Science Discovery
2020; 8(6): 134-138
http://www.sciencepublishinggroup.com/j/sd
doi: 10.11648/j.sd.20200806.13
ISSN: 2331-0642 (Print); ISSN: 2331-0650 (Online)

The Application of Partially Functional Linear Regression Model in Health Science

Weiwei Xiao, Yixuan Wang*

School of Science, North China University of Technology, Beijing, China

Email address:
wwsunny@163.com (Weiwei Xiao), 1426044457@qq.com (Yixuan Wang)
*Corresponding author

To cite this article:
Weiwei Xiao, Yixuan Wang. The Application of Partially Functional Linear Regression Model in Health Science. *Science Discovery*. Vol. 8, No. 6, 2020, pp. 134-138. doi: 10.11648/j.sd.20200806.13

Received: September 30, 2020; **Accepted**: October 28, 2020; **Published**: November 4, 2020

Abstract: With the rapid development of information technology, data information also presents the Characteristics of diversity. Meanwhile more and more datum are presented in the form of functions. Therefore, functional data has become the focus of researchers. Functional data analysis has also proved to be of great value in the fields of biology, medicine and metrology. A partially functional linear regression model is proposed for the regression cases in which the response variables are scalar types and the predictive variables are both variable types and functional types. For the functional predictive variables, we use the functional principal component analysis method to reduce the dimension of the functional data.The least square method is used to calculate the estimate of parameters.With the improvement of people's living standard, people pay more and more attention to health. And an increasing number of people are eager to live a healthy life and keep healthy. Healthy and comfortable sleep has become a topic of increasing concern to researchers. Using data from PhysioNet Databases on activity and sleep in healthy people for this study, we found that the predicted variables in the model could well explain the response variables. The application of partially functional linear model is further extended.

Keywords: Functional Data, Partially Functional Linear Regression, Functional Principal Component Analysis, Health Science

部分函数型线性回归模型在健康科学中的应用

肖维维，王艺璇*

北方工业大学理学院，北京市，中国

邮箱
wwsunny@163.com（肖维维），1426044457@qq.com（王艺璇）

摘要：随着信息技术的迅速发展，数据信息也呈现出多元化的特点，越来越多的数据以函数的形式呈现出来。因此，函数型数据成为广大研究者们关注的焦点。函数型数据分析也被证实在生物学、医学、计量学等领域有很大的应用价值。针对响应变量是标量型、预测变量既有变量型又有函数型的回归情形，提出了一种部分函数型线性回归模型。针对函数型预测变量，我们采用函数型主成分分析法，对函数型数据进行降维；并采用最小二乘法求得参数估计。随着人民生活水平的提高，人们对于健康的重视不断提升，越来越多的人迫切地想得到健康的生活，保持身体的健康状态。健康与舒适的睡眠日益成为广大研究者关注的话题。我们通过应用PhysioNet Databases中提供的有关健康人活动与睡眠的数据进行了研究，研究结果可以看出此模型中的预测变量可以很好的解释响应变量。部分函数型线性模型的应用得到了进一步的推广。

关键词：函数型数据，部分函数型线性回归，函数型主成分分析，健康科学

1. 引言

随着信息技术的发展，越来越多的数据均呈现出曲线或图像的形式。面对一些高纬度的数据，传统的多元回归分析已经不能解决，早在1982年Ramsay就提出了函数型数据这一概念[1]。函数型数据是指定义在紧区间上的实值函数，换言之，以函数形式呈现出来的数据。图1给出了2017年中国郑州、天津、唐山等49个城市的日平均气温变化情况，就属于函数型数据。我们可以看出函数型数据的一个特点：函数型数据是无限维的。

自从Ramsay和Dalzell提出函数型数据分析以来[2]，函数型数据分析一直备受广大研究者的欢迎，主成分分析、聚类分析、自协方差函数等很多统计方法都被应用到函数型数据分析中，函数型数据分析在医学、经济学、生物学及计算机等各个领域都有所发展。

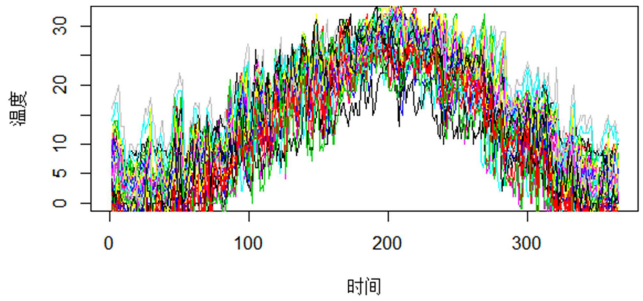

图1 中国50个城市气温数据。

由于函数型数据是无限维的，与标量型数据不同，这两种类型的数据可以分别作为预测变量和响应变量出现在归回模型中，因此，导致了函数型数据回归模型种类众多。在实际应用中，预测变量既有函数型数据又有标量型数据的模型最为居多，我们将这一类线性模型称为部分函数型线性回归模型。部分函数型线性回归模型首先由Zhang et al.提出，主要是研究标量的响应变量和包括向量值和一个函数型的混合协变量之间的线性关系[3]。Shin提出一个基于函数型主成分分析的估计方法并且研究了它的渐近性质[4]。Shin and Lee分别采用函数型主成分分析法和Tikhonov正则化方法研究了预测问题[5]。Yu et al.考虑了参数部分的线性假设检验[6]，他们成功地把广义似然比检验方法推广到部分函数型线性回归模型中[7]。Lu et al.考虑了部分函数型线性分位数回归模型的估计[8]。Li et al.研究了响应变量非忽略性缺失下的部分函数型线性回归[9]。

随着生活生平的提高，人们越来越重视身心健康。人体又是一个复杂的系统，系统各因素之间影响十分密切。本论文旨在研究心率、睡眠前后唾液中皮质醇水平差、体质指数、行为回避/抑制指数对睡眠质量的影响。提出了部分函数型线性模型的概念，通过主成分分析和最小二乘法求得了该模型中的参数系数估计，最终将该模型应用到了健康人睡眠质量的研究。

2. 部分函数型线性模型

第 i 个样本或观测点的数据为 $\left\{\left(X_{i1}\left(t_1\right),t_1\in T_1\right),\left(X_{i2}\left(t_2\right),t_2\in T_2\right),\cdots,\left(X_{id}\left(t_d\right),t_d\in T_d\right),Z_i,Y_i\right\}$ ，$i=1\cdots n$ ，并假设这些数据均是独立同分布的。其中，函数型预测变量 $X_{ij}\left(t\right)$ 是平方可积的随机过程；非函数型预测变量 $Z_i=\left(Z_{i1},Z_{i2},\cdots Z_{iq}\right)^T$ 是 q 维向量；响应变量 Y_i 是标量型数据。我们假设 Y 与 $\left(X_j\left(t_j\right),Z\right),j=1\cdots d$ 存在以下关系：

$$Y=\alpha+\sum_{j=1}^{d}\int_{T_j}X_j\left(t_j\right)\beta_j\left(t_j\right)dt_j+Z^T\gamma+\varepsilon \tag{1}$$

且满足，

$$\mathrm{E}\left(\varepsilon\mid X_j\left(t_j\right),Z\right)=0$$

$$\mathrm{Var}\left(\varepsilon\mid X_j\left(t_j\right),Z\right)=\sigma^2$$

其中，α 是未知截距，是一个常数，$\beta_j\left(\bullet\right)$ 是平方可积的未知斜率函数，$\gamma=\left(\gamma_1,\gamma_2,\cdots,\gamma_q\right)^T$ 是未知的参数向量。

3. 模型估计

3.1. 主成分分析

为了解决函数型数据无限维的问题，我们采用了函数型主成分分析法进行降维[10]。令 $\phi_k\left(\bullet\right),k=1,2,\cdots$ 是函数空间上的一组标准正交基，因此有 $\int_T\phi_k^2\left(t\right)dt=1$ 。函数型预测变量 $X_{ij}\left(t_j\right)$ 和参数函数 $\beta_j\left(t_j\right)$ 可以展开为：

$$X_{ij}\left(t_j\right)=\sum_{k=1}^{\infty}\xi_{ijk}\phi_k\left(t_j\right)$$

$$\beta_j\left(t_j\right)=\sum_{k=1}^{\infty}\beta_{jk}\phi_k\left(t_j\right)$$

那么有

$$\xi_{ijk}=\int_{T_j}X_{ij}\left(t_j\right)\phi_k\left(t_j\right)dt_j$$

$$\beta_{jk}=\int_{T_j}\beta_j\left(t_j\right)\phi_k\left(t_j\right)dt_j$$

其中，ξ_{ijk} 成为函数型主成分得分，并满足 $\mathrm{E}\left(\xi_{ijk}\right)=0$ ；$\sum\beta_{jk}^2<\infty$ 。

可以得到模型（1）的另一个表达式：

$$Y_i=\alpha+\sum_{j=1}^{d}\sum_{k=1}^{\infty}\xi_{ijk}\beta_{jk}+Z_i^T\gamma+\varepsilon_i,i=1,2,\cdots,n \quad (2)$$

为了解决函数型预测变量无限维所带来的的困难，我们将预测变量在 p 处截断，且维数 p 随着 $n\to\infty$ 而渐进增加[11]。

模型（2）可进一步写成：

$$Y_i=\alpha+\sum_{j=1}^{d}\sum_{k=1}^{p}\xi_{ijk}\beta_{jk}+Z_i^T\gamma+\varepsilon_i,i=1,2,\cdots,n \quad (3)$$

3.2．最小二乘法

我们定义一个最小二乘法准则[12][13]：

$$L_n(\alpha,\beta,\gamma)=\sum_{i=1}^{n}\left(Y_i-\alpha-\sum_{j=1}^{d}\sum_{k=1}^{p}\xi_{ijk}\beta_{jk}-Z_i^T\gamma\right)^2 \quad (4)$$

定义 $(\hat{\alpha},\hat{\beta},\hat{\gamma})=\arg\min_{\alpha,\beta,\gamma}L_n(\alpha,\beta,\gamma)$，其中，$\hat{\beta}=\left(\hat{\beta}_1^T,\hat{\beta}_2^T,\cdots,\hat{\beta}_3^T\right)^T$ 和 $\hat{\beta}_j=\left(\hat{\beta}_{j1},\hat{\beta}_{j2},\cdots,\hat{\beta}_{jp}\right)^T$，所以 $\hat{\alpha}$ 是 α 的一个估计，$\hat{\gamma}$ 是 γ 的一个估计，并且将 $\beta_j(t)$ 的估计定义为

$$\hat{\beta}_j(t)=\sum_{k=1}^{p}\hat{\beta}_{jk}\hat{\phi}_k(t)$$

令 $\hat{U}_{ij}=\left(\hat{\xi}_{ij1},\cdots,\hat{\xi}_{ijp}\right)^T$，$\hat{U}_i=\left(\hat{U}_{i1}^T,\cdots,\hat{U}_{id}^T\right)^T$ 和 $\hat{U}=\left(\hat{U}_1,\cdots,\hat{U}_n\right)^T$，$Z=(Z_0,Z_1,\cdots,Z_n)$，其中，$Z_0=(1,1,\cdots,1)^T$ 因此 $\alpha=\gamma_0$。则（4）式可以写成矩阵形式：

$$L_n(\alpha,\beta,\gamma)=\left(Y-Z^T\gamma-\hat{U}\beta\right)^T\left(Y-Z^T\gamma-\hat{U}\beta\right)$$

$L_n(\alpha,\beta,\gamma)$ 的最小解的表达式为

$$\begin{cases}\hat{\gamma}=\left\{Z\left(I-P_{\hat{U}}\right)Z^T\right\}^{-1}Z\left(I-P_{\hat{U}}\right)Y\\ \hat{\beta}=\left(\hat{U}^T\hat{U}\right)^{-1}\hat{U}^T\left(Y-Z\hat{\gamma}\right)\end{cases}$$

其中 $P_{\hat{U}}=\hat{U}\left(\hat{U}^T\hat{U}\right)^{-1}\hat{U}^T$，$I$ 是 $n\times n$ 的单位矩阵。

4．实例研究

4.1．数据来源说明

本部分使用PhysioNet Databases中提供的有关健康人活动与睡眠的数据。数据是由BioBeats（biobeats.com）与比萨大学的研究人员合作收集并提供的。

健康人活动和睡眠的多级监控（MMASH）数据集提供24小时连续的逐次跳动心脏数据，三维加速度计数据，睡眠质量，身体活动和心理特征（例如焦虑状态，压力事件和情感），吸引22位健康参与者。此外，该数据集中还提供了唾液生物标志物（皮质醇和褪黑激素）和活性日志。MMASH数据集将使研究人员能够测试身体活动，睡眠质量和心理特征之间的相关性。

招募了22名健康的年轻成年男性。在开始之前，参加者签署了知情同意书以参加本研究。根据《通用数据保护条例：欧洲议会和欧盟理事会EU 2016/679法规》（2016年4月27日），该指南提供了有关研究方案，可能的风险和数据使用的信息，有关保护个人隐私的规定如下：关于个人数据的处理以及此类数据的自由移动。在按照赫尔辛基宣言修订在2013年，研究批准了比萨大学（#0077455/2018）的伦理委员会。

在数据记录开始时，记录参与者的拟人特征（即年龄，身高和体重）。同时，参与者填写了一组初始调查表，这些调查表提供了有关参与者心理状态的信息：早晨-晚上调查问卷（MEQ），状态-特质焦虑量表（STAI-Y），匹兹堡睡眠质量问卷指数（PSQI）和行为避免/抑制（BIS / BAS）。在测试过程中，参与者连续24小时佩戴了两种设备：心率监测仪（Polar H7心率监测仪-Polar Electro Inc.，美国纽约州贝斯佩奇）记录心跳和心跳间隔，以及活动记录仪（ActiGraph wGT3X-BT-ActiGraph LLC，美国佛罗里达州彭萨科拉）以记录诸如加速度计数据，睡眠质量和体育锻炼之类的书法信息。此外，在一天的不同时间（即第二天的10、14、18、22和9）记录感知的情绪（积极和消极的情绪安排-PANAS）。此外，参加者在入睡前填写了每日压力清单（DSI），以总结当天的压力事件。

一天两次（即上床睡觉前和醒来时），受试者在家中用适当的小瓶收集唾液样本。唾液样本用于提取RNA并测量特定时钟基因的诱导，并评估特定激素。研究参与者需要至少一周的药物洗脱期。

4.2．实例应用

我们旨在研究志愿者的每小时心率、睡眠前后唾液中皮质醇水平差、体质指数（体重(kg)/身高$^2(m)$）、行为回避/抑制指数等因素对睡眠质量的影响[14]。其中，行为回避/抑制指数（BIS/BAS）包含了4个子分量指标，即对厌恶事件的敏感性（Bis）、对奖励的响应性（Reward）、个人的持久性和激励力度（Drive）、寻求刺激的冲动性（Fun）。

本次研究的响应变量是志愿者的匹兹堡睡眠质量调查问卷指数[15]；标量型数据为睡眠前后唾液中的皮质醇水平差、体质指数、行为回避/抑制指数；响应变量为志愿者一天24小时的心率变化。图2给出了某志愿者一天24小时的每分钟心率变化。将数据代入我们的模型中，即可得到函数型参数 $\hat{\beta}$ 和非函数型参数 $\hat{\gamma}$。研究结果见表1和图3。

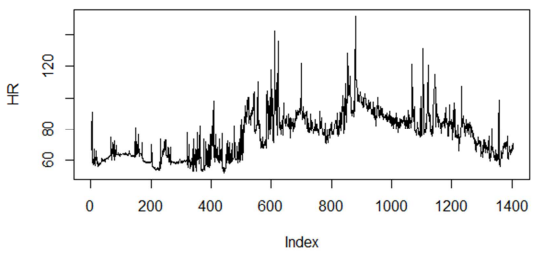

图2 某志愿者每分钟心率变化图。

表1 表格说明应为表的相应说明性文字。

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	-13.74223	13.11028	-1.048	0.3715	
$\hat{\beta}_1$	0.47681	0.31339	1.521	0.2255	
$\hat{\beta}_2$	-0.10631	0.09969	-1.066	0.3644	
$\hat{\beta}_3$	0.30778	0.14241	2.161	0.1194	
$\hat{\beta}_4$	-0.37000	0.22588	-1.638	0.1999	
$\hat{\beta}_5$	-0.23609	0.06148	-3.840	0.0311	*
$\hat{\beta}_6$	0.08040	0.07167	1.122	0.3436	
$\hat{\beta}_7$	0.08272	0.08948	0.924	0.4234	
$\hat{\beta}_8$	-0.42945	0.17103	-2.511	0.0869	.
$\hat{\beta}_9$	0.74572	0.26499	2.814	0.0671	.
$\hat{\beta}_{10}$	-0.34513	0.20328	-1.698	0.1881	
$\hat{\beta}_{11}$	0.19288	0.07918	2.436	0.0928	.
$\hat{\gamma}_{皮质醇}$	46.80798	19.32102	2.423	0.0939	.
$\hat{\gamma}_{体质指数}$	0.67268	0.22606	2.976	0.0588	.
$\hat{\gamma}_{BIS/BAS\_Bis}$	-1.02237	0.23747	-4.305	0.0231	*
$\hat{\gamma}_{BIS/BAS\_Reward}$	0.28971	0.15276	1.897	0.1542	
$\hat{\gamma}_{BIS/BAS\_Drive}$	-0.52872	0.24916	-2.122	0.1239	
$\hat{\gamma}_{BIS/BAS\_Fun}$	-0.74234	0.48433	-1.533	0.2229	

R-sq.(adj) = 0.9655 Deviance explained =87.03%。

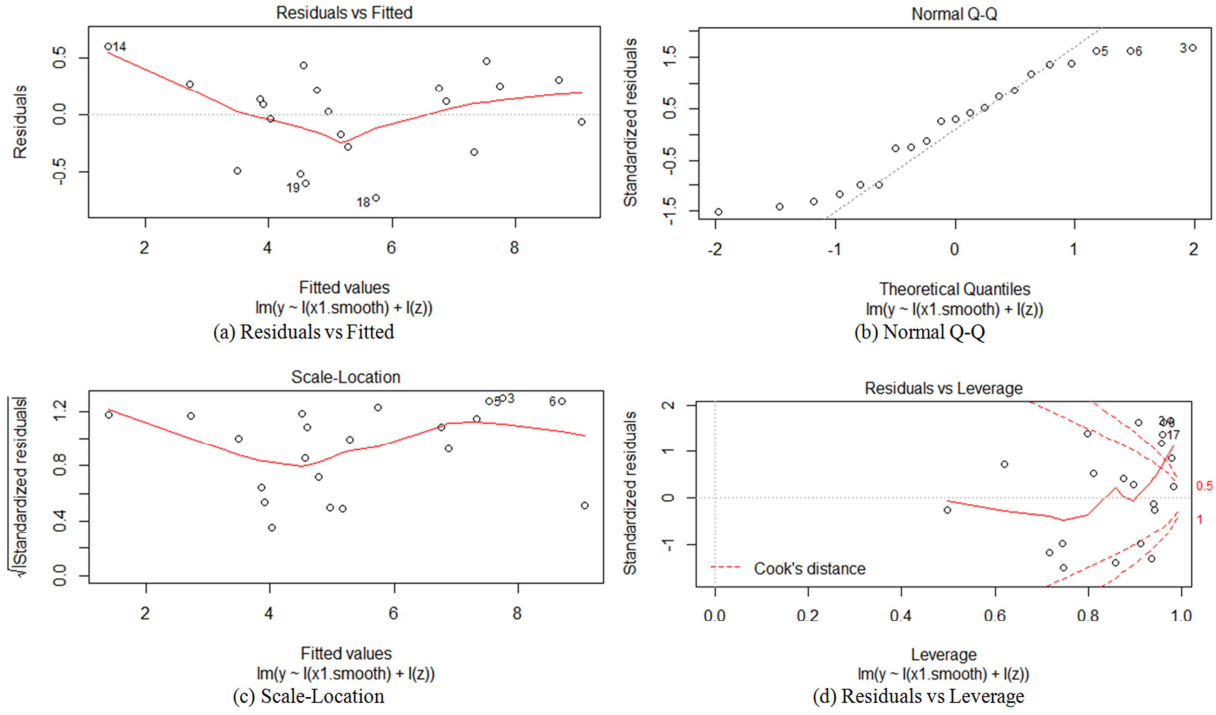

图3 残差图。

由表1可以看出，心率对睡眠质量是有一定影响的，可以粗略地看出，在晚上心率与睡眠质量是负相关的，在白天心率与睡眠质量是正相关的，即晚上心率与差的睡眠质量有关，白天心率与好的睡眠质量有关；行为回避/抑制指数对睡眠质量的影响相对较大，且与睡眠质量是负相关，即回避/抑制指数对差的睡眠质量有影响；早晚皮质醇水平差和体质指数对睡眠质量有一定的影响，且与睡眠质量是正相关的，即与好的睡眠质量有关。

由图3可以看出，(a)图是残差与真实值之间的关系，可以看出残差与真实值是无关的；（b）图是检验残差的正态性，从图可以看出残差服从正态分布；（c）图是检验等方差的假设，从图可以看出方差是一个定值；（d）

图是检验数据分析中特别极端的点，可以看出第3个点和第17个点是极端点。由表1和图3可以清晰的看出此模型中的预测变量可以很好的解释响应变量。部分函数型线性模型的应用得到了进一步的推广。

5．结论

本文提出了部分函数型线性模型，即响应变量是标量型的预测变量既有向量又有函数型的混合变量之间的线性关系。并将部分函数型线性模型应用于健康科学领域。本文研究了健康成年男性的睡眠质量。通过显著性检验，我们得出结论：心率、睡眠前后唾液中皮质醇水平差、体质指数、行为回避/抑制指数对睡眠质量都有不同显著性的影响。心率在晚上与差的睡眠质量有关，在白天与好的睡眠质量有关；行为回避/抑制指数与差的睡眠质量有关；早晚皮质醇水平差和体质指数与好的睡眠质量有关。为了提高我们的睡眠质量，我们建议定期体检，了解自己的身体状态；同时加强体育锻炼，增强体质，保障体内激素的正常分泌；更重要的是关注心理健康，提高自我调节能力，必要时及时就医。因此，身体健康与心理健康对人们睡眠质量都有很大的影响。我们应该倡导健康生活习惯的同时，也要注意自身的心理调节，身体与心理同时健康才是真正的健康。

致谢

本文为北方工业大学人才专项资助(207051360020XN140/004)的阶段性成果之一。

参考文献

[1] RAMSAY J O, When the data are functions [J]. Psychometriha, 1982, 47(4):379-396.

[2] Ramsay J O, Dalzell C J. Some tools for functional data analysis [J]. J Roy Statist Soc Ser B, 1991,53(3): 539-572.

[3] ZHANG D W, LIN X H, SOWERS M. Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome [J].Biometrics, 2007, 63(2):351-362.

[4] SHIN H. Partial functional linear regression [J].Journal of Statistical Planning and inference, 2009, 139(10):3405-3418.

[5] SHIN H, LEE M H. On prediction rate in partial functional linear regression [J]. J Multivariate Anal, 2012, 103(1):93-106.

[6] YU P, ZHANG Z Z, DU J. A test of linearity in partial functional linear regression[J].Metriha, 2016, 79(8):953-969.

[7] FAN J Q, ZHANG C M, ZHANG, J. Generalized likelihood ratio statistics and wilkes phenomenon [J].The Aririals of Statistics, 2001, 29(1):153-193.

[8] LU Y, DU J, SUN 2 M. Functional partially linear quantile regression model [J].Metriha, 2014, 77(2):317-332.

[9] LI T F, XIE F C, FEND X N, et al. Functional linear regression model for nonig-norable missing scalar responses[J].Statistics Sinica, 2018, 28(4):1867-1886.

[10] HSING T, EUBANK R L. Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators [M].John Wiley & Sons, 2015.

[11] H.-G. MÜLLER, U. STADTMÜLLER. Generalized Functional Linear Models [J]. The Annals of Statistics, 2005, 33(2):774–805.

[12] CAI T T, HALL P. Prediction in functional linear regression [J]. The Annals of Statistics, 2006, 34(5):2159-2179.

[13] HALL P, HOOKWITZ J L. Methodology and convergence rates for functional linear regression [J]. The Annals of Statistics, 2007,35(1):70-91.

[14] Carver CS, White TL. "Behavioural inhibition, behavioural activation, and affective responses to impending reward and punishment: The BIS/BAS Scales". J Pers Soc Psychol. 1994,67: 319–333.

[15] Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. "The Pittsburgh Sleep Quality Index: A New Instrument for Psychiatric Practice and Research". Psychiatry Res. 1989,28: 193-213.