Homework 4

Isabella Chittumuri

10/15/2020

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 and can be found in the turtle dataset.

```
# Install packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(faraway)
library(tidyverse)
```

```
## -- Attaching packages --------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.0     v stringr 1.4.0
## v tidyr   1.0.2     v forcats 0.5.0
## v readr   1.3.1
```

```
## -- Conflicts -----------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

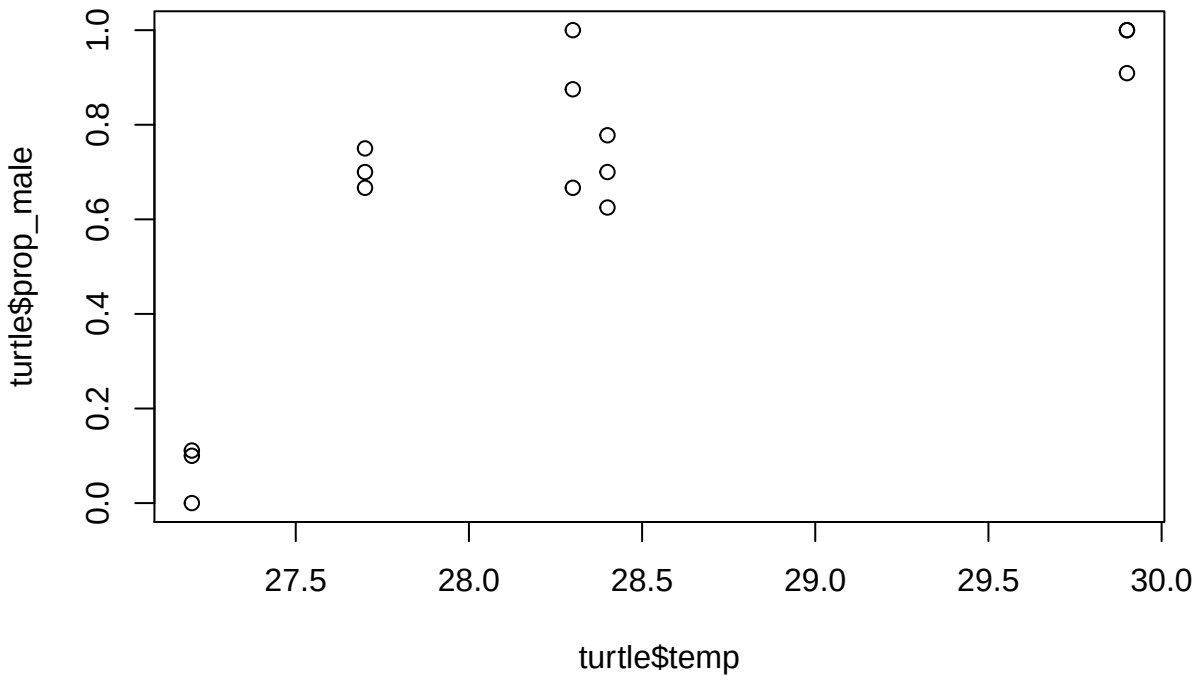
```
# Get dataset
data("turtle")
?turtle
summary(turtle)
```

```
##       temp           male            female
##  Min.   :27.2   Min.   : 0.000   Min.   :0
##  1st Qu.:27.7   1st Qu.: 4.500   1st Qu.:1
##  Median :28.3   Median : 7.000   Median :2
##  Mean   :28.3   Mean   : 6.067   Mean   :3
##  3rd Qu.:28.4   3rd Qu.: 7.500   3rd Qu.:3
##  Max.   :29.9   Max.   :13.000   Max.   :9
```

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

```r
# Get the proportion of males
turtle$prop_male <- ifelse(turtle$female == 0, 1, (turtle$male)/(turtle$male+turtle$female))

# Plot it against temp
plot(turtle$temp, turtle$prop_male)
```



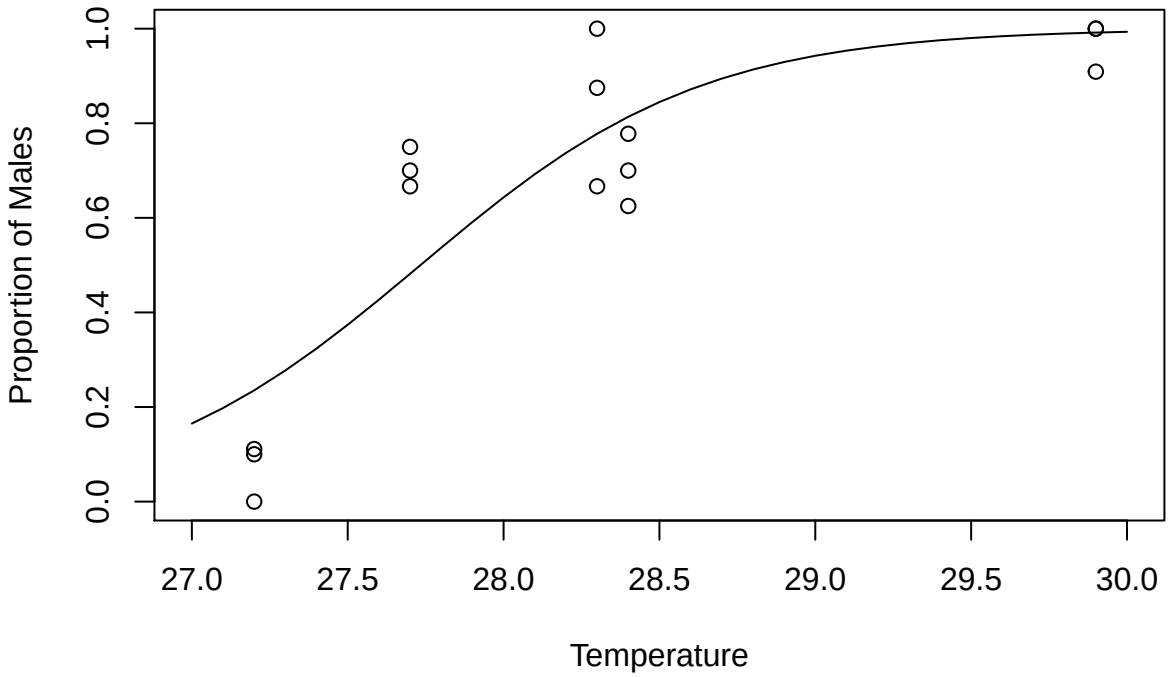
Looking at this graph, we see that the proportion of males hatched increases as temperature increases, even though there isn't a clear linear relationship between the two.

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

```r
# Binomial model
bmod1 <- glm(cbind(male, female) ~ temp, family="binomial", turtle)

# Get temp in .1 increments
x <- seq(27,30, .1)

# Plot the model
plot(prop_male ~ temp, data = turtle, xlim = c(27,30), ylim = c(0,1),
     xlab = "Temperature", ylab = "Proportion of Males")
lines(x, ilogit(-61.3183 + 2.2110 * x))
```



This graph shows the binomial model overlayed with the proportion of males against temperature. Overall the model looks like it fits the data well, regardless of the minor overfitting and underfitting.

```
summary(bmod1)
```

```
##
## 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
```

The null deviance shows how well the response is predicted by the model with nothing but an intercept. The residual deviance shows how well the response is predicted by the model when the predictors are included. The binomial model summary shows that the residual deviance has a value of 24.9, much lower than the null deviance with value of 74.5. This means that model with the predictors preforms better than without the predictors. We can test this using the chi-squared test of the difference between the deviances of the binomial

model and the null model.

```
# Chi-squared test, p-value
pchisq(deviance(bmod1), df.residual(bmod1), lower.tail = FALSE)
```

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

The chi-squared test gives us a p-value of 0.024, which is less than our alpha level of .05. This suggests that something is wrong with the model because the deviance is not following a Chi-squared distribution.

(c) Is this data sparse?

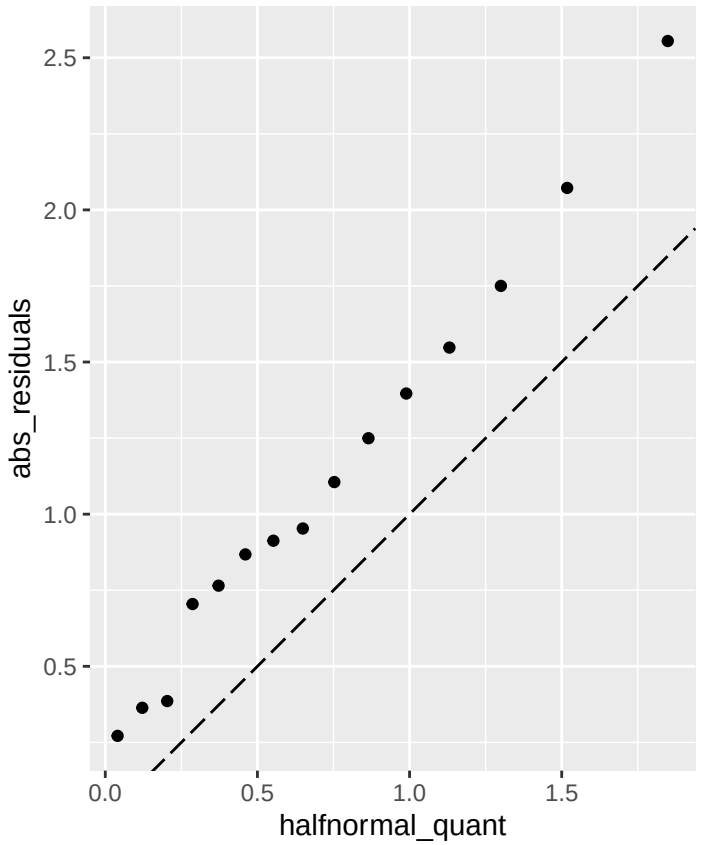
```
head(turtle)
```

```
##   temp male female prop_male
## 1 27.2    1      9 0.1000000
## 2 27.2    0      8 0.0000000
## 3 27.2    1      8 0.1111111
## 4 27.7    7      3 0.7000000
## 5 27.7    4      2 0.6666667
## 6 27.7    6      2 0.7500000
```

No, the data is not sparse because there are more than 5 counts per trial.

(d) Check for outliers.

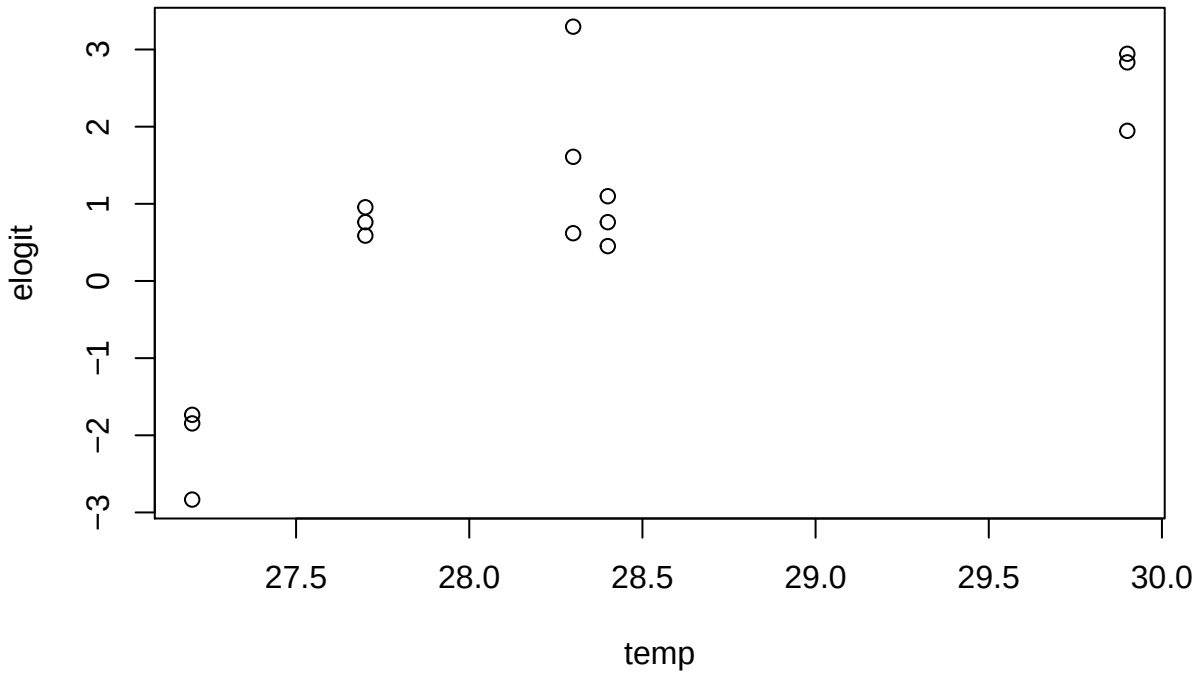
```
# Half normal plot
residuals <- residuals(bmod1)
x <- abs(residuals)
labord <- order(x)
x <- sort(x)
i <- order(x)
n <- length(x)
halfnormal_quant <- qnorm((n + 1:n)/(2 * n + 1))
data.frame(
  halfnormal_quant = halfnormal_quant,
  abs_residuals = x[i]
) %>%
  ggplot(aes(x = halfnormal_quant, y = abs_residuals)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "longdash") +
  coord_equal()
```



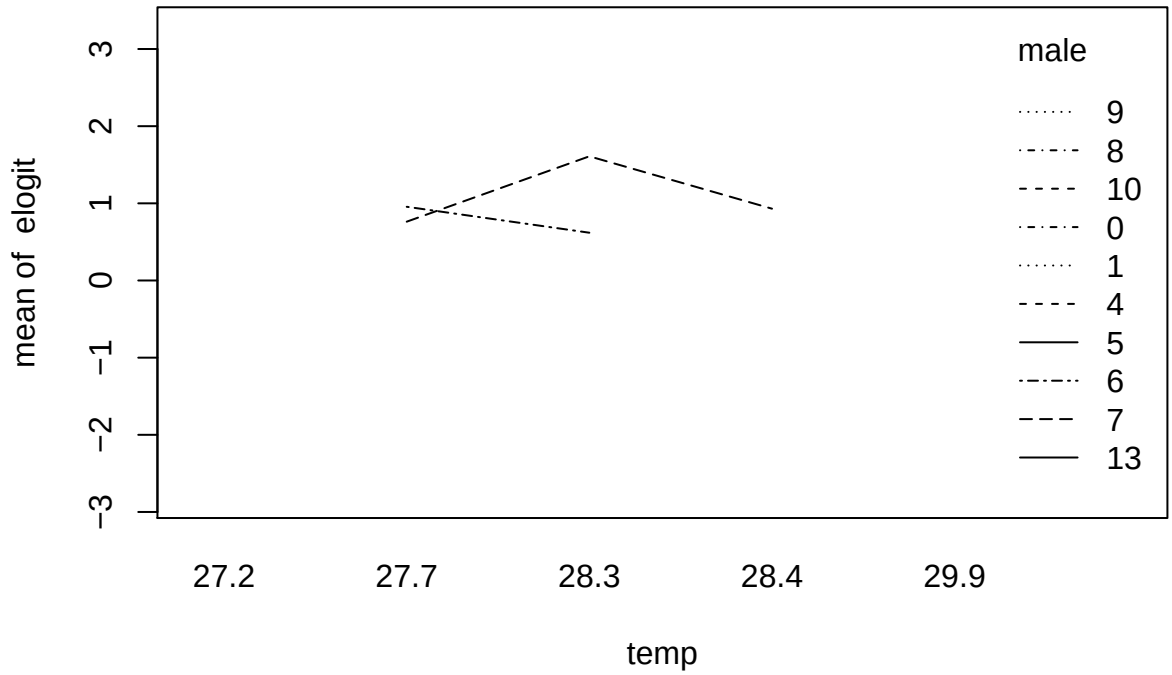
This plot shows a half normal distribution. The absolute value of the residuals is on a relatively straight linear line, adjacent to the half normal distribution line (dashed line). Based on this plot, there are no apparent outliers.

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

```
# Interaction plot
turtle['elogit'] <- with(turtle,log((male + 0.5)/(female + 0.5)))
plot(elogit~temp, turtle)
```

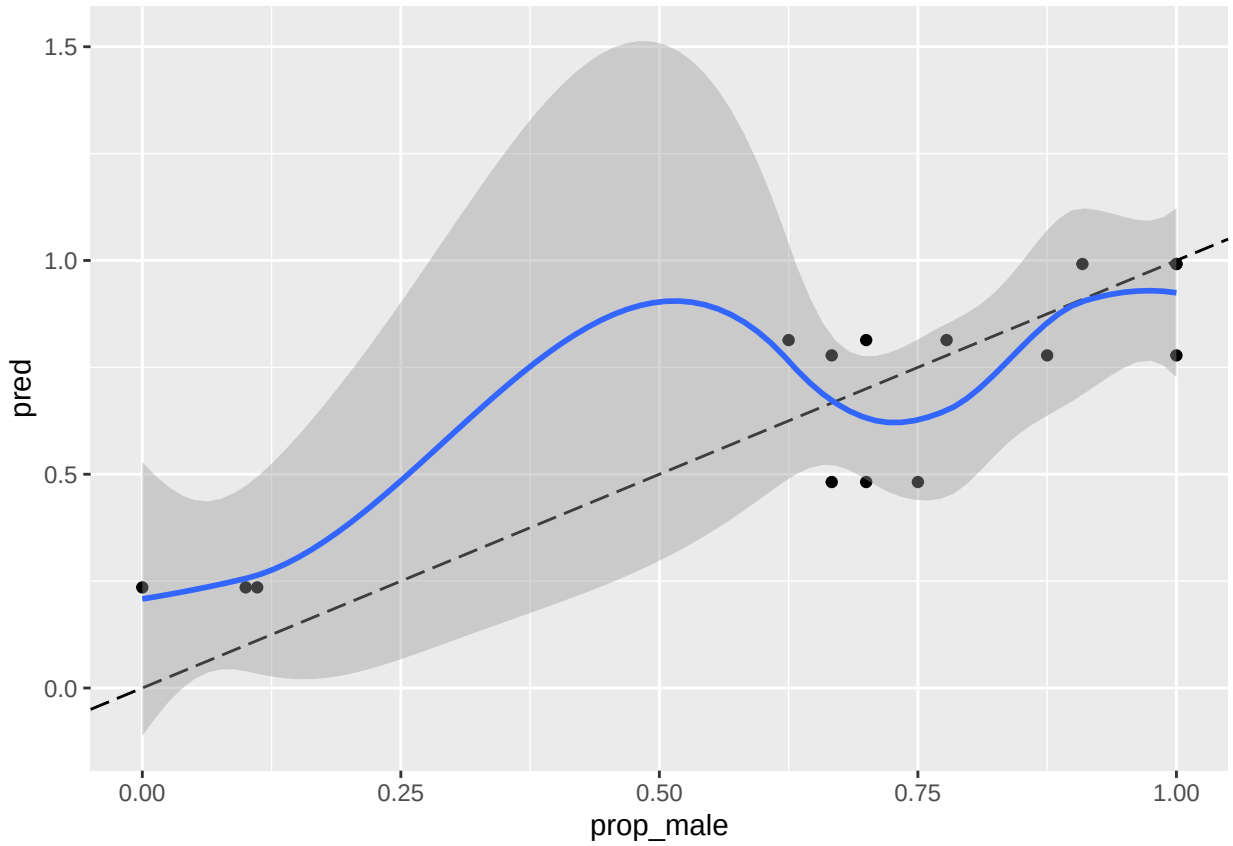



```
with(turtle,interaction.plot(temp,male,elogit))
```




```
# Predict function shows the predictions of the model
turtle %>%
  mutate(pred = predict(bmod1, ., type = "response")) %>%
  ggplot(aes(x = prop_male, y = pred)) +
  geom_point() + geom_abline(slope = 1, intercept = 0, linetype = "longdash") +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

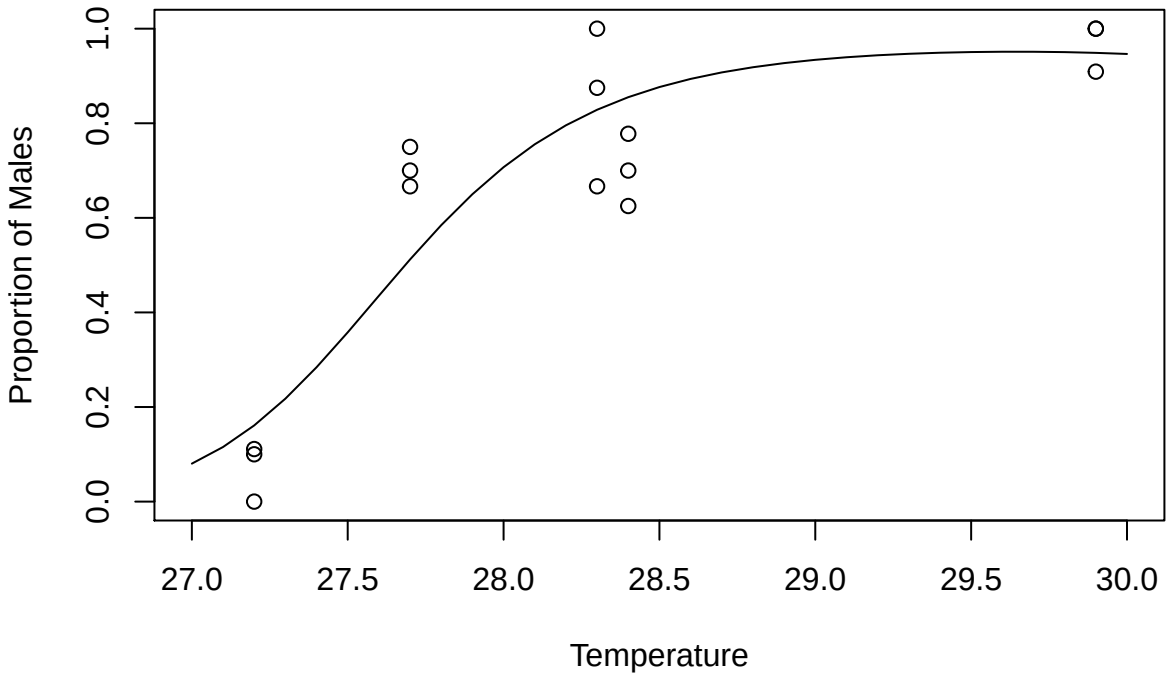


The first plot shows the empirical logits of the proportion of males against temperature. This graph is similar to the our graph in part (a). The second plot shows the interaction between number of males against that of the empirical logit. There seems to be something weird going on in this graph, but it doesn't indicate a lack of fit.

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

```
# Binomial model using quadratic term in temp
bmod2 <- glm(cbind(male, female) ~temp + I(temp^2), family="binomial", turtle)

# Plot the model using temp^2
x <- seq(27,30, .1)
plot(prop_male ~ temp, data = turtle, xlim = c(27,30), ylim = c(0,1),
     xlab = "Temperature", ylab = "Proportion of Males")
lines(x, ilogit(-677.595 +(45.9173*x)- (.7745 * x^2)))
```




```
summary(bmod2)
```

```
##
## 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
```

This quadratic model's summary, we see that the additional term (temp^2) has the same level of significance as the intercept and temperature. It also shows that the residual deviance has a value of 20.25, much lower than the null deviance with value of 74.5. Note, this is also lower than the previous model that had a residual deviance of 24.9. This means that the quadratic model fits the data better. Again, we can test this using the chi-squared test of the difference between the deviances of the binomial model and the null model.

```r
# Compare models using anova test
anova(bmod2, bmod1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(male, female) ~ temp + I(temp^2)
## Model 2: cbind(male, female) ~ temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        12     20.256
## 2        13     24.942 -1  -4.6863   0.0304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Check for significance using Chi-squared deviance test
pchisq(deviance(bmod2), df.residual(bmod2), lower.tail = FALSE)
```

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

In the anova test, we see that the p-value is less than .05, meaning that we are improving the model by adding (temp^2). In the Chi-squared deviance test, we see that the p-value is greater than 0.05, meaning that this model is following a Chi-squared distribution. Therefore, the quadratic model fits the data.