فصل ۱:
پیش درآمد

۱.۱ در شروع کار

آمار با یک مسئله شروع می‌شود. با جمع آوری داده‌ها پیش می‌رود، با تجزیه‌وتحلیل داده‌ها ادامه می‌یابد و با نتیجه‌گیری پایان می‌گیرد. اشتباهی رایج میان آمارگرانِ بی‌تجربه این است که بدون توجه به اهداف، یا اینکه آیا داده‌ها مناسب تجزیه و تحلیل مورد نظر هستند، غرق در یک تحلیل پیچیده می‌شوند. پیش از خیز برداشتن باید نگریست!

تدوین صورت مسئله-که می‌تواند صرفاً یک مهارت محاسباتی یا تجربی باشد- غالباً اساسی‌تر از راه حل آن است. (آلبرت اینشتین)

برای تدوین صحیح صورت یک مسئله:

زمینه عینی آن باید درک شود. آمارگران غالباً در مشارکت با افراد دیگر کار می‌کنند و لازم است درکی از حوزه کاری داشته باشند. این باید به عنوان فرصتی برای یادگیری تلقی شود نه یک روال خسته کننده.

هدف باید درک شود. (باز هم) ممکن است همکاری با شخصی باشد که اهداف برایش روشن نباشد. از ماجراجوئی باید بر حذر بود؛ با مته بر خشخاش گذاشتن این امکان وجود دارد که چیزی یافت شود لیکن ممکن است تصادفی باشد.

باید اطمینان حاصل کرد که درخواست مشتری چیست. می‌توان بر روی یک مجموعه داده یکسان، تجزیه و تحلیل‌های کاملا متفاوتی انجام داد. گاهی اوقات آمارگران، تجزیه و تحلیلی به مراتب پیچیده‌تر از آنچه مشتری به آن نیاز داشته است، انجام می‌دهند. ممکن است که آمارهای توصیفی[۱] ساده، تنها چیزی باشد که مورد نیاز بوده‌است.

صورت مسئله باید در قالب آماری درآید. این یک مرحله‌ی چالش برانگیز است، مرحله‌ای که گاهی خطاهای جبران ناپذیری در آن رخ می‌دهد. هنگامی که مسئله به زبان آمار تبدیل شود، راه حل معمولا مشخص است. مشکلات مربوط به این مرحله نشان می‌دهد که چرا در زمینه آمار، روش‌های هوش مصنوعی[۲] هنوز موثر نیستند. برنامه ریزیِ کامپیوتریِ تعریف مسئله کاری دشوار است.

اینکه یک روش آماری قادر به خواندن و پردازش داده‌ها باشد کافی نیست. نتایج یک تجزیه و تحلیل نامتناسب می تواند بی معنا باشد.

---

[۱] descriptive statistics

[۲] Artificial intelligence

دانستن نحوه جمع آوری داده ها نیز اهمیت دارد.

آیا آن‌ها از طریق پیمایش نمونه‌ای طراحی شده[۱] جمع‌آوری شده‌اند؟ چگونگی جمع‌آوری داده‌ها بر اینکه چه نتیجه‌ای حاصل شود اثری تعیین کننده دارد.

آیا "بی پاسخ[۲]" وجود دارد؟ داده‌های کسب نشده ممکن است دارای اهمیتی برابر با داده‌های موجود باشند.

آیا "مقادیر گم شده[۳]" وجود دارد؟ این مسئله‌ای متداول، مشکل آفرین و زمانبر است.

کد گذاری داده‌ها به چه صورت انجام شده است؟ به ویژه اینکه متغیرهای کیفی چگونه نشان داده شده‌اند؟

واحدهای اندازه‌گیری چه هستند؟

باید مراقب خطا در ورود داده‌ها و دیگر انحرافات بود. این مشکل در هر مجموعه داده واقعی با حجمی حداقل متوسط بسیار شایع -تقریباً حتمی- است. چندین بررسی صحت داده‌ها باید انجام شود.

۱. ۲ تجزیه و تحلیل ابتدائی داده‌ها

تجزیه و تحلیل ابتدائی داده‌ها مرحله‌ای اساسی است که باید همیشه انجام پذیرد. به نظر ساده می‌رسد اما حیاتی است. خلاصه‌های عددی[۴] مانند میانگین[۵]، انحراف معیار[۶] (SD)، ماکزیمم و مینیمم، همبستگی[۷] و هر چیز دیگری مختص آن مجموعه داده خاص، باید ایجاد شود. خلاصه‌های نموداری[۸] به همان اندازه حائز اهمیتند. تنوع گسترده‌ای از روش‌ها برای انتخاب وجود دارد. برای هر متغیر به تنهایی می‌توان نمودار جعبه‌ای[۹]، هیستوگرام[۱۰]، نمودار چگالی[۱۱] ونمودارهای دیگری تهیه کرد. برای دو متغیر،

---

۱ designed sample survey

۲ Nonresponse

۳ missing values

۴ numerical summaries

۵ mean

۶ standard deviation

۷ correlation

۸ graphical summaries

۹ boxplot

۱۰ histogram

۱۱ density plot

ایجاد نمودار پراکندگی (پراکنش)[۱] استاندارد است؛ این در حالی است که برای بیشتر از حتی دو متغیر روش‌های خوب متعددی از جمله نمودارهای تعاملی و پویا[۲] برای عرضه داده‌ها وجود دارد. در نمودارها باید به دنبال نقاط دور افتاده[۳] (داده پرت)، خطا در ورود داده‌ها، توزیع‌ها و ساختارهای چوله[۴] یا غیرمعمول غیرمعمول بود. بررسی اینکه آیا داده‌ها بر اساس انتظارات قبلی توزیع شده‌اند نیز باید انجام پذیرد.

تبدیل داده‌ها به شکلی مناسب برای تجزیه و تحلیل از طریق پاک سازی خطاها و انحرافات، کاری زمان‌بر است. معمولا این کار از خود تجزیه و تحلیل داده‌ها نیز بیشتر زمان می‌گیرد. در این کتاب تمام داده‌ها برای تجزیه و تحلیل آماده خواهندشد، لیکن باید در نظر داشته باشید که این کار در عمل به ندرت انجام می‌شود.

مثالی را می‌نگریم. موسسه ملی بیماری‌های قند،کلیه و هاضمه[۵]، مطالعه‌ای بر روی ۷۶۸ زن بالغ بومی پیما[۶] که نزدیک فونیکس[۷] زندگی می‌کردند، انجام داد. متغیرهای زیر ثبت شدند:

تعداد دفعات بارداری، غلظت قند خون در طی دو ساعت در یک تست خوراکی نوسان گلوکز، فشار خون دیاستولیک[۸] (mmHg) ، ضخامت چین خوردگی پوست عضله سه سر بازو[۹](mm)، سرم انسولین دو ساعته[۱۰] (mu U/ml)، شاخص توده بدنی[۱۱] (bmi) (وزن (kg) / قد (m)) ، شجره نامه خانوادگی ابتلا به دیابت، سن (سال) و آزمایشی که آیا بیمار علائم دیابت را نشان می‌دهد (در صورت منفی بودن با عدد صفر و در صورت مثبت بودن با عدد یک ثبت شد.) [۱۲]

البته قبل از هر کاری، شخص باید هدف مطالعه را دریابد و اطلاعات بیشتری در مورد چگونگی جمع آوری داده‌ها کسب کند. در هر صورت ادامه می‌دهیم و نگاهی به داده ها می‌اندازیم:

```
> library(faraway)
```

---

[۱] scatterplot

[۲] interactive and dynamic graphics

[۳] outliers

[۴] skewed

[5] The National Institute of Diabetes and Digestive and Kidney

[۶] قبیله ای سرخپوست

[۷] Phoenix

[۸] diastolic blood pressure

[۹] triceps skin fold thickness

[۱۰] 2-hour serum insulin

[11] Body mass index

[۱۲] داده‌ها از مرکز اطلاعات UCI از پایگاه داده یادگیری ماشین به آدرس اینترنتی www.ics.uci.edu/~mlearn/MLRepository.html، قابل دسترسی است.

```
> data(pima)
> pima
  pregnant  glucose  diastolic  triceps insulin  bmi diabetes age
1        6      148         72      35       0  33.6    0.627  50
2        1       85         66      29       0  26.6    0.351  31
3        8      183         64       0       0  23.3    0.672  32
…much deleted…
768      1       93         70      31       0  30.4    0.315  23
```

فرمان library (faraway)، داده‌های استفاده شده در این کتاب را قابل دسترسی می‌سازد. همانطور که در ضمیمه A توضیح داده شده است، این بسته نرم افزاری ابتدا باید نصب شود. ما در اینجا به صراحت این فرمان را نوشته‌ایم. لیکن در تمام فصل‌های بعدی فرض بر این است که در صورت تمایل به استفاده از داده‌های این متن، شما قبلا این فرمان را اجرا کرده‌اید. اگر پیغام "خطا در یافتن داده‌ها" را دریافت می‌کنید، ممکن است به دلیل فراموش کردنِ نوشتن این فرمان باشد.

فرمانِ pima این مجموعه داده خاص را فرا می‌خواند. به سادگی با نوشتن نام چارچوب داده‌ای-pima-داده ها منتشر می‌شوند، که آن برای نشان دادن در اینجا بسیار طولانی است. برای مجموعه داده‌ای با این اندازه، شخص می‌تواند صرفاً داده‌ها را جهت یافتن هر چیز خارج از موضع، بازرسیِ چشمی کند اما به طور قطع استفاده از روش‌های خلاصه سازی ساده‌تر است.

با چند خلاصه عددی آغاز می‌کنیم:

```
> summary(pima)
   pregnant         glucose       diastolic         triceps     
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00  
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00  
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54  
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00  
    insulin           bmi          diabetes          age       
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00  
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00  
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00  
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24  
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00  
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00  
      test       
 Min.   :0.000  
```

```
 1st Qu.:0.000
Median :0.000
Mean   :0.349
3rd Qu.:1.000
Max.   :1.000
```

فرمانِ summary( )، راهی سریع جهت دستیابی به اطلاعات خلاصه‌های معمول عددی تک متغیره است. در این مرحله ما در جستجوی هر چیز غیر عادی یا غیر منتظره هستیم که شاید بیانگر خطایی در ورود داده‌ها باشد. از این جهت نگاهی دقیق به مقادیر ماکزیمم و مینیمم هر متغیر ارزشمند است. با شروع از متغیر pregnant، ماکزیمم ۱۷ را می‌بینیم؛ این عدد بزرگی است اما غیر ممکن نیست. اما سپس می‌بینیم که مینیمم ۵ متغیر بعدی صفر است. تنها فشار خون یک مرده می‌تواند صفر باشد!!!!- باید اشتباهی رخ داده باشد. اینک به مقادیر مرتب‌شده می‌نگریم:

```
> sort(pima$diastolic)
  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 [19]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 [37]  0  0  0  0  0  24 30 30 38 40 44 44 44 44 46
…etc…
```

ما می‌بینیم که ۳۵ مقدار اول صفر است. شرحِ همراه داده‌ها چیزی بیان نمی‌کند، اما به نظر می‌رسد که صفر به عنوان کدِ مقادیر گم‌شده[1] استفاده شده است. به هر دلیلی، محققان فشار خون ۳۵ بیمار را نگرفته‌اند. در یک تحقیق واقعی، شخص ممکن است بتواند از محققان بپرسد که در واقع چه اتفاقی رخ داده است. با این اوصاف، می‌توان برخی از سوءتفاهم‌هایی که به سادگی رخ می‌دهند را برطرف نمود. یک آمارگرِ ناشی ممکن است این مقادیر گم شده را نادیده بگیرد و تجزیه و تحلیلی را با این فرض که مقادیر مشاهده شده واقعا صفر بوده‌اند، تکمیل کند. اگر خطا بعداً کشف می‌شد، در این صورت ممکن بود محققان به خاطر استفاده از صفر به عنوان کد داده‌های گم شده (انتخابی نامناسب، به این خاطر که صفر یک مقدار قابل قبول برای بعضی از متغیرها است) و عدم ذکر آن درشرح داده‌ها، مورد سرزنش قرار گیرند. متاسفانه چنین چشم‌پوشی‌هایی، به خصوص در مجموعه داده‌هایی با چنین اندازه و پیچیدگی، غیر معمول نیستند. آمارگر بخشی از مسئولیتِ یافتن اینگونه خطاها را بر عهده دارد.

ما تمامی مقادیر صفر این پنج متغیر را NA قرار می دهیم که کد مقادیر گم شده در نرم افزار R است:

[1] Missing values code

```
> pima$diastolic[pima$diastolic == 0] <- NA
> pima$glucose[pima$glucose == 0] <- NA
> pima$triceps[pima$triceps == 0] <- NA
> pima$insulin[pima$insulin == 0] <- NA
> pima$bmi[pima$bmi == 0] <- NA
```

متغیرِ test کمّی[۱] نیست، بلکه دسته‌ای[۲] است. این گونه متغیرها، عامل[۳] نیز نامیده می‌شوند. هرچند این متغیر به خاطر کد گذاری عددی، به صورت متغیری کمّی مورد استفاده قرار گرفته است. بهتر است که این‌گونه متغیرها را عامل بنامیم تا به طور مناسب مورد استفاده قرار بگیرند. گاهی اوقات این مسئله فراموش می‌شود و آماره‌هایی غیر منطقی مانند "متوسط کد پستی" محاسبه می‌گردد.

```
> pima$test <- factor(pima$test)
> summary(pima$test)
  0   1
500 268
```

حال می‌بینیم که ۵۰۰ مورد جواب آزمایششان منفی و ۲۶۸ مورد، مثبت بوده است. حتی بهتر است که از برچسب‌های توصیفی[۴] استفاده شود:

```
> levels(pima$test) <- c("negetive", "positive")
> summary(pima)
   pregnant         glucose        diastolic        triceps
 Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :29.00
 Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15
 3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
                  NA's   :5       NA's   :35       NA's   :227
   insulin           bmi          diabetes          age
 Min.   : 14.00   Min.   :18.20   Min.   :0.0780   Min.   :21.00
 1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00
 Median :125.00   Median :32.30   Median :0.3725   Median :29.00
 Mean   :155.55   Mean   :32.46   Mean   :0.4719   Mean   :33.24
```

---

[۱] quantitative

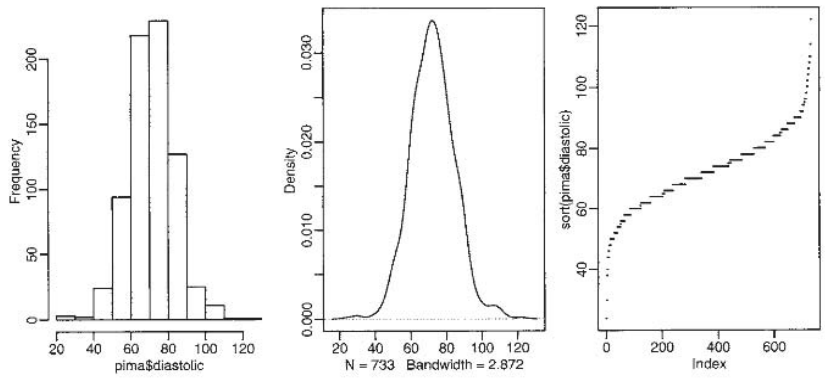
[۲] categorical

[۳] factor

[۴] descriptive labels

```
3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
NA's   :374      NA's   :11
    test
negetive:500
 positive:268
```

حال که تکلیف مقادیر گم شده را مشخص کرده‌ایم و داده‌ها را به طور مناسبی کدگذاری نموده‌ایم، برای رسم نمودار آماده‌ایم. شاید معروفترین نمودار تک متغیره، هیستوگرام باشد:



شکل ۱. ۱: اولین پنل، هیستوگرامی از فشار خون دیاستولیک را نشان می دهد، دومین پنل تخمین چگالی کرنل[۱] این متغیر رانمایش می دهد، در حالی که سومین شکل، شمایی از یک نمودار شاخص[۲] از مقادیر مرتب شده است.

```
> hist(pima$diastolic)
```

همانطور که در اولین پنلِ شکل ۱. ۱ دیده می‌شود، یک توزیع زنگوله‌ای[۳] از فشار خون دیاستولیک می‌بینیم که در حدود ۷۰ مرکزیت دارد. تهیه یک هیستوگرام به تعیین تعداد رده‌ها و جایگاه آن‌ها بر روی محور افقی (طول هر طبقه یا رده) نیازمند است. برخی گزینش‌ها می‌تواند منجر به هیستوگرام‌هایی شود که بعضی از خصوصیات داده‌ها را مبهم می‌سازد. نرم افزار R تعداد و طول طبقه‌ها را با توجه به توزیع و اندازه داده‌ها مشخص می‌کند. اما این گزینش عاری از خطا نیست و باز هم امکان ایجاد

---

۱ kernel density estimates

۲ Index plot

۳ bell-shaped distribution

هیستوگرام های گمراه کننده وجود دارد. به همین دلیل برخی ترجیح می‌دهند که از تخمین‌های چگالی کرنل، که نسخه روان‌تر هیستوگرام است استفاده کنند.[1]

```
> plot (density (pima$diastolic, na . rm=TRUE) )
```

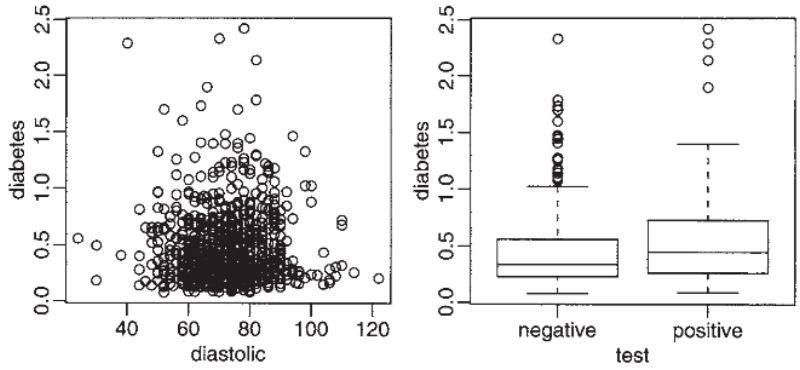
تخمین کرنل در پنل دوم شکل ۱. ۱ قابل مشاهده است. می بینیم که این نمودار از بلوک بندی‌های گیج کننده هیستوگرام اجتناب می‌ورزد. گزینه‌ای دیگر می‌تواند به سادگی، به نمودار کشیدن داده‌های مرتب شده بر حسب متغیر شاخص باشد:

```
> plot (sort (pima$diastolic), pch=".")
```

مزیت این کار در امکان بررسی جداگانه‌ی تمامی موارد، دیدن شکل توزیع و یافتن نقاط دور افتاده است. همچنین گسستگی در اندازه‌گیری‌های فشار خون قابل مشاهده است- مقادیر به نزدیکترین عدد زوج گرد می‌شوند و بنابراین "پله" در نمودار دیده می‌شود.

اینک به تعدادی از نمودارهای دو متغیره (شکل ۱. ۲) توجه کنید:

```
> plot(diabetes ~ diastolic, pima)
> plot(diabetes ~ test, pima)
```



شکل ۱. ۲: اولین پنل، نمودار پراکندگی فشار خون دیاستولیک برحسب عملکرد دیابت را نشان می‌دهد و پنل دوم، دو نمودار جعبه ای از فشار خون دیاستولیک است که بر اساس نتیجه آزمایش جدا شده‌اند.

در ابتدا، یک نمودار پراکندگی استاندارد از دو متغیر کمّی دیده می‌شود. در پنل دوم دو نمودار جعبه‌ای در کنار یکدیگر رسم شده است. نمودار جعبه‌ای برای یک متغیر کمی و متغیری کیفی مناسب

[1] برای مبحث مزیت‌های هیستوگرام و تخمین‌های کرنل به Simonoff (1996) مراجعه شود.

است. همچنین ایجاد یک ماتریسِ نمودار پراکندگی[1] نیز-که در اینجا نشان داده نشده- مفید خواهد بود و با دستور زیر ساخته می‌شود:

```
> pairs(pima)
```

نمودارهای پیشرفته بیشتری در ادامه بررسی خواهد شد، لیکن خلاصه‌های عددی و نموداری ارائه شده در این بخش برای بررسی اولیه داده‌ها کافی هستند.

۱. ۳ چه زمانی باید از تجزیه و تحلیل رگرسیونی استفاده شود:

آنالیز رگرسیون برای توضیح دادن یا مدل کردن رابطه بین متغیر Y- که متغیر پاسخ[2] ، خروجی[3] یا وابسته[4] نامیده می‌شود- و یک یا چند متغیر پیش‌بین[5]، ورودی[6]، مستقل[7] یا توضیح‌دهنده[8] - $X_1, \ldots \ldots, X_p$ -استفاده می‌گردد. در حالت $p = 1$، رگرسیون ساده است. لیکن هنگامی که $p >$ 1 باشد رگرسیونِ چندگانه[9] یا برخی اوقات رگرسیونِ چند متغیره[10] نامیده می‌شود. زمانی که بیش از یک Y وجود دارد، رگرسیون را چندگانه‌ی چندمتغیره[11] می‌نامند، مطلبی که به طور کامل در این کتاب بیان نشده است، با این وجود، برای بررسی این حالت می‌توان روی هر Y ، رگرسیون جداگانه انجام داد.

متغیر پاسخ، الزاماً متغیری پیوسته[12] است، اما متغیر توضیح دهنده می‌تواند پیوسته، گسسته[13] یا دسته‌ای باشد. بررسی متغیرهای توضیح‌دهنده دسته‌ای به قسمت‌های بعدی این کتاب موکول شده است. با توجه به مثال ارائه شده در بخش قبل، رگرسیونی با diastolic و bmi به عنوان X ها و diabetes به عنوان Y، رگرسیونی چندگانه شامل فقط متغیر های کمی است، که نگاهی گذرا به آن خواهیم انداخت. یک رگرسیون با diastolic و test به عنوان X ها و bmi به عنوان Y دارای دو متغیر پیش بین (یکی

---

[1] scatterplot matrix
[2] response
[3] output
[4] dependent
[5] predictor
[6] input
[7] independent
[8] explanatory
[9] multiple regression
[10] Multivariate regression
[11] multivariate multiple regression
[12] continuous
[13] discrete

کمّی و دیگری کیفی) است. این مسئله در فصل ۱۳ در مبحث تجزیه و تحلیل کواریانس[۱] بررسی خواهد شد. یک رگرسیون با test به عنوان X و diastolic به عنوان Y ، تنها شامل متغیرهای پیش‌بین کمّی است.-یک موقعیت ساده دو نمونه‌ای (تیمار)- مبحثی که آنالیز واریانس (ANOVA)[۲] نامیده می‌شود. رگرسیونی با test به عنوان Y بر روی diastolic و bmi به عنوان پیش‌بین‌ها، شامل متغیر پاسخ کیفی خواهد بود. در این مورد، یک رگرسیون لوجستیک[۳] قابل استفاده است، اما این مبحث در این کتاب بررسی نخواهد شد.

تجزیه و تحلیل رگرسیونی اهداف متعددی دارد:

پیش بینی مشاهدات آتی

بررسی تاثیر یا رابطه بین متغیرهای توضیح‌دهنده و متغیر پاسخ.

توصیفی کلی از ساختار داده ها

هم‌چنین، برای به کارگیری متغیرهای پاسخ چندمتغیره، متغیرهای پاسخ باینری[۴] (تجزیه و تحلیل رگرسیون لجستیک) و متغیرهای پاسخ شمارشی[۵] (رگرسیون پواسون) تعمیم‌هایی موجود است.

۱. ۴ تاریخچه

مسائل رگرسیونی برای اولین بار در قرن هجدهم به منظور کمک به ناوبری با استفاده از علم نجوم، مطرح شدند. لژاندر[۶] روش حداقل مربعات را در (سال) ۱۸۰۵ گسترش داد.

گاوس[۷] ادعا کرد که این روش را چندین سال قبل توسعه داده بود و در ۱۸۰۹ نشان داد که هنگامی که خطاها توزیع نرمال دارند، حداقل مربعات راه حل بهینه است. این روش تا اواخر قرن نوزدهم تقریباً به طور انحصاری در علوم فیزیکی مورد استفاده قرار می‌گرفت. فرانسیس گالتون[۸] در سال ۱۸۷۵، اصطلاح

---

[۱] analysis of covariance

[۲] analysis of variance

[۳] logistic regression

[۴] binary

[۵] count

[۶] Legendre

[۷] Gauss

[۸] Francis Galton

"رگرسیون(بازگشت) به میانگین[۱]" را در اشاره به معادله رگرسیون ساده به فرم زیر، برای اولین بار مطرح کرد:

$$\frac{y-\bar{y}}{SD_y} = r\frac{x-\bar{x}}{SD_x}$$

که در آن r همبستگی[۲] بین x و y را نشان می دهد. گالتون برای توضیح دادن این پدیده که "پسرانِ پدرانِ بلند قد، به احتمال زیاد قد بلند می‌شوند، اما نه به بلندی پدرانشان و حال آنکه پسران پدران کوتاه قد به احتمال زیاد قد کوتاه می‌شوند، اما نه به کوتاهیِ پدرانشان" از این معادله استفاده کرد. این پدیده، اثر رگرسیون (بازگشت)[۳] نامیده می‌شود[۴].

این اثر را می توان با نمرات به‌دست آمده از دوره‌ای که با استفاده از این کتاب تدریس شده است، تفهیم نمود. در شکل ۱. ۳، نموداری از نمرات میان ترم در برابر نمرات پایان ترم دیده می‌شود. حدود هر متغیر به گونه‌ای تغییر داده شده که میانگینِ صفر و انحراف معیاری برابر با یک داشته باشد، تا اینکه دشواری هر یک از امتحانات و حداکثر نمره ممکن (هر درس) ذهن را منحرف نکند. هم‌چنین، این کار معادله رگرسیون را به صورت زیر ساده می‌کند:

$$y = rx$$

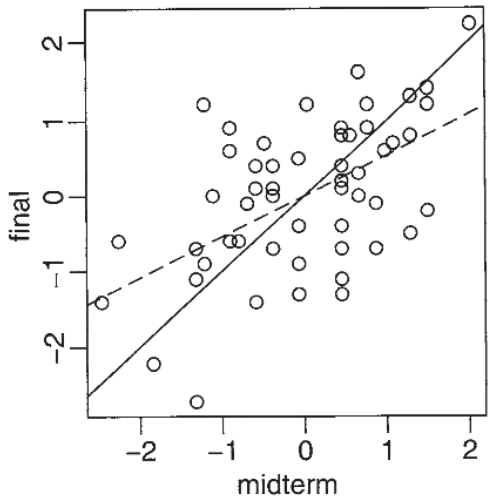
```
> data(stat500)
> stat500 <- data.frame (scale(stat500))
> plot(final ~ midterm, stat500)
> abline(0, l)
```

[۱] regression to mediocrity

[۲] correlation

[۳] regression effect

[۴] برای اطلاعات بیشتر به Stigler (1986) مراجعه گردد.



شکل ۱. ۳: نمرات میان ترم و پایان ترم در واحد استاندارد. برازش حداقل مربعات با خط چین و $y = x$ نیز با خط پر نمایش داده شده است.

خط $y = x$ به نمودار اضافه شده است. حال به عنوان مثال دانش آموزی که نمره‌اش در امتحان میان ترم یک انحراف معیار بیش از میانگین باشد، به طور منطقی می‌توان انتظار داشت که در امتحان پایان ترم نیز به همین خوبی عمل کند. برازش حداقل مربعات رگرسیون را محاسبه کرده و خط رگرسیون را به نمودار می‌کشیم. (جزئیات بیشتر در ادامه). همچنین مقادیر همبستگی نیز محاسبه شده است:

```
> g <- lm(final ~ midterm, stat500)
> abline(coef(g), lty = 5)
> cor(stat500)
    midterm      final    hw    total
midterm 1.00000 0.545228 0.272058 0.84446
final   0.54523 1.000000 0.087338 0.77886
hw      0.27206 0.087338 1.000000 0.56443
total   0.84446 0.778863 0.564429 1.00000
```

برازش رگرسیون، همان خط چین شکل ۱. ۳ است و همیشه از عدم دقت رنج می‌برد. می‌بینیم دانش آموزی که در میان ترم نمره‌اش یک انحراف معیار بالاتر از میانگین قرار دارد، پیش‌بینی می‌شود که فقط ۰/۵۴۵ انحراف معیار نمره در امتحان پایان ترم بالاتر از میانگین بگیرد.

از طرفی، دانش آموزی که در امتحان میان ترم نمره‌اش کمتر از میانگین است، انتظار می‌رود که در امتحان پایان ترم نسبتاً بهتر عمل کند، اگرچه همچنان زیر میانگین.

اگر امتحانات قادر به اندازه‌گیری توانایی دانش‌آموزان به بهترین وجه باشند، و با فرض اینکه این توانایی از میان ترم تا امتحان پایانی بی‌تغییر باقی بماند، آنگاه انتظار دستیابی به یک همبستگی دقیق وجود دارد. البته تصور چنین امتحان بی‌نقصی، انتظار زیادی است و مقداری انحراف به ناچار وجود خواهد داشت. علاوه بر این تلاش افراد نیز ثابت نیست. گرفتن یک نمره‌ی بالا در میان‌ترم تاحدودی می‌تواند به مهارت‌های شخص نسبت داده شود، اما مقداری هم به شانس بستگی دارد. فرد نمی‌تواند انتظار داشته باشد که این شانس تا انتهای ترم باقی بماند. بنابراین "رگرسیون(بازگشت) به میانگین" را در این مثال می‌توان دید.

البته این در تمام موقعیت های (x, y) مانند مثال قبل صادق است – یک مورد، اصطلاحِ "بدبیاریِ سال دوم در ورزش" است.- به این معنی که یک ستاره جدید، بعد از سال اول بسیار خوب، فصلِ دوم متوسطی دارد. هرچند در مثال پدر-فرزندی، نزدیکیِ نسل‌های پی‌درپی به میانگین پیش‌بینی می‌شود؛ لیکن این موضوع برای تمام جامعه صادق نیست. چرا که نوسانات تصادفی، موجب بقای تغییرات می‌شوند. در بسیاری از کاربردهای رگرسیون، اثر بازگشتی مورد نظر نیست لذا تاسف بار است که ما مانده‌ایم و این عنوان غلط انداز.

متودولوژی رگرسیون با ظهور محاسباتِ پیشرفته، به سرعت توسعه یافت. تا قبل از این، تنها برآورد یک مدل رگرسیون نیاز به محاسبات دستی گسترده‌ای داشت. با بهبود یافتن سخت افزارهای محاسباتی، حیطه تجزیه و تحلیل نیز گسترده‌تر شده است.

تمرین ها:

۱. مجموعه داده teengamb مربوط به مطالعه‌ای در مورد نوجوانانی است که در بریتانیا قمار می‌کنند. از داده‌ها خلاصه‌های عددی و نموداری بگیرید، در مورد هر ویژگی (خصوصیتی) که جالب توجه است، نظر بدهید. خروجی‌های ارائه شده را به مقداری محدود کنید که برای یک خواننده پر مشغله، جهت درک پایه‌ایِ داده‌ها مفید باشد.

۲. مجموعه داده usewages از سرشماری جمعیت[۱] مربوط به سال ۱۹۸۸ استخراج شده‌است. مانند سوال قبل از داده‌ها خلاصه‌های عددی و نموداری بگیرید.

۳. مجموعه داده prostate از مطالعه‌ای بر روی ۹۷ مرد با سرطان پروستات، که در آستانه برداشت ریشه‌ای غده پروستات بودند، استخراج شده است. مانند سوال یک، از داده‌ها خلاصه‌های عددی و نموداری بگیرید.

[۱] Current Population Survey

۴. مجموعه داده sat مطالعه‌ای با عنوان "هر قدر بپردازی، همانقدر دریافت می‌کنی: مناظره در مورد عدالت در هزینه‌های مدارس عمومی" استخراج شده است. مانند سوال یک، از داده‌ها خلاصه‌های عددی و نموداری بگیرید.

۵. مجموعه داده divusa، شامل اطلاعاتی در مورد طلاق در آمریکا، از سال ۱۹۲۰ تا ۱۹۹۶ است. مانند سوال یک از داده‌ها خلاصه‌های عددی و نموداری بگیرید.