

# Logistic Regression

ACTL3142 & ACTL5110 Statistical Machine Learning for Risk Applications

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani



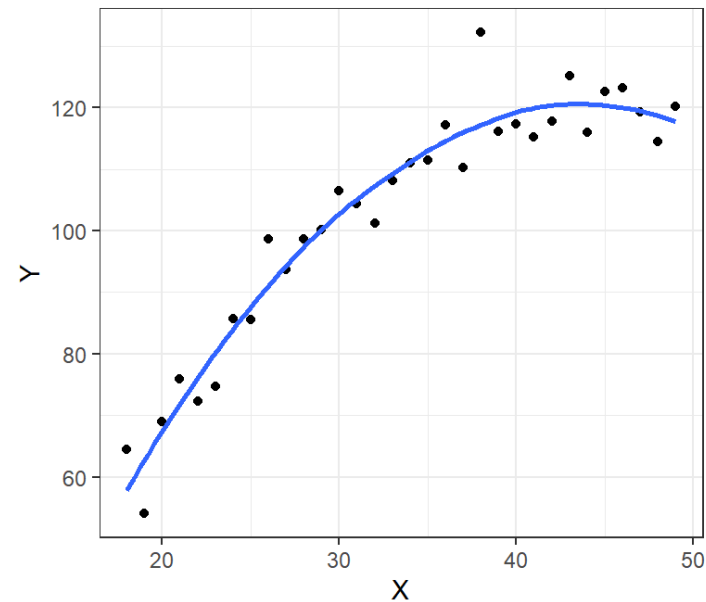
## Lecture Outline

- **An overview of classification**
- Logistic regression
- Poisson regression
- Generalised linear models



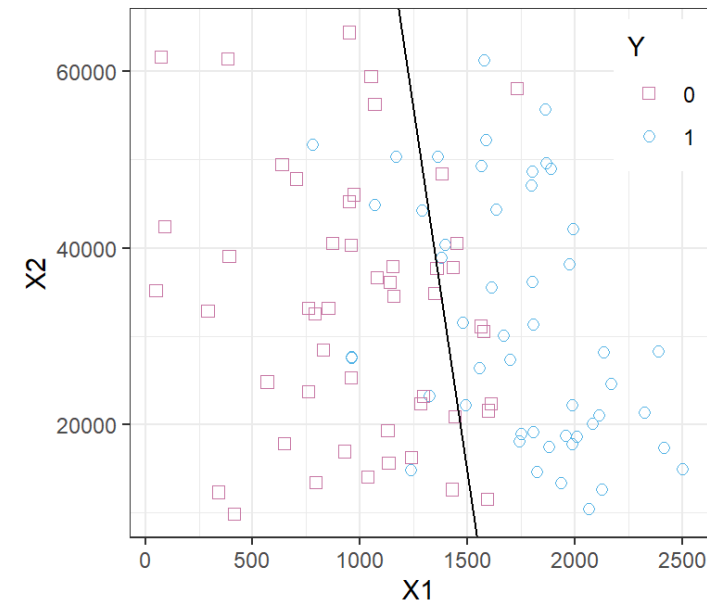
# Regression vs. classification

## Regression



- $Y$  is quantitative, continuous
- Examples: Sales prediction, claim size prediction, stock price modelling

## Classification



- $Y$  is qualitative, discrete
- Examples: Fraud detection, face recognition, accident occurrence, death



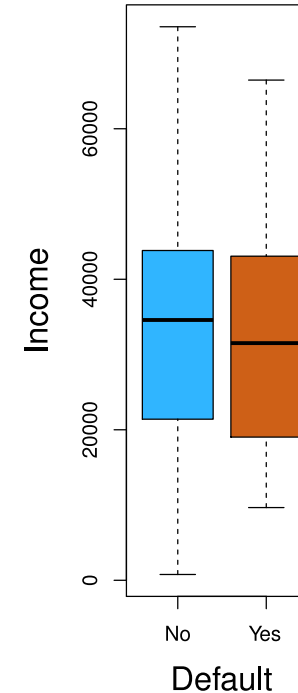
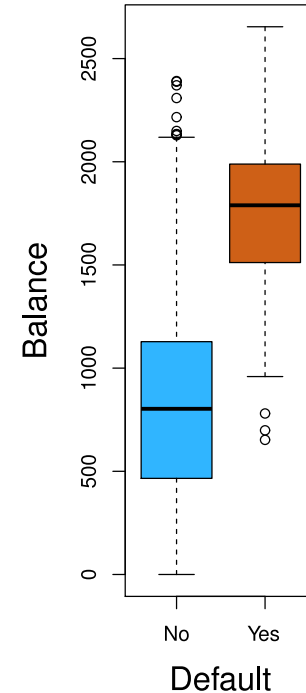
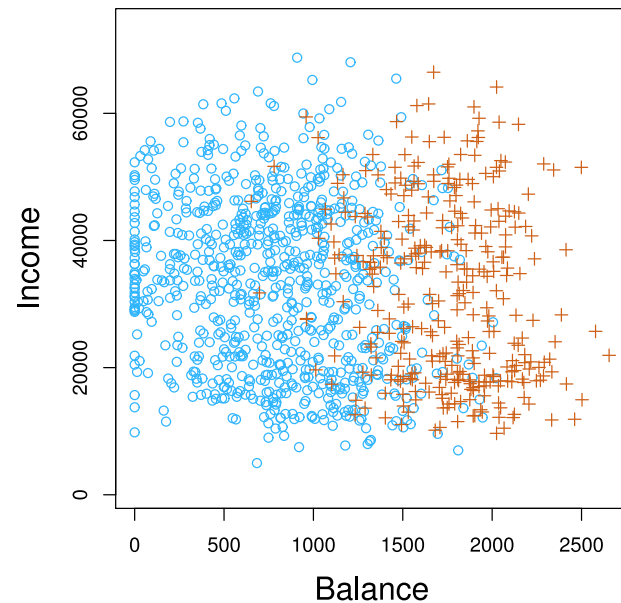
# Some examples of classification problems

- Success/failure of a treatment, explained by dosage of medicine administered, patient's age, sex, weight and severity of condition, etc.
- Vote for/against political party, explained by age, gender, education level, region, ethnicity, geographical location, etc.
- Customer churns/stays depending on usage pattern, complaints, social demographics, etc.



# Example: Predicting defaults (Default from ISLR2)

$$Y = X\beta + \varepsilon$$



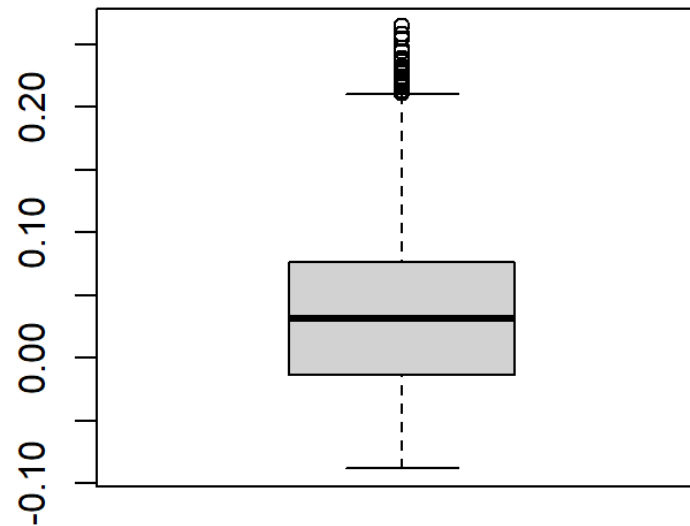
- **default** ( $Y$ ) is a binary variable (yes/no or 0/1)
- Annual **income** ( $X_1$ ) and credit card **balance** ( $X_2$ ) may be continuous predictors



# Example: Predicting defaults - Discussion

Simple linear regression on **Default** data:

Fitted values of default probability  $y = X\beta$  probability of default.



What do you observe?



# Classification problems

- Coding in the binary case is simple


$$Y \in \{0, 1\} \Leftrightarrow Y \in \{\text{blue}, \text{green}\}$$

- Our objective is to find a good predictive model  $f$  that can:

1. Estimate the probability

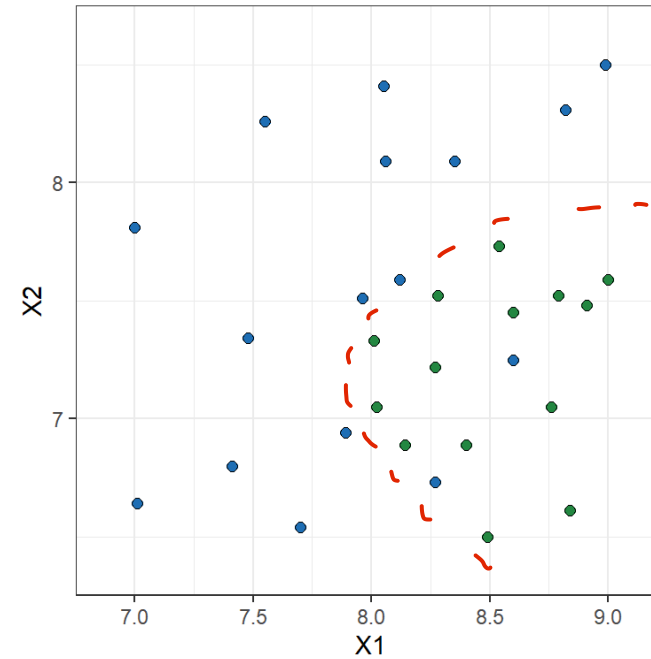
$$\mathbb{P}(Y = 1|X) \in \{0, 1\}$$

Probability of is blue based on  $X$ .

$f(X) \rightarrow$  

2. Classify observation

$$f(X) \rightarrow \hat{Y} \in \{\text{blue}, \text{green}\}$$



## Lecture Outline

- An overview of classification
- **Logistic regression**
- Poisson regression
- Generalised linear models





# Logistic regression

Extend linear regression to model binary categorical variables

$$\underbrace{\ln\left(\frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)}\right)}_{\text{log-odds}} = \underbrace{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}_{\text{linear model}}$$

*Model log-odds as a linear model.*

# Principles of Logistic Regression

- The output is binary  $Y \in \{1, 0\}$
- Each case's  $Y$  variable has a probability between 0 and 1 that depends on the values of the predictors  $X$  such that

$$\mathbb{P}(\underline{Y = 1|X}) + \mathbb{P}(\underline{Y = 0|X}) = 1$$

- Probability can be restated as odds

$$\text{Odds}(Y = 1|X) = \frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)} = \frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)}$$

- Odds are a measure of relative probabilities

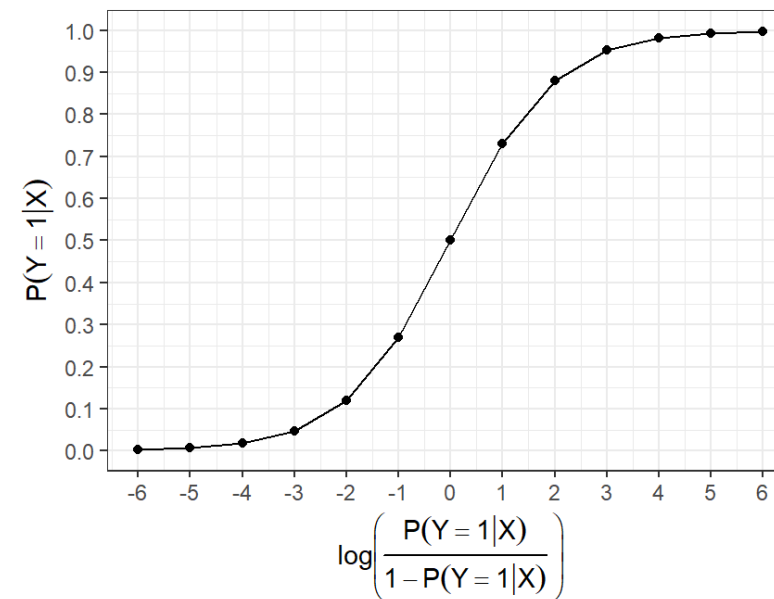
e.g. Odds = 3 ||  $Y=1$  is 3x more likely than  $Y=0$ .  
 $\mathbb{P}(Y=1|X) = 0.75$



# Probabilities, odds and log-odds

**Goal:** Transform a number between 0 and 1 into a number between  $-\infty$  and  $\infty$

probability	odds	logodds
0.001	0.001	-6.907
0.250	0.333	-1.099
0.500	1.000	0.000
0.750	3.000	1.099
0.999	999.000	6.907



If we have log-odds, we have the probabilities!



# Logistic regression

- Perform regression on log-odds

$$\ln \left( \frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

- Use (training) data and maximum-likelihood estimation to produce estimates  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ .
- Predict probabilities using

$$\mathbb{P}(Y = 1|X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p}}$$

$$\ln \left( \frac{p}{1-p} \right) = X\beta$$



# Interpretation of coefficients


- Recall for **multiple linear regression** we model the response as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

An increase of the entry  $x_{ij}$  by 1 in  $X$  we would predict  $Y_i$  to increase by  $\hat{\beta}_j$  on average since

$$\mathbb{E}[Y_i | X] = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_j (x_{ij} + 1) + \cdots + \hat{\beta}_p x_{ip}$$

- For **logistic regression** we have a similar relationship. When  $x_{ij}$  increases by 1 we would expect the **log-odds** for  $Y_i$  to increase by  $\beta_j$ .
- The new predicted probability of success by increasing  $x_{ij}$  by 1 is now



$$\mathbb{P}(Y_i = 1 | X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_j (x_{ij} + 1) + \cdots + \hat{\beta}_p x_{ip}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_j (x_{ij} + 1) + \cdots + \hat{\beta}_p x_{ip}}}.$$

Convince yourself that the probability does increase if  $\beta_j$  is positive!



# How are the coefficients estimated?

- Recall the Bernoulli distribution is parameterised by a parameter  $p$  and has the density

$$f(y) = p^y (1 - p)^{1-y}.$$

- In logistic regression we maximise the likelihood of the data. Denote

$$p(y_i; \beta) = \frac{1}{1 + e^{-x_i \beta}},$$

where  $x_i$  denotes the  $i$ 'th row of  $X$ .

- We maximise the log-likelihood below

$$\ell(\beta) = \sum_{i=1}^n y_i \ln p(y_i; \beta) + (1 - y_i) \ln(1 - p(y_i; \beta)).$$

We take partials w.r.t. to each  $\beta_j$  and set to 0. Needs numerical approximation.

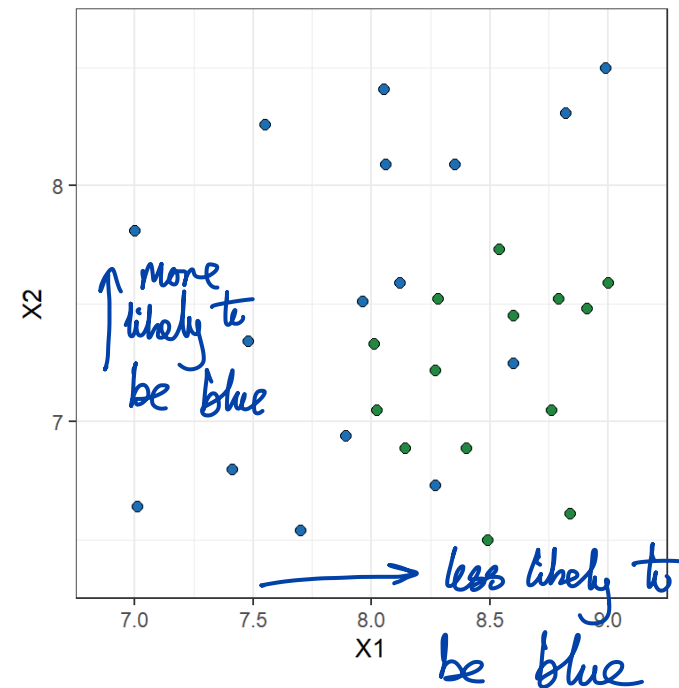
*i.e.*  $\frac{\partial}{\partial \beta_j} \ell(\beta) = 0$



# Toy example: Logistic Regression

$$Y = \begin{cases} 1 & \text{if } \bullet \text{ (blue)} \\ 0 & \text{if } \bullet \text{ (green)} \end{cases} \quad \ln \left( \frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- The parameter estimates are  $\hat{\beta}_0 = 13.671$ ,  $\hat{\beta}_1 = -4.136$ ,  $\hat{\beta}_2 = 2.803$
- $\hat{\beta}_1 = -4.136$  implies that the bigger  $X_1$  the lower the chance it is a blue point
- $\hat{\beta}_2 = 2.803$  implies that the bigger  $X_2$  the higher the chance it is a blue point

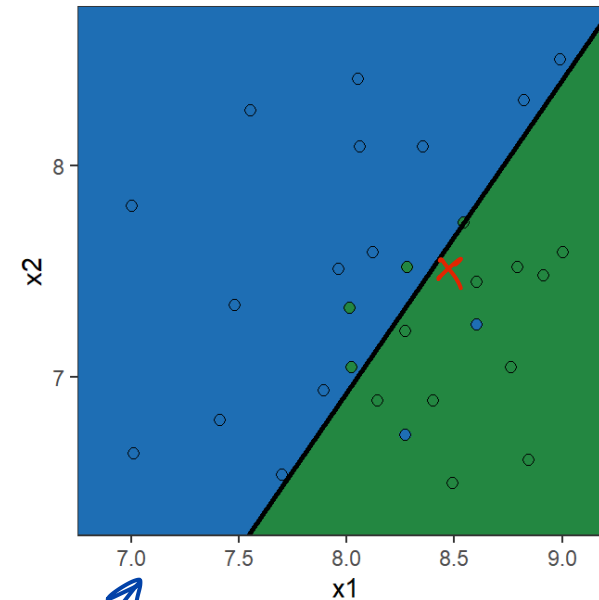


Interpretation is dependent on how we code  $Y$ .

# Toy example: Logistic Regression

$$\ln \left( \frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)} \right) = 13.671 - 4.136X_1 + 2.803X_2$$

X1	X2	log-odds	P(Y=1 X)	prediction
7.0	8.0	7.14	0.9992	blue
8.0	7.5	1.61	0.8328	blue
8.0	7.0	0.20	0.5508	blue
8.5	7.5	-0.46	0.3864	green
9.0	7.0	-3.93	0.0192	green



We classify blue if  $P(\hat{Y} = 1|X) > 0.5$



# Some important points about logistic regression

- Changes in predictor values correspond to changes in the *log-odds*, not the probability
- Evaluating predictors to add / remove is the same as in linear regression. The only change is the form of the response
- As a result, most of the modelling limitations of linear regression (e.g. collinearity) carry over as well
- Possible to do logistic regression on non-binary responses, but not used that often, and not covered here

→ We have our hypothesis testing still!



# Example: Predicting defaults

```
1 glmStudent <- glm(default ~ student, family = binomial(), data = ISLR2::Default)
2 summary(glmStudent)
```

Call:

```
glm(formula = default ~ student, family = binomial(), data = ISLR2::Default)
```

Coefficients:

	Estimate	Std. Error	<i>not t</i> z value	Pr(> z )
(Intercept)	-3.50413	0.07071	-49.55	< 2e-16 ***
studentYes	0.40489	0.11502	3.52	0.000431 ***

---

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: 2908.7 on 9998 degrees of freedom

AIC: 2912.7

Number of Fisher Scoring iterations: 6

*1 is coded as default.*

*Next week*

*F-test equivalent*



# Example: Predicting defaults

```
1 glmAll <- glm(default ~ balance + income + student, family = binomial(), data = ISLR2::Default)
2 summary(glmAll)
```

Call:

```
glm(formula = default ~ balance + income + student, family = binomial(),
     data = ISLR2::Default)
```

Coefficients:

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

---

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

*Better than previous model judging by AIC.*

*2912 vs 1579*



# Example: Predicting defaults - Discussion

Results of logistic regression:

`default` against `student`

Predictor	Coefficient	Std error	Z-statistic	P-value
(Intercept)	-3.5041	0.0707	-49.55	<0.0001
<code>student = Yes</code>	0.4049	0.1150	3.52	0.0004

`default` against `balance`, `income`, and `student`

Predictor	Coefficient	Std error	Z-statistic	P-value
(Intercept)	-10.8690	0.4923	-22.080	< 0.0001
<code>balance</code>	0.0057	2.319e-04	24.738	< 0.0001
<code>income</code>	0.0030	8.203e-06	0.370	0.71152
<code>student = Yes</code>	-0.6468	0.2362	-2.738	0.00619



# Assessing accuracy in classification problems

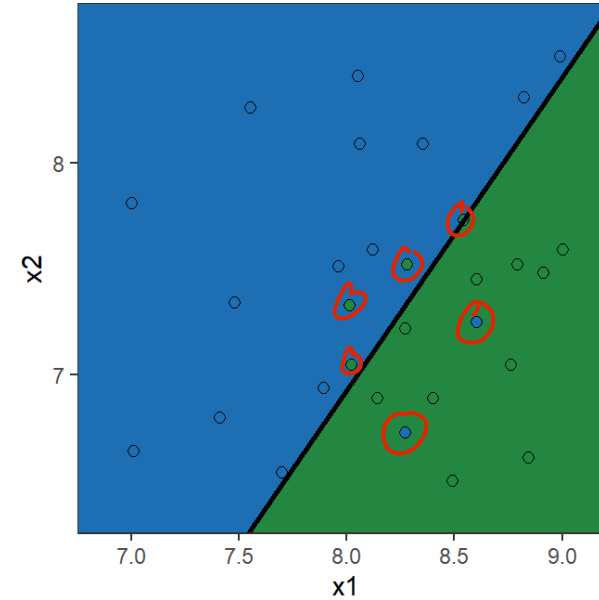
- We assess model accuracy using the error rate

$$\text{error rate} = \frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i)$$

1 if  $\hat{y}_i \neq y_i$

- In our toy example with a 50% threshold

$$\text{training error rate} = \frac{6}{30} = 0.2$$



# Confusion matrix: Toy example (50% Threshold)

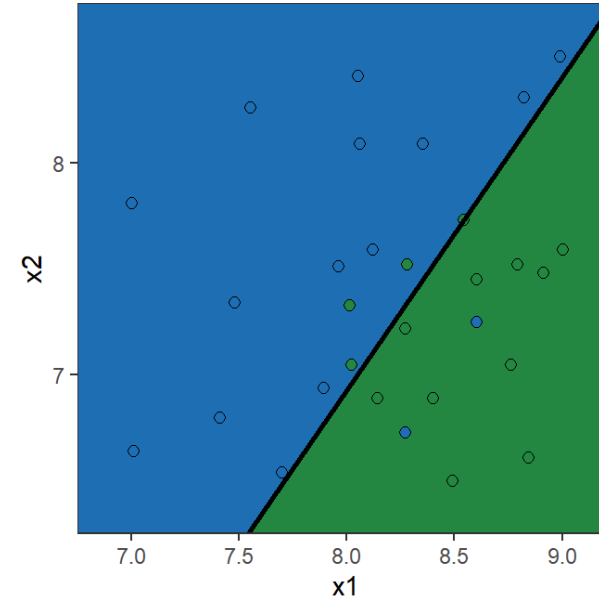
- Confusion matrix

	$Y = 0$	$Y = 1$	Total
$\hat{Y} = 0$	10	2	12
$\hat{Y} = 1$	4	14	18
Total	14	16	30

- True-Positive Rate =  $\frac{14}{16} = 0.875$
- False-Positive Rate =  $\frac{4}{14} = 0.286$

4 predicted  
to be

blue  
when they  
were not



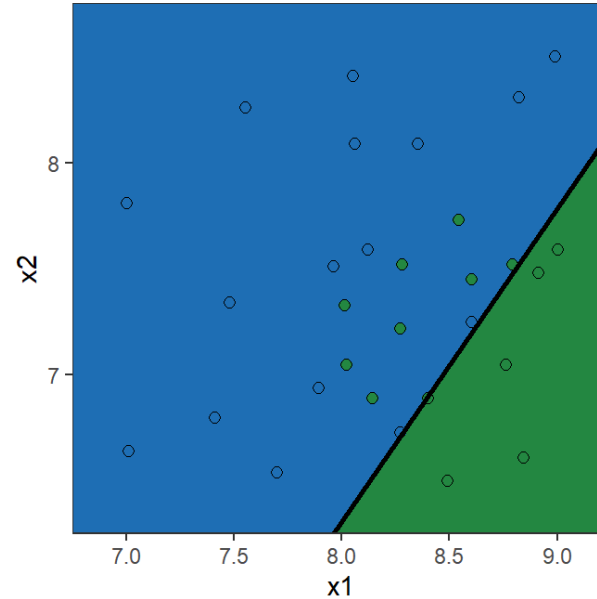
Predict 1 if  $P(\hat{Y}=1|X) > 0.5$

# Confusion matrix: Toy example (15% Threshold)

- Confusion matrix

	$Y = 0$	$Y = 1$	Total
$\hat{Y} = 0$	6	0	6
$\hat{Y} = 1$	8	16	24
Total	14	16	30

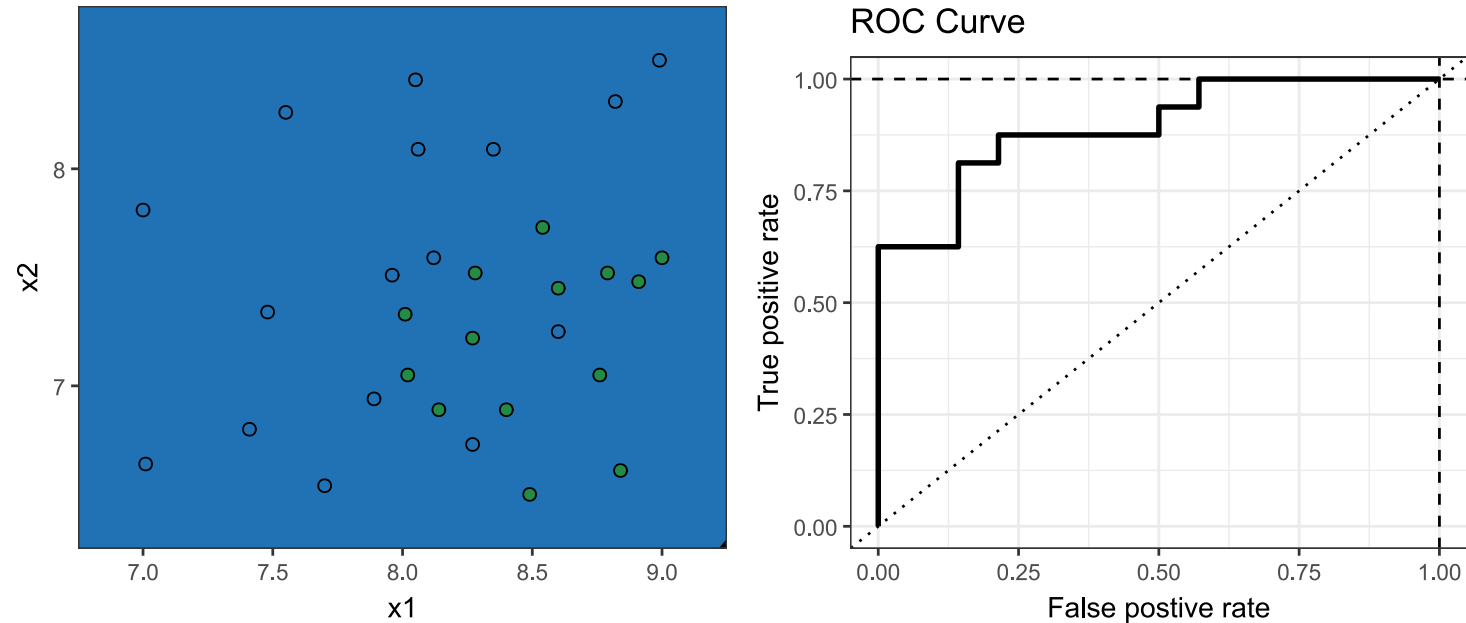
- True-Positive Rate =  $\frac{16}{16} = 1$
- False-Positive Rate =  $\frac{8}{14} = 0.429$



Changing threshold  
effects TPR FPR.

$$1 \text{ if } P(Y=1 | X) > 0.15$$

# ROC Curve and AUC: Toy example



- ROC Curve: Plots the true-positive rate against the false-positive rate
- A good model will have its ROC curve hug the top-left corner more
- AUC is the area under the ROC curve: For this toy example  $AUC = 0.8929$



## Lecture Outline

- An overview of classification
- Logistic regression
- **Poisson regression**
- Generalised linear models



# Poisson regression - Motivation

In many application we need to model count data:

- In mortality studies the aim is to explain the number of deaths in terms of variables such as age, gender and lifestyle.
- In health insurance, we may wish to explain the number of claims made by different individuals or groups of individuals in terms of explanatory variables such as age, gender and occupation.
- In general insurance, the count of interest may be the number of claims made on vehicle insurance policies. This could be a function of the color of the car, engine capacity, previous claims experience, and so on.



# The Bikeshare dataset

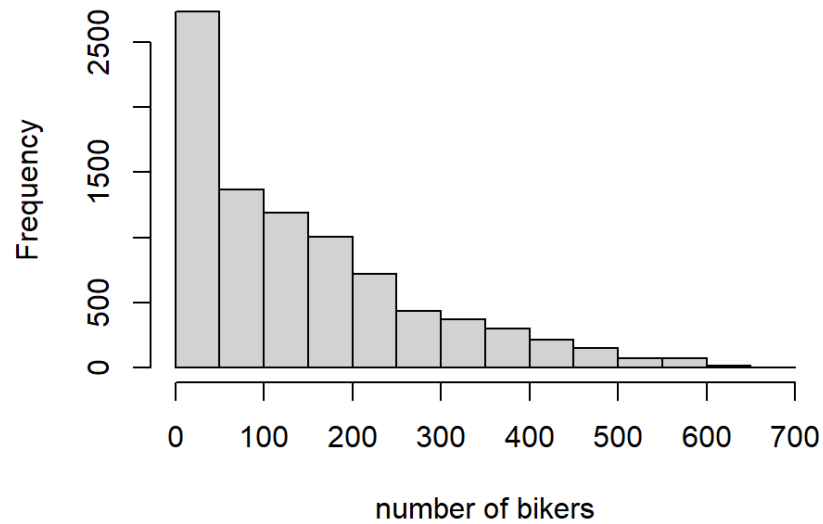
```
1 str(ISLR2::Bikeshare)
```

```
'data.frame': 8645 obs. of 15 variables:
 $ season : num 1 1 1 1 1 1 1 1 1 1 ...
 $ mnth : Factor w/ 12 levels "Jan","Feb","March",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ day : num 1 1 1 1 1 1 1 1 1 1 ...
 $ hr : Factor w/ 24 levels "0","1","2","3",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ holiday : num 0 0 0 0 0 0 0 0 0 0 ...
 $ weekday : num 6 6 6 6 6 6 6 6 6 6 ...
 $ workingday: num 0 0 0 0 0 0 0 0 0 0 ...
 $ weathersit: Factor w/ 4 levels "clear","cloudy/misty",...: 1 1 1 1 1 2 1 1 1 1 ...
 $ temp : num 0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...
 $ atemp : num 0.288 0.273 0.273 0.288 0.288 ...
 $ hum : num 0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...
 $ windspeed : num 0 0 0 0 0 0.0896 0 0 0 0 ...
 $ casual : num 3 8 5 3 0 0 2 1 1 8 ...
 $ registered: num 13 32 27 10 1 1 0 2 7 6 ...
 $ bikers : num 16 40 32 13 1 1 2 3 8 14 ...
```

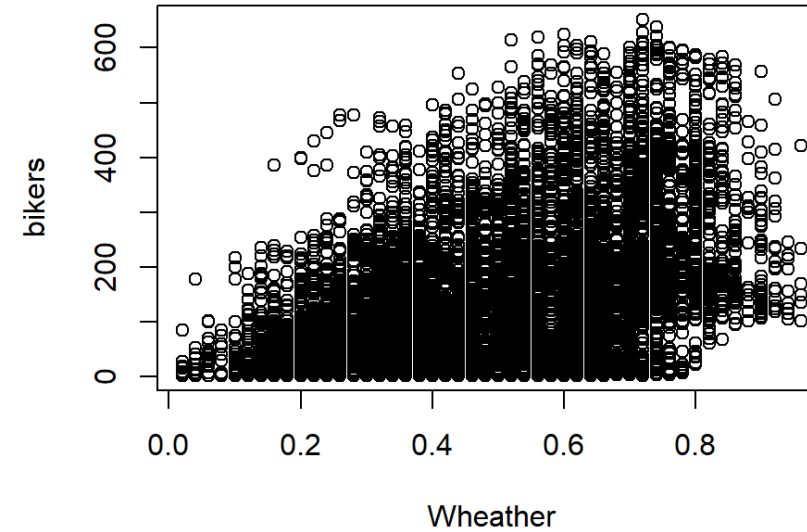


# The Bikeshare dataset - Discussion

Histogram of bikers



Bikers vs. temperature



How could we model the number of **bikers** as function of the other variables?



# Why not use multiple linear regression?

$$\text{✂ } Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

- Could predict negative values
- Constant variance may be inadequate
- Assumes continuous numbers while counts are integers

$$\underline{\log(Y)} = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

- Solves problem of negative values
- May solve constant variance problem
- Assumes continuous numbers while counts are integers
- Not applicable with zero counts



# Poisson regression

- Assume that  $Y \sim \text{Poisson}(\lambda)$

$$\mathbb{P}(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \dots \quad \text{with } \underline{\mathbb{E}[Y]} = \underline{\text{Var}(Y)} = \lambda$$

- Assume that  $\mathbb{E}[Y] = \lambda(X_1, \dots, X_p)$  is log-linear in the predictors

$$\log(\lambda(X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

- Use data and maximum-likelihood estimation to obtain  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

$$\mathcal{L}(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^n \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!} \quad \text{with } \log(\lambda(x_i)) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$



# Some important points about Poisson regression

- Interpretation: An increase in  $X_j$  by one unit is associated with a change in  $\mathbb{E}[Y]$  by a factor  $e^{\beta_j}$ .
- Mean-variance relationship:  $\mathbb{E}[Y] = \text{Var}(Y) = \lambda$  implies that the variance is non-constant and increases with the mean.
- Non-negative fitted values: Predictions are always positive
- Evaluating predictors to add / remove is the same as in linear regression. The only change is the form of the response *(more like logistic than linear)*.
- As a result, most of the modelling limitations of linear regression (e.g. collinearity) carry over as well

$$\log(\lambda_i) = \beta_0 + \dots + \beta_j(x_{ij} + 1) + \dots + \beta_p x_{ip}$$

$$\lambda_i = e^{\beta_j} \cdot e^{x_{ij} \beta}$$

↑  
i'th row of  $X$ .



# Poisson regression - Bikeshare dataset

```

1 glmBikeshare <- glm(bikers ~ workingday + temp + weathersit + mnth + hr, family = poisson(),
2                   data = ISLR2::Bikeshare)
3 summary(glmBikeshare)

```

Call:

```

glm(formula = bikers ~ workingday + temp + weathersit + mnth +
     hr, family = poisson(), data = ISLR2::Bikeshare)

```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	2.693688	0.009720	277.124	< 2e-16	***
workingday	0.014665	0.001955	7.502	6.27e-14	***
temp	0.785292	0.011475	68.434	< 2e-16	***
weathersitcloudy/misty	-0.075231	0.002179	-34.528	< 2e-16	***
weathersitlight rain/snow	-0.575800	0.004058	-141.905	< 2e-16	***
weathersitheavy rain/snow	-0.926287	0.166782	-5.554	2.79e-08	***
mnthFeb	0.226046	0.006951	32.521	< 2e-16	***
mnthMarch	0.376437	0.006691	56.263	< 2e-16	***
mnthApril	0.691693	0.006987	98.996	< 2e-16	***
mnthMay	0.910641	0.007436	122.469	< 2e-16	***
mnthJune	0.893405	0.008242	108.402	< 2e-16	***
mnthJuly	0.773787	0.008806	87.874	< 2e-16	***
mnthAug	0.821341	0.008332	98.573	< 2e-16	***
mnthSept	0.903663	0.007621	118.578	< 2e-16	***
mnthOct	0.937743	0.006744	139.054	< 2e-16	***

expect  
less  
bikes  
compared  
to  
clear  
weather



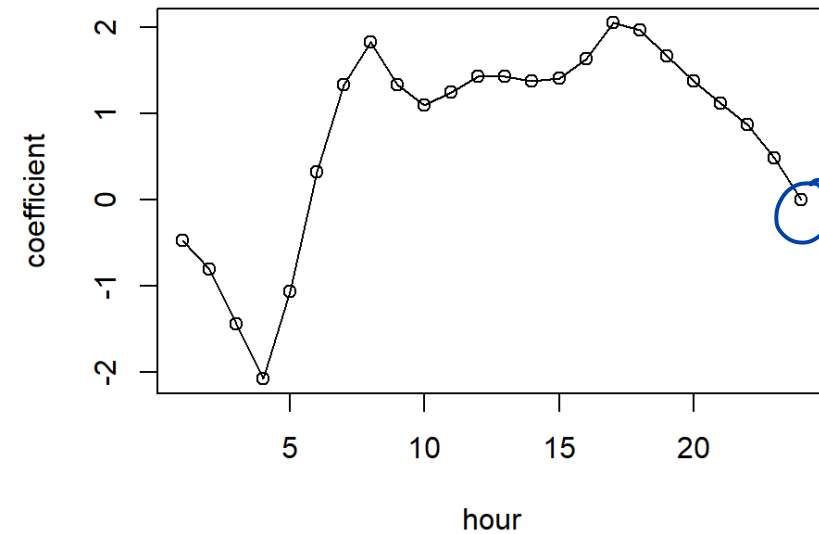
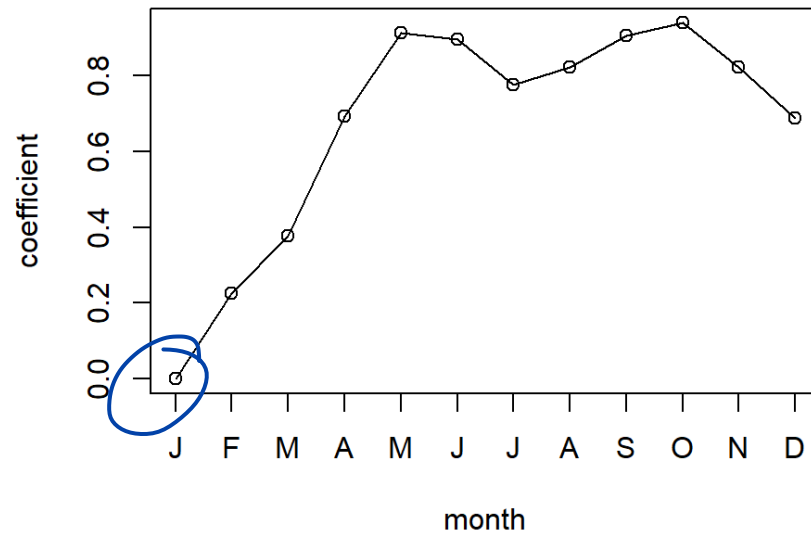


# Poisson regression - Bikeshare dataset

```

1 plot(x = 1:12, y = c(0, glmBikeshare$coefficients[7:17]), type = 'o',
2      xlab = "month", ylab = "coefficient", xaxt = "n")
3 axis(1, at=1:12, labels=substr(month.name, 1, 1))
4 plot(x = 1:24, y = c(glmBikeshare$coefficients[18:40], 0), type = 'o',
5      xlab = "hour", ylab = "coefficient")

```



## Lecture Outline

- An overview of classification
- Logistic regression
- Poisson regression
- **Generalised linear models**



# Generalised linear models

	Linear Regression	Logistic Regression	Poisson Regression	Generalised Linear Models
<b>Type of Data</b>	Continuous	Binary (Categorical)	Count	Flexible
<b>Use</b>	Prediction of continuous variables	Classification	Prediction of the number of events	Flexible
<b>Distribution of Y</b>	Normal	Bernoulli (Binomial for multiple trials)	Poisson	Exponential Family
$\mathbb{E}[Y X]$	$X\beta$	$\frac{e^{X\beta}}{1+e^{X\beta}}$	$e^{X\beta}$	$g^{-1}(X\beta)$
<b>Link Function Name</b>	Identity	Logit	Log	Depends on the choice of distribution
<b>Link Function Expression</b>	$\eta(\mu) = \mu$	$\eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	$\eta(\mu) = \log(\mu)$	Depends on the choice of distribution

