Logistic Regression

ACTL3142 & ACTL5110 Statistical Machine Learning for Risk and Actuarial Applications

Disclaimer

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2021) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Overview

- Introduction to classification
- Logistic regression
- Poisson regression
- Introduction to generalised linear models ← Standard linear regression

Reading

James et al. (2021): Chapters 4.1, 4.2, 4.3, 4.6, 4.7.1, 4.7.2, 4.7.6, 4.7.7

~~quarto::include_file("_classification.qmd")~~

An overview of classification

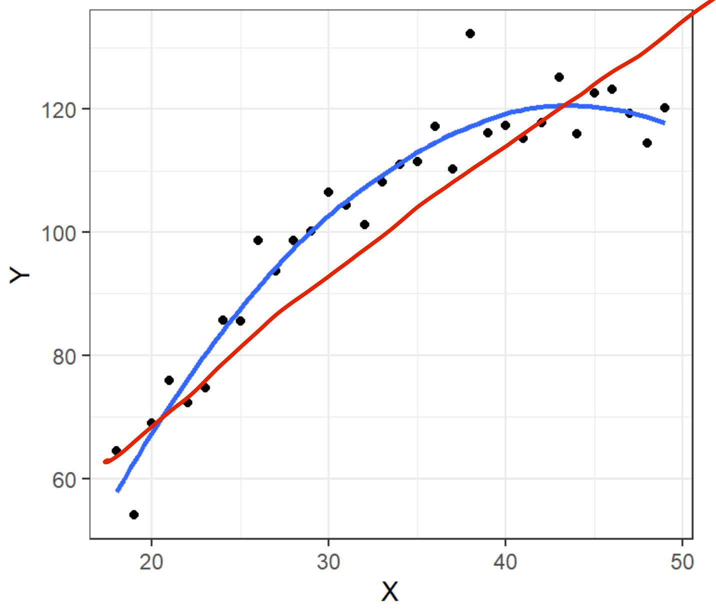
Regression vs. classification

How close do I predict to 0 or 1?

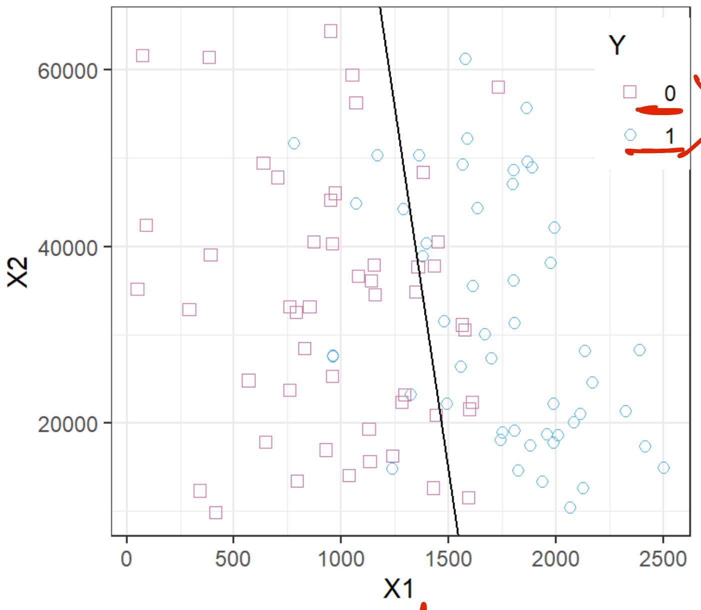
Regression

$Y = \beta_0 + \beta_1 X + \varepsilon$

Continuous



Classification



Reflect types

Cat / Dog.

• Classify as 0 or 1

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

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

– Can we do classification problems?

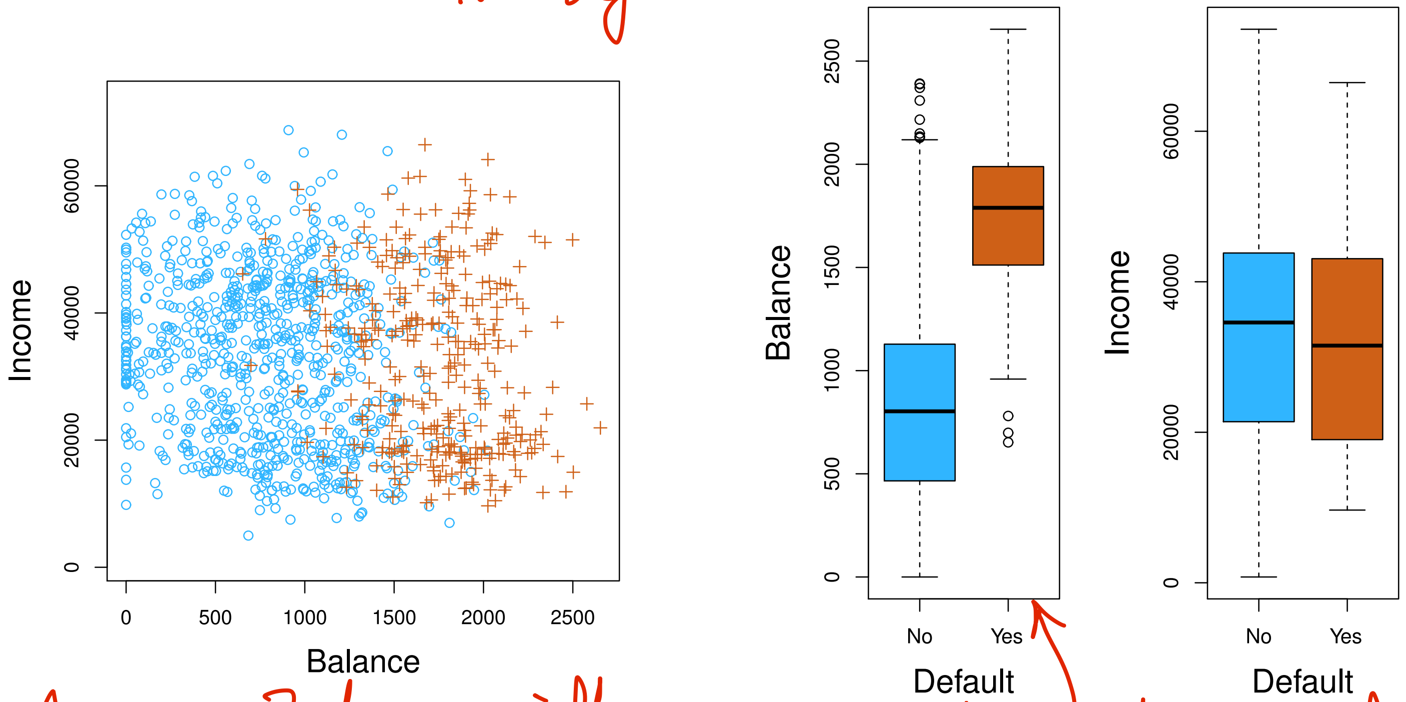
Some examples of classification problems

- 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`)

+ default

o Not default.



- Appears Balance influences default probs significantly

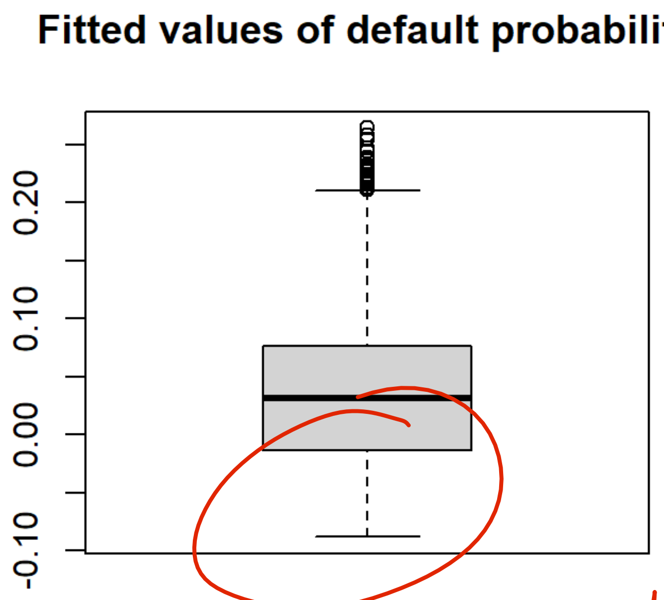
- `default` (Y) is a binary variable (yes/no or 0/1)
- Annual `income` (X_1) and credit card `balance` (X_2) may be continuous predictors

Example: Predicting defaults - Discussion

Simple linear regression on `Default` data:

▶ Show code for Figure

P(Default) ? — $Y = \beta_0 + \beta_1 \cdot \text{Balance} + \beta_2 \cdot \text{Income}$



Negative probabilities make no sense!

What do you observe?

$Y = f(X) + \varepsilon$

Classification problems

- Coding in the binary case is simple

Default   Not default

$$Y \in \{0, 1\} \Leftrightarrow Y \in \{\bullet, \bullet\}$$

Cat   Dog

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

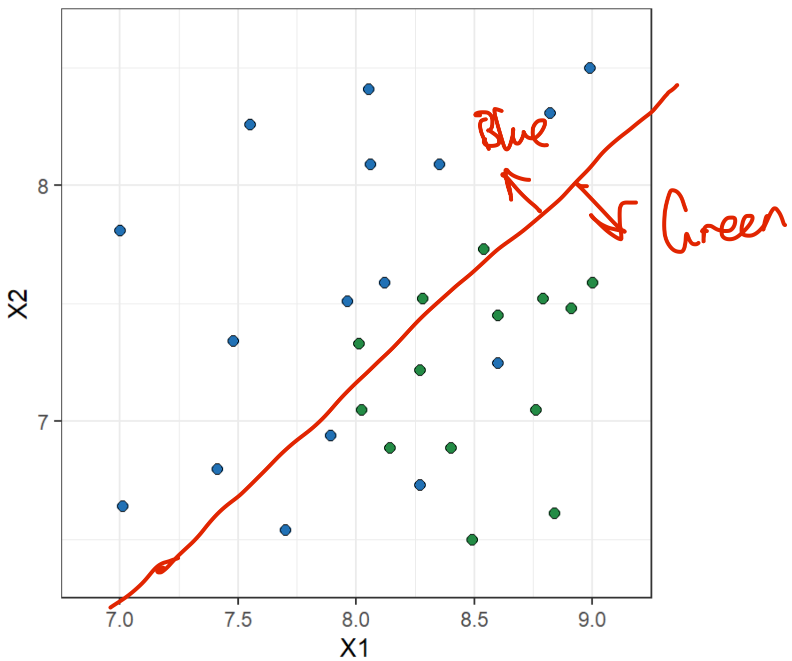
  1. Estimate the probability $\mathbb{P}(Y = 1|X) \in \{0, 1\}$

$$f(X) \rightarrow \bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet$$

  2. Classify observation

$$f(X) \rightarrow \hat{Y} \in \{\bullet, \bullet\}$$

Give a prediction on Y being blue or green.



Logistic regression

Logistic regression

Extend linear regression to model binary categorical variables

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

Log-odds is linear in X.

Before:

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

Response ^{Y} is linear in X.

Principles of Logistic Regression

- The output is binary $Y \in \{1, 0\}$ Code your response to be {0, 1} e.g. Cat = 0, dog = 1
- Each case's Y variable has a probability between 0 and 1 that depends on the values of the predictors X such that

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

• Might use 1 to represent success, 0 failure

- Probability can be restated as odds

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

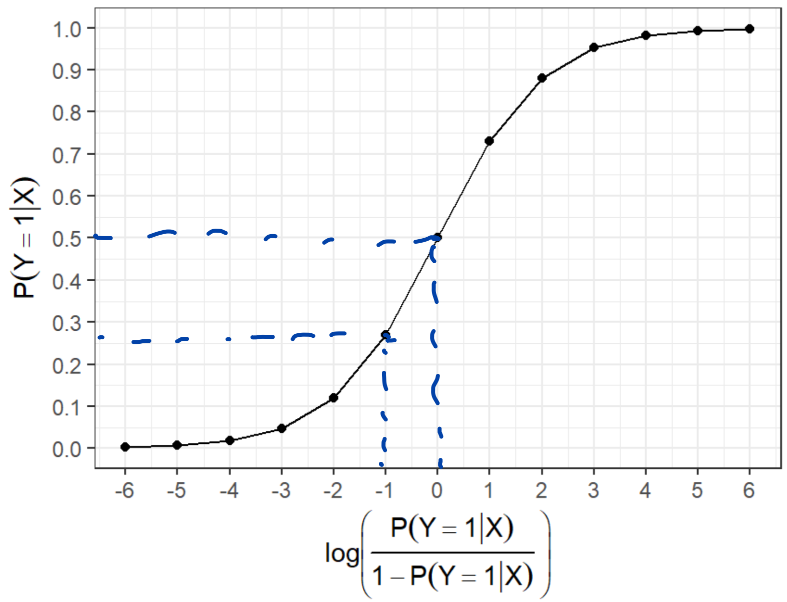
- Odds are a measure of relative probabilities

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

1-1 correspondence



$(0,1) \longrightarrow (-\infty, \infty)$

(Response) for standard lin Reg.

Logistic regression

Assume log-odds are linear in X

- Perform regression on log-odds

$$\ln\left(\frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$$

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

$$\mathbb{P}(Y=1|X)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_pX_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_pX_p}}$$

Interpretation of coefficients

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

(i,j) in X.

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

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 β_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 β_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

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

– If we assume log-odds are linear in X, then this is the predicted probability.

where x_i denotes the i'th row of X.

- We maximise the log-likelihood below

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

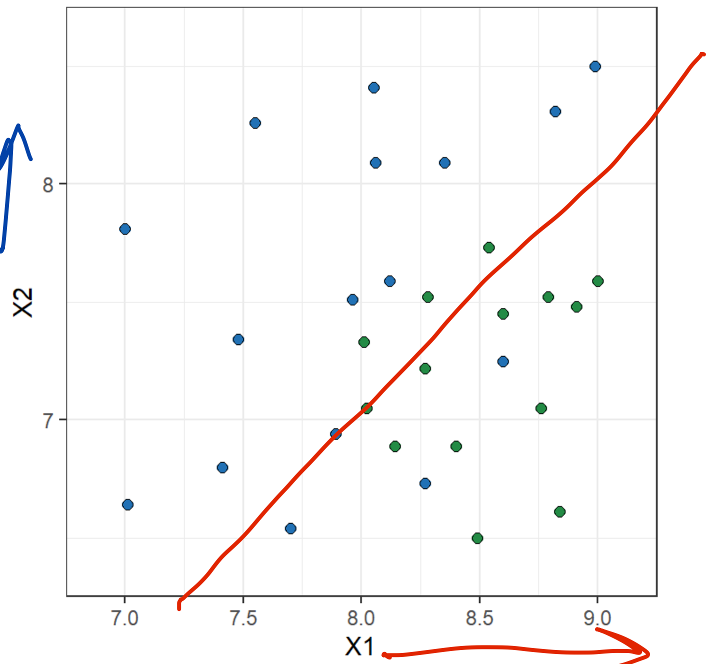
We take partials w.r.t. to each β_j and set to 0. Needs numerical approximation.

Toy example: Logistic Regression

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

- 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

More blue

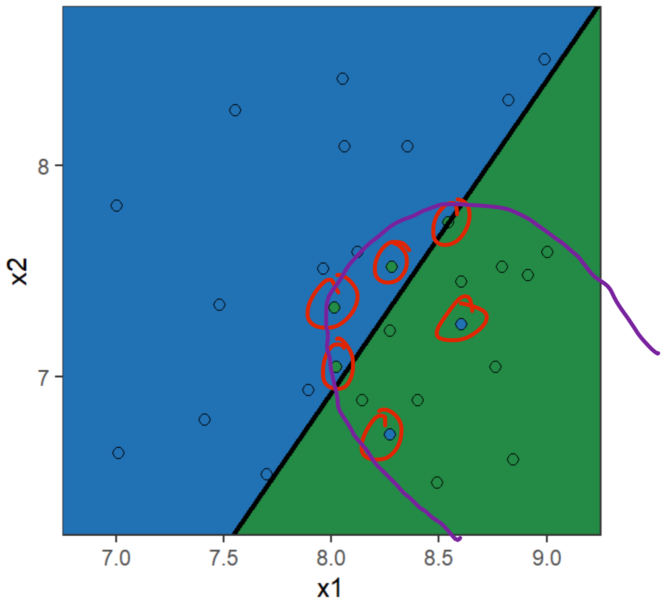


More green.

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



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

  AIC, Forward step, t-tests / z-tests
- 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

  Multinomial regression

Example: Predicting defaults

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

Response

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

Coefficients:
            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 ***
---
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
```

Z-test because Wald test from MLE.

If you are a student, you are more likely to default.

categories Type

animal Dog

Example: Predicting defaults

```r
glmAll <- glm(default ~ balance + income + student, family = binomial(), data = ISLR2::Default)
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      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  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
```

Negative coffecient

- Confounding variables

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
`student = Yes`	+ 0.4049	0.1150	3.52	0.0004

– student was a proxy for Balance

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

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

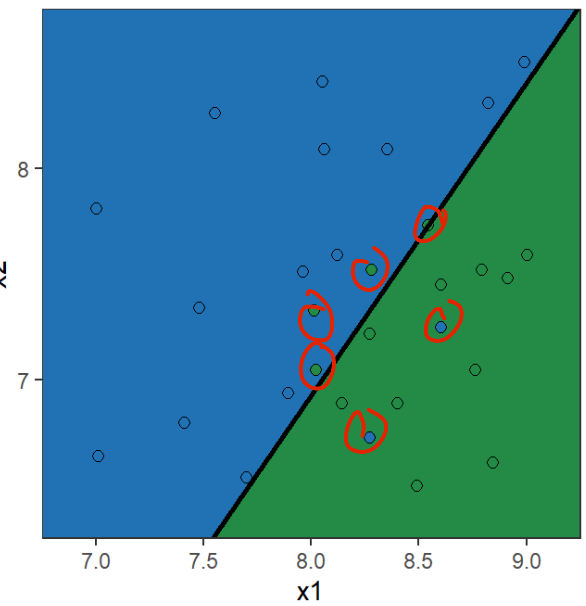
Assessing accuracy in classification problems

- We assess model accuracy using the error rate

$$\text{error rate} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{I(y_i \neq \hat{y}_i)}_{\text{Count up how many incorrect predictions}}$$

- In our toy example with a 50% threshold

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



The error rate we most care about is the test error rate

P(Y=1 | X)   0.01   0.49   0.75

0   0   1

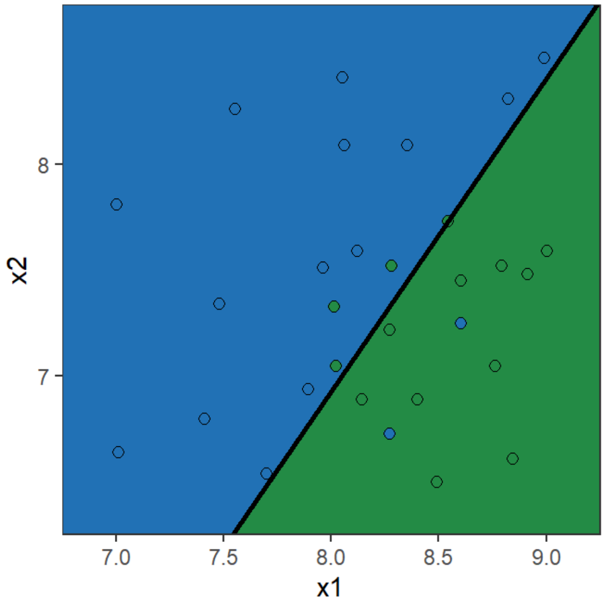
0   1   1

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 $= \frac{4}{14} = 0.286$



What if threshold was lower?

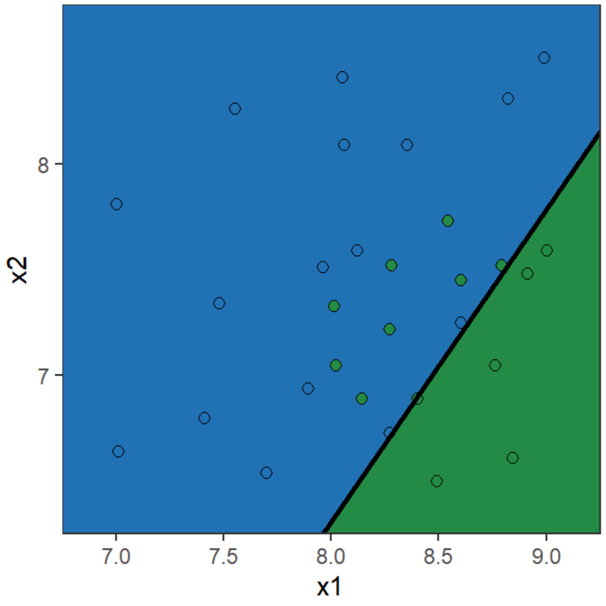
More blue (↑) or less blue

Confusion matrix: Toy example (15% Threshold)

- Confusion matrix

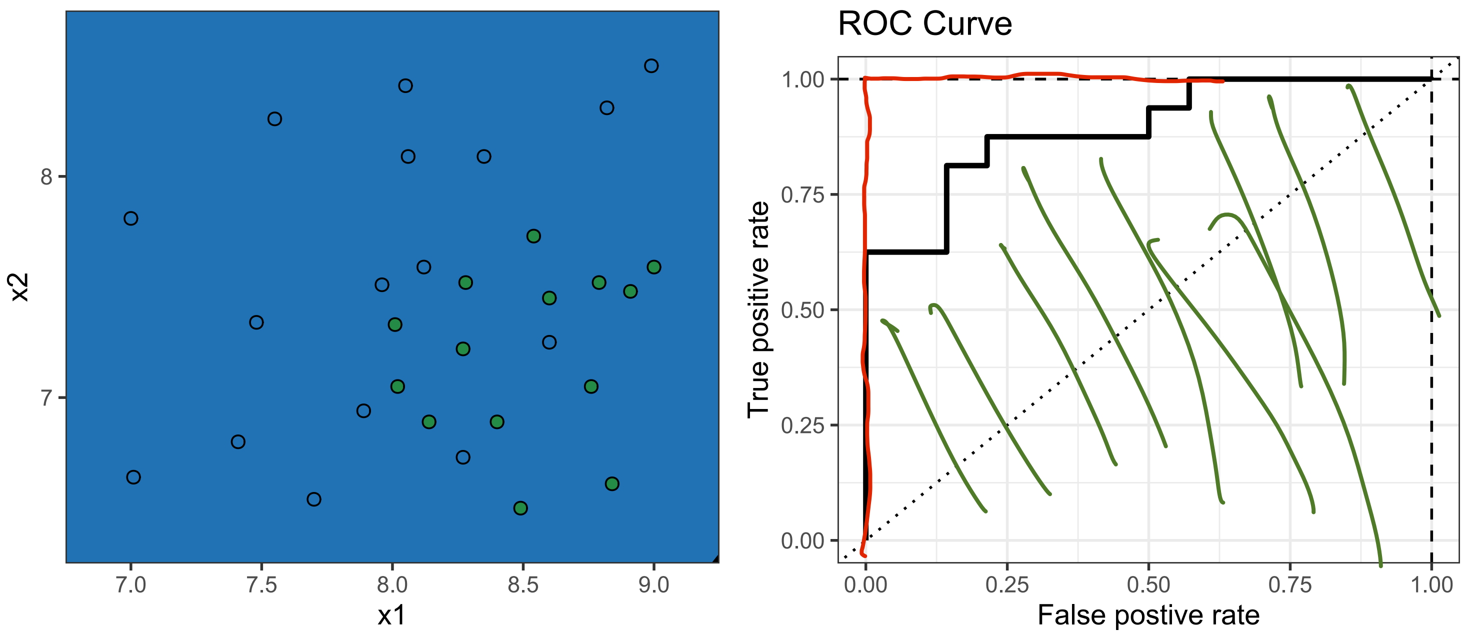
	$Y=0$	$Y=1$	**Total**
$\hat{Y}=0$	6	0	6
$\hat{Y}=1$	8	16	24
Total	14	16	30

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



ROC Curve and AUC: Toy example

Area under (ROC) Curve



- 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

Model with AUC closer to 1

Poisson regression

Poisson regression - Motiviation

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.

- Modelling how many seagulls try to steal my food at the beach

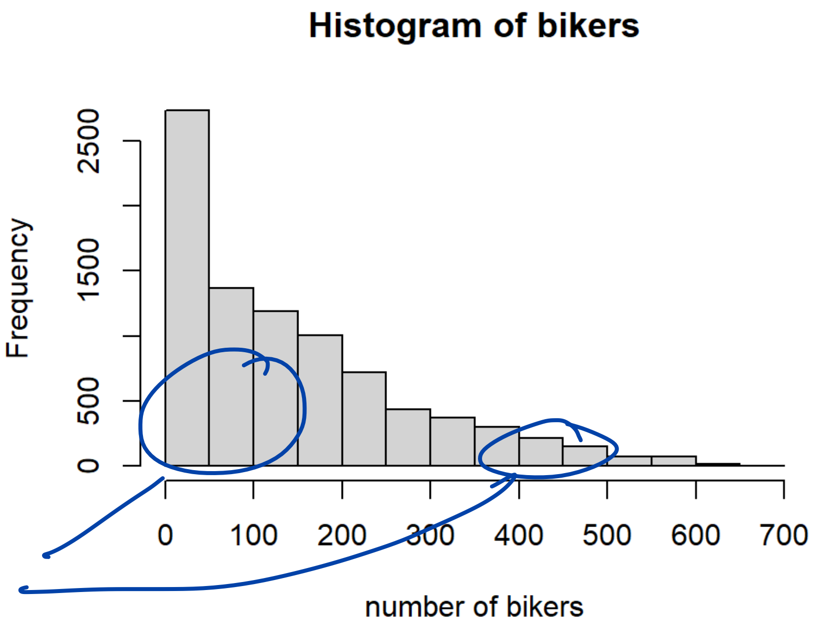
The Bikeshare dataset

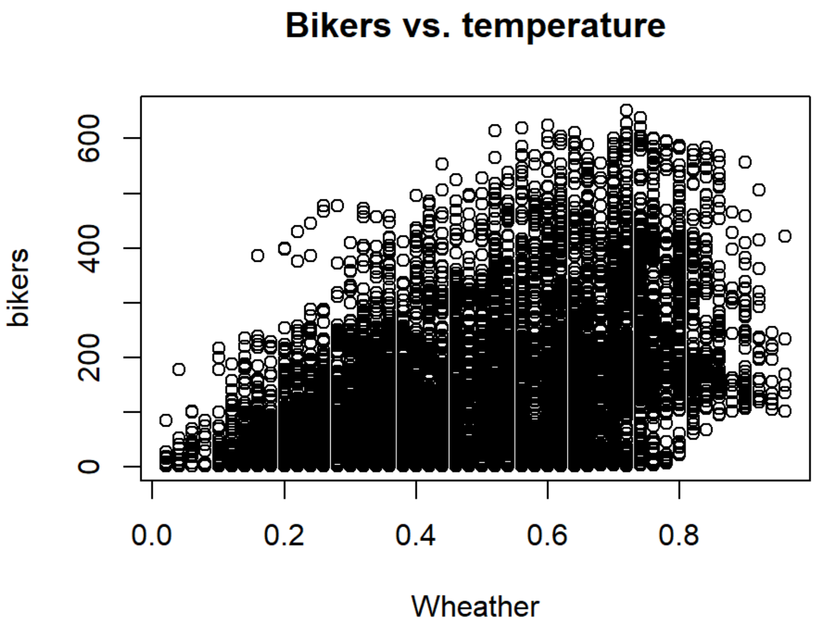
```r
str(ISLR2::Bikeshare)
```

```
'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 ...
```

- Count how many rental bikes are hired

The Bikeshare dataset - Discussion





Same variance assumption is concerning

Standardised temperature

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

Why not use muliple linear regression?

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

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

$$\log(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

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

log(0) does not exist

- Yes you can do ad-hoc solutions.
  - But why not just use a more appropriate model?

Poisson regression

we expect $Y \sim \text{Poi}(e^{100})$

$E[Y] = e^{100}$

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

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

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

$$\log(\lambda(X_1, \ldots, X_p)) = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}_{\text{Suppose this is } +100, \text{ then}}$$

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

$$\mathcal{L}(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \underbrace{\frac{e^{\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}}_{\text{pdf of Poisson}} \quad \text{with} \quad \lambda(x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{p1}$$

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^{β_j}.
- Mean-variance relationship: $\mathbb{E}[Y] = \mathrm{Var}(Y) = \lambda$ 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 AIC, Steps, z-test
- As a result, most of the modelling limitations of linear regression (e.g. collinearity) carry over as well / Some linear relationship

$$\log(\lambda) = \beta_0 + \cdots + \beta_j(X_j + 1) + \cdots + \beta_p X_p$$

$$\lambda = e^{\beta_j} \cdot e^{\beta_0 + \cdots + \beta_p X_p}$$

Poisson regression - Bikeshare dataset

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

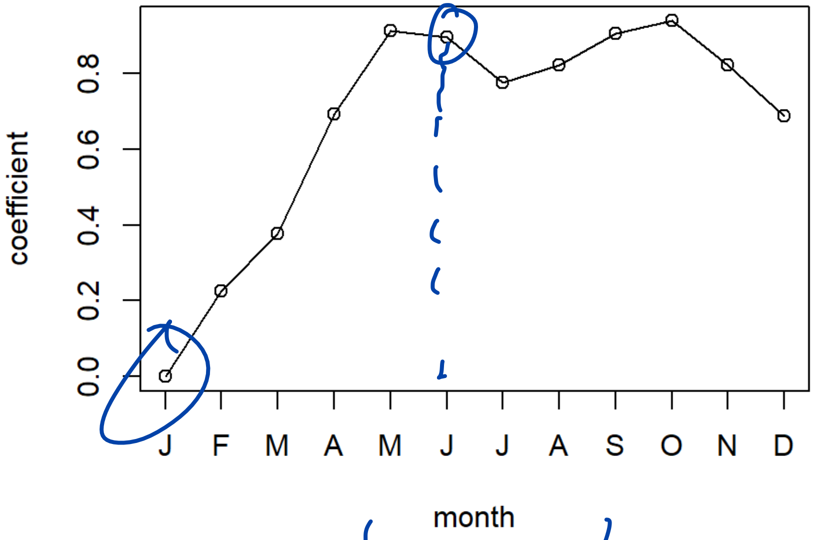
```
Call:
glm(formula = bikers ~ workingday + temp + weathersit + mnth +
    hr, family = poisson(), data = ISLR2::Bikeshare)

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 ***
```

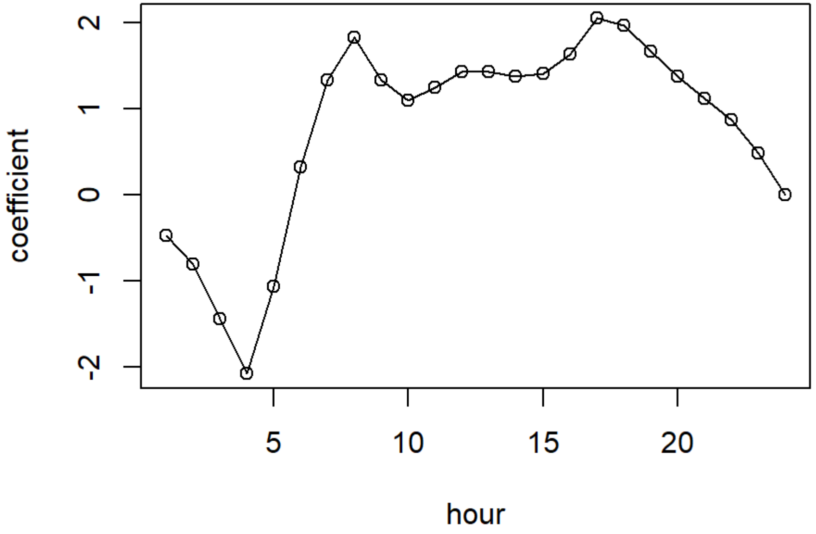
Poisson regression - Bikeshare dataset

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

$\approx 0.9 \rightarrow e^{0.9} = 2.5$



Summer



log mean increases by 0.9 in June

$\rightarrow$ mean increases by a factor of $e^{0.9} \approx 2.5$

Generalised linear models

Generalised linear models

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