STK4900/9900 - Lecture 6

Program

1. Binary data and proportions
2. Comparing two proportions
3. Contingency tables
4. Excess risk, relative risk, and odds ratio
5. Logistic regression with one predictor
6. Some comments on classification

- Section 3.4
- Section 5.1
- Supplementary material on proportions and contingency tables (cf. your introductory statistics textbook)

Binary data and proportions

In the first part of the course, we considered the situation where the outcome was numerical

We will now consider the situation where the outcome is a binary variable (coded as 0 or 1)

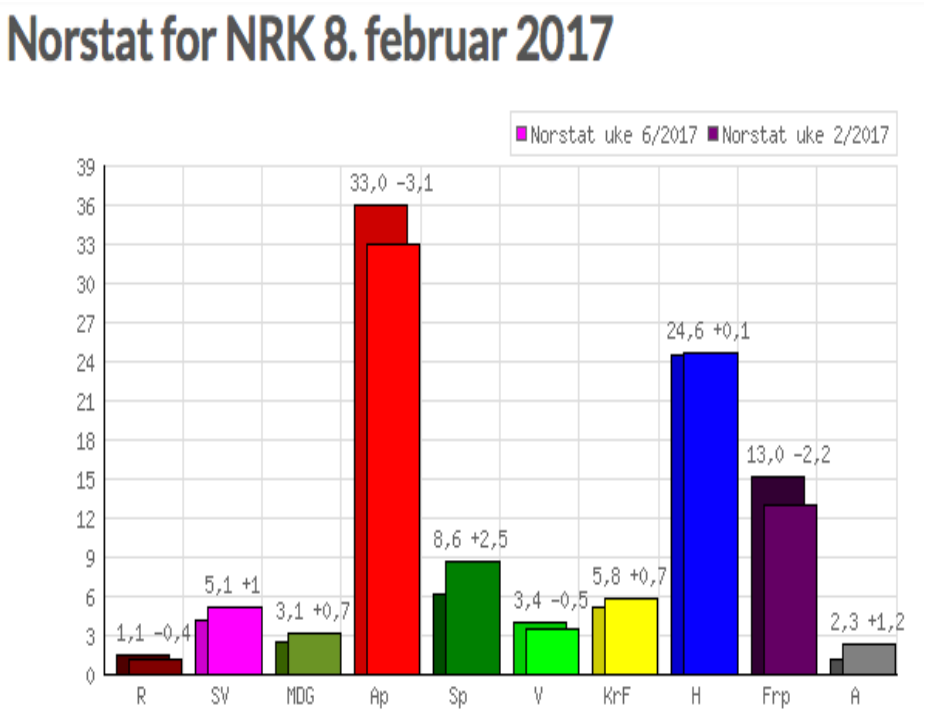
Example: Opinion polls

In February 2017 Norstat asked *n* = 935 individuals which party they would support if there had been election to the parliament tomorrow

309 would have voted Ap

Ap's support on the opinion poll is

$$\frac{309}{935} = 0.330 = 33.0\%$$



In general we have a sample of binary data $y_1, y_2, ..., y_n$ from a population

Here $y_i = 1$ if subject i has a certain property (e.g. vote Ap), while otherwise $y_i = 0$

We let $p = P(y_i = 1)$

Then p is the proportion in the population (with $0 \leq p \leq 1$).

We may estimate p by the sample proportion:

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\#(y_i = 1)}{n}$$

Standard error: $se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

In the example, the standard error becomes

$$se(\hat{p}) = \sqrt{\frac{0.33(1-0.33)}{935}} = 0.0154$$

One may show that $\hat{p}$ is approximately normally distributed (cf. the central limit theorem)

95% confidence interval for the population proportion p :

$$\hat{p} \pm 1.96 \cdot se(\hat{p})$$

In the example a 95% confidence interval becomes:

$$0.33 \pm 1.96 \cdot 0.0154$$

i.e.

$$0.33 \pm 0.03$$

Thus our estimate of Ap's support is 33.0% with a "margin of error" of ± 3%

Comparing two proportions

Assume that we have a random sample of binary data from each of two populations, and that the two samples are independent

Example: "Divorce" among seagulls

Kittiwake (krykkje) is a seagull whose mating behavior is basically monogamous, but some couples do not reunite the next breeding season ("divorce")

Does the "divorce rate" depend on whether breeding was successful or not?

769 kittiwake pair-bonds were studied over two breeding seasons

Of the 160 couples that had not successful breeding the first season, 100 divorced

Of the 609 couples that were successful, 175 divorced

Population 1:

Population proportion: p_1

Sample size: n_1

Sample proportion: $\hat{p}_1$

Population 2:

Population proportion: p_2

Sample size: n_2

Sample proportion: $\hat{p}_2$

We are interested in estimating $p_1 - p_2$ and testing $H_0 : p_1 = p_2$

95% confidence interval for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \cdot se(\hat{p}_1 - \hat{p}_2)$$

where

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{se(\hat{p}_1)^2 + se(\hat{p}_2)^2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

In the example:

Unsuccessful (population 1):

Sample size: $n_1 = 160$

Sample proportion: $\hat{p}_1 = \frac{100}{160} = 0.625$

Successful (population 2):

Sample size: $n_2 = 609$

Sample proportion: $\hat{p}_2 = \frac{175}{609} = 0.287$

We obtain:

$$\hat{p}_1 - \hat{p}_2 = 0.625 - 0.287 = 0.338$$

$$se(\hat{p}_1 - \hat{p}_2) = 0.0424$$

95% confidence interval:

$0.338 \pm 1.96 \cdot 0.0424$ i.e. $0.338 \pm 0.083 = (0.255, 0.421)$

We then consider testing the null hypothesis $H_0 : p_1 = p_2$ versus the (two-sided) alternative $H_A : p_1 \neq p_2$

Test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{se_0(\hat{p}_1 - \hat{p}_2)}$$

Here $se_0(\hat{p}_1 - \hat{p}_2)$ is the estimated standard error under the null hypothesis, obtained by using the sample proportion $\hat{p}$ in the two samples combined:

$$se_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

We reject H_0 for large values of $|z|$

Under H_0 the test statistic is approximately standard normal

P-value (two-sided): $P = 2\,P(Z > |z|)$ where Z is standard normal

In the example:

Unsuccessful (population 1):

$$n_1 = 160 \qquad \hat{p}_1 = \frac{100}{160} = 0.625$$

Successful (population 2):

$$n_2 = 609 \qquad \hat{p}_2 = \frac{175}{609} = 0.287$$

We obtain:

$$\hat{p} = \frac{100 + 175}{160 + 609} = 0.358$$

$$se_0(\hat{p}_1 - \hat{p}_2) = 0.0426$$

The test statistic takes the value

$$z = \frac{0.625 - 0.287}{0.0426} = 7.9$$

which is highly significant

NB! Important to also consider and report effect size NB!

2x2 tables

It is common to summarize the situation with two binary samples in a 2x2 table. For the example we have the 2x2 table:

	divorced	not divorced	Total
Unsuccessful	100	60	160
Successful	175	434	609
Total	275	494	769

An alternative way of formulating the test for the null hypothesis of no difference between the populations (cf. slide 8), is to compare the observed numbers in the table (denoted O's) with the corresponding expected numbers if the null hypothesis is true (denoted E's)

If there is no difference between the two groups we would (e.g.) expect

$$160 \cdot \frac{275}{769} = 57.2$$

divorces among the unsuccessful couples

Expected numbers:

	divorced	not divorced	Total
Unsuccessful	57.2	102.8	160
Successful	217.8	391.2	609
Total	275	494	769

Test statistic: $\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$

We reject H_0 for large values of χ^2

Under H_0 the test statistic is approximately chi-square distributed with 1 degree of freedom (df) provided that all E's are at least 5

P-value: $P(\chi^2 \geq \chi^2_{\text{obs}})$ where χ^2 is chi-square distributed with 1 df

One may show that $\chi^2 = z^2$ so this is a reformulation of the test on slide 8

R commands:

```
kittiwake=matrix(c(100,175,60,434),nrow=2)
dimnames(kittiwake)=list(c("unsuccessfull","successful"),c("divorced","not_divorced"))
kittiwake
chisq.test(kittiwake,correct=F)$expected
prop.test(kittiwake,correct=F)
```

R output (edited):

```
                divorced not_divorced
unsuccessfull        100           60
successful           175          434

                divorced not_divorced
unsuccessfull     57.217      102.783
successful       217.783      391.217

X-squared = 62.8813, df = 1, p-value = 2.196e-15
alternative hypothesis: two.sided
95 percent confidence interval:
 0.25446 0.42082
sample estimates:
   prop 1    prop 2
0.62500     0.28736
```

Contingency tables

The chi-square test may be extended to contingency tables of higher order

Example: Blood pressure

Blood pressure of 92 teenagers according to the blood pressure of their fathers:

Father's blood pressure	Child's blood pressure: Lower third	Middle third	Upper third	Total
Lower third	14	11	8	33
Middle third	11	11	9	31
Upper third	6	10	12	28
Total	31	32	29	92

Does the blood pressure of the children depend on the blood pressure of their fathers?

We will test the null hypothesis that there is no difference between the groups (in the example, that the blood pressure of the children does not depend on the blood pressure of their fathers)

Expected numbers (E's) are computed as for 2x2 tables

Test statistic: $\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$

We reject H_0 for large values of χ^2

Under H_0 the test statistic is approximately chi-square distributed with $df = (\#\text{rows} - 1) \cdot (\#\text{columns} - 1)$ provided that all E's are at least 5

In the example we have $df = (3-1) \cdot (3-1) = 4$

R commands:

```
bloodpr=matrix(c(14,11,6,11,11,10,8,9,12),nrow=3)
dimnames(bloodpr)=list(c("F.low","F.middle","F.upper"), c("C.low","C.middle","C.upper"))
bloodpr
chisq.test(bloodpr,correct=F)$expected
chisq.test(bloodpr,correct=F)
```

R output (edited):

```
          C.low  C.middle  C.upper
F.low        14        11        8
F.middle     11        11        9
F.upper       6        10       12

          C.low    C.middle   C.upper
F.low    11.120    11.478     10.402
F.middle 10.446    10.783      9.772
F.upper   9.435     9.739      8.826


Pearson's Chi-squared test
X-squared = 3.814, df = 4, p-value = 0.432
```

Risk measures

Assume that we have a random sample of binary data from each of two populations, and that the two samples are independent

We assume that population 1 corresponds to an "exposed" population (specified by $x = 1$) and that population 2 corresponds to an "unexposed" population (specified by $x = 0$)

Example: "Divorce" among seagulls

	divorced (y=1)	not divorced (y=0)	Total
Unsuccessful (x=1)	100	60	160
Successful (x=0)	175	434	609
Total	275	494	769

Population 1: $p_1 = p(1) = P(y = 1 \mid x = 1)$

Population 2: $p_2 = p(0) = P(y = 1 \mid x = 0)$

On slide 6 we used the *excess risk*

$$ER = p(1) - p(0)$$

to measure the effect of the "exposure"

An alternative would be to use the *relative risk* given by:

$$RR = \frac{p(1)}{p(0)}$$

In the example, estimates of these two measures of risk are given by (cf. slide 7)

$$ER = \hat{p}(1) - \hat{p}(0) = 0.625 - 0.287 = 0.338$$

$$RR = \frac{\hat{p}(1)}{\hat{p}(0)} = \frac{0.625}{0.287} = 2.18$$

A third risk measure is based on the concept of *odds*, so we will first discuss this concept

Assume that an event has a probability p of occurring

Then the odds for the event is

$$\text{odds} = \frac{p}{1-p}$$

The odds is one if the probability that an event will happen is equal to the probability that it will not happen, cf. the expression "a fifty-fifty chance"

When you throw a die, the odds that it will face six is 1 : 5
(i.e. it is five times more likely that it will not face six than it will face six)

We then return to the situation with two populations:

$$\text{Population 1:} \quad p(1) = P(y = 1 \mid x = 1)$$

$$\text{Population 2:} \quad p(0) = P(y = 1 \mid x = 0)$$

The odds for the two populations are:

$$\text{Population 1:} \quad \frac{p(1)}{1 - p(1)} \qquad \text{Population 2:} \quad \frac{p(0)}{1 - p(0)}$$

Then a third risk measure is the *odds ratio*

$$OR = \frac{p(1)/[1 - p(1)]}{p(0)/[1 - p(0)]}$$

In the example, an estimate for the odds ratio becomes (cf. slide 7)

$$OR = \frac{0.625/(1 - 0.625)}{0.287/(1 - 0.287)} = \frac{1.667}{0.403} = 4.14$$

Why Odds-ratio?

- Turns up in logistic regression!
- Related to relative risk in the following way

  + *RR*=1 when *OR*=1 (and vice versa)

  + if *RR*>1 we have *1<RR<OR*

  + if *RR<1* we have *OR < RR < 1*

  + if *p(1)* and *p(0)* are both small then

OR=1 Exposure does not affect odds of outcome
OR>1 Exposure associated with higher odds of outcome
OR<1 Exposure associated with lower odds of outcome

$$OR = \frac{p(1)/[1-p(1)]}{p(0)/[1-p(0)]} \approx RR = \frac{p(1)}{p(0)}$$

+ Examples: *p(1)=0.2* and *p(0)=0.1* gives *RR=2* and *OR=2.25*
p(1)=0.1 and *p(0)=0.05* gives *RR=2* and *OR=2.11*

Logistic regression with one predictor

When discussing logistic regression, we will to a large extent use the WCGS study for illustration

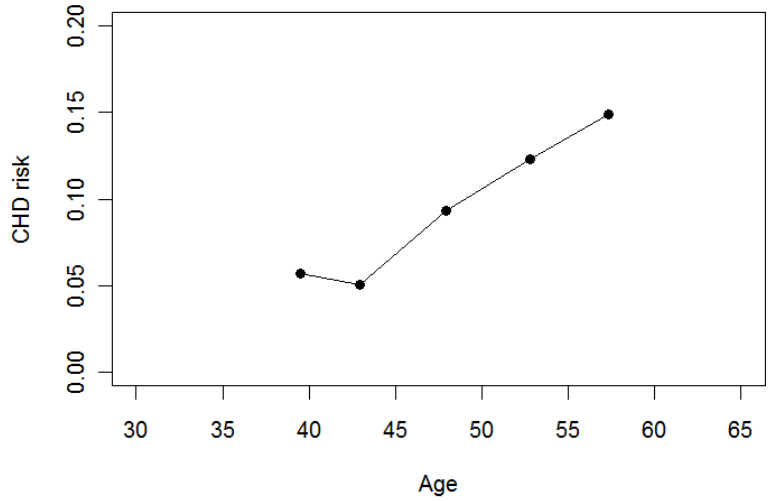
WCGS is a large epidemiological study designed to study risk factors for coronary heart disease (CHD) among middle-aged men

The men were followed for 10 years, and for each man it was recorded if he developed CHD (y=1) or not (y=0) over the course of the study

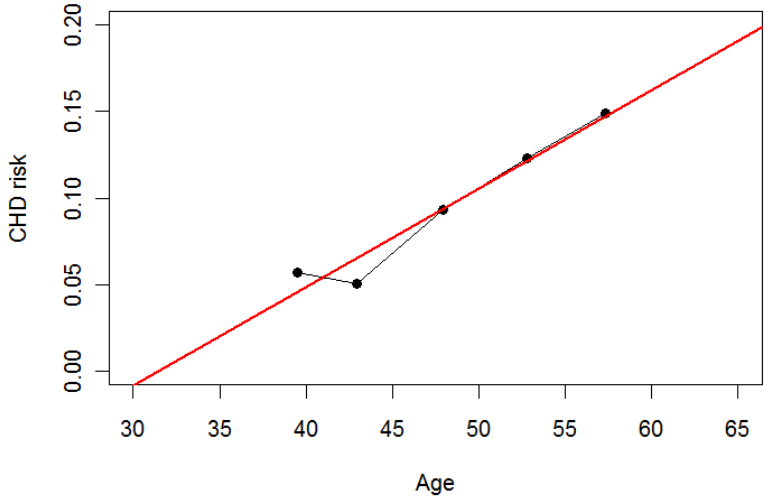
How does the age (at entry to the study) affect the risk (probability) of developing CHD?

Age group (mean)	35-40 (39.5)	41-45 (42.9)	46-50 (47.9)	51-55 (52.8)	56-60 (57.3)
# Total	543	1091	750	528	242
# CHD	31	55	70	65	36
% CHD	5.7 %	5.0 %	9.3 %	12.3 %	14.9 %

The figure shows the observed proportion with CHD plotted versus the mean age in each age group



A least square fit to the observed proportions gives the fitted line



This least squares line may give an all right description of the observed proportions, but there are in general problems with using linear regression for binary data and proportions

In general we have data (x_1,y_1) , ... , (x_n,y_n)

Here y_i is a binary outcome (0 or 1) for subject i and x_i is a predictor for the subject (which may be binary or numerical)

In the WCGS study, $y_i = 1$ if man number i developed CHD during the course of the study, $y_i = 0$ if not, and x_i may be his age (at entry to the study)

In general we let

$$p(x) = E(y \mid x) = P(y = 1 \mid x)$$

We want a model that specifies a relation between $p(x)$ and x

One option would be a linear model:

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

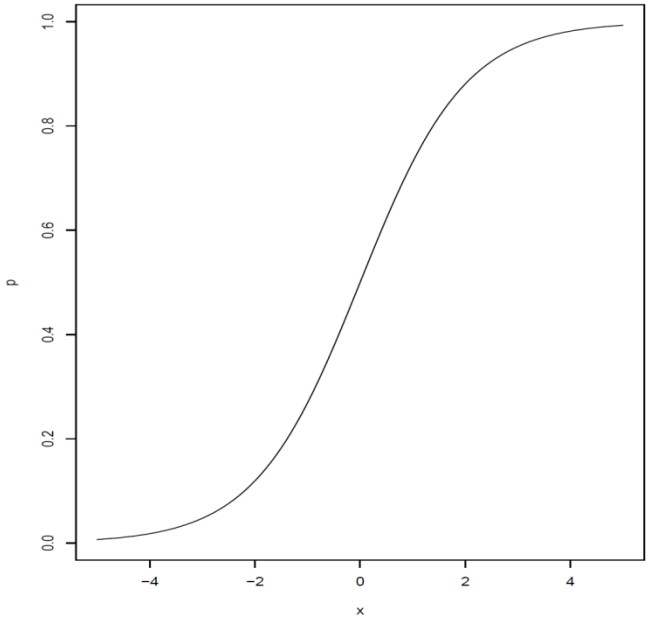
This is an additive risk model, which may be useful in some situations

However, it is a main problem with the additive risk model that it may give impossible values for the probabilities (negative or above 1)

To avoid this problem it is common to consider the logistic regression model given by

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

This gives a "S-shaped" relation between $p(x)$ and x

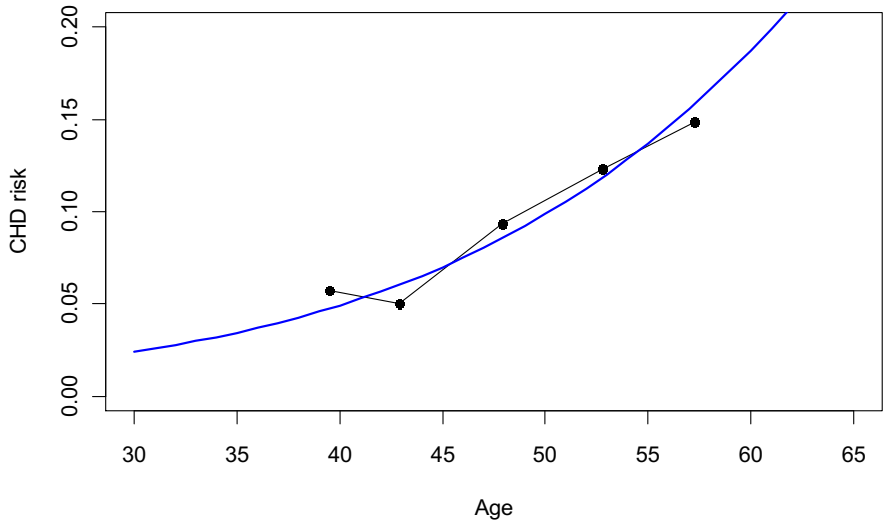


If we fit a logistic regression model for the WCGS data using the mean age in each age group as a numeric covariate, we get

$\hat{\beta}_0 = -5.947$ and $\hat{\beta}_1 = 0.0747$

This gives the fitted model

$$\hat{p}(\text{age}) = \frac{\exp(-5.947 + 0.0747 \cdot \text{age})}{1 + \exp(-5.947 + 0.0747 \cdot \text{age})}$$



The method for estimating the parameters of a logistic regression model will be described in Lecture 7

The logistic model may alternatively be given in terms of the odds:

$$\frac{p(x)}{1-p(x)} = \exp(\beta_0 + \beta_1 x)$$

If we consider two subjects with covariate values $x+\Delta$ and x, respectively, their odds ratio becomes

$$\frac{p(x+\Delta)/[1-p(x+\Delta)]}{p(x)/[1-p(x)]} = \frac{\exp\big(\beta_0 + \beta_1 (x+\Delta)\big)}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1 \Delta)$$

In particular e^{β_1} is the odds ratio corresponding to one unit's increase in the value of the covariate

In the WCGS study the odds ratio for one year increase in age is $e^{0.0747} = 1.078$ while the odds ratio for a ten-year increase is $e^{0.0747 \cdot 10} = 2.11$

(The numbers deviate slightly from those on pp 144-145 in the text book, since we have used mean age for each age group in this illustration; cf. the exercises for the results when actual age is used.)

R commands for logistic regression

Binary CHD data with mean age in each age group as covariate

R commands:

```
wcgs=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/data/wcgs.txt",
                sep="\t",header=T,na.strings=".")
wcgs$agem=39.5*(wcgs$agec==0)+42.9*(wcgs$agec==1)+47.9*(wcgs$agec==2)+
                52.8*(wcgs$agec==3)+57.3*(wcgs$agec==4)
attach(wcgs)
cbind(chd69, agem)
```

R output of binary CHD data (edited):

	chd69	agem
[1,]	0	47.9
[2,]	0	52.8
[3,]	0	57.3
[4,]	0	52.8
[5,]	0	42.9
[6,]	0	47.9
[7,]	0	39.5
[8,]	0	42.9
[9,]	0	47.9

	chd69	agem
[10,]	0	42.9
[11,]	0	57.3
[12,]	0	52.8
[13,]	0	47.9
[14,]	1	39.5
[15,]	0	47.9
[16,]	0	52.8
[17,]	0	42.9
[18,]	0	57.3
[19,]	0	42.9

	chd69	agem
[3145,]	0	42.9
[3146,]	0	42.9
[3147,]	0	52.8
[3148,]	0	42.9
[3149,]	0	42.9
[3150,]	0	47.9
[3151,]	0	42.9
[3152,]	0	52.8
[3153,]	0	52.8
[3154,]	0	47.9

When we use the mean age in each age group as covariate, all information is summarized in the table

Age group (mean)	35-40 (39.5)	41-45 (42.9)	46-50 (47.9)	51-55 (52.8)	56-60 (57.3)
# Total	543	1091	750	528	242
# CHD	31	55	70	65	36

As an alternative to using the individual binary data, we may therefore use the grouped data given in the table

R commands:

```
chd.grouped=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/data/chd_grouped.txt ",
                       header=T)
chd.grouped
```

R output of grouped CHD data:

```
  no  chd agem
 543   31 39.5
1091   55 42.9
 750   70 47.9
 528   65 52.8
 242   36 57.3
```

We may fit the logistic regression model using the individual binary data or by using the grouped data

R commands for binary data:

```
fit.binary=glm(chd69~agem, data=wcgs,family=binomial)
summary(fit.binary)
predict(fit.binary, type = "response", data.frame(agem=50)) #predicts prob. at age 50
```

R commands for grouped data:

```
fit.grouped=glm(cbind(chd,no-chd)~agem, data=chd.grouped, family=binomial)
summary(fit.grouped)
```

The two ways of fitting the logistic regression model give the same estimates and standard errors:

R output (edited):

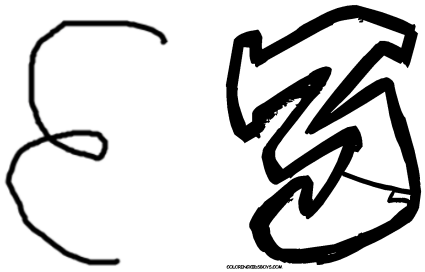
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9466	0.5616	-10.588	< 2e-16
agem	0.0747	0.0116	6.445	1.15e-10

(Other parts of the R output will differ, as we will discuss in Lecture 7)

Binary classification with logistic regression

- We introduced logistic regression as a regression model for qualitative (=categorical) response variables, with two categories (response 0 or 1)
- When there are more than two categories, there exist natural multiple-class extensions (but for simplicity we stick to binary problems here)
- When we use the fitted logistic model to predict a categorical response, we first predict the probability of each of the categories, and can then use this predicted probability to select a category. In this sense logistic regression can be viewed as a CLASSIFIER – we perform CLASSIFICATION
- Logistic regression is one of the most widely used classifiers, and is the basic building block underlying many statistical/machine learning methods such as the deep learning algorithm

Binary classification



← Is this the digit ´3´?

Will this consumer click on my ad? ➔

← Will this patient survive?

Remember we had

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

If $p(x) \geq 0.5$, or equivalenty,

$\beta_0 + \beta_1 x \geq 0$ ➔ classify as y=1

If $p(x) < 0.5$, or equivalenty,

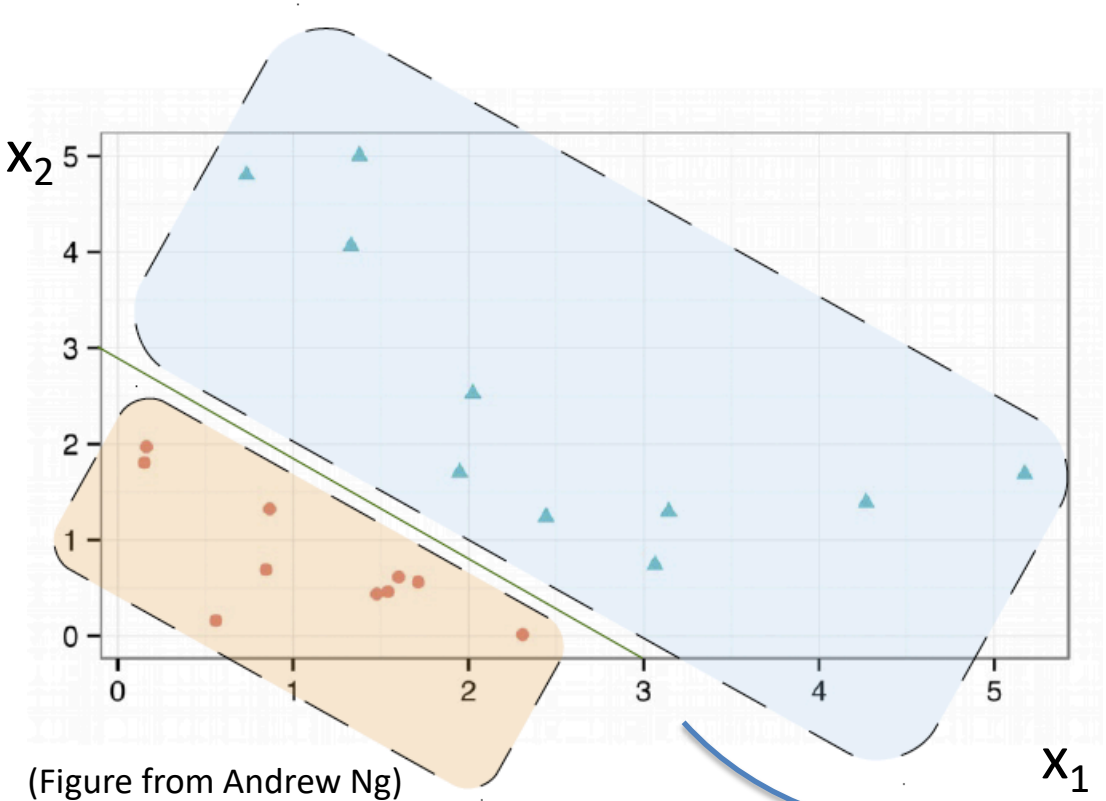
$\beta_0 + \beta_1 x < 0$ ➔ classify as y=0

NB! Not mandatory to divide at 0.5. Depends on the application in question

Extend to two predictors x_1 and x_2

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0$ ➔ classify as y=1



(Figure from Andrew Ng)

Example:

$$\beta = [-3, 1, 1]^T$$

$$-3 + x_1 + x_2 \geq 0$$

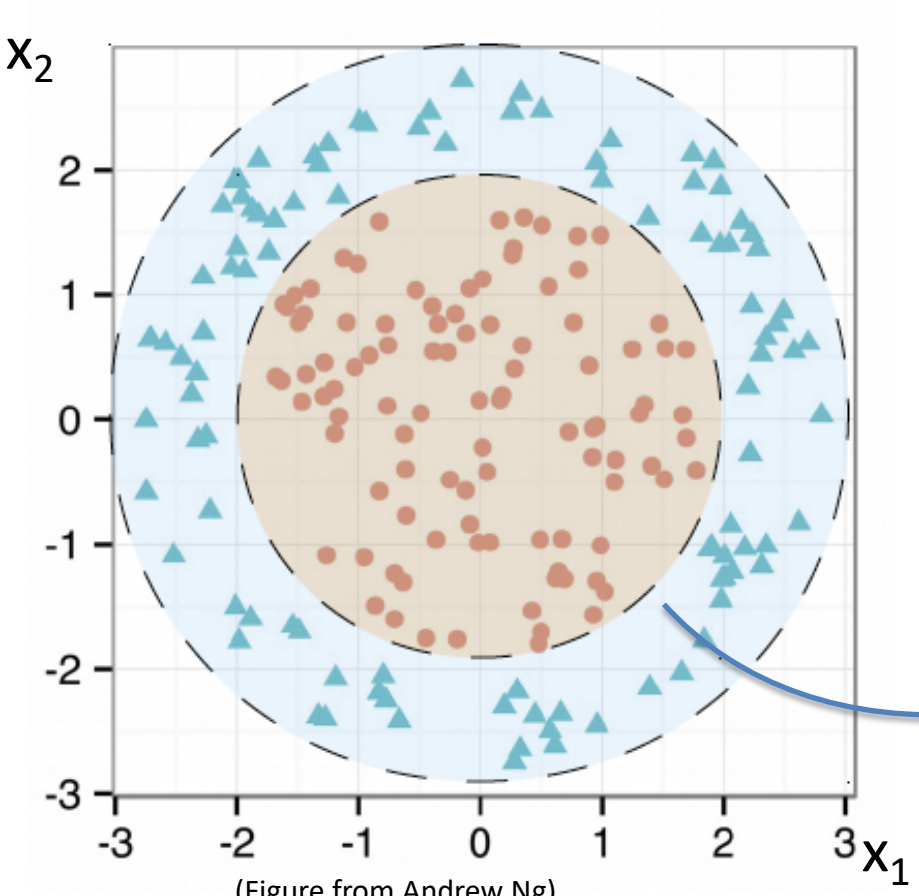
$$x_1 + x_2 \geq 3$$

Linear separator

With higher order terms

$$p(x) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2\right)}$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 \geq 0$$



(Figure from Andrew Ng)

➔ classify as y=1

Example:

$$\beta = [-2, 0, 0, 1, 1]^T$$

$$x_1^2 + x_2^2 \geq 2$$

There exist several different classification techniques that we can use to predict qualitative responses (like above).

The most commonly used are

- logistic regression
- linear discriminant analysis (LDA)
- K-nearest neighbours (KNN)

Logistic regression and LDA are very similar, but the LDA assumes that predictors are normally distributed. KNN is completely non-parametric. No method will systematically dominate the other.

More computer intensive methods are for example

- generalized additive models (tomorrow)
- tree-based methods
- random forests
- boosting
- support vector machines (SVM)

Suggested reading:
An Introduction to Statistical Learning
James, Witten, Hastie, Tibshirani
Springer 2013