Stony Brook University

Residual Logistic Regression

A Dissertation Presented by

Fabiola Berenice Báez-Revueltas

to

The Graduate School

In Partial Fulfillment of the

Requirements

For the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

August 2009

Stony Brook University

The Graduate School

Fabiola Berenice Báez-Revueltas

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

Acceptance of this dissertation.

Wei Zhu – Dissertation Advisor
Professor, Applied Mathematics and Statistics

Stephen Finch - Chairperson of Defense
Professor, Applied Mathematics and Statistics

Hongshik Ahn
Professor, Applied Mathematics and Statistics

Marci Lobel
Associate Professor, Psychology Department

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Residual Logistic Regression

by

Fabiola Berenice Báez-Revueltas

Doctor in Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2009

In biostatistical analysis it is often necessary to filter out potential confounding variables and mature techniques have been developed to do so in the setting of matched data analysis and multiple regression analysis among others, especially when the response variable is continuous. Satisfactory methods, however, have not been developed in the setting of dichotomous outcomes. We propose the residual logistic regression analysis for logistic regression controlling for confounding variables. This method is compared to existing methods including the Pearson residual analysis. The pros and cons of this and other methods are discussed and guidelines are provided for the users.

Another traditional method for controlling confounding variables is through subject matching or pairing. We examined the pros and cons of whether one should analyze the

original independent samples or a selected subset of the paired sample extracted from the original samples. A resampling technique is adopted to examine whether the paired sample selected is unbiased or not. The powers of these two approaches are compared through simulation studies for tests on population means and proportions.

Finally, methods proposed and discussed in this thesis are applied to a study conducted at the NYU Alzheimer's Disease Core Center (ADCC) where the goal is to determine if the mental decline rate is the same for subjects with or without subjective complaints of cognitive impairment. It is shown that the proposed residual logistic regression analysis yielded superior and yet consistent results in comparison to other existing methods.

To my two greatest inspirations

Enrique and Sebastian

Table of Contents

List of Figures

List of Tables

Introduction

In statistical analyses it is often critical to filter out the influence from potential confounding variables. In the past few decades mature techniques have been developed for multiple linear regression where the response variable is continuous. Satisfactory methods, however, have not been developed in the setting of a logistic regression where the response variable is dichotomous.

Confounding variables in a statistical model are those variables correlated to both the independent and dependent variables. Thus they can affect the results obtained from the study by adding a considerable amount of bias and rendering the conclusions meaningless.

Several methods are available to control for potential confounding variables. These include[50]:

- In study designs

- Restriction
- Random allocation of subjects to study groups to even out unknown confounders
- Matching subjects on potential confounders

- In data analysis
  - Stratified analysis using the Mantel Haenszel method to adjust for confounders
  - Case-control studies
  - Model fitting using regression techniques

The advantages and disadvantages of these methods are:

- Matching methods call for subjects with exactly the same characteristics and have a risk of either over or under matching
- In stratified analyses some strata might become too small and thus create loss of information
- For regression methods the techniques already developed can lead to estimation problems when the data is not handled properly

The goal of this work is to provide solutions and guidelines on how to manage the confounding variables especially when the outcome is binary. We will focus on two related issues. First we will examine the potential bias of a selected matched sample from the original independent samples, and compare the power of these two approaches for the inference of two population proportions or means. Secondly, we propose a novel method of logistic regression controlling for confounding variables, the 'residual logistic

regression analysis'. This approach is compared to other existing methods including the Pearson residual analysis and the hierarchical logistic regression analysis. Comparisons are made through simulation studies as well as in a real life data set from the NYU Alzheimer's Disease Core Center comparing the mental decline rates between subjects with or without subjective complaints of cognitive impairment. The latter has motivated our research from the very beginning.

Paired vs. Independent Samples

In the last decades, the problem of whether or not to match data in clinical trials with dichotomous response has gathered much attention[74].The motivation for our work in this area originates from the analysis of data on aging and Alzheimer's Disease collected through the NYU Alzheimer's Disease Core Center[1]; such data consist on subjects that were recruited initially trough clinical referrals or voluntary enrollment (upon reading advertisements from internet or newspapers or other sources). These subjects are then followed up and monitored periodically on the conditions of their mental and Physical health. Given such a study design the existence of confounding variables became unavoidable. For example the individuals considered in the study were classified into different cognitive level using the Global Deterioration Scale [Barry Reisberg, et. al, 1999][56], which consist of seven different stages; the first two levels of this measurement represent subjects with full cognitive capabilities, while the third level

[1] This is one of thirty centers nationwide sponsored by the National Institute on Aging

represents Mild Cognitive Impaired people and levels beyond this are considered Alzheimer sufferers.

Even though the first two levels represent subjects with their full potential in cognitive performance, their difference is that individuals in level 1 (we will call GDS1) are normal and people in level 2 (GDS2) have complains of cognitive impairment; the goal is to compare if, along the study, they stayed within these two levels or if they progressed to higher levels on the Global Deterioration Scale (GDS); i.e. if there was any significant difference on the way subjects achieved higher or equal to GDS3 levels, depending on what was their initial status.

From this we can see that we are dealing with a sample where the subjects are categorized into two different groups which, after doing a preliminary statistical analysis, showed significant differences with respect to their demographic characteristics, such as, age and gender among others[2], therefore, the analysis tat can be performed has two options:

a) To analyze the initial independent samples of GDS 1 and GDS 2 subjects
b) To match the subjects from the two groups on their demographic characteristics and, subsequently, analyze the smaller paired samples.

In this work, we compare these two approaches through straightforward theoretical derivations as well as simulation and resampling studies to elucidate the pros and cons of

[2] For more detailed information see Chapter 5

each method and to provide guidelines to analysts in the field as to when to use which method.

The topic of concern is vital in statistical inferences, and thus much effort has been devoted in such comparisons. Regrettably, little can be found in formal documentations. Empirical study has shown that both the paired and the unpaired methods have their own advantages and disadvantages[39]. For example, the independent samples tend to have smaller standard errors (for the individual samples) and more degrees of freedom than an extracted paired sample. On the other hand, when pairing we are able to reduce the variability between and within subjects. Also, by pairing we control better for confounding variables and often yield more statistically efficient analyses [7, 60].

In our situation, we also wary about the potential bias induced through the matching process[4]. The modern bootstrap resampling methodology is employed to gauge whether a certain matched sample is unbiased.

[4] When subtracting a matched sample from an independent some of the pairs are created in an objective way, specially if we are dealing a 1:1 matched sample where, sometimes for a case, we will have more than one control that can be matched with and the final decision of which one is chosen is left to the person gathering the sample, therefore if we think about it, we could gather a considerable amount if different data sets where at least one pair will not be equal and that might or might not change the conclusions we make about the sample. That is why is important to analyze if the paired sample being studied is a good representation of all the possible paired samples that can be obtained.

Residual Logistic Regression

Among various types of statistical analysis, regression analysis is perhaps the most ideal platform for incorporation and filtration of confounding variables. For multiple regression analysis where the response variable is continuous, the two-stage residual linear regression analysis strategy has been well developed and adopted through the years[10, 33, 42]. Such is not true for logistic regression analysis (or any generalized linear models other than the general linear model) when the response is categorical. In this work, we propose a novel two-stage residual logistic regression analysis in the same spirit as the residual linear regression analysis. However, our model features a more general rationale based on residual link function and can be naturally extended to any generalized linear model. This model is compared to other common strategies for confounding variable controlling including the Pearson residual analysis [36] and the hierarchical logistic regression analysis. Pros and cons are discussed and guidelines provided.

Chapter 1

Paired vs. Independent Data Analysis

As discussed previously, extracting matched pairs from the original independent samples [7, 39] present a solution to the problem of potential confounding variables in certain data analysis scenarios. In this Chapter, we discuss the evaluation of matched sample bias using the bootstrap resampling method. We also present some quick theoretical and simulation studies comparing the power using the paired or the independent samples approaches for inferences on population means and proportions to extend the documentation in published.

In order to perform a paired sample analysis the variables (except the dependent variable of interest) that will be analyzed need to be separated into two categories: potential confounding variables and non-confounding variables. This is usually determined through prior knowledge and a univariate test comparing the distribution of these variables between groups of interest. For example, when comparing the decline rate or

time to decline to GDS stage 3 or above, for subjects who are at the GDS 1 or GDS 2 stage upon initial enrollment, we will first conduct an independent samples tests (parametric or nonparametric) comparing age, gender, years of education etc. between the two groups (GDS 1 and GDS 2). All these factors listed are deemed relevant to mental decline based on field knowledge. The subsequent univariate test indicated significantly different distribution on any of these covariates, they are deemed as potential confounding variables this study. Thus, they should be used to form the strata for the subsequent matching and matched pairs selection [39].

In our particular analysis we will consider 1:1 matching, also it is important to point out that before this paired sample is analyzed we need to validate whether this sample is unbiased. That is, whether or not it is representative of all the possible paired samples that can be obtained through the original independent samples. We employed a resampling method for such evaluations.

1.1 Resampling

Resampling is a nonparametric statistical method tied to Monte Carlo simulations where the main idea is to take samples from the original sample (of the same size or smaller) in order to obtain estimates and confidence intervals for population parameters without making assumptions about the form of the population distribution [4, 8, 25].

There are four main types of resampling [26, 31, 44]: cross validation [61, 69], permutation test [30, 44], jackknife [27, 29, 62] and bootstrap [21, 25, 26, 44, 63]. Among these methods bootstrap is the most general and popular methodology, mainly because the other methods are neither as flexible nor as reliable as this method is; the bootstrap method allows the researcher to draw as many sub-samples as possible (unlike, for example, the jackknife, that is limited by the sample size). Besides, bootstrap usually provides less biased and more consistent results and even though the other methods might be easier to compute for certain situations, this method is usually a better option, specially because it can be applied to any statistic [23].

The main idea behind bootstrap goes back at least two centuries, and it was Bradley Efron [25] who developed it, Moone and Duval in the gave also important applications of this technique [50]. Where the basic algorithm is as follows:

1. Given a sample $X = (x_1, x_2, ..., x_n)$ of size *n*, a bootstrap subsample of it is a sample of the form $X^* = (x_1^*, x_2^*, ..., x_p^*)$, $p \le n$, where each value x_i^* is randomly sampled with replacement from x, therefore for distinctive values

   $$P(x_j^* = x_i) = 1/n \qquad 1 \le i \le n$$

   with independent choices of x_j^* for $1 \le j \le n$. Therefore repeated values are allowed since the sample size of x_j^* is p, then some values in *x* will be left out.

2. Compute the statistic $\hat{\theta}^*$ for the subsample obtained.

3. Repeat steps 1 and 2, B number of times (B=1000 typically [8]) where the bootstrap resampled values for the estimator are:

$$\hat{\theta}_k^* = \hat{\theta}\left(\left(X^*\right)^{(k)}\right) \qquad 1 \le k \le B$$

4. Approximate the standard error and the mean of the bootstrap replications as follows:

$$se_{boot} = \left\{\frac{\sum_{b=1}^{B}\left(\hat{\theta}^* - \hat{\theta}\right)}{B-1}\right\}^{1/2} \qquad \hat{\theta} = \sum_{b=1}^{B}\frac{\hat{\theta}^*}{B}$$

5. So, calculate the confidence intervals, the percentile method is employed, which uses the $\alpha/2$ and the $1-\alpha/2$ quantiles for the $1-\alpha$ level confidence interval.

For example, the 95% confidence interval for θ would be constructed with the upper and lower 2.5% quantiles of the sampled values $\hat{\theta}_k^*$. Specifically, the bootstrap percentile 95% confidence interval for θ is $\left(\hat{\theta}^*_{(U)}, \hat{\theta}^*_{(B+1-U)}\right)$ where for this purpose, the values $\hat{\theta}_k^*$ are sorted in increasing order $\hat{\theta}^*_{(1)} \le \hat{\theta}^*_{(2)} \le \ldots \le \hat{\theta}^*_{(B)}$ where $U = \left(\alpha/2\right)$, with U rounded down to the nearest integer, the only exception to this when $0<U<1/2$, in which U is rounded up to 1.

1.1.1 Resampling method for paired data

The bootstrap method just presented is mainly designed for independent data. Because of that we proposed a method to generate match-paired samples at random [39] and, subsequently an entire set of them, in order to determine if the sample that will be used for the analysis is representative of the underlying distribution of all the possible subsamples in terms of the major variables under study.

The mechanism is as follows:

1. Paired sample database.

   In order to obtain a match-paired sample it is necessary to determine which subjects can be matched based on the matching factors criteria and which will be ignored.

   Let consider 1:1 pairing, for this just a set of individuals from the original dataset will qualify to be matched, i.e. have a potential match, therefore they will be on a "new" sample we will call paired-sample database where, a subject might have more than one option for pairing and therefore which will create clusters of options.

   For example, let's consider $X_1, X_2 . X_3, X_4$ as the matching factors in that order of importance, therefore the process would be as follows: First separate the sample by X_1 (in our particular example is the two GDS levels), creating with this clusters for each value of it; second, each of these clusters is divided by X_2

creating even more groups (in our case this variable is Age and we create Age-GDS level clusters); third, each of these new groups is divided by the third matching factor (in our case this variable is Gender and we will create GDS level – Age – Gender clusters) and so on. It is important to point out that the grouping should finish before the strata become too small and too much information is lost, therefore it is likely that we will ignore the criteria of the least important factors in order to obtain a substantial sample[7]. At the end the clusters created are the matched depending on their characteristics (see Fig.1)

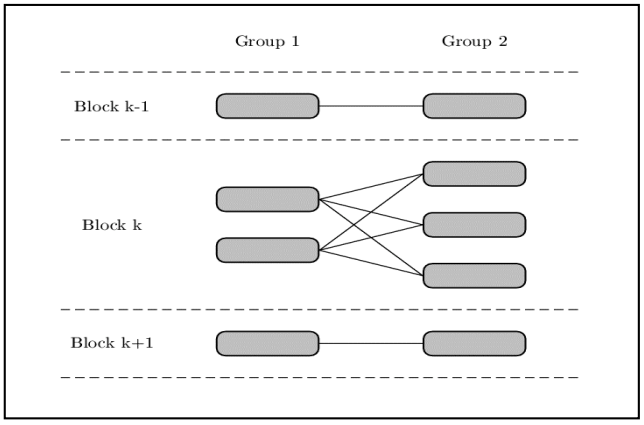

Figure 1: Paired sample database and distribution.
As mentioned before, in our case X_1 is the GDS levels (2 groups, GDS1 and GDS2) and X_2 is the Age; it can be seen that a block of subjects in group 1 can be paired with a block of subjects in group 2, and, given that we are performing a 1:1 matching, we have several options for each subject. This situation can be observed several times in the match-paired sample database together with cases where there is only 1 subject in a block that can be matched with a block that also has only one person.

2. Generate a random sample.

   Because we want to create bootstrap resamples, the following must be done:

   a. Determine how many pairs can be obtained from each matched cluster.

[7] We would like to keep a big percentage of the original data's information on the new match-paired data set.

b. Sample with replacement B times from each cluster (If the cluster contains only one subject then this subject will be always selected for the paired samples. If there are more than one, say three, but only two pairs can be created from it then two subjects will be sampled with replacement each time)

3. Examine the distribution of the paired samples.

   This is done in order to set the standards of a "good" paired sample. Therefore, the standard error, mean and confidence intervals of the main variables are calculated. If the paired sample that will be used on the analysis behaves within the characteristics given by the bootstrap samples, then we can proclaim that it is a representative sample.

   As an additional check, the same analyzes that are performed on the single paired data can also be performed on all these subsamples in order to observe the distribution of the estimators. This can be considered optional given that it requires very extensive computations.

The SAS code that we employed for the creation of this database after clustering the data as we just mentioned in Fig.1 is as follows

Renaming each cluster

```
%MACRO LOOP;
%DO i = 1 %TO 60;
```

```
DATA b&i;
SET RESAMPLE.TEST;
IF b=&i;
RUN; %END; %MEND; %LOOP;
```

For cluster with more subjects than the amount of pairs that can be obtained from them we reduce them randomly

```
DATA b1a;
    SET b1;
    RANDOM = RANUNI(0);
    RUN;
    PROC SORT;
    BY RANDOM;
    RUN;
    DATA b1b;
    SET b1a;
    IF _N_<2;
    DROP RANDO; RUN;
```

Bootstraping per cluster 1000 samples are taken

```
%MACRO BOOT;
%DO i = 1 %TO 1000;
DATA ANALYSIS_BOOT_1;
CHOICE = INT(RANUNI(2765551+&i)*n)+1;
SET bi POINT = CHOICE NOBS = n;
```

```
j+1;
IF j > n THEN STOP;
RUN;
DATA ANALYSIS_BOOT1;
ST ANALYSIS_BOOT_1;
BOOTSAMPLE=&i; RUN;
%IF &i = 1 %THEN %DO;
DATA BOOTS1;
SET ANALYSIS BOOT1;
RUN;
%END;
%ELSE %DO;
DATA BOOTS1;
SET BOOTS1 ANALYSIS_BOOT1; RUN; %END;
```

Merging all the bootstrap samples into a whole data set

```
DATA ALL_BOOT;
MERGE BOOTS1 BOOTS2 BOOTS3 BOOTS4 BOOTS5 BOOTS6 BOOTS7
BOOTS8 BOOTS9 BOOTS10;
BY b; RUN; %END; %MEND; %BOOT;
```

1.2 Power

The power of a statistical test is highly correlated to the sample size, where the test commonly describes how close or different are the subgroups that conform the sample that is being analyzed.

In our analysis we will consider two data sets, the independent and the paired one, we assume each of these samples can be divided by two exclusive strata (case and controls) and that they have a common response variable. We will examine inferences on both means and proportions for large samples.

In hypothesis testing we have the possibility of committing two types of errors and, by definition, the power is the complement of the Type II error rate, defined as:

$$1-\beta = P\left(z \geq Z_{1-\alpha} | H_1\right) \qquad (1)$$

Here α is defined as the Type I error rate or the significance level. Also, for a one-sided test a general equation relating the difference being analyzed Δ ($\Delta = \mu_1 - \mu_2$ for means and $\Delta = \pi_1 - \pi_2$ for proportions), the Type I error and the Type II is [40]:

$$|\Delta| = Z_{1-\alpha}\sigma_0 + Z_{1-\beta}\sigma_1 \qquad (2)$$

1.2.1 Inference on means

Let X be a random sample of size N from a population distributed $N(\mu,\sigma)$ consisting of two exclusive strata $X_1 \sim N(\mu_1,\sigma_1)$ and $X_2 \sim N(\mu_2,\sigma_2)$[8] where n_1 and n_2 are the sample sizes respectively.

- Independent sample design

For a normally distributed population a z-test can be considered where the hypotheses for the comparison of two means is as follows:

$$H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0 \qquad vs \qquad H_1 : \mu_1 > \mu_2 \Rightarrow \mu_1 - \mu_2 = \Delta > 0$$

Let the sample sizes of the two groups be such that $n_1 = n$, $n_2 = kn$ with $k \geq 1$ and $\sigma_1 = \sigma_2 = \sigma$, with σ unknown therefore, because of this last characteristic, the standardized z-test is based on the T statistic[9]

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \Delta}{S\sqrt{1/n_1 + 1/n_2}}$$

where $\overline{X}_i = \sum \frac{x_{ij}}{n_i}$, $S = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$ and $S_i = \sqrt{\frac{\sum (x_{ij} - \overline{X}_i)^2}{n_i - 1}}$ $\qquad i = 1,2$

[8] Because the overall population is distributed normal we can consider each of the stratums to be normally distributed as well.

[9] This test was introduced by William Sealy Gosetin 1908 [27]

Therefore, the power of the test is given by

$1-\beta = P\left(T > t_{n_1+n_2-2,\alpha} \middle| H_1\right)$ where $n_1 + n_2 - 2 = \zeta$, therefore

$$= P\left(\overline{X}_1 - \overline{X}_2 > t_{\zeta,\alpha} S\sqrt{1/n_1 + 1/n_2} \middle| \mu_1 - \mu_2 = \Delta\right)$$

$$= P\left(\frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{1/n_1 + 1/n_2}} > t_{\zeta,\alpha} \middle| \mu_1 - \mu_2 = \Delta\right)$$

$$= P\left(\frac{\overline{X}_1 - \overline{X}_2 - \Delta}{S\sqrt{1/n_1 + 1/n_2}} > t_{\zeta,\alpha} - \frac{\Delta}{S\sqrt{1/n_1 + 1/n_2}} \middle| \mu_1 - \mu_2 = \Delta\right)$$ where $eff = \frac{\Delta}{\sigma} \approx \frac{\Delta}{S}$,

therefore

$$= P\left(T > t_{\zeta,\alpha} - \frac{eff}{\sqrt{1/n_1 + 1/n_2}} \middle| \mu_1 - \mu_2 = \Delta\right)$$

- Paired sample design

Following the same criteria as with the independent data analysis; the hypotheses to be tested are

$H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$ *vs* $H_1 : \mu_1 > \mu_2 \Rightarrow \mu_1 - \mu_2 = \Delta > 0$

For this case, let the sample size be $n_1 = n_2 = n$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Because this is a paired sample we have that

$$\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2Corr(x_1,x_2)\sigma_1\sigma_2 \qquad \text{if } Corr(x_1,x_2) = \rho \quad \text{then}$$

$$= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 = 2\sigma^2(1-\rho)$$

This means that for the paired analysis the variance is increased.

Therefore the test statistic is

$$T = \frac{\bar{d} - \Delta}{S_D/\sqrt{n}} \qquad \text{where } \bar{d} = \frac{\sum d_i}{n} = \frac{\sum (x_{1,i} - x_{2,i})}{n} \text{ and } S_D \approx \sigma_D = 2\sigma(1-\rho) \approx 2S(1-\rho)$$

Given this, the power is

$$1-\beta = P(\bar{X}_D > t_{n-1,\alpha} | H_1)$$

$$= P\left(\bar{X}_D > t_{n-1,\alpha} S_D \sqrt{1/n} | \Delta > 0\right)$$

$$= P\left(\frac{\bar{X}_D}{S_D\sqrt{1/n}} > t_{n-1,\alpha} | \Delta > 0\right)$$

$$= P\left(\frac{\bar{X}_D - \Delta}{S_D\sqrt{1/n}} > t_{n-1,\alpha} - \frac{\Delta}{S_D\sqrt{1/n}} | \Delta > 0\right)$$

where $eff = \frac{\Delta}{\sigma} \approx \frac{\Delta}{S} \Rightarrow \frac{\Delta}{S_D} = \frac{\Delta}{S\sqrt{2(1-\rho)}} = \frac{eff}{\sqrt{2(1-\rho)}}$, therefore

$$1-\beta = P\left(T > t_{n-1,\alpha} - eff\left(\frac{n}{2(1-\rho)}\right)^{1/2} | \Delta > 0\right)$$

- *Trials*

K	n	eff.	Independent	Paired		
				$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$
1	25	0.5	0.5487	0.6279	0.7814	0.9594
2			0.6534			
3			0.6979			
4			0.7222			
1	50		0.8027	0.8896	0.9677	0.9992
2			0.8917			
3			0.9210			
4			0.9347			
1	100		0.9700	0.9916	0.9994	1.0000
2			0.9923			
3			0.9962			
4			0.9976			
1	150		0.9962	0.9995	0.9999	1.0000
2			0.9957			
3			0.9986			
4			0.9999			
1	25	1	0.9676	0.9869	0.9985	0.9992
2			0.9914			
3			0.9927			
4			0.9973			
1	50		0.9994	0.9993	0.9999	1.0000
2			0.9997			
3			1.0000			
4			1.0000			
1	100		1.0000	1.0000	1.0000	1.0000
2			1.0000			
3			1.0000			
4			1.0000			
1	150		1.0000	1.0000	1.0000	1.0000
2			1.0000			
3			1.0000			
4			1.0000			

Table 1: Simulations performed to estimate the power of a paired and an independent sample given the characteristics described on the top of the table. We can observe that when the sample gets bigger and the variance smaller then the power converges to one.

1.2.2 Inferences on Proportions

Consider two Bernoulli populations with parameters π_1 and π_2 . Independent random samples of sizes n_1 and n_2 available for these two populations. Let $X_1 \sim Bin(n_1,\pi_1)$ and $X_2 \sim Bin(n_2,\pi_2)$ [67].

- Independent samples

The test for two proportions is based on the test statistic $T = p_1 - p_2$ under [12, 40]

$$H_0 : \pi_1 = \pi_2 \Rightarrow \pi_1 - \pi_2 = 0 \qquad vs \qquad H_1 : \pi_1 > \pi_2 \Rightarrow \pi_1 - \pi_2 = \Delta$$

where

$$E(T) = \pi_1 - \pi_2$$

$$\sigma_0^2 = \frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2} = \pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \text{ under } H_0$$

$$\sigma_1^2 = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} \qquad \text{and } \mu_1 = \pi_1 - \pi_2 \text{ under } H_1$$

Using equation (2), we obtain that

$$|\Delta| = Z_{1-\alpha}\sigma_0 + Z_{1-\beta}\sigma_1$$

$$= Z_{1-\alpha}\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} + Z_{1-\beta}\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

Subsequently we can obtain the power (1-β) from:

$$Z_{1-\beta} = \frac{\Delta - Z_{1-\alpha}\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} = \frac{\Delta(n_1 n_2)^{1/2} - Z_{1-\alpha}\left[\pi(1-\pi)(n_1+n_2)\right]^{1/2}}{\left[n_2\pi_1(1-\pi_1)+n_1\pi_2(1-\pi_2)\right]^{1/2}}$$

- Paired Samples

In 1947 the American psychologist Quinn McNemar introduced a non-parametric method designed for nominal data to determine the difference between paired proportions. In recent years, researchers like Connet, Smith and McHugh [11] among others, derived the power for the McNemar's [24] test.

For this example, let us consider a match-paired large sample of size M with a 1-1 matching where each member was sampled independently. Each subject had a response $Y_i = 0,1$ (*i=1,2* depending on the group that the subject belongs). Therefore, for each pair there are four possibilities that depend on these responses as follows

		Y_2 = 0	Y_2 = 1
Y_1	0	p_{00}	p_{01}
	1	p_{10}	p_{11}

Where $p_{ij},(i=1,2\ ;\ j=1,2)$, such that $\sum p_{ij}=1$are the probabilities that an individual has a response $Y_1=i$ and $Y_2=j$ with sample sizes $n_1=n_2=n$

The test statistic is based on $T=p_{10}-p_{01}$ under the hypotheses [24]

$$H_0:\pi_{01}=\pi_{10}\Rightarrow\pi_{01}-\pi_{10}=0 \qquad vs \qquad H_1:\pi_{01}>\pi_{10}\Rightarrow\pi_{01}-\pi_{10}=\Delta$$

Let the estimator of the proportion be such that

$$\begin{pmatrix} p_{11} \\ p_{10} \\ p_{01} \\ p_{00} \end{pmatrix} \sim N\left(\begin{pmatrix} \pi_{11} \\ \pi_{10} \\ \pi_{01} \\ \pi_{01} \end{pmatrix}, \frac{1}{n} \begin{pmatrix} \pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{10} & -\pi_{11}\pi_{01} & -\pi_{11}\pi_{00} \\ -\pi_{10}\pi_{11} & \pi_{10}(1-\pi_{10}) & -\pi_{10}\pi_{01} & -\pi_{10}\pi_{00} \\ -\pi_{01}\pi_{11} & -\pi_{01}\pi_{10} & \pi_{01}(1-\pi_{01}) & -\pi_{01}\pi_{00} \\ -\pi_{00}\pi_{11} & -\pi_{00}\pi_{10} & -\pi_{00}\pi_{01} & \pi_{00}(1-\pi_{00}) \end{pmatrix} \right)$$

Therefore

$E(T_0)=0$ under H_0

$E(T_1)=E(p_{10})-E(p_{01})=\pi_{10}-\pi_{01}$ under H_1

Let, for the null hypothesis $\pi_{10}=\pi_{01}=\pi_d/2$, where π_d is the chance of a discordant pair, and

$$\sigma^2=Var(p_{10}-p_{01})=Var(p_{10})+Var(p_{01})-2Corr(p_{10},p_{01})$$

therefore

$\sigma_0^2=\frac{\pi_d}{n}$ under H_0

$$\sigma_1^2 = \frac{1}{n}\left[\pi_{01} + \pi_{10} - (\pi_{01} - \pi_{10})^2\right] = \frac{1}{n}\left[\pi_d - (\pi_{01} - \pi_{10})^2\right] \text{ under } H_1$$

Following equation (2), we obtain

$$|\Delta| = Z_{1-\alpha}\sqrt{\frac{\pi_d}{n}} + Z_{1-\beta}\sqrt{\frac{\pi_d - (\pi_{01} - \pi_{10})^2}{n}}$$

Subsequently we can obtain the power (1-β) from:

$$Z_{1-\beta} = \frac{\Delta - Z_{1-\alpha}\sqrt{\frac{\pi_d}{n}}}{\sqrt{\frac{\pi_d - (\pi_{01} - \pi_{10})^2}{n}}} = \frac{\Delta n^{1/2} - Z_{1-\alpha}\pi_d^{1/2}}{\left[\pi_d - (\pi_{01} - \pi_{10})^2\right]^{1/2}}$$

- *Trials*

n	P_1	P_2	Independent		Paired
			k=1	k=2	
50	0.2	0.2	0.8051	0.8389	0.9505
		0.4	0.9904	0.9985	1.0000
		0.6	1.0000	1.0000	1.0000
		0.8	1.0000	1.0000	1.0000
	0.4	0.2	0.6179	0.7054	0.5753
		0.4	0.8496	0.8749	0.9505
		0.6	0.9916	0.9985	1.0000
		0.8	1.0000	1.0000	1.0000
	0.6	0.2	0.9177	0.9686	0.9554
		0.4	0.5832	0.6700	0.5948
		0.6	0.8461	0.8749	0.9505
		0.8	0.9904	0.9985	0.9980
	0.8	0.2	0.9918	0.9988	1.0000
		0.4	0.9177	0.9686	0.8438
		0.6	0.6179	0.7054	0.6772
		0.8	0.8051	0.8389	0.9505
100	0.2	0.2	0.7291	0.7734	0.9505
		0.4	0.9953	0.9864	1.0000
		0.6	1.0000	1.0000	1.0000
		0.8	1.0000	1.0000	1.0000
	0.4	0.2	0.8599	0.6664	0.8340
		0.4	0.7642	0.8133	0.9505
		0.6	0.9956	0.9875	1.0000
		0.8	1.0000	1.0000	1.0000
	0.6	0.2	0.9956	0.9345	1.0000
		0.4	0.8432	0.6368	0.6406
		0.6	0.7642	0.8133	0.9505
		0.8	0.9953	0.9864	1.0000
	0.8	0.2	1.0000	0.9943	1.0000
		0.4	0.9955	0.9345	0.9846
		0.6	0.8599	0.6664	0.5199
		0.8	0.7291	0.7734	0.9505

Table 2: Simulations for the estimation of the power for a paired and an independent sample given the characteristics described on the top of the table. We can observe that when the sample gets bigger and the variance smaller then the power converges to one. These calculations differ from the ones obtained by Wacholder and Weinberg [74] since we are summing large samples

1.3 Logistic Regression

Regression models are part of Generalized linear Models and represent a powerful device for analyzing data. They allow the researcher to focus on the behavior of a dependent variable as a function of one or more covariates. They can be used for prediction, inference, hypothesis testing and modeling causal relationships.

Logistic regression is a particular kind of regression analysis designed specifically for dichotomous outcomes. For this model we consider the odds of the event occurring for each individual on the population as the dependent variable, which creates the basic logistic regression equation of the form [54, 55]

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{j=1}^{t} x_{ij}\beta_j$$

Where π_i is the probability of success for the i[th] subject. It can be obtained as follows

$$\pi_i = \frac{e^{\alpha+\sum_{j=1}^{t} x_{ij}\beta_j}}{1+e^{\alpha+\sum_{j=1}^{t} x_{ij}\beta_j}}$$

Each observation y_i follows a Bernoulli distribution. Therefore the estimators are obtained using MLE. The likelihood in terms of the probability of success of each individual is [54]

$$L(\beta) = \prod_{i=1}^{n}\left(\frac{e^{\alpha+\sum_{j=1}^{t} x_{ij}\beta_j}}{1+e^{\alpha+\sum_{j=1}^{t} x_{ij}\beta_j}}\right)^{y_i}\left(\frac{1}{1+e^{\alpha+\sum_{j=1}^{t} x_{ij}\beta_j}}\right)^{1-y_i}$$

In the following Chapter, we present a novel method of residual logistic regression for the controlling of confounding variables in logistic regression analysis based on the original independent samples. When we use the matched sample extracted from the original independent samples, we can utilize the conditional logistic regression model customized for paired data as discussed below.

1.3.1 Logistic Regression for Paired Data

This model was designed for retrospectively matched samples of cases and controls, for either 1:1 matching or 1:N. This model assumes that the covariates have a common effect for the odds of response for all the matched sets and allows that each one of them have a unique risk of the response given by the intercept α_i.

There are two possible methods that can be applied for this purpose: conditional and unconditional logistic regression. The first one is used for finely matched case-control studies, i.e. when the number of observations in each matched set is small. The second one might be used when the sample is frequency matched and/or when we are interested in including the matching variables as explanatory variables. That is actually our case, because that way we will be able to analyze better the behavior of the confounding variables and also, that way we will e able to compare the method more accurately.

Now, consider a matched set formed by N pairs where individuals are being matched by a set of covariates (the matching factors). Let us consider a 1:1 case-control study where one subject from the case group is matched with a subject from the control group by having the almost the same, if not identical, characteristics under the matching factors. Denote x_{ij} as the binary indicator of the pairs, where $x_{ij} = 0$ means that the subject *j* (j=1,2) is from the control group and $x_{ij} = 1$ that the subject is from the case group. Let also consider the outcome of each member of the pairs as $y_{ij} = 1$ when the individual has a positive outcome and 0 when it doesn't.

The unconditional logistic regression model for a single covariate with a single binary covariate Cox [17] suggested using a model assuming with a constant odds ratio for the pairs to be tested such as:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha_i + \beta x_{ij} \qquad i = 1,...,N; j = 1,2$$

where

$$\begin{aligned}\pi_{i1} &= P\left(Y_{i1} = 1 \middle| x_{i1} = 1\right) \\ &= \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} \\ \pi_{i2} &= P\left(Y_{i2} = 1 \middle| x_{i2} = 0\right) \\ &= \frac{e^{\alpha_i}}{1 + e^{\alpha_i}}\end{aligned}$$

This leads to a log odds ratio that considers the exposed and non-exposed members as:

$$\log\left(\frac{\pi_{i1}}{1-\pi_{i1}}\right) - \log\left(\frac{\pi_{i2}}{1-\pi_{i2}}\right) = \alpha_i + \beta x_{i1} - \left(\alpha_i + \beta x_{i2}\right) = \beta$$

Now, each pair consist of two members that were sampled independently given the value of the matching variable. Therefore the terms of the members within each pair are conditionally independent [47] so that the likelihood is a binomial product for the i*th* independent pair as follows:

$$L_i(\beta) = \prod_{i=1}^{N}\prod_{j=1}^{2} \pi_{ij}^{\,y_{ij}} \left[1-\pi_{ij}\right]^{1-y_{ij}}$$

$$= \prod_{i=1}^{N} \pi_{i1}^{\,y_{i1}} \left[1-\pi_{i1}\right]^{1-y_{i1}} \pi_{i2}^{\,y_{i2}} \left[1-\pi_{i2}\right]^{1-y_{i2}}$$

$$= \prod_{i=1}^{N} \left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)^{y_{i1}} \left(\frac{1}{1+e^{\alpha_i+\beta}}\right)^{1-y_{i1}} \left(\frac{e^{\alpha_i}}{1+e^{\alpha_i}}\right)^{y_{i2}} \left(\frac{1}{1+e^{\alpha_i}}\right)^{1-y_{i2}}$$

For multiple covariates (let consider t covariates) the likelihood of the model is of the form:

$$L_i(\beta) = \prod_{i=1}^{N}\prod_{j=1}^{2} \left(\frac{e^{\alpha_i+\sum_{p=1}^{t} x_{jip}\beta_p}}{1+e^{\alpha_i+\sum_{p=1}^{t} x_{ijp}\beta_p}}\right)^{y_{ij}} \left(\frac{1}{1+e^{\alpha_i+\sum_{p=1}^{t} x_{ijp}\beta_p}}\right)^{1-y_{ij}}$$

The conditional model actually comes from the model we just presented where a principle of conditioning originally attributed to Fisher is applied; Cox [16] assumed that the common log odds ratio can be estimated without estimating the nuisance parameters through conditioning where the disagreements between the subjects that conform each match are the ones that are being analyzed.

Chapter 2

Residual Logistic Regression

2.1 Residual Linear Regression

It is a very common analysis that implies two stages and the use of residuals. The first stage considers only the confounding factors in the model and the second considers the estimated error from the first stage as a dependent variable for a model where only the non-confounding variables will be tested. To illustrate this method let us consider a population of size N with a continuous outcome Y and t explanatory variables $X_1, X_2, ..., X_t$

Let k ($1 \le k \le t$) of the explanatory variables be potential confounding factors, therefore the model that will be fitted for Stage 1 is as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

The estimated error $e_j = Y_j - \hat{Y}_j$. These residuals [68] follow zero, constant variance, normal distribution and $Corr(e_i, e_j) \neq 0 \,, \forall i \neq j$

Stage two defines a new dependent variable to be tested that is not correlated any more with the confounding factors and is created with the residuals obtained from the first stage of the analysis. Therefore, let $Y^* = e_j$, this will be fitted with the rest of the variables (the non-confounding ones) as follows

$$Y - \hat{Y} = \beta_0^* + \beta_{k+1} x_{k+1} + \beta_{k+2} x_{k+2} + ... + \beta_t x_t + \varepsilon$$

This last model will be analyzed using a variable selection method (the most commonly applied is stepwise selection) in order to see which other variables add explanation to the variability on *Y* after controlling for the confounding factors.

The method just presented has several flaws [20] that come up because it has two possible sources of bias, one of which can make it too conservative and another one that can make it too liberal:

- Conservative bias

  Occurs specially when the independent variable of interest is correlated with the confounding variables and as this correlation increases the chances of this happening are grow.

- Liberal bias (gives significant results more often that it should)

It is strongest when four conditions meet: Small sample size, many confounding variables, independent variable of interest is independent of the confounding variable and there are fixed scores i.e. that the scores do not vary randomly in the sample but are rather fixed.

2.2 Residual Logistic Regression

In logistic regression analysis, it is necessary to account for covariates such as gender and race. The traditional hierarchical logistic regression modeling approach does not facilitate variable selection and suffers from collinearity [65]. Here we propose a novel method of '*residual logistic regression analysis*' for controlling confounding covariates (gender, race, etc.) when the response variable *Y* (whether the subject is a case) is dichotomous. This method would enable best-subset or stepwise variable selection among variables of interest such that the maximum amount of risk explainable by covariates of interest can be estimated while accounting for potential confounding factors. Our procedure is a novel two-stage logistic regression analogous to that of the residual linear regression analysis and can be summarized as follows.

Let $x_1, x_2, ..., x_k$ be potential covariates/confounding variables such as gender, race, etc.

Stage 1.

Fit $\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$

to obtain $T = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k$

Stage 2.

Fit the *residual logistic link function* [38] using genetic variables of interest $x_{k+1}, x_{k+2}, ..., x_t$ as follows:

$$\ln\left(\frac{\pi}{1-\pi}\right) - T = \beta_0^* + \beta_{k+1} x_{k+1} + ... + \beta_t x_t$$

This is equivalent to fitting a new logistic regression of the form:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0^* + T + \beta_{k+1} x_{k+1} + ... + \beta_t x_t$$

where T is a variable with a fixed coefficient of 1. Now one can compile a customized SAS or MATLAB program to perform stepwise or best-subset variable selection among variables of interest (while holding T in the equation) to achieve maximum prediction power.

The merits of this procedure are

1) Easy variable selection

2) Intuitive interpretation of risk odds accounted for by covariates of interest since both levels are logistic regression analysis.

The rationale for our approach in terms of fitting the residual link function on stage 2 is in complete agreement with the residual linear regression analysis where its second stage can be equivalently expressed in residual link function as follows:

$$E(Y) - \hat{Y} = \beta_0^* + \beta_{k+1} x_{k+1} + \ldots + \beta_t x_t$$

2.3 Pearson Residual Analysis

In linear regression the residuals are the difference between the observed and predicted values of *Y* and they are calculated straightforwardly from the regression equation. But in logistic regression the error variance is a function of the conditional mean unlike in linear regression where this error is independent of such conditional mean. For that reason residuals in logistic regression need to be standardized[31] .

The most common residual for logistic regression is the Pearson or standardized or chi residual [49]

$$Z_j = \frac{P(Y_j = 1) - P(\hat{Y}_j = 1)}{\sqrt{P(\hat{Y}_j = 1)\left[1 - P(\hat{Y}_j = 1)\right]}}$$

which, for large samples is normally distributed with mean 0 and standard deviation 1, where large positive or negative values of it indicates a poor fit for case *j* and where the estimated probability that we will be using here is the one obtained from stage 1 as:

$$\hat{\pi}(x) = P\left(\hat{Y} = 1 \middle| x_1,...,x_k\right) = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i}}$$

There are other alternatives for the estimation of the residuals in logistic regressions such as, deviance residual, which is basically the predicted probability of correct group, is the contribution of each case; and the logit residual, which is very similar to the Pearson residual, but it is divided by its variance instead of its standard deviation [49].

For our approach we will consider the Pearson residual due to the similarity that it has to the residual used in Residual linear Regression and because of the advantage that it has when we are analyzing large samples by having an asymptotic distribution normal.

Therefore, the second stage of our method will consist of considering this Pearson residual as the new dependent variable that will be used as dependent variable in stage 2. So, for a sample large enough, this new outcome is normally distributed. Let $Y_j^* = Z_j \in (-\infty,\infty)$. Given this characteristic then the second part of the analysis can be done by using a linear regression model instead of a logistic one of the form:

$$Y^* = \beta_0 + \beta_{k+1} x_{k+1} + \beta_{k+2} x_{k+2} + ... + \beta_t x_t + \varepsilon$$

As we did in Residual Linear Regression, we will perform a variable selection using stepwise selection in order to see which non confounding variables can be added to the explanation of the original dependant variable.

Of course there are possible flaws of this method and they are similar to the ones we had with residual linear regression:

- The Pearson residual is considered a number that summarizes the agreement between the observed and the fitted values. Its advantage (as well as disadvantage) is that is a single number that summarizes too much information [Hosmer and Lemeshow] [36]
- Given that this residual, like the one in linear regression is measuring the possible error of the fit, it has similar potential problems when it is being used as a new dependant variable, specially the one concerning conservative bias. This is because when the residuals are obtained from Stage 1 they are usually uncorrelated with all the confounding factors which makes them unable to be correlate with any variable which itself is correlated to the confounding variables, no matter how much that variable might be contributing to the model.

2.4 Hierarchical Logistic Regression

Multilevel or hierarchical models can be considered as extensions of regressions models [34] where data can be structured in groups and the coefficients can vary depending on the group.

When working to control confounding factors we would like to see two different things; one, to minimize the possible amount of bias that the confounding factors might add to the analysis. Two, to still consider the effect these variables are providing to the model without this affecting the final results. Therefore we need to control then in order to see what else can explain the variability of the outcome.

With simple logistic regression we cannot really do this labor [6], given that all variables are treated equally and therefore we aren't really controlling for any effect. Residual Logistic Regression, on the other hand, does control for the confounding variables effect. It involves several non automatic analysis just to obtain results for one single group of variables.

Hierarchical models are a very new technique that can deal with this problem automatically by grouping the data depending on the confounding factors using indicators that will specify the grouping. These kind of models are called *varying-intercept model*[10], because the model calculates a different intercept within each group[29].

Traditional techniques applied on hierarchical or multilevel data has two big problems. One is that they consider all the observations as independent without any type of correlation among individuals. However patients within the same outcome for a particular confounding variable such as demographic characteristics (that are very common

[10] HILL, G.J. *Data Analysis Using Regression and Multilevel Hierarchical Models*. Cambridge University Press, 2007

confounding effects) may share characteristics and their outcome are, most likely, dependant of one another.

On the other hand a big disadvantage of hierarchical models is that, if the groups created for each level of the confounding factors are very small we might suffer lack of power in the tests we perform and possibly bias [6].

Originally this kind of models where used to analyze nested sources of variability in hierarchical data; taking account of the variability associated with each level of hierarchy, that is, they take account of the variability at each level of hierarchy and thus allow each effect to be analyzed within the models accounting for clustering of observations while traditional logistic regression and residual logistic regression models assume independence of observations.

The basic idea of hierarchical modeling (also known as multilevel modeling, empirical Bayes, random coefficient modeling, or growth curve modeling) is to think of the lowest-level units (smallest and most numerous) as organized into a hierarchy of successively higher-level units. For example, in our data, subjects have an MMS level, with, depending on each MMS level then the Gender of the subject, then the Age and finally the initial GDS level. We can then describe outcomes for an individual as a sum of effects for the individual student, for her/his MMS level, for the gender, for the Age, for the GDS group. Each of these effects can often be regarded as one of an exchangeable

collection of effects drawn from a distribution described by a variance component. There may also be regression coefficients at some or all of the levels.

More formally, this analysis works as follows:

It starts with a traditional logistic regression model $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_{ij}$, where *i* is the subject level indicator and *j* is the level indicator for X.

A simple way to account for effects of higher-level units is to add dummy variables into the initial equation. These dummy variables are added as an intercept for each higher-level unit. The most sophisticated way to do this is to treat the intercept as a random variable with specified probability distribution, which leads to a random intercept model and more conservative estimate. The new model would be

$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_j + \beta x_{ij}$ where $\alpha_j = \alpha + u_j$

Here the effect is measured by the random intercepts α_j, a linear combination of a grand mean α and a deviation u_j from that mean and the intercepts from the independent variable measure the differences between the different levels, controlling for other effects in the models such as the risk factor of the subjects.

So the analysis is performed in several levels:

- Level 1, We express the outcome as the sum of an intercept for the subjects and the subjects risk factor
- Level 2: We specify the level intercept as the sum of an overall mean and the random deviation from that mean
- Finally: The equations for the two levels are combined into the equation

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + u_j + \beta x_{ij}$$

this is called a mixed model given that it has both fixed and random effects.

2.5 Quasi-Complete Separation

We will also consider the case of quasi-complete separation that occurs when a certain categorical covariate/group combination is null. For example, the controls do not have a certain genotype. The (ordinary or residual) logistic regression analysis may fail to converge in this case. If this happens, one should adopt the usual Pearson residual analysis for logistic regression. As discussed above (and recapped briefly here), this is also a two-stage procedure with Stage 1 being identical to our residual logistic regression. In Stage 2, letting $Y = 1$ if the subject is a case and 0 otherwise, one would compute the Pearson residual

$$Y^* = \frac{P\left(Y_j = 1\right) - P\left(\hat{Y}_j = 1\right)}{\sqrt{P\left(\hat{Y}_j = 1\right)\left[1 - P\left(\hat{Y}_j = 1\right)\right]}}$$

where $\hat{Y}$ is the Stage 1 estimate, and subsequently fit a linear regression model:

$$Y^* = \beta_0^* + \beta_{k+1}x_{k+1} + ... + \beta_t x_t + \varepsilon$$

Significant predictors can be chosen similarly using the stepwise or best-subset variable selection method. The drawback is that the intuitive interpretation of risk accounted for by covariates of interest is lost. The derivation of the quasi-complete separation is presented below in a simplified setting.

Proof: Quasi-complete separation

For a logistic regression model of the form:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where x_1, x_2 are two dummy variables representing a certain categorical covariate of interest with 3 possible categories, we have

$$f(\beta) = \frac{e^{\beta_0 n_1}}{\left(1+e^{\beta_0}\right)^{N_1}} \frac{e^{(\beta_0+\beta_2)n_2}}{\left(1+e^{\beta_0+\beta_2}\right)^{N_2}} \frac{e^{(\beta_0+\beta_1)n_3}}{\left(1+e^{\beta_0+\beta_1}\right)^{N_3}}$$

$$\ln f(\beta) = \beta_0 n_1 + (\beta_0+\beta_2)n_2 + (\beta_0+\beta_1)n_3 - N_1\left(1+e^{\beta_0}\right) - N_2\left(1+e^{\beta_0+\beta_2}\right) - N_3\left(1+e^{\beta_0+\beta_1}\right)$$

$$\frac{\partial \ln f(\beta)}{\partial \beta_1} = n_3 - N_3\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} = 0 \quad \Rightarrow \quad n_3\left(1+e^{\beta_0+\beta_1}\right) = N_3 e^{\beta_0+\beta_1} \quad \Rightarrow \quad e^{\beta_0+\beta_1} = \frac{n_3}{N_3-n_3}$$

$$\frac{\partial \ln f(\beta)}{\partial \beta_0} = n_2 + n_3 - N_2\frac{e^{\beta_0+\beta_2}}{1+e^{\beta_0+\beta_2}} = 0 \quad \Rightarrow \quad n_2\left(1+e^{\beta_0+\beta_2}\right) = N_2 e^{\beta_0+\beta_2} \quad \Rightarrow \quad e^{\beta_0+\beta_2} = \frac{n_2}{N_2-n_2}$$

$$\frac{\partial \ln f(\beta)}{\partial \beta_0} = n_1 + n_2 + n_3 - N_1 \frac{e^{\beta_0}}{1+e^{\beta_0}} - N_2 \frac{e^{\beta_0+\beta_2}}{1+e^{\beta_0+\beta_2}} - N_3 \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} = 0$$

$$\Rightarrow \quad n_1 - N_1 \frac{e^{\beta_0}}{1+e^{\beta_0}} = 0 \quad \Rightarrow \quad n_1\left(1+e^{\beta_0}\right) - N_1 e^{\beta_0} = 0 \quad \Rightarrow \quad e^{\beta_0} = \frac{n_1}{N_1 - n_1}$$

We will find that the maximum likelihood estimate of β_0 is ln(0) which is impossible to obtain when one of the categories has probability of 0. One solution to this problem is the Pearson residual analysis where the second stage regression is linear as discussed before.

Chapter 3.

Applications and Results

3.1 Data Overview

Alzheimer's Disease (AD) is one of the most common types of dementia that consist on a neurodegeneration characterized by a progressive cognitive deterioration; that, with time, causes significant decline on the performance of the person on activities of daily living together with several mood and behavioral changes.

In the United States several surveys reveal that an estimated 5.1 million subjects suffer AD. This number includes both, elderly subjects (subjects over 65 years old) and non-elderly subjects (people younger than 65 years old)[2] .

The NYU ADCC is one of 30 centers in the United States where the mission is to work toward better treatment options and care for dementia patients. The data provided here is part of a longitudinal study where the goal is to determine whether the mental decline rate is the same for subjects with complains of cognitive impairment and subjects that don't have any complains.

Two hundred and thirteen healthy persons evaluated from 1/1/1984 to 12/31/1997 with a follow up data obtained up until 12/31/2001 were utilized. The subjects, diagnosed as normal, were classified based on the Global Deterioration Scale (GDS) (Reisberg et al. [57]) as either GDS stage 1 (GDS1) or GDS stage 2 (GDS2)[11] at the onset of the study. For this classification all the subjects were evaluated on a variety of cognitive abilities, which included: verbal recall, associative recall, visual recognition, memory, immediate memory, language function, visuospatial praxis, and psychomotor speed[12]. This creates the two exclusive groups that we have been talking about, with 47 subjects evaluated as GDS1 and 166 subjects as GDS2 with seven covariates to be analyzed Age, Gender, Education, Mini Mental Score (MMS), Psychometric Deterioration Scale (PDS), Brief Cognitive Scale (BCR), Depression Scale (HDT) and Total Amount of Days the Patient Has Been in the Study (DAYSTOT) [3].

[11] A person can be diagnosed as normal even if it has complains of cognitive impairment.
[12] The GDS1 is a stage in which older persons are free of subjective or objective impairments. The GDS2 is a stage in which older persons have subjective cognitive impairment only [57].

An exploratory statistical analysis [1] found significant differences between important demographic characteristics of the two groups, for example, Age[13] and MMS[14] (Independent 2-sided t-test, $p = 0.0202$ and $p < 0.0001$ respectively). It was thought that gender would be also significantly different just by looking at the rates (55% of the 47 GDS1 subjects are female while 65% of the 166 GDS2 persons are) but such difference came up as not significant (Independent 2-sided t-test, $p = 0.2242$).

It is important to point out that during the follow up of the subjects the stages were each of them was classified changed. Many subjects ended up on higher GDS levels (i.e. they declined to higher stages of dementia or AD)[15], while others stayed within their original stages[16].

3.2 Independent vs. Paired Data Analysis

Several reasons have been given to either support or not a matched data analysis (in our particular case a 1:1) and we can keep adding reasons to this list to either go for it or not.

[13] Given the phenomenon we are analyzing, age is an important factor because the chances of a subject to suffer cognitive impairment are bigger with age.
[14] The Mini Mental Score (MMS) is the most commonly used score for complaints of memory problems. Is a series of questions and tests, each of which scores points if answered correctly. If every answer is correct, a maximum score of 30 points is possible. People with Alzheimer's disease generally score 26 points or less. It is important t say that this is not a test for Alzheimer's disease. There are many other reasons why a person might score less than 26 points [2].
[15] The GDS classification scale consists of 7 stages where GDS3 is MCI, and 4 and above are stages that describe the severity of Alzheimer on the patient [3]
[16] In the case of some of the GDS1 subjects they ended up with a GDS2

In the data we will be studying we have one big problem, which is unbalance, ie. one of the groups is considerably bigger than the other which might lead to biased results. Therefore, by matching we are expecting to balance the data and obtain better results, although, as we mention before, by doing this we can lose a considerable amount of information. In order to avoid this we will try to obtain a matched sample that can represent any possible matched sample from the original data set. In order to test this we will employ resampling methods.

The method to obtain a matched sample is widely known and consists basically in the following:

1. Separate the variables into two main categories:
    a. Factors that might be associated with the condition or disease of the patient (medical characteristics that might define the disease such as heart rate, blood pressure, etc.)
    b. Factor that are not associated but that might add variability to the condition (this are usually demographic characteristics such as age, gender, race, etc)

    The second group can be considered as confounding variables given that they can add variability to the data and therefore bias to the analysis if they are not handled with care.

2. Set the level of importance of these factors and match accordingly

In the current study the matching factors are Age, Education, Gender and MMS (in that order of importance), as we mentioned before, it was found that the first two variables

have significant differences between the two groups, besides, all these variables are known to be highly correlated to the decline of cognitive abilities in subjects [48, 58, 64], which makes them potential candidates for confounding factors, something that have to be taken into account given that can bring, also, a big amount of bias to the results

.

3.2.1 Results

Original Data

From the 213 subjects, 54% of the subjects with GDS2 declined while only 15% of GDS1 did. It was found that the declining percentage from GDS2 is significantly larger than the one from GDS1 (Fisher's exact test, $p<0.0001$).

In addition, the average decline time[20] for the 7 GDS1 subjects was 3212 days while the average decline time for the 90 GDS2 individuals was 1919 days. The time to decline to GDS ≥3 for the GDS2 group is significantly shorter than the one for GDS1 (Savage two-sample test for event time, $p=0.0007$).

However, as we mentioned before, these two GDS groups are significantly different with respect to age (mean difference = 3.424 years, std. err.=1.464, $p=0.020$) and Mini Mental

[20] This means the number of days between the start date of the subject in the study to the point where a GDS3 or higher was detected.

Score or MMS (mean difference=0.662 point, std. err.=0.184, p<0.0001). There also appeared to be a gender difference between the two groups (26 of the 47 GDS1, i.e. 55%, subjects are female while 108 of the 166 GDS2 subjects, i.e. 65%, are female) although such difference is not significant (p=0.147). Therefore the observed group difference in decline proportion and time to decline might be potentially confounded by the difference in Age and Mini Mental Score [58].

Paired Data

Twenty eight pairs of subjects (12 male, 16 female) were obtained by matching their Gender, Age and MMS scores between a cohort of 47 GDS1 and a cohort of 166 GDS2 subjects. There are no significant differences (GDS2 – GDS1) between groups for Age (mean difference=0.028 years, std. err.=1.907, p = 0.988) and MMS (perfect match; mean difference = 0, std. err. = 0.257, p = 1.000).

Of the 28 pairs, 15 subjects with GDS2 declined to GDS ≥3 and only 5 persons from GDS1 did. The percentage of individuals that declined from the GDS2 group is significantly larger than the one for GDS1 (McNemar's test, p=0.008).

In addition, the average decline time for the 5 GDS1 subjects was 3643 days while the average decline time for the 15 GDS2 individuals was 2059 days. The time to decline to

GDS ≥3 for the GDS2 group is significantly shorter than the one for GDS1 (Savage two-sample test for event time, p=0.020).

Comparison of the two data sets

So far is seen that the two data sets yield similar results on the difference between decline proportions and decline time for the two groups (GDS1 and GDS2). It can also be said that the two methods support the same theory that is that the GDS2 group has a larger decline proportion with faster decline time to GDS ≥3 than the GDS1 group.

Furthermore, the two data approaches where analyzed more meticulously to evaluate if one of them is a better approach and can give more accurate and better information about the phenomenon. For this the paired and the independent data were examined through logistic regression and conditional logistic regression.

The following logistic model was fitted to both data sets:

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 * Group + \beta_2 * Age + \beta_3 * MMS + \beta_4 * Gender + \beta_5 * Education$$

where p is the probability of declining to GDS ≥3.

There was found a significant group effect for both data sets. The estimated odds ratios for the two groups (GDS2 and GDS1) and their corresponding 95% confidence intervals were:

Original: 4.960 (odds ratio); (2.015, 12.208)

Paired: 7.563 (odds ratio); (1.734, 32.977)

Therefore, when controlling for Gender, Age, MMS and Education the subjects in the GDS2 group are more likely to decline than the individuals of the GDS1 group or, in other words, the odds for the GDS2 group to decline is larger than the odds for the GDS1.

The correlations' matrices for the two data sets are as follows:

Original Data

	Age	Gender	Educ	MMS	PDS	BCRtot	HDTtot	Daystot
Age	1.000	0.0413	-0.0336	-0.2120*	0.3327*	0.3010*	0.0750	-0.0433
Gender		1.0000	-0.0270	-0.1070	0.2020*	-0.0080	-0.0930	0.0100
Educ			1.0000	0.1180*	0.3784*	0.1110	0.0460	0.0667
MMS				1.0000	-0.3780*	-0.3329*	-0.1789*	0.0462
PDS					1.0000	0.2730*	0.0050	-0.1565*
BCRtot						1.0000	0.2000*	0.0224
HDTtot							1.0000	-0.1167
DAYStot								1.0000

Table 3. Pearson correlation coefficients from the correlation matrix for the Alzheimer's data where * means values that are significant for Ho: Rho=0

Paired Data

	Age	Gender	Educ	MMS	PDS	BCRtot	HDTtot	Daystot
Age	1.000	-0.0897	0.0783	-0.1672	0.1813	0.1996	0.0323	-0.0158
Gender		1.0000	-0.0398	-0.1092	0.2055	0.0396	-0.2895*	-0.0624
Educ			1.0000	0.0199	0.3654*	-0.0457	0.0551	0.1699
MMS				1.0000	-0.3236*	-0.1323	-0.1794	0.0918
PDS					1.0000	0.1475	0.1200	-0.2463
BCRtot						1.0000	0.2116	0.1402
HDTtot							1.0000	0.1099
DAYStot								1.0000

Tale 4. Pearson correlation coefficients from the correlation matrix for the Alzheimer's data
where * means values that are significant for Ho: Rho=0

Additional Predictors for Declining

It was shown that Group is a predictor of decline. Using the stepwise variable selection procedure we were able to obtain a model that can define the declining of the subjects

With a model including all the covariates as follows

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 * Group + \beta_2 * Age + \beta_3 * PDS + \beta_4 * BCRTOT + \beta_5 * DAYSTOT$$

where p is the probability of declining to GDS≥3.

The corresponding p-values, estimated odds ratios and their 95% confidence intervals as well as the goodness of fit are:

Original data

Age p= 0.0064; 1.065 (odds ratio); (1.018, 1.115)

PDS: p=<0.0001; 2.145 (odds ratio); (1.461, 3.151)

BCRTOT: p=<0.0001; 1.467 (odds ratio); (1.212, 1.777)

DAYSTOT: p=<0.0001; 1.001 (odds ratio); (1.001, 1.001)

Goodness-of-fit = 0.2033

Paired Data

Age p= 0.0425; 1.133 (odds ratio); (1.004, 1.278)

PDS: p= 0.7624; 1.152 (odds ratio); (0.461, 2.876)

BCRTOT: p= 0.0054; 1.931 (odds ratio); (1.214, 3.070)

DAYSTOT: p= 0.0178; 1.001 (odds ratio); (1.000, 1.002)

Goodness-of-fit = 0.3931

As we can see the models have the same tendency and are quite similar, we can also see that the goodness of fit is greater for the paired data and this might be due to the fact that the confounding variables are not so well controlled as with the paired data model.

Resample

It is critical to examine how representative the selected paired sample (IPD) is of all the possible paired samples. For this purpose, we calculated the mean, standard error and confidence intervals of the results obtained from the bootstrap resampling.

The number of resamples obtained was B=1000. It was found that, like the IPD, there are no significant differences (GDS2–GDS1) where the variables behaved as follows

AGE Original: GDS1= 65.480 GDS2= 65.507

Resample 95% CI: GDS1= (65.45,65.54) GDS2= (65.49,65.54)

EDUC Original: GDS1= 15.500 GDS2= 15.778

Resample 95% CI: GDS1= (15.21,15.79) GDS2= (15.52,16.11)

MMS Original: GDS1= 29.540 GDS2= 29.540

Resample 95% CI: GDS1=SAME ALL GDS2= SAME ALL

For the 1000 resamples we found that the average decline time for the GDS1 subjects was 3643 (in comparison to 3643 for the IPD) days while the average decline time for the GDS2 individuals was 2006 days (and it was 2059 days for the IPD). Therefore the same conclusion as with the IPD can be made, that means that the GDS2 group declines faster than the GDS1.

The estimators of the different regressions and the survival analysis were also checked in order to confirm if the IPD is unbiased. We found the following:

1- For the model

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 * Group + \beta_2 * Age + \beta_3 * MMS + \beta_4 * Gender + \beta_5 * Education$$

There was a significant group effect for the paired data set and the resamples, below we show the original estimator and the 95% confidence interval obtained from the resamples

IPD: Group estimate = 6.293

Resample: 95% CI for group estimate= 5.000, 8.166

For the final model we performed the same variable selection used on the IPD to each subsample and it was found that more than 50% had the exact same conclusions as the IPD for all the models.

3.3 Logistic Regression with Independent Samples

Alzheimer's disease is a type of dementia highly correlated to demographical characteristics, especially during the onset of the disease, and these factors also can affect highly the performance of the subjects during the psychometric tests that the subjects go through, therefore it is important to handle them with care given their potential to be confounding variables.

For the methods proposed for Residual Logistic Regression we have that the stage 1 model is as follows:

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 * Group + \beta_2 * Age + \beta_3 * MMS + \beta_4 * Gender + \beta_5 * Education$$

where p is the probability of declining to GDS ≥3.

Pearson Residual Analysis

After the stage 1 model ws fitted, the estimated probability of decline for each individual is calculated and with this the Pearson residual that we will rename as Y^*, the sample size of our sample is large enough (n=213 subjects), therefore we can assume that this residual is distributed normal with mean zero and standard deviation of one. This estimated value will be the new dependant variable to test the effect of the non-confounding variables as follows:

$$Y^* = \beta_0 + \beta_1 PDS + \beta_2 BCRTOT + \beta_3 HDTTOT + \beta_4 DAYSTOT + \varepsilon$$

Where, as we mentioned before, because the new dependent variable is not dichotomous any more, we apply a linear regression instead of the initial logistic regression.

After a stepwise selection the final model is as follows:

$$Y^* = \beta_0 + \beta_1 DAYSTOT + \varepsilon$$

with the corresponding p-value

DAYSTOT p=< 0.0001

Residual Logistic Regression Analysis

From the stage 1 of the method we obtain a new covariate for our final model, defined as:

$$X^* = \hat{\beta}_0 + \hat{\beta}_1 * Group + \hat{\beta}_2 * Age + \hat{\beta}_3 * MMS + \hat{\beta}_4 * Gender + \hat{\beta}_5 * Education$$

where the estimators are as follows:

$\hat{\beta}_0 = \quad 3.2468$
$\hat{\beta}_1 = \quad 1.6013$
$\hat{\beta}_2 = \quad 0.0655$
$\hat{\beta}_3 = \quad 0.0059$
$\hat{\beta}_4 = -0.1152$
$\hat{\beta}_5 = -0.3070$

This variable compiles all the information that the confounding variables provide to the dependant variable and it will be added as a fixed effect into a "complete" model where all the variables will be tested as follows:

$$\log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 * X^* + \hat{\beta}_1 * PDS + \hat{\beta}_2 * BCRTOT + \hat{\beta}_3 * HDTTOT + \hat{\beta}_4 * DAYSTOT$$

where $\beta_1 = 1$

After a stepwise selection the final model is:

$$\log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 * X^* + \hat{\beta}_1 * PDS + \hat{\beta}_2 * BCRTOT + \hat{\beta}_3 * DAYSTOT$$

with the corresponding p-values and odds ratio as follows:

PDS p= 0.0011; 1.904 (odds ratio); (1.293, 2.805)

BCRTOT p=< 0.0001; 1.473 (odds ratio); (1.214, 1.786)

DAYSTOT p=< 0.0001; 1.001 (odds ratio); (1.001, 1.001)

Hierarchical Logistic Regression Model

For this model the potential confounding variables where set as the different levels of hierarchy, setting Group (the GDS level) as the first level, Age as the second, Education as third, Gender as fourth and finally MMS as fifth. After running this analysis we didn't found any significant variables.

We ran a second model where the fifth level wasn't considered and found significant variables that stayed in the model after a stepwise selection and they are the following, with their corresponding p-values:

PDS p=0.0041

BCRTOT p= 0.0070

DAYSTOT p=0.0027

Chapter 4

Discussion and Conclusions

First it is important to mention that the resample data showed that the IPD is a good representation of all the potential paired samples that can be obtained from the sample. Given this the IPD can be considered as unbiased.

The power analysis showed that for most of the significant variables (the matching factors and the ones that were considered for the final model) the paired sample approach is a better approximation yielding higher power than its independent samples counterpart.

The consistency between the paired and independent samples analysis confirmed that the GDS1 and GDS2 groups decline at significantly different rates. From here we can conclude that people on GDS2 group decline much faster and in a higher proportion than people in the GDS1 group.

The analyses performed give us a general view of the behavior of Alzheimer's disease for these two GDS groups. We can also obtain a reliable conclusion thanks to the similarities among most of the models present, especialy the traditional logistic regression model and the residual logistic regression method, although the latter is more justifiable and flexible (in variable selection etc.). The Pearson residual analysis has a very significantly different final model than any of the other methods and this is due to the high levels of correlation based on the original data, which leads to conservative bias and less significant results. The hierarchical logistic regression has a different issue. If we go ahead and set all the levels of hierarchy necessary to cover all the potential confounding variables we end up with no significant variables. However if we eliminate the least important confounding variable (in this case MMS) then we obtain the exact same model that we obtained from the traditional logistic regression and the residual logistic regression analysis. The reason of this is because, as we set the hierarchy levels we are creating strata that become smaller and smaller within each hierarchy and by leaving MMS as one of the levels the strata became too small and therefore the analysis becomes less powerful.

Besides all the issues mentioned before, it is important to point out as well that both Pearson residual analysis and Hierarchical logistic regression do not provide an odds ratio of the covariates of interest which render them as weaker options unless, of course, that is not an important factor. However, we also point out that when the rare case of quasi-complete separation happens, the Pearson residual analysis is the only viable approach.

References

[1] Agresti, A. (2002). *Categorical Data Analysis*. Wiley, Second Edition.

[2] Alzheimer's Association. (2007). *Alzheimer Disease Facts and Figures 2007*. USA.

[3] Alzheimer's Society. (2002). *Quality Research in Dementia Information Sheet. The Mini Mental State Examination (MMSE) – a guide for people with dementia and their carers*. USA.

[4] Anderson, D.R. (2006). A Resample Method Called the Bootstrap. Colorado State University, USA.

[5] Ashe, F. (1986). An Essay at Measuring the Variance of Estimates of Outstanding Claim Payments. E. S. Knight & Co. Research Centre, Sydney, Australia.

[6] Austin, P.C., Tu, J.V. and Alter, D.A. (2003). Comparing Hierarchical Modeling with Traditional Logistic Regression Analysis Among Patients Hospitalized With Acute Myocardial Infraction: Should we Be Analyzing Cardiovascular Outcomes Data Differently?. Am Heart J; 145 (1): 27-35.

[7] Barker, J.N., Davies, S.M., De For, T., Ramsay, N.K.C., Weisdorf, D.J., and Wagner, J.E. (2001). Survival after transplantation of unrelated donor umbilical cord blood is comparable to that of human leukocyte antigen-matched unrelated donor bone marrow: results of a matched-pair analysis. Blood; 97(10): 2957-61.

[8] Barker, N. (2000). A Practical Introduction to the Bootstrap Using the SAS System. Oxford Pharmaceutical Sciences, Wallingford, UK.

[9] Brooks, R. and Stone, M. (1994). Joint Continuum Regression for Multiple Predictants. Journal of the American Statistical Association. Vol.89. No.428.

[10] Clayton, N. (1996). Development of Food Storing and the Hippocampus in Juvenile Marsh Tits (Parus Paluris). Behavioral Brain Research. 74(1/2):153-159

[11] Connet, J.E., Smith, J.A. and McHugh, R.B. (1987). Sample size and Power for a Pair-Matched Case-Control Studies. *Statistics in Medicine*. 6(10):53-59

[12] Conor, R.J. (1987). Sample Size for Testing Differences in Proportions for the Paired-Sample Design. *Biometrics*. 43(1):207-11.

[13] Cook, D. (1993). Exploring Partial Residual Plots. *Technometrics*; 35:351-362.

[14] Cook, D. and Cross-Dabrera, R. (1998). Partial Residual Plots in Generalized Linear Models. Journal of the American Statistical Association; 93:442442, 730-739.

[15] Cook, R.D., Hawkins, D.M. and Weisberg, S. (1992). Comparison of Model Misspecification Diagnostics Using Residual From Least Median of Squares Fits. Journal of the American Statistical Association; 87(418):419-424.

[16] Cook, R.D. and Weinsberg, S. (1982). Criticism and Influence Analysis in Regression. Sociological Methodology. 13:313-361.

[17] Cox, D.R. (1958). Two Further Applications f a Model for Binary Regression. Biometrika. 45(3/4):562-565.

[18] Dai, J., Li, Z. and Rocke, D. (n.y.). Hierarchical Logistic Regression Modeling With SAS GLIMMIX. University of California, Davis, CA, USA.

[19] Dallas, M.J. (2004). Testing Equality of Survival Functions Based on Both Paired and Unpaired Censored Data. Biometrics; 56(1):154-159.

[20] Darlington, R.B. and Smulders, T.V. (2001) Problems with Residual Analysis. Animal Behavior. 62:599-602.

[21] Derr, R.E. (n.y.). Performing Exact Logistic Regression with the SAS System. SAS Institute. Paper 254-25.

[22] Dickman, P.W. (2003). Logistic Regression in SAS version 8. Department of Medical Epidemiolgy and Biostatistics, Karolinska Institutet.

[23] Dixon, P.M. (2001). *Bootstrap Resampling*. Oxford University Press, Oxford.

[24] Duffy, S.W. (1984). Asymptotic and Exact Power for the McNemar Test and Its Analogue With R Controls Per Case. Biometrics. 40(4):1005-1015.

[25] Efron, B. and Trishirani, R. (1993). An Introduction to the Bootstrap. Chapman & Hall. USA.

[26] Efron, B. (1982). The jackknife, the bootstrap, ad other resampling plans. Society for Industrial Mathematics Press. Philadelphia, Pennsylvania, USA.

[27] Fisher, B.J. (1987). Guinness, Gosset, Fisher and Small Samples. Statistical Science. 2(1):45-52.

[28] Garcia-Fiñana, M., Cruz-Orive, L.M., Mackay, C.E., Pakkenber, B., and Roberts, N. (2003). Comparison of MR Imaging against Physical Sectioning to Estimate the Volume of Human Cerebral Compartments. Neuroimage. 18(2):505-16.

[29] Gelman, A. and Hill, J. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge Univ. Press.

[30] Good, P. (1993). Permutation Tests; A Practical Guide to Resampling Methods forTesting Hypothesis. Springer Series in Statistics, USA.

[31] Good, P.I. (2001). Resampling methods: a practical guide to data analysis. Birkhauser. USA.

[32] Gurmu, S. (1996). Testing for Fixed Effects in Logit and Probit Models Using an Artificial Regression. Econometric Theory. 12:872-874.

[33] Healy, S.D, Gwinner, E. and Kress, J.R. (1996). Hippocampal Volume in Migratory and Non-migratory Warblers: Effects of Age and Experience. Behavioral Brain Research. 81(1/2):61-68.

[34] Hill, G.J. (2007). Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, USA.

[35] Hosmer, D.W., Jovanovic, B. and Lemeshow, S. (1989). Best Subsets Logistic Regression. Biometrics. 45(4):1265-1270.

[36] Hosmer, D.W., and Lemeshow, S. (2000). Applied Logistic Regression. New York; Wiley. USA.

[37] Hosmer, D.W. and Lemeshow, S. (1992). Confidence Interval Estimation of Interaction. Epidemiology. 3(5):452-6.

[38] Jaffrezic, S., White, I.M.S. and Thompson, R. (2000). A Link Function Approach to Model Heterogeneity of Residual Variances Over Time in Lactation Curve Analysis. Journal of Dairy Science.83(5):1089-1093.

[39] Kleimbaum, D. and Klein, M. (2002). Logistic Regression, A self-learning text. Springer. Atlanta, GA, USA, Second Edition.

[40] Lachin, J.M. (2000). Biostatistical Methods. The Assessment of Relative Risks. Wiley Series in Probability and Statistics. USA.

[41] Laramore, G.E., and Spence, A.M. (1996). Boron Neutron CaptureTherapy (BNCT) for High-Grade Gliomas of the Brain: A Cautionary Note. Int. J. Radiation Oncology Biol. Phys. 36(1):241-246.

[42] Lavenex, P., Steele, M.A. and Jacobs, L.F. (2000). Sex Differences but not Seasonal Variation in the Hippocampus of Food Catching Squirrels: A Stereological Study. The Journal of Comparative Neurology. 425(2000):152-166.

[43] Lemeshow, S., Teres, D., Avrunim, J.S. and Pastides, H. (1988). Medical futility: Predicting the Outcome of Intensive Care Unit Patients by Nurses and Doctors – A Prospective Comparative Study. Crit Care Med. 31(2):456-61.

[44] Lutz, M.W., Kenakin, T.P., Cosi, M., Menius, J.A., Krishnamoorthy, C., Rimele, T. and Morgan, P.H. (1995). Use of Resampling Techniques to Estimate the Variance of Parameters in Pharmacological Assays When Experimental Protocols Preclude Independent Replication: An Example Using Schild Regressions. J. Pharmacol Toxicol Methods. 34(1):37-46.

[45] Maechler, M. and Buhlmann, P. (2006) Computational Statistics - Seminar for Statistics. ETH Zurich.

[46] Mallows, C.L. (1995). More Comment on CP. Technometrics. 37(4):362-372.

[47] Mantel, N. and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. J. Natl Center Inst. 22(4):719-48.

[48] Medical News. (2006). A high level of education means Alzheimer hits later, but harder and faster. Medical Research News. 16:14.

[49] Menard, S. (2001). Applied Logstic Regression Analysis. Second Edition. Sage University, USA.

[50] Mooney, C.Z. and Duval, R.D. (1993). Bootstraping: A Non-Parametric Approach to Statistical Inference. SAGE Publications. USA.

[51] O'Gorman, T.W. and Woolson, R.F. (1995). Using Kendall's taub Correlations to Improve Variable Selection Methods in Case-Control Studies. Biometrics. 51(4):1451-60.

[52] O'Neil, T.J. and Barry, S. C. (1995). Truncated Logistic Regression. Biometrics; 51(2):522-541.

[53] Parekh, N. (1999). Validity and Efficiency of Parameter Estimates in Frequency Matched Case-Control Studies. University of Toronto. Canada.

[54] Pregibon, P. (1981). Logistic Regression Diagnostics. The Annals of Statistics; 9(4)4:705-724.

[55] Pregion, P. (1982). Resistant Fits for Some Commonly Used Logistic Models With Some Medical Applications. Biometrics; 38(2):485-98.

[56] Pytynia, K., Grant, J.R., Etzel, C.J., Roberts, D.B., Wei, Q. and Strugis, E.M. (2004). Matched-Pair Analysis of Survival of Never Smokers and Ever Smokers With Squamous Cell Carcinoma of the Head and Neck. Journal of Clinical Oncology; 22(19):3981-3988.

[57] Reinsberg, B. and Franssen E.H. (1999). Clinical Stages of Alzheimer Disease. The Encyclopedia of Visual Medicine Series. An Atlas of Alzheimer Disease, Parthenon, Pearl River, NY.

[58] Reisberg, B., Prichep, L., Mosconi, L., John, E., Glodzkin-Sobanska, L., Boksay I., Monteiro, I. Torossian, C., Venvyas, A., Ashraf, N. (2008). The Pre-mild Cognitive Impairment, subjective cognitive impairment stage of Alzheimer's Disease. Alzheimer's Dementia. The Journal of the Alzheimer's Association; 4(1):S98-S108.

[59] Rieger, R.H., Kaplan, N.L. and Weinberg, C.R. (2001). Efficient Use of Sibling in Testing for Linkage and Association. Genetic Epidemiology; 20(1):175-191.

[60] Rosner, B. (1999). Fundamentals of Biostatistics. 4th Edition. Duxbury Press. USA.

[61] Sarle, W. (1991). What are Cross-Validation and Bootstrapping. SAS Institute.

[62] Sawyer, S. (2005) Resampling Data: Using a Statistical Jackknife. Washington University, USA.

[63] Sawyer, S. (2005). Resampling Data: Using Bootstraps. Washington University, USA.

[64] Simon, H., Cannistra, S.A., Godine, J.E., Huang, E., Heller, D., Shellito, P.C. and Stern, T.A. (2004). Alzheimer's Disease. Reuters Health. March. Nidus Information Services, Inc., New York, NY. USA.

[65] Stern, M.C., Conti, D.V., Siegmund, K.D., Corral, R., Yuan, J.M., Koh, W.P., Yu, M.C. (2007). DNA repair single-nucleotide polymorphisms in colorectal cancer and their role as modifiers of the effect of cigarette smoking and alcohol in the Singapore Chinese Health Study. Cancer Epidemiol Biomarkers Prev; 16(11):2363-72.

[66] Suissa, S. and Shuster, J.J. (1991). The 2X2 Matched-Pairs Trials: Exact Unconditional Design and Analysis. Biometrics; 47(2):361-72.

[67] Tamhane, A.C. and Dunlop, D.D. (1999). Statistics and Data Analysis. Prentice Hall, USA.

[68] Topp, R. and Gomez, G. (2004). Residual Analysis in Linear Regression Models with and Interval-Censored Covariate. Statistics in Medicine; 23(21):3377-3391.

[69] Upton, G. and Cook, I. (2004). Dictionary of Statistics. Oxford University Press, NY, USA.

[70] Van Der Voet, H and Mallows, B.C. (1997). CP and Predictions with Many Regressors: Comments on Mallow (1995). Technometrics; 39(1):115-116.

[71] Yamada, Y., Ackerman, I., Fransen, E., Mackenzie, R. and Thomas, G. (1999). Does the Dose Fractionation Schedule Influence Local Control of Adjuvant Radiotherapy for Early Stage Breast Cancer?. Int. J, Radiation Oncology Biol. Phys. 44(1):99-104.

[72] Ying So. (n.y.). A Tutorial on Logistic Regression. SAS Institute Inc.USA.

[73] Yu, A. (2003). Resampling Methods: Concepts, Applications and Justifications. PAREonline.net.

[74] Wacholder, S. and Weinberg, C.R. (1982). Paired vs. Two Sample for Clinical Trial of Treatment with Dichotomoud Outcomes. Power Considerationsi. Biometricsi; 38(3):801-12.

[75] Witte, J.S., Greenland, S. and Kim, L. (1998). Software for Hierarchical Modeling of Epidemiolgic Data. Epidemiology Resources; 9(5):563-565.

Appendix A

Residual Logistic Regression Variable Selection

The SAS code for the proposed model for our example is as follows:

Stage 1

```
PROC LOGISTIC DATA=AD DESCENDING;
MODEL Y = GROUP AGE MMS GENDER EDUC;
RUN;
```

After the estimators for the parameters of the potential confounding variables are calculated then we create a new variable as follows:

Stage 2

```
DATA AD_T;
SET AD;
```

T= $\hat{\beta}_{AGE} * AGE + \hat{\beta}_{MMS} * MMS + \hat{\beta}_{GENDER} * GENDER + \hat{\beta}_{EDUC} * EDUC$

RUN;

QUIT;

DATA AD_RLR;

SET AD_T;

OFFSETVAL=T;

RUN;

QUIT;

PROC LOGISTIC DATA=AD_RLR DESCENDING;

MODEL Y = PDS HDTTOT BCRTOT/ OFFSET=OFFSETVAL SELECTION=STEPWISE;

RUN;