Flu shots

Elena Shevchenko

Data

A local health clinic sent fliers to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 100 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded Y = 1, and a client who did not receive a flu shot was coded Y = 0. In addition, data were collected on their age (X1) and their health awareness. The latter data were combined into a health awareness index (X2), for which higher values indicate greater awareness. Also included in the data was client gender, where males were coded X3 = 1 and females were coded X3 = 0.

Analysis

First data was split into training set and test set. Test set contains 20 observations and the rest 79 points were assigned to the training data. Then logistic regression was fit using training set. Predictors X1 and X2 are significant. Nagelkerke coefficient of determination is 0.41.

```
Call:
glm(formula = Y ~ ., family = binomial, data = train_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.53037  -0.43201  -0.22870  -0.04191   2.50281

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.30248    4.89580   1.287  0.19798
X1           0.08781    0.04641   1.892  0.05845 .
X2          -0.23964    0.07326  -3.271  0.00107 **
X31          0.50848    0.80544   0.631  0.52784
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1: Logistic regression fit

ROC curve was plot after applying model to the test data. AUC is 0.89. Threshold for predicted probabilities is 0.78. Confusion matrix is presented below. The test error rate is 0.9.

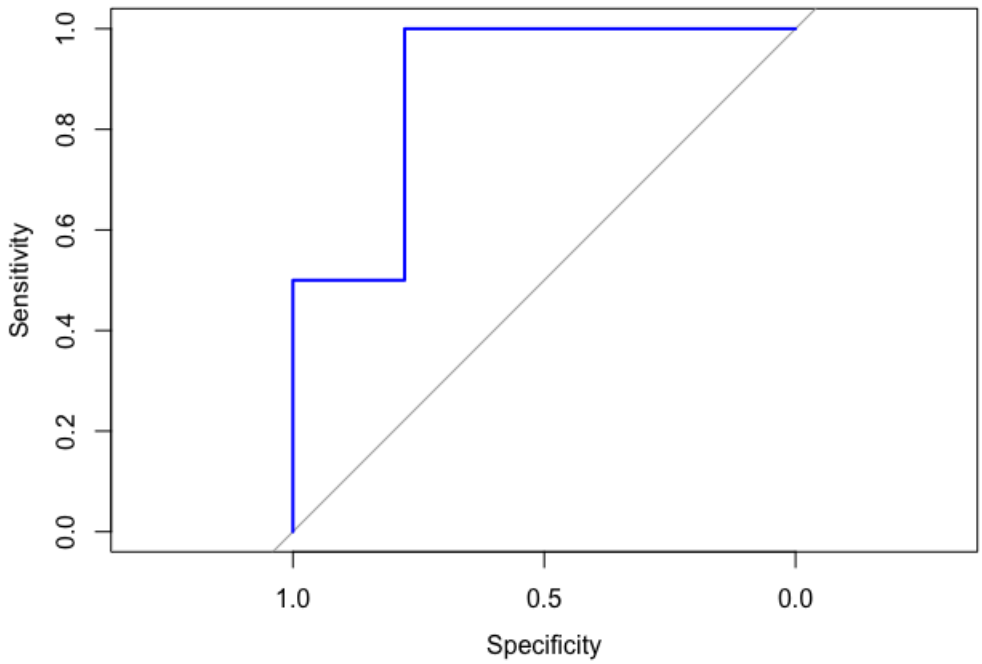

Figure 2: ROC curve

logfit_pred	0	1
0	17	1
1	1	1

Figure 3: Confusion matrix

Now LRT and Wald test are conducted to test whether X3 can be removed from the model. Both tests result in choosing the smaller model without X3.

Likelihood Ratio Test

Chi-Square	DF	Pr > ChiSq
0.4044	3	0.9393

Figure 4: LRT for X3

```
Wald test

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X1 + X2
  Res.Df Df      F Pr(>F)
1     75
2     76 -1 0.3986 0.5298
```

Figure 5: Wald test for X3

Results of stepwise model selection using backward method show that the best model is the model with only two predictors X1 and X2.

```
Step:  AIC=52.57
Y ~ X1 + X2

       Df Deviance    AIC
<none>      46.572 52.572
- X1    1   51.529 55.529
- X2    1   65.758 69.758

Call:  glm(formula = Y ~ X1 + X2, family = binomial, data = train_data)

Coefficients:
(Intercept)           X1           X2
    5.52227      0.09528     -0.22899

Degrees of Freedom: 78 Total (i.e. Null);  76 Residual
Null Deviance:      82.28
Residual Deviance: 46.57        AIC: 52.57
```

Figure 6: Model selection

Prediction

Now prediction is made to determine the probability that male client aged 55 with a health awareness index at the average health awareness score will receive a flu shot using the final 2-factor model. The probability is 0.054, which is much lower than the threshold, meaning that this person will unlikely receive a flu shot.