BSC LAB
Business & Service Computing

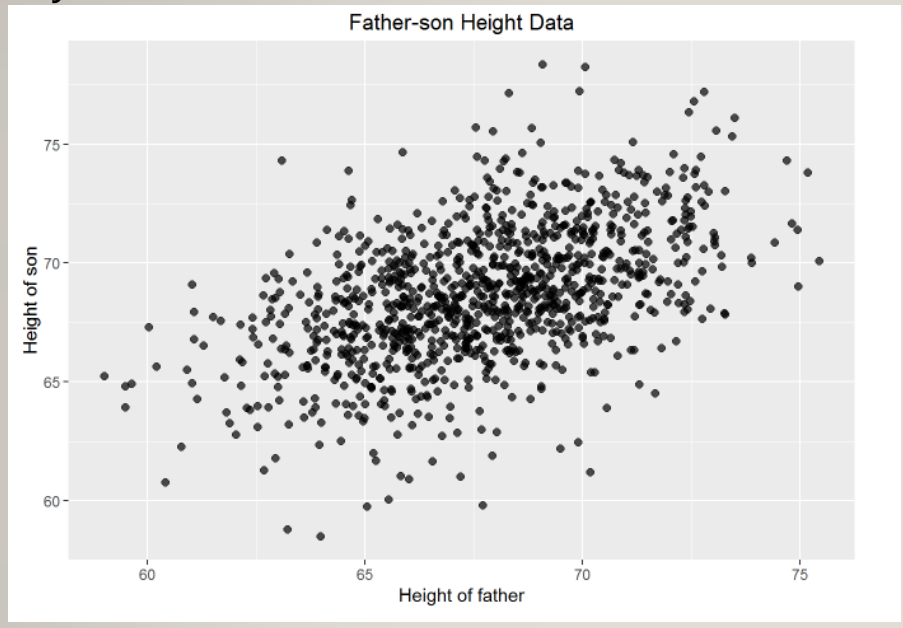
PNU Industrial Data Science

Regression (회귀)

Prof. Hyerim Bae

우리 아들 키가 얼마나 클까?

[By Francis Galton]



[by Orley Ashenfelter]

품질 = 12.145
+ 0.00117 * 겨울 강우량
+ 0.06140 * 생장기 평균 기온
- 0.00386 * 추수기 강우량

Contents

산업데이터과학은 산업현장에서 수집된 데이터를 분석하는데 필요한 기초 소양을 강의합니다.

Regression

Explanatory model vs. Predictive model

- Goal
  - Good EM: Model fits data well(모델이 데이터를 잘 적합시키는 모델)
  - Good PM: Model predicts well for a new case(새로운 사례를 정확하게 예측하는 모델)
- Data set
  - EM: Entire data set(모집단에서 가정된 관계에 대한 정보가 최대한 반영된 최적의 적합 모델을 추정하기 위해서 전체 데이터 세트를 사용)
  - PM: Train data and validation data (데이터는 일반적인 학습세트와 검증세트로 나누어지며, 학습세트는 모델을 추정하는데, 검증세트는 새로운 데이터에 대한 모델의 성능을 평가하는데 사용)
- Performance evaluation
  - EM: Goodness of Fit (데이터가 모델에 얼마나 잘 적합하는가?)
  - PM: Accuracy (모델이 얼마나 새로운 사례를 잘 예측하는가?)

Predictive Modeling

Goal: predict target values in other data where we have predictor values, but not target values

새로운 사례에 대해서 출력값을 알아내기 위함 (DM)

- Classic data mining context
- Model Goal: Optimize predictive accuracy
- Train model on training data
- Assess performance on validation (hold-out) data
- Explaining role of predictors is not primary purpose (but useful)

Estimating the regression equation and prediction

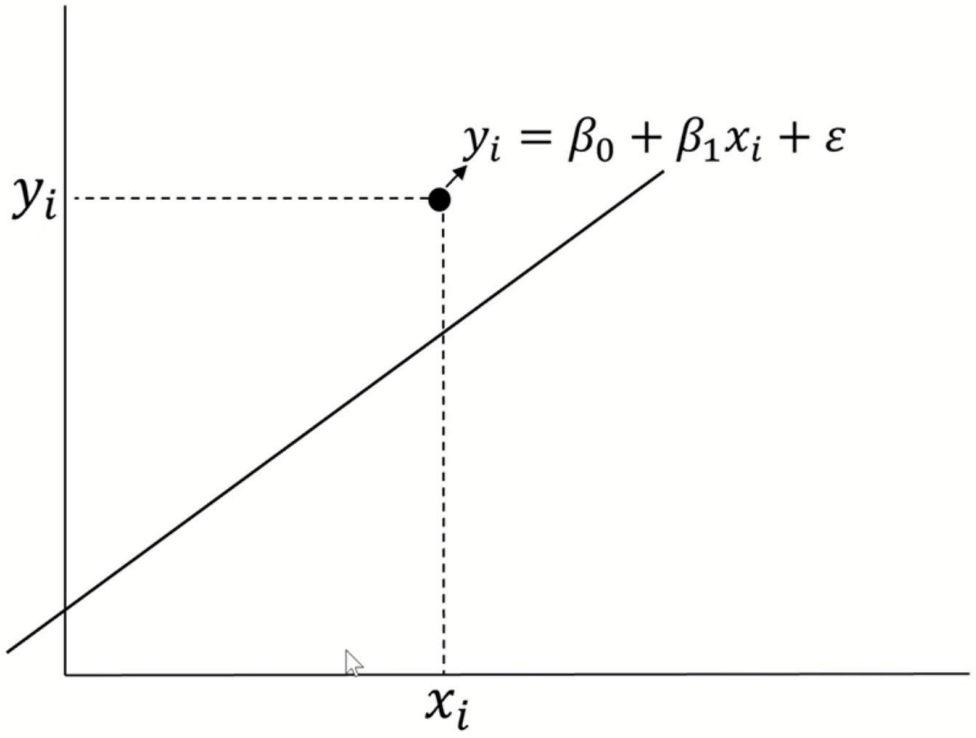
- Regression
  - Simple vs. Multiple
  - Linear vs. Non-linear

- LRM

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

- Unbiased and has the least squared error if
  - (정규성) ε follows a normal distribution
  - (선형성) Predictors and output variable has a linear relation
  - (독립성) Observations are independent of each other
  - (등분산성) The variability of Y values for a given set of predictors is the same (regardless of the values of the predictors. Homoskedasticity))

$$\varepsilon_i \sim (0, \sigma^2)$$



Estimation of the coefficient

- Minimizing the difference between Y_i and $\mathrm{E}(Y_i)$

- Least squares estimators

$$\widehat{\beta_1} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum {x_i}^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

Example: Prices of Toyota Corolla

Goal: predict prices of used Toyota Corollas based on their specification

Data: Prices of 1442 used Toyota Corollas, with their specification information

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

Price in Euros
Age in months as of 8/04
KM (kilometers)
Fuel Type (diesel, petrol, CNG)
HP (horsepower)
Metallic color (1=yes, 0=no)
Automatic transmission (1=yes, 0=no)
CC (cylinder volume)
Doors
Quarterly_Tax (road tax)
Weight (in kg)

The Fitted Regression Model

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	0.00000008	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

Hypthesis

$$H_0: \quad \beta_1 = \beta_{1,0}$$
$$H_1: \quad \beta_1 \neq \beta_{1,0}$$

Test Statistic

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

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

Interval of H_0's acceptance

$$-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}$$

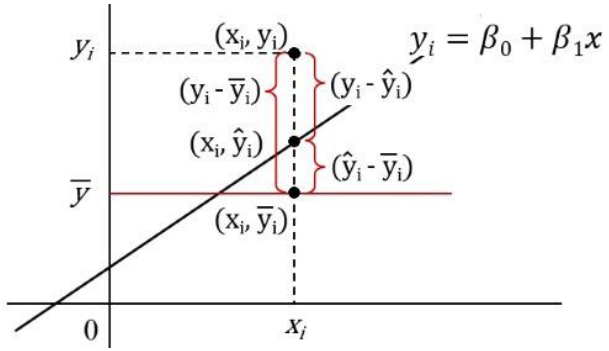
$$pvalue = 2 \times (1 - P(T \leq t_0)$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-8212.5  -839.2   -14.3   831.5  7270.7

Coefficients:
                   Estimate   Std. Error t value             Pr(>|t|)
(Intercept)     -1774.877829  1643.744823  -1.080                0.2807
Age_08_04        -135.430875     4.875906 -27.776 < 0.0000000000000002 ***
KM                 -0.019003     0.002341  -8.116  0.00000000000000283 ***
Fuel_TypeDiesel  1208.339159   534.431400   2.261               0.0241 *
Fuel_TypePetrol  2425.876714   520.587979   4.660  0.00000391697679667 ***
HP                 38.985537     5.587183   6.978  0.00000000000811621 ***
Met_Color          84.792715   126.883452   0.668               0.5042
Automatic         306.684154   289.433138   1.060               0.2898
CC                  0.031966     0.099075   0.323               0.7471
Doors             -44.157742    64.056530  -0.689               0.4909
Quarterly_Tax      16.677343     2.602668   6.408  0.00000000030287017 ***
Weight             12.667487     1.536587   8.244  0.00000000000000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1406 on 588 degrees of freedom
Multiple R-squared:  0.8567,    Adjusted R-squared:  0.854
F-statistic: 319.6 on 11 and 588 DF,  p-value: < 0.00000000000000022
```

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$
$$= \left(\hat{Y}_i - \bar{Y}\right) + e_i$$



Predicted Values

Training Data scoring - Summary Report

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

Validation Data scoring - Summary Report

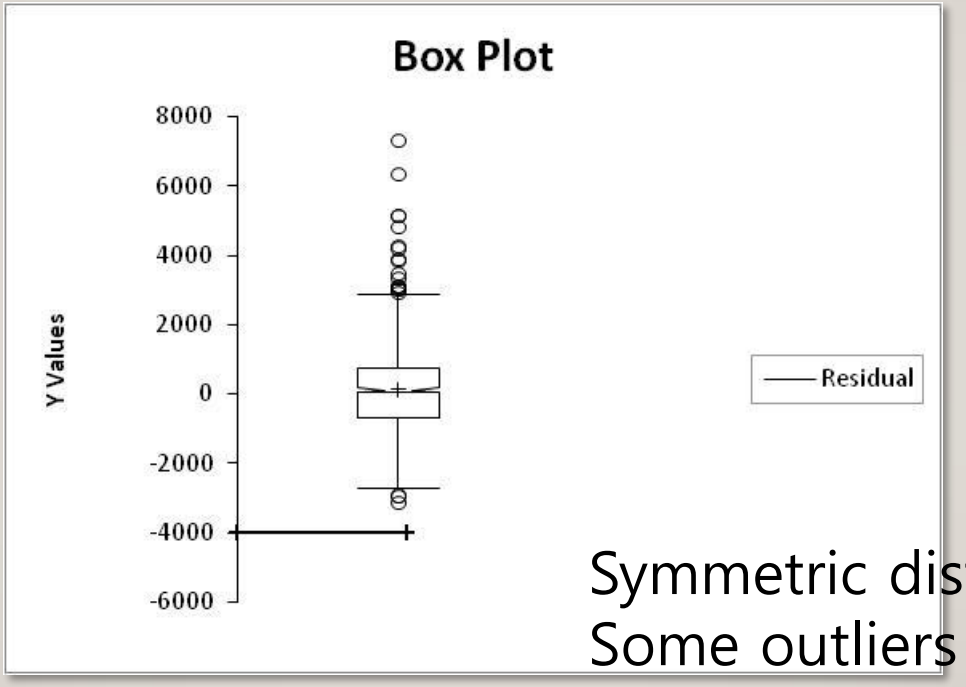
Total sum of squared errors	RMS Error	Average Error
1021587500	1334.079894	116.3728779

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

Predicted price computed using regression coefficients

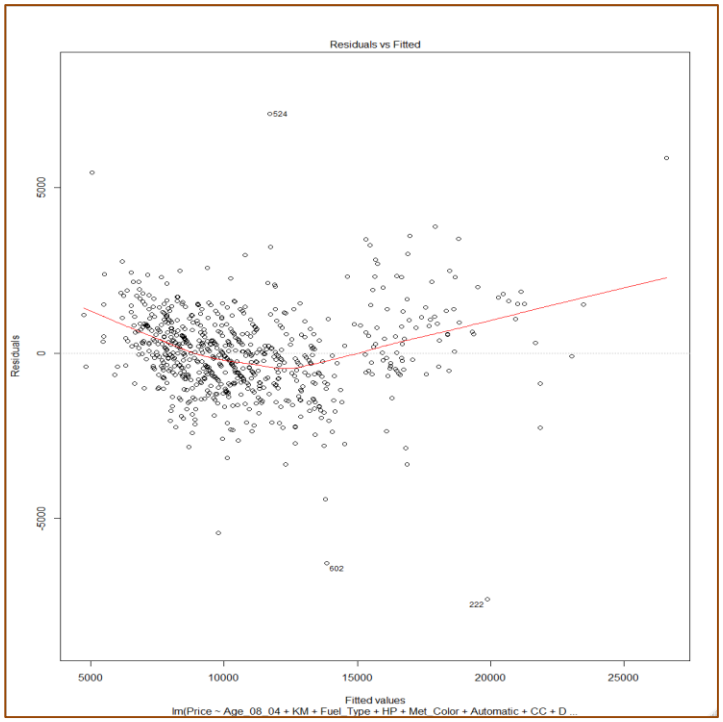
Residuals = difference between actual and predicted prices

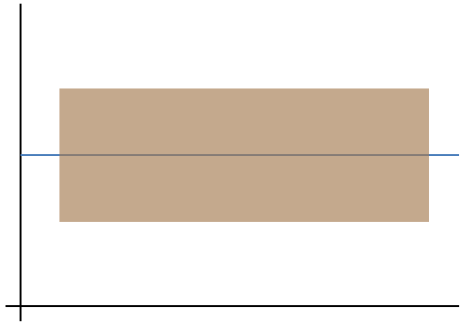
Distribution of Residuals



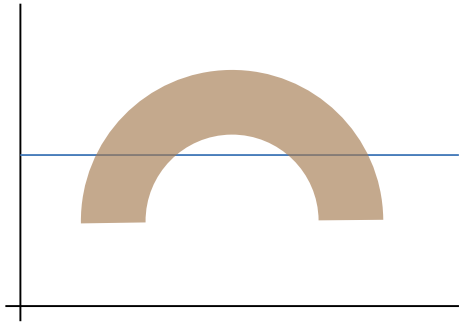
Symmetric distribution
Some outliers

잔차 분석

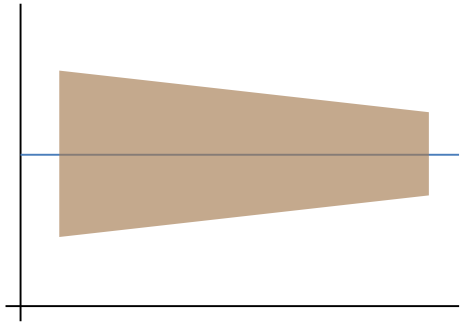

Residuals vs Fitted
524
602
222
5000
0
-5000
Residuals
5000
10000
15000
20000
25000
Fitted values
lm(Price ~ Age_08_04 + KM + Fuel_Type + HP + Met_Color + Automatic + CC + D ...



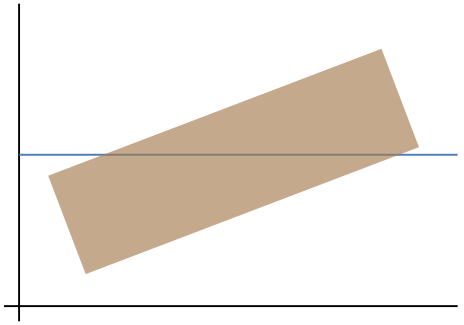
1



2



3



4

Practical reason why we need to reduction the # of predictors

- Many predictors…
  - Cost for data collection (정보수집에 드는 비용)
  - May be able to measure fewer predictors more accurately (평가의 정확성)
  - High chance of missing values (다수의 결측치)
  - Multicollinearity: Estimation of regression coefficients are likely to be unstable (다중공선성)
  - May increase the variance of predictors
  - May cause inaccuracy

1. 미래에 대한 예측을 할 때, 예측변수들 전부를 수집하는 것이 실행 가능하지 않거나 비용이 많이 든다.
2. 적은 수의 예측변수를 사용하면 더 정확한 측정을 할 수 있다.
3. 예측변수가 많을 수록 결측치 존재의 위험성이 높아진다.
4. Parsimony가 좋은 모델의 중요한 성질이다.
5. 많은 변수로 인해 다중공선성(두 개 이상의 예측변수가 종속변수에 동일한 선형 관계를 공유하는 것)이 발생하고, 이로 인해 회귀계수의 추정치들이 불안해질 수 있다. 예측변수가 p개라면 5(p+2)보다 많은 사례를 데이터로 사용해야함
6. 종속변수와 상관없는 예측변수를 사용하면 예측의 분산이 증가할 수 있다.
7. 종속변수와 실제 상관관계가 있는 예측변수를 누락시키면 예측의 평균오차 혹은 편향도가 증가할 수 있다.

Exhaustive Search

- R^2: Ratio of variance that a model can explain(모델에서 설명할 수 있는 변동성의 비율)
  - All possible subsets of predictors assessed (single, pairs, triplets, etc.)
  - Computationally intensive
  - Judge by "adjusted R^2"

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

Penalty for number of predictors

예측변수의 수만 증가시켜도 발생하는 R^2의 인위적인 증가를 배제

Selecting Subsets of Predictors

Goal: Find parsimonious model (the simplest model that performs sufficiently well)

- More robust
- Higher predictive accuracy

Exhaustive Search

Partial Search Algorithms

- Forward
- Backward
- Stepwise

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

Model with only 6 predictors

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

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

Model Fit →

Training Data scoring - Summary Report

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

Predictive performance
(compare to 12-predictor model!) →

Validation Data scoring - Summary Report

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

Summary

- Linear regression models are very popular tools, not only for explanatory modeling, but also for prediction
- A good predictive model has high predictive accuracy (to a useful practical level)
- Predictive models are built using a training data set, and evaluated on a separate validation data set
- Removing redundant predictors is key to achieving predictive accuracy and robustness
- Subset selection methods help find “good” candidate models. These should then be run and assessed.

Logistic regression

Logistic Regression

- Extends idea of linear regression to situation where outcome variable is categorical

- Widely used, particularly where a structured model is useful to explain (=*profiling*) or to predict
  - 고객을 반납/비반납 고객으로 분류(분류)
  - 남자 최고경영진과 여자 최고경영진으로 구별하는 요인찾기 (프로파일링)

- We focus on binary classification
  i.e. Y=0 or Y=1

- 2 steps
  - 각 클래스에 속하는 확률을 추정
  - 각 관측치를 이들 클래스 중 하나로 분류하기 위하여 확률값에 대한 분류기준값을 적용

The Logit

Goal: Find a function of the predictor variables that relates them to a 0/1 outcome

- Instead of *Y* as outcome variable (like in linear regression), we use a function of Y called the ***logit***
- Logit can be modeled as a linear function of the predictors
- The logit can be mapped back to a probability, which, in turn, can be mapped to a class

Step 1: Logistic Response Function

p = probability of belonging to class 1

Need to relate p to predictors with a function that guarantees $0 \leq p \leq 1$

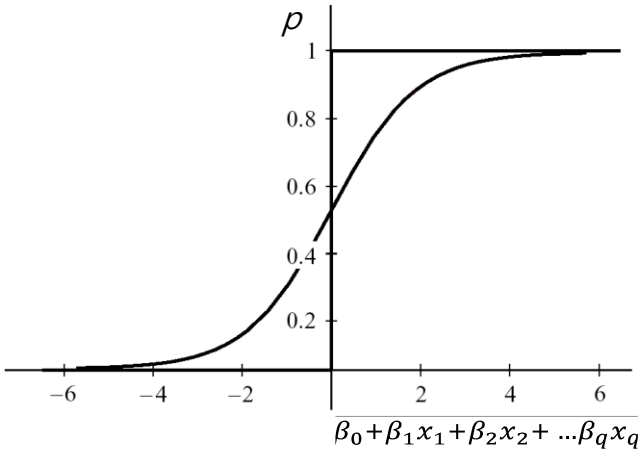
Standard linear function (as shown below) does not:

Want to guarantee that Y exists in [0, 1]

$$p_{LR} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q$$

q = number of predictors

The Fix: use *logistic response function*



$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\ldots\beta_qx_q)}}$$

Step 2: The Odds

The odds(클래스 1에 속할) of an event are defined as:

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

← p = probability of event

"클래스 0에 속할 확률에 대한 클래스 1에 속할 확률"

$$p = \frac{Odds}{1+Odds}$$

We can also relate the Odds to the predictors:

eq. 10.5

$$Odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q}$$

x_j가 한단위 증가하면?

To get this result, substitute 10.2 into 10.4

Step 3: Take log on both sides

This gives us the logit:

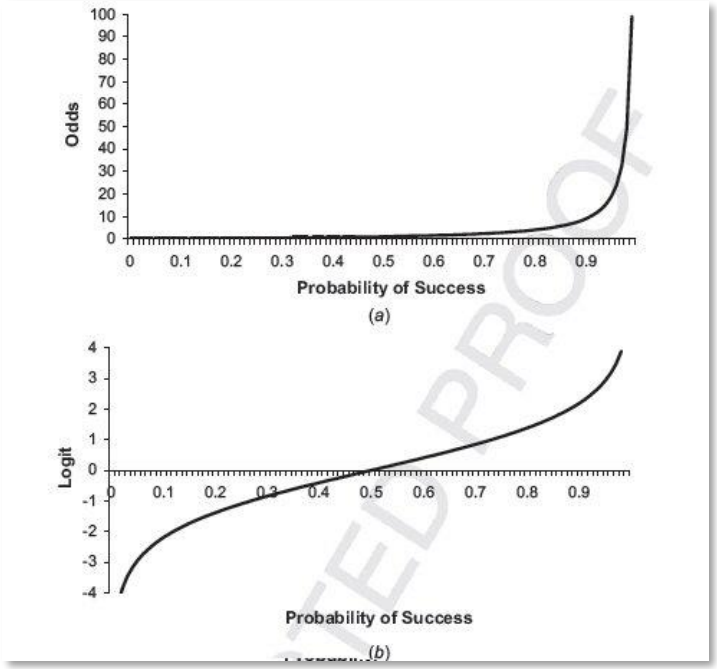
$$\log(Odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

Logit, cont.

So, the logit is a linear function of predictors x_1, x_2, …

- Takes values from -infinity to +infinity

Review the relationship between logit, odds and probability



Personal Loan Offer

Outcome variable: accept bank loan (0/1)

Predictors: Demographic info (연령, 소득 등), and info about their bank relationship (담보대출건, 증권계좌)

Data preprocessing

- Partition 60% training, 40% validation
- Create 0/1 dummy variables for categorical predictors

$$EducProf = \begin{cases} 1 \text{ if education is } Professional \\ 0 \text{ otherwise} \end{cases}$$

$$EducGrad = \begin{cases} 1 \text{ if education is at } Graduate \text{ level} \\ 0 \text{ otherwise} \end{cases}$$

$$Securities = \begin{cases} 1 \text{ if customer has securities account in bank} \\ 0 \text{ otherwise} \end{cases}$$

$$CD = \begin{cases} 1 \text{ if customer has CD account in bank} \\ 0 \text{ otherwise} \end{cases}$$

$$Online = \begin{cases} 1 \text{ if customer uses online banking} \\ 0 \text{ otherwise} \end{cases}$$

$$CreditCard = \begin{cases} 1 \text{ if customer holds Universal Bank credit card} \\ 0 \text{ otherwise} \end{cases}$$

Single Predictor Model

Modeling loan acceptance on income (x)

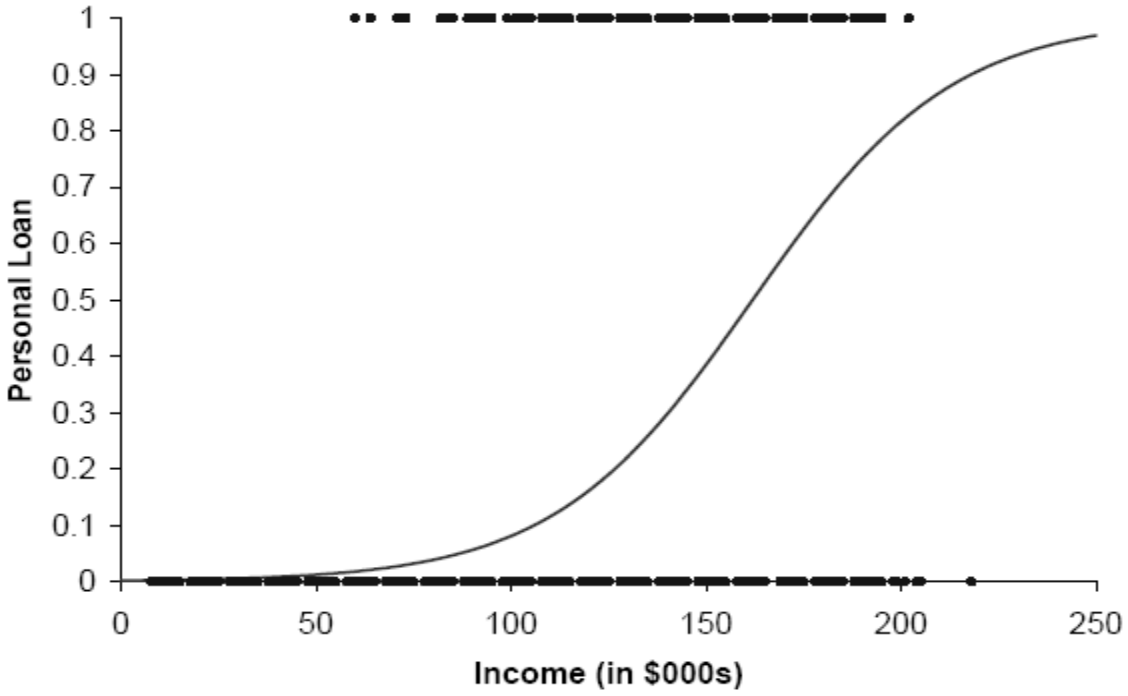
$$\text{Prob}(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Fitted coefficients (more later): $b_0 = -6.3525$, $b_1 = 0.0392$

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.3525 - 0.0392x}}$$

Seeing the Relationship

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.3525 - 0.0392x}}$$



Last step - classify

Model produces an estimated probability of being a "1"

- Convert to a classification by establishing cutoff level
- If estimated prob. > cutoff, classify as "1"

Ways to Determine Cutoff

- 0.50 is popular initial choice

- Additional considerations (see Chapter 5)
  - Maximize classification accuracy
  - Maximize sensitivity (subject to min. level of specificity)
  - Minimize false positives (subject to max. false negative rate)
  - Minimize expected cost of misclassification (need to specify costs)

Example, cont.

- Estimates of βs are derived through an iterative process called *maximum likelihood estimation*
  - 주어진 데이터를 얻을 가능성을 최대로 하는 추정치를 찾는 방법
  - 컴퓨터 프로그램을 사용하여 반복 추정
- Let's include all 12 predictors in the model now
- XLMiner's output gives coefficients for the logit, as well as odds for the individual terms

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

Figure 10.3: Logistic regression coefficient table for personal loan acceptance as a function of 12 predictors.

Estimated Equation for Logit

$$\begin{aligned}\text{logit} = &-13.201 - 0.045\,Age + 0.057\,Experience + 0.066\,Income + 0.572\,Family \\ &+0.18724874\,CCAvg + 0.002\,Mortgage - 0.855\,Securities + 3.469\,CD \\ &-0.844\,Online - 0.964\,Credit\ Card + 4.589\,EducGrad + 4.523\,EducProf\end{aligned}$$

Equation for Odds (Equation 10.10)

음의 계수 양의 계수

$$\begin{aligned}\text{odds}(Personal\ Loan = \text{Yes}) = {} & e^{-13.201}(0.956)^{Age}\ (1.058)^{Experience}\ (1.068)^{Income} \\ & \cdot (1.771)^{Family}\ (1.206)^{CCAvg}\ (1.002)^{Mortgage} \\ & \cdot (0.425)^{Securities}\ (32.105)^{CD}\ (0.430)^{Online} \\ & \cdot (0.381)^{CreditCard}(98.405)^{EducGrad}\ (92.086)^{EducProf}\end{aligned}$$

Converting to Probability

$$p = \frac{Odds}{1 + Odds}$$

Interpreting Odds, Probability

For predictive classification, we typically use probability with a cutoff value

For explanatory purposes, odds have a useful interpretation:

- If we increase x_1 by one unit, holding x_2, x_3 … x_q constant, then
- b_1 is the factor by which the odds of belonging to class 1 increase

–예측변수가 가변수 일때,

- CD에 대한 odds=32.015 => CD계좌를 가지지 않은 고객에 비하여 대출제안을 수락할 odds

–가변수가 아닌 연속형 변수일 때,

- 예: xj가 3->4로 증가할 때와 30->31로 증가할 때, p에 미치는 효과가 상이함

–오즈비

- 두 개의 범주사이의 오즈 값의 비
- 예) 전문교육과 대학원 교육을 받은 고객에 대한 대출제안 수락여부: 오즈비가 1보다 크면, 전문교육을 받은 고객이 대학원 교육을 받은 고객보다 대출 수락할 오즈보다 높다

Summary

- Logistic regression is similar to linear regression, except that it is used with a categorical response
- It can be used for explanatory tasks (=profiling) or predictive tasks (=classification)
- The predictors are related to the response Y via a nonlinear function called the *logit*
- As in linear regression, reducing predictors can be done via variable selection
- Logistic regression can be generalized to more than two classes (not in XLMiner)

부록

3.3.1 선형성(linearity)

진단방법 ①(설명변수와 종속변수) 산점도 ➔ 이차 함수 형태

②잔차와 예측치 산점도 ➔ 이차 함수 형태

해결방법 ①설명 변수의 이차항이나 다차항을 삽입한다.

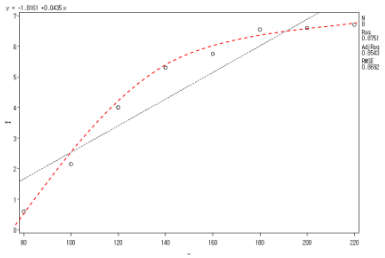
산점도를 보면 종속변수와 설명변수의 직선(산형) 관계를 진단할 수 있다. 잔차와 예측치의 산점도가 일정한 함수 형태를 가지면(일반적으로 이차 함수) 선형성이 무너지게 되는데 이를 해결하려면 설명변수의 이차항을 설명변수로 추가한다. 이차항을 추가할 때는 설명변수를 표준화 한 후 넣으면 다중공선성 문제가 완화된다. (다음 페이지 참고)

EXAMPLE 3-2 선형성 파괴: 이차 관계

```
data quard;
    input y x @@;
    cards;
0.6 80 6.7 220 5.3 140 4 120
6.55 180 2.15 100 6.6 200 5.75 160
run;

proc reg data=quard;
    model y=x;
    plot y*x;
    plot student.*predicted.;
run;
```

종속변수와 설명변수의 산점도를 보면 직선 관계라고 보기 어렵다. 이차 함수 관계에 가깝다.



3.3.2 등분산성(homoscedasticity)

진단방법 ①잔차와 예측치 산점도, 나팔 모양

해결방법 ①가중최소자승법, WLS(Weighted Least Square) 사용한다.

②종속변수변환. 일반적으로 LOG 변환을 하는 것이 일반적이다.

잔차와 예측치 산점도에서 나팔 모양이면 오차의 분산이 예측치가 커짐에 따라 커지거나 작아지고 있음을 의미하므로 등분산 가정이 무너지게 된다. 이런 경우 가중최소자승 추정치를 이용하거나 종속변수변환을 실시한다. 등분산의 경우 일반적으로 오차의 분산은 $V(\epsilon_i)=\sigma_i^2=\sigma^2/w_i$ 으로 가정되고 가중최소자승가중치로 $w_i=1/y_i^2$, 혹은 $w_i=1/x_i^2$ 을 주로 사용한다.

WLS(Weighted Least Square)

$\min_{\alpha,\beta}\sum w_i(y_i-\alpha-\beta x_i)^2$인 $\hat{\alpha},\hat{\beta}$을 WLS 추정치라 한다. 일반적으로 가중치 w_i는 $1/\sigma_i^2$ (σ_i^2

을 알고 있을 때, 그러나 실제 알지 못한다) 혹은 $1/x_i^2$, $1/\hat{y}_i^2$ 등을 사용한다. 단순회귀의 잔차분석은 잔차와 예측치 산점도에 주로 의존하므로 $1/\hat{y}_i^2$을 주로 사용한다. 다중회귀에서는 문제가 되는 설명변수를 이용한 가중치$1/x_i^2$을 사용하기도 하지만 판단이 쉽지 않아 다중회귀모형에서도 $1/\hat{y}_i^2$을 사용한다.

가중회귀 추정치를 구하는 문제는 다음과 같이 생각할 수 있다. 종속변수가 y_i^*, 설명변수가 $1/x_i$인 회귀모형의 OLS 구하는 문제와 동일하다.

$$\min_{\alpha,\beta}\sum\frac{1}{x_i^2}(y_i-\alpha-\beta x_i)^2=\min_{\alpha,\beta}\sum(\frac{y_i}{x_i}-\frac{\alpha}{x_i}-\beta)^2=\min_{\alpha,\beta}\sum(y_i^*-\frac{1}{x_i}\alpha-\beta)^2$$

가중치를 $1/\hat{y}_i^2$ 사용했을 때는 다음 정규방정식에 의해 추정치를 구할 수 있다. 이를 가중회귀추정치이다.

$$\alpha\sum w_i+\beta\sum w_i x_i=\sum w_i y_i$$
$$\alpha\sum w_i x_i+\beta\sum w_i x_i^2=\sum w_i x_i y_i$$