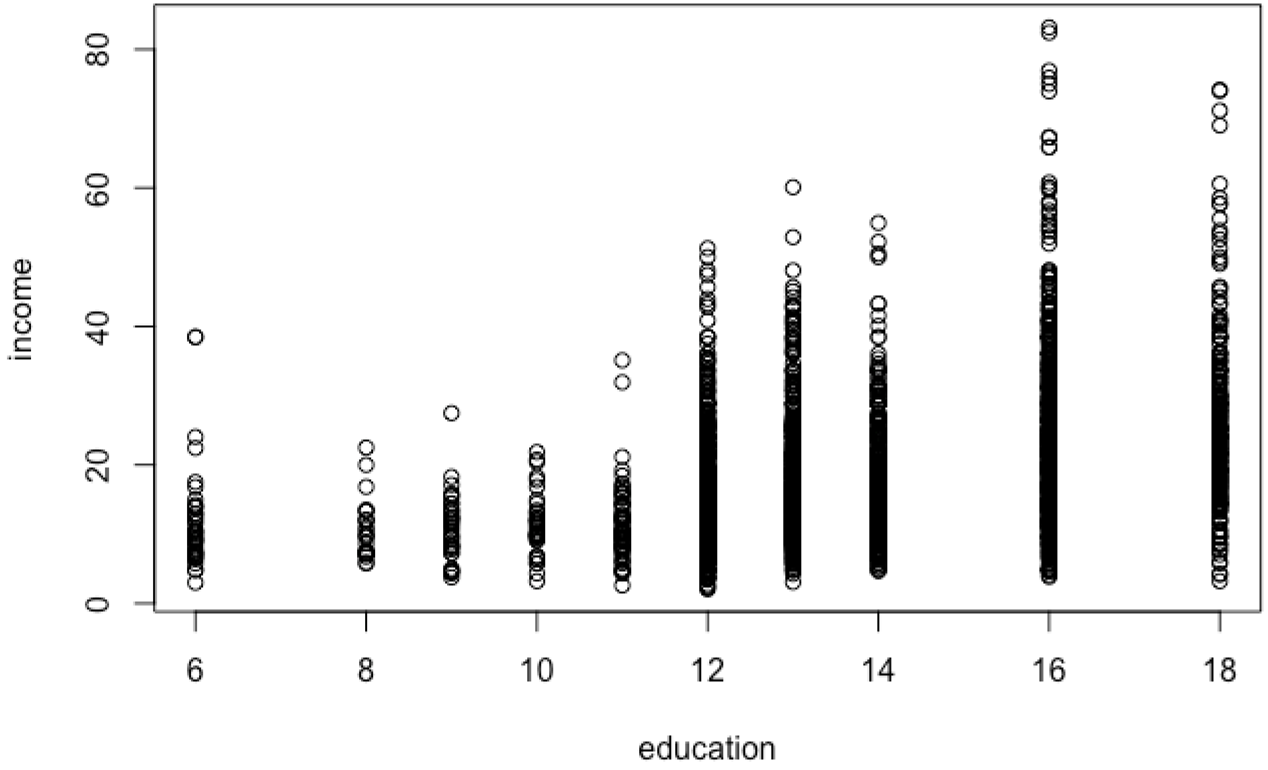
Heteroskedasticity

```
# Set your working directory
setwd("C:/PAPP")

# Verify that your working directory
getwd()

#
rm(list = ls())
```

```
##
install.packages("lmtest")
install.packages("sandwich")

##
library(lmtest)
library(sandwich)
library(texreg)
library(dplyr)
```

```
##
cps <- read.csv("cps2013.csv")

##
head(cps)
```

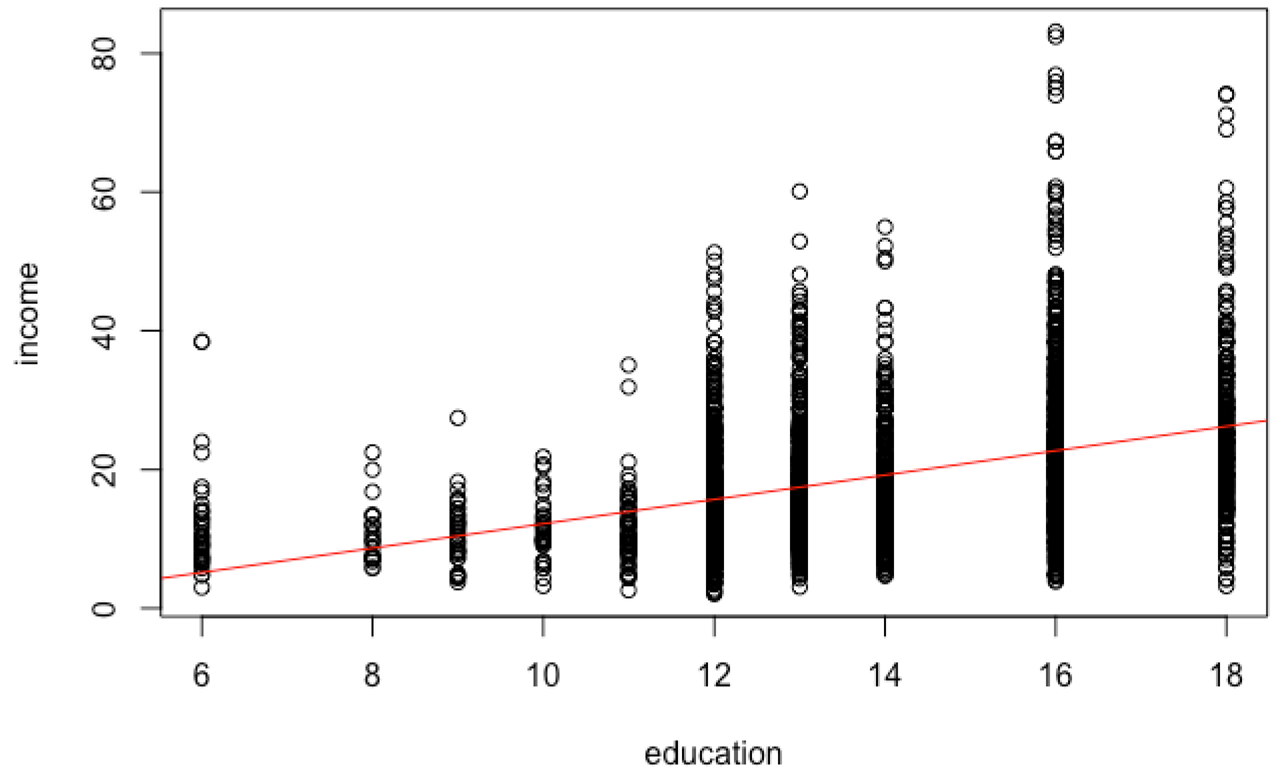
```
##
plot(income ~ education, data = cps)
```


income
education
0
20
40
60
80
6
8
10
12
14
16
18

```
##
model1 <- lm(income ~ education, data = cps)
summary(model1)
```

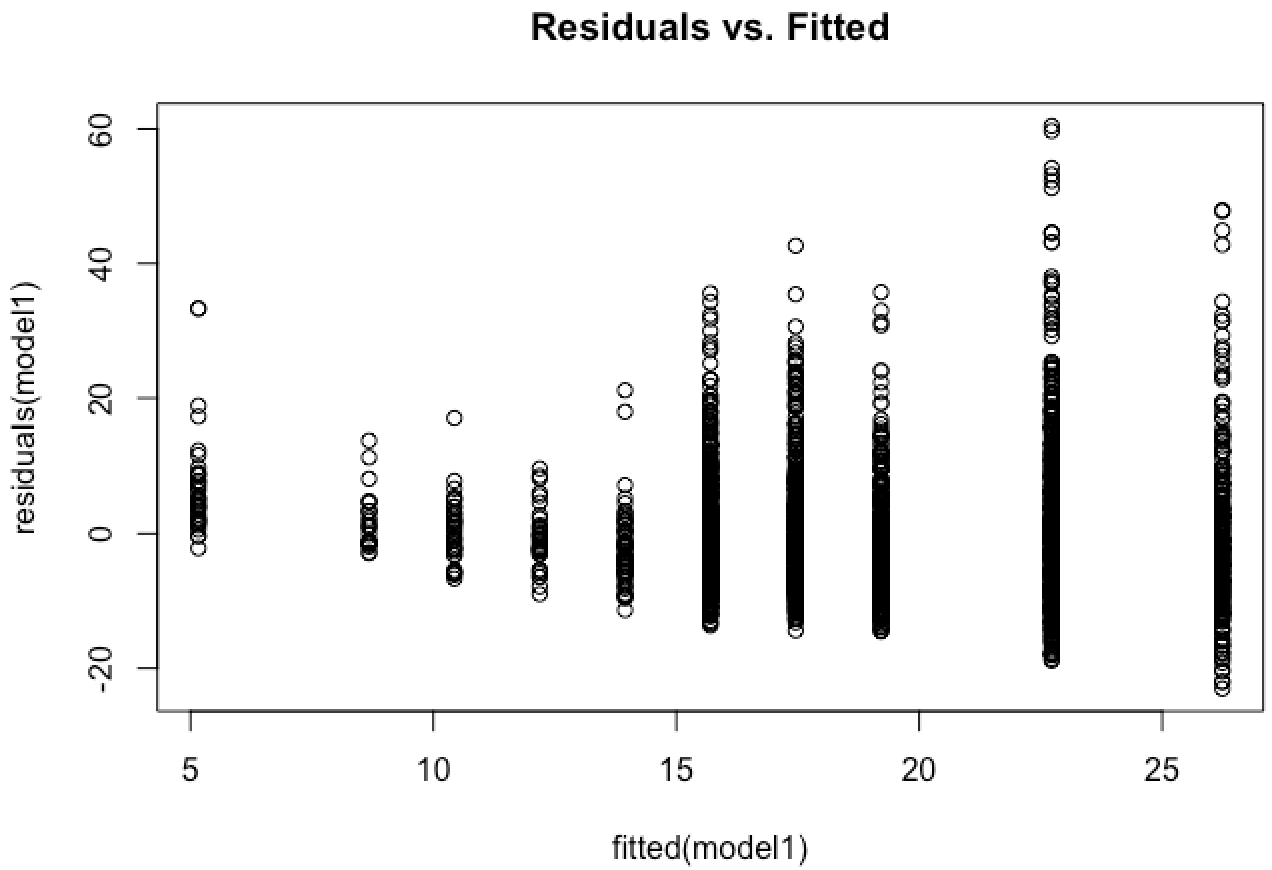
```
Call:
lm(formula = income ~ education, data = cps)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.035  -6.086  -1.573   3.900  60.446 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.37508    1.03258  -5.206 2.07e-07 ***
education    1.75638    0.07397  23.745  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.496 on 2987 degrees of freedom
Multiple R-squared:  0.1588,    Adjusted R-squared:  0.1585 
F-statistic: 563.8 on 1 and 2987 DF,  p-value: < 2.2e-16
```

```
##
plot(income ~ education, data = cps)
abline(model1, col = "red")
```



income
0
20
40
60
80
6
8
10
12
14
16
18
education

```
##
plot(fitted(model1), residuals(model1), main = "Residuals vs.
Fitted")
```


Residuals vs. Fitted
residuals(model1)
60
40
20
0
-20
5
10
15
20
25
fitted(model1)

```
##
bptest(model1)
```

```
 studentized Breusch-Pagan test

data:  model1
BP = 76.679, df = 1, p-value < 2.2e-16
```

```
##
screenreg(coeftest(model1, vcov = vcovHC(model1)))
```

```
======================
              Model 1
----------------------
(Intercept)  -5.38 ***
             (1.05)
education     1.76 ***
             (0.08)
======================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

VIII. 로지스틱 회귀분석

1. 로지스틱 회귀분석의 이해

- 선형 OLS의 등분산성/정규성

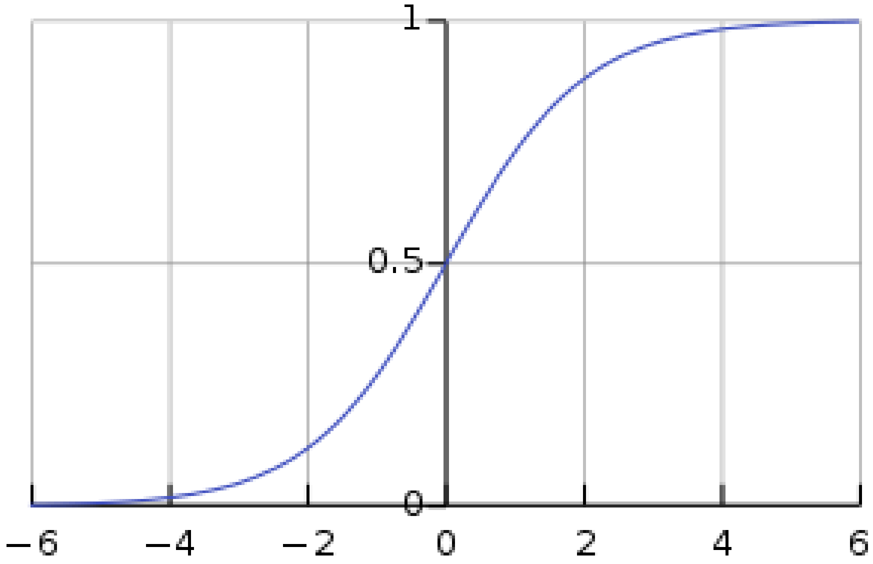
1. 로지스틱 회귀분석의 이해

- 이항분포에 대한 이해

1. 로지스틱 회귀분석의 이해

- 승산 Odds, 그리고 승산비 Odds ratio

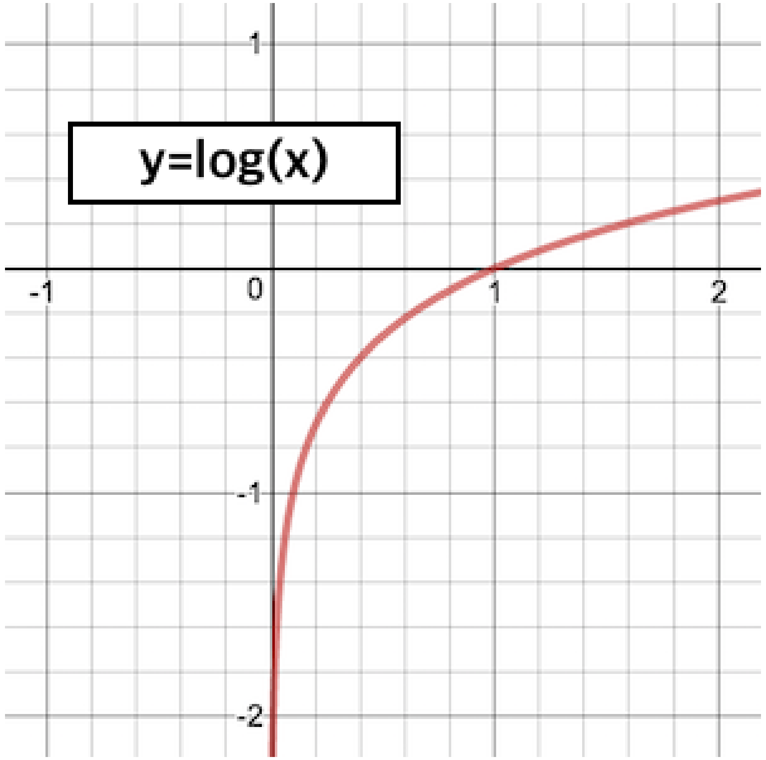
확률	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99	1
승산	0	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4.00	9.00	99.0 0	∞



https://ko.wikipedia.org/wiki/%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1_%ED%9A%8C%EA%B7%80

1. 로지스틱 회귀분석의 이해

- 로그 오즈 (Log Odds)   Logit



https://pythonkim.tistory.com/28

1. 로지스틱 회귀분석의 이해

- 로그 오즈 (Log Odds) Logit

확률	p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
승산	p/1-p	0.111	0.25	0.429	0.667	1	1.5	2.333	4	9
로짓	ln(p/1-p)	-2.198	-1.386	-0.846	-0.405	0	0.405	0.847	1.386	2.197

2. 로지스틱 회귀분석의 추정

$$log\left(\frac{p}{1-p}\right) = a + b_1x_1 + \cdots + b_kx_k$$

- 최대우도추정법 Maximum Likelihood Estimation MLE

3. 로지스틱 회귀분석의 유의도

- Convergence (수렴)과 Separation (분리) 문제

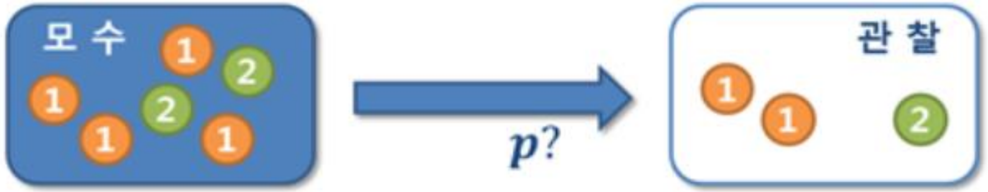
3. 로지스틱 회귀분석의 유의도

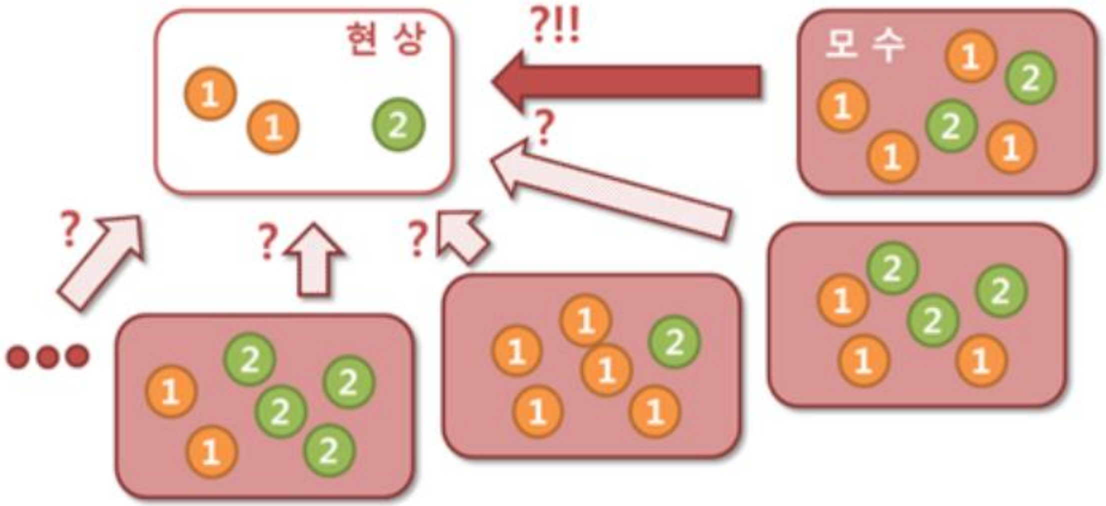
- 유사 R-square 검정

https://pythonkim.tistory.com/28

3. 로지스틱 회귀분석의 유의도

- 우도비 검정 Likelihood ratio test





https://dermabae.tistory.com/188

3. 로지스틱 회귀분석의 유의도

- 우도비 검정 Likelihood ratio test

함수	범위	완벽한 적합도	좋은 적합도	나쁜 적합도
우도(L)	$0 \le L \le 1$	1	1에 근접	0에 근접
ln우도(lnL)	$-\infty < \ln L \le 0$	0	0에 근접	음수로 감소
-2LL	$0 \le -2LL < \infty$	0	0에 근접	양수로 증가

3. 로지스틱 회귀분석의 유의도

- Information Criteria 검정

$$AIC = -2\ln L + 2q$$

$$AIC' = \frac{-2\ln L + 2q}{N}$$

$$BIC = -G + (df)(\ln N)$$

$$G = -2\ln L_0 - (-2\ln L_M)$$

https://pythonkim.tistory.com/28

3. 로지스틱 회귀분석의 유의도

- 예측력

https://pythonkim.tistory.com/28

4. 로지스틱 회귀분석의 해석

- $\ln(\frac{P(찬성)}{1-P(찬성)})$
  = -1.261 - 0.001 * 연령 + 0.171 * 학력 -0.283 * 보수성 – 0.268 * 여성 – 0.414 * 기혼

- $\frac{P(찬성)}{1-P(찬성)}$
  = Exp(-1.261 - 0.001 * 연령 + 0.171 * 학력 -0.283 * 보수성 – 0.268 * 여성 – 0.414 * 기혼)

https://pythonkim.tistory.com/28

4. 로지스틱 회귀분석의 해석

- 회귀계수 (b)
  독립변수가 1단위 증가시, 종속변수가 '1'일 log odds의 변화량
  변화량이 일정한 선형관계
- 100(Exp(b)-1)
  독립변수가 1단위 증가시, 종속변수가 '1'이 될 odds의 변화율 (%)
  변화율이 일정한 비선형관계

https://pythonkim.tistory.com/28

4. 로지스틱 회귀분석의 해석

- 회귀계수 (b)
  독립변수가 1단위 증가시, 종속변수가 ‘1’일 log odds의 변화량
  변화량이 일정한 선형관계
- 100(Exp(b)-1)
  독립변수가 1단위 증가시, 종속변수가 ‘1’이 될 odds의 변화율 (%)
  변화율이 일정한 비선형관계

https://pythonkim.tistory.com/28

5. 다항 로지스틱의 논리

Binomial Logistic

```
# Set your working directory
setwd("C:/PAPP")

# Verify that your working directory
getwd()

#
rm(list = ls())
```

```
#
library(foreign)
library(Zelig)
library(texreg)
library(dplyr)
library(lmtest)

##
rm(list = ls())

## bes <- read.dta("bes.dta")
```

Variable	Description	Range
Turnout	투표여부	No (0); Yes (1)
Gender	성별	1 (male); 0 (female)
LeftRightSelf	이념적 지향	1 (left) - 11 (right)
CivicDutyIndex	시민적 의무에 대해 느끼는 가치	high values mean high civic duty
polinfoindex	정치에 대한 지식	0 (low) - 8 (high)
edu*	교육연수	binary
in.school	학교 재학 여부	binary
in.uni	대학 재학 여부	binary

```
##

bes$Gender <- factor(bes$Gender, levels = c(0, 1), labels =
c("Female", "Male"))


##

head(bes)
```

```
##
bes <- filter(bes,
              !is.na(Turnout),
              !is.na(Income),
              !is.na(polinfoindex),
              !is.na(Gender),
              !is.na(edu15),
              !is.na(edu17),
              !is.na(edu18),
              !is.na(edu19plus),
              !is.na(in_school),
              !is.na(in_uni))
```

```
##
model1 <- glm(Turnout ~ Income + polinfoindex + Gender +
          edu15 + edu17 + edu18 + edu19plus + in_school + in_uni,
          family = binomial(link = "logit"),
          data = bes)

screenreg(model1)
```

```
=============================
                 Model 1
-----------------------------
(Intercept)        -1.14 ***
                   (0.15)
Income              0.03
                   (0.02)
polinfoindex        0.38 ***
                   (0.02)
GenderMale         -0.35 ***
                   (0.08)
edu15               0.38 ***
                   (0.10)
edu17               0.46 **
                   (0.15)
edu18               0.11
                   (0.14)
edu19plus           0.24 *
                   (0.12)
in_school           0.15
                   (0.39)
in_uni             -0.72 **
                   (0.25)
-----------------------------
AIC              4401.20
BIC              4464.53
Log Likelihood  -2190.60
Deviance         4381.20
Num. obs.        4161
=============================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

```
##
table(bes$Turnout)

##
predicted_probs <- predict(model1, type = "response")

##
expected <- as.numeric(predicted_probs > 0.5)
observed <- bes$Turnout

outcome <- table(observed,expected)
outcome
```

```
##
(outcome[1,1] + outcome[2,2]) / sum(outcome)

##
mean(bes$Turnout)
```

```
##

model2 <- glm(Turnout ~ Income + polinfoindex + Influence +
Gender + Age + edu15 + edu17 + edu18 + edu19plus + in_school +
in_uni, family = binomial(link = "logit"), data = bes)


##

screenreg(list(model1, model2))
```

	Model 1	Model 2
(Intercept)	-1.14 ***	-3.90 ***
	(0.15)	(0.22)
Income	0.03	0.15 ***
	(0.02)	(0.02)
polinfoindex	0.38 ***	0.25 ***
	(0.02)	(0.02)
GenderMale	-0.35 ***	-0.36 ***
	(0.08)	(0.08)
edu15	0.38 ***	-0.34 **
	(0.10)	(0.11)
edu17	0.46 **	0.36 *
	(0.15)	(0.16)
edu18	0.11	0.14
	(0.14)	(0.15)
edu19plus	0.24 *	0.01
	(0.12)	(0.13)
in_school	0.15	1.13 **
	(0.39)	(0.40)
in_uni	-0.72 **	-0.05
	(0.25)	(0.27)
Influence		0.21 ***
		(0.02)
Age		0.05 ***
		(0.00)
AIC	4401.20	4003.90
BIC	4464.53	4079.90
Log Likelihood	-2190.60	-1989.95
Deviance	4381.20	3979.90
Num. obs.	4161	4161

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

```
##
lrtest(model1, model2)
```

```
Likelihood ratio test

Model 1: Turnout ~ Income + polinfoindex +
Gender + edu15 + edu17 + edu18 +
    edu19plus + in_school + in_uni
Model 2: Turnout ~ Income + polinfoindex +
Influence + Gender + Age +
    edu15 + edu17 + edu18 + edu19plus +
in_school + in_uni
  #Df  LogLik Df Chisq Pr(>Chisq)
1  10 -2190.6
2  12 -1990.0  2 401.3  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01
'*' 0.05 '.' 0.1 ' ' 1
```