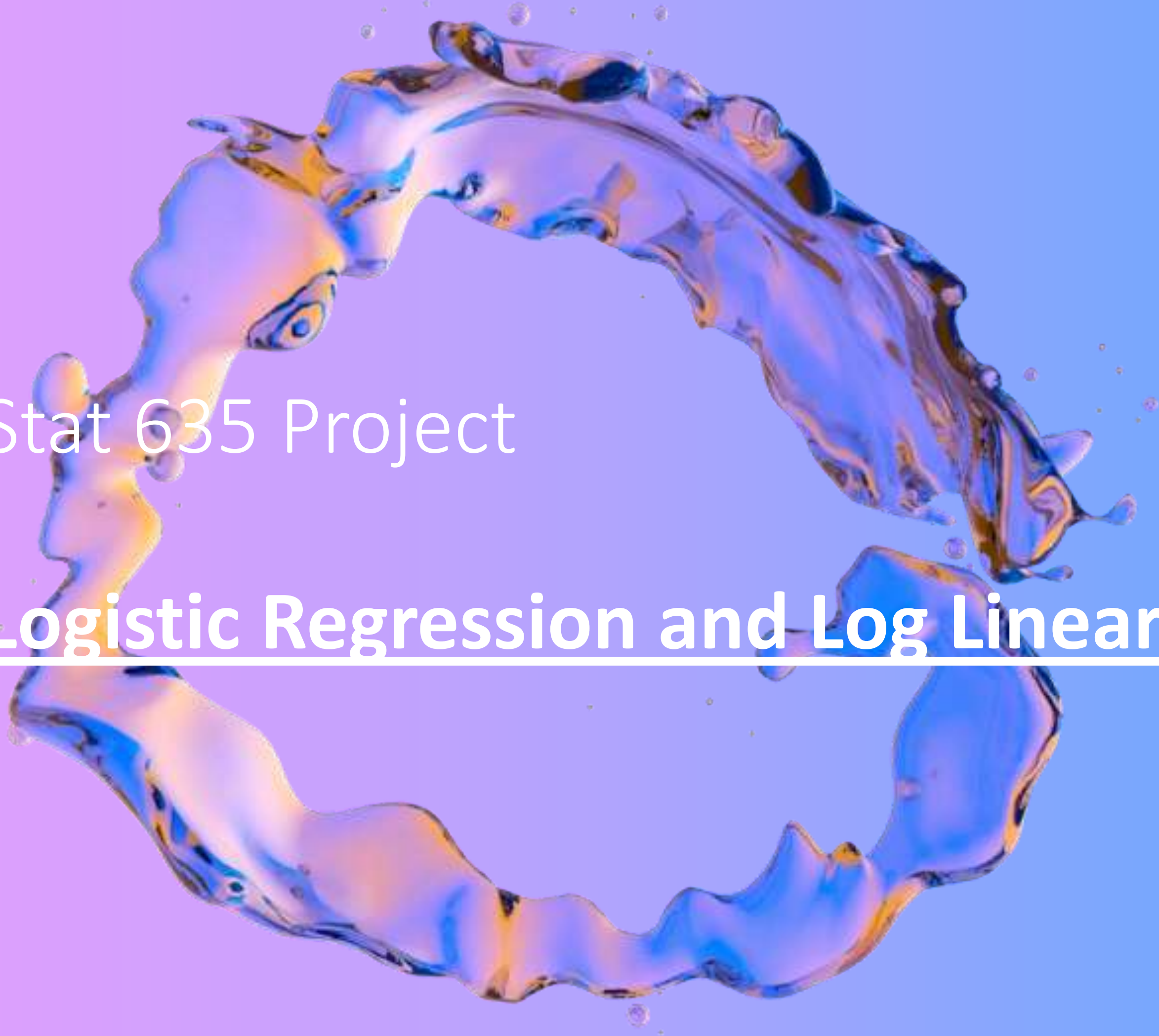
University of Calgary

2022
REPORT



Stat 635 Project

Logistic Regression and Log Linear Regression

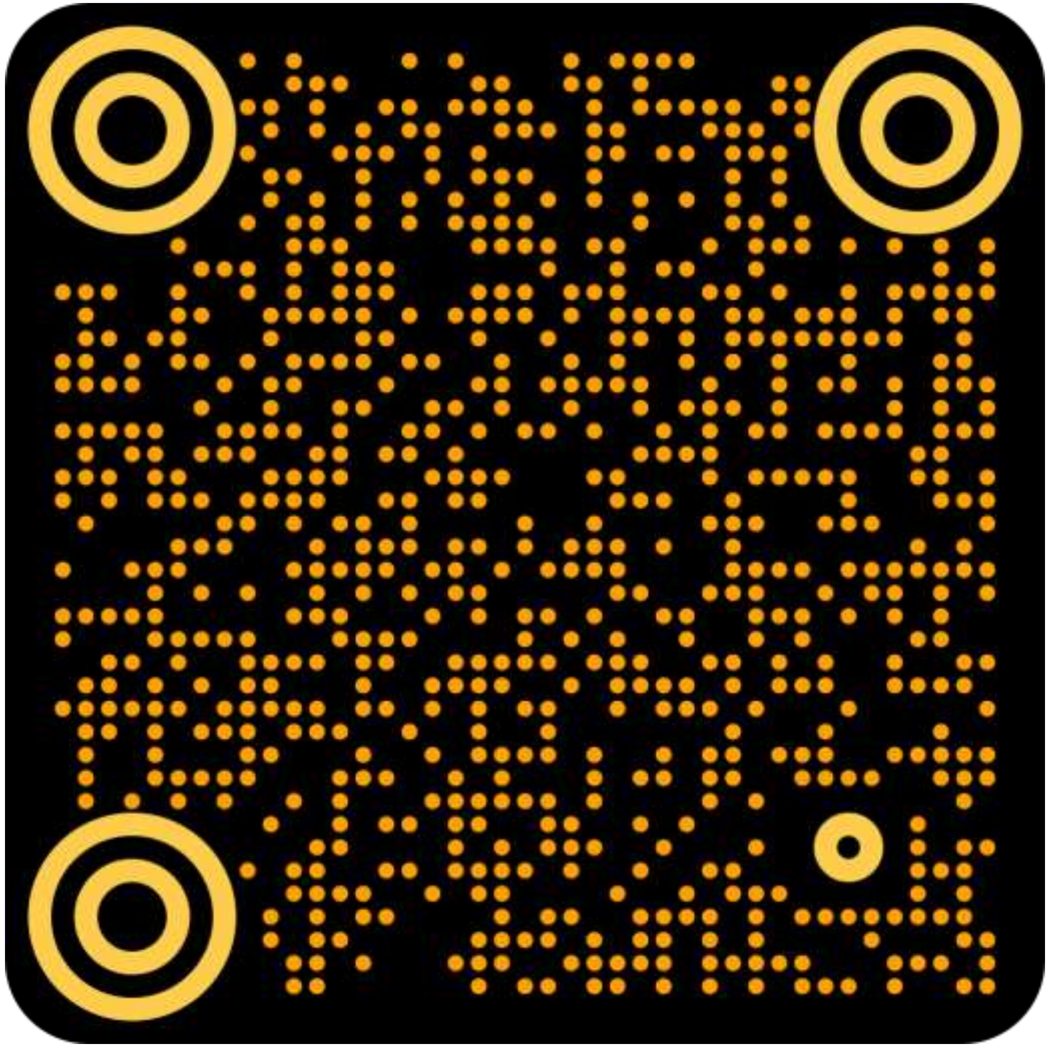
Hao Nan Wang

UCID: 30185958

Email: haonan.wang@ucalgary.ca

Monday, Dec 5th, 2022

Since this presentation contains a variety of models and estimation results, I personally suggest that you could scan this QR code, so that you could enjoy this presentation and view the PowerPoint on your personal devices in the meanwhile.



CONTENTS

Logistic Regression

Data and Preliminary Analysis

Model Comparison and Selection

Study Model

Interpretation of $f(\beta_i \ or \ \beta_i' s)$

Computation of the Estimated Value of Odds Ratio

Confidence Interval for OR

Wald-Based Hypothesis Test for β_k

Prediction

Residuals and Plots

Model Diagnostics and GOF

Issues: Over-dispersion and Outliers

Data and Preliminary Analysis

Since the start of the COVID-19 pandemic, many healthcare activities have been cancelled or delayed. Consequently, referrals of suspected new cancers have reduced, with increases in cancer-related deaths predicted. In addition to it, Northern England has been experiencing a persistent rise in the number of primary liver cancers.

The data was prospectively collected on all patients referred to the Newcastle-upon-Tyne NHS Foundation Trust (NUTH) hepatopancreatobiliary multidisciplinary team (HPB MDT) in the first 12 months of the pandemic (March 2020-February 2021), comparing to a retrospective observational cohort of consecutive patients presenting in the 12 months immediately preceding it (March 2019-February 2020).

The objective is to assess the impact of the COVID-19 pandemic on patients with newly diagnosed liver cancer and use logistic regression to predict the probability of death.

Attribute Information (27 attributes in original dataset)

1.Cancer: Cancer flag [Y/N]

2.Year: Categorical [Prepandemic (March 2019–February 2020)/Postpandemic(March 2020–February 2021)]

3.Month: Month of the year 1-12

4.Age: Age of the patitent

5.Gender: Male or Female [M/F]

6.Cirrhosis: Underlying liver disease [Y/N]

7.Size: Tumour diameter in mm

8.HCC TNM Stage: Hepatocellular carcinoma Tumour node metastasis Stage ("I", "II", "IIIA+IIIB", "IV")

9.HCC BCLC Stage: Hepatocellular carcinoma Barcelona Clinic for Liver Cancer Stage ("0", "A", "B", "C", "D")

10.ICC TNM Stage: Intrahepatic cholangiocarcinoma Tumour node metastasis Stage ("I", "II", "III", "IV")

11.Treatment grps: First-line treatment received ["OLTx" (orthotopic liver transplantation), "Resection", "Ablation", "TACE"" (transarterial chemoembolisation), "SIRT" (selective internal radiation therapy), "Medical", "Supportive care"]

12.Survival from MDM: Survival from Multidisciplinary meeting

13.Alive Dead: "Alive", "Dead"

14.Type of incidental finding: ("Primary care-routine", "Secondary care-routine", "Primary care-acute", "Secondary care-acute")

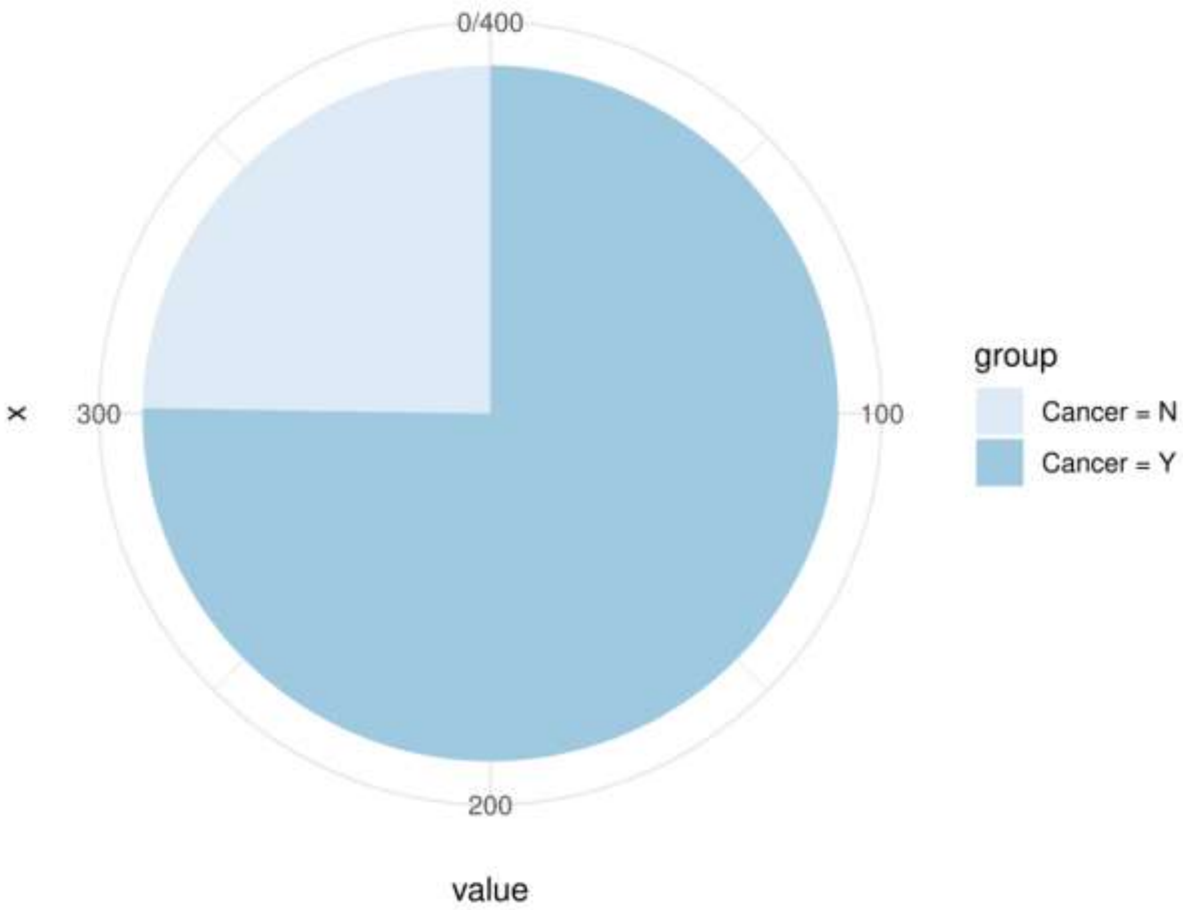
……..

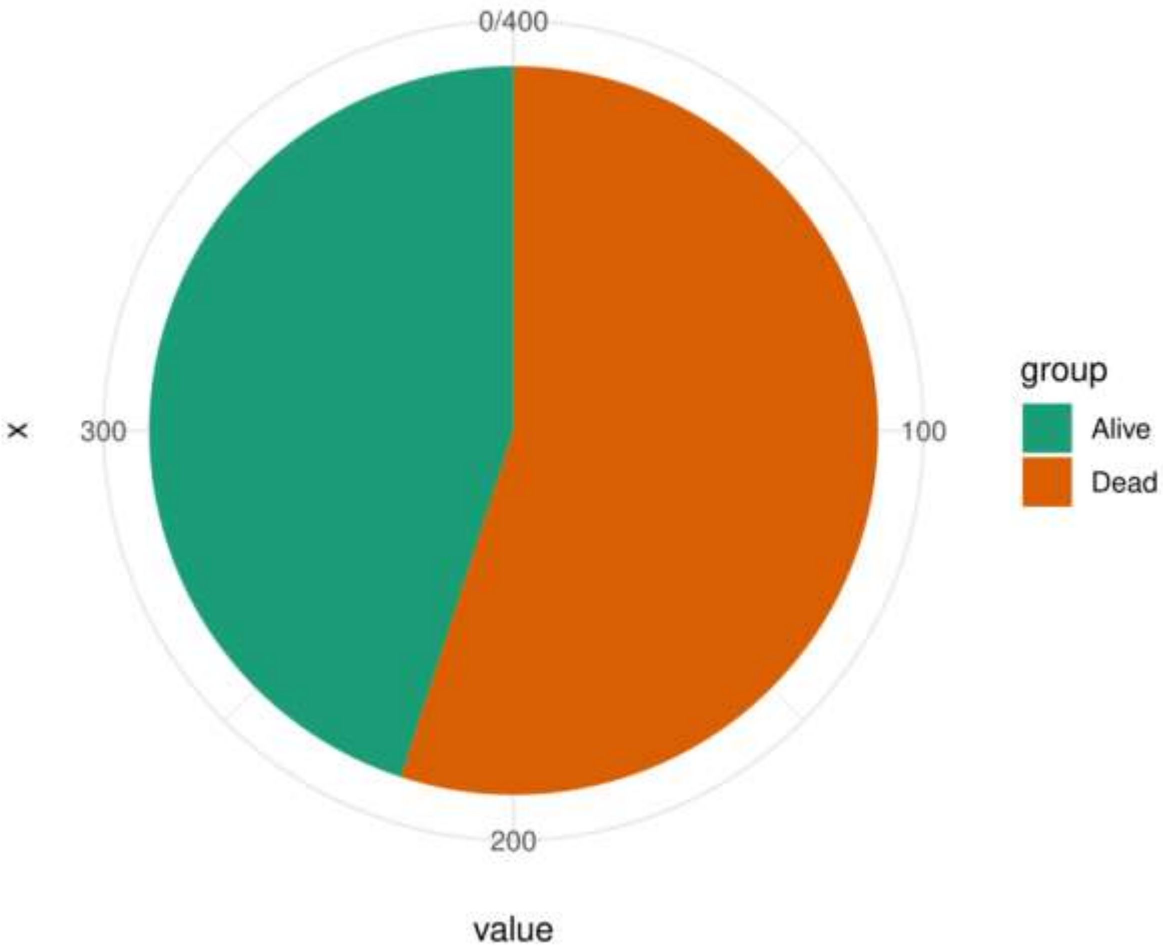
……..

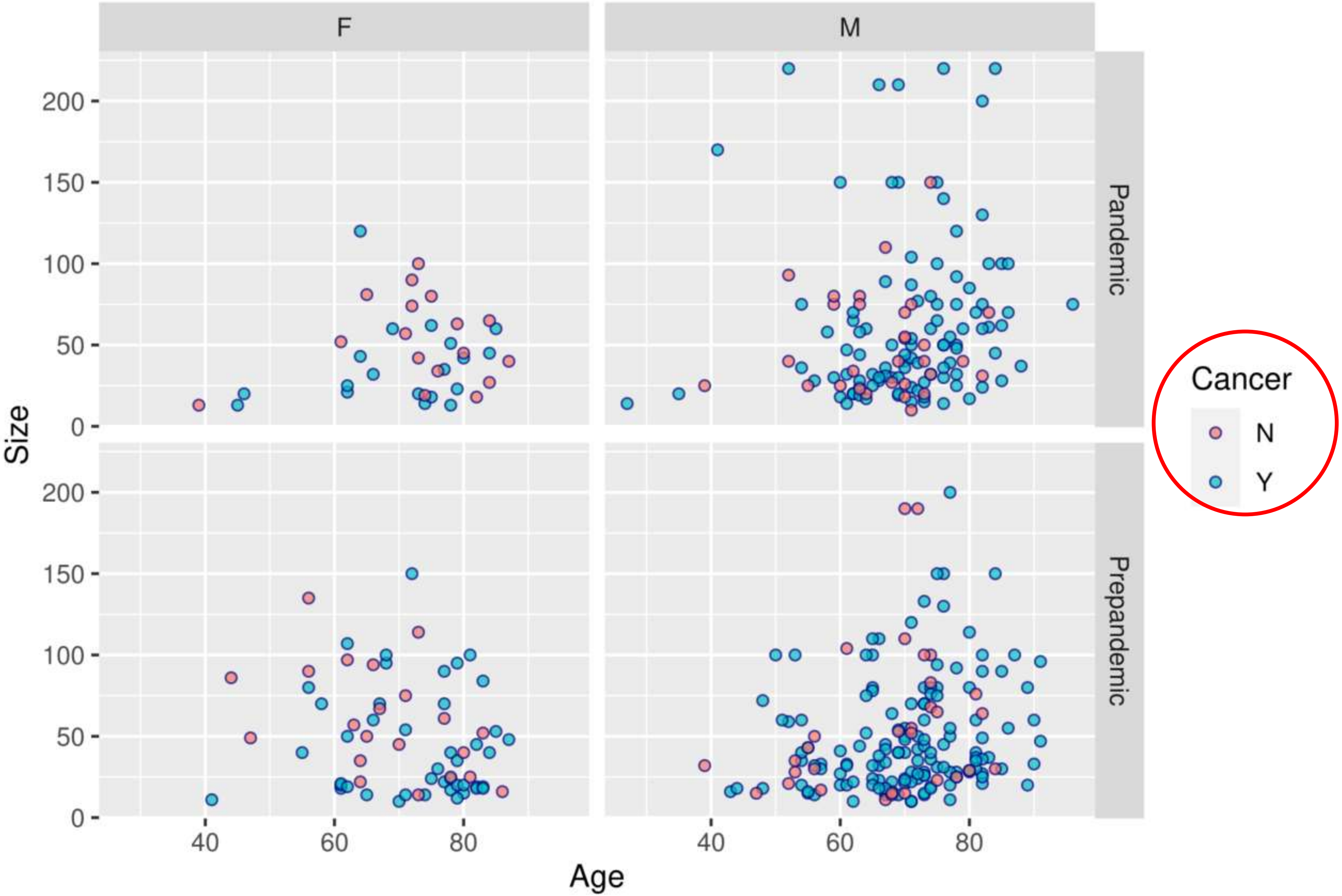
Data and Preliminary Analysis

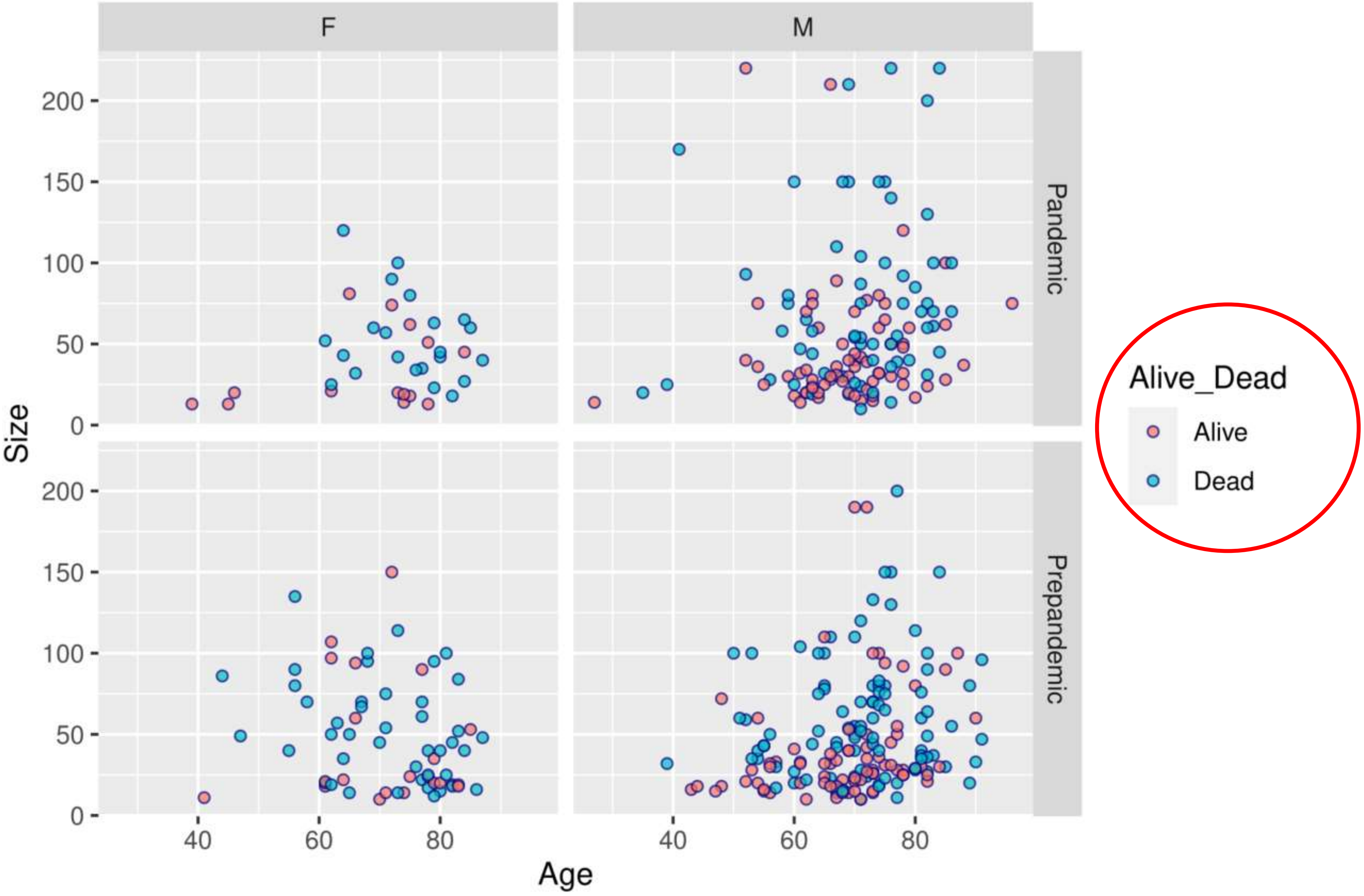
```r
data = read.csv("covid-liver.csv", header = T)
# 1: Cancer; 2: Pandemic; 6: Age; 7:Gender; 10: Tumour Size 16: Alive Dead
liver.data = data[,c(1,2,6,7,10,16)]
# Remove rows with NA's using na.omit()
## originally 450 observations, 148 are NA observations
liver.data <- na.omit(liver.data)
head(liver.data)
```

```
##   Cancer        Year Age Gender Size Alive_Dead
## 1      Y Prepandemic  68      M   22      Alive
## 2      Y Prepandemic  70      M   40       Dead
## 3      Y Prepandemic  64      M   52       Dead
## 4      Y Prepandemic  73      M   80       Dead
## 5      Y Prepandemic  66      F   60      Alive
## 6      Y Prepandemic  70      M   24       Dead
```


0/400
100
200
300
x
value
group
Cancer = N
Cancer = Y


0/400
100
200
300
x
value
group
Alive
Dead


F
M
Pandemic
Prepandemic
Size
0
50
100
150
200
Age
40
60
80
Cancer
N
Y


F
M
Pandemic
Prepandemic
Size
200
150
100
50
0
40
60
80
Age
Alive_Dead
Alive
Dead

```r
# Replace conditionally
liver.data = as.data.table(liver.data)
liver.data[Alive_Dead == "Dead", Alive_Dead := 1]
liver.data[Alive_Dead == "Alive", Alive_Dead := 0]
liver.data[Cancer == "Y", Cancer := 1]
liver.data[Cancer == "N", Cancer := 0]
liver.data[Year == "Pandemic", Year := 1]
liver.data[Year == "Prepandemic", Year := 0]
liver.data[Gender == "F", Gender := 1]
liver.data[Gender == "M", Gender := 0]
head(liver.data,1)
```

```
##     Cancer Year Age Gender Size Alive_Dead
## 1:       1    0  68      0   22          0
```

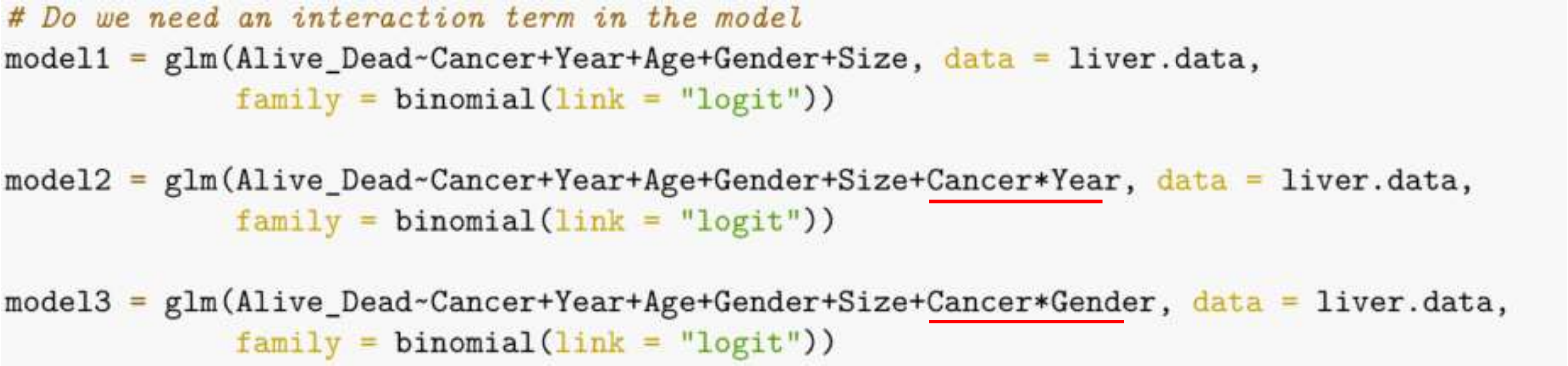
```r
liver.data$Alive_Dead = as.factor(liver.data$Alive_Dead)
liver.data$Cancer = as.factor(liver.data$Cancer)
liver.data$Year = as.factor(liver.data$Year)
liver.data$Gender = as.factor(liver.data$Gender)
```

Model Comparison and Selection

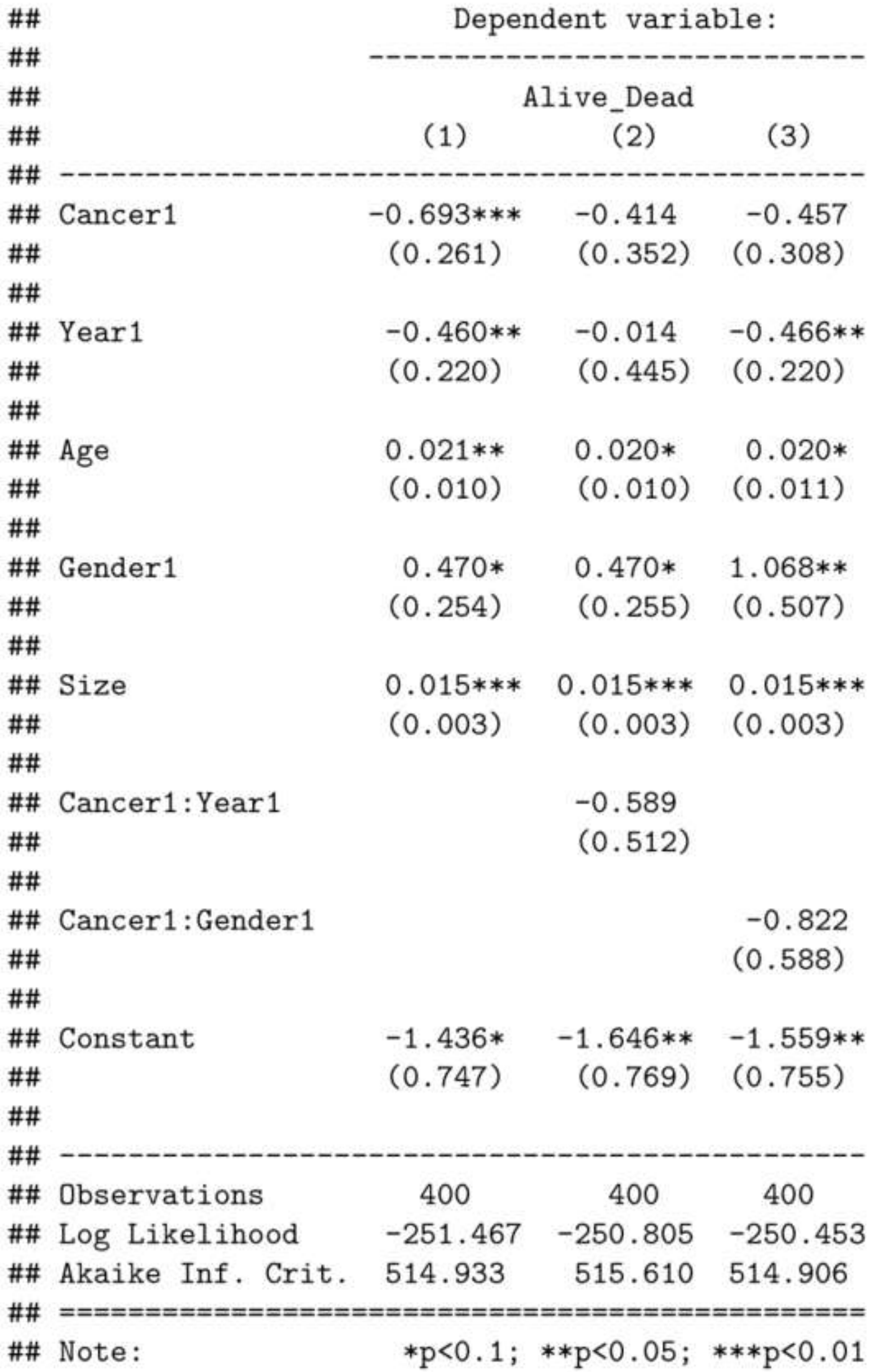
Do we need to include interaction terms in the model?

Why? When? How?

1. They have large main effects.
2. The interaction has been proven in previous studies.
3. You want to test some new hypotheses.



```
# Do we need an interaction term in the model
model1 = glm(Alive_Dead~Cancer+Year+Age+Gender+Size, data = liver.data,
             family = binomial(link = "logit"))

model2 = glm(Alive_Dead~Cancer+Year+Age+Gender+Size+Cancer*Year, data = liver.data,
             family = binomial(link = "logit"))

model3 = glm(Alive_Dead~Cancer+Year+Age+Gender+Size+Cancer*Gender, data = liver.data,
             family = binomial(link = "logit"))
```

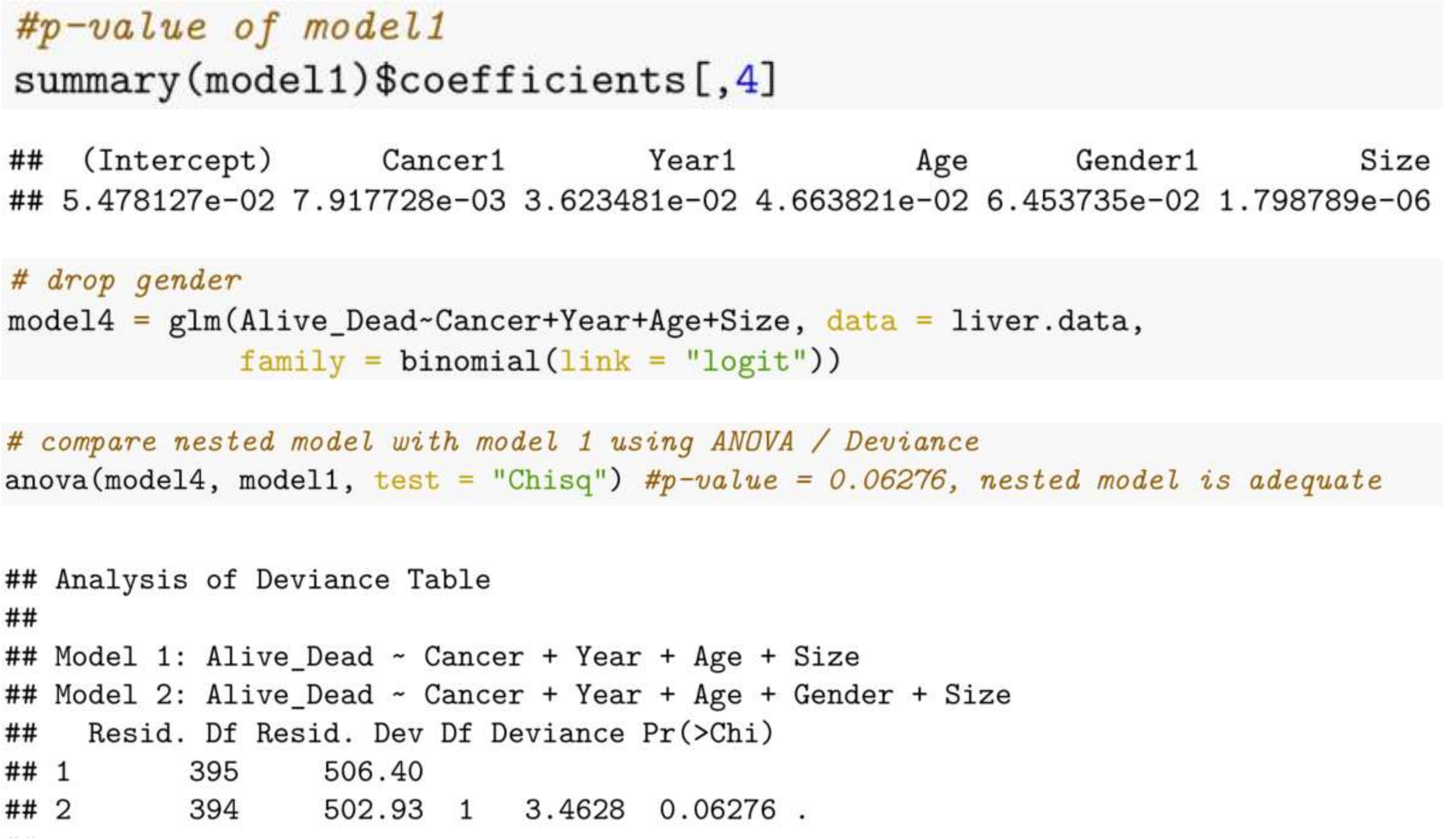


```
##                        Dependent variable:
##                    -----------------------------
##                             Alive_Dead
##                       (1)        (2)       (3)
## ---------------------------------------------------
## Cancer1            -0.693***   -0.414    -0.457
##                     (0.261)   (0.352)   (0.308)
##
## Year1               -0.460**   -0.014   -0.466**
##                     (0.220)   (0.445)   (0.220)
##
## Age                  0.021**   0.020*    0.020*
##                     (0.010)   (0.010)   (0.011)
##
## Gender1              0.470*    0.470*   1.068**
##                     (0.254)   (0.255)   (0.507)
##
## Size                0.015***  0.015***  0.015***
##                     (0.003)   (0.003)   (0.003)
##
## Cancer1:Year1                  -0.589
##                               (0.512)
##
## Cancer1:Gender1                          -0.822
##                                         (0.588)
##
## Constant             -1.436*  -1.646**  -1.559**
##                     (0.747)   (0.769)   (0.755)
##
## ---------------------------------------------------
## Observations          400       400       400
## Log Likelihood      -251.467  -250.805  -250.453
## Akaike Inf. Crit.   514.933   515.610   514.906
## ===================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

Do we need to include interaction terms in the model?

Is every variable individually significant in the model?

Is the proposed model adequate, in compared with early proposed model?



```
#p-value of model1
summary(model1)$coefficients[,4]
```

```
##   (Intercept)       Cancer1         Year1           Age       Gender1          Size
## 5.478127e-02 7.917728e-03 3.623481e-02 4.663821e-02 6.453735e-02 1.798789e-06
```

```
# drop gender
model4 = glm(Alive_Dead~Cancer+Year+Age+Size, data = liver.data,
             family = binomial(link = "logit"))
```

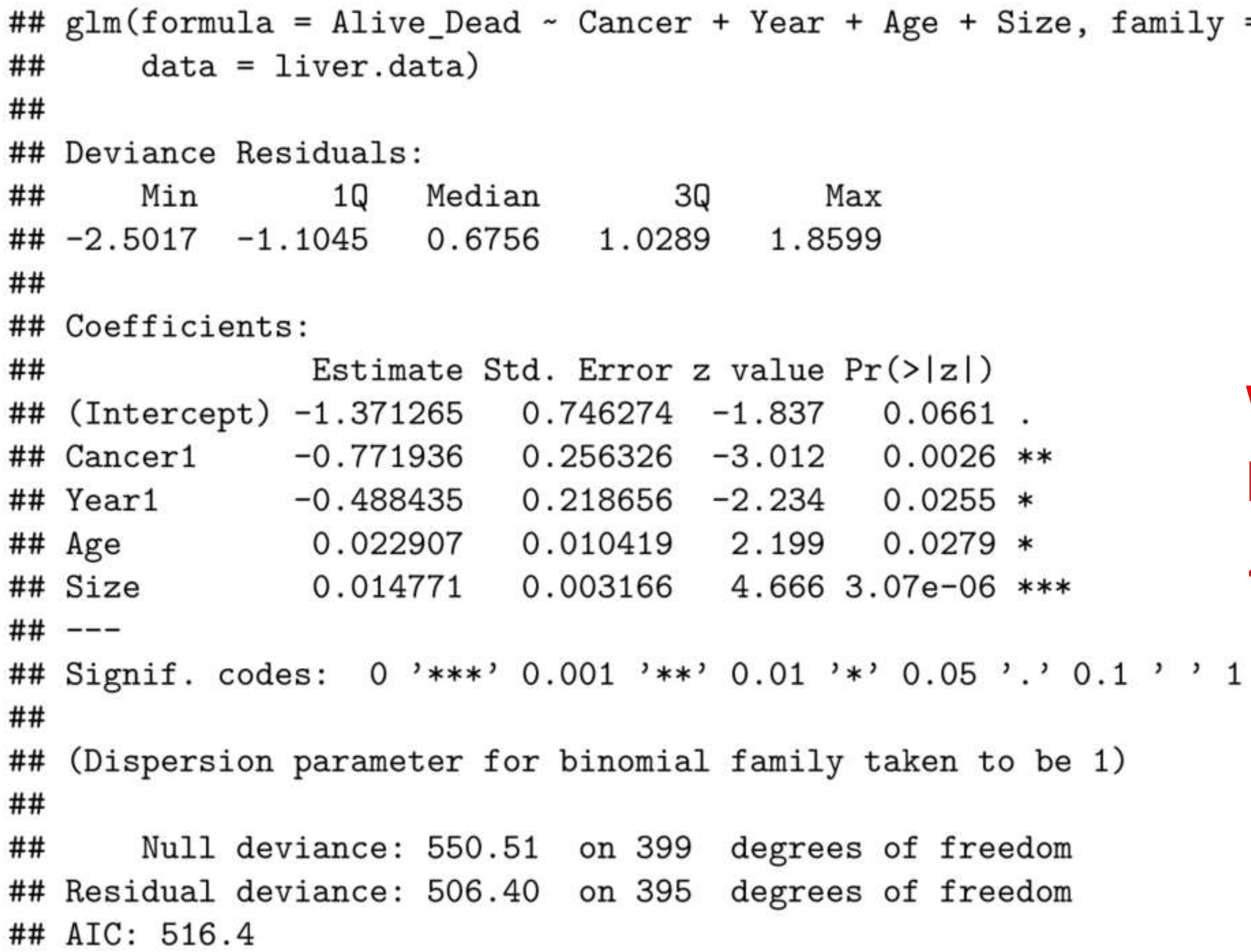
```
# compare nested model with model 1 using ANOVA / Deviance
anova(model4, model1, test = "Chisq") #p-value = 0.06276, nested model is adequate
```

```
## Analysis of Deviance Table
##
## Model 1: Alive_Dead ~ Cancer + Year + Age + Size
## Model 2: Alive_Dead ~ Cancer + Year + Age + Gender + Size
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       395     506.40
## 2       394     502.93  1   3.4628  0.06276 .
```

Study Model

$$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \text{ for i} = 1,2,...,400$$

where π_i is the death probability, $x_{i1} = I[\text{Cancer}]$, $x_{i2} = I[\text{Year}]$, x_{i3} is Age, x_{i4} is Size.



```
## glm(formula = Alive_Dead ~ Cancer + Year + Age + Size, family = binomial(link = "logit"),
##     data = liver.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5017  -1.1045   0.6756   1.0289   1.8599
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.371265   0.746274  -1.837   0.0661 .
## Cancer1     -0.771936   0.256326  -3.012   0.0026 **
## Year1       -0.488435   0.218656  -2.234   0.0255 *
## Age          0.022907   0.010419   2.199   0.0279 *
## Size         0.014771   0.003166   4.666 3.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 550.51  on 399  degrees of freedom
## Residual deviance: 506.40  on 395  degrees of freedom
## AIC: 516.4
```

What is the interpretation of β_i?

How do we do hypothesis test?

…….

Interpretation of Estimated Value of βi

Interpretation of Estimated Value of Odds Ratio

Confidence Interval for OR

Hypothesis Test for βi

Prediction

Residuals and Plots

Model Diagnostics and GOF

Issues: Overdispersion and Outliers

1. Interpretation of $f(\beta_i$ or $\beta_i's)$

```
summary(model4)$coefficients[,1]
```

```
## (Intercept)     Cancer1       Year1         Age        Size
## -1.37126463 -0.77193590 -0.48843458  0.02290729  0.01477093
```

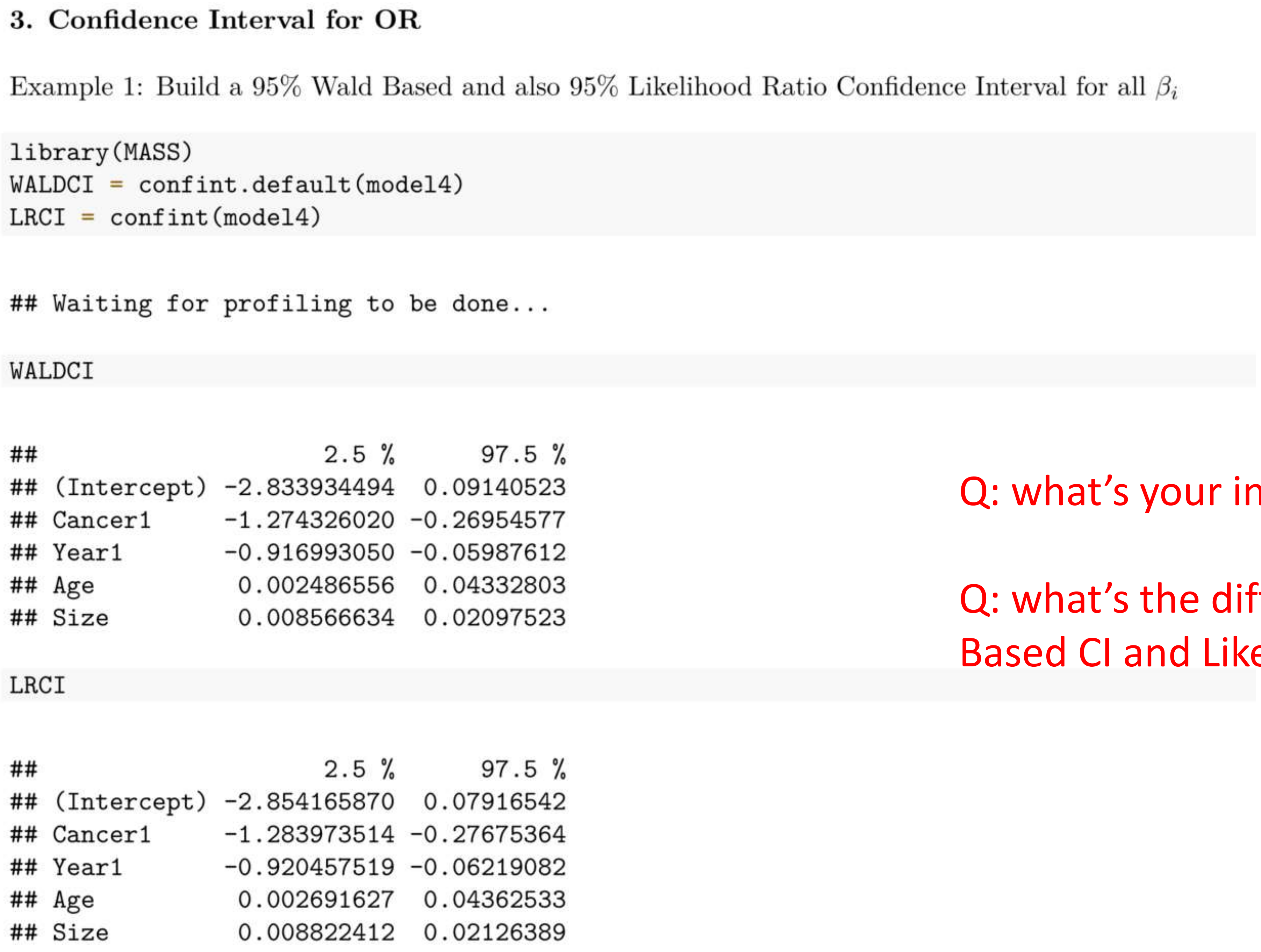
Example 1: β_1 : log odds ratio of death with Cancer versus Non-Cancer, holding all other covariates at a fixed value. **(with dichotomous predictor variable)**

Example 2: β_3 : for a one-unit increase in Age, the expected change in log odds of being dead, holding all other covariates at a fixed value. **(with continuous predictor variable)**

2. Compute the Estimated Value of Odds Ratio

Example 1: the estimated odds ratio of death with Cancer versus Non-Cancer, holding all other covariates at a fixed value, is $exp(\hat{\beta_1}) = exp(-0.7719) = 0.46$, represents the odds of death for Cancer are about 54% less than the odds for Non-Cancer, holding other variable at a fixed value.

Example 2: for a one-unit increase in Age, the expected change in the odds of being dead, holding all other covariates at a fixed value, is $exp(\hat{\beta_3}) = exp(0.0229) = 1.023$, represents for a one-unit in increase in Age, it is expected to see about 2.3% increase in the odds of being dead, holding other at a fixed level.



3. Confidence Interval for OR

Example 1: Build a 95% Wald Based and also 95% Likelihood Ratio Confidence Interval for all β_i

```
library(MASS)
WALDCI = confint.default(model4)
LRCI = confint(model4)
```

```
## Waiting for profiling to be done...
```

```
WALDCI
```

```
##                    2.5 %      97.5 %
## (Intercept) -2.833934494  0.09140523
## Cancer1     -1.274326020 -0.26954577
## Year1       -0.916993050 -0.05987612
## Age          0.002486556  0.04332803
## Size         0.008566634  0.02097523
```

```
LRCI
```

```
##                    2.5 %      97.5 %
## (Intercept) -2.854165870  0.07916542
## Cancer1     -1.283973514 -0.27675364
## Year1       -0.920457519 -0.06219082
## Age          0.002691627  0.04362533
## Size         0.008822412  0.02126389
```

Q: what’s your interpretation?

Q: what’s the difference between Wald Based CI and Likelihood Ratio CI?

Example 2: Without using any R packages, build a 95% Wald Based Confidence Interval for β_1 and corresponding odds ratio ψ

For 95% CI for β_1: $\hat{\beta}_1 \pm 1.96 \times se(\hat{\beta}_1) = [L, U]$

```
# beta CI
c(summary(model4)$coefficients[2,1]-1.96*summary(model4)$coefficients[2,2],
  summary(model4)$coefficients[2,1]+1.96*summary(model4)$coefficients[2,2])
```

```
## [1] -1.2743353 -0.2695365
```

For 95% CI for the Odds Ratio ψ : exp[L,U]

```
# OR CI
exp(c(-1.2743353, -0.2695365))
```

```
## [1] 0.2796168 0.7637334
```

We are 95% confident that the true odds ratio between Cancer and Non-Cancer is contained in this interval.

Since this confidence interval does not contain the value 1, it is statistically significant. This should make sense because this CI excluding unity (1), which highlights β_1 is significantly different from zero.

Question: Build a 95% Confidence Interval for $\beta_1 + \beta_2$ and corresponding odds ratio ψ

Answer: For 95% CI for $\beta_1 + \beta_2$: $(\widehat{\beta_1 + \beta_2}) \pm 1.96 \times se(\widehat{\beta_1 + \beta_2}) = [L, U]$ and similarly, 95% CI for OR is exp[L,U]

note: $se(\widehat{\beta_1 + \beta_2}) = \sqrt{Var(\hat{\beta}_1 + \hat{\beta}_2)} = \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) + 2 \cdot Cov(\hat{\beta}_1, \hat{\beta}_2)}$

4. Wald-Based Hypothesis Test for β_k

Determine if *Year* (indicator variable to show whether pandemic or not) is associated with the risk of death.

Step1: $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$

Step2: State Test Statistics and Compute Its Value $z = \frac{\hat{\beta_2} - 0}{se(\hat{\beta_2})} = \frac{-0.488435}{0.218656} = -2.234$

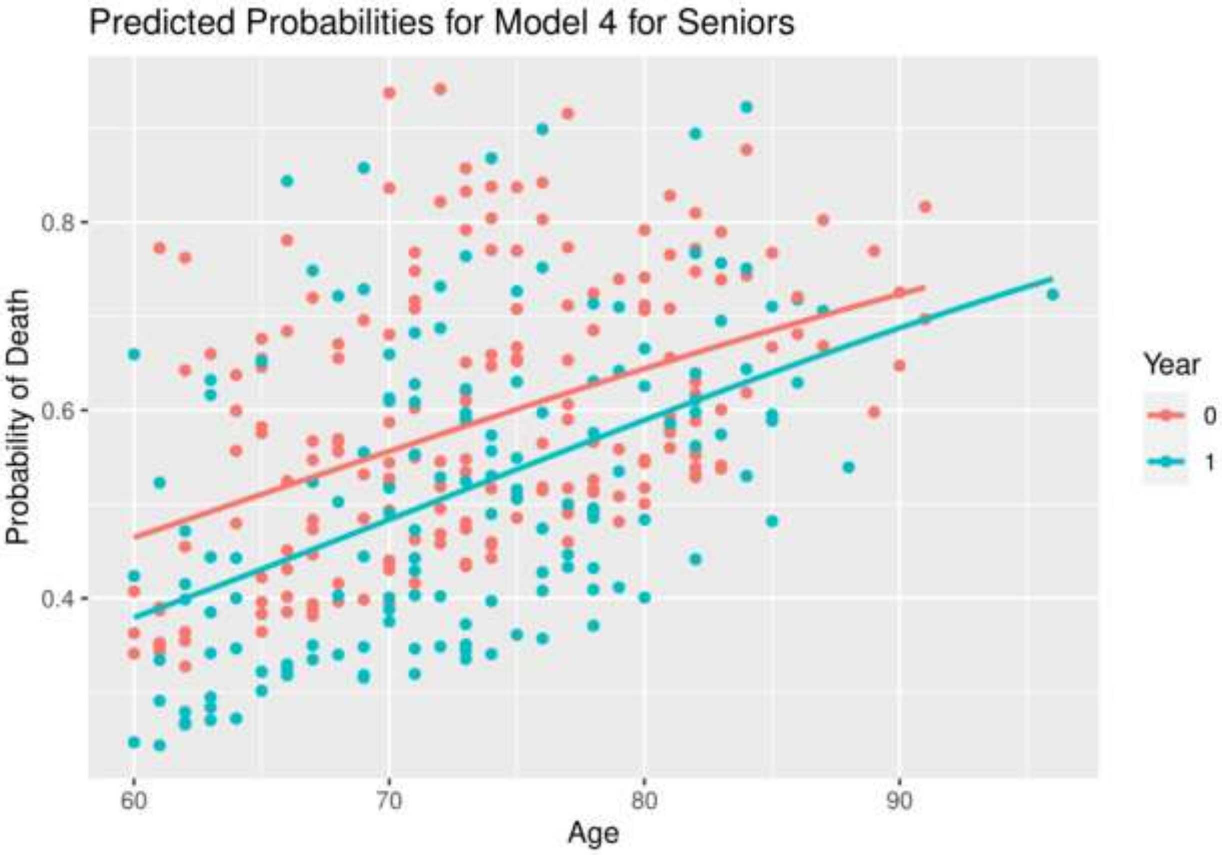
Step3: Compute p-value $2P(Z > |-2.234|) = 0.0255 < \alpha = 0.05$

Step4: Conclusion: We reject the null hypothesis, which underlines the fact that *Year* is statistically significant @ 95% significance level.

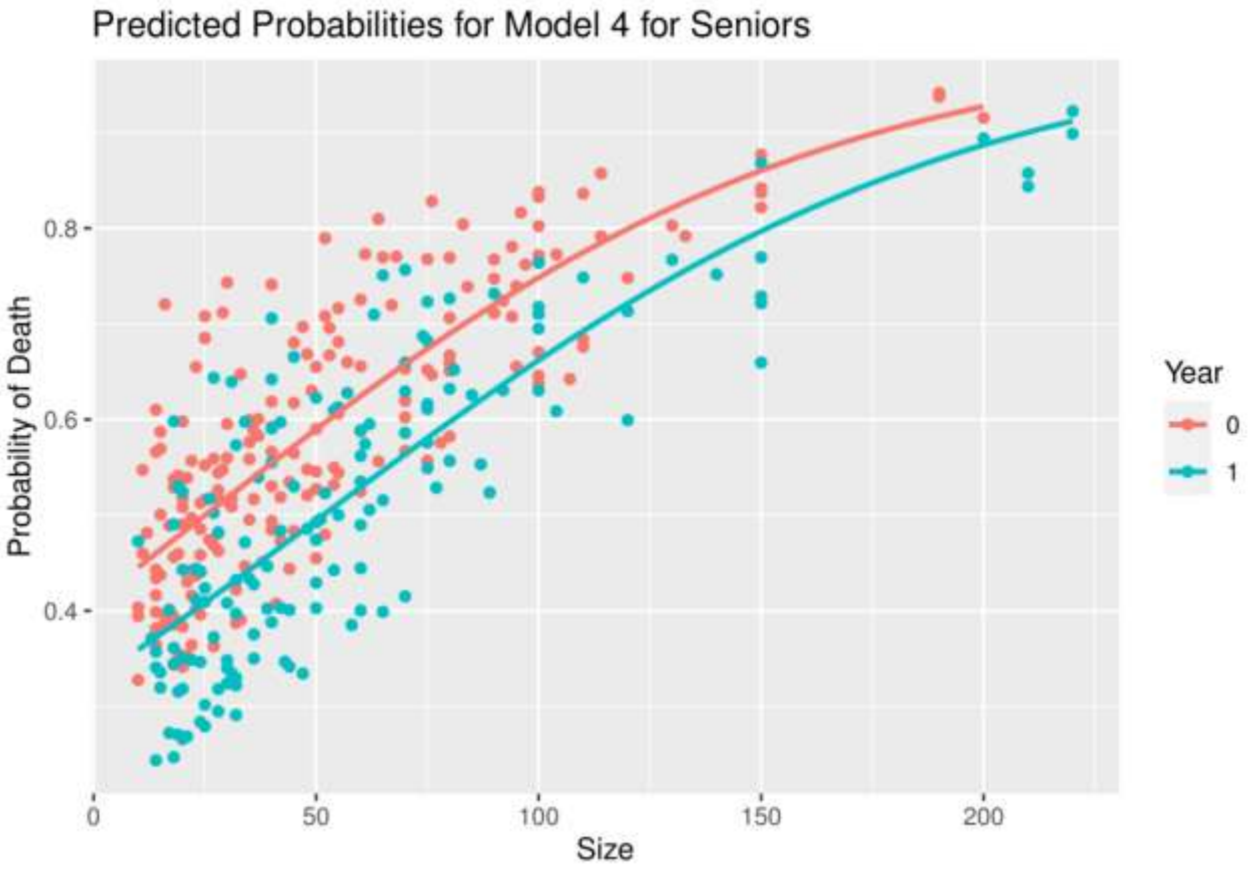
5. Prediction

One should know that in R language, the default binomial model with the default predictions are of log-odds (probabilities on logit scale) and change **type = "response"** gives the predicted probabilities.

Using the proposed study model, draw the predicted death probability $\hat{p}$ for seniors (age>=60) versus their age and draw another plot on predicted death probability versus the size of tumor, color the predicted line with different colors such that pandemic in blue and pre-pandemic in red.



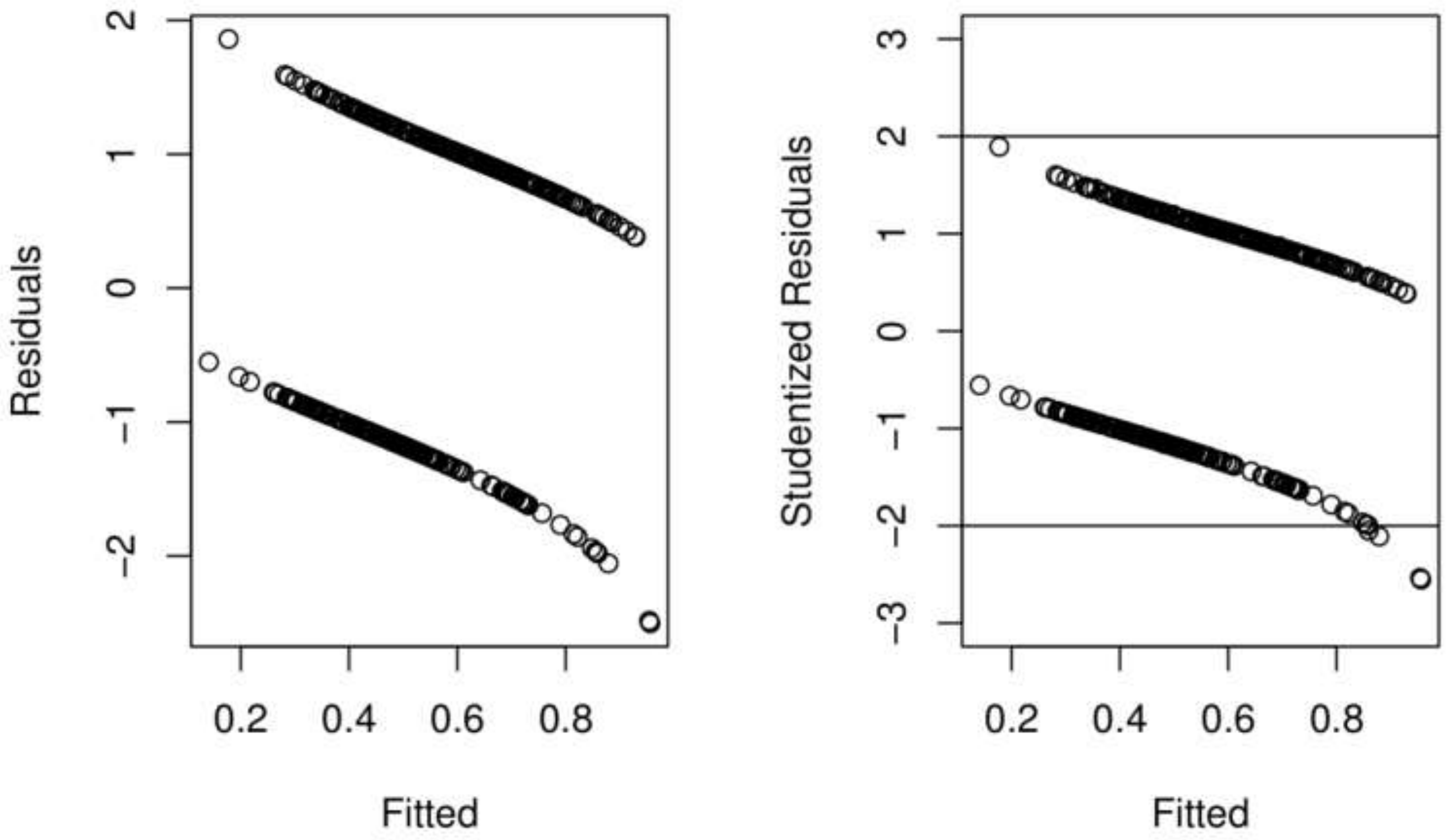
Predicted Probabilities for Model 4 for Seniors

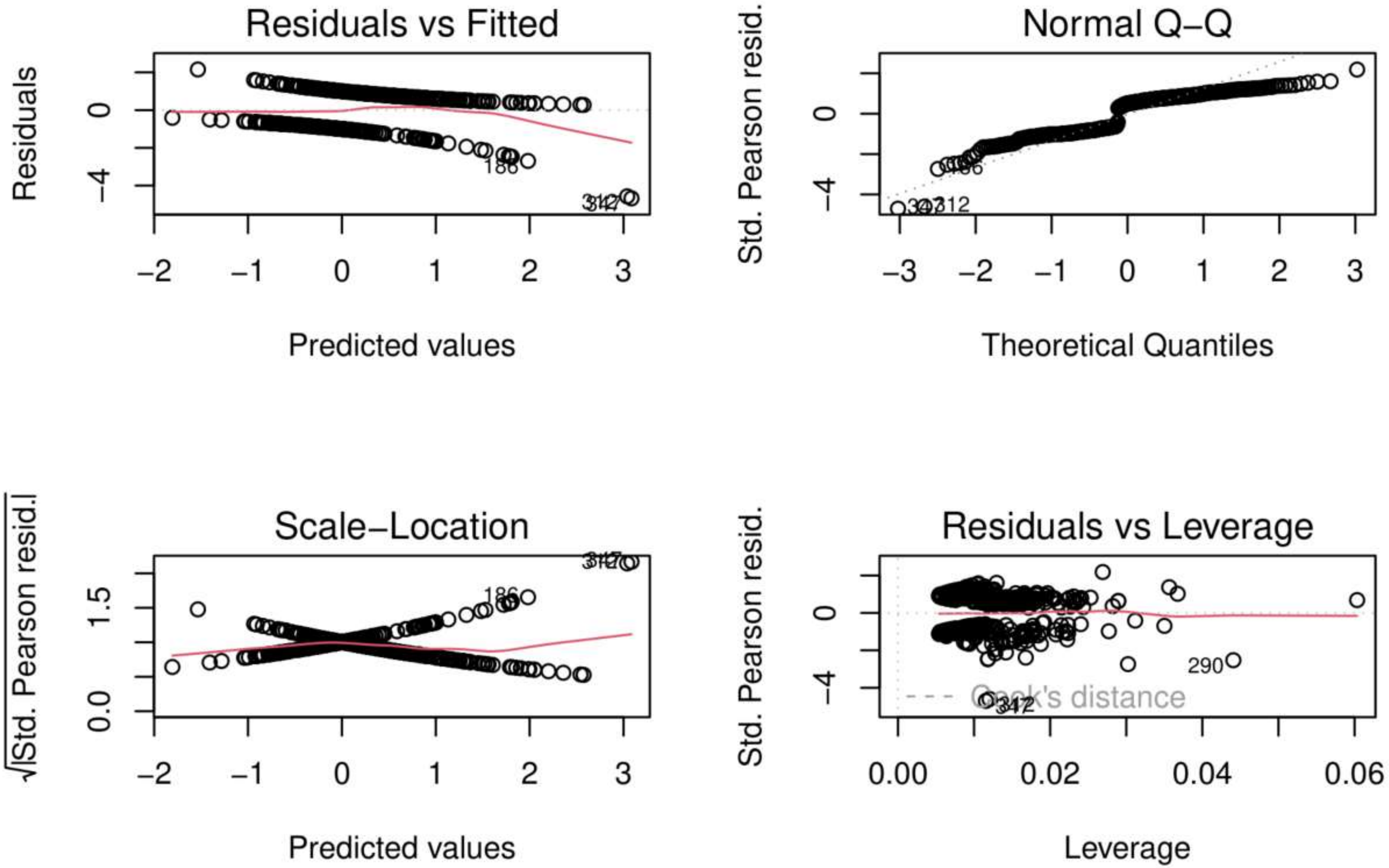


Predicted Probabilities for Model 4 for Seniors

6. Residuals and Plots

```
par(mfrow = c(1,2))
plot(fitted(model4), residuals(model4), xlab = "Fitted", ylab = "Residuals")
plot(fitted(model4), rstudent(model4),ylim = c(-3,3), xlab = "Fitted", ylab = "Studentized Residuals")
abline(2,0)
abline(-2,0)
```



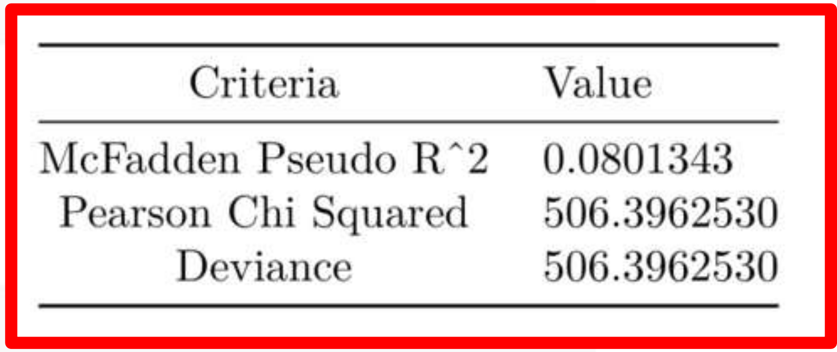

Residuals vs Fitted
Residuals
0
−4
186
342
317
−2
−1
0
1
2
3
Predicted values
Normal Q−Q
Std. Pearson resid.
0
−4
186
342
312
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Scale−Location
√|Std. Pearson resid.|
1.5
0.0
317
186
−2
−1
0
1
2
3
Predicted values
Residuals vs Leverage
Std. Pearson resid.
0
−4
290
Cook's distance
342
0.00
0.02
0.04
0.06
Leverage

7. Model Diagnostics and GOF

Use the McFadden's pseudo-R^2, the Pearson chi-squared goodness of fit statistic X^2 and the deviance D of the fitted model

$$\text{Pseudo-}R^2 = \frac{Null.Dev - Resid.Dev}{Null.Dev} = \frac{Dev(b_{min}) - Dev(b)}{Dev(b_{min})} = \frac{l(b_{min}) - l(b))}{l(b_{min}) - l(b_{max}))}$$

```
dev_b = deviance(model4) #Dev(b)
model4_null = glm(Alive_Dead~1, data = liver.data, family = binomial)
dev_n = deviance(model4_null) #Dev(bmin)
Pseudo_R_squared = (dev_n-dev_b)/dev_n
```



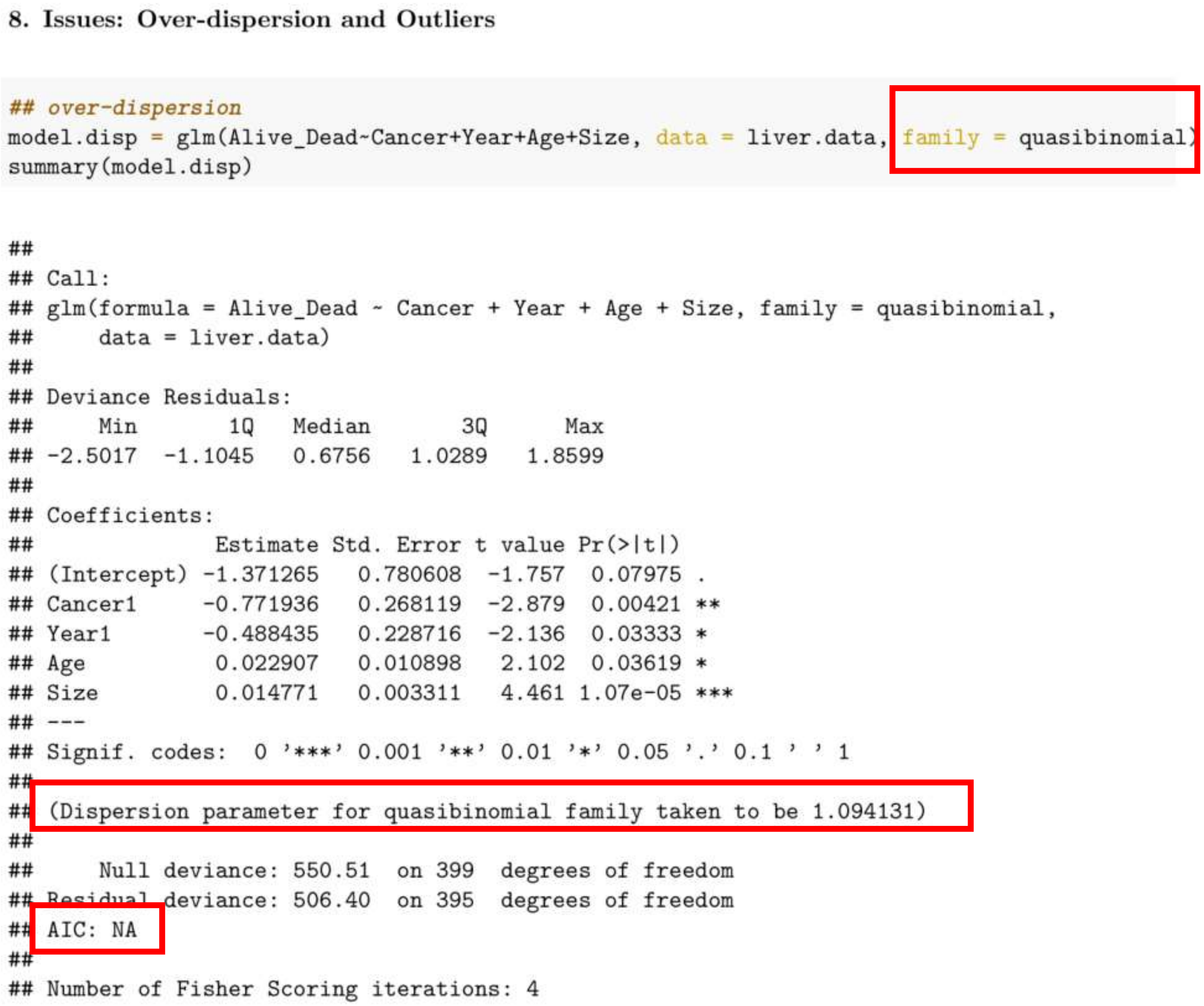
Criteria	Value
McFadden Pseudo R^2	0.0801343
Pearson Chi Squared	506.3962530
Deviance	506.3962530

$$\text{Pearson's } \chi^2 = X^2 = \sum (\text{Pearson Residuals})^2$$

```
pearson = sum(residuals(model4, type = "pearson")^2)
```

$$\text{Deviance } D = 2\Big(l(b_{max}) - l(b)\Big)$$

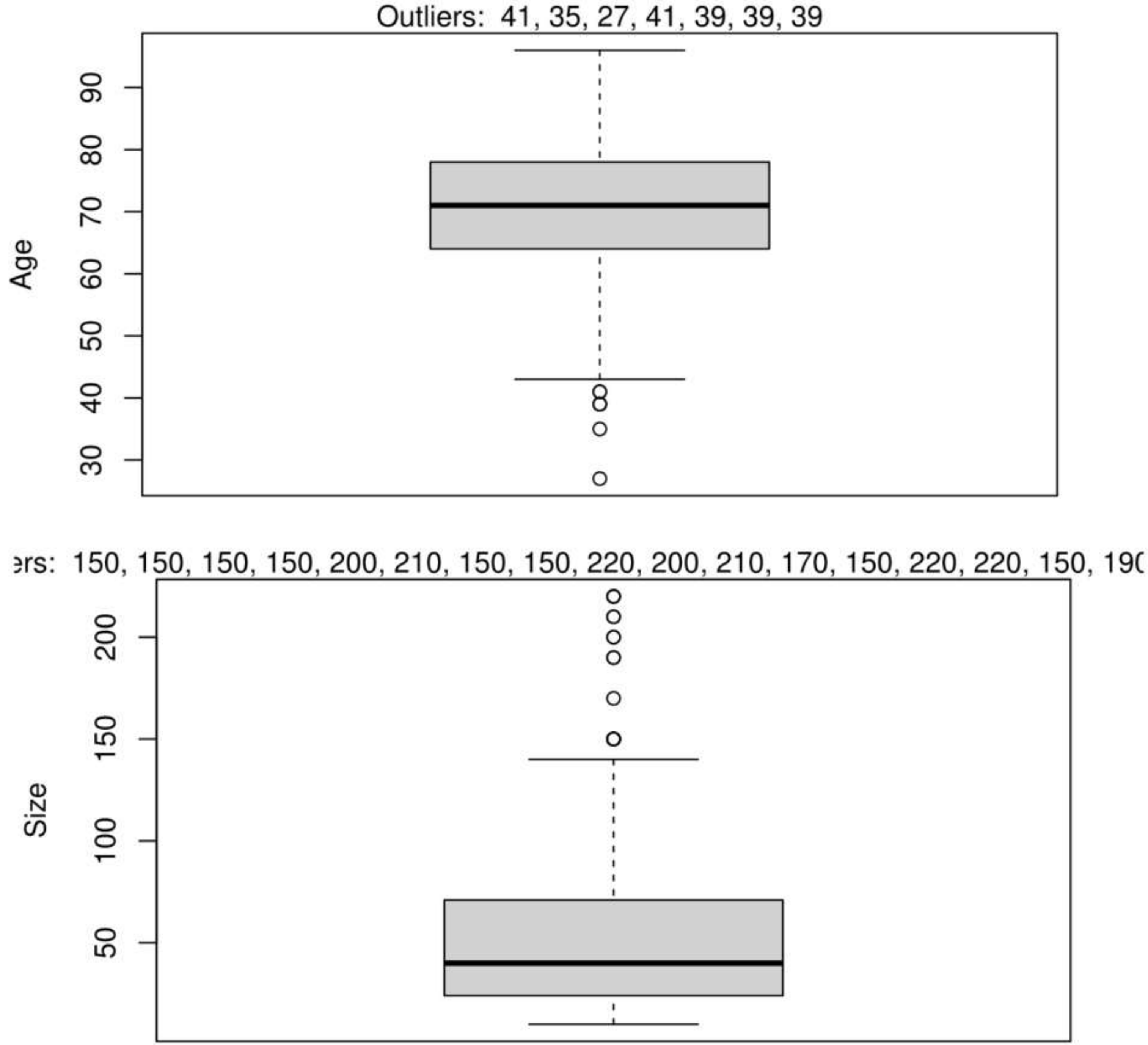
```
dev = pearson = sum(residuals(model4,type = "deviance")^2)
```



8. Issues: Over-dispersion and Outliers

```
## over-dispersion
model.disp = glm(Alive_Dead~Cancer+Year+Age+Size, data = liver.data, family = quasibinomial)
summary(model.disp)
```

```
##
## Call:
## glm(formula = Alive_Dead ~ Cancer + Year + Age + Size, family = quasibinomial,
##     data = liver.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5017  -1.1045   0.6756   1.0289   1.8599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.371265   0.780608  -1.757  0.07975 .
## Cancer1     -0.771936   0.268119  -2.879  0.00421 **
## Year1       -0.488435   0.228716  -2.136  0.03333 *
## Age          0.022907   0.010898   2.102  0.03619 *
## Size         0.014771   0.003311   4.461 1.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.094131)
##
##     Null deviance: 550.51  on 399  degrees of freedom
## Residual deviance: 506.40  on 395  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```


Outliers: 41, 35, 27, 41, 39, 39, 39
Age
30
40
50
60
70
80
90
ers: 150, 150, 150, 150, 200, 210, 150, 150, 220, 200, 210, 170, 150, 220, 220, 150, 19(
Size
50
100
150
200

CONTENTS

Log Linear Regression

Data and Preliminary Analysis

Study Model (Homogeneous Association Model

- Computation of Relative Odds and its 95% Confidence Interval

Model Comparison and Selection

Study Model (Joint Independence Model)

- Computation of Relative Odds and its 95% Confidence Interval
- Hypothesis Test

Independence Analysis

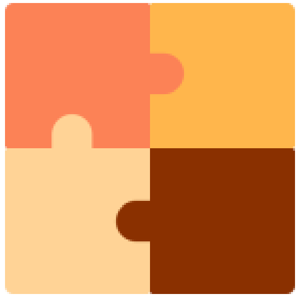
Graphics

Model Diagnostics and GOF

Data and Preliminary Analysis

Psychiatric hospitalization is increasingly specialized these days. Hartford Hospital in the United States conducted a CYP-GUIDES randomized controlled trial (RCT) where 477 patients are assigned to standard therapy (Group S) and 982 to genetically-guided therapy (Group G).

This trial compares the outcome in hospitalized patients with **major depressive disorder or severe depression** treated according to the patient's CYP2D6 genotype and functional status versus standard therapy. The primary outcome was hospital Length of Stay (LOS) in hour.



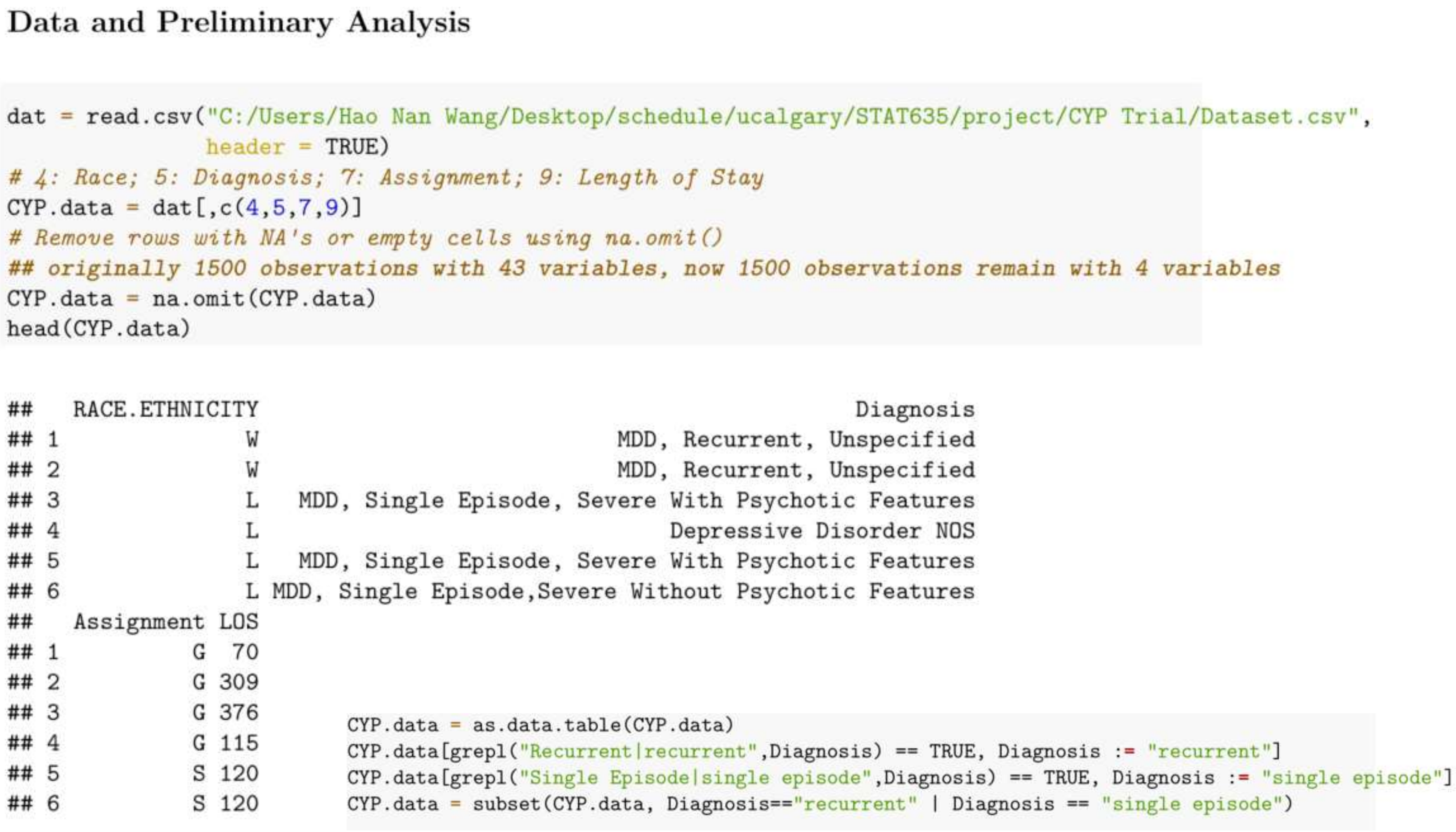
Every person has the right to be free from racial discrimination, some people suspect that there is an association between Assignment and Diagnosis, but this association is not the same in every ethnic group.

Data Attributes (there are 43 attributes in this data set):

ID – Unique identification number assigned to each patient who was enrolled in the trial.
GENDER – Male or female.
AGE – The age, in years, of each patient at the time of enrollment.
RACE/ETHNICITY – Race was self-reported by the patient from a system database, which included "White", "Black", "Latino", and "Other/Unknown" as options. Therefore, this column is referred to as "Race/Ethnicity".
DIAGNOSIS – The diagnosis given upon admission to the hospital, which was used by the trial coordinator to evaluate each patient for inclusion in the trial.
MD – Alphabetic code assigned to each hospital physician who cared for the patients in the trial.
ASSIGNMENT – Patients were randomly assigned to *Group S* or *Group G* in a *1:2* ratio. Patients in *Group S* received standard care, whereas patients in *Group G* had their psychotropic prescriptions guided by their CYP2D6 functional status.
ELECTRONIC MEDICAL RECORD (EMR) – During the course of the trial, 2 platforms for EMR were employed: the Clinical Evaluation and Monitoring System *(C)* and the Epic® *(E)* EMR. The first 856 patients were recruited and followed under CEMS, and the subsequent 644 under Epic®.
LOS – Length of Stay (in hours) at the psychiatric hospital (IOL), defined and calculated as the date/time of discharge minus the date/time of admission.

……..

……..

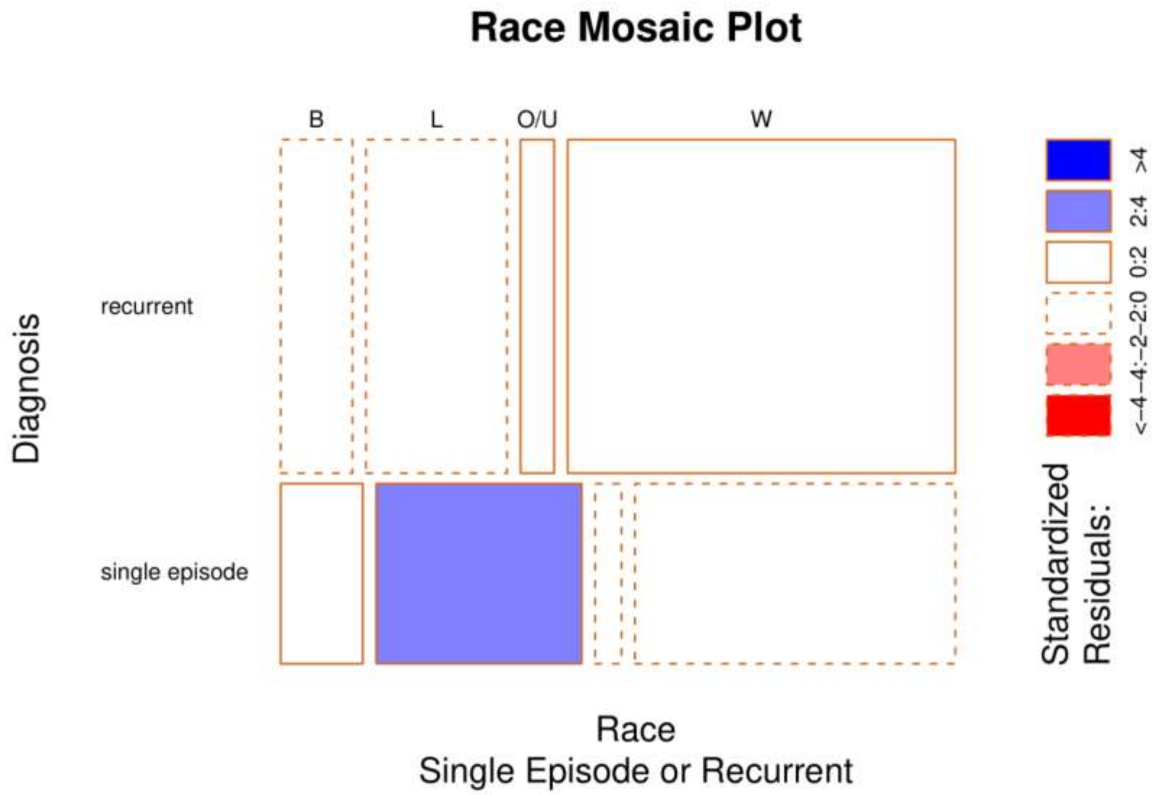


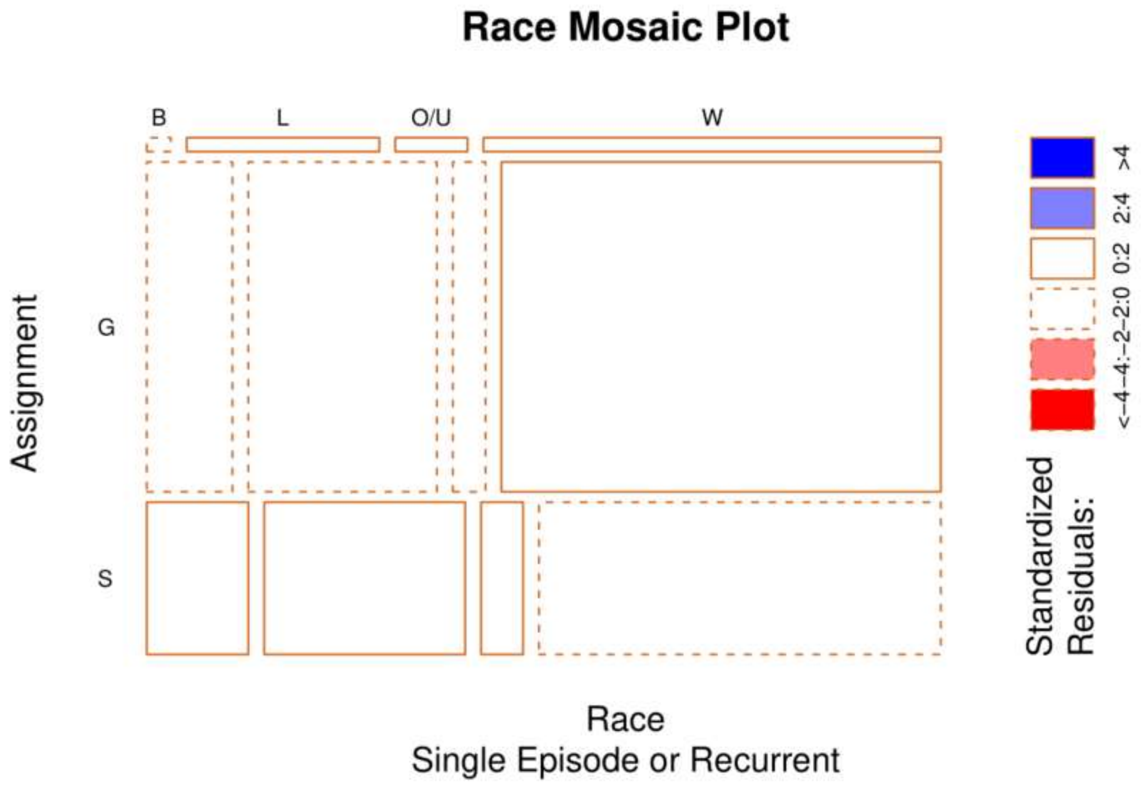
Data and Preliminary Analysis

```r
dat = read.csv("C:/Users/Hao Nan Wang/Desktop/schedule/ucalgary/STAT635/project/CYP Trial/Dataset.csv",
               header = TRUE)
# 4: Race; 5: Diagnosis; 7: Assignment; 9: Length of Stay
CYP.data = dat[,c(4,5,7,9)]
# Remove rows with NA's or empty cells using na.omit()
## originally 1500 observations with 43 variables, now 1500 observations remain with 4 variables
CYP.data = na.omit(CYP.data)
head(CYP.data)
```

```
##   RACE.ETHNICITY                                              Diagnosis
## 1              W                           MDD, Recurrent, Unspecified
## 2              W                           MDD, Recurrent, Unspecified
## 3              L   MDD, Single Episode, Severe With Psychotic Features
## 4              L                              Depressive Disorder NOS
## 5              L   MDD, Single Episode, Severe With Psychotic Features
## 6              L MDD, Single Episode,Severe Without Psychotic Features
##   Assignment LOS
## 1          G  70
## 2          G 309
## 3          G 376
## 4          G 115
## 5          S 120
## 6          S 120
```

```r
CYP.data = as.data.table(CYP.data)
CYP.data[grepl("Recurrent|recurrent",Diagnosis) == TRUE, Diagnosis := "recurrent"]
CYP.data[grepl("Single Episode|single episode",Diagnosis) == TRUE, Diagnosis := "single episode"]
CYP.data = subset(CYP.data, Diagnosis=="recurrent" | Diagnosis == "single episode")
```

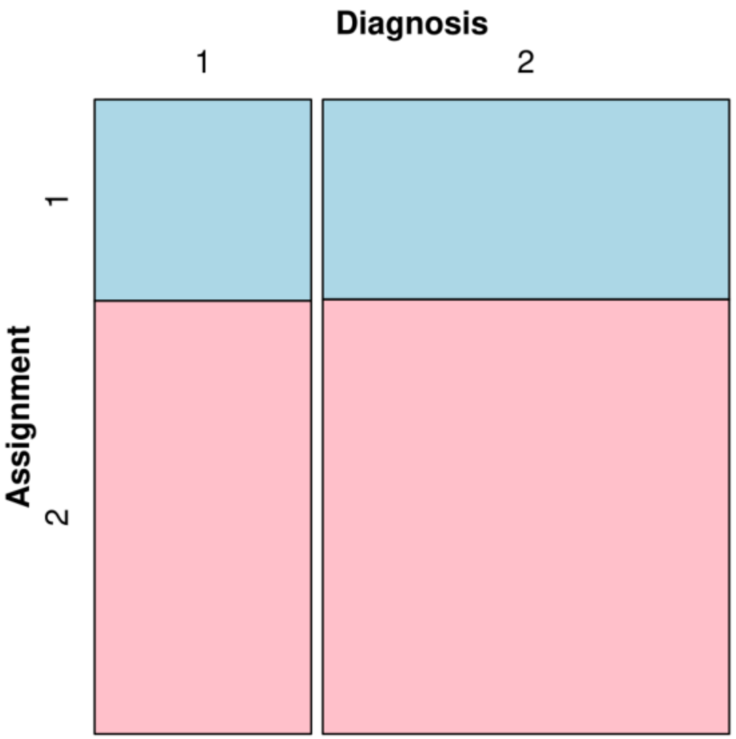

Race Mosaic Plot
B
L
O/U
W
Diagnosis
recurrent
single episode
Race
Single Episode or Recurrent
Standardized Residuals:
>4
2:4
0:2
-2:0
-4:-2
<-4

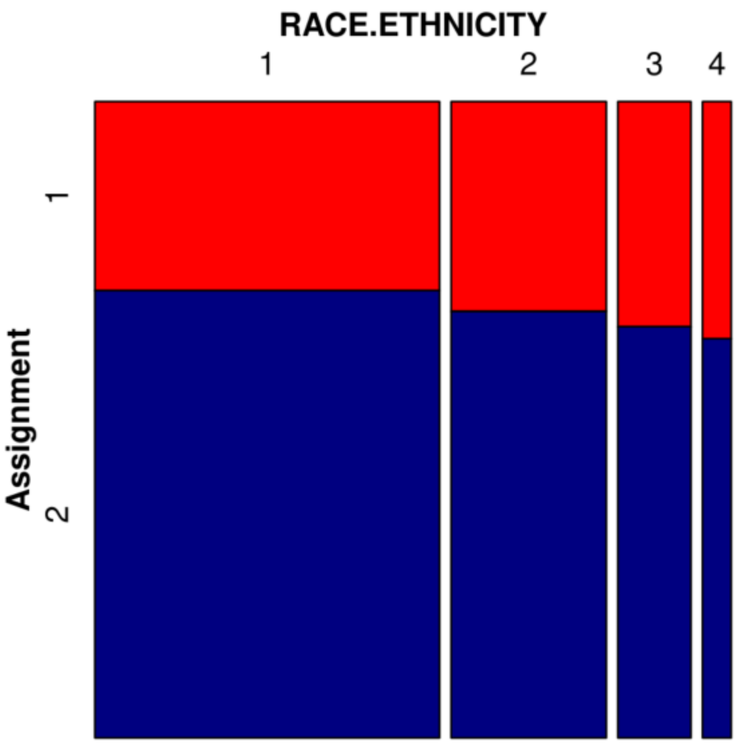

Race Mosaic Plot
B
L
O/U
W
Assignment
G
S
Race
Single Episode or Recurrent
Standardized Residuals:
>4
2:4
0:2
-2:0
-4:-2
<-4

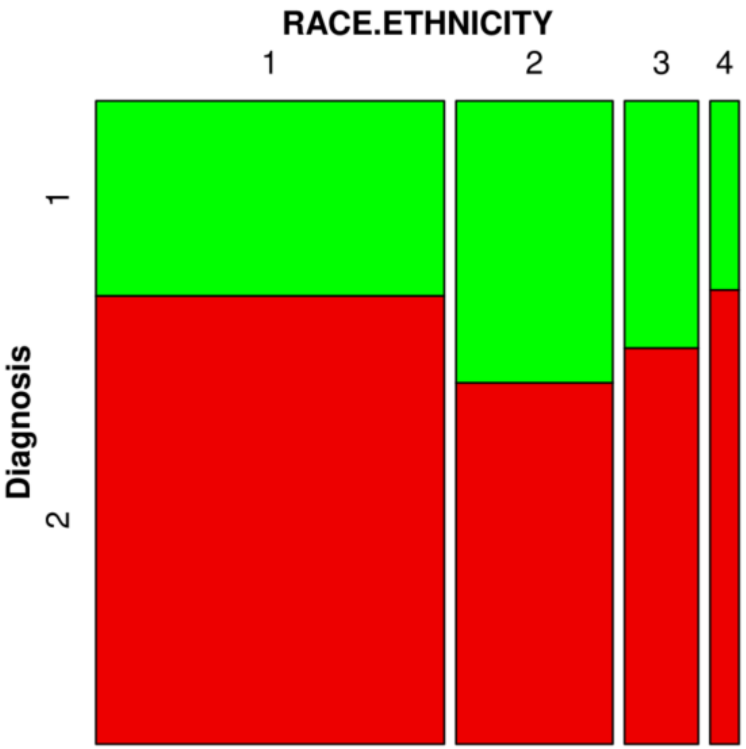
```
# data recode and regroup
CYP.data = as.data.table(CYP.data)
CYP.data[Assignment == "G", Assignment := 2] # psychotropic prescriptions guided by CYP2D6
CYP.data[Assignment == "S", Assignment := 1] # standard care
CYP.data = subset(CYP.data, Assignment==1 | Assignment == 2)

CYP.data[RACE.ETHNICITY == "W", RACE.ETHNICITY := 1] # White
CYP.data[RACE.ETHNICITY == "L", RACE.ETHNICITY := 2] # Latino
CYP.data[RACE.ETHNICITY == "B", RACE.ETHNICITY := 3] # Black
CYP.data[RACE.ETHNICITY == "O/U", RACE.ETHNICITY := 4] # Otherwise/Unknown

CYP.data[grepl("Recurrent|recurrent",Diagnosis) == TRUE, Diagnosis := 2]
CYP.data[grepl("Single Episode|single episode",Diagnosis) == TRUE, Diagnosis := 1]
CYP.data = subset(CYP.data, Diagnosis==1 | Diagnosis == 2)
# Now only 1070 observations are left at this moment
head(CYP.data)
```

```
##    RACE.ETHNICITY Diagnosis Assignment LOS
## 1:              1         2          2  70
## 2:              1         2          2 309
## 3:              2         1          2 376
## 4:              2         1          1 120
## 5:              2         1          1 120
## 6:              1         1          2 113
```

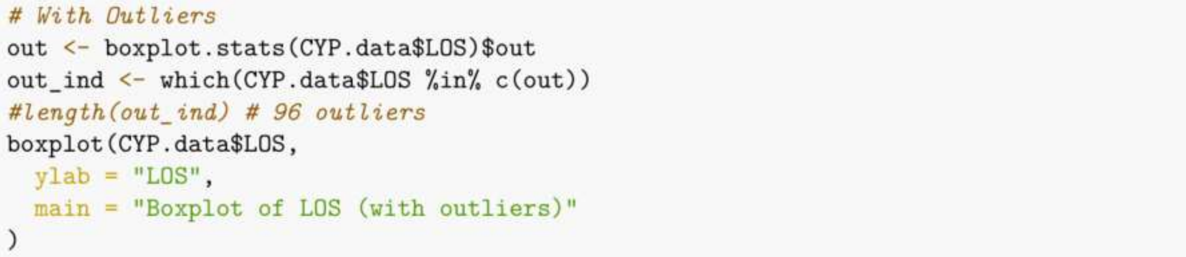
A mosaic plot is a graphical representation of a contingency table that shows percentages of data in groups



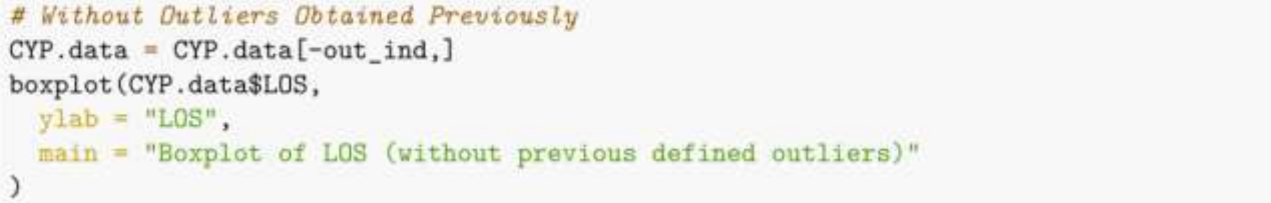




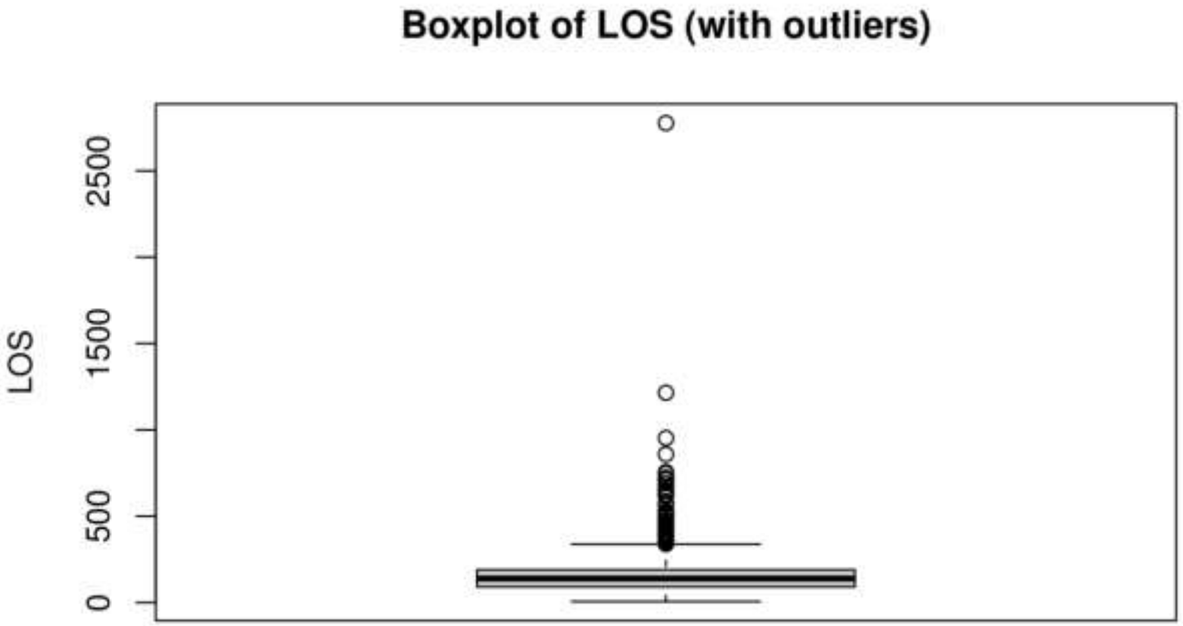
Outliers

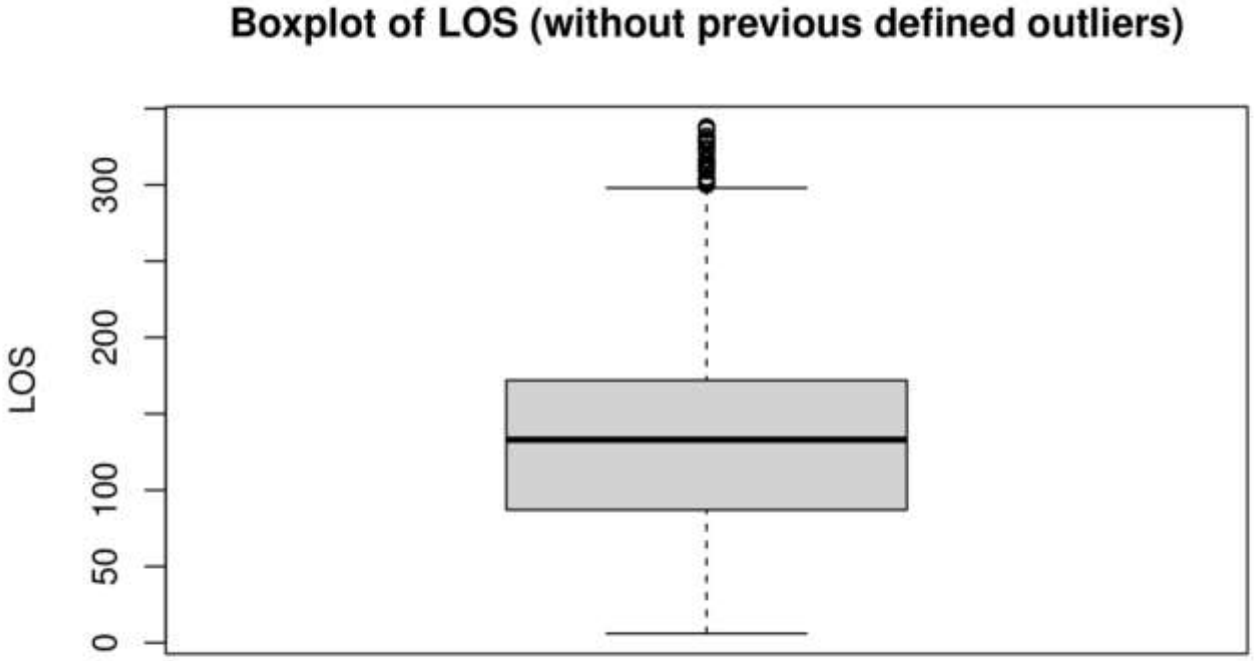


```r
# With Outliers
out <- boxplot.stats(CYP.data$LOS)$out
out_ind <- which(CYP.data$LOS %in% c(out))
#length(out_ind) # 96 outliers
boxplot(CYP.data$LOS,
  ylab = "LOS",
  main = "Boxplot of LOS (with outliers)"
)
```

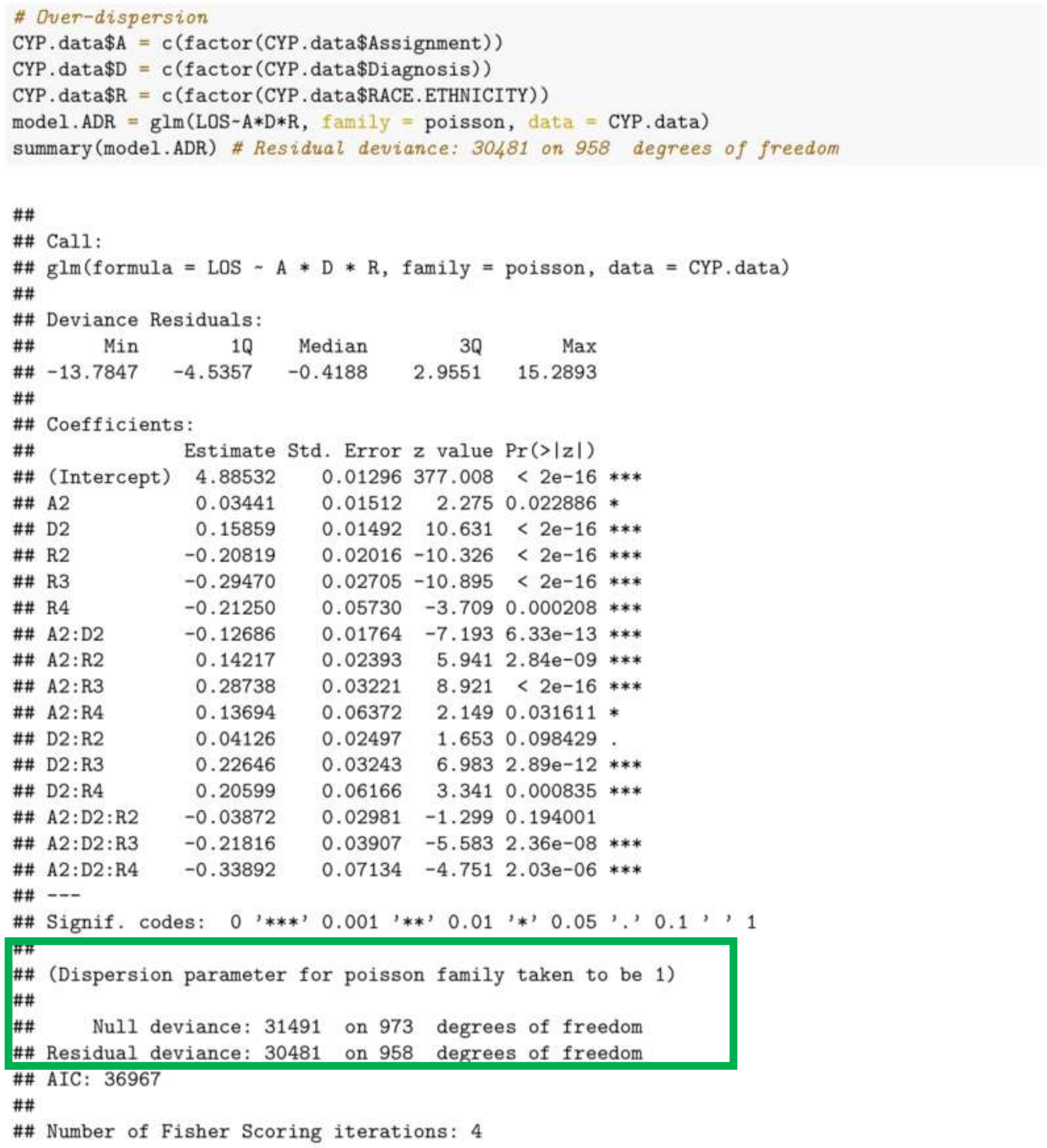


```r
# Without Outliers Obtained Previously
CYP.data = CYP.data[-out_ind,]
boxplot(CYP.data$LOS,
  ylab = "LOS",
  main = "Boxplot of LOS (without previous defined outliers)"
)
```





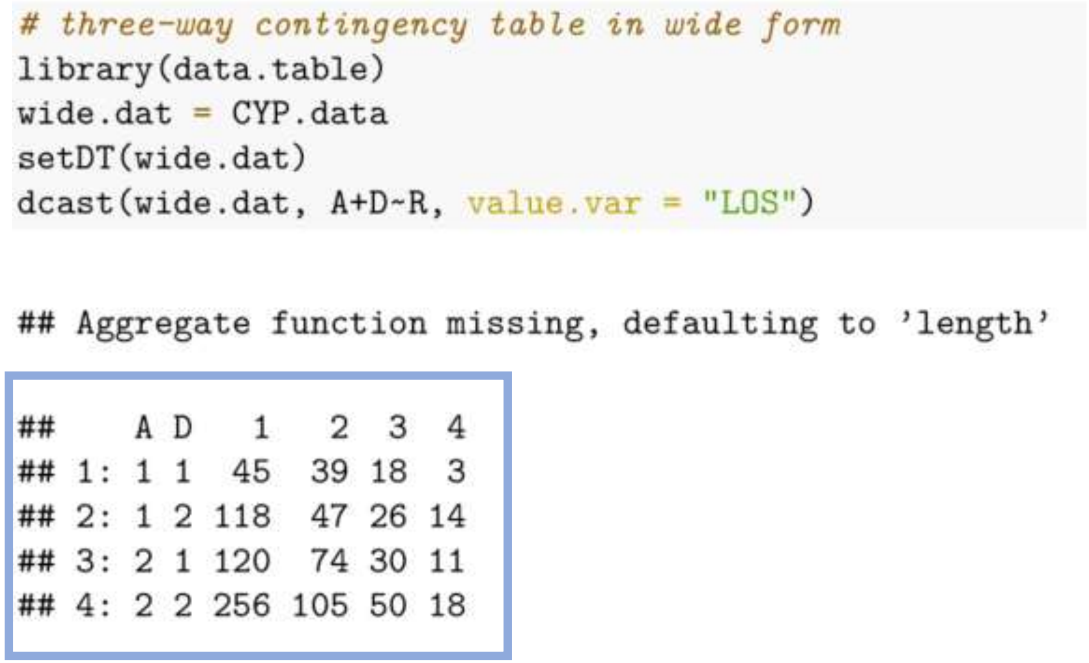
Over-Dispersion



```
# Over-dispersion
CYP.data$A = c(factor(CYP.data$Assignment))
CYP.data$D = c(factor(CYP.data$Diagnosis))
CYP.data$R = c(factor(CYP.data$RACE.ETHNICITY))
model.ADR = glm(LOS~A*D*R, family = poisson, data = CYP.data)
summary(model.ADR) # Residual deviance: 30481 on 958  degrees of freedom
```

```
##
## Call:
## glm(formula = LOS ~ A * D * R, family = poisson, data = CYP.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.7847   -4.5357   -0.4188    2.9551   15.2893
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.88532    0.01296 377.008  < 2e-16 ***
## A2           0.03441    0.01512   2.275 0.022886 *
## D2           0.15859    0.01492  10.631  < 2e-16 ***
## R2          -0.20819    0.02016 -10.326  < 2e-16 ***
## R3          -0.29470    0.02705 -10.895  < 2e-16 ***
## R4          -0.21250    0.05730  -3.709 0.000208 ***
## A2:D2       -0.12686    0.01764  -7.193 6.33e-13 ***
## A2:R2        0.14217    0.02393   5.941 2.84e-09 ***
## A2:R3        0.28738    0.03221   8.921  < 2e-16 ***
## A2:R4        0.13694    0.06372   2.149 0.031611 *
## D2:R2        0.04126    0.02497   1.653 0.098429 .
## D2:R3        0.22646    0.03243   6.983 2.89e-12 ***
## D2:R4        0.20599    0.06166   3.341 0.000835 ***
## A2:D2:R2    -0.03872    0.02981  -1.299 0.194001
## A2:D2:R3    -0.21816    0.03907  -5.583 2.36e-08 ***
## A2:D2:R4    -0.33892    0.07134  -4.751 2.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 31491  on 973  degrees of freedom
## Residual deviance: 30481  on 958  degrees of freedom
## AIC: 36967
##
## Number of Fisher Scoring iterations: 4
```

Three-Way Contingency Table



```
# three-way contingency table in wide form
library(data.table)
wide.dat = CYP.data
setDT(wide.dat)
dcast(wide.dat, A+D~R, value.var = "LOS")
```

```
## Aggregate function missing, defaulting to 'length'
```

```
##    A D   1   2  3  4
## 1: 1 1  45  39 18  3
## 2: 1 2 118  47 26 14
## 3: 2 1 120  74 30 11
## 4: 2 2 256 105 50 18
```

Interpretation of the Estimated Value

Model Comparison

Confidence Interval

Hypothesis Test

Independence Analysis

Graphics

Model Diagnostics and GOF

Study Model (Homogeneous Association Model)

$$log(\mu_{ijk}, \phi) = \mu + \mu_i^A + \mu_j^D + \mu_k^R + \mu_{ij}^{AD} + \mu_{ik}^{AR} + \mu_{jk}^{DR} \text{ for i = 1,2, j = 1,2, k = 1,2,3,4}$$

Computation of Relative Odds and its 95% Confidence Interval

Example 1: Based on the model with all the two way interactions (the study model), for **White**(R=1), please estimate the relative odds of **Single Episode**(D=1) versus **Recurrent**(D=2) patients **Length of Stay** for **standard treatment**(A=1) versus **psychotropic prescriptions guided by CYP2D6**(A=2)

```
# Use Quasipoisson to deal with the over-dispersion problem
model.AD.DR.AR = glm(LOS~A*D+D*R+A*R, family = quasipoisson, data = CYP.data)
summary(model.AD.DR.AR)$coefficient
```

```
##                 Estimate Std. Error     t value    Pr(>|t|)
## (Intercept)  4.855878830 0.06421474 75.61937684 0.000000000
## A2           0.074316752 0.06991898  1.06289808 0.288095189
## D2           0.197436504 0.06947799  2.84171285 0.004582095
## R2          -0.185105130 0.08432232 -2.19520907 0.028387126
## R3          -0.191181356 0.10949566 -1.74601761 0.081127432
## R4          -0.003967516 0.18842691 -0.02105599 0.983205361
## A2:D2       -0.180946308 0.07330433 -2.46842603 0.013743982
## D2:R2        0.013070134 0.07735427  0.16896460 0.865860041
## D2:R3        0.076646524 0.10244877  0.74814489 0.454555814
## D2:R4       -0.041551942 0.17344313 -0.23957099 0.810713943
## A2:R2        0.111423754 0.08078509  1.37926143 0.168135051
## A2:R3        0.139383441 0.10324992  1.34996171 0.177346139
## A2:R4       -0.128151869 0.16020995 -0.79989956 0.423966668
```

```
# mu 22 AD
# A2:D2        -0.1809
# est.OR
exp(-0.1809)
```

```
## [1] 0.8345188
```

```
# 95% for = mu 22 AD...does not contain zero, statistically significant
c(summary(model.AD.DR.AR)$coefficients[7,1]-1.96*summary(model.AD.DR.AR)$coefficients[7,2],
summary(model.AD.DR.AR)$coefficients[7,1]+1.96*summary(model.AD.DR.AR)$coefficients[7,2])
```

```
## [1] -0.32462279 -0.03726983
```

```
# 95% for OR
exp(c(-0.1944246,0.2776185)) # contain 1, it is not statistically significant
```

```
## [1] 0.8233083 1.3199825
```

Example 2: Based on the model with all the two way interactions (the study model), for **Recurrent**(D=2), please estimate the relative odds of **White**(R=1) versus **Latino**(R=2) patients **Length of Stay** for **standard treatment**(A=1) versus **psychotropic prescriptions guided by CYP2D6**(A=2)

```
# mu 22 AR + mu 22 DR: 0.11142+0.01307=0.12449
# est.OR
exp(0.12449)
```

```
## [1] 1.132571
```

```
# 95% for = mu 22 AR + mu 22 DR...contain zero, not statistically significant
c(summary(model.AD.DR.AR)$coefficients[11,1]-1.96*summary(model.AD.DR.AR)$coefficients[11,2],
summary(model.AD.DR.AR)$coefficients[11,1]+1.96*summary(model.AD.DR.AR)$coefficients[11,2])
```

```
## [1] -0.04691502  0.26976252
```

```
# 95% for OR
exp(c( -0.04691502, 0.26976252)) # contain 1, it is not statistically significant
```

```
## [1] 0.9541685 1.3096534
```

Can we use Conditional Independence Model or other models instead?

Model Comparison and Selection

Saturated Model -> Homogeneous Association Model -> Conditional Independence Model ->

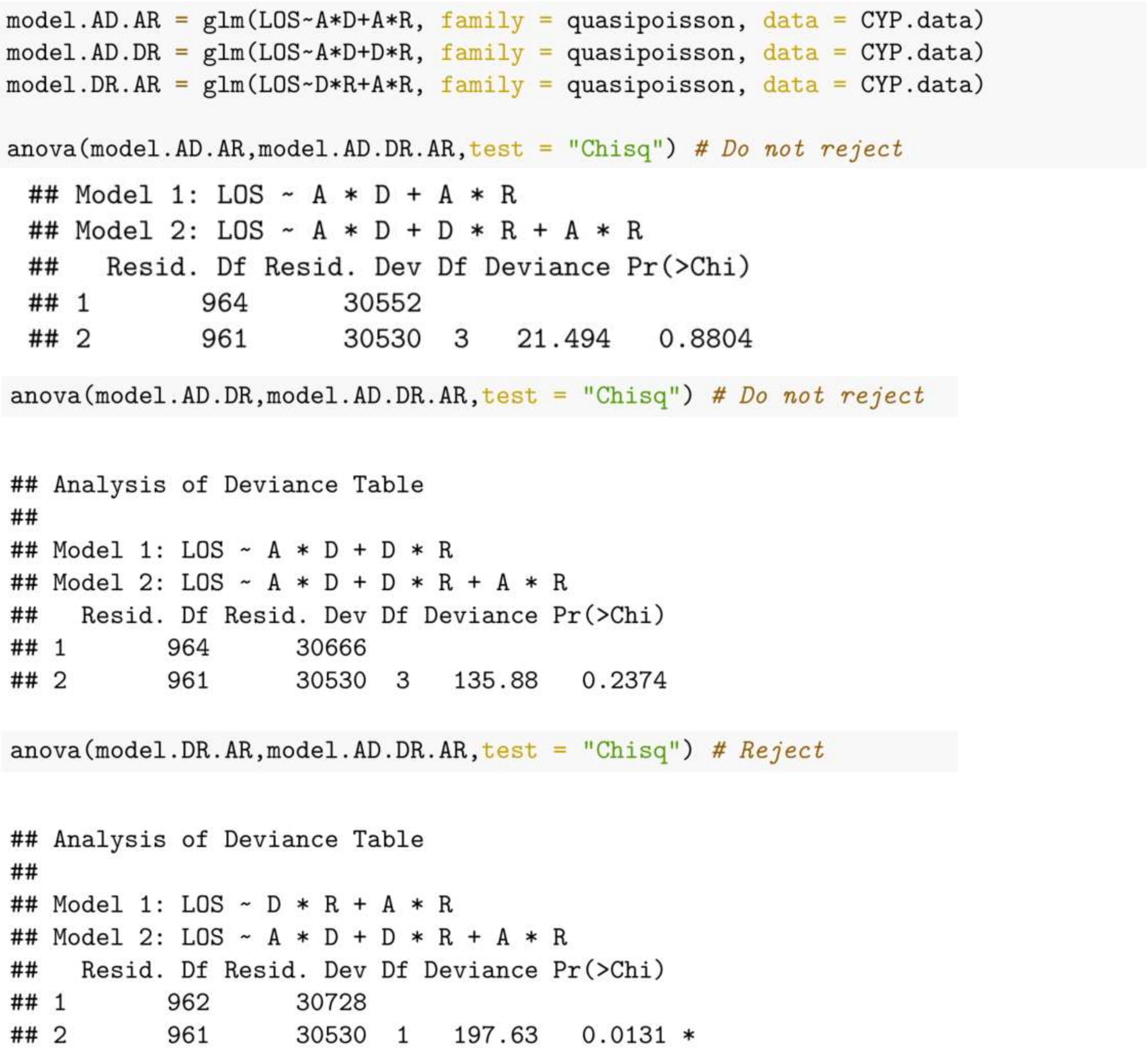
Joint Independence Model -> Mutual Independence Model

(ADR) -> (AD, AR, DR) -> (AD, AR), (AR, DR), (AD, DR) ->

(AD, R) or (AR, D) or (DR, A) -> (A,D,R)

```
model.ADR = glm(LOS~A*D*R, family = quasipoisson, data = CYP.data)
model.AD.DR.AR = glm(LOS~A*D+D*R+A*R, family = quasipoisson, data = CYP.data)
# Deviance Test -> Do Not Reject H0
anova(model.AD.DR.AR, model.ADR, test = "Chisq") # Do Not Reject
```

```
## Analysis of Deviance Table
##
## Model 1: LOS ~ A * D + D * R + A * R
## Model 2: LOS ~ A * D * R
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       961      30530
## 2       958      30481  3   49.695   0.6708
```



```
model.AD.AR = glm(LOS~A*D+A*R, family = quasipoisson, data = CYP.data)
model.AD.DR = glm(LOS~A*D+D*R, family = quasipoisson, data = CYP.data)
model.DR.AR = glm(LOS~D*R+A*R, family = quasipoisson, data = CYP.data)

anova(model.AD.AR,model.AD.DR.AR,test = "Chisq") # Do not reject
```

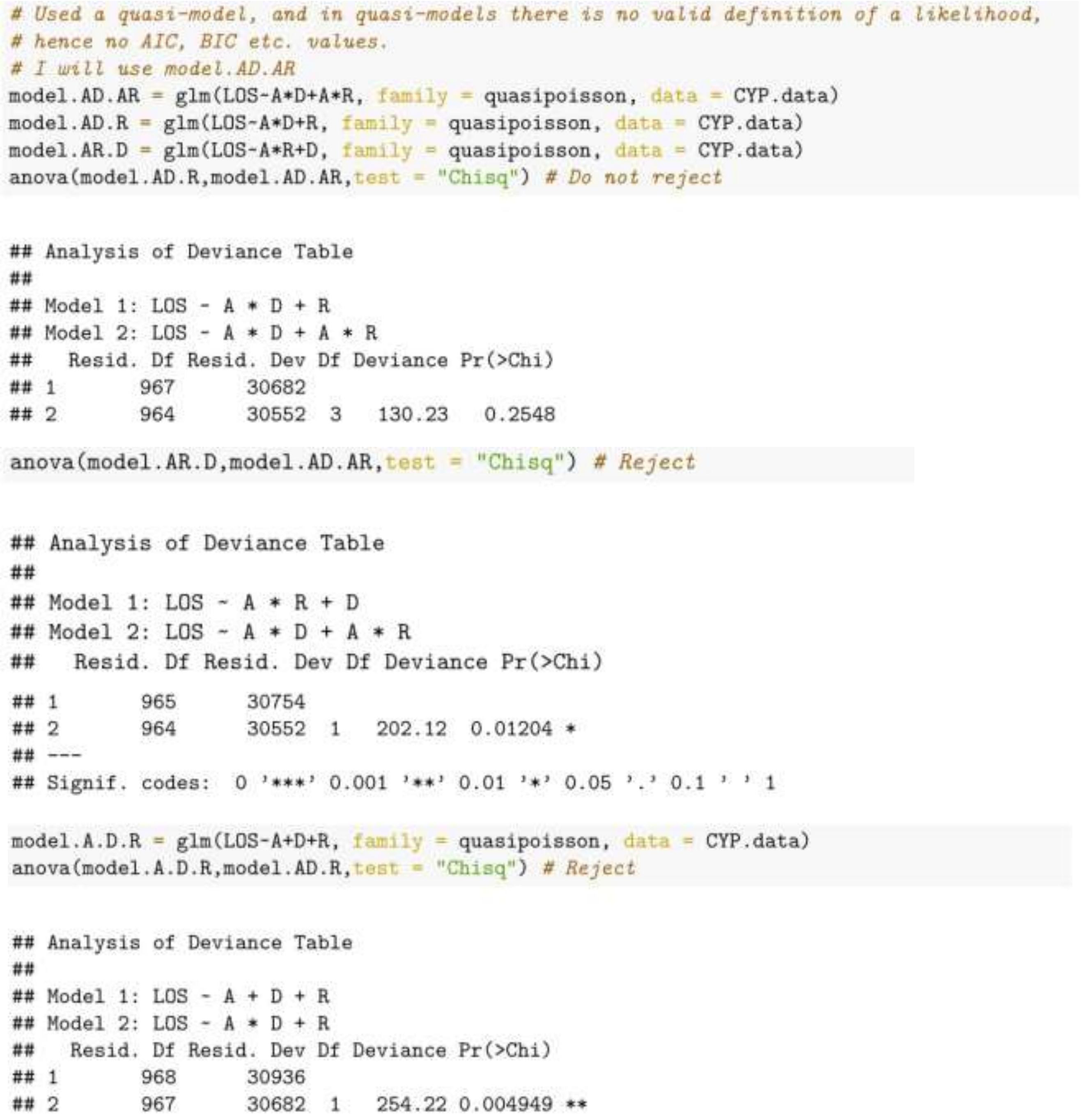
```
## Model 1: LOS ~ A * D + A * R
## Model 2: LOS ~ A * D + D * R + A * R
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       964      30552
## 2       961      30530  3   21.494   0.8804
```

```
anova(model.AD.DR,model.AD.DR.AR,test = "Chisq") # Do not reject
```

```
## Analysis of Deviance Table
##
## Model 1: LOS ~ A * D + D * R
## Model 2: LOS ~ A * D + D * R + A * R
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       964      30666
## 2       961      30530  3   135.88   0.2374
```

```
anova(model.DR.AR,model.AD.DR.AR,test = "Chisq") # Reject
```

```
## Analysis of Deviance Table
##
## Model 1: LOS ~ D * R + A * R
## Model 2: LOS ~ A * D + D * R + A * R
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       962      30728
## 2       961      30530  1   197.63   0.0131 *
```



```
# Used a quasi-model, and in quasi-models there is no valid definition of a likelihood,
# hence no AIC, BIC etc. values.
# I will use model.AD.AR
model.AD.AR = glm(LOS~A*D+A*R, family = quasipoisson, data = CYP.data)
model.AD.R = glm(LOS~A*D+R, family = quasipoisson, data = CYP.data)
model.AR.D = glm(LOS~A*R+D, family = quasipoisson, data = CYP.data)
anova(model.AD.R,model.AD.AR,test = "Chisq") # Do not reject
```

```
## Analysis of Deviance Table
##
## Model 1: LOS ~ A * D + R
## Model 2: LOS ~ A * D + A * R
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       967      30682
## 2       964      30552  3   130.23   0.2548
```

```
anova(model.AR.D,model.AD.AR,test = "Chisq") # Reject
```

```
## Analysis of Deviance Table
##
## Model 1: LOS ~ A * R + D
## Model 2: LOS ~ A * D + A * R
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       965      30754
## 2       964      30552  1   202.12  0.01204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.A.D.R = glm(LOS~A+D+R, family = quasipoisson, data = CYP.data)
anova(model.A.D.R,model.AD.R,test = "Chisq") # Reject
```

```
## Analysis of Deviance Table
##
## Model 1: LOS ~ A + D + R
## Model 2: LOS ~ A * D + R
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       968      30936
## 2       967      30682  1   254.22 0.004949 **
```

This underlines that model.A.D.R is not adequate and therefore, we should use the joint independence model model.AD.R as our new study model

Study Model (Joint Independence Model)

$$log(\mu_{ijk};\phi) = \mu + \mu_i^A + \mu_j^D + \mu_k^R + \mu_{ij}^{AD} \text{ for i = 1,2, j = 1,2, k = 1,2,3,4}$$

Computation of Relative Odds and its 95% Confidence Interval

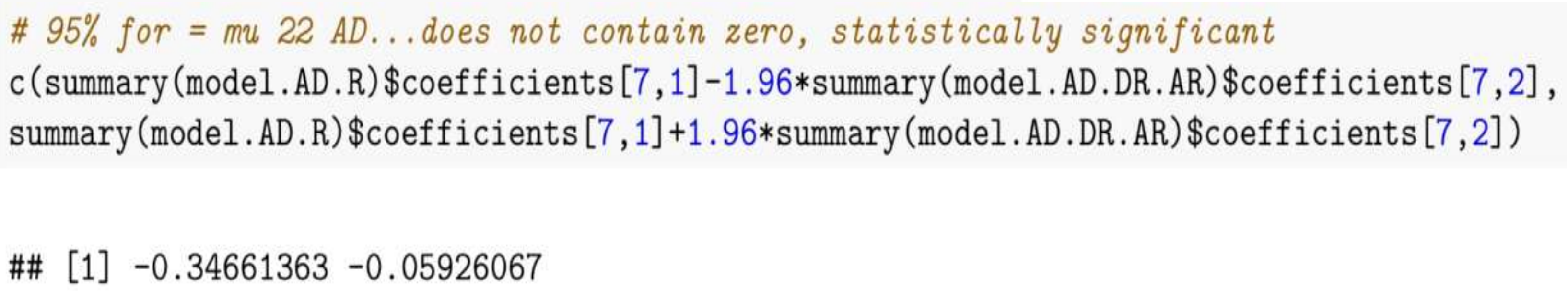
Based on the new proposed study model(model.AD.R), for **White**(R=1), please estimate the relative odds of **Single Episode**(D=1) versus **Recurrent**(D=2) patients **Length of Stay** for **standard treatment**(A=1) versus **psychotropic prescriptions guided by CYP2D6**(A=2)

```
summary(model.AD.R)$coefficient
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  4.80543704 0.05421675 88.6338104 0.0000000000
## A2           0.13066830 0.06054651  2.1581474 0.0311619502
## D2           0.22844047 0.06105412  3.7416063 0.0001936136
## R2          -0.09861142 0.03726269 -2.6463848 0.0082678901
## R3          -0.04759404 0.04855390 -0.9802309 0.3272173394
## R4          -0.10794837 0.07648537 -1.4113597 0.1584602475
## A2:D2       -0.20293715 0.07262780 -2.7942076 0.0053055559
```

```
# mu 22 AD
# A2:D2 -0.20293715
# est.OR
exp(-0.20293715)
```

```
## [1] 0.8163295
```



```
# 95% for = mu 22 AD...does not contain zero, statistically significant
c(summary(model.AD.R)$coefficients[7,1]-1.96*summary(model.AD.DR.AR)$coefficients[7,2],
summary(model.AD.R)$coefficients[7,1]+1.96*summary(model.AD.DR.AR)$coefficients[7,2])
```

```
## [1] -0.34661363 -0.05926067
```

Hypothesis Test

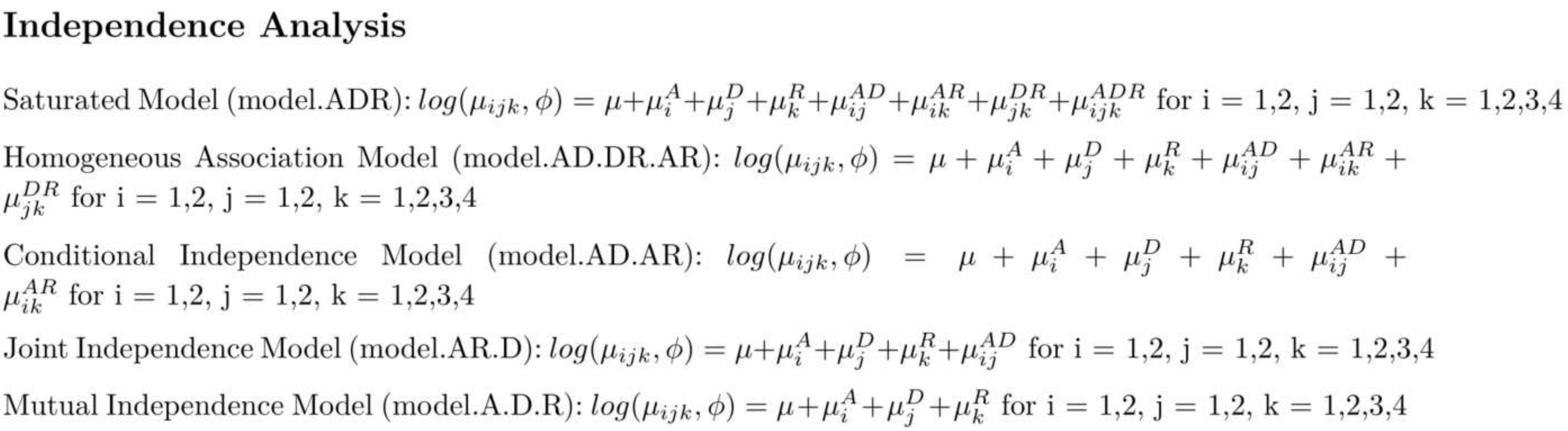
Every person has the right to be free from racial discrimination, some people suspect that there is an association between Assignment and Diagnosis but this association is not the same in every ethnic group. Use the joint independence model (the study model) and do a hypothesis test for it. How many parameters are fit in the model?

H_0 : the study model is adequate Against H_a : the study model is not adequate

```
1-pchisq(deviance(model.AD.R)-deviance(model.ADR),
         df.residual(model.AD.R)-df.residual(model.ADR))
```

```
## [1] 0
```

number of parameters = 1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) = 1 + 1 + 1 + 3 + 1 = 7



Independence Analysis

Saturated Model (model.ADR): $log(\mu_{ijk}, \phi) = \mu + \mu_i^A + \mu_j^D + \mu_k^R + \mu_{ij}^{AD} + \mu_{ik}^{AR} + \mu_{jk}^{DR} + \mu_{ijk}^{ADR}$ for i = 1,2, j = 1,2, k = 1,2,3,4

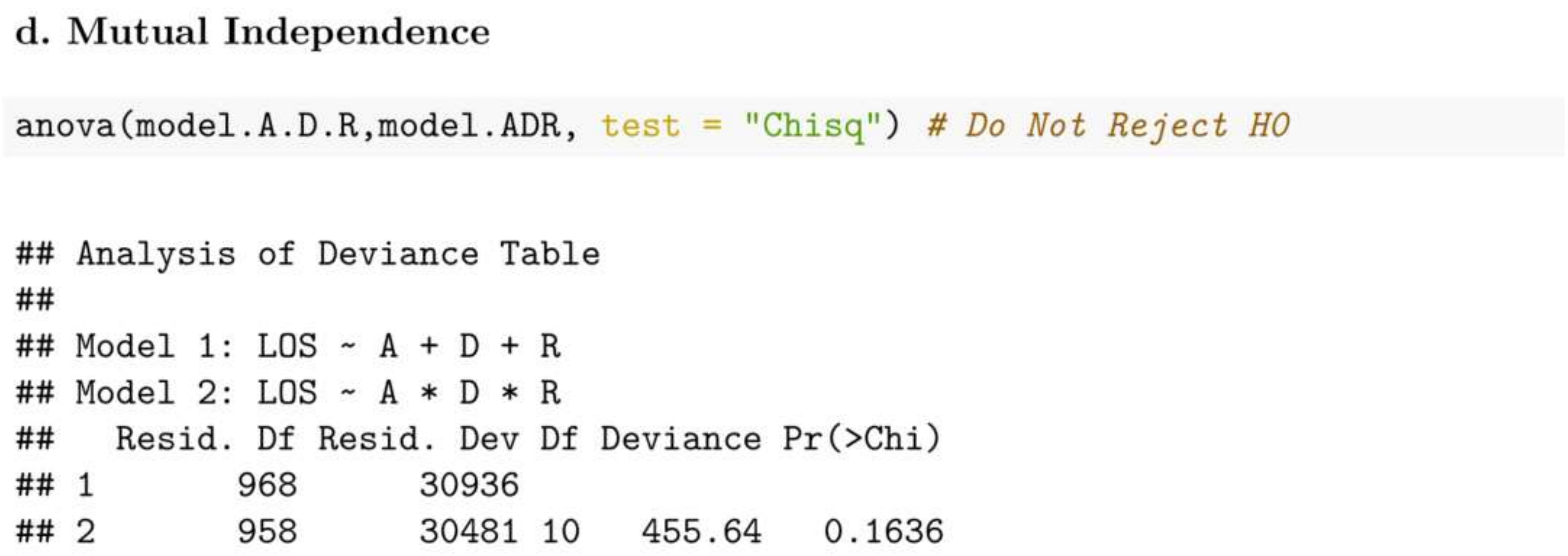
Homogeneous Association Model (model.AD.DR.AR): $log(\mu_{ijk}, \phi) = \mu + \mu_i^A + \mu_j^D + \mu_k^R + \mu_{ij}^{AD} + \mu_{ik}^{AR} + \mu_{jk}^{DR}$ for i = 1,2, j = 1,2, k = 1,2,3,4

Conditional Independence Model (model.AD.AR): $log(\mu_{ijk}, \phi) = \mu + \mu_i^A + \mu_j^D + \mu_k^R + \mu_{ij}^{AD} + \mu_{ik}^{AR}$ for i = 1,2, j = 1,2, k = 1,2,3,4

Joint Independence Model (model.AR.D): $log(\mu_{ijk}, \phi) = \mu + \mu_i^A + \mu_j^D + \mu_k^R + \mu_{ij}^{AD}$ for i = 1,2, j = 1,2, k = 1,2,3,4

Mutual Independence Model (model.A.D.R): $log(\mu_{ijk}, \phi) = \mu + \mu_i^A + \mu_j^D + \mu_k^R$ for i = 1,2, j = 1,2, k = 1,2,3,4

For example,

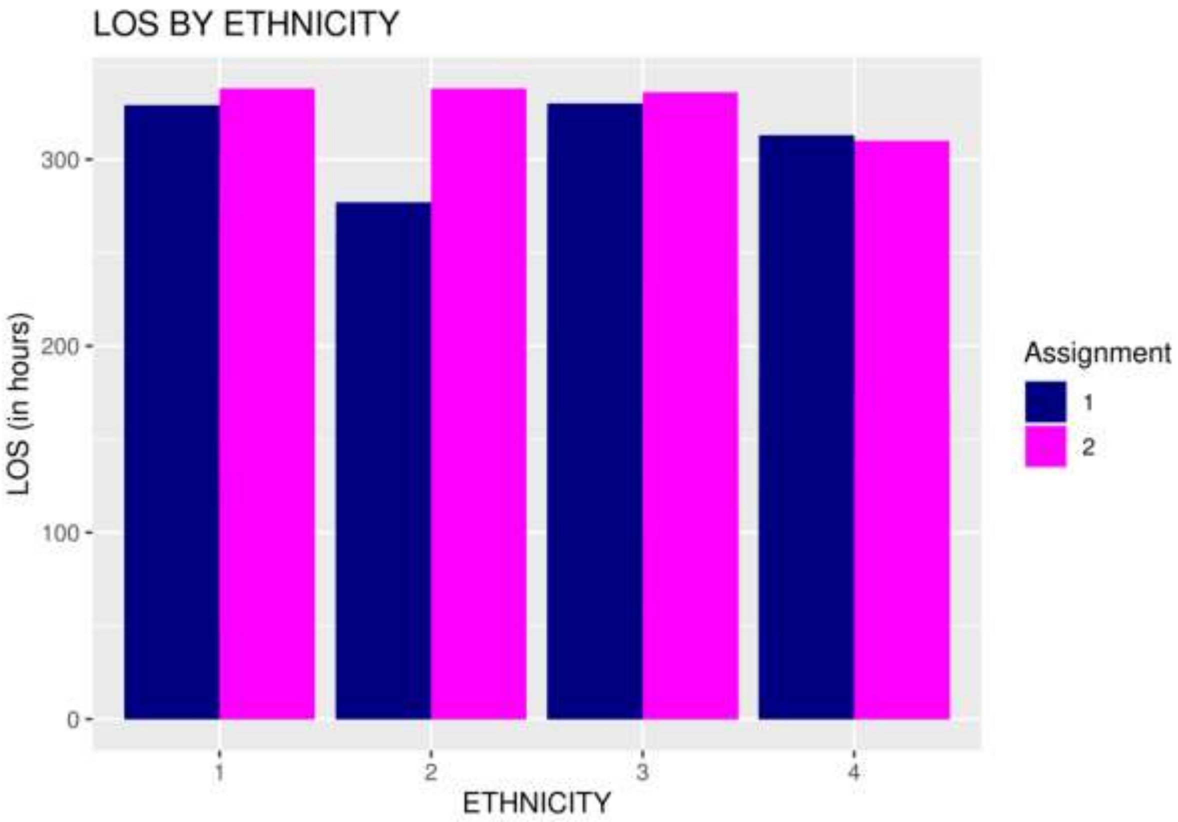


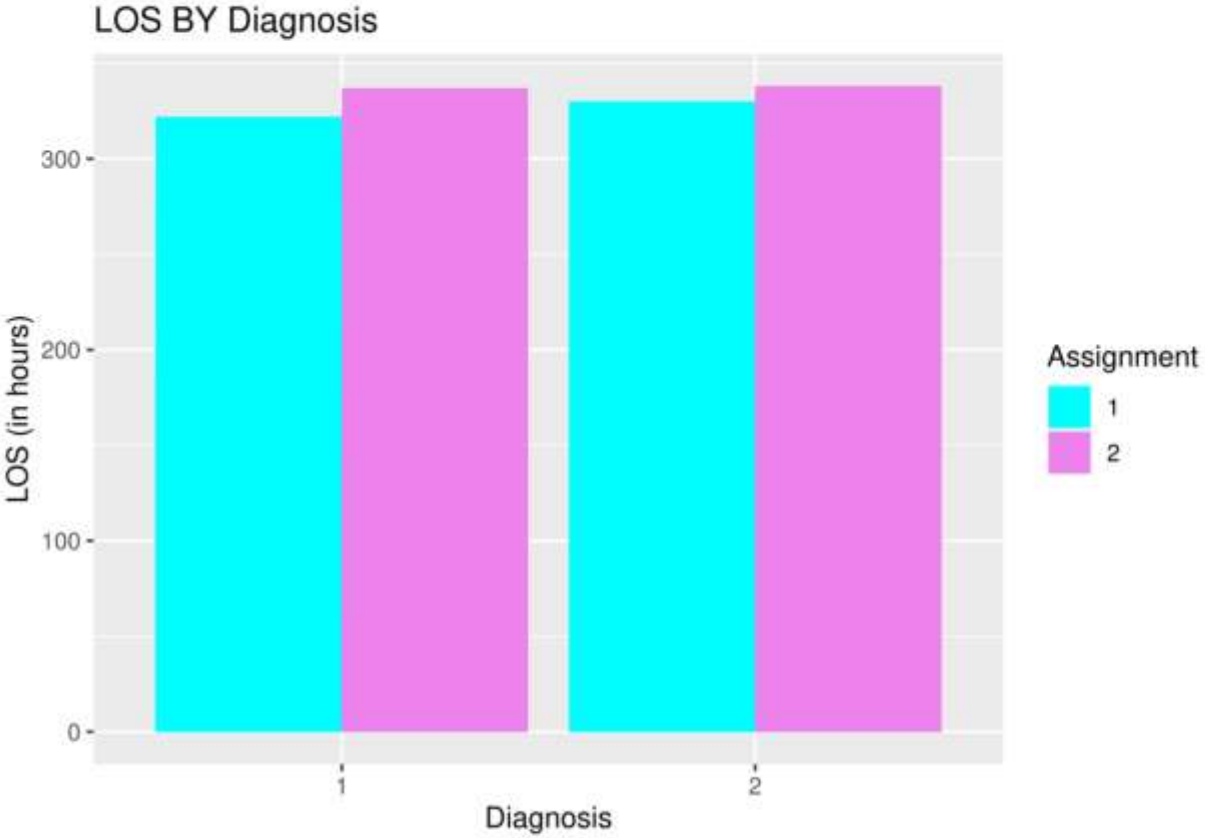
d. Mutual Independence

```
anova(model.A.D.R,model.ADR, test = "Chisq") # Do Not Reject H0
```

```
## Analysis of Deviance Table
##
## Model 1: LOS ~ A + D + R
## Model 2: LOS ~ A * D * R
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       968      30936
## 2       958      30481 10   455.64   0.1636
```

Graphics


LOS BY ETHNICITY
LOS (in hours)
300
200
100
0
1
2
3
4
ETHNICITY
Assignment
1
2


LOS BY Diagnosis
LOS (in hours)
300
200
100
0
1
2
Diagnosis
Assignment
1
2

Model Diagnostics and GOF

Calculate the fitted values (cell counts), the Pearson and deviance residuals, and the goodness of fit statistics X^2 and D. Which of the cells of the table contribute most to X^2 and D? Do X^2 and D indicate the model is a good fit to the data?

```r
table_4 = data.frame(CYP.data$LOS,predict(model.AD.R, type = "response"),
resid(model.AD.R, "pearson"), resid(model.AD.R, "deviance"))
names(table_4) = c("Original", "Fitted Value", "Pearson Res", "Deviance Res")
# Only show the first 30 observation
table_4[1:30,]
```

```
##    Original Fitted Value Pearson Res Deviance Res
## 1        70     142.8234  -6.0935603   -6.7683714
## 2       309     142.8234  13.9049787   12.0240380
## 3       120     110.7002   0.8838927    0.8719309
## 4       120     110.7002   0.8838927    0.8719309
## 5       113     139.2270  -2.2227271   -2.2986390
## 6        67     129.4115  -5.4862827   -6.0506346
## 7       135     132.7558   0.1947763    0.1942313
## 8       104     142.8234  -3.2485798   -3.4154702
## 9        18     128.2088  -9.7332390  -12.2367986
## 10      116     153.5272  -3.0286781   -3.1666493
## 11      142     139.2270   0.2350152    0.2342415
```

GOF Statistics Chi-squared

```
sum(resid(model.AD.R, "pearson")^2)
```

```
## [1] 31125.79
```

GOF Statistics Deviance

```
sum(resid(model.AD.R, "deviance")^2)
```

```
## [1] 30682.06
```

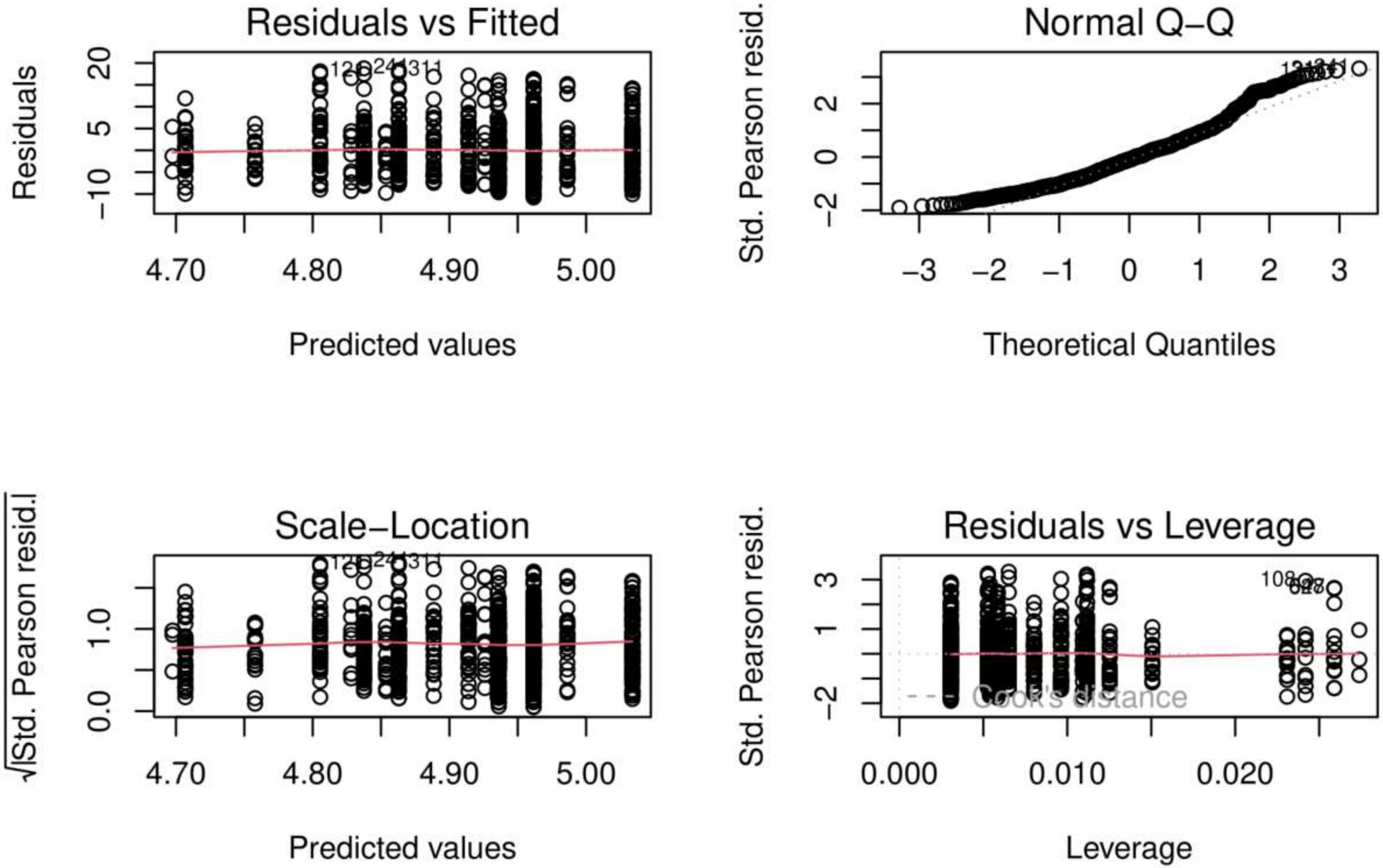
Contribution the Most

```
table_4[table_4$`Pearson Res` == max(abs(table_4$`Pearson Res`)),]
```

```
##     Original Fitted Value Pearson Res Deviance Res
## 241      337     126.1528    18.77238     15.51034
```

```
table_4[table_4$`Deviance Res` == max(abs(table_4$`Deviance Res`)),]
```

```
##     Original Fitted Value Pearson Res Deviance Res
## 241      337     126.1528    18.77238     15.51034
```


Residuals vs Fitted
Residuals
20
5
−10
126
241
311
4.70
4.80
4.90
5.00
Predicted values
Normal Q−Q
Std. Pearson resid.
2
0
−2
131
241
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Scale−Location
√|Std. Pearson resid.|
1.0
0.0
126
241
311
4.70
4.80
4.90
5.00
Predicted values
Residuals vs Leverage
Std. Pearson resid.
3
1
−2
108
627
Cook's distance
0.000
0.010
0.020
Leverage

References

Fedesoriano. (2022, September 26). *Covid-19 effect on liver cancer prediction dataset*. Kaggle. Retrieved December 4, 2022, from https://www.kaggle.com/datasets/fedesoriano/covid19-effect-on-liver-cancer-prediction-dataset

Geh D, Watson R, Sen G, et al COVID-19 and liver cancer: lost patients and larger tumours BMJ Open Gastroenterology 2022;9:e000794. doi: 10.1136/bmjgast-2021-000794

Penn State Eberly College of Science. (n.d.). 5: Three-Way Tables: Types of Independence. STAT 504 Analysis of Discrete Data . Retrieved December 4, 2022, from https://online.stat.psu.edu/stat504/book/export/html/720

Tortora, Joseph; Robinson, Saskia; Baker, Seth; Ruaño, Gualberto (2020), “Clinical Dataset of the CYP-GUIDES Trial”, Mendeley Data, V1, doi: 10.17632/25yjwbphn4.1