Classification I: Introduction and GLMs

-Applied Multivariate Analysis-

Lecturer: Darren Homrighausen, PhD

What is Classification?

All the previous regression material presumes that the response Y is quantitative

(Drug susceptibility, area burned by forest fires, etc)

However, in many cases, the responses are qualitative

(Digit recognition, cancer status, etc)

An Overview of Classification

Some examples:

- A person arrives at an emergency room with a set of symptoms that could be 1 of 3 possible conditions. Which one is it?
- A online banking service must be able to determine whether each transaction is fraudulent or not, using a customer's location, past transaction history, etc.
- Given a set of individuals sequenced DNA, can we determine whether various mutations are associated with different phenotypes?

All of these problems are not regression problems. They are classification problems.

The Set-up

It begins just like regression: suppose we have observations

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Again, we want to estimate a function that maps X into Y that helps us predict as yet observed data.

(This function is known as a classifier)

The same constraints apply:

- We want a classifier that predicts test data, not just the training data.
- Often, this comes with the introduction of some bias to get lower variance and better predictions.

How do we measure quality?

In regression, we have $Y_i \in \mathbb{R}$ and (generally) use squared error loss

However, when Y_i only takes a few possible values, we will use zero-one prediction risk instead:

$$\text{pred} = \mathbb{E}[\mathbf{1}(Y_0 \neq \hat{Y})] = \mathbb{P}(\hat{Y} \neq Y_0),$$

where

$$\mathbf{1}(A) = \begin{cases} 1 & \text{if statement } A \text{ is true} \\ 0 & \text{if statement } A \text{ is true} \end{cases}$$

is an indicator function

We want to label or classify a new observation (X_0, Y_0) such that $\hat{Y} = Y$ as often as possible.

(Implicitly, $\hat{Y}$ is a function of X_0 and the sample)

Best classifier

When we did regression, we measured quality by squared error loss:

$$\text{pred} = \mathbb{E}(Y - \hat{Y})^2$$

(This is still an unknown quantity!)

It turns out, the closest procedure we can find, is:

$$\mathbb{E}\,Y|X = \underset{\hat{Y}}{\text{argmin}}\, \mathbb{E}(Y - \hat{Y})^2$$

known as the regression function

Best classifier

We can define the same quantity for classification

$$\underset{\hat{Y}}{\operatorname{argmin}}\, \mathbb{E}[\mathbf{1}(Y_0 \neq \hat{Y})] = \underset{\hat{Y}}{\operatorname{argmin}}\, \mathbb{P}(\hat{Y} \neq Y_0),$$

In classification this is known as the Bayes' rule

In analogy to the regression function, the Bayes' rule looks like:

$$0 \text{ if } \mathbb{P}(Y = 0|X) \geq \mathbb{P}(Y = 1|X)$$

or

$$1 \text{ if } \mathbb{P}(Y = 1|X) \geq \mathbb{P}(Y = 0|X)$$

(That is, we want to maximize the conditional probability)

Emphasis: The Bayes' rule, like the regression function, is unknown/unknowable

We can try to estimate them, however.

Introductory example

An Introductory Example

Suppose we work for a credit card company and we wish to identify people that are likely to default on their credit card debt

We have predictors (for 10,000 people):

- Student status
- Income
- Balance

Along with their default status:

$$Y = \begin{cases} 1 \text{ if person defaults} \\ 0 \text{ if person doesn't default} \end{cases}$$

Let's look at some plots.

An Introductory Example

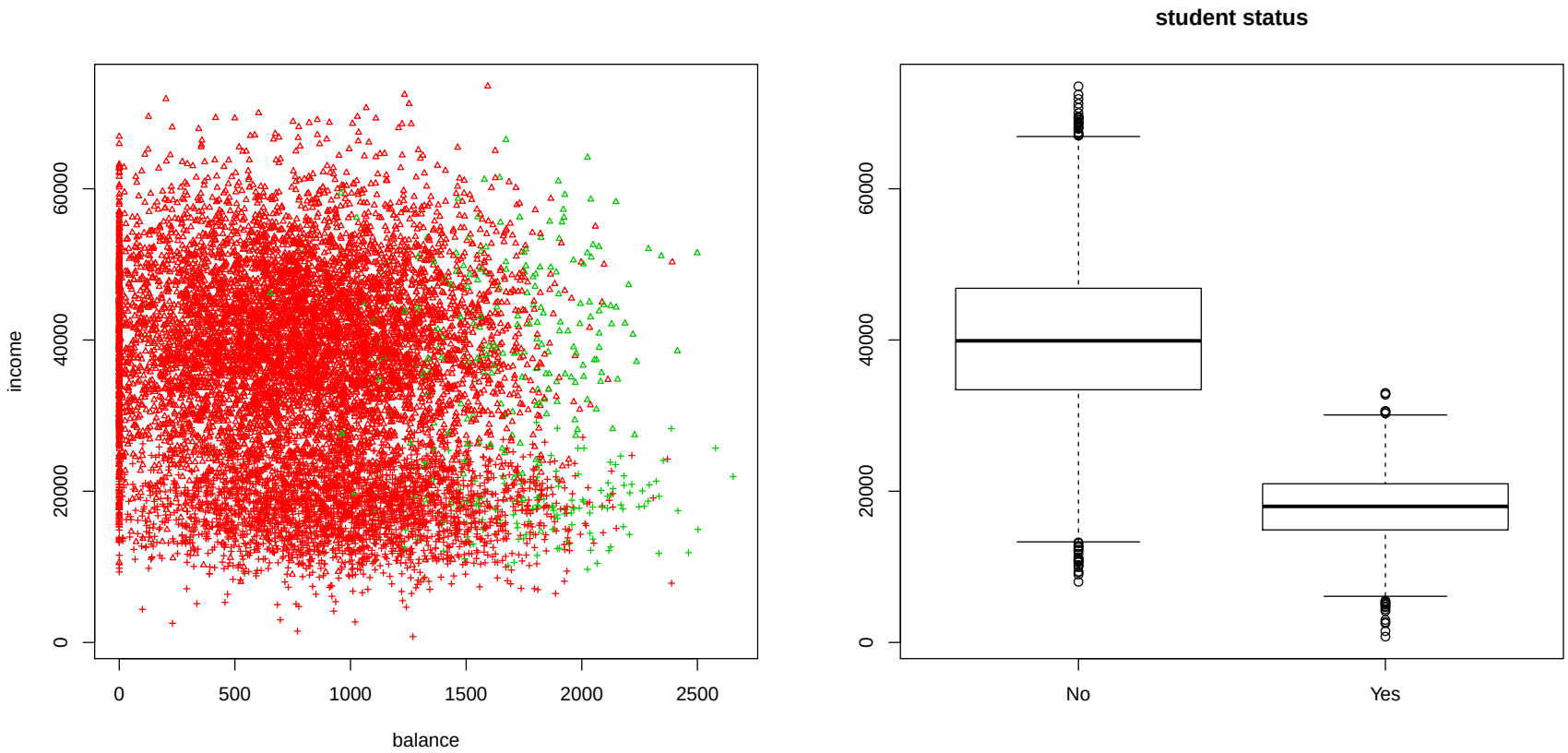
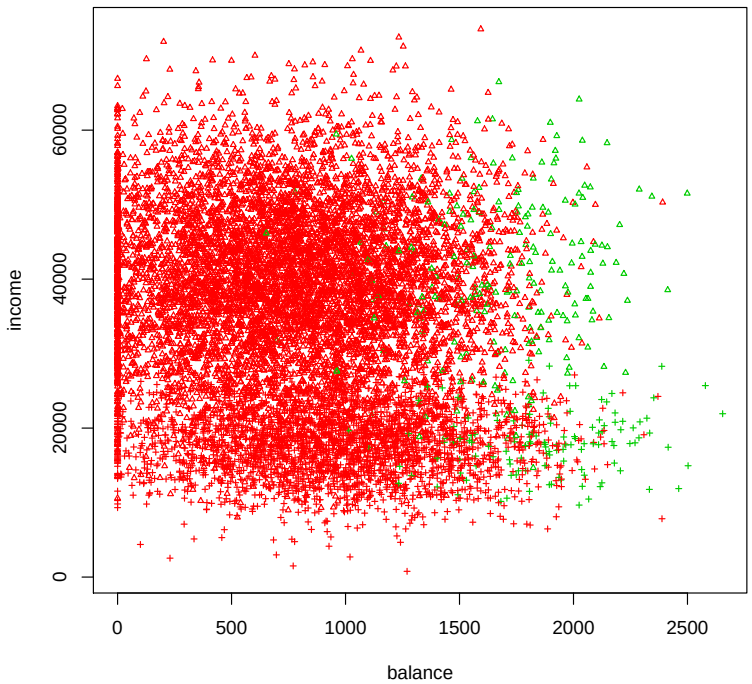
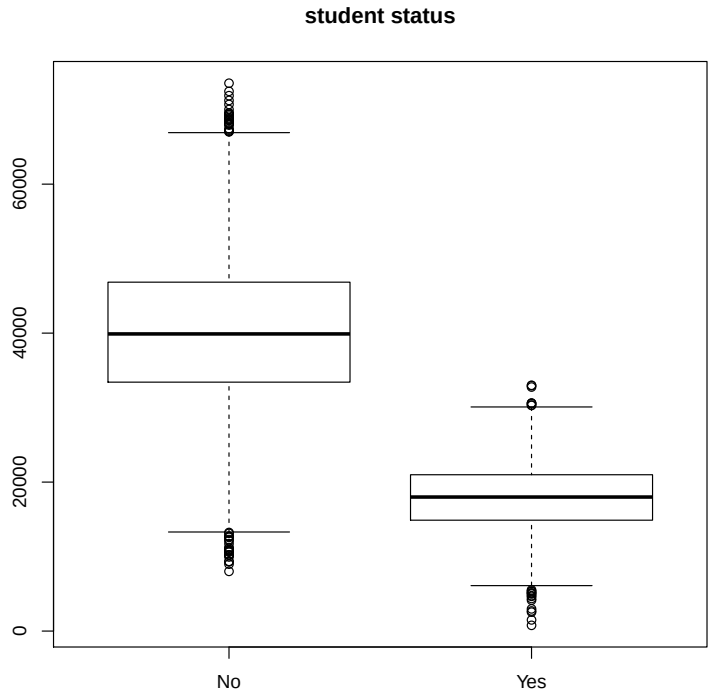


Figure: The red are people without defaults, green are defaults. The '+' are students, the 'Δ' are not students.

An Introductory Example





Some comments:

- Income doesn't seem to be related to defaults
- Student status is also unrelated to defaults, but highly related to income
- Balance seems to strongly predict default status.

An Introductory Example: Why not Use Regression?

Suppose for a moment we only consider balance. Then, we can run a simple linear regression of default status on balance

```
Y = rep(0,n)
Y[default == 'Yes'] = 1
out.lm = lm(Y~balance)
summary(out.lm)
```

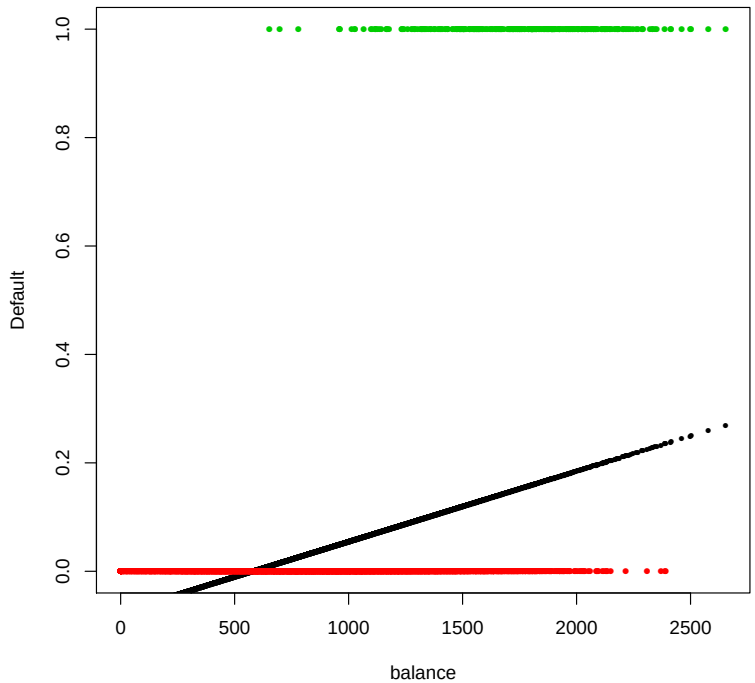
R will happily do this.

An Introductory Example: Why not Use Regression?

```
> summary(out.lm)
Call:
lm(formula = Y ~ balance)
Residuals:
     Min       1Q   Median       3Q      Max
-0.23533 -0.06939 -0.02628  0.02004  0.99046
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.519e-02  3.354e-03  -22.42   <2e-16 ***
balance      1.299e-04  3.475e-06   37.37   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.1681 on 9998 degrees of freedom
Multiple R-squared: 0.1226, Adjusted R-squared: 0.1225
F-statistic:  1397 on 1 and 9998 DF,  p-value: < 2.2e-16
```

An Introductory Example: Why not Use Regression?

Let's plot our data with estimated regression function:



Not so great..

An Introductory Example: Generalized Linear Models (GLMs)

GLMs differ from ordinary regression by modeling the *probabilities* as opposed to the outcomes themselves. To wit:

Regression:

$$Y_i = X_i^\top \beta + \epsilon_i$$

Logistic regression (with logit link): Let $\pi(X_i) = Pr(Y_i = 1|X_i)$,

$$\log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = X_i^\top \beta$$

This is known as the logistic function.

It is differentiable, maps [0,1] to $\mathbb{R}$, and is invertible. Its inverse is:

$$\pi(X_i) = \frac{\exp\{X_i^\top \beta\}}{1+\exp\{X_i^\top \beta\}}.$$

An Introductory Example: Generalized Linear Models (GLMs)

Let's look at each of these terms

- $\pi(X_i) = Pr(Y_i = 1|X_i)$ is the probability Y is equal to 1 at a given level of $X = X_i$
- $$\frac{\pi(X_i)}{1 - \pi(X_i)}$$

  is known as the odds that Y is equal to 1.
- $$\log\left(\frac{\pi(X_i)}{1 - \pi(X_i)}\right)$$

  is the log odds.

Effectively, we are assuming that the log odds of $Y = 1$ (which is always called a success) is linear in the predictors X

An Introductory Example: Generalized Linear Models (GLMs)

With regression, there was a closed form solution for an estimate of β:

$$\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top Y.$$

This was due to estimation via least squares

(This is also the maximum likelihood estimator (MLE) under Gaussian errors)

For GLMs, the likelihood is different, but we still use the MLE

There isn't any closed form solution and all solution methods are iterative maximizers.

Of course, this make no real difference, since we are going to use someone else's code.

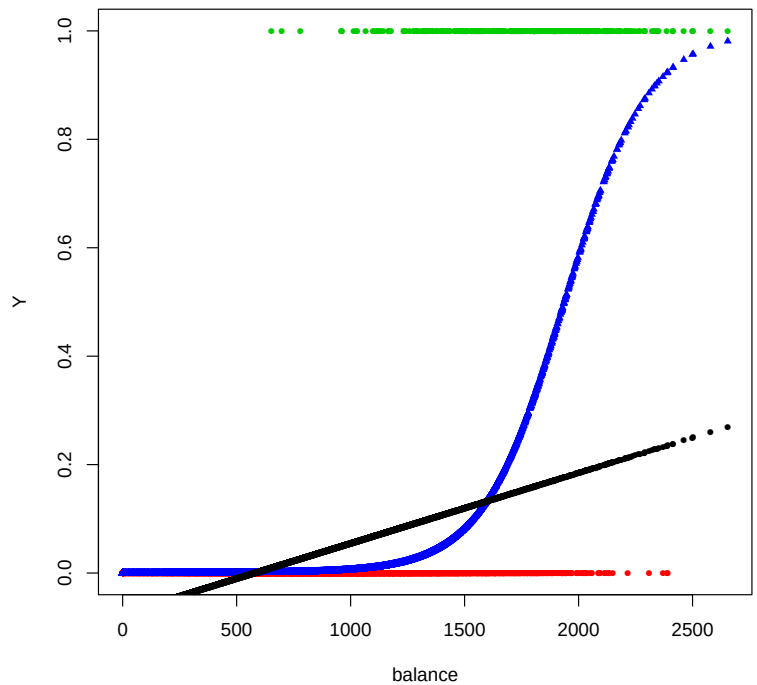
An Introductory Example: Generalized Linear Models (GLMs)

```
out.glm  = glm(default~balance,family='binomial')
> summary(out.glm)
Call:
glm(formula = default ~ balance, family = "binomial")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5
Number of Fisher Scoring iterations: 8
```

An Introductory Example: Compare GLM to Regression.

Let's plot our data with

- Simple Linear Regression (black)
- GLM (blue)



An Introductory Example: Making Predictions

Once we get $\hat{\beta}$, making predictions is a simple matter.

Suppose we want to estimate the probability that someone with a balance of \$1,000 will default. We form:

$$\hat{\pi}(1,000) = \frac{\exp\{-10.65 + 0.0055 * 1000\}}{1 + \exp\{-10.65 + 0.0055 * 1000\}} = 0.00576.$$

Pretty small..

Maybe look at \$2,000 instead:

$$\hat{\pi}(2,000) = \frac{\exp\{-10.65 + 0.0055 * 2000\}}{1 + \exp\{-10.65 + 0.0055 * 2000\}} = 0.586.$$

Much larger..

An Introductory Example: Classification

But, Darren, I thought we were classifying!

To form a classifier out of these predictions round the probabilities!

$$\hat{\pi}(1,000) = \frac{\exp\{-10.65 + 0.0055 * 1000\}}{1 + \exp\{-10.65 + 0.0055 * 1000\}} = 0.00576.$$

Thus, a balance of \$1,000 would be classified as no default

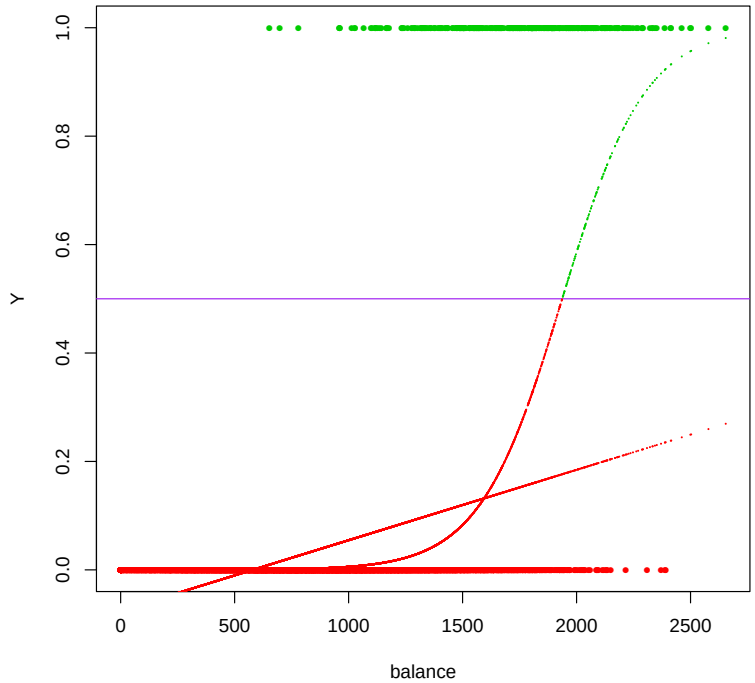
Maybe look at \$2,000 instead:

$$\hat{\pi}(2,000) = \frac{\exp\{-10.65 + 0.0055 * 2000\}}{1 + \exp\{-10.65 + 0.0055 * 2000\}} = 0.586.$$

A balance of \$2,000 would be classified as default

An Introductory Example: Compare GLM to Regression.

Results of using a cut-off of 0.5



Sensitivity and Specificity

We need two concepts:

Sensitivity: Classifying a person as a 'default' given that they defaulted.

(This is like correctly rejecting the null hypothesis, i.e. power)

Specificity: Classifying a person as 'no default' given that they did not default

(this is like **not** committing a type I error i.e. 1- $\mathbb{P}$(type I error))

As suggested by the notation, it is easiest to think of this in terms of hypothesis testing

Sensitivity and Specificity

	Training Error	Training Sensitivity	Training Specificity
Linear Reg.	0.033	0.000	1.000
GLM	0.027	0.300	0.995

where

- Training Error: $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_i \neq \hat{Y}_i)$
- Training Sensitivity: $\frac{1}{\#\text{ default}}\sum_{i\in\text{default}}\mathbf{1}(\hat{Y}_i = \text{default})$
- Training Specificity:

$$\frac{1}{\#\text{ no default}}\sum_{i\in\text{no default}}\mathbf{1}(\hat{Y}_i = \text{no default})$$

Discussion on confounding

GLMs and Multiple Logistic Regression: An Apparent Paradox

Let's look at just including student as a predictor:

```
> out.glm.student  = glm(default~student,
family='binomial')
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

Versus including income and balance as well.

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
income       3.033e-06  8.203e-06   0.370  0.71152
```

GLMs and Multiple Logistic Regression: An Apparent Paradox

Let's look at just including student as a predictor:

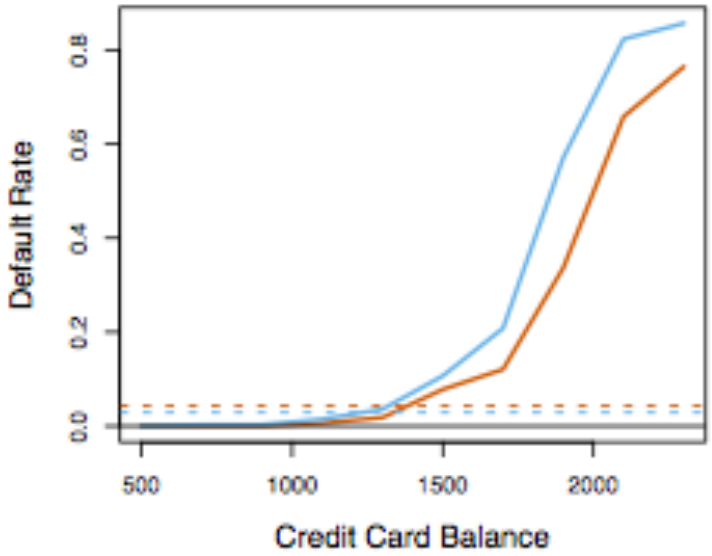
```
> out.glm.student  = glm(default~student,
family='binomial')
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

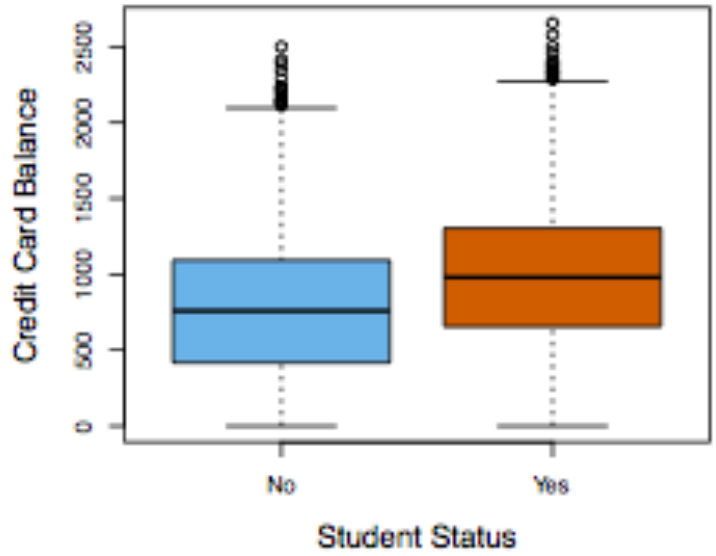
Versus including income and balance as well.

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
income       3.033e-06  8.203e-06   0.370  0.71152
```

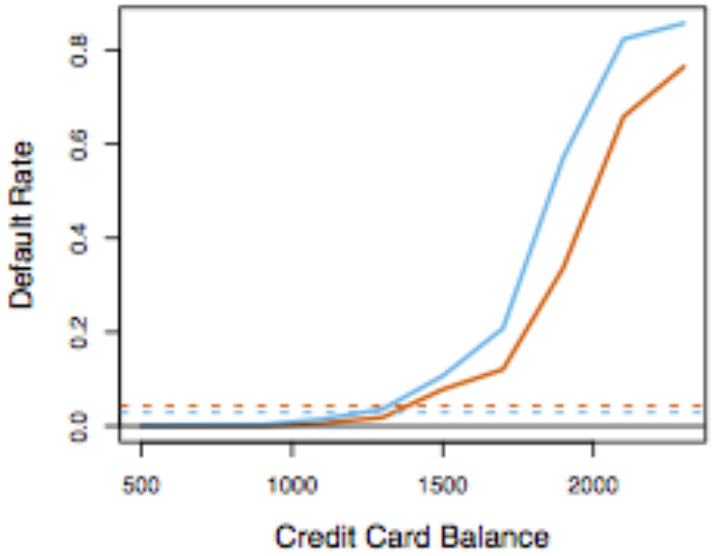
The sign on student is different!

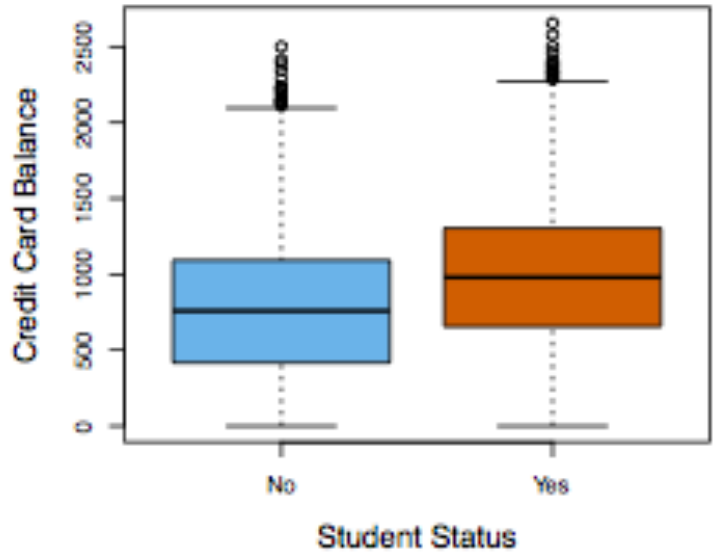
GLMs and Multiple Logistic Regression: An Apparent Paradox





GLMs and Multiple Logistic Regression: An Apparent Paradox





Some observations:

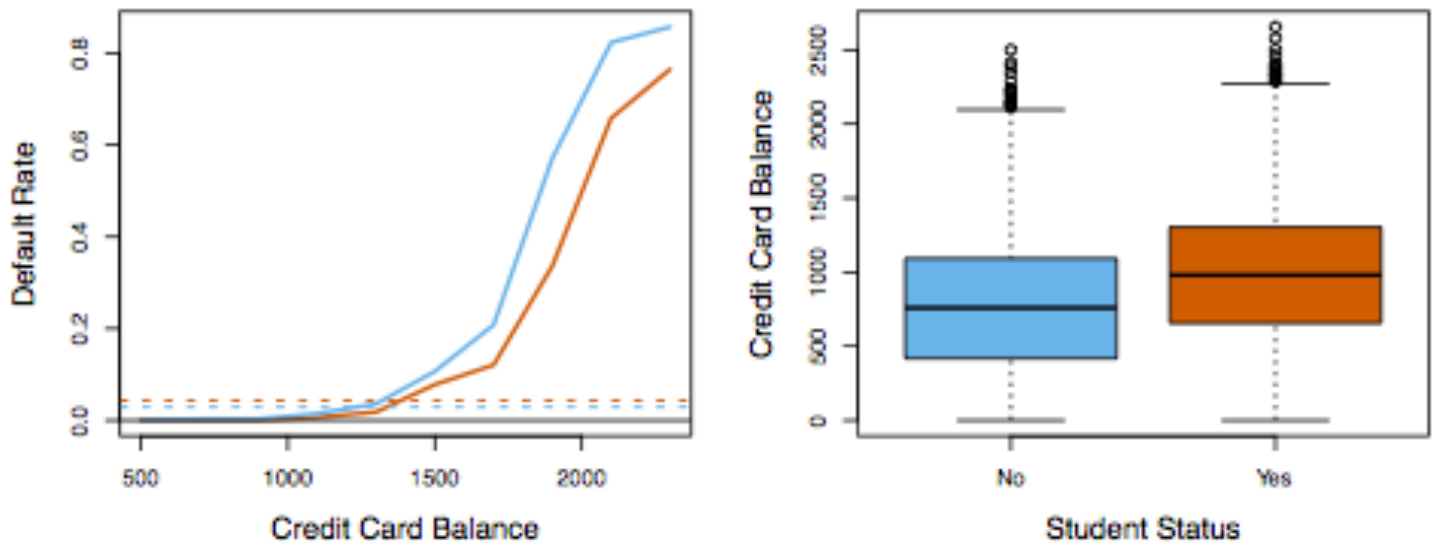
- Students have slightly higher balances
- Non-students have slightly higher default rate (for a given balance)
- Students have a slightly higher default rate.

How Could this Be?

The answer is called confounding

For a fixed value of income and balance, students are less risky (negative coefficient estimate) while overall, students are riskier (positive coefficient).

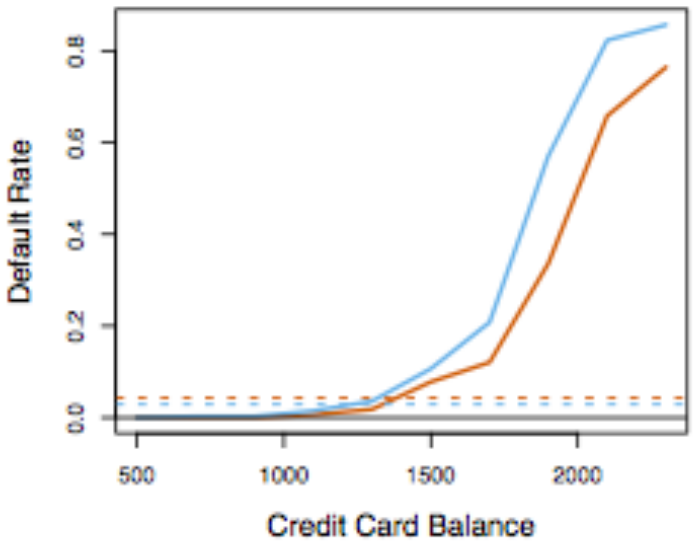
The boxplot tells the tale. Students have more debt, which is associated with more defaults.

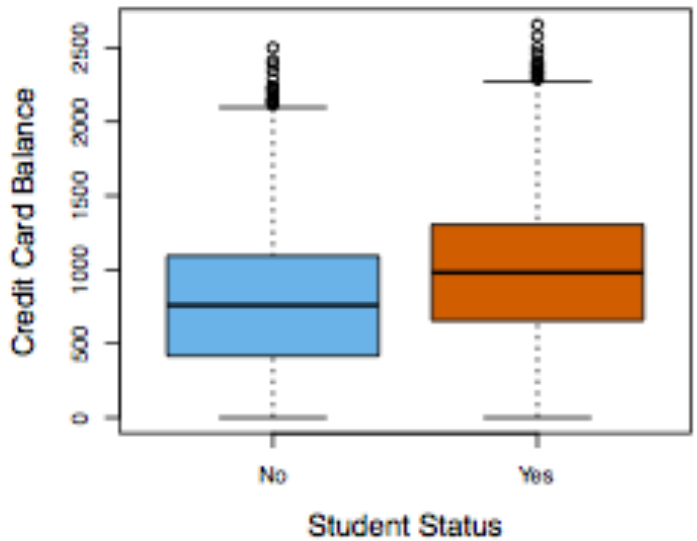


What Does this Mean?

If you are a credit card company, then:

- If all you know is that they are a student, then you should be wary.
- However, if you have two candidates with the same balance, then the student is less risky!





More than two levels to the response

You can use logistic regression when your response has more than two levels. There are two cases:

Unordered response: Called multinomial logistic regression. Essentially, you fit logistic regressions for each level versus a reference level (examples: eye color or political preference)

Ordered Response: These are common slopes or proportional odds model (examples: how strongly do you agree with a statement or number of malformed limbs in an experiment with mice)

For More Information

See the files:

- glmLectureHandwritten.pdf (introduction to GLMs)
- glmMultilevelResponsehandwritten.pdf (Overview of multilevel GLMs)

Both are on the website

However, we won't discuss these in depth in this class as there are other options that are more natural.