Data Analytics for Social Science
Logistic regression

Johan A. Elkink

School of Politics & International Relations
University College Dublin

7 March 2017

Outline

Levels of measurement

	Discreet	Continuous
Nominal	party choice	-
Ordinal		-
		-
Interval		
Ratio		

Categories in no particular order

(examples in cells)

Levels of measurement

	Discreet	Continuous
Nominal	party choice	-
Ordinal	education level	-
		-
Interval		
Ratio		

Categories in a specific order

(examples in cells)

Levels of measurement

	Discreet	Continuous
Nominal	party choice	-
Ordinal	education level	-
		-
Interval	how likely to vote ...	temperature
Ratio		

All values possible

(examples in cells)

Levels of measurement

	Discreet	Continuous
Nominal	party choice	-
Ordinal	education level	-
		-
Interval	how likely to vote ...	temperature
Ratio	deaths in war	ideological distance

All values possible, with a meaningful zero point

(examples in cells)

Levels of measurement

	Discreet	Continuous
Nominal	party choice	-
Ordinal	education level	-
Binary	turnout	-
Interval	how likely to vote ...	temperature
Ratio	deaths in war	ideological distance

Two categories, coded as 0 and 1

(examples in cells)

Binary models

Binary models have a dependent variable consisting of two categories.

For example,

- Vote on a particular law
- Turning out in an election
- Approval in a referendum
- Bankrupt or not

Classification

With linear regression, we try to:

1. estimate the relationship between two variables, and
2. predict the mean of y_i based on the values in $\mathbf{x}_i$.

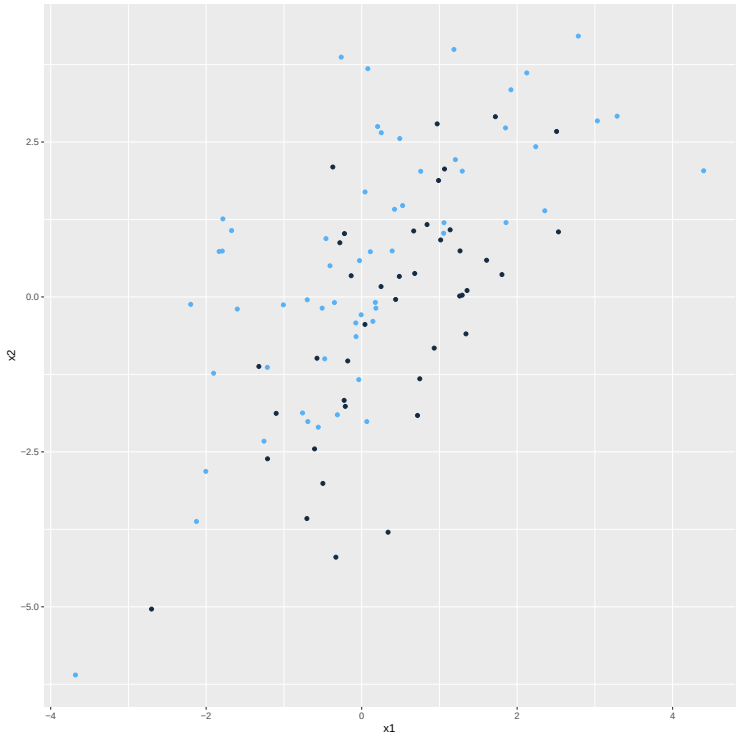
When the dependent variable is categorical, we instead try to:

1. estimate the relationship between two variables, and
2. predict the category of y_i based on the values in $\mathbf{x}_i$.

The former is typically called **regression**, the latter **classification**.

Note that we are still in a **supervised learning** context: we estimate the model to classify cases, based on observed categorisation $\mathbf{y}$.

Classification: example data



For example, lets assume we have two continuous variables $\mathbf{x}_1$ and $\mathbf{x}_2$ and try to understand a classify based on a binary dependent variable $\mathbf{y}$.

Outline

Linear regression

We could regress a dependent variable **y** coded as ones and zeros on **X** using linear regression and then apply a **decision rule** that we classify observation i as 1 if $\hat{y}_i > 0.5$ and as 0 otherwise.

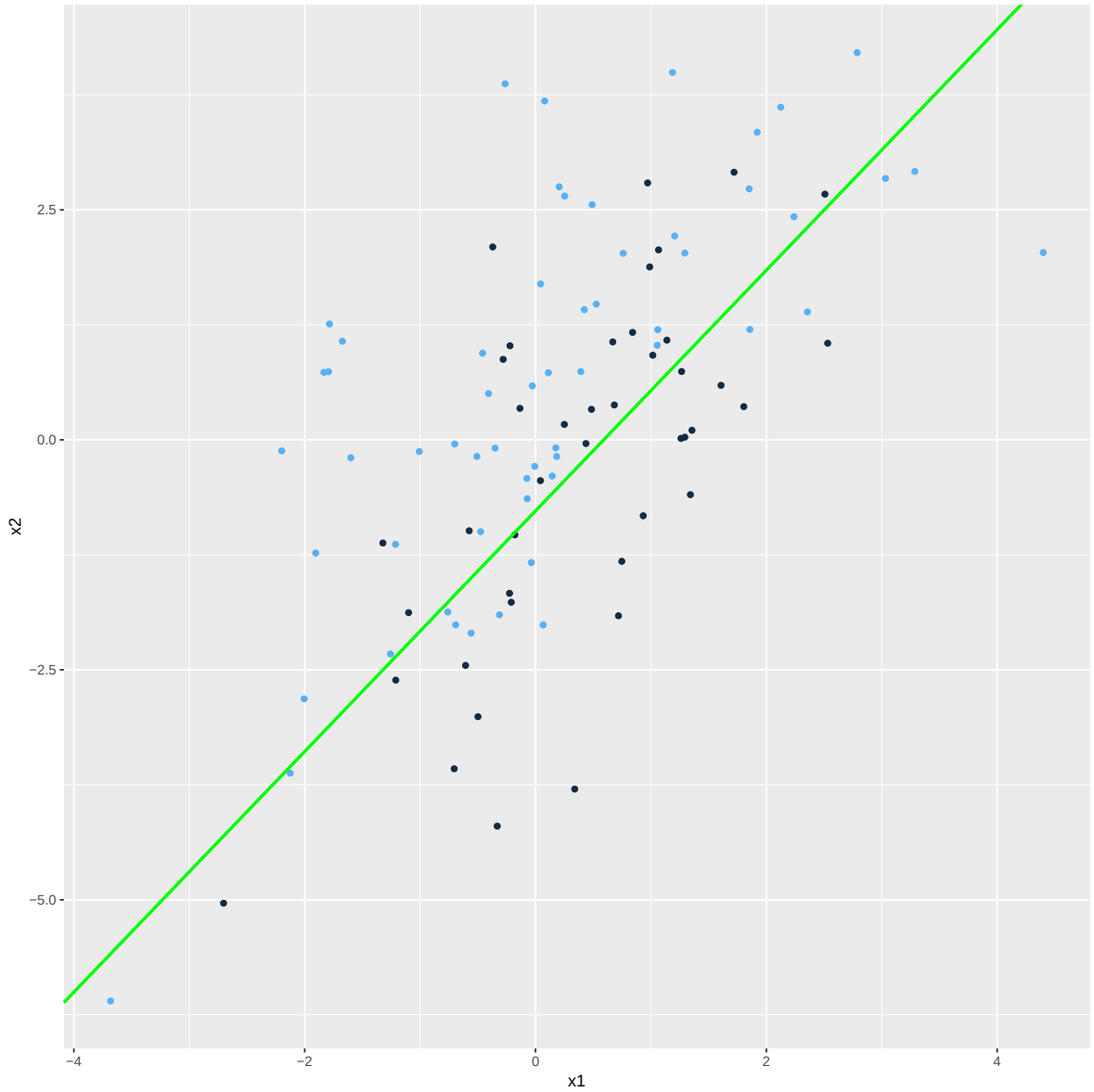
```
model <- lm(y ~ x1 + x2)
predicted <- ifelse(predict(model) > 0.5, 1, 0)
```

The **decision boundary** is then a straight line, as in:

$$y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} = 0.5$$

$$x_{i2} = -\frac{\beta_1}{\beta_3} - \frac{\beta_2}{\beta_3} x_{i1}$$

Linear regression

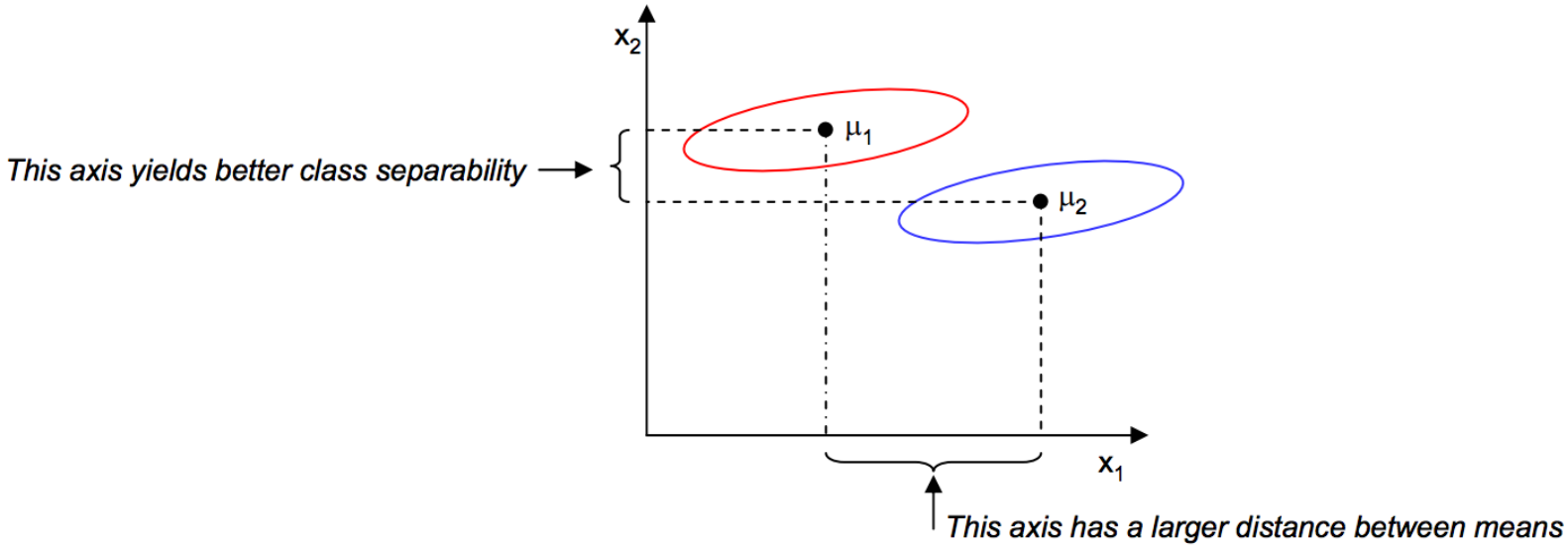


Outline

Linear discriminant analysis

An alternative might be to find the mean on **X** for each group in **y** and then classify based on the nearest mean.

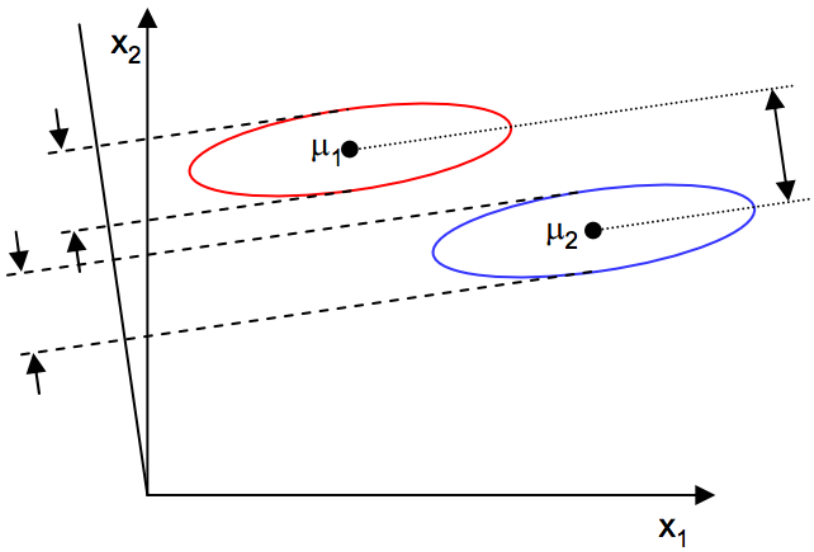
This is problematic, as the direction in which the means are most distinct, there might also be a lot more overlap in the data, e.g. x_1 in this plot:



Gutierrez-Osuna, http://www.nada.kth.se/∼stefanc/DATORSEENDE_AK/pr_l10.pdf

Linear discriminant analysis

Due to the variation within each group, a different projection of the means might be more suitable.



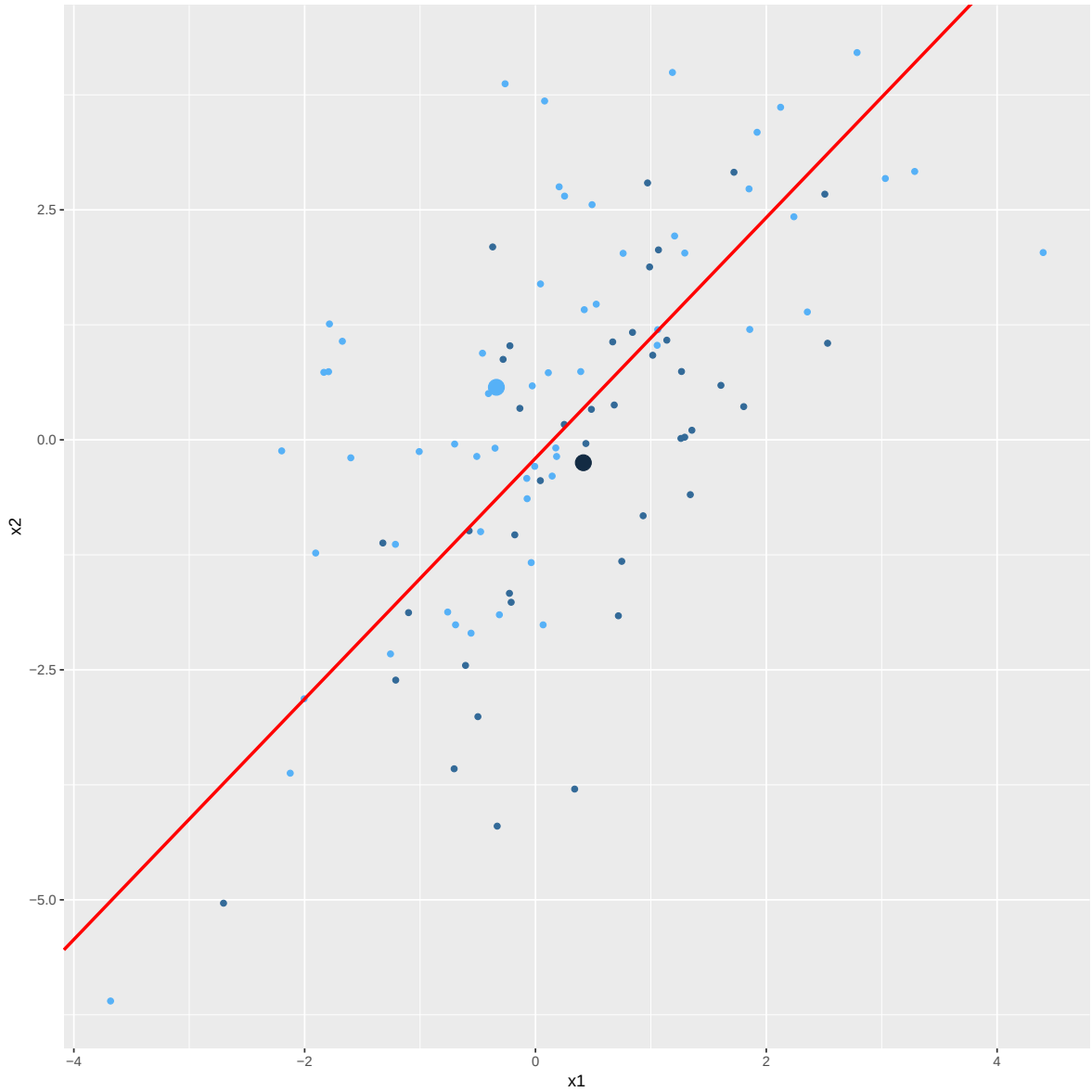
Linear discriminant analysis (**LDA**) projects the data such that the difference in means is maximized and the variances minimized.

The **decision boundary** is then orthogonal to this projection.

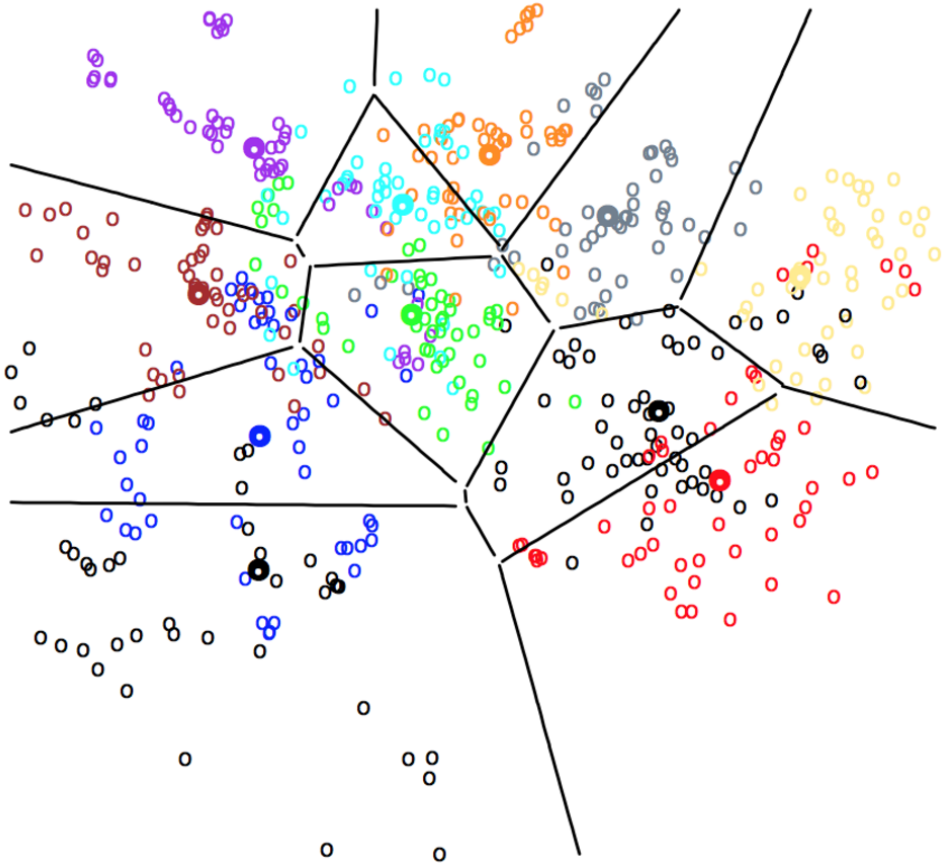
```
model <- lda(y ~ x1 + x2)
predicted <- predict(model)$class
```

Gutierrez-Osuna, http://www.nada.kth.se/∼stefanc/DATORSEENDE_AK/pr_l10.pdf

Linear discriminant analysis



Linear discriminant analysis



Unlike the linear regression example, LDA can be used for categorical variables with more than two categories.

(Hastie, Tibshirani and Friedman, 2009, 118)

Outline

Limited dependent variables

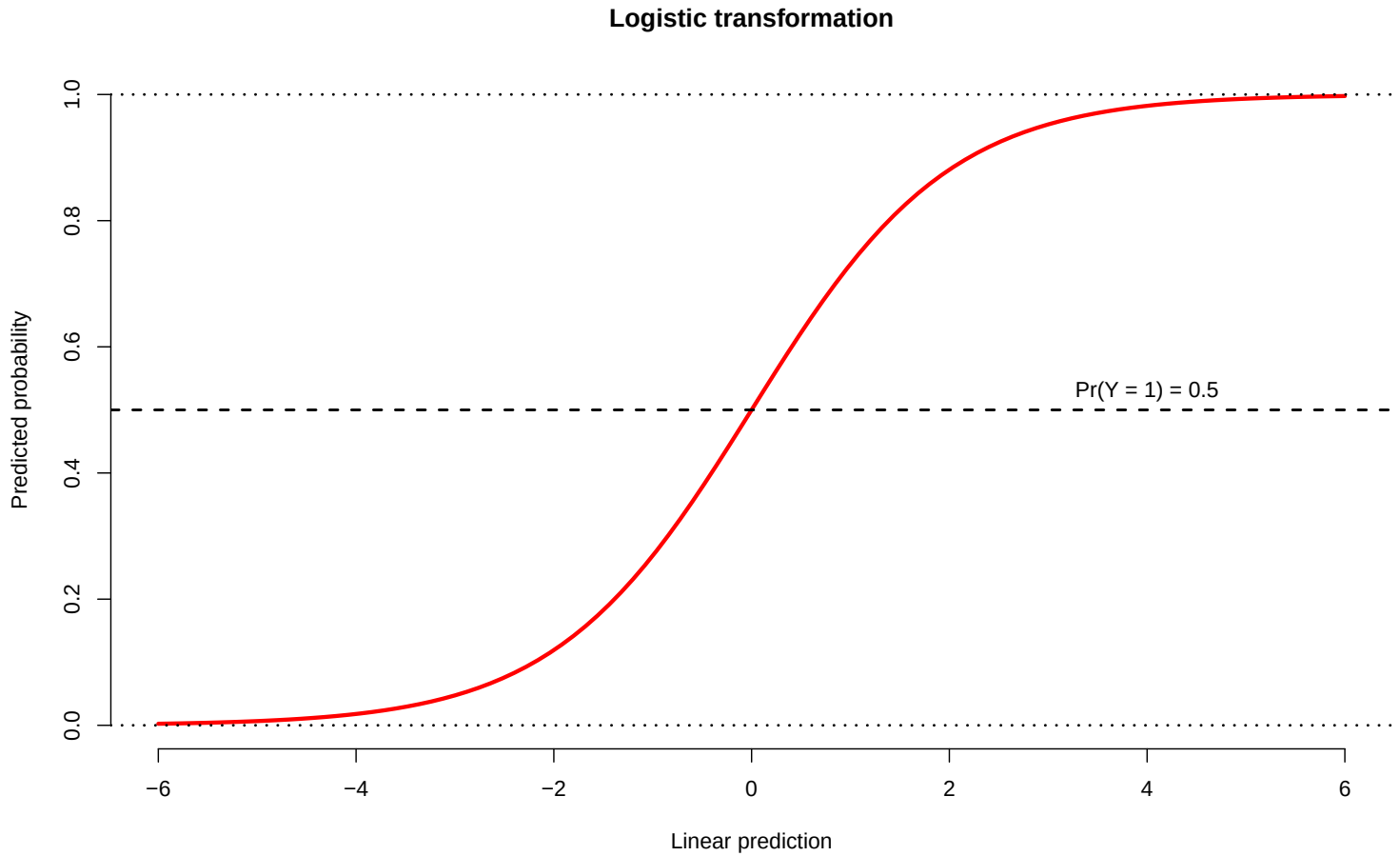
For binary dependent variables we might estimate the **probability of observing a one**.

When a dependent variable is not continuous, or is truncated for some reason, a linear model would lead to implausible predictions, i.e. impossible probabilities.

- Prediction below 0 and above 1 would not make sense.
- For any case where the predicted probability is already high, it cannot increase much with a change in $\boldsymbol{X}$ (and vice versa for low probabilities).
- A linear regression implies high levels of **heteroskedasticity** and therefore unreliable t-tests.

Estimators

A typical approach is to have an estimator that is "linear in the parameters" – i.e. it generates a linear prediction based on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ – but then transforms this linear prediction into one bounded between 0 and 1.



Logistic regression

The most common transformation is the logistic transformation, which relates to the log-odds:

$$\log\left(\frac{Pr(y_i = 1)}{Pr(y_i = 0)}\right) = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2},$$

which can also be formulated as:

$$Pr(y_i = 1) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2})}}.$$

```
model <- glm(y ~ x1 + x2, family =
                  binomial(link = "logit"))
predicted <- ifelse(predict(model) > 0, 1, 0)
```

Estimating a logistic regression

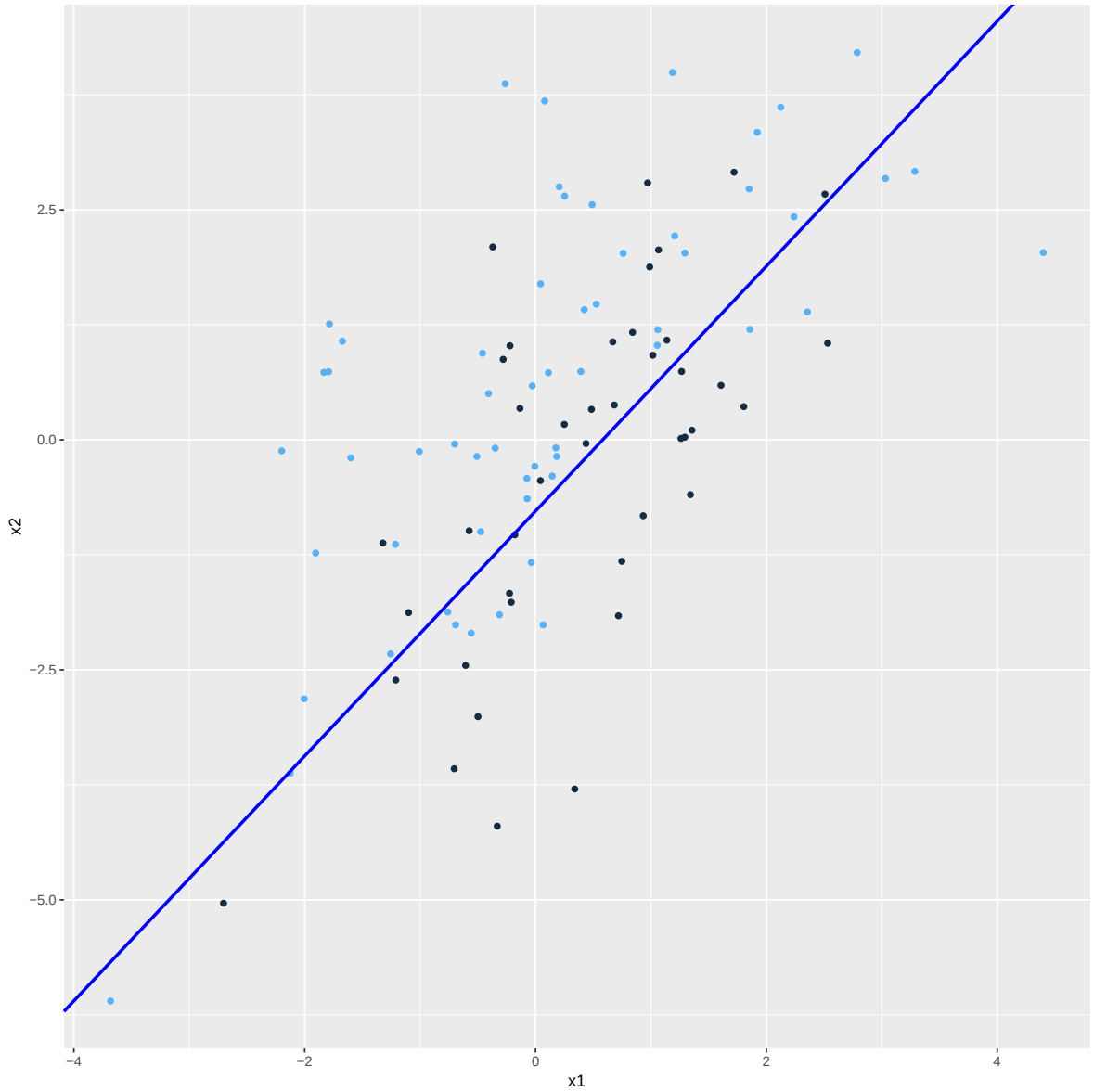
Estimating a logistic regression is straightforward and output will look similar to that of linear regression.

E.g. explaining "Yes" in the Marriage Equality Referendum.

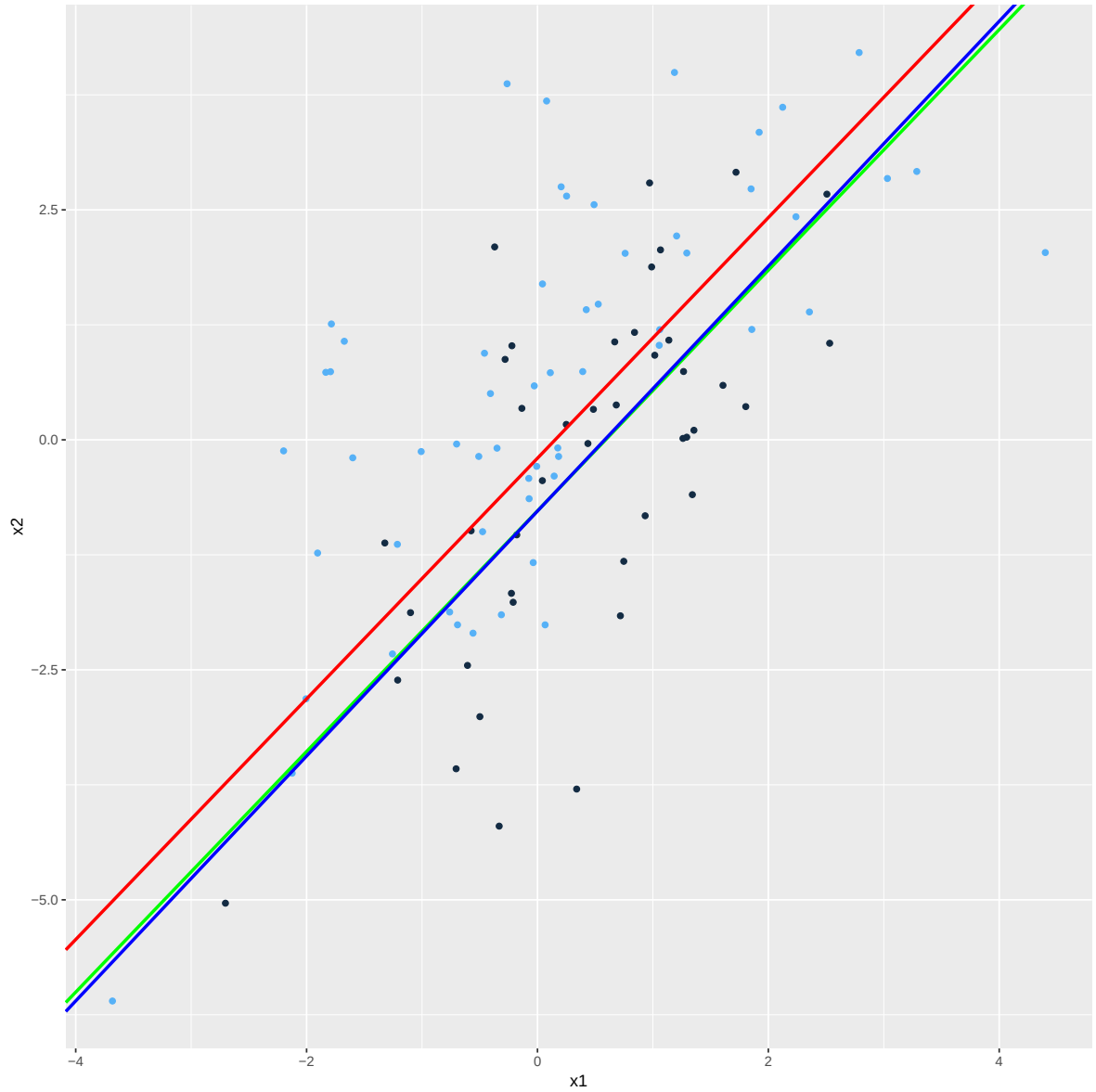
Note the use of continuous and discreet independent variables.

Age 25-34	-0.152
	(0.410)
35-44	-0.707^{*}
	(0.386)
45-54	-0.865^{**}
	(0.390)
55-64	-1.084^{***}
	(0.399)
65+	-1.857^{***}
	(0.374)
Urban	0.305^{*}
	(0.168)
Pro-abortion attitude	0.221^{***}
	(0.028)
intercept	0.358
	(0.372)
N	851

Logistic regression



Linear, logistic, and discriminant



Choosing between LDA and logistic

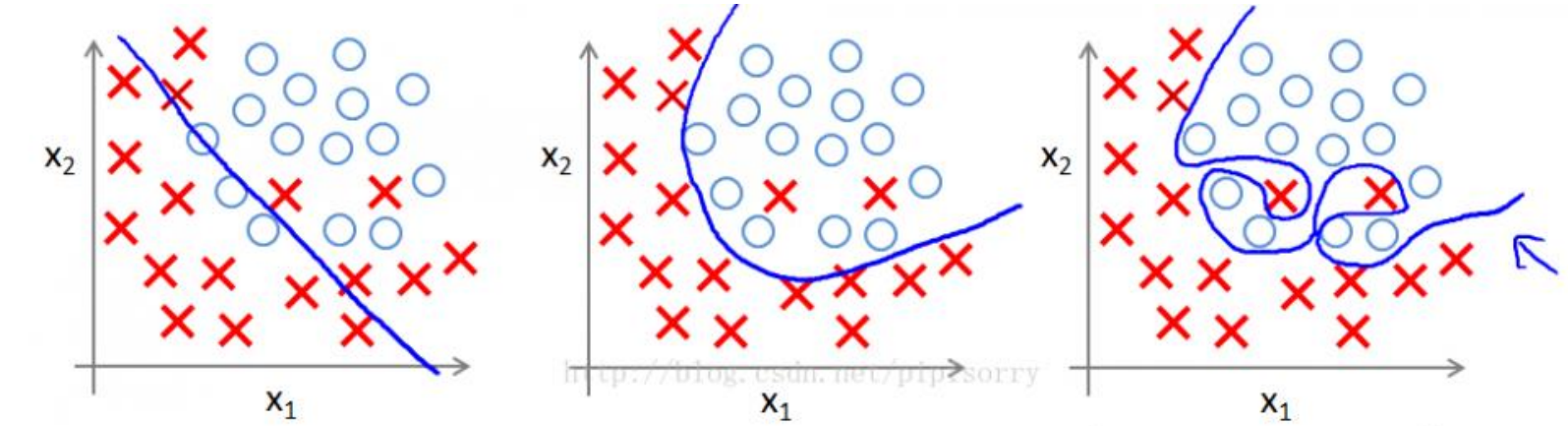
Generally, they provide very similar results, but:

- LDA is more sensitive to **outliers**, logistic regression more robust.
- LDA assumes normally distributed $\mathbf{X}$, while logistic regression makes no such assumption.
- If the assumptions are satisfied, LDA is more precise, has lower estimation variance.
- If the two classes are perfectly separable, logistic regression will fail, but LDA will not.
- Logistic regression more appropriate for imbalanced categories (e.g. 80% ones).

(Hastie, Tibshirani and Friedman, 2009, 128)

Stepwise and regularized

As with linear regression, stepwise variable selection approaches and regularized versions for logistic regression and discriminant analysis exist, but this is beyond the scope of this module.



$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

(g = sigmoid function)

"Underfit"

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$

"Overfit"

https://read01.com/jEdyx.html

Outline

Presentation of results

Contrary to linear regression, the coefficients for LDA or logistic regression are not straightforward to interpret.

It is therefore useful to try to graphically represent the effect.

Note that a quick method to interpret logistic regression coefficients is to divide them by 4 to get the slope at the point where $P(\mathbf{y} = 1|\mathbf{X}) = 0.5$.

Note that the graphical visualisation in the lab is a simplification and not really suitable for publication. The problem is that we present results from the simple regression (one independent variable), not the multiple regression—the latter is more involved due to the non-linear nature of the relation between $\mathbf{X}$ and $\hat{\mathbf{y}}$.

Outline

Confusion matrix

Evaluating the performance of the binary model can be done by using the **confusion matrix**:

		True value	
		1	0
Prediction	1	True positive (TP)	False positive (FP)
	0	False negative (TN)	True negative (FN)

Confusion matrix

Evaluating the performance of the binary model can be done by using the **confusion matrix**:

		True value 1	True value 0	
Prediction	1	True positive (TP)	False positive (FP)	Precision: $\frac{TP}{TP+FP}$
Prediction	0	False negative (TN)	True negative (FN)	
		Sensitivity: $\frac{TP}{TP+FN}$	Specificity: $\frac{TN}{FP+TN}$	

Receiver Operating Characteristic curve

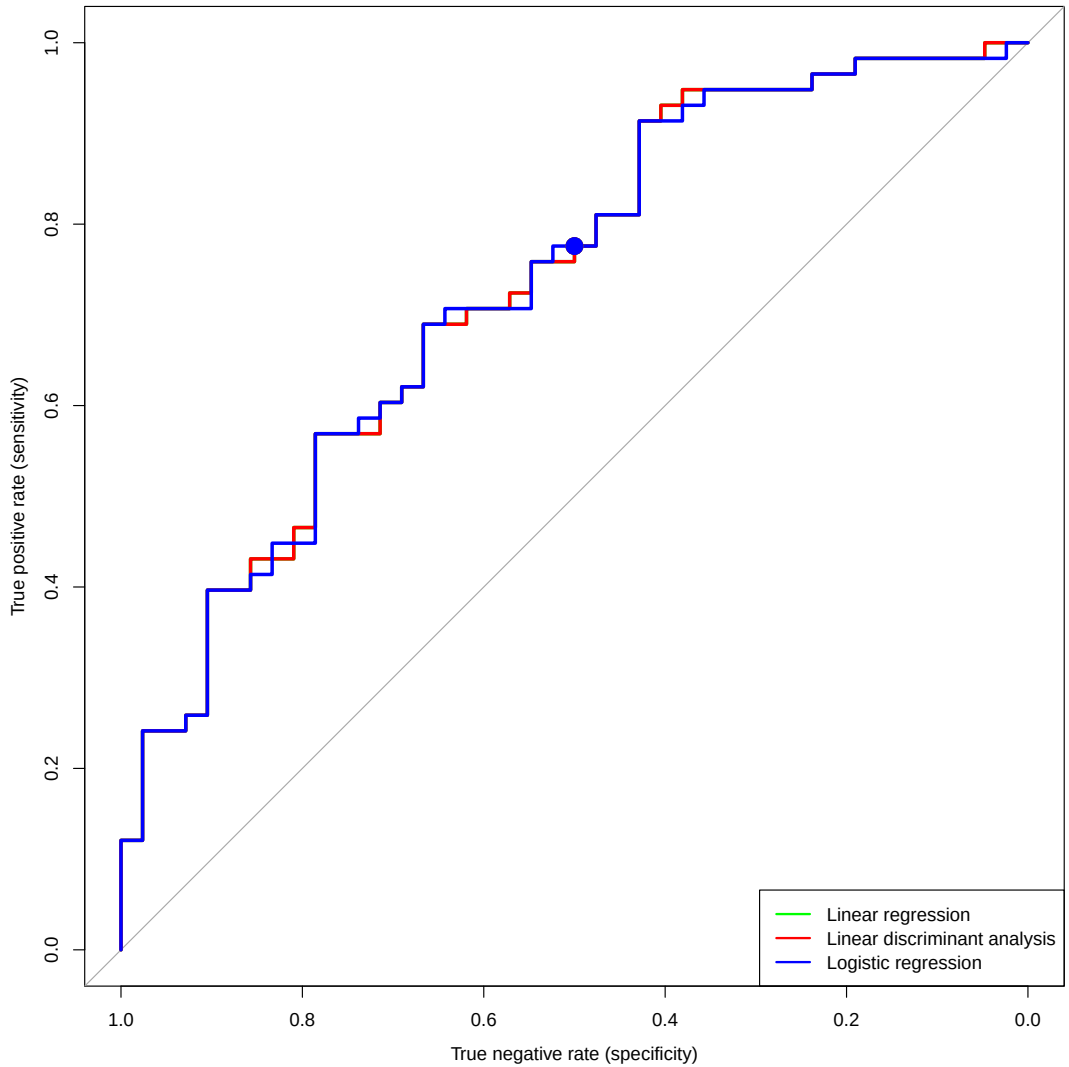
The accuracy of predictions will depend on the threshold probability—variations on default of $\hat{\pi} = 0.5$ are possible.

Depending on the application, it might be better or worse to over- or underestimate ones relative to zeros.

The ROC-curve plots, for all possible thresholds, the true positive rate against the false positive rate.

An ROC-curve further from (above) the 45 degree line indicates a better predictive performance; any predictions under this line indicate worse than random prediction.

ROC curves



Area Under Curve (AUC)

Given the above, we can also calculate the area under the ROC-curve as a measure of prediction quality.

This is somewhat related to the Gini coefficient for income distributions ($G = 2AUC - 1$).

```
library(pROC)
model <- lm(y ~ x1 + x2)
auc(y, predict(model))
plot(roc(y, predict(model)))
```

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition ed. New York: Springer.