

Albatross - Statistical Package in the Web

(<http://cheoling.snu.ac.kr:3838/DHGLM/>)

Albatross Analytics

Data Import

Data Management ▾

Basic Analysis ▾

Regression ▾

Random Effect Model ▾

Survival Analysis ▾

Materials ▾

Upload File


Browse...

No file selected

☒ Header

Separator

Comma ▾

 Download Data

Reset Data

Albatross - Menu

Data Import			
Data Management	Data Handling	Random Effect Model	Double HGLM
	Merge Dataset		MDHGLM
Basic Analysis	Descriptive Statistic	Survival Analysis	Kaplan-Meier Estimate
	t-test		Cox Model
	ANOVA		Frailty Model
	Frequency Analysis		Competing Risk Model
	Correlation Analysis		Joint Model
Regression	Linear Model	Materials	Data-sets & Manual
	GLM		
	- Logit		

Materials – data_sets and manual

download

data_sets ▼

<http://cheoling.snu.ac.kr:3838/data-sets/>

Index of /data-sets/

- 2월25일강의데이터.zip
- data-sets.zip

download

manual ▼

<http://cheoling.snu.ac.kr:3838/Manual/>

Index of /Manual/

- Manual_2002118.pptx
- Manual_200212.pdf
- Manual(Korean).pdf
- 접속방법.pptx

변수종류에 따른 통계분석법

반응변수(y)	설명변수(x)	통계분석법	비모수 검정
연속형(혈압)	범주형(2개 범주)	t-검정	윌콕슨 순위합 검정
		대응표본 t-검정	윌콕슨 부호순위 검정
연속형(혈압)	범주형(3개 이상)	분산분석(ANOVA)	크루스칼-왈리스 검정
범주형(병발생여부)	범주형(투약여부)	빈도분석(카이제곱 검정)	피셔 정확 검정
연속형(아기체중)	연속형(임신기간)	상관분석, 단순회귀분석	스피어만 상관계수
연속형	연속형..+ 범주형..	선형모형(회귀분석)	
개수,이진, ...	연속형..+ 범주형..	일반화선형모형	
이진형	연속형..+ 범주형..	로짓모형	
생존시간(수술 후 재발시간)	범주형	카플란-마이어 생존곡선,	
		로그-순위 검정	
생존시간(수술 후 재발시간)	연속형..+ 범주형..	콕스 모형	

Contents

1. Basic Statistics
 - t-test, ANOVA, Frequency, Correlation
2. Linear Model
3. Generalized Linear Model
4. Joint Model
5. Linear Mixed Model
6. Hierarchical Generalized Linear Model
7. Survival Data Analysis

1. Basic Statistics

Crabs.csv

Albatross Analytics | Data Import | Data Management ▾ | Basic Analysis ▾ | Regression ▾ | Random Effect Model ▾ | Survival Analysis ▾

Upload File

Browse...

Crabs.csv

Upload complete

☒ Header

Separator

Comma ▾

Download Data

Reset Data

✕ Delete

+ Add New

🔧 Edit Data

Data Selection ☒ single ☐ multiple

Show

10 ▾

 entries

	crab ↕	sat ↕	y ↕	weight ↕
1	1	8	1	3.05
2	2	0	0	1.55
3	3	9	1	2.3
4	4	0	0	2.1
5	5	4	1	2.6
6	6	0	0	2.1

- crab : 174 female crabs
- sat : number of satellites
- y : =1 (sat>0), =0 (sat=0)
- weight : Crab weight in kig
- width : Crab carapace width in cm
- color : 1=light medium,
2=medium, 3=dark medium,
4=dark
- spline : 1=both good, 2=one worn
or broken, 3=both worn or broken

Descriptive Statistics

Albatross Analytics Data Import Data Management ▾ **Basic Analysis ▾** Regression ▾ Random Effect Model ▾ Survival Analysis ▾

Descriptive Statistics

Run

Variable (Numeric Only)

width ▾

☐ Use Group

Variable : width

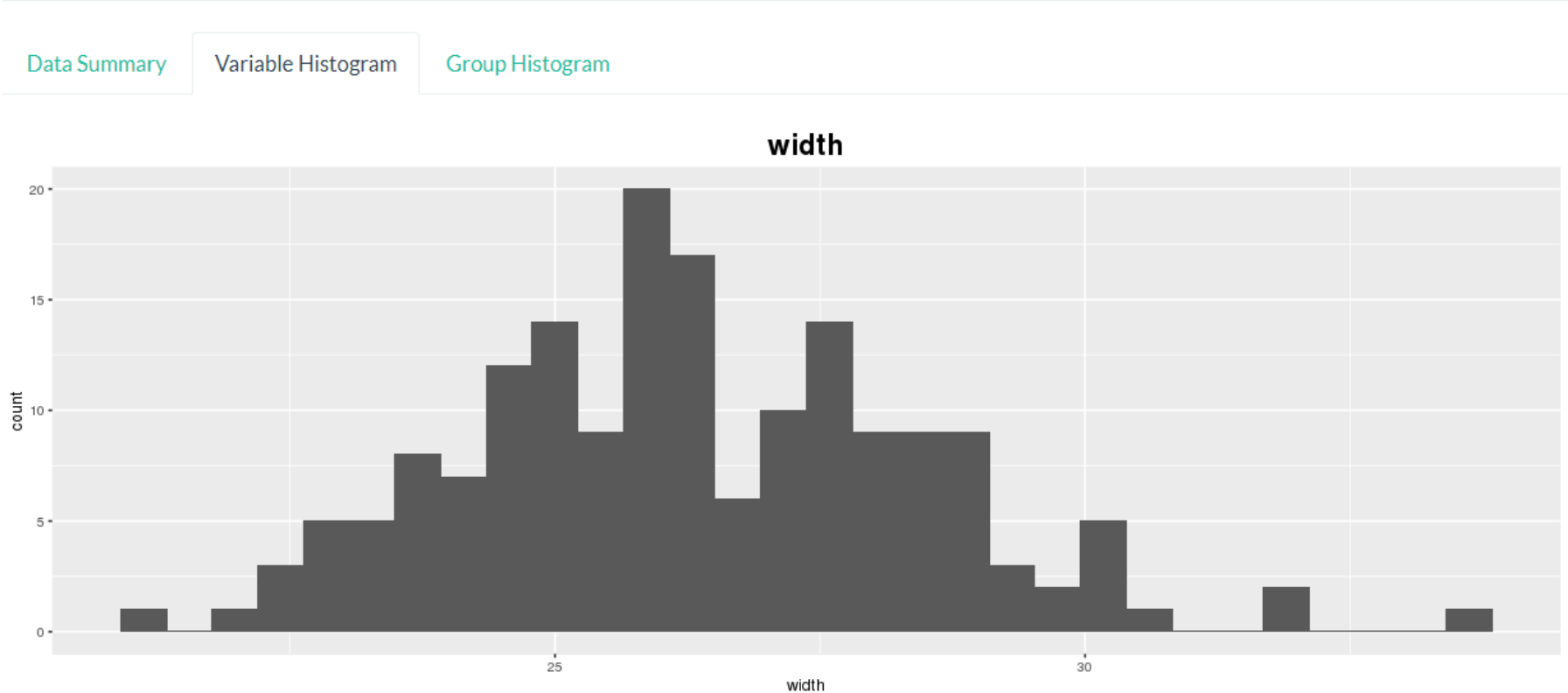
Data Summary

[Variable Histogram](#)

Variable Results

	n	mean	min	median	max	sd	se
width	173.0000	26.2988	21.0000	26.1000	33.5000	2.1091	0.1603

Variable : width



Variable : width

Data Summary

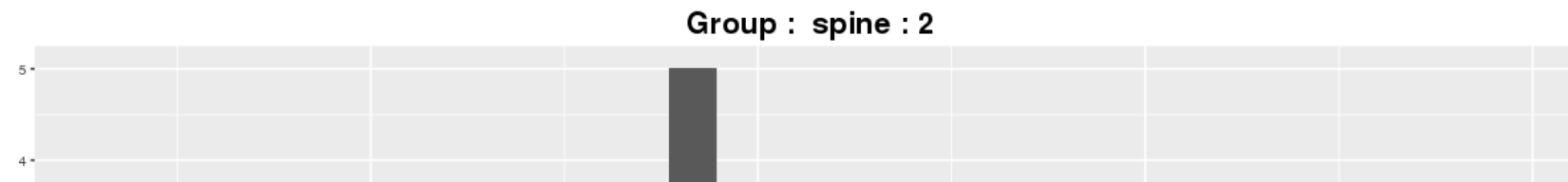
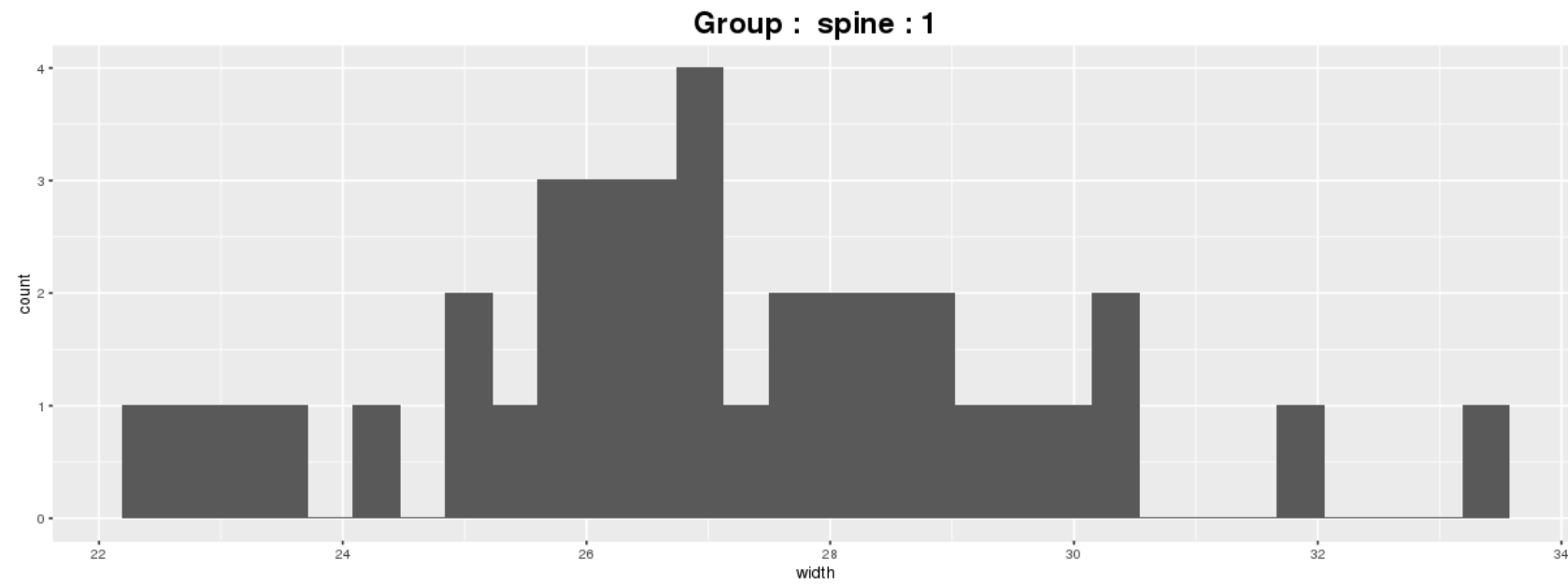
Variable Histogram

Group Histogram

Select Histogram

spine : 1 spine : 2

Append Histogram



one sample t-test

Albatross Analytics

Data Import

Data Management ▾

Basic Analysis ▾

Regression ▾

Random Effect Model ▾

Survival Analysis ▾

t-test

Run

Type of t-test

☒ One sample ☐ Paired ☐ Unpaired

Variable (Numeric Only)

weight ▾

Null Value

3

Alternative Hypothesis

two.sided ▾

Significance Level

0.05 ▴ ▾

☐ Shapiro-Wilk Test (Normality)

One Sample t-test

Model Summary

Data View

Descriptive Statistics

	n	mean	median	sd	se
weight	173	2.4372	2.3500	0.5770	0.0439

t-test

t	df	p.value	mean.of.x	se	lowerCI	upperCI
-12.8289	172.0000	0.0000	2.4372	0.0439	2.3506	2.5238

☒ Wilcoxon Test

Wilcoxon Test

W	p-value
1081.0000	0.0000

paired t-test

[Albatross Analytics](#)[Data Import](#)[Data Management ▾](#)[Basic Analysis ▾](#)[Regression ▾](#)[Random Effect Model ▾](#)[Survival Analysis ▾](#)

t-test

[Run](#)

Type of t-test

☐ One sample ☒ Paired ☐ Unpaired

Variable 1 (Numeric Only)

weight ▾

Variable 2 (Numeric Only)

width ▾

Null Difference

-25

Alternative Hypothesis

two.sided ▾

Significance Level

0.05

☐ Shapiro-Wilk Test (Normality)

Paired t-test

[Model Summary](#)[Data View](#)

Descriptive Statistics

	n	mean	median	sd	se
weight	173	2.4372	2.3500	0.5770	0.0439
width	173	26.2988	26.1000	2.1091	0.1603

t-test

t	df	p.value	mean.of.the.differences	se	lowerCI	upperCI
9.2457	172.0000	0.0000	-23.8617	0.1231	-24.1047	-23.6186

☒ Single Rank Test

Wilcoxon Test

V	p-value
12665.5000	0.0000

t-test

t-test

Run

Type of t-test

☐ One sample ☐ Paired ☒ Unpaired

Variable (Numeric Only)

width

Group Variable

spine

Group 1

1

Group 2

3

Null Value

0

Equal Variances

TRUE

Alternative Hypothesis

two.sided

Significance Level

0.05

☐ Shapiro-Wilk Test (Normality)

☐ Levene Test (Variance Equality)

Unpaired t-test

Model Summary

[Data View](#)

Descriptive Statistics

	n	mean	median	sd	se
1	37	27.1108	26.8000	2.4119	0.3965
3	121	26.2455	26.2000	1.9251	0.1750

t-test

t	df	p.value	mean.of.x	mean.of.y	se	lowerCI	upperCI
2.2495	156.0000	0.0259	27.1108	26.2455	0.3847	0.1055	1.6252

☒ Wilcoxon Rank Sum Test

Wilcoxon Test

W	p-value
2716.0000	0.0501

ANOVA

ANOVA

Run

Variable (Numeric Only)

sat

Group Variable

Make Interaction Variable

Append

☐ Shapiro-Wilk Test (Normality)

☐ Breusch-Pagan Test (Homoscedasticity)

☐ Kruskal-Wallis Test

ANOVA Formula : sat ~ color

Model Summary

[Model Checking Plots](#)

ANOVA

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
Model	1	62.0550	62.0550	6.4593	0.0119
Residuals	171	1642.8120	9.6071	NA	NA
Total	172	1704.8671	NA	NA	NA

Specific Results

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
color	1	62.0550	62.0550	6.4593	0.0119
Residuals	171	1642.8120	9.6071	NA	NA
Total	172	1704.8671	NA	NA	NA

☒ Kruskal-Wallis Test

Kruskal-Wallis Results

Statistic	df	p-value
10.7692	3.0000	0.0130

ANOVA

[Albatross Analytics](#)[Data Import](#)[Data Management ▾](#)[Basic Analysis ▾](#)[Regression ▾](#)[Random Effect Model ▾](#)[Survival Analysis ▾](#)

ANOVA

[Run](#)

Variable (Numeric Only)

sat ▾

Group Variable

color spine color:spine

Make Interaction Variable

color spine

[Append](#)

☐ Shapiro-Wilk Test (Normality)

☐ Breusch-Pagan Test

ANOVA Formula : NULL

[Model Summary](#)[Model Checking Plots](#)

ANOVA

[Run](#)

Variable (Numeric Only)

sat ▾

Group Variable

Make Interaction Variable

color spine

[Append](#)☐ Shapiro-Wilk Test (Normality)☐ Breusch-Pagan Test

ANOVA Formula : sat ~ color+spine+color:

[Model Summary](#)[Model Checking Plots](#)

ANOVA

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
Model	11	147.8008	13.4364	1.3893	0.1824
Residuals	161	1557.0662	9.6712	NA	NA
Total	172	1704.8671	NA	NA	NA

Specific Results

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
color	3	67.5212	22.5071	2.3272	0.0766
spine	2	19.2533	9.6266	0.9954	0.3718
color:spine	6	61.0263	10.1711	1.0517	0.3941
Residuals	161	1557.0662	9.6712	NA	NA
Total	172	1704.8671	NA	NA	NA

Frequency Analysis

Frequency Analysis

Run

Row Variable

color

☒ Use Column Variable

Column Variable

y

☒ Row Percent

☐ Column Percent

☐ Percent

☒ Chi-squared Test

Row Variable : color , Colu

Data Summary

Chart

Frequency Table

	0	1	Total
1	3	9	12
2	26	69	95
3	18	26	44
4	15	7	22
Total	62	111	173

Row Percent Table

	0	1
1	25.0000	75.0000
2	27.3684	72.6316
3	40.9091	59.0909
4	68.1818	31.8182

☒ Chi-squared Test

☒ Fisher Exact Test

Pearson's Chi-squared

X-squared	df	p-value
14.0775	3.0000	0.0028

Fisher Exact Test

p-value	Alternative
0.0032	two.sided

Correlation Analysis

Correlation Analysis

Run

Variable

crab
y

Selected

sat
weight
width
colspi

→

←

Coefficient Type

Pearson

Plot Type

Scatter

Correlation Analysis Results

Correlation Results [Correlation Plots](#)

Correlation Matrix

	sat	weight	width	colspi
sat	1.0000	0.3692	0.3399	-0.1479
weight	0.3692	1.0000	0.8869	-0.2552
width	0.3399	0.8869	1.0000	-0.2236
colspi	-0.1479	-0.2552	-0.2236	1.0000

Sample Size

	Count
Total	173

Probability Values

	sat	weight	width	colspi
sat	-0.0000	0.0000	0.0000	0.0521
weight	0.0000	-0.0000	0.0000	0.0021
width	0.0000	0.0000	-0.0000	0.0062
colspi	0.0521	0.0007	0.0031	-0.0000

Coefficient Type

Spearman



Correlation Matrix

	sat	weight	width	colspi
sat	1.0000	0.4048	0.3744	-0.2115
weight	0.4048	1.0000	0.8991	-0.2509
width	0.3744	0.8991	1.0000	-0.2130
colspi	-0.2115	-0.2509	-0.2130	1.0000

Sample Size

	Count
Total	173

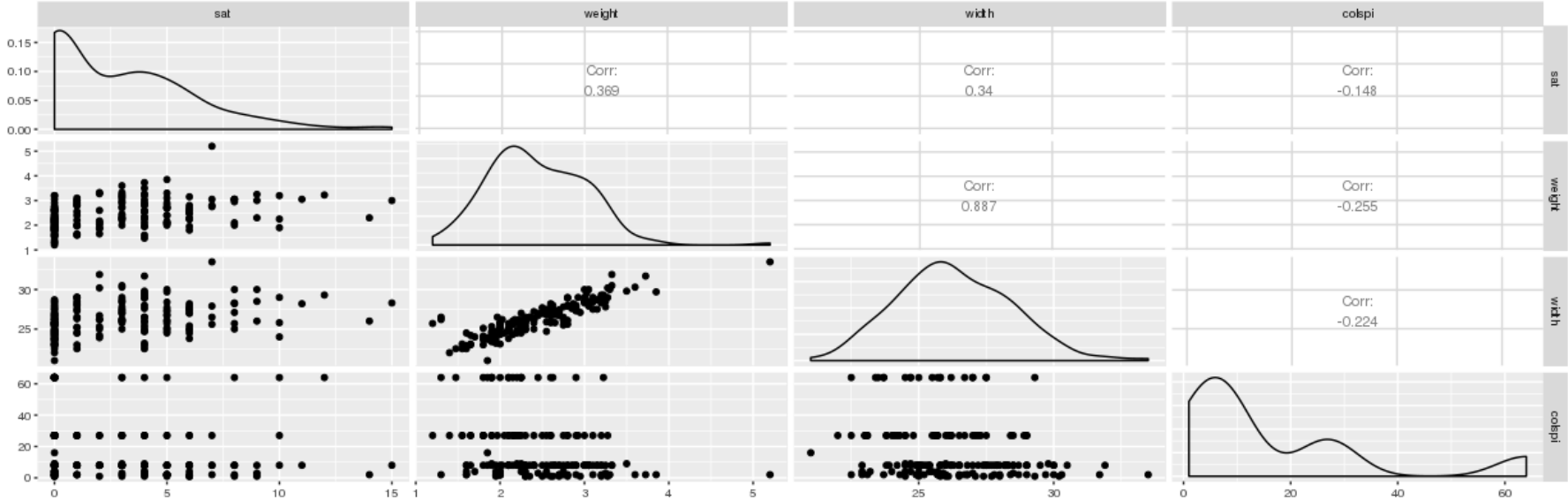
Probability Values

	sat	weight	width	colspi
sat	-0.0000	0.0000	0.0000	0.0098
weight	0.0000	-0.0000	0.0000	0.0026
width	0.0000	0.0000	-0.0000	0.0098
colspi	0.0052	0.0009	0.0049	-0.0000

Correlation Analysis Results

Correlation Results

Correlation Plots



2. Linear Model

반응변수 (Response) y 는 다음 3가지 조건을 만족

y : $n \times 1$ 열벡터 (column vector)

① 정규성 (normality) : $y \sim \text{normal}$

② 선형가법성 (linear additivity) : $\mu = E(y) = X\beta$

μ : y 의 평균, X : 모형행렬 (model matrix), β : 모수벡터

③ 등분산성 (constant variance) : $\text{var}(y) = \phi I$ (free of μ)

crackgrowth data(crackgrowth(page72).csv)

- y : increment of crack length measured on a compact tension steel test
- crack0 : initial value of crack
- cycle : number of cycles/ 10^6

Model

$$y = \beta_0 + \beta_1 \text{crack0} + \beta_2 \text{cycle} + e$$

	ID ↕	y ↕	crack0 ↕	specimen ↕	cycle ↕	phi ↕	lambda ↕
1	1	0.05	0.9	1	0.01	1	1
2	2	0.05	0.95	1	0.02	1	1
3	3	0.05	1	1	0.03	1	1
4	4	0.07	1.05	1	0.04	1	1
5	5	0.07	1.12	1	0.05	1	1

Linear Model

Run

Model

y ~ crack0+cycle

Response Variable

y

Variable

ID
y
specimen
phi
lambda

Selected

crack0
cycle

Interaction

Model Summary

Model Summary

Model Checking Plot

Prediction

Model: y ~ crack0+cycle

ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
crack0	1	0.20821	0.20821	1796.20648	0.00000
cycle	1	0.03158	0.03158	272.41154	0.00000
Residuals	238	0.02759	0.00012	NA	NA

Coefficients

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.24423	0.00806	-30.32044	0.00000
crack0	0.31493	0.00959	32.82705	0.00000
cycle	-0.81232	0.04922	-16.50489	0.00000

Model Summary

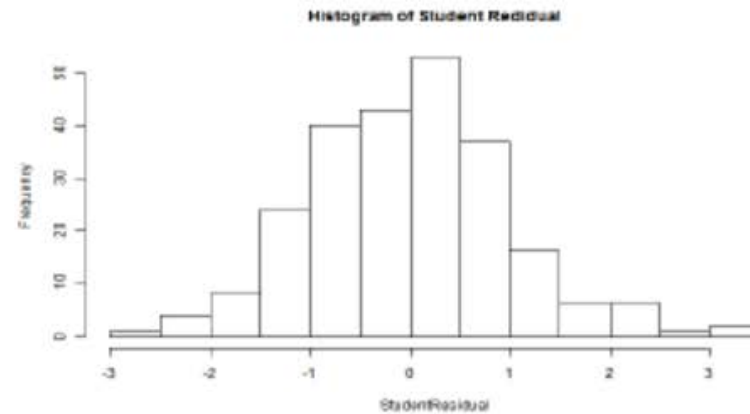
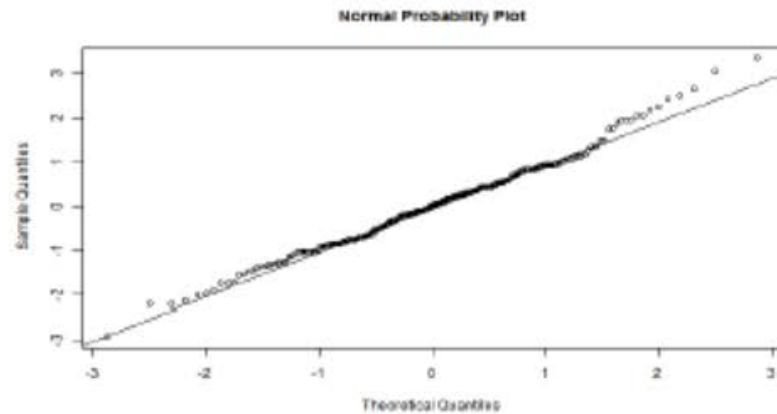
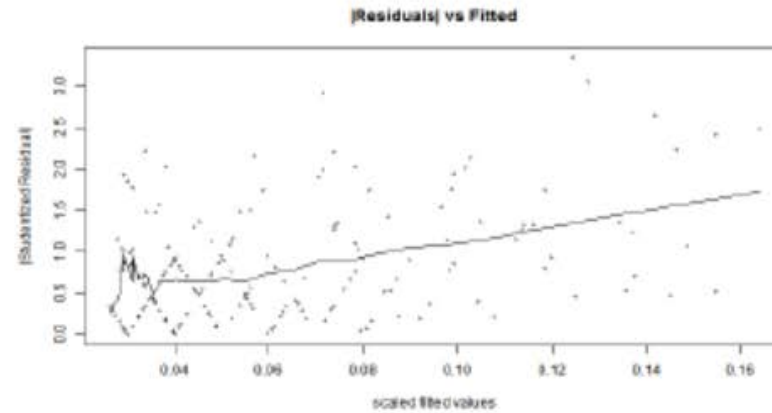
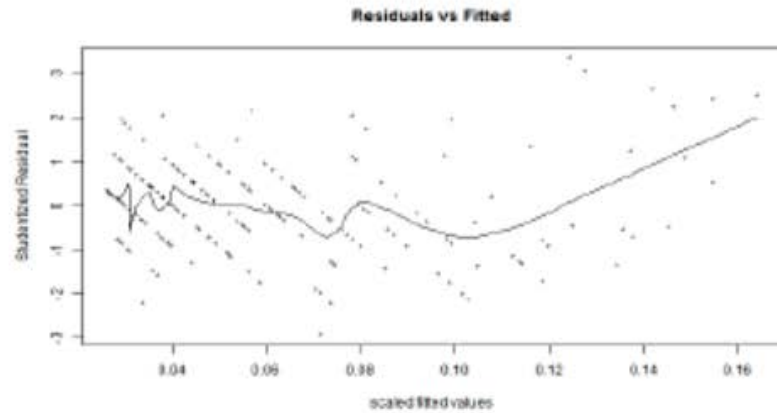
Num of obs	241.00000
F(2, 238)	1034.30901
Prob > F	0.00000
R-squared	0.89682
Adj R-squared	0.89595
Residual Std.Error	0.01077

Model Checking Plot

Model Summary

Model Checking Plot

Prediction



Normal Q-Q plot
: 정규성, 이상점

Residuals vs Fitted
: 등분산성

Ozone data(ozone.csv)

- Ozone relation to meteorological variables
- data gives ozone concentration(y in relation to nine meteorological variables (x_1 x_9), there being 330 units. These data have been used by Breiman (1995), among others. He derived the following four regression models for y .)

Model

- model1 : $x^6 + x_4^2 + x_6^2 + x_2x_4 + x_2x_5$
- model2 : $x^1 + x_5 + x_2^2 + x_4^2x_6^2 + x_2x_4 + x_5x_7$
- model3 : $x^2 + x_4x_5 + x_6^2 + x_2x_4 + x_2x_5$
- model4 : $x^1 + x_2 + x_4 + x_5 + x_6 + x_7x_2^2 + x_4^2x_6^2 + x_2x_4 + x_5x_7$

Upload your data file

Browse...

Ozone.csv

Upload manually

☒ Header

separator

Comma

Delete

Add New

Edit Data

Data Selection

☒ single ☐ multiple

Show 10 entries

Search:

	x1	x2	x3	x4	x5	x6	x7	x8	x9	y
1	5710	4	28	40	2693	-25	87	250	3	3
2	5700	3	37	45	590	-24	128	100	4	5
3	5760	3	51	54	1450	25	139	60	5	5
4	5720	4	69	35	1568	15	121	60	6	6
5	5790	6	19	45	2631	-33	123	100	7	4
6	5790	3	25	55	554	-28	182	250	8	4
7	5700	3	73	41	2083	23	114	120	9	6
8	5700	3	59	44	2654	-2	91	120	10	7
9	5770	8	27	54	5000	-19	92	120	11	4
10	5720	3	44	51	111	9	173	150	12	6

Showing 1 to 10 of 330 entries

Previous

1

2

3

4

5

..

☒ Make Variable

New Variable Name

x5x7

Make Variable

x5*x7

variable

operator

☐ Add Variable

	x1	x2	x3	x4	x5	x6	x7	x8	x9	y	phi	x4x4	x6x6	x2x4	x2x5	x2x
1	5710	4	28	40	2693	-25	87	250	3	3	1	1600	625	160	10772	1
2	5700	3	37	45	590	-24	128	100	4	5	1	2025	576	135	1770	
3	5760	3	51	54	1450	25	139	60	5	5	1	2916	625	162	4350	
4	5720	4	69	35	1568	15	121	60	6	6	1	1225	225	140	6272	1

Model 1

Linear Model

Run

Model

$y \sim x6 + x4x4 + x6x6 + x2x4 + x2x5$

Response Variable

y

Variable

x1
x2
x3
x4
x5
x7
x8
x9
y
phi
x2x2
x5x7

x6
x4x4
x6x6
x2x4
x2x5

→
←

Interaction

☐ Offset Variable

☐ Comparison with other model

☐ Margins

Model Summary

Model Checking Plot

Prediction

Model: y ~ x6+x4x4+x6x6+x2x4+x2x5

ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x6	1	967.42046	967.42046	48.18482	0.00000
x4x4	1	12618.98881	12618.98881	628.52065	0.00000
x6x6	1	623.61224	623.61224	31.06058	0.00000
x2x4	1	67.12079	67.12079	3.34312	0.06841
x2x5	1	333.22309	333.22309	16.59702	0.00006
Residuals	324	6505.04067	20.07729	NA	NA

Coefficients

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.59144	0.86650	1.83664	0.06718
x6	0.04378	0.00877	4.99197	0.00000
x4x4	0.00280	0.00023	11.97586	0.00000
x6x6	-0.00092	0.00018	-5.21183	0.00000
x2x4	0.00393	0.00280	1.40413	0.16124
x2x5	-0.00012	0.00003	-4.07394	0.00006

Model Summary

Num of obs	330.00000
F(5, 324)	145.54124
Prob > F	0.00000
R-squared	0.69193
Adj R-squared	0.68717
Residual Std. Error	4.48077
AIC	1934.30863

Model 2

Linear Model Run

Model
 $y \sim x1+x5+x2x2+x4x4+x6x6+x2x4+x5x7$

Response Variable
Y

Variable

x2
x3
x4
x7
x8
x9
y
phi
x6
x2x5

→

←

Selected

x1
x5
x2x2
x4x4
x6x6
x2x4
x5x7

Interaction

☐ Offset Variable
☐ Comparison with other model
☐ Margins
☐ VIF

Model Summary Model Checking Plot Prediction

Model: $y \sim x1+x5+x2x2+x4x4+x6x6+x2x4+x5x7$

ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	7788.76560	7788.76560	386.74904	0.00000
x5	1	2268.32619	2268.32619	112.63312	0.00000
x2x2	1	290.51496	290.51496	14.42544	0.00017
x4x4	1	4087.07969	4087.07969	202.94283	0.00000
x6x6	1	118.33502	118.33502	5.87589	0.01590
x2x4	1	2.55049	2.55049	0.12664	0.72217
x5x7	1	75.05395	75.05395	3.72678	0.05442
Residuals	322	6484.78016	20.13907	NA	NA

Coefficients

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	17.09917	27.11466	0.63062	0.52873
x1	-0.00274	0.00480	-0.57151	0.56806
x5	-0.00058	0.00027	-2.14195	0.03295
x2x2	0.00319	0.01358	0.23472	0.81458
x4x4	0.00342	0.00029	11.63760	0.00000
x6x6	-0.00036	0.00015	-2.36319	0.01871
x2x4	-0.00016	0.00039	-0.04656	0.96290
x5x7	-0.00000	0.00000	-1.93049	0.05442

Model Summary

Num of obs	330.00000
F(7, 322)	103.78282
Prob > F	0.00000
R-squared	0.69289
Adj R-squared	0.68621
Residual Std.Error	4.48766

AIC 1937.27921

- 29 -

Model 3

Linear Model

Run

Model

$y \sim x2 + x4 + x5 + x6 + x4x4 + x6x6 + x2x4 + x2x5$

Response Variable

y

Variable

x3
x7
x8
x9
y
phi
x1
x2x2
x5x7

Selected

x2
x4
x5
x6
x4x4
x6x6
x2x4
x2x5

Interaction

☐ Offset Variable

☐ Comparison with other model

☐ Margins

☐ VIF

☐ Robust standard errors

☐ Confidence intervals for coefficients

Model Summary

Model Checking Plot

Prediction

Model: $y \sim x2 + x4 + x5 + x6 + x4x4 + x6x6 + x2x4 + x2x5$

ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x2	1	3.79914	3.79914	0.20401	0.65181
x4	1	12868.83170	12868.83170	691.03646	0.00000
x5	1	960.32294	960.32294	51.56786	0.00000
x6	1	165.90159	165.90159	8.90866	0.00306
x4x4	1	703.32800	703.32800	37.76763	0.00000
x6x6	1	350.58579	350.58579	18.82592	0.00002
x2x4	1	19.08001	19.08001	1.02457	0.31220
x2x5	1	65.73200	65.73200	3.52971	0.06118
Residuals	321	5977.82490	18.62251	NA	NA

Coefficients

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.35229	5.64935	1.12443	0.26167
x2	1.46658	0.72790	2.01481	0.04476
x4	-0.23283	0.14230	-1.63618	0.10278
x5	-0.00036	0.00037	-0.97104	0.33226
x6	0.03709	0.00654	4.34287	0.00002
x4x4	0.00540	0.00101	5.35182	0.00000
x6x6	-0.00077	0.00017	-4.42060	0.00001
x2x4	-0.01897	0.00961	-1.97421	0.04922
x2x5	-0.00014	0.00008	-1.87875	0.06118

Model Summary

Num of obs	330.00000
F(8, 321)	101.60810
Prob > F	0.00000
R-squared	0.71690
Adj R-squared	0.70984
Residual Std. Error	4.31538

AIC 1912.41683

Model 4

Linear Model

Model

$y \sim x1 + x2 + x4 + x5 + x6 + x7 + x2x2 + x4x4 + x6x6 + x2x4 + x5x7$

Response Variable

y

Variable

x3

x6

x9

y

phi

c2x5

x1

x2

x4

x5

x6

x7

x2x2

x4x4

x6x6

x2x4

x5x7

Interaction

☐ Other Variable
☐ Comparison with other model
☐ Marginal
☐ VIF
☐ Robust standard errors
☐ Confidence intervals for coefficients
☐ Breusch-Pagan test

Model Summary Model Checking Plot Prediction

Model: $y \sim x1 + x2 + x4 + x5 + x6 + x7 + x2x2 + x4x4 + x6x6 + x2x4 + x5x7$

ANOVA

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	7788.76560	7788.76560	420.23338	0.00000
x2	1	406.54096	406.54096	21.93442	0.00000
x4	1	4708.33779	4708.33779	254.03264	0.00000
x5	1	991.14313	991.14313	53.67592	0.00000
x6	1	109.68713	109.68713	5.91804	0.01554
x7	1	107.45253	107.45253	5.79967	0.01862
x2x2	1	0.15915	0.15915	0.00859	0.92623
x4x4	1	621.69093	621.69093	33.54258	0.00000
x6x6	1	426.46190	426.46190	23.00923	0.00000
x2x4	1	32.11465	32.11465	1.73271	0.18901
x5x7	1	29.11915	29.11915	1.57009	0.21997
Residuals	318	5893.92316	18.53438	NA	NA

Coefficients

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-27.53582	21.06789	-0.08623	0.97611
x1	0.06634	0.00558	1.13557	0.25689
x2	0.63880	0.00163	1.06191	0.28908
x4	-0.21179	0.15756	-1.99158	0.04727
x5	-0.00010	0.00068	-0.29124	0.81284
x6	0.05406	0.01130	4.94261	0.00000
x7	0.03062	0.01539	1.98926	0.04753
x2x2	-0.01206	0.01833	-0.65800	0.51100
x4x4	0.00475	0.00121	3.91291	0.00011
x6x6	-0.00086	0.00038	-4.77101	0.00000
x2x4	-0.00962	0.00836	-1.13122	0.25953
x5x7	-0.00000	0.00060	-1.25343	0.21697

Model Summary

Num of obs 330.00000

F(11, 318) 74.65964

Prob > F 0.00000

R-squared 0.72087

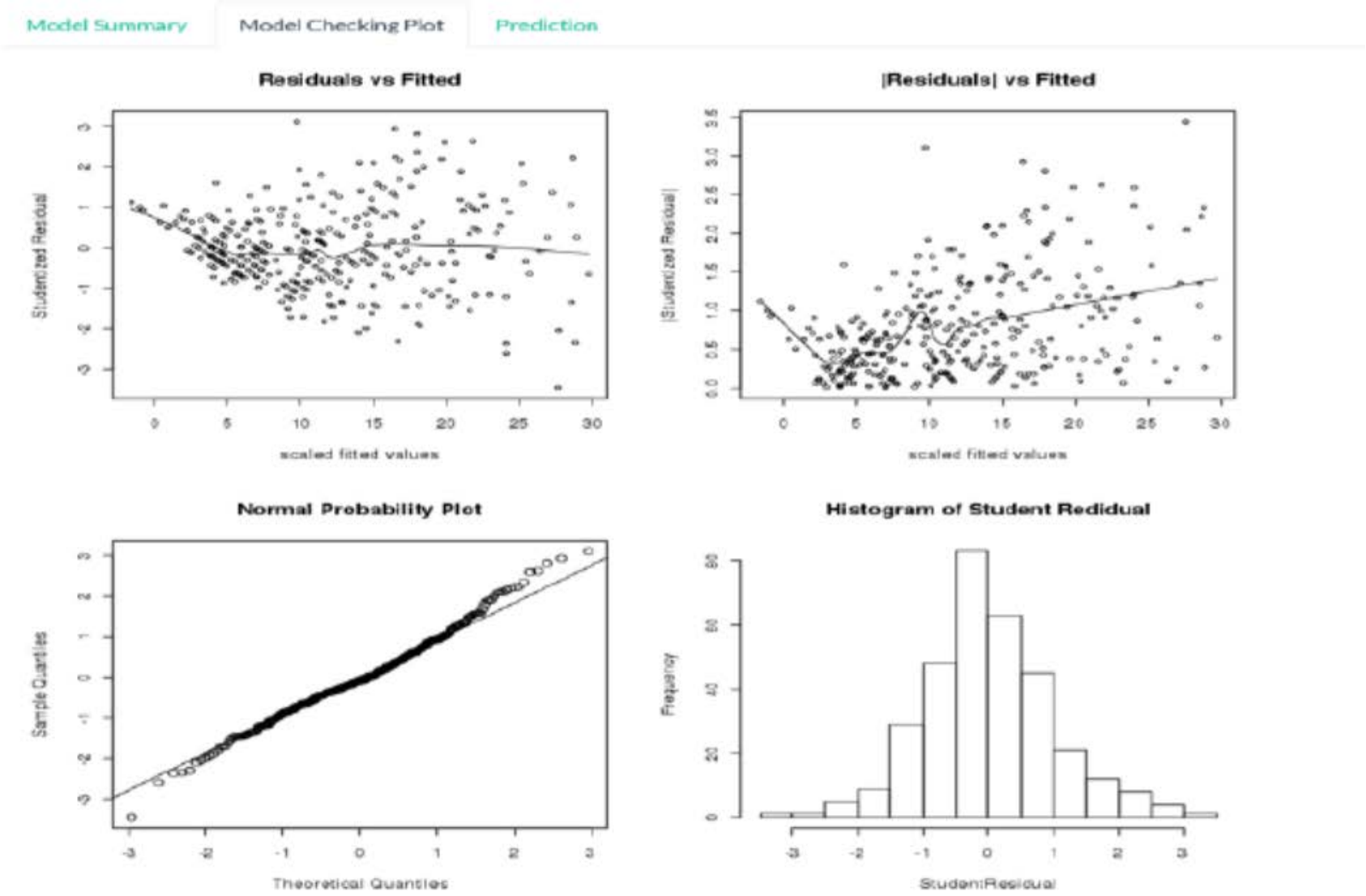
Adj R-squared 0.71122

Residual Std. Error 4.30516

AIC 1913.75287

model	Model 1	Model 2	Model 3	Model 4
AIC	1934.3	1937.3	1912.4	1913.8

Residual plots for Model 3



3. Generalized Linear Model

반응변수 y 가 이산형 (비율 또는 개수) 일때는 회귀분석의 3가지 가정을 만족하지 않는다.

	<i>Proportion</i>	<i>Count</i>
정규성	$y \sim B(m, p)$	$y \sim \text{Poisson}(\mu)$
선형가법성	$0 < \mu = mp \neq X\beta \in \mathbb{R}^n$	$0 < \mu \neq X\beta \in \mathbb{R}^n$
등분산성	$\text{var}(y) = mp(1-p)$	$\text{var}(y) = \mu$

y 가 포아송 분포를 따른다고 할 때 세 조건을 동시에 만족시키는 자료 변환은 존재하지 않는다.

$y \sim \text{Poisson}(\mu)$

- ① 정규성 : $y^{2/3}$
- ② 가법성 : $\log(y)$
- ③ 등분산성 : $y^{1/2}$

① $y \sim$ 지수족 (exponential family)

: 정규, 이항, 포아송, 감마분포 등

② $\eta = g(\mu) = X\beta$

g : 연결함수 (link function), $\mu = E(y)$

η : 선형예측변수 (linear predictor)

③ $\text{var}(y) = \phi V(\mu)$

ϕ : $\text{var}(y)$ 중 μ 와는 상관없는 산포(dispersion) 모수

$V(\mu)$: 분산함수 (variance function)

GLM classes

(1) $y \sim B(n,p)$: 비율 (proportion) 자료

1) logit : $\log(p/(1-p))=X\beta \rightarrow$ 로짓모형 (logit model)

2) probit : $\Phi^{-1}(p) = X\beta$, $\Phi()$: cumulative distribution of $N(0,1)$

3) complimentary log-log : $\log(-\log(1-p))=X\beta$

(2) $y \sim \text{Poisson}(\mu)$: 개수 (count) 자료

1) log : $\log(\mu) = X\beta \rightarrow$ 로그선형모형 (log-linear model)

(3) $y \sim \text{gamma}(\mu)$: 양의 자료. 평균이 증가하면서 분산이 증가.

1) inverse : $1/\mu = X\beta$

2) log : $\log(\mu) = X\beta$

Ozone continued

GLM

Run

Model (e.g. $y \sim x + (x^2)$)

 $y \sim x2 + x4 + x7 + x8 + x9 + x8x8 + x9x9$

Response Variable

y

Variable

Selected

x1
x3
x5
x6
y

x2
x4
x7
x8
x9
x8x8
x9x9

Interaction

Distribution

gamma

☐ Binomial denominator

Link Function

log

☐ No intercept model

☐ Offset Variable

☐ Comparison with other model

Model Summary Model Checking Plot Prediction

Model: $y \sim x2 + x4 + x7 + x8 + x9 + x8x8 + x9x9$

Model Summary

	data	Num of obs	130.00000
Family	Gamma	Res. deviance(df=322)	42.19945
Link	log	Null deviance(df=329)	167.03022
Optimization	RWLS	Loglikelihood	-862.69038
Num of iteration	5	AIC	1743.38076
		BIC	1777.57260

Coefficients

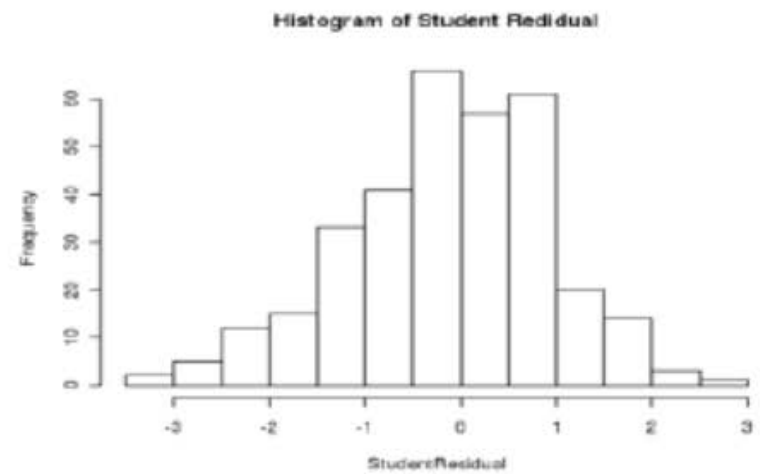
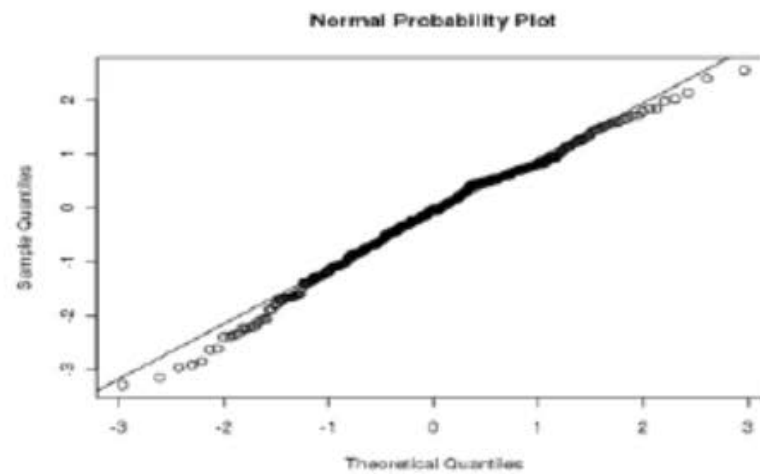
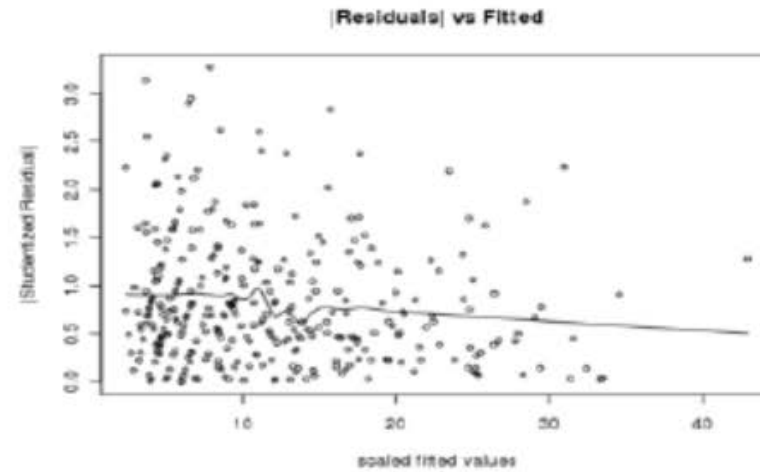
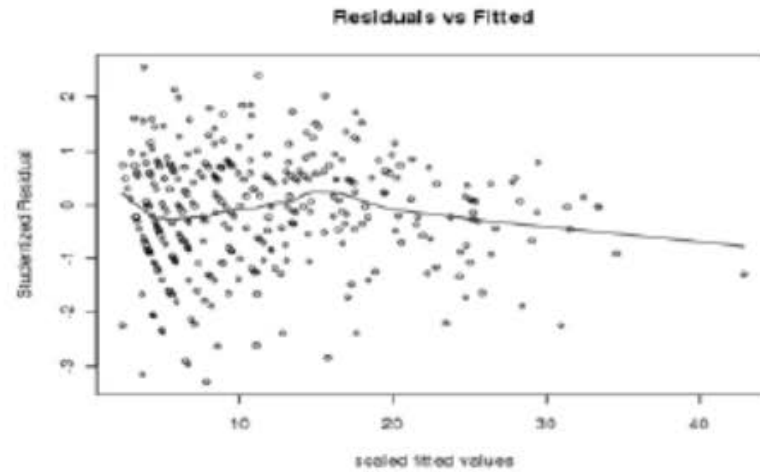
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.21151	0.13667	8.86439	0.00000
x2	-0.03010	0.00955	-3.15300	0.00177
x4	0.01000	0.00330	3.01061	0.00201
x7	0.00335	0.00055	6.05534	0.00000
x8	-0.00525	0.00089	-5.92531	0.00000
x9	0.00461	0.00107	8.94541	0.00000
x8x8	0.00001	0.00000	4.24622	0.00003
x9x9	-0.00003	0.00000	-10.30293	0.00000

Model Checking Plot

Model Summary

Model Checking Plot

Prediction



Cancer data(Cancer.csv)

- survival and death for 539 males diagnosed with lung cancer
- count : number of deaths from lung cancer
- Histology : type I, II, III
- stage : stage of disease (value : 1, 2, 3)
- time : follow-up time interval after the diagnosis grouped into 2 month
- risktime : total time at risk
- model : $\eta = \log \mu_{ijk} = \log t_{ijk} + \beta_0 + \beta_i^H + \beta_j^S + \beta_k^T$

	time ⌄	histology ⌄	stage ⌄	count ⌄	risktime ⌄
1	1	1	1	9	157
2	1	2	1	5	77
3	1	3	1	1	21
4	2	1	1	2	139
5	2	2	1	2	68

☒ Make Variable

New Variable Name

logrisktime

Make Variable

log(risktime)

variable

risktime

operator

log

☒ Add Variable

```
[1] "log(risktime)"
```

```
time histology stage count risktime logrisktime
1    1         1     1     9     157    5.056246
2    1         2     1     5     77    4.343805
```


Convert factor variables

- ☒ time
- ☒ histology
- ☒ stage
- ☐ count
- ☐ risktime
- ☐ logrisktime

Label

Name	Type	Label
time	factor	NA
histology	factor	NA
stage	factor	NA
count	integer	NA
risktime	integer	NA
logrisktime	numeric	NA

GLM

Model (e.g. $y \sim x + I(x^2)$)

1 count ~ histology+stage+time

Response Variable

count

Variable

count

risktime

logrisktime

Selected

histology

stage

time

Interaction

2

Distribution

poisson

☐ Binomial denominator

Link Function

log

☐ No intercept model

☒ 3 Offset Variable

Offset Variable

logrisktime

☐ Comparison with other model

☐ Margins

☐ VIF

☐ Robust standard errors

☐ Confidence intervals for coefficients

☐ Exponential scale

4

Run

Model Summary

Model: count ~ histology+stage+time

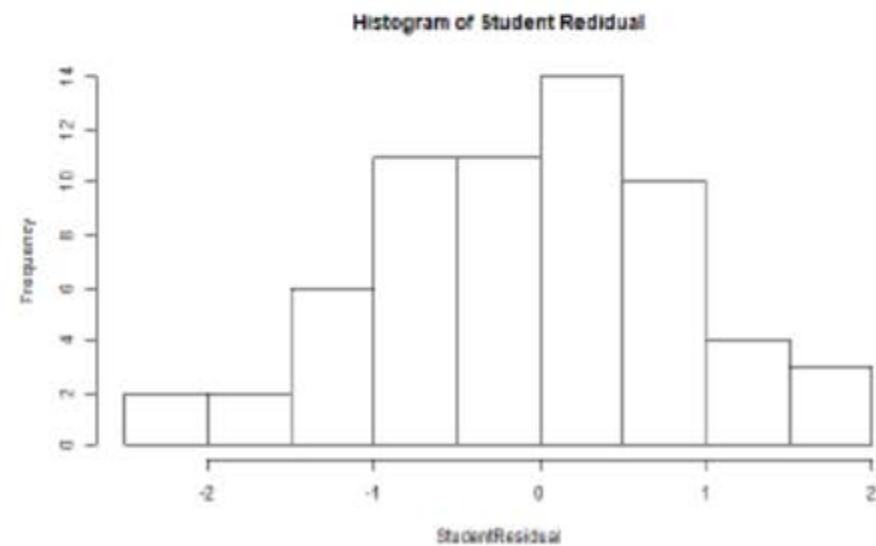
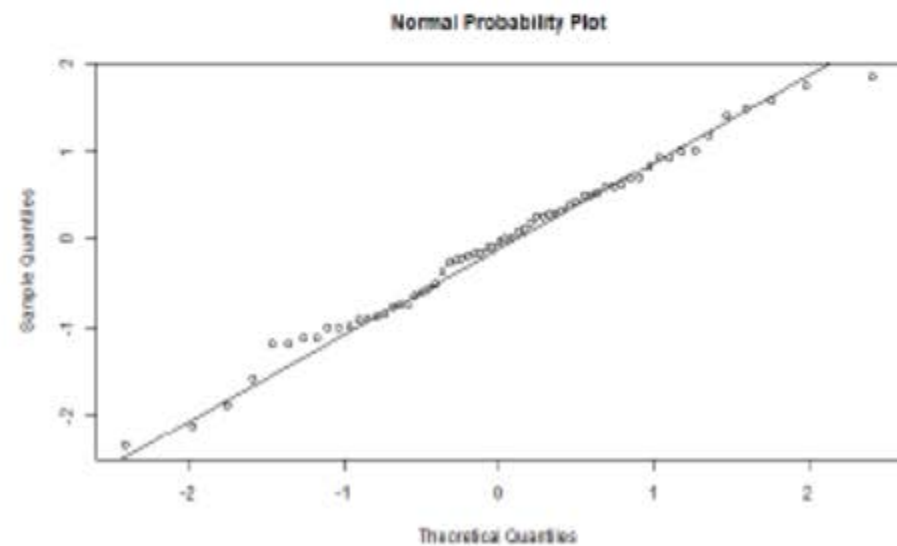
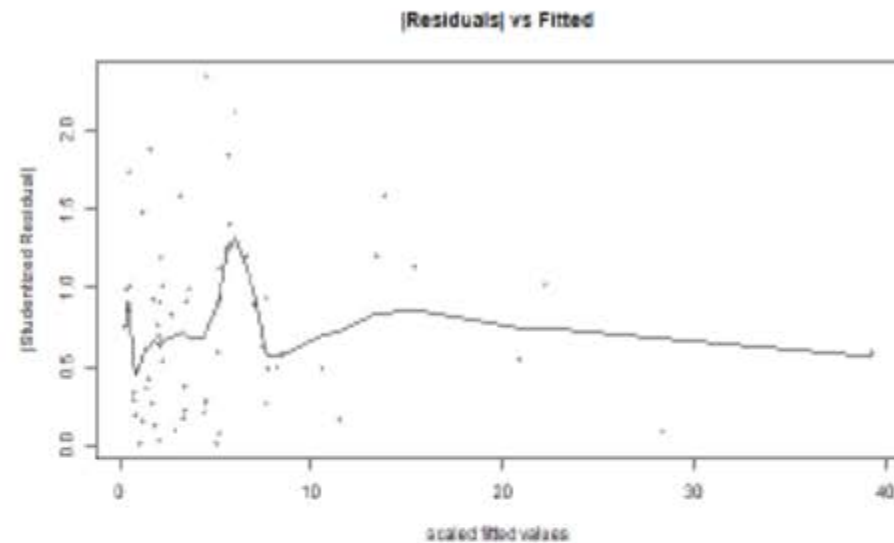
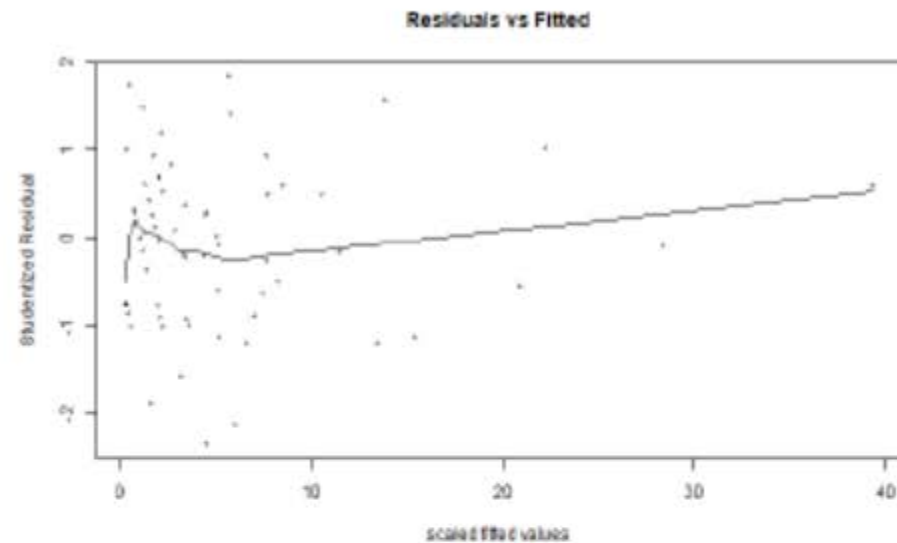
Model Summary

	data		
Family	poisson	Num of obs	63.00000
Link	log	Res. deviance(df=52)	43.92253
Optimization	IWLS	Null deviance(df=62)	175.71781
Num of iteration	5	Log likelihood	-114.87133
		AIC	251.74266
		BIC	275.31715

Coefficients

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.00928	0.16651	-18.07254	0.00000
histology2	0.16244	0.12195	1.33202	0.18285
histology3	0.10754	0.14745	0.72933	0.46580
stage2	0.47001	0.17444	2.69439	0.00705
stage3	1.32431	0.15205	8.70949	0.00000
time2	-0.12745	0.14908	-0.85494	0.39259
time3	-0.07973	0.16352	-0.48758	0.62585
time4	0.11892	0.17107	0.69518	0.48694
time5	-0.66511	0.26061	-2.55210	0.01071
time6	-0.35015	0.24348	-1.43810	0.15040
time7	-0.17518	0.24985	-0.70115	0.48321

Model Checking Plot



Predictions

Model Summary				Model Checking Plot				Prediction			
time	histology	stage	count	risktime	logrisktime	pred	pred95LL	pred95UL	StudentResidual	leverage	cook
1	1	1	9	157	5.06	7.74	5.59	10.73	0.50	0.21	0.01
1	2	1	5	77	4.34	4.47	3.18	6.29	0.27	0.14	0.00
1	3	1	1	21	3.04	1.15	0.77	1.73	-0.15	0.05	0.00
2	1	1	2	139	4.93	6.04	4.26	8.56	-2.13	0.19	0.07
2	2	1	2	68	4.22	3.47	2.41	5.01	-0.92	0.12	0.01
2	3	1	1	17	2.83	0.82	0.54	1.25	0.19	0.04	0.00
3	1	1	9	126	4.84	5.74	3.98	8.27	1.40	0.20	0.05
3	2	1	3	63	4.14	3.38	2.30	4.95	-0.22	0.13	0.00
3	3	1	1	14	2.64	0.71	0.46	1.10	0.33	0.03	0.00
4	1	1	10	102	4.62	5.67	3.91	8.21	1.84	0.20	0.10
4	2	1	2	55	4.01	3.59	2.43	5.33	-0.99	0.14	0.01
4	3	1	1	12	2.48	0.74	0.48	1.15	0.29	0.04	0.00
5	1	1	1	88	4.48	2.23	1.31	3.79	-1.01	0.16	0.01
5	2	1	2	50	3.91	1.49	0.86	2.58	0.42	0.12	0.00
5	3	1	0	10	2.30	0.28	0.16	0.50	-0.76	0.02	0.00
6	1	1	3	82	4.41	2.85	1.74	4.68	0.10	0.18	0.00
6	2	1	2	45	3.81	1.84	1.10	3.07	0.12	0.13	0.00
6	3	1	1	8	2.08	0.31	0.18	0.54	0.99	0.02	0.00
7	1	1	1	76	4.33	3.15	1.91	5.19	-1.59	0.21	0.04
7	2	1	2	42	3.74	2.05	1.22	3.44	-0.03	0.14	0.00
7	3	1	0	6	1.79	0.28	0.16	0.48	-0.75	0.02	0.00
1	1	2	12	134	4.90	10.58	7.84	14.26	0.49	0.25	0.01
1	2	2	4	71	4.26	6.59	4.82	9.02	-1.20	0.17	0.02
1	3	2	1	22	3.09	1.93	1.33	2.82	-0.77	0.07	0.00
2	1	2	7	110	4.70	7.64	5.52	10.58	-0.27	0.21	0.00

	Death	Total
Treatment	41 (p_1)	733
Placebo	60 (p_2)	742

Binomial Distribution / Logit Model

$$X_1 \sim \text{Bin}(733, p_1), \quad X_2 \sim \text{Bin}(742, p_2)$$

$$\text{Logit Model : } \log\{p_1/(1-p_1)\}=\beta_0, \quad \log\{p_1/(1-p_1)\}=\beta_0+\beta_1$$

$$\text{Trt group의 odds (=사망확률/정상확률) : } p_1/(1-p_1)=\exp(\beta_0)$$

$$\text{Placebo group의 odds : } p_2/(1-p_2)=\exp(\beta_0+\beta_1)$$

$$\exp(\beta_1)=\frac{p_2/(1-p_2)}{p_1/(1-p_1)} : \text{OR (오즈비; Odds Ratio) : Trt 대비 Placebo의 사망에 대한 오즈비}$$

logistic1.csv

✕ Delete

+ Add New

🔧 Edit Data

Data Selection

☒ single ☐ multiple

Show

10

 entries

Search:

	y	n	trt
1	41	733	1
2	60	742	2

Showing 1 to 2 of 2 entries

Model (e.g. $y \sim x + I(x^2)$)

`cbind(y, n - y) ~ trt`

Response Variable

y

Variable

y
n

Selected

trt

Interaction

Distribution

binomial

☒ Binomial denominator

Binomial denominator

n

Link Function

logit

Model: cbind(y , n - y) ~ trt

Model Summary

	data
Family	binomial
Link	logit
Optimization	IWLS
Num of iteration	3

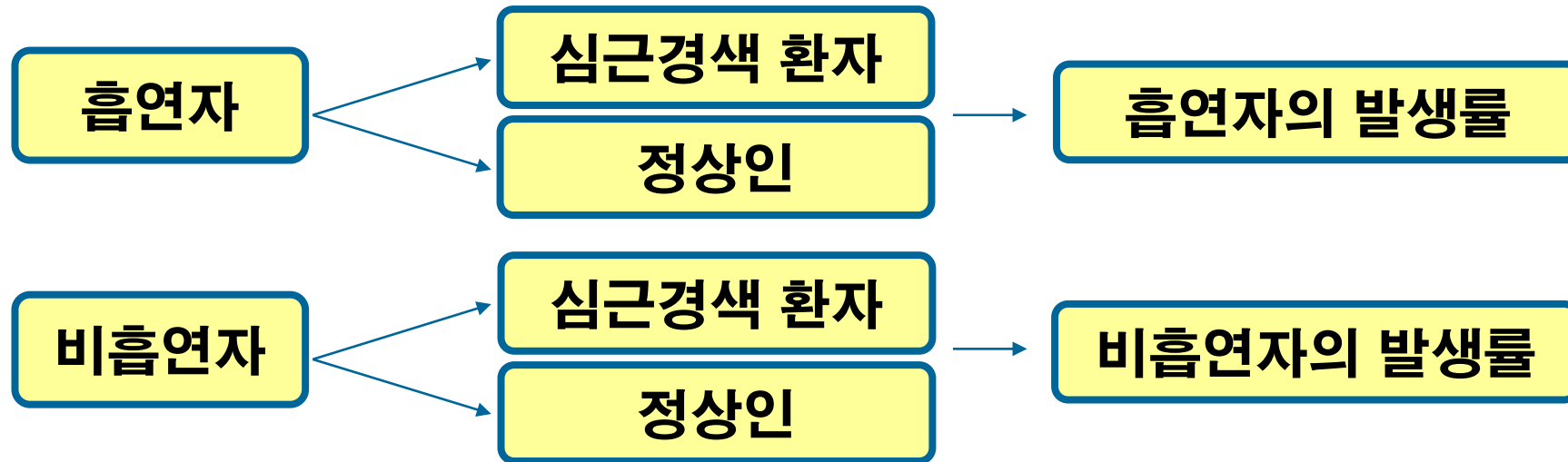
Coefficients

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.22134	0.34853	-9.24261	0.00000
trt	0.39533	0.20969	1.88534	0.05938

$$\begin{aligned} \text{OR} &= \exp(0.3953) \\ 95\% \text{ CI} &: \exp(0.3953 \pm 1.96 * 0.209) \\ \text{OR} &= 1.51 \text{ (0.99, 2.24)} \end{aligned}$$

Trt 는 placebo에 비하여 사망률에 대한 오즈를 1.51 (95% CI : 0.99, 2.24) 배 감소!

(1) 전향적 연구 (prosepctive study) : 집단별로 질병발생율을 조사

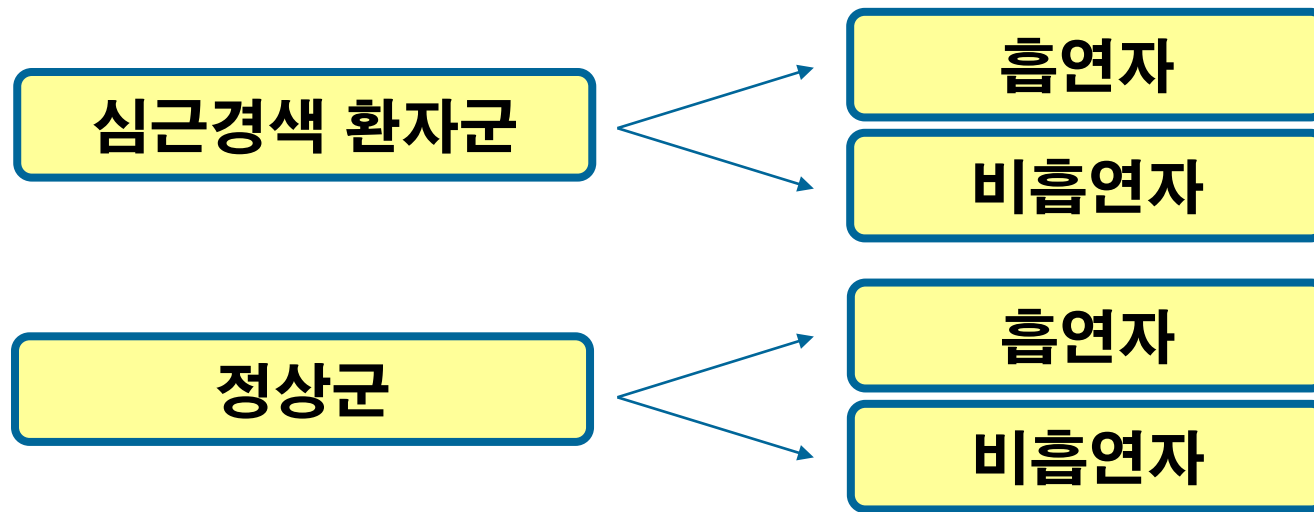


→ 상대위험도를 계산

◆ 전향적 연구의 장단점

- 장 점 : 집단간의 발생률을 바로 계산
 - 흡연자의 심근경색 발생률은 1%,
비흡연자의 심근경색 발생률은 0.1% 이다.
 - 상대위험도를 통해서 두 집단간 발생률을 비교
- 단 점
 - 많은 표본이 필요 : 질병 발생률은 낮기 때문에
(0.1% : 1/1,000, 10/10,000)
 - 조사 및 질병검사를 위한 시간과 비용이 많이 든다.

(2) 후향적 연구 (retrospective study) : 환자군과 정상군(대조군)의 과거를 되돌아보는 연구



◆ 후향적 연구의 예 : 흡연여부와 심근경색증

집단	심근경색	정상
흡연	170 (85%)	70 (35%)
비흡연	30	130
합계	200	200

- 심근경색 200 명 : 심근경색 전문의는 쉽게 얻을 수 있음
 - 정상 200 명 : 환자와 비슷한 성향의 일반인
- ➔ 총 표본수 : 400명 (전향적 연구 20,000명)

◆ 후향적 연구의 장단점

➤ 장 점

- 비용이 적게 든다 : 적의 수의 표본으로 연구가 가능

➤ 단 점

- 각 집단에서 질병 발생률을 바로 계산하지 못함
- 환자군의 흡연율 = 85%, 정상군의 흡연율 = 35%

→ 의학적으로는 관심 있는 비율이 아님

→ 관심 : 흡연, 비흡연 집단의 발생률 비교

→ 오즈비로 흡연, 비흡연 집단간 발생률은 비교할 수 있음

(2) 후향적 연구에서의 오즈비

- 환자 집단에서의 오즈 : $\frac{\pi_1}{1 - \pi_1}$
 - π_1 : 환자의 흡연율

- 정상 집단에서의 오즈 : $\frac{\pi_2}{1 - \pi_2}$
 - π_2 : 정상인의 흡연율

$$OR_2 = \frac{\text{환자 집단의 오즈}}{\text{정상 집단의 오즈}} = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)}$$

◆ 전향적 연구에서의 오즈비 = 후향적 연구에서의 오즈비

$$OR_2 = OR_1$$

예) 흡연과 심근경색의 비교 (후향적 연구)

$$OR_2 = \frac{0.85 / 0.15}{0.35 / 0.65} = 10.5 = OR_1 = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$

→ 정상집단보다 환자집단의 흡연에 대한 오즈비가 10.5배 높다.

→ 흡연자가 비흡연자보다 질병발생에 대한 오즈비가 10.5배 높다.

retro.csv

Show 10 entries

Search:

	y	n	x
1	170	240	1
2	30	160	2

Showing 1 to 2 of 2 entries

Model (e.g. $y \sim x + I(x^2)$)

`cbind(y, n - y) ~ x`

Response Variable

y

Variable

y

n

Selected

x

Interaction

Distribution

binomial

☒ Binomial denominator

Binomial denominator

n

Link Function

logit

Model: cbind(y , n - y) ~ x

Model Summary

	data
Family	binomial
Link	logit
Optimization	IWLS
Num of iteration	3

OR=exp(0.23536))
95% CI : exp(2.3536±1.96 * 0.2474)
OR=10.5 (6.5, 17.1)

Coefficients

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.24094	0.34885	9.29030	0.00000
x	-2.35364	0.24737	-9.51452	0.00000

4. Joint Model

- 식이요법에 따른 체중감량 (kg)

식이요법 1	식이요법 2
5.3	7
5.7	9.9
4.7	8.5
3.5	7.1
7.7	10.3
4.9	8.8
7.6	8.9
5.5	8.1
2.8	8.3
8.4	9.1

weight.csv

	y ↕	x ↕
1	5.3	1
2	5.7	1
3	4.7	1
4	3.5	1
5	7.7	1
6	4.9	1
7	7.6	1
8	5.5	1
9	2.8	1
10	8.4	1

- 평균에 대한 회귀분석

- 분산에 대한 회귀분석

Model

y ~ x

Response Variable

y ▼

Variable

y

Selected

x

→

←

Model

phi ~ 1 + x

Model for phi(residual variance)

phi ▼

Variable

y

Selected

x

→

←

Interaction

Coefficients of the model: $y \sim x$

link: identity

Dist: gaussian

	Estimate	Std. Error	t-value	p_val	LL	UL
(Intercept)	2.6200	1.2007	2.1820	0.0291	0.2666	4.9734
x	2.9900	0.6669	4.4834	0.0000	1.6829	4.2971

- 식이요법 2는 식이요법 1에 비하여 평균적으로 3kg 감량

Coefficients of the model: $\phi_i \sim 1 + x$

link: log

Dist: gaussian

	Estimate	Std. Error	t-value
(Intercept)	2.2846	1.0541	2.1673
x	-1.0836	0.6667	-1.6255

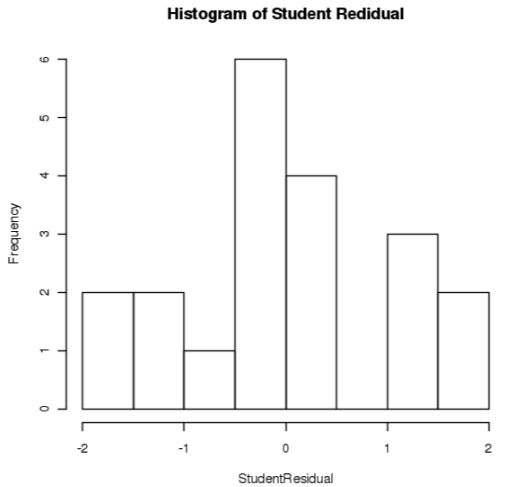
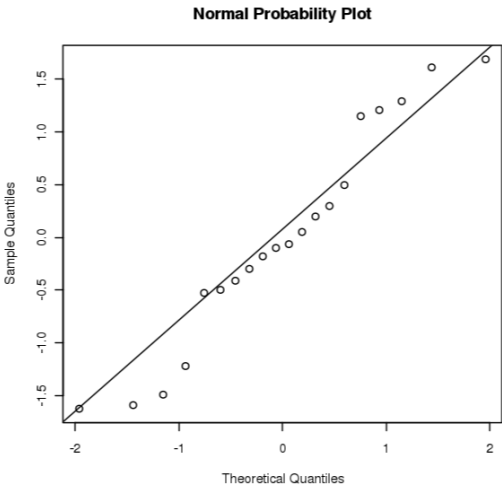
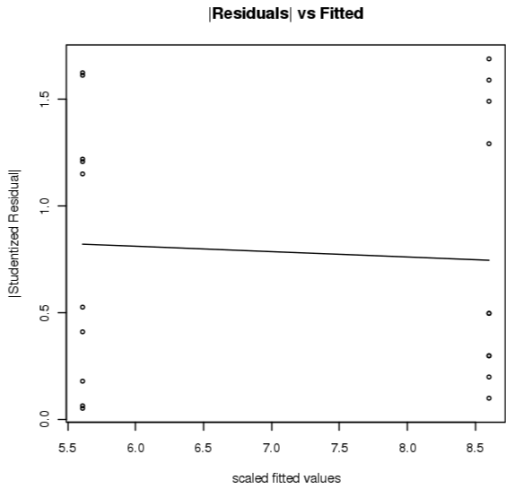
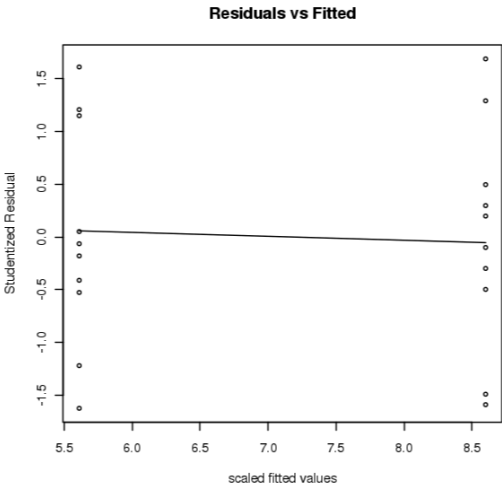
- 식이요법 2는 식이요법 1에 비하여 분산이 0.34배 (=exp(-1.084))

Likelihood Function Values and Conditional AIC

-2ML (-2 h) :	67.9398
-2RL (-2 p_beta (h)) :	67.5510
cAIC :	71.9398
Scaled Deviance :	18.0000
df :	18.0000

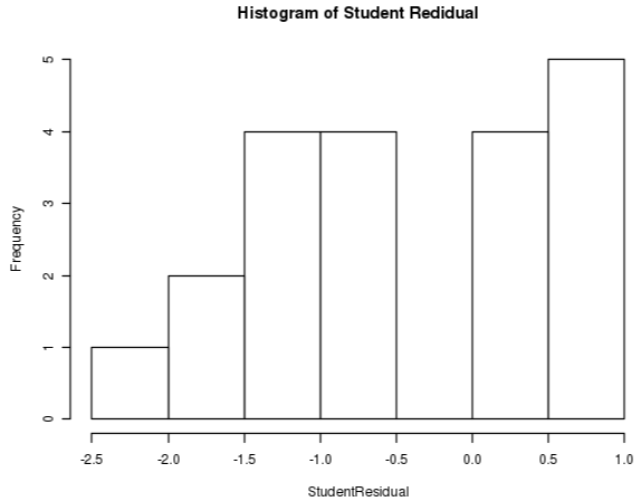
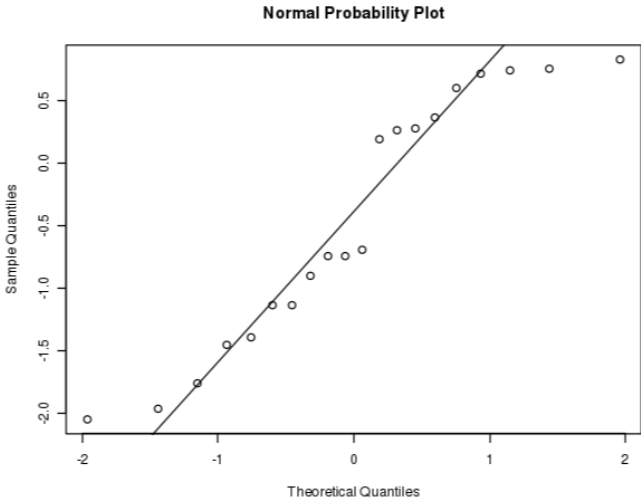
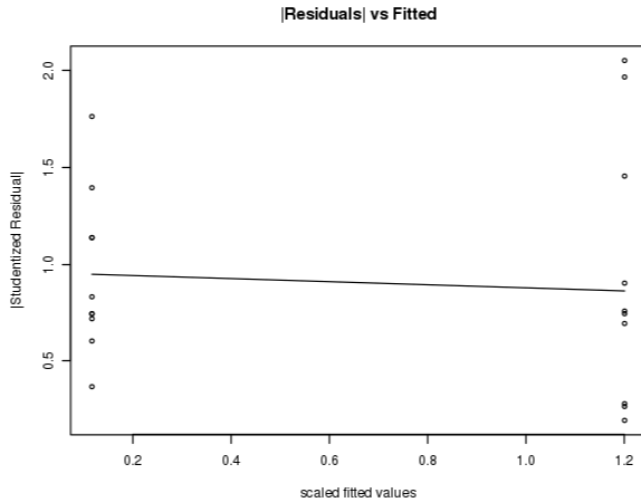
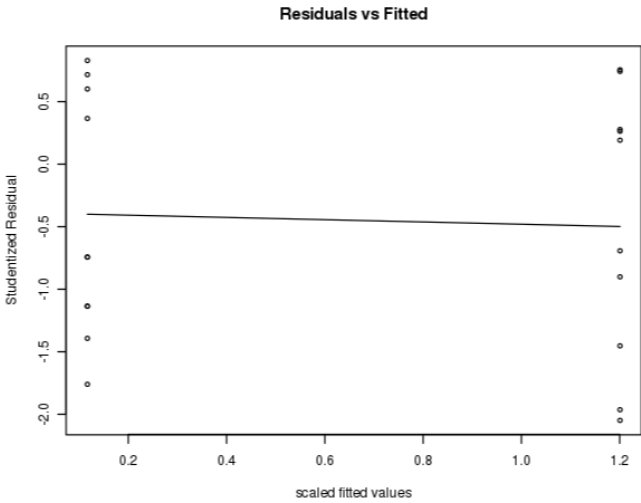
Model checking plots for mean

residuals



Model checking plots for phi

residuals



5. Liner mixed model

- 반복 측정되어 서로 상관된 자료
- 고정효과 + 변량효과
- y : 정규분포 \rightarrow 선형혼합모형

6명의 환자의 치료 효과

Patient	Treatment		Differences A-B	Patient Mean
	A	B		
1	20	12	8	16.0
2	26	24	2	25.0
3	16	17	-1	16.5
4	29	21	8	25.0
5	22	21	1	21.5
6	24	17	7	20.5
Mean	22.83	18.67	4.17	20.75

treatmenteffect.csv

	y	x	patient
1	20	1	1
2	26	1	2
3	16	1	3
4	29	1	4
5	22	1	5
6	24	1	6
7	12	2	1
8	24	2	2
9	17	2	3
10	21	2	4

- 선형혼합모형

반응변수 = 전체평균 + 치료효과

+ 개인별 변량효과 + error

$$y_{ij} = \mu + t_j + p_i + e_{ij} \quad (i^{th} \text{ subject } j^{th} \text{ treatment})$$

- 모델 가정

- Errors $e_{ij} \sim N(0, \sigma^2)$

- 개인별 변량효과 $p_i \sim N(0, \sigma_p^2)$

DHGLM
Run
Model
y ~ x +(1| patient)
Response Variable
y
Variable
y
patient
Selected
x
Interaction
Random Effects
patient

Coefficients of the model: $y \sim x + (1 \mid \text{patient})$

link: identity

Dist: gaussian

	Estimate	Std. Error	t-value	p_val	LL	UL
(Intercept)	27.0000	3.1572	8.5519	0.0000	20.8119	33.1881
x	-4.1667	1.8631	-2.2364	0.0253	-7.8184	-0.5149

Coefficients of the model: phi ~ 1

link: log

Dist: gaussian

	Estimate	Std. Error	t-value
(Intercept)	2.2185	0.5341	4.1537

Coefficients of the model lambda

	Estimate	Std. Error	t-value
patient	2.2246	0.8180	2.7195

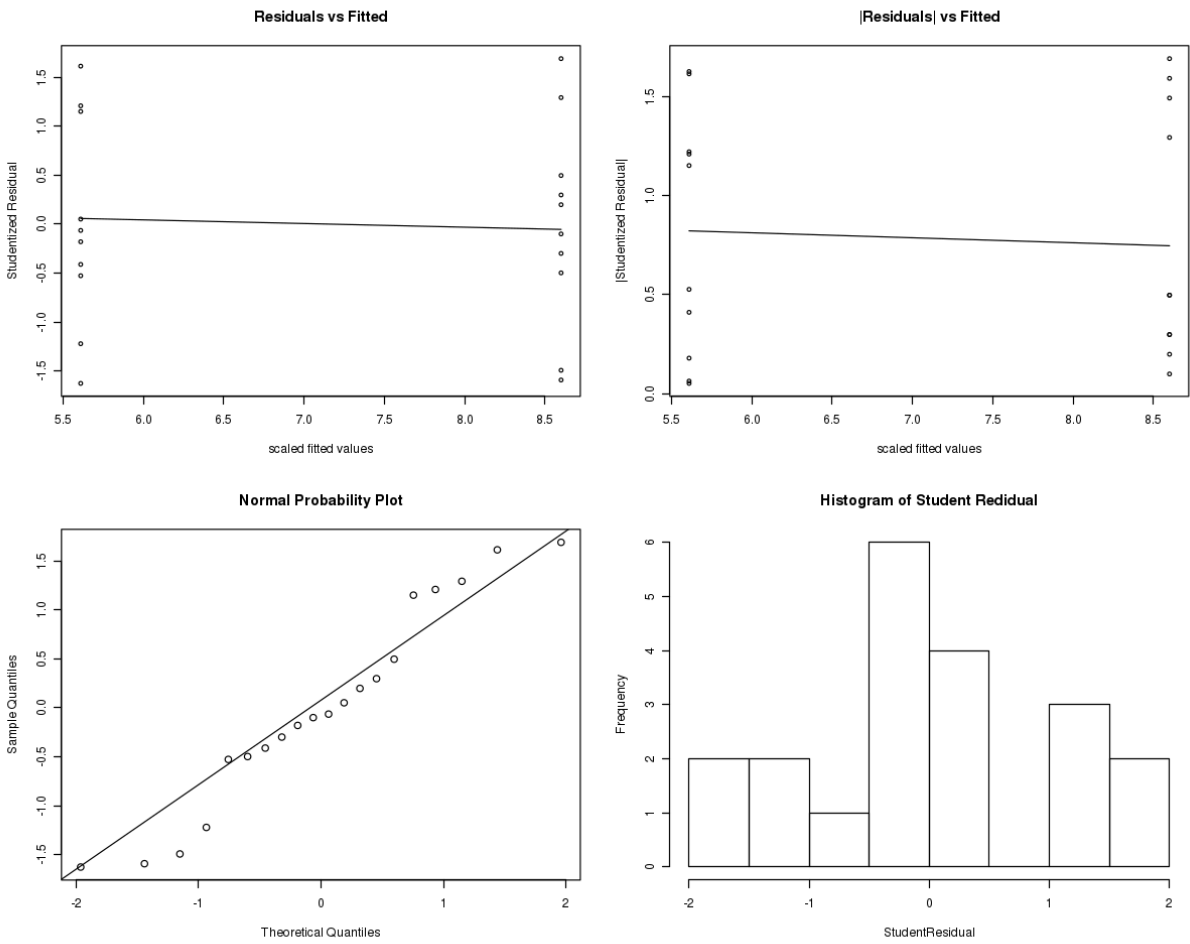
Likelihood Function Values and Conditional AIC

-2ML (-2 p_v(mu) (h)) :	64.7245
-2RL (-2 p_beta(mu),v(mu) (h)) :	59.0353
cAIC :	65.6649
Scaled Deviance :	7.0113
df :	7.0113

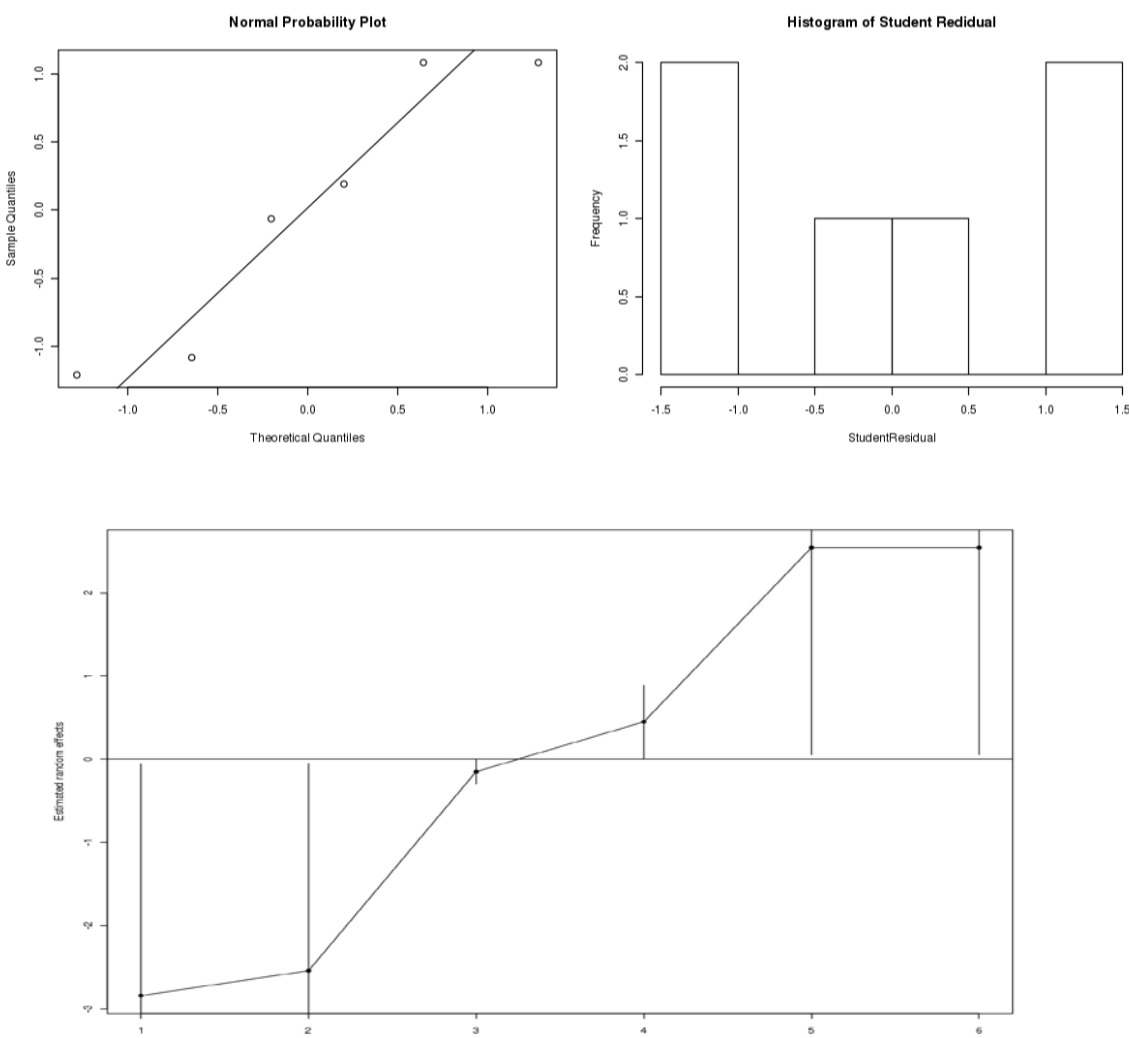
- patient effect

Model checking plots for mean

residuals



random effect




6. Hierarchical generalized liner model

- 반복 측정되어 서로 상관된 자료
- 고정효과+변량효과
- y : 정규분포 \rightarrow 선형혼합모형
- y : 비정규 (개수 or 비율) \rightarrow HGLM


Epilepsy data(epilepsy(page66).csv)

- Thall and Vail (1990)
- longitudinal data from a clinical trial of 59 epileptics
- who were randomized to a new drug or a placebo ($T=1$ or $T=0$)
- The trial included the logarithm of the average number of epileptic seizures
recorded in the 8-week period preceding the trial (B)
- the logarithm of age (A)
- number of clinic visit (V : a linear trend, coded(-3,-1,1,3))
- A multivariate response variable (y) consists of the seizure counts
2-week
periods before each of four visits to the clinic



59명
환자

4회 반복



ID	Trt	Sev	Age	W2	W4	W6	W8
1	0	11	31	5	3	3	3
2	0	11	30	3	5	3	3
3	0	6	25	2	4	0	5
4	0	8	36	4	4	1	4
5	0	66	22	7	18	9	21
6	0	27	29	5	2	8	7
.
.
.
55	1	16	32	3	5	4	3
56	1	22	26	1	23	19	8
57	1	25	21	2	3	0	1
58	1	13	36	0	0	0	0
59	1	12	37	1	4	3	2

Trt : 0 , 새로운 약을 복용

1, 가짜약

Baseline : 약을 복용하기 8주
전부터 약을 복용하기까지의
발작회수.

반응변수 : 약 복용을 시작한 후
2주간격으로 관측한 간질병 환
자의 발작 회수.

W2 : 처음 2주의 발작회수

W4 : 2-4 주 사이의 발작회수

W6 : 4-6 주 사이의 발작회수

W8 : 6-8 주 사이의 발작회수

epilepsy(page66).csv

	y ↕	T ↕	B ↕	A ↕	V ↕	patient ↕	id ↕
1	5	0	1.011600912	3.43399	-3	1	1
2	3	0	1.011600912	3.43399	-1	1	2
3	3	0	1.011600912	3.43399	1	1	3
4	3	0	1.011600912	3.43399	3	1	4
5	3	0	1.011600912	3.4012	-3	2	5
6	5	0	1.011600912	3.4012	-1	2	6
7	3	0	1.011600912	3.4012	1	2	7
8	3	0	1.011600912	3.4012	3	2	8
9	2	0	0.405465108	3.21888	-3	3	9
10	4	0	0.405465108	3.21888	-1	3	10

Showing 1 to 10 of 236 entries

Previous 1 2 3 4 5 ... 24 Next

y : 약 복용 후 관측한 간질병 환자의 발작 회수

T : 0 (새로운 약), 1(가짜약)

B : log(Baseline)

A : log(Age)

V : linear trend for visit (-3, -1, 1, 3)

patient : 59명 환자 번호

id : 236개 (=59*4) 관측치 번호

y_{ij} : response variable for the i th patient and j th visit

$$E(y_{ij}|v_i, v_{ij}) = \mu_{ij} \quad \text{and} \quad \text{var}(y_{ij}|v_i, v_{ij}) = \mu_{ij}$$

$$\log(\mu_{ij}) = \beta_0 + \beta_B B_i + \beta_T T_i + \beta_{B*T} B_i * T_i + \beta_V V_j + v_i + v_{ij}$$

v_i : patient effect, v_{ij} : patient and visit effect

Poisson - normal HGLM

$$v_i \sim N(0, \lambda_1) \quad v_{ij} \sim N(0, \lambda_2) \quad \text{변량효과 (random effect)}$$

NB (Negative Binomial) - gamma HGLM : allowing different distributions for random effects

$$\exp(v_i) \sim G(1, \lambda_1) \quad \exp(v_{ij}) \sim G(1, \lambda_2) : \text{gamma distribution with mean 1} \\ \text{and variance } \lambda_1, \lambda_2$$

GLM

DHGLM

4

Run

Model

y ~ B+T+A+B:T+V

Response Variable

2

y

Variable

y patient id

Selected

B T A B:T V

Interaction

1

B T

Random Effects

Cubic Spline

Covariance Kernel

Precision Kernel

Random slope model

Interaction in the Random effect

Distribution for Random effects 1

gaussian

Upload neighborhood file

3

Distribution

poisson

Link Function

log

Binomial denominator

No intercept model

Offset Variable

Comparison with other model

VIF

Robust standard errors

Confidence intervals for coefficients

Exponential scale

Model Summary

Model Checking Plot

Prediction

Coefficients of the model: $y \sim B + T + A + B:T + V$

link: log

Dist: gaussian

	Estimate	Std. Error	t-value	p_val	exp(LL)	exp(UL)
(Intercept)	-2.80	0.41	-6.87	0.00	0.03	0.14
B	0.95	0.04	21.80	0.00	2.37	2.81
T	-1.34	0.16	-8.56	0.00	0.19	0.36
A	0.90	0.12	7.70	0.00	1.95	3.08
V	-0.03	0.01	-2.90	0.00	0.95	0.99
B:T	0.56	0.06	8.85	0.00	1.55	1.99

Coefficients of the model: $\phi \sim 1$

link: log

Dist: gaussian

Likelihood Function Values and Conditional AIC

-2ML (-2 h) :	1635.90
-2RL (-2 p_beta (h)) :	1664.75
cAIC :	1647.90
Scaled Deviance :	869.91
df :	230.00

Poisson-Normal HGLM

DHGLM

4

Run

Model

y ~ B+T+A+B:T+V +(1| patient)

Response Variable

2

y

Variable

y
patient
id

Selected

B
T
A
B:T
V

Interaction

1

B T

Random Effects

patient

Cubic Spline

Covariance Kernel

Precision Kernel

☐ Random slope model

☐ Interaction in the Random effect

Distribution for Random effects 1

gaussian

Upload neighborhood file

Distribution

3

poisson

☐ Binomial denominator

Link Function

log

☐ No intercept model

☐ Offset Variable

☐ Comparison with other model

☐ VIF

☐ Robust standard errors

Model Summary

Model Checking Plot

Prediction

Coefficients of the model: $y \sim B + T + A + B:T + V + (1 | \text{patient})$

link: log

Dist: gaussian

	Estimate	Std. Error	t-value	p_val	exp(LL)	exp(UL)
(Intercept)	-1.20	0.02	-0.76	0.45	0.01	6.70
B	0.87	0.00	4.90	0.00	1.69	3.38
T	-0.94	0.01	-1.77	0.08	0.14	1.10
A	0.44	0.00	0.95	0.34	0.63	3.86
V	-0.03	0.00	-2.89	0.00	0.95	0.99
B:T	0.33	0.00	1.23	0.22	0.82	2.38

Coefficients of the model: $\phi_i \sim 1$

link: log

Dist: gaussian

	Estimate	Std. Error	t-value
patient	-1.16	0.01	-5.61

Coefficients of the model lambda

	Estimate	Std. Error	t-value
patient	-1.16	0.01	-5.61

Likelihood Function Values and Conditional AIC

-2ML (-2 p_v(mu) (h)) :	1339.26
-2RL (-2 p_beta(mu),v(mu) (h)) :	1348.58
cAIC :	1272.67
Scaled Deviance :	399.17
df :	182.25

- 73 -

NB-Gamma HGLM

DHGLM

Model

$y \sim B+T+A+B:T+V+(1|patient)+(1|id)$

Response Variable

y

Variable

y
patient
id

Selected

B
T
A
B:T
V

Interaction

B T

Random Effects

patient id

Cubic Spline

Covariance Kernel

Precision Kernel

Run

Cubic Spline

Covariance Kernel

Precision Kernel

Random slope model

Interaction in the Random effect

Distribution for Random effects 1

gamma

Distribution for Random effects 2

gamma

Upload neighborhood file

Distribution

poisson

Binomial denominator

Link Function

log

No intercept model

Offset Variable

Comparison with other model

VIF

Robust standard errors

Confidence intervals for coefficients

Exponential scale

Model Summary

Model Checking Plot

Prediction

Coefficients of the model: $y \sim B + T + A + B:T + V + (1 | patient) + (1 | id)$

link: log

Dist: gamma Dist: gamma

	Estimate	Std. Error	t-value	p_val	exp(LL)	exp(UL)
(Intercept)	-1.30	0.01	-1.04	0.30	0.02	3.20
B	0.89	0.00	6.21	0.00	1.84	3.24
T	-0.84	0.00	-2.08	0.04	0.19	0.95
A	0.49	0.00	1.33	0.18	0.79	3.38
V	-0.03	0.00	-1.48	0.14	0.94	1.01
B:T	0.33	0.00	1.58	0.11	0.92	2.08

Coefficients of the model: $\phi \sim 1$

link: log

Dist: gaussian

Coefficients of the model lambda

	Estimate	Std. Error	t-value
patient	-1.29	0.01	-5.79
id	-1.98	0.01	-12.29

Likelihood Function Values and Conditional AIC

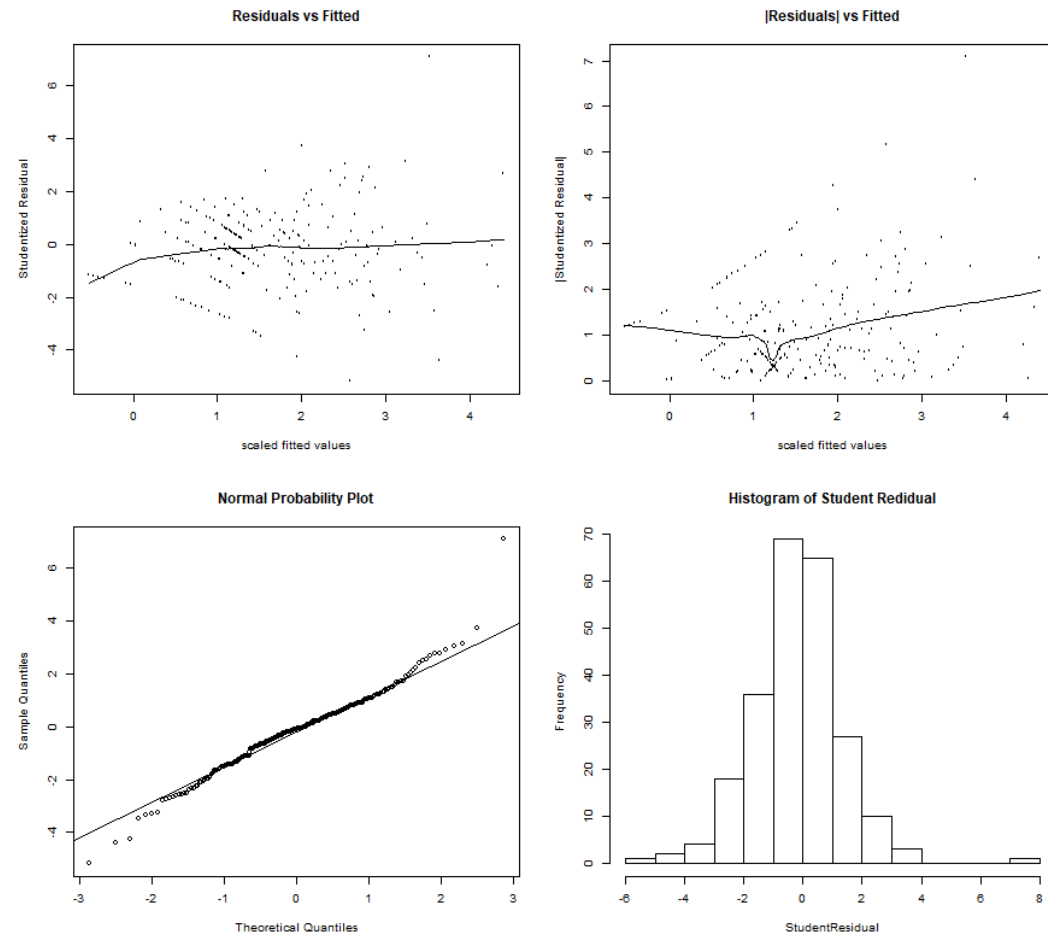
-2ML (-2 p_v(mu) (h)) :	1255.58
-2RL (-2 p_beta(mu),v(mu) (h)) :	1270.83
cAIC :	1163.92
Scaled Deviance :	142.57
df :	108.34

Poisson-Normal HGLM

Model Summary Model Checking Plot Prediction

Model checking plots for mean

residuals

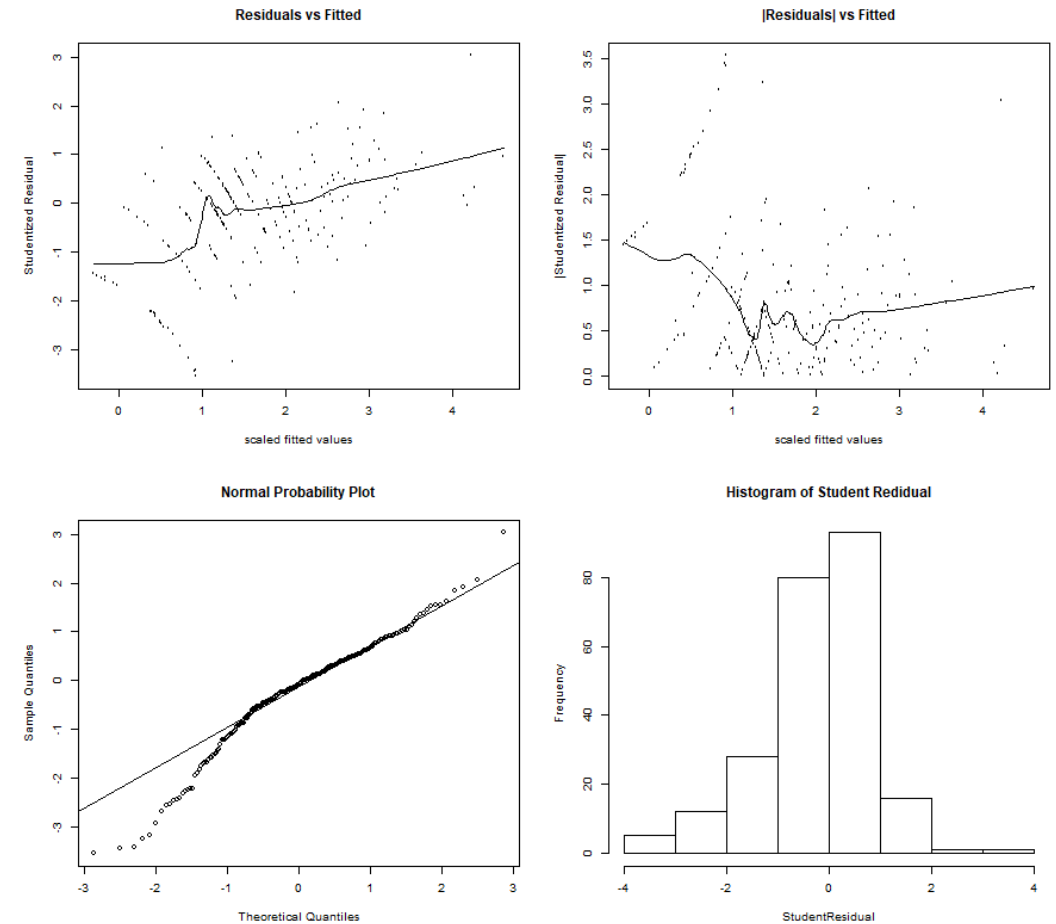


NB-Gamma HGLM

Model Summary Model Checking Plot Prediction

Model checking plots for mean

residuals



respiratory(page111,159).csv

	patient ↕	treatment ↕	sex ↕	age ↕	center ↕	baseline ↕	past ↕	y ↕	trt ↕	msex ↕	base ↕
1	1	placebo	male	46	1	0	0	0	0	1	0
2	1	placebo	male	46	1	0	0	0	0	1	0
3	1	placebo	male	46	1	0	0	0	0	1	0
4	1	placebo	male	46	1	0	0	0	0	1	0
5	2	placebo	male	28	1	0	0	0	0	1	0
6	2	placebo	male	28	1	0	0	0	0	1	0
7	2	placebo	male	28	1	0	0	0	0	1	0
8	2	placebo	male	28	1	0	0	0	0	1	0
9	3	active	male	23	1	1	1	1	1	1	1
10	3	active	male	23	1	1	1	1	1	1	1

Showing 1 to 10 of 444 entries

Previous

1

2

3

4

5

...

45

Next

patient : 111 patients

treatment : 54 active, 57 placebo

center : center 1 (56 patients), 2 (55 patients)

baseline, base : previous y

y : good=1, poor=0

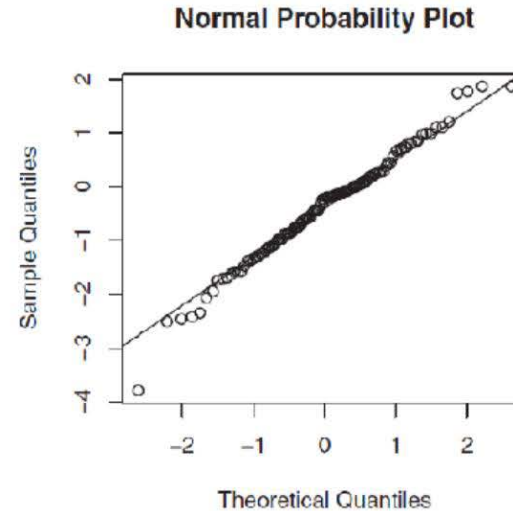
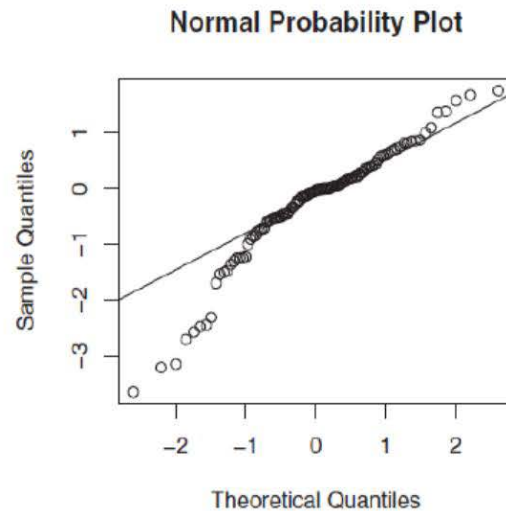
msex : male=1, female=0

- With binary data it is difficult to identify the distribution of random effects.
- The use of a heavy-tailed distribution for random effects, by allowing random effects for λ , removes sensitivity of the parameter estimation to the choice of random-effect distribution.
- For binary data, GLMM estimators can give serious biases if the true distribution is not normal.
- Taking advantage of DHGLM is recommended.

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_0^{(\mu)} + \beta_1^{(\mu)} trt_i + \beta_2^{(\mu)} msex_i + \beta_3^{(\mu)} age_i + \beta_4^{(\mu)} center_i + \beta_5^{(\mu)} base_i + \beta_6^{(\mu)} y_{i(j-1)} + v_i^{(\mu)}$$

$$\log(\lambda_i) = \beta_0^{(\lambda)} + age_i \beta_1^{(\lambda)} + v_i^{(\lambda)}$$

where $v_i^{(\mu)} \sim N(0, \lambda_i)$ and $v_i^{(\lambda)} \sim N(0, \tau)$



Binomial-Normal HGLM

Model for phi ☐ Use

Model
 $y \sim \text{trt} + \text{msex} + \text{age} + \text{center} + \text{base} + \text{past} + (1 | \text{patient})$

Response Variable
y

Variable
patient
treatment
sex
baseline
y

Selected
trt
msex
age
center
base
past

Interaction

Random Effects
patient

Cubic Spline ☐
Covariance Kernel ☐
Precision Kernel ☐
Distribution of Random Effects
gaussian

Model for lambda ☐ Use

Cubic Spline ☐
Covariance Kernel ☐
Precision Kernel ☐
☒ Random slope model
Random slope

☐ Interaction in the Random effect
Distribution for Random effects 1
gaussian
Upload neighborhood file ☐

Distribution
binomial
☐ Binomial denominator
Link Function
logit

Model Summary Prediction Model Checking Plot

Binomial-Normal DHGLM

DHGLM

Run

Model for phi

☐ Use

Model

y ~ trt+msex+age+center+base+past +(1| patient)

Response Variable

y

Variable

patient
treatment
sex
baseline
y

Selected

trt
msex
age
center
base
past

Interaction

Random Effects

patient

Cubic Spline

☐

Covariance Kernel

☐

Precision Kernel

☐

☒ Random slope model

Random slope

☐ Interaction in the Random effect

Distribution for Random effects 1

gaussian

Upload neighborhood file

☐

Distribution

binomial

☐ Binomial denominator

Link Function

logit

☐ No intercept model

Cubic Spline

☐

Covariance Kernel

☐

Precision Kernel

☐

Distribution of Random Effects

gaussian

Model for lambda

☒ Use

Model

lambda ~ 1 + age +(1| patient)

Model for variance of random effect in the mean model

lambda

Variable

patient
treatment
sex
center
baseline
past
y
trt
msex
base

Selected

age

Interaction

Random Effects

patient

Link Function

log

Distribution for Random effects 1

gaussian

☐ Offset Variable

Model Summary

Prediction

Model Checking Plot

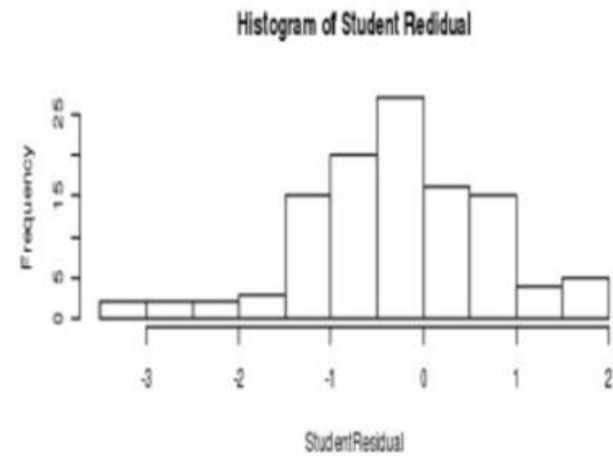
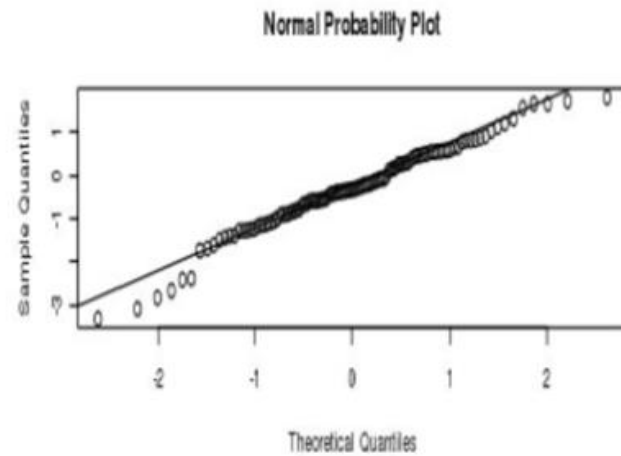
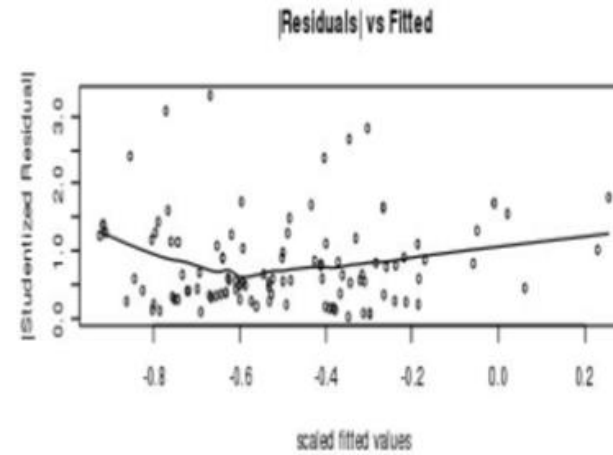
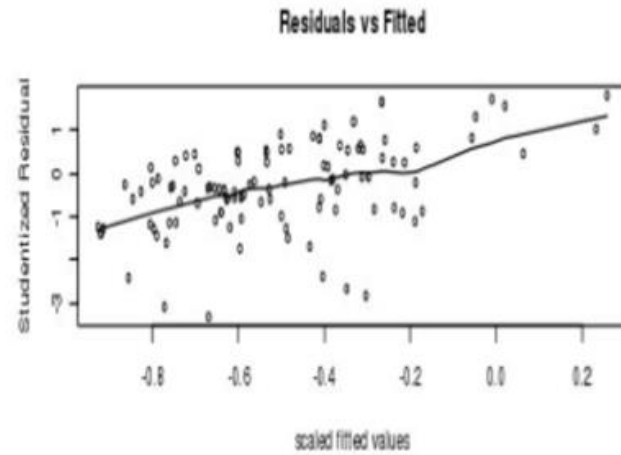
- The likelihood ratio test for $H_0 : \tau = 0$, based on the restricted likelihood, rejects the null hypothesis (the deviance difference for HGLM and DHGLM is $3.3 > \chi^2_{2\delta}(1)=2.71$)
- Furthermore, in this example there are apparent differences between parameter estimates.

Table 6.7 HGLM and DHGLM results for the respiratory data

	HGLM			DHGLM		
	Estimate	SE	t-value	Estimate	SE	t-value
model for the mean						
$\beta_0^{(\mu)}$	-1.111	1.033	-1.075	-0.290	1.683	-0.172
$\beta_1^{(\mu)}$	1.256	0.415	3.028	1.601	0.640	2.503
$\beta_2^{(\mu)}$	-0.261	0.597	0.437	-0.541	1.030	0.525
$\beta_3^{(\mu)}$	-0.035	0.019	-1.889	-0.060	0.032	-1.873
$\beta_4^{(\mu)}$	0.682	0.419	1.626	0.672	0.654	1.027
$\beta_5^{(\mu)}$	1.821	0.446	4.079	2.411	0.672	3.586
$\beta_6^{(\mu)}$	0.575	0.304	1.891	-0.051	0.338	-0.152
model for the random effect variance						
$\beta_0^{(\lambda)}$	-0.683	0.737	-0.927	0.015	0.345	0.042
$\beta_1^{(\lambda)}$	0.047	0.020	2.339	0.067	0.010	6.976
$\log(\tau)$				-1.246	3.020	-0.413
likelihood values and cAIC						
$-2 \log(\text{likelihood})$	431.0			428.2		
$-2 \log(\text{restricted likelihood})$	438.4			435.1		
cAIC	422.5			413.6		

- In this case, we should report the results from the DHGLM because a distributional assumption of random effects is hard to identify with the binary data.

Model checking plot for lambda



7. Survival data analysis

- T: 생존시간 (음이 아닌 실수값)
- $f(t)$: T의 한 분포인 확률밀도함수

⇒ **생존율(survival rate, survival function):**

- $S(t) = \Pr(T > t)$
- 환자가 t시점 이상 생존할 확률
- 환자가 t시점까지 사망하지 않고 생존할 확률
- No censoring: $S(t)$ 의 추정=t시점 이상 생존한 대상수/총 연구대상수

⇒ **위험함수(hazard rate, hazard function):**

- $h(t) = f(t)/S(t)$
- 어떤 환자가 t시점까지는 생존했다가 t시점 바로 직후에 사망하게 되는 순간위험률

(Note) 생존율이 높을수록 위험률이 대체로 낮아지는 경향

- T: 생존시간 (음이 아닌 실수값)
- $f(t)$: T의 한 분포인 확률밀도함수

⇒ **생존율(survival rate, survival function):**

- $S(t) = \Pr(T > t)$
- 환자가 t시점 이상 생존할 확률
- 환자가 t시점까지 사망하지 않고 생존할 확률
- No censoring: $S(t)$ 의 추정=t시점 이상 생존한 대상수/총 연구대상수

⇒ **위험함수(hazard rate, hazard function):**

- $h(t) = f(t)/S(t)$
- 어떤 환자가 t시점까지는 생존했다가 t시점 바로 직후에 사망하게 되는 순간위험률

(Note) 생존율이 높을수록 위험률이 대체로 낮아지는 경향

[kidney\(page224\).csv](#)

- id : patients identifier for 38 patients
- time : time until infection since the insertion of the catheter
- Status : event status (=1; infection, =0; censoring)
- Age : age of patients in years
- Sex : 1=male, 2=female
- Disease : a factor for disease type with levels Other, GN, AN, and PKD

	id	time	status	age	sex	disease
1	1	8	1	28	1	Other
2	1	16	1	28	1	Other
3	2	23	1	48	2	GN
4	2	13	0	48	2	GN
5	3	22	1	32	1	Other

- Survival Time (time): time until the first and second recurrences of kidney infection ($n_i = 2$) since the insertion of the catheter.
- $q = 38$ patients.
- The survival times for the same patient are likely to be correlated because of a shared frailty describing the common patient's effect.
- 24% of the data was censored.

Kaplan-Meier Estimator

Run

Survival time (Numeric Only)

time

☐ Initial Time

Indicator (Numeric Only)

status

☐ Use Group

Result

Data Summary

Kaplan-Meier Estimate

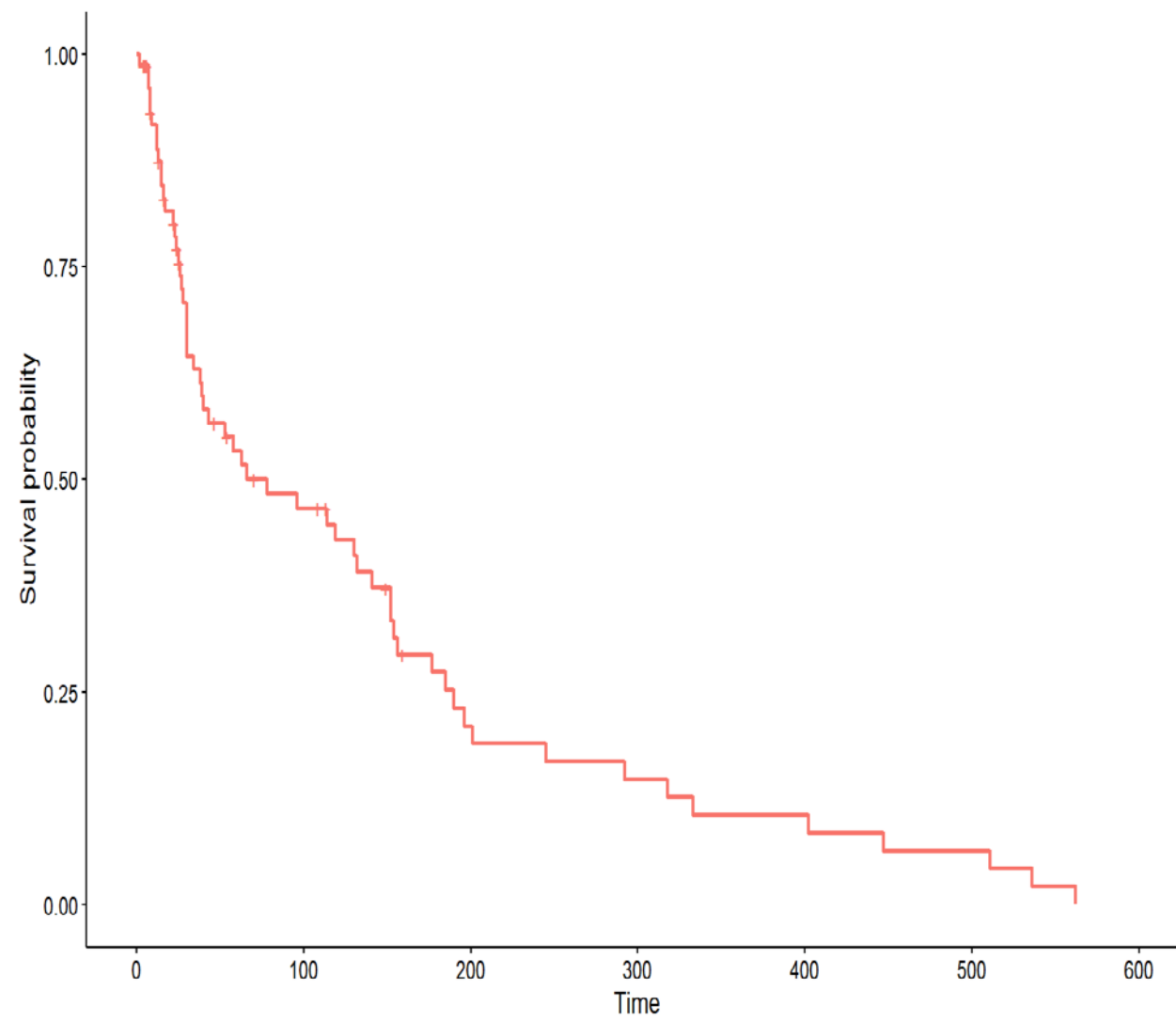
Kaplan-Meier Curve

Data Summary

records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
76.00	76.00	76.00	58.00	137.02	19.77	78.00	38.00	141.00

Kaplan-Meier Estimates

time	n.risk	n.event	censored	survival	std.err	lower 95% CI	upper 95% CI
2.000	76.000	1.000	0.000	0.987	0.013	0.961	1.000
4.000	75.000	0.000	1.000	0.987	0.013	0.961	1.000
5.000	74.000	0.000	2.000	0.987	0.013	0.961	1.000
6.000	72.000	0.000	1.000	0.987	0.013	0.961	1.000
7.000	71.000	2.000	0.000	0.959	0.024	0.914	1.000
8.000	69.000	2.000	2.000	0.931	0.032	0.873	0.989
9.000	65.000	1.000	0.000	0.917	0.035	0.853	0.981
12.000	64.000	2.000	0.000	0.888	0.042	0.815	0.961
13.000	62.000	1.000	1.000	0.874	0.045	0.797	0.951
15.000	60.000	2.000	0.000	0.845	0.051	0.760	0.929
16.000	58.000	1.000	1.000	0.830	0.054	0.743	0.918



Kaplan-Meier Estimator

Run

Survival time (Numeric Only)

time

☐ Initial Time

Indicator (Numeric Only)

status

☒ Use Group

Groups

sex

☐ Log-Rank Test

☒ Log-Rank Test

Result

Data Summary

Kaplan-Meier Estimate

Kaplan-Meier Curve

Group K-M Estimates

Group K-M Curve

Data Summary

records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
76.00	76.00	76.00	58.00	137.02	19.77	78.00	38.00	141.00

Group Data Summary

	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
sex=1	20.00	20.00	20.00	18.00	65.29	29.85	22.00	12.00	30.00
sex=2	56.00	56.00	56.00	40.00	161.58	22.94	130.00	66.00	185.00

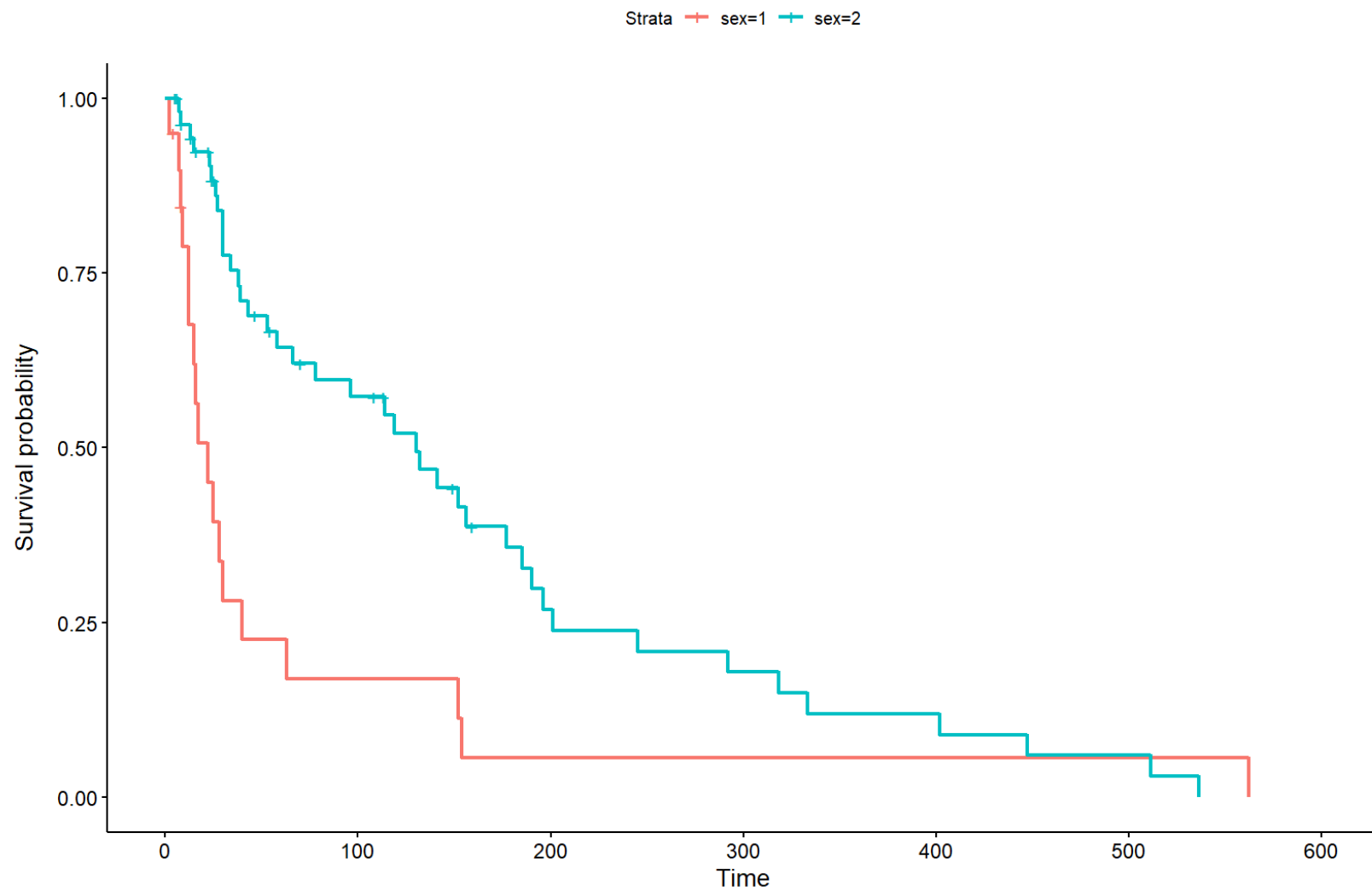
Log-Rank Test Table

N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
20.000	18.000	10.187	5.993	8.308
56.000	40.000	47.813	1.277	8.308

Log-Rank Test Results

Chisq	Degrees of freedom	p-value
8.308	1.000	0.004

Group K-M Curve



Model

```
Surv( time , status ==1)~  
age+sex
```

Survival Time

time

☐ Initial Time

Variable

id
time
status
disease
frail



Selected

age
sex

Censoring Indicator

status

Surv(time , status ==1)~ age+sex

Coefficients

	coef	exp(coef)	se(coef)	z	Pr(> z)
age	0.002032	1.002034	0.009246	0.219749	0.826067
sex	-0.829314	0.436349	0.298955	-2.774043	0.005536

Confidence Interval

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.0020	0.9980	0.9840	1.0204
sex	0.4363	2.2917	0.2429	0.7840

Proportional Hazard Assumption Check

	chisq	df	p
age	0.106	1.000	0.744
sex	11.100	1.000	0.001
GLOBAL	12.218	2.000	0.002

n	number of events
76.00	58.00

Test Results

	Statistic	df	p-value
Likelihood ratio test	7.116	2.000	0.028
Wald test	8.020	2.000	0.018
Score(logrank) test	8.445	2.000	0.015

Concordance	se
0.662	0.045

분석 모형: 프레일티 모형(frailty models)

- ▶ 고려사항: 상관성, 이질성, 중도절단성

Vaupel et al.(1979), Clayton(1978, 1985)

- ▶ 정의
각 개인의 공통적인 프레일티 U 가 주어질 때 생존시간의 위험률:

$$\lambda(t) = \lambda_0(t) \exp(x_1\beta_1 + \cdots + x_p\beta_p)U,$$

$\lambda_0(\cdot)$ 는 임의의 기저(baseline)위험률,

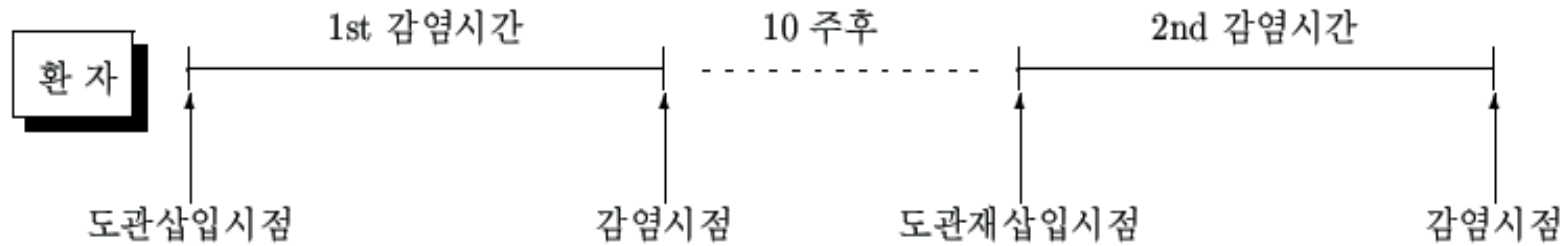
U 는 평균이 1인 적절한 분포(예: LN or gamma)를 가진다.

▶ 의미

- (i) $U = 1$ 인 사람은 위험률이 표준적인 경향
(모든 $U = 1$ 이면 FM은 Cox모형)
- (ii) $U < 1$ 인 사람은 표준적인 사람보다 위험률이 낮은 경향
- (iii) $U > 1$ 인 사람은 표준적인 사람보다 위험률이 높은 경향

<신장환자의 재발 감염시간에 관한 임상 연구설계>

(McGilchrist & Aisbett, 1991)



log-normal model (HL(0,1))

Frailty Model Run

Model

Surv(time ,status ==1)~ sex+age +(1|id)

Survival Time

time

☐ Initial Time

Variable

id
time
status
disease
frail

Selected

sex
age

Interaction

Random Effects

id

Distribution for Random Effect

gaussian

Censoring Indicator

status

☐ grouped duration

☐ Comparison with other model

☐ Robust standard errors

☐ Confidence intervals for coefficients

☐ Exponential scale

Additional Settings

Order of Laplace Approximation for Likelihood(mean) and Restricted Likelihood(Dispersion)

Order for Mean

0

Order for Dispersion

1

Model Summary

Random Effect Inferences

Models for conditional hazard: $\text{Surv}(\text{time}, \text{status} == 1) \sim \text{sex} + \text{age} + (1$
Coefficients of the mean model

	Estimate	Std. Error	t-value	p-value
sex	-1.38	0.43	-3.20	0.00
age	0.00	0.01	0.40	0.69

Coefficients of the dispersion model

	Estimate	Std. Error
id	0.53	0.34

Likelihood Function Values

-2*lo	330.40
-2*hp	390.77
-2*p_b.v(hp)	371.54

log-normal model (HL(0,1))

Model Summary

Random Effect Inferences

Models for conditional hazard: $\text{Surv}(\text{time}, \text{status} == 1) \sim \text{sex} + \text{age} + (1 | \text{id})$

Coefficients of the mean model

	Estimate	Std. Error	t-value	p-value
sex	-1.38	0.43	-3.20	0.00
age	0.00	0.01	0.40	0.69

Coefficients of the dispersion model

	Estimate	Std. Error
id	0.53	0.34

Likelihood Function Values

-2h0	330.40
-2*hp	390.77
-2*p_b,v(hp)	371.54

AIC

cAIC	362.46
mAIC	370.70
rAIC	373.54

gamma model (HL(0,2))

Frailty Model

Run

Model

Surv(time,status==1)~ sex+age +(1 | id)

Survival Time

time

☐ Initial Time

Variable

id

time

status

disease

frail

Selected

sex

age

→

←

Interaction

Random Effects

id

Distribution for Random Effect

gamma

Censoring Indicator

status

☐ grouped duration

☐ Comparison with other model

☐ Robust standard errors

☐ Confidence intervals for coefficients

☐ Exponential scale

Additional Settings

Order of Likelihood Approximation for Likelihood(mean) and Restricted Likelihood(Dispersion)

Order for Mean

0

Order for Dispersion

2

Model Summary

Random Effect Inferences

Models for conditional hazard: Surv(time, status == 1) ~ sex + age + (1 | id)

Coefficients of the mean model

	Estimate	Std. Error	t-value	p-value
sex	-1.69	0.48	-3.50	0.00
age	0.01	0.01	0.52	0.60

Coefficients of the dispersion model

	Estimate	Std. Error
id	0.56	0.28

Likelihood Function Values

-2*0	324.08
-2*hp	391.74
-2*p_b,v(hp)	370.89
-2*s_b,v(hp)	368.88

gamma model (HL(0,2))

Model Summary

Random Effect Inferences

Models for conditional hazard: $\text{Surv}(\text{time}, \text{status} == 1) \sim \text{sex} + \text{age} + (1 | \text{id})$

Coefficients of the mean model

	Estimate	Std. Error	t-value	p-value
sex	-1.69	0.48	-3.50	0.00
age	0.01	0.01	0.52	0.60

Coefficients of the dispersion model

	Estimate	Std. Error
id	0.56	0.28

Likelihood Function Values

-2h0	324.08
-2*hp	391.74
-2*p_b,v(hp)	370.89
-2*s_b,v(hp)	368.88

AIC

cAIC	358.93
mAIC	370.34
rAIC	372.89