Chad Huntebrinker’s Homework 9

Chad Huntebrinker

2025-04-07

# Problem 1

We need to use a logistic model to find the effect of temperature and interpret the effect.

## Part a

We will find a logistic regression model to find the effect temperature has on the distress of the O-rings.

## (Intercept) Temp   
## 15.0429016 -0.2321627

##   
## Call:  
## glm(formula = TD ~ Temp, family = binomial, data = shuttle\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 15.0429 7.3786 2.039 0.0415 \*  
## Temp -0.2322 0.1082 -2.145 0.0320 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28.267 on 22 degrees of freedom  
## Residual deviance: 20.315 on 21 degrees of freedom  
## AIC: 24.315  
##   
## Number of Fisher Scoring iterations: 5

Temp coefficient is -0.2322. That means a one-degree increase in temperature decreases the logit by 0.2322. Since p = 0.0320 < 0.05, temperature has a significant effect on our model.

## Part b

Now, we need to estimate the probability of thermal distress at 31 degrees.

## Estimated probability of distress at 31 degrees F: 0.9996088

So, according to the estimated probability, there was a high likelihood of one of the O-rings failing.

## Part c

At what temperature does the estimated probability equal 0.50? And give linear approximation for the change in the estimated probability per degree increase in temperature.

## [1] "Temperature at 0.5: 64.7946409245488"

## [1] "Approximate change in probability per degree at 64.7946409245488 degrees F: -0.0580406860546001"

We find that the estimated probability of thermal distress is 0.5 at about 64.79 degrees F. And at 64.79 degrees F, the probability decreases by about 5.8 percent per 1 degree increase in temperature.

## Part d

## [1] "Odds multiplier per degree increase in temperature: 0.792817086409902"

For every 1 degree increase in temperature, the odds of thermal distress are multiplied by about 0.793. In other words, they are reduced by about 21%

## Part e

We will now be testing the null hypothesis of if temperature has no effect on the model. And the alternative hypothesis is that temperature does have an effect.

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: TD  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 22 28.267   
## Temp 1 7.952 21 20.315 0.004804 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We find the p-value is 0.0048. Because 0.0048 < 0.05, that means that temperature affects the probability of thermal distress. This means that the alternative hypothesis is correct.

# Problem 2

We will be fitting a logistic regression model for the probability of a satellite, using color alone as the predictor. ## Part a While treating color as a nominal-scale factor, we need to give the prediction equation and explain how to compare the first and fourth colors.

##   
## Call:  
## glm(formula = y ~ factor(crabs$color, levels = c("4", "1", "2",   
## "3")), family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) -0.7621 0.4577  
## factor(crabs$color, levels = c("4", "1", "2", "3"))1 1.8608 0.8087  
## factor(crabs$color, levels = c("4", "1", "2", "3"))2 1.7382 0.5123  
## factor(crabs$color, levels = c("4", "1", "2", "3"))3 1.1299 0.5509  
## z value Pr(>|z|)   
## (Intercept) -1.665 0.095910 .   
## factor(crabs$color, levels = c("4", "1", "2", "3"))1 2.301 0.021393 \*   
## factor(crabs$color, levels = c("4", "1", "2", "3"))2 3.393 0.000692 \*\*\*  
## factor(crabs$color, levels = c("4", "1", "2", "3"))3 2.051 0.040289 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 212.06 on 169 degrees of freedom  
## AIC: 220.06  
##   
## Number of Fisher Scoring iterations: 4

The prediction equation is: logit(π)= −0.7621 + 1.8608c1 + 1.7482c2 + 1.1299c3

c1 = 1 if the crab is color 1, so it would be -0.7621 + 1.8608 ~ 1.10 when it’s color 1. When it’s color 4, all the other covariates are 0 except for the intercept so it would be about -0.76. That means that color 1 has a lower odds of having a satellite compared to color 4.

## Part b

Using that model, we will be conducting a likelihood-ratio test of the null hypothesis that color has no effect on the probability of a satellite. The alternative hypothesis is that color does have an effect on there being a satellite.

## Analysis of Deviance Table  
##   
## Model 1: y ~ 1  
## Model 2: y ~ factor(crabs$color, levels = c("4", "1", "2", "3"))  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 172 225.76   
## 2 169 212.06 3 13.698 0.003347 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We find the p-value is about 0.003 (which is < 0.05). So we reject the null hypothesis and confirm the alternative hypothesis, which is that color does have an effect there being a satellite.

## Part c

Now, we will treat color as a quantitative, get an equation, interpret the coefficient of color and test the hypothesis that color has no effect. Our null hypothesis is that color has no effect on the equation while the alternative is that the color does have an effect.

##   
## Call:  
## glm(formula = y ~ as.numeric(color), family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.3635 0.5551 4.257 2.07e-05 \*\*\*  
## as.numeric(color) -0.7147 0.2095 -3.412 0.000645 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 213.30 on 171 degrees of freedom  
## AIC: 217.3  
##   
## Number of Fisher Scoring iterations: 4

## Analysis of Deviance Table  
##   
## Model 1: y ~ 1  
## Model 2: y ~ as.numeric(color)  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 172 225.76   
## 2 171 213.30 1 12.461 0.0004156 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Our equation is 2.3635 - 0.7147x and our p-value is 0.0004. Since the p-value < 0.05, we reject the null hypothesis and confirm the alternative (that color does have a significant effect on the model).

## Part d

The advantage of treating color as quantitative is that it reduces the number of parameters; as a result, there is an increase in the power which helps us with our hypothesis testing.

The disadvantages of treating color as quantitative is that it could lead to making incorrect assumptions (like assuming the color is linear when it isn’t) which could lead to a lack of model fit.

## Part e

We will now be using weight and quantitative color as explanatory variables, find standardized coefficients, and interpret.

##   
## Call:  
## glm(formula = y ~ z.weight + as.numeric(z.color), family = binomial,   
## data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.7430 0.1851 4.013 5.98e-05 \*\*\*  
## z.weight 1.9077 0.4414 4.322 1.55e-05 \*\*\*  
## as.numeric(z.color) -0.8247 0.3582 -2.302 0.0213 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 190.27 on 170 degrees of freedom  
## AIC: 196.27  
##   
## Number of Fisher Scoring iterations: 4

We find the color coefficient is -0.8247. That means that for every one standard deviation increase in color, the logit of being a satellite crab decrease by about 0.8.

#Code Appendix:  
#Chad Huntebrinker  
#Problem 4.5  
#Create the dataset  
shuttle\_data <- data.frame(  
 Ft = 1:23,  
 Temp = c(66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78, 67, 53, 67,   
 75, 70, 81, 76, 79, 75, 76, 58),  
 TD = c(0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0,   
 0, 0, 0, 0, 0, 1, 0, 1)  
)  
  
logit\_model <- glm(TD ~ Temp, data = shuttle\_data, family = binomial)  
summary(logit\_model)

##   
## Call:  
## glm(formula = TD ~ Temp, family = binomial, data = shuttle\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 15.0429 7.3786 2.039 0.0415 \*  
## Temp -0.2322 0.1082 -2.145 0.0320 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28.267 on 22 degrees of freedom  
## Residual deviance: 20.315 on 21 degrees of freedom  
## AIC: 24.315  
##   
## Number of Fisher Scoring iterations: 5

#a)  
coef(logit\_model)

## (Intercept) Temp   
## 15.0429016 -0.2321627

#Temp coefficient is -0.2322. That means a one-degree increase in Temp decreases the logit by  
#0.2322. Since p = 0.0320 < 0.05, Temp has a significant effect on our model.  
  
#b)  
new\_data <- data.frame(Temp = 31)  
prob\_31 <- predict(logit\_model, newdata = new\_data, type = "response")  
cat("Estimated probability of distress at 31 degrees F:", prob\_31)

## Estimated probability of distress at 31 degrees F: 0.9996088

#c)  
logit\_function <- function(temp) {  
 coef(logit\_model)[1] + coef(logit\_model)[2] \* temp  
}  
  
temp\_50 <- uniroot(function(temp) logit\_function(temp) - log(0.5 / (1 - 0.5)), lower = 0, upper = 100)$root  
print(paste("Temperature at 0.5:", temp\_50))

## [1] "Temperature at 0.5: 64.7946409245488"

#Compute linear approximation for change in probability per degree  
#B \* p \* (1 - p)  
slope\_at\_temp\_50 <- coef(logit\_model)[2] \* (0.5 \* (1 - 0.5))  
print(paste("Approximate change in probability per degree at", temp\_50, " degrees F:", slope\_at\_temp\_50))

## [1] "Approximate change in probability per degree at 64.7946409245488 degrees F: -0.0580406860546001"

#d)  
odds\_multiplier <- exp(coef(logit\_model)[2])  
print(paste("Odds multiplier per degree increase in temperature:", odds\_multiplier))

## [1] "Odds multiplier per degree increase in temperature: 0.792817086409902"

#For every 1 degree increase in temperature, the odds of thermal distress are multiplied by about 0.793.  
#In other words, they are reduced by 79.28%  
  
#e)  
anova\_result <- anova(logit\_model, test="Chisq") # Test if temperature has a significant effect  
print(anova\_result)

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: TD  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 22 28.267   
## Temp 1 7.952 21 20.315 0.004804 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#p-value = 0.0048 (< 0.05), so we reject the null hypothesis. As a result, temperature significantly  
#impacts the probability of thermal distress.  
  
#Chad Huntebrinker  
#Problem 4.9  
library(arm)  
library(readxl)  
#Get the dataset  
crabs <- read\_excel("crab\_data.xlsx")  
  
#a)  
model\_1 <- glm(y ~ factor(crabs$color, levels = c("4", "1", "2", "3")), family = binomial, data = crabs)  
summary(model\_1)

##   
## Call:  
## glm(formula = y ~ factor(crabs$color, levels = c("4", "1", "2",   
## "3")), family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) -0.7621 0.4577  
## factor(crabs$color, levels = c("4", "1", "2", "3"))1 1.8608 0.8087  
## factor(crabs$color, levels = c("4", "1", "2", "3"))2 1.7382 0.5123  
## factor(crabs$color, levels = c("4", "1", "2", "3"))3 1.1299 0.5509  
## z value Pr(>|z|)   
## (Intercept) -1.665 0.095910 .   
## factor(crabs$color, levels = c("4", "1", "2", "3"))1 2.301 0.021393 \*   
## factor(crabs$color, levels = c("4", "1", "2", "3"))2 3.393 0.000692 \*\*\*  
## factor(crabs$color, levels = c("4", "1", "2", "3"))3 2.051 0.040289 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 212.06 on 169 degrees of freedom  
## AIC: 220.06  
##   
## Number of Fisher Scoring iterations: 4

#The prediction equation is: logit(π)= −0.76 + 1.86c1 + 1.74c2 + 1.13c3  
  
#c1 = 1 if the crab is color 1  
#So it would be -0.7621 + 1.8608 = 1.10  
#When it's color 4, all the other covariates are 0 except for the intercept. So it would be -0.76  
  
#b)  
model\_null <- glm(y ~ 1, family = binomial, data = crabs)  
anova(model\_null, model\_1, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: y ~ 1  
## Model 2: y ~ factor(crabs$color, levels = c("4", "1", "2", "3"))  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 172 225.76   
## 2 169 212.06 3 13.698 0.003347 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#c)  
model\_2 <- glm(y ~ as.numeric(color), family = binomial, data = crabs)  
summary(model\_2)

##   
## Call:  
## glm(formula = y ~ as.numeric(color), family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.3635 0.5551 4.257 2.07e-05 \*\*\*  
## as.numeric(color) -0.7147 0.2095 -3.412 0.000645 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 213.30 on 171 degrees of freedom  
## AIC: 217.3  
##   
## Number of Fisher Scoring iterations: 4

anova(model\_null, model\_2, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: y ~ 1  
## Model 2: y ~ as.numeric(color)  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 172 225.76   
## 2 171 213.30 1 12.461 0.0004156 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#d)  
#Advantage (Power): Treating color as quantitative reduces the number of parameters, increasing power   
#for detecting an effect.  
  
#Disadvantage (Lack of Fit): Treating color as quantitative could lead to making assumptions (like   
#assuming the color is linear when it isn't) which could lead to a model lack of fit.  
  
#e)  
model\_full <- glm(y ~ weight + as.numeric(color), family = binomial, data = crabs)  
  
# Standardize coefficients  
model\_standardized <- standardize(model\_full)  
summary(model\_standardized)

##   
## Call:  
## glm(formula = y ~ z.weight + as.numeric(z.color), family = binomial,   
## data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.7430 0.1851 4.013 5.98e-05 \*\*\*  
## z.weight 1.9077 0.4414 4.322 1.55e-05 \*\*\*  
## as.numeric(z.color) -0.8247 0.3582 -2.302 0.0213 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 190.27 on 170 degrees of freedom  
## AIC: 196.27  
##   
## Number of Fisher Scoring iterations: 4