Chad Huntebrinker’s Homework 10

Chad Huntebrinker

2025-04-14

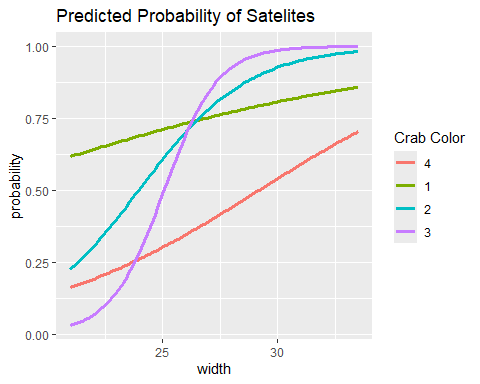
# Homework 4.19

For this problem, we will be creating a model with color and width predictors and add three terms to permit interaction between color and width

## Part a

We need to report the prediction equations relating width to the probability of a satellite, for each color. We also need to plot our findings and report them.

##   
## Call:  
## glm(formula = y ~ width + color + width \* color, family = binomial,   
## data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.85383 6.69393 -0.874 0.3818   
## width 0.20043 0.26166 0.766 0.4437   
## color1 4.10122 13.27532 0.309 0.7574   
## color2 -4.18613 7.58093 -0.552 0.5808   
## color3 -15.66423 9.56065 -1.638 0.1013   
## width:color1 -0.09443 0.50042 -0.189 0.8503   
## width:color2 0.21844 0.29524 0.740 0.4594   
## width:color3 0.65794 0.37533 1.753 0.0796 .  
## ---  
## 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: 183.08 on 165 degrees of freedom  
## AIC: 199.08  
##   
## Number of Fisher Scoring iterations: 5



The equation is logit() = -5.85 + 0.20x + 4.10c1 - 4.19c2 - 15.66c3 - 0.09(x \* c1) + 0.22(x \* c2) + 0.66(x \* c3) where c1 is color 1, c2 is color 2, c3 is color 3, and x is width.

What we see on the graph is that color 1 and color 4 have a steady relation with width. As width increases, so does the probability that they will have a satellite. Color 2 and 3 have a sharp increase around 22.5cm width and it starts faltering off around 28cm width.

## Part b

Next, we test whether the interaction model fits better than the simpler model without interaction terms. We’ll need to compare their predictive power by finding the correlation R between the observed and fitted values for each model and interpret.

Our null hypothesis is that the interaction terms do not improve the model while our alternative hypothesis is that the interaction terms do improve the model.

## AIC of interaction model: 199.0806

## AIC of simplier model: 197.457

## Analysis of Deviance Table  
##   
## Model 1: y ~ width + color  
## Model 2: y ~ width + color + width \* color  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 168 187.46   
## 2 165 183.08 3 4.3764 0.2236

## Correlation for the simplier model: 0.4522131

## Correlation for the interaction model: 0.472076

What we find is that their AICs are extremly similar (only different by about 2). We also see that the likelihood ratio statistic is 4.4 with a p-value of 0.22. Since our p-value is > 0.05, we fail to reject the null hypothesis; thus, left with the conclusion that our interaction terms don’t help with the model. And we see that the correlation for the simpler model is 0.45 while the correlation for the interaction model is 0.47.

All of these support the idea that the simpler model is fits just as well as the interaction model and is fine to use.

# Problem 5.1

For the horseshoe crabs data file, we will be fitting a model using weight and width as explanatory variables for the probability of a satellite.

## Part a

We need to conduct a likelihood-ratio test of H0: B1 = B2 = 0 and interpret. Our alternative hypothesis is that at least one predictor is significant to whether there’s a satellite to the crab.

## Analysis of Deviance Table  
##   
## Model 1: y ~ 1  
## Model 2: y ~ width + weight  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 172 225.76   
## 2 170 192.89 2 32.867 7.296e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The LR statistic is 32.9 and our p-value is 7.296e-08. Since our p-value < 0.05, we reject H0. That means that we have strong evidence that at least one of our predictors has a significant role on the probability of there being a satellite for the crab.

## Part b

Next, we will be conducting separate likelihood-ratio tests for the partial effects of each variable. Our null hypothesis for both is that the variable we are removing (weight for one test and width for the other) doesn’t have an effect on the probability of the crab having a satellite while our alternative is that it does.

## Analysis of Deviance Table  
##   
## Model 1: y ~ width  
## Model 2: y ~ width + weight  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 171 194.45   
## 2 170 192.89 1 1.5608 0.2116

## Analysis of Deviance Table  
##   
## Model 1: y ~ weight  
## Model 2: y ~ width + weight  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 171 195.74   
## 2 170 192.89 1 2.8453 0.09164 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The LR statistic is 1.56 for weight and 2.85 for width, with p-values of 0.21 and 0.09. Since both of these p-values are > 0.05, we fail to reject H0 for both of these LR tests. So the question becomes: why does neither of these LR tests show evidence of an effect when the test in part a shows very strong evidence? One possible reason why neither test shows evidence of an effect while the previous test did is that when combined together, the variables provide valuable insight that can’t be seen in a LR-test. So when they are separate, they seem to be useless to the model. But when you include both in the model, the combination explains more about the probability that there is a satellite for the crab.

## Part c

Finally, we will be using purposeful selection to build a model when weight, the spine condition, and color factors are the potential explanatory variables. We will be doing this by comparing each model’s AIC and deviance.

First, we will build each model with one variable and compare their AICs:

## Reduced model's AIC: 227.7585

## Weight model's AIC: 199.7371

## Spine model's AIC: 229.6243

## Color model's AIC: 220.0608

We will keep weight and color with our models because their AICs are the best (in other words, the lowest).

Next, we will build a model with weight and color and compare them to their models with only weight or color.

## Weight and color model's AIC: 198.5423

## Analysis of Deviance Table  
##   
## Model 1: y ~ weight  
## Model 2: y ~ weight + color  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 171 195.74   
## 2 168 188.54 3 7.1949 0.06594 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table  
##   
## Model 1: y ~ color  
## Model 2: y ~ weight + color  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 169 212.06   
## 2 168 188.54 1 23.519 1.237e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We find that the AIC for weight and color is better than the model with just color and slightly improves when compared to the model with just weight. We also see that the deviance decreases with the model of weight and color. As a result, we will move ahead with the model that includes both weight and color in it due to the fact that when included together, weight and color seem to do a better job of explaining the probability of a crab of having a satellite.

Next, we add spine to our model of weight and color.

## Weight, color, and spine condition model's AIC: 199.5665

## Analysis of Deviance Table  
##   
## Model 1: y ~ weight + color  
## Model 2: y ~ weight + color + spine  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 168 188.54   
## 2 167 187.57 1 0.97577 0.3232

The AIC isn’t better by the addition of spine (compared to weight and color model’s AIC of 198.5423) and we find that the deviance doesn’t decrease by much either. So, we remove spine from the model.

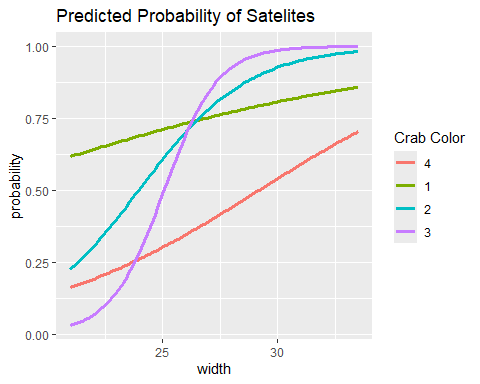
Finally, we add an interaction between color and weight into our color and weight model.

## Weight and color's interaction model's AIC: 197.6563

## Analysis of Deviance Table  
##   
## Model 1: y ~ weight + color  
## Model 2: y ~ weight + color + weight \* color  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 168 188.54   
## 2 165 181.66 3 6.886 0.07562 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We see the AIC improves slightly compared to the AIC of just weight and color (weight and color model’s AIC: 198.5423) but we also see the deviance has a small decrease (of about 7). As a result, there is a marginal improvement when we add an interaction; however, it doesn’t seem to be enough to justify complicating the model by adding an interaction term. As a result, we find our best model is with weight and color but no interaction.

#Code Appenix  
#Chad Huntebrinker  
#Problem 4.19  
library(ggplot2)  
library(readxl)  
  
#a)  
crabs <- read\_excel("crab\_data.xlsx")  
crabs$color <- factor(crabs$color, levels = c("4", "1", "2", "3"))  
  
  
model\_1 <- glm(y ~ width + color + width \* color, family = binomial, data = crabs)  
sum\_model\_1 <- summary(model\_1)  
  
#The equation is -5.85 + 0.20x + 4.10c1 - 4.19c2 - 15.66c3 - 0.09(x\*c1) + 0.22(x\*c2) + 0.66(x\*c3)  
  
#Next, we will be plotting and interpreting the prediction equation:  
#First, need to create a data frame with our predictions by giving it  
#a range of widths  
width\_seq <- seq(min(crabs$width), max(crabs$width), length.out = 100)  
num\_colors <- levels(crabs$color)  
  
prediction\_grid <- expand.grid(  
 width = width\_seq,  
 color = num\_colors  
)  
  
prediction\_grid$probability <- predict(model\_1, newdata = prediction\_grid, type = "response")  
  
ggplot(prediction\_grid, aes(x = width, y = probability, color = color)) +  
 geom\_line(size = 1.2) +  
 labs(  
 title = "Predicted Probability of Satelites",  
 xlab = "Width",  
 ylab = "Probability",  
 color = "Crab Color"  
 )



#What we see on the graph is that color 1 and even color 4 have a steady relation with width. As width  
#increases, so does the probability that they will have a satelite. Color 2 and 3 have a sharp increase around  
#22.5 width and it starts faltering off around 28 width.  
   
#b)  
model\_2 <- glm(y ~ width + color, family = binomial, data = crabs)  
sum\_model\_2 <- summary(model\_2)  
  
cat("AIC of interaction model: ", sum\_model\_1$aic, "\n")

## AIC of interaction model: 199.0806

cat("AIC of simplier model: ", sum\_model\_2$aic, "\n")

## AIC of simplier model: 197.457

anova(model\_2, model\_1)

## Analysis of Deviance Table  
##   
## Model 1: y ~ width + color  
## Model 2: y ~ width + color + width \* color  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 168 187.46   
## 2 165 183.08 3 4.3764 0.2236

cat("Correlation for the simplier model: ", cor(crabs$y, fitted(model\_2)), "\n")

## Correlation for the simplier model: 0.4522131

cat("Correlation for the interaction model: ", cor(crabs$y, fitted(model\_1)), "\n")

## Correlation for the interaction model: 0.472076

#The likelihood ratio statistic is 4.4 and the p-value is 0.22 (which is greater than 0.05).   
#The correlation for the simpler model is 0.452 and the correlation for the complex model is 0.472.  
#As a result of these findings, we can say that simpler model fits almost as well as the interaction model  
#and would be fine to use.  
  
#Chad Huntebrinker  
#Problem 5.1  
library(readxl)  
  
#a)  
crabs <- read\_excel("crab\_data.xlsx")  
  
model\_1 <- glm(y ~ width + weight, family = binomial, data = crabs)  
sum\_model\_1 <- summary(model\_1)  
  
model\_reduced <- glm(y~1, family = binomial, data = crabs)  
  
#Our null hypothesis is that B1 = B2 = 0 while our alternative hypothesis is that at least one  
#predictor is significant to the model (either B1 != 0, B2 != 0 or both)  
anova(model\_reduced, model\_1)

## Analysis of Deviance Table  
##   
## Model 1: y ~ 1  
## Model 2: y ~ width + weight  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 172 225.76   
## 2 170 192.89 2 32.867 7.296e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#The LR statistic is 32.9. Since our p-value < 0.05, we reject H0. That means that at least  
#one of our predictors has a significant role on the model  
  
#b)  
model\_no\_weight <- glm(y ~ width, family = binomial, data = crabs)  
model\_no\_width <- glm(y ~ weight, family = binomial, data = crabs)  
  
anova(model\_no\_weight, model\_1)

## Analysis of Deviance Table  
##   
## Model 1: y ~ width  
## Model 2: y ~ width + weight  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 171 194.45   
## 2 170 192.89 1 1.5608 0.2116

anova(model\_no\_width, model\_1)

## Analysis of Deviance Table  
##   
## Model 1: y ~ weight  
## Model 2: y ~ width + weight  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 171 195.74   
## 2 170 192.89 1 2.8453 0.09164 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#The LR statistic is 1.56 for weight and 2.85 for width, with p-values of 0.21 and 0.09.  
#One possible reason why neither test shows evidence of an effect while the previous step's test did is that  
#when combined together, the variables provide valuable insight that can't be seen in a LR-test. So when they  
#are separate, they seem useless to the model. But putting them together in the same model allows them to  
#better explain what is going on in the model.  
  
#c)  
#First, build each model with one variable and compare their AICs  
model\_weight <- glm(y ~ weight, family = binomial, data = crabs)  
model\_spine <- glm(y ~ spine, family = binomial, data = crabs)  
model\_color <- glm(y ~ color, family = binomial, data = crabs)  
  
summary(model\_reduced)

##   
## Call:  
## glm(formula = y ~ 1, family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.5824 0.1585 3.673 0.000239 \*\*\*  
## ---  
## 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: 225.76 on 172 degrees of freedom  
## AIC: 227.76  
##   
## Number of Fisher Scoring iterations: 4

summary(model\_weight)

##   
## Call:  
## glm(formula = y ~ weight, family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.6947 0.8802 -4.198 2.70e-05 \*\*\*  
## weight 1.8151 0.3767 4.819 1.45e-06 \*\*\*  
## ---  
## 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: 195.74 on 171 degrees of freedom  
## AIC: 199.74  
##   
## Number of Fisher Scoring iterations: 4

summary(model\_spine)

##   
## Call:  
## glm(formula = y ~ spine, family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.75947 0.51165 1.484 0.138  
## spine -0.07104 0.19460 -0.365 0.715  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 225.62 on 171 degrees of freedom  
## AIC: 229.62  
##   
## Number of Fisher Scoring iterations: 4

summary(model\_color)

##   
## Call:  
## glm(formula = y ~ color, family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.3635 0.5551 4.257 2.07e-05 \*\*\*  
## 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

#We will keep weight and color because their AICs are the best.  
  
#Next, we will build a model with weight and color and compare them to their models with only weight or color  
model\_weight\_color <- glm(y ~ weight + color, family = binomial, data = crabs)  
summary(model\_weight\_color)

##   
## Call:  
## glm(formula = y ~ weight + color, family = binomial, data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.0316 1.1161 -1.820 0.0687 .   
## weight 1.6531 0.3825 4.322 1.55e-05 \*\*\*  
## color -0.5142 0.2234 -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

anova(model\_weight, model\_weight\_color)

## Analysis of Deviance Table  
##   
## Model 1: y ~ weight  
## Model 2: y ~ weight + color  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 171 195.74   
## 2 170 190.27 1 5.4684 0.01936 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(model\_color, model\_weight\_color)

## Analysis of Deviance Table  
##   
## Model 1: y ~ color  
## Model 2: y ~ weight + color  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 171 213.30   
## 2 170 190.27 1 23.029 1.596e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#We find that the AIC for weight and color is better than when they are on their own and we  
#see that there is increase in deviance when we remove color or weight from the model  
  
#Next, add spine to our model  
model\_weight\_color\_spine <- glm(y ~ weight + color + spine, family = binomial, data = crabs)  
summary(model\_weight\_color\_spine)

##   
## Call:  
## glm(formula = y ~ weight + color + spine, family = binomial,   
## data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.5318 1.1934 -2.122 0.03387 \*   
## weight 1.6619 0.3826 4.344 1.4e-05 \*\*\*  
## color -0.6203 0.2395 -2.590 0.00959 \*\*   
## spine 0.2973 0.2381 1.249 0.21181   
## ---  
## 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: 188.72 on 169 degrees of freedom  
## AIC: 196.72  
##   
## Number of Fisher Scoring iterations: 4

anova(model\_weight\_color, model\_weight\_color\_spine)

## Analysis of Deviance Table  
##   
## Model 1: y ~ weight + color  
## Model 2: y ~ weight + color + spine  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 170 190.27   
## 2 169 188.72 1 1.5519 0.2129

#The AIC isn't better by this addition and we find that the deviance doesn't increase by much. So, we  
#remove spine from the model  
  
#Finally, add an interaction between color and weight  
model\_weight\_color\_interaction <- glm(y ~ weight + color + weight\*color, family = binomial, data = crabs)  
summary(model\_weight\_color\_interaction)

##   
## Call:  
## glm(formula = y ~ weight + color + weight \* color, family = binomial,   
## data = crabs)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.1890 3.1959 -0.372 0.710  
## weight 1.2885 1.3487 0.955 0.339  
## color -0.8506 1.2227 -0.696 0.487  
## weight:color 0.1468 0.5239 0.280 0.779  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 225.76 on 172 degrees of freedom  
## Residual deviance: 190.19 on 169 degrees of freedom  
## AIC: 198.19  
##   
## Number of Fisher Scoring iterations: 4

anova(model\_weight\_color, model\_weight\_color\_interaction)

## Analysis of Deviance Table  
##   
## Model 1: y ~ weight + color  
## Model 2: y ~ weight + color + weight \* color  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 170 190.27   
## 2 169 190.19 1 0.079107 0.7785

#We see the AIC doesn't improve and we also see the deviance has only a small increase. As a result,  
#we find our best model is with weight and color but no interaction.