Linear Modelling in Stata

Session 6: Further Topics in Linear Modelling

Contents

1. Categorical Predictors in Linear Models

So far we have considered models in which both the outcome and predictor variables are continuous. However, none of the assumptions of the linear model impose any conditions on the x-variables. The predictor variables in a linear model can have any distribution. This makes it possible to include categorical predictors (sometimes referred to as *factors*) in a linear model.

1.1. A Dichotomous Variable

Suppose that we are involved in a clinical trial, in which subjects are given either an active treatment or placebo. We have an outcome measure, Y, that we wish to compare between the two treatment groups. We can create a variable, x, which has the value 0 for the subjects on placebo and 1 for the subjects on active treatment. We can then form the linear model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

In this model, β_0 is the expected value of Y if $x = 0$, i.e. in the placebo group. β_1 is the expected increase in Yg as x increases by 1, i.e. the expected difference between the placebo and active treatment groups.

Does the above model satisfy the assumptions of the linear model ? Well, the mean of Y is certainly a linear function of x. Whether the individual observations of Y are independent depends on the experimental design, but if they all represent different individuals they should be independent. Whether the variance of Y is constant remains to be seen, but since x can only take two values in this case, it amounts to the variance of Y in the placebo group being the same as the variance of Y in the active treatment group. Finally, a linear model would only be appropriate if Y followed a normal distribution. Again, this needs to be tested, but the assumptions of the linear model can be met as easily for a dichotomous predictor as for a continuous one.

Here is an example of using a linear model in the above scenario. The data used for this example was simulated, using the commands

```
set obs 40
gen x = _n > 20
gen Y = 3 + x + invnorm(uniform())
```

This creates a set of 40 observations, 20 with x=0 and 20 with x=1. Y is normally distributed with variance 1, and has a mean of 3 if x=0 and 4 if x=1.

When I analysed this data, using `regress Y x`, I got the following output[a]:

[a] Note that the random number generator is used in generating Y through the function `uniform`, so if you repeat this analysis you will not get exactly the same results, but they should be similar.

```
. regress Y x

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =   10.97
       Model |  9.86319435     1  9.86319435           Prob > F      =  0.0020
    Residual |  34.1679607    38   .89915686           R-squared     =  0.2240
-------------+------------------------------           Adj R-squared =  0.2036
       Total |   44.031155    39  1.12900398           Root MSE      =  .94824

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .9931362   .2998594     3.31   0.002     .3861025     1.60017
       _cons |     3.0325   .2120326    14.30   0.000     2.603262    3.461737
------------------------------------------------------------------------------
```

What do we conclude from the above output ? Firstly, we can say that x is a significant predictor of Y ($p = 0.002$), which in this case means that the outcome differs significantly between the placebo and active treatment groups. The estimated difference between the two groups is 0.993, with a 95% confidence interval of (0.39, 1.60), which is close to the true value of 1. The mean value among the placebo group is 3.03, with a confidence interval of (2.60, 3.46), which is again close to the true value of 3.

Note that the values that x takes need not be 0 and 1: it can take any two values. However, the interpretation of the coefficients depends on the values that it takes. β_0 is the expected value of Y when $x = 0$, so if the two groups were given x-values of 1 and 2, β_0 would not correspond the mean value of Y in either of the groups. For this reason, dichotomous variables are always best coded as 0 and 1.

1.1.1. T-Test

We saw previously that the way to test for a difference in a normally distributed outcome between two groups is to use a t-test. In fact, the above analysis is *exactly* equivalent to a t-test, as the following stata output shows.

Note that the difference between the two groups and its standard error are *exactly* the same as in the linear model above, and consequently the p-value from the test the the difference is not equal to 0 is also exactly the same. Hence we can say that the t-test is a special case of a linear model. It requires the same assumptions and leads to the same conclusions. However, the linear model has the advantage that it can incorporate adjustment for other variables.

```
. ttest Y, by(x)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      20      3.0325    .2467866    1.103663    2.515969     3.54903
       1 |      20    4.025636    .1703292    .7617355    3.669133    4.382139
---------+--------------------------------------------------------------------
combined |      40    3.529068    .1680033    1.062546    3.189249    3.868886
---------+--------------------------------------------------------------------
    diff |           -.9931362    .2998594               -1.60017   -.3861025
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -3.3120
Ho: diff = 0                                     degrees of freedom =       38

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0010         Pr(|T| > |t|) = 0.0020          Pr(T > t) = 0.9990
```

1.2. A Categorical Variable with Several Categories

.

The above method works if there only two categories, but what happens if there are more than two ? Clearly, we cannot use $x = 1, 2, \ldots k$ to represent the k categories, since this is treating x as a continuous variable: the expected value of Y would be $\beta_0 + \beta_1$ in group 1, $\beta_0 + 2\beta_1$ in group 2 etc.

Instead, we use a series of "dummy" or "indicator" variables. Indicator variables take the values 0 or 1, and we need to have enough variables that each group has a different combination, which requires $k - 1$ variables if we have k groups. We have already seen that we require a single variable if we have two groups, and table 1.1 below shows how we can use two variables to determine which one of three groups an observation belongs to.

Group	x_1	x_2
A	0	0
B	1	0
C	0	1

Table 1.1.: Use of indicator variables to identify several groups

Here we have 3 groups, so we need 2 indicator variables. The linear model is therefore

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

In this model, β_0 is the expected value of Y when $x_1 = 0$ and $x_2 = 0$, i.e. the expected value of Y in group A. β_1 represents the change in Y as x_1 increases by 1, i.e. the difference between group A and group B. β_2 represents the change in Y as x_2 increases by 1, i.e. the difference between group A and group C.

Here is an example to illustrate the use of indicator variables, again using simulated data. There are three groups, with indicator variables `x1` and `x2` defined as in table 1.1. `Y` is normally distributed with variance 1 and mean 3 in group A, 5 in group B and 4 in group C. The results of analysing this data in stata are given below

```
. regress Y x1 x2

      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  2,    57) =   16.82
       Model |  37.1174969     2  18.5587485           Prob > F      =  0.0000
    Residual |  62.8970695    57  1.10345736           R-squared     =  0.3711
-------------+------------------------------           Adj R-squared =  0.3491
       Total |  100.014566    59  1.69516214           Root MSE      =  1.0505

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.924713   .3321833     5.79   0.000     1.259528    2.589899
          x2 |   1.035985   .3321833     3.12   0.003     .3707994    1.701171
       _cons |   3.075665   .2348891    13.09   0.000     2.605308    3.546022
------------------------------------------------------------------------------
```

From this output we can conclude that there are highly significant differences in Y between the three groups (from the F-statistic, 16.82, which gives a p-value of 0.0000). The mean value in group A (i.e. when `x1` and `x2` are both 0) is 3.08, with a 95% confidence interval of (2.61, 3.55), close to the true value of 3. The difference between group A and group B is estimated by the coefficient of `x1`, which is 1.92 (95% CI; 1.26, 2.59), compared to the true value of 2, and the difference between group A and group C is estimated by the coefficient of `x2`, which is 1.04 (95% CI; 0.37, 1.70), compared to the true value of 1.

In the above analysis, groups B and C are both compared to group A. This is how such analyses are generally performed: one group is chosen as a baseline or reference group (all of the indicator variables are set to 0 in this group), and the other groups are compared to it. However, it may be that we are also interested in the difference between group B and group C, which is not given directly in the above output. We know that the expected value in group B is $\beta_0 + \beta_1$, whilst the expected value in group C is $\beta_0 + \beta_2$. Hence the expected difference between the groups is $(\beta_0+\beta_1)-(\beta_0+\beta_2) = \beta_1-\beta_2$. This difference can be calculated from the regression output as 1.92 - 1.04 = 0.88, but testing whether this is significantly different from 0, or putting a confidence interval around it, is less straightforward. Fortunately, stata can do it for us using the command `lincom`, short for "Linear Combination", since $\beta_1 - \beta_2$ is a linear combination of the parameters β_1 and β_2. How this is done is illustrated below

```
. lincom x1 - x2

 ( 1)  x1 - x2 = 0

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .8887284   .3321833     2.68   0.010     .2235428    1.553914
------------------------------------------------------------------------------
```

Thus we can conclude that the difference between groups B and C is significant ($p = 0.01$), with a 95% confidence interval of (0.22, 1.55), compared to a true value of 1.

Other linear combinations may be of interest. For example, the mean value of Y in group B is given by $\beta_0 + \beta_1$. Calculating this using `lincom` gives the following output

```
. lincom _cons + x1

 ( 1)  x1 + _cons = 0

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.000378   .2348891    21.29   0.000     4.530021    5.470736
------------------------------------------------------------------------------
```

In other words, the mean value of Y in group B is 5.00, with a 95% confidence interval (4.53, 5.47). This is close to the true value of 5.

Generating indicator variables yourself can be tedious and error-prone, particularly if there are a large number of categories. However, Stata can do it for you. Suppose that in the above example, we did not have x_1 and x_2, but only a variable `group`, which took the values "1", "2" and "3". We can tell stata to treat `group` as a categorical variable by adding `i.` as a prefix to its name.

The regression model can be fitted with the command

```
regress Y i.group
```

which produces the following output:

```
. regress Y i.group

      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  2,    57) =   16.82
       Model |  37.1174969     2  18.5587485           Prob > F      =  0.0000
    Residual |  62.8970695    57  1.10345736           R-squared     =  0.3711
-------------+------------------------------           Adj R-squared =  0.3491
       Total |  100.014566    59  1.69516214           Root MSE      =  1.0505

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
          2  |   1.924713   .3321833     5.79   0.000     1.259528    2.589899
          3  |   1.035985   .3321833     3.12   0.003     .3707994    1.701171
             |
       _cons |   3.075665   .2348891    13.09   0.000     2.605308    3.546022
------------------------------------------------------------------------------
```

This can be seen to be identical to the previous analysis, except for the labelling of the groups.

1.2.1. ANOVA

This section can safely be skipped over if you are unfamiliar with the technique of Analysis of Variance, or ANOVA

Just as linear regression with a single indicator variable is equivalent to a t-test, linear regression with several indicator variables is equivalent to an ANOVA test. This is illustrated by the stata output below, analysing the same data as in section 1.2.

```
. oneway Y group

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      37.1174969      2   18.5587485     16.82     0.0000
 Within groups      62.8970695     57   1.10345736
------------------------------------------------------------------------
    Total           100.014566     59   1.69516214

Bartlett's test for equal variances:  chi2(2) =   0.3023  Prob>chi2 = 0.860
```

The anova table above is identical to that produced by the linear regression analysis. Again, the assumptions required for performing ANOVA are the same as those required for performing linear regression with several indicator variables: the `regress` command is simply a convenient way of producing exactly the same analysis. Bartlett's test, given at the bottom of the above printout, is a test that the variance of Y is the same in all three groups. Since x only takes three values, this is testing that the variance is the same for all values of x, which is one of the assumptions of the linear model.

1.3. Mixing Categorical & Continuous Variables

In the preceding sections we have only dealt with situations in which there was only a single categorical variable in the linear model. However, it is also possible to mix categorical and continuous predictors in the same model. For example, suppose that in the clinical trial discussed in section 1.1, we expect the outcome variable Y to vary with age, as well as with treatment. In such a case, we would want to fit both age and treatment in our linear model, for two reasons

1. It may be that the two treatment groups differ in age. If this is the case, a difference in Y between the two groups may be because of the age difference, rather than a treatment effect. This is known as confounding, and will be discussed in detail in chapter 2. Here, I will just say that if we fit both age and treatment, the treatment coefficient measures what the difference between the two groups would have been if the groups had not differed in age.

2. If age is an important predictor of Y, then by including age in our model, we are reducing the amount of unexplained variation, and with therefore achieve narrower confidence intervals around our estimated effect of treatment.

Again, it is easiest to illustrate this using simulated data. Suppose that we have recruited subjects aged 20 – 40, and that in the placebo group, Y is normally distributed with mean 20 - age/2 and variance 1, and that in the active treatment group, the mean of Y is 21 - age/2. A scatter-plot of this simulated data is given in Figure 1.1.

In this case, if we simply fit treatment as our predictor variable, we find the difference between the two groups is not significant.

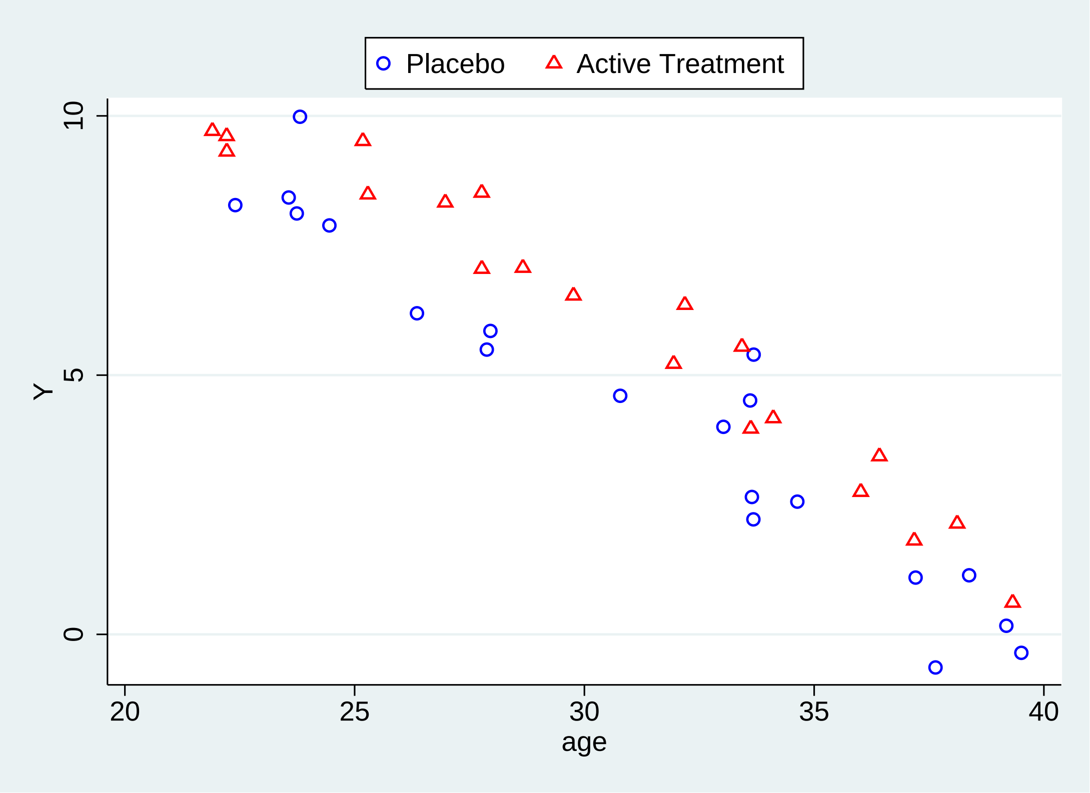

Figure 1.1.: Variation in outcome with age in simulated trial data

```
. regress Y treat

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =    2.86
       Model |  26.5431819     1  26.5431819           Prob > F      =  0.0989
    Residual |  352.500943    38  9.27634061           R-squared     =  0.0700
-------------+------------------------------           Adj R-squared =  0.0456
       Total |  379.044125    39  9.71908013           Root MSE      =  3.0457

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1.629208   .9631376     1.69   0.099    -.3205623    3.578978
       _cons |   4.379165   .6810411     6.43   0.000      3.00047    5.757861
------------------------------------------------------------------------------
```

Here, although the observed effect of treatment (1.63) is close to its hypothetical value (1.00), the confidence interval is very wide, because of the variation with age which is being treated as random variation in this model. If we fit age as well as treatment, the observed effect changes very little (to 1.24), but it is now very highly significant, because its standard error is reduced greatly. We have seen that the standard error of a coefficient is equal to $\frac{\sigma}{\sqrt{n}s_x}$: $\sqrt{n}$ and s_x are unchanged, but by including age in the model, we have decreased σ greatly, and consequently decreased the standard error.

```
. regress Y treat age

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  2,    37) =  262.58
       Model |  354.096059     2   177.04803           Prob > F      =  0.0000
    Residual |  24.9480658    37  .674272049           R-squared     =  0.9342
-------------+------------------------------           Adj R-squared =  0.9306
       Total |  379.044125    39  9.71908013           Root MSE      =  .82114

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1.238752   .2602711     4.76   0.000     .7113924    1.766111
         age |  -.5186644   .0235322   -22.04   0.000    -.5663453   -.4709836
       _cons |   20.59089   .7581107    27.16   0.000     19.05481    22.12696
------------------------------------------------------------------------------
```

One way to think of this is that in the first analysis, we are selecting a subject at random from each of the two treatment groups, and measuring the difference in Y between them. This varies enormously, depending on the ages of the two subjects. However, if we select two subjects *of the same age* and measure the difference in Y between them, there will be much less variation: this is what we do when we include age in the regression equation.

From the second analysis, the expected value of Y in the placebo group is given by $20.59 - 0.52 \times age$, whilst the expected value in the active treatment group is given by $20.59 - (0.52 \times age) + 1.24$. These predictions can be added to the plot of the data given in Figure 1.1 above.

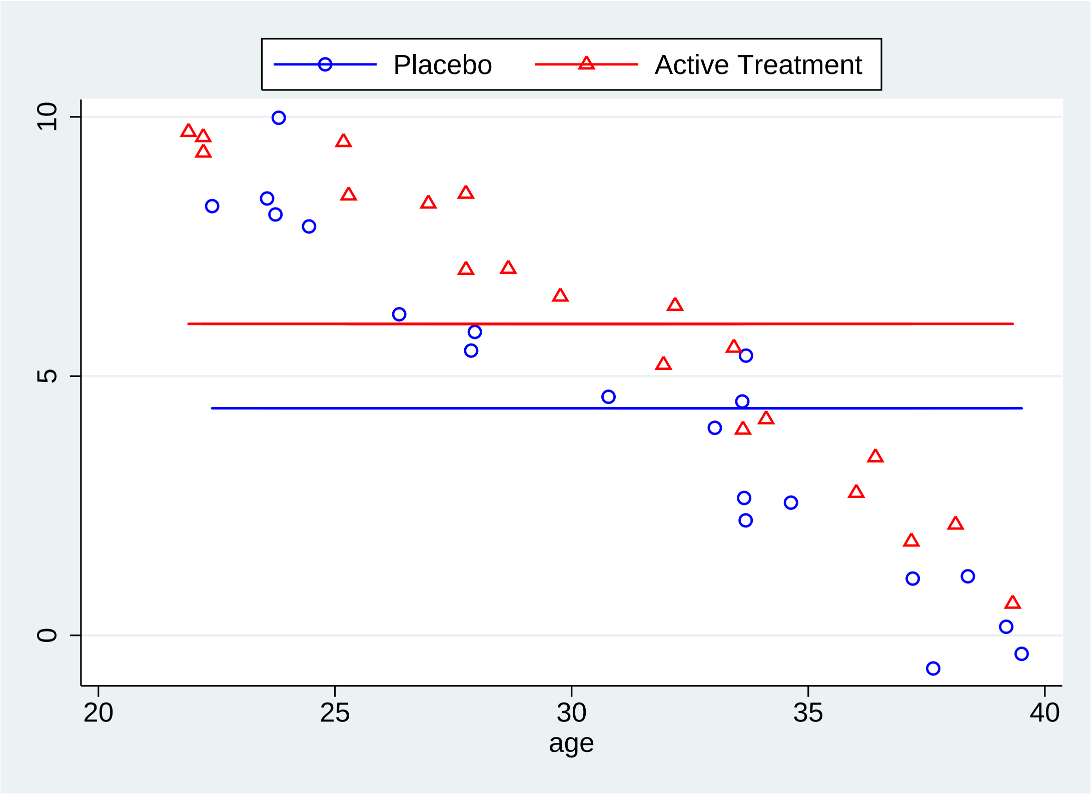

Figure 1.2.: Variation in outcome with age in simulated trial data, and predicted values

1.4. Interactions

In the above analysis, we have assumed that the effect of treatment is the same at all ages. Geometrically, this means that we have fitted parallel lines to the plot of Y against age, as seen in Figure 1.2. However, this assumption may not be true: it may be that the treatment is more effective in the older patients than in the younger ones, so the fitted lines should be further apart in the older subjects. A situation like this, in which the effect of one variable on the outcome depends on the value of another variable, is called an "interaction": we say that there is an interaction between age and treatment.

For example, consider the data illustrated in Figure 1.3. In this example, the effect of treatment is to reverse the effect of age, so that the outcome in the treated group is normally distributed with mean 10 and variance 1. Thus the predicted values in the treated and untreated groups get further apart with increasing age. Fitting parallel lines to these two groups does not provide a good fit to the data, as Figure 1.3 shows.

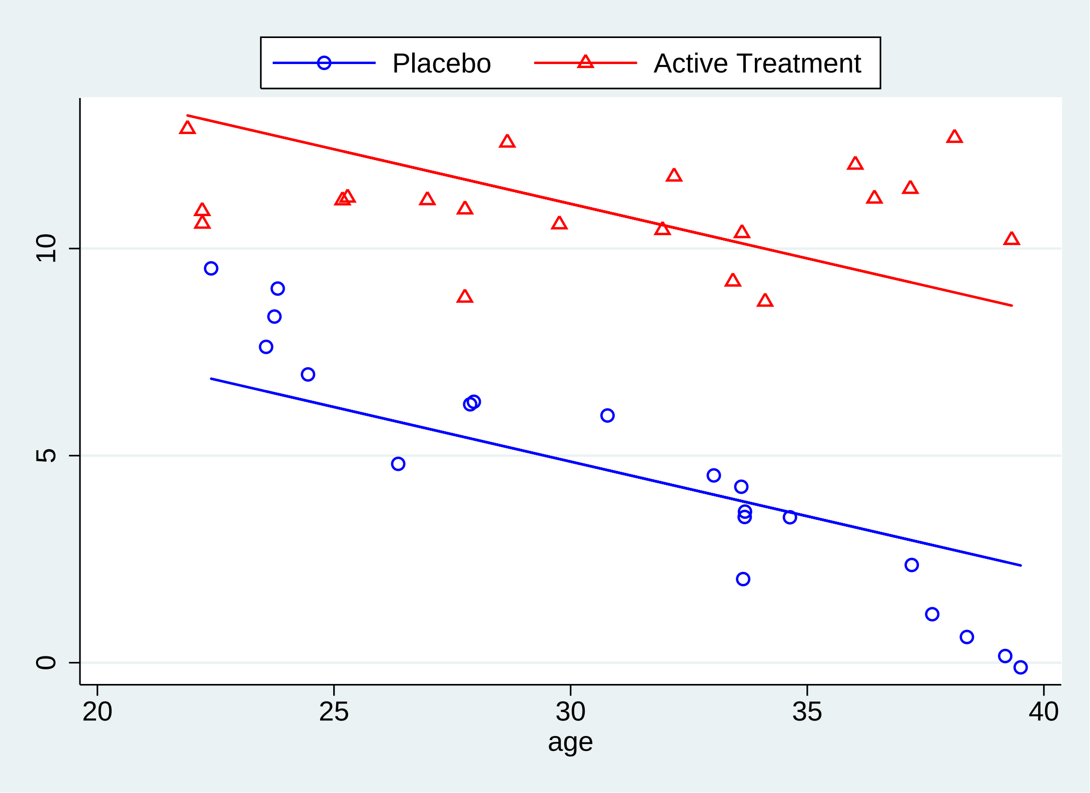

Figure 1.3.: Variation in outcome with age in simulated trial data, in which age interacts with treatment, with predicted values from regression model excluding interaction.

Fitting an interaction term can be thought of as fitting the two separate equations

$$Y = \begin{cases} \beta_{00} + \beta_{10} \times \texttt{age} + \varepsilon & \text{if } \texttt{treat} = 0 \\ \beta_{01} + \beta_{11} \times \texttt{age} + \varepsilon & \text{if } \texttt{treat} = 1 \end{cases} \tag{1.1}$$

The two groups each have different intercepts (β_{00} and β_{01}) and slopes with age (β_{10} and β_{11}). However, it is possible to combine the two above equations into the single equation

$$Y = \beta_{00} + \beta_{10} \times \text{age} + (\beta_{01} - \beta_{00}) \times \text{treat} + (\beta_{11} - \beta_{10}) \times \text{age} \times \text{treat} + \varepsilon. \tag{1.2}$$

Thus fitting the linear model with an interaction term amounts to including as predictors the variables age, treat, and a new variable formed by multiplying age and treat together. The intercept in this model is the intercept in placebo group (β_{00}), the coefficient of age is the slope with age in the placebo group (β_{10}), the coefficient of treat is the difference between the intercept in the active group and the intercept in the placebo group ($\beta_{01} - \beta_{00}$), and the coefficient of the interaction term is the difference in the slopes between the active and treatment groups ($\beta_{11} - \beta_{10}$).

Just as Stata can automatically generate indicator variables, it can also automatically generate interaction terms. To request an interaction between to variables, you include the symbol # between them. So if we had two categorical variables and wanted to consider the interaction between them, we would add i.var1#i.var2 to the regression model. There is also a shorthand we can use: i.var1##i.var2 is equivalent to i.var1 i.var2 i.var1#i.var2. If we wish to include a continuous variable in an interaction, we precede the variable name with c. rather than i.

So, to fit a linear model containing age, treatment and their interaction to the data in Figure 1.3, we would enter

```
regress Y i.treat##c.age
```

and obtain the following output.

```
. regress Y i.treat##c.age

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  3,    36) =  173.38
       Model |  563.762012     3  187.920671           Prob > F      =  0.0000
    Residual |  39.0189256    36  1.08385904           R-squared     =  0.9353
-------------+------------------------------           Adj R-squared =  0.9299
       Total |  602.780938    39  15.4559215           Root MSE      =  1.0411

------------------------------------------------------------------------------------
                 Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
             treat |
Active Treatment   |  -8.226356   1.872952    -4.39   0.000    -12.02488   -4.427833
               age |  -.4866572   .0412295   -11.80   0.000    -.5702744     -.40304
                   |
       treat#c.age |
Active Treatment   |   .4682374   .0597378     7.84   0.000     .3470836    .5893912
                   |
             _cons |   19.73531   1.309553    15.07   0.000     17.07942    22.39121
------------------------------------------------------------------------------------
```

The interaction term (written treat#c.age) is highly significant ($p = 0.000$), so we have good evidence that the rate of change of Y with age is different between the two groups. The coefficient for age is still the slope of the graph of Y against age in the placebo group (i.e. when treat $= 0$). However, the slope of Y against age in the treated group is now given by the sum of the coefficients age + treat#c.age This slope can be calculated using lincom as before, to give

```
. lincom age + 1.treat#c.age

 ( 1)  age + 1.treat#c.age = 0

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0184198   .0432288    -0.43   0.673    -.1060919    .0692523
------------------------------------------------------------------------------
```

This is very nearly 0, suggesting that Y does not change with age in the treated group, although it does reduce with age in the placebo group. Figure 1.4 shows the same data as Figure 1.3, but the predicted values in this case include the interaction term. This shows clearly that the slopes in the treatment and placebo groups are quite different: no change with age in the treated group, reduction with age in the placebo group.

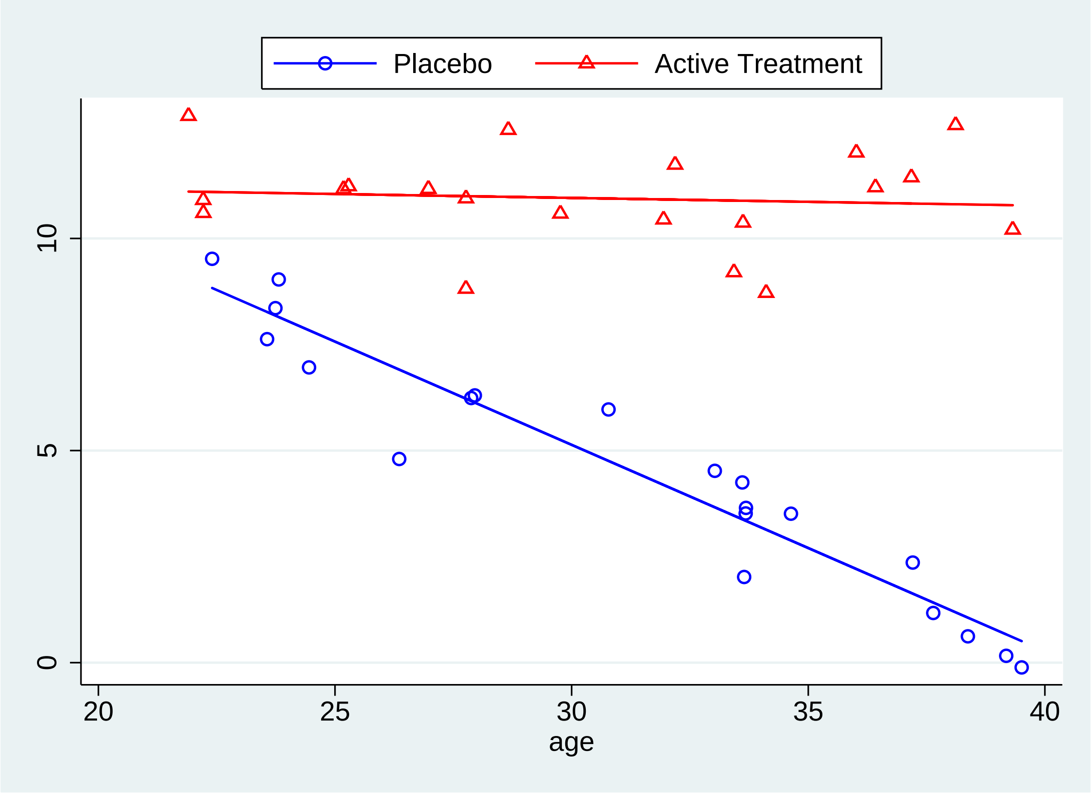

Figure 1.4.: Variation in outcome with age in simulated trial data, in which age interacts with treatment, with predicted values

As in the previous examples, the intercept (`_cons`) is the expected value of Y when all the other variables are equal to 0: i.e. the expected value at age 0 in the placebo group. The coefficient for `treat` measures the difference between the treated and untreated groups *at age 0*. It cannot be interpreted as the difference between the placebo and treated groups in this case, since the treatment effect varies with age, as seen in Figure 1.4. In fact, it is of little intrinsic meaning in this case, since

1. It only applies to subjects of age 0, and we are unlikely to be interested in them

2. The youngest subjects in our sample were of age 20, so we are extrapolating a long way beyond our data.

The effect of treatment at age a can be calculated as `treat` $+ a \times$ `treat#c.age`. Again, `lincom` can be used to perform the calculations, so the treatment effect at age 20 would be given by:

```
. lincom 1.treat + 20*1.treat#c.age

 ( 1)  1.treat + 20*1.treat#c.age = 0

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.138392   .7279832     1.56   0.127    -.3380261     2.61481
------------------------------------------------------------------------------
```

The treatment effect at age 40 would be given by:

```
 ( 1)  1.treat + 40*1.treat#c.age = 0

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   10.50314   .6378479    16.47   0.000     9.209524    11.79676
------------------------------------------------------------------------------
```

In other words, as shown in Figure 1.4, there is little difference between the treated & placebo groups at age 20, but considerable difference at age 40.

1.5. Testing Several Parameters With `testparm`

Often when dealing with categorical variables, we wish to test whether several parameters are signficantly different from 0 at the same time. The command that enables us to do so is `testparm`, which is a contraction of "test parameters". The syntax for testparm is simply `testparm` *`varlist`*, where *varlist* is a list of variables which all appeared in the last regression model. `testparm` tests the hypothesis that $\beta = 0$ for all variables in *varlist*.

2. Confounding

2.1. What is Confounding ?

A linear model can be used to show that two variables are associated, i.e. if one increases, the other also tends to increase (or tends to decrease, if the association is negative). It may be that changes in one cause the changes in the other, but it may be that there is a third factor which is associated with both.

As an example of confounding, consider the data in `auto.dta` which we looked at in a previous practical. Imagine that we are interested in whether US-built cars differ from non-US-built cars in their fuel consumption. We have already seen that the fuel consumption, in miles per gallon, is stored in the variable `mpg`. There is also a variable `foreign`, which is 0 for US vehicles and 1 for non-US vehicles. So we can test for a difference using a linear model to predict `mpg` from `foreign`. The output is shown below.

```
. regress mpg foreign

  Source |       SS       df       MS                  Number of obs =      74
---------+------------------------------               F(  1,    72) =   13.18
   Model |  378.153515     1  378.153515               Prob > F      =  0.0005
Residual |  2065.30594    72  28.6848048               R-squared     =  0.1548
---------+------------------------------               Adj R-squared =  0.1430
   Total |  2443.45946    73  33.4720474               Root MSE      =  5.3558

------------------------------------------------------------------------------
     mpg |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 foreign |   4.945804   1.362162      3.631   0.001       2.230384    7.661225
   _cons |   19.82692   .7427186     26.695   0.000       18.34634    21.30751
------------------------------------------------------------------------------
```

This shows a significant difference in fuel consumption between US and non-US vehicles, with foreign vehicles averaging nearly 5 miles per gallon more than US vehicles. However, remember that weight is a very important determinant of fuel consumption, and consider Figure 2.1. This shows a plot of fuel consumption against weight for both US and non-US vehicles. Clearly, the number of miles travelled per gallon decreases as the weight of the vehicle increases. However, notice that most of the lightest vehicles are non-US, whilst all of the heaviest vehicles are US-built. Could the advantage in fuel consumption of non-US vehicles be due to their lighter weight ?

We can test this by fitting both `weight` and `foreign` as predictors in our linear model. Then, the coefficient of `foreign` measures the difference in fuel consumption between a US and a non-US vehicle *of the same weight*, rather than just a randomly selected US and non-US vehicle as we did before. The results of fitting this regression are given below.

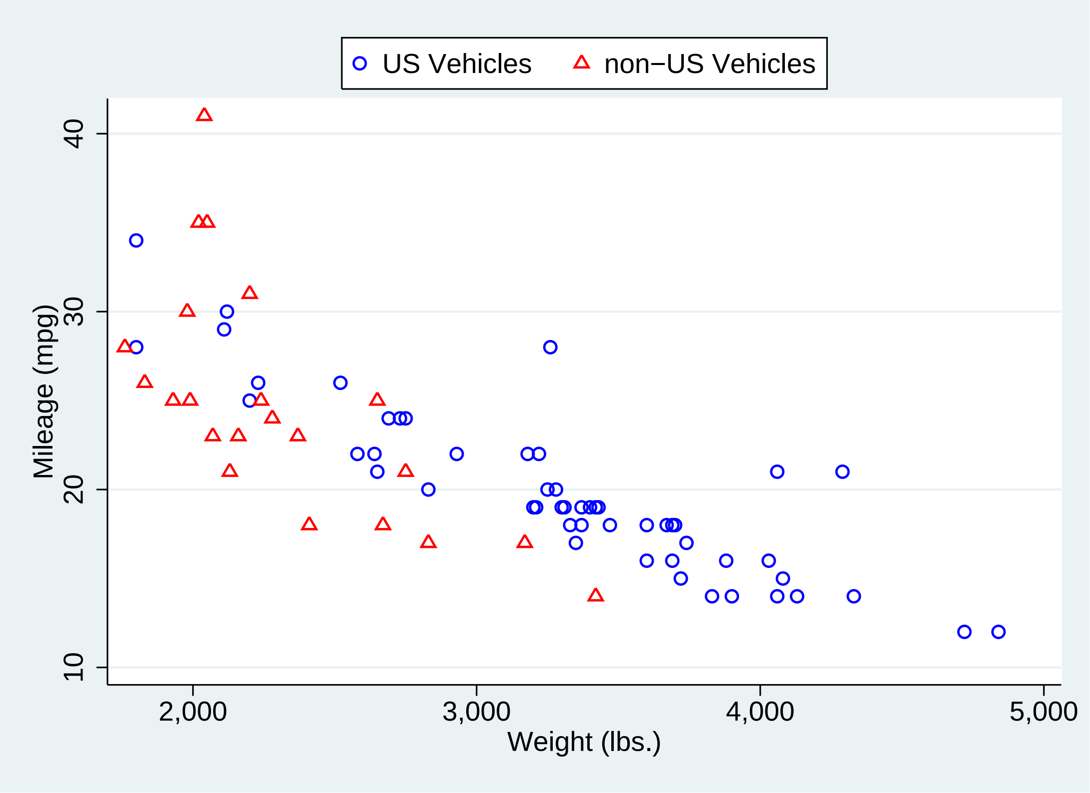

Figure 2.1.: Plot of m.p.g. against weight for US and non-US vehicles

```
. regress mpg foreign weight

  Source |       SS       df       MS              Number of obs =      74
---------+------------------------------           F(  2,    71) =   69.75
   Model |  1619.2877     2  809.643849            Prob > F      =  0.0000
Residual |  824.171761    71   11.608053           R-squared     =  0.6627
---------+------------------------------           Adj R-squared =  0.6532
   Total |  2443.45946    73  33.4720474           Root MSE      =  3.4071

------------------------------------------------------------------------------
     mpg |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 foreign |  -1.650029   1.075994     -1.533   0.130         -3.7955    .4954421
  weight |  -.0065879   .0006371    -10.340   0.000       -.0078583   -.0053175
   _cons |    41.6797   2.165547     19.247   0.000        37.36172    45.99768
------------------------------------------------------------------------------
```

This model fits much better than the previous model ($R^2 = 66\%$, as opposed to 15% previously). However, the effect of `foreign` is no longer statistically significant ($p = 0.13$), suggesting that the difference between US and non-US vehicles we saw before could be explained by differences in weight. We can conclude that the apparent difference in fuel consumption between US and non-US cars is due to differences in weight between US and non-US cars.

2.2. How do we recognise a confounder ?

In order to confound the association between a predictor variable and an outcome, a variable has to satisfy the following conditions:

1. The confounder must be associated with the outcome.

2. The association between outcome and confounder must be independent of the predictor variables.

3. The confounder cannot be on the causal path linking the predictor to the outcome.

Determining whether a variable satisfies the above conditions, particularly the last, can be tricky. It is not possible without some understanding of the mechanism by which the predictor is associated with the outcome.

For example, consider the difference in fuel consumption between US-made and non-US-made cars considered earlier. We have seen that the non-US made cars have better fuel consumption because they are lighter, but why are they lighter ? If they are lighter because non-US car designers have clever tricks to reduce the weight of a car without affecting its capacity, comfort etc, then weight is a path variable, not a confounder, and non-US car designers produce cars with better fuel consumption. However, if the US and non-US designers were given different briefs: the US designers were asked to produce larger cars since that is what sells in the US, then weight is not a path variable but a confounder. In this case, the second analysis above is appropriate and suggests that the US designers are every bit as good at producing cars with low fuel consumption as non-US designers, given their different briefs.

2.3. How do we allow for confounding ?

Dealing with a confounder once it has been identified is very straightforward. Simply adding a confounding variable as a predictor in the linear model will allow for confounding. Remember that a coefficient in the linear model measures the effect of a variable *when all other variables do not change*, so the coefficient for the predictor we are interested in will be a measure of its effect on the outcome when the confounding variable does not change.

It should be noted that adjusting for confounding in this way assumes that

1. The confounder has been measured perfectly

2. The association between confounder and outcome is perfectly linear.

If either of these assumptions are not true, the linear model will remove *some*, but not *all*, of the effect of confounding. The *residual* confounding may still be enough to affect the parameter estimates.

Ultimately, confounding is a property of the real-world system we are attempting to model, not of the model itself. There is no way of determining from the data whether a variable is a confounder and should be adjusted for, or whether it is on the causal pathway and should not be adjusted for. The correct course of action depends on the mechanism by which the predictors and putative confounders affect the outcome, which requires a thorough knowledge of the subject area. Statistical methods for allowing for confounding are simple and straightforward to apply, the difficult problem is to know whether they are appropriate.

Confounding is a major problem for observational studies. When a significant association is shown in an observational study, it could be due to a causal effect (i.e. one variable is directly related to the other) or to confounding (i.e. one variable is correlated with another variable which directly affects the outcome).

3. Other Considerations

3.1. Variable Selection

Suppose that we have a number of predictor variables, from which to predict the outcome. Should we include all of the variables in our linear model, or only a subset of them ? And if we are to use a subset, how do we select that subset ?

If we are only interested in the predictions from the linear model, we can use all of the variables. However, if we are interested in the mechanisms by which the predictors affect the outcome, it is of use to remove those "predictors" which are only predictive because they are correlated with other variables which genuinely affect Y. In addition, if we wish to use the model on further samples, it may be that there is a financial or logistical advantage to having a smaller number of variables to measure.

There are a number of ways which have been suggested for selecting variables in this context.

3.1.1. Forward Selection

Suppose we are selecting from k predictors. First, we chose a significance level to use to determine whether a variable is a significant predictor or not. We will call this p_e, the significance level at which a predictor enters the regression model. Then we use each of the k predictors in turn to predict Y, and the one which explains the greatest proportion of the variance is our candidate for inclusion at this stage. We then look at the significance level of the t-test for $\beta = 0$ for this variable: if it is less than p_e, the variable is selected.

Once a variable has been selected in this way, each of the remaining variables are added to the model in turn. If any of these variables are significant at the p_e level after adjusting for all the variables selected so far, the most significant variable is added to the selection. This process repeats until either all of the variables have been selected, or none of them are significant at the p_e level after adjusting for the selected variables.

3.1.2. Backward Elimination

An alternative to forward selection is backward elimination. In this procedure, we chose a significance level at which a variable will be *removed* from the model, p_r. Then we produce a single linear model containing all k predictor variables, and compare the significance level of the least significant model to p_r. If it is greater than p_r, that variable is removed from the model and a linear model formed with the remaining $k - 1$ variables. The process is repeated until all remaining variables have significance levels less than p_r.

Backward selection can be better than forward selection if there are variables that are jointly good predictors, but not individually (as seen in the dataset `wood73`). However, if there are a number of strongly correlated variables, the results can be unreliable.

3.1.3. Stepwise Selection

Stepwise selection is a combination of the two preceding methods. The procedure starts as forward selection, but each time a variable is added to the model, all of the variables are tested to see if any can now be removed. Obviously, when using stepwise selection, p_r must be greater than or equal to p_e, otherwise a variable added in the "forward" part of the procedure could be removed in the "backward" part.

3.1.4. All Subsets

All three of the above methods have the drawback that not every possible combination of predictors is considered. In some cases it may be possible to consider every possible combination, and choose the one with the highest adjusted R^2. However, this can be very time consuming: if we have 10 predictors, there are 1023 possible subsets to be fitted, compared with at most 10 using forwards or backwards selection. Since all subsets regression is not implemented in stata, it will not be discussed further.

3.1.5. Some Caveats Concerning Variable Selection

Significance Levels in Variable Selection

The main problem with variable selection is that the significance levels of the various parameters are no longer correct. The hypotheses we are testing to include or exclude variables are not independent of each other, and they are not randomly selected (we always test either the most significant variable or the least significant). How greatly the p-values differ from their nominal levels depends on the ratio of the sample size to the number of predictor variables: if there are many times as many observation as predictor variables, this problem will not be of great importance.

Differences in Models Selected

It is not uncommon for the model chosen using forward selection to differ from that chosen using backward elimination. A measure of common sense is needed to determine which model to use. For example, if two variables are highly correlated, there will be little difference between a model containing one and a model containing the other.

Inappropriate variable choice

Statistical significance should not be the only criterion for selecting predictor variables in a model. It may be that you know a particular variable is a predictor of your outcome, even though your sample is too small for the effect to be statistically significant. You should therefore have no qualms about forcing certain variables to be in the regression model even if the automatic variable selection methods would exclude them[a].

3.1.6. Variable selection in Stata

In stata, the `sw` command can be used to give either forward, backward or stepwise selection. The syntax for this command is

[a] However, you will have to explain which variables were forced into the model and your reasons for doing so in the methods section of any resulting publication

```
sw regress yvar xvars
```

The significance levels p_e and p_r are set using the options `pe()` and `pr()`: at least one of these must be set. If only `pe()` is set, forward selection is performed, whilst if only `pr()` is set, backwards elimination is performed. If both are set, stepwise selection is used.

For example, consider the `auto` dataset distributed with stata. Imagine we wish to determine which variables contribute to the weight of the vehicle. To use forwards selection, we would type

```
sw regress weight price headroom trunk length turn displ gear_ratio, pe(0.05)
```

to get the output

```
p = 0.0000 <  0.0500  adding   length
p = 0.0000 <  0.0500  adding   displ
p = 0.0015 <  0.0500  adding   price
p = 0.0288 <  0.0500  adding   turn

  Source |       SS       df       MS              Number of obs =      74
---------+------------------------------           F(  4,    69) =  293.75
   Model |  41648450.8     4  10412112.7           Prob > F      =  0.0000
Residual |  2445727.56    69  35445.3269           R-squared     =  0.9445
---------+------------------------------           Adj R-squared =  0.9413
   Total |  44094178.4    73  604029.841           Root MSE      =  188.27

------------------------------------------------------------------------------
  weight |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  length |   19.38601   2.328203      8.327   0.000       14.74137    24.03064
   displ |   2.257083    .467792      4.825   0.000       1.323863    3.190302
   price |   .0332386   .0087921      3.781   0.000       .0156989    .0507783
    turn |   23.17863   10.38128      2.233   0.029       2.468546    43.88872
   _cons |  -2193.042   298.0756     -7.357   0.000      -2787.687   -1598.398
------------------------------------------------------------------------------
```

Using forward selection, first `length`, then `displ`, then `price` and finally `turn` are selected as significant predictors. The other variables, `headroom`, `trunk` and `gear_ratio` are not selected as significant.

To use backward elimination, we would type

```
sw regress weight price headroom trunk length turn displ gear_ratio, pr(0.05)
```

and would get the output

```
p = 0.6348 >= 0.0500  removing headroom
p = 0.5218 >= 0.0500  removing trunk
p = 0.1371 >= 0.0500  removing gear_ratio

  Source |       SS       df       MS                  Number of obs =      74
---------+------------------------------               F(  4,    69) =  293.75
   Model |  41648450.8     4  10412112.7               Prob > F      =  0.0000
Residual |  2445727.56    69  35445.3269               R-squared     =  0.9445
---------+------------------------------               Adj R-squared =  0.9413
   Total |  44094178.4    73  604029.841               Root MSE      =  188.27

------------------------------------------------------------------------------
  weight |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   price |   .0332386   .0087921      3.781   0.000       .0156989    .0507783
    turn |   23.17863   10.38128      2.233   0.029       2.468546    43.88872
   displ |   2.257083    .467792      4.825   0.000       1.323863    3.190302
  length |   19.38601   2.328203      8.327   0.000       14.74137    24.03064
   _cons |  -2193.042   298.0756     -7.357   0.000      -2787.687   -1598.398
------------------------------------------------------------------------------
```

The final model here is exactly the same as we obtained from forward selection. First all variables are fitted, then `headroom`, `trunk` and `gear_ratio` are eliminated (in that order).

One important feature of `sw` is the ability to treat several variables as a single term. For example, if there is a categorical variable with 4 levels, it will be fitted using three indicator variables, as we have seen. Normally, these three variables should either all be included or all excluded. This can be achieved by putting parentheses around the variables to be treated together. For example, suppose we wish to predict the urban population of various American states from the total population and the region. If we are using forward selection, we would type[b]

```
xi:  sw regress popurban pop (i.region), pe(0.05)
```

and obtain the output

[b] Note that the older stata syntax for generating indicator variables needs to be used with `sw`, since this command does not understand factor variables. More information about the old syntax is given in Appendix A

```
i.region          _Iregion_1-4        (naturally coded; _Iregion_1 omitted)
                      begin with empty model
p = 0.0000 <  0.0500  adding   pop
p = 0.0003 <  0.0500  adding   _Iregion_2 _Iregion_3 _Iregion_4

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  4,    45) =  794.57
       Model |  8.0830e+14     4  2.0208e+14           Prob > F      =  0.0000
    Residual |  1.1444e+13    45  2.5432e+11           R-squared     =  0.9860
-------------+------------------------------           Adj R-squared =  0.9848
       Total |  8.1975e+14    49  1.6730e+13           Root MSE      =  5.0e+05

------------------------------------------------------------------------------
    popurban |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pop |   .8655805   .0154854    55.90   0.000     .8343914    .8967696
  _Iregion_2 |    -383325   222541.4    -1.72   0.092    -831546.4    64896.47
  _Iregion_3 |  -529247.1   210480.1    -2.51   0.016    -953175.8   -105318.3
  _Iregion_4 |   313714.2     221173     1.42   0.163      -131751    759179.4
       _cons |  -402777.4   188162.4    -2.14   0.038      -781756   -23798.79
------------------------------------------------------------------------------
```

Here, pop enters the model first, then _Iregion_2, _Iregion_3 and _Iregion_4. This is despite the fact that _Iregion_2 and _Iregion_4 are not individually statistically significant (p = 0.09 & 0.16 respectively). The three variables taken as a whole are highly significant, as can be shown using testparm:

```
. testparm _I*

 ( 1)  _Iregion_2 = 0.0
 ( 2)  _Iregion_3 = 0.0
 ( 3)  _Iregion_4 = 0.0

       F(  3,    45) =    7.67
            Prob > F =    0.0003
```

If the parentheses were not used around i.region, only _Iregion_3 would have been added to the model.

Finally, it may be that there are certain variables that are known to be important within the process, and we may want to ensure they are included in the model irrespective of their statistical significance. This can be achieved using the lockterm1 option, which ensures that the predictor variable is retained in the model irrespective of its significance. Again, parentheses can be put around several variable names to make sw treat them as a single unit.

3.2. Multicollinearity

Strictly, multicollinearity occurs when one predictor variable can be calculated from one or more other predictors variables. For example, if the linear model is $Y = x_1 - 3x_2 + x_3$, but $x_3 = x_1 + 2x_2$, then we can replace the regression equation with $Y = 2x_1 - x_2$ to get exactly the same predicted values. Thus the β parameters are no longer uniquely defined. If multicollinear predictors are used with the regress command, stata will usually drop one or more predictors until the β parameters are uniquely determined. However, it may be unable to determine which variables are involved, in which case it will be unable to calculate standard errors for some of the variables.

In practice, such exact multicollinearity is rare. However, if two or more variables are highly correlated,[c] this can still cause problems with estimating the β parameters. A common consequence is that the standard errors of the β parameters becomes very large. It can also happen that the β parameters themselves take very unusual values, and may even have the opposite sign to that expected.

The only easy way around this problem is to use less correlated variables as predictors. For example, rather than using both diastolic blood pressure and systolic blood pressure as predictors (which are highly correlated), use systolic and (diastolic - systolic), which will be much less strongly correlated.

If it is not possible to remove the multicollinearity in this way, more complex solutions are available, such as ridge regression and principal components regression. However, these techniques are beyond the scope of this course.

3.3. Polynomial Regression

If the relationship between Y and x is non-linear, polynomial terms (i.e. x^2, x^3 etc.) can be added to the model. This enables curved lines defined by quadratic or cubic equations to be fitted to the data, which may improve the fit. Polynomial terms should be fitted until one of them is not significant (rather like a forward selection procedure) at a chosen significance level. The power of the highest polynomial is referred to as the *degree* of the polynomial model.

It should be remembered that the individual coefficients are meaningless in polynomial regression. In other words, if $Y = \beta_0 + \beta_1 x + \beta_2 x^2$, β_1 and β_2 do not have a simple interpretation. The change in Y for a change of 1 in x is no longer constant, but depends on the initial value of x. The only way to understand the association between Y and x is to draw the function $Y = \beta_0 + \beta_1 x + \beta_2 x^2$.

It should be noted that, depending on the range of values of x, there may be strong correlations between the various powers of x. This can lead to the problem of multicollinearity outlined in section 3.2. This can be avoided by using orthogonal polynomials, i.e. rather than fit x^2, fit $(x-c)^2$, where c is chosen so that the correlation between x and $(x-c)^2$ is 0. Orthogonal polynomials can be calculated by stata: see the command `orthpoly`.

An alternative, suggested by Altman, is to use fractional polynomials. For details of how to use fractional polynomials in stata, see the help for the command `fracpoly`. Another alternative approach is to use splines: details can be found in the help for the command `mkspline`.

3.4. Transformations

If the variance of Y is not constant, the only solution is to transform Y. A transformation may also be called for if the distribution of Y is not normal. Transformations may also be used on the x-variables, but this is less common: usually it is simpler to use polynomials if the association between Y and x is non-linear but there is no other reason to transform Y.

If Y has a positively skewed distribution (i.e. there are some unusually large values), it may be worth transforming Y by taking logs and fitting the model

$$\log(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

This transformation is commonly used with measurements of biochemical parameters, which tend to be skewed. This transformation is also appropriate if the variance of Y increases as Y increases.

[c] strictly speaking, if there are two linear combinations of variables that are highly correlated, for example if $x_1 + 3x_2$ is highly correlated with $x_3 + 4x_4$

The major drawback with transforming the data is that the parameters are harder to interpret. If we fit the linear model $f(Y) = \beta_0 + \beta_1 x_1$, where $f(Y)$ is some transformation of Y, then if x increases by 1, it is $f(Y)$, not Y which will increase by β_1. The amount by which Y increases will depend on its initial value, which makes it difficult to simply summarise the effect of x on Y.

For this reason, transformations other than the log transformation are largely avoided. The reason that the log transformation is used is that if $\log(Y)$ increases by β, that is equivalent to multiplying Y by e^{β}. Hence the parameters from a linear model for $\log(Y)$ are easy to interpret on the original log scale. For example, if the regression equation is $\log(Y) = 0.5x$, then for each unit increase in x, Y is multiplied by $e^{0.5} = 1.65$. In other words, when x increased by 1, Y increases by 65%.

3.5. Regression Through the Origin

It may be that there is an *a priori* reason to assume that $Y = 0$ when all of the predictor variables are 0. In this case, it is possible to fit a linear model in which β_0 is forced to be 0, i.e. the regression line is forced through the origin. However, this should only be used if data is available for x-values close to 0. If we have no data from near the origin, we should not be concerned with trying to predict Y near the origin: we can only be sure our linear model holds in the region in which we have collected data.

If we are certain regression through the origin is justified, we can force stata to force β_0 to be equal to zero with the option `nocons` to the `regress` command. However, it should be noted that R^2 is calculated differently when performing regression through the origin, and cannot be compared to the R^2 value obtained normally.

4. Practical

4.1. Datasets

The datasets that you will use in this practical can be accessed via http from within stata. However, the directory in which they are residing has a very long name, so you can save yourself some typing if you create a global macro for this directory. You can do this by entering

```
global basedir http://personalpages.manchester.ac.uk/staff/mark.lunt
global datadir $basedir/stats/6_LinearModels2/data
```

(In theory, the global variable `datadir` could have been set with a single command, but fitting the necessary command on the page would have been tricky. Far easier to use two separate commands as shown above). If you wish to run the practical on a computer without internet access, you would need to:

1. Obtain copies of the necessary datasets

2. Place them in a directory on your computer

3. Define the global macro `$datadir` to point to this directory.

4.2. Categorical Variables

4.2.1. Dichotomous Variables and T-Tests

Load the `auto` dataset distributed with stata using the command `sysuse auto, clear`. We are going to confirm that US-made vehicles tend to be heavier than non-US made vehicles.

2.1 Fit a linear model with the command `regress weight foreign`. Are foreign vehicles heavier or lighter than US vehicles ? Is the difference signficant ?

2.2 Fit the above model again, but this time use `regress weight i.foreign`
Does this make any difference ?

2.3 Check that `regress` gives the same results as `ttest` by entering the command
`ttest weight, by(foreign)` Look at the difference between the means and its standard error: are they the same as those obtained by `regress`

2.4 Make box-and-whisker plots for the weight of both US and non-US vehicles with the commands

```
sort foreign
graph box weight, by(foreign)
```

Does the variance of `weight` look the same for US and non-US vehicles ?

2.5 You may wish to confirm your ideas in the last question by typing `by foreign: summ weight`. Are the standard deviations similar between US and non-US vehicles ?

2.6 Finally, use `hettest` to confirm whether or not the variances are the same. Was a linear model an appropriate way of analysing this data ?

4.2.2. Multiple Categories and ANOVA

The dataset `$datadir/soap` gives information on the scores given to 90 bars of soap for their appearance. The scores are on a scale of 1-10, the higher the score the better. Three operators each produced 30 bars of soap. We wish to determine if the operators are all equally proficient.

2.7 Sort the data by operator (type `sort operator`), and produce box-and-whisker plots of appearance for each operator by typing `graph box appearance, by(operator)`. Which operator appears to have the highest scores ?

2.8 What are the mean scores for each of the 3 operators ? The command to use is `by operator: summ appearance`

2.9 Fit a linear model to the data with `regress appearance i.operator`. Are there significant differences between the three operators ?

2.10 What is the p-value for the differences between the operators ?

2.11 Which operator was used as the baseline for the linear model ? (Check that the constant term in the model is the same as this operator's mean score you calculated earlier)

2.12 Use `lincom` to calculate the mean score for operator 2. Is this the same as the score you calculated earlier ?

2.13 Use `lincom` to calculate the difference in mean score between operators 2 and 3. Is this difference statistically significant ?

4.3. Interactions and Confounding

The dataset `$datadir/cadmium` give data on the ages, vital capacities and durations of exposure to cadmium (in three categories) in 88 workers. We wish to see if exposure to cadmium has a detrimental effect on vital capacity. However, we know that vital capacity decreases with age, and that the subjects with the longest exposures will tend to be older than those with shorter exposures. Thus, age could confound the relationship between exposure to cadmium and vital capacity.

3.1 Plot a graph of vital capacity against age, to satisfy yourself that vital capacity decreases with increasing age ? (use `scatter capacity age`)

3.2 In case you are not satisfied, fit a linear model to predict vital capacity from age, with `regress capacity age`

It would be nice to be able to tell to which exposure group each point on the plot of vital capacity against age belonged. This can be done by using the commands

```
gen cap1 = capacity if exposure == 1
gen cap2 = capacity if exposure == 2
gen cap3 = capacity if exposure == 3
scatter cap1 cap2 cap3 age
```

This graph clearly shows that the group with the highest exposure tend to be older (they are towards the right of the graph).

3.3 Is there a difference between the three exposure groups in vital capacity ? (use `regress capacity i.exposure`)

3.4 We have seen that a lower vital capacity in the most exposed may be due to their age, rather than their exposure. Adjust the previous example for age, using `regress capacity age i.exposure` Now use `testparm i.exposure` to test whether there are significant differences between groups.

To get a visual idea of the meaning of the previous regression model, type

```
predict ppred, xb
gen ppred1 = ppred if exposure == 1
gen ppred2 = ppred if exposure == 2
gen ppred3 = ppred if exposure == 3
sc cap1 cap2 cap3 age || line ppred1 age || line ppred2 age || line ppred3 age
```

Note that the final command (`scatter ...`) must all be entered on one line. This will show the same plot of capacity against age we saw before, but with three parallel regression lines, one for each group.

3.5 Finally, we wish to test the hypothesis that subjects with high exposure lose their vital capacity quicker as they age, i.e. that there is an interaction between age and vital capacity. We can do this with the command `regress capacity i.exposure##c.age` followed by `testparm i.exposure#c.age` Is there a significant difference between the slopes with age in the three groups ?

To see the regression lines in this case, type

```
predict ipred, xb
gen ipred1 = ipred if exposure == 1
gen ipred2 = ipred if exposure == 2
gen ipred3 = ipred if exposure == 3
sc cap1 cap2 cap3 age || line ipred1 age || line ipred2 age || line ipred3 age
```

Again, the final command (`scatter ...`) must all be entered on one line.

3.6 From the regression output, which group has the steepest slope with age and which group the least steep ?

3.7 Use `lincom age + 3.exposure#c.age` to calculate the decrease in vital capacity per year increase in age in the highest exposure group.

4.4. Variable Selection

Use the dataset `$datadir/hald`. This contains data on the amount of heat evolved as cement hardens, as a function of the proportions of 4 chemicals in the composition of the cement.

4.1 Use forward selection to chose a model for predicting the amount of heat evolved. (Use `sw regress y x1 x2 x3 x4, pe(0.05)`) Which variables are included.

4.2 Now use backward elimination, using the command `sw regress y x1 x2 x3 x4, pr(0.05)` Does this select the same model ?

4.3 Choose a model using stepwise selection, with the command `sw regress y x1 x2 x3 x4, pe(0.05) pr(0.0500005)` Is this model the same as any or all of the previous models ?

4.4 Produce a correlation matrix for the x-variables using the command `corr x1 x2 x3 x4` What is the correlation between `x2` and `x4` ?

4.5 Does this help to explain why the different methods of variable selection produced different models ?

4.6 Fit all 4 predictors in a single model with `regress y x1 x2 x3 x4` Look at the F-statistic: is the fit of the model statistically significant ?

4.7 Look at the R^2 statistic: is this model good at predicting how much heat will be evolved ?

4.8 Look at the table of coefficients: how many of them are significant at the p= 0.05 level ?

4.5. Polynomial Regression

Use the data in `$datadir/growth`. This dataset gives the weight, in ounces, of a baby weighed weekly for the first 20 weeks of its life.

5.1 Plot the data with `scatter weight week` Does the plot appear to be linear, or is there evidence of curvature ?

5.2 Fit a straight line to the data with `regress weight week` Produce a partial residual plot with `cprplot week` Does this confirm what you thought previously ?

5.3 Generate an `week`^{2} term with the command `gen week2 = week*week` Add this term to the regression model, with `regress weight week week2` Does this improve the fit of the model ? (i.e. is `week2` a significant predictor ?)

5.4 Generate predicted values from this model with `predict pred2, xb`. Produce a graph of the observed and predicted values with `twoway scatter weight week || line pred2 week`

5.5 Continue to generate polynomial terms and add them to the regression model until the highest order term is no longer significant. What order of polynomial is required to fit this data ?

5.6 Produce a correlation matrix for the polynomial terms with `corr week*`. What is the correlation between `week` and `week2` ?

A. Old Stata Syntax

Before Stata had factor variables, there was a different process for generating indicator variables. I do not recommend *using* the old method, but it is probably worth being a little familiar with it, since you may well see it in existing do-files.

The main difference was that the old way required a prefix to any command using indicator variables with `xi:` (short for "eXpand Interactions". The `xi` command would generate a series of indicator variables, beginning with `_I`, and include them in the command in place of the categorical variable.

The second main difference was that the symbol for creating an interaction was `*` rather than `#`. So the interaction between two categorical variables `var1` and `var2` would be written as `i.var1*i.var2`, rather than `i.var1#i.var2`. There was no equivalent of the `c` prefix, so if `var2` was continuous, the interaction would be written `i.var1*var2`, rather than `i.var1#c.var2`.

Another major difference is that string variables could be used with `xi`, but only numeric variables can be used as factors. This is not a major drawback, since Stata 14 can show string labels for the levels of the factors in the output, rather than the numbers themselves (I prepared my notes with Stata 12, so only the numbers appear in my output).

A final difference is in how the reference category is selected, which we will meet in a few weeks. The new method is relatively intuitive and simple, the old method I still need to look up the details of the syntax 20 years after I first used it. Not that intuitive.

The differences between the two syntaxes are outlined in Table A.1.

	New syntax	Old Syntax
Prefix	none required	`xi:`
Variable type	Numeric	String or numeric
Interaction	#	*
Creates new variables	No	Yes

Table A.1.: Differences between old and new syntaxes for defining indicators

For full details, see `help fvvarlist` for the new syntax and `help xi` for the old syntax.