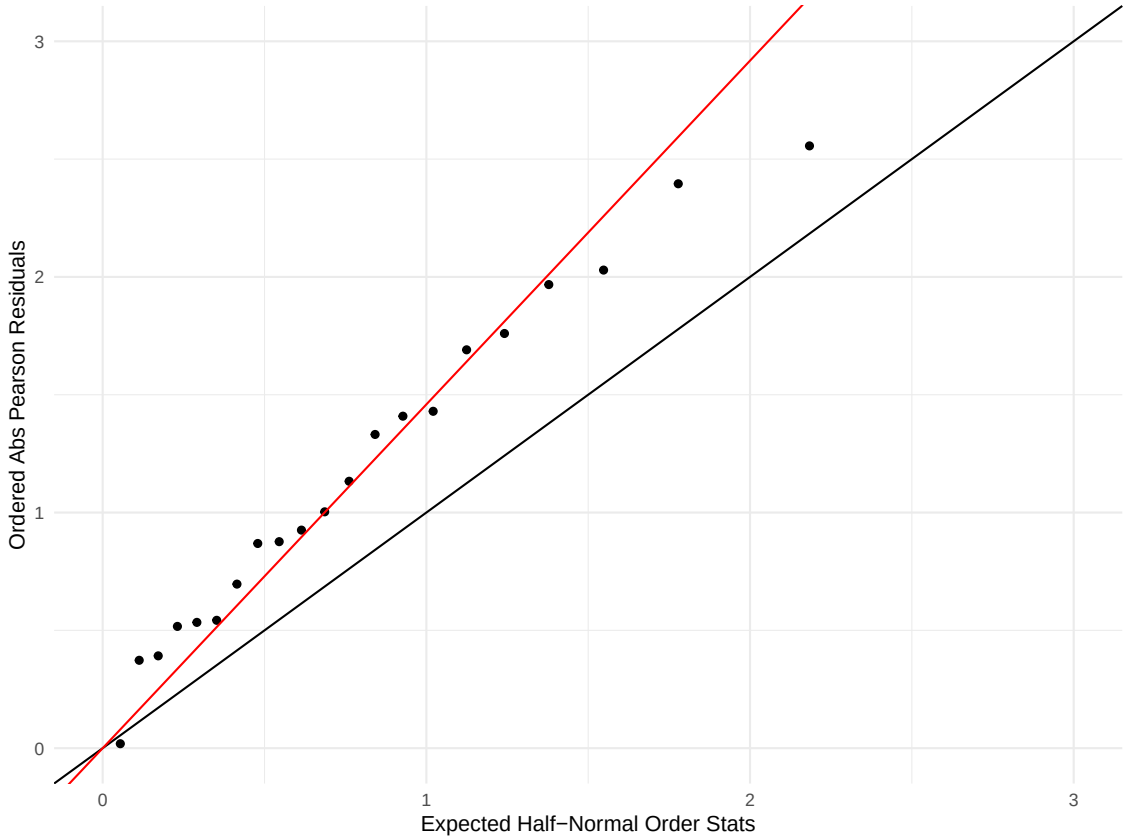
Update model with dispersion parameter and check plot again

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

```
##
## Call:
## glm(formula = germ_resp ~ species + root, family = binomial(link = "logit"),
##     data = germ)
##
## 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     0.2199  -3.186  0.00144 **
## speciesoa_75   0.2705     0.2257   1.198  0.23081
## rootcucumber   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
```

```
resid_points %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_abline(slope = 1) +
  geom_abline(slope = sqrt(phi), color = 'red') +
  scale_x_continuous(limits = c(0, 3)) +
  scale_y_continuous(limits = c(0, 3)) +
  labs(
    x = "Expected Half-Normal Order Stats",
    y = "Ordered Abs Pearson Residuals"
  )
```



From the plot with a line with slope equal to the square root of the dispersion parameter (red line), we can see this line fits the data much better than our original line.

In the updated model, our estimates and their interpretations are still valid. In the updated model, the std. error has been inflated by 2.1283678. This has caused the z value and Pr(>|z|) to change. The coefficient for the term comparing species of seed in the model has now clearly become insignificant, indicating there may not be a significant association between seed species and germinating when accounting for root extract media.

3)

A possible source of the overdispersion is correlation within each group of seeds. Because the seeds are germinated in groups, there may be different external factors (i.e. different temperature) for different groups that affect the rate of germination.