Nodal Involvement in Prostate Cancer

Kevin Dang

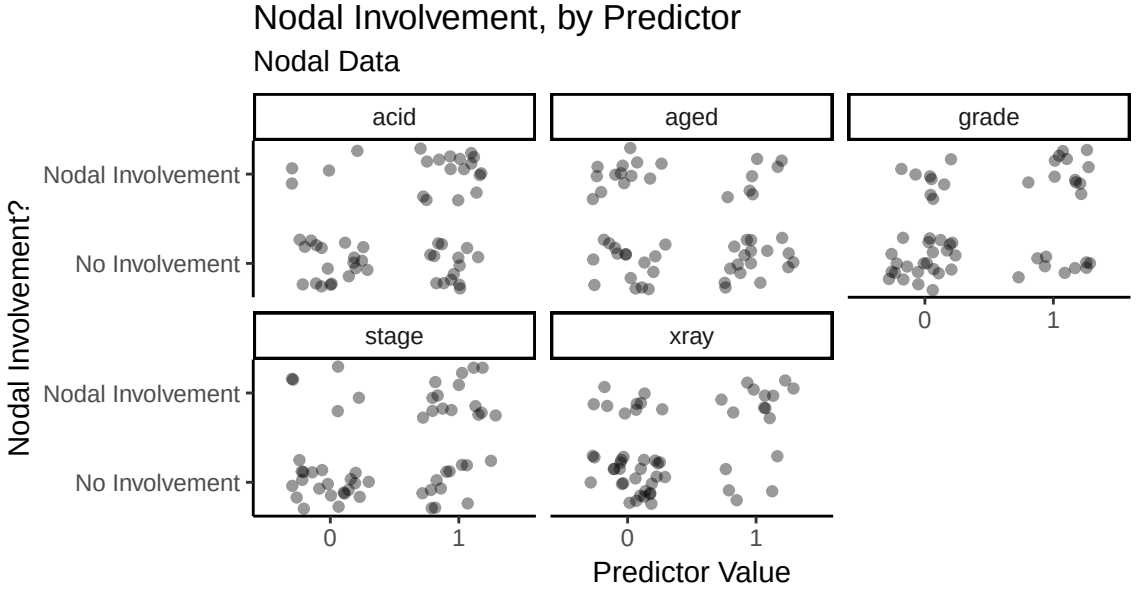
When deciding on how to treat prostate cancer, physicians use a cancer staging system which takes into account the presence of cancer in the surrounding lymph nodes, referred to as nodal involvement. My analysis involves determining whether prostate cancer has spread to the lymph nodes based on certain characteristics. Starting with the *Nodal Involvement by Predictor*[1] graph, it is difficult to tell whether any of the five characteristics are successful in predicting nodal involvement. Upon closer inspection, it appears as though `stage`, `acid` and `xray` have more true positive and true negative data points than false positive and false negative data points, which means that they may have a higher success rate when predicting nodal involvement. An initial *binary logistic regression model*[2] shows that `acid` and `xray` are considered somewhat significant, `stage` is close to the standard significance level of 0.05, while `age` and `grade` are not close to the significance level at all. To explore the potentially significant predictors further, a second *binary logistic regression model*[3] was fit, with nodal involvement ("`r`") as the response and `stage`, `acid` and `xray` as the predictors. The *analysis of deviance table*[4] for the second model shows a significant reduction in the residual deviance as each of the three variables are added to the null model. In regards to the model assumptions, the values are discrete (0 or 1) and there are also no outliers in the data since the z-value for each predictor is under 3. Also, there is low intercorrelation among the predictors, as shown in the *correlation matrix*[5]. To clarify what each predictor represents, `stage` is a measure of the size and position of the tumour, `xray` indicates how serious the cancer is from an X-ray reading, and `acid` represents the level of acid phosphatase in the blood serum. These three variables may be helpful indicators of nodal involvement in prostate cancer, from evidence provided by the model. However, physicians should proceed with caution as there are some observations which incorrectly predict nodal involvement.

[1] Appendix A: Nodal Involvement, by Predictor
[2] Appendix A: Binary Logistic Regression Model 1
[3] Appendix A: Binary Logistic Regression Model 2
[4] Appendix A: Binary Logistic Regression Model 2, Analysis of Deviance Table
[5] Appendix A: Correlation Matrix

Appendix A

Nodal Data

```
## Observations: 53
## Variables: 7
## $ m     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ r     <dbl> 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,...
## $ aged  <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,...
## $ stage <fct> 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,...
## $ grade <fct> 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0,...
## $ xray  <fct> 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ acid  <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1,...
```



Binary Logistic Regression Model 1

```
##
## Call:
## glm(formula = r ~ aged + stage + grade + xray + acid, family = binomial,
##     data = nodal_tbl)
##
## 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 **
## aged1        -0.2917     0.7540  -0.387   0.6988
## stage1        1.3729     0.7838   1.752   0.0799 .
## grade1        0.8720     0.8156   1.069   0.2850
## xray1         1.8008     0.8104   2.222   0.0263 *
## acid1         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
```

Binary Logistic Regression Model 2

```
##
## Call:
## glm(formula = r ~ stage + xray + acid, family = binomial, data = nodal_tbl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1231  -0.6620  -0.3039   0.4710   2.4892
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.0518     0.8420  -3.624  0.00029 ***
## stage1        1.6453     0.7297   2.255  0.02414 *
## xray1         1.9116     0.7771   2.460  0.01390 *
## acid1         1.6378     0.7539   2.172  0.02983 *
## ---
## 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: 49.180  on 49  degrees of freedom
## AIC: 57.18
##
## Number of Fisher Scoring iterations: 5
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: r
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                     52     70.252
## stage  1   7.6995        51     62.553 0.005524 **
## xray   1   8.0901        50     54.463 0.004451 **
## acid   1   5.2822        49     49.180 0.021544 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correlation Matrix

```
##         aged stage grade  xray  acid
## aged    1.00  0.06 -0.19 -0.10 -0.20
## stage   0.06  1.00  0.41  0.15  0.13
## grade -0.19  0.41  1.00  0.22  0.01
## xray  -0.10  0.15  0.22  1.00  0.16
## acid  -0.20  0.13  0.01  0.16  1.00
```