Conditional vs. marginal estimators of within-pair regression effects in individually-matched case-control studies and twin cohorts

Lyle C Gurrin
Centre for Epidemiology and Biostatistics
Melbourne School of Population and Global Health

ViCBiostat, Melbourne, November 2014

Acknowledgements No. 1

Jack and Jill...

John Carlin
Jonathan Sterne
John Hopper

and

Gillian Dite

Acknowledgements No. 2

Collaborators on the more recent work on binary data...

Martin Hazelton
Fizz Williamson
Sabria Khan

Outline

- Background and motivation
- Regression models for continuously-valued paired exposure and outcome data
- History
- Conditional estimators
- Binary data (both exposure and outcome)
- Estimators of within-pair effect
- Simulation results
- Extensions

Paired data

- Twins provide naturally matched pairs for studies of human health, although paired data goes beyond just twins.
- We can can exploit within-pair comparisons of data to avoid confounding associations between outcomes and exposures by shared factors.
- Specific assumptions about shared factors allow the determination of genetic and environmental contributions to disease risk.

Twin Studies - 1

- Tradition of focussing on genetic hypotheses
- Decompose variation in a quantitative trait
- Compare within-pair correlation of DZ with MZ ($\frac{1}{2}$ under the additive genetic model)
- *Classical Twin Model* assumes that variation attributable common or shared environment is the same for DZ and MZ twins
- Lower DZ than MZ within-pair correlation provides evidence that a trait is determined by genetic factors

Twin Studies - 2

- Can the twin context provide greater insight on associations?
  - Cardiovascular risk (blood pressure) with birthweight
  - Cancer risk (breast density) with physical measures (height, weight, BMI)
- Ideally like to separate the effect of shared and individual factors (*eg* maternal versus placental)
- When can a regression relationship be said to have a genetic basis?

Individual twin regression

Exposure variable x_{ij} and binary outcome y_{ij} for $i = 1, \ldots, n$ and $j = 1, 2$. A cross-sectional or individual-level regression model might propose that

$$\begin{aligned} E(y_{i1}) &= \alpha + \beta x_{i1} \\ E(y_{i2}) &= \alpha + \beta x_{i2} \end{aligned}$$

Individual twin regression

$$\begin{aligned} E(y_{i1}) &= \alpha + \beta x_{i1} \\ E(y_{i2}) &= \alpha + \beta x_{i2} \end{aligned}$$

If we take the difference between the two equations we get

$$E(y_{i1} - y_{i2}) = \beta(x_{i1} - x_{i2})$$

If we take the average between the two equations we get

$$E(\overline{y}_i) = \alpha + \beta\overline{x}_i$$

Between- and within-pair regression

These are special cases of a model general model

$$
\begin{aligned}
E(y_{i1}) &= \beta_0 + \beta_w(x_{i1} - \overline{x}_i) + \beta_b \overline{x}_i \\
E(y_{i2}) &= \beta_0 + \beta_w(x_{i2} - \overline{x}_i) + \beta_b \overline{x}_i
\end{aligned}
$$

where $\overline{x}_i = (x_{i1} + x_{i2})/2$. Since

$$
\begin{aligned}
x_{i1} - \overline{x}_i &= (x_{i1} - x_{i2})/2 \\
x_{i2} - \overline{x}_i &= (x_{i2} - x_{i1})/2 \\
&= -(x_{i1} - x_{i2})/2
\end{aligned}
$$

we can re-write the multivariable between- and within-pair model as...

Between- and within-pair regression

These are special cases of a model general model

$$
\begin{aligned}
E(y_{i1}) &= \beta_0 + \beta_w(x_{i1} - x_{i2}) + \beta_b \overline{x}_i \\
E(y_{i2}) &= \beta_0 + \beta_w(x_{i2} - x_{i1}) + \beta_b \overline{x}_i
\end{aligned}
$$

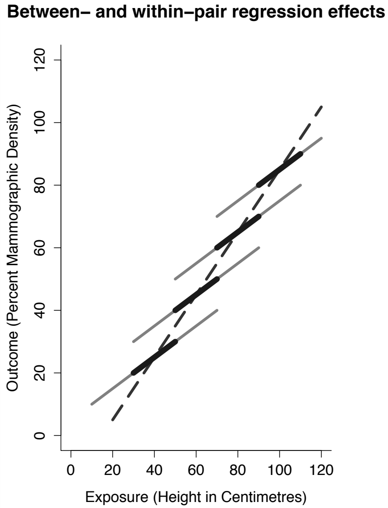
Univariate regressions of the within-pair differences and within-pair means yield estimates of β_w and β_b respectively.

Simultaneous estimation of β_w and β_b from the multivariable model generates the same estimates for OLS and GLS (but standard errors will differ).

Interpretation - 1

- β_w is the expected change in the the outcome y for a unit change in the *deviation* of the exposure x from the pair mean, holding this pair mean constant.
- β_b is the expected change in the outcome y for a unit change in the *pair mean* $\overline{x}$, holding the within-pair deviation (difference) constant.

Illustration of between- and within-pair effects



Interpretation - 2

What we are postulating is

- A model for the expected value of the outcome y can be improved by using data from pairs.
- A good way to do this is to relate the expected value of y_{i1} not just to x_{i1} (the twin's own exposure value) but also to their co-twins exposure value x_{i2}.
- The expected difference in outcome y comparing between two x values may depend on whether we are comparing (i) co-twins with each other within-pair; or (ii) unrelated twins between pairs.

Twin – Co-Twin regression

The multivariable model re-expressed

$$
\begin{aligned}
E(y_{i1}) &= \beta_0 + \beta_t x_{i1} + \beta_c x_{i2} \\
E(y_{i2}) &= \beta_0 + \beta_t x_{i2} + \beta_c x_{i1}
\end{aligned}
$$

where

$$
\begin{aligned}
\beta_t &= (\beta_w + \beta_b)/2 \\
\beta_c &= (\beta_w - \beta_b)/2
\end{aligned}
$$

from which we can see that $\beta_c = 0$ ($E(y_{i1})$ does not depend on x_{i2} and *vice versa*) is equivalent to $\beta_w = \beta_b$ (between- & within-pair reg. effects are the same).

Individual twin regression

Recall the individual-level regression model

$$E(y_{i1}) = \alpha + \beta x_{i1}$$
$$E(y_{i2}) = \alpha + \beta x_{i2}$$

If the multivariable between- and within-pair regression model is correct, then fitting the individual-level regression (again, by either OLS or GLS) produces and estimate of β that is a weighted average of the corresponding estimates of β_w and β_b with weights that depend on ρ_x and ρ_y, the observed within-pair correlation of x and y respectively.

Weighted average estimates of β

This results first came to my attention through **biostatistics** via the seminal paper by Neuhaus & Kalbfleisch (1998) in *Biometrics*.

Neuhaus & Kalbfleisch (1998), however, quote Scott & Holt (1982) in *J. Amer. Stat. Assoc.*, a paper on two-stage **sample surveys**.

Scott & Holt (1982) in turn trace the result back to Maddala (1971) in *Econometrica*, so we're now in **economics** where the interest at the time was "pooling cross section and time series data".

Weighted average estimates of β

There's more: Maddala (1971) references this

Wallace TD & Hussain A (1969). The use of error components models in combining cross section with time series data. *Econometrica*, **37**, 55–72,

which in turn refers to this

Hildreth C (1950). Combining Cross-Section Data and Time Series. Cowles Commission Discussion Paper: Statistics No. 347, May 15, 1950.

Weighted average estimates of β

This led me to propose **Gurrin's Law**: One can always find a reference to the between- and within-cluster "beta is weighted average" result published before one was born *regardless* of how old one is!

Papers by John Neuhaus et al.

- **Neuhaus JM & Jewell N** (1990). The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*, **46**, 977 – 990.
- **Neuhaus JM & Kalbfleisch JD** (1998). Between- and within-cluster covariate effect in the analysis of clustered data. *Biometrics*, **54**, 638 – 645.
- **Neuhaus JM & McCulloch CE** (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J. R. Statist. Soc. B.*, **68**, 859 – 872.

Papers on interpretation

- **Begg MD & Parides MK** (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, **22**, 2591 – 2602.
- **Carlin JB, Gurrin LC, Sterne JAC, Morley R & Dwyer T** (2005). Regression models for twin studies: a critical review. *Int. J. Epidem.*, **34**, 1089 – 1099.
- **Dwyer T, Blizzard CL**. (2005). A discussion of some statistical methods for separating within-pair associations among all twins in research on fetal origins of disease. *Paediatric and Perinatal Epidemiology*, **19**(1), 48 – 53.
- **Gurrin LC, Carlin JB, Sterne JAC, Dite GS & Hopper JL** (2006). Using bivariate models to understand between- and within-cluster regression coefficients with application to twin data. *Biometrics*, **62**, 745 – 751.

The Sjolander Show

- **Sjolander A, Lichtenstein P, Larsson H & Pawitan Y**. (2012). Between-within models for survival analysis. *Statistics in Medicine*.
- **Sjolander A, Frisell T & Oberg S**. (2012). Causal interpretation of between-within models for twin research. *Epidemiologic Methods*, **1**(1), No. 10.
- **Sjolander A, Johansson ALV, Lundholm C, Altman D, Almqvist C & Pawitan Y**. (2012). Analysis of 1:1 matched cohort studies and twin studies, with binary exposures and binary outcomes. *Statistical Science*, **27**(3), 395-411.

Binary x and y

- We now consider the scenario where both x_{ij} and y_{ij} are binary 0/1 variables.
- $x_{i1} - \overline{x}_i = (x_{i1} - x_{i2})/2 = -(x_{i2} - x_{i1})/2$ can take only three values: $-\frac{1}{2}$, 0 or $\frac{1}{2}$.
- $\overline{x}_i$ can take only three values: 0, $\frac{1}{2}$ or 1.
- A pair is *exposure-concordant* if $x_{i1} = x_{i2}$, so $\overline{x}_i = 0$ or $\overline{x}_i = 1$, otherwise it is *exposure-discordant* where $\overline{x}_i = \frac{1}{2}$.
- A pair is *outcome-concordant* if $y_{i1} = y_{i2}$, otherwise it is *outcome-discordant*.

Individual twin regression

Exposure variable x_{ij} and binary outcome y_{ij} with expectation p_{ij} for pair i and individual j. An ordinary logistic regression model implies:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta x_{ij}$$

So for each pair we have:

$$\log\left(\frac{p_{i1}}{1-p_{i1}}\right) = \beta_0 + \beta x_{i1}$$
$$\log\left(\frac{p_{i2}}{1-p_{i2}}\right) = \beta_0 + \beta x_{i2}$$

Individual twin regression

Within-pair difference in the two regression equations:

$$\log\left(\frac{p_{i1}}{1-p_{i1}}\right) - \log\left(\frac{p_{i2}}{1-p_{i2}}\right) \;=\; \beta(x_{i1} - x_{i2})$$

Average the two regression equations:

$$\frac{1}{2}\left[\log\left(\frac{p_{i1}}{1-p_{i1}}\right) + \log\left(\frac{p_{i2}}{1-p_{i2}}\right)\right] \;=\; \beta_0 + \beta\overline{x}_i$$

- Both equations depend on the exposure x through the regression coefficient β.
- We can, however, generalise to allow the exposure effect on the within-pair log odds ratio and the between-pair average log odds to be distinct.

Between- and within-pair regression

- The model proposed by Neuhaus & Kalbfleisch (1998) where $p_{ij} = E(y_{ij})$ is

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_w(x_{ij} - \bar{x}_i) + \beta_b\bar{x}_i$$

for $j = 1, 2$ and $\bar{x}_i = (x_{i1} + x_{i2})/2$.

- We can also write

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_w\tfrac{1}{2}(x_{ij} - x_{ik}) + \beta_b\bar{x}_i$$

where $k = (3 - j)$, indicting that this model has terms for both between- and within-pair regression effects.

Paired data in matched studies

- Binary outcome data from individual matched pairs is typically analysed for associations with exposures using conditional logistic regression (CLR).
- CLR uses the likelihood *conditional* on the sum of the pair's outcome (0, $\frac{1}{2}$ or 1) to estimate the regression parameter β.

Paired data in matched studies

- Pairs that are either outcome-concordant or exposure-concordant do not contribute to the conditional likelihood.
- So in an individually matched case-control study, where $y_{i1} \neq y_{i2}$ by design, only exposure-discordant pairs contribute to the estimation of β, an inherently within-pair regression parameter.

Paired data in cohort studies

- But what about outcome-concordant twin- or sib-pairs appearing in cohort studies?
- Can their inclusion improve the precision of estimation of regression parameters while still maintaining the benefits of a paired analysis?

Paired data in cohort studies

- We know their inclusion does not influence the results of CLR.
- But what about the multivariable between- and within-pair regression model?
- What is the relationship between estimates of β from CLR and estimates of β_w from ordinary logistic regression with the multivariable model?

Estimation of β_b and β_w

- Use ordinary logistic regression (OLR) on *all* pairs to estimate β_0, β_w and β_b, and take β_w as our association parameter. The OLR estimate of β_w depends on the assumptions for β_0 & β_b (*ie* whether we estimate them or set them to zero).
- Use conditional logistic regression (CLR) on pairs that are both exposure-discordant *and* outcome-discordant (so "doubly-discordant") pairs to estimate β_w. The CLR model does not have parameters β_0 and β_b.

There is a "close empirical correspondence" (Neuhaus & Kalbfleisch (1998), Ten Have *et al.* (1995), Sjolander *et al.* (2012)) between the OLR and CLR estimates...
...but they are **not** formally identical.

Table : Summary of paired binary data for one pair member exposed and one pair member unexposed

	Unexposed member ($x_2 = 0$) Event ($y_2 = 1$)	No event ($y_2 = 0$)	Totals
Exposed member ($x_1 = 1$)			
Event ($y_1 = 1$)	n_{11}	n_{10}	$n_{11} + n_{10}$
No event ($y_1 = 0$)	n_{01}	n_{00}	$n_{01} + n_{00}$
Totals	$n_{11} + n_{01}$	$n_{10} + n_{00}$	$\sum\sum n_{ij}$

CLR estimator of β_w

For a single binary exposure the CLR estimate of β_w is the log ratio of number of exposure-outcome concordant pairs (n_{10}) to exposure-outcome discordant pairs (n_{01}) among **outcome-discordant** pairs:

$$\hat{\beta}_w(\text{CLR}) = \log(n_{10}/n_{01})$$

with standard error

$$\text{s.e.}(\hat{\beta}_w(\text{CLR})) = \sqrt{n_{10}^{-1} + n_{01}^{-1}}.$$

OLR estimators of β_w: $\beta_b = \beta_0 = 0$

For OLR there is an addition contribution from the $n_{11} + n_{00}$ exposure-discordant pairs that are **outcome-concordant** pairs. For $\beta_b = \beta_0 = 0$ we have

$$\hat{\beta}_w(\text{OLR}, \beta_b = \beta_0 = 0) \;=\; 2\log\left(\frac{n_{10} + \frac{1}{2}(n_{11} + n_{00})}{n_{01} + \frac{1}{2}(n_{11} + n_{00})}\right)$$

with s.e.$(\hat{\beta}_w(\text{OLR}), \beta_b = \beta_0 = 0) =$

$$\sqrt{\left(n_{10} + \frac{1}{2}(n_{11} + n_{00})\right)^{-1} + \left(n_{01} + \frac{1}{2}(n_{11} + n_{00})\right)^{-1}}.$$

OLR estimators of β_w: $\beta_b \neq 0, \beta_0 \neq 0$

For $\overline{x}_i$ *categorical*, β_b and β_0 replaced with their MLE's we have

$$\hat{\beta}_w(\text{OLR}) = \log\left(\frac{n_{10}+n_{11}}{n_{01}+n_{11}} \times \frac{n_{10}+n_{00}}{n_{01}+n_{00}}\right)$$

with s.e.$(\hat{\beta}_w(\text{OLR}))^2 =$

$$(n_{10}+n_{11})^{-1}+(n_{01}+n_{11})^{-1}+(n_{10}+n_{00})^{-1}+(n_{01}+n_{00})^{-1}$$

where

$$n = n_{00}+n_{01}+n_{10}+n_{11}.$$

OLR estimators of β_w: General Form

More generally we have $\hat{\beta}_w =$

$$\log\left[\frac{{n_{10}}^{1-\alpha_1/2}+\alpha_2(\alpha_3 n_{11}+(1-\alpha_3)n_{00})}{{n_{01}}^{1-\alpha_1/2}+\alpha_2(\alpha_3 n_{11}+(1-\alpha_3)n_{00})}\right]$$

$$\times\left[\frac{{n_{10}}^{1-\alpha_1/2}+\alpha_2((1-\alpha_3)n_{11}+\alpha_3 n_{00})}{{n_{01}}^{1-\alpha_1/2}+\alpha_2((1-\alpha_3)n_{11}+\alpha_3 n_{00})}\right]$$

where $(\alpha_1, \alpha_2, \alpha_3) =$
$(1, 0, 0)$ for CLR;
$(0, \frac{1}{2}, \frac{1}{2})$ for OLR with $\beta_b = \beta_0 = 0$; and
$(0, 1, 0)$ for OLR with $\beta_b = \hat{\beta}_b$ and $\beta_0 = \hat{\beta}_0$.

Unanswered questions

- The result when $\beta_b \neq 0$ and $\beta_0 \neq 0$ is from Sjolander *et al.* (*Stat. Sci.*, 2012). They state "...the decomposition into within- and between-effects is a legitimate method for binary exposures, which was questioned by Carlin *et al.* (2005)".
- **BUT** this result only applies when we model the mean effect $m(\overline{x}_i)$ as

$$\beta_0 \mathbf{I}(\overline{x}_i = 0) + \beta_{0.5}\mathbf{I}(\overline{x}_i = 0.5) + \beta_1 \mathbf{I}(\overline{x}_i = 1)$$

that is, the mean appears in the regression equation as a categorical variable.

Unanswered questions

- In this case only exposure-concordant pairs contribute to estimating β_w, so it's not much of a "decomposition into within- and between-effects".
- Explicit results for $\overline{x}_i$ as a continuously valued exposure (albeit with only three possible values) are not available, but one can show the precision of $\hat{\beta}_w$ is (slightly) greater than for the above categorical model.

Simulation: $\beta_0 = \beta_b = 0$, $\beta_w = 1$

Table : Summary of simulation results for 100 datasets with 500 twin pairs. True values are $\beta_0 = 0$, $\beta_b = 0$, $\beta_w = 1$. $z \sim \mathrm{N}(0, \sigma^2)$ with between- and within-pair std dev $\sigma = 2$. The observed binary covariate is $x = I(z > 0)$.

	Mean Estimate	Empirical Std Err	Model Std Err
CLR	0.98	0.25	0.28
OLR ($\beta_b = \beta_0 = 0$)	0.96	0.24	0.26
OLR ($\beta_b = \hat{\beta}_b$, $\beta_0 = \hat{\beta}_0$)	0.97	0.24	0.26

Simulation: $\beta_0 = 0$, $\beta_b = \beta_w = 1$

Table : Summary of simulation results for 100 datasets with 500 twin pairs. True values are $\beta_0 = 0$, $\beta_b = 1$, $\beta_w = 1$. $z \sim \mathrm{N}(0, \sigma^2)$ with between- and within-pair std dev $\sigma = 2$. The observed binary covariate is $x = I(z > 0)$.

	Mean Estimate	Empirical Std Err	Model Std Err
CLR	1.02	0.28	0.29
OLR ($\beta_b = \beta_0 = 0$)	0.92	0.24	0.26
OLR ($\beta_b = \hat{\beta}_b$, $\beta_0 = \hat{\beta}_0$)	0.99	0.26	0.27

Conclusions

- CLR and OLR estimators of the within-pair regression effect are specific examples of a general estimator that assigns weights to count data from outcome-concordant exposure-discordant pairs.
- OLR estimators are potentially more efficient than CLR estimators since they use data from all exposure-concordant pairs, rather than just those that are outcome-discordant, an assertion that is born out by our simulation studies.

Extension 1: Two time-points

Let y_{ijk} (x_{ijk}) be the outcome (exposure) at time $k = 1, 2$ for the j^{th} twin in the i^{th} twin pair.

Then we can decompose x_{ijk} as using a two-way analysis of variance

$$\begin{aligned} x_{ijk} &= 0.5(x_{ijk} - \overline{x}_{ij.}) + 0.5(x_{ijk} - \overline{x}_{i.k}) \\ &\quad + 0.5(\overline{x}_{ij.} - \overline{x}_{i..}) + 0.5(\overline{x}_{i.k} - \overline{x}_{i..}) \\ &\quad + \overline{x}_{i..} \end{aligned}$$

Extension 2: Count data

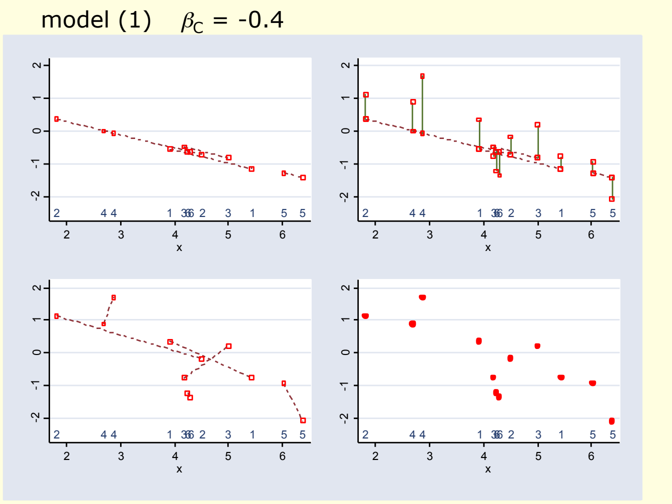
Between- and within-pair effects for **Poisson regression of count outcomes**. The unconditional estimate (OPR?) of β_w is the difference in the log of the mean count between exposed and unexposed.

The conditional estimate (CPR?) of β_w is (approximately) the log of the mean difference in counts d between exposed and unexposed (if $d > 0$). More formally

$$\exp(\hat{\beta}_w) = d/2 + \sqrt{(d/2)^2 + 1}$$

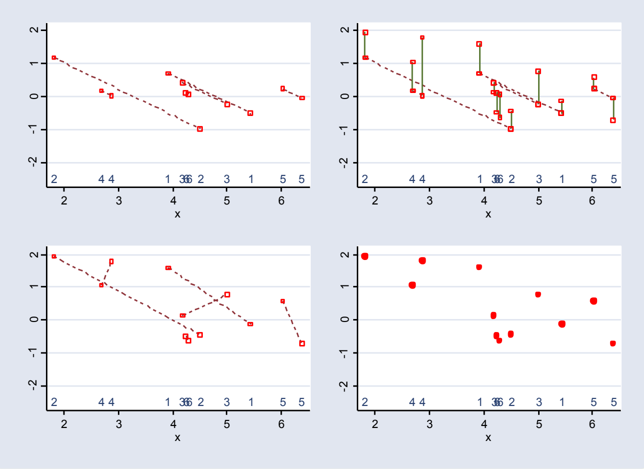
so $\exp(\hat{\beta}_w) > 0$ even if $d < 0$.

Graph 1

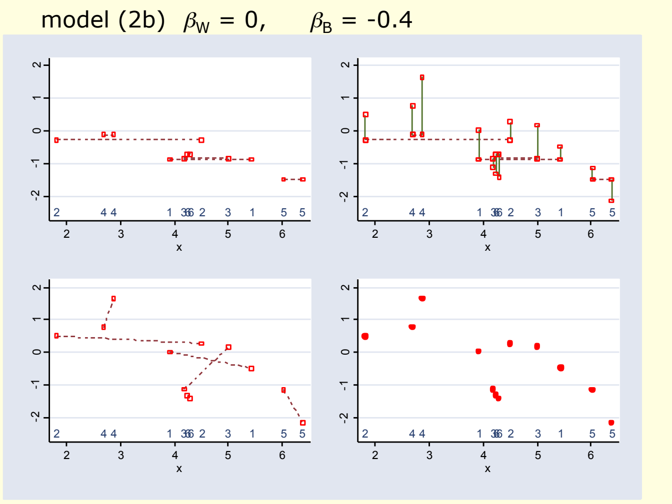


Graph 2

model (2a) β_W = -0.8, β_B = 0

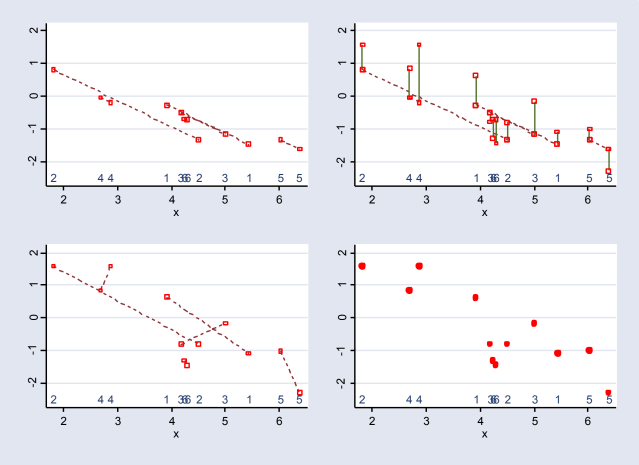


Graph 3



Graph 4

model (2c) β_W = -0.8, β_B = -0.4



References

- **Neuhaus JM & Jewell N** (1990). The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*, **46**, 977 – 990.
- **Ten Have TR, Landis JR & Weaver SL** (1995). Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Stat. Med.*, **14**, 413 – 429.
- **Neuhaus JM & Kalbfleisch JD** (1998). Between- and within-cluster covariate effect in the analysis of clustered data. *Biometrics*, **54**, 638 – 645.
- **Begg MD & Parides MK** (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, **22**, 2591 – 2602.
- **Sjolander A, Johansson ALV, Lundholm C, Altman D, Almqvist C & Pawitan Y** (2012). Analysis of 1:1 matched cohort studies and twin studies, with binary exposures and binary outcomes. *Stat. Sci.*, **27**, 395–411.

References

- **Carlin JB, Gurrin LC, Sterne JAC, Morley R & Dwyer T** (2005). Regression models for twin studies: a critical review. *Int. J. Epidem.*, **34**, 1089 – 1099.
- **Gurrin LC, Carlin JB, Sterne JAC, Dite GS & Hopper JL** (2006). Using bivariate models to understand between- and within-cluster regression coefficients with application to twin data. *Biometrics*, **62**, 745 – 751.
- **Stone J, Gurrin LC, . . ., Hopper JL & Byrnes GB** (2010). Sibship analysis of associations between SNP haplotypes and a continuous trait with application to mammographic density. *Genetic Epidemiology*, **34**, 309 – 318.
- **Jamsen KM, Zaloumis SG, Scurrah KJ & Gurrin LC** (2012). Specification of generalized linear mixed models for family data using Markov Chain Monte Carlo methods. *Journal of Biometrics and Biostatistics*. S1-003.
- **Zaloumis SG, Scurrah KJ, Ellis J, Harrap S & Gurrin LC** (2014). Non-proportional odds multivariate logistic regression of ordinal family data. *Biometrical Journal*, in press.