STAT 131A Final Project

Eric Ho and Joseph Gitlin

```
knitr::opts_chunk$set(echo = TRUE, tidy.opts=list(width.cutoff=60), tidy=TRUE)
```

```
pkgTest <- function(x) {
  if (!require(x,character.only = TRUE)) {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}
packages = c("tidyverse", "patchwork")
loading <- lapply(packages, pkgTest)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: patchwork
```

```
library(tidyverse)
library(patchwork)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
##
## The following object is masked from 'package:datasets':
##
##     rivers
```

Question 2a: Reading Data

```
cholangitis <- read.csv("cholangitis-data.csv")
cat_vars <- c("status", "drug", "sex", "ascites", "hepatomegaly", "spiders", "edema", "stage")
cholangitis[, cat_vars] <- lapply(cholangitis[, cat_vars], factor)
cholangitis[, "age"] <- cholangitis[, "age"] / 365 # converting from days to years
head(cholangitis)
```

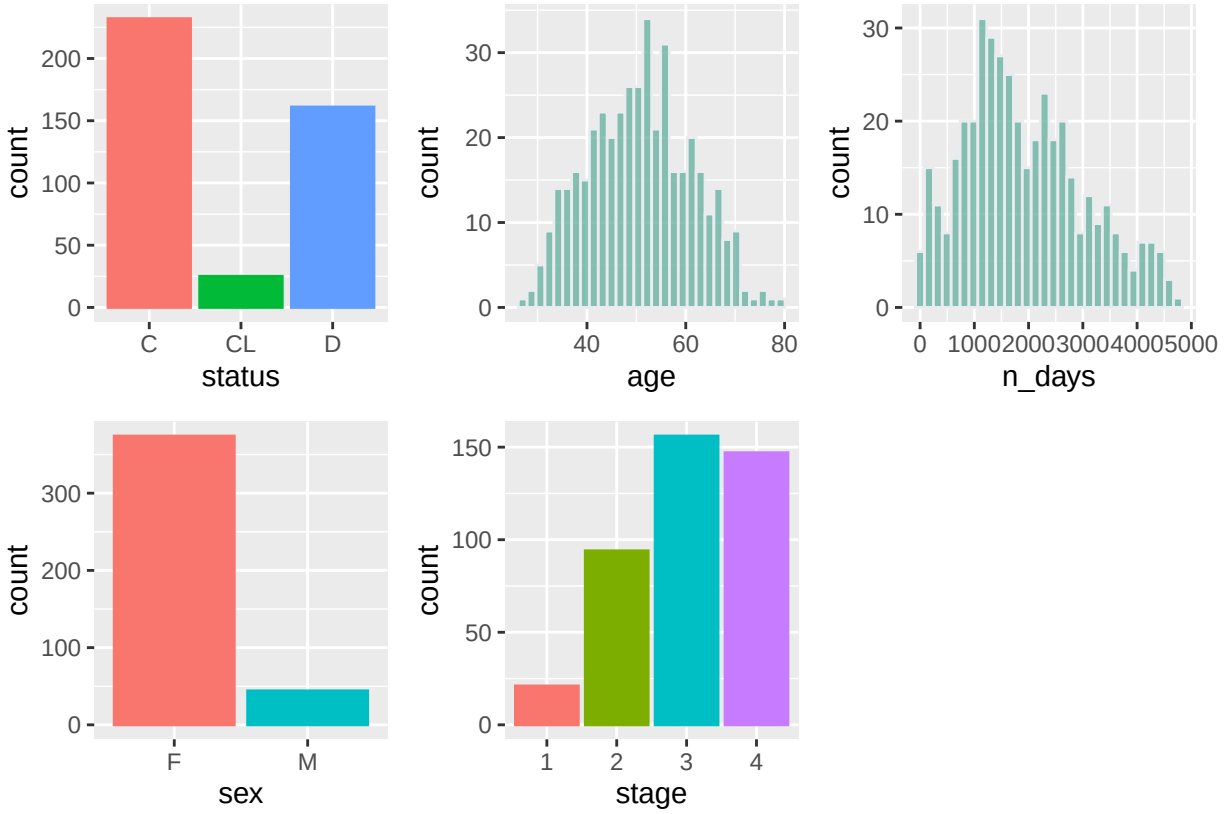
```
##   id n_days status           drug      age sex ascites hepatomegaly spiders
## 1  1    400      D D-penicillamine 58.80548   F       Y            Y       Y
## 2  2   4500      C D-penicillamine 56.48493   F       N            Y       Y
## 3  3   1012      D D-penicillamine 70.12055   M       N            N       N
## 4  4   1925      D D-penicillamine 54.77808   F       N            Y       Y
## 5  5   1504     CL         Placebo 38.13151   F       N            Y       Y
## 6  6   2503      D         Placebo 66.30411   F       N            Y       N
##   edema bilirubin cholesterol albumin copper alk_phos   sgot tryglicerides
## 1     Y      14.5         261    2.60    156   1718.0 137.95           172
## 2     N       1.1         302    4.14     54   7394.8 113.52            88
## 3     S       1.4         176    3.48    210    516.0  96.10            55
## 4     S       1.8         244    2.54     64   6121.8  60.63            92
## 5     N       3.4         279    3.53    143    671.0 113.15            72
## 6     N       0.8         248    3.98     50    944.0  93.00            63
##   platelets prothrombin stage
## 1       190        12.2     4
## 2       221        10.6     3
## 3       151        12.0     4
## 4       183        10.3     4
## 5       136        10.9     3
## 6       361        11.0     3
```

Question 2b: EDA

Through exploratory data analysis (EDA), we're looking to identify trends or potential relationships between variables, whether it is independent variables with other independent variables, or more likely, independent variables to the final status.

```
status_bar <- ggplot(cholangitis, aes(x=status, color=factor(status), fill=factor(status))) +
  geom_bar() +
  theme(legend.position="none")

sex_bar <- ggplot(cholangitis, aes(x=sex, color=factor(sex), fill=factor(sex))) +
  geom_bar() +
  theme(legend.position="none")

stage_bar <- ggplot(cholangitis, aes(x=stage, color=factor(stage), fill=factor(stage))) +
  geom_bar() +
  theme(legend.position="none")

age_hist <- ggplot(cholangitis, aes(x=age)) +
  geom_histogram(fill="#69b3a2", color="#e9ecef", alpha=0.8)

n_days_hist <- ggplot(cholangitis, aes(x=n_days)) +
  geom_histogram(fill="#69b3a2", color="#e9ecef", alpha=0.8)

status_bar + age_hist + n_days_hist + sex_bar + stage_bar
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



When looking at the bar plots and histogram of the status, age, and sex respectively, we notice that a majority of the patients survived following the treatments, however there is a significant number of deaths as well. The patients who received transplants (denoted by CL) are not accurate representatives of the data or the subject of this trial and also should be dropped when analyzing the data.

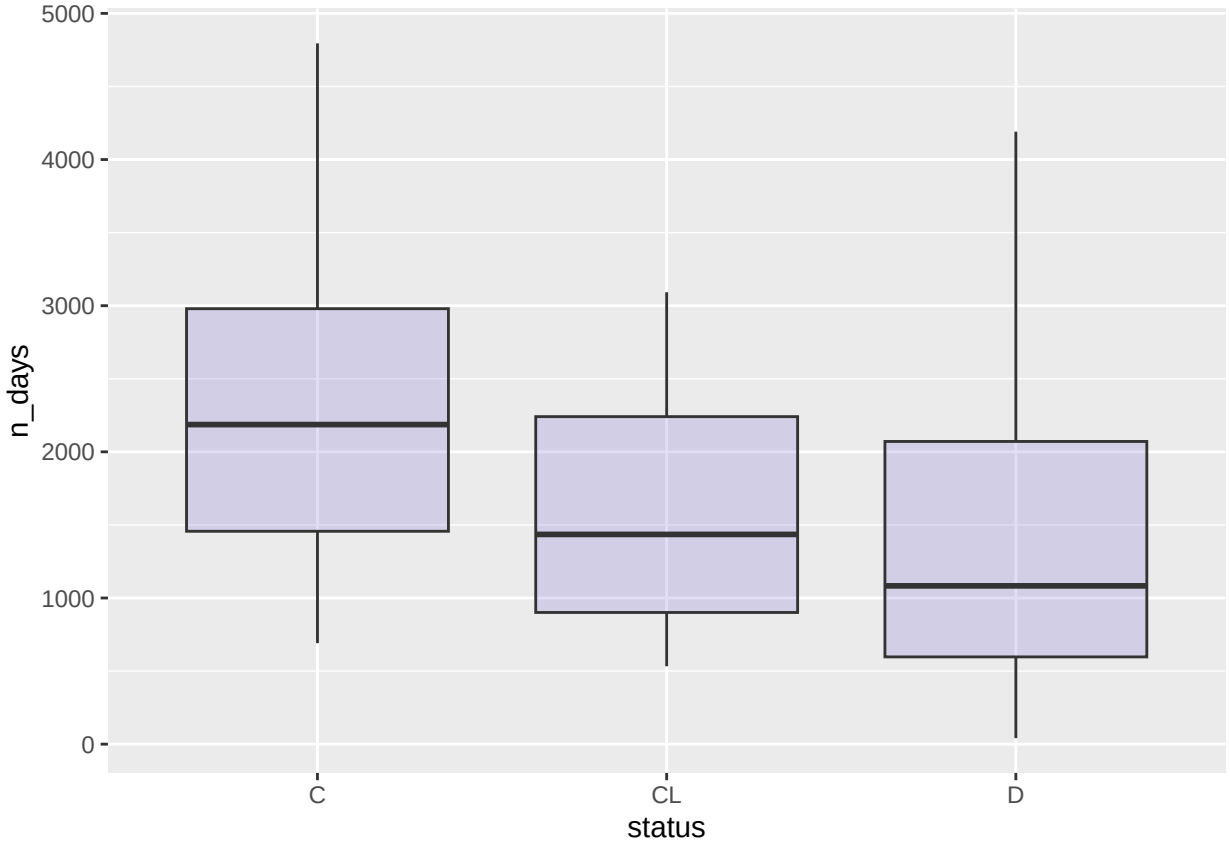
Age is widely distributed but most concentrated in the 50s.

With regards to the n_days variable, there is a larger concentration between the 1000 and 2000 day mark which can either be attributed to more patients dying during that time or the study ending early. This can be explored further through a visualization relating the number of days with the status of the patient (see below).

There are a significant number of female patients compared to male patients which is something to note about this data–any results or predictions are likely to be more accurate for women than men.

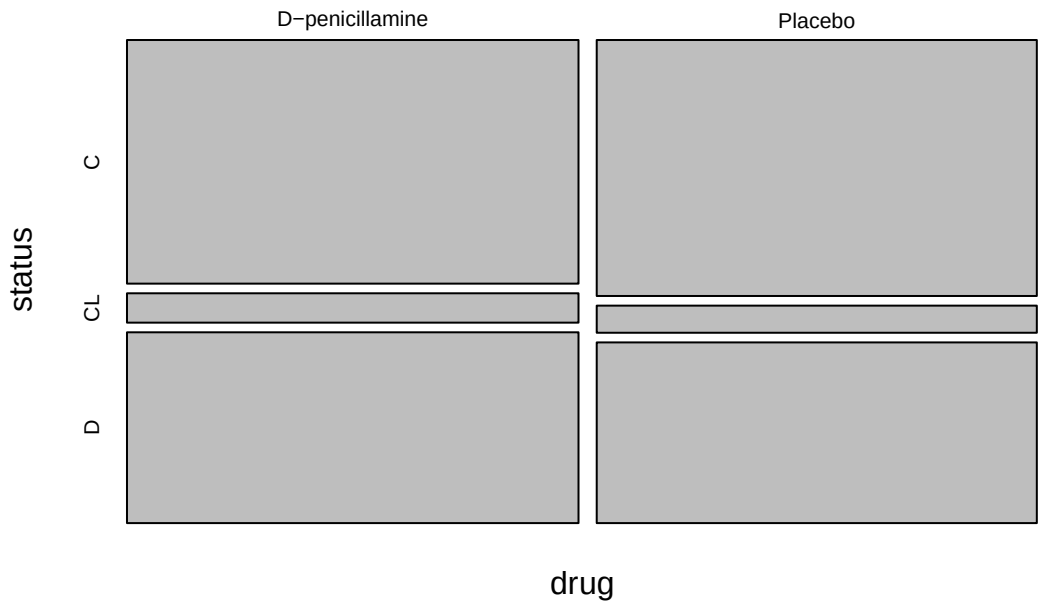
It looks like this drug trial happened with patients mostly in stage 3 or stage 4. This is likely because patients did not catch the diagnoses earlier in the stage of the disease.

```
ggplot(cholangitis, aes(x=status, y=n_days)) +
    geom_boxplot(fill="slateblue", alpha=0.2)
```



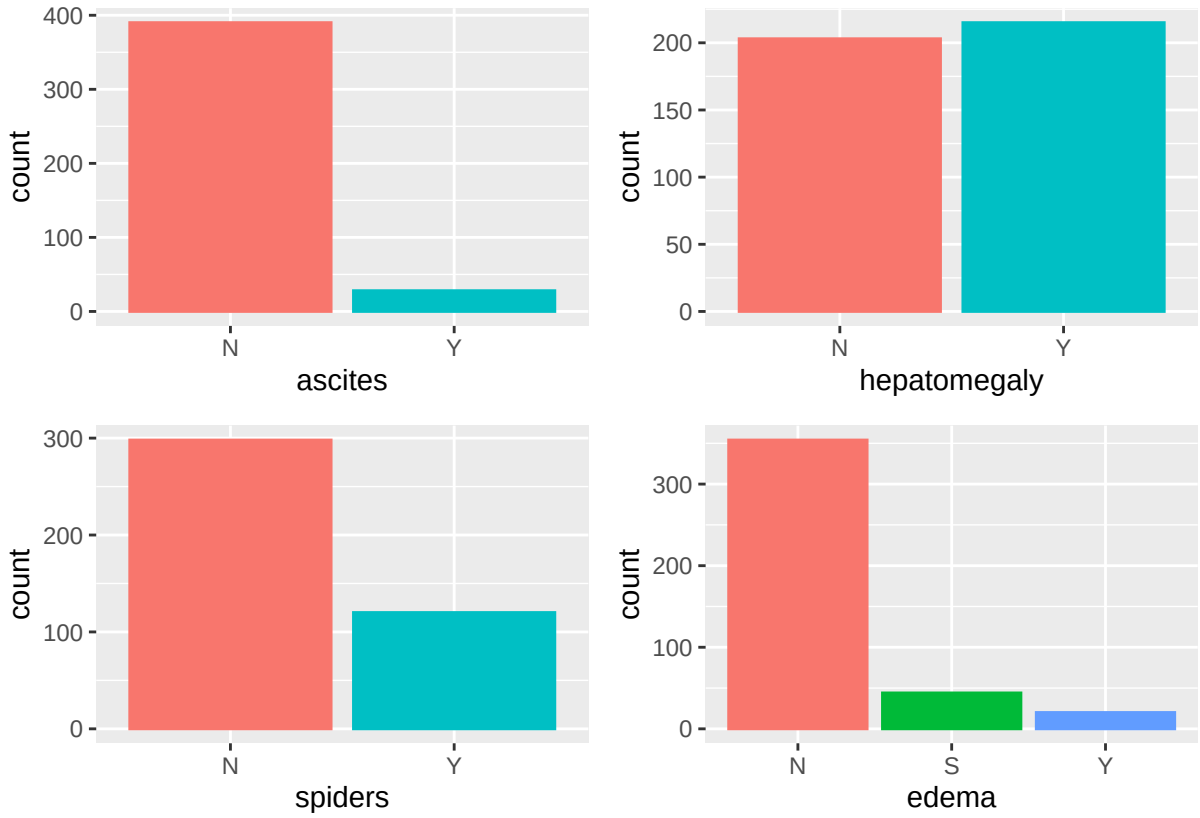
From this boxplot, the median date for patients who died is around 1000, while the median release time for patients who survived was close to 2000 days. This coincides with the histogram previously seen and suggests that a large number of patients will survive at least 1000 days.

```
mosaicplot(drug~status, cholangitis)
```



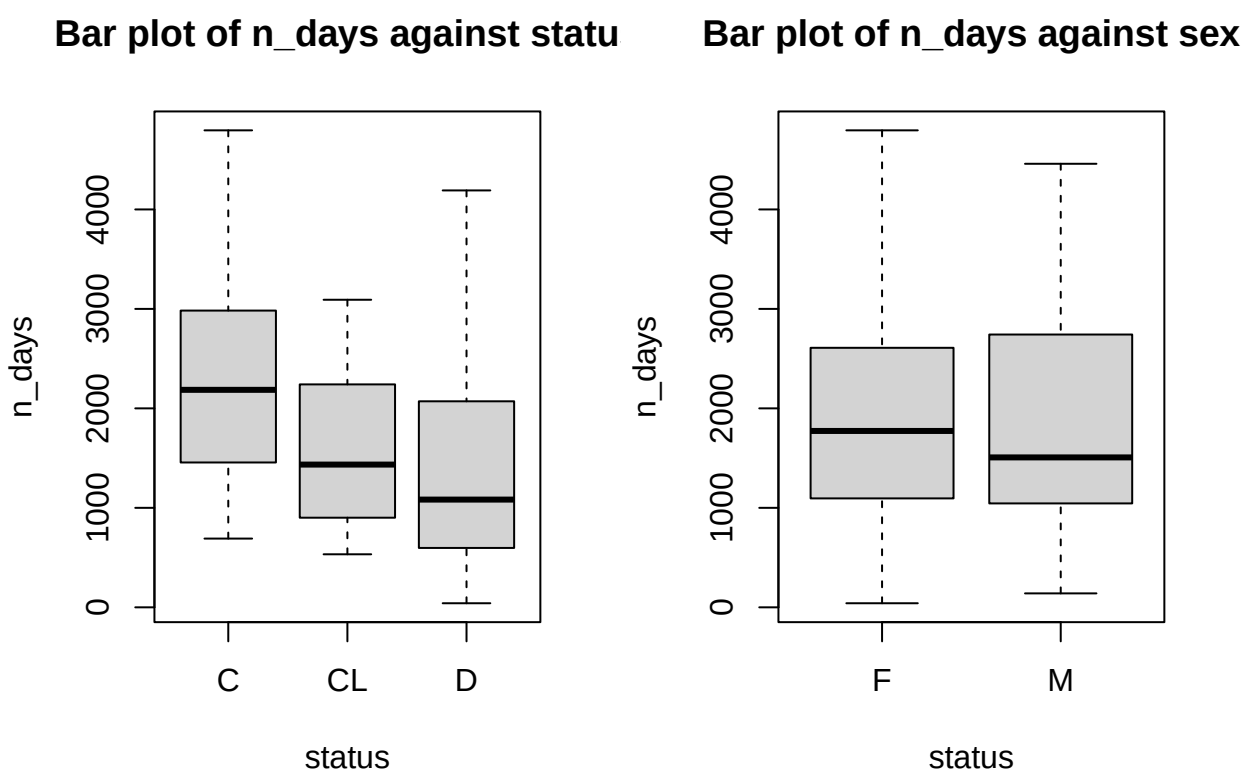
Based on this mosaic plot, the ratio of people who received the placebo that survived is higher than those who received the trial drug.

```
ascites_bar <- ggplot(cholangitis, aes(x=ascites, color=factor(ascites), fill=factor(ascites))) +
  geom_bar() +
  theme(legend.position="none")

hepatomegaly_bar <- ggplot(cholangitis, aes(x=hepatomegaly, color=factor(hepatomegaly), fill=factor(hepa

spiders_bar <- ggplot(cholangitis, aes(x=spiders, color=factor(spiders), fill=factor(spiders))) +
  geom_bar() +
  theme(legend.position="none")

edema_bar <- ggplot(cholangitis, aes(x=edema, color=factor(edema), fill=factor(edema))) +
  geom_bar() +
  theme(legend.position="none")

ascites_bar + hepatomegaly_bar + spiders_bar + edema_bar
```

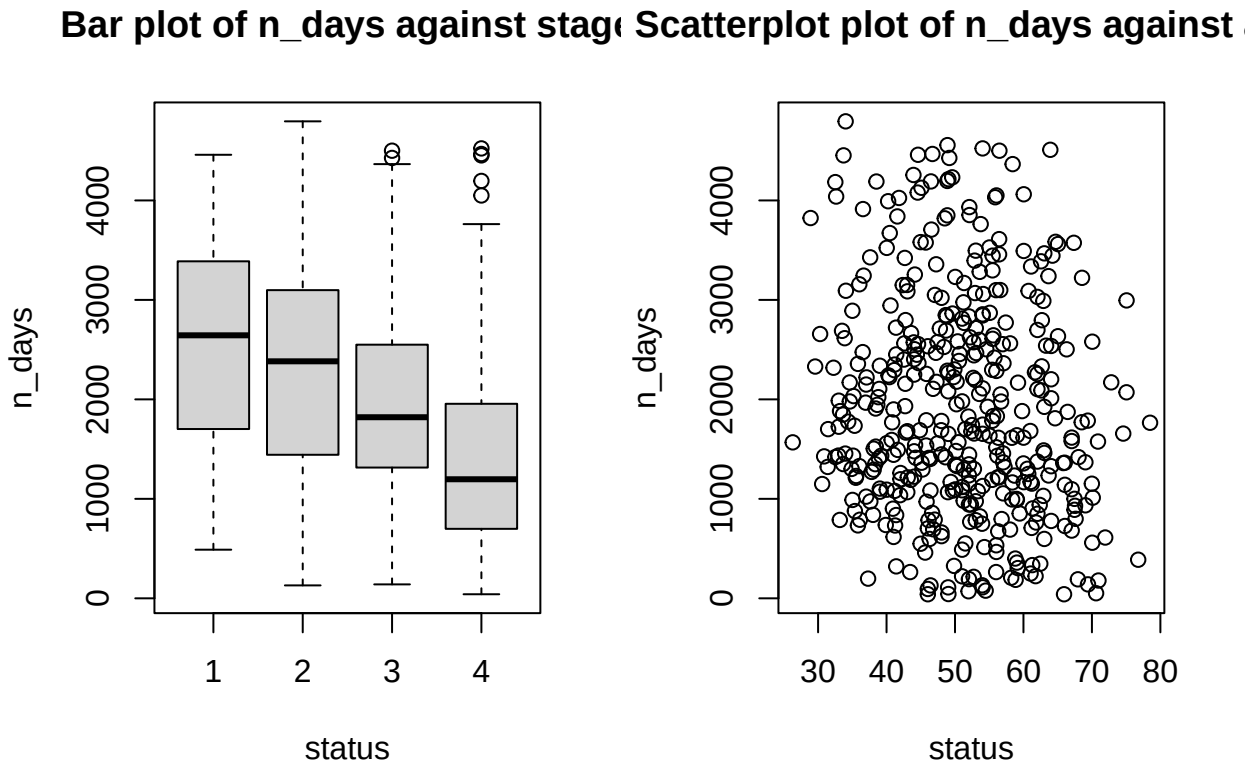


Based on these plots, a majority of the patients did not have a presence of spider angiomas or ascites. There was essentially an even balance of patients who had and did not have hepatomegaly. A majority of patients did not have edema nor have diuretic therapy.

```
par(mfrow=c(1, 2))
plot(cholangitis[,"status"], cholangitis[,"n_days"], main="Bar plot of n_days against status", xlab="sta
plot(cholangitis[,"sex"], cholangitis[,"n_days"], main="Bar plot of n_days against sex", xlab="status",
```

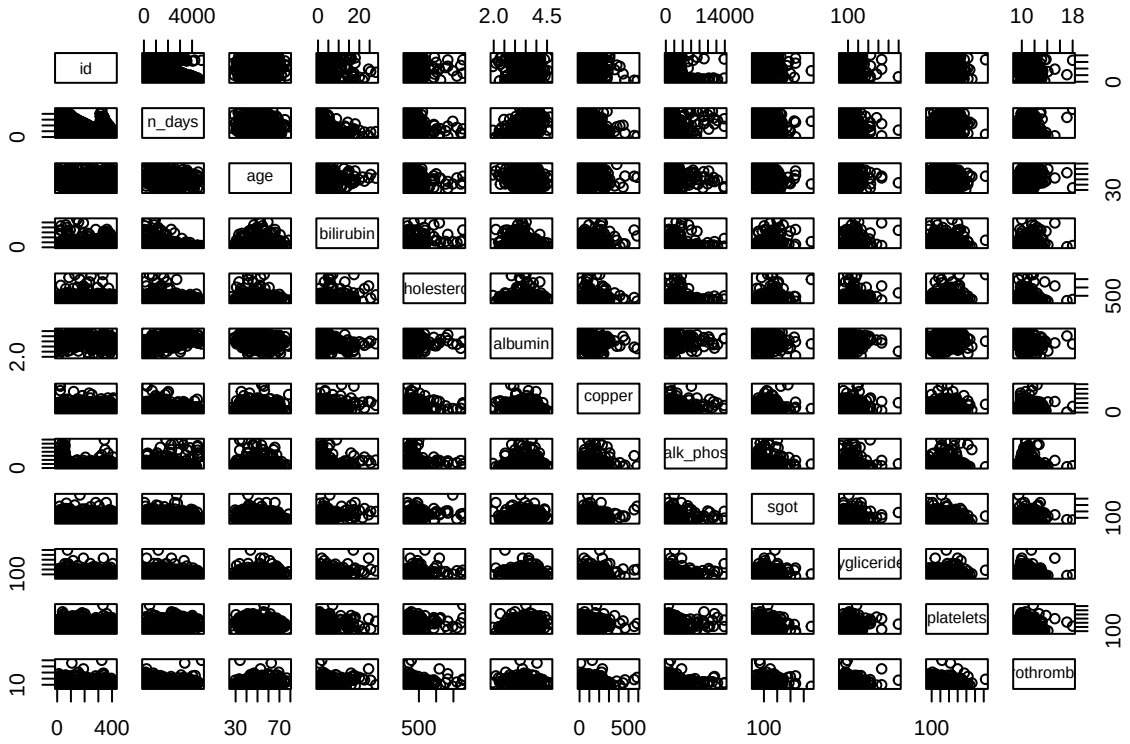



```
par(mfrow=c(1, 2))
plot(cholangitis[,"stage"], cholangitis[,"n_days"], main="Bar plot of n_days against stage", xlab="statu
plot(cholangitis[,"age"], cholangitis[,"n_days"], main="Scatterplot plot of n_days against age", xlab="
```

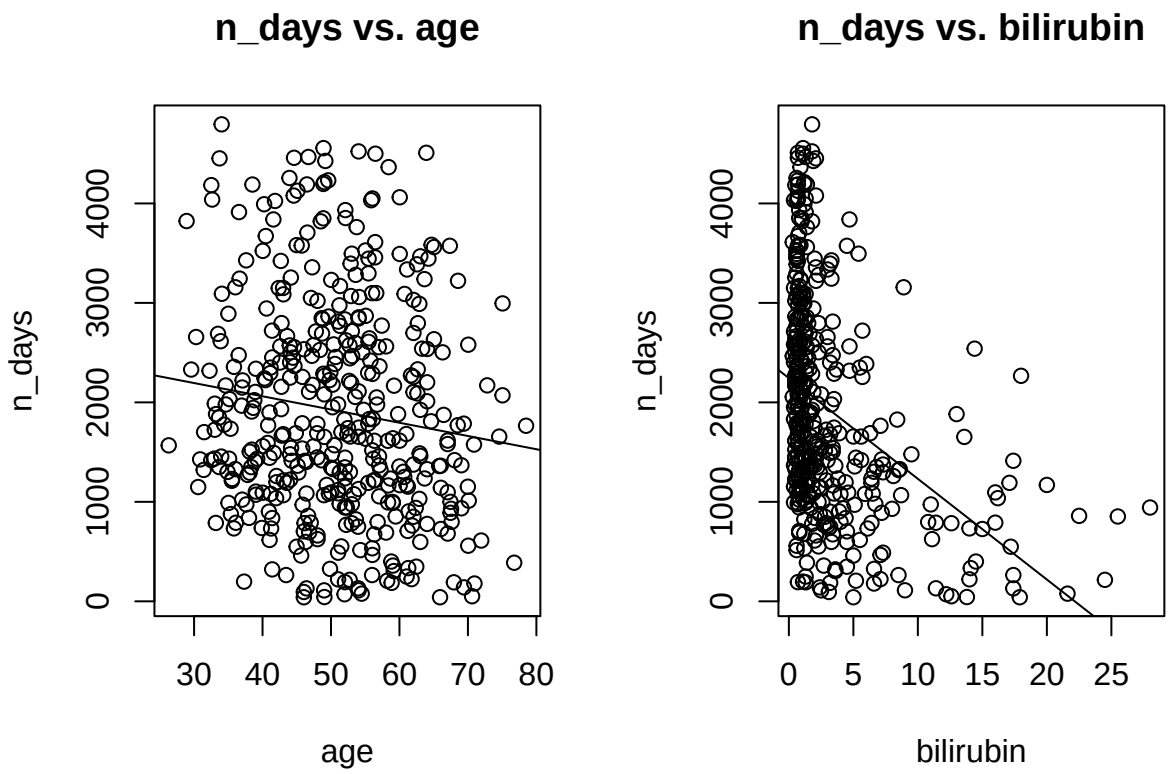


We transformed the covariates because we had to plot the number of days, a numeric variable, on the y-axis against the covariates, categorical variables that acted as the independent variables, on the x-axis. We've included the plots of n_days against each categorical variable to preface the regression run on n_days correlating to all of the categorical variables shown in a linear model. We believe that showing the individual plots in addition to the regression analysis of n_days correlated to all categorical variables would paint a more complete picture of just how n_days can be modeled and linked/related to the categorical variables in the dataset.

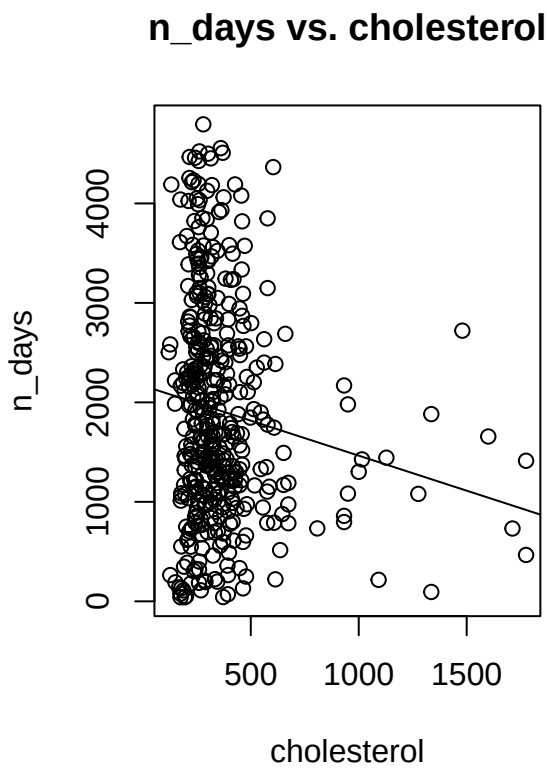
```
chol_continuous <- select(cholangitis, -all_of(cat_vars))
pairs(chol_continuous)
```

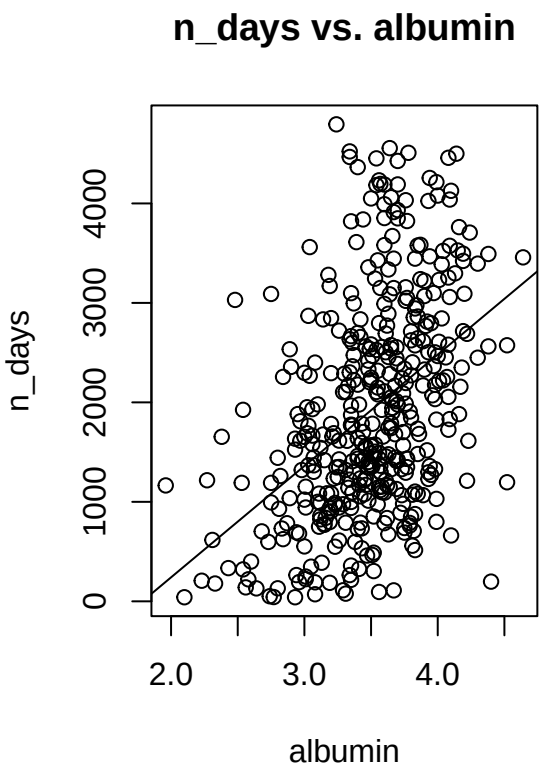



```
par(mfrow=c(1,2))
plot(chol_continuous$age, chol_continuous$n_days, main = "n_days vs. age", xlab = "age", ylab = "n_days
abline(lm(chol_continuous$n_days ~ chol_continuous$age))
plot(chol_continuous$bilirubin, chol_continuous$n_days, main = "n_days vs. bilirubin", xlab = "bilirubi
abline(lm(chol_continuous$n_days ~ chol_continuous$bilirubin))
```

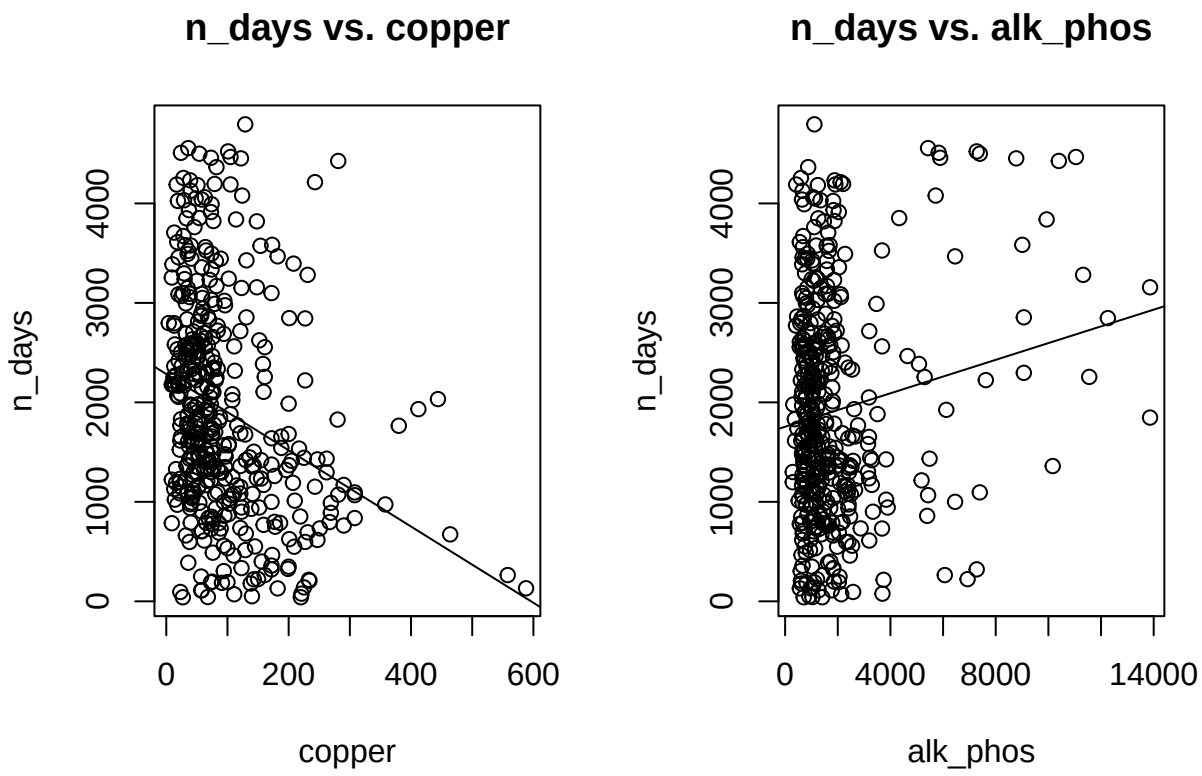



```
par(mfrow=c(1,2))
plot(chol_continuous$cholesterol, chol_continuous$n_days, main = "n_days vs. cholesterol", xlab = "chol
abline(lm(chol_continuous$n_days ~ chol_continuous$cholesterol))
plot(chol_continuous$albumin, chol_continuous$n_days, main = "n_days vs. albumin", xlab = "albumin", yl
abline(lm(chol_continuous$n_days ~ chol_continuous$albumin))
```

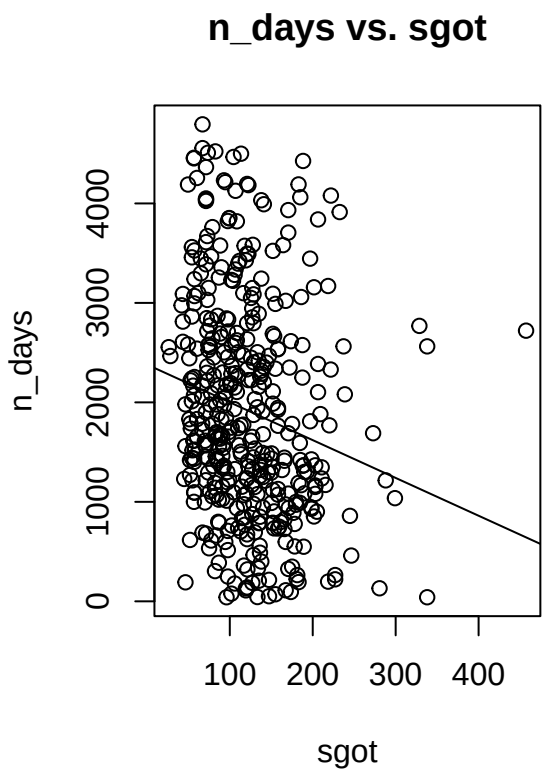


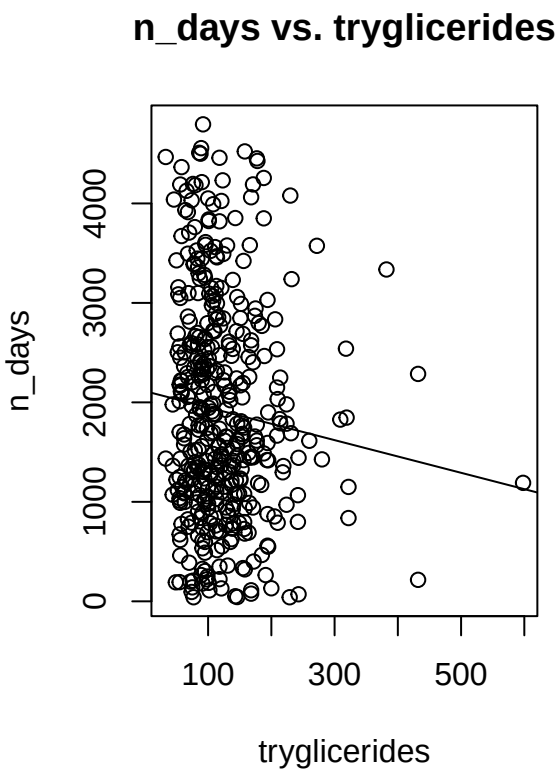



```
par(mfrow=c(1,2))
plot(chol_continuous$copper, chol_continuous$n_days, main = "n_days vs. copper", xlab = "copper", ylab =
abline(lm(chol_continuous$n_days ~ chol_continuous$copper))
plot(chol_continuous$alk_phos, chol_continuous$n_days, main = "n_days vs. alk_phos", xlab = "alk_phos",
abline(lm(chol_continuous$n_days ~ chol_continuous$alk_phos))
```

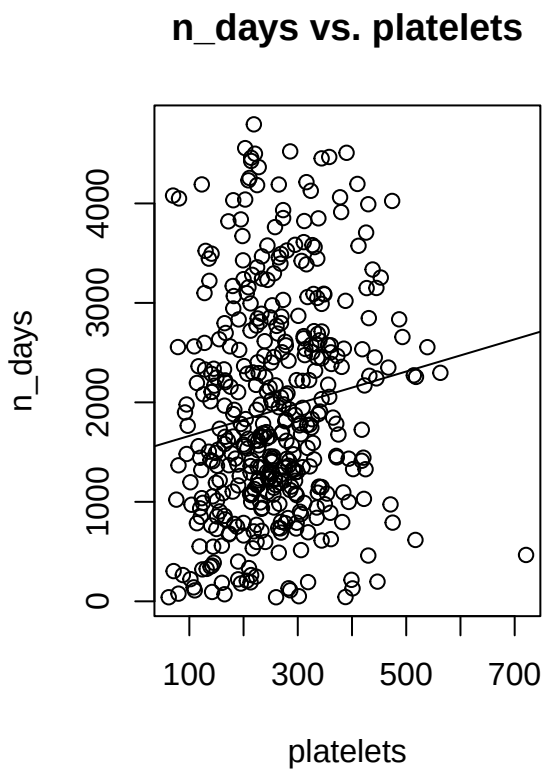


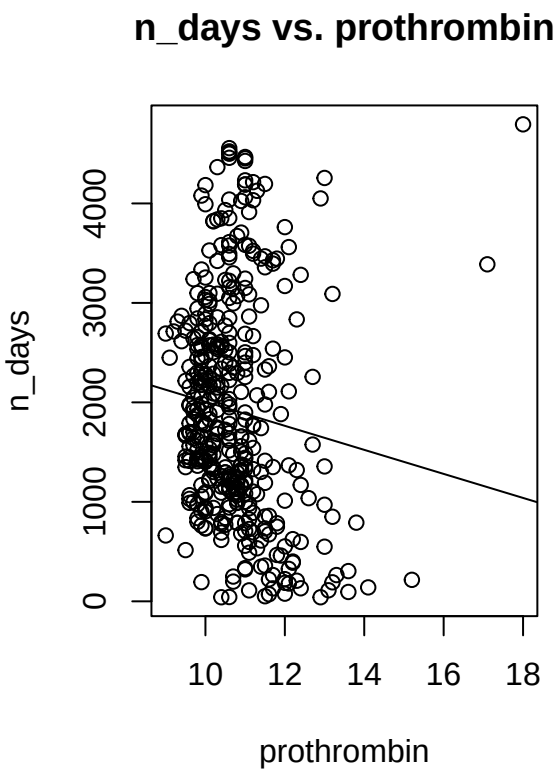

```
par(mfrow=c(1,2))
plot(chol_continuous$sgot, chol_continuous$n_days, main = "n_days vs. sgot", xlab = "sgot", ylab = "n_da
abline(lm(chol_continuous$n_days ~ chol_continuous$sgot))
plot(chol_continuous$tryglicerides, chol_continuous$n_days, main = "n_days vs. tryglicerides", xlab = "
abline(lm(chol_continuous$n_days ~ chol_continuous$tryglicerides))
```





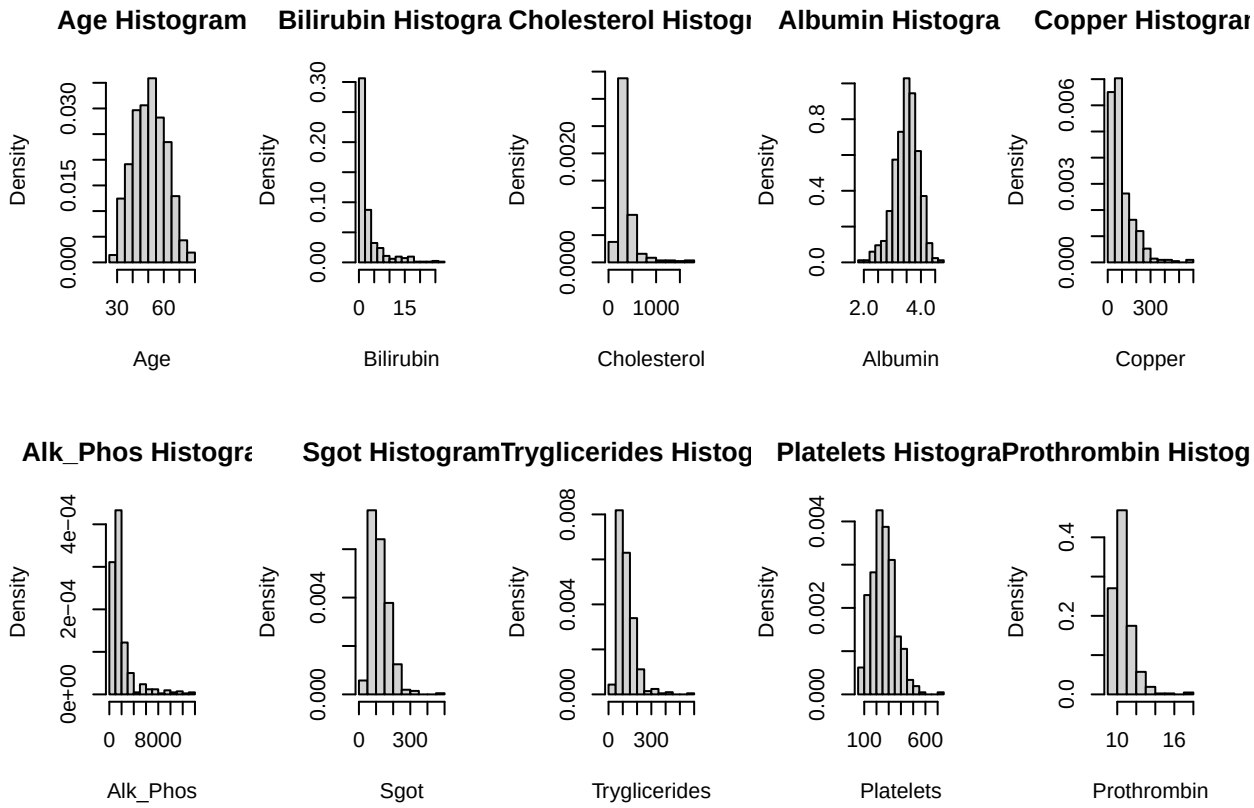

```
par(mfrow=c(1,2))
plot(chol_continuous$platelets, chol_continuous$n_days, main = "n_days vs. platelets", xlab = "platelets
abline(lm(chol_continuous$n_days ~ chol_continuous$platelets))
plot(chol_continuous$prothrombin, chol_continuous$n_days, main = "n_days vs. prothrombin", xlab = "proth
abline(lm(chol_continuous$n_days ~ chol_continuous$prothrombin))
```





Explanations for scatter plots: Albumin and Platelets seem to be the only two continuous explanatory variables that change at least moderately from the time of registration to the final day recorded. Most of the points in most of the graphs are clustered at the left edge, indicating that there was little change in the continuous explanatory variable from the time of registration to the final day recorded. Only Albumin and Platelets seemed to increase with more/higher n_days.

```
par(mfrow = c(2, 5))
hist(chol_continuous$age, freq = FALSE, main = "Age Histogram", xlab = "Age")
hist(chol_continuous$bilirubin, freq = FALSE, main = "Bilirubin Histogram", xlab = "Bilirubin")
hist(chol_continuous$cholesterol, freq = FALSE, main = "Cholesterol Histogram", xlab = "Cholesterol")
hist(chol_continuous$albumin, freq = FALSE, main = "Albumin Histogram", xlab = "Albumin")
hist(chol_continuous$copper, freq = FALSE, main = "Copper Histogram", xlab = "Copper")
hist(chol_continuous$alk_phos, freq = FALSE, main = "Alk_Phos Histogram", xlab = "Alk_Phos")
hist(chol_continuous$sgot, freq = FALSE, main = "Sgot Histogram", xlab = "Sgot")
hist(chol_continuous$triglicerides, freq = FALSE, main = "Triglicerides Histogram", xlab = "Tryglicerid
hist(chol_continuous$platelets, freq = FALSE, main = "Platelets Histogram", xlab = "Platelets")
hist(chol_continuous$prothrombin, freq = FALSE, main = "Prothrombin Histogram", xlab = "Prothrombin")
```



We see from the histograms that Age is roughly normal as expected, Albumin has a slight left skew, Platelets has a slight right skew, and every other explanatory variable has a strong right skew. There's a heavier concentration of lower values for most of the explanatory variables except for Platelets and Albumin. So from start to finish of the treatment period, since Albumin and Platelets are less skewed, the mean is closer to the median than it is in the other explanatory variables, meaning that those two variables change more during the period than the other variables do.

Question 3: Multivariate Regression

```
cholangitis_clean <- cholangitis %>%
  subset(status != "CL")
```

Before performing regression analysis on the data, we want to first clean and remove any extraneous data which would not add value. First, the patients who received a liver transplant can be dropped from the data as they do not accurately represent the rest of the patients or our target patients.

Another important aspect to consider will be which variables are important to our data. In particular, the id variable adds no value as it just assigns a number to each patient, as if indexing them. This does not contribute to the prediction of the number of days, our predictor, and this can be excluded.

Based on the mosiac plot seen in the EDA, it seems as though the drug does not have much of an impact as well. Finally, the status of an individual cannot be known until they are discharged from the study, at which point the number of days they survived is also known. Because of this redundancy, they can both be removed.

In making a model, looking at different combinations of explanatory variables is useful to gauge different

fits. Below are three models, one of all the continuous variables, one of all variables except drugs and status, and the last is looking at the status, sex, stage, and age.

```
continuous_fit <- lm(n_days ~ . - id, chol_continuous)
summary(continuous_fit)
```

```
##
## Call:
## lm(formula = n_days ~ . - id, data = chol_continuous)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2863.55  -618.31   -46.45   555.46  2728.36
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.339e+03  7.639e+02  -1.753 0.080327 .
## age           -7.080e+00  4.514e+00  -1.568 0.117577
## bilirubin     -6.937e+01  1.289e+01  -5.383 1.25e-07 ***
## cholesterol   -4.072e-01  2.247e-01  -1.812 0.070744 .
## albumin        7.816e+02  1.150e+02   6.795 3.91e-11 ***
## copper        -2.337e+00  6.076e-01  -3.846 0.000139 ***
## alk_phos       1.211e-01  2.225e-02   5.446 9.00e-08 ***
## sgot          -2.813e-02  9.158e-01  -0.031 0.975507
## tryglicerides  7.777e-01  7.720e-01   1.007 0.314344
## platelets      1.036e+00  4.876e-01   2.126 0.034143 *
## prothrombin    8.219e+01  4.732e+01   1.737 0.083150 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 906.4 on 402 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.3467, Adjusted R-squared:  0.3304
## F-statistic: 21.33 on 10 and 402 DF,  p-value: < 2.2e-16
```

```
fit <- lm(n_days ~ . - id - drug - status, cholangitis_clean)
summary(fit)
```

```
##
## Call:
## lm(formula = n_days ~ . - id - drug - status, data = cholangitis_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2748.28  -579.77   -28.04   530.89  2462.21
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -40.65562  989.10088  -0.041 0.967244
## age             -6.44485    5.80105  -1.111 0.267573
## sexM            62.75787  187.14980   0.335 0.737635
## ascitesY      -140.04002  277.74717  -0.504 0.614535
## hepatomegalyY  -84.78046  129.82319  -0.653 0.514287
```

```
## spidersY      -113.24261  136.12580  -0.832 0.406208
## edemaS        -144.68505  203.14623  -0.712 0.476947
## edemaY        -369.76061  299.00056  -1.237 0.217298
## bilirubin      -56.97116   17.75823  -3.208 0.001498 **
## cholesterol     -0.27108    0.29786  -0.910 0.363584
## albumin        570.07451  157.23726   3.626 0.000345 ***
## copper          -2.28609    0.79010  -2.893 0.004124 **
## alk_phos         0.11885    0.02564   4.636 5.55e-06 ***
## sgot             0.04091    1.11614   0.037 0.970788
## tryglicerides    0.67952    0.98114   0.693 0.489176
## platelets        0.41715    0.62529   0.667 0.505263
## prothrombin     96.60946   63.21132   1.528 0.127603
## stage2        -303.35609  263.20209  -1.153 0.250118
## stage3        -452.44744  256.67822  -1.763 0.079090 .
## stage4        -631.96335  278.39037  -2.270 0.023998 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 904.7 on 268 degrees of freedom
##   (105 observations deleted due to missingness)
## Multiple R-squared:  0.415,  Adjusted R-squared:  0.3735
## F-statistic: 10.01 on 19 and 268 DF,  p-value: < 2.2e-16
```

```
responseModel <- lm(n_days ~ status + sex + stage + age, cholangitis_clean)
summary(responseModel)
```

```
##
## Call:
## lm(formula = n_days ~ status + sex + stage + age, data = cholangitis_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1870.5  -745.9  -181.3   578.0  2636.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2880.312    316.615   9.097  < 2e-16 ***
## statusD     -774.125    109.072  -7.097 6.14e-12 ***
## sexM         201.881    166.181   1.215  0.22518
## stage2       -94.009    240.480  -0.391  0.69607
## stage3      -414.318    231.816  -1.787  0.07468 .
## stage4      -773.761    240.314  -3.220  0.00139 **
## age           -3.853      5.042  -0.764  0.44523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 984.9 on 386 degrees of freedom
## Multiple R-squared:  0.2383, Adjusted R-squared:  0.2265
## F-statistic: 20.13 on 6 and 386 DF,  p-value: < 2.2e-16
```

Looking at the residuals, all of them are quite spread out suggesting that the models are not performing the greatest, or that our dataset is not as conducive to linear models as seen through the less linear pattern of the continuous variables. The median closest to zero can be seen in the model taking into account more explanatory variables.

###Question 3b: Variable Selection

```
step_model <- ols_step_forward_p(fit)
step_model
```

```
##
##                                Selection Summary
## ------------------------------------------------------------------------------
##         Variable                     Adj.
## Step       Entered       R-Square    R-Square     C(p)         AIC         RMSE
## ------------------------------------------------------------------------------
##    1    albumin            0.1874      0.1854    99.0111    6557.1680    1010.7338
##    2    bilirubin          0.2739      0.2701    49.1162    6514.9833     956.7074
##    3    alk_phos           0.3179      0.3126    24.6716    6492.4001     928.4392
##    4    stage              0.3560      0.3459     3.8084    6475.8315     905.6589
##    5    copper             0.3706      0.3592    -3.0005    6468.7780     896.4490
##    6    prothrombin        0.3775      0.3645    -5.1125    6466.4795     892.7199
##    7    edema              0.3842      0.3681    -7.1866    6466.1736     890.1639
##    8    cholesterol        0.3848      0.3669    -2.9413    6388.2792     893.8902
##    9    tryglicerides      0.3880      0.3684    -2.7990    6388.3105     892.8133
##   10    age                0.3900      0.3688    -2.0230    6389.0079     892.5068
## ------------------------------------------------------------------------------
```

Utilizing the olsrr package, we can perform stepwise regression on the model containing all of the varaibles to determine the ideal composition of variables from the dataset. The metrics R-Squared, Adj. R-Squared, AIC, C(p), and RMSE. The stepwise regression model is selected by choosing to include and exclude variables in subsequent steps and calculating the metrics for the model. The olsrr package simplifies this to easily output the variables for the ideal model.

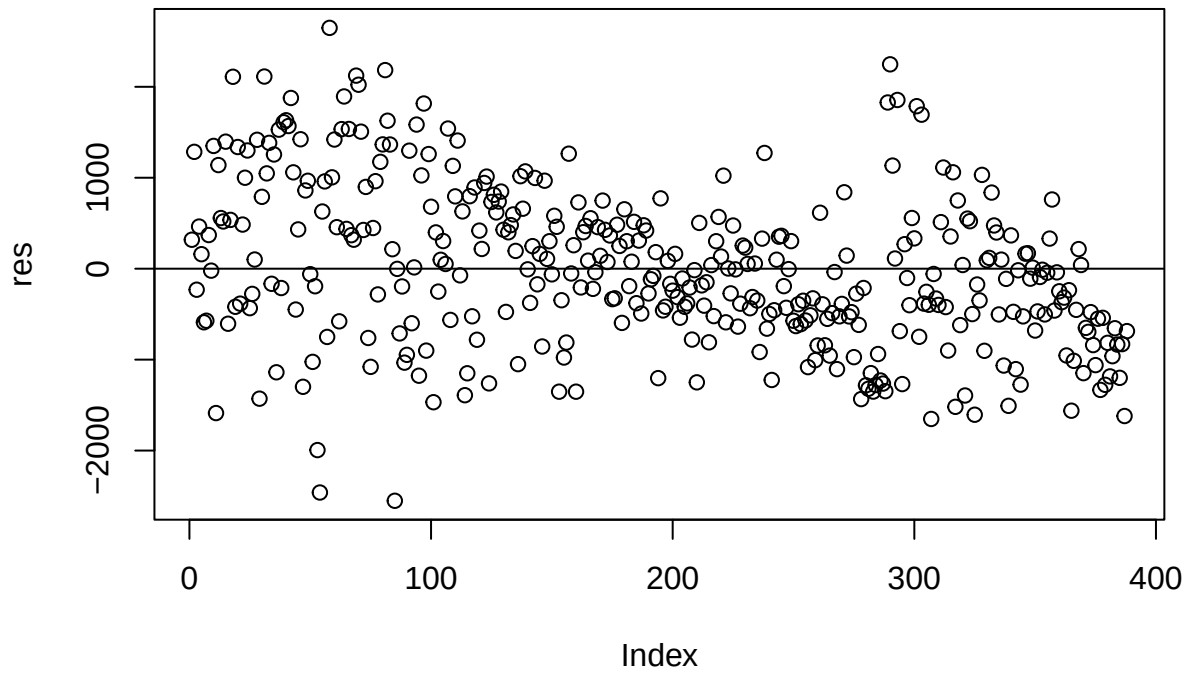
```
covariates <- step_model$predictors
step_fit <- lm(n_days ~ ., cholangitis_clean[, c("n_days", covariates)])
summary(step_fit)
```

```
##
## Call:
## lm(formula = n_days ~ ., data = cholangitis_clean[, c("n_days",
##     covariates)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2551.97  -570.51   -55.34   516.13  2647.91
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -805.60274  775.19903  -1.039  0.29937
## albumin        647.04288  122.98204   5.261 2.41e-07 ***
## bilirubin      -67.06789   13.19657  -5.082 5.91e-07 ***
## alk_phos         0.12158    0.02215   5.488 7.52e-08 ***
## stage2        -120.51697  218.34918  -0.552  0.58131
## stage3        -399.28495  211.72772  -1.886  0.06009 .
## stage4        -639.46428  219.14069  -2.918  0.00374 **
## copper          -1.95185    0.63384  -3.079  0.00223 **
## prothrombin    124.80335   49.29472   2.532  0.01176 *
```

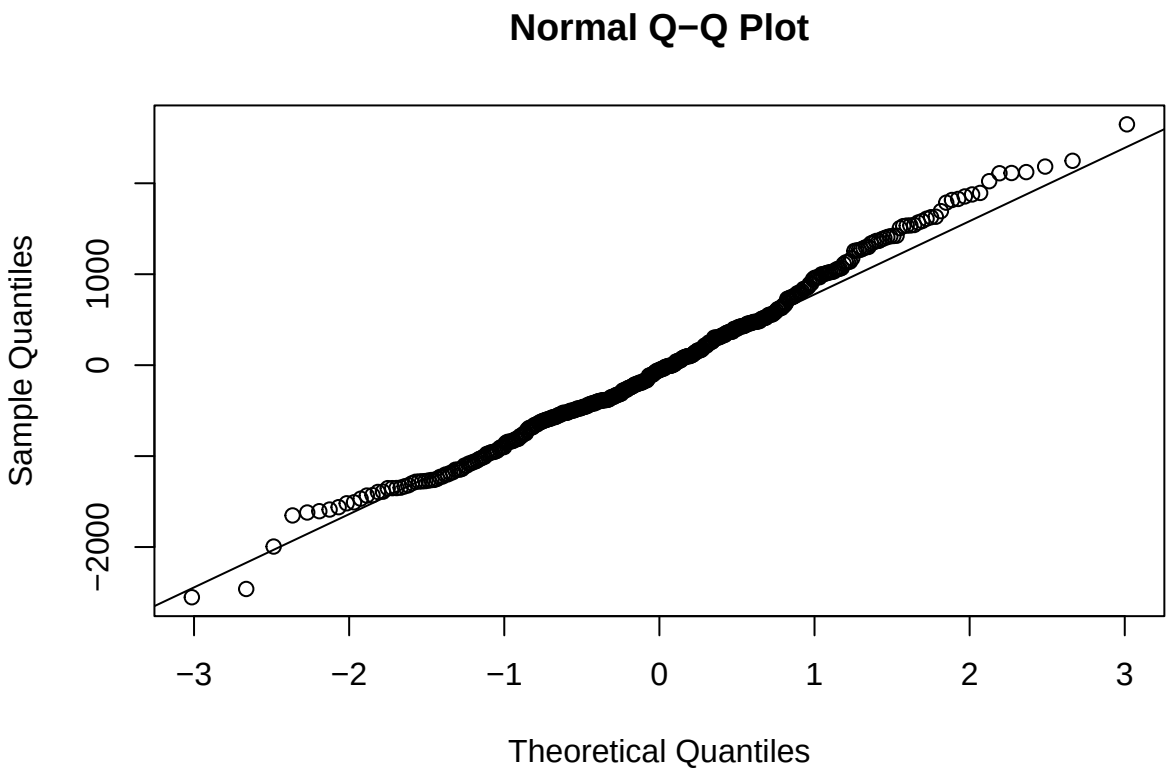
```
## edemaS        -252.72177  157.76266  -1.602  0.11002
## edemaY        -405.26755  237.31054  -1.708  0.08851 .
## cholesterol     -0.34371    0.22954  -1.497  0.13513
## tryglicerides    1.13132    0.78817   1.435  0.15201
## age             -5.26434    4.69425  -1.121  0.26282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 892.5 on 374 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:   0.39,  Adjusted R-squared:  0.3688
## F-statistic: 18.39 on 13 and 374 DF,  p-value: < 2.2e-16
```

###Question 3c: Regression Diagnostics

```
res <- resid(step_fit)
plot(res)
abline(0, 0)
```




```
qqnorm(res)
qqline(res)
```



The residual plot and qq plot give us an indication on the distribution the residuals. This plot shows that the residuals deviate somewhat from the normal distribution. Judging from the EDA previously, we saw that there was not a large number of continuous variables which had linear relationships which suggests that the linear assumption of regression is violated. Furthermore, the R-Squared and Adj. R-Squared value is low which indicates that the fit for the model on n_days is not very good. Polynomial regression or some model would likely fit better with the data.