Model Selection with 근본 완성편

Roh

Nested Model

- Two models are said to be **‘nested’** if one of the models constitutes a special case of the other model
- A linear regression model which contains ‘temperature’ as a covariate, is, for example, nested within an otherwise identical model that includes both ‘temperature’ and ‘rainfall’ as covariates, because the former model can be obtained by fixing the coefficient associated with ‘rainfall’ in the latter model to be zero.
- Model A : Homeless = Gender
- Model B : Homeless = Gender + Substance
- 모델 A 는 B에 nested 되어 있다

Stepwise / Forward / Backward Selection 은 태생적으로 nested model들의 비교이다

Intercept only

- Model A:
  Homeless = b0

…

Full model

- Model B:
  Homeless = b0 + b1 Gender + b2 Substance + b3 i1 + b4 Sexrisk + b5 indtot

Comparing two nested models : Likelihood

Model A is nested within B

- L(B) > L(A) as Model A is a special case of Model B

Because logistic regression predicts probabilities, rather than just classes, we can fit it using likelihood. For each training data-point, we have a vector of features, x_i, and an observed class, y_i. The probability of that class was either p, if $y_i = 1$, or $1 - p$, if $y_i = 0$. The likelihood is then

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \tag{12.6}$$

What can we do with likelihood?

변수를 추가 했을경우 (모델 complexity 상승)

- 복잡한 모델 B가 간단한 모델 A보다 Likelihood는 높음
  - 왜냐면 둘다 Likelihood가 Maximized 되게끔 Parameter Set이 추정됨
  - 왜냐면 B의 스페셜 case가 모델 A
  - 그럼 항상 복잡한 모델이 좋은건가?
- Likelihood를 가지고 할 수 있는것
  - n = sample size, k = num of variable, L = likelihood
  - AIC = 2k - 2log(L)
  - BIC = klog(n) - 2log(L)
- 하지만, AIC/BIC는 테스트가 아님! (not test statistics)
  - Model selection through AIC and BIC is completely different paradigm

AIC/BIC 는 non-nested model에 좀더 유용함

How to test? - Overall model evaluation

1. Likelihood ratio test

$$-2\ \log\Lambda = -2\ \log\left(\frac{L_0}{L_1}\right) = -2\left(\ell_0 - \ell_1\right) \xrightarrow{d} \chi_k^2$$

2. Rao's Score test (a.k.a Lagrange Multiplier test)

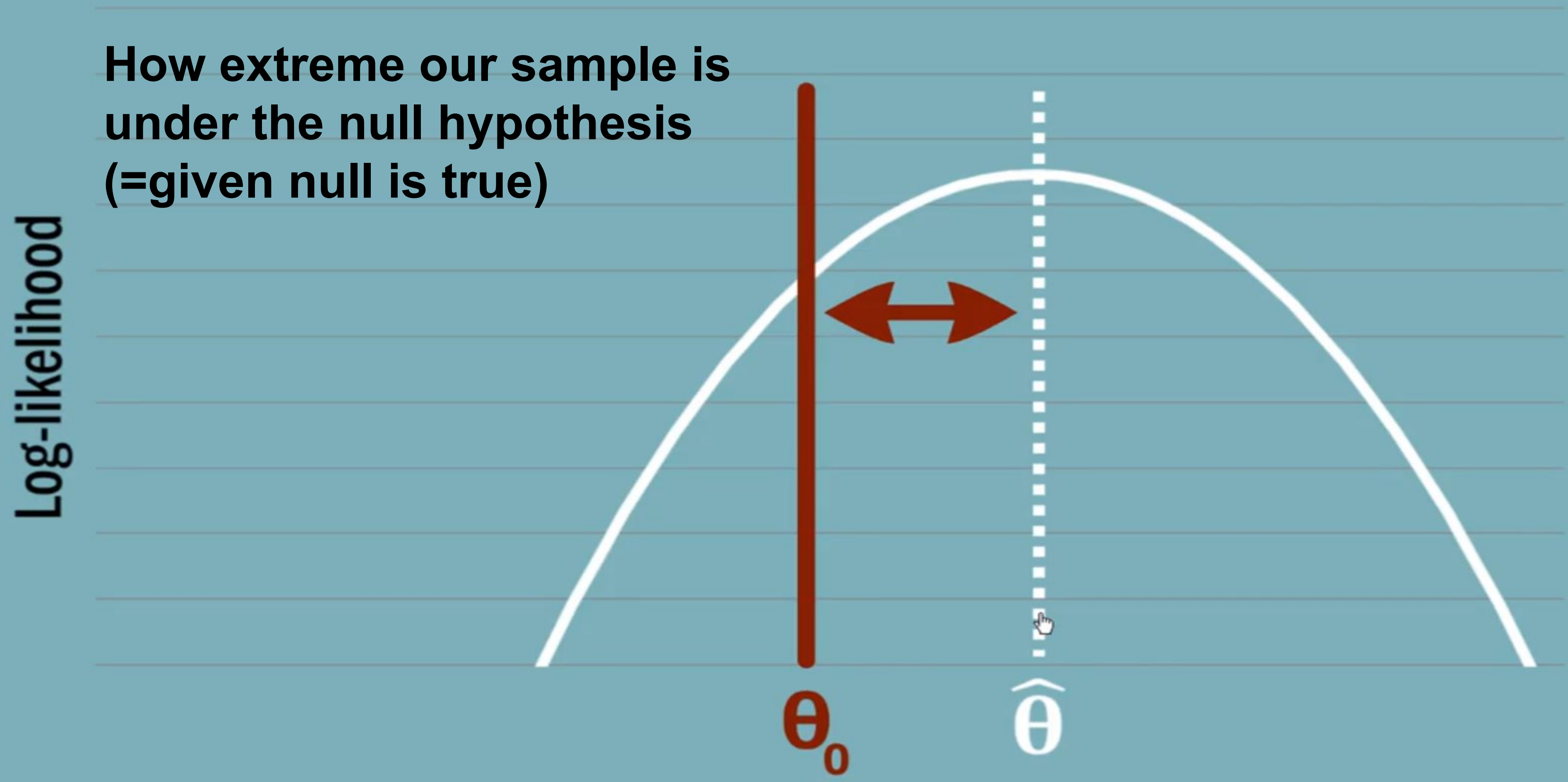
$$S = \left(\frac{\partial L(\hat{\beta}_0)}{\partial \beta}\right)^T - E\left(\frac{\partial^2 L(\hat{\beta}_0)}{\partial \beta \partial \beta'}\right)^{-1} \frac{\partial L(\hat{\beta}_0)}{\partial \beta} \xrightarrow{d} \chi_k^2$$

3. Wald test

$$W = (\hat{\beta} - \beta_0)^T \left[Cov(\hat{\beta})\right]^{-1} (\hat{\beta} - \beta_0) \xrightarrow{d} \chi_p^2$$

Log-likelihood function $\ell(\theta)$

How extreme our sample is under the null hypothesis (=given null is true)



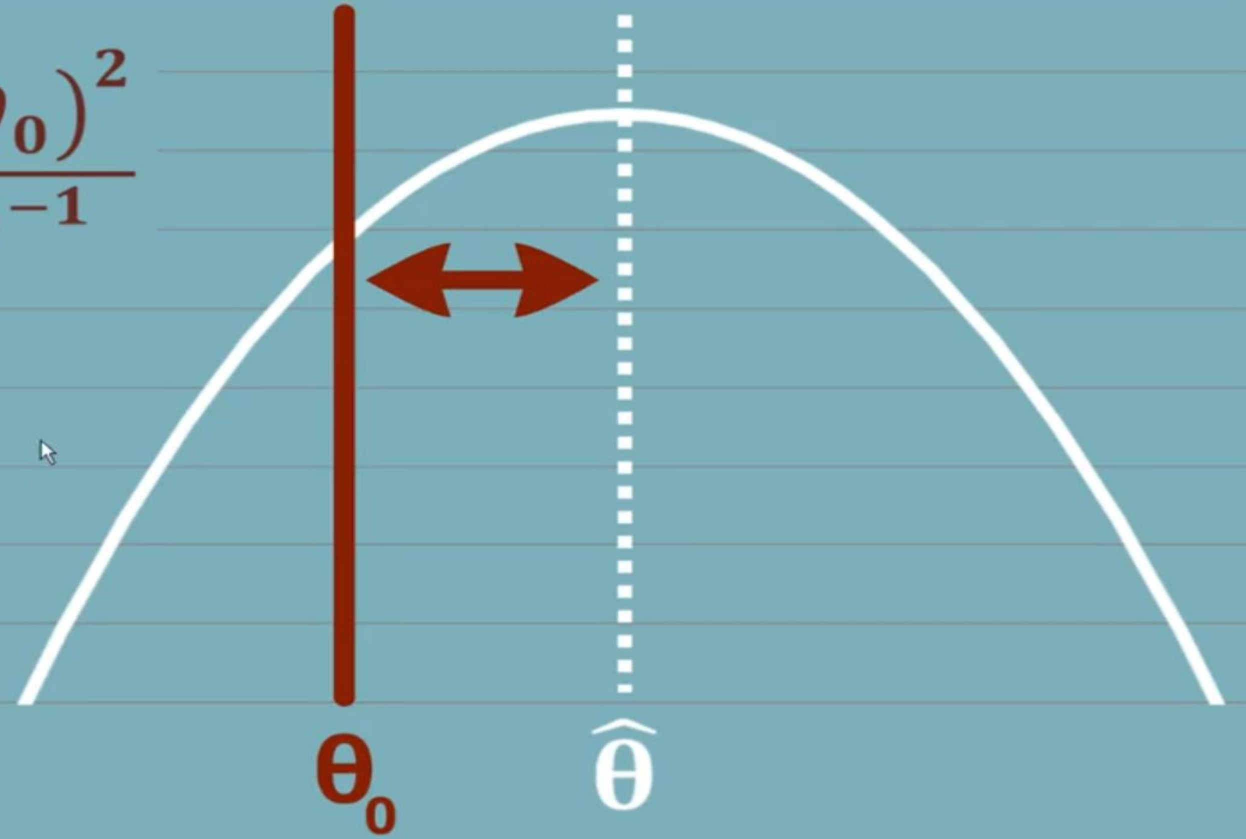
Log-likelihood function $\ell(\theta)$

Wald test

$$t_W = \frac{(\hat{\theta} - \theta_0)^2}{I(\theta_0)^{-1}}$$

$$\sim \chi^2(1)$$

Log-likelihood

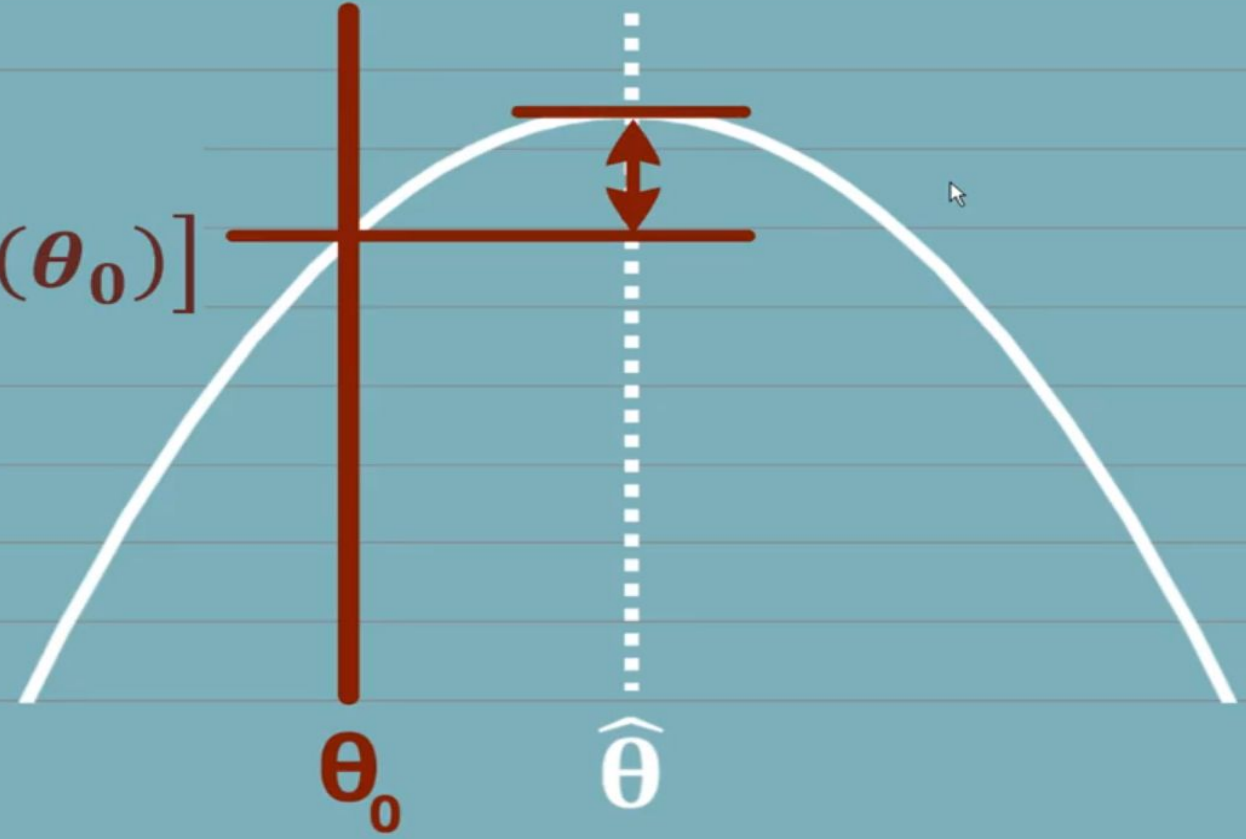


Log-likelihood function $\ell(\theta)$

Likelihood ratio test

$$t_{LR} = 2\left[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta_0})\right]$$

$$\sim \chi^2(1)$$



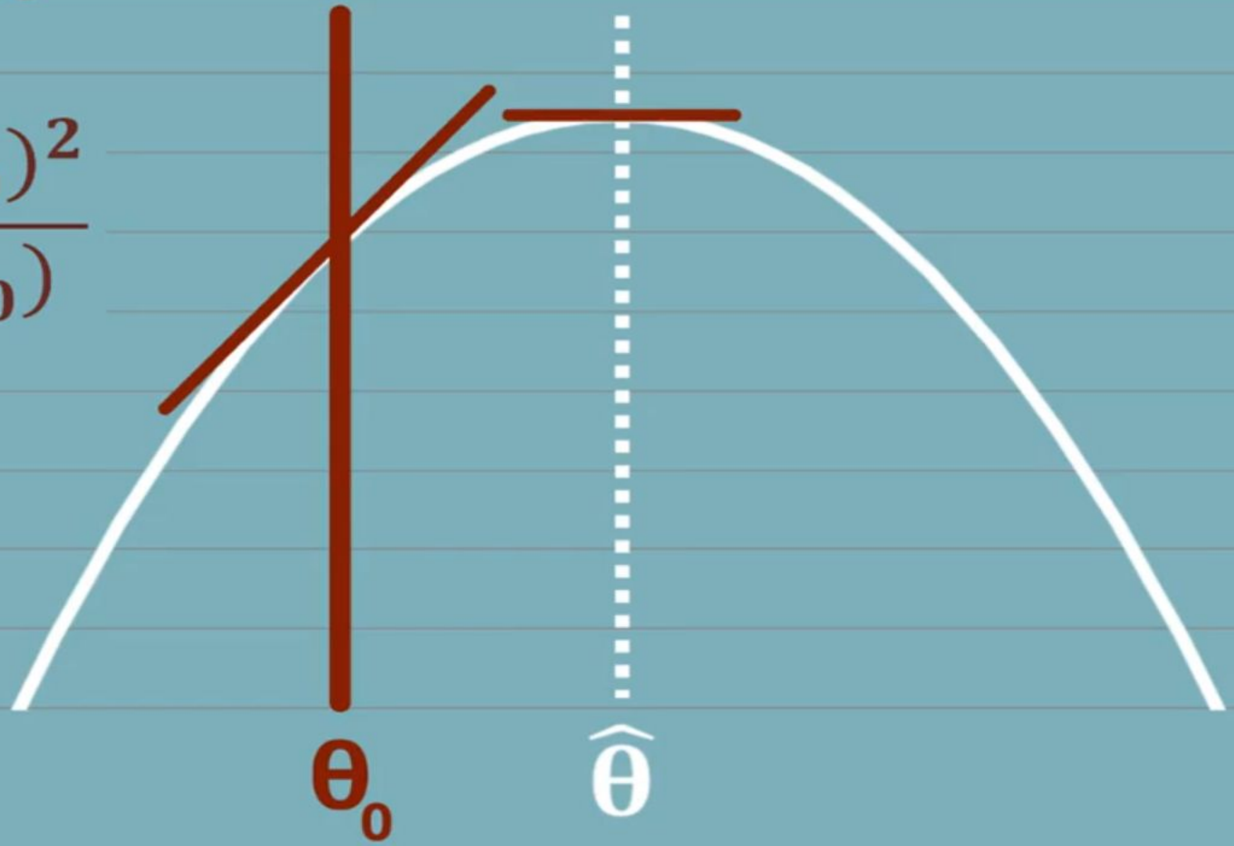
Log-likelihood function $\ell(\theta)$

Score test

$$t_S = \frac{S(\theta_0)^2}{I(\theta_0)}$$

$$\sim \chi^2(1)$$

Log-likelihood



3 개 중에 뭘 써야하나?

- n이 클때는 3개 다 유사하다 (asymptotic)
- SAS Logistic 에서 모델 Selection 을 할때는 Rao Score를 쓴다
- Global test 를 할때는 Wald 를 쓴다
- R에서 Rao나 LRT를 쓰는 방법은 anova 함수를 사용하면 된다
- R에서 Wald를 쓸때는 ‘aod’ package를 쓰면 된다
    - test (beta_i =0) 하고자 하는 파라미터가 전체라면 (0, 0, 0, … , 0) 인 joint test 가 된다

Health Evaluation and Linkage to Primary Care (HELP 데이터)

Homeless	0=No, 1=Yes	지난 6개월동안 길거리 노숙한 적있는 사람
Gender	0=male, 1=female	
i1	0-142	지난 한달간 일일 평균 주량
Substance	술, 코카인, 헤로인	주요 약물 종류
Sexrisk	0-21	성관련 리스크 스코어, 높을수록 안좋음
indtot	0-184	중독약물 관련 설문지 스코어
Age		나이
Racegrp	white, black, hispanic, others	

SAS 결과물
R 에서 똑같이 만들어보기

1

```sas
proc logistic data=help descending;
   class substance (param=ref ref='alcohol');
   model homeless = female i1 substance sexrisk indtot;
run;
```

````
```{r}
df$substance = relevel(df$substance, ref = 'alcohol')
logres =
  glm(homeless ~ female + i1 + substance + sexrisk + indtot,
      contrasts = list(substance = contr.treatment), #dummy coding
      binomial, data=df)
summary(logres)
```
````

2

Class Level Information

Class	Value	Design Variables	
SUBSTANCE	alcohol	0	0
	cocaine	1	0
	heroin	0	1

Dummy Coding (left) & Effect Coding (right)

```
$substance
        2 3
alcohol 0 0
cocaine 1 0
heroin  0 1
```

```
$substance
        [,1] [,2]
alcohol    1    0
cocaine    0    1
heroin    -1   -1
```

3

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	627.284	590.652
SC	631.400	619.463
-2 Log L	625.284	576.652

```
```{r}
n = dim(df)[1]
k = 1
aic = -2*logLik(logres_intcept)[1] + 2*k
bic = -2*logLik(logres_intcept)[1] + log(n)*k
sprintf('aic is %.3f', aic)
sprintf('bic is %.3f', bic)
sprintf('-2LogL is %.3f', -2*logLik(logres_intcept)[1])
```
```

```
[1] "aic is 627.284"
[1] "bic is 631.400"
[1] "-2LogL is 625.284"
```

```
```{r}
n = dim(df)[1]
k = dim(df)[2]+1
aic = -2*logLik(logres)[1] + 2*k
bic = -2*logLik(logres)[1] + log(n)*k
sprintf('aic is %.3f', aic)
sprintf('bic is %.3f', bic)
sprintf('-2LogL is %.3f', -2*logLik(logres)[1])
```
```

```
[1] "aic is 590.652"
[1] "bic is 619.463"
[1] "-2LogL is 576.652"
```

4

```
          Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        48.6324        6         <.0001
Score                   45.6522        6         <.0001
Wald                    40.7207        6         <.0001
```

Globaltests output assess the joint null hypothesis that all parameters except the intercept equal 0

```
score = anova(logres_intcept,logres, test = 'Rao')
cbind(score)
```

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Rao <dbl>	Pr(>Chi) <dbl>
1	452	625.2845	NA	NA	NA	NA
2	446	576.6520	6	48.63245	45.65242	3.471746e-08

```
wald = wald.test(b=coef(logres), Sigma = vcov(logres), Terms = 1:dim(df)[2]+1)
wald$result$chi2
```

```
        chi2           df            P
4.072111e+01 6.000000e+00 3.286007e-07
```

5

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
FEMALE	1	1.0831	0.2980
I1	1	7.6866	0.0056
SUBSTANCE	2	4.2560	0.1191
SEXRISK	1	3.4959	0.0615
INDTOT	1	8.2868	0.0040

The ODS type3 output contains tests for each covariate (including joint tests for class variables with two or more values) conditional on all other covariates being included in the model.

Effect <chr>	chi2 <dbl>	df <dbl>	P <dbl>
female	1.083131	1	0.297998255
i1	7.686783	1	0.005562669
substance	4.256031	2	0.119073377
sexrisk	3.495973	1	0.061518259
indtot	8.286941	1	0.003993119

SAS는 Wald로 했는데 Rao로 해도 무방함!
Function 만들어서 자동으로 돌리기에는 Rao가 편함!

6

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.1319	0.6335	11.3262	0.0008
FEMALE	1	-0.2617	0.2515	1.0831	0.2980
I1	1	0.0175	0.00631	7.6866	0.0056
SUBSTANCE cocaine	1	-0.5033	0.2645	3.6206	0.0571
SUBSTANCE heroin	1	-0.4431	0.2703	2.6877	0.1011
SEXRISK	1	0.0725	0.0388	3.4959	0.0615
INDTOT	1	0.0467	0.0162	8.2868	0.0040

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.13192130	0.633470892	-3.365461	0.0007641599
female	-0.26170295	0.251459584	-1.040736	0.2979982548
i1	0.01748948	0.006308187	2.772505	0.0055626693
substancecocaine	-0.50334542	0.264529994	-1.902791	0.0570677590
substanceheroin	-0.44313645	0.270302572	-1.639409	0.1011281272
sexrisk	0.07250902	0.038780017	1.869752	0.0615182589
indtot	0.04668849	0.016218576	2.878705	0.0039931193

7

sexrisk 1유닛 증가할때마다 homeless 될 odd가 1.075배 증가 (기존에는 log-odd)

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
FEMALE	0.770	0.470	1.260
I1	1.018	1.005	1.030
SUBSTANCE cocaine vs alcohol	0.605	0.360	1.015
SUBSTANCE heroin vs alcohol	0.642	0.378	1.091
SEXRISK	1.075	0.997	1.160
INDTOT	1.048	1.015	1.082

```
exp(cbind(OR = coef(logres), confint(logres)))
```

```
Waiting for profiling to be done...
                      OR      2.5 %    97.5 %
(Intercept)      0.1186092 0.03307666 0.3988986
female           0.7697396 0.46813610 1.2573692
i1               1.0176433 1.00548459 1.0307037
substancecocaine 0.6045050 0.35921120 1.0148768
substanceheroin  0.6420196 0.37707881 1.0898444
sexrisk          1.0752025 0.99697625 1.1610664
indtot           1.0477956 1.01573979 1.0826271
```

Stepwise in R

Selection Entry: 0.3
Selection Stay: 0.35

Age 와 Race를 추가한 모델

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -2.772789   0.834716  -3.322 0.000894 ***
female           -0.281222   0.252795  -1.112 0.265945    
i1                0.016883   0.006351   2.658 0.007853 ** 
substancecocaine -0.293095   0.289485  -1.012 0.311313    
substanceheroin  -0.408693   0.278440  -1.468 0.142160    
sexrisk           0.084311   0.039726   2.122 0.033810 *  
indtot            0.042709   0.016399   2.604 0.009205 ** 
age               0.013004   0.013715   0.948 0.343060    
racegrpwhite      0.429851   0.256989   1.673 0.094397 .  
racegrphispanic   0.259021   0.359435   0.721 0.471134    
racegrpother      0.260721   0.445219   0.586 0.558143    
```

Step1: Intercept only ⇒ Choose li

Effect <chr>	Df <dbl>	Deviance <dbl>	Rao <dbl>	Pr(>Chi) <dbl>	Lik <chr>
female	1	4.365390	4.319663	0.0376744	-310.46
i1	1	27.186776	25.543496	0.0000004	-299.05
substance	2	17.066582	17.012163	0.0002022	-304.11
sexrisk	1	4.199706	4.189231	0.0406816	-310.54
indtot	1	22.416662	21.046221	0.0000045	-301.43
age	1	3.349754	3.344173	0.0674434	-310.97
racegrp	3	8.538131	8.525780	0.0363078	-308.37

Step2: Intercept + li ⇒ Choose indtot

Effect <chr>	Df <dbl>	Deviance <dbl>	Rao <dbl>	Pr(>Chi) <dbl>	Lik <chr>
female	1	2.9647237	2.9394924	0.0864379	-297.57
substance	2	3.8534595	3.8965744	0.1425180	-297.12
sexrisk	1	3.0045201	2.9953111	0.0835059	-297.55
indtot	1	14.3507174	13.7559005	0.0002082	-291.87
age	1	0.7278456	0.7287849	0.3932773	-298.68
racegrp	3	5.2103688	5.2252045	0.1560301	-296.44

Step2-2: Intercept + li + indtot ⇒ Remove Nothing

Step3: Intercept + li + indtot ⇒ Choose sexrisk

Effect <chr>	Df <dbl>	Deviance <dbl>	Rao <dbl>	Pr(>Chi) <dbl>	Lik <chr>
female	1	0.6051738	0.6041358	0.4370044	-291.57
substance	2	2.9934185	3.0214244	0.2207527	-290.38
sexrisk	1	1.8510630	1.8471898	0.1741106	-290.95
age	1	0.8375619	0.8386890	0.3597720	-291.45
racegrp	3	2.6049349	2.6150177	0.4548628	-290.57

Step3-2: Intercept + li + indtot + sexrisk⇒ Remove Nothing

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.729780   0.575900  -4.740 2.14e-06 ***
i1           0.023359   0.005775   4.045 5.24e-05 ***
sexrisk      0.048709   0.035921   1.356 0.175092
indtot       0.053891   0.015503   3.476 0.000508 ***
```

Step4: Intercept + li + indtot + sexrisk ⇒ Choose substance

Effect <chr>	Df <dbl>	Deviance <dbl>	Rao <dbl>	Pr(>Chi) <dbl>	Lik <chr>
female	1	1.000547	0.9977198	0.3178629	-290.45
substance	2	4.154531	4.1965955	0.1226651	-288.87
age	1	1.258767	1.2602753	0.2615990	-290.32
racegrp	3	3.886806	3.8932461	0.2732248	-289

Step4-2: Intercept + li + indtot + sexrisk + substance ⇒ Remove Nothing

```
wald = wald.test(b=coef(logres), Sigma = vcov(logres), Terms = 5:6)
wald
```

```
Wald test:
----------

Chi-squared test:
X2 = 4.2, df = 2, P(> X2) = 0.12
```

Step5: Intercept + li + indtot + sexrisk + substance ⇒ Choose age

Effect <chr>	Df <dbl>	Deviance <dbl>	Rao <dbl>	Pr(>Chi) <dbl>	Lik <chr>
female	1	1.000547	0.9977198	0.3178629	-290.45
age	1	1.258767	1.2602753	0.2615990	-290.32
racegrp	3	3.886806	3.8932461	0.2732248	-289

Step5-2: Intercept + li + indtot + sexrisk + substance + age ⇒ Remove age

```
wald = wald.test(b=coef(logres), Sigma = vcov(logres), Terms = 7)
wald
```
```

Wald test:
----------

Chi-squared test:
X2 = 0.64, df = 1, P(> X2) = 0.42
```

Step6: Intercept + li + indtot + sexrisk + substance ⇒ Choose racegrp

Effect <chr>	Df <dbl>	Deviance <dbl>	Rao <dbl>	Pr(>Chi) <dbl>	Lik <chr>
female	1	1.000547	0.9977198	0.3178629	-290.45
age	1	1.258767	1.2602753	0.2615990	-290.32
racegrp	3	3.886806	3.8932461	0.2732248	-289

Step5-2: Intercept + li + indtot + sexrisk + substance + racegrp ⇒ Remove racegrp

```
wald = wald.test(b=coef(logres), Sigma = vcov(logres), Terms = 7:9)
wald
```

```
Wald test:
----------

Chi-squared test:
X2 = 2.7, df = 3, P(> X2) = 0.44
```

Done.

Useful Website

- [Logistic Regression Model Comparison]
- [How to test for simultaneous equality of choosen coefficients in logit or probit model?]
- [Odd Ratios]