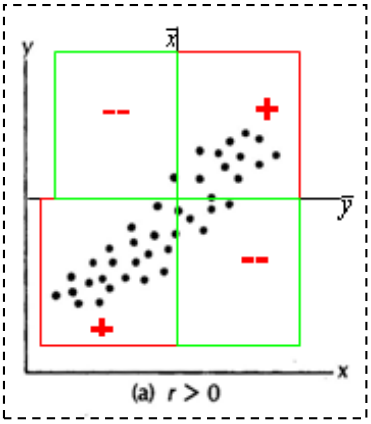
## 상관계수

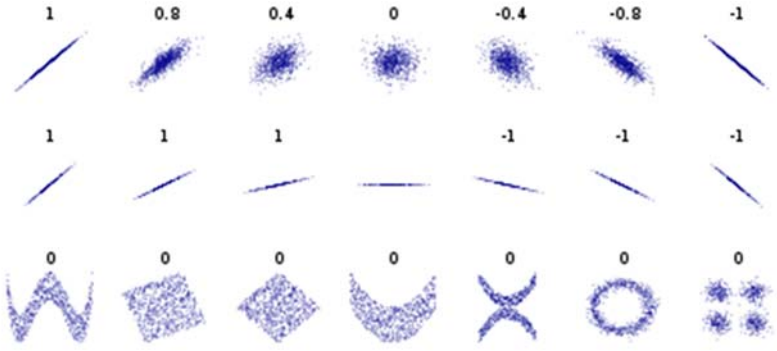
### 상관계수 정의

- 두 변수 간의 선형 관계 정도를 나타내는 값
- 정의 : $\rho = \frac{COV(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{E(X-E(X))(Y-E(Y))}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{E(XY)-E(X)E(Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$
- 표본 상관계수 : $r = \hat{\rho} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})/(n-1)}{\sqrt{\sum(x_i-\bar{x})^2/(n-1)}\sqrt{\sum(y_i-\bar{y})^2/(n-1)}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$




### 해석

- 측정형 metric 변수 간의 선형 관계 척도
- 순서형은 변수 간 선형 관계 정도는 Spearman 상관계수, Kendall $\tau$
- 상관계수는 최대값은 $\pm 1$ (양/음의 완전한 직선 관계), 최소값은 0 이다.
- 타원의 길이가 길고 폭이 좁을수록 상관계수는 $\pm 1$에 가깝다.
- $\pm 0.8$ 이 강한 상관관계, $\pm 0.5$ 약한 상관관계





Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

### 유의성 검정

귀무가설 : 두 변수의 선형함수관계는 유의하지 않다. $H_0 : \rho = 0$

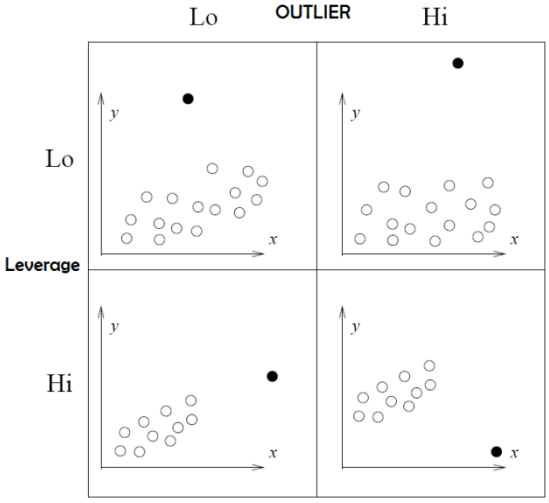
대립가설 : 두 변수의 선형함수관계는 유의하지 않다. $H_0 : \rho = 0$

검정통계량 : $T = \sqrt{n-2}\,\dfrac{r}{\sqrt{1-r^2}} \sim t(n-2)$

## 산점도 scatter plot

두 측정형 변수 중 하나를 X-축에 다른 하나를 Y-축으로 하여 2 차원 공간에 관측치를 표현 (인관관계의 회귀분석에서는 설명변수를 X-축, 종속변수를 Y-축)

- 두 변수간의 함수 관계를 시각적 진단
- 선형관계에서 이상치 outlier, 영향치 influential 진단




Influential


Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

## 프로그램 활용

블루크로스 보험회사의 잉여금과 보험청구 상관관계가 존재

SAS | Page 469, 상관분석

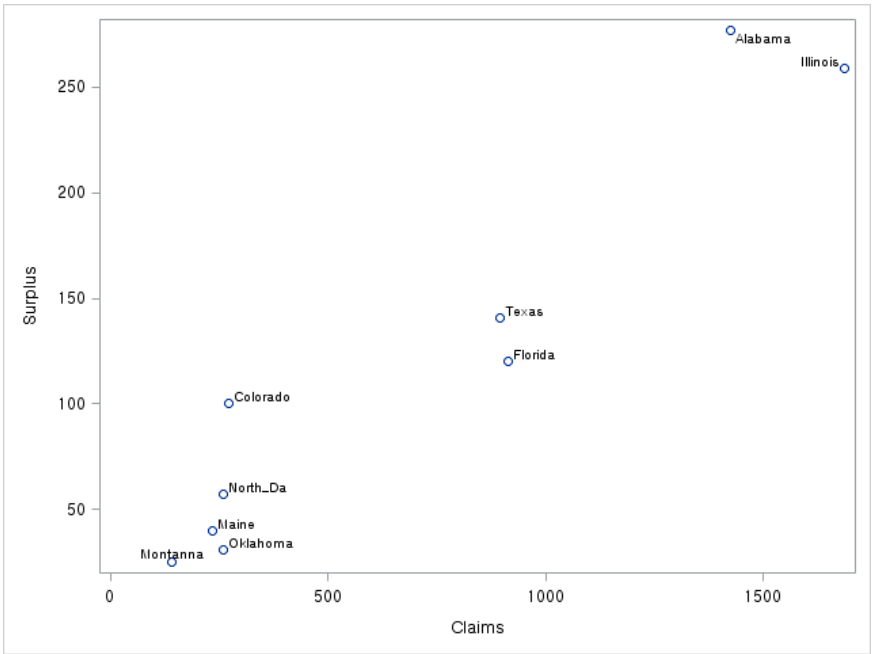
```
data p124;
input State $   Claims   Surplus;
cards;
Alabama         1425     277
Colorado        273      100
Florida         915      120
Illinois 1687    259
Maine 234        40
Montanna        142      25
North_Dakoda    259      57
Oklahoma        258      31
Texas 894        141
run;

proc corr data=p124;
        var claims surplus;
    run;
```

Pearson Correlation Coefficients, N = 9
Prob > |r| under H0: Rho=0

| | Claims | Surplus |
|---|---|---|
| Claims | 1.00000 | 0.95684<br><.0001 |
| Surplus | 0.95684<br><.0001 | 1.00000 |

```
proc sgplot data=p124;
        scatter x=claims y=surplus /
datalabel=state;
        run;
```

- 유의확률이 0.0001 미만이므로 상관관계는 매우 유의하다. (이상치 없음)
- 보험청구액으로 높을수록 잉여금은 많아진다. (알라마바, 일리노이는 영향치) 결정계수(상관계수의 제곱)를 높인다. 하여, 반드시 산점도 그리기 필수




단순회귀분석 개요

**인과관계와 상관관계**

적용하는 방법에 상관없이, 두 변수간의 강한 수학적(또는 그래프상의) 관계가 어떤 하나가 다른 것의 원인이 된다는 것을 의미하지는 않음

*변수들이 상호 강하게 연관되어 있다고 해서 상호 인과관계가 있다고는 할 수 없음*

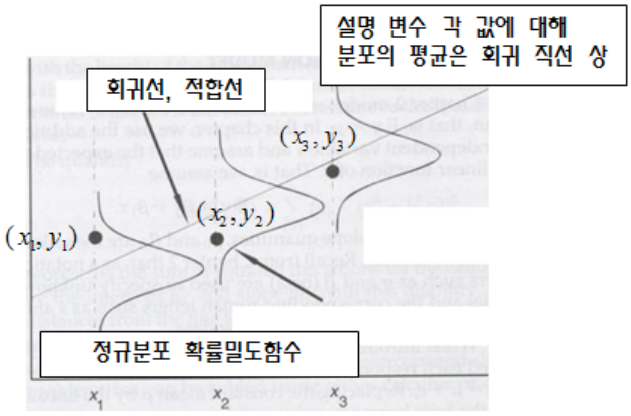
근본원인이 검증되려면 다음 두 가지 요건이 모두 충족되어야 함 :

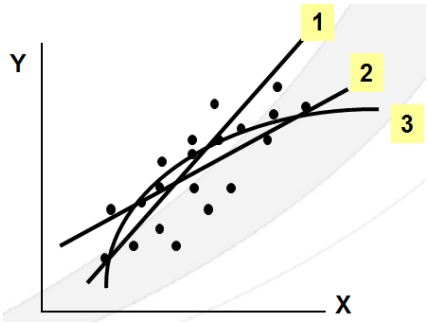
- 잠재적 근본 원인과 결과간의 통계적인 유의성이 있는 관계
- 프로세스 지식의 검증을 통하여 인과 관계가 확정된 관계
- 위 두 개 조건 중 하나만으로는 충분한 인과관계 검증이 이루어지지 않음

## 단순회귀모형

$y_i = \beta_0 + \beta_1 x_i + e_i$ (단순(직선)회귀모형)

- 회귀계수, 모수 parameter ($\beta_0, \beta_1$), unknown but constant
- 종속변수 dependent, response, target $y_i$
- 독립변수 independent, exploratory, predictor $x_i$
- 오차 error (가정) $e_i \sim iidN(0, \sigma^2)$ 독립성, 등분산성, 정규성

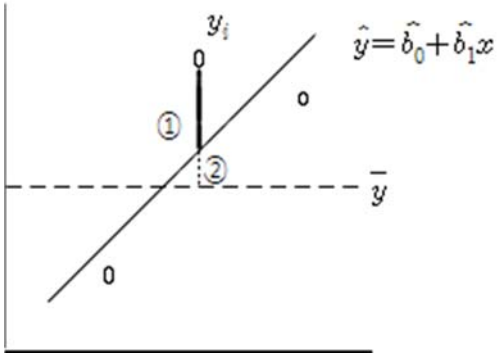




Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

## 회귀계수 추정




$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{S_{(xy)}}{S_{(xx)}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i y_i) - n(\bar{x})(\bar{y})}{\sum_{i=1}^{n}(x_i)^2 - n(\bar{x})}$$

OLS 추정치 $\min_{\beta_0,\beta_1}\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$ , $\varepsilon_i = (1)$

MLE 추정치 오차의 가정 $e_i \sim iidN(0,\sigma^2)$ => $y_i \sim iidN(a+bx_i,\sigma^2)$

$$L(y_1,y_2,...,y_n;a,b) = \prod_i f(y_i;a,b) = \prod_i \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(y_i-a-bx_i)^2}{2\sigma^2}\}$$

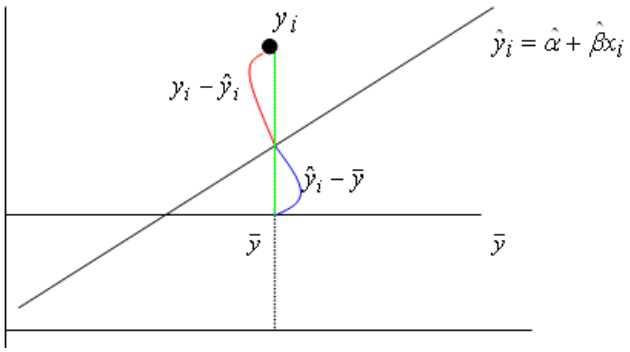
**Gauss-Markov 정리 : OLS is BLUE**

## 모형 적합성 (분산분석 ⇔ F-검정)

**통계적 가설**

귀무가설 : $(\mathrm{mod}\,el)\,y = a+bx$ 적합하지 않음

대립가설 : $(\mathrm{mod}\,el)\,y = a+bx$ 은 적합하다.




$$SST = \sum(y_i-\bar{y})^2$$
$$= \sum(y_i-\hat{y}_i)^2 + \sum(\hat{y}_i-\bar{y})^2$$
$$= SSE\,(\text{unexplaine d})$$
$$+ SSR\,(\text{explianed by model})$$
$$T = \frac{SSR/df_r}{SSE/df_e} = \frac{MSR}{MSE}$$

- 총변동 $SSTO = \sum(y_i-\bar{y})^2$ (초록색 부분)


Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

- 회귀(모형)변동: $SSR = \sum(\hat{y}_i - \bar{y}_i)^2$ (파란부분)
- 오차 변동: $SSE = \sum(y_i - \hat{y}_i)^2$ (빨간 부분)

**분산분석 표**

| 변동 | 자유도 | 자승합 SS | 평균자승합 MS | F-통계량 |
|---|---|---|---|---|
| 모형 model | p | SSR | MSR=SSR/p | F=MSR/MSE |
| 오차 error | n-p-1 | SSE | MSE=SSE/(n-p-1) | (유의확률) |
| 총변동 | n-1 | SST | | |

**오차분산 $\sigma^2$ 추정치**

(수리적 증명 생략) $\hat{\sigma^2} = MSE = \frac{SSE}{(n-p-1)}$, p=설명변수 개수

## 회귀계수 (기울기) 유의성

**통계적 가설**

귀무가설 : $\beta_1 = 0$ ⇔ 설명변수 X와 종속변수 Y의 선형 함수관계는 유의하지 않다 ⇔ 설명변수 X의 설명력을 유의하지 않다.

*) 절편 $\beta_0$의 유의성 검정을 하지 않음, *) 절편 $\beta_0 = 0$이면 원점을 지나는 직선임

**검정통계량**

$$T = \frac{\hat{\beta}_1 - \beta_1(=0)}{S(\hat{\beta}_1)} \sim t(n-2)$$

$$S(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2} = \frac{MSE}{SSx}$$

**모형 적합성 F-검정과 관계**

설명변수가 하나이므로 모형의 자유도=1 => $F(1,n) = (t(n))^2$

그러므로 단순회귀모형에서는 분산분석과 기울기 유의성 검정은 동일하다

## 결정계수 coefficient of determination

**정의 :** $R^2 = \frac{SSR}{SST}$

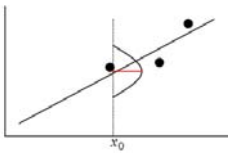
**특성**

- 회귀모형이 종속변수의 총변동을 설명하는 비율 ⇔ 모형 적합성 대표 값
- 회귀모형에 고려된 설명변수의 종속변수 변동을 선형적으로 설명하는 정도
- 단순회귀에서는 상관계수 제곱 $r^2$ 은 결정계수이다.
- 결정계수는 검정통계량이 아니므로 유의성 검정 불가, 일반적으로 70% 이상이면 종속변수 설명이 충분한 설명변수(들)를 선정하였음.
- 회귀계수 추정치와 상관계수 관계 $r^2 = b^2 \frac{SXX}{SYY}$
- 수정 adjusted 결정계수 : $R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$

## 추정 (예측치)

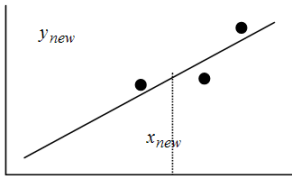
**예측치** $y = \hat{a} + \hat{b}x$

$E(y \mid x) = a + bx$ **신뢰구간 confidence interval**

$\hat{E}(y_0) = \hat{\mu}_{y|x_0} = \hat{\alpha} + \hat{\beta}x_0$  $s^2(\hat{E}(Y_0)) = MSE[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}]$,

$\frac{\hat{E}(Y_0) - E(Y_0)}{s\{\hat{E}(Y_0)\}} \sim t(n-2)$

$y = a + bx_{new} + e$ **예측구간 prediction interval**

$\hat{y}_{new} = \hat{\alpha} + \hat{\beta}x_{new}$  $E(\hat{y}_{new}) = \alpha + \beta x_{new}$,

$\sigma^2\{\hat{Y}_{new}\} = \sigma^2[1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum(x_i - \bar{x})}]$

$\frac{\hat{Y}_{new} - E(\hat{Y}_{new})}{s\{\hat{Y}_{new}\}} \sim t(n-2)$, $s\{\hat{Y}_{new}\} = MSE[1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{\sum(X_i - \bar{X})}]$



Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

## 잔차분석

### Residual 잔차

- 오차항의 추정치 : $r_i = \hat{e}_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$
- 표준화 잔차 : $z_i = \frac{r_i}{\sqrt{MSE}}$
- 스튜던트 잔차 : $r_i = \frac{r_i}{\sqrt{MSE/1-h_{ii}}}$

### 활용 (진단)

(1) 모형은 선형인가? => 잔차 그래프 (pattern 가짐)

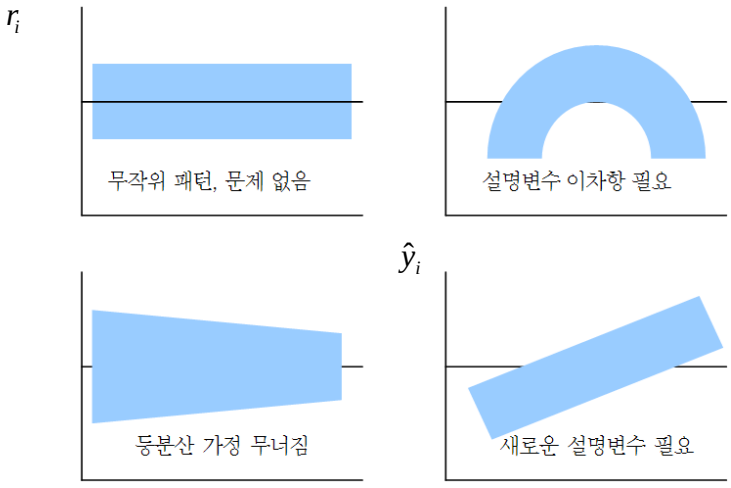
(2) 오차의 가정

- 정규성 : 문제는 되지 않음, n>20 이상 (like CLT) => 잔차의 정규성 검증
- 독립성 : 시계열 데이터만 검증 => DW 통계량
- 등분산성 : => 잔차 그래프 Fan 모양

(3) 이상치 outliers, 영향치 influential observation

### 도구

(1) 잔차 (일반적으로 스튜던트 잔차)와 종속변수 예측치 산점도 (설명변수가 두 개 이상인 경우에는 각 설명변수를 X-축으로 한 잔차 산점도 필요




SAS 보험회사 예제 계속

```
proc reg data=p124;
      model surplus=claims;
   output out=out1 p=yhat rstudent=res L95m=L95m U95m=U95m L95=L95
```

```
U95=U95;
run; QUIT;
```

**Analysis of Variance**

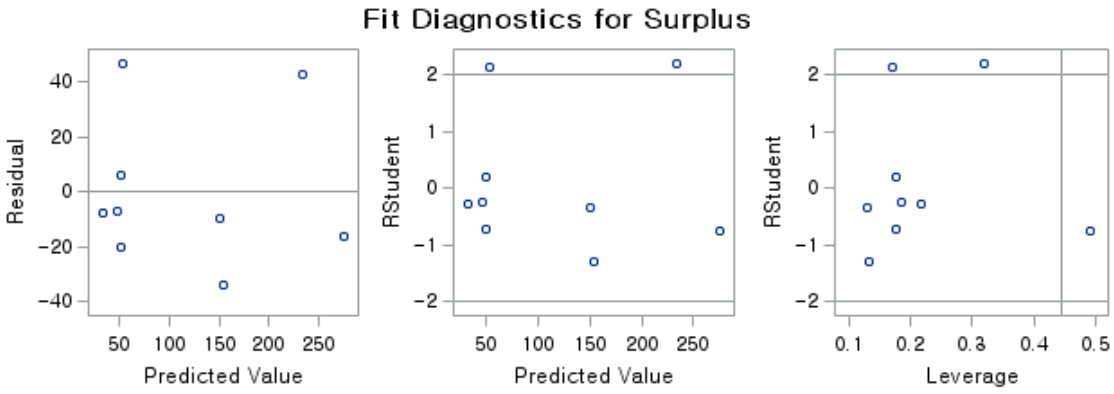
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 65942 | 65942 | 75.87 | <.0001 |
| Error | 7 | 6083.66306 | 869.09472 | | |
| Corrected Total | 8 | 72026 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 29.48041 | R-Square | 0.9155 |
| Dependent Mean | 116.66667 | Adj R-Sq | 0.9035 |
| Coeff Var | 25.26893 | | |

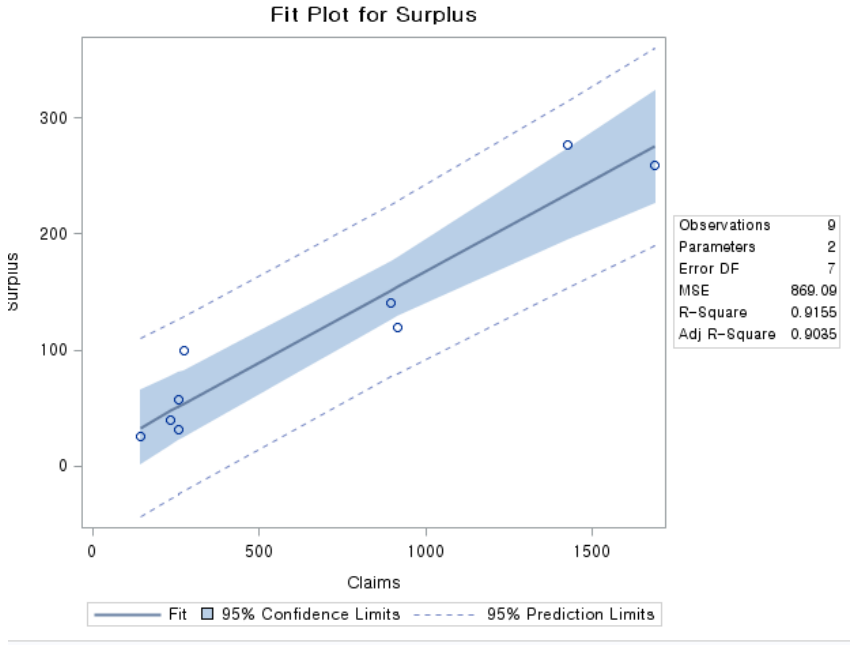
**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 10.56270 | 15.65065 | 0.67 | 0.5214 |
| Claims | 1 | 0.15688 | 0.01801 | 8.71 | <.0001 |

추정회귀모형 : $\underset{(t=8.71,\ p<0.0001)}{\text{잉여금} = 10.56 + 0.157 * \text{청구액}}$, $R^2 = 91.6\%$




| State | Claims | Surplus | Predicted Value of Surplus | Lower Bound of 95% C.I. for Mean | Upper Bound of 95% C.I. for Mean | Lower Bound of 95% C.I.(Individual Pred) | Upper Bound of 95% C.I.(Individual Pred) | Studentized Residual without Current Obs |
|---|---|---|---|---|---|---|---|---|
| Alabama | 1425 | 277 | 234,1183766 | 194,6654426 | 273,5713105 | 154,0182404 | 314,2185128 | 2,1920139228 |
| Colorado | 273 | 100 | 53,39125749 | 24,49497522 | 82,28753976 | -22,0706067 | 128,8531217 | 2,1325973652 |
| Florida | 915 | 120 | 154,1089749 | 128,7464718 | 179,471478 | 79,92842228 | 228,2895275 | -1,302454379 |
| Illinois | 1687 | 259 | 275,2212457 | 226,3074757 | 324,1350157 | 190,062249 | 360,3802424 | -0,747532917 |
| Maine | 234 | 40 | 47,27289148 | 17,35946268 | 77,18632029 | -28,5842838 | 123,1300668 | -0,254224891 |
| Montanna | 142 | 25 | 32,83982294 | 0,316270689 | 65,3633752 | -44,0840308 | 109,7636767 | -0,280173874 |
| North_Da | 259 | 57 | 51,19492098 | 21,94028902 | 80,44955294 | -24,4048893 | 126,7947313 | 0,2015268585 |
| Oklahoma | 258 | 31 | 51,0380398 | 21,75751468 | 80,31856492 | -24,5717941 | 126,6478737 | -0,72299659 |
| Texas | 894 | 141 | 150,8144701 | 125,7969614 | 175,8319789 | 76,75116228 | 224,877778 | -0,333259294 |




## 다중회귀 모형

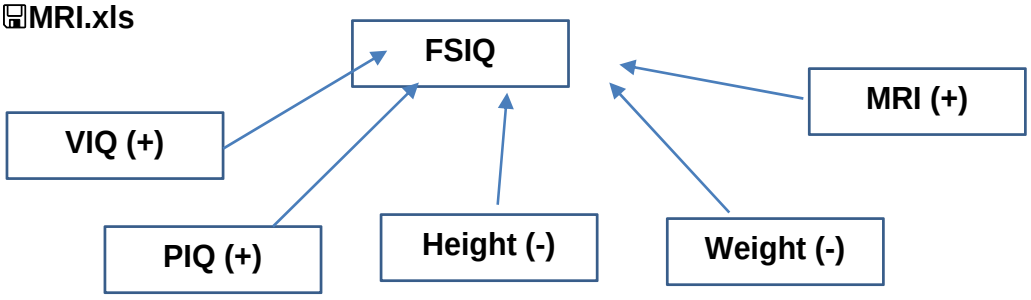
### 회귀모형

$$Y_i = a + b_1 x_{1i} + b_2 x_{2i} + ... + b_p x_{pi} + e_i,$$

### 가정

$e_i \sim iidN(0, \sigma^2)$ 독립성, 등분산성, 정규성

### 예제 MRI.xls



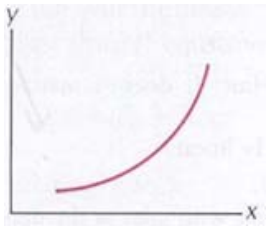
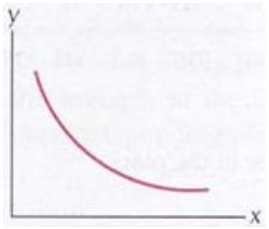
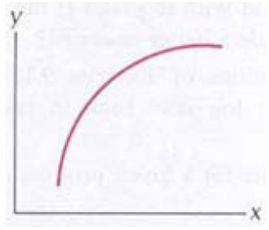
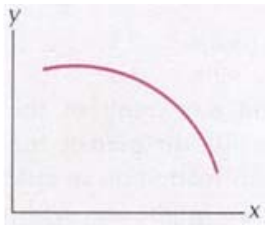

총 IQ 에 영향을 미치는 요인으로 언어 verbal IQ, 수행 performance IQ, 몸무게, 키 (작을수록 똑똑하다는 속설), 뇌의 크기=두뇌정보를 고려, 이론적/경험적 부호 표시

### (순서 1) 산점도 그리기

- 종속변수와 설명변수 간의 함수관계 보기
- 설명변수 간 상관관계가 높은 경우 => 다중공선성

```
proc sgscatter data=mri;
  title "Scatterplot Matrix for Iris Data";
  matrix fsiq viq piq mri weight height
      / group=gender;
run;

proc corr data=mri nosimple;
   var fsiq viq piq mri weight height ;
run;
```

■ 변수변환 방법 (두 변수의 함수 관계의 선형화)

| 산점도 유형 | 변수변환 | 산점도 유형 | 변수변환 |
|---|---|---|---|
|  | $x$ : $x^2$, $x^3$ ; $y$ : $\log y$, $-\frac{1}{y}$ |  | $x$ : $\log x$, $-\frac{1}{x}$ ; $y$ : $\log y$, $-\frac{1}{y}$ |
|  | $x$ : $\log x$, $-\frac{1}{x}$ ; $y$ : $y^2$, $y^3$ |  | $x$ : $x^2$, $x^3$ ; $y$ : $y^2$, $y^3$ |

○ 변수변환 방법 (변수의 정규분포 변환) Power Data Transformation $y^* = y^\alpha$

| $\alpha$ | | 해결내용 |
|---|---|---|
| 3 | 세제곱 | Severe 좌로 치우침 |
| 2 | 제곱 | mild 좌로 치우침 |
| 1/2 | 제곱근 $\sqrt{}$ | mild 우로치우침 |
| log | 로그 | 우로 치우침 |
| -1 | 역변환 | severe 우로 치우침 |

**(순서 2) 회귀모형 추정**

```
proc reg data=mri;
    model fsiq = viq piq mri weight height;
run;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 5 | 20895 | 4179.02973 | 1481.83 | <.0001 |
| Error | 32 | 90.24611 | 2.82019 | | |
| Corrected Total | 37 | 20985 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.67934 | R-Square | 0.9957 |
| Dependent Mean | 113.55263 | Adj R-Sq | 0.9950 |
| Coeff Var | 1.47891 | | |

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | -9.31128 | 5.61322 | -1.66 | 0.1069 |
| VIQ | VIQ | 1 | 0.57655 | 0.01933 | 29.82 | <.0001 |
| PIQ | PIQ | 1 | 0.54964 | 0.02052 | 26.78 | <.0001 |
| MRI | MRI | 1 | -0.00001361 | 0.00000567 | -2.40 | 0.0224 |
| Weight | Weight | 1 | -0.00729 | 0.01683 | -0.43 | 0.6680 |
| Height | Height | 1 | 0.15285 | 0.11169 | 1.37 | 0.1807 |

- 전체 모형 (F-검정 결과)의 유의함.
- 그러나 회귀계수 검정결과 Weight, height 유의하지 않음

### (순서 3) 변수선택

종속변수에 유의한 영향을 미치는 설명변수를 찾는 방법이다. 다중공선성을 먼저 시행한 후 변수선택을 하는 것이 일반적이나 결과는 대부분 동일하다.

(F-검정통계량 이용 방법)

- 후진선택 Backward: 유의하지 않은 순서대로(F-통계량 값 가장 적음) 제거하는 방법
- 전진선택 Forward: 유의한 순서대로(F-통계량 값 가장 큼) 선택하는 방법
- 단계선택 Stepwise: 전진선택과 유사하나 선택된 설명변수의 유의성도 추가 선택된 설명변수에 의해 검증
- 수작업: 유의하지 않은 설명변수 제거 / 선택 순서를 분석자가 결정 (권장)

(모형 적합성 관련 통계량 이용)

- 결정계수 $R^2 = \frac{SSR}{SST}$ (보고서 제시) 적합 모형이 종속변수 변동의 설명부분
- 수정된 결정계수 $R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$, 유의하지 않은 설명변수가 삽입되어도 결정계수가 커지는 문제 보완, 사로 다른 모형 비교 시 사용
- AIC(Akaike Information Criterion) $AIC = n\ln(SSE/n) + 2(p-1)$
- SBC(Schwarz Bayes Criterion) $SBC = n\ln(SSE/n) + (p-1)\ln np)$
- PRESS $PRESS = \sum(y_i - \hat{Y}_{(i)})^2$
- AIC, SBC, PRESS 값이 작은 모형

```
proc reg data=mri;
    model fsiq = viq piq mri weight height/selection=backward;
run;
```

Summary of Backward Elimination

| Step | Variable Removed | Label | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
|---|---|---|---|---|---|---|---|---|
| 1 | Weight | Weight | 4 | 0.0000 | 0.9957 | 4.1874 | 0.19 | 0.6680 |
| 2 | Height | Height | 3 | 0.0002 | 0.9954 | 3.9889 | 1.85 | 0.1834 |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | -3.26299 | 3.48702 | 2.46865 | 0.88 | 0.3560 |
| VIQ | 0.57462 | 0.01908 | 2556.96484 | 906.96 | <.0001 |
| PIQ | 0.54290 | 0.01995 | 2087.05556 | 740.28 | <.0001 |
| MRI | -0.00000889 | 0.00000411 | 13.21044 | 4.69 | 0.0375 |

- 몸무게, 키는 유의하지 않아 제외됨
- MRI 회귀계수 부호가 이상하다. ⇔ 상관계수 부호와 일치해야 하는데???
  이는 무슨 일? 바로 다중공선성 문제 발생

```
proc reg data=mri;
   model fsiq= viq piq mri weight height/selection=adjrsq;
run;
```

| Number in Model | Adjusted R-Square | R-Square | Variables in Model |
|---|---|---|---|
| 4 | 0.9952 | 0.9957 | VIQ PIQ MRI Height |
| 3 | 0.9950 | 0.9954 | VIQ PIQ MRI |
| 5 | 0.9950 | 0.9957 | VIQ PIQ MRI Weight Height |
| 4 | 0.9949 | 0.9954 | VIQ PIQ MRI Weight |
| 2 | 0.9945 | 0.9948 | VIQ PIQ |
| 3 | 0.9945 | 0.9949 | VIQ PIQ Weight |

- 수정결정계수 크기로 가장 높은 모형 순으로, 물론 포함된 설명변수의 유의성은 검증된 것이 아님, 그래서 height 가 포함되어 있음

### (순서 4) 다중공선성 Multicollinearity

문제

- 설명변수들간의 높은 상관관계로 인하여 $|X'X| \approx 0$이 되고 $(X'X)^{-1}$의 값이 불안정
- 추정회귀계수 $\hat{b}=(X'X)^{-1}X'y$의 분산이 불안정해져 추정 회귀계수의 부호까지 바뀌는 문제 발생

진단방법

- 산점도 행렬과 상관계수 이용, 두 설명변수의 관계에 의한 문제 발생 진단
- VIF 이용: $VIF_k = \frac{1}{1-R_k}$, $R_k$는 설명변수 $X_k$를 종속변수로 하고 나머지 다른 변수들을 설명변수로 하여 계산된 결정계수, 3 이상이면 문제, 두 변수간 (pairwise) 문제를 발견하지 못하는 문제가 있다. 여러 설명변수가 동시에 고려되므로... 이에 대한 보완으로 상태지수가 있음.
- Condition Index: $CI_k = \sqrt{\lambda_k / \lambda_{\max}}$, 주성분분석 개념, 설명변수의 상관관계가 높으면 제일 주성분의 고유치(원변수 변동에 대한 설명 기여율)가 커진다. 10 이상이면 문제. 문제의 발견은 변동 기여율에 의해 하게 된다. 문제가 되는 행의 변동 기여율이 큰 값을 찾아 공통 설명변수를 찾으면 된다.

해결방법

- (변수 제외) 문제가 되는 설명변수 제외, 종속변수와 상관관계가 낮은 설명변수 제외
- (주성분분석 이용) 원변수의 선형결합으로 만들어진 주성분 변수(서로 독립), 그러나 주성분 변수의 의미가 불분명하여 자주 사용하지 않음

```
proc reg data=mri;
    model fsiq = viq piq mri/collin vif;
run;
```

- VIF 크기는 이상 없음. VIQ, PIQ 의 변동 기여율이 3 번째 행에서 동시에 크므로 두 변수의 상관관계로 인하여 다중공선성 문제가 발생하고 이로 인하여 MRI 의 부호가 바뀌는 문제가 발생하였음
- (PIQ, VIQ) 중 하나를 제외하자. FSIQ 와 상관계수 낮은 PIQ 를 제외하는 것이 적절함.

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | Intercept | 1 | -3.26299 | 3.48702 | -0.94 | 0.3560 | 0 |
| VIQ | VIQ | 1 | 0.57462 | 0.01908 | 30.12 | <.0001 | 2.51425 |
| PIQ | PIQ | 1 | 0.54290 | 0.01995 | 27.21 | <.0001 | 2.66837 |
| MRI | MRI | 1 | -0.00000889 | 0.00000411 | -2.16 | 0.0375 | 1.16665 |

| Collinearity Diagnostics | | | | | | |
|---|---|---|---|---|---|---|
| Number | Eigenvalue | Condition Index | Proportion of Variation | | | |
| | | | Intercept | VIQ | PIQ | MRI |
| 1 | 3.95707 | 1.00000 | 0.00038561 | 0.00098204 | 0.00091212 | 0.00033642 |
| 2 | 0.03122 | 11.25746 | 0.05650 | 0.13381 | 0.10702 | 0.03595 |
| 3 | 0.00875 | 21.26036 | 0.00520 | 0.85393 | 0.83883 | 0.00264 |
| 4 | 0.00295 | 36.62942 | 0.93792 | 0.01127 | 0.05324 | 0.96107 |

**(순서 5) 회귀진단 및 활용**

(스튜던트) 잔차와 예측치 산점도

- 잔차와 설명변수 간 산점도
- 잔차에 대한 정규성 검정 (n 이 충분히 크다면 CLT 에 의해 큰 문제가 되지 않음)

영향치 및 이상치 진단통계량

○ 이상치(outlier)

- 표준화 잔차 (standardized) $z_i = \frac{r_i}{s(r_i) = \sqrt{MSE}}$
- 스튜던트 잔차 (studentized) $st_i = \frac{r_i}{\sqrt{MSE(1-h_{ii})}}$

(hat 행렬 $X(X'X)^{-1}X'$의 대각원소가 $h_{ii}$)

- 스튜던트 제외 잔차 $rt_{(i)} = \frac{r_{(i)}}{\sqrt{MSE_{(i)}(1-h_{ii})}}$

○ 영향치(Influential obs.) 통계량

- Hat 행렬 $H = X(X'X)^{-1}X'$ 관측치가 예측치에 미치는 영향 정도, 대각원소인 $h_{ii}$를 Leverage 라 한다. 관측치가 관측점의 중심으로부터 떨어진 정도를 의미한다. 기준: $2(p+1)/n$이상
- COV ratio $COV = \dfrac{|MSE_{(i)}(X_{(i)}'X_{(i)})^{-1}|}{|MSE(XX)^{-1}|}$, 기준값=$1 \pm 3(p+1)/n$

○ 이상치 & 영향치 통계량

- Cook's distance: $C_i = \dfrac{\sum(\hat{Y}_{F,j} - \hat{Y}_{(j),j})}{MSE(p+1)}$, 기준값=1
- DFBETAS (Difference of Betas), 기준값=$2/\sqrt{n}$

```
proc reg data=mri;
   model fsiq = viq mri/influence;
run;
```

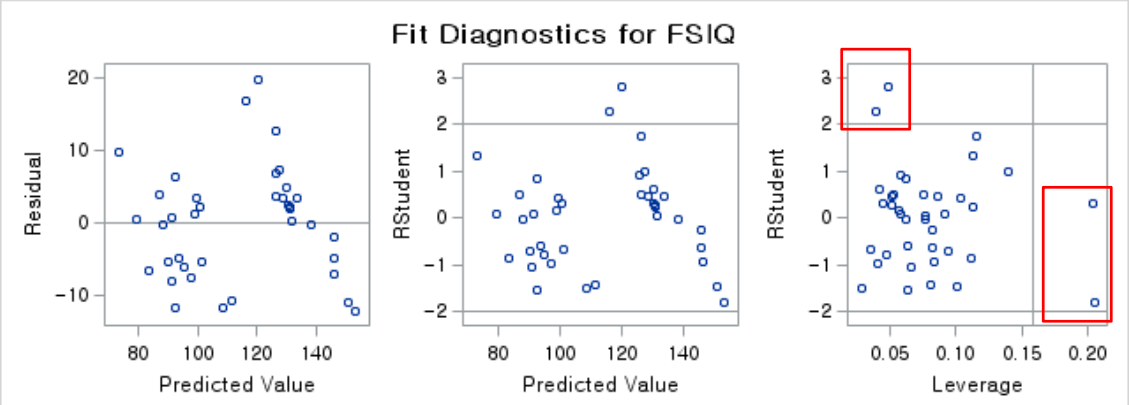
| | | | |
|---|---|---|---|
| Root MSE | 7.89740 | R-Square | 0.8960 |
| Dependent Mean | 113.55263 | Adj R-Sq | 0.8900 |
| Coeff Var | 6.95483 | | |

**Parameter Estimates**

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -10.88005 | 16.34802 | -0.67 | 0.5101 |
| VIQ | VIQ | 1 | 0.96409 | 0.05934 | 16.25 | <.0001 |
| MRI | MRI | 1 | 0.00001801 | 0.00001876 | 0.96 | 0.3437 |

**Output Statistics**

| Obs | Residual | RStudent | Hat Diag H | Cov Ratio | DFFITS | DFBETAS | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Intercept | VIQ | MRI |
| 1 | 1.9097 | 0.2533 | 0.1132 | 1.2232 | 0.0905 | 0.0523 | 0.0574 | -0.0695 |
| 2 | 12.5981 | 1.7450 | 0.1155 | 0.9535 | 0.6305 | -0.5236 | -0.0228 | 0.5348 |
| 3 | 2.1295 | 0.2731 | 0.0512 | 1.1421 | 0.0634 | -0.0356 | 0.0238 | 0.0284 |
| 4 | 3.4858 | 0.4477 | 0.0504 | 1.1286 | 0.1031 | -0.0455 | 0.0539 | 0.0283 |
| 5 | 6.3873 | 0.8314 | 0.0619 | 1.0947 | 0.2136 | -0.0169 | -0.1562 | 0.0878 |




```
proc reg data=mri;
   model fsiq = viq mri/influence stb;
   reweight obs.=9;
   reweight obs.=13;
run;
```

| | | | |
|---|---|---|---|
| Root MSE | 6.67677 | R-Square | 0.9259 |
| Dependent Mean | 112.27778 | Adj R-Sq | 0.9214 |
| Coeff Var | 5.94665 | | |

**Parameter Estimates**

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -12.47789 | 13.98213 | -0.89 | 0.3786 | 0 |
| VIQ | VIQ | 1 | 0.95187 | 0.05041 | 18.88 | <.0001 | 0.94109 |
| MRI | MRI | 1 | 0.00002016 | 0.00001611 | 1.25 | 0.2197 | 0.06235 |

- 이상치 제거로 결정계수가 92.1% 증가
- $FSIQ = -12.48 + 0.952 * VIQ + 0.00002 * MRI$
- (잉, MRI 유의 않음) 그래도 일단 해석을 한다면…
- MRI, VIQ 회귀계수 부호가 양이므로 언어 IQ, MRI 가 커지면 FSIQ 값이 커진다.
- 표준화 회귀계수 크기 이용 : VIQ 의 영향력이 MRI 에 비해 10 배 크기로 FSIQ 에 영향을 준다.