Homework 4 - Chapter 3 Question 2

Nickhil Sethi

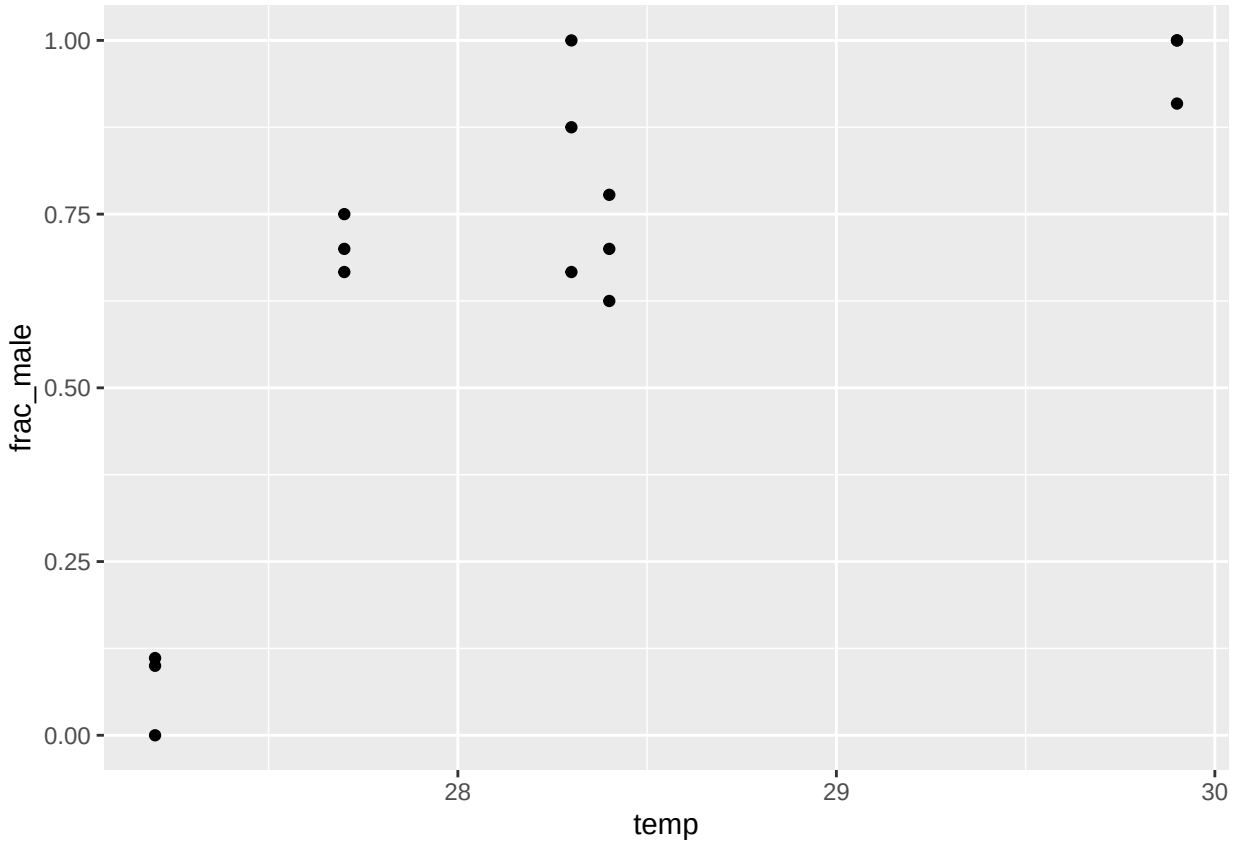
Load pima dataset:

```
data("turtle")
?turtle
```

```
## Incubation temperature and the sex of turtles
##
## Description:
##
##      Incubation temperature and the sex of turtles
##
## Usage:
##
##      data(turtle)
##
## Format:
##
##      A data frame with 15 observations on the following 3 variables.
##
##      temp temperature in degrees centigrade
##
##      male number of male turtles hatched
##
##      female number of female turtles hatched
##
## Details:
##
##      Incubation temperature can affect the sex of turtles. There are 3
##      independent replicates for each temperature.
##
## Source:
##
##      Beyond Traditional Statistical Methods Copyright 2000 D. Cook, P.
##      Dixon, W. M. Duckworth, M. S. Kaiser, K. Koehler, W. Q. Meeker and
##      W. R. Stephenson. Developed as part of NSF/ILI grant DUE9751644.
##
## Examples:
##
##      data(turtle)
```

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

```
turtle$frac_male <- turtle$male / (turtle$male + turtle$female)
ggplot(data = turtle, mapping = aes(temp, frac_male)) + geom_point()
```



Broadly speaking the relationship appears to be a positive correlation, with the majority of the relationship being driven by three high leverage points at the lower end of the temperature axis. The relationship could even be non-linear.

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

```
lmod <- glm(cbind(male, female) ~ temp, family=binomial, turtle)
sumary(lmod)
```

```
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -61.31832   12.02237 -5.1004 3.390e-07
## temp          2.21103    0.43085  5.1317 2.871e-07
##
## n = 15 p = 2
## Deviance = 24.94249 Null Deviance = 74.50804 (Difference = 49.56555)
```

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

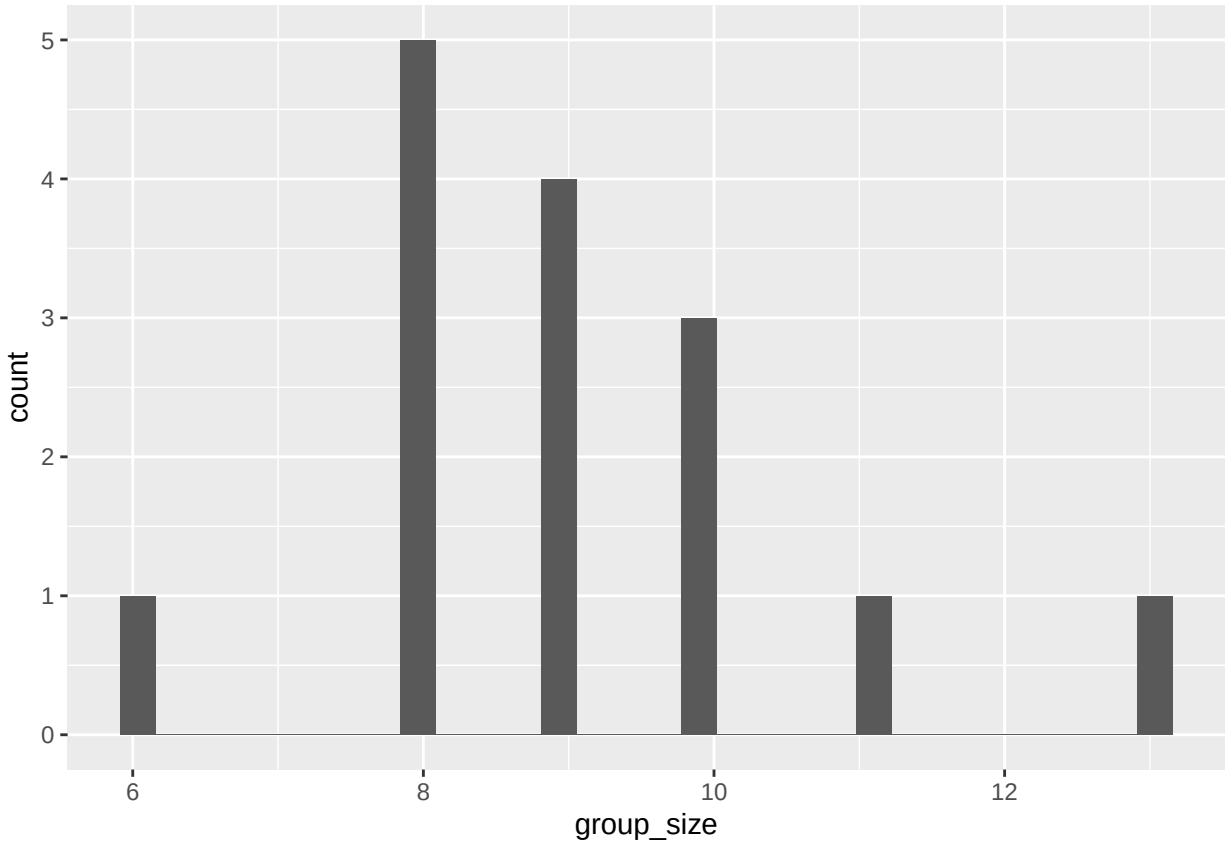
```
## [1] 0.02348863
```

Under the assumption that the model is correct, the deviance is distributed as $\chi^2(n-q-1)$; we see that the p-value of the deviance observed is very low, and thus the model is not a good fit.

(c) Is this data sparse?

```
turtle$group_size <- turtle$male + turtle$female
ggplot(data = turtle, aes(x=group_size), binwidth=1) + geom_histogram()
```

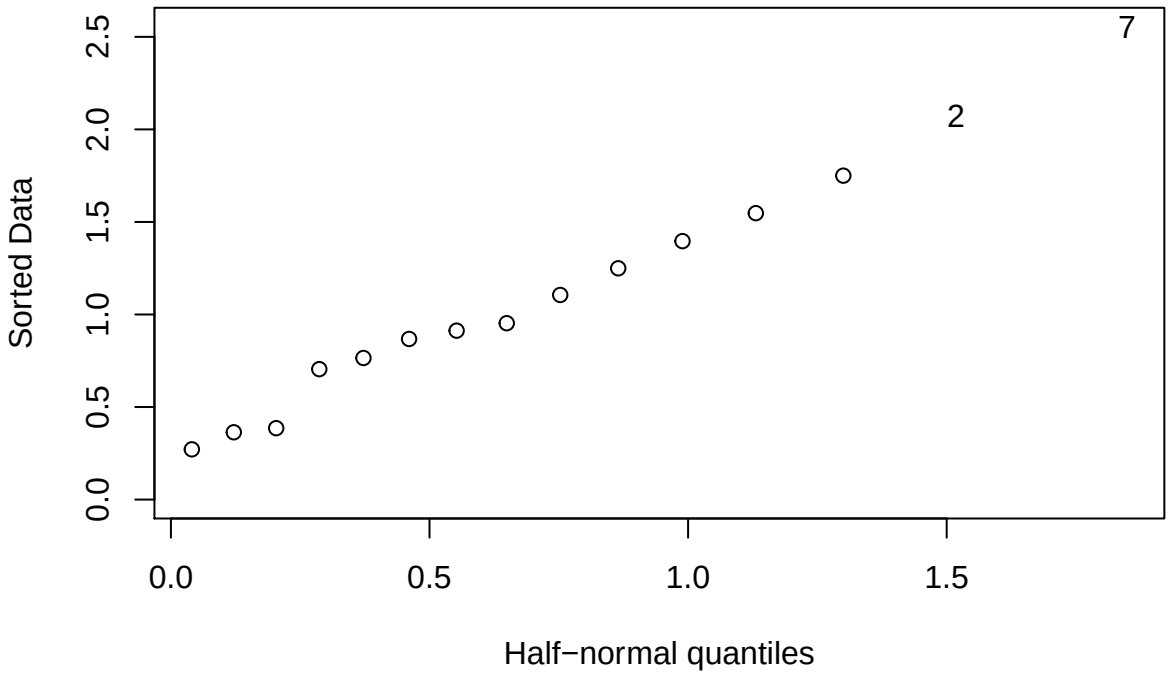
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The data does not seem to be sparse – the group sizes hover around 8 or higher.

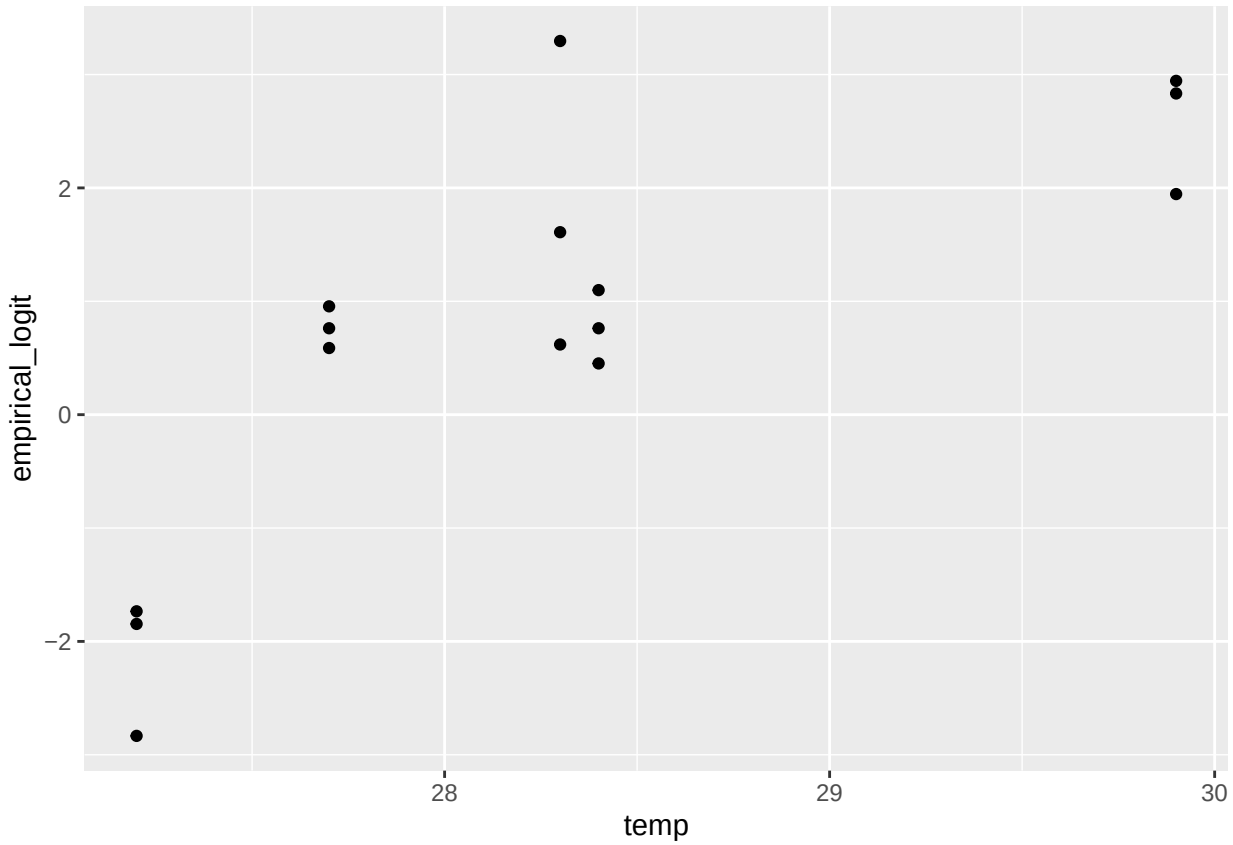
(d) Check for outliers.

```
halfnorm(residuals(lmod))
```



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

```
turtle$empirical_logit <- log((turtle$male + .5) / (turtle$female + .5))
ggplot(data = turtle, mapping = aes(x = temp, y = empirical_logit)) + geom_point()
```



We see strong non-linearity here, indicating a lack of fit. Empirical logit is an estimate of the log odds, which should be linear with respect to the inputs. However, we do not see a linear relationship here.

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

```
turtle$temp_sq <- turtle$temp^2
lmod2 <- glm(cbind(male, female) ~ temp + temp_sq, family=binomial, turtle)
sumary(lmod2)
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -677.59495  268.79836 -2.5208  0.01171
## temp          45.91727   18.91693  2.4273  0.01521
## temp_sq       -0.77451    0.33267 -2.3282  0.01990
##
## n = 15 p = 3
## Deviance = 20.25621 Null Deviance = 74.50804 (Difference = 54.25183)
```

```
pchisq(deviance(lmod2), df.residual(lmod2), lower=FALSE)
```

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

This term is significant in the model, with the coefficient on *temp_sq* having a p-value of .019. The model fit is now slightly better than the others, but still not a great fit with a p-value of .06

Bonus Question (2 points)

(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?

```
turtle_by_temp <- group_by(turtle, temp)
turtle_by_temp %>%
  summarise(GroupVariance=var(frac_male))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

temp	GroupVariance
27.2	0.0037449
27.7	0.0017593
28.3	0.0283565
28.4	0.0058359
29.9	0.0027548

(h) If the three replicates are homogenous, they could be combined so that the dataset would have only five cases in total. Create this dataset and fit a model linear in temperature. Compare the fit seen for this model with that found in (b).