Smoking, Age and Death

Kevin Dang

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*[1] shows that a greater proportion of smokers in the study were alive after 20 years than non-smokers. In addition, the *binomial regression model*[2] 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*[3] 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*[4] 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.

[1] Appendix B, Table 1
[2] Appendix B, Binomial Regression 1
[3] Appendix B, Table 2
[4] 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
```