



서강대학교 정보통신 대학원 정 화민 교수 (MIS Ph.D.)

Linear Regression Analysis

선형회귀분석이란? 회귀식을 통하여 특정변수(독립변수 또는 설명변수)의 변화가 다른 변수(종속변수)의 변화와 어떤 관련성이 있는지 변수의 원인 및 결과 등에 관한 사항 선형적으로 분석하는 대표적인 방법.

단순회귀분석: 독립변수가 **1**개, 종속변수 **1**개

다중회귀분석: 독립변수가 **2**개 이상, 종속변수가 **1**개

회귀직선 공식: $Y = a + bX$ **Y**는 결과, **x**는 원인

a 는 Y의 절편, **b**는 기울기

a 값과 **b** 값을 구한다음 원인 **x**에 데이터를 넣어 **Y**의 결과값을 예측할 수 있다.

다음페이지의 데이터를 입력하고 각각의 공식에 맞게 답을 구해보자.

Linear Regression Analysis

	A	B	C	D	E	F	G	H	I
1	번호	노동력(X)	생산량(Y)	X ²	XY				
2	1	267	428	71,289	114,276				
3	2	263	430	69,169	113,090				
4	3	238	417	56,644	99,246				
5	4	219	384	47,961	84,096				
6	5	274	432	75,076	118,368				
7	6	257	425	66,049	109,225				
8	7	321	474	103,041	152,154				
9	8	305	462	93,025	140,910				
10	9	285	449	81,225	127,965				
11	10	247	405	61,009	100,035				
12	합계	2,676	4,306	724,488	1,159,365				
13		A	B	C	D				
14									
15									
16	N =	10							
17									
18	a =	204.8125							
19									
20	b =	0.8438							
21									

$$Y = a + bX$$

$$a = \frac{BC - AD}{nC - A^2}$$

$$b = \frac{nD - AB}{nC - A^2}$$

$$=(B*_C - A*D) / (N*_C - A^2)$$

$$=(N*D - A*B) / (N*_C - A^2)$$

Linear Regression Analysis

1. 번호, 노동력, 생산량 데이터를 엑셀에 입력하고 X의 제곱, XY를 구해보자.

(여기서 X축은 원인이므로 노동력이 되고, Y 축은 결과가 됨으로 X,Y의 축이 변하면 안된다.)

Y의 절편 a, 기울기 b를 구하는 공식을 보고 값을 구해보자.

그러면 : 최종적으로 회귀직선공식은 다음과 같이 된다.

$Y = 204.81 + 0.84X$ 가 된다. (소수점 2자리 버림으로 계산)

그럼 노동력이 **350**명일 때 생산량은 몇 개가 되는 것일까? **498**개

여기 까지는 우리가 예측함수 **FORECAST**를 써도 가능하다.

노동력이 **350**일 때 **498**개가 나올 확률은 ? 추정치와 잔차를 구한다음
R제곱 값을 구하자 .

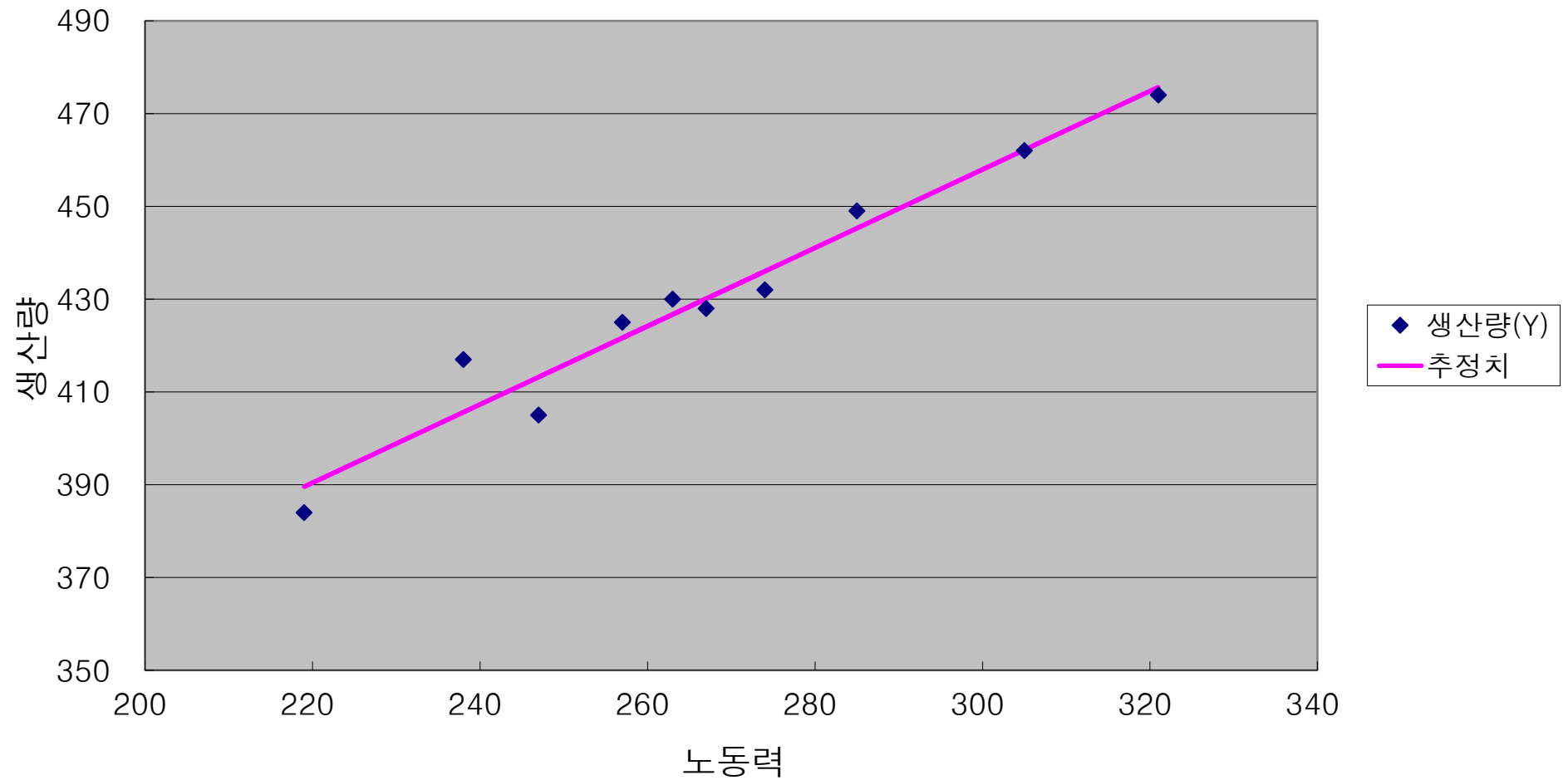
Linear Regression Analysis

추정치(도출된 회귀식에 의한 예측치)와 잔차(평균이 아닌 추정된 회귀식과의 차이)를 구하고 추정치 분산, 잔차분산을 구한다음 R제곱을 구하자.

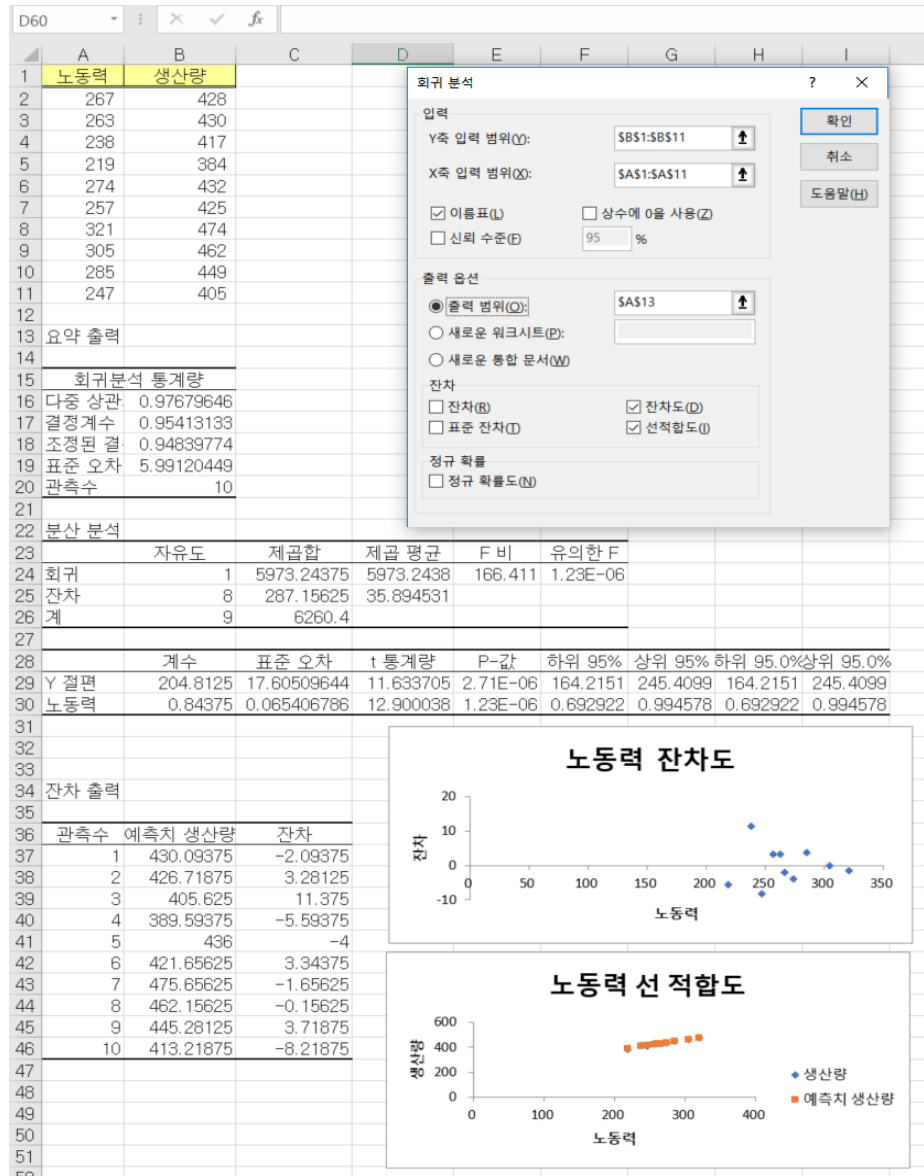
F2 \times \checkmark f_x $=\$B\$18+\$B\$20*B2$								
	A	B	C	D	E	F	G	H
1	번호	노동력(X)	생산량(Y)	X^2	XY	추정치	잔차	
2	1	267	428	71,289	114,276	430	-2	
3	2	263	430	69,169	113,090	427	3	
4	3	238	417	56,644	99,246	406	11	
5	4	219	384	47,961	84,096	390	-6	
6	5	274	432	75,076	118,368	436	-4	
7	6	257	425	66,049	109,225	422	3	
8	7	321	474	103,041	152,154	476	-2	
9	8	305	462	93,025	140,910	462	0	
10	9	285	449	81,225	127,965	445	4	
11	10	247	405	61,009	100,035	413	-8	
12	합계	2,676	4,306	724,488	1,159,365			
13								
14								
15				Y 분산	626.0400	←	=VARP(C2:C11)	
16	N =	10						
17				추정치분산	597.3244	←	=VARP(F2:F11)	
18	a =	204.8125						
19				잔차분산	28.7156	←	=VARP(G2:G11)	
20	b =	0.8438						
21				R^2	0.9519	←	=1 - (E19/E17)	

Linear Regression Analysis

추정치 적합도



Linear Regression Analysis



엑셀의 데이터 분석 메뉴를 이용하여 회귀분석 결과값 도출.

- Y절편
- 기울기
- R제곱값

노동력이 **350** 일 때 생산량 **498**개가 나왔을 때의 해석 ->

노동력을 독립변수, 생산량을 종속변수로 한 선형회귀분석에서 Y절편 값은 **204.81**, 기울기는 **0.84**로 나타났다. R제곱값(결정계수)이 **.95**이므로 노동력이 **350**일때 생산량이 **498**개가 나올 확률이 **95%**로 나타났다. 모형적합도인 분산분석, 노동력, Y의 절편 P값도 통계적으로 유의하여 회귀모형이 적합한 것으로 나타났다.

Linear Regression Analysis (R에서의 결과와 동일)

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

regression.R x regression1.R x x_y x

	y	x	x
1	428	267	
2	430	263	
3	417	238	
4	384	219	
5	432	274	
6	425	257	
7	474	321	
8	462	305	
9	449	285	
10	405	247	

Showing 1 to 10 of 10 entries

Console D:\백데이터기획에서분석/Data/

```
> library(readxl)
> x_y <- read_excel("D:/x_y.xlsx")
> view(x_y)
> lm(x_y)

Call:
lm(formula = x_y)

Coefficients:
(Intercept)          x
    204.8125      0.8438

> m <- lm(x_y)
> summary(m)

Call:
lm(formula = x_y)

Residuals:
    Min       1Q   Median       3Q      Max
-8.2188 -3.5234 -0.9063  3.3281 11.3750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  204.8125    17.60510   11.63 2.71e-06 ***
x             0.84375     0.06541    12.90 1.23e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.991 on 8 degrees of freedom
Multiple R-squared:  0.9541, Adjusted R-squared:  0.9484
F-statistic: 166.4 on 1 and 8 DF, p-value: 1.233e-06
> |
```

Environment History Connections

Global Environment

Data

- m List of 12
- x_y 10 obs. of 2 variables

Values

x	chr [1:4]	"a"	"a"	"b"	"c"
---	-----------	-----	-----	-----	-----

Files Plots Packages Help Viewer

Install Update

Name	Description	Version
System Library		
assertthat	Easy Pre and Post Assertions	0.2.0
BH	Boost C++ Header Files	1.66.0-1
bindr	Parametrized Active Bindings	0.1.1
bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2.2
bit	A Class for Vectors of 1-Bit Booleans	1.1-14
bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBs')	1.1.1
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-20
cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
class	Functions for Classification	7.3-14
cli	Helpers for Developing Command Line Interfaces	1.0.0
cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.7-1
codetools	Code Analysis Tools for R	0.2-15
colorspace	Color Space Manipulation	1.3-2
compiler	The R Compiler Package	3.5.1
corplot	Visualization of a Correlation Matrix	0.84
crayon	Colored Terminal Output	1.3.4
curl	A Modern and Flexible Web Client for R	3.2
datasets	The R Datasets Package	3.5.1
DBI	R Database Interface	1.0.0
devtools	Tools to Make Developing R Packages Easier	1.13.6
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create Compact Hash Digests of R Objects	0.6.15
dplyr	A Grammar of Data Manipulation	0.7.6
fansi	ANSI Control Sequence Aware String Functions	0.2.3

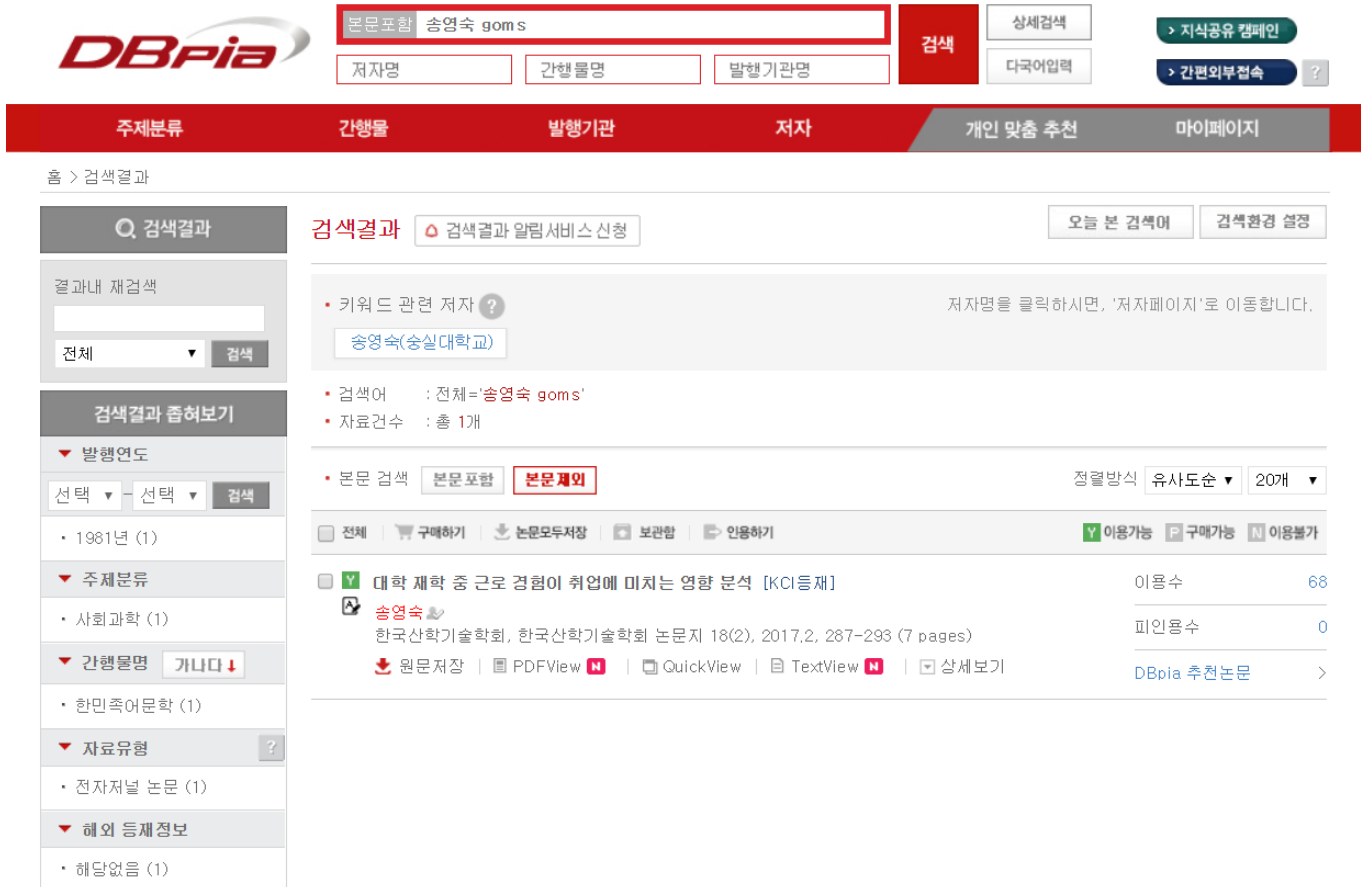
logistic Regression Analysis

로지스틱 회귀(logistic regression)는 D.R.Cox가 1958년 에 제안한 확률 모델로서 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계 기법이다.

로지스틱 회귀는 종속변수가 이항형 문제(즉, 유효한 범주의 개수가 두개인 경우)를 지칭할 때 사용된다. 이외에, 두 개 이상의 범주를 가지는 문제가 대상인 경우엔 다항 로지스틱 회귀 ([multinomial logistic regression](#)) 또는 분화 로지스틱 회귀 (polytomous logistic regression)라고 하고 복수의 범주이면서 순서가 존재하면 서수 로지스틱 회귀 ([ordinal logistic regression](#)) 라고 한다.^[2] 로지스틱 회귀 분석은 의료, 통신, 데이터마이닝과 같은 다양한 분야에서 분류 및 예측을 위한 모델로서 폭넓게 사용되고 있다.

source : 위키백과

logistic Regression Analysis



The screenshot shows the DBpia search results page. The top navigation bar includes '주제분류' (Subject Classification), '간행물' (Publications), '발행기관' (Publishing Institutions), '저자' (Authors), '개인 맞춤 추천' (Personalized Recommendations), and '마이페이지' (My Page). The search bar at the top contains '논문포함 송영숙 goms'. The left sidebar shows filters for '발행연도' (1981 (1)), '주제분류' (사회과학 (1)), '간행물명' (가나다 ↓), '자료유형' (전자저널 논문 (1)), and '해외 등재 정보' (해당없음 (1)). The main content area displays search results for '송영숙(송실대학교)'. The first result is '대학 재학 중 근로 경험이 취업에 미치는 영향 분석 [KCI등재]' by 송영숙, published in '한국산학기술학회, 한국산학기술학회 논문지 18(2), 2017.2, 287-293 (7 pages)'. The result is marked as '이용수 68' (Usage 68) and 'DBpia 추천논문' (DBpia Recommended Paper).

학교 전자도서관으로 들어가 그림의 논문을 다운받습니다.

(GOMS 공공자료를 이용한 로지스틱 회귀분석방법을 이용하였습니다.)

logistic Regression Analysis

Table 1. Characteristics of Respondents

Spec.		N(%)
Gender	Male	9528(52.5)
	Female	8632(47.5)
Field of Study	Humanities	2087(11.5)
	Society	3524(19.4)
	Education	1513(8.3)
	Engineering	4986(27.5)
	Natural Science	2290(12.6)
	Medical Science	1151(6.3)
	Art and Sports	2609(14.4)
Institution Type -Ownership	National	3573(19.7)
	Public	171(0.9)
	Private	14412(79.4)
	Other	4(0.0)
Institution Type -Year	2 Year	5395(29.7)
	4 Year	12325(67.9)
	University of Education	440(2.4)
Institution Area	Seoul	4075(22.4)
	Kyunggi	4711(25.9)
	Chungchung	2904(16.0)
	Youngnam	4299(23.7)
	Honam	2171(12.0)
All		18160(100.0)

Table 3. Difference in Employment and Permanent Employment according to Internship

Spec.		Internship		χ^2/p
		Yes	No	
Employment	Yes	426(2.3)	12236(67.4)	4.200* /.044
	No	153(0.8)	5345(29.4)	
Permanent employment	Yes	342(1.9)	9218(50.8)	9.901** /.002
	No	237(1.3)	8386(46.1)	
All		579(3.2)	17581(96.8)	

*p< .05, **p< .01, ***p< .001

인구통계적 특성이
빈도분석으로 잘 나타나
있습니다.
원 데이터는 GOMS 2013년
데이터 입니다.

재학중 근로경험유무, 인턴경험유무의
집단간 분포의 차이를 카이제곱
검증으로 봤습니다.

logistic Regression Analysis

Table 5. The Effect of Internship and The Relatedness of Work Experience and Employment on College Graduates' Permanent Employment

Variables	B	S.E.	Wald	p	Exp(B)
Internship	.003	.015	.048	.826	1.003
The Relatedness of Work Experience and Employment	.199	.088	5.172	.023*	1.220
Constants	.158	.039	16.307	.000	1.171

*p< .05, **p< .01, ***p< .001

관련전공 일경험유무는 유의 확률 값이 .023으로 유의수준 .05 미만으로 나타나 정규직 취업에 통계적으로 유의한 영향이 있는 것으로 나타났다. 그러나 인십턴경험은 정규직 취업에 통계적으로 유의한 영향을 주지 않는 것으로 나타났다.

(관련전공 일경험한 사람이 그렇지 못한 사람에 비하여 정규직으로 갈 확률이 1.22배 높다는 것을 알 수 있다.