Archive of SID

Journal of Statistical Sciences, Spring and Summer, 2022
Vol. 16, No. 1, pp 189-207
DOI: 10.29252/jss.16.1.189

Using Markov Latent Class Models in Estimating the Classification Error of Iranian Labor Flow Statistics

Fallah Mohsekhani, Z., , Kalhori Nadrabadi, L.
Statistical Research and Training Center, Tehran , Iran.

Corresponding author: L. Kalhori, kalhori@srtc.ac.ir

Received: 14 March 2022 **Revised:** 4 July 2022 **Accepted and Published Online:** 7 July 2022.

Introduction

In analyzing labor market statistics, policies and plans are usually based on changes in stock variables. The number of changing statuses can be estimated in countries where labor force surveys are based on rotational samples and ordinary people in different periods. Flow statistics present the input and output of the labor force and the number of changing status labor force between periods. One of the essential non-sampling errors in labor force statistics is the response error, which may lead to the incorrect classification of individuals. To use flow statistics, their classification error must be specified so that knowledge of this error and the analyzing on flow statistics can be done. Usually, the error of classifying flow statistics is estimated using the interview method, which is costly and complicated. In this paper, the feasibility of each model is examined while presenting the process of estimating flow statistics and appropriate models for calculating the classification error according to the sample rotation pattern in Iran.

Material and Methods

Markov Latent Class Models (MLC) for panel data analysis exploit panel surveys' repeating nature to extract classification error information. MLC analysis may be the only way to evaluate flow error in panel surveys where are-interview program is not possible from the interview data. Let A_t be the observed data at wave t and X_t denote the observable true value at wave t. Also, the cross-classification of the variable A at three waves is denoted by $A_1A_2A_3$. The MLC model contains two components: (1) the structural component $\pi^{X_1X_2X_3}_{x_1x_2x_3}$ that describes the interdependencies between the

www.SID.ir

Archive of SID

X_t,$t = 1, 2, 3$ and the model covariates (grouping variables) which represent the time-to-time transitions among the true classifications and (2) the error component $\pi^{A_1A_2A_3|X_1X_2X_3}_{a_1a_2a_3|x_1x_2x_3}$ describing the interactions between A_t at each wave $t = 1, 2, 3$ and X_t and other model covariates which represent the deviations between the true and observed classifications. To reduce the number of parameters, the transition probabilities can be assumed to be stationary. The Iranian Labour Force Survey (LFS) is a seasonal multi-stage survey. Weighting is carried out in three stages. To estimate changes between periods without losing efficiency of current level estimation, rotation sampling (2-2-2) is used. In this paper, we applied the MLC Model to estimate the flow error in the Labore Force Survey of Iran in the 2019 and 2020 years.

Results and Discussion

We assumed four different scenarios to evaluate the LFS flow error due to the limitation of 2-2-2 rotation pattern. The observations of four sampling periods were considered for each sample unit, assuming no changes in the situation between the second and third periods. Also, the flow statistics of four different time sequences were studied to assess the seasonal effects on flow error. The results demonstrated that the highest error is related to the unemployed status, and the lowest error is related to inactive status.

Conclusion

According to the pattern of sampling rotation in LFS of Iran, it is possible to estimate the flow statistics using labor force statistics. Non-sampling error is one of the most influential errors in flow statistics. one of the essential non-sampling errors in these statistics is response errors that may lead to misclassification of individuals into the workforce status. MLC models can be used to estimate the classification error of flow statistics. According to the proposed model, it is possible to provide labor force statistics and the classification error without spending money on re-interviewing.

Keywords: Labor Force Survey, Flow Statistic, Markov Latent Class Model.

Mathematics Subject Classification (2010): 62P25, 03C98.

©The Author(s). The Publisher is Iranian Statistical Society.
This is an open access article distributed under the terms and conditions of (CC BY-NC 4.0)

www.SID.ir

Archive of SID

مجله علوم آماری، بهار و تابستان ۱۴۰۱
جلد ۱۶، شماره ۱، ص ۱۸۹ – ۲۰۷
DOI: 10.29252/jss.16.1.189
مقاله پژوهشی

استفاده از مدل‌های رده نهان مارکوف دربرآورد خطای رده‌بندی آمارهای جریان نیروی کار ایران

زهره فلاح محسن‌خانی، لیدا کلهری ندرآبادی

پژوهشکده آمار

نویسنده مسئول : لیدا کلهری، kalhori@srtc.ac.ir

تاریخ دریافت: ۱۴۰۰/۱۲/۲۳ تاریخ بازنگری: ۱۴۰۱/۰۴/۱۳ تاریخ پذیرش و انتشار: ۱۴۰۱/۰۴/۱۶

چکیده: در کشورهایی که آمارگیری نیروی کار بر اساس نمونه‌گیری چرخشی است، به علت وجود افراد مشترک در دوره‌های مختلف آمارگیری، آمار تغییر وضعیت، قابل برآورد است و با عنوان آمارهای جریان ارائه می‌شود. یکی از مهم‌ترین خطاهای غیرنمونه‌گیری در آمارهای بازارکار، خطای پاسخ است. در آمارهای جریان این خطا مضاعف است. معمولا خطای رده‌بندی آمارهای جریان با استفاده از روش پر هزینه و دشوار مصاحبه مجدد برآورد می‌شود. در این مقاله ضمن ارایه فرایند برآورد آمار جریان و مدل‌های مناسب برای محاسبه خطای رده‌بندی در آن، با توجه به الگوی چرخش نمونه‌گیری در آمارگیری نیروی کار ایران هر یک از روش‌ها امکان سنجی شده است. در خاتمه مدل رده نهان مارکوف با فرض نابرابری احتمال‌های انتقال بر اساس الگوی چرخش نمونه‌های نیروی کار ایران به‌عنوان مدلی مناسب برای برآورد خطای رده‌بندی در آمار جریان ایران با استفاده از داده های آمارگیری نیروی کار سال های ۱۳۹۸ و ۱۳۹۹ معرفی شده است.

واژه‌های کلیدی: آمارگیری نیروی کار، آمار جریان، مدل رده نهان مارکوف.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62P25، 03C98.

۱ مقدمه

اشتغال پایدار یکی از مهمترین دستاوردهای هر دولتی است و ارائه برآوردهای آن با آمارهای تغییرات نیروی کار امکان‌پذیر است. در ایران آمارگیری نیروی کار پس از تجدید نظر همه جانبه از سال ۱۳۸۴ با استفاده از نمونه‌گیری

©نویسند(گان). ناشر انجمن آمار ایران است.
این مقاله با دسترسی آزاد تحت شرایط و ضوابط (CC BY-NC 4.0) توزیع شده است.

www.SID.ir

Archive of SID

چرخشی و با هدف دستیابی به شاخص‌های فصلی و سالانه نیروی کار و تغییرات آن در سطح کل کشور و استان‌ها، به صورت فصلی اجرا شده است. علاوه بر نرخ‌های اشتغال و بیکاری و همچنین برآوردهای تعداد بیکاران و شاغلان که مرکز آمار ایران به صورت فصلی و سالانه اعلام می‌کند به‌دلیل روش نمونه‌گیری استفاده شده در این آمارگیری ارائه آمار تغییر وضع فعالیت بین فصول نیز امکان‌پذیر است. نمونه‌گیری چرخشی با ثابت نگه داشتن بخشی از واحدهای نمونه بین دو فصل آمارگیری و تغییر بقیه واحدها، به نحو مطلوب امکان برآورد سطوح و تغییرات را فراهم می‌کند. انجام بهینه نمونه‌گیری چرخشی، مستلزم بکارگیری نمونه پایه است. منظور از نمونه پایه، نمونه‌ای است که می‌توان از آن برای تأمین نیازهای چند آمارگیری یا چند دوره از یک آمارگیری، زیرنمونه‌هایی انتخاب کرد (گرامی و همکاران، ۱۳۸۱). روش نمونه‌گیری در آمارگیری نیروی کار، نمونه‌گیری خوشه‌ای دو مرحله‌ای با طبقه‌بندی است که واحد نمونه‌گیری مرحله اول، خوشه و واحد نمونه‌گیری مرحله دوم، گروهی شامل سه خانوار معمولاً مجاور است. الگوی چرخش ۲-۲-۲ یعنی از هر خانوار نمونه چهار بار آمارگیری به عمل می‌آید، به این ترتیب که خانوار، دو فصل متوالی در نمونه است سپس به طور موقت برای دو فصل متوالی از نمونه خارج می‌شود، بعد مجدداً برای دو فصل متوالی به نمونه باز می‌گردد و پس از آن برای همیشه از نمونه خارج می‌شود. در حقیقت آدرس خانوار واحد نمونه گیری است. در نتیجه اگر خانوار نمونه تغییر محل داده باشد این خانوار دنبال نمی‌شود بلکه خانواری که جدیداً در آن آدرس ساکن شده در نمونه قرار می‌گیرد. میزان تداخل نمونه بین دو فصل متوالی و دو فصل یکسان از دو سال متوالی، ۵۰ درصد و بین دو سال متوالی، ۵۵ درصد است. برای ارائه برآوردها، داده‌های مربوط به هر یک از خانوارهای نمونه و اعضای آن‌ها در سه مرحله «اعمال وزن پایه (عکس احتمال انتخاب)»، «تعدیل وزن برای بی‌پاسخی» و «تعدیل وزن بر اساس پیش‌بینی‌های جمعیتی» وزن‌دهی می‌شوند. وزنی که پس از اعمال هر سه مرحله وزن‌دهی به دست می‌آید تعیین می‌کند که «هر فرد در نمونه، نماینده چند نفر در جامعه» است. پس از وزن‌دهی، برآورد تعداد از جمع وزن‌ها و برآورد نرخ از تقسیم برآوردهای تعداد به دست می‌آید. مقایسه این برآوردها در زمان‌های مختلف، تغییرات بازار کار را طی زمان نشان می‌دهند اما جابه‌جایی‌هایی که در داخل این مقاطع زمانی رخ می‌دهد با این شاخص‌ها مشخص نمی‌شود. به‌منظور آگاهی از جابه‌جایی‌های داخل مقاطع یا به‌عبارتی اندازه و ویژگی‌های جمعیتی افرادی که وضعیت نیروی کار آن‌ها طی زمان در نظر گرفته شده، تغییر کرده است، آمارهای جریان نیروی کار محاسبه می‌شود.

از آنجا که مرکز آمار ایران در آمارگیری نیروی کار از نمونه‌های چرخشی استفاده نموده، امکان محاسبه آمارهای جریان فراهم شده است. علی رغم کاربردهای بسیار آمار جریان [1] در بحث تحلیل نیروی کار هنوز یک مشکل وجود دارد و آن خطای پاسخ است که ممکن است به رده‌بندی نادرست افراد به داخل رده‌های نیروی کار منجر شود. خطای پاسخ به دلایل بسیاری به وجود می‌آید، مانند درک نادرست سؤال در ثبت اطلاعات توسط مأمور، نداشتن دانش کافی و آگاهانه یا به عمد پاسخ نادرست دادن. در تحقیقات بین‌المللی مشاهده شده است که برای سؤال‌های وضع فعالیت اقتصادی، خطاها سیستماتیک نیستند و وقتی که به مجموعه داده‌های مقطعی توجه شود، می‌توان از این خطاها چشم‌پوشی کرد. اما وقتی افراد در دوره‌های آمارگیری مختلف متصل می‌شوند (آمارهای جریان)، این خطاها به یک تغییر در وضع فعالیت اقتصادی منجر می‌شوند که غلط است و از این‌رو جریان‌های ناخالص بین هر وضعیت

[1] Flow Statistic

www.SID.ir

Archive of SID

را اشتباه نشان می‌دهد (اداره آمار کار، ۲۰۱۰). این خطا، منبع کم‌برآوردی قطر اصلی ماتریس انتقال نیروی کار و به تبع بیش‌برآوردی عناصر غیر قطر این ماتریس است که لازم است کاربران را از آن مطلع ساخت. همان گونه که کاربران با دانش از خطاهای نمونه‌گیری به تحلیل داده‌ها می‌پردازند می‌بایست از خطاهای رده‌بندی در آمارهای جریان نیز آگاه باشند. یکی از مهمترین روش‌هایی که برای خطاهای رده‌بندی ارائه شده، روش تعدیل خطای حاشیه‌ای است که توسط آبود و زلنر (۱۹۸۵) بیان شده است. در این روش، مدلی برای تعدیل جریان ناخالص ارائه می‌شود. همچنین فولر و چائو (۱۹۸۵)، تصحیحی بر اساس خطاهای اریبی گروه چرخش و رده‌بندی را با هم پیشنهاد کردند. بعد از آن هم بررسی‌ها و مطالعات بسیاری انجام شد ولی همچنان ارجاعات به روش آبود و زلنر متداول‌تر بود. از آنجایی که همواره دسترسی به مقادیر واقعی امکان‌پذیر نیست، روش‌های دیگری برای ارزیابی تغییرات پاسخ‌های واحدهای نمونه به عنوان معیاری از خطای اندازه‌گیری مورد استفاده قرار می‌گیرد. یکی از این روش‌ها، برآورد خطای اندازه‌گیری با استفاده از روش‌های مبتنی بر مدل است. (علی‌اکبری صبا و همکاران، ۱۳۹۴). روش دیگر، مصاحبه مجدد با فرض اجرای مجدد آمارگیری با خطای بسیار کمتر از اجرای اصلی است. در آمریکا از سال ۱۹۵۴ برای به دست آوردن خطاهای مقطع آمارهای نیروی کار مصاحبه مجدد انجام می‌شود فاینبرگ و تانور (۲۰۱۶). مشکل اصلی مصاحبه مجدد هزینه اجرای قابل توجه و فراهم نمودن شرایط ایده‌آل در اجرای آمارگیری است (آمارگیر خبره، نظارت دقیق) که در عمل اجرای سالانه آن برای ایران امکان‌پذیر نیست.

در آمارگیری‌های طولی، مدل‌های تحلیل رده نهان که توسط بایمر (۲۰۱۱) پیشنهاد شده است برای تفکیک تغییرات واقعی از تغییرات جعلی استفاده می‌شود. بایمر از الگوریتم بیشینه‌سازی امید ریاضی (EM) برای برآورد پارامترها در این مدل‌ها به منظور تحلیل خطای اندازه‌گیری در آمارگیری‌های پانلی استفاده کرده است. برای آمارگیری نیروی کار می‌توان فرض کرد متغیر نهان همان جواب واقعی است که در عمل نداریم در این صورت برای برآورد خطای رده‌بندی آمارهای جریان نیروی کار با توجه به الگوی چرخش نمونه‌ها، می‌توان مدل‌های تحلیل رده نهان را بکار برد. همان طور که باسی و همکاران در برآورد جابجایی نیروی کار ایتالیا از رویکرد رده نهان آمیخته مارکوف با نشانگرهای چندگانه استفاده نمودند (باسی و همکاران، ۲۰۱۷).

در این مقاله مدل‌های تحلیل رده نهان با فرض‌های مختلف به‌عنوان جایگزینی برای روش مصاحبه مجدد با هدف برآورد خطاهای رده‌بندی در آمارهای جریان نیروی کار ایران به‌کار گرفته شده است. در بخش ۲ آمارهای جریان نیروی کار درایران ارائه می شود. در بخش ۳ روش‌شناسی خطاهای رده‌بندی، رویکرد رده نهان مارکوف با فروض مختلف ارائه شده است. در بخش ۴ خطاهای رده‌بندی در آمارهای جریان نیروی کار ایران با استفاده از مدل‌های رده نهان مارکوف با چهار فرض مختلف و با بکارگیری برنامه ℓEM و لحاظ الگوی چرخش نمونه‌ها در آمارگیری نیروی کار ایران برای سال‌های ۱۳۹۸ و ۱۳۹۹ ارائه شده است. در انتها به بحث ونتیجه‌گیری پرداخته شده است.

www.SID.ir

Archive of SID

۲ آمار جریان نیروی کار در ایران

از سال ۱۳۸۴ آمارگیری نیروی کار ایران با استفاده از نمونه‌گیری چرخشی با هدف دستیابی به شاخص‌های فصلی و سالانه نیروی کار در سطح کل کشور و استان‌ها اجرا می‌شود. الگوی چرخش در این آمارگیری ۲-۲-۲ است که در مقدمه توضیح داده شده است. با استفاده از نتایج این آمارگیری نه تنها برآورد شاخص‌های وضع فعالیت مقاطع زمانی که برآورد تغییر وضع فعالیت‌های افراد در طول زمان یا به‌عبارتی آمار جریان نیز قابل برآورد است. جدول ۱ ساختار یک ماتریس انتقال را به صورت ساده نشان می‌دهد مهران (۲۰۱۰). همانطور که ملاحظه می‌شود در ماتریس

جدول ۱. ماتریس انتقال نیروی کار

داده‌های نیروی کار بر اساس زمان گذشته و زمان فعلی		زمان فعلی		
	$E(t-1)$	EE	EU	EN
زمان گذشته	$U(t-1)$	UE	UU	UN
	$N(t-1)$	NE	NU	NN

انتقال با یک حرکت مواجه هستیم. حرت از فصل اول به فصل دوم، منظور از E [1] تعداد شاغل‌ها، U[2] بیکارها، و N [3] غیر فعال‌ها است. منظور از EE تعداد افرادی است که در دو زمان شاغل بوده‌اند و EU، تعداد افراد شاغل در زمان گذشته که در زمان فعلی بیکار شده‌اند را نشان می‌دهد. منظور از EN، تعداد افراد شاغل در زمان گذشته است که غیر فعال شده‌اند، به همین ترتیب نیز بقیه خانه‌ها تعریف می‌شوند (فلاح محسن‌خانی و همکاران، ۱۳۹۰). با استفاده از آمارهای جریان نیروی کار، احتمالات ورود به اشتغال، بیکاری، غیر فعالی و خروج از اشتغال، بیکاری، غیر فعالی و ثبات وضعیت در اشتغال، بیکاری و غیر فعالی محاسبه می‌شود. این احتمالات از تقسیم جمعیتی که از وضعیت مورد نظر در فصل قبل به وضعیت مورد نظر در فصل جاری رسیده‌اند بر افرادی که در فصل قبل وضعیت مورد نظر را داشته‌اند، به‌دست می‌آید. ساختار ماتریس احتمال، متناظر با ساختار ماتریس انتقال بوده و به‌صورت فصلی قابل محاسبه است.

۳ روش‌شناسی خطاهای رده‌بندی

برای استفاده از آمارهای جریان لازم است مقدار خطاهای رده‌بندی این آمارها برآورد شوند، از روش‌های مختلفی برای ارائه خطاهای رده‌بندی در آمارهای جریان می‌توان استفاده کرد که در این قسمت به ارائه روش‌های کاربردی و عملی بر اساس الگوی نمونه‌گیری چرخشی ایران می‌پردازیم. یکی از مهم‌ترین روش‌هایی که برای خطاهای رده‌بندی ارائه شده، روش تعدیل خطای حاشیه‌ای است که توسط آبود و زلنر (۱۹۸۵) بیان شده است، در این روش، مدلی برای تعدیل

[1] Employment
[2] Unemployment
[3] Not In Labor Force

www.SID.ir

Archive of SID

جریان ناخالص ارائه می‌شود. در این روش نیاز است مقادیر واقعی در دسترس باشند، در عمل داده‌های مصاحبه مجدد با فرض اجرای مجدد آمارگیری با خطای بسیار کمتر از اجرای اصلی به جای مقایر واقعی در نظر گرفته می‌شوند. برای افزایش دقت در مصاحبه مجدد نسبت به آمارگیری اولیه لازم است از مصاحبه‌گران خبره استفاده و فرآیند نظارت نیز توسط بهترین ناظران انجام شود. امکان انجام مصاحبه مجدد به دلایل مختلف همه ساله وجود ندارد، از جمله اشکالات اجرای آن، بودجه اجرا و دسترسی به مصاحبه‌گران و ناظران خبره در تمام استان‌ها است (مهران و همکاران، ۱۳۹۱). برای برآورد خطای اندازه‌گیری داده‌های طولی می‌توان از مدل‌های تحلیل رده نهان استفاده کرد (بایمر، ۲۰۱۱). در واقع این روش می‌تواند برای تفکیک تغییرات واقعی از تغییرات جعلی در آمارگیری‌های طولی استفاده شود. بایمر (۲۰۱۱) از الگوریتم بیشینه‌سازی امید ریاضی (EM) برای برآورد پارامترها در این مدل‌ها به منظور تحلیل خطای اندازه‌گیری در آمارگیری‌های طولی استفاده کرده است. در این روش منظور از متغیر نهان همان جواب واقعی است که در عمل از آن اطلاع نداریم.

۳.۱ خطای رده‌بندی با استفاده از مدل رده نهان مارکوف

در آمارگیری‌هایی که‌به‌صورت پانلی انجام می‌شود، می‌توان از مدل رده نهان مارکوف[1] (MLC) برای اندازه‌گیری تغییرات پاسخ در دوره‌های مختلف استفاده کرد. مدل مارکوف نهان، یک روش تحلیلی برای برآورد رده‌بندی در آمارگیری‌های پانلی است که از تکرار پاسخ‌ها در آمارگیری‌های پانلی برای استخراج اطلاعات در مورد خطای رده‌بندی از داده‌های مصاحبه، بهره می‌گیرد. برقراری شرط مارکوف مرتبه اول (احتمال شرطی رخداد آینده فقط به آخرین رخداد موجود یعنی رخداد کنونی وابسته باشد) برای استفاده از این روش الزامی است. مزیت مهم روش MLC برای برآورد خطای اندازه گیری این است که تنها نیاز به داده‌های پانلی دارد و به داده‌های مصاحبه مجدد یا سایر منابع مکمل نیاز ندارد. بنابراین از آن جایی که در نمونه‌گیری نیروی کار از الگوی چرخش استفاده می‌شود می‌توان برآورد خطای رده‌بندی در آمارهای جریان ایران را با این روش محاسبه نمود.

فرض کنید هدف از مطالعه، اندازه‌گیری متغیر X است اما در واقعیت ممکن است مقادیر مشاهده شده با مقادیر واقعی متغیر X تفاوت داشته باشند. با استفاده از مدل‌های تحلیل رده نهان[2] (LCA) می‌توان میزان خطا در اندازه‌گیری متغیر X را ارزیابی نمود. مقدار واقعی متغیر X را به عنوان یک متغیر نهان[3] که قابل مشاهده نیست در نظر می‌گیریم. همچنین فرض می‌کنیم که مقدار مشاهده شده برای متغیر X با متغیرهای نشانگر نمایش داده می‌شود. برای هر متغیر X حداقل یک متغیر نشانگر مانند A، وجود دارد. اگر X به روش‌های متفاوتی اندازه‌گیری شود، مقادیر مشاهده شده برای متغیر X بیش از یک مقدار است که هر کدام از آن‌ها توسط یک متغیر نشانگر مجزا مانند B و C نمایش داده می‌شود. مدل LCA روش مناسبی را برای مدل‌بندی ارتباط بین مقدار واقعی متغیر و مقدار مشاهده شده آن فراهم می کند.

[1] Markov Latent Class
[2] Latent Class Analysis
[3] Latent Variable

www.SID.ir

Archive of SID

برای مشاهدات، متغیرهایی مانند A ، B و C به عنوان متغیرهای آشکار در نظر گرفته می‌شوند. اگر متغیرهای آشکار به طور خاص برای اندازه‌گیری X طراحی شده باشند، آن‌ها را متغیر نشان‌گر X گویند. فرض کنید که سه متغیر نشان‌گر A، B و C برای اندازه‌گیری متغیر نهان X در نظر گرفته شده‌اند. تابع درستنمایی برای جدول سه‌طرفه ABC در اندازه‌گیری متغیر نهان X به‌صورت

$$\pi_{abc}^{ABC} = \sum_{x} \pi_{x}^{X} \pi_{(a|x)}^{(A|X)} \pi_{(b|x)}^{(B|X)} \pi_{(c|x)}^{(C|X)}$$

نوشته می‌شود، که حالتی خاص از مدل‌های LCA است. فرض کنید $\pi_x^X = P(X = x)$ و احتمال شرطی متغیر نشان‌گر به‌شرط متغیر نهان $\pi_{(a|x)}^{(A|X)}$ به عنوان احتمال پاسخ[1] در نظر گرفته شود. در این حالت اگر $a \neq x$ باشد، $\pi_{(a|x)}^{(A|X)}$ احتمال خطا[2] ، و اگر $a = x$ باشد، $\pi_{(a|x)}^{(A|X)}$ احتمال درستی[3] است.

فرض‌های مدل LCA استاندارد

۱. نمونه n واحدی که از نمونه تصادفی ساده بدون جایگذاری از یک جامعه بزرگ به‌دست آمده است.

۲. متغیرهای نشان‌گر به ازای کلیه مقادیر متغیرهای نشان‌گر a، b، و c و مقدار x دارای استقلال موضعی هستند، یعنی $\pi_{abc}^{ABC} = \pi_{(a|x)}^{(A|X)} \pi_{(b|x)}^{(B|X)} \pi_{(c|x)}^{(C|X)}$.

۳. احتمالات شرطی $\pi_{(a|x)}^{(A|X)}$، $\pi_{(b|x)}^{(B|X)}$ و $\pi_{(c|x)}^{(C|X)}$ همگن هستند، یعنی مقادیر آن‌ها برای همه واحدهای نمونه یکسان است.

۴. متغیرهای نشان‌گر A، B و C همگی متغیر نهان X را اندازه‌گیری می‌کنند.

معمولاً در آمارگیری‌های پانلی، یک اندازه‌گیری واحد از X در چندین نقطه زمانی ثبت می‌شود. مدل LC شناسایی‌پذیر برای هر آمارگیری پانلی با حداقل سه دوره مناسب است. فرض کنید X_t به مقدار متغیر تصادفی X در زمان t، $t = ۱, \ldots, T$ اشاره می‌کند. بر طبق ویژگی مارکوف برای $t \geq ۲$ داریم:

$$P(X_t = x_t | X_۱ = x_۱, \ldots, X_{t-۱} = x_{t-۱}) = P(X_t = x_t | X_{t-۱} = x_{t-۱}). \qquad (۱)$$

در این مدل مقدار X در زمان t تنها وابسته به X در زمان $t - ۱$ منظور شده است. گرچه این فرض در بسیاری از کاربردها محتمل است، اما می‌توان آنرا برای وضعیت‌های دارای بیش از سه دوره، تعدیل کرد.

[1] Response Probability
[2] Error Probability
[3] Accuracy probability

www.SID.ir

Archive of SID

۳.۱.۱ مدل‌های مارکوف

برای درک مفهوم مدل‌های رده نهان مارکوف (MLC)، ابتدا مدل‌های مارکوف را بدون متغیرهای نهان، تحت عنوان مدل‌های مارکوف آشکار[1] MM را در نظر می‌گیریم. فرض کنید که اندازه‌گیری‌های X بدون خطای اندازه‌گیری باشد (گودمن، ۱۹۷۳). چنانچه خواهیم دید مدل‌های MM مؤلفه‌های ساختاری مدل‌های MLC است. در حقیقت، مدل‌های MLC مدل‌های MM هستند که در آن X با خطا ثبت شده‌اند (بایمر، ۲۰۱۱). در ادامه با اضافه کردن یک مؤلفه اندازه‌گیری، چارچوب مدل‌بندی MMبه مدل‌های MLC تعمیم داده خواهد شد.

۳.۱.۲ مدل‌های مارکوف آشکار

فرض کنید که X متغیر رسته‌ای مورد نظر باشد که در هر دوره بدون خطا ثبت شده است. برای هر واحد دلخواه در جامعه، فرض کنید x_t مقدار مشاهده شده X_t، $t = ۱, \ldots, T$ باشد. وضعیت واحد در دوره t، با x_t نشان داده شده است. فرض کنید K تعداد وضعیت‌هایی است که واحد در زمان t می‌تواند بگیرد. همچنین، فرض کنید که تعداد وضعیت‌ها در همه زمان‌ها یکسان باشد. تغییرات از وضعیت $x_{t-۱}$ به x_t را انتقال و $P(X_t = x_t|X_{t-۱} = x_{t-۱})$ را احتمال انتقال گویند و با $\pi_{x_t|x_{t-۱}}^{X_t|X_{t-۱}}$ نشان می‌دهند. فرض کنید سه دوره پیاپی در آمارگیری پانلی وجود دارد ($T = ۳$). همچنین $n_{x_۱x_۲x_۳}$ به تعداد افرادی که در دوره اول در وضعیت $x_۱$، در دوره دوم در وضعیت $x_۲$ و در دوره سوم در وضعیت $x_۳$ هستند، اشاره می‌کند. داده‌ها در یک جدول سه طرفه مانند جدول ۲ قرار می‌گیرد. فرض کنید $\pi_{x_۱x_۲x_۳}^{X_۱X_۲X_۳}$ به احتمال اینکه یک مشاهده متعلق به خانه $(x_۱, x_۲, x_۳)$ باشد،

جدول ۲. رده‌بندی متقاطع وضعیت اشتغال در جدول سه طرفه اشتغال (E)، بیکاری (U) و غیر فعال بودن (N)

		دوره سوم									
		N			U			E			
	دوره دوم	N	U	E	N	U	E	N	U	E	
دوره اول	E	$n_{۱۳۲}$	$n_{۱۲۳}$	$n_{۱۱۳}$	$n_{۱۳۲}$	$n_{۱۲۲}$	$n_{۱۱۲}$	$n_{۱۳۱}$	$n_{۱۲۱}$	$n_{۱۱۱}$	
	U	$n_{۲۳۳}$	$n_{۲۲۳}$	$n_{۲۱۳}$	$n_{۲۳۲}$	$n_{۲۲۲}$	$n_{۲۱۲}$	$n_{۲۳۱}$	$n_{۲۲۱}$	$n_{۲۱۱}$	
	N	$n_{۳۳۳}$	$n_{۳۲۳}$	$n_{۳۱۳}$	$n_{۳۳۲}$	$n_{۳۲۲}$	$n_{۳۱۲}$	$n_{۳۳۱}$	$n_{۳۲۱}$	$n_{۳۱۱}$	

اشاره می‌کند. عام‌ترین شکل مدل به صورت

$$\pi_{x_۱x_۲x_۳}^{X_۱X_۲X_۳} = \pi_{x_۱}^{X_۱} \pi_{x_۲|x_۱}^{X_۲|X_۱} \pi_{x_۳|x_۱x_۲}^{X_۳|X_۱X_۲}, \qquad (۲)$$

[1] Manifest Markov

www.SID.ir

Archive of SID

است. با در نظر گرفتن فرض ویژگی مارکوف مدل به صورت

$$\pi^{X_۱X_۲X_۳}_{x_۱x_۲x_۳} = \pi^{X_۱}_{x_۱}\pi^{X_۲|X_۱}_{x_۲|x_۱}\pi^{X_۳|X_۲}_{x_۳|x_۲} \qquad (۳)$$

در نظر گرفته می‌شود و به آن مدل MM استاندارد گویند. یک زنجیر مارکوف مانا است اگر

$$\pi^{X_{t+۱}|X_t}_{x|x'} = \pi^{X_{t+۲}|X_{t+۱}}_{x|x'} \quad \forall t, x, x'.$$

پارامترهای احتمال در مدل (۲) می‌تواند به طور معادل به صورت یک مدل لگ‌خطی برای فراوانی‌های مورد انتظار، $m_{x_۱x_۲x_۳}$، به صورت

$$\log(m_{x_۱x_۲x_۳}) = u + u^{X_۱}_{x_۱} + u^{X_۲}_{x_۲} + u^{X_۳}_{x_۳} + u^{X_۱X_۲}_{x_۱x_۲} + u^{X_۱X_۳}_{x_۱x_۳} + u^{X_۲X_۳}_{x_۲x_۳} + u^{X_۱X_۲X_۳}_{x_۱x_۲x_۳},$$

در نظر گرفته شود، که در آن برای پارامترهای u قیدهای

$$\sum_{x_t} u^{X_t}_{x_t} = \sum_{x_t} u^{X_tX_{t'}}_{x_tx_{t'}} = \sum_{x_t} u^{X_۱X_۲X_۳}_{x_۱x_۲x_۳} = ۰, \quad t = ۱, ۲, ۳, \quad t \neq t',$$

فرض می‌شود. توجه کنید که در این وضعیت مدل سلسله مراتبی می‌تواند به صورت استاندارد $\{X_۱X_۲X_۳\}$ نوشته شود. به همین ترتیب، مدل (۳) با حذف اثرهای متقابل سه تایی به صورت

$$\log(m_{x_۱x_۲x_۳}) = u + u^{X_۱}_{x_۱} + u^{X_۲}_{x_۲} + u^{X_۳}_{x_۳} + u^{X_۱X_۲}_{x_۱x_۲} + u^{X_۱X_۳}_{x_۱x_۳} + u^{X_۲X_۳}_{x_۲x_۳},$$

قابل ارائه است که برای سادگی به صورت $\{X_۱X_۲ \;\; X_۱X_۳ \;\; X_۲X_۳\}$ با قید مجموع مساوی صفر نوشته می‌شود. با قرار دادن $u^{X_۲}_{x} = u^{X_۳}_{x}$، $u^{X_۱X_۲}_{x_۱x_۲} = u^{X_۲X_۳}_{x_۲x_۳}$ و $(x_۱, x_۲) = (x_۲, x_۳)$ فرض مانایی برای همه مقادیر x، $x_۱$، $x_۲$ و $x_۳$ برقرار است و احتمال‌های مربوط به هر یک از مدل‌ها به صورت

$$\begin{aligned}
\pi^{X_۱}_{x_۱} &= \frac{\exp(u^{X_۱}_{x_۱})}{\sum_{x_۱}\exp(u^{X_۱}_{x_۱})},\\
\pi^{X_۲|X_۱}_{x_۲|x_۱} &= \frac{\exp(u^{X_۲}_{x_۲} + u^{X_۱X_۲}_{x_۱x_۲})}{\sum_{x_۲}\exp(u^{X_۲}_{x_۲} + u^{X_۱X_۲}_{x_۱x_۲})},\\
\pi^{X_۳|X_۲}_{x_۳|x_۲} &= \frac{\exp(u^{X_۳}_{x_۳} + u^{X_۲X_۳}_{x_۲x_۳})}{\sum_{x_۳}\exp(u^{X_۳}_{x_۳} + u^{X_۲X_۳}_{x_۲x_۳})}.
\end{aligned}$$

محاسبه خواهند شد، که در آنها مقادیر u با برآوردهای ML جایگزین خواهند شد (بایمر، ۲۰۱۱).

www.SID.ir

Archive of SID

۳.۱.۳ مدل‌های رده نهان مارکوف

مدل‌های رده نهان مارکوف ابتدا توسط ویگینز (۱۹۷۳) پیشنهاد شد، با این وجود پولسن (۱۹۸۲)، وندِ پال و دِ لیو (۱۹۸۶) و وندِ پال و لنگ‌هاین (۱۹۹۰) سهم بسزایی در تعمیم آن در مسائل کاربردی داشته‌اند. تحلیل رده نهان مارکوف ($MLCA$) یک مجموعه از روش‌ها بر مبنای مدل MLC است که می‌تواند برای تفکیک تغییرات واقعی از تغییرات جعلی در آمارگیری‌های طولی استفاده شود. بایمر و بوشری (۲۰۰۰) از این روش برای تحلیل خطای اندازه‌گیری در اندازه‌گیری‌های آمارگیری پانلی رسته‌ای استفاده کردند. مدل $MLCA$ فرض می‌کند که مشاهدات X در معرض خطای رده‌بندی قرار دارد. مانند آنچه در LCA بحث شد، $MLCA$ با X مانند یک متغیر نهان در مدل‌بندی رفتار می‌کند. احتمال‌های خطای رده‌بندی که بر اندازه‌های X در هر دوره تأثیر می‌گذارد، هنگامی قابل برآورد است که دو یا بیشتر از دو اندازه مکرر در یک یا بیشتر از یک دوره وجود داشته باشد. شبیه به مدل LC ، مدل MLC شامل دو مؤلفه است: (۱) مؤلفه ساختاری که وابستگی‌های داخلی بین X_t، $t = ۱, \ldots, T$ و متغیرهای کمکی را توصیف می‌کند و رابطه (۲) مؤلفه خطا که وابستگی درونی بین مشاهدات A_t، B_t، C_t و غیره در هر دوره را مشخص و اثر متقابل آن‌ها با X_t و دیگر متغیرهای کمکی مدل را توصیف می‌کند. با این وجود چنانچه خواهیم دید، مدل‌های MLC انعطاف‌پذیرتر از مدل‌های LC هستند زیرا برای برآورد احتمال‌های خطا برای شناساپذیری به تکرار اندازه‌گیری‌ها نیازی نیست.

فرض‌های پایه مدل MLC شامل انتقال‌های مارکوفی، استقلال خطاهای رده‌بندی بین دوره‌ها و زمان همگنی احتمال‌های خطا است. با حداقل سه دوره پانلی، $MLCA$ برآوردهایی را از احتمال‌های شیوع دوره ۱، احتمال‌های انتقال دوره t به $t+۱$ و احتمال‌های خطای رده‌بندی مربوط به A_t که تنها اندازه‌گیری X_t است، ارائه می‌دهد. مدل‌های استاندارد MLC فرض می‌کند که برخی ویژگی‌های مورد علاقه، X، برای $T \geq ۳$ دوره پانلی اندازه‌گیری شده است.

۳.۱.۴ الگوریتم EM برای مدل‌های MLC

برای وضعیتی با سه دوره و K رده نهان، تعداد پارامترهای مدل MLC برابر $(K-۱)(۳K+۱)$ است. بنابراین برای داده‌های دودویی مدل شامل ۷ پارامتر (علاوه بر میانگین کلی) است. کاربرد الگوریتم EM را برای یک مدل MLC ساده با سه دوره بیان می‌شود. داده‌های مشاهده شده در یک جدول توافقی با متغیرهای A در سه دوره به‌صورت $A_۱A_۲A_۳$ نشان داده می‌شود. حال فرض کنید که X_t به مقدار مشاهده واقعی A_t در دوره t اشاره می‌کند. اگر X_t برای هر آزمودنی در هر دوره معلوم باشد، داده‌های کامل را می‌توان به فرم $X_۱X_۲X_۳A_۱A_۲A_۳$ که شامل $K^۶$ خانه است و K به تعداد رده‌های نهان اشاره می‌کند، در نظر گرفت. فرض کنید $(x_۱, x_۲, x_۳, a_۱, a_۲, a_۳)$ به یک خانه دلخواه از جدول اشاره می‌کند و $\pi^{X_۱X_۲X_۳A_۱A_۲A_۳}_{x_۱x_۲x_۳a_۱a_۲a_۳}$ به احتمال تعلق به آن خانه اشاره می‌کند. تابع لگاریتم

www.SID.ir

Archive of SID

درستنمایی داده‌های کامل در این وضعیت به صورت

$$\ell = \sum n_{x_۱,x_۲,x_۳,a_۱,a_۲,a_۳} \log \sum \pi^{X_۱X_۲X_۳A_۱A_۲A_۳}_{x_۱x_۲x_۳a_۱a_۲a_۳},$$

است، که در آن $n_{x_۱,x_۲,x_۳,a_۱,a_۲,a_۳}$ به تعداد مشاهدات در خانه $(x_۱, x_۲, x_۳, a_۱, a_۲, a_۳)$ اشاره می‌کند. تحت فرض‌های مدل مارکوف استاندارد داریم:

$$\pi^{X_۱X_۲X_۳A_۱A_۲A_۳}_{x_۱x_۲x_۳a_۱a_۲a_۳} = \left(\pi^{X_۱}_{x_۱}\pi^{X_۲|X_۱}_{x_۲|x_۱}\pi^{X_۳|X_۲}_{x_۳|x_۲}\right)\left(\pi^{A_۱|X_۱}_{a_۱|x_۱}\pi^{A_۲|X_۲}_{a_۲|x_۲}\pi^{A_۳|X_۳}_{a_۳|x_۳}\right),$$

و برای همه $a_۱ = a_۲ = a_۳$ و $x_۱ = x_۲ = x_۳$ طبق قید احتمال‌های خطای همگن

$$\pi^{A_۱|X_۱}_{a_۱|x_۱} = \pi^{A_۲|X_۲}_{a_۲|x_۲} = \pi^{A_۳|X_۳}_{a_۳|x_۳},$$

است. الگوریتم EM شامل دو گام در حلقه‌ای تکرار شونده است که عبارتند از: الف) گام E (امیدگیری) و ب) گام M(بیشینه‌سازی). در گام E، $\hat{n}_{x_۱x_۲x_۳|a_۱a_۲a_۳} = n_{a_۱a_۲a_۳}\hat{\pi}^{X_۱X_۲X_۳|A_۱A_۲A_۳}_{x_۱x_۲x_۳|a_۱a_۲a_۳}$ می‌تواند محاسبه شود، که $n_{a_۱a_۲a_۳}$ تعداد مشاهدات در خانه $(a_۱, a_۲, a_۳)$ و $\hat{n}_{x_۱x_۲x_۳a_۱a_۲a_۳}$ تعداد برآورد شده برای خانه $(x_۱, x_۲, x_۳, a_۱, a_۲, a_۳)$ است. این تعداد در گام M برای یافتن برآورد جدید $\pi^{X_۱X_۲X_۳|A_۱A_۲A_۳}_{x_۱x_۲x_۳|a_۱a_۲a_۳}$ استفاده می‌شود (فلاح محسن‌خانی و همکاران، ۱۳۹۶).

۳.۱.۵ نرم‌افزار برای تحلیل داده‌های رسته‌ای

برنامه ℓEM ورمانت (۱۹۹۷) برای مدل‌های خطی و تحلیل سابقهٔ پیشامد با داده‌های گمشده با استفاده از لگاریتم EM تهیه شده است و برای تحلیل داده‌های رسته‌ای اسمی، ترتیبی و یا سطوح بازه‌ای است که می‌تواند برای برآورد پارامترها استفاده شود. از آن جا که استفاده از این برنامه به کارگیری مدل‌های رده نهان مارکوف را بدون نیاز به برنامه نویسی جدید مقدور می‌سازد، و برای برآورد خطاهای رده‌بندی در داده‌های طولی نیز توصیه شده است (بایمر، ۲۰۱۱)، در این مقاله نیز برای ارائه خطاهای رده‌بندی از این نرم‌افزار استفاده شده است.

۴ خطاهای رده‌بندی در آمارهای جریان ایران

در این قسمت خطاهای رده‌بندی در آمارهای جریان نیروی کار ایران با استفاده از مدل‌های رده نهان مارکوف با چهار فرض مختلف و با بکارگیری برنامه ℓEM و لحاظ الگوی چرخش نمونه‌ها در آمارگیری نیروی کار ایران با استفاده از داده های آمارگیری نیروی کار سال‌های ۱۳۹۸ و ۱۳۹۹ محاسبه می‌شود.

www.SID.ir

Archive of SID

۴.۱ مدل‌های رده نهان مارکوف با احتمال‌های انتقال برابر و اطلاعات سه دوره

در مدل‌های رده نهان مارکوف اشاره شد، سه دوره متوالی برای مشاهدات در نظر گرفته شده و برآورد پارامترها بر این اساس به دست آمده است. در صورتی که الگوی چرخش نمونه‌ها در ایران از الگوی ۲-۲-۲ پیروی می‌کند. در الگوی چرخش به کار گرفته شده توسط ایران بعد از تکرار خانوارها در دو فصل متوالی، خانوارها برای دو فصل از نمونه خارج می‌شوند و دوباره برای دو فصل وارد نمونه می‌شوند به عبارتی گام‌های بین فصول مشترک یکسان نیستند. نمونه‌های مشترک در دو فصل اول پشت هم هستند در صورتی که نمونه‌های مشترک بار دوم تکرار با نمونه‌های مشترک بار سوم تکرار، دو فصل فاصله دارد. برای مثال فرض کنیم خانوارهایی که در زمستان ۱۳۹۲ وارد نمونه شده‌اند در بهار ۱۳۹۳ نیز در نمونه هستند و بعد از دو فصل خروج موقت از نمونه در فصل‌های زمستان ۱۳۹۳ و بهار ۱۳۹۴ وارد نمونه می‌شوند. ولی با این وجود، مدل با لحاظ فرض یکسان بودن احتمال انتقال بین دوره‌ها یعنی، $P(X_2|X_1) = P(X_3|X_2)$ تعریف شده است. فرض برابری احتمال‌های انتقال برای الگوی چرخش ۲-۲-۲ منطبق با واقعیت اجرای آمارگیری نیست زیرا برای مثال احتمال بیکار به بیکار بودن در دو فصل متوالی با دو فصل با فاصله شش ماه یکسان به دست آمده است. بنابراین برای مدل‌بندی با سه فصل متوالی شرط برابر بودن احتمال‌های انتقال را نمی‌توان در نظر گرفت و باید شرط نابرابری احتمال‌های انتقال در نظر گرفته شود.

۴.۲ مدل‌های رده نهان مارکوف با احتمال‌های انتقال نابرابر واطلاعات سه دوره

در این سناریو از اطلاعات سه دوره استفاده می‌شود ولی احتمال‌های انتقال دوره‌ها را نابرابر فرض می‌شوند. در برنامه ℓEM لحاظ کردن فرض نابرابری احتمال‌های انتقال $P(X_2|X_1) \neq P(X_3|X_2)$ نیز در نظر گرفته شده است. در دومین سناریو، فرض نابرابری احتمال‌های انتقال بین دوره‌ها تعریف شده است. نبود فرض برابری احتمال‌های انتقال منجر به افزایش پارامترهای مدل می‌شود و وابستگی نتایج به دست آمده به مقدار اولیه در نظر گرفته شده در الگوریتم EM می‌شود. این اشکال در برآوردهای دوره‌های متفاوت نیز مشاهده می‌شود. یعنی اگر برآوردهای احتمال‌های انتقال چند دوره محاسبه شوند نیز ملاحظه می‌شود برآوردها تفاوت قابل ملاحظه‌ای با یکدیگر دارند. بنابراین نتایج حاصل از این سناریو نیز پذیرفته نمی‌شود. می‌توان با ایده برابری احتمال‌های انتقال، سناریو دیگری که با الگوی چرخش ایران سازگار باشد، ارائه کرد.

۴.۳ مدل‌های رده نهان مارکوف با احتمال‌های انتقال برابر و اطلاعات چهار دوره

در این سناریو از فرض احتمال‌های انتقال برابر استفاده می‌شود ولی با راهکاری مشکل عدم برابری گام‌ها بین دوره‌ها را برطرف می‌کنیم. طبق الگوی چرخش ۲-۲-۲ برای هر مشاهده چهار تکرار وجود دارد ولی فاصله زمانی بین تکرارها یکسان نیست. دو تکرار اول دو فصل پشت سرهم، تکرار دوم و سوم شش ماه فاصله زمانی و تکرار سوم و چهارم مجدد دو فصل پشت سر هم قرار می‌گیرند. یک فرض اضافه در مدل در نظر گرفتیم و آن این است که فرض می‌کنیم بین دوره‌های دوم و سوم تغییر وضعیتی رخ نداده است بنابراین در داده‌های اصلی مشاهداتی را در نظر

www.SID.ir

Archive of SID

می‌گیریم که وضعیت‌های فصل دوم و سوم یکسان هستند. این فرض اولا سبب می‌شود مشکل گام‌های بین دوره‌ها حل شود و مهمتر این که بسیاری از تغییر وضعیت‌هایی که در بین تکرارهای دوم و سوم رخ می‌دهد، به علت طولانی بودن زمان بین دو دوره واقعی است و نباید به عنوان خطا لحاظ شود. بنابراین در سومین سناریو با لحاظ برابری احتمال‌های انتقال بین دوره‌ها، مشاهدات مربوط به چهار دوره در سه دوره خلاصه می‌شود. بنابراین می‌توان فرض نابرابری احتمال‌های انتقال بدون خطای رده‌بندی را با چهار دوره تعریف کرد.

۴.۴ مدل‌های رده نهان مارکوف با احتمال‌های انتقال نابرابر واطلاعات چهار دوره

سناریوی چهارم، استفاده از ایده در نظر گرفتن احتمال‌های انتقال نابرابر است ولی اطلاعات بر اساس چهار دوره آمارگیری است. در این سناریو با چهار دوره آمارگیری در نظر گرفته شده سه احتمال انتقال داریم که احتمال انتقال اول با احتمال انتقال سوم برابر و با احتمال انتقال دوم نابرابر در نظر گرفته می‌شود.

$$P(X_2|X_1) = P(X_4|X_3), \quad P(X_2|X_1) \neq P(X_3|X_2).$$

همان‌طور که ذکر شد در مدل و برنامه ℓEM احتمال‌های انتقال متفاوت در نظر گرفته شد ولی در تئوری مدل یکسان نبودن فواصل زمانی دوره‌های انتقال لحاظ نشده است. یعنی در مدل نمی‌توان مشخص نمود که احتمال‌های انتقال در زمستان۱۳۹۲ به بهار ۱۳۹۳ و زمستان ۱۳۹۳ به بهار ۱۳۹۴ برای دو فصل متوالی است ولی احتمال‌های انتقال بهار ۱۳۹۳ به زمستان ۱۳۹۳، برای دو فصل با فاصله زمانی شش ماه است به عبارتی وقتی فاصله زمانی دو دوره شش ماه در نظر گرفته می‌شود احتمال این که بسیاری از تغییر وضعیت‌ها واقعی باشد بیشتر می‌شود، در صورتی که این مساله قابل تعریف در مدل نیست.

سناریوی سوم بهترین انتخاب بر اساس الگوی چرخش نمونه‌های آمارگیری نیروی کار مرکز آمار ایران است. زیرا فرض در آن با الگوی چرخش ۲-۲-۲ که دو فصل خانوار در نمونه حضور دارد، دو فصل حضور ندارد و مجدداً در فصل‌های پنجم و ششم در نمونه حاضر است، سازگارتر است.

برای ملاحظه اثرات فصلی، آمارهای جریان با شروع فصل‌های مختلف نیز بررسی می‌شوند. داده‌های سال ۱۳۹۸ با چهار شروع مختلف به صورت تابستان ۹۹-بهار ۹۹-تابستان ۹۸-بهار ۹۸، پاییز ۹۹- تابستان ۹۹-پاییز ۹۹- تابستان ۹۸، زمستان ۹۹- پاییز ۹۹-زمستان ۹۸- پاییز ۹۸ و بهار ۱۴۰۰- زمستان ۹۹- بهار ۹۹- زمستان ۹۸ در نظر گرفته می‌شود. از آن جایی که هر فرد در هر دوره وضعیت‌های شاغل، بیکار و غیرفعال را می‌تواند اختیار کند بنابراین در چهار دوره ۸۱ وضعیت داریم. حال افراد مشترک برای هر یک از چهار دوره استخراج می‌شود (n_{ijkl}) سپس افرادی که در دوره‌های دوم و سوم تغییر وضعیت دادند از مشاهدات خارج می‌شوند بدین صورت چهار دوره به سه دوره با ۲۷ وضعیت تقلیل می‌یابد. این مشاهدات ورودی‌های برنامه ℓEM در نظر گرفته می‌شوند. با فرض $P(X_2|X_1) = P(X_3|X_2)$ ماتریس احتمال‌های انتقال بعد از تصحیح خطای رده‌بندی و ماتریس احتمال‌های تغییرات فصلی برای این داده‌ها عبارت است از: تفاوت مقادیر روی قطر اصلی جداول ۳ و ۴ خطای رده‌بندی برای

www.SID.ir

Archive of SID

جدول ۳. ماتریس احتمال‌های انتقال بعد از تصحیح خطاهای رده‌بندی در بهار به تابستان ۱۳۹۸

		تابستان		
		E	U	N
بهار	E	۰٫۹۴۹	۰٫۰۱	۰٫۰۴۱
	U	۰٫۱۹۱	۰٫۶۷۱	۰٫۱۳۸
	N	۰٫۰۱۶	۰٫۰۰۴	۰٫۹۷۹

جدول ۴. ماتریس احتمال تغییرات فصلی وضعیت نیروی کار در بهار به تابستان ۱۳۹۸

		تابستان		
		E	U	N
بهار	E	۰٫۸۸۸	۰٫۰۲۳	۰٫۰۸۹
	U	۰٫۳۱۲	۰٫۴۸۷	۰٫۲۰۱
	N	۰٫۰۶۳	۰٫۰۱۷	۰٫۹۲۰

وضعیت‌های شاغل به شاغل، بیکار به بیکار و غیرفعال به غیرفعال را نشان می‌دهد. این خطا برای وضعیت شاغل به شاغل ۶ درصد (۰٫۹۴۹-۰٫۸۸۸) ، بیکار به بیکار حدود ۱۸ درصد و غیرفعال به غیرفعال حدود ۶ درصد است. بنابراین بالاترین خطا مربوط به وضعیت بیکار به بیکار است. تفاوت مقادیر روی قطر اصلی جداول ۵ و ۶ خطای

جدول ۵. ماتریس احتمال‌های انتقال بعد از تصحیح خطاهای رده‌بندی در تابستان به پاییز ۱۳۹۸

		پاییز		
		E	U	N
تابستان	E	۰٫۹۴۳	۰٫۰۱۰	۰٫۰۴۷
	U	۰٫۱۷۸	۰٫۰۶۵	۰٫۱۷۲
	N	۰٫۰۱۸	۰٫۰۰۴	۰٫۹۷۸

جدول ۶. ماتریس احتمال تغییرات فصلی وضعیت نیروی کار در تابستان به پاییز ۱۳۹۸

		پاییز		
		E	U	N
تابستان	E	۰٫۸۷۶	۰٫۰۲۸	۰٫۰۹۵
	U	۰٫۰۳۰۶	۰٫۴۷۷	۰٫۲۱۶
	N	۰٫۰۶۵	۰٫۰۱۶	۰٫۹۲۰

رده‌بندی را نشان می‌دهد که برای وضعیت‌های شاغل به شاغل حدود ۷ درصد، برای بیکار به بیکار حدود ۱۷ درصد و برای غیرفعال به غیرفعال حدود ۶ درصد است. تفاوت مقادیر روی قطر اصلی جداول ۷ و ۸ خطای رده‌بندی را

www.SID.ir

Archive of SID

جدول ۷. ماتریس احتمال‌های انتقال بعد از تصحیح خطاهای رده‌بندی در پاییز به زمستان ۱۳۹۸

		زمستان		
		N	U	E
پاییز	E	۰٫۰۴۷	۰٫۰۰۸	۰٫۹۴۵
	U	۰٫۱۴۸	۰٫۷۴	۰٫۱۱۲
	N	۰٫۹۸	۰٫۰۰۴	۰٫۰۱۶

جدول ۸. ماتریس احتمال تغییرات فصلی وضعیت نیروی کار در پاییز به زمستان ۱۳۹۸

		زمستان		
		N	U	E
پاییز	E	۰٫۱۰۷	۰٫۰۳۵	۰٫۸۵۸
	U	۰٫۲۲۹	۰٫۵۲۳	۰٫۲۴۸
	N	۰٫۹۴	۰٫۰۱۲	۰٫۰۴۸

نشان می‌دهد که برای وضعیت‌های شاغل به شاغل حدود ۹ درصد، برای بیکار به بیکار حدود ۲۲ درصد و برای غیرفعال به غیرفعال حدود ۴ درصد است. تفاوت مقادیر روی قطر اصلی جداول ۹ و ۱۰ خطای رده‌بندی را نشان

جدول ۹. ماتریس احتمال‌های انتقال بعد از تصحیح خطاهای رده‌بندی در زمستان ۱۳۹۸ به بهار ۱۳۹۹

		بهار ۱۳۹۹		
		N	U	E
زمستان ۱۳۹۸	E	۰٫۰۳۸	۰٫۰۰۷	۰٫۹۴۵
	U	۰٫۱۱۹	۰٫۶۴۴	۰٫۲۳۷
	N	۰٫۹۷۹	۰٫۰۰۳	۰٫۰۱۷

جدول ۱۰. ماتریس احتمال تغییرات فصلی وضعیت نیروی کار در زمستان ۱۳۹۸ به بهار ۱۳۹۹

		بهار ۱۳۹۹		
		N	U	E
زمستان ۱۳۹۸	E	۰٫۱۰۸	۰٫۰۳۵	۰٫۸۵۷
	U	۰٫۲۶۵	۰٫۴۱۶	۰٫۳۱۸
	N	۰٫۹۳۶	۰٫۰۱۱	۰٫۰۵۳

می‌دهد که برای وضعیت‌های شاغل به شاغل حدود ۹ درصد، برای بیکار به بیکار حدود ۲۳ درصد و برای غیرفعال به غیرفعال حدود ۴ درصد است. بنابراین بالاترین خطا مربوط به وضعیت بیکار به بیکار است.

همان‌طور که جداول ۳ تا ۱۰ نشان می‌دهد بالاترین خطا مربوط به وضعیت بیکار به بیکار و کمترین خطا مربوط

www.SID.ir

Archive of SID

به وضعیت غیرفعال به غیرفعال است و این مسئله در فصول مختلف سال نیز مشاهده می‌شود. به‌عبارتی شروع فصل‌های مختلف در خطای رده‌بندی آمارهای جریان تاثیر چشم‌گیری ندارد.

آمارهای جریان بر اساس نمونه‌های انتخاب شده از نمونه پایه است. بنا بر این خطای رده‌بندی آمارهای جریان را در نمونه‌های برگرفته از نمونه پایه دیگری نیز می‌بایست محاسبه و نتایج را بررسی کرد. بدین منظور از نمونه پایه مربوط به سال‌های ۱۳۸۸ تا ۱۳۹۱ احتمال‌های انتقال بعد از تصحیح خطاهای رده‌بندی برای سال‌های ۱۳۸۹ و ۱۳۹۰ نیز استخراج شد (فلاح محسن‌خانی و همکاران، ۱۳۹۶) و مشاهده شد که خطای رده‌بندی آمارهای جریان در این سال‌ها نیز روندی مشابه با سال ۱۳۹۸ و ۱۳۹۹ که از نمونه پایه دیگری برگرفته است، داراست.

بحث و نتیجه‌گیری

در آمارگیری نیروی کار مرکز آمار ایران که از سال ۱۳۸۴ به صورت فصلی اجرا می‌شود از روش نمونه‌گیری چرخشی استفاده می‌شود. این روش نمونه‌گیری علاوه بر برآوردهای مقاطع زمانی، اطلاعات مربوط به جریان نیروی کار را نیز ارائه می‌نماید. اهمیت این آمارها برای برنامه‌ریزان نیروی کار نه تنها کم‌تر از آمارهای مقطع نیست بلکه در مواردی به مراتب بیش‌تر است. در استخراج این آمارها ابتدا می‌بایست افراد مشترک در دوره آمارگیری را مشخص و اطلاعات‌شان را در دو دوره با هم متصل نمود. طبق الگوی چرخش در ایران، نیمی از افراد نمونه در دو فصل متوالی مشترک هستند، بنابراین برآورد آمارهای جریان بااستفاده از نتایج آمارگیری نیروی کار امکان پذیر است. با توجه به نیاز و نگاه کاربران، جداول مختلفی برای آمارهای جریان نیروی کار می‌توان محاسبه نمود و بر اساس این جداول، تحلیل‌های مفید و متنوعی می‌توان انجام داد. خطای غیرنمونه‌گیری از خطاهای تأثیرگذار در آمارهای جریان است. یکی از مهم‌ترین خطاهای غیرنمونه‌گیری در این آمارها، خطای پاسخ است که ممکن است به رده‌بندی نادرست افراد به داخل رده‌های نیروی کار منجر شود. بخشی از خطای رده‌بندی در داده‌های هر دوره آمارگیری معمولاً خنثی می‌شود ولی غالباً در آمارهای جریان این خطاها مضاعف می‌شود. برای بکار بردن این آمارها بهتر است خطای رده‌بندی آن‌ها نیز بررسی و مشخص شود تا با اطلاع از حدود این خطا تحلیل‌های برگرفته از این آمارها انجام شوند. معروفترین روش برای محاسبه خطای رده‌بندی آمارهای جریان انجام مصاحبه مجدد است که روشی پر هزینه و دشوار است. در این مقاله ابتدا نحوه محاسبه آمارهای جریان ارائه و سپس مدل‌های رده نهان مارکوف که برای اندازه‌گیری خطاهای رده‌بندی در داده‌های طولی معرفی شده‌اند، بیان می‌شود. سرانجام برای آمارگیری نیروی کار ایران با توجه به الگوی چرخش نمونه‌ها روش‌های مختلف بحث و بررسی می‌شود، و در خاتمه روش مناسب محاسبه خطای رده‌بندی آمار جریان نیروی کار ایران معرفی و برای آخرین سال موجود محاسبه می‌شود. بنابراین با توجه به روش منتخب معرفی شده برای محاسبه خطای رده‌بندی در آمار جریان که در مقاله ارائه شده است، امکان ارائه آمارهای نیروی کار به همراه خطای رده‌بندی آن‌ها بدون صرف هزینه برای اجرای مصاحبه مجدد میسر است.

www.SID.ir

Archive of SID

تقدیر و تشکر

نویسندگان از نکات ارزنده داوران، سردبیر و ویراستار محترم مجله که باعث بهبود هرچه بیشتر مقاله شده‌اند کمال تشکر و قدردانی را دارند.

مراجع

علی‌اکبری صبا، ر.، زاهدیان، ع. و اربابی، م. (۱۳۹۴). روشی برای برآورد خطای اندازه‌گیری درآمد، مجله علوم آماری، نشریه علمی پژوهشی انجمن آمار ایران ،۹(۱)، ۱۰۰-۷۷.

فلاح محسن‌خانی، ز.، جوادی، م.، رفیع، س. و مهران، ف. (۱۳۹۰)، آمارهای جریان حاصل از آمارگیری نیروی کار، پژوهشکده آمار، تهران.

فلاح محسن‌خانی، ز.، حسین‌پوری، م. و کلهری ندرآبادی، ل. (۱۳۹۶)، خطاهای طبقه‌بندی در آمارهای جریان ، پژوهشکده آمار، تهران.

گرامی، ع.، سخاوت، م.ب.، هرندی، ف. و فلاح محسن‌خانی، ز. (۱۳۸۱)، نمونه پایه و کاربرد آن در آمارگیری‌ها به ویژه آمارگیری‌های خانواری، پژوهشکده آمار، تهران.

مهران، ف.، فلاح محسن‌خانی، ز.، جوادی، م.، قائمی، ذ.، آبی، ن.، بنی‌هاشمی، ف.س.، حق‌گو، ن.، زنگنه، م.، طاهری، ق.، عرفانی، ج.، فتاحی، م.، نکویی، ا.، نوراللهی، ط. و یزدانی، ا. (۱۳۹۱)، بررسی خطاهای آماری آمارگیری نیروی کار مرکز آمار ایران، پژوهشکده آمار، تهران.

Abowd, J. M. and Zellner, A. (1985), Estimating Gross Labor-Force Flows, *Journal of Business & Economic Statistics*, **3**, 254-283.

Bassi, F., Croon, M. and Vidotto, D. (2017), A Mixed Latent Class Markov Approach for Estimating Labour Market Mobility with Multiple Indicators and Retrospective Interrogation, *Survey Methodology*, **43**, 107-124.

Biemer, P. (2011), *Latent Class Analysis of Survey Error*, John Wiley & Sons.

Biemer, P. and Bushery, J. (2000), On the Validity of the Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data, *Survey Methodology*, **26**, 139-152.

www.SID.ir

Archive of SID

Fienberg, S. E., and Tanur, J. M. (2016), *Current Population Survey*, US Bureau of Labor Statistics Washington, DC., 1-11.

Fuller, W. and Chua, T. (1985), Gross Change Estimation in the Presence of Response Error, In *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, US Bureau of Labor Statistics Washington, DC., 65-77.

Goodman, L. A.(1973), Causal Analysis of Data from Panel Studies and other Kinds of Surveys, *American Journal of Sociology*, **78**, 1135-1191.

Mehran, F. (2010), *Youth Employment: A Training Module on Data Requirements and Analysis*, ILO, Geneva.

Poulsen, C. A. (1982), *Latent Structure Analysis with Choice Modeling Applications*, Aarhus School of Business Administration and Economics, Arhus, Denmark.

Statistics, U. B. O. L. (2010), Research Series on the Labor Force Status Flows from the Current Population Survey.

Van de Pol ,F, and Langeheine, R. (1990), Mixed Markov Latent Class Models , in C. C. logg, ed., Sociological Methodology , Blackwell , Oxford.

Van de Pol, F. and De Leeuw, J. A. N. (1986), A Latent Markov Model to Correct for Measurement Error, *Sociological Methods and Research*, **15**, 118-141.

Vermunt, J. (1997), Log -Linear Models for Event Histories, Sage Publications. Thousand Oaks, CA.

Wiggins, L. M .(1973), Panel Analysis, Latent Probability Models for Attitude and Behavior Processing, Elsevier SPC, Amsterdam.

www.SID.ir