

# 선형 예측 모델링

회귀분석

Multiple Linear Regression

# 보르도 와인



1년6개월



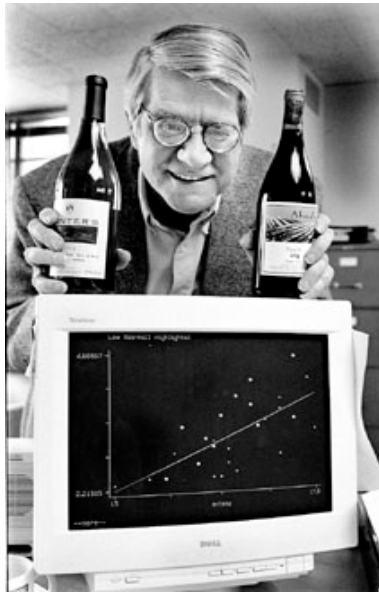
Robert Parker



6개월



# 와인의 품질은?



Prof Orley Ashenfelter



Bordeaux farmer

와인의 품질  
= 포도의 품질  
= 온도, 햇빛, 강수량...

# 와인의 품질 예측 가능?

## Prediction 으로 품질 예측 시도

X: 1952~1980 사이의 30년간의 보르도 지방 기후 데이터 (월별온도, 햇빛, 강수량)

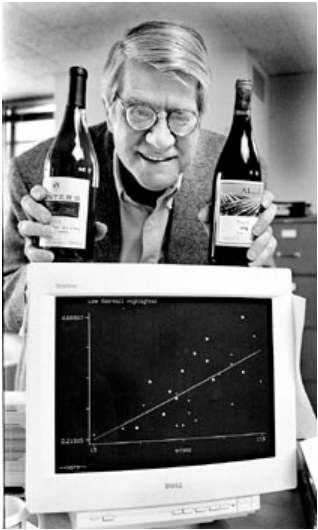
VS

Y: 당해 보르도 와인 평균 가격 (품질)

혹시,  $Y = f(X)$  인  $f$  를 찾을 수 있을까?

찾을 수 있다면,  $x$  로  $y$  예측 가능!!!

# 와인의 품질 예측 가능!!!

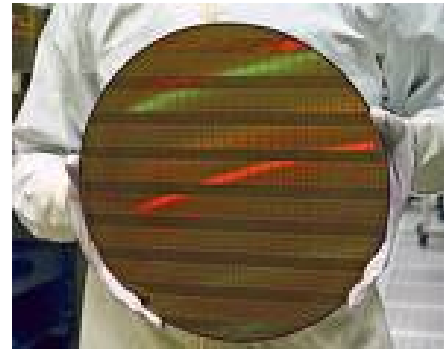


$$\begin{aligned} \text{품질} &= 12.145 \\ &+ 0.00117 * \text{전년도 겨울 강수량} \\ &+ 0.06140 * \text{당해 년도 평균 기온} \\ &- 0.00386 * \text{수확기 강수량} \end{aligned}$$

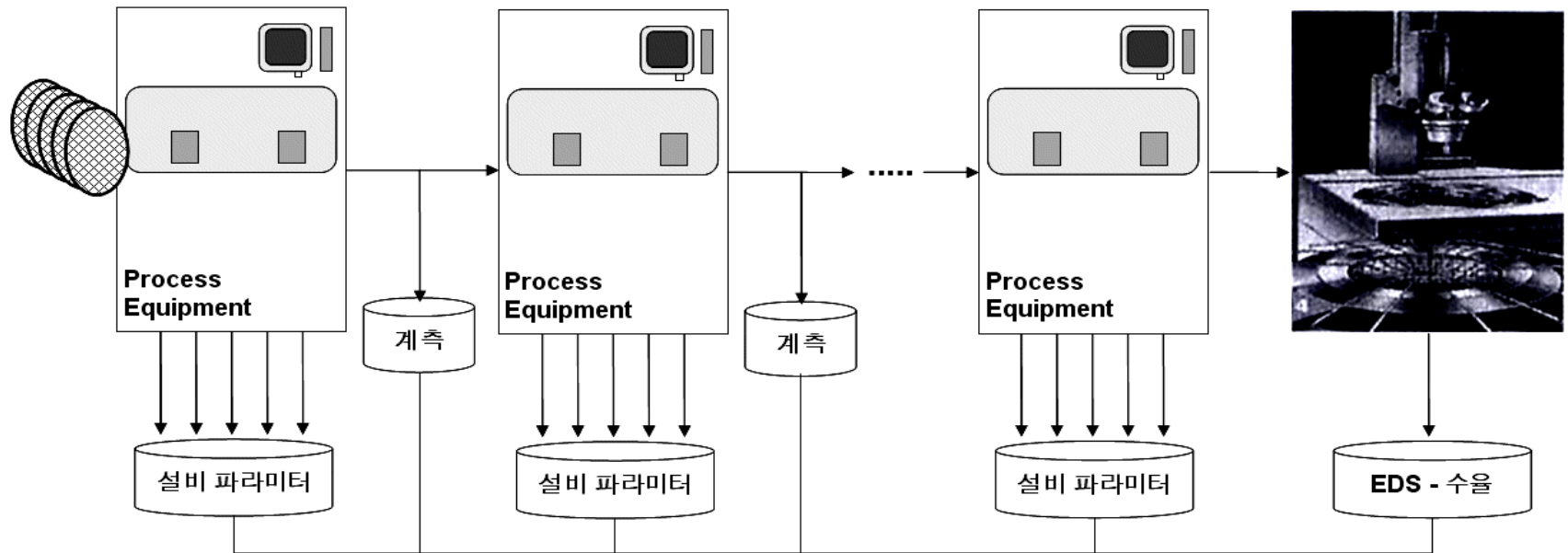
품질 및 가격 예측 가능

또한, 예측 시점이 매년 가을 수확 직 후로 전문가 테이스팅 6개월 전으로 앞 당겨짐 (와인 선물)

# 반도체 공정 품질 관리

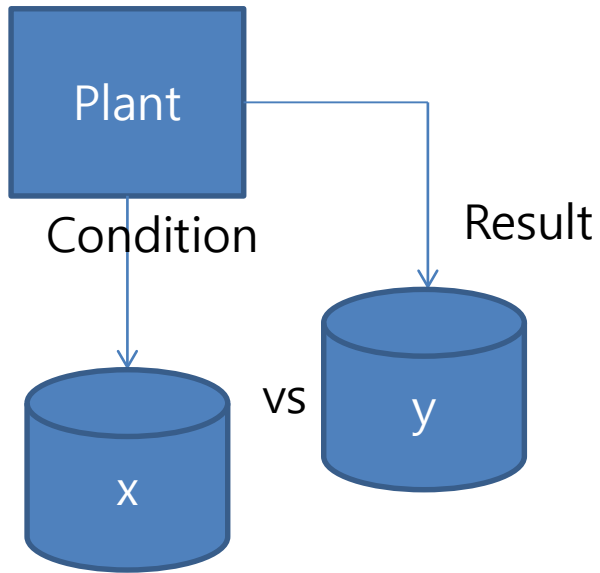


# 반도체 공정 품질 관리



- 한 Lot에 포함된 25개의 웨이퍼들 중 하나의 웨이퍼에 대해서만 계측 정보를 추출하기 때문에 계측 정보의 활용도가 떨어짐.
- 모든 웨이퍼에 대한 계측 정보 수집은 비효율적이며 고 비용 발생.

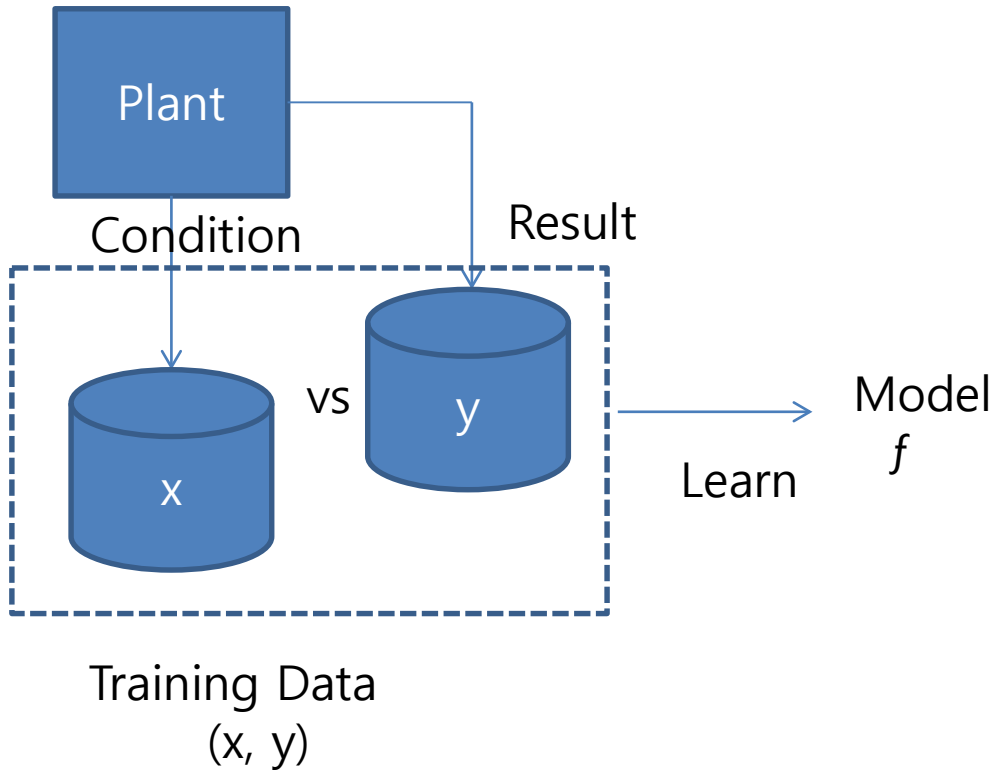
# 교사 학습 Supervised Learning



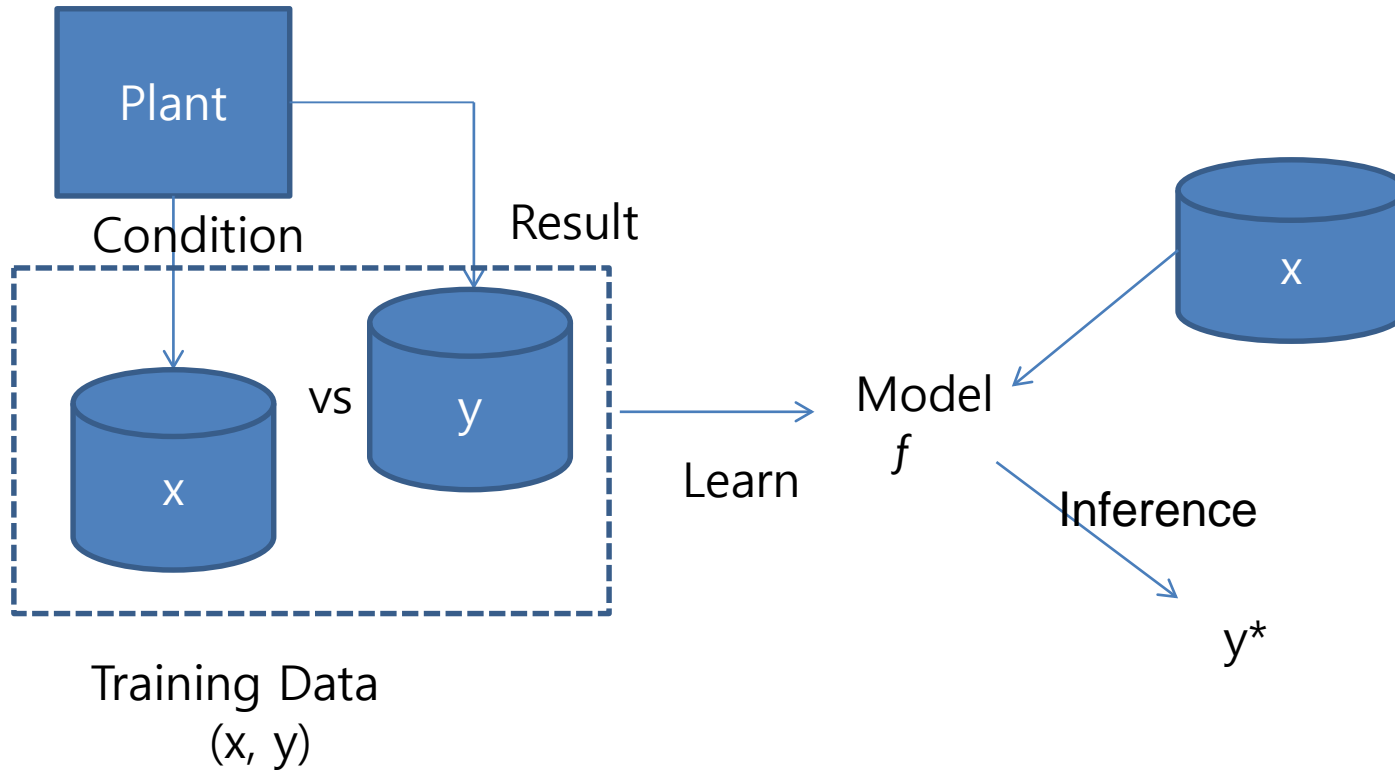
Training Data  
( $x, y$ )



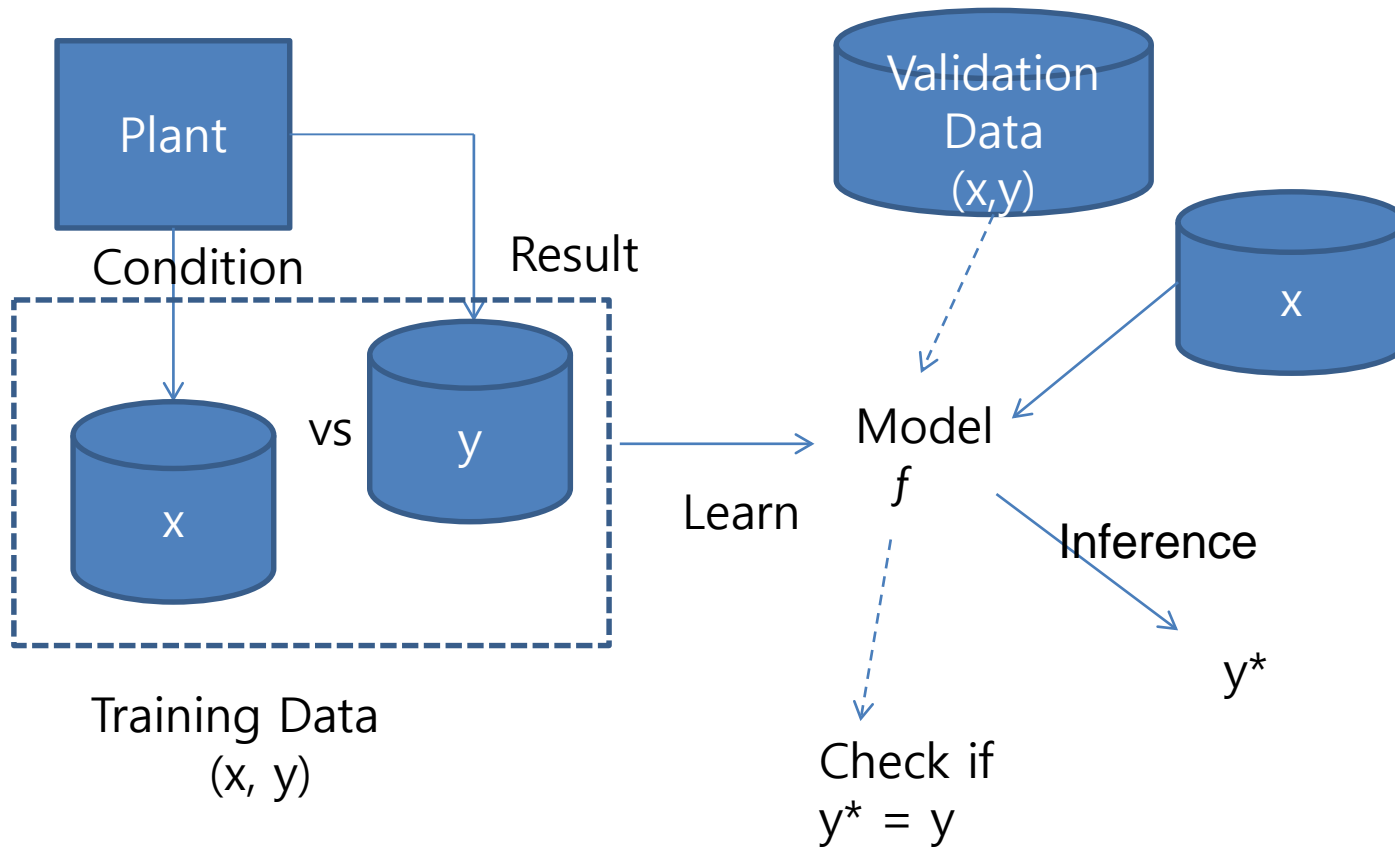
# 교사 학습 패러다임



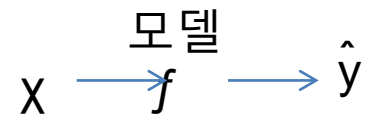
# 교사 학습 패러다임



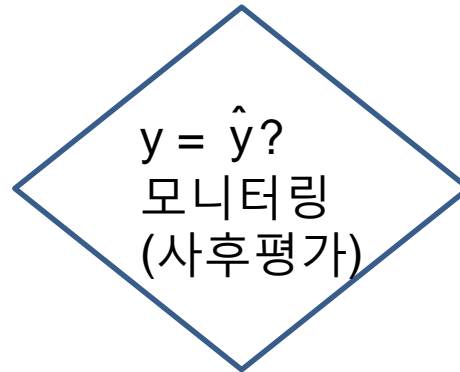
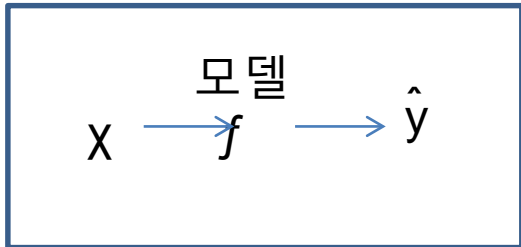
# 교사 학습 패러다임



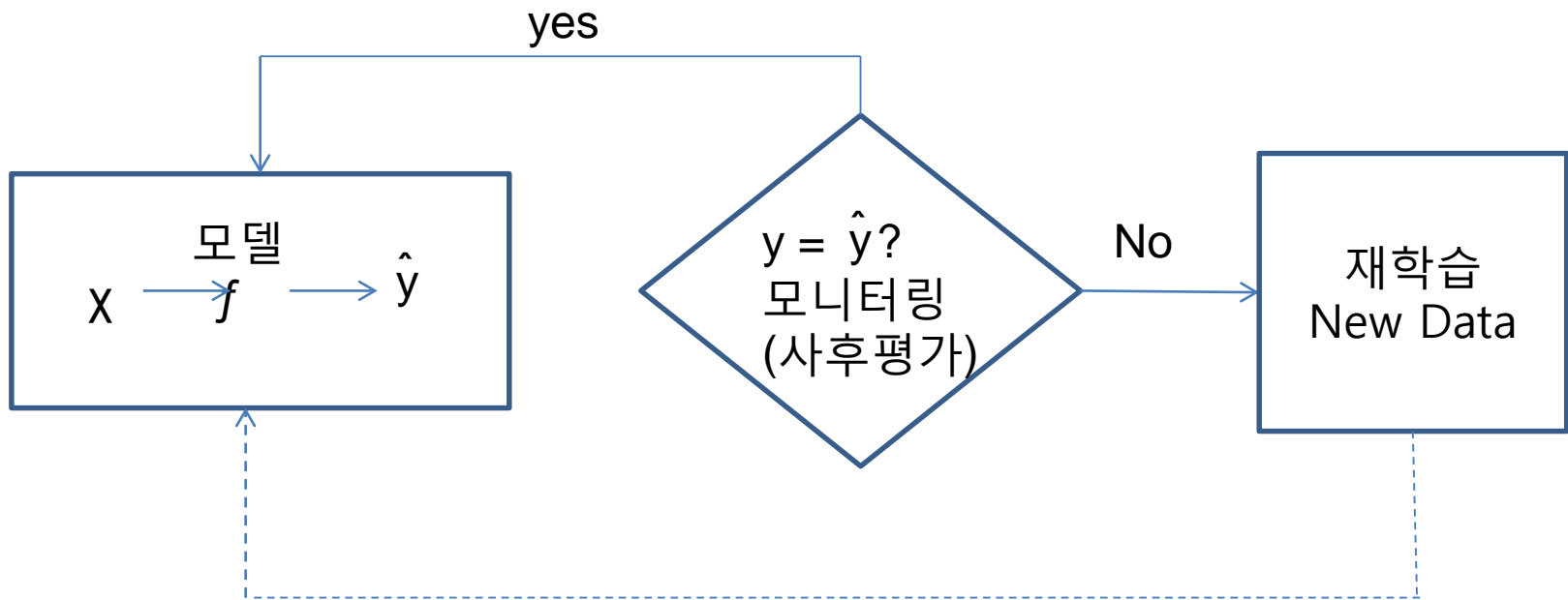
# 모델의 현업 배치, 모니터링 및 재학습



# 모델의 현업 배치, 모니터링 및 재학습



# 모델의 현업 배치, 모니터링 및 재학습



# Prediction

- 종속변수( $y$ )를 독립 변수 ( $x$ ) 들의 함수 ( $f$ ) 로 적합,
- 즉 데이터  $\{(x,y)\}$  로 부터  $y = f(x)$  의  $f$  를 찾는다
- **회귀분석**, 신경회로망, 사례기반추론, **의사결정나무**
- 예:  $y$  무엇을 예측할 수 있는가?
  - 소비자가 마케팅 캠페인에 반응할 확률
  - 휴대폰 고객이 향후 6개월 내에 이탈할 확률
  - 와인의 품질
  - 반도체 웨이퍼의 수율
  - 선박 건조 기간

# Prediction

- Y가 결정이 된 후에는...
- 무엇으로  $y$ 를 예측하려고 하는가? 즉,  $x$  ???
- X는 독립 변수 또는 “예측 변수 predictive variable”
  
- X 선택 기준
  - Y와의 정확한 함수 관계를 알고 있다.
  - Y와의 정확한 함수 관계는 모르지만, 영향을 준다는 걸 100% 확실
  - Y와의 정확한 함수 관계는 모르지만, 영향을 줄 수 있는 가능성이 있다.



# Predictive Analytics

- A new fancy name of “Supervised” data mining, or Regression and Classification
- “Find a mapping/function  $f$  such that  $y = f(x)$  given data set  $D = \{(x,y)\}$ ”
- Regression when  $y$  is continuous
- Classification when  $y$  is categorical/binary

# Predictive Analytics

- Regression
  - Multiple Linear Regression
  - k-NN
  - Decision Tree Regression
  - Neural Networks
- Classification
  - Logistic Linear Regression, Discriminant Analysis
  - k-NN, Naïve Bayese
  - Decision Tree Classifier
  - Neural Networks
  - SVM

# Topics

- Explanatory vs. predictive modeling with regression
- Example: prices of Toyota Corollas
- Fitting a predictive model
- Assessing predictive accuracy
- Selecting a subset of predictors

# “선형” 회귀분석 vs 비선형

- $y = f(x)$  에서  $f$  는 계수들의 선형식
- 예: 품질 = 12.145  
+ 0.00117 \* 전년도 겨울 강수량  
+ 0.06140 \* 당해 년도 평균 기온  
- 0.00386 \* 수확기 강수량

Q)  $x$  대신  $x^2$ ,  $x^3$ ,  $\log x$  등 사용 가능?

Q) 선형식이 아닌 다른  $f$  는? 비선형?

# 데이터로부터 모델의 계수 추정

- 수식  $f$ 의 계수 값을 데이터 세트  $D = \{(\text{기후 변수 값들}, \text{품질 값})\}$  이용하여 찾는 과정
  - 12.145, 0.00117, 0.0614, -0.00386
- 품질 = 12.145
  - + 0.00117 \* 전년도 겨울 강수량
  - + 0.06140 \* 당해 년도 평균 기온
  - 0.00386 \* 수확기 강수량
- 어떻게 찾아내는가?
  - Ordinary Least square => matrix inversion

# 설명모델

**목표:** 예측변수들(설명적 변수들)과 타겟 사이의 관계 설명

- 데이터 분석에서 회귀분석을 사용하는 데 많이 쓰임
- 모델 목표: 데이터를 잘 적합하고 모델에 대한 설명적 변수들의 기여를 이해
- “적합도 검증”:  $R^2$ , 잔차 분석, p-values

# 예측 모델

**목표:** 예측변수 값은 있지만, 타깃 값은 없는 경우  
다른 데이터에서 타깃 값을 예측

- 전통적 데이터 마이닝 맥락
- 모델 목표: 예측 정확성 최적화
- 학습 데이터에서 학습 모델
- 검증(홀드아웃) 데이터에서 성능 평가
- 예측변수의 설명 역할을 주요한 목적이  
아님(유용하긴 함)

예: 도요타 코롤라 중고차 가격

ToyotaCorolla.xls

**목표:** 사양에 따라 도요타 코롤라 중고차의  
가격 예측

**데이터:** 사양 정보에 따른, 도요타 코롤라  
중고차 1,442대의 가격



# 데이터 샘플

(분석에 사용되는 변수만 나옴)

Price	Age	KM	Fuel_Type	HP	Metallic	Automatic	cc	Doors	Quarterly_Tax	Weight
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185

# 사용된 변수들

판매가격 (Euros)

내용연수 (2004년 8월 현재)

연비 (kilometers)

연료유형 (diesel, petrol, CNG)

마력 (horsepower)

금속색상 (1=yes, 0=no)

자동변속 (1=yes, 0=no)

실린더 부피 (cylinder volume)

자동차 문의 개수

분기별 도로 사용세(road tax)

무게 (kg)

# 전처리

연료유형은 범주형, 반드시 이항변수로 전환되어야 함

Diesel (1=yes, 0=no)

CNG (1=yes, 0=no)

None needed for “Petrol” (reference category)

# 학습 분할을 위해 선택된 레코드들의 하위 집합 (제한된 변수의 수만 보임)

Id	Model	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	Fuel_Type_Diesel	Fuel_Type_Petrol
1	RRA 2/3-Doors	13500	23	10	2002	46986	1	0
4	RRA 2/3-Doors	14950	26	7	2002	48000	1	0
5	SOL 2/3-Doors	13750	30	3	2002	38500	1	0
6	SOL 2/3-Doors	12950	32	1	2002	61000	1	0
9	VT I 2/3-Doors	21500	27	6	2002	19700	0	1
10	RRA 2/3-Doors	12950	23	10	2002	71138	1	0
12	BNS 2/3-Doors	19950	22	11	2002	43610	0	1
17	ORT 2/3-Doors	22750	30	3	2002	34000	0	1

60% 학습 데이터/ 40% 검증 데이터

# 적합된 회귀분석 모델

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.0000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

# 예측 성능 평가

# 예측 오류 측정

- “적합도 검증”과 같지 않음
- 모델이 얼마나 잘 새로운 데이터를 예측하는가를 알고자 함. 그것이 훈련되는 데이터에 얼마나 잘 맞는가를 알고자 함이 아님
- 대부분 척도의 핵심 성분은 실제  $y$ 와 예측  $\hat{y}$  (“오류”) 사이의 차이

# 몇몇 오류 척도들

**MAE or MAD:** 평균절대오류(편차)

오류의 크기를 제공

## 평균 오류

예측을 넘기는지 미달되는지를 표시

**MAPE:** 평균절대 백분율오류

- $|y-y'|/y * 100$
- $y$  는 실제 값,  $y'$  는 모델 예측 값
- 정답 대비 몇 % 나 틀리는가?

**RMSE** (근의 평균제곱 오류): 오류를 제곱, 평균을  
찾은 후, 제곱근

**총 SSE:** 제곱오류의 총합



# 다른 에러 측정치

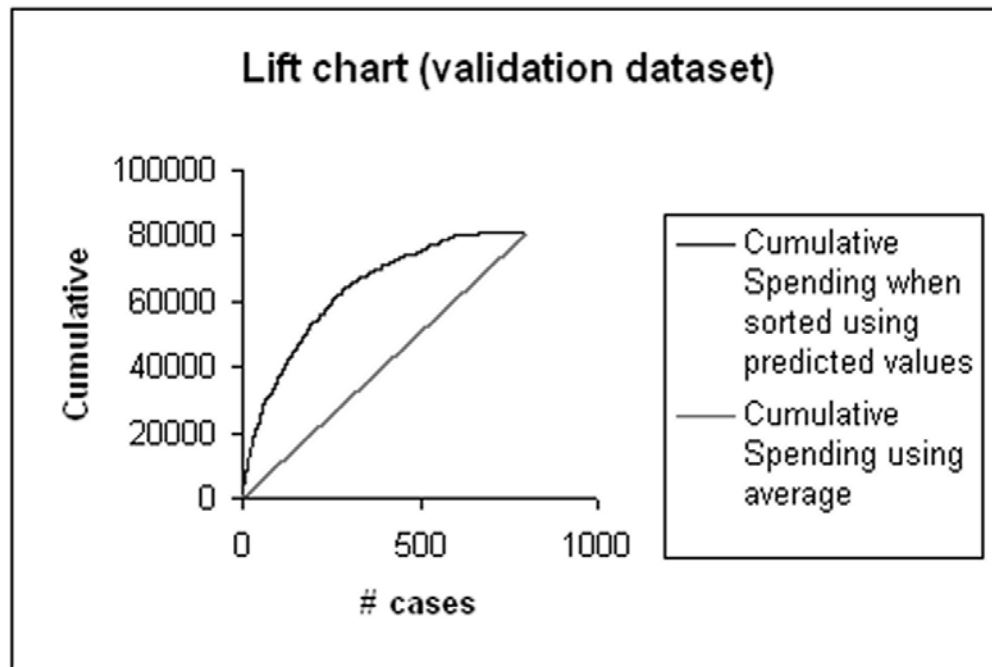
- ( $y$  실제 값,  $y$  예측 값) 의 plot
  - $y' = y$  선 상에 존재하는 데이터는 잔차 = 0 이고,
  - 그 선과 멀리 존재하는 데이터는 잔차 값이 큼
  - 시각적으로 이해하기 용이함

# 예측오류를 위한 향상차트

분류를 위한 향상차트와 유사, 다만...

Y축은, “응답”의 누적 도수 대신, 수치형 타깃 변수(e.g., 수익)의 누적값

# 향상차트 예 - 지출



# 오류 보고

## Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1514553377	1325.527246	-0.000426154

## Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1021587500	1334.079894	116.3728779

# 예측 값

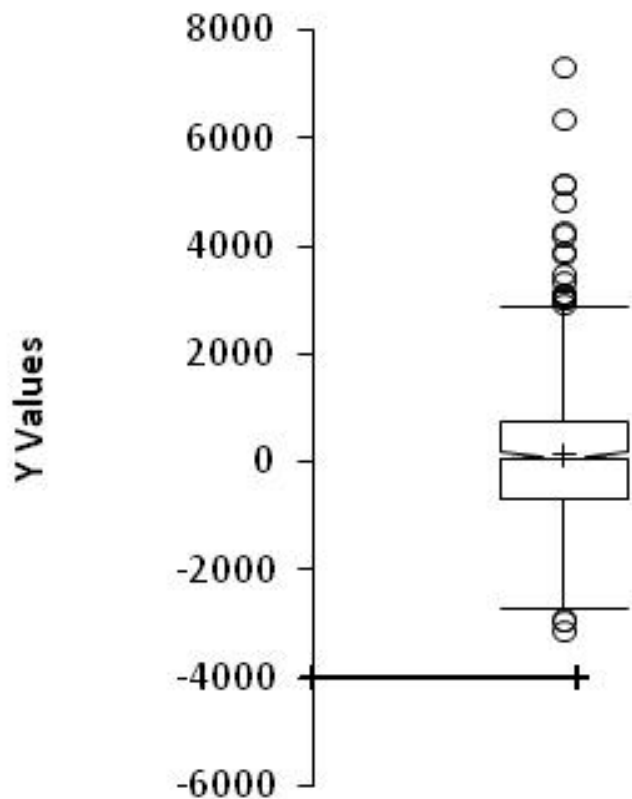
Predicted Value	Actual Value	Residual
15863.86944	13750	-2113.869439
16285.93045	13950	-2335.930454
16222.95248	16900	677.047525
16178.77221	18600	2421.227789
19276.03039	20950	1673.969611
19263.30349	19600	336.6965066
18630.46904	21500	2869.530964
18312.04498	22500	4187.955022
19126.94064	22000	2873.059357
16808.77828	16950	141.2217206
15885.80362	16950	1064.196384
15873.97887	16250	376.0211263
15601.22471	15750	148.7752903
15476.63164	15950	473.3683568
15544.83584	14950	-594.835836
15562.25552	14750	-812.2555172
15222.12869	16750	1527.871313
17782.33234	19000	1217.667664

예측 가격은  
회귀분석 계수를  
사용해 계산

잔차 = 실제 가격과  
예측 가격 사이의  
차이

# 잔차의 분포

## Box Plot



대칭분포  
몇몇 이상치

# 예측변수의 부분 집합 선택

**목표:** 간명한 모델 찾기(충분히 잘 작동할 가장 간단한 모델)

- 좀더 탄탄
- 예측 정확성이 높음

# 예측 변수의 부분 집합 선택

William of Ockham (c. 1285–1349), an influential nominalism philosopher monk, said “It is futile to do with more things that which can be done with fewer.”

## Occam’s Razor:

a principle that generally recommends selecting the competing hypothesis that makes **the fewest new assumptions**, when they both sufficiently explain available data.



# 예측 변수의 부분 집합 선택

전역 탐색

Computationally expensive, 비현실적

부분 탐색 알고리즘

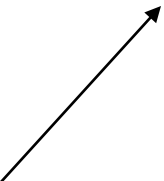
- 전방선택
- 후방제거
- 단계적 선택

# 전역 탐색

- 예측변수들의 모든 가능한 하위집합 평가(하나, 쌍, 삼중항 등)
- 계산적으로 집중됨
- “adjusted  $R^2$ ”로 판단

$$R_{adj}^2 = 1 - \frac{n-1}{n-p-1} (1-R^2)$$

예측변수의 수에  
대한 벌점



# 전방선택 Forward Selection

- 예측변수 없이 시작
- 예측변수를 하나씩 추가(가장 크게 기여하는 변수 추가)
- 추가가 통계적으로 무의미해질 때 중단

# 후방제거 Backward Elimination

- 모든 예측변수로 시작
- 가장 무용한 예측변수를 하나씩 연속적으로 제거
- 남아 있는 모든 예측변수들이 통계적으로 중요한 기여를 할 때 중단

# 단계적 Stepwise

- 전방선택과 같음
- 다만 각 단계에서 중요하지 않은 예측변수들을 버리는 것도 고려

# 후방제거(마지막 7개 모델만 보임)

1	2	3	4	5	6	7	8
Constant	Age_08_04	*	*	*	*	*	*
Constant	Age_08_04	Weight	*	*	*	*	*
Constant	Age_08_04	KM	Weight	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Weight	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Quarterly_Tax	Weight	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Quarterly_Tax	Weight	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Automatic	Quarterly_Tax	Weight

맨 위 모델은 하나의 예측변수를  
가짐(Age\_08\_04)

두 번째 모델은 두 개의 예측변수를 가짐 등등

# 12개 모델 전부

Model (Constant present in all models)											
1	2	3	4	5	6	7	8	9	10	11	12
Constant	Age_08_04	*	*	*	*	*	*	*	*	*	*
Constant	Age_08_04	Weight	*	*	*	*	*	*	*	*	*
Constant	Age_08_04	KM	Weight	*	*	*	*	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Weight	*	*	*	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Quarterly_Tax	Weight	*	*	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Quarterly_Tax	Weight	*	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Automatic	Quarterly_Tax	Weight	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Met_Color	Automatic	Quarterly_Tax	Weight	*	*	*
Constant	Age_08_04	KM	el_Type_Diesel	el_Type_Petrol	HP	Met_Color	Automatic	Quarterly_Tax	Weight	*	*
Constant	Age_08_04	KM	el_Type_Diesel	el_Type_Petrol	HP	Met_Color	Automatic	Doors	Quarterly_Tax	Weight	*
Constant	Age_08_04	KM	el_Type_Diesel	el_Type_Petrol	HP	Met_Color	Automatic	cc	Doors	Quarterly_Tax	Weight

# 12개 모델을 위한 진단

	#Coeffs	RSS	Cp	R-Squared	Adj. R-Squared
<u>Choose Subset</u>	2	2538203648	566.4946289	0.759902259	0.759623076
<u>Choose Subset</u>	3	2245803264	404.393219	0.787561455	0.787066837
<u>Choose Subset</u>	4	1796573056	154.2755432	0.830055744	0.829461533
<u>Choose Subset</u>	5	1689283456	96.06230164	0.84020465	0.839458814
<u>Choose Subset</u>	6	1555462272	22.9589653	0.852863273	0.85200383
<u>Choose Subset</u>	7	1516825984	3.27544785	0.856518017	0.855511126
<u>Choose Subset</u>	8	1515638144	4.60880661	0.856630379	0.855455219
<u>Choose Subset</u>	9	1515206272	6.36643076	0.856671232	0.855326999
<u>Choose Subset</u>	10	1514873088	8.1794405	0.856702749	0.855189045
<u>Choose Subset</u>	11	1514592768	10.02211857	0.856729265	0.855045708
<u>Choose Subset</u>	12	1514553344	11.99999332	0.856732995	0.854878951

좋은 모델은 다음을 갖는다:

높은  $\text{adj-R}^2$ ,  $C_p = \#$  예측변수



Mallow's  $C_p$  ( $P$ =# of variables)

$$C_p = \frac{SSE_p}{S^2} - N + 2P,$$

$$SSE_p = \sum_{i=1}^N (Y_i - Y_{pi})^2$$

# 다음 단계

- 부분 집합 선택 방법은 후보 모델들에게 “좋은 모델”이 될 기회를 준다.
- “최고의” 모델이 정말 가장 좋다는 것을 보증하지는 않는다.
- 또한, “최고의” 모델은 여전히 불충분한 예측 정확성을 가질 수 있다.
- 반드시 후보들을 실행하고 예측 정확성을 평가해야 한다(“choose subset” 클릭).

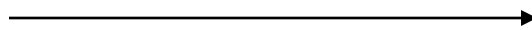
# 6개의 예측변수만을 가진 모델

## The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3874.492188	1415.003052	0.00640071	97276411904
Age_08_04	-123.4366303	3.33806777	0	8033339392
KM	-0.01749926	0.00173714	0	251574528
Fuel_Type_Petrol	2409.154297	319.5795288	0	5049567
HP	19.70204735	4.22180223	0.00000394	291336576
Quarterly_Tax	16.88731384	2.08484554	0	192390864
Weight	15.91809368	1.26474357	0	281026176

### Training Data scoring - Summary Report

모델 적합



Total sum of squared errors	RMS Error	Average Error
1516825972	1326.521353	-0.000143957

### Validation Data scoring - Summary Report

예측 성능

(12개의 예측 모델을 비교하라!)



Total sum of squared errors	RMS Error	Average Error
1021510219	1334.029433	118.4483556

# Danger of Fitting

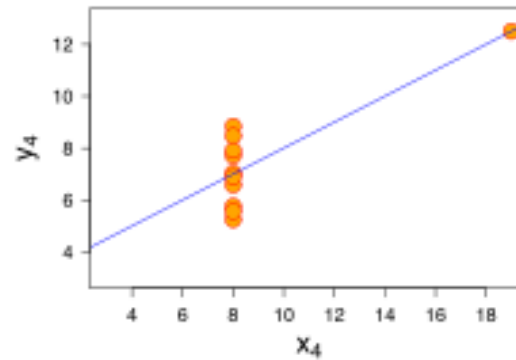
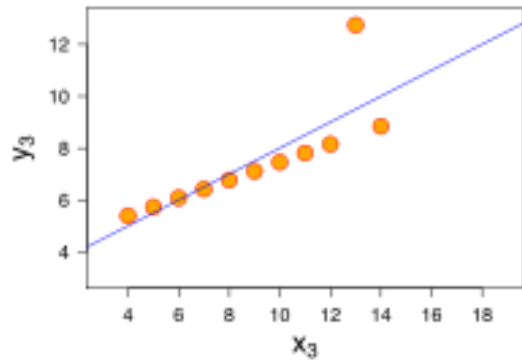
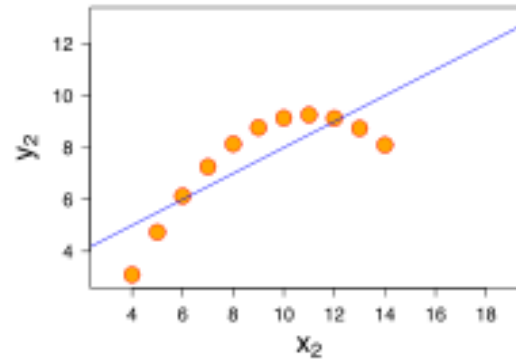
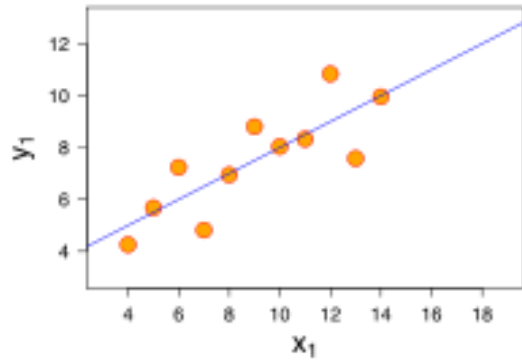
# Anscomb's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

# Anscomb's Quartet

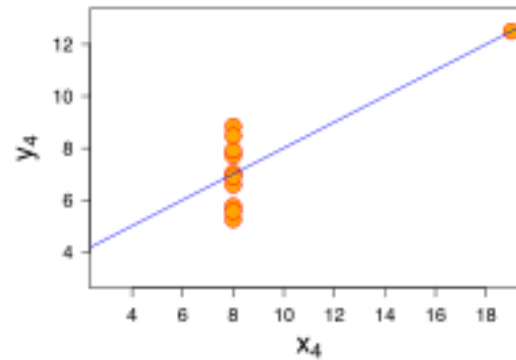
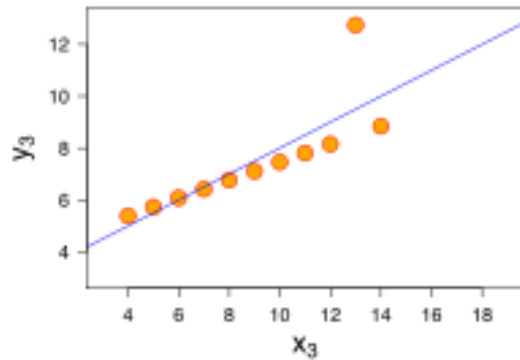
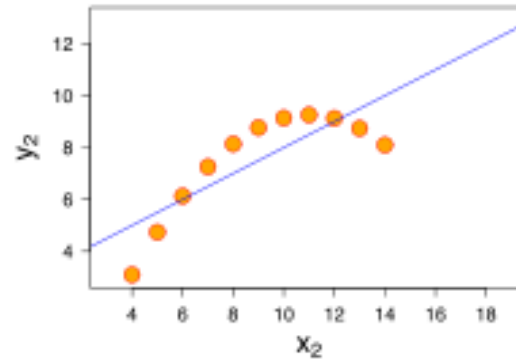
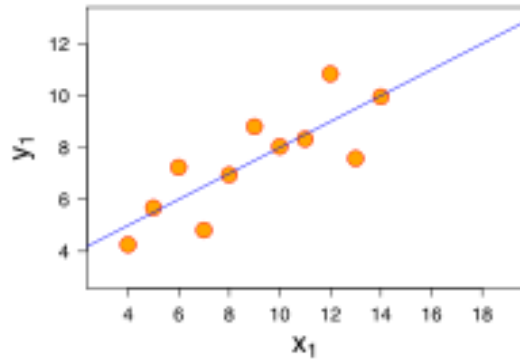
- Mean of  $x$  : 9
  - Variance of  $x$  : 10
  - Mean of  $y$ : 7.5
  - Variance of  $y$  : 3.75
- 
- Correlation between  $x$  and  $y$  : 0.816
  - Linear Regression line:  $y = 3.0 + 0.5 x$

# Anscomb's Quartet



# Anscomb's Quartet

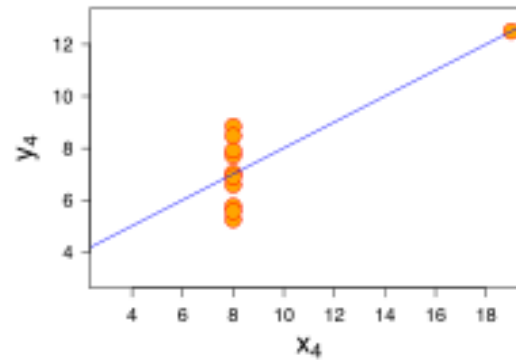
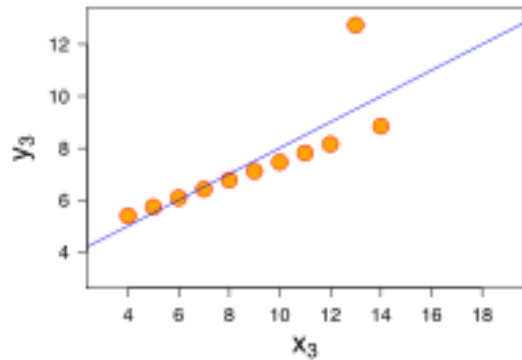
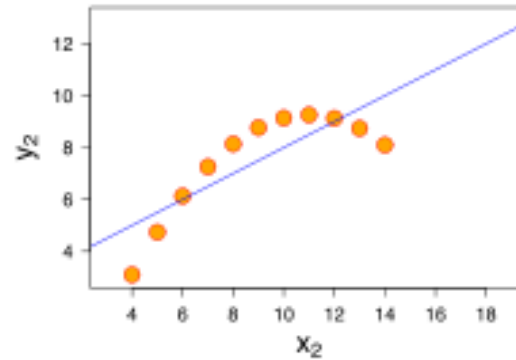
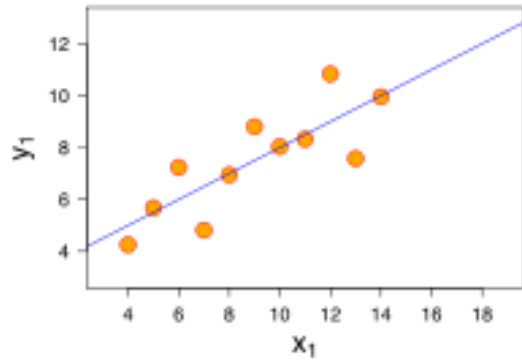
okay



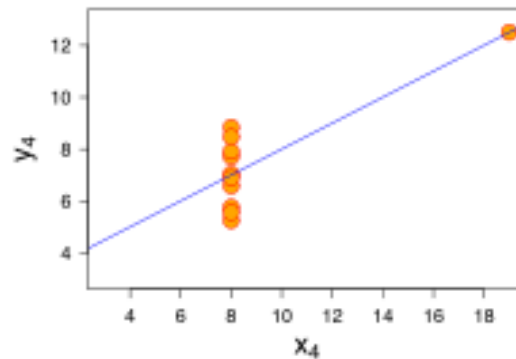
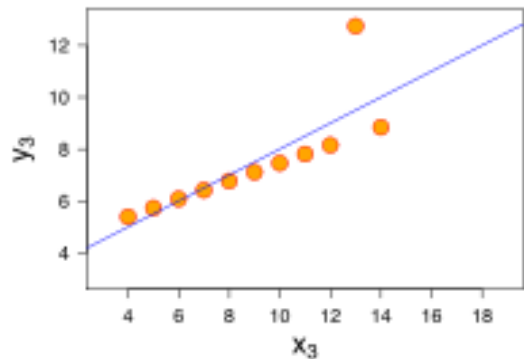
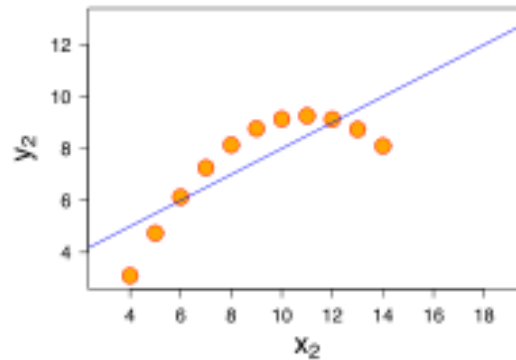
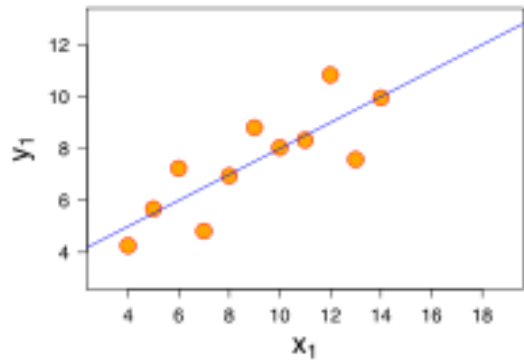


# Anscomb's Quartet

clearly quadratic, 0.816 meaningless

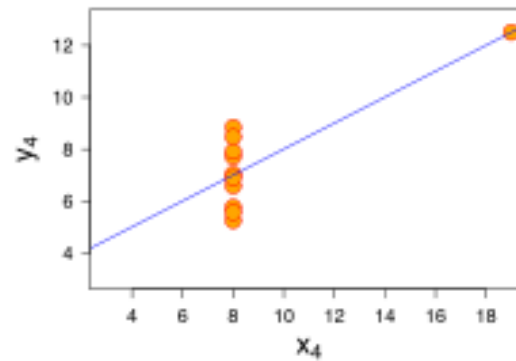
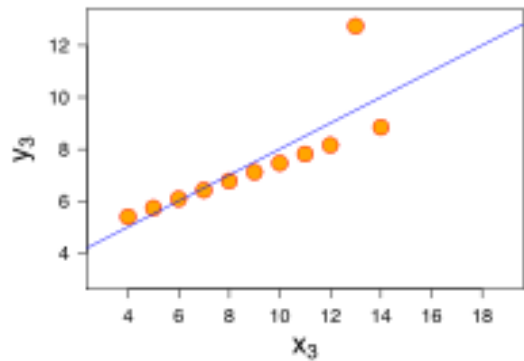
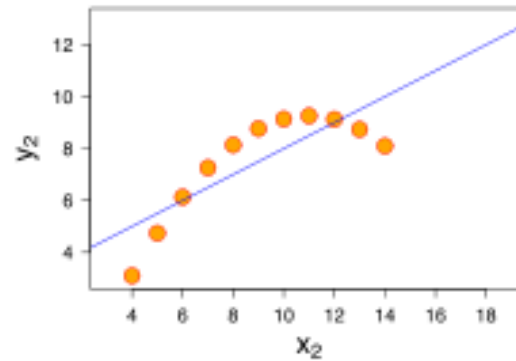
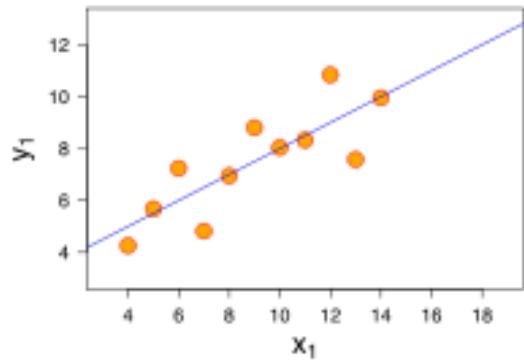


# Anscomb's Quartet



a clear linear relation but with a slightly different slope. But an outlier “incorrects” it.

# Anscomb's Quartet



clearly no linear relation,  
but one outlier makes it

# 요약

- 선형 회귀분석 모델은 설명모델뿐만 아니라 예측모델에서도 매우 유명한 도구다.
- 좋은 예측모델은 높은 예측 정확도를 갖는다(실용적 수준에 유용)
- 예측모델은 학습 데이터 세트를 사용하고 별개의 검증 데이터 세트에서 평가하도록 만들어졌다.
- 여분의 예측변수를 제거하는 것은 예측 정확도와 튼튼함을 얻는 데 핵심적이다.
- 하위집합 선택 방법은 “좋은” 후보 모델을 찾도록 돕는다. 이것들은 반드시 실행 및 평가되어야 한다.