Contents

List of Figures

List of Tables

1 Survey of Quantitative Methods

The bread and butter of quantitative consulting is the ability to apply quantitative methods to business problems in order to obtain actionable insight. Clearly, it is impossible (and perhaps inadvisable, in a more general sense) for any given individual to have expertise in every field of mathematics, statistics, and computer science.

We believe that the best consulting framework is reached when a small team of consultants possesses expertise in 2 or 3 areas, as well as a decent understanding of related disciplines, and a passing knowledge in a variety of other domains: this includes keeping up with trends, implementing knowledge redundancies on the team, being conversant in non-expertise areas, and knowing where to find detailed information (online, in books, or through external resources).

In this section, we present an introduction for 9 "domains" of quantitative analysis:

- survey sampling and data collection;
- data processing;
- data visualisation;
- statistical methods;
- queueing models;
- data science and machine learning;
- simulations;
- optimisation, and
- trend extraction and forecasting;

Strictly speaking, the domains are not free of overlaps. Large swaths of data science and time series analysis methods are quite simply statistical in nature, and it's not unusual to view optimisation methods and queueing models as sub-disciplines of operations research. Other topics could also have been included (such as Bayesian data analysis or signal processing, to name but two), and might find their way into a second edition of this book.

Our treatment of these topics, by design, is brief and incomplete. Each module is directed at students who have a background in other quantitative methods, but not necessarily in the topic under consideration. Our goal is to provide a quick "reference map" of the field, together with a general idea of its challenges and common traps, in order to highlight opportunities for application in a consulting context. These subsections are emphatically NOT meant as comprehensive surveys: they focus on the basics and talking points; perhaps more importantly, a copious number of references are also provided.

We will start by introducing a number of motivating problems, which, for the most part, we have encountered in our own practices. Some of these examples are reported on in more details in subsequent sections, accompanied with (partial) deliverables in the form of charts, case study write-ups, report extract, etc.).

As a final note, we would like to stress the following: it is **IMPERATIVE** that quantitative consultants remember that acceptable business solutions are not always optimal theoretical solutions. Rigour, while encouraged, often must take a backseat to applicability. This lesson can be difficult to accept, and has been the downfall of many a promising candidate.

1.4 Statistical Analysis

Loosely speaking, a **statistic** is any function of a sample from the distribution of a random variable; statistics aim to extract information from an observed sample to summarise the essential features of a dataset.

In general, statistics can be divided into two categories based on their purposes: **descriptive statistics** and **inferential statistics**.

As its name implies, descriptive statistics aim to describe the collected data; examples include:

- sample size (overall and/or subgroups);
- demographic breakdowns of participants;
- measures of central tendencies (e.g., mean, median, mode, etc.), and
- measures of variability (e.g., sample variance, minimum, maximum, interquartile range, etc.).

They can be presented as a single number, in a summary table, or even in graphical representations (e.g., histogram, pie chart, etc.) Descriptive statistics can be extended to summarise **multivariate** behaviours, *via* sample correlations, contingency tables, scatter plots, etc.

Descriptive statistics not only provide an easy-to-understand overview of the dataset, but they also give the consultant a chance to study the collected sample and investigate two important questions:

does the sample make sense? and **is the sample representative?**

Inferential statistics, on the other hand, aim to facilitate the process of inference (**induction**) to the general population from which the sample is drawn.

In this (criminally) brief tour of a far-reaching and ubiquitous subject, we will highlight ten areas of particular interest for consultants; further details can be found in [1–7].

1.4.1 Hypothesis Testing

In a very broad sense, most of statistical inference is done through **hypothesis testing** – are the client's conjectures about their business situation compatible with the evidence provided by the data? Is there a way to get a quantitative ruling in favour of competing hypotheses that relies on something other than the client's gut feeling?

Suppose that a researcher wants to determine if, as she believes, a new teaching method enables students to understand elementary statistical concepts better than the traditional lectures given in a university setting. She recruits $N = 80$ second-year students to test her claim. The students are randomly assigned to one of two groups: students in group A are given the traditional lectures, whereas students in group B are taught using the new teaching method. After three weeks, a short quiz is administered to the students in order to assess their understanding of statistical concepts – Table 1 summarises the results.

Group	Sample Size	Sample Mean	Sample Variance
A	$N_A = 40$	$\bar{y}_A = 75.2$	$S_A^2 = 6.3$
B	$N_B = 40$	$\bar{y}_B = 79.1$	$S_B^2 = 5.4$

Table 1: Summary of teaching method study example

If we assume that both groups have similar background knowledge prior to being taught (which we attempt to do by randomising the group assignment), then the effectiveness of the teaching methods may be compared using two hypotheses, the **null hypothesis** H_0 and the **alternative** H_a. **One-sided testing** pits

$$H_0 : \mu_A \geq \mu_B \quad \text{against} \quad H_a : \mu_A < \mu_B$$

(or the reverse); in **two-sided testing**, we have

$$H_0 : \mu_A = \mu_B \quad \text{against} \quad H_a : \mu_A \neq \mu_B.$$

Intuitively, testing for inequality of method seems looser than testing for the superiority of a specific method over the other.

Hypothesis testing can generate two types of error: we can mistakenly reject H_0 when it is, in fact, correct (**type I error**), or we can mistakenly accept H_0 when it is actually false (**type II error**). In order to control the probability of making a type I error (called **significance level**, and denoted by α), we usually let the hypothesis of interest be the alternative hypothesis.

Since the researcher wants to claim that the new method is more effective than the traditional ones, then it is most appropriate for her to use one-sided hypothesis testing with

$$H_0 : \mu_A \geq \mu_B \quad \text{against} \quad H_1 : \mu_A < \mu_B;$$

The testing procedure is simple

1. calculate a **test statistic** under H_0;
2. reject H_0 in favour of H_1 if the test statistic falls in the **critical region** (also called **rejection region**) of an associated distribution, and
3. accept H_0 otherwise – or rather, fail to reject it (see Figure 1).

Using the summary table above, we can test the researcher's claim by using the **two-sample** t **test**. Assuming that variability in two groups are roughly the same, the test statistic is given by:

$$t_0 = \frac{\bar{y}_B - \bar{y}_A}{S_p \sqrt{\frac{1}{N_A} + \frac{1}{N_B}}},$$

where the **pooled variance** S_p^2 is

$$S_p^2 = \frac{(N_A - 1)S_A^2 + (N_B - 1)S_B^2}{N_A + N_B - 2}.$$

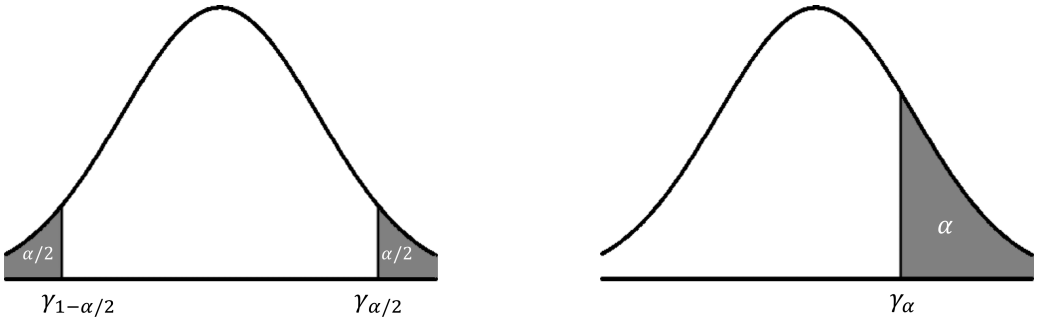


Figure 1: Critical regions for hypothesis testing at α (in grey); two-sided on the left, one-sided on the right; γ_k represent the critical value for the given test and underlying distribution.

In the example, the test statistic is $t_0 = 7.211$. To reject or accept the null hypothesis, we need to compare it against the **critical value** of the Student T distribution with $N-2 = 78$ degrees of freedom at $\alpha = 0.05$, which is

$$t^* = t_{1-\alpha,N-2} = t_{0.95,78} = 1.665.$$

Since $t_0 > t^*$ at $\alpha = 0.05$, we can conclude that we have enough evidence to believe that new teaching method is indeed more effective than the traditional methods, at $\alpha = 0.05$.

IMPORTANT NOTE: in general, the challenge is to recognise which test statistic to use and how it is distributed under H_0. Various scenarios have been explored in the literature (see [2], for instance) and it would be important for statistical consultants to be able to derive their own tests when the client's data does not meet the various assumptions. Ad-hoc solutions come at a price, however – a fair number of clients (and reviewers), if they are familiar with statistical tests at all, do not understand how they are derived and thus only trust 'tried, tested, and true' methods (this applies to other fields of quantitative analysis). New tests and approaches are likely to be treated with **suspicion**.

Questions to Ponder

1. Distribution assumptions:
   - what distribution assumptions are we making by using a $t-$test?
   - how can we verify them?
   - if such assumptions are violated, what is our recourse?
2. Assumption of equal variance:
   - how can we verify the appropriateness of using pooled variance?
   - if it is not appropriate, can we modify the test to overcome the problem?
3. One-sided vs. two-sided tests:
   - when is it appropriate to use a one-sided test, and when is it better to employ a two-sided test?
   - are there drawbacks in using a two-sided test when a one-sided test would be indicated?

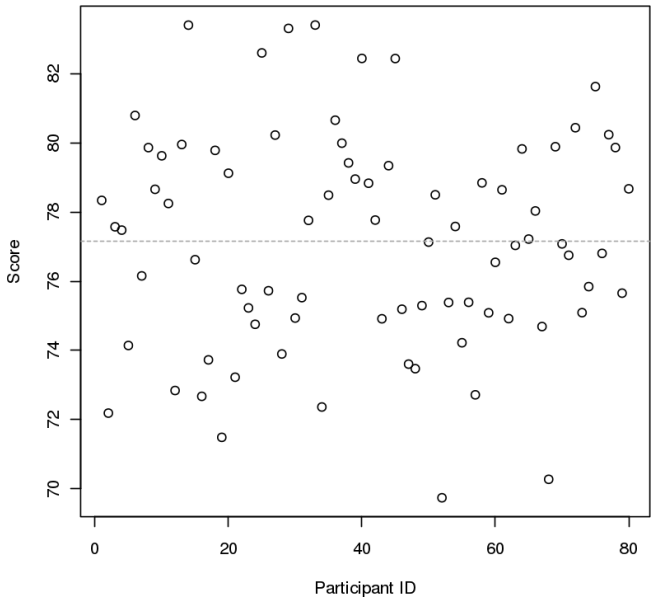


Figure 2: Effectiveness of new teaching method; the grey line is the overall sample mean.

1.4.2 Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a statistical method that partitions a dataset's variability into **explainable variability** (model-based) and **unexplained variability** (error) using various statistical models, to determine whether (multiple) treatment groups have significantly different group means.

The **total sample variability** of a feature y in a dataset is defined as

$$\text{SS}_{\text{tot}} = \sum_{k=1}^{N} (y_k - \bar{y})^2,$$

where $\bar{y}$ is the overall mean of the data.

Let us return to the teaching method example given in Section 1.4.1. Figure 2 shows all the students' scores, ordered by participant ID. Since the assignment of ID is **arbitrary** (at least, in theory), we do not observe any patterns – if we were to guess someone's score with no knowledge except for their participant ID, then picking the sample mean is as good as any other reasonable guesses.

Statistically speaking, this means that the **null model**

$$y_{i,j} = \mu + \varepsilon_{i,j},$$

where μ is the **overall mean**, $i = A, B$, and $j = 1, \ldots, 40$, does not explain any of the variability in the student scores (as usual, $\varepsilon_{i,j}$ represents the departure or noise from the model prediction).

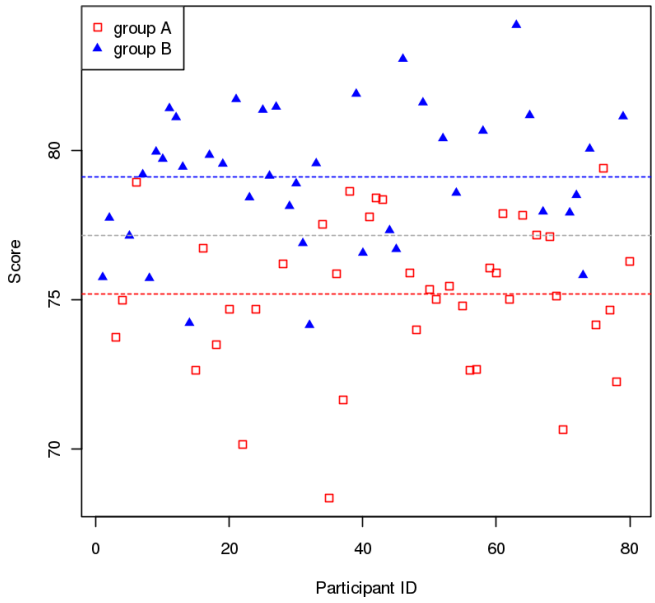


Figure 3: Effectiveness of new teaching method for two groups. the grey line is the overall sample mean, while the red and blue lines represent the average score for groups A and B, respectively.

But the students did not all receive the same treatment – 40 randomly selected students were assigned to group A, and the other 40 to group B. When we add this information onto Figure 2, we clearly see that the two study groups show different characteristics in term of their average scores (see Figure 3). Using their group assignment information, we can refine our null model into the **treatment-based model**

$$y_{i,j} = \mu_i + \varepsilon_{i,j},$$

where μ_i, $i = A, B$ represent the group means. Using this model, we can decompose SS_{tot} into **between-treatment sum of squares** and **error (within-treatment) sum of squares** as

$$\begin{aligned}\text{SS}_{\text{tot}} &= \sum_{i,j}(y_{i,j} - \bar{y})^2 = \sum_{i,j}(y_{i,j} - \bar{y}_i + \bar{y}_i - \bar{y})^2 \\ &= \sum_i N_i(\bar{y}_i - \bar{y})^2 + \sum_{i,j}(y_{i,j} - \bar{y}_i)^2 = \text{SS}_{\text{treat}} + \text{SS}_{\text{e}}\end{aligned}$$

The SS_{treat} component looks at the difference between each of the treatment means and the overall mean, which is explainable; the SS_{e} component, on the other hand, looks at the difference between each observation and its group mean. Clearly, the treatment-based model on its own cannot explain the cause of this variability.

In short, using a treatment-based model, we can explain $\text{SS}_{\text{treat}}/\text{SS}_{\text{tot}} \times 100\%$ of the total variability. This ratio is called the **coefficient of variation**, and is denoted by R^2.

Formally, the ANOVA table incorporates a few more items – Table 2 summarises all the information it contains; the ANOVA table for the teaching methodology example is shown in 3.

The test statistic F_0 follows an F-distribution with $(\text{d.f.}_{\text{treat}}, \text{d.f.}_{\text{e}}) = (1, 78)$ degrees of freedom. At a significance level of $\alpha = 0.05$, the critical value $F^* = F_{0.95,1,78} = 3.96$ is substantially smaller than the test statistic $F_0 = 52$, implying that the two-treatment model is statistically significant.

Source	Sum of Squares	d.f.	Mean Square	F_0
Treatment (Model)	SS_{treat}	$p-1$	$MS_{treat} = SS_{treat}/(p-1)$	MS_{treat}/MS_e
Error	SS_e	$N-p$	$MS_e = SS_e/(N-p)$	
Total	SS_{tot}	$N-1$		

Table 2: A simple ANOVA table, with p treatments and N observations.

Source	Sum of Squares	d.f.	Mean Square	F_0
Treatment (Model)	304.2	1	304.2	52.0
Error	456.3	78	5.85	
Total	760.5	79		

Table 3: ANOVA table for the teaching methodology example, with $p = 2$ and $N = 80$.

This, in turn, means that the model recognises a statistically significant difference between the students scores, based on the teaching methods.

The coefficient R^2 provides a way to measure the model's **significance**. From Table 3, we can compute $R^2 = \frac{304.2}{760.5} = 0.4$, which means that 40% of the total variation in the data can be explained by our two-treatment model. Is this good enough? That depends on the project, and on the client's needs.

Diagnostic Checks As with most statistical procedures, ANOVA relies on certain assumptions for the its result to be valid. Recall that our model is given by

$$y_{i,j} = \mu_i + \varepsilon_{i,j}$$

What assumptions are being imposing? The main assumption is that the error terms follow independently and identically distributed (i.i.d.) normal distributions (i.e., $\varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$). We are thus required to verify three assumptions:

- normality of the error terms;
- constant variance (within treatment groups), and
- equal variances (across treatment groups).

Normality can be tested visually with the help of a **normal-QQ plot**, which compares the standardized residuals quantiles against the theoretical quantiles of the standard normal distribution (a straight line indicates normality). Figure 4 shows some departure in the lower tail, however, moderate departure from normality is usually acceptable as long as it is mostly a tail phenomenon.

To test the assumption of constant variance, we can run visual inspection using (a) residuals vs. fitted values, and (b) residuals vs. order/time. Figure 5 shows that variability from the mean in each treatment group is reasonably similar. If a distinct difference arises and we cannot conclude that the group variances are constant, we will need to apply a **variance stabilising transformation**, such as a **logarithmic transformation** or **square-root transformation** before proceeding.

Formally, equality of variance is often tested using **Bartlett's test** (when normality of the residuals is met) or the **modified Levene's test** (when it is not).

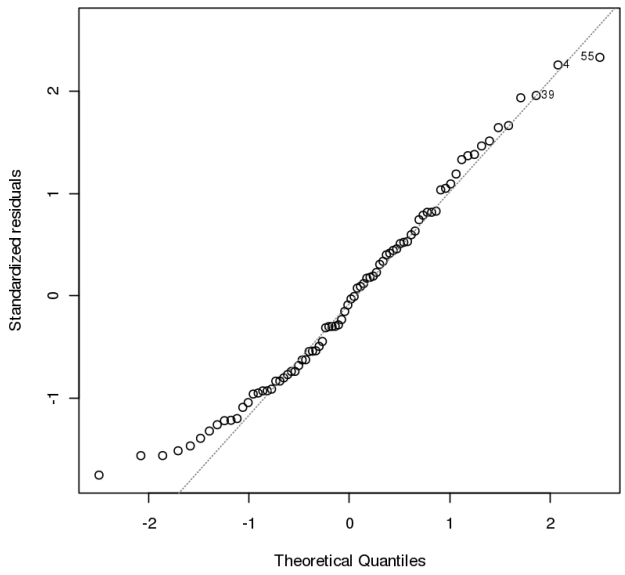


Figure 4: Normal QQ-plot for the two-treatment teaching model (standardised residuals); note the moderate (but acceptable) departure in the lower tail.

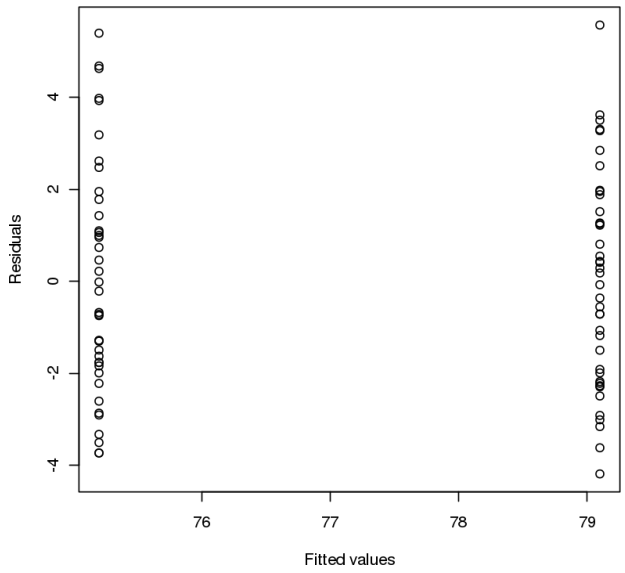


Figure 5: Diagnostic check for constant variance in the two-treatment teaching model. The spread is fairly similar; we can safely assume constant variance (as well as equal variance across treatment groups).

IMPORTANT NOTES: when there are more than $p > 2$ treatment groups, ANOVA provides a test for $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$ vs. $H_1 : \mu_i \neq \mu_j$ for at least one of $i \neq j$. A significant F_0 value indicates that *there is at least one group which differs from others*, but it does not specify which one(s) that may be. Specialised comparison methods such as **Scheffe's method** and **Tukey's test** can be used to identify the statistically different treatments.

Finally, while ANOVA can accommodate unequal treatment group sizes, it is recommended to keep those sizes equal across all groups – this makes the test statistic less sensitive to violations of the assumption of equal variances across treatment groups, providing yet another reason to involve the consultant with the **data collection process**.

1.4.3 Multiple Linear Regression

In sections 1.4.1 and 1.4.2, we considered a scenario where a single, categorical, explanatory variable (Treatment A vs. Treatment B) was used to model a desired response variable (score y). Real-world data is, of course, much more intricate and complex, typically consisting of multiple response variables, with multiple quantitative and categorical/qualitative explanatory features. In this section, we will review how such cases can be handled.

Multiple Linear Regression in Matrix Form Throughout, we suppose that the dataset consists of N observations with a single response output Y and p explanatory variables $X_1, \ldots, X_p$. The **first-order linear model** describing this scenario can be represented in matrix from by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y} = [y_1, \cdots, y_N]^\top$, $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]^\top$, and $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_N]^\top$ are the **response vector**, the **coefficient vector**, and the **error vector**, respectively, and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,p} \end{bmatrix}$$

is the **design matrix**, with the further assumption that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, where $\boldsymbol{I}$ is the $N \times N$ **identity matrix**.

Qualitative Explanatory Variables It has been said that the colour of a vehicle is part of the assessment for car insurance premiums (whether this is true or not, we are not qualified to discuss). Such a variable is **qualitative** (nominal, in fact) in nature, as there is no reasonable way to order colours for insurance purposes. If we want to incorporate this feature in an insurance premium model taking into account k possible colour choices, then we need $k-1$ dummy variables $X_1, \ldots, X_{k-1}$ defined according to the form of

$$X_1 = \begin{cases} 1 & \text{if colour} = \text{red} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if colour} = \text{black} \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$X_{k-1} = \begin{cases} 1 & \text{if colour} = \text{forest green} \\ 0 & \text{otherwise} \end{cases}$$

With **ordinal variables** (e.g., *on scale of 1 to 5, how likely are you to buy a new phone this year?*), we may choose to have 4 dummy variables as above, or a single continuous variable. While the latter approach saves 4 degrees of freedom, we are imposing an assumption that equal spacings on the ordinal have an equal impact on the outcome, which is not always the case – in which case dummy variables might be indicated.

Source	Sum of Squares	d.f.	Mean Square	F_0
Regression	SS_{reg}	$p-1$	$MS_{reg} = SS_{reg}/(p-1)$	MS_{reg}/MS_e
Error	SS_e	$N-p$	$MS_e = SS_e/(N-p)$	
Total	SS_{tot}	$N-1$		

Table 4: ANOVA table for first-order multiple regression model (1); with p explanatory variables and N observations.

Overall Significance of the Model For the model presented in (1), **ordinary least square** (OLS) estimation yields **fitted values**

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

and residuals

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}^\top)\boldsymbol{y}.$$

The ANOVA table has the same form as Table 2 (see Table 4); it is used in testing

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{against} \quad H_1 : \beta_i \neq 0 \text{ for at least one } i.$$

If the test statistic F_0 is **significant**, it does not necessarily imply that all the independent variables $X_1, \ldots, X_p$ are useful in predicting **y**, only that at least one of them is. We can examine significance of the β coefficients individually (using t-test), or multiple coefficients simultaneously (e.g., **Bonferroni simultaneous confidence interval**). Choosing the best subset of the model will be discussed in Sections 1.4.4 and **??**.

Model Adequacy Checks There are some rare examples for which OLS does not yield a unique solution; but in the vast majority of instances, the data can be fitted to the model. How can we tell if the model is **adequate** to the situation at hand?

- **Assumptions on Residuals** – We cannot emphasise enough that **the model is not necessarily valid when it is statistically significant** (i.e. when F_0 is in the critical region); the conclusion only follows once the model has been determined to be an **adequate** fit for the data. A normal-QQ plot can help verify the assumption of normality, for instance, while the assumptions of independence and constant variance can be tested using scatterplots of fitted values against residuals.

- **Outliers and Influential Points** – In addition, **outliers** and **influential points** could affect the fitted values. While it is typically easier to classify some observations as outliers, influential points can distort the regression line significantly. Figure 6 shows the clear impact of an influential point. Outliers and influential points should be studied carefully, as there are a number of possible mechanisms that can account for their presence; it may be that these anomalies are due to data entry error, in which case we may try to correct/impute with a reasonable alternative, if possible (see Section **??**). It may be the case that these unusual observations are worth studying on their own merit.

- **Multicollinearity** and **Variance Inflation Factor** (VIF) – Last but not least, it is important to take a look at the scatterplot matrix and the correlation matrix of the explanatory variables to detect **multicollinearity**. While it is hoped that the explanatory variables have some

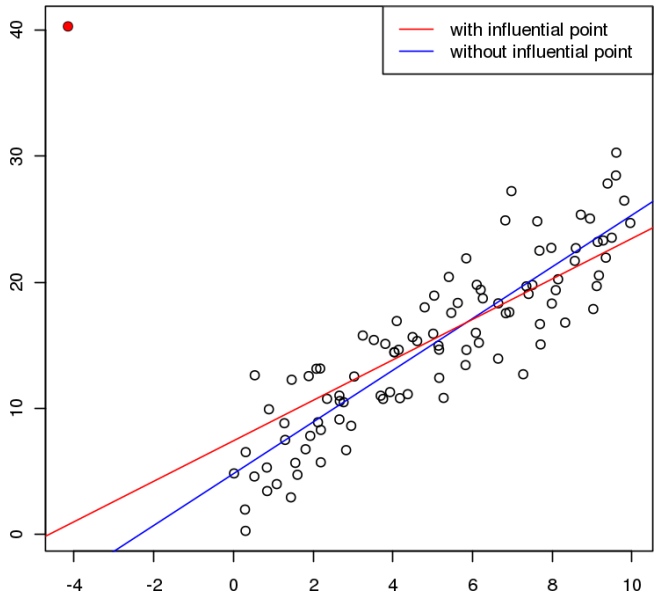


Figure 6: Illustrative example of the effect of an influential point. The red dot in the top left corner is an influential point – the slope of the regression line when it is included in the data (red) is quite different from the slope when it is not (blue).

relationship with the response variable (otherwise any model is bound to be fruitless), high correlations and/or dependencies among the explanatory variables is contraindicated as it introduces instability in the estimates of the regression coefficients are unstable. We can formally test for presence of multicollinearity using **variance inflation factors** (VIF); in its presence, data reduction and data transformation strategies might need to be implemented.

1.4.4 Data Reduction/Model Selection

In a good model, a balance has been struck between its **predictive ability** and its **simplicity**. Clients look for the **simplest model** that explains the behaviour of the response variable Y in a **reasonably adequate manner** (a version of *Occam's Razor*). If there are p predictor variables $X_1, \ldots, X_p$, then there are 2^p possible models from which to select the "best", ranging from the **simple average model** $y_i = \beta_0 + \varepsilon_i$ to the **full model** $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i$.

Step-wise Regression As the number of predictors p grows, it is not feasible to fit all 2^p possible models to determine the optimal model. **Step-wise regression** is an automated model selection procedure that builds a succession of models from which a choice can be made. There are numerous variants – this particular algorithm is called **forward selection**, for reasons that will shortly become clear (to fix the problem in conceptual space, assume that there are $p = 10$ predictor variables).

1. **Selecting the first variable:** Fit p simple linear regressions

$$y_i = \beta_0 + \beta_j x_{i,j} + \varepsilon_i, \quad j = 1, \ldots, p$$

and choose the model with highest R^2 value. In other words, select the variable X_j that best describes the behaviour of Y **on its own**. If X_5 turns out to be that variable, for instance,

then the tentative model is

$$y_i = \beta_0 + \beta_5 X_{i,5} + \varepsilon_i.$$

If this model is not statistically significant (tested at predetermined significance level α), then the final model selection is

$$y_i = \beta_0 + \varepsilon_i$$

and the search is complete. Otherwise, proceed to step 2.

2. **Selecting the second variable:** Fit all two-parameter regression models

$$y_i = \beta_0 + \beta_5 x_{i,5} + \beta_j x_{i,j} + \varepsilon_i, \quad j = 1, \ldots, p, \quad j \neq 5.$$

Select the model that has the highest value of the test statistic

$$t'_k = \sqrt{\frac{\text{MSR}(X_k|X_5)}{\text{MSE}(X_5, X_k)}}.$$

Say that $k = 3$ yields the largest such value. If the associated model's $p-$value is smaller than α, then our tentative model is updated to

$$y_i = \beta_0 + \beta_3 X_{i,3} + \beta_5 X_{i,5} + \varepsilon_i$$

and we proceed to step 3. Otherwise, the final model selection is

$$y_i = \beta_0 + \beta_5 X_{i,5} + \varepsilon_i$$

and the search is complete.

3. **All subsequent steps:** Repeat step 2 using

$$t''_k = \sqrt{\frac{\text{MSR}(X_k|X_5, X_3)}{\text{MSE}(X_5, X_3, X_k)}}$$

and so forth, until no additional term improves the model significantly.

In contrast to forward selection which starts with the simple average model

$$y_i = \beta_0 + \varepsilon_i$$

and build a nested sequence of increasingly complex models, **backward elimination** begins with the full model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i$$

and keeps removing terms until removal of *any* variable causes a significant loss of its predictive power (calculated using $t_k^{(\ell)}$). In general, forward selection and backward elimination will not select the same final model.

In the **combined approach**, the process starts from the simple average model as in forward selection, but each time a new variable is added to the tentative model, a backward elimination search is performed to test whether any of the previously added variables are no longer significant. This approach enables the model to be better tuned to the data and has been known to prevent **overfitting** (more on this in Section **??**). In either case, the step-wise selection methods are **expensive**, computing-wise.

The test statistic $t_k^{(\ell)}$ is the square root of the ratio of conditional MSR over MSE. In everyday terms, it is testing *whether the addition of X_k provides a significant improvement in predictive ability over the current tentative model's*. Other alternative include the **Akaike Information Criterion** (AIC), the **Bayesian Information Criteria** (BIC), **Mallow's C_p Criterion**, and the **R^2 criterion** – simply pick the model which optimises the desired criterion.

IMPORTANT NOTE: step-wise regression is **flawed** in many ways which we will not explore at the moment; in practice, it has slowly started being replaced by **regularisation methods** such as ridge regression and the LASSO (see Section **??**). From a consulting standpoint, this is a development over which it is worth trying to educate clients.

1.4.5 Basics of Multivariate Statistics

Up until this point, we have been considering situations the response has been **univariate**. In applications, especially those that require data science methods, the situation often calls for **multivariate** responses, where the response variables are thought to have some relationship (e.g. a **correlation structure**). It remains possible to analyse each response variable independently, but the dependence structure can be exploited to make **joint** (or simultaneous) inferences.

Properties of the Multivariate Normal Distribution The probability density function of a random vector $\mathbf{X} \in \mathbb{R}^p$ that follows a **multivariate normal distribution** with **mean vector $\boldsymbol{\mu}$** and **covariance matrix $\boldsymbol{\Sigma}$**, denoted by $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is given by

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})},$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p} \end{bmatrix}.$$

For such an $\boldsymbol{X}$, the following properties hold:

1. any linear combination of its components are normally distributed;
2. all subsets of components follow a (modified) multivariate normal distribution;
3. a diagonal covariance matrix implies the independence of its components;
4. conditional distributions of components follow a normal distribution, and
5. the quantity $(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$ follows a χ_p^2 distribution.

These properties make the multivariate normal distribution attractive, from a theoretical point of view (if not entirely realistic). For instance,

- using the first property, we can use **contrasts** to test which components are distinct from the others;
- the fifth property is the multivariate analogue of the square of a standard normal random variable $Z \sim \mathcal{N}(0,1)$ following a $Z^2 \sim \chi^2_1$ distribution;
- but two univariate normal random variables with zero covariance are not necessarily independent (the joint p.d.f. of two such variables is not necessarily the p.d.f. of a multivariate normal distribution).

A number of univariate approaches generalise nicely.

Hypothesis Testing for Mean Vectors When the sample comes from a univariate normal distribution, we can test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0$$

by using a $t-$statistic. Analogously, if the sample comes from a $p-$variate normal distribution, we can test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0} \quad \text{against} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$$

by using **Hotelling's T^2 test statistic**, mathematically defined as

$$T^2 = N \cdot (\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

where $\bar{\boldsymbol{X}}$ denotes the **sample mean** and $\boldsymbol{S}$ is the **sample covariance matrix**. Under H_0,

$$T^2 \sim \frac{(N-1)p}{(N-p)} F_{p,N-p}.$$

Thus, we do not reject H_0 at a significance level of α if

$$N \cdot (\bar{\boldsymbol{X}} - \boldsymbol{\mu_0})^\top \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu_0}) \leq \frac{(N-1)p}{(N-p)} F_{p,N-p}(\alpha) \tag{2}$$

and reject it otherwise.

Confidence Region and Simultaneous Confidence Intervals for Mean Vectors In the $p-$variate normal distribution, any $\boldsymbol{\mu}$ that satisfies the condition

$$N \cdot (\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq \frac{(N-1)p}{(N-p)} F_{p,N-p}(\alpha) \tag{3}$$

resides inside a $(1-\alpha)100\%$ **confidence region** (an ellipsoid in this case). **Simultaneous Bonferroni confidence intervals** with overall error rate α can also be derived, using

$$(\bar{x}_j - \mu_j) \pm t_{N-1}(\alpha/p) \sqrt{\frac{s_{j,j}}{N}} \text{ for } j = 1, \ldots, p$$

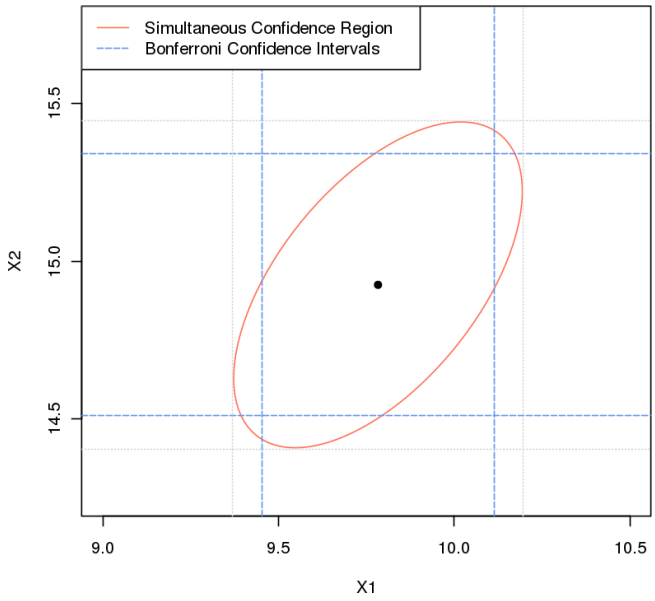


Figure 7: 95% confidence ellipse, Bonferroni and Hotelling's T^2 simulatenous confidence intervals for a bivariate normal random sample.

Another approach is to use **Hotelling's T^2 simultaneous confidence intervals**, given by

$$(\bar{x}_j - \mu_j) \pm \sqrt{\frac{p(N-1)}{N-p} F_{p,N-p}(\alpha)} \sqrt{\frac{s_{j,j}}{N}} \text{ for } j = 1, \ldots, p$$

Figure 7 shows these regions for a bivariate normal random sample. Notice that the Hotelling's T^2 simultaneous confidence intervals form a rectangle that confines the confidence region, while the Bonferroni confidence intervals are slightly narrower. Given that all the components of the mean vector are correlated (according to a generally non-diagonal covariance matrix), the confidence region should be used if the goal is to study the **plausibility of the mean vector as a whole**, while Bonferroni confidence intervals may be more suitable when **component-wise confidence intervals** are of interest.

1.4.6 Multivariate Analysis of Variance (MANOVA)

As shown in Section 1.4.2, ANOVA is often used in a first pass to determine whether the means from every sub-population are identical.

One-Way MANOVA ANOVA can test means from more than two populations; the **multivariate ANOVA** (MANOVA) is quite simply a multivariate extension of ANOVA which tests whether the mean vectors from all sub-populations are identical.

Let there be I sub-populations in the population, from each of which N_i $p-$dimensional responses are drawn, $i = 1, \ldots, I$. Mathematically, each observation can be expressed as:

$$\boldsymbol{X}_{i,j} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_{ij}$$

Source	SSP	d.f.
Treatment	$\boldsymbol{B} = \sum_{i=1}^{I} N_i(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})^\top(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})$	$I-1$
Error	$\boldsymbol{W} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)^\top(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)$	$\sum_{i=1}^{I} N_i - I$
Total	$\boldsymbol{B}+\boldsymbol{W} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\boldsymbol{X}_{i,j} - \bar{\boldsymbol{X}})^\top(\boldsymbol{X}_{i,j} - \bar{\boldsymbol{X}})$	$\sum_{i=1}^{I} N_i - 1$

Table 5: One-way MANOVA table; with I sub-populations.

1. Some HS or Less	2. HS	3. College/University	4. Post-Graduate or higher
16	55	83	46

Table 6: Respondents' educational achievements, from a (fictitious) 2017 survey.

where $\boldsymbol{\mu}$ is the **overall mean vector**, $\boldsymbol{\tau}_i$ is the i^{th} **population-specific treatment effect**, and $\boldsymbol{\varepsilon}_{ij}$ is the **random error**, which follows a $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. It is important to note that the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for each sub-population, and that

$$\sum_{i=1}^{I} N_i \boldsymbol{\tau}_i = \boldsymbol{0}$$

to ensure that the estimates are uniquely identifiable.

To test the hypothesis

$$H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_I = \boldsymbol{0} \quad \text{against} \quad H_1 : \text{at least one of } \boldsymbol{\tau}_i \neq \boldsymbol{0},$$

we decompose the total sum of squares and cross-products SSP_{tot} into

$$\text{SSP}_{\text{tot}} = \text{SSP}_{\text{treat}} + \text{SSP}_{\text{e}}.$$

Based on this decomposition, we compute the test statistic known as **Wilk's lambda**

$$\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|}$$

and reject H_0 if Λ^* is too small.

1.4.7 Goodness-of-Fit Tests

A (fictitious) 2017 survey asked a sample of $N = 200$ adults between the age of 25 to 35 about their highest educational achievement. The result is summarised in Table 6. In 1997, it was found that $p_1 = 13\%$ of adults had not complete high school, $p_2 = 32\%$ had obtained a high school degree but not a post-secondary degree, $p_3 = 37\%$ had either an undergraduate college or university diploma but no post-graduate degree, and $p_4 = 18\%$ had at least one post-graduate degree. Based on the result of this survey, is there sufficient evidence to believe that educational backgrounds of the population have changed since 1997?

Since each respondent's educational achievement can only be classified into one of these categories, they are **mutually exclusive**. Furthermore, since these categories cover all possibilities

Category	O_j	$p_{j,0}$	$m_{j,0}$	$(O_j - m_{j,0})^2/m_{j,0}$
1	16	0.13	26	3.846
2	55	0.32	64	1.266
3	83	0.37	74	1.095
4	46	0.18	36	2.778
Total	200	1	200	7.815

Table 7: Summary table for goodness-of-fit data for educational achievements under H_0.

on the educational front, they are also **exhaustive**. We can thus view the distribution of educational achievements as being **multinomial**. For such a distribution, with parameters $p_1, \cdots, p_k$, the expected frequency in each category is $m_j = Np_j$.

Let O_j denote the observed frequency for the j^{th} category. If there has been no real change since 1997, we would expect the sum of squared differences between the observed 2017 frequencies and the expected frequencies based on 1997 data to be small. We can use this information to test the **goodness-of-fit** between the observations and the expected frequencies *via* Pearson's χ^2 test statistic

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - m_j)^2}{m_j}$$

which follows a χ^2 distribution with $k-1$ degrees of freedom.

In the above example, the hypotheses of interest are

$$H_0 : p_1 = 0.13, p_2 = 0.32, p_3 = 0.37, p_4 = 0.18 \quad \text{against} \quad H_1 : \text{not } H_0.$$

Table 7 summarises the information under H_0. Pearson's test statistic is $X^2 = 7.815$, with an associated $p-$value of 0.0295, which implies that there is enough statistical evidence (at the $\alpha = 0.05$ level) to accept that the population's educational achievements have changed over the last 20 years.

1.4.8 Paired Comparisons and Analysis of Covariance (ANCOVA)

In Section 1.4.1, we looked at the effectiveness of new teaching method by assigning each group to a specific treatment and comparing the mean test scores. A crucial assumption for that model is that subjects in each group have **similar background knowledge** about statistics prior to the three week lectures. If this assumption is wrong, however, we may be making incorrect decisions based on the model. Even if each group had similar background knowledge *on average*, there may be large variability from person-to-person, masking the true treatment effect.

Paired Comparison One way to avoid such **subject-to-subject variability** is to administer both treatments to each individual, and then compare treatment effects by looking at the **difference in the outcomes**. If a grocery chain is interested in measuring the effectiveness of two advertising campaigns, for instance, it is reasonable to assume that there is a large variability in total sales, as well as popular items sold, at each store – it may then be preferable to run both campaigns in

each store and analyse the resulting data rather than to split the stores into two groups (in each of which a different advertising campaign is run) and then to compare the mean outcomes in the two groups.

Formally, let $X_{i,1}$ denote the total sales with campaign A and $X_{i,2}$ the total sales with campaign B. The quantity of interest is $D_i = X_{i,1} - X_{i,2}$ for each store $i = 1, \ldots, N$. Assuming that the differences D_i follow an i.i.d. normal distribution with mean δ and variance σ_d^2, then we can test

$$H_0 : \delta = 0 \quad \text{against} \quad H_1 : \delta \neq 0$$

by using the test statistic

$$t_0 = \sqrt{N}\frac{\bar{D}}{s_d},$$

which follows a Student's t distribution with $N-1$ degrees of freedom; thus we reject H_0 if the observed test statistic t_0 has p-value less than the pre-specified significance level $\alpha/2$.

Analysis of Covariance (ANCOVA) ANOVA compares multiple group means and tests whether any of the group means differ from the rest, by breaking down the total variability into a treatment (explainable) variability component and an error (unexplained) variability component, and building a ratio F_0 to determine whether or not to reject H_0.

Analysis of covariance (ANCOVA) introduces **concomitant variables** (or **covariates**) to the ANOVA model, splitting the total variability into 3 components: SS_{treat}, SS_{con}, and SS_e, aiming to reduce error variability. The choice of covariates is thus crucial in running a successful ANCOVA.

In order to be useful, a concomitant variable must be related to response variable in some way, otherwise it not only fails to reduce error variability, but it also increases the model complexity. In the teaching method example, we could consider administering a pre-study test to measure the prior knowledge level of each participant and use this score as a concomitant variable. In the advertising campaign example, we could have used the previous month's sales as a covariate. In medical studies, we could use the age and weight of subjects as covariates.

But concomitant variables should not be affected by treatments. Suppose that, in a medical study, patients were asked *how strongly they believed that they were given actual medication rather than a placebo*. If the treatment is indeed effective, then a participant's response to this question could be **markedly different** in the treatment group than in the placebo group (perhaps the medication has strong side-effects which cannot be ignored). This means that true treatment effect may be masked by concomitant variable due to unequal effects on treatment groups.

Qualitative covariates (such as gender, say) are not part of the ANCOVA framework. Indeed, such a covariate just creates new ANOVA treatment groups.

Figure 8 shows a potential breakdown of the total variability when moving from an ANOVA to an ANCOVA model – the error variability is further split into an error and a covariate component, while the treatment variability remains unchanged.

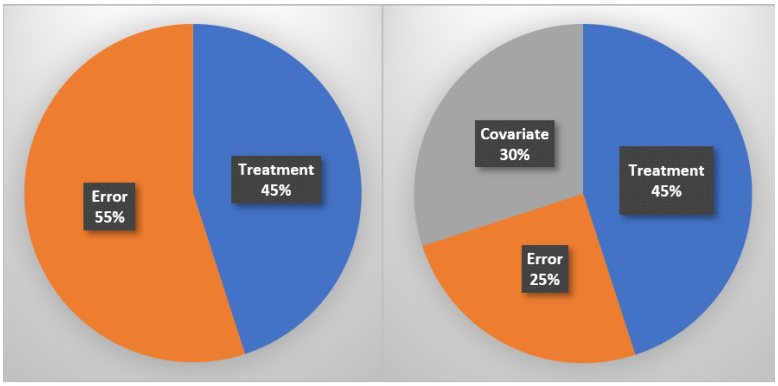


Figure 8: Breakdown of variability for ANOVA and ANCOVA.

ANCOVA Model and Its Assumptions Suppose that we are testing the effect of p treatments, with N_j subjects in each group. Then the ANCOVA model takes the form

$$y_{i,j} = \mu + \tau_j + \gamma(x_{i,j} - \bar{x}) + \varepsilon_{i,j} \tag{4}$$

where

- $y_{i,j}$ is the response of the i^{th} subject in the j^{th} treatment group;
- μ is the overall mean;
- τ_j is the j^{th} treatment effect subject to a constraint $\sum_{j=1}^{p} \tau_j = 0$;
- γ is the coefficient for the **covariate effect**;
- $(x_{i,j} - \bar{x})$ is the covariate value of the i^{th} subject in the j^{th} treatment group, adjusted by the mean, and
- $\varepsilon_{i,j}$ is the error of i^{th} subject in the j^{th} treatment group.

Four assumptions must be satisfied:

- **independence and normality of residuals** – the residuals are thought to follow an $i.i.d.$ normal distribution with mean of 0 and variance σ_ε^2;
- **homogeneity of residual variances** – the variance of the residuals must be uniform across treatment groups;
- **homogeneity of regression slopes** – the regression effect (slope) must be uniform across treatment groups, and
- **linearity of regression** – the regression relationship between the response and the covariate must be linear.

The first of these assumptions can be tested with the help of a QQ-plot and a scatter-plot of residuals vs. fitted values, while the second may use the Bartlett or the Levene test. The final assumption is not as crucial as the other three assumptions. Various remedial methods can be applied should any of these assumptions fail.

The third assumption, however, is **crucial** to the ANCOVA model; it can be tested with the **equal slope test**, which requires an ANCOVA regression on equation (4) with an additional interaction term $x \times \tau$. If the interaction is not significant, the third assumption is satisfied. In the event that the interaction term is statistically significant, a different approach (e.g. moderated regression analysis, mediation analysis) is required since using the original ANCOVA model is not prescribed. An in-depth application of an ANCOVA model is highlighted in Section 1.4.11.

Dose levels (d)	0	7000	15000	30000
Sample size (n)	50	35	65	50
Number of observed adverse effect (y)	3	6	33	39
Rate of observed adverse effect (p)	0.06	0.17	0.51	0.78

Table 8: Summary of experimental results involving C.I. Acid Red 114; $N = 200$.

1.4.9 Nonlinear Regression

From the use of tooth paste, cosmetics, cleaning solutions and so forth, we are exposed to numerous chemicals on a daily basis; thousands of new chemicals are introduced into commercial products each year, and government agencies (such as Health Canada and the Environmental Protection Agency in the U.S.) must determine whether these chemicals are safe for humans, animals, and the environment.

To test whether a chemical poses a risk of adverse effects, we must first determine whether it triggers adverse effects over a range of potential exposure levels, and if so, how much is considered safe (or how much would pose an unacceptable risk). Traditionally (and not necessarily ethically), rodents were used to study whether a chemical is carcinogenic or not.

Suppose that N laboratory rodents are divided into k groups, with each group consisting of N_i rodents. Over the course of the experiment, each group was given a certain amount of exposure to the chemical under investigation. The outcome of the experiment is whether each rodent eventually develops a tumour or not; that is, the outcome is expressed as 0 (tumour absent) or 1 (tumour present). Table 8 summarises the outcome of an experiment.

Clearly, we cannot fit an ordinary linear regression to the data as the outcome is **dichotomous**. How could we model the relationship between the adverse effect and the dose levels?

For each dose level d, the probability of adverse effect is $p_d = P(y = 1|d)$. The **conditional expectation** given the dose level is also $E(y = 1|d) = p_d$. Since the relationship resembles an $S-$shaped curve, we may use a logistic distribution to model the data:

$$E(y = 1|d) = p_d = \frac{\exp[\beta_0 + \beta_1 d]}{1 + \exp[\beta_0 + \beta_1 d]}$$

To obtain **maximum likelihood estimates** for β_0 and β_1, we need to rely on numerical methods such as the **Newton-Raphson method**; the dose-response model for the above example is shown in Figure 9.

Relationship to Linear Regression Since p_d is a probability, it has to lie in $[0, 1]$. By taking the odds of having an adverse effect, defined by $\omega_d = p_d/(1 - p_d)$, the boundary of the response is changed to $[0, \infty)$. Taking the log odds will span $\mathbb{R}$, and the functional form of the **logistic regression model** is

$$\log(\omega_d) = \log\left(\frac{p_d}{1 - p_d}\right) = \beta_0 + \beta_1 d, \tag{5}$$

which is a simple linear regression model.

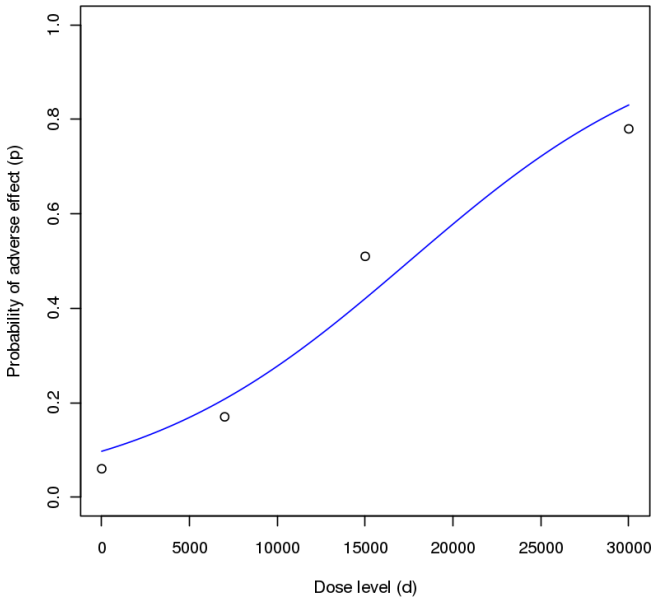


Figure 9: Dose-response model for C.I. Acid Red 114 using logistic regression.

Other non-linear regression models Other **sigmoidal curves** can be used to model the relationship between predictors and a binary response variable. Popular alternatives include the **probit** link $P(y|x) = \Phi(\beta_0 + \beta_1 x)$, where Φ is the cumulative distribution function of the standard normal distribution, or the **complementary log-log** link $P(y|x) = 1 - \exp(-\exp(\beta_0 + \beta_1 x))$. In toxicology studies, one of the most widely used model is called the **Hill**, and it is defined *via*

$$P(y|d, \alpha, \kappa, \eta) = \alpha + (1-\alpha)\frac{d^\eta}{d^\eta + \kappa^\eta};$$

part of its appeal to health scientists is the interpretation of its parameters – α represents the **background rate for adverse effect**, while κ denotes ED_{50} (the **effective dose at which** 50% **of participants would exhibit the response of interest**) and η provides the **steepness of the dose-response curve**.

Figure 10 compares the simple logistic model to the Hill model; we observe that the Hill model provides a closer fit to the observed proportions, and the curvature is more pronounced compared to the logistic model.

1.4.10 Bayesian Statistics

In classical statistics, model parameters such as μ and σ are treated as constants; **Bayesian statistics**, on the other hand assume that **model parameters are random variables**. As the name implies, Bayes' Theorem lies at the foundation of Bayesian statistics:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}, \tag{6}$$

where H represents the hypothesis and D denotes the observed data, which is sometimes written in shorthand as **posterior** $= P(H|D) \propto P(D|H) \times P(H) =$ **evidence** $\times$ **prior**. In other words, our degree of belief in a hypothesis should be updated by the evidence provided by data.

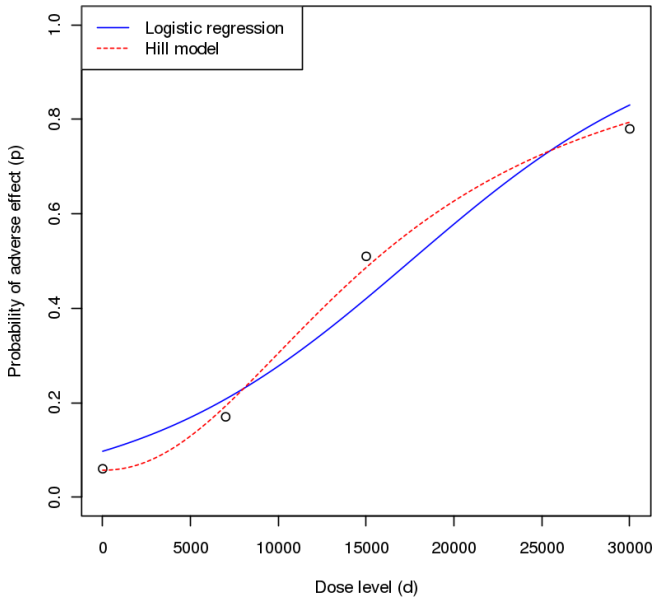


Figure 10: Dose-response model for C.I. Acid Red 114 using logistic regression (blue) and the Hill model (red).

IMPORTANT NOTE: the use of Bayesian statistics is controversial in many quarters, and your clients (or fellow consultants) might have strong **frequentist** leanings. Navigate with care.

Bayes' Theorem escapes the controversy – nobody disputes its validity – and has proven to be a useful component in various models and algorithms, such as email spam filters, and the following example.

Suppose we are interested in diagnosing whether a tumour is begin or malignant, based on several measurements obtained from video imaging. Bayes' Theorem (6) can be recast in a tumour data mould:

- **posterior:** $P(H|D) =$ based on collected data, how likely is a given tumour to be benign (or malignant)?
- **prior:** $P(H) =$ in what proportion are tumours benign (or malignant) in general?
- **likelihood:** $P(D|H) =$ knowing a tumour is benign (or malignant), how likely is it that these particular measurements would have been observed?
- **evidence:** $P(D) =$ regardless of a tumour being benign or malignant, what is the chance that a tumour has the observed characteristics?

To answer the above question (that is, to compute the posterior), we will use a **naïve Bayes classifier** (see Section **??** for other classification methods).

Naïve Bayes Classification for Tumour Diagnoses

1. **Objective function:** a simple way to determine whether a tumour is benign or malignant is to compare **posterior probabilities** and choose the one with highest probability. That is,

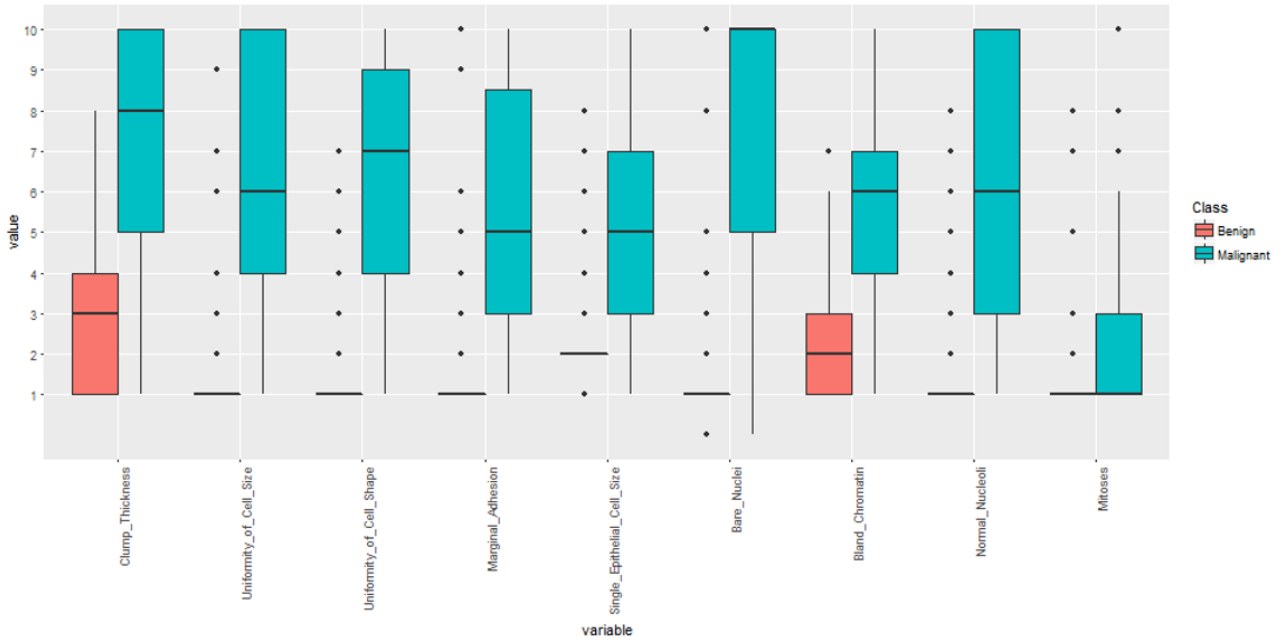


Figure 11: Boxplot visualisation of measurements for benign and malignant tumours.

Obs.	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	3	1	1	1	2	2	3	1	1

Table 9: Scores for an undiagnosed tumour.

we diagnose a tumour as **malignant** if

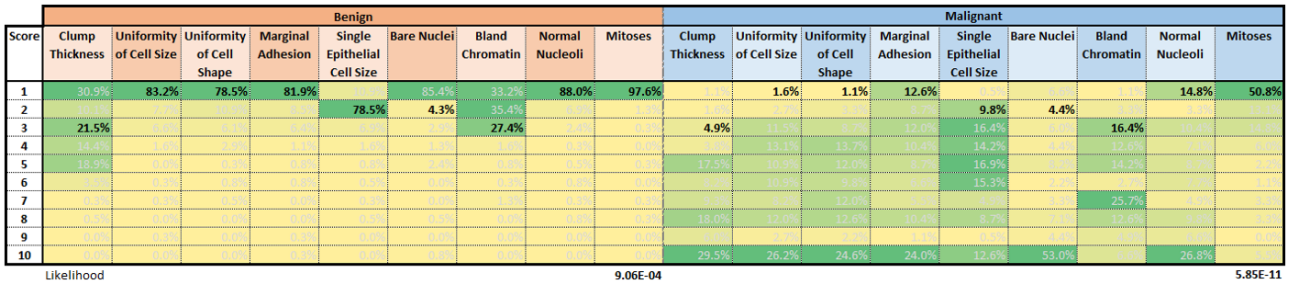
$$\frac{P(\text{malignant}|D)}{P(\text{benign}|D)} = \frac{P(D|\text{malignant}) \times P(\text{malignant})}{P(D|\text{benign}) \times P(\text{benign})} > 1,$$

and as **benign** otherwise.

2. **Dataset:** the classifier is built on a sample of $N = 458$ tumours with nine measurements, each scored on a scale of 1 to 10. The measurements include items such as *clump thickness* and *bare nuclei*; boxplots of these measurements are shown in Figure 11. We also have one undiagnosed case with these measurements, with its explanatory signature scores given in Table 9; this is the observation for which a prediction is required.
3. **Assumptions:** we assume that the scores of each measurement are independent of one another (hence the *naive* qualifier); this assumption simplifies the likelihood function to

$$P(H|D) = P(H|x_1, x_2, \cdots, x_9) = P(H|x_1) \times \cdots \times P(H|x_9).$$

4. **Prior distribution:** we can ask subject matter experts to provide a rough estimate for the general ratio of benign to malignant tumours, or use the proportion of benign tumours in the sample as our prior. In situations where we have no knowledge about this distribution, we may simply assume a **non-informative prior** (in this case, the prevalence rates being the same for both responses).



	Benign									Malignant								
Score	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	30.9%	83.2%	78.5%	81.9%	35.7%	85.4%	33.2%	88.0%	97.6%	1.1%	1.6%	1.1%	12.6%	0.5%	6.8%	1.1%	14.8%	50.8%
2	69.1%	16.8%	21.5%	18.1%	78.5%	4.3%	39.4%	8.0%	2.4%	4.9%	4.9%	4.9%	4.9%	9.8%	4.4%	4.9%	4.9%	4.9%
3	21.5%	0.0%	0.0%	0.0%	0.0%	0.0%	27.4%	0.0%	0.0%	4.9%	4.9%	4.9%	10.0%	16.4%	0.0%	16.4%	4.9%	4.9%
4	14.4%	1.8%	2.9%	1.1%	1.6%	1.3%	1.6%	0.3%	0.0%	4.9%	13.1%	13.7%	10.4%	14.2%	4.4%	12.6%	4.9%	4.9%
5	13.2%	0.0%	0.0%	0.0%	2.4%	2.4%	0.0%	0.0%	0.0%	17.3%	10.0%	12.0%	9.8%	16.9%	0.0%	14.8%	4.9%	4.9%
6	9.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	17.3%	10.0%	9.8%	9.8%	15.3%	0.0%	4.9%	4.9%	4.9%
7	6.3%	0.3%	0.3%	0.0%	0.3%	0.0%	1.3%	0.3%	0.0%	17.3%	9.8%	12.0%	9.8%	9.8%	1.3%	25.7%	4.9%	4.9%
8	0.3%	0.0%	0.0%	0.0%	0.3%	0.3%	0.0%	0.0%	0.0%	18.0%	12.0%	12.6%	10.4%	8.7%	0.0%	4.9%	4.9%	4.9%
9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.9%	4.9%	4.9%	4.9%	4.9%	0.0%	4.9%	4.9%	4.9%
10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	29.5%	26.2%	24.6%	24.0%	12.6%	53.0%	4.9%	26.8%	4.9%
Likelihood									9.06E-04									5.85E-11

Figure 12: Multinomial probabilities for benign and malignant tumours.

Class	Prior	Likelihood	Posterior	Ratio
Malignant	0.327	5.85×10^{-11}	1.92×10^{-11}	3.15×10^{-8}
Benign	0.673	9.06×10^{-4}	6.09×10^{-4}	

Table 10: Computation of posterior probabilities in the undiagnosed case.

5. **Computation of likelihoods:** under independence, each measurement is assumed to follow a multinomial distribution (since scores are on scale from 1 to 10). Multiplying probabilities from each multinomial distribution (one each for both classes) provides the overall likelihoods for benign and malignant tumours, respectively. The likelihood of the undiagnosed case being a benign tumour is given to be 9.06×10^{-4}, while the likelihood of being a malignant tumour is 5.85×10^{-11}, based on the multinomial probabilities given in Table 12
6. **Computation of Posterior:** Multiplying the prior probability and likelihood, we get a quqntity that is proportional to the respective posterior probabilities. Looking at Table 10, we conclude that the tumour in the undiagnosed case is**likely benign** (note that we have no measurement on how much more likely it is to be benign than to be malignant – the classifier is **not calibrated**).

1.4.11 Case Study: Covariance Analysis of the Effect of a Probiotic Agent on IBS

Irritable Bowel Syndrome (IBS) is a functional colonic disease with high prevalence. Typical symptoms include "chronic abdominal pain, discomfort, bloating, and alteration of bowel habits" [Wikipedia]; it has been linked to chronic pain, fatigue, and work absenteeism and is considered to have a severe impact on quality of life [Paré *et al.* (2006), Maxion-Bergemann *et al.* (2006)]. Although there is no known cure for IBS, there are treatments that attempt to relieve symptoms, including dietary adjustments, medication and psychological interventions.

In 2010, the *Canadian College of Naturopathic Medicine* (CCNM) was commissioned to conduct a study to investigate the effect of a probiotic agent on IBS. The study's details and a preliminary data analysis using **hierarchical linear models** (HLM) can be found in a preliminary report – it's key findings are that a strong placebo/expectation effect is present in the early stages of the study (which is not entirely surprising given the nature of the phenomenon under study), and that there is no strong statistical evidence to suspect that the agent itself has much of an effect on mild to moderate IBS [Herman, Cooley, Seely (2011)].

The sponsor has expressed interest in determining whether these findings still hold when the trial data is examined using **analysis of covariance** (ANCOVA), a general linear model which

evaluates whether the population means of a dependent/response variables (in this case, *IBS Severity* or a measure of *Quality of Life* (QoL)) are equal across levels of a categorical independent variable (in this case, two treatment effects over time), while statistically controlling for the effects of covariates (in this case, the baseline scores for IBSS and QoL). By comparison with the more traditional analysis of variance (ANOVA), ANCOVA can be used to increase the likelihood of finding a significant difference between treatment groups (when one exists) by reducing the within-group error variance.

While some of the results looked promising (in particular for severe IBS sufferers), no statistical evidence for treatment effect was found at the 95% significance level; furthermore, even had evidence been found at that level, design and recruitment issues would have called their practical significance into question.

In 2013, CCNM conducted a second study to investigate the effect of the probiotic agent, this time focusing on severe IBS. The results are provided in the report "Covariance Analysis of IBS Study II".

References

[1] Sahai, H., Ageel, M.I. [2000], The Analysis of Variance: Fixed, Random and Mixed Models, Birkhäuser.
[2] Kutner, M.H., Nachtsheim, C.J., Neter, J, Li. W. Ageel, M.I. [2004], Applied Linear Statistical Models, 5th ed., McGraw-Hill-Irwin.
[3] Hollander, M., Wolfe, D.A. [1999], Nonparametric Statistical Methods, 2nd ed., Wiley.
[4] Bruce, P., Bruce, A. [2017], Practical Statistics for Data Scientists: 50 Essential Concepts, O'Reilly.
[5] Sivia, D.S., Skilling, J. [2006], Data Analysis: a Bayesian Tutorial, 2nd ed., Oxford.
[6] Rizzo, M.L. [2007], Statistical Computing with R, CRC Press.
[7] Reinhart, A. [2015], Statistics Done Wrong: the Woefully Complete Guide, No Starch Press.
[8] Hogg, R., McKean, J., Craig, A., [2005], Introduction to Mathematical Statistics, 6th ed., Pearson.
[9] Johnson, R., Wichern, D., [2007], Applied Multivariate Statistical Analysis, 6th ed., Pearson.
[10] Montegomery, D., [2009], Introduction to Mathematical Statistics, 7th ed., Wiley.
[11] **https://en.wikipedia.org/wiki/Naive_Bayes_classifier**
[12] **https://en.wikipedia.org/wiki/Hill_equation_(biochemistry)**

Covariance Analysis of Irritable Bowel Syndrome Study II

by **Shintaro Hagiwara**, M.Sc,, and **Patrick Boily**, Ph.D.
Centre for Quantitative Analysis and Decision Support
Carleton University
cqads@carleton.ca

presented to the Canadian College of Naturopathic Medicine

October 2014

This report presents key findings of the covariance analysis that was performed to test the effect of the probiotic agent on the severe sufferers of Irritable Bowel Syndrome (IBS).

It relies in many ways on work previously done by CQADS; as such, large chunks of this report follow the structure and content of "**Covariance Analysis for the 2010 CCNM Pilot Study on Irritable Bowel Syndrome**" [9], a report produced by CQADS in August of 2013.

Background and Executive Summary

Irritable Bowel Syndrome (IBS) is a functional colonic disease with high prevalence. Typical symptoms include "chronic abdominal pain, discomfort, bloating, and alteration of bowel habits" [1]; it has been linked to chronic pain, fatigue, and work absenteeism and is considered to have a severe impact on quality of life [2, 3]. Although there is no known cure for IBS, there are treatments that attempt to relieve symptoms, including dietary adjustments, medication and psychological interventions.

In 2010, the Canadian College of Naturopathic Medicine (CCNM) was commissioned to conduct a pilot study to investigate the effect of a probiotic agent on IBS. The study's details and a preliminary data analysis using hierarchical linear models (HLM) can be found in a preliminary report: it's key findings are that a strong placebo/expectation effect is present in the early stages of the study which is not entirely surprising given the nature of the phenomenon under study, and that there is no strong statistical evidence to suspect that the agent itself has much of an effect on mild to moderate IBS [4]. Furthermore, the key findings from covariance analyses (ANCOVA) on the above data conducted by the Centre for Quantitative Analysis and Decision Support (CQADS) aligned with the analysis using HLM [4,9]; the main ANCOVA results are summarized in the table below.

ANCOVAs for IBS and QoL measures (original dataset)			Sample Size	Initial		At 3 months		p-value
				mean	SD	mean	SD	
All subjects	IBS severity	Placebo	57	273.8	73.7	204.0	97.2	0.095 (0.137†)
		Probiotics	59	268.9	76.4	175.3	78.6	
	QoL	Placebo	58	42.0	20.4	33.4	21.0	**0.056**
		Probiotics	59	40.2	18.6	26.4	17.5	
Severe subjects*	IBS severity*	Placebo	16	363.0	57.9	281.4	121.4	(0.049†)
		Probiotics	19	351.0	44.0	206.3	104.5	
	QoL*	Placebo	17	55.8	21.6	50.6	21.8	0.007
		Probiotics	19	48.3	16.1	29.9	18.0	

Due to the small sample size (and because of issues associated with positively determining membership in the severe sufferer category), the analyses marked with a "*" were not endorsed by CQADS, and are provided for completeness. The significance of the treatment is measured by the *p*-value (*p*-values obtained after analysis on the reduced dataset, for which outliers have been removed, are indicated by a "†").

While some of the results looked promising, no statistical evidence for treatment effect was found at the 95% significance level; furthermore, even had evidence been found at that level, design and recruitment issues would have called their practical significance into question [9].

In 2013, CCNM conducted a second study to investigate the effect of a probiotic agent, this time focusing on severe IBS sufferers. Potential participants were considered to be severe IBS sufferers if they had total IBS severity scores of 300 or higher, with the highest possible score being 500. The study sponsor has expressed interest in analyzing this new data using Analysis of Covariance (ANCOVA) in order to determine whether there is a statistically significant difference between the placebo and the probiotic agent.

ANCOVA is a general linear model which evaluates whether the population means of a dependent/response variable (in this case, total IBS severity score, five IBS sub-scores, and a measure of Quality of Life) are equal across levels of a categorical independent variable (in this case, two treatment effects over time), while statistically controlling for the effects of covariates (in this case, the baseline scores). By comparison with the more traditional analysis of variance (ANOVA), ANCOVA can be used to increase the likelihood of finding a significant difference between treatment groups (when one exists) by reducing the within-group error variance.

The main results of the 7 ANCOVAs (for the new data, imputed with Last Observation Carried Forward, see next page) and the 5 IBS sub-scores ANCOVAs (for the original data, imputed with LOCF, below). Detailed explanations are found in the body of the report.

ANCOVA for the 5 IBS sub-scores (original dataset)		Group	Sample Size	Initial		At 3 months		p-value
				mean	SD	mean	SD	
All subjects	Abdominal pain	Placebo	57	45.26	23.50	30.68	24.51	0.106
		Probiotics	59	43.95	22.79	23.49	21.41	
	Abdominal distension	Placebo	57	51.28	22.93	34.18	26.48	0.445
		Probiotics	59	48.35	25.28	30.19	22.25	
	Satisfaction	Placebo	57	67.79	20.89	56.95	23.40	**0.085**
		Probiotics	59	69.60	23.53	50.42	21.38	
	Interference	Placebo	57	65.81	18.63	47.63	21.16	0.158
		Probiotics	59	59.67	18.13	40.14	20.07	
	Frequency	Placebo	57	43.68	24.32	34.56	27.37	0.347
		Probiotics	59	47.37	28.26	31.04	28.78	

As shown in these tables, the ANCOVA of the two clinical trials to study the effect of the probiotic agent on IBS do not reveal a statistically significant treatment effect. That being said, even though we conclude that there is no evidence to differentiate the treatment effect from the placebo effect, there were some instances when the difference in improvements between the two treatment groups (Probiotics over Placebo in the first study, I over K in the second) were nearly significant (e.g., patients' satisfaction with their bowel movement habits in the first study, and their quality of life in both studies, with *p*-values reaching 0.085, 0.056 and 0.061, respectively).

While the *p*-values themselves may look encouraging, the large placebo effect and high fluctuating nature of IBS on a day-to-day basis make it very difficult to control for the uncertainty in the data. Furthermore, it is far from obvious that these results can be generalized to a larger population due to the non-probabilistic nature of samples collected for the clinical trials, as well as the possibility of a self-reporting bias.

ANCOVA for the 7 core analyses (new dataset)		Group	Sample Size	Initial		End (at 3 month)		p-value
				mean	SD	mean	SD	
All subjects	Total IBS severity	I	45	350.41	42.91	265.75	100.62	0.310
		K	42	351.82	53.83	245.10	106.21	
	Abdominal pain	I	45	61.92	17.52	43.30	23.08	0.603
		K	42	64.56	17.64	39.96	26.18	
	Satisfaction	I	45	82.74	15.43	65.54	22.13	0.330
		K	42	76.58	16.79	57.61	23.55	
	Interference	I	45	74.41	13.97	56.22	22.60	0.327
		K	42	75.38	15.05	56.42	23.01	
	Frequency	I	45	62.89	23.22	52.22	32.11	0.358
		K	42	62.98	23.58	45.95	31.00	
	Abdominal distension	I	45	68.44	16.91	48.46	25.77	0.902
		K	42	72.32	16.26	45.17	27.88	
	QOL	I	43	52.91	18.52	40.43	23.33	**0.061**
		K	41	52.59	15.63	47.66	20.35	

1. Understanding the Structure of the Data

1.1 Recruitment

100 participants were recruited for the study, where 50 of which were assigned to group K, and 50 to the group I: one of these groups represent the active treatment group, while the other group is administered a placebo treatment (CQADS analysts do not know which label corresponds to which group).

The objective of this study is to examine the effect of the treatment against the (placebo) control group on severe IBS patients. It should be noted that there were 16 participants who were not classified as a severe IBS sufferer according to their pre-treatment total IBS severity scores. Participant ID 68, who had a severity score of 158, was discarded from the study; however, 15 patients whose baseline IBS severity scores ranging from 259.6 to 298 were kept for this study as the severity of IBS is known to fluctuate rather frequently.

1.2 Randomization

In order to facilitate a balanced representation in the active treatment group and the placebo group in terms of their demographical characteristics, participants were first categorized by their gender group (M/F) and age group (< or ≧50 years). Within each subgroup, participants were then randomly assigned to the treatment group or the placebo group, in a double-blind fashion (i.e. neither the examiners nor the participants were aware of the groups to which they had been assigned). As the number of treatment/placebo assignments in each group was not intended to be even, this randomization process leads us to (Unbalanced) Randomized Complete Block Design.

1.3 Outcome Measures

The two main response variables under considerations are the total IBS severity score and the IBS Quality of Life (QoL) measure. Furthermore, we will be examining the effect of treatment on each of the five questions that constitute the total IBS severity score. These questions measure the levels of abdominal pain, abdominal distension and bloating, satisfaction, interference, and frequency. All scores are collected at the beginning of the study (baseline) and at one-month intervals for three months. As a side note, all of these response variables are computed using self-reported data.

1.4 Drop-outs, Missing Observations, and Imputation

Eight participants did not deliver any information after the baseline measure: four participants from the group K and four from group I. As there was no information regarding the treatment effects for those participants, they were eliminated from the remaining analysis. Furthermore, six participants failed to follow-up after the first or the second month of the study. **Table 1** summarizes the breakdown of those participants.

Table 1 – IBSS drop-out data. Only those participants that remain after the first two months are retained

	Total # of recruited participants	Dropped out after Baseline	Dropped out after Month 1	Dropped out after Month 2	Remaining after Month 3
Treatment K	49	4	3	2	40 (81.6%)
Treatment I	50	4	1	0	45 (90.0%)
Total	**99**	**8**	**4**	**2**	**85 (85.8%)**

Since the covariance analysis requires the dataset to be free of missing observations, imputations must be performed prior to proceeding with the analysis.

In general, it is difficult to study the exact reasons why some participants terminate the follow-up prematurely; however it could be conjectured that participants who complete the study are either more likely to believe in the effect of the active agent or to actually be feeling the effect of the treatment than those who fail to complete the treatment. In fact, taking a look at drop-outs with partial information, it is often the case that these observations do not follow the general downward trend seen in the participants with the complete information. In an attempt to test this conjecture, partial non-respondents should be kept in the analysis.

Therefore, for those participants with recorded observations up to the second follow-up, the Last Observation Carried Forward (LOCF) imputation was favoured over the regression imputation [5], and implemented for the analysis. However, it should be noted that four participants dropped out of the study after the first follow-up. Due to the observed month-to-month fluctuation in the scores within each patient, it may not be reasonable to assume that the IBS severity scores and QoL measures for these participants stay constant over a two month period. Therefore, the decision was made to eliminate these participants from subsequent analysis.

To compensate for the fact that the imputation was done prior to the covariance analysis, one degree of freedom is docked for each imputation. Note that only the missing observations at the third month into the study are imputed, as we are interested in comparing the baseline measures and the final measures.

For the IBS severity score and its five sub-scores, there were no partial non-respondent; however, subjects 19, 22, and 32 did not complete some questions on the QoL questionnaire at the baseline. For this reason, these participants are removed from the covariance analysis for the QoL scores. **Table 2** summarizes the participants who dropped out prior to completion of the study and who were kept for the analysis with imputed scores.

Table 2 – Number of participants used in covariance analyses for IBS severity measure and QoL measure

	IBS		QoL	
Treatment group	**K**	**I**	**K**	**I**
Removed	7	5	8	7
Completed (+ imputed)	40 (42)	45	39 (41)	43
Total (Recruited)	49	50	49	50

1.5 Outlier Detection

Outlying observations frequently have a dramatic effect on the fitted values of the selected model; should such extreme points be found in the dataset, they need to be studied carefully in order to determine whether they should be retained or removed [6]. If influential observations are identified, remedial measures may need to be applied in order to minimize their undue effects.

Given that we have at most four data points per participant, and due to the large observed within-participant variability over time, it is near impossible to identify within-participant observations which we could deemed to be "extreme". It is, however, significantly easier to identify any abnormal between-participant observations.

Numerous methods can be used to find outliers; none of them are foolproof and good judgement must be used. For this reason, the box-and-whisker plots can help in the search for possible outliers: data points falling below $Q_1 - 1.5 \cdot \text{IQR}$ or above $Q_3 + 1.5 \cdot \text{IQR}$, (where Q_1, Q_3, IQR are the first quartile, the third quartile and the inter-quartile range, respectively)

require a more in-depth analysis (see **Figure 1**, on page 6). From the box-and-whisker plots, we observe that medians for treatment groups I and K usually do not differ greatly at the third follow-up. Furthermore, the variability of the data (given by the range of the whisker) tends to be greater at the last follow-up compared to the variability observed at the pre-treatment assessment.

2. Model Selection

As mentioned in **Section 1.2**, the participants were stratified according to their gender (M/F) and age group (< or ≥50 years), and then randomized within each block in an effort to promote balanced representation between two treatment groups. From a statistical perspective, blocking is used to isolate controllable variables that are not of the primary interest: since participants were randomized within each block (subgroup), and the number of treatment/placebo assignments in each group was not intended to be even, this randomization process leads us to unbalanced Randomized Complete Block Design (RCBD).

2.1 ANCOVA Models

On top of the treatment and the block effects, ANCOVA models involve the linear effect of a continuous covariate [7]: the models that we use are of the following form:

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma x_{ijk} + \varepsilon_{ijk},$$

where

- y_{ijk} is the k^{th} **response variable** in the i^{th} treatment group and j^{th} block (the scores at third follow-up);
- μ is the **overall mean**;
- τ_i is the i^{th} **treatment effect**;
- β_j is the j^{th} **block effect**;
- γ is the **covariate (or regression) effect**;
- $x_{ijk} = X_{ijk} - \bar{X}$ is the k^{th} **covariate (or concomitant variable)** in the i^{th} treatment group and j^{th} block (the baseline IBSS or QoL value adjusted for the mean), and
- ε_{ijk} is the k^{th} **residual** in the i^{th} treatment group and j^{th} block.

The indices correspond to $i = 1,2,\ j = 1,\dots,4,\ k = 1,\dots,n_{ij},\ \sum_i \sum_j n_{ij} = N$, where N is the number of participants.

2.2 ANCOVA Assumptions

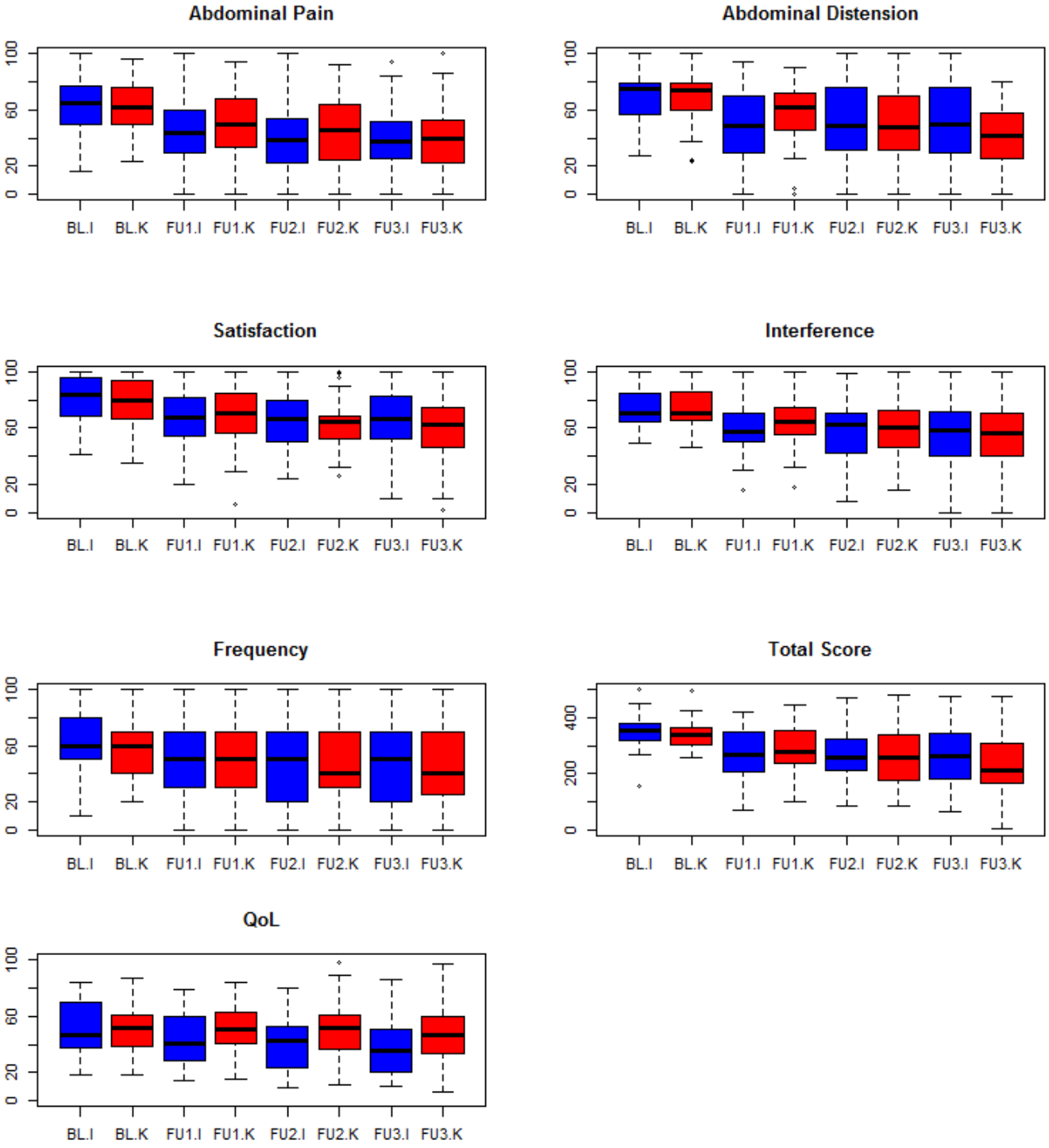
In order to use an ANCOVA model, four assumptions must be satisfied:

1. *Independence and Normality of Residuals*: the residuals are thought to be independently and identically distributed random variables following a normal distribution with zero mean (i.e. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_{\varepsilon}^2 \boldsymbol{I})$);
2. *Homogeneity of Residual Variances*: the variance of the residuals must be uniform across treatment groups;
3. *Homogeneity of Regression Slopes*: the regression effect (slope) must be uniform across treatment groups, and
4. *Linearity of Regression*: the regression relationship between the response and the covariate must be linear.

The first of these assumptions can be tested with the help of a **QQ-plot** and a scatter plot of **residual vs. fitted values**, while the second may use the **Bartlett's** or the **Levene's** test. The final assumption is not as crucial as the other three assumptions. Various remedial methods can be applied should any of these assumptions fail [6].

The third assumption, however, is critical to the ANCOVA model. It can be tested with the **equal slope test**: we run an ANCOVA regression on the models given in Sections 4 and 5 with an additional interaction term $x \times \tau$. If the interaction is not significant, the third assumption is satisfied. In the event that the interaction term is statistically significant, a different approach (e.g. moderated regression analysis, mediation analyses) is required as using the original ANCOVA model is not prescribed [8]. ANCOVA assumptions will be verified for both IBSS and QoL response variables in sections 4 and 5 respectively.

Figure 1 – Box-and-whisker plots for IBS severity scores at each time point. The blue and red columns represent the scores for treatment groups I, and K, respectively, while circles represent outlying values according to the box-and-whisker test



3. Covariance Analysis for the IBS Severity Score

A total of 100 participants were recruited for the study. One subject did not meet the recruitment criteria, and eight of which dropped out after the baseline assessment. A further three drop-outs were removed (see **Section 1.4**), leaving a total of $N = 88$ participants for the analyses for the IBS severity score and its sub scores. In order to accommodate the two imputations (again, see Section 2.4), two degrees of freedom are docked from the residual source in the ANCOVA analyses.

3.1 Total IBS Severity Score

The ANOVA table for the **ANCOVA Model on the total IBS severity score** is found in **Table 3**. At first glance, as the *p*-value for the treatment effect is 0.310, we conclude that there is not enough evidence to suggest that the two treatment effects differ at 0.05 significance level. Since the 95% confidence interval for the difference in the treatment effects include 0, the estimated treatment effects are not presented.

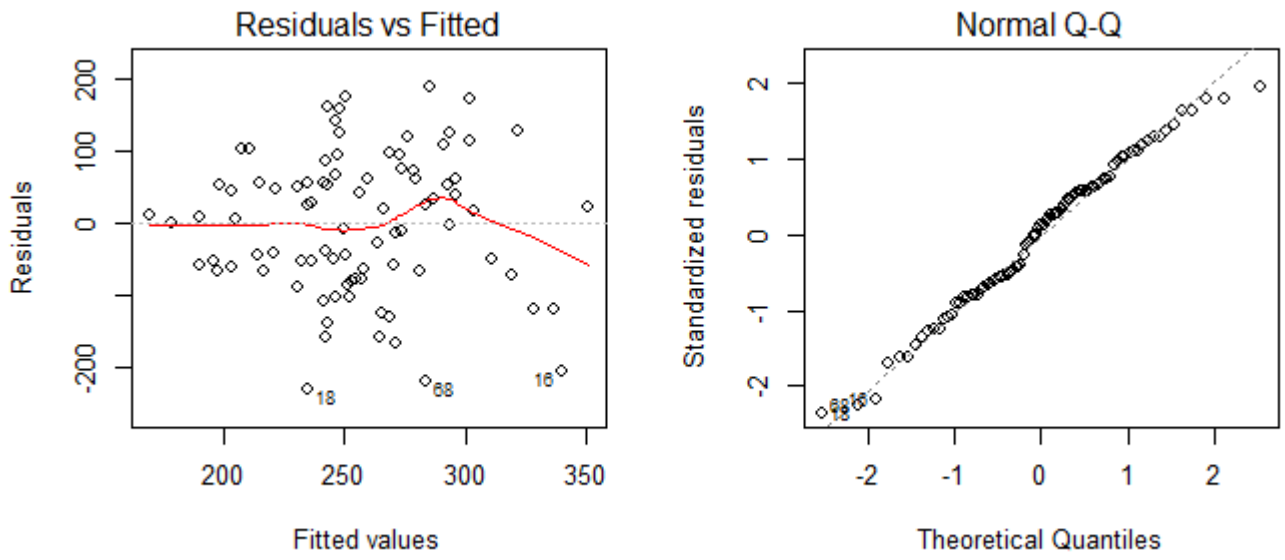
Table 3 – ANOVA table for the variance analysis on the total IBS severity score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	10521	10521	1.043	0.310
β (Block)	3	19551	6517	0.646	0.588
γ (Covariate)	1	89895	89895	8.911	0.004
ε (Residual)	81-2=79	796996	10088.56		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 2**. The data is well behaved on the normal Q-Q plot, verifying that the assumption of normality is met.

Bartlett's test is used to assess the homogeneity of the residual variances in groups K and I. The test statistic $X^2 = 0.265$, with a corresponding *p*-value of 0.60, implies that there is insufficient evidence to reject the assumption of homogeneity of variances. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 3**).

Figure 2 – Normality and independence of the residuals from ANCOVA for the total IBS severity score



Furthermore, with a *p*-value of 0.004 for the covariate effect, it seems reasonable to assume that the relationship between the response and the covariate is indeed linear.

Finally, the test for equal slopes compares the original model $y \sim \tau + \beta + \gamma x$ to the modified interaction model

$$y \sim \tau + \beta + \gamma x + \rho(x \times \tau).$$

The lack of significance of the interaction term is interpreted as favourable to the third assumption. The appropriate ANOVA table is shown in **Table 4**; the corresponding *p*-value of 0.937 indicates that that it is reasonable to assume the homogeneity of regression slopes.

Figure 3 – Homogeneity of variance between treatment groups I and K for the total IBS severity score based on ANCOVA

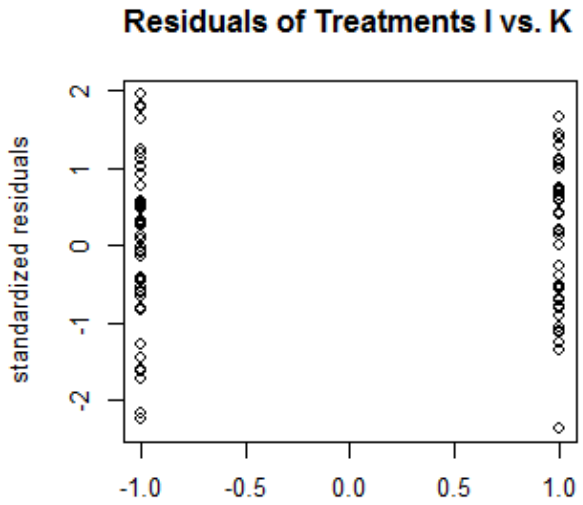


Table 4 – Homogeneity of regression slopes across treatment groups for the covariance model for the total IBS severity score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	796996				
Interaction	78	796932	1	64	0.006	0.937

The plot of residuals vs. fitted values (**Figure 2**, left) shows three outliers based on the covariance analysis. **Table 5** summarizes treatment effects on these participants. This combination provides an impetus to study the effect of possible influential observations. Note that all three outliers in **Table 5** have large reduction in the IBS severity score to categorize those participants as either not suffering from IBS (scores ranging from 0 to 75) or mildly suffering from IBS (scores ranging from 75 to 175). While their rate of reduction is anomalous compared to the rest of the participants, since not all three participants belong to one group, the covariance analysis on reduced dataset (i.e., IDs 16, 18, and 68 removed) should not alter the results significantly. Hence, no further analyses are conducted for the total IBS severity score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Table 5 – Outliers based on the covariance analysis on the total IBS severity score

ID	Group	Baseline score	Final score	Difference
16	I	448	134	-314
18	K	326	6	-320
68	I	365	65	-300

3.2 Abdominal Pain Score

The ANOVA table for the **abdominal pain score using ANCOVA Model** is found in **Table 6**. It should be noted that the *p*-value for the covariate effect is 0.630, the result suggests that analysis of variance would be more appropriate than analysis of covariance to test the difference in the abdominal pain scores in two treatment groups.

Table 6 – ANOVA table for the covariance analysis on the abdominal pain score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	192	192	0.314	0.577
β (Block)	3	3003	1001	1.639	0.187
γ (Covariate)	1	143	143	0.233	0.630
ε (Residual)	81-2=79	48261	611		

Table 7, which provides the ANOVA table for the analysis of variance on the abdominal pain score, indicates that the treatment effects do not differ as the *p*-value for the difference in the treatment effects is 0.603. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 4**. The normal Q-Q plot shows a slight deviation from the assumption of normality; however, as ANOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Table 7 – ANOVA table for the variance analysis on abdominal pain scores with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	163.31	163.31	0.273	0.603
β (Block)	3	3147.44	1049.15	1.759	0.162
ε (Residual)	83-2=81	48404	597.58		

To assess the homogeneous variances of the residuals in the groups I and K, Bartlett's test is used. There is insufficient evidence to conclude that the variances are non-homogeneous across treatment groups as the statistic is $X^2 = 0.239$ with a corresponding *p*-value of 0.625. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 5**).

Figure 4 – Normality and independence of the residuals from ANOVA for the abdominal pain scores

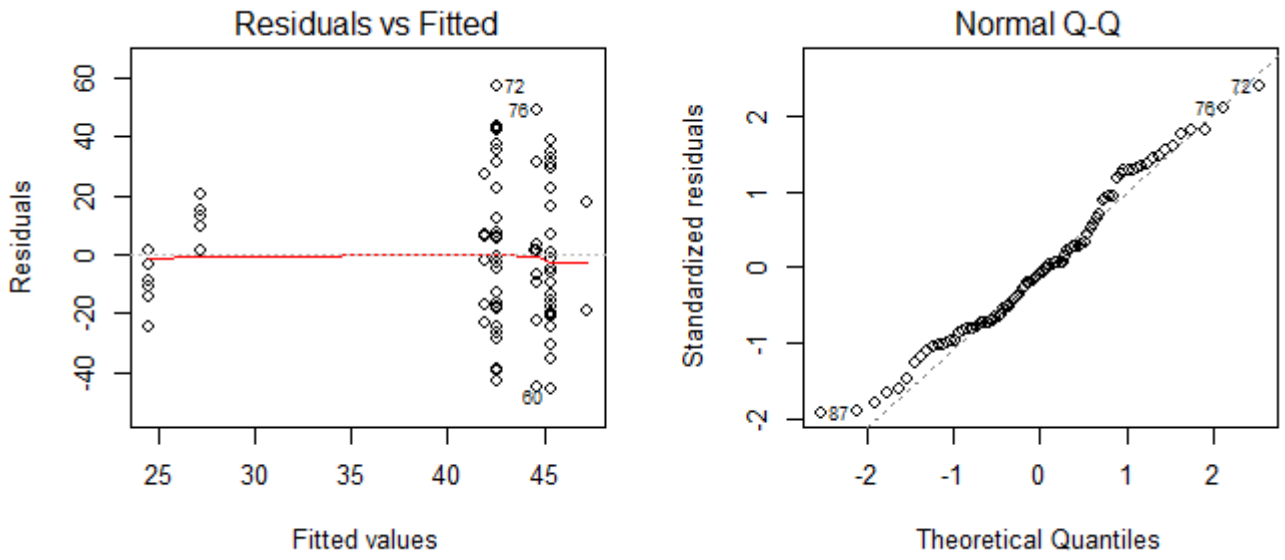
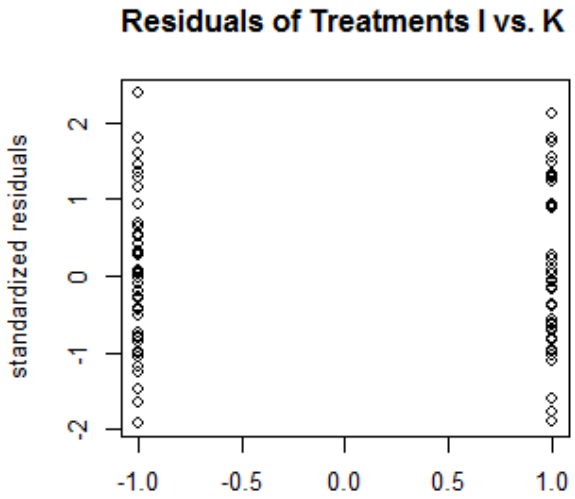


Figure 5 – Homogeneity of variance between treatment groups I and K for the ANOVA of the abdominal pain score



The plot of residuals vs. fitted values (**Figure 4**, left) shows three outliers based on the variance analysis. **Table 8** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the *p*-value associated with the treatment effect on the abdominal pain score is 0.603, analysis on the reduced dataset (i.e., potential influential observations removed) should not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Table 8 – Outliers based on the analysis of variance on the abdominal pain score

ID	Group	Baseline score	Final score	Difference
73	K	50	100	50
77	I	78	94	16
88	I	76	0	-76

3.3 Satisfaction Score

Table 9 provides the ANOVA table for the **satisfaction score using ANCOVA Model**. As the *p*-value for the treatment effect is given to be 0.330, we conclude that there is not enough evidence to suggest that the treatment has an effect at the 0.05 significance level.

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 6**. The normal Q-Q plot demonstrates deviation from the assumption of normality on both tails; however, as ANCOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Due to a moderate deviation from the normality assumption, Levene's test is used to assess the homogeneous variances of the residuals in the groups I and K. The test statistic is $W = 0.072$ with a corresponding *p*-value of 0.790. There is thus insufficient evidence to conclude that the variances are non-homogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 7**).

Table 9 – ANOVA table for the covariance analysis on satisfaction score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
$\boldsymbol{\tau}$ (Treatment)	1	837	837	0.961	0.330
$\boldsymbol{\beta}$ (Block)	3	4089	1363	1.565	0.205
$\boldsymbol{\gamma}$ (Covariate)	1	13078	13078	15.013	<0.001
$\boldsymbol{\varepsilon}$ (Residual)	81-2=79	68815	871		

Furthermore, with a *p*-value for the covariate effect being less than 0.001, it seems reasonable to assume that the relationship between the response and the covariate is linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 10**; the corresponding *p*-value of 0.261 indicates that that it is reasonable to assume the homogeneity of regression slopes.

The plot of residuals vs. fitted values (**Figure 6**, left) shows three outliers based on the covariance analysis. **Table 11** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the p-value associated with the treatment effect on the satisfaction score is 0.330, analysis on the reduced dataset (i.e., potential influential observations removed) should not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Figure 6 – Normality and independence of the residuals from ANCOVA for the satisfaction score

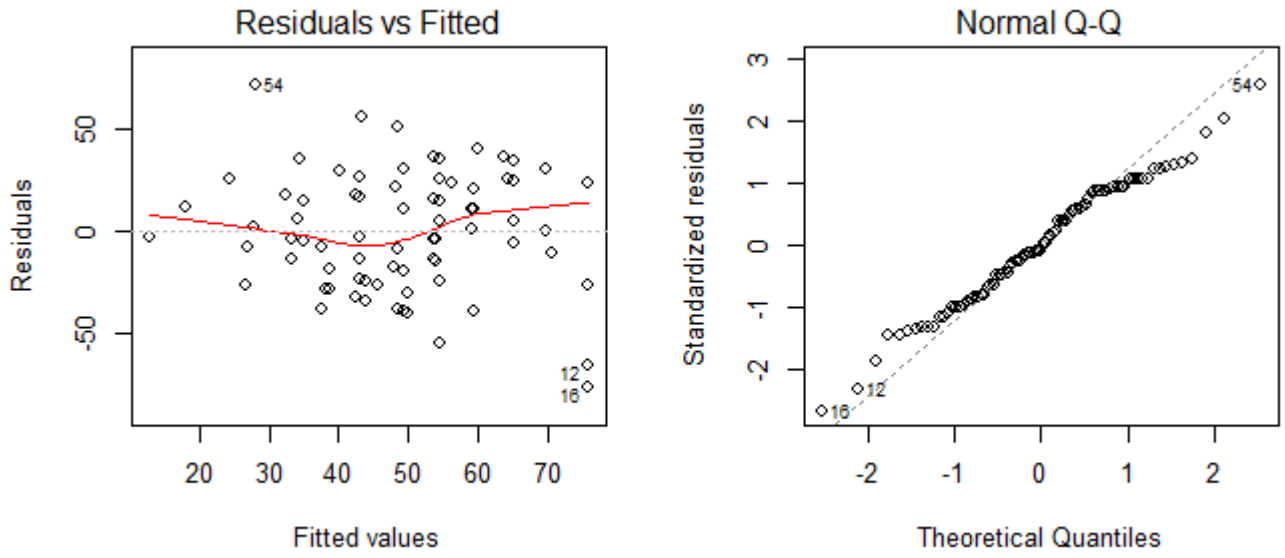


Figure 7 – Homogeneity of variance between treatment groups I and K for the ANCOVA of the satisfaction score

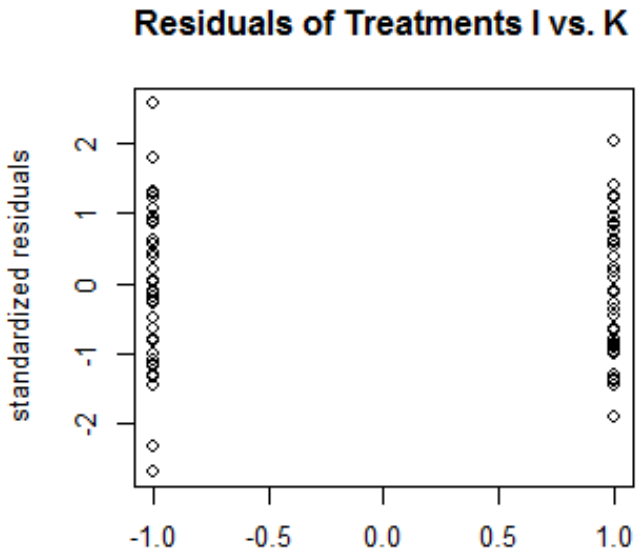


Table 10 – Homogeneity of regression slopes across treatment groups for the covariance model for the satisfaction score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	68815				
Interaction	78	67700	1	1115	1.280	0.261

Table 11 – Outliers based on the analysis of variance on the satisfaction score

ID	Group	Baseline score	Final score	Difference
12	I	100	10	-90
16	K	100	0	-100
55	I	10	100	90

3.4 Interference Score

Table 12 provides the ANOVA table for the **interference score using ANCOVA Model**. As the *p*-value for the treatment effect is given to be 0.327, we conclude that there is not enough evidence to suggest that the treatment has an effect at the 0.05 significance level.

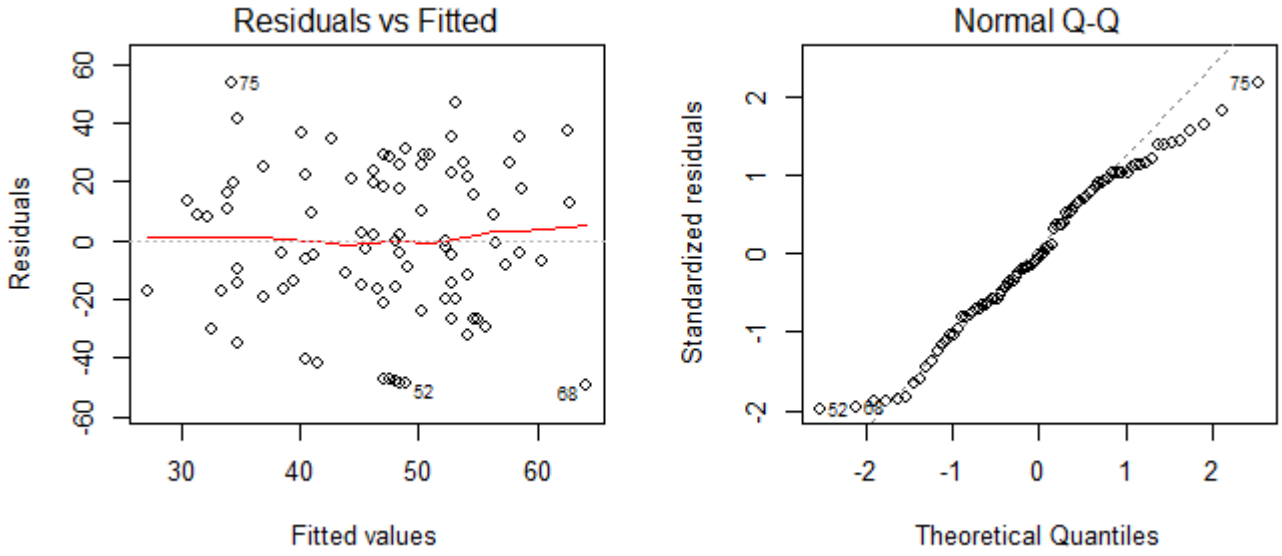
Table 12 – ANOVA table for the covariance analysis on interference score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	680	680	0.973	0.327
β (Block)	3	878	293	0.419	0.740
γ (Covariate)	1	4899	4899	7.013	0.010
ε (Residual)	81-2=79	55183	699		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 8**. The normal Q-Q plot demonstrates deviation from the assumption of normality on both tails; however, as ANCOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Due to a moderate deviation from the normality assumption, Levene's test is used to assess the homogeneous variances of the residuals in the groups I and K. The test statistic is $W = 0.068$ with a corresponding *p*-value of 0.795. There is thus insufficient evidence to conclude that the variances are non-homogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 9**).

Figure 8 – Normality and independence of the residuals from ANCOVA for the interference score



Furthermore, with a *p*-value for the covariate effect being 0.01, it seems reasonable to assume that the relationship between the response and the covariate is linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 13**; the corresponding *p*-value of 0.261 indicates that that it is reasonable to assume the homogeneity of regression slopes.

The plot of residuals vs. fitted values (**Figure 8**, left) shows three outliers based on the covariance analysis. **Table 14** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the p-value associated with the treatment effect on the interference score is 0.327, analysis on the reduced dataset (i.e., potential influential observations removed) should not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Figure 9 – Homogeneity of variance between treatment groups I and K for the ANCOVA of the interference score

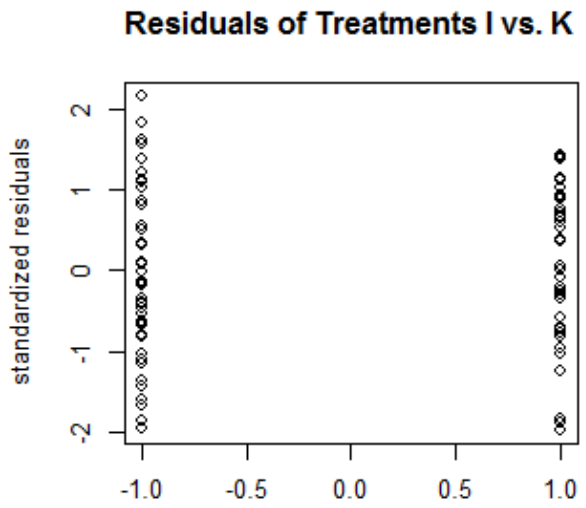


Table 13 – Homogeneity of regression slopes across treatment groups for the covariance model for the interference score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	68815				
Interference	78	67700	1	1115	1.280	0.261

Table 14 – Outliers based on the analysis of variance on the interference score

ID	Group	Baseline score	Final score	Difference
53	K	87	0	-87
69	I	100	15	-85
76	K	56	88	32

3.5 Frequency Score

Table 15 provides the ANOVA table for the **frequency score using ANCOVA Model**. As the *p*-value for the treatment effect is given to be 0.358, we conclude that there is not enough evidence to suggest that the treatment has an effect at the 0.05 significance level.

Table 15 – ANOVA table for the covariance analysis on frequency score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
$\boldsymbol{\tau}$ (Treatment)	1	596	596	0.854	0.358
$\boldsymbol{\beta}$ (Block)	3	1116	372	0.533	0.661
$\boldsymbol{\gamma}$ (Covariate)	1	3588	3588	7.083	0.009
$\boldsymbol{\varepsilon}$ (Residual)	81-2=79	40014	507		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 10**. The normal Q-Q plot demonstrates a slight deviation from the assumption of normality; however, as ANCOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Due to a minor deviation from the normality assumption, Levene's test is used to assess the homogeneous variances of the residuals in the groups I and K. The test statistic is $W = 0.321$ with a corresponding *p*-value of 0.573. There is thus

insufficient evidence to conclude that the variances are non-homogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 11**).

Figure 10 – Normality and independence of the residuals from ANCOVA for the frequency score

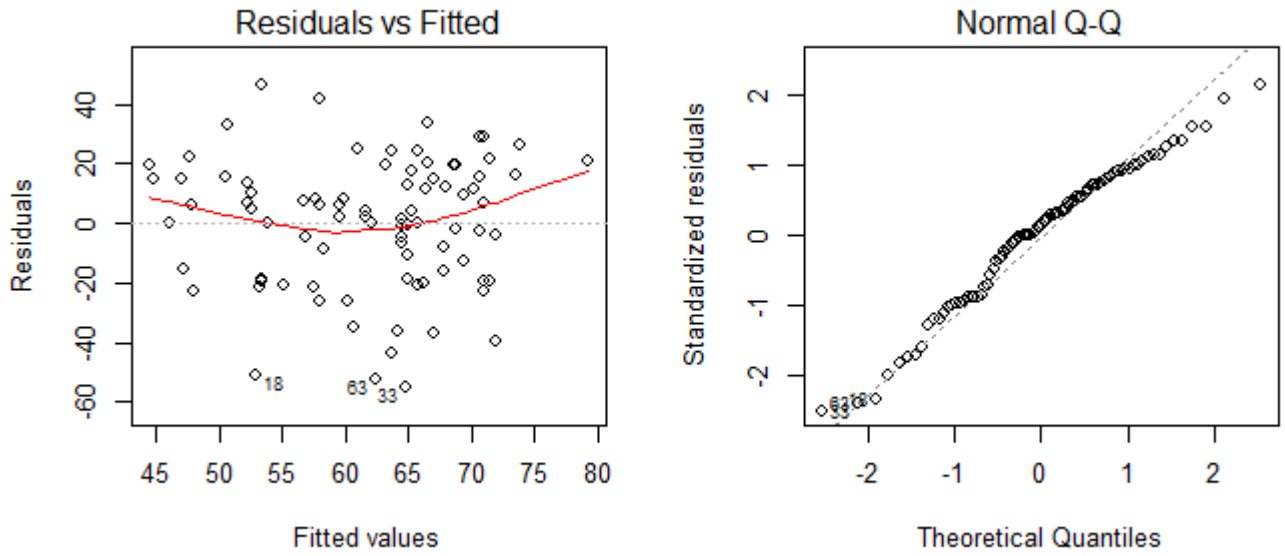
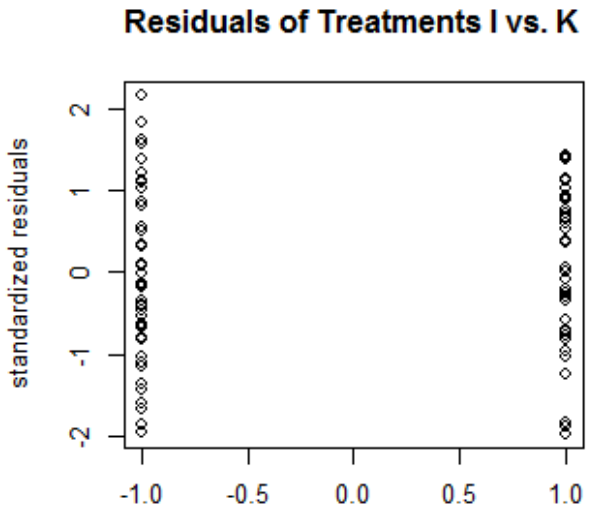


Figure 11 – Homogeneity of variance between treatment groups I and K for the ANCOVA of the frequency score



Furthermore, with a *p*-value for the covariate effect being 0.009, it seems reasonable to assume that the relationship between the response and the covariate is linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 16**; the corresponding *p*-value of 0.427 indicates that that it is reasonable to assume the homogeneity of regression slopes.

The plot of residuals vs. fitted values (**Figure 10**, left) shows three outliers based on the covariance analysis. **Table 17** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the p-value associated with the treatment effect on the satisfaction score is 0.358, analysis on the reduced dataset (i.e., potential influential observations removed) should not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Table 16 – Homogeneity of regression slopes across treatment groups for the covariance model for the frequency score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	40014				
Frequency	78	40006	1	8.000	0.016	0.427

Table 17 – Outliers based on the analysis of variance on the frequency score

ID	Group	Baseline score	Final score	Difference
18	K	66.7	2	-64.7
34	I	82.7	10	-72.7
64	K	90	10	-80

3.6 Abdominal Distension Score

Table 18 provides the ANOVA table for the **abdominal distension score using ANCOVA Model**. As the *p*-value for the treatment effect is given to be 0.902, we conclude that there is not enough evidence to suggest that the treatment has an effect at the 0.05 significance level.

Table 18 – ANOVA table for the covariance analysis on abdominal distension score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	7	7	0.015	0.902
β (Block)	3	847	282	0.586	0.626
γ (Covariate)	1	5383	5383	11.182	0.001
ε (Residual)	81-2=79	38028	481		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 12**. The normal Q-Q plot demonstrates a slight deviation from the assumption of normality; however, as ANCOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Due to a minor deviation from the normality assumption, Levene's test is used to assess the homogeneous variances of the residuals in the groups K vs. I. The test statistic is $W = 0.059$ with a corresponding *p*-value of 0.809. There is thus insufficient evidence to conclude that the variances are non-homogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 13**).

Furthermore, with a *p*-value for the covariate effect being 0.001, it seems reasonable to assume that the relationship between the response and the covariate is linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 19**; the corresponding *p*-value of 0.835 indicates that that it is reasonable to assume the homogeneity of regression slopes.

The plot of residuals vs. fitted values (**Figure 12**, left) shows three outliers based on the covariance analysis. **Table 20** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the p-value associated with the treatment effect on the satisfaction score is 0.358, analysis on the reduced dataset (i.e., potential influential observations removed) would not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Figure 12 – Normality and independence of the residuals from ANCOVA for the abdominal distension score

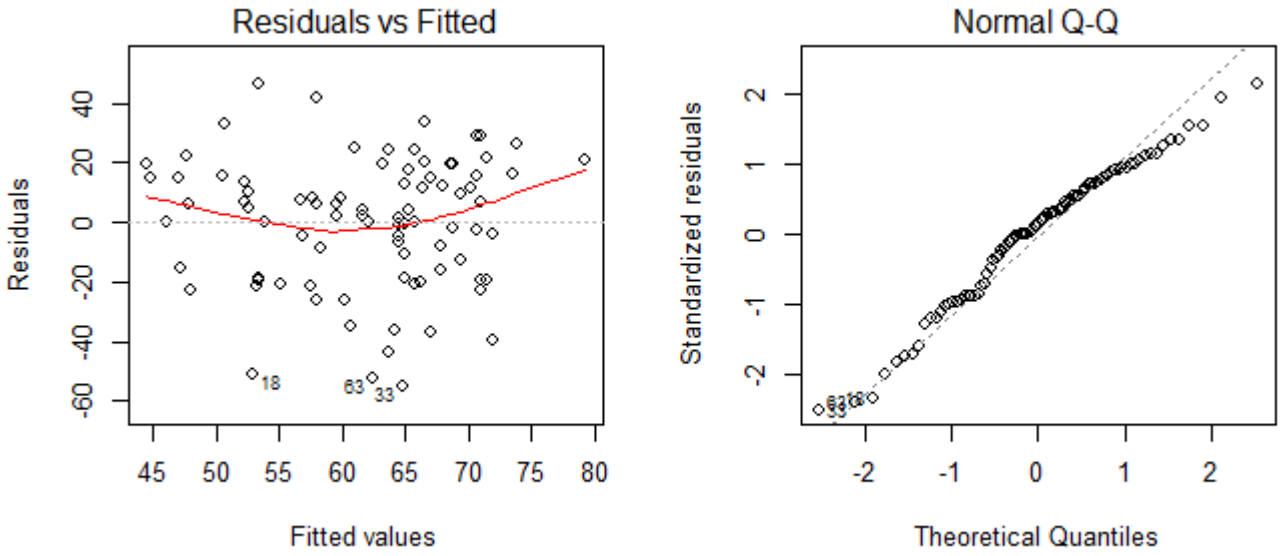


Figure 13 – Homogeneity of variance between treatment groups I and K for the ANCOVA of the abdominal distension score

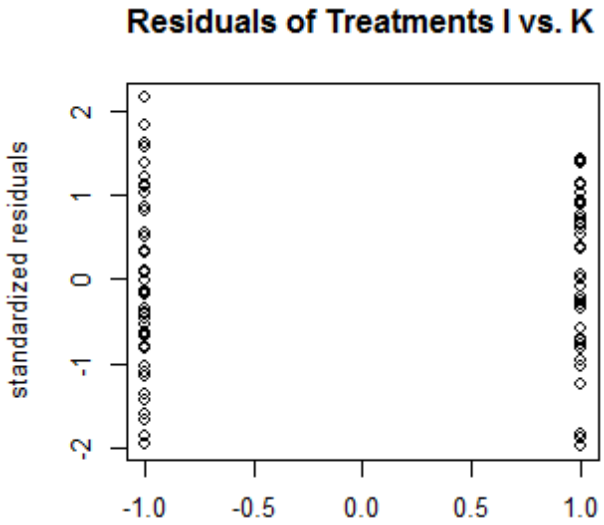


Table 19 – Homogeneity of regression slopes across treatment groups for the covariance model for the abdominal distension score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	38028				
Interaction	78	38007	1	21.000	0.044	0.835

Table 20 – Outliers based on the analysis of variance on the abdominal distension score

ID	Group	Baseline score	Final score	Difference
18	K	66.7	0	-66.7
64	I	80	10	-70
69	I	75	5	-70

4. Covariance Analysis for the QoL Score

As before, a total of 100 participants were recruited for the study, where one subject did not meet the recruitment criteria, three subjects had incomplete baseline measure for QoL, and eight of which dropped out after the baseline assessment. A further four drop-outs were removed (see Section 2.4), leaving a total of $N = 84$ participants for the analyses for the IBS severity score and its sub scores. In order to accommodate the two imputations (again, see Section 2.4), two degrees of freedom are docked from the residual source in the ANCOVA analyses.

4.1 QoL Score on Full Dataset

The ANOVA table for the **ANCOVA Model on the QoL score** is found in **Table 21**. At first glance, as the *p*-value for the treatment effect is 0.061, we conclude that there is not enough evidence to suggest that the two treatment effects differ at 0.05 significance level; however, it should be noted that the point estimate yields that, on average, participants in treatment group I have lost an extra 7.26 QoL score over the course of three months treatment period.

Table 21 – ANOVA table for the variance analysis on the QoL score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	1099	1099	3.629	0.061
β (Block)	3	370	123	0.407	0.748
γ (Covariate)	1	14847	14847	49.031	<0.001
ε (Residual)	78-2=76	23013	303		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 14**. The data is well behaved on the normal Q-Q plot, verifying that the assumption of normality is met.

Bartlett's test is used to assess the homogeneous variances of the residuals in the groups K vs. I. The test statistic is $X^2 = 0.006$, with a corresponding *p*-value of 0.937 imply that there is insufficient evidence to conclude that the variances are heterogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 15**).

Figure 14 – Normality and independence of the residuals from ANCOVA for the QoL score

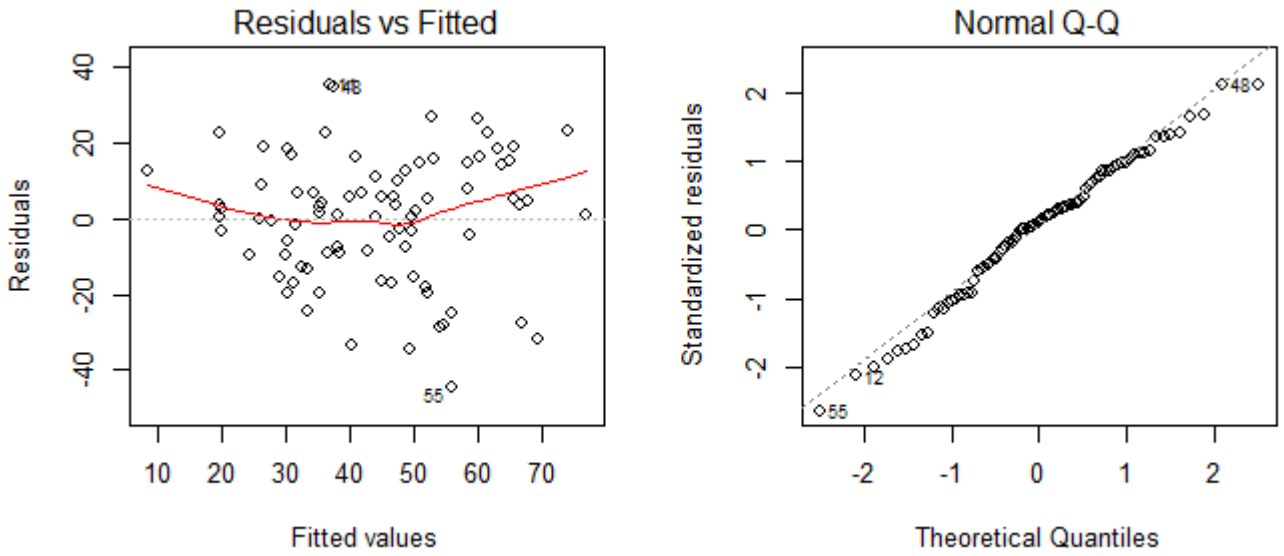
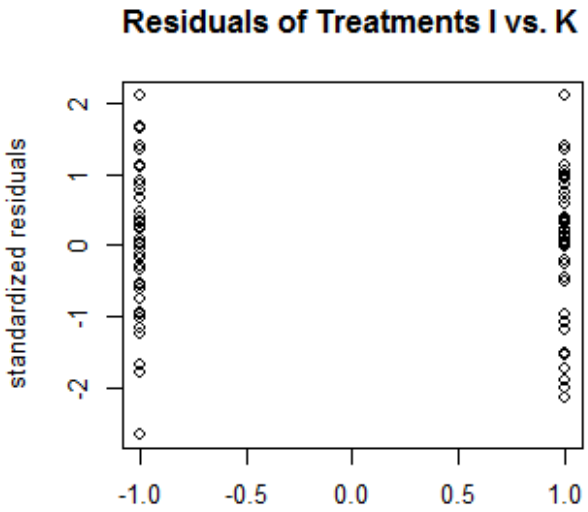


Figure 15 – Homogeneity of variance between treatment groups I and K for the QoL score based on ANCOVA



Furthermore, with a *p*-value for the covariate effect being less than 0.001, it seems reasonable to assume that the relationship between the response and the covariate is indeed linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 22**; the corresponding *p*-value of 0.481 indicates that that it is reasonable to assume the homogeneity of regression slopes.

Table 22 – Homogeneity of regression slopes across treatment groups for the covariance model for the QoL score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	76	23014				
Interaction	75	22862	1	152.000	0.502	0.481

The plot of residuals vs. fitted values (**Figure 14**, left) shows three outliers based on the covariance analysis. **Table 23** summarizes treatment effects on these participants. This combination provides an impetus to study the effect of possible influential observations. Since the p-value associated with the difference in the effects of the two treatment groups is 0.061, we will examine whether the treatment effect would be statistically significant, under the removal of the potential influential observations.

Table 23 – Outliers based on the covariance analysis on the QoL score

ID	Group	Baseline score	Final score	Difference
14	K	37.5	72.1	34.6
59	I	54.4	72.1	11.7
69	I	70.2	11.4	-58.8

4.2 QoL Score on Reduced Dataset

The ANOVA table for the **ANCOVA Model on the QoL score based on a reduced dataset** is found in **Table 24**. At first glance, as the *p*-value for the treatment effect is increased to 0.093, we conclude that there is not enough evidence to suggest that the two treatment effects differ at a 0.05 significance level; however, it should be noted that the point estimate yields that, on average, participants in treatment group I have lost an extra 6.04 QoL score over the course of three months treatment period.

Table 24 – ANOVA table for the variance analysis on the QoL score on a reduced dataset with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	733	733	2.906	0.093
β (Block)	3	730	243	0.965	0.414
γ (Covariate)	1	16494	16494	65.381	<0.001
ε (Residual)	75-2=73	18416	252		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 16**. The data shows a minor deviation from the assumption of normality on the normal Q-Q plot; however, as the ANCOVA is moderately robust to the deviation from the normality assumption, the level of deviation seen here is no concern.

The Levene's test is thus used to assess the homogeneous variances of the residuals in the groups I and K. The test statistic is $W = 0.023$, with a corresponding p-value of 0.881, implying that there is insufficient evidence to conclude that the variances are heterogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 17**).

Figure 16 – Normality and independence of the residuals from ANCOVA for the QoL score

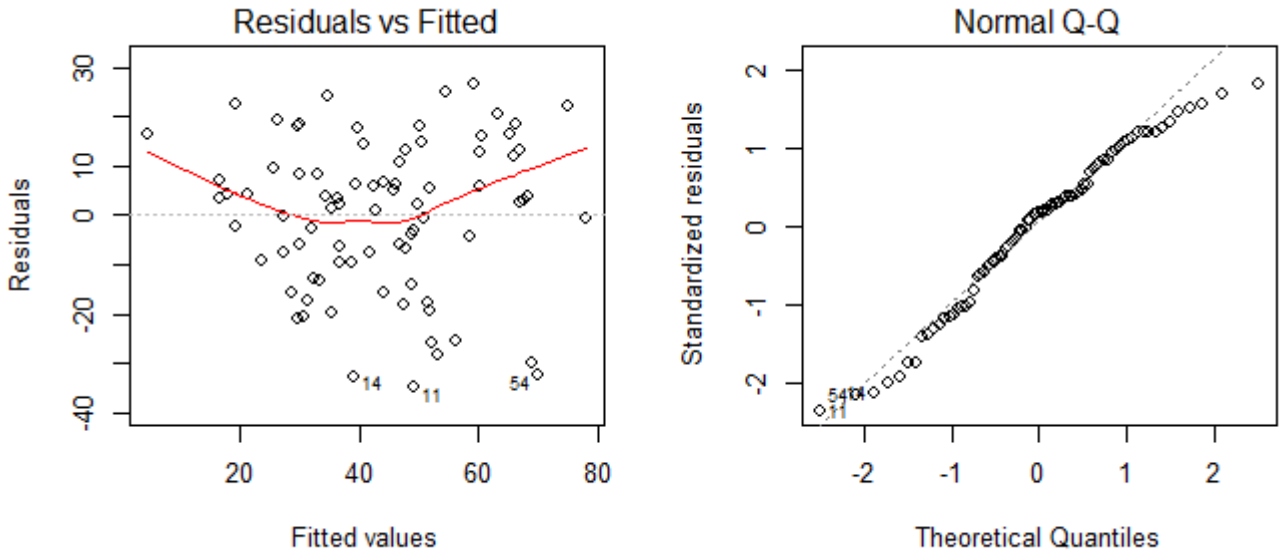
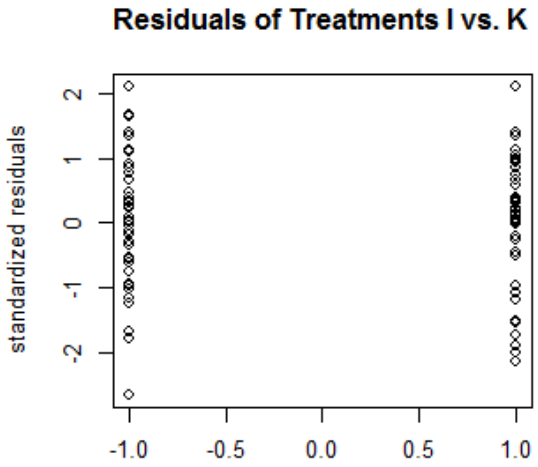


Figure 17 – Homogeneity of variance between treatment groups I and K for the QoL score based on ANCOVA



Furthermore, with a *p*-value for the covariate effect being less than 0.001, it seems reasonable to assume that the relationship between the response and the covariate is indeed linear. The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 25**; the corresponding *p*-value of 0.467 indicates that that it is reasonable to assume the homogeneity of regression slopes.

Table 25 – Homogeneity of regression slopes across treatment groups for the covariance model for the QoL score on a reduced dataset with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	73	18416				
Interaction	72	18281	1	135.000	0.535	0.467

The plot of residuals vs. fitted values (**Figure 16**, left) shows three outliers based on the covariance analysis. However, since these observations are classified as an outlier only due to the removal of the three outliers from the original dataset, we will not perform any further analyses on the QoL score, and conclude that, at 95% significance level, two treatment groups do not differ in their treatment effects.

5. IBS Sub-Score Analyses for 2010 Dataset

Extremely similar analyses were conducted for the sub-scores of the IBS data collected during the 2010 pilot study; in the interest of readability, the results were condensed and placed in a table format in the Executive Summary. While none of the sub-scores showed statistically significant improvement under the probiotic agent, one of them (Statisfaction, *p*-value: 0.085) was nearly significant.

6. Conclusions and Recommendations

We end the report with key findings of our analysis, as well as some recommendations for future investigations.

6.1 Blocking and Balanced Designs

In this report, we have found that blocking (or subgrouping) the participants according to their gender and age does not play an important role in the ANCOVA. In future studies involving this probiotic agent, blocking should only be used if there are compelling reasons to suspect that treatment effects are different for at least one subgroup, as blocking results in fewer degrees of freedom.

Special care should also be taken to have a balanced design (i.e., equal number of replicates for each subgroup), especially if subgroup analyses are of interest: for instance, the overwhelming number of female participants and small number of male participants make any conclusions about male subgroups statistically unsound.

6.2 Recruitment Process

In the 2013 IBS Study, participants needed to come forward to be selected. The recruitment process used advertisements on the radio, in local newsletters and newspapers, on the web and social media, as well as posters with which local MDs and NDs could encourage patient referrals.

The elephant in the room is that this type of recruitment process leads to self-selection biases: the participants in the 2013 IBS Study may not constitute a representative sample of IBS sufferers, which makes it difficult to generalize the result of the analyses beyond the collected sample, even when there is a significant impact.

This is a problem that plagues numerous clinical studies – unfortunately, it is quite difficult to counter this situation.

6.3 Practical Significance of Results

With the caveat brought up in section 6.2, our interpretation of the covariance analyses results is that there is simply not enough evidence to conclude that the agent is effective against IBS.

It is true that the difference in the treatment effects between the two groups on the (self-reported) QoL score is nearly statistically significant at the 0.05 significance level. The corresponding estimated difference in the treatment effects is 7.26 under the using full dataset, which means that on average, participants in the group I seem to have lost an extra 7.26 QoL points over the course of three months, compared to those in the group K. However, given the amount of variability in individuals from month to month, we are reluctant to conclude that the agent under investigation provides a practically significant improvement in the average participant's quality of life.

Further investigation may shed some light on the situation and will help us determine if the relationship between the agent and QoL is causal or spurious.

6.4 Publication of Results

Even though this study did not find any statistically significant improvement for IBS, it should be published in order to counter publication bias.

References

[1] "Irritable Bowel Syndrome," [Online]. Available: http://en.wikipedia.org/wiki/Irritable_bowel_syndrome. [Accessed 5 May 2013].

[2] P. Paré, J. Gray, S. Lam, R. Balshaw, S. Khorasheh, M. Barbeau, S. Kelly and C. R. McBurney, "Health-related quality of life, work productivity, and health care resource utilization of subjects with irritable bowel syndrome: baseline results from LOGIC (Longitudinal Outcomes Study of Gastrointestinal Symptoms in Canada), a naturalistic study," Clinical Therapeutics, vol. 28, no. 10, pp. 1726-35, 2006.

[3] S. Maxion-Bergemann, F. Thielecke, F. Abel and R. Bergemann, "Costs of irritable bowel syndrome in the UK and US," PharmacoEconomics, vol. 24, no. 1, pp. 21-37, 2006.

[4] P. Herman, C. Kooley and D. Seely, "Double-blind placebo-controlled pilot study to investigate the effects of an investigational Probiotic on Irritable Bowel Syndrome," 2011.

[5] S. Hagiwara, "Nonresponse Error in Survey Sampling: Comparison of Different Imputation Methods," Ottawa, 2012.

[6] M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li, Applied Linear Statistical Models, 5th ed., New York: McGraw-Hill/Irwin, 2004.

[7] P. W. John, Statistical Design and Analysis of Experiments, New York: Macmillan, 1971.

[8] S. Green and N. Salkind, Using SPSS for Windows and Macintosh: Analyzing and Understanding Data, 6th ed., Upper Saddle River, NJ: Prentice Hall, 2011.

[9] Hagiwara, S. and Boily, P., "Covariance Analysis for the 2010 CCNM Pilot Study on Irritable Bowel Syndrome," Internal Report to the CCNM (available from CQADS), 2013.