Logistic Regression

Overview

Type of Predictors Type of Response	Categorical	Continuous	Categorical and Continuous
Continuous	ANOVA	OLS Regression	OLS Regression or ANCOVA
Categorical	Contingency Table	Logistic Regression	Logistic Regression

- **Y is binary** (0/1, pass/fail, having a disease/ not) with **continuous or categorical predictors**

- Goal: Predict the probability of having an event (e.g., having disease, pass exam etc.) based on given information or to see how predictor are related to an event

Examples of logistic regression:

- Amazon -> whether you will make a purchase
  - ✓ Outcome: Make a purchase vs. do not make a purchase
- Netflix -> whether you will like a movie
  - ✓ Outcome: Watch a movie vs. Not watch a movie
- Insurance Companies -> what is your risk
  - ✓ Outcome: Accident vs. No accident
- UTSA -> who will graduate from their program
  - ✓ Outcome: Graduated vs. did not graduate

Why not linear regression?

- Continuous response y and Predictors x_j;

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$$

- In logistic regression, our response variable is binary (0 or 1/ Having a disease or not/ Pass the exam or not, etc.)
- Normality assumption on Y does not make sense
  - Need different assumptions on Y
- What happens if we fit linear regression model?

Logistic Regression

- Bernoulli assumption on Y
  - $Y_i \sim Bernoulli\ (P_i)$
  - Examples of Bernoulli include **Flipping a coin;** observing a head ~ Bernoulli (0.5)
- We want to model the probability of having an event P, $(0 \leq P \leq 1)$
  - $$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots + \beta_p x_p$$

    where $\frac{p}{1-p}$ is called as **odds**
- *Why do we define response variables as log (odds)?*
- *What is difference on right-hand side compared to multiple linear regression?*

Logistic regression

- No longer Normality assumption on Y
- Can no longer separate the error term from the model
- Estimate $\beta's$ based under Bernoulli assumption on Y
- Interpretation is not straightforward
  - Understand **odds** and **odds ratio**
- Diagnostics still needed
- **glm**(y~ x, data = data, **family = "binomial"**) in R

Odds

- Odds = p/(1-p)
- Ratio of probability event happens to probability it doesn't happen
  - Odds of 1 means equally likely to happen or not
  - What if odds(car accident) > 1 ?
  - What if odds(car accident) < 1 ?

Odds Ratio

- Compares odds for event under different conditions
- For example,

$$\text{odds ratio}(\text{accident}; M, F) = \frac{odds(accident|M)}{odds(accident|F)}$$

  - Odds ratio of 1 means odds are the same
  - What if odds ratio (car accident; M,F) > 1
  - What if odds ratio (car accident; M,F) < 1

Expected Odds and Probabilities

Under logistic regression model:

- Expected odds:

$$\frac{\hat{p}}{1-\hat{p}} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots\right)$$

- Expected probability (fitted $\hat{p}$) :

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots)}$$

  - **Always fall between 0 and 1**

Interpretation: (i) predictor is categorical

- Consider the simple linear regression with one categorical variable x

$$\log\left(\frac{p}{1-p}\,\middle|\,x\right) = \beta_0 + \beta_1 x$$

- Suppose that x=1 (female) or 0 (male), then
  - $\log\left(\frac{p}{1-p}\,\middle|\,male\right) = \beta_0, \quad \log\left(\frac{p}{1-p}\,\middle|\,female\right) = \beta_0 + \beta_1$
- It implies
  - $\log\left(\frac{odds|\,female}{odds|\,male}\right) = \beta_1$, equiv. to $\frac{odds|\,female}{odds|\,male} = e^{\beta_1}$
- **The odds of having an event is e^{β_1} times for female group compared to that of male group.**
  - **What does positive and negative β_1 mean?**

Interpretation: (i) predictor is categorical

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_a: \beta_1 \neq 0$$

- **If p-value ≥.05**, the predictor is not statistically significant in the model (β_1=0), hence odds ratio $\frac{odds|\, female}{odds|\, male} = e^0 = 1$
- **If p-value <.05,** the predictor is statistically significant in the model ($\beta_1 \neq 0$),
  - **If β_1>0** then odds ratio> 1, group 1 has higher odds than group 0
  - **If β_1 <0** then odds ratio< 1, group 1 has lower odds than group 0

Interpretation: (ii) predictor is continuous

- Suppose that x is continuous (e.g., age),

$$\log\left(\frac{p}{1-p}\,\middle|\,x\right) = \beta_0 + \beta_1 x, \qquad \log\left(\frac{p}{1-p}\,\middle|\,x+1\right) = \beta_0 + \beta_1(x+1)$$

- It implies

$$-\log\left(\frac{odds|\,x+1}{odds|\,x}\right) = \beta_1 \text{ , equiv. to } \frac{odds|\,x+1}{odds|\,x} = \mathrm{e}^{\beta_1}$$

- **The odds of having an event change by a factor of e^{β_1} with one unit increase in x.**

  (**multiplicative change** in the odds for a one unit change in the predictor variable)

- Additive change <-> Multiplicative change

Interpretation: (ii) predictor is continuous

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_a: \beta_1 \neq 0$$

- **If p-value ≥.05**, the predictor is not statistically significant in the model (β_1=0), hence odds ratio $\frac{odds|\, x+1}{odds|\, x} = e^0 = 1$
- **If p-value <.05,** the predictor is statistically significant in the model ($\beta_1 \neq 0$),
  - **If β_1>0** then the odds of having an event increases by a multiple of e^{β_1} with one unit increase in predictor x.
  - **If β_1 <0** then the odds of having an event decreases by a multiple of e^{β_1} with one unit increase in predictor x.

Interpretation: (iii) multiple regression

- The interpretation is the same as the previous single continuous or categorical predictor under the condition, “when all other predictors are same”.

Goodness-of-fit

- Goodness-of-fit of a statistical model describes how well it fits a set of observations

1. Hosmer and Lemeshow goodness-of-fit test
   - H0: Model is adequate. H1: Model is not adequate
   - hoslem.test() in R – package "ResourceSelection"
2. Pseudo R-Squares (modified R-squares)
   - PseudoR2() in R

Exercise 1: ESR and Plasma Data

- Response: **esr** 0|1 indicator for healthy or unhealthy erythrocyte sedimentation rate(ESR)
- 0 is healthy and 1 is unhealthy
- Predictors: **fibrinogen** and **gamma** globulin plasma levels
- Want to model unhealthy ESR as a function of plasma levels
  - **make sure 1 is coded for the even of interest!**
- Hosmer and Lemeshow goodness-of-fit test

Exercise 1: ESR and Plasma Data

- Model unhealthy **esr** as a function of **gamma** and **fibrinogen** levels
- Which terms seem significant?
- What does the model tell us about odds ratios for changes in plasma levels?
- **Significance check** of individual predictor;
  - P-value from parameter estimates
- **Interpretation** of estimated coefficients
  - Odds Ration estimates
- Remove insignificant terms, refit and interpret
- Which person is more likely to be in unhealthy **esr** status?
- Diagnostics – residual plots and cook's distance

Exercise 1: ESR and Plasma Data

```
glm.plasma0 = glm(esr ~ fibrinogen+gamma, data = plasma,
family = "binomial")
summary(glm.plasma0)
```

$H_0: \beta_{fib} = 0 \ vs.$
$H_a: \beta_{fib} \neq 0$

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.7921     5.7963  -2.207   0.0273 *
## fibrinogen    1.9104     0.9710   1.967   0.0491 *
## gamma         0.1558     0.1195   1.303   0.1925
```

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -12.79 + 1.91 * fib + 0.156 * gamma$$

Exercise 1: Logistic ESR Models

```
glm.plasma = glm(esr ~ fibrinogen, data = plasma, family
= "binomial")
summary(glm.plasma)

## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.8451     2.7703  -2.471   0.0135 *
## fibrinogen    1.8271     0.9009   2.028   0.0425 *
## ---

OR2=exp(glm.plasma$coefficients)
round(OR2, 3)
## (Intercept)  fibrinogen
##       0.001       6.216
```

Calculate odds ratio for the interpretation

The odds of being unhealthy status change by a factor of exp(1.827)=6.216 with one unit increase in fibrinogen. *Thus, a person with higher fibrinogen level is more likely to be unhealthy esr*

Exercise 1: Logistic ESR Models

```
HosmerLemeshowTest(fitted(glm.plasma), plasma$esr)$C
```

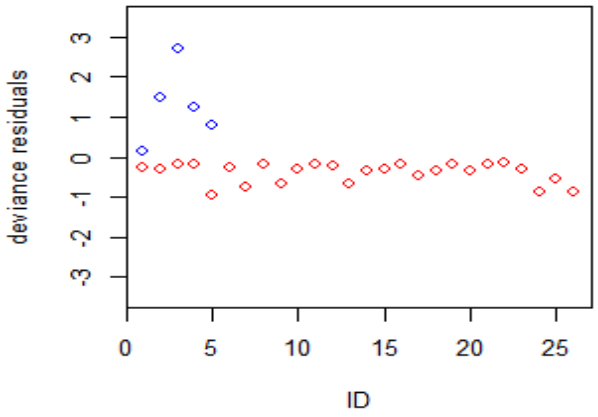
```
##
##  Hosmer-Lemeshow C statistic
##
## data:  fitted(glm.plasma) and plasma$esr
## X-squared = 8.9002, df = 8, p-value = 0.3508
```

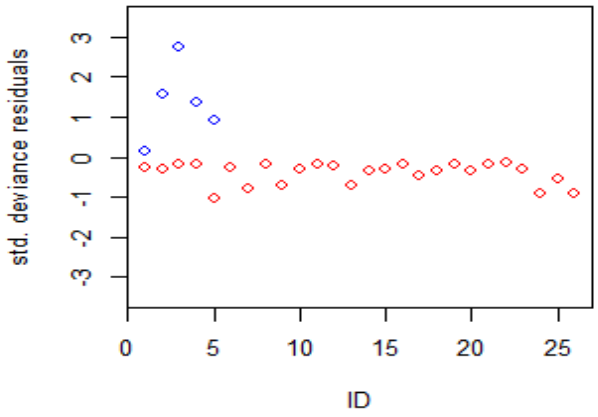
H0: model is adequate vs. Ha: model is inadequate

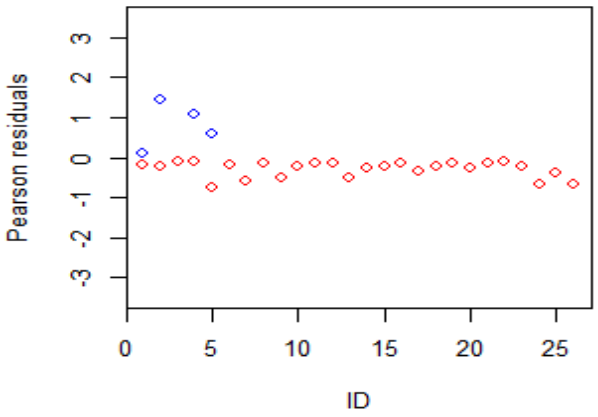
```
## psuedo R^2
pseudo.r2<-PseudoR2(glm.plasma, which =  c("McFadden
", "Nagel", "CoxSnell"))
round(pseudo.r2, 3)
```

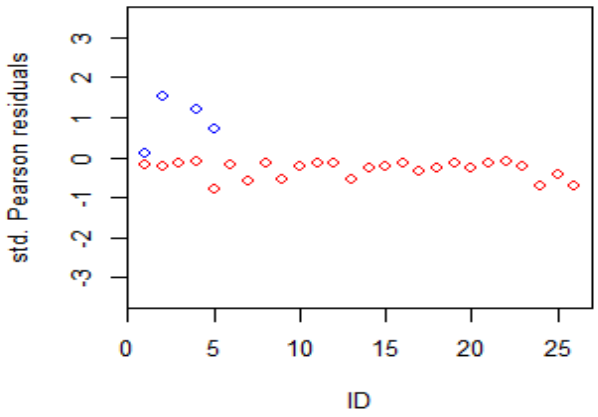
```
##   McFadden Nagelkerke   CoxSnell
##      0.196      0.278      0.172
```

Exercise 1: Logistic ESR Models



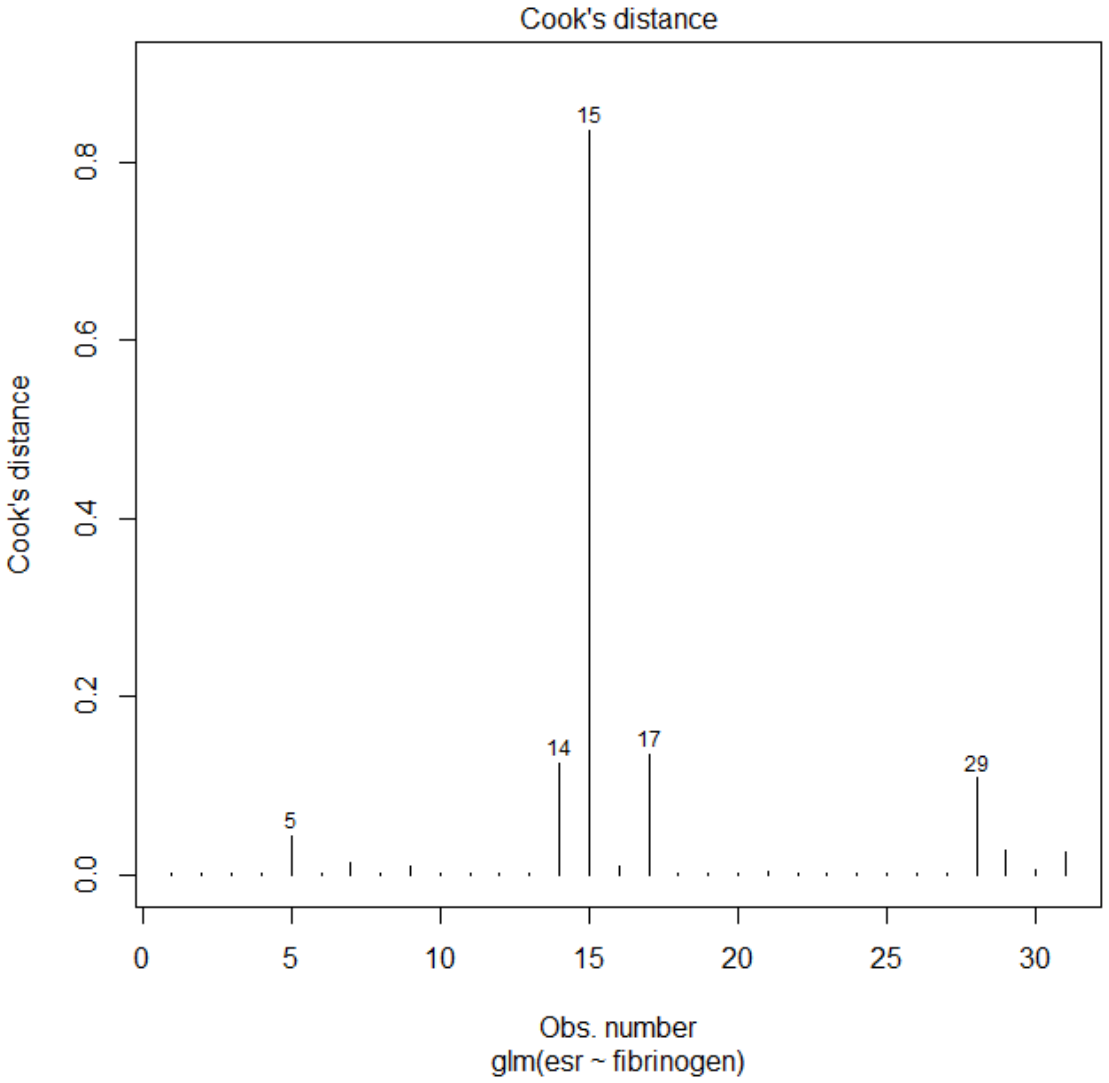






Check if any patterns found. If not, assumption is valid

Exercise 1: Logistic ESR Models



Cook's distance – observation with large cook's d is an influential point

Exercise 2: Amputation

- Response: **amputation** 0|1 indicator (0=not amputated, 1=amputated)
- Predictors:
  - **illness_severity** (three categories: low, moderate, and high)
  - **diabetes** (two categories: uncontrolled and controlled)
  - **ulcers**: (two categories: no ulcer as 0, at least on ulcers as 1
- Want to model if a patient gets an amputation or not
- Model selection with other predictors

Exercise 2: Amputation

- Which predictors are significant?
- Interpretation of Odds Ratio for categorical variable with more than two levels
- What kind of patient has the highest probability to gets an amputation?

Exercise 2: Amputation

```
glm.amputation <- glm(AMPUTATION ~ factor(ILLNESS_SEVERITY)+factor(diabet
es)+factor(Ulcers), data = amputation, family = "binomial")
summary(glm.amputation)
```

```
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.3049     0.4219 -10.203  < 2e-16
## factor(ILLNESS_SEVERITY)Low       -2.1956     0.6106  -3.596 0.000323
## factor(ILLNESS_SEVERITY)Moderate  -0.6745     0.4087  -1.651 0.098831
## factor(diabetes)uncontrolled       1.0397     0.3763   2.763 0.005730
## factor(Ulcers)1                    2.1879     0.3757   5.823 5.77e-09
```

Exercise 2: Amputation

```
# Odds Ratio
round(exp(glm.amputation$coefficients),3)
```

```
##                          (Intercept)      factor(ILLNESS_SEVERITY)Low
##                                0.014                            0.111
## factor(ILLNESS_SEVERITY)Moderate      factor(diabetes)uncontrolled
##                                0.509                            2.828
##                   factor(Ulcers)1
##                                8.917
```

A person with high illness_severity, uncontrolled diabetes and at least one Ulcers has the highest chance to have the amputation

Exercise 2: Amputation

- Model selection (stepwise selection via AIC)
- Consider more predictors
  - ILLNESS_SEVERITY: High/ Moderate/ Low
  - SEX_CODE: F/ M
  - AGEGROUP: 0-17/ 18-44/ 45-64/ 65-74/ 75+
  - Diabetes: controlled/ Uncontrolled
  - Hypertension: 1/0
  - Ulcers: 1/0

Exercise 2: Amputation

```
model.null = glm(AMPUTATION ~ 1, data=amputation, family = binomial) # null
model : no predictor
model.full = glm(AMPUTATION ~ ., data=amputation, family = binomial) # full
model: all predictors

step.models<-step(model.null, scope = list(upper=model.full),
                  direction="both",test="Chisq", trace = F)

summary(step.models)   # summary of stepwise selection

##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -4.5915     0.4739  -9.688  < 2e-16 ***
## Ulcers                     2.1628     0.3765   5.744 9.23e-09 ***
## ILLNESS_SEVERITYLow       -1.9942     0.6268  -3.182  0.00146 **
## ILLNESS_SEVERITYModerate  -0.6195     0.4116  -1.505  0.13231
## diabetesuncontrolled       1.0332     0.3770   2.740  0.00614 **
## hypertension               0.5706     0.3781   1.509  0.13129
```

Predictions and Classifications

- Logistic regression: model a probability of having an event (e.g., having a heart attack)
- Can be related to **classification**. what if you get estimated probability as 0.8? or if 0.2?
  - ✓ if we observed their status, we can evaluate classification performance (plasma example)
- Can be used as a prediction tool for new input

Predictions and Classifications

- **0.5 cut-off** is reasonable
- **Sample proportion** can also be a cut-off
  - ✓ What about for classification on **rare disease**?
  - ✓ If sample data have only a few 1's (disease) and majority 0's (no disease), probabilities tends to be underestimated and 0.5 threshold can be too high

Predictions and Classifications

- Revisit plasma data: (esr ~ fibrinogen)
- We will obtain those results for the plasma model to see:
  - Predicted probabilities of unhealthy ESR
  - Frequencies for correctly and incorrectly classified observations
- How good is the classification?

Predictions and Classifications

	fibrinogen	esr	fit.prob	pred.class.1
12	2.35	0	0.07233014	0
13	5.06	1	0.91682381	1
14	3.34	1	0.32243042	0
15	2.38	1	0.07609529	0
16	3.15	0	0.25166154	0
17	3.53	1	0.40239891	0
18	2.68	0	0.12471720	0
19	2.60	0	0.10961646	0
20	2.23	0	0.05892903	0

Given data and true "esr" class

Calculated

$$\hat{p} = \frac{\exp(-6.85+1.83*fib)}{1+\exp(-6.85+1.83*fib)}$$

Estimated class of "esr" with 0.5 threshold

Predictions and Classifications

- Misclassification rate:
  - # of misclassified obs / total # of obs

```
# misclassification rate from 0.5 threshold
mean(plasma$esr != pred.class.1)
## [1] 0.125


# misclassification rate from sample proportion thre
shold
mean(plasma$esr != pred.class.2)
## [1] 0.28125
```

Predictions and Classifications

- Choose the optimal threshold is important topic in classification/ data mining
  - Goal: accurate prediction for **future observation**
  - Split the data into **train** and **test** set
  - Train set as a given information and test set as a future set
  - Find the threshold which minimizes the **"test error"**