COMP6237 – Logistic Regression

Shoaib Ehsan

s.ehsan@soton.ac.uk

Lecture slides available here:

http://comp6237.ecs.soton.ac.uk/

(Thanks to Jason Noble and Cosma Shalizi whose lecture materials I used to prepare)

COMP6237: Logistic Regression

- Outline:
  - Introduction
  - Basic ideas of logistic regression
  - Logistic regression using R
  - Some underlying maths and MLE
  - The multinomial case
  - How to deal with non-linear data
    - Model reduction and AIC
  - How to deal with dependent data
  - Summary

Introduction

- Previous lecture: Linear regression
  - tried to predict a continuous variable from variation in another continuous variable (e.g., basketball ability from height)
- Here: Logistic regression
  - Try to predict results of a binary (or categorical) outcome variable Y from a predictor variable X
  - This is a classification problem: classify X as belonging to one of two classes
  - Occurs quite often in science … e.g., medical trials (will a patient live or die dependent on medication?)

Dependent variable Y

Predictor Variables X

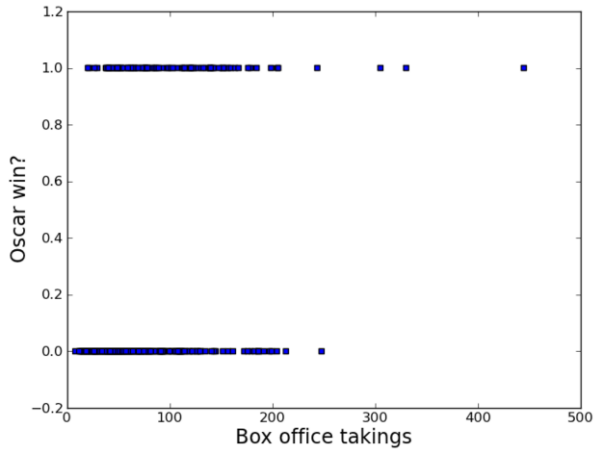
Fate	Treatment	Age(2005)
alive	placebo	45
dead	drug A	61
dead	placebo	29
alive	drug B	33
dead	placebo	70
dead	drug A	61
dead	placebo	44
alive	drug B	50

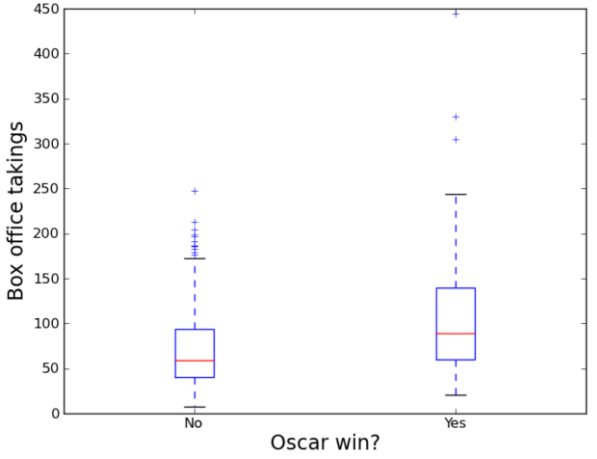
The Oscars Example

- A fictional data set that looks at what it takes for a movie to win an Oscar
- Outcome variable: Oscar win, yes or no?
- Predictor variables:
  - Box office takings in millions of dollars
  - Budget in millions of dollars
  - Country of origin: US, UK, Europe, India, other
  - Critical reception (scores 0 … 100)
  - Length of film in minutes
  - This (fictitious) data set is available here:
  - https://www.southampton.ac.uk/~mb1a10/stats/filmData.txt

Predicting Oscar Success

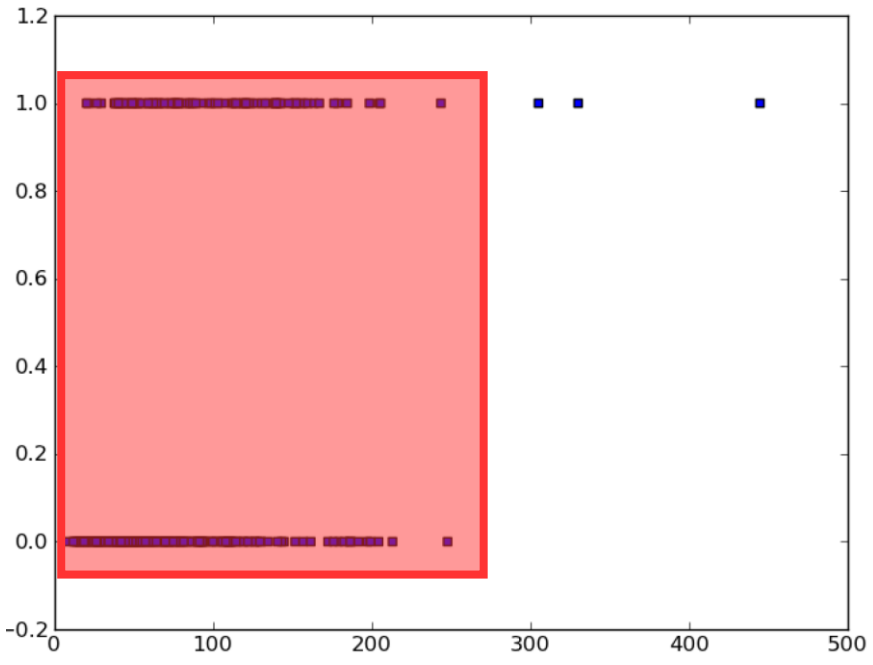
- Let's start simple and look at only one of the predictor variables
- Do big box office takings make Oscar success more likely?
- Could use same techniques as below to look at budget size, film length, etc.

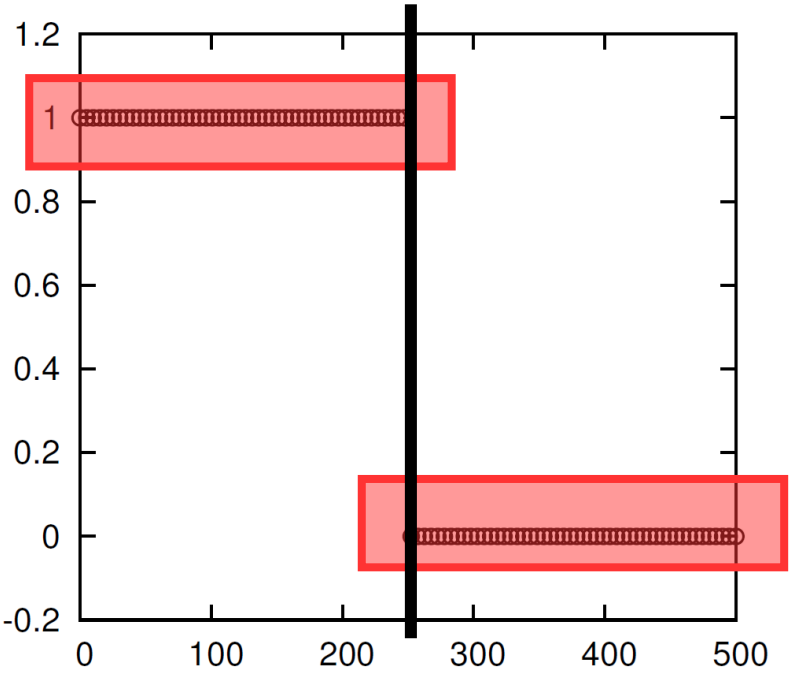




Introduction (1)

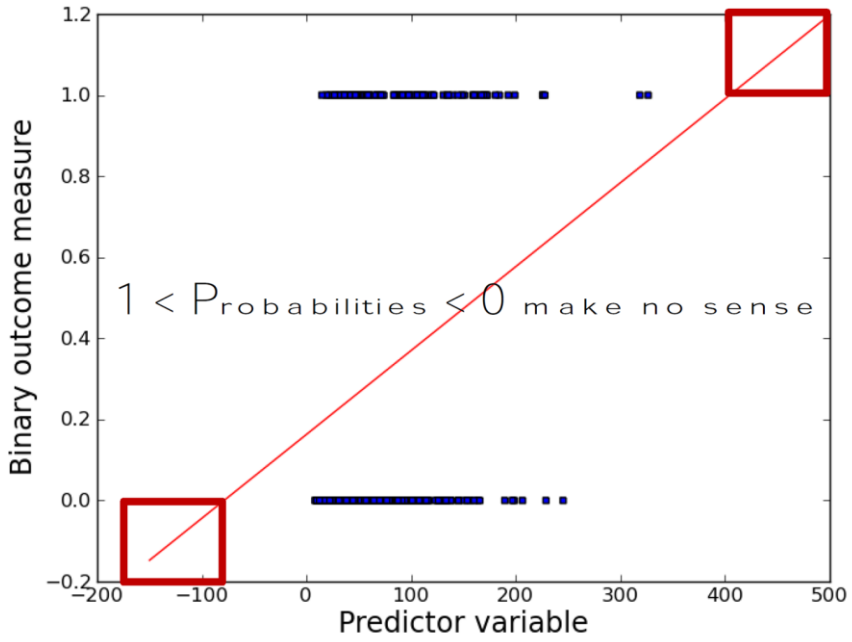
- Could use a linear classifier …
  - But this does not give us probabilities, which are desirable if
    - We want to handle different error costs between classes
    - We need some indication of confidence
    - Perfect classification is not possible





Introduction (2)

- Naive approach:
  - Code binary variable as 0 or 1, do linear regression and interpret outcomes as probabilities …
  - Avoid range problem by assuming
    - P(x)=0 x<0
    - P(x)=1 x>1 ?
  - Problems with saturation: once we reach boundaries, we cannot discriminate any more, model becomes insensitive to predictor



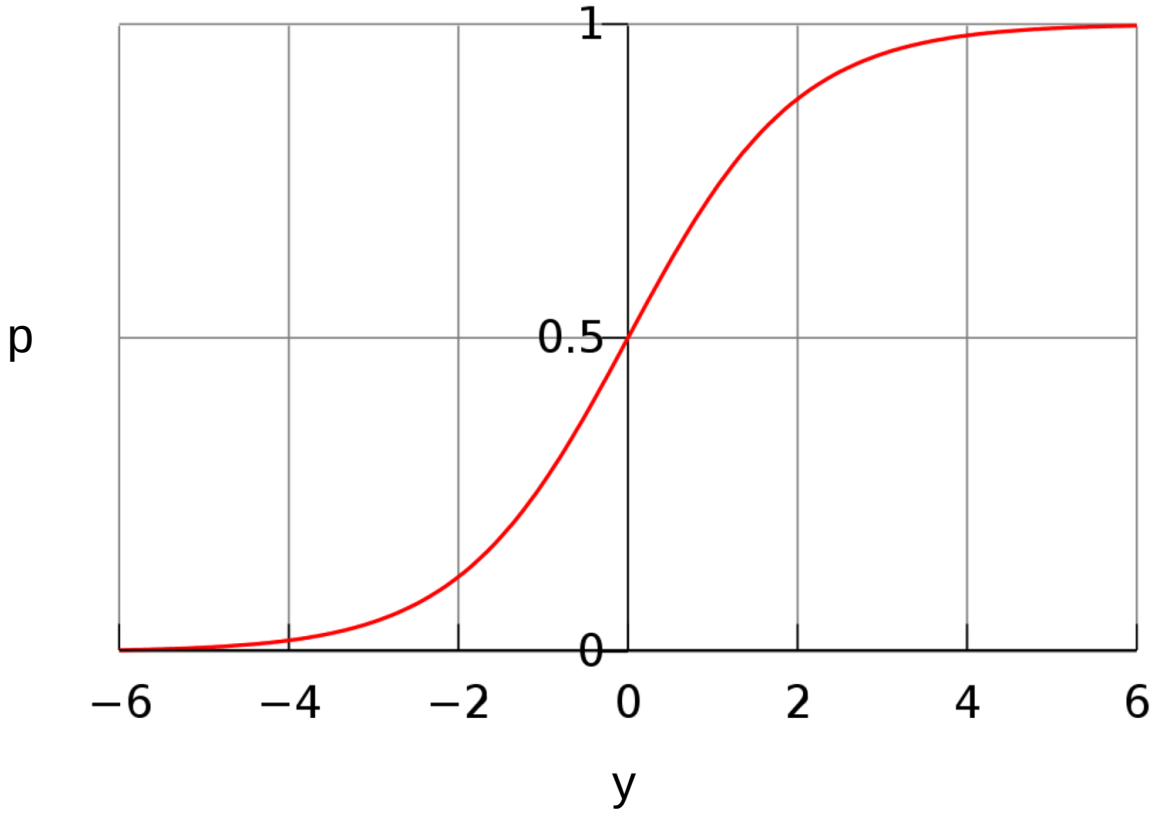
The Idea

- Can transform predictor variable to something we can do linear regression on!
- Want to find a probability Pr(Y=1|X=x)=p(x) for Y to be in class 1 (or 0)
  - Cannot do linear regression on p directly because of range issues
  - What about doing regression with log p(x) linear in x? → log's only unbounded in one direction (whereas linear functions are not)
  - Easiest modification of log p that has an unbounded range is the logistic transformation $\log p(\mathbf{x})/\big(1-p(\mathbf{x})\big)$
    - p/(1-p) is also called “odds”

An Aside: From Probability to Odds

- Odds are often used in gambling/betting
  - Odds = “probability of event”/”probability of not event”
- E.g.:
  - “9 to 1 against” → p=0.1
  - “even odds” → p=0.5
  - “3 to 1 on” → p=0.75
- Not scientific parlance, don't write your work up like this.

The Logistic Function



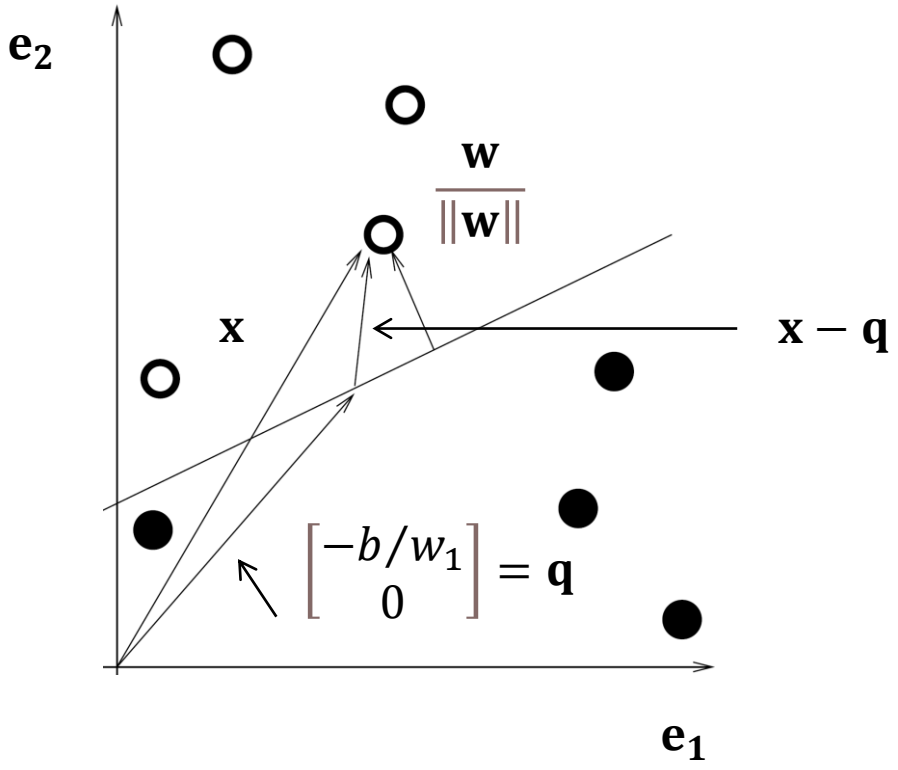
$$y = \log \frac{p}{(1-p)}$$

Is the easiest transform to solve our range problems.

Logistic Regression

- Formally, the logistic regression model is

$$\log \frac{p(\mathbf{x})}{(1-p(\mathbf{x}))} = b + \mathbf{wx} \qquad p(\mathbf{x}) = \frac{1}{1+\exp(-(b+\mathbf{wx}))}$$



1.) May want to say Y=1 iff p(x)>=1/2, which is iff b+wx>=0, i.e., b+wx=0 gives a decision boundary

2.) Distance from decision boundary is

$$D(\mathbf{x}) = (\mathbf{x}-\mathbf{q})\,\mathbf{w}/\|\mathbf{w}\|$$
$$= \mathbf{x}\,\mathbf{w}/\|\mathbf{w}\| + b/\|\mathbf{w}\|$$

so logistic regression says that probs depend on that distance

3.) Boltzmann weights ...

How to Run Logistic Regression in R

- Will use the Oscar example here, variables of interest are Oscar and BoxOffice
- Build the regression model in R and then use summary command to see information:

```
boxOfficeModel = glm( Oscar ~ BoxOffice,
family=binomial (link="logit"))
```

```
summary(boxOfficeModel)
```

family is binomial, will understand this later in the lecture

generalised linear model

R Output

p value; likelihood for the data to arise if there Was no relationship between the variables

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6432  -0.8316  -0.6997   1.2380   1.8546

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.750349   0.256883  -6.814 9.50e-12 ***
BoxOffice    0.011306   0.002507   4.510 6.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logit increment per $million in Box Office takings

logit score for a film that made $0

Making Sense of the R Output

- What is the chance of a film with Box Office takings of $50 million to win an Oscar?
  - Logit score = -1.75+0.011*50=-1.2
  - p = 1/(1+exp(-logit)) =0.23
  - → the model predicts that such a film has a 23% chance to win an Oscar

Using the Other Variables

- What about the other predictor variables?
  - Could look at each variable separately with a logistic regression model and check if it has any explanatory value …
  - Better include them jointly to fit all the variables in the same model:

```
fullModel = glm( Oscar ~ BoxOffice + Budget + Country +
Critics + Length, family=binomial (link="logit"))
```

  - (could handle this for normal linear regression in the same way)

Likelihood Functions for Log Regression

- Can fit the model using MLE
- For each of n training points we have
  - A vector of features $\mathbf{x_i}$
  - An observed class $y_i = 0,1$
  - Probability of class yi=1 is p(xi) and yi=0 is 1-p(xi)
- Likelihood function

$$L(Y_i; \mathbf{w}, b) = \prod_{i=1}^{n} p\,(\mathbf{x_i})^{y_i}\big(1 - p(\mathbf{x_i})\big)^{1-y_i}$$

$$l(\mathbf{w}, b) = \log L(Y_i; \mathbf{w}, b) = \sum_{i=1}^{n} y_i \log p(\mathbf{x_i}) + (1 - y_i)\log\big(1 - p(\mathbf{x_i})\big)$$

$$= \sum_{i=1}^{n} \log\big(1 - p(\mathbf{x_i})\big) + \sum_{i=1}^{n} y_i \log \frac{p(\mathbf{x_i})}{1 - p(\mathbf{x_i})}$$

MLE for Logistic Regression

$$l(\mathbf{w}, b) = \sum_{i=1}^{n} \log\left(1 - p(\mathbf{x_i})\right) + \sum_{i=1}^{n} y_i \log \frac{p(\mathbf{x_i})}{1 - p(\mathbf{x_i})}$$

$$= \sum_{i=1}^{n} \log\left(1 - p(\mathbf{x_i})\right) + \sum_{i=1}^{n} y_i \left(b + \mathbf{x_i w}\right)$$

$$= \sum_{i=1}^{n} -\log\left(1 + \exp(b + \mathbf{x_i w})\right) + \sum_{i=1}^{n} y_i \left(b + \mathbf{x_i w}\right)$$

- To maximise … need to find

$$\partial l\left(\mathbf{w}, b\right) / \partial w_j = 0$$

$$\partial l\left(\mathbf{w}, b\right) / \partial w_j = -\sum_{i=1}^{n} \frac{\exp(b + \mathbf{w x_i})}{1 + \exp(b + \mathbf{w x_i})} x_{i,j} + \sum_{i=1}^{n} y_i \, x_{i,j}$$

$$= \sum_{i=1}^{n} \left(y_i - p(\mathbf{x_i}; \mathbf{w}, b)\right) x_{i,j}$$

- This is a transcendental equation, cannot solve analytically;
- use some numerical optimisation scheme, e.g., Newton's method.
- More details, cf. http://czep.net/stat/mlelr.pdf

The Multinomial Case

- What if we have more than one possible category?
- For each of n training points we have
  - A vector of features $\mathbf{x_i}$
  - An observed class $y_i = 1,2,\ldots,J$
  - Want to estimate probability that data point i belongs to class j $p_{i,j} = Pr\{Y_i = j\}$
- Each data point must belong to one class $\sum_{j=1}^{J} p_{i,j} = 1$
  - $\rightarrow$ have J-1 parameters
  - Typically nominate one of the response categories as reference mode and estimate ratios $p_{i,j}/p_{i,J}$

The Multinomial Case (2)

- Then make linear ansatz for logs

$$\eta_{i,j} = \log p_{i,j}/p_{i,J} = \alpha_j + x_i\beta_j$$

- Can write this in terms of the pij's:

$$p_{i,j} = p_{i,J}\exp\left(\alpha_j + x_i\beta_j\right)$$

$$\sum_{j=1}^{J} p_{i,j} = 1 \quad \longrightarrow \quad p_{i,J}\left(\sum_{j=1}^{J-1}\exp\left(\alpha_j + x_i\beta_j\right) + 1\right) = 1$$

$$\longrightarrow \qquad p_{i,j} = \frac{\exp\left(\alpha_j + x_i\beta_j\right)}{\sum_{j=1}^{J-1}\exp\left(\alpha_j + x_i\beta_j\right) + 1}$$

Also sometimes called the "softmax" function.

If we need to give a category we can take j for which pij is largest.

The Multinomial Case (2)

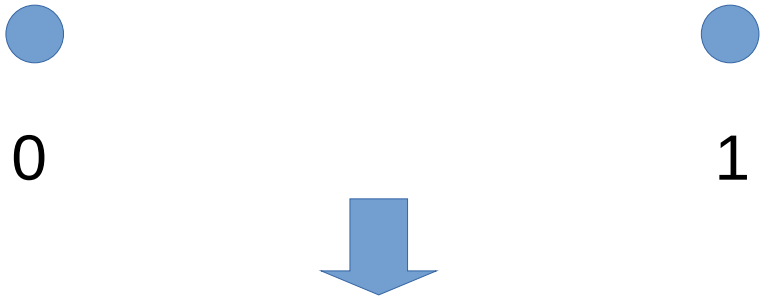
- MLE?
  - Introduce “Iverson brackets” [yi=j] = 1 if yi is in category j and zero otherwise (for better notation)

$$L(Y_i; \mathbf{w}, b) = \prod_{i=1}^{n} p_{i,1}^{[y_i=1]} p_{i,1}^{[y_i=2]} \dots p_{i,1}^{[y_i=J]}$$

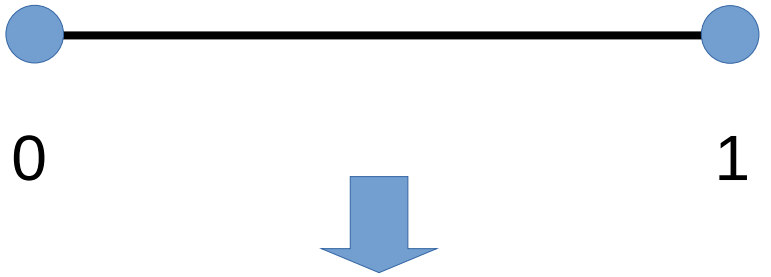
  - We can then proceed as before and pluck in the softmax function ans solve a numerical optimization problem to calculate max log L.
  - This is often done via stochastic gradient ascent

Review of our Strategy

Original Y

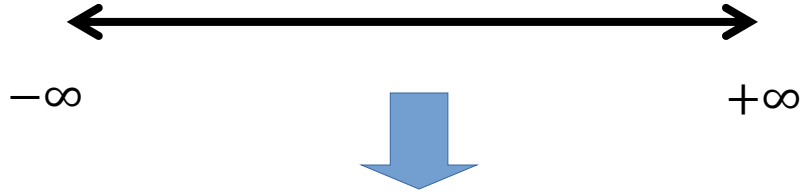


Y as a probability



Transformed

$g(Y) \in (-\infty, +\infty)$



Linear model for g $\quad g(Y) = mX + b$

Comments

- Logistic regression is a modelling choice, posit it, then check whether it works or has systematic flaws
- Reasons for using it:
  - Tradition
  - Often works surprisingly well as a classifier (But many simple techniques do that …)
  - Closely related to exponential family distributions (which e.g. arise out of maxent)
  - Is problematic if data can be linearly separated (if b,w perfectly separates linearly, so does cb,cw with c>0; so there is no parameter vector that maximises likelihood)

A Note on Generalised Linear Models

- Log regression is part of a family of generalised linear models (GLMs)
  - Conditional distribution of the response falls in some parametric family and parameters are set by a linear predictor
  - E.g.:
    - Ordinary least squares: response is Gaussian with mean equal to linear predictor and variance constant
    - Log regression: response is binomial with n equal to the number of data points with a given x and p given by the logistic function
  - Changing relationship between parameters and linear predictor is called changing the **link function**; in R this can be specified in glm – all fit with same numerical likelihood maximisation technique

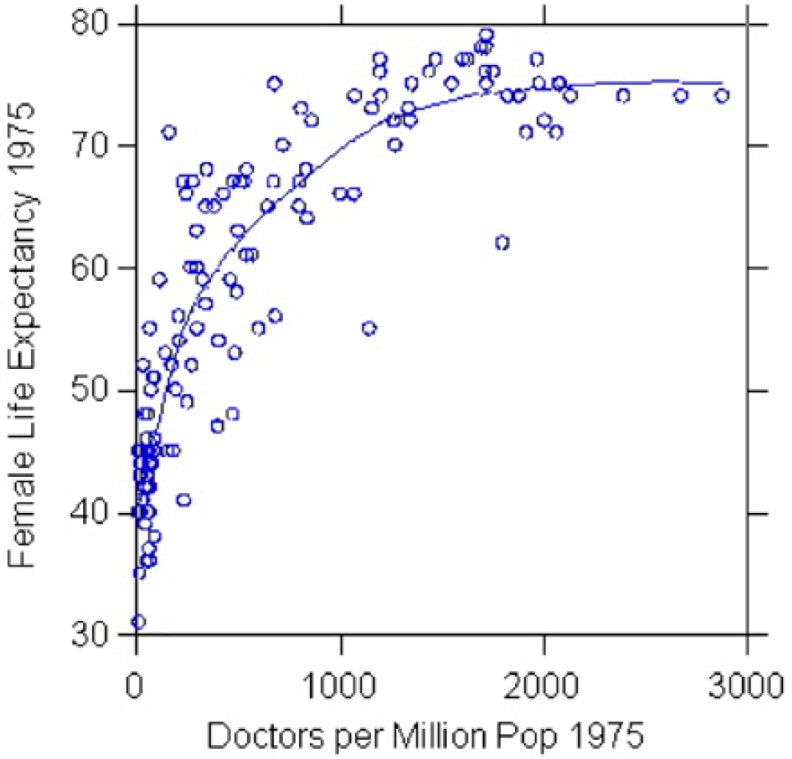
Summary Logistic Regression

- Use for categorical outcome variables
- Probabilistic interpretation → logistic transform
- Underlying stochastic model: binomial
- Idea of the maths
- Link to GLMs

Non-Linear Data

Problems with Linear Regression

- Linear regression assumes a linear relationship between outcome variables and predictors; can we deal with non-linear data like those? →
- (→ transformations of variables,
- fitting polynomials)

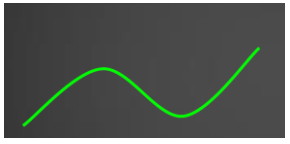


- Independence assumption, linear regression assumes that effects of predictors are independent of each other (→ interaction terms)

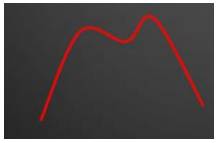
Non-linear Data: Fitting Polynomials

- Fit a polynomial instead of a straight line
- Same idea as linear regression, just turning one predictor (X) into several (X^2,X^3, …)
- Allows to deal with obvious non-linearity without having to specify in advance what transformation is
- Intuition for degree of polynomial should come from shape of relationship on scatterplot

parabolic
(2nd order)



cubic
(3rd order)



quartic
(4th order)

R practicalities

- 1st way:

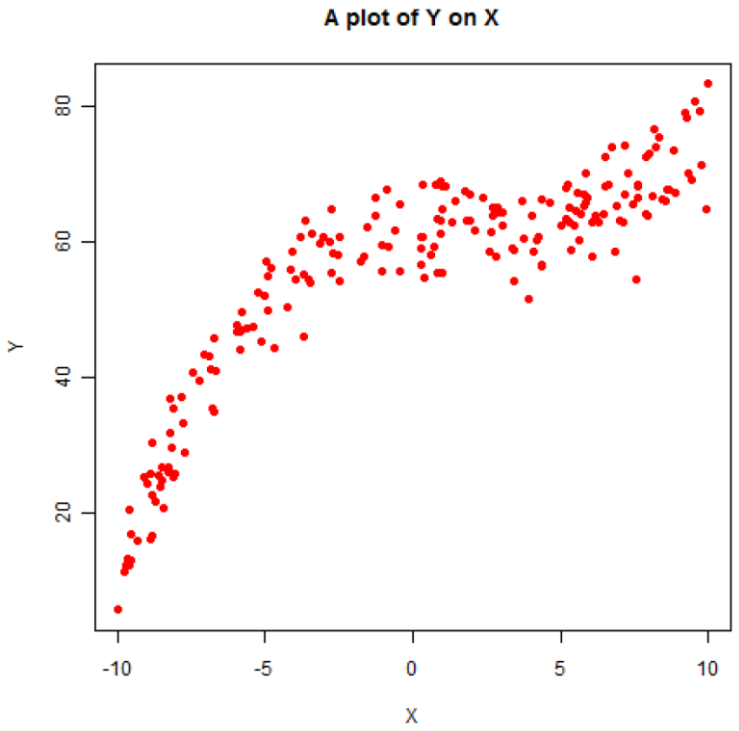
```
model = lm(y~x + I(x^2))
```

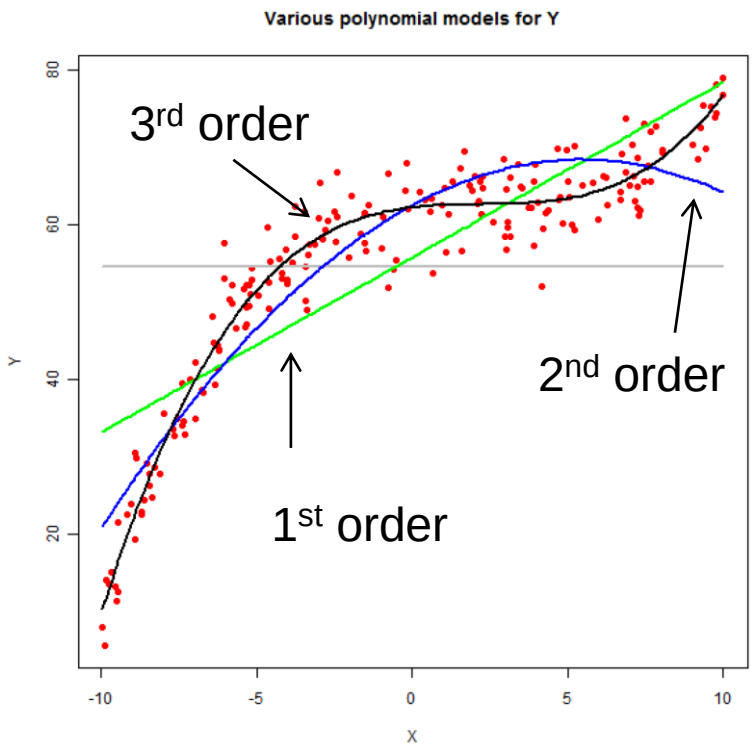
- (easy to interpret)
- 2nd way:

```
model = lm (y ~ poly(x,2) )
```

  - Uses orthogonal polynomials which are more efficient for fitting, but resulting coefficients not straightforward to interpret

Example





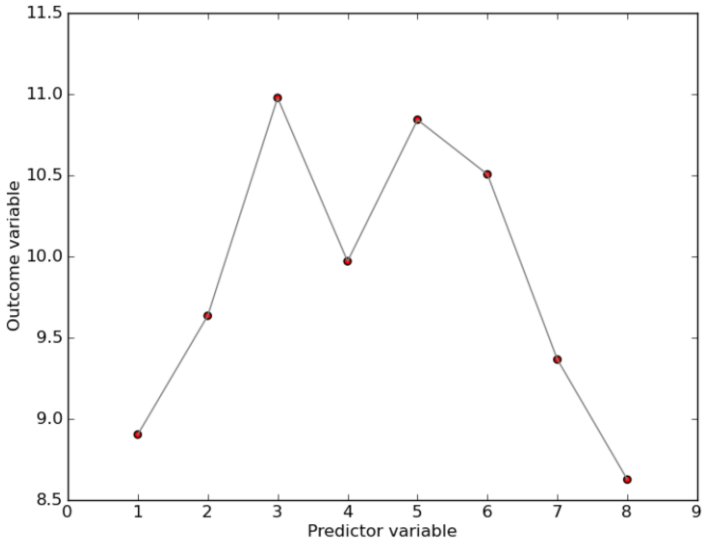

```
m0 = lm ( Y ~ 1 )
m1 = lm ( Y ~ X )
m2 = lm ( Y ~ X + I(X^2) )
m3 = lm ( Y ~ X + I(X^2) + I(X^3) )
m4 = lm ( Y ~ X + I(X^2) + I(X^3) + I(X^4) )
```

Problem: Higher order models will always do better …end up with matching number of data points?

Occam's Razor

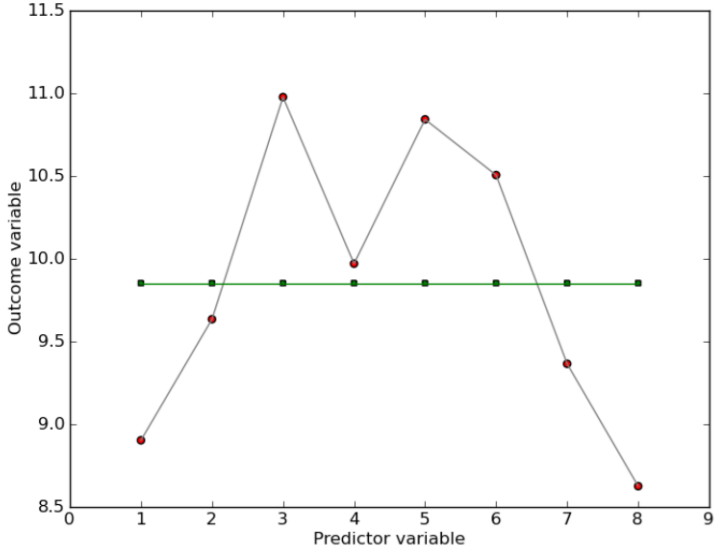
- William of Occam (1288-1348)
- All else being equal, the simplest explanation is the best one.
- In statistics this means:
  - A model with fewer parameters is to be preferred to one with more parameters.
  - … but this needs to be weighed against a model's ability to predict.

Over-fitting



- N-1 predictors enough to replicate data. But this is an absurd model without explanatory power.
- (if you ask me the height of somebody I ask for information about his gender, DOB, occupation, address, parents, neighbours, friends, ...)

Under-fitting



- Lowest number of predictors we can use is to just explain all the variation by the mean.
- (no variation explained, if you ask me the height of somebody in the UK, I say 1.67m ...)

The Machine learning way – Regularization

- Two reasons to do this:
  - Problem ill-posed (more variables than observations)
  - Solution does not generalize well
- Redefine optimization problem: add regularization term to residuals

$$E = \sum_{i=1}^{N} (y_i - \tilde{\mathbf{w}}\tilde{\mathbf{x}}_i)^2 = \left\|\mathbf{y} - \tilde{X}\tilde{\mathbf{w}}\right\|^2 \rightarrow \left\|\mathbf{y} - \tilde{X}\tilde{\mathbf{w}}\right\|^2 + \lambda\|\tilde{\mathbf{w}}\|_{1/2}$$

- Popular forms of norm:
  - Ridge (Tikhonov): L^2 norm (sum squares)
    - → force some elements of w to be small
  - Lasso: L^1 norm (sum absolute values)
    - → typically some elements of w can be forced to zero

Stats Methods

The old Ways of doing it

- Do model reduction through a series of F-tests, asking if models with more parameters explain significantly more variation in the outcome variable.
- → “Step-wise” model reduction.
  - Either start with full model and reduce or start with simplest model and build up complexity.
  - Caveats:
    - Complicated ...
    - Model one ended up could depend on the starting point ...

A better Way?

- Kullback-Leibler divergence:
  - Measure of the informational distance between two probability distributions
  - K-L distance between a real-world distribution and a model distribution tells us how much information is lost if we describe the real-world distribution with the model distribution.
  - A good idea to obtain a good model is to minimise the K-L distance to the real-world.

Akaike's Information Criterion

- If we had a true distribution F and two models G1 and G2 we could figure out which model we prefer by calculating K-L distances F-G1 and F-G2;don't know F in real world cases, but can estimate F-G1 and F-G2 from the data
- AIC is an estimator for the K-L divergence
- Akaike's information criterion:

$$AIC = 2K - 2\log(L)$$

AIC score: the lower the better.

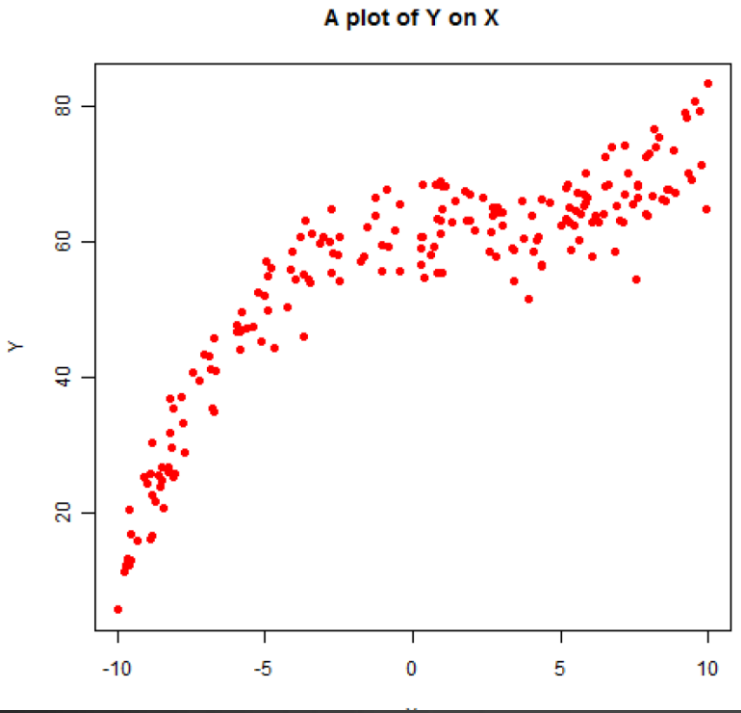
Number of predictors
– punishes models with many predictors

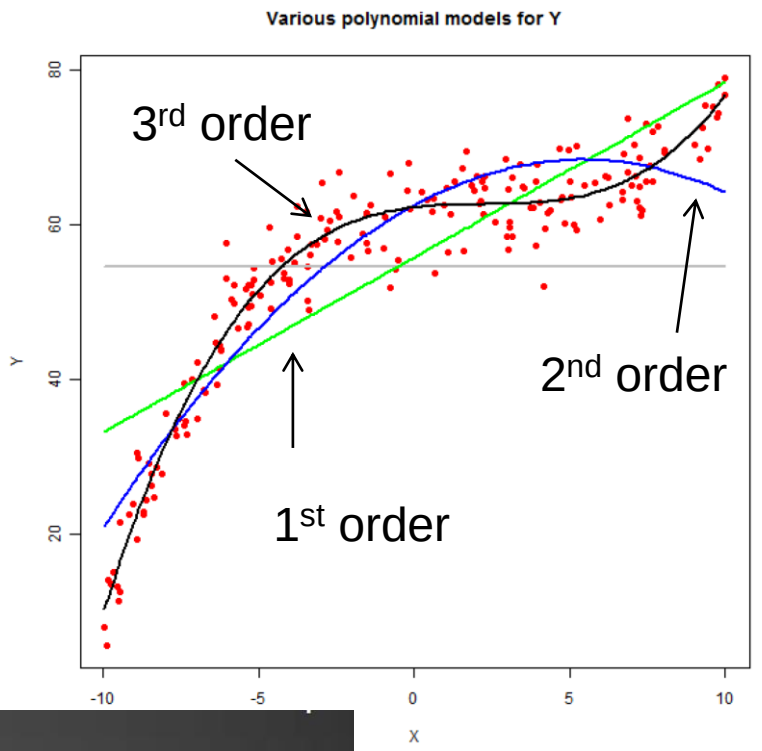
Maximised likelihood value
(using these predictors)
– rewards fit of model to data

R practicalities

- Very easy to use in R. Suppose we have (regression) models m1, m2, m3, …
- invoke: AIC (m1,m2,m3,…)
  - This will list AIC values of all the models, simply pick the lowest AIC score.
- drop1 (model)
  - Is also quite useful. I returns AIC scores for dropping each predictor in turn.

Back to the Example






```
m0 = lm ( Y ~ 1 )
m1 = lm ( Y ~ X )
m2 = lm ( Y ~ X + I(X^2) )
m3 = lm ( Y ~ X + I(X^2) + I(X^3) )
m4 = lm ( Y ~ X + I(X^2) +
```

```
AIC(m0,m1,m2,m3,m4)

    df       AIC
m0   2 1691.918
m1   3 1435.907
m2   4 1278.603
m3   5 1110.864
m4   6 1111.103
```

m3, the cubic model, has the lowest AIC score

Limitations of AIC

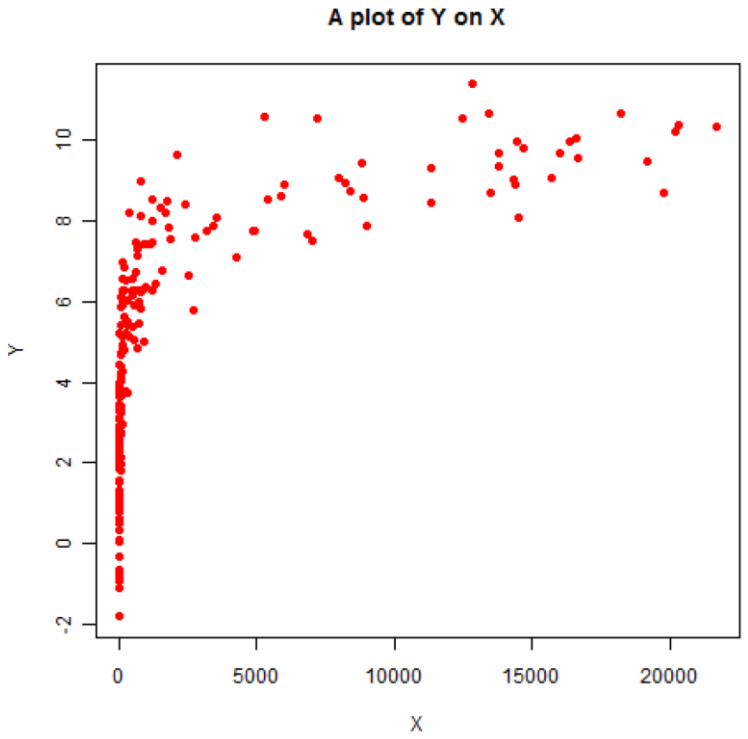
- AIC is an asymptotic approximation
  - Number of params must be small compared to number of data points
- True model must be in the parameterised family
- Every model f in our family must map to a unique conditional probability distribution p(data|f)
- Likelihood L(f) function must be twice differentiable
- Use it:
  - Linear regression, generalised linear models, constant bin width histogram estimation
- Don't use it for
  - Multi-layer neural networks (uniqueness of p)
  - Mixture models
  - The uniform distribution (differentiability)
- More details, e.g.:
- http://www.csse.monash.edu.au/~dschmidt/ModelSectionTutorial1_SchmidtMakalic_2008.pdf

A more pragmatic strategy ...

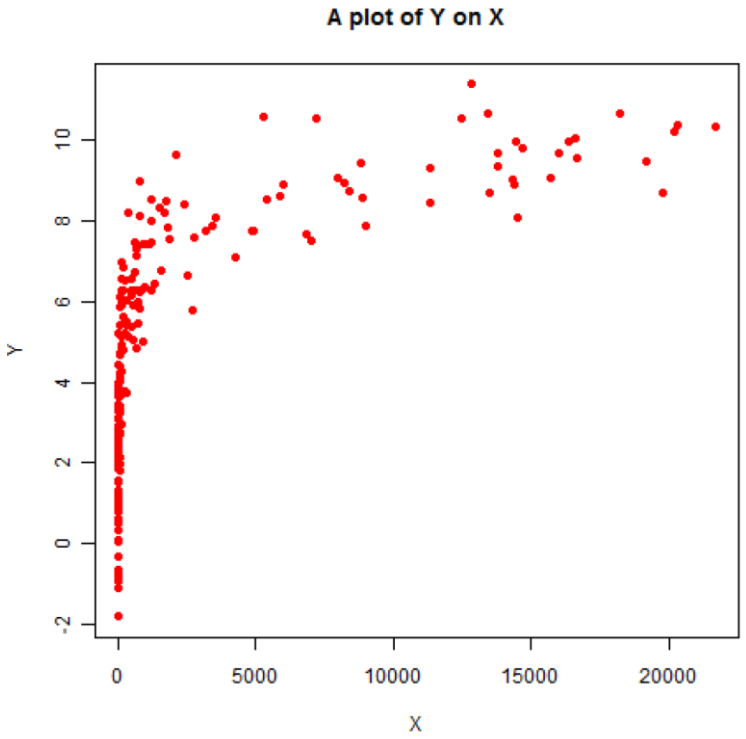
Nonlinear Data, Transformations

- Don't work with the raw data for X but with some transformation f(X) which will hopefully be closer to linear
  - This is a bit like you were finding you were measuring the wrong thing, e.g., surface area might be a more direct measure of wind exposure than height
- How to know which transformation to use?
  - Sometimes there is a theoretical reason
  - Some relationships on a scatterplot may look familiar and suggest a fix
  - Can always experiment with different functions and assess them with R^2 or AIC

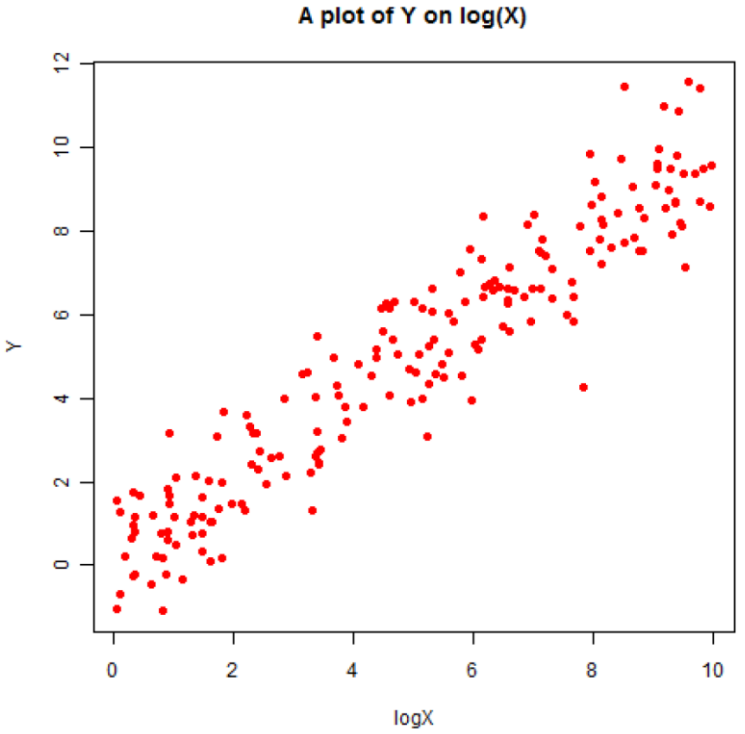
Example



Example



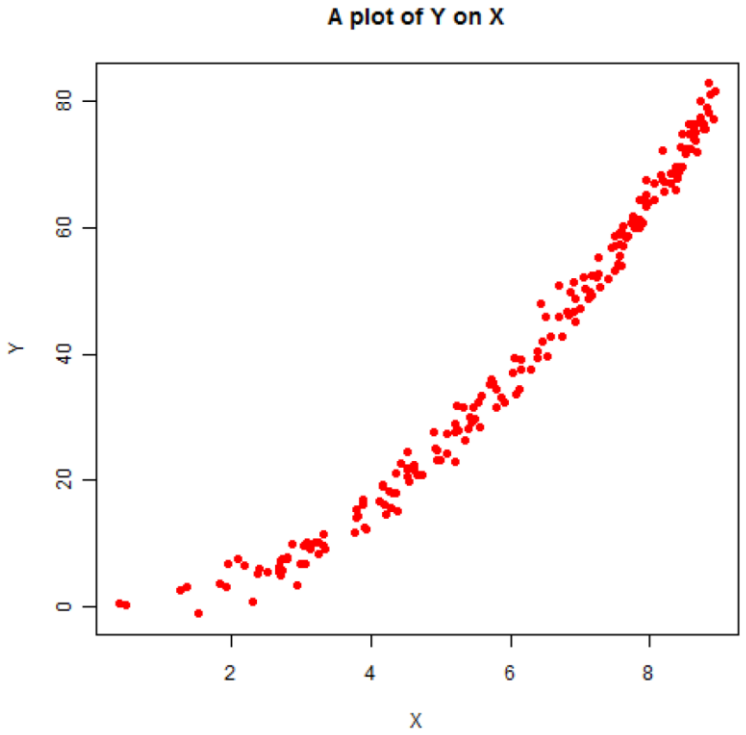
f(X)=log(X)



Correlation coefficient “improved” from 0.65 to 0.95

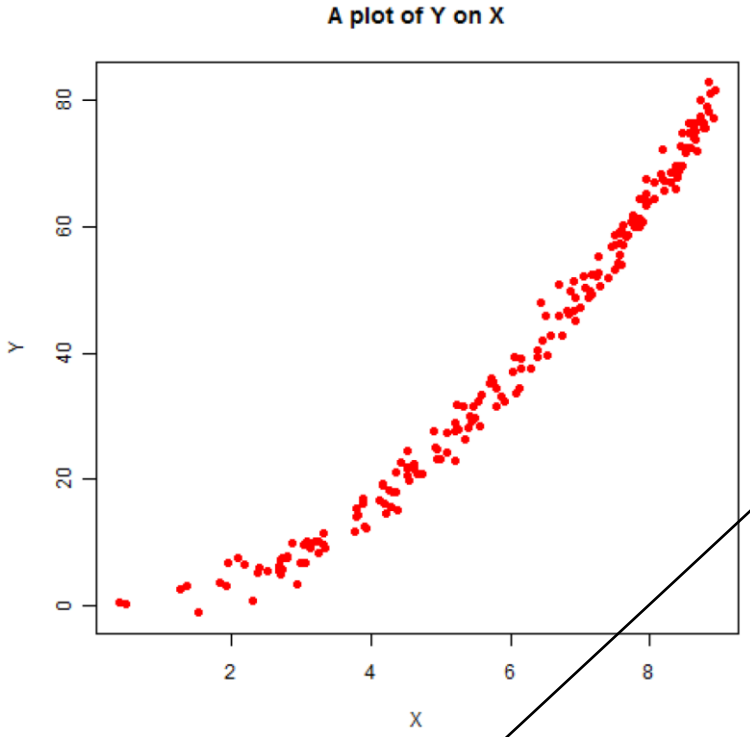
If relationship with transformed variables “looks more linear”, we will get a better fit with linear regression.

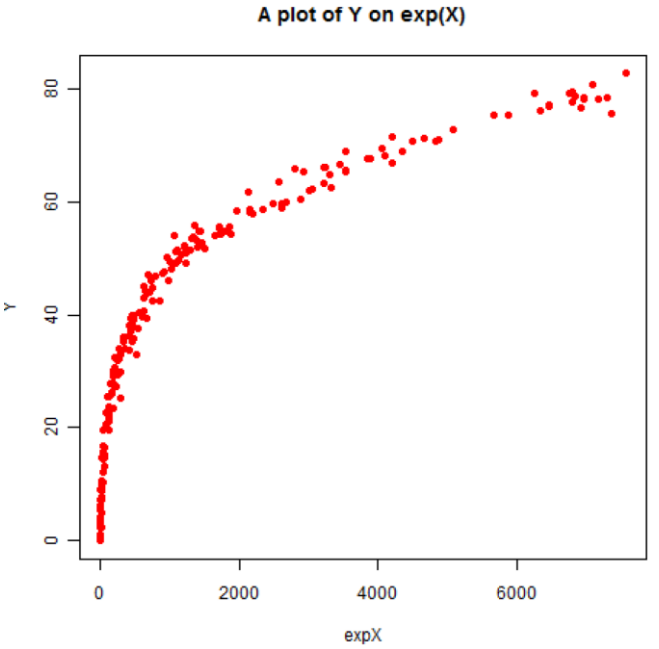
Example (2)

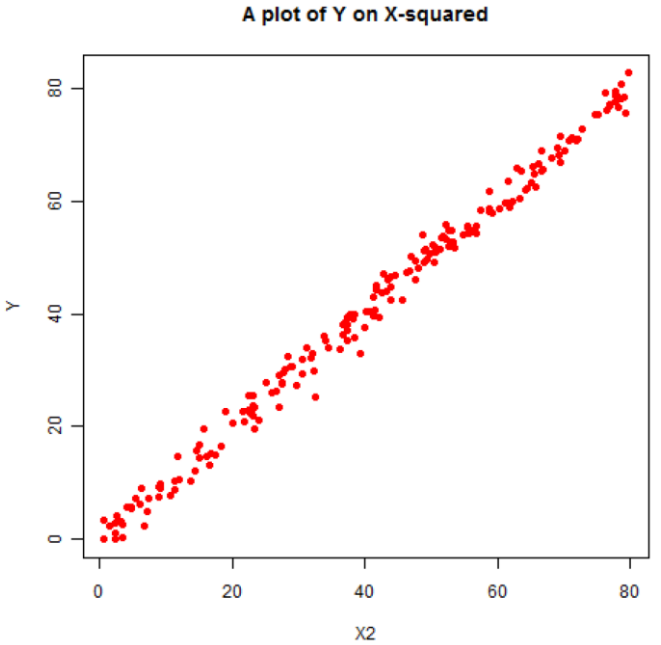


This time Y increases much faster than X.

Example (2)







This time Y increases much faster than X.
Could try …

exp(X) ?

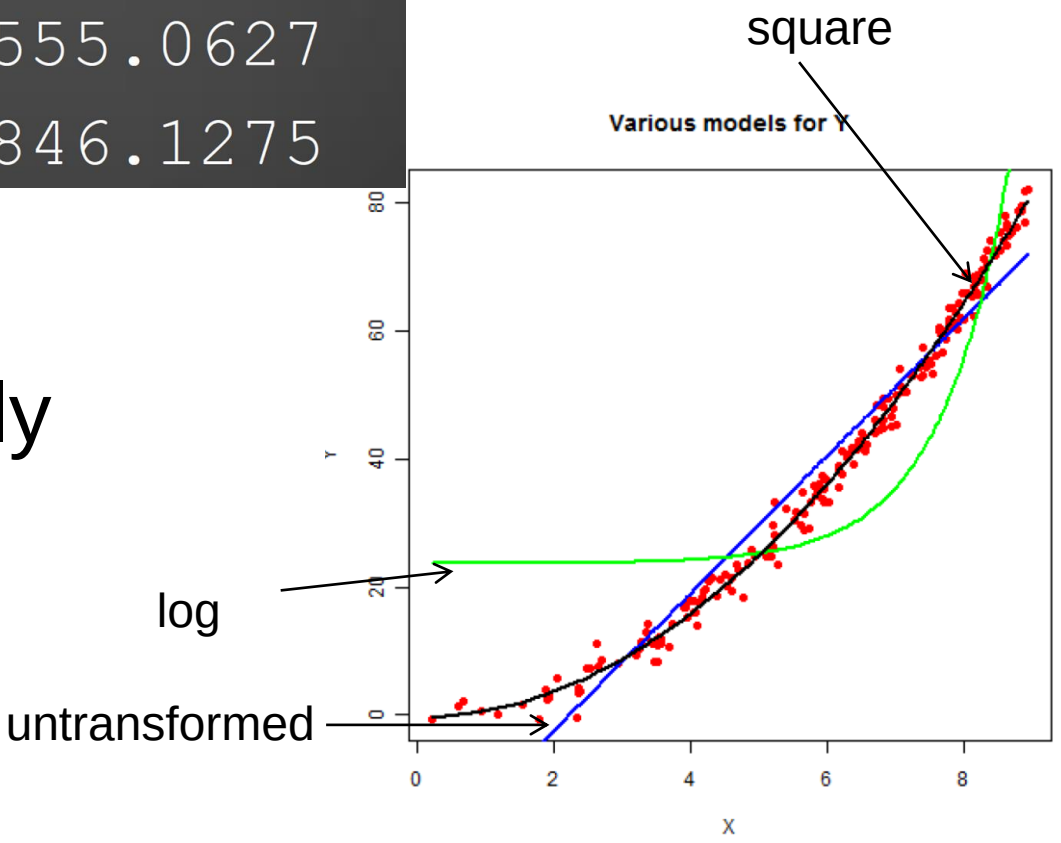
X^2?

Example, cont.

- Can fit regression models for all three cases and then use AIC to assess …

	df	AIC
unchanged	3	1227.2688
expTrans	3	1555.0627
squareTrans	3	846.1275

- But this is also visually
- apparent ...



Dependent Data

Interaction Terms

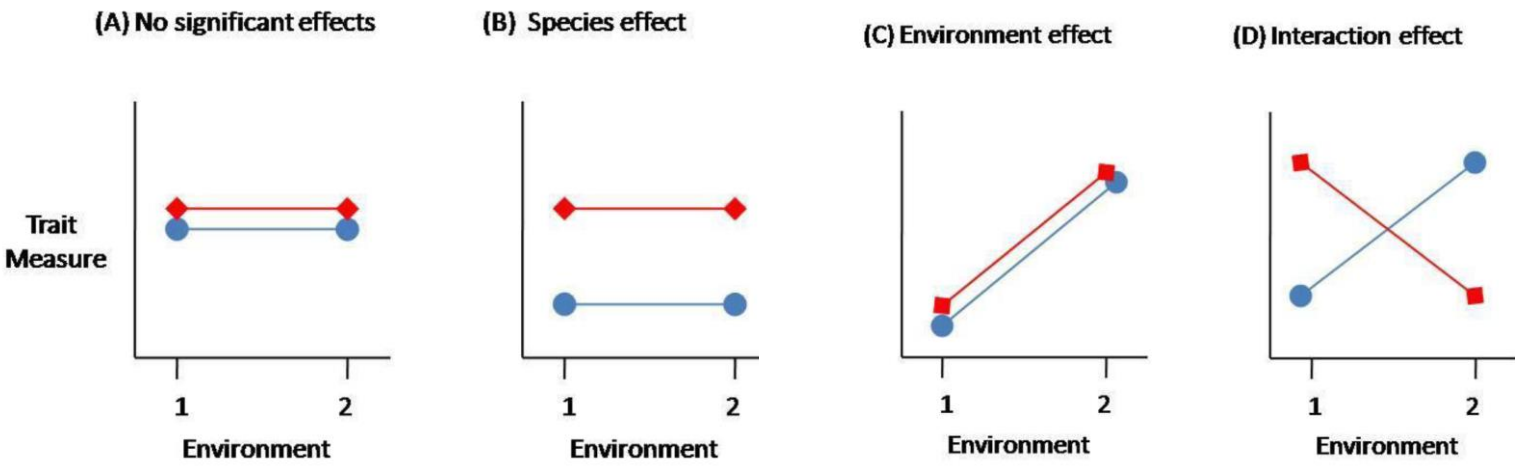
- Regression on multiple variables assumes that these variables are independent
  - i.e. if a predictor variable affects the outcome variable then its effect is independent of all other predictor variables
  - E.g. the linear relationship between X1 and Y is supposed to hold whatever values all other variables X2,X3, … assume
  - E.g.: Want to predict happiness from length of marriage
    - For men happiness increases with length of marriage
    - For women it decreases
    - The relationship may be linear, but it is not independent of sex

Interaction Terms (2)

- If two predictor variables influence the outcome in a way that is **not additive** we need to include an interaction term in the model to capture this effect
- This is the same as **epistasis** in NK landscapes (i.e., cannot say if anchovies improves a pizza without knowing if it has prawns on it)

A Species Example

- Say we have two species, red and blue and two environments 1 and 2
- Species effect: red does better or worse than blue
- Environment effect: avg. fitness in 1 != avg. fitness 2



How to do it in R?

```
model = lm( Y ~ X1 * X2 )
```

- Will regress Y on X1 and X2 including an interaction term X1 by X2
- Equivalent to

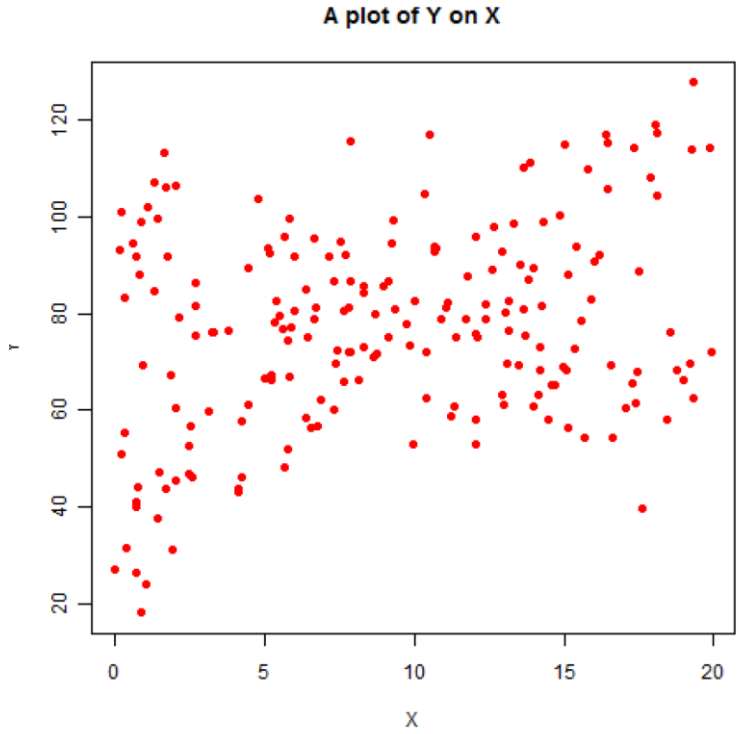
```
model = lm(Y ~ X1 + X2 + X1:X2)
```

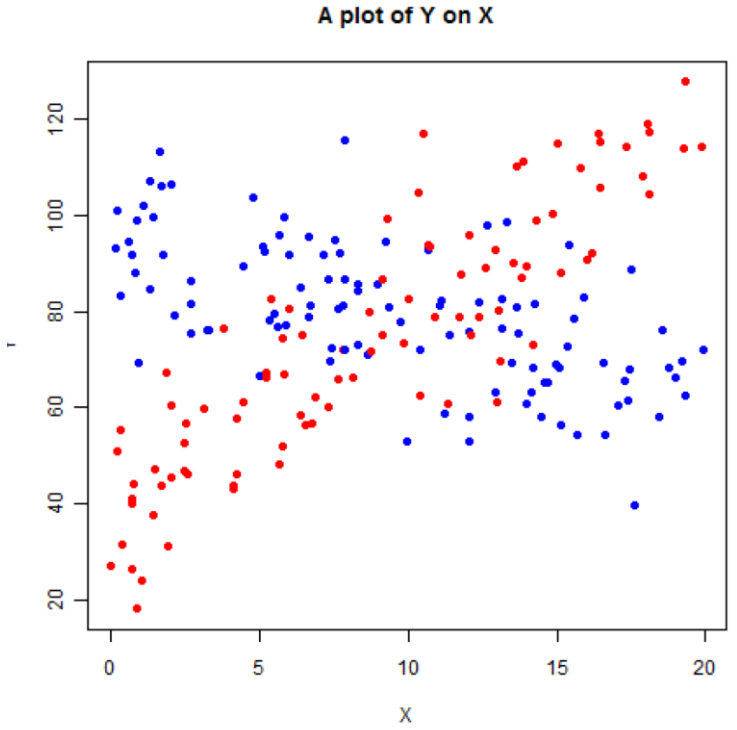
- For syntax for more complex situations see
- http://stat.ethz.ch/R-manual/R-patched/library/stats/html/formula.html

When to include Interaction Terms

- If you have a large number of predictors, it is not practical to include all interactions
  - Due to the combinatorics, you'll quickly have more parameters than data points …
  - Want to keep them to a minimum
- Include when:
  - Theoretical reasons or direct questions that need to be answered; or suggested by other descriptive stats
  - Once in there, up for elimination by model reduction

Example Interaction Terms





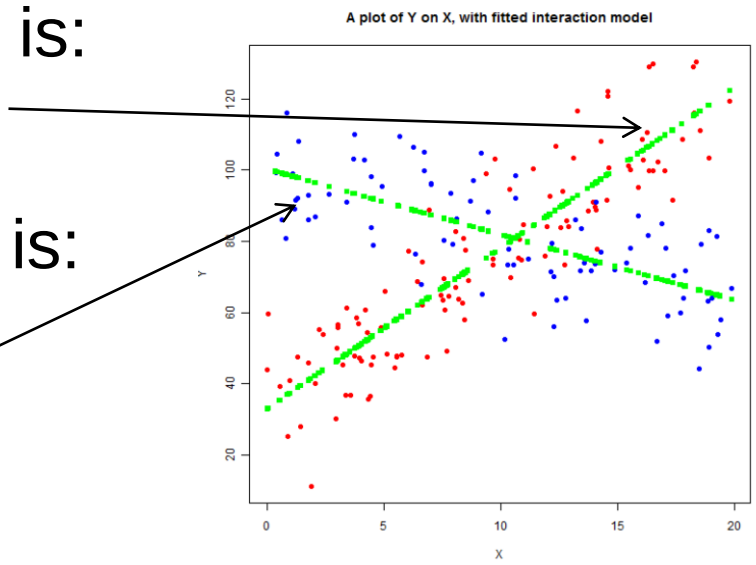
- The outcome measure Y (happiness) is dependent on the predictor X (length of marriage) but also on a categorical variable Group (male or female)

R practicalities

```
lm(formula = Y ~ X * Group)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.4776     2.5465   13.15   <2e-16 ***
X              4.3830     0.2144   20.44   <2e-16 ***
GroupB        67.2934     3.2639   20.62   <2e-16 ***
X:GroupB      -6.3115     0.2756  -22.91   <2e-16 ***
---
```

If you are in group A your prediction line is:
Y=33.47+4.38X

If you are in group B your prediction line is:
Y=(33.47+67.29)+(4.38-6.31)X i.e.
Y=100.77-1.77X



R practicalities (2)

Can the model be reduced?

```
> drop1(fullModel)
Single term deletions

Model:
Y ~ X * Group
         Df Sum of Sq   RSS      AIC
<none>                 23674   962.76 ←
X:Group  1      63366 87039 1221.16
```

AIC analysis confirms that this is not the case, i.e., our preferred model is the one with interactions.

Summary

- What you should remember:
  - Aims of logistic regression
    - When to apply it.
  - Logit transform and why it is used
  - Interpreting (R) outputs of logistic regression
  - How it actually works.
  - You should be able to use it, have a play with the R example.
  - Ideas how to deal with non-linear and dependent data
  - Model reduction, AIC

Some Problems

- If you want to experiment a bit more with predicting movie success using linear/logistic regression models, explore the movie dataset I used in the lectures:

  https://www.southampton.ac.uk/~mb1a10/stats/filmData.txt

- I can also recommend two good step-by-step tutorials:
  - https://www.r-bloggers.com/predicting-movie-ratings-with-imdb-data-and-r/
  - https://rpubs.com/DocOfi/223687

Some Problems

- A data set has been collected to relate the age of a learner to the outcome of driving tests. Carrying out logistic regression, somebody obtains a slope of w=0.01 and an intercept of b=0.1. What are the chances of a 100 years old applicant to pass the test?

Some Problems

- Somebody collects a data set to analyze examination outcomes (discriminating between fail, pass, and repeat) of students on a three-year BSc degree and carries out multinomial logistic regression the predict the outcome dependent on the year of study. Results give:
- intercept (fail)=1 slope fail=-1
- intercept (pass)=3 slope (pass)=-1/2
- What is the chance of a student having to repeat the 3rd year?

Some Problems

- Consider the ridge regression problem (slide 34). Derive an expression for the optimal (augmented) weight vector w.
- In the formulation for ridge regression on the slide also the bias term in w (i.e. component 1) is penalized. This is not always desirable. How would the procedure (and the result derived above) have to be modified to avoid this penalization?
- What are the differences between L1 and L2 regularization?

Some Problems

- Consider the problem of kernel regression (slide 49). Derive the expression for the optimal weight vector given a transformation ϕ.