محمد سعادتی ۸۱۰۱۹۸۴۱۰

بخش اول – ماتریس همبستگی و دادههای گمشده

۲-

روش‌های مختلفی برای برخورد با چارچوب های اطلاعاتی و مشاهداتی که دارای داده گمشده ثبت یا نشان داده **NA** به صورت **R** هستند، وجود دارد. این مقدارها در محیط نرم‌افزاری .می‌شوند. دو رویکرد برای مدیریت داده گمشده می‌توان در نظر گرفت

- خروج مشاهدات یا متغیرها با مقدارهای گمشده از محاسبات و تحلیل‌های آماری
- جایگزنی داده‌های گمشده با مقدار جایگزین (مثلا میانگین یا میانه مقدارهای متغیر)

یکی از توابع موثر در **R** که برای کار روی داده گمشده مناسب است، تابع **()mutate** از کتابخانه **dplyr** است. کار با این تابع بسیار ساده بوده و پارامترهای محدودی دارد .به کمک تابع **()mutate** می‌توانید یک متغیر جدید براساس محاسبات تعیین شده، ایجاد کنید.

خارج کردن داده گمشده از مجموعه داده

اگر می‌خواهید داده گمشده از مجموعه اطلاعاتی خارج شود و در تحلیل‌های آماری نقشی نداشته باشد، کافی است از تابع **()na.omit**استفاده کنید.

جایگزین مقدار برای داده گمشده

در بعضی از تحلیل‌های آماری به دلیل کمبود مشاهدات، گاهی داده گمشده را با میانگین **(Mean)**یا میانه **(Median)** جایگزین می‌کنند تا تعداد نمونه، کاهش نیابد.

خلاصه

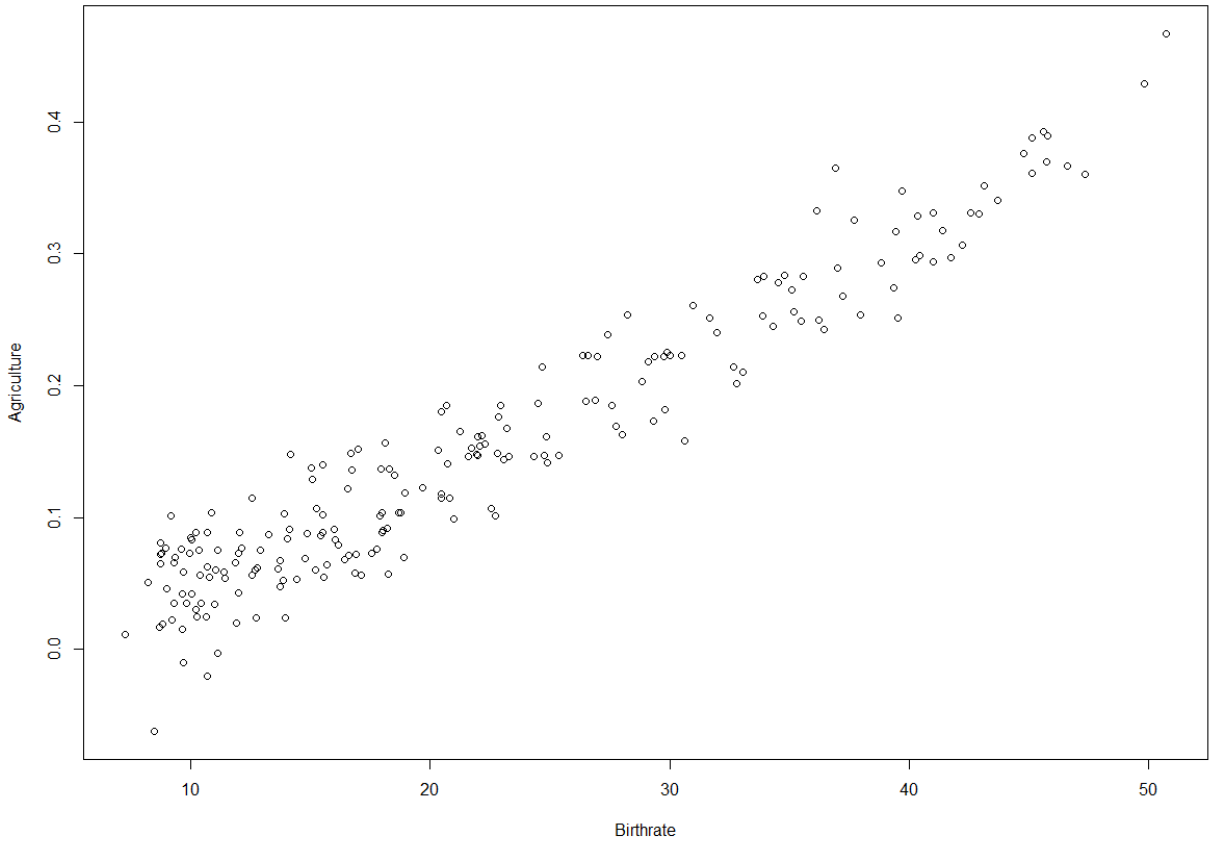
در این نوشتار، رویکرد ما برای داده گمشده به دو صورت بیان شد: «حذف (Deletion) «یا «جایگزینی (Imputation) «با مقدار دلخواه. جدول‌های زیر به بررسی توابع و عملکرد آن‌ها به این منظور پرداخته است.

حذف		
کتابخانه	شرح	کد
base	نمایش مشاهدات گمشده	colnames(df)[apply(df, ۲, anyNA)]
dplyr	حذف داده‌های گمشده از چارچوب داده (df)	na.omit(df)

جایگزینی			
تابع	شرح	مزایا	معایب
apply()	بررسی ستون‌ها با مقدار گمشده و محاسبه شاخص‌های آماری	وضوح محاسبات انجام شده و خروجی به صورت چارچوب داده	زمان زیاد برای اجرا روی چند ستون یا متغیر
sapply()	بررسی ستون‌ها با مقدار گمشده و محاسبه شاخص دلخواه	سرعت انجام محاسبات و اجرای همزمان برای چندین ستون	بدون ایجاد چارچوب داده

۳- ماتریس همبستگی داده ها در فایل 1.r ، print شده است

۴-



۵-

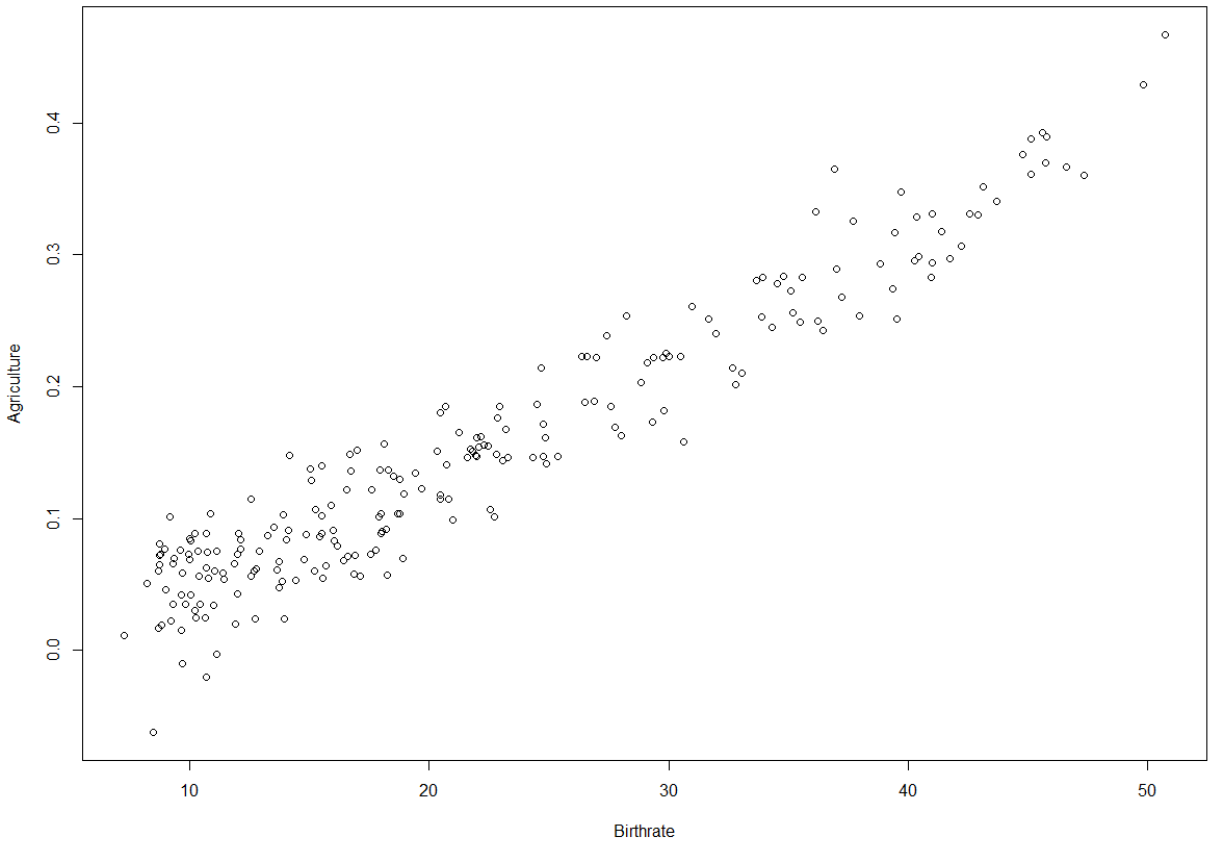
با توجه به برقراری رابطه خطی نسبی بین دو متغییر نرخ زاد و ولد و کشاورزی، برای تخمین مقادیر گمشده کشاورزی به این صورت عمل می کنیم:

ابتدا میانگین مقادیر این دو متغییر را بدون در نظر گرفتن داده هایی که برابر NA هستند ، حساب می کنیم

سپس برای محاسبه مقادیر گمشده کشاورزی، مقدار متناظر نرخ زاد و ولد مربوط به آن را ضرب در نسبت دو میانگینی که در مرحله قبل حساب کردیم می کنیم. یعنی روش محاسبه به این صورت است:

```
mean_of_Agriculture = mean(countries$Agriculture, na.rm =TRUE)
mean_of_Birthrate = mean(countries$Birthrate, na.rm =TRUE)
if (is.na(countries$Agriculture[i]))
countries$Agriculture[i] = countries$Birthrate[i] *
mean_of_Agriculture / mean_of_Birthrate
```

نمودار **scatter plot** نهایی به این شکل است:



بخش دوم – شبیه سازی متغیر تصادفی

۲- اگر خروجی تابع runif(1) کمتر از p = 0.6 باشد متغییر تصادفی با توزیع برنولی برابر 1 و در غیر اینصورت برابر 0 می شود

۳- برای ساخت یک متغییر تصادفی با توزیع دو جمله ای با n= 10، ابتدا ۱۰ متغییر تصادفی با توزیع برنولی می سازیم سپس این ۱۰ مقدار را با یکدیگر جمع می کنیم.

در ادامه تعداد ۱۰۰ نمونه از این متغییر تصادفی می سازیم و میانگین و واریانس انها را حساب می کنیم

در روابط تئوری داشتیم که میانگین توزیع دو جمله ای برابر np و واریانس توزیع دو جمله ای برابر npq است. حال می بینیم که نتایج بدست امده برابر با مقادیری است که از طریق روابط تئوری محاسبه می شود.

بخش سوم – تبدیل معکوس

۱-

$$X = F^{-1}(U) \xrightarrow{?} X \sim F \quad (\text{CDF of } X \text{ is } F)$$

اثبات:

$$P(X \le x) = P(F^{-1}(U) \le x) = P(U \le F(x)) = F(x)$$

راه دوم:

$$F(x) = P(X \le x) \rightarrow F(x) \overset{?}{=} P(F^{-1}(U) \le x)$$

$$U = F(F^{-1}(U)) \le F(x) \rightarrow F^{-1}(U) \le x \rightarrow$$

$$P(U \le F(x)) = F(x) \rightarrow P(F^{-1}(U) \le x) = F(x)$$

2- درستی رابطه مورد نظر با استفاده از قسمت قبل ثابت می شود

توزیع فراوانی به شکل زیر است: