4211 Homework 7

Matthew DeSantis

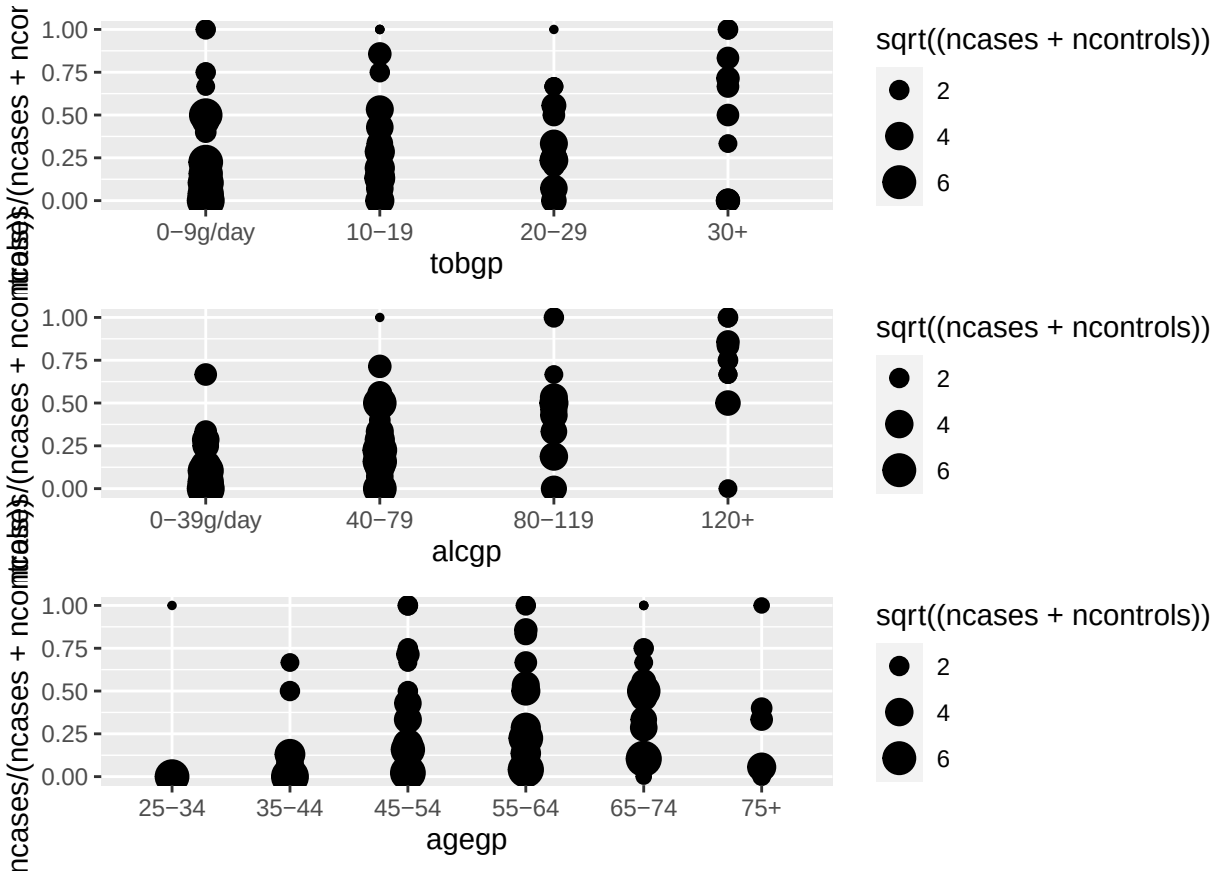
2023-03-30

1 (3.1)

(a)

```
tob = ggplot(esoph,aes(y=(ncases/(ncases+ncontrols)), x=tobgp, size=sqrt((ncases+ncontrols)))) + geom_p
alc = ggplot(esoph,aes(y=(ncases/(ncases+ncontrols)), x=alcgp, size=sqrt((ncases+ncontrols)))) + geom_p
age = ggplot(esoph,aes(y=(ncases/(ncases+ncontrols)), x=agegp, size=sqrt((ncases+ncontrols)))) + geom_p

grid.arrange(tob, alc, age, ncol=1)
```



The plots seem to show that higher alcohol usage, tobacco usage, and age seem to increase the chance of getting cancer. This relationship seems to be very strong for alcohol usage, and weaker for age.

(b)

```
mod1 = glm(cbind(ncases, ncontrols) ~ agegp*alcgp*tobgp, family = binomial, data = esoph)
mod2 = step(mod1, trace = FALSE)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mod2)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
##     family = binomial, data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9507  -0.7376  -0.2438   0.6130   2.4127
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.19039    0.20737  -5.740 9.44e-09 ***
## agegp.L      3.99663    0.69389   5.760 8.42e-09 ***
## agegp.Q     -1.65741    0.62115  -2.668  0.00762 **
## agegp.C      0.11094    0.46815   0.237  0.81267
## agegp^4      0.07892    0.32463   0.243  0.80792
## agegp^5     -0.26219    0.21337  -1.229  0.21915
## alcgp.L      2.53899    0.26385   9.623  < 2e-16 ***
## alcgp.Q      0.09376    0.22419   0.418  0.67578
## alcgp.C      0.43930    0.18347   2.394  0.01665 *
## tobgp.L      1.11749    0.24014   4.653 3.26e-06 ***
## tobgp.Q      0.34516    0.22414   1.540  0.12358
## tobgp.C      0.31692    0.21091   1.503  0.13294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.953  on 87  degrees of freedom
## Residual deviance:  82.337  on 76  degrees of freedom
## AIC: 221.39
##
## Number of Fisher Scoring iterations: 6
```

(c)

```
mod3 = glm(cbind(ncases, ncontrols) ~ unclass(agegp)*unclass(alcgp)*unclass(tobgp), family = binomial, d
mod4 = step(mod3, trace = FALSE)
summary(mod4)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp), family = binomial, data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6478  -0.9246  -0.4338   0.6740   2.4568
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.16395    0.50931 -14.066  < 2e-16 ***
## unclass(agegp)  0.74375    0.08179   9.094  < 2e-16 ***
## unclass(alcgp)  1.10255    0.10317  10.687  < 2e-16 ***
## unclass(tobgp)  0.43085    0.09394   4.587 4.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.95  on 87  degrees of freedom
## Residual deviance: 108.78  on 84  degrees of freedom
## AIC: 231.83
##
## Number of Fisher Scoring iterations: 4
```

(d)

```
mod5 = glm(cbind(ncases, ncontrols) ~ unclass(agegp)+unclass(alcgp)+unclass(tobgp)+I(unclass(agegp)^2),
summary(mod5)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp) + I(unclass(agegp)^2), family = binomial,
##     data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2757  -0.7828  -0.2313   0.5679   2.4646
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -10.10233    1.03074  -9.801  < 2e-16 ***
```

```
## unclass(agegp)        2.50576    0.50188   4.993 5.95e-07 ***
## unclass(alcgp)        1.06511    0.10458  10.185  < 2e-16 ***
## unclass(tobgp)        0.43951    0.09559   4.598 4.27e-06 ***
## I(unclass(agegp)^2)  -0.23417    0.06402  -3.658 0.000255 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.953  on 87  degrees of freedom
## Residual deviance:  93.172  on 83  degrees of freedom
## AIC: 218.23
##
## Number of Fisher Scoring iterations: 5
```

(e)

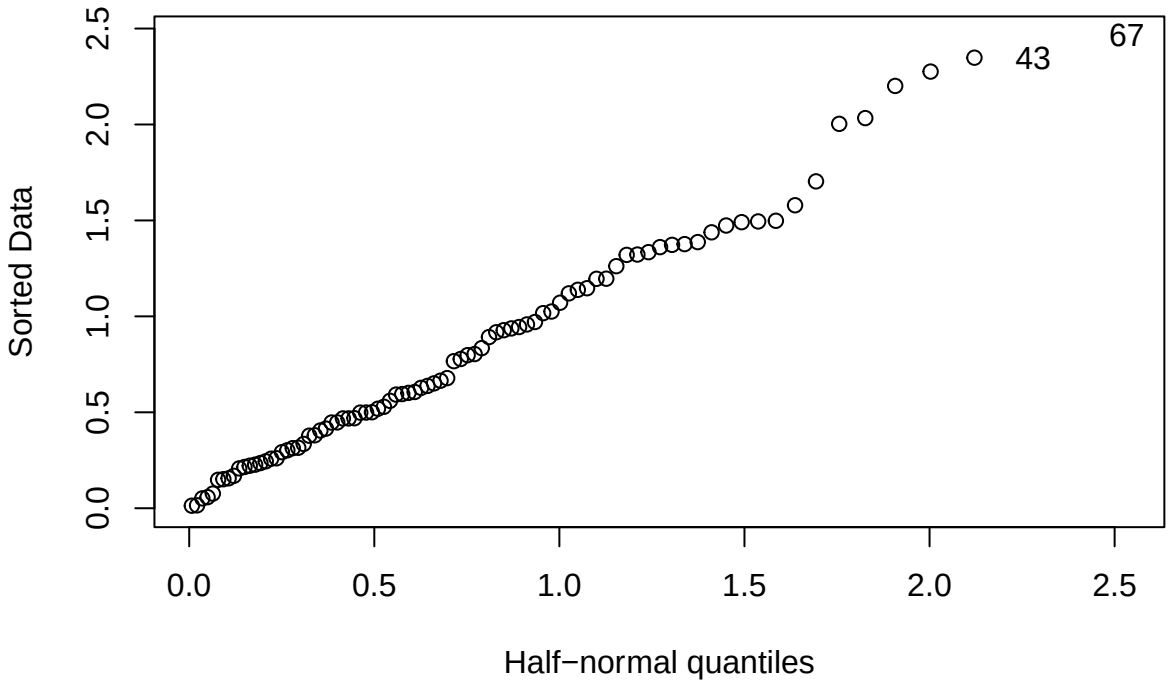
```
drop1(mod5)
```

```
## Single term deletions
##
## Model:
## cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp) + I(unclass(agegp)^2)
##                      Df Deviance    AIC
## <none>                    93.172 218.23
## unclass(agegp)        1  126.099 249.15
## unclass(alcgp)        1  215.963 339.02
## unclass(tobgp)        1  114.342 237.40
## I(unclass(agegp)^2)   1  108.779 231.83
```

This model appears to be the best model; dropping any of its parameters would result in a higher AIC.

(f)

```
halfnorm(residuals(mod5))
```



The model doesn't appear to have outliers.

(g)

```
exp(mod5$coefficients[3])
```

```
## unclass(alcgp)
##       2.901154
```

The odds go up by a factor of the listed number.

(h)

```
exp(confint(mod5)[3])
```

```
## Waiting for profiling to be done...
```

```
## [1] 2.373678
```

```
exp(confint(mod5)[8])
```

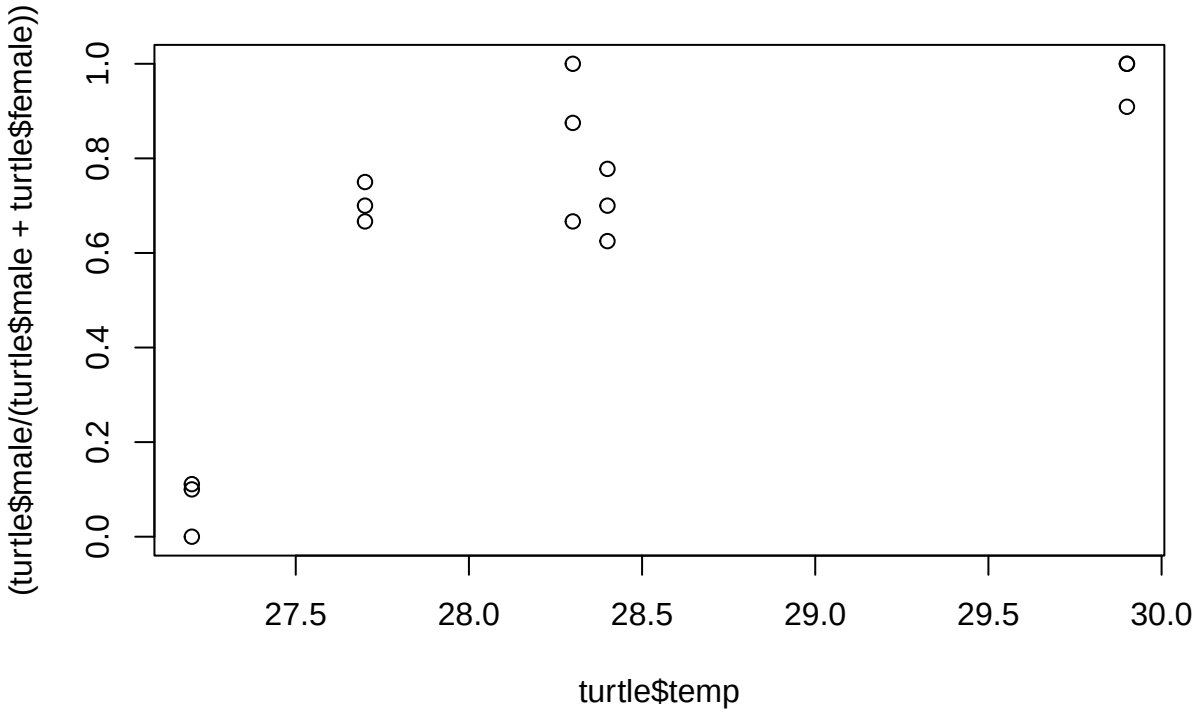
```
## Waiting for profiling to be done...
```

```
## [1] 3.578623
```

2 (3.2)

(a)

```
data("turtle")
plot(x=turtle$temp, y=(turtle$male/(turtle$male+turtle$female)))
```



There seems to be a positive relationship between the temperature and the proportion of males born.

(b)

```
tmod1 = glm(cbind(male, female) ~ temp, family = binomial, data = turtle)
summary(tmod1)
```

```
##
## Call:
## glm(formula = cbind(male, female) ~ temp, family = binomial,
##     data = turtle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0721  -1.0292  -0.2714   0.8087   2.5550
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
## temp          2.2110     0.4309   5.132 2.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.508  on 14  degrees of freedom
## Residual deviance: 24.942  on 13  degrees of freedom
## AIC: 53.836
##
## Number of Fisher Scoring iterations: 5
```

```
pchisq(24.942, 13, lower.tail = FALSE)
```

```
## [1] 0.02349208
```

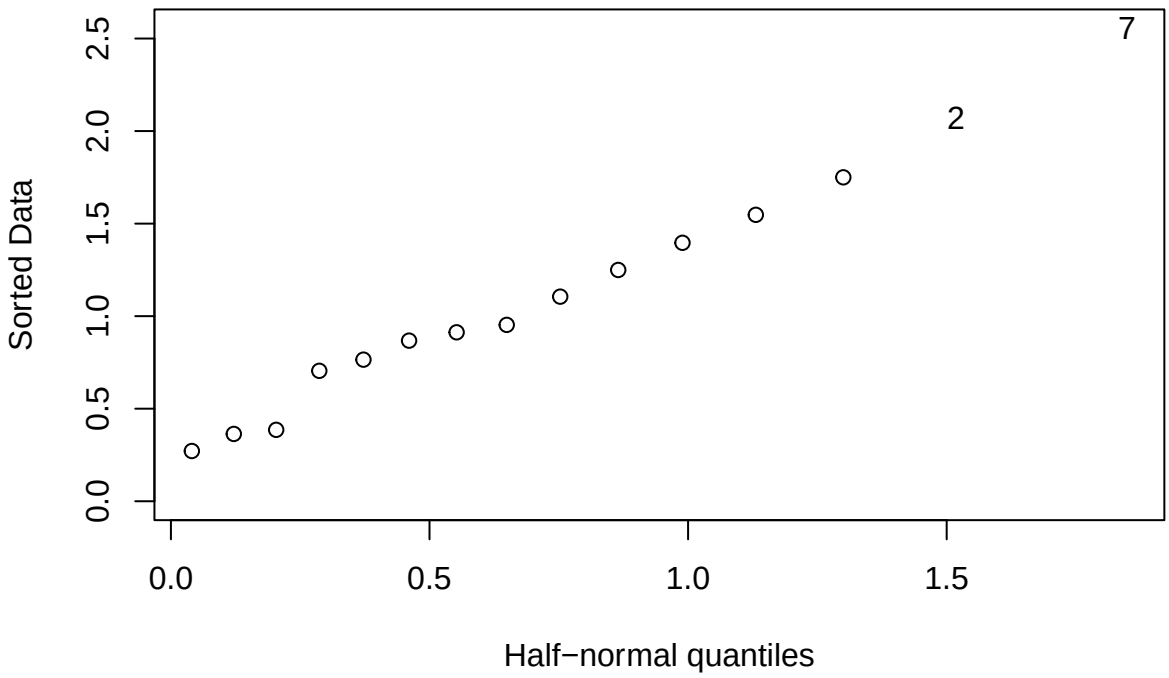
It does not appear to be a good fit.

(c)

No, most of the values are not zero.

(d)

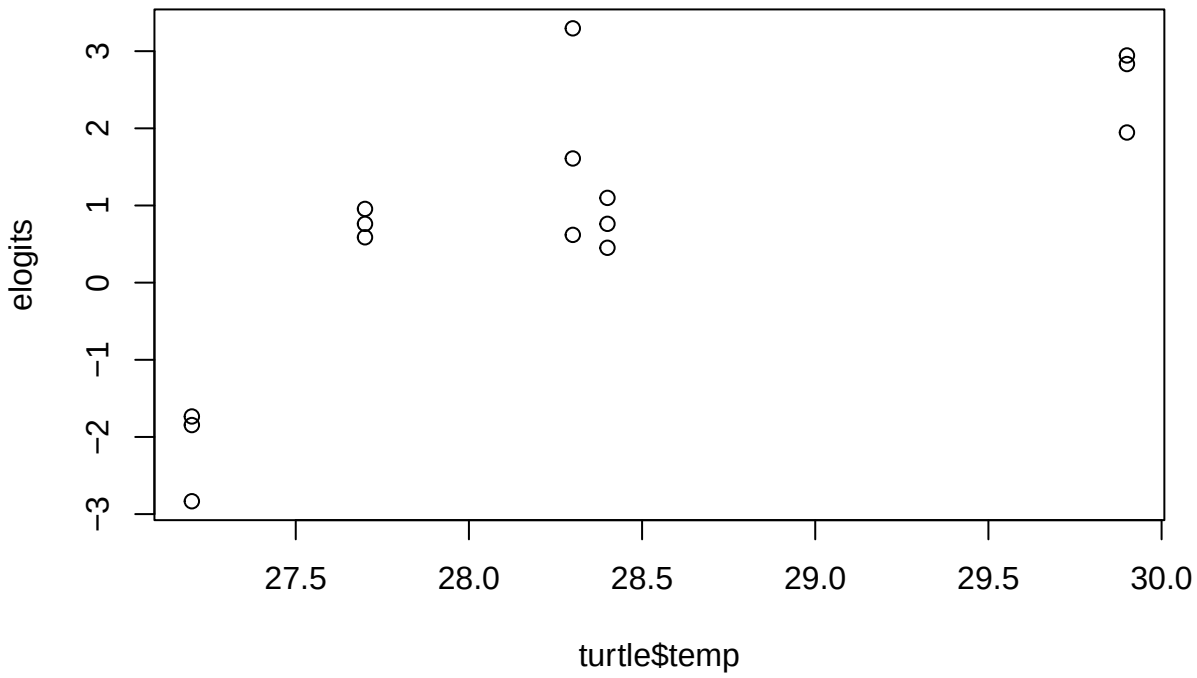
```
halfnorm(residuals(tmod1))
```



There do not appear to be any outliers.

(e)

```
elogits <- with(turtle,log((male+0.5)/(female+0.5)))
plot(turtle$temp,elogits)
```



The relationship doesn't appear to be linear, so this indicates a lack of fit.

(f)

```
tmod2 = glm(cbind(male, female) ~ temp + I(temp^2), family = binomial, data = turtle)
summary(tmod2)
```

```
##
## Call:
## glm(formula = cbind(male, female) ~ temp + I(temp^2), family = binomial,
##     data = turtle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6703  -0.8875  -0.4194   0.9481   2.2198
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -677.5950   268.7984  -2.521   0.0117 *
## temp          45.9173    18.9169   2.427   0.0152 *
## I(temp^2)     -0.7745     0.3327  -2.328   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.508  on 14  degrees of freedom
## Residual deviance: 20.256  on 12  degrees of freedom
## AIC: 51.15
##
## Number of Fisher Scoring iterations: 4
```

It is significant.

(g)

```
vars = data.frame(temp = c(), expected = c(), observed = c())

for(x in c(27.2,27.7,28.3,28.4,29.9)){
  male = sum(turtle$male[turtle$temp==x])
  female = sum(turtle$female[turtle$temp==x])
  total = male+female
  expected = (male/total)*(female/total)/3
  observed = var(c(rep(1,male), rep(0,female)))
  vars = rbind(vars,c(x,expected,observed))
  colnames(vars) = c("temp","expected variance", "observed variance")
  #vars$temp = append(vars$temp, x)
  #vars$expected = append(vars$expected, expected)
  #vars$observed = append(vars$observed, observed)
}
vars
```

```
##   temp expected variance observed variance
## 1 27.2        0.02286237        0.07122507
## 2 27.7        0.06886574        0.21557971
## 3 28.3        0.03851852        0.11954023
## 4 28.4        0.06950160        0.21652422
## 5 29.9        0.01147959        0.03571429
```

The expected variance would be nqp, which would be (male/total)x(female/total)/3. I calculate the observed variance with var(). As can be seen, there appears to be much more observed variance at every level. This indicates overdispersion. ## (h)

```
turt2 = data.frame()
for(x in c(27.2,27.7,28.3,28.4,29.9)){
  male = sum(turtle$male[turtle$temp==x])
  female = sum(turtle$female[turtle$temp==x])
  turt2 = rbind(turt2,c(x,male,female))
  colnames(turt2) = c("temp","male", "female")
}
tmod3 = glm(cbind(male, female) ~ temp, family = binomial, data = turt2)
summary(tmod3)
```

```
##
## Call:
```

```
## glm(formula = cbind(male, female) ~ temp, family = binomial,
##     data = turt2)
##
## Deviance Residuals:
##      1       2       3       4       5
## -2.224   2.248   1.239  -1.382  -1.191
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
## temp          2.2110     0.4309   5.132 2.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.429  on 4  degrees of freedom
## Residual deviance: 14.863  on 3  degrees of freedom
## AIC: 33.542
##
## Number of Fisher Scoring iterations: 5
```

This model seems to perform better, with a lower AIC.