Program `ge_trend_v2`

Roger Logan and Donna Spiegelman

January 7, 2002

Abstract

The program `ge_trend_v2` is designed to calculate the power and minimum required sample size for case-control studies testing hypotheses about gene-environment interactions with a polytomous exposure variable. This program extends the original program `ge_trend` by permitting the investigator the freedom to allow the main effect odds ratio for gene and exposure to vary in a user-specified interval under the alternative hypothesis.

Keywords: gene-environment interaction, power, sample size calculations, polytomous exposure variable

Contents

1 Description

Assume D and G are binary indicators of disease status and the genotype. As in Foppa & Spiegelman, (1997), let E denote ordinal exposure levels $0, 1, \ldots, q-1$, for some $q \geq 2$, where $q-1$ is the highest exposure level. We assume the log odds ratio of disease to be a linear function of the ordinal scores for exposure. That is, the log odds ratio between adjacent exposure levels is assumed to be constant for those with the genotype, $OR(E|G = 1)$ and for those without the genotype $OR(E|G = 0)$. Denote the top-to-bottom quantile exposure effect in those without the genotype as OR^{tb}(E|G=0)

$= \text{OR(E|G=0)}^{q-1}$. Let θ denote the gene-environment interaction effect, where $\theta = \dfrac{OR(E|G=1)}{OR(E|G=0)}$, and let the top-to-bottom odds ratio for the gene-environment interaction to be $\theta^{tb} = \theta^{q-1}$. These definitions imply the logistic model

$$\text{logit}\left[\text{P(D} = 1|\text{E, G})\right] = \beta_0 + \beta_{\text{g}}\text{G} + \beta_{\text{e}}\text{E} + \beta_{\text{eg}}\text{EG},$$

where $\beta_g = \log\left[OR(G|E=0)\right]$, $\beta_e = \log\left[OR(E|G=0)\right]$, and $\beta_{eg} = \log(\theta)$.

The program `ge_trend_v2` can be used to find the minimum sample size, N, in a case-control study for which the hypothesis $H_0 : \beta_{eg} = 0$ versus the test of $H_A : \beta_{eg} \neq 0$ has power ψ and size α, using the power function $\psi(N, \beta_{eg|H_A})$, where

$$\psi(N, \beta_{eg|H_A}) = 1 - \Phi\left(\frac{z_{\alpha/2}\sigma_{GE|H_0} - \beta_{EG|H_A}}{\sigma_{GE|H_A}}\right).$$

In the expression for $\psi(N, \beta_{eg|H_A})$, Φ is the standard cumulative normal distribution, $z_{\alpha/2}$ is the value of a standard normal variable, Z, such that $P(Z \geq z) = 1 - \alpha/2$, α is the Type 1 error rate of the test, and $\sigma_{GE|H_0}$ and $\sigma_{GE|H_A}$ are the asymptotic standard errors of the estimates of β_{eg} under H_0 and H_A, respectively. The values of β_e and β_g are input by the user under the null hypothesis by inputing the values of $OR^{tb}(E|G=0)$ and $OR(G|E=0)$. Under the alternative hypothesis, these values are allowed to vary in an interval specified by the user. By setting the upper and lower bounds for the intervals to be the same value as provided under the null hypothesis, the user can obtain the results for the original `ge_trend` program.

Under the null hypothesis the values of $OR^{tb}(E|G=0)$ and $OR(G|E=0)$ are approximately equal to the marginal odds ratios $OR^{tb}(E)$ and $OR(G)$. To aid the user in determining suitable intervals for $OR^{tb}(E|G=0)$ and $OR(G|E=0)$ under the alternative hypothesis, the program calculates the values of $OR^{tb}(E|G=0)$ and $OR(G|E=0)$ under the assumption that the marginal odds ratios assumed under the null hypothesis are now those that would be obtained when the alternative hypothesis is true, under the assumption of independence of the distributions of G and E in the study base. The user can then use these new conditional odds ratios to determine the location of the required intervals. Because the assumption of independence between G and E is likely to be at least moderately violated, the user is prermitted to give the assumed values of the two odds ratios over a range, rather than a fixed point.

In addition, `ge_trend_v2` can be used for determining the power of testing the above hypotheses for a specified sample size. In this case, the program calculates the maximum and minimum power when $OR^{tb}(E|G=0)$ and $OR(G|E=0)$ are allowed to vary in specified intervals.

For more information concerning the methods used by this program, we refer the user to the Spiegelman and Logan (2001), which can be obtained at
`http://www.hsph.harvard.edu/faculty/spiegelman/manuscripts/manu.html` .

The program assumes that the size of the test is fixed at a level of $\alpha = 0.05$.

2 Invocation

To use the program `ge_trend_v2` for calculating sample sizes, the user must provide the following information.

1. The desired power.
2. The distribution of gene prevalence and exposure in the population from which the cases will arise. Under the common assumption that the distribution of genotype and exposure are independent in the population from which the cases will arise, when the exposure is a quantile, then only then number of quantiles and the marginal prevalence, $P(G = 1)$, of the genotype is needed.
3. The control-case ratio (must be greater than 0).
4. The assumed odds ratio for genotype in the lowest exposure level $OR(G|E = 0)$ (must be greater than 0).
5. The top-to-bottom quantile odds ratio for exposure $OR^{tb}(E|G = 0)$ (must be greater than 0).
6. The top-to-bottom quantile ratio of odds ratios for gene-environment interaction, ROR, under the alternative hypothesis (must be greater than 0).
7. An interval for $OR^{tb}(E|G = 0)$ assumed under H_A (lower bound must be greater than 0).
8. An interval for $OR(G|E = 0)$ assumed under H_A (lower bound must be greater than 0).

When the user wants to calculate the power for a test at a given sample size, the user must provide the number of cases in addition to items 2-8 needed for calculating sample sizes.

The program is then invoked by executing the following command at the Unix prompt:

```
ge_trend_v2
```

3 Examples

In this section we present three examples of the use of the program `ge_trend_v2`. The first two are sample size calculations for two possible exposure level distributions and the third example is a power calculation when the exposure levels are assumed to be quantiles. In each example we will make the following assumptions.

- Desired power in samplesize calculations is 0.8.
- The marginal prevalence of the genotype, P(G=1) = 0.5.
- Number of exposure levels is 5 .
- Control-case ratio = 1.
- $OR^{tb}(E|G = 0) = 1.5$ under null hypothesis .
- $OR(G|E = 0) = 1.5$ under null hypothesis.
- ROR = 1.5 under alternative hypothesis.

1. In this example, we assume that we are using quantiles for the exposure levels.

```
rosella:strol 43% ge_trend_v2

 Do you want power or sample size calculations?
 (type "p" for power or "s" for sample size) s
Enter the power to be achieved ..................:   .8
Are you using exposure quantiles (y/n)?..........:   y
Enter the number of quantiles (k: integer).......:   5
Enter the gene prevalence in controls ...........:   .5
Enter the control-case ratio (c) ................:   1
Enter OR^tb(E|G=0) under the null................:   1.5
Enter OR(G|E=0) under the null...................:   1.5
Enter ROR^tb under the alternative...............:   1.5



 Assuming that the alternative hypothesis is true, that Pr(E,G)=Pr(E)Pr(G),
 and these values of OR(E) and OR(G) for the marginal odds ratios,
 the following conditional odds ratios under the
 alternative are implied. You can use  these values
 as  an indicator of the center of the range of possible values
 for OR^tb(E|G=0) and OR(G|E=0).

Marginal odds ratio     Conditional odds ratio

 OR^tb(E)  OR(G)        OR^tb(E|G=0) OR(G|E=0)

 1.50      1.50              1.20      1.20
Enter range for OR^tb(E|G=0)  under the alternative  : 1.0 1.5
Enter range for OR(G|E=0)  under the alternative  : 1.0 1.5

    Size  Min power  OR(G|E=0)  OR^tb(E|G=0)  Max power  OR(G|E=0)  OR^tb(E|G=0)
-------------------------------------------------------------------------------
    6386    0.8000     1.5000     1.5000     0.8075     1.0000     1.0000

The values of OR(G|E=0) and OR^tb(E|G=0)
are the values under the alternative that produced the
minimum and maximum power over the interval given as input.

Do you wish another calculation (y/n): y
```

2. In this example we assume that the distribution of the exposure levels is given by P(E=0)=0.4, P(E=1)=0.3, P(E=2)= 0.1, P(E=3)=0.1, and P(E=4)=0.1.

```
 Do you want power or sample size calculations?
 (type "p" for power or "s" for sample size) s
Enter the power to be achieved ..................:   .8
Are you using exposure quantiles (y/n)?..........:    n
Enter the number of exposure categories..........:    5
Enter the prevalence of exp. categories in controls,
 hitting "return" after each value
.4
```

```
.3
.1
.1
.1
Enter the gene prevalence in controls ...........:   .5
Enter the control-case ratio (c) ................:   1
Enter OR^tb(E|G=0) under the null................:   1.5
Enter OR(G|E=0) under the null...................:   1.5
Enter ROR^tb under the alternative...............:   1.5


 Assuming that the alternative hypothesis is true, that Pr(E,G)=Pr(E)Pr(G),
 and these values of OR(E) and OR(G) for the marginal odds ratios,
 the following conditional odds ratios under the
 alternative are implied. You can use  these values
 as  an indicator of the center of the range of possible values
 for OR^tb(E|G=0) and OR(G|E=0).

Marginal odds ratio     Conditional odds ratio

 OR^tb(E)  OR(G)        OR^tb(E|G=0) OR(G|E=0)

 1.50      1.50              1.19      1.30
Enter range for OR^tb(E|G=0)  under the alternative  : 0.75 1.4
Enter range for OR(G|E=0)  under the alternative  : 1 1.5

    Size  Min power  OR(G|E=0)  OR^tb(E|G=0)  Max power  OR(G|E=0)  OR^tb(E|G=0)
-------------------------------------------------------------------------------
    6922    0.8000     1.5000     0.7500       0.8097     1.0000     1.0000

The values of OR(G|E=0) and OR^tb(E|G=0)
are the values under the alternative that produced the
minimum and maximum power over the interval given as input.
```

3. In this example we find the power of the test that there is a gene-environment interaction when the sample size is 1000 with 500 cases.

```
Do you want power or sample size calculations?
 (type "p" for power or "s" for sample size)
p
 Are you using exposure quantiles (y/n)?.........:    y
 Enter the number of quantiles (k: integer).....:    5
 Enter the gene prevalence in controls ..........:   .5
 Enter the control-case ratio (c) ...............:    1
 Enter the number of cases ......................:    500
 Enter OR^tb(E|G=0) under the null...............:  1.5
 Enter OR(G|E=0) under the null..................:  1.5
 Enter ROR(E|G) under the alternative............:  1.5
```

```
 Assuming that the alternative hypothesis is true, that Pr(E,G)=Pr(E)Pr(G),
 and these values of OR(E) and OR(G) for the marginal odds ratios,
 the following conditional odds ratios under the
 alternative are implied. You can use  these values
 as  an indicator of the center of the range of possible values
 for OR^tb(E|G=0) and OR(G|E=0).

Marginal odds ratio     Conditional odds ratio

 OR^tb(E)  OR(G)        OR^tb(E|G=0) OR(G|E=0)

 1.50      1.50              1.20      1.20
 Enter range for OR(E|G=0)  under the alternative  : 1 2
 Enter range for OR(G|E=0)  under the alternative  : 1 2

    Size  Min power  OR(G|E=0)  OR^tb(E|G=0)  Max power  OR(G|E=0)  OR^tb(E|G=0)
-------------------------------------------------------------------------------
    1000    0.1969     1.0000     1.0000     0.2142     2.0000     2.0000

The values of OR(G|E=0) and OR^tb(E|G=0)
are the values under the alternative that produced the
 minimum and maximum power over the interval given as input.
```

4 Warnings

If the required inputs are not correctly input, the program will inform the user and prompt the user to use the correct input. For example, if the distribution of the exposure levels does not add to one, the program will respond as follows.

```
Enter the number of exposure categories:4
Enter the prevalence of exp. categories in controls,
 hitting "return" after each value
.3
.3
.2
.1
 Probabilities do not add to 1.0.   0.90000000000000
Enter the prevalence of exp. categories in controls,
 hitting "return" after each value
```

If incorrect bounds for $OR^{tb}(E|G=0)$ are used the program responds as follows.

```
Enter range for OR^tb(E|G=0)  under the alternative  : 2 1
 Lower bound should be less or equal to the upper bound
 Try again
```

```
Enter range for OR^tb(E|G=0)  under the alternative  : -1 2
 Lower bound should be positive.
 Try again
Enter range for OR^tb(E|G=0)  under the alternative  :
```

In the calculation of the expected cell counts when performing the sample size estimates it is possible for some cells to contain less than 5 subjects. In this case the resulting values should be questioned since exact analysis rather than asymptotic methods should probably be used. The program will include the following output along with the main results:

```
Suspect values :  some cell counts when d=1 were less than 5.
```

For further explanation, see Foppa and Spiegelman.

5 References

Foppa I, Spiegelman D. Power and sample size calculations for case–control studies of gene–environment interactions with a polytomous exposure variable. *Am J Epidemiol* 1997;146:596-604.

Spiegelman D., Logan R. Power and sample size for case-control studies of gene-environment interactions: a new method with comparizon to the old. Submitted 2001. Can be obtained at `http://www.hsph.harvard.edu/faculty/spiegelman/manuscripts/manu.html`

The FORTRAN source code can be obtained at
`http://www.hsph.harvard.edu/faculty/spiegelman/ge_trend_v2.html`

6 Credits

Written by Roger Logan, Ph.D., Harvard School of Public Health, Boston MA.

Questions can be directed to Roger Logan, `strol@channing.harvard.edu`, or Donna Spiegelman `stdls@channing.harvard.edu`.