- 예제: Ericksen

  - 변수 설명

    1) minority: Percentage black or Hispanic
    2) crime: Rate of serious crimes per 1000 population
    3) poverty: Percentage poor
    4) language: Percentage having difficulty speaking or writing English
    5) highschool: Percentage age 25 or older who had not finished high school
    6) housing: Percentage of housing in small, multiunit buildings
    7) city: A factor with levels: city, major city; state, state or state-remainder

  - 실습 내용

    1) crime을 반응변수, 나머지를 설명변수로 하는 회귀모형 적합
       - 임의로 선택한 6개 케이스는 test data / 나머지 60개 케이스로 적합
    2) 가정 만족 여부 확인
    3) 필요한 추론 실시
       - 모수의 유의성 검정
       - test data에 대한 예측 실시 및 실제 자료와의 오차 비교

- 자료 확인

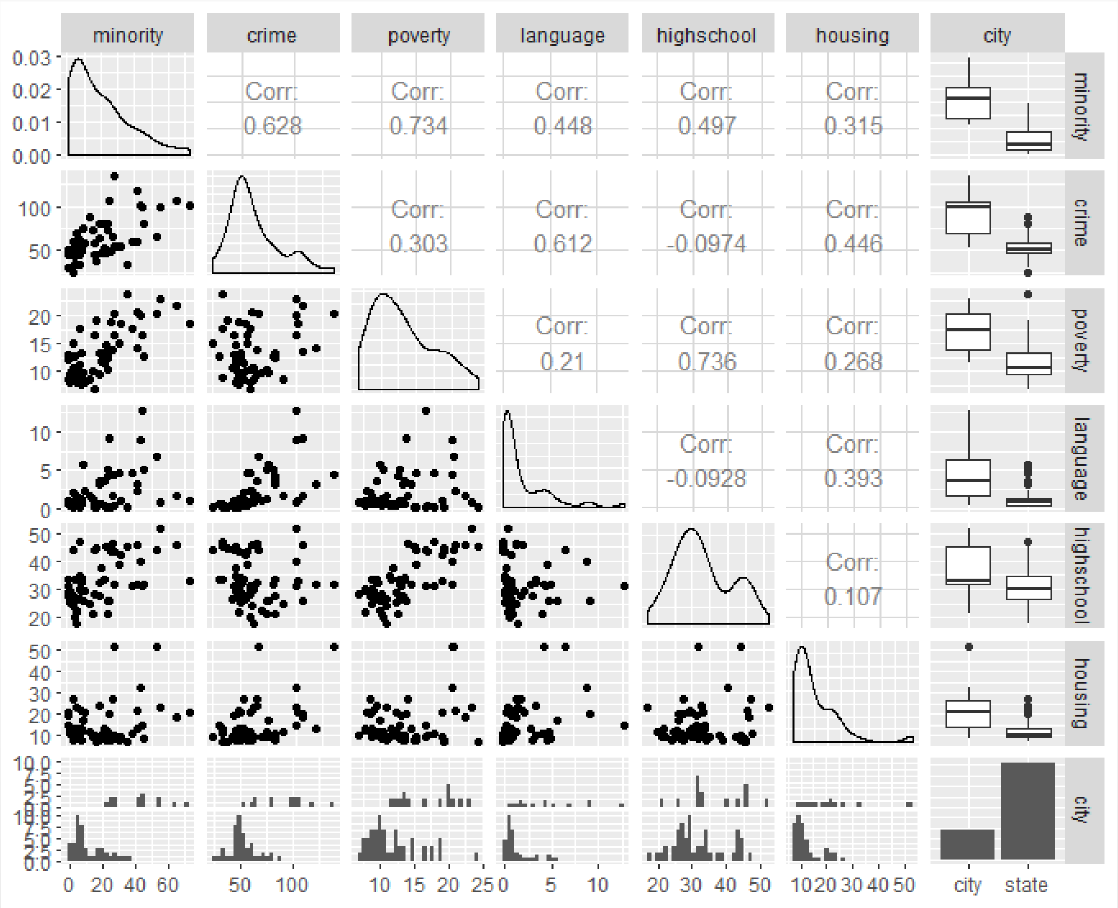
```
> library(car)

> str(Ericksen)
'data.frame':   66 obs. of  9 variables:
 $ minority    : num  26.1 5.7 18.9 16.9 24.3 15.2 10.8 17.5 22.3 ...
 $ crime       : int  49 62 81 38 73 73 58 68 81 55 ...
 $ poverty     : num  18.9 10.7 13.2 19 10.4 10.1 8 11.8 13.4 ...
 $ language    : num  0.2 1.7 3.2 0.2 5 1.2 2.4 0.7 3.6 0.3 ...
 $ highschool  : num  43.5 17.5 27.6 44.5 26 21.4 29.7 31.4 33.3 ...
 $ housing     : num  7.6 23.6 8.1 7 11.8 9.2 21 8.9 10.1 10.2 ...
 $ city        : Factor w/ 2 levels "city","state": 2 2 2 2 2 2 2 ...
 $ conventional: int  0 100 18 0 4 19 0 0 0 0 ...
 $ undercount  : num  -0.04 3.35 2.48 -0.74 3.6 1.34 -0.26 -0.16 ...
```

- 전체 data set을 test data와 training data로 분리

```
> library(dplyr)

> Ericksen_1 <- select(Ericksen, -undercount, -conventional)

> set.seed(1234)

> x.id <- sample(1:nrow(Ericksen_1), size=6)

> df_1 <- Ericksen_1[x.id,]

> df_2 <- Ericksen_1[-x.id,]
```

- Scatterplot matrix 작성

```
> library(GGally)
> ggpairs(df_2)
```




우측으로 치우친 분포
crime, language, housing

변환이 필요하거나 혹은 영향력이 큰 관찰값이 있는 변수

- 회귀모형 적합

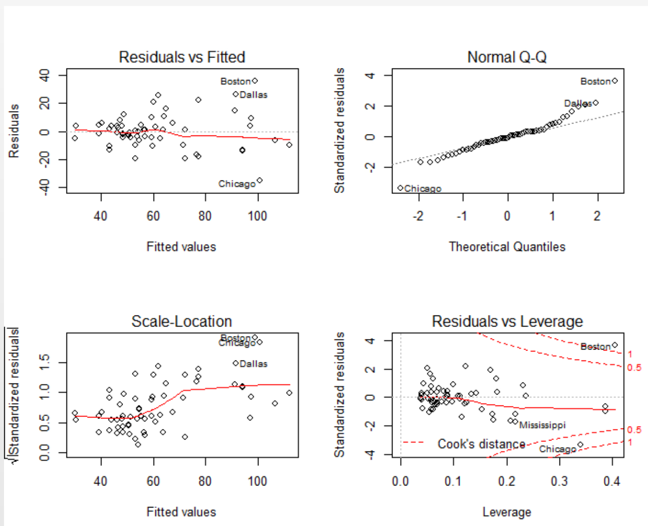
```
> fit_1 <- lm(crime ~ . , data=df_2)
> summary(fit_1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.4297    11.4709   8.145  6.57e-11 ***
minority      0.6436     0.1939   3.320  0.00163 **
poverty       0.5446     0.7337   0.742  0.46117
language      1.1743     0.8572   1.370  0.17652
highschool   -1.3885     0.3299  -4.208  9.99e-05 ***
housing       0.2883     0.2242   1.286  0.20408
citystate   -15.7358     6.8293  -2.304  0.02516 *
---
Residual standard error: 12.85 on 53 degrees of freedom
Multiple R-squared:   0.72,   Adjusted R-squared:  0.6883
F-statistic: 22.72 on 6 and 53 DF,  p-value: 4.665e-13
```

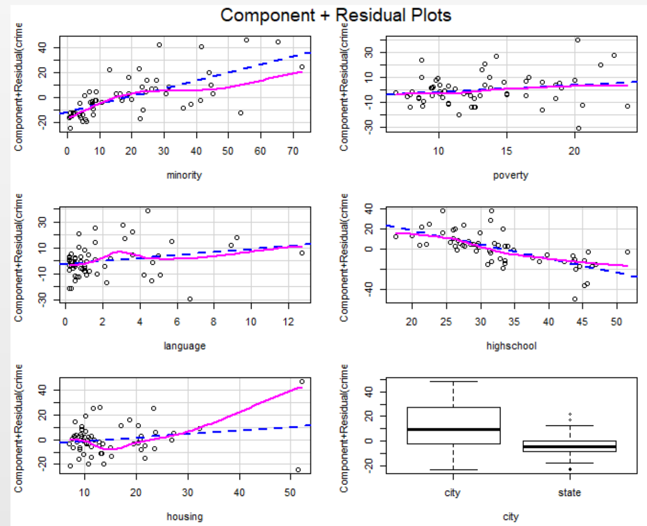
- 변수 선택 과정 생략

- 가정 만족 여부 확인

```
> par(mfrow=c(2,2))
> plot(fit_1)
> par(mfrow=c(1,1))
```

```
> crPlots(fit_1)
```




- 분산 증가에 대한 의심
- 독립성을 포함한 다른 가정에는 큰 문제가 없는 것으로 보임
- 다중공선성도 없는 것으로 보임

- test data에 대한 예측과 실제 자료와의 비교

```
                    pred crime
Delaware        55.17261    68
South.Dakota    46.83667    32
Rhode.Island    44.73651    59
South.Carolina  44.91122    53
Cleveland       79.71923   101
Saint.Louis     75.81259   143
```

pred: 예측값

crime: 실제 자료

```
> pred <- predict(fit_1, newdata=df_1[-2])

> res <- cbind(pred, df_1[2])

> forecast::accuracy(pred, df_1[[2]])
              ME     RMSE      MAE      MPE     MAPE
Test set 18.13519 30.60655 23.08075 13.33179 28.78665
```

df_1[-2]: crime을 제외한 데이터 프레임
df_1[2]: crime만으로 이루어진 데이터 프레임
df_1[[2]]: crime을 벡터로 추출

# 예측 정확성 측도

- 자료의 크기에 영향을 받는 측도
  - Mean absolute error(MAE) = $mean(\ |e_i|\ )$
  - Root mean squared error(RMSE) = $\sqrt{mean\left(e_i^2\right)}$

- 자료의 크기에 영향을 받지 않는 측도
  - Mean absolute percentage error(MAPE) = $mean\left(100 \times \left|\frac{e_i}{Y_i}\right|\right)$

# 2. 특이한 관찰값 탐지

- 특이한 관찰값

  1) 이상값: 추정된 회귀모형으로 설명이 잘 안 되는 관찰값
  2) 영향력이 큰 관찰값: 회귀계수 추정에 과도한 영향을 미치는 관찰값

- 특이한 관찰값 탐지

  - 유용한 통계량: DFBETAS, DFFITS, Covariance ratio, Cook's distance, Leverage

  - 유용한 R 함수: influence.measures( ), infIndexPlot( ), dfbetaPlots( ), avPlots( ), influencePlot( ), outlierTest( )

- Leverage

  - Hat matrix: $H = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
  - $h_{ii}$: Hat matrix의 $i$번째 대각원소
  - $\hat{Y}_i = h_{ii}Y_i + \sum_{j\neq i}^{n} h_{ij}Y_j$: $h_{ii}$는 $\hat{Y}_i$에 대한 $Y_i$의 영향력 또는 일종의 가중값
  - $i$번째 설명변수 자료($\boldsymbol{x}_i$)와 전체 설명변수 자료의 평균($\bar{\boldsymbol{x}}$) 사이의 표준화된 거리의 개념
  - $\frac{1}{n} \leq h_{ii} \leq 1$
  - $\bar{h} = \sum h_{ii}/n = (k+1)/n$

- DFFITS

$$DFFITS = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

- $Var(\hat{Y}_i) = h_{ii}\sigma^2$
- $h_{ii}$: Hat matrix의 $i$번째 대각 원소.
  변수 $X_i$의 leverage

- DFFTIS 값이 큰 관찰값: 영향력이 크다고 할 수 있음
- 통계량의 분포를 알 수 없음: 절대적인 판단 기준을 사용하는 것은 바람직하지 않음
- 상대적으로 큰 값의 관찰값 확인

- DFBETAS

$$DFBETAS_j = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)} c_{j+1,j+1}}}$$

- $\hat{\beta}_{j(i)}$: $(X_i, Y_i)$를 제외한 나머지 자료로 추정된 $\hat{\beta}_j$
- $Var(\hat{\beta}_j) = \sigma^2 c_{j+1,j+1}$
- $c_{j+1,j+1}$: $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$의 $j+1$번째 대각 원소

- $DFBETAS_j$의 값이 큰 관찰값: $\beta_j$의 추정에 큰 영향을 주었다는 의미
- 절대적인 판단 기준을 사용하는 것은 바람직하지 않음

- Cook's Distance

$$COOKD = \frac{\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)^T \boldsymbol{X}^T \boldsymbol{X}\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)}{(k+1)MSE}$$

- 모든 회귀계수에 대한 영향력
- Cook's Distance가 큰 관찰값: 회귀계수 벡터의 추정에 큰 영향력
- DFBETAS를 종합한 하나의 숫자: 더 보편적인 통계량

- Covariance ratio

$$COVRATIO = \frac{\left|\left(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)}\right)^{-1} MSE_{(i)}\right|}{\left|(\boldsymbol{X}^T \boldsymbol{X})^{-1} MSE\right|}$$

  - $Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1}\sigma^2$
  - 하나의 숫자로 precision 표시
  - 공분산 행렬을 하나의 숫자로 표현하는 방법: 행렬식

- COVRATIO의 값 > 1: $(X_i, Y_i)$가 $\widehat{\boldsymbol{\beta}}$의 precision을 향상시키는 방향으로 영향
- COVRATIO의 값 < 1: $(X_i, Y_i)$가 $\widehat{\boldsymbol{\beta}}$에 나쁜 영향

- R 함수의 활용: 함수 influence.measures( )

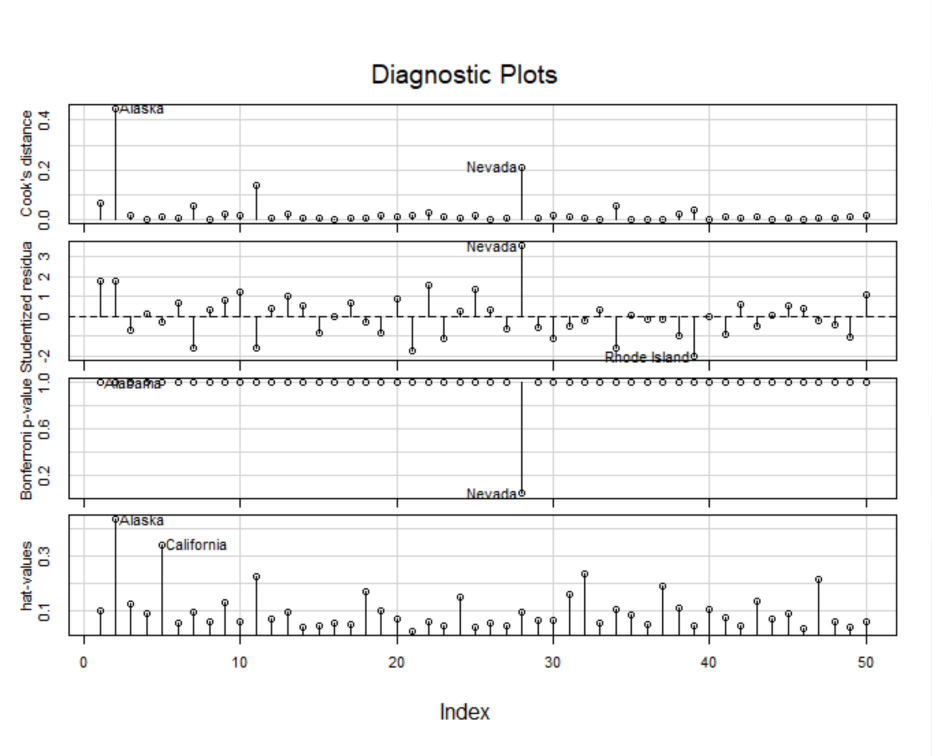
```
> fit_w <- lm(weight ~ height, women)

> influence.measures(fit_w)
Influence measures of
         lm(formula = weight ~ height, data = women) :

    dfb.1_  dfb.hght    dffit cov.r   cook.d    hat inf
1   1.0106 -9.73e-01  1.14329 0.860 5.28e-01 0.2417   *
2   0.2893 -2.77e-01  0.34099 1.348 6.06e-02 0.1952
3   0.1222 -1.16e-01  0.15310 1.362 1.26e-02 0.1560
4   0.0123 -1.15e-02  0.01687 1.339 1.54e-04 0.1238
5  -0.0527  4.82e-02 -0.08447 1.288 3.84e-03 0.0988
6  -0.0789  6.91e-02 -0.16460 1.214 1.43e-02 0.0810
7  -0.0688  5.36e-02 -0.23753 1.119 2.88e-02 0.0702
8  -0.0212 -4.53e-16 -0.31959 1.004 4.94e-02 0.0667
9   0.0350 -4.92e-02 -0.21800 1.140 2.45e-02 0.0702
10  0.1203 -1.41e-01 -0.33506 1.044 5.50e-02 0.0810
11  0.1252 -1.39e-01 -0.24336 1.193 3.07e-02 0.0988
12  0.0854 -9.22e-02 -0.13567 1.311 9.86e-03 0.1238
13 -0.0035  3.72e-03  0.00491 1.390 1.31e-05 0.1560
14 -0.4406  4.64e-01  0.57152 1.179 1.59e-01 0.1952
15 -1.3653  1.43e+00  1.67669 0.514 8.78e-01 0.2417   *
```

- 너무 많은 숫자 출력
- 그래프 작성이 필요함

- 각 통계량마다 각기 통용되는 임계값이 있음
- 어느 하나의 통계량 값이라도 임계값 초과하면 inf에 별표가 붙음

- R 함수의 활용: 패키지 car의 함수 infIndexPlot( )

```
> infIndexPlot(fit_s)
```




Index plot of

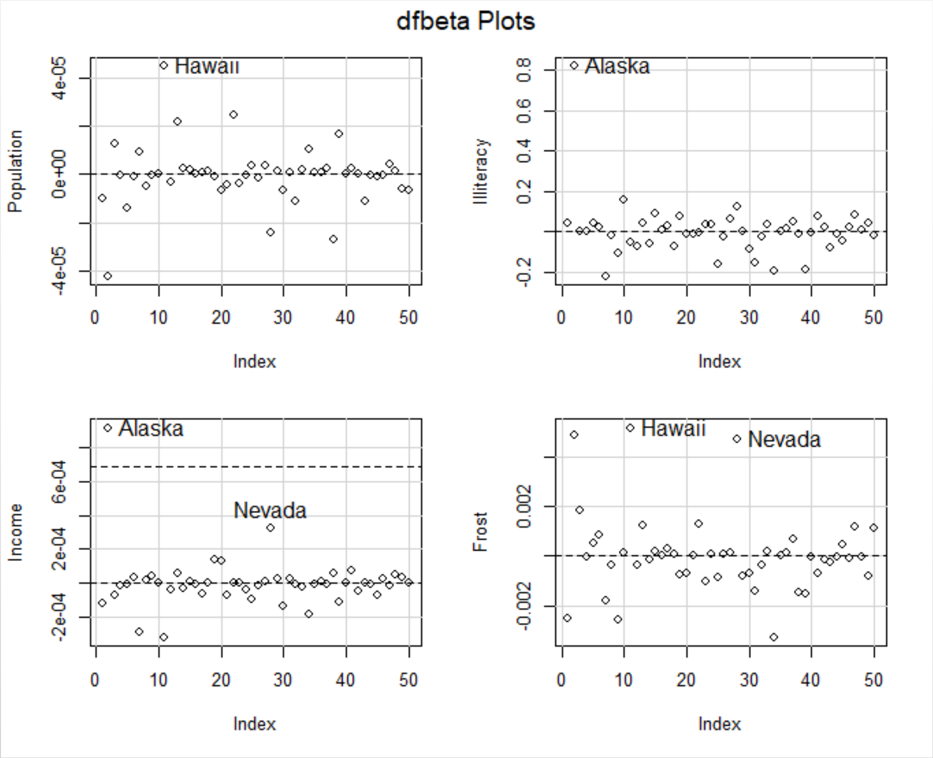
- Cook's distance
- Studentized residuals
- p-value of Bonferroni outlier test
- leverage

- Alaska, Nevada : 특이한 관찰값으로 분류 가능

- R 함수의 활용: 패키지 car의 함수 dfbetaPlots( )

```
> dfbetaPlots(fit_s, id.method="identify")
```

- id.method="identify" → 특이한 관찰값의 라벨을 그래프에 추가 가능
- 마우스를 Plots 창으로 이동 / 마우스 포인터 십자로 변환 / 특정 점 위로 이동하고 마우스 왼쪽 단추 클릭 / Esc 키 누르면 다음 그래프 작성



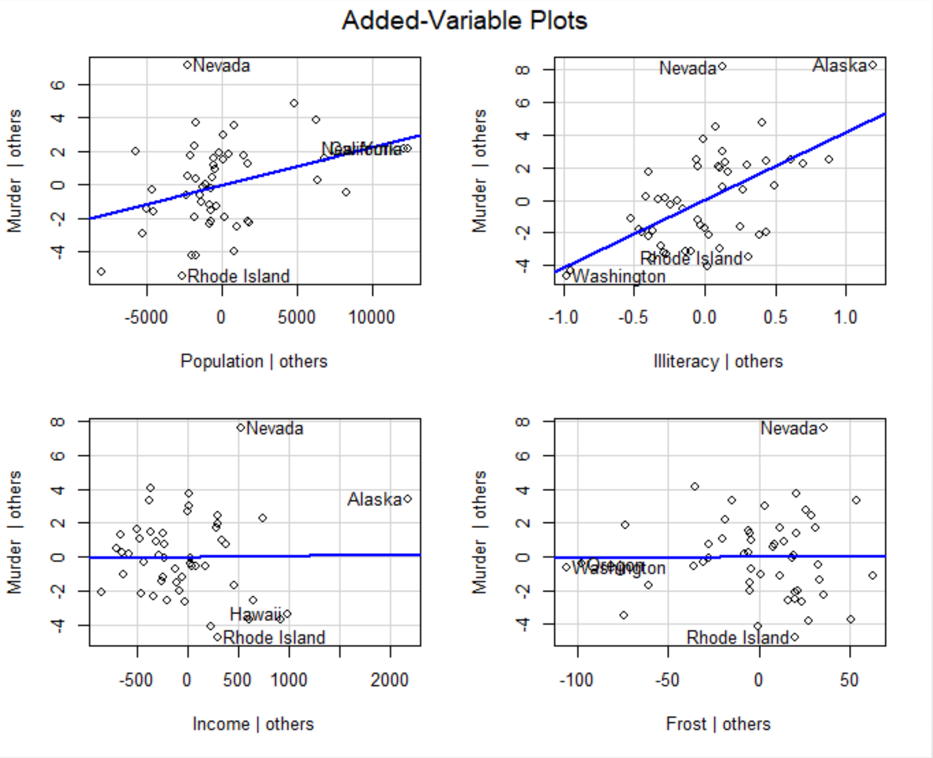

- Alaska, Nevada, Hawaii: 특이한 관찰값으로 분류 가능

- R 함수의 활용: 패키지 car의 함수 avPlots( )

변수 $X_1$의 added variable plot(partial regression plot) 작성 절차:

- 잔차 $e_{Y|2,\ldots,k}$ 계산: 반응변수 $Y$와 설명변수 $X_2, \ldots, X_k$의 회귀모형의 잔차
- 잔차 $e_{1|2,\ldots,k}$ 계산: 반응변수 $X_1$과 설명변수 $X_2, \ldots, X_k$의 회귀모형의 잔차
- 두 잔차의 산점도 작성: 변수 $X_1$의 added variable plot
- 다른 변수의 영향력이 제거되고 순수하게 $Y$와 $X_1$의 관계 확인
- 이상값과 영향력이 큰 관찰값 탐지에 유용한 그래프

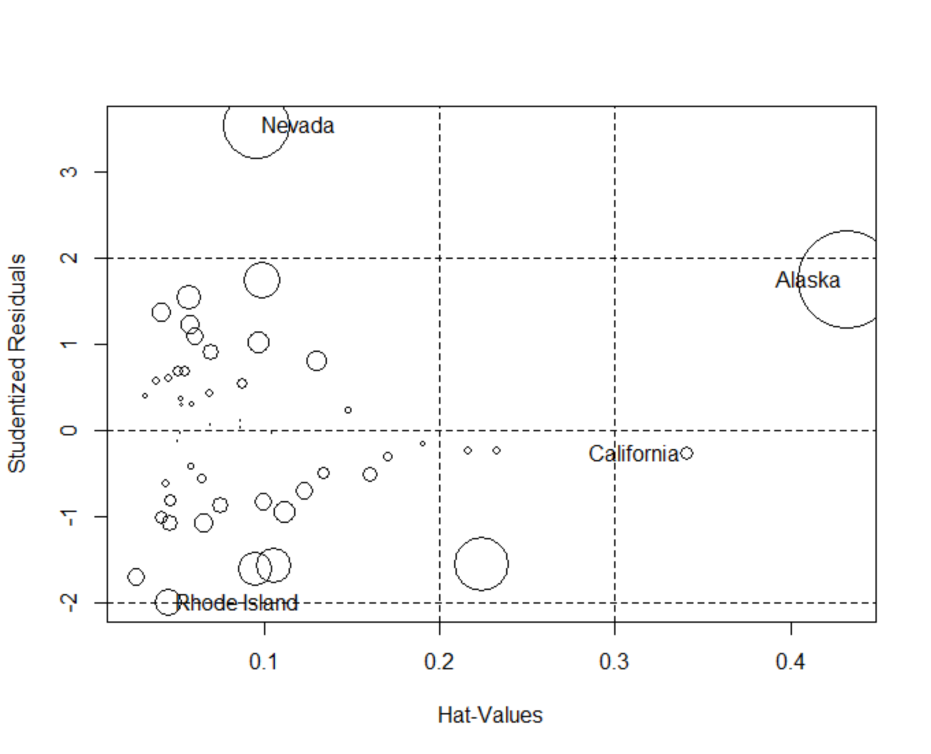
```
> avPlots(fit_s)
```




- 각 그래프마다 X축과 Y축상에서 가장 극단적인 두 점에 대한 라벨 제공
- X축: partial leverage
- Y축: 잔차

- R 함수의 활용: 패키지 car의 함수 influencePlot( )

```
> influencePlot(fit_s)
                 StudRes        Hat       CookD
Alaska         1.7536917 0.43247319 0.448050997
California    -0.2761492 0.34087628 0.008052956
Nevada         3.5429286 0.09508977 0.209915743
Rhode Island  -2.0001631 0.04562377 0.035858963
```




- X축: leverage
- Y축: Studentized residual
- 점의 크기: Cook's distance 에 비례하여 결정
- 수직 점선: leverage 평균값의 2배와 3배

- R 함수의 활용: 패키지 car의 함수 outlierTest( )

  - Studentized residual: 이상값 탐지에 유용함 / $t_i \sim t(n-k-2)$

  - 각 studentized residual이 $t(n-k-2)$에서의 표본인지 여부 검정 가능

  - Nevada 이상값 여부 확인

```
> 2*pt(max(rstudent(fit_s)), df=fit_s$df.resid-1,
       lower.tail=FALSE)
[1] 0.0009508836
```

  - 함수 rstudent( ): studentized residual 계산
  - 함수 pt(x, df=, lower.tail=TRUE)
  - fit_s$df.resid=n-k-1

  - Studentized residual이 큰 몇 개의 관찰값에 대해 검정이 진행되는 것이 일반적

  - 사실상 모든 관찰값이 검정 대상 → 실질적인 다중 검정

  - 다중 검정의 문제: 엄청나게 큰 일종 오류 확률

  - 일종 오류 확률을 조절하는 Bonferroni 수정이 필요한 상황

- 다중 검정의 문제: Type 1 error rate의 증가
  - 3개 귀무가설이 모두 사실이라고 가정
  - 3개 독립된 검정을 각각 유의수준 α에서 실시
  - 3개 검정 모두에서 옳은 결정을 내릴 확률

$$P(3\text{개 검정에서 모두 } H_0 \text{ 기각 못함 }) = (1-\alpha)^3 < (1-\alpha)$$

  - 다중 검정의 일종 오류 확률

$$P(3\text{개 다중 검정에서의 일종 오류})$$
$$= P(\text{적어도 한 번은 } H_0 \text{ 기각})$$
$$= 1 - P(3\text{개 검정에서 모두 } H_0 \text{ 기각 못함}) = 1-(1-\alpha)^3 > \alpha$$

예: $1-(1-0.05)^3 = 0.1426, \;\; 1-(1-0.05)^{15} = 0.5367$

- Bonferroni correction
  - 유의수준을 $\alpha/K$로 조절
  - $K$: 다중 검정에 포함된 검정의 개수
  - Bonferroni p-value: p-value $\times$ $K$

```
> outlierTest(fit_s)
       rstudent unadjusted p-value Bonferonni p
Nevada 3.542929         0.00095088     0.047544
```

- 모든 관찰값이 검정 대상
- 검정 개수: 관찰값의 개수와 동일
- unadjusted p-value $\times$ 50 = Bonferroni p-value

- 예제: Ericksen

  - 실습 내용

    1) crime을 반응변수, 나머지를 설명변수로 하는 회귀모형 적합
        - 임의로 선택한 6개 케이스는 test data / 나머지 60개 케이스로 적합
        - set.seed(1234) 사용

    2) 이상값 혹은 영향력이 큰 관찰값 탐지

    3) 관찰값 제거가 필요하다면 제거 후 다시 모형 적합 / 가정 만족 여부 확인 / 이상값 탐지 작업 시행

- 예제: 패키지 carData의 Bfox
  - 자료: 1946년에서 1975년 사이의 캐나다 여성 직업 참여에 대한 연도별 자료
  - 변수 설명

1) partic: Percent of adult women in the workforce.

2) tfr: (Total fertility rate) expected births to a cohort of 1000 women at current age-specific fertility rates.

3) menwage: Men's average weekly wages, in constant 1935 dollars and adjusted for current tax rates.

4) womwage: Women's average weekly wages.

5) debt: Per-capita consumer debt, in constant dollars.

6) parttime: Percent of the active workforce working 34 hours per week or less.

- 자료 분석 내용

1) 1945년~1972년 자료를 사용하여 모형 적합
2) 적합된 모형의 가정 만족 여부 확인
3) 이상값 혹은 영향력이 큰 관찰값 탐지
4) 제거가 필요한 관측값이 있는 경우, 제거 후 적합 및 가정 확인
5) 적합된 모형을 이용하여 1973~1975년 변수 partic의 평균값 예측하고 실제 값과 비교