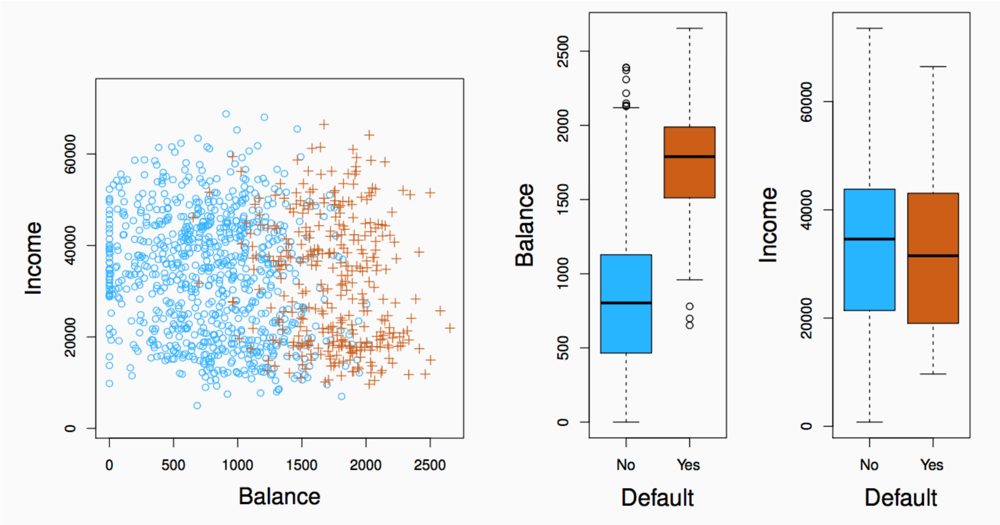


Figure 1: ‘Default‘ data.

Contents

1 Logistic regression

1.1 Example

Example: Default dataset

Data (Fig. 1) consist of annual income and monthly credit card balance for a subset of 10,000 individuals for whom we know who defaulted on their credit card payments.

We are interested in **predicting whether an individual will default** on his or her credit card payment, on the basis of annual income and monthly credit card balance.

Data can be loaded from the ISLR library of the companion textbook.

```
#>  default     student       balance           income     
#>  No :9667   No :7056   Min.   :   0.0   Min.   :  772  
#>  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340  
#>                        Median : 823.6   Median :34553  
#>                        Mean   : 835.4   Mean   :33517  
#>                        3rd Qu.:1166.3   3rd Qu.:43808  
#>                        Max.   :2654.3   Max.   :73554  
```

`Default` dataset: The annual incomes and monthly credit card balances of a number of individuals.

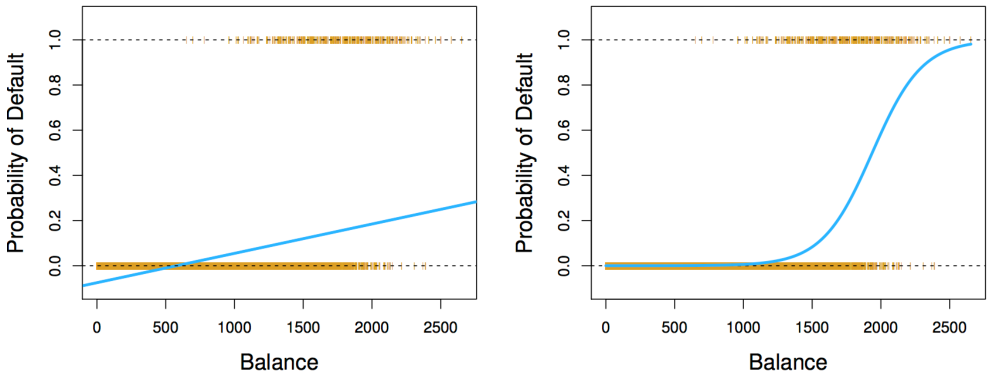


Figure 2: ‘Default‘ data: Linear vs Logistic regression on ‘Balance‘ covariate.

Can we use Linear Regression?

Suppose for the Default classification task that we code

$$Y = \begin{cases} 0 & \text{No} \\ 1 & \text{Yes} \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as **Yes** if $\hat{Y} > 0.5$?

- For a binary response with a 0/1 coding, regression by least squares does make sense; since in the population $E(Y|X = x) = Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.

Let’s write $p(X) = Pr(Y = 1|X)$ for short

$$p(X) = \beta_0 + \beta_1 X$$

- However, *linear* regression **might produce probabilities less than zero or bigger than one**.
- *Logistic regression* is more appropriate.
- It models $p(X)$ using a function, the **logistic function** that gives outputs between 0 and 1 for all values of X.
- *Curiously, it turns out that the classifications that we get if we use linear regression to predict a binary response will be the same as for the linear discriminant analysis (LDA) procedure we discuss later.*

Linear versus Logistic Regression

The orange marks indicate the response Y, either 0 or 1 (Fig. 2).

Linear regression does not estimate $Pr(Y = 1|X)$ well (some estimated probabilities are negative). Logistic regression seems well suited to the task (all probs are between 0 and 1).

Linear Regression for three class response

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

This coding suggests an ordering, and in fact implies that the difference between stroke and drug overdose is the same as between drug overdose and epileptic seizure. But if that is not appropriate? In general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression.

Linear regression is not appropriate here. *Multiclass Logistic Regression* or *Discriminant Analysis* are more appropriate.

1.2 The logistic model

1.2.1 Logistic Regression

Let's write $p(X) = Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

that is the **logistic function** ($e \approx 2.71828$ is a mathematical constant [Euler's number.])

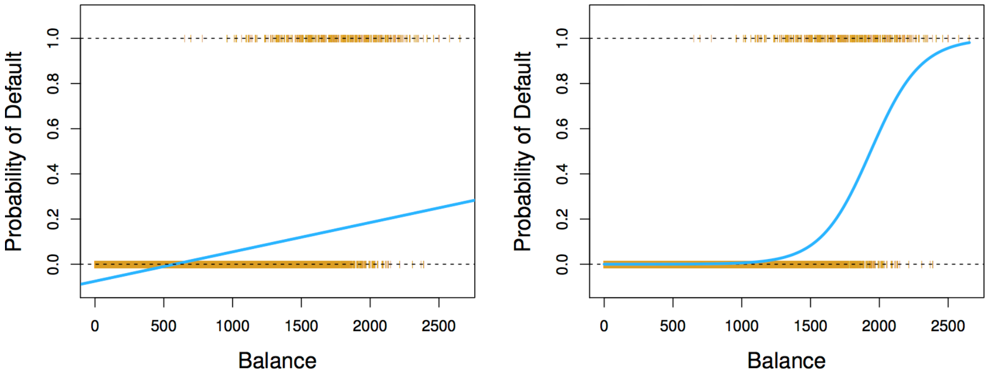
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

A bit of rearrangement gives

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X\ .$$

This monotone transformation is called the **log odds** or **logit** transformation of $p(X)$.

Linear versus Logistic Regression



Logistic regression ensure that our estimate for $p(X)$ lies between 0 and 1.

1.2.1.1 Interpretation

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \qquad \Leftarrow \quad \text{log-odds} \qquad (-\infty, \infty)$$

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \qquad \Leftarrow \quad \text{odds} \qquad (0, \infty)$$

$$p(X) = e^{\beta_0+\beta_1 X}(1-p(X))$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \qquad \Leftarrow \quad \text{logistic function} \qquad (0, 1)$$

Interpretation of β_1

- a one-unit increase of X changes the log odds by β_1
- a one-unit increase of X multiplies the odds by e^{β_1}
- a one-unit increase of X changes $p(X)$ according to the sign of β_1 and depending on the value of X (the relationship is not linear, is an “S”)

1.2.1.2 Generalized Linear Models (GLM)

GLM generalize (Normal) linear models.

Logistic regression is a GLM,

- a Binomial distribution for Y
  $$Y \sim \text{Binomial}(p = p(X), n = 1)$$
  $p(X)$ is the expected value of Y
- linear predictor for a function of the expected value of Y
  $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Function $g(\cdot)$ that applies to the expected value for Y is called the **link function**.

In logistic regression, the link function is the **logit**.

1.2.1.3 Maximum Likelihood (ML)

We use **ML** to estimate the parameters,

$$lik(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1-p(x_i))$$

This likelihood gives the probability of the observed zeros and ones in the data.

We pick β_0 and β_1 to maximize the likelihood of the observed data.

Most statistical packages can fit linear logistic regression models by maximum likelihood. In `R` we use the **`glm`** function.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

1.2.1.4 Making Predictions

What is our estimated probability of default for someone with a balance of $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

With a balance of $2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Then, for any given value of balance, a prediction can be made for default.

For example, one might predict `default` = `Yes` for any individual for whom $p(\text{balance}) > 0.5$.

Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as $p(\text{balance}) > 0.1$.

A categorical predictor

Let's do it again, using `student` as the predictor

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student	0.4049	0.1150	3.52	0.0004

$$\hat{p}(\texttt{student} = \texttt{Yes}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\times 1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\times 1}} = \frac{e^{-3.5041+0.4049}}{1+e^{-3.5041+0.4049}} = 0.0431$$

$$\hat{p}(\texttt{student} = \texttt{No}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\times 0}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\times 0}} = \frac{e^{-3.5041}}{1+e^{-3.5041}} = 0.0292$$

1.3 Multiple Logistic regression

Logistic regression with several variables

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \,.$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

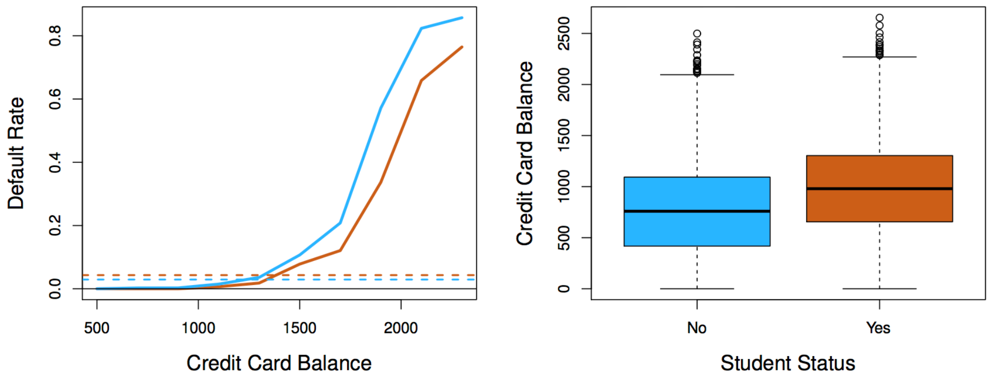


Figure 3: ‘Default‘ data: Logistic regression on ‘Student‘ covariate.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student	-0.6468	0.2362	-2.74	0.0062

Why is coefficient for `student` negative, while it was positive before?

Remember the concept of **confounding**.

- (right) Students tend to have higher balances than non-students,
- so their marginal default rate is higher than for non-students (left: horizontal broken lines display the overall default rates for students and non-students).
- (left) But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.

Prediction

For example, a student with a credit card balance of $1,500 and an income of $40, 000 has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\times 1500+\hat{\beta}_2\times 40+\hat{\beta}_3\times 1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\times 1500+\hat{\beta}_2\times 40+\hat{\beta}_3\times 1}} = 0.058$$

A non-student with the same balance and income has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\times 1500+\hat{\beta}_2\times 40+\hat{\beta}_3\times 0}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\times 1500+\hat{\beta}_2\times 40+\hat{\beta}_3\times 0}} = 0.105$$

```
#>               Estimate Std. Error  z value   Pr(>|z|)
#> (Intercept) -1.087e+01  4.923e-01 -22.0801 4.911e-108
#> studentYes  -6.468e-01  2.363e-01  -2.7376  6.188e-03
#> balance      5.737e-03  2.319e-04  24.7376 4.220e-135
#> income       3.033e-06  8.203e-06   0.3698  7.115e-01
#> AIC: 1579.54
#>                      Estimate Std. Error  z value  Pr(>|z|)
#> (Intercept)        -1.099e+01  5.667e-01 -19.3986 7.926e-84
#> studentYes         -2.856e-01  8.239e-01  -0.3466 7.289e-01
#> balance             5.817e-03  2.938e-04  19.8005 2.947e-87
#> income              3.016e-06  8.226e-06   0.3667 7.139e-01
#> studentYes:balance -2.184e-04  4.781e-04  -0.4568 6.478e-01
#> AIC: 1581.34
```

1.4 Logistic regression with more than two classes

So far we have discussed logistic regression with two classes.

It is easily generalized to more than two classes.

One version (used in the R package `glmnet`) has the symmetric form

$$Pr(Y=k|X)=\frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{l=1}^{K}e^{\beta_{0l}+\beta_{1l}X_1+\ldots+\beta_{pl}X_p}}$$

Here there is a linear function for *each* class.

Multiclass logistic regression is also referred to as *multinomial regression.*

In practice they tend not to be used all that often. One of the reasons is that *discriminant analysis* is popular for multiple-class classification.