Assignment 2

Kevin Dang

August 8, 2018

Important Note: Footers are used in this report. Italicized text containing a footer notation indicates that there is a plot or code output that is to be referred to in the appendix.

Nodal Involvement in Prostate Cancer

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

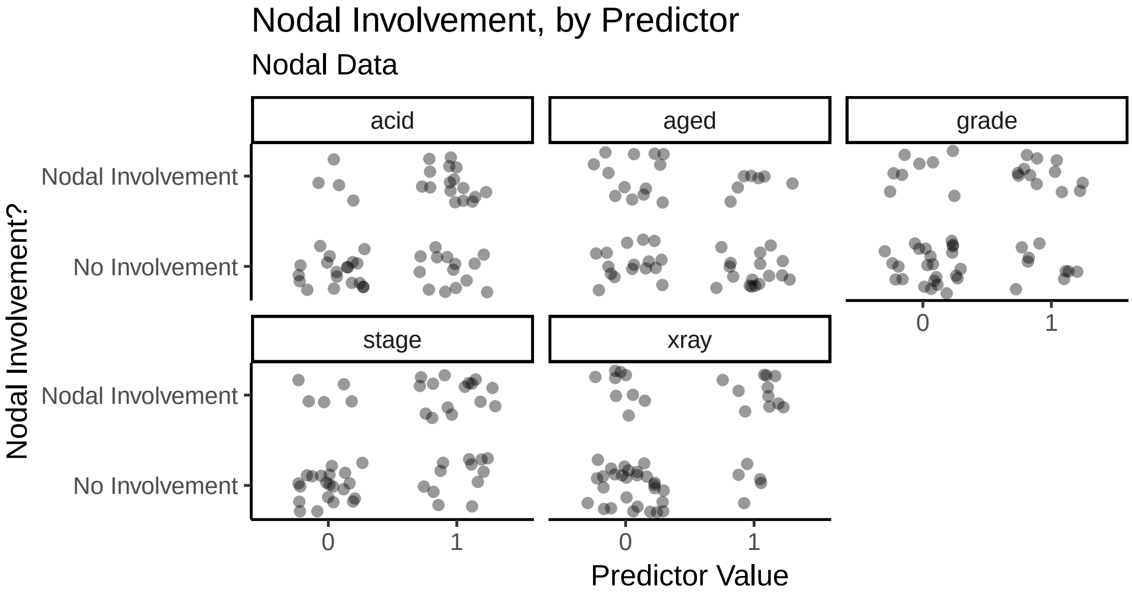
[4] Appendix A: Binary Logistic Regression Model 2, Analysis of Deviance Table

[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

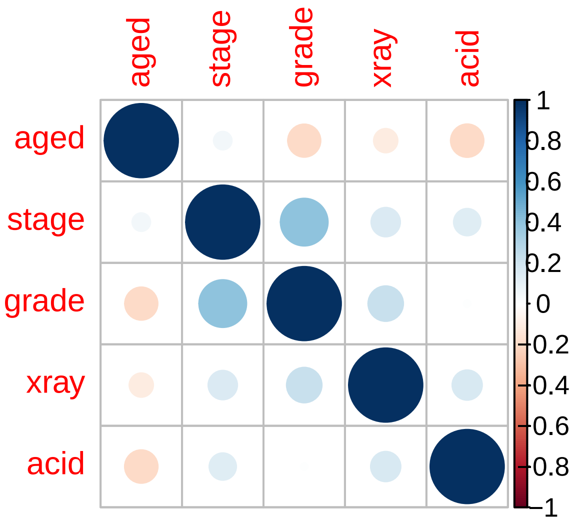
```
##
## Call:
## glm(formula = r ~ aged + stage + grade + xray + acid, family = binomial,
##     data = nodal_tbl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3317  -0.6653  -0.2999   0.6386   2.1502
##
## Coefficients:
##              Estimate Std. Error z value 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
```



Smoking, Age and Death

Smoking is a major health concern among the population, however many individuals of numerous age groups continue to smoke. The goal is to analyze potential relationships between age group, smoking status and mortality rate among women. Looking at potential relationships, the first *table*[6] shows that a greater proportion of smokers in the study were alive after 20 years than non-smokers. In addition, the *binomial regression model*[7] for mortality against smoking shows a significant negative relationship between the variables, which indicates that smoking decreases mortality rate. This is unexpected, but another factor (age) has not been taken into account, which could explain this unusual relationship. Also, the residual deviance is quite large compared to its degrees of freedom, so this model is not a good fit. To investigate this unintuitive relationship, a second *table*[8] was created to show the relationship between smoking and age in groups of `dead` or `alive`. In this table, there is a larger proportion of younger women who smoke, relative to older women who smoke. Many of these younger women who smoke were still alive after 20 years into the study, while many of the older women passed away. Another *binomial regression model*[9] is fit to the data, this time containing age groups as a predictor. This model is a very strong fit since the residual deviance is quite small relative to its degrees of freedom. Now that `age` has been accounted for, the `smoker` variable is positively correlated with mortality; this is an example of Simpson's paradox. The dependence of smoking status and mortality rate are explained by their respective relationship with age (i.e. smoking and mortality are dependent, conditional on age). If investigators in this study did not measure age, they may have incorrectly concluded that smoking correlates with a lower risk of death. In observational studies such as this one, investigators need to be careful in drawing conclusions before considering other factors that can influence relationships between the variables of interest.

[6] Appendix B, Table 1

[7] Appendix B, Binomial Regression 1

[8] Appendix B, Table 2

[9] Appendix B, Binomial Regression 2

Appendix B

Smoking Data

```
## Observations: 14
## Variables: 4
## $ age    <fct> 18-24, 18-24, 25-34, 25-34, 35-44, 35-44, 45-54, 45-54,...
## $ smoker <dbl> 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0
## $ alive  <dbl> 53, 61, 121, 152, 95, 114, 103, 66, 64, 81, 7, 28, 0, 0
## $ dead   <dbl> 2, 1, 3, 5, 14, 7, 27, 12, 51, 40, 29, 101, 13, 64
```

Table 1

```
##
## smoker      dead     alive
##      0 0.3142077 0.6857923
##      1 0.2388316 0.7611684
```

Binomial Regression 1

```
##
## Call:
## glm(formula = cbind(dead, alive) ~ smoker, family = binomial,
##     data = smoking_tbl)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -9.052  -5.674  -1.869   5.776  12.173
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.78052    0.07962  -9.803  < 2e-16 ***
## smoker      -0.37858    0.12566  -3.013  0.00259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 641.5  on 13  degrees of freedom
## Residual deviance: 632.3  on 12  degrees of freedom
## AIC: 683.29
##
## Number of Fisher Scoring iterations: 4
```

Table 2

```
## , ,  = dead
##
##       age
## smoker       18-24       25-34       35-44       45-54       55-64
##      0 0.008547009 0.017793594 0.030434783 0.057692308 0.169491525
##      1 0.017094017 0.010676157 0.060869565 0.129807692 0.216101695
##       age
## smoker       65-74         75+
##      0 0.612121212 0.831168831
##      1 0.175757576 0.168831169
##
## , ,  = alive
##
##       age
## smoker       18-24       25-34       35-44       45-54       55-64
##      0 0.521367521 0.540925267 0.495652174 0.317307692 0.343220339
##      1 0.452991453 0.430604982 0.413043478 0.495192308 0.271186441
##       age
## smoker       65-74         75+
##      0 0.169696970 0.000000000
##      1 0.042424242 0.000000000
```

Binomial Regression 2

```
##
## Call:
## glm(formula = cbind(dead, alive) ~ age + smoker, family = binomial,
##     data = smoking_tbl)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.72545  -0.22836   0.00005   0.19146   0.68162
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.8601     0.5939  -6.500 8.05e-11 ***
## age25-34       0.1201     0.6865   0.175 0.861178
## age35-44       1.3411     0.6286   2.134 0.032874 *
## age45-54       2.1134     0.6121   3.453 0.000555 ***
## age55-64       3.1808     0.6006   5.296 1.18e-07 ***
## age65-74       5.0880     0.6195   8.213  < 2e-16 ***
## age75+        27.8073 11293.1430   0.002 0.998035
## smoker          0.4274     0.1770   2.414 0.015762 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 641.4963  on 13  degrees of freedom
## Residual deviance:   2.3809  on  6  degrees of freedom
## AIC: 65.377
##
## Number of Fisher Scoring iterations: 20
```