STAT5703 HW1 Exercise 4

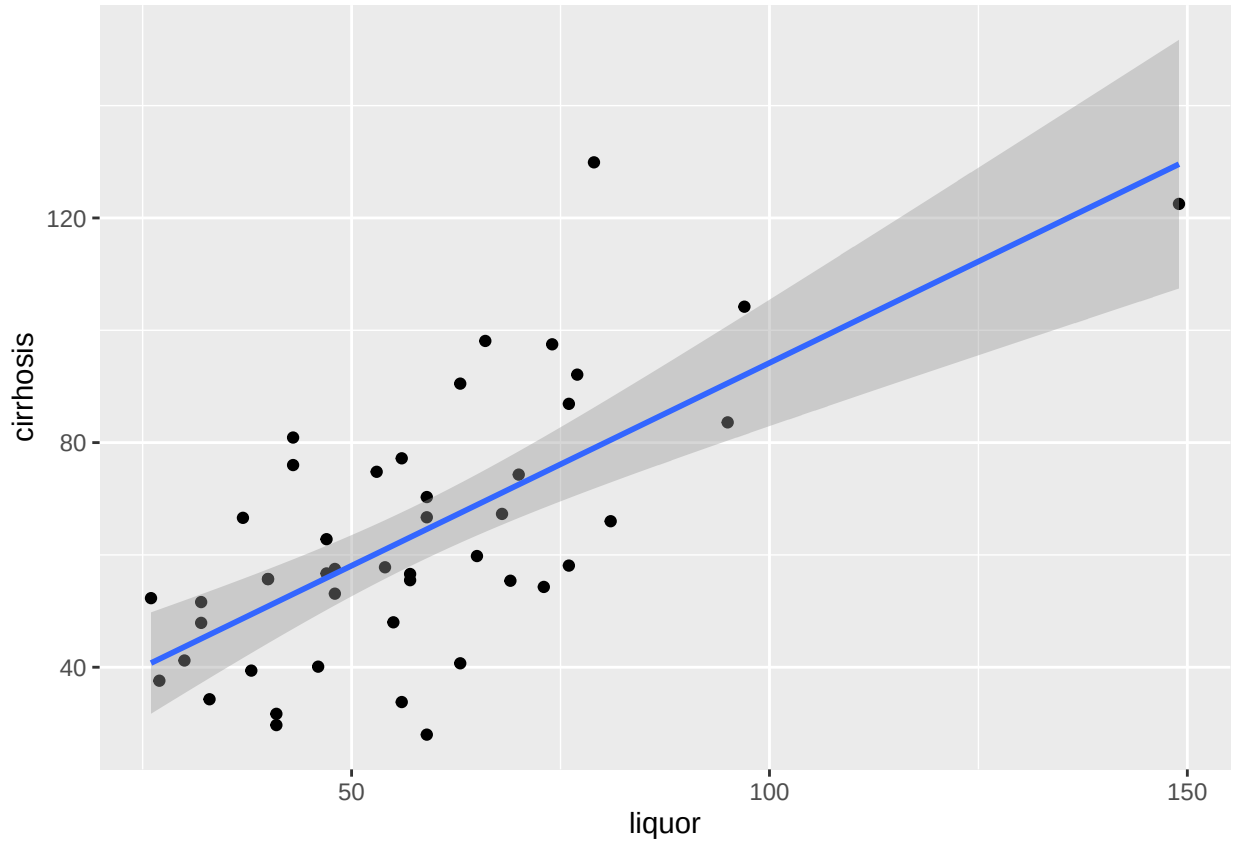
Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)

Exercise 4.

```
df_liver <- read.csv(file ="liver.csv")
names(df_liver) <- c('liquor','cirrhosis')
```

Question 1.

```
library(ggplot2)
ggplot(df_liver, aes(x=liquor, y=cirrhosis)) +
  geom_point() +
  geom_smooth(method='lm')
```



By visualizing the data, we can observe that 'cirrhosis' and 'liquor' have a positive relationship and fit a straight line.

Question 2.

Because cirrhosis is affected by liquor consumption, 'cirrhosis' should be the response variable in this measurement.

Question 3.

Intuitively, the higher the liquor consumption per capita, the higher the cirrhosis mortality rate. When alcohol consumption is zero, the cirrhosis mortality rate should be positive because there are other factors besides alcohol that can cause long-term damage to the liver. Therefore, when cirrhosis is our response variable, the slope should be positive and the intercept should also be positive. We can prove our intuition by looking at the graph of question one, first of all, the line is going up, so the slope is positive and correlation between liquor and cirrhosis shows a high positive relationship. Second, when liquor is 0, the line points to a value above 0.

```
cor(df_liver$liquor, df_liver$cirrhosis)
```

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

Question 4.

Model A: α is the expected Y value when X variable is 0. β is the slope, which means how much value of Y changed when X is increased by 1. Model B: α is the expected Y for the mean of X. β is how much value of Y changed if difference between X_i and the mean is increased by 1.

Question 5.

Model A:

```
linearModA <- lm(cirrhosis ~ liquor, data=df_liver)
summary(linearModA)
```

```
##
## Call:
## lm(formula = cirrhosis ~ liquor, data = df_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
```

For model A, $\alpha = 21.9649$, $\beta = 0.7222$. As our interpretation in question 4, when liquor consumption is 0, cirrhosis mortality rate is 21.9649. When liquor increases 1 unit, cirrhosis mortality rate increases 0.7222. The linear model is statistical significance if both p-values of individual predictor variables and model are less than 0.05. For model A, p-values of both parameters and F-statistic are less than 0.05, so model A is statistically significant.

Model B:

```r
df_liver$DiffX <- df_liver$liquor-mean(df_liver$liquor)
linearModB <- lm(cirrhosis ~ DiffX, data=df_liver)
summary(linearModB)
```

```
##
## Call:
## lm(formula = cirrhosis ~ DiffX, data = df_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)  63.4935     2.5571  24.830  < 2e-16 ***
## DiffX         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
```

For model B, $\alpha = 63.4935$, $\beta = 0.7222$. When liquor consumption is equal to mean, cirrhosis mortality rate is 63.4935. When difference between liquor and the mean increases 1 unit, cirrhosis mortality rate increases 0.7222 which is same as the model A. Model B is statistical significance because p-values of both parameters and F-statistic are less than 0.05.

Question 6.

We can interpret the least squares estimates from their equations:$\hat{\alpha} = \bar{y}_n - \bar{x}_n\hat{\beta}$, $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ can be written as,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{Sample\ Covariance\ between\ X\ and\ Y}{Sample\ Variance\ of\ X}$$

Therefore, the higher the covariance between X and Y, the higher$\hat{\beta}$. Negative covariances cause negative $\hat{\beta}$, positive covarances cause positive $\hat{\beta}$. The higher $\hat{\beta}$ causes lower $\hat{\alpha}$.

Question 7.

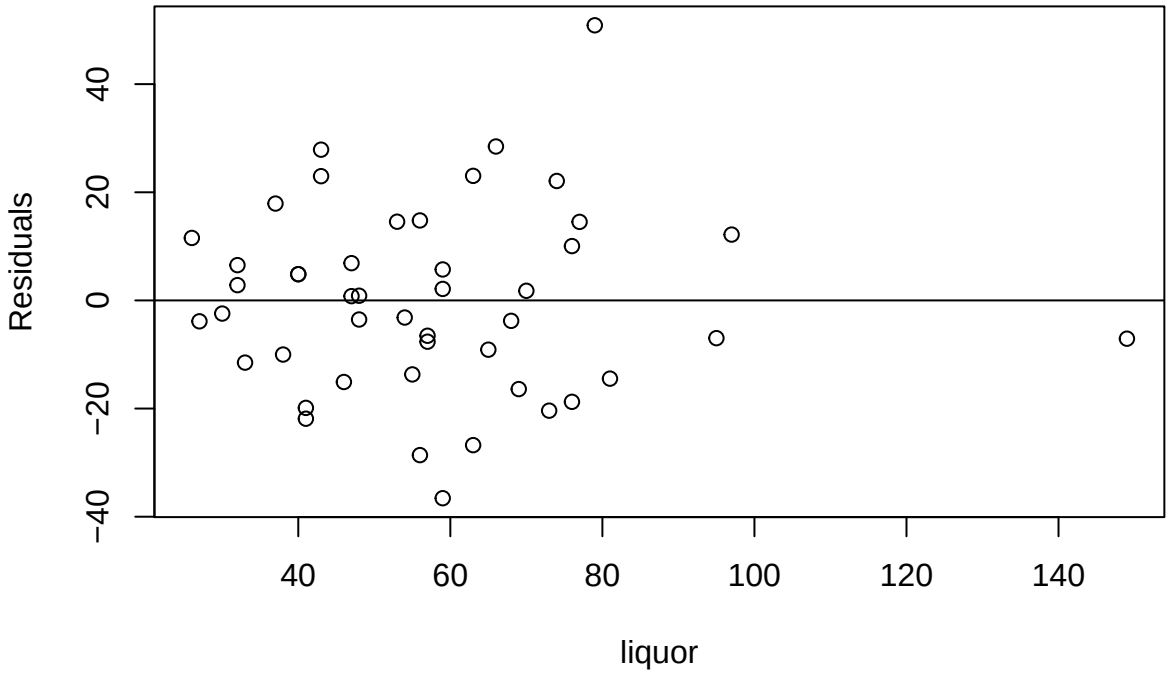
```r
new.liquor <- data.frame(
  liquor = c(180)
)
linearMod <- lm(cirrhosis ~ liquor, data=df_liver)
predict(linearMod, new.liquor)
```

```
##        1
## 151.9673
```

In our prediction, cirrhosis mortality rate is 151.9673 with per capita liquor consumption of 180 ounces. This is a good predictor, since from the plot in question 1, the line should around 150 when liquor is 180.

Question 8.

```
linearMod <- lm(cirrhosis ~ liquor, data=df_liver)
res = resid(linearMod)
plot(df_liver$liquor, res, ylab="Residuals", xlab="liquor")
abline(0, 0)
```



The residuals seem reasonable to assume that errors ϵ_i are i.i.d. because the plot shows that residuals are randomly distributed around 0 and $\mathbf{E}[\epsilon_i]$ approximately equal to 0.

Question 9.

```
confint(linearMod)
```

```
##                2.5 %     97.5 %
## (Intercept) 7.485087 36.4448077
## liquor      0.486901  0.9575696
```

The confidence interval of β is $[0.486901, 0.9575696]$ if noise terms are i.i.d. normal. We have $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, therefore,

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{n}(x_i-\bar{x})\epsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Let's assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

In our case, replace σ^2 with unbiased estimator $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\hat{\epsilon}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2}$. Therefore $\frac{\hat{\beta}-\beta}{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim \mathcal{T}_{n-2}$. The confidence interval is,

$$\hat{\beta} \pm t_{\alpha/2,n-2} \times \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}$$

To compute asymptotic confidence interval, We know that,

$$\mathbf{E}[\hat{\beta}] = \beta$$

$$Var[\hat{\beta}|X] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Hence by CLT,

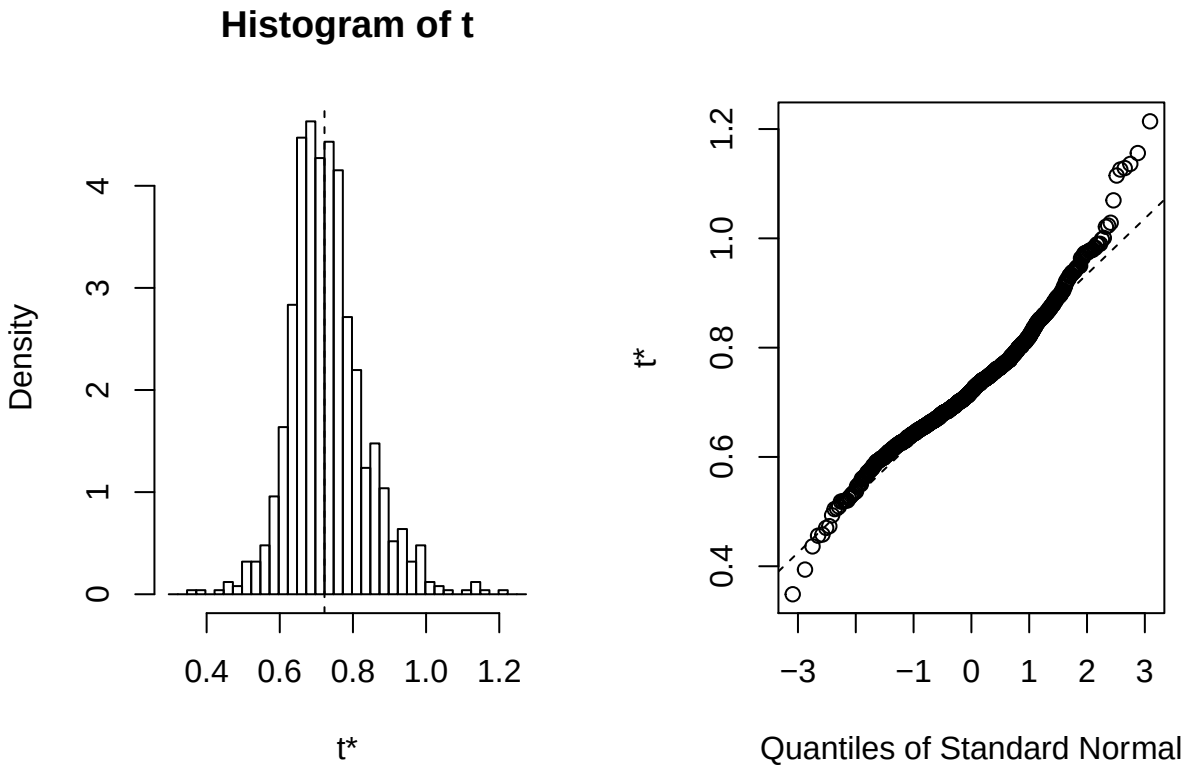
$$\frac{\hat{\beta} - \beta}{\sqrt{Var[\hat{\beta}|X]}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1)$$

$$\hat{\beta} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

which is same as the distribution with normality assumption. Hence the asymptotic confidence interval should be same as the exact confidence interval. If we assume that the noise is exactly normal and independent, we can formally get the properties of estimators but those are just for small sample. If we assume that the noise is not normal just independent, we can obtain a normal distribution asymptotically as $n \to \infty$, but those properties are not appropriate when n is finite.

Question 10.

```r
library(boot)
rsq <- function(formula, data, indices) {
  d <- data[indices,] # allows boot to select sample
  fit <- lm(formula, data=d)
  return(coef(fit)["liquor"])
}
results <- boot(data=df_liver, statistic=rsq,
   R=1000, formula=cirrhosis ~ liquor)
plot(results)
```



We can see the bootstrap distribution from the histogram above.

```
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%   ( 0.5671,  0.9827 )
## Calculations and Intervals on Original Scale
```

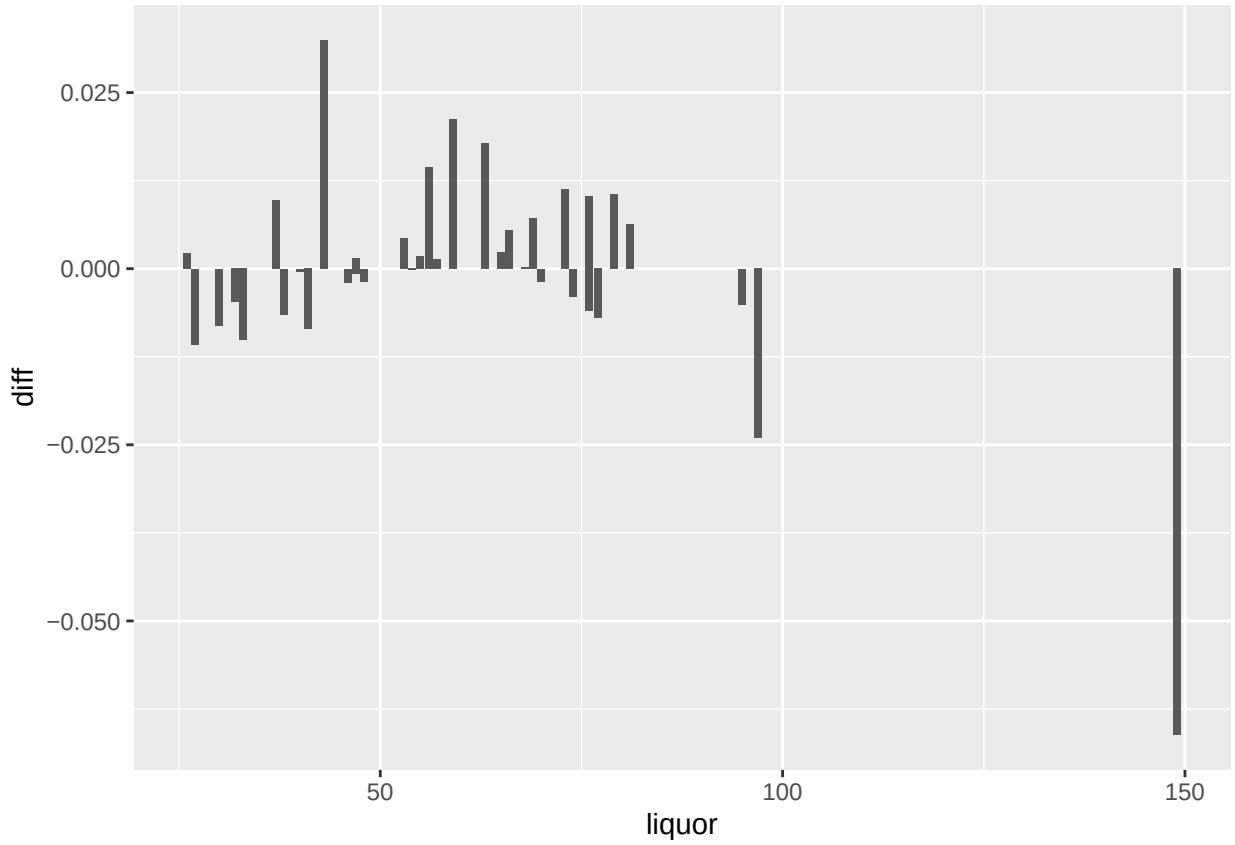
The 95% bootstrap confidence interval for β is $[0.5451, 1.0081]$ which is slightly wider than the confidence interva in question 9:$[0.486901, 0.9575696]$.

Question 11.

```
corr = cor(df_liver$liquor, df_liver$cirrhosis)
df_liver$diff <- NULL
leave_one_out_corr<- function (idx) {
df_tmp <- df_liver[-c(idx),]
return(cor(df_tmp$liquor, df_tmp$cirrhosis) - corr)
```

```
}
df_liver$diff<-unlist(Map(leave_one_out_corr, seq.int(1, nrow(df_liver))))
```

```
ggplot(df_liver) + geom_col(aes(x=liquor,y=diff))
```



In the plot above, we can clearly observe that the difference $\hat{\rho}_{(i)} - \hat{\rho}$ of some liquor values is larger than others, especially the liquor value which close to 150, indicating that these points are being particularly influential in the analysis.