Quiz7 - MA478

Clark

Once again let's consider the `wbca` dataset. Create a model to predict malignency of tumor based on marginal adhesion. Compare a model with a probit link to a logit link. Which one is preferred?

```
library(faraway)
library(tidyverse)
data(wbca)

wbca_mod <- wbca %>%
  mutate(malig = ifelse(Class==0,1,0))

mod_prob <- glm(malig~Adhes,data=wbca_mod,family=binomial(link="probit"))
mod_log <- glm(malig~Adhes,data=wbca_mod,family=binomial(link="logit"))

AIC(mod_prob)
```

```
## [1] 469.8879
```

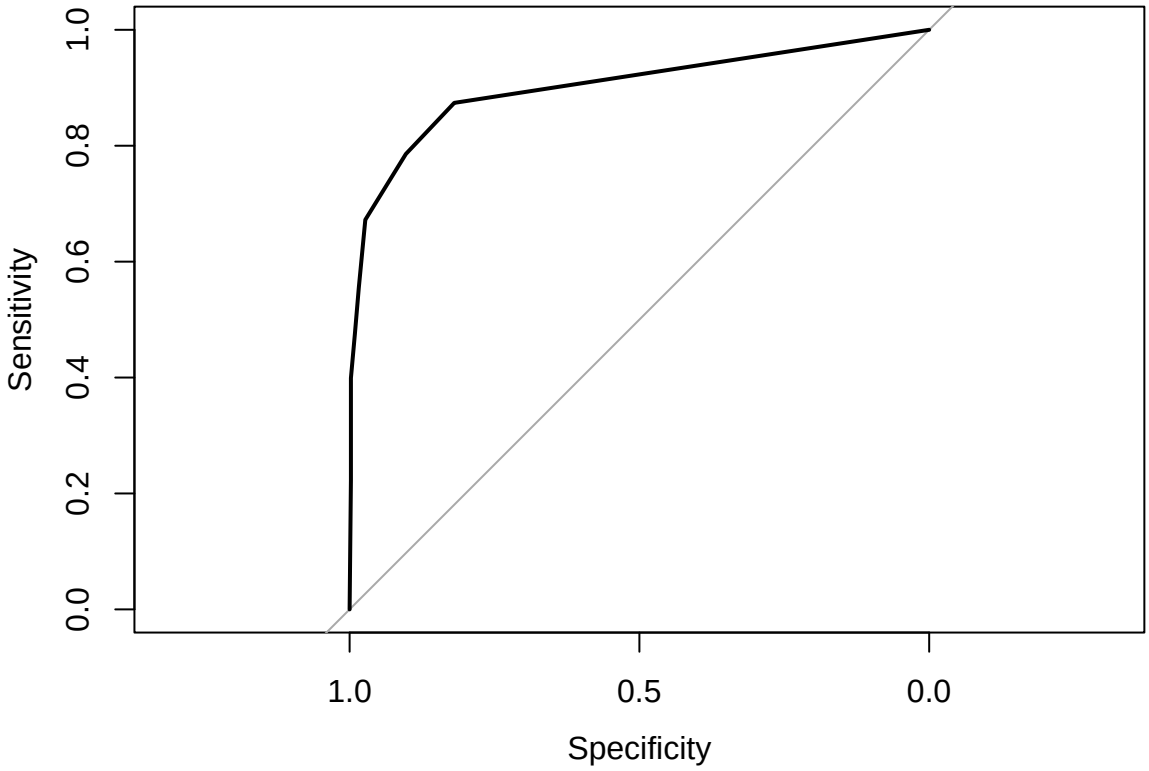
```
AIC(mod_log)
```

```
## [1] 463.1017
```

create an ROC curve for your preferred model. Determine an appropriate threshold where the sensitivity is over 60% and the specificity is over 97%.

```
library(pROC)
ROC_Curve <- roc(as.factor(wbca_mod$malig),predict(mod_log, type="response"))

plot(ROC_Curve)
```




```
ROC_Curve$sensitivities
```

```
##  [1] 1.0000000 0.8739496 0.7857143 0.6722689 0.5546218 0.4747899 0.3991597
##  [8] 0.3445378 0.2394958 0.2226891 0.0000000
```

```
ROC_Curve$specificities
```

```
##  [1] 0.0000000 0.8194131 0.9029345 0.9729120 0.9841986 0.9909707 0.9977427
##  [8] 0.9977427 0.9977427 0.9977427 1.0000000
```

```
ROC_Curve$thresholds
```

```
##  [1]      -Inf 0.1662232 0.3549260 0.5985040 0.8036191 0.9202717 0.9707144
##  [8] 0.9896844 0.9964224 0.9987661       Inf
```

Explain why the specificities are so much higher than the sensitivities for this dataset.

There's only a single predictor so the model likely is not sufficient, which we can see:

```
library(glmtoolbox)
hltest(mod_log)
```

```
##
```

```
##    The Hosmer-Lemeshow goodness-of-fit test
##
##  Group Size Observed Expected
##      1  393       30 37.79350
##      2   58       21 13.70422
##      3   58       27 27.46720
##      4   76       65 63.40859
##      5   42       42 41.66062
##      6   54       53 53.96587
##
##          Statistic =  34.81031
## degrees of freedom =  4
##            p-value =  5.0813e-07
```

Because of this, the model is likely defaulting to predicting most of our data as the smallest class (0s)

```
predict(mod_log, type="response")%>%median()
```

```
## [1] 0.09616668
```

Inflating our specificity for most threshold values. BL is, this isn't a very good model.