Model Fitting

Logistic Regression I

Learning Objectives

1. Use `glm()` function in R to fit Logistic Regression model.
2. Interpret the summary output of `glm()` function for Logistic Regression.
3. Use Akaike Information Criterion (AIC) to compare models.

Multiple Logistic Regression Model

- Suppose, we want to fit a multiple Logistic Regression model for credit default dataset. Response variable is `Y=default` and predictor variables are X_1=`balance`, X_2=`income` and X_3=`student`.
- The logit of `default` is linear as shown here:

$$ln(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 \times balance + \beta_2 \times income + \beta_3 \times student \tag{1}$$

- The odds of `default` is:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 \times balance+\beta_2 \times income+\beta_3 \times student} \tag{2}$$

- The logistic model (S-shaped curve) or the probability of `default = Yes` for a given X, i.e. $p(X) = Pr(Y = 1|X)$ is:

$$p(X) = \frac{e^{\beta_0+\beta_1 \times balance+\beta_2 \times income+\beta_3 \times student}}{1 + e^{\beta_0+\beta_1 \times balance+\beta_2 \times income+\beta_3 \times student}} \tag{3}$$

Fit Logistic Regression Model

- Use a 80-20 split, to get `train` and `test` data. Apply `glm()` function with the `train` data.
- As our response variable, `default` is categorical with a binary yes/no value, the `family` argument is set to `binomial`.

```
model1 = glm(default ~ balance + income + student,
             data = train,
             family = binomial)
summary(model1)
```

glm() Summary

```
Call:
glm(formula = default ~ balance + income + student, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5257  -0.2620  -0.0640   0.1927   3.2673

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.878e+00  2.725e-01 -36.251  < 2e-16 ***
balance      6.630e-03  1.414e-04  46.870  < 2e-16 ***
income       4.767e-06  4.309e-06   1.106    0.269
studentYes  -6.286e-01  1.232e-01  -5.103 3.35e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12604.9  on 10663  degrees of freedom
Residual deviance:  4855.8  on 10660  degrees of freedom
AIC: 4863.8

Number of Fisher Scoring iterations: 7
```

1

2

3

`glm()` Summary: Estimates of Coefficients

1

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.878e+00  2.725e-01 -36.251  < 2e-16 ***
balance      6.630e-03  1.414e-04  46.870  < 2e-16 ***
income       4.767e-06  4.309e-06   1.106    0.269
studentYes  -6.286e-01  1.232e-01  -5.103 3.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Model equation is:

$$ln(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 \times balance + \beta_2 \times income + \beta_3 \times student \tag{4}$$

- Substituting the values from the `glm()` summary:

$$ln(\frac{p(X)}{1-p(X)}) = -9.878 + 0.00663 \times balance + 4.767 \times 10^{-6} \times income - 0.6286 \times student \tag{5}$$

glm() Summary: Significance of Coefficients

Using t-test

2

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -9.878e+00  2.725e-01 -36.251  < 2e-16 ***
balance      6.630e-03  1.414e-04  46.870  < 2e-16 ***
income       4.767e-06  4.309e-06   1.106    0.269    
studentYes  -6.286e-01  1.232e-01  -5.103 3.35e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Significance test for each coefficient β_j assumes a null hypothesis H_0: $\beta_j = 0$, and alternate hypothesis H_1: $\beta_j \neq 0$, where $j = 1, 2, 3$.
- β_0, β_1 and β_3 are significant, as their p-value is less than 0.05. This means `balance` (β_1) and `student` (β_3) are significant.
- β_2 is not significant, as its p-value is greater than 0.05. This means `income` is not significant.

Multicollinearity

Variance Inflation Factor

- `income` and `student` are correlated with correlation coefficient $= -0.76$
- Check for multicollinearity of each predictor variable compared to others, using the Variance Inflation Factor, VIF.
- VIF of 1 is ideal and represents no multicollinearity. VIF of 5 or more implies multicollinearity.

```
library(car)
vif(model1)
```

```
balance   income  student
1.058775 2.495272 2.535116
```

- There is no multicollinearity in the predictor variables in this data, as all VIF values are below 5.

Model selection

AIC: Akaike Information Criterion

- Variance/bias tradeoff: simplest model (low variance) with best fit (low bias)
- Akaike Information Criterion or AIC, judges a model by how close its fitted values are to the observed values in the data (i.e. bias), while also penalising more complex models (i.e. variance).
- AIC is given by:

$$AIC = -2loglik(\hat{\beta}) + 2k \tag{6}$$

  Where $\hat{\beta}$ are the coefficient estimates of the model, k is the number of parameters in the model and *loglik* represents the maximum log likelihood of the model.
- Optimal model has lowest AIC.

```
AIC(model1)
```

```
[1] 4863.814
```

Model selection

AIC

- When comparing models with different number of parameters, k, select the one with lowest AIC.

	formula	k	aic
1	balance + income + student	3	4863.814
2	balance + student	2	4863.038
3	income + student	2	12532.277
4	income + balance	2	4888.153
5	balance	1	4948.994
6	student	1	12534.186
7	income	1	12579.773

Model Selection

```
summary(model2)
```

```
Call:
glm(formula = default ~ balance + student, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5109  -0.2632  -0.0642   0.1928   3.2757

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.6836497  0.2062891 -46.942   <2e-16 ***
balance      0.0066269  0.0001413  46.890   <2e-16 ***
studentYes  -0.7328038  0.0795607  -9.211   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12605  on 10663  degrees of freedom
Residual deviance:  4857  on 10661  degrees of freedom
AIC: 4863

Number of Fisher Scoring iterations: 7
```

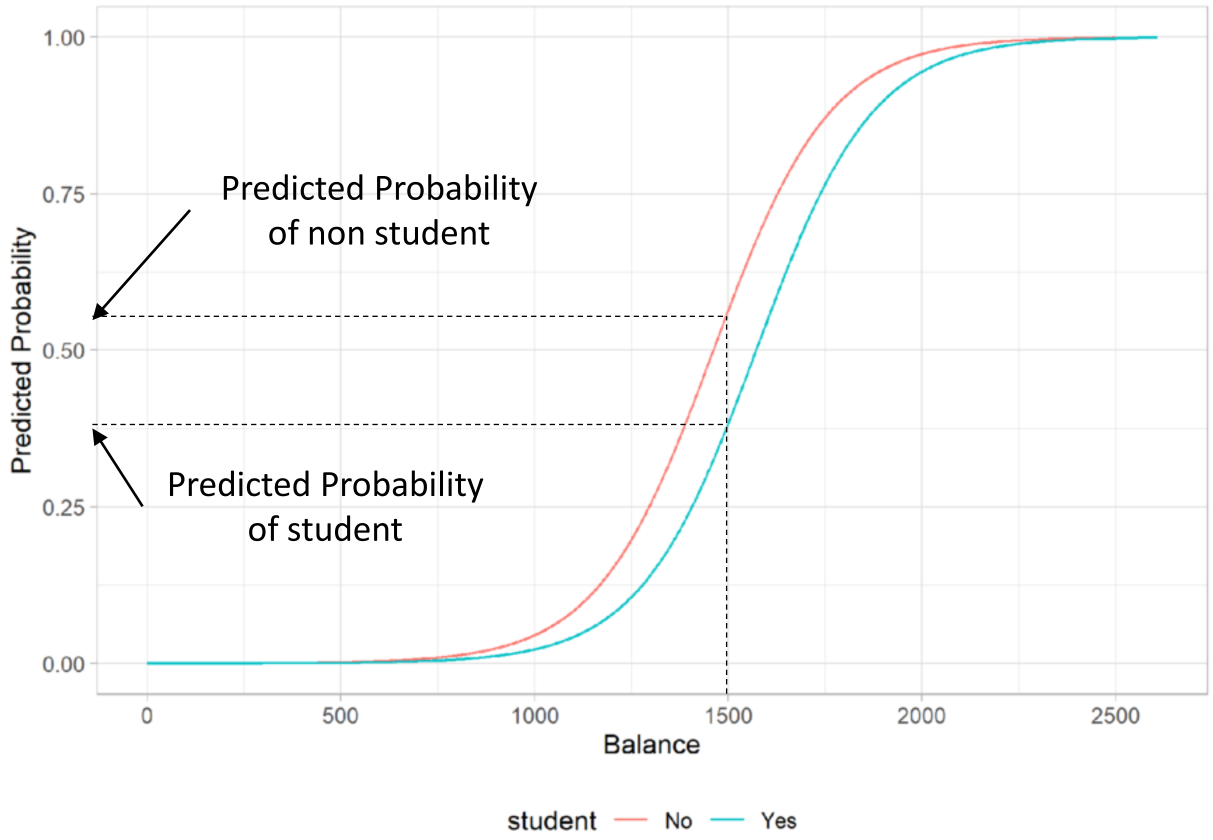
Interpret Model Coefficients

- Here's our model equation:

$$ln(\frac{p(X)}{1-p(X)}) = -9.684 + 0.00663 \times balance - 0.7328 \times student \tag{7}$$

- For each unit increase in `balance`, holding other predictors *fixed*, on average:
  - Log odds or logit of `default` changes by 0.00663.
  - Odds are multiplied by $e^{0.00663} = 1.0067$, or an increase of 0.67% in odds of default.
- For `student = Yes`, holding other predictors *fixed*, on average:
  - Log odds or logit of `default` changes by -0.7328.
  - Odds are multiplied by $e^{-0.7328} = 0.481$, or a decrease of 52% in odds of default.

Visualise Model



- Given a balance value, customers who are students are less likely to default, and therefore less risky.

Visualise Model

Code

```r
probs <- fitted(model2)

pp <- train %>% mutate(pred_prob = probs)

ggplot() +
        geom_line(data = pp,
                  aes(x=balance, y=pred_prob, color=student))+
        scale_x_continuous(breaks = seq(0, 3500, 500)) +
        labs(y="Predicted Probability", x="Balance") +
        theme_light() +
        theme(legend.position = "bottom")
```

References I

Agresti, A. (2018).
An introduction to categorical data analysis.
John Wiley & Sons.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
An Introduction to Statistical Learning: with Applications in R.
Springer.