Session 4  
Linear Models in STATA and ANOVA

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strengths of Linear Relationships</td>
<td>4-2</td>
</tr>
<tr>
<td>A Note on Non-Linear Relationships</td>
<td>4-4</td>
</tr>
<tr>
<td>Multiple Linear Regression</td>
<td>4-5</td>
</tr>
<tr>
<td>Removal of Variables</td>
<td>4-8</td>
</tr>
<tr>
<td>Independent Samples t-test</td>
<td>4-10</td>
</tr>
<tr>
<td>f-test: Two Sample for Variances</td>
<td>4-12</td>
</tr>
<tr>
<td>Paired Samples t-test</td>
<td>4-13</td>
</tr>
<tr>
<td>One way ANOVA</td>
<td>4-15</td>
</tr>
<tr>
<td>Practical Session 4</td>
<td>4-17</td>
</tr>
</tbody>
</table>
SESSION 4: Linear Models in STATA and ANOVA

Strengths of Linear Relationships

In the previous session we looked at relationships between variables and the Line of Best Fit through the points on a plot. Linear Regression can tell us whether any perceived relationship between the variables is a significant one.

But what about the strength of a relationship? How tightly are the points clustered around the line?

The strength of a linear relationship can be measured using the Pearson Correlation Coefficient.

The values of the Correlation Coefficient can range from –1 to +1. The following table provides a summary of the types of relationship and their Correlation Coefficients:

<table>
<thead>
<tr>
<th>Linear Relationship</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect Negative</td>
<td>-1</td>
</tr>
<tr>
<td>Negative</td>
<td>-1 to 0</td>
</tr>
<tr>
<td>None</td>
<td>0</td>
</tr>
<tr>
<td>Positive</td>
<td>0 to +1</td>
</tr>
<tr>
<td>Perfect Positive</td>
<td>+1</td>
</tr>
</tbody>
</table>

The higher the Correlation Coefficient, regardless of sign, the stronger the linear relationship between the two variables.

From the GSS data set ‘gss91t.dta’, we can look at the linear relationships between the education of the respondent (educ), that of the parents (maeduc and paeduc), the age of the respondent (age), and the Occupational Prestige Score (prestg80).

In STATA, click on

Statistics ➢ Summaries, tables & tests ➢ Summary Statistics ➢ Pairwise correlations
All possible pairs of variables from your chosen list will have the Correlation Coefficient calculated.

Notice that, for each pair of variables, the number of respondents, \( N \), differs. This is because the default is to exclude missing cases \textbf{pairwise}; that is, if a respondent has missing values for some of the variables, he or she is removed from the Correlation calculations involving those variables, but is included in any others where there are valid values for both variables.
Using the Sig. (2-tailed) value, we can determine whether the Correlation is a significant one. The Null Hypothesis is that the Correlation Coefficient is zero (or close enough to be taken as zero), and we reject this at the 5% level if the significance is less than 0.05.

STATA flags the Correlation Coefficients with an asterisk if they are significant at the 5% level.

We can see in our example that there are significant positive Correlations for each pair of the education variables; age is significantly negatively correlated with each of them, and prestg80 has significant positive correlations with each. All these correlations are significant at the 1% level, with the education of mothers and fathers having the strongest relationship.

The remaining variable pairing, age and prestg80, does not have a significant linear relationship; the correlation coefficient of 0.007 is not significantly different from zero, as indicated by the significance level of 0.799. This is a formal test of what we saw in the scatter plot of prestg80 against age in the previous session, the points seemed randomly scattered.

A Note on Non-Linear Relationships

It must be emphasised that we are dealing with Linear Relationships. You may find that the correlation coefficient indicates no significant linear relationship between two variables, but they may have a Non-Linear Relationship which we are not testing for.

The following is the result of the correlation and scatter plot procedures performed on some hypothetical data.

As can be seen, the correlation coefficient is not significant, indicating no linear relationship, while the plot indicates a very obvious quadratic relationship. It is
always a good idea to check for relationships visually using graphics as well as using formal statistical methods!

**Multiple Linear Regression**

Simple Linear Regression looks at one dependent variable in terms of one independent (or explanatory) variable. When we want to 'explain' a dependent variable in terms of two or more independent variables we use Multiple Linear Regression.

Just as in Simple Linear Regression, the Least Squares method is used to estimate the Coefficients (the constant and the Bs) of the independent variables in the now more general equation:

\[ \text{dependent variable} = B_0 + B_1(\text{Independent Var1}) + B_2(\text{Independent Var2}) + ... \]

Use the dataset ‘gss91t.dta’ to investigate the effect of the respondent's age (age), sex (sex), education (educ) and spouse’s education (speduc) on the Occupational Prestige score (prestg80).

Firstly, we will produce scatter plots of the continuous variables by clicking on

**Graphics ➤ Scatterplot matrix**

Then we can produce some correlation coefficients by clicking on

**Statistics ➤ Summaries, tables & tests ➤ Summary Statistics ➤ Pairwise correlations**
We cannot see any unusual patterns in the Scatter Plots that would indicate relationships other than linear ones might be present. The correlations indicate that there are significant linear relationships between *prestg80* and the two education variables, but not *age*. However, there are also significant correlations between what will be our 3 continuous independent variables (*educ*, *speduc* and *age*). How will this affect the Multiple Regression?

We follow the same procedure as Simple Linear Regression; we click on:

**Statistics ➔ Linear Regression and related ➔ Linear regression**

Choose *prestg80* as the dependent variable

Choose *educ*, *speduc*, *age* and *sex* as the independent variables

Click OK
sex is not a continuous variable, but, as it is a binary variable, we can use it if we interpret the results with care. The following output is obtained.

The 2nd table is the Model Summary table, which tells us how well we are explaining the dependent variable, prestg80, in terms of the variables we have entered into the model; the figures here are sometimes called the Goodness of Fit statistics.

The figure in the row headed R-Squared is the proportion of variability in the dependent variable that can be explained by changes in the values of the independent variables. The higher this proportion, the better the model is fitting to the data.

The 1st table is the ANOVA table and it also indicates whether there is a significant Linear Relationship between the Dependent variable and the combination of the Explanatory variables; an F-Test is used to test the Null Hypothesis that there is no Linear Relationship. The F-Test is given as a part of the 2nd table. We can see in our example that, with a Significance value (Prob>F) of less than 0.05, we have evidence that there is a significant Linear Relationship.

In the 3rd table, the table of the coefficients, we have the figures that will be used in our equation. All 4 explanatory variables have been entered, but should they all be there? Looking at the 2 columns, headed t and P>|t|, we can see that the significance level for the variable speduc is more than 0.05. This indicates that, when the other variables (a constant, educ, age and sex) are used to explain the variability in prestg80, using speduc as well doesn’t help to explain it any better; the coefficient of speduc is not significantly different from zero. It is not needed in the model.

Recall that, when we looked at the correlation coefficients before fitting this model, educ and speduc were both significantly correlated with prestg80, but educ had the stronger relationship (0.520 compared to 0.355). In addition, the correlation between educ and speduc, 0.619, showed a stronger linear relationship. We should not be surprised, therefore, that the Multiple Linear
Regression indicates that using \textit{educ} to explain \textit{prestg80} means you don't need to use \textit{speduc} as well.

On the other hand, \textit{age} was not significantly correlated with \textit{prestg80}, but \textit{was} significantly correlated with both education variables. We find that it appears as a significant effect when combined with these variables in the Multiple Linear Regression.

\textbf{Removal of variables}

We now want to remove the insignificant variable \textit{speduc}, as its presence in the model affects the coefficients of the other variables.

We follow the same procedure as before and click on:

\textbf{Statistics} \textgreater{} \textbf{Linear Regression and related} \textgreater{} \textbf{Linear regression}

The output obtained now is as follows:

The output obtained now is as follows:
We can now see that R-squared has decreased to 0.293 from 0.3318. This is because we have removed the variable *speduc* from the regression model. The ANOVA table also shows that the combination of variables in each model has a significant Linear Relationship with *prestg80*.

Both *educ* and *age* remain significant in the model, however we see that *sex* has now become not significant. So we repeat the procedure but this time we remove *sex* from the model. Our final model is shown in the following output:

Therefore the regression equation is:

\[
prestg80 = 5.582 + (2.47 * educ) + (0.114 * age)
\]

So, for example, for a person aged 40 with 12 years of education, we estimate the Occupational Prestige score *prestg80* as:

\[
prestg80 = 5.582 + (2.47 * 12) + (0.114 * 40) = 39.782
\]
Independent Samples T-Test

Under the assumption that the variables are normal, how can we investigate relationships between variables where one is continuous?

For these tests, we will use the data set 'statlab.dta'.

In this data set, the children were weighed and measured (among other things) at the age of ten. We want to know whether there is any difference in the average heights of boys and girls at this age. We do this by performing a t-test.

We start by stating our Null Hypothesis:

\[ H_0: \text{We assume there is no difference between boys and girls in terms of their height} \]

The Alternative Hypothesis is the one used if the Null Hypothesis is rejected.

\[ H_a: \text{We assume there is difference between boys and girls in terms of their height} \]

To perform the t-test, click on:

Statistics ➤ Summaries, tables & tests ➤ Classical tests of hypotheses ➤ Group mean comparison test

We want to test for differences in the mean HEIGHTS of the children;

Move the variable \( cth \) to the Variable name area.

We want to look at differences in the heights of the two groups BOYS and GIRLS, and so the Group variable name is \( sex \).

Click OK.

Click or not? We need to do an F-test.
The first part of the output gives some summary statistics; the numbers in each group, and the mean, standard deviation, standard error and the confidence interval of the mean for the height. **STATA** also gives out the combined statistics for the 2 groups.

In the second part of the output, we have the actual t-test. **STATA** gives out two null hypotheses as well as all the possible alternative hypotheses that we could have. Depending on which test you are after, you could either use a 1-tailed t-test \((Ha: \text{diff}<0)\) or Ha: \text{diff}>0 or a 2-tailed t-test \((Ha: \text{diff} \neq 0)\).

Our Null Hypothesis says that there is no difference between the boys and girls in terms of their heights; in other words, we are testing whether the difference of -0.357, is **significantly different from zero**. If it is, we must reject the Null Hypothesis, and instead take the Alternative.

**STATA** calculates the t-value, the degrees of freedom and the Significance Level; we can then make our decision quickly based on the displayed Significance Level. We will use the 2-tailed test in our example.

If the Significance Level is less than 0.05, we reject the Null Hypothesis and take the Alternative Hypothesis instead.

In this case, with a Significance Level of 0.012, we say that there is evidence, at the 5% level, to suggest that there is a difference between the heights of boys and girls at age ten (the Alternative Hypothesis).

(From the output, you can see that we can also conclude that this difference is negative).
f-test: Two Sample for Variances

The f-Test performs a two-sample f-test to compare two population variances. To be able to use the t-test, we need to determine whether the two populations have the same variance or not. In such a case, use the f-test. The f-test compares the f-score to the f distribution.

In this case, the null hypothesis (H0) and the alternative hypothesis (Ha) are:

\[
H_0 : \text{the two populations have the same variance} \\
H_a : \text{the two populations do not have the same variance}
\]

If we look at the same variable cth, we can now determine whether we should have ticked the option ‘Unequal variance’ or not. This decision is based on an F-test which will check on the variance of the 2 populations.

To use the f-test click on

Statistics ➤ Summaries, tables & tests ➤ Classical tests of hypotheses ➤ Group variance comparison test

The following output is obtained.
The 1st table contains some summary statistics of the two groups.

In the 2nd part of the output, we have the F-test. A significance value (P>F) of 0.05 or more means that the Null Hypothesis of assuming equal variances is acceptable, and we therefore can use the default option 'Equal Variances' in the previous t-test; a significance value of less than 0.05 means that we have to check the option 'Unequal variances' when performing the t-test.

In this case, the significance value is comfortably above this threshold, and therefore equal variances are assumed.

**Paired Samples t-test**

Imagine you want to compare two groups that are somehow paired; for example, husbands and wives, or mothers and daughters. Knowing about this pairing structure gives extra information, and you should take account of this when performing the t-test.

In the data set ‘statlaba.dta’, we have the weights of the parents when their child was aged 10 in ftw and mtw. If we want to know if there is a difference between males and females in terms of weight, we can perform a Paired Samples T-Test on these two variables.

We start by stating our Null Hypothesis:

\[ H_0: \text{We assume there is no difference between the weights of the parents.} \]

The Alternative Hypothesis, is the one used if the Null Hypothesis is rejected.

\[ H_a: \text{We assume there is difference between the weights of the parents} \]
To perform the **t-test**, click on:

**Statistics ➔ Summaries, tables & tests ➔ Classical tests of hypotheses ➔ Two-sample mean comparison test**

![Image of t-test interface]

Choose the 2 variables that you want to test.

Do not choose if observations are paired.

Click OK.

---

As with the **Independent Samples T-Test**, we are first given some summary statistics. The **Paired Samples Test** table shows that the difference between the weights of the males and females is 34.09 – is this significantly different from zero?

We use this table just as we did in the **Independent Samples T-Test**, and since the **Sig. (2-tailed)** column shows a value of less than 0.05, we can say that there is evidence, at the 5% level, to reject the **Null Hypothesis** that there is no difference between the mothers and fathers in terms of their weight.
One-Way ANOVA

We now look at the situation where we want to compare several independent groups. For this we use a One-Way ANOVA (ANALYSIS OF VARIANCE).

We will make use of the data set ‘gss91t.dta’. We can split the respondents into three groups according to which category of the variable life they fall into; exciting, routine or dull. We want to know if there is any difference in the average years of education of these groups. Our Null Hypothesis is that there is no difference between them in terms of education.

We start by stating our Null Hypothesis:

\[ H_0: \text{We assume there is no difference between the level of education of the 3 groups.} \]

The Alternative Hypothesis, is the one used if the Null Hypothesis is rejected.

\[ H_a: \text{We assume there is difference between the level of education of the 3 groups.} \]

To perform the one way ANOVA, click on:

Statistics ➔ ANOVA/MANOVA ➔ One-way analysis of variance

Choose the response variable.

Choose the group or factor variable

Click to obtain some summary statistics

Click OK.
**STATA** produces output that enables us to decide whether to accept or reject the Null Hypothesis that there is no difference between the groups. But if we find evidence of a difference, we will not know where the difference lies.

For example, those finding life exciting may have a significantly different number of years in education from those finding life dull, but there may be no difference when they are compared to those finding life routine.

We therefore ask **STATA** to perform a further analysis for us, called **Bonferroni**.

The output produced by **STATA** is below.

<table>
<thead>
<tr>
<th>LIFE</th>
<th>Summary of EDUC</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Exciting</td>
<td>13.642032</td>
<td>3.1082011</td>
<td>433</td>
<td></td>
</tr>
<tr>
<td>Routine</td>
<td>12.44132</td>
<td>2.7258611</td>
<td>503</td>
<td></td>
</tr>
<tr>
<td>Dull</td>
<td>10.487805</td>
<td>3.2567004</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12.891505</td>
<td>3.0215245</td>
<td>977</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Analysis of Variance</th>
<th></th>
<th></th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>SS</td>
<td>582.720694</td>
<td>2</td>
<td>291.360347</td>
<td>34.08</td>
</tr>
<tr>
<td>Within groups</td>
<td>df</td>
<td>8327.7879</td>
<td>974</td>
<td>8.5500809</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>MS</td>
<td>8910.49949</td>
<td>976</td>
<td>9.12961013</td>
<td></td>
</tr>
</tbody>
</table>

Bartlett’s test for equal variances: chi2(2) = 9.0369  Prob>chi2 = 0.041

Comparison of EDUC by LIFE (Bonferroni)

<table>
<thead>
<tr>
<th>Row Mean-Column Mean</th>
<th>Exciting</th>
<th>Routine</th>
<th>Dull</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routine</td>
<td>-1.20068</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Dull</td>
<td>-3.15423</td>
<td>-1.95355</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The 1st table gives some summary statistics of the 3 groups.

The 2nd table gives the results of the **One-Way ANOVA**. A measure of the variability found between the groups is shown in the **Between Groups** line, while the **Within Groups** line gives a measure of how much the observations within each group vary. These are used to perform the **f-test** which we use to test our Null Hypothesis that there is no difference between the three groups in terms of their years in education.

We interpret the **f-test** in the same way as we did the **t-test**; if the significance (in the Prob>F column) is less than 0.05, we have evidence, at the 5% level, to reject the Null Hypothesis, and say that there is some difference between the groups. Otherwise, we accept our Null Hypothesis.
We can see from the output that the f-value of 34.08 has a significance of less than 0.0005, and therefore we reject the Null Hypothesis. The 3rd table then shows us where these differences lie.

**Bonferroni** creates subsets of the categories; if there is no difference between two categories, they are put into the same subset. We can say that, at the 5% level, all 3 categories are different as all significance levels are less than 0.05.

**Practical Session 4**

Use the data set `statlab.dta`.

1. Use correlation and regression to investigate the relationship between the weight of the child at age 10 (**ctw**) and some physical characteristics:
   - **cbw** child’s weight at birth
   - **cth** child’s height at age 10
   - **sex** child’s gender (coded 1 for girls, 2 for boys)

2. Repeat Question 1, but instead use the following explanatory variables:
   - **fth** Father’s height
   - **ftw** Father’s weight
   - **mth** Mother’s height
   - **mtw** Mother’s weight

Use the data set `gss91t.dta`.

3. Investigate the Linear Relationships between the following variables using Correlations:
   - **educ** Education of respondent
   - **maeduc** Education of respondent’s mother
   - **paeduc** Education of respondent’s father
   - **speduc** Education of respondent’s spouse

4. Using Linear Regression, investigate the influence of education and parental education on the choice of marriage partner (Dependent variable **speduc**). Use the variable **sex** to distinguish between any gender effects.

5. It is thought that the size of the family might affect educational attainment. Investigate this using **educ** and **sibs** (the number of siblings) in a Linear Regression.

6. Also investigate whether the education of the parents (**maeduc** and **paeduc**) affects the family size (**sibs**).
7. How does the result of question 6 influence your interpretation of question 5? Are you perhaps finding a spurious effect? Test whether sibs still has a significant effect on educ when maeduc and paeduc are included in the model.

8. Compute a new variable \( \text{pared} = (\text{maeduc} + \text{paeduc}) / 2 \), being the average years of education of the parents. By including pared, maeduc and paeduc in a Multiple Linear Regression, investigate which is the better predictor of educ; the separate measures or the combined measure.

Use the data set ‘statlab.a.dta’.

At the age of ten, the children in the sample were given two tests; the Peabody Picture Vocabulary Test and the Raven Progressive Matrices Test. Their scores are stored in the variables ctp and ctr.

Create a new variable called tests which is the sum of the two tests; this new variable will be used in the following questions.

In each of the questions below, state your Null and Alternative Hypotheses, which of the two you accept on the evidence of the relevant test, and the Significance Level.

9. Use an Independent Samples T-Test to decide whether there is any difference between boys and girls in terms of their scores.

10. By pairing the parents of the child, decide whether there is any difference between fathers and mothers in terms of the heights. (Use fth and mth).

11. The fathers’ occupation is stored in the variable fto, with the following categories:

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Professional</td>
</tr>
<tr>
<td>1</td>
<td>Teacher / Counsellor</td>
</tr>
<tr>
<td>2</td>
<td>Manager / Official</td>
</tr>
<tr>
<td>3</td>
<td>Self-employed</td>
</tr>
<tr>
<td>4</td>
<td>Sales</td>
</tr>
<tr>
<td>5</td>
<td>Clerical</td>
</tr>
<tr>
<td>6</td>
<td>Craftsman / Operator</td>
</tr>
<tr>
<td>7</td>
<td>Labourer</td>
</tr>
<tr>
<td>8</td>
<td>Service worker</td>
</tr>
</tbody>
</table>

Recode fto into a new variable, occgrp, with categories:
Attach suitable variable and value labels to this new variable.

Using a **One-Way ANOVA**, test whether there is any difference between the occupation groups, in terms of the test scores of their children.

Open the data set ‘sceli.dta’.

In the **SCELi** questionnaire, employees were asked to compare the current circumstances in their job with what they were doing five years previously. Various aspects were considered:

- **effort**: Effort put into job
- **promo**: Chances of promotion
- **secur**: Level of job security
- **skill**: Level of skill used
- **speed**: How fast employee works
- **super**: Tightness of supervision
- **tasks**: Variety of tasks
- **train**: Provision of training

They were asked, for each aspect, what, if any, change there had been. The codes used were:

1. Increase
2. No change
3. Decrease
7. Don’t know

The sex of the respondent is stored in the variable **gender**, (code 1 is male, and code 2 is female) and the age in **age**.

For each of the job aspects, change code 7, 'Don't know' to a missing value.

Choose one or more of the job aspects. For each choice, answer the following questions:

12. What proportion of the employees sampled are employees perceiving a decrease in the job aspect?
13. What proportion of the employees sampled are female employees perceiving an increase or no change in the job aspect?

14. Use a bar chart to illustrate graphically any differences in the pattern of response between males and females.

15. Is there a significant difference in the average ages of the male and female employees in this sample?

16. Choose one or more of the job aspects. For each choice, investigate whether the employees falling into each of the categories have differences in terms of their ages.