Final Project

STAT 131A, Fall 2021

Due: Wednesday, December 15, 2021 at 11:59 PM

1 Dataset

The dataset (*cholangitis.csv*) comes from a randomized, double-blinded, placebo-controlled clinical trial of the immunosuppressive drug D-penicillamine at the Mayo Clinic. The study consisted of patients living with primary biliary cholangitis, a fatal chronic autoimmune disease of unknown cause affecting the liver. The details of the dataset are provided in the document *cholangitis.pdf*. There are 418 observations of 20 variables, both numeric and categorical. The study lasted about 12 years. The goals of the project are to:

1. Fit a linear regression equation to the **number of days** a patient survives from the time of registration.
2. Fit a logistic regression to model the **status** of a patient at the end of the study (you only need to consider alive vs dead, and may ignore the 25 patients who received a liver transplant).

2 Visualization

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

3 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**).
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**).
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**).

4 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**)

5 Format for Submission

You are expected to create a *Rmd* file for this project from scratch. The text from this instructions pdf should not be part of your *Rmd* file. You will turn in only a compiled *pdf* to gradescope. An actual analysis would blend these components together into a single narrative. However, for grading purposes, we have divided the project into the specific tasks described above, and each of the tasks should be addressed in a separate section and appropriately labeled so that you can tell Gradescope which pages correspond to which task.

This project is intentionally more open-ended than the homework, so as to be more reflective of an actual analysis of the data. For each of the specific tasks, provide commentary on what you are doing, provide R code and output, and also provide commentary on what you deduce from the output. The commentary should be just regular text typed in your Rmd file (it does not need a > in front of it like the homework). DO NOT put any commentary in the comments of your R code.

To make your code in the compiled pdf wrap nicely, add the following into a code chunk at the beginning of your Rmd file:

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

Final Project

Xinxiaomeng Liu

Visualization

1. Importing the data:

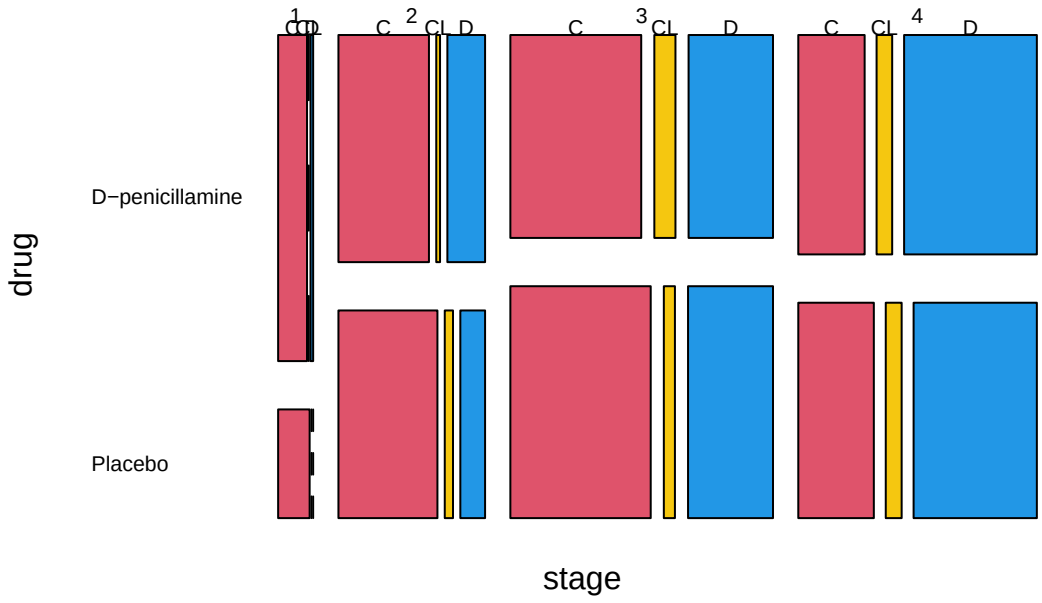
```
cholangitis<-read.csv("cholangitis.csv",header=TRUE)

cholangitis$status = factor(cholangitis$status)
cholangitis$drug = factor(cholangitis$drug)
cholangitis$sex = factor(cholangitis$sex)
cholangitis$ascites = factor(cholangitis$ascites)
cholangitis$hepatomegaly = factor(cholangitis$hepatomegaly)
cholangitis$spiders = factor(cholangitis$spiders)
cholangitis$edema = factor(cholangitis$edema)
cholangitis$stage = factor(cholangitis$stage)
cholangitis <- cholangitis[,-1]
```

> I also excluded the column named "id". Since this trail is randomized, it doesn't make sense to expect "id" to be an explanatory variable.

2. Basic exploratory data analysis:

> First, we would like to know whether D-penicillamine has any effect on patience.

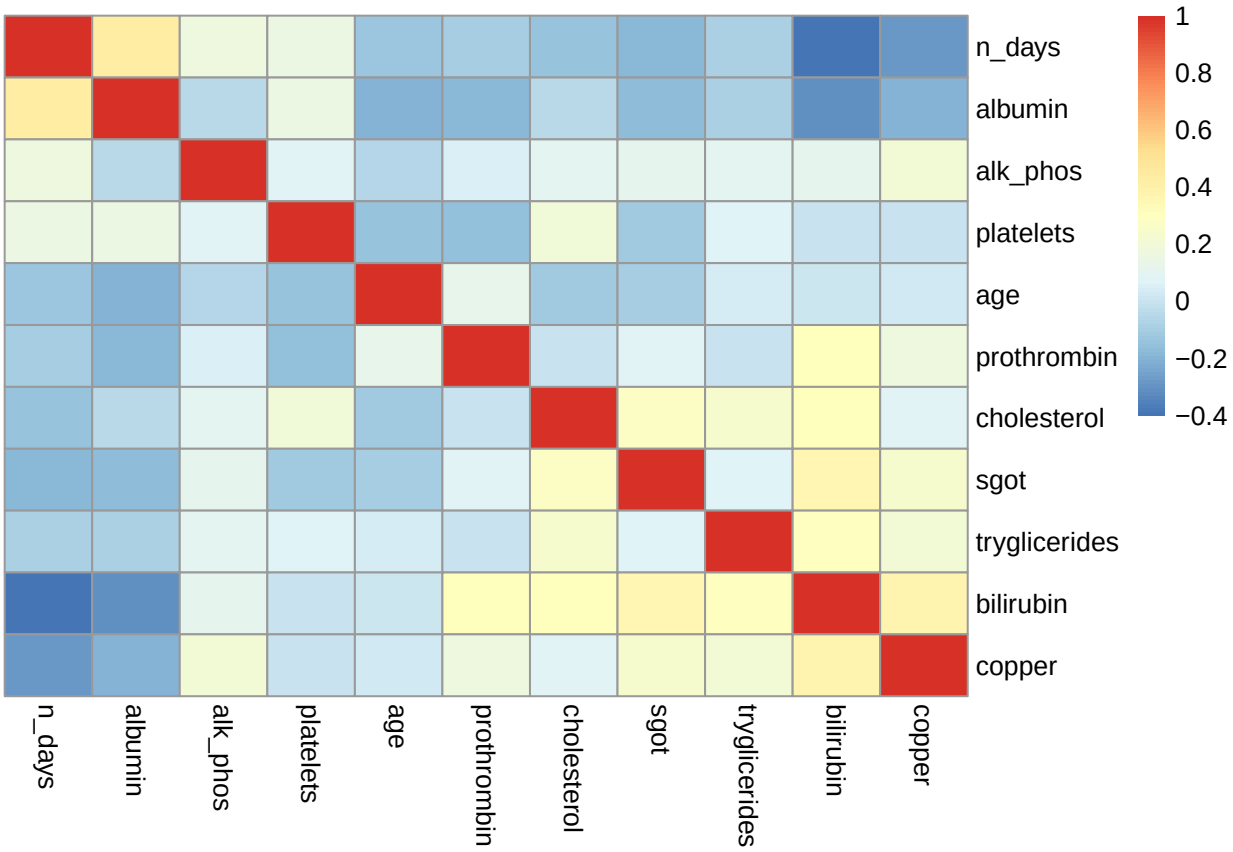
```
mosaicplot(~ stage + drug + status, data = cholangitis[1:312,], las = 1, col = c("2","7","4"))
```



We use a mosaic plot to visualize the relationships between "stage", "drug", and "status". We use the data from the first enrolled 312 people, because the later enrolled ones did not take D-penicillamine or placebo. From this plot, we can see that for people in each stage of disease, the proportion of people who survived in the "D-penicillamine" group is almost the same as the proportion of people who survived in the "Placebo" group. Thus, there's no evidence that D-penicillamine has effect on people in any stage of disease.

Second, We use a heatmap plot to visualize the relationships between all the continuous variables (except for "id").

```
library(pheatmap)
corMat <- cor(na.omit(cholangitis[, -c(2,3,5:9,19)]))
pheatmap(corMat, cluster_rows = TRUE, cluster_cols = TRUE, treeheight_row = 0, treeheight_col = 0)
```



From this plot, we can see that "n_days" is negatively related with "bilirubin" and "copper". A possible explanation is that since high bilirubin and high urine copper implies liver disease, people with high bilirubin or high copper might passed away early in this trail. We can also see that "bilirubin", "copper", "tryglicerides" and "sgot" are almost all positively related. It is probably because that people would have high bilirubin, high urine copper, high triglycerides and high sgot once they have liver disease.

Multivariate Regression

1. Multivariate regression analysis:

We first exclude those data with missing values. Then we perform a regression analysis:

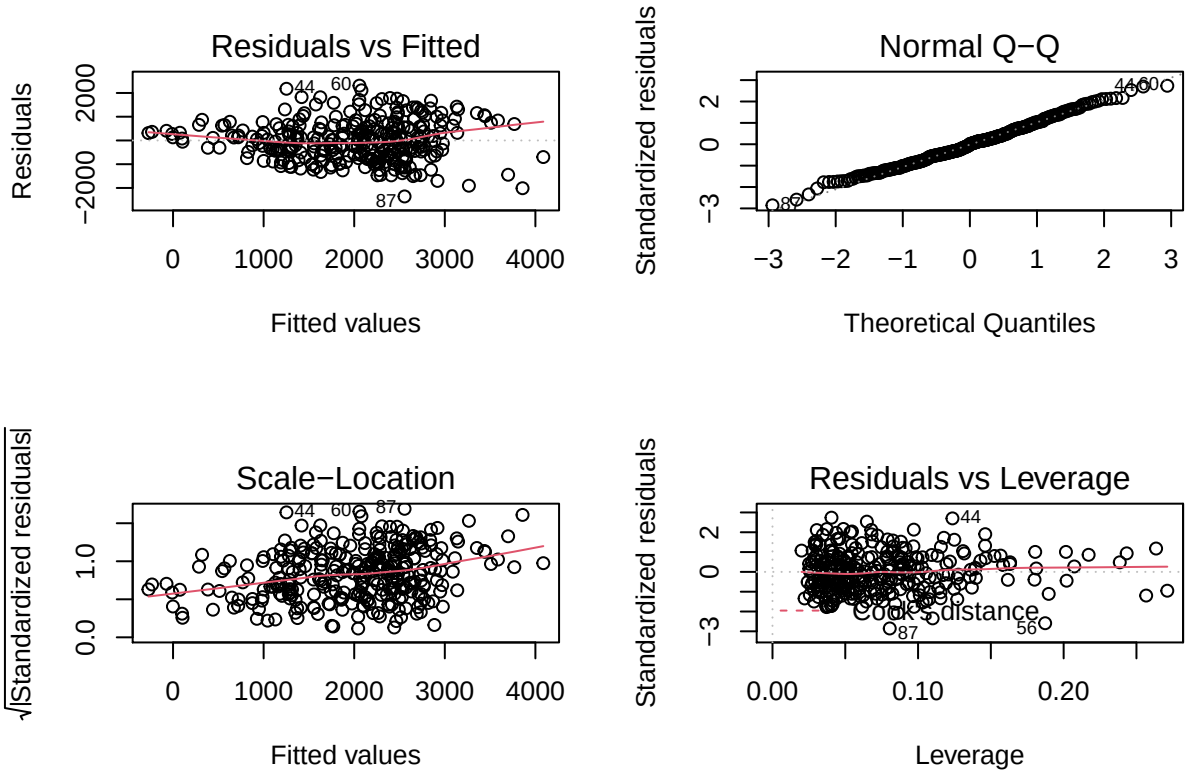
```
fit <- lm(n_days ~., data = na.omit(cholangitis))
summary(fit)
```

```
##
## Call:
## lm(formula = n_days ~ ., data = na.omit(cholangitis))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2357.74  -611.01    11.27   566.13  2307.18
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.024e+03  9.495e+02  -1.078 0.281765
## statusCL      -6.687e+02  2.213e+02  -3.022 0.002742 **
## statusD       -5.862e+02  1.336e+02  -4.387 1.62e-05 ***
## drugPlacebo    3.302e+00  1.027e+02   0.032 0.974381
## age           -7.459e-03  1.528e-02  -0.488 0.625805
## sexM           9.031e+01  1.728e+02   0.523 0.601573
## ascitesY      -4.968e+01  2.639e+02  -0.188 0.850831
## hepatomegalyY -2.614e+01  1.205e+02  -0.217 0.828365
## spidersY      -9.687e+01  1.265e+02  -0.766 0.444456
## edemaS        -1.383e+02  1.860e+02  -0.744 0.457716
## edemaY        -3.488e+02  2.827e+02  -1.234 0.218255
## bilirubin     -5.304e+01  1.679e+01  -3.159 0.001757 **
## cholesterol   -1.943e-01  2.649e-01  -0.733 0.463964
## albumin        5.623e+02  1.454e+02   3.868 0.000136 ***
## copper        -1.756e+00  7.245e-01  -2.424 0.015964 *
## alk_phos       1.357e-01  2.460e-02   5.515 7.85e-08 ***
## sgot           7.147e-01  1.053e+00   0.679 0.497668
## trygliceride  7.907e-01  8.977e-01   0.881 0.379190
## platelets      4.470e-01  5.851e-01   0.764 0.445468
## prothrombin    1.615e+02  6.109e+01   2.643 0.008671 **
## stage2        -2.237e+02  2.516e+02  -0.889 0.374521
## stage3        -3.429e+02  2.463e+02  -1.392 0.164966
## stage4        -5.233e+02  2.650e+02  -1.975 0.049278 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 861.2 on 284 degrees of freedom
## Multiple R-squared:  0.4593, Adjusted R-squared:  0.4174
## F-statistic: 10.97 on 22 and 284 DF,  p-value: < 2.2e-16
```

Then we do regression diagnostics.

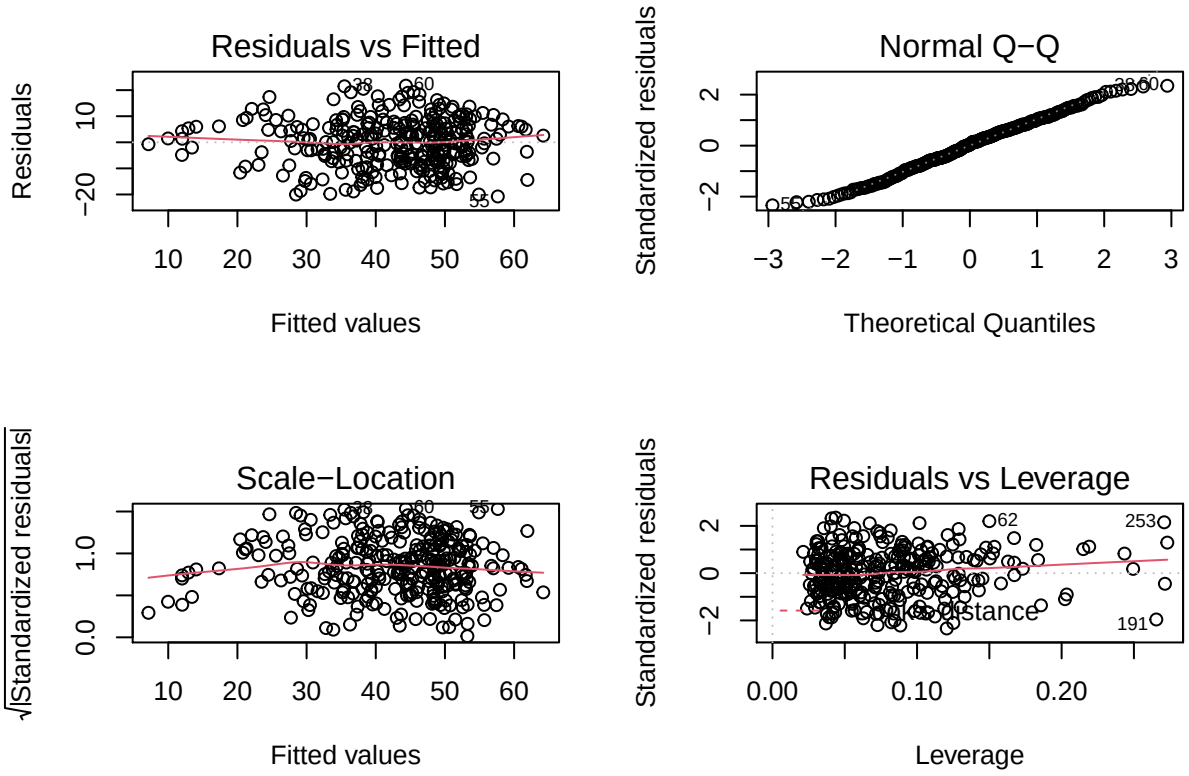
```
par(mfrow = c(2,2))
plot(fit)
```



The Residual vs Fitted plot indicates non-linearity. We can add age^2 as an additional explanatory variable to fix this. The Residual vs Fitted plot doesn't indicate heteroscedasticity. However, the incresing pattern in the Scale-Location plot indicates heteroscedasticity. To solve this problem, we can try transforming the reponse by the log or squre-root. (After trying them both, I found that square-root is a better choice for this case.) The Q-Q plot indicates that the data is normal. The Residual vs Leverage plot indicates that there are outliers. Three points flagged here are observations i = 44, 87, 56. So we can perform the regression analysis after dropping these outliers.

We perform the regression diagnostics again after adding age^2 as an additional explanatory, transforming the reponse by the squre-root and removing the outliers:

```
par(mfrow = c(2,2))
plot(lm(sqrt(n_days) ~ I(age^2) +., data = na.omit(cholangitis[-c(44,56,87),])))
```



Now the outcome is satisfactory. Then we perform the regression analysis again:

```
cholangitis_modified <- na.omit(cholangitis[-c(44,56,87),])
fit1 <- lm(sqrt(n_days) ~ I(age^2) +., data = cholangitis_modified)
summary(fit1)
```

```
##
## Call:
## lm(formula = sqrt(n_days) ~ I(age^2) + ., data = cholangitis_modified)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.7649  -6.3194   0.4623   6.2787  21.7147
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.955e+00  1.487e+01  -0.400 0.689141
## I(age^2)      -5.534e-08  3.202e-08  -1.728 0.085031 .
## statusCL      -7.774e+00  2.432e+00  -3.197 0.001549 **
## statusD       -6.653e+00  1.492e+00  -4.459 1.19e-05 ***
## drugPlacebo   -6.116e-01  1.137e+00  -0.538 0.591164
## age            1.919e-03  1.194e-03   1.607 0.109210
## sexM           1.417e+00  1.917e+00   0.739 0.460252
## ascitesY      -2.021e-01  2.944e+00  -0.069 0.945332
## hepatomegalyY -1.270e-01  1.331e+00  -0.095 0.924073
## spidersY      -9.795e-01  1.401e+00  -0.699 0.484937
## edemaS        -2.284e+00  2.062e+00  -1.108 0.268811
## edemaY        -8.897e+00  3.245e+00  -2.742 0.006500 **
```

```
## bilirubin     -7.281e-01  1.862e-01  -3.911 0.000116 ***
## cholesterol   -1.057e-03  2.922e-03  -0.362 0.717919
## albumin        6.980e+00  1.613e+00   4.328 2.10e-05 ***
## copper        -2.414e-02  7.988e-03  -3.022 0.002739 **
## alk_phos       1.592e-03  2.846e-04   5.592 5.33e-08 ***
## sgot           1.069e-02  1.167e-02   0.916 0.360456
## tryglicerides  1.345e-02  1.009e-02   1.332 0.183866
## platelets      9.077e-03  6.494e-03   1.398 0.163304
## prothrombin    1.200e+00  6.715e-01   1.788 0.074924 .
## stage2        -2.117e+00  2.766e+00  -0.765 0.444740
## stage3        -3.271e+00  2.713e+00  -1.205 0.229035
## stage4        -5.270e+00  2.921e+00  -1.804 0.072283 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.462 on 280 degrees of freedom
## Multiple R-squared:  0.5565, Adjusted R-squared:  0.5201
## F-statistic: 15.27 on 23 and 280 DF,  p-value: < 2.2e-16
```

7 of 24 coefficients are significant. The Adjusted R-squared is 0.5201, which is higher than that before modifying, but is still not high enough. Thus, the regression equation doesn't fit the data well enough. The p-value of the F-statistic is quite small, which is a highly significant result. But most tests of general fit are highly significant. So we can conclude that it's not quite a good model. Since we have found in "Basic exploratory data analysis" part that some variables are correlated (e.g. "bilirubin", "copper", "tryglicerides" and "sgot"), the poor performance is probably caused by multicollinearity. Therefore, we need to do variable selection.

2. Variable selection

We use the `step` function to do variable selection.

```
reg <- step(fit1, trace = 0, direction = "both")
summary(reg)
```

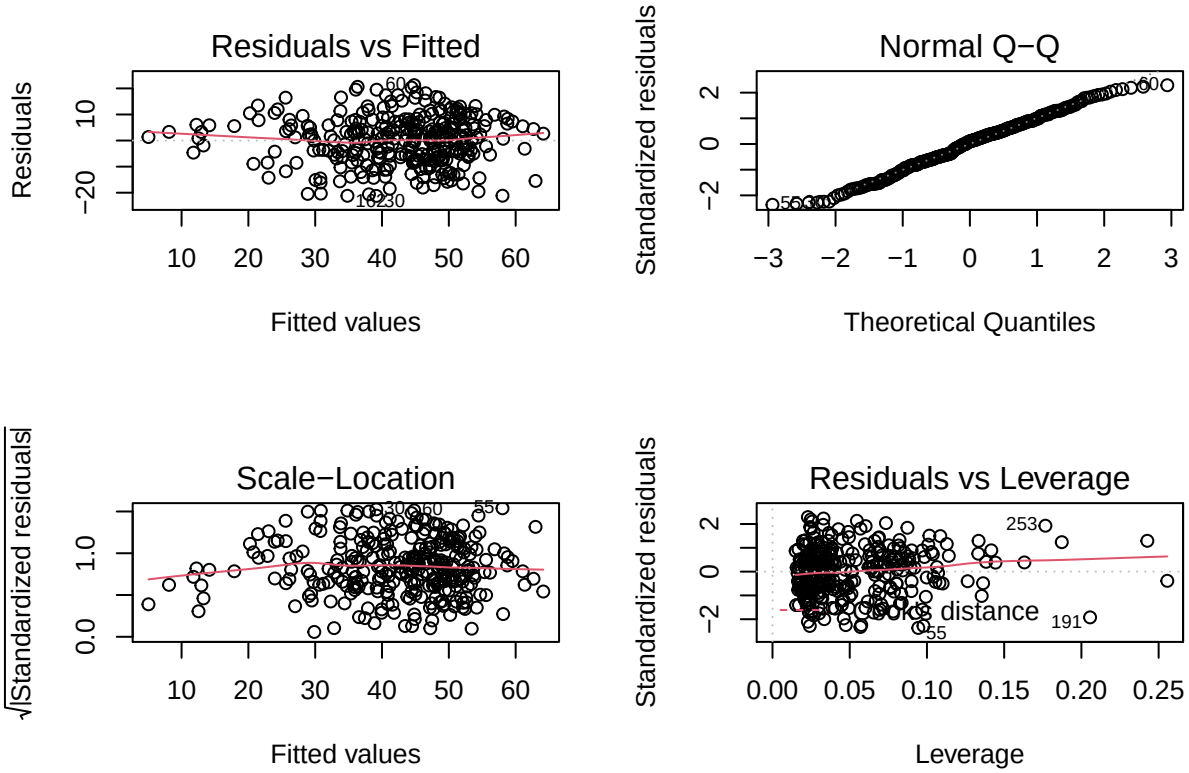
```
##
## Call:
## lm(formula = sqrt(n_days) ~ I(age^2) + status + age + edema +
##     bilirubin + albumin + copper + alk_phos + tryglicerides +
##     prothrombin + stage, data = cholangitis_modified)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2328  -6.2991   0.9162   6.3344  21.2543
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.679e+00  1.413e+01  -0.119 0.905501
## I(age^2)      -5.049e-08  3.120e-08  -1.618 0.106732
## statusCL      -7.168e+00  2.378e+00  -3.014 0.002805 **
## statusD       -6.485e+00  1.435e+00  -4.519 9.09e-06 ***
## age            1.771e-03  1.168e-03   1.516 0.130616
## edemaS        -2.567e+00  1.997e+00  -1.285 0.199699
```

```
## edemaY        -9.371e+00  2.703e+00  -3.467 0.000607 ***
## bilirubin     -7.355e-01  1.611e-01  -4.564 7.43e-06 ***
## albumin        7.278e+00  1.546e+00   4.707 3.92e-06 ***
## copper        -2.230e-02  7.491e-03  -2.977 0.003163 **
## alk_phos       1.689e-03  2.762e-04   6.113 3.17e-09 ***
## tryglicerides  1.414e-02  9.737e-03   1.453 0.147415
## prothrombin    1.056e+00  6.525e-01   1.619 0.106487
## stage2        -2.067e+00  2.684e+00  -0.770 0.441784
## stage3        -3.748e+00  2.586e+00  -1.449 0.148417
## stage4        -6.068e+00  2.691e+00  -2.255 0.024876 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.396 on 288 degrees of freedom
## Multiple R-squared:  0.5501, Adjusted R-squared:  0.5267
## F-statistic: 23.48 on 15 and 288 DF,  p-value: < 2.2e-16
```

A suitable submodel is sqrt(n_days) ~ I(age^2) + status + age + edema + bilirubin + albumin + copper + alk_phos + tryglicerides + prothrombin + stage

3. Regression diagnostics

```
par(mfrow = c(2,2))
fit2 <- lm(sqrt(n_days) ~ I(age^2) + status + age + edema +
    bilirubin + albumin + copper + alk_phos + tryglicerides +
    prothrombin + stage, data = cholangitis_modified)
plot(fit2)
```



The Residual vs Fitted plot doesn't indicate non-linearity or heteroscedasticity. Also, the incresing pattern in the Scale-Location plot doesn't indicate heteroscedasticity. The Q-Q plot indicates that the data is normal The Residual vs Leverage plot indicates that there are outliers like i=253,191. However, they are within a resonable range. Thus, the outcome of the regrassion diagnostic is satisfactory.

Logistic Regression

```
cholangitis_binary <- droplevels(subset(cholangitis_modified, cholangitis_modified$status != "CL"))
fit_logistic <- glm(status ~., family = binomial, data = cholangitis_binary)
summary(fit_logistic)
```

```
##
## Call:
## glm(formula = status ~ ., family = binomial, data = cholangitis_binary)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3505  -0.5850  -0.2582   0.4729   2.4290
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.744e+01  4.173e+00  -4.180 2.92e-05 ***
## n_days        -7.777e-04  2.172e-04  -3.581 0.000342 ***
## drugPlacebo   -3.561e-01  3.605e-01  -0.988 0.323182
## age            1.166e-04  5.113e-05   2.280 0.022637 *
## sexM           9.940e-01  6.003e-01   1.656 0.097726 .
## ascitesY       2.064e+00  1.324e+00   1.559 0.119046
## hepatomegalyY  3.621e-01  4.024e-01   0.900 0.368262
## spidersY       2.157e-01  4.281e-01   0.504 0.614424
## edemaS        -2.292e-01  7.005e-01  -0.327 0.743582
## edemaY        -1.027e+00  1.421e+00  -0.723 0.469986
## bilirubin      6.542e-02  8.033e-02   0.814 0.415415
## cholesterol    8.137e-04  1.048e-03   0.776 0.437499
## albumin        2.353e-01  5.318e-01   0.442 0.658202
## copper         2.076e-03  2.901e-03   0.716 0.474276
## alk_phos       3.133e-04  9.658e-05   3.244 0.001177 **
## sgot           5.961e-03  3.428e-03   1.739 0.082073 .
## tryglicerides  3.943e-03  3.584e-03   1.100 0.271243
## platelets     -3.113e-05  2.113e-03  -0.015 0.988248
## prothrombin    9.364e-01  2.335e-01   4.010 6.06e-05 ***
## stage2         2.835e+00  1.557e+00   1.821 0.068663 .
## stage3         3.063e+00  1.556e+00   1.969 0.048942 *
## stage4         2.942e+00  1.549e+00   1.900 0.057460 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 387.96  on 284  degrees of freedom
## Residual deviance: 217.24  on 263  degrees of freedom
## AIC: 261.24
```

```
##
## Number of Fisher Scoring iterations: 6
```

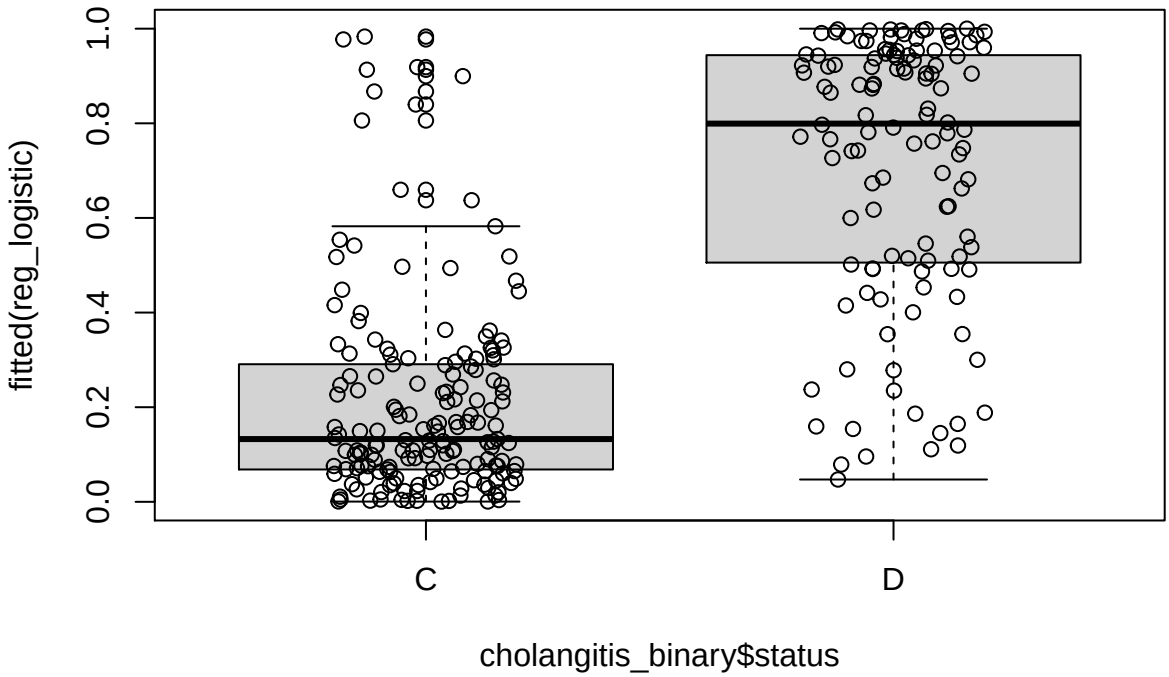
```
reg_logistic <- step(fit_logistic, trace = 0, direction = "both")
summary(reg_logistic)
```

```
##
## Call:
## glm(formula = status ~ n_days + age + sex + ascites + cholesterol +
##     alk_phos + sgot + tryglicerides + prothrombin + stage, family = binomial,
##     data = cholangitis_binary)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8604  -0.6038  -0.2690   0.4716   2.4712
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.717e+01  3.492e+00  -4.918 8.76e-07 ***
## n_days        -8.587e-04  1.866e-04  -4.601 4.21e-06 ***
## age            1.150e-04  4.851e-05   2.370 0.017810 *
## sexM           1.178e+00  5.424e-01   2.172 0.029845 *
## ascitesY       1.559e+00  1.104e+00   1.413 0.157713
## cholesterol    1.342e-03  9.685e-04   1.385 0.165957
## alk_phos       3.465e-04  9.034e-05   3.836 0.000125 ***
## sgot           6.610e-03  3.181e-03   2.078 0.037744 *
## tryglicerides  5.040e-03  3.265e-03   1.544 0.122659
## prothrombin    9.738e-01  2.172e-01   4.484 7.32e-06 ***
## stage2         2.951e+00  1.579e+00   1.869 0.061654 .
## stage3         3.282e+00  1.568e+00   2.092 0.036402 *
## stage4         3.337e+00  1.524e+00   2.189 0.028607 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 387.96  on 284  degrees of freedom
## Residual deviance: 221.77  on 272  degrees of freedom
## AIC: 247.77
##
## Number of Fisher Scoring iterations: 6
```

A good submodel is status ~ n_days + age + sex + ascites + cholesterol + alk_phos + sgot + tryglicerides + prothrombin + stage Most of the coefficients are significant.

The following plots the fitted values against the actual response:

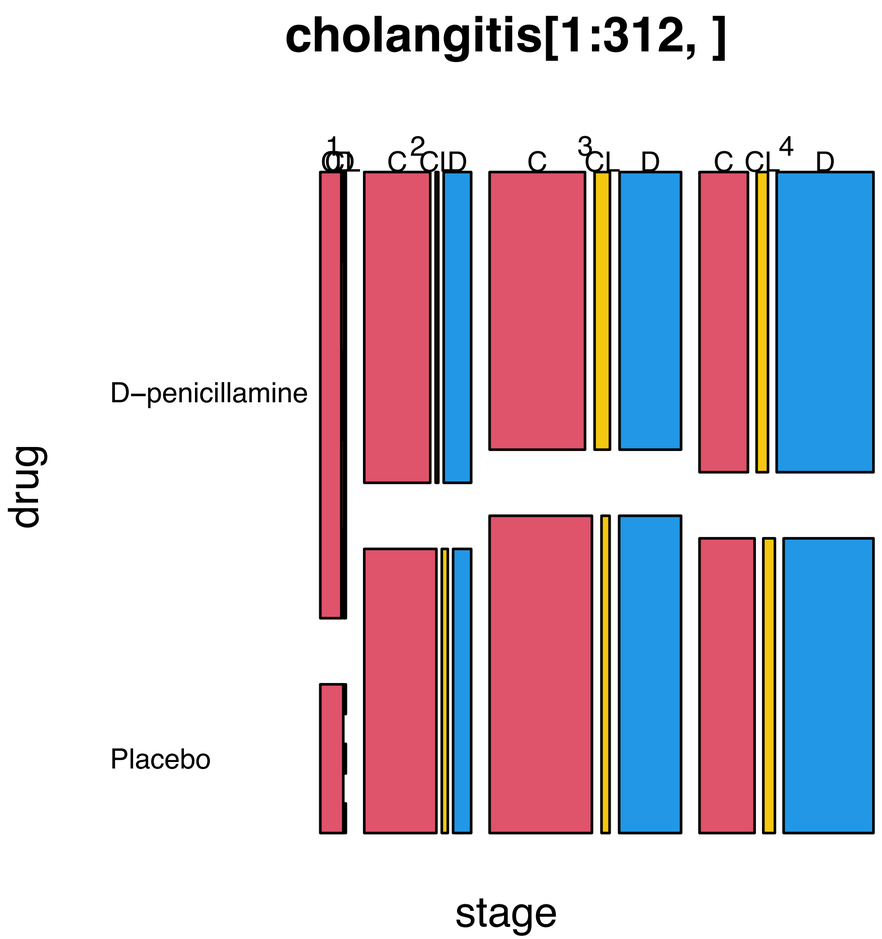
```
boxplot(fitted(reg_logistic) ~ cholangitis_binary$status, at = c(0, 1))
numeric_status <- as.numeric(cholangitis_binary$status)-1
points(x = jitter(numeric_status), fitted(reg_logistic))
```

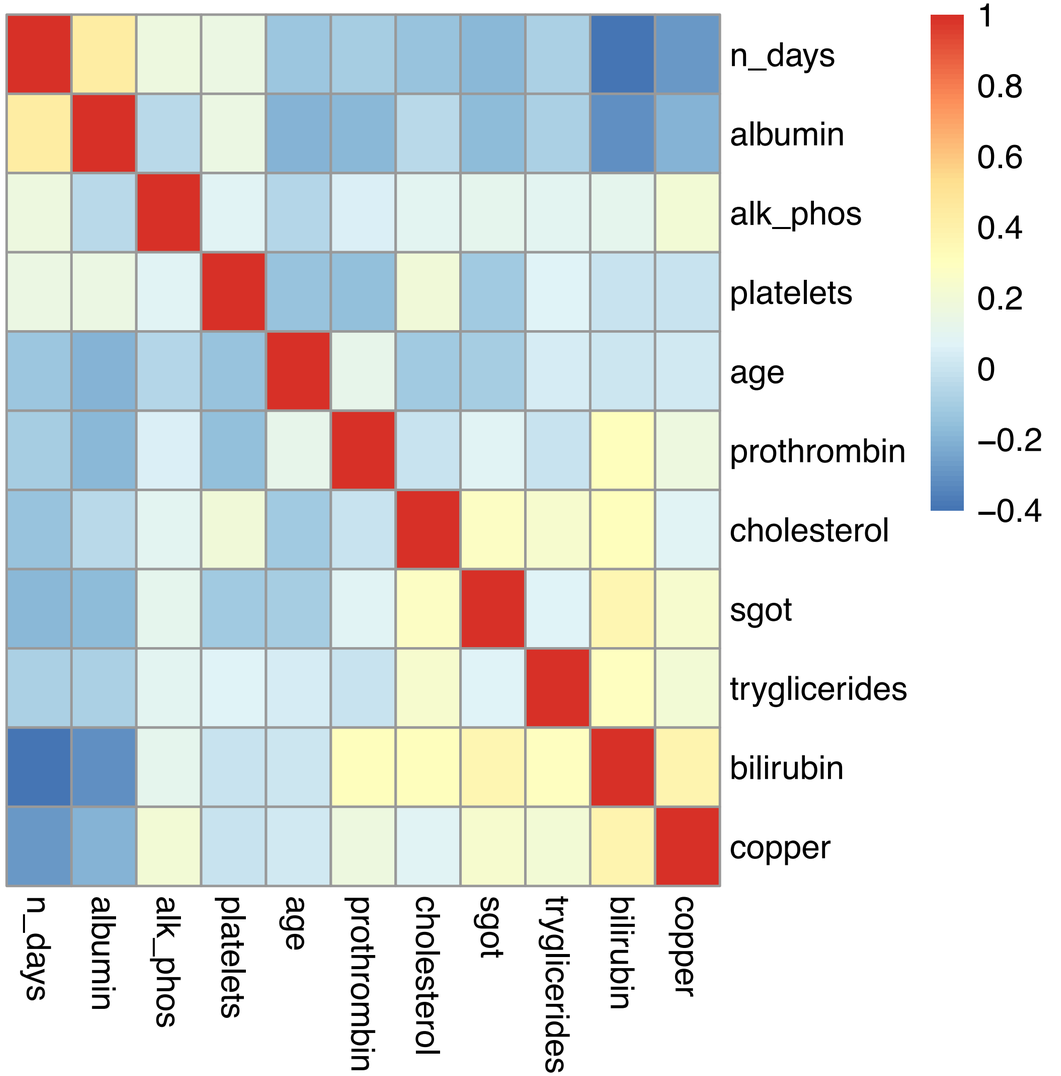


From the plot we can see that this model is overall good, since most points with "C" status have received low fitted probability under the model, and most points with "D" status have received high fitted probability under the model. But there are still some points with "C" status have received very high fitted probability, and some points with "D" status have received very low fitted probability.

Exploratory Data Analysis

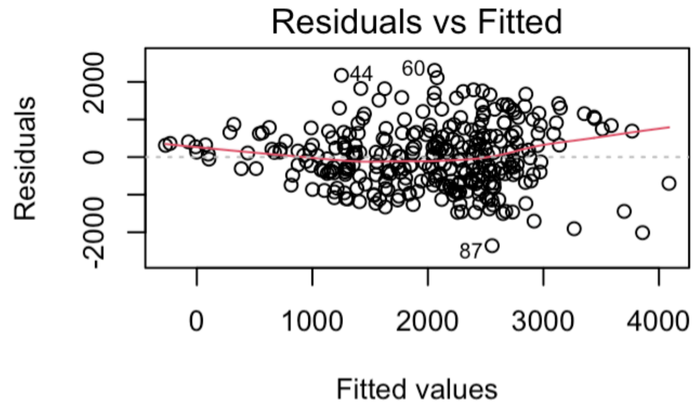
Mosaic Plot & Heat Map

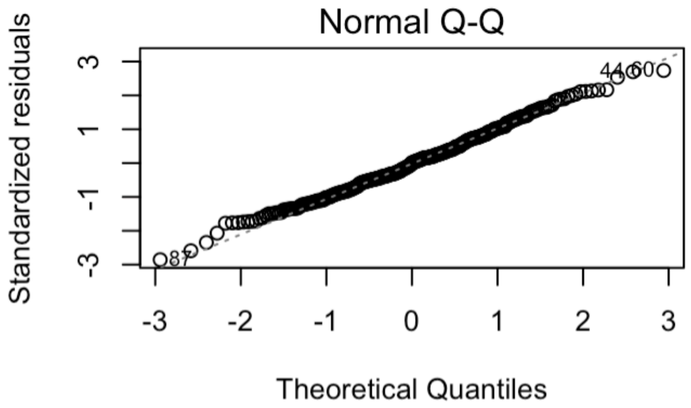


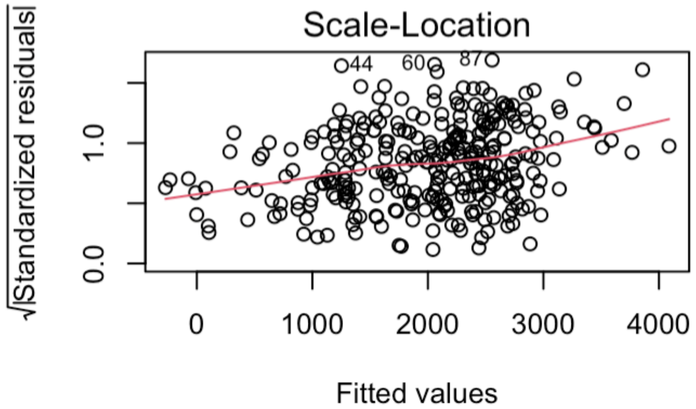


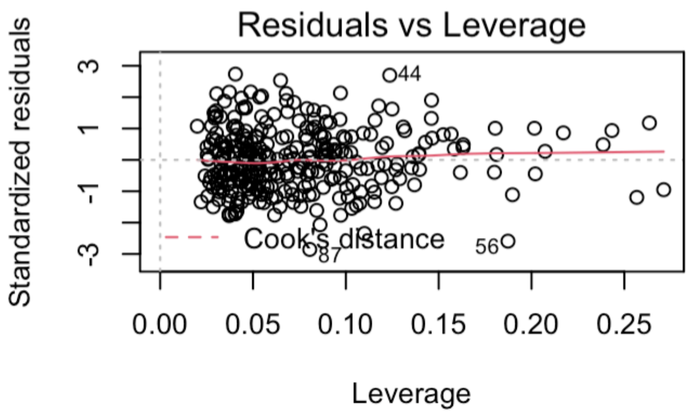
Multivariate Regression Analysis

Regression Diagnostics









Nonlinearity:
add age^2 as an additional explanatory variable

Heteroscedasticity:
transform the response by the square-root

Outliers:
drop i = 44, 56, 87