역사 데이터 시각화 분석

Regression Analysis with R

What is Regression Analysis?

- 개념(Concept)

- 독립변수와 종속변수 간에 존재하는 연관성을 분석하기 위하여 관측된 자료에서 이들간의 함수적 관계식을 통계적으로 추정하는 방법이다.
- 독립변수의 수에 따라 단순회귀분석과 다중회귀분석으로 나뉜다.
- 회귀분석을 시행하기에 앞서 기본 가정(정규성, 등분산성, 독립성)들이 모두 충족 되어야 한다.
- 회귀분석은 잔차(측정값-실제값)의 제곱의 합을 최소로 하는 최소 제곱법을 사용한다.
- 독립변수에 의해 설명되는 종속변수의 비율 값으로 결정계수를 사용한다.
- 회귀모형의 유의성을 검증하기 위해 F값을 사용하고 회귀계수의 유의성을 검증하기 위해 T값을 활용한다.

- 회귀 분석 용어 정리

- 종속변수 : 분석의 대상이 되는 변수
- 독립변수 : 종속변수에 영향을 미치는 변수
- 잔차 : 추정 값과 실제 값의 차이값
- 정규성 : Q-Q plot에서 두 변수가 유사한 정도
- 등분산성 : 잔차들을 사용한 산점도에서 잔차들이 고루 퍼져 있는 정도
- 독립성 : 더빈 왓슨 값이 0~4이내이고 2일 때는 가장 독립성을 만족한다.

What is Regression Analysis?

- 회귀모형식

$$\widehat{Y_i}=\widehat{b_0}+\widehat{b_1}X_i+e_i \qquad b_0=\bar{Y}-b_1\bar{X},\ b_1=\frac{S_{xy}}{S_{xx}},\ e_i=\widehat{Y_i}-Y_i$$

- 분산 분석표

Source(요인)	Df(자유도)	SS(제곱합)	Ms(평균제곱)	F
Reg(회귀)	K-1	SSR	MSR=$\frac{SSR}{k-1}$	F=$\frac{MSR}{MSE}$
Error(오차)	N-(k-1)-1	SSE	MSE=$\frac{SSE}{n-k-1}$	
Total	N-1	SST		

K=변수의 수
N=총 데이터의 수

$$S_{xy} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum {X_i}^2 - n\bar{X}^2$$
$$S_{xx} = (X_i - \bar{X})^2 = \sum {X_i}^2 - n\bar{X}^2$$
$$S_{yy} = (Y_i - \bar{Y})^2 = \sum {Y_i}^2 - n\bar{Y}^2$$

$SST=S_{yy}$

$SSR={b_1}^2 S_{xx}=b_1 S_{xy}$

$SSE=SST-SSR$

What is Regression Analysis?

- T값, F값, R^2(결정계수)값 관련 공식

$\mathrm{SE}(b_1)$, (표준오차)$= \sqrt{\frac{MSE}{S_{xx}}}$

T값$= \frac{\widehat{b_1} - 0}{SE(\widehat{b_1})} = \frac{b_1}{SE(b_1)}$

F값$= (t\text{값})^2$

R^2(결정계수)$= \frac{SSR}{SST}$

- 잔차에 대한 가정사항

- 잔차들의 합은 0이다.
- 잔차들의 X_i에의한 가중합은 0이다.
- 잔차들의 $\widehat{X_i}$에의한 가중합은 0이다.
- 잔차에 의해 생성된 회귀식은 항상 평균점$(\overline{X}, \overline{Y})$을 지난다.

- 예제(Example)
  - 9575명의 몸무게와 체질량 지수에 대하여 기록된 데이터

- 변수 설명
  - Weight = 몸무게에 대한 수치형 데이터(kg)
  - BMI = 체질량 지수($\frac{Weight}{Height^2}$)

```
> getwd()
[1] "/Users/Seongmin_M/Downloads"
>
> setwd("/Users/Seongmin_M/Downloads")
>
> NHANES<-read.csv("NHANES_1.csv",head=T)
>
> head(NHANES)
  X Weight  BMI
1 1   82.7 29.4
2 2   85.6 29.6
3 3   71.5 23.1
4 4   93.8 30.0
5 5   81.6 29.7
6 6   68.3 21.9
>
> str(NHANES)
'data.frame':   9575 obs. of  3 variables:
 $ X     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Weight: num  82.7 85.6 71.5 93.8 81.6 68.3 67.7 86.5 61.7 85 ...
 $ BMI   : num  29.4 29.6 23.1 30 29.7 21.9 26.3 31.9 20.9 26.7 ...
```

- 데이터 확인

```
> getwd()
[1] "/Users/Seongmin_M/Downloads"
>
> setwd("/Users/Seongmin_M/Downloads")
>
> NHANES<-read.csv("NHANES_1.csv",head=T)
>
> head(NHANES)
  X Weight  BMI
1 1   82.7 29.4
2 2   85.6 29.6
3 3   71.5 23.1
4 4   93.8 30.0
5 5   81.6 29.7
6 6   68.3 21.9
>
> str(NHANES)
'data.frame':   9575 obs. of  3 variables:
 $ X     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Weight: num  82.7 85.6 71.5 93.8 81.6 68.3 67.7 86.5 61.7 85 ...
 $ BMI   : num  29.4 29.6 23.1 30 29.7 21.9 26.3 31.9 20.9 26.7 ...
```

- Head함수와 str함수를 사용하여 데이터의 형태를 확인하여 준다.

- 회귀식 생성

```
> lm.NHANES=lm(NHANES$BMI~NHANES$Weight)
>
> lm.NHANES

Call:
lm(formula = NHANES$BMI ~ NHANES$Weight)

Coefficients:
  (Intercept)  NHANES$Weight
       5.3798         0.2618
```

  - 회귀식은 BMI=5.3798+0.2618Weight이다.

- 회귀모형 검증

```
> anova(lm.NHANES)
Analysis of Variance Table

Response: NHANES$BMI
                Df Sum Sq Mean Sq F value    Pr(>F)
NHANES$Weight    1  48813   48813   13305 < 2.2e-16 ***
Residuals     3705  13593       4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

  - p값이 0.05이하 이므로 회귀모형은 유의 하다.

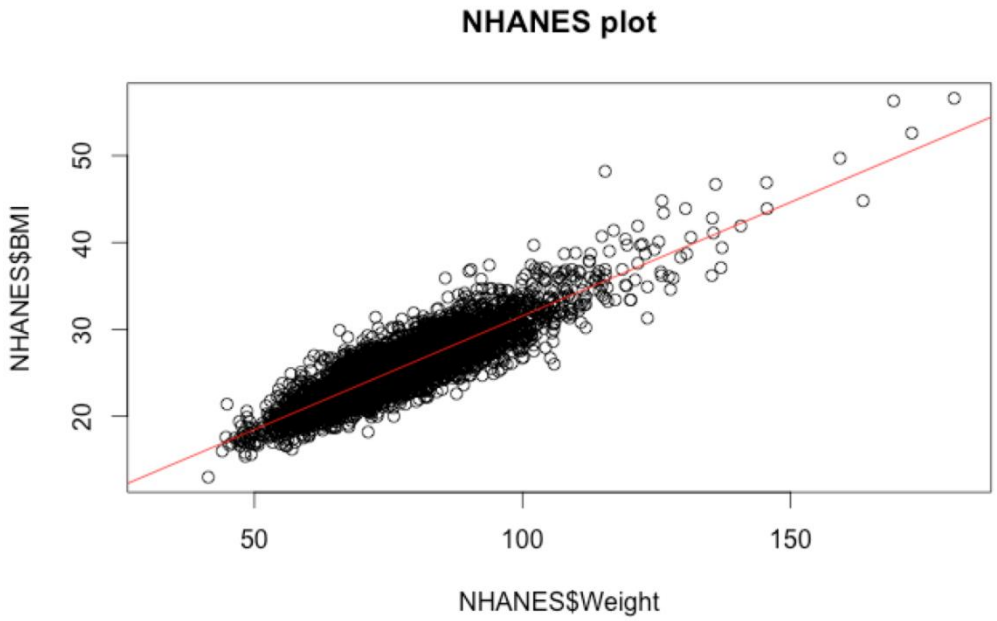
- 회귀계수 검증

```
> summary(lm.NHANES)

Call:
lm(formula = NHANES$BMI ~ NHANES$Weight)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1021 -1.3043 -0.0928  1.2004 12.6110

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.37983    0.17969   29.94   <2e-16 ***
NHANES$Weight  0.26178    0.00227  115.35   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.915 on 3705 degrees of freedom
  (5868 observations deleted due to missingness)
Multiple R-squared:  0.7822,    Adjusted R-squared:  0.7821
F-statistic: 1.33e+04 on 1 and 3705 DF,  p-value: < 2.2e-16
```

- 상수항과 독립변수의 p값이 모두 0.05이하이므로 회귀계수는 유의하다.
- 설명력은 78.21%의 설명력을 지니고 있다.

- 공분산성 검증

```
> vcov(lm.NHANES)
                 (Intercept) NHANES$Weight
(Intercept)    0.0322871504 -4.015012e-04
NHANES$Weight -0.0004015012  5.150687e-06
```



- 공분산 값(절대값)이 10 이하이고 점들이 한 직선상에 일치하지 않으므로 다중 공선성은 없다고 볼 수 있다.

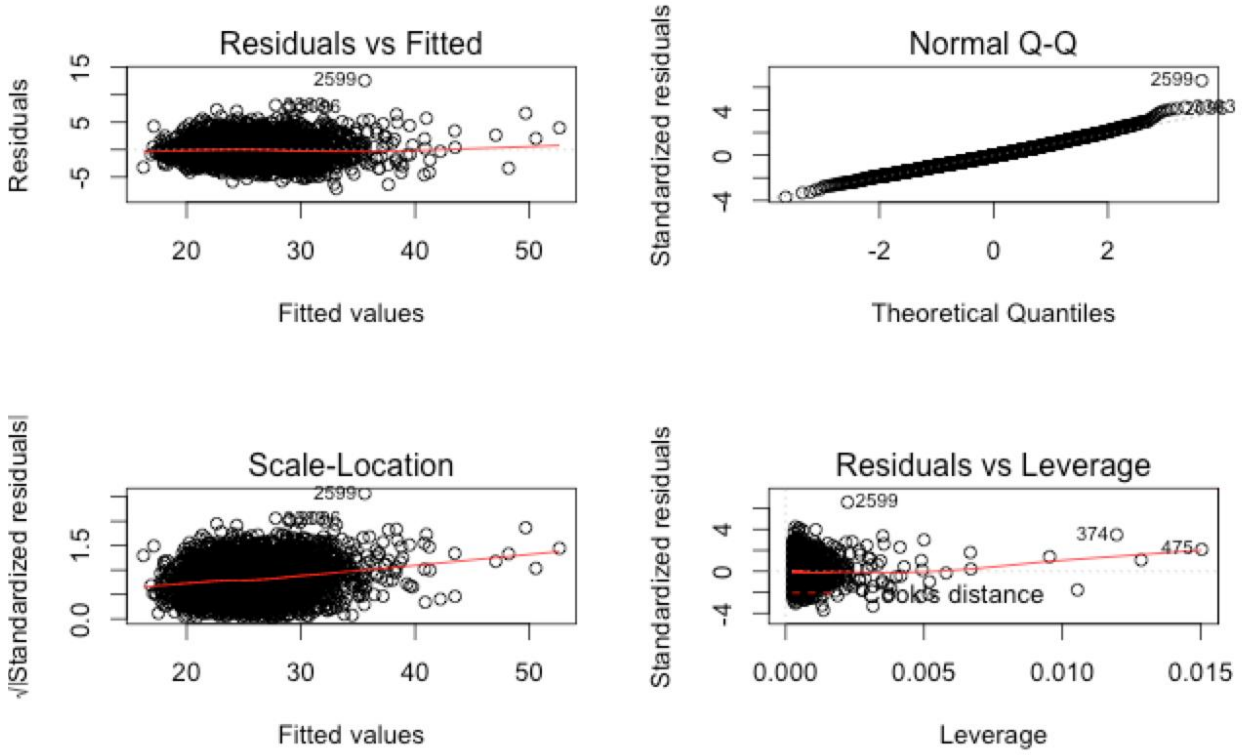
- 독립성 검증

```
> install.packages("lmtest")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/lmtest_0.9-33.tgz'을 시도합니다
Content type 'application/x-gzip' length 266752 bytes (260 Kb)
URL을 열었습니다
==================================================
downloaded 260 Kb


The downloaded binary packages are in

/var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpXqG9Ju/downloaded_packages
> library(lmtest)
필요한 패키지를 로딩중입니다: zoo

다음의 패키지를 부착합니다: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric


다음의 패키지를 부착합니다: 'lmtest'

The following object is masked from 'package:RCurl':

    reset

>
> dwtest(lm.NHANES)

        Durbin-Watson test

data:  lm.NHANES
DW = 1.9466, p-value = 0.05209
alternative hypothesis: true autocorrelation is greater than 0
```

- Durbin-watson값이 1.9466이므로 독립성을 충족한다.

- 등분산성, 정규성 검증

```
> par(mfrow=c(2,2))
>
> plot(lm.NHANES)
```



- 잔차가 상하에 고루 분포하므로 등분산성이 만족한다고 할 수 있다.
- QQ도표는 직선의 형태를 띄므로 정규성은 만족한다.

- 이상치 확인

```
> install.packages("car")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/car_2.0-25.tgz'을 시도합니다
Content type 'application/x-gzip' length 1386779 bytes (1.3 Mb)
URL을 열었습니다
==================================================
downloaded 1.3 Mb


The downloaded binary packages are in

/var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpzB2XaL/downloaded_packages
> library(car)
경고메시지:
패키지 'car'는 R 버전 3.1.3에서 작성되었습니다
>
> outlier.test(lm.NHANES)
     rstudent unadjusted p-value Bonferonni p
2599 6.629329         3.8603e-11    1.431e-07
```

- 2599번이 이상치 임을 확인 할 수 있으며 표에서도 나타내 주고 있다.

- 이상치 확인

```
> NHANES[2599,]
        X Weight  BMI
2599 2599  115.4 48.2
>
> lm.NHANES$fitted[2599]
    2599
35.58903
>
> lm.NHANES$residuals[2599]
    2599
12.61097
```

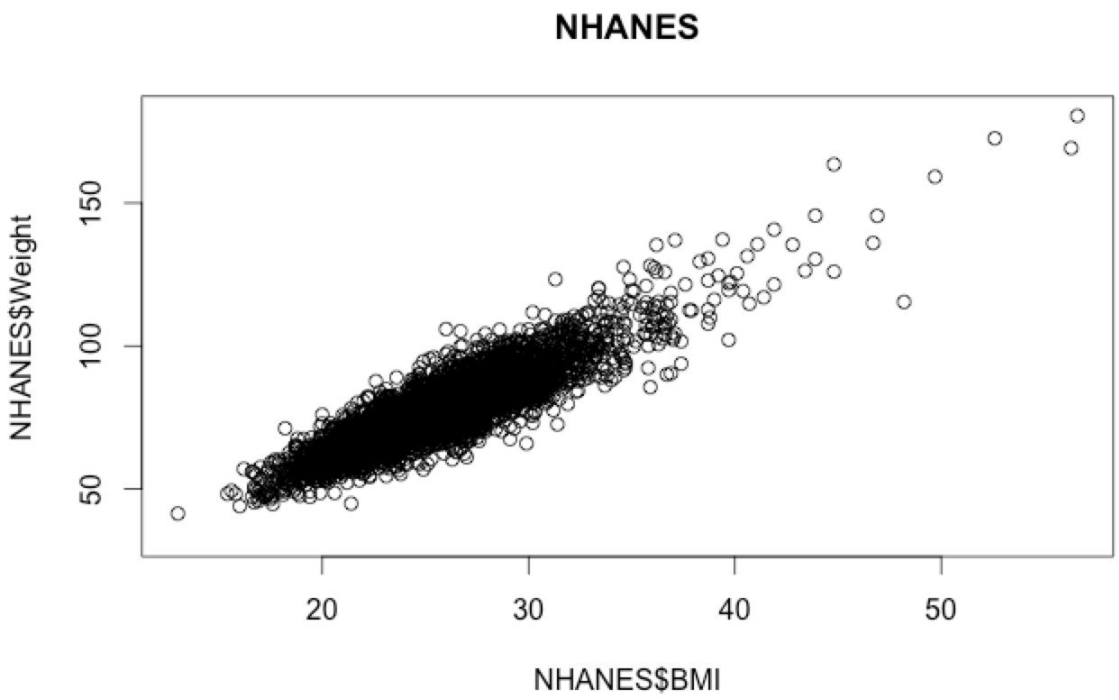
  - 이상치의 BMI는 48.2이고 Weight는 115.4이다.
  - 회귀식을 대입하여 BMI를 구한 결과 35.58903이 나올 것으로 예상 되었고 이는 실제값과 많은 차이가 난다.
  - 2599번의 잔차는 12.61097 로 매우 높다.
  - 그러므로 이상치를 빼고 다시 분석 할 필요가 있다.

- 결론

  - 계수와 모형, 독립성, 등분산성, 정규성을 만족하므로 회귀 모형을 사용할 수 있다.

- 일반 시각화(산점도)

```
> par(mfrow=c(1,1))
>
> plot(NHANES$BMI,NHANES$Weight, main="NHANES")
```

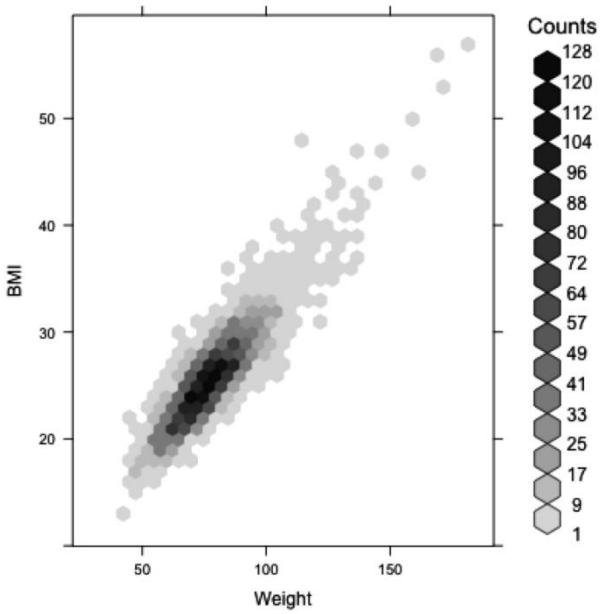


- 데이터의 수가 많고 겹치는 부분이 많아 데이터의 특성을 파악하기 힘들다.

Hebin visualization with R

- Hebin패키지를 활용한 시각화

```
> install.packages("hebin")
Warning in install.packages :
  package 'hebin' is not available (for R version 3.1.2)
>
> library(hexbin)
>
> hexbinplot(BMI ~ Weight, data=NHANES)
```

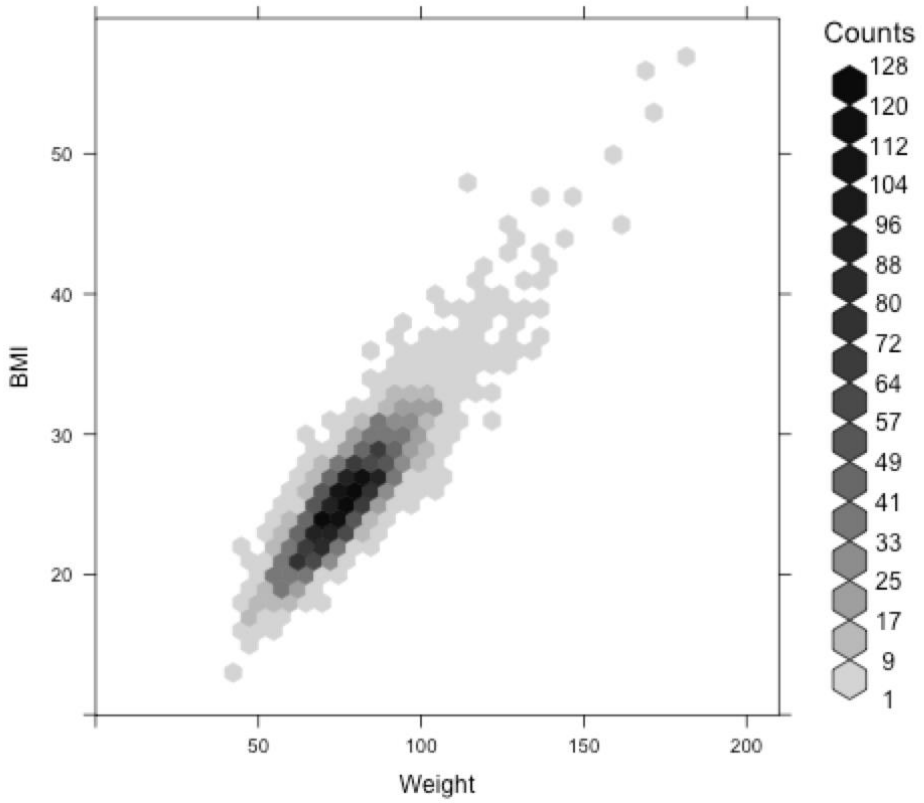


- 데이터가 겹치는 부분에 명도를 달리하여 표시하였다.

Hebin visualization with R

- 창크기 확대

```
> hexbinplot(BMI ~ Weight, data=NHANES, xlim=c(0,210))
```

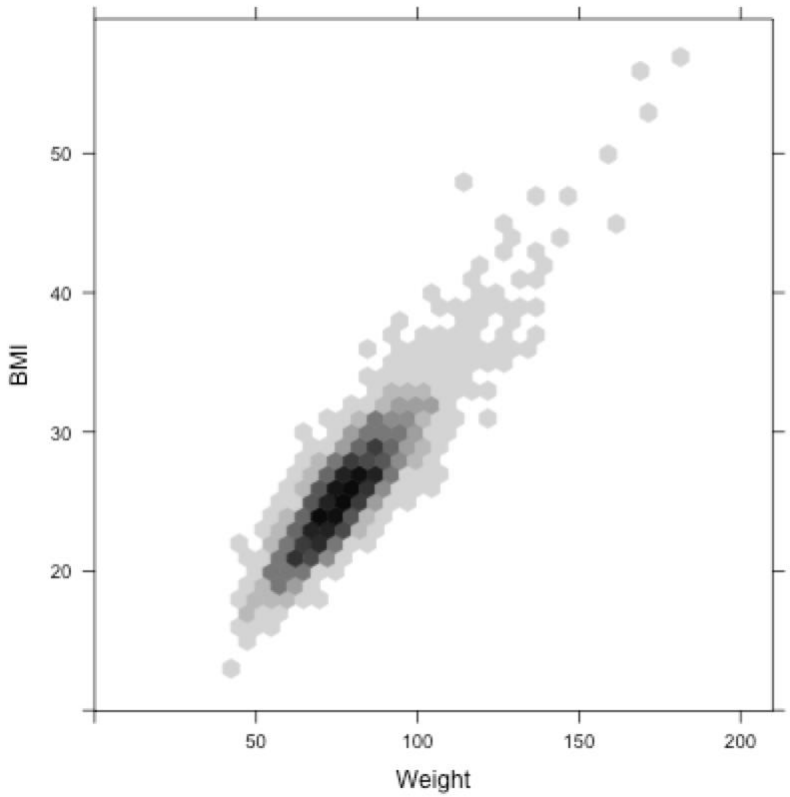


- 창의 크기를 확대하였다.

Hebin visualization with R

- 범례 삭제

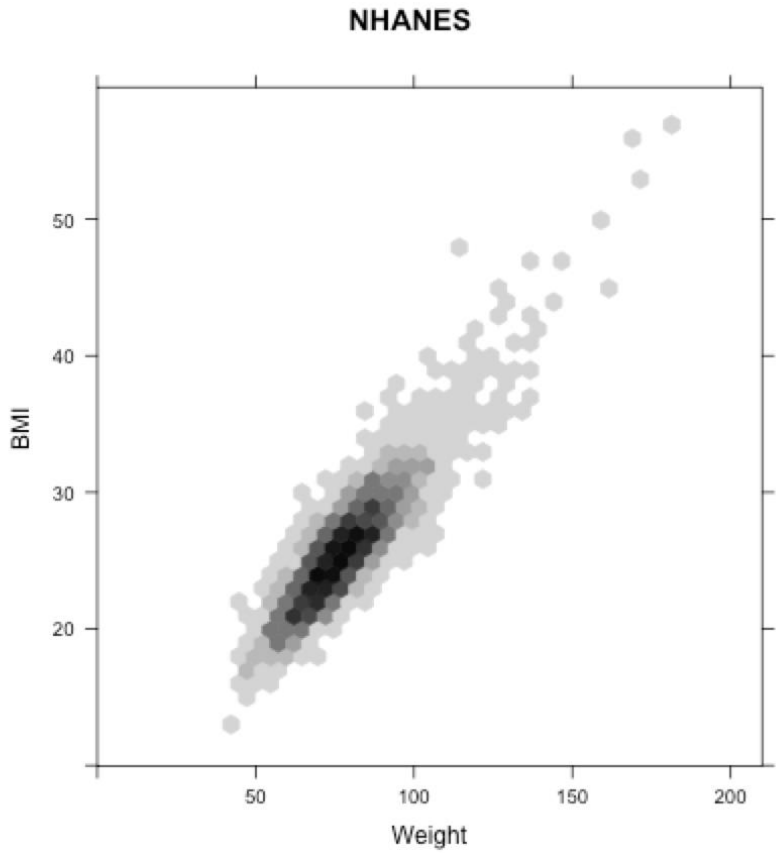
```
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, data=NHANES)
```



- 범례를 삭제하였다.

- 제목 삽입

```
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, data=NHANES, main="NHANES")
```

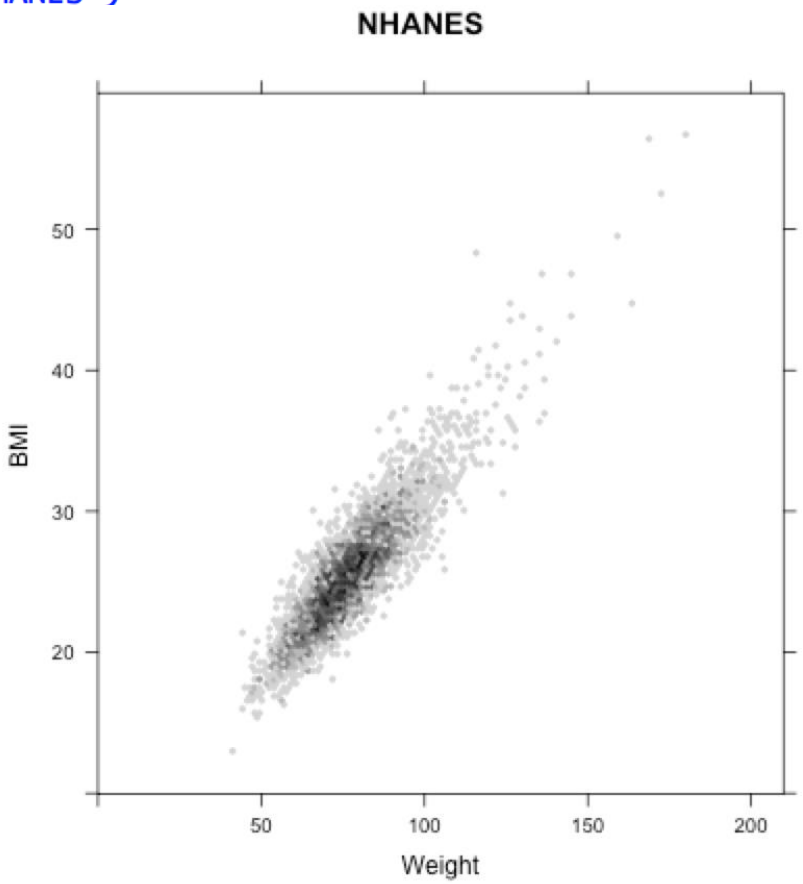


- 제목을 삽입하였다.

Hebin visualization with R

- 점크기 조정

```
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, xbins=100, data=NHANES, main="
NHANES")
```

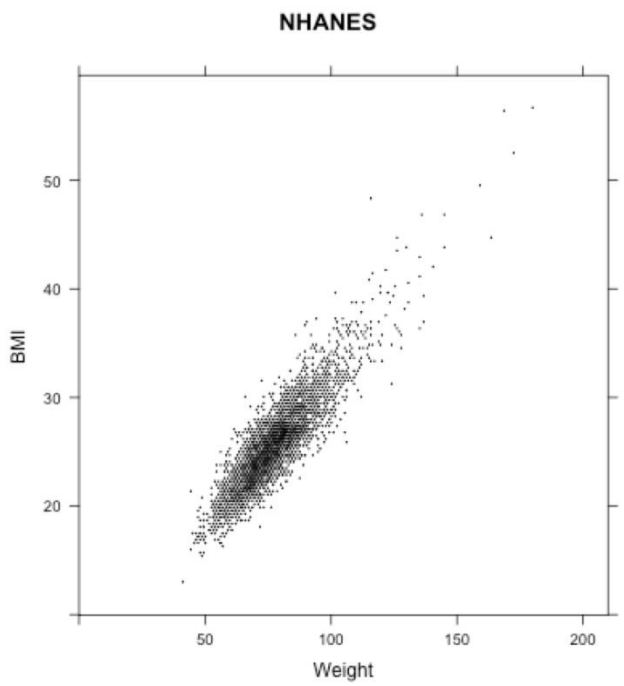


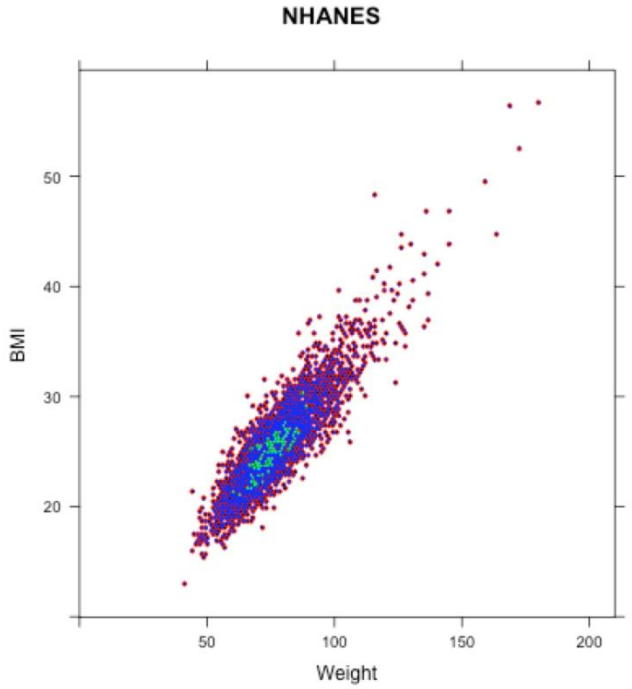
- 점의 크기를 축소하였다.

Hebin visualization with R

- 스타일 조정

```
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, xbins=100, data=NHANES, style
= "lattice",main="NHANES")
>
> hexbinplot(BMI ~ Weight, xlim=c(0,210), colorkey=F, xbins=100, data=NHANES, style
= "nested.centroids",main="NHANES")
```





- 스타일 변경에 따라 표현법이 상이해 지는 것을 확인 할 수 있다.