STOR 590 HW3 Solution

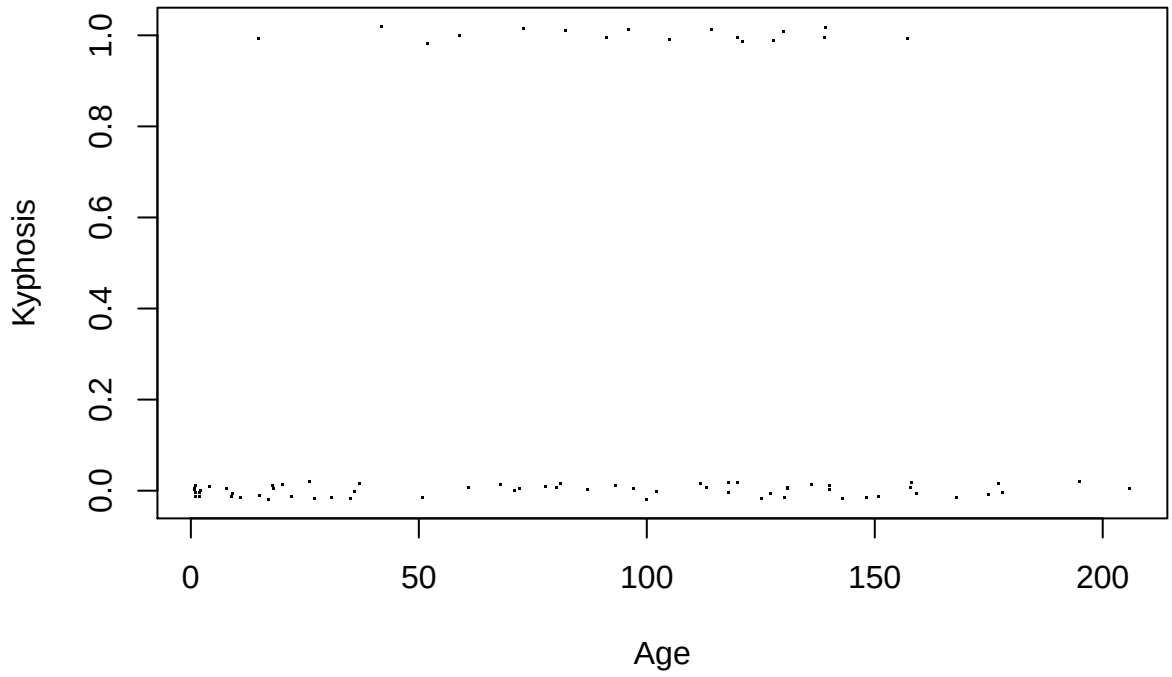
Taebin Kim

Page 47 Exercise 3

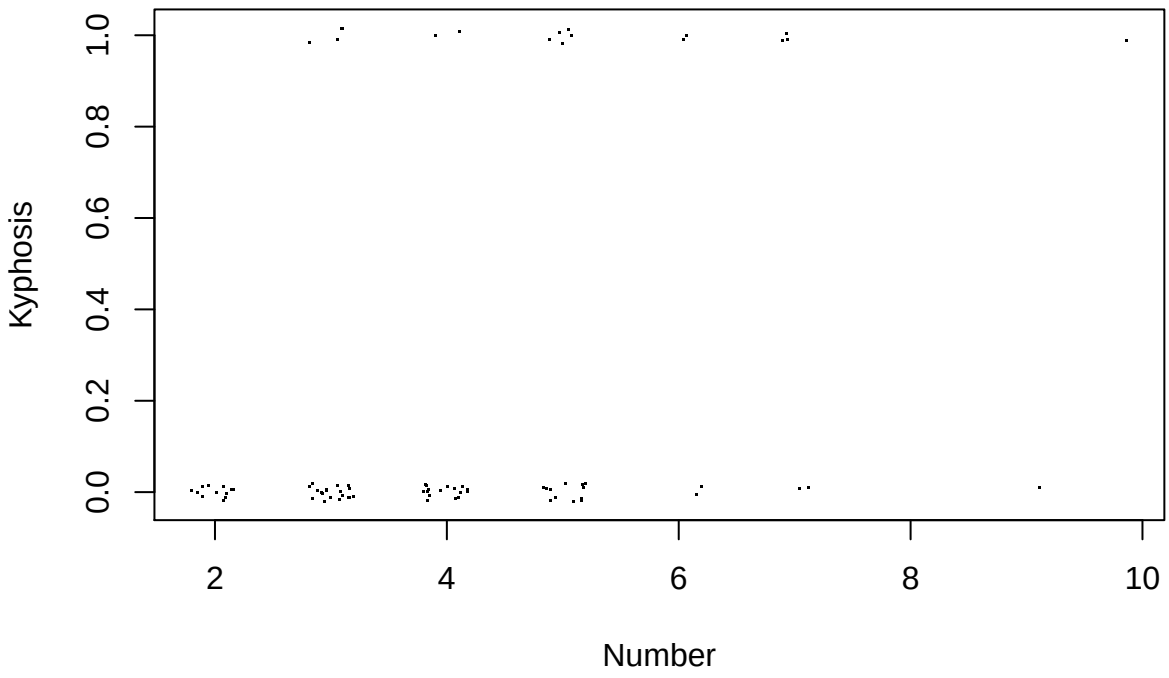
Part (a)

We make plots of the response as it relates to each of the three predictors.

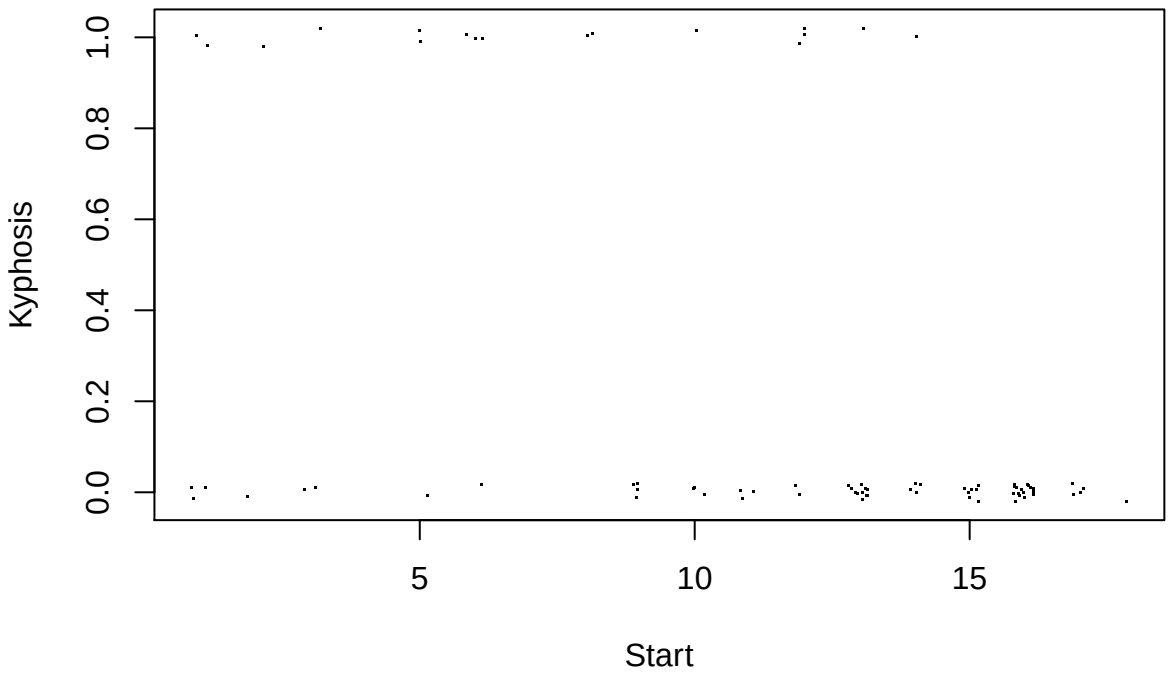
```
data(kyphosis, package="rpart")
kyphosis$y <- ifelse(kyphosis$Kyphosis == "absent", 0, 1)
plot(jitter(y, 0.1) ~ jitter(Age), kyphosis, xlab='Age', ylab='Kyphosis', pch='.')
```




```
plot(jitter(y, 0.1) ~ jitter(Number), kyphosis, xlab='Number', ylab='Kyphosis', pch='.')
```




```
plot(jitter(y, 0.1) ~ jitter(Start), kyphosis, xlab='Start', ylab='Kyphosis', pch='.')
```



The plots show that a person with `kyphosis` tend to have a larger value of `Number` and a smaller value of `Start`. The relationship between `Kyphosis` and `Age` is hard to tell from the plot above.

Part (b)

We fit a GLM with the kyphosis indicator as the response and the other three variables as predictors.

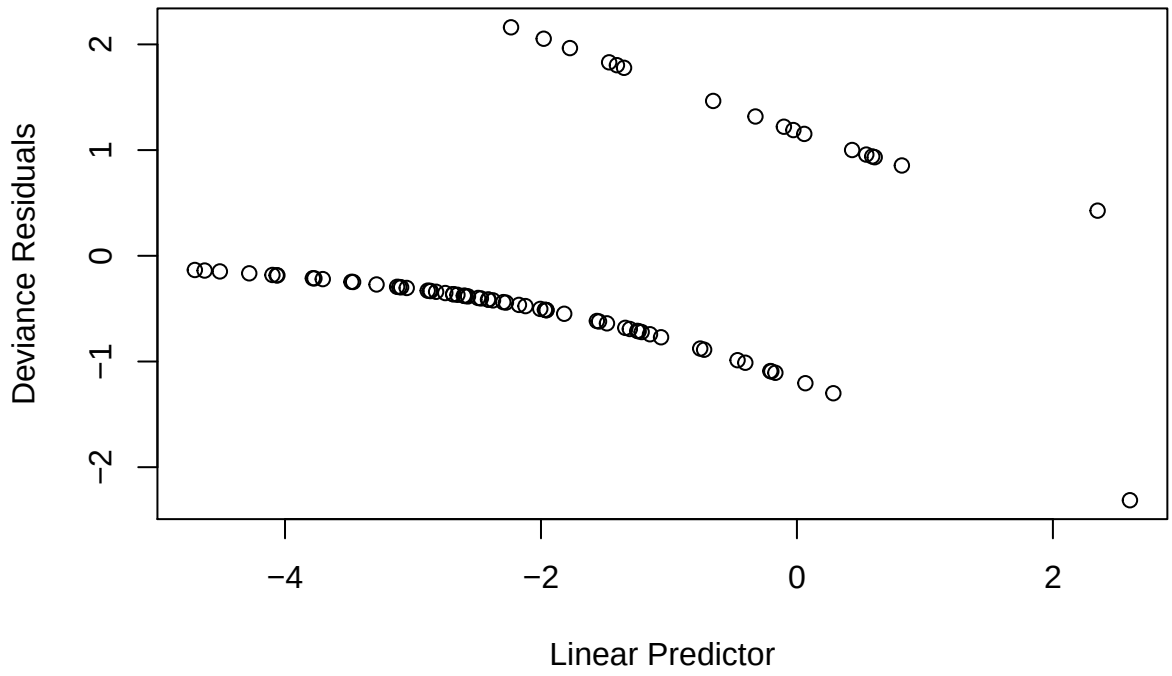
```
lmod <- glm(Kyphosis ~ Age + Number + Start, family=binomial, data=kyphosis)
summary(lmod)
```

```
##
## Call:
## glm(formula = Kyphosis ~ Age + Number + Start, family = binomial,
##     data = kyphosis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3124  -0.5484  -0.3632  -0.1659   2.1613
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.036934   1.449575  -1.405  0.15996
## Age          0.010930   0.006446   1.696  0.08996 .
## Number       0.410601   0.224861   1.826  0.06785 .
## Start       -0.206510   0.067699  -3.050  0.00229 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.234  on 80  degrees of freedom
## Residual deviance: 61.380  on 77  degrees of freedom
## AIC: 69.38
##
## Number of Fisher Scoring iterations: 5
```

We also plot the deviance residuals against the fitted values.

```
residuals <- residuals(lmod) #Should not add type = "response"
linpred <- predict(lmod)
plot(residuals ~ linpred, xlab="Linear Predictor", ylab="Deviance Residuals")
```



We gain no insight into the fit of the model.

Part (c)

We produce a binned residual plot.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

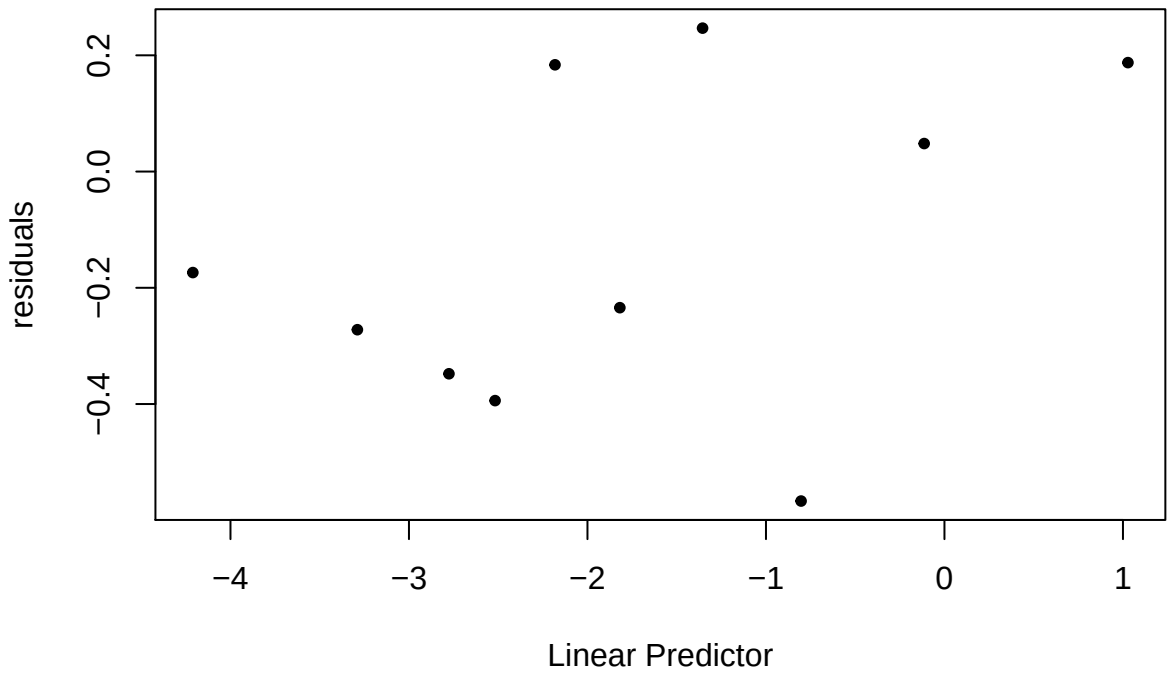
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(faraway)
```

```
kyphosis <- mutate(kyphosis, residuals=residuals(lmod), linpred=predict(lmod))
gdf <- group_by(kyphosis, ntile(linpred,10)) # 10 bins
diagdf <- summarise(gdf, residuals=mean(residuals), linpred=mean(linpred))
plot(residuals ~ linpred, diagdf, xlab="Linear Predictor", pch=20)
```



There is no distinctive pattern in the binned residual plot. We can also draw a similar plot with the following code.

```r
library(arm)
```

```
## Warning: package 'arm' was built under R version 3.6.2
```

```
## Loading required package: MASS
```

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

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

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

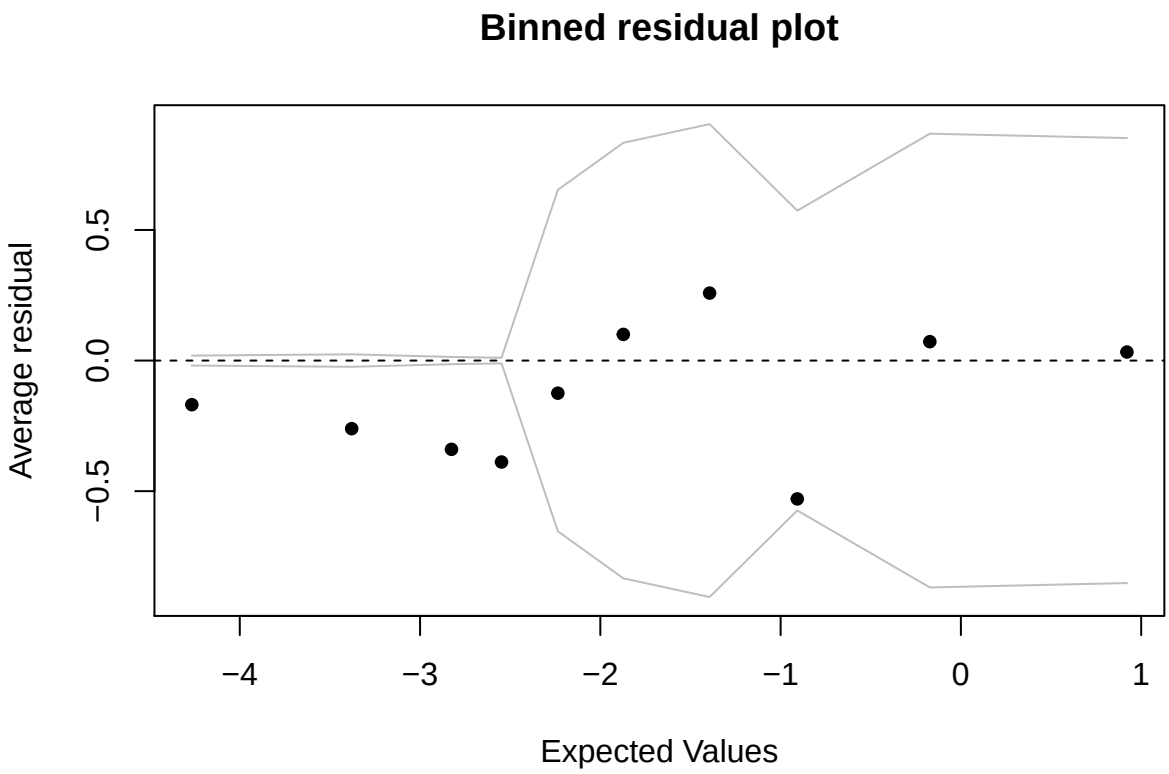
```
##
## arm (Version 1.10-1, built: 2018-4-12)
```

```
## Working directory is C:/Users/taebi/Desktop/UNC/2020 Spring/STOR 590 (TA)
```

```
##
## Attaching package: 'arm'
```

```
## The following objects are masked from 'package:faraway':
##
##     fround, logit, pfround
```

```
binnedplot(linpred, residuals) #Automatically chooses # bins
```



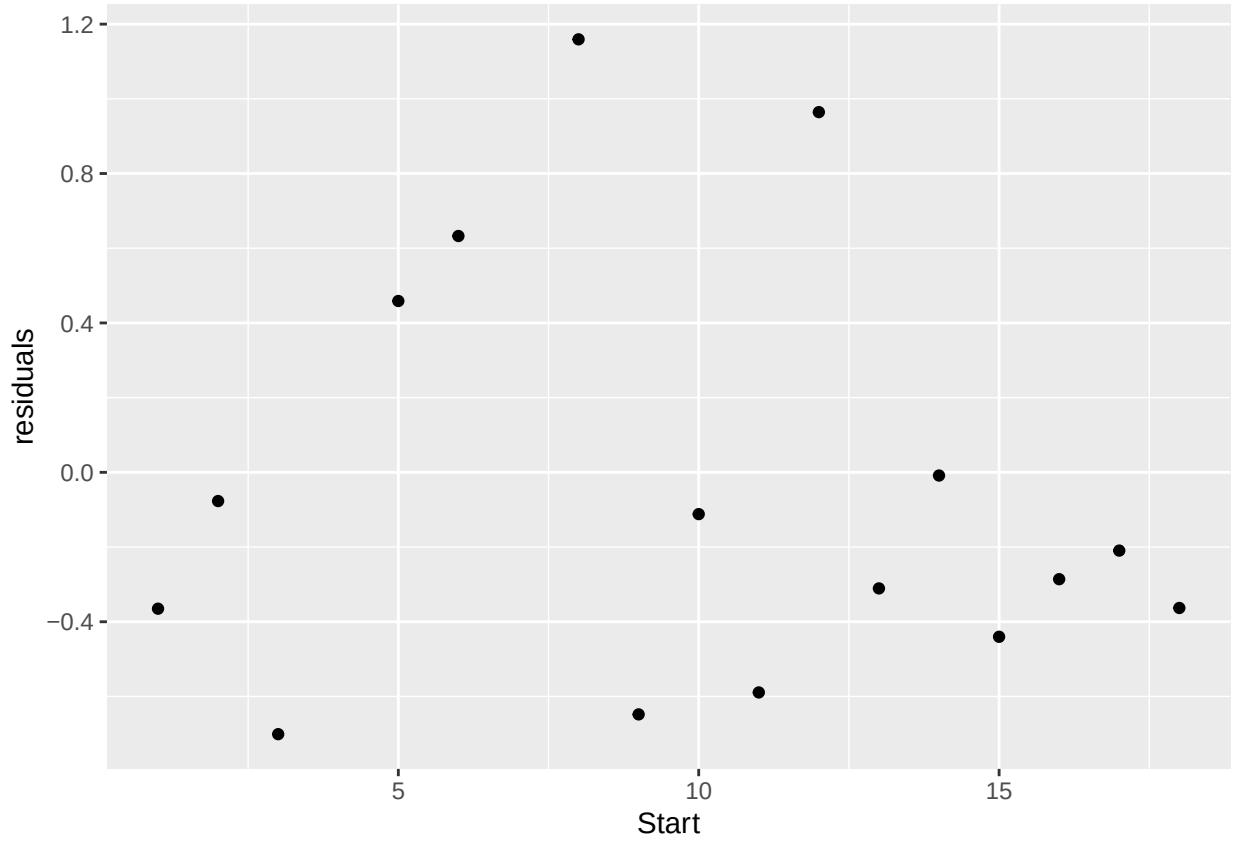
Part (d)

We plot the residuals against the `Start` predictor, using binning as appropriate.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
group_by(kyphosis, Start) %>%
  summarise(residuals=mean(residuals)) %>%
  ggplot(aes(x=Start, y=residuals,)) + geom_point()
```

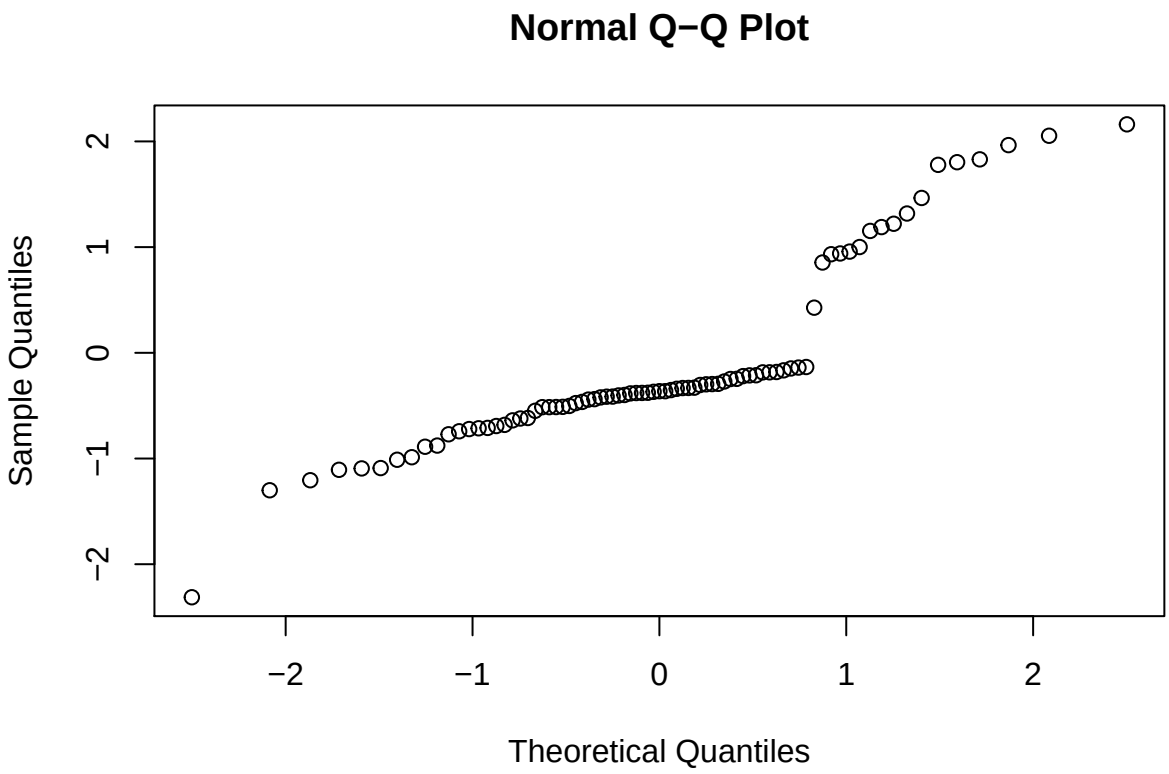


The plot shows nothing remarkable, except maybe a trend that slightly curves downward.

Part (e)

We produce a normal QQ plot for the residuals.

```
qqnorm(residuals(lmod))
```

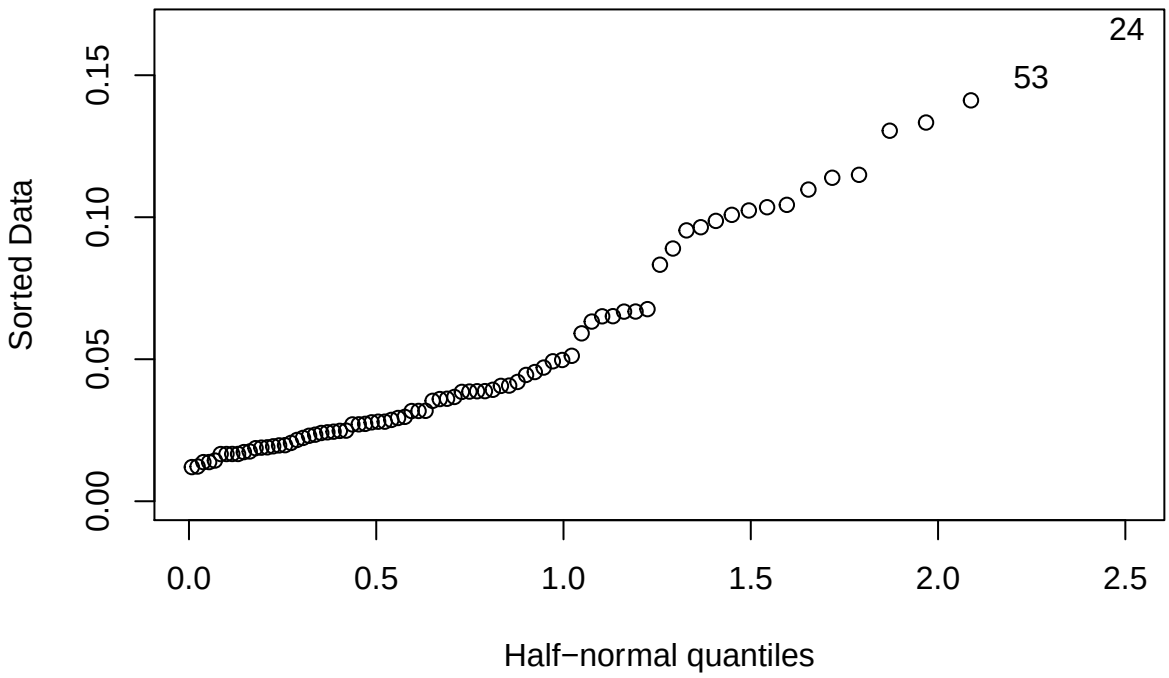


We see that the plot is very far from the desired linear relationship. We see two clusters of points corresponding to $y = 0$ and $y = 1$. But there is no reason to expect these residuals to be normally distributed so this does not raise any concern.

Part (f)

We make a plot of the leverages.

```
halfnorm(hatvalues(lmod))
```

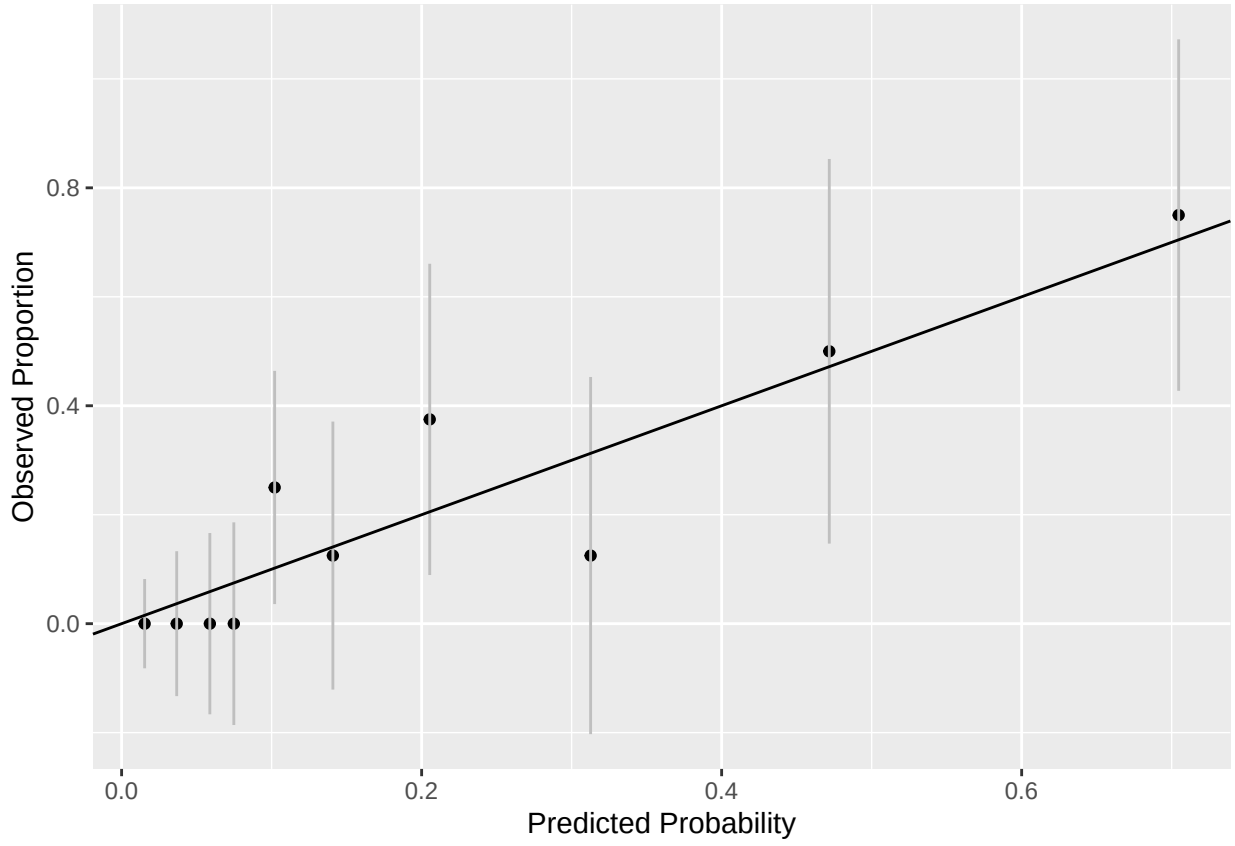


Observation 24 and 53 are two possible outliers. However, they are considerably linear along with other points, and thus we may ignore their influence.

Part (g)

We check the goodness of fit for the model. We also compute the Hosmer-Lemeshow statistic and associated p-value.

```
kyphosis <- mutate(kyphosis, predprob=predict(lmod, type="response"))
gdf <- group_by(kyphosis, ntile(linpred,10))
hldf <- summarise(gdf, y=sum(y), ppred=mean(predprob), count=n())
hldf <- mutate(hldf, se.fit=sqrt(ppred*(1-ppred)/count))
ggplot(hldf, aes(x=ppred, y=y/count, ymin=y/count-2*se.fit, ymax=y/count+2*se.fit)) +
  geom_point() + geom_linerange(color=grey(0.75)) + geom_abline(intercept=0, slope=1) +
  xlab("Predicted Probability") + ylab("Observed Proportion")
```




```
hlstat <- with(hldf, sum((y-count*ppred)^2/(count*ppred*(1-ppred))))
c(hlstat,nrow(hldf))
```

```
## [1]  6.346379 10.000000
```

```
1-pchisq(6.346379, 10-1)
```

```
## [1] 0.7048128
```

The plot shows no consistent deviation from what is expected, and the p-value is around 0.7, so we detect no lack of fit.

Part (h)

We use the model to classify the subjects into predicted outcomes using a 0.5 cutoff. We also produce cross-tabulation of these predicted outcomes with the actual outcomes.

```
kyphosis <- mutate(kyphosis, predout=ifelse(predprob < 0.5, "no", "yes"))
xtabs( ~ Kyphosis + predout, kyphosis)
```

```
##          predout
## Kyphosis  no yes
##   absent  61   3
##   present 10   7
```

The result shows that when kyphosis is actually present, the probability that this model would predict a present outcome is around $\frac{7}{10+7} = 0.41$. The name for this characteristic of the test is *sensitivity*.

Page 47 Exercise 4

First, load in the data and learn about the variables by:

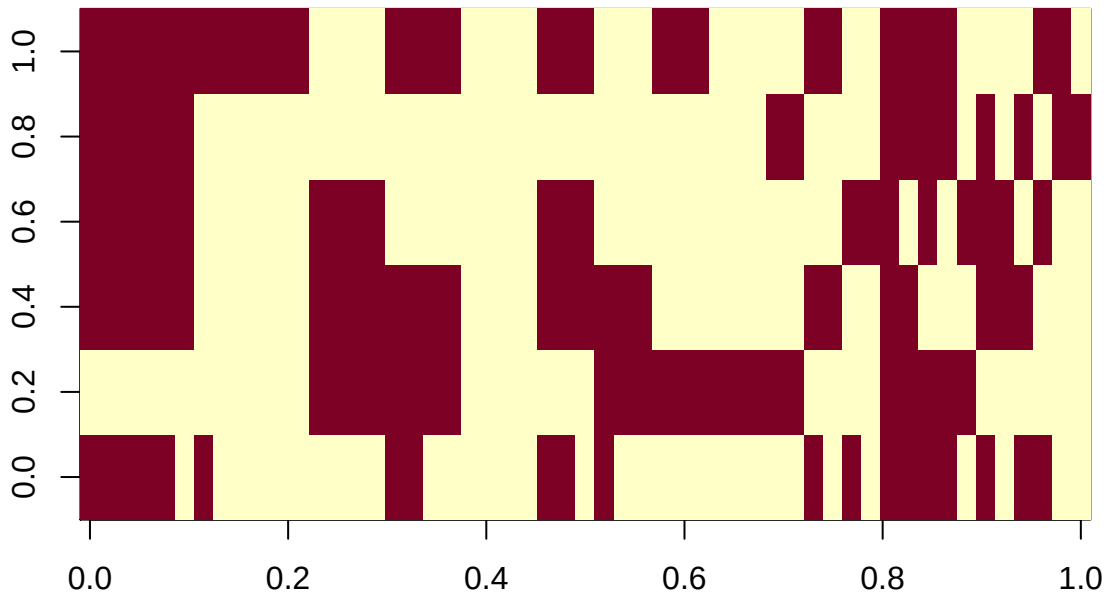
```
data(nodal, package="boot")
```

```
help(nodal, package="boot")
```

Part (a)

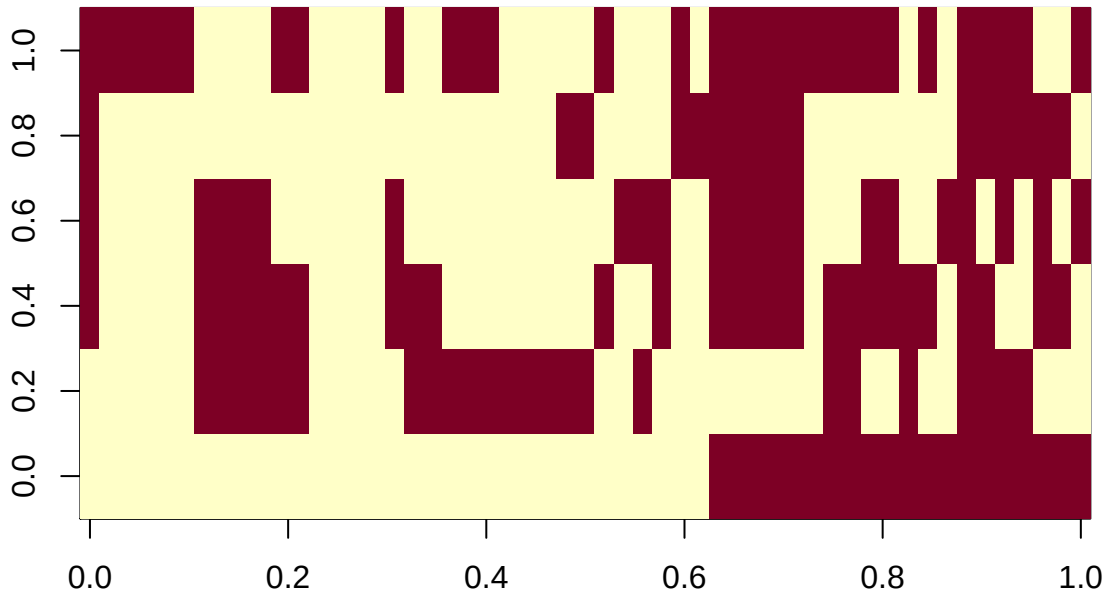
We construct a plot consisting of a binary image of the data as:

```
nodal$m <- NULL
image(as.matrix(nodal))
```

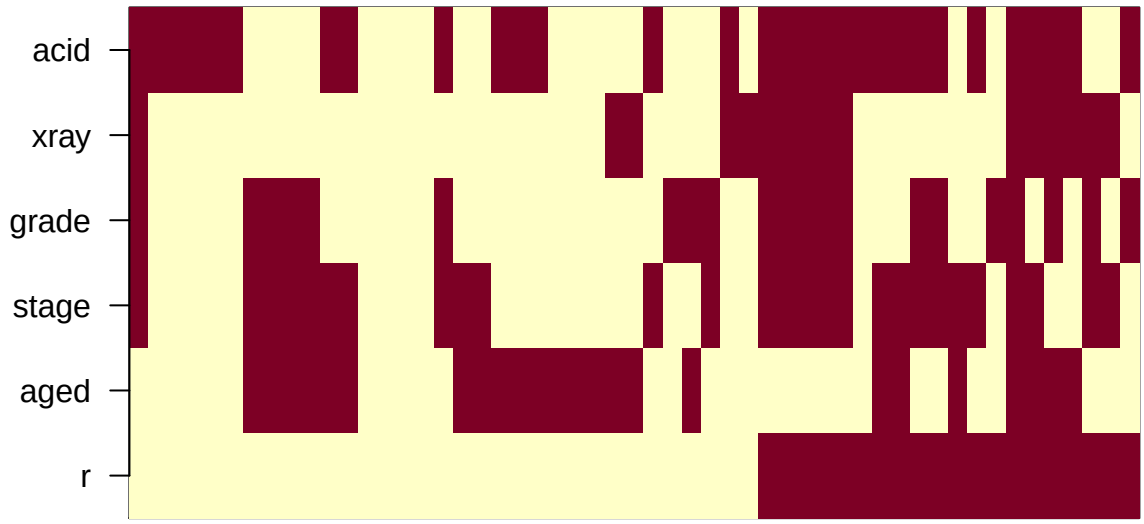


We can improve this plot by ordering the cases on the response and labeling the axes informatively using the `axis` command.

```
nodal <- nodal[order(nodal$r),]
image(as.matrix(nodal))
```




```
image(as.matrix(nodal), xaxt= "n", yaxt= "n")
axis(2, at=seq(0,1,length.out=ncol(nodal) ), labels= colnames(nodal), las= 2)
```



Note that we did not label the x-axis since the index of `nodal` is not in ascending order anymore.

Part (b)

We fit an appropriate model with nodal outcome as the response and the other five variables as predictors.

```
lmod <- glm(r ~ ., family=binomial, data=nodal)
summary(lmod)
```

```
##
## Call:
## glm(formula = r ~ ., family = binomial, data = nodal)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3317  -0.6653  -0.2999   0.6386   2.1502
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.0794     0.9868  -3.121   0.0018 **
## aged         -0.2917     0.7540  -0.387   0.6988
## stage         1.3729     0.7838   1.752   0.0799 .
## grade         0.8720     0.8156   1.069   0.2850
## xray          1.8008     0.8104   2.222   0.0263 *
## acid          1.6839     0.7915   2.128   0.0334 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.252  on 52  degrees of freedom
## Residual deviance: 47.611  on 47  degrees of freedom
## AIC: 59.611
##
## Number of Fisher Scoring iterations: 5
```

```
1-pchisq(21.072,5)
```

```
## [1] 0.0007850749
```

Since the p-value is so small, we are confident that there is some relationship between the predictors and the response.

Part (c)

We fit a smaller model that removes `aged` and `grade` from the model.

```
lmodr <- glm(r ~ stage + xray + acid, family=binomial, data=nodal)
anova(lmod, lmodr, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: r ~ aged + stage + grade + xray + acid
## Model 2: r ~ stage + xray + acid
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        47     47.611
## 2        49     49.180 -2  -1.5696   0.4562
```

We see that `aged` and `grade` are not significant in a model that already includes other predictors.

Part (d)

We calculate the increase of the odds of nodal involvement by having a serious x-ray result compared to a nonserious result, using the smaller model. We also give a 95% confidence interval for the increse of the odds.

```
exp(lmodr$coefficients[3]*1) - 1
```

```
##     xray
## 5.764077
```

```
exp(confint(lmodr)[3,]*1) - 1
```

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

```
##      2.5 %     97.5 %
##  0.5796698 34.4906523
```

Part (e)

We fit a model with all five predictors and all their two-way interactions.

```
lmod_int <- glm(r ~ .*., family=binomial, data=nodal)
summary(lmod_int)
```

```
##
## Call:
## glm(formula = r ~ . * ., family = binomial, data = nodal)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.89302  -0.00016   0.00000   0.00015   1.89302
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -38.55   31256.83  -0.001    0.999
## aged          -54.81   32392.29  -0.002    0.999
## stage          74.21   11937.81   0.006    0.995
## grade          38.55   31256.83   0.001    0.999
## xray           17.18   16447.57   0.001    0.999
## acid           36.94   31256.83   0.001    0.999
## aged:stage     18.46    7928.05   0.002    0.998
## aged:grade     31.97   38433.53   0.001    0.999
## aged:xray      56.96   17366.75   0.003    0.997
## aged:acid      36.35   31407.11   0.001    0.999
## stage:grade   -92.34   14996.24  -0.006    0.995
## stage:xray    -17.06   28388.60  -0.001    1.000
## stage:acid    -72.60   11937.81  -0.006    0.995
## grade:xray     36.50   31778.73   0.001    0.999
## grade:acid     54.48   31841.25   0.002    0.999
## xray:acid     -35.70    8157.65  -0.004    0.997
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.252  on 52  degrees of freedom
## Residual deviance: 29.542  on 37  degrees of freedom
## AIC: 61.542
##
## Number of Fisher Scoring iterations: 20
```

Note that the standard errors of the coefficients are very large due to the multicollinearity.

Part (f)

We use the bias-reduced model fitting method to fit the model of the previous question.

```
library(brglm)
```

```
## Warning: package 'brglm' was built under R version 3.6.2
```

```
## Loading required package: profileModel
```

```
## Warning: package 'profileModel' was built under R version 3.6.2
```

```
## 'brglm' will gradually be superseded by 'brglm2' (https://cran.r-project.org/package=brglm2), which
```

```
bmod <- brglm(r ~ .*., family=binomial, data=nodal)
summary(bmod)
```

```
##
## Call:
## brglm(formula = r ~ . * ., family = binomial, data = nodal)
##
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6438     1.6586  -1.594    0.111
## aged         -0.6995     1.8290  -0.382    0.702
## stage         2.3156     1.8196   1.273    0.203
## grade         2.6023     1.9506   1.334    0.182
## xray          0.6900     2.0954   0.329    0.742
## acid          1.2729     1.8124   0.702    0.483
## aged:stage    0.3728     1.8119   0.206    0.837
## aged:grade   -1.5480     1.8594  -0.833    0.405
## aged:xray     1.4587     1.9780   0.737    0.461
## aged:acid     0.3634     1.7722   0.205    0.838
## stage:grade  -2.8219     1.8258  -1.546    0.122
## stage:xray    0.8375     2.2944   0.365    0.715
## stage:acid   -0.9608     1.6387  -0.586    0.558
## grade:xray   -0.1890     2.2626  -0.084    0.933
## grade:acid    0.5575     1.8918   0.295    0.768
## xray:acid    -0.2560     1.8611  -0.138    0.891
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.450  on 52  degrees of freedom
## Residual deviance: 38.997  on 37  degrees of freedom
## Penalized deviance: 45.20152
## AIC:  70.997
```

It shows that the interaction between `stage` and `grade` is the largest.

Part (g)

We use the bias-reduced model to classify the cases in the dataset, and then compare these to the actual classifications.

```
bprob <- predict(bmod, type="response")
nodal <- mutate(nodal, bpred=ifelse(bprob>0.5, 1, 0))
sum(nodal$r != nodal$bpred)
```

```
## [1] 8
```

```r
mean(nodal$r != nodal$bpred)
```

```
## [1] 0.1509434
```

We repeat this comparison for the model in (b).

```r
lprob <- predict(lmod, type="response")
nodal <- mutate(nodal, lpred=ifelse(lprob>0.5, 1, 0))
sum(nodal$r != nodal$lpred)
```

```
## [1] 10
```

```r
mean(nodal$r != nodal$lpred)
```

```
## [1] 0.1886792
```

These misclassification rates are not perfectly reasonable estimates of how these models will perform in the future since we have not calculated the rates using the cross-validation.