STATISTICAL MODELLING: Theory and practice

Project 2: Survival data

GOALS: Binary data

AZT	AIDS_yes	Total
Yes	25	170
No	44	168

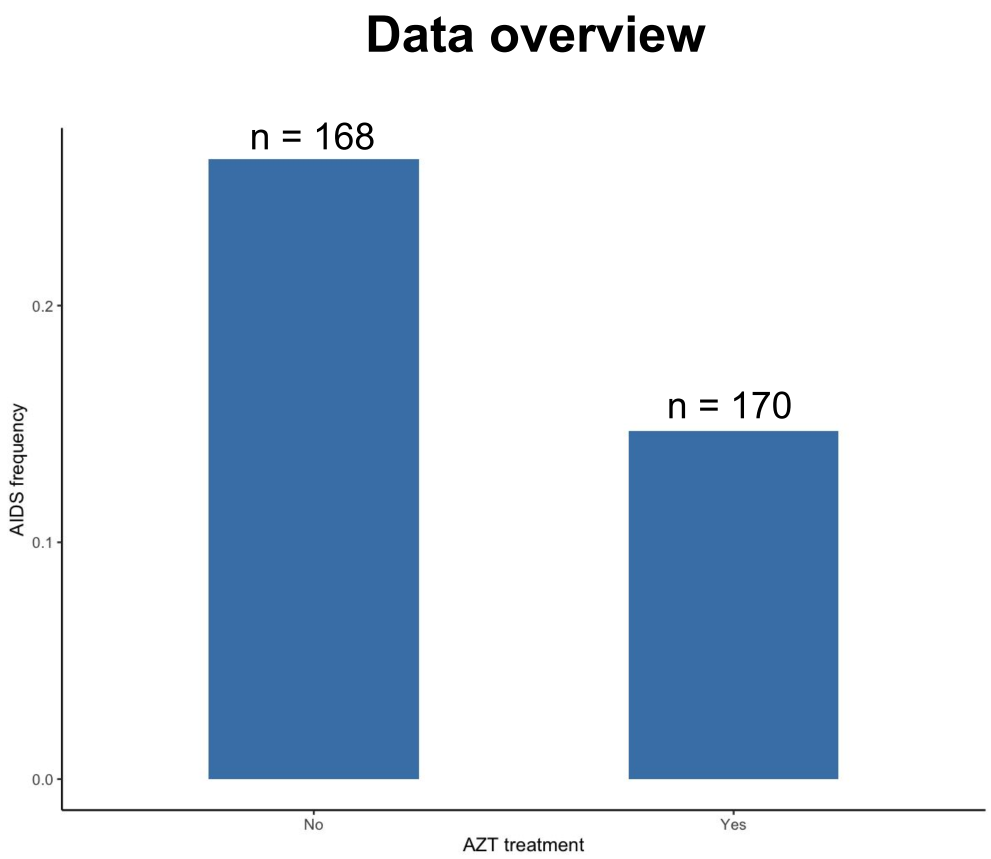
Assignment 1

1. Data overview
2. Fit a **binomial distribution** to the data
3. Fit the binomial separately to the two distributions and **test group difference**
4. Estimate parameters in the model using **log odds-ratio** and report **confidence interval**

Assignment 2

1. Fit a **logistic regression** for the binary outcome “AIDS” = yes vs “AIDS” = no and present the odds ratio for the AZT effect on AIDS.
2. Test the hypothesis (H0) of no **effect of AZT** using:
   a. Likelihood ratio test
   b. Wald test
   c. Score test

Assignment 1: DATA OVERVIEW



Is this visual difference significant ?

$H_0: \quad p_{AZT} = p_{noAZT}$

$H_1: \quad p_{AZT} \neq p_{noAZT}$

Test of equal proportions : p_AZT and p_noAZT

```
prop.test(no_AZT, AZT)

p-value = 0.01299
```

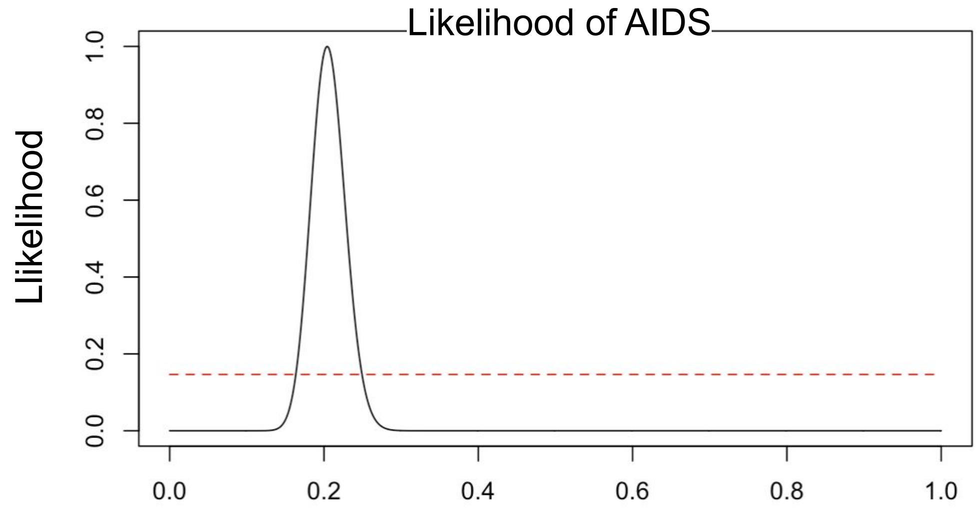
Based on this we can reject the null hypothesis

Assignment 1: FIT BINOMIAL DISTRIBUTION, TEST GROUP DIFFERENCE

All data regardless of treatment

$$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

x = 338 , n = 69

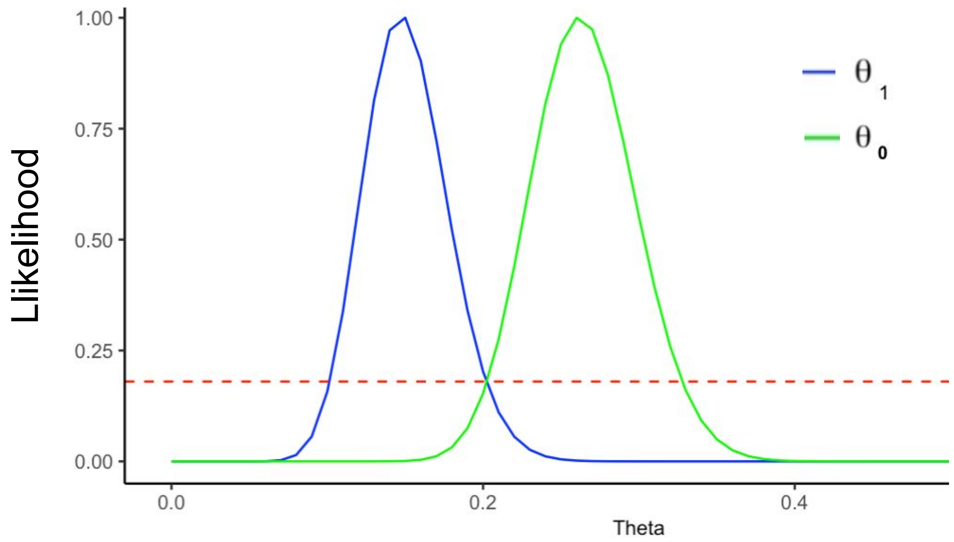


Group data per treatment

$$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

$\theta_1 \longrightarrow$ Treatment : n = 25 ; x = 170

$\theta_0 \longrightarrow$ No treatment : n = 44 ; x = 168



θ_1 = 0.1470588 CI = [0.09926 , 0.2054156]

θ_0 = 0.2619048 CI = [0.199347 , 0.331655]

Assignment 1:
LOG ODDS-RATIO

Log odds-ratio

$$p_0 = \frac{e^{\eta}}{1+e^{\eta}}$$

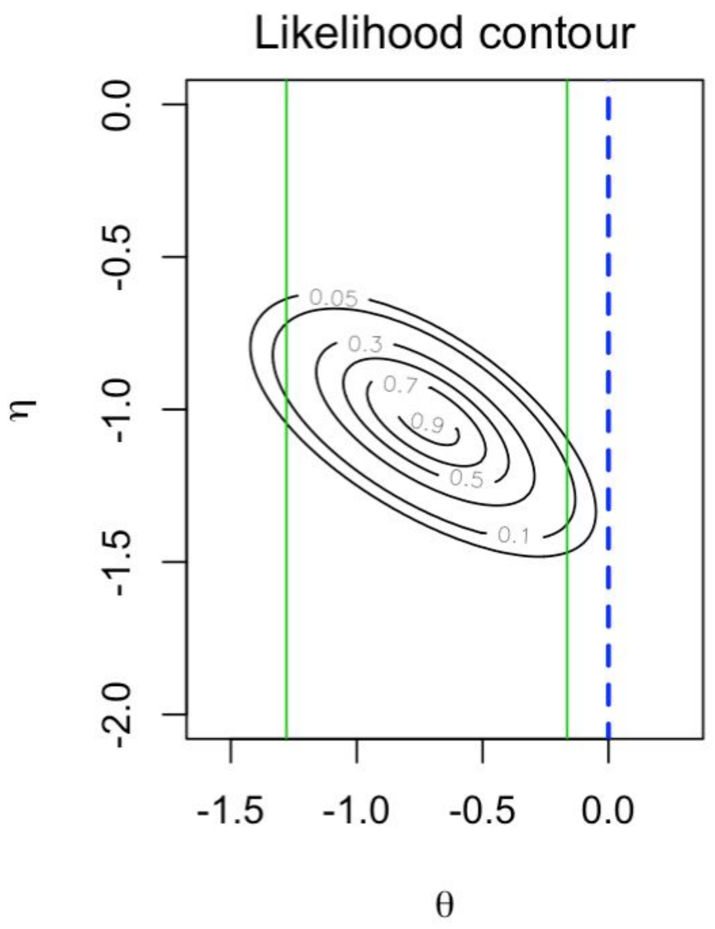
$$p_1 = \frac{e^{\theta+\eta}}{1+e^{\theta+\eta}}$$

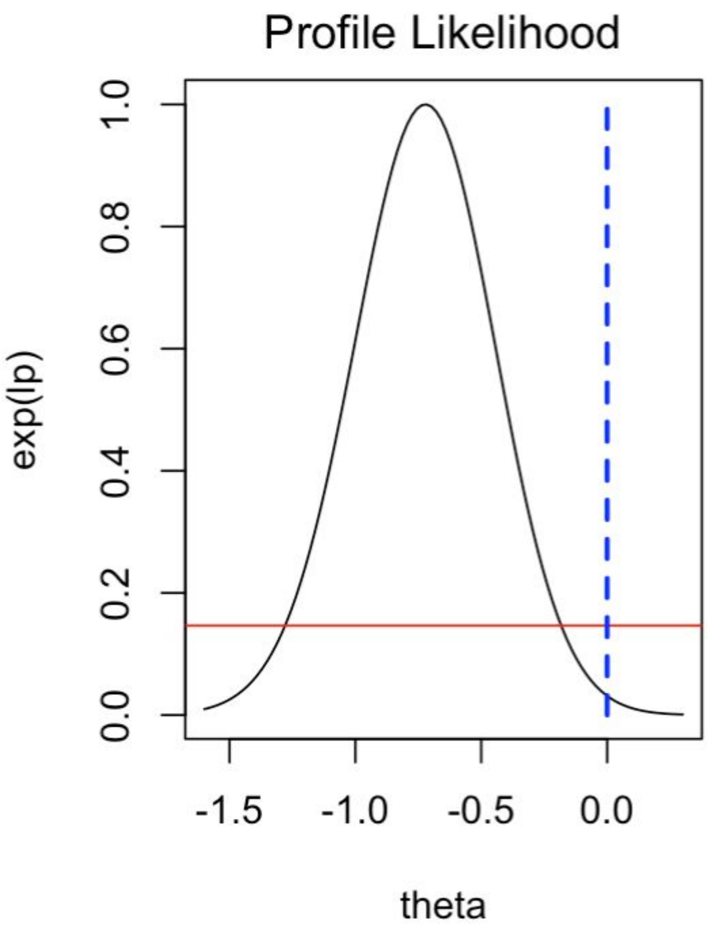
Likelihood of Log odds-ratio

$$L(\theta,\eta) =$$

$$= e^{\theta x} e^{\eta(x+y)} (1+e^{\theta+\eta})^{-m} (1+e^{\eta})^{-n}$$

$$\widehat{\theta} = \log \frac{x/(m-x)}{y/(n-y)}. \qquad \mathrm{se}(\widehat{\theta}) = \left(\frac{1}{x}+\frac{1}{y}+\frac{1}{m-x}+\frac{1}{n-y}\right)^{1/2}$$





η **= - 1.0360920**
θ **= - 0.7217664** **CI = [-0.1643134 , -1.279219]**

Assignment 2:
FIT THE REGRESSION MODEL

AZT	AIDS_yes	Total
Yes	25	170
No	44	168

Logistic regression

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Log-odds ratio

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

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_x x_x)}{1 + exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_x x_x)}$$

$model_0$ = `glm(formula = aids ~ 1, family = binomial)`

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.3606     0.1349  -10.08   <2e-16 ***

AIC: 344.12
```

$model_1$ = `glm(formula = aids ~ tx, family = binomial)`

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.0361     0.1755  -5.904 3.54e-09 ***
x             -0.7218     0.2787  -2.590  0.00961 **

AIC: 339.19
```

	Log-odds	2.5%	97.5%
$model_1$	0.4859	0.2783	0.833

Assignment 2:

BINARY DATA

HYPOTHESIS TO TEST

H_0 : $model_0 = model_1$

H_1: $model_0 \neq model_1$

LIKELIHOOD RATIO TEST

$$\tilde{Q} = \; = -2\log\left(\frac{L(\theta_0)}{L(\theta_1)}\right) \longrightarrow \chi^2 \longrightarrow$$

`p-value = 0.00848(df=2) **`

WALD TEST

$$z = \frac{\hat{\theta} - \theta_0}{se\left(\hat{\theta}\right)} \longrightarrow N(0,1) \longrightarrow$$

`p-value = 0.0048 **`

Assignment 2:

BINARY DATA

HYPOTHESIS TO TEST

H_0: $model_0$ = $model_1$

H_1: $model_0$ ≠ $model_1$

SCORE TEST

1. Calculate probability of a patient having AIDS

$$\theta_i = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$$

2. Calculate S(θ) and I(θ)

3. Solve the equation:

$$\underbrace{t(S(\hat{\beta}))}_{\text{transpose(S(θ))}} \underbrace{V(S(\hat{\beta}))^{-1}}_{\text{Information matrix (I(θ))}} \underbrace{S(\hat{\beta})}_{\text{S(θ)}}$$

4. Calculate p-value

$\chi^2 \longrightarrow$ **p-value = 0.0088 ****

GOALS: Survival time series

Assignment 1

1. Overview of **AIDS with treatment effect**
2. Fit **exponential distribution** to time:
   a. All data
   b. For the two treatments
3. **Likelihood comparison**
4. Find MLE of a **log-odds model** and compare with previous model
5. Find **Wald interval** for the treatment parameter
6. Derive theoretical results

Assignment 2

1. Descriptive statistics
2. Fit parametric survival models: Exponential, Weibull and Log-logistic
3. Choose best model:
   a. Present model
   b. Calculate Time ratio and hazard ratio
   c. Asses model with Cox-Snell residual

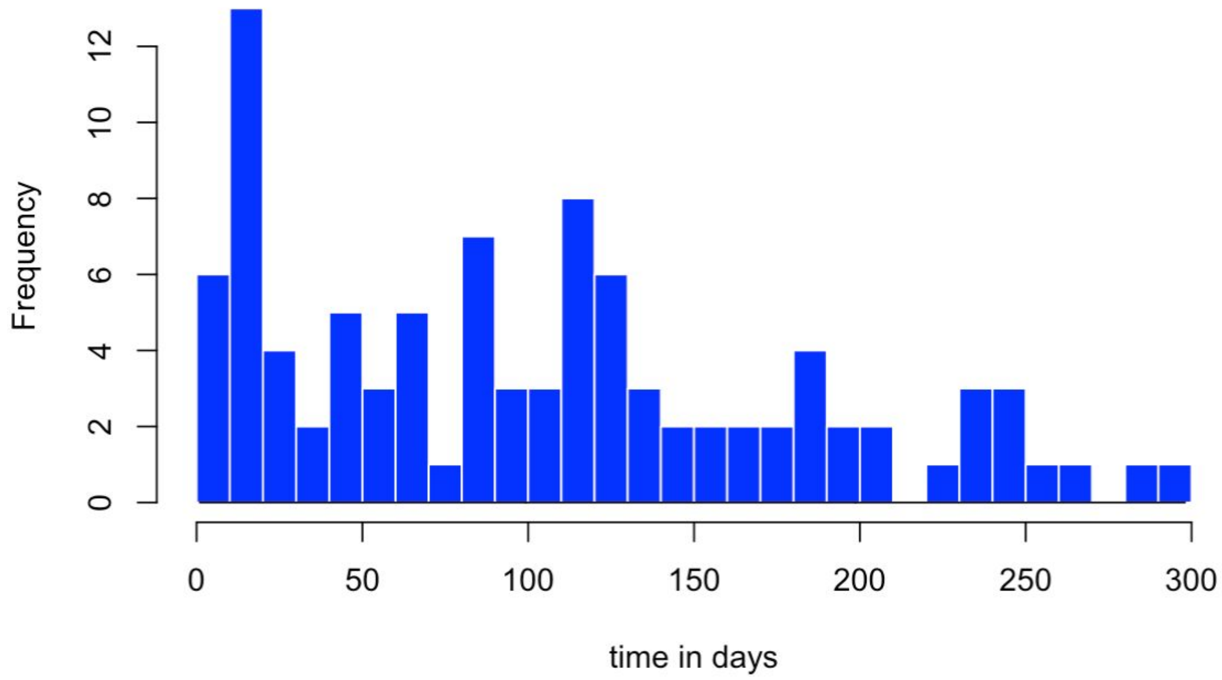
Assignment 1
DESCRIPTIVE STATISTICS

Study length:

No event: 364 days

Event: 298 days

Treatment	Event	Number	Proportion
Yes	Yes	514	0.446
Yes	No	63	0.055
No	Yes	541	0.470
No	No	33	0.028



Event = AIDS or death
Treatment = AZT

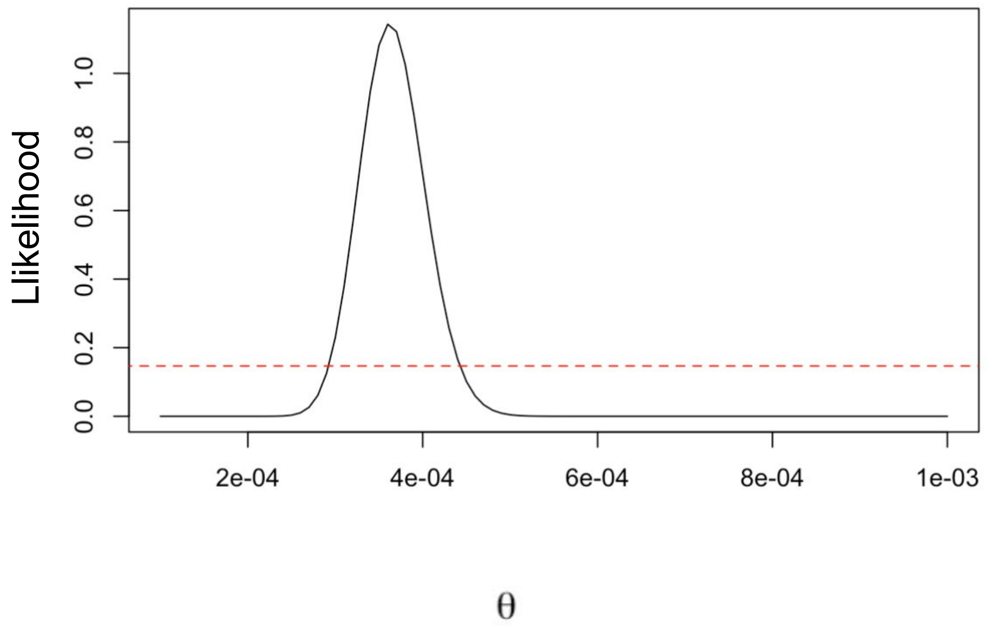
Assignment 1

FIT EXPONENTIAL DISTRIBUTION, GROUP DIFFERENCE

All data regardless of treatment

$$f(y) = \frac{1}{\lambda} e^{-y/\lambda}$$

$$LogL(x \mid \theta) = \sum_{i=1}^{n} ln(\theta) - \theta x_i$$

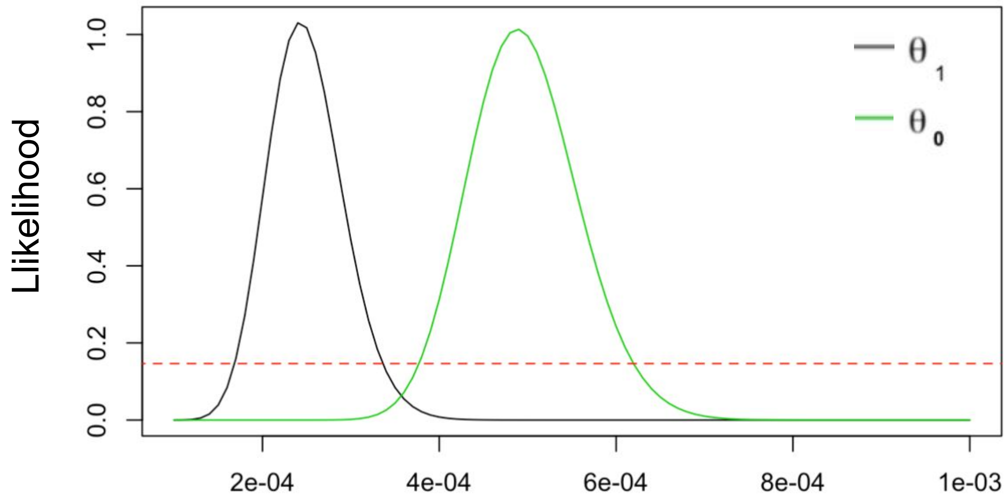


Group data per treatment

$$LogL(x \mid \theta) = \sum_{i=1}^{n} ln(\theta) - \theta x_i$$

$\theta_1 \longrightarrow$ Treatment

$\theta_0 \longrightarrow$ No Treatment



θ_1 = 0.0004986311 CI = [0.00039 , 0.000667]

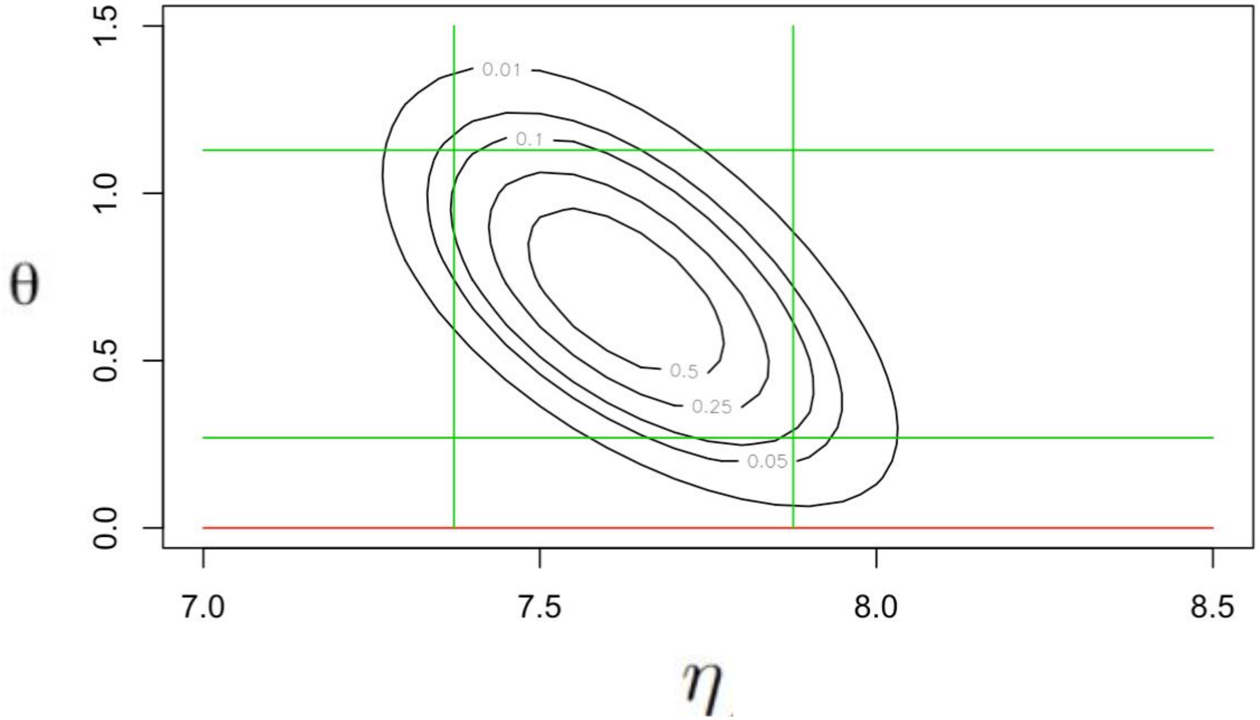
θ_0 = 0.0002535525 CI = [0.00018, 0.00036]

Assignment 1
LOG ODDS-RATIO

Log odds-ratio

$$\pi_y = \frac{e^{\eta}}{1 + e^{\eta}}$$

$$\pi_x = \frac{e^{\theta+\eta}}{1 + e^{\theta+\eta}}$$



Likelihood of Log odds-ratio

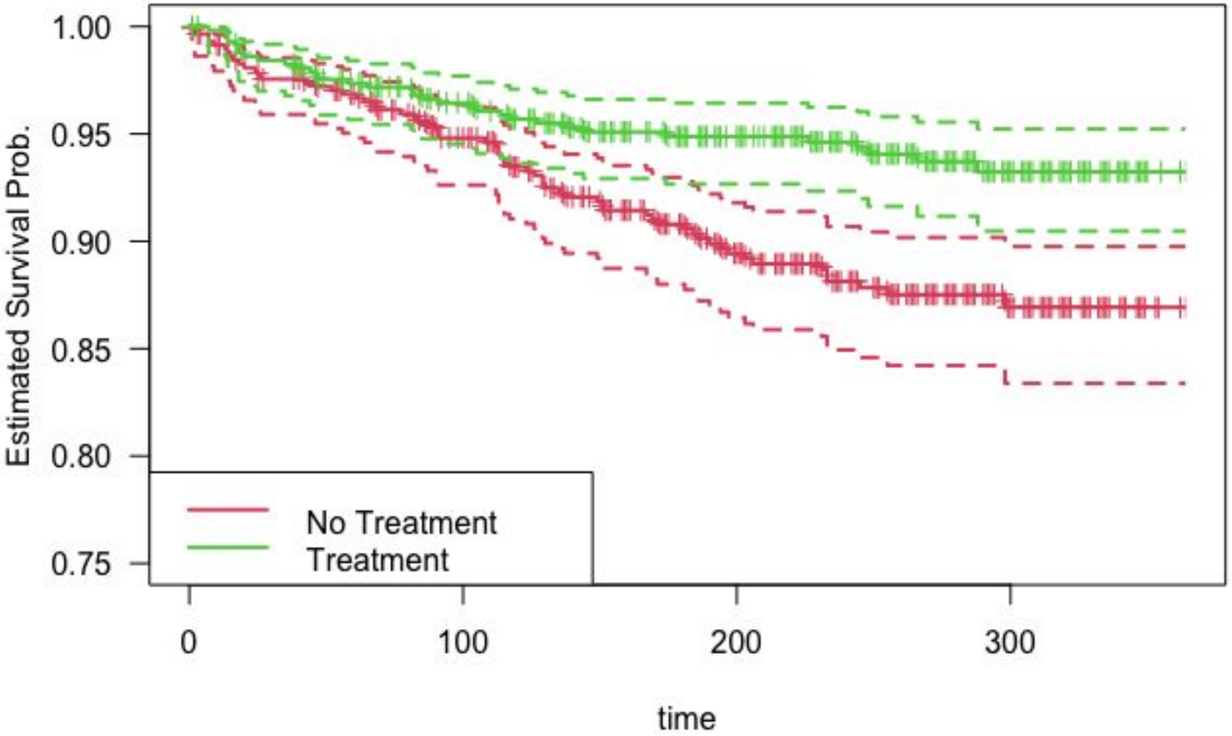
$$L(\theta, \eta) = e^{\theta x} e^{\eta(x+y)} (1 + e^{\theta+\eta})^{-m} (1 + e^{\eta})^{-n}$$

$$\widehat{\theta} = \log \frac{x/(m-x)}{y/(n-y)}. \qquad \mathrm{se}(\widehat{\theta}) = \left(\frac{1}{x} + \frac{1}{y} + \frac{1}{m-x} + \frac{1}{n-y}\right)^{1/2}$$

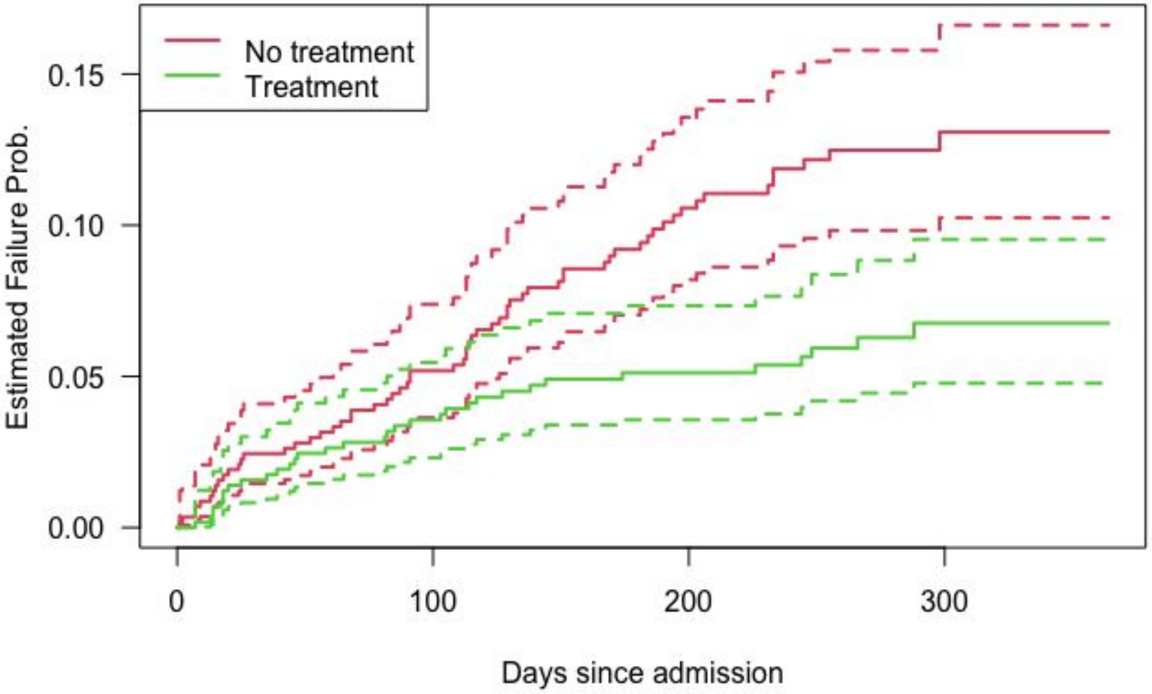
η = - 1.0360920 CI = [0.2780036 , 1.120342]
θ = - 0.7217664 CI = [7.377432 , 7.871296]

Assignment 2:
SURVIVAL and CUMULATIVE incidence

Survival



Cumulative incidence



Event = AIDS or death
Treatment = AZT

Assignment 2:
SURVIVAL COMPARISON

HYPOTHESIS TO TEST

H_0 : $S(t)_{tx}$ = $S(t)_{no\_tx}$

H_0 = Both groups survive the same, thus the treatment has no effect

Log-Rank test

$$Q = \frac{\left(\sum_{i=1}^{m} w_i(d_{1i} - \hat{e}_{1i})\right)^2}{\sum_{i=1}^{m} w_i \hat{v}_{1i}} \qquad w_i = 1,$$

```
Call:
    survdiff(formula = Surv(time, event == 1) ~ tx,
data = survival,
          rho = 1)

    N Observed Expected (O-E)^2/E (O-E)^2/V
    tx=0 577      60.1      45.1      5.02       10.3
    tx=1 574      31.7      46.8      4.84       10.3

 Chisq= 10.3  on 1 degrees of freedom,  p= 0.001
```

Reject null hypothesis

Assignment 2:
EXPONENTIAL MODEL FITTING

```
Call:
survreg(formula = Surv(time, event == 1) ~  cd4 + tx,
        data = survival, dist = "exponential")

              Value Std. Error    z        p
(Intercept) 6.71473    0.15647 42.9  < 2e-16
cd4         0.01609    0.00251  6.4  1.5e-10
tx1         0.66680    0.21489  3.1   0.0019

Scale fixed at 1

Exponential distribution

Loglik(model)= -819.9
Loglik(intercept only)= -856.6
Chisq= 73.36 on 2 degrees of freedom,  p= 1.2e-16
Number of Newton-Raphson Iterations: 7
n= 1151
```

EXPONENTIAL REGRESSION MODEL

$$S(t) = exp\left(-\frac{t}{exp\left(\beta_0 + \beta_1 x + \beta_2 x\right)}\right)$$

Confidence intervals

	2.5%	97.5%
b0	6.408	7.021
Cd4 (b1)	0.011	0.021
Tx (b2)	0.246	1.088

Assignment 2: WEIBULL MODEL FITTING

```
Call:
survreg(formula = Surv(time, event == 1) ~  cd4 + tx,
        data = survival, dist = "exponential")

                Value Std. Error    z       p
(Intercept) 6.71473    0.15647 42.9  < 2e-16
cd4         0.01609    0.00251  6.4  1.5e-10
tx1         0.66680    0.21489  3.1   0.0019

Scale fixed at 1

Exponential distribution

Loglik(model)= -819.9
Loglik(intercept only)= -856.6
Chisq= 73.36 on 2 degrees of freedom, p= 1.2e-16
Number of Newton-Raphson Iterations: 7
n= 1151
```

WEIBULL REGRESSION MODEL

$$S(t) = exp\left(-t^{1/\sigma} exp\left(-\frac{1}{\sigma} x^T \beta\right)\right)$$

Confidence intervals

	2.5%	97.5%
bo	6.563	7.552
Cd4 (b1)	0.013	0.028
Tx1 (b2)	0.27	1.4

Assignment 2:
LOG-LOGISTIC MODEL FITTING

```
Call:
survreg(formula = Surv(time, event == 1) ~  cd4 + tx,
        data = survival, dist = "loglogistic")
              Value Std. Error     z       p
(Intercept) 6.82584    0.25453 26.82 < 2e-16
cd4         0.02080    0.00375  5.55 2.9e-08
tx1         0.84295    0.28980  2.91  0.0036
Log(scale)  0.20259    0.09558  2.12  0.0340

Scale= 1.22

Log logistic distribution
Loglik(model)= -815.8
Loglik(intercept only)= -852.7
Chisq= 73.73 on 2 degrees of freedom, p= 9.8e-17
Number of Newton-Raphson Iterations: 6
n= 1151
```

LOG-LOGISTIC REGRESSION MODEL

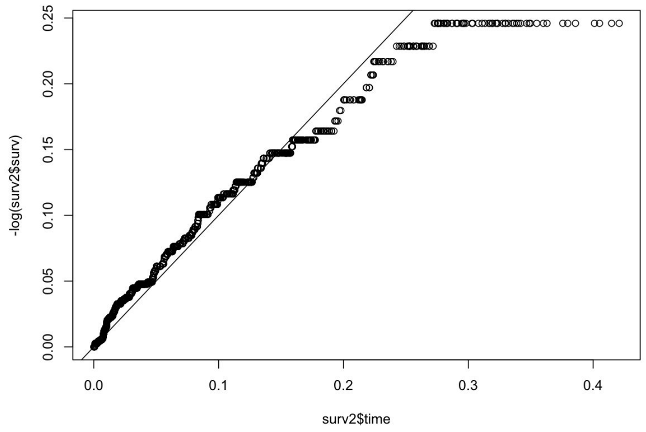
$$S(t) = \frac{1}{1 + exp\left(\frac{\log(t) - \left(\beta_0 + \beta_1 x + \beta_2 x\right)}{\sigma}\right)}$$

Confidence intervals

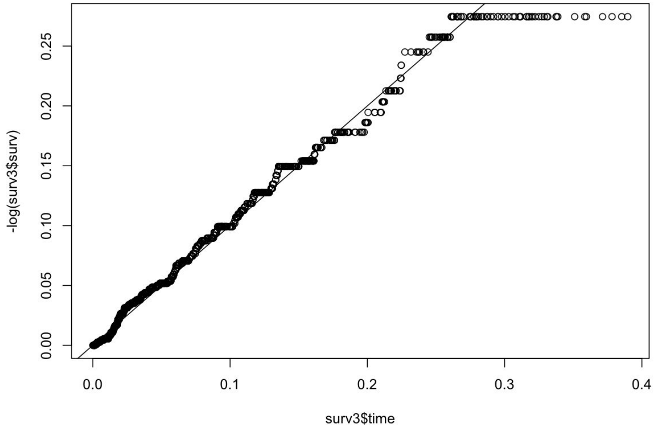
	2.5%	97.5%
bo	6.327	7.324
Cd4 (b1)	0.013	0.028
Tx1 (b2)	0.275	1.411

Assignment 2:
MODEL COMPARISON

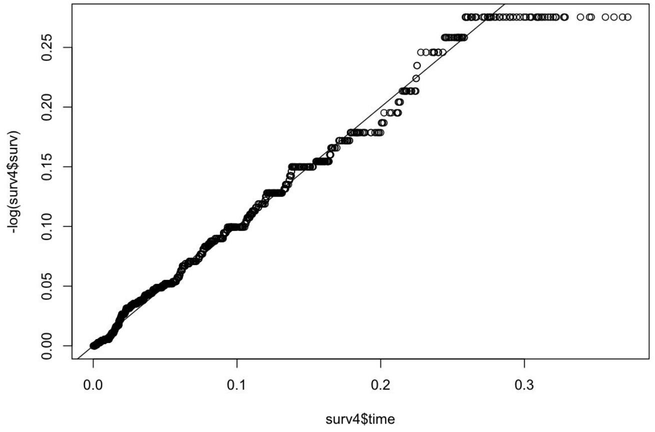
	AIC
Exponential	1645.838
Weibull	1640.671
Log-Logistic	**1639.655**



Exponential



Weibull



Log-logistic

Assignment 2:
LOG-LOGISTIC MODEL

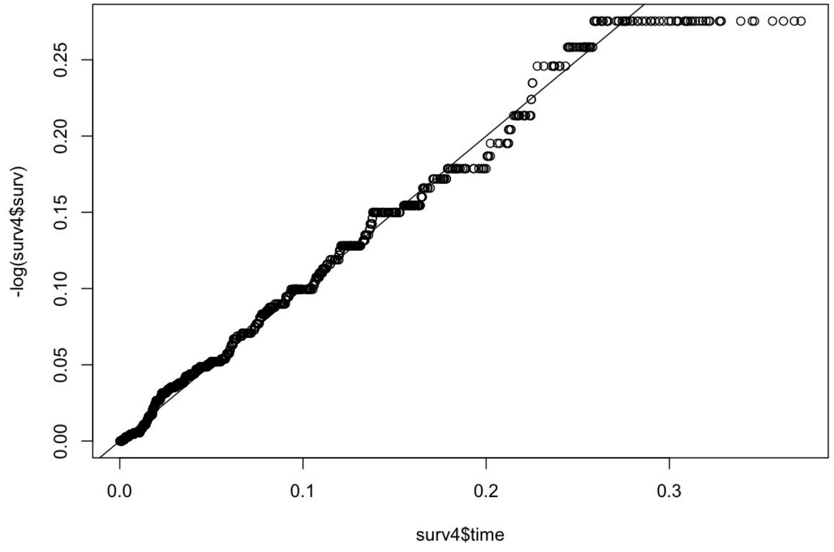
Time Ratio

	TR	2.5%	97.5%
Intercept	921.35	559.46	1517.33
cd4	1.021	1.013	1.028
tx	2.323	1.316	4.1
cd4*50	**2.829**	**1.959**	**4.086**

Hazard Ratio

	HR	2.5%	97.5%
Intercept	0.001	0.0007	0.002
cd4	0.979	0.972	0.99
tx	0.430	0.244	0.76
cd4*50	**0.353**	**0.510**	**0.244**

$$Cox\ Snell\ Residuals = r_i = -\log(S(t))$$



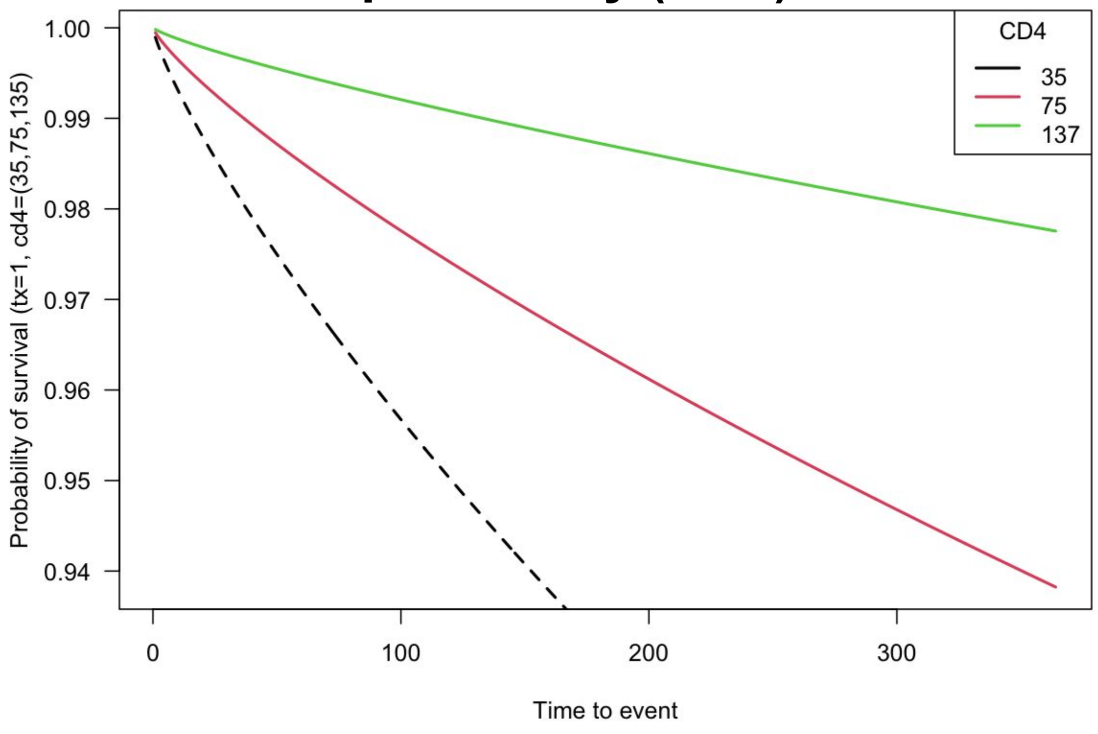
Cox-Snell Diagnostic plot

- When we increase the CD4 (cells/ml) by 50 the median survival time increases by **2.829**.

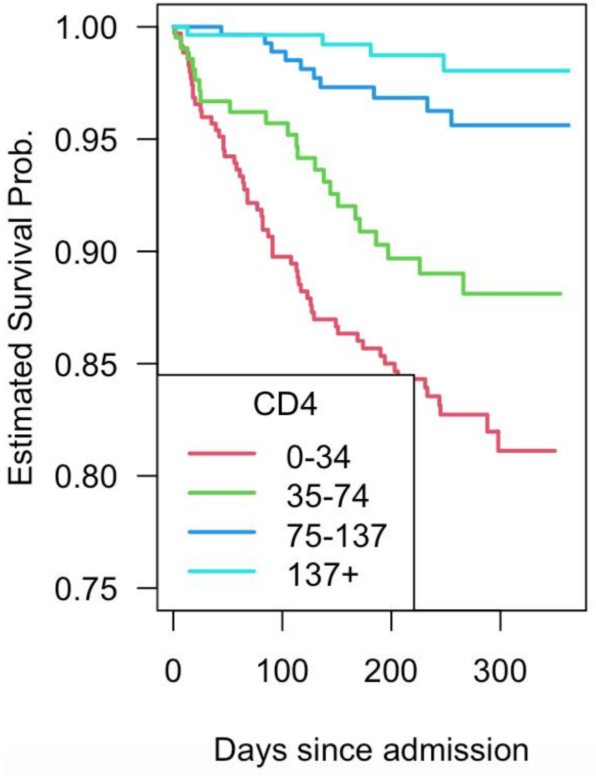
We can conclude that the more CD4 cells number is increased, the longer the patient will go without suffering an event

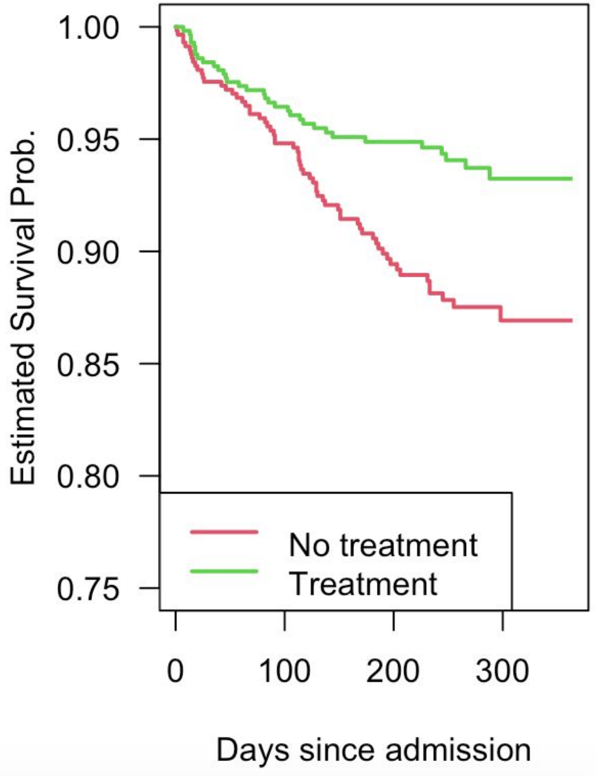
Assignment 2:
LOG-LOGISTIC MODEL

Survival probability (tx=1)



Model representation





References

Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. OUP Oxford; 2001. (Oxford science publications)

Code for the project can be found at [Statistical Modelling]

DTU