Fortified Fish Sauce for Combating Infantile Beriberi in Rural Cambodia

A Statistical Report

Department of Human Nutrition, The University of British Columbia

Gal Av-Gay

March 30th, 2016

Contents

1 Introduction

Infantile beriberi (Vitamin B_1/thiamine deficiency) is a serious health concern in rural Cambodia. This is primarily due to limited thiamine in the diets of most rural Cambodians. You have developed a thiamine-fortified version of fish sauce, a common condiment used ubiquitously throughout Cambodia, intended to combat this problem. To investigate the health benefits of this fortified fish sauce, two randomized controlled clinical trials were performed on mothers and their children in rural Cambodia.

Three concentrations of thiamin-fortified fish sauce were randomly assigned and distributed to the test subjects, who consumed the fish sauce over a duration of six months. Baseline and endpoint measures of blood/breast-milk thiamine concentrations were obtained for mothers and their children who consumed the fish sauce. Some additional demographics were also collected for each respondent including age, village, and occupation, among others. The general objective of these clinical trials is to assess the effectiveness of fortifying fish sauce on combating beriberi in rural Cambodia.

This report outlines approaches for the statistical analysis of these clinical trials. How the data was collected is described, and how it may be explored is explained. The nature of the missingness in the data is discussed, as well as how it affects the analysis. Some basic (ANOVA) and more comprehensive (ANCOVA/ Linear Regression) approaches to analyzing the data are introduced, with a focus on checking the model assumptions. The conditional change model is explained, which is recommended for analysis of data with baseline and endpoint measurements of the response of interest.

Beyond providing guidance for statistical analysis of the data, this document addresses a number of previously posed questions. The implications of the different assumptions for each model are explained. A method for investigating the relationship between mother's blood thiamine concentration, breast-milk thiamine

concentration, and infant blood thiamine concentration is also included.

2 Data Description

Two separate randomized controlled clinical trials were carried out with mothers and children in rural Cambodia. One trial enrolled pregnant mothers who gave birth during the six-month duration of the clinical trial, while the other enrolled mothers who had at least one child below the age of five. In both trials, families were randomized into one of three treatment groups:

1. Control (C) - Families in this treatment group were given normal, unfortified fish sauce;
2. Low-Concentration (LC) - Families in this treatment group were given fish sauce fortified with a low concentration of thiamin;
3. High-Concentration (HC) - Families in this treatment group were given fish sauce fortified with a high concentration of thiamin.

Details of the data collected for each trial are summarized in Table 1,

Table 1: Summary of Data for the two Clinical Trials

Non-Pregnant	Pregnant
- 270 mothers at beginning - Fish-sauce consumption data - Various demographics - Blood T concentration at baseline and endpoint for mother and child - 197 mothers and 191 children after accounting for missing data	- 90 mothers at beginning - Fish-sauce consumption data - Various demographics - Mother's blood T concentration at baseline and endpoint - Infant's blood T concentration at endpoint - Mother's breastmilk T concentration at endpoint - 77 mothers and 65 children after accounting for missing data

Altogether the clinical trials appear to have been executed well. For example, it appears that the women were individually randomized to the treatment groups, meaning that bias in treatment assignment is effectively eliminated. Without randomization, this would not be an experiment, and claims regarding causality (i.e. ingestion of fortified fish sauce increases blood thiamine) could not be supported. Randomization may have been carried out across all 42 villages as a single process, or rather within each village in a stratified fashion. If randomization was carried out in a non-stratified fashion, there may be a more apparent effect of village on the response variable.

It is important to identify the study population by asking who the women from these 42 villages represent, as the results of the analysis can only be extended to that population. If the 42 villages are viewed as representative of all rural Cambodian villages, the study population might be considered to be 'Cambodian mothers living in rural villages'. It is also important to identify whether women in one village are 'more alike' than women in other villages as this would suggest the necessity for inclusion of village as a factor in the analysis. Examination of this issue can be done by visual inspection as described in Section 2.3.

Aside from this, the primary issue regarding how the data was collected concerns potential patterns in the 'missingness' of the data.

2.1 Missing Data

The main issue here is missing response variable data, which is a fundamental concern. The issue of missing covariate data is secondary albeit still important. The following section explains how to deal with missing values in the response variables. In order to determine what can be done about missing data, it is important to first examine whether there is a pattern to what data is missing.

If the response data is 'Missing at Random', i.e, the missing response (blood thiamine) values do not correlate with the treatment levels (fish sauce thiamine

concentration), or in other words, the distribution of missing values is fairly uniform across the treatment levels, then a number of issues are avoided. In this case, missing data can be omitted and analysis can be performed on only cases with known outcomes without having to worry about non-response bias. Interpretation can be appropriately extended to the study population (of rural Cambodian mothers who satisfy the specified eligibility criteria). Furthermore, imputation methods for filling in missing values based on the distribution of available values can be implemented effectively, as will be explained later.

Consider however, a situation where response data is missing more often for families with low income than for families with high income. In this case, the part of the sample with response data available is not representative of the study population, and the results of a statistical analysis will likely be biased due to the non-response bias present in the data. This means that interpretation of a statistical analysis cannot properly be extended to the study population without considering what sort of bias is introduced. However, given that data is available for a number of covariates, regression imputation is still an option in this case, and may reduce non-response bias (this will be discussed later)

It is beneficial to identify whether the missing response data appears to be missing at random. This can be done by comparing the distributions of covariates for the individuals with missing response data and for the individuals with available response data. For example, if the mean income for test subjects with missing endpoint blood thiamine concentrations is clearly different from the mean income for the remaining test subjects, the response data is not missing at random. The question of whether there is a significant difference between the two means can be assessed using a two sample t-test, however significance is not necessarily the issue here. It should also be noted that since the exact reasons for why many of the data points are missing are known, the mechanism by which data is missing does not necessarily have to be investigated statistically. It may be possible to make a judgement based on subject-matter expertise regarding whether the data

is actually missing at random.

2.2 Imputation

Different methods exist for imputing data. The most basic method involves imputation using the mean value of the dependent variable for the particular treatment group. Consider a situation where the endpoint blood thiamine concentration is missing for a particular mother in the control treatment group. This endpoint blood thiamine concentration is imputed with the mean value over mothers in the control treatment group; this is called 'mean imputation'. This is a very basic approach to imputation that can only be recommended if the data is missing at random. If, for example, the data is not missing at random, and it is missing more often for mothers from a particular town, or from a specific occupation, then this pattern in the missing data will be retained even after mean-imputation.

Another method, regression imputation, predicts each missing value based on the remaining covariates, by regressing the variable containing missing values against the remaining covariates. For example, for a given mother with a missing value for endpoint blood thiamine concentration, the missing blood thiamine concentration is imputed by evaluating the fitted regression equation at this mother's values of the remaining covariates (income, age, level of education, etc.). This method helps eliminate some of the non-response bias associated with what data is missing. For example, if data is missing often for mothers at a low education level, the non-response bias implies that the data is more representative of mothers with a high education level than those with a low education level. If regression imputation is used to fill in those missing values, the non-response bias associated with this unbalanced representation across education levels is corrected for, at least to a certain extent. However, a different problem arises: since missing values are predicted based on available covariates, this method artificially strengthens

relationships that otherwise may not be very strong, as predictions will always lie on the fitted regression line.

A possible remedy for this issue is to introduce an error term to the predictions that is proportional to the regression residual error. This is called stochastic imputation and unfortunately still results in some issues. For example, it could be that the relationship between income and endpoint blood thiamine concentration is not actually very strong for the entire population. However, if the relationship is strong enough in the subset of non-missing data that is being used to predict the missing data, then the strength of this relationship will persist in the imputed dataset, even though it may be incorrect. In addition, the noise that is introduced by using stochastic imputation may be too small, resulting in a similar (albeit reduced) issue to that encountered via regression imputation. This method is recommended when data is missing for only the response variable.

When data is missing across all variables, multiple imputation is recommended. This method assumes that the data come from a multivariate distribution (often the normal), and that missing values can occur for any variable. Often this form of imputation will be implemented on a transformed scale for some of the variables to make the multivariate distributional assumption reasonable. Multiple imputation involves creating multiple imputed data-sets using stochastic imputation. Each of these imputed data-sets is analyzed separately, and the results are averaged over all these data-sets. Multiple imputation does not necessarily attempt to estimate each missing value, but essentially simulates a sample of the missing values. Multiple imputation does this in such a way so that valid statistical inferences (e.g., finding valid confidence intervals) can still be made. This is the main benefit of multiple imputation.

When the data is missing at random, there is no need for imputation, but any of the imputation methods can still be implemented without risking introduction of bias. In either case, when the data is missing at random or not, stochastic imputation is recommended if data is missing from only the response variable, and

multiple imputation is recommended if data is missing from all variables. There exist methods for implementing imputation in SPSS, and instructions for doing so can be found in (1).

2.3 Exploratory Data Analysis

It is important to first check whether the presence of missing response variable data correlates with other covariates, as this will give an indication of whether the response is missing at random. One method for doing this involves contingency tables, such as Table 2, where it can be seen that mothers with a higher level of education are more likely to complete the study. Another option utilizes box-plots as in Figure 1, where the distribution of household incomes is compared between mothers who are missing the response or not. Visual inspection of such tables and plots will clarify disparities between distributions of missing and non-missing data, however, assessing significance of these differences requires hypothesis testing via a two-sample t-test. Once again, significance is not necessarily the issue here. The question is whether the assumption of randomly distributed missing values is reasonable.

Table 2: Contingency table for mother's education level versus missingness of response, for non-pregnant trial. Individuals with a missing value for education level are not included.

Response	**Primary-School**	**Lower Secondary-School**	**Upper Secondary-School**	**Higher Education**	**Total**
Available	96 (70.1 %)	57 (67.9%)	21 (80.1 %)	2 (100 %)	197
Missing	41 (29.9 %)	27 (32.1 %)	5 (19.9 %)	0 (0 %)	79
Total	137	84	26	2	249

Next, additional trends can be visualized. Most importantly, the distribution of the response variable for different treatment groups can be illustrated using side by side box-plots, as in Figure 2, from which it can be seen that the non-control

treatment groups result in a markedly improved response. This is evident from the fact that the IQRs (represented by the distance between the top and bottom of the box in the boxplot) for the treatment groups do not overlap with that of the control group.

Finally, relationships between the remaining variables and the response or the treatment groups can be explored. One might want to investigate possible relationships between treatment groups and other variables such as education, income, or town, to ensure that randomization was properly implemented. Box-plots are recommended for a continuous variable versus a categorical variable, and contingency tables and bar charts are recommended for relating categorical variables. Scatter-plots are useful for plotting pairs of continuous variables against one another.

Figure 1: Boxplots for 12-month household income for the non-pregnant trial. There is no strong indication of a relationship between income and missingness from this plot.

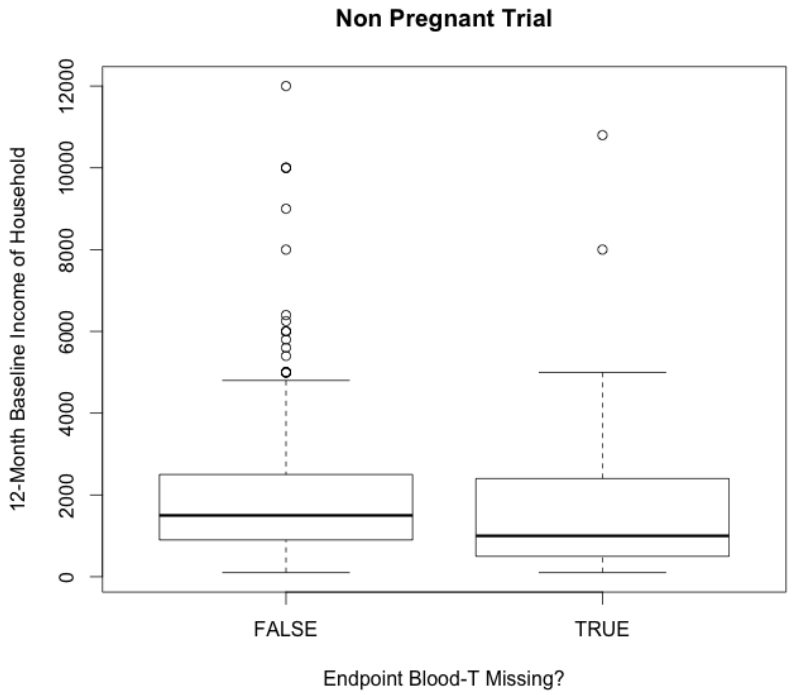
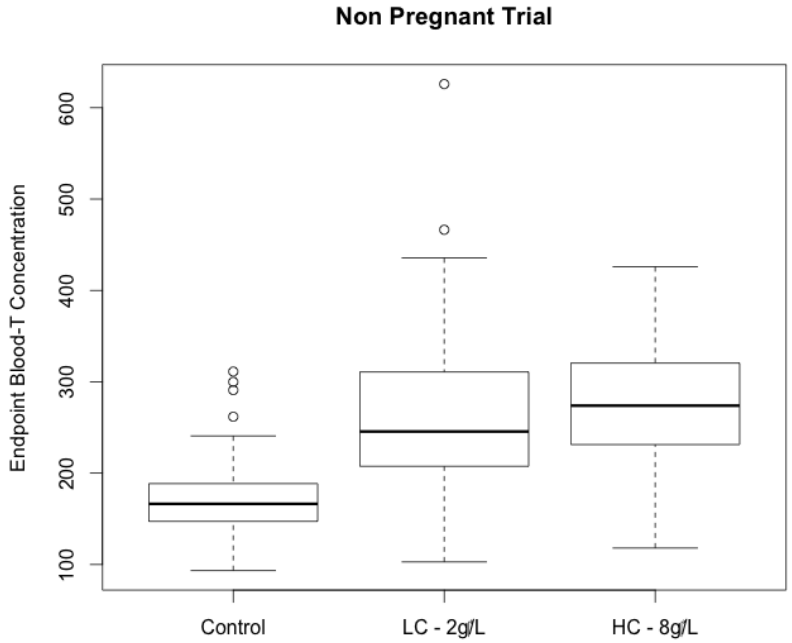


Figure 2: Boxplots for endpoint blood thiamine concentrations for the non-pregnant trial. It can be seen that average endpoint blood thiamine concentration is greater for the non-control treatment groups.



3 Statistical Questions

In this section, some preliminary statistical concerns are addressed. The proposed approach that is outlined in Section 4 involves the use of ANOVA and ANCOVA, so the following information is presented under the assumption that these models will be used. In addition, this section assumes that individuals with missing values on the variables being investigated have been removed or that values have been imputed, so that the analysis is based on a 'complete' data set.

3.1 Dependent Variables

It is important to specify that the dependent variable of interest should be the difference between the endpoint and baseline thiamine concentrations (breastmilk

or blood),

$$Y_{diff} = Y_{EP} - Y_{BL},$$

where Y_{EP} is the endpoint thiamine concentration and Y_{BL} is the baseline thiamine concentration. This way, individual effects are controlled for. If a particular individual is more likely to have a higher or lower blood thiamine concentration than the typical individual, this systematic difference will be eliminated when we consider the difference as our response.

To illustrate this, suppose that the systematic deviation of the blood thiamine level of individual i from μ, the mean blood thiamine level for their group, is represented by an effect α_i, both at baseline and at endpoint. Then the baseline and endpoint blood thiamine concentration of individual i can be represented using the mean baseline and endpoint blood thiamine concentrations plus the systematic effect of individual i, plus a measurement error term ϵ_i:

$$Y_i^{EP} = \mu^{EP} + \alpha_i + \epsilon_i^{EP},$$

$$Y_i^{BL} = \mu^{BL} + \alpha_i + \epsilon_i^{BL}.$$

When we take the difference between the endpoint and baseline levels for individual i, the α_i values in Y_i^{EP} and Y_i^{BL} cancel out, and we have eliminated these individual effects as a source of variation in the response. This is the reason for using the difference between endpoint and baseline measurements as our dependent variable.

There are a couple of responses, however, where baseline values are not available. These include the breast-milk thiamine concentrations for mothers and the endpoint blood thiamine concentrations for infants in the pregnant-mother clinical trial. In these cases, one possible approach for attempting to control for systematic individual-level effects involves regressing against the mother's baseline blood thiamine concentration. This will be discussed in Section 3.3.

3.2 Normality Assumptions

In ANOVA, ANCOVA, and linear regression, normality of the response is assumed and, at least strictly speaking, required for making statistical inferences (constructing confidence intervals, interpreting p-values, etc). This means that the values of the response variable, conditional on the explanatory variables, are assumed to be normally distributed. Although ANOVA, ANCOVA, and linear regression are quite robust to modest violations of normality, it is nevertheless recommended to investigate the appropriateness of the normality assumption.

Consider the case where thiamine concentration is the response variable and the only predictor (explanatory variable) is the treatment group (fish-sauce thiamine concentration), as in ANOVA. The response data might then be represented by the following linear model:

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where Y_{ij}, the response for the jth individual in the ith treatment group, is the difference between endpoint and baseline blood-thiamine (as specified previously), μ_i is the mean response for the ith treatment group, and ϵ_{ij} are independent and identically distributed error terms that have a mean of zero. The error term, ϵ_{ij}, is expressed as:

$$\epsilon_{ij} = Y_{ij} - \mu_i.$$

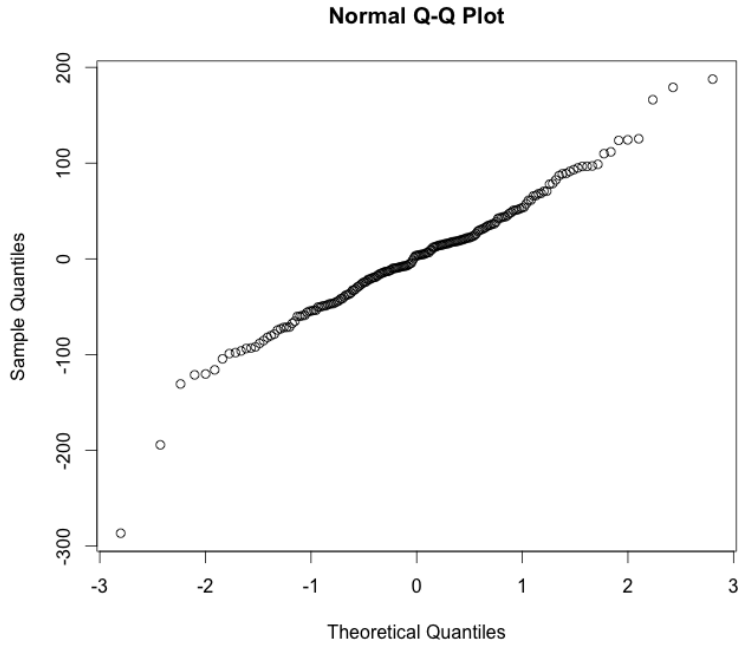
Since our estimate for μ_i is $\bar{Y}_i$, the average response for the ith treatment group, our observed residuals are:

$$r_{ij} = Y_{ij} - \bar{Y}_i.$$

The observed residual, r_{ij}, is an estimate of the error term, ϵ_{ij}. A normal QQ-plot of the observed residuals (see Figure 3) provides a check if it is reasonable to assume that the error terms are independent and identically normally distributed. The QQ-plot will look similar to a straight line if the residuals are normally distributed.

For more complicated models involving more than one explanatory variable, such as ANCOVA, the residuals are expressed slightly differently, but the same general idea applies.

Figure 3: Normal QQ-plot of the residuals from an ANOVA model with mother's blood thiamine concentration difference as the response and treatment group as the explanatory variable, exemplifying normally distributed residuals. This plot was generated using your data from the non-pregnant trial.



When the residuals are drastically non-normal, a Box-Cox transformation can be applied to make them closer to normal. The Box-Cox method finds a transformation that results in approximately normally distributed residuals, but this can hinder interpretation of the results of the analysis. For example, suppose that from an ANOVA context, the Box-Cox method suggests a transformation of the dependent variable, Y, by a power of 1.5 will result in residuals that can be reasonably approximated as normally distributed. ANOVA on these transformed responses no longer tests the differences between the means of the dependent variable; instead ANOVA now tests for differences between the means of $Y^{1.5}$. If ANOVA finds significant differences between these new means, it is difficult to extend this interpretation to the un-transformed data.

Due to results such as in Figure 3, which show that the observed residuals

appear approximately normal, transformation of the data is not recommended. However, to have confidence in subsequent statistical inferences, it is important to show that the residuals can reasonably be assumed to be normally distributed.

3.3 Differences in Baseline Means across Treatments

Randomization of the test subjects to the treatment groups is meant to control for selection bias among test subjects. This means that test subjects are not more or less likely to be placed in a particular treatment group based on the values of their other covariates. For example, after randomization, it is not expected that mothers with high baseline blood thiamine are more likely to be in the control treatment than mothers with low baseline blood thiamine. Sometimes, even after randomization, such differences between treatment groups are still observed due to chance. As differences at baseline in covariates that are predictive of the response can result in selection bias, preliminary exploration of the data should include visual examination to identify such differences (using box-plots).

There is another issue regarding differences in baseline measurements across the treatments. In this study, for example, it may be that individuals with high baseline blood thiamine concentrations are less likely to have an increase in blood thiamine by the endpoint even after consumption of fortified fish sauce. This should not mean that the treatment does not work, as the thiamine levels of the individual were already high enough to begin with and the treatment may not have been necessary for that individual.

To deal with this issue, the baseline measurements should be included as a covariate in the analysis, leading to what is sometimes called the conditional change model. This allows for the baseline measurements to account for some of the variation that cannot be attributed to the treatment effects. As mentioned earlier, there are certain cases where baseline measurements are not available (mother's breast-milk thiamine and infant blood thiamine). In these cases, it might make sense to

add mother's baseline blood thiamine as a covariate, as it is likely indicative of the theoretical baseline thiamine concentrations for mother's breastmilk thiamine and infant blood thiamine concentrations. Implementing this in the analysis requires the use of ANCOVA or linear regression models, rather than the simpler ANOVA.

4 Proposed Analysis

Before describing any statistical models, it is useful to first note the different available response (dependent) variables. There are a total of five different response variables that can be used, summarized in Table 3 below. Each of these continuous response variables will require a separate but similar analysis. The response will be referred to as "Thiamine Concentration" in all of the examples provided below; this can refer to any one of the responses listed in Table 3.

Table 3: Summary of response variables for the two clinical trials, where the number of available responses are indicated by 'n'.

Non-Pregnant Mothers	Pregnant Mothers
- Endpoint minus Baseline Blood Thiamine Concentration for Mothers ($n = 197$)	- Endpoint minus Baseline Blood Thiamine Concentration for Mothers ($n = 77$)
- Endpoint minus Baseline Blood Thiamine Concentration for Children ($n = 191$)	- Endpoint Blood Thiamine Concentration for Children ($n = 65$)
	- Endpoint Breast-Milk Thiamine Concentration for Mothers ($n = 67$)

4.1 ANOVA

Analysis of Variance (ANOVA) is a simple form of linear regression that can be used to assess differences between the means of a response in different categories. We describe the details of ANOVA for the completely randomized design used in your two clinical trials. The assumptions for ANOVA are similar to that of a linear regression model:

1. Independence of responses;
2. Approximate normality of the responses;
3. Equality of variances of the responses.

The randomization in your clinical trials bolsters the assumption that the observations are independent by balancing both known and unknown influential factors in the assignment of the treatments to the subjects. Nevertheless, known factors may be unbalanced after randomization, and therefore it is useful to check this independence assumption by plotting the residuals against other covariates. If a pattern in the distribution of residuals is observed, it may be that this independence assumption is being violated. For example, randomization is supposed to balance the age distribution of mothers across the different treatment groups. It is important to check if the distribution of mother's age is similar for each treatment group, particularly if the mother's age is predictive of the response.

Normality of the residuals should be a reasonable approximation to rely on the confidence intervals for the estimated coefficients, or the corresponding p-values (discussed later) that the software will provide when fitting a regression model. Section 3.2 described how to check this assumption using a QQ-plot.

Equality of variances of the responses in each treatment group is difficult to test. In your randomized controlled trial setting, the group sizes should be more-or-less the same, in which case this assumption is not so critical for ANOVA. Hence, visual inspection of the relative spread of the observations in each treatment

group via box-plots should suffice. More detail on this assumption is provided in Section 4.3 on linear regression.

In the context of your design, ANOVA corresponds to fitting a simple linear model with Thiamine Concentration as the dependent variable and treatment group as the independent variable, resulting in the average response in each treatment group as the predicted value for the mean Thiamine Concentration for each treatment group. The differences between each of the individual Thiamine Concentrations and the predicted mean Thiamine Concentration for the group are the 'residuals' and are estimates of the random errors ϵ_{ij}, which are supposed to be normally distributed with a mean of zero and a common variance. If μ_i is the true (unknown) mean Thiamine Concentration for the ith treatment group, the null hypothesis (H_0) assessed by the ANOVA is:

$$H_0 : \ \mu_1 = \mu_2 = \mu_3$$

with the alternative hypothesis (H_A) being that at least one of μ_1, μ_2, and μ_3 is not equal to the others.

As it is easy to implement ANOVA (for example in SPSS) without understanding what is being done, a short qualitative explanation is provided. ANOVA attempts to explain the total variability in the response by partitioning the 'Total Sum of Squares' into the sum of the variability that can be explained by differences in average levels corresponding to the treatment groups (the Sum of Squares Treatments) and the remaining variability that corresponds to random error (the Sum of Squares Residual). The ratio of the corresponding Mean Squares (Mean Square = Sum of Squares / degrees of freedom) is an 'F-statistic', which is compared to an F-distribution with the same degrees of freedom. The degrees of freedom (df) are determined by the design. In your case, $df_{treatments} = 2$ and $df_{Residual} = n-3$ where n is the number of observations of the response being analysed. This comparison yields a p-value, which is the probability under the null hypothesis of observing a value of the F-statistic at least as extreme as that for the data. If the p-value is

small, say smaller than some specified significance level α (often 0.05), we reject the null hypothesis at significance level α meaning there is strong evidence of a difference between the mean response values for the different treatment groups.

The main issue with this simple ANOVA approach is that it does not incorporate additional covariates that could potentially explain some of the variation in the data, thereby leading to a more precise assessment of the treatment effects. Adding covariates to the model would also allow for implementation of the conditional change model, described in Section 3.3. In addition, even if the normality assumption for the residuals is not reasonable under a simple linear model, such as that corresponding to the ANOVA approach, there is a chance that it will be reasonable under a more complex model with more covariates. For these reasons, we recommend implementing an ANCOVA (Analysis of Covariance) model to test the previously stated hypothesis.

4.2 ANCOVA

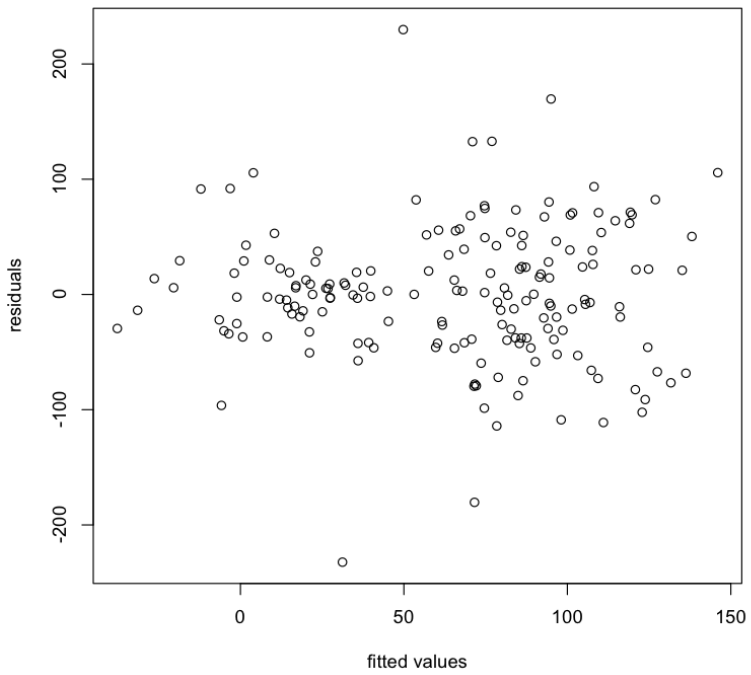
ANCOVA is an extension of ANOVA where additional covariates are added that may be affecting the response variable. Other covariates that may explain some of the variation in Thiamine Concentration, such as income or age, can be incorporated in the analysis via ANCOVA. Furthermore, ANCOVA can control for baseline differences across treatments by the addition of baseline Thiamine Concentration as a covariate, thereby implementing the conditional change model. The assumptions for this model are the same as for ANOVA, with the additional assumption that the dependent variable is linearly related to each of the independent variables. As using ANCOVA will provide a better assessment of differences across the treatment groups than ANOVA, in that additional sources of variability will be accounted for, this method is recommended for your hypothesis testing. Performing ANCOVA in SPSS is described in (2).

4.3 Linear Regression Analysis

Linear regression will fit a model for predicting a continuous response variable given a set of covariates. The assumptions for linear regression are the same as those for ANOVA and ANCOVA, as specified above.

Recall that equality of the variances of the residuals is difficult to test for, and therefore visual inspection will have to suffice. In this case, it makes sense to plot the residuals against the fitted values to check that the distribution of the residuals against the fitted values is seemingly random, with no presence of a 'funnel' shape. The plot in Figure 4 is an example for which the variances of residuals appear close enough to being equal for the equality assumption to be reasonable.

Figure 4: Plot of randomly distributed residuals against fitted values from a linear model



An important consideration for linear regression is that highly correlated independent variables should not be included in a model. The reason is that this will result in imprecise estimation of the regression coefficients (discussed later) for

these covariates. For example, the variable "$Any - School - Woman$", a binary variable corresponding to whether a mother has had any education, is expected to be highly correlated with the variable "$Highest - School - Woman$", which indicates the level of education of each mother. In the same vein, the variable corresponding to the total number of fish sauce bottles received throughout the study may be highly correlated with the variable corresponding to the total number of fish sauce bottles consumed throughout the study. In both of these cases, if highly correlated, only one of each of these variables should be included in the linear model.

Interpreting the coefficients that are obtained by fitting a linear model is the final step of the analysis. For categorical variables, such as the treatment group factor, one of the groups will always be assigned as the reference group and will not have a coefficient. This means that the coefficients for the other groups describe a comparison to that reference group. Consider the linear model coefficients in Table 4: although there are three treatment groups, coefficients appear only for the low-concentration group and the high-concentration group. The coefficients for these groups correspond to the expected difference from the control group, which is the reference group. For example, from Table 4, we see that the expected adjusted Thiamine Concentration difference for mothers in the low-concentration Thiamine-fortified fish sauce treatment group is approximately 81 units higher than for mothers in the control group. This can be interpreted as: if there are two mothers who have identical values of the other covariates, where one is assigned the low-concentration fish sauce, and the other is assigned the unfortified fish sauce, the Thiamine Concentration (in this case mother's blood thiamine) of the mother who was assigned the low-concentration fish sauce is expected to be approximately 81 units higher than that of the mother who was assigned the unfortified fish sauce, after six months. Note that because the baseline value of the response is included in the model (BL_TDP_Woman), interpretation can be in terms of Thiamine Concentration after six months rather than in terms of change

from baseline after six months.

The remaining columns in Table 4 all provide further descriptions of the coefficient values estimated by the linear model. The standard error for the coefficient corresponds to the variability of the estimated coefficient. If this standard error is high relative to the coefficient estimate, the magnitude of the t-value will be small, and the confidence interval for the coefficient will be quite wide. This is easier to understand if we consider only the p-value, which will in turn also be large if the standard error is large compared to the coefficient estimate. Each p-value in this case corresponds to a hypothesis test where the null hypothesis is that the true value of that coefficient is actually equal to zero. P-values less than some significance level α provide strong evidence that the coefficient is not equal to zero; that is, that the corresponding covariate has a (linear) relationship with the response. The smallest p-values in Table 4, corresponding to the two treatment levels, are both less than 0.001. There is therefore strong evidence that the treatment effect coefficients are not equal to zero. The coefficient corresponding to the total number of fish sauce bottles consumed, on the other hand, has a p-value of approximately 0.185. This high p-value indicates that, when the effects of the other covariates included in the model are taken into account, these data provide no evidence to suggest that the number of fish sauce bottles consumed has an effect on the Thiamine Concentration.

From Table 4, it is evident that the treatment effects are significant. Furthermore, the high concentration treatment group coefficient is slightly larger, suggesting that the effect of this treatment group might be stronger. However, upon inspection, the difference between the coefficients for the low-concentration and high-concentration treatment groups is very much smaller than the standard errors for both of these coefficients. This indicates that the difference between these coefficients is negligible; that is, the low-dose and high-dose treatments appear to be roughly equally effective. In other cases, where the difference between coefficients is not obviously negligible, hypothesis testing of whether there is a

Table 4: Sample output from a linear model for the non-pregnant trial with difference between endpoint and baseline mother's blood thiamine concentration as the dependent variable.

Coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	−17.01248	33.65077	−0.506	0.61384
ArmLC - 2g/L	81.34797	11.22746	7.245	< 0.001
ArmHC - 8g/L	82.96653	11.76621	7.051	< 0.001
BL_TDP_Woman	−0.26608	0.08781	−3.030	0.00284
Village_Number	0.75665	0.37635	2.011	0.04600
Highest_School_Woman2. Lower Secondary school	−5.55653	10.67740	−0.520	0.60348
Highest_School_Woman3. Upper Secondary school	−20.23888	15.27421	−1.325	0.18698
Highest_School_Woman4. Higher education	−122.40843	44.51842	−2.750	0.00663
BL_Woman_Age	1.32298	0.77766	1.701	0.09077
FSbottles_CONSUMED	1.37384	1.03192	1.331	0.18490

difference between the two corresponding treatments can be done by fitting an additional model that considers the responses from the low-dose and high-dose treatment groups as being from one treatment group. We refer to this model as the reduced model, as it is nested within the original model. The output of ANCOVA fits (as in Table 4) will provide values for the 'Sum of Squares Residuals' (SSR) for each of the models. In order to test the null hypothesis $H_0 : \beta_1 = \beta_2$ against $H_A : \beta_1 \neq \beta_2$, we compare the two models using the F-statistic:

$$F = \frac{(SSR_{reduced} - SSR_{original})}{SSR_{original}/(n - p - 1)}$$

where p is the number of coefficients in the original model (in the case of Table 4 the number of coefficients is 9). The value obtained for the F-statistic is compared to F_{α,v_1,v_2}, the F-value at significance level α with $v_1 = n - p - 2$ and $v_2 = n - p - 1$ degrees of freedom. If $F \geq F_{\alpha,v_1,v_2}$, we reject H_0 at significance level α and conclude that there is strong evidence that the coefficients are not equal. It should be noted that the degrees of freedom v_1 and v_2 correspond to n minus the number of coefficients (including the intercept) in the reduced and original models

respectively.

4.3.1 Intent-to-Treat versus As-Treated Analyses

Under an 'Intent-to-Treat' analysis, there is an underlying assumption that every mother consumes the amount of fish sauce in the number of bottles distributed. This is the type of analysis we have described so far in this report. However, there may be a discrepancy between the amount of fish sauce distributed and the amount of fish sauce actually consumed by each mother. An 'As-Treated' analysis uses the data collected for the number of fish sauce bottles each mother indicated she consumed to determine the 'dosage' actually received. This approach assumes that the self-reported data is reliable.

For the 'As-Treated' analysis, we suggest regressing Thiamine Concentration against the covariate for the total fish sauce reported as consumed throughout the study. One option is to consider the total fish sauce consumed as a continuous covariate, however, this would require more complicated models (e.g., perhaps using quadratic terms) in order to find the optimal dose of fish sauce. Therefore we recommend separating the data for the total fish sauce reported as consumed into an ordered factor with a small number of levels, such as:

1. 1-5 bottles consumed;
2. 5-10 bottles consumed;
3. 10-15 bottles consumed;
4. more than 15 bottles consumed.

Something like this would ease the process of finding the optimal dose of fish sauce; you could then examine which of these dosages yields the best response using the analysis approaches already described.

4.3.2 Relationships between Responses

To investigate the relationship between breast-milk, child's blood, and mother's blood Thiamine Concentrations, three different linear regression models can be fit with each of the specified variables as a response and the remaining as covariates. Other covariates can be included in each model if they are suspected to have an effect on the given response. For example, say child's blood Thiamine Concentration is being regressed against mother's blood Thiamine Concentration. One might want to include the age of the mother as a covariate in this model, as it may have an effect on the relationship between these Thiamine Concentrations. Interpretation of each of these models as above should provide an understanding of how the three variables are related.

5 Conclusion

It is important to perform exploratory data analysis to better understand the trends in the response variable between treatment levels. Differences between response distributions should be visualized before they are tested statistically. It is also essential to understand the patterns of the missing data. Imputation, specifically multiple imputation, is useful for filling in missing values, and can sometimes help eliminate some non-response bias. ANCOVA should be used to test if there are significant differences between the treatment levels. In terms of an optimal dose of Thiamine in the fish sauce, informal and formal methods for assessing the difference between the low concentration and high concentration doses are provided.

References

[1] CK Enders *Multiple Imputation in SPSS* 2010.

http://www.appliedmissingdata.com/spss-multiple-imputation.pdf

[2] Laerd Statistics *ANCOVA in SPSS* 2010.

https://statistics.laerd.com/spss-tutorials/ancova-using-spss-statistics.php