Appendix to: Novel insights into activity patterns in children, found using functional data analyses

Jeff Goldsmith, Xinyue Liu, Judith Jacobson, Andrew Rundle

## A.1 Estimation of the Function-on-Scalar Regression model

In Section 3, we describe the formulation and interpretation of the linear function-on-scalar regression model. The purpose of this Appendix is to provide the technical details for the estimation of this model, and our presentation closely follows that of (Ramsay and Silverman 2005) and (Reiss, Huang, and Mennes 2010).

Briefly, for functional responses $y\_{i}(t)$ and scalar covariates $x\_{i1},…,x\_{ip}$ (e.g. age, sex, asthma diagnosis), the linear function-on-scalar regression model is

$$y\_{i}(t)=β\_{0}(t)+\sumβ\_{k}\_{k=1}^{p}(t)x\_{ik}+ϵ\_{i}(t)$$

with coefficient functions $β\_{k}(t)$ and residuals term $ϵ\_{i}(t)$. For estimation of this model, we expand the coefficient functions using a cubic B-spline basis $Θ(t)$ with $M$ basis functions, so that $β\_{k}(t)=β\_{k}^{T}Θ(t)$; notationally, $Θ(t)$ contains basis functions as rows and $β\_{k}$ is a length-$M$ column vector.

The inclusion of the parameter $t$ in the preceding emphasizes the conceptual framework for functional data analyses, but in practice functional responses are observed on a finite grid; in our example, activity counts are averaged into 10 minute epochs between midnight and midnight. Similarly, coefficient functions and residuals are estimated on a finite grid. With this in mind, we re-express the function-on-scalar regression using matrices:

$$Y=Xβ^{T}Θ+ϵ$$

where $Y$ is the matrix constructed by row-stacking the observed functional responses over the finite grid, $X$ is the design matrix constructed by row-stacking the observed vectors of covariates (the first column is understood to be 1 for all subjects to allow the intercept), $Θ$ is the matrix containing B-spline basis functions evaluated over the finite grid as rows, $β$ is the matrix containing spline coefficient vectors as columns, and $ϵ$ is the matrix of row-stacked residual functions evaluated over the finite grid. Define, for matrices $A$ and $B$, $vec(A)$ to be the vector constructed by concatenating the columns of $A$ and $A⊗B$ to be the Kronecker product of $A$ and $B$. The matrix formulation of the function-on-scalar regression can be re-expressed as

$$vec(Y^{T})=(X⊗Θ^{T})vec(β)+vec(ϵ^{T}).$$

In this formulation the vector of unknown parameters is $vec(β)$; the response vector $vec(Y^{T})$ and design matrix $(X⊗Θ^{T})$ are observed, and standard techniques can be used for estimation of spline coefficients. In particular, two options are available:
 \* Ordinary least squares, which assumes that errors in $ϵ$ are uncorrelated and share a common error variance across all times of day, and
 \* Generalized least squares, which allows for correlation over time and also for non-constant variance. We prefer to use GLS for estimation and inference, because doing so is more accurate and produces estimates with lower variance. However, we do note that (for completely observed curves) the use of OLS will result in perfect agreement between $∫β\_{k}(t) dt=\overline{β}\_{k}$ and regression coefficients $β\_{k}$ resulting from a multiple linear regression analysis fit to the average activity count. This perhaps surprising result stems from the equivalence between the residual sum of squares minimized in both cases.

Hypothesis tests for coefficient functions $β\_{k}(t)$ can be performed as $F$-tests of the null hypothesis $H\_{0}:β\_{k}=0$. In practice, GLS requires the estimation of the residual covariance, which can introduce an additional source of uncertainty. For this reason, an alternative approach to testing is to focus on the test statistic $∫^\_{k}^{2}(t)dt$ and to derive the null distribution of the test statistic using a bootstrap procedure (Zhu et al. 2007). In our analysis, we focus on the latter approach to testing and inference.

This estimation strategy assumes that $M$ basis functions suffice to estimate the coefficient function $β\_{k}(t)$, and the constant $M$ should be chosen with some care: if $M$ is too small, the resulting estimates may "over-smooth" the coefficient functions, while if $M$ is too large, the resulting estimates may "under-smooth". Alternatives that use a large $M$ but explicitly impose smoothness constraints are available, although hypothesis tests and inference for such approaches are not yet well-established.

# References

Ramsay, J O, and B W Silverman. 2005. *Functional Data Analysis*. New York: Springer.

Reiss, P T, L Huang, and M Mennes. 2010. “Fast Function-on-Scalar Regression with Penalized Basis Expansions.” *International Journal of Biostatistics* 6: Article 28.

Zhu, H, J G Ibrahim, N Tang, D B Rowe, X Hao, R Bansal, and B S Peterson. 2007. “A Statistical Analysis of Brain Morphology Using Wild Bootstrapping.” *IEEE Transactions onMedical Imaging* 26 (7): 954–66.