Task 1: The EUxPancreas Cohort Study

The EUxPancreas study investigates regional and demographic patterns in pancreatic cancer incidence across Europe. The dataset includes demographic, lifestyle, and regional predictors for new cancer cases, with the goal of identifying key risk factors.

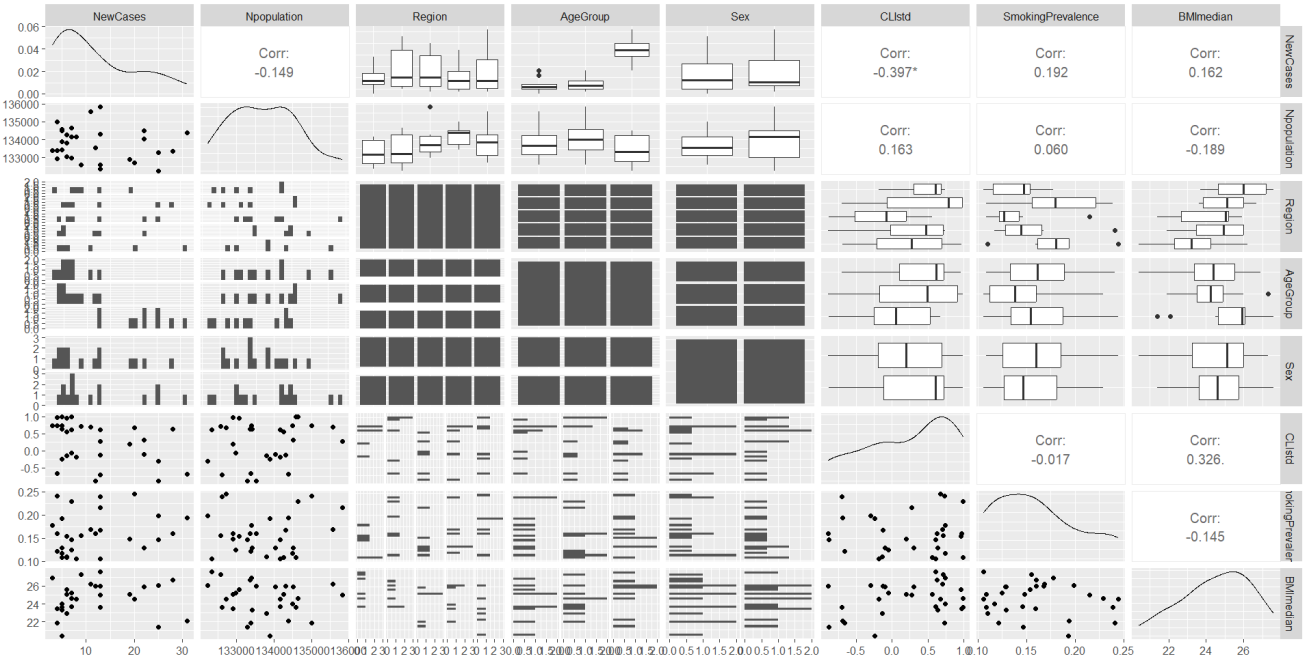
Explore the dataset:

```
> head(data)
  X NewCases Npopulation  Region AgeGroup  Sex CLIstd SmokingPrevalence BMImedian
1 1        7      134167 Region1    20-39 Male  0.608           0.14618      23.7
2 2        6      133057 Region2    20-39 Male  0.961           0.15393      25.6
3 3        7      132978 Region3    20-39 Male -0.055           0.12506      25.2
4 4        5      133420 Region4    20-39 Male  0.637           0.12901      23.3
5 5       11      135585 Region5    20-39 Male  0.705           0.16821      26.2
6 6        8      134167 Region1    40-59 Male -0.182           0.10592      24.5
```

```
> summary(data)
       X            NewCases      Npopulation        Region
 Min.   : 1.00   Min.   : 3.00   Min.   :132219   Length:30
 1st Qu.: 8.25   1st Qu.: 5.25   1st Qu.:132998   Class :character
 Median :15.50   Median : 8.50   Median :133660   Mode  :character
 Mean   :15.50   Mean   :11.93   Mean   :133744
 3rd Qu.:22.75   3rd Qu.:17.50   3rd Qu.:134366
 Max.   :30.00   Max.   :31.00   Max.   :135836
   AgeGroup             Sex                CLIstd        SmokingPrevalence
 Length:30          Length:30          Min.   :-0.8910   Min.   :0.1058
 Class :character   Class :character   1st Qu.:-0.1705   1st Qu.:0.1259
 Mode  :character   Mode  :character   Median : 0.4350   Median :0.1545
                                       Mean   : 0.2368   Mean   :0.1596
                                       3rd Qu.: 0.7133   3rd Qu.:0.1891
                                       Max.   : 0.9970   Max.   :0.2447

   BMImedian
 Min.   :20.40
 1st Qu.:23.52
 Median :24.80
 Mean   :24.55
 3rd Qu.:25.98
 Max.   :27.60
```

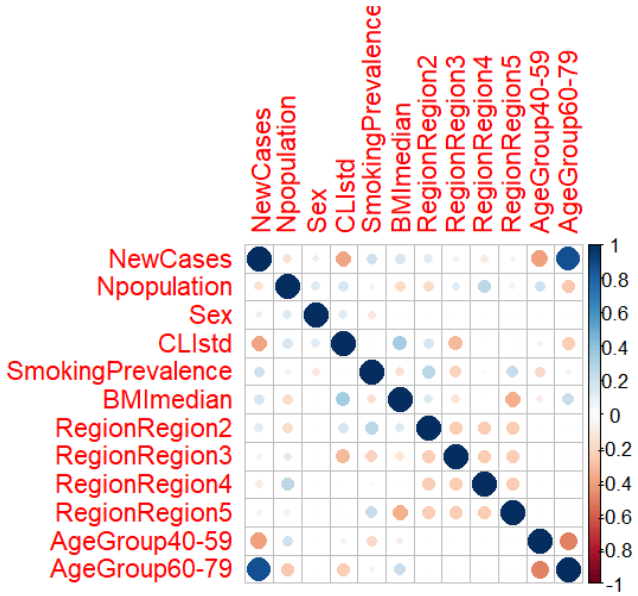
```
> colSums(is.na(data)) #Check for missing data
                X          NewCases       Npopulation            Region
                0                 0                 0                 0
         AgeGroup               Sex           CLIstd SmokingPrevalence
                0                 0                 0                 0
        BMImedian
                0
```

The dataset contained 30 observations across 8 variables with no missing data.



In the data set there are some nonnumerical values. We convert AgeGroup and Region to categorical values and sex to numerical (0 and 1) and plot them in a corrplot for better visibility

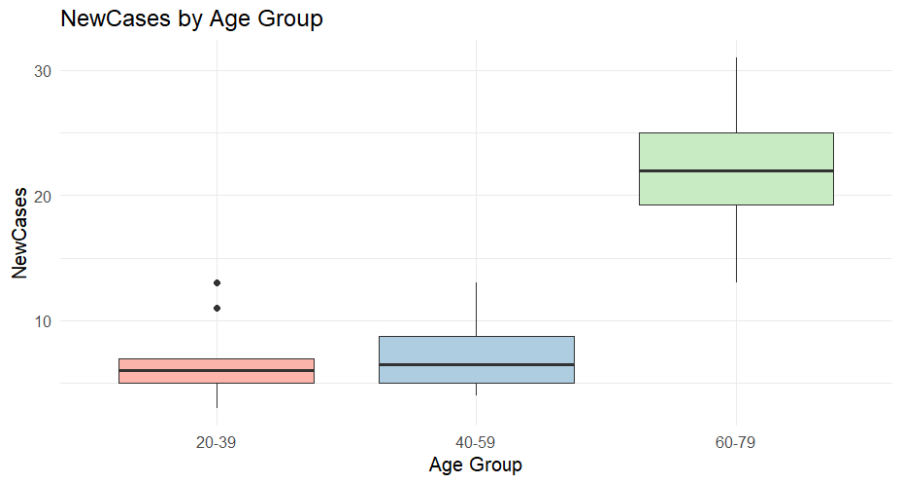
```r
#Ensure the columns are factors
data$Region <- as.factor(data$Region)
data$AgeGroup <- as.factor(data$AgeGroup)


# Convert Sex to binary
data$Sex <- ifelse(data$Sex == "Male", 1, 0)
```



The correlation plot suggests that AgeGroup has a strong positive association with NewCases, while CLIstd (composite lifestyle index) is negatively correlated, suggesting protective effects of healthier lifestyles

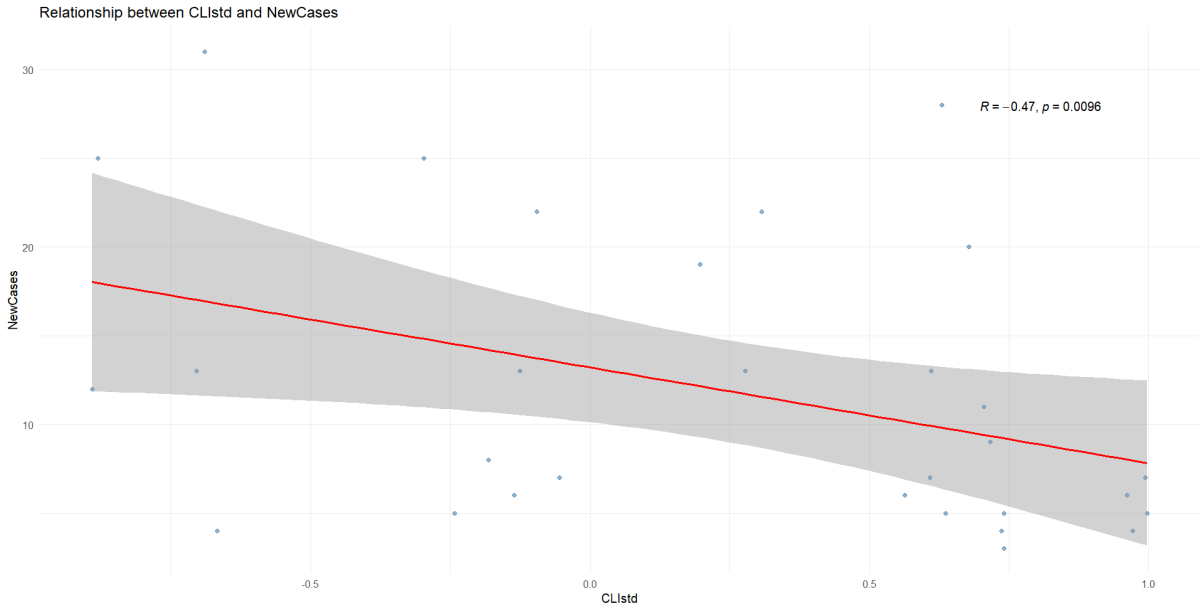
We can therefore plot NewCases by Age Group in a boxplot.

```
> ggplot(data, aes(x = AgeGroup, y = NewCases, fill = AgeGroup)) +
+   geom_boxplot() +
+   theme_minimal() +
+   labs(title = "NewCases by Age Group",
+        x = "Age Group",
+        y = "NewCases") +
+   scale_fill_brewer(palette = "Pastel1") +
+   theme(legend.position = "none")
```



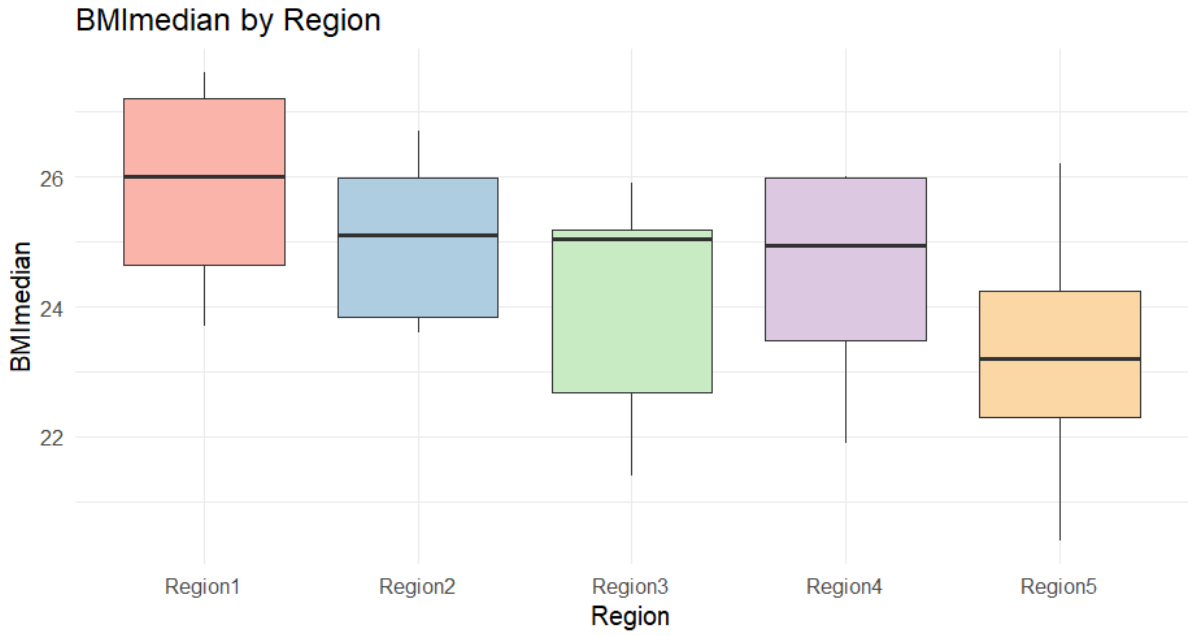
We can clearly see that older people have higher cancer prevalence.

We can also see that the continuous variable CLIstd seems to have a negative correlation with New Cases

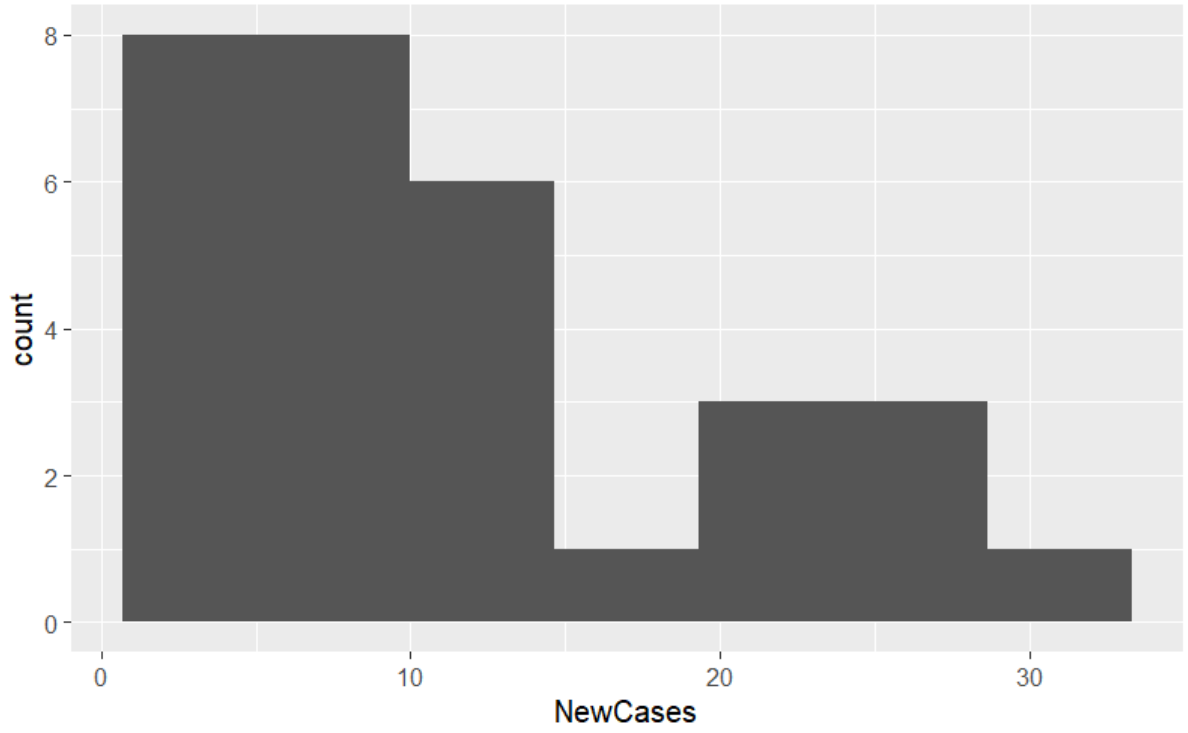


A better lifestyle indicates less cancer prevalence.

We also notice that there is not that strong of a relation between variables. However there is a significant negative intervariable correlation between Region and BMI



NewCases are right-skewed, supporting Poisson modeling.



Exploratory data analysis revealed that the number of new pancreatic cancer cases (NewCases) follows a count distribution, making the Poisson family a suitable modeling choice. Higher age groups showed significantly higher incidence rates, while healthier lifestyle scores (CLIstd) were negatively associated with new cases. Regional differences were observed in median BMI. These findings guided the subsequent model development, focusing on Poisson regression with population offsets and selected interaction terms.

Develop a model to examine the trends in new cases of pancreatic cancer across the recorded population variables and Analyse the model outputs and examine model performance:

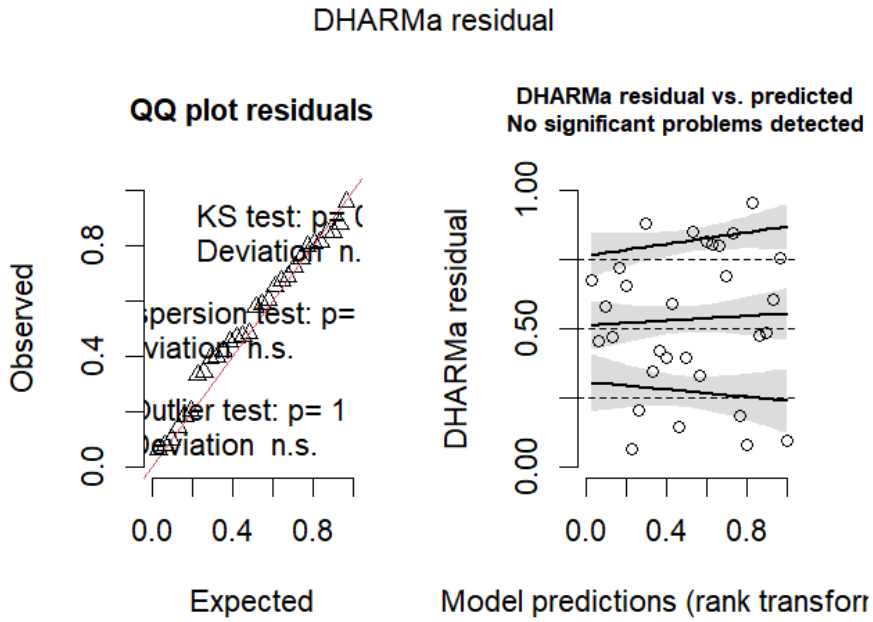
Since the outcome variable (NewCases) represents count data and is approximately Poisson-distributed, we used a Poisson GLM with a log link. The population size was included as an offset to account for varying subgroup sizes. For the first model we included all the predictors.

In Model 1 all predictors were included. AgeGroup and CLIstd were significant ($p < 0.05$), while BMI, region and SmokingPrevalence were not

```
Call:
glm(formula = NewCases ~ Sex + AgeGroup + Region + CLIstd + BMImedian +
    SmokingPrevalence, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -12.24995    1.10368 -11.099  < 2e-16 ***
Sex                 0.12479    0.10854   1.150  0.25028
AgeGroup40-59       0.09439    0.17133   0.551  0.58168
AgeGroup60-79       1.02073    0.15183   6.723 1.78e-11 ***
RegionRegion2       0.21867    0.19230   1.137  0.25549
RegionRegion3       0.18079    0.19382   0.933  0.35092
RegionRegion4      -0.04542    0.18872  -0.241  0.80981
RegionRegion5       0.23155    0.22066   1.049  0.29400
CLIstd             -0.32063    0.11642  -2.754  0.00589 **
BMImedian           0.07719    0.04176   1.849  0.06453 .
SmokingPrevalence   2.33964    1.64550   1.422  0.15507
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  20.459  on 19  degrees of freedom
AIC: 165.89

Number of Fisher Scoring iterations: 4

> BIC(model1)
[1] 181.3069
> rsq(model1, adj=TRUE)
[1] 0.7722274
```

DHARMa residual diagnostics showed no overdispersion and a uniform distribution of simulated residuals, indicating a good model fit.

```
> testDispersion(model1)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.86085, p-value = 0.728
alternative hypothesis: two.sided
```



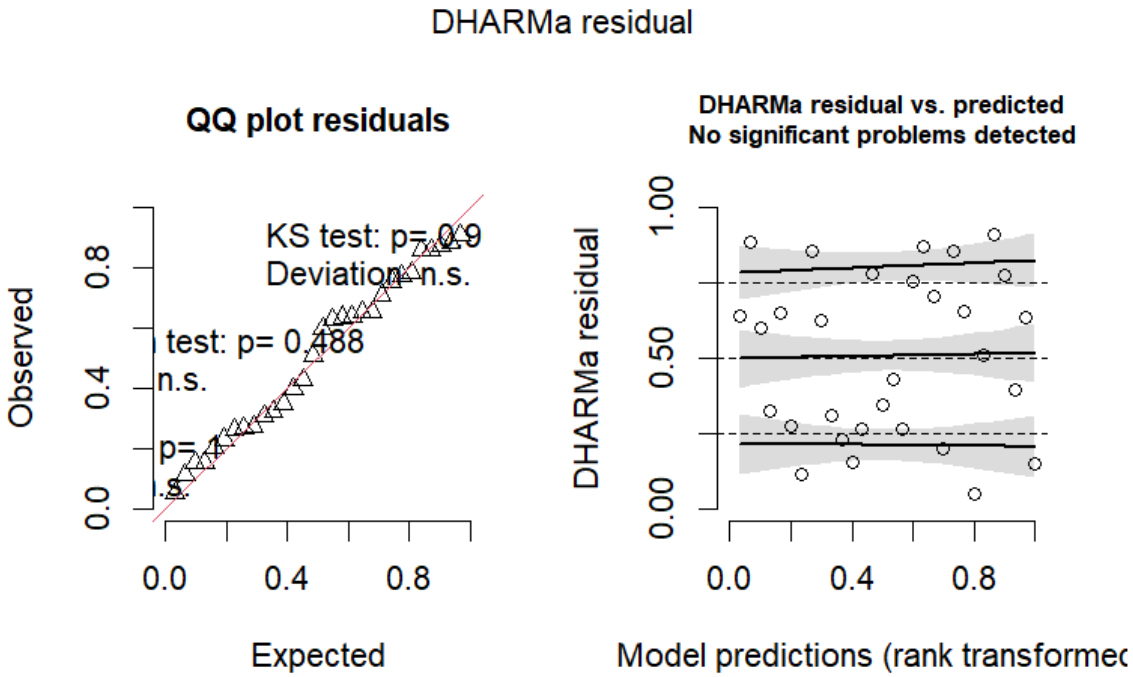
We moved on by trying CLIstd*Region as an interaction term.

In Model 2, only AgeGroup were significant ($p < 0.05$), while BMI, sex, region, CLIstd, SmokingPrevalence and the interaction term were not significant. The performance was clearly worse than Model 1.

```
Call:
glm(formula = NewCases ~ Sex + AgeGroup + CLIstd * Region + BMImedian +
    SmokingPrevalence, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -12.83625    1.94048  -6.615 3.72e-11 ***
Sex                    0.12488    0.11798   1.058    0.290
AgeGroup40-59          0.04086    0.18625   0.219    0.826
AgeGroup60-79          0.96615    0.18734   5.157 2.51e-07 ***
CLIstd                -0.78950    0.49850  -1.584    0.113
RegionRegion2         -0.02977    0.29447  -0.101    0.919
RegionRegion3          0.03496    0.25662   0.136    0.892
RegionRegion4         -0.23240    0.26257  -0.885    0.376
RegionRegion5          0.07642    0.28062   0.272    0.785
BMImedian              0.10476    0.07635   1.372    0.170
SmokingPrevalence      3.12162    1.91066   1.634    0.102
CLIstd:RegionRegion2   0.59176    0.54062   1.095    0.274
CLIstd:RegionRegion3   0.45336    0.49503   0.916    0.360
CLIstd:RegionRegion4   0.49343    0.63513   0.777    0.437
CLIstd:RegionRegion5   0.30808    0.46539   0.662    0.508
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  18.715  on 15  degrees of freedom
AIC: 172.15

Number of Fisher Scoring iterations: 4
```

```
> testDispersion(model2)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.72836, p-value = 0.488
alternative hypothesis: two.sided
```

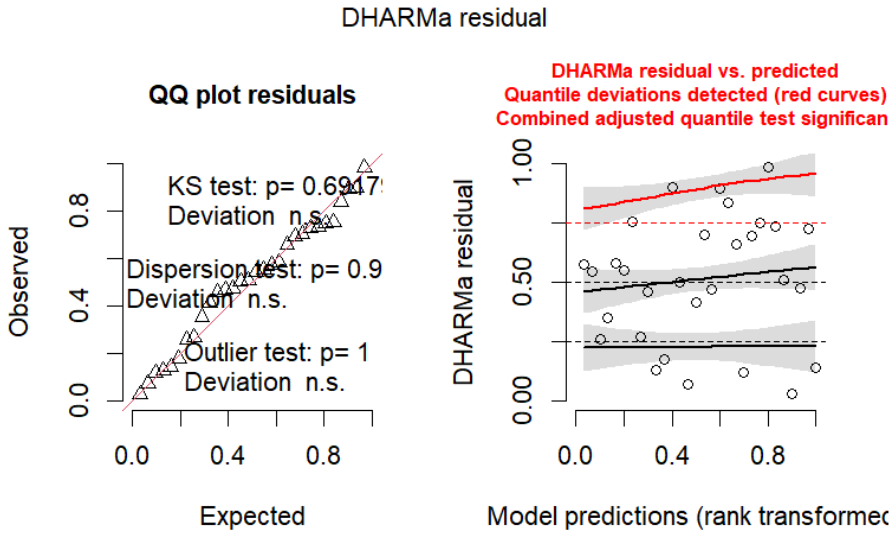
DHARMa residual diagnostics showed no overdispersion and a uniform distribution of simulated residuals, indicating a good model fit.



We try using CLIstd*Sex as an interaction term and we drop region as a predictor.
In Model 3, AgeGroup, BMI and SmokingPrevalence were significant ($p < 0.05$), while, sex, CLIstd, and the interaction term were not significant.

```
Call:
glm(formula = NewCases ~ AgeGroup + CLIstd * Sex + BMImedian +
    SmokingPrevalence, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -12.22217    0.96727 -12.636  < 2e-16 ***
AgeGroup40-59      0.10061    0.17204   0.585   0.5587
AgeGroup60-79      0.95984    0.16431   5.842 5.16e-09 ***
CLIstd            -0.20450    0.14733  -1.388   0.1651
Sex                0.17298    0.11164   1.549   0.1213
BMImedian          0.07565    0.03726   2.030   0.0423 *
SmokingPrevalence  3.19854    1.35674   2.358   0.0184 *
CLIstd:Sex        -0.25151    0.22087  -1.139   0.2548
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  22.637  on 22  degrees of freedom
AIC: 162.07

Number of Fisher Scoring iterations: 4
```

```
> testDispersion(model3)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.95334, p-value = 0.936
alternative hypothesis: two.sided
```

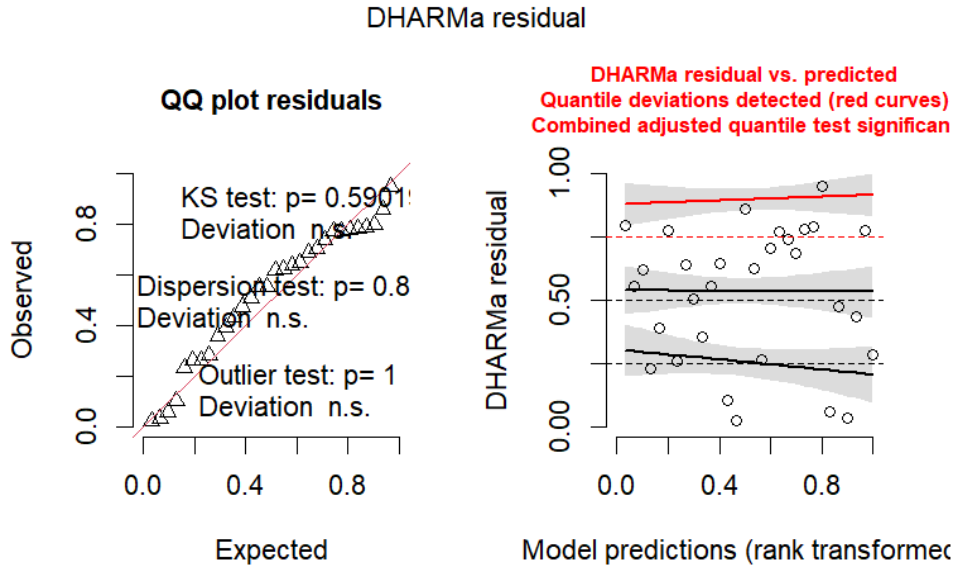


There is no overdispersion and e residuals seem normally distributed but we get some quantile deviation indicating a worse fit.

We continue our model fitting with CLIstd*BMI as an interaction term.
In Model 4, AgeGroup and SmokingPrevalence were significant ($p < 0.05$), while sex, BMI, region, CLIstd, and the interaction term were not significant.

```
Call:
glm(formula = NewCases ~ AgeGroup + CLIstd * BMImedian + Sex +
    SmokingPrevalence, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -12.23217    0.97542 -12.540  < 2e-16 ***
AgeGroup40-59      0.12090    0.17311   0.698  0.48491
AgeGroup60-79      1.01068    0.15146   6.673 2.51e-11 ***
CLIstd            -1.91563    1.34035  -1.429  0.15295
BMImedian          0.07047    0.03605   1.955  0.05059 .
Sex                0.12376    0.10969   1.128  0.25921
SmokingPrevalence  3.80180    1.40104   2.714  0.00666 **
CLIstd:BMImedian   0.06538    0.05462   1.197  0.23128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  22.497  on 22  degrees of freedom
AIC: 161.93

Number of Fisher Scoring iterations: 4
```

```
> testDispersion(model4)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.9045, p-value = 0.848
alternative hypothesis: two.sided
```



There is no overdispersion and the residuals seem less normally distributed and we get some quantile deviation indicating a worse fit.

Now we can plot the different model parameters in a matrix for better visibility.

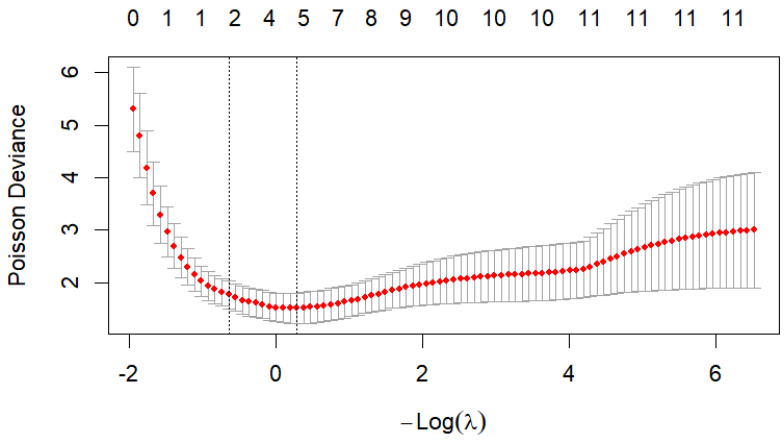
```
   Model      AIC      BIC   Rsq_adj
1 model1 165.8937 181.3069 0.7722274
2 model2 172.1497 193.1677 0.7580223
3 model3 162.0713 173.2809 0.7805745
4 model4 161.9316 173.1412 0.7893618
```

AIC evaluates how well your model fits the data while penalizing complexity (the number of parameters).
BIC - Similar to AIC but penalizes complexity more strongly,
R^2 - The proportion of variation in the response variable that your model explains.

We see that Model 3 and Model 4 have the best performance.

We applied LASSO regularization to identify the most relevant predictors and reduce potential multicollinearity between correlated variables

```
> cv_lasso <- cv.glmnet(X, y,
+                       family = "poisson",
+                       offset = offset_var,
+                       alpha = 1)  # alpha = 1 -> LASSO
> plot(cv_lasso)
> cv_lasso$lambda.min     # lambda that minimizes cross-validation error
[1] 0.7503966
> coef(cv_lasso, s = "lambda.min")
12 x 1 sparse Matrix of class "dgCMatrix"
                    lambda.min
(Intercept)       -9.882868809
AgeGroup40-59      .
AgeGroup60-79      0.967443219
RegionRegion2      0.070845699
RegionRegion3      .
RegionRegion4     -0.004433446
RegionRegion5      .
CLIstd            -0.153208338
BMImedian          .
Sex                .
SmokingPrevalence  0.824450956
CLIstd:BMImedian   .
```

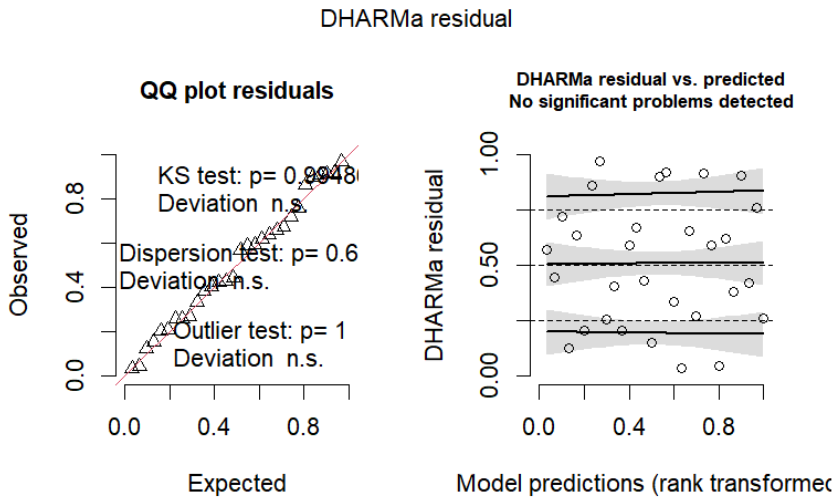


Moving right -> stronger penalty (simpler model, fewer nonzero coefficients).
y-axis -The cross-validated error. It shows how well the model performs on unseen data for each λ.

```
Call:
glm(formula = NewCases ~ AgeGroup + CLIstd + SmokingPrevalence +
    Region, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -10.22060    0.29818 -34.277  < 2e-16 ***
AgeGroup40-59      0.11090    0.17258   0.643   0.5205
AgeGroup60-79      1.11941    0.14376   7.787 6.88e-15 ***
CLIstd            -0.22617    0.10083  -2.243   0.0249 *
SmokingPrevalence  1.90042    1.61120   1.180   0.2382
RegionRegion2      0.22945    0.18856   1.217   0.2237
RegionRegion3      0.08436    0.18807   0.449   0.6538
RegionRegion4     -0.07578    0.18747  -0.404   0.6861
RegionRegion5      0.07318    0.19940   0.367   0.7136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  24.758  on 21  degrees of freedom
AIC: 166.19

Number of Fisher Scoring iterations: 4
```

LASSO selected AgeGroup and CLIstd as main predictors, supporting previous GLM findings

```
> BIC(model5)
[1] 178.8029
> rsq(model5, adj=TRUE)
[1] 0.8035233
> testDispersion(model5)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.82572, p-value = 0.616
alternative hypothesis: two.sided
```

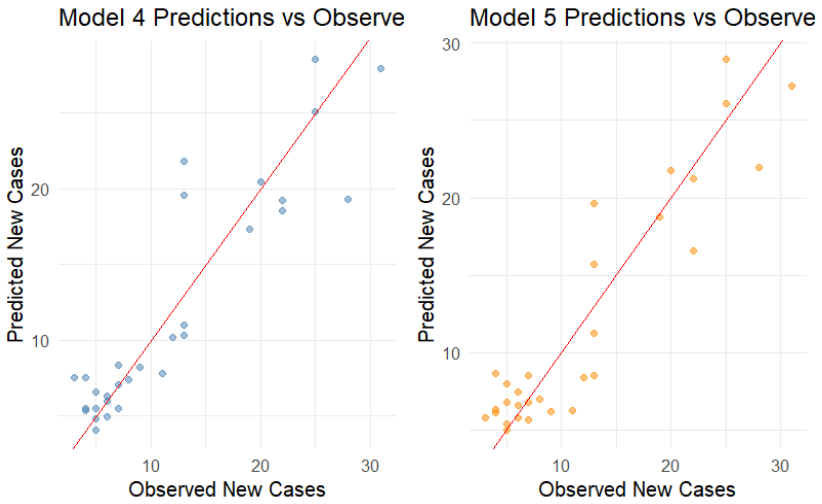


In the LASSO model, the interaction term (CLIstd:BMImedian) was penalized to zero, meaning the algorithm considered it non-essential for improving predictive accuracy under cross-validation. However, in model 4, including this interaction improved model fit. This difference arises because LASSO prioritizes model simplicity and generalization, whereas standard regression prioritizes in-sample explanatory power.

If we try an elastic net we can see that the model actually keeps the interaction term, however it is practically zero.

```
> #We try using Elastic net
> set.seed(123)
> cv_elastic <- cv.glmnet(
+   X, y,
+   family = "poisson",
+   offset = offset_var,
+   alpha = 0.5,  # Elastic Net
+   nfolds = 10
+ )
> # Coefficients
> coef(cv_elastic, s = "lambda.min")
12 x 1 sparse Matrix of class "dgCMatrix"
                    lambda.min
(Intercept)       -9.868209685
AgeGroup40-59      .
AgeGroup60-79      0.927054773
RegionRegion2      0.076877188
RegionRegion3      .
RegionRegion4     -0.016654776
RegionRegion5      .
CLIstd            -0.139097022
BMImedian          .
Sex                .
SmokingPrevalence  0.894556968
CLIstd:BMImedian  -0.001008695
```

Finally we use the two best models (4 and 5) to see how well they can predict new cases in pancreatic cancer.



Overall, this analysis highlights the strong influence of age and lifestyle on pancreatic cancer incidence and suggests that promoting healthier lifestyles could have a measurable impact on reducing cancer risk. However, given the cross-sectional design, causality cannot be established, and future research should extend these findings using longitudinal data and additional behavioral or genetic covariates.

task2

zhexuan

2025-11-14

Task 2: Tumor Models in Mice for Studying Treatment Effect

The dataset consists of 60 mice with repeated tumor volume measurements (for 90 days). Each mouse belongs to one of two disease models: induced tumors or xenograft mice, and is assigned either to a control or treatment group receiving the compound ATC10X.

set up

```
dat <- read.csv("Data_T2.csv")

str(dat)
```

```
## 'data.frame':    5460 obs. of  6 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Time     : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ DV       : num  481 525 474 538 519 ...
##  $ Treatment: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Model    : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(dat)
```

```
##        X              ID             Time           DV           Treatment
##  Min.   :   1   Min.   : 1.00   Min.   : 0   Min.   : 361.8   Min.   :0.0
##  1st Qu.:1366   1st Qu.:15.75   1st Qu.:22   1st Qu.: 559.2   1st Qu.:0.0
##  Median :2730   Median :30.50   Median :45   Median : 635.6   Median :0.5
##  Mean   :2730   Mean   :30.50   Mean   :45   Mean   : 673.2   Mean   :0.5
##  3rd Qu.:4095   3rd Qu.:45.25   3rd Qu.:68   3rd Qu.: 760.3   3rd Qu.:1.0
##  Max.   :5460   Max.   :60.00   Max.   :90   Max.   :1381.5   Max.   :1.0
##      Model
##  Min.   :0.0
##  1st Qu.:0.0
##  Median :1.0
##  Mean   :0.6
##  3rd Qu.:1.0
##  Max.   :1.0
```

```
dat <- dat %>%
  mutate(
    ID        = factor(ID),
    Treatment = factor(Treatment, levels = c(0, 1),
                       labels = c("Control", "Treatment")),
    Model     = factor(Model, levels = c(0, 1),
                       labels = c("Induced", "Xenograft")),
    logDV     = log(DV)
  )

summary(dat$DV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   361.8   559.2   635.6   673.2   760.3  1381.5
```

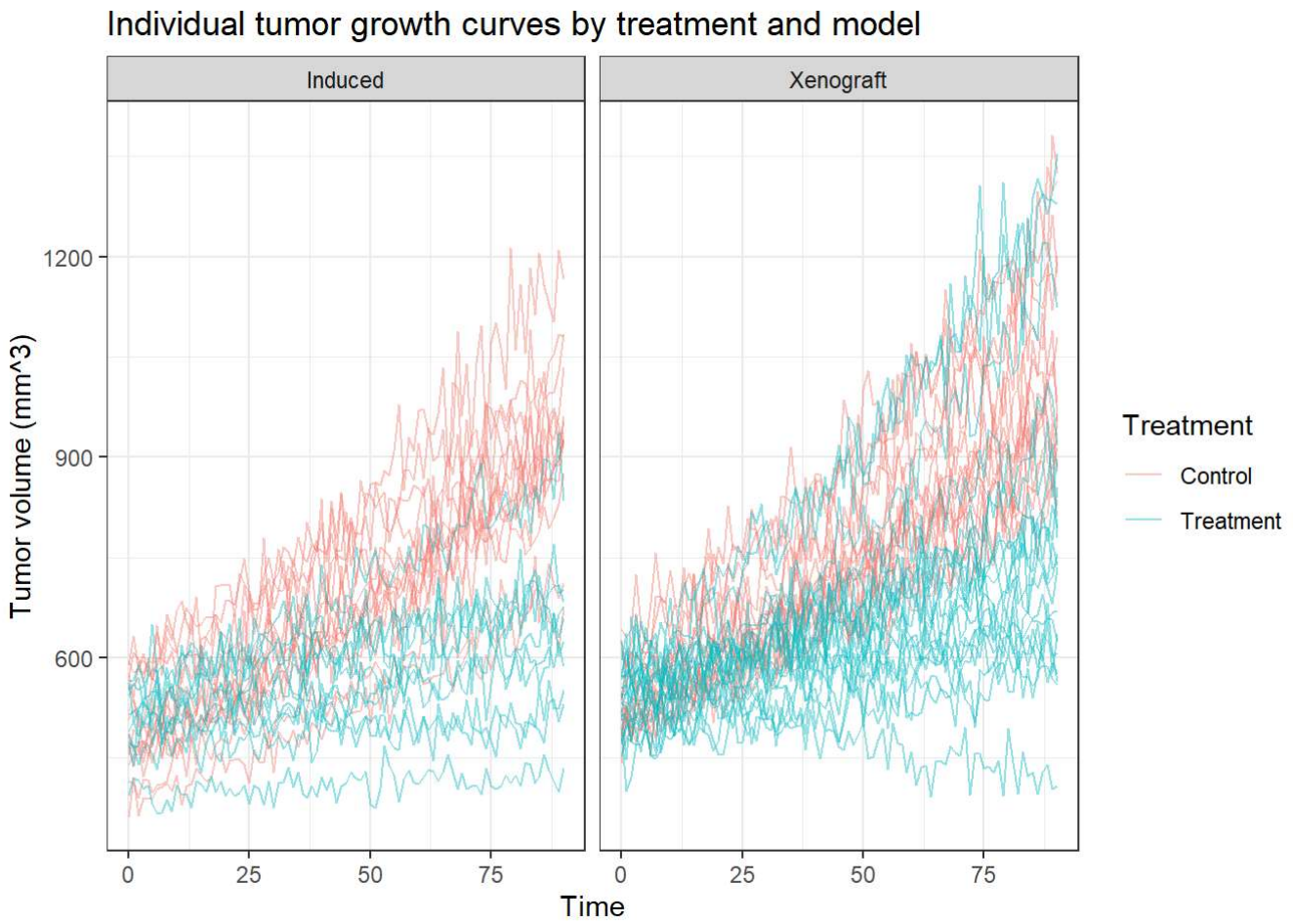
```
summary(dat$logDV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.891   6.327   6.455   6.485   6.634   7.231
```

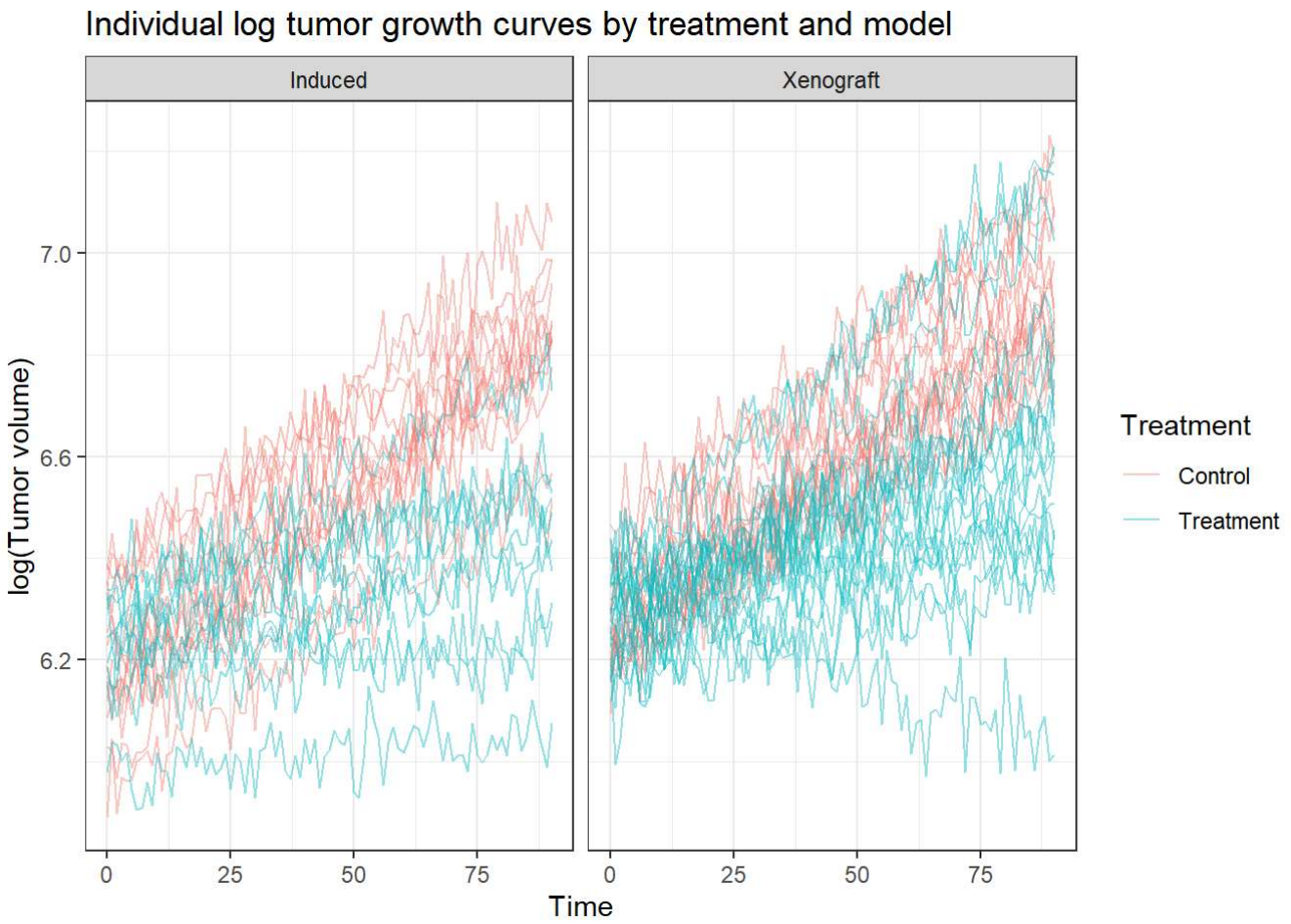
visualization

Initial visualisations of raw DV and log-transformed DV showed that tumor growth is broadly exponential on the original scale but near-linear on the log scale. Average curves further suggested that tumors grew over time in all groups, with notably slower growth in the treatment group, and with broadly similar growth patterns across the two mouse models.

```
# raw DV
ggplot(dat, aes(x = Time, y = DV, group = ID,
                colour = Treatment)) +
  geom_line(alpha = 0.4) +
  facet_wrap(~ Model) +
  scale_y_continuous("Tumor volume (mm^3)") +
  labs(title = "Individual tumor growth curves by treatment and model") +
  theme_bw()
```

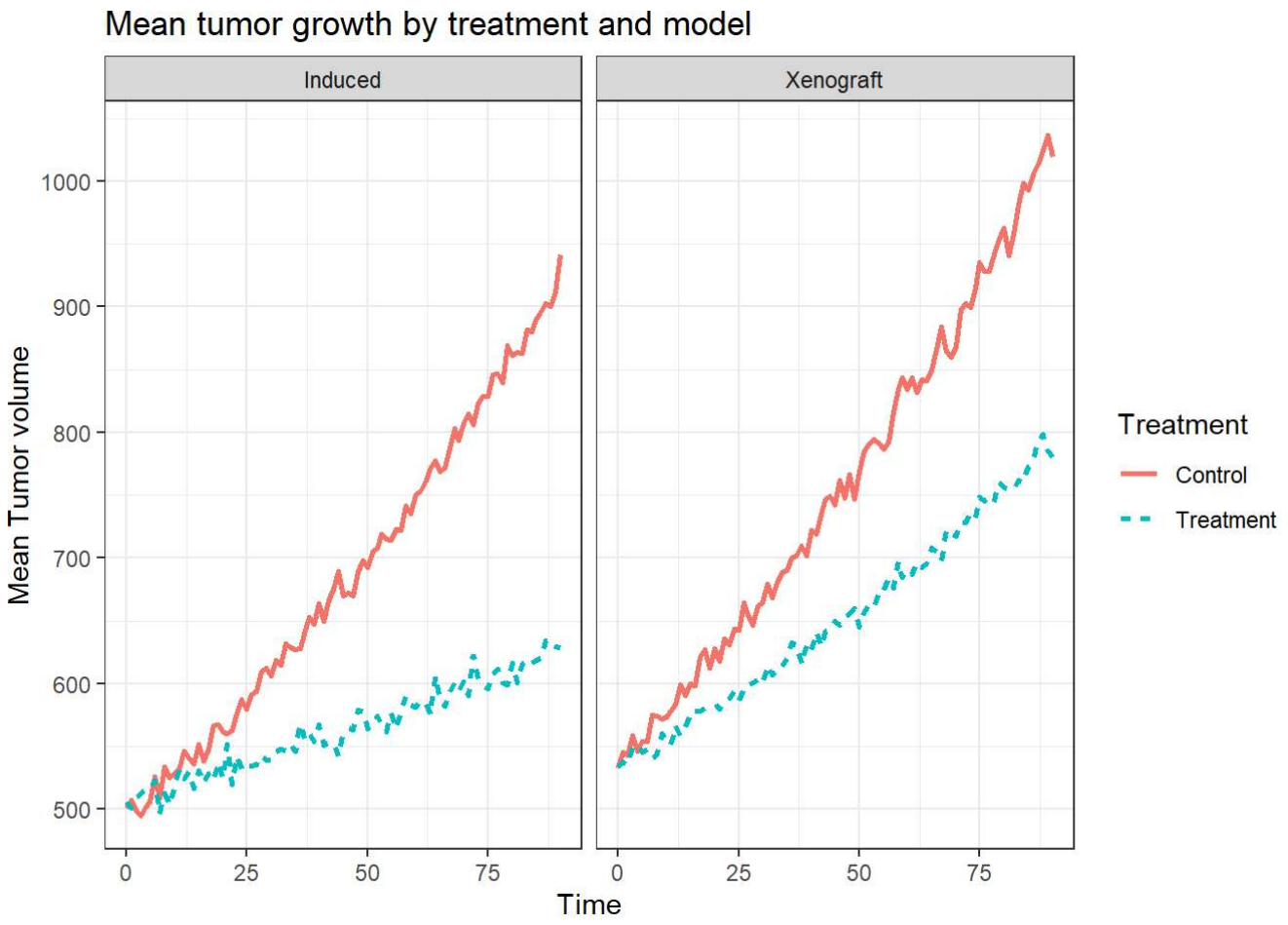



```
# log(DV)
ggplot(dat, aes(x = Time, y = logDV, group = ID,
                colour = Treatment)) +
  geom_line(alpha = 0.4) +
  facet_wrap(~ Model) +
  scale_y_continuous("log(Tumor volume)") +
  labs(title = "Individual log tumor growth curves by treatment and model") +
  theme_bw()
```

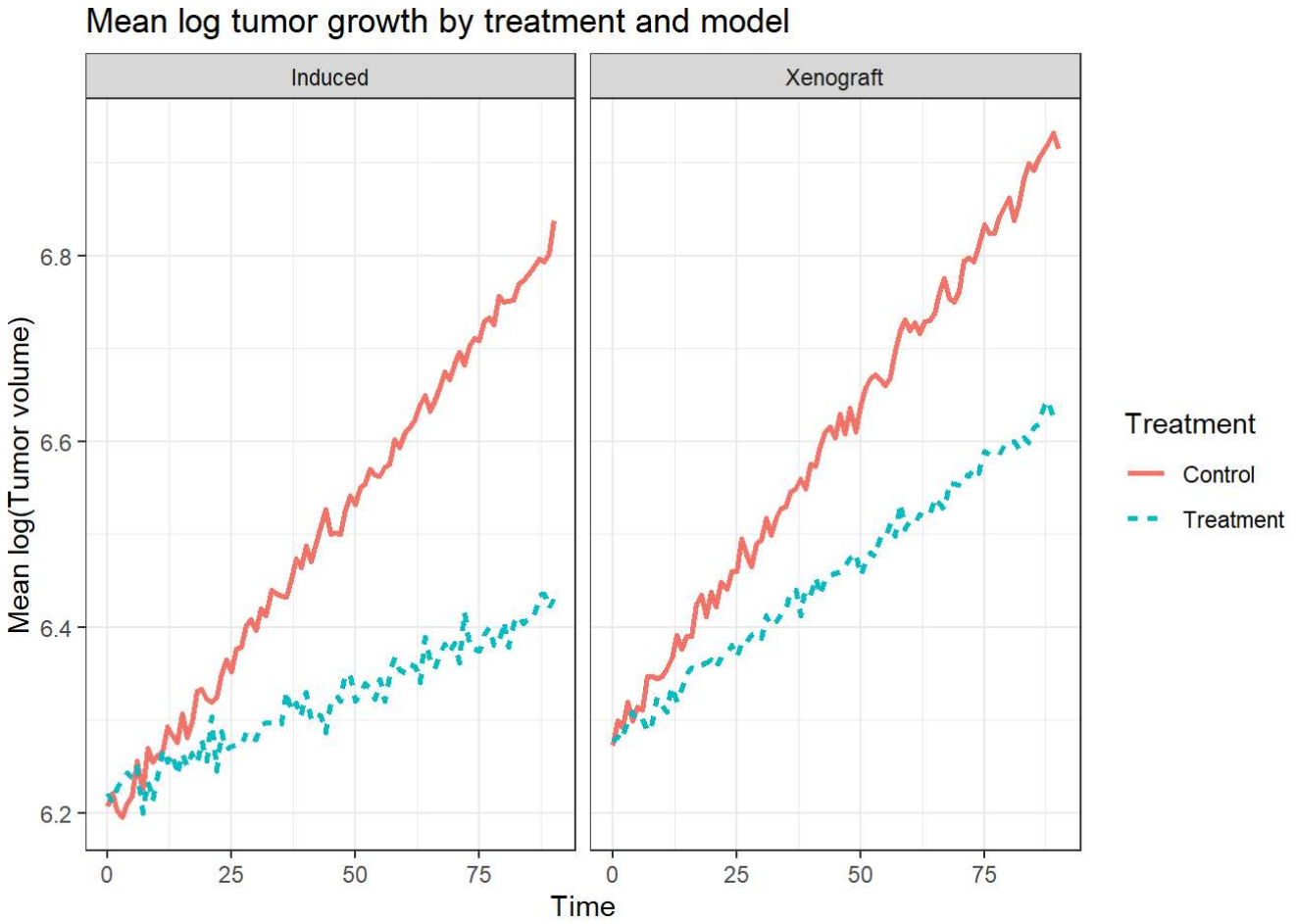



```
ggplot(dat, aes(x = Time, y = DV,
                colour = Treatment, linetype = Treatment)) +
  stat_summary(fun = mean, geom = "line", size = 1) +
  facet_wrap(~ Model) +
  labs(y = "Mean Tumor volume",
       title = "Mean tumor growth by treatment and model") +
  theme_bw()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
ggplot(dat, aes(x = Time, y = logDV,
                colour = Treatment, linetype = Treatment)) +
  stat_summary(fun = mean, geom = "line", size = 1) +
  facet_wrap(~ Model) +
  labs(y = "Mean log(Tumor volume)",
       title = "Mean log tumor growth by treatment and model") +
  theme_bw()
```



linear mixed-effects model

To stabilize variance and linearize growth, tumor volume was log-transformed before modelling.

Linear mixed-effects models were appropriate here because the goal was to characterise tumor growth over time while accounting for repeated measurements.

Model 1: only time factor

This baseline model captures average growth over time while allowing each mouse to have its own intercept.

```
m1 <- lmer(logDV ~ Time + (1 | ID), data = dat)
summary(m1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: logDV ~ Time + (1 | ID)
##    Data: dat
##
## REML criterion at convergence: -11011.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7812 -0.6267 -0.0084  0.6217  4.9920
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.026651 0.16325
##  Residual             0.007281 0.08533
## Number of obs: 5460, groups:  ID, 60
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 6.251e+00  2.120e-02 6.004e+01   294.9   <2e-16 ***
## Time        5.205e-03  4.396e-05 5.399e+03   118.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## Time -0.093
```

Model 2: time x treatment

This model evaluates whether treatment affects baseline volume or growth rate.

```
m2 <- lmer(logDV ~ Time * Treatment + (1 | ID), data = dat)
summary(m2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: logDV ~ Time * Treatment + (1 | ID)
##    Data: dat
##
## REML criterion at convergence: -13111.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7869 -0.5578  0.0017  0.5414  5.0000
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.021543 0.1468
##  Residual             0.004928 0.0702
## Number of obs: 5460, groups:  ID, 60
##
## Fixed effects:
##                            Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)              6.242e+00  2.693e-02  5.886e+01 231.782   <2e-16 ***
## Time                     7.042e-03  5.115e-05  5.398e+03 137.676   <2e-16 ***
## TreatmentTreatment       1.822e-02  3.808e-02  5.886e+01   0.478    0.634
## Time:TreatmentTreatment -3.674e-03  7.233e-05  5.398e+03 -50.789   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Time   TrtmnT
## Time        -0.085
## TrtmntTrtmn -0.707  0.060
## Tm:TrtmntTr  0.060 -0.707 -0.085
```

```
anova(m1, m2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: dat
## Models:
## m1: logDV ~ Time + (1 | ID)
## m2: logDV ~ Time * Treatment + (1 | ID)
##    npar    AIC    BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
## m1    4 -11027 -11001 5517.7    -11035
## m2    6 -13146 -13107 6579.2    -13158  2123  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 3: time x treatment x model

This extends m2 by testing whether disease model influences growth or treatment effects.

```
m3 <- lmer(logDV ~ Time * Treatment * Model + (1 | ID), data = dat)
summary(m3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: logDV ~ Time * Treatment * Model + (1 | ID)
##    Data: dat
##
## REML criterion at convergence: -13338.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.1913 -0.5747 -0.0055  0.5531  5.4582
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.018788 0.13707
##  Residual             0.004699 0.06855
## Number of obs: 5460, groups:  ID, 60
##
## Fixed effects:
##                                      Estimate Std. Error         df t value
## (Intercept)                          6.193e+00  3.683e-02  5.690e+01 168.151
## Time                                 6.938e-03  7.312e-05  5.396e+03  94.885
## TreatmentTreatment                   2.747e-02  5.706e-02  5.690e+01   0.482
## ModelXenograft                       9.107e-02  5.043e-02  5.690e+01   1.806
## Time:TreatmentTreatment             -4.709e-03  1.133e-04  5.396e+03 -41.576
## Time:ModelXenograft                  1.949e-04  1.001e-04  5.396e+03   1.947
## TreatmentTreatment:ModelXenograft   -3.209e-02  7.343e-02  5.690e+01  -0.437
## Time:TreatmentTreatment:ModelXenograft  1.515e-03  1.458e-04  5.396e+03  10.390
##                                     Pr(>|t|)
## (Intercept)                           <2e-16 ***
## Time                                  <2e-16 ***
## TreatmentTreatment                    0.6320
## ModelXenograft                        0.0762 .
## Time:TreatmentTreatment               <2e-16 ***
## Time:ModelXenograft                   0.0516 .
## TreatmentTreatment:ModelXenograft     0.6637
## Time:TreatmentTreatment:ModelXenograft   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Time   TrtmnT MdlXng Tm:TrT Tm:MdX TrT:MX
## Time        -0.089
## TrtmntTrtmn -0.645  0.058
## ModelXngrft -0.730  0.065  0.471
## Tm:TrtmntTr  0.058 -0.645 -0.089 -0.042
## Tm:MdlXngrf  0.065 -0.730 -0.042 -0.089  0.471
## TrtmntTr:MX  0.502 -0.045 -0.777 -0.687  0.069  0.061
## Tm:TrtmT:MX -0.045  0.502  0.069  0.061 -0.777 -0.687 -0.089
```

```
anova(m2, m3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: dat
## Models:
## m2: logDV ~ Time * Treatment + (1 | ID)
## m3: logDV ~ Time * Treatment * Model + (1 | ID)
##    npar    AIC    BIC logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m2    6 -13146 -13107 6579.2    -13158
## m3   10 -13407 -13341 6713.4    -13427 268.36  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 4: Random slope model: time x treatment x model + (Time | ID))

This model further allows each mouse to have its own growth rate (random slope).

Anova tests demonstrated improvement across model steps, with a particularly large improvement from m3 to m4, confirming that individual-specific growth rates are essential for accurately modelling these data.

```
m4 <- lmer(logDV ~ Time * Treatment * Model + (Time | ID), data = dat)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00704387 (tol = 0.002, component 1)
```

```
summary(m4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: logDV ~ Time * Treatment * Model + (Time | ID)
##    Data: dat
##
## REML criterion at convergence: -17631.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2265 -0.6684  0.0051  0.6463  3.8377
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  ID       (Intercept) 1.031e-02 0.101516
##           Time        4.099e-06 0.002025 0.01
##  Residual             2.029e-03 0.045045
## Number of obs: 5460, groups:  ID, 60
##
## Fixed effects:
##                                      Estimate Std. Error         df t value
## (Intercept)                         6.1931944  0.0272466 56.0881214 227.301
## Time                                0.0069376  0.0005432 55.9830698  12.772
## TreatmentTreatment                  0.0274739  0.0422103 56.0881151   0.651
## ModelXenograft                      0.0910747  0.0373090 56.0881165   2.441
## Time:TreatmentTreatment            -0.0047094  0.0008415 55.9830728  -5.596
## Time:ModelXenograft                 0.0001949  0.0007438 55.9830723   0.262
## TreatmentTreatment:ModelXenograft  -0.0320936  0.0543227 56.0881133  -0.591
## Time:TreatmentTreatment:ModelXenograft  0.0015147  0.0010830 55.9830739   1.399
##                                     Pr(>|t|)
## (Intercept)                          < 2e-16 ***
## Time                                 < 2e-16 ***
## TreatmentTreatment                    0.5178
## ModelXenograft                        0.0178 *
## Time:TreatmentTreatment              6.84e-07 ***
## Time:ModelXenograft                   0.7942
## TreatmentTreatment:ModelXenograft     0.5570
## Time:TreatmentTreatment:ModelXenograft   0.1675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Time   TrtmnT MdlXng Tm:TrT Tm:MdX TrT:MX
## Time         0.004
## TrtmntTrtmn -0.645 -0.002
## ModelXngrft -0.730 -0.003  0.471
## Tm:TrtmntTr -0.002 -0.645  0.004  0.002
## Tm:MdlXngrf -0.003 -0.730  0.002  0.004  0.471
## TrtmntTr:MX  0.502  0.002 -0.777 -0.687 -0.003 -0.003
## Tm:TrtmT:MX  0.002  0.502 -0.003 -0.003 -0.777 -0.687  0.004
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00704387 (tol = 0.002, component 1)
```

```
anova(m3, m4)
```

```
## refitting model(s) with ML (instead of REML)
```

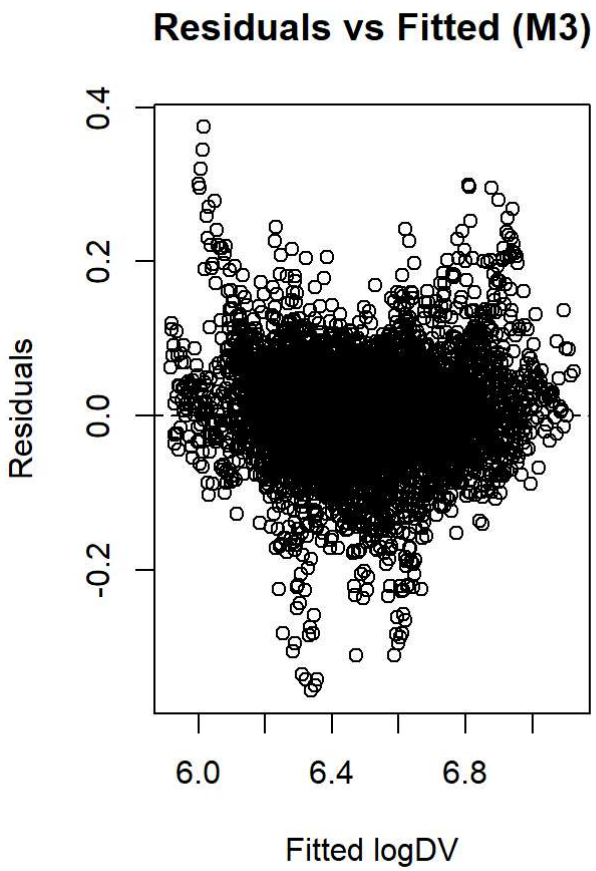
```
## Data: dat
## Models:
## m3: logDV ~ Time * Treatment * Model + (1 | ID)
## m4: logDV ~ Time * Treatment * Model + (Time | ID)
##    npar    AIC    BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
## m3   10 -13407 -13341 6713.4    -13427
## m4   12 -17683 -17604 8853.4    -17707  4280  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

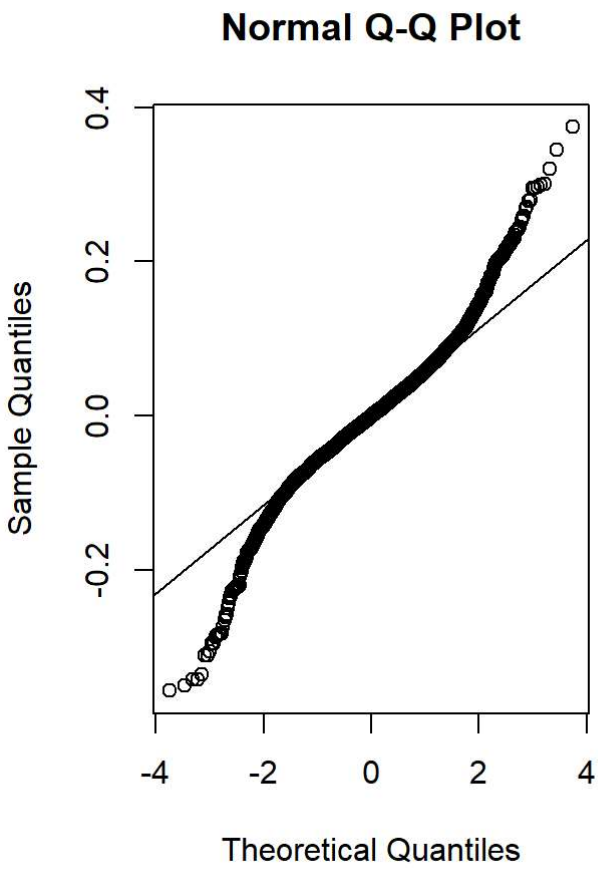
Residual diagnostics showed clear differences between m3 and m4.

For m3, the residuals-fitted plot showed systematic structure, and the Q–Q plot showed significant deviations in both tails, indicating poor adherence to model assumptions.

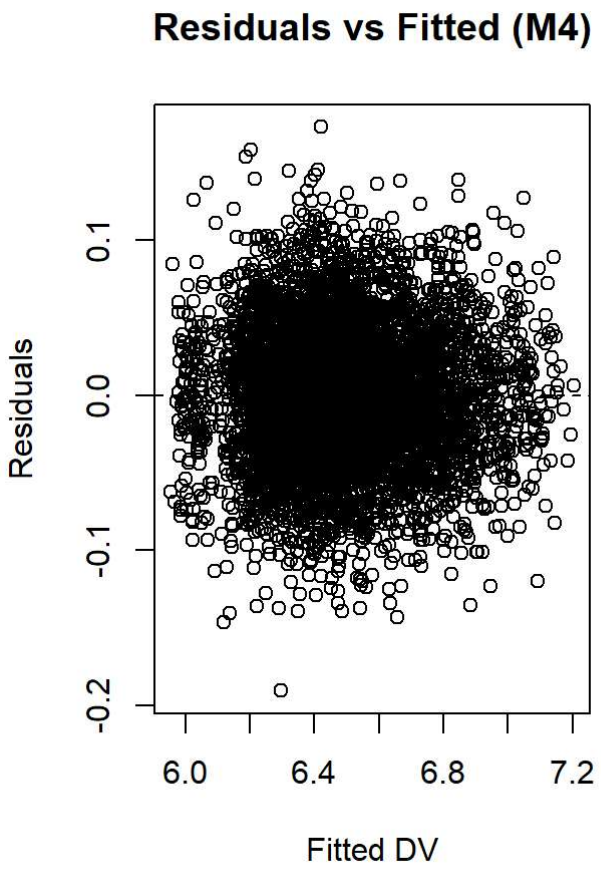
By contrast, m4 yielded centred, approximately homoscedastic residuals forming a symmetric cloud and a Q–Q plot closely following the reference line. This indicates that random slopes successfully captured within-mouse variability, leaving residuals that behave approximately as assumed (independent and normally distributed).

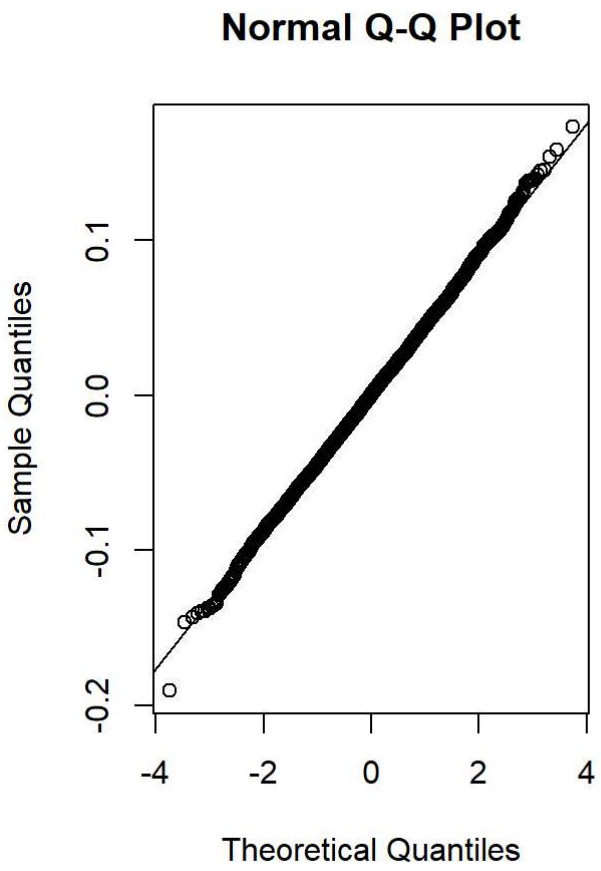
```
# residual vs fitted & qq plot
# m3
par(mfrow = c(1, 2))
plot(resid(m3) ~ fitted(m3),
     main = "Residuals vs Fitted (M3)", xlab = "Fitted logDV", ylab = "Residuals")
abline(h = 0, lty = 2)

qqnorm(resid(m3))
qqline(resid(m3))
```






```
par(mfrow = c(1, 1))


# m4
par(mfrow = c(1, 2))
plot(resid(m4) ~ fitted(m4),
     main = "Residuals vs Fitted (M4)", xlab = "Fitted DV", ylab = "Residuals")
abline(h = 0, lty = 2)

qqnorm(resid(m4))
qqline(resid(m4))
```






```
par(mfrow = c(1, 1))
```

Results and predictions & Impact of treatment and disease model

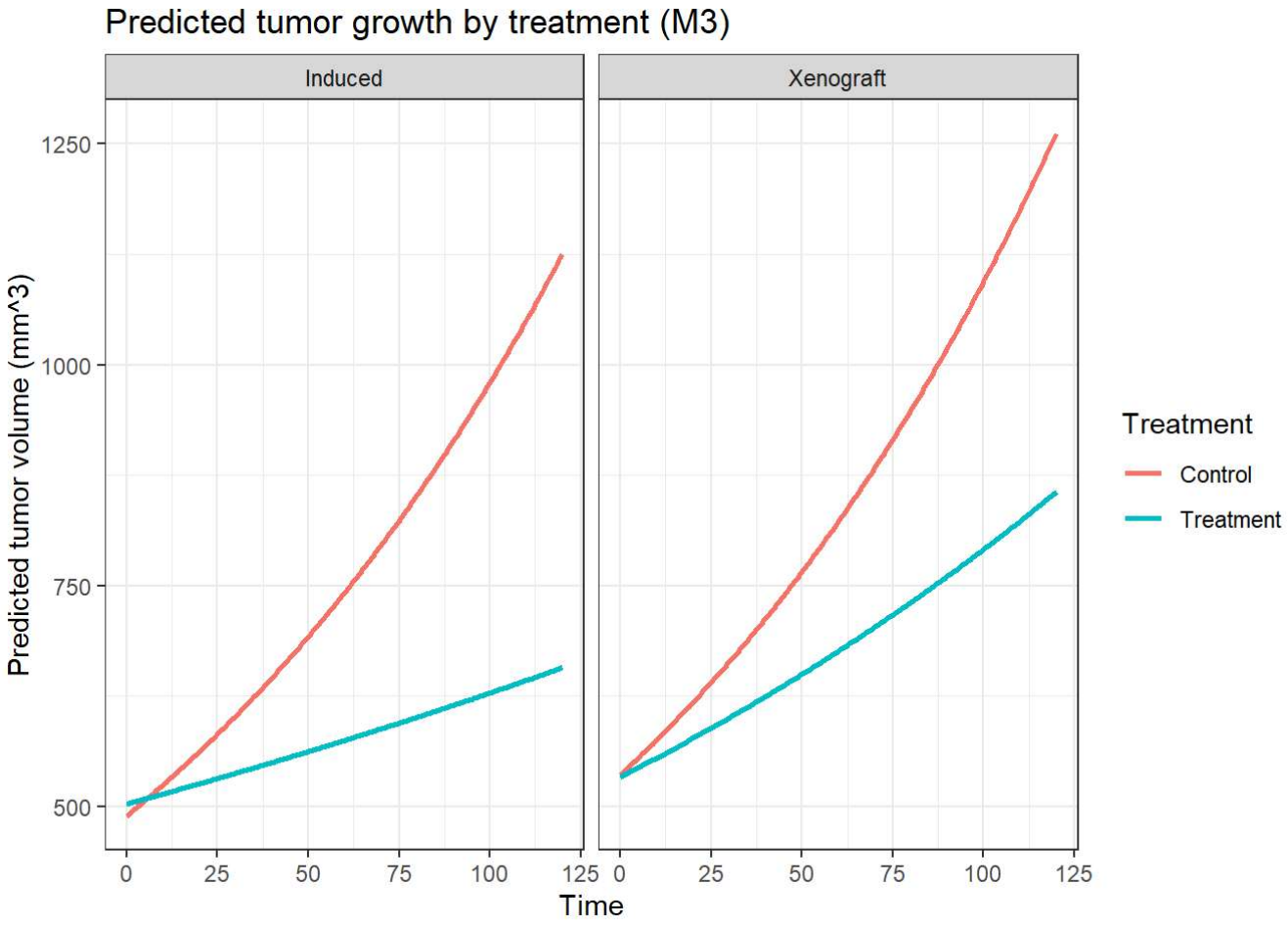
The anova(m4) table evaluates the significance of each fixed effect. The effect of Time, Model and interaction Time x Treatment were highly significant. And there was no significant difference in treatment effect among different models. Also, the treatment slows down the growth rate of the tumor, but this change doesn't depend on the model.

```
anova(m4)
```

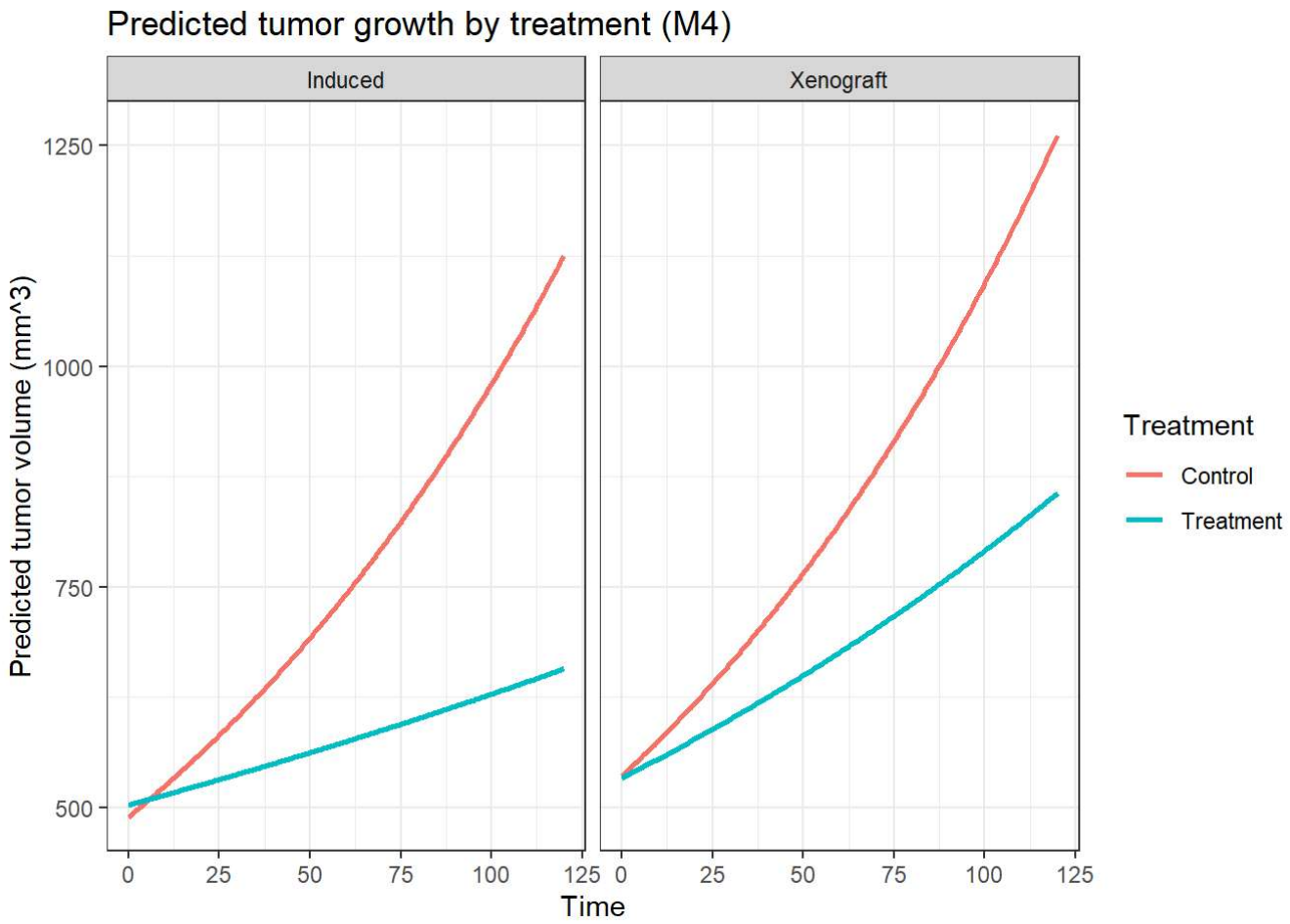
```
## Type III Analysis of Variance Table with Satterthwaite's method
##                       Sum Sq Mean Sq NumDF  DenDF  F value    Pr(>F)
## Time                 0.70840 0.70840     1 55.983 349.1299 < 2.2e-16 ***
## Treatment            0.00036 0.00036     1 56.088   0.1770  0.675573
## Model                0.01548 0.01548     1 56.088   7.6303  0.007746 **
## Time:Treatment       0.10808 0.10808     1 55.983  53.2640 1.114e-09 ***
## Time:Model           0.00627 0.00627     1 55.983   3.0925  0.084121 .
## Treatment:Model      0.00071 0.00071     1 56.088   0.3490  0.557031
## Time:Treatment:Model 0.00397 0.00397     1 55.983   1.9560  0.167462
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We did interpolation and extrapolation up to 120 days to predict results of models. The predicted tumor growth from m3 and m4 produced same results because both models have the same fixed-effects structure. They differ only in how they model specific variability. It is obvious to see that the increase rate of tumor volume in the treatment group of the induced tumor mouse model is much lower than that in the xenograft mouse model.

```
# interpolation & extrapolation
time_seq <- seq(from = min(dat$Time), to = 120, by = 0.5)

newdat <- expand.grid(
  Time      = time_seq,
  Treatment = levels(dat$Treatment),
  Model     = levels(dat$Model)
)

newdat$pred_logDV <- predict(m3, newdata = newdat, re.form = NA)
newdat$pred_DV <- exp(newdat$pred_logDV)

ggplot(newdat, aes(Time, pred_DV, colour = Treatment)) +
  geom_line(size = 1) +
  facet_wrap(~ Model) +
  labs(y = "Predicted tumor volume (mm^3)",
       title = "Predicted tumor growth by treatment (M3)") +
  theme_bw()
```




```
newdat$pred_logDV <- predict(m4, newdata = newdat, re.form = NA)
newdat$pred_DV <- exp(newdat$pred_logDV)

ggplot(newdat, aes(Time, pred_DV, colour = Treatment)) +
  geom_line(size = 1) +
  facet_wrap(~ Model) +
  labs(y = "Predicted tumor volume (mm^3)",
       title = "Predicted tumor growth by treatment (M4)") +
  theme_bw()
```

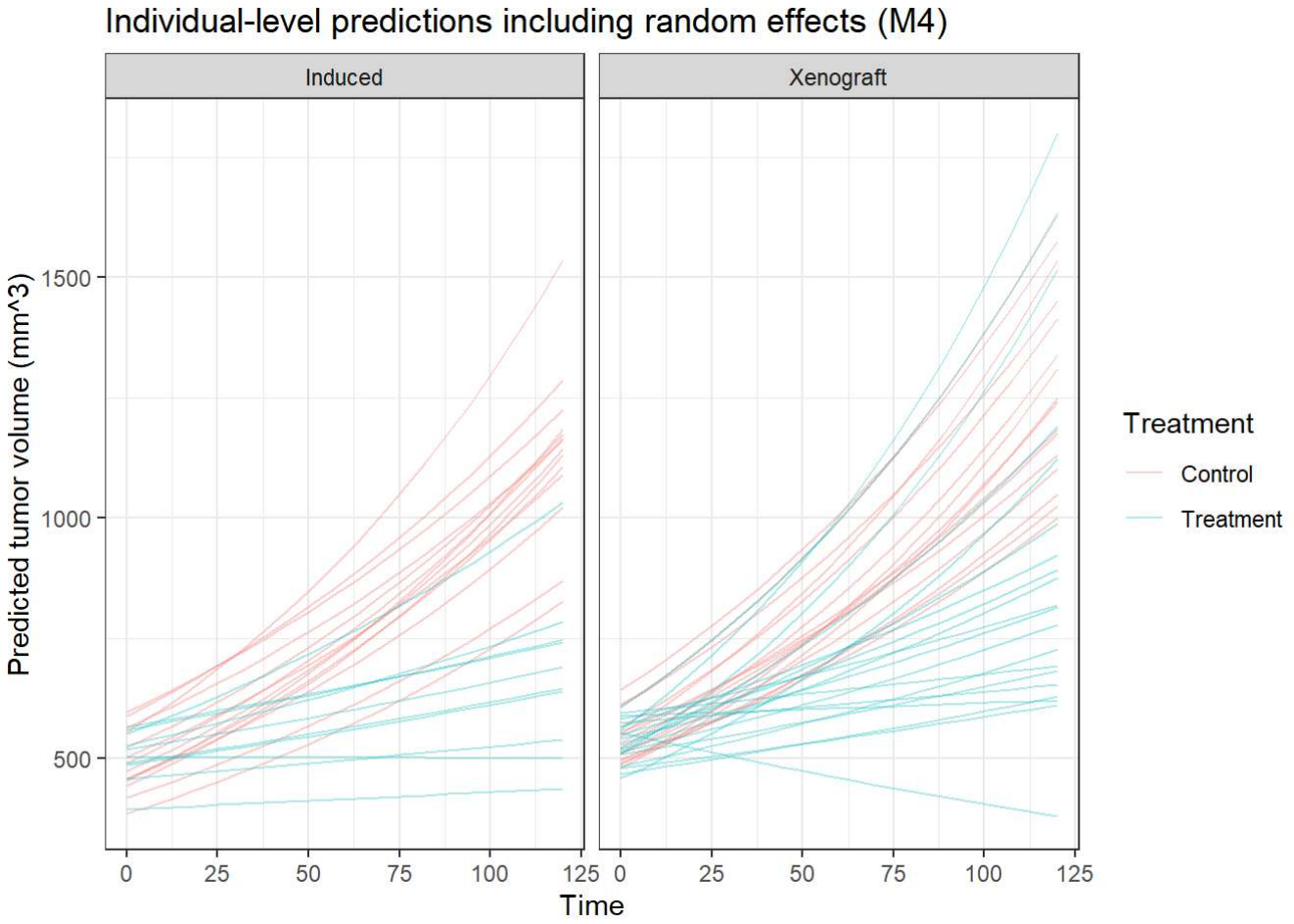


To illustrate how m3 and m4 differ in their handling of individual mice, predictions were generated for new data with 60 subjects. Under m3, all mice shared the same growth rate and differed only in their intercepts, resulting in parallel trajectories. In contrast, m4 produced curves with differing slopes as well as intercepts, reflecting its random-slope structure. This more flexible pattern can align better with the observed variability in tumor progression.

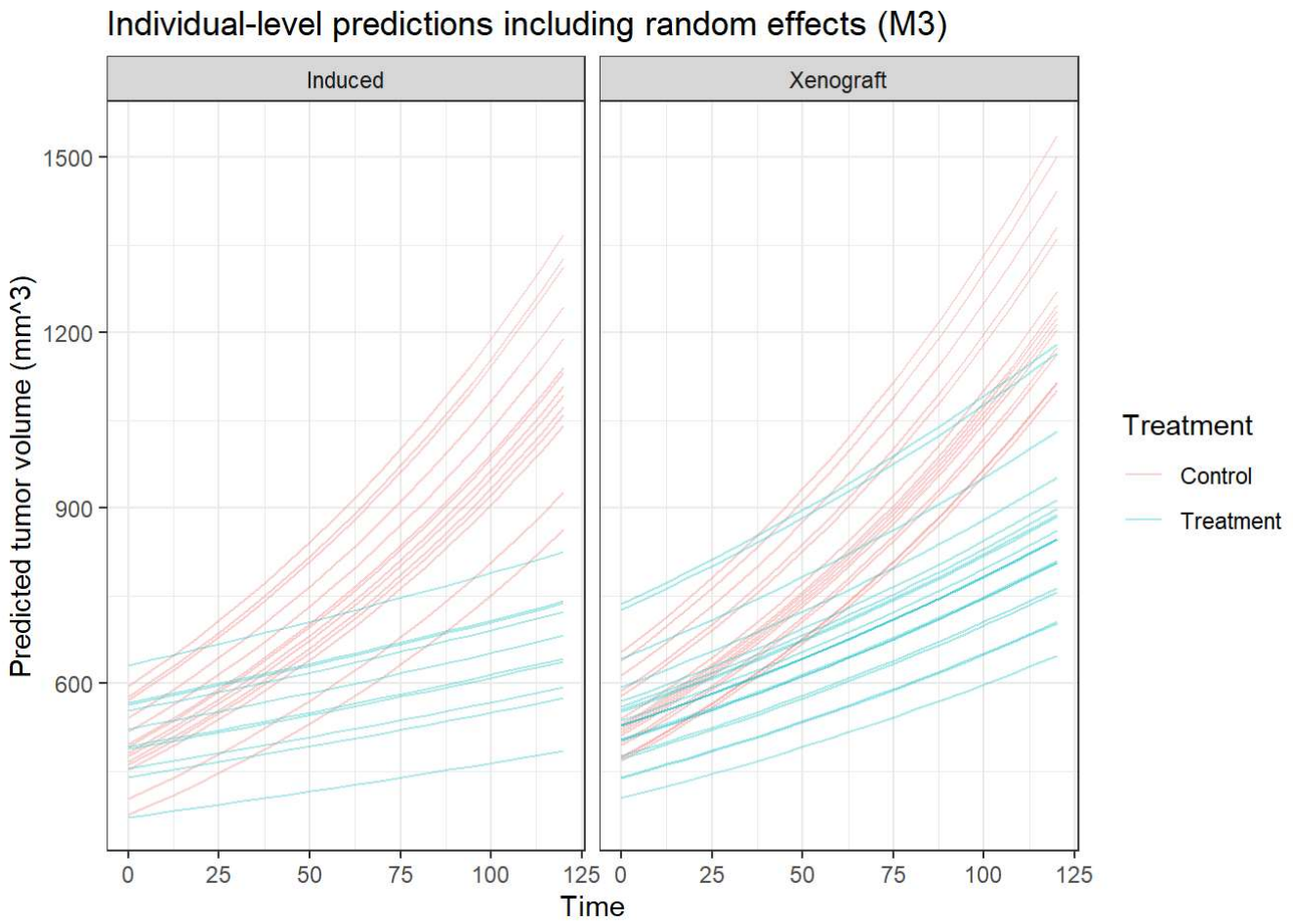
```
# new data with ID
newdat_id <- dat %>%
  group_by(ID, Treatment, Model) %>%
  summarise(Time = seq(min(Time), 120, length.out = 20),
            .groups = "drop")
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## ℹ Please use `reframe()` instead.
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
newdat_id$pred_logDV <- predict(m4, newdata = newdat_id, re.form = NULL)
newdat_id$pred_DV <- exp(newdat_id$pred_logDV)

ggplot(newdat_id, aes(Time, pred_DV, group = ID, colour = Treatment)) +
  geom_line(alpha = 0.3) +
  facet_wrap(~ Model) +
  labs(y = "Predicted tumor volume (mm^3)",
       title = "Individual-level predictions including random effects (M4)") +
  theme_bw()
```




```
newdat_id$pred_logDV <- predict(m3, newdata = newdat_id, re.form = NULL)
newdat_id$pred_DV <- exp(newdat_id$pred_logDV)

ggplot(newdat_id, aes(Time, pred_DV, group = ID, colour = Treatment)) +
  geom_line(alpha = 0.3) +
  facet_wrap(~ Model) +
  labs(y = "Predicted tumor volume (mm^3)",
       title = "Individual-level predictions including random effects (M3)") +
  theme_bw()
```



The estimated trends from emtrends(m4) further quantified these differences. Treated mice showed a consistently smaller Time slope than controls, confirming that ATC10X slows tumor growth on the log scale. The two tumor models exhibited slightly different slopes, with induced tumors growing marginally slower than xenografts in the absence of treatment and displaying a slightly larger reduction under treatment.

Overall, ATC10X suppresses the growth of tumors, and the two disease models differ slightly in both baseline size and growth dynamics. The fixed effects reveal similar conclusions across models, while the random-slope structure of m4 provides a more realistic representation of the heterogeneous growth rates observed in individual mice.

```
emtrends(m4, ~ Treatment | Model, var = "Time")
```

```
## Note: D.f. calculations have been disabled because the number of observations exceeds 3000.
## To enable adjustments, add the argument 'pbkrtest.limit = 5460' (or larger)
## [or, globally, 'set emm_options(pbkrtest.limit = 5460)' or larger];
## but be warned that this may result in large computation time and memory use.
```

```
## Note: D.f. calculations have been disabled because the number of observations exceeds 3000.
## To enable adjustments, add the argument 'lmerTest.limit = 5460' (or larger)
## [or, globally, 'set emm_options(lmerTest.limit = 5460)' or larger];
## but be warned that this may result in large computation time and memory use.
```

```
## Model = Induced:
##  Treatment Time.trend       SE  df asymp.LCL asymp.UCL
##  Control      0.00694 0.000543 Inf  0.005873   0.00800
##  Treatment    0.00223 0.000643 Inf  0.000968   0.00349
##
## Model = Xenograft:
##  Treatment Time.trend       SE  df asymp.LCL asymp.UCL
##  Control      0.00713 0.000508 Inf  0.006137   0.00813
##  Treatment    0.00394 0.000454 Inf  0.003047   0.00483
##
## Degrees-of-freedom method: asymptotic
## Confidence level used: 0.95
```

The ability to carry out predictions, within and beyond the ranges of the data

To evaluate predictive performance within the observed data range, overall predictions were generated using the fixed effects of m3 and m4. For both m3 and m4, the root-mean-square error on the log scale was approximately 0.15, corresponding to an average prediction error of around 16% on the original tumor volume scale. This indicates that both models capture the main growth dynamics of the data reasonably well at the population level.

```
# RMSE
# m4
dat$pred_logDV <- predict(m4, re.form = NA)
dat$resid_logDV <- dat$logDV - dat$pred_logDV

rmse <- sqrt(mean(dat$resid_logDV^2))
rmse
```

```
## [1] 0.1490912
```

```
p = exp(rmse)-1
p
```

```
## [1] 0.1607788
```

```
# m3
dat$pred_logDV <- predict(m3, re.form = NA)
dat$resid_logDV <- dat$logDV - dat$pred_logDV

rmse <- sqrt(mean(dat$resid_logDV^2))
rmse
```

```
## [1] 0.1490912
```

```
p = exp(rmse)-1
p
```

```
## [1] 0.1607788
```

The marginal R^2 was around 0.56 for both models, showing that the fixed effects explained more than half of the systematic variation in tumor growth. The conditional R^2 reached 0.96 for m4 and 0.91 for m3, reflecting the substantial contribution of random effects in capturing individual differences. The higher R^2c of m4 indicates that allowing each mouse to have its own growth rate provides a better overall representation of data.

```
library(MuMIn)
```

```
## Warning: package 'MuMIn' was built under R version 4.5.2
```

```
r.squaredGLMM(m4)
```

```
##           R2m       R2c
## [1,] 0.558666 0.9621573
```

```
r.squaredGLMM(m3)
```

```
##            R2m       R2c
## [1,] 0.5605041 0.9120662
```

When examining prediction intervals over time, the random-intercept model (m3) showed almost constant 95% interval widths, with only a very slight U-shaped pattern and values remaining around 0.14 on the log scale. This behaviour is typical for a model that assumes a common growth rate across mice and therefore does not allow uncertainty to accumulate strongly with time.

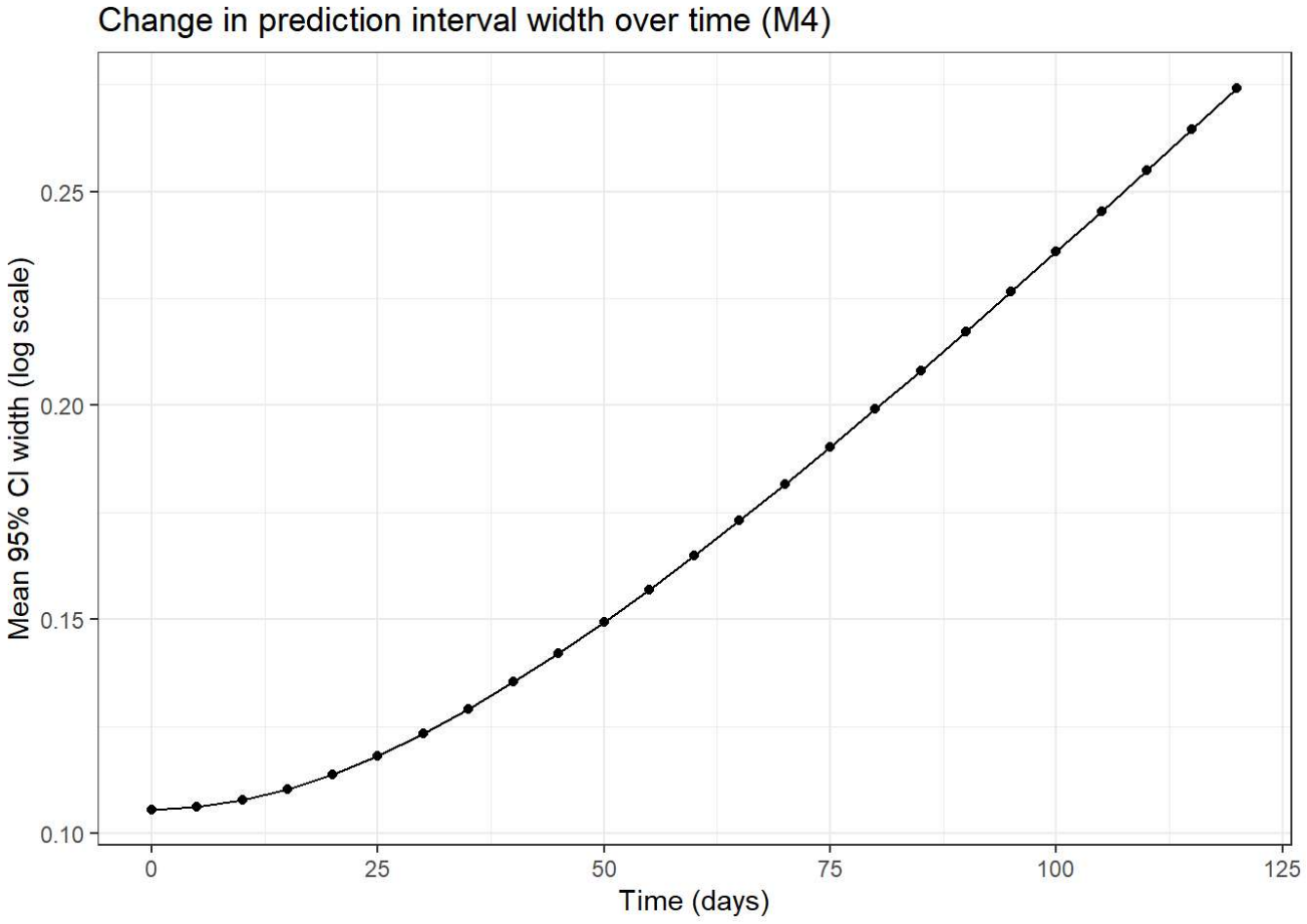
In contrast, the random-slope model (m4) produced intervals that widened more substantially, from about 0.10 within the observed range to over 0.25 by day 120. This reflects the fact that m4 explicitly models between-mouse variability in growth rates, so uncertainty about the population mean naturally increases as predictions are pushed further away from the data. Thus, m4 does not perform worse. Instead, it provides a more realistic representation of long-term uncertainty, whereas m3 may underestimate extrapolation uncertainty by assuming identical slopes for all animals.

```
# prediction intervals
time_grid <- seq(0, 120, by = 5)
# m4
emm_long <- emmeans(
  m4,
  ~ Treatment * Model | Time,
  at = list(Time = time_grid)
)
```

```
## Note: D.f. calculations have been disabled because the number of observations exceeds 3000.
## To enable adjustments, add the argument 'pbkrtest.limit = 5460' (or larger)
## [or, globally, 'set emm_options(pbkrtest.limit = 5460)' or larger];
## but be warned that this may result in large computation time and memory use.
```

```
## Note: D.f. calculations have been disabled because the number of observations exceeds 3000.
## To enable adjustments, add the argument 'lmerTest.limit = 5460' (or larger)
## [or, globally, 'set emm_options(lmerTest.limit = 5460)' or larger];
## but be warned that this may result in large computation time and memory use.
```

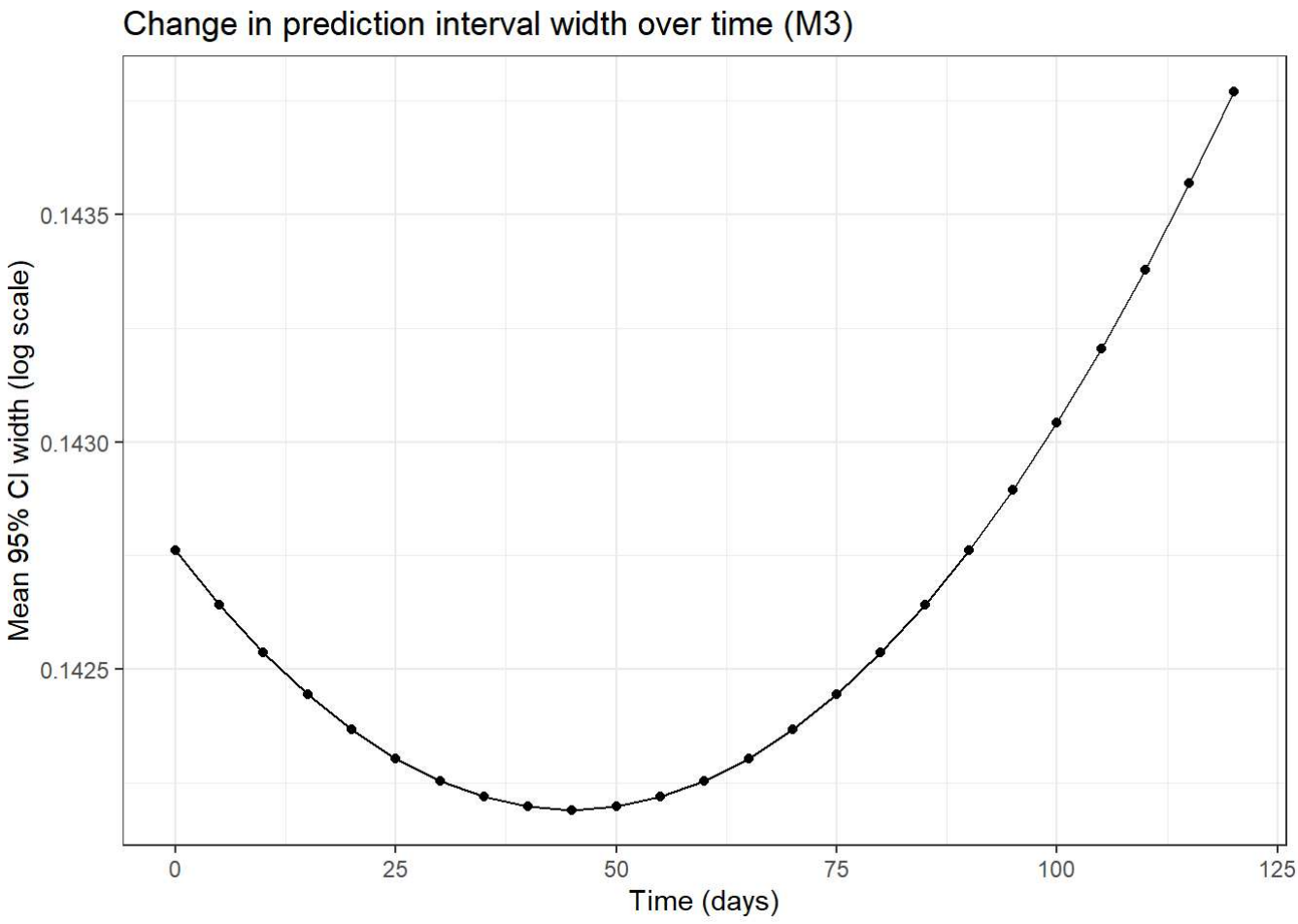
```
emm_df <- as.data.frame(emm_long)
emm_df$CI_width <- emm_df$asymp.UCL - emm_df$asymp.LCL

ci_by_time <- emm_df %>%
  group_by(Time) %>%
  summarise(mean_CI_width = mean(CI_width), .groups = "drop")

ggplot(ci_by_time, aes(x = Time, y = mean_CI_width)) +
  geom_line() +
  geom_point() +
  labs(
    x = "Time (days)",
    y = "Mean 95% CI width (log scale)",
    title = "Change in prediction interval width over time (M4)"
  ) +
  theme_bw()
```




```
# m3
emm_long <- emmeans(
  m3,
  ~ Treatment * Model | Time,
  at = list(Time = time_grid)
)
```

```
## Note: D.f. calculations have been disabled because the number of observations exceeds 3000.
## To enable adjustments, add the argument 'pbkrtest.limit = 5460' (or larger)
## [or, globally, 'set emm_options(pbkrtest.limit = 5460)' or larger];
## but be warned that this may result in large computation time and memory use.
## Note: D.f. calculations have been disabled because the number of observations exceeds 3000.
## To enable adjustments, add the argument 'lmerTest.limit = 5460' (or larger)
## [or, globally, 'set emm_options(lmerTest.limit = 5460)' or larger];
## but be warned that this may result in large computation time and memory use.
```

```
emm_df <- as.data.frame(emm_long)
emm_df$CI_width <- emm_df$asymp.UCL - emm_df$asymp.LCL

ci_by_time <- emm_df %>%
  group_by(Time) %>%
  summarise(mean_CI_width = mean(CI_width), .groups = "drop")
ggplot(ci_by_time, aes(x = Time, y = mean_CI_width)) +
  geom_line() +
  geom_point() +
  labs(
    x = "Time (days)",
    y = "Mean 95% CI width (log scale)",
    title = "Change in prediction interval width over time (M3)"
  ) +
  theme_bw()
```



Because the model is linear in time on the log scale, the predicted growth rate remains constant. This reflects a structural feature of the model rather than biological reality. Real tumors are unlikely to follow such sustained exponential expansion due to biological constraints such as limited resource. Therefore, while the model

provides reliable interpolation within the observed time range, predictions far outside the data should be interpreted cautiously and viewed as illustrative rather than biologically realistic.

Task_3_report

November 14, 2025

[46]:
```
# a first look at the data
head(data)
summary(data)
```

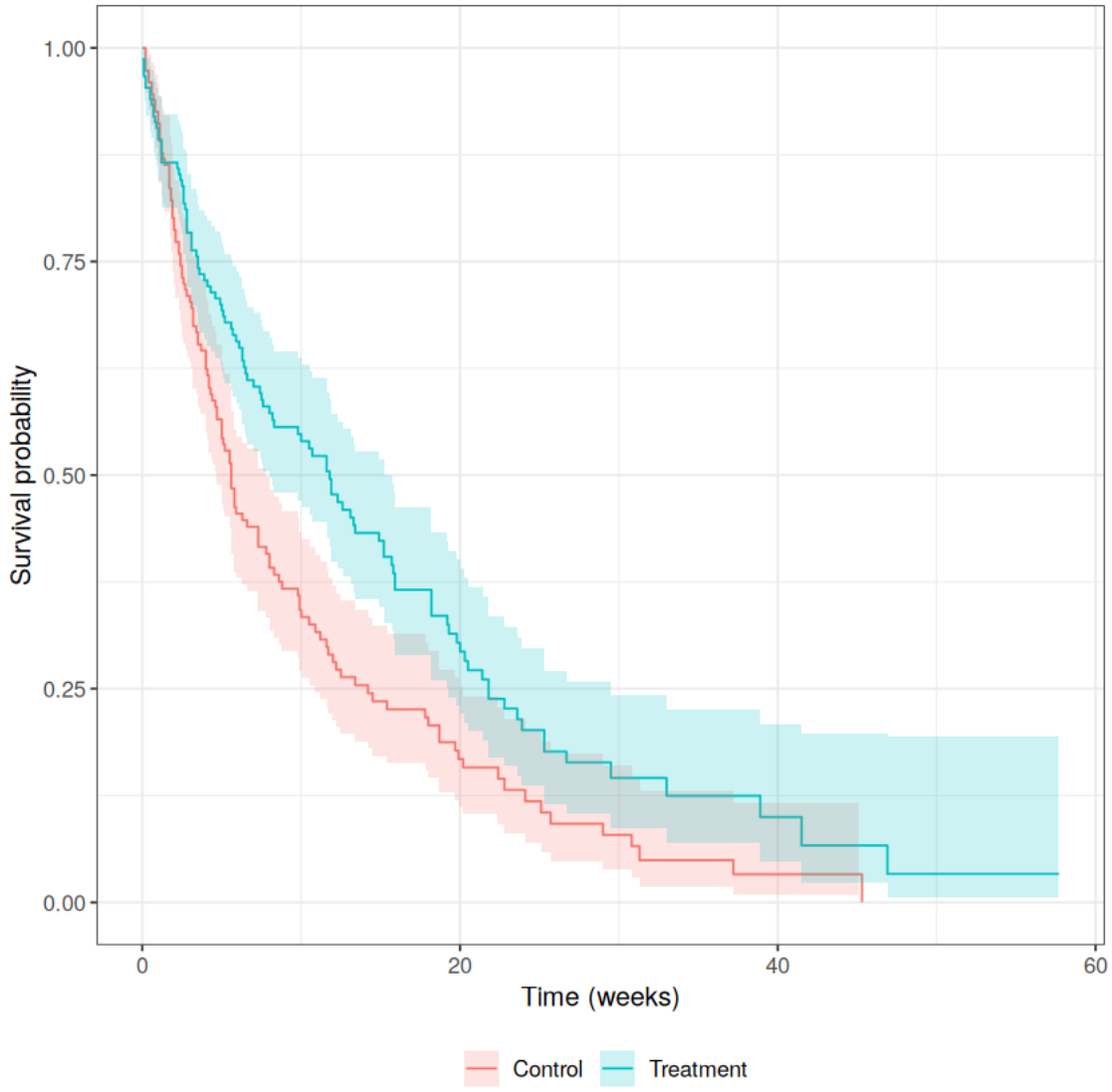
A data.frame: 6 × 9

	X <int>	ID <int>	TreatmentGroup <chr>	Age <int>	Sex <chr>	ECOG_PS <int>	GFR <int>	Time <dbl>	E <
1	1	1	Treatment	73	Female	2	83	0.6	1
2	2	2	Treatment	62	Male	0	86	6.3	1
3	3	3	Treatment	61	Female	0	98	57.7	0
4	4	4	Treatment	66	Male	1	58	0.7	1
5	5	5	Treatment	88	Female	0	120	30.7	0
6	6	6	Treatment	73	Male	0	55	6.5	1

```
       X                ID         TreatmentGroup          Age       
 Min.   :  1.00   Min.   :  1.00   Length:300         Min.   :38.0  
 1st Qu.: 75.75   1st Qu.: 75.75   Class :character   1st Qu.:59.0  
 Median :150.50   Median :150.50   Mode  :character   Median :65.0  
 Mean   :150.50   Mean   :150.50                      Mean   :65.3  
 3rd Qu.:225.25   3rd Qu.:225.25                      3rd Qu.:72.0  
 Max.   :300.00   Max.   :300.00                      Max.   :91.0  
     Sex                ECOG_PS           GFR              Time       
 Length:300         Min.   :0.0000   Min.   : 30.00   Min.   : 0.000  
 Class :character   1st Qu.:0.0000   1st Qu.: 66.00   1st Qu.: 2.600  
 Mode  :character   Median :1.0000   Median : 81.00   Median : 5.900  
                    Mean   :0.8167   Mean   : 78.89   Mean   : 9.673  
                    3rd Qu.:1.0000   3rd Qu.: 91.00   3rd Qu.:13.600  
                    Max.   :2.0000   Max.   :120.00   Max.   :57.700  
     Event       
 Min.   :0.00  
 1st Qu.:1.00  
 Median :1.00  
 Mean   :0.76  
 3rd Qu.:1.00  
 Max.   :1.00  
```

From the visual graphs we can observe quite a even split between the control and treatment group, further more the number of male are more in number making the data is skewed. For GFR, Age the data distribution look quite normal but time distribution is heavily left skewed.

[57]:
```
# Kaplan-Meier survival curves with confidence intervals by treatment group
survfit2(Surv(Time, Event) ~ TreatmentGroup, data = data) |>
  ggsurvfit() +
  labs(
    x = "Time (weeks)",
    y = "Survival probability"
  )+
  add_confidence_interval() #+
  # add_risktable()
```



[68]:
```
summary(survfit2(Surv(Time, Event) ~ TreatmentGroup, data = data), times = 5)
```

```
Call: survfit(formula = Surv(Time, Event) ~ TreatmentGroup, data = data)
```

```
                TreatmentGroup=Control
       time       n.risk      n.event     survival      std.err lower 95% CI
      5.000       77.000       65.000        0.543        0.042        0.467
upper 95% CI
      0.632

                TreatmentGroup=Treatment
       time       n.risk      n.event     survival      std.err lower 95% CI
     5.0000      99.0000      45.0000       0.6927       0.0382       0.6217
upper 95% CI
     0.7718
```

We are using Kaplan-Meier survival modeling here, a thing to observer here the early treatment for first 4 weeks we can see the control group having better survival odds but with increasing time the treatment shows effect and we can see uplift of the treatment course.

Can be observed in the graphs.

[53]:
```r
# Fit the Cox PH model
cox_model <- coxph(Surv(Time, Event) ~ TreatmentGroup + Age + Sex + ECOG_PS +␣
 ↪GFR, data = data)

# Summarize the model
summary(cox_model)
```

```
Call:
coxph(formula = Surv(Time, Event) ~ TreatmentGroup + Age + Sex +
    ECOG_PS + GFR, data = data)

  n= 300, number of events= 228

                            coef exp(coef)  se(coef)      z Pr(>|z|)
TreatmentGroupTreatment -0.525216  0.591428  0.138716 -3.786 0.000153 ***
Age                     -0.010326  0.989727  0.006697 -1.542 0.123080
SexMale                 -0.086113  0.917490  0.137160 -0.628 0.530115
ECOG_PS                  0.444375  1.559514  0.091175  4.874 1.09e-06 ***
GFR                     -0.013653  0.986439  0.003674 -3.716 0.000202 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                        exp(coef) exp(-coef) lower .95 upper .95
TreatmentGroupTreatment    0.5914     1.6908    0.4506    0.7762
Age                        0.9897     1.0104    0.9768    1.0028
SexMale                    0.9175     1.0899    0.7012    1.2005
ECOG_PS                    1.5595     0.6412    1.3043    1.8647
GFR                        0.9864     1.0137    0.9794    0.9936
```

```
Concordance= 0.646  (se = 0.019 )
Likelihood ratio test= 54.07  on 5 df,   p=2e-10
Wald test            = 54.51  on 5 df,   p=2e-10
Score (logrank) test = 55.88  on 5 df,   p=9e-11
```

Key observations: After analy the Pr(>|z|), we find 3 factors to be significant - TreatmentGroup, ECOG_PS & GFR * TreatmentGroup = Treatment, P value is quite small and we can see there is a negative coefficient meaning there is an inverse correlation to the risk of death, meaning the person having the treatment has a higher chance of recovery and by exp(coef) reduces it by 41% * ECOG_PS, p value is quite small and is significant furthermore it has a positive correlation and by looking the exp(coef) we can can it increases the chance of death. * GFR, is relevant due to the small p value, and has as negative relation i.e. 1 unit increase in GFR adds decreased risk by 1.4%.

[55]:
```r
# Extract the hazard ratios and 95% CI
hazard_ratios <- coef(cox_model)
conf_int <- confint(cox_model)

# Print the results
print(paste("Hazard Ratios and 95% CI:"))
print(hazard_ratios)
print(conf_int)
```

```
[1] "Hazard Ratios and 95% CI:"
TreatmentGroupTreatment                     Age                  SexMale
             -0.52521603             -0.01032611             -0.08611317
                 ECOG_PS                     GFR
              0.44437455             -0.01365335
                                  2.5 %       97.5 %
TreatmentGroupTreatment -0.79709423 -0.253337834
Age                     -0.02345135  0.002799134
SexMale                 -0.35494193  0.182715596
ECOG_PS                  0.26567435  0.623074745
GFR                     -0.02085457 -0.006452126
```

Looking at the HR ratios of the Cox PH model, we draw similar conclusions as before. Furthermore there are some slightly significant observations that can be drawn are being Male reduced the risk marginally or with age the risk goes down as well. But looking at the data imbalance we can see the data to be favoring men as well as higher ages.

S3T4

Group A3

2025-11-05

Step 1: Joining datasets

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
library(tableone)
```

```
## Warning: package 'tableone' was built under R version 4.5.2
```

```
library(broom)
library(ResourceSelection) # Hosmer-Lemeshow test
```

```
## Warning: package 'ResourceSelection' was built under R version 4.5.2
```

```
## ResourceSelection 0.3-6   2023-06-27
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.5.2
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(car) # VIF
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(survival) # time-to-event
library(cmprsk)
```

```
## Warning: package 'cmprsk' was built under R version 4.5.2
```

```
library(tidyr)
library(GGally)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(knitr)

data_t3 <- read_csv("Data_T3.csv")
```

```
## New names:
## * `` -> `...1`
```

```
## Rows: 300 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): TreatmentGroup, Sex
## dbl (7): ...1, ID, Age, ECOG_PS, GFR, Time, Event
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_t4 <- read_csv("Data_T4.csv")
```

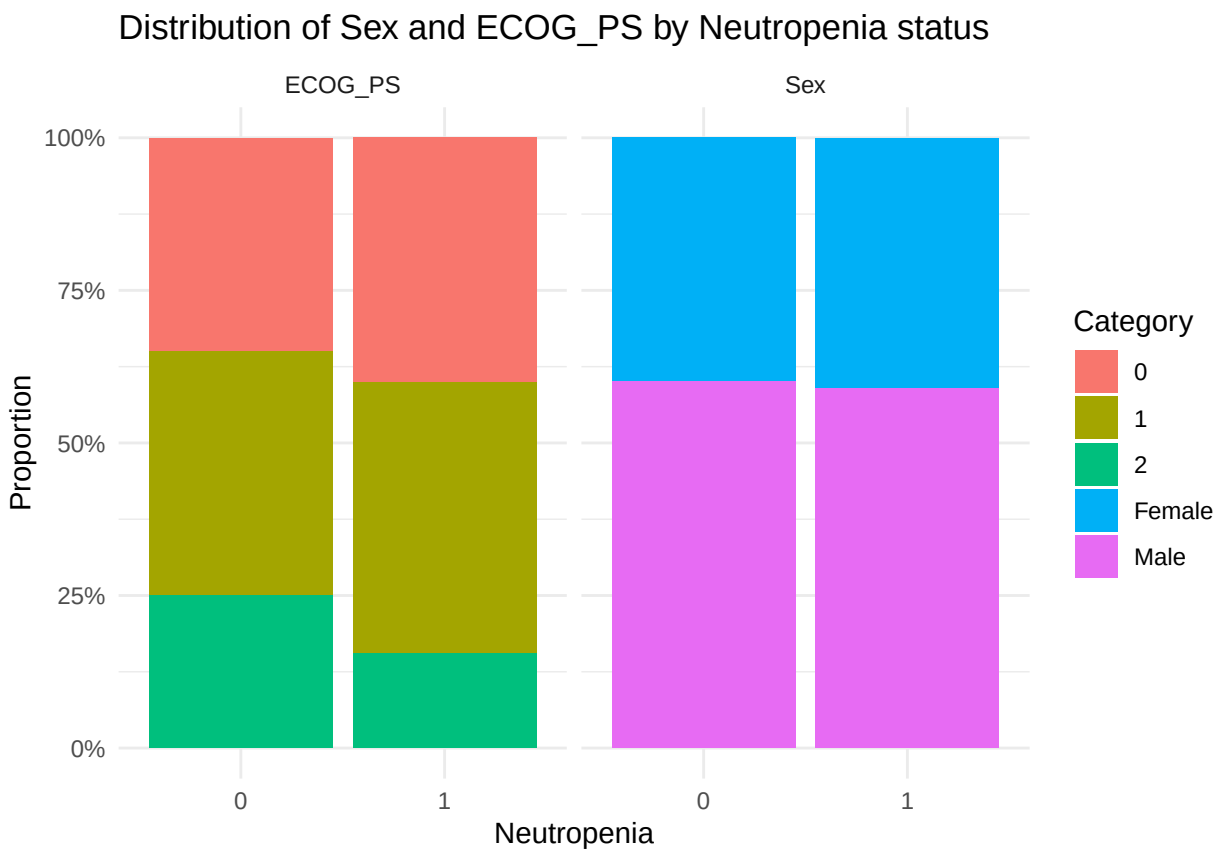
```
## New names:
## Rows: 150 Columns: 3
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (3): ...1, ID, Neutropenia
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

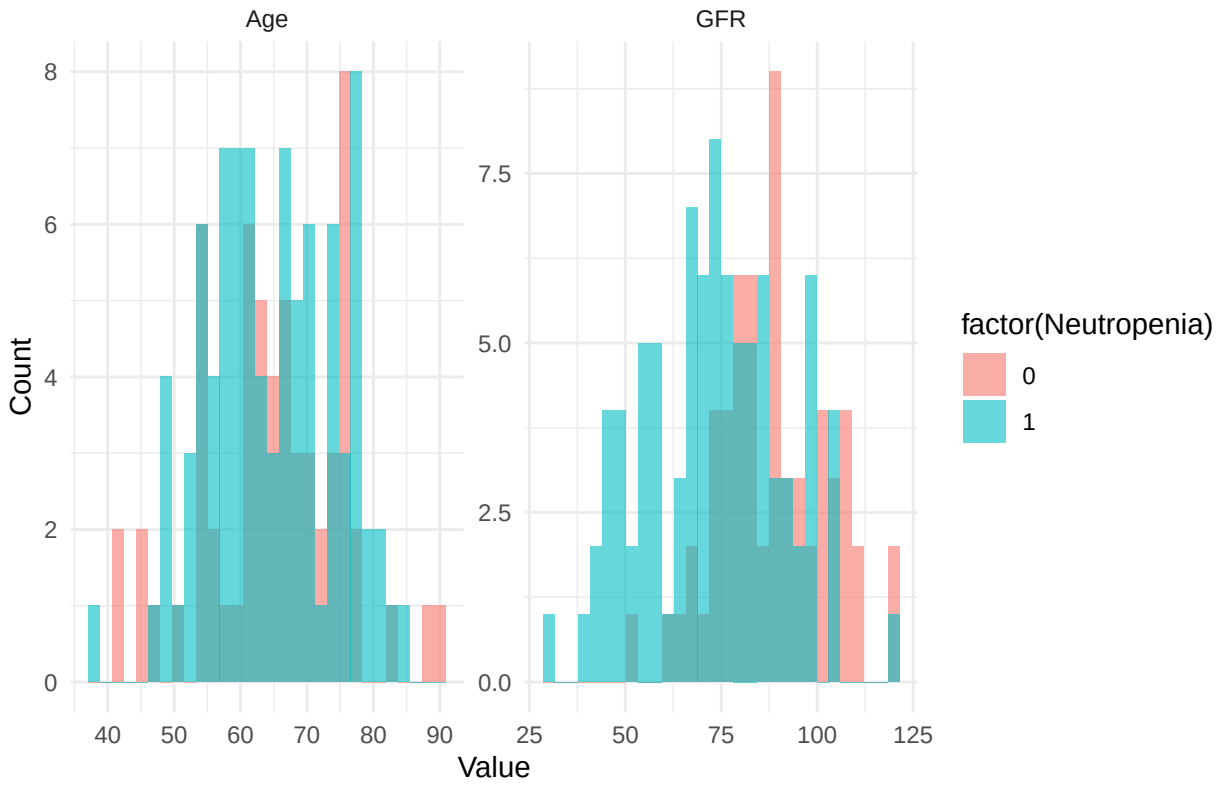
```r
treat_group <- data_t3 %>% filter(TreatmentGroup == "Treatment")
merged <- treat_group %>%
  left_join(data_t4, by = "ID")
```

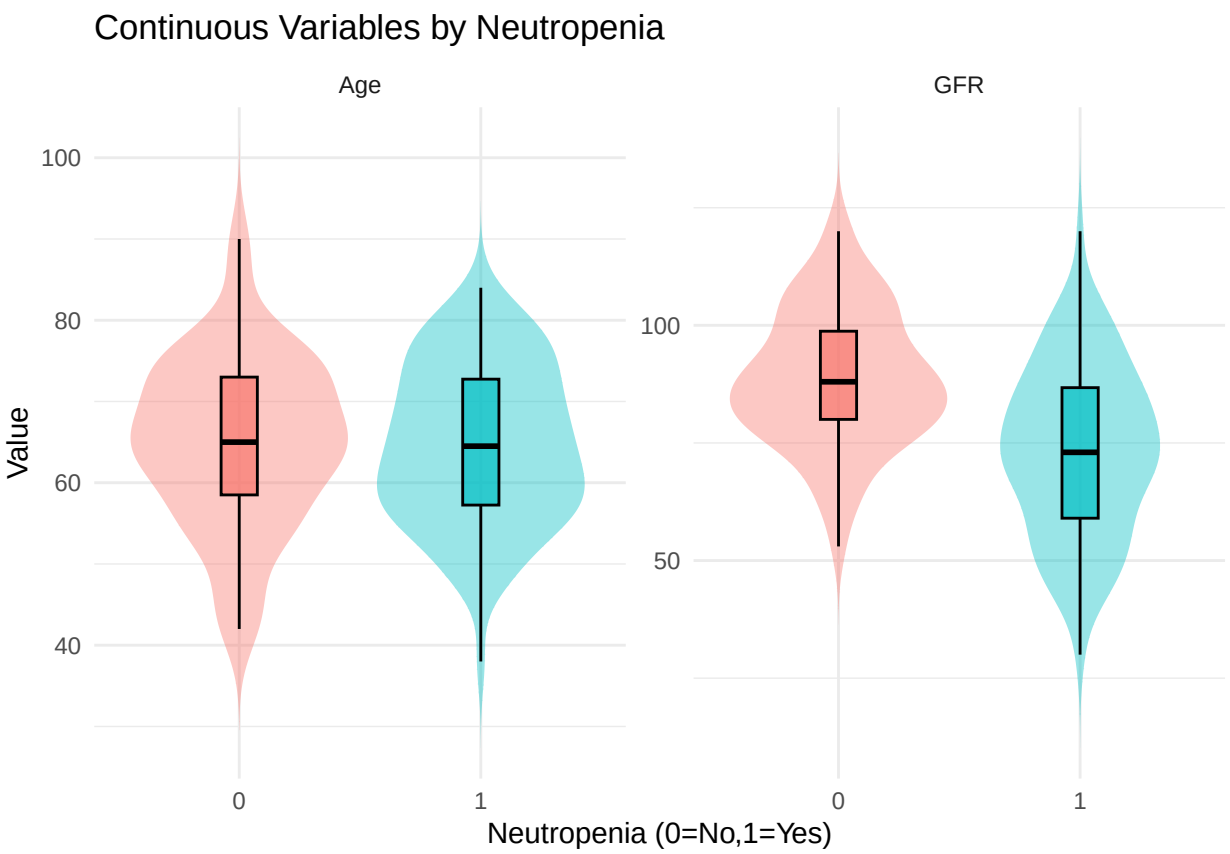
Step 2: Data exploration

We examined the data, visualized the distribution of categorical data, including Sex and EcoG_PS, and continuous variables, including GFR and Age, by Neutropenia status, and the distribution of Neutropenia Status in categorical data.

```
##                   Stratified by Neutropenia
##                    level  0             1             p      test
##   n                            60            90
##   Age (mean (SD))         65.05 (10.46) 64.68 (9.71)   0.824
##   Sex (%)          Female     24 (40.0)     37 (41.1)   1.000
##                    Male       36 (60.0)     53 (58.9)
##   ECOG_PS (%)      0          21 (35.0)     36 (40.0)   0.356
##                    1          24 (40.0)     40 (44.4)
##                    2          15 (25.0)     14 (15.6)
##   GFR (mean (SD))         88.52 (14.50) 73.28 (18.09) <0.001
```

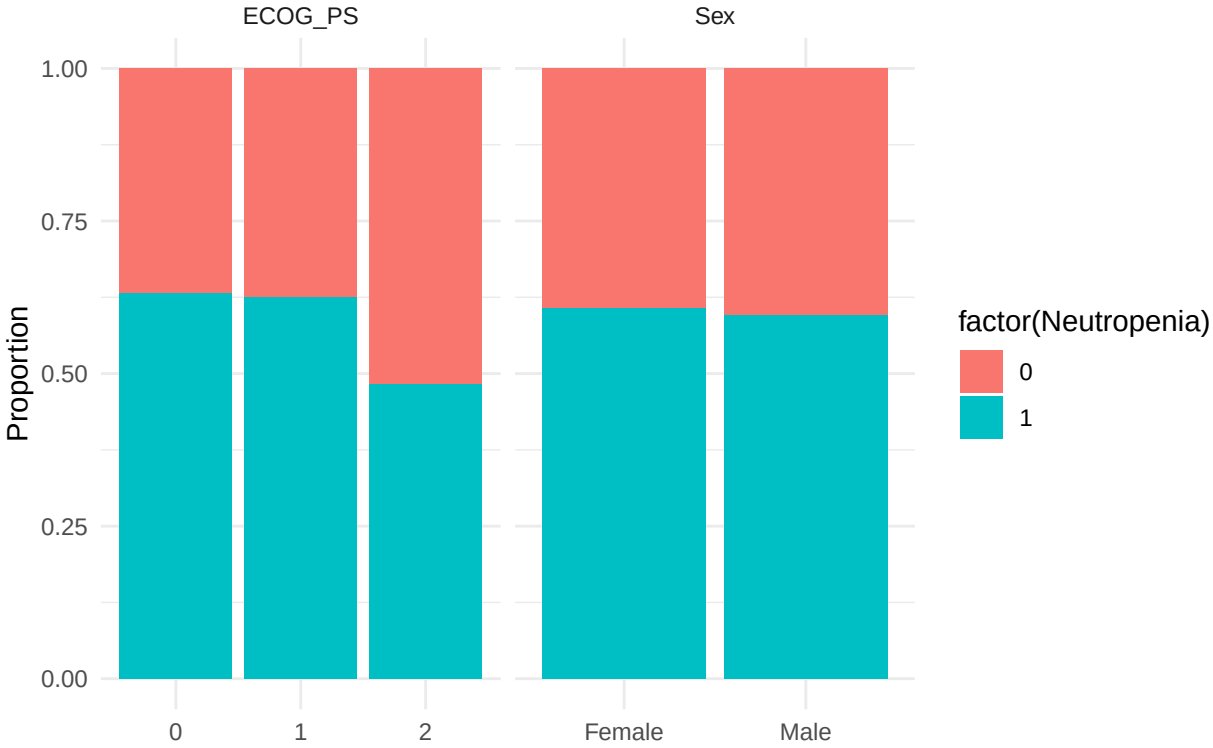

Distribution of Sex and ECOG_PS by Neutropenia status
ECOG_PS
Sex
100%
75%
50%
25%
0%
Proportion
0
1
0
1
Neutropenia
Category
0
1
2
Female
Male


Distribution of Continuous Variables by Neutropenia
Age
GFR
8
6
4
2
0
Count
7.5
5.0
2.5
0.0
40
50
60
70
80
90
25
50
75
100
125
Value
factor(Neutropenia)
0
1




```
## Warning: There were 2 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘text = text_fn(.data$x, .data$y)‘.
## i In group 1: ‘color = 0‘.
## Caused by warning in ‘cor()‘:
## ! the standard deviation is zero
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 1 remaining warning.
## There were 2 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘text = text_fn(.data$x, .data$y)‘.
## i In group 1: ‘color = 0‘.
## Caused by warning in ‘cor()‘:
## ! the standard deviation is zero
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 1 remaining warning.
```

Categorical Variables by Neutropenia



Results shows that continuous variables are seemingly in a normal distribution, GRF distribution looks differently in different Neutropenia case, while Age doesn't seem to have significant influence on Neutropenia.

Percentage of high-level EcoG_PS is higher in Neutropenia cases, while sex distribute similarly in with and without Neutropenia.

Next, we carry out statistic tests on continuous variables.

```
##
##     Female Male
##   0     24   36
##   1     37   53
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(merged$Neutropenia, merged$Sex)
## X-squared = 0, df = 1, p-value = 1
```

```
##
##      0  1  2
##   0 21 24 15
##   1 36 40 14
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(merged$Neutropenia, merged$ECOG_PS)
## X-squared = 2.0644, df = 2, p-value = 0.3562
```

Table 1: Comparison of Continuous Variables by Neutropenia

	Variable	Normality_Group0_p	Normality_Group1_p	Test	Statistic	p_value
t...1	Age	0.4044	0.2335	t-test	0.2196	0.8265
t...2	GFR	0.8557	0.8876	t-test	5.7022	0.0000

From the results of the Pearson's Chi-squared tests, p-value of sex is 1, showing Neutropenia is statistically independent from Sex. The p-value for test of EcoG is 0.3562, higher than 0.05, which means that Neutropenia is also approximately independent from EcoG.

From Shapiro Tests, both continuous variables fit the normal distribution, which also accord with the histogram. While the p-value of age is higher than 0.05 showing significant difference in age with or without Neutropenia, the p-value of GFR is close to 0, showing no significant difference in GFR between different Neutropenia state. This result does not fit the violin plot and boxplot. The reason might be that the medium is shifting from mean value.

Step 3: Modeling

```
##
## Call:
## glm(formula = Neutropenia ~ Age + Sex + ECOG_PS + GFR, family = binomial,
##     data = merged)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.73875    1.51573   3.126  0.00177 **
## Age          0.01113    0.01921   0.579  0.56242
## SexMale     -0.18031    0.38281  -0.471  0.63764
## ECOG_PS     -0.33155    0.25237  -1.314  0.18894
## GFR         -0.05752    0.01235  -4.658 3.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 201.90  on 149  degrees of freedom
## Residual deviance: 172.05  on 145  degrees of freedom
## AIC: 182.05
##
## Number of Fisher Scoring iterations: 4
```

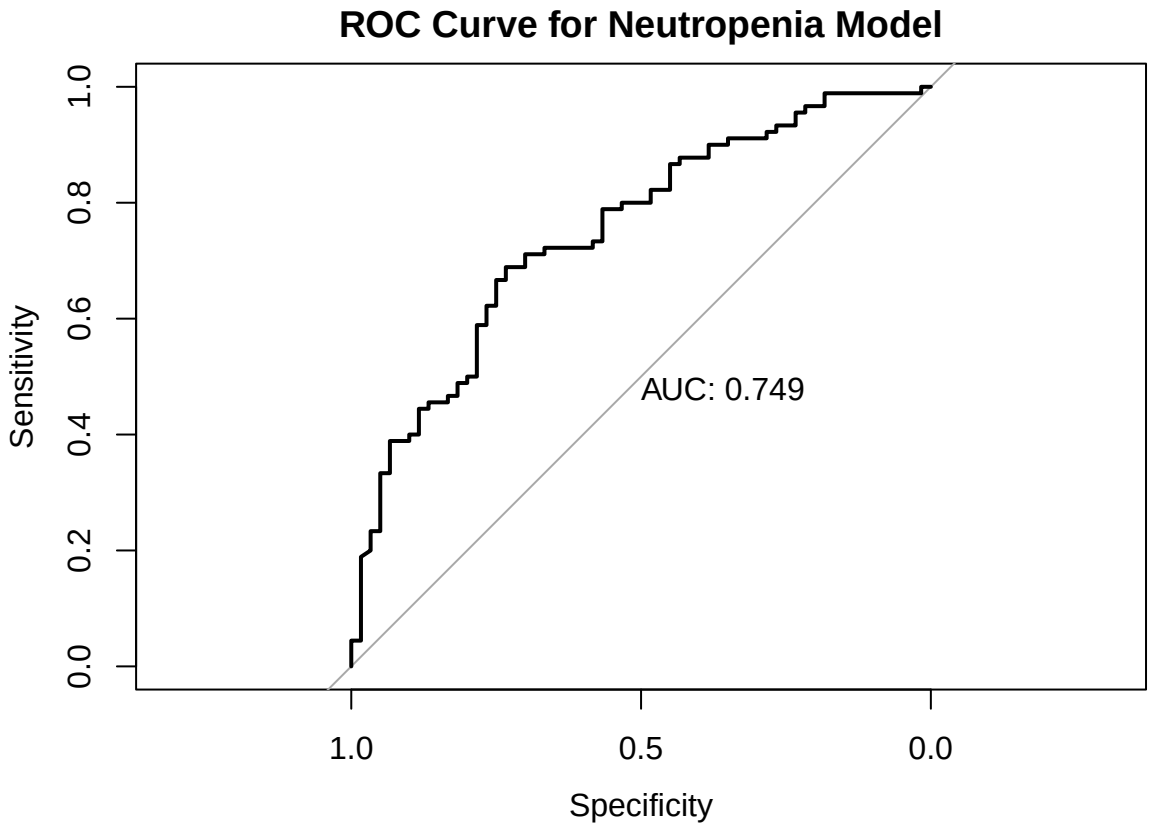
```
## # A tibble: 5 x 7
##   term        estimate std.error statistic    p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>    <dbl>     <dbl>
## 1 (Intercept)  114.       1.52       3.13  0.00177       6.62   2608.
## 2 Age            1.01     0.0192     0.579 0.562         0.974     1.05
## 3 SexMale        0.835    0.383     -0.471 0.638         0.390     1.76
## 4 ECOG_PS        0.718    0.252     -1.31  0.189         0.434     1.17
## 5 GFR            0.944    0.0123    -4.66  0.00000319    0.920     0.966
```

```
##      Age      Sex  ECOG_PS      GFR
## 1.044859 1.028674 1.008695 1.040921
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  merged$Neutropenia, fitted(model_logit)
## X-squared = 8.6298, df = 8, p-value = 0.3745
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



We used logistic regression to establish a model for this case. The log-odds and odds-ratio shows that only GFR has a significant influence on Neutropenia state, better GFR leads to lower risk of Neutropenia. VIF value are all close to 1, which means that there are no obvious colinearality problem in this model. To access the performance, we accessed the AUC value, which is 0.749, showing the model is acceptable.

For treatment, it is suggested by this model that renal function can support treatment to a better result.

Step 4: Future Work

In the future, maybe we can include factor of time of Neutropenia, to build a time-to-event model.