STAT 654 Problem Set 2

Nikko Cleri cleri@tamu.edu

March 30, 2020

Reading in our data from the 'Default' dataset from ISLR2

```
rm(list = ls())
library(ISLR2)
names(Default)
```

```
## [1] "default" "student" "balance" "income"
```

```
attach(Default)
```

Logistic Regression

The Logistic regression to the Default data set with response 'default' and predictors 'status', 'balance', and 'income' yields a model with the following details:

```
glm.fits <- glm(
  default ~ student + balance + income,
  data = Default, family = binomial
)
summary(glm.fits)
```

```
##
## Call:
## glm(formula = default ~ student + balance + income, family = binomial,
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

where the intercept, student status and balance predictors are all significant to (at least) the 0.001 level. The model predicts that students are less likely to default (negative coefficient), and higher balance means more likely to default (positive coefficient). The income predictor is insignificant with p=0.71152.

The following predicts the probability that the student defaults given the three predictors. We also show the contrasts, which interprets defaulting ('Yes') as 1 and not defaulting ('No') as 0.

```r
glm.probs <- predict(glm.fits, type = "response")
contrasts(default)
```

```
##     Yes
## No    0
## Yes   1
```

Here we see the performance of the logistic predictions.

```r
glm.pred <- rep("No", length(default))
glm.pred[glm.probs > .5] = "Yes"
table(glm.pred, default)
```

```
##         default
## glm.pred   No  Yes
##      No  9627  228
##      Yes   40  105
```

```r
mean(glm.pred == default)
```

```
## [1] 0.9732
```

The logistic model predicts greater than 97 percent of the student's default statuses correctly, but does a poor job at predicting those who did default, getting only 105 out of 333 'Yes' classifications correct.

Probit Regression

Following similarly, we have the probit regression

```r
glm.fits.probit <- glm(
  default ~ student + balance + income,  data = Default,
  family = binomial(link = "probit")
)
summary(glm.fits.probit)
```

```
##
## Call:
## glm(formula = default ~ student + balance + income, family = binomial(link = "probit"),
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2226  -0.1354  -0.0321  -0.0044   4.1254
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.475e+00  2.385e-01 -22.960   <2e-16 ***
## studentYes  -2.960e-01  1.188e-01  -2.491   0.0127 *
## balance      2.821e-03  1.139e-04  24.774   <2e-16 ***
## income       2.101e-06  4.121e-06   0.510   0.6101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1583.2  on 9996  degrees of freedom
## AIC: 1591.2
##
## Number of Fisher Scoring iterations: 8
```

which again shows that the intercept, student status and balance predictors are all significant to (at least) the 0.01 level. The model predicts that students are less likely to default (negative coefficient), and higher balance means more likely to default (positive coefficient). The income predictor is insignificant with p=0.6101.

The following shows the performance of the probit predictions where again we interpret defaulting ('Yes') as 1 and not defaulting ('No') as 0.

```
glm.probs.probit <- predict(glm.fits.probit, type = "response")
contrasts(default)
```

```
##     Yes
## No    0
## Yes   1
```

```
glm.pred.probit <- rep("No", length(default))
glm.pred.probit[glm.probs.probit > .5] = "Yes"
table(glm.pred.probit, default)
```

```
##                default
## glm.pred.probit   No  Yes
##             No  9639  238
##             Yes   28   95
```

```
mean(glm.pred.probit == default)
```

```
## [1] 0.9734
```

We again see that this probit prediction is greaater than 97 percent accurate, but similarly to the logistic regression, does a poor job of predicting the 'Yes' classifications correctly (only 98 out of 333).