Final Project

Yu Gu

STAT 131A Fall 2021”

Visualization

1. Importing the data: Read the data into R. Make sure your categorical variables are factors. (5 points).

1.1 Import data Note that ID is removed here, since it will not be useful for data analysis.

```r
#import data
cholang <- read.csv('cholangitis.csv', header = T, stringsAsFactors = T, na.strings = "NA")
cholang <- cholang[,-1]
summary(cholang)
```

```
##      n_days     status               drug          age        sex     ascites
##  Min.   :  41   C :232   D-penicillamine:158   Min.   : 9598   F:374   N:390
##  1st Qu.:1093   CL: 25   Placebo        :154   1st Qu.:15644   M: 44   Y: 28
##  Median :1730   D :161   NA's           :106   Median :18628
##  Mean   :1918                                  Mean   :18533
##  3rd Qu.:2614                                  3rd Qu.:21273
##  Max.   :4795                                  Max.   :28650
##
##  hepatomegaly spiders edema     bilirubin       cholesterol        albumin
##  N:203        N:298   N:354   Min.   : 0.300   Min.   : 120.0   Min.   :1.960
##  Y:215        Y:120   S: 44   1st Qu.: 0.800   1st Qu.: 248.0   1st Qu.:3.243
##                       Y: 20   Median : 1.400   Median : 310.0   Median :3.530
##                               Mean   : 3.221   Mean   : 365.5   Mean   :3.497
##                               3rd Qu.: 3.400   3rd Qu.: 400.0   3rd Qu.:3.770
##                               Max.   :28.000   Max.   :1775.0   Max.   :4.640
##                                                NA's   :5
##      copper           alk_phos            sgot         tryglicerides
##  Min.   :  4.00   Min.   :  289.0   Min.   : 26.35   Min.   : 33.0
##  1st Qu.: 41.25   1st Qu.:  857.2   1st Qu.: 82.04   1st Qu.: 84.0
##  Median : 72.50   Median : 1257.0   Median :114.11   Median :109.0
##  Mean   : 95.71   Mean   : 1937.1   Mean   :121.75   Mean   :122.9
##  3rd Qu.:123.00   3rd Qu.: 2039.0   3rd Qu.:151.90   3rd Qu.:151.0
##  Max.   :588.00   Max.   :13862.4   Max.   :457.25   Max.   :598.0
##                                                      NA's   :5
##    platelets      prothrombin        stage
##  Min.   : 62.0   Min.   : 9.00   Min.   :1.000
##  1st Qu.:190.0   1st Qu.:10.00   1st Qu.:2.000
##  Median :250.0   Median :10.60   Median :3.000
##  Mean   :257.4   Mean   :10.73   Mean   :3.026
##  3rd Qu.:318.0   3rd Qu.:11.10   3rd Qu.:4.000
##  Max.   :721.0   Max.   :18.00   Max.   :4.000
##
```

1.2 Data Cleaning I first change the NA data in the ‘drug’ column into ‘NotParticipated’, representing another factor level. Then, I change the NA data in numerical columns into the median value and the NA data in categorical columns into the most frequent factor level. Actually, we can directly remove these NA data, but since the dataset is rather small, I somehow don't want to kick out some rows randomly, therefore, I choose to do some transformation in the NA data.

```
#change the NA data into another factor level in the 'drug' column
cholang$drug <- as.character(cholang$drug)
cholang$drug <- ifelse(is.na(cholang$drug), 'NotParticipated', cholang$drug)
cholang$drug <- as.factor(cholang$drug)

#change the NA data in numerical columns into the median value
cholang_num <- select_if(cholang, is.numeric)
head(cholang_num)
```

```
##   n_days   age bilirubin cholesterol albumin copper alk_phos   sgot
## 1    400 21464      14.5         261    2.60    156   1718.0 137.95
## 2   4500 20617       1.1         302    4.14     54   7394.8 113.52
## 3   1012 25594       1.4         176    3.48    210    516.0  96.10
## 4   1925 19994       1.8         244    2.54     64   6121.8  60.63
## 5   1504 13918       3.4         279    3.53    143    671.0 113.15
## 6   2503 24201       0.8         248    3.98     50    944.0  93.00
##   tryglicerides platelets prothrombin stage
## 1           172       190        12.2     4
## 2            88       221        10.6     3
## 3            55       151        12.0     4
## 4            92       183        10.3     4
## 5            72       136        10.9     3
## 6            63       361        11.0     3
```

```
cholang_num <- as.data.frame(apply(cholang_num, 2, function(x){
  x[is.na(x)] <- median(x, na.rm = T)
  return(x)
}))

#change the NA data in categorical columns into the most frequent factor level
cholang_cate <- select_if(cholang, is.factor)
head(cholang_cate)
```

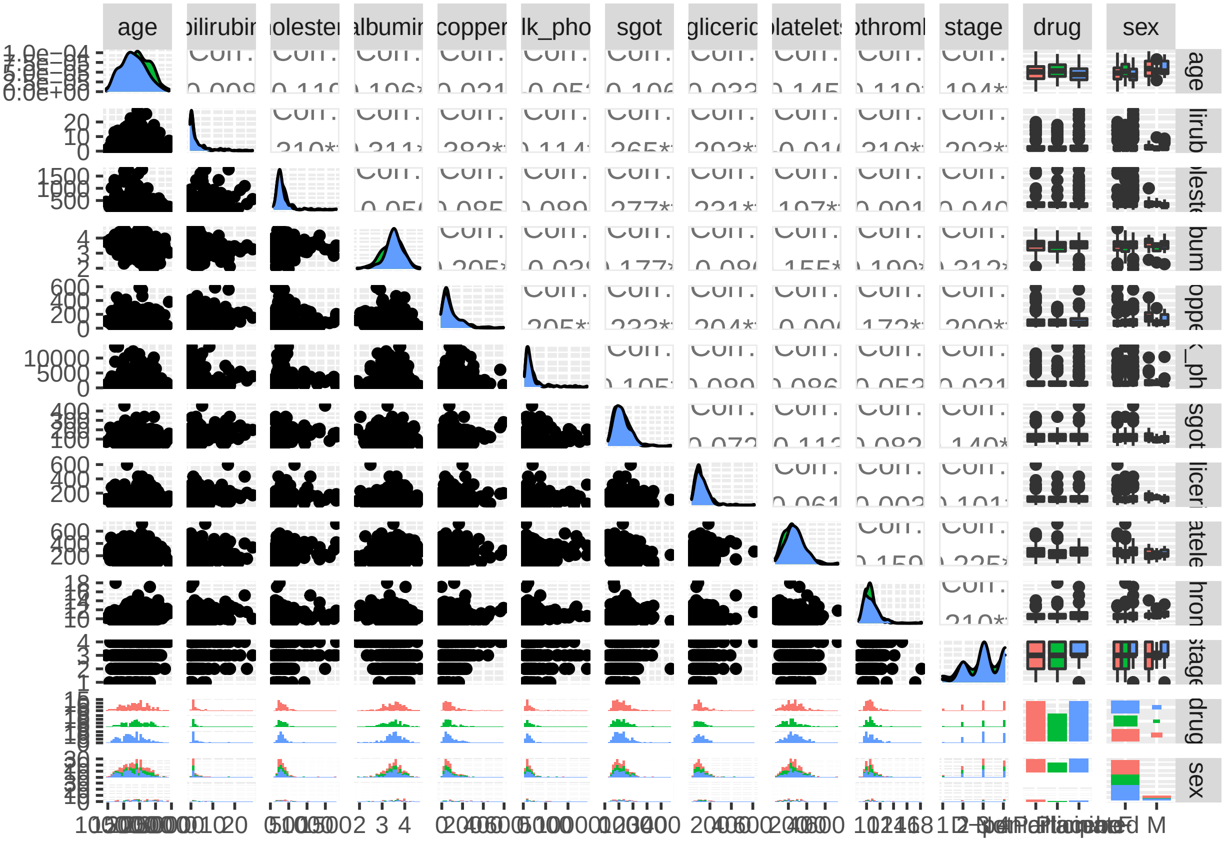
```
##   status            drug sex ascites hepatomegaly spiders edema
## 1      D D-penicillamine   F       Y            Y       Y     Y
## 2      C D-penicillamine   F       N            Y       Y     N
## 3      D D-penicillamine   M       N            N       N     S
## 4      D D-penicillamine   F       N            Y       Y     S
## 5     CL         Placebo   F       N            Y       Y     N
## 6      D         Placebo   F       N            Y       N     N
```

```
cholang_cate <- as.data.frame(apply(cholang_cate, 2, function(x){
  x[is.na(x)] <- names(which.max(table(x)))
  return(x)
}))
```

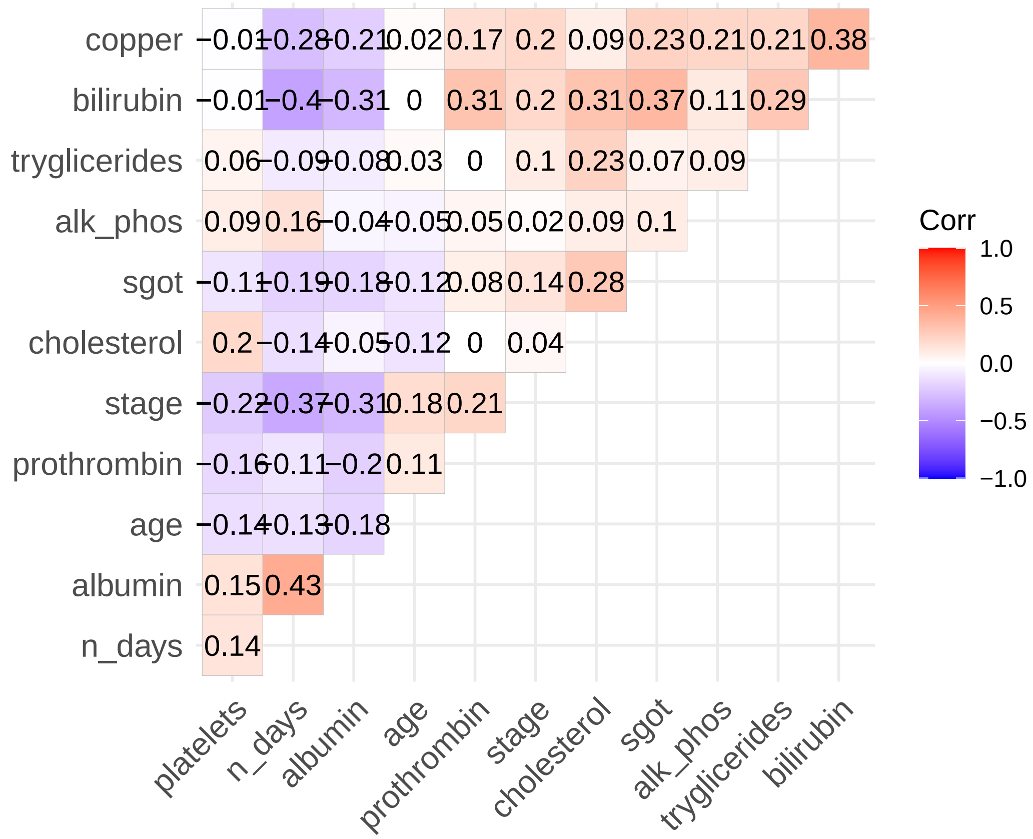
2. Basic exploratory data analysis: Perform exploratory data analysis of the data, using any appropriate tools we have learned. Note any interesting features of the data. (20 points).

```
# pairs plot
na.omit(cholang) %>%
  select(age, bilirubin, cholesterol, albumin, copper, alk_phos, sgot, tryglicerides, platelets, prothr
ggpairs(aes(fill = drug))
```

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

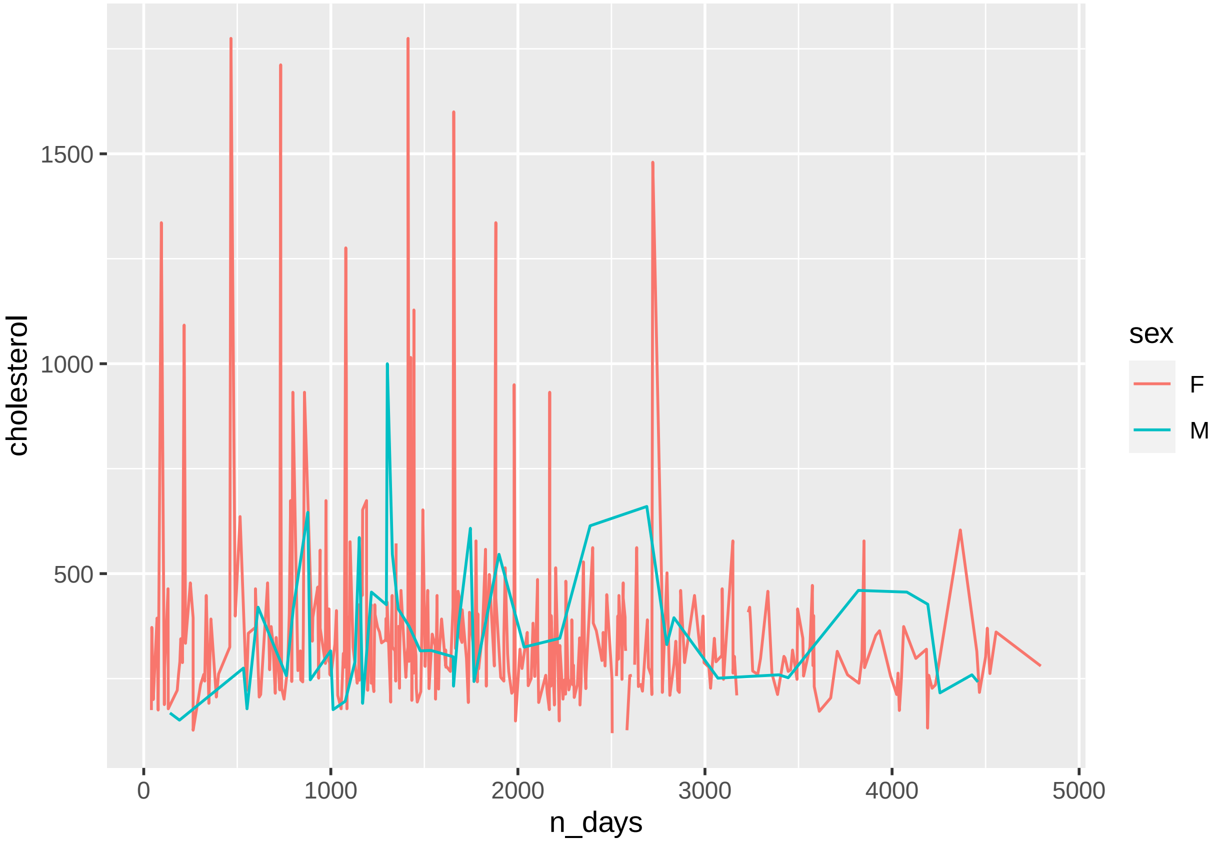


```
cholang_num %>%
  cor() %>%
  ggcorrplot(type = "upper",
             hc.order = T,
             lab = T,
             sig.level = .5)
```

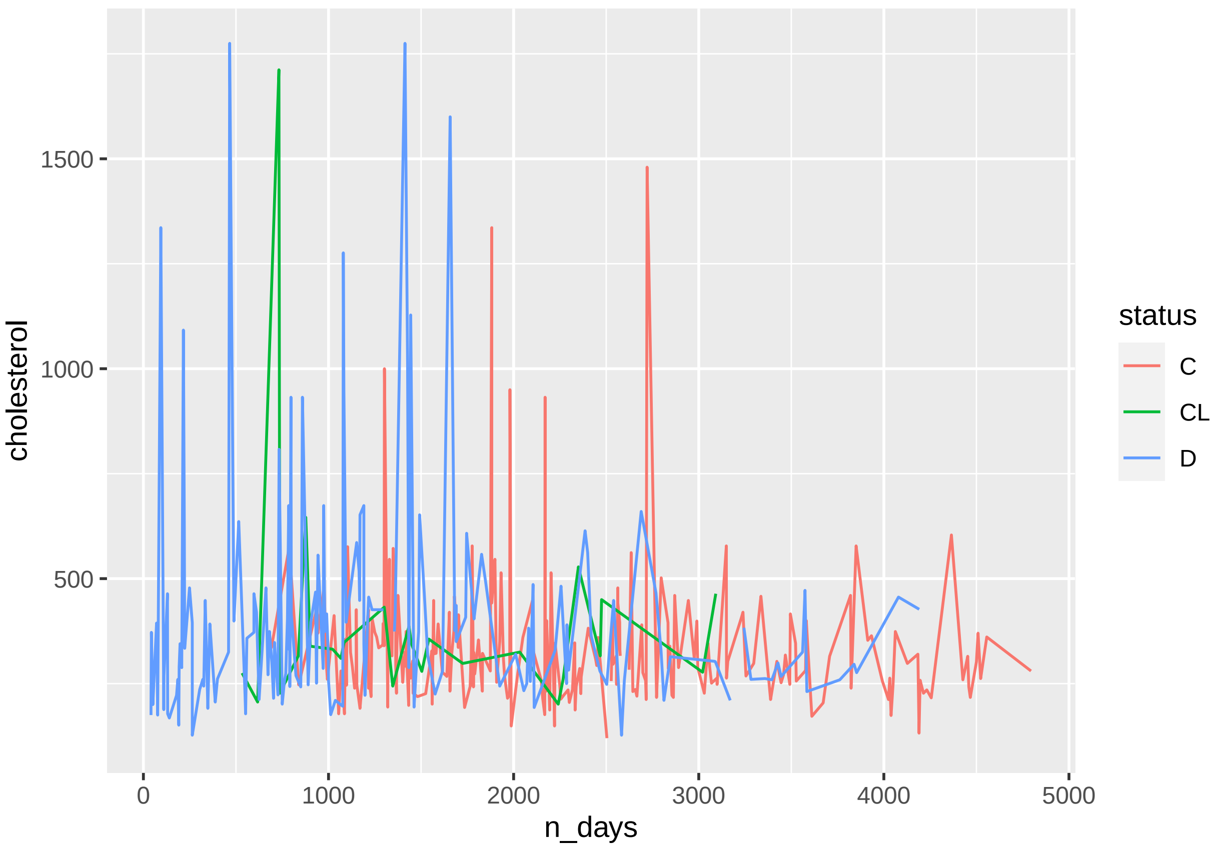


From the corrolation plot of numeric data, we can see nearly all numeric data has a small correlation value, therefore, they can be considered as independent variables.

```
ggplot(cholang, aes(n_days, cholesterol, color = sex)) + geom_line()
```

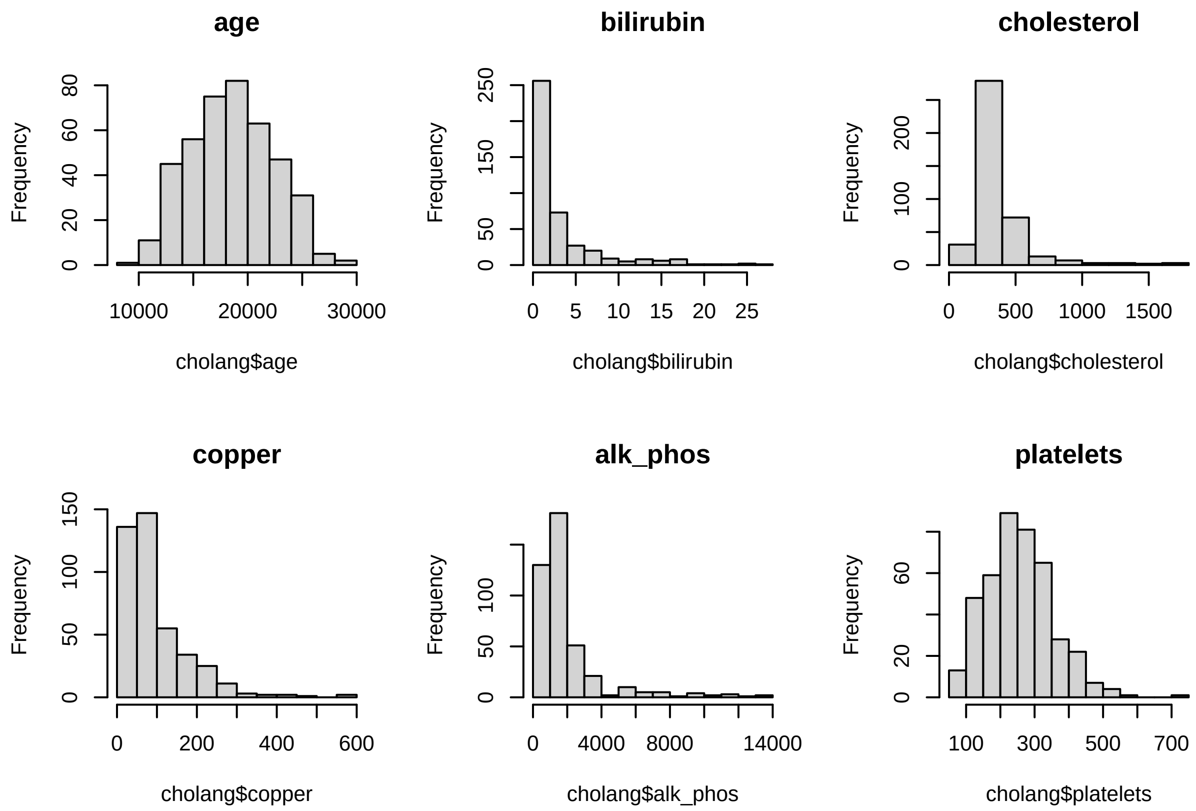


```r
ggplot(cholang, aes(n_days, cholesterol, color = status)) + geom_line()
```



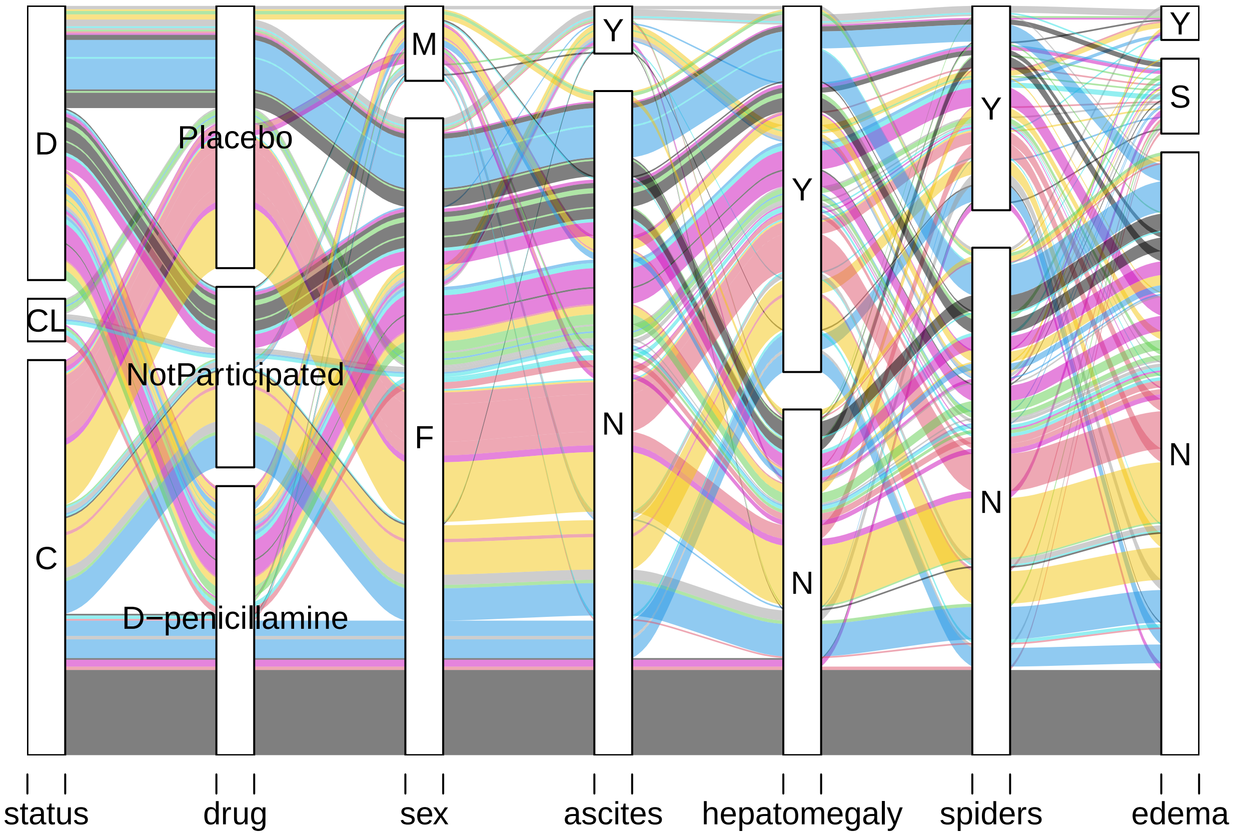
From the previous plot of cholesterol change with days, we can see that female seems to have a higher cholesterol level than man, and the cholesterol level drops gradually as time goes, ignoring some outliers.

```r
par(mfrow=c(2,3))
hist(cholang$age,main="age")
hist(cholang$bilirubin,main="bilirubin")
hist(cholang$cholesterol,main="cholesterol")
hist(cholang$copper,main="copper")
hist(cholang$alk_phos,main="alk_phos")
hist(cholang$platelets,main="platelets")
```

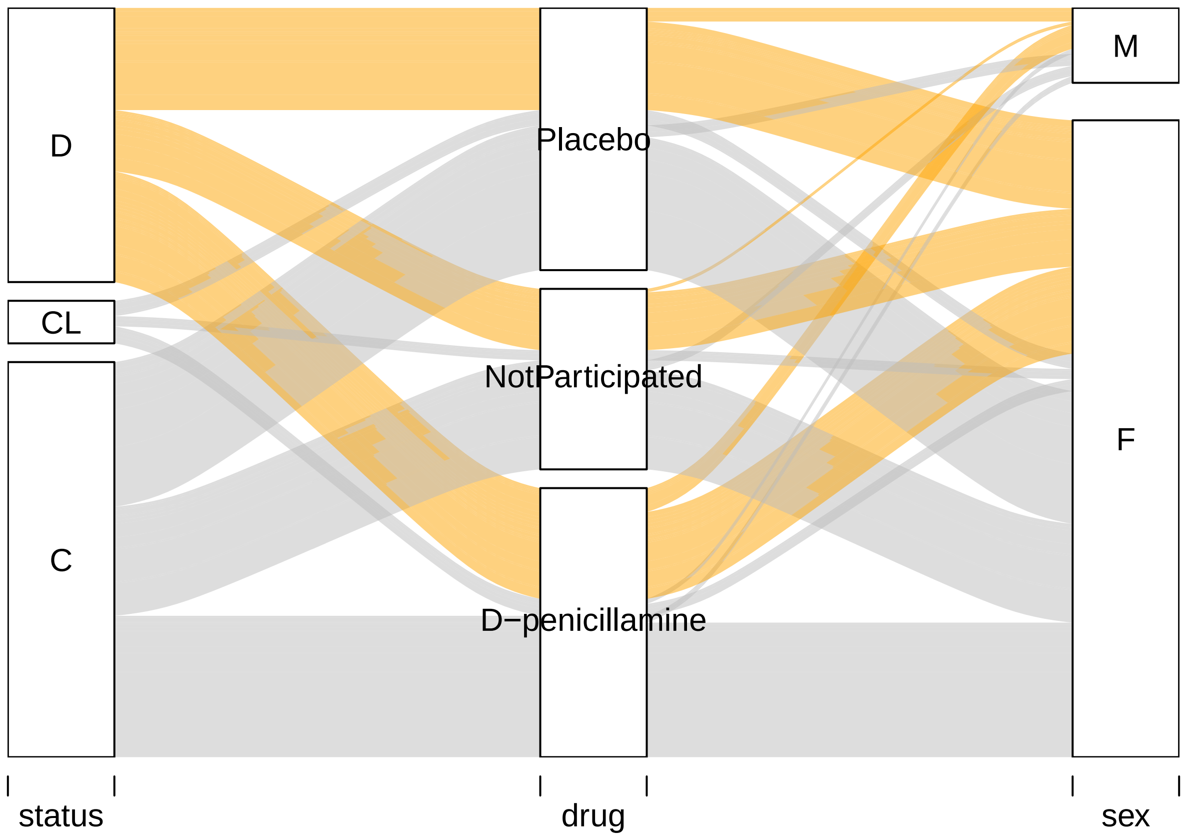


From the histogram before, we can see that most of the numerical values are not normally distributed. The age is normally distributed, which will be good for analysis. And most of the chemicals are right-skewed.(I didn't show all of the plots, but the trend is basically the same.)

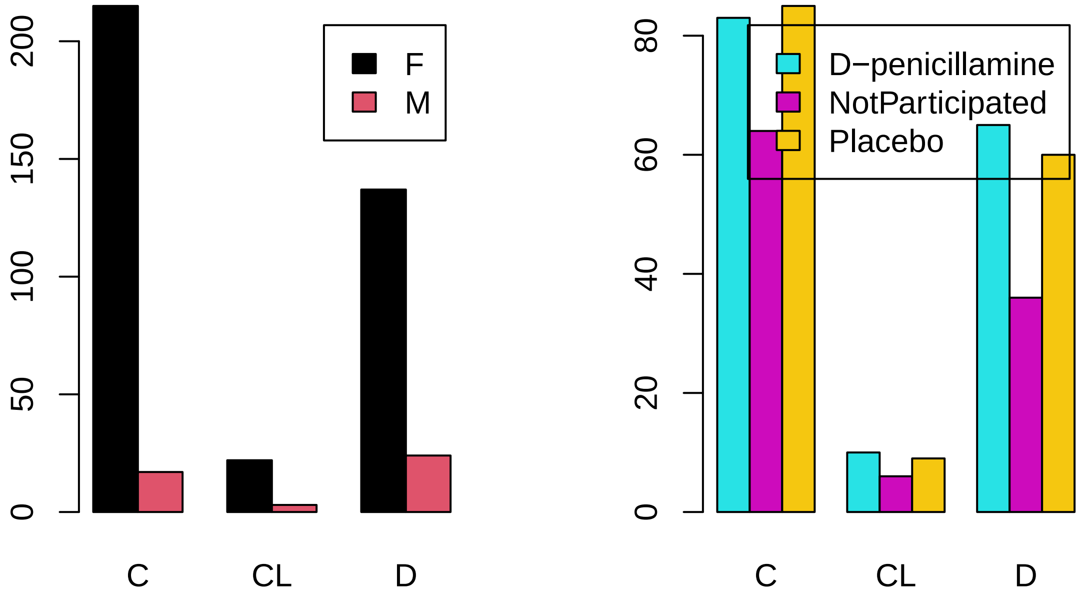
```
cholang_cate$Freq<-1
cholangCateAggregates<-aggregate(Freq ~ .,data=cholang_cate,FUN=sum)
alluvial(cholangCateAggregates[,-ncol(cholangCateAggregates)], freq=cholangCateAggregates$Freq,
         col=palette())
```



```
alluvial(cholangCateAggregates[,c(1:3)], freq=cholangCateAggregates$Freq,
         col= ifelse(cholangCateAggregates$status == "D", "orange", "grey"))
```

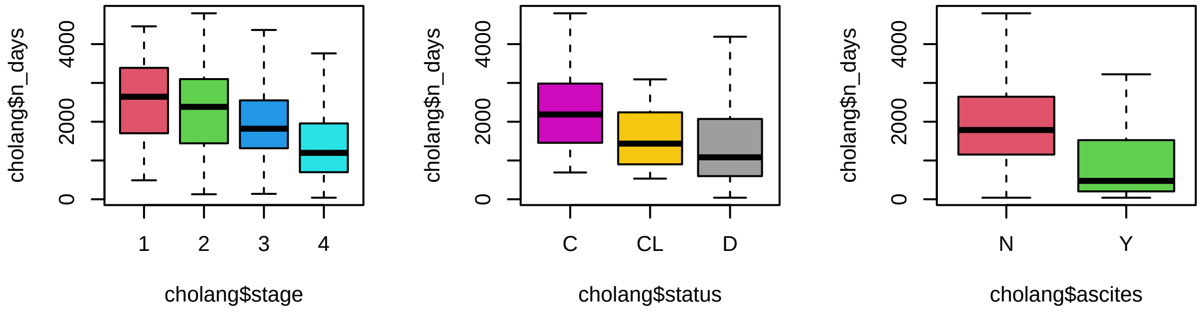


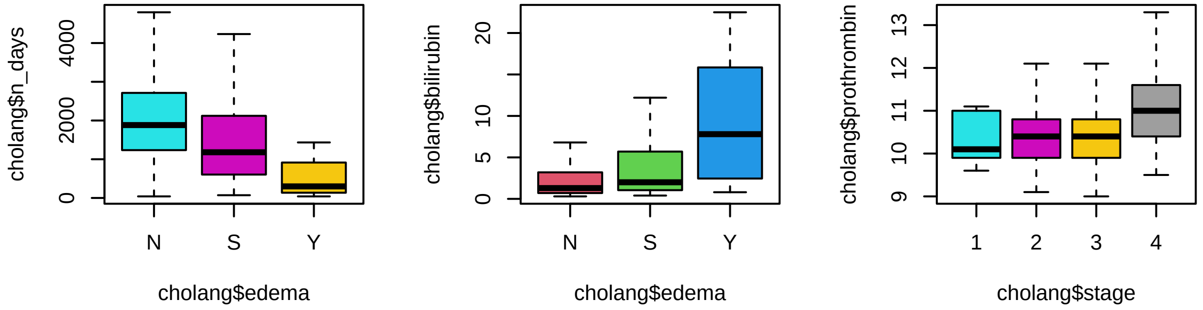
```
par(mfrow = c(1, 2))
barplot(with(cholang,table(sex,status)),beside=TRUE,legend=TRUE,col=palette()[1:2])
barplot(with(cholang,table(drug,status)),beside=TRUE,legend=TRUE,col=palette()[5:7])
```



From the alluvial plot and barplot, we can see some characteristics for categorical variables. (1) Female patients are a lot more than male patients in this dataset, and seems to have a higher rate to survive. Since the male data is small, this judgement may be biased. (2) The patients have D-penicillamine or placebo does not show a high deviation for the rate of survival. (3) A large porportion of patients who are cured later don't have symptoms like ascites,hepatomegaly,spiders,edema.

```
par(mfrow=c(2,3))
boxplot(cholang$n_days~cholang$stage,col=palette()[2:5],outline=FALSE)
boxplot(cholang$n_days~cholang$status,col=palette()[6:8],outline=FALSE)
boxplot(cholang$n_days~cholang$ascites,col=palette()[2:3],outline=FALSE)
boxplot(cholang$n_days~cholang$edema,col=palette()[5:7],outline=FALSE)
boxplot(cholang$bilirubin~cholang$edema,col=palette()[2:5],outline=FALSE)
boxplot(cholang$prothrombin~cholang$stage,col=palette()[5:8],outline=FALSE)
```





From the boxplot before we can see some trends: (1) The latter stage will have shorter days of living (which is consistent with intuition) on average. (2) The patients that are dead later have shorter days till the end of the survey on average. (3) and (4) The patients don't have ascites or edema will have a longer day for living on average. (5) The patients who have edema will have a higher level of bilirubin on average. (6) The patients in the 4th stage will have a higher level of prothrombin on average.

Multivariate Regression

1. Multivariate regression analysis

Perform a regression analysis of the response (number of days) on the explanatory variables. Describe here whether you transformed your data or covariates, or excluded any observations, and why. Here you might include diagnostic plots (i.e. for transformations you considered but did not use), but only show those that are necessary for explaining your choices. (20 points).

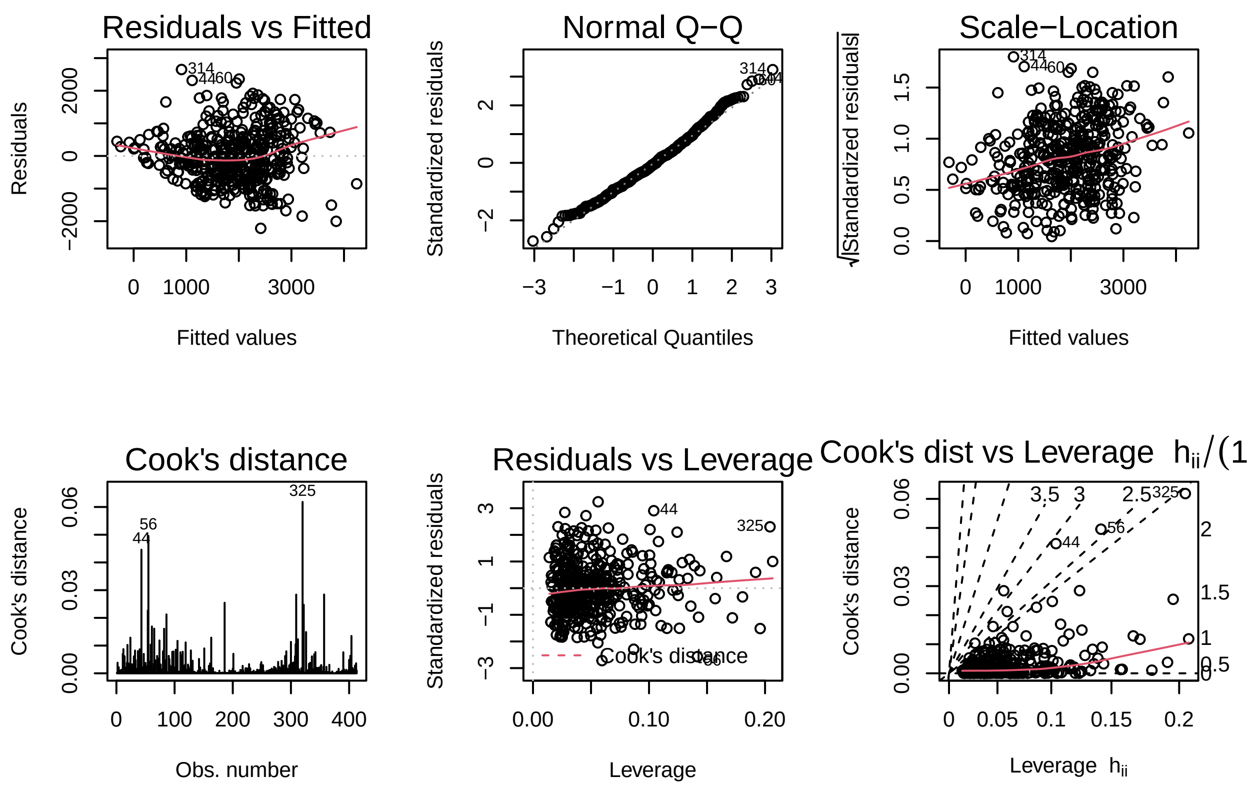
(The data cleaning part was done before.)

```r
lmFull = lm(n_days ~ ., cholang)
summary(lmFull)
```

```
##
## Call:
## lm(formula = n_days ~ ., data = cholang)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2219.1  -570.8   -52.1   509.4  2650.5
##
## Coefficients:
```

```
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         -1.061e+03  7.522e+02  -1.411 0.159123    
## statusCL            -4.520e+02  1.872e+02  -2.415 0.016203 *  
## statusD             -6.034e+02  1.070e+02  -5.639 3.28e-08 ***
## drugNotParticipated -3.359e+02  1.095e+02  -3.067 0.002312 ** 
## drugPlacebo          1.086e+01  9.854e+01   0.110 0.912274    
## age                  4.030e-03  1.260e-02   0.320 0.749248    
## sexM                 8.172e+01  1.455e+02   0.562 0.574564    
## ascitesY             6.023e+01  2.167e+02   0.278 0.781186    
## hepatomegalyY       -2.362e+01  9.197e+01  -0.257 0.797465    
## spidersY             1.102e+01  1.017e+02   0.108 0.913700    
## edemaS              -2.131e+02  1.447e+02  -1.473 0.141519    
## edemaY              -4.966e+02  2.585e+02  -1.921 0.055424 .  
## bilirubin           -4.594e+01  1.269e+01  -3.621 0.000332 ***
## cholesterol         -3.248e-01  2.128e-01  -1.526 0.127711    
## albumin              5.625e+02  1.131e+02   4.975 9.78e-07 ***
## copper              -1.864e+00  5.974e-01  -3.121 0.001936 ** 
## alk_phos             1.291e-01  2.101e-02   6.146 1.96e-09 ***
## sgot                 4.750e-01  8.597e-01   0.553 0.580857    
## tryglicerides        8.949e-01  7.273e-01   1.231 0.219219    
## platelets            5.360e-01  4.749e-01   1.129 0.259696    
## prothrombin          1.692e+02  4.651e+01   3.638 0.000312 ***
## stage               -2.014e+02  5.543e+01  -3.634 0.000316 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 841.7 on 391 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.4521, Adjusted R-squared:  0.4227 
## F-statistic: 15.36 on 21 and 391 DF,  p-value: < 2.2e-16
```

```
# Code for diagnostics
par(mfrow = c(2, 3))
plot(lmFull, which=1:6)
```



## 2. Variable selection: Perform variable selection to select a suitable model involving a subset of your explanatory variables. You can use either stepwise methods or regression subsets in conjunction with cross validation. (10 points). 1. Variable selection with categorical variables

```
bCholang = regsubsets(n_days ~ ., cholang)
summary(bCholang)$out
```

```
##          statusCL statusD drugNotParticipated drugPlacebo age sexM ascitesY
## 1  ( 1 ) " "      " "     " "                 " "         " " " "  " "     
## 2  ( 1 ) " "      "*"     " "                 " "         " " " "  " "     
## 3  ( 1 ) " "      "*"     " "                 " "         " " " "  " "     
## 4  ( 1 ) " "      "*"     " "                 " "         " " " "  " "     
## 5  ( 1 ) " "      "*"     " "                 " "         " " " "  " "     
## 6  ( 1 ) " "      "*"     " "                 " "         " " " "  " "     
## 7  ( 1 ) " "      "*"     "*"                 " "         " " " "  " "     
## 8  ( 1 ) " "      "*"     "*"                 " "         " " " "  " "     
##          hepatomegalyY spidersY edemaS edemaY bilirubin cholesterol albumin
## 1  ( 1 ) " "           " "      " "    " "    " "       " "         "*"    
## 2  ( 1 ) " "           " "      " "    " "    " "       " "         "*"    
## 3  ( 1 ) " "           " "      " "    " "    " "       " "         "*"    
## 4  ( 1 ) " "           " "      " "    " "    "*"       " "         "*"    
## 5  ( 1 ) " "           " "      " "    " "    "*"       " "         "*"    
## 6  ( 1 ) " "           " "      " "    " "    "*"       " "         "*"    
## 7  ( 1 ) " "           " "      " "    " "    "*"       " "         "*"    
## 8  ( 1 ) " "           " "      " "    " "    "*"       " "         "*"    
##          copper alk_phos sgot tryglicerides platelets prothrombin stage
## 1  ( 1 ) " "    " "      " "  " "           " "       " "         " "  
## 2  ( 1 ) " "    " "      " "  " "           " "       " "         " "
## 3  ( 1 ) " "    "*"      " "  " "           " "       " "         " "
## 4  ( 1 ) " "    "*"      " "  " "           " "       " "         " "
## 5  ( 1 ) " "    "*"      " "  " "           " "       " "         "*"
## 6  ( 1 ) " "    "*"      " "  " "           " "       "*"         "*"
## 7  ( 1 ) " "    "*"      " "  " "           " "       "*"         "*"
## 8  ( 1 ) "*"    "*"      " "  " "           " "       "*"         "*"
```

2. Variable selection without categorical variables

```
bCholang2 = regsubsets(n_days ~ ., cholang_num)
summary(bCholang2)$out
```

```
##          age bilirubin cholesterol albumin copper alk_phos sgot tryglicerides
## 1  ( 1 ) " " " "       " "         "*"     " "    " "      " "  " "
## 2  ( 1 ) " " "*"       " "         "*"     " "    " "      " "  " "
## 3  ( 1 ) " " "*"       " "         "*"     " "    " "      " "  " "
## 4  ( 1 ) " " "*"       " "         "*"     " "    "*"      " "  " "
## 5  ( 1 ) " " "*"       " "         "*"     "*"    "*"      " "  " "
## 6  ( 1 ) " " "*"       " "         "*"     "*"    "*"      " "  " "
## 7  ( 1 ) " " "*"       "*"         "*"     "*"    "*"      " "  " "
## 8  ( 1 ) " " "*"       "*"         "*"     "*"    "*"      " "  " "
##          platelets prothrombin stage
## 1  ( 1 ) " "       " "         " "
## 2  ( 1 ) " "       " "         " "
## 3  ( 1 ) " "       " "         "*"
## 4  ( 1 ) " "       " "         "*"
## 5  ( 1 ) " "       " "         "*"
## 6  ( 1 ) " "       "*"         "*"
## 7  ( 1 ) " "       "*"         "*"
## 8  ( 1 ) "*"       "*"         "*"
```

```
set.seed(78912)
permutation<-sample(1:nrow(cholang_num))
folds <- cut(1:nrow(cholang_num),breaks=10,labels=FALSE)
predErrorMat<-matrix(nrow=10,ncol=nrow(summary(bCholang2)$which))

for(i in 1:10){
    #Segement your data by fold using the which() function
    testIndexes <- which(folds==i,arr.ind=TRUE)
    testData <- cholang_num[permutation,][testIndexes, ]
    trainData <- cholang_num[permutation,][-testIndexes, ]
    #Use the test and train data partitions however you desire...
    predError<-apply(summary(bCholang2)$which[,-1],1,function(x){
        lmObj<-lm(trainData$n_days ~ .,data=trainData[,-1][,x,drop=FALSE])
        testPred<-predict(lmObj,newdata=testData[,-1])
        mean((testData$n_days-testPred)^2)
    })
    predErrorMat[i,]<-predError
}
colMeans(predErrorMat)
```

```
## [1] 996713.3 900722.7 846634.2 803706.2 786712.3 790430.4 791306.2 791770.2
```

```
LOOCV<-function(lm){
    vals<-residuals(lm)/(1-lm.influence(lm)$hat)
    sum(vals^2)/length(vals)
}
calculateCriterion<-function(x=NULL,y,dataset,lmObj=NULL){
    #dataset contains only explanatory variables
    #x is a vector of logicals, length equal to number of explanatory variables in dataset, telling us
    #sigma2 is estimate of model on full dataset
    # either x or lmObj must be given to specify the smaller lm model
    sigma2=summary(lm(y~.,data=dataset))$sigma^2
    if(is.null(lmObj)) lmObj<-lm(y ~ ., data=dataset[,x,drop=FALSE]) #don't include intercept
    sumlmObj<-summary(lmObj)
    n<-nrow(dataset)
    p<-sum(x)
    RSS<-sumlmObj$sigma^2*(n-p-1)
    c(R2=sumlmObj$r.squared,
        R2adj=sumlmObj$adj.r.squared,
        "RSS/n"=RSS/n,
        LOOCV=LOOCV(lmObj),
        Cp=RSS/n+2*sigma2*(p+1)/n,
        CpAlt=RSS/sigma2-n+2*(p+1),
        AIC=AIC(lmObj), # n*log(RSS/n)+2*p +constant,
        BIC=BIC(lmObj) # n*log(RSS/n)+p*log(n) + constant
    )
}

critCholang<-apply(summary(bCholang2)$which[,-1],1,calculateCriterion,
    y=cholang$n_days,
    dataset=cholang_num[,-1])

critCholang<-t(critCholang)
critCholang
```

```
##          R2     R2adj    RSS/n     LOOCV       Cp      CpAlt      AIC      BIC
## 1 0.1856133 0.1836556 991420.6 1000386.1 998818.2 122.074926 6963.514 6975.621
## 2 0.2656586 0.2621196 893974.8  904839.4 905071.2  71.384616 6922.268 6938.409
## 3 0.3106336 0.3056382 839222.9  853629.2 854018.2  43.779531 6897.849 6918.027
## 4 0.3526484 0.3463787 788074.8  808595.5 806568.9  18.123036 6873.564 6897.777
## 5 0.3680608 0.3603917 769312.0  792271.3 791504.9   9.977729 6865.492 6893.740
## 6 0.3744837 0.3653521 761493.0  792619.4 787384.6   7.749841 6863.222 6895.506
## 7 0.3766822 0.3660402 758816.6  793121.0 788407.1   8.302677 6863.750 6900.069
## 8 0.3796461 0.3675120 755208.3  793400.9 788497.6   8.351651 6863.758 6904.112
```

```
data.frame(
  AIC = which.min(abs(critCholang[,"AIC"])),
  LOOCV = which.min(abs(critCholang[,"LOOCV"]))
)
```

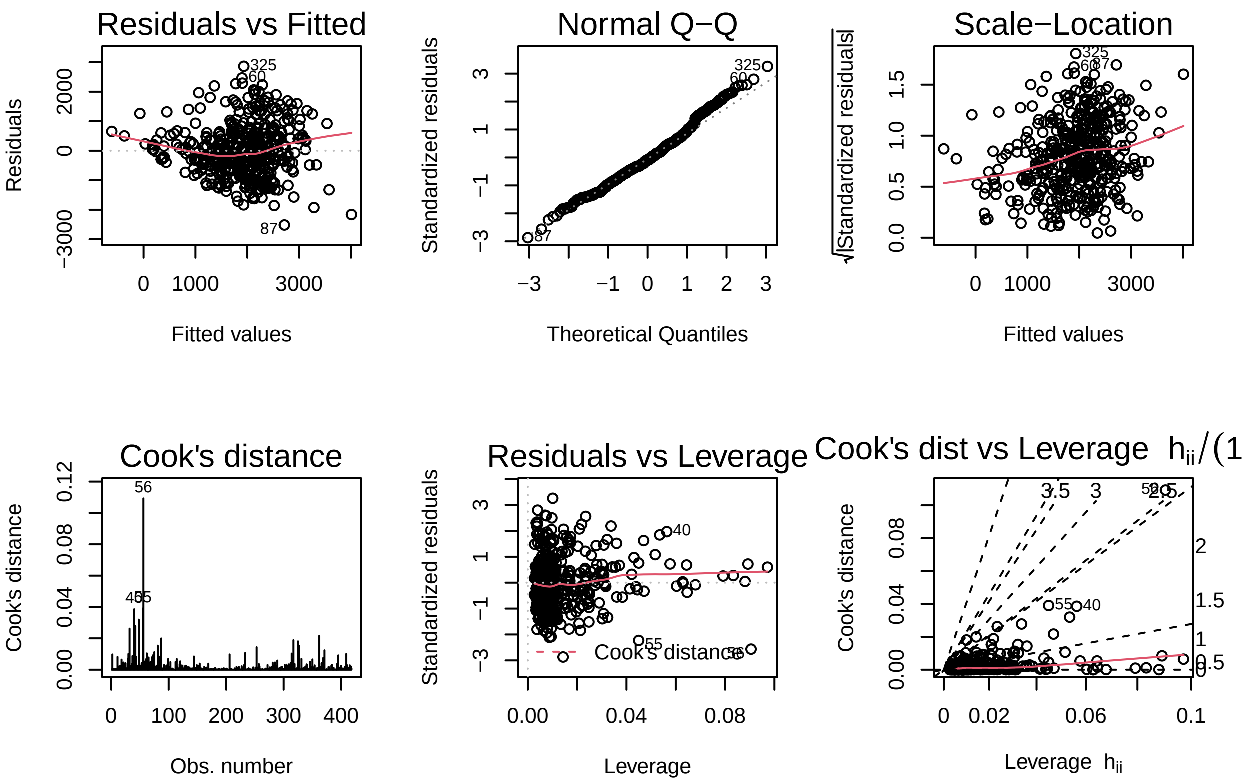
```
##   AIC LOOCV
## 6   6     5
```

Based on AIC, model (6) is the best model with the lowest AIC. However, based on LOOCV, model (5) is the best model with the lowest LOOCV. Jointly consider the prediction error of cross-validation, I decide to choose the 5th model, which is n_days ~bilirubin+albumin+copper+alk_phos+stage

3. Regression diagnostics:

Look at diagnostic plots of this final model and comment on whether any of the regression assumptions are obviously violated for this dataset and the final model. (10 points)

```
# Code for diagnostics
par(mfrow = c(2, 3))
plot(lm(n_days ~bilirubin+albumin+copper+alk_phos+stage,data = cholang), which=1:6)
```



There are still some problems with this model. We can see there's a non-linear (quadratic) relation in the Residuals vs Fitted plot, and increasing pattern the Scale-Location plot, suggesting that the distribution of residuals is heteroscedastic. From the QQ plot, we can see that residuals are normally distributed, suggesting the distribution is normal and the model is valie.

From the Cook's distance plot and Residuals vs Leverage plot, we can see that there are some outliers such as 56, 40 and 55.

The next steps can be: (1)remove outliers; (2) change the model into a non-linear model, such as a quadratic model.

Logistic Regression

Fit a logistic regression model for the survival status of a patient at the end of the study, given all the explanatory variables (remember, you are considering status as binary, ignoring the patients who receive transplants). You may also perform variable selection. Comment on your model, with visualizations, as in the , text. (15 points)

```
cholang2 = cholang[-which(cholang$status == "CL")]
set.seed(123)
nTest<-.1*nrow(cholang2)
```

```r
whTest<-sample(1:nrow(cholang2),size=nTest)
test<-cholang2[whTest,]
train<-cholang2[-whTest,]
glm <- glm(status ~.,family=binomial(link='logit'),data=train)
summary(glm)
```

```
##
## Call:
## glm(formula = status ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4463  -0.7726  -0.3964   0.7055   2.4779
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -9.727e+00  2.474e+00  -3.932 8.42e-05 ***
## n_days                 -7.631e-04  1.607e-04  -4.748 2.06e-06 ***
## drugNotParticipated -7.882e-01  3.500e-01  -2.252 0.024315 *
## drugPlacebo            -2.350e-01  3.211e-01  -0.732 0.464197
## age                     6.534e-05  3.686e-05   1.773 0.076264 .
## ascitesY                6.330e-01  8.624e-01   0.734 0.462943
## hepatomegalyY           2.466e-01  2.863e-01   0.861 0.389050
## spidersY                1.160e-01  3.173e-01   0.366 0.714682
## edemaS                  3.757e-01  4.442e-01   0.846 0.397697
## edemaY                  7.874e-01  1.396e+00   0.564 0.572777
## bilirubin               1.749e-01  5.877e-02   2.976 0.002921 **
## cholesterol             2.696e-04  8.014e-04   0.336 0.736524
## albumin                 3.388e-01  3.738e-01   0.906 0.364761
## copper                  4.854e-04  1.944e-03   0.250 0.802853
## alk_phos                1.721e-04  6.654e-05   2.586 0.009705 **
## sgot                    2.733e-03  2.727e-03   1.002 0.316197
## tryglicerides           3.160e-03  2.581e-03   1.224 0.220774
## platelets               1.238e-03  1.535e-03   0.806 0.420005
## prothrombin             5.315e-01  1.504e-01   3.533 0.000411 ***
## stage                   3.077e-01  1.768e-01   1.740 0.081843 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 511.81  on 371  degrees of freedom
## Residual deviance: 351.05  on 352  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 391.05
##
## Number of Fisher Scoring iterations: 6
```

```r
anova(glm, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
```

```
## 
## Response: status
## 
## Terms added sequentially (first to last)
## 
## 
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
## NULL                            371     511.81              
## n_days         1   73.411       370     438.40 < 2.2e-16 ***
## drug           2    5.406       368     433.00 0.0670021 .  
## age            1    4.355       367     428.64 0.0369027 *  
## ascites        1    5.906       366     422.73 0.0150889 *  
## hepatomegaly   1    6.649       365     416.09 0.0099212 ** 
## spiders        1    1.330       364     414.76 0.2488732    
## edema          2    4.961       362     409.79 0.0836882 .  
## bilirubin      1   30.283       361     379.51 3.734e-08 ***
## cholesterol    1    0.482       360     379.03 0.4876636    
## albumin        1    0.081       359     378.95 0.7762586    
## copper         1    2.182       358     376.77 0.1395984    
## alk_phos       1    8.412       357     368.35 0.0037265 ** 
## sgot           1    0.313       356     368.04 0.5758295    
## tryglicerides  1    0.500       355     367.54 0.4792870    
## platelets      1    0.055       354     367.49 0.8143559    
## prothrombin    1   13.380       353     354.11 0.0002544 ***
## stage          1    3.061       352     351.05 0.0801991 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that n_days, bilirubin, alk_phos, prothombin is sigificant variable and are useful for decreasing the deviation.

```r
fitted.results <- predict(glm,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
new.status <- ifelse(test$status == "D",1,0)
misClasificError <- mean(fitted.results != new.status)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.780487804878049"
```

The prediction accuracy is 78%, a pretty good result for survival prediction. Therefore, the model is credible. In fact, I've tried to do the stepwise selection here, but found out the prediction accuracy becomes worse once we do variable selection. Since the original dataset is not big, I decide to leave it with the original model.