HW 3

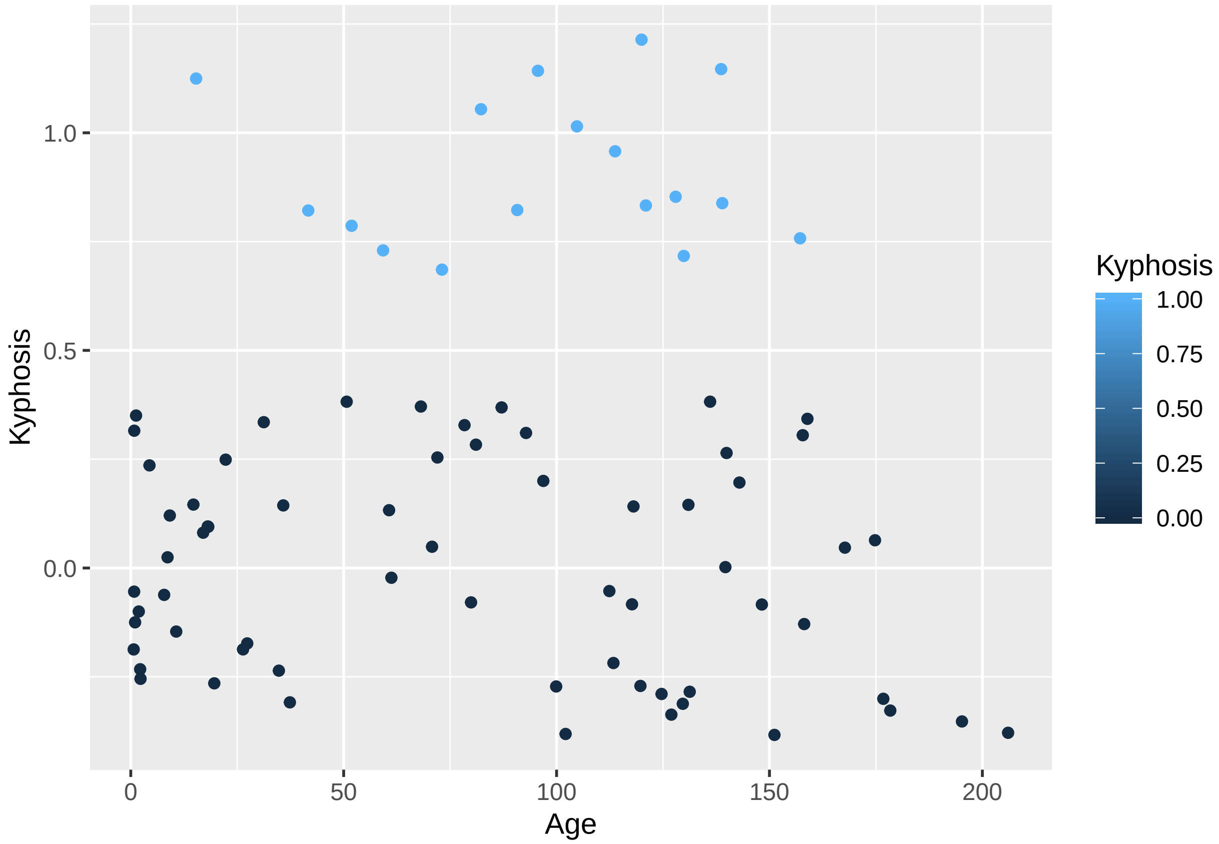
Ted Henson

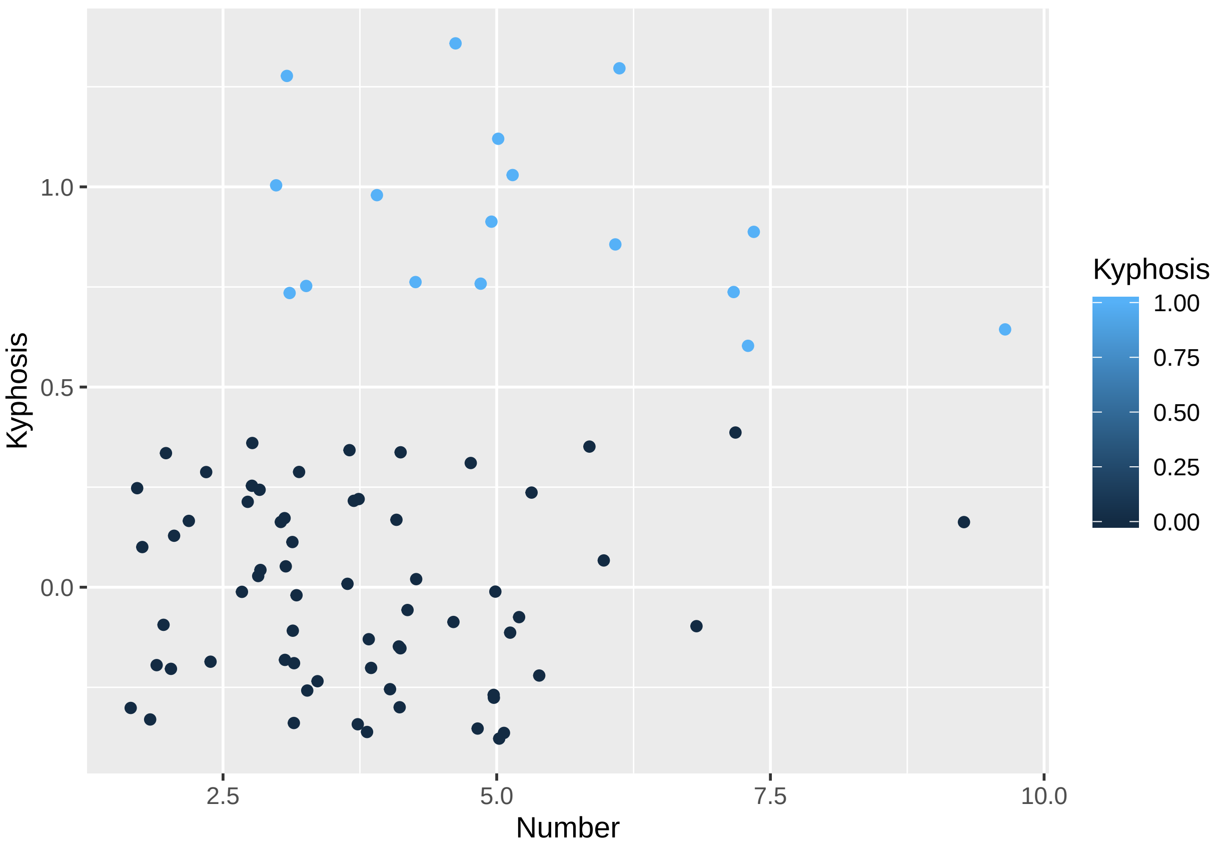
1/31/2020

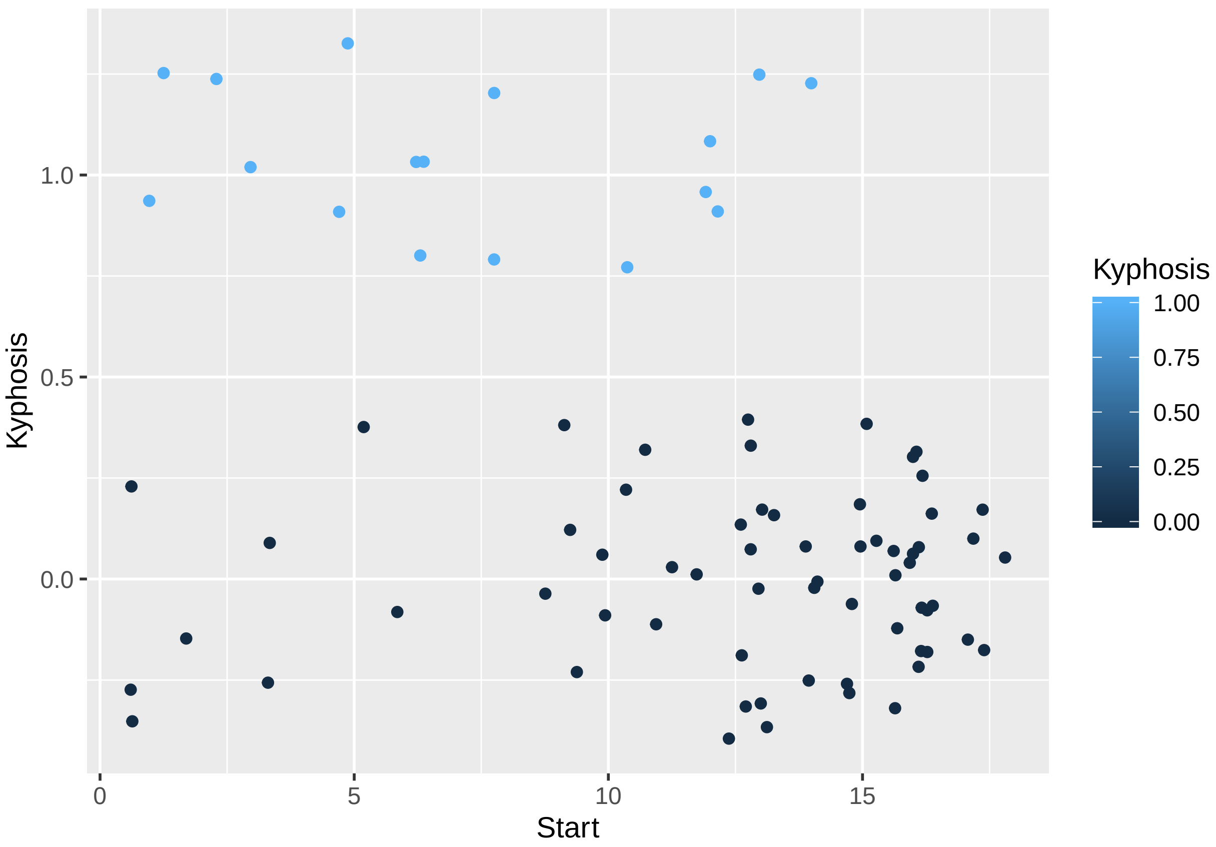
Question 3

a)

```
## [1] "Plots of Kyphosis versus predictor variables"
```

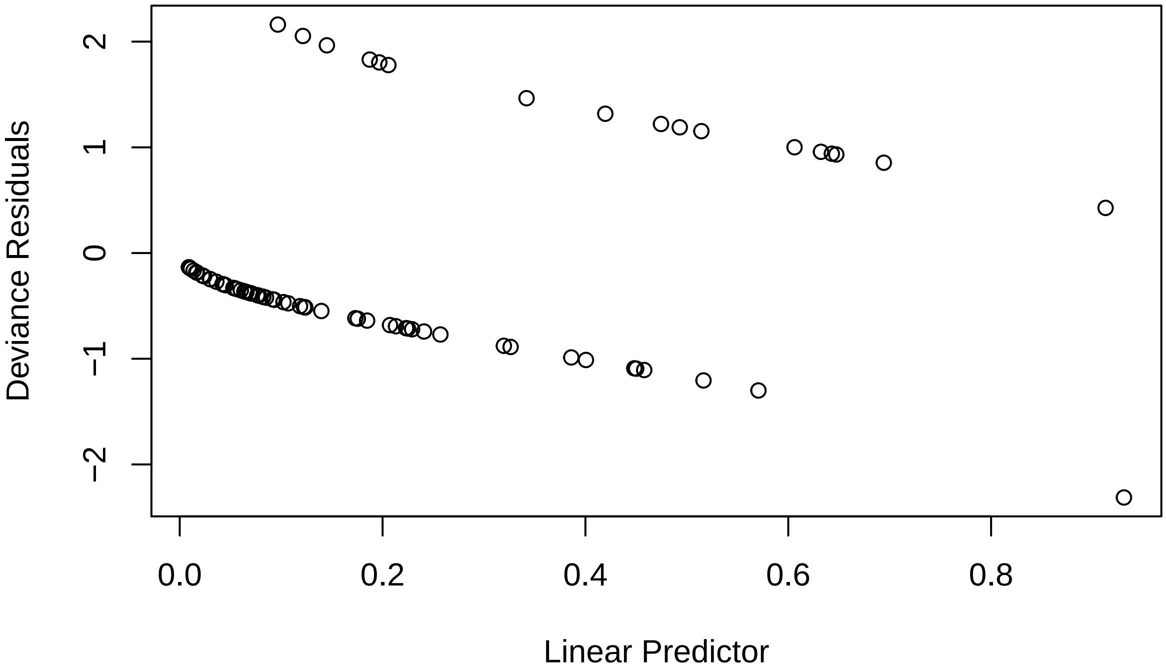



Kyphosis
1.0
0.5
0.0
2.5
5.0
7.5
10.0
Number
Kyphosis
1.00
0.75
0.50
0.25
0.00



The presence of Kyphosis appears to be rare for higher values of the Start and Number variables.. It seems to be most common for middle values of the Age variable.

b)



There are only a few cases where Kyphosis is predicted with a probability over 50%. Most of the larger deviance residuals are on the lower end of the predicted response, although there was one outlier in the far upper quantile.

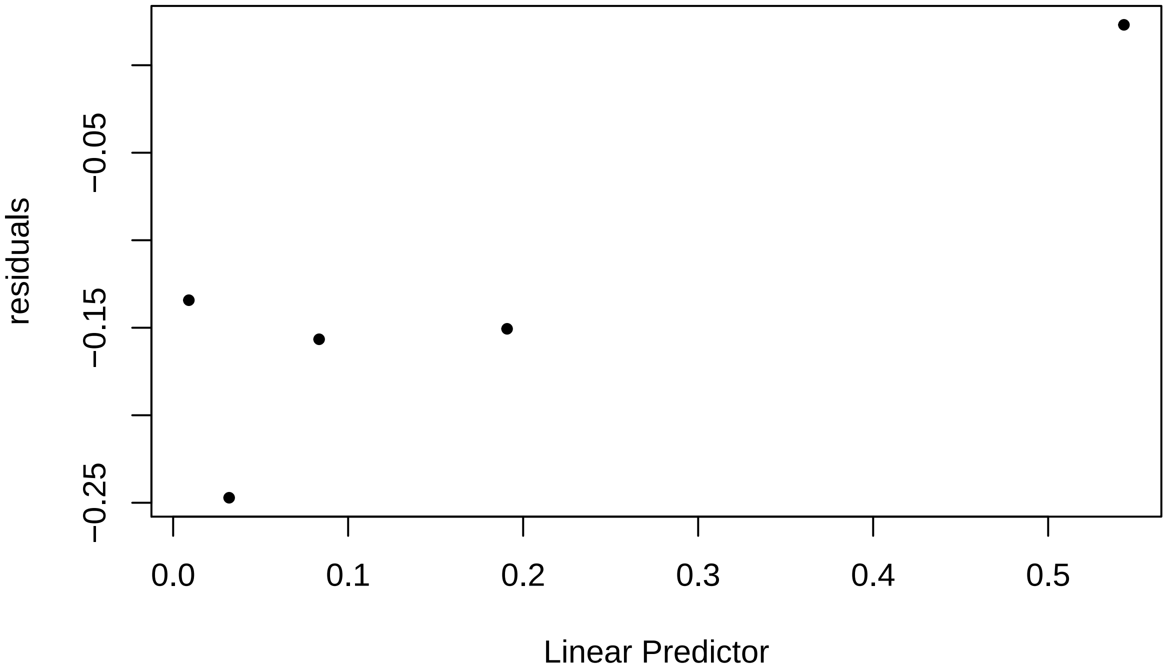
c)

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

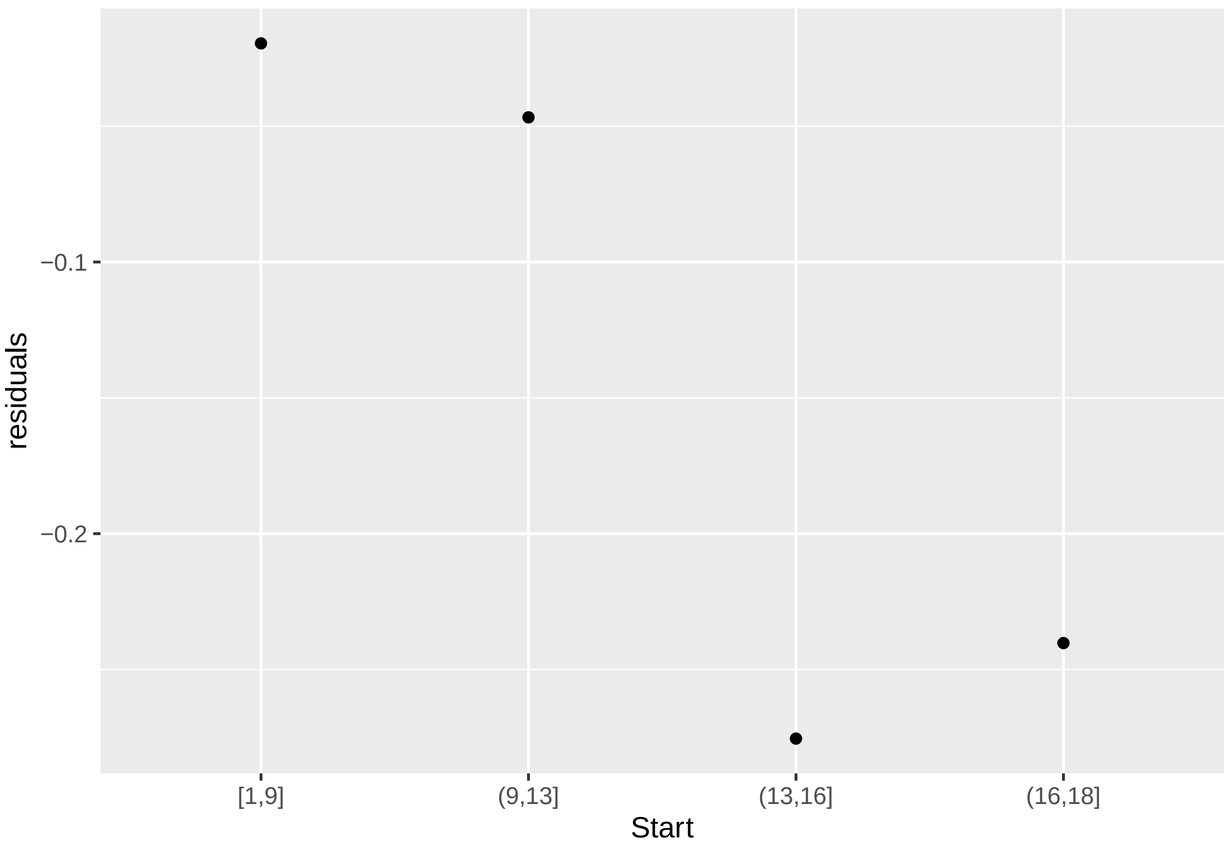
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
## Warning: Factor `cut(linpred, breaks = unique(quantile(linpred)))` contains
## implicit NA, consider using `forcats::fct_explicit_na`
```

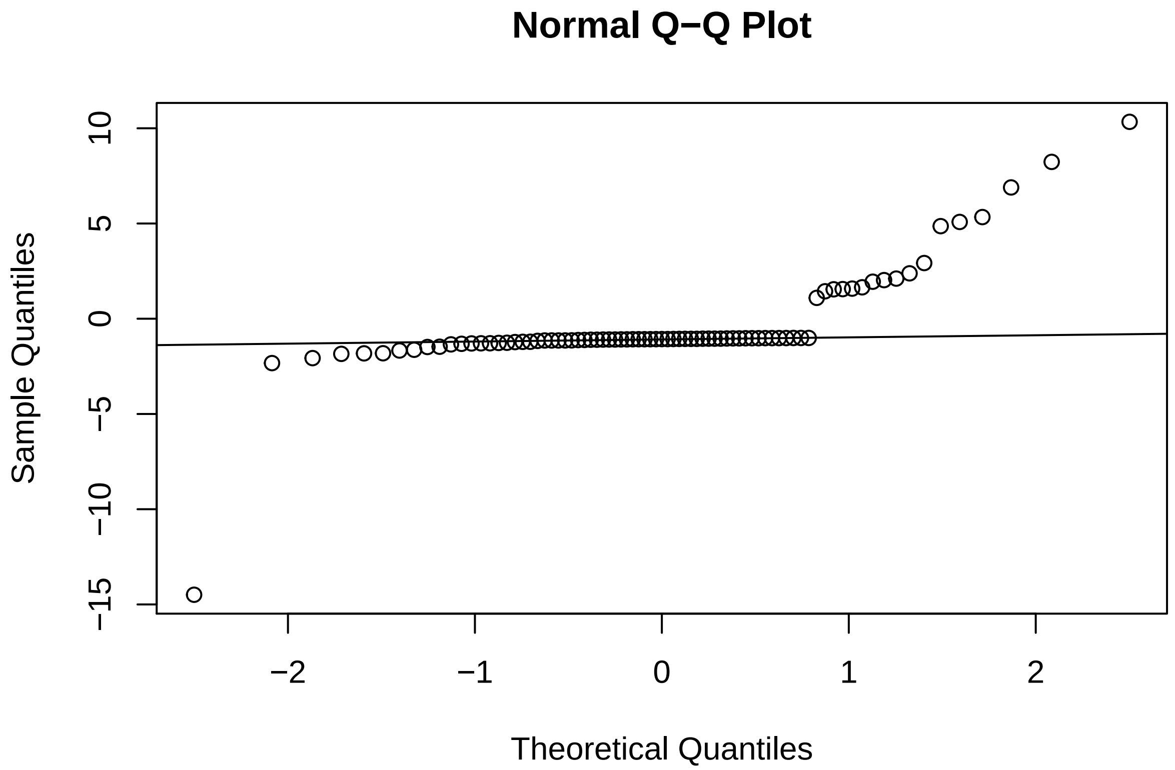


The mean of the residuals seem to be largest for in the lower quantiles of the predicted response values.



#d

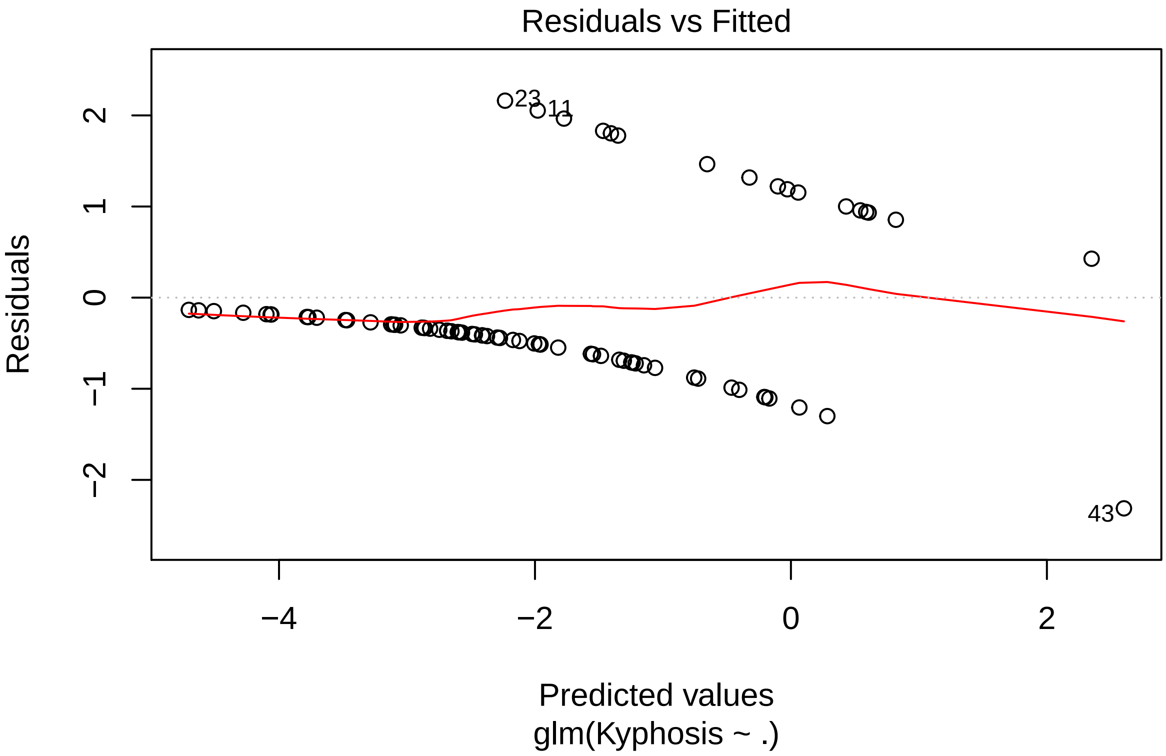
The mean of the residuals are larger at larger values of Start variable.

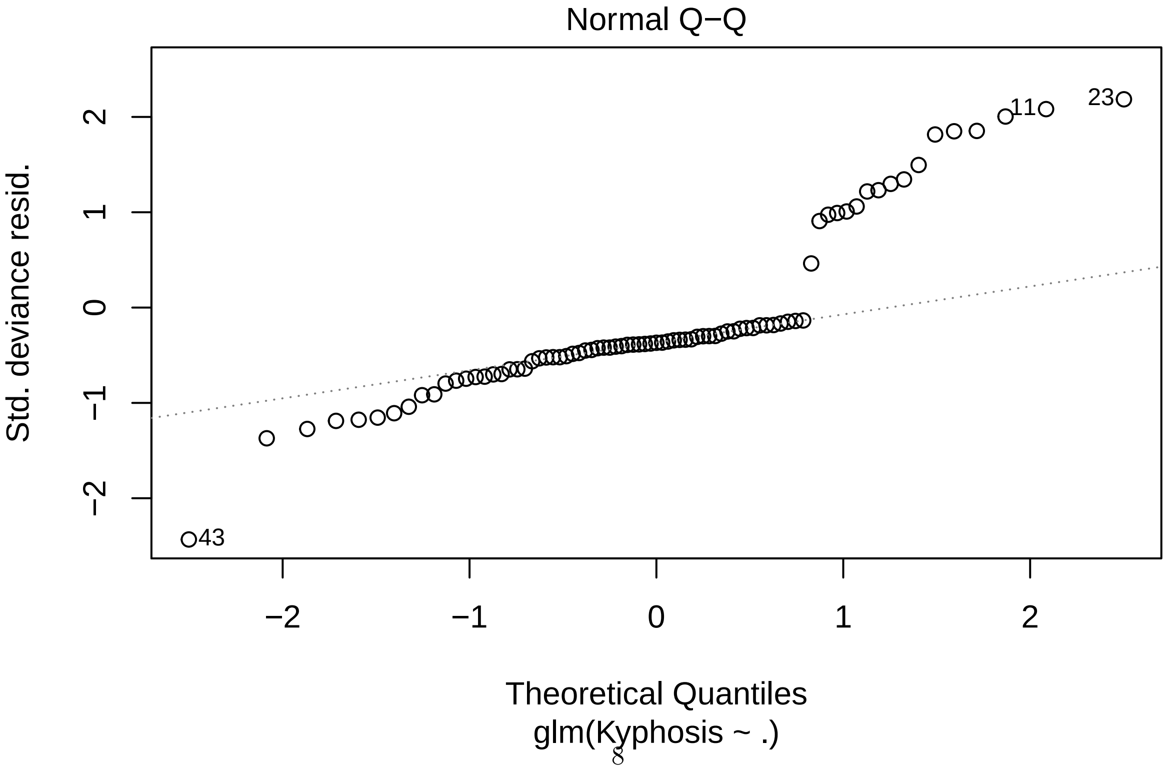


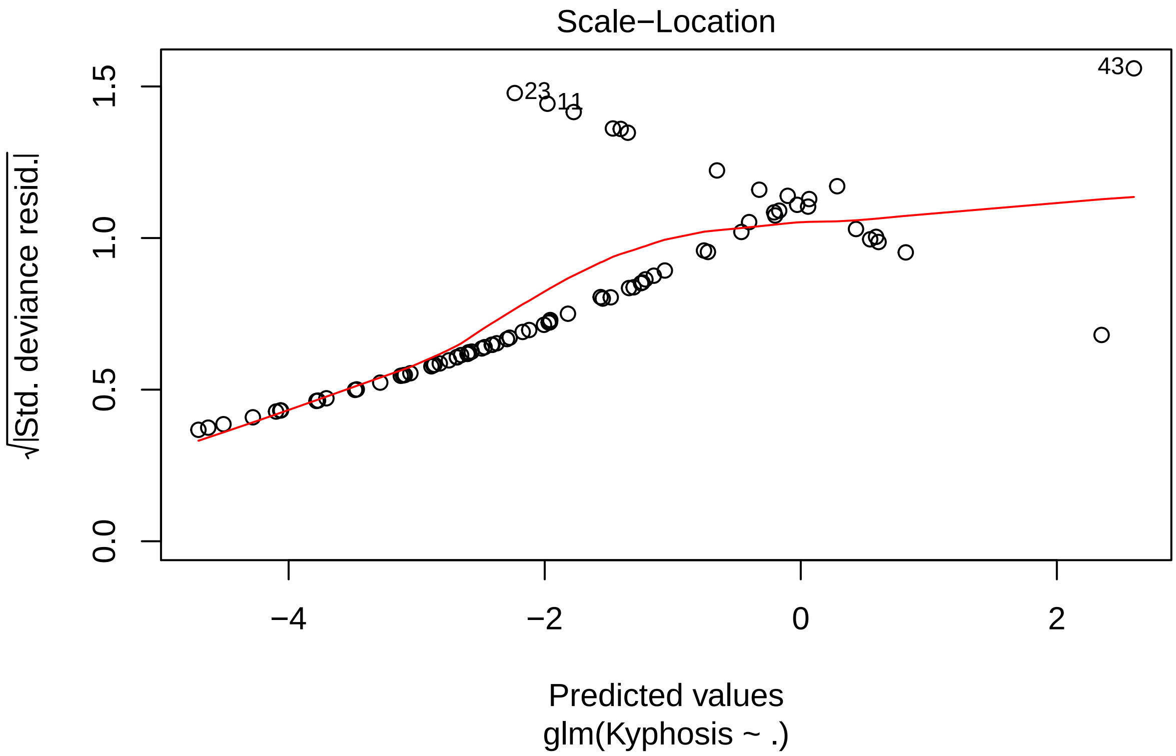
#e

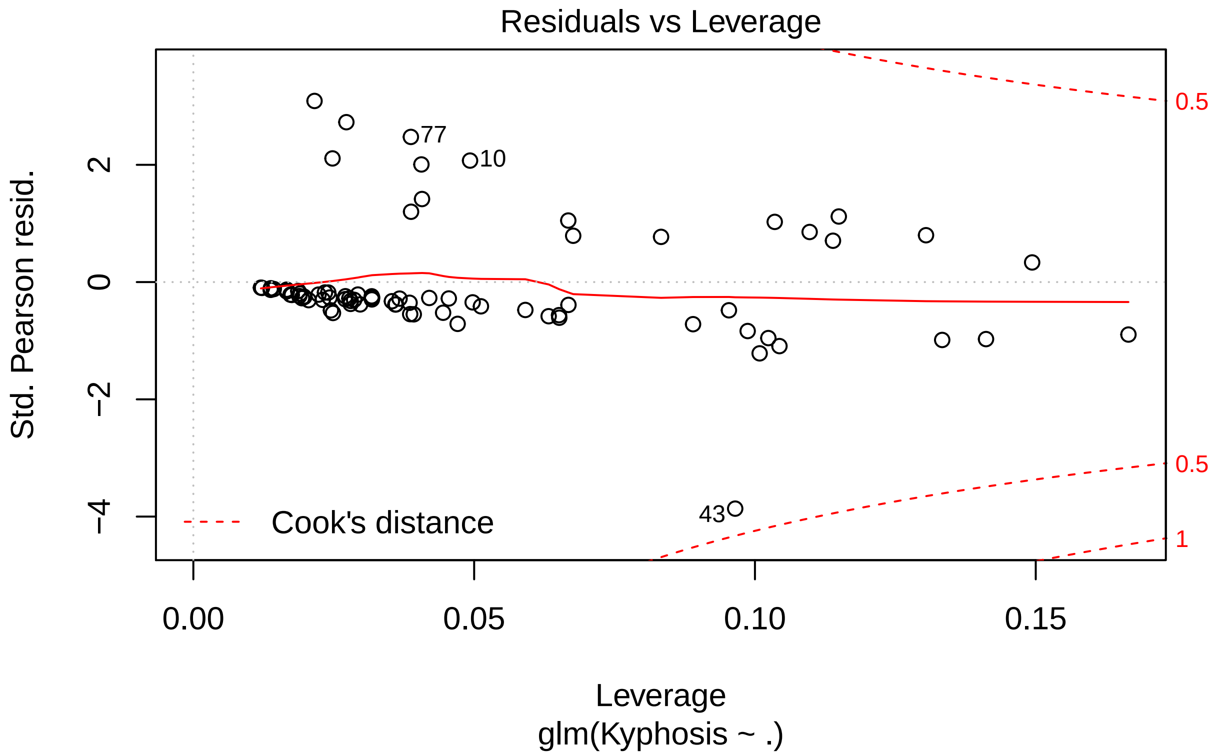
The residuals are much larger in the upper quantiles of the response variable.

f)


Residuals vs Fitted


Normal Q−Q

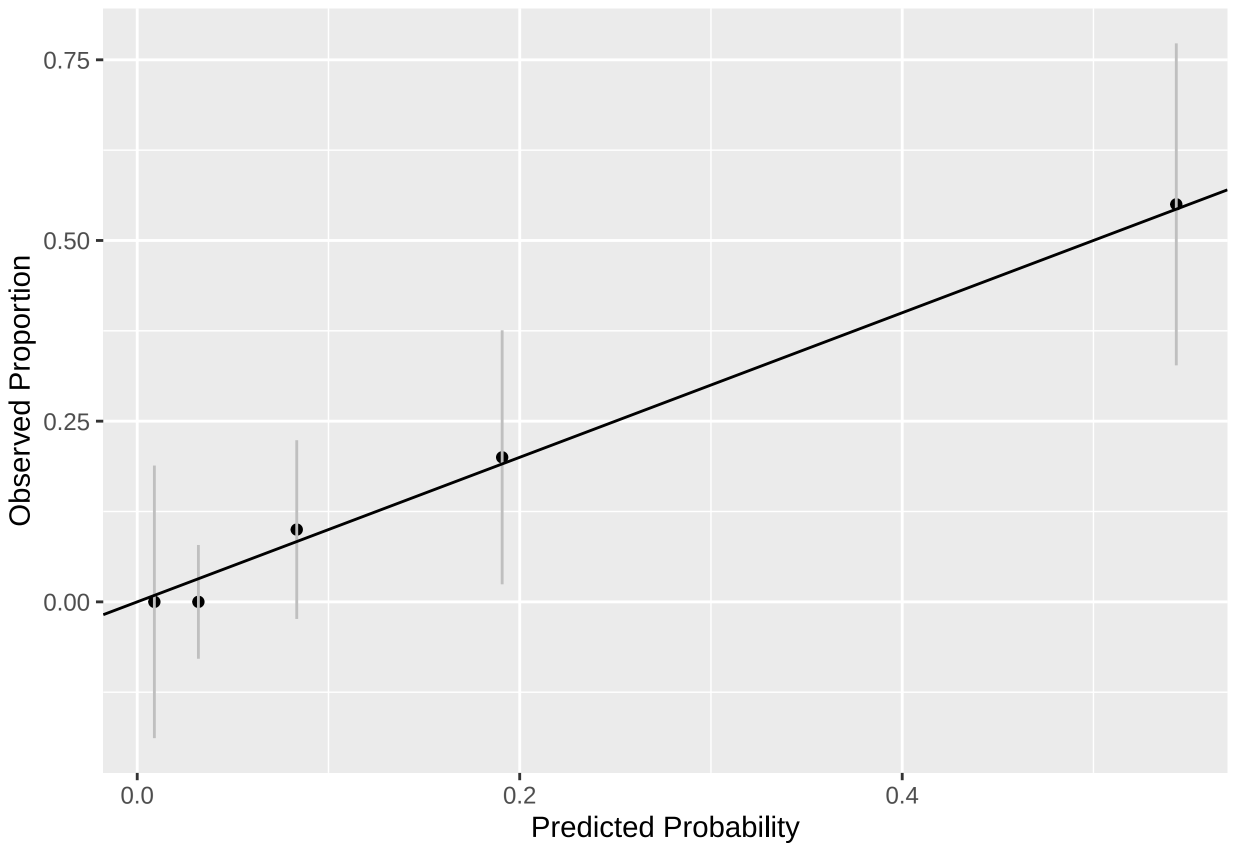

Scale−Location
1.5
1.0
0.5
0.0
√|Std. deviance resid.|
23
11
43
−4
−2
0
2
Predicted values
glm(Kyphosis ~ .)



As the fourth plot shows, there are no points that are influential as they are within the dashed cooked lines, although point 43 is very close.

g)

```
## Warning: Factor `cut(linpred, breaks = unique(quantile(linpred)))` contains
## implicit NA, consider using `forcats::fct_explicit_na`
```



```
## Loading required package: reshape

##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
##     rename

## Loading required package: MASS

##
## Attaching package: 'MASS'

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

##
##  Hosmer and Lemeshow test (binary model)
##
## data:  kyphosis$Kyphosis, kyphosis$predprob
## X-squared = 0.056836, df = 0, p-value < 2.2e-16
```

The p value for the hosmer-lemeshow test was about .05 using 3 quantile groups and basically zero using two quantile groups. Therefore, the model predicting the two groups has statistical significance of a good fit.

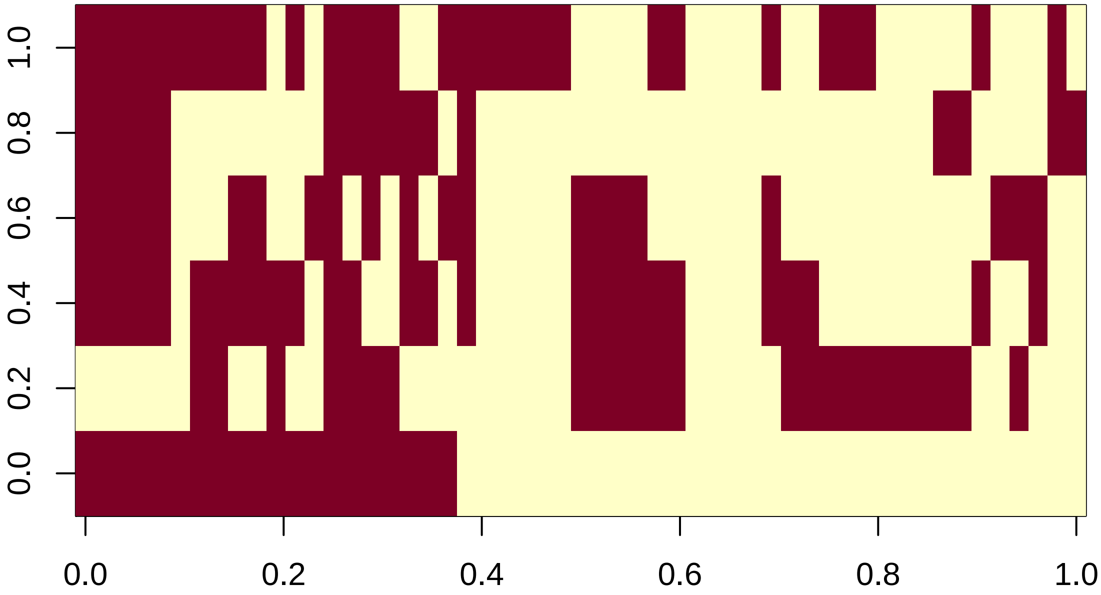
h)

```
##          predout
## Kyphosis no yes
##        0 61   3
##        1 10   7
```

When Kyphosis is present, the model would predict it is present with probability 0.4117647. This would be the true positive rate.

Question 4

a)



b)

```
##
## 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
```

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

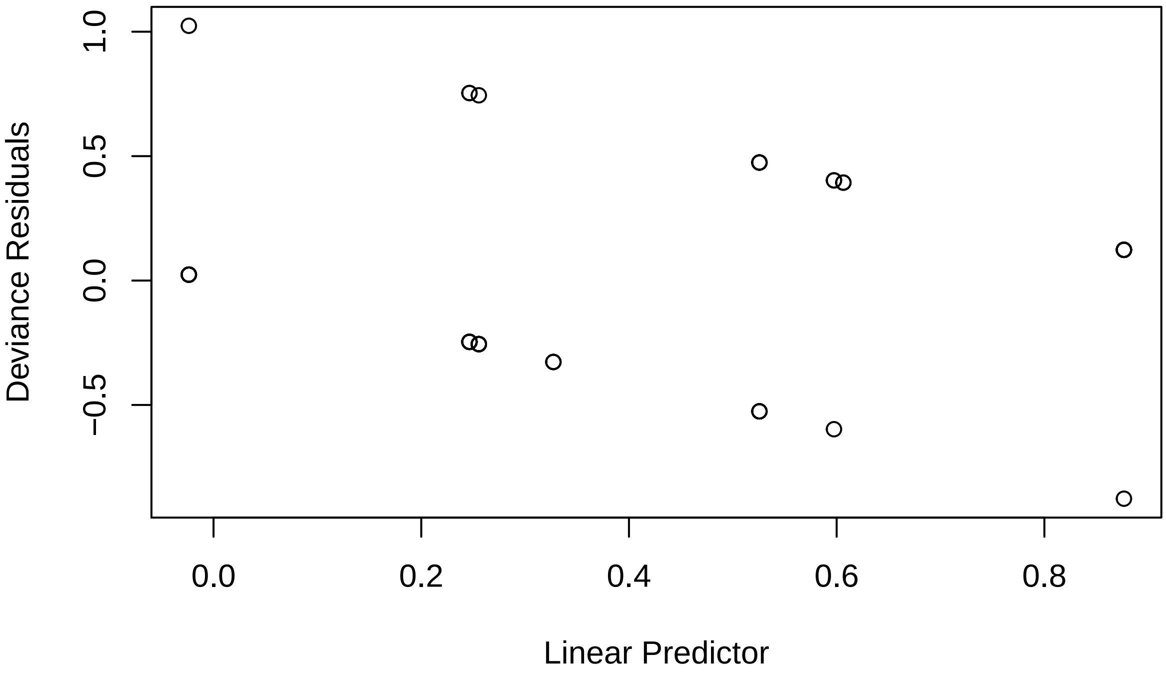
```
##                  2.5 %    97.5 %
## (Intercept) -5.3002078 -1.362046
## aged        -1.7956919  1.214897
## stage       -0.1313106  3.000236
## grade       -0.7346704  2.539458
## xray         0.2669083  3.523458
## acid         0.2089821  3.378604
```

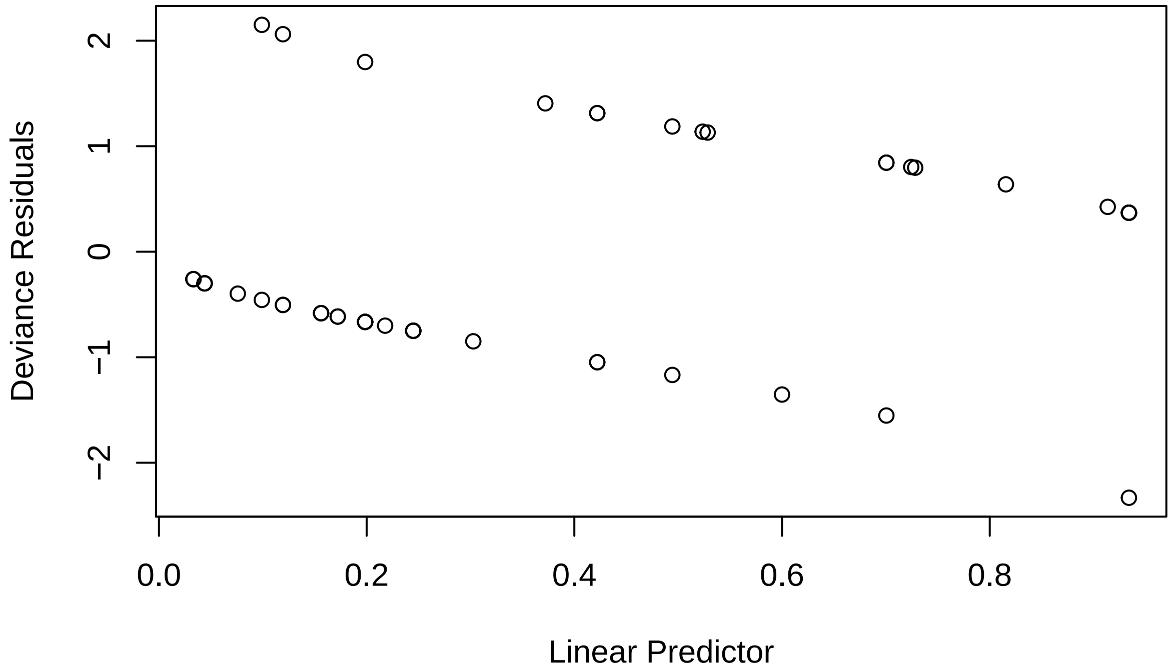
Yes the Xray and Acid variables both have p values below the .05 threshold and strictly positive 95% confidence intervals.

c)

```
##
## Call:
## glm(formula = r ~ stage + xray + acid, data = nodal)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8766  -0.2554   0.0238   0.1234   1.0238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0238     0.1000  -0.238  0.81298
## stage         0.2792     0.1137   2.455  0.01768 *
## xray          0.3510     0.1244   2.822  0.00688 **
## acid          0.2702     0.1149   2.352  0.02274 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.165372)
##
##     Null deviance: 12.4528  on 52  degrees of freedom
## Residual deviance:  8.1032  on 49  degrees of freedom
## AIC: 60.872
```

```
##
## Number of Fisher Scoring iterations: 2
```





Yes the smaller model could be preferred over the larger model because the deviance residuals are much smaller.

d)

A serious x ray result increases the odds of having nodal involvement by 1.4205216 compared to a non serious x ray result. The 95% confidence interval for the change in odds is 1.0579312, 1.6522249

e)

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = r ~ .^2, 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   31256.82  -0.001    0.999
## aged          -54.81   32392.28  -0.002    0.999
## stage          74.21   11937.81   0.006    0.995
## grade          38.55   31256.82   0.001    0.999
## xray           17.18   16447.57   0.001    0.999
## acid           36.94   31256.82   0.001    0.999
## aged:stage     18.46    7928.05   0.002    0.998
## aged:grade     31.97   38433.52   0.001    0.999
## aged:xray      56.96   17366.75   0.003    0.997
## aged:acid      36.35   31407.10   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   31841.24   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
```

The standard errors are very large because there is not enough data (information) to estimate all of the parameters. Some of the two way interactions are probably collinear with the original variables as well.

f)

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

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

```
##
## Call:
## brglm(formula = r ~ .^2, 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
```

The stage and grade interaction has the largest coefficient.

g)

```
##    predout
## r    0  1
##   0 29  4
##   1  4 16
```

```
##    predout
## r    0  1
##   0 30  3
##   1  7 13
```

The bias reduced model had one less miss classification than the full model. These models are probably over estimates of the classification rates of what one could expect in the future. Cross validation or boot strapping would need to be implemented to get a better idea of the future classification accuracy.