S632 HW3

Erik Parker

Feburary 3rd, 2018

1. ELM 3.2: Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male and female turtles born was recorded.

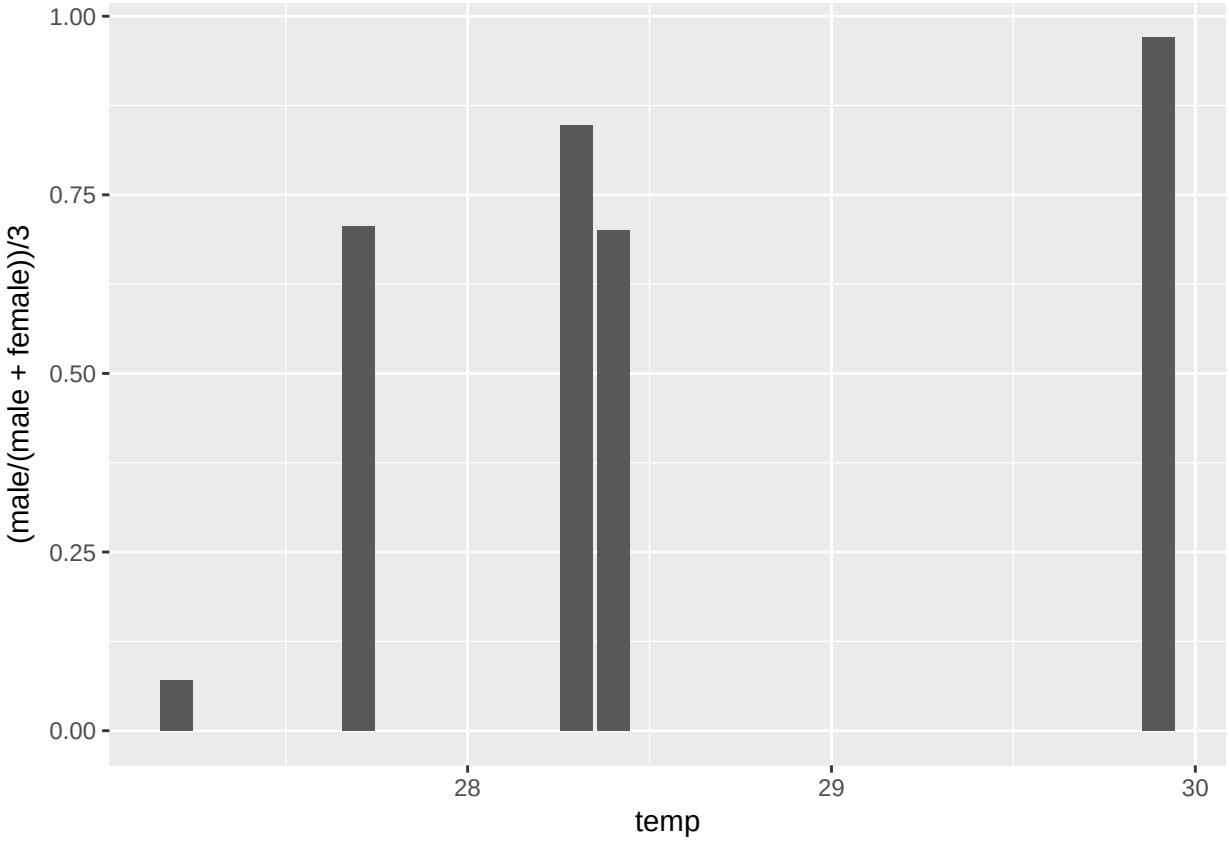
```
rm(list = ls())

library(ggplot2)
library(faraway)
library(dplyr)
library(pscl)

turtles <- turtle
```

a) Plot the proportion of males against the temperature. Comment on the nature of the relationship.

```
ggplot(turtles, aes(x = temp, y = (male/(male + female))/3)) + geom_col()
```



This plot shows us that generally, as the temperature increases, the proportion of males born increases.

b) Fit a binomial response model with a linear term in termperature. Does this model fit the data?

```
m1 <- glm(cbind(male, female) ~ temp, family = binomial, turtles)
summary(m1)
```

```
##
## Call:
## glm(formula = cbind(male, female) ~ temp, family = binomial,
##     data = turtles)
##
## 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(74.508 - 24.942, 1, lower = FALSE)
```

```
## [1] 1.918092e-12
```

```
pchisq(deviance(m1), df.residual(m1), lower = FALSE)
```

```
## [1] 0.02348863
```

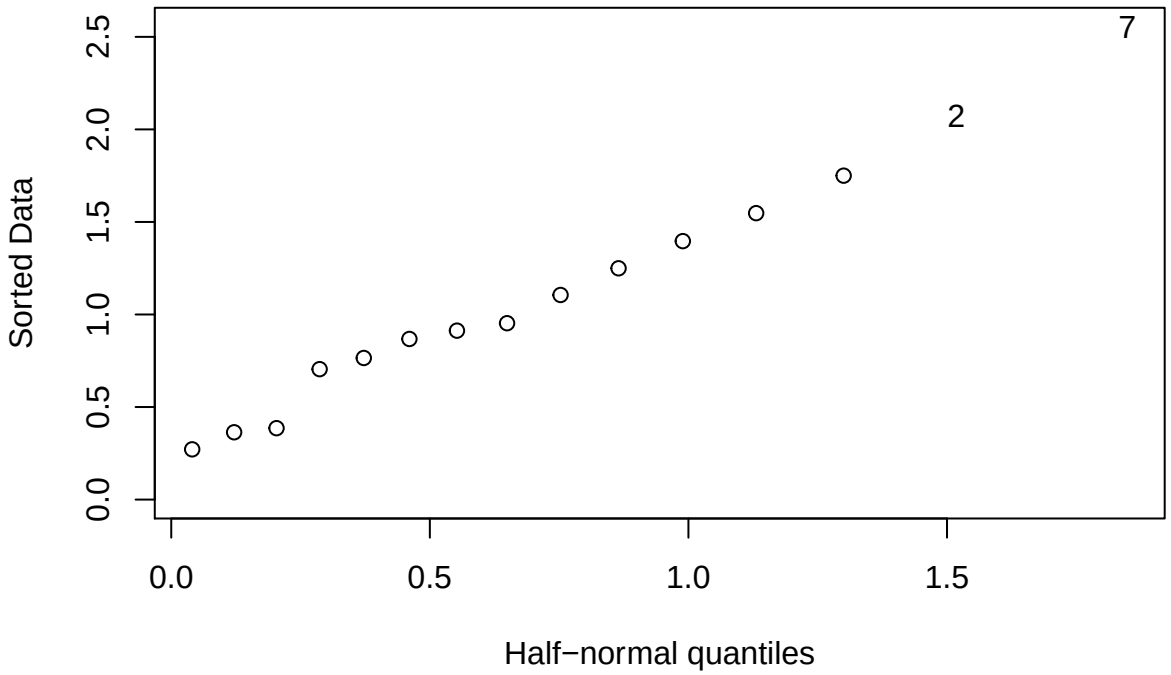
From the first p-value of near zero, we can see that the inclusion of temperature in the model is significant - it significantly reduces the deviance. However, this conclusion is contradicted by the second p-value of 0.023, which as it is less than 0.05, tells us that the model does not fit the data adequately well. So, the inclusion of *temp* to this model improves it, but even with that regressor, the model still does not fit the data well.

c) Is this data sparse?

As defined in class: data can be considered sparse when there are fewer than 5 entries per observation, as a general rule of thumb. For these data, the fewest entries we ever see is 6, so if we go by this general rule, then no - this data is not sparse.

d) Check for outliers

```
halfnorm(residuals(m1))
```

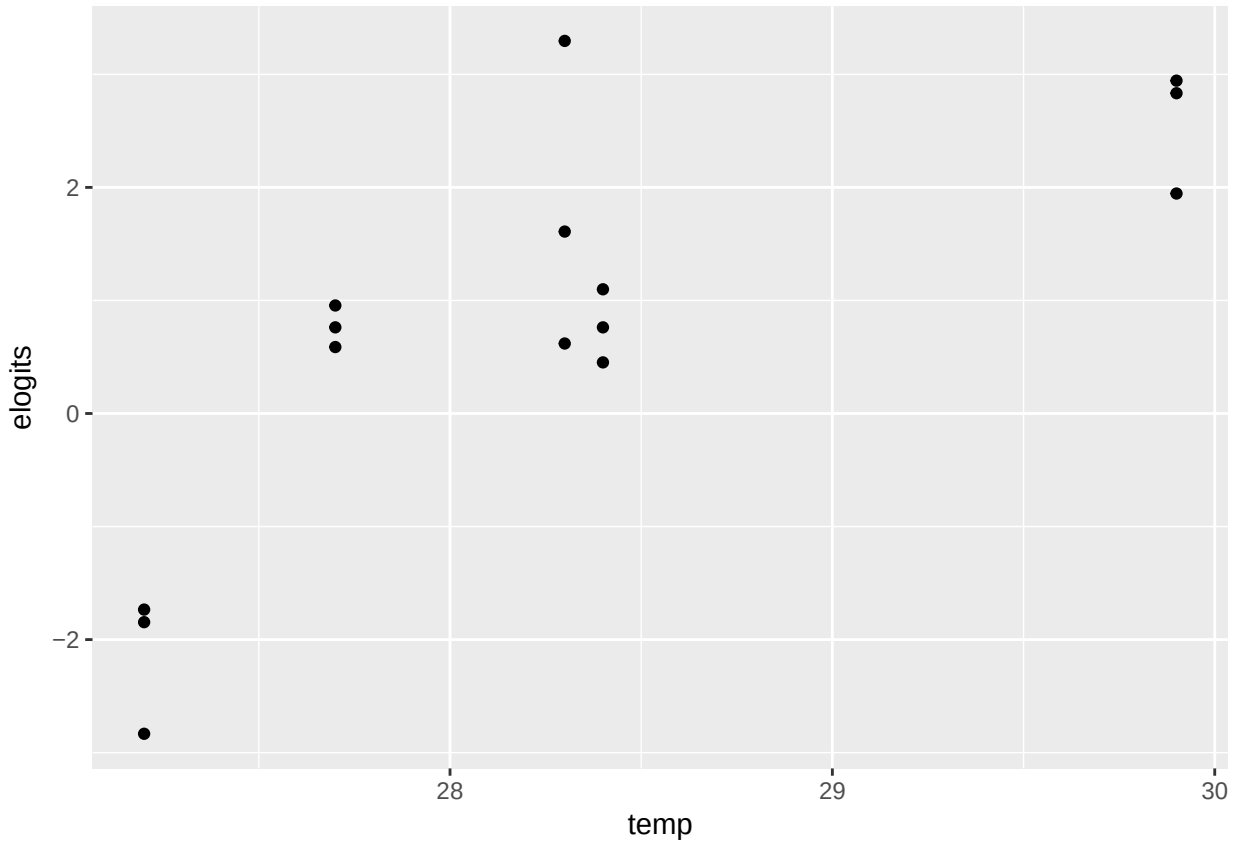


The half-normal plot clearly shows that the residuals of these data are well described by a straight line, meaning that there are no obvious outliers in this dataset.

e) Compute the empirical logits and plot these against temperature. Does this indicate lack of fit?

```
elogits <- with(turtles, log((male + 0.5)/((male + female) - male + 0.5)))

ggplot(turtles, aes(x = temp, y = elogits)) + geom_point()
```



This plot of the empiricial logits against temperature does indicate a lack of fit, as their distribution is not well described by a straight line.

f) Add a quadratic term in temperature. Is this additional term a significant predictor of the response? Does the quadratic model fit the data?

```
m2 <- glm(cbind(male, female) ~ temp + I(temp^2), family = binomial, turtles)
summary(m2)
```

```
##
## Call:
## glm(formula = cbind(male, female) ~ temp + I(temp^2), family = binomial,
##     data = turtles)
##
## 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
```

```
pchisq(74.508 - 20.256, 2, lower = FALSE)
```

```
## [1] 1.657021e-12
```

```
pchisq(deviance(m2), df.residual(m2), lower = FALSE)
```

```
## [1] 0.06239194
```

Using the chi-squared test to compare the deviance between this new model with a quadratic term for temperature and the original (reduced) model lacking a quadratic term, we see that the p-value of near 0 leads us to the conclusion that the effect of the new quadratic term is statistically significant.
Furthermore, we can see from the second test (with a p-value just above 0.05) that in general this quadratic model does fit the data, if just barely at the conventional cutoff of $\alpha = 0.05$. So overall, the model does fit the data, but it's a very marginal case and doesn't fit the data well.

g) There are three replicates for each value of temperature. Assuming independent binomial variation, how much variation would be expected in the three proportions observed? Compare this to the observed variation in these proportions. Do they approximately agree or is there evidence of greater variation?

```
turtles$propmales <- (turtles$male)/(turtles$male + turtles$female)

turtles$probsuccess <- predict(m2, type = "response")

turtles$expectedvar <- turtles$probsuccess * (1 - turtles$probsuccess)

expected <- c(sum(turtles$expectedvar[1:3]), sum(turtles$expectedvar[4:6]),
    sum(turtles$expectedvar[7:9]), sum(turtles$expectedvar[10:12]), sum(turtles$expectedvar[13:15]))

observed <- c(var(turtles$propmales[1:3]), var(turtles$propmales[4:6]), var(turtles$propmales[7:9]),
    var(turtles$propmales[10:12]), var(turtles$propmales[13:15]))

final <- cbind(expected, observed)
final
```

```
##       expected    observed
## [1,] 0.4032151 0.003744856
## [2,] 0.7496974 0.001759259
## [3,] 0.4285067 0.028356481
## [4,] 0.3737966 0.005835905
## [5,] 0.1466677 0.002754821
```

Here we see that the expected variance for each of the replicates is higher than what was actually observed. This means that there is no evidence for greater variation than expected.

h) Combine all three replicates, so there are 5 cases total. Fit a model linear in temperature. Compare the fit seen for this model with that found in b.

```r
turtles2 <- turtles %>% group_by(temp) %>% summarise_all(funs(sum(.)))

m3 <- glm(cbind(male, female) ~ temp, family = binomial, turtles2)
summary(m3)
```

```
##
## Call:
## glm(formula = cbind(male, female) ~ temp, family = binomial,
##     data = turtles2)
##
## 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
```

```r
pchisq(deviance(m3), df.residual(m3), lower = FALSE)
```

```
## [1] 0.001937595
```

```r
pchisq(64.429 - 14.863, 1, lower = FALSE)
```

```
## [1] 1.918092e-12
```

From the first two p-values, we see that the model with 5 cases total and the linear regressor temperature does not fit the data sufficiently well (first p-value of 0.002), but that the inclusion of temperature in this model is considered statistically sufficient compared to the null model (second p-value of near 0). This means that the model does not fit the data, though it doesn't fit slightly better than a null model would. We can then compare this with the results of b), where we also saw that overall the model did not fit the data well, though the addition of *temp* was deemed to be significant. Interestingly, we see from the summary tables of the two models that the estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ coefficients are the same, as are their associated values for standard error, and their z-values. I'm not exactly sure how to interpret this, but it seems to be further evidence that the two models are very similar.