

Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

16

선형모델의 일반화

경북대학교 배준현 교수
(joonion@knu.ac.kr)



16. 선형모델의 일반화

- 선형모델의 일반화가 필요한 이유:
 - 선형회귀분석을 위한 조건:
 - 결과변수가 연속형 변수이면서 정규분포를 따라야 한다.
 - 선형회귀분석을 위한 조건에 맞지 않는 경우:
 - 결과변수가 범주형 변수일 때: 로지스틱 회귀분석
 - 결과변수가 어떤 사건이 발생하는 횟수일 때: 포아송 회귀분석



16. 선형모델의 일반화

- 일반화 선형모델: *generalized* linear model
 - 선형회귀모델을 확장: 정규분포를 따르지 않는 결과변수에 대한 회귀모델 생성
 - 표준 선형회귀모델: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
 - μ_y : 결과변수의 조건부 평균, x_m : 예측변수, β_m : 회귀계수, m : 변수의 개수
 - 일반 선형회귀모델: $f(\mu_y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
 - $f(\mu_y)$: 결과변수의 조건부 평균의 함수 (*link function*)
 - 표준 선형회귀모델은 일반선형모델의 한 특수한 경우
 - 링크함수가 항등함수: $f(\mu_y) = \mu_y$
 - 확률분포는 정규분포를 따름
 - 회귀계수의 추정: 최대우도법(*MLE, Maximum Likelihood Estimation*)



16. 선형모델의 일반화

- 일반화 선형모델: *generalized linear model*
 - 로지스틱 회귀분석: *logistic regression* analysis
 - 결과변수가 범주형 변수일 때: 정규분포를 따르지 않음
 - 이분 변수(*binary* variable): 예/아니오, 성공/실패, 생존/사망 등
 - 다중 변수(*multicategory* variable): 우수/보통/미흡, A/B/AB/O 등
 - 포아송 회귀분석: *Poisson regression* analysis
 - 결과변수가 어떤 사건이 발생하는 횟수일 때: 포아송 분포를 따름
 - 연간 철도사고횟수, 월간 빈집털이횟수, 일간 상담횟수 등
 - 횟수변수는 포아송 분포를 따르고, 평균과 분산은 종종 상관관계를 가짐



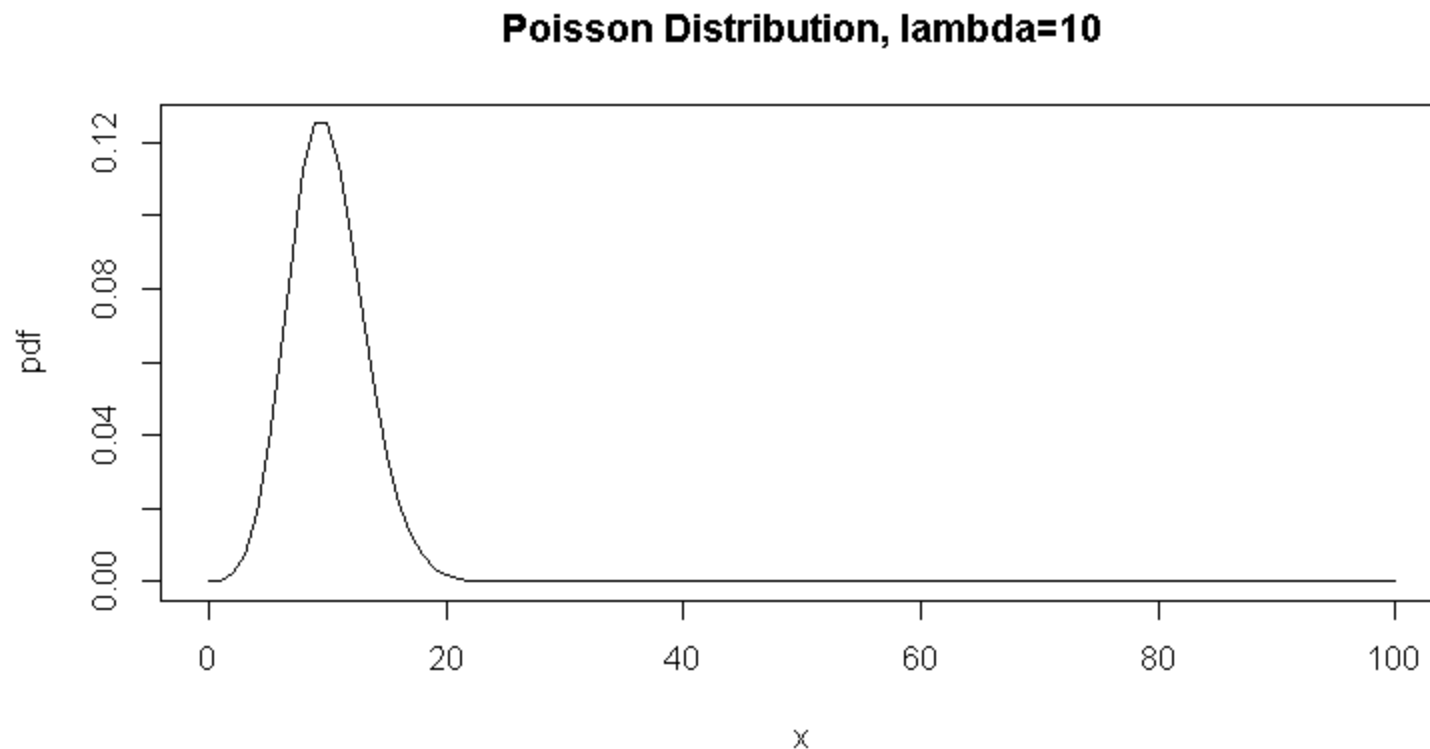
16. 선형모델의 일반화

- 포아송 회귀분석: *Poisson* regression analysis
 - 결과변수가 특정 기간 동안의 **사건발생횟수**(또는 개수)인 경우에 적용
 - 한 시간 동안 걸려오는 상담전화 횟수
 - 하루 동안 발생하는 범죄 횟수
 - 한 달 동안 발생하는 교통사고 횟수 등
 - 포아송 회귀모델: *Poisson* regression model
 - 링크함수는 $\ln(\lambda)$ 이며, 확률분포는 포아송 분포를 따름
 - $\ln(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
 - λ : 결과변수 y 의 평균



16. 선형모델의 일반화

```
> x <- 0:100  
> pdf <- dpois(x, lambda=10)  
> plot(x, pdf, type='l', main="Poisson Distribution, lambda=10")
```





16. 선형모델의 일반화

- robust 패키지의 breslow.dat 데이터셋: 뇌전증 환자의 투약 전/후 8주간의 발작횟수

```
> library(robust)
> data(breslow.dat)
> str(breslow.dat)
'data.frame':  59 obs. of  12 variables:
 $ ID      : int  104 106 107 114 116 118 123 126 130 135 ...
 $ Y1      : int   5  3  2  4  7  5  6 40  5 14 ...
 $ Y2      : int   3  5  4  4 18  2  4 20  6 13 ...
 $ Y3      : int   3  3  0  1  9  8  0 23  6  6 ...
 $ Y4      : int   3  3  5  4 21  7  2 12  5  0 ...
 $ Base    : int  11 11  6  8 66 27 12 52 23 10 ...
 $ Age     : int  31 30 25 36 22 29 31 42 37 28 ...
 $ Trt     : Factor w/ 2 levels "placebo","progabide": 1 1 1 1 1 1 1 1 1 1 ...
 $ Ysum    : int  14 14 11 13 55 22 12 95 22 33 ...
 $ sumY    : int  14 14 11 13 55 22 12 95 22 33 ...
 $ Age10   : num   3.1 3  2.5 3.6 2.2 2.9 3.1 4.2 3.7 2.8 ...
 $ Base4   : num   2.75 2.75 1.5 2 16.5 6.75 3 13 5.75 2.5 ...
```



16. 선형모델의 일반화

- 항노전증제를 투약 후 8주 동안 발생하는 발작횟수에 미치는 영향 분석
- 분석에 필요한 네 가지 변수만을 추출하여 요약통계량 확인

```
> seizure <- breslow.dat[c("Base", "Age", "Trt", "sumY")]
```

```
> summary(seizure)
```

Base	Age	Trt	sumY
Min. : 6.00	Min. :18.00	placebo :28	Min. : 0.00
1st Qu.: 12.00	1st Qu.:23.00	progabide:31	1st Qu.: 11.50
Median : 22.00	Median :28.00		Median : 16.00
Mean : 31.22	Mean :28.34		Mean : 33.05
3rd Qu.: 41.00	3rd Qu.:32.00		3rd Qu.: 36.00
Max. :151.00	Max. :42.00		Max. :302.00

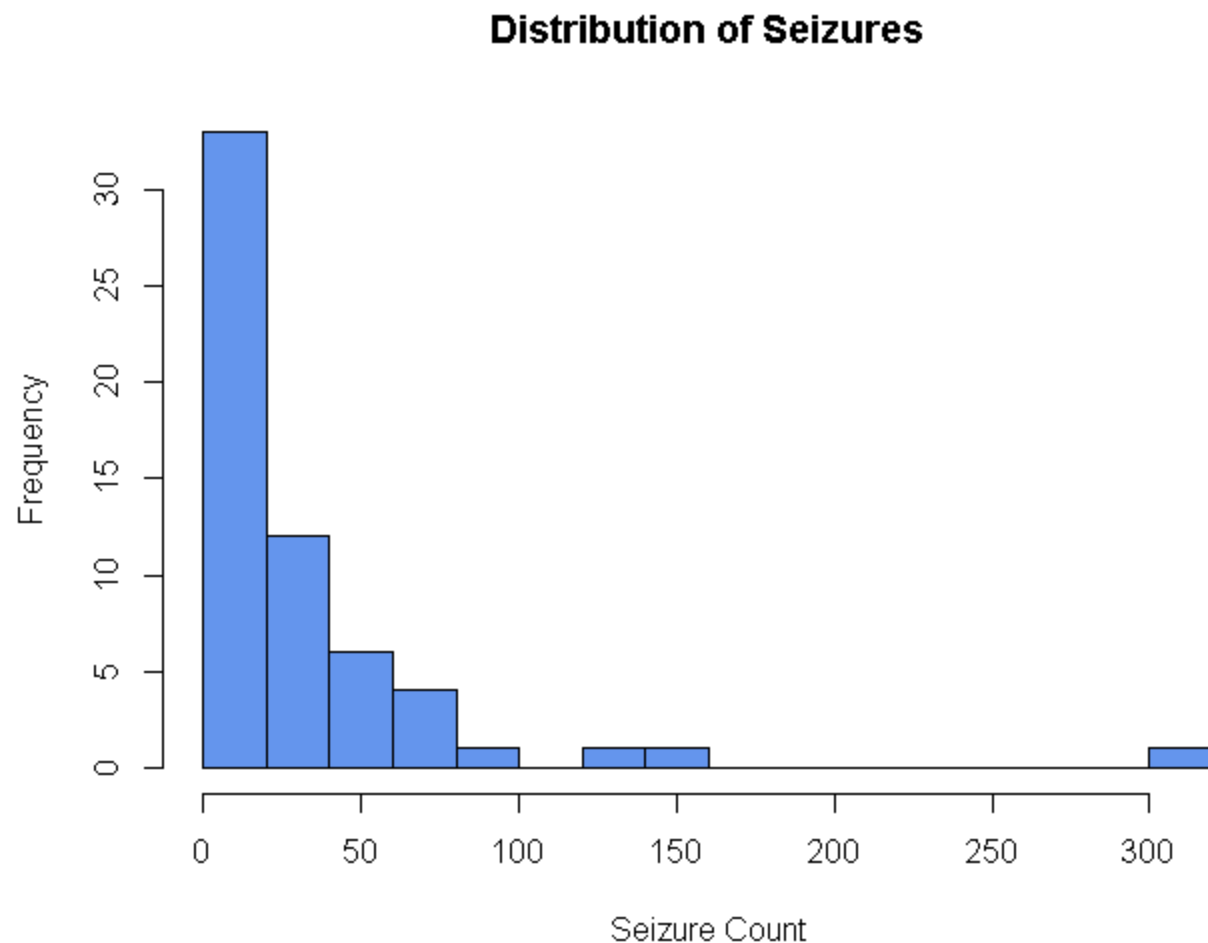
- 치료전 발작횟수와 치료 후 발작횟수 모두 중위수가 평균에 비해 작다



16. 선형모델의 일반화

- 발작횟수에 대한 히스토그램: 오른쪽으로 꼬리가 긴 편향된 분포(포아송 분포)

```
hist(seizure$sumY, breaks=20, col="cornflowerblue",  
      xlab="Seizure Count", main="Distribution of Seizures")
```





16. 선형모델의 일반화

- 포아송 회귀분석 수행: glm() 함수 이용

```
> seizure.poisson <- glm(sumY ~ Base + Age + Trt, data=seizure, family=poisson)
> summary(seizure.poisson)
```

Call:

```
glm(formula = sumY ~ Base + Age + Trt, family = poisson, data = seizure)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-6.0569	-2.0433	-0.9397	0.7929	11.0061

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.9488259	0.1356191	14.370	< 2e-16 ***
Base	0.0226517	0.0005093	44.476	< 2e-16 ***
Age	0.0227401	0.0040240	5.651	1.59e-08 ***
Trtprogabide	-0.1527009	0.0478051	-3.194	0.0014 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



16. 선형모델의 일반화

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2122.73 on 58 degrees of freedom

Residual deviance: 559.44 on 55 degrees of freedom

AIC: 850.71

Number of Fisher Scoring iterations: 5



16. 선형모델의 일반화

- 포아송 회귀분석의 결과 확인: 회귀계수 확인

```
> coef(seizure.poisson)
```

(Intercept)	Base	Age	Trtprogabide
1.94882593	0.02265174	0.02274013	-0.15270095

- 결과변수의 원래 척도로 예측변수의 회귀계수 해석:

```
> exp(coef(seizure.poisson))
```

(Intercept)	Base	Age	Trtprogabide
7.0204403	1.0229102	1.0230007	0.8583864

- 항노전증제를 처방받은 환자 집단은 위약을 복용한 환자 집단에 비해 발작횟수가 14.2% 감소



16. 선형모델의 일반화

- 이항 로지스틱 회귀분석: *binomial* logistic regression analysis
 - 결과변수가 이분형 범주일 때 특정 사건이 발생할 확률을 직접 추정
 - 결과변수의 예측값이 항상 1(사건발생)과 0(미발생) 사이의 확률값
 - 확률값이 0.5보다 크면 사건이 발생, 0.5보다 작으면 발생하지 않음
 - 예) 기업부도가 발생할 확률
 - 로지스틱 변환: *logistic transformation*
 - 예측변수의 선형결합을 로그 변환한 결과변수로 나타냄
 - 이항 로지스틱 회귀모델: binomial logistic regression model
 - $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
 - p : 이항 사건의 성공 확률(사건발생), $1 - p$: 이항 사건의 실패 확률(미발생)



16. 선형모델의 일반화

■ 이항 로지스틱 회귀분석:

- 오즈: *odds*

- $odds = \frac{p}{1-p}$: 사건 발생확률 대 사건 미발생 확률의 비율

- 로짓(*logit*): 오즈에 로그를 취한 값 = $\ln\left(\frac{p}{1-p}\right)$

- 로지스틱 회귀모델:

- 로그오즈(log odds=logit)에 대한 선형모델

- 링크함수가 로그오즈이며, 확률분포는 이항분포

- 사건발생확률 p 에 대해서 정리:

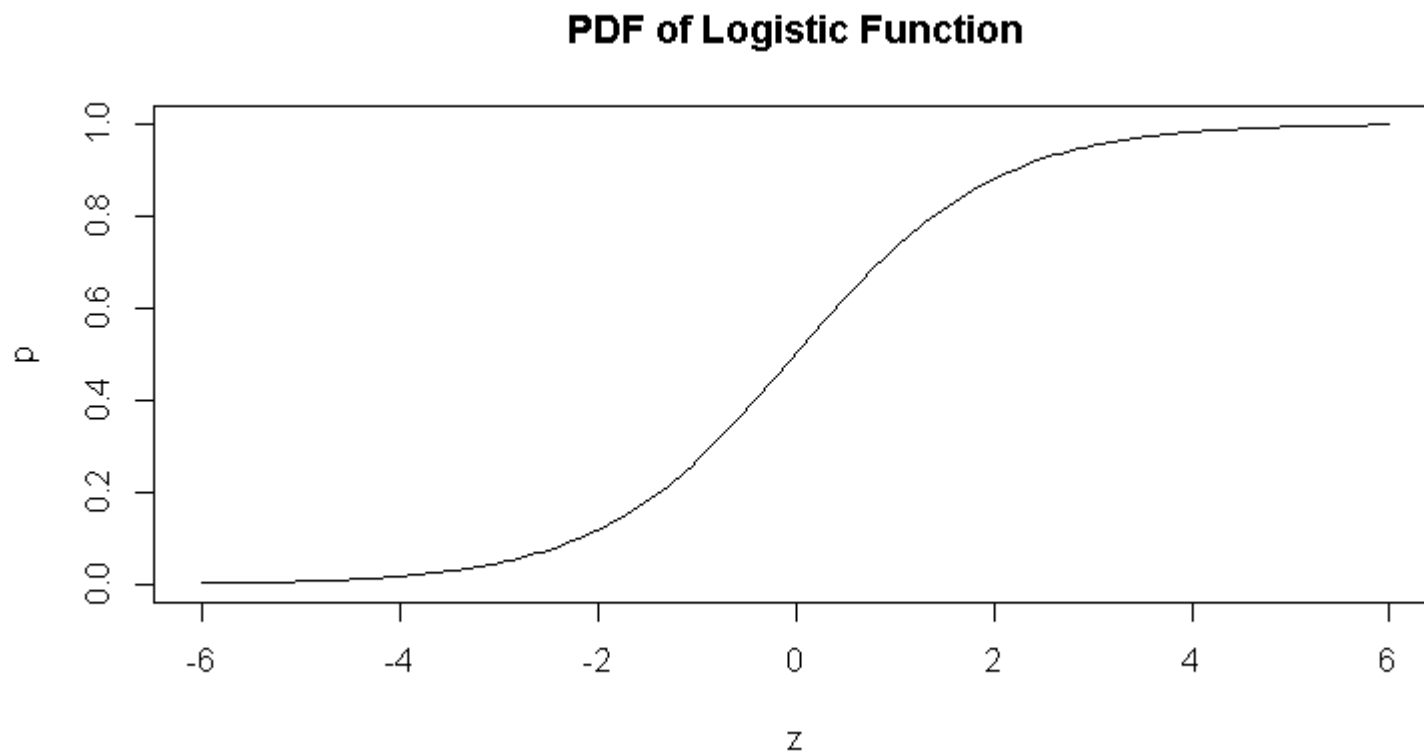
- $p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}, z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$

- 회귀계수를 알면 결과변수의 사건발생확률을 구할 수 있음



16. 선형모델의 일반화

```
> e <- exp(1)
> z <- seq(-6, 6, length=200)
> p <- 1 / (1 + e^(-z))
> plot(z, p, type="l", main="PDF of Logistic Function")
```





16. 선형모델의 일반화

- modeldata 패키지의 mlc_churn 데이터셋을 이용한 로지스틱 회귀분석
- mlc_churn 데이터셋: 이동통신회사의 고객이탈(customer churn) 데이터

```
> library(modeldata)
```

```
> data(mlc_churn)
```

```
> str(mlc_churn)
```

```
'data.frame': 5000 obs. of 20 variables:
```

```
$ state          : Factor w/ 51 levels "AK","AL","AR",...: 17 36 32 36 37  
$ account_length : int  128 107 137 84 75 118 121 147 117 141 ...  
$ area_code      : Factor w/ 3 levels "area_code_408",...: 2 2 2 1 2 3 3 2  
$ international_plan : Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 2 1 2 ...  
$ voice_mail_plan  : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 1 1 2 ...  
$ number_vmail_messages : int  25 26 0 0 0 0 24 0 0 37 ...  
$ total_day_minutes : num  265 162 243 299 167 ...
```

...(중략)

```
$ number_customer_service_calls: int  1 1 0 2 3 0 3 0 1 0 ...  
$ churn                        : Factor w/ 2 levels "yes","no": 2 2 2 2 2 2 2 2 2 2 ...
```




16. 선형모델의 일반화

- 고객의 거주지역(state), 지역코드(area_code)는 분석에서 제외
- churn 변수값: 고객미이탈을 1, 고객이탈을 2로 변환 (관심 사건이 고객이탈이므로)

```
> churn <- mtc_churn[, -c(1, 3)]
```

```
> levels(churn$churn)
```

```
[1] "yes" "no"
```

```
> churn$churn <- factor(ifelse(churn$churn=="no", 1, 2),  
                        levels=c(1, 2), labels=c("no", "yes"))
```

```
> str(churn)
```

```
> levels(churn$churn)
```

```
[1] "no" "yes"
```



16. 선형모델의 일반화

- 훈련용 데이터와 시험용 데이터를 나눈 후 이탈고객의 비율 확인

```
> churn.train <- churn[1:3333,]  
> churn.test <- churn[3334:5000,]  
> table(churn.train$churn)  
no    yes  
2850  483  
> prop.table(table(churn.train$churn))  
no      yes  
0.8550855 0.1449145  
> table(churn.test$churn)  
no    yes  
1443  224  
> prop.table(table(churn.test$churn))  
no      yes  
0.8656269 0.1343731
```



16. 선형모델의 일반화

- glm() 함수를 이용하여 로지스틱 회귀분석 수행

```
> churn.logit <- glm(churn ~ ., data=churn.train, family=binomial(link="logit"))
> summary(churn.logit)
```

Call:

```
glm(formula = churn ~ ., family = binomial(link = "logit"), data = churn.train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1532	-0.5132	-0.3402	-0.1953	3.2528

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-8.6515638	0.7243142	-11.944	< 2e-16	***
account_length	0.0008458	0.0013912	0.608	0.543199	
international_planyes	2.0427543	0.1454974	14.040	< 2e-16	***
voice_mail_planyes	-2.0250146	0.5740840	-3.527	0.000420	***
number_vmail_messages	0.0358803	0.0180108	1.992	0.046355	*



16. 선형모델의 일반화

```
total_eve_calls      0.0010579  0.0027826   0.380 0.703817
total_eve_charge     -9.5463678 19.2437266  -0.496 0.619840
total_night_minutes  -0.1238287  0.8764906  -0.141 0.887650
total_night_calls     0.0006993  0.0028419   0.246 0.805628
total_night_charge    2.8338084 19.4769043   0.145 0.884319
total_intl_minutes   -4.3377914  5.3009719  -0.818 0.413185
total_intl_calls     -0.0929680  0.0250603  -3.710 0.000207 ***
total_intl_charge    16.3900316 19.6323938   0.835 0.403804
number_customer_service_calls 0.5135638  0.0392678 13.079 < 2e-16 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2158.7  on 3315  degrees of freedom
AIC: 2194.7
```

Number of Fisher Scoring iterations: 6



16. 선형모델의 일반화

■ 오즈비: *odds ratio*

- 다른 독립변수가 동일하다는 가정하에서
 - 특정 독립변수 한 단위 증가에 따른
 - 사건 발생확률 대 미발생확률 비율의 **변화율**
- 오즈비는 오즈의 정의로부터 도출 가능
 - $\ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
 - $odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m}$
 - x_1 **변수의 오즈비** = $\frac{e^{\beta_0 + \beta_1 \times 1 + \beta_2 x_2 + \cdots + \beta_m x_m}}{e^{\beta_0 + \beta_1 \times 0 + \beta_2 x_2 + \cdots + \beta_m x_m}} = \frac{e^{\beta_1 \times 1}}{e^{\beta_1 \times 0}} = e^{\beta_1}$



16. 선형모델의 일반화

- 오즈비로 고객의 이탈확률 확인: total_day_charge값이 1 증가하면 이탈확률이 약 4.5배 증가

```
> coef(churn.logit)
```

```
> exp(coef(churn.logit))
```

(Intercept)	account_length
1.748532e-04	1.000846e+00
international_planyes	voice_mail_planyes
7.711821e+00	1.319919e-01
number_vmail_messages	total_day_minutes
1.036532e+00	7.833315e-01
total_day_calls	total_day_charge
1.003201e+00	4.539006e+00
total_eve_minutes	total_eve_calls
2.267538e+00	1.001058e+00
total_eve_charge	total_night_minutes
7.146035e-05	8.835312e-01
total_night_calls	total_night_charge
1.000700e+00	1.701012e+01

...(이하 생략)



16. 선형모델의 일반화

■ 로지스틱 회귀분석과 예측:

- 로지스틱 회귀계수를 알면 사건발생의 확률을 계산 가능
 - $P(\text{사건발생}) = \frac{1}{1+e^{-z}}$, $z = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_mx_m$
- 예측: *prediction*
 - 새로운 예측변수의 값을 입력하면 결과변수값이 1인 확률을 예측가능
 - 훈련 데이터로부터 회귀계수를 학습한 결과로
 - 시험 데이터의 고객이 이탈할 확률을 예측할 수 있음



16. 선형모델의 일반화

- 로지스틱 회귀를 이용한 고객이탈 확률 예측

```
> library(modeldata)
> data(mlc_churn)
> churn <- mlc_churn[, -c(1, 3)]
> churn$churn <- factor(ifelse(churn$churn=="no", 1, 2),
                        levels=c(1, 2), labels=c("no", "yes"))
> churn.train <- churn[1:3333,]
> churn.test <- churn[3334:5000,]
> churn.logit <- glm(churn ~ ., data=churn.train, family=binomial(link="logit"))
```




16. 선형모델의 일반화

- 범주형 변수 두 개를 더미변수로 변환하여 고객 이탈 확률 계산

```
> churn.test$international_plan <- ifelse(churn.test$international_plan=="no", 0, 1)
> churn.test$voice_mail_plan <- ifelse(churn.test$voice_mail_plan=="no", 0, 1)
> z <- coef(churn.logit)[1] +
  (as.matrix(churn.test[-18]) %*% coef(churn.logit)[-1])
> p <- 1/(1+exp(-z))
> head(p)
```

	[,1]
[1,]	0.07236813
[2,]	0.05774332
[3,]	0.22650409
[4,]	0.15289153
[5,]	0.07078500
[6,]	0.05880824



16. 선형모델의 일반화

- predict() 함수 이용: 학습한 모델을 바탕으로 예측을 수행

```
> churn.test <- churn[3334:5000,]  
> churn.logit.pred <- predict(churn.logit, newdata=churn.test, type="response")  
> head(churn.logit.pred)
```

1	2	3	4	5	6
0.07236813	0.05774332	0.22650409	0.15289153	0.07078500	0.05880824



16. 선형모델의 일반화

- 분류(*classification*): 계산된 예측확률로 이탈고객과 미이탈고객으로 분류

```
> churn.logit.pred <- factor(churn.logit.pred > 0.5,  
                             levels=c(FALSE, TRUE), labels=c("no", "yes"))
```

```
> head(churn.logit.pred)
```

```
1  2  3  4  5  6  
no no no no no no  
Levels: no yes
```

```
> table(churn.logit.pred)
```

```
churn.logit.pred  
   no   yes  
1595    72
```



16. 선형모델의 일반화

- 예측 결과의 비교: 혼동 행렬(confusion matrix)

```
> table(churn.test$churn, churn.logit.pred, dnn=c("Actual", "Predicted"))
```

	Predicted	
Actual	no	yes
no	1414	29
yes	181	43

```
> mean(churn.test$churn==churn.logit.pred)
```

```
[1] 0.8740252
```

- 로지스틱 회귀 모델을 사용한 이진 분류기는 87.4%의 정확도를 나타냄



16. 선형모델의 일반화


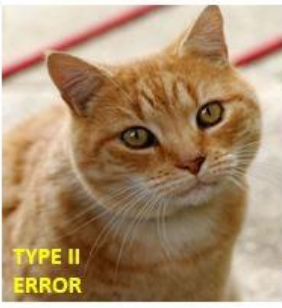


■ 혼동 행렬: *Confusion Matrix*





- 이진 분류기의 분류 결과를 2×2 행렬로 표시한 행렬
- 이진 분류기가 분류(예측)할 때, 얼마나 많이 헛갈렸는가를 나타냄

		예측값	
		Positive	Negative
실제값	Positive	TP	FN
	Negative	FP	TN



16. 선형모델의 일반화

		PREDICTED VALUES	
		Positive (CAT)	Negative (DOG)
ACTUAL VALUES	Positive (CAT)	 TRUE POSITIVE 6 YOU ARE A CAT	 FALSE NEGATIVE 1 YOU ARE A DOG TYPE II ERROR
	Negative (DOG)	 FALSE POSITIVE 2 YOU ARE A CAT TYPE I ERROR	 TRUE NEGATIVE 11 YOU ARE NOT A CAT

		Actual Values	
		1	0
Predicted Values	1	 TRUE POSITIVE You're pregnant	 FALSE POSITIVE You're pregnant TYPE 1 ERROR
	0	 FALSE NEGATIVE You're not pregnant TYPE 2 ERROR	 TRUE NEGATIVE You're not pregnant



16. 선형모델의 일반화

■ 분류 모델의 성능 평가 지표: *Evaluation Metric*

- 정확도: *CA*, Classification *Accuracy*

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

- 정밀도: *Precision*

- $Precision = \frac{TP}{TP+FP}$, 분류기가 양성으로 판정한 것이 얼마나 정확한가?

- 재현율: *Recall*

- $Recall = \frac{TP}{TP+FN}$, 분류기가 양성으로 판정한 것의 비율은 얼마인가?

- *F1*-Score

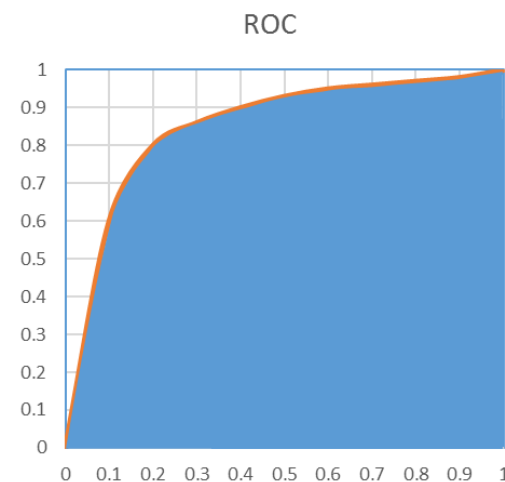
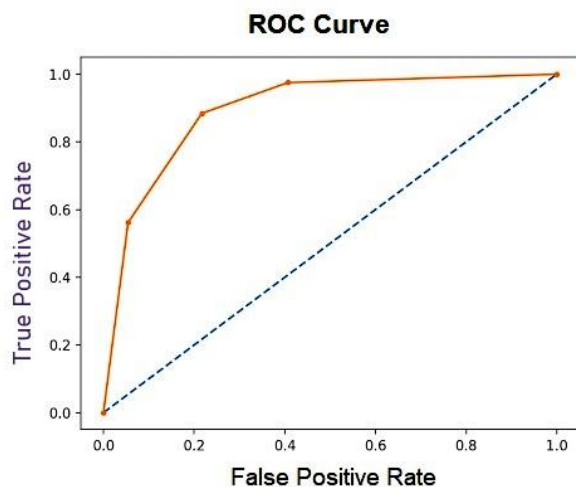
- $F1 = \frac{2 \times precision \times recall}{precision + recall}$, 정밀도와 재현율의 조화평균



16. 선형모델의 일반화

■ 분류 모델의 성능 평가 지표:

- ROC 곡선: Receiver Operation Characteristic Curve
 - 이진 분류의 결과에서 FP 비율과 TP 비율의 관계를 그린 곡선
- **AUC**: Area Under Curve
 - ROC 곡선의 하부 면적으로 표현하는 성능 평가 지표





16. 선형모델의 일반화

- 다항 로지스틱 회귀분석: *multinomial* logistic regression analysis
 - 예측변수로부터 세 개 이상의 사건(범주)을 갖는 결과변수의 사건발생확률 예측
 - 다항 로지스틱 회귀모델: g 개의 범주가 있을 때
 - $\ln\left(\frac{p_k}{p_1}\right) = \beta_{0k} + \beta_{1k}x_1 + \beta_{2k}x_2 + \cdots + \beta_{mk}x_m, k = 2, 3, \cdots, g$
 - p_1 : 기준범주인 사건 1의 발생확률, p_k : 사건 k 의 발생확률
 - 각 범주별 발생확률로 표현:
 - $p_k = \frac{e^{z_k}}{1 + \sum_{h=2}^g e^{z_h}}, k = 2, 3, \cdots, g$
 - 기준범주($k=1$)에 대한 발생확률
 - $p_1 = \frac{1}{1 + \sum_{h=2}^g e^{z_h}}, k = 2, 3, \cdots, g, \ln\left(\frac{p_k}{p_1}\right) = \ln\left(\frac{p_1}{p_1}\right) = 0, e^{z_1} = 1$



16. 선형모델의 일반화

- EffectsStars 패키지의 PID 데이터셋: 미국 유권자 944명의 정치성향 데이터

```
> library(EffectStars)
```

```
> data(PID)
```

```
> str(PID)
```

```
'data.frame': 944 obs. of 6 variables:
```

```
$ TVnews      : int  7 1 7 4 7 3 7 1 7 0 ...
```

```
$ PID         : Factor w/ 3 levels "Democrat","Independent",...: 3 1 1 1 1 1 1 2 2
```

```
1 ...
```

```
$ Income      : num  1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 ...
```

```
$ Education   : Factor w/ 2 levels "low","high": 1 2 2 2 2 2 2 2 2 1 ...
```

```
$ Age         : int  36 20 24 28 68 21 77 21 31 39 ...
```

```
$ Population: int  0 190 31 83 640 110 100 31 180 2800 ...
```

```
> head(PID)
```

	TVnews	PID	Income	Education	Age	Population
1	7	Republican	1.5	low	36	0
2	1	Democrat	1.5	high	20	190
3	7	Democrat	1.5	high	24	31
4	4	Democrat	1.5	high	28	83
5	7	Democrat	1.5	high	68	640
6	3	Democrat	1.5	high	21	110



16. 선형모델의 일반화

- 다항 로지스틱 회귀분석: VGAM 패키지의 `vglm()` 함수 이용

```
> library(VGAM)
> pid.mlogit <- vglm(PID ~ ., family=multinomial(), data=PID)
> summary(pid.mlogit)
```

Call:

```
vglm(formula = PID ~ ., family = multinomial(), data = PID)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept):1	1.2296119	0.3031106	4.057	4.98e-05	***
(Intercept):2	0.1275830	0.3405777	0.375	0.70795	
TVnews:1	0.0440935	0.0321897	1.370	0.17075	
TVnews:2	0.0247123	0.0350497	0.705	0.48077	
Income:1	-0.0165464	0.0027760	-5.960	2.51e-09	***
Income:2	-0.0002418	0.0027864	-0.087	0.93085	
Educationhigh:1	-0.2886055	0.1759813	-1.640	0.10101	
Educationhigh:2	-0.3530642	0.1971199	-1.791	0.07328	.
Age:1	-0.0077751	0.0052743	-1.474	0.14044	
...(이하 생략)					



16. 선형모델의 일반화

- 오즈비를 통해 예측변수 한 단위의 증가에 따른 오즈의 변화 비율 확인

```
> exp(coef(pid.mlogit))
```

(Intercept):1	(Intercept):2	TVnews:1	TVnews:2
3.4199020	1.1360792	1.0450800	1.0250202
Income:1	Income:2	Educationhigh:1	Educationhigh:2
0.9835898	0.9997582	0.7493078	0.7025321
Age:1	Age:2	Population:1	Population:2
0.9922550	0.9933500	1.0002592	1.0002052

- 뉴스 시청일수가 하루 늘어나면 Democrat일 가능성이 Republican일 가능성에 비해 4.5% 증가
- 소득이 한 단위가 증가하면 Democrat일 가능성이 Republican일 가능성에 비해 1.6% 감소
- 단, 뉴스 시청일수의 영향은 통계적으로 유의하지 않은 반면, 소득의 영향은 통계적으로 유의함



16. 선형모델의 일반화

- fitted() 함수: 다항 로지스틱회귀모델에 대한 각 범주별 예측확률 계산

```
> pid.mlogit.pred <- fitted(pid.mlogit)
```

```
> head(pid.mlogit.pred)
```

	Democrat	Independent	Republican
1	0.6247928	0.1932306	0.1819766
2	0.5739020	0.1817883	0.2443097
3	0.6109039	0.1745194	0.2145766
4	0.5843473	0.1772105	0.2384421
5	0.5839453	0.1694467	0.2466080
6	0.5856824	0.1794368	0.2348808



16. 선형모델의 일반화

- 다른 예측변수는 평균으로 고정하고, 한 예측변수의 값을 변화시키는 가상의 데이터셋을 생성

```
> testdata <- data.frame(Education=c("low", "high"),  
                          TVnews=mean(PID$TVnews),  
                          Income=mean(PID$Income),  
                          Age=mean(PID$Age),  
                          Population=mean(PID$Population))
```

```
> testdata  
  Education  TVnews  Income    Age Population  
1      low  3.727754 46.57574 47.04343    306.3814  
2      high  3.727754 46.57574 47.04343    306.3814
```



16. 선형모델의 일반화

- 예측변수의 수준에 따라 사건발생의 확률이 어떻게 변화하는지 확인

```
> pid.mlogit.pred <- predict(pid.mlogit, newdata=testdata, type="response")
```

```
> cbind(testdata, pid.mlogit.pred)
```

	Education	TVnews	Income	Age	Population	Democrat	Independent	Republican
1	low	3.727754	46.57574	47.04343	306.3814	0.4169951	0.2852971	0.2977078
2	high	3.727754	46.57574	47.04343	306.3814	0.3854667	0.2472630	0.3672703



16. 선형모델의 일반화

- 교육수준 등의 다른 변수를 통제하고 소득수준을 달리하여 변화의 추이 확인

```
> testdata <- data.frame(Education=rep("low", 5),
  TVnews=mean(PID$TVnews),
  Income=seq(20, 100, 20),
  Age=mean(PID$Age),
  Population=mean(PID$Population))

> testdata
> pid.mlogit.pred <- predict(pid.mlogit, newdata=testdata, type="response")
> cbind(testdata, pid.mlogit.pred)
```

	Education	TVnews	Income	Age	Population	Democrat	Independent	Republican
1	low	3.727754	20	47.04343	306.3814	0.5253435	0.2330383	0.2416182
2	low	3.727754	40	47.04343	306.3814	0.4434690	0.2725630	0.2839680
3	low	3.727754	60	47.04343	306.3814	0.3645531	0.3104445	0.3250024
4	low	3.727754	80	47.04343	306.3814	0.2923033	0.3448868	0.3628100
5	low	3.727754	100	47.04343	306.3814	0.2292065	0.3747050	0.3960885

Any Questions?

