A joint regression analysis for genetic association studies with outcome stratified samples

Colin O. Wu[1*], Gang Zheng[1] and Minjung Kwak[2]
[1] National Heart, Lung and Blood Institute, Bethesda, MD 20892, USA.
[2] Yeungnam University, Kyeongbuk, 712-749, SOUTH KOREA.
[*] Corresponding author: Colin O. Wu, email:*wuc@nhlbi.nih.gov*

Abstract

Genetic association studies in practice often involve multiple traits resulting from a common disease mechanism, and samples for such studies are often stratified based on some trait outcomes. In such situations, statistical methods using only one of these traits may be inadequate and lead to under-powered tests for detecting genetic associations. We propose in this paper an estimation and testing procedure for evaluating the shared-association of a genetic marker on the joint distribution of multiple traits of a common disease. Specifically, we assume that the disease mechanism involves both quantitative and qualitative traits, and our samples could be stratified based on the qualitative trait. Through a joint likelihood function, we derive a class of estimators and test statistics for evaluating the shared genetic association on both the quantitative and qualitative traits. Our simulation study shows that the joint likelihood test procedure is potentially more powerful than association tests based on separate traits. Application of our proposed procedure is demonstrated through the rheumatoid arthritis data provided by the Genetic Analysis Workshop 16 (GAW16).

Keywords: genetic association study, joint regression model, pleiotropic analysis, qualitative trait, quantitative trait, stratified sample.

1 Introduction

We propose in this paper a joint regression approach to evaluate the genetic associations and covariate effects on both a quantitative trait and a multinomial qualitative trait. In order to include combined data from multiple studies, we focus on samples stratified based on the qualitative trait with the quantitative trait observed on all or some of the strata. The quantitative trait may not be measured on some strata, partly due to the practical issues of high cost and low scientific values of having its observations on these strata. Our motivating example is the Rheumatoid Arthritis (RA)

data provided by the Genetic Analysis Workshop 16 (GAW16) (Amos et al., 2009). This case-control study combines 868 RA positive patients (cases) from the North American Rheumatoid Arthritis Consortium and 1194 RA negative subjects from the New York Cancer Project (controls) and contains genotype data of 545,080 SNPs. Since the anti-cyclic citrullinated peptide (anti-CCP) antibody is a potentially important surrogate marker for diagnosis and prognosis in RA and higher anti-CCP levels have been linked to increased severity of RA (Huizinga et al., 2005), the GAW16 data contain anti-CCP measurements for the RA positive patients but not for the RA negative subjects. By treating the anti-CCP value as a quantitative trait and the RA status as a qualitative trait, our methodology is capable of estimating and testing the associations of SNPs or candidate genes jointly with these two traits. Numerical studies and detail and extension of our method are presented in the main paper published (Wu et al. 2013).

2 Data Structure and Joint Regression Models

Let $\{X, Y, \mathbf{G}, \mathbf{Z}\}$ be the random variables being considered, where X is a qualitative trait with values in $\{0, 1, \ldots, K\}$, Y is a real valued quantitative trait, $\mathbf{G}$ is a categorical variable denoting the value of genetic markers or SNPs, and $\mathbf{Z} = (Z^{(1)}, \ldots, Z^{(D)})^T$ is an R^D-valued, $D \geq 1$, covariate matrix. Specific choices for coding $\mathbf{G}$ depend on whether the biallalic genetic model is recessive (REC), additive (ADD) or dominant (DOM), and $\mathbf{G}$ may involve multiple SNPs or candidate genes. The outcome stratified sample $\mathcal{D}$ based on X is the combination of $K+1$ subsamples $\mathcal{D}_0, \ldots, \mathcal{D}_K$ with sample sizes $n_0, \ldots, n_K$, respectively, and the overall sample size $n = \sum_{k=0}^{K} n_k$. When $K = 1$, $\mathcal{D} = \{\mathcal{D}_0, \mathcal{D}_1\}$ is a case-control study with $\mathcal{D}_1$ and $\mathcal{D}_0$ being the samples for cases and controls, respectively. Since $\mathcal{D}$ is stratified based on the values of the qualitative trait X, X_i's are observed for all the subjects $i = 1, \ldots, n$. Because Y given certain levels of X may not be scientifically important, Y_i may not be measured in some of the subsamples. We assume that there is a constant $0 \leq K_0 \leq K$, such that $\mathcal{D}_k = \{X_i = k, \mathbf{G}_i, \mathbf{Z}_i; i = \sum_{l=0}^{k-1} n_l + 1, \ldots, \sum_{l=0}^{k} n_l\}$ if $0 \leq k < K_0$, and $\mathcal{D}_k = \{X_i = k, Y_i, \mathbf{G}_i, \mathbf{Z}_i; i = \sum_{l=0}^{k-1} n_l + 1, \ldots, \sum_{l=0}^{k} n_l\}$ if $K_0 \leq k \leq K$.

Since the observations in $\mathcal{D}$ are stratified based on the values of X_i, a joint regression model for $(X_i, Y_i)^T$ with covariates $\{\mathbf{G}_i, \mathbf{Z}_i\}$ can be constructed by modeling the conditional probability $P(X_i = k|\mathbf{G}_i, \mathbf{Z}_i)$, and, for each $X_i = k$, the conditional mean of Y_i given $\{\mathbf{G}_i, \mathbf{Z}_i\}$. If a linear model for Y_i is used, a linear joint model for $\mathcal{D}$ is

$$\begin{cases} g\Big[P(X_i = k|\mathbf{G}_i, \mathbf{Z}_i)\Big] = \beta_0^{(k)} + (\beta_1^{(k)})^T \mathbf{G}_i + (\beta_2^{(k)})^T \mathbf{Z}_i, \quad \text{for } 1 \leq k \leq K, \\ P(X_i = 0|\mathbf{G}_i, \mathbf{Z}_i) = 1 - \sum_{k=1}^{K} P(X_i = k|\mathbf{G}_i, \mathbf{Z}_i), \\ Y_i|_{X_i=k} = \alpha_0^{(k)} + (\alpha_1^{(k)})^T \mathbf{G}_i + (\alpha_2^{(k)})^T \mathbf{Z}_i + \epsilon_i^{(k)}, \ K_0 \leq k \leq K, \end{cases} \tag{1}$$

where $g(\cdot)$ is a known link function, $\beta_1^{(k)}$ and $\beta_2^{(k)}$ describe the genetic association of $\mathbf{G}_i$ and the covariate effect of $\mathbf{Z}_i$ on the probability of $X_i = k$, $\alpha_1^{(k)}$ and $\alpha_2^{(k)}$ describe the additional genetic association of $\mathbf{G}_i$ and the covariate effect of $\mathbf{Z}_i$ on the quantitative trait Y_i within each given level of $X_i = k$, and $\epsilon_i^{(k)}$ are the mean zero random errors with variances σ_k^2. Since Y_i is not observed when $X_i = k$ with $0 \leq k < K_0$, the effects of $\{\mathbf{G}_i, \mathbf{Z}_i\}$ on Y_i are not identifiable if there are no further assumptions on Y_i given $X_i = k$ and $0 \leq k < K_0$.

An important assumption for the models (1) is that the disease status is first categorized by the qualitative trait X, then, within certain levels of X, the disease severity is further measured

by the quantitative trait Y. The parameters $\beta_1^{(k)}$ describe the associations of $\mathbf{G}_i$ with the disease categories in the first step, and $\alpha_1^{(k)}$ describe the further genetic associations of $\mathbf{G}_i$ with the disease severity within a given disease category. Thus, to evaluate the shared genetic associations of $\mathbf{G}_i$ with both X and Y, we would like to test:

$$\begin{cases} H_0\text{: } \beta_1^{(k)} = \mathbf{0} \text{ for all } 1 \le k \le K, \text{ and } \alpha_1^{(k)} = \mathbf{0} \text{ for all } K_0 \le k \le K; \\ H_1\text{: } \beta_1^{(k)} \ne \mathbf{0} \text{ for some } 1 \le k \le K, \text{ or } \alpha_1^{(k)} \ne \mathbf{0} \text{ for some } K_0 \le k \le K, \end{cases} \tag{2}$$

where $\mathbf{0}$ denotes the column vector of 0 with the appropriate length. When the observations of Y_i are ignored in (1), (2) reduces to the test: "H_0^b: $\beta_1^{(k)} = \mathbf{0}$ for all $1 \le k \le K$" versus "H_1^b: $\beta_1^{(k)} \ne \mathbf{0}$ for some $1 \le k \le K$". If only Y_i is used for testing the genetic associations of $\mathbf{G}_i$, we would like to test: "H_0^a: $\alpha_1^{(k)} = \mathbf{0}$ for all $K_0 \le k \le K$" versus "H_1^a: $\alpha_1^{(k)} \ne \mathbf{0}$ for some $K_0 \le k \le K$" based on the quantitative trait part of (1). Since the parameters in H_0 are subsets of the ones in H_0^b or H_0^a, the true pleiotropic associations of $\mathbf{G}_i$ with the disease, which may be categorized by both disease categories and disease severity, are more likely to be detected by incorporating the information of X_i and Y_i in the joint models (1). Extensions of (1) to partially linear joint models where the linear structure assumption is not appropriate is dealt in detail in the main paper (Wu et al. 2013).

3 Likelihood based Estimation and Testing

Let $f_x(y|\mathbf{g},\mathbf{z})$ be the conditional density of $Y = y$ given $\{x, \mathbf{g}, \mathbf{z}\}$, $\pi(\mathbf{g},\mathbf{z})$ the joint density of $\{\mathbf{g},\mathbf{z}\}$, and $f(y, x|\mathbf{g},\mathbf{z}) = f_x(y|\mathbf{g},\mathbf{z})\, P(x|\mathbf{g},\mathbf{z})$ the conditional density of $(Y, X)^T$ given $\{\mathbf{g},\mathbf{z}\}$ in the population. Following Equation (1.2) of Wu (2000), the joint density of $\{Y_i, X_i, \mathbf{G}_i, \mathbf{Z}_i\}$ for the observed stratified sample $\mathcal{D}_k$ is

$$f(y, x, \mathbf{g}, \mathbf{z}) = \frac{f_x(y|\mathbf{g},\mathbf{z})\, P(x|\mathbf{g},\mathbf{z})\, \pi(\mathbf{g},\mathbf{z})}{W_k} 1_{[x=k]}, \tag{3}$$

where $W_k = P(X = k)$ and $1_{[\cdot]}$ is the indicator function with $1_{[x=k]} = 1$ if $x = k$, and 0 otherwise. Since $\{X_i, \mathbf{G}_i, \mathbf{Z}_i\}$ are observed for $X_i = k$, $0 \le k < K_0$, and $\{Y_i, X_i, \mathbf{G}_i, \mathbf{Z}_i\}$ are observed for $X_i = k$, $K_0 \le k \le K$, we define $L_{X,k,i}(X_i, \mathbf{G}_i, \mathbf{Z}_i) = \log P(X_i = k|\mathbf{G}_i, \mathbf{Z}_i)$, and $L_{Y,k,i}(Y_i, \mathbf{G}_i, \mathbf{Z}_i) = \log f_k(Y_i|\mathbf{G}_i, \mathbf{Z}_i)$ if $K_0 \le k \le K$, and 0 if $0 \le k < K_0$. The full log-likelihood function for the observed data $\mathcal{D}$ is

$$L(\mathcal{D}) = \sum_{k=0}^{K} \Big[L_{X,k}(\mathcal{D}) + L_{Y,k}(\mathcal{D}) \Big] + \sum_{k=0}^{K} L_{W,k}(\mathcal{D}), \tag{4}$$

where, with $m_{-1} = 0$, $m_k = n_0 + \cdots + n_k$, $L_{W,k}(\mathcal{D}) = \sum_{i=m_{k-1}+1}^{m_k} \log\left[\pi(\mathbf{G}_i, \mathbf{Z}_i)/W_k\right]$,

$$L_{X,k}(\mathcal{D}) = \sum_{i=m_{k-1}+1}^{m_k} L_{X,k,i}(X_i, \mathbf{G}_i, \mathbf{Z}_i) \quad \text{and} \quad L_{Y,k}(\mathcal{D}) = \sum_{i=m_{k-1}+1}^{m_k} L_{Y,k,i}(Y_i, \mathbf{G}_i, \mathbf{Z}_i).$$

We assume throughout that $\pi(\cdot,\cdot)$ and W_k do not depend on the model parameters, so that $L(\mathcal{D})$ of (4) depends on the model through the partial log-likelihood function

$$L_{X,Y}(\mathcal{D}) = L_X(\mathcal{D}) + L_Y(\mathcal{D}), \tag{5}$$

where $L_X(\mathcal{D}) = \sum_{k=0}^{K} L_{X,k}(\mathcal{D})$, $L_Y(\mathcal{D}) = \sum_{k=0}^{K} L_{Y,k}(\mathcal{D})$, and the estimation and inference can be acheived by maximizing $L_{X,Y}(\mathcal{D})$ over the parameters.

Let $\theta = (\alpha^T, \beta^T, \sigma^T)^T$ be the vector of parameters of (1) within the parameter space Θ. Under the assumption that $\pi(\cdot,\cdot)$ and W_k do not depend on $\theta \in \Theta$, the maximum likelihood estimator (MLE) $\widehat{\theta} = (\widehat{\alpha}^T, \widehat{\beta}^T, \widehat{\sigma}^T)^T$ satisfies $S_{X,Y}(\mathcal{D}; \widehat{\theta}) = \partial L_{X,Y}(\mathcal{D}; \theta)/\partial\theta|_{\theta=\widehat{\theta}} = \mathbf{0}$, which are equivalent to

$$S_X(\mathcal{D}; \widehat{\beta}) = \frac{\partial L_X(\mathcal{D}; \beta)}{\partial \beta}\Big|_{\beta=\widehat{\beta}} = \mathbf{0} \quad \text{and} \quad S_Y(\mathcal{D}; \widehat{\alpha}, \widehat{\sigma}) = \frac{\partial L_Y(\mathcal{D}; \alpha, \sigma)}{\partial (\alpha^T, \sigma^T)^T}\Big|_{\alpha=\widehat{\alpha}, \sigma=\widehat{\sigma}} = \mathbf{0}. \tag{6}$$

It follows from (6) that θ can be computed by separately maximizing $L_X(\mathcal{D}; \beta)$ and $L_Y(\mathcal{D}; \alpha, \sigma)$ with respect to β and $(\alpha^T, \sigma^T)^T$.

Asymptotic properties of $\widehat{\theta}$, such as consistency and asymptotically normality, can be developed using the usual derivations for the MLEs (Serfling, 1980, Section 4.2). If we assume that (a) $n_k/n \to c_k$ for some constant $0 < c_k < 1$ and all $k = 0, \ldots, K$, (b) the Fisher information matrix $\mathcal{I}(\theta) = -E[\partial^2 \log f(Y, X, \mathbf{G}, \mathbf{Z}|\theta) \big/ (\partial\theta_u\, \partial\theta_v)]$ is positive definite with inverse $\mathcal{I}^{-1}(\theta)$, where θ_a is the ath element of θ, and (c) the regularity conditions of MLEs, such as Serfling (1980, Section 4.2, R1-R3), are satisfied, then $\widehat{\theta}$ has approximately the $\mathcal{N}(\theta, n^{-1}\mathcal{I}^{-1}(\theta))$ distribution when n is sufficiently large. The approximate $100 \times (1-\alpha)\%$ confidence interval for $C^T\theta$ is $C^T\widehat{\theta} \pm Z_{1-\alpha/2} n^{-1} C^T \mathcal{I}^{-1}(\widehat{\theta}) C$, where C is a known column vector and Z_α is the $(100 \times \alpha)$th quantile of the standard normal distribution.

Since the genetic associations in (2) are linear hypotheses of the form, H_0: $\Gamma^T\theta = \mathbf{0}$ versus H_1: $\Gamma^T\theta \neq \mathbf{0}$, for an appropriate matrix Γ such that $\Gamma^T\theta = ((\alpha_1^{(K_0)})^T, \ldots, (\alpha_1^{(K)})^T, (\beta_1^{(1)})^T, \ldots, (\beta_1^{(K)})^T)^T$, we can constuct three test statistics by comparing the log-likelihood functions and the MLEs computed under H_0 and H_1. Let $\widetilde{\theta} = (\widetilde{\alpha}^T, \widetilde{\beta}^T, \widetilde{\sigma}^T)^T$ be the MLE of θ computed by (6) under the null hypothesis H_0, that is, $\widetilde{\alpha}_1^{(k)} = \alpha_1^{(k)} = \mathbf{0}$ for $K_0 \leq k \leq K$ and $\widetilde{\beta}_1^{(k)} = \beta_1^{(k)} = \mathbf{0}$ for $1 \leq k \leq K$. We have the following three asymptotically equivalent test statistics:

(a) the likelihod ratio test statistic: $LRT(\mathcal{D}) = 2[L_{X,Y}(\mathcal{D}; \widehat{\theta}) - L_{X,Y}(\mathcal{D}; \widetilde{\theta})]$;

(b) the Wald statistic $W(\mathcal{D}) = n(\Gamma^T\widehat{\theta})^T[\Gamma^T\mathcal{I}^{-1}(\widehat{\theta})\Gamma]^{-1}(\Gamma^T\widehat{\theta})$;

(c) the score statistic: $SC(\mathcal{D}) = n^{-1} S_{X,Y}^T(\mathcal{D}; \widetilde{\theta}) \mathcal{I}^{-1}(\widetilde{\theta}) S_{X,Y}(\mathcal{D}; \widetilde{\theta})$.

Under partially linear joint models, the parameter vector and parameter space are $\theta^* = ((\alpha^*)^T, \beta^T, \sigma^T, \mu^T)^T$ and Θ^*, respectively. When $\pi(\cdot,\cdot)$ and W_k do not depend on θ^* and the elements of μ are smooth functions, a commonly used approach for nonparametric analysis is to approximate $\mu_{k,d}(\cdot)$ for each (k, d) by a basis expansion of the form

$$\mu_{k,d}(z) \approx \sum_{l=1}^{l_{k,d}} \gamma_{k,d,l} B_{k,d,l}(z) = \gamma_{k,d}^T \mathbf{B}_{k,d}(z), \tag{7}$$

where $\{B_{k,d,l}(\cdot); l = 1, \cdots, l_{k,d}\}$ is a set of basis functions, $\gamma_{k,d} = (\gamma_{k,d,1}, \ldots, \gamma_{k,d,l_{k,d}})^T$ and $\mathbf{B}_{k,d}(z) = (B_{k,d,1}(z), \ldots, B_{k,d,l_{k,d}}(z))^T$ (e.g., Stone, 1994). Although any basis approximation may be considered, appropriate basis choices in practice may depend on the specific nature of the data. For example, a Fourier basis may be used when the underlying functions have periodicity, global polynomials may be adequate for smooth functions, and B-splines (polynomial splines) may be desirable for exhibiting local features. Because of their local flexibility and numerical stability, we use B-spline bases in our simulation study.

Let $f_k^*(Y_i;\alpha^*,\sigma,\gamma|\mathbf{G}_i,\mathbf{Z}_i)$ be the approximate conditional density of Y_i obtained by substituting $\mu_{k,d}(\cdot)$ with $\gamma_{k,d}^T\mathbf{B}_{k,d}(\cdot)$, where $\gamma=(\gamma_{K_0,1}^T,\ldots,\gamma_{K,D_0}^T)^T$. The log-likelihood function $L_Y(\mathcal{D};\alpha^*,\sigma,\mu)$ can then be approximated by

$$L_Y^*(\mathcal{D};\alpha^*,\sigma,\gamma)=\sum_{k=0}^{K}\sum_{i=m_{k-1}+1}^{m_k}L_{Y,k,i}^*(Y_i,\mathbf{G}_i,\mathbf{Z}_i;\alpha^*,\sigma,\gamma), \tag{8}$$

where $L_{Y,k,i}^*(Y_i,\mathbf{G}_i,\mathbf{Z}_i;\alpha^*,\sigma,\gamma)=\log f_k^*(Y_i;\alpha^*,\sigma,\gamma|\mathbf{G}_i,\mathbf{Z}_i)$ if $K_0\le k\le K$, and 0 otherwise. Let $\xi=(\alpha^{*T},\beta^T,\sigma^T,\gamma^T)^T$ be the vector of parameters involved in the approximate log-likelihood function. We can approximate the likelihood by $L_{X,Y}^*(\mathcal{D};\xi)=L_X(\mathcal{D};\beta)+L_Y^*(\mathcal{D};\alpha^*,\sigma,\gamma)$. Similar to (6), the approximate MLE $\widehat{\xi}=(\widehat{\alpha}^{*T},\widehat{\beta}^T,\widehat{\sigma}^{*T},\widehat{\gamma}^T)^T$ of ξ maximizes $L_{X,Y}^*(\mathcal{D};\xi)$ and satisfy the normal equations $S_{X,Y}^*(\mathcal{D};\widehat{\xi})=\mathbf{0}$, or equivalently

$$S_X(\mathcal{D};\widehat{\beta})=\mathbf{0}\ \text{ and }\ S_Y^*(\mathcal{D};\widehat{\alpha}^*,\widehat{\sigma}^*,\widehat{\gamma})=\mathbf{0}. \tag{9}$$

Because of (7), the approximate MLE $\widehat{\sigma}^*$ of σ in (9) may be different from the MLE $\widehat{\sigma}$ obtained in (6). Substituting $\widehat{\gamma}=(\widehat{\gamma}_{K_0,1}^T,\ldots,\widehat{\gamma}_{K,D_0}^T)^T$ back to (7), the predicted value for $\mu_{k,d}(z)$ is $\widehat{\mu}_{k,d}(z)=\widehat{\gamma}_{k,d}^T\mathbf{B}_{k,d}(z)$. The estimation for $\theta^*=((\alpha^*)^T,\beta^T,\sigma^T,\mu^T)^T$ depends on the approximation (7), where $l_{k,d}$ may increase as the sample size n increases.

Since the asymptotic distributions of $\widehat{\xi}$ and $\widehat{\mu}_{k,d}(z)$ have not been established, we propose a bootstrap procedure based on the approximate likelihood ratio statistics from (7) for testing with additive nonparametric components $\mu_{k,d}(\cdot)$. Let $\widetilde{\xi}=(\widetilde{\alpha}^{*T},\widetilde{\beta}^T,\widetilde{\sigma}^{*T},\widetilde{\gamma}^T)^T$ be the approximate MLEs from $L_{X,Y}^*(\mathcal{D},\xi)$ under the null hypothesis H_0, that is, $\widetilde{\alpha}_1^{(k)}=\alpha_1^{(k)}=\mathbf{0}$ for $K_0\le k\le K$ and $\widetilde{\beta}_1^{(k)}=\beta_1^{(k)}=\mathbf{0}$ for $1\le k\le K$, and let $LRT^*(\mathcal{D})=L_{X,Y}^*(\mathcal{D};\widehat{\xi})-L_{X,Y}^*(\mathcal{D};\widetilde{\xi})$ be the approximate likelihood ratio test statistic. We reject H_0 if $LRT^*(\mathcal{D})>c_\alpha$ for some α-level critical value c_α. Details of the bootstrap procedure for computing c_α and extensive numerical studies are given in the main paper (Wu et al. 2013).

4 Discussion

We have developed a joint model approach for analyzing genome-wide association study data with a qualitative trait and a quantitative trait, and proposed a likelihood-based method to estimate and test the pleiotropic genetic associations with the traits. Our estimation and inference procedures may be applied to stratified samples where the quantitative trait is measured on some of the strata. By jointly modeling both the quantitative and qualitative traits in a single analysis, a joint association testing procedure may lead to more powerful tests than procedures based on single-trait association tests. Further research is needed to develop estimation and testing procedures which may incorporate additional linear or additive nonparametric terms such as gene-gene and gene-environmental interactions. For regression models with nonparametric components, further investigation is warranted to compare the joint model association tests with various potential approaches of combining separate single-trait association tests.

References

[1] Amos, C. I., Chen, W. V., Seldin, M. F., Remmers, E. F., Taylor, K. E., Criswell, L. A., Lee, A. T., Plenge, R. M., Kastner, D. L. and Gregersen, P. K. (2009). Data for genetic analysis workshop 16 Problem 1, association analysis of rheumatoid arthritis data. *BMC Proceedings*, **3** (Suppl 7), S2.

[2] Huizinga, T. W., Amos, C. I., van der Helm-van Mil, A. H., Chen, W., van Gaalen, F. A., Jawaheer, D., Schreuder, G. M., Wener, M., Breedveld, F. C., Ahmad, N., Lum, R. F., de Vries, R. R., Gregersen, P. K., Toes, R. E., and Criswell, L. A. (2005). Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis Rheumatoid*, **52**, 3433–3438.

[3] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

[4] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Annals of Statistics*, **22**, 118–171.

[5] Wu, C. O. (2000). Local polynomial regression with selection biased data. *Statistica Sinica*, **10**, 789–817.

[6] Wu, C. O., Zheng, G., and Kwak, M. (2013) A joint regression analysis for genetic association studies with outcome stratified samples. *Biometrics*, DOI:10.1111/biom.12012.