Logistic Regression

History of Logistic Regression

- Logistic function was invented in the 19th century to describe the growth of populations and the course of autocatalytic chemical reactions.
- Quetelet and Verhulst
- Population growth was described easiest by exponential growth but led to impossible values

History of Logistic Regression

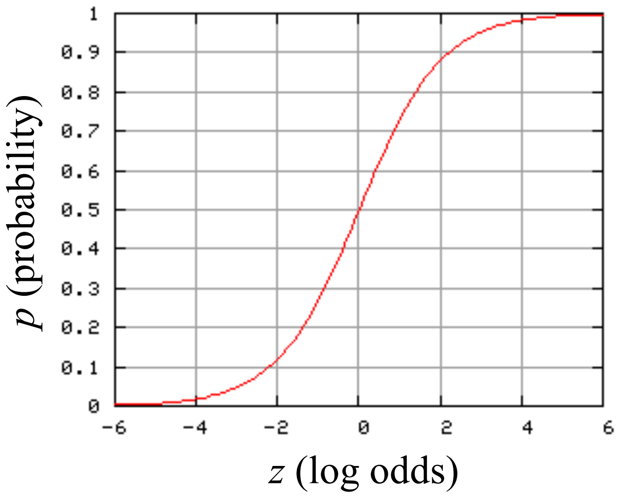
- Logistic function was the solution to a differential equation that was examined from trying to dampen exponential population growth models.

History of Logistic Regression

- Published in 3 different papers around the 1840’s. The first paper showed how the logistic models agreed very well with the actual course of the populations of France, Belgium, Essex, and Russia for periods up to the early 1830’s.

The Logistic Curve

$$LOGIT(p) = \ln\left(\frac{p}{(1-p)}\right) = z \quad \Leftrightarrow \quad p = \frac{\exp(z)}{1+\exp(z)}$$



Modeling discrete response variables

In a very large number of problems in cognitive science and related fields

the response variable is categorical, often *binary* (yes/no; acceptable/not acceptable; phenomenon takes place/does not take place)

potentially explanatory factors (independent variables) are categorical, numerical or both

Binary outcomes

- Linear regression is appropriate for continuous outcomes
- in biomedical research, our outcomes are more commonly of different forms
- Binary is probably the most prevalent 흔한, 유행하는
  - disease versus not disease
  - cured versus not cured
  - progressed versus not progressed
  - dead versus alive

Examples: multinomial responses

Discrete response variable without natural ordering:

Subject decides to buy one of 4 different products

We have brain scans of subjects seeing 5 different objects, and we want to predict seen object from features of the scan

We model the chances of developing 4 different (and mutually exclusive) psychological syndromes in terms of a number of behavioural indicators

Binomial and multinomial logistic regression models

Problems with binary (yes/no, success/failure, happens/does not happen) dependent variables are handled by (binomial) logistic regression

Problems with more than one discrete output are handled by

- ordinal logistic regression, if outputs have natural ordering
- multinomial logistic regression otherwise

The output of ordinal and especially multinomial logistic regression tends to be hard to interpret, whenever possible I try to reduce the problem to a binary choice

- E.g., if output is yes/maybe/no, treat “maybe” as “yes” and/or as “no”

Don’t be afraid of logistic regression!

Logistic regression seems less popular than linear regression

This might be due in part to historical reasons

the formal theory of generalized linear models is relatively recent: it was developed in the early nineteen-seventies

the iterative maximum likelihood methods used for fitting logistic regression models require more computational power than solving the least squares equations

Results of logistic regression are not as straightforward to understand and interpret as linear regression results

Finally, there might also be a bit of prejudice against discrete data as less “scientifically credible” than hard-science-like continuous measurements

Classic multiple regression

The by now familiar model:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_n \times x_n + \epsilon$$

Why will this not work if variable is binary (0/1)?

Why will it not work if we try to model proportions instead of responses (e.g., proportion of YES-responses in condition C)?

Properties of the residuals (with linear regression)

- Nonnormal error terms
  - Each error term can only take one of two values:

$$\begin{aligned} e_i &= 1 - \beta_0 - \beta_1 x_i \quad \text{if } y_i = 1 \\ e_i &= -\beta_0 - \beta_1 x_i \quad \text{if } y_i = 0 \end{aligned}$$

- Nonconstant error variance: the variance depends on X:

$$\begin{aligned} &Var(\hat{p}) = p(1-p) \\ &\sigma^2 = p(1-p) \\ &\sigma^2 = (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i) \end{aligned}$$

Modeling log odds ratios

Following up on the "proportion of YES-responses" idea, let's say that we want to model the probability of one of the two responses (which can be seen as the population

- proportion of the relevant response for a certain choice of the values of the dependent variables)

Probability will range from 0 to 1, but we can look at the logarithm of the odds ratio instead:

$$\text{logit(p)} = \log\frac{P}{1-P}$$

This is the logarithm of the ratio of probability of 1-response to probability of 0-response

- It is arbitrary what counts as a 1-response and what counts as a 0-response, although this might hinge on the ease of interpretation of the model (e.g., treating YES as the 1-response will probably lead to more intuitive results than treating NO as the 1-response)

Log odds ratios are not the most intuitive measure (at least for me), but they range continuously from −1 to +1

Relationship between Odds & Probability

$$\text{Odds}\left(\text{event}\right) = \frac{\text{Probability}\left(\text{event}\right)}{\text{1-Probability}\left(\text{event}\right)}$$

$$\text{Probability}\left(\text{event}\right) = \frac{\text{Odds}\left(\text{event}\right)}{1+\text{Odds}\left(\text{event}\right)}$$

The logistic regression model

Predicting log odds ratios:

$$logit(p) = \beta_0 + \beta_1 \times \mathrm{x}_1 + \beta_2 \times \mathrm{x}_2 + \ldots + \beta_n \times \mathrm{x}_\mathrm{n}$$

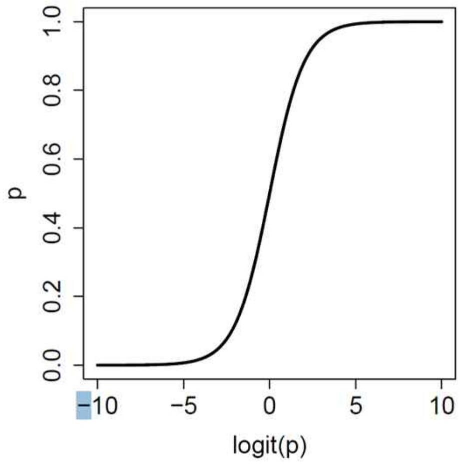
Back to probabilities:

$$p = \frac{e^{logit(p)}}{1 + e^{logit(p)}}$$

Thus:

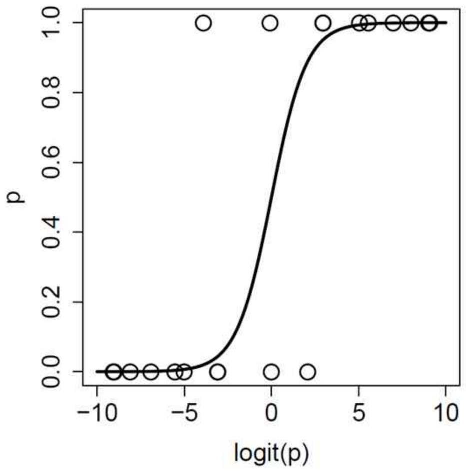
$$p = \frac{e^{\beta_0+\beta_1\times\mathrm{x}_1+\beta_2\times\mathrm{x}_2+\ldots+\beta_n\times\mathrm{x}_\mathrm{n}}}{1 + e^{\beta_0+\beta_1\times\mathrm{x}_1+\beta_2\times\mathrm{x}_2+\ldots+\beta_n\times\mathrm{x}_\mathrm{n}}}$$

From log odds ratios to probabilities



- $\boxed{p = \frac{e^{x\beta}}{1+e^{x\beta}}}$
- $\Phi^{-1}(p)$

Probabilities and responses



A subtle point: no error term

- NB:

$= x\beta$

$$logit(p) = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_n \times x_n$$

The outcome here is not the observation, but (a function of) *p*, the expected value of the *probability* of the observation given the current values of the dependent variables

This probability has the classic “coin tossing” Bernoulli distribution, and thus variance is not free parameter to be estimated from the data, but model-determined quantity given by $p(1 - p)$

Notice that errors, computed as observation $- p$, are not independently normally distributed: they must be near 0 or near 1 for high and low *p*s and near .5 for *p*s in the middle

The generalized linear model

glm

Logistic regression is an instance of a “generalized linear model”

Somewhat brutally, in a generalized linear model

- a weighted linear combination of the explanatory variables models a function of the expected value of the dependent variable (the “link” function)
- the actual data points are modeled in terms of a distribution function that has the expected value as a parameter

General framework that uses same fitting techniques to estimate models for different kinds of data

Linear regression as a generalized linear model

Linear prediction of a function of the mean:

$$g(E(y)) = X\beta$$ ⋯GLM model

“Link” function is identity:

⇒normality assumption 만족.

$$g(E(y)) = E(y)$$

Given mean, observations are normally distributed with variance estimated from the data

- This corresponds to the error term with mean 0 in the linear regression model

Logistic regression as a generalized linear model

Linear prediction of a function of the mean:

$$g(E(y)) = X\beta$$

“Link” function is :

$$g(E(y)) = \log\frac{E(y)}{1-E(y)}$$ ⇒ $\boxed{-\infty \sim \infty\text{의 값을 가질 수 있다.}}$

Given $E(y)$, i.e., p, observations have a Bernoulli distribution with variance $p(1-p)$

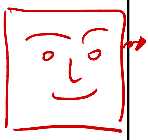
“Link” functions: P(Y=1)

- Logit link:

$$\text{logit}(P(Y=1)) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$$

- Probit link:

$$probit(P(Y=1)) = \Phi^{-1}(P(Y=1))$$



→ [illegible]

- Complementary log-log:

$$c\log\log(P(Y=1)) = \log[-\log(1-P(Y=1))]$$

$$(1-P(Y=1))^{-1}$$

Relationship between Odds & Probability

$$\text{Odds}(\text{event}) = \frac{\text{Probability}(\text{event})}{1\text{-Probability}(\text{event})}$$

$$\text{Probability}(\text{event}) = \frac{\text{Odds}(\text{event})}{1+\text{Odds}(\text{event})}$$

The Odds Ratio

Definition of Odds Ratio: Ratio of two odds estimates.

So, if Pr(response | trt) = 0.40 and Pr(response | placebo) = 0.20

Then:

$$\widehat{\text{Odds}}(\text{response}|\ \underset{\text{treatment}}{\underline{\text{trt group}}}) = \frac{0.40}{1-0.40} \doteq 0.667$$

$$\widehat{\text{Odds}}(\text{response}\ |\ \underset{\text{Control group}}{\underline{\text{placebo group}}}) = \frac{0.20}{1-0.20} = 0.25$$

$$\Rightarrow \quad \widehat{\text{OR}}(\text{Trt vs. Placebo}) \doteq \frac{0.667}{0.25} \doteq 2.67$$

Interpretation of the Odds Ratio

$\frac{90}{70}$ ··· 현역 $\frac{10}{30}$] 삼수3에

⇒ How 현역이 3배나 잘은 약은 쓰여(?) 국어

•Example cont'd:

Outcome = response, $\widehat{\text{OR}}_{(\text{trt vs. plb})} = \boxed{2.67}$

Then, the odds of a response in the treatment group were estimated to be 2.67 times the odds of having a response in the placebo group.

Alternatively, the odds of having a response were 167% higher in the treatment group than in the placebo group.

Assumptions in logistic regression

- Assumptions in logistic regression
  - Y_i are from Bernoulli or binomial (n_i, μ_i) distribution
  - Y_i are independent
  - Log odds $P(Y_i = 1)$ or logit $P(Y_i = 1)$ is a linear function of covariates

- Relationships among probability, odds and log odds

Measure	Min	Max	Name
Pr(Y=1)	0	1	prob
$\frac{\Pr(Y=1)}{1-\Pr(Y=1)}$	0	∞	odds
$\log\left(\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right)$	-∞	∞	log odds

Commonality between linear and logistic regression

- Operating on the logit scale allows a linear model that is similar to linear regression to be applied
- Both linear and logistic regression are apart of the family of Generalized Linear Models (GLM)

Logistic Regresion is a General Linear Model (GLM)

- Family of regression models that use the same general framework
- Outcome variable determines choice of model

Outcome	GLM Model
Continuous	Linear regression
(binary) = Dichotomous	Logistic regression
Counts	Poisson regression

Estimation of logistic regression models

Minimizing the sum of squared errors is not a good way to fit a logistic regression model

The least squares method is based on the assumption that errors are normally distributed and independent of the expected (fitted) values

As we just discussed, in logistic regression errors depend on the expected (*p*) values (large variance near .5, variance approaching 0 as *p* approaches 1 or 0), and for each *p* they can take only two values (1 – *p* if response was 1, *p* – 0 otherwise)

$y = \beta_0 + \beta_1 x + \varepsilon$ ~ 오차

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + e$ ~ 잔차

$\varepsilon = \hat{e}$

→ 오차는 잔차의 추정가능

$\sum \varepsilon = \sum e = \sum y - \hat{y}$

Logistic Regression Models are estimated by Maximum Likelihood

- Using this estimation gives model coefficient estimates that are asymptotically consistent, efficient, and normally distributed.
- Thus, a 95% Confidence Interval for β_K is given by:

$$\hat{\beta}_K \pm z_{\alpha/2}\left(SE_{\hat{\beta}_K}\right)$$

$$= (L\ ,\ U)$$

⇒ 95% 신뢰구간」

Logistic Regression Models are estimated by Maximum Likelihood

- The Odds Ratio for the kth model coefficient is:

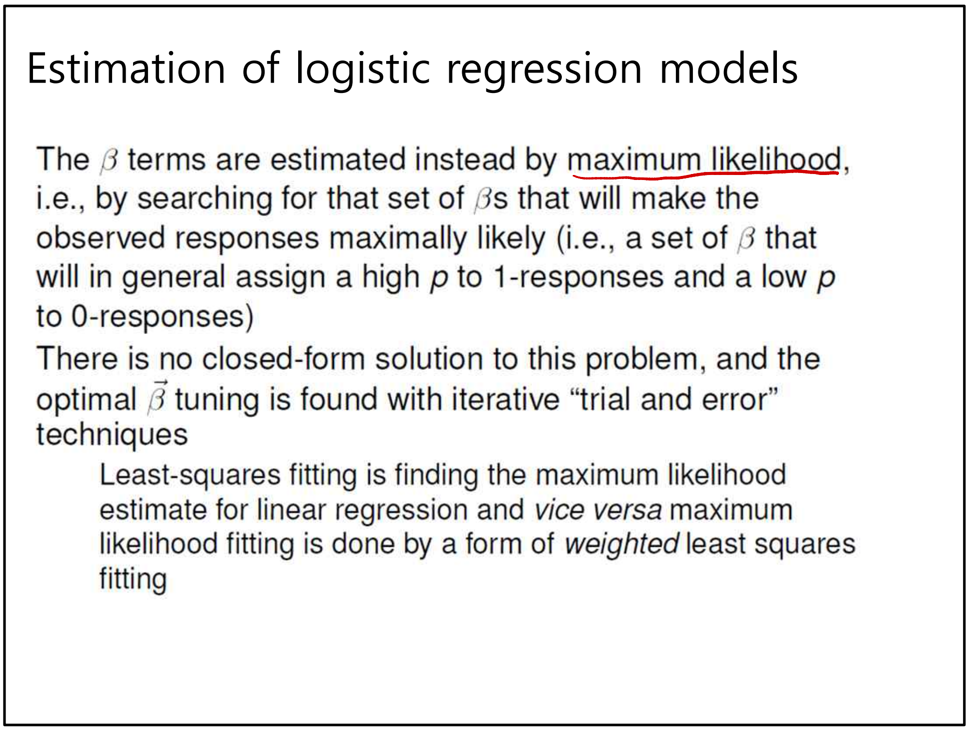
$$\hat{\text{OR}} = \exp\left(\hat{\beta}_{\text{K}}\right)$$

「Odds Ratio의 신뢰구간」

- We can also get a 95% CI for the OR from:

$$= \left(e^{\text{L}},\ e^{\text{U}}\right)$$

where $(\text{L}\ ,\ \text{U})$ is a 95% CI for β_K

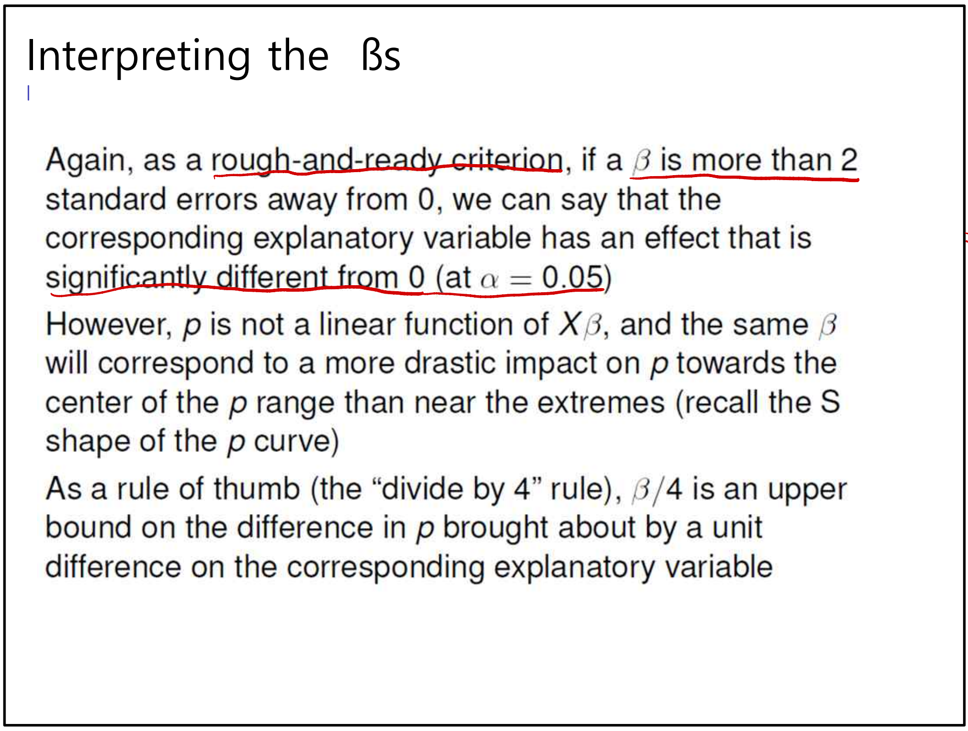


Estimation of logistic regression models

The β terms are estimated instead by maximum likelihood, i.e., by searching for that set of βs that will make the observed responses maximally likely (i.e., a set of β that will in general assign a high *p* to 1-responses and a low *p* to 0-responses)

There is no closed-form solution to this problem, and the optimal $\vec{\beta}$ tuning is found with iterative "trial and error" techniques

Least-squares fitting is finding the maximum likelihood estimate for linear regression and *vice versa* maximum likelihood fitting is done by a form of *weighted* least squares fitting



Interpreting the ßs

Again, as a rough-and-ready criterion, if a β is more than 2 standard errors away from 0, we can say that the corresponding explanatory variable has an effect that is significantly different from 0 (at $\alpha = 0.05$)

However, *p* is not a linear function of $X\beta$, and the same β will correspond to a more drastic impact on *p* towards the center of the *p* range than near the extremes (recall the S shape of the *p* curve)

As a rule of thumb (the "divide by 4" rule), $\beta/4$ is an upper bound on the difference in *p* brought about by a unit difference on the corresponding explanatory variable

P_value

=) 귀무가설이 맞다는 가정하에 현재에 있는 데이터보다 미래에 얻어질 데이터가 귀무가설을 뒷받침할 확률

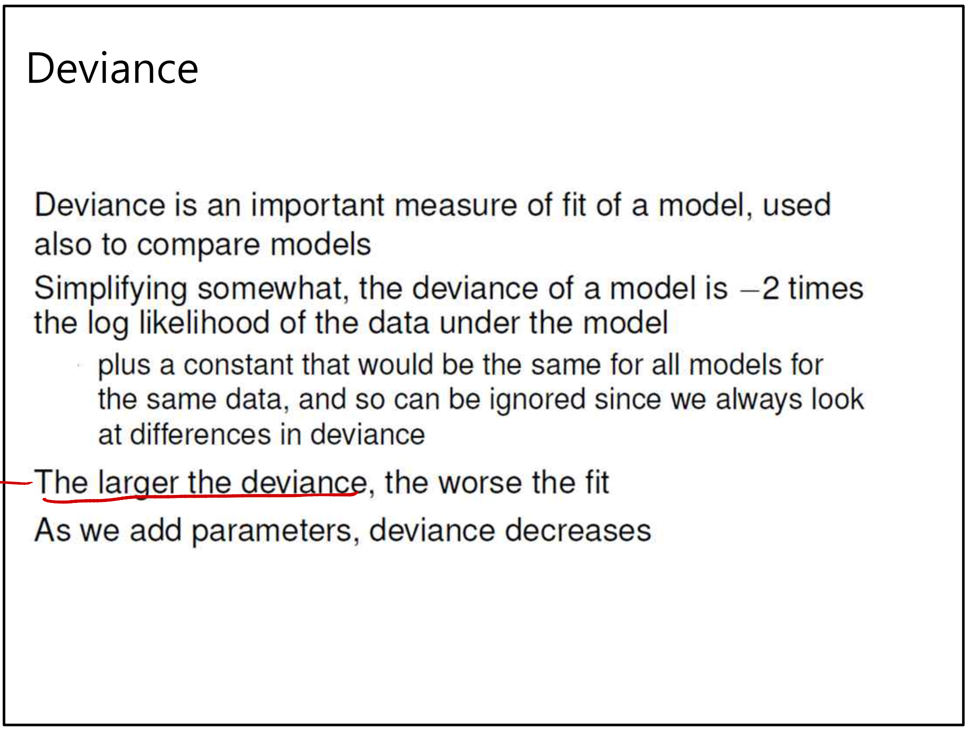
Goodness of fit

⇒ 관측값과 예측값이 얼마나 맞는지

- Again, measures such as R^2 based on residual errors are not very informative
- One intuitive measure of fit is the *error rate*, given by the proportion of data points in which the model assigns $p > .5$ to 0-responses or $p < .5$ to 1-responses
    - This can be compared to baseline in which the model always predicts 1 if majority of data-points are 1 or 0 if majority of data-points are 0 (baseline error rate given by proportion of minority responses over total)
- Some information lost (a .9 and a .6 prediction are treated equally)
- Other measures of fit proposed in the literature, no widely agreed upon standard

Binned goodness of fit

- Goodness of fit can be inspected visually by grouping the *p*s into equally wide bins (0-0.1,0.1-0.2, . . .) and plotting the average *p* predicted by the model for the points in each bin vs. the observed proportion of 1-responses for the data points in the bin
- We can also compute a R^2 or other goodness of fit measure on these binned data



Deviance

Deviance is an important measure of fit of a model, used also to compare models

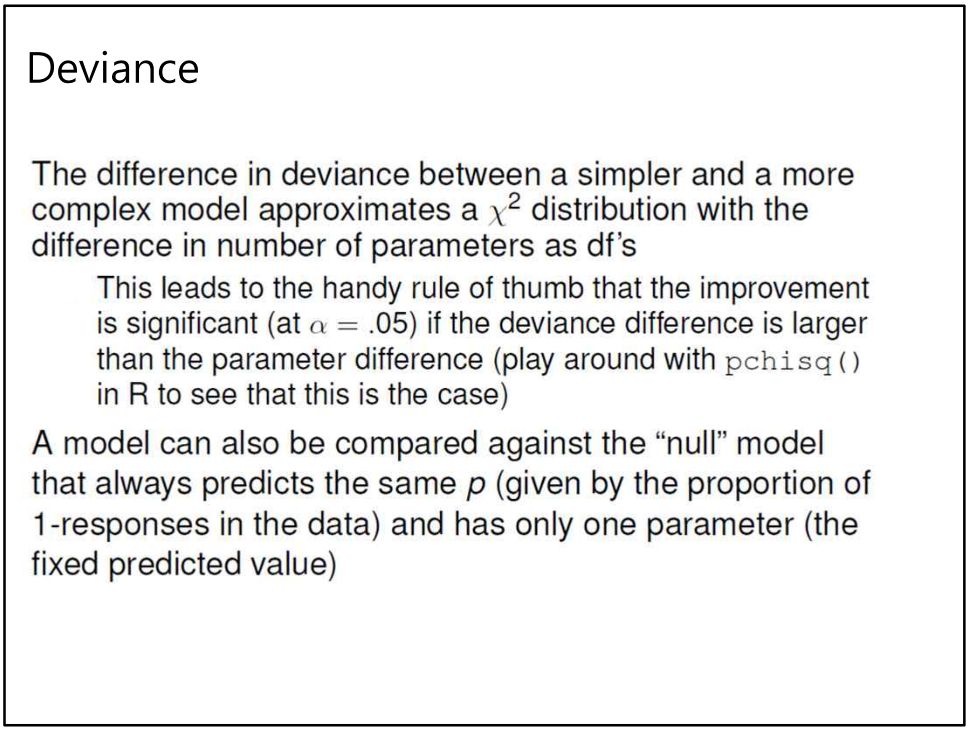
Simplifying somewhat, the deviance of a model is -2 times the log likelihood of the data under the model

- plus a constant that would be the same for all models for the same data, and so can be ignored since we always look at differences in deviance

The larger the deviance, the worse the fit

D
deviance는
분산개념

As we add parameters, deviance decreases



Deviance

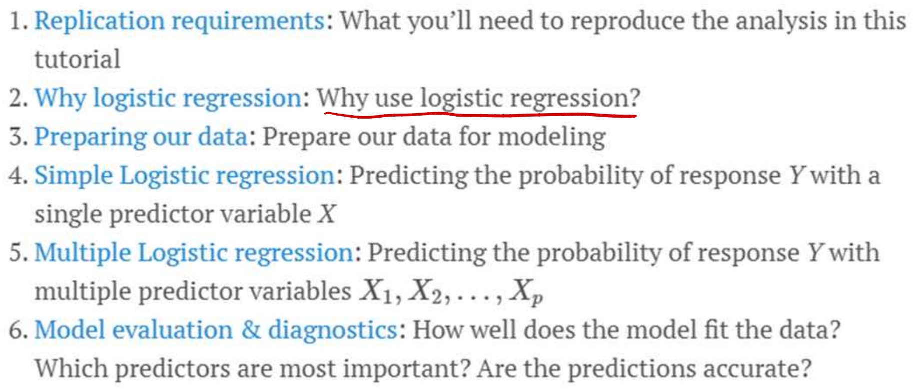
The difference in deviance between a simpler and a more complex model approximates a χ^2 distribution with the difference in number of parameters as df's

- This leads to the handy rule of thumb that the improvement is significant (at $\alpha = .05$) if the deviance difference is larger than the parameter difference (play around with `pchisq()` in R to see that this is the case)

A model can also be compared against the "null" model that always predicts the same *p* (given by the proportion of 1-responses in the data) and has only one parameter (the fixed predicted value)

$np>5$ && $nq>5$ 일때

$\Rightarrow \boxed{\text{proportion} \sim N}$



1. Replication requirements: What you'll need to reproduce the analysis in this tutorial
2. Why logistic regression: Why use logistic regression?
3. Preparing our data: Prepare our data for modeling
4. Simple Logistic regression: Predicting the probability of response Y with a single predictor variable X
5. Multiple Logistic regression: Predicting the probability of response Y with multiple predictor variables $X_1, X_2, \ldots, X_p$
6. Model evaluation & diagnostics: How well does the model fit the data? Which predictors are most important? Are the predictions accurate?

Replication Requirements

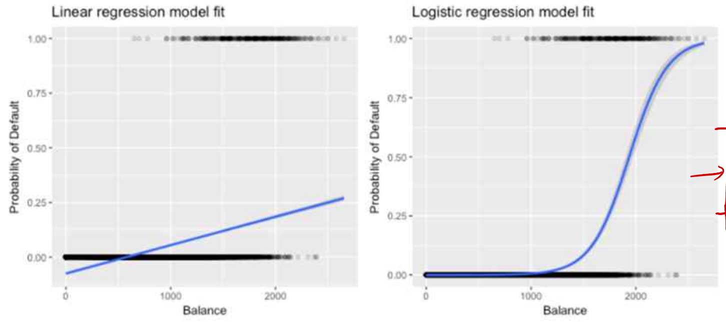
- # Packages
- library(tidyverse) # data manipulation and visualization
- library(modelr) # provides easy pipeline modeling functions
- library(broom) # helps to tidy up model outputs

- # Load data
- (default <- as_tibble(ISLR::Default))

Why Logistic Regression

$$Y = \begin{cases} 1, & \text{if epileptic seizure;} \\ 2, & \text{if stroke;} \\ 3, & \text{if drug overdose.} \end{cases}$$

Probit은 odds ratio
계산하기 어려움
$\Phi^{-1}(p) = x$이므로 ~ 역변환



$\rightarrow p = \frac{e^{x\beta}}{1+e^{x\beta}}$

Preparing Data

- split data into a training (60%) and testing (40%) data sets so we can assess how well our model performs on an out-of-sample data set.

⇒ cross validation

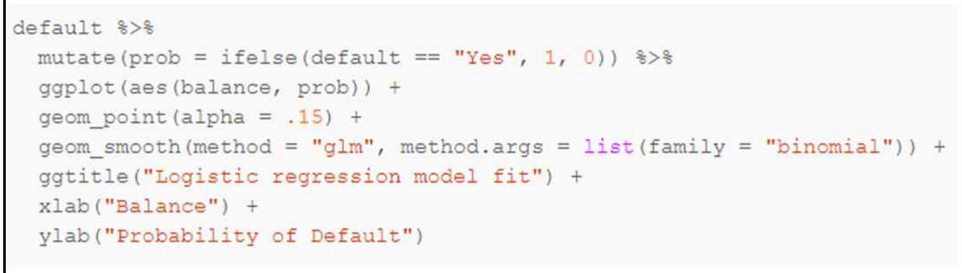
- set.seed(123)
- sample <- sample(c(TRUE, FALSE), nrow(default), replace = T, prob = c(0.6,0.4))
- train <- default[sample,]
- test <- default[!sample,

⇒ true false를 뽑고 집합을

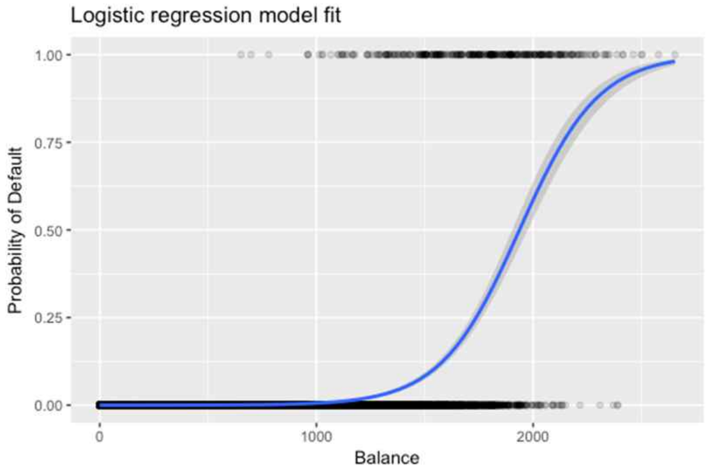
Simple Logistic Regression

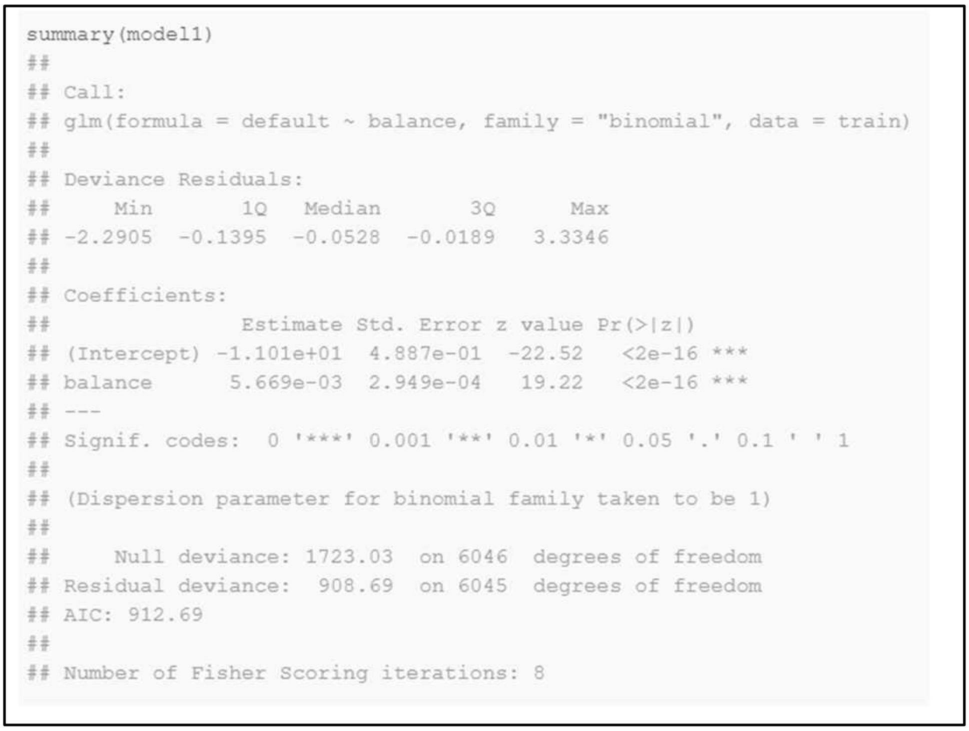
- model1 <- glm(default ~ balance, family = "binomial", data = train)

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y'_i=0} (1 - p(x'_i))$$

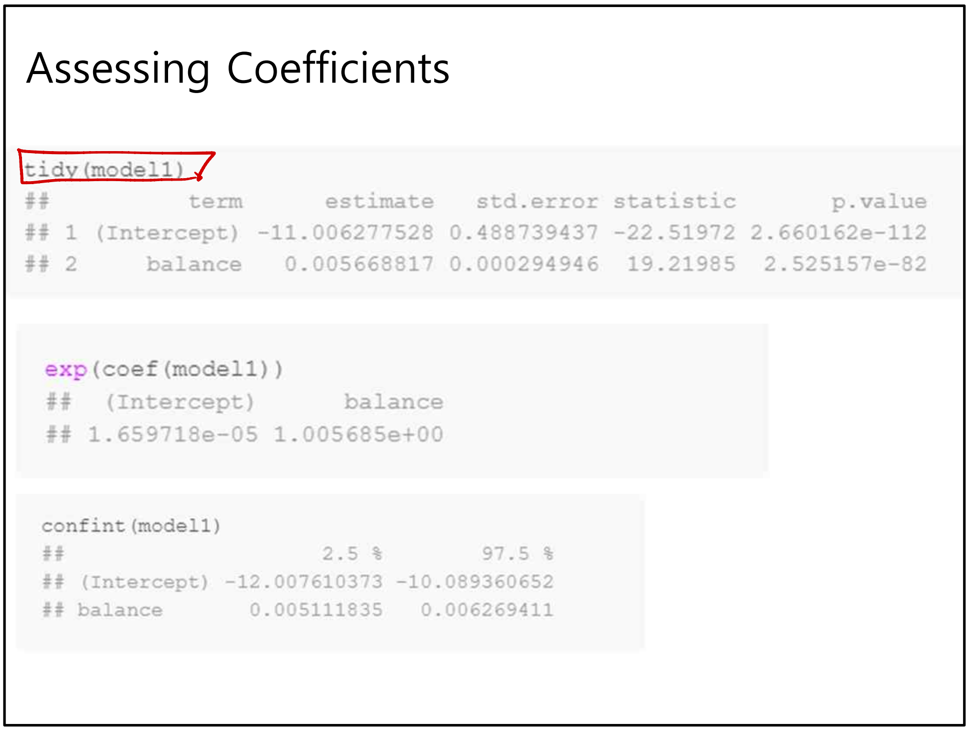


```
default %>%
  mutate(prob = ifelse(default == "Yes", 1, 0)) %>%
  ggplot(aes(balance, prob)) +
  geom_point(alpha = .15) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  ggtitle("Logistic regression model fit") +
  xlab("Balance") +
  ylab("Probability of Default")
```





```
summary(model1)
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2905  -0.1395  -0.0528  -0.0189   3.3346
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.101e+01  4.887e-01  -22.52   <2e-16 ***
## balance      5.669e-03  2.949e-04   19.22   <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: 1723.03  on 6046  degrees of freedom
## Residual deviance:  908.69  on 6045  degrees of freedom
## AIC: 912.69
##
## Number of Fisher Scoring iterations: 8
```



Assessing Coefficients

```
tidy(model1)
##          term      estimate   std.error statistic       p.value
## 1 (Intercept) -11.006277528 0.488739437 -22.51972 2.660162e-112
## 2     balance   0.005668817 0.000294946  19.21985  2.525157e-82
```

```
exp(coef(model1))
##  (Intercept)      balance
## 1.659718e-05 1.005685e+00
```

```
confint(model1)
##                     2.5 %       97.5 %
## (Intercept) -12.007610373 -10.089360652
## balance       0.005111835   0.006269411
```