6장 데이터 종류에 따른 분석 기법

6.1 분석기법

- **80:20 파레토법칙**
- 80%의 데이터분석 실제 문제는 20% 정도의 통계기법으로 처리할 수 있다.

6.2 모든 데이터에 행해야 할 분석

1. 데이터의 내용, 구조 타입을 파악함
2. 데이터의 요약 통계량을 파악함
3. 결측값의 존재여부를 파악함
4. 시각화를 시도함

6.3 수량형 변수의 분석

1. 데이터 분포의 시각화: 히스토그램, 상자그림, 줄기그림, 바이올린그림, 확률밀도추정량
2. 요약 통계량의 계산
3. 데이터의 정규성 검토
4. 가설검정과 신뢰구간: t-검정
5. 이상점[특이점] 찾기

6.4 성공-실패값 범주형 변수의 분석

1. 요약 통계량 계산: 도수분포표
2. 데이터 분포의 시각화: 막대그래프
3. 가설검정과 신뢰구간: 이항검정

6.5 수량형 x, 수량형 y의 분석

1. 산점도를 통하여 관계의 모양을 파악함
2. 상관계수를 구함
3. 선형모형을 적합함
4. 잔차분석 (모형적합성 검토)
5. 이상값이 존재하는 경우 로버스트 회귀분석을 시행함
6. 비선형 데이터인 경우 LOESS 등의 평활법을 사용함

High breakdown regression

M-estimators are not very resistant to leverage points, unless they have redescending ψ-functions. In this case, they need a good starting point. Several robust estimators of regression have been proposed in order to remedy this shortcoming, and are high breakdown point estimators of regression. We remember that the breakdown point of the Huber-type and least residuals estimators is 0% (they cannot cope with leverage points), and in general it cannot exceed $1/p$ for other robust M-estimators (that is, it decreases with increasing dimension where there are more opportunities for outliers to occur).

The first estimator to become popular was the *least median of squares* (LMS) estimate, defined as the p-vector

$$\hat{\beta}_{lms} = \min med(y_i - x_i^{\mathrm{T}}\beta)^2 \ .$$

This fit is very resistant and needs no scale estimate. The breakdown point of the LMS estimator is 50%, but this estimator is highly inefficient we the central model is the Gaussian one.

Another proposal is the *least trimmed squares* (LTS) estimate, defined as the p-vector

$$\hat{\beta}_{lts} = \min \sum_{i=1}^{k} r_{[i]}^2 \ ,$$

where $r_{[1]}^2 \leq r_{[2]}^2 \leq \ldots \leq r_{[n]}^2$ are the ordered squared residuals $r_i^2 = (y_i - x_i^{\mathrm{T}}\beta)^2$, $i = 1, \ldots, n$, and k is the largest integer such that $k \leq n/2+1$. This estimator is more efficient than LMS, having breakdown point 50%. Its main disadvantage is that it requires a heavy computational effort.

A further proposal is the S-estimator, in which the coefficients are chosen to find the solution of

$$\sum_{i=1}^{n} \chi\left(\frac{r_i}{c_o s}\right) = (n-p)k_2$$

with smallest scale s. Usually, $\chi(\cdot)$ is usually chosen to the integral of Tukey's bisquare function, $c_0 = 1.548$ and $k_2 = 0.5$ is chosen for consistency at the Gaussian distribution. S-estimates have breakdown point 50%, such as for the bisquare family. However, the asymptotic efficiency of an S-estimate under normal errors is 0.29, which is not very satisfactory but is better than LMS and LTS.

It is possible to combine the resistance of these high breakdown estimators with the efficiency of M-estimation. The MM-estimates (see Yohai, 1987, and Marazzi, 1993) have an asymptotic efficiency as close to one as desired, and simultaneously breakdown point 50%. Formally, the MM-estimate $\hat{\beta}_{MM}$ it consists in: (1) compute an S-estimate with breakdown point 1/2, denoted by $\hat{\beta}^*$; (2) compute the residuals $r_i^* = y_i - x_i^{\mathrm{T}}\beta^*$ and find $\hat{\sigma}^*$ solution of $\sum_{i=1}^n \chi(r_i^*/\sigma) = k_2(n-p)$; (3) find the minimum $\hat{\beta}_{MM}$ of $\sum_{i=1}^n \rho((y_i - \beta^{\mathrm{T}} x_i)/\hat{\sigma}^*)$, where $\rho(\cdot)$ is a suitable function. The function `rlm` has an option that allows to implement MM-estimation:

```
rlm(formula, data, ...,method="MM")
```

To use other breakdown point estimates, in `R` there exists the function `lqs` to fit a regression model using resistant procedures, that is achieving a regression estimator with a high breakdown point (see Rousseeuw and Leroy, 1987, Marazzi, 1993, and Venables and Ripley, 2002, Sec. 6.5). The usage of the function `lqs` is:

```
lqs(formula, data, method = c("lts", "lqs", "lms", "S"),...)
```

where

- `formula`: is a formula of the form `lm`;
- `data`: (optional) is the data frame used in the analysis;
- `method`: the method to be used for resistant regression. The first three methods minimize some function of the sorted squared residuals. For methods `lqs` and `lms` is the quantile squared residual, and for `lts` it is the sum of the quantile smallest squared residuals.

Several other arguments are available for this function, such as the tuning constant `k0` used for $\chi(\cdot)$ and $\psi(\cdot)$ functions when `method = "S"`, currently corresponding to Tukey's biweight.

For summarizing the output of function `lqs` the function `summary` cannot be used. Function `lqs` gives a list with usual components, such as `coefficients`, `scale` `residuals`, `fitted.values`. In particular, `scale` gives the estimate(s) of the scale of the error. The first is based on the fit criterion. The second (not present for `method == "S"`) is based on the variance of those residuals whose absolute value is less than 2.5 times the initial estimate. Finally, we remember that high breakdown procedures do not usually provide standard errors. However, these can be obtained by a data-based simulation, such as a bootstrap.

6.6 범주형 x, 수량형 y

1. 데이터시각화: 병렬 상자그림
2. ANOVA 선형모형 적합
3. 잔차분석 (모형적합성 검토)

6.7 수량형, 범주형 (성공-실패)

1. 산점도
2. 일반화선형모형 적합
3. 잔차분석 (모형적합성 검토)