Chapter 12: Linear models with two explanatory factor variables

STATS 201/8

University of Auckland

Learning Outcomes

In this chapter you will learn about:

- Two explanatory factors—Two-way analysis of variance
- Relevant R-code.

Two explanatory factors (Two-way analysis of variance)

Exam score vs test success and attendance

Here we are using the same two explanatory variables as in Chapter 8 but we are going to change the explanatory test score variable so that it only has two states — did they pass the test or not?

That is, we are going to use the dichotomous factor variable "test success", rather than the raw test score value.

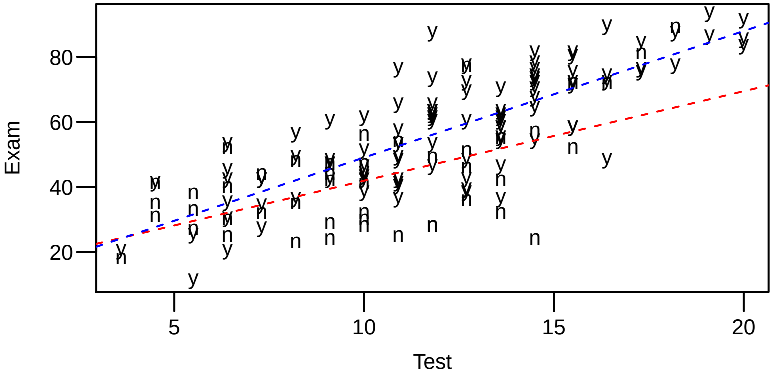
We shall also be using attendance as a second explanatory factor.

NOTE: When people use the term ANOVA (Analysis of Variance), they are typically referring to a linear regression in which all the explanatory variables are factors, as is the case here.

The example we are using here would be called a two-way ANOVA, as there are two explanatory factors.

Exam score vs test success and attendance...

Plotting the data



In Case Study 9 we investigated whether the effect of a student's test mark on exam score changed depending on whether they regularly attend or not.

We saw that those who attended regularly (blue line and "y" for "yes") got more 'return' for their test mark.

Example—Exam vs test success and attendance

Here we are going to turn the `Test` variable into a factor with two levels: did they pass the test or not?

We can then ask whether passing the test results in better exam marks and vice-versa, on average. We will also ask the same question of regular attendance.

Let us create the new factor variable `Pass.test`:

```
> Stats20x.df$Pass.test = with(Stats20x.df,
+                              factor(ifelse(Test>=10,"pass","nopass")))
> ## Check to see if the call above does what we expect
> min(Stats20x.df$Test[Stats20x.df$Pass.test=="pass"])
[1] 10
> max(Stats20x.df$Test[Stats20x.df$Pass.test=="nopass"])
[1] 9.1
```

Exam vs test success and attendance

`interactionPlots()`

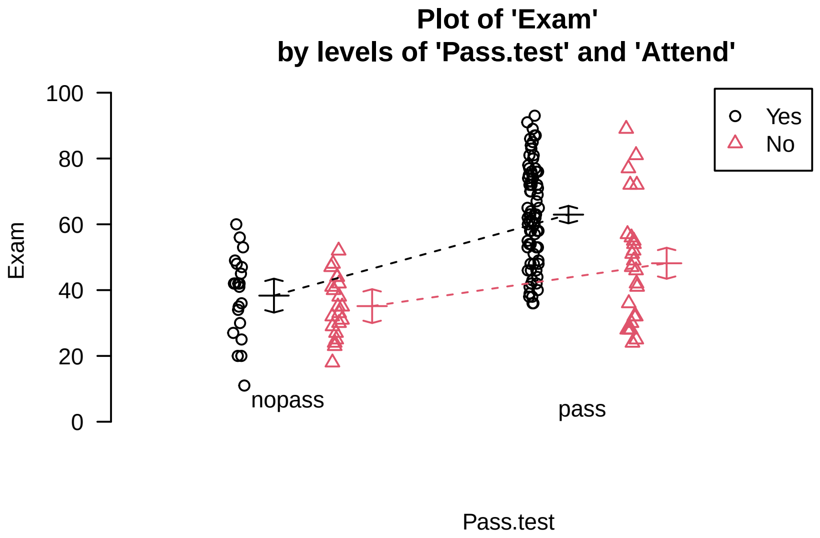
Let us see how these data explain `Exam` by using an `s20x` function `interactionPlot()`.

This is designed specifically for plotting a continuous Y (in our case `Exam`) against two factor variables (here they are `Attend` and the newly created `Pass.test`).

Exam vs test success and attendance

`interactionPlots()...`

```
> interactionPlots(Exam ~ Pass.test + Attend, data = Stats20x.df)
```



Exam vs test success and attendance

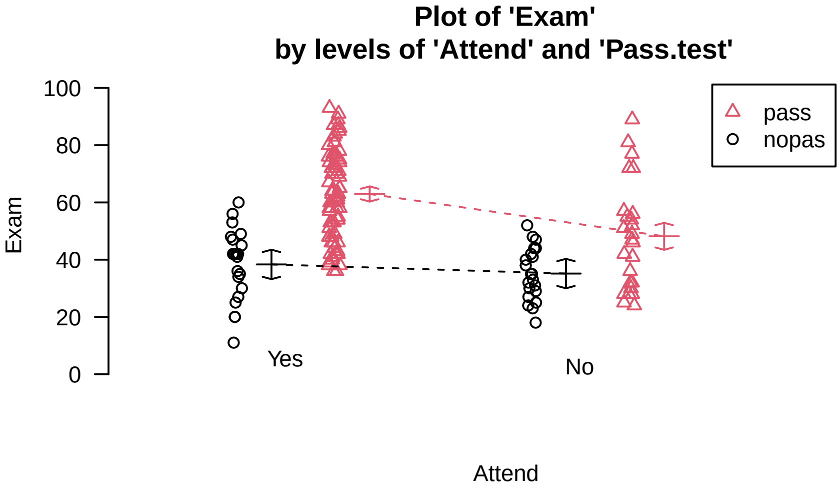
Here we see that 'attenders' who pass the test seem to be doing markedly better than most other students. Note that we do not have parallel lines here, indicating interaction between these two factors.

This means the effect of passing the test on exam may depend on whether a student regularly attended or not.

Exam vs test success and attendance

We can look at it in the opposite order, but would still conclude the same insights as above.

```
> interactionPlots(Exam ~ Attend + Pass.test, data = Stats20x.df)
```



Exam vs test success and attendance

Two explanatory factor variables each with 2 levels...

The *reference cell* model[1] formula for the two-way ANOVA is written as:

$$\texttt{Exam} = \beta_0 + \beta_1 \times \texttt{Attend}_{\texttt{Yes}} + \beta_2 \times \texttt{Pass.test}_{\texttt{pass}} + \beta_3 \times \texttt{Attend}_{\texttt{Yes}} \times \texttt{Pass.test}_{\texttt{pass}} + \varepsilon,$$

where $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$.

This model is relative to the baseline, or reference[2], levels of the `Attend` and `Pass.test` factor variables.

[1] We first encountered the reference cell model in Chapter 11.

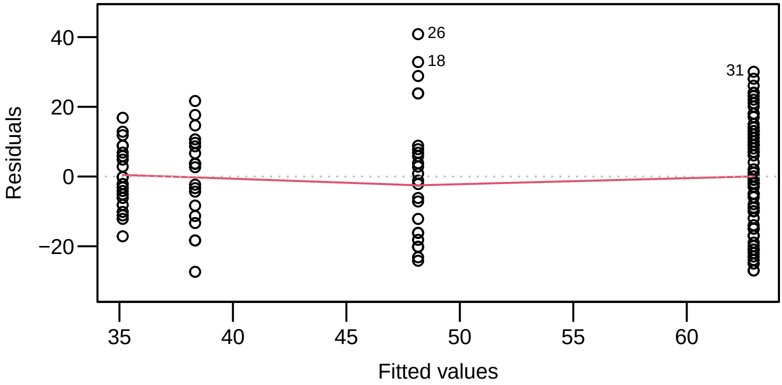
[2] By default, `R` sets the label with the lowest alphanumeric value as the reference level for each factor variable.

Exam vs test success and attendance

Assumption checks...

Let us fit the model with interaction, and check the assumptions.

```
> Exam.fit = lm(Exam ~ Attend * Pass.test, data = Stats20x.df)
> plot(Exam.fit, which = 1)  # needs sub.caption = '' to reproduce exactly the belo
```

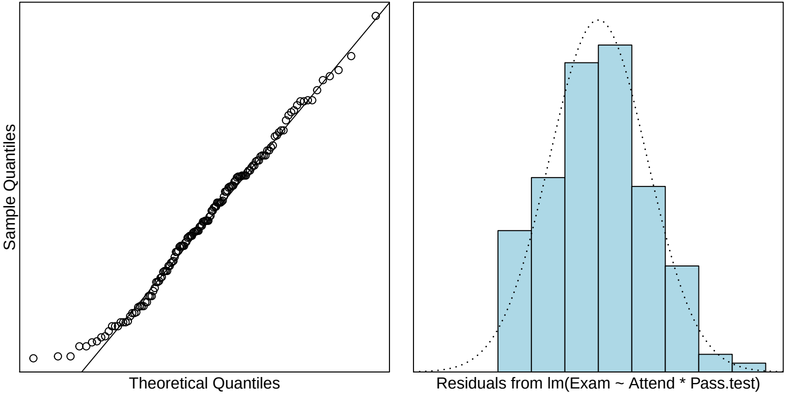


The **EOV** assumption seems to be okay.

Exam vs test success and attendance

Assumption checks...

```
> normcheck(Exam.fit)
```

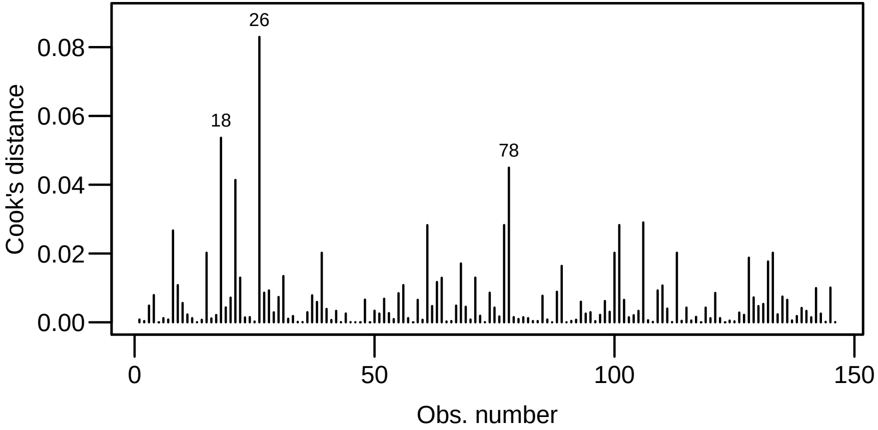


The Normality assumption seems to be okay.

Exam vs test success and attendance

Assumption checks. . .

```
> plot(Exam.fit, which = 4, cex.lab = 1.5)
```



No unduly influential data points.

Exam vs test success and attendance

We conclude that we can trust the output. Let us see what it is telling us.

```
> anova(Exam.fit)
Analysis of Variance Table

Response: Exam
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Attend            1  7630.8  7630.8  34.990 2.364e-08 ***
Pass.test         1 11076.9 11076.9  50.791 4.763e-11 ***
Attend:Pass.test  1   909.7   909.7   4.171   0.04297 *
Residuals       142 30968.4   218.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This simply confirms what we thought: The effect of passing the test depends on whether they have attended or not.

We cannot simply state the effect of passing the test, because the size of this effect depends on whether the student attended or not.

One way to think of this is that we have to consider all 4 (2×2) different test success/attendance possibilities separately.

Exam vs test success and attendance

Let us investigate what our model tells us in terms of the estimated parameters:

```
> summary(Exam.fit)
```

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               35.143      3.223  10.905  < 2e-16 ***
AttendYes                  3.190      4.557   0.700  0.48504
Pass.testpass             13.017      4.371   2.978  0.00341 **
AttendYes:Pass.testpass   11.599      5.679   2.042  0.04297 *
---
Residual standard error: 14.77 on 142 degrees of freedom
Multiple R-squared:  0.3878,Adjusted R-squared:  0.3749
F-statistic: 29.98 on 3 and 142 DF,  p-value: 4.452e-15
```

The P-value for interaction is the same as before.

Note also that the $R^2 = 39\%$ can be obtained from the ANOVA table above as follows: $R^2 = 100 \times \left(1 - \frac{30968}{30968+910+11077+7631}\right)$ is the proportion of variability that is explained by our model terms.

Exam vs test success and attendance

Interpreting the output. . .

In studies in which *all* of the explanatory variables are factors, our interest lies in making statistical inferences about the sizes of pairwise differences between means. So, for our Exam Score study, we could do this using the above output.[3] However, we will use the `emmeans()` function to perform these calculations for us.

We are interested in comparisons, or *contrasts*, between pairs of means for each treatment combination.[4] To do this we simply supply the in-built `pairwise` function in a formula when specifying the `specs` argument, i.e.

`specs = pairwise ~ Attend:Pass.test`.

[3] We will examine how to do this later in this chapter.

[4] This is because the interaction between `Attend` and `Pass.test` is statistically significant (p-value $= 0.04297$). See ANOVA table on Slide 15.

Exam vs test success and attendance

Interpreting the output...

First, we use the emm_options() function to tell emmeans we want to use a colon (":") to separate the factor levels in each treatment combination, i.e.

```
> emm_options(sep = ":")
```

The emmeans() function creates and stores a list of two objects. We store the results in exam_intn.emm so that we can print the contents of each object separately.

```
> exam_intn.emm <- emmeans(Exam.fit, specs = pairwise ~ Attend:Pass.test)
```

Exam vs test success and attendance

Interpreting the output. . .

The first object, named `emmeans`, contains four rows: one per treatment[5] combination of the levels of `Attend` and `Pass.test`.

```
> exam_intn.emm$emmeans
 Attend Pass.test emmean   SE  df lower.CL upper.CL
 No     nopass      35.1 3.22 142     28.8     41.5
 Yes    nopass      38.3 3.22 142     32.0     44.7
 No     pass        48.2 2.95 142     42.3     54.0
 Yes    pass        62.9 1.66 142     59.7     66.2

Confidence level used: 0.95
```

Each row of the table contains information corresponding to one of the treatment combinations: the estimated mean (`emmean`), the standard error of the mean (`SE`), the number of degrees of freedom (`df`) used to estimate the standard error, and the lower (`lower.CL`) and upper (`upper.CL`) confidence limits of the mean.

[5] The word *treatment* is often used on its own to refer to a combination of the levels of two or more factor variables.

Exam vs test success and attendance

Interpreting the output...

The second object contains information corresponding to the simple `contrasts`[6] of the treatment means, i.e. combinations of pairwise differences between the four means.

```
> exam_intn.emm$contrasts[-c(3:4)]
 contrast                  estimate   SE  df t.ratio p.value
 No:nopass - Yes:nopass       -3.19 4.56 142 -0.700  0.4850
 No:nopass - No:pass         -13.02 4.37 142 -2.978  0.0034
 Yes:nopass - Yes:pass       -24.62 3.63 142 -6.789  <.0001
 No:pass - Yes:pass          -14.79 3.39 142 -4.364  <.0001
```

We see from the second row of the above `contrasts` table that the *effect* (difference between the means) of the two levels of `Pass.test` conditional on the level of `Attend` = `No` is `-13.02`. So, among those students who did not regularly attend lectures, those who passed the test scored an average of 13 points higher in the exam than those who failed.

[6] Only those contrasts which involve conditioning on the level of one the two factors have been retained.

Exam vs test success and attendance

Interpreting the output...

We require confidence intervals for the pairwise differences between means to write our Executive Summary. These are easily obtained by supplying the `contrasts` object to the `confint()` function, i.e.

```
> confint(exam_intn.emm$contrasts)
 contrast                 estimate   SE  df lower.CL upper.CL
 No:nopass - Yes:nopass     -3.19 4.56 142    -15.0     8.66
 No:nopass - No:pass       -13.02 4.37 142    -24.4    -1.65
 No:nopass - Yes:pass      -27.81 3.63 142    -37.2   -18.38
 Yes:nopass - No:pass       -9.83 4.37 142    -21.2     1.54
 Yes:nopass - Yes:pass     -24.62 3.63 142    -34.0   -15.19
 No:pass - Yes:pass        -14.79 3.39 142    -23.6    -5.98

Confidence level used: 0.95
Conf-level adjustment: tukey method for comparing a family of 4 estimates
```

Notice that we have not excluded rows 3 and 4 of the `contrasts` table when generating the above confidence intervals. This is because doing so would not yield the required Tukey-adjusted p-values.

Exam vs test success and attendance

Interpreting the output. . .

We interpret this output as follows (noting that the effect is always conditional on the level of the other factor):

- We estimate that for students who attend regularly, those who pass the test can expect to get 15 to 34 more marks in the exam than those who do not pass the test.
- For students who do not attend regularly, those who pass the test can expect to get 2 to 24 more marks in the exam than those who do not pass the test.
- For students who pass the test, those who regularly attend can expect to get between 6 and 24 more marks in the exam than those who do not attend regularly.
- And, for those who do not pass the test, those who regularly attend can expect to get between 9 marks less and 15 more marks than those who do attend regularly.

Exam vs test success and attendance

NOTE: The above statements are made about the population of students from which the STATS20x data are assumed to be a random sample.

— this means YOU!

Alternative parametrisations of the two-factor linear model

The reference cell model

Recall the reference cell model we used to represent `Exam` score:

$$\texttt{Exam} = \beta_0 + \beta_1 \times \texttt{Attend}_{\texttt{Yes}} + \beta_2 \times \texttt{Pass.test}_{\texttt{pass}} + \beta_3 \times \texttt{Attend}_{\texttt{Yes}} \times \texttt{Pass.test}_{\texttt{pass}} + \varepsilon,$$

where $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$.

The parameter β_0 denotes the overall true baseline mean exam score. Notice that neither the `no` level of `Attend` nor the `nopass` level of `Pass.test` appear as subscripts in the above model. This tells us that these are the baseline levels, i.e. β_0 denotes the mean over `Exam` scores from students who neither regularly attended lectures nor passed the test.

So, what do the parameters $\beta_1, \beta_2,$ and β_3 represent? To help us answer this question we consider the means model[7] formulation for `Exam` score.

[7] We first encountered the means model for the single factor male fruitflies study in Chapter 11.

Alternative parametrisations of the two-factor linear model

The means model

The means model parametrisation for exam score is

$$\texttt{Exam}_{ijk} = \mu_{ij} + \varepsilon_{ijk},$$

where μ_{ij} denotes the true mean exam score of 20x students who are in the ith level of `Attend` and jth level of `Pass.test` ($i = \texttt{no}$ or `yes`; $j = \texttt{nopass}$ or `pass`). The error term $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$ denotes the deviation of the kth student's exam score from the mean exam score, μ_{ij}.

But, how can we use μ_{ij} to assess whether one or both of `Attend` and `Pass.test` have an effect on `Exam` score?

Alternative parametrisations of the two-factor linear model

Relating the means and reference cell models

We decompose each mean response, μ_{ij}, into four terms:

1. μ_{11}, the baseline or reference-level mean response;
2. $\mu_{i1} - \mu_{11}$, the *main effect*[8] of the ith level of the first factor, where i does not equal the baseline level;
3. $\mu_{1j} - \mu_{11}$, the *main effect* of the jth level of the second factor, where j does not equal the baseline level;
4. Interaction, the part of μ_{ij} that is left over after eliminating the contributing components defined by terms 1–3 above, i.e.

$$\begin{aligned}\text{Interaction} &= \mu_{ij} - \mu_{11} - (\mu_{i1} - \mu_{11}) - (\mu_{1j} - \mu_{11}) \\ &= \mu_{ij} - \mu_{i1} - \mu_{1j} + \mu_{11}\end{aligned}$$

[8] A main effect is defined as the difference between the mean response when all factors, except the one of interest, are at the baseline level and the reference-level mean.

Alternative parametrisations of the two-factor linear model

Relating the means and reference cell models

We now have the tools to re-express the mean `Exam` score in terms of the main effects of the ith level of `Attend` and the jth level of `Pass.test`, and their interaction, i.e.

$$\mu_{ij} = \mu_{11} + (\mu_{1j} - \mu_{11}) + (\mu_{i1} - \mu_{11}) + (\mu_{ij} - \mu_{i1} - \mu_{1j} + \mu_{11}).$$

The following two-way table[9] illustrates how each term in the above decomposition relates to each combination of the levels of `Attend` and `Pass.test`:

Attend	Pass.test	
	`nopass`	`pass`
`no`	μ_{11}	$\mu_{i1} - \mu_{11}$
`yes`	$\mu_{1j} - \mu_{11}$	$\mu_{ij} - \mu_{11} - (\mu_{i1} - \mu_{11}) - (\mu_{1j} - \mu_{11})$

[9] More generally, differences in the first row of a two-way reference model decomposition table correspond to the main effects of the column factor. The differences in the first column correspond to the row factor main effects. The terms in each of the of the remaining cells, except the reference cell, correspond to interaction effects.

Alternative parametrisations of the linear model

Relating the means and reference cell models

Factors		Parametrisation			
Attend	Pass.test	Means	Estimate[10]	Reference cell	Estimate[11]
no	nopass	μ_{11}	35.1	$\beta_0 = \mu_{11}$	35.1
yes	nopass	μ_{21}	38.3	$\beta_1 = \mu_{21} - \mu_{11}$	3.2
no	pass	μ_{12}	48.2	$\beta_2 = \mu_{12} - \mu_{11}$	13.1
yes	pass	μ_{22}	62.9	$\beta_3 = \mu_{22} - \mu_{21} - \mu_{12} + \mu_{11}$	11.5

From the above table we see that:

- β_1 represents the effect of `Attend = yes` at the reference level of `Pass.test = nopass`
- β_2 represents the effect of `Pass.test = pass` at the reference level of `Attend = no`
- β_3 represents the `Attend` *x* `Pass.test` interaction effect when `Attend = yes` and `Pass.test = pass`

[10] See estimates of `Attend`×`Pass.test` treatment means on slide 19.

[11] See regression coefficients table on slide 16.

Alternative parametrisations of the linear model

The reference cell model

The values in the `Estimate` column of the regression summary table[12] result in the following equation for predicted longevity:

$$\widehat{\texttt{Exam}} = 35.14 + 3.19 \times \texttt{Attend}_{\texttt{Yes}} + 13.02 \times \texttt{Pass.test}_{\texttt{pass}}$$
$$+ 11.60 \times \texttt{Attend}_{\texttt{Yes}} \times \texttt{Pass.test}_{\texttt{pass}}$$

[12] See slide 16; Coefficients rounded to 2 decimal places.

Alternative parametrisations of the two-factor linear model

Relating the means and effects models

We saw in Chapter 11 that the effects model offers an alternative to the reference model parametrisation.[13] To relate the means and effects models we use an alternative decomposition of each mean response, μ_{ij}, into:

1. $\mu = \mu_{\cdot\cdot}$, the reference overall mean response;
2. $\alpha_i = \mu_{i\cdot} - \mu$, the *main effect* of the ith level of the first factor[14];
3. $\pi_j = \mu_{\cdot j} - \mu$, the *main effect* of the jth level of the second factor[14];
4. $(\alpha\tau)_{ij}$, the interaction effect reflecting the component of μ_{ij} left over after eliminating the components corresponding to the terms defined in 1–3 above, i.e. $(\alpha\tau)_{ij} = \mu_{ij} - \mu - \alpha_i - \pi_j$.

---

[13] The effects model is the default parametrisation used in the analysis of data from designed experiments and, therefore, in STATS 240 (Design and Structured Data).

[14] The distance between the mean response of the first (second) factor at the ith (jth) level and the overall mean.

Alternative parametrisations of the two-factor linear model

The effects model

It directly follows from point 4 above that we can re-express the mean `Exam` score, μ_{ij}, in terms of α_i, the effect of the ith level of `Attend`, π_j, the effect of the jth level of `Pass.test`, and their interaction, $(\alpha\pi)_{ij}$, i.e.

$$\mu_{ij} = \mu + \alpha_i + \pi_j + (\alpha\pi)_{ij}.$$

This means that the effects model formula for the two-way ANOVA is

$$\texttt{Exam}_{ijk} = \mu + \alpha_i + \pi_j + (\alpha\pi)_{ij} + \epsilon_{ijk},$$

where $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$.

Alternative parametrisations of the linear model

Relating the means and effects models

The following two-way layout illustrates how the above decomposition of μ_{ij} relates to each combination of the levels of `Attend` and `Pass.test`:

Attend	Pass.test		Row mean	
	`nopass`	`pass`		
`no`	$\mu_{11} = 35.1$	$\mu_{12} = 48.2$	$\mu_{1\cdot}$	$\alpha_1 = \mu_{1\cdot} - \mu$
`yes`	$\mu_{21} = 38.3$	$\mu_{22} = 62.9$	$\mu_{2\cdot}$	$\alpha_2 = \mu_{2\cdot} - \mu$
Column mean	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	μ	
Column effect	$\pi_1 = \mu_{\cdot 1} - \mu$	$\pi_2 = \mu_{\cdot 2} - \mu$		

Alternative parametrisations of the linear model

Relating the means and effects models

Factors		Parametrisation			
Attend	Pass.test	Means	Estimate[15]	Reference cell	Estimate[16]
no	nopass	μ_{11}	35.1	$\beta_0 = \mu_{11}$	35.1
yes	nopass	μ_{21}	38.3	$\beta_1 = \mu_{21} - \mu_{11}$	3.2
no	pass	μ_{12}	48.2	$\beta_2 = \mu_{12} - \mu_{11}$	13.1
yes	pass	μ_{22}	62.9	$\beta_3 = \mu_{22} - \mu_{21} - \mu_{12} + \mu_{11}$	11.5

From the above table we see that:

- β_1 represents the effect of `Attend = yes` at the reference level of `Pass.test = nopass`
- β_2 represents the effect of `Pass.test = pass` at the reference level of `Attend = no`
- β_3 represents the `Attend` *x* `Pass.test` interaction effect when `Attend = yes` and `Pass.test = pass`

[15] See estimates of `Attend`×`Pass.test` treatment means on slide 19.

[16] See regression coefficients table on slide 16.