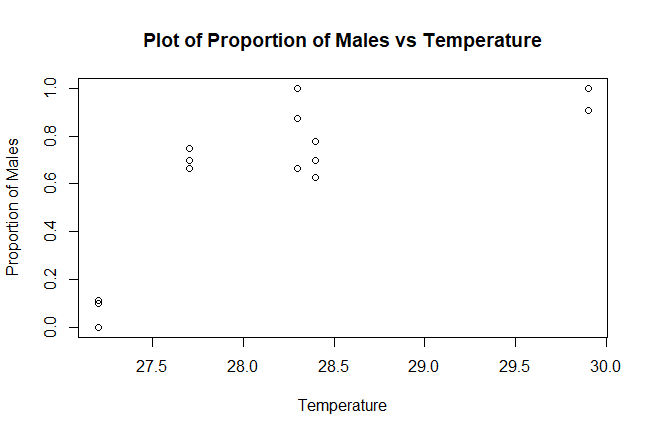
**STAT34700-HW2**

**Problem 2**

**(a)**



From the plot we can see that under higher incubation temperature the proportion of males is larger, which mean that under higher incubation temperature, more male turtles born. In addition, we can see a roughly non-linear but near quadratic trend in this plot.

**(b)**

We fit a binomial GLM with *male (number of male turtle born)* as the binomial response, *Temp (incubation temperature)* as a numerical predictor. In this generalized linear model, we use logit link function and binomial distribution family.

> fit\_2b = glm(cbind(male,female)~temp,family=binomial,data=turtle)

**The GLM formula:**

**Output:**

> sumary(fit\_2b)

Estimate Std. Error z value Pr(>|z|)

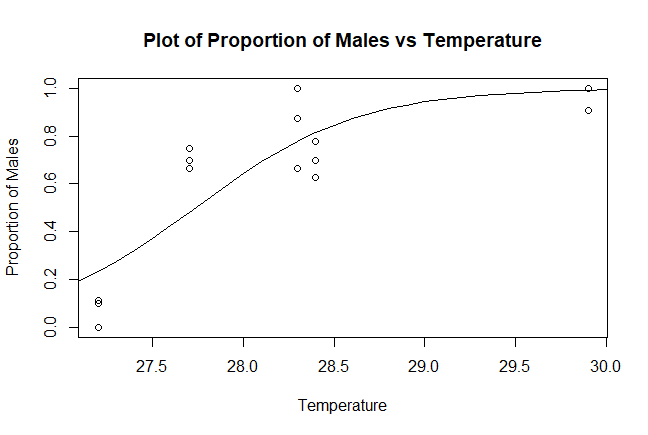
(Intercept) -61.31832 12.02237 -5.1004 3.390e-07

temp 2.21103 0.43085 5.1317 2.871e-07

n = 15 p = 2

Deviance = 24.94249 Null Deviance = 74.50804 (Difference = 49.56555)

After fitting the binomial GLM, we plot the fitted value on the plot of proportion of males vs temperature:



From this plot, we can say that the model not fit the data well.

**Test Good-of-fit:**

Provided that the response is binomial distributed and that each group size is relatively large, the deviance is approximately Chi-squared distributed with 13 degrees of freedom if the model is

correct. Thus, we can use the deviance to test whether the model is an adequate fit.

**Output:**

> pchisq(deviance(fit\_2b),df.residual(fit\_2b),lower=FALSE)

[1] 0.02348863

**Conclusion:**

The null hypothesis is that the model has a good fit on the data. Because the p-value is smaller than significance level 0.05, then we can reject the null hypothesis and conclude that the model is not an adequate fit.

**(c)**

We use assume a binomial distribution on the response, then to make the deviance follow an approximately Chi-squared distribution, each group sample size should be greater than or equal to 5. If the group sample size is less than 5, we say that the data is sparse.

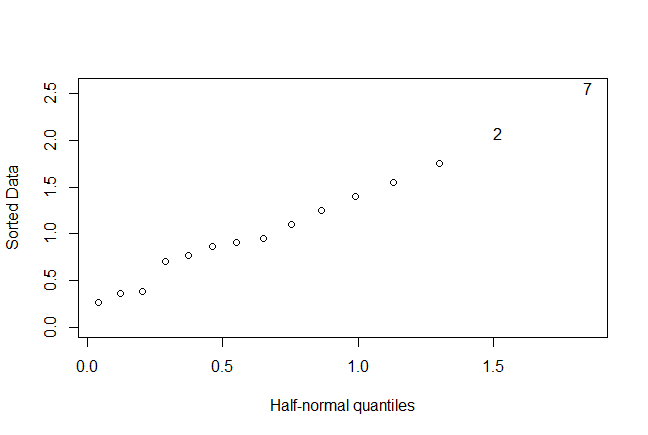
In this data, for each observation, the total number of turtle born is greater than 5. As a result, the data here is nor sparse.

**Conclusion:**

The data is not sparse.

**(d)**

To check outliers in this data, we plot the half-normal plot:



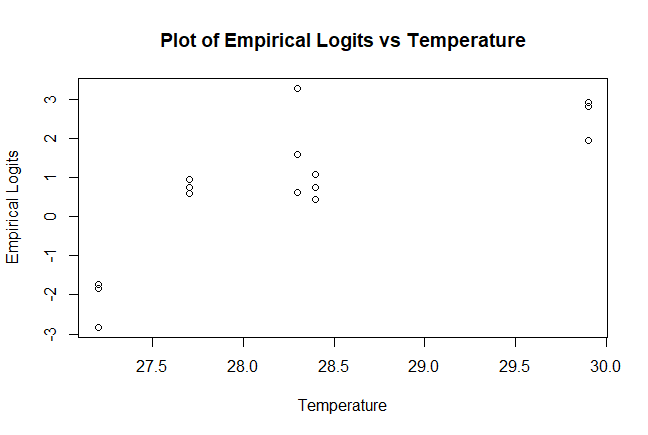
From the Half-normal plot, we can see that there is no obvious outliers in this data because there is no big distinction between data point 2, 7 and other points.

**Conclusion:**

There is no obvious outlier.

**(e)**

First, we compute the empirical logits and plot these against temperature.



The empirical logits = .

From this plot, we can see a nonlinear and near quadratic trend. If the binomial GLM fit the data well, the empirical logit should have a linear relationship with incubation temperature. As a result, the quadratic trend between empirical logits and incubation temperature indicates a lack of fit for binomial GLM.

**Conclusion:**

The plot indicates a lack of fit for binomial GLM.

**(f)**

We fit a binomial GLM with *male (number of male turtle born)* as the binomial response, *Temp (incubation temperature)* and *Temp^2* as two numerical predictors. In this generalized linear model, we use logit link function and binomial distribution family.

> fit\_2f = glm(cbind(male,female)~ temp + I(temp^2),family=binomial,turtle)

**The GLM formula:**

**Output:**

> sumary(fit\_2f)

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

I(temp^2) -0.77451 0.33267 -2.3282 0.01990

n = 15 p = 3

Deviance = 20.25621 Null Deviance = 74.50804 (Difference = 54.25183)

**Test significance of quadratic term:**

To test the significance of the quadratic term, we can use the deviance to compare two models: model 1 with quadratic term and model 2 without quadratic term.

The likelihood test statistic is Deviance of model 2 – Deviance of model 1, which asymptotically Chi-squared distributed with df = 1 (assuming that the model 2 is correct and the distributional assumptions hold).

**Output:**

> pchisq(deviance(fit\_2b)-deviance(fit\_2f),1,lower=FALSE)

[1] 0.03040446

**Conclusion:**

The null hypothesis is that the model without quadratic term is correct. Because the p-value is smaller than significance level 0.05, then we can reject the null hypothesis and conclude that the model with quadratic term is correct. Hence, the quadratic term is a significant predictor to the response.

**Test Good-of-fit:**

Provided that the response is binomial distributed and that each group size is relatively large, the deviance is approximately Chi-squared distributed with 12 degrees of freedom if the model is

correct. Thus, we can use the deviance to test whether the model is an adequate fit.

**Output:**

> pchisq(deviance(fit\_2f),df.residual(fit\_2f),lower=FALSE)

[1] 0.06239194

**Conclusion:**

The null hypothesis is that the model with quadratic term has a good fit on the data. Because the p-value is greater than significance level 0.05, then we cannot reject the null hypothesis and conclude that the model with quadratic term is an adequate fit.

**(g)**

**Calculate the observed variation:**

For each incubation temperature i, there are three proportions observed:

The mean of these proportions is

The observed variation of these proportions is

Assuming independent binomial variation, **the expected variation** is

By using R, we can get the observed variation and expected variation of proportions for each temperature:

**Output:**

> var\_ex

[1] 0.015058343 0.032625718 0.014904138 0.013959688 0.005329255

> var\_ob

[1] 0.003744856 0.001759259 0.028356481 0.005835905 0.002754821

**Conclusion:**  
From the results, we can see that there is only one observed variation greater than the expected variation. For other temperatures, the observed variations are much less than the expected variations. As a result, there is no obvious evidence of greater variation and these two variations not approximately agree with each other.

**(h)**

Combine the three replicates under each incubation temperature by adding up the male, female and total values and recalculate the proportions of males.

The new data frame is:

| **temp** | **male** | | **female** | | **total** | | **proportion** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | |  | |  | |  |
| **1** | 27.2 | 2 | | 25 | | 27 | | 0.07407407 |
| **2** | 27.7 | 17 | | 7 | | 24 | | 0.70833333 |
| **3** | 28.3 | 26 | | 4 | | 30 | | 0.86666667 |
| **4** | 28.4 | 19 | | 8 | | 27 | | 0.70370370 |
| **5** | 29.9 | 27 | | 1 | | 28 | | 0.96428571 |

We fit a binomial GLM with *male (number of male turtle born)* as the binomial response, *Temp (incubation temperature)* as numerical predictor. In this generalized linear model, we use logit link function and binomial distribution family.

> fit\_2h = glm(cbind(male,female) ~ temp, family=binomial, data=newturtle)

**The GLM formula:**

**Output:**

> sumary(fit\_2h)

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)

Compared with the fit in part (b), the estimated coefficients, standard errors of coefficients, corresponding z-values and p-values in these two fits are all the same. The deviances, degree of freedom and sample sizes of these two fits are different.

**Test Good-of-fit:**

Provided that the response is binomial distributed and that each group size is relatively large, the deviance is approximately Chi-squared distributed with 3 degrees of freedom if the model is

correct. Thus, we can use the deviance to test whether the model is an adequate fit.

**Output:**

> pchisq(deviance(fit\_2h),df.residual(fit\_2h),lower=FALSE)

[1] 0.001937595

The null hypothesis is that the model has a good fit on the data. Because the p-value is smaller than significance level 0.05, then we can reject the null hypothesis and conclude that the model is not an adequate fit.

**Conclusion:**

Compared to the p-value of fit in part (b), the p-value here is smaller. Then we are more likely to reject the null hypothesis and conclude that the model does not fit well here. As a result, the fit of this model is worse than the fit of model in part (b).

**R Code**

# Problem 2 --------------------------------------------------------------

install.packages("faraway")

library(faraway)

#a

data(turtle)

View(turtle)

turtle$total = turtle$male + turtle$female

turtle$proportion = turtle$male / turtle$total

plot(turtle$temp, turtle$proportion, main="Plot of Proportion of Males vs Temperature", xlab="Temperature", ylab="Proportion of Males")

#b

fit\_2b = glm(cbind(male,female) ~ temp, family=binomial, data=turtle)

sumary(fit\_2b)

x <- seq(27,30,0.1)

lines(x,ilogit(-61.31832+2.21103\*x))

#good-of-fit test

pchisq(deviance(fit\_2b),df.residual(fit\_2b),lower=FALSE)

#d

halfnorm(residuals(fit\_2b))

#e

elogits = with(turtle,log((male+0.5)/(total- male+0.5)))

plot(turtle$temp,elogits,main="Plot of Empirical Logits vs Temperature", xlab="Temperature", ylab="Empirical Logits")

#f

fit\_2f = glm(cbind(male,female)~ temp + I(temp^2),family=binomial,turtle)

sumary(fit\_2f)

#test the significance of temp^2

pchisq(deviance(fit\_2b)-deviance(fit\_2f),1,lower=FALSE)

#good-of-fit test

pchisq(deviance(fit\_2f),df.residual(fit\_2f),lower=FALSE)

#g

#calculate observed variation

var\_ob = c(var(turtle[1:3,5]),var(turtle[4:6,5]),var(turtle[7:9,5]),var(turtle[10:12,5]),var(turtle[13:15,5]))

#calculate expected variation

for (i in 1:15){

turtle$var[i] = (fitted(fit\_2f)[i]\*(1-fitted(fit\_2f)[i]))/turtle$total[i]

}

var\_ex = NULL

for (i in 1:15){

if (i %% 3 == 1) {

var\_ex = c(var\_ex, 1/3\*(turtle$var[i]+turtle$var[i+1]+turtle$var[i+2]))

}

}

#h

#create new combined data set

newturtle = data.frame(matrix(rep(0,25),nrow=5,ncol=5))

newturtle$X1 = c(27.2,27.7,28.3,28.4,29.9)

male=NULL

for (i in 1:15){

if (i %% 3 ==1){

male = c(male,turtle$male[i]+turtle$male[i+1]+turtle$male[i+2])

}

}

newturtle$X2 = male

total=NULL

for (i in 1:15){

if (i %% 3 ==1){

total = c(total,turtle$total[i]+turtle$total[i+1]+turtle$total[i+2])

}

}

newturtle$X4 = total

newturtle$X3 = newturtle$X4 - newturtle$X2

newturtle$X5 = newturtle$X2 / newturtle$X4

colnames(newturtle)=c("temp","male","female","total","proportion")

#fit the binomial GLM to the new data

fit\_2h = glm(cbind(male,female) ~ temp, family=binomial, data=newturtle)

summary(fit\_2h)

#good-of-fit test

pchisq(deviance(fit\_2h),df.residual(fit\_2h),lower=FALSE)