گزیده‌مطالب آماری
سال ۱۹، شماره‌ی ۱، بهار و تابستان ۱۳۸۷، صص ۱۴۴-۱۱۹

جبران داده‌های گم‌شده در مطالعات طولی با استفاده از WinBUGS

گریتچن کاریگان، آدریان گ. بارنت، آنتی ج. داسون، گیتا میشرا[†]

دانشگاه کوییزلند

مترجمان: لیلی تاپاک،[۱،*] امید حمیدی[۲]

[۱] مؤسسه‌ی آموزش عالی آبادانی و توسعه‌ی روستاها
[۲] دانشگاه پیام نور

چکیده. داده‌های گم‌شده مشکلی رایج در پژوهش‌های مبتنی بر آمارگیری است. بسته‌های بسیاری وجود دارند که داده‌های گم‌شده را جبران می‌کنند اما تعداد کمی می‌توانند به‌راحتی اطلاعات طولی گم‌شده را جبران کنند. WinBUGS داده‌های گم‌شده را با استفاده از جانهی چندگانه جبران می‌کند و قادر است ساختار طولی را با استفاده از اثرات تصادفی یکپارچه کند. ما برتری جانهی طولی بر جانهی مقطعی را با استفاده از WinBUGS نشان می‌دهیم. ما از اطلاعات مثالی مربوط به مطالعه‌ی طولی استرالیا درباره‌ی سلامت زنان استفاده می‌کنیم. همچنین نتایج نرم افزار SAS را ارائه می‌دهیم که از WinBUGS برای تحلیل مدل‌های طولی با اطلاعات گم‌شده‌ی متغیر کمکی استفاده می‌کند و استفاده از آن را در مطالعه‌ی طولی بیماران سرطانی لاعلاج و پرستاران آن‌ها نشان می‌دهیم.

[†] Carrigan, G.; Barnett, A.G.; Dobson, A.J.; Mishra, G. (2007). Compensating for Missing Data from Longitudinal Studies Using WinBUGS. *Journal of Statistical Software*, **19**, 1-17.

واژگان کلیدی: جانهی چندگانه؛ داده‌های طولی؛ داده‌های گم‌شده؛ SAS؛ WinBUGS.
دریافت: ۱۳۸۷/۳/۶ پذیرش: ۱۳۸۷/۱۰/۱

[*] عهده‌دار مکاتبات

۱- مقدمه

داده‌های گم‌شده مشکل عمومی در تحقیقات مبتنی بر آمارگیری است. نادیده گرفتن هر گونه داده‌های گم‌شده با استفاده از یک تحلیل موردی کامل می‌تواند نتایجی اریب را ایجاد کند. اریبی هنگامی اتفاق می‌افتد که شرکت‌کنندگان دارای داده‌های کامل، به‌طور سیستماتیک با شرکت‌کنندگان دارای داده‌های گم‌شده تفاوت داشته باشند. به‌خصوص مطالعات طولی مستعد چنین اریبی‌هایی هستند زیرا داده‌های گم‌شده به‌مرور زمان درنتیجه‌ی بی‌پاسخی و انصراف شرکت‌کنندگان انباشته می‌شود. یکی از روش‌های جبران داده‌های گم‌شده جانهی است. طی بیست سال گذشته انبوه نوشتگان درباره‌ی نظریه و روش‌شناسی جانهی رشد قابل ملاحظه‌ای داشته است و نرم افزارهایی نیز در راستای آن تکامل یافته است. ولی در زمینه‌ی جانهی داده‌های طولی کار نسبتاً کمی صورت گرفته است.

برای جانهی چند رهیافت نظری وجود دارد. راگوناتان [۱۵] این قبیل رهیافت‌ها را بازنگری و سه رده را شناسایی کرده است: معادلات براورد موزون، جانهی چندگانه و فرمول‌های مبتنی بر درستنمایی. ابراهیم و دیگران [۸] رهیافت تماماً بیزی را به‌عنوان رده‌ی چهارم درنظر گرفتند.

معادلات براوردگر موزون (WEE)، آمارهای ثبتی دارای داده‌های کامل را وزن‌دار می‌کنند تا موارد مشابه دارای داده‌های گم‌شده را جبران کنند. در این اواخر نوشتگان بر بهبود براورد واریانس ([۱۹] و [۲۰]) تمرکز داشته‌اند زیرا WEE، هنگامی‌که تعدیل نشده باشد، واریانس واقعی داده‌ها را کم‌براورد می‌کند. اجرای WEE در حال حاضر به جای شیوه‌های استاندارد در بسته‌های نرم‌افزار آماری اصلی، بر الگوریتم‌های مدل-ویژه و کاربر-تعریف‌شده تکیه دارد. جانهی چندگانه (MI) از شبیه‌سازی بیز برای پرکردن داده‌های گم‌شده استفاده می‌کند که از تلفیق نتایج به دست آمده از چندین مجموعه‌ی جانهی‌شده‌ی مکرر به دست می‌آیند. برای مشاهده‌ی پوشش فراگیر جانهی چندگانه به رابین [۲۲] مراجعه کنید. مدل‌های تماماً بیزی (FB) روش‌های MI را با شبیه‌سازی توأم توزیع متغیرهای دارای داده‌های گم‌شده و نیز پارامترهای نامعلوم در یک معادله‌ی رگرسیون گسترش می‌دهند. در FB مدل‌های تحلیل و جانهی کاملاً و به‌طور هم‌زمان

مشخص هستند. همچنین فنون ماکسیمم درستنمایی نیز بر مدل‌های کاملاً مشخص تکیه دارند اما از این جهت که براوردهای پارامترها با استفاده از تقریب‌های مبتنی بر درستنمایی به جای شبیه‌سازی بیزی ایجاد شده‌اند، با FB فرق دارند.

رهیافت‌های ماکسیمم درستنمایی برای جانهی اغلب در بسته‌های نرم‌افزاری اصلی، غیر قابل کنترل‌اند. اجرای روش‌ها بر فرض‌های محکم در مورد الگوهای گم‌شدگی که مکرراً در بررسی داده‌های پیچیده زیر پا گذاشته می‌شوند، تکیه دارد. در حالی که شیوه‌های MI در تعدادی بسته‌های نرم‌افزاری مانند SAS [۲۳]، Stata [۲۷]، S-Plus [۹] و R [۱۸] وجود دارند، این روش‌ها عموماً بر این فرض تکیه دارند که داده‌ها، نرمال چندمتغیره‌اند یا می‌توانند با توزیع نرمال چندمتغیره تقریب زده شوند [۲۴]. بیش‌تر کارهای اخیر درباره‌ی معادلات رگرسیونی زنجیره‌ای، منجر به افزوده شدن بسته‌هایی شده است که داده‌های رسته‌ای را نیز شامل می‌شوند: MICE در S-PLUS [۲۹]، Ice در Stata [۲۱] و IVEware برای SAS [۱۶]. ولی تألیف‌کنندگان هنوز مشکلاتی در اضافه کردن داده‌های طولی در روش‌های جانهی با استفاده از این برنامه‌ها دارند. فنون FB در جانهی طولی بسیار مناسب هستند، زیرا می‌توانند ساختار سلسله مراتبی را در فرایند مدل‌سازی ادغام کنند و مانند رگرسیون زنجیره‌ای، این قابلیت را دارند که به‌طور سیستماتیک به داده‌های رسته‌ای نیز بپردازند. بسته‌های نرم‌افزار WinBUGS [۲۶] و MLwiN [۱۷] هر دو از چارچوب FB استفاده می‌کنند. کاولس [۵] و وودورت [۳۰] هر دو شرح مختصر و مفیدی برای WinBUGS تهیه کرده‌اند، درحالی‌که کارپنتر و کنوارد [۲] وکانگدون [۳] مثال‌های مقدماتی جانهی FB با داده‌های گم‌شده را مطرح کرده‌اند. پتیت [۱۳] و کیو و همکاران [۱۴] تحلیل‌های دقیقی درزمینه‌ی داده‌های رسته‌ای گم‌شده ارائه کرده‌اند.

هدف این مقاله نشان دادن توانایی WinBUGS برای جبران داده‌های طولی گم‌شده با تمرکز خاص بر داده‌های گم‌شده‌ی متغیر کمکی است. ما این کار را با نگاه بر تحلیل طولی میزان وقوع دیابت در زنان استرالیایی انجام می‌دهیم. در بخش ۲ مثالی تشویق‌کننده از مطالعه‌ی طولی استرالیا درباره‌ی سلامت زنان را معرفی می‌کنیم. در بخش ۳ یک مدل کاملاً بیزی برای میزان وقوع دیابت بدون و با داده‌های گم‌شده‌ی متغیر کمکی را تعیین می‌کنیم. در بخش ۴ اجرای آن را در WinBUGS ارائه کرده و

نتـایج را در بخـش ۵ توصـیف مـی‌کنـیم. در بخـش ۶ یـک مـاکروی کلـی SAS (کـه WinBUGS نامیده می‌شود) برای تحلیل مدل‌های طولی با داده‌های گـم‌شـده‌ی متغیـر کمکی ارائه می‌دهیم. با یک بحث و تعدادی توصیه در بخش ۷ نتیجه‌گیری می‌کنیم.

۲- مثال تشویق‌کننده

زنانی که اضافه وزن دارند بیش‌تر در معرض خطر ابتلا به دیابـت قـرار دارنـد. ولـی تـأثیر نسبی چاقی بلندمدت‌تر و تغییرات وزنی کوتاه‌مدت در میزان وقوع دیابـت مـورد علاقـه‌ی دانشمندان است [۱۱]. مطالعه‌ی طولی استرالیا درمورد سـلامت زنـان (ALSWH) بـرای پاسخ به چنین سؤالاتی طراحی شده است زیرا بهداشت و سلامت یک نمونه‌ی نمایانگر از زنان استرالیا را در گذر زمان دنبال می‌کند [۱۰].

مطالعه‌ی ALSWH داده‌هایی از آمارگیری‌های پستی خود-گزارش‌شده را هـر۲ تـا ۳ سال جمع‌آوری می‌کند. برای این تحلیل از داده‌هایی مربوط به گروهی از زنان میـان‌سـال که در زمان آمارگیری آغازین در سال ۱۹۹۶ (S1) ۴۵ تـا ۵۰ سـاله بودنـد اسـتفاده شـده است. آمـارگیری‌هـای بعـدی در سـال ۱۹۹۸ (S2)، ۲۰۰۱ (S3) و ۲۰۰۴ (S4) رخ داده است. در آمارگیری آغازین (S1) تعداد ۱۳۷۱۶ زن با شرکت در مطالعه‌ی طـولی موافقـت کردند و تا آمارگیری چهارم (S4) تعـداد ۱۰۹۰۵ از زنـان بـاقی مانـده بودنـد. متغیرهـای کلیدی برای تحلیل میزان وقوع دیابت و وزن در زیر نشان داده شده است.

در S1 از زنان پرسیده شد که آیا هرگز دیابتی تشخیص داده شـده‌انـد. در S2 و S3 و S4 از زنان پرسیده شد که از آمارگیری قبلی تاکنون دیـابتی تـشخیص داده شـده‌انـد. بـا استفاده از این داده‌ها زنان در یکی از گروه‌های زیر طبقه‌بندی شدند: مورد موجـود در S1، مورد وقوع بین S1 و S2، مورد وقوع بین S2 و S3 و مورد وقـوع بـین S3 و S4، بـدون دیابت، یا نامعلوم.

در هر آمارگیری از زنان خواسته می‌شد تا وزن و قـد خـود را گـزارش کننـد. قـدهای خود-گزارش‌شده از سه آمارگیری اول برای به دست آوردن یک مقدار براورد شده‌ی تکی برای هر زن با متوسط‌گیری از داده‌های موجود مورد استفاده قـرار گرفـت. شـاخص جـرم بدن (BMI) برای هر زن در S1 با استفاده از وزن خود-گـزارش‌شـده (کیلـوگرم) در S1

تقسیم بر مربع قد براوردشده (متر) محاسبه شد. BMI به‌صورت زیر (طبق سازمان بهداشت جهانی [۳۱]) رسته‌بندی شد: « کمتر از وزن معمول »، یعنی کمتر از ۱۸/۵ کیلوگرم بر مترمربع ($۱۸/۵ >$)؛ « وزن سالم »، [۲۵ , ۱۸/۵) کیلوگرم بر مترمربع؛ « اضافه وزن »، [۳۰ , ۲۵) کیلوگرم بر مترمربع؛ « چاق »، [۳۵ , ۳۰) کیلوگرم بر مترمربع یا « بسیار چاق »، بزرگ‌تر یا مساوی ۳۵ کیلوگرم بر مترمربع ($۳۵ \geq$). کمتر از ۲٪ زنان به‌عنوان گروه « کمتر از وزن معمول » در S1 طبقه‌بندی شدند، بنا بر این این رسته‌ی مزبور با گروه « وزن سالم »، ترکیب شده است.

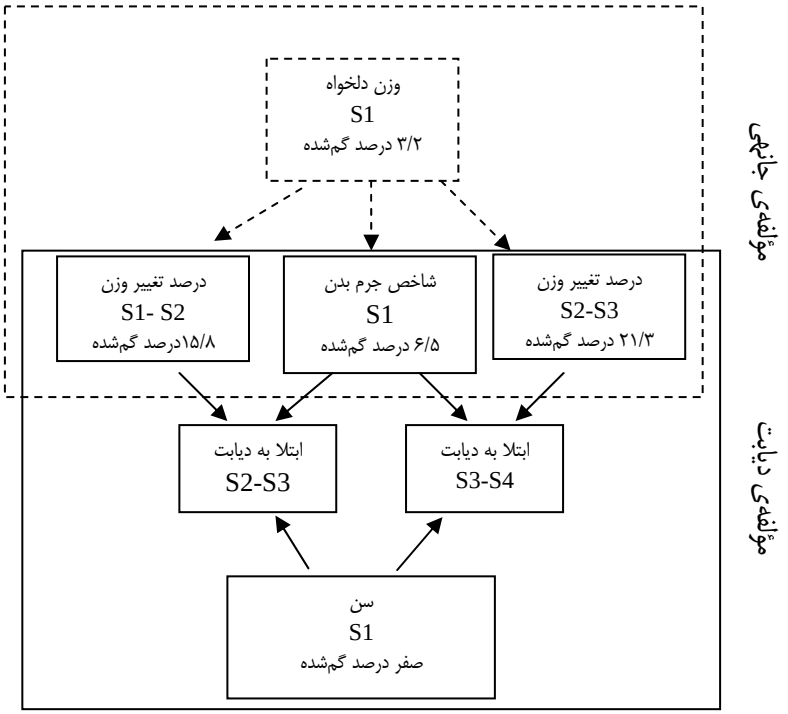
از زنان S1 پرسیده شد که دوست دارند چه وزنی داشته باشند پاسخ‌ها به این ترتیب رسته‌بندی شدند: راضی \ دوست دارند وزن بیش‌تری داشته باشند، دوست دارند بین ۰ تا ۵ کیلوگرم وزن کمتر داشته باشند، دوست دارند بین ۵ تا ۱۰ کیلوگرم وزن کمتر داشته باشند، دوست دارند بیش‌تر از ۱۰کیلوگرم وزن کمتر داشته باشند.

۳- مشخصات مدل

بررسی ارتباط بین سلامتی و وزن اغلب در بررسی داده‌ها مشکل است زیرا وزن سؤال حساسی است و گاهی گزارش نمی‌شود. برای مثال در ALSWH در S1 تعداد ۴۵۴ زن (۴٪) وزن خود را گزارش نکردند در حالی که برای دیگر متغیرهای استفاده‌شده در این گزارش درصد میانگین داده‌های گم‌شده ۱/۳٪ بود. اگر زنانی که اضافه وزن دارند کمتر تمایل به گزارش وزن خود داشته باشند در این صورت یک تحلیل موردی داده‌های کامل می‌تواند ارتباط واقعی بین وزن و میزان وقوع دیابت را بسیار کم‌براورد کند.

شکل ۱ خلاصه‌ی گرافیکی مدل ما را ارائه میدهد. مدل به مؤلفه‌های جانهی و دیابت تقسیم می‌شود. در مؤلفه‌ی دیابت ارتباط بین BMI در S1 و درصد تغییرات سالانه‌ی وزن در مورد وقوع دیابت که برای سن در خط مبنا تعدیل شده، بررسی شده است. BMI در S1 نمایانگر چاقی بلندمدت‌تر است درحالی‌که درصد تغییر وزن سالیانه نمایانگر تغییر وزن کوتاه‌مدت است. چون ما به تغییر وزن قبل از آغاز بیماری دیابت علاقه‌مند بودیم، تغییر وزن در دوره‌ی آمارگیری قبل از گزارش شدن میزان وقوع بیماری اندازه‌گیری شد (تا از مخاطره‌ی « علت معکوس » جلوگیری شود که به موجب آن زنانی که دیابتی

تشخیص داده شده بودند بعدها وزن کم می‌کردند). بنا بر این جمعیت مورد مطالعه برای تحلیل به زنانی که مورد وقوع بین S2 و S3 یا S3 وS4 بودند و یا کسانی که دیابت نداشتند محدود شد. زنانی که مورد ابتلا بین S2 و S3 بودند از تحلیل در ادامه‌ی دوره حذف شدند زیرا دیگر در جمعیت مورد مخاطره نبودند.



شکل۱- مدل ارتباط بین ابتلا به دیابت و BMI بلندمدت و تغییرات کوتاه‌مدت در وزن

همان‌طور که در شکل ۱ نشان داده شده، مقدار متغیرهای کمکی گم‌شده متفاوت است که دلایل احتمالی این‌که چرا تکمیل نشده‌اند نیز متفاوت است. رابین سه الگوی بالقوه‌ی گم‌شدگی را تعریف می‌کند: گم شدن کاملاً تصادفی (MCAR) که در آن تفاوت سیستماتیک بین خصوصیات داده‌های گم‌شده و گم‌نشده وجود ندارد؛ گم شدن تصادفی (MAR) که در آن تفاوت سیستماتیک وجود دارد اما می‌توان آن را با دیگر داده‌های

مشاهده‌شده توضیح داد و گم شدن غیرتصادفی (MNAR) که تفاوت‌ها نمی‌توانند با داده‌های مشاهده‌شده توضیح داده شوند.

ما برای ایجاد یک مؤلفه‌ی جانهی در مدل با استفاده از سؤال « شما مایل هستید چه قدر وزن داشته باشید؟ » فرض MAR را در نظر گرفته‌ایم. در مقایسه با وزن واقعی، پاسخ‌های گم‌شده به پرسش « وزن دلخواه » به‌مراتب کمتر بود. همچنین متغیر « وزن دلخواه » در S1 با میزان وزن خود–گزارش‌شده در همه‌ی آمارگیری‌ها به‌شدت همبستگی داشت [۱۲]. از این رو از « وزن دلخواه » برای جانهی وزن‌های گم‌شده در هر آمارگیری استفاده کردیم. برای هر وزن جانهی‌شده مقدار BMI در S1 و درصد تغییر وزن را دوباره حساب کردیم. زنانی که قد آن‌ها نامعلوم بود (۳/۲٪) یا آن‌هایی که به سوال « وزن دلخواه » پاسخ ندادند (۳/۲٪)، از تحلیل حذف شدند.

به‌منظور توضیح، سه مدل جداگانه طراحی کرده‌ایم؛

(i) مدل موردی کامل (۷۱۱۳ زن)؛

(ii) مدل جانهی مقطعی (۹۵۵۷ زن)؛

(iii) مدل جانهی طولی با استفاده از اثرات تصادفی برای لحاظ کردن همبستگی‌های درون–موضوعی (۹۵۵۷ زن).

مدل‌های (ii) و (iii)، هر دو مؤلفه‌ی دیابت و جانهی را داشتند (شکل۱). مدل (i) فقط مؤلفه‌ی دیابت دارد.

اکنون سه مدل را با جزئیات بیش‌تری توضیح می‌دهیم.

۳-۱- مدل موردی کامل

فرض کنیم Y_{it} یک متغیر دودویی باشد که وقوع دیابت برای فرد i ($i = ۱,\ldots,۷۱۱۳$) در زمان t که $t = ۱,۲$ را مشخص می‌کند. درنتیجه مدل موردی کامل به‌صورت زیر است.

$$Y_{it} \sim Bernoulli\,(p_{it}),$$

$$\text{logit}(p_{it}) = \alpha_t + X_i^T \beta + Z_{i\,(t-۱)}\Psi$$

که $\mathbf{X}$ ماتریس متغیرهای کمکی زمان–ناوردا (سن در S1 و BMI در S1)، Z برداری شامل تنها متغیر کمکی دارای تغییر در طول زمان (تغییر درصد وزن در دوره‌ی پیش از وقوع گزارش شده) و α عرض از مبدأ است که بر اساس زمان آمارگیری تغییر می‌کند.

۳-۲- مدل جانهی مقطعی

مؤلفه‌ی دیابت این مدل همان ساختار مدل موردی کامل را دنبال می‌کند. در مؤلفه‌ی جانهی این مدل، فرض کرده‌ایم که وزن فرد i ($i = ۱,\ldots,۹۵۵۷$) در آمارگیری s ($s = ۱,\ldots,۴$) به‌صورت زیر توزیع شده است.

$$W_{is} \sim Normal(\mu_{is}, \sigma^{۲}),$$
$$\mu_{is} = \gamma + \varphi t + \mathbf{L}_i^T \phi$$

که $\mathbf{L}$ برداری شامل پاسخ‌های هر فرد i به سؤال « وزن دلخواه » است. بنا بر این وزن برای فرد i در آمارگیری s با میانگین جمعیت γ به علاوه‌ی یک افزایش φ در هر آمارگیری توصیف و طبق پاسخ فرد i به « وزن دلخواه » در S1 (ϕ) تعدیل شده است. براورد γ، φ و ϕ براساس آمارهای ثبتی دارای داده‌های کامل یا تا حدودی کامل متکی است. وزن زنان دارای وزن گم‌شده (W_{is}) از روی یک توزیع نرمال با میانگین $\hat{\mu}_{is}$ و واریانس $\hat{\sigma}^{۲}$ جانهی می‌شود.

توجه داشته باشید که مؤلفه‌ی دیابت در دو دوره‌ی زمانی ($t = ۱,۲$ آمارگیری‌های ۳ و ۴) ارزیابی می‌شود در حالی که مؤلفه‌ی وزن در چهار دوره‌ی زمانی ($s = ۱,\ldots,۴$، آمارگیری‌های ۱ تا ۴) ارزیابی می‌شود. این بدان معنی است که از حد اکثر اطلاعات برای جانهی وزن استفاده کرده‌ایم، در حالی که آمارگیری‌های ۱ و ۲ از مؤلفه‌ی دیابت حذف شده بودند زیرا ما تنها به موردهای وقوع بیماری علاقه‌مند بودیم.

۳-۳- مدل جانهی طولی

مؤلفه‌ی دیابت مدل همان ساختار مدل موردی کامل را دنبال می‌کند. ما یک عرض از مبدأ تصادفی را در مؤلفه‌ی جانهی مدل وارد کردیم تا همبستگی‌های درون–موضعی را در وزن بگنجانیم و درنتیجه طرح مطالعه‌ی طولی را به حساب آوریم.

مؤلفه‌ی جانهی برای وزن به قرار زیر است.

$$W_{is} \sim Normal(\mu_{is}, \sigma_b^2),$$
$$\mu_{is} = \gamma_i + \varphi\, t + \mathbf{L}_i^T \phi,$$
$$\gamma_i \sim Normal(\lambda, \sigma_w^2)$$

به جای میانگین وزن جمعیت، هر شخص براورد خود را دارد (γ_i به‌عنوان یک عرض از مبدأ تصادفی شناخته شده است). واریانس کل وزن از مدل قبلی (σ^2) به دو بخش واریانس درون–موضعی σ_w^2 و واریانس بین موضوعی σ_b^2 تقسیم شده است. همبستگی درون–موضعی با $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$ داده شده است.

۳-۴- استنباط با استفاده از نمونه‌گیری گیبس

مدل‌هایی که در بالا ارائه شده‌اند از دو توزیع پارامتری (برنولی و نرمال) با پارامترهای بسیاری در چندین سطح سلسله مراتبی استفاده می‌کنند. درنتیجه راه حل تحلیلی برای این مدل غیر قابل کنترل است. خوشبختانه می‌توانیم در مورد پارامترها با استفاده از نمونه‌گیری گیبس استنباط کنیم (روش پیش‌فرض در WinBUGS). در نمونه‌گیری گیبس هر پارامتر نامعلوم به شرط همه‌ی داده‌های دیگر مشاهده‌شده و دیگر پارامترهای براورد شده، براورد می‌شود (برای جزئیات نمونه‌گیری گیبس به گلمن و همکاران [۶] مراجعه کنید). برای مثال در مدل جانهی سلسله مراتبی، ممکن است یک وزن گم‌شده را از یک توزیع نرمال با میانگین $\hat{\mu}_{is}$ و واریانس $\hat{\sigma}^2$ نمونه‌گیری کنیم. تکرار کامل هنگامی اتفاق می‌افتد که همه‌ی پارامترها و داده‌های گم‌شده براورد شده باشد. در نتیجه تکرار بعدی بر اساس این براوردها و داده‌ها صورت می‌گیرد. برای آغاز تکرارها یک

مجموعه‌ی مقادیر اولیه برای هر مشاهده و پارامتر نامعلوم مشخص می‌شود. تکرارهای زیادی (معمولاً بیش از ۱۰۰۰ تکرار) در تلاش برای همگرا شدن در یک راه حل صورت می‌گیرد. برخی از مسائل عملی اجرای این قبیل تکرارها در بخش بعدی مورد بحث قرار گرفته‌اند.

۴- برنامه‌ی WinBUGS

برای اجرای تحلیل در WinBUGS چهار شرط اساسی وجود دارد: تعیین یک مدل، وارد کردن داده‌ها، تعیین مقادیر اولیه و اجرای نمونه‌گیر گیبس. این فرایند وقتی بسیار کارامد است که اطلاعات بالا در چهار دسته فایل ذخیره شوند: فایل داده‌های ورودی، فایل شامل مشخصات مدل، فایل مقادیر اولیه و فایلی که فرمان‌های WinBUGS را اجرا می‌کند. در این جا روی فایل مشخصات مدل تمرکز داریم که تبدیل هر یک از سه مدل را به فرمت WinBUGS نشان می‌دهد. مشخصات مدل در WinBUGS با سایر بسته‌های آماری استاندارد تفاوت دارد از این لحاظ که در آن به جای وارد کردن مشخصات مدل در شیوه‌های آماری از پیش برنامه‌ریزی‌شده باید مدل به‌طور کامل و روشن توسط کاربر مشخص شود. اطلاعات مربوط به ایجاد فایل‌های بسته‌ای باقی‌مانده در پیوست ۱ موجود است.

۴-۱- مدل موردی کامل

اولین خطوط برنامه به‌صورت زیر است.

```
model{
    for (i in 1:7311){
        for (t in 1:n survey[i]){
```

گزاره‌ی مدل، فایل مشخصات مدل را باز می‌کند. ما از آمارهای ثبتی مربوط به ۷۳۱۱ نفر در دو نقطه‌ی زمانی حلقه‌ای درست کرده‌ایم به جز جایی که زنی اولین تشخیص دیابت را در S3 گزارش داد که در این مورد او دیگر مشمول جامعه‌ی مورد مخاطره در S3 نبود و داده‌ها از تنها یک نقطه‌ی زمانی مورد استفاده قرار گرفت. این شرط با استفاده

از یک متغیر نشانگر، nsurvey، به دست آمده است که وقتی ابتلا بـه دیابـت بـین *S2* و *S3* اتفاق می‌افتد مقدار ۱ و در غیر این صورت مقدار ۲ را می‌گیرد.

در این مدل $t = ۱$ به ابتلای دیابت بین S2 و S3 و تغییر وزن بـین S1 و S2 اشـاره دارد. به همین ترتیب $t = ۲$ به ابتلای دیابت بین S3 و S4 و تغییـر وزن بـین S2 و S3 اشاره دارد. BMI و سن در S1 با گذشت زمان تغییر نمی‌کنند.

برای وقوع دیابت **(diab)** توزیع برنولی در نظر گرفته شده است،

```
diab[i,t]~dbern(diab.prob[i,t]);
```

پایان دستورات با یک نقطه–ویرگول در WinBUGS اختیاری است.

ما به پارامتر **diab.prob** علاقه‌مندیم که احتمال مبتلا شدن بـه دیابـت را نـشان می‌دهد. رابطه‌ی بین احتمال دیابت و دیگر متغیرهای توصیفی را به‌صورت زیر مدل‌بنـدی کرده‌ایم.

```
logit(diab.prob[i,t]) <- d.int + (d.time*equals(t,2))
        +(d.wtspc*wtspc[i,t])
        + (d.bmi[1]*equals(bmi[i],2))
        + (d.bmi[2]*equals(bmi[i],3))
        +(d.bmi[3]*equals(bmi[i],4))
        + (d.age*age[i]);
```

تابع **equals** با مقدار عدد صحیح t (نمایانگر زمان) اسـتفاده مـی‌شـود کـه پـارامتر **d.time** برای زمان ۲ « روشن » و برای همه‌ی زمان‌های دیگر « خـاموش » شـود. از این رو پارامتر **d.time** تغییر در خطر ابتلا به دیابت در S4 را در مقایـسه بـا S3 بـراورد می‌کند (همه‌ی پارامترهای براورد شـده بـرای مؤلفـه‌ی دیابـت یـک پیـشوند « **d.** » دارند). همچنین از تابع **equals** برای ایجاد متغیرهای نشانگر رسته‌های BMI اسـتفاده کرده‌ایم.

برای هر پارامتر مجهول در سطح پایین سلسله مراتب، پیشین‌هایی غیر آگـاهی‌بخـش تعیین کرده‌ایم. این کار با در نظر گرفتن یک توزیع نرمال با میـانگین صـفر و دقـت کـم

dnorm(0.0,1.0E-6) برای هر پارامتر مجهول انجام می‌گیرد، که دقت برابر عکس واریانس است.

۴-۲- مدل جانهی مقطعی

مشخصات مؤلفه‌ی دیابت مدل بسیار شبیه به مدل موردی کامل بود. همانند بالا، ما دیابت مدل‌سازی شده را یک متغیر برنولی در نظر گرفته‌ایم؛

```
diab[i,t]~ dbern(diab.prob[i,t]);
```

ولی، برای کمک به همگرایی، یک محدودیت روی حداقل مقداری که **diab.prob** می‌تواند بگیرد، اعمال کرده‌ایم (زیرا احتمال‌های خیلی کوچک منجر به غیر قابل براورد شدن درست‌نمایی‌ها می‌شود).

```
diab.prob [i,t]<- max (0.0001 , diab.temp[i,t]);
logit(diab.temp [i,t]) <- d.int+ (d.time*equals(t,2))
                +(d.wtspc*wtspc[i,t])
                +(d.bmi[1]*step(1.9- bmi[i,t]))
                +(d.bmi[2]*step(2.9- bmi[i,t]))
                +(d.bmi[3]*step(3.9- bmi[i,t]))
                + (d.age*age[i]);
```

با وارد کردن داده‌های گم‌شده در وزن، BMI در مدل یک متغیرتصادفی شده است. به این دلیل، ما نمی‌توانیم از تابع **equals** برای ایجاد یک تابع نشانگر برای BMI استفاده کنیم. مشکل با استفاده از تابع **step** و با مجموعه‌ی آستانه هایی با مقداری بین رسته‌های صحیح رفع می‌شود. توضیحات بیش‌تر در استفاده از تابع **step** و **equals** در کتاب راهنمای کاربر WinBUGS اشپیگل هالتر و همکاران [۲۶] یافت می‌شود.

مؤلفه‌ی جانهی این مدل صرفاً بر جانهی وزن در هر آمارگیری تمرکز دارد که برای محاسبه‌ی BMI و در صد تغییر وزن هر دو استفاده می‌شود. جایی که وزن برای فرد i در آمارگیری s مجهول باشد، توزیع آن نرمال در نظر گرفته می‌شود. علاقه‌ی ما به

رابطه‌ی بین مقدار میانگین وزن و دیگر اطلاعات موجود در داده‌ها بود که به‌صورت زیر نشان داده می‌شود.

```
for (s in 1:4){
wtkg[i ,s]<-cut(wtkg.uncut[i,s]);
wtkg. uncut[i,s]~dnorm(wt.mu[i,s], wt.tau);
wt.mu[i,s]<-w.int +(w.slo*s)
      +(w.like[1]*equals(like[i],2))
      +(w.like[2]*equals(like[i], 3))
      +(w.like[3]*equals(like[i],4));}
```

از تابع **cut** برای جلوگیری از پس خوراندن نتایج از دیابت‌های مؤثر بر وزن‌های جانهی‌شده استفاده کردیم. به عبارت دیگر برای حفظ جریان اطلاعات همان‌طور که در شکل ۱ نشان داده شده است.

این برنامه‌ها در حلقه‌ی « فرد »، یا « i »، کار گذاشته شده‌اند اما در قسمت بیرونی حلقه‌ی « t » که در آن مؤلفه‌ی دیابت وجود دارد، قرار گرفته‌اند. این باعث شد که وزن بتواند با استفاده از داده‌های هر ۴ آمارگیری مدل‌سازی شود و با این کار اطمینان به دست آید که مدل جانهی همه‌ی اطلاعات موجود را مورد استفاده قرار داده است. وزن، متغیری نبود که مورد علاقه‌ی مستقیم ما باشد بنا بر این ما از توابع منطقی برای محاسبه‌ی مجدد متغیرهای **bmi** (رسته‌ای) و **wtspc** (پیوسته) به‌صورت زیر استفاده کرده‌ایم.

```
wtspc[i,t]<-((wtkg[i,t+1]-wtkg[i,t])/wtkg[i,t])/(t+1)
          *100;
bmic[i,t] <-wtkg[i,1]/(height [i]*height[i]);
bmi[i,t]<-1+step(bmic[i,t]-25)+step(bmic[i,t]-30)
        +step(bmic[i,t]-35);
```

۴-۳- مدل جانهی طولی

مؤلفه‌ی دیابت این مدل درست مانند مدل مقطعی است. مؤلفه‌ی جانهی تنها بـه تعـدیلی اندک در عبارت عرض از مبدأ و یک خط جدید برنامه برای توصیف توزیع عرض از مبـدأ تصادفی نیاز داشت.

```
for(s in 1:4){
wtkg[i,s]<-cut(wtkg.uncut[i,s]);
wtkg.uncut[i,s]~dnorm(wt.mu[i,s],wt.tou);
wt.mu[i,s]<-w.int[i]+(w.slo*s)
        +(w.like[1]*equals(like [i],2))
        + (w.like[2]*equals(like[i],3))
        +(w.like[3]*equals (like[i],4));}
w.int[i]~ dnorm(w.mu,w.tau);
```

برازش سه مدل مزبور با استفاده از معیار اطلاع انحـراف (DIC) [۲۵] و اعتبارسـنجی متقابل ده برابر [۱]، مقایسه شده است.

۵- نتایج

تحلیل هر یک از این سه مدل در WinBUGS نسخه‌ی ۱.۴.۱ انجـام شـده اسـت. هـر مدل به تعداد ۲۵۰۰۰ بار تکرار شده است و تعداد ۵۰۰۰ تکرار اضافی برای دوره‌ی داغیدن در نظـر گرفتـه شـده اسـت. زمـان لازم بـرای اجـرای مـدل جـانهی طـولی (بـسیار پیچیـده) در WinBUGS برابـر ۵۳ دقیقـه بـود کـه از یـک سـرور Microsoft Windows Server 2003 Enterprise Edition و پردازنده‌ی دوگانه‌ی Xeon با ظرفیت ۳/۶ GHz و RAM با ظرفیت ۶ GB استفاده شده است. زمان اجرا بـه تعداد مشاهدات در داده‌ها و تعداد تکرارهای لازم بستگی زیادی دارد.

برای مقایسه‌ی نتایج WinBUGS با دیگر بسته‌ها، تحلیل موردی کامل در SAS، نسخه‌ی ۹.۱.۳ با استفاده از **Proc genmod** با اختیار **type=exch** و **d=binomial** و **Link= logit** نیز اجرا شده است.

نسبت بخت‌ها برای وقوع دیابت در هر یک از سه مدل در جدول ۱ نشان داده شده است. در نسبت بخت‌ها برای دیابتی‌ها یا حدود پسین آن‌ها، بین مدل‌های مختلف و دو بسته، تفاوت کمی وجود داشت. تفسیر نتایج نیز با نتیجه‌ی اصلی مبنی بر این که چاقی بلندمدت (BMI) پیش‌گوی قوی‌تری برای وقوع دیابت درمقایسه با افزایش وزن کوتاه‌مدت است تغییر نکرده است [۱۱].

مدل طولی، برازش بهتری برای داده‌ها نسبت به مدل مقطعی بود زیرا مدل طولی DIC کوچک‌تری داشت (جدول ۲). مدل طولی از پارامترهای به‌مراتب بیشتری استفاده می‌کرد زیرا هر زن عرض از مبدأ خود را داشت. این افزایش زیاد پارامترها برازش مؤلفه‌ی جانهی مدل را بهبود بسیار زیادی بخشید. این جانهی ارتقا یافته، برازش مؤلفه‌ی دیابت را اندکی بهتر کرد (DIC برابر ۲۵۷۰/۱ در مقابل ۲۵۷۳/۸).

جدول ۱- نسبت بخت‌ها (و حدود اطمینان/پسین ۹۵٪) برای وقوع دیابت

	حالت کامل (N= ۷۳۱۱)		جانهی‌شده (N= ۹۵۵۷)	
	SAS	WinBUGS	مقطعی	طولی
زمان (سال)	NA	۱/۳۴(۱/۰۰,۱/۷۶)	۱/۴۸(۱/۱۶,۱/۸۷)	۱/۵۰(۱/۱۷,۱/۹۴)
شاخص جرم بدن				
وزن سالم	مرجع	مرجع	مرجع	مرجع
اضافه وزن	۲/۹۵(۲/۰۰,۴/۳۶)	۳/۰۲(۲/۰۰,۴/۴۱)	۲/۸۸(۲/۰۳,۳/۹۷)	۳/۰۸(۲/۱۵,۴/۳۵)
چاق	۷/۵۹(۵/۱۰,۱۱/۲۱)	۷/۷۳(۵/۱۳,۱۱/۲۰)	۶/۶۳(۴/۳۷,۹/۰۳)	۶/۷۴(۴/۶۱,۹/۶۹)
بسیار چاق	۱۳/۴۳(۸/۵۰,۲۱/۲۳)	۱۲/۹۷(۸/۶۱,۱۸/۲)	۱۳/۶۸(۹/۱۳,۱۹/۶۷)	۱۳/۵۵(۹/۱۱,۱۹/۷۴)
تغییرات وزن ٪	۱/۰۳(۰/۹۹,۱/۰۷)	۱/۰۳(۰/۹۹,۱/۰۸)	۱/۰۲(۰/۹۹,۱/۰۵)	۱/۰۳(۰/۹۹,۱/۰۷)
سن (سال)	۱/۱۰(۱/۰۰,۱/۲۲)	۱/۱۰(۱/۰۰,۱/۲۱)	۱/۱۳(۱/۰۴,۱/۲۲)	۱/۱۳(۱/۰۳,۱/۲۳)

نتایج اعتبارسنجی متقابل ۱۰ برابر شبیه به نتایج حاصل از DIC بود. برای مدل مقطعی، خطای متوسط برای یک وزن جانهی شده ۱۰/۴ کیلوگرم (انحراف استاندارد SD=۰/۲۳ کیلوگرم) بوده است. برای مدل طولی، متوسط خطا به مراتب کمتر از ۵/۳

کیلوگرم ($SD=۰/۱۷kg$) به دست آمده است. اعتبارسنجی متقابل تفاوت کمی بین این مدل‌ها از نظر برازش آن‌ها با مؤلفه‌ی دیابتی پیدا کرد. برای مدل مقطعی، متوسط مساحت زیر منحنی مشخصه‌ی عملگر دریافت‌کننده (ROC) برای دیابت‌های پیشگویی‌شده (بلی/خیر) برابر ۰/۵۴۸ بود. برای مدل طولی متوسط مساحت زیر منحنی ROC برابر ۰/۵۴۳ بود.

جدول۲- مقایسه ی برازش مدل مقطعی و طولی (معیار اطلاعات انحراف = DIC)

	مقطعی		طولی	
مؤلفه	**تعداد پارامترها**	**DIC**	**تعداد پارامترها**	**DIC**
دیابت	۷/۷	۲۵۷۳/۸	۸/۹	۲۵۷۰/۱
جانهی	۶/۰	۲۶۱۲۷۱	۹۰۱۶/۲	۲۱۱۴۱۸
کل	۱۳/۶	۲۶۳۸۴۴	۹۰۲۵/۱	۲۱۳۹۸۸

۶- یک ماکروی SAS کلی‌تر

در این بخش ما یک ماکروی SAS را برای تحلیل مدل‌های طولی عمومی با داده‌های گم‌شده‌ی متغیرکمکی توضیح می‌دهیم. ماکرو یک مجموعه‌ی داده‌های SAS را به فرمت WinBUGS تبدیل می‌کند و برنامه‌های WinBUGS را برای اجرای مدل طولی می‌نویسد. سپس نتایج را از WinBUGS می‌خواند و آن‌ها را در SAS خلاصه می‌کند.

این ماکروی خاص به مدل‌هایی با متغیر وابسته‌ی پیوسته محدود می‌شود و بنا بر این محدودتر از مدل‌های استفاده شده در بخش ۴ است. در مدل، امکان حضور متغیرکمکی چندگانه وجود دارد که ممکن است رسته‌ای یا پیوسته باشند.

ولی متغیرکمکی دارای داده‌های گم‌شده باید پیوسته و وابسته به زمان باشد (یعنی، در طول زمان تغییر کند). این ماکرو از مدل ساده‌تری نسبت به مدلی که در بخش قبل برای مجموعه‌ی داده‌های دیابتی بحث شد، استفاده می‌کند. مدل برای دیابت شامل توابع خاصی (مانند محاسبه‌ی شاخص جرم بدن از روی وزن) است و از دوره‌های زمانی متفاوتی برای مدل مورد علاقه و مدل جانهی استفاده کرده است. این گونه برنامه‌نویسی

خاص را می‌توان به برنامه‌ی WinBUGS که بـه وسـیله‌ی مـاکروی SAS تولیـد شـده است، اضافه کرد.

در این جا ماکروی SAS بـرای یـک متغیـر وابسـته‌ی پیوسـته بـا یـک مجموعـه‌ی داده‌های طولی کوچک‌تر، مربوط به مطالعه‌ی بیماران سرطانی لاعلاج و پرسـتاران آن‌هـا توضیح داده می‌شود [۴]. برامد مورد نظر سطح نگرانی پرستاران است که بر اساس نگرانی بیمارستان و میزان افسردگی اندازه‌گیری می‌شود (HADS) و در این مقیاس امتیاز بـالاتر نشان‌دهنده‌ی نگرانی بیش‌تر است [۳۲]. در این مطالعـه بیمـاران و پرستارانـشان بـه‌طـور منظم در آخرین سال زندگی مصاحبه شده‌اند. در این مطالعه موضوع مورد علاقـه‌ی ویـژه این بود که، اضطراب بیمار و چگونه بر اضطراب اثر می‌گذارد. هم اضـطراب بیمـار و هـم اضطراب پرستار دارای مقادیر گم‌شده هستند. متغیرهای دیگر جنـسیت پرسـتار، زمـان تـا مرگ (برحسب هفته)، و تعداد علائم بیماری بیمار است. تعداد علائـم بیمـاری و امتیـازات اضطراب، متغیرهای‌کمکی وابسته به زمان هستند. مجموعه‌ی داده‌ها شامل ۵۱۴ مـصاحبه از ۱۰۹ پرستار است.

برای داده‌های سـرطان لاعـلاج پایـانی مـاکروی SAS، **longimp**، بـا اسـتفاده از گزاره‌ی زیر فراخوانده می‌شود (توجه کنید که متن بین « **/*** » و « ***/** » توضـیح اسـت که می‌تواند حذف شود).

```
%longimp (dsetin = data. Carer, /* Input data set */
Depvar = hadsanx, /* Dependent variable (model of interest)*/
Vars = phadsanx gender deathwks,
      /* Explanatory var(s)(model of interest )*/
Class = gender, /* Class explanatory variable(s)
                (model of interest)*/
Depvari = phadsanx,
      /* Dependent variable(imputation model , continuous) */
varsi = count, /* Explanatory variable(s)(imputation model)*/
```

```
classi =, /* class explanatory variable(s)
            (imputation model)*/
time = interno, /* Time variable */
repeated = carerid, /* Repeated variable (e.g. subject )*/
centre = Y, */ Centre continuous explanatory variables (Y/N),
              default = Y*/
MCMC = 5000/* Number of MCMC iterations & burn - in,
             default = 1000*/
);
```

متغیر وابسته در مدل، نگرانی پرستار است **(depvor = hadsanx)**. این متغیر وابسته به جنسیت آن‌ها (که یک متغیر رسته‌ای است، **class = gender**)، سطح نگرانی بیمار آن‌ها **(Phadsanx)** و زمان تا مرگ بیمار **(deathwks)** است. سطح نگرانی بیمار یک متغیرکمکی وابسته به زمان است و چندین مقدار گم‌شده دارد. این مقادیر گم‌شده را با تبدیل آن‌ها به متغیرهای وابسته‌ی مدل جانهی **(depvari = phadsanx)** جانهی کرده‌ایم. سطح نگرانی بیمار وابسته به تعداد علائم بیماری آن‌ها است **(count)**. تعداد علائم بیماری دارای مقادیر گم‌شده نیستند.

دیگر ورودی‌های لازم یک متغیر تعریف‌کننده‌ی زمان است که در این مورد تعداد مصاحبه‌ها است **(time=interno)** و یک شمارنده‌ی شناسایی برای هر پرستار است که نتایج تکرارشده‌ی آن‌ها را متصل می‌کند **(repeated=carerid)**.

گزینه‌ای برای مرکزی کردن متغیرهای توضیحی پیوسته با تفریق میانگین آن‌ها وجود دارد **(centre)**. این کار به‌طور کلی در WinBUGS به صلاح است زیرا اغلب، همگرایی الگورتیم MCMC را بهبود می‌بخشد. گزینه‌ی نهایی انتخاب تعداد تکرارهایی MCMC و دوره‌ی داغیدن آن است **(MCMC)**.

فراخواندن ماکروی SAS بالا خروجی چهار صفحه‌ای زیر را تولید می‌کند.

Longitudinal model using WinBUGS 1

Model in formation

Input data set: work.carer
Dependent variable: hadsanx
Observations: 514
Subjects/Clusters: 109

Longitudinal model using Win BUGS 2
Geweke’s MCMC converge diagnostic

Variable	Mean difference	df	t-value	p-value
Count	-0.00057	2998	-0.24445	0.80690
Deathwks	0.00082	2998	1.47354	0.14071
Gender-1	0.11865	2998	3.30014	0.00098
Intercept	-0.01108	2998	-0.46386	0.64278
Intercept-i	-0.01178	2998	-0.86887	0.38499
phadsanx	0.00506	2998	1.90966	0.05627

Longitudinal model using Win BUGS 3
Parameter estimates – Model of interest

Variable	Mean	SD	95% Posterior Interval Lower	Upper
Intercept	6.646	0.486	5.556	7.479
deathwks	0.039	0.011	0.016	0.062
Gender-1	-2.649	0.756	-4.112	-1.203
Phadsanx	0.288	0.054	0.182	0.390
Sigma2	8.377	0.610	7.278	9.656
Rho	0.597	0.047	0.502	0.684

```
Longitudinal Model Using WinBUGS                    4
Parameter estimates - Imputation model

                                          95% Posterior Interval
Variable      mean      SD      Lower     Upper
intercept-i   0.263     0.271   -0.281    0.792
Count         0.351     0.046   0.261     0.443
```

اولین صفحه‌ی خروجی برخی اطلاعات اساسی در باره‌ی مدل ارائه می‌دهد. صفحه‌ی دوم تشخیص همگرایی MCMC [۷] را ارائه می‌دهد. این ۱۰ درصد اول زنجیره را با ۵۰ درصد آخر آن مقایسه می‌کند. برای یک استنباط معتبر، داشتن زنجیره‌های MCMC پایدار و همگراشده مهم است. زنجیره‌ای که همگرا شده است باید یک میانگین ثابت داشته باشد و خروجی، میانگین‌های اولین و آخرین بخش‌های زنجیره را با استفاده از t–تست جفت‌نشده با هم مقایسه کند. در این مورد به نظر می‌رسد میانگین متغیر جنسیت کمی افزایش داشته است. زنجیره باید دوباره اجرا شود (درصورت امکان مدت طولانی‌تر) تا براوردهای پایدارتری برای جنسیت تولید کند.

صفحات سوم و چهارم براوردهای پارامتر را با استفاده از مدل مطلوب و مدل جانهی ارائه می‌دهد. نتایج نشان می‌دهد که همچنان که مرگ بیمار نزدیک می‌شود اضطراب پرستار افزایش می‌یابد **(deathwks)**. همچنین اضطراب پرستاران مرد به مراتب کمتر **(gender)** و افزایش نگرانی بیمار با افزایش نگرانی پرستار مرتبط بود **(phadsanx)**. میانگین امتیاز اضطراب **(intercept)** ۶/۴۶۴ و واریانس آن **(sigma2)** ۸/۳۷۷ بود همبستگی درون–موضوعی نسبتاً قوی برای پرستاران که برابر **(rho)** ۰/۵۹۷ بوده نشان‌دهنده‌ی تشابه بسیار زیاد در هر سطح نگرانی پرستاران در گذر زمان است. برای مدل جانهی تعداد علائم بیمار با سطح اضطراب آن‌ها ارتباط مثبت دارد.

۷- بحث

داده‌های آمارگیری اغلب تاحدودی کامل می‌شوند و استفاده از تحلیل موردی کامل می‌تواند نتایج اریب تولید کند. در مثال مربوط به مخاطره‌ی وقوع دیابت که این جا استفاده شد چنین نبود زیرا تحلیل موردی کامل و تحلیل پس از جانهی براوردهای مشابهی را تولید کرده‌اند. ولی تحلیل موردی کامل از داده‌های مربوط به ۷۳۱۱ زن استفاده کرده است در حالی که تحلیل‌هایی که از جانهی استفاده می‌کنند داده‌های مربوط به ۹۵۵۷ زن را به‌کار برده‌اند (۳۱ درصد افزایش). استفاده از نمونه‌ی بزرگ‌تر عموماً نتایجی را ارائه می‌دهند که بیش‌تر نمایان‌گرند.

در مثال ما در باره‌ی وقوع دیابت بین پارامترهای براوردشده از مدل جانهی مقطعی و مدلی که ساختار طولی داده‌ها را مورد استفاده قرار می‌داد، تفاوتی وجود نداشت. ولی، همان‌طورکه اعتبارسنجی متقابل نشان داده است مدل جانهی طولی براوردهای بهتری از وزن‌های گم‌شده ارائه می‌دهد (میانگین خطای ۵/۳ کیلوگرمی برای مدل طولی در مقایسه با ۱۰/۴ کیلوگرم برای مدل مقطعی). بنا بر این هنوز قویاً استفاده از جانهی طولی را هنگامی‌که ساختار داده‌ها طولی هستند، توصیه می‌کنیم.

بسیاری از نرم‌افزارهای موجود برای جانهی باید ظرفیت‌های خود را برای جانهی طولی گسترش داده و به جای آن از جانهی مقطعی استفاده نکنند. از طرف دیگر، WinBUGS می‌تواند ساختار طولی را با سادگی نسبی در خود بگنجاند. همان‌طورکه در بخش ۴ نشان داده شد، تفاوت‌های کمی در برنامه‌ی جانهی مقطعی و طولی وجود داشت. WinBUGS همچنین قادر است به‌آسانی داده‌های رسته‌ای گم‌شده را حل و فصل کند، در حالی که بسیاری از دیگر بسته‌ها بر فرض نرمال بودن تکیه دارند. ماکروی SAS برای تحلیل مطالعات طولی با متغیرهای کمکی گم‌شده تهیه شده است که از WinBUGS استفاده می‌کند اما قابل اجرا شدن در SAS است.

مرجع‌ها

[1] Breiman, L.; Friedman. J.; Olshen, R.; Stone, C. (1984). *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth

International Group, Belmont, California.

[2] Carpenter, J.; Kenward, M. (2005). Example Analyses Using WinBUGS 1.4. URL http://www.missingdata.org.uk/.

[3] Congdon, P. (2001). *Bayesian Statistical Modelling*. Wiley Series in Probability and Statistics. Wiley, Chichester, New York.

[4] Correa-Velez, I.; Clavarino, A.; Eastwood, H.; Barnett, A. (2003). Use of complementary and alternative medicine and quality of life: changes at the end of life. *Palliative Medicine*, **17**, 695–703.

[5] Cowles, M.K. (2004). Review of WinBUGS 1.4. *The American Statistician*, **58**, 330–336.

[6] Gelman, A.; Carlin, J.B.; Stern, HS.; Rubin, D.B. (2004). *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, Fla., 2nd edition.

[7] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J Bernado, J Berger, A Dawid, A Smith (eds.), Bayesian Statistics 4, Clarendon Press, Oxford, UK.

[8] Ibrahim, J.G.; Chen, M.H.; Lipsitz, S.R.; Herring, A.H. (2005). Missing-data methods for generalized linear models: A comparative review. *J. Amer. Statist. Assoc.*, **100**, 332–346.

[9] Insightful Corp (2003). *S-PLUS Version 6.2*. Seattle, WA. URL http://www.insightful.com/.

[10] Lee, C.; Dobson, A.J.; Brown, W.J.; Bryson, L.; Byles, J.; Warner-Smith, P.; Young, A.F. (2005). Cohort profile: The Australian longitudinal study on women's health. *International Journal of Epidemiology*, **34**, 987–991.

[11] Mishra, G.D.; Carrigan, G.; Brown, W.J.; Barnett, A.G.; Dobson, A.J. (2007). Short-term weight change and the incidence of diabetes in midlife: results from the Australian longitudinal study of women's health. *Diabetes Care*.

[12] Mishra, G.D.; Dobson, A.J. (2004). Multiple imputation for body mass index: lessons from the Australian longitudinal study on women's health. *Statistics in Medicine*, **23**, 3077–3087.

[13] Pettitt, A.N.; Tran, T.T.; Haynes, M.A.; Hay, J.L. (2006). A Bayesian hierarchical model for categorical longitudinal data from a social survey of immigrants. *J. R. Stat. Soc. Ser. A*, **169**, 97–114.

[14] Qiu, Z.; Song, P.X.K.; Tan, M. (2002). Bayesian hierarchical models for multi-level repeated ordinal data using WinBUGS. *Journal of Biopharmaceutical Statistics*, **12**, 121–135.

[15] Raghunathan, T.E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, **25**, 99–117.

[16] Raghunathan, T.E,; Solenberger, P.; Van Hoewyk, J. (eds.) (2002). *IVEware: Imputation and Variance Estimation Software Users Guide*. University of

Michigan: Survey Research Center, Institute for Social Research.

[17] Rasbash, J.; Steele, F.; Browne, W.; Prosser, B. (2005). *A User's Guide to MLwiN Version 2.0*. University of Bristol, Bristol. URL http://www.cmm.bris.ac.uk/MLwiN/download/manuals.shtml.

[18] R Development Core Team (2007). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[19] Robins, J.M.; Rotnitzky, A.; Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, **89**, 846–866.

[20] Robins, J.M.; Rotnitzky, A.; Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.*, **90**, 106–121.

[21] Royston, P. (2005). Multiple imputation of missing values: update. *The Stata Journal*, **5**, 1–14.

[22] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, NewYork.

[23] SAS Institute Inc (2003). *SAS/STAT Software, Version 9.1*. Cary, NC. URL http://www.sas.com/.

[24] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

[25] Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; Van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (With Discussion). J. R. Stat. Soc.
Ser. B, **64**, 583–640.

[26] Spiegelhalter, D.J.; Thomas, A.; Best, N.G.; Lunn, D. (2003). *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge. URL http://www.mrc-bsu.cam.ac.uk/bugs/.

[27] StataCorp (2003). *Stata Statistical Software: Release 8*. StataCorp LP, College Station, TX. URL http://www.stata.com/.

[28] Sturtz, S.; Ligges, U.; Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, **12**, 1–16.

[29] Van Buuren, S.; Oudshoorn, K. (1999). Flexible Multiple Imputation by MICE. *Technical Report PG/VGZ/99.054*, TNO Prevention and Health. URL http://www.multiple-imputation.com/.

[30] Woodworth, G.G. (2004). *Biostatistics: A Bayesian Introduction*. Wiley-Interscience, Hoboken, NJ.

[31] World Health Organization (2000). Obesity: Preventing and Managing the Global Epidemic. *Technical Report 894*, WHO Technical Report Series.

[32] Zigmond, A.S.; Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, **67**, 361–370.

پیوست

۱- فایل‌های دسته‌ای در WinBUGS

۱-۱- فایل داده‌های ورودی

داده‌های ردیفی مقادیر متغیرها هستند: دیابت، وزن (در هر آمارگیری)، سن، قد، تمایل به تنظیم وزن و متغیر نشانگر: **nsurvey**.

داده‌ها می‌توانند به یکی از این دو روش وارد WinBUGS شوند: لیست فرمت S-PLUS یا به‌صورت یک سری از آرایه‌های یک‌بعدی در یک فایل مجزاشده‌ی جدول‌بندی. در هر دو مورد، ابعاد سلسله مراتبی داده‌ها باید توسط کاربر تعیین شود. داده‌ها به‌صورت فایل txt. ذخیره می‌شوند. داده‌های گم‌شده به‌صورت « NA » وارد می‌شوند.

WinBUGS برای مدیریت داده‌ها مناسب نیست، بنا بر این برای مجموعه‌ی داده‌های بزرگ‌تر بسته‌های دیگری لازم است. ما از SAS، نسخه‌ی ۹.۱.۳ استفاده کردیم تا فایل داده‌ها را به‌صورت متن مجزا و جدول‌بندی‌شده ایجاد کنیم. همچنین می‌توان از R و بسته‌ی R2WinBUGS در مدیریت مجموعه‌ی داده‌های بزرگ‌تر استفاده کرد [۲۸].

با استفاده از آرایه‌ی ترتیب داده‌ها، داده‌ها تحت سرستون‌های زیر وارد می‌شوند (سه خط مثال از داده‌ها نیز نشان داده شده‌اند).

```
Diab [,1] diab [,2] wtkg[,1] wtkg[,2] wtkg[,3] wtkg[,4]
 age[] hight[] like[] nsurrey[]
1  NA  40.5  30.2  NA  35.5  51  140  1  1
0  1  89.8  89.0  75.5  99.2  52  160  4  2
0  0  65.2  65.4  66.7  NA  50  175  2  2
...
```

آخرین خط فایل داده‌های این فرمت باید کلمه‌ی « END » باشد و کلید بازگشت باید برای WinBUGS وارد شود تا فایل درست خوانده شود.

۲-۱- فایل مقادیر اولیه

مقادیر اولیه می‌توانند به یکی از دو روش تعیین شوند: توسط کاربر یا به‌صورت مقادیر تولیدشده به‌صورت تصادفی با استفاده از **`gen.inits()`** در فایل دست نوشته‌ی WinBUGS. همچنین ممکن است بعضی مقادیر اولیه تعیین و سپس از تابع **`gen.inits()`** برای ایجاد بقیه استفاده شود. معیارهای خاص کاربر در فایل txt. ذخیره می‌شود که ساختار آن از همان پروتوکل فایل داده‌ها پیروی می‌کند. اغلب ضروری است که کاربر مقادیر اولیه‌ی تخمینی را تعیین کند زیرا **`gen.inits()`** می‌تواند مقادیر آغازین نامعقولی را تولید کند که به این معنی است که WinBUGS نمی‌تواند آغاز به تکرار نمونه‌گیر گیبس کند.

WinBUGS با هر رکود مشاهده‌نشده در متغیر دارای داده‌های گم‌شده به‌صورت یک متغیر تصادفی رفتار می‌کند که توزیع آن باید براورد شود. برای تعیین مقادیر اولیه برای این رکوردهای مشاهده‌نشده، باید یک ماتریس از مقادیر را وارد کنیم که با ابعاد متغیر جور باشد. اگر یک رکورد مشاهده شود، ورودی در ماتریس مقادیر اولیه « NA » است و گرنه، اگر رکورد مشاهده نشود، ورودی همان مقدار اولیه‌ی مشخص شده توسط کاربر است.

۳-۱- فایل نوشتاری

فایل نوشتاری فهرستی از فرمان‌های WinBUGS است که به‌طور متوالی اجرا می‌شوند و با انتخاب « Script » در منوی Model فعال می‌شود. فرمان‌های زیر باید به ترتیبی که در زیر نشان داده شده، در این فایل وجود داشته باشد تا مدل اجرا شود.

`check('path/filename')`: فایل مشخصات مدل را فرا می‌خواند و خطاها را در نحو برنامه را پویش می کند. به خط کج روبه‌رو در فایل path توجه کنید.

`Load('path/file')`: فایل داده‌ها را فرا می‌خواند.

compile(n): اطمینان می‌دهد که فایل مشخصات مدل و فایل داده‌ها سازگارند. n نشان‌دهنده‌ی تعداد زنجیره‌هایی است که باید هم‌زمان اجرا شوند.

inits(n,path/filename): فایل مقادیر اولیه را فرا می‌خواند.

gen.inits(): مقادیر اولیه را که در فایل مقادیر اولیه تعیین نشده‌اند تولید می‌کند. اگر مقادیر اولیه به‌طور کامل به‌وسیله‌ی کاربر تعیین شده باشند این فرمان لازم نیست.

update(n): شبیه‌سازی MCMC را آغاز و n تکرار را اجرا می‌کند.

فرمان‌های اضافی برای دیدن خروجی لازم است. مجموعه‌ی (name) نشان می‌دهد که نام یک گره مورد علاقه است. WinBUGS همچنان که زنجیره را تکرار می‌کند اطلاعاتی درباره‌ی نام ذخیره خواهد کرد. چند فرمان بسیار مفید برای دیدن خروجی در زیر نشان داده شده است.

stat(node): میانگین‌های پسین، فاصله‌ها و دیگر آماره‌ها را برای گره‌های مورد علاقه ارائه می‌دهد.

history(node): یک نمودار از همه‌ی مقادیری است که زنجیره در طول چرخه‌ی تکرار گرفته است.

autoc(node): یک نمودار از خود همبستگی بین مقادیر مجاور در زنجیره تا ۴۰ تأخیر.

کتاب راهنمای کاربر WinBUGS یک فهرست سودمند و جامع از فرمان‌ها تهیه کرده است [۲۶].

لیلی تاپاک
فوق لیسانس آمار
همدان، میدان دانشگاه، چهارراه عارف، مؤسسه‌ی آموزش عالی آبادانی و توسعه‌ی روستاها.
پیام‌نگار: leylytapak81@yahoo.com

امید حمیدی
فوق لیسانس آمار
استان همدان، رزن، دانشگاه پیام نور واحد رزن.
پیام‌نگار: