Use of R in Genetic Epidemiology Designs
-Power/Sample Size Considerations

Jing Hua Zhao 1,2

1MRC Epidemiology Unit
2Institute of Metabolic Science
Addenbrooke’s Hospital
Cambridge CB2 0QQ
United Kingdom

http://www.mrc-epid.cam.ac.uk/~jinghua.zhao
E-mail: jinghua.zhao@mrc-epid.cam.ac.uk

July 21, 2010
Table of contents

1 Preliminaries
2 Study designs
   - Main results
   - Example - EPIC-Norfolk obesity project
   - Power of mediation
   - Further variations
3 Summary
4 Credits
5 Appendix
   - Power estimation based on proportion of variance explained
   - Gene-environment interaction (GEI)
Some terminology

- Genes, Chromosome, markers
- Alleles, genotypes, haplotypes
- Phenotypes, mode of inheritance, penetrance
- Mendelian laws of inheritance
- Hardy-Weinberg equilibrium
- linkage disequilibrium
- Gene-environment interaction
Genetic epidemiology

- It is the study of the role of genetic factors in determining health and disease in families and in populations, and the interplay of such genetic factors with environmental factors, or “a science which deals with the aetiology, distribution, and control of diseases in groups of relatives and with inherited causes of disease in populations” (http://en.wikipedia.org).

- It customarily includes study of familial aggregation, segregation, linkage and association. It is closely associated with the development of statistical methods for human genetics which deals with these four questions. The last two questions can only be answered if appropriate genetic markers available (Elston & Ann Spence. Stat Med 2006;25:3049-80).
Linkage studies

- It is the study of cosegregation between genetic markers and putative disease loci, and has been very successful in localizing rare, Mendelian disorders but since has difficulty for traits which do not strictly follow Mendelian mode of inheritance, considerable linkage heterogeneity and it has limited resolution.

- It typically involves parametric (model-based) and nonparametric (model-free) methods, the latter most commonly refers to allele-sharing methods.

- The underlying concepts are nevertheless very important. It can still be useful in providing candidates for fine-mapping and association studies.

- With availability of whole genome data, it is possible to infer relationship or correlation between any individuals in a population.
Association studies

- They focus on association between particular allele and trait; it is only feasible with availability of dense markers.

- It has traditionally applied to both relatives in families and population sample. For the latter there has been serious concern over spurious association due to difference in allele frequencies between hidden sub-populations in a sample.

- A range of considerations has been made (Balding. Nat Rev Genet 2006;7:781-91) but the availability of whole genome data again refresh views including statistical examination of population substructure.
Published genome-wide associations
Major study designs

Three common genetic association designs involving unrelated individuals (left), nuclear families with affected singletons (middle) and affected sib-pairs (right). Males and females are denoted by squares and circles with affected individuals filled with black colors and unaffected individuals being empty.

A conceptual picture based on a test of $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1 > \mu_0$ from a normal distribution
Sample size calculation based on normal distribution

Let $T \sim N(\mu_1, \sigma_1^2)$, we have the following steps,

1. $Z = \frac{T - \mu_0}{\sigma_0} \sim N\left(\frac{\mu_1 - \mu_0}{\sigma_0}, \frac{\sigma_2^2}{\sigma_0^2}\right)$.

2. $\beta = P(Z < Z_{1-\alpha} | \mu_1, \sigma_1^2) = \Phi \left( \frac{Z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma_0}}{\sigma_1} \right)$ and

$$Z_\beta = \frac{Z_{1-\alpha} \sigma_0 - (\mu_1 - \mu_0)}{\sigma_1}.$$

Since $Z_\beta = -Z_{1-\beta}$ and we are interested in $1 - \beta$,

$$Z_{1-\beta} = \frac{(\mu_1 - \mu_0) - Z_{1-\alpha} \sigma_0}{\sigma_1}, \quad |\mu_1 - \mu_0| = Z_{1-\alpha} \sigma_0 + Z_{1-\beta} \sigma_1.$$

3. As $\sigma_i \equiv \sigma_i / N$, $i = 1, 2$. $\sqrt{N}|\mu_1 - \mu_0| = Z_{1-\alpha} \sigma_0 + Z_{1-\beta} \sigma_1$.

$$N = \left( \frac{Z_{1-\alpha} \sigma_0 + Z_{1-\beta} \sigma_1}{\mu_1 - \mu_0} \right)^2$$
Sample size estimation for affected sib-pair linkage and association

The mean and variance for the designs considered above were given in Risch & Merikangas (1996) and described in Zhao J Stat Soft 2007; 23(8):1-8, both under multiplicative model.

Let $\gamma =$ genotypic risk ratio; $p =$ frequency of disease allele A; $Y =$ probability of allele sharing; $N_L =$ number of ASP families required for linkage; $P_A =$ probability of transmitting disease allele A; $H_1, H_2 =$ proportions of heterozygous parents; $N_{tdt} =$ number of family trios; $N_{asp}^* =$ number of ASP. families

The following tables were based on refined pbsize and fbsize functions in R/gap.
### Power of linkage versus association

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$p$</th>
<th>$Y$</th>
<th>$N_{asp}$</th>
<th>$P_A$</th>
<th>$H_1$</th>
<th>$N_{tdt}$</th>
<th>$H_2$</th>
<th>$N_{asp}^*$</th>
<th>$\lambda_o / \lambda_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>0.01</td>
<td>0.52</td>
<td>6402</td>
<td>0.80</td>
<td>0.05</td>
<td>1201</td>
<td>0.11</td>
<td>257</td>
<td>1.08/1.09</td>
</tr>
<tr>
<td>4.0</td>
<td>0.10</td>
<td>0.60</td>
<td>277</td>
<td>0.80</td>
<td>0.35</td>
<td>165</td>
<td>0.54</td>
<td>53</td>
<td>1.48/1.54</td>
</tr>
<tr>
<td>4.0</td>
<td>0.50</td>
<td>0.58</td>
<td>446</td>
<td>0.80</td>
<td>0.50</td>
<td>113</td>
<td>0.42</td>
<td>67</td>
<td>1.36/1.39</td>
</tr>
<tr>
<td>4.0</td>
<td>0.80</td>
<td>0.53</td>
<td>3024</td>
<td>0.80</td>
<td>0.24</td>
<td>244</td>
<td>0.16</td>
<td>177</td>
<td>1.12/1.13</td>
</tr>
<tr>
<td>2.0</td>
<td>0.01</td>
<td>0.50</td>
<td>445964</td>
<td>0.67</td>
<td>0.03</td>
<td>6371</td>
<td>0.04</td>
<td>2155</td>
<td>1.01/1.01</td>
</tr>
<tr>
<td>2.0</td>
<td>0.10</td>
<td>0.52</td>
<td>8087</td>
<td>0.67</td>
<td>0.25</td>
<td>761</td>
<td>0.32</td>
<td>290</td>
<td>1.07/1.08</td>
</tr>
<tr>
<td>2.0</td>
<td>0.50</td>
<td>0.53</td>
<td>3753</td>
<td>0.67</td>
<td>0.50</td>
<td>373</td>
<td>0.47</td>
<td>197</td>
<td>1.11/1.11</td>
</tr>
<tr>
<td>2.0</td>
<td>0.80</td>
<td>0.51</td>
<td>17909</td>
<td>0.67</td>
<td>0.27</td>
<td>701</td>
<td>0.22</td>
<td>431</td>
<td>1.05/1.05</td>
</tr>
<tr>
<td>1.5</td>
<td>0.01</td>
<td>0.50</td>
<td>6944779</td>
<td>0.60</td>
<td>0.02</td>
<td>21138</td>
<td>0.03</td>
<td>8508</td>
<td>1.00/1.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.10</td>
<td>0.51</td>
<td>101926</td>
<td>0.60</td>
<td>0.21</td>
<td>2427</td>
<td>0.25</td>
<td>1030</td>
<td>1.02/1.02</td>
</tr>
<tr>
<td>1.5</td>
<td>0.50</td>
<td>0.51</td>
<td>27048</td>
<td>0.60</td>
<td>0.50</td>
<td>1039</td>
<td>0.49</td>
<td>530</td>
<td>1.04/1.04</td>
</tr>
<tr>
<td>1.5</td>
<td>0.80</td>
<td>0.51</td>
<td>101926</td>
<td>0.60</td>
<td>0.29</td>
<td>1820</td>
<td>0.25</td>
<td>1030</td>
<td>1.02/1.02</td>
</tr>
</tbody>
</table>
Sample sizes required for association detection using population data with given prevalences

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$p$</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>0.01</td>
<td>46681</td>
<td>8959</td>
<td>4244</td>
<td>1887</td>
</tr>
<tr>
<td>4.0</td>
<td>0.10</td>
<td>8180</td>
<td>1570</td>
<td>744</td>
<td>331</td>
</tr>
<tr>
<td>4.0</td>
<td>0.50</td>
<td>10891</td>
<td>2091</td>
<td>991</td>
<td>441</td>
</tr>
<tr>
<td>4.0</td>
<td>0.80</td>
<td>31473</td>
<td>6041</td>
<td>2862</td>
<td>1272</td>
</tr>
<tr>
<td>2.0</td>
<td>0.01</td>
<td>403970</td>
<td>77530</td>
<td>36725</td>
<td>16323</td>
</tr>
<tr>
<td>2.0</td>
<td>0.10</td>
<td>52709</td>
<td>10116</td>
<td>4792</td>
<td>2130</td>
</tr>
<tr>
<td>2.0</td>
<td>0.50</td>
<td>35285</td>
<td>6772</td>
<td>3208</td>
<td>1426</td>
</tr>
<tr>
<td>2.0</td>
<td>0.80</td>
<td>79391</td>
<td>15237</td>
<td>7218</td>
<td>3208</td>
</tr>
<tr>
<td>1.5</td>
<td>0.01</td>
<td>1599920</td>
<td>307056</td>
<td>145448</td>
<td>64644</td>
</tr>
<tr>
<td>1.5</td>
<td>0.10</td>
<td>192105</td>
<td>36869</td>
<td>17465</td>
<td>7762</td>
</tr>
<tr>
<td>1.5</td>
<td>0.50</td>
<td>98013</td>
<td>18811</td>
<td>8911</td>
<td>3961</td>
</tr>
<tr>
<td>1.5</td>
<td>0.80</td>
<td>192105</td>
<td>36869</td>
<td>17465</td>
<td>7762</td>
</tr>
</tbody>
</table>
The initial design of the study was case-control (e.g., WTCCC with seven cases and controls) with 3425 cases and 3400 controls.

- It is potentially more powerful.
- Controls are selected, however.

It has therefore been changed into case-cohort design, in which cases are defined to be individuals whose BMI above 30 and controls are a random sample (sub-cohort) of the EPIC-Norfolk cohort which includes obese individuals.

- The sub-cohort is representative of the whole population and allows for a range of traits to be examined.
- There is more work to do with two-stage design.
- The problem of Mendelian randomisation can be considered in a general framework.
It started with assessment of how the power is compromised relative to the original case-control design.

This was followed by power/sample size calculation using methods established by Cai & Zeng (Biometrics 2004, 60:1015-1024) as implemented in the R/gap function ccsize, noting a number of assumptions.

More practically, we took the subcohort sample size to be 2,500, i.e., 10% of a total of 25,000 individuals as a rough representative sample.
The case-cohort sample of the EPIC-Norfolk obesity genetics project is a combination of the sub-cohort sample and case sample which is truncated from the whole cohort at BMI=30.
N=25,000, $\alpha = 5 \times 10^{-8}$ ($p_D$=prevalence, $p_1$=frequency, hr=hazard ratio)

<table>
<thead>
<tr>
<th>$p_D$</th>
<th>$p_1$</th>
<th>hr</th>
<th>ssize</th>
<th>$p_D$</th>
<th>$p_1$</th>
<th>hr</th>
<th>ssize</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>1.3</td>
<td>14391</td>
<td>0.2</td>
<td>0.2</td>
<td>1.4</td>
<td>3164</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>1.4</td>
<td>5732</td>
<td>0.2</td>
<td>0.3</td>
<td>1.3</td>
<td>4548</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>1.2</td>
<td>21529</td>
<td>0.2</td>
<td>0.3</td>
<td>1.4</td>
<td>2152</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>1.3</td>
<td>5099</td>
<td>0.2</td>
<td>0.4</td>
<td>1.2</td>
<td>20131</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>1.4</td>
<td>2613</td>
<td>0.2</td>
<td>0.4</td>
<td>1.3</td>
<td>3648</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3</td>
<td>1.2</td>
<td>11095</td>
<td>0.2</td>
<td>0.4</td>
<td>1.4</td>
<td>1805</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>3490</td>
<td>0.2</td>
<td>0.5</td>
<td>1.2</td>
<td>17120</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3</td>
<td>1.4</td>
<td>1882</td>
<td>0.2</td>
<td>0.5</td>
<td>1.3</td>
<td>3422</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
<td>1.2</td>
<td>8596</td>
<td>0.2</td>
<td>0.5</td>
<td>1.4</td>
<td>1713</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
<td>1.3</td>
<td>2934</td>
<td>0.1</td>
<td>0.2</td>
<td>1.4</td>
<td>8615</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
<td>1.4</td>
<td>1611</td>
<td>0.1</td>
<td>0.3</td>
<td>1.4</td>
<td>3776</td>
</tr>
<tr>
<td>0.3</td>
<td>0.5</td>
<td>1.2</td>
<td>7995</td>
<td>0.1</td>
<td>0.4</td>
<td>1.3</td>
<td>13479</td>
</tr>
<tr>
<td>0.3</td>
<td>0.5</td>
<td>1.3</td>
<td>2786</td>
<td>0.1</td>
<td>0.4</td>
<td>1.4</td>
<td>2824</td>
</tr>
<tr>
<td>0.3</td>
<td>0.5</td>
<td>1.4</td>
<td>1538</td>
<td>0.1</td>
<td>0.5</td>
<td>1.3</td>
<td>10837</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>1.4</td>
<td>9277</td>
<td>0.1</td>
<td>0.5</td>
<td>1.4</td>
<td>2606</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>1.3</td>
<td>7725</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Does it work? The LDL example (Sandhu et al. Lancet 2008, 371:483-91)

<table>
<thead>
<tr>
<th>Study 1 (EPIC-Norfolk subcohort) n=2269</th>
<th>Study 2 (EPIC-Norfolk obese set) n=1009</th>
<th>Study 3 (1958 British birth cohort) n=1375</th>
<th>Study 4 (CoLaus) n=5367</th>
<th>Study 5 (GEMS study) n=1665</th>
</tr>
</thead>
<tbody>
<tr>
<td>β coeff (SE)</td>
<td>p value</td>
<td>β coeff (SE)</td>
<td>p value</td>
<td>β coeff (SE)</td>
</tr>
<tr>
<td>rs4420638</td>
<td>0.24 (0.04)</td>
<td>0.14 (0.06)</td>
<td>0.02</td>
<td>0.25 (0.04)</td>
</tr>
<tr>
<td>rs599839</td>
<td>-0.15 (0.04)</td>
<td>-0.23 (0.06)</td>
<td>7.6×10⁻⁴</td>
<td>-0.14 (0.04)</td>
</tr>
<tr>
<td>rs4970834</td>
<td>-0.13 (0.04)</td>
<td>-0.18 (0.06)</td>
<td>5.5×10⁻³</td>
<td>-0.11 (0.04)</td>
</tr>
<tr>
<td>rs562338</td>
<td>-0.17 (0.04)</td>
<td>-0.11 (0.06)</td>
<td>0.07</td>
<td>-0.18 (0.05)</td>
</tr>
<tr>
<td>rs7575840</td>
<td>0.15 (0.03)</td>
<td>0.15 (0.05)</td>
<td>2.4×10⁻³</td>
<td>0.04 (0.04)</td>
</tr>
<tr>
<td>rs478442</td>
<td>-0.16 (0.04)</td>
<td>-0.07 (0.06)</td>
<td>0.25</td>
<td>-0.16 (0.04)</td>
</tr>
<tr>
<td>rs4591370</td>
<td>-0.17 (0.04)</td>
<td>-0.06 (0.06)</td>
<td>0.28</td>
<td>-0.16 (0.04)</td>
</tr>
<tr>
<td>rs4560142</td>
<td>-0.16 (0.04)</td>
<td>-0.06 (0.06)</td>
<td>0.27</td>
<td>-0.16 (0.04)</td>
</tr>
<tr>
<td>rs576203</td>
<td>-0.16 (0.04)</td>
<td>-0.07 (0.06)</td>
<td>0.25</td>
<td>-0.16 (0.04)</td>
</tr>
<tr>
<td>rs506585</td>
<td>-0.16 (0.04)</td>
<td>-0.06 (0.06)</td>
<td>0.31</td>
<td>-0.16 (0.04)</td>
</tr>
<tr>
<td>rs488507</td>
<td>-0.14 (0.04)</td>
<td>-0.07 (0.06)</td>
<td>0.25</td>
<td>-0.16 (0.04)</td>
</tr>
<tr>
<td>rs538928</td>
<td>-0.16 (0.04)</td>
<td>-0.01 (0.06)</td>
<td>0.92</td>
<td>-0.16 (0.04)</td>
</tr>
<tr>
<td>rs10402271</td>
<td>0.04 (0.03)</td>
<td>0.11 (0.05)</td>
<td>0.02</td>
<td>0.12 (0.04)</td>
</tr>
<tr>
<td>rs693</td>
<td>-0.12 (0.03)</td>
<td>-0.07 (0.05)</td>
<td>0.15</td>
<td>-0.06 (0.03)</td>
</tr>
</tbody>
</table>

*Table 3: Associations between Affymetrix SNPs with a combined p value of <1.0×10⁻⁹ and circulating concentrations of LDL cholesterol in independent study populations.*
In analogy to the *FTO*-T2D association mediated by BMI, metabolic traits have been considered (Freathy et al. Diabetes 2008, 57:1419-26) in a so-called Mendelian randomisation study of causal association.

There are *FTO*-BMI ($a$) and BMI-metabolic traits ($b$) associations, and BMI is a mediator between *FTO* and metabolic traits ($c = a \times b$).
Power based on ab

- We can of course perform simulations to obtain power estimate but it would be somewhat involved.
- We implement this in ab function in R/gap.
- We have for EPIC-Norfolk 25,000, SNP-BMI regression coefficient (SE) of 0.15 (0.01), and BMI-T2D log(1.19) (0.01). We consider $\alpha = 0.05$.
- Criticism arisen from this posthoc power calculation could be alleviated when we allow for a range of sample sizes to be considered in the next slide.
SNP-BMI-T2D in EPIC-Norfolk study

![Graph showing power of mediation](image_url)
How have we obtained the unexpected with T2D?

We have customarily used \( T2D = i_1 + cSNP + e_1 \), \( T2D = i_2 + c' SNP + bBMI + e_2 \), \( BMI = i_3 + aSNP + e_3 \); \( \hat{\alpha}b \) is called the “expected” and \( \hat{\alpha} - \hat{\alpha}' \) the “observed” which does not change even though the size of the mediated effect increases, i.e., \( \hat{\sigma}^2_{T2D} = \hat{c}^2 \hat{\sigma}^2_{SNP} + \frac{\pi^2}{3} \), \( \hat{\sigma}^2_{T2D} = \hat{c}'^2 \sigma^2_{SNP} + \hat{b}^2 \hat{\sigma}^2_{BMI} + 2\hat{c}' \hat{b} \hat{\sigma}_{SNPBMI} + \frac{\pi^2}{3} \).

It is recommended that \( \hat{\alpha}_{corrected} = \hat{\alpha} \sqrt{1 + \frac{\hat{b}^2 \hat{\sigma}^2_{SNPBMI}}{\pi^2/3}} \) be used instead.

We replace \( \frac{\pi^2}{3} \) with 1 for probit regression.
One may need to consider a range of outcomes such as binary, continuous, count, time-to-event and longitudinal data.

A unified framework was discussed by Vittinghoff E, et al. (2009) Stat Med 28:541-57. The model takes into account binary or continuous primary and mediation factors by both analytic and approximation methods. Nevertheless, this leads to a large number of combinations and functions.

A single function masize was created in R/gap to simplify this.
Two-stage design on main effect

- The goal is to reduce cost without compromising efficiency. Given our study sample and SNPs of interest are defined, a staged design furnishes collection of all information in several steps.

- In the simplest and well-studied two-staged design of genetic case-controls studies, a proportion of individuals is genotyped at all of the SNPs and a proportion of the most significant ones is selected and to be carried over as replication study at the second stage. Skol et al. Nat Genet 2006; 38(2):209-13 (check the associate website for a program called CaTS).

- It was implemented in the function tscc within R/gap.
Some complications-two-stage GEI

- A case-only design is used as the first stage.
- This is to be followed by a second stage involving both cases and controls.
- Some recent references are given here:
  - Kass PH, Gold EB. in Ahrens W, Pigeot I (Eds) Handbook of Epidemiology 2005; 1.7
  - Li D, Conti DV. Am J Epidemiol 2008; 169:497-504
  - Schaid DJ, Sinnwell JP. Hum Genet 2010; 127:659-68
A flaw in the case-cohort framework for extremely large cohort?

The right panels show when the cohort of 6.5 million, the power/sample size is unstable such that a change of p1 from 0.04 to 0.03 led to sample size increase from 3968 to 211,480!
Moreover

- The formula is only appropriate for the case of dominant model, and it would be much preferable to consider the most widely used additive model.

- We will need probability weighting to allow for general genetic models to be specified. This has not been established but would be analogous to the framework as implemented in the computer program Quanto which is widely used.
Other packages

- **powerGWASinteraction** (Kooperberg & LeBlanc 2008). It calculates power for SNP × SNP and SNP × environment interactions in genome-wide association studies. It assumes a two-stage analysis, where only SNPs that are significant at a marginal significance level $\alpha_1$ are investigated for interactions, and a binary environmental covariate.

- **trex** (Schaid & Sinnwell 2010). It implements truncated exact test for two-stage case-control design for studying rare genetic variants. It consists of a screening stage focusing on rare variants in cases by which number of case-carriers of any rare variants exceeds a user-specified threshold will have additional cases and controls genotyped and analysed for all cases and controls in the second stage.
Power calculation is an integrated part of in designing epidemiological studies and closely linked with the statistical analysis to be carried out.

There are certain basic principles to follow whenever new problem comes along. We can then implement in R.

Customised software, however, would greatly facilitate the calculation, e.g.,

- R - for the comparison between association/linkage including the Mendelian randomisation example.
- SAS - for the regression example.
- Quanto - for GEI here.
The statistical understanding is on the evolving, e.g., author of Quanto is currently conducting a survey, which should lead to amendment to our calculation for gene-environment interaction here.

The mult-stage approach includes case-only coupled with case-control samples and current focus on rare variants.

There appears to have problems with case-cohort sampling from extremely large cohort.

The case of Mendelian randomisation should be considered in a broader framework, e.g., survival outcome.
References for Genetic Epidemiology

References for Mendelian Randomisation

Acknowledgements

The results presented here/hereafter were based on work recently done at MRC.

- Functions pbsize and fbsize were originally written in C at IoP and refined before and after the JSS paper.
- Function ccsize for case-cohort design attributes to EPIC-Norfolk obesity study, and function tscc for stage design was implemented as a result of an internal journal club.
- The Mendelian randomisation example was used in preparation for MRC QQR.
- $R^2$ regression methods attributes to ELSA-DNAR applications from UCL (http://www.natcen.ac.uk/elsa/).
- Case-control results were obtained for InterAct (http://www.inter-act.eu/).
Power estimation based on proportion of variance explained

<table>
<thead>
<tr>
<th>Sample size</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>0.10</td>
<td>0.52</td>
<td>0.86</td>
<td>0.97</td>
<td>1</td>
</tr>
<tr>
<td>15,000</td>
<td>0.29</td>
<td>0.86</td>
<td>0.99</td>
<td>1.00</td>
<td>1</td>
</tr>
<tr>
<td>20,000</td>
<td>0.52</td>
<td>0.97</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
</tr>
<tr>
<td>25,000</td>
<td>0.72</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ a = 10^{-5} \]

\[ a = 5 \times 10^{-7} \]
SAS code

```
proc power;
ods output output=op;
multreg model = fixed
  alpha = 0.00001 0.000001 0.0000005
  nfullpred = 1
  ntestpred = 1
  rsqfull = 0.001 to 0.005 by 0.001
  rsqdiff = 0.001 to 0.005 by 0.001
  ntotal = 10000 to 25000 by 1000 power = .;
run;
```

Note that ods can suppress/save all outputs to databases.
Let $D=$ disease, $E=$ exposure, and $g=$ genotype at a candidate locus with susceptibility allele $A$ and normal allele $a$. The population prevalence of $A$ and exposure will be denoted by $q_A$ and $p_E$. Under HWE, the genotypes $AA$, $Aa$, $aa$ have frequencies $P(g|q_A) = q_A^2$, $2q_A(1-q_A)$, $(1-q_A)^2$. We assume particular model e.g., $G(g) = 0, 1, 2$ and relate disease to genetic and environmental covariates through logistic and log-linear models

$$P(D = 1|G, E) = \frac{e^{\alpha + \beta_g G + \beta_e E + \beta_{ge} GE}}{1 + e^{\alpha + \beta_g G + \beta_e E + \beta_{ge} GE}}$$

and

$$P(D = 1|G, E) = e^{\alpha + \beta_g G + \beta_e E + \beta_{ge} GE}$$

so that the baseline probabilities of disease in the population is given by $e^\alpha/(1 + e^\alpha)$ and $e^\alpha$ whereas $e_g$, $e_e$, $e_{ge}$ are the genetic, environmental and interactive relative risks.
Power calculation under matched design

- We can use conditional logistic regression model

\[
L(\beta_g, \beta_e, \beta_{ge}) = \prod_{i=1}^{N} \frac{e^{\beta_g G_{ij} + \beta_e E_{ij} + \beta_{ge} G_{ij} E_{ij}}}{\sum_{j \in M(i)} e^{\beta_g G_{ij} + \beta_e E_{ij} + \beta_{ge} G_{ij} E_{ij}}}
\]

where \(M(i)\) includes all subjects in matched set \(i\).

- Power/sample size calculation can proceed with contrasting \(l^1 = \ln[L(\beta_g, \beta_e, \beta_{ge})]\), \(l^0 = \ln[L(\beta_g, \beta_e)]\) with

\[
\Lambda = 2(\hat{l}^1 - \hat{l}^0)
\]

and \(N\Lambda\) being the non-centrality parameter of chi-squared distribution under the alternative hypothesis.

- The required sample size is obtained via equating the noncentrality parameter to theoretical values under a given significant level and power (the previous conceptual picture still applies).
Legends in the project manual were perhaps confusing so it is worthwhile to re-present here.

- Matched case-control study
- Type I error rate ($\alpha$) = 0.00001 (two-sided)
- Continuous environmental factors with standard deviation 1, and interaction odds ratio ($R_{ge}$) = 1.2 - 4
- $K = 0.05$ (0.1 - 0.15)
- Sample size ($N$) = 500 - 12,000
- Additive model
- Allele frequency ($p$) = 0.05, 0.1, 0.2, 0.3

We supplied these to Quanto 1.0 (http://hydra.usc.edu/gxe, now available on the Epidemiology Unit machines) Gauderman WJ. Stat Med 21:35-50, 2002
Scenario 2

![Graph showing power for allele frequency](image)
Scenario 3

Power for allele frequency = 0.2

Rge
- 1.1
- 1.2
- 1.3
- 1.4
- 1.5
- 1.6
- 1.7
- 1.8
- 1.9
- 2

Power (y-axis) vs Sample size (x-axis)
Scenario 4