StatsHw7

Trevor Freeland

May 6, 2018

```
library(lme4)
library(pander)
library(stargazer)
library(tidyverse)
```

```
medicare <- read.csv("http://math.carleton.edu/Chihara/Stats345/Medicare.csv")
turtle <- read.csv("http://math.carleton.edu/Chihara/Stats345/Turtle.csv")
teratology <- read.csv("http://math.carleton.edu/Chihara/Stats345/Teratology.csv")
```

58

Final model:

LogOdds(Survival) = .75 -.62(Emergency) -.31(Urgent)

Type-Emergency = -.62 -> Holding everything else constant, a patient admitted for an emergency has an odds of survival that is 54% ($e^{-.62}$) that of a patient that was admitted by their choice.

Type-Urgent = -.31 -> Holding everything else constant, a patient admitted for an urgent reason has an odds of survival 73% ($e^{-.31}$) of that of a patient that was admitted by their choice.

```
med.glm <- glm(alive/m~white + hmo + type, family = binomial, data = medicare, weights = m)
summary(med.glm)
med.glm2 <- glm(alive/m~white + type, family = binomial, data = medicare, weights = m)
anova(med.glm, med.glm2, test = "Chisq")
summary(med.glm2)
med.glm3 <- glm(alive/m~type, family = binomial, data = medicare, weights = m)
anova(med.glm3, med.glm2, test = "Chisq")
#.07 for my p-value,I will not include it in my model based on the .05 significance level.
summary(med.glm3)
exp(coef(med.glm3))
```

59

After running the above simulations 100 times we can see that there is a difference in the mean and the standard deviations of the two different models, the independent assuming model and the non-independent model. In the Table below we can see that on average in the independent trials we got a p-value of .3. In the model that didn't assume independence, we got a mean p-value of .33, slightly higher than that of the independent trials. This makes sense beause without independence we were assuming that moths were at least at a probability of .3 of being eaten, but if a moth "next to" them were eaten then the probability of that moth getting eaten rose. However it is also worth looking at the average variances in the table as well. We can see that the Var increase in the non-independent model on average compared to the independent assuming model. Again this makes sense because we were adding a randomly uniform increase in the probability that a moth was eaten aftera moth next to them was eaten, which means we are adding more variablity into our model.

See table below for full results

```r
set.seed(12346789)
N <- 10^3
NN <- 100
pihat <- numeric(NN)
expectHat <- numeric(NN)
VarHat <- numeric(NN)
pihat2 <- numeric(NN)
expectHat2 <- numeric(NN)
VarHat2 <- numeric(NN)

#Running our intitial simulation 100 times.
#Keeping Track of p-values,
#expected values and variances each time.

for (j in 1:NN){
Removed <- rbinom(N, 1, .3) #Vector of outcomes 1= eaten, 0 = live!
pihat[j] <- mean(Removed) #pi^
expectHat[j] <- N*pihat[j] #E[Y]
VarHat[j] <- N*pihat[j]*(1-pihat[j]) #variance

ep <- 0 #epsilon = 0
Removed2 <- numeric(N)
Removed2[1] <- rbinom(1, 1, .3) #First moth
for (i in 2:N){
#The if statement is saying that if the moth before the selected one was eaten,
#then we are going to increase the probability that the next moth gets eaten.
if (Removed2[i-1] == 1) ep <- runif(1, 0, .2) else ep <- 0
Removed2[i] <- rbinom(1, 1, .3 + ep)
}
pihat2[j] <- mean(Removed2)
expectHat2[j] <- N*pihat2[j]
VarHat2[j] <- N*pihat2[j]*(1-pihat2[j])
}
```

```r
ind <- data.frame(meanP = mean(pihat), sdP = sd(pihat), meanEX = mean(expectHat), sdEX = sd(expectHat),
no_ind <- data.frame(meanP = mean(pihat2), sdP = sd(pihat2), meanEX = mean(expectHat2), sdEX = sd(expec
table <- rbind(ind, no_ind)
row.names(table) <- c("Independent", "Not Independent")
table <- round(table, 2)
table
```

```
                meanP  sdP meanEX  sdEX meanVar sdVar
Independent      0.30 0.01 298.13 13.75  209.06  5.51
Not Independent  0.33 0.02 333.02 16.27  221.86  5.49
```

60

	$\hat{\beta}$	SE	z	P-value
1a.	.97	.46	2.11	.03
1b.	.97	.52	1.87	.06
2a.	1.70	.46	3.70	.0002
2b.	1.70	.52	3.27	.001
3a.	1.70	.80	2.13	.03
3b.	1.70	.91	1.87	.06

If we have overdispersion in our data and we get an overdispersion parameter larger than 1, we should be worried if we have values that are on the border of being significant. The larger the dispersion parameter the wider that "border" becomes, but in the table above we can see multiple instances with just a dispersion parameter of 1.3 made problems 1 and 3 both have coefficients go from being "statistically significant" to not (based on .05 p-value). The p-value still changes in problem 2, but since the p-value was so small at first, even the change makes the coefficient still significant. We should always be checking for overdispersion because even values close to 1 can change our models and interpretations.

61

(A)

LogOdds(Males) = -61.3 + 2.21(Temp)

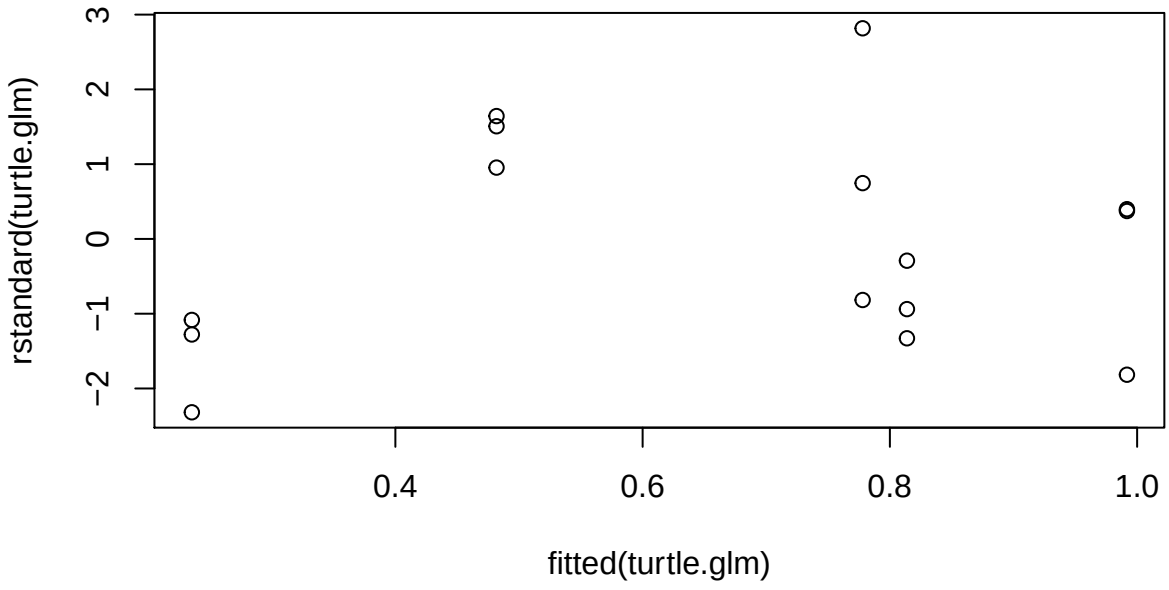
When conducting a goodness of fit test using the residual deviance output from the summary table we get a p-value of .024 which indicates that our model is not adequate.

```
turtle.glm <- glm(cbind(male, female)~temp, family=binomial, data = turtle)
summary(turtle.glm)
good.fit <- 1 - pchisq(24.9, 13)
good.fit
```

(B)

There does not appear to be any outliers, all of the standard residuals are within +/- 3.

```
plot(rstandard(turtle.glm)~ fitted(turtle.glm))
```



(C)

The dispersion parameter is about 2, so we should rerun the model using a quasibinomial setting and check for differences. When re-running the model while accounting for overdispersion there does not appear to be an changes in significance in any of the variables. Obviously our standard errors increased for all our of terms but we still got t-values greater than 3.5 the terms do still seem to be significant.

```
sum(residuals(turtle.glm, type = "pearson")^2)/turtle.glm$df.residual
turtle.glm2 <- glm(cbind(male, female)~temp, family=quasibinomial, data = turtle)
summary(turtle.glm2)
summary(turtle.glm)
good.fit <- 1 - pchisq(24.9, 13)
good.fit
```

62

(A)

A 95% confidence interval for the coefficient of temperature in the log odds is between 1.16 and 3.55. On the odds scale this corresponds with a 95% confidence interval of 3.17 to 34.93. This means that for a 1 degree increase in temperature the odds that an offspring is male increase between 317% and 3,493%.

(B)

Overall all of the estimates between the three models are very similar. The only real difference is in the standard deviations of the coefficients in the models. The mixed-effects model seems to have standard errors for the coefficients in between that of our binomial model without accounting for overdispersion and our model that accounts for overdispersion.

```
exp(confint(turtle.glm2))
```

```
turtle$id <- 1:15
turtle.glmer <- glmer(cbind(male, female)~temp + (1|id), family=binomial, data = turtle)
#summary(turtle.glmer)
stargazer(turtle.glmer, turtle.glm, turtle.glm2, type = "text")
```

```
=================================================================
                              Dependent variable:
                    ---------------------------------------------
                              cbind(male, female)
                    generalized linear  logistic  glm: quasibinomial
                      mixed-effects                 link = logit
                           (1)            (2)            (3)
-----------------------------------------------------------------
temp                     2.216***       2.211***       2.211***
                         (0.529)        (0.431)        (0.612)

Constant                -61.518***     -61.318***     -61.318***
                         (14.805)       (12.022)       (17.081)

-----------------------------------------------------------------
Observations                15             15             15
```

```
Log Likelihood           -24.122        -24.918
Akaike Inf. Crit.         54.244         53.836
Bayesian Inf. Crit.       56.368
=================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

63

(A)

LogOdds(Dead) = 1.14 - 3.32(GroupB) - 4.48(GroupC) - 4.13(groupD)

Holding all else constant, fetuses that were from a rat in group B had a 96.4% (1 - $e^{-3.32}) decrease in the odds of dying than of a fetus from a rat in the placebo group.

Holding all else constant, fetuses that were from a rat in group B had a 98.9% (1 - $e^{-4.48}) decrease in the odds of dying than of a fetus from a rat in the placebo group.

Holding all else constant, fetuses that were from a rat in group B had a 98.4% (1 - $e^{-4.13}) decrease in the odds of dying than of a fetus from a rat in the placebo group.

(B)

We got a overdispersion parameter of 2.86, which indicates that overdispersion is a problem. After re-running the model while accounting for overdispersion we still find that all of the different groups are significant. Our largest t value went from 6.12 to 3.6 after accounting for overdispersion.

(C)

When fitting a GLMM to our data, we get larger estimates for the effects of all of the different groups, and all of them are still significant. We can see the side by side fixed effect estimates of the three different models in the table below, where we can see that the mixed-effects implys that the groups have larger effects, but also higher standard deviations on some of the coefficients.

```
teratology$alive <- teratology$n - teratology$y
tera.glm <- glm(cbind(y, alive)~group, data = teratology, family = binomial)
summary(tera.glm)
```

```
sum(residuals(tera.glm, type = "pearson")^2)/tera.glm$df.residual
tera.glm2 <- glm(cbind(y, alive)~group, data = teratology, family = quasibinomial)
summary(tera.glm2)
```

```
tera.glmer <- glmer(cbind(y, alive)~group + (1|litter), data = teratology, family = binomial)
#summary(tera.glmer)
stargazer(tera.glmer, tera.glm, tera.glm2, type = "text")
```

```
=================================================================
                              Dependent variable:
                    ---------------------------------------------
                                  cbind(y, alive)
                    generalized linear logistic  glm: quasibinomial
                      mixed-effects                 link = logit
                           (1)            (2)           (3)
```

groupB	-4.540***	-3.323***	-3.323***
	(0.735)	(0.331)	(0.560)
groupC	-5.883***	-4.476***	-4.476***
	(1.175)	(0.731)	(1.238)
groupD	-5.606***	-4.130***	-4.130***
	(0.908)	(0.476)	(0.806)
Constant	1.809***	1.144***	1.144***
	(0.362)	(0.129)	(0.219)
Observations	58	58	58
Log Likelihood	-92.081	-122.461	
Akaike Inf. Crit.	194.163	252.923	
Bayesian Inf. Crit.	204.465		
Note:		*p<0.1; **p<0.05; ***p<0.01	