데이터 분석(Python / R)

통계분석 기초

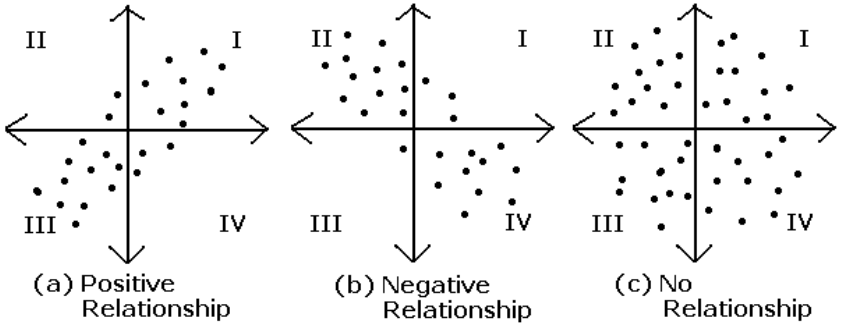
삼성전자공과대학교 3학년 3학기

		Features			
		Two group	Multiple Group	One Variable	Multiple Variable
Target Variable	Number	T test	ANOVA	Linear regression	Multiple linear regression
	Category	Chi square Test		Logistic regression	

Correlation(상관계수) vs. Covariance(공분산)

$$\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

- 공분산은 X의 편차와 Y의 편차를 곱한 것의 평균

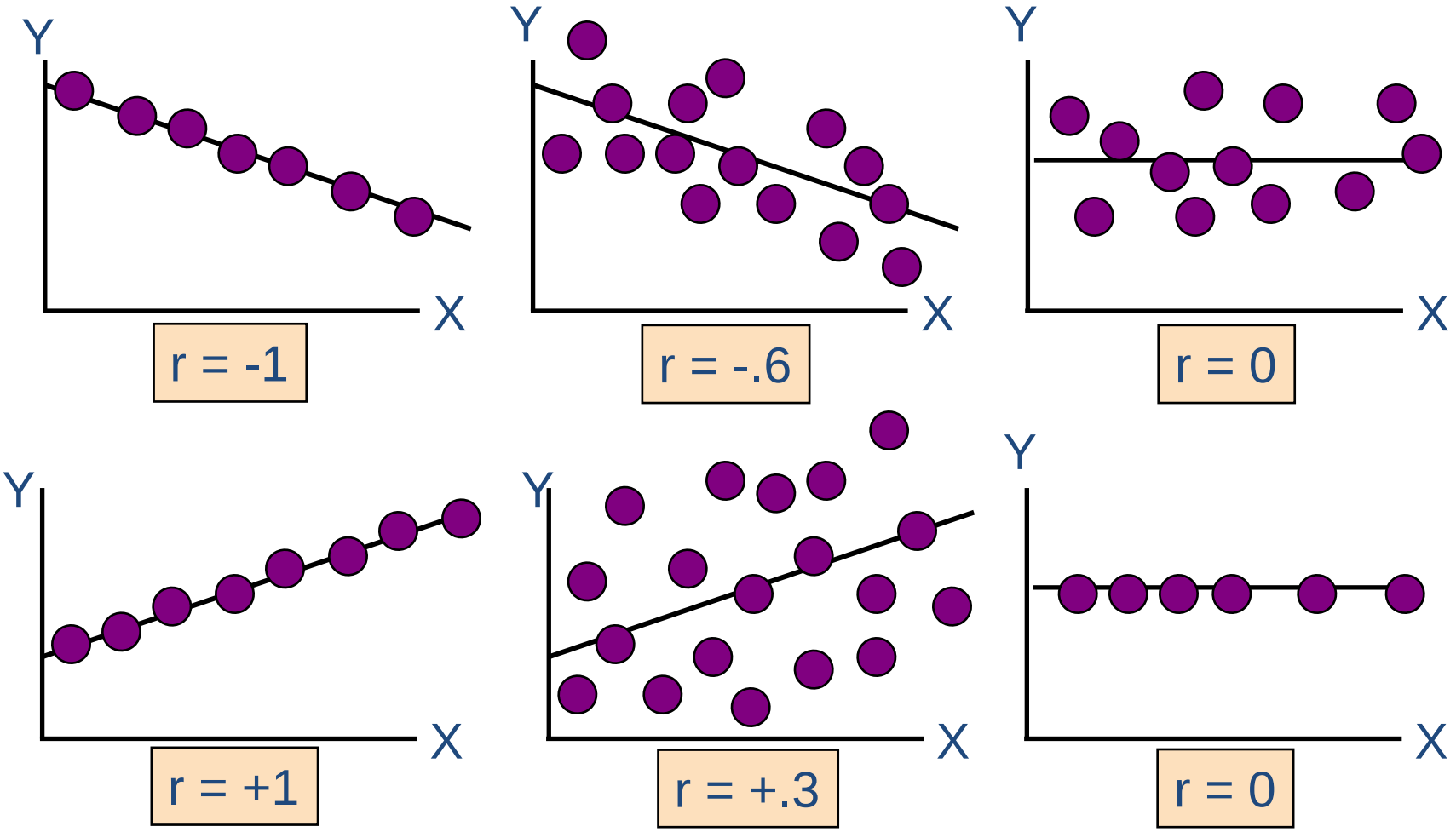


- cov(X,Y) > 0 X and Y are positively correlated
- cov(X,Y) < 0 X and Y are inversely correlated
- cov(X,Y) = 0 X and Y are independent

$$r = \frac{\mathrm{cov}\,ariance(x,y)}{\sqrt{\mathrm{var}\,x}\sqrt{\mathrm{var}\,y}}$$

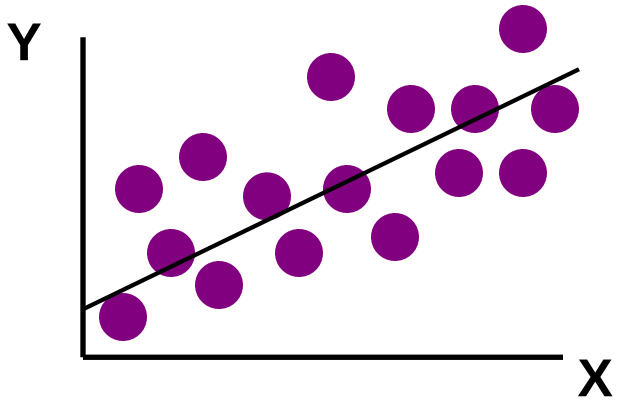
- 공분산은 단위의 영향을 받음
- 공분산을 scaling
- -1 <= 상관계수 <= 1

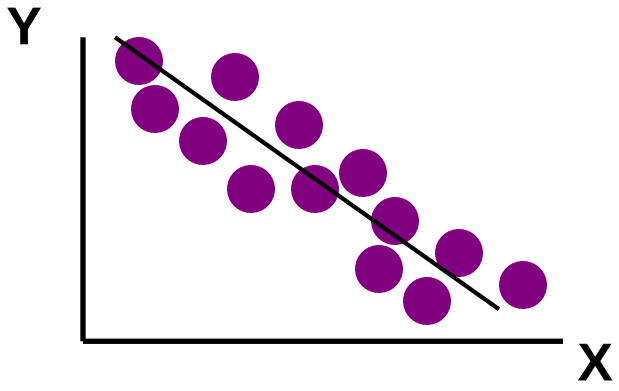
Scatter plot of correlation coefficient



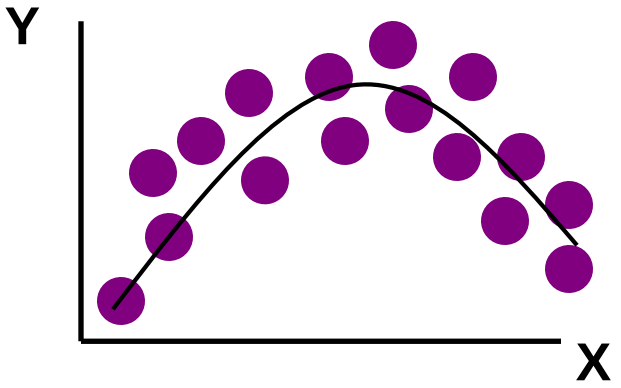
Linear correlation

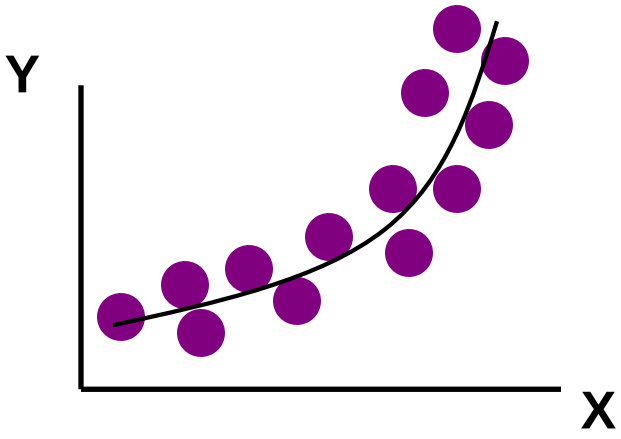
Linear relationships





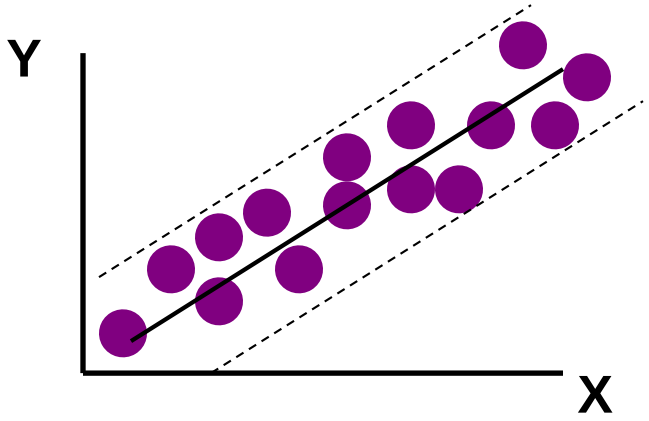
Curvilinear relationships

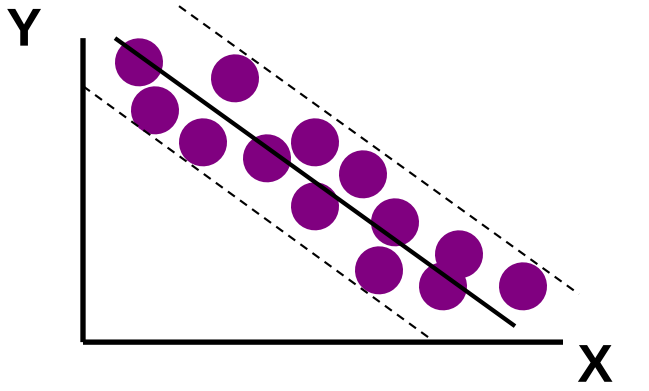




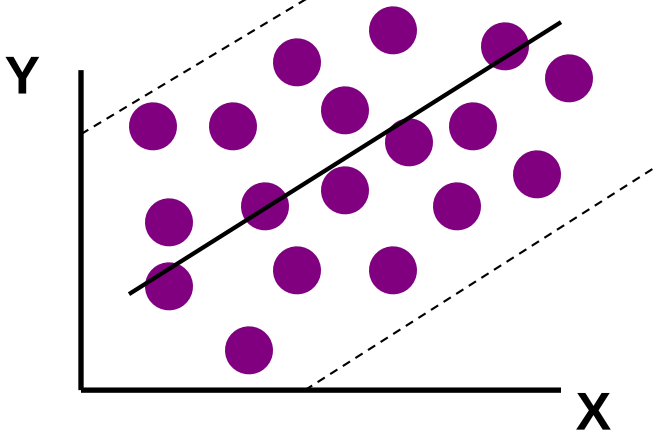
Linear correlation

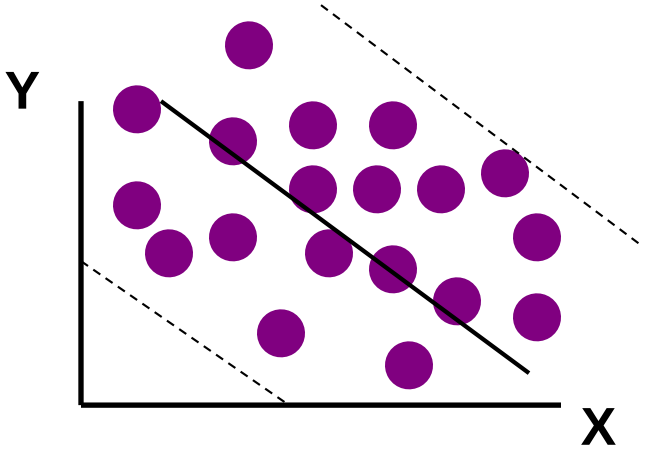
Strong relationships



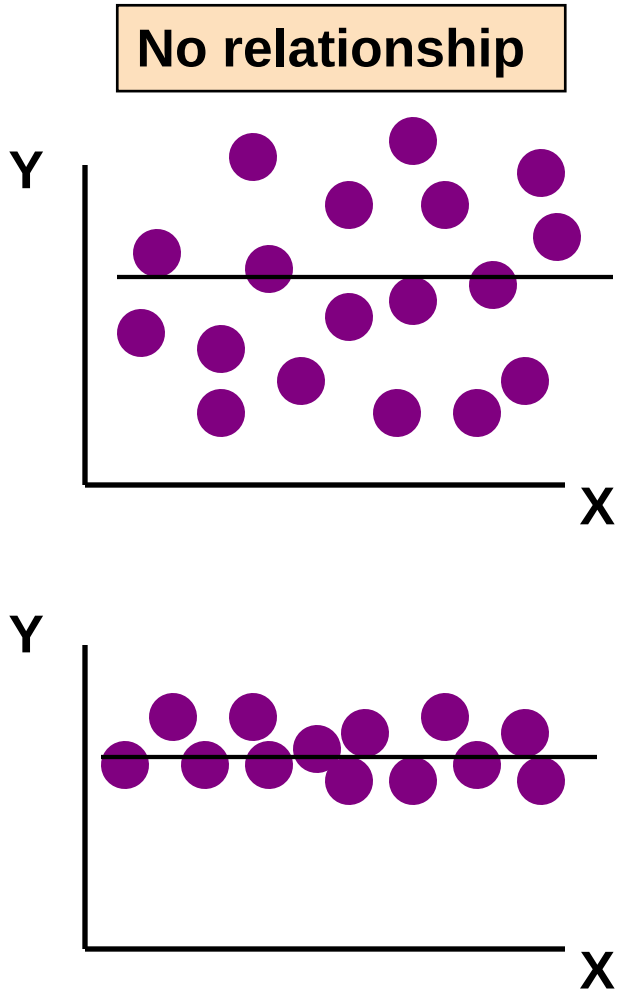


Weak relationships

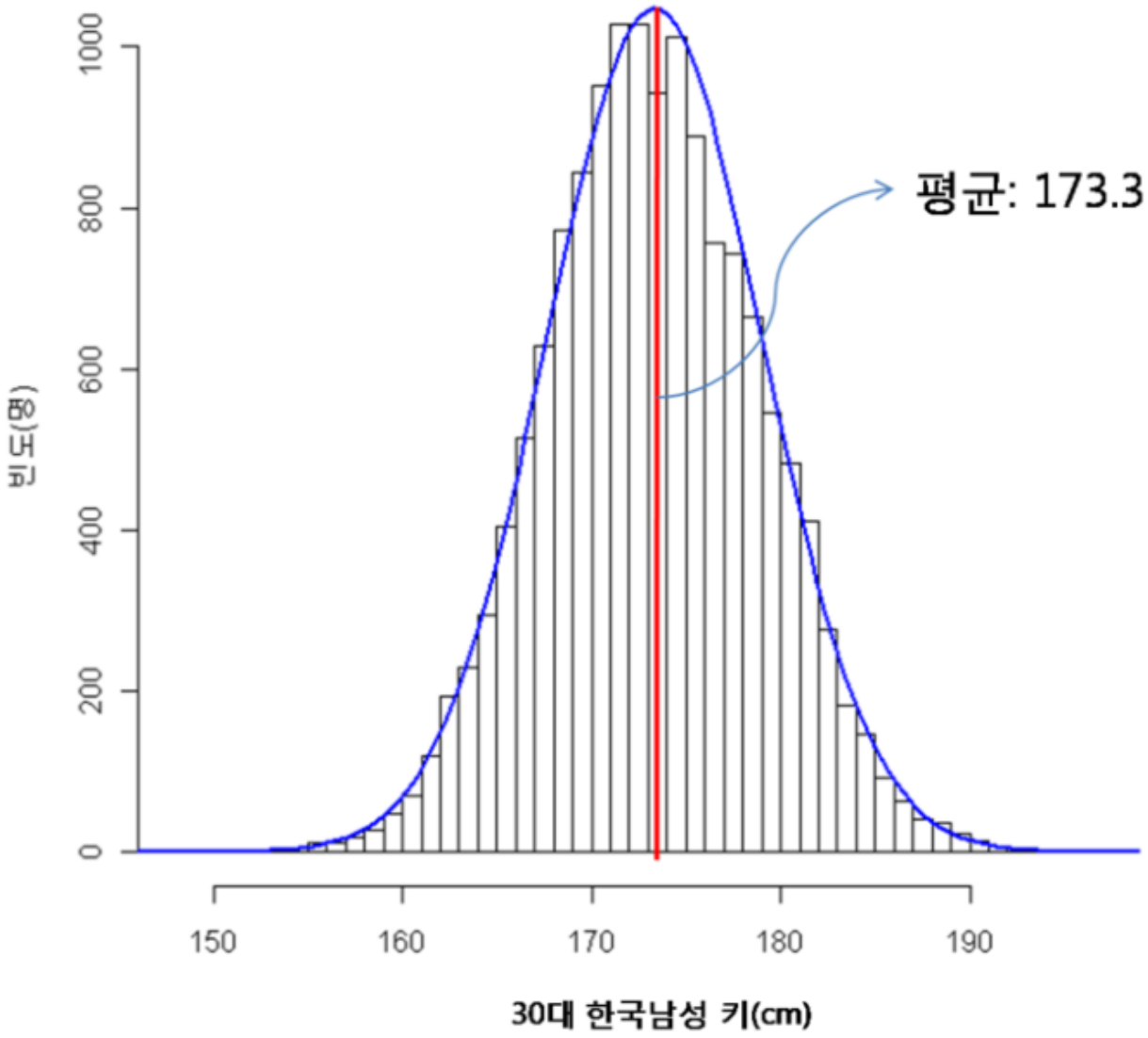




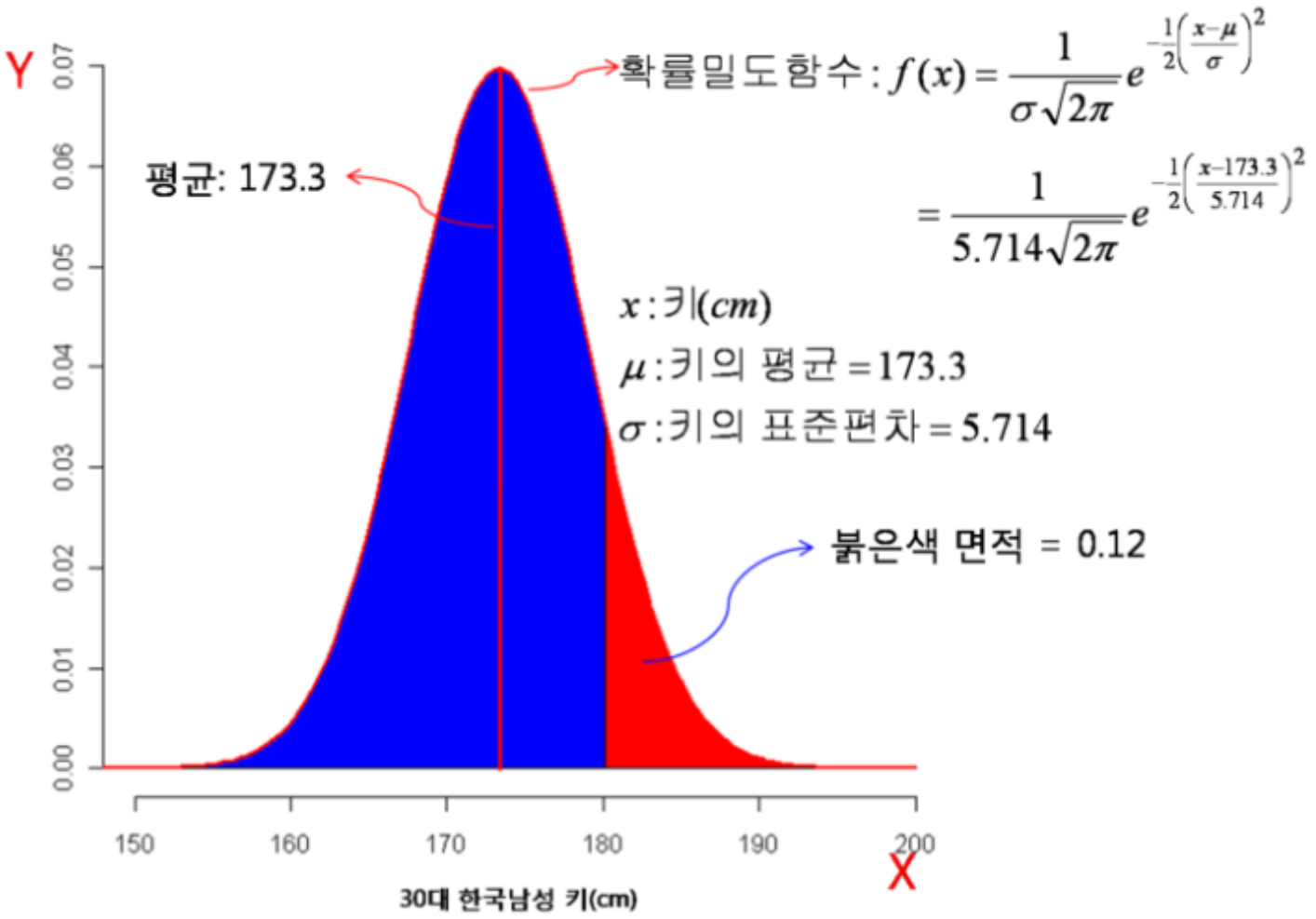
Linear correlation



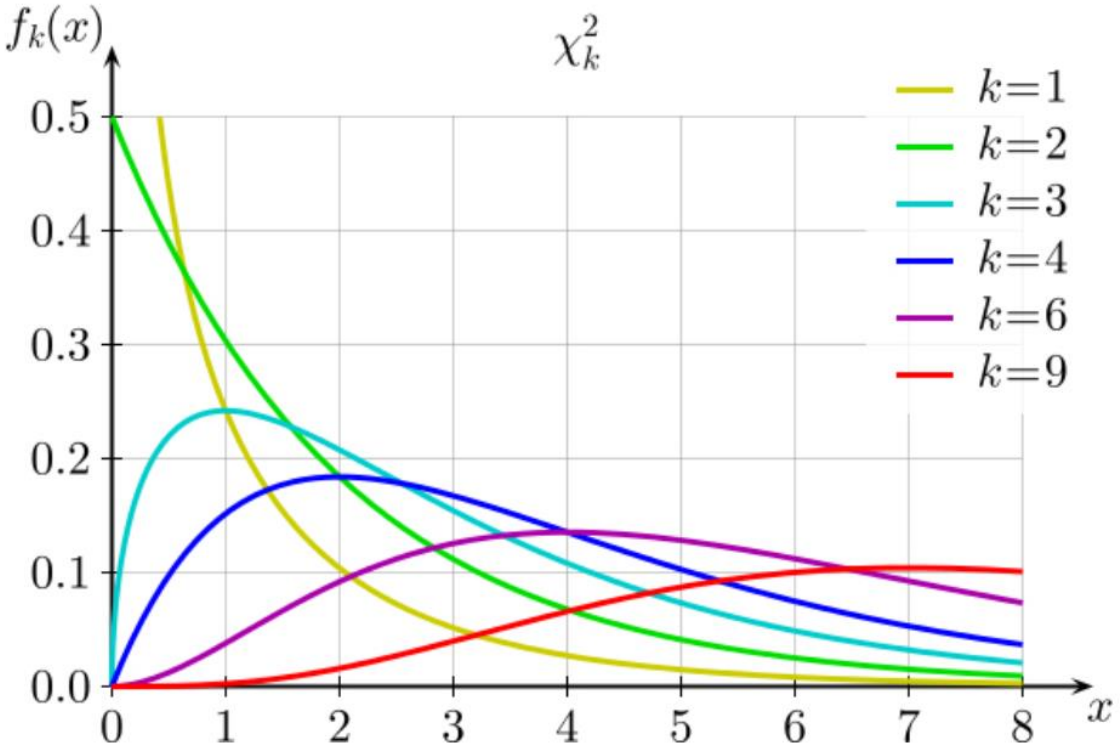
Normal Distribution(정규 분포)



Normal Distribution(정규 분포)

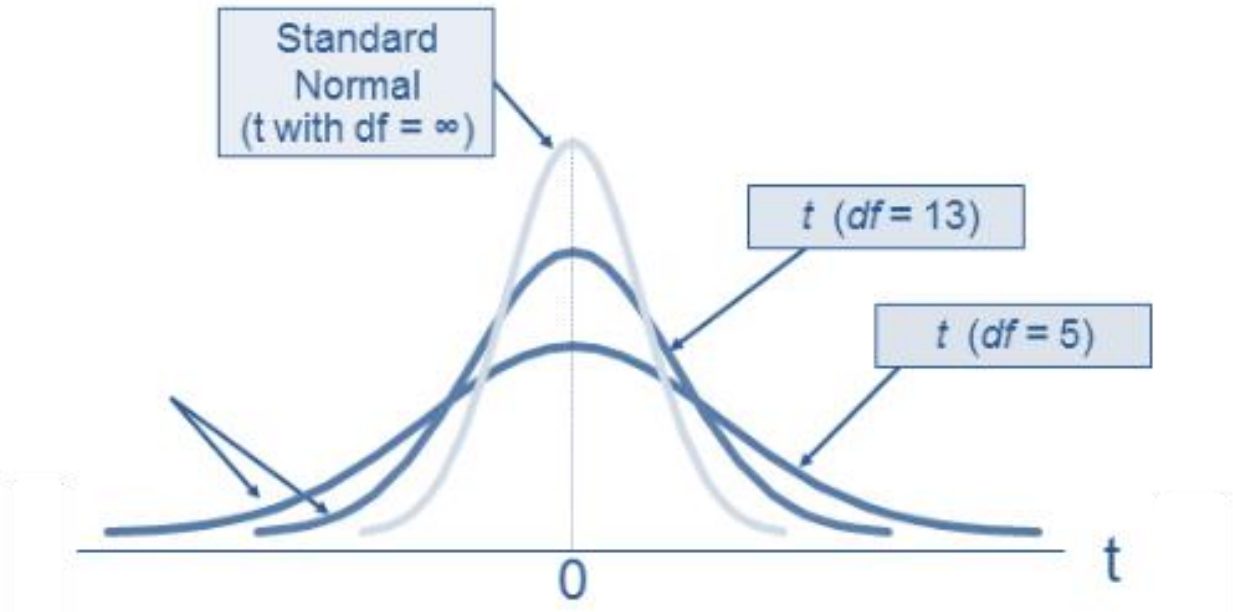


Chi-Square Distribution



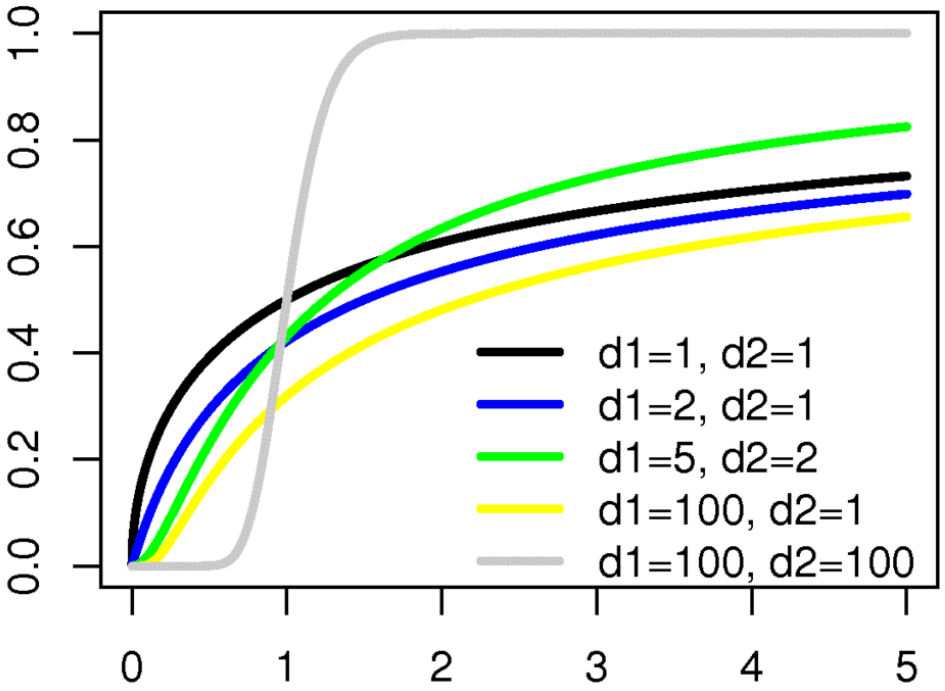
- k개의 샘플 데이터 제곱의 합의 분포
- 범주형 자료에 대한 적합성 검정
- 복수 집단에 대한 독립성 검정
- k : 자유도
- 자유도가 커지면 분포의 모양이 대칭에 가까워짐

T Distribution



- 정규분포와 후사한 분포로 t=0에 대해서 종 모양으로 대칭
- 모표준편차를 알 수 없고 표본크기가 작은 경우(n < 30)에 평균 차이 추정
- df : 자유도 (표본개수 - 1) - 자유도가 커지면 정규분포와 유사

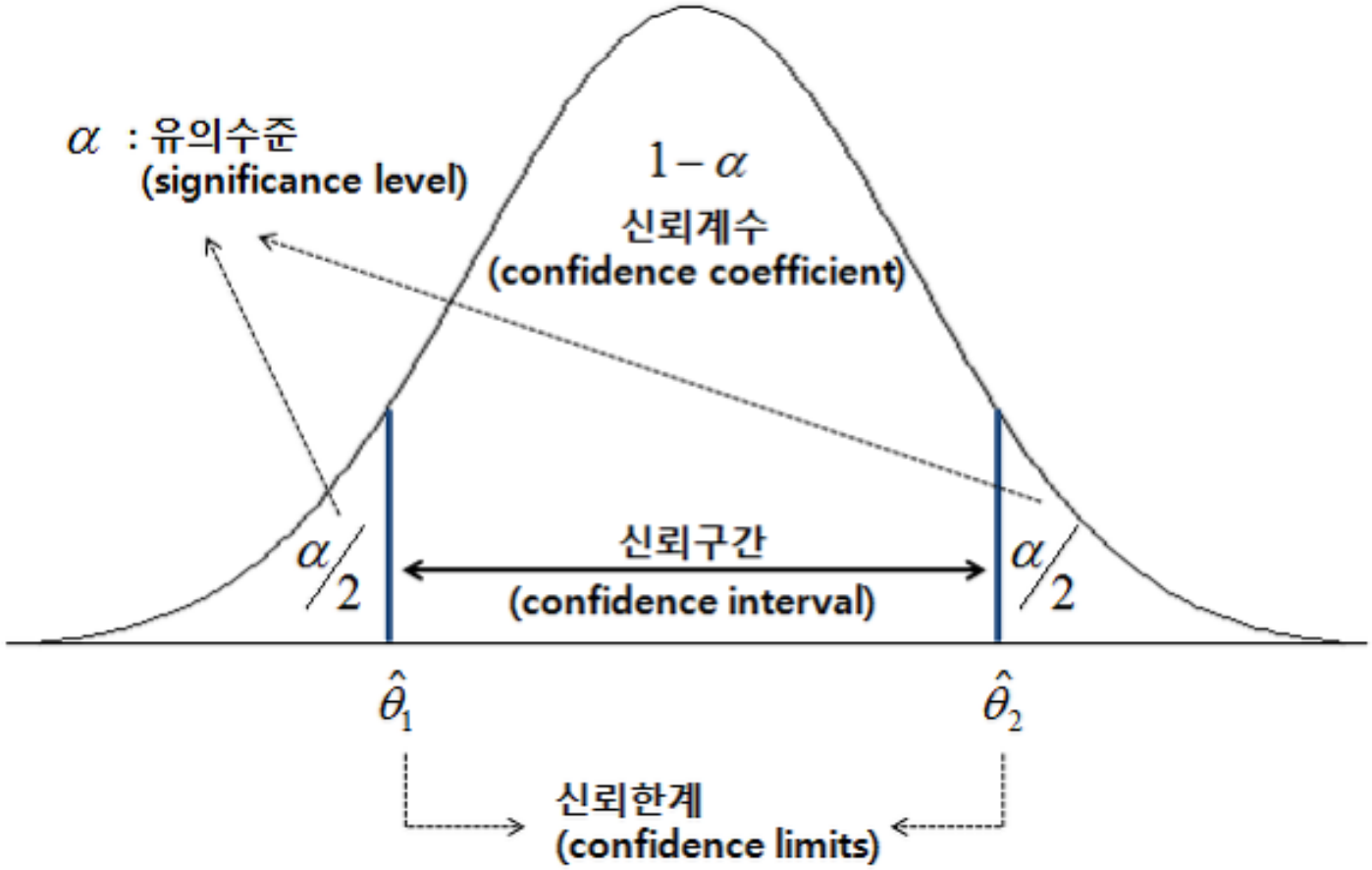
F Distribution



- 두 모집단의 분산에 대한 비의 추정
- 분산분석(ANOVA)과

  회귀분석(Regression)의 통계적 기준
- d1, d2 : 자유도
- d1 : 집단수 - 1
- d2 : 표본수 -1

Interval Estimation(구간 추정)




특집 KBS 뉴스9
신년 여론조사 : 경제
KBS 신년 여론조사
▶기관 : 미디어리서치
▶시기 : 2014년 12월 30일
▶대상 : 만 19세 이상 남녀 천 명
▶지역 : 17개 시도
▶오차 : 95% 신뢰 수준 ±3.1%p
▶방법 : 유·무선 RDD 전화 조사
▶응답률 : 17.5%
▶여론조사 공개 : KBS 홈페이지
REPUBLIC
KBS1
LLOYD
Holika Holika
삼성 부동산
특집
9

통계적 가설 검정(Statistical Hypothesis Test)

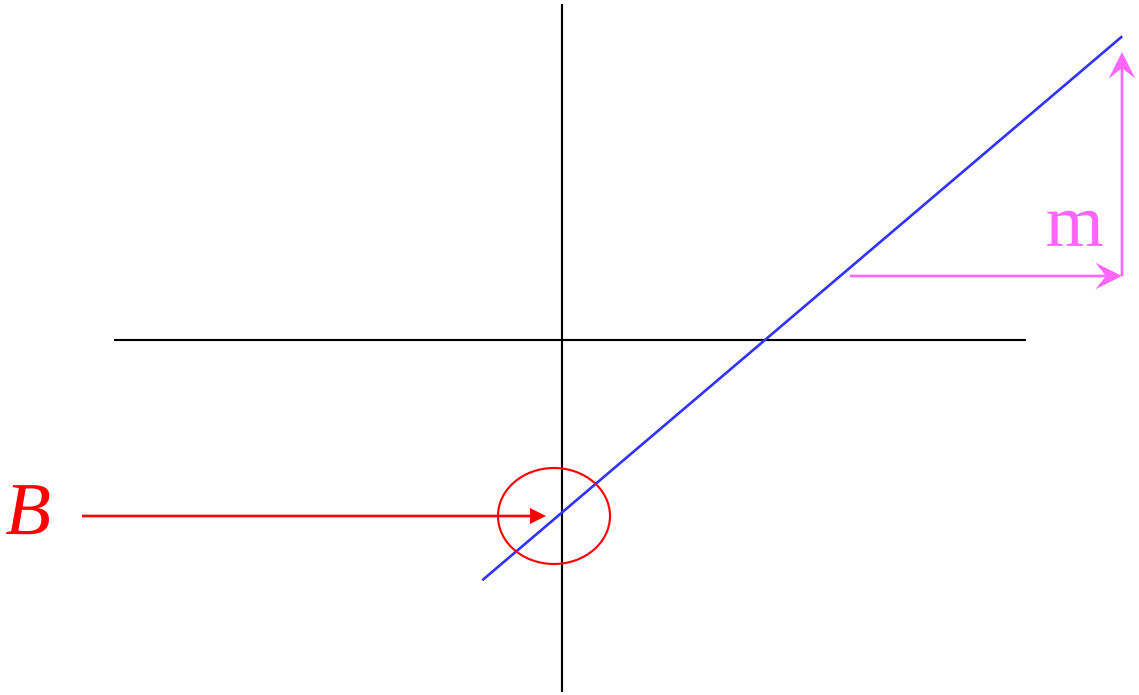
- T-Test : 두 개의 집단간의 차이(평균)가 의미가 있는지 확인
  - 두 개의 집단 간의 평균 비교 : 남자와 여자의 소득 비교
  - 하나의 집단에 대한 비교 : 과외를 하기 전과 후의 성적 변화
  - 특정 집단의 평균이 어떤 숫자와 같은지 다른지 비교
- ANOVA-Test(ANalysis Of VAriance; 분산 분석) : 여러 그룹간(2개 이상)의 평균의 차이가 통계적으로 유의미 한지를 판단
  - 정규성 : 각각의 그룹에서 변인은 정규분포.
  - 분산의 동질성 : Y의 모집단 분산은 각각의 모집단에서 동일.
  - 관찰의 독립성: 각각의 모집단에서 크기가 각각인 표본들이 독립적으로 표집
- Chisquare-Test : 그룹간에 차이가 있는지의 여부에 대해 Chisquare분포를 사용하는 가설검정
  - 그룹간에 비율(proportion) 차이가 있는지의 여부

Linear regression

In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (X) and the other the dependent (=outcome) variable Y.

What is “Linear”?

- Remember this:
- *Y=mX+B?*



What's Slope?

A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

Prediction

If you know something about X, this knowledge helps you predict something about Y. (Sound familiar?…sound like conditional probabilities?)

Regression equation

Expected value of y at a given level of *x*=

$$E(y_i / x_i) = \alpha + \beta x_i$$

Predicted value for an individual…

$$y_i = \alpha + \beta * x_i + \text{random error}_i$$

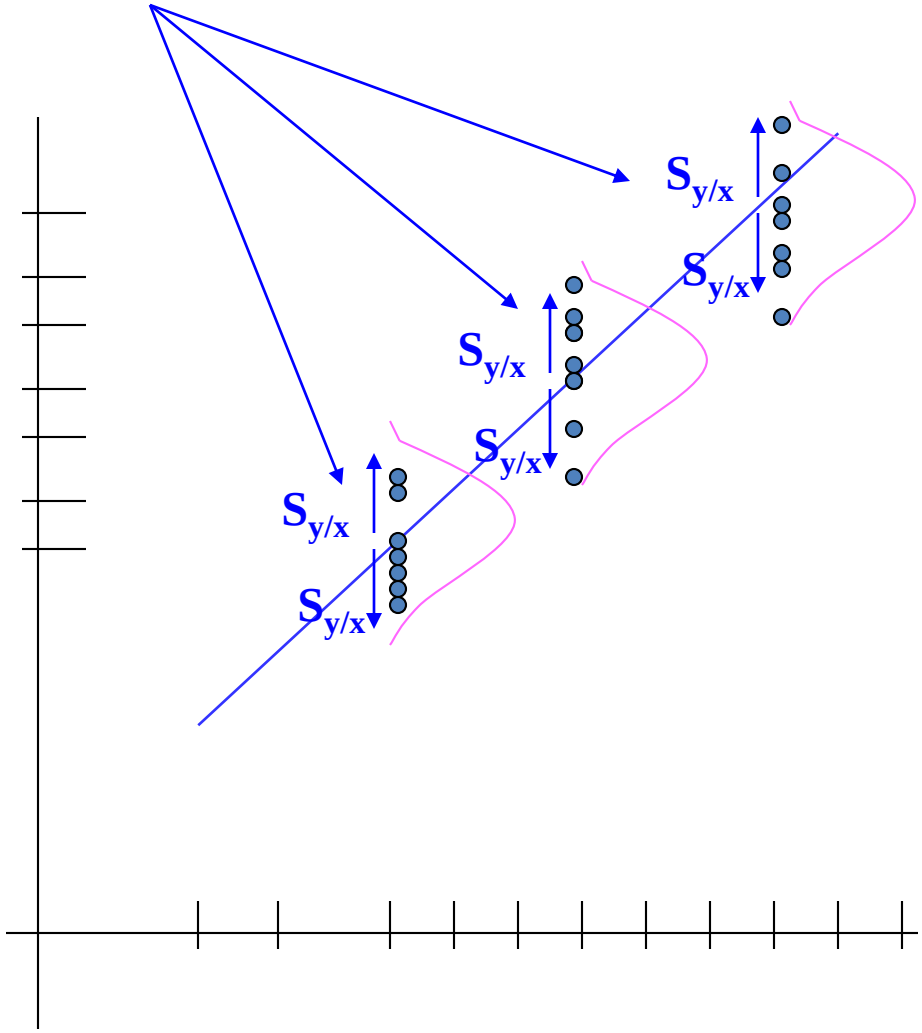
Fixed – exactly on the line

Follows a normal distribution

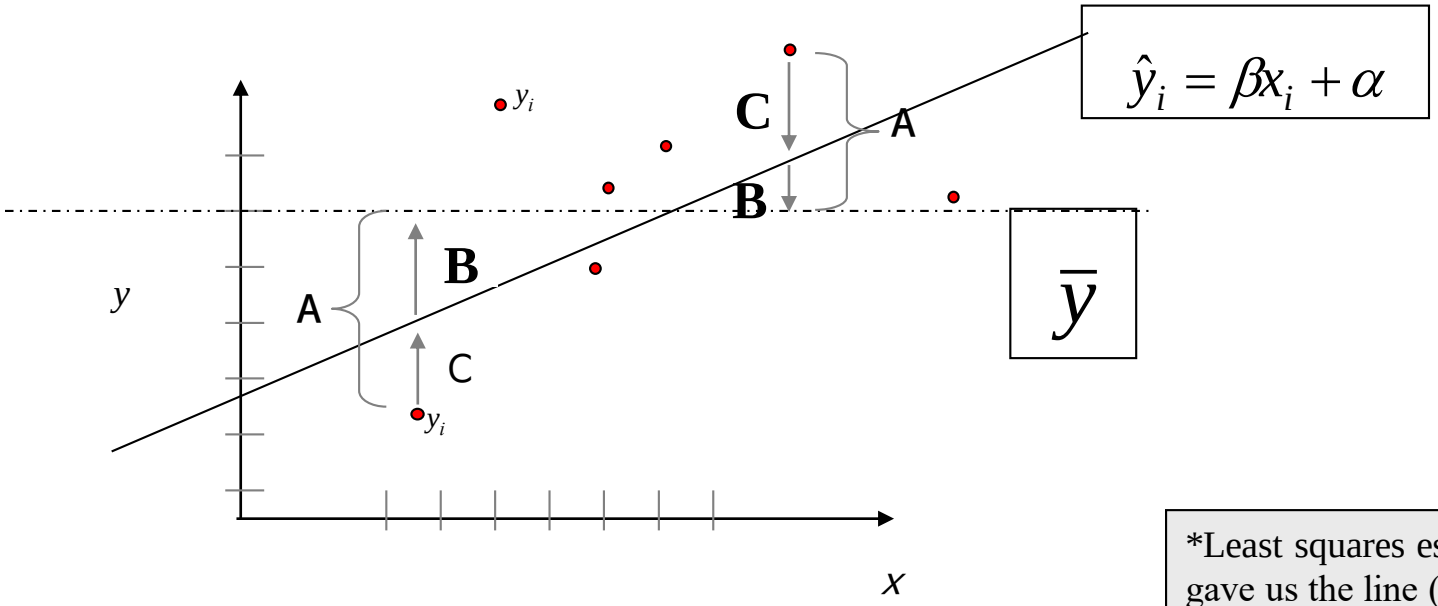
Assumptions (or the fine print)

- Linear regression assumes that…
  - 1. The relationship between X and Y is linear
  - 2. Y is distributed normally at each value of X
  - 3. The variance of Y at every value of X is the same (homogeneity of variances)
  - 4. The observations are independent

The standard error of Y given X is the average variability around the regression line at any given value of X. It is assumed to be equal at all values of X.



Regression Picture



*Least squares estimation gave us the line (β) that minimized C^2

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$R^2 = SS_{reg}/SS_{total}$

A^2

SS_{total}

Total squared distance of observations from naïve mean of y

Total variation

B^2

SS_{reg}

Distance from regression line to naïve mean of y

Variability due to x (regression)

C^2

$SS_{residual}$

Variance around the regression line

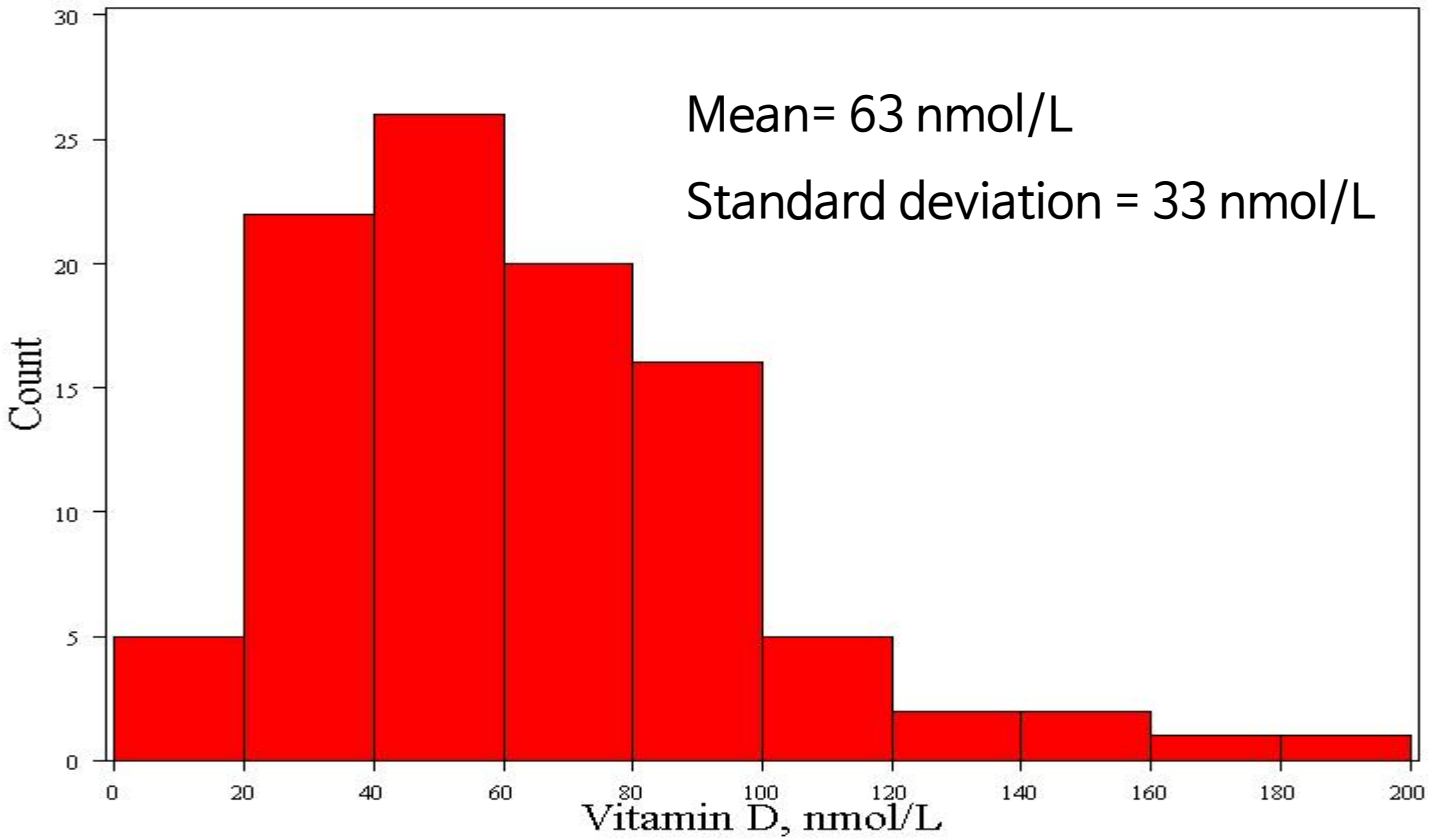
Additional variability not explained by x—what least squares method aims to minimize

Recall example: cognitive function and vitamin D

- Hypothetical data loosely based on [1]; cross-sectional study of 100 middle-aged and older European men.
  - Cognitive function is measured by the Digit Symbol Substitution Test (DSST).

1. Lee DM, Tajar A, Ulubaev A, et al. Association between 25-hydroxyvitamin D levels and cognitive performance in middle-aged and older European men. J Neurol Neurosurg Psychiatry. 2009 Jul;80(7):722-9.

Distribution of vitamin D

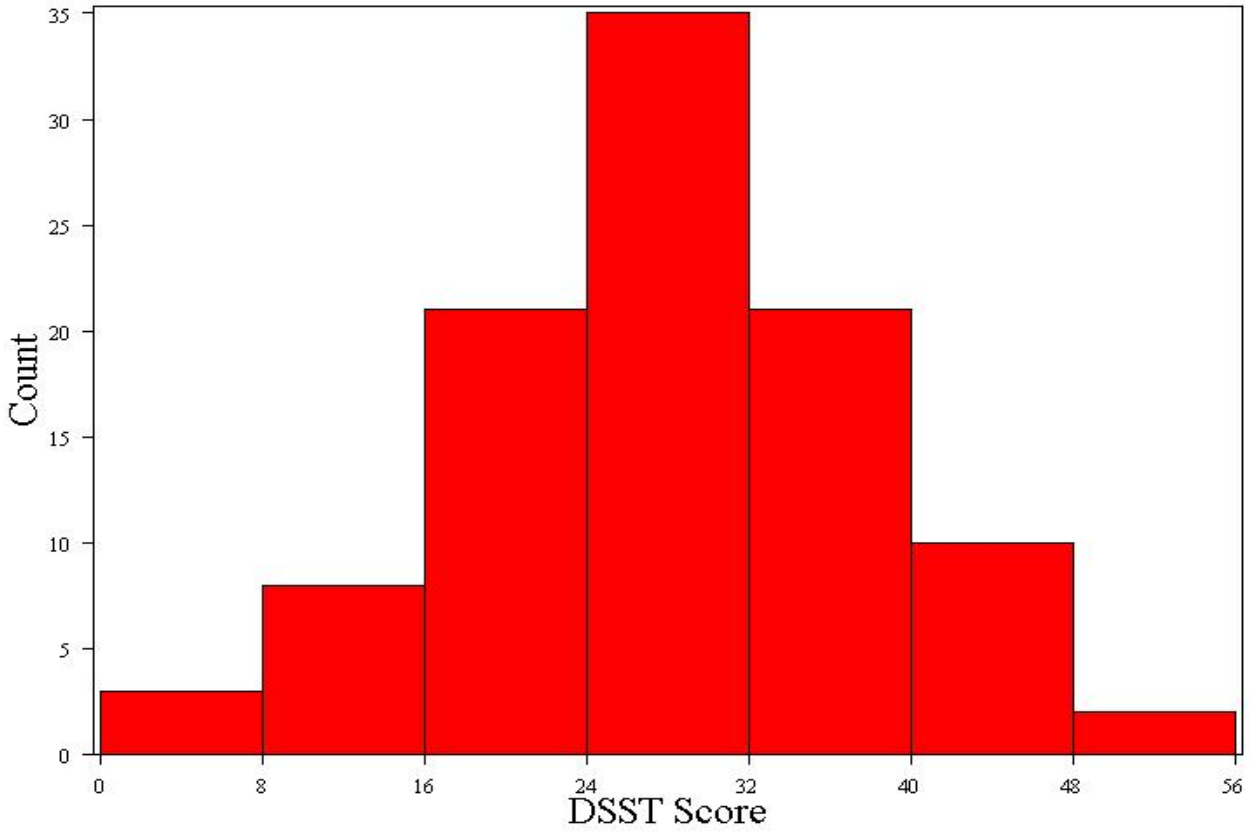


Distribution of DSST

Normally distributed

Mean = 28 points

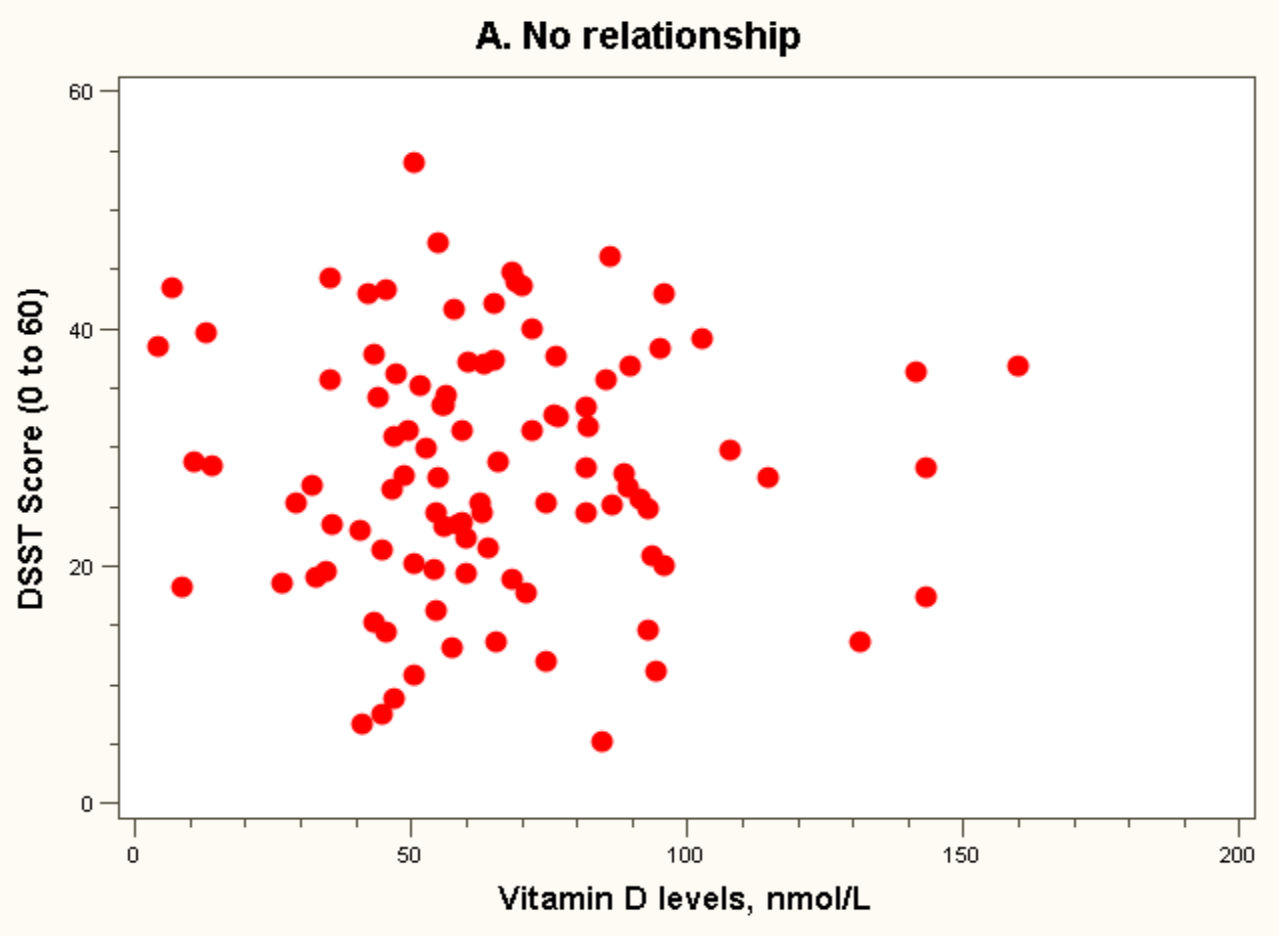
Standard deviation = 10 points



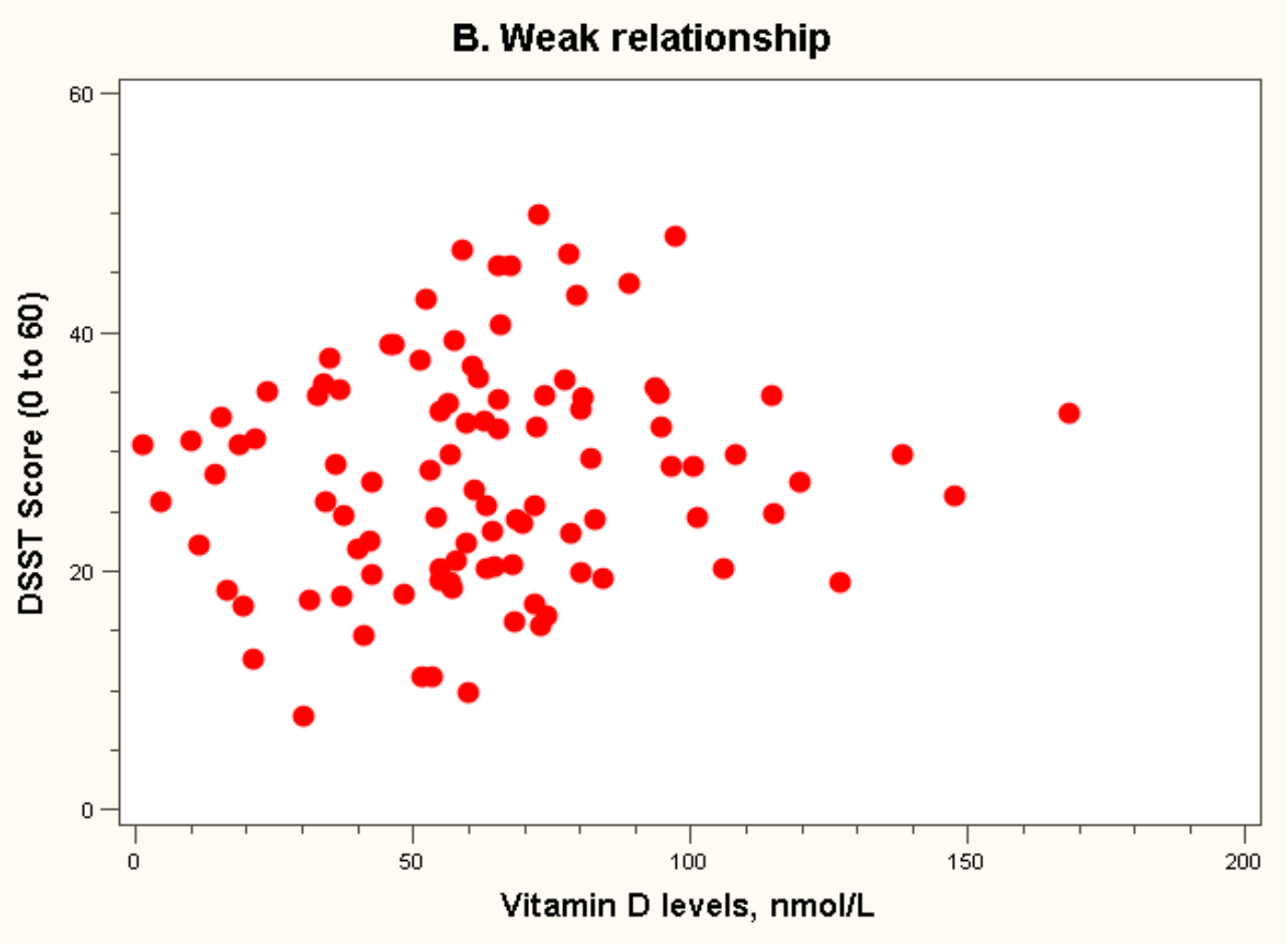
Four hypothetical datasets

- I generated four hypothetical datasets, with increasing TRUE slopes (between vit D and DSST):
  - 0
  - 0.5 points per 10 nmol/L
  - 1.0 points per 10 nmol/L
  - 1.5 points per 10 nmol/L

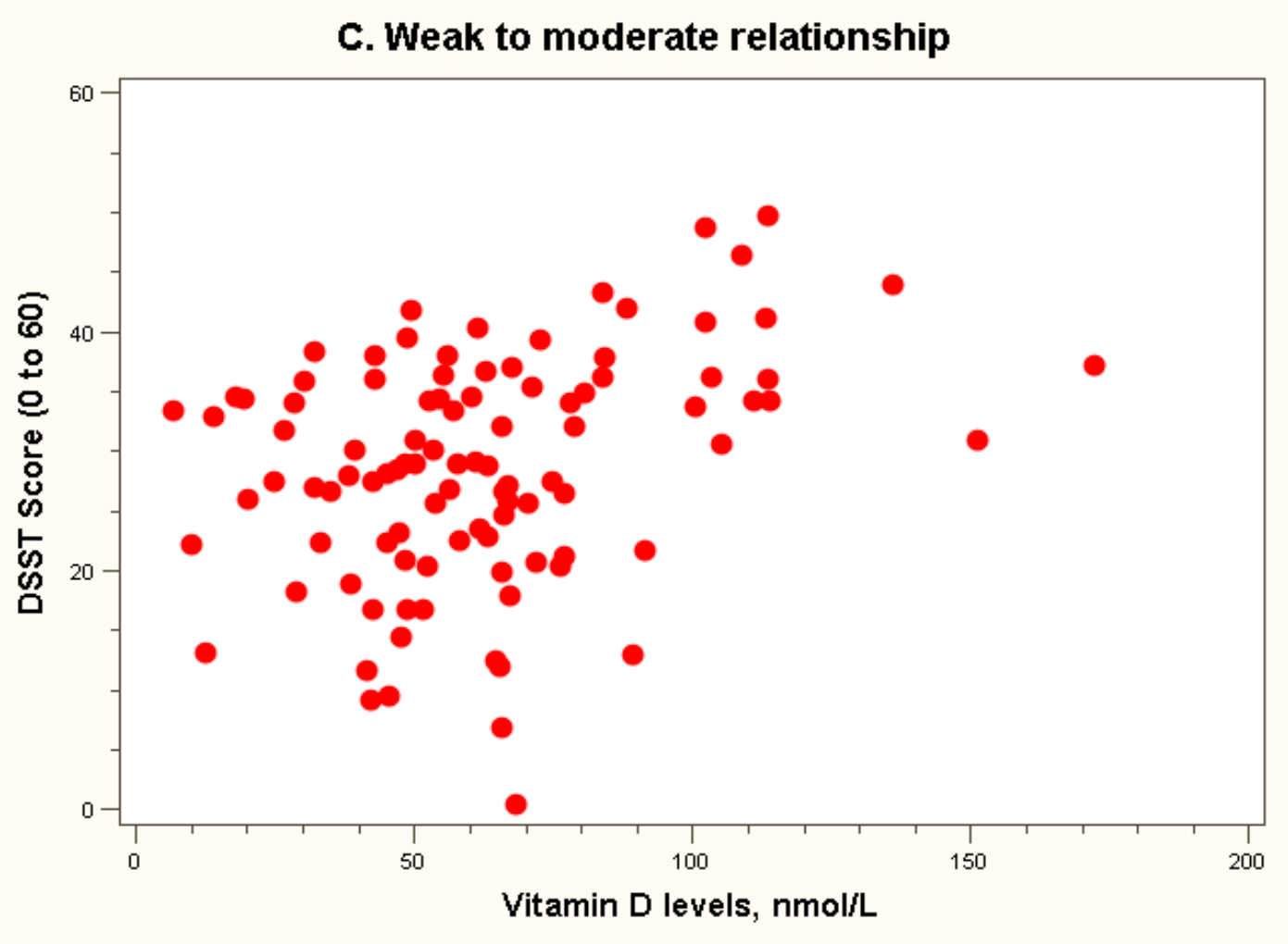
Dataset 1: no relationship



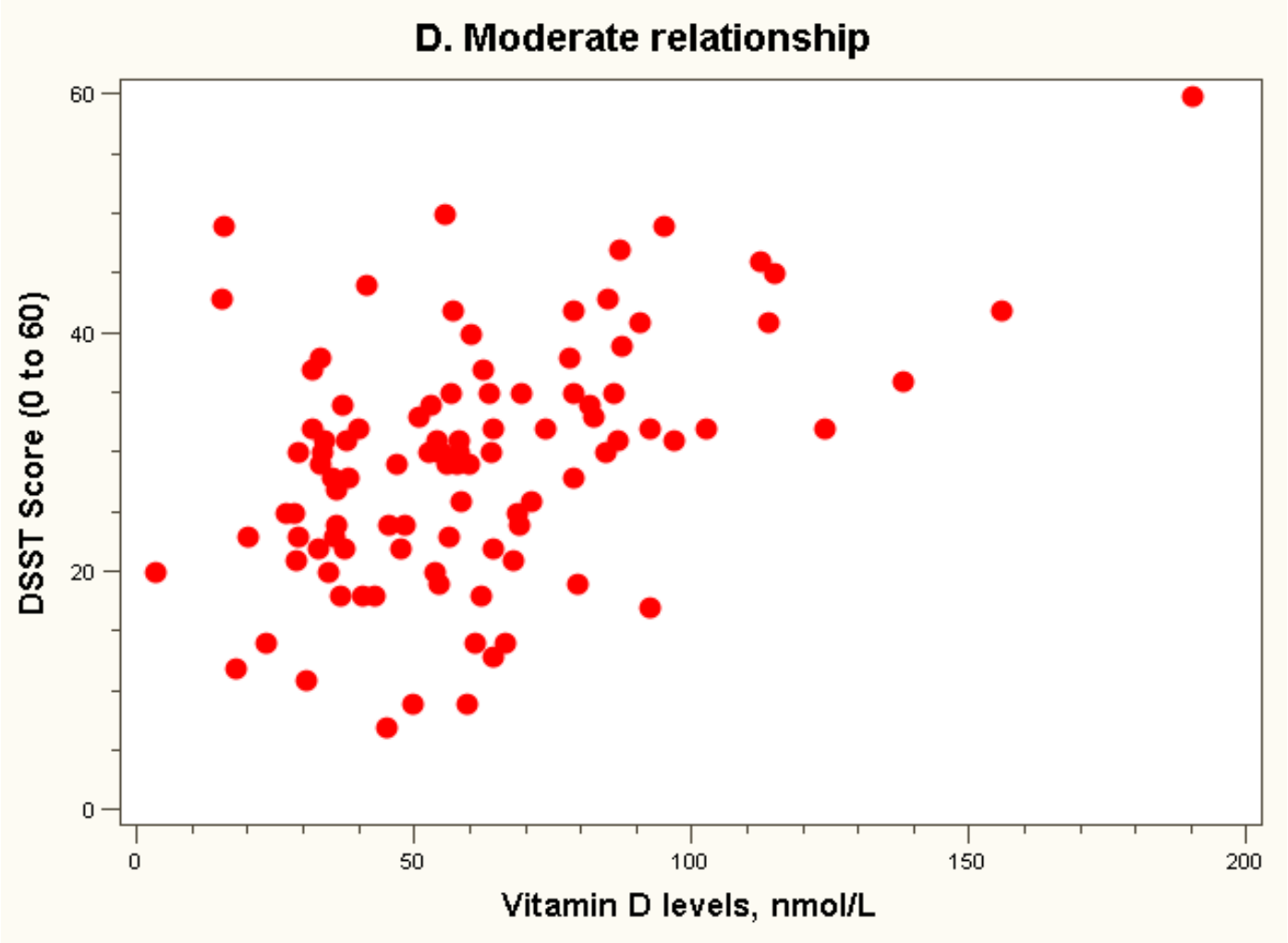
Dataset 2: weak relationship



Dataset 3: weak to moderate relationship



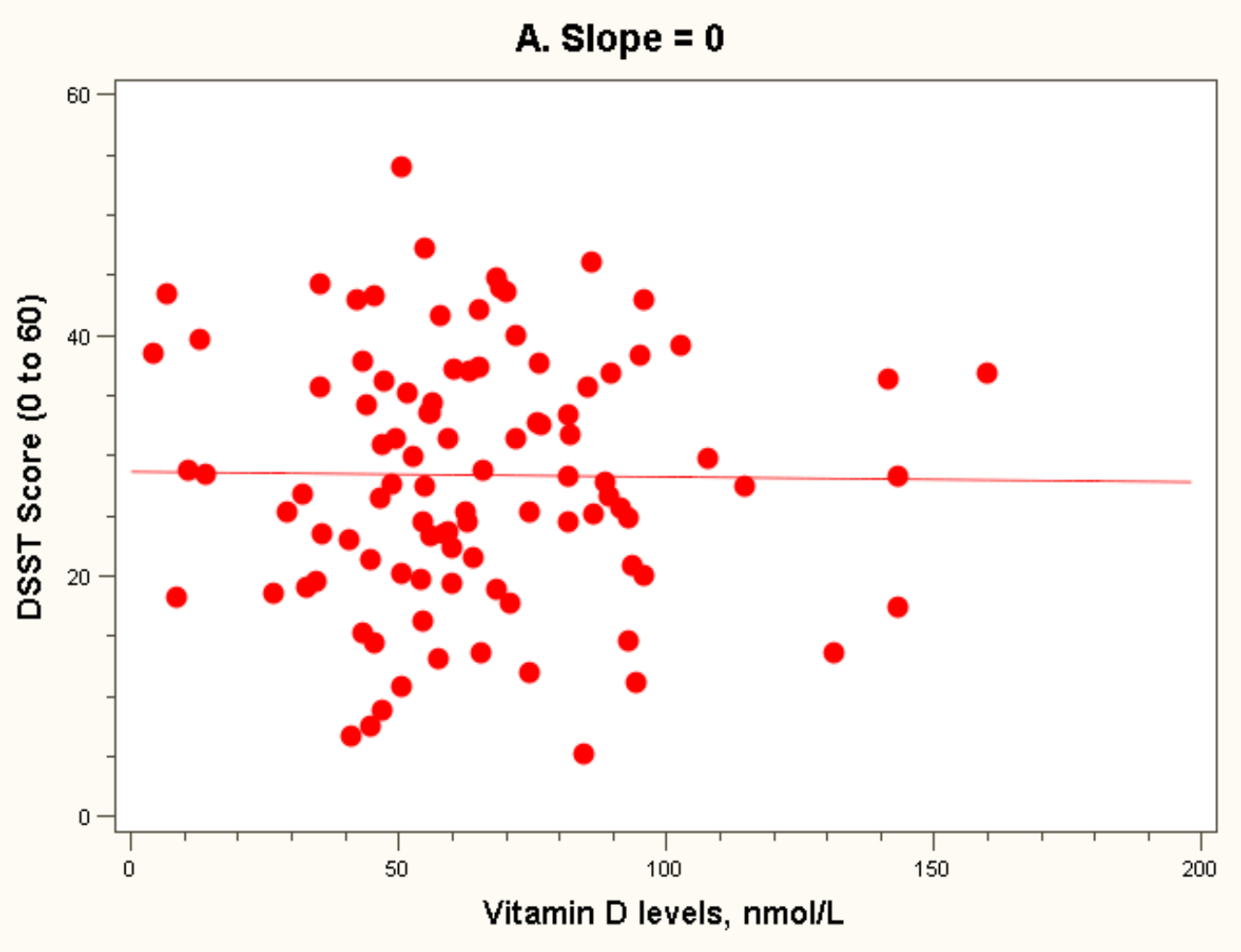
Dataset 4: moderate relationship



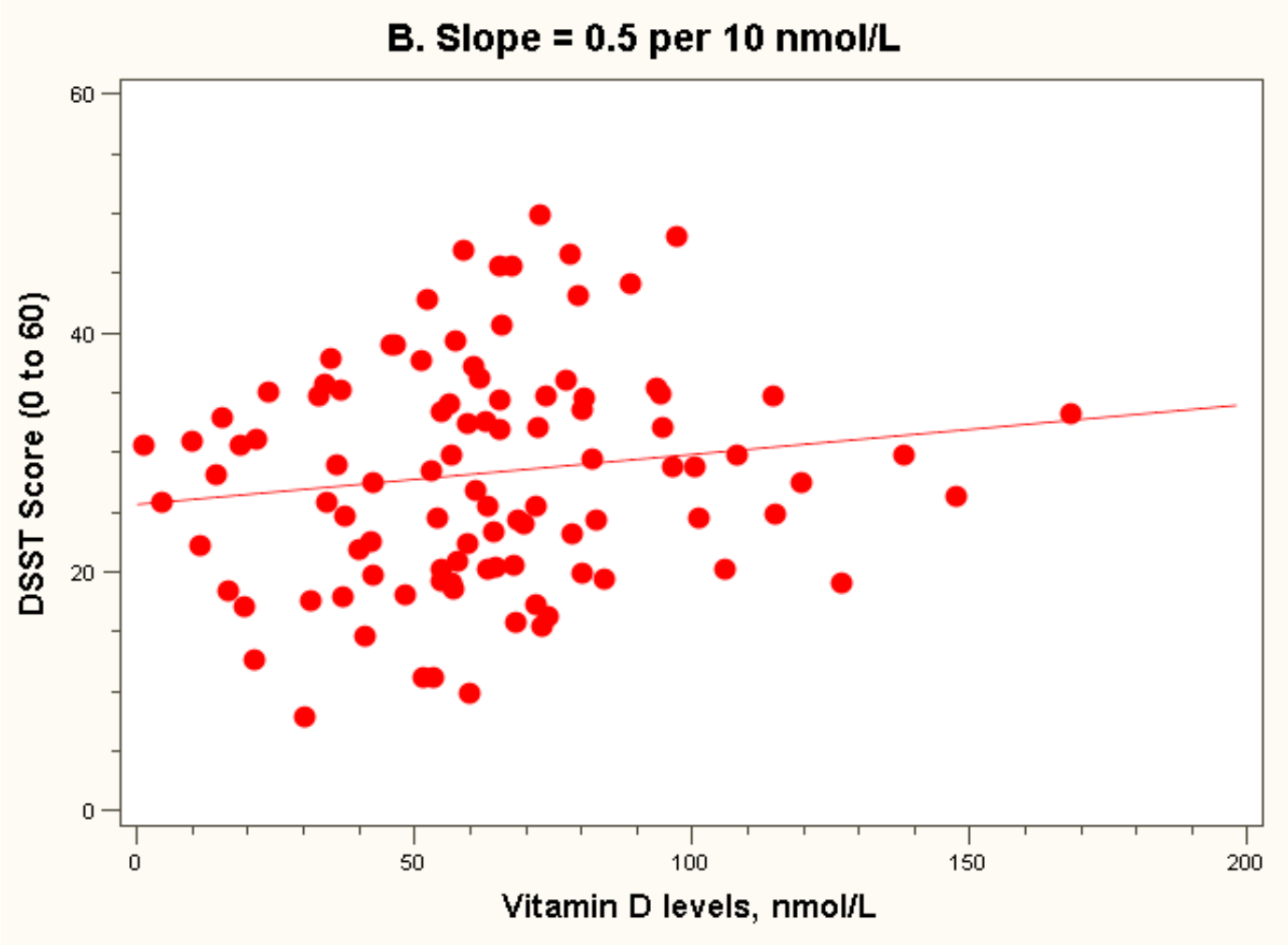
The “Best fit” line

Regression equation:

$E(Y_i) = 28 + 0*\text{vit } D_i$ (in 10 nmol/L)



The “Best fit” line

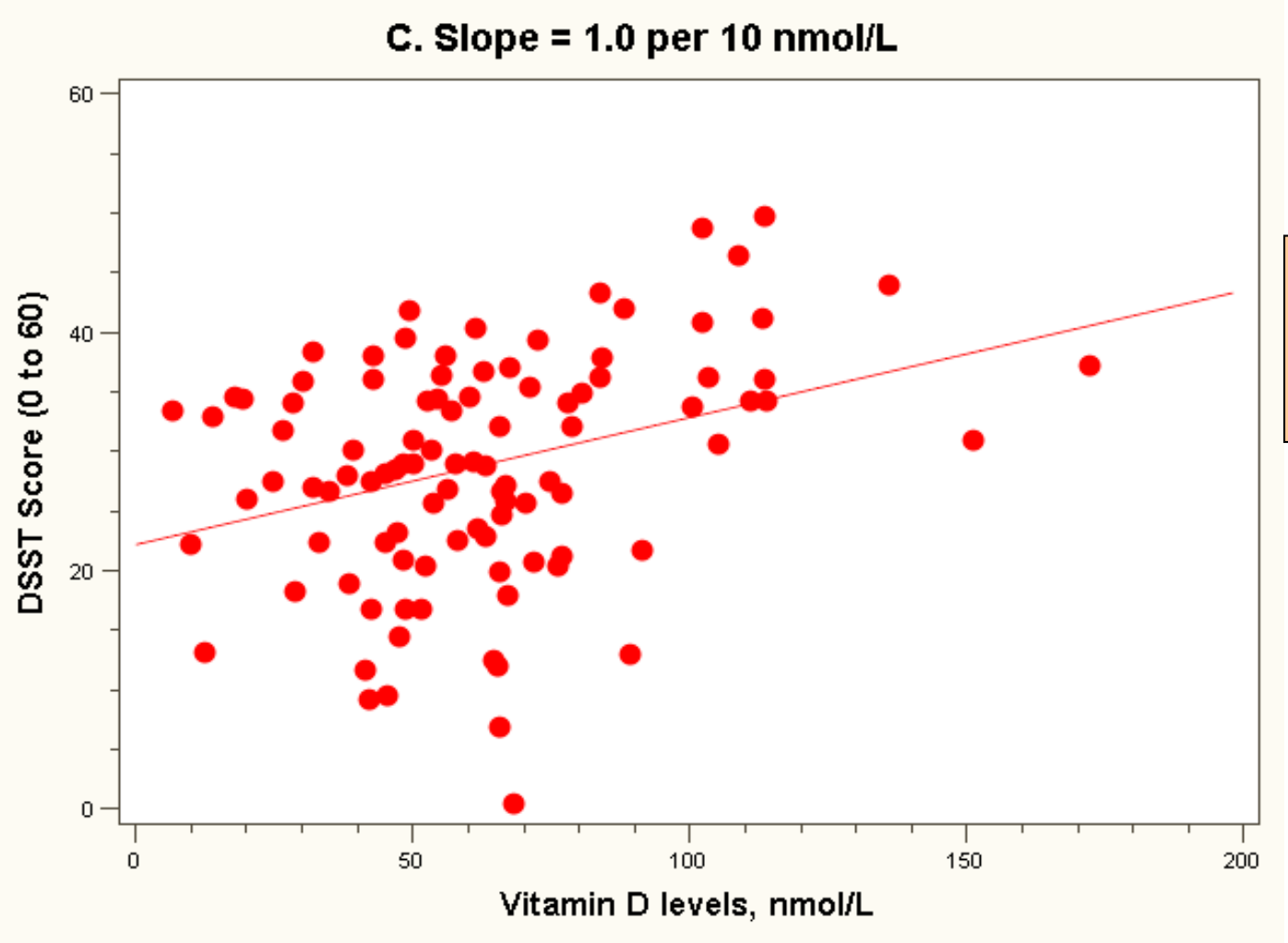


Note how the line is a little deceptive; it draws your eye, making the relationship appear stronger than it really is!

Regression equation:

$E(Y_i) = 26 + 0.5*vit D_i$ (in 10 nmol/L)

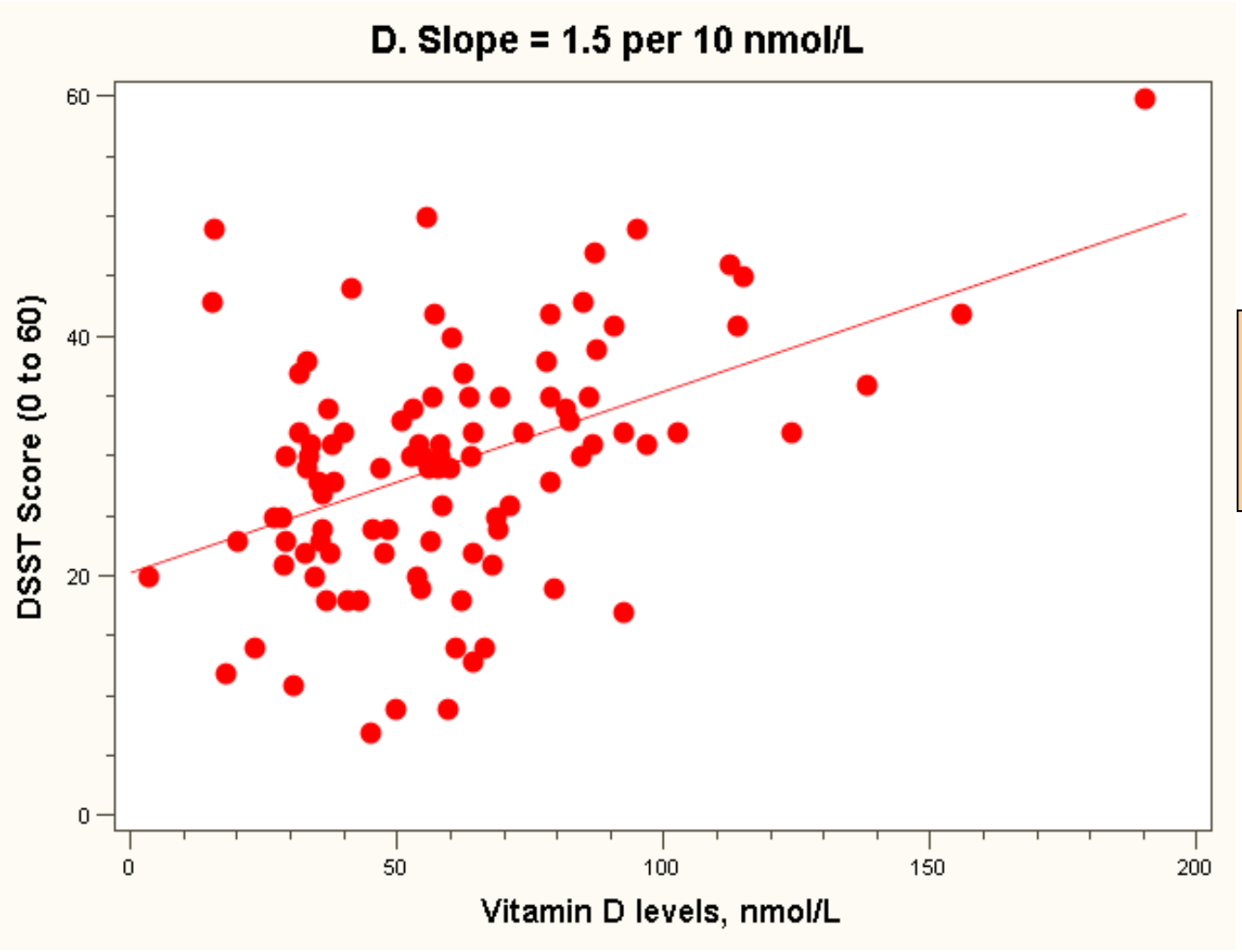
The “Best fit” line



Regression equation:

$E(Y_i) = 22 + 1.0*vit D_i$ **(in 10 nmol/L)**

The “Best fit” line



Regression equation:

$E(Y_i) = 20 + 1.5*\text{vit } D_i$ **(in 10 nmol/L)**

Note: all the lines go through the point (63, 28)!

Estimating the intercept and slope: least squares estimation

** Least Squares Estimation
A little calculus….
What are we trying to estimate? **β, the slope,** from

What's the constraint? We are trying to minimize the squared distance (hence the "least squares") between the observations themselves and the predicted values , or (also called the "residuals", or left-over unexplained variability)

$\text{Difference}_i = y_i - (\beta x + \alpha)$ $\text{Difference}_i^2 = (y_i - (\beta x + \alpha))^2$

Find the β that gives the minimum sum of the squared differences. How do you maximize a function? Take the derivative; set it equal to zero; and solve. Typical max/min problem from calculus….

$$\frac{d}{d\beta}\sum_{i=1}^{n}(y_i - (\beta x_i + \alpha))^2 = 2(\sum_{i=1}^{n}(y_i - \beta x_i - \alpha)(-x_i))$$

$$2(\sum_{i=1}^{n}(-y_i x_i + \beta x_i^{\,2} + \alpha x_i)) = 0\ldots$$

From here takes a little math trickery to solve for β…

Resulting formulas…

Slope (beta coefficient) = $\hat{\beta} = \frac{Cov(x, y)}{Var(x)}$

Intercept= $\text{Calculate}: \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

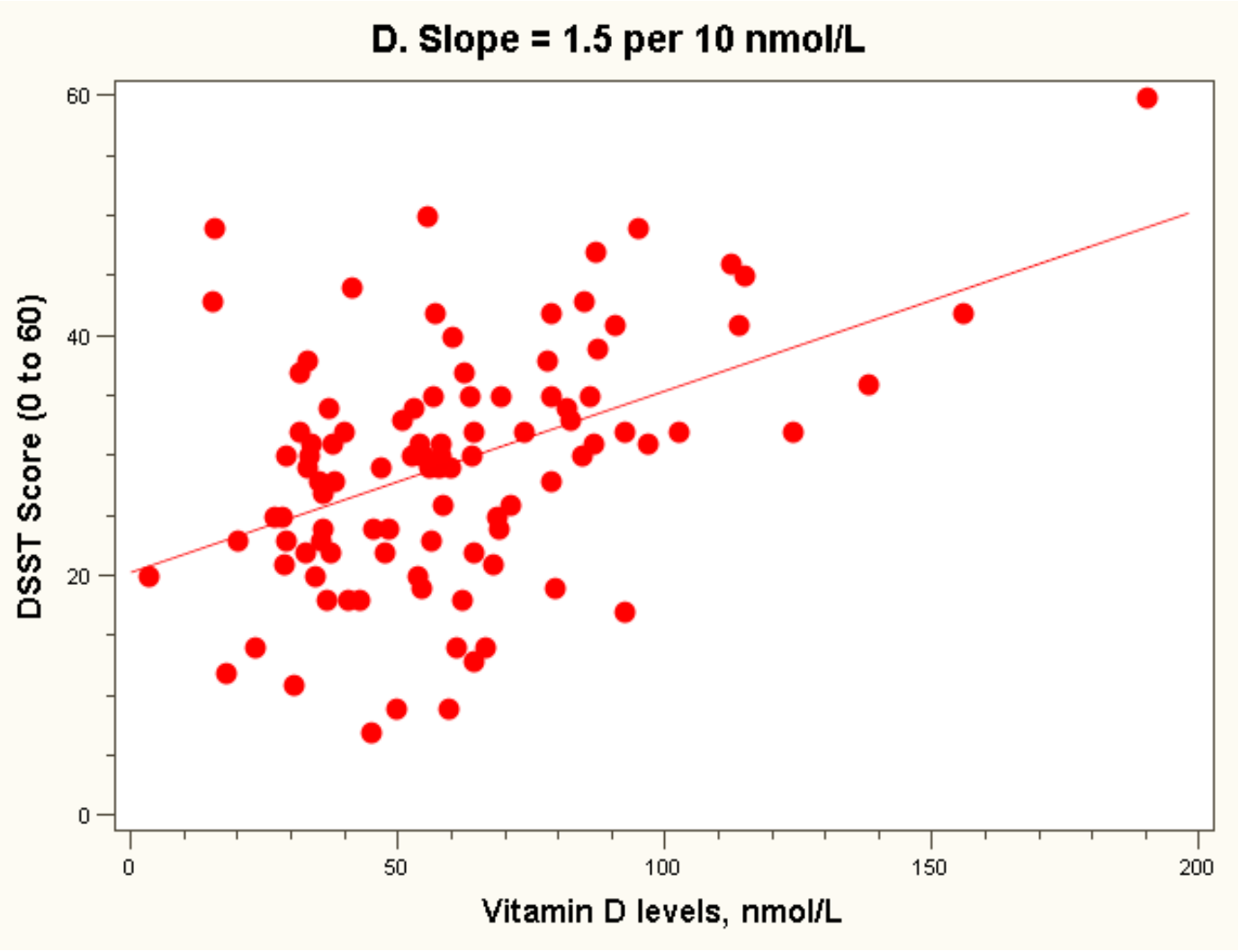
Regression line always goes through the point: $(\bar{x}, \bar{y})$

Relationship with correlation

$$\hat{r} = \hat{\beta} \frac{SD_x}{SD_y}$$

In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (*X*) and the other the dependent (=outcome) variable *Y*.

Example: dataset 4



SDx = 33 nmol/L

SDy= 10 points

Cov(X,Y) = 163 points*nmol/L

Beta = $163/33^2 = 0.15$ points per nmol/L

= 1.5 points per 10 nmol/L

r = 163/(10*33) = 0.49

Or

r = 0.15 * (33/10) = 0.49

Residual Analysis: check assumptions

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
  - Evaluate normal distribution assumption
  - Evaluate independence assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

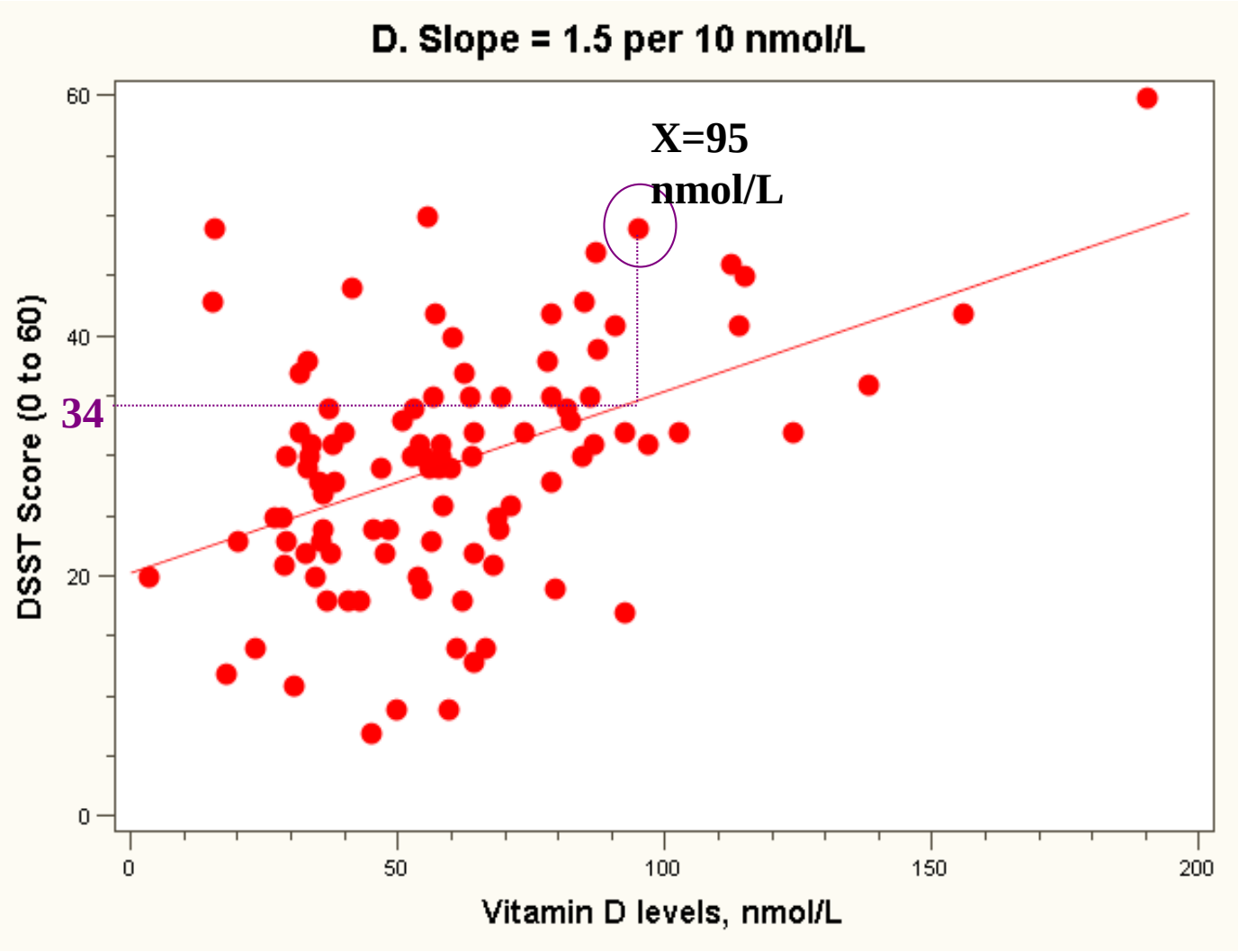
Predicted values…

$$\hat{y}_i = 20 + 1.5x_i$$

For Vitamin D = 95 nmol/L (or 9.5 in 10 nmol/L):

$$\hat{y}_i = 20 + 1.5(9.5) = 34$$

Residual = observed − predicted

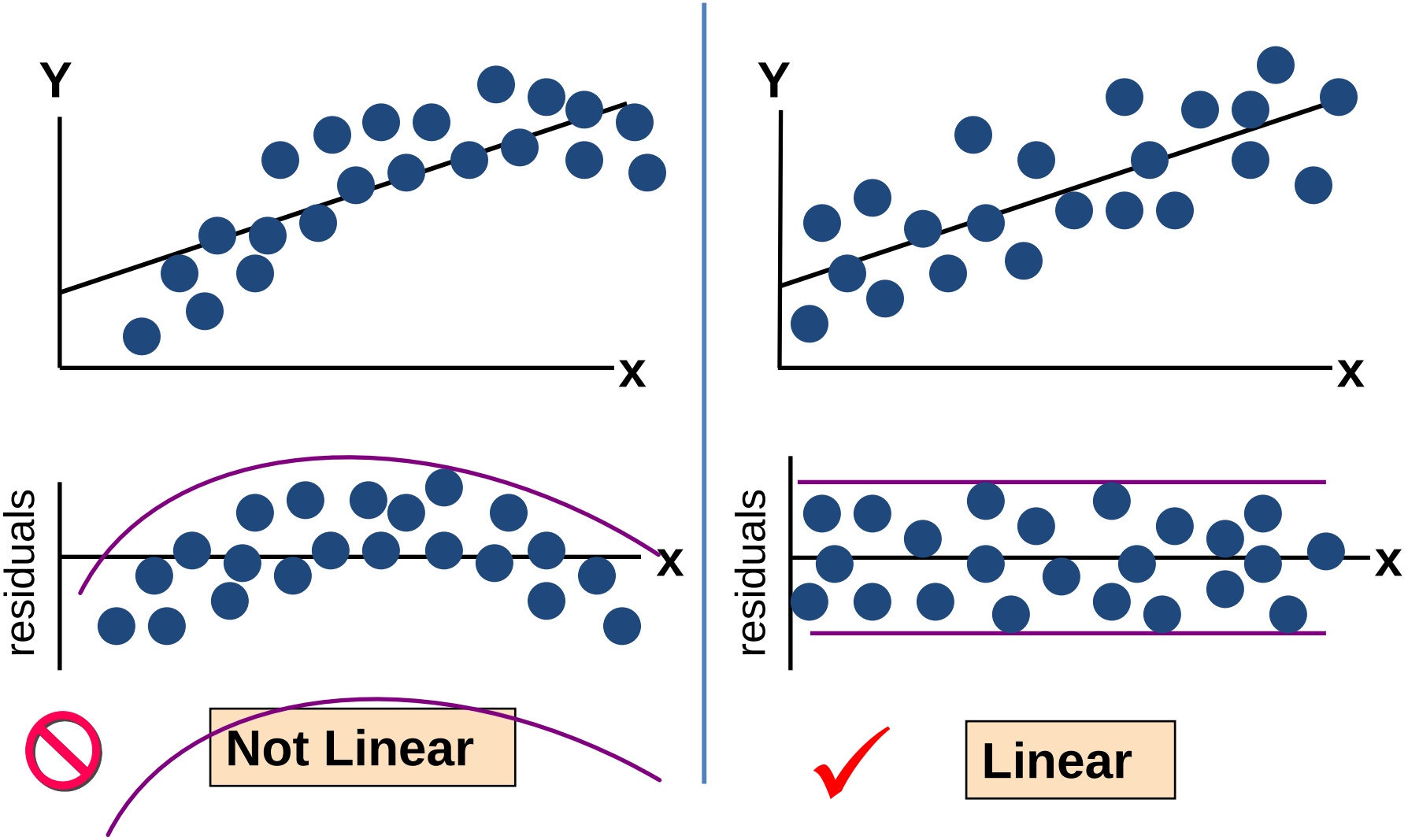


$$y_i = 48$$

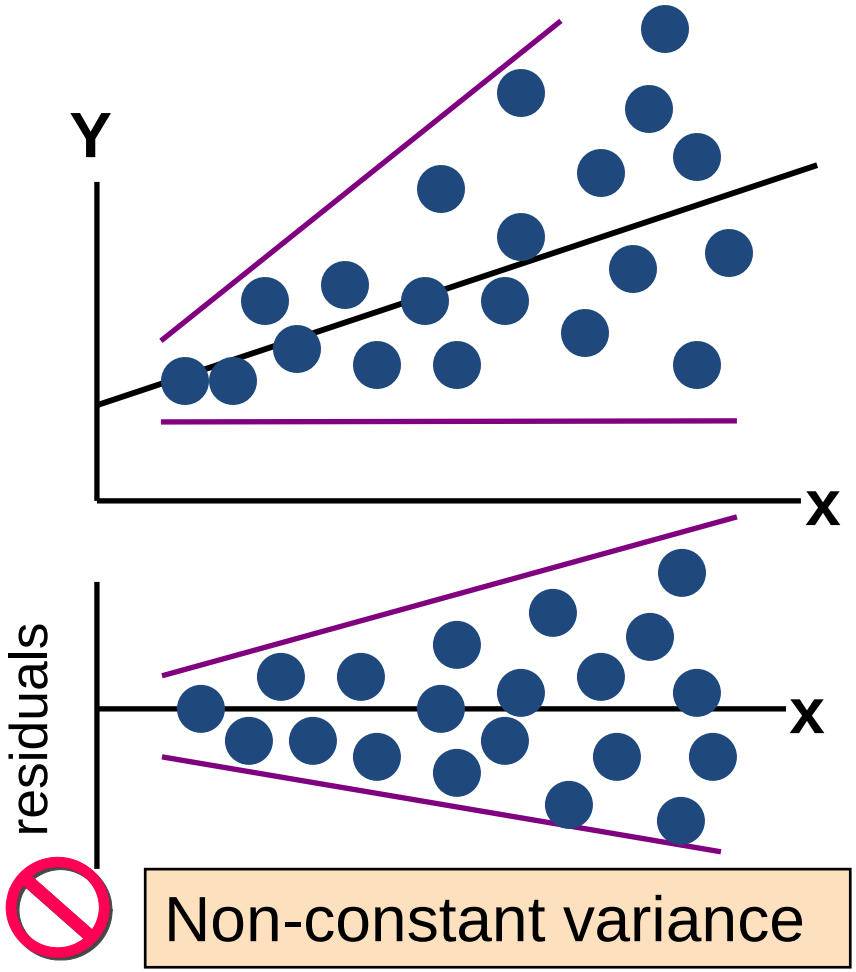
$$\hat{y}_i = 34$$

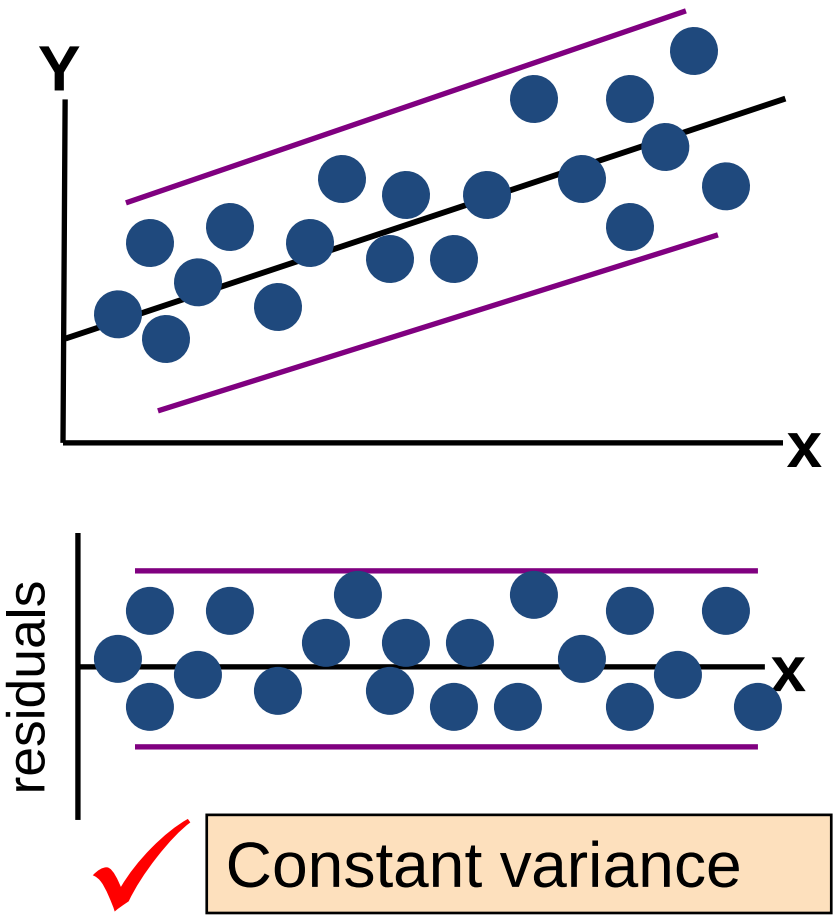
$$y_i - \hat{y}_i = 14$$

Residual Analysis for Linearity



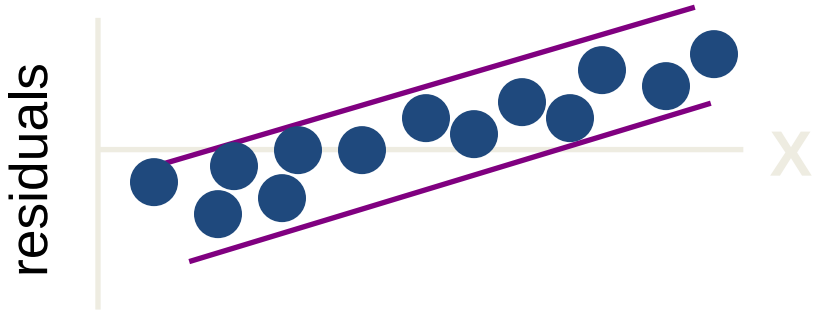
Residual Analysis for Homoscedasticity

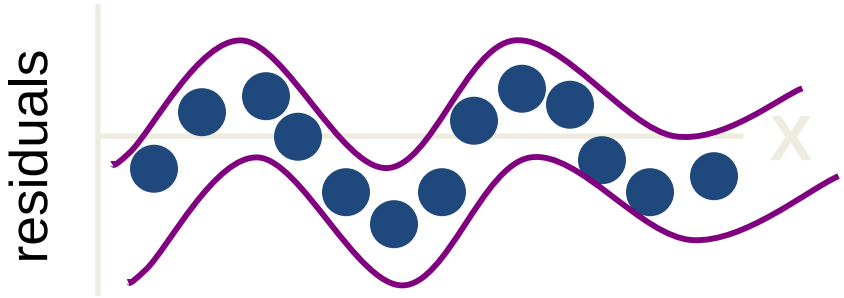




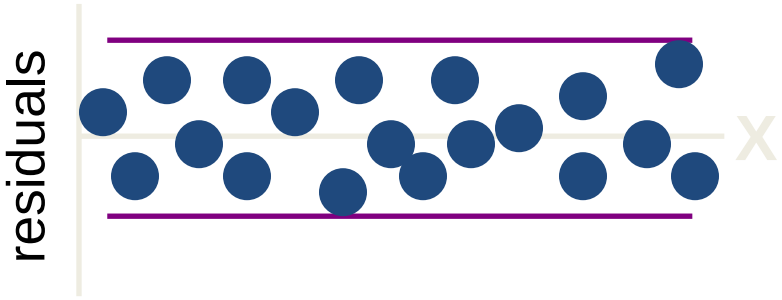
Residual Analysis for Independence

Not Independent

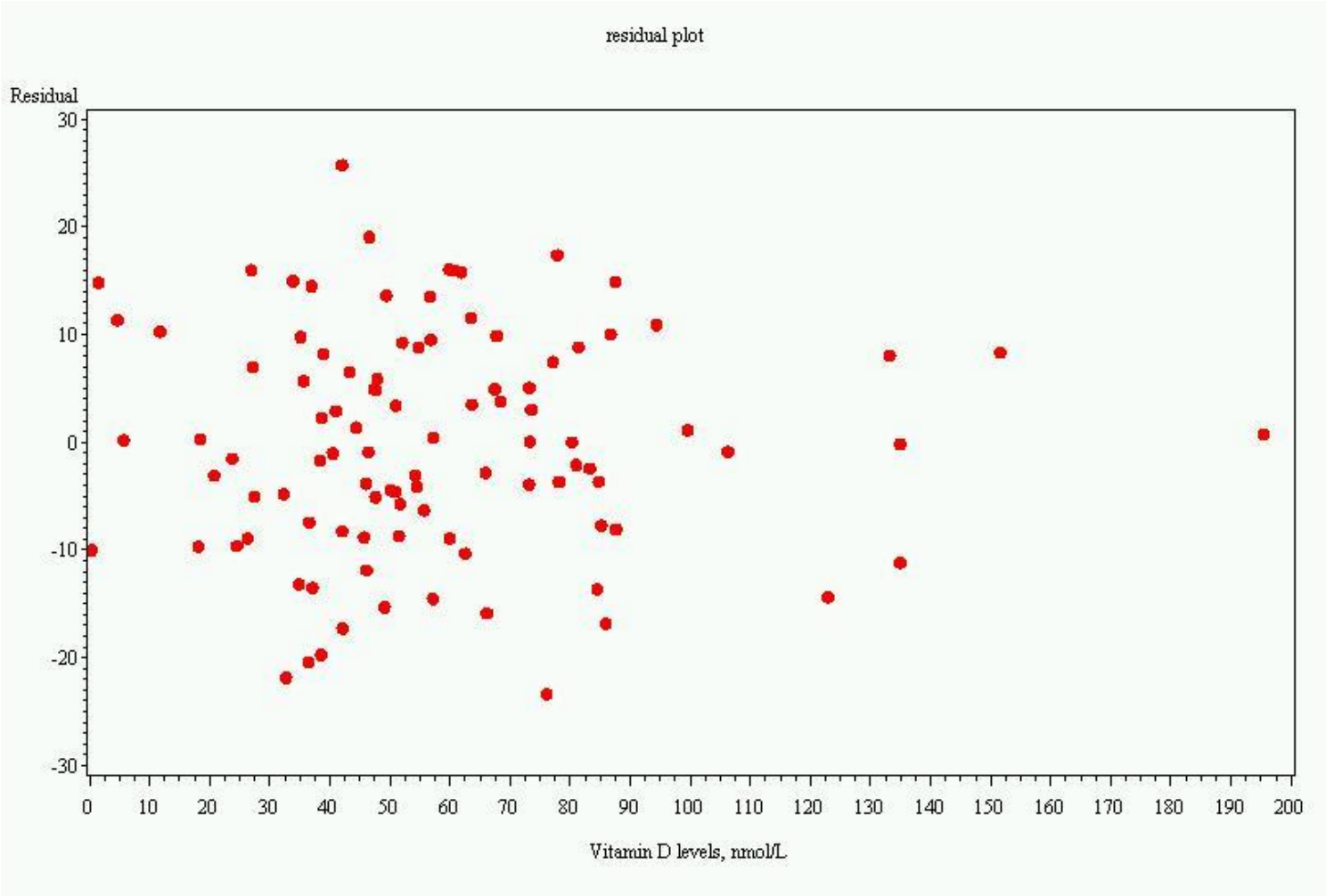




Independent



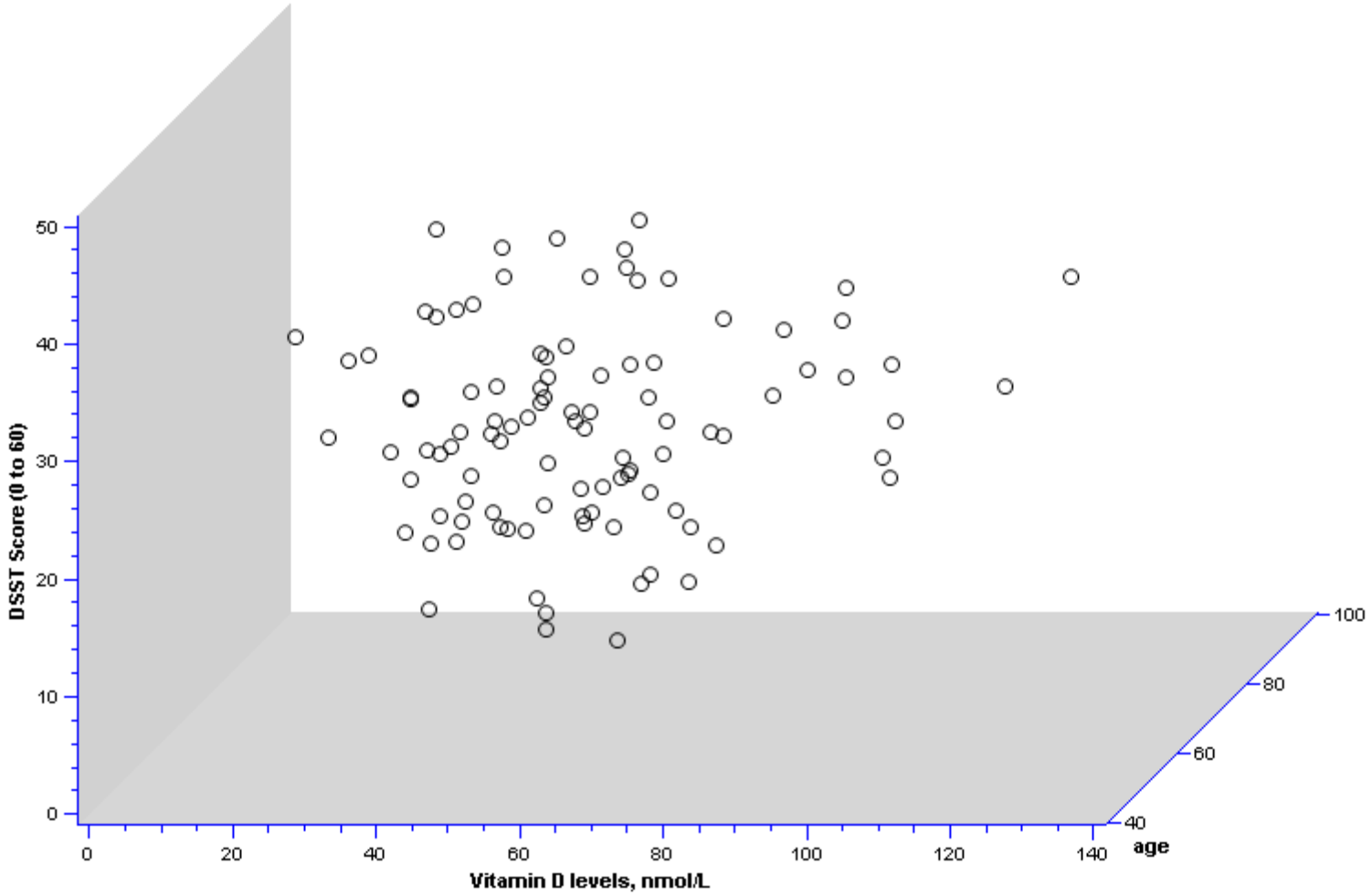
Residual plot, dataset 4



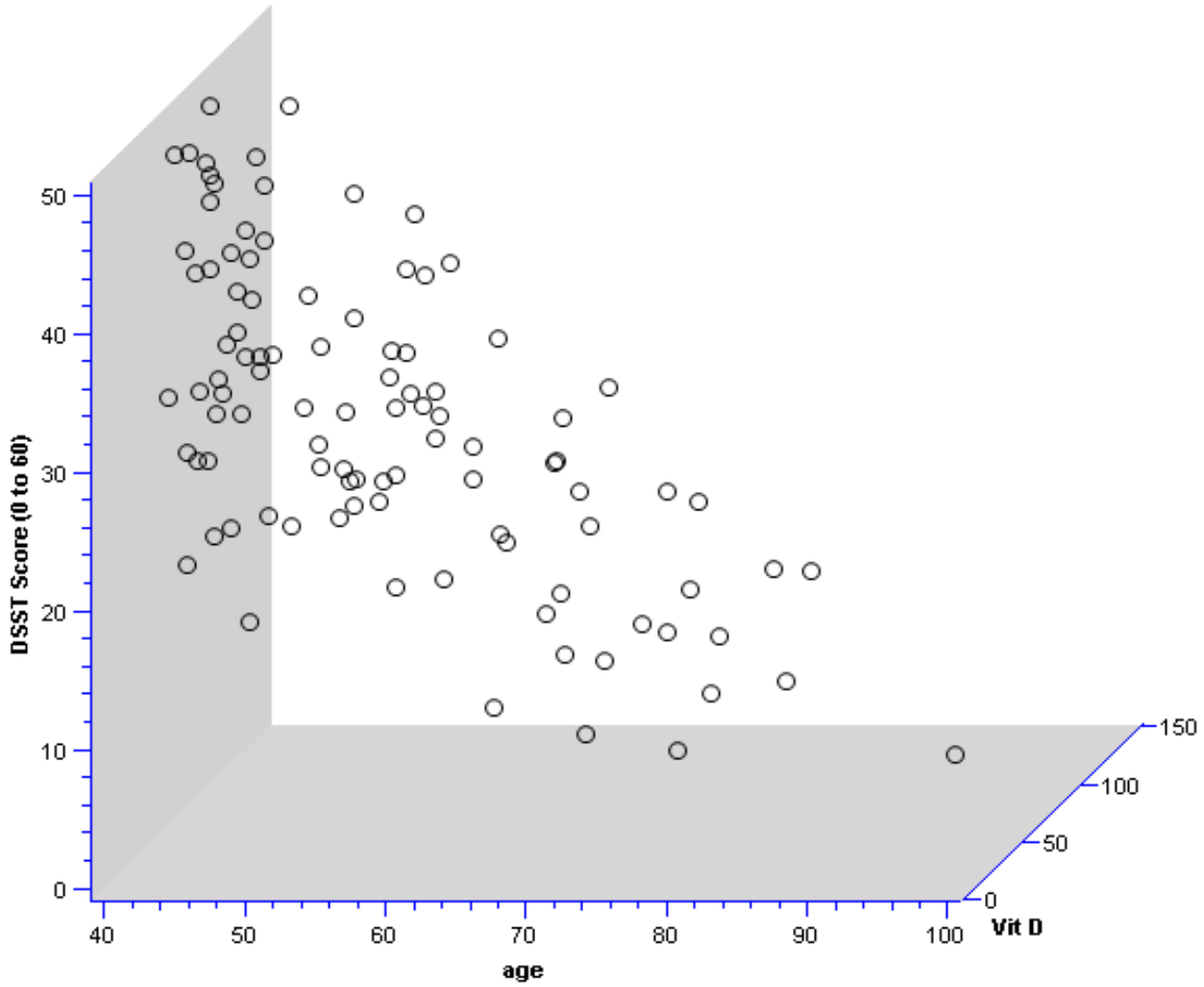
Multiple linear regression…

- What if age is a confounder here?
  - Older men have lower vitamin D
  - Older men have poorer cognition
- "Adjust" for age by putting age in the model:
  - DSST score = intercept + $slope_1 x$vitamin D + $slope_2 x$age

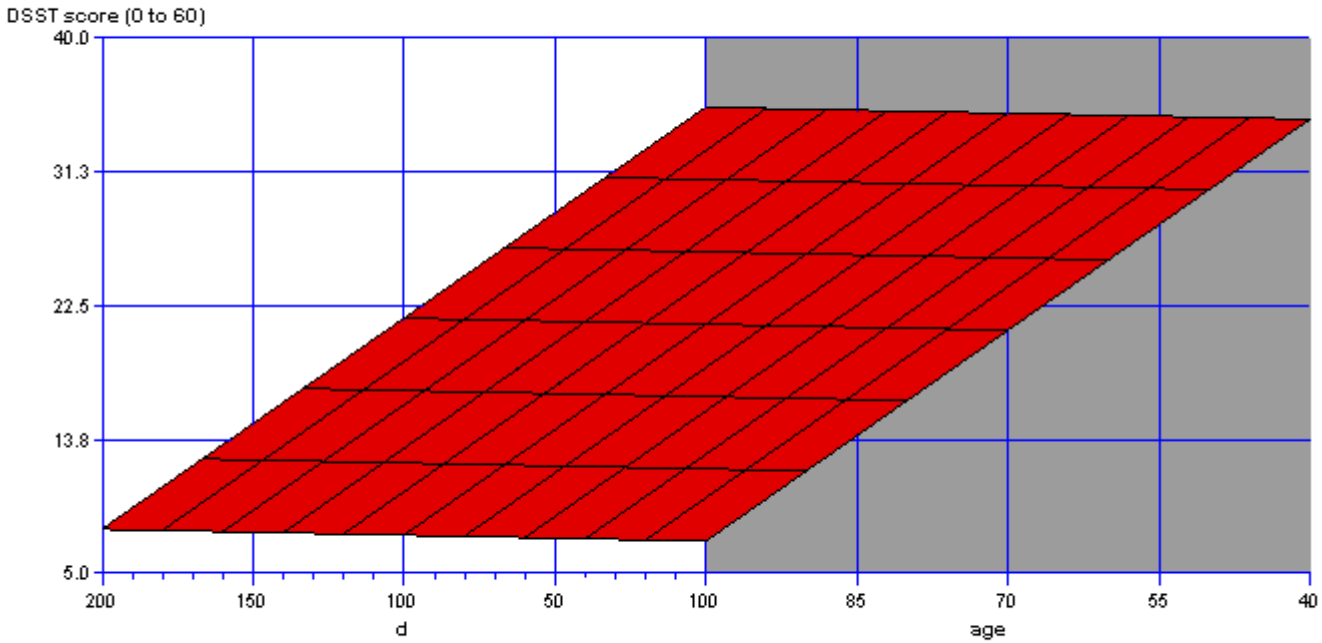
2 predictors: age and vit D…



Different 3D view…



Fit a plane rather than a line…



On the plane, the slope for vitamin D is the same at every age; thus, the slope for vitamin D represents the effect of vitamin D when age is held constant.

Equation of the “Best fit” plane…

- DSST score = 53 + 0.0039*x*vitamin D (in 10 nmol/L) - 0.46 *x*age (in years)

- P-value for vitamin D >>.05
- P-value for age <.0001

- Thus, relationship with vitamin D was due to confounding by age!

Multiple Linear Regression

- More than one predictor…

E(y)= α + β_1*X + β_2*W + β_3*Z…

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

A t-test is linear regression!

- Divide vitamin D into two groups:
  - Insufficient vitamin D (<50 nmol/L)
  - Sufficient vitamin D (>=50 nmol/L), reference group
- We can evaluate these data with a t-test or a linear regression…

$$T_{98} = \frac{40 - 32.5 = 7.5}{\sqrt{\frac{10.8^2}{54} + \frac{10.8^2}{46}}} = 3.46;\ p = .0008$$

As a linear regression…

Intercept represents the mean value in the sufficient group.

Slope represents the difference in means between the groups. Difference is significant.

Parameter Variable	Estimate	Standard Error	t Value	Pr> |t|
Intercept	40.07407	1.47511	27.17	<.0001
insuff	-7.53060	2.17493	-3.46	0.0008

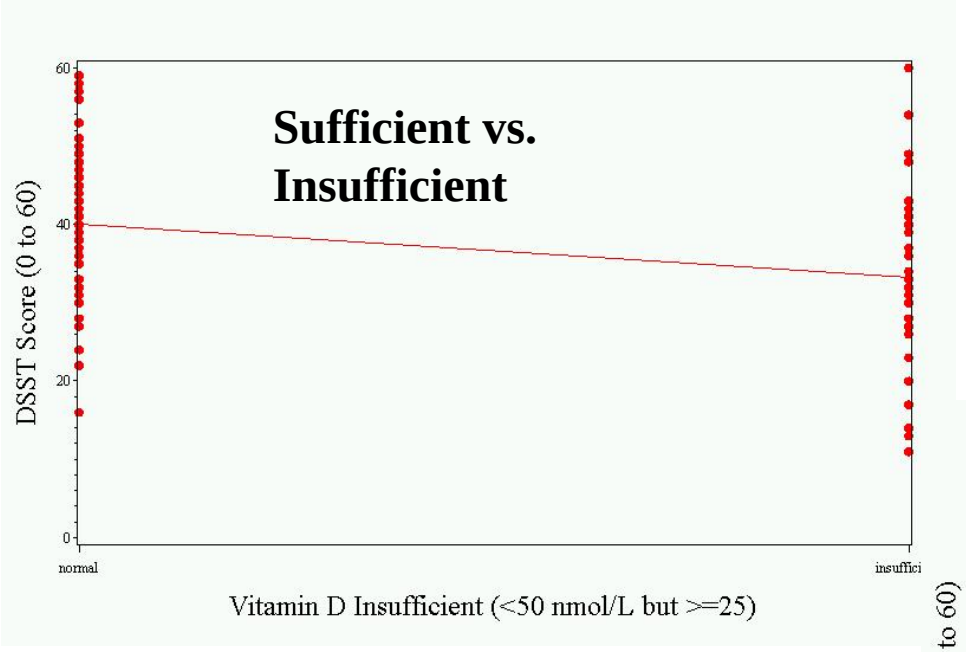
ANOVA is linear regression!

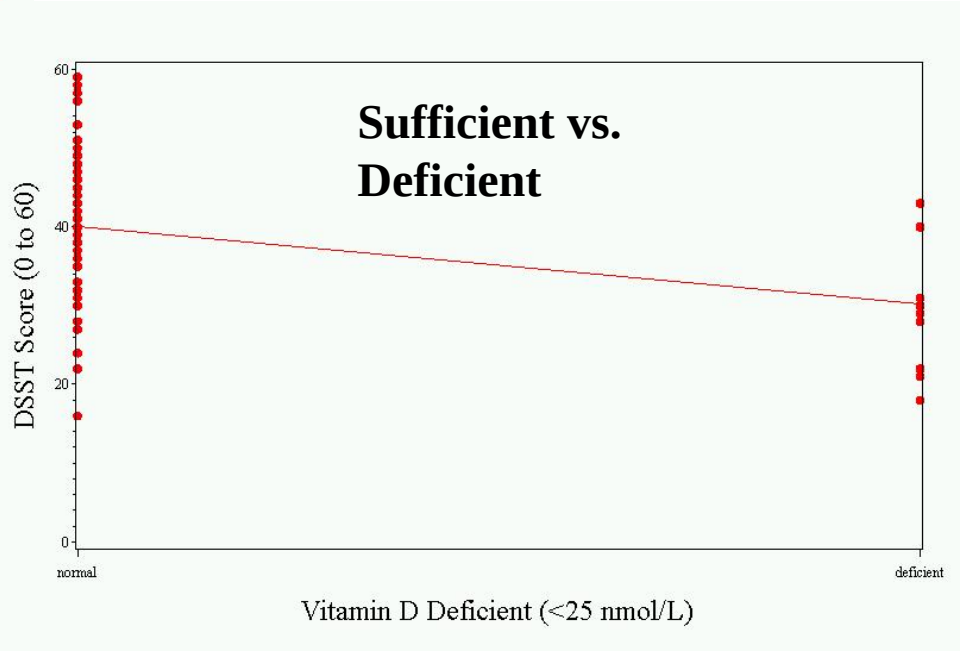
- Divide vitamin D into three groups:
  - Deficient (<25 nmol/L)
  - Insufficient (>=25 and <50 nmol/L)
  - Sufficient (>=50 nmol/L), reference group

DSST= α (=value for sufficient) + $\beta_{\text{insufficient}}$*(1 if insufficient) + β_2 *(1 if deficient)

This is called “dummy coding”—where multiple binary variables are created to represent being in each category (or not) of a categorical variable

The picture…


Sufficient vs.
Insufficient
DSST Score (0 to 60)
60
40
20
0
normal
insuffici
Vitamin D Insufficient (<50 nmol/L but >=25)


Sufficient vs.
Deficient
DSST Score (0 to 60)
60
40
20
0
normal
deficient
Vitamin D Deficient (<25 nmol/L)

Results…

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	40.07407	1.47817	27.11	<.0001
deficient	**1**	**-9.87407**	**3.73950**	**-2.64**	**0.0096**
insufficient	**1**	**-6.87963**	**2.33719**	**-2.94**	**0.0041**

- Interpretation:
  - The deficient group has a mean DSST 9.87 points lower than the reference (sufficient) group.
  - The insufficient group has a mean DSST 6.87 points lower than the reference (sufficient) group.

Multivariate regression pitfalls

- **Multi-collinearity**
- **Residual confounding**
- **Overfitting**

Multicollinearity

- **Multicollinearity** arises when two variables that measure the same thing or similar things (e.g., weight and BMI) are both included in a multiple regression model; they will, in effect, cancel each other out and generally destroy your model.

- Model building and diagnostics are tricky business!

Residual confounding

- You cannot completely wipe out confounding simply by adjusting for variables in multiple regression unless variables are measured with zero error (which is usually impossible).

Overfitting

- In multivariate modeling, you can get highly significant but meaningless results if you put too many predictors in the model.
- The model is fit perfectly to the quirks of your particular sample, but has no predictive ability in a new sample.

R-Square(결정 계수)

$$R^2 = 1 - \frac{\Sigma(\text{ 오차}^2\ )}{\Sigma(\text{ 편차}^2\ )}$$

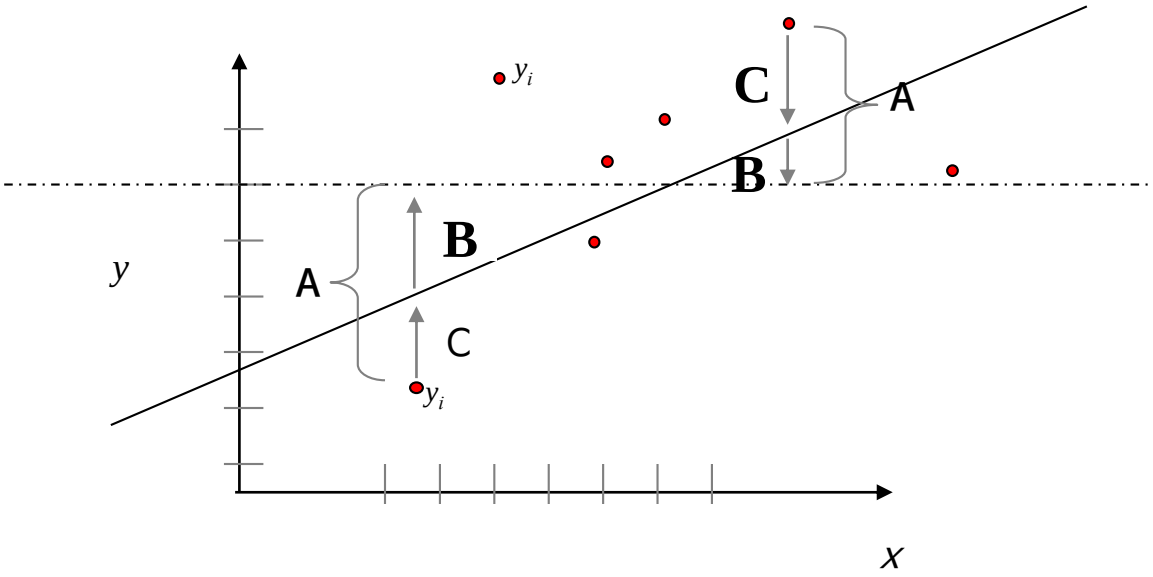
오차 = $(t_i - y_i)$, t_i : 실제값 , y_i : 예측값

편차 = $(t_i - \overline{t_i})$, t_i : 실제값 , $\overline{t_i}$: 평균값

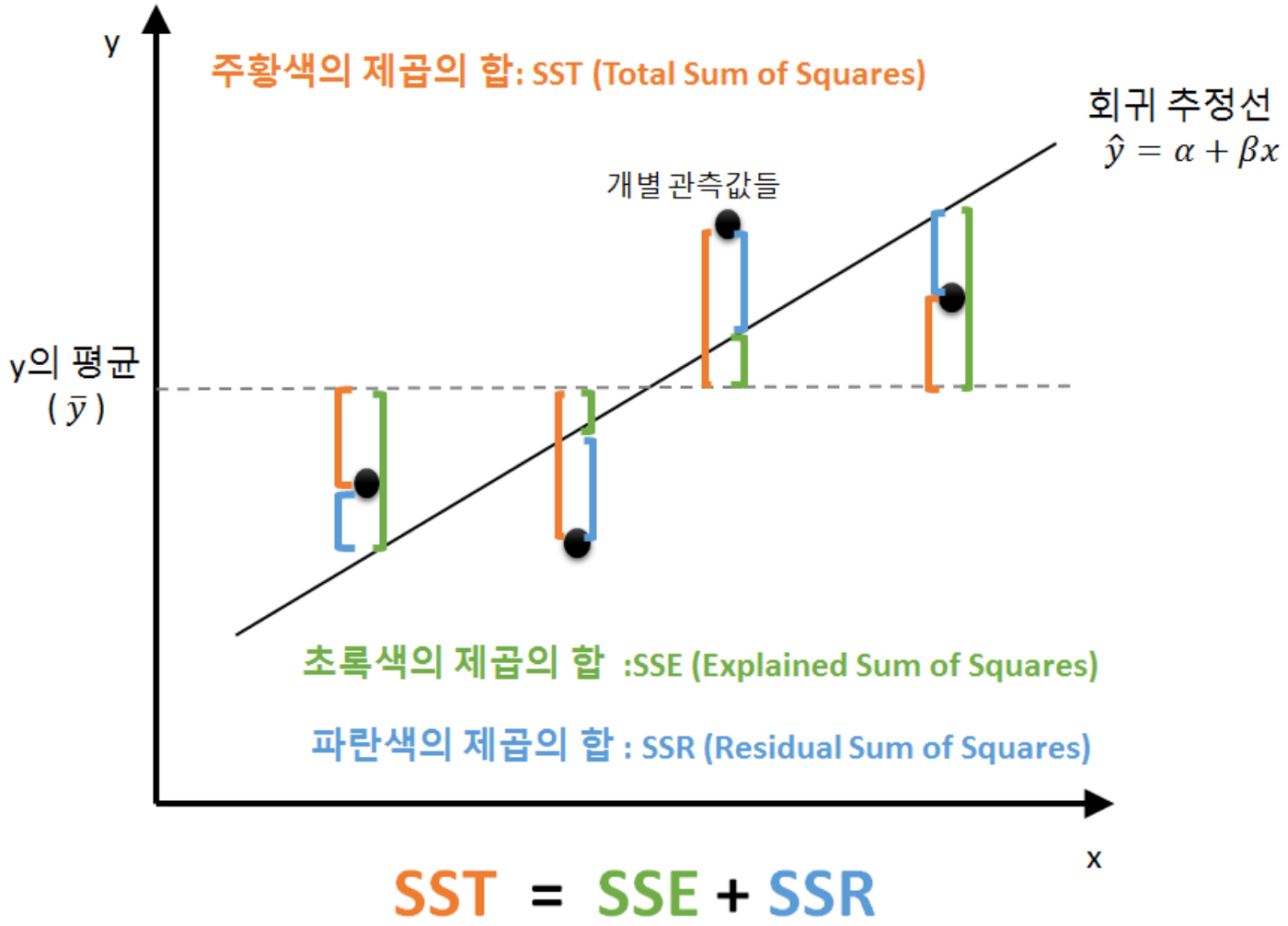
$$R^2 = \frac{\Sigma(\text{ 예측값에 대한 편차}^2\ )}{\Sigma(\text{ 편차}^2\ )}$$

예측값에 대한 편차 = $(y_i - \overline{t_i})$, y_i : 예측값 , $\overline{t_i}$: 평균값

편차 = $(t_i - \overline{t_i})$, t_i : 실제값 , $\overline{t_i}$: 평균값



R-Square(결정 계수)



R-Square(결정 계수)

총 변동 Total SS : total sum of squares (SST) : $\mathrm{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2.$: 개별 y의 편차제곱의 합

설명된 변동 Model SS :explained sum of squares (SSE) : $\mathrm{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$: 회귀식 추정 y의 편차제곱의 합

설명 안된 변동 Residual SS : residual sum of squares (SSR) : $\mathrm{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2.$: 잔차의 제곱의 합

$$\frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = r^2$$

$$R^2 \equiv \mathrm{SSE}/\mathrm{SST} = 1 - \mathrm{SSR}/\mathrm{SST}.$$

Adjusted R-Square(수정된 결정계수)

- 결정계수는 독립변수 개수가 많아질수록 그 값이 커지게 되어 종속변수의 변동을 별로 설명해 주지 못하는 변수가 모형에 추가된다고 하더라도 결정계수값이 커질 수 있음

$$\text{adjusted } R^2 = 1 - \frac{n-1}{(n-p-1)(1-R^2)}$$

n: 표본 데이터에서 자료의 개수

p: 독립 변수의 개수; 모형에서 사용한 상수를 제외한 변수의 개수

F-Statistics(F 검정 통계량)

❖ F-Test : F-검정은 두 모집단의 분산의 차이가 있는가를 검정할 때 사용

n 개의 관측치와 p 개의 독립변수에 대한 다중회귀분석

- 귀무가설(H0) : β1=β2=⋯=βp=0 즉, 모든 독립변수가 종속변수과 관계 없다.
- 대립가설(H1) : β1,β2,⋯,βp 중 적어도 하나는 00 이 아니다. 즉, 종속변수와 관계 있는 독립변수가 존재한다

$$F = \frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)}$$ 는 자유도가 $(p, n-p-1)$ 인 F분포를 따른다.

n 개의 관측치와 p 개의 독립변수에 대한 다중회귀분석에서 전체모형 FM이라고 하고, FM에서 k 개의 독립변수가 제거된 모형을 축소모형 RM

- 귀무가설(H0) : RM 이면 충분
- 대립가설(H1) : FM이 더 좋은 모델

$$F = \frac{[\mathrm{SSR}(RM) - \mathrm{SSR}(FM)]/(p+1-k)}{\mathrm{SSE}(FM)/(n-p-1)}$$ 는 자유도가 $(p+1-k, n-p-1)$ 인 F분포를 따른다.

모델 간 비교에 사용 : 첫번째 F <= 두번째 F, RM 사용 가능

T-Statistics(T 검정 통계량)

n 개의 관측치와 p 개의 독립변수에 대한 다중회귀분석

- 귀무가설(H0) : β1=0 즉, 독립변수와 종속변수는 관계 없다.
- 대립가설(H1) : β1!=0 즉, 독립변수는 종속변수에 영향을 준다(독립변수의 회귀계수가 유의하다)

회귀계수의 추정치 $\hat{\beta}_i$ 와 표준오차 $\mathrm{se}(\hat{\beta}_i)$ 에 대해 $t_i = \dfrac{\hat{\beta}_i}{\mathrm{se}(\hat{\beta}_i)}$ 는 자유도가 $n-p-1$ 인 t분포를 따른다.

P-Value < 0.05, 독립변수는 종속변수에 영향을 준다(독립변수의 회귀계수는 유의하다)

Logistic Regression

분석하고자 하는 대상들이 두 집단 혹은 그 이상의 집단으로 나누어진 경우에 개별 관측치들이 어느 집단에 분류될 수 있는가를 분석하고 이를 예측하는 모형을 개발하는데 사용되는 통계기법

Linear regression vs. Logistic regression

	일반선형 회귀분석	로지스틱 회귀분석
종속변수	연속형 변수	이산형 변수
모형 탐색 방법	최소자승법	최대우도법, 가중최소자승법
모형 검정	F-test, t-test	χ^2 test

로지스틱 회귀분석 과정

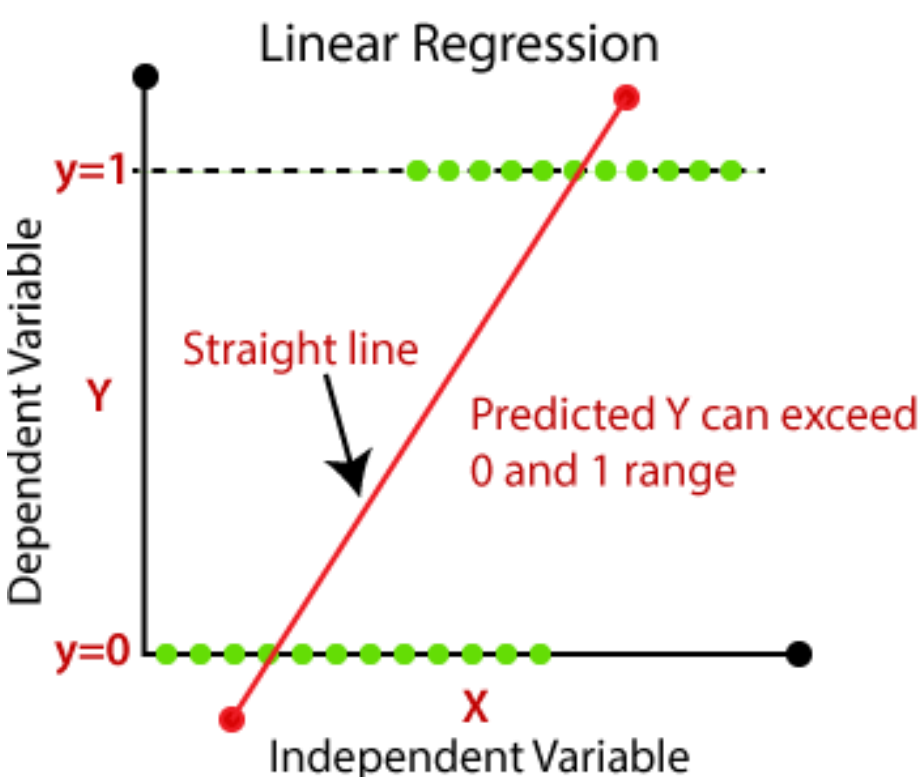
1단계: 각 집단에 속하는 확률의 추정치를 예측. 이진분류의 경우 집단 1에 속하는 확률 P(Y=1)의 추정치로 얻음.

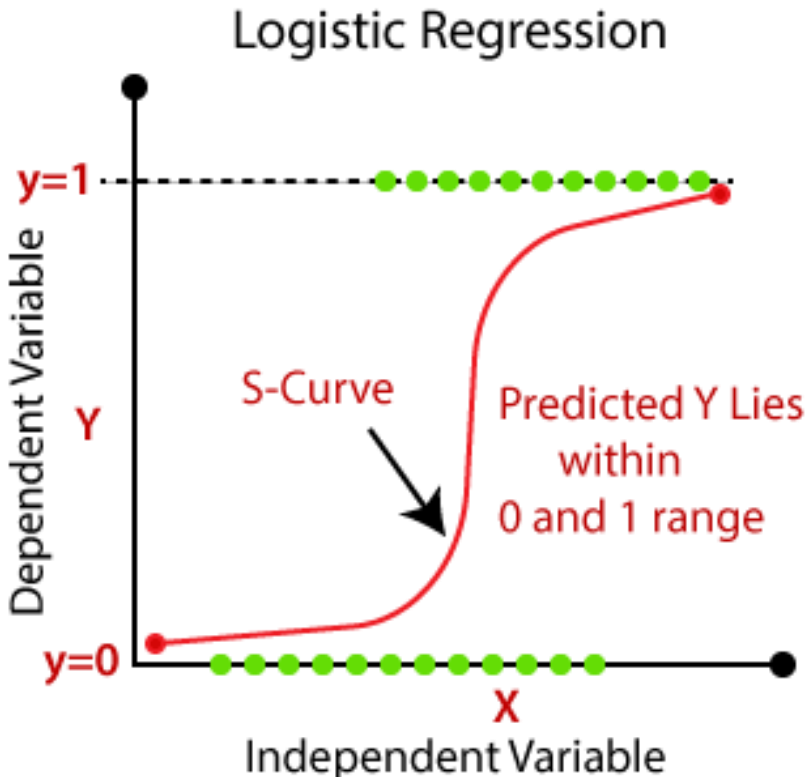
2단계: 추정확률 ➔ 분류기준값(cut-off) 적용 ➔ 특정범주로 분류

예) $P(Y=1) \geq 0.5$ ➔ 집단 1로 분류

$P(Y=1) < 0.5$ ➔ 집단 0으로 분류

Logistic Regression





Logistic Regression

- 로짓(logit)함수: log(odds)
  - 종속변수로 Y를 사용하는 대신에 로짓함수를 사용
  - 집단 1에 속하는 확률인 p를 구한다
  - P는 [0,1]사이의 값을 갖는다
  - 그러나 만약 p를 다음의 식과 같이 q개의 예측변수들의 선형함수로 표현한다면 우변이 0과 1사이의 값을 갖는다는 것을 보장할 수 없게 됨

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

  - 따라서 다음과 같은 비선형 함수를 이용

$$p = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q\right)}}$$

Logistic Regression

- 로지스틱 반응함수(logistic response function)

$$p = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q\right)}}$$

  - x1, x2, … , xq가 어떤 값을 갖는다 하더라도 우변은 항상 0과 1의 값을 갖는다

- odds(승산비)

$$odds = \frac{p}{1 - p}$$

  - 집단 1(Y=1)에 속하는 승산은 집단 1에 속하는 확률에서 집단 0에 속하는 확률을 나눈 비율

Logistic Regression

- odds
  - 사건의 odds가 주어졌을 때 사건의 확률을 계산할 수 있음

$$odds = \frac{p}{1-p} \quad \Rightarrow \quad odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q}$$

- 로짓(logit): log(odds)

$$\log\left(odds\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q$$

  - 로지스틱 회귀분석이란 로짓을 종속변수로 정의하고, 이 로짓과 q개의 예측변수와의 관계를 선형으로 모형화한 것을 말함

Maximum Likelihood Estimation(최대우도법)

- 어떤 확률변수에서 샘플링한 값들을 토대로 그 확률변수의 모수를 구하는 방법
- 즉, 어떤 모수가 주어졌을 때, 원하는 값들이 나올 가능도를 최대로 만드는 모수를 선택하는 방법

- 입력값 X와 모델의 파라메터 θ가 주어졌을 때 정답 Y가 나타날 확률을 최대화하는 θ를 찾는 것입니다.
  우리가 가지고 있는 데이터가 학습 과정에서 바뀌는 것은 아니므로 X와 Y는 고정된 상태입니다.
- 모델에 X를 넣었을 때 실제 Y에 가장 가깝게 반환하는 θ를 찾아내는 것

$$\theta_{ML} = arg\max_{\theta} P_{model}\left(Y|X;\theta\right)$$

$$= arg\max_{\theta} \sum_{i=1}^{m} \log P_{model}\left(y_i|x_i;\theta\right)$$

$$\sum_{i=1}^{m} \log P_{model}\left(y_i|x_i;\theta\right) = -m\log\sigma - \frac{m}{2}\log 2\pi - \sum_{i=1}^{m} \frac{\left\|\hat{y}_i - y_i\right\|^2}{2\sigma^2}$$

AIC(Akaike Information Criterion), BIC(Bayesian Information Criterion)

$$\mathrm{AIC}_p := n \ln\left(\frac{\mathrm{SSE}_p}{n}\right) + 2(p+1) \qquad \mathrm{BIC}_p := n \ln\left(\frac{\mathrm{SSE}_p}{n}\right) + (p+1)\ln n\,.$$

- p : 독립변수 개수
- 두 모형을 비교했을 때 더 작은 쪽이 더 좋다고 판단
- 변수가 많아질수록 패널티 부여

Regularization

- Norm : 두 벡터 간의 거리

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

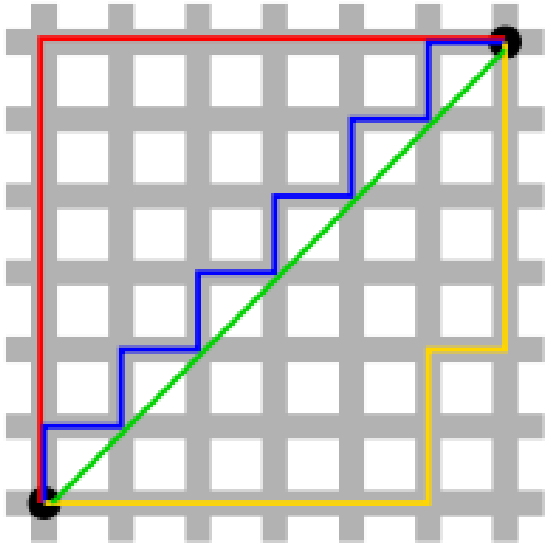
* 여기서 p 는 Norm 의 차수.
* p = 1 이면 L1 Norm 이고, P = 2 이면 L2 Norm
* n은 해당 벡터의 원소 수

- L1 Norm : 벡터 p, q 의 각 원소들의 차이의 절대값의 합

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|, \text{ where } (\mathbf{p}, \mathbf{q}) \text{ are vectors } \mathbf{p} = (p_1, p_2, \ldots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \ldots, q_n)$$

- L2 Norm : p, q 의 직선 거리

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}.$$



Regularization

- L1 Loss : 오차 절대값의 합, Least absolute deviations(LAD), Least absolute Errors(LAE), Least absolute value(LAV), Least absolute residual(LAR), Sum of absolute deviations

$$L = \sum_{i=1}^{n} |y_i - f(x_i)|$$

* y : 실제 값
* f(x) : 예측 값
* n : 데이터 개수

- L2 Loss : 오차 제곱 합, Least squares error(LSE)

$$L = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- L2 Loss 는 직관적으로 오차의 제곱을 더하기 때문에 Outlier 에 더 큰 영향을 받음
- Outlier 가 적당히 무시되길 원한다면 L1 Loss 를 사용하고, Outlier 의 등장에 신경 써야 하는 경우라면 L2 Loss 를 사용
- L1 Loss 는 0인 지점에서 미분이 불가능하다는 단점 또한 가지고 있습니다.

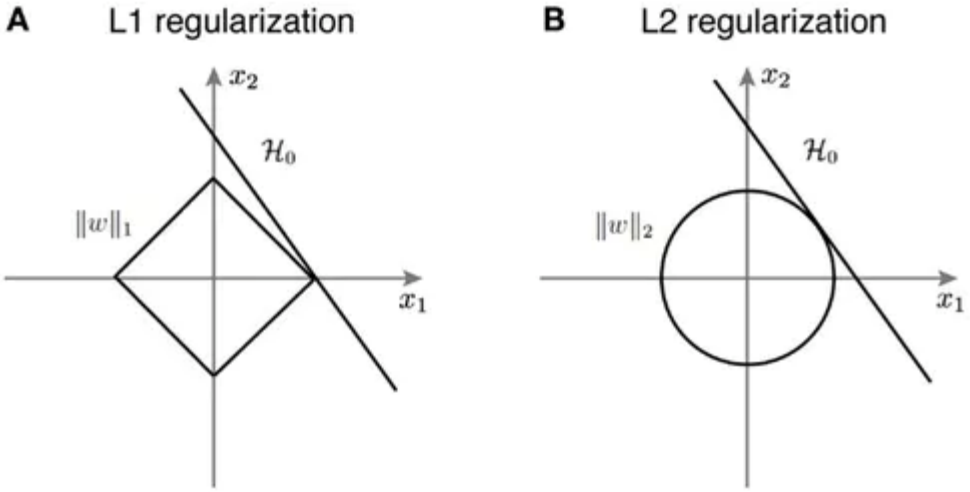
Regularization

- L1 Regularization : 기존 cost function에 가중치(weight)의 절대값을 포함하여 가중치가 너무 크지 않은 방향으로 학습되도록 함

$$Cost = \frac{1}{n}\sum_{i=1}^{n}\{L(y_i, \hat{y}_i) + \frac{\lambda}{2}|w|\}$$

* λ : learning rate

- L2 Regularization : $$Cost = \frac{1}{n}\sum_{i=1}^{n}\{L(y_i, \hat{y}_i) + \frac{\lambda}{2}|w|^2\}$$



* L1 Regularization은 특정 Feature 를 0으로 처리하는 것이 가능하므로 Feature selection 이 가능

Ridge, Lasso, Elastic Net Regression

- Ridge Regression : L2 Regularization 적용
  - 다중 공선성 방지
  - 계수 축소를 통한 모델 복잡도 감소
- Lasso Regression : L1 Regularization 적용
  - 변수 선택 가능, 많은 변수를 가진 데이터에 사용
- Elastic Net Regression : L1, L2 적용