缺失数据情况下的血压预测：使用 NHANES 分析缺失数据模式和插补方法

Harish Chauhan*，**Nikunj Gupta***，**Zoe Haskell-Craig***

纽约大学

arXiv:2305.01655v1 [cs.LG] 1 May 2023

动机

世界卫生组织将心血管疾病（CVD）定义为“心脏和血管的一组疾病”，包括冠心病和卒中。（世界卫生组织，2021年）心血管疾病受到血压升高、血糖升高、血脂升高和肥胖等“中间风险因素”的影响。这些因素主要受生活方式和行为影响，包括缺乏身体活动、不健康饮食、盐摄入过量以及烟草和酒精使用。然而，遗传因素以及贫困、压力和种族主义等社会和环境因素也发挥着重要作用。研究与这些“中间风险因素”相关的行为和环境因素的研究人员，需要获取关于饮食和身体活动的高质量、详尽信息。

然而，缺失数据是临床和公共卫生研究中普遍存在的问题，会影响随机试验和观察性研究。由于失访、错过研究访视、拒绝回答调查问题，或门诊就诊期间未记录测量值，各项研究中缺失数据的原因可能有很大差异。处理缺失值的一种方法是直接删除存在缺失的观测值（称为完整病例分析）。这种方法很少使用，因为删除包含缺失数据的数据点（逐案删除）会减少样本数量，进而影响准确性。处理缺失数据还有其他方法，例如根据变量的已观测值汇总变量（可用病例分析）。

鉴于NHANES数据集中缺失数据普遍存在，我们将分析在不同模拟缺失数据模式下的插补方法。随后，我们将应用这些插补方法创建一个完整数据集，并以此使用普通最小二乘法，根据饮食和身体活动预测血压。

方法

Data

我们使用了2013—2014年美国国家健康与营养调查（NHANES）的数据。该研究对美国具有全国代表性的总体样本进行了调查。研究人员对各年龄段的参与者进行了健康访谈和调查问卷，同时开展了实验室测试和体格检查。（美国疾病控制与预防中心与美国国家卫生统计中心，2017年）收集的数据包括人口统计数据（年龄、性别、种族与族裔、教育程度、公民身份、家庭人数和收入）、身体比例和体重、脉搏和血压、糖尿病病史、饮食习惯（例如使用食盐的情况）、热量和宏量营养素摄入量（如膳食纤维、脂肪等）、身体活动的类型和持续时间，以及血胆固醇水平。

2013—2014年的最终样本量为 10 175。在将样本限定为年龄不低于 20 岁的参与者后（因为有些测试仅针对成年人进行），我们获得了 n = 5769 个观测值。进一步排除第 1 天收缩压数据缺失的参与者后，最终得到 5111 行数据。高水平的低密度脂蛋白（LDL）胆固醇对心血管疾病尤其危险，因为它可能在血管中积聚并造成阻塞。我们将血液中 LDL 胆固醇超过 160 mg/dL（毫克/分升）的参与者归类为 LDL 偏高。（Goldman, R & Clark, C. 2021年）

通过考察变量之间的两两皮尔逊相关系数进行探索性数据分析。此外，还绘制了缺失模式。

缺失数据机制

我们将分析限定在 16 个变量（包括饮食列 DR1TKCAL，见表 1）以及每个变量均有观测数据的参与者（无缺失数据），并据此模拟完全随机缺失（MCAR）、随机缺失（MAR）和非随机缺失（MNAR）模式。

完全随机缺失（MCAR）如果某个变量的缺失概率对所有单位都相同，也就是说，缺失与所有已观测和未观测变量都无关，那么该变量就是完全随机缺失。（Mack C 2018）如果数据是完全随机缺失的，那么剔除缺失数据的案例不会使推断产生偏差。我们通过从样本中随机移除观测值来模拟完全随机缺失。

随机缺失（MAR）从数据本身可以看出，大多数缺失并非完全随机，正如数据本身所示——

*这些作者对本研究的贡献相同。

变量名称	描述
RIDAGEYR	Age
RIAGENDR	性别
RIDRETH1[3]	种族/族裔
DMDCITZN	公民身份
BMXLEG	腿围
BPXPULS	脉搏
DIQ010	糖尿病诊断
DIQ050	胰岛素使用
HIQ011	医疗保险
PAQ635	步行或骑自行车
PAQ650	剧烈运动
PAQ665	中等强度运动
PAD680	久坐分钟数
PAQ710	电视使用
DR1TKCAL	热量摄入

表 1：用于缺失数据分析的数据子集中包含的变量。

如此。更一般的假设是随机缺失，即变量缺失的概率仅取决于可用信息。也就是说，缺失与已观测到的变量有关，而与未观测到的变量无关。（Mack C 2018）因此，如果调查中所有受访者的性别、年龄、BMI 和饮食信息都有记录，那么当对该问题不作答的概率仅取决于这些其他已完整记录的变量时，热量摄入便是随机缺失的。我们通过从饮食列 DR1TKCAL 中删除数据来模拟随机缺失数据：如果受访者在身体活动问题上的回答值为 2，则删除数据的概率加倍。

非随机缺失 (MNAR) 如果缺失取决于未被记录的信息，而这些信息同时也能预测缺失值，那么这种缺失就不再是“随机”的。例如，假设“脾气暴躁”的人较少回答有关“饮食”的问题，脾气暴躁与血压相关，且“脾气暴躁”这一特征未被观测到。那么，血压值就是非随机缺失的。医学研究中一个常见的例子是：如果某种治疗会引起不适，患者更有可能退出研究。这种缺失并非随机缺失（除非对所有患者都测量并观测了“不适”）。如果缺失并非随机缺失，就必须对其进行明确建模，否则就必须接受推断中存在一定偏倚。我们模拟非随机缺失时，若受访者的 BMI 大于 25（即超重的定义），就以两倍的概率从饮食列中移除数据。请注意，虽然较大的数据集中包含 BMI，但我们没有在缺失模拟样本中纳入这一变量，因此它相当于一个未观测变量。

插补方法

我们为本次分析选择了以下插补方法。

单变量特征插补方法采用均值、中位数或该列中最常见的值来填补缺失数据。这种策略可能严重扭曲该变量的分布，导致汇总指标出现问题，尤其会低估标准差。此外，均值插补会通过将相关性估计值“拉向”零来扭曲变量之间的关系。

最近邻插补通过查找集合中与目标样本“最接近”的样本，并对这些邻近点取平均来填补缺失值。与单变量特征插补相比，这种方法确有优势，但计算开销较大，而且对离群值相当敏感。

低秩近似是使用奇异值分解（SVD）将矩阵近似为秩低于原矩阵的矩阵。该技术通过迭代，利用完整的低秩矩阵估算缺失观测值，然后重新计算 SVD。我们在训练数据集上进行了 SVD，并选择使测试集误差最小的秩。

多变量特征插补是一种更复杂的方法。它将每个含缺失值的特征建模为其他特征的函数，并使用估算结果进行插补。该方法以轮询方式迭代执行：每一步都将一个特征列指定为输出 Y，并将其他特征列作为输入 X。对于 Y 已知的样本，使用 (X,Y) 拟合一个回归器。随后，利用该回归器预测 Y 的缺失值。针对每个特征，以迭代方式重复这一过程。

评估：均方误差我们使用均方误差（MSE）评估插补方法的效果，将插补值与原始子样本中的值进行比较。

预测

对插补数据进行 OLS 回归 为了解生活方式与血压之间的关联（这是心血管疾病的一个重要“中间”风险因素），我们将普通最小二乘法（OLS）回归与插补方法结合起来，应用于完整数据集。OLS 是一种用于估计线性回归模型未知参数的线性最小二乘法。所得估计量可以用一个简单公式表示，尤其是在简单线性回归的情况下，此时回归方程右侧只有一个回归器。当观测值数量远多于特征数量（如本研究所示）时，如果误差同方差且序列不相关，OLS 回归在线性无偏估计量中具有最优性。当误差具有有限方差时，OLS 可提供最小方差的均值无偏估计。在这些条件下，OLS 的表现显著优于套索回归和岭回归（使用 $\lambda > 0$）。

我们首先应用六种插补方法，创建一个不含缺失观测值的完整数据集。随后按 70/30 的比例将数据划分为训练集和测试集。我们使用训练集训练普通最小二乘法（OLS）模型，再将该模型应用于测试集数据，以预测血压。我们使用均值

平方误差（均方误差，MSE）和均方根误差（RMSE）评估各种插补方法的性能。

结果与讨论

探索性数据分析： 相关性图显示，大多数变量之间的相关性从中等到较低，但膳食摄入变量、两个不同时间点测量的舒张压和收缩压，以及腿围、臂围和腰围的测量值除外（图1）。膳食变量同时出现缺失，心血管疾病病史问卷数据也同时出现缺失。通过卡方检验，我们确定膳食数据缺失与健康保险状态和性别相互独立。膳食数据缺失与身体锻炼状态（变量 PAQ650）并非相互独立（图2）。

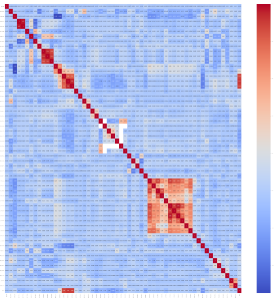


图1：特征之间的相关性

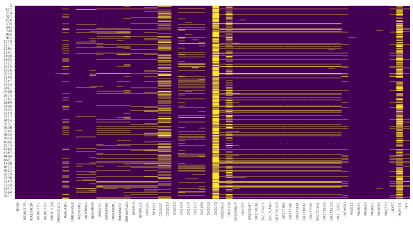


图2：NHANES数据集中的缺失数据

不同缺失模式下插补方法的均方误差： 我们考虑了三种常见的“缺失机制”：完全随机缺失、随机缺失和非随机缺失。我们计算插补后的缺失值与相应真实值之间的均方误差（MSE）。

均方误差			
插补方法	MCAR	MAR	MNAR
Mean	0.831	1.149	1.045
中位数	0.910	1.262	1.070
众数插补	1.157	1.484	1.188
KNN	0.861	1.473	0.926
低秩模型	1.021	1.060	1.010
回归	**0.560**	**1.013**	**0.853**

表2：模拟数据上的插补方法分析

表2总结了我们对模拟缺失数据上各种插补方法表现的研究结果。我们的研究表明，多变量回归特征插补优于其他所有插补方法，且不受缺失模式的影响。除回归外，当数据为随机缺失或非随机缺失时，表现次佳的是低秩近似插补。然而，当数据为完全随机缺失时，均值插补的表现位居第二，与K近邻的表现相差无几。

均方根误差		
插补方法	训练集	Test
Mean	14.516	14.783
中位数	14.776	**14.227**
众数插补	14.573	14.619
KNN	14.619	14.642
低秩模型	18.232	17.854
回归	**14.361**	15.088

表3：使用插补数据集进行血压回归

预测结果： 表3总结了使用完整数据集进行线性回归以预测血压所得的结果。我们发现，使用多变量特征插补处理缺失数据时，模型表现最佳。令人意外的是，使用各列中位数进行插补后，模型在测试集上的表现最好。

总结： 这一意外结果可能有多种原因。首先，我们仅使用了特征和观测值的一个小子集来构建模拟分析的数据集。完整数据集中的数据分布以及特征之间的关系可能大不相同，因此，不同插补方法在完整数据集和子集上的表现可能存在差异。其次，回归插补在完整数据集中的表现不佳，可能是因为完整数据集中各特征之间的相关性有限。如果无法根据一个变量预测另一个变量，那么回归插补很可能表现不佳。最后，使用各列的中位数填补缺失值，本质上会使数据更加平滑。普通最小二乘法模型在特征更加“平滑”时，可能具有更好的预测效果。

结论

总而言之，我们发现，在该数据集模拟的三种缺失模式中，多元回归插补的表现都最佳。我们能够根据饮食和身体活动，以合理的均方根误差预测血压。使用列中位数插补缺失值所得到的预测模型，其均方根误差最佳。

本研究有几项优势。首先，我们评估了插补方法在不同缺失模式下的表现，而不是简单地假设所有数据都属于完全随机缺失（MCAR）。其次，我们考察了不同插补方法对插补值准确性以及血压预测能力的影响。本研究也有几项局限。我们在模拟研究中只考察了部分特征和观测值。此外，我们只在一列（热量摄入）中设置了缺失。未来研究应模拟不同的缺失值频率，以及其他特征中的缺失模式。

参考文献

美国疾病控制与预防中心和国家卫生统计中心。2017年。《国家健康与营养检查调查：概览》。https://www.cdc.gov/nchs/data/nhanes/nhanes 13 14/ NHANES Overview Brochure.pdf。访问日期：2022年3月2日。

Goldman, R 和 Clark, C。2021年。按年龄划分的推荐胆固醇水平。https://www.healthline.com/health/high-cholesterol/levels-by-age# children。访问日期：2022年3月2日。Mack C, W. D.、Su Z。2018年。《患者登记系统中的缺失数据管理：患者结局评估登记系统：用户指南》补编。载于缺失数据类型。马里兰州罗克维尔：医疗保健研究与质量局。世界卫生组织。2021年。心血管疾病（CVD）。

h

t

t

p

s://

w

w

w.

who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)。访问日期：2022年3月2日。