Chapter 11

Two-Way ANOVA

An analysis method for a quantitative outcome and two categorical explanatory variables.

If an experiment has a quantitative outcome and two categorical explanatory variables that are defined in such a way that each experimental unit (subject) can be exposed to any combination of one level of one explanatory variable and one level of the other explanatory variable, then the most common analysis method is **two-way ANOVA**. Because there are two different explanatory variables the effects on the outcome of a change in one variable may either not depend on the level of the other variable (additive model) or it may depend on the level of the other variable (interaction model). One common naming convention for a model incorporating a k-level categorical explanatory variable and an m-level categorical explanatory variable is "k by m ANOVA" or "k x m ANOVA". ANOVA with more that two explanatory variables is often called **multi-way ANOVA**. If a quantitative explanatory variable is also included, that variable is usually called a **covariate**.

In two-way ANOVA, the error model is the usual one of Normal distribution with equal variance for all subjects that share levels of both (all) of the explanatory variables. Again, we will call that common variance σ^2. And we assume independent errors.

Two-way (or multi-way) ANOVA is an appropriate analysis method for a study with a quantitative outcome and two (or more) categorical explanatory variables. The usual assumptions of Normality, equal variance, and independent errors apply.

The structural model for two-way ANOVA *with* interaction is that each combination of levels of the explanatory variables has its own population mean with no restrictions on the patterns. One common notation is to call the population mean of the outcome for subjects with level a of the first explanatory variable and level b of the second explanatory variable as μ_{ab}. The interaction model says that any pattern of μ's is possible, and a plot of those μ's could show any arbitrary pattern.

In contrast, the no-interaction (additive) model does have a restriction on the population means of the outcomes. For the no-interaction model we can think of the mean restrictions as saying that the effect on the outcome of any specific level change for one explanatory variable is the same for every fixed setting of the other explanatory variable. This is called an **additive model**. Using the notation of the previous paragraph, the mathematical form of the additive model is $\mu_{ac} - \mu_{bc} = \mu_{ad} - \mu_{bd}$ for any valid levels a, b, c, and d. (Also, $\mu_{ab} - \mu_{ac} = \mu_{db} - \mu_{dc}$.)

A more intuitive presentation of the additive model is a plot of the population means as shown in figure 11.1. The same information is shown in both panels. In each the outcome is shown on the y-axis, the levels of one factor are shown on the x-axis, and separate colors are used for the second factor. The second panel reverses the roles of the factors from the first panel. Each point is a population mean of the outcome for a combination of one level from factor A and one level from factor B. The lines are shown as dashed because the explanatory variables are categorical, so interpolation "between" the levels of a factor makes no sense. The parallel nature of the dashed lines is what tells us that these means have a relationship that can be called additive. Also the choice of which factor is placed on the x-axis does not affect the interpretation, but commonly the factor with more levels is placed on the x-axis. Using this figure, you should now be able to understand the equations of the previous paragraph. In either panel the change in outcome (vertical distance) is the same if we move between any two horizontal points along any dotted line.

Note that the concept of interaction vs. an additive model is the same for ANCOVA or a two-way ANOVA. In the additive model the effects of a change in

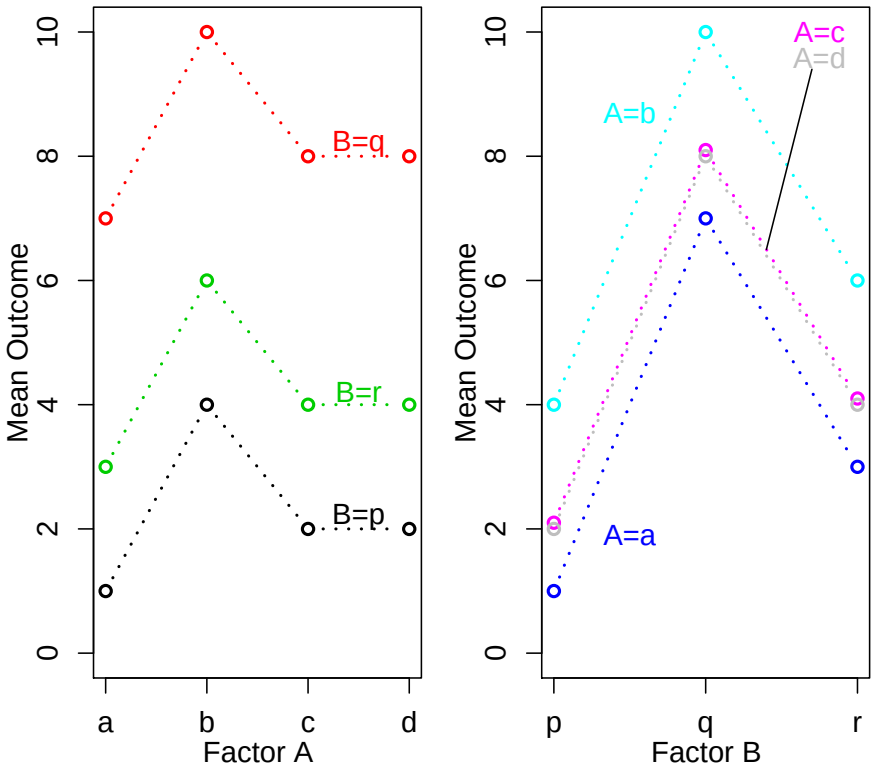

Figure 11.1: Population means for a no-interaction two-way ANOVA example.

one explanatory variable on the outcome does *not* depend on the value or level of the other explanatory variable, and the effect of a change in an explanatory variable can be described while not stating the (fixed) level of the other explanatory variable. And for the models underlying both analyses, if an interaction is present, the effects on the outcome of changing one explanatory variable *depends* on the specific value or level of the other explanatory variable. Also, the lines representing the mean of y at all values of quantitative variable x (in some practical interval) for each particular level of the categorical variable are all parallel (additive model) or not all parallel (interaction) in ANCOVA. In two-way ANOVA the order of the levels of the categorical variable represented on the x-axis is arbitrary and there is nothing between the levels, but nevertheless, if lines are drawn to aid the eye, these lines are all parallel if there is no interaction, and not all parallel if there is an interaction.

The two possible means models for two-way ANOVA are the additive model and the interaction model. The additive model assumes that the effects on the outcome of a particular level change for one explanatory variable does not depend on the level of the other explanatory variable. If an interaction model is needed, then the effects of a particular level change for one explanatory variable *does* depend on the level of the other explanatory variable.

A **profile plot**, also called an **interaction plot**, is very similar to figure 11.1, but instead the points represent the *estimates* of the population means for some data rather than the (unknown) true values. Because we can fit models with or without an interaction term, the same data will show different profile plots depending on which model we use. It is very important to realize that a profile plot from fitting a model without an interaction always shows the best possible parallel lines for the data, regardless of whether an additive model is adequate for the data, so this plot should not be used as EDA for choosing between the additive and interaction models. On the other hand, the profile plot from a model that includes the interaction shows the actual sample means, and is useful EDA for choosing between the additive and interaction models.

A profile plot is a way to look at outcome means for two factors simultaneously. The lines on this plot are meaningless, and only are an aid to viewing the plot. A plot drawn with parallel lines (or for which, given the size of the error, the lines could be parallel) suggests an additive model, while non-parallel lines suggests an interaction model.

11.1 Pollution Filter Example

This example comes from a statement by Texaco, Inc. to the Air and Water Pollution Subcommittee of the Senate Public Works Committee on June 26, 1973. Mr. John McKinley, President of Texaco, cited an automobile filter developed by Associated Octel Company as effective in reducing pollution. However, questions had been raised about the effects of filters on vehicle performance, fuel consumption, exhaust gas back pressure, and silencing. On the last question, he referred to the data in CarNoise.dat as evidence that the silencing properties of the Octel filter were at least equal to those of standard silencers.

This is an experiment in which the treatment "filter type" with levels "standard" and "octel" are randomly assigned to the experimental units, which are cars. Three types of experimental units are used, a small, a medium, or a large car, presumably representing three specific car models. The outcome is the quantitative (continuous) variable "noise". The categorical experimental variable "size" could best be considered to be a blocking variable, but it is also reasonable to consider it to be an additional variable of primary interest, although of limited generalizability due to the use of a single car model for each size.

A reasonable (initial) statistical model for these data is that for any combination of size and filter type the noise outcome is normally distributed with equal variance. We also can assume that the errors are independent if there is no serial trend in the way the cars are driven during the testing or in possible "drift" in the accuracy of the noise measurement over the duration of th experiment.

The means part of the structural model is either the additive model or the interaction model. We could either use EDA to pick which model to try first, or we could check the interaction model first, then switch to the additive model if the

		TYPE		
		Standard	Octel	Total
SIZE	small	6	6	12
	medium	6	6	12
	large	6	6	12
Total		18	18	36

Table 11.1: Cross-tabulation for car noise example.

interaction term is not statistically significant.

Some useful EDA is shown in table 11.1 and figures 11.2 and 11.3. The cross-tabulation lets us see that each **cell** of the experiment, i.e., each set of outcomes that correspond to a given set of levels of the explanatory variables, has six subjects (cars tested). This situation where there are the same number of subjects in all cells is called a **balanced design**. One of the key features of this experiment which tells us that it is OK to use the assumption of independent errors is that a different subject (car) is used for each test (row in the data). This is called a **between-subjects design**, and is the same as all of the studies described up to this point in the book, as contrasted with a within-subjects design in which each subject is exposed to multiple treatments (levels of the explanatory variables). For this experiment an appropriate within-subjects design would be to test each individual car with both types of filter, in which case a different analysis called within-subjects ANOVA would be needed.

The boxplots show that the small and medium sized cars have more noise than the large cars (although this may not be a good generalization, assuming that only one car model was testing in each size class). It appears that the Octel filter reduces the median noise level for medium sized cars and is equivalent to the standard filter for small and large cars. We also see that, for all three car sizes, there is less car-to-car variability in noise when the Octel filter is used.

The error bar plot shows mean plus or minus 2 SE. A good alternative, which looks very similar, is to show the 95% CI around each mean. For this plot, the standard deviations and sample sizes for each of the six groups are separately used to construct the error bars, but this is less than ideal if the equal variance assumption is met, in which case a pooled standard deviation is better. In this example, the best approach would be to use one pooled standard deviation for

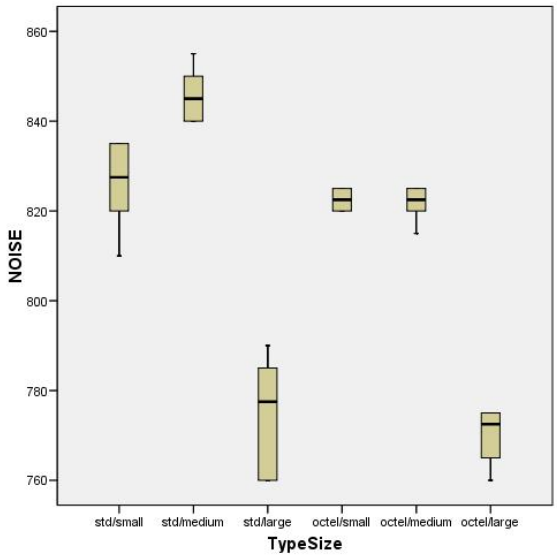

Figure 11.2: Side-by-side boxplots for car noise example.

each filter type.

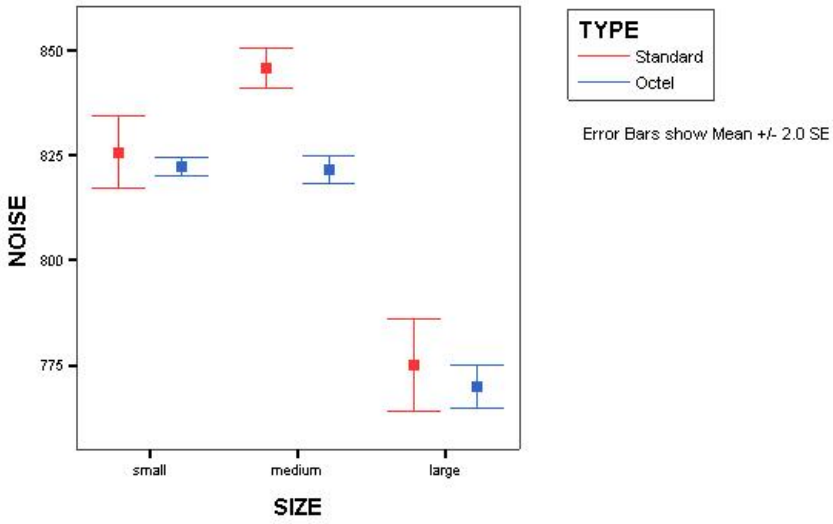

Figure 11.3: Error bar plot for car noise example.

Source	Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	27912	5	5582	85.3	<0.0005
SIZE	26051	2	13026	199.1	<0.0005
TYPE	1056	1	1056	16.1	<0.0005
SIZE*TYPE	804	2	402	6.1	<0.0005
Error	1962	30	65		
Corrected Total	29874	35			

Table 11.2: ANOVA for the car noise experiment.

11.2 Interpreting the two-way ANOVA results

The results of a two-way ANOVA of the car noise example are shown in tables 11.2 and 11.3. The ANOVA table is structured just like the one-way ANOVA table. The SS column represents the sum of squared deviations for each of several different ways of choosing which deviations to look at, and these are labeled "Source (of Variation)" for reasons that are discussed more fully below. Each SS has a corresponding df (degrees of freedom) which is a measure of the number of independent pieces of information present in the deviations that are used to compute the corresponding SS (see section 4.6). And each MS is the SS divided by the df for that line. Each MS is a variance estimate or a variance-like quantity, and as such its units are the squares of the outcome units.

Each F-statistic is the ratio of two MS values. For the between-groups ANOVA discussed in this chapter, the denominators are all MS_{error} (MSE) which corresponds exactly to $\text{MS}_{\text{within}}$ of the one-way ANOVA table. MSE is a "pure" estimate of σ^2, the common group variance, in the sense that it is unaffected by whether or not the null hypothesis is true. Just like in one-way ANOVA, a component of SS_{error} is computed for each treatment cell as deviations of individual subject outcomes from the sample mean of all subjects in that cell; the component df for each cell is $n_{ij} - 1$ (where n_{ij} is the number of subjects exposed to level i of one explanatory variable and level j of the other); and the SS and df are computed by summing over all cells.

Each F-statistic is compared against it's null sampling distribution to compute a p-value. Interpretation of each of the p-values depends on knowing the null hypothesis for each F-statistic, which corresponds to the situation for which the

numerator MS has an expected value σ^2.

The ANOVA table has lines for each main effect, the interaction (if included) and the error. Each of these lines demonstrates MS=SS/df. For the main effects and interaction, there are F values (which equal that line's MS value divided by the error MS value) and corresponding p-values.

The ANOVA table analyzes the total variation of the outcome in the experiment by decomposing the SS (and df) into components that add to the total (which only works because the components are what is called orthogonal). One decomposition visible in the ANOVA table is that the SS and df add up for "Corrected model" + "Error" = "Corrected Total". When interaction is included in the model, this decomposition is equivalent to a one-way ANOVA where all of the ab cells in a table with a levels of one factor and b levels of the other factor are treated as ab levels of a single factor. In that case the values for "Corrected Model" correspond to the "between-group" values of a one-way ANOVA, and the values for "Error" correspond to the "within-group" values. The null hypothesis for the "Corrected Model" F-statistic is that all ab population cell means are equal, and the deviations involved in the sum of squares are the deviations of the cell sample means from the overall mean. Note that this has $ab - 1$ df. The "Error" deviations are deviations of the individual subject outcome values from the group means. This has $N - ab$ df. In our car noise example $a = 2$ filter types, $b = 3$ sizes, and $N = 36$ total noise tests run.

SPSS gives two useless lines in the ANOVA table, which are not shown in figure 11.2. These are "Intercept" and "Total". Note that most computer programs report what SPSS calls the "Corrected Total" as the "Total".

The rest of the ANOVA table is a decomposition of the "Corrected Model" into main effects for size and type, as well as the interaction of size and type (size*type). You can see that the SS and df add up such that "Corrected Model" = "size" + "type" + "size*type". This decomposition can be thought of as saying that the deviation of the cell means from the overall mean is equal to the size deviations plus the type deviations plus any deviations from the additive model in the form of interaction.

In the presence of an interaction, the p-value for the interaction is most im-

portant and the main effects p-values are generally ignored if the interaction is significant. This is mainly because if the interaction is significant, then some changes in *both* explanatory variables must have an effect on the outcome, regardless of the main effect p-values. The null hypothesis for the interaction F-statistic is that there is an additive relationship between the two explanatory variables in their effects on the outcome. If the p-value for the interaction is less than alpha, then we have a statistically significant interaction, and we have evidence that any non-parallelness seen on a profile plot is "real" rather than due to random error.

A typical example of a statistically significant interaction with statistically non-significant main effects is where we have three levels of factor A and two levels of factor B, and the pattern of effects of changes in factor A is that the means are in a "V" shape for one level of B and an inverted "V" shape for the other level of B. Then the main effect for A is a test of whether at all three levels of A the mean outcome, averaged over both levels of B are equivalent. No matter how "deep" the V's are, if the V and inverted V are the same depth, then the mean outcomes averaged over B for each level of A are the same values, and the main effect of A will be non-significant. But this is usually misleading, because changing levels of A has big effects on the outcome for either level of B, but the effects differ depending on which level of B we are looking at. See figure 11.4.

If the interaction p-value is statistically significant, then we conclude that the effect on the mean outcome of a change in one factor *depends* on the level of the other factor. More specifically, for at least one pair of levels of one factor the effect of a particular change in levels for the other factor depends on which level of the first pair we are focusing on. More detailed explanations require "simple effects testing", see chapter 13.

In our current car noise example, we explain the statistically significant interaction as telling us that the population means for noise differ between standard and Octel filters for at least one car size. Equivalently we could say that the population means for noise differ among the car sizes for at least one type of filter.

Examination of the plots or the Marginal Means table suggests (but does not prove) that the important difference is that the noise level is higher for the standard

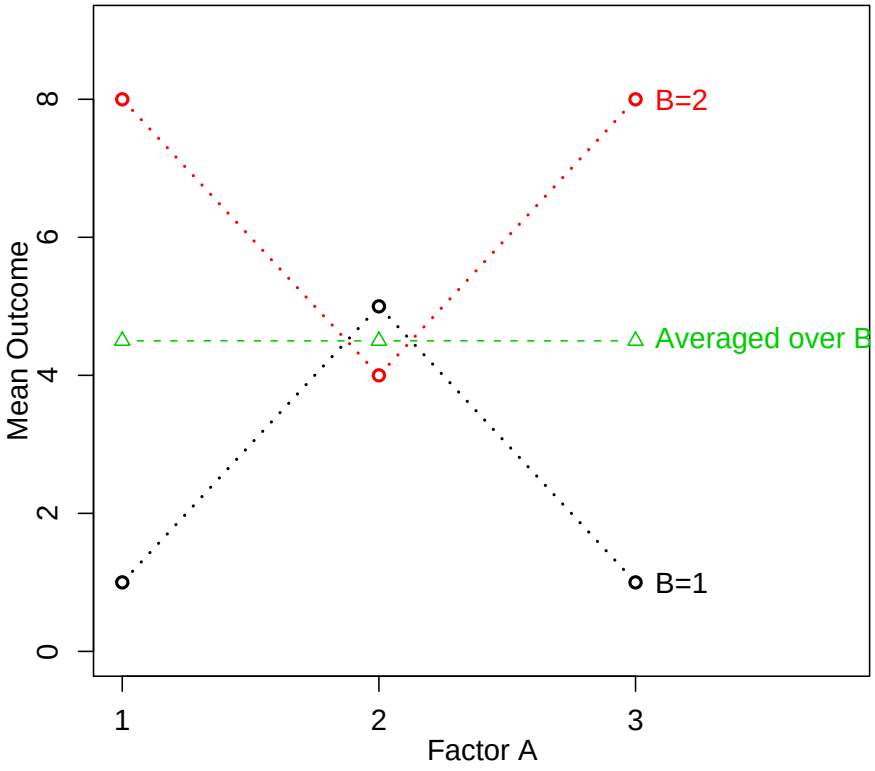

Figure 11.4: Significant interaction with misleading non-significant main effect of factor A.

				95% Confidence Interval	
SIZE	TYPE	Mean	Std. Error	Lower Bound	Upper Bound
small	Standard	825.83	3.30	819.09	832.58
	Octel	822.50	3.30	815.76	829.24
medium	Standard	845.83	3.30	839.09	852.58
	Octel	821.67	3.30	814.92	828.41
large	Standard	775.00	3.30	768.26	781.74
	Octel	770.00	3.30	763.26	776.74

Table 11.3: Estimated Marginal Means for the car noise experiment.

filter than the Octel filter for the medium sized car, but the filters have equivalent effects for the small and large cars.

If the interaction p-value is not statistically significant, then in most situations most analysts would re-run the ANOVA without the interaction, i.e., as a main effects only, additive model. The interpretation of main effects F-statistics in a non-interaction two-way ANOVA is easy. Each main effect p-value corresponds to the null hypothesis that population means of the outcome are equal for all levels of the factor ignoring the other factor. E.g., for a factor with three levels, the null hypothesis is that $H_0 : \mu_1 = \mu_2 = \mu_3$, and the alternative is that at least one population mean differs from the others. (Because the population means for one factor are averaged over the levels of the other factor, unbalanced sample sizes can give misleading p-values.) If there are only two levels, then we can and should immediately report which one is "better" by looking at the sample means. If there are more than two levels, we can only say that there are some differences in mean outcome among the levels, but we need to do additional analysis in the form of "contrast testing" as shown in chapter 13 to determine which levels are statistically significantly different.

Inference for the two-way ANOVA table involves first checking the interaction p-value to see if we can reject the null hypothesis that the additive model is sufficient. If that p-value is smaller than α then the adequacy of the additive model can be rejected, and you should conclude that both factors affect the outcome, and that the effect of changes in one factor *depends* on the level of the other factor, i.e., there is an interaction between the explanatory variables. If the interaction p-value is larger than α, then you can conclude that the additive model is adequate, and you should re-run the analysis without an interaction term, and then interpret each of the p-values as in one-way ANOVA, realizing that the effects of changes in one factor are the same at every fixed level of the other factor.

It is worth noting that a transformation, such as a log transformation of the outcome, would not correct the unequal variance of the outcome across the groups defined by treatment combinations for this example (see figure 11.2). A log transformation corrects unequal variance only in the case where the variance is larger for groups with larger outcome means, which is not the case here. Therefore,

other than using much more complicated analysis methods which flexibly model changes in variance, the best solution to the problem of unequal variance in this example, is to use the "Keppel" correction which roughly corrects for moderate degrees if violation of the equal variance assumption by substituting $\alpha/2$ for α. For this problem, we still reject the null hypothesis of an additive model when we compare the p-value to 0.025 instead of 0.05, so the correction does not change our conclusion.

Figure 11.5 shows the 3 by 3 residual plot produced in SPSS by checking the Option "Residual plot". The middle panel of the bottom row shows the usual residual vs. fit plot. There are six vertical bands of residual because there are six combinations of filter level and size level, giving six possible predictions. Check the equal variance assumption in the same way as for a regression problem. Verifying that the means for all of the vertical bands are at zero is a check that the mean model is OK. For two-way ANOVA this comes down to checking that dropping the interaction term was a reasonable thing to do. In other words, if a no-interaction model shows a pattern to the means, the interaction is probably needed. This default plot is poorly designed, and does not allow checking Normality. I prefer the somewhat more tedious approach of using the Save feature in SPSS to save predicted and residual values, then using these to make the usual full size residual vs. fit plot, plus a QN plot of the residuals to check for Normality.

Residual checking for two-way ANOVA is very similar to regression and one-way ANOVA.

11.3 Math and gender example

The data in mathGender.dat are from an observational study carried out to investigate the relationship between the ACT Math Usage Test and the explanatory variables gender (1=female, 2=male) and level of mathematics coursework taken (1=algebra only, 2=algebra+geometry, 3=through calculus) for 861 high school seniors. The outcome, ACT score, ranges from 0 to 36 with a median of 15 and a mean of 15.33. An analysis of these data of the type discussed in this chapter can be called a 3x2 ("three by two") ANOVA because those are the numbers of levels of the two categorical explanatory variables.

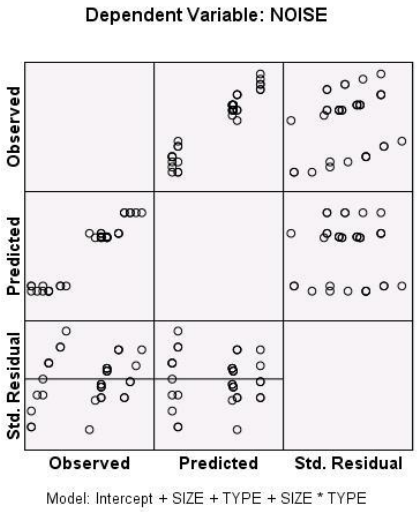

Figure 11.5: Residual plots for car noise example.

The rows of the data table (experimental units) are individual students. There is some concern about independent errors if the 861 students come from just a few schools, with many students per school, because then the errors for students from the same school are likely to be correlated. In that case, the p-values and confidence intervals will be unreliable, and we should use an alternative analysis such as mixed models, which takes the clustering into schools into account. For the analysis below, we assume that student are randomly sampled throughout the country so that including two students from the same school would only be a rare coincidence.

This is an observational study, so our conclusions will be described in terms of association, not causation. Neither gender nor coursework was randomized to different students.

The cross-tabulation of the explanatory variables is shown in table 11.4. As opposed to the previous example, this is not a balanced ANOVA, because it has unequal cell sizes.

Further EDA shows that each of the six cells has roughly the same variance for the test scores, and none of the cells shows test score skewness or kurtosis suggestive of non-Normality.

		Gender		
		Female	Male	Total
Coursework	algebra	82	48	130
	to geometry	387	223	610
	to calculus	54	67	121
Total		523	338	861

Table 11.4: Cross-tabulation for the math and gender example.

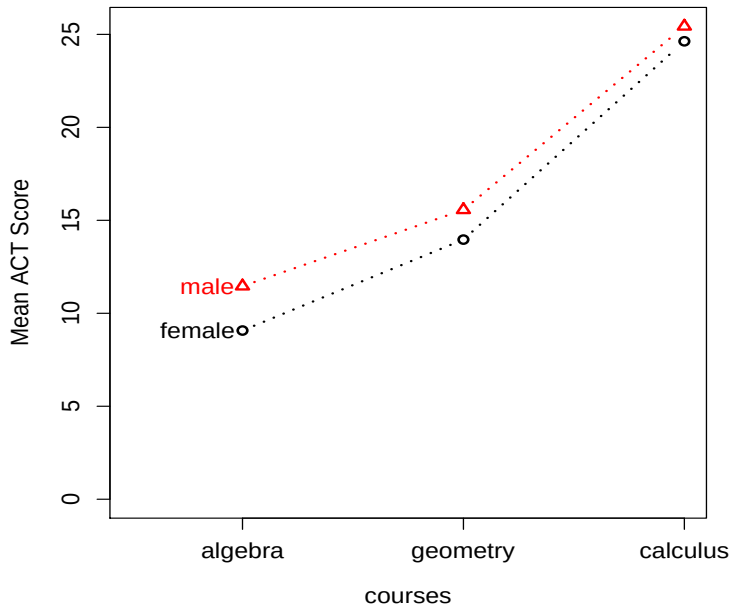

Figure 11.6: Cell means for the math and gender example.

Source	Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16172.8	5	3234.6	132.5	<0.0005
courses	14479.5	2	7239.8	296.5	<0.0005
gender	311.9	1	311.9	12.8	<0.0005
courses*gender	37.6	2	18.8	0.8	0.463
Error	20876.8	855	24.4		
Corrected Total	37049.7	860	43.1		

Table 11.5: ANOVA with interaction for the math and gender example.

A profile plot of the cell means is shown in figure 11.6. The first impression is that students who take more courses have higher scores, males have slightly higher scores than females, and perhaps the gender difference is smaller for students who take more courses.

The two-way ANOVA with interaction is shown in table 11.5.

The deviations used in the sums of squared deviations (SS) in a two-way ANOVA with interaction are just a bit more complicated than in one-way ANOVA. The main effects deviations are calculated as in one-way interaction, just ignoring the other factor. Then the interaction SS is calculated by using the main effects to construct the best "parallel pattern" means and then looking at the deviations of the actual cell means from the best "parallel pattern means".

The interaction line of the table (courses*gender) has 2 df because the difference between an additive model (with a parallel pattern of population means) and an interaction model (with arbitrary patterns) can be thought of as taking the parallel pattern, then moving any two points for any one gender. The formula for interaction df is $(k-1)(m-1)$ for any k by m ANOVA.

As a minor point, note that the MS is given for the "Corrected Total" line. Some programs give this value, which equals the variance of all of the outcomes ignoring the explanatory variables. The "Corrected Total" line adds up for both the SS and df columns but not for the MS column, to either "Corrected Model" + "Error" or to all of the main effects plus interactions plus the Error.

Source	Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16135.2	3	5378.4	220.4	<0.0005
courses	14704.7	2	7352.3	301.3	<0.0005
gender	516.6	1	516.6	21.2	<0.0005
Error	20914.5	857	24.4		
Corrected Total	37049.7	860			

Table 11.6: ANOVA without interaction for the math and gender example.

The main point of this ANOVA table is that the interaction between the explanatory variables gender and courses is not significant (F=0.8, p=0.463), so we have no evidence to reject the additive model, and we conclude that course effects on the outcome are the same for both genders, and gender effects on the outcome are the same for all three levels of coursework. Therefore it is appropriate to re-run the ANOVA with a different means model, i.e., with an additive rather than an interactive model.

The ANOVA table for a two-way ANOVA without interaction is shown in table 11.6.

Our conclusion, using a significance level of $\alpha = 0.05$ is that both courses and gender affect test score. Specifically, because gender has only two levels (1 df), we can directly check the Estimated Means table (table 11.7) to see that males have a higher mean. Then we can conclude based on the small p-value that being male is associated with a higher math ACT score compared to females, for each level of courses. This is not in conflict with the observation that some females are better than most males, because it is only a statement about means. In fact the estimated means table tells us that the mean difference is 2.6 while the ANOVA table tells us that the standard deviation in any group is approximately 5 (square root of 24.4), so the overlap between males and females is quite large. Also, this kind of study certainly cannot distinguish differences due to biological factors from those due to social or other factors.

Looking at the p-value for courses, we see that at least one level of courses differs from the other two, and this is true separately for males and females because the additive model is an adequate model. But we cannot make further important statements about which levels of courses are significantly different without additional analyses, which are discussed in chapter 13.

courses	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
algebra	10.16	0.44	9.31	11.02
to geometry	14.76	0.20	14.36	15.17
to calculus	14.99	0.45	24.11	25.87

gender	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
female	14.84	0.26	15.32	16.36
male	17.44	0.30	16.86	18.02

Table 11.7: Estimated means for the math and gender example.

We can also note that the residual (within-group) variance is 24.4, so our estimate of the population standard deviation for each group is $\sqrt{24.4} = 4.9$. Therefore about 95% of test scores for any gender and level of coursework are within 9.8 points of that group's mean score.

11.4 More on profile plots, main effects and interactions

Consider an experiment looking at the effects of different levels of light and sound on some outcome. Five possible outcomes are shown in the profile plots of figures 11.7, 11.8, 11.9, 11.10, and 11.11 which include plus or minus 2 SE error bars (roughly 95% CI for the population means).

Table 11.8 shows the p-values from two-way ANOVA's of these five cases.

In case A you can see that it takes very little "wiggle", certainly less than the size of the error bars, to get the lines to be parallel, so an additive model should be OK, and indeed the interaction p-value is 0.802. We should re-fit a model without an interaction term. We see that as we change sound levels (move left or right), the mean outcome (y-axis value) does not change much, so sound level does not affect the outcome and we get a non-significant p-value (0.971). But changing light levels (moving from one colored line to another, at any sound level) does change the mean outcome, e.g., high light gives a low outcome, so we expect a significant p-value for light, and indeed it is <0.0005.

Case	light	sound	interaction
A	<0.0005	0.971	0.802
B	0.787	0.380	0.718
C	<0.0005	<0.0005	<0.0005
D	<0.0005	<0.0005	0.995
E	0.506	<0.0005	0.250

Table 11.8: P-values for various light/sound experiment cases.

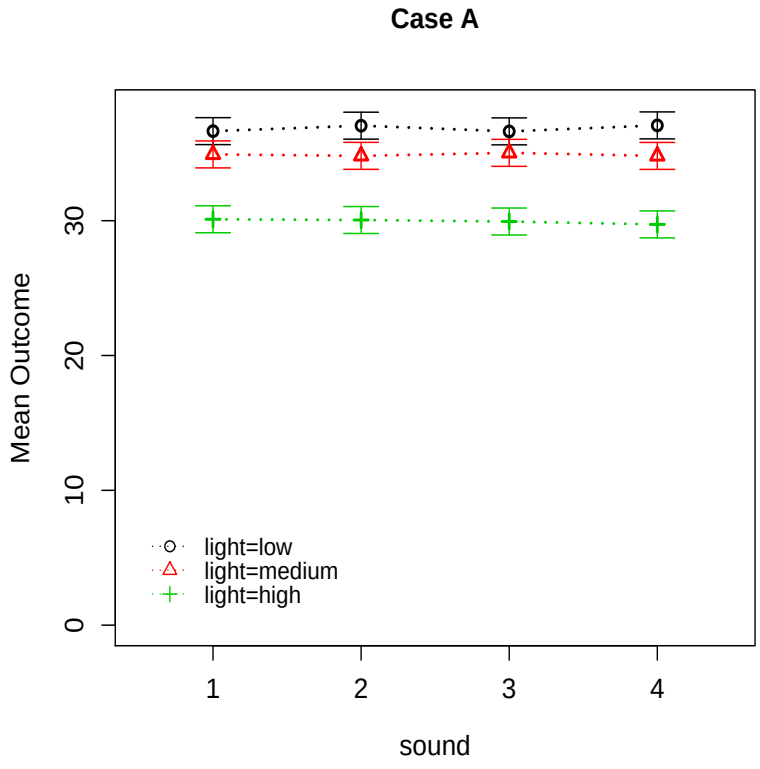

Figure 11.7: Case A for light/sound experiment.

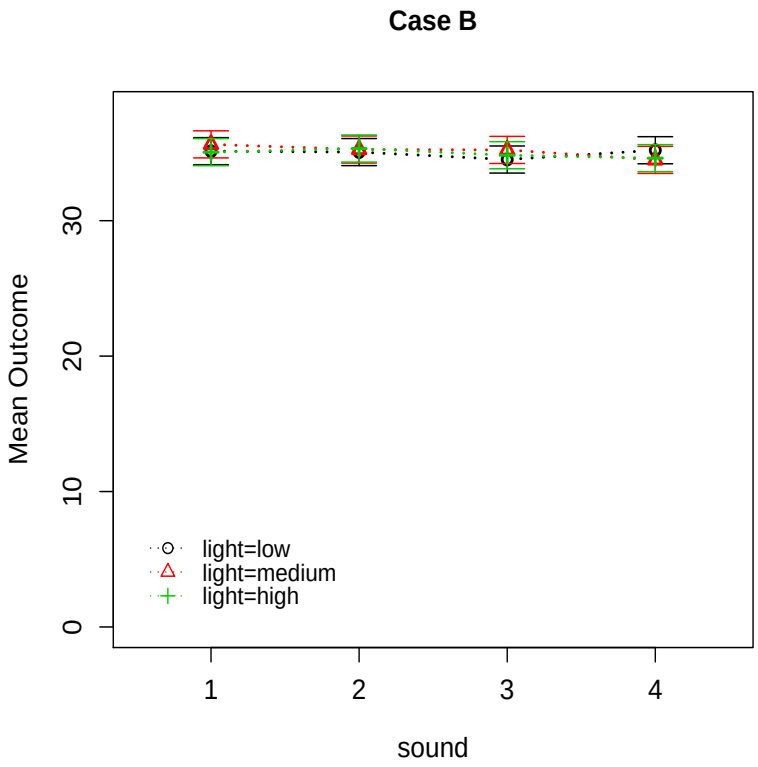

Figure 11.8: Case B for light/sound experiment.

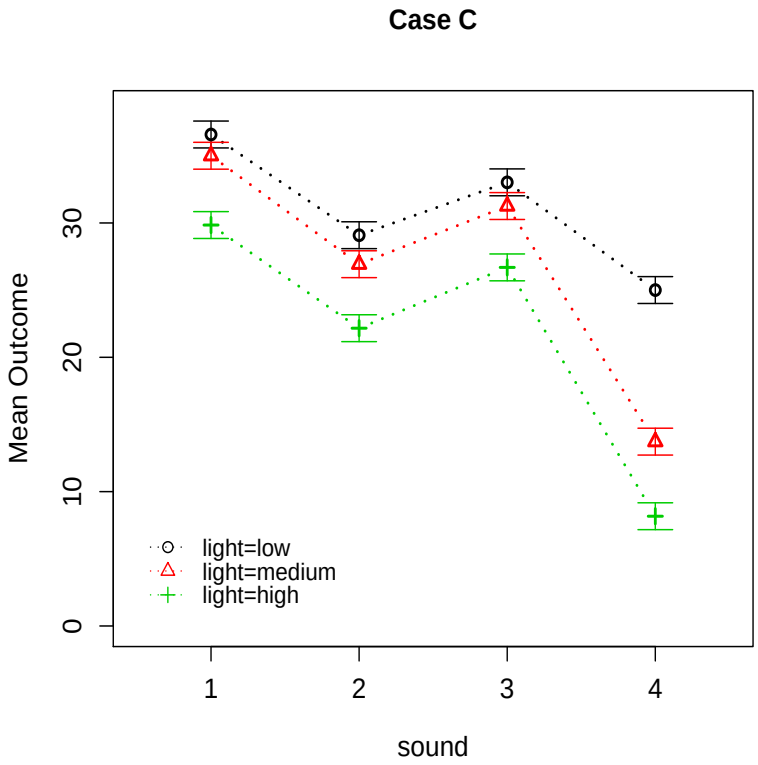

Figure 11.9: Case C for light/sound experiment.

In case B, as in case A, the lines are nearly parallel, suggesting that an additive, no-interaction model is adequate, and we should re-fit a model without an interaction term. We also see that changing sound levels (moving left or right on the plot) has no effect on the outcome (vertical position), so sound is not a significant explanatory variable. Also changing light level (moving between the colored lines) has no effect. So all the p-values are non-significant (>0.05).

In case C, there is a single cell, low light with sound at level 4, that must be moved much more than the size of the error bars to make the lines parallel. This is enough to give a significant interaction p-value (<0.0005), and require that we stay with this model that includes an interaction term, rather than using an additive model. The p-values for the main effects now have no real interest. We know that both light and sound affect the outcome because the interaction p-value is significant. E.g., although we need contrast testing to be sure, it is quite obvious

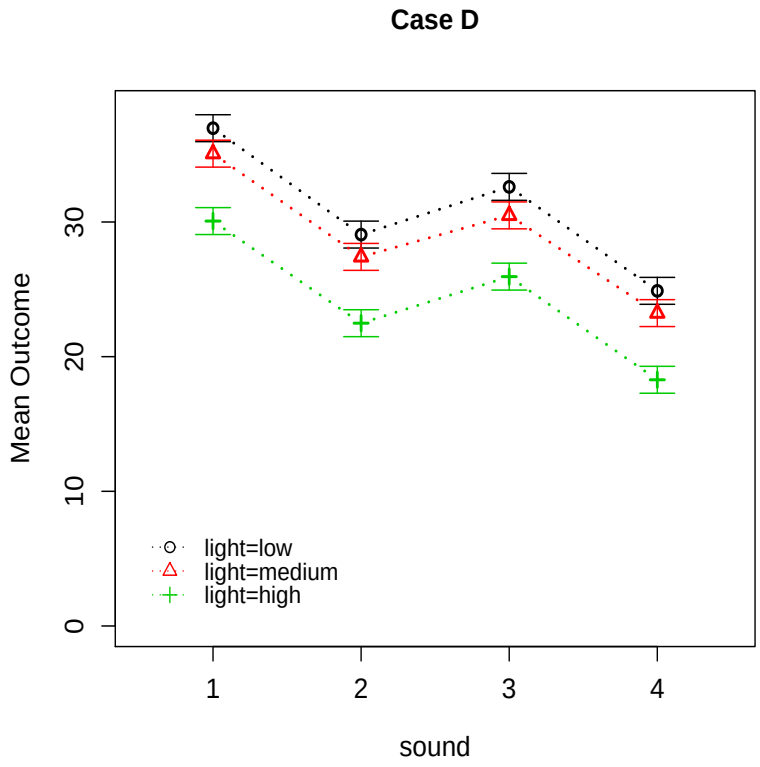

Figure 11.10: Case D for light/sound experiment.

that changing from low to high light level for any sound level lowers the outcome, and changing from sound level 3 to 4 for any light level lowers the outcome.

Case D shows no interaction (p=0.995) because on the scale of the error bars, the lines are parallel. Both main effects are significant.because for either factor, at at least one level of the other factor there are two levels of the first factor for which the outcome differs.

Case E shows no interaction. The light factor is not statistically significant as shown by the fact that for any sound level, changing light level (moving between colored lines) does not change the outcome. But the sound factor is statistically significant because changing between at least some pairs of sound levels for any light level does affect the outcome.

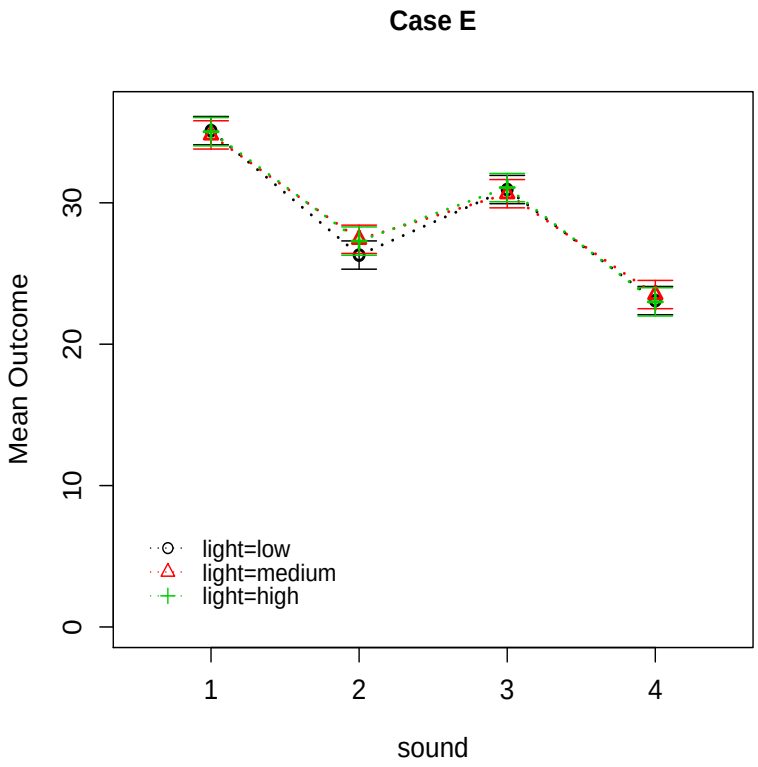

Figure 11.11: Case E for light/sound experiment.

Taking error into account, in most cases you can get a good idea which p-values will be significant just by looking at a (no-interaction) profile plot.

11.5 Do it in SPSS

To perform two-way ANOVA in SPSS use Analyze/GeneralLinearModel/Univariate from the menus. The "univariate" part means that there is only one kind of outcome measured for each subject. In this part of SPSS, you do *not* need to manually code indicator variables for categorical variables, or manually code interactions.

The Univariate dialog box is shown in figure 11.12. Enter the quantitative outcome in the Dependent Variable box. Enter the categorical explanatory variables in the Fixed Factors box. This will fit a model *with* an interaction.

Figure 11.12: SPSS Univariate dialog box.

To fit a model without an interaction, click the Model button to open the Univariate:Model dialog box, shown in figure 11.13. From here, choose "Custom"

instead of "Full Factorial", then do whatever it takes (there are several ways to do this) to get both factors, but not the interaction into the "Model" box, then click Continue.

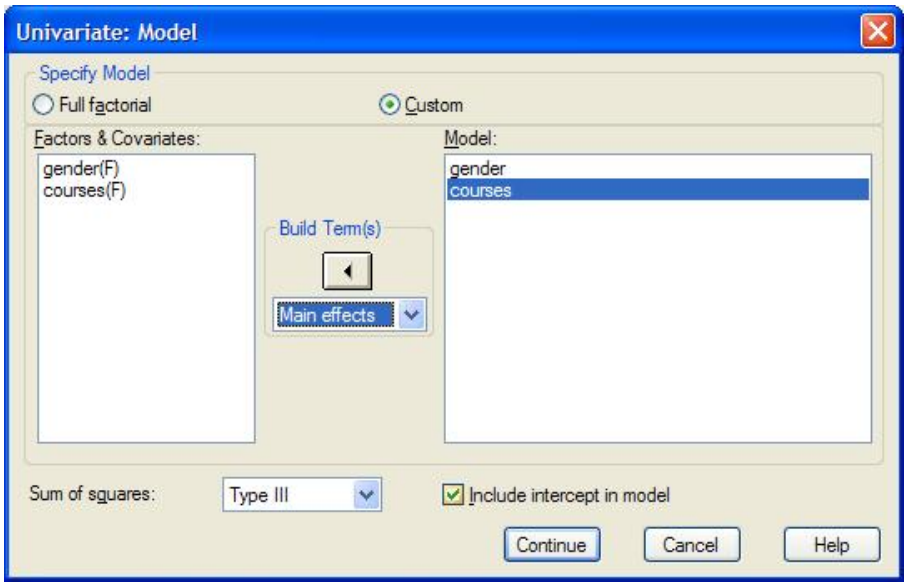

Figure 11.13: SPSS Univariate:Model dialog box.

For either model, it is a good idea to go to Options and turn on "Descriptive statistics", and "Residual plot". The latter is the 3 by 3 plot in which the usual residual vs. fit plot is in the center of the bottom row. Also place the individual factors in the "Display Means for" box if you are fitting a no-interaction model, or place the interaction of the factors in the box if you are fitting a model with an interaction.

If you use the Save button to save predicted and residual values (either standardized or unstandardized), this will create new columns in you data sheet; then a scatter plot with predicted on the x-axis and residual on the y-axis gives a residual vs. fit plot, while a quantile-normal plot of the residual column allows you to check the Normality assumption.

Under the Plots button, put one factor (usually the one with more levels) in the "Horizontal Axis" box, and the other factor in the "Separate Lines" box, then click Add to make an entry in the Plots box, and click Continue.

Finally, click OK in the main Univariate dialog box to perform the analysis.