R_Activity_Assignment_9

Dahee Ahn

2024-11-01

```
rm(list=ls())
```

5.1 Part A

Let's pretend that commercial Chilean bass cannot have a Mercury concentration higher than 0.354 (PPM). Fish above this level will be considered contaminated (and thus not marketable), whereas fish below the level will not be contaminated (= marketable). Let's determine the probability of a fish being marketable as a function of the pH of the lake they were sourced from. This is not my area of expertise, but we are hypothesizing here that pH levels might moderate levels of mercury contamination.

1. Load in the bass.txt dataset.

```
bass <- read.table("C:/Users/chemk/Desktop/Classes/ENT6707_DataAnalysis/week11/bass.txt", header
=TRUE, sep="\t")
nrow(bass)
```

```
## [1] 53
```

```
str(bass)
```

```
## 'data.frame':    53 obs. of  12 variables:
##  $ ID               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Lake             : chr  "Alligator" "Annie" "Apopka" "Blue Cypress" ...
##  $ Alkalinity       : num  5.9 3.5 116 39.4 2.5 19.6 5.2 71.4 26.4 4.8 ...
##  $ pH               : num  6.1 5.1 9.1 6.9 4.6 7.3 5.4 8.1 5.8 6.4 ...
##  $ Calcium          : num  3 1.9 44.1 16.4 2.9 4.5 2.8 55.2 9.2 4.6 ...
##  $ Chlorophyll      : num  0.7 3.2 128.3 3.5 1.8 ...
##  $ AvgMercury       : num  1.23 1.33 0.04 0.44 1.2 0.27 0.48 0.19 0.83 0.81 ...
##  $ NumSamples       : int  5 7 6 12 12 14 10 12 24 12 ...
##  $ MinMercury       : num  0.85 0.92 0.04 0.13 0.69 0.04 0.3 0.08 0.26 0.41 ...
##  $ MaxMercury       : num  1.43 1.9 0.06 0.84 1.5 0.48 0.72 0.38 1.4 1.47 ...
##  $ ThreeYrStdMercury: num  1.53 1.33 0.04 0.44 1.33 0.25 0.45 0.16 0.72 0.81 ...
##  $ AgeData          : int  1 0 0 0 1 1 1 1 1 1 ...
```

```
head(bass)
```

```
##   ID        Lake Alkalinity  pH Calcium Chlorophyll AvgMercury NumSamples
## 1  1   Alligator        5.9 6.1     3.0         0.7       1.23          5
## 2  2       Annie        3.5 5.1     1.9         3.2       1.33          7
## 3  3      Apopka      116.0 9.1    44.1       128.3       0.04          6
## 4  4 Blue Cypress       39.4 6.9    16.4         3.5       0.44         12
## 5  5       Brick        2.5 4.6     2.9         1.8       1.20         12
## 6  6      Bryant       19.6 7.3     4.5        44.1       0.27         14
##   MinMercury MaxMercury ThreeYrStdMercury AgeData
## 1       0.85       1.43              1.53       1
## 2       0.92       1.90              1.33       0
## 3       0.04       0.06              0.04       0
## 4       0.13       0.84              0.44       0
## 5       0.69       1.50              1.33       1
## 6       0.04       0.48              0.25       1
```

```
tail(bass)
```

```
##    ID        Lake Alkalinity  pH Calcium Chlorophyll AvgMercury NumSamples
## 48 47    Trafford       81.5 8.9    20.5         9.6       0.27          6
## 49 48       Trout        1.2 4.3     2.1         6.4       0.94         10
## 50 49 Tsala Apopka       34.0 7.0    13.1         4.6       0.40         12
## 51 50        Weir       15.5 6.9     5.2        16.5       0.43         11
## 52 52     Wildcat       17.3 5.2     3.0         2.6       0.25         12
## 53 53        Yale       71.8 7.9    20.5         8.8       0.27         12
##    MinMercury MaxMercury ThreeYrStdMercury AgeData
## 48       0.04       0.40              0.27       0
## 49       0.59       1.24              0.98       1
## 50       0.08       0.90              0.31       1
## 51       0.23       0.69              0.43       1
## 52       0.15       0.40              0.28       1
## 53       0.15       0.51              0.25       1
```

```
summary(bass)
```

```
##        ID          Lake              Alkalinity          pH       
##  Min.   : 1   Length:53          Min.   :  1.20   Min.   :3.600  
##  1st Qu.:14   Class :character   1st Qu.:  6.60   1st Qu.:5.800  
##  Median :27   Mode  :character   Median : 19.60   Median :6.800  
##  Mean   :27                      Mean   : 37.53   Mean   :6.591  
##  3rd Qu.:40                      3rd Qu.: 66.50   3rd Qu.:7.400  
##  Max.   :53                      Max.   :128.00   Max.   :9.100  
##     Calcium      Chlorophyll       AvgMercury       NumSamples   
##  Min.   : 1.1   Min.   :  0.70   Min.   :0.0400   Min.   : 4.00  
##  1st Qu.: 3.3   1st Qu.:  4.60   1st Qu.:0.2700   1st Qu.:10.00  
##  Median :12.6   Median : 12.80   Median :0.4800   Median :12.00  
##  Mean   :22.2   Mean   : 23.12   Mean   :0.5272   Mean   :13.06  
##  3rd Qu.:35.6   3rd Qu.: 24.70   3rd Qu.:0.7700   3rd Qu.:12.00  
##  Max.   :90.7   Max.   :152.40   Max.   :1.3300   Max.   :44.00  
##    MinMercury       MaxMercury     ThreeYrStdMercury    AgeData      
##  Min.   :0.0400   Min.   :0.0600   Min.   :0.0400    Min.   :0.0000  
##  1st Qu.:0.0900   1st Qu.:0.4800   1st Qu.:0.2500    1st Qu.:1.0000  
##  Median :0.2500   Median :0.8400   Median :0.4500    Median :1.0000  
##  Mean   :0.2798   Mean   :0.8745   Mean   :0.5132    Mean   :0.8113  
##  3rd Qu.:0.3300   3rd Qu.:1.3300   3rd Qu.:0.7000    3rd Qu.:1.0000  
##  Max.   :0.9200   Max.   :2.0400   Max.   :1.5300    Max.   :1.0000  
```

2. Using R, create a new column called marketable in which each observation of marketable is a 1 when AvgMercury is less than 0.354 and 0 otherwise (make sure marketable is numeric!).

```r
library(tidyverse)
```

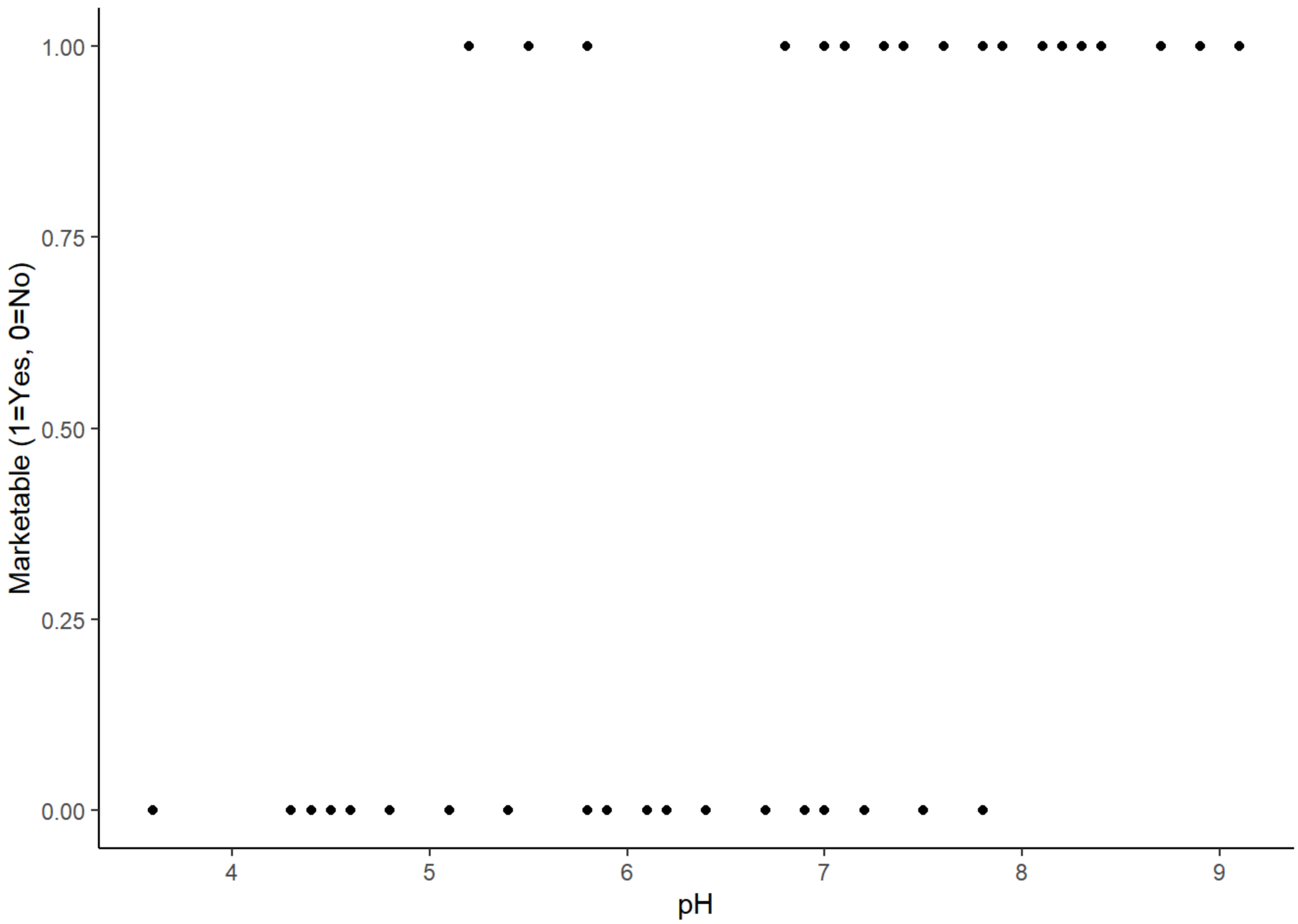
```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2     
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```r
bass$marketable <- ifelse(bass$AvgMercury < 0.354, 1, 0)
class(bass$marketable)
```

```
## [1] "numeric"
```

3. Plot marketable as function of pH.

```
library(ggplot2)
ggplot(bass, mapping=aes(y=marketable, x=pH))+
  geom_point()+
  theme_classic()+
  ylab("Marketable (1=Yes, 0=No)")+
  xlab("pH")
```



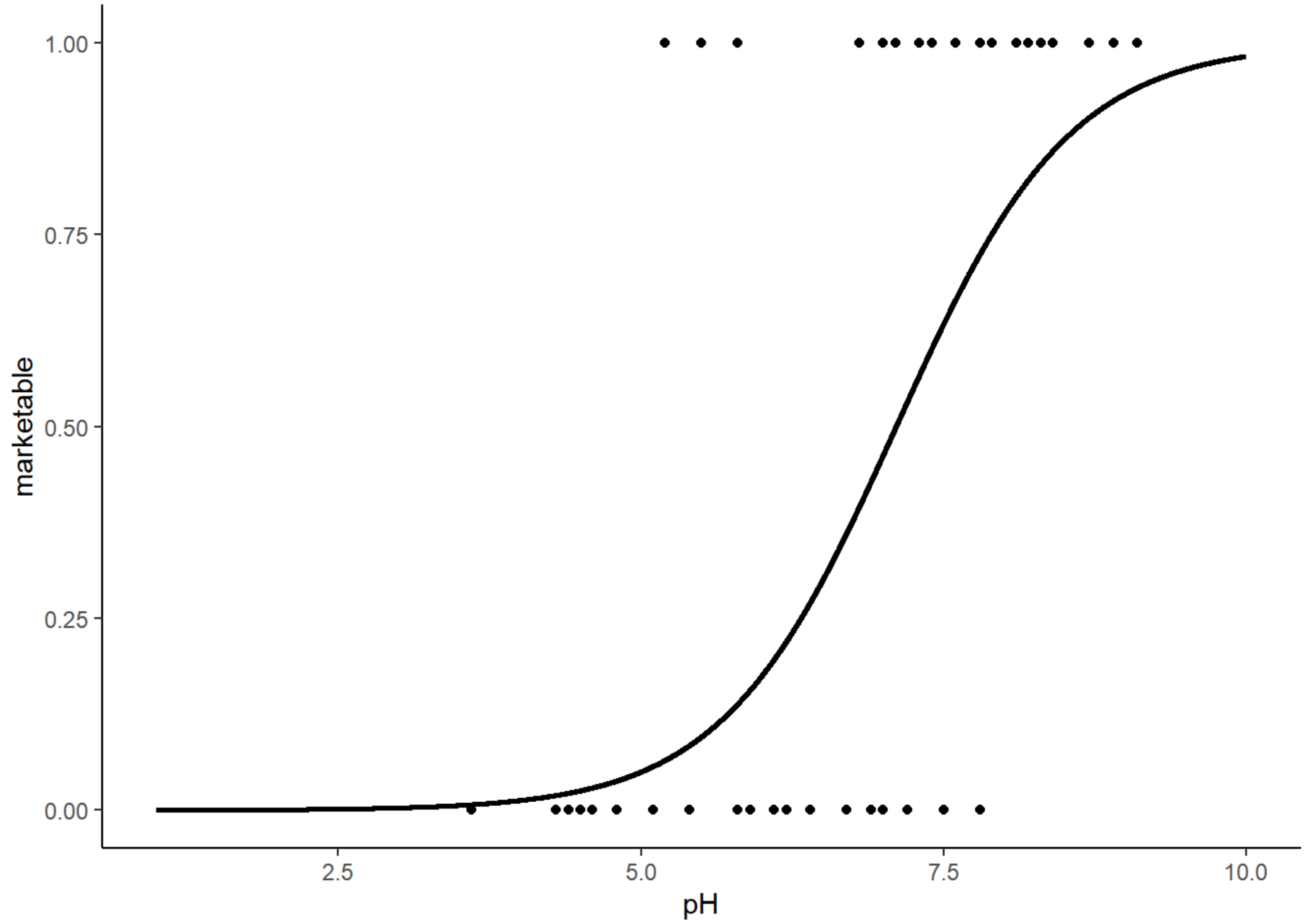
4. Fit a logistic regression modeling the effect of pH on marketable.

```
fit_bass_logistic_1 <- glm(marketable~pH, data=bass, family=binomial(link="logit"))
summary(fit_bass_logistic_1)
```

```
##
## Call:
## glm(formula = marketable ~ pH, family = binomial(link = "logit"),
##     data = bass)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.9380     2.8891  -3.440 0.000582 ***
## pH            1.3987     0.4128   3.388 0.000703 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 71.174  on 52  degrees of freedom
## Residual deviance: 50.444  on 51  degrees of freedom
## AIC: 54.444
##
## Number of Fisher Scoring iterations: 5
```

5. Reproduce the graph of marketable as a function of pH and overlay the fit/predicted line from your logistic regression.

```
new_data<- data.frame(pH=seq(1,10, 0.001))
new_data$Predicted_bass_logistic <- predict(fit_bass_logistic_1, newdata=new_data, type="respons
e")
ggplot(data=bass, mapping=aes(x=pH, y=marketable))+
  geom_point()+
  theme_classic()+
  geom_line(data=new_data, aes(x=pH, y=Predicted_bass_logistic), linewidth=1)
```



6. Using your model, find the pH values at which there is a 50% chance of fish being marketable with mercury.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
dose.p(fit_bass_logistic_1, p=0.5)
```

```
##              Dose        SE
## p = 0.5: 7.105234 0.2516347
```

7. Write 2-3 sentences interpreting the model and provide summary statistics (e.g., odds ratios, z-values) to support any claims you make about statistical significance.

Answer: With a one unit increase in pH, the odds of chance of fish being marketable increase by a factor of 4.05 (z-value = 3.338, $p < 0.001$). The 95% confidence interval for this oadds ratio ranges from about 2.014 to 10.519, suggesting that one unit increase in pH could significantly increase marketable probability. The intercept's significant negative value (z = -3.440, $p < 0.001$) suggests that at very low pH levels, the odds of marketability are low.

```
round(exp(cbind(coef(fit_bass_logistic_1), confint(fit_bass_logistic_1))),3)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 % 97.5 %
## (Intercept) 0.00 0.000  0.006
## pH          4.05 2.014 10.519
```

5.2 Part B

For this second part, you will analyze cricket chirps per unit of time as a function of temperature in degrees Fahrenheit.

1. Load in the chirps.txt dataset.

```
chirps <- read.table("C:/Users/chemk/Desktop/Classes/ENT6707_DataAnalysis/week11/chirps.txt", he
ader=TRUE, sep="\t")
nrow(chirps)
```

```
## [1] 7
```

```
str(chirps)
```

```
## 'data.frame':    7 obs. of  2 variables:
##  $ Temperature: num  54.5 59.5 63.5 67.5 72 78.5 83
##  $ Chirps     : int  81 97 103 123 150 182 195
```

```
head(chirps)
```

```
##   Temperature Chirps
## 1        54.5     81
## 2        59.5     97
## 3        63.5    103
## 4        67.5    123
## 5        72.0    150
## 6        78.5    182
```

```
tail(chirps)
```

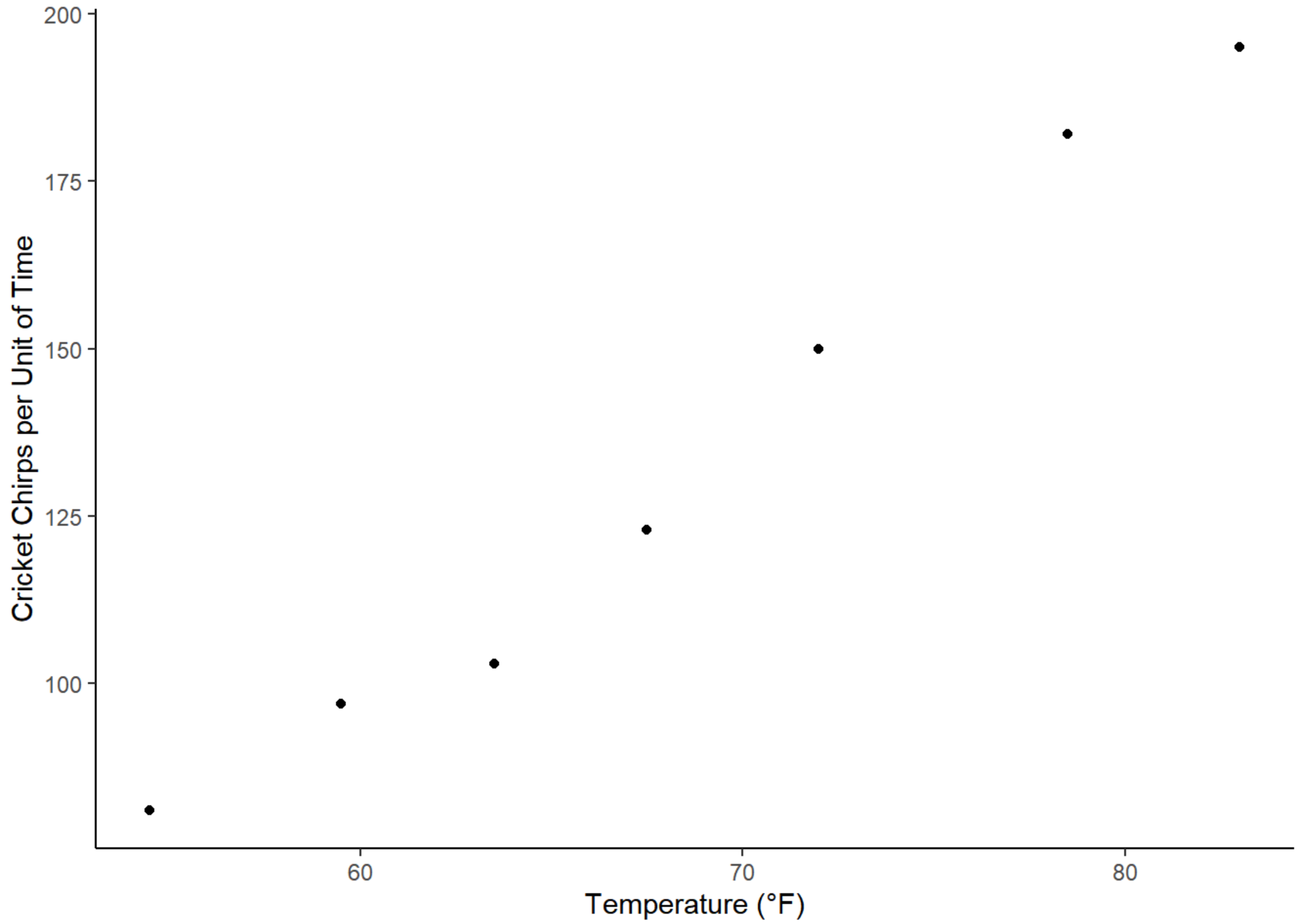
```
##   Temperature Chirps
## 2        59.5     97
## 3        63.5    103
## 4        67.5    123
## 5        72.0    150
## 6        78.5    182
## 7        83.0    195
```

```
summary(chirps)
```

```
##   Temperature        Chirps
##  Min.   :54.50   Min.   : 81
##  1st Qu.:61.50   1st Qu.:100
##  Median :67.50   Median :123
##  Mean   :68.36   Mean   :133
##  3rd Qu.:75.25   3rd Qu.:166
##  Max.   :83.00   Max.   :195
```

2. Plot Chirps as a function of Temperature.

```
ggplot(chirps, aes(x=Temperature, y=Chirps))+
  geom_point()+
  theme_classic()+
  ylab("Cricket Chirps per Unit of Time")+
  xlab("Temperature (°F)")
```



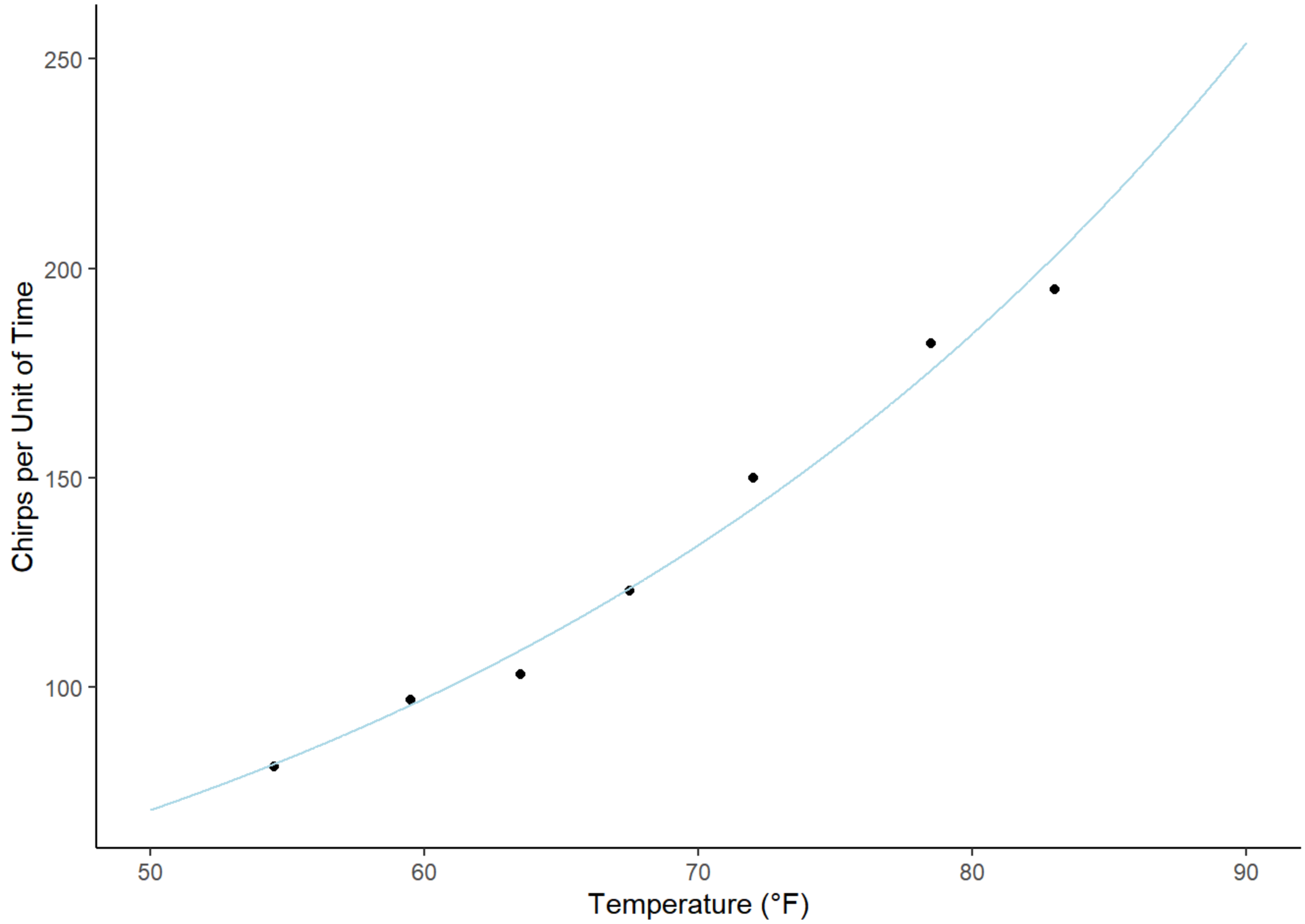
3. Fit a poisson regression modeling Chirps as a function of Temperature.

```
fit_chirps_poisson_1 <- glm(Chirps~Temperature, data=chirps, family=poisson(link="log"))
summary(fit_chirps_poisson_1)
```

```
##
## Call:
## glm(formula = Chirps ~ Temperature, family = poisson(link = "log"),
##     data = chirps)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.659220   0.251297  10.582   <2e-16 ***
## Temperature 0.031969   0.003499   9.138   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 86.0350  on 6  degrees of freedom
## Residual deviance:  1.2335  on 5  degrees of freedom
## AIC: 52.012
##
## Number of Fisher Scoring iterations: 3
```

4. Reproduce the graph of Chirps as a function of Temperature and overlay the fit/predicted line from your Poisson regression.

```
new_chirps <- data.frame(Temperature = seq(50, 90, 0.001))
new_chirps$Predicted_Chirps_poisson <- predict(fit_chirps_poisson_1, newdata=new_chirps, type="r
esponse")
ggplot(data=chirps, mapping=aes(x=Temperature, y=Chirps))+
  geom_point()+
  geom_line(data=new_chirps, aes(x=Temperature, y=Predicted_Chirps_poisson), color="lightblue")+
  ylab("Chirps per Unit of Time")+
  xlab("Temperature (°F)")+
  theme_classic()
```



5. Write 1-2 sentences interpreting the results.

Answer: The expected number of chirps per unit of time changes by a multiplicative factor of 1.03(= exp(0.031969)), or 3.1% increase, with an increase in 1°F increase in temperature (°F).