Kaylin Purvor (UG)

Newcastle University

Student Number: 110163574

Supervisor: John Matthews

An Analysis of Nutritional Intakes of Secondary School Children

An Assessment of the Impact of Revised School Food Standards

May 2015

Abstract

The aim of this project is to perform a statistical analysis of a dataset containing nutritional information on a sample of secondary school children, which was collected at two time points in different years, 2000 and 2009.

It was hoped that, as a result of revised school food standards that were implemented between 2006 and 2009, the quality of children's diets would have improved in this time period. Specifically, this referred to reducing intake of energy, sodium and saturated fat, and increasing intake of vitamin C.

This project analyses these data using linear regression techniques in order to assess the impact of the new standards. In doing so, various pitfalls and downsides of commonly used methods are identified and discussed, as well as ways in which to overcome them.

Table of Contents

1 Introduction

1.1 The childhood obesity crisis

A balanced, nutritious diet is a crucial factor for good overall health. A person's diet should provide all of the nutrients, vitamins and minerals needed for optimal growth, maintenance and repair. It is important for children in particular to maintain such a diet, as their bodies are developing and thus have even more nutrient requirements and high metabolic rates. Furthermore, a good diet can help to prevent numerous diseases, and to enhance brain development, which can result in a higher IQ and improved concentration levels (Wilson K. L., accessed 2015).

Obesity is a potential consequence of long-term poor diet, caused by energy excess. It occurs when an individual frequently consumes much more energy than they use, typically from foods that are high in fat and sugar, over a long period of time (House of Commons Health Committee, 2004, p. 3). This leads the body to store the unused excess energy as fat.

The impacts of obesity on health – both physical and psychological – are numerous. Examples of physical effects include high risk of cardiovascular disease, a range of cancers, type II diabetes, strokes, high blood pressure, osteoarthritis, fertility problems and reduced life expectancy (House of Commons Health Committee, 2004, p. 16-21). Psychological consequences include depression, anxiety and low self-esteem. (House of Commons Health Committee, 2004, p. 16-21).

In the early 2000s, there was growing national concern about childhood obesity levels, which were rapidly increasing: in 2002, it was reported that 21.8% of boys and 27.5% of girls aged 2-15 years were clinically overweight or obese (National Centre for Social Research, 2002). This was a major cause for public concern, as the adverse outcomes of obesity were well known at the time. In addition, it was estimated at the time that the direct cost of obesity to the NHS was £46-49 million per year, and that the cost of treating obesity-related conditions was £945-1075 million per year (House of Commons Health Committee, 2004, p. 21). As a result, there was nationwide public demand to improve children's diets.

1.2 The demand to review school food standards

"Every mum and dad knows that if you want your child to do well at school, and particularly to concentrate well in the classroom in the afternoon, a healthy meal at lunchtime is vital." –Nick Clegg (Gove, 2014)

In 2005, a campaign to review school food standards, and thus improve school meals, was launched by TV chef Jamie Oliver. For many pupils, school meals serve as the main meal of the day, making them a crucial source of essential nutrients (Mucavele et al., 2013, p.10). Oliver believed that if children received healthier school meals (which comprise roughly a third of daily food intake) and were educated in nutrition, they would be encouraged to make better dietary choices outside of school and in adulthood. He believed that improving the quality of school meals was a means to improve children's diets overall and to tackle the obesity crisis (Jamie Oliver Food Foundation, 2015).

The very first school food standards were introduced in 1941, but were removed in 1980 to cut government expenditure in schools – permitting cheap, but unhealthy, 'convenience food' was seen as a way to do this (School Meals Review Panel, 2005, p.17-18). It was not until 2001 that standards were re-established, but the adequacy of these standards was widely criticized after the broadcast of Oliver's documentary series, *Jamie's School dinners*, which subsequently led to his campaign, entitled *Feed Me Better*. His campaign made an online petition to thoroughly re-assess national school food standards and to commit long-term funding to this initiative. Signed by 271,677 people, the petition was delivered to 10 Downing Street on 20th March 2005 (Jamie's School Dinners, 2015).

In response to the campaign and to other similar pressure groups, the Department for Education and Skills created the School Food Trust (now known as the Children's Food Trust), a £60 million trust fund for the initiative. They sought recommendations for renewing the standards from an association called the School Meals Review Panel, which consisted of a variety of professions including dieticians, head teachers, school caterers, parents and governors (School Meals Review Panel, 2005, p.16). The Panel constructed 14 nutrient-based standards and 9 food-based standards, which the government implemented in schools across the country between September 2006 and September 2009 (School Meals Review Panel, 2005, p.27-29). These standards are given in Appendix 7.1.

The aim of the initiative was to improve school meals, and ultimately, the standard of children's diets in general. There was particular emphasis on reducing intake of energy, salt and saturated fat and increasing intake of vitamin C. It was therefore essential to study the average nutrient intake of schoolchildren before and after the introduction of revised standards, to determine whether the initiative had been effective.

Much evaluation of the impact of these standards has already been carried out, with positive findings (Gove, 2014; Mucavele et al., 2013, p.10). For instance, the number of primary school children eating the required amount of vegetables increased from 59% to 74% between 2005 and 2009, and secondary school children in 2011 were found to consume 30% less sugar, salt, fat and saturated fat in their school meals than in 2004 (Department for Education, 2014, p.4). Overall, the initiative was largely considered to be successful in improving the nutritional quality of school meals, although there were a number of operational drawbacks, such as the standards being too restrictive, too confusing and too expensive for school chefs to implement (Mucavele et al., 2013, p.4).

1.3 Project objectives

The purpose of this project is to perform a statistical analysis on the impact of the revised standards, using data collected from schools in the North East of England. Data were collected from both primary and secondary schools, to be considered separately. This project focusses only on the secondary school data, as the time-constraint of the project does not permit a thorough analysis of both.

The key objectives of the project are:

- To assess the efficacy of the initiative by comparing the consumption of various nutrients, pre- and post- implementation of new standards
- To establish whether other variables (such as if a child has school or packed lunch) have an impact on the effectiveness of the standards, and to investigate such impacts
- To present the findings and results in a clear and comprehensible format

1.4 The data

There were 6 secondary schools that participated in this study. A cross-sectional study design was used, where surveys were carried out at two separate time-points, one in 2000 and one in 2009, to represent pre- and post-changes. In total, 513 children across the 6 schools and the 2 time-points were surveyed.

The surveys involved collecting nutritional data from the participants via 'food diaries' – self-written records of everything consumed by the individual, over two 4-day periods. Nutritionists then converted the lists of food in these diaries to numerical quantities representing the mean daily intake of various macro- and micro-nutrients. In addition, they looked at what was eaten specifically at lunch time, and subsequently quantified the mean lunchtime intake of these nutrients. The quantified nutrients are listed below, including their units of measurement:

- Energy (kcal)
- Carbohydrates (g)
- Protein (g)
- Fat (g)
- Saturated fat (g)
- Non-milk extrinsic sugars (mg)

MACRO-NUTRIENTS

- Sodium (mg)
- Vitamin C (mg)
- Iron (mg)

MICRO-NUTRIENTS

In addition to estimating each child's daily and lunchtime intake of the above nutrients, nutritionists also estimated the percentage of their total energy intake that was accounted for by each nutrient.

General personal data were also collected from the subjects, including their sex and post code of their home address.

Furthermore, parents of secondary school children have the choice of providing them with a packed lunch or signing them up to receive school lunches, so each subject was labelled as either a school or packed lunch participant. This information was important because the revised school food standards obviously apply only to school lunches, and will not have had a direct impact on packed lunches, making it necessary to take account of lunch type.

2 Simple analysis – illustrative example

This chapter comprises of some initial, basic analysis of the energy intake data. The purpose of the chapter is purely to illustrate the shortcomings of naive analyses, so only the data for lunchtime energy intake will be

considered, as a demonstrative example. Due to soaring obesity levels, children's energy intake was considered too high and so the standards aimed to cause a reduction.

2.1 Effects of year

The effect of year on lunchtime energy intake is the key indicator of the impact of the new standards, as any substantial changes that occurred between 2000 and 2009 were most likely due to their implementation. The most basic form of analysis is simply to compare the mean energy intake from 2000 with the mean energy intake from 2009, and observe the difference. Table 2.1 shows these means, calculated from the data.

	2000 (pre-implementation) $\boldsymbol{n = 298}$	**2009 (post-implementation)** $\boldsymbol{n = 215}$
Average lunchtime energy intake (kcal)	692.5	545.8

Table 2.1: *Mean lunchtime energy intake (kcal) of participants in 2000 and those in 2009, representing pre- and post-implementation of standards. The sample sizes in each year are also given, denoted by n*

Average lunchtime energy intake decreased by 146.7 kcal from 2000 to 2009, which is a fairly large reduction, in terms of the RDA (recommended daily amounts) for 11-14 year olds. The RDA for all nutrients is shown in Appendix 7.2. Roughly speaking, lunch should provide about a third of these values. Hence, boys should not consume more than $2220/3 = 720$ kcal during lunch and girls should not exceed $1845/3 = 615$ kcal.

The 2000 mean of 692.5 kcal is for both sexes pooled together, but in either case, this value is either much too large or bordering on being too large. In contrast, the 2009 mean of 545.8 kcal is well within the recommended limits for both sexes. This simple analysis indicates that children consumed considerably less food at lunch in 2009 than 2000, which suggests that there has been an improvement in the calorie content of lunches.

However, this approach is much too simple to provide worthwhile conclusions. For instance, other factors such as lunch type have not been taken into account, as school food regulations only apply to school lunches. This is dealt with in the next section.

2.2 Effects of lunch type on energy intake

As aforementioned, packed lunches will not have changed as a direct result of the revised standards unlike school lunches. It is indeed possible that some parents could have altered the contents of their packed lunches, in accordance with public concern, but *a priori*, the standards were mainly expected to affect the school lunch children. In any case, the effects of lunch type require investigation.

The mean lunchtime energy intake for each lunch type in each year are listed separately in Table 2.2.

	2000	**2009**	**Difference 2009-2000**
School lunch	711.9	495.9	-216.0
Packed lunch	612.3	574.2	-38.2

Table 2.2: *Mean lunchtime energy intake in kcal for school lunch and packed lunch participants separately in 2000 and 2009, along with the difference between the years*

As predicted, the decline in average energy intake was substantial in school lunch children, and slight in packed lunch children. This suggests that the standards made a big improvement to the calorie content of school lunches, but they appear to have affected packed lunches less.

This analysis is still too simple, however, to provide meaningful conclusions. Although lunch type has now been investigated, no consideration has been given to the sex of the participants, which is another factor affecting energy intake. Even though this study places no concern on differences between boys' and girls' nutrient intake – in fact, nutritionists would prefer to pool the sexes together for their analyses – it is known and accepted that a difference exists, and a problem arises when it is not accounted for. The details of this are explained in the next section.

2.3 The problem caused by imbalance in sex ratio

A person's sex affects how much energy they consume; specifically boys tend to eat more than girls. This is illustrated generally by the fact that boys have a greater RDA for energy, and for these data specifically, by Tables 2.3 and 2.4, which display the mean energy intake separately for each sex, averaged over the two time points.

	Males $\boldsymbol{n = 247}$	**Females** $\boldsymbol{n = 266}$
Average lunchtime energy intake (kcal)	645.0	618.0

Table 2.3: *Mean lunchtime energy intake (kcal) separately for males and females, averaged over the 2 years*

	2000		**2009**	
	Males $n = 139$	**Females** $n = 159$	**Males** $n = 108$	**Females** $n = 107$
School lunch	731.3	697.1	467.4	527.4
Packed lunch	653.7	549.4	615.2	534.9

Table 2.4: *Mean lunchtime energy intake (kcal) of males and females, categorised by lunch type and year*

Boys clearly consumed more energy than girls (in all but one of the categories – 2009 school lunch). This reinforces the existence of a sex effect but since there is no interest in comparing the boy means with girl means, because the sex effect is not related to the implementation of new standards, this effect needs to be corrected for. The comparisons of interest are the ones between the four treatment combinations:

- 2000 school lunch
- 2000 packed lunch
- 2009 school lunch
- 2009 packed lunch

Since energy intake depends on sex, these groups are only comparable if they contain the same ratio of boys to girls.

For instance, if a group contained mostly boys, then its sample mean for energy intake would probably be higher to reflect this large proportion of boys. Comparing this group to one that contains mostly girls, or with any slight different sex ratio for that matter, would lead to distorted results: the effects of year and lunch type would be confounded with the effects of sex and it would not be possible to know how much of the differences to attribute to each. Hence, these naïve comparisons between imbalanced groups can produce misleading results that do not accurately reflect the impact of the new standards.

The sex ratio of the groups must therefore be investigated. Table 2.5 shows the proportions of girls in each category. Unfortunately, none of the proportions are equal to one another. The observed means given before in Table 2.2 are therefore not comparable for investigating the differences between the groups, as speculated.

	2000	**2009**
School lunch	$136/240 = \mathbf{0.58}$	$37/78 = \mathbf{0.47}$
Packed lunch	$23/58 = \mathbf{0.40}$	$70/137 = \mathbf{0.51}$

Table 2.5: *Proportion of girls in each category*

To isolate the effects of year and lunch type in each category, the imbalances in sex ratio can be adjusted for. The method for doing so is explained in detail in the next chapter.

3 Detailed analysis – illustrative example

Having identified the shortcomings of the simple analysis, this chapter describes a more suitable method for comparing the means for the four combinations of year and lunch type, overcoming the issue of sex imbalance. Again, only the data for lunchtime energy intake shall be used to demonstrate the methodology.

3.1 The linear regression model

The first stage of this analysis is to fit a linear regression model to the data. The model contains year, lunch type and sex as covariates, as well as the year by lunch interaction. This interaction is included because it is likely that the impact of year will be greater for school lunches than for packed lunches. The other two-way interactions and the three-way interaction are not considered further.

Letting $Y_{ijk\ell}$ denote the lunchtime energy intake for the ℓ^{th} subject, who participated in the i^{th} year, had j^{th} lunch type and was from the k^{th} sex, where

$$\text{Year: } i = \begin{cases} 0, & 2000, \\ 1, & 2009, \end{cases} \qquad \text{Lunch: } j = \begin{cases} 0, & \text{school lunch}, \\ 1, & \text{packed lunch}, \end{cases} \qquad \text{Sex: } k = \begin{cases} 0, & \text{male}, \\ 1, & \text{female}, \end{cases}$$

gives the linear regression model below:

$$Y_{ijk\ell} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + \epsilon_{ijk\ell}, \tag{3.1}$$

where μ is the overall mean, α_i is the effect of the i^{th} year, β_j is the effect of the j^{th} lunch type, $(\alpha\beta)_{ij}$ is the effect of the $(ij)^{th}$ combination of year and lunch type, γ_k is the effect of the k^{th} sex and $\epsilon_{ijk\ell}$ represents the error of the ℓ^{th} individual. To ensure model identifiability (Jacquez and Greif, 1985), the parameter constraints $\alpha_0 = \beta_0 = 0$, $(\alpha\beta)_{0j} = (\alpha\beta)_{i0} = 0$ and $\gamma_0 = 0$ are imposed.

The R summary output below shows estimates for the model parameters.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    734.50      13.96  52.606  < 2e-16 ***
yearf9        -219.71      22.27  -9.867  < 2e-16 ***
lunchf1       -106.37      25.08  -4.241 2.65e-05 ***
genderf1       -39.88      15.16  -2.631  0.00878 **
yearf9:lunchf1 186.12      34.89   5.335 1.44e-07 ***
```

For each covariate, R has produced an estimate of its coefficient along with its standard error. It has also provided a p-value for a two-tailed Student's t-test, whose null hypothesis is that the population means for the levels of each factor are equal. Tests cannot prove hypotheses with certainty, but the p-value represents the strength of the evidence against the null hypothesis; it can be interpreted as the probability of obtaining a result that is at least as extreme on a repeated application, assuming truth of the null hypothesis. Therefore, very small p-values indicate that the differences in sample means are unlikely to have occurred by chance. P-values smaller than 0.05 are deemed statistically significant. In this case, the p-values for all the variables are very small, so each covariate is significant and should be retained in the model.

3.2 Fitted means

The parameter estimates of the regression model permit the calculation of fitted means. For an individual with $(ijk)^{th}$ combination of covariates, the fitted mean, denoted $\hat{Y}_{ijk}$, is the expected value of their lunchtime energy intake. This is obtained by taking the expectation of (3.1). The expectation of the error terms is zero since the errors follow a $N(0,1)$ distribution, and so the fitted mean for an individual of type (ijk) is

$$\begin{aligned}\hat{Y}_{ijk} = E[Y_{ijk}] = 734.50 - 219.71\, I[Year = 2009] - 106.37\, I[Lunch = school\,] \\ - 39.88\, I[Sex = female] \\ + 186.12\, I[Year = 2009\ \&\ Lunch = school].\end{aligned} \tag{3.2}$$

Here, $I[A]$ denotes an indicator function of event A, which is equal to one if A is true and equal to zero otherwise. Hence, the fitted mean for each type of individual is obtained by substituting 0 or 1 as appropriate for the indicator variables into equation (3.2).

However, the fitted means for each type of individual are not actually of interest – the study is focussed on comparing the combinations of year and lunch type, with no relevance attached to the sex of an individual. Hence, it is the fitted mean for a $(ij)^{th}$ type of individual, denoted $\hat{Y}_{ij}$, which is required. These are calculated by substituting either 0 or 1 as appropriate for year and lunch type, and then averaging over sex, by using the proportion of females in the category for the sex value. In other words, the fitted mean of an individual of type (ij) is given by

$$\begin{aligned}\hat{Y}_{ij} = 734.50 - 219.71\, I[Year = 2009] - 106.37\, I[Lunch = packed\,] \\ - 39.88 FemaleProportion \\ + 186.12\, I[Year = 2009\ \&\ Lunch = packed].\end{aligned} \tag{3.3}$$

Using equation (3.3), with the appropriate female proportions from each category, the calculations for the fitted means of the groups are shown below.

(I) 2000 school lunch

$$\begin{aligned}\hat{y}_{0,0} &= 734.50 - (219.71 \times 0) - (106.37 \times 0) - \left(39.88 \times {}^{136}/_{240}\right) + (186.12 \times 0 \times 0) \\ &= 711.9013\end{aligned}$$

(II) 2000 packed lunch

$$\begin{aligned}\hat{y}_{0,1} &= 734.50 - (219.71 \times 0) - (106.37 \times 1) - \left(39.88 \times {}^{23}/_{58}\right) + (186.12 \times 0 \times 1) \\ &= 612.3155\end{aligned}$$

(III) 2009 school lunch

$$\begin{aligned}\hat{y}_{1,0} &= 734.50 - (219.71 \times 1) - (106.37 \times 0) - \left(39.88 \times {}^{37}/_{78}\right) + (186.12 \times 1 \times 0) \\ &= 495.8726\end{aligned}$$

(IV) 2009 packed lunch

$$\begin{aligned}\hat{y}_{1,1} &= 734.50 - (219.71 \times 1) - (106.37 \times 1) - \left(39.88 \times {}^{70}/_{137}\right) + (186.12 \times 1 \times 1) \\ &= 574.1634\end{aligned}$$

These fitted means are displayed in Table 3.1, next to the corresponding raw means from the previous chapter (from Table 2.2). On comparison, the fitted means appear to be very similar to the observed means. The values have been given to 4 decimal places so that the differences can be observed, as they are in fact identical to 2 decimal places. This indicates that this model provides an excellent fit to the data.

	2000		2009	
	Observed mean	**Fitted mean**	**Observed mean**	**Fitted mean**
School lunch	711.9010	711.9013	495.8709	495.8726
Packed lunch	612.3197	612.3155	574.1635	574.1634

Table 3.1: *Fitted means and observed means (from Table 2.2) of lunchtime energy intake (kcal) for each category*

If the saturated model had been used instead – a model including all terms and all interactions – the observed and fitted means would actually be identical, as all variation would be accounted for. The removal of a term (or terms) causes the fitted means to deviate from the observed means; with the extent of deviation depending on the importance of the removed variable(s).

It is evident from the closely matched values in Table 3.1 that the omission of the various two-way interactions and the three- way interaction has only caused minute departures from the observed means. Therefore, these interactions clearly do not account for much of the variation and are unimportant. It is therefore sensible not to include them for the sake of a simpler model.

Although the fitted means closely match the observed means, this does not address the problem of the varying sex ratio between groups, as shown in calculations (I) – (IV) above. This effect can be annulled by adjusting the means for the sex variable, which is demonstrated in the next section.

3.3 Adjusted means (or least squares means)

Having identified the need to correct for the sex imbalances, the concept of adjusted means is now introduced. Adjusted means are within-group averages, for which the imbalance in a certain covariate has been corrected (Dallal, 2001). They are based on the estimated value of the model parameters, and so they are also known as least squares means (Dallal, 2001).

3.3.1 Manual calculation

The fitted means for group (ij) were calculated using the female proportion of the group as the sex value, which were inconsistent amongst the groups. In contrast, the least squares means are found by using a *fixed* value of sex (usually the mean value) for all categories (Dallal, 2001). Letting $m = 0.5185185$, the mean sex value from the dataset, the adjusted mean for an $(ij)^{th}$ covariate combination is calculated using the following equation.

$$\hat{Y}_{ij} = 734.50 - 219.71\, I[Year = 2009] - 106.37\, I[Lunch = packed\,] - 39.88m + 186.12\, I[Year = 2009\ \&\ Lunch\ type = packed]. \quad (3.4)$$

Thus, the calculations for the least squares means are as follows.

(V) 2000 school lunch

$$\hat{y}_{0,0} = 734.50 - (219.71 \times 0) - (106.37 \times 0) - 39.88m + (186.12 \times 0 \times 0)$$
$$= 713.8215$$

(VI) 2000 packed lunch

$$\hat{y}_{0,1} = 734.50 - (219.71 \times 0) - (106.37 \times 1) - 39.88m + (186.12 \times 0 \times 1)$$
$$= 607.4515$$

(VII) 2009 school lunch

$$\hat{y}_{1,0} = 734.50 - (219.71 \times 1) - (106.37 \times 0) - 39.88m + (186.12 \times 1 \times 0)$$
$$= 494.1115$$

(VIII) 2009 packed lunch

$$\hat{y}_{1,1} = 734.50 - (219.71 \times 1) - (106.37 \times 1) - 39.88m + (186.12 \times 1 \times 1)$$
$$= 573.8615$$

These means are shown in Table 3.2, alongside the corresponding fitted means from Table 3.1.

	2000		2009	
	Fitted	**Least squares**	**Fitted**	**Least squares**
School lunch	711.90	713.82	495.87	494.11
Packed lunch	612.32	607.45	574.16	573.86

Table 3.2: *Least squares means, adjusted for sex, of lunchtime energy intake (kcal) for each category, alongside the fitted means calculated previously (Table 3.1)*

On comparison to the fitted means calculated beforehand, there have been fairly notable shifts in each category, suggesting that the unadjusted means were indeed skewed by sex imbalances. For instance, the raw mean of 612.3 kcal for the 2000-packed lunch group, which contained a small female proportion of less than 40%, was likely to be dominated by the boys' intake and thus be too large. As expected, the least squares mean has corrected for this by decreasing the mean to 607.45 kcal. Similar effects have occurred in the other groups in accordance with their proportions of males and females.

Thus, despite the sex imbalance in the groups, these least squares means for the year and lunch type combinations have the desired comparability.

3.3.2 Calculation by `lsmeans`

A package called `lsmeans` can be downloaded in R to compute the least squares means more efficiently. The output summary below gives the least squares means for lunchtime energy intake, adjusted for sex. The model has been called 'energy' and the year and lunch type variables (called yearf and lunchf to show that they have been defined as factors) are included in the formula. This instructs the program to compute least squares means for combinations of these factors, adjusting for any leftover covariate(s) in the model – in this case sex.

```
> lsmeans(energy,~yearf+lunchf)
 yearf lunchf   lsmean       SE  df lower.CL upper.CL
 0     0      714.5599 11.05279 508 692.8451 736.2747
 9     0      494.8482 19.31055 508 456.9099 532.7866
 0     1      608.1938 22.44414 508 564.0990 652.2885
 9     1      574.6002 14.56874 508 545.9778 603.2226
```

These adjusted means are displayed in Table 3.3, next to the manually derived adjusted means that were given in Table 3.2.

	2000		2009	
	`lsmeans`	**Manual**	**`lsmeans`**	**Manual**
School lunch	714.56	713.82	494.85	494.11
Packed lunch	608.19	607.45	574.60	573.86

Table 3.3: *Least squares means, adjusted for sex, computed manually alongside those computed by* `lsmeans`

It appears that the least squares means calculated by `lsmeans` differ from those obtained manually. The differences are small, but for each group, they differ by the same amount, which requires further investigation:

- $714.56 - 713.82 = 0.74$
- $494.84 - 494.11 = 0.74$
- $608.19 - 607.45 = 0.74$
- $574.60 - 573.86 = 0.74$

3.3.3 The difference between the manual method and `lsmeans`

Recall that the manual method removed the sex effect by using a constant value of m for all groups. This gave predictions for lunchtime energy intake at this uniform sex value. The mean sex value from the data set was used as the value of m, but in actual fact any choice of m would be possible. Any m would succeed in removing the sex effect, provided it is fixed, thus any m would achieve the desired group-comparability.

Using different values of m would inevitably produce different adjusted means for the four groups, since they would substitute different values into (3.4). This is an identifiability issue – the adjusted means obviously vary depending on the value of m, making them not unique (Jacquez and Greif 1985). As such, they are not estimable quantities. In essence, the adjusted means themselves are unsuitable to compare due to their dependence on the subjective choice of m.

However, this study is concerned with the comparison of groups, and the *differences* between the group means are in fact unique. This is because when one adjusted mean is subtracted from another to obtain the difference, the m's cancel out, and so the differences are independent of m. These are therefore estimable quantities that can be given to the nutritionists for assessment.

Consequently, the choice of m is arbitrary. With that said however, a choice of m that produces plausible means is favourable, because they are to be assessed by nutritionists, who may not have an in-depth knowledge of statistics and may be disconcerted by implausible mean intakes.

The discrepancies between the manually adjusted means and those found by `lsmeans` are explained by the fact it applies a different choices of m. This section now proceeds to show which value is used by `lsmeans`.

For models where all covariates are categorical, the `lsmeans` command begins by calculating the expected response for all combinations of all factor levels. In this case, it computes the expected lunchtime energy intake for each combination of year, lunch type and gender, of which there are eight. This is shown in the following output.

```
> summary(ref.grid(energy))
 yearf lunchf genderf prediction       SE  df
 0     0      0         734.5017 13.96239 508
 9     0      0         514.7901 20.60255 508
 0     1      0         628.1356 23.18240 508
 9     1      0         594.5420 16.49927 508
 0     0      1         694.6180 12.81805 508
 9     0      1         474.9064 20.88665 508
 0     1      1         588.2519 24.18631 508
 9     1      1         554.6584 16.34603 508
```

Thus, for each group (i.e. each year and lunch combination), there are two associated predictions instead of one – one for males and one for females. When `lsmeans` is called, by default it combines these two predictions by computing their simple average, giving a least squares mean for each group. Essentially, `lsmeans` adjusts the group means for sex by averaging the mean of the boys' intakes and the mean of the girls' intakes. For example, `lsmeans` gives the least squares mean for the 2000-school lunch group (year = 0 and lunch = 0) by averaging the boy mean from this group and the girl mean from the group (highlighted in yellow above). Thus the adjusted mean lunchtime energy intake in 2000 for school lunch children is

$$\hat{Y}_{LSMEANS,0,0} = \frac{734.5017 + 694.6180}{2} = 714.56$$

This matches the adjusted mean obtained from `lsmeans` given in Table 3.3. Similar calculations show that this is true for each of the groups, which confirms that this is indeed the method used by the `lsmeans` command. This method is equivalent to using m = ½ in (3.4), the proof of which is as follows.

Proof. The `lsmeans` package computes the least squares means for a group of type (ij) by averaging the boy and girl intakes of that group, i.e.

$$\begin{aligned}
\hat{Y}_{LSMEANS,ij} &= \frac{1}{2}(\hat{Y}_{ij0} + \hat{Y}_{ij1}) \\
&= \frac{1}{2}\left\{\left(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} + \hat{\gamma}_0\right) + \left(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} + \hat{\gamma}_1\right)\right\} \\
&= \frac{1}{2}\left(2\hat{\mu} + 2\hat{\alpha}_i + 2\hat{\beta}_j + 2\widehat{(\alpha\beta)}_{ij} + \hat{\gamma}_0 + \hat{\gamma}_1\right) \\
&= \frac{1}{2}\left(2\hat{\mu} + 2\hat{\alpha}_i + 2\hat{\beta}_j + 2\widehat{(\alpha\beta)}_{ij} + 0 + \hat{\gamma}\right)
\end{aligned}$$

(where the parameter constraint $\hat{\gamma}_0 = 0$ has been imposed and so the gender effect is equal to the female effect, i.e. $\hat{\gamma} = \hat{\gamma}_1$)

$$= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} + \frac{\hat{\gamma}}{2}. \qquad (3.5)$$

The adjusted mean for the $(ij)^{th}$ group can also be given, as before, by

$$\hat{Y}_{MANUAL,ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} + m\hat{\gamma}. \qquad (3.6)$$

Equating (3.5) and (3.6) gives

$$\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} + \frac{\hat{\gamma}}{2} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} + m\hat{\gamma}$$

$$\Rightarrow \frac{\hat{\gamma}}{2} = m\hat{\gamma}$$

$$\Rightarrow m = {}^{1}/_{2}. \qquad \square$$

The contrasting choices, $m_{MANUAL} = 0.5185185$ and $m_{LSMEANS} = 0.5$, account for the discrepancies in the least squares means from the two methods.

Since both choices of m give plausible values, neither option is preferable to the other. So, for ease of calculation, `lsmeans` will be used throughout the rest of this project, as it is more time-efficient than manual computation.

3.4 Diagnostic checks

Diagnostic checks of the model must be performed throughout the analysis, to test its adequacy. This is because the multiple linear regression model relies on some fundamental assumptions (Poole & O'Farrell 1970), which will be discussed in the upcoming section, and any inferences made from the model are not valid unless these assumptions are satisfied.

3.4.1 Assumptions of the multiple linear regression model

Note that in this section, subscript i now indexes all data points, regardless of covariate value, instead of denoting the year variable as in the previous section.

The underlying assumptions of linear regression are as follows.

1. Linearity of the parameters. It is assumed that the response variable, Y, is a linear combination of the parameters, $\beta_0, \beta_1, \dots, \beta_p$, but not necessarily a linear combination of the explanatory variables $X_1, X_2, \dots, X_p$ (Williams et al., 2013). If this assumption is not met, then the coefficients can lead to inaccurate conclusions about the relationships between the variables.
2. Zero conditional mean of errors. For any combination of covariate values, the errors (defined later) are assumed to have a mean of zero. Breach of this assumption can cause the coefficients to be biased (Williams et al., 2013). A reason for which this may occur is violation of the linearity assumption above.
3. Independence of errors. It is assumed that the error terms are independent of one another, otherwise the estimates of the standard errors of the coefficients and significance may be biased. The estimates of the coefficients themselves would remain unbiased, despite violation of the assumption, but would be inefficient. Note that the assumption only requires independence of the errors – not the observations themselves (Williams et al., 2013).

4. Homoscedasticity of errors. The errors are assumed to have an unknown, constant variance across all levels of the explanatory variables. Violation of this assumption would not cause the ordinary least squares estimates to be biased or inconsistent, but would make them inefficient (Williams et al., 2013).
5. Normality of errors. The error terms are assumed to follow a Normal distribution. This assumption is not actually required to provide unbiased, consistent and efficient coefficient estimates, but rather to make significant tests and confidence intervals valid. Due to the central limit theorem, this assumption becomes less important as the sample size is increased (Williams et al., 2013).

The error terms are defined as the differences between the observed values of the response variable and the values of the response predicted by the true regression model for the whole population. Since the true values of the parameters are rarely known, the errors cannot usually be calculated. It is therefore difficult to check that the assumptions on the errors hold. Instead, the assumptions can be checked for the estimated residuals, denoted by e_i, which are the differences between the observed response values and the values predicted by the estimated regression model. If the assumptions hold for the estimated residuals, it is reasonable to assume that they hold for the errors as well.

However, it is not normally the raw estimated residuals themselves that are examined, but a standardized version instead. In this analysis, the studentized residuals will be used, given by

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}},$$

where MSE denotes the mean square error and h_{ii} is known as the leverage, given by the i^{th} diagonal entry of the hat matrix, H

$$H = X(X^T X)^{-1} X^T$$

where X is the data matrix.

Checking for homoscedasticity of residuals

To check for homoscedasticity of the residuals, scatterplots will be produced of the (studentized) residuals against the fitted values, and against each of the covariates. For the assumption to hold, the residuals should be randomly scattered around zero, throughout the full length of the plot. The appearance of any systematic trend or pattern would indicate a departure from the assumption, suggesting non-constant variance, which would weaken, but not completely invalidate, the analysis.

Continuing with lunchtime energy intake as an example, the plots of the residuals of these data against the fitted values and covariates are shown in Figure 3.1. A graphical technique called 'jittering' has been implemented, whereby a random variable generated from a Uniform $(-0.1, 0.1)$ distribution is added to each data point. This is because the covariates, year and lunch type, are 2-level factors, thus can only take one of two values. Consequently, simply plotting them against the residuals does not produce random scatter across the plots, but instead produces a 'column' of points at each value, which make the data difficult to view.

Similarly, the fitted values can only take a limited number of values, for each combination of levels. Jittering the data means that the covariates and fitted values are no longer restricted to limited values, but are only changed inconsequentially. The jittered residual plots in Figure 3.1 show randomly scattered clouds' of points around each value, with no apparent pattern or trend. Therefore, there is no evidence to suggest that the variances are not constant, so the assumption holds for lunchtime energy intake.

Checking for Normality of residuals

The assumption of Normality can be checked by producing a Normal probability plot. This is done by plotting the ordered residuals of the sample against the corresponding order statistics from a standard Normal distribution. The assumption of Normality is justified if the points fit reasonably well to a straight line; large deviations from the line are evidence to suggest non-Normality. Mild non-Normality is safe to ignore, but strong non-Normality should be addressed.

For lunchtime energy intake, the Normal probability plot is shown in Figure 3.2. Most of the points lie on the straight line, with a few small deviations at the tails, but overall there is no apparent curvature and there is no reason to suspect non-Normality of the residuals. Thus this assumption holds for lunchtime energy intake.

3.4.2 Influential observations

As well as checking some of the model assumptions on the errors, it is important to check for the presence of influential observations, as part of the assessment of model adequacy. Influential observations are types of outliers that not only diverge from the general pattern, but significantly affect the parameter estimates. They are therefore considered disadvantageous to the analysis, because ideally, conclusions should not be dominated by individual observations that are extreme and not representative of the rest of the population.

To detect influential observations, outliers that seem potentially influential should be flagged up and temporarily removed. If performing the regression without these points has a sizeable effect on the regression coefficients, this suggests that they are indeed influential and their role in the analysis requires careful scrutiny.

The identification of an influential observation does not provide justification for excluding it, unless there is good reason to believe that the observation is invalid – for example, if there has been a recording error or if the subject was not from the intended population (Williams et al., 2013).

The discrepancy between the current parameter estimate and the estimate obtained if that value was removed, can be measured by Cook's distance. Plotting the points against their Cook's distance is a good indicator of potentially influential observations; points with large distances in comparison to the others should be investigated.

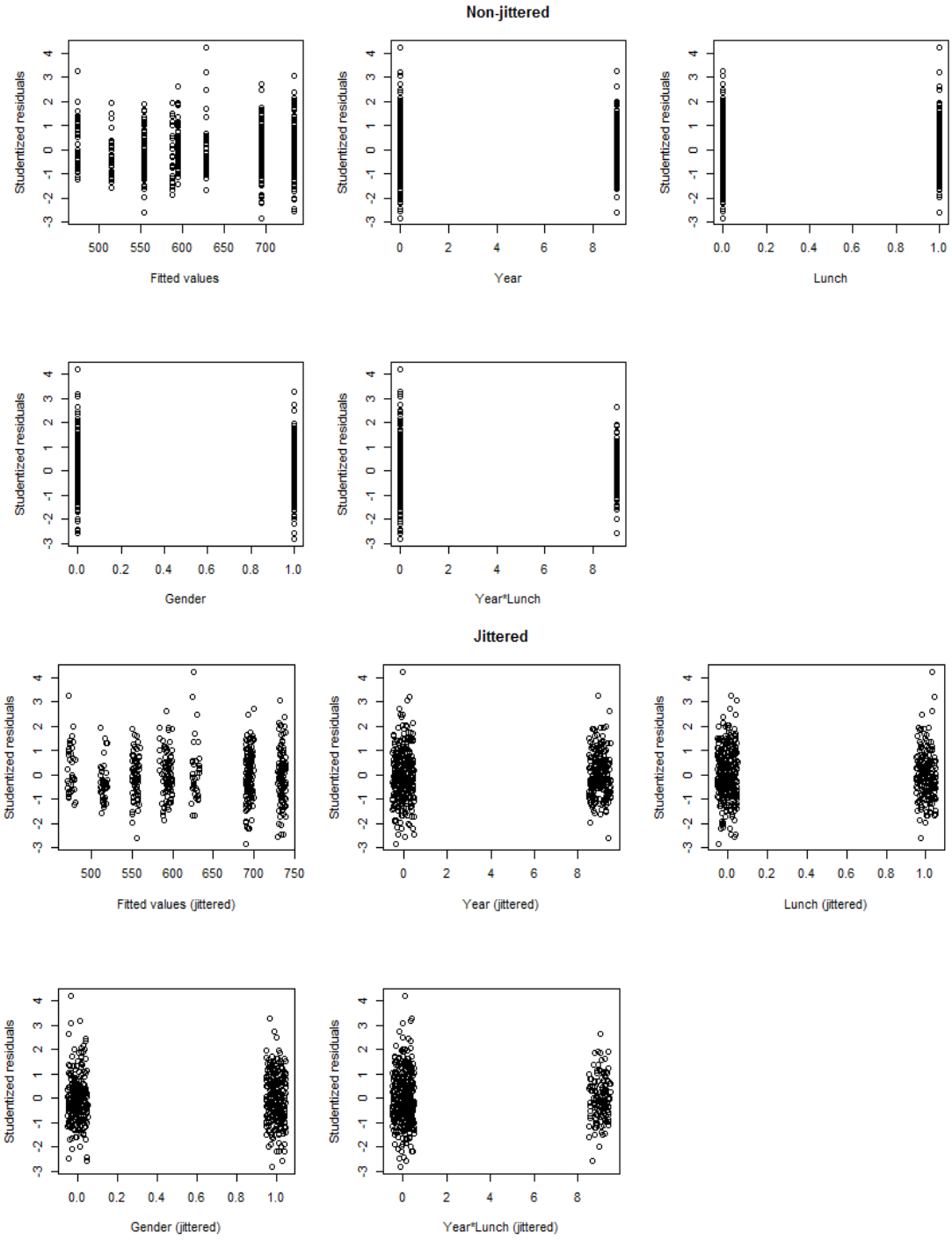

***Figure 3.1**: Residual plots for lunchtime energy intake*

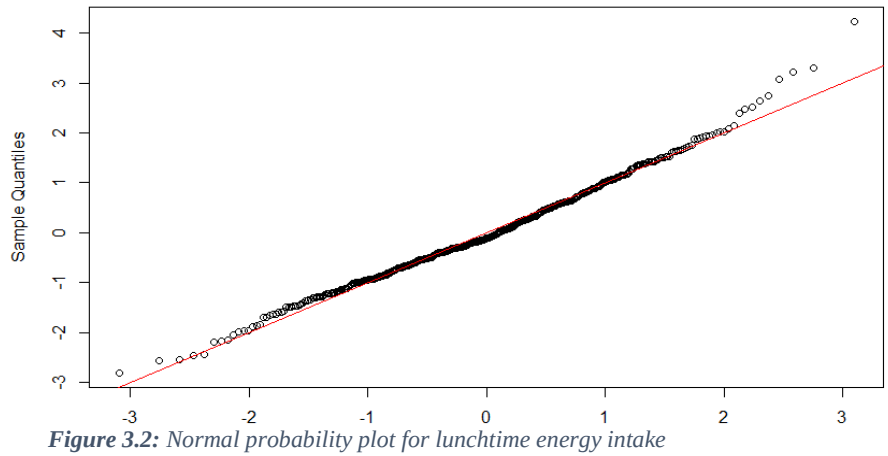

***Figure 3.2:** Normal probability plot for lunchtime energy intake*

Returning to the lunchtime energy intake

data as an example, Figure 3.3 shows the plot of their Cook's distances. Evidently, the 207th observation has a large value of Cook's distance relative to the rest, so it is sensible to investigate the effects of removing it. Table 3.4 compares the original parameter estimates with the coefficients obtained having excluded the point in question.

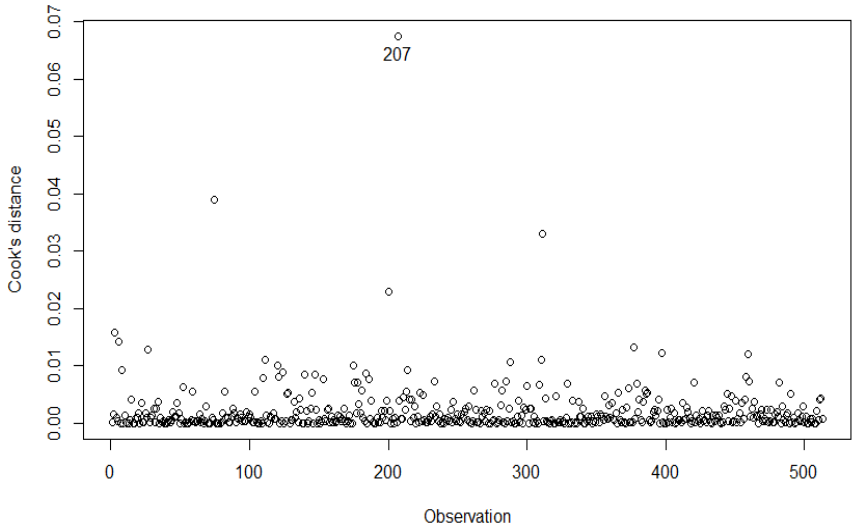

Figure 3.3: *Observations plotted against their Cook's distance*

	Coefficients based on total data set	**Coefficients having omitted observation 207**	**Difference**
Intercept	734.50	733.21	-1.29
Year	-219.71	-219.50	0.21
Lunch	-106.37	-118.54	-12.17
Gender	-39.88	-37.60	2.28
Year*Lunch	186.12	198.20	12.08

Table 3.4: *Coefficients from model having removed the point with large Cook's distance compared with the coefficients from the model based on total data set, along with the differences*

Most of the changes in parameter estimates are not severe and do not cause concern. The magnitude of the most drastic change is 12.17, in the coefficient for lunch type, followed by 12.08, in the coefficient for the interaction. These are actually fairly large, and perhaps warrant some further investigation. However, the 207th point will be retained in the analysis as we have no basis on which to omit it.

3.5 Summary

In summary, this chapter has addressed the problem of sex imbalance in the sample by making the groups comparable. This was achieved by fitting a linear regression model to the lunchtime energy data to obtain a regression equation, then adjusting the fitted mean of each group for the effects of sex, by fixing the sex variable at a uniform value. The choice of this value was found to be arbitrary, since it has no effect on the estimable quantities – the differences between the group means. Nevertheless, a choice that leads to credible means is preferable, so that they can be interpreted straightforwardly by nutritionists who do not have statistical backgrounds. Typically, the mean of the imbalanced variable is used as the fixed value, but the R package `lsmeans` uses 0.5 by default. Both are satisfactory choices as they lead to plausible means. For the

purpose of timesaving, `lsmeans` with default setting will be used for the upcoming analyses of other nutrients.

Diagnostic checks of model adequacy was also covered in this chapter. Assumptions 4 and 5 about the errors can be checked using residual and Normal probability plots. In addition, plots of Cook's distance were used to detect influential observations.

The methodology developed in this chapter will be implemented to analyse some of the key nutrients in the next chapter.

4 Detailed analysis of key nutrients

This chapter applies the methodology described in Chapter 3 to analyse several key nutrients: energy, sodium, saturated fat and vitamin C (other nutrients not included due to space constraints). For each nutrient, linear models will be fitted to both the lunchtime and daily intakes, enabling calculation of adjusted means for each group and then inferences will be made on their differences. Diagnostic checks will be performed throughout. The aim of this chapter is to present the findings in an accessible format for nutritionists to interpret.

4.1 Presentation of data: is the interaction significant?

Table 4.1 shows the p-values of each predictor variable, after fitting the linear regression model (3.1) to the lunchtime and daily intakes of each nutrient. The presentation of the results will vary, depending on whether the interaction term between year and lunch type is significant or not.

	Lunchtime intake			Daily intake		
	Interaction	**Year**	**Lunch**	**Interaction**	**Year**	**Lunch**
Energy	<0.0001			0.172	<0.0001	0.438
Sodium	<0.0001			0.992	<0.0001	0.066
Sat. fat	<0.0001			0.120	<0.0001	0.991
Vitamin C	0.361	0.093	0.513	0.783	0.015	0.409

Table 4.1: *p-values for predictor variables, obtained from fitting linear regression models to the data for lunchtime and daily nutrient intake*

Significant interaction term

If the interaction term is significant, this suggests that the effects of year are different for each lunch type (and conversely, the effects of lunch type are different in each year). In this case, no importance should be attached to the p-values of the year and lunch variables, which correspond to their main effects, because there is no such thing as a 'main' effect of year if it behaves differently for each lunch type (and similarly there is no 'main' effect of lunch). To this effect, if the interaction is indeed significant, this means that the year and lunch variables are clearly significant, regardless of their p-values, via their interaction. This is why the p-values for year and lunch for nutrients with significant interaction are deliberately not provided in the table above. For these nutrients, a two-way table is necessary to present the means for each year and lunch type separately, along with their differences.

Non-significant interaction term

On the contrary, if the p-value for the interaction term is larger than 0.05, then there is no evidence to suggest that year affects intake differently for each lunch type (this does not necessarily mean that this is not the case, but it means that these data do not provide enough evidence to verify this). In this case, the inclusion of an interaction term complicates the presentation and visualisation of the results, yet does not add anything worthwhile to the model. For this reason, if the interaction is not significant, then the model will be re-fitted without it. This simplifies the presentation, as a two-way table is no longer required – instead, a one-way table for the difference between years and a separate one-way table for the difference between lunch types will be used to present the means.

4.2 Energy

At the time of the new standards, childhood obesity was a prevalent problem and so it was supposed that children were exceeding the RDA of energy. For this reason, it was intended that the standards would reduce average energy intake.

4.2.1 Lunchtime energy intake

When the model is fitted to lunchtime energy intake, the significant interaction term ($p=1.44\times10^{-7}$) between year and lunch signifies that a two-way table is required to present the means. These are shown in Table 4.2, along with the differences and the 95% confidence interval of these differences.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
School lunch	714.6	494.8	-219.7	(-263.4,-176.1)
Packed lunch	608.2	574.6	-33.6	(-86.1 18.9)
Difference	-106.4	79.8	186.1	(117.6, 254.7)
95% confidence interval of difference	(-155.5, 57.2)	(32.33, 127.2)		

Table 4.2: *Least squares means, adjusted for sex, for lunchtime energy intake (kcal) of each year and lunch type combination*

It appears that between 2000 and 2009, school lunches decreased in energy content by around 220 kcal, which is quite a substantial decline, suggesting that an improvement was made to school lunches in this time period. The energy content of packed lunches decreased too, but only by around 34 kcal, which is not strong evidence for an improvement in packed lunches.

In 2000, school lunches were on average about 106 kcal more calorific than packed lunches, but by 2009 they had become less calorific than packed lunches by about 80 kcal. These are reasonably large quantities, which is moderate evidence to suggest that packed lunches were the healthier option in 2000, but school lunches were the healthier option in 2009 (in terms of calorie content).

The difference between the differences is 186.12 kcal. This is a little more complex to interpret. It means that the change in energy intake over the years was roughly 186 kcal more in school lunch children than in packed lunch children, and that the difference in energy intake between packed and school lunch children was 186 kcal more in 2000 than it was in 2009.

4.2.2 Daily energy intake

When the model is fitted to the daily energy intakes, the large p-value for the interaction term (p=0.172) suggests that it is not significant and it is therefore removed from the model. A two-way table is thus not required.

Table 4.3 shows the adjusted means for the two years, with the difference and corresponding confidence interval; Table 4.4 shows the adjusted means for the two lunch types, also with the difference and confidence interval.

	2000	**2009**	**Difference**	**95% confidence interval of difference**

Average daily energy intake	1904.9	1640.2	-264.7	(-337.7, -191.7)

Table 4.3: *Least squares means, adjusted for sex, for daily energy intake (kcal) in each year*

	School lunch	**Packed lunch**	**Difference**	**95% confidence interval of difference**
Average daily energy intake	1766.8	1778.3	11.5	(-62.8, 85.9)

Table 4.4: *Least squares means, adjusted for sex, for daily energy intake (kcal) of each lunch type*

It appears that the mean daily energy intake decreased by about 265 kcal from 2000 to 2009, which is a reasonably large reduction, relative to the RDA. This suggests that the changes to school food standards had an overall positive effect on daily energy consumption for all children, regardless of lunch type.

There is a marginal difference of just under 12 kcal between the daily energy intake of school and packed lunch children. Thus, neither lunch type appears to be the obvious better option, as there seems to be not much difference between them in terms of daily energy intake.

4.2.3 Diagnostic checks

The residual plots and the Normal probability plots for the lunchtime and daily energy analysis (not provided due to space constraints) do not give reason to suspect a departure from model assumptions. Likewise, the regression coefficients excluding points with large Cook distances do not suggest that any point is influential.

4.3 Sodium intake

As for energy intake, it was hoped that the new standards would reduce the average sodium intake of schoolchildren, as it was believed that they were consuming too much. Sodium in the diet comes largely from salt, and this has negative effects on health, such as causing high blood pressure.

4.3.1 Lunchtime sodium intake

For the lunchtime sodium intake model, the p-value for the interaction term is very small ($p=2.2\times10^{-7}$), so a two-way table is required. Table 4.5 shows the adjusted means, along with their differences and confidence intervals, and the difference between the differences.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
School lunch	891.8	515.9	-375.9	(-447.3, -304.5)
Packed lunch	958.2	881.7	-76.5	(-162.3, 9.3)
Difference	66.4	365.8	-299.4	(-411.5, -187.4)
95% confidence interval of difference	(-14.0, 146.8)	(288.3, 443.4)		

Table 4.5: *Least squares means, adjusted for sex, for lunchtime sodium intake (mg) of each year and lunch type combination*

There was a significant decline in mean lunchtime sodium intake, of almost 376 mg, from 2000 to 2009 in school lunch children, implying a big improvement was made to school lunches regarding sodium content. The mean for packed lunch children decreased by 76 mg, which is a slight reduction, giving no strong evidence of an improvement to packed lunches in the same time period.

In 2000, the average lunchtime sodium intake of school and packed lunch children in the sample differed by just 66 mg, giving no indication of a significant difference between them. However, by 2009, the mean for school lunch children was 366 mg less than that of packed lunch children, suggesting that school lunches were the healthier option in 2009 in terms of lunchtime sodium intake.

The difference between the differences is about -300 mg. This means that the change in lunchtime sodium intake from 2000 to 2009 was 300 mg more substantial in school lunch children than packed lunch children. Likewise, the difference between mean lunchtime sodium intake of school and packed lunch children was 300 mg more in 2009 than in 2000.

4.3.2 Daily sodium intake

Fitting the model to the daily sodium intake results in a large p-value for the interaction term (p=0.992), and so it is excluded. Table 4.6 shows the mean daily sodium intakes for each year; Table 4.7 shows the mean daily sodium intake for each lunch type.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Average daily sodium intake	2630.8	2150.3	-480.5	(-594.7, -366.3)

Table 4.6: *Least squares means, adjusted for sex, for daily sodium intake (mg) of each year*

	School lunch	**Packed lunch**	**Difference**	**95% confidence interval of difference**
Average daily sodium intake	2311.4	2469.7	158.3	(42.0, 274.6)

Table 4.7: *Least squares means, adjusted for sex, for daily sodium intake (mg) of each lunch type*

Due to the insignificant interaction, the daily sodium intake changes by the same amount across the years for both the lunch types. The change for both lunch types is a decrease of 481 mg, which is a considerable reduction. This is evidence to suggest that children's overall diets improved in terms of daily sodium intake in this time period.

In both 2000 and 2009, the average daily sodium intake of school lunch children was 158 mg less than that of packed lunch children. This is a reasonably large amount, so there is some evidence to suggest that having a school lunch caused children to consume less sodium per day compared to packed lunch.

4.3.3 Diagnostic checks

The plots for the diagnostic checks of both sodium models show no cause for concern. These models are therefore adequate.

4.4 Saturated fat intake

As for energy and sodium intake, the revised standards intended to reduce saturated fat intake, as it is associated with high cholesterol and cardiovascular disease.

4.4.1 Lunchtime saturated fat intake

The significant p-value of the interaction term ($p=6.22\times10^{-6}$) necessitates the two-way table to present the means, which are shown in Table 4.8.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
School lunch	10.2	6.3	-4.0	(-5.0, -3.0)
Packed lunch	9.7	9.3	-0.3	(-1.5, 0.9)
Difference	-0.6	3.1	-3.6	(-5.2, -2.1)
95% confidence interval of difference	(-1.7, 0.6)	(2.0, 4.2)		

Table 4.8: *Least squares means, adjusted for sex, for lunchtime saturated fat intake (g) of each year and lunch type combination*

For school lunch children, there has been a decrease of almost 4 grams in their lunchtime saturated fat consumption, which is considerably large relative to a third of the RDA, suggesting that the standards have improved the saturated fat content of school lunches. For packed lunch children, the mean lunchtime saturated fat intake changed by less than a third of a gram, so there is no evidence of a change to saturated fat content in packed lunches.

In 2000, the school lunch children consumed about half a gram more saturated fat during lunchtime than the packed lunch children, giving no evidence of a real difference between the lunch types, whereas in 2009, the school lunch children actually consumed about 3 grams less than the packed lunch children, which is a reasonably large amount. This suggests that the improvement to school lunches caused them to become the healthier option by 2009, in regards to saturated fat content.

The difference of differences is 3.64 grams, so the change in saturated fat intake from 2000 to 2009 was almost 4 grams more amongst school lunch children than packed lunch children – a fairly large quantity. Similarly, the difference between mean saturated fat intake of school and packed lunch children was almost 4 grams larger in 2009 than 2000.

4.4.2 Daily saturated fat intake

The large p-value for the interaction term (p=0.120) means that it can be removed from the model and that two-way tables are not necessary to display the means. The means for the years of daily saturated fat intake are displayed in Table 4.9; the means for the lunch types are shown in Table 4.10.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Average daily saturated fat intake	27.9	23.8	-4.1	(-5.6, -2.6)

Table 4.9: *Least squares means, adjusted for sex, for daily saturated fat intake (g) of each year*

	School lunch	**Packed lunch**	**Difference**	**95% confidence interval of difference**
Average daily saturated fat intake	25.2	26.4	1.2	(-0.3, 2.7)

Table 4.10: *Least squares means, adjusted for sex, for daily saturated fat intake (g) of each lunch type*

It seems that in 2009, children consumed 4 grams less saturated fat a day compared to 2000, which is not a relatively large amount, giving no strong evidence for an improvement in saturated fat intake.

School lunch children consumed around 1 gram less of saturated fat a day compared to packed lunch children, which is a marginal difference, so there is no evidence to suggest that either lunch type leads to in a smaller daily intake of saturated fat.

4.4.3 Diagnostic checks

From the plots for the diagnostic checks, these models satisfy the assumptions of linear regression.

4.5 Vitamin C intake

The revised standards aimed to increase school children's intake of vitamin C, as it is essential for the growth and repair of tissues.

4.5.1 Lunchtime vitamin C intake

The interaction term for this model has a large p-value (p=0.361) and so it is removed. Table 4.11 shows the mean lunchtime intakes of vitamin C for the years; Table 4.12 shows the means for the lunch types.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Average lunchtime vitamin C intake	27.5	34.3	6.8	(2.3, 11.30)

Table 4.11: *Least squares means, adjusted for sex, for lunchtime vitamin C intake (mg) of each year*

	School lunch	**Packed lunch**	**Difference**	**95% confidence interval of difference**
Average lunchtime vitamin C intake	28.6	33.1	4.4	(-0.2, 9.0)

Table 4.12: *Least squares means, adjusted for sex, for lunchtime vitamin C intake (mg) of each lunch type*

The average vitamin C content increased by a little less than 7 mg from 2000 to 2009. This is not a large amount, and so does not provide strong evidence to suggest that an improvement occurred to the vitamin C content of children's lunches.

Likewise, the difference in vitamin C intake between the lunch types was also small; school lunches provided about 4 mg less vitamin C than packed lunches, which is not enough evidence to suggest that packed lunches were the better option in regards to vitamin C.

4.5.2 Daily vitamin C intake

There is a large p-value for the interaction term in this model (p=0.783), meaning the interaction is removed. Tables 4.13 and 4.14 shows the mean daily intakes of vitamin C for the years and lunch types respectively.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Average daily vitamin C intake	78.9	93.6	14.7	(6.3, 23.1)

Table 4.13: *Least squares means, adjusted for sex, for daily vitamin C intake (mg) of each year*

	School lunch	Packed lunch	Difference	95% confidence interval of difference
Average daily vitamin C intake	83.0	89.5	6.5	(-2.1, 15.0)

Table 4.14: *Least squares means, adjusted for sex, for daily vitamin C intake (mg) of each lunch type*

It appears that throughout the day, children consumed around 15 mg more vitamin C in 2009 than in 2000, regardless of what lunch type they had. This quantity is relatively small compared to the RDA so this does not provide evidence for a change.

In both years, children who had a packed lunch consumed about 6 mg more vitamin C throughout the whole day than those who had a school lunch. Once again, this quantity is not large enough to support the conclusion that one lunch type is a better option.

4.5.3 Diagnostic checks

For both the lunchtime and daily intakes, the diagnostic checks of these models reveal model inadequacies. Although the residual plots are satisfactory with randomly scattered points, on the other hand the Normal probability plot has numerous points that deviate far from the straight line. This can be seen in Figure 4.1. Curvature is evident in both plots, providing strong evidence of non-Normality of the residuals. This breaks the model assumption of Normality, which invalidates the significance tests and the confidence intervals above. Consequently, the conclusions that were made about the differences between years and lunch types are invalidated. The procedure to address the non-Normailty issue is described in the next chapter.

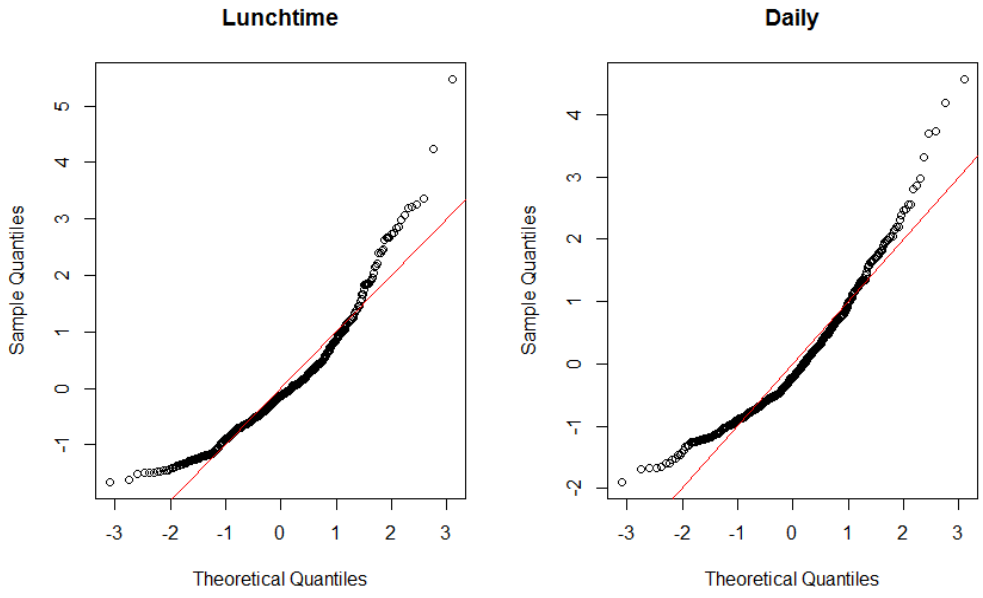

Figure 4.1: *Normal probability plots for lunchtime and daily vitamin C intake*

4.6 Summary

Before proceeding to resolve the non-Normality issue, the findings of this section will first be summarized.

For the lunchtime data of most nutrients, the year*lunch interaction was significant, whereas for the daily data it was not. This makes intuitive sense because the way that lunch type interacts with year is likely to be diluted when considering the consumption throughout the whole day rather than just during lunch. This was true for all nutrients except Vitamin C, but since these data broke the Normality assumption, the conclusions about it lost validity. This means that for lunchtime analysis, the lunch types are considered separately, but for daily analysis, they are pooled together.

The lunchtime consumption of energy, sodium and saturated fat declined significantly over the years in the case of school lunch children, but not significantly for packed lunch children, which was expected due to standards being applied to school lunch only.

The daily consumption of energy and sodium fell significantly over the years, but there was no evidence to suggest the same outcome for saturated fat. There was also no evidence to suggest that the daily consumption of any nutrients differed for school lunch and packed lunch children, which is an indicator that the revised standards had a similar impact on the daily diets of all children, irrespective of lunch type.

5 Departures from model assumptions

This chapter describes how to deal with models that do not satisfy the Normality assumption of linear regression. Considerable departures from model assumptions, such as the ones for the lunchtime and daily vitamin C intakes, should not simply be ignored, as they can invalidate conclusions.

5.1 Data transformation

Data transformation, which means performing the same mathematical operation on all observations, is a commonly-used tool that has many roles in data analysis. One such role is to improve the Normality of a dataset. The aim is to determine a transformation such that when it is applied, the data become approximately Normally distributed – enough so that Normality of the residuals can be assumed. A transformation must not change the order of the values, but can alter the distance between successive points to modify the overall shape of the distribution and achieve a 'bell curve' that is typical of a Normal distribution.

5.2 Box-Cox power transformation

In 1964, statisticians George Box and David Cox devised a procedure for transforming non-Normal data to Normality (Box & Cox, 1964). Supposedly, the pair decided to collaborate on a paper whilst both in Wisconsin, due to the similarity of their surnames and the fact that they were both British. Their method, known as the Box-Cox power transform, is now one of the most popular and commonly used methods to remedy the breakdown of the Normality assumption (Asar et al., 2014).

The Box-Cox power transformation for data $Y_i, i = 1,2, \dots, n$, is as follows.

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda}, & if\ \lambda \neq 0, \\ \log(Y_i), & if\ \lambda = 0, \end{cases} \tag{5.1}$$

where the exponent λ is known as the transformation parameter and requires estimation. Equation (5.1) is the 'conventional,' one-parameter version of the transformation. Clearly, it can be applied to positive data only. To address this restriction, a two-parameter version was devised for non-positive data, which allows for a shift before transformation. This is given by

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{(Y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1}, & if\ \lambda_1 \neq 0, \\ \log(Y_i + \lambda_2), & if\ \lambda_1 = 0. \end{cases} \tag{5.2}$$

In this case, both the transformation parameter λ_1 and the shift parameter λ_2 require estimation, and the condition $Y_i > -\lambda_2$, for all i, is imposed so that the data are made positive by the shift (note that (5.1) is simply the case of the two-parameter version (5.2) with a shift equal to zero, i.e. $\lambda_2 = 0$).

Usually, the Box-Cox parameters are estimated using maximum likelihood, although alternative methods have been proposed, both by Box and Cox (1964) – for example, they developed a method that incorporates Bayesian techniques – and by other authors since then (Sakia, 1992).

Generally, the estimate for the transformation parameter is rounded to the nearest sensible quantity, so that a practical and recognisable transformation (such as square root or inverse transformation, for example) can be implemented.

Once the parameter(s) have been estimated and the Box-Cox transformation applied, a setback is that the transformed variable will typically be on a scale that is unfamiliar to practitioners. Since results are reported most accessibly when on the original scale, the characteristics of the inverse transformation need consideration.

As demonstrated in the previous chapter, the models for both lunchtime and daily vitamin C violate the Normality assumption, and so Box Cox transformations shall be applied to both.

5.3 Lunchtime vitamin C intake

For the lunchtime case, the two-parameter version (5.2) is required, because there were some children that consumed no vitamin C during lunch, meaning that the dataset includes several zeroes.

5.3.1 Box-Cox: estimation of parameters

The Box-Cox parameters can be estimated in R, using a package called `geoR`, which allows use of the `boxcoxfit` command. When called, it returns a maximum likelihood estimate for the transformation parameter, λ_1. Since a shift parameter is required too, the command `lambda2=TRUE` is put in the argument to call for an estimate of λ_2 also. The covariate values for the fitted model are also provided in the argument in the matrix `xmat`, which instructs R to return estimates for the model parameters as well. When this code is run, the following output (truncated to 4 decimal places) is returned.

```
> boxcoxfit(av_vitc_L,xmat,lambda2=TRUE)
Fitted parameters:
 lambda  lambda2 beta0   beta1   beta2   beta3   beta4   sigmasq
 0.4022  0.0015  6.4378  0.0302 -0.1398 -0.5743  0.1154  9.4643
```

This suggests that to achieve approximate Normality, the data could be transformed by a shift of 0.0015, followed by exponentiation to a power of around 0.5. In effect, the Box-Cox method has suggested a square root transformation, preceded by a minute shift. This shift is rounded down to zero without problem, since all observations including the zeroes can be square rooted. Hence, the Box-Cox transformation for these data is a simple square root transformation, which will now be applied.

5.3.2 Square root transformation

Model

After square rooting the data and fitting the model to these transformed data, the interaction term is removed from the model due to its insignificance (p=0.1310). The other covariates are all of key interest and are retained regardless of their p-values, but for completeness, the year variable is significant (p=0.0153) and lunch type is not (p=0.1618).

Diagnostic checks

Diagnostic checks of this model present no cause for concern: the residuals are randomly scattered and the Normal probability plot now shows an acceptable fit (Figure 5.1). This suggests that the square rooted data are indeed Normal, which means that the square root transformation was successful in Normalizing the data.

As another indicator of Normality, the residual skewness can be considered. If data follow an exact Normal distribution, then the skewness of their residuals is obviously zero. Of course, real data is never exactly Normal, so at least some residual skewness is always present in real life situations, but as long as it is small then Normality can be assumed. In this case, the residual skewness is calculated by R to be 0.136 (to 3 d.p.). This small value suggests that the transformed data are at least approximately Normal, which corresponds with the conclusion from the Normal probability plot.

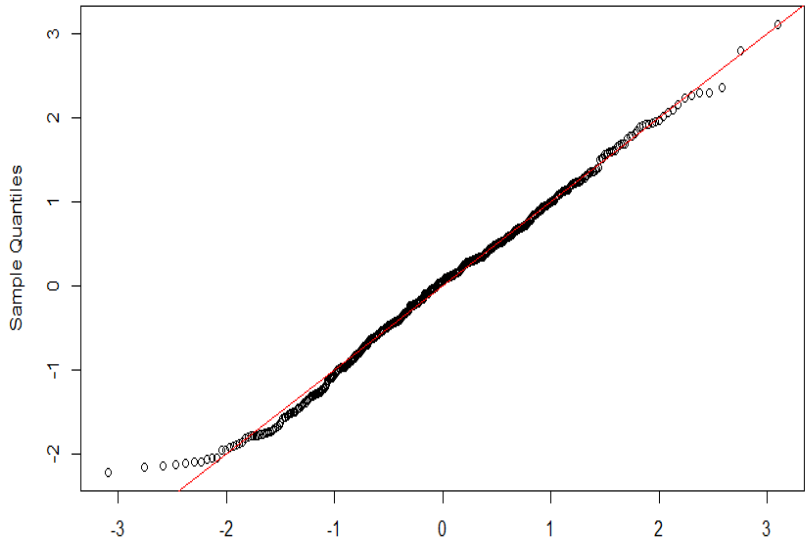

Figure 5.1: *Normal probability plot for square-rooted lunchtime vitamin C intakes*

Inference

Hence, the adjusted means for the square-rooted data can now be calculated and compared. They are shown in Tables 5.1 and 5.2. These values are not particularly useful to nutritionists, however, as they are summaries that have been obtained from transformed data, so they do not represent typical values for mean vitamin C intake. To provide values on an appropriate scale for nutritionists, the values can be back-transformed, by squaring. These back-transformed values are also shown in Tables 5.1 and 5.2, along with their differences, for which the confidence intervals have been left blank deliberately. The reason for this will be explained shortly.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Means of square-rooted data	4.8	5.4	0.5	(0.1, 0.9)
Back-transformed means	23.5	28.7	5.2	

Table 5.1: *Least squares means, adjusted for sex, of square rooted lunchtime vitamin C intake (mg) for each year, along with the back-transformed values*

	School lunch	Packed lunch	Difference	95% confidence interval of difference
Means of square-rooted data	5.0	5.3	0.3	(-0.1, 0.7)
Back-transformed means	25.0	28.1	3.1	

Table 5.2: *Least squares means, adjusted for sex, of square rooted lunchtime vitamin C intake (mg) for each lunch type, along with the back-transformed values*

It appears that vitamin C intake increased by about 5 mg from 2000 to 2009 and that packed lunches contained about 3 mg more vitamin C than school lunches.

Problem with back-transformation

However, there is a problem with interpreting the back-transformed means and their differences. Back-transforming by squaring the means of the transformed data ensures that they are converted back to the original scale. This is beneficial for nutritionists' understanding, but it has been made apparent already (in section 4.5) that data on this scale violates the Normality assumption, which is the reason a transformation was sought in the first place. This violation means that confidence intervals for the difference of the back-transformed means cannot be found, hence why they are left blank in the tables. This is because confidence intervals are only valid if a number of assumptions can be reasonably made, one of which is that the population distribution is approximately Normal (Williams, 2013), which is not the case for data on the original scale. As a result, valid conclusions can be drawn from data on the square-root scale only, which makes this transformation redundant. Tables 5.3 and 5.4, which contain algebraic expressions for the contents of the tables above, are provided to enhance the clarity of the explanation.

	2000	2009	Difference	95% confidence interval of difference
Means on square-root scale	$\widehat{\sqrt{Y_{2000}}}$	$\widehat{\sqrt{Y_{2009}}}$	$\hat{d} = \widehat{\sqrt{Y_{2009}}} - \widehat{\sqrt{Y_{2000}}}$	$\hat{d} \pm s.e.(d)$
Means on original scale	$\hat{Y}_{2000} = \left(\widehat{\sqrt{Y_{2000}}}\right)^2$	$\hat{Y}_{2009} = \left(\widehat{\sqrt{Y_{2009}}}\right)^2$	$\left(\widehat{\sqrt{Y_{2009}}}\right)^2 - \left(\widehat{\sqrt{Y_{2000}}}\right)^2$	

Table 5.3: *Interpretation of square-root transformation for the least squares means of each year and their difference*

	School lunch	Packed lunch	Difference	95% confidence interval of difference
Means on square-root scale	$\widehat{\sqrt{Y_{SL}}}$	$\widehat{\sqrt{Y_{PL}}}$	$\hat{d} = \widehat{\sqrt{Y_{PL}}} - \widehat{\sqrt{Y_{SL}}}$	$\hat{d} \pm s.e.(d)$
Means on original scale	$\hat{Y}_{SL} = \left(\widehat{\sqrt{Y_{SL}}}\right)^2$	$\hat{Y}_{PL} = \left(\widehat{\sqrt{Y_{PL}}}\right)^2$	$\left(\widehat{\sqrt{Y_{PL}}}\right)^2 - \left(\widehat{\sqrt{Y_{SL}}}\right)^2$	

Table 5.4: *Interpretation of square-root transformation for the means of each lunch type and their difference*

Therefore, although a square root transformation is successful in Normalizing these data as required, it is not particularly useful as it does not enable practical interpretation of the estimable quantities on the original scale.

5.3.3 Log-transformation

Although the Box-Cox parameter estimates suggested a square-root transformation for these data, the previous section showed that this possesses difficulties with interpretation. This section investigates a log-transformation, which involves taking the natural logarithm of each observation, because it has the beneficial property of having an intuitive interpretation when the data are back-transformed.

Unlike the square-root transformation however, a log-transformation cannot be applied to the zero observations in these data, as the logarithm of zero is undefined. This is resolved by adding a constant quantity to each observation before taking logs, but a sensible choice of constant must be determined – whichever one minimizes the residual skewness is the logical choice, since Normality corresponds to zero residual skewness.

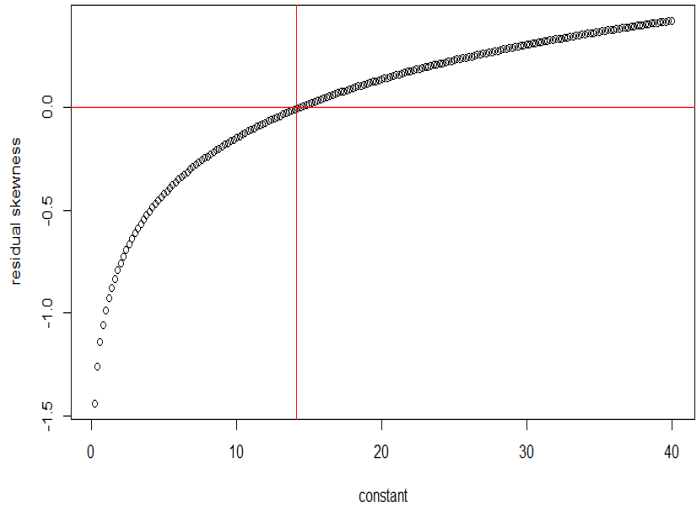

Figure 5.2: *Various constants plotted against their residual skewness obtained when added to the data prior to log-transformation*

Figure 5.2 shows various constants plotted against the residual skewness that is obtained by adding it to the data prior to log-transformation. Minimal residual skewness is achieved with a constant of approximately 15, which gives the following transformation

$$Y^*_{ijk} = \log(Y_{ijk} + 15) \tag{5.3}$$

where Y_{ijk} denotes the lunchtime vitamin C intake of an individual of type (ijk). The transformed data will be referred to as the logged shifted data.

Model

Fitting the model to the transformed data leads to the removal of the interaction (p=0.1242). After this, year appears to be significant (p=0.0150), whilst lunch type appears not to be (p=0.1771).

Diagnostic checks

Diagnostic checks are now carried out to assess the Normality of the logged shifted data. The points on the Normal probability plot exhibit a close fit to the straight line, as shown in Figure 5.3, making it reasonable to accept the Normality assumption for the transformed data.

Furthermore, the residual skewness of the model on these transformed data is only 0.018, indicating that the log-transformation has achieved approximate Normality. The residual skewness here is in actual fact smaller than that of the square rooted data, so in some sense, the log-transformation (with shift) is better than the square-root at Normalizing these data.

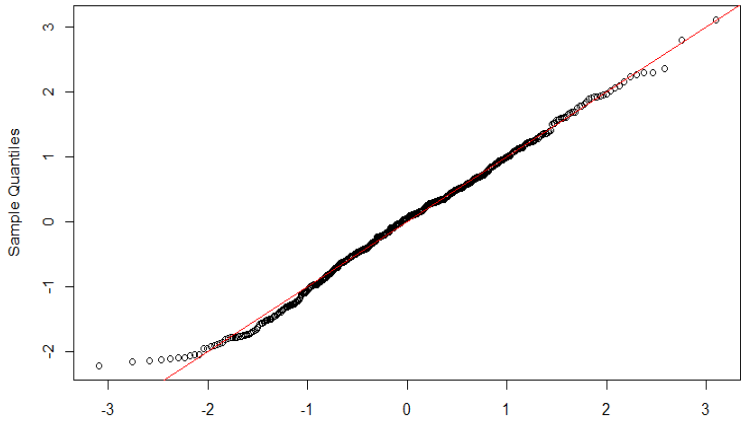

Figure 5.3: *Normal probability plot for lunchtime vitamin C intake*

Useful property of log-transformation

An attractive quality of the log-transformation is that, upon back-transformation, an intuitive interpretation is possible, as mentioned at the beginning of this section. This is owing to the following general relationship between the geometric mean and arithmetic mean of some general data $Y_1, Y_2, \ldots, Y_n$,

$$GM(Y_i) = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} = \exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log(Y_i)\right\} = \exp\{AM(\log(Y_i)\}, \tag{5.4}$$

where $GM(.)$ and $AM(.)$ denote the geometric and arithmetic means respectively of the data in their argument.

The arithmetic mean of the logged shifted lunchtime vitamin C intake in 2000, for instance, is given by

$$\widehat{\log(Y_{2000})} = \frac{1}{n_{2000}}\sum_{\ell=1}^{n_{2000}}\log(Y_{2000\ell}), \tag{5.5}$$

where ℓ denotes the individual and n_{2000} is the number of subjects in the year 2000. There is of course a similar expression for the mean in 2009 and for the means of each of the lunch types. From (5.4), the arithmetic mean for the logged shifted lunchtime vitamin C intake is clearly equal to the logged geometric mean, i.e.

$$\widehat{\log(Y_{2000})} = AM(\log(Y_{2000})) = \log(GM_{2000}). \tag{5.6}$$

Taking the difference between the years, i.e. between the means of logged shifted data for 2000 and 2009, which is the estimable quantity, gives

$$\widehat{\log(Y_{2009})} - \widehat{\log(Y_{2000})} = \log(GM_{2009}) - \log(GM_{2000}) = \log\left(\frac{GM_{2009}}{GM_{2000}}\right). \tag{5.7}$$

Rather usefully, when this difference is back-transformed by anti-logging, it simply gives the ratio of the geometric means of the shifted data, which is a straightforward and comprehensible expression. This ratio can then be used as the 'difference' between the means on the original scale, which nutritionists will be able to interpret more easily.

Due to the asymmetry of the log-transformation, the confidence interval of this ratio can be found directly by anti-logging the confidence interval of the difference, since this is not symmetrical about the ratio.

Tables 5.5 and 5.6 summarize the above by giving algebraic expressions for each cell.

	2000	**2009**	**Difference / Ratio**	**95% confidence interval of difference**
Means on log scale	$\log(GM_{2000})$	$\log(GM_{2009})$	**2009 - 2000:** $\log\left(\frac{GM_{2009}}{GM_{2000}}\right)$	(L, U)
Means on original scale	GM_{2000}	GM_{2009}	**2009 / 2000:** $\frac{GM_{2009}}{GM_{2000}}$	$(\exp(L), \exp(U))$

Table 5.5: *Interpretation of log-transformation for the means of each year and their difference*

	School lunch	**Packed lunch**	**Difference/Ratio**	**95% confidence interval of difference**
Means on log scale	$\log(GM_{SL})$	$\log(GM_{PL})$	**PL – SL:** $\log\left(\frac{GM_{PL}}{GM_{SL}}\right)$	(L, U)
Means on original scale	GM_{SL}	GM_{PL}	**PL / SL:** $\frac{GM_{PL}}{GM_{SL}}$	$(\exp(L), \exp(U))$

Table 5.6: *Interpretation of log-transformation for the means of each lunch type and their difference*

Hence, when using the log-transformation, it is possible to produce a comprehensible expression for the estimable quantity on the original scale of the data, along with its confidence interval. This easy and intuitive interpretation makes the log-transformation a much more appealing option than the square-root transformation, despite the results of the Box-Cox method, particularly as the residual skewness is actually smaller when using the log-transformation. Due to this, the log-transformation will be used in favour of the square-root transformation for this project.

Inference

Having justified use of the log-transformation instead of the square-root transformation, inference is now made on the back-transformed shifted means and their differences, given in Tables 5.7 and 5.8.

	2000	**2009**	**Difference/Ratio**	**95% confidence interval of difference**
Means on log-scale	3.638	3.756	**PL – SL:** 0.118	(0.023, 0.212)
Means on original scale	38.0	42.8	**PL / SL:** 1.12	(1.02, 1.24)

Table 5.7: *Least squares means, adjusted for sex, of logged lunchtime vitamin C intake (mg) for each year, along with back-transformed values*

	School lunch	**Packed lunch**	**Difference/Ratio**	**95% confidence interval of difference**
Means on log scale	3.664	3.730	**PL – SL:** 0.0664	(-0.0299, 0.1626)
Means on original scale	39.0	41.7	**PL / SL:** 1.07	(0.97, 1.18)

Table 5.8: *Least squares means, adjusted for sex, of logged lunchtime vitamin C intake (mg) for each lunch type, along with back-transformed values*

It appears that mean vitamin C intake in 2009 was 1.12 times as large as the mean in 2000, indicating that a reasonable increase has occurred. The lower limit of the 95% confidence interval for this ratio is greater than 1, and so it is statistically very likely that the intake was higher in 2009 than in 2000. This suggests that the new food standards caused an improvement during lunchtime, in terms of vitamin C intake.

Also, it seems that packed lunches on average contained 1.07 times as much vitamin C than school lunches. The span of the confidence interval for this ratio is mostly greater than 1, so there is strong evidence to suggest that the average packed lunch contains a little more vitamin C than the average school lunch.

5.4 Transforming daily vitamin C intake

Having analysed the lunchtime vitamin C values, the daily intakes now need consideration, as they too broke the Normality assumption. Unlike for lunchtime, the daily intakes are all non-zero, so individuals clearly obtained vitamin C outside of school, if not during lunch. Consequently, these data do not require a shift and the conventional one-parameter version of the Box-Cox method can be used.

5.4.1 Box-Cox: estimation of parameter

Once again, the `boxcoxfit` command can be used in R to obtain an estimate for the exponent. Since the shift parameter is not required, the `lambda2` argument is not included here – R defaults it to 'false' in its absence. The matrix `xmat` is used again to obtain model parameter estimates, which produces the following output for the daily vitamin C intakes.

```
> boxcoxfit(av_vitc_TD,xmat)
Fitted parameters:
 lambda   beta0   beta1   beta2   beta3   beta4   sigmasq
 0.1997  6.6449  0.0458  0.1472 -0.0183  0.0073  1.4561
```

The estimate for the transformation parameter is approximately zero, indicating that a log-transformation is optimal. This is advantageous because, as mentioned, the log-transformation has the unique property of having a straightforward interpretation of the differences on the original scale. It is also advantageous to use the same transformation for daily data as for lunchtime data for presentational purposes.

5.4.2 Log-transformation

Having applied the log-transformation to these data and found that the transformed data do indeed satisfy model assumptions (graphs not shown), adjusted means and differences are obtained, and then back-transformed by anti-logging. The results, the geometric means of each category along with their ratio, are displayed in Tables 5.11 and 5.12.

	2000	**2009**	**Ratio (2009/2000)**	**95% confidence interval of ratio**
Average daily vitamin C intake	68.9	83.1	1.21	(1.09, 1.33)

Table 5.11: *Adjusted mean daily vitamin C intake in mg for each year, where the averages are the geometric means*

	School lunch	Packed lunch	Ratio (PL/SL)	95% confidence interval of ratio
Average daily vitamin C intake	72.8	78.6	1.08	(0.97, 1.20)

Table 5.12: *Adjusted mean daily vitamin C intake in mg for each lunch type, where the averages are the geometric means*

The average daily intake of vitamin C was 1.21 times greater in 2009 than in 2000, which is a fairly large improvement. Moreover, the 95% confidence interval spans a range that is above 1, suggesting that an increase from 2000 to 2009 is highly likely.

The average daily vitamin C intake of children who eat packed lunches is 1.08 times as large compared to those who have school lunches, suggesting that lunch type is a contributing factor to how much vitamin C is consumed daily, and that packed lunches affect vitamin C positively. Once again, the span of the confidence interval is mostly greater than 1, so it is likely that packed lunch children do indeed consume slightly more vitamin C throughout the day than school lunch children.

5.5 Summary

The aim of this chapter was to resolve the non-Normality of lunchtime vitamin C intake and daily vitamin C intake. This was done by taking the Box-Cox approach. The Box-Cox transformation for the lunchtime data was a square-root transformation, and for the daily data it was a log-transformation. However, after investigating the transformations further, it was found that the log-transformation possesses a very useful feature that makes it uniquely usable when making inferences on back-transformed values. Consequently, it was decided that a log-transformation would be used for both datasets. This could be justified for the lunchtime data as the log-transformed data still appeared to satisfy the Normal assumption and had small residual skewness.

6 The relationship between sodium intake and energy intake

This section draws attention to a potential flaw in the previous analysis of sodium intake, and proceeds to investigate ways in which to overcome it.

6.1 The problem with fitting the previous model to the sodium data

A key objective of the revised school food standards was to reduce the overall amount of sodium that children consume, and the results in Chapter 4 on daily sodium intake suggest that they have been successful in doing so. Table 4.6 below (reproduced directly from Chapter 4) shows how the average seemed to improve over the years – daily sodium intake dropped considerably between 2000 and 2009, by over 480 mg. However, the reasons for this sodium decrease merit further investigation. Table 4.3 (also from Chapter 4) shows that daily energy intake also fell considerably over the same time period, by almost 265 kcal. Since energy intake is a proxy for amount eaten, and obviously the more food one consumes, the higher the sodium intake, the issue to consider here is whether the reduction in sodium intake is simply attributed to the reduction in energy intake. If this is indeed the case, then the salt-density of children's diets may have been unchanged over the years.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Average daily sodium intake	2630.8	2150.3	-480.5	(-594.7, -366.3)

Table 4.6: *Adjusted mean daily sodium intake in mg for each year*

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Average daily energy intake	1904.9	1640.2	-264.7	(-337.7, -191.7)

Table 4.3: *Adjusted mean daily energy intake in kcal for each year*

To aid visualisation, Figure 6.1 shows the scatterplot of sodium versus energy, with different coloured symbols for the two years: black circles represent 2000, red crosses represent 2009. The strong positive correlation between the two variables is apparent. The red crosses generally have smaller sodium values than the black circles, showing that sodium intake was indeed lower in 2009, but it can be seen that they have smaller energy values too. The fitted lines for the two years are very similar in gradient, which implies that the relationship between sodium and energy was similar in the two years. Indeed, more will be said about this shortly.

Of course, one might argue that as long as average sodium intake is at an acceptable level, then the salt-density of diet is unimportant. However, it is advisable to maintain a less salt-dense diet, because food intake can vary considerably on a day-to-day basis, depending on a wide range of factors. Therefore, making a habit of choosing less salty foods ensures that one's overall sodium intake is reduced in the long term, even on days when food intake is higher than normal.

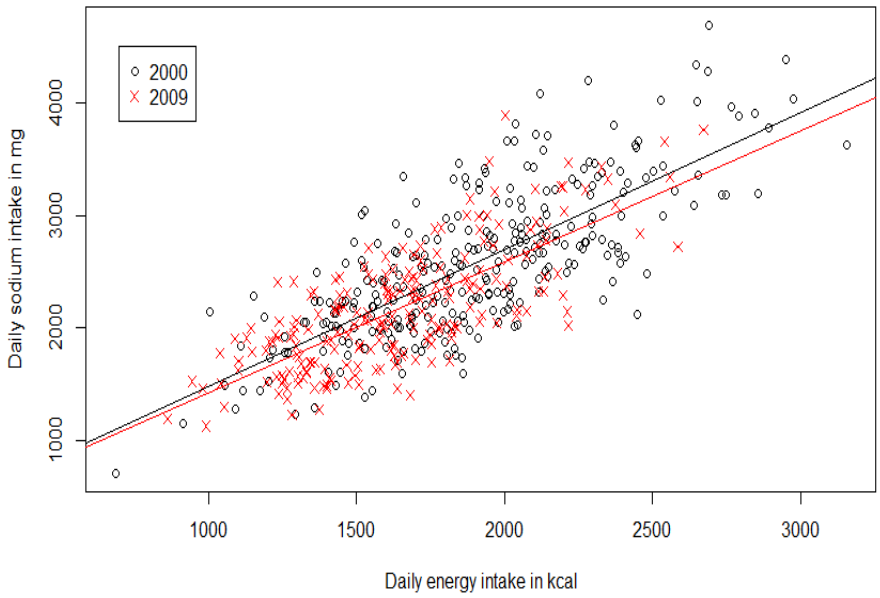

Figure 6.1: *Daily sodium intake versus daily energy intake, with different points for the years*

In summary, although the previous analysis suggests there has been a significant decrease (and therefore an improvement) in daily sodium intake, there may not have been such a substantial improvement in the *density* of salt in diet, if the variation in sodium intake is solely due to the variation in energy intake. There is, therefore, a need to investigate this further.

6.2 Model for sodium intake with energy as covariate

The most natural and obvious way to investigate how heavily sodium intake depends on energy intake, is to include energy intake as an explanatory variable in the model. This model is given by

$$Na = \alpha + \beta E + \epsilon, \tag{6.1}$$

where Na = daily sodium intake, E = daily energy intake, ϵ = error, with $\epsilon \sim N(0, \sigma^2)$, and α incorporates the effects of all other covariates as well as the general mean.

Having fit this model to the data, the year and lunch interaction is removed due to its insignificant p-value (p=0.1352). The p-value for daily energy intake is extremely small and remains so after removal of the interaction (p=2.2 $\times 10^{-16}$), suggesting that it has a substantial effect on daily sodium intake, as expected. The year, lunch type and sex covariates are also very significant (all $p < 0.001$), suggesting that they too affect the value of sodium intake considerably.

The introduction of energy as a covariate gives rise to another interaction of interest – it is plausible that energy intake affected sodium intake differently in the two years, and so the year and energy interaction requires consideration also. Assessing the p-value for this term is essentially a test for parallelism: if it is not significant it is removed, and so the regression lines for each year will have equal gradients, whereas if the interaction is present, the gradients will differ. When this interaction is incorporated into the model, its p-value is not significant (p=0.3344), so the term is excluded. This suggests that the way in which energy affected sodium

was not significantly different in the two years. This in turn implies that the slight difference between the gradient of the fitted lines of the two years in Figure 6.1 is simply down to chance.

The diagnostic checks for this model provide no reason to believe that the usual assumptions do not hold. The plots of the residuals against the covariate values, jittered where appropriate (for the categorical covariates), show randomly scattered points across the plot, with no apparent trend and the Normal probability plot exhibits a relatively straight line of points, with only some minor deviations.

Since the model is adequate, it can be used to compute least squares means for the daily sodium intake in each year, which can then be compared. These are displayed in Table 6.1. This time, the means are not only adjusted for sex as before, but also for energy intake, meaning that energy is fixed at an arbitrary value. In other words, the least squares means have been calculated as if the energy intake was the same in each year. (As discussed in Chapter 3, any value for energy can be used, as it will not affect the difference between years – making this an estimable quantity). This allows for comparisons to be made between the mean sodium intakes of each year, as the effects of year will no longer be confounded with the variations in energy intake.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Average daily sodium intake	2497.5	2323.8	-173.7	(-254.2, -93.2)

Table 6.1: *Adjusted mean daily sodium intake (mg) for each year, adjusted for sex and energy*

It appears that the mean daily sodium intake, for any fixed value of energy intake, has decreased by 173.70 mg. This means that on average, even if a child consumed the same amount of energy in both years, their average daily sodium intake will still have decreased by over 170 mg. Compared to the difference of 480 mg between the means that were unadjusted for energy intake, this is still a reasonably large amount, suggesting that a reasonable amount of the reduction in sodium is not just due to the reduced energy intake. This in turn implies that there has indeed been a reduction in salt density, which is of course a desired effect.

6.3 Model for proportion of energy intake provided by sodium

Nutritionists often like to analyse the proportion of daily energy intake that comes from sodium, so this section fits a model to this proportion, given by

$$\frac{Na}{E} = \alpha + \epsilon, \tag{6.2}$$

where Na, E, α and ϵ have the same interpretation as in section 6.1. The idea of this approach is to incorporate energy intake into the response, by division, in order to achieve a model that does not depend on energy intake but still takes account of it.

Fitting this model causes the p-values for all covariates, except the interaction term, to be small (all $p < 0.01$). Removal of the interaction does not drastically alter the other p-values (all remain less than 0.01), so all covariates seem to have a significant effect on the proportion.

The diagnostic checks for this model do not reveal any inadequacies, suggesting that it is a satisfactory model that complies with the Normality and constant variance assumptions about the residuals. It is thus possible to proceed with analysis of the means.

To help the reader become attuned to values that are typical for the ratio of sodium (in mg) to energy (in kcal), the raw means of this proportion, for each year and lunch type combination, are given in Table 6.2.

	2000	**2009**	**Difference**
School lunch	1.3360	1.3034	-0.0325
Packed lunch	1.4780	1.3477	-0.1303

Table 6.2: *Raw means of the ratio of sodium intake (mg) over energy intake (kcal) for each year and lunch type combination in mg/kcal*

With these means as an approximate benchmark, it seems that the proportion of sodium over energy in mg/kcal is typically somewhere between 1.0 and 1.5.

The least squares means of the proportions, adjusted for sex as usual, can now be calculated using the model and then compared. They are given in Table 6.3, along with their difference plus a confidence interval for the difference. It appears that between 2000 and 2009, the proportion decreased by about 0.07 mg/kcal. With respect to the typical values shown before in Table 6.2, this is a fairly reasonable decrease. This suggests that the proportion of sodium in the children's total daily energy intake decreased slightly between the years, which is compatible with the conclusion in section 6.2, that the salt-density of the children's diets decreased.

	2000	**2009**	**Difference**	**95% confidence interval of difference**
Mean of proportion	1.3923	1.3194	-0.0729	(-0.1169, -0.0289)

Table 6.2: *Mean proportion of daily sodium (mg) over daily energy (kcal) for each year*

Fitting a model to the sodium over energy proportion, as done above, is a technique favoured by nutritionists. However, whilst this model is valid, it does not include energy intake as a covariate, meaning that none of its variation has been accounted for, even though it may well affect the Na/E proportion. Energy should therefore be added to the model to see if there are truly grounds on which to not include it.

The new model is then given by

$$\frac{Na}{E} = \alpha + \beta E + \epsilon. \tag{6.3}$$

When this is fitted, the year* lunch and year*energy interactions are again insignificant, and thus excluded. The p-value for energy is indeed very small ($p=1.19\times10^{-5}$), indicating that energy intake has a significant effect on the dependent variable, as expected. As a result, there are no grounds on which to remove it and it should be retained in the model, making (6.3) preferable to (6.2).

However, the idea of fitting a model to Na/E was to use a model that takes account of energy intake but does not depend on it, allowing nutritionists to analyse changes in sodium that are not attributed to changes in energy. However, from a statistician's viewpoint, energy intake should be included in this model anyway due to its evident significance ($p=1.19\times10^{-5}$). Although the regression coefficient of energy is very small (-1.161

$\times 10^{-4}$), which means that its effect is only minor, its very small p-value means that this small effect is still very significant and so the model is not free of energy. It could be argued however, that for practical purposes, since the effect of energy is so small, it could be ignored.

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.603e+00  5.452e-02  29.408  < 2e-16 ***
yearf9      -1.036e-01  2.312e-02  -4.484 9.08e-06 ***
lunchf1      8.969e-02  2.245e-02   3.995 7.41e-05 ***
genderf1    -6.935e-02  1.986e-02  -3.491 0.000523 ***
av_kcal_TD  -1.161e-04  2.623e-05  -4.424 1.19e-05 ***
```

6.4 Modification of model

It would be ideal if the techniques above could be combined in some way, so that nutritionists could use a model for the specified proportion, that is free of energy intake and with which statisticians agree. To do this, an attempt can be made to adjust the model so that it no longer depends on energy.

Looking again at model (6.1), which includes energy as a covariate, and dividing throughout by energy intake, E, gives

$$\frac{Na}{E} = \frac{\alpha}{E} + \beta. \tag{6.3}$$

The response is now the specified proportion, but the right hand side is clearly not free of energy intake. Instead, an alternative model can be considered, given by

$$\log(Na) = \alpha + \beta \log(E). \tag{6.4}$$

(The reader might notice that taking the log of the sodium data is potentially problematic, since the sodium data is assumed to be Normal, and taking the log of Normal data does not produce Normal data – see Appendix 7.3 for an explanation as to why this is not an issue.) This model takes the variation in energy into account, through the $\log(E)$ term. Subtracting $\beta \log(E)$ from both sides gives

$$\log\left(\frac{Na}{E^{\beta}}\right) = \alpha. \tag{6.5}$$

Clearly, if β is equal to 1 (or at least close to 1), then (6.5) provides a model whose response is a function of Na/E and is independent of E. Thus, the estimate of β, the coefficient of log(E), needs to be obtained and checked. Fitting model (6.4) in R gives the following output.

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.41785    0.25705   5.516 5.53e-08 ***
yearf9          -0.07465    0.01710  -4.364 1.55e-05 ***
lunchf1          0.06568    0.01663   3.950 8.92e-05 ***
genderf1        -0.04944    0.01467  -3.371 0.000807 ***
log(av_kcal_TD)  0.85265    0.03394  25.119  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.163 on 508 degrees of freedom
Multiple R-squared:  0.6251,   Adjusted R-squared:  0.6221
F-statistic: 211.7 on 4 and 508 DF,  p-value: < 2.2e-16
```

The estimate of β is 0.853, which for practical purposes is close enough to 1. Thus, (6.5) is indeed a model whose response is a function of the desired proportion and is free of energy, so it can now be fitted and means obtained.

When (6.5) is fitted, the p-value for the year and lunch interaction is large (p=0.1308), so the term is excluded usual. After this, the p-values for the other covariates, year, lunch and sex, are all small (all p<0.01), indicating that they all significantly affect the logged proportion of sodium over energy.

Diagnostic checks of the residuals give no cause for concern, indicating that this model satisfies the residual assumptions. The least squares means of the logged proportion are then calculated from the model and are shown in Table 6.3, along with their back-transformed values, so that they can be compared.

	2000	**2009**	**Difference/Ratio**	**95% confidence interval of difference/ratio**
Means on log scale	0.316	0.264	**2009 - 2000:** -0.053	(-0.085, -0.020)
Means on original scale	1.37	1.30	**2009 / 2000:** 0.95	(0.92, 0.98)

Table 6.3: *Least squares means, adjusted for sex, for each year of proportion of sodium (mg) over energy (kcal)*

It appears that between 2000 and 2009, salt density of children's diets, represented by sodium over energy proportion, experienced a decrease, so that in 2009 it was 0.95 times the amount that it was in 2000. In other words, it has decreased by about 5%. This is a reasonable decrease, and supports the conclusions from the other models – that sodium intake in children's diets has indeed decreased and not simply because the amount eaten has decreased.

6.5 Summary

This section has detailed several techniques for investigating how much of the sodium decrease is attributed to the energy decrease. The same techniques could be applied to test if the decrease in saturated fat was largely due to energy decrease, but due to space-constraints this analysis is not included.

The three models that were used gave similar conclusions: that children's sodium intake decreased over the years and not purely because their overall food consumption had lessened. This in turn indicates that the salt-density of their diets may have decreased.

7 Appendices

7.1 Revised school food standards

Nutrient Standards	
Energy	30% of the estimated average requirement (EAR)[36] This standard is linked to the recommendation that schools need to promote healthy levels of physical activity
Protein	Not less than 30% of reference nutrient intake (RNI)
Total carbohydrate	Not less than 50% of food energy
Non-milk extrinsic sugars	Not more than 11% of food energy
Fat	Not more than 35% of food energy
Saturated fat	Not more than 11% of food energy
Fibre	Not less than 30% of the calculated reference value *Note: calculated as Non Starch Polysaccharides*
Sodium	Not more than 30% of the SACN[37] recommendation
Vitamin A	Not less than 40% of the RNI
Vitamin C	Not less than 40% of the RNI
Folate/folic acid	Not less than 40% of the RNI
Calcium	Not less than 40% of the RNI
Iron	Not less than 40% of the RNI
Zinc	Not less than 40% of the RNI

Food Standards	
Fruit and vegetables	Not less than 2 portions per day per child, at least one of which should be salad or vegetables, and at least one of which should be fruit
Oily fish	On the school lunch menu at least once every 3 weeks
Deep fried products	Meals should not contain more than two deep fried products in a single week
Processed foods[1]	Should not be reformed/reconstituted foods made from "meat slurry"
Bread (without spread)	Available unrestricted throughout lunch
Confectionery and savoury snacks[2]	Not available through school lunches
Salt/highly salted condiments	Not available at lunch tables or at the service counter
Drinks	The only drinks available should be water (still or fizzy), skimmed or semi-skimmed milk, pure fruit juices, yoghurt and milk drinks with less than 10% added sugar, or combinations of these (e.g. smoothies)
Water	Easy access to free, fresh, chilled drinking water

7.2 Recommended Daily Amounts (RDAs)

Food/Nutrient	Recommendations			
	4-7y		11-14y	
	Male	Female	Male	Female
Energy (kcals)	1715	1545	2220	1845
Fat (%)	No more than 35% food energy			
Saturated Fat (%)	No more than 11% food energy			
NMES (%)	No more than 11% food energy			
Protein (g)	19.7		42.1	41.2
Sodium (g)	700		1600	1600
Calcium (mg)	450		1000	800
Iron (mg)	6.1		11.3	14.8
Zinc (mg)	6.5		9.0	9.0
Vitamin C (mg)	30		35	35
Vitamin A (µg)	400		600	600
Folate (µg)	100		200	200
Fruit & Vegetables[1] (portion/g)	At least 5 portions per day (equivalent to 400g)			

7.3 Taking logs of a Normal distribution

Let X be a random variable. Assume that $X \sim N(\mu, \sigma^2)$.

Then, standardising X gives

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1),$$

$$\Rightarrow X = \mu + \sigma Z, \;\; Z \sim N(0, 1).$$

Taking the log of the Normally distributed variable, X, gives

$$\log(X) = \log(\mu + \sigma Z)$$

$$= \log\left\{\mu\left(1 + \frac{\sigma}{\mu} Z\right)\right\}$$

$$= \log(\mu) + \log\left(1 + \frac{\sigma}{\mu} Z\right).$$

Provided that the standard deviation is small, relative to the mean, then $\sigma Z / \mu$ is a small number, and in that case $\log(1 + \sigma Z / \mu)$ is equivalent to the log of a number that is just over 1, which is also a small number. Hence,

$$\log(X) \approx \log(\mu) + \frac{\sigma}{\mu} Z.$$

This implies that $Y = \log(X) \sim N(\log(\mu), \sigma / \mu)$ approximately. In other words, if X is Normally distributed with mean μ and standard deviation σ, then provided that $\sigma \ll \mu$, the variable $Y = log(X)$ will also be approximately Normally distributed.

8 References

ASAR, O., OZLEM, I. and DAG, O. (2014), *Estimating Box-Cox Power Transformation Parameter via Goodness of Fit*. Available from: <http://arxiv.org/ftp/arxiv/papers/1401/1401.3812.pdf> [Accessed: 7th April 2015]

BOX, G. E. P. and COX, D. R. (1964), An Analysis of Transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**(2)

DALLAL, G. E. (2001), *Adjusted Means: Adjusting for Categorical Variables*. Available from: <http://www.jerrydallal.com/lhsp/lsmeans2.htm> [Accessed: 3rd February 2015]

Department for Education (2014), *Revised Standards for Food in Schools: Consultation document*. UK government. Available from: <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/287073/Revised_standards_for_food_in_schools_-_Consultation_document.pdf > [Accessed: 13th February 2015]

GOVE, M. (2014), *Press Release. New School Food Standards.* UK Government. Available from: <https://www.gov.uk/government/news/new-school-food-standards> [Accessed: 22nd January 2015]

House of Commons Health Committee (2004), *Obesity – third report, together with formal minutes*, 1. London: The stationary office. Available from: <http://www.publications.parliament.uk/pa/cm200304/cmselect/cmhealth/23/23.pdf> [Accessed: 20th January 2015]

JACQUEZ, J. A. and GREIF, P. (1985), Numerical Parameter Identifiability and Estimability. *Mathematical Biosciences.* Elsevier Science Publishing. New York.

Jamie Oliver Food Foundation (2015), *Our Programmes.* Available from: <http://www.jamieoliverfoodfoundation.org.uk/> [Accessed: 6th April 2015]

MUCAVELE, P., NICHOLAS, J and SHARP, L. (2013), *Development and Pilot Testing of Revised Standards for School Lunches*. Available from: < http://www.schoolfoodplan.com/wp-content/uploads/2014/02/School-Food-Plan-Pilot-study-EVALUATION-REPORT-Final-V3.pdf > [Accessed: 10th February 2015]

National Centre for Social Research and University College London (2002), *Health Survey for England, 2002*. 2nd Edition. Colchester, Essex: UK Data Archive.

OSBOURNE, J. W. (2010), Improving Your Data Transformations: Applying the Box-Cox Transformation. *Practical Assessment, Research & Evaluation*, **15(**12)

POOLE, M. A. and O'FARRELL, P. N. (1970), *The Assumptions of the Linear Regression Model.* Available from: <http://people.uleth.ca/~towni0/PooleOfarrell71.pdf> [Accessed: 7th April 2015]

SAKIA, R. M. (1992), The Box-Cox Transformation Technique: A Review. *The Statistician,* **41**(169)

School Meals Review Panel (2005), *Turning the Tables: Transforming School Food. Main Report.* Available from: <http://dera.ioe.ac.uk/5584/3/SMRP%20Report%20FINAL.pdf> [Accessed: 7th January 2015]

Wikipedia (2015), *Jamie's School Dinners*. Available from: <http://en.wikipedia.org/wiki/Jamie%27s_School_Dinners > [Accessed: 6th April 2015]

WILSON, K. L., *What is the importance of good nutrition for kids?* Healthy Eating. Available from: <http://healthyeating.sfgate.com/importance-good-nutrition-kids-6236.html > [Accessed: 15th January 2015]

WILLIAMS, M. N., GRAJALES, C. A. G. and KURKIEWICZ, D. (2013), Assumptions of Multiple Regression: Correcting Two Misconceptions. *Practical Assessment, Research & Evaluation,* **18**(11)