Score Tests for Association Under Response-dependent Sampling Designs for Expensive Covariates

ANDRIY DERKACH

Department of Statistical Sciences,
University of Toronto, Toronto, ON, M5S 3G3, Canada
E-mail: derkach@utstat.toronto.edu

JERALD F. LAWLESS

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: jlawless@uwaterloo.ca

LEI SUN

Department of Statistical Sciences,
University of Toronto, Toronto, ON, M5S 3G3, Canada
E-mail: sun@utstat.toronto.edu

Summary

Response-dependent sampling is widely used in settings where certain variables are expensive to obtain. Estimation has been thoroughly investigated but recent applications have emphasized tests of association for expensive covariates and a response variable. We consider testing and provide easily implemented likelihood score tests for generalized linear models under a broad range of sampling plans. We show that when there are no additional covariates, the score statistics are identical for conditional and full likelihood approaches, and are of the same form as for ordinary random sampling. Applications in genetics are discussed briefly.

Keywords: Generalized linear model, Genetic association study, Incomplete data, Maximum likelihood estimation, Tests of independence, Two-phase sample

This is a pre-copyedited, author-produced PDF of an article accepted for publication in *Biometrika* following peer review. The version of record "Derkach, A., Lawless, J.F. and Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, 102 (4), 988–994" is available online at: DOI: 10.1093/biomet/asv038.

1 Introduction

Tests of a null hypothesis H_0 of no association between a response variable Y and a $p \times 1$ vector of covariates X in a population of independent units are of interest in many settings. With either simple random samples of size n or sampling conditional on X, test statistics are often based on the vector $S = \sum_{i=1}^{n}(Y_i - \bar{Y})X_i$, which arises as a likelihood score statistic in several linear or generalized linear models, including Gaussian, binary logistic and Poisson log-linear models. Recently, response-dependent sampling designs have become of much interest in contexts where the covariates X are expensive to measure (Lawless et al., 1999, Scott and Wild, 2011). An important application in genetics is in connection with tests for association between genetic variants X and a specific quantitative trait Y (Barnett et al., 2013, Lin et al., 2013, Lee et al., 2014). Two-phase studies are frequently used in such cases (Chatterjee et al., 2003, Breslow et al., 2009). In particular, phase 1 data on Y and additional covariates Z are available for all individuals in a cohort or population, and then in phase 2 information on X is obtained for a subset of individuals whose probability of selection may depend on their (Y, Z) values.

When there are no additional covariates Z, the statistic S can be used to test the null hypothesis H_0 under Y-dependent phase 2 sampling, and a permutation distribution can be used to obtain p-values. However, a model relating Y and X is often proposed, and then model-based likelihood or estimating function methods can be used (Lawless et al., 1999, Scott and Wild, 2011). Moreover, a model is needed if estimation of the effect of X is also of interest or if it is necessary to adjust for Z. Although there is a large literature on estimation under Y-dependent two-phase sampling, testing has received little attention. The purpose of this note is to address testing. We show that score tests based on a semiparametric likelihood take a simple form for a wide range of generalized linear models and sampling plans; this occurs because the test statistics are specified, and their variances estimated, under H_0. We also show that conditional and full likelihood approaches produce the same score statistic when there are no additional covariates. This explains previously reported simulation results where tests based on conditional and full likelihoods have similar power. The full likelihood approach can be applied for many sampling plans for which the conditional likelihood approach is unavailable, such as sampling designs based on ranks or on residuals from fitted models of Y given Z. Previous articles have not provided universally valid variance estimates for statistics based on full likelihood and we give them here.

2 Y-Dependent Sampling

We assume that the distribution of Y, given observed covariate vectors X and Z, has probability density or mass function of the form

$$f(y \mid x, z; \theta) = f_0\{y \mid \mu(x, z); \theta\}, \tag{1}$$

where f_0 is a known function, $\theta = (\beta_0, \beta', \gamma', \sigma)'$, $\mu(x, z) = \beta_0 + \beta' x + \gamma' z$ and σ may be a scalar or vector containing scale and shape parameters. This form covers location-scale models, exponential family generalized linear models, proportional hazards models and other families of distributions. The probability density or mass function for (X, Z) is written as $g(x, z)$. Following the general framework of (Lawless et al., 1999), we assume that a cohort or population of units $(Y_i, X_i, Z_i)(\ i = 1, ..., N)$ is generated from $f(y \mid x, z; \theta)g(x, z)$, and that Y_i and Z_i are observed for all units. Covariate X_i, however, is ascertained only for certain individuals and we define the binary indicator $R_i = I(X_i \text{ is ascertained})$. The selection of units for measurement of X can depend on the observed Y and Z; thus, the X_i are missing at random (Rubin, 1976): $\text{pr}(R \mid Y, Z, X) = \text{pr}(R \mid Y, Z)$. Studies that employ specified selection probabilities are termed two-phase studies (Chatterjee et al., 2003, Breslow et al., 2009); phase

1 refers to the collection of data on (Y, Z) and phase 2 to the collection of data on X. We focus on this setting here, but the methodology extends to situations where X values are missing at random for certain units, and missingness probabilities have to be estimated.

Two types of sampling that use specified selection probabilities are seen in many applications (Lawless et al., 1999). Both are based on partitioning the range of (Y, Z) into strata $S_1, \ldots, S_K$:

(i) Basic stratified sampling, in which the known (Y, Z) for each of the N cohort units are assigned to their appropriate strata; then simple random samples of specified sizes n_j ($j = 1, ..., K$) are taken from the units in each stratum j, and their X values are obtained. The stratum sizes N_j are known and the sampling fractions are therefore $p_j = n_j/N_j$.

(ii) variable probability sampling, in which units from the cohort or population are considered as their (Y, Z) values are generated. Then, if a unit's (Y, Z) lies in stratum j, it is selected with specified probability p_j. In some contexts variable probability sampling is termed preferential sampling (Diggle et al., 2010). In this case the total phase 2 sample size and the number of units selected from each stratum are random variables.

Usually basic stratified sampling is used when a cohort of N units exists at the time when sampling begins; variable probability sampling is used in settings where units are generated over time. These sampling designs can be extended to allow selection probabilities $\pi(y, z) = \text{pr}(R = 1 \mid Y = y, Z = z)$ that are arbitrary functions of y and z. Forms of quota sampling (McCullagh, 2008) can also be handled (Lawless et al. (1999), Section 2). It is also possible to base the strata on residuals r from the fit of a regression model for Y given Z; the missing X remain missing at random in this case. A third type of sampling used in some areas is what we term rank-based sampling. In this case individuals are selected according to their Y ranks or their residual r ranks. In many studies units with large and/or small values of Y or r are over-sampled or, in some cases, sampled exclusively.

Many methods of estimation have been proposed for two-phase studies or missing data more generally; these include estimating functions that incorporate weights based on the selection probabilities or calibration (Robins et al., 1994, Lipsitz et al., 1999, Chatterjee et al., 2003, Breslow et al., 2009, Scott and Wild, 2011) and maximum likelihood methods (Lawless et al., 1999, Little and Rubin, 2002, Ibrahim et al., 2005, Zhao et al., 2009). For the testing context, likelihood methods are preferred because of their efficiency and generality. Unlike weighted estimating function methods, they can be applied to situations where some units have zero probability of selection in phase 2. Both full and conditional likelihood methods are widely used. Full likelihood estimation (Lawless et al., 1999) is based on the fact that

$$L(\theta, g) = \prod_{R_i=1} f(y_i \mid z_i, x_i; \theta) g(x_i \mid z_i) \prod_{R_i=0} f_1(y_i \mid z_i; \theta, g) \tag{2}$$

is the likelihood function for the observed data under a wide range of sampling plans for which X_i values for units with $R_i = 0$ are missing at random. Each of basic stratified, variable probability and rank based sampling formed on (Y, Z) or on regression residuals from $(Y \mid Z)$ satisfy the missing at random condition (Rubin, 1976). In (2), we redefine g for simplicity to denote the unknown distribution of X given Z; it is treated as an unknown parameter since it is needed for $f_1(y \mid z) = \int f(y \mid z, x) g(x \mid z) dx$, which denotes the distribution of Y given Z.

Conditional likelihoods can also be used in some settings. The likelihoods can vary according to the exact sampling plan, but are based on the distribution of Y_i for units with $R_i = 1$, given Z_i, X_i and the fact that the units were sampled. For example, in extreme Y sampling there are values C_l, C_u and individuals sampled come from either stratum S_1 consisting of observations with $Y \leq C_l$ or S_2 consisting of those with $Y \geq C_u$. In recent genetics-related work, Lin et al. (2013) have used full likelihood, whereas a number of other authors (Huang and Lin, 2007, Li et al., 2011, Barnett et al., 2013) have used the conditional likelihood

$$L_C(\theta) = \prod_{R_i=1} \frac{f(y_i \mid z_i, x_i; \theta)}{\text{pr}(Y_i \leq C_l \mid z_i, x_i; \theta) + \text{pr}(Y_i \geq C_u \mid z_i, x_i; \theta)}. \tag{3}$$

This is based on the conditional distribution of Y_i given X_i, Z_i and $R_i = 1$; a general formulation is given in Section 3.2. Strictly speaking, (3) only applies for variable probability sampling, but Lawless et al. (1999) show that (3) can be extended to deal with basic stratified sampling by placing terms $p_1 = n_1/N_1$ and $p_2 = n_2/N_2$ in front of the two terms in the denominator; in this case $L_C(\theta)$ is a pseudo-likelihood. It is an advantage that $g(x \mid z)$ does not appear in the conditional likelihood and so need not be estimated. On the other hand, conditional likelihood may be less efficient than full likelihood, and for sampling designs that are residual-based or rank-based, conditional likelihoods that are independent of $g(x \mid z)$ do not exist.

3 Likelihood-based Tests for Association

3.1 Full likelihood

We assume that interest centres on $f(y \mid x, z)$ for a model of the form (1) and in particular, on testing $H_0 : \beta = 0$. We let $\mu_i = \beta_0 + \beta' x_i + \gamma' z_i$, with x_i and z_i $p \times 1$ and $q \times 1$ vectors, respectively, and $\beta = (\beta_1, ..., \beta_p)', \gamma = (\gamma_1, ..., \gamma_q)'$. The full likelihood function (2) under missing at random sampling schemes has response model parameters $\theta = (\beta_0, \beta', \gamma', \sigma)'$. We seek to avoid parametric modelling assumptions for g and as is conventional (Zhao et al., 2009, Lin et al., 2013), we treat g as a discrete distribution with support determined by the distinct pairs (x_i, z_i) in the observed data. For asymptotic properties, however, we require that Z be discrete.

We let $\phi_0(y \mid \mu, \sigma) = \log\{f_0(y \mid \mu, \sigma)\}$ and $\phi_0'(y \mid \mu, \sigma) = \partial\phi_0(y \mid \mu, \sigma)/\partial\mu$. The likelihood (2) gives a score function for β whose components may then be expressed

$$U_r(\theta, g) = \frac{\partial \text{log} L(\theta, g)}{\partial \beta_r} = \sum_{i \in V} \phi_0'(y_i \mid \mu_i, \sigma)\, x_{ir} + \sum_{i \notin V} E\{\phi_0'(y_i \mid \mu_i, \sigma)\, X_{ir} \mid y_i, z_i\} \;\; (r = 0, 1, ..., p),$$

where $x_{i0} = 1$ and for convenience we let V denote as the set of units with $R_i = 1$. Under the null hypothesis $H_0 : \beta = 0$, the partial score test statistic is the vector

$$U(\hat{\theta}, \hat{g}) = \sum_{i \in V} \phi_0'(y_i \mid \hat{\mu_i}, \hat{\sigma})\, x_i + \sum_{i \notin V} \phi_0'(y_i \mid \hat{\mu_i}, \hat{\sigma})\, \hat{E}(X_i \mid z_i), \tag{4}$$

where $\hat{\theta} = (\hat{\beta}_0, 0, \hat{\gamma}, \hat{\sigma})$ and $\hat{g}$ are the maximum likelihood estimates under H_0, and $\hat{\mu}_i = \hat{\beta}_0 + \hat{\gamma}' z_i$. The expectation in (4) is based on the estimate $\hat{g}$, and we consider two cases. First, if X and Z are independent, then $g(x \mid z)$ becomes just $g(x)$, and it can be shown that $\hat{g}(x) = \sum_{i \in V} I(X_i = x)/n$ and $\hat{E}(X_i \mid y_i, z_i) = \hat{E}(X_i) = \bar{x}_n$, where n is the size of the phase 2 sample V. The first component of (4) with $r = 0$ and $x_{i0} = 1$ equals zero and this implies that

$$U_0(\hat{\theta}, \hat{g}) = \sum_{i \in V} \phi_0'(y_i \mid \hat{\mu_i}, \hat{\sigma}) + \sum_{i \notin V} \phi_0'(y_i \mid \hat{\mu_i}, \hat{\sigma}) = 0,$$

so the test statistic can be rewritten as

$$U = \sum_{i \in V} \phi_0' \left(y_i \mid \hat{\mu}_i, \hat{\sigma}\right) (x_i - \bar{x}_n). \tag{5}$$

We note that $r_i = \phi_0' \left(y_i \mid \hat{\mu}_i, \hat{\sigma}\right)$ is a generalized score residual for the fitted model (1) under H_0, and that (5) can also be written as

$$U = \sum_{i \in V} (r_i - \bar{r}_n) x_i, \tag{6}$$

where $\bar{r}_n$ is the average of the residuals for $i \in V$. This statistic has the same form as if the response-dependent sample V were a random sample.

In the case where X and Z are not independent, it can be shown from equations (12) and (13) of Zhao et al. (2009) that $\hat{E}(X_i \mid z_i) = \bar{x}_n(z_i)$, the mean value of X for individuals in V having $Z = z_i$. This gives, from (4), that

$$U(\hat{\theta}, \hat{g}) = \sum_{i \in V} \phi_0' \left(y_i \mid \hat{\mu}_i, \hat{\sigma}\right) x_i + \sum_{i \notin V} \phi_0' \left(y_i \mid \hat{\mu}_i, \hat{\sigma}\right) \bar{x}_n(z_i), \tag{7}$$

and there is no further simplification. A problem in this case is that a z_i value observed for a unit $i \notin V$ might not occur in V, and then $\bar{x}_n(z_i)$ in (7) is undefined. Purely nonparametric estimation of g requires for asymptotic results, see Section 3.4, that Z be discrete in any case, but for specific samples we may need to discretize Z further so that an undefined $\bar{x}_n(z_i)$ does not occur.

Various test statistics for H_0 can be based on U, for example linear statistics $1'U$ or quadratic statistics such as Hotelling's $U'\text{var}(U)^{-1}U$ (Derkach et al., 2014). The choice of statistic, and its potential power, depends on consideration of plausible alternatives to H_0. For estimation of $\text{var}(U)$ under H_0, if X and Z are independent we can employ the permutation variance estimate $X_c' X_c \sum_{i \in V} (r_i - \bar{r}_n)^2/(n-1)$, where X_c denotes the $n \times p$ centred X matrix. This can be used along with asymptotic normal or chi-squared approximations to obtain p-values for tests, or in small to moderate size samples we could obtain p-values by sampling from the permutation distribution for the test statistic, which arises from randomly permuting the X_i across the units in V. If X and Z are dependent, a permutation distribution does not apply, although when dependence is weak Type 1 errors are not distorted much (Anderson and Robinson, 2001). A model-based variance estimate can be used in this case; we discuss this in Section 3.4.

3.2 CONDITIONAL LIKELIHOOD

Consider a variable probability sampling scheme where $\text{pr}(R = 1 \mid Y = y, Z = z, X = x) = \text{pr}(R = 1 \mid Y = y, Z = z) = \pi(y, z)$ is a known function. A conditional likelihood for θ is based on $\text{pr}(Y \mid X = x, Z = z, R = 1)$,

$$L_c(\theta) = \prod_{i \in V} \frac{f(y_i \mid x_i, z_i; \theta)}{B(\mu_i, \sigma)}, \tag{8}$$

where $B(\mu, \sigma) = \int f(y \mid x, z; \theta) \pi(y, z) dy = \int f_0 \left(y \mid \mu, \sigma\right) \pi(y, z) dy$. Then,

$$U_r(\theta) = \frac{\partial \text{log} L_c(\theta)}{\partial \beta_r} = \sum_{i \in V} \{\phi_0' \left(y_i \mid \mu_i, \sigma\right) - A\left(\mu_i, \sigma\right)\} x_{ir}, \text{ for } r = 0, 1, ..., p, \tag{9}$$

where $A(\mu, \sigma) = \partial \text{log} B(\mu, \sigma)/\partial \mu$. Under $H_0 : \beta = 0$, the statistic $U = \left(U_1(\hat{\theta}), ..., U_p(\hat{\theta})\right)'$ is then

$$U = \sum_{i \in V} \left\{\phi_0' \left(y_i \mid \hat{\mu}_i, \hat{\sigma}\right) - A\left(\hat{\mu}_i, \hat{\sigma}\right)\right\} x_i,$$

where $\hat{\mu}_i = \hat{\beta}_0 + \hat{\gamma}' z_i$.

Variance estimation for U can be based on observed or expected information from the conditional likelihood in the case of variable probability sampling. Conditional likelihood can be extended to basic stratified sampling, in which case the likelihood becomes a pseudo likelihood, and a sandwich variance estimator should be used. Lawless et al. (1999) provide results for conditional likelihood and pseudo likelihood.

3.3 Case with No Covariates Z

When there are no supplementary covariates Z, γ disappears in Section 3.2 and we note that with $r = 0$ and $x_{i0} = 1$ the statistic $U_0(\theta)$ equals zero at $\hat{\beta}_0, \hat{\sigma}$. This implies that $A(\hat{\beta}_0, \hat{\sigma}) = n^{-1} \sum_{i \in V} \phi'_0 \left(y_i \mid \hat{\beta}_0, \hat{\sigma} \right)$ and therefore U is exactly the same as the statistic (5) based on full likelihood when γ is dropped.

In the Gaussian case, the test statistic (5) is proportional to $S = \sum_{i=1}^{n} (Y_i - \bar{Y}) X_i$, and the equality of the conditional and full likelihood statistics has previously been shown by (Tang, 2010). Our result shows this equivalence holds for the very general model of form (1). This indicates that tests of H_0 based on full or conditional likelihood will have the same local power within family (1), and explains why in some simulation studies (Huang and Lin, 2007) Wald or score statistics based on the two likelihoods have had similar power. It can also be shown that asymptotic model-based variance estimates for U are equivalent under the general model (1); a proof for this is given in the 2014 University of Toronto PhD thesis of the first author. Variance estimation is outlined in the following section.

3.4 Variance Estimates

The model-based covariance matrix of the full likelihood score statistic (7) is described here; the derivation is outlined in the Supplementary Material. Variance estimation has not previously been addressed in the full generality of the sampling plans considered here. We assume Z is discrete and with a slight abuse of notation we denote the distinct values of Z by $z(1), ..., z(K)$; we also let $\Delta_{ik} = I\{z_i = z(k)\}$ and $\epsilon_i = \phi'_0(y_i \mid \mu_i, \sigma)$. In the Supplementary Material, see equation (S.7), we show how the asymptotic covariance matrix for U can be obtained:

$$\text{var}(U) = \text{var}_1(U) + \sum_{k=1}^{K} \left[E\left(\sum_{i \in V} \epsilon_i^2 \Delta_{ik} \right) - \frac{\{E\left(\sum_{i \in V} \epsilon_i \Delta_{ik}\right)\}^2}{E(n_k)} \right] \text{var}\{X \mid z(k)\}, \tag{10}$$

where the first term in this expression is the covariance matrix of U when values of X for all individuals $i = 1, ..., N$ are set equal to $E(X \mid Z = z_i)$ and treated as known. Also, $n_k = \sum_{i=1}^{N} I\{z_i = z(k)\} I(R_i = 1)$ is the number of occurrences of $Z = z(k)$ among units in V. We estimate the first term in (10) by replacing expected values such as $E\{\phi''_{\mu\mu}(y_i \mid \mu_i, \sigma)\}$ with corresponding observed values $\phi''_{\mu\mu}(y_i \mid \hat{\mu}_i, \hat{\sigma})$, and by replacing $E(X_i \mid z_i)$ with $\hat{E}(X_i \mid z_i) = \bar{x}_n(z_i)$. Conditional variance var$\{X \mid z(k)\}$ is replaced by the sample covariance matrix based on $\{x_i, i \in V : z_i = z(k)\}$. With $r_i = \phi'_0(y_i \mid \hat{\mu}_i, \hat{\sigma})$, the remaining second terms in (10) are estimated by $\sum_{i \in V} \left(r_i \Delta_{ik} - \sum\limits_{i \in V} r_i \Delta_{ik} / n_k \right)^2$ for $k = 1, \ldots, K$.

In the Supplementary Material, we also show that when there are no covariates Z, the estimate of (10) is equivalent to the permutation covariance matrix: var$(U) = X'_c X_c s_r^2$. When additional covariates Z are present then provided they are independent of X, the permutation covariance matrix can still be used, with r_i now the residuals from the fit of Y on Z for the full phase 1 sample of size N.

4 Concluding Remarks

The equivalence of score test statistics based on full and conditional likelihoods explains simulation results in the literature that demonstrate their near equivalence under Y-dependent sampling in the Gaussian case. Similar results will occur for a broad range of models (1). This note shows that tests of association between expensive covariates X and a response variable Y based on two-phase, response-dependent samples can be derived from full semiparametric maximum likelihood across the family of models. Our approach assumes that any additional covariates Z are categorical, and continuous covariates are handled by discretizing them. Simulation results presented in Supplementary Material and in the 2014 University of Toronto PhD thesis of the first author show that for many practical testing scenarios, the effect of doing this is slight. Alternative methods that seek to estimate the conditional probabilities $g(x \mid z)$ for continuous z by smoothing can be developed (Zeng and Lin, 2014) but they also involve approximations where finite samples are concerned. We also showed that when additional covariates Z are absent, the full likelihood score statistics are identical to conditional likelihood score statistics. When covariates Z are present, conditional likelihood tests do not require estimation of $g(x \mid z)$; however, they do not apply to certain types of sampling plans and are less powerful than tests based on full likelihood, which we have shown are easy to apply. Finally, we note that the statistic (7) has mean zero under the null hypothesis when (Y, Z) and X are independent, whether or not (1) is the correct distribution for Y given X and Z. The tests are thus robust with respect to Type 1 error, but there will be some power loss under model misspecification.

We remark that family (1) includes many models used in dealing with survival or event times, including proportional hazards, accelerated failure time, proportional odds, and more general transformation family models (Kalbfleisch and Prentice, 2002, Lawless, 2003). Survival time data are typically subject to right censoring, but it is easily seen that the methods and results in this paper apply to censored data likelihood functions, the main effect being that the score residuals in our test statistics have a slightly more complex form. Mendolia et al. (2014) and Chen et al. (2014) have considered genetic association testing with survival time outcomes for the case of proportional hazards models; our treatment here greatly expands the scenarios and models that can conveniently be considered. Although we have considered fully parametric models, it is possible to extend the discussion to semi-parametric generalized linear models by the use of semi-parametric maximum likelihood (Zeng and Lin, 2007).

We reiterate that since H_0 involves a p-dimensional covariate X, various test statistics could be based on the score vectors (7) or (4); see for example Li and Lagakos (2006) and Derkach et al. (2014). The choice of statistic as well as the precise type of Y-dependent sampling can have a major effect on power; see, Li and Lagakos (2006). Basu and Pan (2011) and Derkach et al. (2014) provide extensive simulation results for genetic association studies of rare variants based on random or covariate stratified samples, and Lee et al. (2014) summarize current methodology and related issues.

Finally, we remark that inverse probability-weighted, Horvitz–Thompson estimating functions are sometimes used with biased sampling plans (Lawless et al., 1999, Scott and Wild, 2011). These methods do not apply when some individuals have zero probability of selection, as in the case of extreme Y sampling plans, so they have not been considered in this paper. They also tend to be less efficient than the likelihood-based methods discussed here, although augmented inverse probability-weighted estimating functions that have efficiencies closer to those of maximum likelihood can be found (Scott and Wild, 2011).

Supplementary Material

Supplementary material available at *Biometrika* online includes derivation of model-based variance estimates for full likelihood presented in Section 3.4. We also present simulation results for genetic association studies with rare variants, to evaluate these variance estimates.

Acknowledgement

The authors thank the editor, associate editor and reviewer for helpful comments. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada and a Canadian Institutes of Health Research Strategic Training for Advanced Genetic Epidemiology fellowship.

References

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88.

Barnett, I. J., Lee, S., and Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic Epidemiology*, 37(2):142–151.

Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7):606–619.

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L., and Kulich, M. (2009). Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*, 1(1):32–49.

Chatterjee, N., Chen, Y.-H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168.

Chen, H., Lumley, T., Brody, J., Heard-Costa, N. L., Fox, C. S., Cupples, L. A., and Dupuis, J. (2014). Sequence kernel association test for survival traits. *Genetic Epidemiology*, 38(3):191–197.

Derkach, A., Lawless, J. F., and Sun, L. (2014). Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science*, 29(2):302–321.

Diggle, P. J., Menezes, R., and Su, T.-L. (2010). Geostatistical inference under preferential sampling (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.

Huang, B. E. and Lin, D. Y. (2007). Efficient association mapping of quantitative trait loci with selective genotyping. *American Journal of Human Genetics*, 80(3):567–76.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models. *Journal of the American Statistical Association*, 100(469):332–346.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York, 2nd edition.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. John Wiley, 2nd edition.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, 95(1):5–23.

Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E., and Conti, D. (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic Epidemiology*, 35(8):790–799.

Li, Q. H. and Lagakos, S. W. (2006). On the relationship between directional and omnibus statistical tests. *Scandinavian Journal of Statistics*, 33(2):239–246.

Lin, D.-Y., Zeng, D., and Tang, Z.-Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences*, 110(30):12247–12252.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94(448):1147–1160.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York, 2nd edition.

McCullagh, P. (2008). Sampling bias and logistic models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):643–677.

Mendolia, F., Klein, J. P., Petersdorf, E. W., Malkki, M., and Wang, T. (2014). Comparison of statistics in association tests of genetic markers for survival outcomes. *Statistics in Medicine*, 33(5):828–844.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Scott, A. J. and Wild, C. J. (2011). Fitting regression models with response-biased samples. *Canadian Journal of Statistics*, 39(3):519–536.

Tang, Y. (2010). Equivalence of three score tests for association mapping of quantitative trait loci under selective genotyping. *Genetic Epidemiology*, 34(5):522–527.

Zeng, D. and Lin, D. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association*, 109(505):371–383.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):507–564.

Zhao, Y., Lawless, J. F., and McLeish, D. L. (2009). Likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal*, 51(1):123–136.