Homework 4 - Chapter 3 Question 2

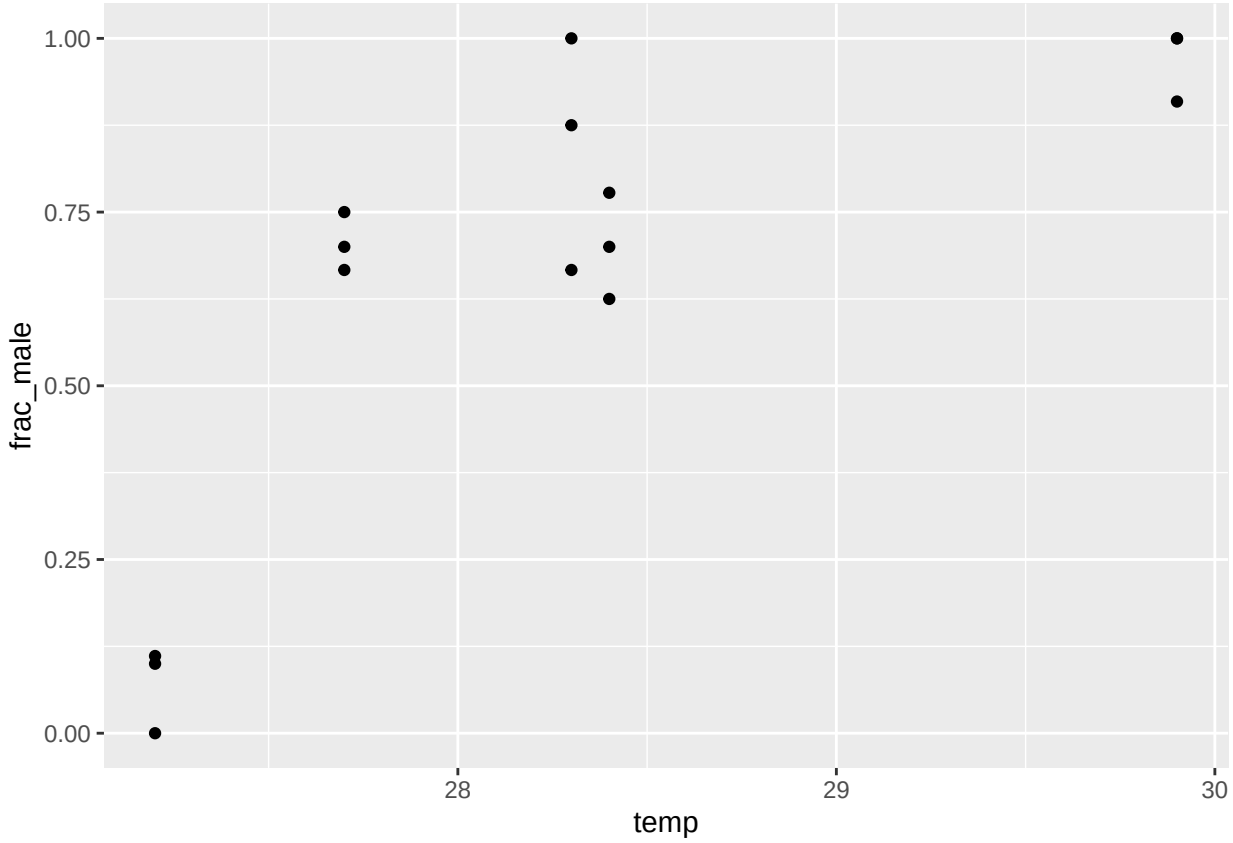
Nickhil Sethi

Load turtle dataset:

```
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)
```

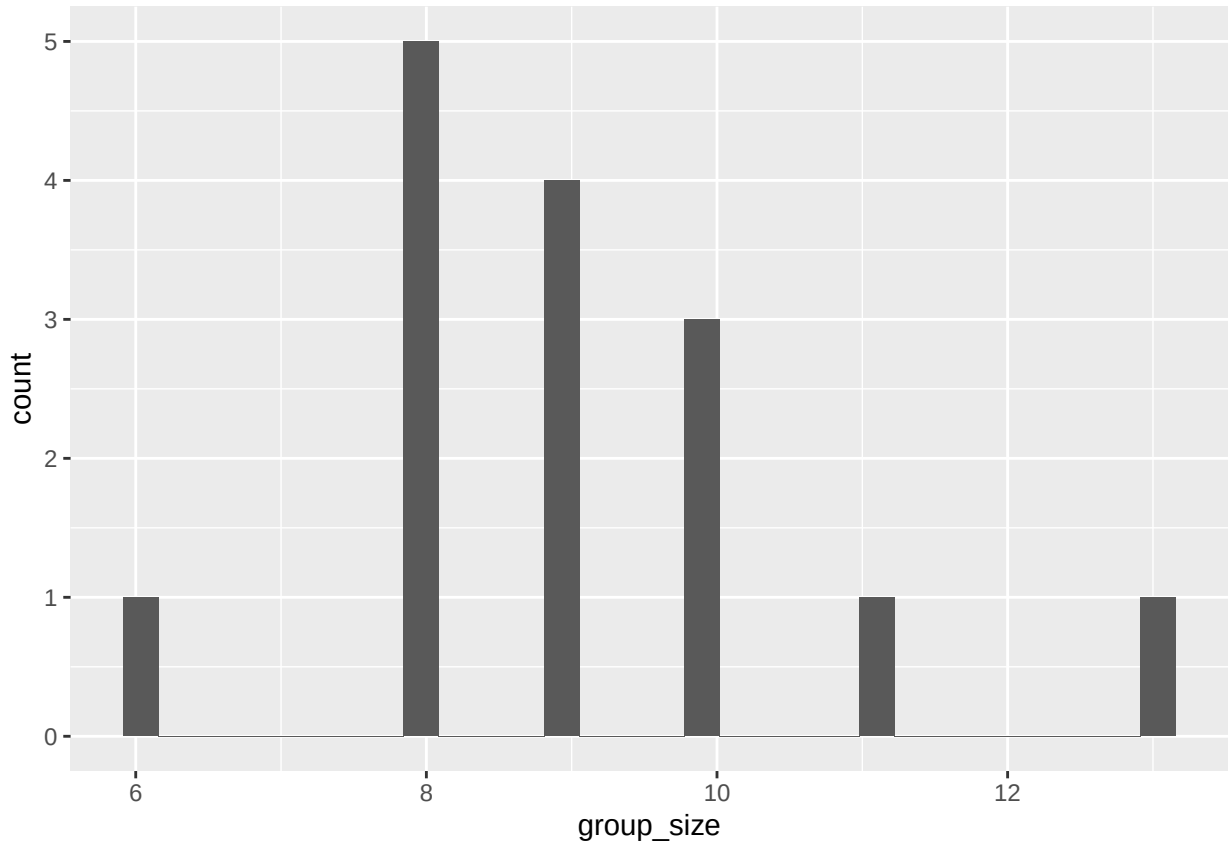
```
## [1] 0.02348863
```

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()
```

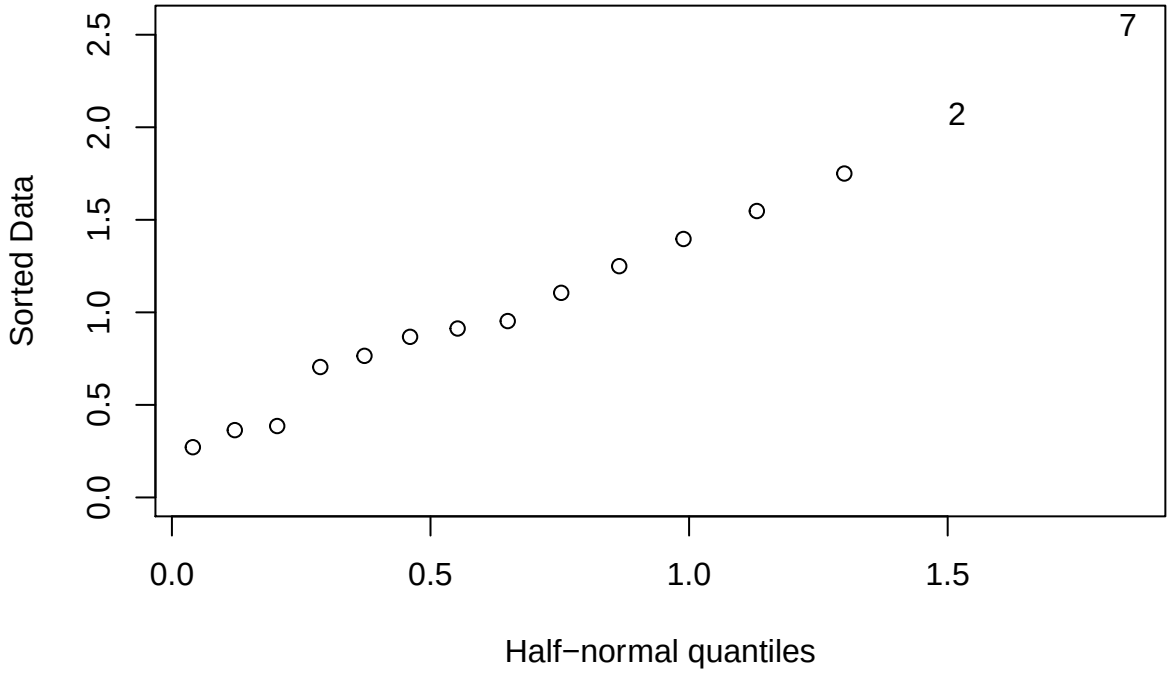
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

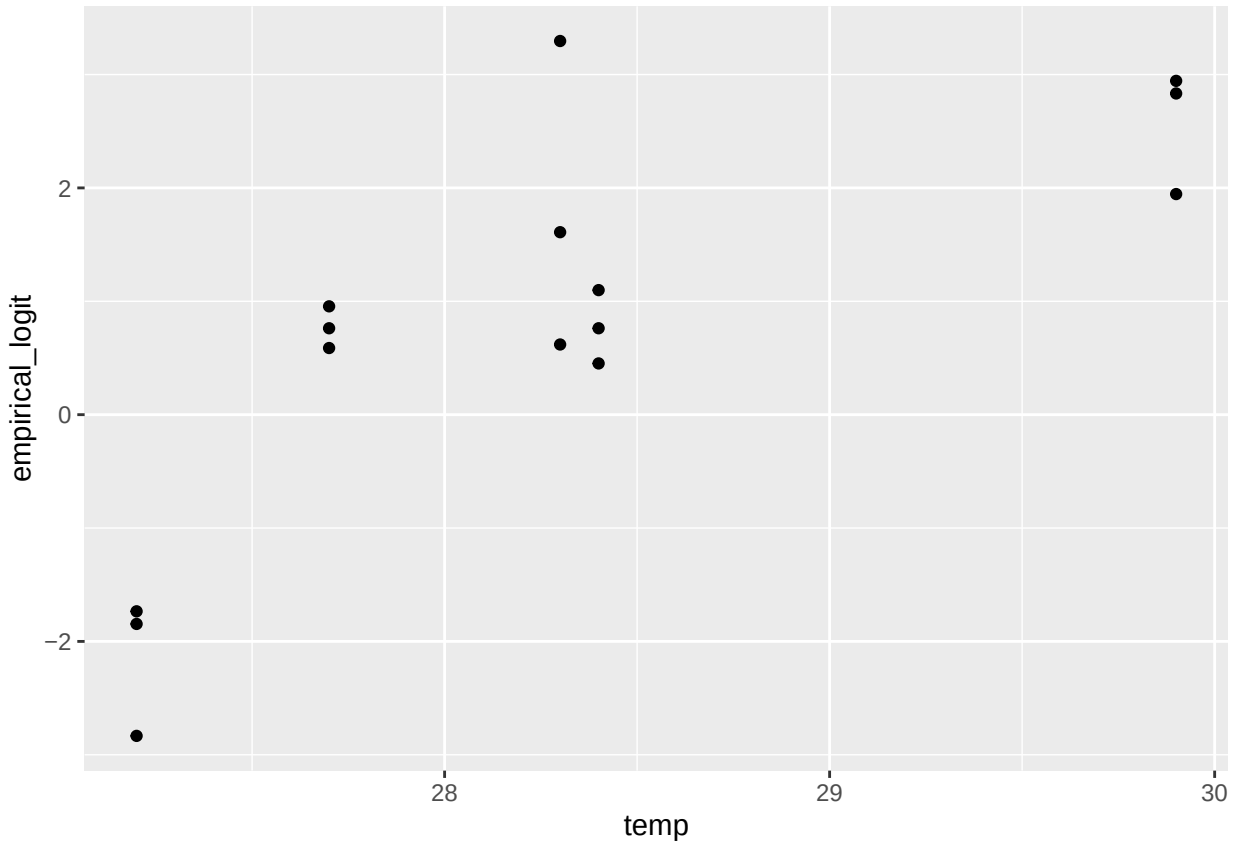
(d) Check for outliers.

```
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; however, it does pass significance at the level of .05.

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?

As the question above outlines, there are three replicates per temperature, each with a different group size. For each temperature, we can compute the empirical variance of the observed proportion $frac\_male$ over the three replicates.

As for the theoretical variance, we can derive a formula for this as follows; for each replicate in a temperature class, let $\hat{p}$ be the observed proportion of males, Y be the number of males observed and m be the total group size. Given the random pair (Y, m), clearly $\hat{p} = Y/m$, and so by the conditional variance formula we have:

$$\begin{aligned}
\text{var}(\hat{p}) &= \text{var}(Y/m) \\
&= \text{E}[\text{var}(\text{Y/m|m})] + \text{var}[\text{E}(\text{Y/m|m})] \\
&= \text{E}[\text{var}(Y/m|m)] + 0 \\
&= \text{E}[\frac{1}{m^2}\text{var}(Y|m)] \\
&= \text{E}[p * (1-p)/m]
\end{aligned}$$

```
turtle$est_prob <- predict(lmod2, type="response")

turtle_by_temp <- group_by(turtle, temp)
turtle_by_temp %>%
  summarise(
    EmpiricalGroupVariance=var(frac_male),
    TheoreticalGroupVariance=(
      mean((1 / (male + female)) * est_prob * (1 - est_prob))
    ))
```

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

temp	EmpiricalGroupVariance	TheoreticalGroupVariance
27.2	0.0037449	0.0150583
27.7	0.0017593	0.0326257
28.3	0.0283565	0.0149041
28.4	0.0058359	0.0139597
29.9	0.0027548	0.0053293

If the above calculations are correct, the empirical group variance is systematically lower than the theoretical variance we would expect, perhaps indicating under-dispersion.

(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).

```
turtle_by_temp <- group_by(turtle, temp)
turtle_temp_sum <- summarize(turtle_by_temp, male = sum(male), female = sum(female))
```

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

```
# aggregated model
lmod_agg <- glm(cbind(male, female) ~ temp, family=binomial, turtle_temp_sum)
sumary(lmod_agg)
```

```
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -61.31832   12.02240 -5.1003 3.390e-07
## temp          2.21103    0.43086  5.1317 2.871e-07
##
## n = 5 p = 2
## Deviance = 14.86295 Null Deviance = 64.42850 (Difference = 49.56555)
```

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

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

```
# previous model in (b)
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
```

The aggregated model is actually a far worse fit than the previous model, with a p-value of .002. It's potentially worth noting that the aggregation reduces the number of data points in the model and thus reduces the significance, making it difficult to compare the two.