**Donald Day**

**Stat Models MSDS 5350**

**Project 2: Logistic Regression for COPD**

Here are the following tables:

Table 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Male | | | Females | | |
|  | Controls | COPD | All | Controls | COPD | All |
| Smoking Status |  |  |  |  |  |  |
| Never | 229 (0.494) | 44 (0.095) | 273 (0.589) | 171 (0.369) | 20 (0.043) | 191 (0.412) |
| Former | 96 (0.282) | 51 (0.150) | 147 (0.432) | 156 (0.459) | 37 (0.109) | 193 (0.568) |
| Current | 165 (0.361) | 72 (0.158) | 237 (0.519) | 180 (0.394) | 40 (0.088) | 220 (0.482) |
|  | Controls | | | COPD | | |
| Years Smoked | Mean Years Smoked: 15.81 | | | Mean Years Smoked: 24.02 | | |
|  | Standard Deviation: 17.13 | | | Standard Deviation: 18.22 | | |
| Age | Mean Age: 56.31 | | | Mean Age: 57.33 | | |
|  | Standard Deviation: 13.93 | | | Standard Deviation: 12.70 | | |

Table 2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Male | | | Female | | |
|  | OR | 95% CI | p-value | OR | 95% CI | p-value |
| Smoking Status | 1.47559 | (1.2058, 1.8113) | < 0.000 | 1.3266 | (1.0156, 1.7435) | < 0.000 |
| Years Smoked | 1.02672 | (1.0166, 1.0371) | <0.000 | 1.0300 | (1.0175, 1.043) | <0.000 |
| Age | 1.00770 | (0.9951, 1.0206) | 0.2356 | 1.0026 | (0.9864, 1.0191) | 0.7557 |

Table 3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Male | | | Female | | |
|  | Adjusted OR | 95% CI | p-value | Adjusted OR | 95% CI | p-value |
| Smoking Status | 0.964864 | (0.672765, 1.3762) | 0.8444 | 0.51544 | (0.32265, 0.8107) | 0.004 |
| Years Smoked | 1.028327 | (1.010429, 1.0470) | 0.002 | 1.06235 | (1.03851, 1.0883) | < 0.000 |
| Age | 0.998636 | (0.983064, 1.0144) | 0.8642 | 0.96769 | (0.94447, 0.9901) | 0.006255 |

When deciding on which predicting variables would work best, I decided to use the AUROC, the area under the ROC curve. When looking at these graphs, typically the one that covers the largest area has the best fit. Another statistic used to determine best fit between variables is the AIC. Typically the lower the better. Here are the results for the different variables.

Smoking Status – AIC = 1282

1: sample.smkstatus <- sample(c(TRUE, FALSE), nrow(copd), replace=TRUE, prob=c(0.8,0.2))

2: train.smkstatus <- copd[sample.smkstatus, ]

3: test.smkstatus <- copd[!sample.smkstatus, ]

4: model.smkstatus <- glm(copd\_num ~ smkstatus, family="binomial", data=train)

5: predicted <- predict(model, test, type="response")

6: plotROC(test.smkstatus$copd\_num, predicted)

Chart

Description automatically generated

Years Smoked – AIC = 1254

Same code as before

Chart

Description automatically generated

Age – AIC = 1297

Chart

Description automatically generated

COPD ~ smkstatus + SmkYrs + Age – AIC = 1019

Chart

Description automatically generated

Looking at the AIC numbers, the multivariable logistic regression came out on top. However, it’s area under the ROC curve had the smallest area out of all the logistic regression variables. Years smoked seems to have the best fit between all the variables. It has the largest area under the ROC curve and the second lowest AIC score. The years smoked variable should be the main predictor variable.