Homework 5 Solutions

Load packages:

```
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(mosaic))
```

Load data

```
diag <- read.csv("https://www.macalester.edu/~ajohns24/data/breastcancer2.csv")
head(diag)
```

```
##   X Class Adhes BNucl Chrom Epith Mitos NNucl Thick UShap USize new
## 1 1   yes     8    10     9     7     1     7     8    10    10 yes
## 2 2   yes     3     3     4     2     1     4     5     3     3 yes
## 3 3   yes    10     9     5     7     4     5     8     5     7 yes
## 4 4   yes     4     1     4     6     1     3     7     6     4 yes
## 5 5   yes     6    10     4     4     2     1    10     7     7 yes
## 6 6   yes    10    10     5     5     4     4     7     2     3 yes
```

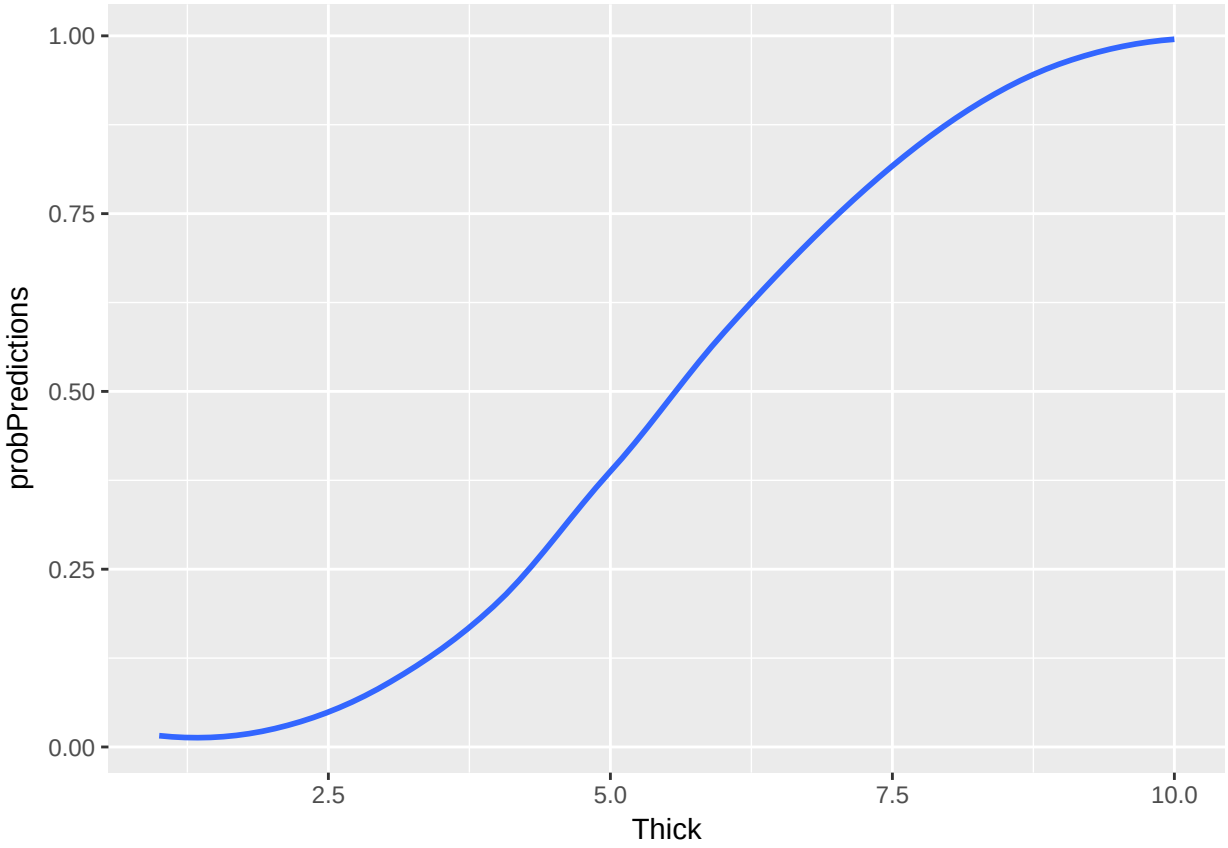
Exercise 1

```
#a
#log(odds cancer) = -5.18199 + 0.94611 Thick
mod1 <- glm(Class ~ Thick, diag, family="binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = Class ~ Thick, family = "binomial", data = diag)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2248  -0.4281  -0.1695   0.1659   2.9156
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.18199    0.38666  -13.40   <2e-16 ***
## Thick        0.94611    0.07591   12.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.39  on 680  degrees of freedom
## Residual deviance: 451.69  on 679  degrees of freedom
## AIC: 455.69
##
## Number of Fisher Scoring iterations: 6
```

```
#b
diag$probPredictions <- mod1$fitted.values
ggplot(diag, aes(x=Thick, y=probPredictions)) +
    stat_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```




```
#c
#For every 1 point increase on the thickness abnormality scale,
#the odds of having cancer increase by 157.6% on average
100*(exp(0.94611) - 1)
```

```
## [1] 157.5671
```

```
#d
#Yes - the p-value is very small.

#e
probFun <- makeFun(mod1)
probFun(Thick = 8)
```

```
##         1
## 0.9158194
```

```
#f
#answers may vary
```

Exercise 2

```
#a
#odds
0.5 / (1-0.5)
```

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

```
#log odds
log(1)
```

```
## [1] 0
```

```
#b
#0 = log(odds) = -5.18199 + 0.94611 Thick when Thick=5.477154
5.18199/0.94611
```

```
## [1] 5.477154
```

```
#c
#classify as malignant if Thick > 5.477154
```

Exercise 3

```
#a
#sensitivity = P(classify malignant as malignant)
#specificity = P(classify benign as benign)


#b
table(diag$Class, mod1$fitted >= 0.5)
```

```
##
##       FALSE TRUE
##   no    424   19
##   yes    75  163
```

```
#sensitivity
163/(163+75)
```

```
## [1] 0.6848739
```

```
#specificity
424/(424+19)
```

```
## [1] 0.9571106
```

Exercise 4

```
#a
(log(0.2/0.8) + 5.18199)/(0.94611)
```

```
## [1] 4.011897
```

```
#b
table(diag$Class, mod1$fitted >= 0.2)
```

```
##
##       FALSE TRUE
##   no    341  102
##   yes    31  207
```

```r
#sensitivity
207/(207+31)
```

```
## [1] 0.8697479
```

```r
#specificity
341/(102+341)
```

```
## [1] 0.7697517
```

```r
#c
#Increasing specificity results in decreased specificity
```

Exercise 5

```r
mod2 <- glm(Class ~ Thick * USize, diag, family="binomial")
summary(mod2)
```

```
##
## Call:
## glm(formula = Class ~ Thick * USize, family = "binomial", data = diag)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7603  -0.1738  -0.0322   0.1069   2.8984
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.11229    1.37578  -8.077 6.63e-16 ***
## Thick         1.34208    0.23033   5.827 5.65e-09 ***
## USize         2.42525    0.36872   6.577 4.79e-11 ***
## Thick:USize  -0.21664    0.05485  -3.950 7.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.39  on 680  degrees of freedom
## Residual deviance: 178.41  on 677  degrees of freedom
## AIC: 186.41
##
## Number of Fisher Scoring iterations: 7
```

```r
#a
#check out the plot: cases with small Thick AND Usize are
#classified as benign

#b
table(diag$Class, mod2$fitted >= 0.5)
```

```
##
##       FALSE TRUE
##   no    429   14
##   yes    18  220
```

```r
#sensitivity
220/(220+18)
```

```
## [1] 0.9243697
```

```r
#specificity
429/(429+14)
```

```
## [1] 0.9683973
```

```r
#c
#if we removed the interaction term, the classification
#border would be linear, not curved
```