In the first part of this session we will look at a regression model that includes an interaction between the explanatory variables.

The second part (optional) covers fitting and interpreting a quadratic regression model.

We will use a dataset called armitage.dta

The aim of our analysis in this session is to investigate whether vital capacity among workers exposed to cadmium fumes differs from that in unexposed participants. Vital capacity is the maximum volume of air that can be exhaled after maximum inhalation.

In this dataset people are divided into three groups. Group 1 are workers not exposed to cadmium fumes. Group 2 are workers exposed for less than 10 years and Group 3 are workers exposed for 10 or more years. Respiratory function is known to decline with age and so age is a potential confounder in this analysis.

The dataset is available Moodle in the folder "Linear Regression Datasets"

The variables are as follows:

group Group (1 = not exposed; 2 = exposed <10 years; 3 = exposed for 10+ years)

age Age in years

vitcap Vital capacity in litres

# Initial analysis

1) Start a new Do file, open the dataset, and explore each of the variables.

2) Produce some plots or tables to investigate the relationship between vital capacity and each variable. Note whether it seems that mean vital capacity differs by group or by age.

# ANOVA 1

1) First create three dummy variables to identify Group membership.

2) Use your dummy variable to fit a linear regression model comparing vital capacity between the groups.

3) Use the i. prefix to fit the same linear regression model. Compare the output to convince yourself that these models are identical. Fill in X with these options in this sentence (0.037, 0.090, 0.01, 0.96, -0.51): The global F-test p-value is X. The difference in mean vital capacity between group 2 and group 1 is X litres, p = X. The difference in mean vital capacity between group 3 and group 1 is X litres, p = X

4) Fit a model to compare vital capacity between the groups, controlling for age.

5) Use Stata to conduct a partial F test on the two group coefficients. Write up the results in a couple of sentences

The ANOVA comparing vital capacity from betwen the groups adjusting for age can be written as

Vi=α+β1g2i+β2g3i+β3agei+ϵi

Where,

Vi is vital capacity for the ith person  
g2i=1 if ith person is in group=2 (exposed <10 years) and 0 otherwise  
g3i=1 if ith person is in group=3 (exposed 10+ years) and 0 otherwise  
agei is age of the ith person in years

Write down the equations for groups 1, 2, 3

# Pairwise differences

1) Write down the difference in mean vital capacity holding age constant between:

* Group 3 and group 1
* Group 2 and group 1
* Group 3 and group 2

2) Note that the Stata regress output shows the confidence intervals for only the first two comparisons. How could we calculate a confidence interval for the difference between group 3 and group 2? You’ll need the commands lincom, and possibly using bx. to select which group (x value) to be the baseline. Also, try calculating the confidence interval by hand, using the covariance matrix created by mat list e(V) from a regress command

3) plot the data & fitted line for each group, is the fit good?

# Interactions

To demonstrate the model we will manually create interaction variables. This shows how Stata fits the interaction for us when we use the i. prefix.

1) Make three variables for the interaction between age and the dummy variable for each group

gen g1age = dummy1 \* age

gen g2age = dummy2 \* age

gen g3age = dummy3 \* age

2) Use these variables to fit a regression model where the effect of age differs between groups:

regress vitcap dummy2 dummy3 age g2age g3age

3) This is equivalent to the below Stata command. Run this model and check you understand how the output compares.

regress vitcap i.group age i.group#c.age

The "#" tells Stata to include variables for the interaction between each level of group and age. Note the "c." before age. When you ask Stata to interact two variables it assumes they are both categorical. To tell Stata that age is a continuous variable, we add the "c.". We don't need to include the "i'" for group, but it is a good habit to include both "i." and "c." when you run an interaction to avoid any errors.

Write down the overall model equation, and the equations for each group

# Quadratic terms & centering

We will now run a regression model including a quadratic term to investigate if there is evidence for non-linearity in the relationship with age in the group who were not exposed to cadmium fumes

1) Create a new variable that is the square of age

gen agesq = age^2

2) Run the quadratic regression model in the unexposed group

regress vitcap age agesq if group==1

Interpret the output

In the last model the coefficient for age did not have a meaningful interpretation. We can centre age in the model to give a more interpretable output.

3) Create two new variables: age centred on the mean in group 1, and the square of the centred age

sum age if group==1

gen cage = age - r(mean)

label var cage "Centred age (around the mean for group 1)"

gen cagesq = cage^2

label var cagesq "Squared centred age"

4) Run the quadratic regression model in the unexposed group using the centred variables

regress vitcap cage cagesq if group==1

interpret the centered age coefficient