Examining associations (Regression)

Outcome is a measure of risk Endpoints: counting people, and time to event

Logistic regression –counting people

- To evaluate association between an exposure and an disease
- Similar principles as linear regression
- Difference:
  - outcome has two levels (ie binary, eg: disease and no disease). With linear regression the outcome is continuous.
  - No Normality assumptions as in linear regression
- 'Odds ratio' is the effect size.
- It has some useful mathematical properties that allow easier modelling (compared to relative risk)

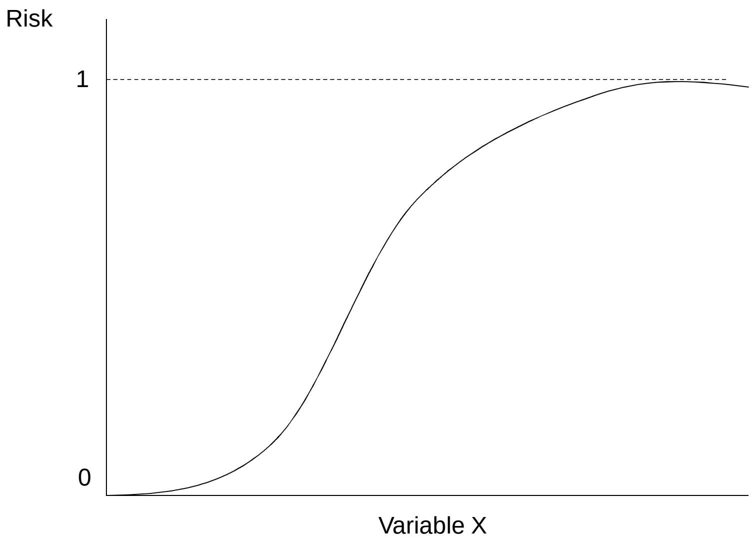
Odds Ratio (OR)

$$\text{Odds Ratio (OR)} = \frac{\text{odds of event in exposed}}{\text{odds of event in unexposed}}$$

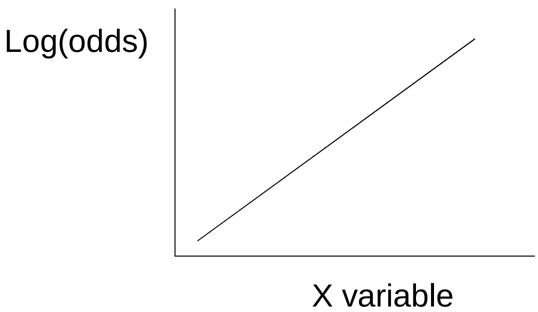
Event= a disease, death, or any well-defined occurrence (eg admission to hospital)

Logistic compared to linear

- In linear regression the y-value (a continuous measurement) can in theory have any value (negative and positive) on an infinite scale
- But in logistic regression we have risk; which can only be between 0 and 1
- The relationship between risk and a factor therefore looks like the following figure:



- This is not easy to model.
- To overcome this, instead of having the y-axis as risk, it uses a transformation:
- If p=risk (eg the proportion of people with a disorder)
- Then $\log_e(\text{odds}) = \log_e[p/(1-p)]$
- This is then on the same infinite scale as we'd get with a continuous measurement, and we can fit a line as with linear regression, ie a linear relationship
- Occasionally, we need to take a transformation of X (eg log) to get linear relationship



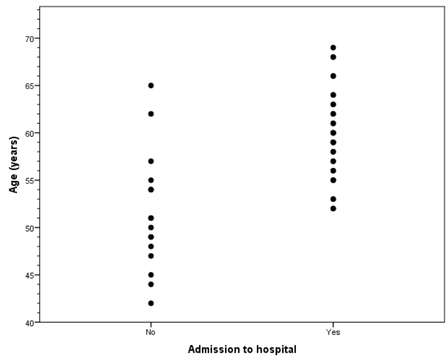
Logistic regression assumptions

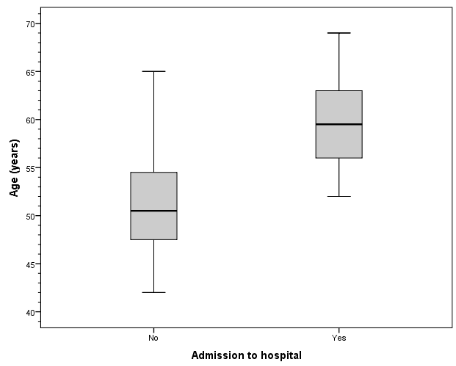
- The outcome measure is usually binary (dead/alive, disease/no disease), but can be extended to measures that have ≥3 levels
- The x-variables can be categorical or continuous (but not time-to-event unless everyone has had the event)
- Continuous x-variables do not have to be Normally distributed (but sometimes may help, ie to get a better/more reliable model, so a transformation such as logs can be done)
- The observations must be independent
- The variables should not be linear combinations of each other (eg 3 factors: height, weight and Z, where Z=heightx2 + weight)

Examples

- Outcome= admission to hospital (yes or no)
- Three factors (called exposures, covariates, or x-variables) to be examined separately:
- Age (continuous)
- Sex (binary, ie 2 levels)
- Social class (categorical, ie 4 levels)

Look at the data first - age





SPSS output for logistic regression of risk of hospital admission y/n = age

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	.304	.089	11.662	1	.001	1.355	1.138	1.613
	Constant	-16.143	4.900	10.855	1	.001	.000		

a. Variable(s) entered on step 1: Age.

This is the model: $\log_e$ odds of CV = -16.143 + 0.304xAge
[NB: odds = risk/1-risk]

So, we can predict a risk value for anyone
Eg, if age = 60 years
Log odds = -16.143 + 0.304x60 = 2.097
$\log_e$ (risk/1-risk) = 2.097, so re-arranging gives risk = 89%
(which seems about right, if you look at the scatter plot above)

SPSS output for logistic regression of risk of hospital admission y/n = age

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	.304	.089	11.662	1	.001	1.355	1.138	1.613
	Constant	-16.143	4.900	10.855	1	.001	.000		

a. Variable(s) entered on step 1: Age.

The log odds ratio for age is 0.304.
But we want to work with the anti-log scale
Hence use Exp(B) = 1.355
As age increases by 1 unit (i.e. 1 year) the risk of hospital admission increases by 35%

The expected true odds ratio could be between 1.14 and 1.61
P-value for this odds ratio = 0.001

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.836	8	.118

This bit of the output provides a test of whether the model is a good fit to the data (small p-values <0.05, indicate it might not be)

Converting a continuous x-factor into categorical

- If there isn't a sufficiently clear linear relationship using age as a continuous factor, we can examine it as a categorical factor
- Also, it is sometimes easier to <u>interpret</u> the results using categories
- But a problem with converting a continuous x-factor into categorical one, is that we lose information on variability
- Should always use the factor as continuous first, and do another logistic regression using the categories and check that the results are generally similar

SPSS output for logistic regression of risk of hospital admission y/n = age, divided into categories

Variables in the Equation									
								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Age_group (<60)			15.701	2	.000			
	Age_group(60-65)	3.239	.945	11.757	1	.001	25.500	4.004	162.381
	Age_group(>65)	2.970	.954	9.693	1	.002	19.500	3.006	126.515
	Constant	-1.099	.577	3.621	1	.057	.333		

a. Variable(s) entered on step 1: Age_group.

So, odds ratio for getting admitted to hospital is:
OR=25.5 among 60-65 compared to <60 years
OR=19.5 among >65 compared to <60 years

We're looking for a trend, though you may not always see this clearly

One advantage of this, is that you can plot the two ORs and 95% CI on a diagram (as error bars)

Look at the data first - sex

Hospital admission * Gender Crosstabulation				
Count				
		Gender		Total
		Female	Male	
Hospital admission	No	11	5	16
	Yes	14	20	34
Total		25	25	50

So, 56% (14/25) females were admitted to hospital compared to 80% (20/25) males
We can see that the risk is lower, so the logistic regression should also show this

SPSS output for logistic regression of risk of hospital admission y/n = sex

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Gender(1)	-1.145	.642	3.180	1	.075	.318	.090	1.120
	Constant	1.386	.500	7.687	1	.006	4.000		

a. Variable(s) entered on step 1: Gender.

The log odds ratio for sex is -1.145
But we want to work with the anti-log scale
Hence use Exp(B) = 0.318
The odds of hospital admission is 0.32 times lower in females than in males
(ie the odds has been reduced by 68%)
Make sure you know which level of the factor is made the comparison
(here females vs males, not males vs females)

The expected true odds ratio could be between 0.09 and 1.20 (wide)
The 95% CI includes the no effect value of 1
P-value for this odds ratio = 0.075 (ie not statistically significant)

There may be a difference in the odds of hospital admission between the two
Groups but the evidence not strong here

Look at the data first – social class

Hospital admission * SE_class Crosstabulation

Count

		SE_class				Total
		Low	Lower middle	Upper middle	High	
Hospital admission	No	3	5	2	6	16
	Yes	10	7	10	7	34
Total		13	12	12	13	50

So the 4 risks are:
10/13 7/12 10/12 7/13
77% 58% 83% 54%
No clear pattern/association

SPSS output for logistic regression of risk of hospital admission y/n = social class

Variables in the Equation								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Low			3.311	3	.346			
	Lower middle	1.050	.862	1.484	1	.223	2.857	.528	15.473
	Upper middle	.182	.808	.051	1	.821	1.200	.246	5.844
	High	1.455	.954	2.329	1	.127	4.286	.661	27.785
	Constant	.154	.556	.077	1	.782	1.167		

a. Variable(s) entered on step 1: SE_class.

One level of social class has to be the comparison group (here 'Low')
So, with 4 levels, there will be 3 odds ratios

The odds of admission among Lower middle is 2.857 times higher compared to Low
The odds of admission among Upper middle is 1.200 times higher compared to Low
The odds of admission among High is 4.28 times higher compared to Low

All of the 95% CIs include the no effect value 1.0

SPSS output for logistic regression of risk of hospital admission y/n = social class

Variables in the Equation								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Low			3.311	3	.346			
	Lower middle	1.050	.862	1.484	1	.223	2.857	.528	15.473
	Upper middle	.182	.808	.051	1	.821	1.200	.246	5.844
	High	1.455	.954	2.329	1	.127	4.286	.661	27.785
	Constant	.154	.556	.077	1	.782	1.167		

a. Variable(s) entered on step 1: SE_class.

There are 3 p-values, 1 for each of the odds ratios
These are not easy to interpret, because they <u>only</u> relate to that specific level of the factor.
Also, if some are small and others big, it is difficult to understand what is happening (eg what if we had p=0.32 (Lower middle), 0.001 (Upper middle), 0.15 (High)
Might not make sense that some levels have an association with risk of hospital admission risk while other levels do not)

Therefore, what we want is one p-value for the factor 'social class'
After interpreting that p-value, we might then look at individual ones from the above table (<u>but they should be for pre-specified comparisons</u>)

SPSS output for logistic regression of
risk of hospital admission y/n = social class

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Low			3.311	3	.346			
	Lower middle	1.050	.862	1.484	1	.223	2.857	.528	15.473
	Upper middle	.182	.808	.051	1	.821	1.200	.246	5.844
	High	1.455	.954	2.329	1	.127	4.286	.661	27.785
	Constant	.154	.556	.077	1	.782	1.167		

a. Variable(s) entered on step 1: SE_class.

You can get the 1 p-value for social class from this part of the output:

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3.583	3	.310
	Block	3.583	3	.310
	Model	3.583	3	.310

This indicates no evidence of association between hospital admission and social class

- A method called a ‘likelihood ratio test’ (sometimes referred to as a ‘change in deviance’ analysis) will provide a single p-value for judging whether the factor is significant or not
- A regression model without social class (model A) is compared with a model with social class (model B).
- The test then judges if model B explains the original data better than model A.
- If it does, then social class is associated with risk of hospital admission

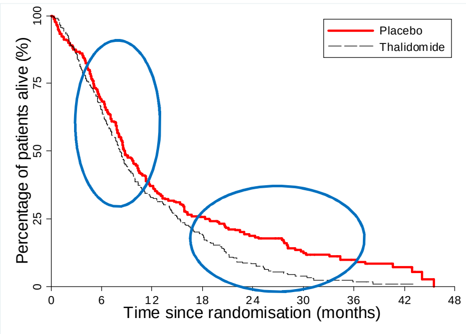
Outcome measure: time-to-event

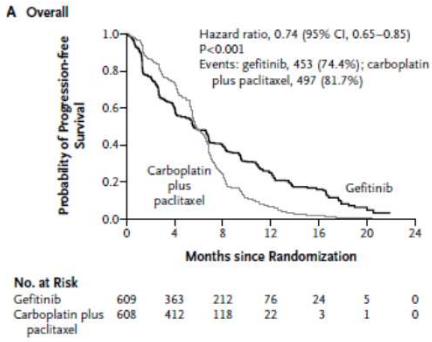
- Cox regression
- Can examine association between eg time until death with any type of variable
  - Age (‘taking measurements’)
  - Social class (‘counting people’)
- Continuous data (eg age) better if Normally distributed. If not use a suitable transformation to make it normal

- Cox’s Regression is also called the Proportional Hazards model (PH)
- It assumes that the **ratio of events** (the hazard) between 2 individuals or 2 groups is the same **over time**
- For example, if a treatment reduces the risk of dying by 15% at 6 months, then it should also reduce the risk by the same amount at 2 years (or at any other timepoint during which there is data)
- Having hazards that remain constant over time is very useful (simplifies things greatly), but may be unrealistic. Sometimes there are clear differences over time which increase or decrease the risk of an event
- If the PH assumption is very clearly violated then one approach is to look at time-dependency, ie. build into the model that the hazard for a specified variable changes over time
- There are also other methods, such as restricted mean survival times (but you should seek advice on this)

Proportional Hazards (PH) Assumption

- PH assumption is violated when the curves clearly cross each other
- Or in example below, where the curves are much more separated later on
- However, the survival model is quite robust to a moderate violation
- Need to examine the curves by eye; alternatively some software does a statistical test for the assumption
- If assumption clearly does not hold then examine risk difference at a timepoint





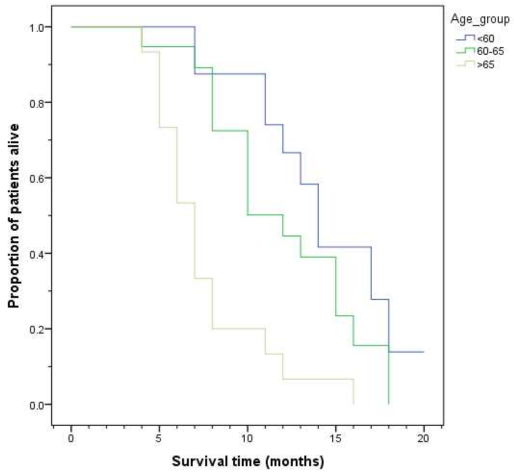
You can clearly see that the space between the 2 curves is very different

The curves clearly cross (crossing at the very end usually doesn't matter much)

Examples

- Outcome= time until death
- Three factors (called exposures, covariates, or x-variables) to be examined separately:
- Age (continuous)
- Sex (binary, ie 2 levels)
- Social class (categorical, ie 4 levels)

Look at the data first - age



Impossible to show Kaplan-Meier curve with age as a continuous factor.
Therefore, turn into say 2 or 3 categories to have a look

SPSS output for Cox regression of time until death = age

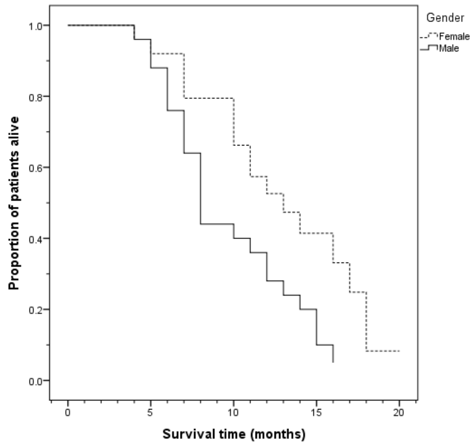
Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	.161	.035	21.562	1	.000	1.175	1.098	1.258

The log hazard ratio for age is 0.161.
But we want to work with the anti-log scale
Hence use Exp(B) = 1.175
As age increases by 1 unit (ie 1 year) the chance of dying increases by 17%

The expected true hazard ratio could be between 1.098 and 1.258
P-value for this hazard ratio is <0.0001 (under ‘Sig.’)

Look at the data first - sex



Curves are clearly separated, and % alive is greater for females (ie their chance of dying is lower)
So the Cox model should also reflect this

SPSS output for Cox regression of time until death = sex

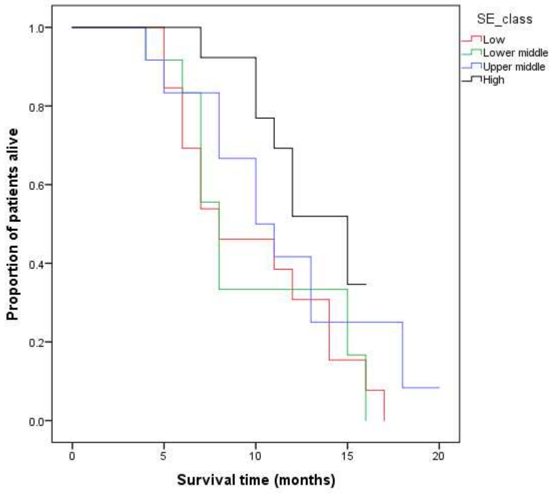
Variables in the Equation								
	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Gender(Female)	-.761	.341	4.972	1	.026	.467	.240	.912

The log hazard ratio for sex is -0.761
But we want to work with the anti-log scale
Hence use Exp(B) = 0.467
The risk of death is 0.467 (round to 0.47) times lower in females than in males
(ie the risk has been reduced by 53%)
Make sure you know which level of the factor is made the comparison
(here females vs males, not males vs females)

The expected true hazard ratio could be between 0.24 and 0.91 (wide)
The 95% CI excludes the no effect value of 1
P-value for this odds ratio = 0.026 (ie statistically significant)

There is a difference in the chance of dying between the two groups

Look at the data first – social class



Curves are generally separated, and with approximate trend

SPSS output for Cox regression of time until death = social class

Variables in the Equation								
	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
SE_class (Low)			5.167	3	.160			
SE_class (Lower middle)	-.019	.437	.002	1	.965	.981	.416	2.311
SE_class (Upper middle)	-.536	.437	1.501	1	.221	.585	.248	1.379
SE_class (High)	-.940	.472	3.969	1	.046	.391	.155	.985

One level of social class has to be the comparison group (here ‘Low’)
So, with 4 levels, there will be 3 hazards ratios

The risk of death among Lower middle is 0.98 times lower compared to Low
The risk of death among Upper middle is 0.585 times lower (42%) compared to Low
The risk of death among High is 0.391 times lower (61%) compared to Low
Looks like nice trend; higher social class, decreased risk of dying

But two 95% CIs include the no effect value 1.0, whilst one is statistically significant (High).

SPSS output for Cox regression of
time to death = social class

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
SE_class(Low)			5.167	3	.160			
SE_class(Lower middle)	-.019	.437	.002	1	.965	.981	.416	2.311
SE_class(Upper middle)	-.536	.437	1.501	1	.221	.585	.248	1.379
SE_class(High)	-.940	.472	3.969	1	.046	.391	.155	.985

There are 3 p-values, 1 for each of the hazard ratios
These are not easy to interpret, because they only relate to that specific level of the factor.
Same issue as with logistic regression

Therefore, what we want is one p-value for the factor 'social class'
And ignore the ones in the table above, for the time being

- There is a p-value for each level of social class, and it can be difficult to interpret if some are small (eg p<0.01) and others are large (eg p>0.10)
- Large p-values could be due to having only a few events in a specific group
- We need to only examine one p-value for the factor social class
- As we did with logistic regression, a likelihood ratio test could be done (for a categorical variable with ≥3 levels), for Cox regression

SPSS output for Cox regression of
time until death = social class

You can get the 1 p-value for social class from this part of the output:

Omnibus Tests of Model Coefficients[a]									
-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
246.603	5.451	3	.142	5.508	3	.138	5.508	3	.138

a. Beginning Block Number 1. Method = Enter

This p-value of 0.138 indicates insufficient evidence of a clear association between risk of dying and social class

But the hazards ratios showed a trend, which seemed clinically important.
The p-value being >0.05 is probably due to the study not being big enough

- Note that the 3 factors examined were the same in logistic and Cox
- The difference was that logistic used a binary event outcome (hospital admission or not), while Cox used time until event (how long it took the event to occur)
- But both regressions can be interpreted in a similar way
- The effect sizes (odds ratio or hazard ratio) often have a similar interpretation
- Also, the same considerations apply when converting a continuous factor into a categorical one
- And the issue of interpreting 1 p-value when a categorical factor has ≥3 levels (and try to avoid interpreting the p-value for each level of the factor)