مقدمه

کتب متعددی در حوزه رگرسیون[1] و آنالیز واریانس[2] موجودند. این کتب نیازمند سطوح آمادگی متفاوتی هستند و به صورت‌های مختلفی بر محتوا تاکید دارند. این کتاب مقدماتی نیست و مطالعه آن نیازمند قدری دانش در حوزه تئوری آماری پایه و تمرین است.

از خوانندگان انتظار می‌رود که با ملزومات استنباط آماری مانند برآورد[3]، آزمون‌های فرض[4] و بازه‌های اطمینان[5] آشنا باشند. دانشی ابتدائی در حوزه تحلیل داده، قدری اطلاعات در مورد جبرخطی و محاسبات نیز مورد نیاز است.

تاکید این نوشتار بر تمرین رگرسیون و آنالیز واریانس است. هدف این است که بدانیم چه روش‌هایی وجود دارند و از آن مهم‌تر این روش‌ها را چه زمانی باید به کار برد. برای شفاف‌سازی کاربرد این روش‌ها و نمایاندن نتایج حاصل از آن‌ها، مثال‌های متعددی ارائه شده است. بر مباحث تئوری ریاضی تاکید چندانی نشده است، چراکه انتظار می‌رود خوانندگان تا حدی در این مورد اطلاعات قبلی داشته باشند و از طرفی مناسب‌تر است در جایی دیگر به این موضوعات پرداخته شود. مباحث تئوری حائز اهمیتند؛ چراکه رویکردهای انتخابیمان را هدایت می‌کنند. به تئوری آماری به طور عمیق‌تری پرداخته‌ایم. این تنها شامل قضایای مرسوم نیست. مفاهیم آماری کیفی نیز به همان میزان مهمند چراکه مارا قادر می‌سازند به جای تنها صحبت کردن در مورد موضوعی، دست به عمل بزنیم. یادگیری این قواعد کیفی سخت‌تر است، چراکه بیان دقیق آنها دشوار می‌باشد؛ اما این قواعد آمارگران موفق و با تجربه را یاری می‌نمایند.

تحلیل داده را نمی‌توان بدون تمرین آموخت؛ یعنی باید از یک بسته محاسباتی آماری استفاده شود. گزینه‌های زیادی برای انتخاب از میان این بسته‌ها وجود دارد. این بسته‌ها برای استفاده افراد مختلفی طراحی شده‌اند و دارای نقاط قوت و ضعف متفاوتی هستند. در اینجا ما R[6] را انتخاب کرده‌ایم. دلایل متعددی برای انتخاب R وجود دارد:

تطبیق‌پذیری: R یک زبان برنامه‌نویسی نیز است، در نتیجه در مورد فرآیندهای از پیش برنامه‌نویسی شده در یک بسته نرم‌افزاری دچار محدودیت نمی‌شویم. برنامه‌نویسی روش‌های جدید در R نسبتا ساده است.

---

[1] regression
[2] analysis of variance
[3] estimation
[4] hypothesis
test

[5] confidence interval

[6] Ref. Ihaka and Gentleman (1996) و تیم مرکزی توسعه R (۲۰۰۳)

تعامل: تحلیل داده به طور ذاتی فرآیندی تعاملی است. برخی از بسته‌های آماری قدیمی زمانی طراحی شدند که محاسبات گران‌تر بوده و پردازش دسته‌ای باب بود. با وجود پیشرفت سخت‌افزارها، پارادایم قدیمی پردازش دسته‌ای در استفاده از آنها باقی مانده است. R در هر قدم یک عمل انجام می‌دهد و ما را قادر می‌سازد که بر اساس آنچه حین تحلیل می‌بینیم تغییراتی اعمال کنیم.

آزادی: R بر پایه S-که بسته تبلیغاتی S-plus از آن مشتق گشته- ایجاد شده‌است. R نرم افزاری متن-باز[1] است و به صورت رایگان در اختیار افراد قرار می‌گیرد.[2] R و S-plus با یکدیگر سازگارند؛ بدین معنی که برای بسیاری از مثال‌های استفاده شده در این کتاب می‌توان از S-plus استفاده کرد.

محبوبیت: در کاربردهای عمومی، SAS رایج‌ترین بسته آماری است، اما R یا S میان محققان آماری محبوب‌ترند. با نگاهی به نشریات آماری این محبوبیت تایید می‌شود. R همچنین برای کاربردهای کمی در امور مالی دارای شهرت است.

شروع کار با R

یادگیری R نیازمند اندکی کوشش است. این کوشش با افزایش بهره‌وری پاسخ داده خواهد شد. شما می‌توانید طریقه دسترسی به R و همچنین نصب نرم‌افزارها و داده‌های فرعی دیگری که در این کتاب استفاده شده‌اند را در ضمیمه ۱ بیاموزید.

این کتاب مقدمه‌ای بر R نیست. در ضمیمه ۲ زبان برنامه‌نویسی به طور مختصر توضیح داده شده است، اما این به تنهایی کفایف نمی‌کند. تمامی فرمان‌هایی که برای تولید نتایج موجود در این کتاب استفاده شده‌اند به طور عمدی در متن قرار داده شده‌اند. این یعنی شما می‌توانید این تحلیل‌ها را دوباره تولید کرده و تغییرات را پیش از آشنایی کامل با R تجربه کنید.[3]

در انتها جا دارد از سازندگان این نرم‌افزار -که بدون آنها تهیه این کتاب امکان‌پذیر نبود- سپاس‌گزاری کنیم.[4]

---

[1] open-source software

[2] نسخه‌های لینوکس، مکینتاش، ویندوز و سایر نسخ UNIX را می‌توان از www.r-project.org تهیه کرد.

[3] می‌توانید برای دسترسی به راهنمای مقدماتی مجانی R به وبسایت www.r-project.org مراجعه نمایید.

[4] داده‌های ذکر شده در متن کتاب را می‌توانید در وبسایت www.stat.lsa.umich.edu/~faraway/LMR بیابید.