Linear Models in Stata: Binary Logistic Regression

Dr Cameron Hurst
cphurst@gmail.com

CEU, ACRO and DAMASAC, Khon Kaen University

25th June 2556

What we will cover....

1. Basics on categorical outcomes
   - Two categorical variables and χ^2 test of independence
   - Other classical methods for categorical outcomes
2. Case study: The DMHT dataset
3. Summary statistics
4. Binary Logistic Regression using Stata
   - Data requirements
   - Running the models
5. Disseminating results from Binary Logistic regression

Conventions

Before we start, I will just point out a few conventions I will use:

Note:.....

Things to note will occur in a green box

Pitfalls:.....

Common mistakes and things to watch out for will occur in a red box

SYNTAX:....

All Stata syntax will be in purple boxes and be in `courier` font. This will help you find it easily when you have to refer back to these notes.

Motivating example

Case study 1: Cholesterol

Before we consider the various tests, let's consider the following dataset. We have 200 patients on which the following variables are measured:

1. id (Patient ID)
2. age
3. sbp (Systolic blood pressure)
4. dbp (diastolic blood pressure)
5. ses (Socio-economic status <- education and income)
6. bmi (body mass index)
7. cholesterol

Associations among categorical variables

Two categorical variables

There are a large number of methods that can be used for analysing relationships among categorical variables, we will consider three:

1. χ^2 test of independence
2. tests for (two) proportions
3. (binary) logistic regression

χ^2 test of independence

Revision: The test statistic

A test statistic (for any method) is a statistic (measure calculated from the sample) that gives us an idea whether we should (or should not) reject the null hypothesis (Ho).

For the χ^2 test of independence.....

$$\chi^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i}$$

where O = observed and E = expected cell frequencies

Typically, large test statistics suggest that we have strong evidence to reject the null hypothesis

χ^2 test of independence

Cholesterol data: BMI vs Socio economic status

- For our cholesterol data our main research question might concern the effect of body mass index on cholesterol.
- However, it is well established that many lifestyle (risk) factors can also play a role in both BMI and cholesterol levels.
- We want to investigate whether there is an association between **BMI** and socio-economic status (**SES**)
  - i.e. Is SES a potential confounder that may need to be adjusted for in any analysis concerning Cholesterol and BMI

χ^2 test of independence:Observed frequencies

Eye-balling the data

SES	*Under − Normal*	*Over − Obese*	$\sum$
Low	10	12	22
Mid-Low	16	19	35
Middle	56	47	103
Mid-High	5	13	18
High	8	14	22
$\sum =$	95	105	200

This is what the data shows us in reality, WHAT WOULD IT LOOK LIKE IF THERE WAS NO ASSOCIATION BETWEEN SES AND BMI

χ^2 test of independence

Expected frequencies:What we WOULD observe under Ho (no association)

SES	*Under – Normal*	*Over – Obese*	$\sum$
Low	10.45	11.5	22
Mid-Low	16.625	18.375	35
Middle	48.925	54.075	103
Mid-High	8.55	9.45	18
High	10.45	11.55	22
$\sum =$	95	105	200

If we compare these frequencies from those on the last slide, we can see that low and mid-low frequencies are what would be expected (under Ho), but middle, mid-high and high groups differ (somewhat) from expectation...but different ENOUGH??

χ^2 test of independence

Comparing observed and expected frequencies

- A simplistic way of gauging whether there is a relationship (between BMI and SES)would be to look at the difference between the observed and expected frequencies
- For example,there are 7.075 more middle SES, under-normal weight individuals then we would expect (if SES does not relate to BMI)
- Is this a large number? ANS: depends. 7.1 people isn't a large difference when we are talking about 1000 people, but it is when we are talking about 10.
- We need standardized measure to gauge the magnitude of the difference

χ^2 test of independence

Back to the test statistic

A χ^2 test statistic is a standard measure to gauge to magnitude of the difference (between observed and expected frequencies)

$$\chi^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i}$$

where O = observed and E = expected cell frequencies

Note: $E = \frac{Rowtotal x Coltotal}{Grandtotal}$

The expected frequencies (in the denominator) are used to 'adjust' for the number of people in each cell.

χ^2 test of independence

Partial χ^2

SES	*Under − Normal*	*Over − Obese*	
Low	0.01937799	0.01753247	
Mid-Low	0.02349624	0.02125850	
Middle	1.02310935	0.92567037	
Mid-High	1.47397661	1.33359788	
High	0.57440191	0.51969697	
		$\chi^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i} =$	5.932118

If we accumulate this standardized difference we get $\chi^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i} = 5.932118$ Is this large??

χ^2 test of independence

The p-value

- To evaluate whether the χ^2 value of 5.932118, we need to calculate the p-value
- The p-value represents the probability (chance) we would see a (standardized) difference as high as 5.93, if the null hypothesis (no association) is true.
- If such a high difference (5.93) is highly unlikely (say $<$0.05), then we would reject the null hypothesis and go with the alternative hypothesis (there is a relationship between BMI and SES).
- *In fact, for all tests, the p-value represents The chance of seeing a tests statistic so extreme, if Ho is true*

χ^2 test of independence using Stata

Prepare data and run χ^2 test of independence

Stata syntax

```
*Read in data (from comma delimited text file)
insheet using "D:\mydirectory\Cholestrol.csv", comma *Convert
BMI(contiuous) into 2 class variable BMIclass
egen byte BMIclass = cut(bmi), at(0,25,100) icodes
*Create labels for new BMIclass
label define BMInames 0 "Normal" 1 "Overweight"
*Associate labels with BMIclass
label values BMIclass BMInames
*Same for SES
label define SESnames 1 "Low" 2 "Midlow" 3 "Mid" 4 "MidHigh" 5
"High"
label values ses SESnames
*Generate table and Chi2

tabulate ses BMIclass, chi2
```

χ^2 test of independence(Stata)

Running this code gives.....

ses	BMIclass Normal	Overweigh	Total
Low	10	12	22
Mid low	16	19	35
Middle	56	47	103
Mid High	5	13	18
High	8	14	22
Total	95	105	200

Pearson chi2(4) = 5.9321 Pr = 0.204

χ^2 test of independence

Problem with the χ^2 test of independence

- The problem with the χ^2 test is that it gives us a yes/no answer
- It doesn't tell us much about the direction or magnitude of the association
- For this we have to go to other tests. In particular:
  - Test for difference between two proportions
  - (Binary) Logistic regression (and odds ratios)
- Both of these tests are specifically for binary (e.g. Yes/No) outcomes

Other tests for categorical outcomes

Measures of association for Binary outcomes

Before we consider these tests we should revise different measures of association for binary outcomes:

- Difference in proportions $(p_1 - p_2)$
- Risk ratios (RR)
- Odds ratios (OR)

Consider the table below:

	Outcome			
Exposed	yes	no	Total	Risk of outcome
yes	a	b	a + b	a/(a + b)
no	c	d	c + d	c/(c + d)
Total	a +c	b + d	a + b + c + d	

Other tests for categorical outcomes

Measures of association for Binary outcomes

Difference in proportions (prevalence)

$$p_1 - p_2 = \frac{a}{a+b} - \frac{c}{c+d}$$

Risk Ratio

$$RR = \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{c}{c+d}\right)}$$

Odds Ratio

$$OR = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)}$$

A reduced dataset

Cholesterol data: BMI in Low and Middle income families

Let's subset our data, and consider only the Low and Middle SES groups:
This gives us a new table with 125 individuals:

SES	*Under – Normal*	*Over – Obese*	*sum*
Low	10	12	22
Middle	56	47	103
sum	66	59	125

We will consider SES the "exposure" and BMI the "outcome"

Other tests for categorical outcomes

Difference in proportions (prevalence)

$$p_1 - p_2 = \frac{a}{a+b} - \frac{c}{c+d} = \frac{12}{22} - \frac{47}{103}$$
$$= 0.5454545 - 0.4563107 = 0.08914387$$

The prevalence of being overweight-obese is approx 9 % higher in low SES groups compared to the middle SES group

Other tests for categorical outcomes

Difference in proportions (prevalence)

Difference in proportions (prevalence) is a nice measure of association as:

- It is readily interpretable
- It gives a idea of effect size (unlike χ^2 test)

However, also has a major disadvantage:

- As an ABSOLUTE measure, the magnitude of the effect depends on the (baseline) prevalence.

Difference between two proportions

For example, consider two cases of a randomized controlled trials of a new therapy. In both cases prevalence is reduced by 5

- In the first case prevalence in control is 6 % and in treatment group is 1 %
- In the second case prevalence in control is 56 % and in treatment group is 51 %

Do you think this is the same effect size??????

Test for difference between two proportions

The test statistic for a test of difference between proportions is:

$$t = \frac{p_1 - p_2}{S_p}$$

where S_p is the standard error of the pooled sample proportions given by:

$$S_p = \sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}$$

And the pooled estimate of the sample proportions is:

$$p = \frac{(p_1 n_1 + p_2 n_2)}{(n_1 + n_2)}$$

Test for difference between two proportions

Confidence interval: A better way of hypothesis testing

- A more meaningful way of testing hypothesis (than using and p-values), is using confidence interval (CI)
- The CI gives us a 'feasible range' for the true (population) value of our statistic of interest
- In this case we want the true (population) difference in proportion between our two groups

Test for difference between two proportions

Confidence interval for the difference between two proportions

The Confidence Interval for $p_1 - p_2$ is:

$$p_1 - p_2 \pm k \times S_p$$

where k is governed by the levels of confidence
For 99% confidence $k = 2.326$ Corresponds to p<0.01
For 95% confidence $k = 1.960$ Corresponds to p<0.05
For 90% confidence $k = 1.645$ Corresponds to p<0.1
For example the 95% CI would be:

$$[p_1 - p_2 - 1.96 \times S_p, p_1 - p_2 + 1.96 \times S_p]$$

Tests for difference in proportions

Cholesterol data: BMI in Low and Middle income families

Recall our example comparing prevalence of over-obese people from low and middle income groups

$$p_1 - p_2 = \frac{12}{22} - \frac{47}{103} = 0.5454545 - 0.4563107 = 0.08914387$$

We want to know whether there is a significant difference in prevalence of obesity.

- H_0: There is no difference in the prevalence of overweight-obesity between low and middle income earners.
- H_A: Prevalence differs between low and middle income earners.

Tests for difference in proportions: Using Stata

The Stata code for a tests (and confidence interval) for a difference in proportion is:

Stata syntax

```
*Input total number, and total obese in each group

prtesti 22 12 103 47, count
```

Two-sample test of proportion — x: Number of obs = 22; y: Number of obs = 103

Variable	Mean	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x	.5454545	.1061589			.3373868	.7535223
y	.4563107	.049078			.3601195	.5525018
diff	.0891439	.1169546			-.1400829	.3183706
	under Ho:	.11725	0.76	0.447		

Tests for difference in proportions: Interpretation

First (from p-value)

- we can see $p = 0.447$ (not < 0.05) we cannot reject H_0.
- i.e. conclude no (statistically) significant difference in prevalence of obesity between income groups

Second (using confidence intervals)

- we have a (very wide) 95% confidence interval of $[-0.14, 0.318]$
- likely the population difference in prevalence of obesity between -14% and +32% (from Stata output)
- From this we cannot rule out a difference of 0% as a feasible value for the true difference in prevalence

Measures of association

Risk ratios and Odds ratios

Recall the other two measures of association we have discussed

Risk Ratio

$$RR = \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{c}{c+d}\right)}$$

Odds Ratio

$$OR = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)}$$

Measures of association

Risk ratios and Odds ratios

- advantage of the RR and OR (over diff in proportions) is they are relative
- i.e. they account for the baseline prevalence:
  EG
  - RR and OR would show a lower association where 5% prevalence difference from control = 56% and treatment = 51%
  - RR and OR would show a higher association where 5% prevalence difference from from control = 6% and treatment = 1%

RR and OR would show a lower association

Measures of association

Risk ratios

The relative risk (of obesity) in the lower group relative to the middle income group is:

$$RR = \frac{(\frac{a}{a+b})}{(\frac{c}{c+d})} = \frac{\frac{12}{22}}{\frac{47}{103}} = \frac{0.5454545}{0.4563107} = 1.195358$$

This means, (in our sample) the risk of obesity in lower income individuals is 1.2 times that of middle income individuals. Does this seem like a substantial increase in risk to you??

Measures of association
Risk ratios

The relative risk has some advantages:

- it is a relative measure
- it is intuitive

Unfortunately it has some mathematical properties that make it difficult to model (especially when we want to account for more than one risk factor etc.)
Also, the RR should not be used in case-control studies (where the sample 'prevalence' is artificial)

Measures of association

Odds ratios

Finally we come to the odds ratio.

- not as intuitive as the RR and difference in proportions
- but has desirable mathematical properties
- In fact, ORs are the measure of association underpinning Logistic Regression

In our sample

$$OR = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)} = \frac{\frac{12}{10}}{\frac{47}{56}} = \frac{1.2}{0.8392857} = 1.429787$$

Measures of association

Interpretation: Odds ratios

How do we interpret $OR = 1.429787$

- The **ODDS** of obesity in lower income individuals is 1.43 times the odds of obesity in middle income earners
- Note: That an OR is always more extreme than RR (except when prevelance in both groups is low)

Alerts

Be careful with words like chance, likelihood, probability and risk when discussings odds ratios. Note that the odds of an event is different from the probability (risk) of an event

Moving towards a statistical 'model'

- All of the above methods have the advantage of being (comparitively) intuative.
- Unfortunetly (like t-test etc) they are naive; They assume we live in a 'bivariate universe'
- As soon as we consider real-life, observational data. All of the above methods are rendered inaccurate and misleading
- We need a multivariable modelling approach to consider:
  1. Other (independnat) risk factors;
  2. Confounders
  3. Effect modifiers

Let's quickly consider this problem in terms of our DMHT data

DMHT: Study background

- Collaborative clinical study supported by the Thailand National Health Security Office (NHSO) and the Thailand Medical Research Network (MedResNet)
- Official title: An Assessment on Quality of Care among Patients Diagnosed with Type 2 Diabetes and Hypertension Visiting Ministry of Public Health and Bangkok Metropolitan Administration Hospitals in Thailand (Thailand DM/HT)
- In short, main research objective is to assess quality of care of (Type 2) Diabetic and Hypertenisve patients in Thatiland
- At present the study involves about 60000 patients from all across Thailand, sampled from 2553 to 2555
- Three main groups consiidered: (1) Patients with Type 2 Diabetes Mellitus (T2DM), (2) Patients with Hypertention (HT), and (3) Patients with both (DMHT)

(Our) research questions

For the purpose of our excerises today, our research questions are:

1. Among diabetics, does the hypertentsion comorbidity represent an additional burden for achieving clinical (quality of care) goals?
2. What other diabetes patient characteristics (e.g. demographic/lifestyle) might influence achieving the clinical goals?

Data we will consider

- We will consider a random sample of 5000 patients (sampled in 2554), and only those pateints with Diabetes Mellitus [i.e. (1) T2DM and (3) DMHTs]
- Of the hundreds of variables measured in the study, we will only consider a subset:

Outcomes	**Study effect**	**Other covariates**
a1cyn	ht	sex
bpyn		age
ldlcyn		religion
all3yn		duradm
any3yn		smoke
		bmigroup

Detailed

description of variables next few slides...

Variable description: Outcomes

The outcomes in our 'subset' of the dataset are the "ABC" clincal goals often used to assess the quaility of diabetes care. These are:

(a) `a1cyn`→Hemoglobin **A**1C: yes: $< 7\%$; no: $\geq 7\%$

(b) `bpyn`→**B**lood pressure: yes: $< \frac{130}{80}$ mmHg; no: $\geq \frac{130}{80}$ mmHg;

(c) `ldlcyn`→Low density lipid-**C**holesterol: yes: LDL-C < 100 mg/dL; no: LDL-C≥ 100 mg/dL

We will also consider the collective quality performance outcomes:

- `all3yn`: **All** three (ABC) clinical goals are met: yes; no
- `any3yn`: **Any** of the three (ABC) goals are met: yes; no

Question: What type of measurement scale do all of these 5 variables have?

Study effect

We are interested in whether there is any difference in the achievement of treatment goals between those with diabetes alone (T2DM) and those diabetics who ALSO have hypertension; Does a hypertention comorbidity represent an additonal burden specifically in terms of diabetes care.

We will use the variable, `ht` to measure this: no (T2DM alone); yes (DM+HT)

Note: Study Effects

Remember a **study effect** is the explanatory variable (X) that is of **primary interest** (in terms of our research hypothesis)

Covariates

In observational studies (such as the DMHT study), many other explanatory (X) variables need to be considered. I will make a dinstiction between two different types of covariates here:

1. Independant risk factors: Other important independnant explanatory variables of the outcome
2. Confounders: Variables that are associated with BOTH the outcome AND the study effect which, if ignored, can misleadingly enhance/diminish the true relationship between the otucome (Y) and the study effect (X)
3. Effect modifier: Are the levels of association between our outcome and our study effect the same for different LEVELS of a third variable

Covariates

In our case study, I have included 6 covariates which can be split into two different groups:

- Demographic variables"
  - `sex`: Patient gender (binary)
  - `age`: Age in years (continuous)
  - `religion`: Buddhist/Non-buddhist (binary)
- Lifestyle/patient history variables:
  - `duradm`: Duration of T2DM; How long (years) since patient diagnosed with T2DM (continuous)
  - `smoke`: Smoking history (Nominal/Ordinal): Current, Never, Previous, Unknown
  - `bmigroup`: Underweight, Normal, Overweight, Obese (Ordinal)

Bivariate analysis

Cross tabulations of two categorical variables

To generate a cross-tabulation of two catergical variables, we simply add another variable to the `tab` command:

SYNTAX: X-tabs for two categorical variables

```
 tab all3yn ht
*Put col or row after ',' to get percentages
*Eg:
tab all3yn ht, row
```

RECODE of all3g	ht patient		
	DM alone	DM and HT	Total
0	80	140	220
1	977	2,057	3,034
Total	1,057	2,197	3,254

RECODE of all3g	ht patient		
	DM alone	DM and HT	Total
0	80	140	220
	36.36	63.64	100.00
1	977	2,057	3,034
	32.20	67.80	100.00
Total	1,057	2,197	3,254
	32.48	67.52	100.00

Using syntax in Stata

Remember when using Stata (or any other) syntax:

1. Very few people remember syntax the first time (you are not expected to memorize it): Use these notes as a reference manual
2. Syntax means a little pain now for a BIG payoff later
3. The more use it, the easier it is to remember
4. Always document your work (using comments) to remember what your syntax does (you will often come back to it later)

Logistic regression in Stata

The good news is that logistic regression is very easy in Stata. For a Bivariate association we will use:

SYNTAX: Binary logistic regression with a single predictor

```
logistic myy myx
```

...and for a multivariable (more than one predictor), we will use:

SYNTAX: Multivariable logistic regression

```
logistic myy myx1 myx2 myx3 etc etc
```

Logistic regression in Stata

Mostly, people are interesrted in the odds ration (OR) in logistic regression, but sometimes it may be useful to get the raw model coeffcients (log odds ratio), to this we just ask for them in the options

SYNTAX: Binary logistic regression options

```
logistic myy myx, coef
```

Other options can be specified (after the ',') in this way too

Data requirements for logistic regression

You should have covered this in previous sessions, but I will repeat here that in Binary logistic regression:

- The outcome in **Binary** Logistic Regression must be a **Binary** categorical variable
- Predictors can be either Categorical or Continuous (but we interpret their odds ratios a little differently)
- Stata has a VERY SPECIFIC way of understanding whether a predictor is categorical (the "i." prefix), or continuous (no prefix) (but we need the "c." prefix if we are considering an interaction effect)

Our first logistic regression in Stata

Recall our dmht5000 dataset. We will investigate the patient characteristics associated with acheiving all three of the "ABC" treatment goals. Specifically:

- A single (binary) outcome: **all3yn** (yes[0], or no[1])
- Our study effect (binary), **ht**: (DM alone[0], DM+HT[1])
- Two covariates:
  - **age** (Continuous)
  - **sex** (male[0], female[1])

Logistic regression in Stata

First, we'll investigate the crude effects of `ht`, `age` and `sex`

SYNTAX: Models to get crude ORs

```
 *Study effect
logistic all3yn i.ht

*covariates
logistic all3yn age
logistic all3yn i.sex
```

Pitfalls: Categorical predictors in Stata

Note the **i.** in front of the categorical predictors (this tells Stata that these variables are categorical)

We will run these analyses and talk about the results later

Logistic regression in Stata

Now let's put all three predictors into a single multivariable model

SYNTAX: Full (adjusted) model

```
 *Full model
logistic all3yn i.ht age i.sex
```

Note: To adjust, or not to adjust

The Bivariate model (`all3yn` vs `ht`) provides a **CRUDE** estimate of the association between these two variables, whereas the multivariable model (above) provides an (age and sex) **ADJUSTED** Odds ratio. That is, age and sex differences between the T2DM and DMHT groups are statistically removed. We will talk more about this soon.

Stata output: all3yn vs ht

The output from the first logistic regression model is given below.

```
Logistic regression                             Number of obs   =       3254
                                                LR chi2(1)      =       1.59
                                                Prob > chi2     =     0.2071
Log likelihood =  -804.2765                     Pseudo R2       =     0.0010

------------------------------------------------------------------------------
      all3yn | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        1.ht |     1.2031   .1749778     1.27   0.204     .9046981    1.599925
       _cons |    12.2125   1.420201    21.52   0.000      9.72339     15.3388
------------------------------------------------------------------------------
```

For the moment, just note the output (we will discuss the results later)

Stata output: all3yn vs age

```
Logistic regression                               Number of obs   =       3254
                                                  LR chi2(1)      =       4.72
                                                  Prob > chi2     =     0.0299
Log likelihood = -802.71347                       Pseudo R2       =     0.0029

------------------------------------------------------------------------------
      all3yn | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9860409   .0064054    -2.16   0.030     .9735661    .9986754
       _cons |   32.18035    12.9453     8.63   0.000     14.62767    70.79563
------------------------------------------------------------------------------
```

Stata output: all3yn vs sex

```
Logistic regression                               Number of obs   =       3254
                                                  LR chi2(1)      =       0.42
                                                  Prob > chi2     =     0.5158
Log likelihood = -804.86111                       Pseudo R2       =     0.0003

------------------------------------------------------------------------------
      all3yn | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |    1.10303   .1655651     0.65   0.514     .8219056    1.480311
       _cons |   12.88571   1.598782    20.60   0.000     10.10407    16.43314
------------------------------------------------------------------------------
```

Stata output: full model

Logistic regression — Number of obs = 3254
LR chi2(3) = 8.59
Prob > chi2 = 0.0353
Log likelihood = -800.77845 — Pseudo R2 = 0.0053

all3yn	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.ht	1.330818	.2003319	1.90	0.058	.9907972	1.787526
age	.9830933	.0065768	-2.55	0.011	.9702872	.9960684
2.sex	1.098842	.1652317	0.63	0.531	.8183539	1.475467
_cons	29.92411	12.45339	8.17	0.000	13.23672	67.64915

Presenting the results

Obviously, we can't present the raw output in a manuscript or thesis. Let's use our output to generate some tables that provide in our reports. I will present two tables:

1. A table with crude associations (ORs)
2. A table with both crude and adjusted ORs

In a thesis, I might present both tables, but in many publications (where space is at a premium), I might only present the second(depending on the journal, or the importances of differences between crude and adjusted estimates to the 'theme' of the paper)

Table of crude results

Table : Crude effects of Hypertention, Age and Sex on achievement of ABC treatment goals

Effect	OR_{Crude}	**95%CI**
Hypertension	1.203	0.905-1.600
Age(years)	0.986*	0.974-0.999
Sex(Males)	1.102	0.822-1.480

* p<0.05; **p<0.01; ***p<0.001

Interpretation:

- We could not demonstrate an effect of the hypertentsion comorbidity on acheiving diabeties treatment goals ($OR_{ht} = 1.203$, $95\% CI : 0.91 - 1.6$)
- Age can be shown to influence the achievment of treatment goals with the odds of achievment drecreasing by 1.4% with every extra year older a patient got ($OR_{Age} = 0.986, 95\% CI : 0.974 - 0.999$)

Table : Effects of Hypertention, Age and Sex on achievement of ABC treatment goals

Effect	OR_{Crude}	$OR_{Adusted}$	**95%CI**
Hypertension	1.203	1.331	0.991-1.787
Age(years)	0.986*	0.983*	0.970-0.996
Sex(Males)	1.102	1.099	0.818-1.475

* p<0.05; **p<0.01; ***p<0.001

- Although adjusting for age did increase the assocation between ht and acheievement of ABC goals ($OR_{Crude} = 1.203$ vs $OR_{Adusted} = 1.331$) we could still not show a significant assocation ($95\% CI : 0.991 - 1.787$).....although nearly (p=0.058)
- The substantial change between the crude and adjusted effects of ht suggest that sex and/or age did have a confounding effect
- Age (but not Sex) was still shown to have an effect of achieving goals ($OR_{Age} = 0.983, 95\% CI : 0.97 - 0.996$)

THANK-YOU!!

Questions??

YOUR TURN

Exercise

We will retain the hypertension comordbidity as the **Study effect**, but this time:

- Choose one of the 3 individual treatment goals as your outcome (A: a1cyn; B: bpyn; or C: ldlcyn). You could even cross tabulate ht with each to see which is the msot promising. AND
- Choose any of the covariates you like. The only suggestion I have is that you consider AT LEAST TWO (One continuous and one categorical). You could even consider all six if you like.

DON'T FORGET TO...

1. Use syntax and a do file
2. Save you work
3. Document your do files (USE COMMENTS MAK MAK!!!!)