

بسم الله الرحمن الرحيم

روش‌های رگرسیونی برای تحلیل داده‌های با ابعاد بزرگ

(به کمک روش‌های MCMC و با استفاده از نرم‌افزار R)

تألیف:

دکتر سید کامران قریشی

سرشناسه	: قریشی، سید کامران، ۱۳۵۰ - Ghoreishi, Seyed Kamran
عنوان و نام پدیدآور	: روش‌های رگرسیونی برای تحلیل داده‌های با ابعاد بزرگ (به کمک روش‌های MCMC و با استفاده از نرم‌افزار R) / مؤلف سید کامران قریشی.
مشخصات نشر	: قم: دانشگاه قم، انتشارات، ۱۳۹۷.
مشخصات ظاهری	: ۲۴۰ ص: جدول.
شابک	: 978-600-8436-33-1
وضعیت فهرست نویسی	: فیبا
یادداشت	: کتابنامه.
موضوع	: داده‌های کلان -- روش‌های آماری
موضوع	: Big data -- Statistical methods
موضوع	: آمار ریاضی -- داده‌پردازی
موضوع	: Mathematical statistics -- Data processing
موضوع	: آر (زبان برنامه‌نویسی کامپیوتر)
موضوع	: R (Computer program language)
شناسه افزوده	: دانشگاه قم. انتشارات
رده بندی کنگره	: ۱۳۹۷ ق/ ۴/ ۲۵/ ۶/ ۹/ QAV
رده بندی دیویی	: ۷۴/ ۰۰۵
شماره کتابشناسی ملی	: ۵۳۹۴۷۹۱



انتشارات دانشگاه قم

عنوان: روش‌های رگرسیونی برای تحلیل داده‌های با ابعاد بزرگ

مؤلف: دکتر سید کامران قریشی

ویراستار: دکتر عطار دسات مصطفوی‌نیا

چاپ اول: پاییز ۱۳۹۷

شمارگان: ۵۰۰

چاپ و صحافی: هوشنگی

طراح جلد: حیدری

ناظر فنی: علیرضا معظمی

بهاء: ۲۰۰۰۰۰ ریال

شابک: ۹۷۸-۶۰۰-۸۴۳۶-۳۳-۱

ISBN: 978-600-8436-33-1

آدرس الکترونیکی: Publication@ Qom. ac.ir

کلیه حقوق مادی و معنوی برای ناشر محفوظ است.

قم، بلوار الغدیر، دانشگاه قم، اداره چاپ و انتشارات دانشگاه

تلفن: ۰۲۵-۳۳۴۴۰۳۳۱۰۳۳۴۵-۲۵ نمابر: ۰۲۵-۳۳۱۰۳۳۴۵

کاش ز کویت خبری داشتیم کاش به پای تو سری داشتیم
کاش که مانند خراباتیان ناله صاحب اثری داشتیم
تابه هوای تو بگیریم اوج کاش که مابال و پری داشتیم

اللهم عجل لولیک الفرج

فهرست مطالب

۱	مفاهیم پایه‌ای	۱
۱	توزیع‌های احتمال	۱.۱
۸	امید ریاضی	۲.۱
۱۰	فضاهای برداری	۳.۱
۱۵	ماتریس و خواص آن	۴.۱
۲۰	تمرین‌ها	۵.۱
۲۳	مقدمه‌ای بر شبیه‌سازی و تولید نمونه تصادفی	۲
۲۴	مقدمه	۱.۲
۲۴	تولید نمونه تصادفی از توزیع یکنواخت	۲.۲
۲۷	آزمون تصادفی بودن نمونه‌های تولید شده	۳.۲
۲۸	آزمون نیکویی برازش برای نمونه‌های تولید شده	۴.۲
۲۹	روش خی‌دوی پیرسون	۱.۴.۲
۳۰	روش خی‌دوی نسبت درستیابی	۲.۴.۲
۳۲	تولید نمونه از توزیع‌های آماری	۵.۲
۳۲	تولید نمونه از توزیع‌های پیوسته	۱.۵.۲
۳۴	تولید نمونه از توزیع‌های گسسته	۲.۵.۲
۳۶	روش‌های عمومی تولید نمونه از توزیع‌های دلخواه	۶.۲
۳۶	روش ER	۱.۶.۲
۴۲	روش AR	۲.۶.۲
۴۵	تولید نمونه تصادفی از توزیع‌های خاص	۷.۲
۶۹	انتگرال‌گیری به روش مونت کارلو	۸.۲
۷۳	تمرین‌ها	۹.۲

۷۷	مدل‌های رگرسیونی و رگرسیون منظم شده	۳
۷۸	۱.۳ رگرسیون خطی	
۸۰	۲.۳ رگرسیون خطی چند گانه	
۸۳	۳.۳ ملاک‌های انتخاب مدل در رگرسیون خطی چندگانه	
۸۴	۱.۳.۳ ضریب تعیین تعدیل شده R_a^2	
۸۵	۲.۳.۳ ملاک اطلاع آکائیکه	
۸۷	۳.۳.۳ ملاک اطلاع بیزی	
۸۷	۴.۳ رگرسیون لگ-خطی	
۹۳	۵.۳ رگرسیون لوژستیک	
۹۶	۶.۳ انواع رگرسیون خطی منظم شده	
۹۷	۱.۶.۳ رگرسیون ستیغی منظم شده	
۱۰۱	۲.۶.۳ رگرسیون لاسوی منظم شده	
۱۰۷	۳.۶.۳ رگرسیون شبکه ارتجاعی منظم شده	
	۴.۶.۳ تفاوت‌ها و شباهت‌های رگرسیون ستیغی، لاسو و شبکه ارتجاعی منظم شده	
۱۰۸	۵.۶.۳ الگوریتمی برای برآورد پارامترهای انواع رگرسیون منظم شده	
۱۱۰	۶.۶.۳ رگرسیون موزون منظم شده	
۱۱۲	۷.۶.۳ رگرسیون لوژستیک منظم شده	
۱۱۵	۸.۶.۳ رگرسیون پواسون منظم شده	
۱۲۱	۷.۳ تمرین‌ها	
۱۲۴		
۱۲۷	تحلیل داده‌های با ابعاد بزرگ	۴
۱۲۷	۱.۴ مقدمه	
۱۳۱	۲.۴ روش‌های رگرسیونی داده‌های با ابعاد بزرگ	
۱۳۲	۱.۲.۴ روش غربالگری	
۱۳۳	۲.۲.۴ الگوریتم حریص	
۱۳۴	۳.۲.۴ روش رگرسیون منظم با تابع جریمه محدب	
۱۳۸	۳.۴ روش‌های برآورد پارامترها در مدل رگرسیونی با نرم L_1	
۱۳۸	۱.۳.۴ روش پروگزیمال	
۱۴۳	۲.۳.۴ روش کاهش بُعد	

۴.۴	مفاهیم پایه در مدل رگرسیونی با بردار پارامتر تُنک	۱۴۶
۱.۴.۴	مدل رگرسیونی خطی بدون خطا	۱۴۷
۲.۴.۴	مدل رگرسیونی خطی با خطا	۱۵۳
۵.۴	بازه اطمینان و آزمون فرض برای پارامترهای مدل رگرسیونی	۱۵۶
۶.۴	تمرین‌ها	۱۶۵

۵ رگرسیون تعمیم یافته منظم شده و خواص برآوردکننده‌ها در داده‌های با

۱۶۷	ابعاد بزرگ
۱.۵	مقدمه
۲.۵	تابع جریمه گوی L_q
۳.۵	برآوردهای M -منظم شده
۴.۵	برآوردهای M -مدل رگرسیونی لاسوی تعمیم یافته
۵.۵	تابع زیان در مدل رگرسیونی لاسوی تعمیم یافته
۶.۵	تمرین‌ها

۶ مونت کارلوی زنجیره مارکوفی و کاربرد آن برای تحلیل داده‌های با ابعاد

۱۸۵	بزرگ
۱.۶	روش‌های عمومی MCMC
۱.۱.۶	روش متروپولیس-هاستینگس
۲.۱.۶	روش گیبز
۳.۱.۶	روش برشی
۲.۶	تحلیل بیزی به کمک روش‌های MCMC
۳.۶	تحلیل بیزی مدل رگرسیونی لاسو
۴.۶	برآورد پارامتر تنظیم در رگرسیون لاسو به روش بیزی
۱.۴.۶	روش بیز تجربی برای برآورد پارامتر تنظیم
۲.۴.۶	روش فرض چگالی پیشین برای λ
۵.۶	چگالی پیشین انقباضی برای بردار پارامتر تُنک
۶.۶	تمرین‌ها

فهرست مطالب

ت

پیشگفتار

امروزه داده‌های فراوانی در زمینه‌های مختلف کاربردی نظیر ژنتیک، سیگنال‌ها و غیره تولید می‌شوند که از ابعاد بزرگ برخوردار بوده و روش‌های مرسوم و شناخته شده آماری نمی‌توانند برای تحلیل آنها به کار روند لذا لازم است از روش‌های جدیدتری برای تحلیل این نوع از داده‌ها استفاده کنیم. کتاب حاضر به معرفی روش‌هایی خواهد پرداخت که برای تحلیل رگرسیونی داده‌های با ابعاد بزرگ مورد استفاده قرار می‌گیرند.

به دلیل آن که این روش‌ها شامل هر دو دیدگاه فراوانی گرا و بیزی می‌باشند لذا در فصل ۲ بعضی از روش‌های تولید نمونه تصادفی، که اغلب برای تحلیل بیزی داده‌ها و بویژه برای تحلیل بیزی داده‌های با ابعاد بزرگ نیز به کار می‌روند، معرفی شده‌اند. فصول ۳-۵ به معرفی روش‌های فراوانی گرا و فصل ۶ به معرفی روش‌های بیزی برای تحلیل رگرسیونی داده‌های با ابعاد بزرگ اختصاص دارند.

هرچند در نوشتن این کتاب سعی وافر شده است که از لفاظی و یا کلی‌گویی خودداری شود لیکن کتاب خالی از اشکال نبوده و نویسنده با آغوش باز از نظرات و پیشنهادات کارشناسان، آمار شناسان و اساتید معزز برای اصلاح کتاب استقبال می‌کند.

در پایان لازم می‌دانم از خانواده عزیزم که با صبر و حوصله با من همراه بودند تشکر و قدردانی کنم.

سید کامران قریشی

آبان ۱۳۹۷