cmmid centre for mathematical modelling of infectious diseases
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE
MIDSEA NETWORK
CERM Centre for Epidemic Research & Modelling
NUS National University of Singapore
Saw Swee Hock School of Public Health

TM-CM02 Biostatistics for Public Health

Lecture 4

Logistic regression

Kiesha Prem

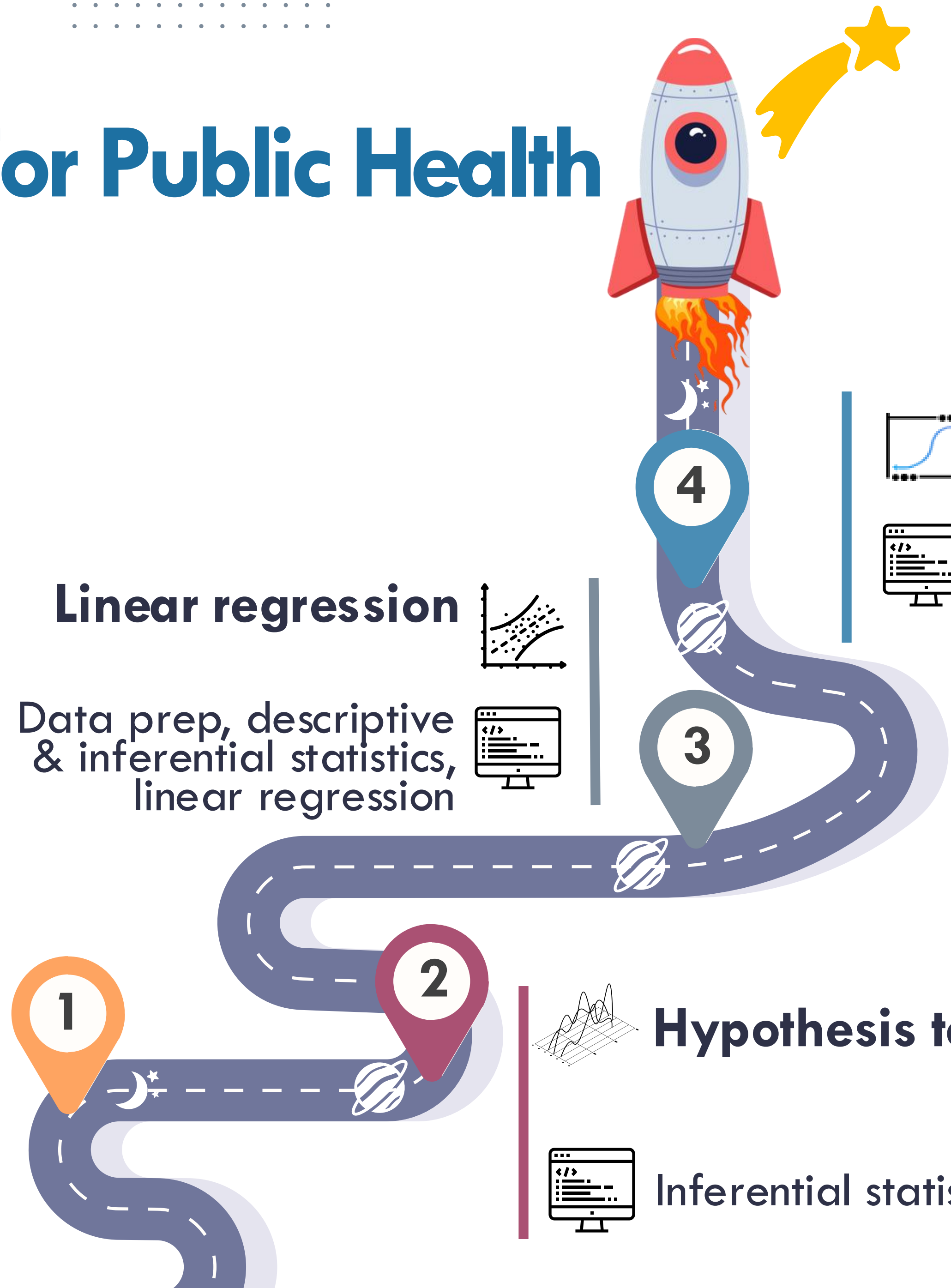
Saw Swee Hock School of Public Health, National University of Singapore

TM-CM02

Biostatistics for Public Health

quizzes

2 assignments



Categorical data analysis

Data prep, descriptive & inferential statistics

Hypothesis testing

Inferential statistics

Linear regression

Data prep, descriptive & inferential statistics, linear regression

Logistic regression

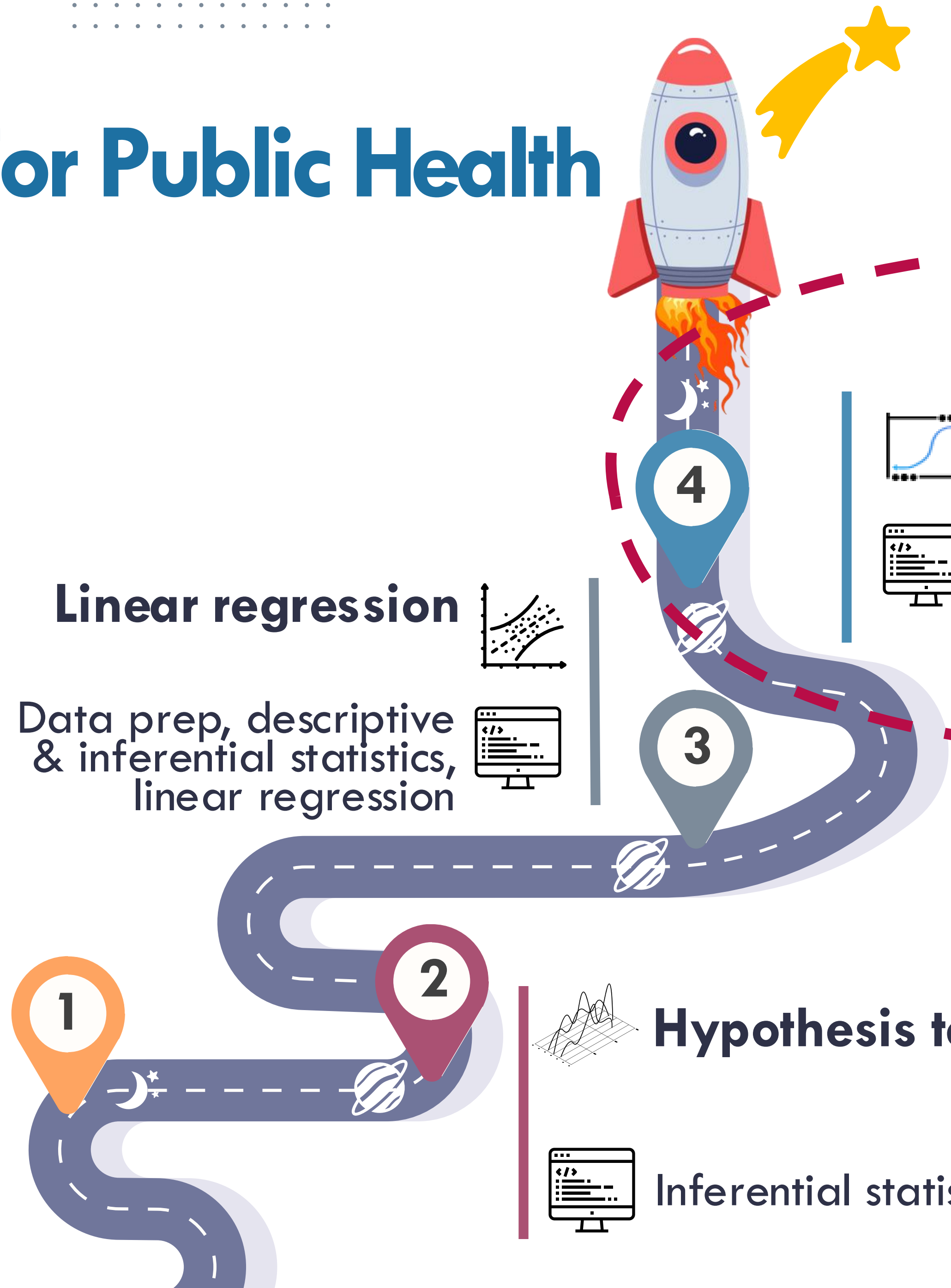
Data prep, descriptive & inferential statistics, logistics regression and others if we have time

NUS National University of Singapore | Saw Swee Hock School of Public Health

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments



1 **Categorical data analysis**

Data prep, descriptive & inferential statistics

2 **Hypothesis testing**

Inferential statistics

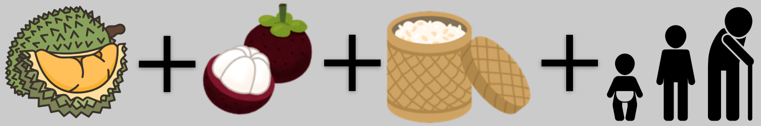
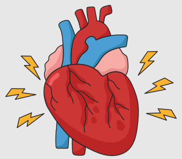
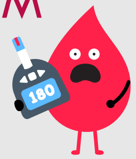
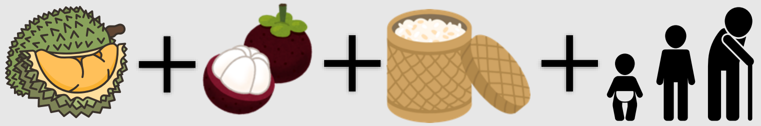
3 **Linear regression**

Data prep, descriptive & inferential statistics, linear regression

4 **Logistic regression**

Data prep, descriptive & inferential statistics, logistics regression and others if we have time

Linear and logistic regression

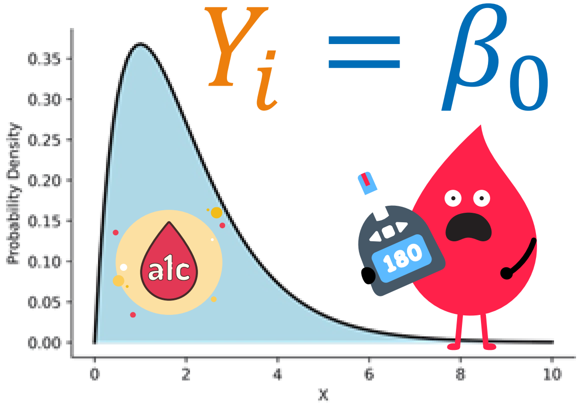
	Outcome of interest (Dependent variable)	**Simple** (1 independent variable)	**Multiple** (>1 independent variables)
Linear	The outcome is **continuous** and on the real line - Weight* - LDL cholesterol* - hbA1c* *may need to be transformed 	 $Y_i = \beta_0 + \beta_1 X_{i,1} + \varepsilon_i$	 $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$
Logistic	The outcome is **binary**/dichotomous - CVD (1/0 Yes/No) - T2DM (1/0 Yes/No) **0:** no T2DM **1:** T2DM   	 $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i,1} + \varepsilon_i$	 $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$

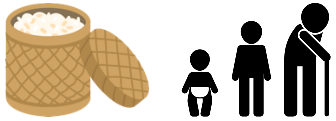
Logistic regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$





Random error

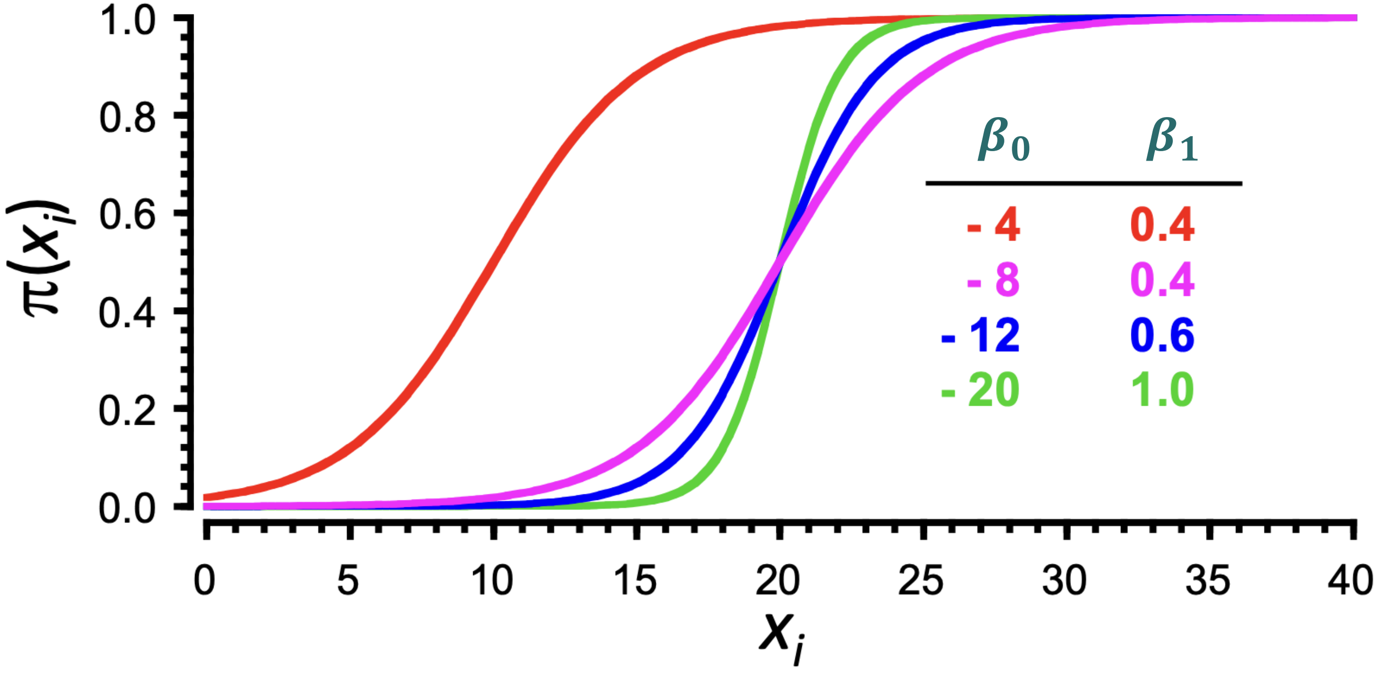
Dependent variable

0: no T2DM

1: T2DM

Modelling disease outcome

Modelling log odds: Simple logistic regression



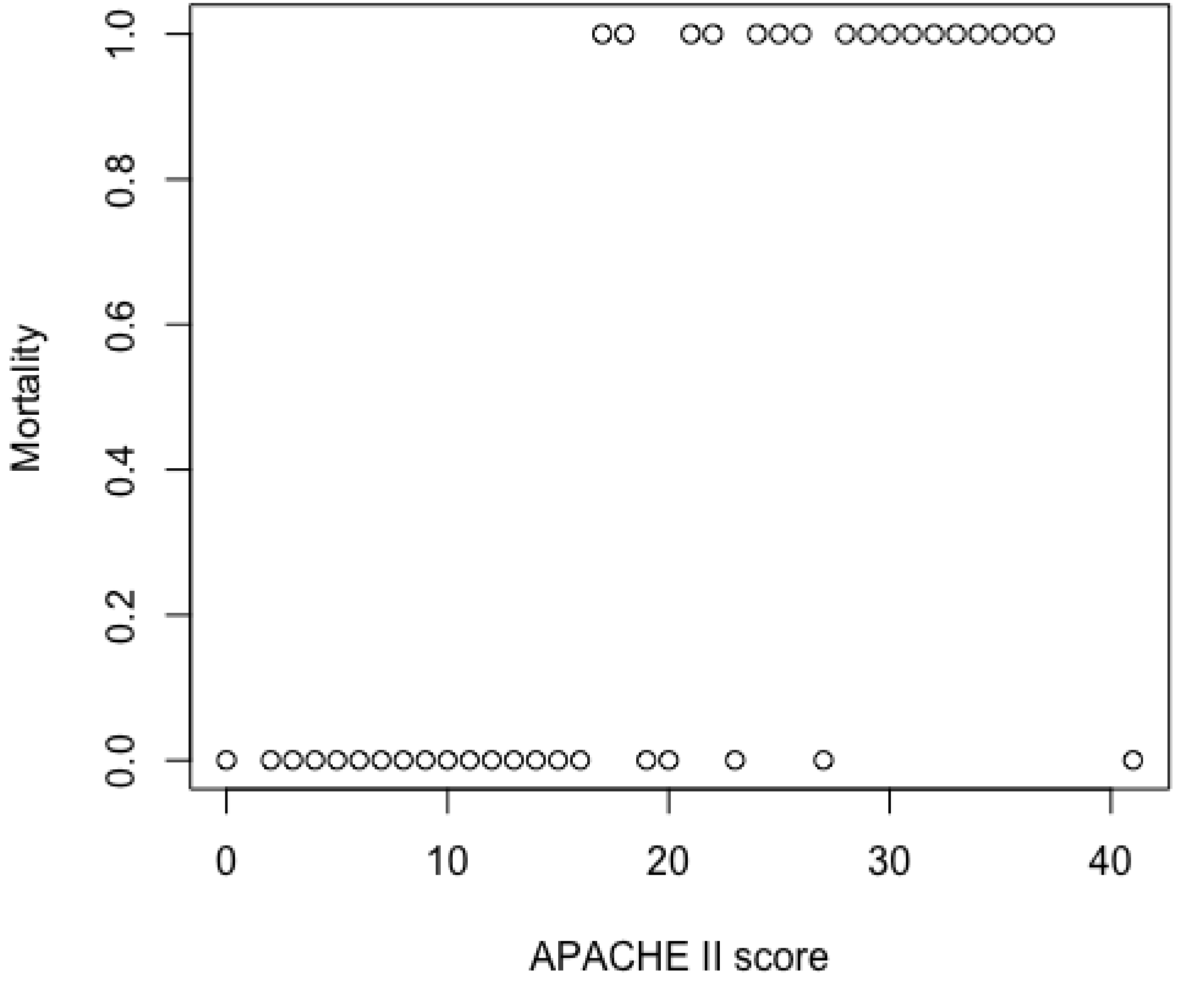
Logistic regression fits probability functions of the following form:

$$\pi(x_i) = \frac{\exp\{\boldsymbol{\beta_0} + \boldsymbol{\beta_1} x_i\}}{1 + \exp\{\boldsymbol{\beta_0} + \boldsymbol{\beta_1} x_i\}}$$

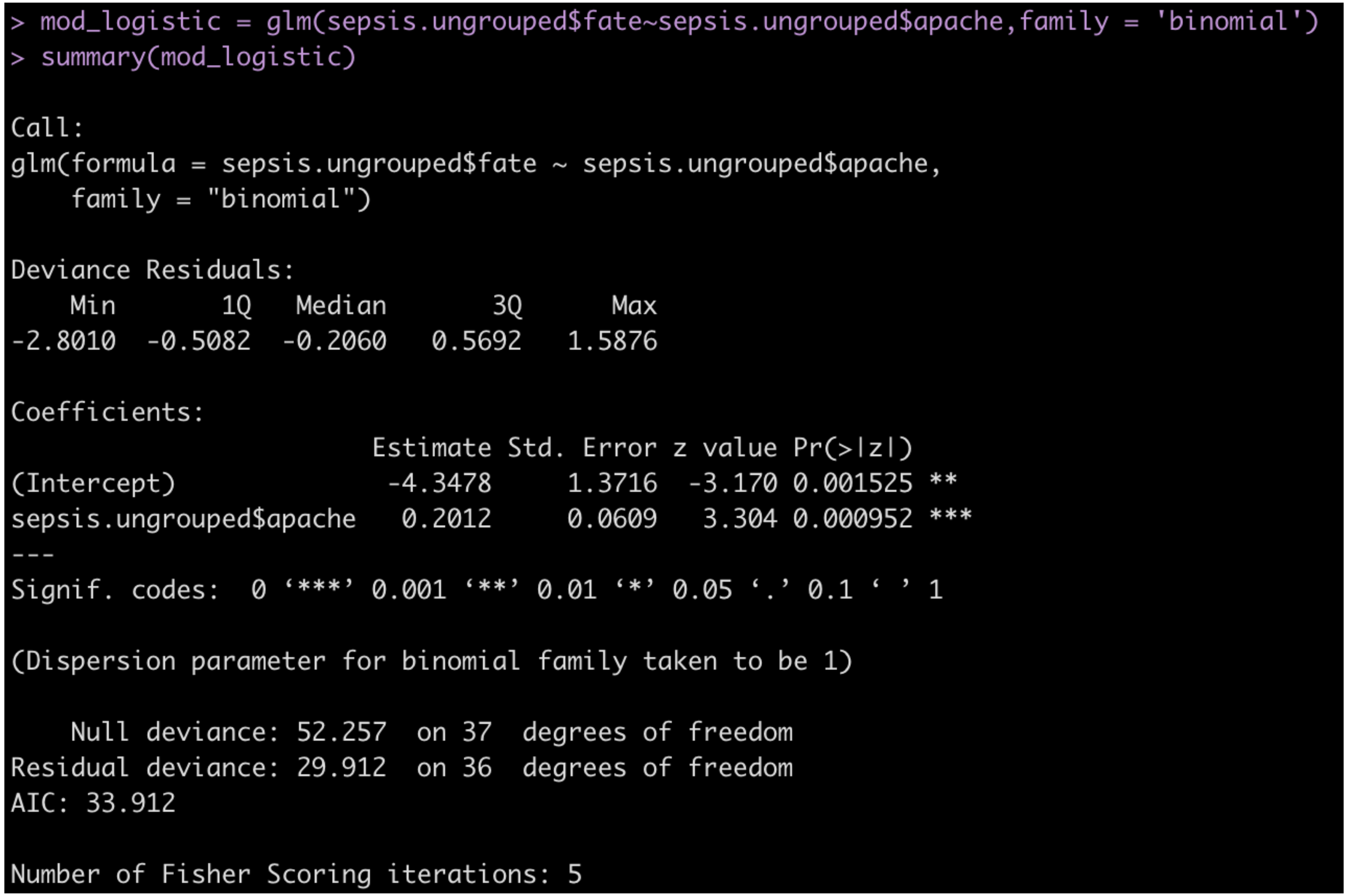
$$\mathrm{logit}\big(\pi(x_i)\big) = \log\left\{\frac{\pi(x_i)}{1 - \pi(x_i)}\right\} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} x_i$$

Sepsis mortality

The APACHE II Score and Mortality in Sepsis (sepsis.ungrouped.RData) is used to **model mortality within 30 days (i.e., fate) with APACHE II score (i.e., apache) as the predictor.**

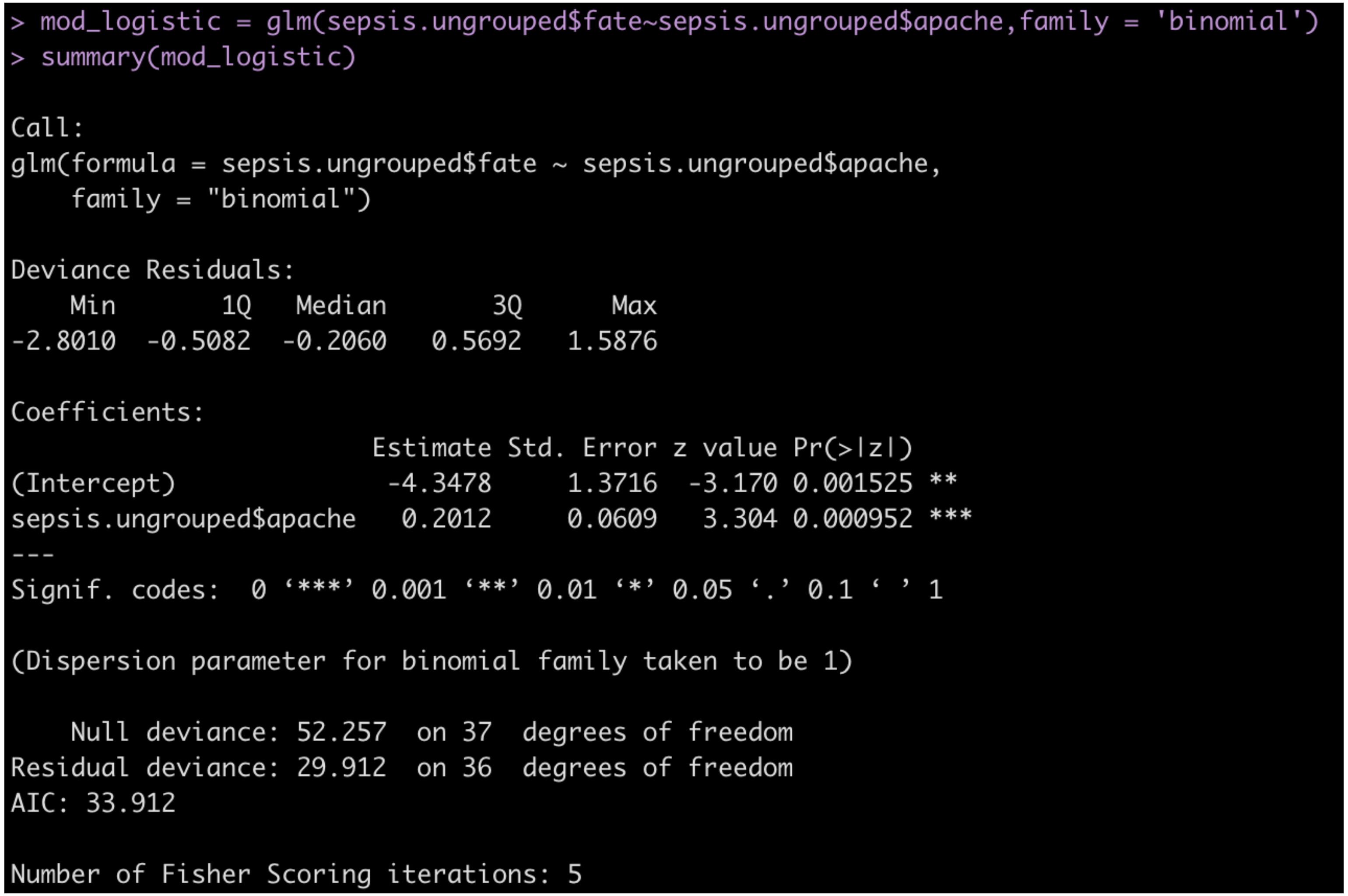


Simple logistic regression



```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8010  -0.5082  -0.2060   0.5692   1.5876

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.3478     1.3716  -3.170 0.001525 **
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

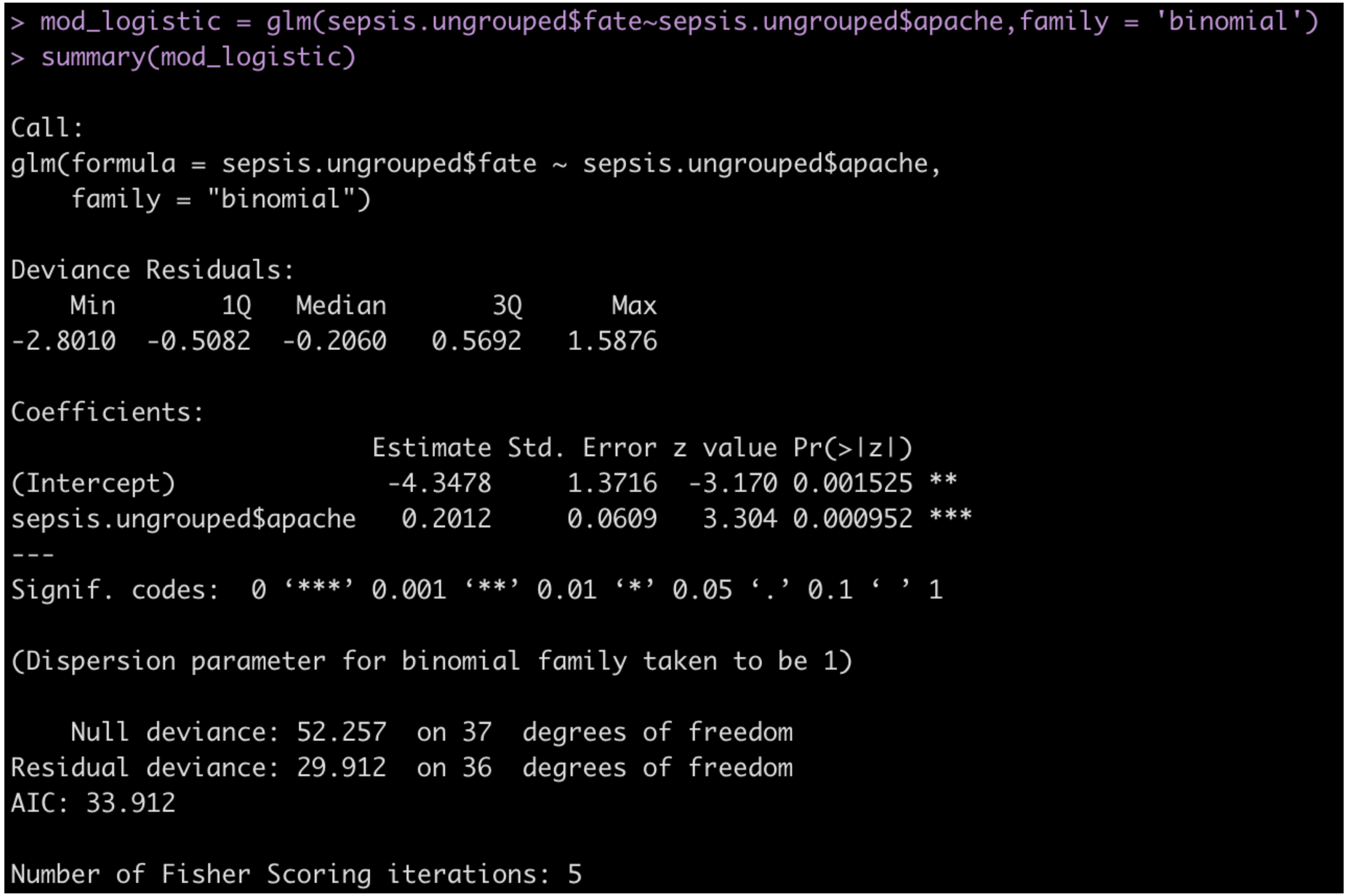
Simple logistic regression



```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8010  -0.5082  -0.2060   0.5692   1.5876

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.3478     1.3716  -3.170 0.001525 **
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

Can you write the linear predictor of the simple logistic regression model?

Simple logistic regression



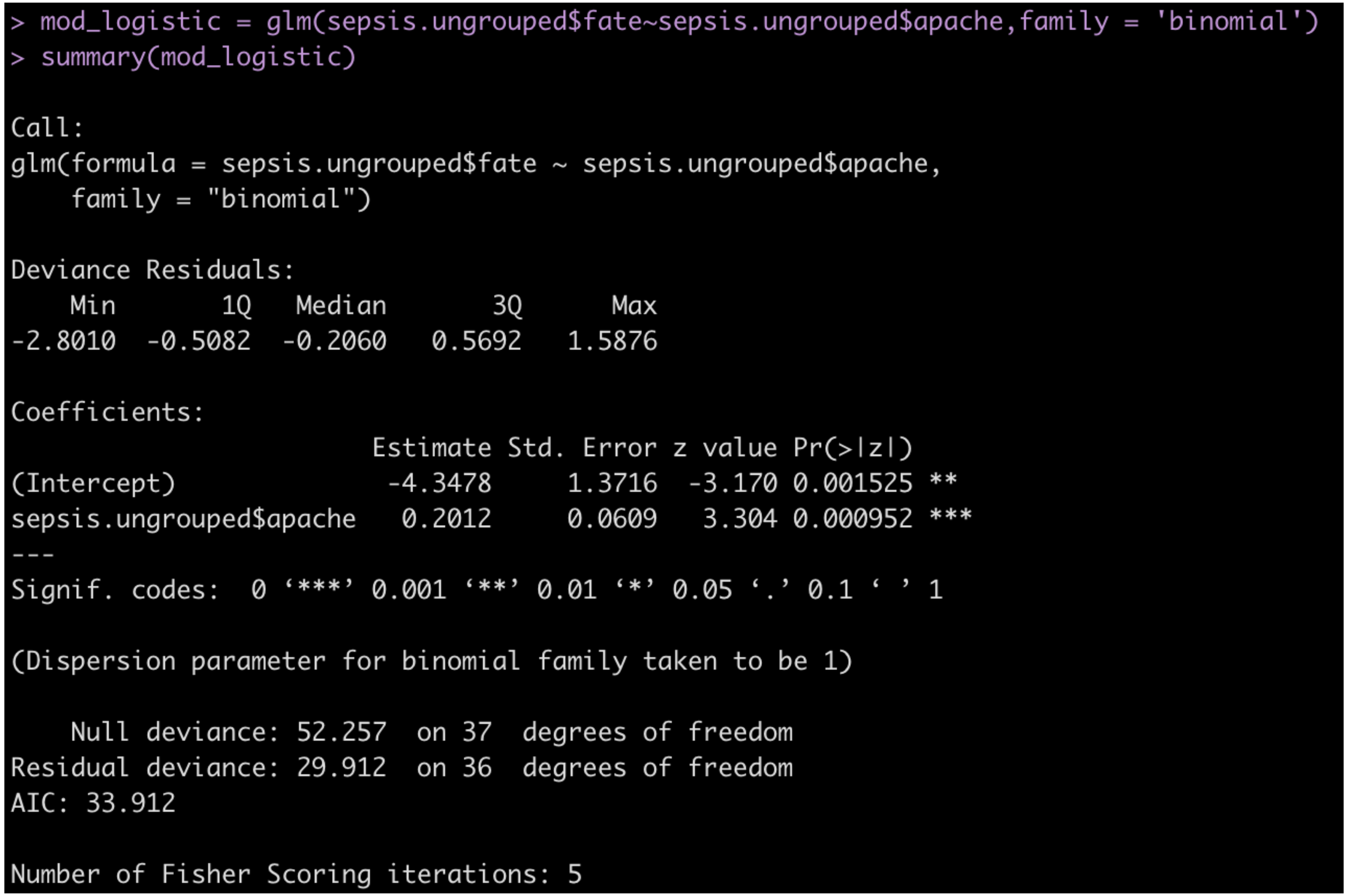
```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8010  -0.5082  -0.2060   0.5692   1.5876

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.3478     1.3716  -3.170 0.001525 **
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

Can you write the linear predictor of the simple logistic regression model?

The linear predictor of the simple logistic regression model is:

$$\beta_0 + \beta_1 \times \text{ apache}_i$$

Simple logistic regression



```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8010  -0.5082  -0.2060   0.5692   1.5876

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.3478     1.3716  -3.170 0.001525 **
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

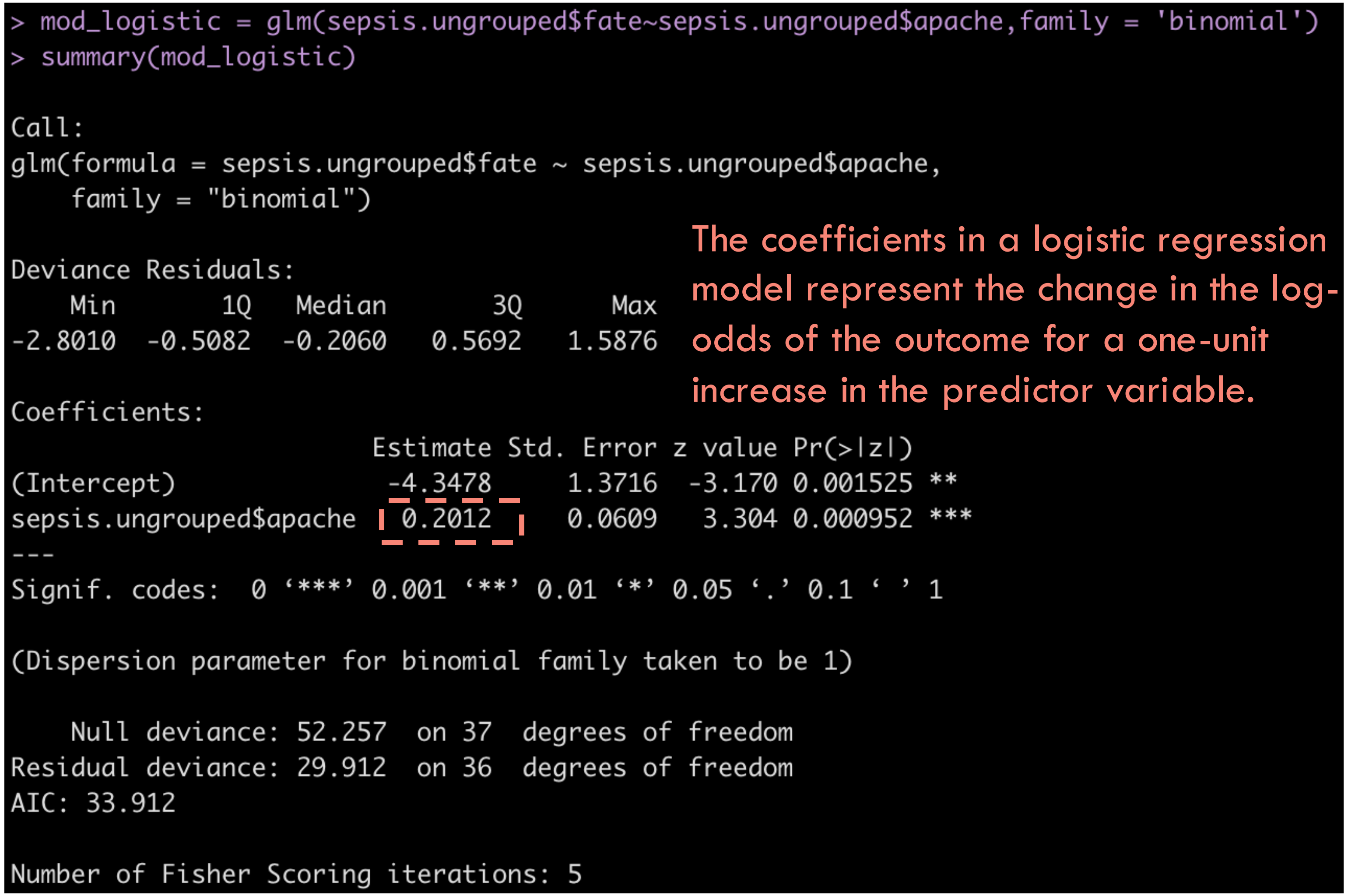
Can you write the linear predictor of the simple logistic regression model?

The linear predictor corresponding to the log odds for death within 30 days is:

$$\log\left\{\frac{\pi(\text{apache}_i)}{1-\pi(\text{apache}_i)}\right\}$$

$$= \text{logit}\{\pi(\text{apache}_i)\}$$

Simple logistic regression



```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8010  -0.5082  -0.2060   0.5692   1.5876

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.3478     1.3716  -3.170 0.001525 **
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

The coefficients in a logistic regression model represent the change in the log-odds of the outcome for a one-unit increase in the predictor variable.

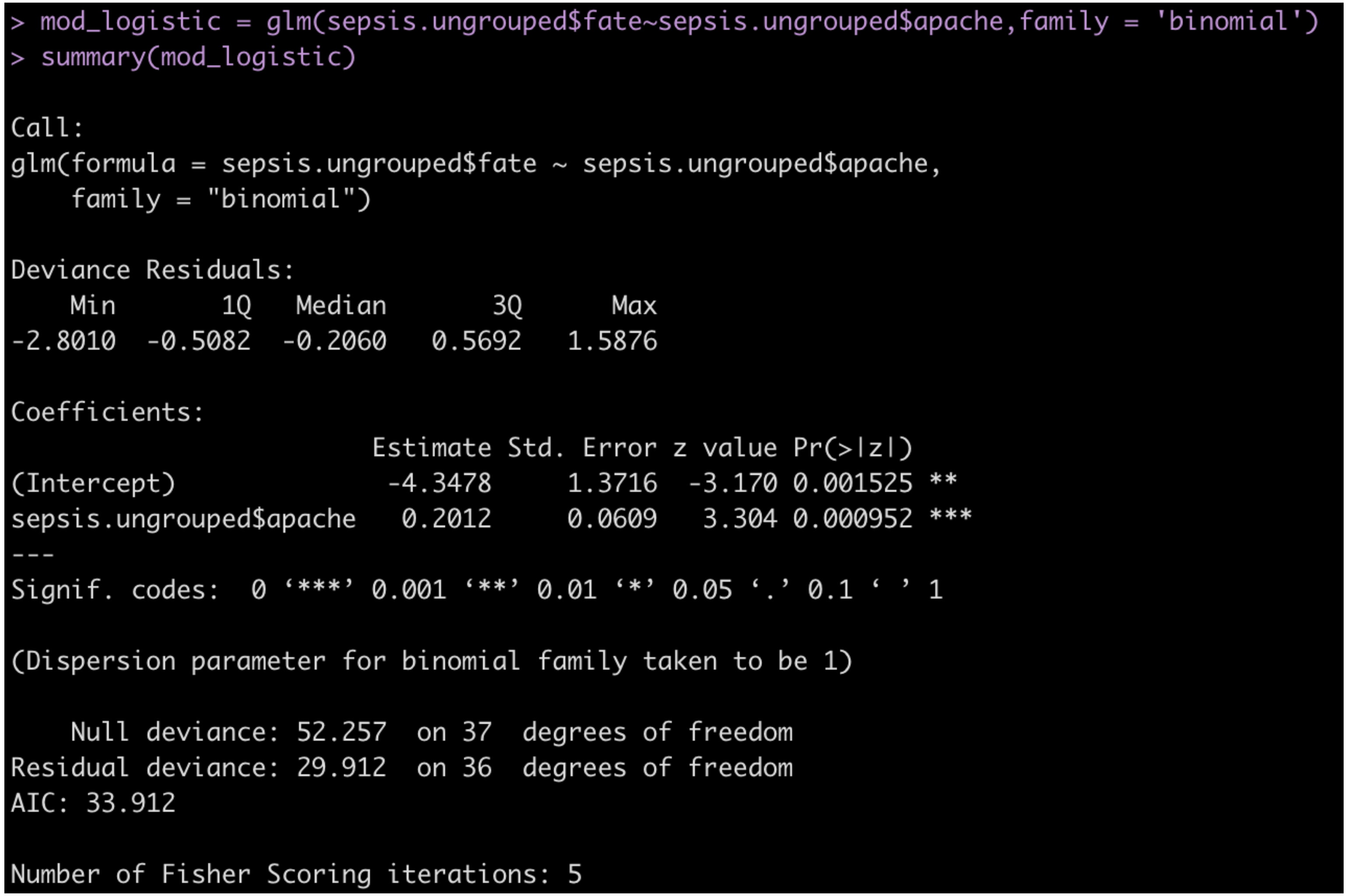
Can you write the linear predictor of the simple logistic regression model?

The linear predictor corresponding to the log odds for death within 30 days is:

$$\log\left\{\frac{\pi(\text{apache}_i)}{1-\pi(\text{apache}_i)}\right\} = \text{logit}\{\pi(\text{apache}_i)\}$$

Probability of death within 30 days at a specific APACHE II score

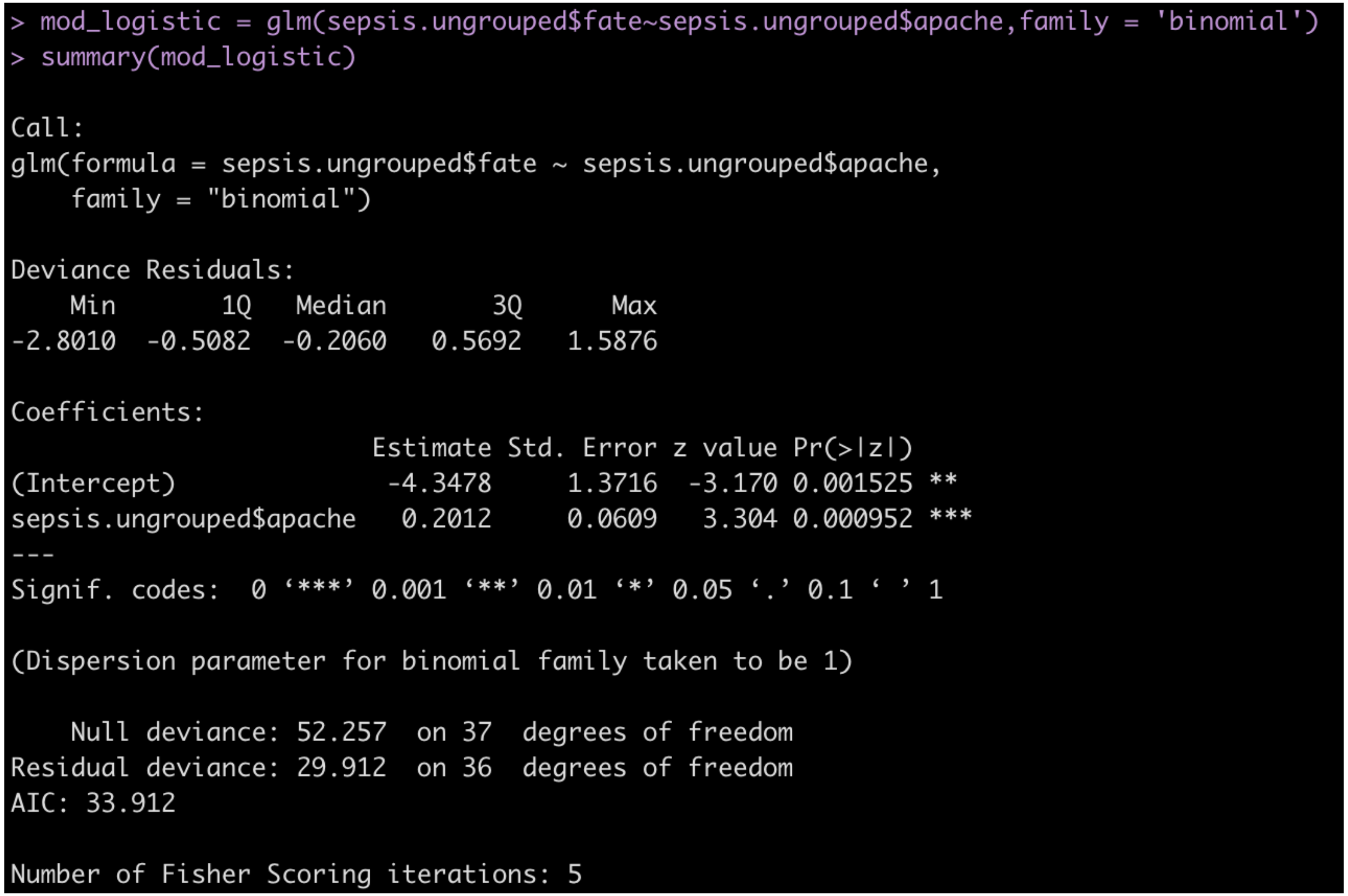
Simple logistic regression



```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8010  -0.5082  -0.2060   0.5692   1.5876

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.3478     1.3716  -3.170 0.001525 **
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

What is the value of the APACHE II score when 50% mortality is achieved?

Simple logistic regression



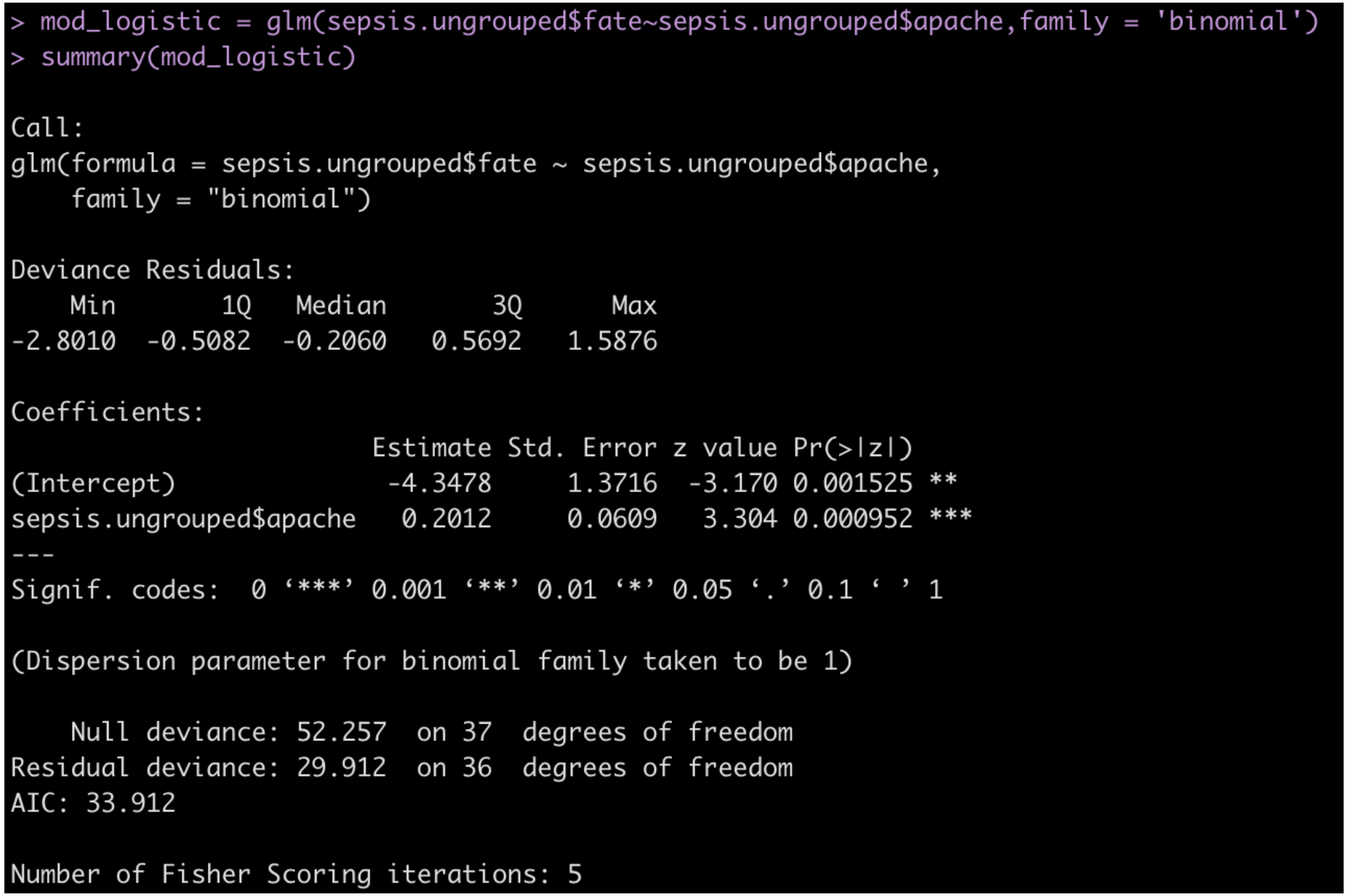
```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8010  -0.5082  -0.2060   0.5692   1.5876

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.3478     1.3716  -3.170 0.001525 **
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

What is the value of the APACHE II score when 50% mortality is achieved?

$$-\widehat{\beta_0}/\widehat{\beta_1} = -\frac{-4.348}{0.201} = 21.632$$

Point of inflection of a logistic regression model

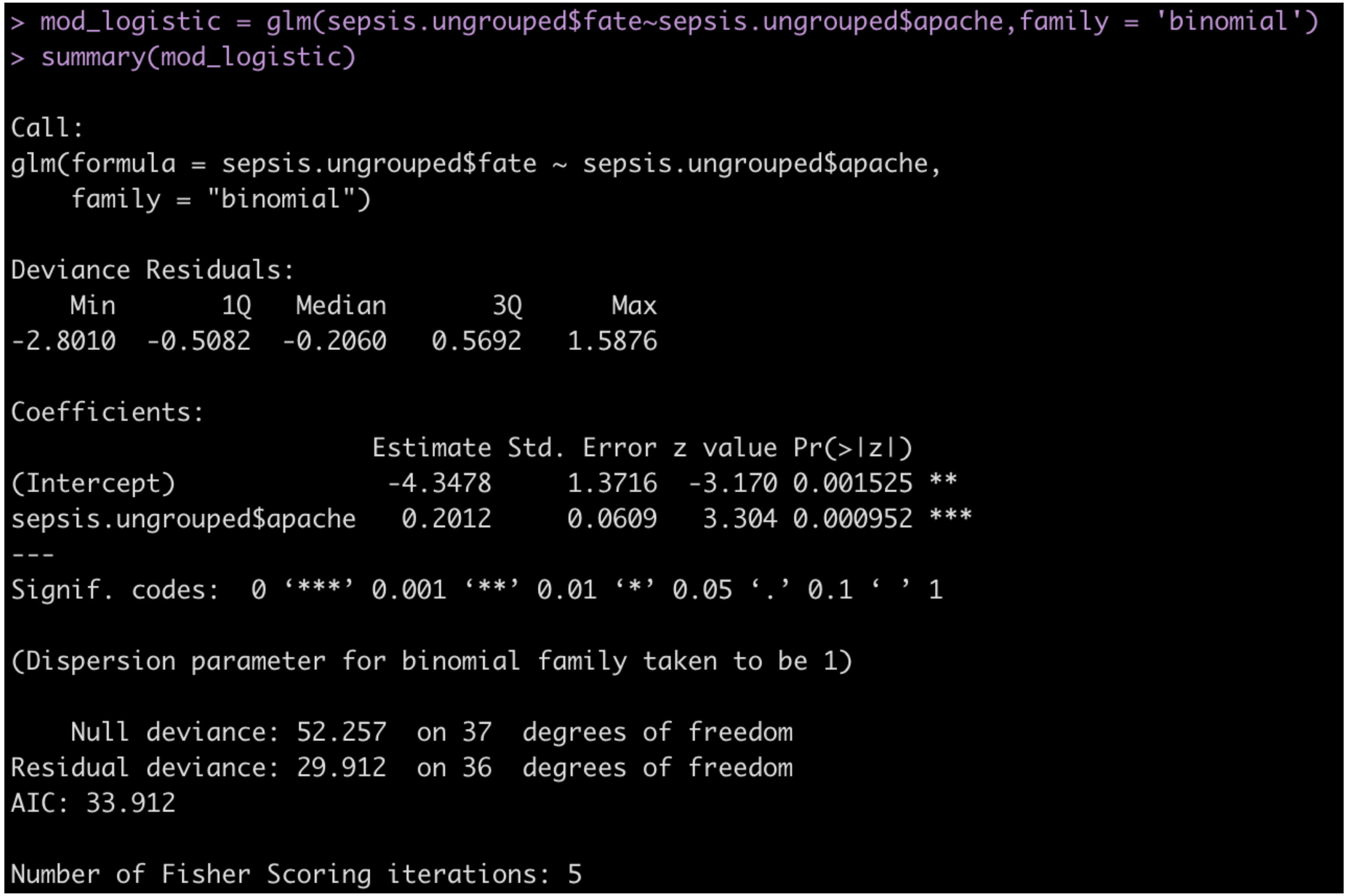
Simple logistic regression



```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8010  -0.5082  -0.2060   0.5692   1.5876

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.3478     1.3716  -3.170 0.001525 **
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

A patient with sepsis was recently admitted to hospital with an APACHE II score of 25. How would you rate the patient’s survival chances?

Simple logistic regression



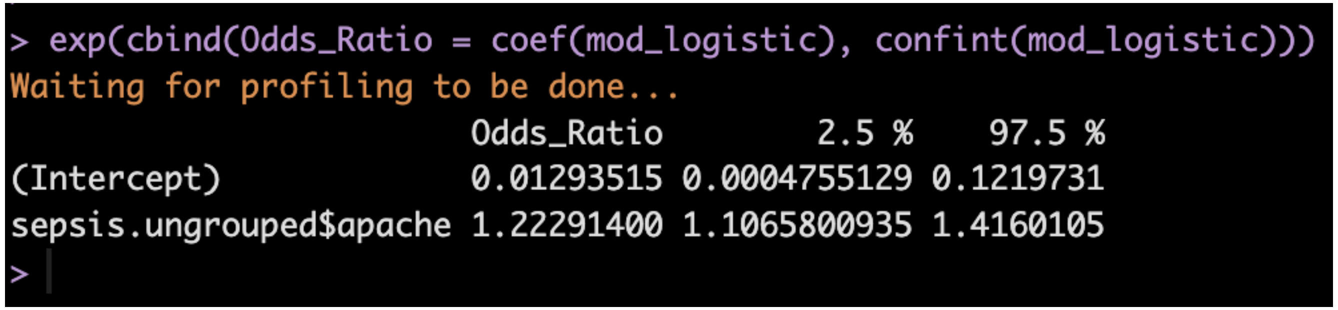
```
> mod_logistic = glm(sepsis.ungrouped$fate~sepsis.ungrouped$apache,family = 'binomial')
> summary(mod_logistic)

Call:
glm(formula = sepsis.ungrouped$fate ~ sepsis.ungrouped$apache, 
    family = "binomial")

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.8010  -0.5082  -0.2060   0.5692   1.5876  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -4.3478     1.3716  -3.170 0.001525 ** 
sepsis.ungrouped$apache   0.2012     0.0609   3.304 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.257  on 37  degrees of freedom
Residual deviance: 29.912  on 36  degrees of freedom
AIC: 33.912

Number of Fisher Scoring iterations: 5
```

A patient with sepsis was recently admitted to hospital with an APACHE II score of 25. How would you rate the patient’s survival chances?

Low; 50% mortality is achieved with an APACHE II score is 21.632. Given that the MLE for β_1 is 0.201>0 (i.e. positive association between mortality and APACHE II score), an APACHE II score greater than 21.632 (e.g. 25) will have more than 50% chance of mortality within 30 days.

Simple logistic regression

What is the odds ratio and its corresponding 95% confidence interval?

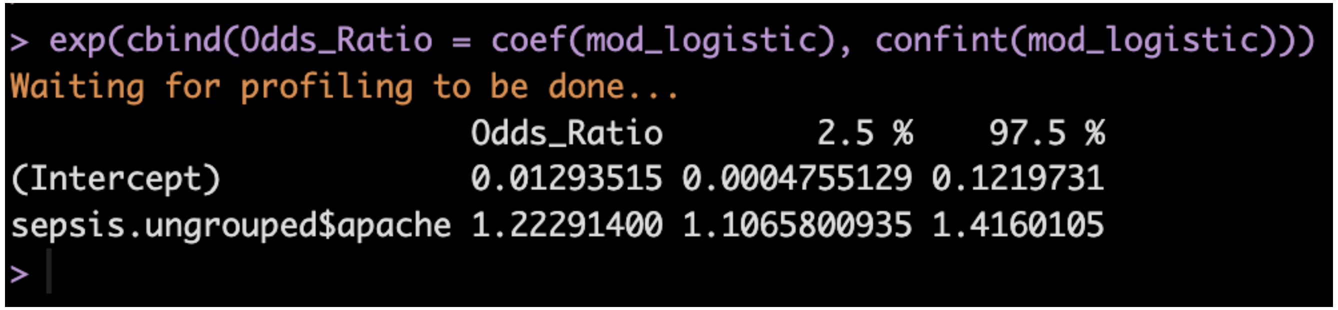


```
> exp(cbind(Odds_Ratio = coef(mod_logistic), confint(mod_logistic)))
Waiting for profiling to be done...
                        Odds_Ratio        2.5 %    97.5 %
(Intercept)             0.01293515 0.0004755129 0.1219731
sepsis.ungrouped$apache 1.22291400 1.1065800935 1.4160105
> 
```

Hint exponentiate the coefficients in a logistic regression to obtain odds ratios, representing the multiplicative change in the odds of the outcome for a one-unit increase in the predictor.

Simple logistic regression

What is the odds ratio and its corresponding 95% confidence interval?



```
> exp(cbind(Odds_Ratio = coef(mod_logistic), confint(mod_logistic)))
Waiting for profiling to be done...
                       Odds_Ratio        2.5 %    97.5 %
(Intercept)            0.01293515 0.0004755129 0.1219731
sepsis.ungrouped$apache 1.22291400 1.1065800935 1.4160105
> 
```

The **odds ratio** represents the change in the odds of success for a one-unit increase in the predictor variable while holding other variables constant.

Odds ratio: 1.223

95% CI: 1.107–1.1416

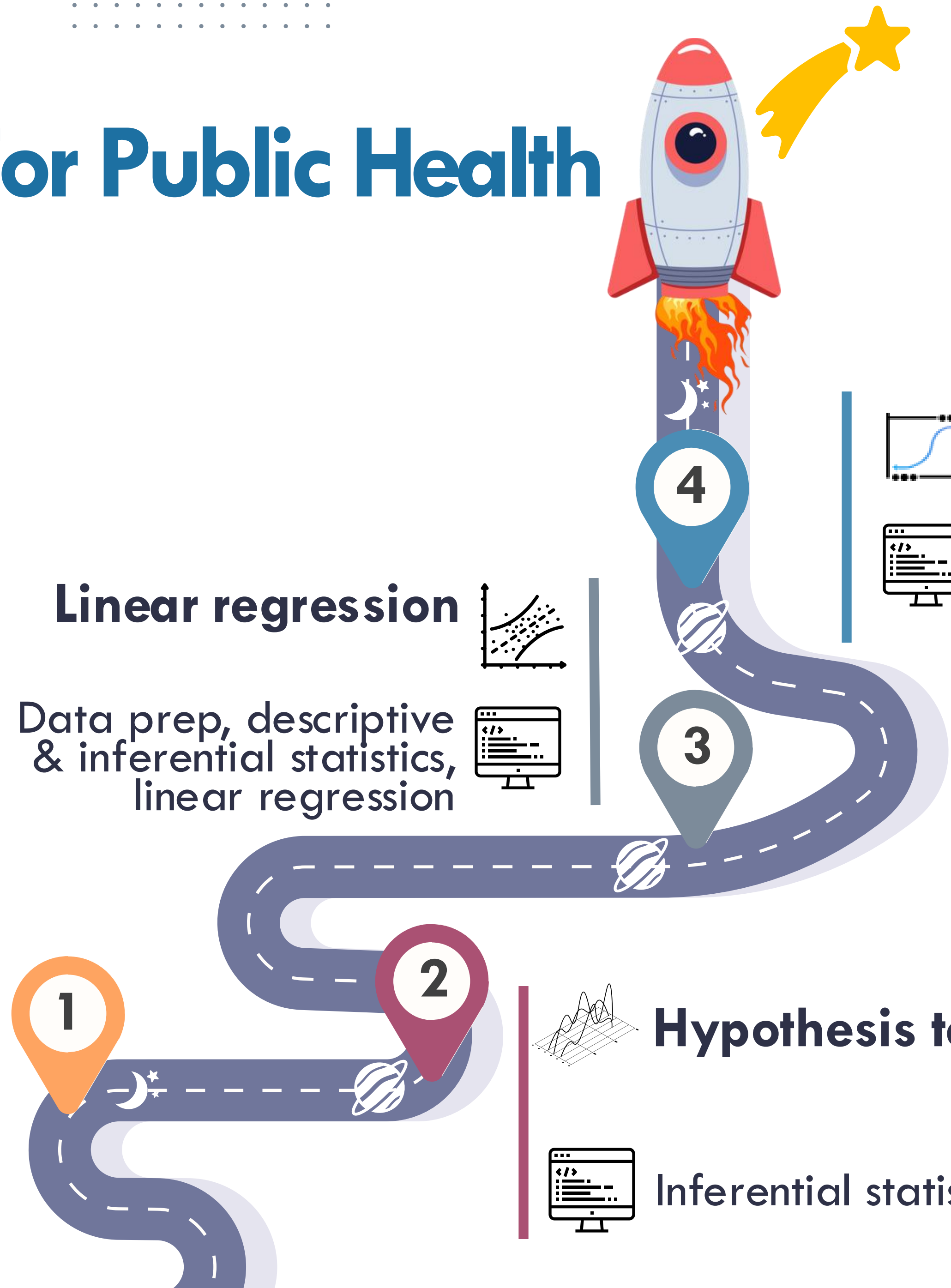
The odds ratio (OR) for death when the APACHE II score increases by 1 unit is 1.223 (95%CI: 1.107–1.1416; p-value<0.001). Therefore, APACHE II score and mortality within 30 days have a significant and positive association as the p-value is less than 0.05 (or 95%CI does not include the null value 1 which corresponds to the situation where the odds are the same regardless of the APACHE II), suggesting APACHE II score is a risk factor for mortality within 30 days.

NUS National University of Singapore | Saw Swee Hock School of Public Health

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments



1. **Categorical data analysis**
   Data prep, descriptive & inferential statistics

2. **Hypothesis testing**
   Inferential statistics

3. **Linear regression**
   Data prep, descriptive & inferential statistics, linear regression

4. **Logistic regression**
   Data prep, descriptive & inferential statistics, logistics regression and others if we have time

TM-CM02 Biostatistics for Public Health
Practical for logistic regression

Kiesha Prem

Saw Swee Hock School of Public Health, National University of Singapore

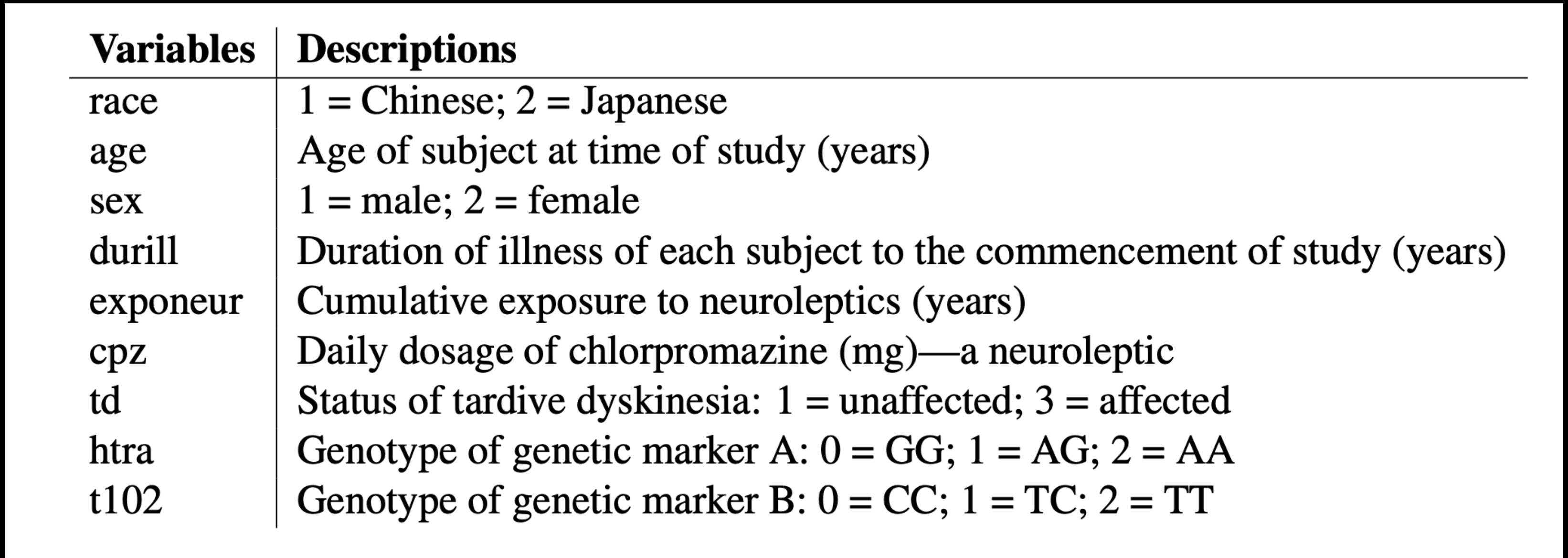
Tardive dyskinesia

Tardive dyskinesia is a movement disorder that develops in approximately 10–20% of patients on long-term neuroleptic treatment. It has been established previously that factors like age, sex, and duration of exposure to neuroleptics are risk factors. In addition, family history is believed to be a risk factor, indicating possible links to genetic factors.

Tardive dyskinesia

A research study investigating the effects of two genes was conducted (Tan et al. (2003) *Schizophrenia Research 65: 61–63*).

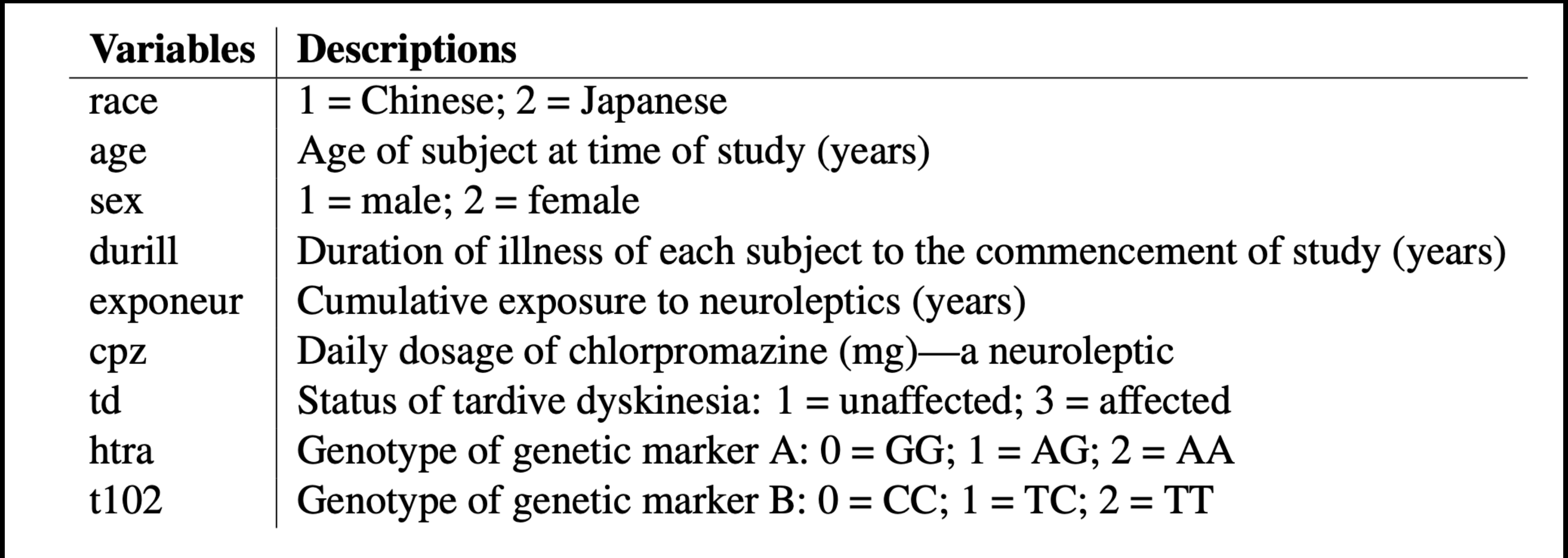
Part of the data for the study can be found in tardive.txt, consisting of the following variables:



Variables	Descriptions
race	1 = Chinese; 2 = Japanese
age	Age of subject at time of study (years)
sex	1 = male; 2 = female
durill	Duration of illness of each subject to the commencement of study (years)
exponeur	Cumulative exposure to neuroleptics (years)
cpz	Daily dosage of chlorpromazine (mg)—a neuroleptic
td	Status of tardive dyskinesia: 1 = unaffected; 3 = affected
htra	Genotype of genetic marker A: 0 = GG; 1 = AG; 2 = AA
t102	Genotype of genetic marker B: 0 = CC; 1 = TC; 2 = TT

Tardive dyskinesia

Carefully perform an exploratory data analysis before attempting to identify the factors associated with the onset of tardive dyskinesia.

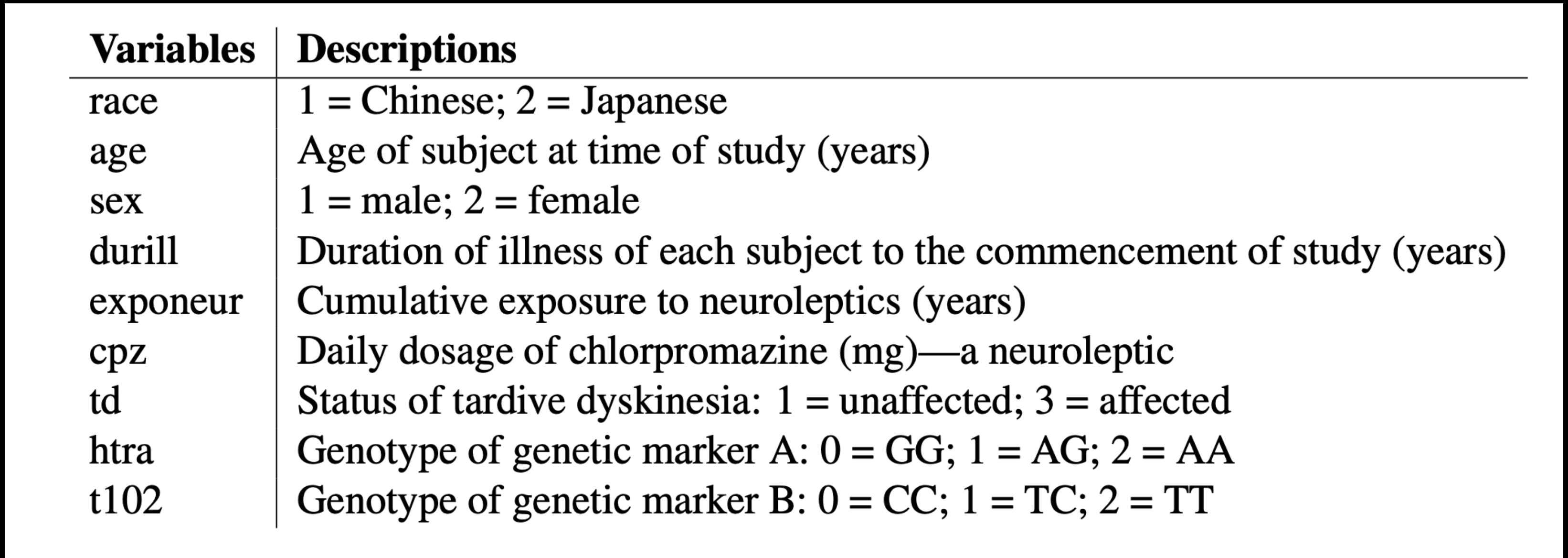


Variables	Descriptions
race	1 = Chinese; 2 = Japanese
age	Age of subject at time of study (years)
sex	1 = male; 2 = female
durill	Duration of illness of each subject to the commencement of study (years)
exponeur	Cumulative exposure to neuroleptics (years)
cpz	Daily dosage of chlorpromazine (mg)—a neuroleptic
td	Status of tardive dyskinesia: 1 = unaffected; 3 = affected
htra	Genotype of genetic marker A: 0 = GG; 1 = AG; 2 = AA
t102	Genotype of genetic marker B: 0 = CC; 1 = TC; 2 = TT

Tardive dyskinesia

Perform a series of univariate analyses before the use of a regression-based approach, to compare those that are affected with tardive dyskinesia with those that are unaffected.

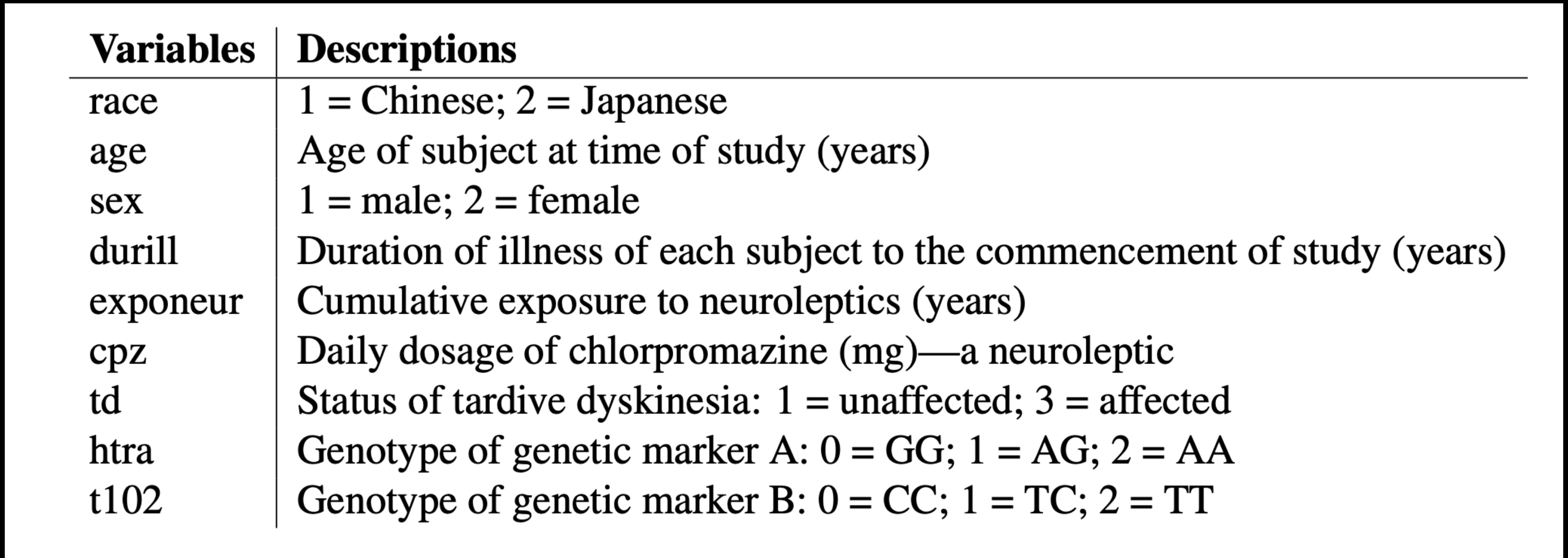
What can you say about the distributions of the numerical variables, especially for the variable cpz? How does that affect subsequent analyses?



Variables	Descriptions
race	1 = Chinese; 2 = Japanese
age	Age of subject at time of study (years)
sex	1 = male; 2 = female
durill	Duration of illness of each subject to the commencement of study (years)
exponeur	Cumulative exposure to neuroleptics (years)
cpz	Daily dosage of chlorpromazine (mg)—a neuroleptic
td	Status of tardive dyskinesia: 1 = unaffected; 3 = affected
htra	Genotype of genetic marker A: 0 = GG; 1 = AG; 2 = AA
t102	Genotype of genetic marker B: 0 = CC; 1 = TC; 2 = TT

Tardive dyskinesia

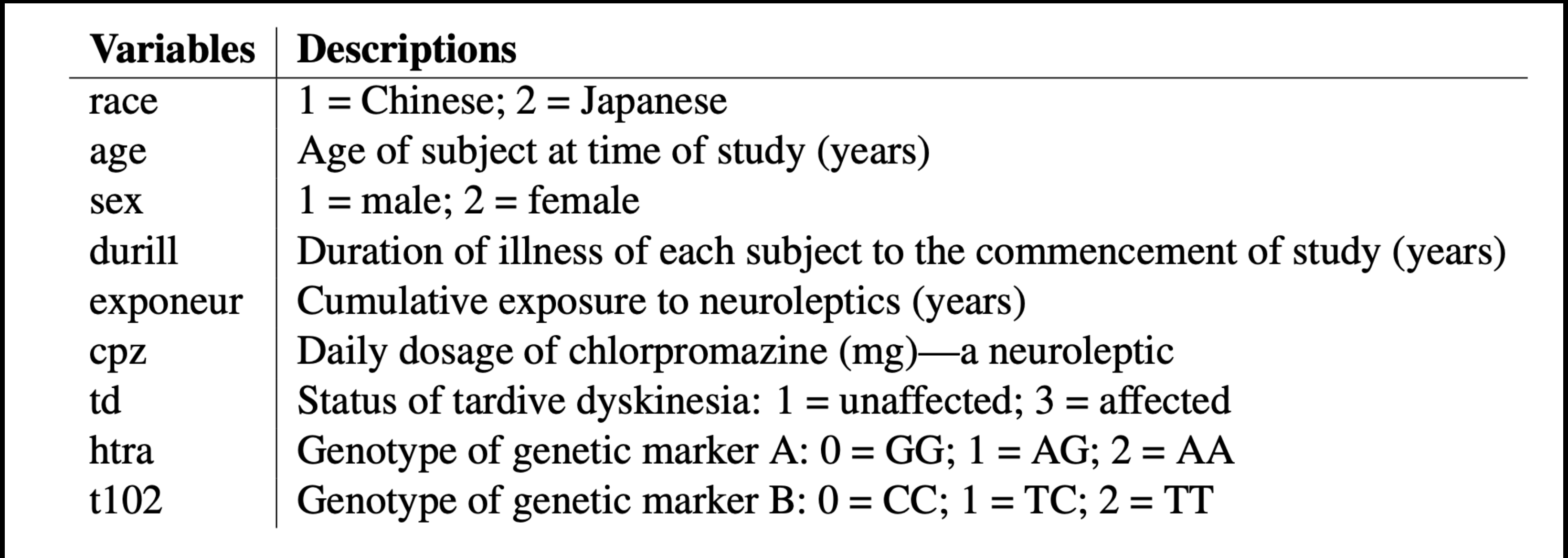
What are the differences between the unaffected and the affected?



Variables	Descriptions
race	1 = Chinese; 2 = Japanese
age	Age of subject at time of study (years)
sex	1 = male; 2 = female
durill	Duration of illness of each subject to the commencement of study (years)
exponeur	Cumulative exposure to neuroleptics (years)
cpz	Daily dosage of chlorpromazine (mg)—a neuroleptic
td	Status of tardive dyskinesia: 1 = unaffected; 3 = affected
htra	Genotype of genetic marker A: 0 = GG; 1 = AG; 2 = AA
t102	Genotype of genetic marker B: 0 = CC; 1 = TC; 2 = TT

Tardive dyskinesia

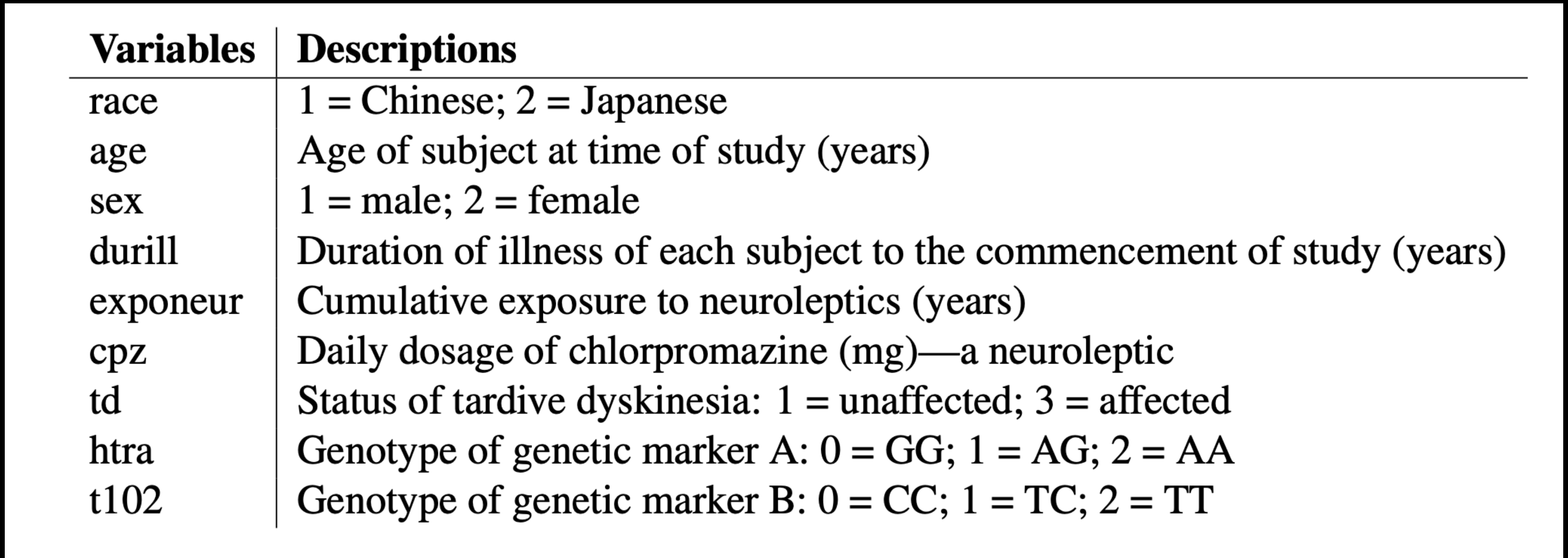
Explore the relationship between some of the numerical variables.
What can you say about the relationships between some of these variables?



Variables	Descriptions
race	1 = Chinese; 2 = Japanese
age	Age of subject at time of study (years)
sex	1 = male; 2 = female
durill	Duration of illness of each subject to the commencement of study (years)
exponeur	Cumulative exposure to neuroleptics (years)
cpz	Daily dosage of chlorpromazine (mg)—a neuroleptic
td	Status of tardive dyskinesia: 1 = unaffected; 3 = affected
htra	Genotype of genetic marker A: 0 = GG; 1 = AG; 2 = AA
t102	Genotype of genetic marker B: 0 = CC; 1 = TC; 2 = TT

Tardive dyskinesia

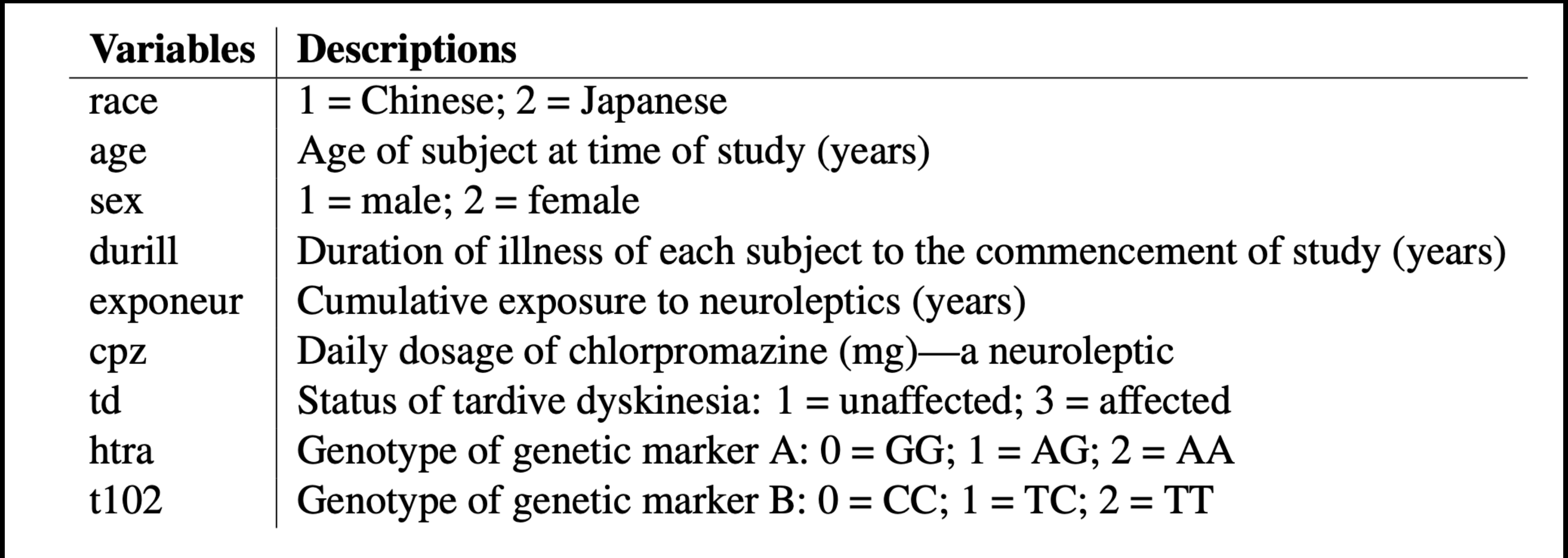
Perform logistic regression.



Variables	Descriptions
race	1 = Chinese; 2 = Japanese
age	Age of subject at time of study (years)
sex	1 = male; 2 = female
durill	Duration of illness of each subject to the commencement of study (years)
exponeur	Cumulative exposure to neuroleptics (years)
cpz	Daily dosage of chlorpromazine (mg)—a neuroleptic
td	Status of tardive dyskinesia: 1 = unaffected; 3 = affected
htra	Genotype of genetic marker A: 0 = GG; 1 = AG; 2 = AA
t102	Genotype of genetic marker B: 0 = CC; 1 = TC; 2 = TT

Tardive dyskinesia

What is your conclusion for the factors that influence tardive dyskinesia onset? How are the significant variables associated with the risk of disease onset? For example, are younger people at higher risk of the disease or are older people at higher risk? And what is the difference? How do you quantify this?



Variables	Descriptions
race	1 = Chinese; 2 = Japanese
age	Age of subject at time of study (years)
sex	1 = male; 2 = female
durill	Duration of illness of each subject to the commencement of study (years)
exponeur	Cumulative exposure to neuroleptics (years)
cpz	Daily dosage of chlorpromazine (mg)—a neuroleptic
td	Status of tardive dyskinesia: 1 = unaffected; 3 = affected
htra	Genotype of genetic marker A: 0 = GG; 1 = AG; 2 = AA
t102	Genotype of genetic marker B: 0 = CC; 1 = TC; 2 = TT

Assignment

Activity

Refer to the assignment questions and supporting R code

I have written some R codes for some of the questions. But you will still need to write some R codes to complete the questions – submit your completed assignment file before 21 March 2025, 12 pm lunchtime in Laos.

Saw Swee Hock School of Public Health

Thank you