STAT 5703 Homework 1

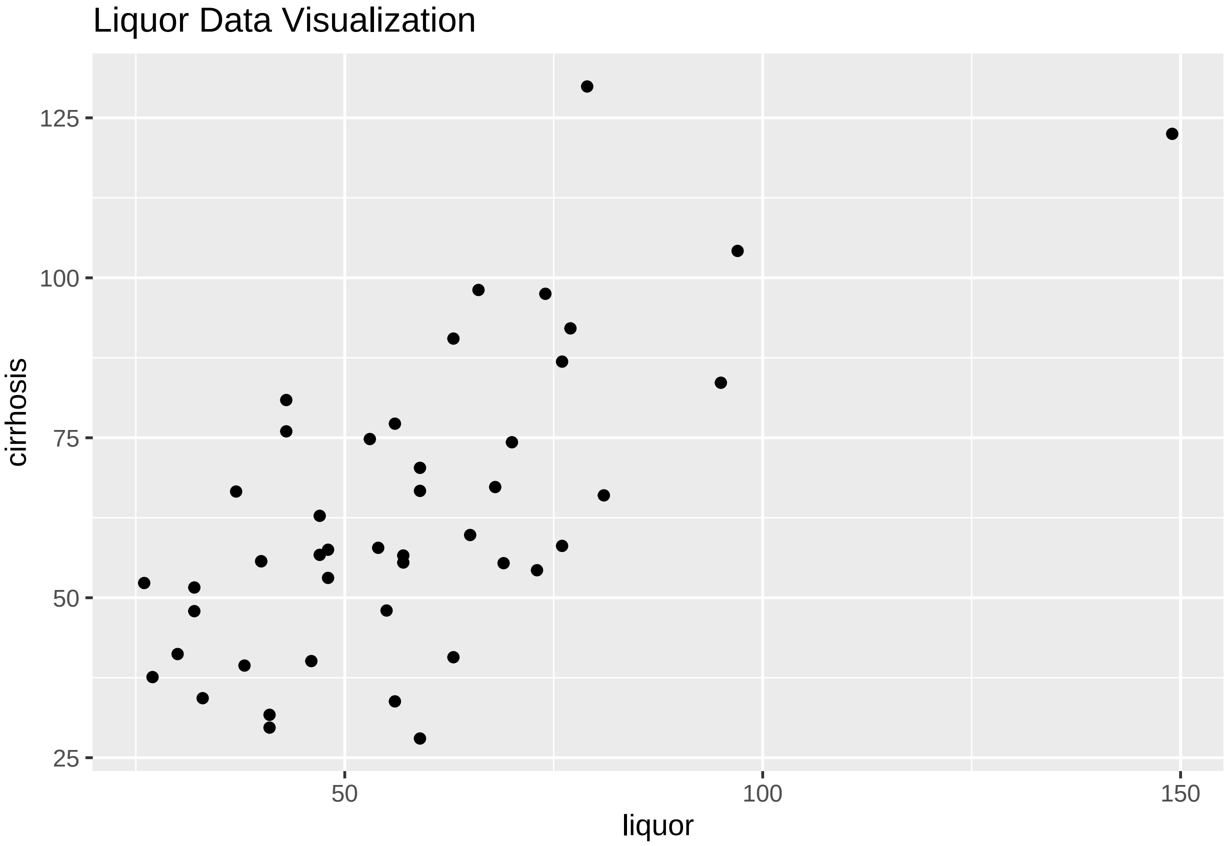
Shijie He(sh3975), Yunjun Xia(yx2569), Shuyu Huang(sh3967)

2/6/2020

Exercise 4

1. Visualize the data and discuss the pertinence of fitting a straight line to this data set.

```
ggplot(data=liver, aes(x = liquor, y =cirrhosis) )+
  geom_point()+
  ggtitle('Liquor Data Visualization')
```



Accoring to the scatter plot, we could observe a positive linear correlation betweem two variables that the more liquor per capita comsumption, the higher the cirrhosis mortality rate. Therefore, it is reasonable to fit a straight line to the dataset.

2. Which of the two variables would you interpret as your response varible.

I would interpret the variable cirrhosis mortality rate as response variable since we are going to investigate how cirrhosis mortality rate change according to the change of liquor per capita consumption.

3. What sign do you expect your two parameters to have? Justify this intuition and interpret the meaning of these data.

I would expect signs for slope and intercept are both positive.

The positive slope refers that the more liquor per capita consumption, the higher cirrhosis mortality rate would be.

The positive intercept refes that even though liquor per capita is zero, we would still have cirrhosis mortality rate according to the natural law.

4.

(a) α refers to cirrhosis mortality rate when liquor per capita consumption in the area is 0. β refers to units the cirrhosis mortality rate change according to one unit liquor per capita change.

(b) α represents to cirrhosis mortality rate when liquor per capita consumption equals to mean liquor per capita consumption. β refers to units the cirrhosis mortality rate change according to one unit liquor per capita change.

5.

```r
liquor_fit<-lm(cirrhosis~liquor, data=liver)
summary(liquor_fit)
```

```
##
## Call:
## lm(formula = cirrhosis ~ liquor, data = liver)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.577 -11.127  -0.821  11.179  50.878
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.9649     7.1847   3.057  0.00379 **
## liquor        0.7222     0.1168   6.185  1.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.34 on 44 degrees of freedom
## Multiple R-squared:  0.4651, Adjusted R-squared:  0.4529
## F-statistic: 38.26 on 1 and 44 DF,  p-value: 1.803e-07
```

Least Squares estimator of α is 21.9649, and $\beta = 0.7222$.

According to the p-values, both are statistically significant.

Interpretation: In the area where the liquor per capita consumption is 0, the cirrhosis mortality rate in this area would be 21.9649.

When one unit of liquor per capita consumption increase, cirrhosis mortality rate will be increase by 0.7222 unit.

6. If you do not want to assume a linear model as in point 4, how do you interpret the least squares estimates that you gave in the previous question?

Without linear model assumption, the least squares estimates refers to be the best linear (fit) relationship between two varibales. The produced slope and intercept explains the increasing in cirrhosis mortalilty rate as liquor per capita consumption increases. The significance level only indicates how important the parameters to the model. Since we do not make assumption on model, the significance level of paramenters has no significance in our interpretation.

7. Give a prediction for the cirrhosis mortality rate, cirrhosis, for a region with per capita liquor consumption of 180 ounces. Is this a good predictor given the data at hand?

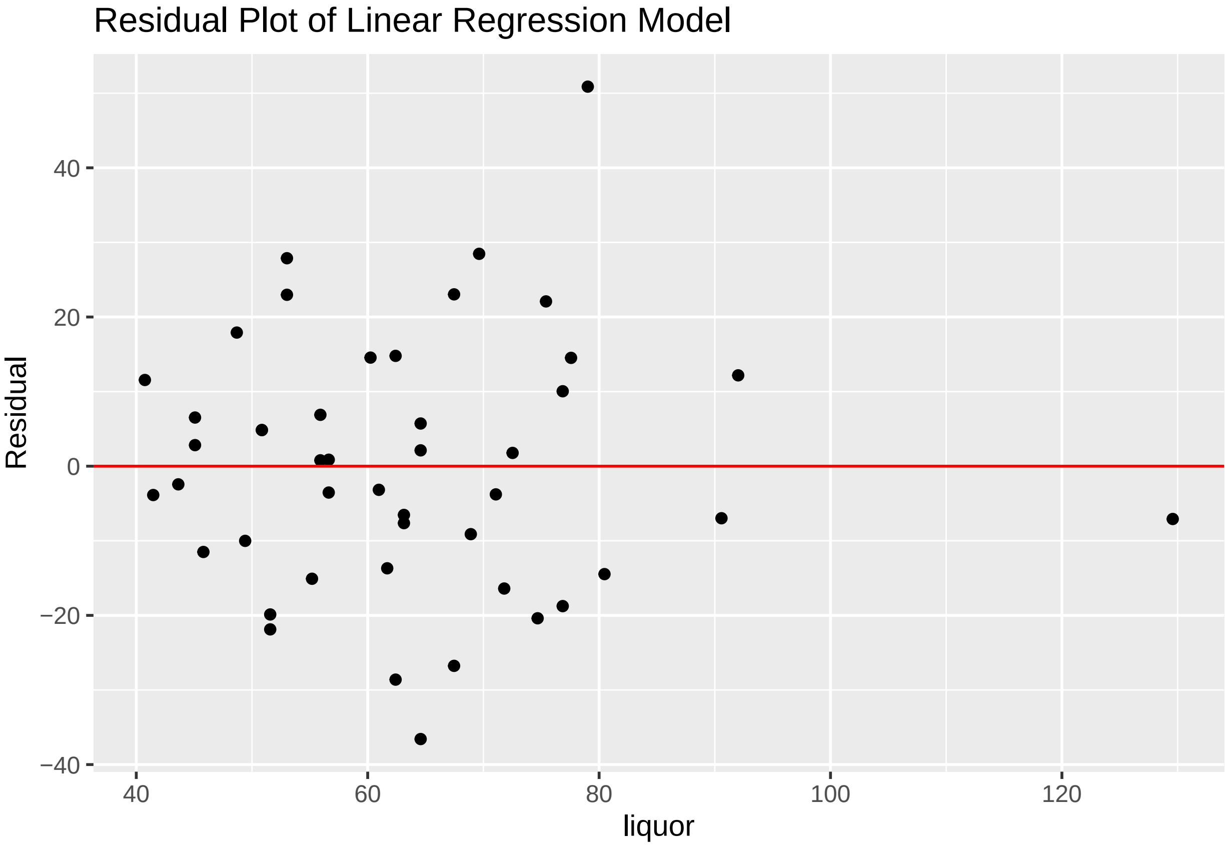
$$y\hat{=}\alpha\hat{+}100\beta\hat{=}21.9649 + 0.7222(180) = 151.9609$$

It is not a good predictor given the data at hand because most of our data points of liquor per capita consumption are less than 100 and only one large data point for liquor per capita consumption is 150, which is also less than our goal of prediction. We may need more large data in order to do better prediction.

8. Plot the residuals of your least squares fit.

Does it seem reasonable to assume that errors ϵ_i are iid?

```
ggplot(liquor_fit)+
  geom_point(aes(x=.fitted, y=.resid))+
  ylab('Residual')+
  xlab('liquor')+
  ggtitle("Residual Plot of Linear Regression Model")+
  geom_hline(yintercept=0, color='red')
```



According to the residual plot, most of residuals are randomly distributed around 0. The even distributions refers to constant variance of residuals. The randomness in distribution indicates indenpendency between response and independent variables. Therefore, it is reasonable to assume errors ϵ_i are iid.

9. Give an exact 95% confidence interval for β assuming that the noise terms are i.i.d. normal. Compare it with a 95% asymptotic confidence interval that does not assume that the errors are normal. Discuss briefly their relative merits.

95% confidence interval for β assuming that the noise terms are i.i.d. normal:

$$[\beta - t_{n-2,0.975}\sqrt{\frac{SS_R}{(n-2)S_{xx}}}, \beta + t_{n-2,0.975}\sqrt{\frac{SS_R}{(n-2)S_{xx}}}]$$
$$= [0.7222 - 2.015368 * StandardError, 0.7222 + 2.015368 * StandardError]$$
$$= [0.7222 - 2.015368 * 0.1168, 0.7222 + 2.015368 * 0.1168]$$
$$= [0.4868051, 0.9575949]$$

```
alpha = 0.05
n=46
t=qt(1-alpha/2, n-2)
se = 0.1168 #according to output in question 5
LB=0.7222-t*se
UB=0.7222+t*se
c(LB, UB)
```

```
## [1] 0.4868051 0.9575949
```

95% asymptotic confidence interval (By Ex4.4 in PS3):

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} N(0, \frac{\sigma^2}{\sigma_X^2})$$

Known that:

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

By Slutsky's Theorem,

$$\frac{\sigma^2}{\hat{\sigma}^2} \xrightarrow{\mathcal{P}} 1$$

, we use $\hat{\sigma}$ to estimate σ.

The confidence interval is in the form:

$$[\hat{\beta} - z_{0.975}\frac{\hat{\sigma}}{\sqrt{n}\sigma_X}, \hat{\beta} + z_{0.975}\frac{\hat{\sigma}}{\sqrt{n}\sigma_X}] = [0.4984011, 0.9460696]$$

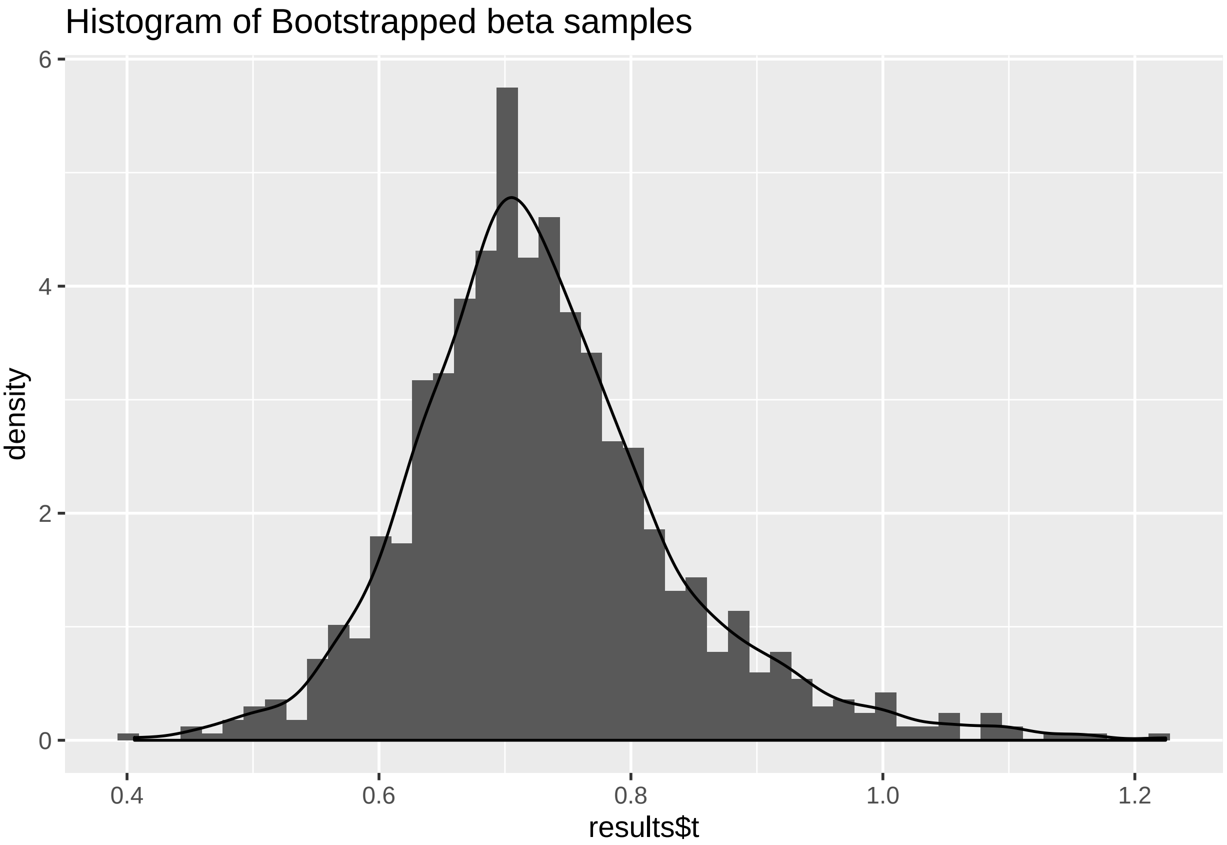
```
n=nrow(liver)
sighat=sqrt(sum(resid(liquor_fit)^2)/n)
Sxx=sum((liver$liquor-mean(liver$liquor))^2)
interval = qnorm(0.975)*sighat/sqrt(Sxx)
betahat = as.numeric(coef(liquor_fit)['liquor'])
LB2= betahat-interval
UB2= betahat+interval
c(LB2,UB2)
```

```
## [1] 0.4984011 0.9460696
```

The two confidence interval is quite similar. The advantage of first method is that it does not require a large sample size. The advantage of second method is that it does not requre normal asssumption of data.

10. Generate 1000 bootstrap samples and use them to compute a 95% bootstrap confidence interval for β. Plot the bootstrap distribution that you obtained and compare your bootstrap confidence interval with the two obtained in point 9.

```
library(boot)
set.seed(5703)
bootstrap <- function(liver, indices) {
  d <- liver[indices,] # allows boot to select sample
  fit<-lm(cirrhosis~liquor, data=d)
  return(coef(fit)['liquor'])
}
results <- boot(data=liver, statistic=bootstrap, R=1000)
ggplot()+
  geom_histogram(aes(x=results$t, y=..density..), bins=50)+
  geom_density(aes(x=results$t, y=..density..))+
  ggtitle('Histogram of Bootstrapped beta samples')
```



```
boot.ci(results, type="bca")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 0.5667,  1.0430 )
## Calculations and Intervals on Original Scale
```

The bootstrap distribution obtained follows a kernel shape. It seems like a normal distribution with slight right-skewness.

The 95% CI produced with bootstrap method is [0.5667, 1.0430]. It is slightly greater than Confidence interval we obtained in Question9.

11.

```
p1 =cor(liver$cirrhosis,liver$liquor)
p1
```

```
## [1] 0.6819694
```

Correlation between liquor and cirrhosis

$$\hat{\rho} = corr(y, x) = 0.6819694$$

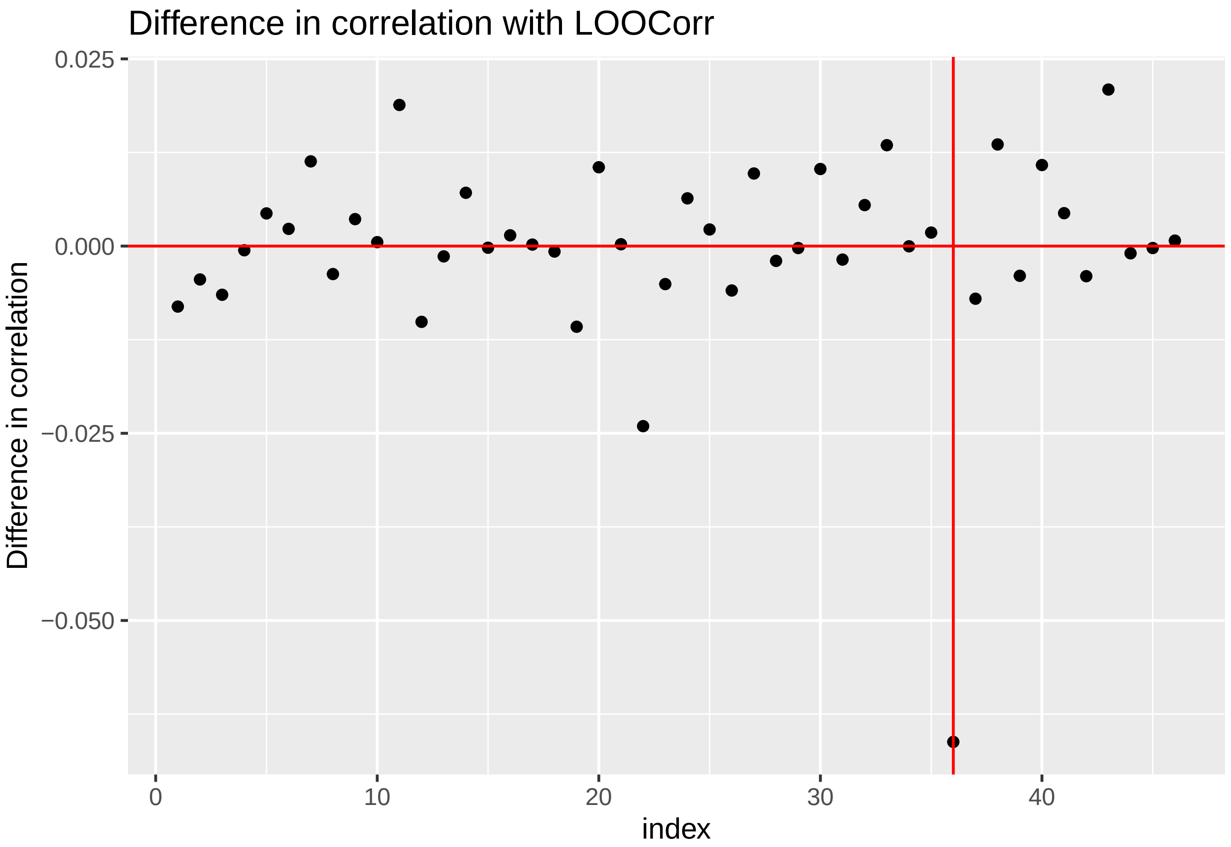
```
library(dplyr)
```

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

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

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

```
LOOCorr <- function(i){
  dftmp <- liver[-c(i),]
  cor(dftmp$cirrhosis, dftmp$liquor) - cor(liver$cirrhosis,liver$liquor)
}
corr_diff <- unlist(Map(LOOCorr, 1:nrow(liver)))
livercopy<-liver
livercopy$corr_diff<-corr_diff
maxindex=which.max(abs(livercopy$corr_diff))
ggplot(livercopy)+
  geom_point(aes(x=seq.int(1,nrow(livercopy)), y=corr_diff))+
  xlab('index')+
  ylab('Difference in correlation')+
  ggtitle('Difference in correlation with LOOCorr')+
  geom_hline(yintercept=0, color='red')+
  geom_vline(xintercept=maxindex, color='red')
```



```
livercopy[maxindex,]
```

```
##    liquor cirrhosis   corr_diff
## 36    149     122.5 -0.06621551
```

According to the plot, after leaving an obeservation index at 36, the corrlation is largely affected.

The data that (liquor=149, cirrhosis= 122.5) is particularly influential in the analysis.