

POSTGRADUATE COURSE IN
LINEAR AND LOGISTIC REGRESSION

Home work

The home exercise is divided into two. The first part is a follow-up on Day 1 and 2, and the second part is an introduction to Day 3 and 4. Datasets are available at the homepage.

Part One: Multiple regressions.

Consider the dataset: “serumcho1.dta”, which is a subset of the dataset

2.20 .fr a.minghamb.dta used in Dupont. In this exercise, focus is on the dependent variable serum cholesterol (sc1) and possible explanatory variables systolic blood pressure (sbp), diastolic blood pressure (dbp), Body Mass Index (bmi), and sex (sex=1 men, sex=2 women). See Dupont chapter 2 and 3 for help and inspiration.

1.1 Create a categorical variable from bmi according to the WHO definitions by

`egen bmi_who=cut(bmi), at(10, 18.5, 25, 30, 60) label`

Is the distribution of sc1 the same in the groups defined by bmi_who?

(Hint: make a box plot or histogram of sc1 for each category of bmi_who).

1.2 Estimate a model (**Model 1**) with sc1 as the dependent variable and sbp, bmi_who and sex as the independent variables; bmi_who and sex should be entered as a categorical variable in the model with $BMI < 10 \text{ kg/m}^2$ and men as reference.

(Hint: use the prefix `xi` : to the `regress` command).

1.3 Explain the coefficients for sbp, bmi_who=3 and sex=2.

Make a plot of the relationship between the expected serum cholesterol and systolic blood pressure for the eight different combination of sex and bmi_who.

Make a plot of the relationship between the expected serum cholesterol and BMI for a man with systolic blood 130 mmHg.

1.4 Find the expected value with 95% confidence interval for a subject with $sbp=85$, $sex=2$ and $bmi_who=1$

(Hint: make the reference group equal to subjects with the values given above).

Next we will focus on the independent variables sbp and dbp.

1.5 Estimate 3 models (**Model 1, 2, 3**) with sc1 as the dependent variable and the following independent variables:

Model 1: sbp, bmi_who, sex

Model 2: sbp, dbp, bmi_who, sex

Model 3: dbp, bmi_who, sex

Compare the coefficients and standard errors for sbp in **Model 2** and **Model 1**.

Compare the coefficients and standard errors for dbp in **Model 2** and **Model 3**.

What do you see? And why does this happen?

[Back to Model 1](#)

1.6 Create a new variable sbp^2 equal to the square of sbp.

Add sbp^2 to **Model 1** and estimate this model (**Model 4**).

Explain the coefficient of sbp^2 .
Find the expected value for sc1 with 95% confidence interval for a subject with values given in 1.4. Compare the result with the one you found in 1.4.

1.7 Estimate a model (**Model 5**) with sc1 as the dependent variable and sbp, sex and bmi as independent variables (that is, BMI in a non-categorized version).

Make a plot of the relationship between the expected serum cholesterol and BMI for a man with systolic blood 130 mmHg. Decide from this and the estimates (from **Model 1** and **5**) whether BMI as a continuous variable is preferable to as good as BMI as a categorized according to WHO.

We stick with **Model 1**.

1.8 Use the *explanatory* variables in **Model 1** as a basis for an investigation of whether the *dependent* variable would benefit from a transformation. (Hint: plots of the distribution of the residuals, residual versus fitted values, and residual versus independent variable should be made.)

In order to estimate a more realistic model possible interactions should perhaps be included.
Here we focus on two: an interaction between sex and sbp and an interaction between sex and bmi_who.

1.9 Estimate a model (**Model 6**) with $\ln(sc1)$ as the dependent variable and sbp, bmi_who, sex and both interactions as independent variables. (Hints: the interactions can be create as dummy variables or by including the products $i.sex*sbp$ and $i.sex*i.bmi_who$ as independent variables in the `regress` command)

1.10 Explain the coefficient to the interaction between sex and sbp.
Test the hypothesis that the interaction is zero.

1.11 Explain the coefficient to the interaction between sex and bmi_who=2.
Test the hypothesis that all coefficients to the interaction between sex and bmi_who are zero. (Hints: use the command `testparm`)

Part Two: Logistic regression.

First a short introduction to or a reminder of *case-control* studies:

Here, let a *case* denote a subject with the disease of interest (e.g. esophageal cancer), a *control* a subject without the disease, and let *exposure* status denote whether the subject has experienced a possible cause (e.g. severe smoking) of the disease or not. The case-control study is often a retrospective study i.e. both cases and controls are asked if they have been exposed at some point in their life. The interest is on estimating an *association* between the exposure and the disease. The preferable measure of association, the relative risk (RR, i.e. the risk of getting the disease among the exposed divided by the risk of getting the disease among the non-exposed) cannot be estimated because the cases together with the controls are not representative for the population. But when the disease is rare then the odds ratio (OR) is an approximation to RR.

For further information, see Bland p. 37-40 and Dupont p. 131-133.

2.1 Read parts 4.1.9.1 and 4.1.9.2 in Dupont.

The formulas for the estimated OR and 95% confidence interval are listed below the table.

exposed	case	control	total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	N
$OR = \frac{odds_{case}}{odds_{control}} = \frac{a/c}{b/d} = \frac{a*d}{c*b}$			
$se(\ln OR) = \sqrt{1/a + 1/b + 1/c + 1/d}$			
$95\% CI(\ln OR) = \ln OR \pm 1.96 * se(\ln OR)$			
$95\% CI(OR) = \exp(\ln OR \pm 1.96 * se(\ln OR))$			
$odds = \frac{p}{1-p}, p = \frac{odds}{odds+1}$			

A 2x2 table with esophageal cancer and severe smokers in age group 45-54 years:

Age 45-54 years

		severe smoker	
		0-19 gm/day	>= 20 gm/day
		total	
cancer	no	88	28
cancer	yes	1	1
	total	89	29
		118	

35-44:

		severe smoker	
		0-19 gm/day	>= 20 gm/day
		total	
cancer	no	149	41
cancer	yes	6	3
	total	155	44
		199	

45-54:

		severe smoker	
		0-19 gm/day	>= 20 gm/day
		total	
cancer	no	134	33
cancer	yes	27	19
	total	161	52
		213	

OR=2.86
95% CI=(1.42, 5.75)

55-64:

		severe smoker	
		0-19 gm/day	>= 20 gm/day
		total	
cancer	no	134	32
cancer	yes	48	28
	total	182	60
		242	

OR=2.44
95% CI=(1.33, 4.47)

2.1 Calculate, by hand, the OR with 95% confidence intervals from the above table.

2.2 Use the dataset `case-control.dta`.

Run command `cc cancer smoker if age==3, woolf` in STATA.

What do you get?

Run command `logistic cancer smoker if age==3` in STATA.

What do you get?

This was the OR in age group 45-54 years.

A crude/unadjusted OR not adjusting for age, can be found using same two commands without "if age==3".

Do this!

In the example above age is a possible confounder, and we wish to estimate an OR adjusted for age.

The classical method for obtaining such an estimate is the Mantel-Haenszel method.

For the 6 age groups the data summarize to the tables given below.

65-74:

		severe smoker		total	
		0-19 gm/day	≥ 20 gm/day		
cancer	no	94	12	106	OR=2.19
	yes	43	12	55	95% CI=(0.91, 5.26)
total	137	24	161		

≥ 75:

		severe smoker		total	
		0-19 gm/day	≥ 20 gm/day		
cancer	no	26	5	31	OR= 0.95
	yes	11	2	13	95% CI=(0.16, 5.63)
total	37	7	44		

2.4 Write the following command in STATA:

cc cancer smoker, by(age) wcoolf

Find the OR and 95% CI for each age group in the output.

2.5 Find the crude OR in the output.

What is the adjusted (Mantel-Haenszel) estimate

Compare the crude and the MH estimate of OR.

Is age a 'confounder' variable for which there should be adjusted?

2.6 Write the following commands in STATA:

xi: logistic cancer smoker i.age

Compare the coefficient for cancer with Mantel-Haenszel estimate.