Code

Bilirubin concentrations and Ascities levels in Primary Biliary Cirrhosis.

0.1 Preliminaries

Hide

```
library(skimr)
library(rms)
library(simputation)
library(broom)
library(modelr)
library(arm)
library(pander)
library(ROCR)
library(pROC)
library(forcats)
library(car)
library(tidyverse)
```

1 Task 1: Data Source

This dataset was found in appendix D of Fleming and Harrington, Counting Processes and Survival Analysis, Wiley, 1991. I have taken it from http://lib.stat.cmu.edu/datasets/ The dataset contains the data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984.

A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants. Missing data items are denoted by `.'. Thus, since many of the values were missing for last 112 people, I chose the first 312 values for the project. A more extended discussion can be found in Dickson, et al., Hepatology 10:1-7 (1989) and in Markus, et al., N Eng J of Med 320:1709-13 (1989).

2 Task 2: Load and Tidy the Data

Hide

```
pbc <- read.csv("C:\\Users\\mani\\Desktop\\Case Acad\\Spring 2018\\tidypbc1.csv", header=T, na.strings=c(".",
na.strings=c("",". ","NA") %>% tbl_df()

pbc <- pbc %>% rename(alk_phos = alk.phos)

map_df(pbc, function(x) sum(is.na(x)))
```

```
# A tibble: 1 x 19
     ID fu.days status  drug  chol   alb copper alk_phos  sgot triglyc
  <int>   <int>  <int> <int> <int> <int>  <int>    <int> <int>   <int>
1     0       0      0     0    28     0      2        0     0      30
```

```
# ... with 9 more variables: plat <int>, protime <int>, stage <int>,
#   sex <int>, Bili <int>, ascities <int>, hepatem <int>, spiders <int>,
#   edema <int>
```

Hide

```
## As we can see, Cholesterol has 28 missing values, Copper has 2, platelets has 4, and triglycerides has 30.

set.seed(40009)
pbc1 <- pbc %>% select(chol, copper, drug, fu.days, ID, plat, sex, stage, status, triglyc, alb, alk_phos, Bil

pbc2 <- pbc1
pbc2 <- pbc2 %>% mutate(status = as.factor(ifelse(status < 2, "Censored", "Death")))
pbc2 <- pbc2 %>% rename(female = sex)
pbc2 <- pbc2 %>%  mutate(drug = ifelse(drug == 1, "D-penicillamine", "Placebo"))

pbc2 <- pbc2 %>% mutate(stage = as.factor(ifelse(stage == 1, "Early", ifelse(stage == 2, "Mid", ifelse(stage
pbc2 <- pbc2 %>% mutate(edema = as.factor(ifelse(edema < 0.5, "No Edema", "Edema")))
```

1. Step 1: Converted all the "." to NA values in order for skim to work.
2. Step 2: Checked for the missing values, if any. Found that Cholesterol has 28 missing values, Copper has 2, platelets has 4, and triglycerides has 30.
3. Step 3: Performed simple imputation to add the missing values in the numeric variables. The reason I performed simple imputation was that the number of missing values isn't very large in the variables.
4. Step 4: Converted Status to a binary variable. Renamed Sex as Female Converted Drug into a character variable. Converted Stage into a factor variable with multiple levels. Converted edema into a factor variable with two levels.

3 Task 3: Listing of My Tibble

Hide

```
pbc2 %>% tbl_df()
```

```
# A tibble: 312 x 19
    chol copper drug      fu.days    ID  plat female stage  status triglyc
   <dbl>  <dbl> <chr>       <int> <int> <dbl>  <int> <fct>  <fct>    <dbl>
 1   261  156   D-penici~     400     1   190      1 Extre~ Death    172  
 2   302   54.0 D-penici~    4500     2   221      1 Advan~ Censo~    88.0
 3   176  210   D-penici~    1012     3   151      0 Extre~ Death     55.0
 4   244   64.0 D-penici~    1925     4   183      1 Extre~ Death     92.0
 5   279  143   Placebo      1504     5   136      1 Advan~ Censo~    72.0
 6   248   50.0 Placebo      2503     6   296      1 Advan~ Death     63.0
 7   322   52.0 Placebo      1832     7   204      1 Advan~ Censo~   213  
 8   280   52.0 Placebo      2466     8   373      1 Advan~ Death    189  
 9   562   79.0 D-penici~    2400     9   251      1 Mid    Death     88.0
10   200  140   Placebo        51    10   302      1 Extre~ Death    143  
# ... with 302 more rows, and 9 more variables: alb <dbl>, alk_phos <dbl>,
#   Bili <dbl>, protime <dbl>, sgot <dbl>, edema <fct>, spiders <int>,
#   hepatem <int>, ascities <int>
```

The tibble has 312 observations(rows) in 19 columns, that is, 19 variables.

4 Task 4: Code Book

Variable	Type	Details
ID	Integer	ID(case number) of the people

Variable	Type	Details
`fu.days`	Integer	Number of days between registration and the earlier of death,transplantion, or study analysis time in July, 1986
`female`	Integer	Here, 1 means female, and 0 male
`stage`	factor	The stage of PBC (has 4 levels)
`status`	factor	Only two levels- Censored or Death
`drug`	Character	Two categories: D-penicillamine or Placebo
`alb`	numeric	values in gm/dl, ranging from 1.96 to 4.64 gm/dl
`plat`	numeric	values in cubic ml/1000, ranging from 62 to 563 cubic ml.
`chol`	numeric	Values in mg/dl, ranging from 120 to 1775 mg/dl
`Copper`	numeric	Values in ug/day, ranging from 4 to 588 ug/day
`triglyc`	numeric	In mg/dl, ranging from 33 to 598 mg/dl
`alk_phos`	numeric	In U/l.Ranges from 289 to 13862 U/l.
`Bili`	numeric	In mg/dl. Ranges from 0.3 to 28 mg/dl
`protime`	numeric	In seconds. Ranges from 9 to 17.2 seconds
`sgot`	numeric	In U/ml. Ranges from 26.35 to 457.25 U/ml
`edema`	factor	Presence or absence of edema
`hepatem`	integer	Presence of hepatomegaly. Binary variable
`ascities`	integer	Presence of ascities. Binary variable
`spiders`	integer	Presence of spider angiomas. Binary variable.

5 Task 5: My Subjects

This dataset is about the PBC(primary biliary cirrhosis) trial conducted in 312 patients from 1974-1984. One of the purposes of the study was to make survival models for patients with PBC, using Serum Bilirubin and albumin concentrations and prothrombin time. Further information is provided in the paper: Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D. and Langworthy, A. (1989), Prognosis in primary biliary cirrhosis: Model for decision making. Hepatology, 10: 1-7. doi:10.1002/hep.1840100102

6 Task 6: My Variables

There are 19 variables (or columns) in the dataset:

1. ID: Specifes the case number of the patients. A total of 312 patients in this study.
2. fu.days: Number of days between registration and the earlier of death, transplantion, or study analysis time in July, 1986
3. female: Gender of the patients involved in the study.
4. stage: The stage at which the disease was at. It is a multicategorical variable with 4 different levels.
5. status: Status of the patients when the trial ended. Either dead or censored
6. drug: The drug patients were given. They were either given D-penicillamine, or placebo.
7. alb: The concentration of albumin present in the serum. Given in gm/dl.
8. plat: The concentration of platelets in the patients. It is given in cubic ml/1000
9. chol: The concentration of cholesterol in the patients. It is given in mg/dl.
10. copper: The concentration of copper removed through urine. It is given in ug/day.
11. triglyc: The concentration of triglycerides in the patients. Given in mg/dl.
12. alk_phos: The concentration of alkaline phosphatase, given in U/l

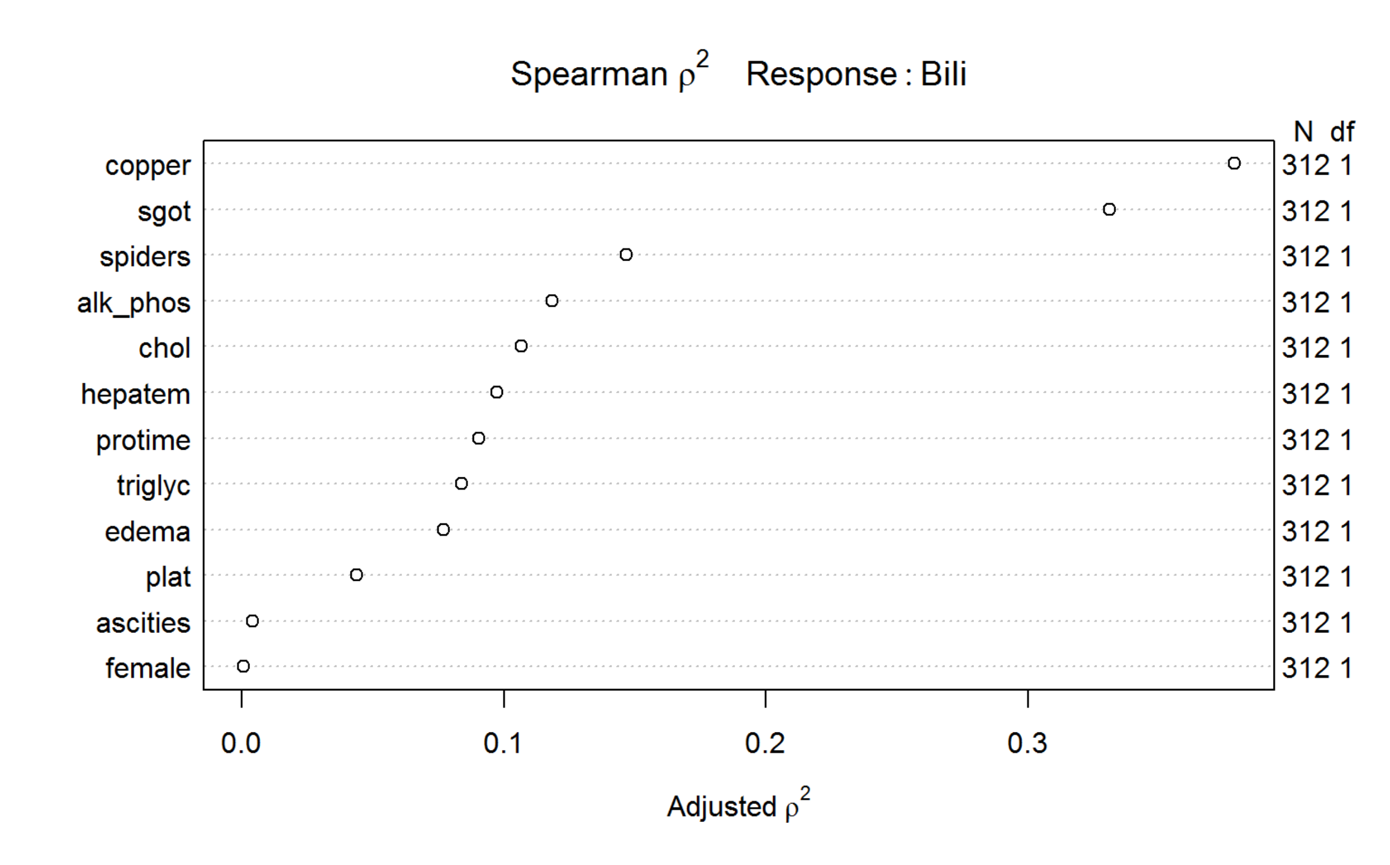
13. sgot: Serum glutamic oxaloacetic transaminase, an enzyme secreted by the liver. It’s concentration is provided in U/ml
14. protime: It is the time taken by prothrombin to form. It is provided in seconds.
15. Bili: Bilirubin concentrations in the serum. Given in mg/dl
16. ascities: It denotes the presence or absence of ascities, which is abnormal accumulation of fluids
17. hepatem: Hepatomegaly , is the abnormal enlargement of liver, and is given as whether present or absent.
18. spiders: spider angiomas, is a disease caused in the liver. It is a binary variable here.
19. edema: Also refers to accumulation of abnormal quantity of fluids, but in different areas. It is converted into a binary factor here.

7 Task 7: My Planned Linear Regression Model

Higher bilirubin levels are associated with occruance of PBC. I plan to see concentrations of other variables and how they affect the Bilirubin levels. Thus, I plan on having the variable “bilirubin” as the outcome variable. My predicting variables shall be: 1. Copper 2. SGOT 3. Triglyc 4. Protime 5. Albumin 6. Hepatem

Hide

```
spear.Bili <- spearman2(Bili ~ alk_phos + protime + triglyc + chol + female + copper + sgot + plat + ascities

plot(spear.Bili)
```



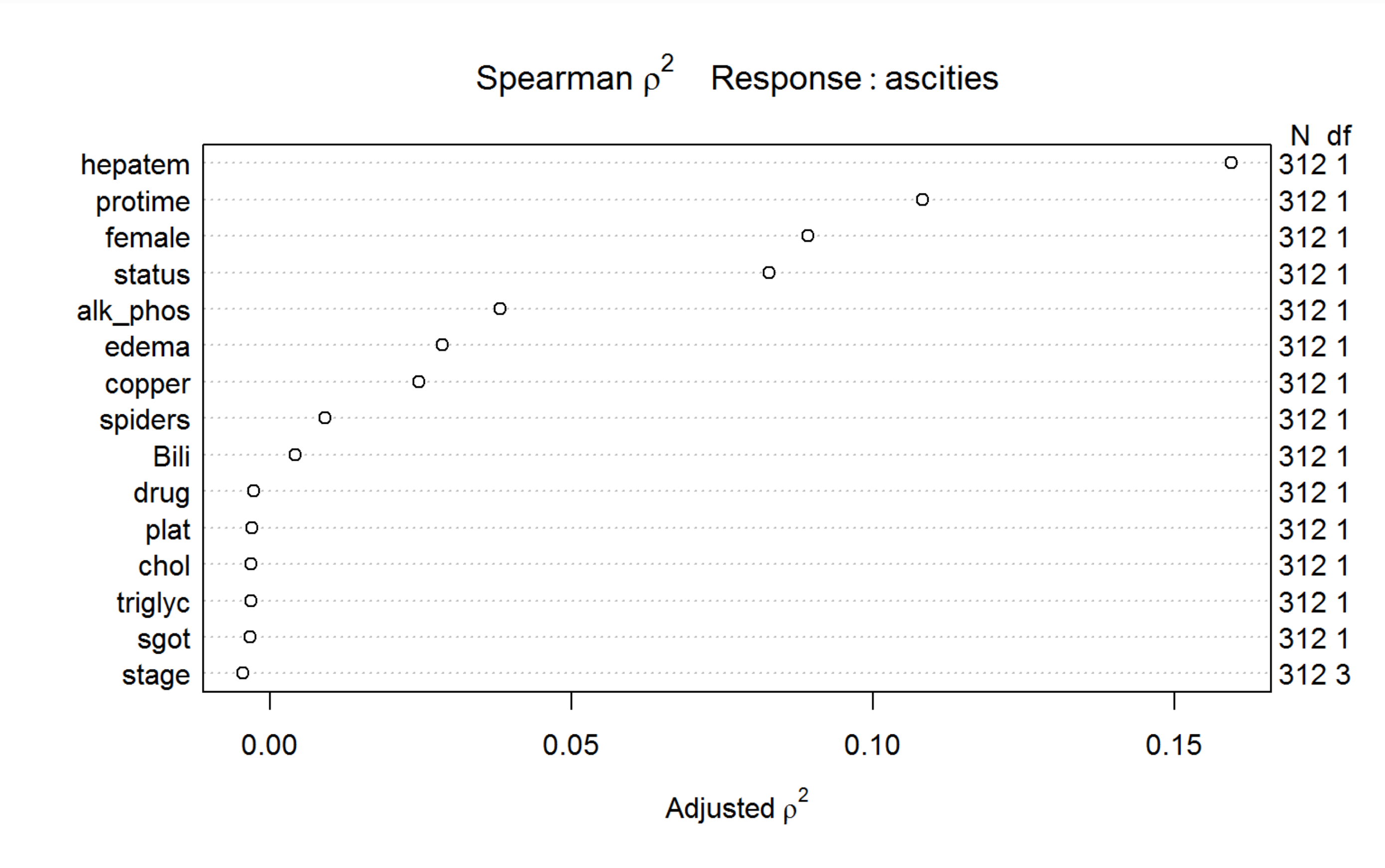
Here, Copper and SGOT are the two most important variables according to the spearman Rho square plot

8 Task 8: My Planned Logistic Regression Model

I plan on having “Ascities” as the binary outcome variable. My other predictors shall be: 1. hepatem 2. Protime 3. Female 4. Status 5. Alkaline Phosphatase.

Hide

```
spear.ascities <- spearman2(ascities ~ alk_phos + protime + triglyc + chol + female + copper + sgot + plat +

plot(spear.ascities)
```



Hepatem is supposed to be the most important variable here, according to the Spearman Rho square plot. The multi-categorical variable is the stage variable. Predictions using stage variable can be made fr ascities.

9 Task 9: Affirmation

The dataset fulfills all the necessary requirements of the project. It has more than 100 observations in 19 variables.

I am certain that it is completely appropriate for this data to be shared with anyone, without any conditions. There are no concerns about privacy or security.

10 Task 10: Linear Regression

10.1 Exploratory Analysis

Hide

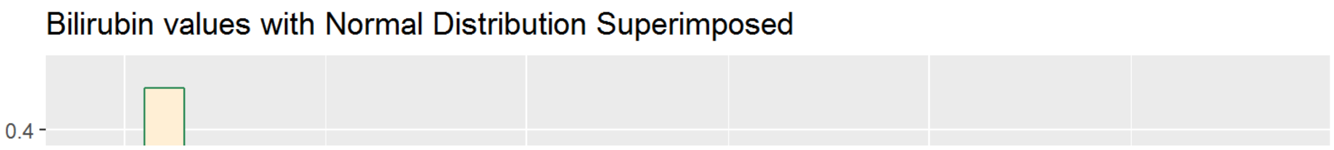
```
skim(pbc2)
```

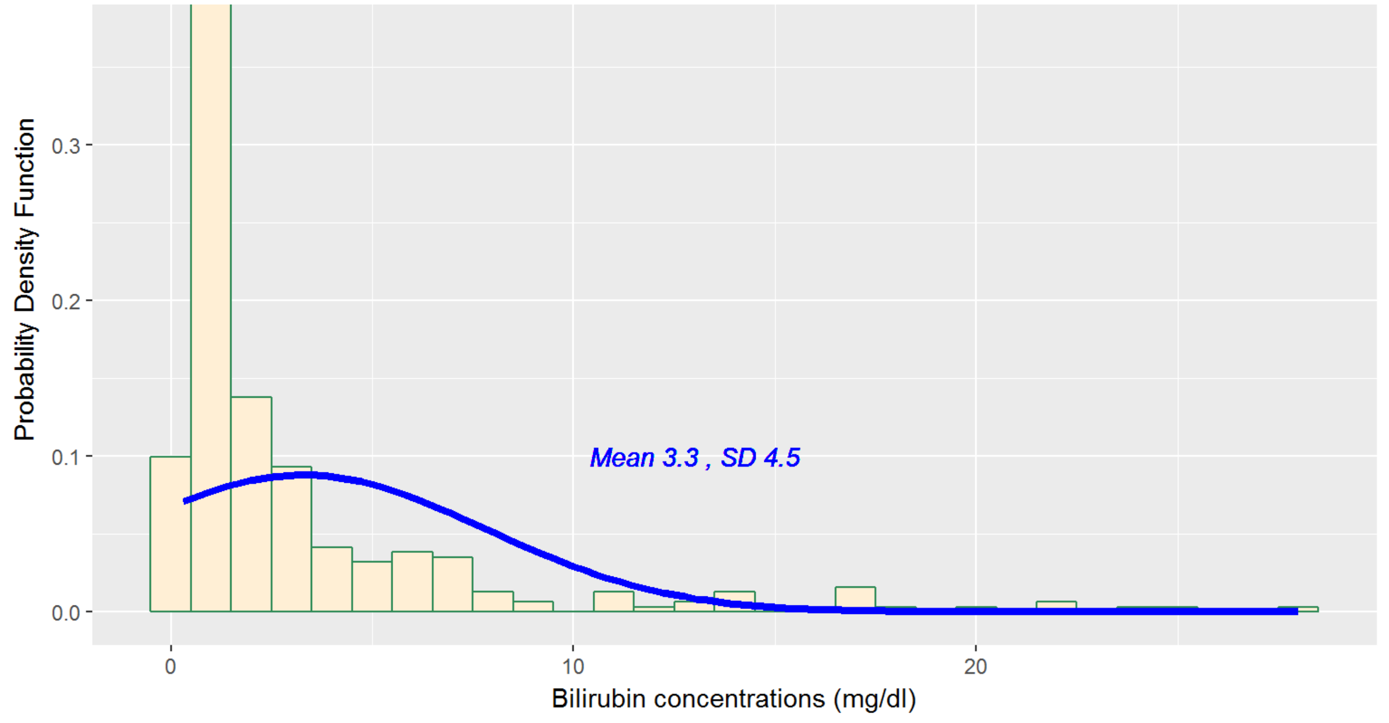
```
Skim summary statistics
 n obs: 312
 n variables: 19

Variable type: character
 variable missing complete   n min max empty n_unique
     drug       0      312 312   7  15     0        2

Variable type: factor
 variable missing complete   n n_unique
    edema       0      312 312        2
    stage        0      312 312        4
   status        0      312 312        2
                           top_counts ordered
             No : 230, Ede: 82, NA: 0   FALSE
 Adv: 120, Ext: 109, Mid: 67, Ear: 16   FALSE
            Cen: 187, Dea: 125, NA: 0   FALSE

Variable type: integer
 variable missing complete   n    mean      sd p0     p25 median     p75
 ascities        0      312 312    0.65    0.48  0     0       1      1
   female        0      312 312    0.88    0.32  0     1       1      1
  fu.days        0      312 312 2006.36 1123.28 41 1191    1839.5 2697.25
  hepatem        0      312 312    0.21    0.41  0     0       0      0
       ID        0      312 312   156.5    90.21  1   78.75  156.5  234.25
  spiders        0      312 312    0.44    0.5   0     0       0      1
 p100     hist
    1 <U+2585><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2587>
    1 <U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2587>
 4556 <U+2583><U+2586><U+2587><U+2586><U+2586><U+2583><U+2582><U+2582>
    1 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2582>
  312 <U+2587><U+2587><U+2587><U+2587><U+2587><U+2587><U+2587><U+2587>
    1 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2586>

Variable type: numeric
 variable missing complete   n    mean      sd      p0     p25   median
      alb        0      312 312    3.52    0.42   1.96    3.31    3.55
 alk_phos        0      312 312 1982.66 2140.39 289     871.5  1259
     Bili        0      312 312    3.28    4.53   0.3     0.8     1.4
     chol        0      312 312  364.58  223.86 120     249.5   308.5
   copper        0      312 312   97.55   85.38   4      41.75   73
     plat        0      312 312  261.87   95.44  62     199.75  258
  protime        0      312 312   10.73    1      9      10      10.6
     sgot        0      312 312  122.56   56.7  26.35   80.6   114.7
  triglyc        0      312 312  122.33   63.09 33      84     108
    p75     p100     hist
    3.8     4.64 <U+2581><U+2581><U+2581><U+2583><U+2587><U+2586><U+2583><U+2581>
 1980   13862.4  <U+2587><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
    3.5    28    <U+2587><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
 396     1775    <U+2587><U+2585><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
 123      588    <U+2587><U+2583><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581>
  322.5   563    <U+2582><U+2585><U+2587><U+2587><U+2585><U+2582><U+2581><U+2581>
   11.1    17.1  <U+2585><U+2587><U+2583><U+2581><U+2581><U+2581><U+2581><U+2581>
  151.9   457.25 <U+2585><U+2587><U+2585><U+2582><U+2581><U+2581><U+2581><U+2581>
  146     598    <U+2587><U+2587><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581>
```

I then checked whether this was a normal distribution or not.

Hide

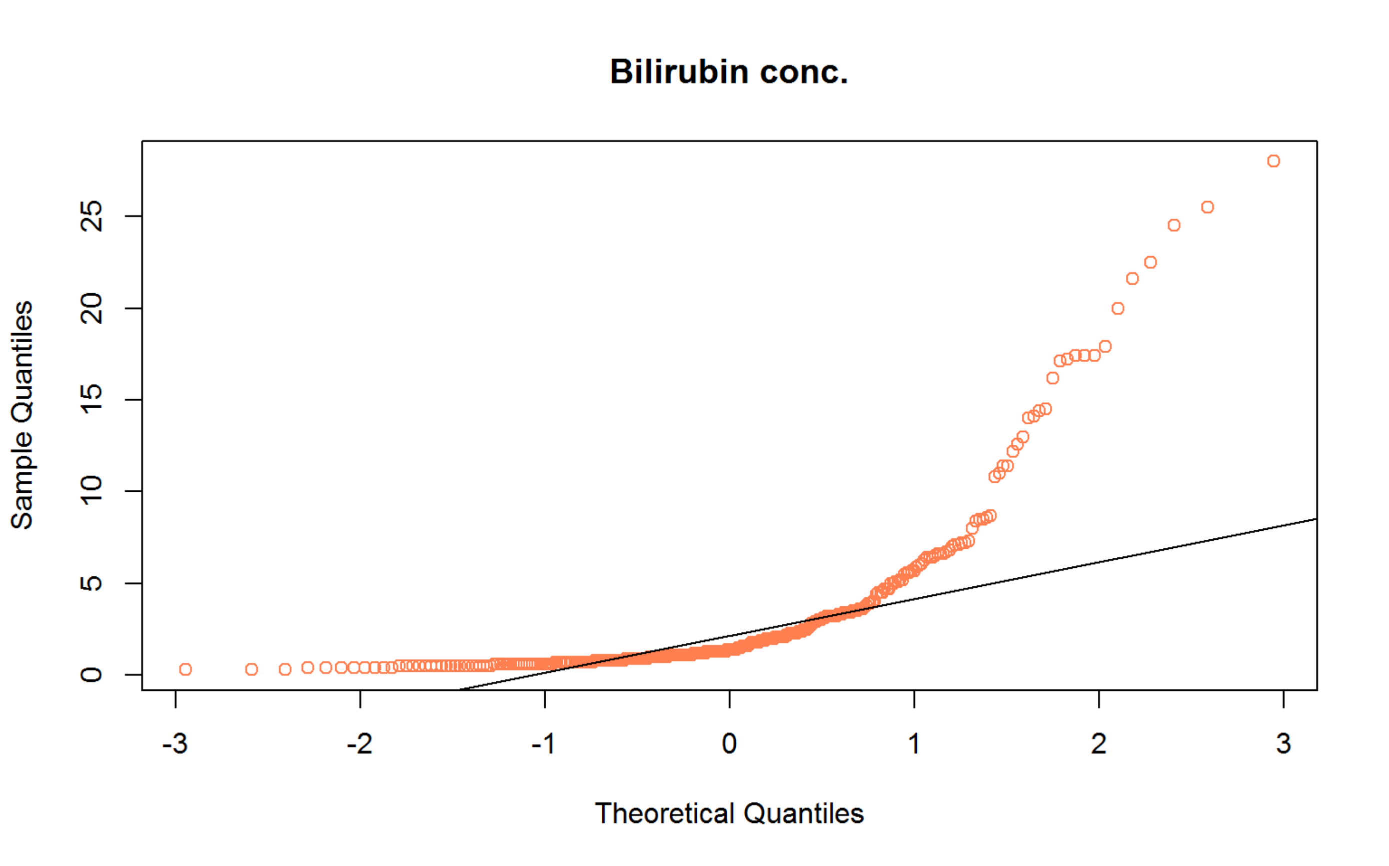
```
ggplot(pbc2, aes(x=Bili)) +
geom_histogram(aes(y = ..density..), binwidth=1, fill = "papayawhip", color = "seagreen") +    stat_function(
args = list(mean = mean(pbc2$Bili),                                                    sd = sd(pbc2$Bili)),
  lwd = 1.5, col = "blue") +
geom_text(aes(label = paste("Mean", round(mean(pbc2$Bili),1),
                                  ", SD", round(sd(pbc2$Bili),1))),
                x = 13, y = 0.1, color="blue", fontface = "italic") +
    labs(title = "Bilirubin values with Normal Distribution Superimposed",
           x = "Bilirubin concentrations (mg/dl)", y = "Probability Density Function")
```





Hide

```
# Checking the QQ plot

qqnorm(pbc2$Bili, main="Bilirubin conc.", col="coral")
qqline(pbc2$Bili)
```

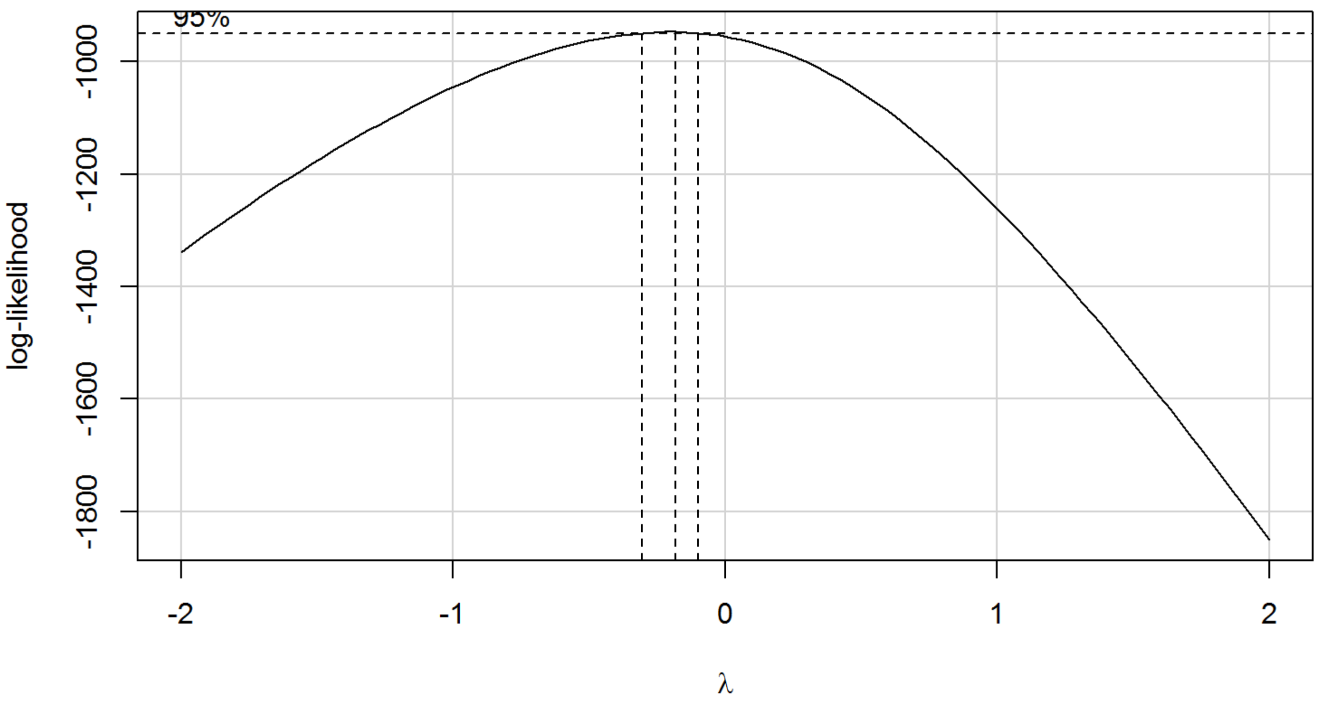


As we can see, the histogram and the QQ plot show that the distribution is not normal. Thus, I made a box cox plot to check for the Y1 value.

10.1.1 Transformation

Hide

```
boxCox(lm(Bili ~ copper + female + sgot + alk_phos + stage + drug + hepatem + protime + plat + triglyc + alb,
```



Hide

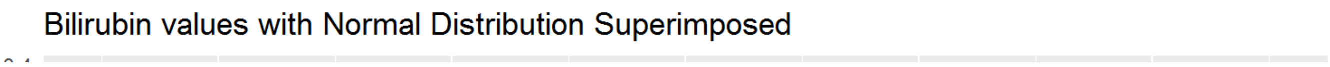
```
powerTransform(lm(Bili ~ copper + female + sgot + alk_phos + stage + drug + hepatem + protime + plat + trigly
```

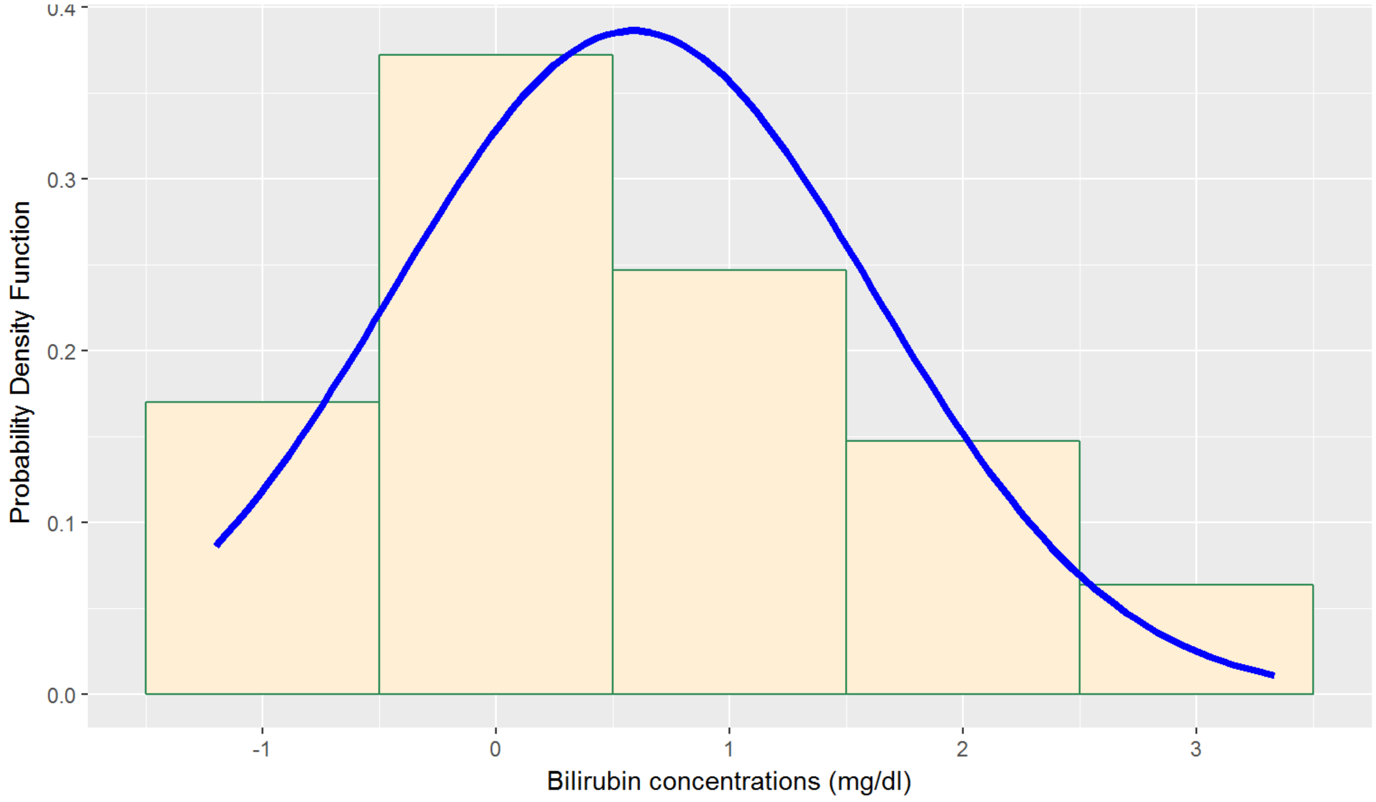
```
Estimated transformation parameters
        Y1
-0.2016253
```

The Y1 value was found to be -0.2, which is closest to 0. Therefore, I converted the Bilirubin values into its natural logarithm.

Hide

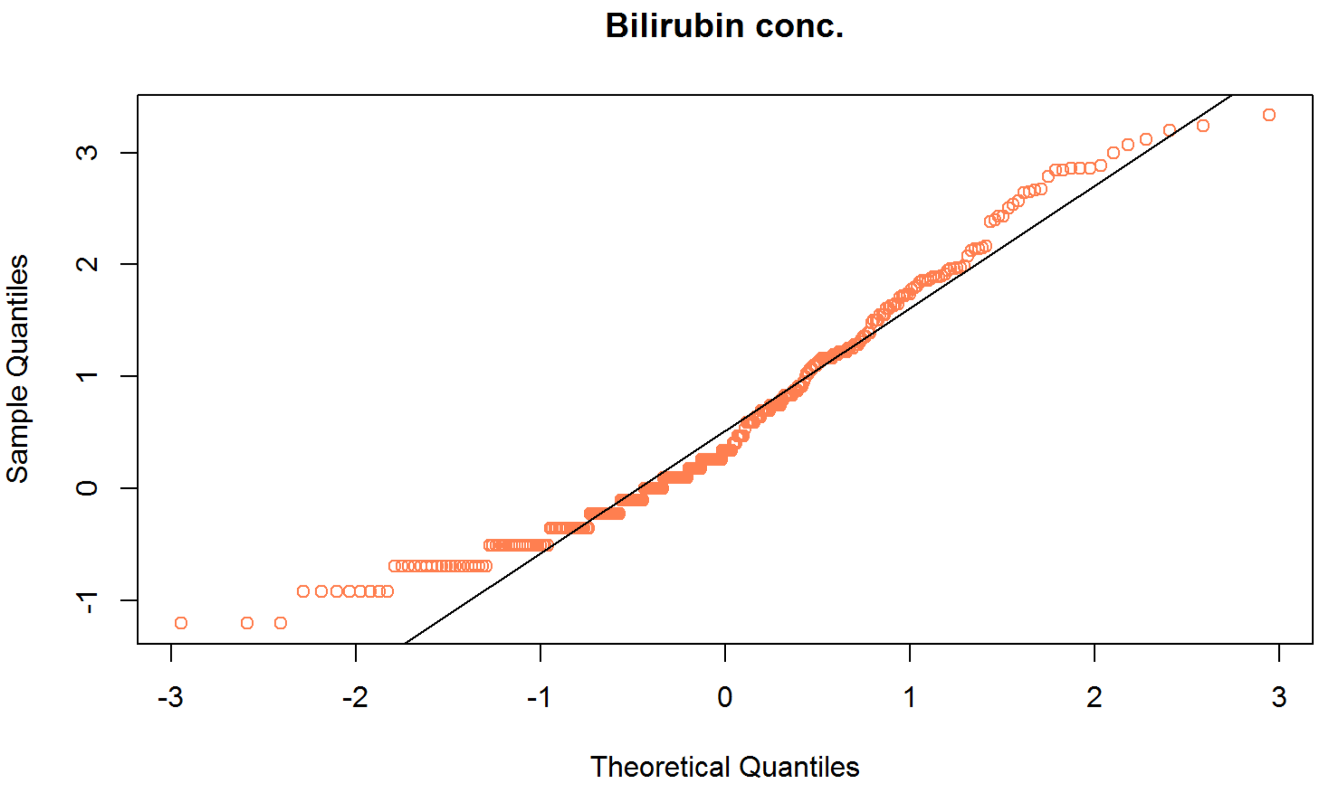
```
pbc4 <- pbc2
pbc4 <- pbc4 %>% mutate(Bili = log(Bili))
ggplot(pbc4, aes(x=Bili)) +
geom_histogram(aes(y = ..density..), binwidth=1,
                     fill = "papayawhip", color = "seagreen") +
      stat_function(fun = dnorm,
                    args = list(mean = mean(pbc4$Bili),
                                sd = sd(pbc4$Bili)),
                    lwd = 1.5, col = "blue") +
      geom_text(aes(label = paste("Mean", round(mean(pbc4$Bili),1),
                                  ", SD", round(sd(pbc4$Bili),1))),
                x = 13, y = 0.1, color="blue", fontface = "italic") +
     labs(title = "Bilirubin values with Normal Distribution Superimposed",
          x = "Bilirubin concentrations (mg/dl)", y = "Probability Density Function")
```





Hide

```
# Now, I checked whether transformation had an effect on the Q-Q plot or not.


qqnorm(pbc4$Bili, main="Bilirubin conc.", col="coral")
qqline(pbc4$Bili)
```



Hence, I proceeded with the transformed values to make my model.

10.3 First Model: Fit

10.2 First Model: Kitchen Sink

I first made a Kitchen Sink Model

Hide

```
# Kitchen sink
model_ks <- lm(Bili~copper + female + sgot + alk_phos + stage + drug + hepatem + protime + plat + triglyc + a
```

I then decided to reduce the number of variables, since the degrees of freedom used in the kitchen sink model would be high.

Hide

```
# Stepwise Forward Regression
with(pbc4,
     step(lm(Bili ~ 1),
           scope=(~ copper + female + sgot + alk_phos + stage + drug + hepatem + protime + plat + triglyc + a
```

```
Start:  AIC=21.47
Bili ~ 1

           Df Sum of Sq    RSS     AIC
+ sgot      1    94.259 237.83 -80.696
+ copper    1    93.561 238.53 -79.781
+ alb       1    47.574 284.51 -24.775
+ triglyc   1    42.435 289.65 -19.190
+ stage     3    44.880 287.21 -17.835
+ protime   1    38.026 294.06 -14.477
+ hepatem   1    35.068 297.02 -11.354
+ plat      1    10.847 321.24  13.105
+ alk_phos  1     7.412 324.67  16.423
<none>                  332.09  21.466
+ drug      1     0.608 331.48  22.894
+ female    1     0.378 331.71  23.110

Step:  AIC=-80.7
Bili ~ sgot

           Df Sum of Sq    RSS      AIC
+ copper    1    50.929 186.90 -153.881
+ triglyc   1    30.429 207.40 -121.410
+ protime   1    26.109 211.72 -114.979
+ hepatem   1    25.757 212.07 -114.460
+ stage     3    28.105 209.72 -113.934
+ alb       1    23.821 214.01 -111.624
+ plat      1     4.725 233.10  -84.958
+ alk_phos  1     2.701 235.13  -82.259
<none>                  237.83  -80.696
+ female    1     0.346 237.48  -79.151
+ drug      1     0.138 237.69  -78.878

Step:  AIC=-153.88
Bili ~ sgot + copper

           Df Sum of Sq    RSS     AIC
+ protime   1   14.3005 172.60 -176.72
+ hepatem   1   14.2369 172.66 -176.60
+ triglyc   1   14.2303 172.67 -176.59
+ alb       1   11.7924 175.11 -172.22
+ stage     3   13.2036 173.69 -170.74
+ plat      1    3.7388 183.16 -158.19
```

```
<none>                  186.90 -153.88
+ female    1    0.4820 186.42 -152.69
+ alk_phos  1    0.2371 186.66 -152.28
+ drug      1    0.2309 186.67 -152.27

Step:  AIC=-176.72
Bili ~ sgot + copper + protime

           Df Sum of Sq    RSS     AIC
+ triglyc   1   15.7562 156.84 -204.58
+ alb       1    7.8048 164.79 -189.15
+ hepatem   1    7.3802 165.22 -188.35
+ stage     3    7.8791 164.72 -185.29
+ plat      1    1.3580 171.24 -177.18
<none>                  172.60 -176.72
+ female    1    0.7051 171.89 -175.99
+ alk_phos  1    0.0951 172.50 -174.89
+ drug      1    0.0414 172.56 -174.79

Step:  AIC=-204.58
Bili ~ sgot + copper + protime + triglyc

           Df Sum of Sq    RSS     AIC
+ alb       1    7.3243 149.52 -217.50
+ stage     3    7.2046 149.64 -213.26
+ hepatem   1    4.6688 152.17 -212.01
+ plat      1    2.9967 153.84 -208.60
<none>                  156.84 -204.58
+ female    1    0.5677 156.27 -203.72
+ alk_phos  1    0.0338 156.81 -202.65
+ drug      1    0.0181 156.82 -202.62

Step:  AIC=-217.51
Bili ~ sgot + copper + protime + triglyc + alb

           Df Sum of Sq    RSS     AIC
+ hepatem   1    3.0358 146.48 -221.91
+ stage     3    4.1341 145.38 -220.25
+ plat      1    1.8102 147.71 -219.31
<none>                  149.52 -217.50
+ female    1    0.2937 149.22 -216.12
+ alk_phos  1    0.0787 149.44 -215.67
+ drug      1    0.0468 149.47 -215.60

Step:  AIC=-221.91
Bili ~ sgot + copper + protime + triglyc + alb + hepatem

           Df Sum of Sq    RSS     AIC
+ plat      1   1.47757 145.00 -223.07
+ stage     3   3.05316 143.43 -222.48
<none>                  146.48 -221.91
+ alk_phos  1   0.50304 145.98 -220.98
+ female    1   0.19857 146.28 -220.33
+ drug      1   0.10287 146.38 -220.12

Step:  AIC=-223.07
Bili ~ sgot + copper + protime + triglyc + alb + hepatem + plat

           Df Sum of Sq    RSS     AIC
<none>                  145.00 -223.07
+ stage     3   2.49314 142.51 -222.48
+ female    1   0.35532 144.65 -221.83
+ alk_phos  1   0.24380 144.76 -221.59
+ drug      1   0.14560 144.86 -221.38
```

```

Call:
lm(formula = Bili ~ sgot + copper + protime + triglyc + alb + 
    hepatem + plat)

Coefficients:
(Intercept)         sgot       copper      protime      triglyc  
 -1.4021439    0.0067010    0.0032284    0.1582353    0.0036073  
        alb      hepatem         plat  
 -0.3244971    0.2511950   -0.0007621  
```

The variables obtained from forward regression were: sgot, copper, Protime, triglyc, alb, hepatem, plat.

I then made a model using these variables:

Hide

```
model_fw2 <- lm(Bili~ sgot + copper + protime + triglyc + alb + hepatem + plat, data = pbc4)
summary(model_fw2)
```

```

Call:
lm(formula = Bili ~ sgot + copper + protime + triglyc + alb + 
    hepatem + plat, data = pbc4)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.75768 -0.44398 -0.04732  0.42181  2.18988 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.4021439  0.6453766  -2.173 0.030583 *  
sgot         0.0067010  0.0007341   9.129  < 2e-16 ***
copper       0.0032284  0.0005163   6.253 1.36e-09 ***
protime      0.1582353  0.0431831   3.664 0.000293 ***
triglyc      0.0036073  0.0006627   5.443 1.08e-07 ***
alb         -0.3244971  0.1018311  -3.187 0.001589 ** 
hepatem      0.2511950  0.1055180   2.381 0.017901 *  
plat        -0.0007621  0.0004330  -1.760 0.079407 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6906 on 304 degrees of freedom
Multiple R-squared:  0.5634,	Adjusted R-squared:  0.5533 
F-statistic: 56.03 on 7 and 304 DF,  p-value: < 2.2e-16
```

The R square value was found to be 0.56, and sgot, copper, protime, triglyc were seen to significantly affect the Bilirubin concetrations.

I then compared the Forward regression model with the Kitchen Sink Model

10.3 Comparisons

Hide

```
anova(model_ks, model_fw2)
```

```
Analysis of Variance Table
```

```
Model 1: Bili ~ copper + female + sgot + alk_phos + stage + drug + hepatem +
    protime + plat + triglyc + alb
Model 2: Bili ~ sgot + copper + protime + triglyc + alb + hepatem + plat
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    298 141.94
2    304 145.00 -6   -3.0641 1.0722 0.3792
```

Hide

```
glance(model_ks)
```

```
  r.squared adj.r.squared     sigma statistic      p.value df    logLik
1  0.572582     0.5539362 0.6901499  30.70844 1.666812e-47 14 -319.8428
       AIC      BIC deviance df.residual
1 669.6856 725.8307 141.9395         298
```

Hide

```
glance(model_fw2)
```

```
  r.squared adj.r.squared     sigma statistic      p.value df    logLik
1 0.5633553      0.553301 0.6906412  56.03117 4.240373e-51  8 -323.1746
       AIC      BIC deviance df.residual
1 664.3492 698.0362 145.0035         304
```

Here, the R squared value for the Kitchen Sink model is higher than the Forward regression model, but the kitchen sink model uses more degrees of freedom.

For Kitchen sink

Hide

```
set.seed(43201)

cv_model_ks <- pbc4 %>%
crossv_kfold(k = 10) %>%
mutate(model = map(train,  ~ lm(Bili ~ sgot + copper + protime + alk_phos + female + triglyc + alb + hepatem

cv_model_pred2 <- cv_model_ks %>%
unnest(map2(model, test, ~ augment(.x, newdata = .y)))
 cv_model_results2 <- cv_model_pred2 %>% dplyr::summarize(
           RMSE_ks = sqrt(mean((Bili - .fitted) ^2)),
            MAE_ks = mean(abs(Bili - .fitted))) %>% round(., 3)
head(cv_model_pred2, 3)
```

```
# A tibble: 3 x 22
  .id    chol copper drug  fu.days    ID  plat female stage status triglyc
  <chr> <dbl>  <dbl> <chr>   <int> <int> <dbl>  <int> <fct> <fct>    <dbl>
1 01      235   39.0 D-pe~    4232    19 209        1 Adva~ Censo~     123
2 01      374  140   Plac~    1356    20 322        1 Extr~ Death      135
3 01      456  124   D-pe~    4079    24  70.0      0 Mid   Death      230
# ... with 11 more variables: alb <dbl>, alk_phos <dbl>, Bili <dbl>,
#   protime <dbl>, sgot <dbl>, edema <fct>, spiders <int>, hepatem <int>,
#   ascities <int>, .fitted <dbl>, .se.fit <dbl>
```

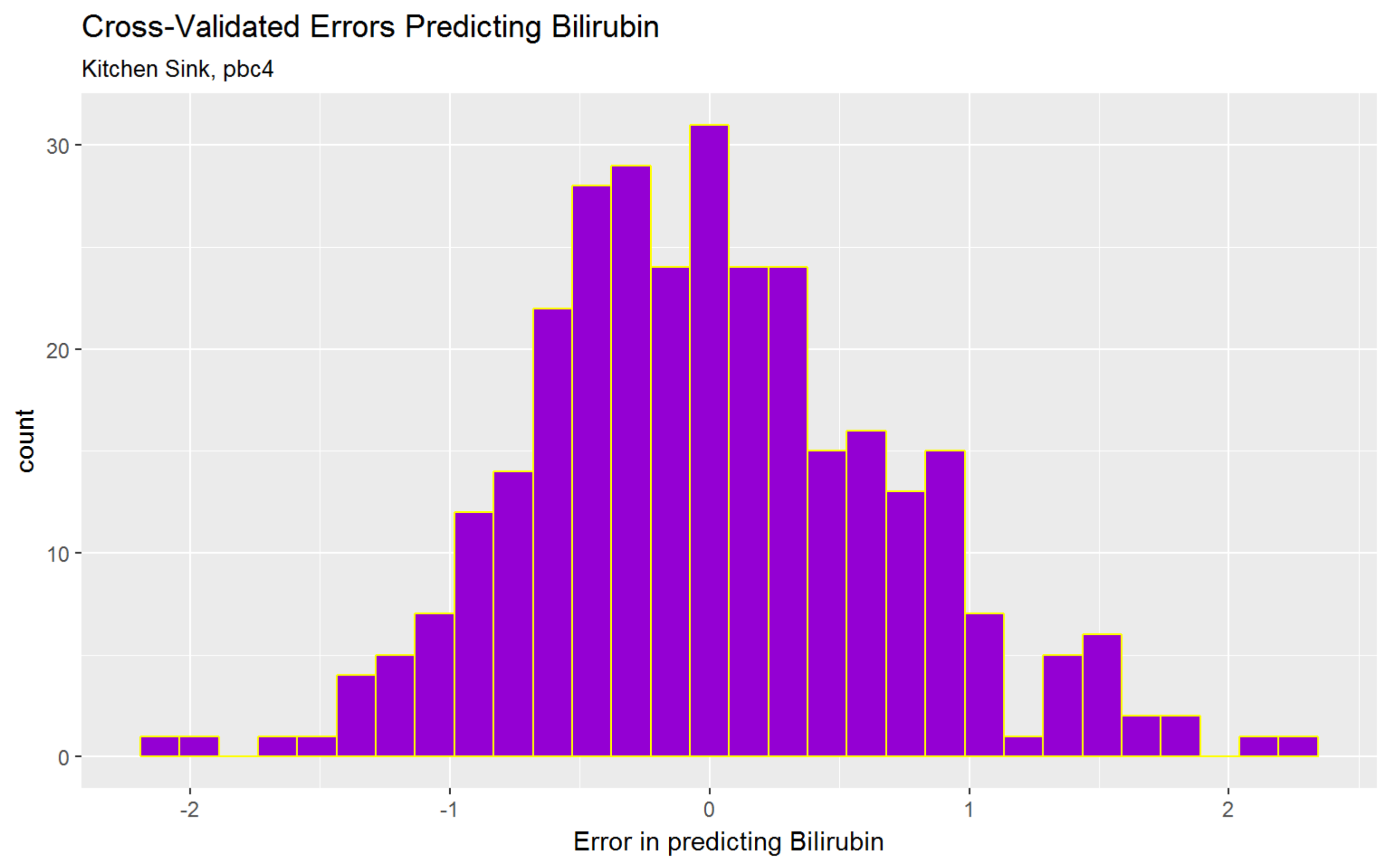
Hide

```
cv_model_results2
```

```
# A tibble: 1 x 2
  RMSE_ks MAE_ks
    <dbl>  <dbl>
1   0.705  0.555
```

Hide

```
# The RMS and MAE values for the kitchen sink model are 0.705 and 0.555 respectively

cv_model_pred2 %>%
mutate(errors = Bili - .fitted) %>%
ggplot(., aes(x = errors)) +
geom_histogram(bins = 30, fill = "darkviolet", col = "yellow") + labs(title = "Cross-Validated Errors Predict
x = "Error in predicting Bilirubin")
```



Hide

```
# FOr the Forward regression model

set.seed(543210)

cv_model_fw <- pbc4 %>%
crossv_kfold(k = 10) %>%
mutate(model = map(train,
 ~ lm(Bili ~ sgot + copper + protime + triglyc + alb + hepatem + plat, data= .)))

cv_model_pred <- cv_model_fw %>%
unnest(map2(model, test, ~ augment(.x, newdata = .y)))

cv_model_results <- cv_model_pred %>% dplyr::summarize(
          RMSE = sqrt(mean((Bili - .fitted) ^2)),
          MAE = mean(abs(Bili - .fitted))) %>% round(., 3)
head(cv_model_pred, 3)
```

```
# A tibble: 3 x 22
  .id    chol copper drug  fu.days    ID  plat female stage status triglyc
  <chr> <dbl>  <dbl> <chr>   <int> <int> <dbl>  <int> <fct> <fct>    <dbl>
1 01      259   46.0 Plac~    3762    11   258      1 Extr~ Death     79.0
2 01      235   39.0 D-pe~    4232    19   209      1 Adva~ Censo~   123
3 01      260  231   D-pe~    3282    57   216      1 Adva~ Death     94.0
# ... with 11 more variables: alb <dbl>, alk_phos <dbl>, Bili <dbl>,
#   protime <dbl>, sgot <dbl>, edema <fct>, spiders <int>, hepatem <int>,
#   ascities <int>, .fitted <dbl>, .se.fit <dbl>
```

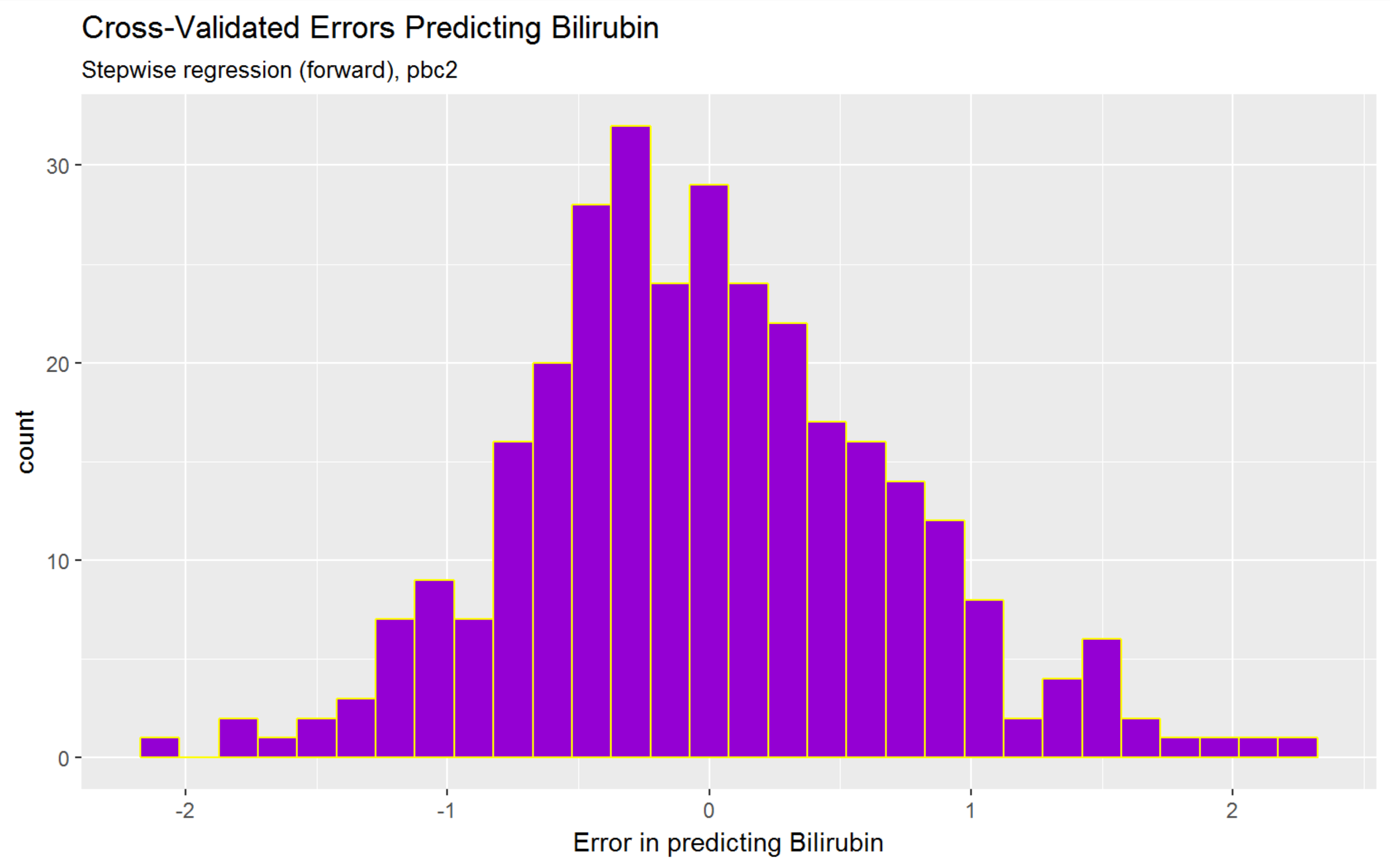
Hide

```
cv_model_results
```

```
# A tibble: 1 x 2
   RMSE   MAE
  <dbl> <dbl>
1 0.708 0.555
```

Hide

```
# The RMSE and MAE values for Forward regression model are 0.700 and 0.551

cv_model_pred %>%
mutate(errors = Bili - .fitted) %>%
ggplot(., aes(x = errors)) +
geom_histogram(bins = 30, fill = "darkviolet", col = "yellow") + labs(title = "Cross-Validated Errors Predict
x = "Error in predicting Bilirubin")
```



The RMSE and MAE values for the kitchen sink model are only slightly higher than the forward regression model, and the distribution of errors is quite similar. Thus, it was on the basis of degrees of freedom that I chose the forward regression model.

10.4 Validation

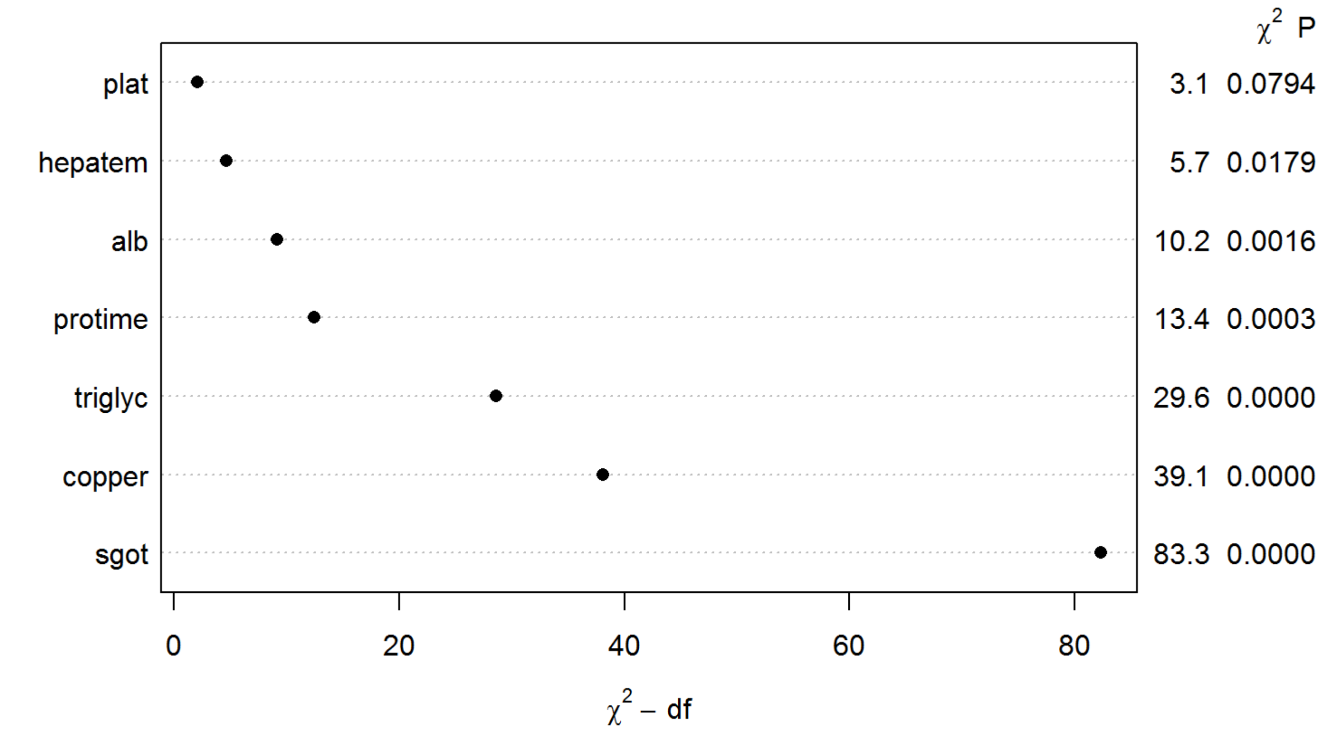
Hide

```
model_fw2ols <- ols(Bili~ sgot + copper + protime + triglyc + alb + hepatem + plat, data = pbc4, x = TRUE, y

validate(model_fw2ols)
```

```
          index.orig training   test optimism index.corrected  n
R-square      0.5634   0.5691 0.5498   0.0193          0.5441 40
MSE           0.4648   0.4562 0.4792  -0.0229          0.4877 40
g             0.8312   0.8333 0.8241   0.0092          0.8220 40
Intercept     0.0000   0.0000 0.0074  -0.0074          0.0074 40
Slope         1.0000   1.0000 0.9981   0.0019          0.9981 40
```

Hide

```
plot(anova(model_fw2ols))
```

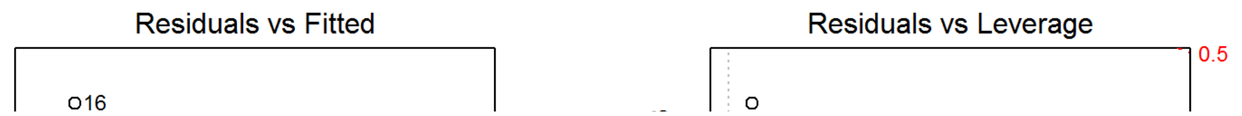


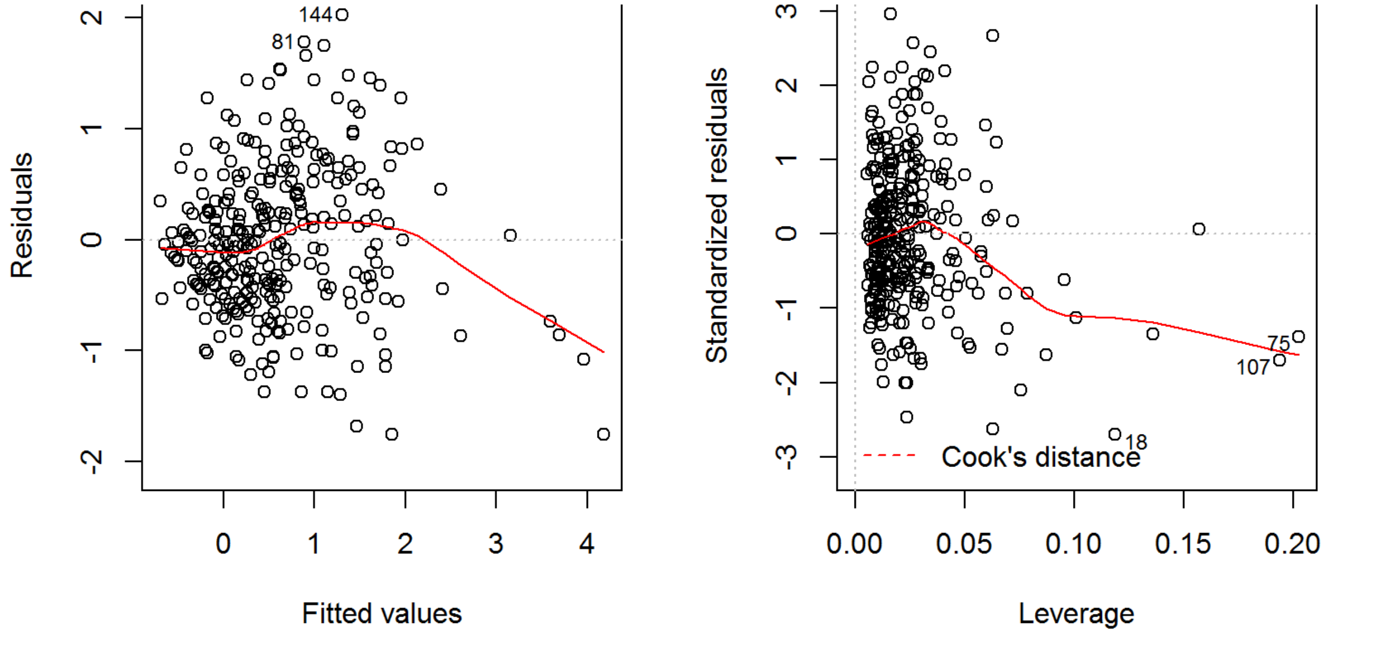
According to the anova here, sgot has the highest predictive power amongst all variables, followed by copper and triglyc.

10.5 Improving the model

Hide

```
par(mfrow = c(1,2)); plot(model_fw2, which = c(1, 5))
```





There were some issues with the outlier values, with some observations going above 2 Residuals, and thus, I decided to remove them in order to see whether there was any increase in the R square value or not, though all of them fell within the Cook's distance. The observations removed were: 144, 67, 18 and 16.

Hide

```
model_fw2del2 <- lm(Bili~ sgot + copper + protime + triglyc + alb + hepatem + plat, data = pbc4[-16,])
summary(model_fw2del2)
```

```

Call:
lm(formula = Bili ~ sgot + copper + protime + triglyc + alb + 
    hepatem + plat, data = pbc4[-16, ])

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7879 -0.4401 -0.0403  0.4286  2.0185 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.4224959  0.6355715  -2.238  0.02594 *  
sgot         0.0067943  0.0007234   9.392  < 2e-16 ***
copper       0.0032617  0.0005085   6.414 5.42e-10 ***
protime      0.1563371  0.0425290   3.676  0.00028 ***
triglyc      0.0036820  0.0006530   5.639 3.93e-08 ***
alb         -0.3262313  0.1002805  -3.253  0.00127 ** 
hepatem      0.2587039  0.1039357   2.489  0.01334 *  
plat        -0.0007075  0.0004267  -1.658  0.09832 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6801 on 303 degrees of freedom
Multiple R-squared:  0.5756,	Adjusted R-squared:  0.5658 
F-statistic:  58.7 on 7 and 303 DF,  p-value: < 2.2e-16
```

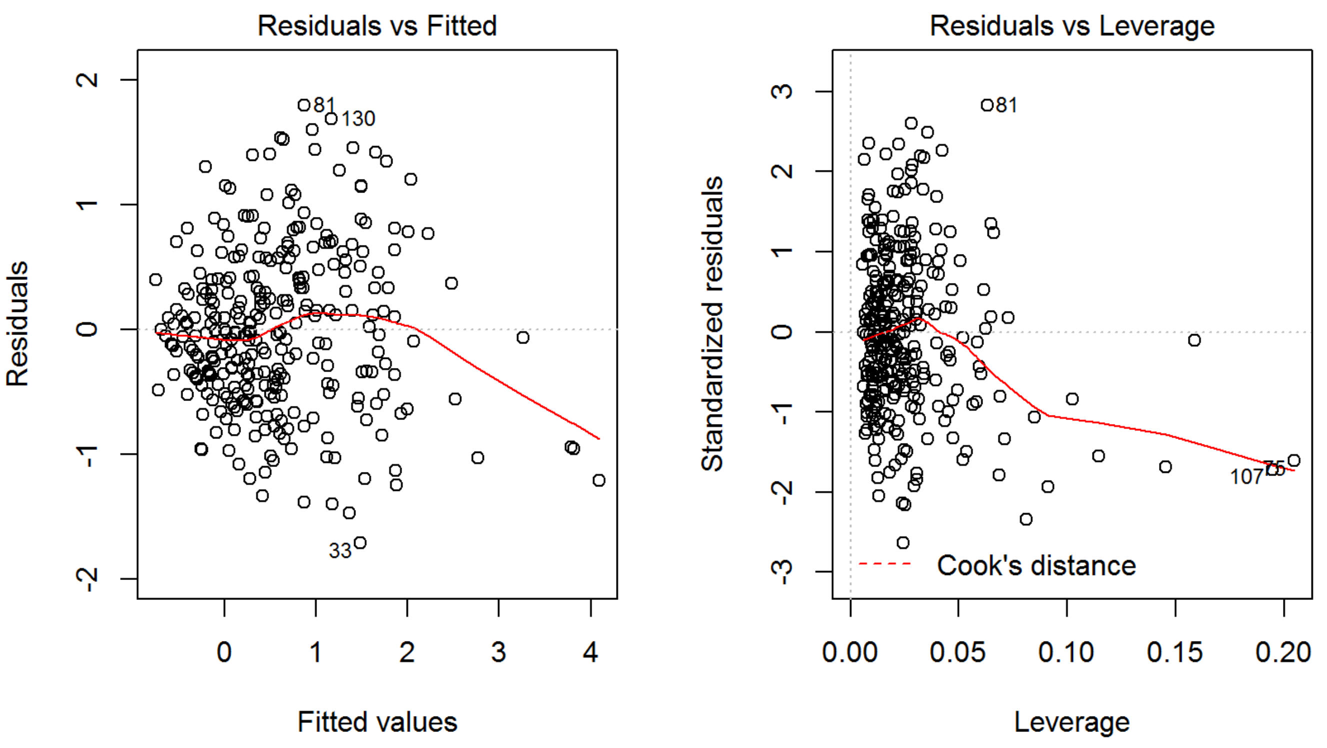
Hide

```
# There was a slight increase in R squared value

model_fw2del3 <- lm(Bili~ sgot + copper + protime + triglyc + alb + hepatem + plat, data = pbc4[-c(144,67,18,
summary(model_fw2del3)
```

```

Call:
lm(formula = Bili ~ sgot + copper + protime + triglyc + alb + 
    hepatem + plat, data = pbc4[-c(144, 67, 18, 16), ])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.71126 -0.44546 -0.03502  0.41381  1.79616 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.6510779  0.6151885  -2.684  0.00768 ** 
sgot         0.0072521  0.0007210  10.058  < 2e-16 ***
copper       0.0035702  0.0005117   6.977 1.94e-11 ***
protime      0.1593541  0.0410947   3.878  0.00013 ***
triglyc      0.0036630  0.0006320   5.796 1.72e-08 ***
alb         -0.2942032  0.0971443  -3.029  0.00267 ** 
hepatem      0.2768070  0.1004704   2.755  0.00623 ** 
plat        -0.0007010  0.0004129  -1.698  0.09059 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6564 on 300 degrees of freedom
Multiple R-squared:  0.5948,	Adjusted R-squared:  0.5853 
F-statistic:  62.9 on 7 and 300 DF,  p-value: < 2.2e-16
```

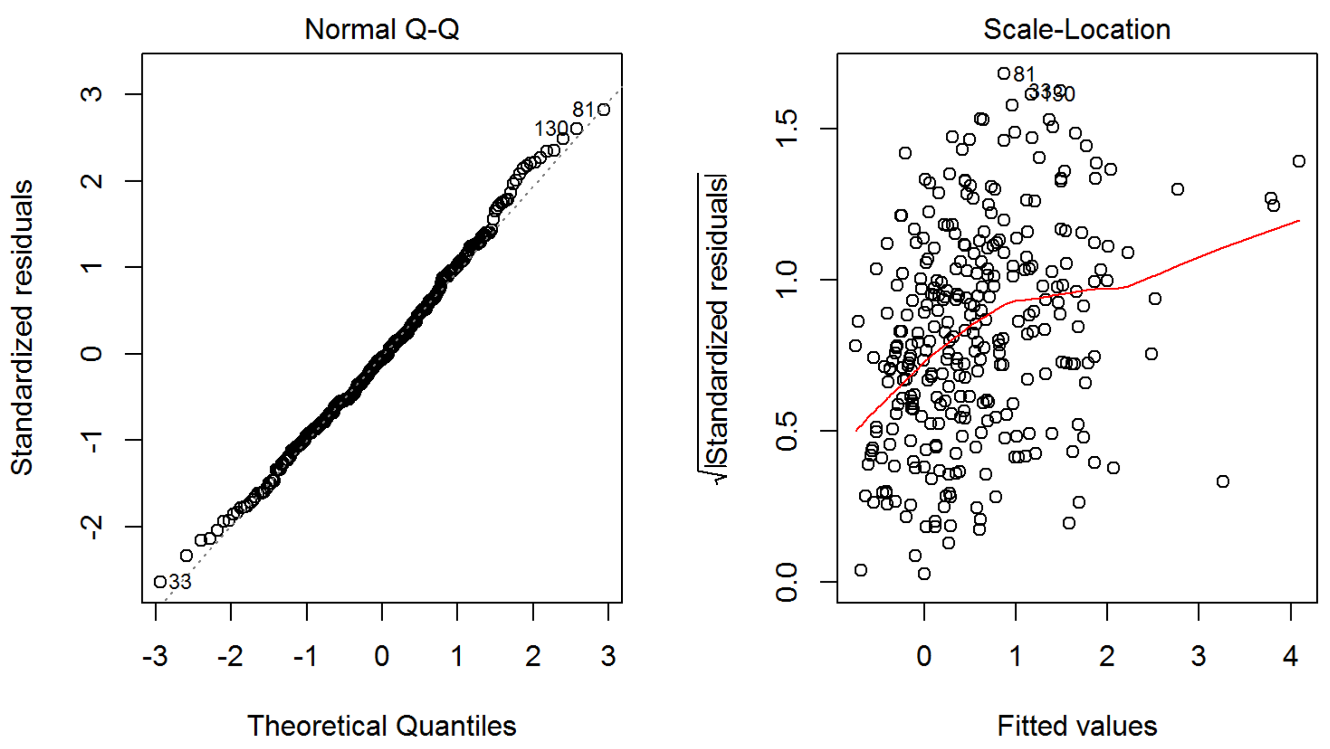
Hide

```
# The R squared value was thus increased by removing the outlier values,

par(mfrow = c(1,2)); plot(model_fw2del3, which = c(1, 5))
```



Hide

```r
par(mfrow = c(1,2)); plot(model_fw2del3, which = c(2, 3))
```



Hide

```r
summary(model_fw2del3)
```

```

Call:
lm(formula = Bili ~ sgot + copper + protime + triglyc + alb + 
    hepatem + plat, data = pbc4[-c(144, 67, 18, 16), ])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.71126 -0.44546 -0.03502  0.41381  1.79616 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.6510779  0.6151885  -2.684  0.00768 ** 
sgot         0.0072521  0.0007210  10.058  < 2e-16 ***
copper       0.0035702  0.0005117   6.977 1.94e-11 ***
protime      0.1593541  0.0410947   3.878  0.00013 ***
triglyc      0.0036630  0.0006320   5.796 1.72e-08 ***
alb         -0.2942032  0.0971443  -3.029  0.00267 ** 
hepatem      0.2768070  0.1004704   2.755  0.00623 ** 
plat        -0.0007010  0.0004129  -1.698  0.09059 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6564 on 300 degrees of freedom
Multiple R-squared:  0.5948,	Adjusted R-squared:  0.5853 
F-statistic:  62.9 on 7 and 300 DF,  p-value: < 2.2e-16
```

Hide

```r
exp(coef(model_fw2del3))
```

```
(Intercept)        sgot      copper     protime     triglyc         alb 
  0.1918430   1.0072785   1.0035766   1.1727532   1.0036697   0.7451251 
    hepatem        plat 
  1.3189118   0.9992993 
```

Hide

```
exp(confint(model_fw2del3))
```

```
                 2.5 %    97.5 %
(Intercept) 0.05717096 0.6437489
sgot        1.00585026 1.0087087
copper      1.00256649 1.0045877
protime     1.08164588 1.2715345
triglyc     1.00242220 1.0049188
alb         0.61546709 0.9020976
hepatem     1.08230289 1.6072473
plat        0.99848766 1.0001115
```

10.6 Predictions

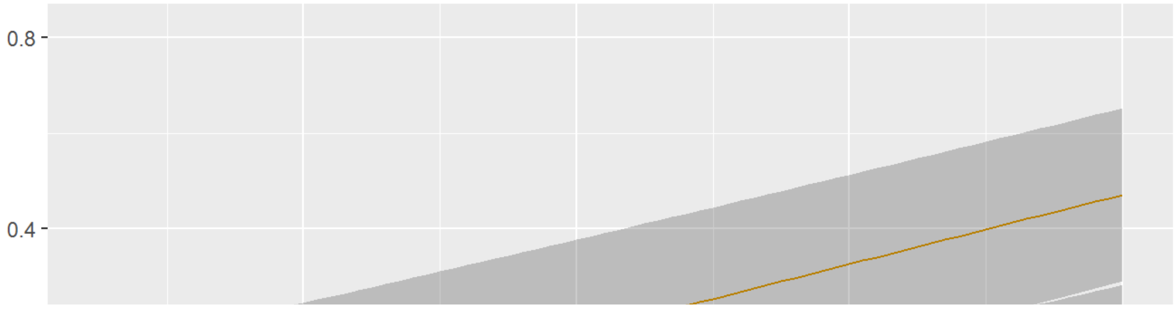
Hide

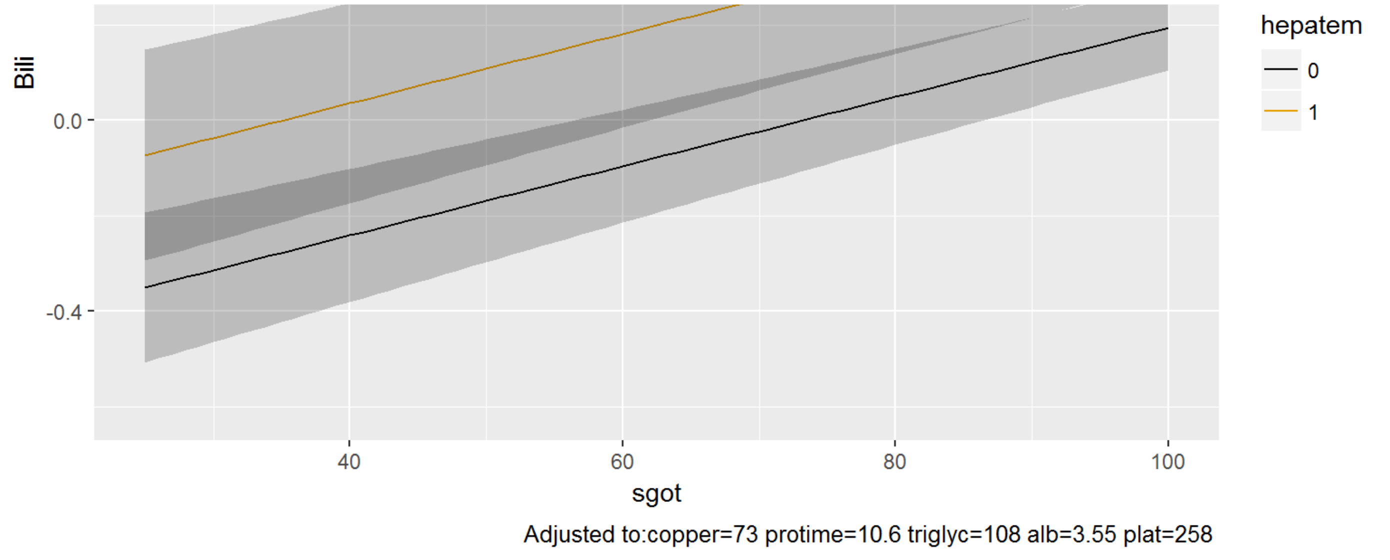
```
p <- datadist(pbc4)
options(datadist = "p")
model_fw2del3ols <- ols(Bili~ sgot + copper + protime + triglyc + alb + hepatem + plat, data = pbc4[-c(144,67

predictions <- Predict(model_fw2del3ols, hepatem = c(0,1), sgot = seq(25, 100) )
tbl_df(predictions)
```

```
# A tibble: 152 x 10
    sgot copper protime triglyc   alb hepatem  plat   yhat  lower  upper
   <int>  <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
 1    25   73.0    10.6     108  3.55       0   258 -0.350 -0.506 -0.193
 2    26   73.0    10.6     108  3.55       0   258 -0.342 -0.498 -0.187
 3    27   73.0    10.6     108  3.55       0   258 -0.335 -0.490 -0.181
 4    28   73.0    10.6     108  3.55       0   258 -0.328 -0.481 -0.175
 5    29   73.0    10.6     108  3.55       0   258 -0.321 -0.473 -0.169
 6    30   73.0    10.6     108  3.55       0   258 -0.313 -0.464 -0.163
 7    31   73.0    10.6     108  3.55       0   258 -0.306 -0.456 -0.157
 8    32   73.0    10.6     108  3.55       0   258 -0.299 -0.447 -0.150
 9    33   73.0    10.6     108  3.55       0   258 -0.292 -0.439 -0.144
10    34   73.0    10.6     108  3.55       0   258 -0.284 -0.431 -0.138
# ... with 142 more rows
```

Hide

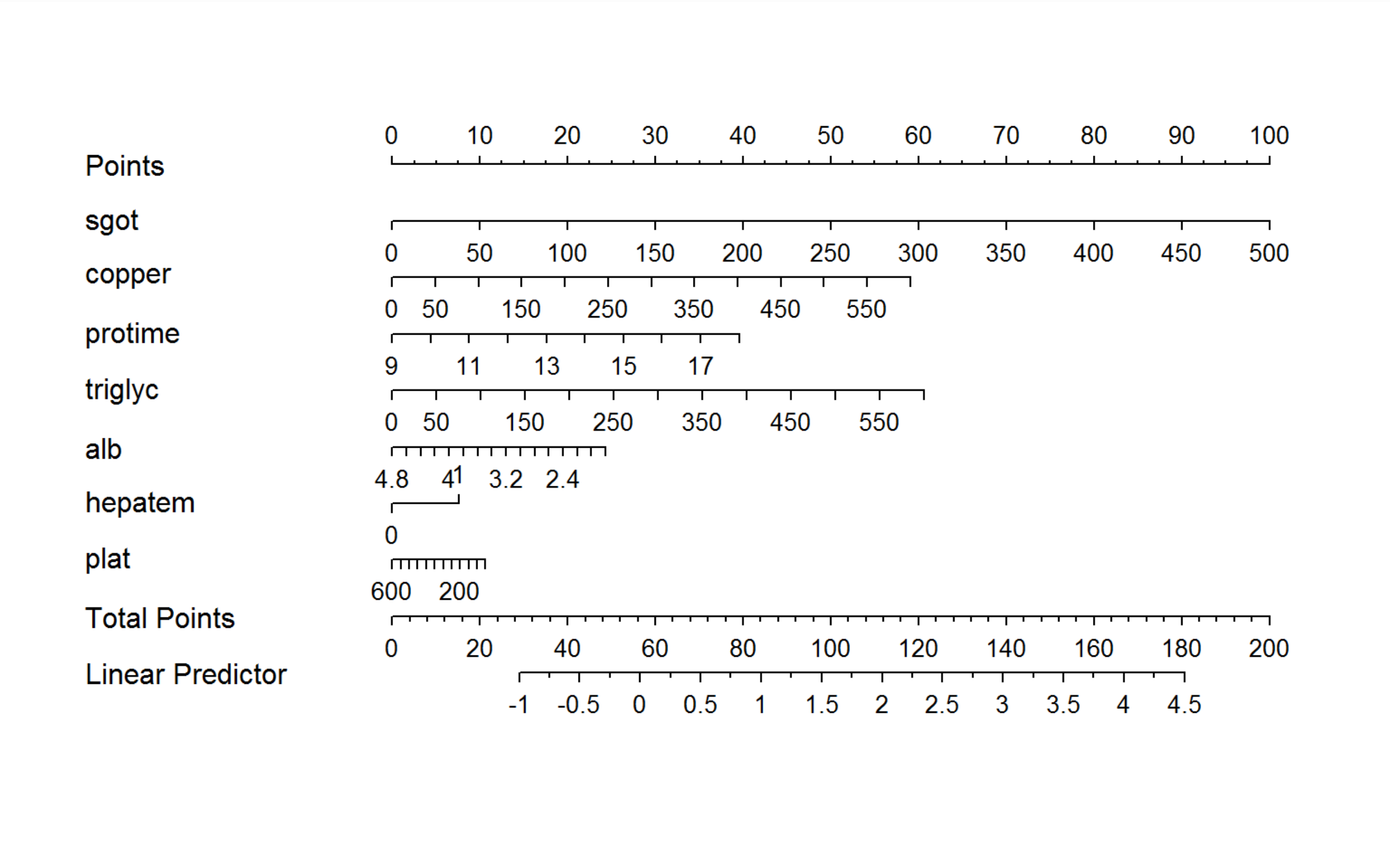
```
ggplot(Predict(model_fw2del3ols, sgot = 25:100, hepatem = c(0,1)))
```





Hide

```
plot(nomogram(model_fw2del3ols))
```



Here, sgot has the highest impact on the prediction of Bilirubin concentrations, followed by triglyc and copper.

10.7 Final Model

Hide

```
model_fw2del3ols
```

```
Linear Regression Model

 ols(formula = Bili ~ sgot + copper + protime + triglyc + alb +
     hepatem + plat, data = pbc4[-c(144, 67, 18, 16), ], x = TRUE,
     y = TRUE)
```

```
                Model Likelihood     Discrimination
                   Ratio Test           Indexes
 Obs     308    LR chi2    278.20    R2       0.595
 sigma0.6564    d.f.            7    R2 adj   0.585
 d.f.    300    Pr(> chi2) 0.0000    g        0.852

 Residuals

      Min       1Q   Median       3Q      Max
 -1.71126 -0.44546 -0.03502  0.41381  1.79616


           Coef    S.E.   t     Pr(>|t|)
 Intercept -1.6511 0.6152 -2.68 0.0077
 sgot       0.0073 0.0007 10.06 <0.0001
 copper     0.0036 0.0005  6.98 <0.0001
 protime    0.1594 0.0411  3.88 0.0001
 triglyc    0.0037 0.0006  5.80 <0.0001
 alb       -0.2942 0.0971 -3.03 0.0027
 hepatem    0.2768 0.1005  2.76 0.0062
 plat      -0.0007 0.0004 -1.70 0.0906
```

Hide

```
summary(model_fw2del3ols)
```

```
             Effects              Response : Bili

 Factor  Low    High  Diff.  Effect    S.E.     Lower 0.95 Upper 0.95
 sgot     80.60 151.9  71.30  0.517080 0.051409  0.415910   0.618240
 copper   41.75 123.0  81.25  0.290080 0.041578  0.208260   0.371900
 protime  10.00  11.1   1.10  0.175290 0.045204  0.086332   0.264250
 triglyc  84.00 146.0  62.00  0.227110 0.039184  0.149990   0.304220
 alb       3.31   3.8   0.49 -0.144160 0.047601 -0.237830  -0.050486
 hepatem   0.00   1.0   1.00  0.276810 0.100470  0.079091   0.474520
 plat    199.75 322.5 122.75 -0.086044 0.050681 -0.185780   0.013691
```

Hide

```
exp(confint(model_fw2del3ols))
```

```
               2.5 %    97.5 %
Intercept 0.05717096 0.6437489
sgot      1.00585026 1.0087087
copper    1.00256649 1.0045877
protime   1.08164588 1.2715345
triglyc   1.00242220 1.0049188
alb       0.61546709 0.9020976
hepatem   1.08230289 1.6072473
plat      0.99848766 1.0001115
```

Hide

```
exp(coef(model_fw2del3ols))
```

```
Intercept      sgot    copper   protime   triglyc       alb   hepatem 
0.1918430 1.0072785 1.0035766 1.1727532 1.0036697 0.7451251 1.3189118 
     plat 
0.9992993 
```

The final model obtained is: log(Bilirubin) = -1.65 + 0.0073(sgot) + 0.0036(copper) + 0.16(protime) + 0.003(triglyc) - 0.294(alb) + 0.27(hepatem) - 0.007(plat) The adjusted R squared value is 0.585, implying that 58.5 % of the variance is explained by this transformed model.

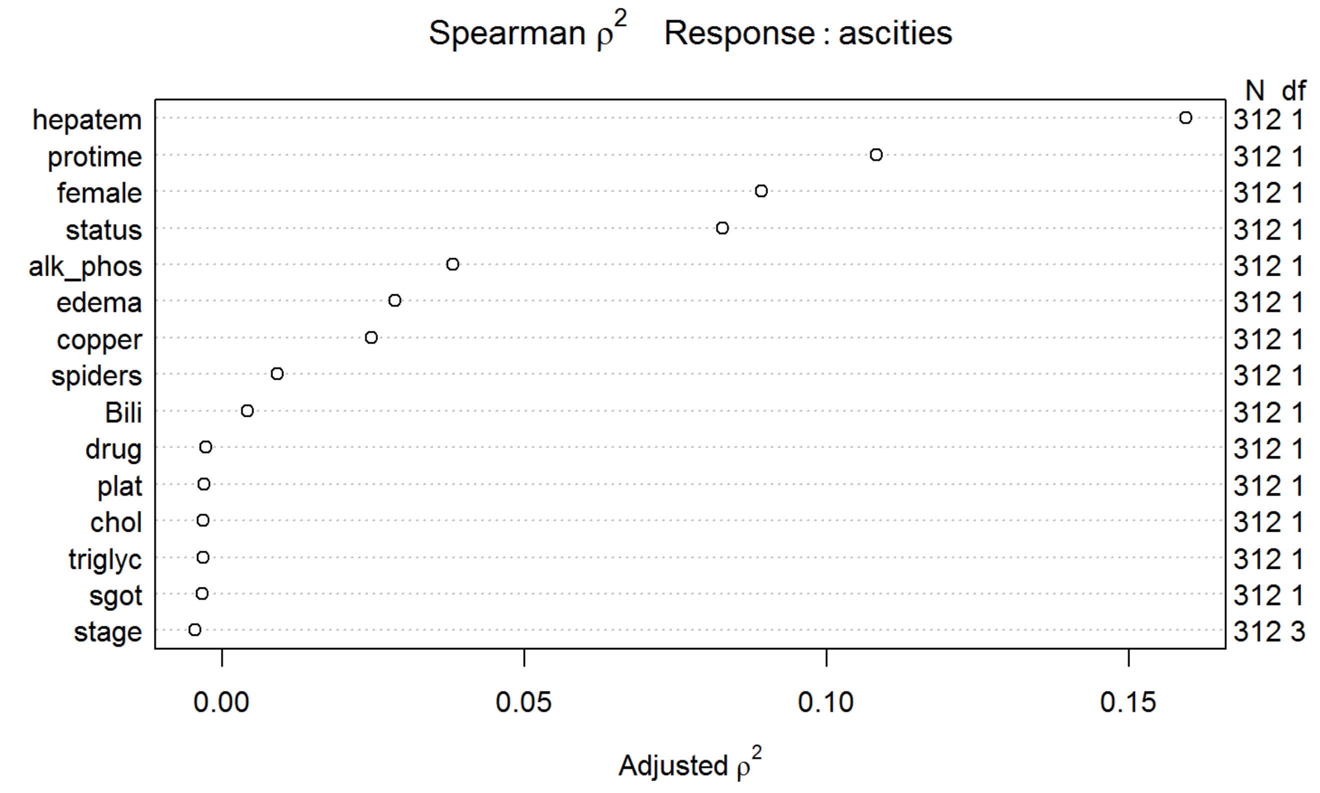
For every 1 increase in the log Bili value, The sgot, copper, protime, triglyc and hepatem values are going to increase, while albumin and plat values are supposed to go down. Those who have had hepatem had a significant increase in the transformed Bili concentrations of about 1.319. The 95% CI was (1.08, 1.60) As the the log Bili concentrations go up by 1mg/dl, the triglyc, plat, sgot and copper values are increased by almost 1 mg/dl, 1 cubic ml/1000, 1 U/ml and 1 ug/day respectively, (The 95% C.I. of (1.002, 1.004), (0.99, 1.00), (1.005, 1.008) and (1.002, 1.004) respectively) For a unit increase in log Bili, the albumin concentrations go down, and there is a slight increase in the protime (by 1.17 seconds). (95% CI of (0.61, 0.9) and (1.08,1.27) respectively).

11 Task 11 Logistic Regression

11.1 Spearman Rho Squared

Hide

```
spear.ascities <- spearman2(ascities ~ alk_phos + protime + triglyc + chol + female + copper + sgot + plat +

plot(spear.ascities)
```



On the basis of the spearman Rho squared plot, I decided to go ahead with the first 5 variables, since I had a small number of observations and limited degrees of freedom to spend. I then made a kitchen sink model using these 5 predictors.

11.2 Kitchen Sink Model

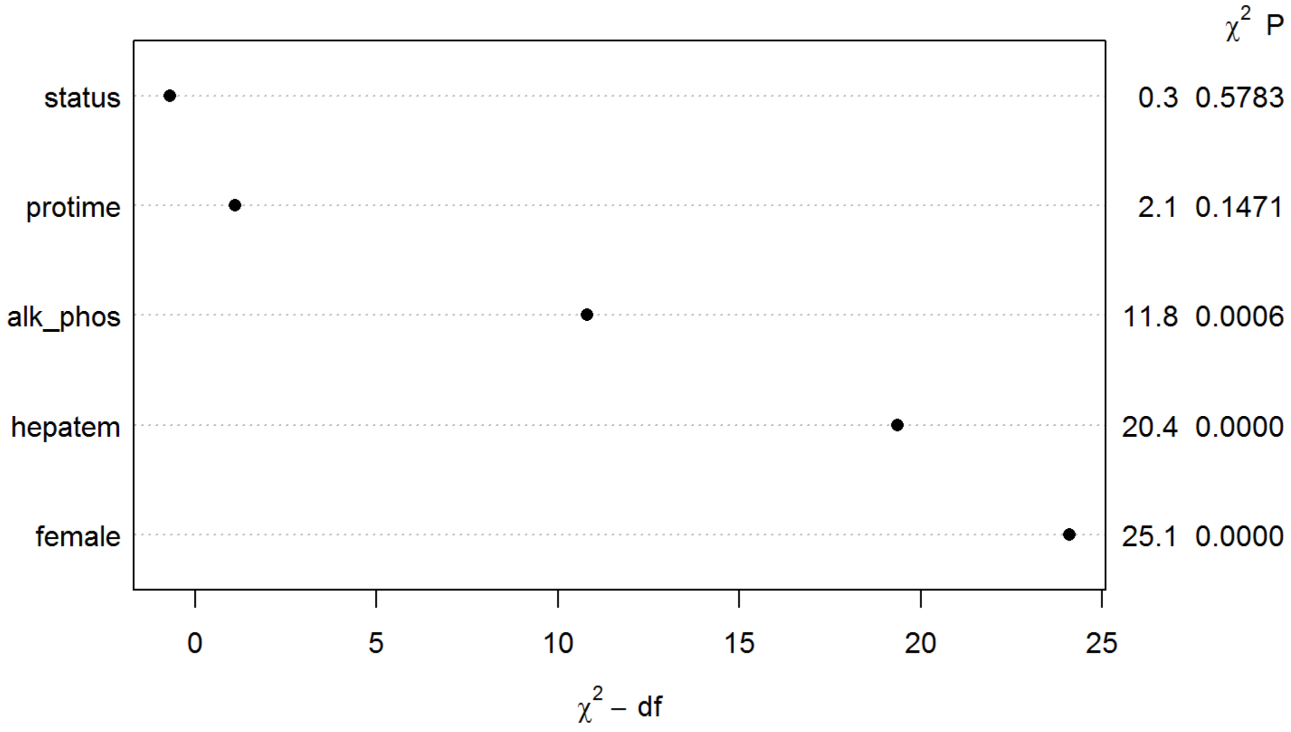
Hide

```
logmodel_ks_lrm <- lrm(ascities~ hepatem + protime + female + status + alk_phos, data = pbc2, x = T, y = T)
anova(logmodel_ks_lrm)
```

```
                Wald Statistics          Response: ascities 

 Factor     Chi-Square d.f. P     
 hepatem    20.37      1    <.0001
 protime     2.10      1    0.1471
 female     25.11      1    <.0001
 status      0.31      1    0.5783
 alk_phos   11.82      1    0.0006
 TOTAL      69.37      5    <.0001
```

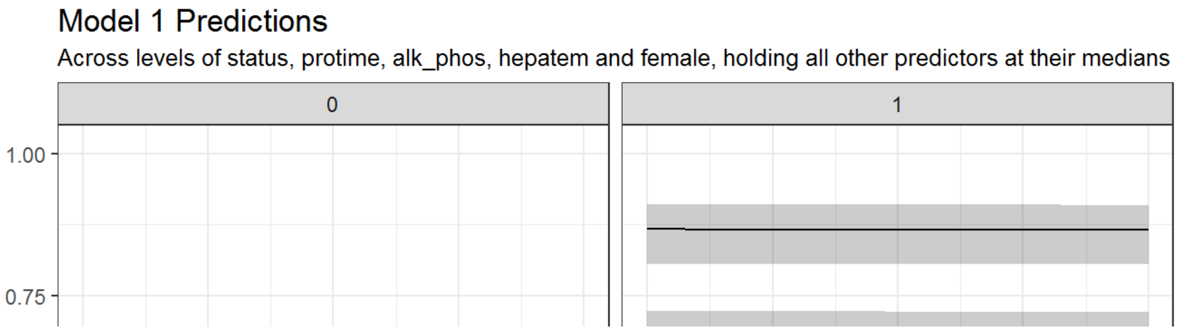
Hide

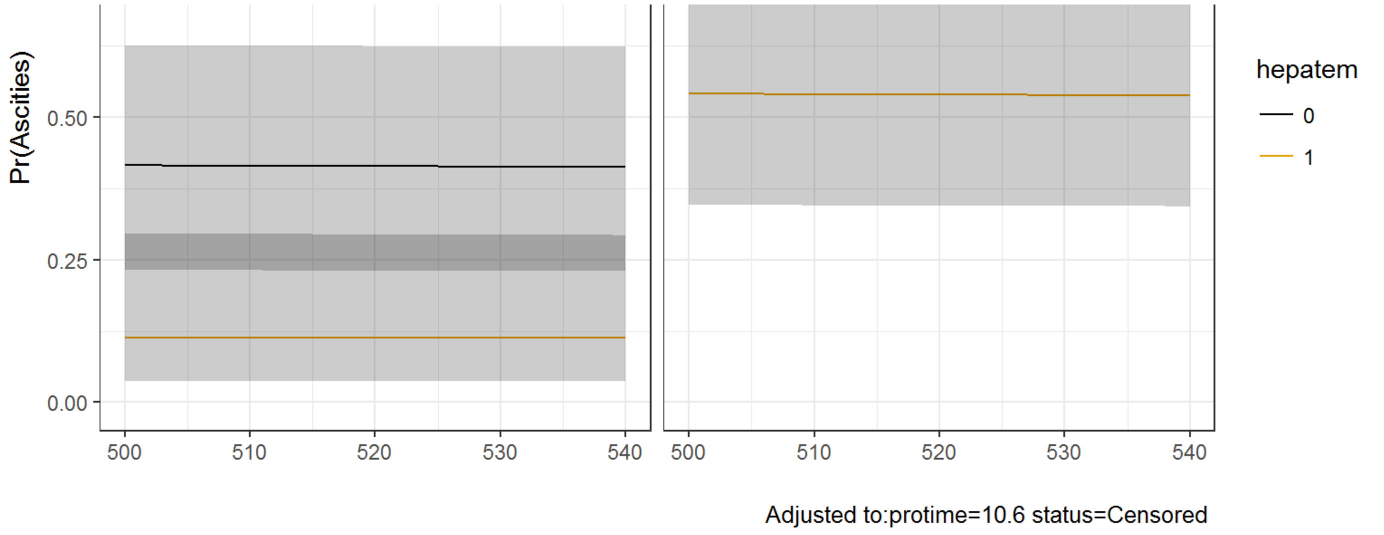
```
plot(anova(logmodel_ks_lrm))
```



Hide

```
ggplot(Predict(logmodel_ks_lrm, alk_phos = 500:540, hepatem = c(0,1), female =c(0,1), fun=plogis)) +
  theme_bw() +
  labs(x = "",
  y = "Pr(Ascities)", title = "Model 1 Predictions", subtitle = "Across levels of status, protime, alk_phos,
```





Adjusted to:protime=10.6 status=Censored

Hide

```
# Making a glm model for the same variables

logmodel_ks_glm <- glm(ascities~ hepatem + protime + female + status + alk_phos, family = binomial, data = pbc
summary(logmodel_ks_glm)
```

```


Call:
glm(formula = ascities ~ hepatem + protime + female + status + 
    alk_phos, family = binomial, data = pbc2)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0373  -0.7899   0.5377   0.6379   2.2283  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.135e+00  1.652e+00   1.292 0.196330    
hepatem     -1.712e+00  3.792e-01  -4.513 6.39e-06 ***
protime     -2.198e-01  1.516e-01  -1.450 0.147130    
female       2.217e+00  4.425e-01   5.011 5.43e-07 ***
statusDeath -1.872e-01  3.368e-01  -0.556 0.578332    
alk_phos    -2.947e-04  8.573e-05  -3.438 0.000586 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 403.76  on 311  degrees of freedom
Residual deviance: 301.96  on 306  degrees of freedom
AIC: 313.96

Number of Fisher Scoring iterations: 4
```

Hide

```
anova(logmodel_ks_glm)
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: ascities
```

```
Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev
NULL                       311     403.76
hepatem   1   48.860       310     354.90
protime   1    4.252       309     350.65
female    1   31.804       308     318.84
status    1    1.112       307     317.73
alk_phos  1   15.770       306     301.96
```

The Kitchen sink model shows that while female, hepatem and alk_phos significantly affect the prediction ability for ascities, status and protime do not appear to do so. I thus did a stepwise backward regression to see if the number of variables could be brought down.

11.3 Stepwise backward regression

Hide

```
step(logmodel_ks_glm)
```

```
Start:  AIC=313.96
ascities ~ hepatem + protime + female + status + alk_phos

           Df Deviance    AIC
- status    1   302.27 312.27
<none>          301.96 313.96
- protime   1   304.01 314.01
- alk_phos  1   317.73 327.73
- hepatem   1   323.63 333.63
- female    1   330.65 340.65

Step:  AIC=312.27
ascities ~ hepatem + protime + female + alk_phos

           Df Deviance    AIC
<none>          302.27 312.27
- protime   1   305.05 313.05
- alk_phos  1   318.84 326.84
- hepatem   1   329.00 337.00
- female    1   331.94 339.94
```

```

Call:  glm(formula = ascities ~ hepatem + protime + female + alk_phos, 
    family = binomial, data = pbc2)

Coefficients:
(Intercept)      hepatem      protime       female     alk_phos  
  2.3288571   -1.7884439   -0.2448006    2.2435730   -0.0003008  

Degrees of Freedom: 311 Total (i.e. Null);  307 Residual
Null Deviance:      403.8 
Residual Deviance: 302.3    AIC: 312.3
```

The stepwise regression gave the following variables for this model: ascities ~hepatem + protime + female + alk_phos

Hide

```
logmodel_ks_lrm2 <- lrm(ascities ~hepatem + protime + female + alk_phos, data = pbc2, x = T, y = T)

# Mkaing a glm model of the same
# Mkaing a glm model of the same

logmodel_ks_glm2 <- glm(ascities ~ hepatem + protime + female + alk_phos, family = binomial, data = pbc2)
```

11.4 Comparisons

11.4.1 Anova Comparison

Hide

```
anova(logmodel_ks_glm, logmodel_ks_glm2)
```

```
Analysis of Deviance Table

Model 1: ascities ~ hepatem + protime + female + status + alk_phos
Model 2: ascities ~ hepatem + protime + female + alk_phos
  Resid. Df Resid. Dev Df Deviance
1       306     301.96            
2       307     302.27 -1 -0.30583
```

On the basis of anova, I would say that model 1 is slghtly better, but uses more degrees of freedom.

11.4.2 AIC/BIC Comparison

Hide

```
glance(logmodel_ks_glm)
```

```
  null.deviance df.null    logLik      AIC      BIC deviance df.residual
1      403.7585     311 -150.9803 313.9606 336.4186 301.9606         306
```

Hide

```
glance(logmodel_ks_glm2)
```

```
  null.deviance df.null    logLik      AIC      BIC deviance df.residual
1      403.7585     311 -151.1332 312.2664 330.9814 302.2664         307
```

The AIC and BIC values have clearly gone down for model 2.

11.4.3 ROC Comparison

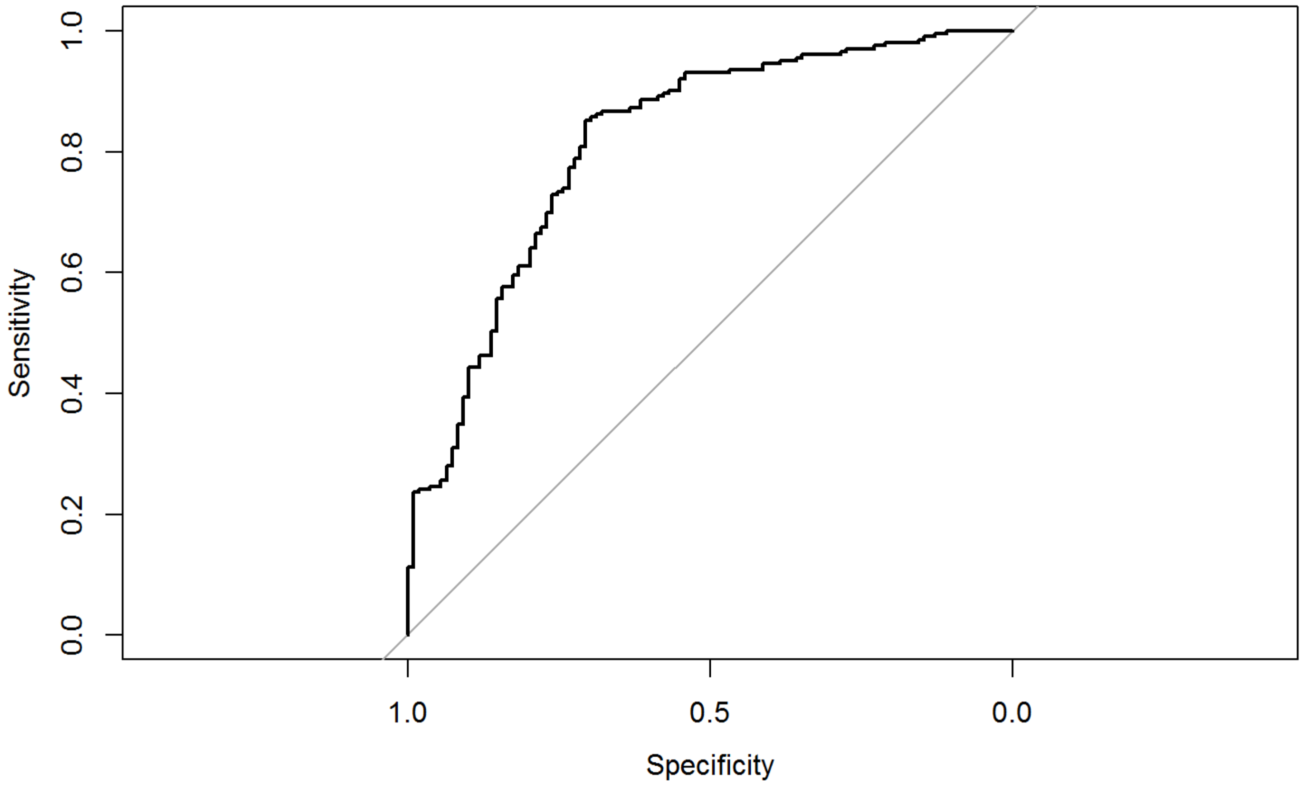
Hide

```
roc_model_ks_glm <- roc(pbc2$ascities ~ predict(logmodel_ks_glm, type = "response"), ci = TRUE)
roc_model_ks_glm
```

```


Call:
roc.formula(formula = pbc2$ascities ~ predict(logmodel_ks_glm,     type = "response"), ci = TRUE)

Data: predict(logmodel_ks_glm, type = "response") in 109 controls (pbc2$ascities 0) < 203 cases (pbc2$ascities
Area under the curve: 0.8153
95% CI: 0.7643-0.8664 (DeLong)
```

Hide

```
plot(roc_model_ks_glm)
```



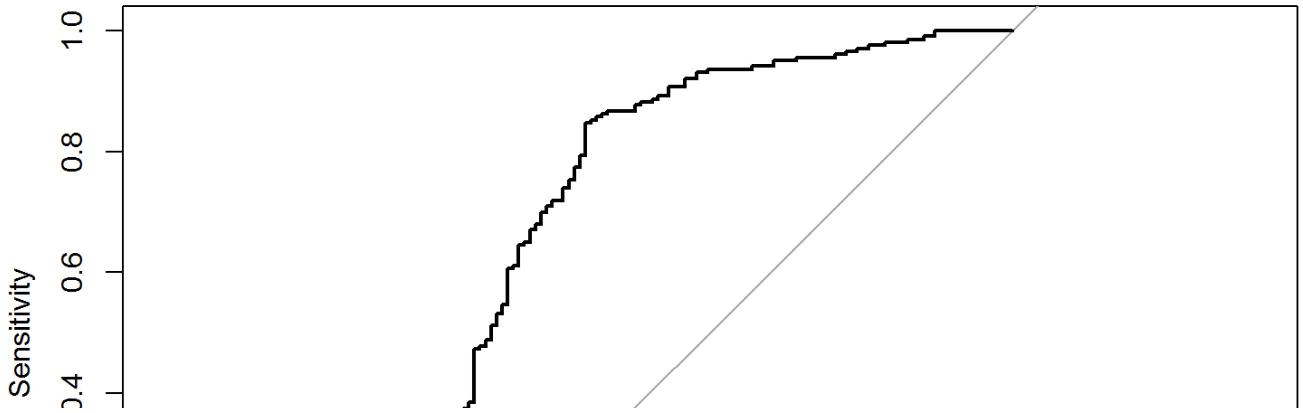
Hide

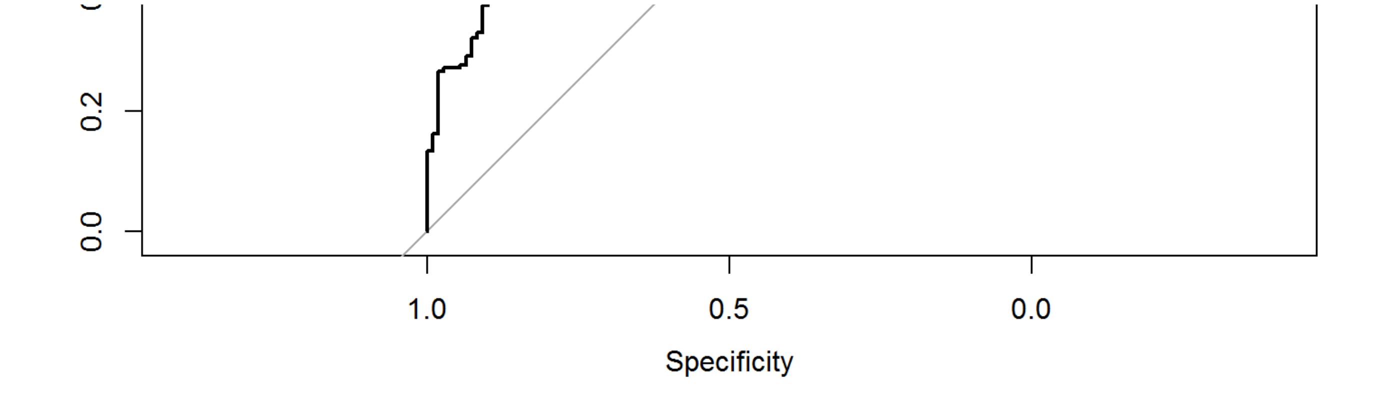
```
roc_model_ks_glm2 <- roc(pbc2$ascities ~ predict(logmodel_ks_glm2, type = "response"), ci = TRUE)
roc_model_ks_glm2
```

```

Call:
roc.formula(formula = pbc2$ascities ~ predict(logmodel_ks_glm2,     type = "response"), ci = TRUE)

Data: predict(logmodel_ks_glm2, type = "response") in 109 controls (pbc2$ascities 0) < 203 cases (pbc2$ascitie
Area under the curve: 0.8161
95% CI: 0.7653-0.8669 (DeLong)
```

Hide

```
plot(roc_model_ks_glm2)
```



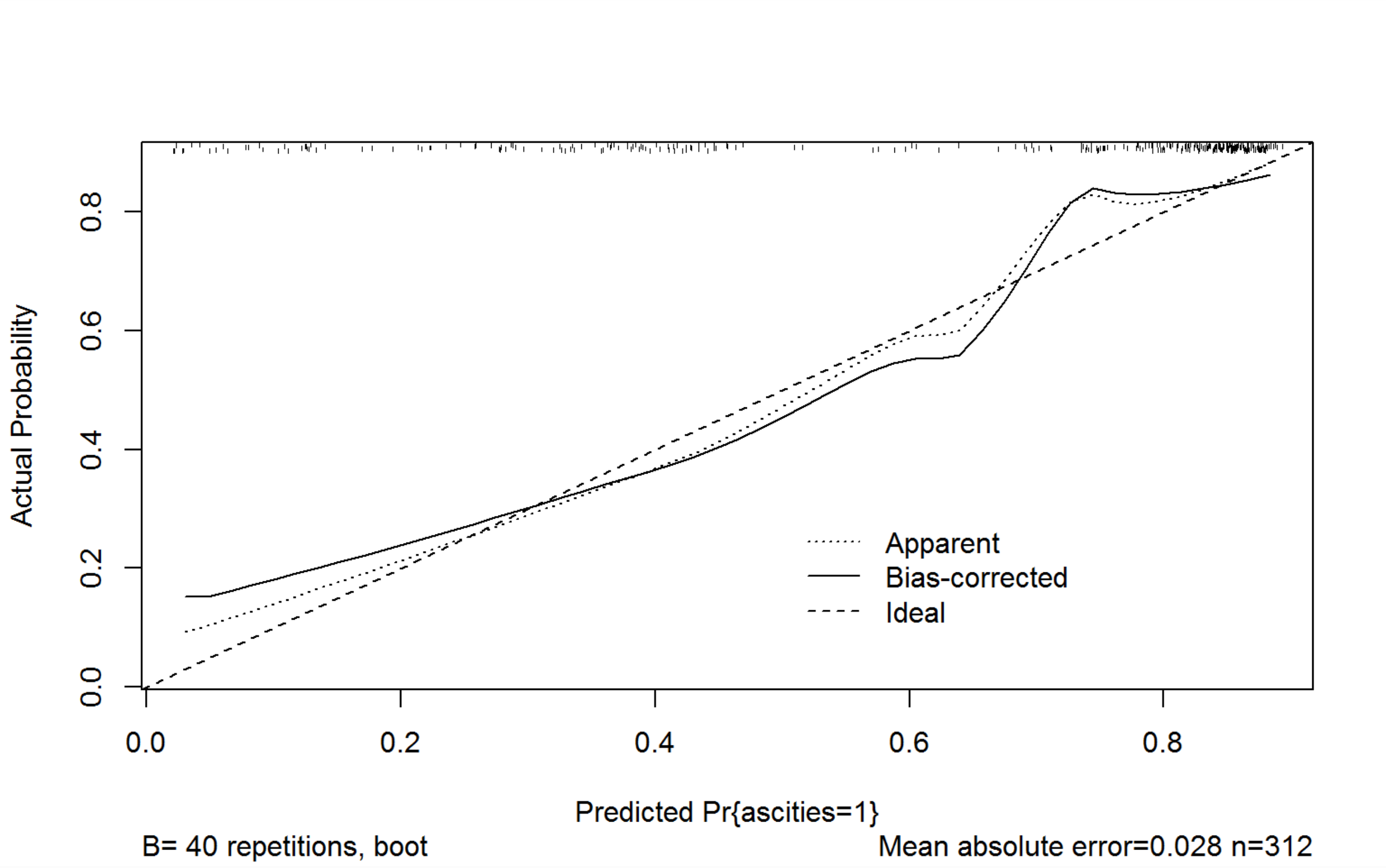


The ROC values are not very different, and model 2 has slightly higher ROC value (0.8161 and 0.8153 for model 2 and model 1 respectively).

11.4.4 Calibration

Hide

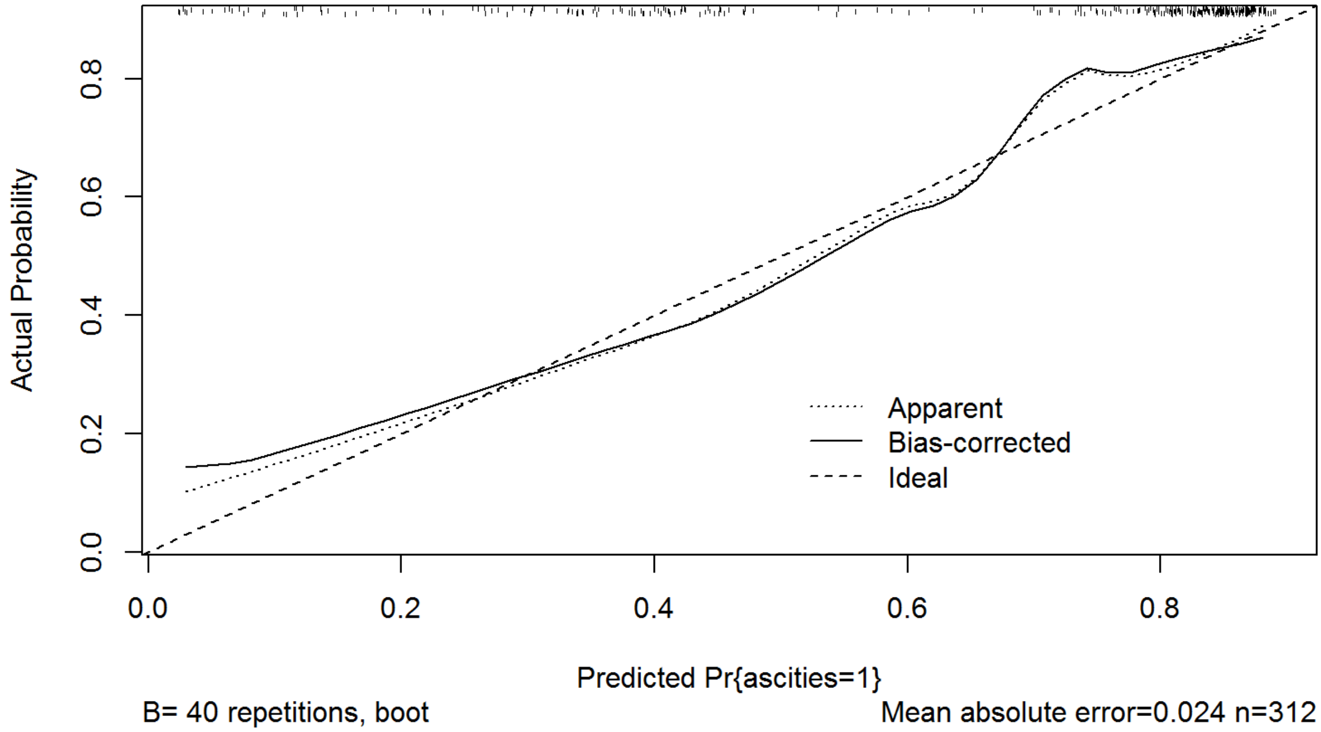
```
plot(calibrate(logmodel_ks_lrm))
```



```
n=312   Mean absolute error=0.028   Mean squared error=0.00154
0.9 Quantile of absolute error=0.073
```

Hide

```
plot(calibrate(logmodel_ks_lrm2))
```



```

n=312   Mean absolute error=0.024   Mean squared error=0.00108
0.9 Quantile of absolute error=0.058
```

The calibration plot for both the models isn’t great.The bias corrected line is both above and below the ideal line, and there are problems in predictions if the predicted values go up. Both the graphs, however, are similar.

11.4.5 Validation

Hide

```
validate(logmodel_ks_lrm)
```

```
          index.orig training   test optimism index.corrected  n
Dxy           0.6307   0.6521 0.6215   0.0306          0.6001 40
R2            0.3835   0.4091 0.3714   0.0377          0.3459 40
Intercept     0.0000   0.0000 0.0058  -0.0058          0.0058 40
Slope         1.0000   1.0000 0.9340   0.0660          0.9340 40
Emax          0.0000   0.0000 0.0162   0.0162          0.0162 40
D             0.3231   0.3493 0.3110   0.0382          0.2848 40
U            -0.0064  -0.0064 0.0026  -0.0090          0.0026 40
Q             0.3295   0.3557 0.3085   0.0473          0.2822 40
B             0.1554   0.1485 0.1589  -0.0104          0.1658 40
g             1.4688   1.5637 1.4296   0.1342          1.3346 40
gp            0.2789   0.2865 0.2739   0.0126          0.2663 40
```

Hide

```
validate(logmodel_ks_lrm2)
```

```
          index.orig training   test optimism index.corrected  n
Dxy           0.6322   0.6344 0.6247   0.0097          0.6225 40
R2            0.3826   0.4012 0.3737   0.0275          0.3550 40
Intercept     0.0000   0.0000 0.0406  -0.0406          0.0406 40
Slope         1.0000   1.0000 0.9445   0.0555          0.9445 40
Emax          0.0000   0.0000 0.0196   0.0196          0.0196 40
D             0.3221   0.3417 0.3132   0.0285          0.2936 40
```

```
U              -0.0064   -0.0064 0.0006   -0.0070          0.0006 40
Q               0.3285    0.3481 0.3126    0.0355          0.2930 40
B               0.1557    0.1510 0.1582   -0.0073          0.1629 40
g               1.4585    1.5231 1.4274    0.0958          1.3627 40
gp              0.2772    0.2823 0.2733    0.0090          0.2681 40
```

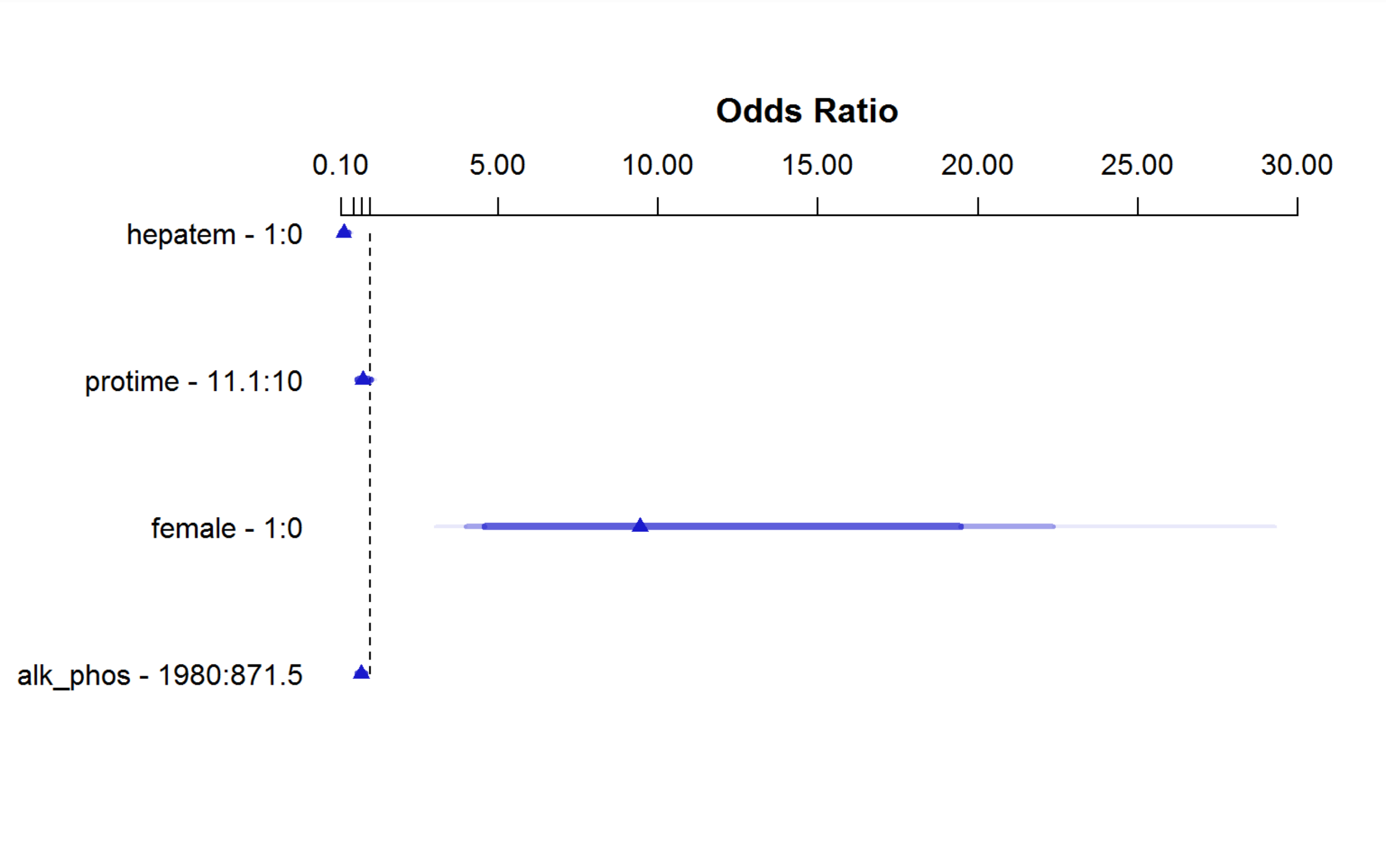
The C-statistic for MOdel 1 is: 0.5 + (0.6307/2) = 0.81535 The C-statistic for model 2 is: 0.5 + (0.6322/2) = 0.8161

Hence, based on all these factors, and the fact that model 2 is easier and spends lesser degrees of freedom, I have decided to go forward with model 2.

11.5 Plots

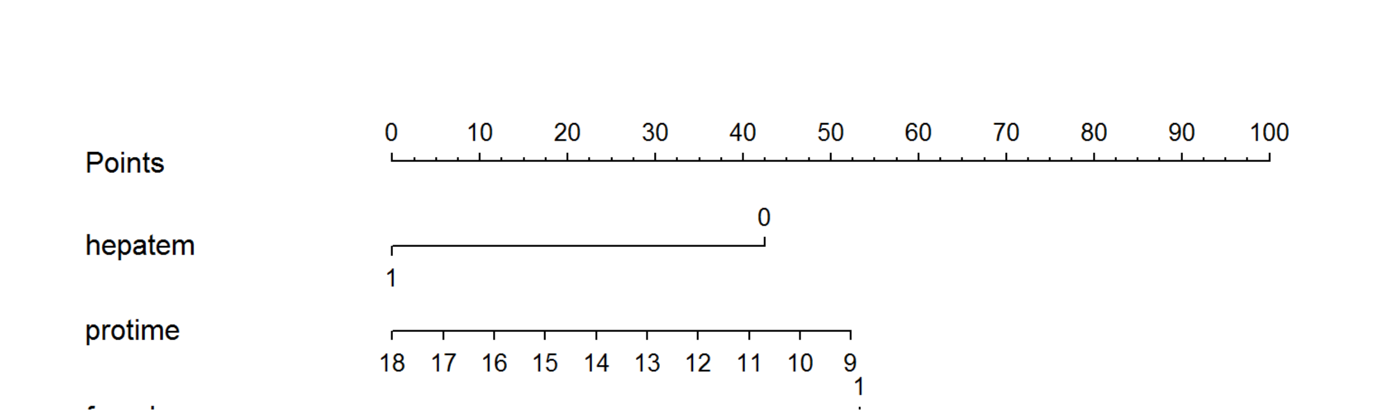
Hide

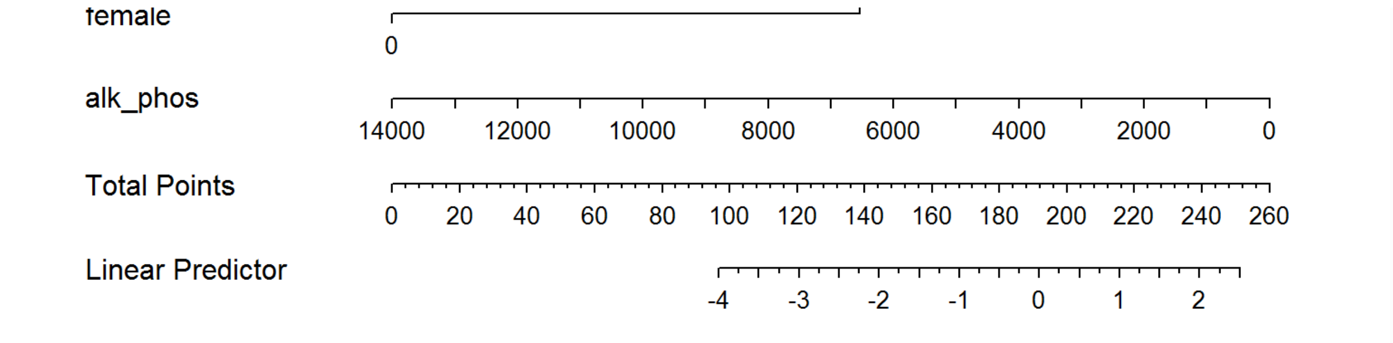
```
plot(summary(logmodel_ks_lrm2))
```

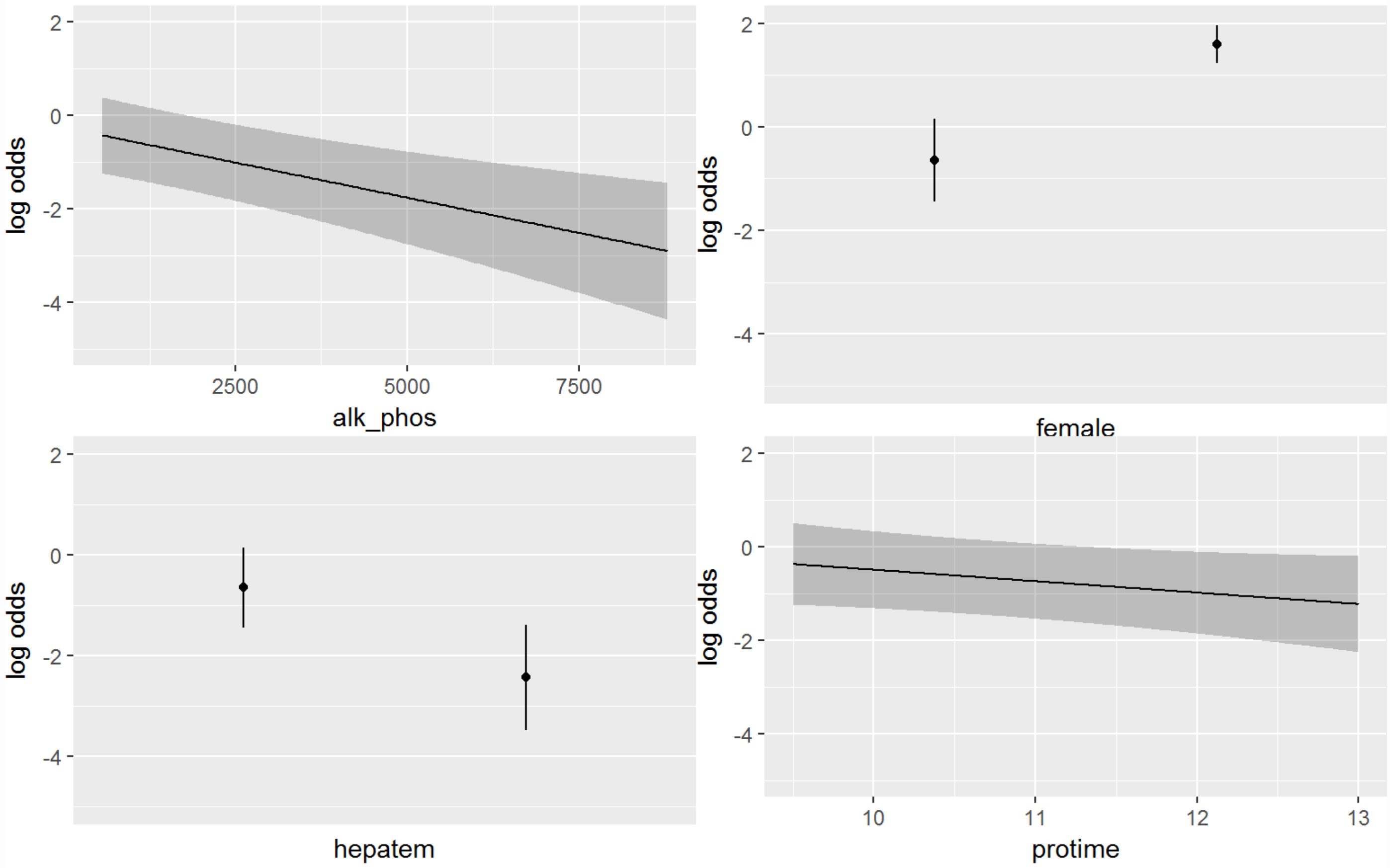


Hide

```
plot(nomogram(logmodel_ks_lrm2))
ggplot(Predict(logmodel_ks_lrm2))
```







Here, we can see that female has a major impact on the odds ratio, and alk_phos is quite the important predictor, as shown by the nomogram, followed by female and protime.

11.6 Odds Ratio and Confidence Interval

Hide

```
logmodel_ks_lrm2
```

```
Logistic Regression Model

 lrm(formula = ascities ~ hepatem + protime + female + alk_phos,
     data = pbc2, x = T, y = T)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs            312    LR chi2      101.49    R2        0.383    C       0.816
  0             109    d.f.              4    g         1.458    Dxy     0.632
  1             203    Pr(> chi2) <0.0001    gr        4.299    gamma   0.632
 max |deriv| 2e-08                           gp        0.277    tau-a   0.288
                                             Brier     0.156
```

```
          Coef    S.E.   Wald Z Pr(>|Z|)
Intercept  2.3289 1.6167  1.44  0.1497  
hepatem   -1.7884 0.3549 -5.04  <0.0001 
protime   -0.2448 0.1449 -1.69  0.0912  
female     2.2436 0.4404  5.09  <0.0001 
alk_phos  -0.0003 0.0001 -3.52  0.0004  
```

Hide

```
summary(logmodel_ks_lrm2)
```

```
             Effects              Response : ascities 

Factor      Low   High   Diff.  Effect   S.E.     Lower 0.95 Upper 0.95
hepatem       0.0    1.0    1.0 -1.78840 0.354890 -2.484000  -1.092900 
 Odds Ratio   0.0    1.0    1.0  0.16722       NA  0.083407   0.335250 
protime      10.0   11.1    1.1 -0.26928 0.159430 -0.581750   0.043191 
 Odds Ratio  10.0   11.1    1.1  0.76393       NA  0.558920   1.044100 
female        0.0    1.0    1.0  2.24360 0.440380  1.380400   3.106700 
 Odds Ratio   0.0    1.0    1.0  9.42700       NA  3.976700  22.347000 
alk_phos    871.5 1980.0 1108.5 -0.33342 0.094854 -0.519330  -0.147500 
 Odds Ratio 871.5 1980.0 1108.5  0.71647       NA  0.594920   0.862860 
```

Hide

```
exp(coef(logmodel_ks_glm2))
```

```
(Intercept)     hepatem     protime      female    alk_phos 
 10.2662011   0.1672202   0.7828606   9.4269537   0.9996993 
```

Hide

```
exp(confint(logmodel_ks_glm2))
```

```
                 2.5 %      97.5 %
(Intercept) 0.40802122 250.6897071
hepatem     0.08206123   0.3316688
protime     0.58701942   1.0449249
female      4.09925808  23.3925816
alk_phos    0.99951465   0.9998535
```

The final equation for the model is: Log odds of Ascities happening = 2.32 - 1.7(hepatem) - 0.24(protime) + 2.24(female) - 0.0003(alk_phos)

The odds ratio indicate that: Females had more odds (9.42 times) of having ascities as compared to males.The 95% CI was (4.099, 23.39) If a person had hepatem (hepatem=1), they had lesser odds of developing ascities (0.16 times). The 95% CI was (0.08, 0.33) If a person's body took more time for prothrombin formation, they had lesser odds of developing ascities. The 95% CI was (0.58, 1.04) Female, hepatem and alk_phos were all statistically significant in determining whether a person had ascities or not.

12 Task 12

For me, the best subsets didn't work, and my R crashed a couple of times. Thus, I decided to do away with using best subsets, and focussed on stepwise regression. I thought making linear model would be easier, but it was slightly more difficult due to the transformation. I wish I had all the ways of calibration and validation at the back of my head, since I had to look up in the slides every time for this. I also wish I had known how to improve a pre-existing model, since I devoted a lot of time for that. I had to re-read the analysis for models, since I had forgotten how to provide the model summary. Holding onto everything together was really confusing, as I did a lot of analysis, and in-between forgot a lot of stuff which I had planned. Assembling everything together was also confusing. I believe the most useful things I learnt from this project were to calibrate and validate the models and to reduce the number of variables and improve a pre-existing model.