MATH411 | Fall 2018 | Chapter 6: Models Involving Factors

Dr. Yongtao Cao

TBD, 2018

Contents

5.1 Introduction **1**

5.2 One-Way ANOVA **5**

5.3 Two-way ANOVA **29**

5.4 Analysis of Covariance: Comparing Regression Lines **49**

5.1 Introduction

The variables we considered so far were all **numerical**, i.e. temperature, distance, pressure, et cetera. While the response must be numerical, it is perfectly valid to use **categorical predictors**, such as e.g. **gender** (`male` or `female`), **status** variables (`employed` or `unemployed`), **shifts** (`day`, `evening`, `night`). In general, these categorical variables have no natural scale of measurement. Thus, we must assign a set of levels to a categorical variable to account for the effect that the variable may have on the response. This is done through the use of indicator variables. In the regression context, they are better known as **dummy variables**. In the following sections, we will study the use of categorical predictors.

Beer and Mosquitoes Revisited

Does consuming beer attract mosquitoes? A study done in Burkino Faso, Africa, about the spread of malaria investigated the connection between beer consumption and mosquito attraction. In the experiment, 25 volunteers consumed a liter of beer while 18 volunteers consumed a liter of water. **The volunteers were assigned to the two groups randomly**. The attractiveness to mosquitoes of each volunteer was tested twice: before the beer or water and after. Mosquitoes were released and caught in traps as they approached the volunteers. For the beer group, the total number of mosquitoes caught in the traps before consumption was 434 and the total was 590 after consumption. For the water group, the total was 337 before and 346 after. We have seen, in Chapter 3, that this problem can be solved with a 2-sample T-test technique, i.e., the hypotheses are:

$$H_0 : \mu_b = \mu_w \qquad VS \qquad H_a : \mu_b > \mu_w$$

The test statistic, assuming equal variance, is given by

$$t = \frac{\bar{x}_b - \bar{x}_w}{s_p\sqrt{1/n_b + 1/n_w}} \overset{H_0}{\sim} t_{n_b+n_w-2}$$

where $s_p = \sqrt{\frac{(n_b-1)\cdot s_b^2+(n_w-1)\cdot s_w^2}{n_b+n_w-2}}$

What does this have to do with regression analysis? More than you think. We can achieve the very same quantitative results by fitting a regression of $y \sim x$. Because regression is a technique for numerical variables, we need to replace the categorical predictor x by an indicator variable that takes values 0 and 1 to identify the drink types – this is a so-called **dummy variable**.

$$x = \begin{cases} 1, & \text{beer} \\ 0, & \text{water} \end{cases}$$

The choice of 0 and 1 to identify the levels of this categorical predictor is arbitrary. In fact, any two distinct values for x would be satisfactory, although 0 and 1 are the normal choice. Then, if we consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

this becomes

$$y_i = \beta_0 + \epsilon_i$$

for observations i with drinking water and hence $x_i = 0$.

Then, for observations j with drinking beer, $x_j = 1$ and the regression equates to

$$y_j = \beta_0 + \beta_1 + \epsilon_i$$

Consequently, β_0 **is the expected number of mosquitoes after drinking water**, and $\beta_0 + \beta_1$ **is the one for drinking beer**. Or we can also say that β_1 **is the difference in the two mosquito attractiveness expectations**.

With R, fitting regression models with categorical predictors is straightforward. **We do not even need to take care of the generating of the dummy variable, but can just provide a**

factor variable, i.e. `class(mosquito$group) = "factor"`.

We observe that the regression coefficients are identical to the results from the `t-test` procedure above, where arithmetic means were drawn. Furthermore, the test for the null hypothesis $\beta_1 = 0$ addresses exactly the same question as the `t-test` for non-paired data does. However, not only the question is identical, but also the answer (and the methodology behind). **The p-values with both approaches are one and the same**. Hence, if we can do regression, we could in fact retire the non-paired `t-test` altogether.

5.2 One-Way ANOVA

We now generalize our discussion above for **comparing two means to more than two means**. In what follows, you should know that categorical variables are also called **factors**. The different values of a factor are called **levels** (treatments, groups).

5.2.1 Completely Randomized Design: Formal Setup

- Compare ____ treatments
- Available resources: ____ experimental units

- Need to **assign** the N experimental units to k different **treatments** (groups) having ____ observations each, i.e., we have ____________. This is a so-called **completely randomized design** (CRD).

- Use randomization:

  – Choose n_1 units **at random** and assign to treatment 1,

  – n_2 units **at random** and assign to treatment 2,

  – …

- If all the treatment groups have the same number of experimental units we call the design **balanced**.

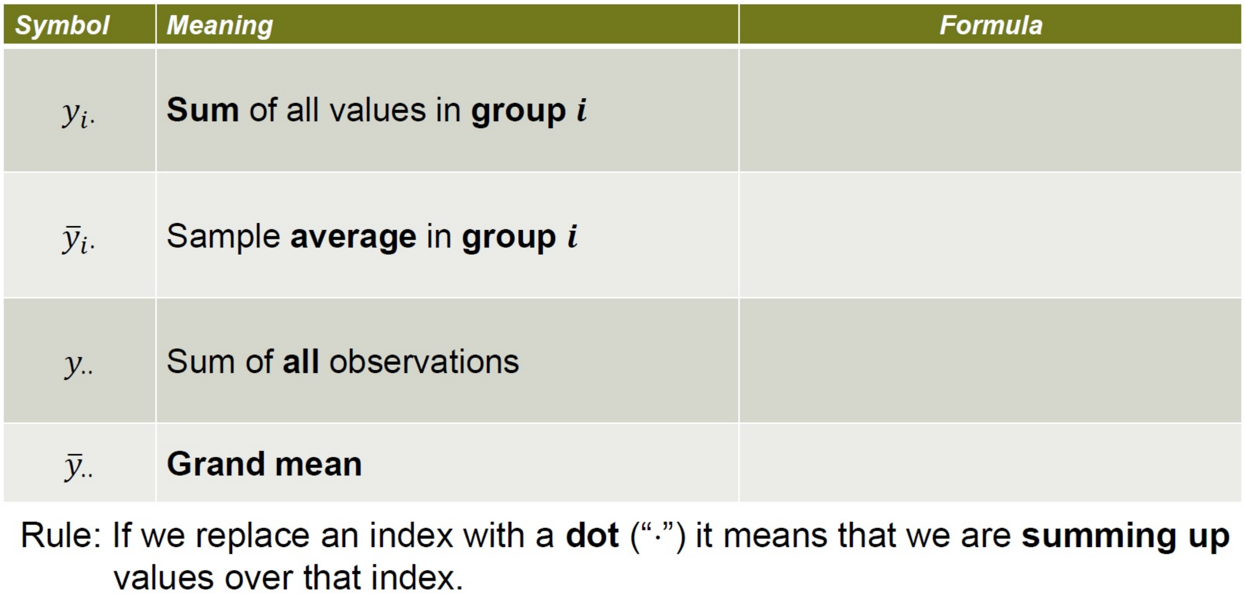
Data Representations

The data from CRD can generally be represented as below:

Table 1: The General Sample Size Case

Treatments (levels)	1	2	…	k
	y_{11}	y_{21}	$\cdots$	y_{k1}
	y_{12}	y_{22}	$\cdots$	y_{k2}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Treatments (levels)	1	2	$\dots$	k
	y_{1n_1}	y_{2n_2}	$\cdots$	y_{kn_k}
——————————-	————	————	————	————
Treatment Totals	$y_{1\cdot}$	$y_{2\cdot}$	$\dots$	$y_{k\cdot}$
Treatment Means	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	$\dots$	$\bar{y}_{k\cdot}$



Symbol	Meaning	Formula
$y_{i\cdot}$	**Sum** of all values in **group *i***	
$\bar{y}_{i\cdot}$	Sample **average** in **group *i***	
$y_{\cdot\cdot}$	Sum of **all** observations	
$\bar{y}_{\cdot\cdot}$	**Grand mean**	

Rule: If we replace an index with a **dot** ("·") it means that we are **summing up** values over that index.

Figure 1: Onw-way ANOVA Notations

Assume the k levels of the factor are **fixed** by the experimenter. This implies the levels are specifically chosen by the experimenter. The goal is to **determine if there exist any differences in the set of k treatment means (or effects)**. So, we want to check the (global) null hypothesis that $\mu_1, \mu_2, \dots, \mu_k$, are **all equal** against the alternative that they are not all equal,

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_a : \mu_i \neq \mu_j \quad \text{for some } \; i \neq j$$

If H_0 is rejected, **then we wish to know which means differ, and by how much**.

5.2.2 Models

We start by formulating a parametric model for our data. Let y_{ij} be the jth observation in the ith treatment group, where $i = 1, \ldots, k$; $j = 1, \ldots, n_i$.

Cell Means Model (Every treatment has its own mean)

In the cell means model **we allow each treatment group to have its own expected value** and **we assume that the observations are independent and fluctuate around this value according to a normal distribution**, i.e.,

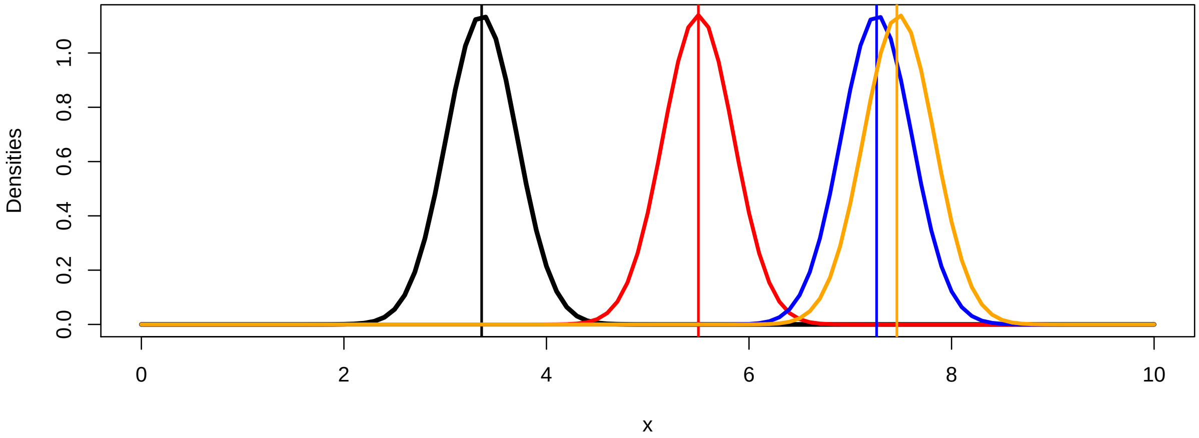
- $\mu_i =$

- $\sigma^2 =$

As for the (standard) two-sample t-test, the variance is **equal** for all groups. We can re-write Model (6.1) as

with (random) errors $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. We simply separated the normal distribution around μ_i into a deterministic part μ_i and a stochastic part ϵ_{ij} fluctuating around zero. We also say that y is the **response** and the grouping information is a **categorical predictor**. Hence, this is nothing else than a regression model with a categorical predictor and normally distributed errors (for those that are already familiar with linear regression models).

- Graphically,



Treatment Effects Model

We can also write Model (6.2) as

- E.g., think of μ as a “**global mean**” and α_i as the corresponding **deviation from the global mean**.
- α_i is also called the ith **treatment effect**.

If we carefully inspect the parameters of models (6.2) and (6.3) we observe the following: in (6.2) we have the parameters [redacted] and [redacted] while in (6.3) we have [redacted] and [redacted]. We (silently) introduced one additional parameter. In fact, model (6.4) is **not identifiable** anymore because we have $k+1$ parameters to model k mean values μ_i. Or in other words: we can “shift around” effects between μ and the α_i’s without changing the resulting values of μ_i, e.g., we can adjust $\mu + 10$ and $\alpha_i + 10$ leading to the same μ_i’s. Hence, we need a side-constraint on the α_i’s that “removes” that additional parameter. Typical choices for such a constraint are

Table 2: Choices for Constraints

Name	Side-constraint	Interpretation of μ	`R`
weighted sum-to-zero			
sum-to-zero			
reference group			

where we have also listed the interpretation of the parameter μ and the **R** naming convention. For all of the above choices it holds that μ determines some sort of "global level" of the data and α_i contains information about differences between the groups mean μ_i from that "global level".

Only $k-1$ elements of the treatment effects are allowed to vary freely. In other words: If we know $k-1$ of the α_i values, we automatically know the remaining α_i. We also say that the treatment effect has $k-1$ **degrees of freedom (df)**.

In **R** the side-constraint is set using the option `contrasts` (see examples). The default value is `contr.treatment` which is the side-constraint "reference group" from above.

5.2.3 Parameter Estimation

We estimate the parameters using the least squares criterion which ensures that the model fits the data well in the sense that the squared deviation from the observed data y_{ij} to the model values $\mu + \alpha_i$ are minimized, i.e.

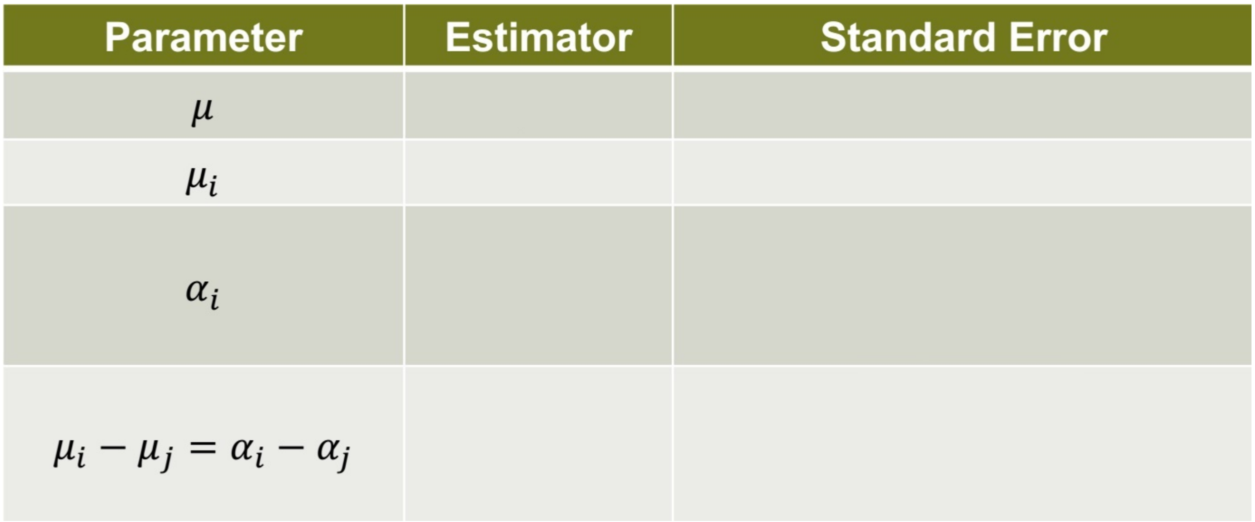
In the parametrization of the cell means model this means

Using the notations in Figure 1, we can independently estimate the values of the different groups, we get [redacted], hence we have

Depending on the side-constraint that we use we get different results for $\widehat{\alpha}_i$.

The estimate of the error variance $\widehat{\sigma}^2$ is also called **mean squared error** MSE. It is given by

where SSE is the residual or (error) sum of squares



Parameter	Estimator	Standard Error
μ		
μ_i		
α_i		
$\mu_i - \mu_j = \alpha_i - \alpha_j$		

Figure 2: Notations

Matrix Presentation

- The data vector will be labelled as $\mathbf{Y}$. $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_k \end{pmatrix}$, where $\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix}$
- The **design** (model) matrix will be denoted by $\mathbf{X}$.
- The parameter vector will be denoted as $\boldsymbol{\beta}$.
- and the error vector will be denoted as $\boldsymbol{\epsilon}$.

- So, we have the linear model:

Example: Meat Storage Study (Kuehl, 2000, Example 2.1)

Read the handout and answer the following questions:

- Response:
- Factor:
- Treatments (levels):
- Experimental units:
- Group size:
- Define the parameters and set up the null and alternative hypothesis:

- See R code for exploring this data.

5.2.4 Tests

With a two-sample t-test we could test whether two samples shared the same mean. We will now extend this to the $k > 2$ situation.

Saying that **all** groups share the same mean is equivalent to model the data as

This is the so-called **single mean model**. It is actually a special case of the cell means model where ______________ which is equivalent to ______________.

Hence, our question boils down to comparing two models, the single mean and the cell means model. More formally, we have the (global) null hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

versus the alternative

$$H_a : \mu_i \neq \mu_j \quad \text{for some} \;\; i \neq j$$

We call it a global null hypothesis because it affects all parameters at the same time.

Equivalently, we can test for no significant treatment effects

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$$

$$H_a : \alpha_i \neq 0 \quad \text{for some } \; i$$

In terms of the linear models, we have

$$H_0 : \mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

$$H_a : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- See the R code for comparing the two models.

We will construct a statistical test by **decomposing the variation of the data**. Total variation of the response around the overall mean can be decomposed into variation "**between groups**" and variation "**within groups**". More formally, we have

where

- $SST =$

- $SStrt =$
- $SSE =$

Hence, we have $SST = SS_{\text{trt}} + SSE$.

As can be seen from the decomposition, SS_{trt} can also be interpreted as the **reduction in residual sum of squares when comparing the cell means with the single mean model** (this will be a useful interpretation later).

This information can be summarized in a so-called ANOVA table, where ANOVA stands for analysis of variance:

Table 3: ANOVA Table

Source	df	Sum of squares (SS)	Mean Squares (MS)	F-ratio
Treatment				
Error				
Total				

The mean squares are sum of squares that are normalized with the corresponding degrees of freedom. This is a so-called **one-way analysis of variance** (or short: **one-way ANOVA**) because there is **only one factor involved**. Another rule of thumb for getting the degrees of freedom for the error is as follows: Total sum of squares has $N-1$ degrees of freedom (we have N observations that fluctuate around the global mean), $k-1$ degrees of freedom are "used" for the treatment effect (k groups minus one side-constraint). Hence, the remaining part (the error) has $N-1-(g-1)=N-g$ degrees of freedom.

If all the groups share the same (theoretical) mean, we expect the treatment sum of squares to be small: Just due to the random nature of the data we expect small differences between the different (empirical) group means. The idea is now to compare the variation **between** groups with the variation **within** groups. **If the variation between groups is substantially larger than the variation within groups we have evidence against the null hypothesis**.

In fact, it can be shown that under H_0 it holds that

where we denote by $F_{n,m}$ the so-called F-distribution with n and m degrees of freedom. The F-distribution has two degrees of freedom parameters, one from the numerator (here: $k-1$) and one from the denominator (here: $N-k$). The $F_{1,m}$-distribution is a special case that we already know, it is the square of a t_m-distribution (a t-distribution with m degrees of freedom).

As with any other statistical test, we reject the null hypothesis if the observed value of the F-ratio (our test statistics) lies in an "extreme" region of the corresponding F-distribution. This is a one-sided test, which means that we reject H_0 in favor of H_a if F is larger than the 95% quantile (if we use a 5% significance level). We also use the notation $F_{n,m,\alpha}$ for the $(\alpha \times 100)$% quantile.

The F-test is also called an **omnibus test** (Latin for "for all") as it compares all group means simultaneously.

CI for a Treatment Mean

Recall: If we have a single sample of size n coming from a normally distributed population $\mathcal{N}(\mu, \sigma_\epsilon^2)$, then

$$\frac{\bar{y} - \mu}{s/\sqrt{n}} \sim$$

An analogous result is true when considering a treatment mean from a single factor ANOVA. That is, if the normality and homogeneity of variance assumptions for the single factor ANOVA are valid, then

Thus to obtain a $(1-\alpha)\,100\%$ CI for μ_i, calculate:

- See R code.

5.2.5 Contrasts and Multiple Testing

The F-test is rather unspecific. It basically gives us a "Yes/No" answer for the question "is there any treatment effect at all?". It doesn't tell us what specific treatment (or treatment combination) is significant. Quite often we have a more specific question than the aforementioned global null hypothesis. E.g.,

- we might want to compare a set of new treatments vs. a control treatment,
- or we want to do pairwise comparisons between many (or all) treatments.

Such kinds of questions can typically be formulated as a so-called **contrast**. As a simple example, say we want to compare group 2 with group 1 (don't care about the remaining $k-2$ groups for the moment). How do you set up H_0 and H_a?

$$H_0:$$

$$H_a :$$

We can encode this with a vector $\mathbf{c} \in \mathbb{R}^k$

In this example the vector $\mathbf{c}$ is equal to $\mathbf{c} =$ ____________ (with respect to `trt1`, `trt2`, $\ldots$, `trtk`). Hence, a contrast is an encoding of our own, specific research question.

Contrasts

Typically, we have the side-constraint

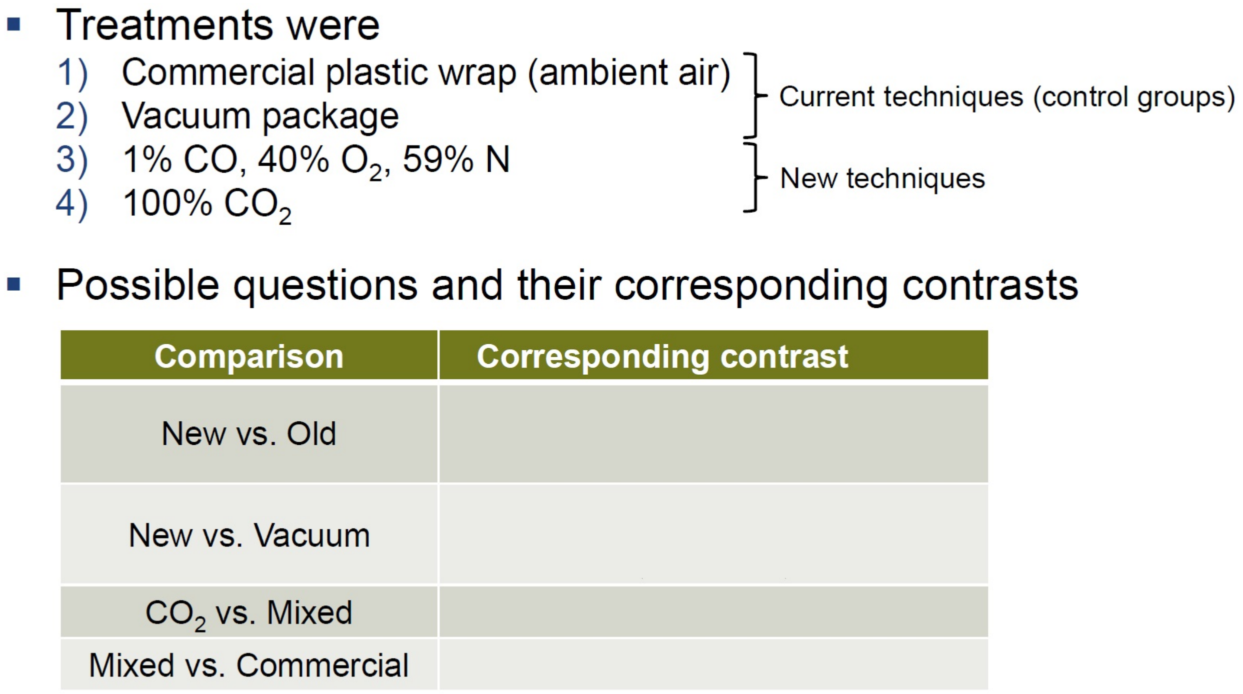
$$\sum_{i=1}^{k} c_i = 0$$

which ensures that the contrast is about differences between treatments and not about the overall level of our response.

We estimate a contrasts true (but unknown) value $\sum_{i=1}^{k} c_i \mu_i$ (a **linear combination of model parameters**!) with

$$\sum_{i=1}^{k} c_i \widehat{\mu}_i$$

In addition, we could derive its accuracy (standard error), construct confidence intervals and do tests. We omit the theoretical details



- Treatments were
  1) Commercial plastic wrap (ambient air)
  2) Vacuum package
  3) 1% CO, 40% O_2, 59% N
  4) 100% CO_2

Current techniques (control groups)

New techniques

- Possible questions and their corresponding contrasts

Comparison	Corresponding contrast
New vs. Old	
New vs. Vacuum	
CO_2 vs. Mixed	
Mixed vs. Commercial	

Figure 3: Contrast Example Using Meat Storage Data

and continue with our example.

Meat Storage Data Revisited

In **R** we use the function `glht` (general linear hypotheses) of the add-on package `multcomp` (Hothorn, Bretz, and Westfall 2016).

- See R code.

Some Technical Details

Every contrast has an associated **sum of squares**

$$SS_c = \frac{\left(\sum_{i=1}^{k} c_i \overline{y}_{i\cdot}\right)^2}{\sum_{i=1}^{k} \frac{c_i^2}{n_i}}$$

having [redacted] degree of freedom, hence [redacted]. This looks unintuitive at first sight but it is nothing else than the square of the t-statistic of the corresponding null hypothesis for the special model parameter $\sum_{i=1}^{k} c_i \mu_i$ (without the MSE factor). Under H_0 it holds that

Because $F_{1,m} = t_m^2$ (the square of a t_m-distribution with m degrees of freedom), this is nothing else than the "squared version" of the t-test. You can think of MS_c as the "part" of MS_{trt} in "direction" of $\mathbf{c}$.

Two contrasts $\mathbf{c}$ and $\mathbf{c}^*$ are called **orthogonal** if

If two contrasts $\mathbf{c}$ and $\mathbf{c}^*$ are orthogonal, the corresponding estimates are (statistically) independent. This means that if we know something about one of the contrasts, this does not help us in making a statement about the other one.

If we have k treatments, we can find $k - 1$ different orthogonal contrasts (one dimension is already used by the global mean (1, 1, $\ldots$, 1). A set of orthogonal contrasts **partitions** the treatment sum of squares meaning that if $c^{(1)}, \ldots, c^{(k-1)}$ are orthogonal contrasts it holds that

Intuition: “We get all information about the treatment by asking the right $k - 1$ questions”.

Multiple Testing

The problem with all statistical tests is the fact that the (overall) error rate increases with increasing number of tests. Assume that we perform m (independent) tests $H_{0,j}$ $j = 1, \ldots, m$, using an individual significance level of α. If all $H_{0,j}$ are true, the probability to make **at least one** false rejection is given by

Even for α small this is close to 1 if m is large. This means: **if we perform many tests, we expect to find some significant results, even if all null hypotheses are true**. Hence, we need a way to control the overall error rate. Let us first list the potential outcomes of a total of m tests, whereof m_0 null hypotheses are true.

Outcome	H_0 true	H_0 false	Total
Significant	V	S	R
Non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

For example, V is the number of wrongly rejected null hypotheses.

The **family-wise error rate** is defined as the probability of rejecting at least one of the true H_0's:

The family-wise error rate is very strict in the sense that we are not considering the actual number of **wrong decisions**, we are just interested whether there is **at least one**. This means the situation where we make $V = 1$ error is equally "bad" as the situation where we make $V = 20$ errors.

We say that a procedure controls the family-wise error rate in the **strong sense** at level α if FWER $\leq \alpha$ for **any** configuration of true and non-true null hypotheses. A typical choice would be $\alpha = 0.05$.

Another error rate is the **false discovery rate** (FDR) which is the expected fraction of false discoveries,

$$\text{FDR} = E\left[\frac{V}{R}\right].$$

Controlling FDR at level 0.2 means that in our list of "significant findings" we expect only 20% that are not "true findings" (so called false positives). If we can live with a certain amount of false positives the relevant quantity to control is the false discovery rate.

If a procedure controls **FWER** at level α, **FDR** is automatically controlled at level α too. On the other side, a procedure that controls **FDR** at level α might have a much larger error rate regarding **FWER**. Hence, **FWER** is a much more strict (conservative) criterion (meaning: leading to fewer rejections).

We can also control the error rates for confidence intervals. We call a set of confidence intervals **simultaneous confidence intervals** at level $(1-\alpha)$ if the probability that all intervals cover the corresponding true parameter value is $(1-\alpha)$. This means: We can look at all confidence intervals at the same time and get the correct "big picture" with probability $(1-\alpha)$.

In the following, we focus on the family-wise error rate (FWER) and simultaneous confidence intervals.

We typically start with "individual" p-values that we modify (or adjust) such that the appropriate **overall** error rate (like FWER) is being controlled. Interpretation of an individual p-value is as you learned it in your introductory course ("the probability to observe an event as extreme as…"). The modified p-values should be interpreted as the smallest overall error rate such that we can reject the corresponding null hypothesis.

Bonferroni

The Bonferroni correction is a generic but very conservative approach. The idea is to use a more restrictive (individual) significance level of ________. It controls the family-wise error rate in the strong sense. Equivalently, we can also multiply the "original" p-values by m and keep using the original α. Especially for large m the Bonferroni correction is very conservative. The confidence intervals based on the adjusted significance level are simultaneous.

Tukey Honest Significant Difference (HSD)

A special case for a multiple testing problem is the comparison between all possible pairs of treatments. There are a total of ________ pairs that we can inspect. We could perform all pairwise t-tests with the function `pairwise.t.test` (it uses a pooled standard deviation estimate from all groups).

There is a better (more powerful) alternative which is called **Tukey Honest Significant Difference**. Think of a procedure that is custom tailored for this situation. It gives us both p-values and confidence intervals. In R this is implemented in the function `TukeyHSD` or in the package `multcomp`.

Multiple Comparison with a Control (MCC)

In the same spirit, if we want to **compare all treatment groups with a control group**, we have a so called **multiple comparisons with a control problem**. The corresponding procedure is called **Dunnett** procedure which is implemented in the add-on package `multcomp`. By default, the first level of the factor is taken as the control group.

FAQ

1. Should I only do individual tests if the global F-test is significant?

No, although this still suggested in many textbooks. The above mentioned procedures have a built-in correction regarding multiple testing and do **not** rely on a significant F-test. Conditioning on the F-test leads to a very conservative approach regarding type I error rate. In addition, coverage rate of e.g. Tukey HSD confidence intervals can be very low.

2. Could it be the case that the F-test is significant but Tukey HSD yields only insignificant pairwise tests?

Yes, because the F-test can combine groups, Tukey HSD cannot.

3. Could it be the case that Tukey HSD yields a significant difference but the global F-test is not significant?

Yes, because Tukey has larger power for some alternatives because it tries to answer a more precise question.

5.3 Two-way ANOVA

In the completely randomized designs that we have seen so far, the k different treatments had no special "structure". In practice, **treatments are often combinations of the levels of two or more factors**. Think for example of a plant experiment using combinations of `light exposure` and `fertilizer`. We call this a ______________________________. If we see **all possible combinations** of the levels of two (or more) factors, we call them **crossed**. An illustration of two crossed factors can be found in Table below.

Table : Example of two Crossed Factors. With "x" we mean that we have data for the specific combination of exposure level (`low` / `medium` / `high`) and fertilizer brand (`A` / `B`).

exposure / fertilizer	A	B
low	x	x
medium	x	x
high	x	x

5.3.1 Two Factor Factorial Designs: Set Up

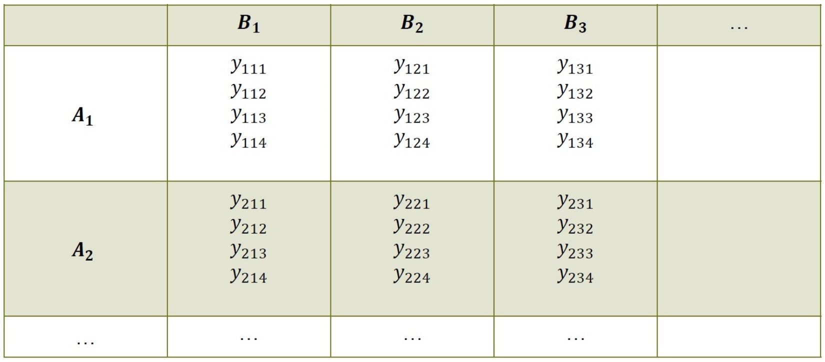
- Factor A with a levels
- Factor B with b levels
- n replicates for each of the $a \times b$ treatment combinations.
- The design size is $N =$.

Typically, we have research questions about **both** factors and their possible interactions (interplay).

- **Main effect**: The main effect of a factor is defined to be the **average change in the response associated with a change in the level of the factor**.
- **Interaction**: If the average change in response across the levels of one factor are not the same at all levels of the other factor, then we say there is an **interaction** between the factors.

Notations

We denote by y_{ijk} the kth response of the treatment formed by the ith level of factor A and the jth level of factor B.



	B_1	B_2	B_3	…
A_1	y_{111} y_{112} y_{113} y_{114}	y_{121} y_{122} y_{123} y_{124}	y_{131} y_{132} y_{133} y_{134}	
A_2	y_{211} y_{212} y_{213} y_{214}	y_{221} y_{222} y_{223} y_{224}	y_{231} y_{232} y_{233} y_{234}	
…	…	…	…	

Figure 4: Generic Data Table for Two Factor Design

TYPE	TOTALS	MEANS	(if $n_{ij} = n$)
Cell(i, j)			
i^{th} level of A			
j^{th} level of B			
Overall			

where n_{ij} is the number of observations in cell (i, j).

Figure 5: Calculations for Two Factor Design

Example: (A 2×2 experiment):

A virologist is interested in studying the effects of $a = 2$ different culture media (M) and $b = 2$ different times (T) on the growth of a particular virus. She performs $n = 6$ (balanced design) replicates for each of the 4 $M * T$ combinations. The $N = 24$ measurements were taken in a completely randomized order. The results:

THE DATA

		M	
		Medium 1	Medium 2
	12 hours	21 23 20 22 28 26	25 24 29 26 25 27
T			
	18 hours	37 38 35 39 38 36	31 29 30 34 33 35

Figure 6: A 2-by-2 Experiment Data

TOTALS

	$T = 1$	$T = 2$	
$T = 12$	$y_{11\cdot} =$	$y_{12\cdot} =$	$y_{1\cdot\cdot} =$
$T = 18$	$y_{21\cdot} =$	$y_{22\cdot} =$	$y_{2\cdot\cdot} =$
	$y_{\cdot 1\cdot} =$	$y_{\cdot 2\cdot} =$	$y_{\cdot\cdot\cdot} =$

Figure 7: Example Data – Totals

MEANS

	$M = 1$	$M = 2$	
$T = 12$	$\overline{y}_{11\cdot} =$	$\overline{y}_{12\cdot} =$	$\overline{y}_{1\cdot\cdot} =$
$T = 18$	$\overline{y}_{21\cdot} =$	$\overline{y}_{22\cdot} =$	$\overline{y}_{2\cdot\cdot} =$
	$\overline{y}_{\cdot 1\cdot} =$	$\overline{y}_{\cdot 2\cdot} =$	$\overline{y}_{\cdot\cdot\cdot} =$

Figure 8: Example Data – Means

The effect of changing T from 12 to 18 hours on the response depends on the level of M

- For medium 1, the T effect =
- For medium 2, the T effect =

The effect on the response of changing M from medium 1 to 2 depends on the level of T

- For T = 12 hours, the M effect =

- For T = 18 hours, the M effect =

If either of these pairs of effects are significantly different then we say there exists a **significant interaction** between factors M and T. For the 2×2 example:

- If 13.83 is significantly different than 6 for the M effects, then we have a significant $M * T$ interaction.
- If 2.6 is significantly different than -5.16 for the T effects, then we have a significant $M * T$ interaction.

There are two ways of defining an interaction between two factors A and B:

- If the change in response between the levels of factor A is not the same at all levels of factor B, then an **interaction** exists between factors A and B.
- The lack of additivity of factors A and B, or the nonparallelism of the mean profiles of A and B, is called the **interaction** of A and B.
- When there is no interaction between A and B, we say the effects are **additive**.

An **interaction plot** or **treatment means plot** is a graphical tool for checking for potential interactions between two factors. To make an interaction plot,

1. Calculate the cell means for all $a \times b$ combinations of the levels of A and B.
2. Plot the cell means against the levels of factor A.
3. Connect and label means for the same levels of factor B.

- The roles of A and B can be reversed to make a second interaction plot.

Interpretation of the interaction plot:

- **Parallel lines usually indicate no significant interaction**.
- Severe lack of parallelism usually indicates a significant interaction.
- Moderate lack of parallelism suggests a possible significant interaction may exist.

Statistical significance of an interaction effect **depends on the magnitude of the MSE**: for small values of the MSE, even small interaction effects (less nonparallelism) may be significant.

When an $A*B$ interaction is large, the corresponding main effects A and B may have little practical meaning. Knowledge of the $A*B$ interaction is often more useful than knowledge of the main effect.

It is possible to have a significant interaction between two factors, while the main effects for both factors are not significant. This would happen when the interaction plot shows interactions in different directions that balance out over one or both factors (such as an X pattern). This type of interaction, however, is uncommon.

5.3.2 Two-Way ANOVA Model

The **two-way ANOVA model with interaction** is:

where

- α_i is
- β_j is
- $(\alpha\beta)_{ij}$ is
- ϵ_{ijk} are

Typically, we use **sum-to-zero side constraints**, i.e. $\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$, $\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0$, and $\sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$. Other choices are also possible. Hence, the main effects have $a-1$ and $b-1$ degrees of freedom, respectively. The degrees of freedom of the interaction term are $(a-1)(b-1)$ which is the product of the degrees of freedom of the involved main effects.

The effects can be interpreted as follows: Think of main effects as "average effects" on the expected value of the response when changing the level of a factor (keeping the other factor fixed). The interaction effect can be thought of as a correction factor to the main effects model. The interaction effect tells us how much an effect of a certain factor changes, when we "switch" the level of the other factor. Strictly speaking, interpretation depends on the side-constraint that we apply. An illustration of the model can be found in Figure 9. It is nothing else than the "theoretical" version of the interaction plot where we have the expected value instead of the sample average on the y-axis.

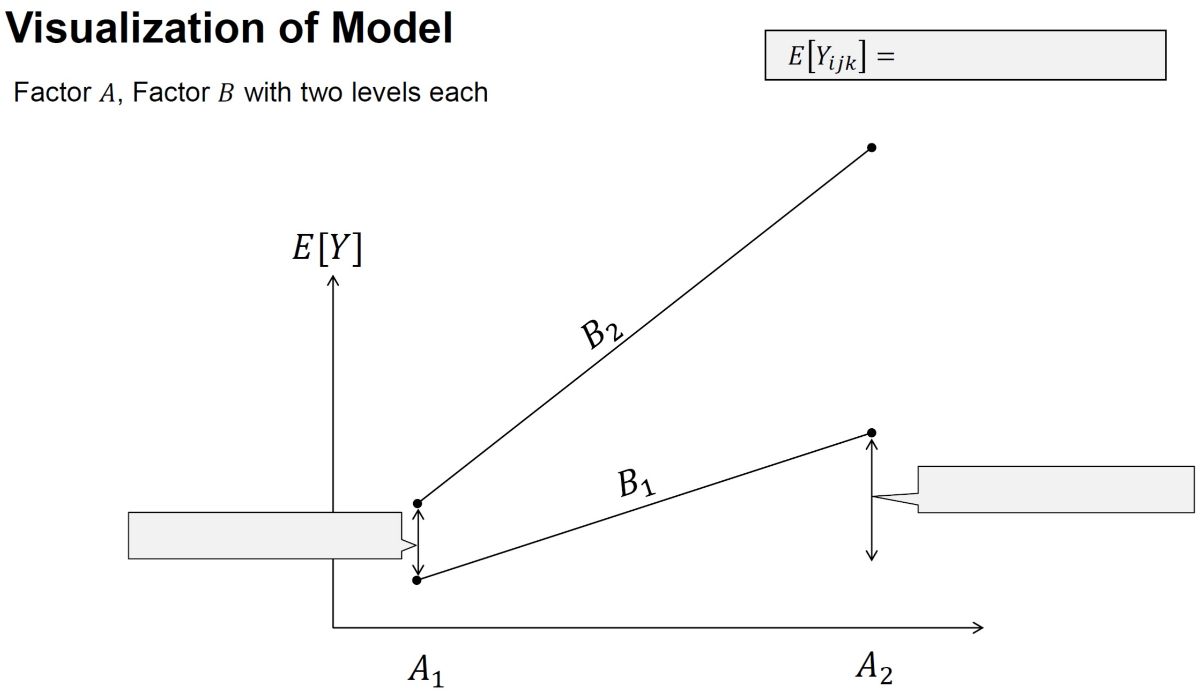
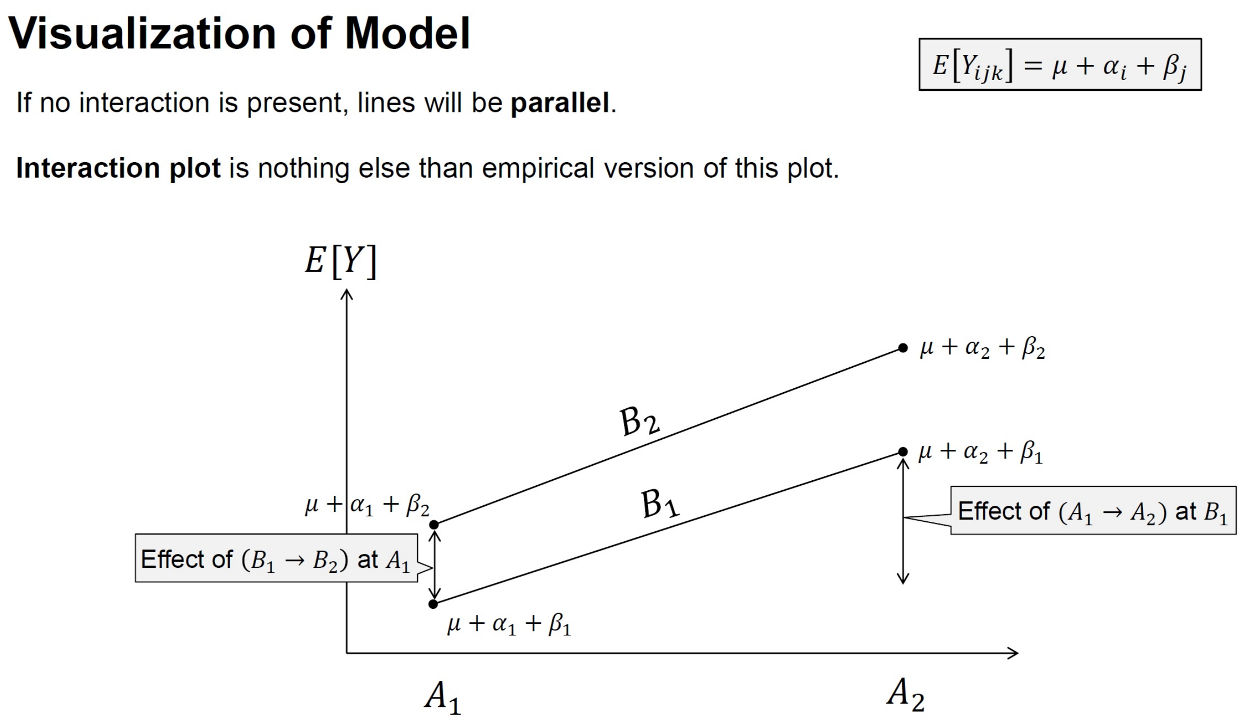


Figure 9: The illustration of an Interactive Model

A model without interaction term is **additive**. This means that the effect of A does not depend on the level of B (and vice versa), “it is always the same, no matter what the level of the other factor”. An illustration can be found in Figure 10, where we see that the lines are parallel. This means, changing B from `level 1` to `level 2` has always the same effect. Similarly, the effect of changing A does not depend on the level of B.



Visualization of Model

$E[Y_{ijk}] = \mu + \alpha_i + \beta_j$

If no interaction is present, lines will be **parallel**.

Interaction plot is nothing else than empirical version of this plot.

$E[Y]$
$\mu + \alpha_2 + \beta_2$
B_2
$\mu + \alpha_2 + \beta_1$
$\mu + \alpha_1 + \beta_2$
B_1
Effect of $(A_1 \rightarrow A_2)$ at B_1
Effect of $(B_1 \rightarrow B_2)$ at A_1
$\mu + \alpha_1 + \beta_1$
A_1
A_2

Figure 10: The illustration of an Additive Model

Parameter Estimates

As before, we estimate parameters using the principles of least squares. Using sum-to-zero side constraints we get

Parameter	Estimate
μ	
α_i	
β_j	
$(\alpha\beta)_{ij}$	

As before: if we replace an index with a dot, we take the mean (or the sum) over that "dimension". For example, $\overline{y}_{1\cdot\cdot}$ is the mean over all observations y_{1jk}, $j = 1, \ldots, b$ and $k = 1, \ldots, n$. Hence, we estimate the expected value of the response y_{ijk} for A at level i and B at level j as

which is nothing else than the mean of the observations in the corresponding "cell" (which is no surprise). Note however that we "untangled" the effect with respect to the two main effects and the interaction.

If we carefully inspect the parameter estimates from above, we see that for the main effects, we use an estimate that completely ignores the other factor. We basically treat the problem as a one-way ANOVA model. This is a consequence of the balanced design. For all levels of A we have the "same population" of B settings. Hence, if we compare $\overline{y}_{1\cdot\cdot}$ (think of taking the average over the first row) with $\overline{y}_{2\cdot\cdot}$ (average over second row) the effect is only due to changing A from level 1 to level 2. In regression terminology we would call this an **orthogonal design**.

5.3.3 Tests

As for the one-way ANOVA case, the total sum of squares SST can be partioned into different sources

where

Source	Sum of squares	Comment
A		
B		
AB		
Error		
Total		

The degrees of freedom of the error term is nothing else than the degrees of freedom of the total sum of squares ($abn - 1$, number of observations minus 1) minus the degrees of freedom of the other effects.

We can construct an ANOVA table based on this information

Source	df	SS	MS	F-ratio
A				

Source	df	SS	MS	F-ratio
B				
AB				
Error				
Total				

As before we can construct tests based on the corresponding F-distributions.

1. **interaction**

- H_0 :
- H_a :
- Under H_0 :

2. **main effect A**

- H_0 :
- H_a :
- Under H_0 :

3. **main effect B**

- H_0 :
- H_a :
- Under H_0 :

Typically, the F-test are analyzed from **bottom to top** (in the ANOVA table). Here, this means **we start with the F-test of the interaction**.

- **If this test indicates that there is not a significant interaction, then continue testing the hypotheses for the two main effects**:
- **If this test indicates that there is a significant interaction, then the interpretation of significant main effects hypotheses can be masked. To draw conclusions about a main effect, we will fix the levels of one factor and vary the levels of the other. Using this approach (combined with interaction plots) we may be able to provide an interpretation of main effects**.

5.3.4 Examples

Example 1: Beetle Insecticide Two-factor Design

Figure 11 gives survival times of groups of **4** beetles randomly allocated to twelve treatment groups obtained by crossing the levels of **four insecticides** (A, B, C, D) at each of **three concentrations** of the insecticides (1 = Low, 2 = Medium, 3 = High). This is a balanced 4-by-3 factorial design (two-factor design) that is replicated 4 times. The unit of measure for the survival times is 10 hours, that is, 0.3 is a survival time of 3 hours.

	dose	insecticide	t1	t2	t3	t4
1	low	A	0.3100	0.4500	0.4600	0.4300
2	low	B	0.8200	1.1000	0.8800	0.7200
3	low	C	0.4300	0.4500	0.6300	0.7600
4	low	D	0.4500	0.7100	0.6600	0.6200
5	medium	A	0.3600	0.2900	0.4000	0.2300
6	medium	B	0.9200	0.6100	0.4900	1.2400
7	medium	C	0.4400	0.3500	0.3100	0.4000
8	medium	D	0.5600	1.0200	0.7100	0.3800
9	high	A	0.2200	0.2100	0.1800	0.2300
10	high	B	0.3000	0.3700	0.3800	0.2900
11	high	C	0.2300	0.2500	0.2400	0.2200
12	high	D	0.3000	0.3600	0.3100	0.3300

Figure 11: Beetle Data

Interpretation of the Dose and Insecticide Effects

The interpretation of the dose and insecticide **main effects** depends on whether interaction is present. The distinction is important, so I will give both interpretations to emphasize the differences.

Given the test for interaction, I would likely summarize the main effects assuming no interaction.

- The average survival time decreases as the dose increases, with estimated mean survival times of 0.618, 0.544, and 0.276, respectively.
- A Bonferroni comparison shows that the population mean survival time for the high dose (averaged over insecticides) is significantly less than the population mean survival times for the low and medium doses (averaged over insecticides). The two lower doses are not significantly different from each other. This leads to two dose groups: `Low` and `High`.

If dose and insecticide interact, you can conclude that beetles given a high dose of the insecticide typically survive for shorter periods of time averaged over insecticides.

- You can not, in general, conclude that the highest dose yields the lowest survival time regardless of insecticide.

- For example, the difference in the medium and high dose marginal means ($0.544 - 0.276 = 0.268$) estimates the typical decrease in survival time achieved by using the high dose instead of the medium dose, averaged over insecticides. If the two factors interact, then the difference in mean times between the medium and high doses on a given insecticide may be significantly greater than 0.268, significantly less than 0.268, or even negative. In the latter case the medium dose would be better than the high dose for the given insecticide, even though the high dose gives better performance averaged over insecticides.

- An interaction forces you to use the cell means to decide which combination of dose and insecticide gives the best results (and the multiple comparisons as they were done above do not give multiple comparisons of cell means; a single factor variable combining both factors would need to be created). Of course, our profile plot tells us that this hypothetical situation is probably not tenable here, but it could be so when a significant interaction is present.

If dose and insecticide do not interact, then the difference

in marginal dose means averaged over insecticides also estimates the difference in population mean survival times between two doses, regardless of the insecticide.

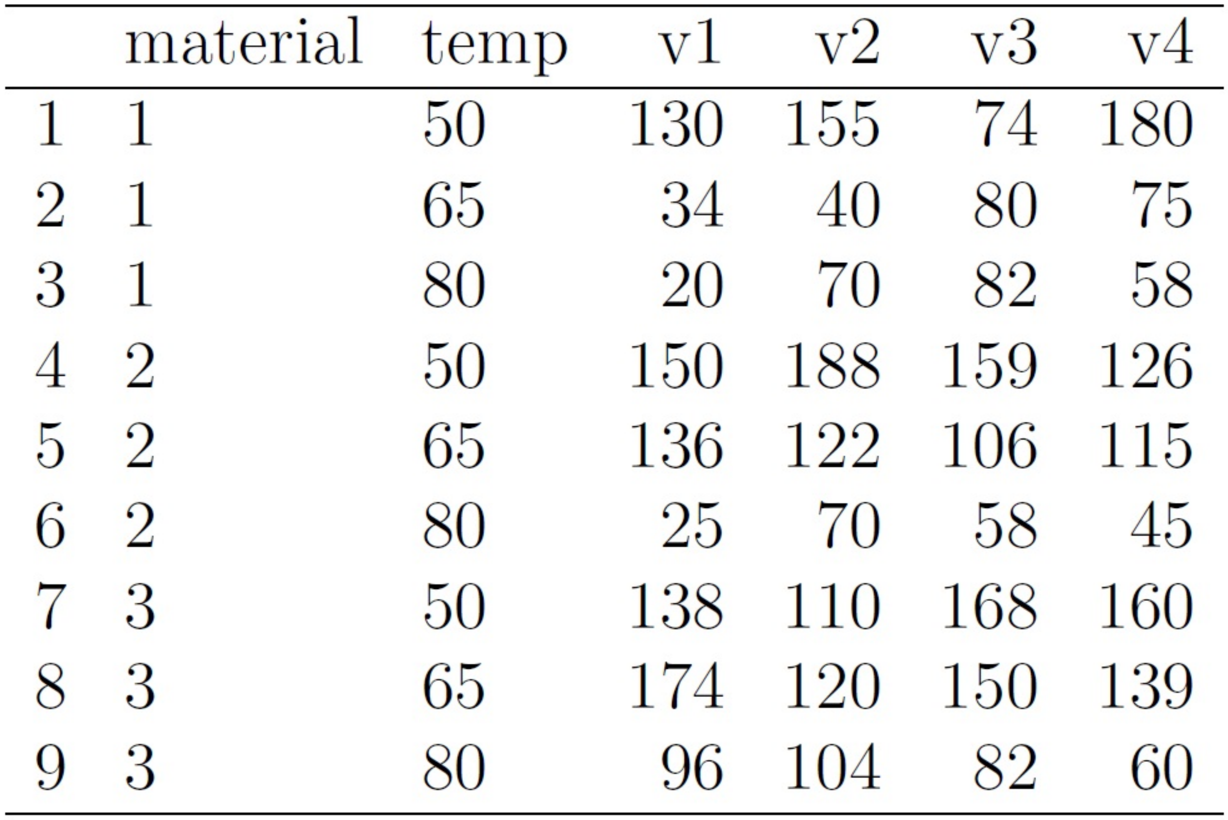
- This follows from the parallel profiles definition of no interaction.
- Thus, the difference in the medium and high dose marginal means ($0.544 - 0.276 = 0.268$) estimates the expected decrease in survival time anticipated from using the high dose instead of the medium dose, regardless of the insecticide (and hence also when averaged over insecticides).
- A practical implication of no interaction is that you can conclude that the high dose is best, regardless of the insecticide used. The difference in marginal means for two doses estimates the difference in average survival expected, regardless of the insecticide.

The same rule applies for analyzing the `insecticide` effect.

Example 2: Output Voltage for Batteries

The `maximum output voltage` for storage batteries is thought to be influenced by the `temperature` in the location at which the battery is operated and the `material` used in the plates.

A scientist designed a two-factor study to examine this hypothesis, using **three temperatures** (50, 65, 80), and **three materials** for the plates (1, 2, 3). Four batteries were tested at each of the 9 combinations of temperature and material type. The maximum output voltage was recorded for each battery. This is a balanced 3-by-3 factorial experiment with 4 observations per treatment.



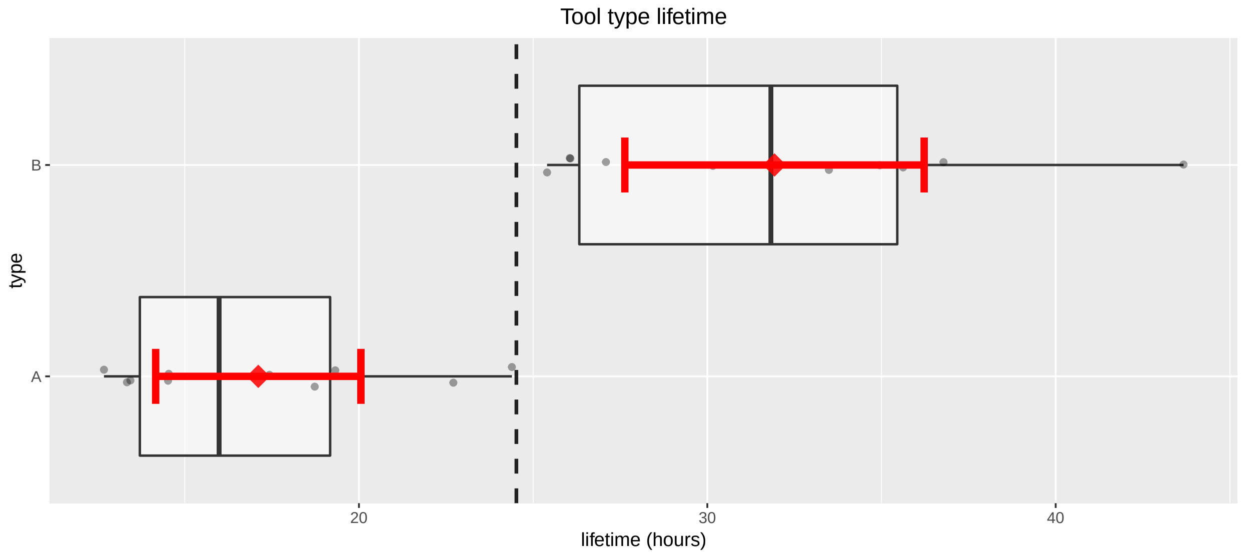
	material	temp	v1	v2	v3	v4
1	1	50	130	155	74	180
2	1	65	34	40	80	75
3	1	80	20	70	82	58
4	2	50	150	188	159	126
5	2	65	136	122	106	115
6	2	80	25	70	58	45
7	3	50	138	110	168	160
8	3	65	174	120	150	139
9	3	80	96	104	82	60

Figure 12: Battery Data

5.4 Analysis of Covariance: Comparing Regression Lines

Suppose that you are interested in comparing the typical `lifetime` (hours) of two `tool types` (A and B). A simple analysis of the data given below would consist of making side-by-side boxplots followed by a two-sample test of equal means (or medians). The standard two-sample test using the pooled variance estimator is a special case of the one-way ANOVA with two groups. The summaries suggest that the distribution of lifetimes for the tool types are different. In the output below, μ_i is population mean lifetime for tool type i $(i = A, B)$.

```
#>    lifetime  rpm type
#> 1     18.73  610    A
#> 2     14.52  950    A
#> 3     17.43  720    A
#> 4     14.54  840    A
#> 5     13.44  980    A
#> 6     24.39  530    A
#> 7     13.34  680    A
#> 8     22.71  540    A
#> 9     12.68  890    A
#> 10     19.32  730     A
#> 11     30.16  670     B
#> 12     27.09  770     B
#> 13     25.40  880     B
#> 14     26.05 1000     B
#> 15     33.49  760     B
#> 16     35.62  590     B
#> 17     26.07  910     B
#> 18     36.78  650     B
#> 19     34.95  810     B
#> 20     43.67  500     B
```

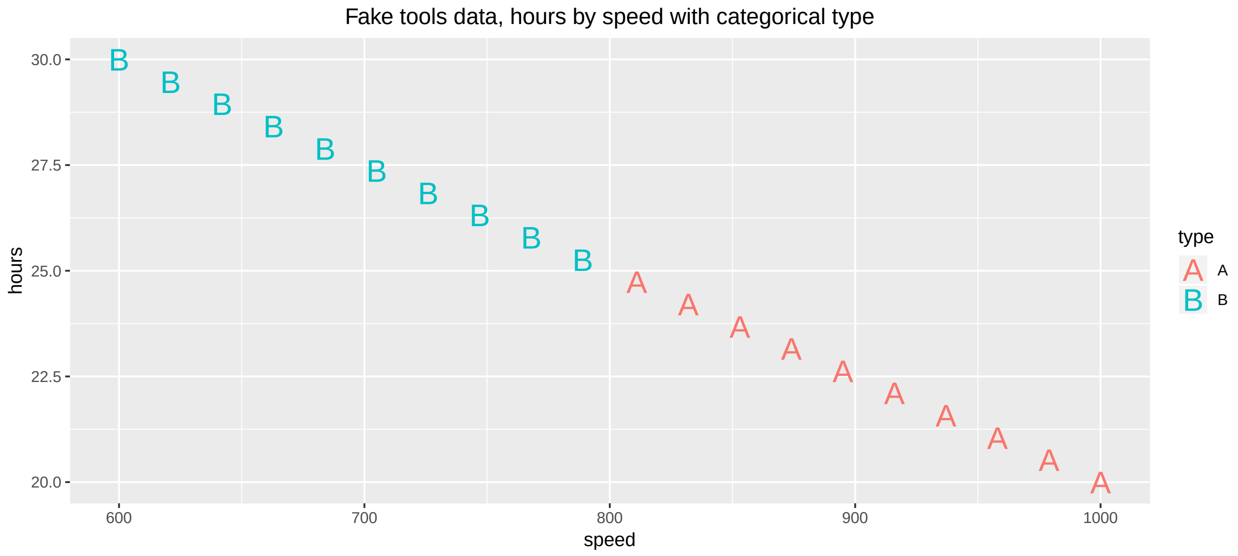


A two sample t-test comparing mean lifetimes of tool types indicates a difference between means.

```
t.summary = t.test(lifetime ~ type, data = tools)
t.summary
#>
#>  Welch Two Sample t-test
#>
#> data:  lifetime by type
#> t = -6.435, df = 15.93, p-value = 8.422e-06
#> alternative hypothesis: true difference in means is not equal
#> 95 percent confidence interval:
#>  -19.70128  -9.93472
#> sample estimates:
#> mean in group A mean in group B
#>          17.110          31.928
```

This comparison is potentially misleading because the samples are not comparable. **A one-way ANOVA is most appropriate for designed experiments where all the factors influencing the response, other than the treatment (tool type), are fixed by the experimenter**. The tools were operated at different speeds. If `speed` influences lifetime, then the observed differences in lifetimes could be due to differences in speeds at which the two tool types were operated.

Fake example: For example, suppose **speed is inversely related to lifetime of the tool**. Then, the differences seen in the boxplots above could be due to tool type B being operated at lower speeds than tool type A. To see how this is possible, consider the data plot given below, where the relationship between lifetime and speed is identical in each sample. A simple linear regression model relating hours to speed, ignoring tool type, fits the data exactly, yet the lifetime distributions for the tool types, ignoring speed, differ dramatically. (The data were generated to fall exactly on a straight line). **The regression model indicates that you would expect identical mean lifetimes for tool types A and B, if they were, or could be, operated at identical speeds**. This is not exactly what happens in the actual data. However, I hope the point is clear.



warning: you should **be wary of group comparisons where important factors that influence the response have not been accounted for or controlled**. For the tool lifetime problem, **you should compare groups (tools) after adjusting the lifetimes to account for the influence of a measurement variable,** **`speed`**. The appropriate statistical technique for handling this problem is called **analysis of covariance (ANCOVA)**.

5.4.1 ANCOVA

A natural way to account for the **effect of speed** is through a multiple regression model with `lifetime` as the response and two predictors, `speed` and tool type, A binary categorical variable.

Consider the model:

where type B is 0 for type A tools, and 1 for type B tools.

This ANCOVA model fits two regression lines, one for each tool type, but restricts the slopes of the regression lines to be identical. To see this, let us focus on the interpretation of the regression coefficients. For the ANCOVA model,

- $\beta_2 =$

- $\beta_0 =$

- $\beta_0 + \beta_1 =$

- $\beta_1 =$

A picture of the population regression lines for one version of the model is given below.

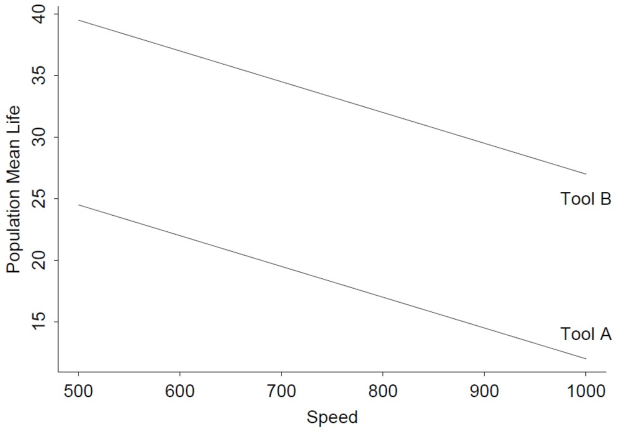


Figure 13: ANCOVA Regression Lines

An important feature of the ANCOVA model is that β_1 measures the difference in mean response for the `tool types`, regardless of the `speed`. A test of $H_0 : \beta_1 = 0$ is the primary interest, and is interpreted as a **comparison of the tool types, after adjusting or allowing for the speeds at which the tools were operated**.

The ANCOVA model is plausible. The relationship between `lifetime` and `speed` is roughly linear within tool types, with similar slopes but unequal intercepts across groups. The plot of the studentized residuals against the fitted values shows no gross abnormalities, but suggests that the variability about the regression line for tool type A is somewhat smaller than the variability for tool type B. The model assumes that the variability of the responses is the same for each group. The QQ-plot does not show any gross deviations from a straight line.

- See R code for the analysis.
- The fitted relationship for the combined data set is
- The fitted relationship for tool type B:

- The fitted relationship for tool type A:

The t-test of $H_0 : \beta_1 = 0$ checks whether the intercepts for the population regression lines are equal, assuming equal slopes. The t-test p-value < 0.0001 suggests that the population regression lines for tools A and B have **unequal intercepts**. The LS lines indicate that the average lifetime of either type tool decreases by 0.0266 hours for each increase in 1 RPM. Regardless of the lathe speed, the model predicts that type B tools will last 15 hours longer (i.e., the regression coefficient for the type B predictor) than type A tools. Summarizing this result another way, the t-test suggests that **there is a significant difference between the lifetimes of the two tool types, after adjusting for the effect of the speeds at which the tools were operated**. The estimated difference in average lifetime is 15 hours, regardless of the lathe speed.

Generalizing the ANCOVA Model to Allow Unequal Slopes

I will present a flexible approach for checking equal slopes and equal intercepts in ANCOVA-type models. The algorithm also provides a way to build regression models in studies where the primary interest is **comparing the regression lines across groups rather than**

comparing groups after adjusting for a regression effect. The approach can be applied to an arbitrary number of groups and predictors. For simplicity, I will consider a problem with three groups and a single regression effect.

The `twins` data are the IQ scores of identical twins, one raised in a foster home (`IQF`) and the other raised by natural parents (`IQN`). The 27 pairs are divided into three groups by social status of the natural parents (H = high, M = medium, L = low). We will examine the regression of `IQF` on `IQN` for each of the three social classes.

There is no **a priori** reason to assume that the regression lines for the three groups have equal slopes or equal intercepts. These are, however, reasonable hypotheses to examine. The **easiest way to check these hypotheses is to fit a multiple regression model to the combined data set, and check whether certain carefully defined regression effects are zero**. The most general model has six parameters, and corresponds to fitting a simple linear regression model to the three groups separately ($3 \times 2 = 6$).

Two indicator variables are needed to uniquely identify each observation by social class. For example, let $I_1 = 1$ for H status families and $I_1 = 0$ otherwise, and let $I_2 = 1$ for M status families and $I_2 = 0$ otherwise. The indicators I_1 and I_2 jointly assume 3 values:

Table 9: Making Indicator Variables

Status	I_1	I_2
L		
M		
H		

Given the indicators I_1 and I_2 and the predictor `IQN`, define two **interaction** or **product effects**: $I_1\times$ IQN and $I_2\times$ IQN.

Unequal slopes ANCOVA model

The most general model allows separate slopes and intercepts for each group:

This model is best understood by considering the three status classes separately.

- If status = L, then $I_1 = I_2 = 0$. For these families

- If status = M, then $I_1 = 0$ and $I_2 = 1$. For these families

Finally, if status = H, then $I_1 = 1$ and $I_2 = 0$. For these families

The regression coefficients β_0 and β_3 are the intercept and slope for the L status population regression line. The other parameters measure differences in intercepts and slopes across the three groups, using L status families as a **baseline** or **reference group**. In particular:

- $\beta_1 =$
- $\beta_2 =$
- $\beta_4 =$
- $\beta_5 =$

The plot gives a possible picture of the population regression lines corresponding to the general model.

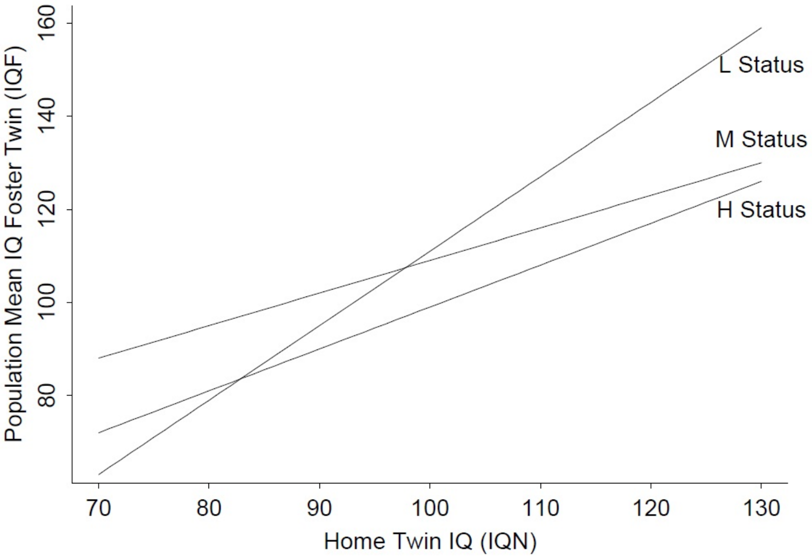


Figure 14: ANCOVA Twins General Model

- Now, write your R code to fit this model.

OK, let us write down the fitted model:

- for the baseline group with status = L,

- For the M status group with indicator I_2 and product effect $I_2\times$ IQN:

- For the H status group with indicator I_1 and product effect $I_1\times$ IQN:

The LS lines are identical to separately fitting simple linear regressions to the three groups.

There are three other models of potential interest besides the `general model` $(*)$.

Equal slopes ANCOVA model

The **equal slopes** ANCOVA model

is a special case of the `general model` $(*)$ with $\beta_4 = \beta_5 = 0$ (no interaction). In the ANCOVA model, β_3 is the slope for all three regression lines. The other parameters have the same interpretation as in the `general model`, see the plot below.

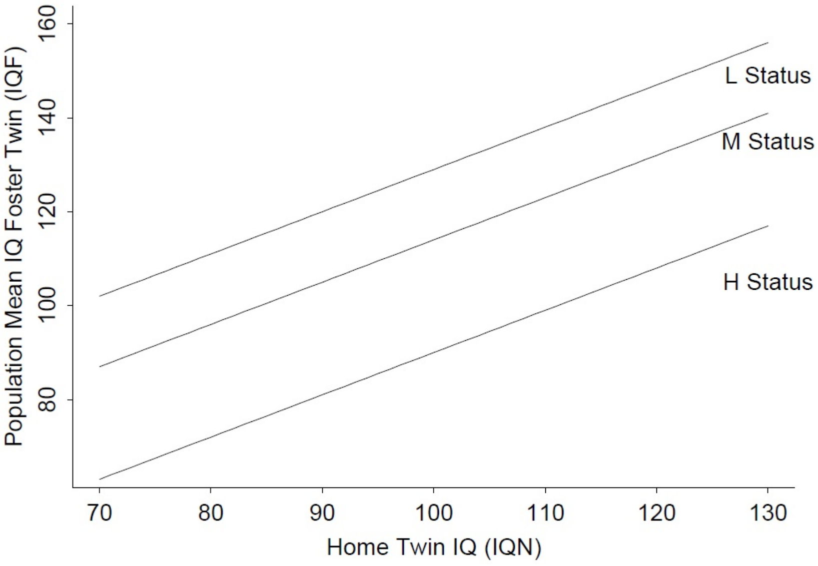


Figure 15: ANCOVA Twins Equal Slopes Model

- Again, fit this model in R.

To write down the fitted model:

- For L status families:

- For M status families:

- For H status families:

Equal slopes and equal intercepts ANCOVA model

The model with **equal slopes** and **equal intercepts**

is a special case of the ANCOVA model with $\beta_1 = \beta_2 = 0$. This model does not distinguish among social classes. The common intercept and slope for the social classes are β_0 and β_3, respectively. The predicted IQF for this model is

for each social class.

No slopes, but intercepts ANCOVA model

The model with **no predictor** (IQN) effects

is a special case of the ANCOVA model with $\beta_3 = 0$. In this model, social status has an effect on IQF but IQN does not. This model of **parallel regression lines** with **zero slopes** is identical to a one-way ANOVA model for the three social classes, where the intercepts play the role of the population means, see the plot below.

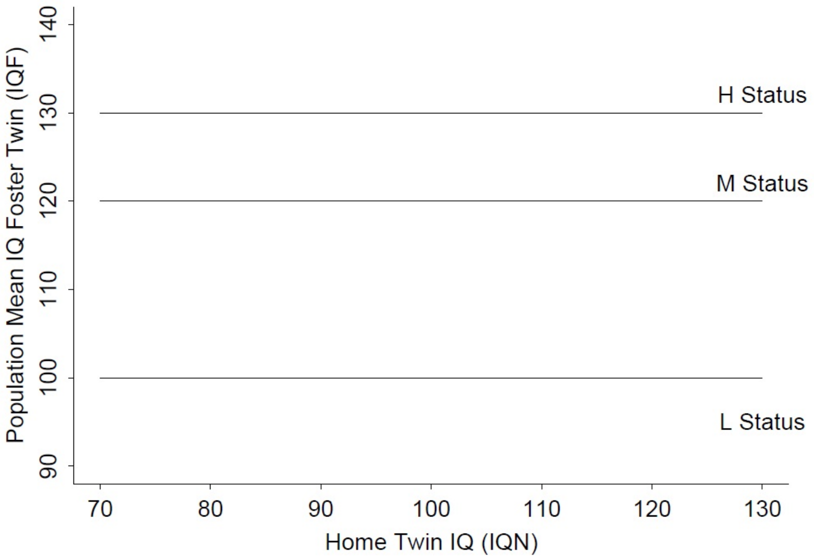


Figure 16: ANCOVA Twins Zero Slopes Model

To write down the fitted model:

- For L status families:

- For M status families:

- For H status families:

The predicted IQFs are the mean IQFs for the three groups.

Choosing Among Models

I will suggest a **backward sequential method** to select which of models $(*)$, $(**)$, and $(***)$ fits best. You would typically be interested in the one-way ANOVA model only when the effect of IQN was negligible.

Step 1: Fit the full model $(*)$ and test the hypothesis of equal slopes $H_0 : \beta_4 = \beta_5 = 0$. (aside: t-tests are used to test either $\beta_4 = 0$ or $\beta_5 = 0$.) To test H_0, eliminate the predictor variables $I_1 \cdot IQN$ and $I2 \cdot IQN$ associated with β_4 and β_5 from the full model $(*)$. Then fit the reduced model $(**)$ with equal slopes. Reject $H_0 : \beta_4 = \beta_5 = 0$ if the increase in the `Residual SS` obtained by deleting $I_1 \cdot IQN$ and $I2 \cdot IQN$ from the full model is significant. A p-value for F-test is obtained from `library(car)` with `Anova(aov(LMOBJECT), type = 3)` for the interaction. If H_0 is rejected, stop and conclude that the population regression lines have different slopes (and then I do not care whether the intercepts are equal). Otherwise, proceed to step 2.

Step 2: Fit the equal slopes or ANCOVA model ($**$) and test for equal intercepts $H0 : \beta_1 = \beta_2 = 0$. Follow the procedure outlined in Step 1, treating the ANCOVA model as the full model and the model $IQF = \beta_0 + \beta_3 IQN + \epsilon$ with equal slopes and intercepts as the reduced model. See the intercept term using `library(car)` with `Anova(aov(LMOBJECT), type = 3)`. If H_0 is rejected, conclude that that population regression lines are parallel with unequal intercepts. Otherwise, conclude that regression lines are identical.

Step 3: Estimate the parameters under the appropriate model, and conduct a diagnostic analysis. Summarize the fitted model by status class.

The plot of the twins data shows fairly linear relationships within each social class. The linear relationships appear to have similar slopes and similar intercepts. The p-value for testing the hypothesis that the slopes of the population regression lines are equal is essentially 1. The observed data are consistent with the reduced model of equal slopes.

The p-value for comparing the model of equal slopes and equal intercepts to the ANCOVA model is 0.238, so there is insufficient evidence to reject the reduced model with equal slopes and intercepts. The estimated regression line, regardless of social class, is:

$$\widehat{IQF} = 9.21 + 0.901 \cdot IQN$$

There are no serious inadequacies with this model, based on a diagnostic analysis (not shown).

An interpretation of this analysis is that the natural parents' social class has no impact on the relationship between the IQ scores of identical twins raised apart. What other interesting features of the data would be interesting to explore? For example, what values of the intercept and slope of the population regression line are of intrinsic interest?

5.4.2 Simultaneous Testing of Regression Parameters

In the twins example, we have this full interaction model,

$$IQF = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 IQN + \beta_4 I_1 \cdot IQN + \beta_5 I_2 \cdot IQN + \epsilon$$

Consider these two specific hypotheses:

1. H_0 : equal regression lines for status M and L
2. H_0 : equal regression lines for status M and H

That is, the intercept and slope for the regression lines are equal for the pairs of status groups.

First, it is necessary to formulate these hypotheses in terms of testable parameters. That is, find the β values that make the null hypothesis true in terms of the model equation.

1.

2.

Using linear model theory, there are methods for testing these multiple-parameter hypothesis tests.

One strategy is to use the Wald test of null hypothesis ____________ where (r) is a matrix of contrast coefficients (typically 1 or -1), $\boldsymbol{\beta}$ is our vector of regression coefficients, and r is a hypothesized vector of what the linear system $(r)\boldsymbol{\beta}$ equals.

For our first hypothesis test, the linear system we're testing in matrix notation is

In hypothesis 1 we are testing $\beta_2 = 0$ and $\beta_5 = 0$, which are the 3rd and 6th position for coefficients in our original model equation. However, we need to choose the correct positions based on the `coef()` order, and these are positions 4 and 6. The large p-value = 0.55 suggests that M and L can be described by the same regression line, same slope and intercept.

In hypothesis 2 we are testing $\beta_1 - \beta_2 = 0$ and $\beta_4 - \beta_5 = 0$ which are the difference of the 2nd and 3rd coefficients and the difference of the 5th and 6th coefficients. However, we need to choose the correct positions based on the `coef()` order, and these are positions 3 and 4, and 5 and 6. The large p-value = 0.91 suggests that M and H can be described by the same regression line, same slope and intercept.

The results of these tests are not surprising, given our previous analysis where we found that the status effect is not significant for all three groups.

Any simultaneous linear combination of parameters can be tested in this way.