Week 3. Types of Data—Visualizing and Transforming

Fox., John. 2016. *Applied Regression Analysis and Generalized Linear Models*, 3rd Ed.
Ch. 3-4. Examining and Transforming Data

Sanghoon Park

Department of Political Science
University of South Carolina

2020-09-20

Introduction

변수의 측정 수준, 시각화, 그리고 기초통계

- 앞으로 다룰 여러 모델들은 데이터의 특성에 대해서 서로 다른 가정을 가지고 있음.
- 이번 주는 다른 유형의 데이터와 변수들의 관계를 통계적 개념으로 시각화하는 것에 목적이 있음.

사회과학자들은 연구과정에 데이터 그 자체를 들여다보는 것에 시간을 할애하지 않는 경향이 있음. 왜 데이터를 보아야 할까?

- 첫째, 데이터는 우리에게 현실 세계가 어떻게 돌아가고 있는지를 알려주는 단서임.
- 둘째, 연구자로 하여금 현실에 불가피하게 존재하는 오차(errors)를 확인할 수 있게 함.
- 셋째, 데이터를 통해 이론화되지 않은 무언가를 파악할 수 있음.

Introduction

변수의 측정 수준, 시각화, 그리고 기초통계

- 앞으로 다룰 여러 모델들은 데이터의 특성에 대해서 서로 다른 가정을 가지고 있음.
- 이번 주는 다른 유형의 데이터와 변수들의 관계를 통계적 개념으로 시각화하는 것에 목적이 있음.

사회과학자들은 연구과정에 데이터 그 자체를 들여다보는 것에 시간을 할애하지 않는 경향이 있음. 왜 데이터를 보아야 할까?

- 첫째, 데이터는 우리에게 현실 세계가 어떻게 돌아가고 있는지를 알려주는 단서임.
- 둘째, 연구자로 하여금 현실에 불가피하게 존재하는 오차(errors)를 확인할 수 있게 함.
- 셋째, 데이터를 통해 이론화되지 않은 무언가를 파악할 수 있음.

Introduction

변수의 측정 수준, 시각화, 그리고 기초통계

- 앞으로 다룰 여러 모델들은 데이터의 특성에 대해서 서로 다른 가정을 가지고 있음.
- 이번 주는 다른 유형의 데이터와 변수들의 관계를 통계적 개념으로 시각화하는 것에 목적이 있음.

사회과학자들은 연구과정에 데이터 그 자체를 들여다보는 것에 시간을 할애하지 않는 경향이 있음. 왜 데이터를 보아야 할까?

- 첫째, 데이터는 우리에게 현실 세계가 어떻게 돌아가고 있는지를 알려주는 단서임.
- 둘째, 연구자로 하여금 현실에 불가피하게 존재하는 오차(errors)를 확인할 수 있게 함.
- 셋째, 데이터를 통해 이론화되지 않은 무언가를 파악할 수 있음.

Introduction

주어진 시간에 대해 하나 이상의 변수들을 분석하는 것은 연구에서 매우 중요

- 데이터 탐색(data exploration), 모델 특정(model specification), 진단(diagnostics)
- Fox (2016) 의 Figure 3은 왜 우리가 데이터를 들여다 보아야 하는지에 관한 직관적인 예시를 제공함.

Table 1: Example data: Anscombe (1973)

x	y1	y2	y3	x4	y4
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04

Introduction

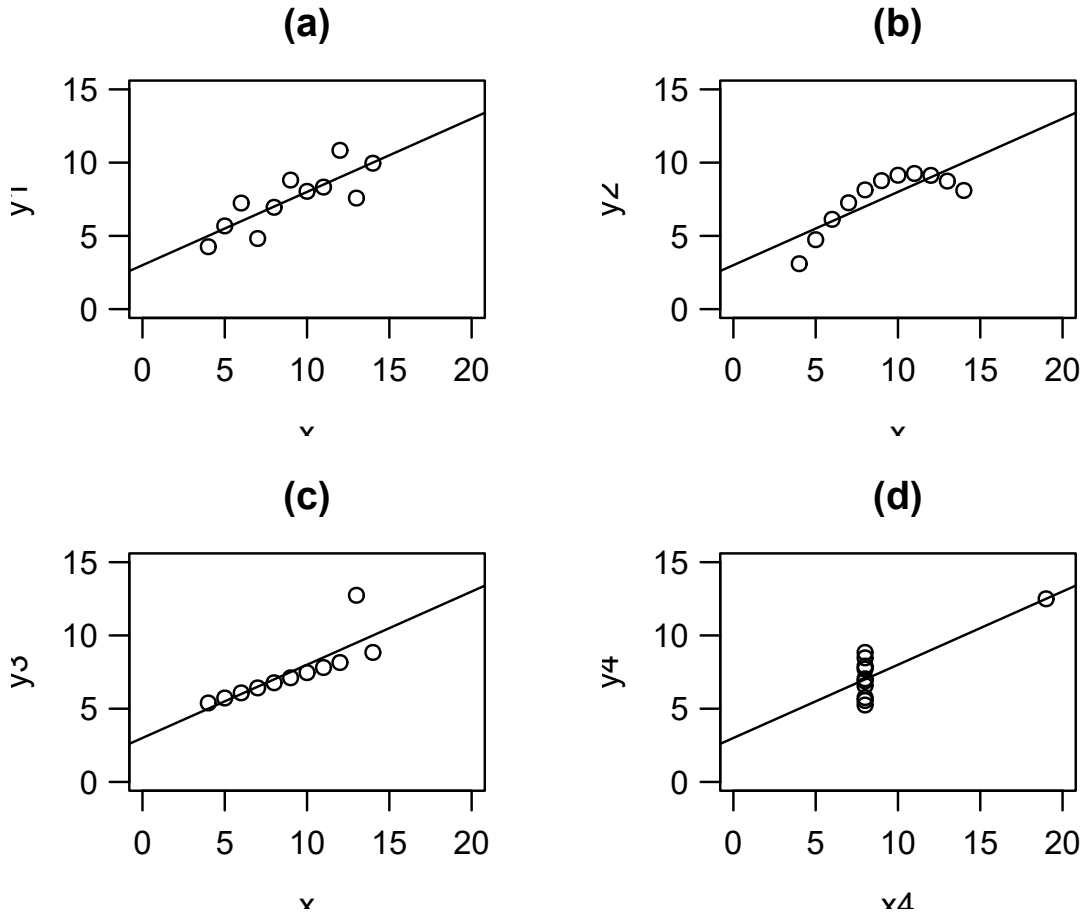
Anscombe (1973)에서 가져온 이 예제는 최소자승법으로 그린 선형회귀선이 동일한 네 쌍의 두 변수를 양변량 플롯으로 보여줌.

Table 2: Descriptive Statistics of Anscombe (1973)

	Obs.	Mean	Std.	Min	Max
x	11	9.000000	3.316625	4.00	14.00
y1	11	7.500909	2.031568	4.26	10.84
y2	11	7.500909	2.031657	3.10	9.26
y3	11	7.500000	2.030424	5.39	12.74
x4	11	9.000000	3.316625	8.00	19.00
y4	11	7.500909	2.030578	5.25	12.50

Introduction

데이터를 제대로 들여다보지 않으면, 아마도 회귀분석을 하고 회귀선을 그린 뒤 이 네 쌍의 변수들이 모두 유사한 관계를 가지고 있다고 결론내릴 수 있을 것.



Types of Data

측정 수준(levels of measurement)

- 명목형(Nominal)
  - 변수에 속한 값들이 서로 배타적이며(mutually exclusive), 구별이 가능하며, 순서를 가지지 않음.
  - 이항(binary) 변수 또는 이산(dichotomous) 변수
- 순서형(Ordinal)
  - 변수에 속한 값들이 순위/순서를 가짐.
  - 다만 상대적 순위만을 알려줄 뿐, 순위 간의 차이가 정확하게 어느 정도인지는 알려주지 않음.

Types of Data

- 등간형(Interval)
  - 변수의 값들 간의 차이가 서로 일정한 간격을 가짐.
- 비율형(Ratio)
  - 절대영(absolute zero)이 존재
  - 예) 경제성장율 0%일 경우, 경제가 '전혀' 성장하지 않았다는 것을 의미. 반면 등간형 변수인 온도가 0도라고 해서 온도가 존재하지 않는 것은 아님.

명목형-, 순위형-, 그리고 등간형 데이터는 보통 이산형일 가능성이 크고, 비율형 데이터는 연속형 자료일 가능성이 큼.

- 다만 학자들은 대개 등간형과 비율형 데이터를 거의 같은 것처럼 취급하고는 함.

Measures of Central Tendency

평균(Mean)

$$\bar{x} = \frac{\sum_{i=1}^{n}}{n}$$

- 일종의 균형점이라고 생각하면 이해가 편함.
- 값들 간의 거리 차이의 제곱을 최소화한 결과라고 생각할 수 있음.
- 위와 같은 평균을 산술평균(arithmetic mean)이라 하며 이외에도 기하평균(geometric mean), 조화평균(harmonic mean) 등이 있음.
  - 기하평균: $\bar{x} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$
  - 조화평균: $\bar{x} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}}$

Measures of Central Tendency

평균(Mean)

그렇다면 평균을 왜 보는 것일까?

- 평균은 중심(central)을 보여줄 수 있는 하나의 지표에 지나지 않음.
- 예를 들어, A, B, C 학급의 영어 실력을 비교하고 싶다고 하자.
  - 아무런 추가 정보가 없을 때, 우리는 각 학급의 영어 실력을 무엇으로 파악할까?
  - 평균이란 어떠한 집단의 정보를 요약하여 그것을 대표하는 값이라는 의미를 가짐.
  - 우리는 잘 모를 때, 그나마 틀릴 가능성이 제일 낮은 값—평균을 제시하곤 함.
- 만약 90점이 8명, 20점이 2명인 학급이 있다고 하자. 평균은 얼마일까?
  - 76점—이 평균이 과연 이 학급의 영어 실력을 잘 보여준다고 할 수 있을까?

Measures of Central Tendency

중앙값(Median) & 최빈값(Mode)

만약 평균이 우리가 기대하는 대표값으로서 제대로 기능하지 못한다면 어떻게 될까?

- 평균에 대한 대안들: 중앙값과 최빈값
- 중앙값: 말 그대로 중앙에 놓인 값. 5개의 값이 있다면 3번째에 해당하는 값이 중앙값이라고 할 수 있음.
  - 구체적으로는 편차(deviations)의 절대값의 합이 최소가 되게 하는 x의 값
  - 평균과는 다르게 이탈치(outliers)에 의해 크게 영향받지 않음.
- 최빈값: 데이터에서 가장 자주 나타나는 값
  - 측정 수주에 상관없이 사용될 수 있다는 점에서 가장 범용성이 높음.
  - 그러나 실질적으로 분석적 함의를 크게 가지고 있지 못함.
  - 대개 이항 변수일 경우에만 사용

Variance

범위(Range)와 분위(Centiles)

- 범위: 표본에서 최소값과 최대값 사이의 공간을 의미
- 분위: 특정한 값이 분포의 어디에 속하는지를 정량화하여 나타낸 결과
- 다른 측정지표들과는 달리, 범위와 분위는 모든 정보량에서 사용되지는 않음.
    - 예) 명목형 변수인 종교가 있다고 하자. 천주교, 기독교, 불교 등으로 코딩된 이 변수의 범위와 분위를 구할 수 있을까?

Variance

편차(Deviations)

- 편차($x_i - \bar{x}$)란 개별 값이 평균으로부터 떨어져 있는 단순 거리를 의미
- 분포에서 모든 편차의 총합은 0
- 모든 값의 편차를 제곱하여 그 평균을 구하면 표본의 분산(variance, σ_x)
- 표준편차는 분산의 제곱근 값
    - 분산이 편차를 제곱하여 더한 것의 평균이었다면, 그러한 분산에 제곱근을 씌워줌으로써 원래의 측정 단위로 원상복귀시키는 것과 같음.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- n이 아니라 $n-1$로 나눠주는 이유는 Bessel의 제안에 따른 것[1]

[1] $n-1$로 나눠줄 경우, 모집단의 분산을 추정할 때 나타날 수 있는 편향(bias)을 완화할 수 있음.

Variance

오차(Errors)

- 표본의 크기에 의해 가중치가 주어진(weighted) 표준 편차

$$\sigma_x = \frac{s}{\sqrt{n}}$$

- 표본평균의 정확성을 보여주는 신뢰구간(confidence intervals)을 계산할 때 사용

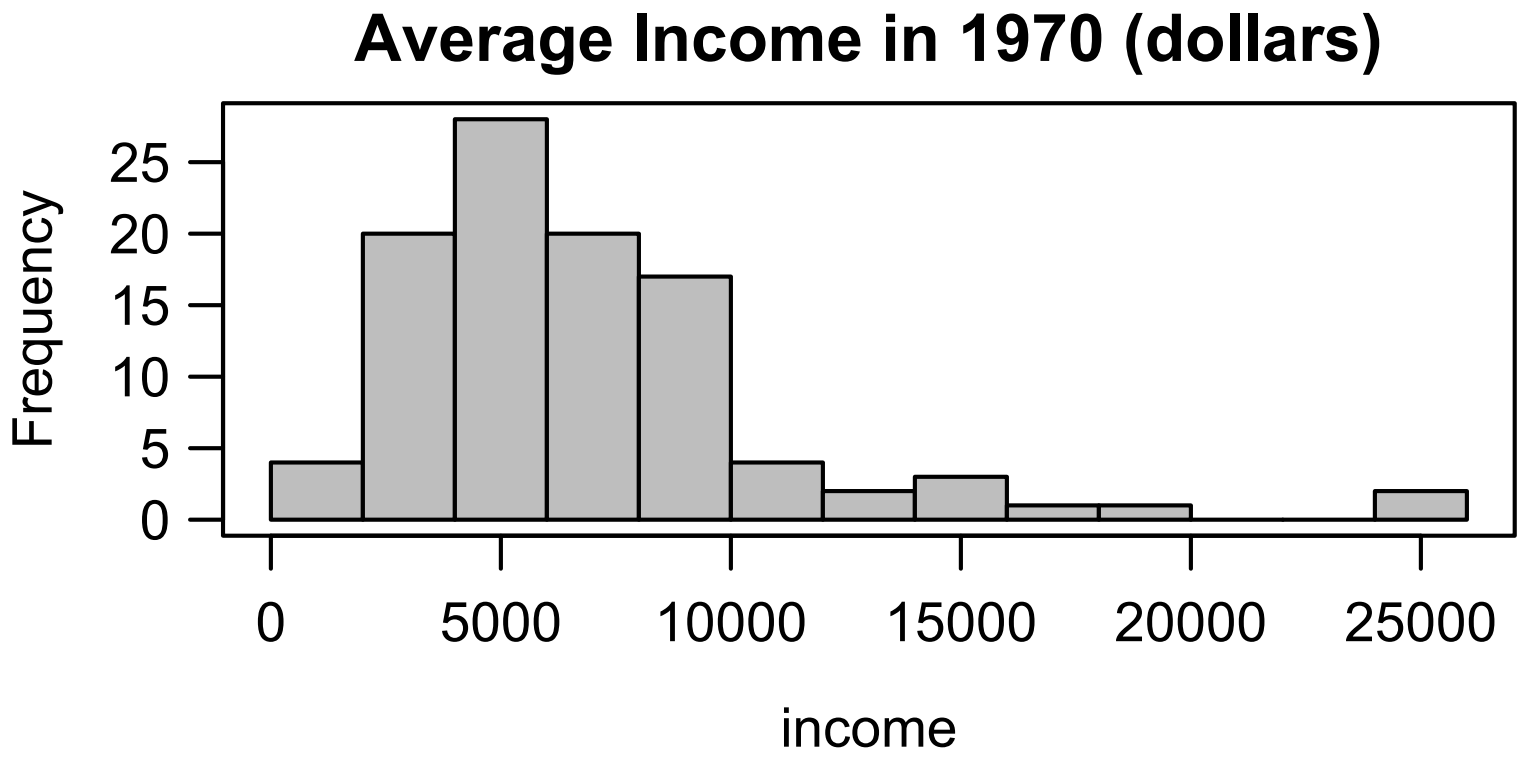
Moment

분포에 대한 일련의 속성들을 보다 일반적으로 보여줌.

- 어떤 분포의 K 번째(K_th) 모멘트 = M_K = $E[(x-\mu)^k]$
- $M_1 = E(x) = \bar{x}$ (평균 또는 중심경향성을 보여주는 다른 지표)
- $M_2 = E[(x-\mu)]^2 = \sigma^2$ (분산)
- $M_3 = E[(x-\mu)]^3$ = 왜도 (분포가 어떻게 기울어있는지)
  - 만약 $M_3 \geq 0$, 오른쪽으로 긴 꼬리를 가진 분포(right-skewed)
  - 만약 $M_3 \leq 0$, 왼쪽으로 긴 꼬리를 가진 분포(left-skewed)
- $M_4 = E[(x-\mu)^4]$ = 첨도 (kurtosis, 분포가 얼마나 뾰족한지)
  - lepto-: 매우 분포가 뾰족한 (한 쪽으로 집중되어 있는)
  - meso-: 분포가 비교적 정규형태로 잘 분포되어 있는
  - platy-: 분포가 상대적으로 평평한

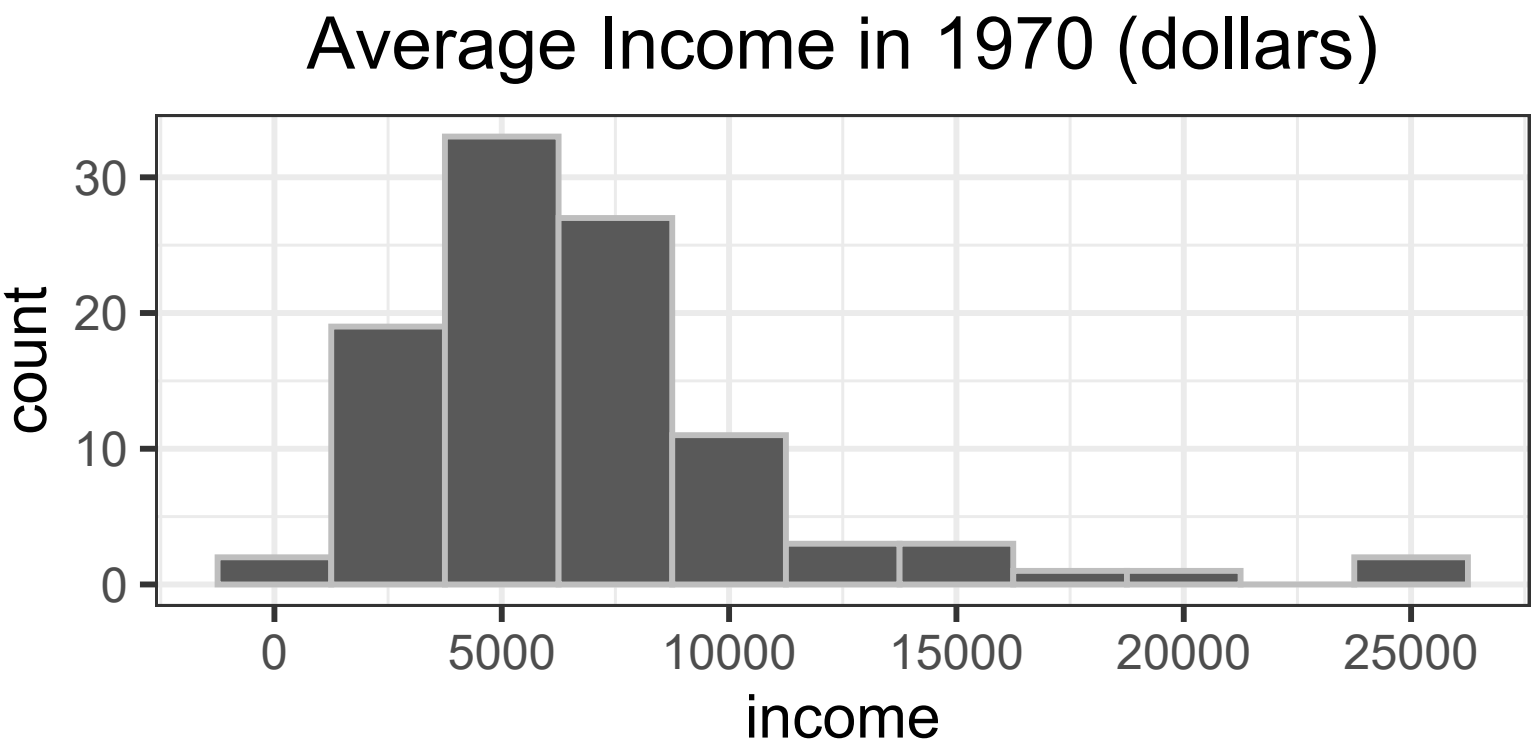
Histograms

```
library(car)
with(Prestige, hist(income,
          breaks="FD", col="gray",
          main="Average Income in 1970 (dollars)")); box()
```



Histograms

```
library(car);library(ggplot2);library(tidyverse)
Prestige %>% ggplot(aes(x = income)) +
  geom_histogram(color = "gray", binwidth = 2500) +
  scale_x_continuous(breaks = seq(0, 25000, 5000)) +
  labs(title = "Average Income in 1970 (dollars)") +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```



Histograms

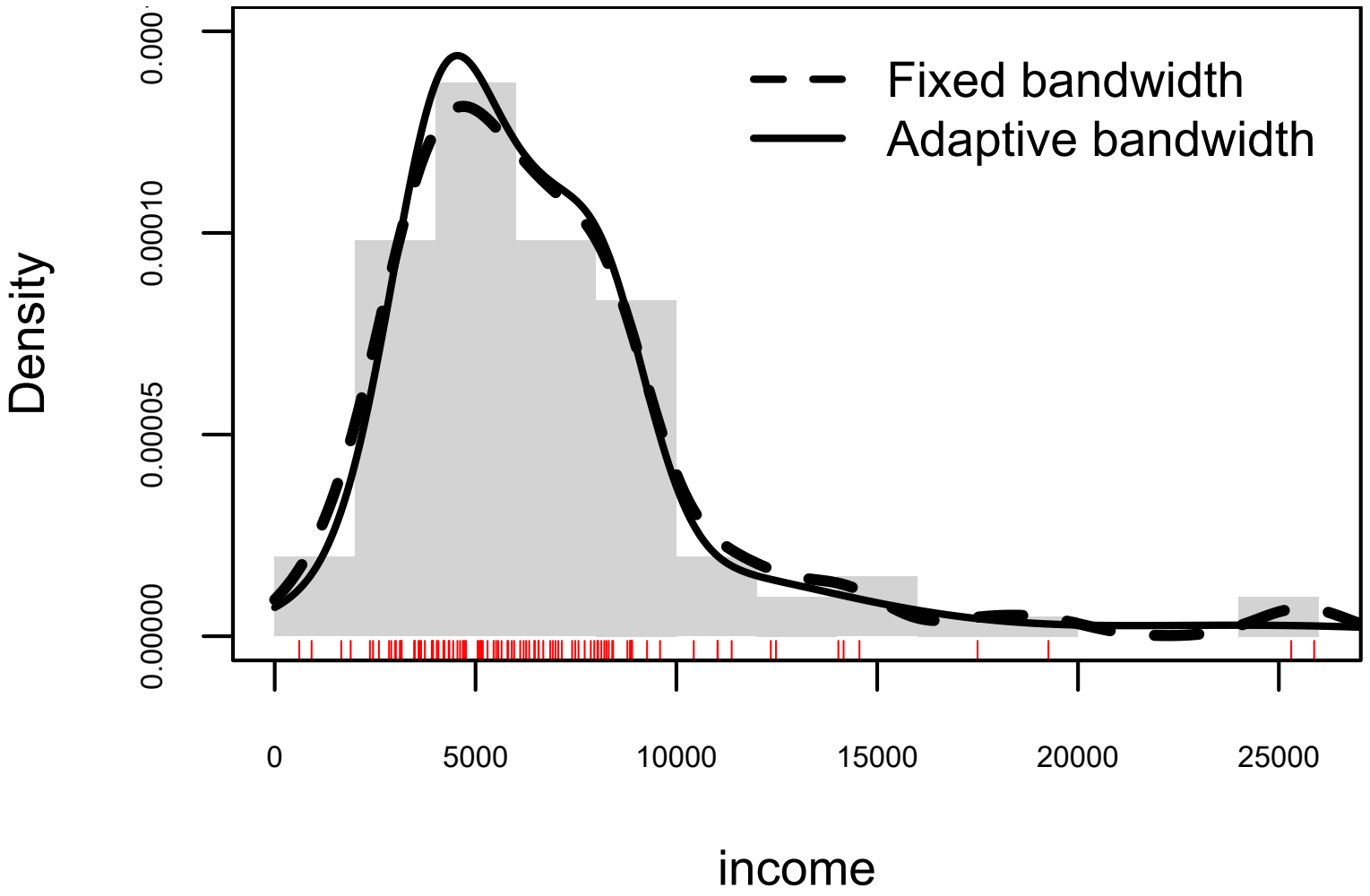
히스토그램이란?

- 히스토그램은 X 축에 따라 데이터의 값 플롯으로 나타냄.
- 데이터의 밀도(빈도)는 Y 축에 나타남.
- 모든 측정 수준의 변수는 히스토그램으로 표현할 수 있음.
- 다만 히스토그램을 그릴 때, "균형"을 잘 파악하는 것이 필요
  - 너무 많은 구간으로 데이터를 나누어서 막대를 그리게 되면 히스토그램을 통해 개략적인 추세를 파악하는 것이 어려워짐; 정보량이 과다
  - 너무 적은 구간으로 데이터를 나눌 경우에는 히스토그램이 데이터에서 파악할 필요가 있는 정보를 감추어버림.
  - 주로 관측치 수에 제곱근을 취한 값($\sqrt{N}$)을 구간의 기준(bin)으로 설정하고는 하지만 정해진 답은 없음.

Kernal Density Plot

```
with(Prestige, {
  hist(income, freq=FALSE, ylim=c(0, 1.5e-4),
       breaks="FD", main="", lty=0)
  lines(density(income, from=0), lwd=3, lty=2)
  lines(adaptiveKernel(income, from=0), lwd=2, lty=1)
  rug(income, col="red")
  legend("topright", c("Fixed bandwidth",
                       "Adaptive bandwidth"),
         lty=2:1, lwd=2, inset=0.02, bty="n");box()})
```

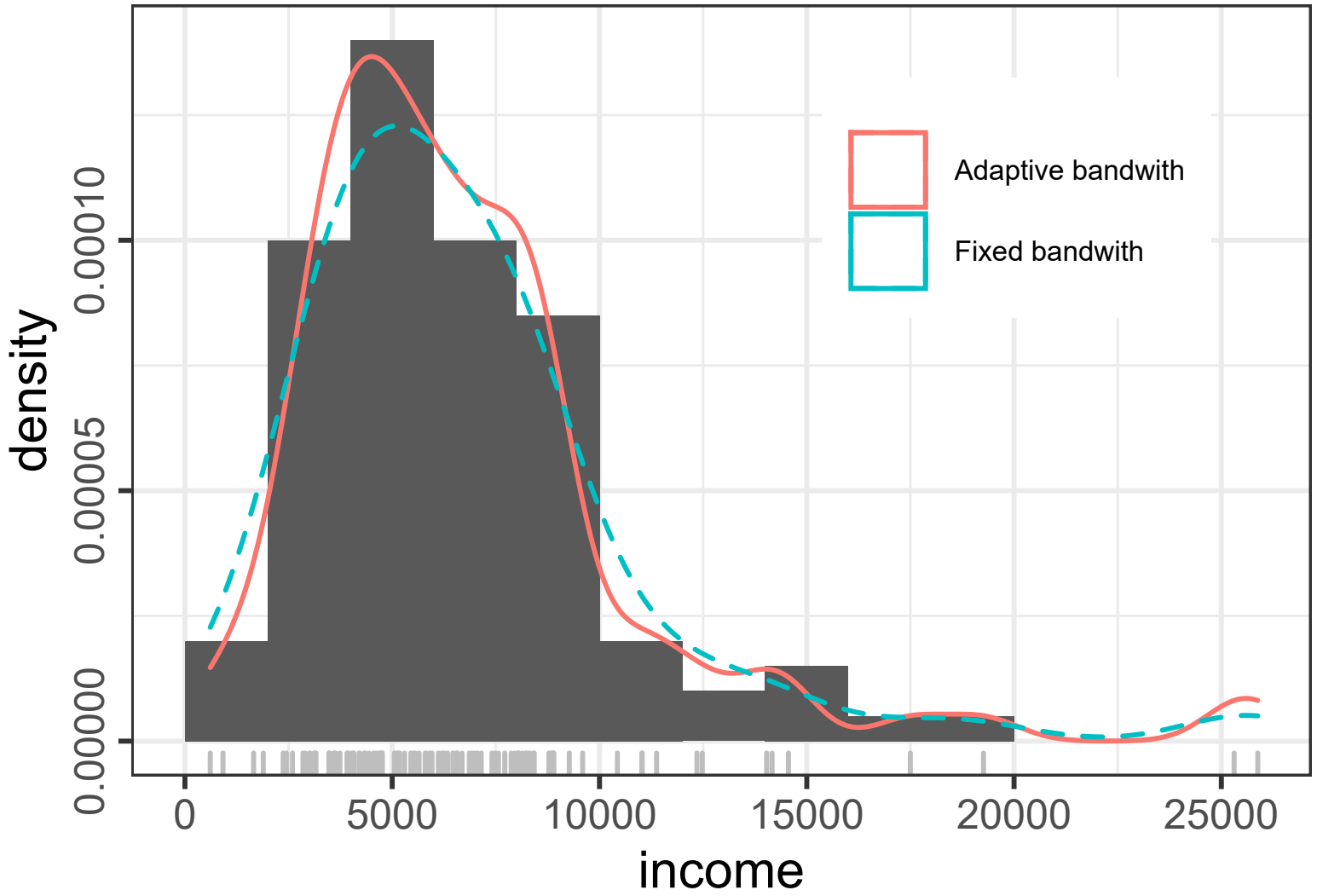
Kernal Density Plot



Kernal Density Plot

```
Prestige %>% ggplot(aes(x = income)) +
  geom_histogram(aes(y = ..density..),
                 breaks=seq(0,max(Prestige$income),2000)) +
  geom_density(aes(color = "Adaptive bandwith"),
                   kernel = "gaussian",
                linetype = 1, adjust = 0.8) +
  geom_density(aes(color = "Fixed bandwith"),
                linetype = 2,
                position = "stack") +
  scale_x_continuous(breaks = c(seq(0, 30000, 5000))) +
  scale_y_continuous(breaks = c(seq(0, 15e-5, 5e-5))) +
  theme_bw() + theme(legend.title = element_blank(),
                     legend.position = c(0.8, 0.9))
```

Kernal Density Plot



Kernal Density Plot

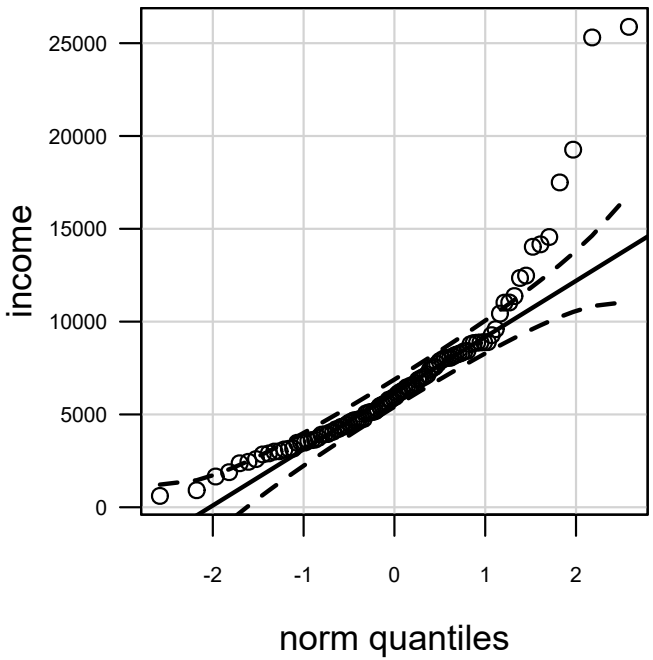
커널밀도플롯이란?

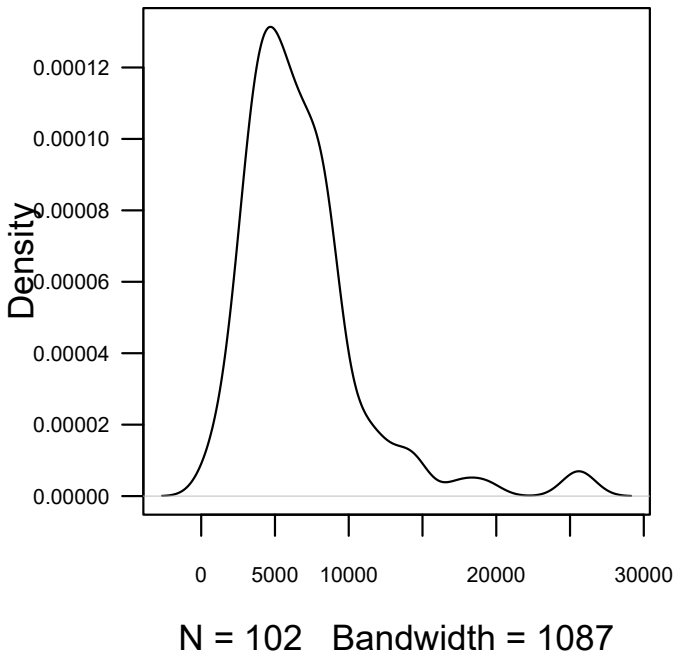
- 히스토그램의 구간(bin)을 부드럽게 만들어 연속적으로 이은 플롯

  비모수적(Nonparametric) 밀도 추정치는 평균화와 평탄화(smoothing)를 통해 전통적인 히스토그램의 결함을 다룬다. . . 공식적으로 밀도 추정치는 표본에 기초하여 어떤 변수의 확률밀도함수를 추정함으로써 얻을 수 있지만, 비공식적으로는 일종의 부드럽게 만든 히스토그램을 만들기 위한 기술적(descriptive) 기법이라고도 볼 수 있다(Fox, 2016, 33).

- 히스토그램 위에 덧씌워 그릴 수 있음.
  - 다만 데이터의 각 관측치가 어디에 위치하는지를 보여주지 않기 때문에 별도의 플롯이 필요함 (rug, 러그).

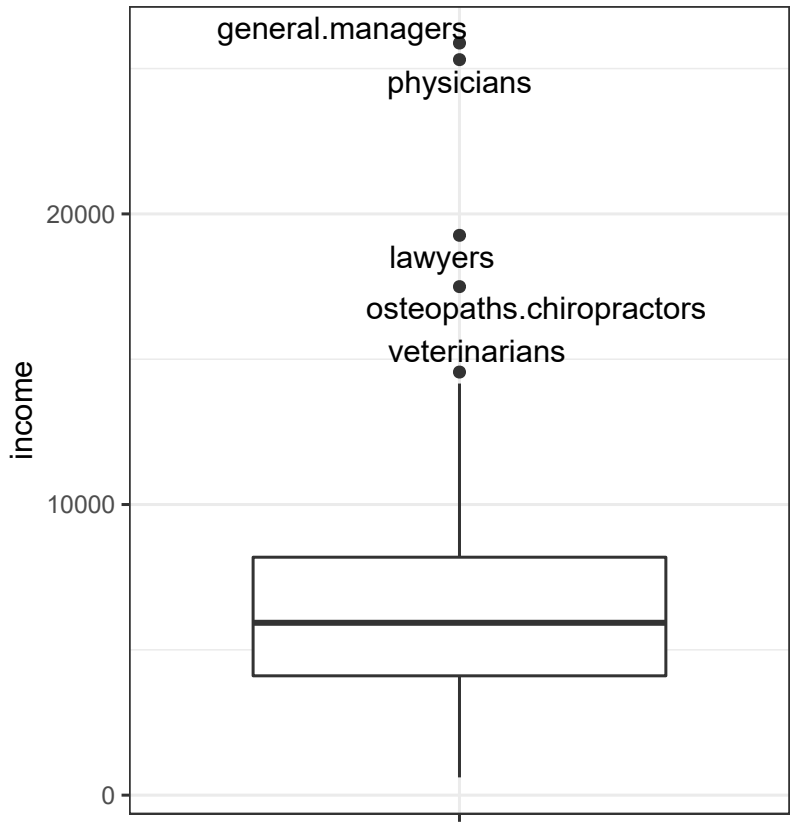
Q-Q Plot





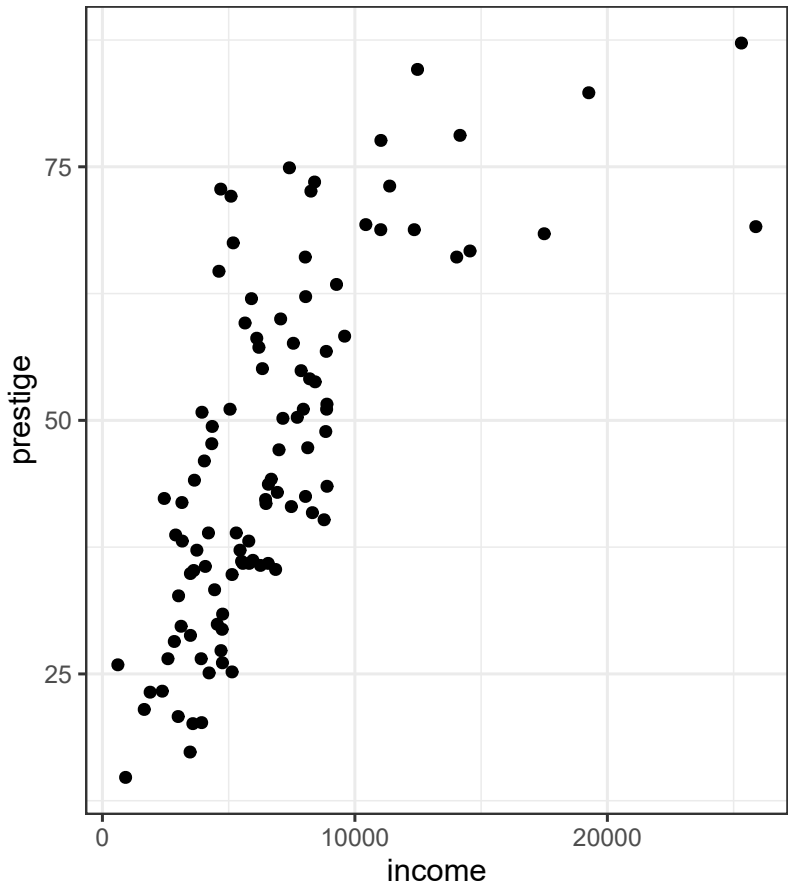
- 사분면(quadrant)에 따라 나눈 관측치들을 분포에 대한 이론적 기대와 비교하여 그린 플롯
- 가장 많이 사용되는 이론적 기대는 정규분포이지만, 분포 자체는 연구자 마음대로 설정할 수 있음(uniform, chi-square)

Boxplots



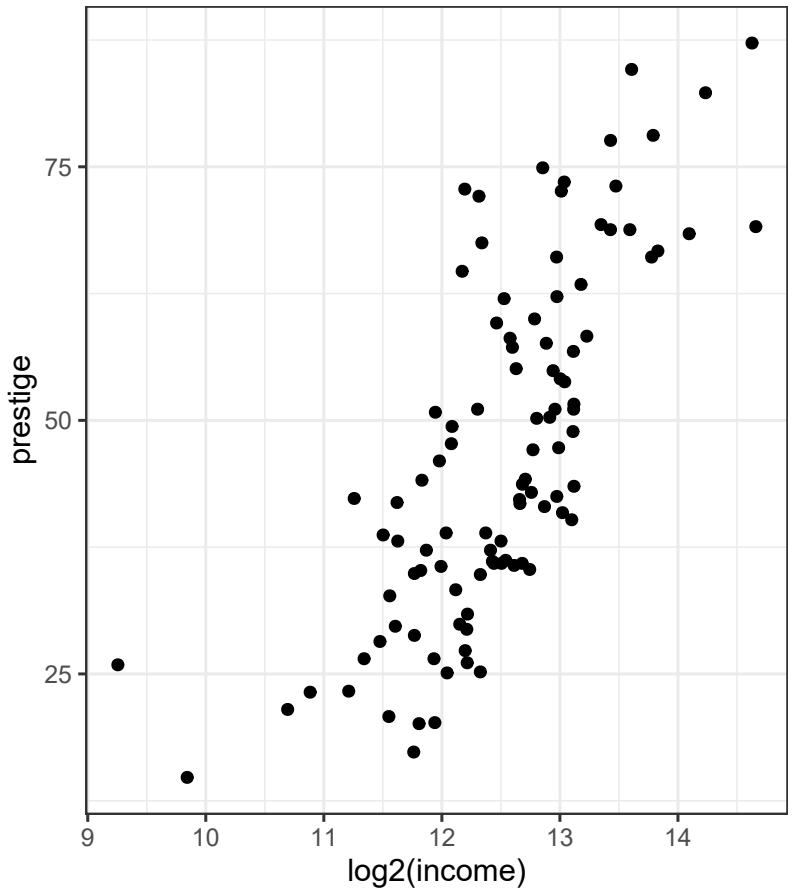
- 데이터의 요약 통계치를 보여주는 플롯
  - 플롯의 크기는 최소, 최대값에 따름.
  - `box`는 100분위 중 25분위와 75분위 사이에 해당하는 결과를 보여줌.
  - 가운데 선은 중앙값(평균이 아님!)
  - 박스 위 아래의 구간(whisker)은 이탈치가 아닌 값들 중 최대/최소값을 보여줌.
  - 이탈치들은 점/동그라미로 표현됨
- 박스플롯으로 데이터의 분포 그 자체에 대해서는 알 수 없음.

Scatterplots



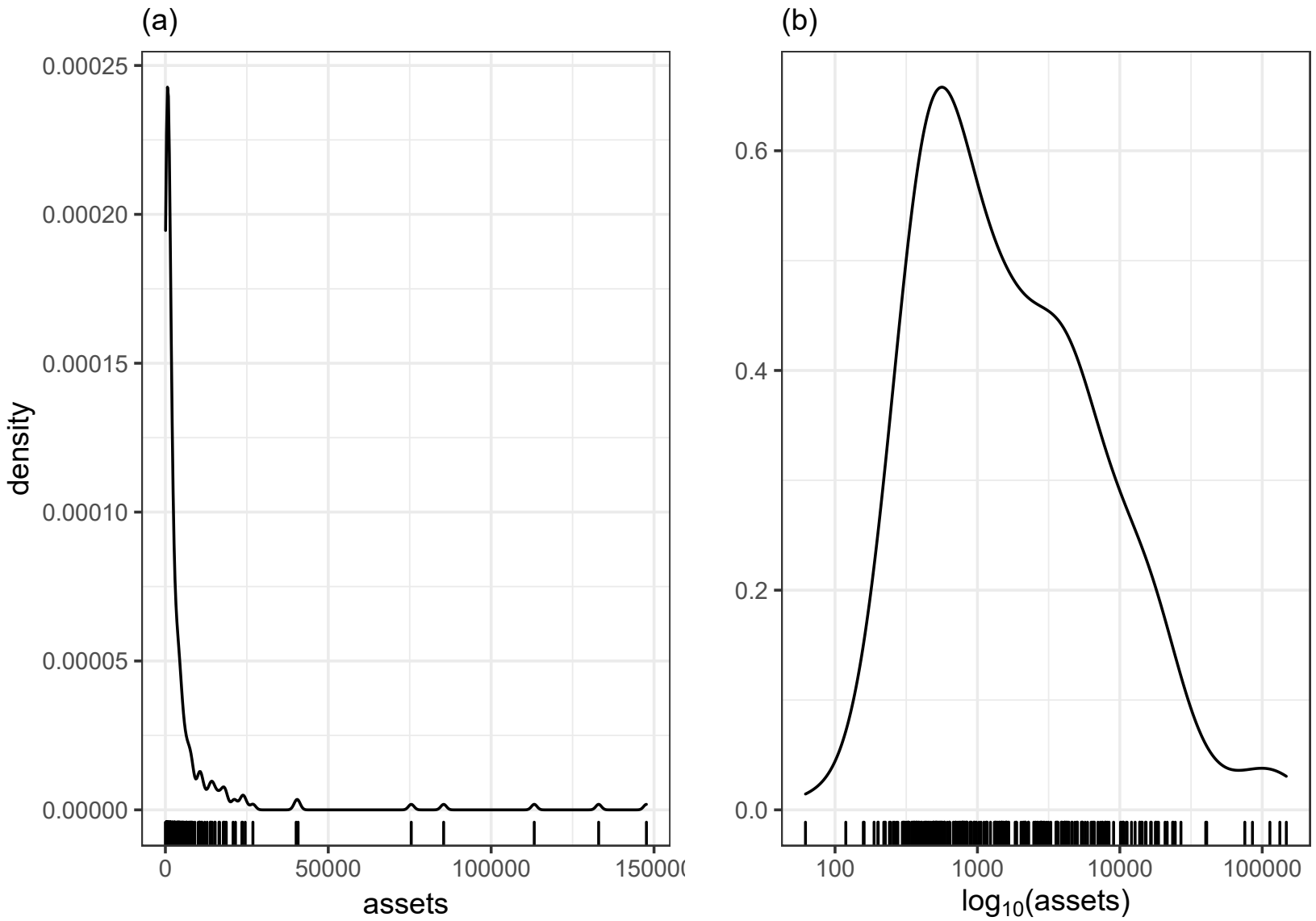
- 산포도(scatterplots)은 두 변수 간의 관계를 점으로 보여주는 플롯
- 일반적으로 연구대상인 변수를 Y 축에 놓고, X 축의 조건에 따른 변화를 살펴봄.
- 산포도를 조금만 손보면 더 많은 정보량을 제공할 수 있음.
  - 왜냐하면 한 변수의 분포가 굉장히 한 쪽으로 치우친 경우, 일반적인 산포도로는 두 변수 간의 관계를 제대로 보여주기 힘들 수 있음.
  - 이 경우, 문제가 되는 변수를 원자료 그대로 쓰기보다는 로그변환(비율) 등을 통해 척도 변환을 해줄 필요가 있음.

Scatterplots



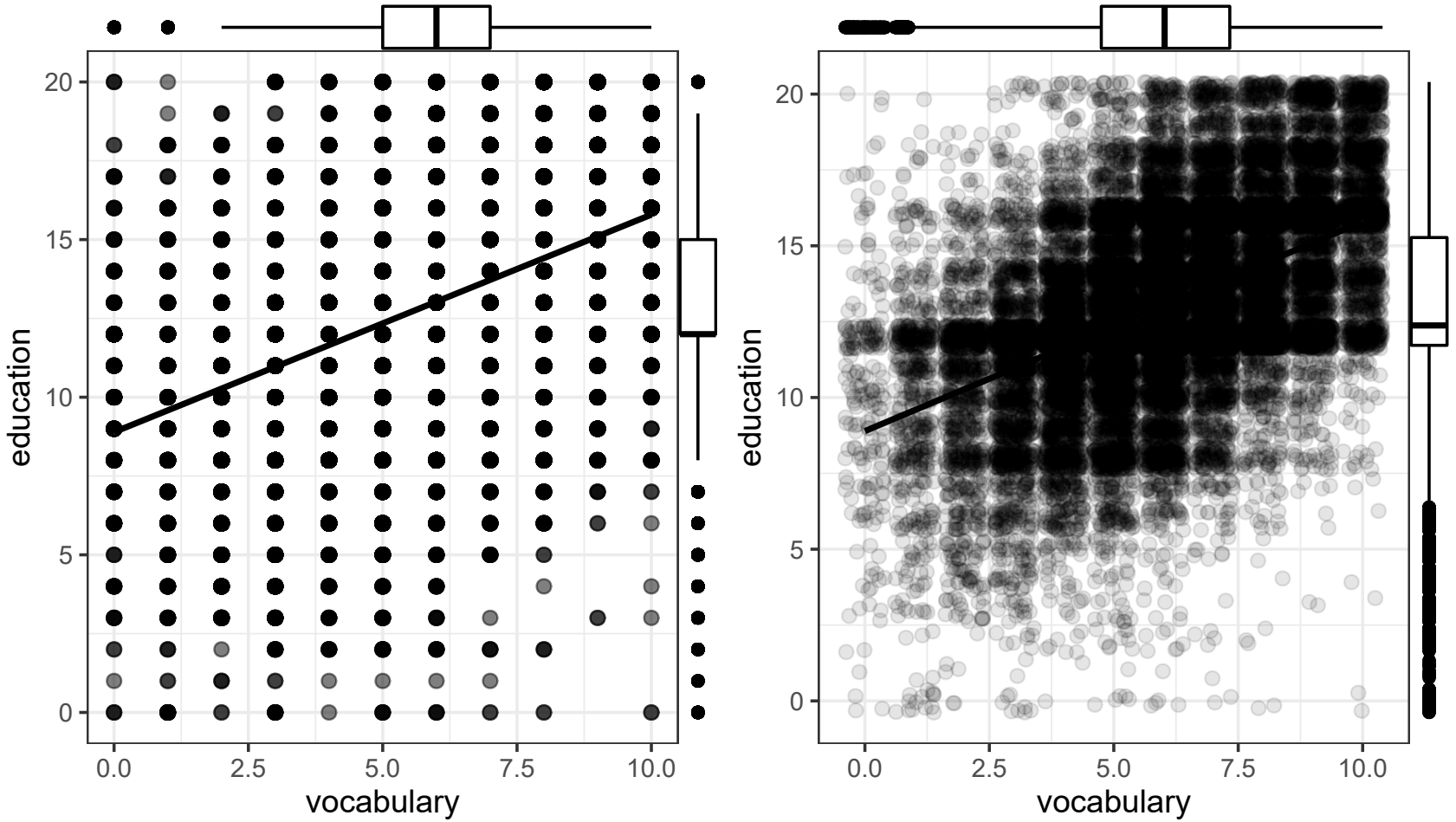
- 산포도(scatterplots)은 두 변수 간의 관계를 점으로 보여주는 플롯
- 일반적으로 연구대상인 변수를 Y 축에 놓고, X 축의 조건에 따른 변화를 살펴봄.
- 산포도를 조금만 손보면 더 많은 정보량을 제공할 수 있음.
    - 왜냐하면 한 변수의 분포가 굉장히 한 쪽으로 치우친 경우, 일반적인 산포도로는 두 변수 간의 관계를 제대로 보여주기 힘들 수 있음.
    - 이 경우, 문제가 되는 변수를 원자료 그대로 쓰기보다는 로그변환(비율) 등을 통해 척도 변환을 해줄 필요가 있음.

Scatterplots



Scatterplots

이산형 데이터의 경우, 관측치들이 중첩되어 퍼져있는 정도를 파악하기 어려울 수 있는데, 이 경우 `jitter`로 해결.

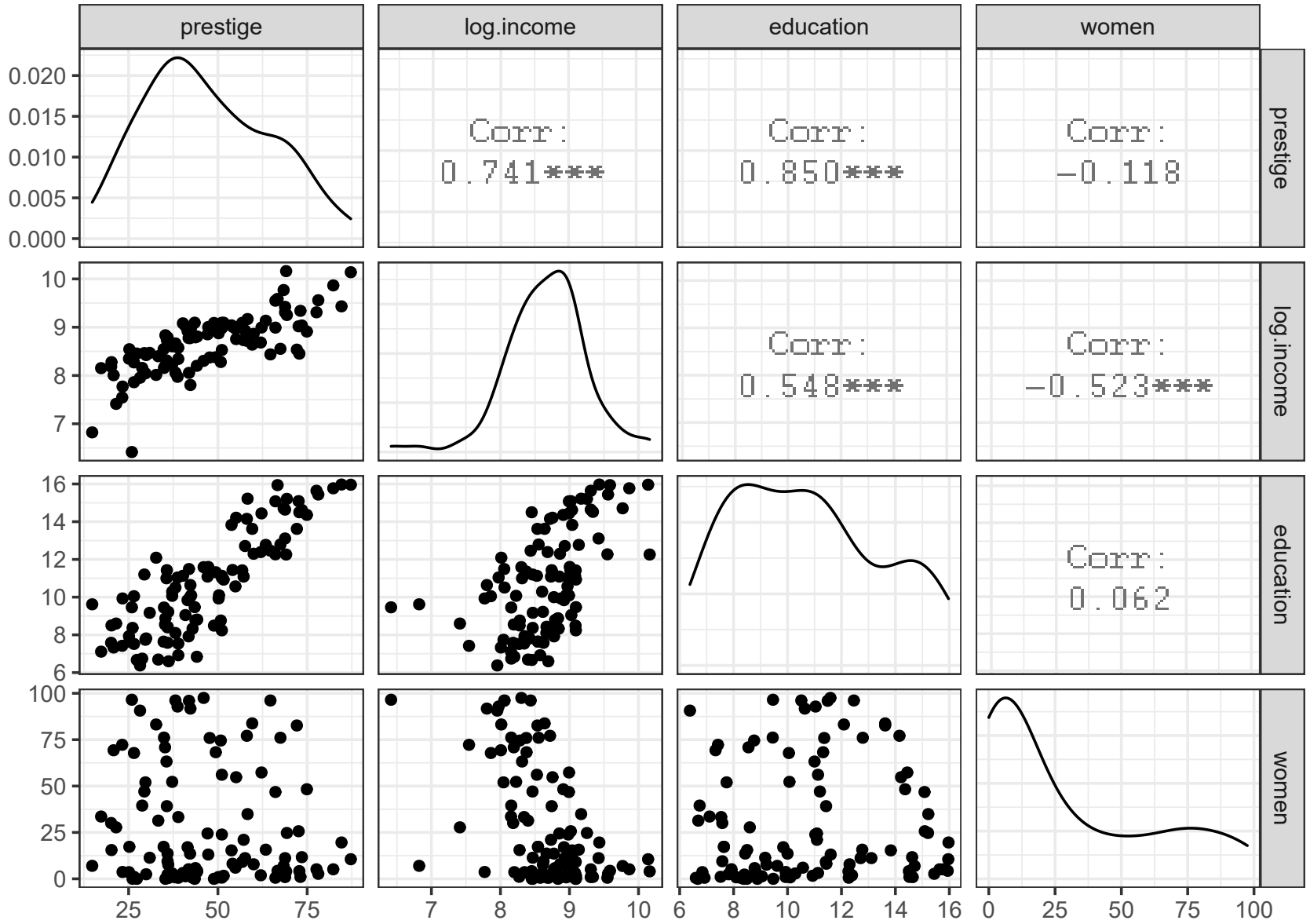


Transforming Data

왜 데이터를 변환(transform) 해야할까?

- 한쪽으로 치우친 분포를 가지고 있는 데이터는 많은 관측치들이 좁은 범위에 모여있기 때문에 분석이 어려움.
- 치우친 분포에서는 대개 비정상적으로 큰 값과 작은 값들이 나머지 값들을 살펴보기 어렵도록 짓누름(suppress).
- 분포를 요약해서 보여주는 통상의 통계 방법들은 대개 평균을 이용
    - 문제는 평균이 극단적인 값, 이탈치들에 민감한 통계치
    - 따라서 서로 다른 척도/범위를 가진 변수들을 비교하기란 어려울 수 있음.

Transforming Data



Transforming Data

거듭곱 변환(Power transformations)도 관계를 명확하게 보여주는 데 도움이 될 수 있음.

단순한 비선형관계는 종종 X, Y 또는 둘 모두를 거듭곱 변환을 함으로써 바로 잡을 수 있다. Mosteller와 Tukey의 's bulging rule은 선형화 변환을 선택하는 데 도움을 준다(Fox, 2016).

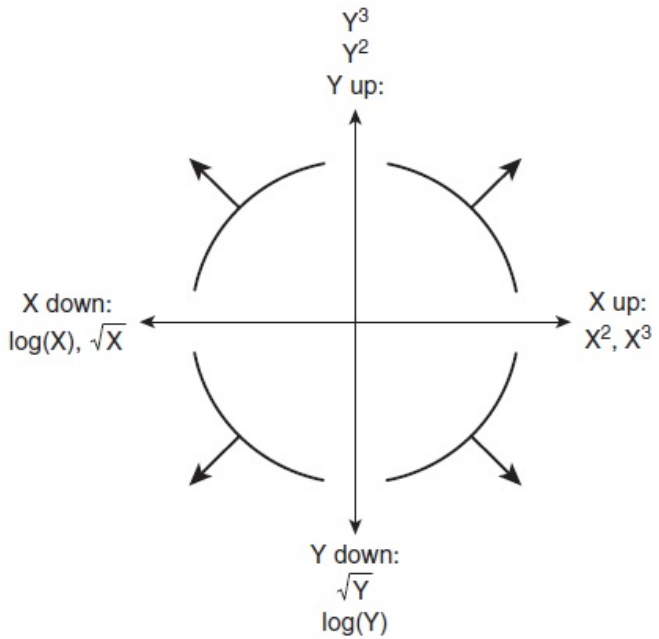


Figure 1: Illustration of Mosteller and Tukey's bulging rule

Transforming Data

```
fakedataMT <- function(p=1,q=1,n=99,s=.1){
  set.seed(1);
  X=seq(1/(n+1),1-1/(n+1),length=n)
  Y=(5+2*X^p+rnorm(n,sd=s))^ (1/q)
  return(data.frame(x=X,y=Y))}

test <- fakedataMT(p=3, q=5)
```

Transforming Data

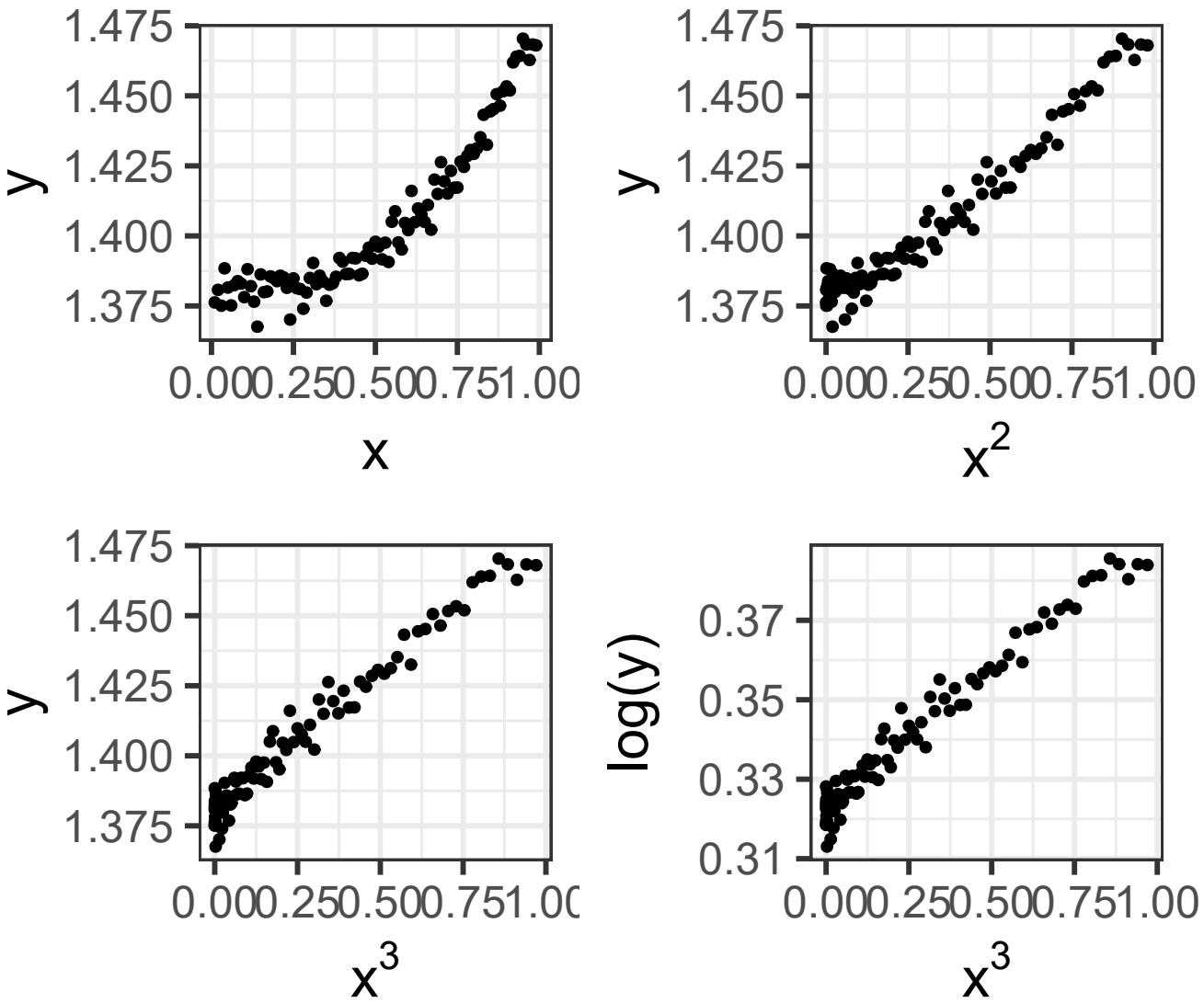


Figure 2: The effect of different power transformations of X and Y

Transformation Data

정규화(Normalization)

마지막으로 살펴볼 데이터 변환: 정규화 & 표준화

- 둘 모두 서로 다른 척도/단위의 변수를 동일한 척도로 변환하여 비교할 수 있게 해줌.
- 정규화: [0, 1] 사이의 범주로 데이터를 변환
  - 각 값에서 최소값을 뺀 이후에 최대값에서 최소값을 뺀 값으로 나누어줌.
  - 최소최대값이 계산에 반영되는 정규화는 이탈치에 민감
  - 대개 머신러닝에서는 사용하지만, 통계모델에서는 사용하지 않음.

$$\text{Normalization} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Transformation Data

표준화(Standarization)

표준화(또는 z-스코어 정규화)는 변수를 0으로 중심화(centering)하고 분산을 1로 표준화한다는 것을 의미

- 표준화: 각 관측치들로부터 평균을 빼고 그 값들을 표준편차로 나눔
  - 다른 척도를 가진 변수들이 동일한 표준정규분포의 특성을 가지도록 함.
  - 표준화 결과로 나타나는 최소값과 최대값은 변수가 어떻게 퍼져있는지에 따라서 다르고 이탈치(outliers)의 존재 여부에 매우 크게 영향을 받음.

$$\text{Standardization} = \frac{x - \bar{x}}{s}$$

Fox, John Jr. 2016. Applied Regression Analysis and Generalized Linear Models. 3 ed. CA: Thousand Oaks, SAGE Publications.

Table of Contents