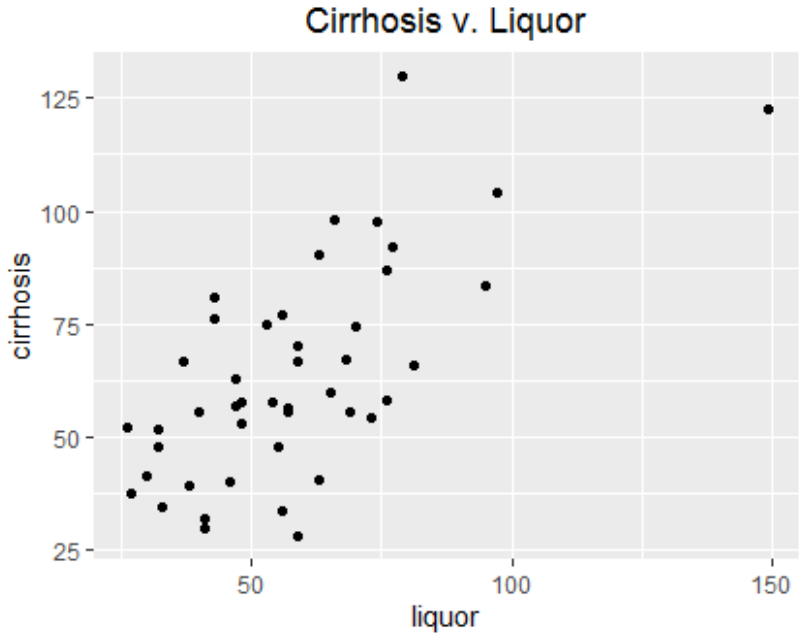
Group: Zining Fan, Mutian Wang, Siyuan Wang
UNI: zf2234, mw3386, sw3418

Graded Homework 1 - Exercise 4

1. Fitting a straight line to the data seems reasonable, since there seems to be a positive linear relationship between cirrhosis mortality rate ("cirrhosis") and liquor consumption per capita ("liquor"). Yet the data points look somewhat dispersed, so we guess R^2 might not be very good.



2. Cirrhosis mortality rate is the response variable.

3. We expect the two parameters to have the same sign, because from the scatter plot we can clearly see a positive correlation between cirrhosis and liquor. That is, the more liquor a person drinks, the more likely he/she dies of cirrhosis.

4. (a) Basically, this model is to find a straight line to best fit the data. α is the intecept of the line, and β is the slope of the line.

   (b) This model will move all the data points to the left for the distance of $\bar{X}_n$, then it will find a straight line to best fit the data. α is the intecept of the line, and β is the slope of the line. More importantly, $\hat{\alpha} = \bar{Y}$.

   This is because $Y_i = (\alpha - \beta\bar{X}_n) + \beta X_i$ and $\hat{\alpha} - \hat{\beta}\bar{X}_n = \bar{Y} - \bar{X}_n\hat{\beta}$ should be the LS estimates. Thus we have $\hat{\alpha} = \bar{Y}$.

5. We use $lm()$ function in R for both models, and part of the output is as follows.

```
fit = lm(df$cirrhosis ~ df$liquor)
summary(fit)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.9649     7.1847   3.057  0.00379 **
df$liquor     0.7222     0.1168   6.185  1.8e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
x = df$liquor - mean(df$liquor)
y = df$cirrhosis
fit1 = lm(y ~ x)
summary(fit1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.4935     2.5571  24.830  < 2e-16 ***
x             0.7222     0.1168   6.185  1.8e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

From the output we can see, both estimates in model (a) and (b) are statistically significant under the 5% significance level. Thus there is a relationship between corrhosis and liquor.

We can see model (a) and (b) have identical estimated values for β. Also we checked the estimated value of α in model (b), and indeed it equals to $\bar{Y}$.

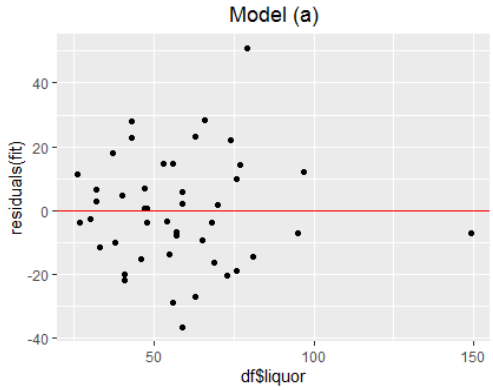
6. The sign of $\hat{\beta}$ can roughly tell us the correlation is positive (if $\hat{\beta} > 0$) or negative (if $\hat{\beta} < 0$). If $\hat{\beta} = 0$, it is likely that there is no relationship between X and Y. This is because the sign of $\hat{\beta}$ is the same as the sign of the covariance.

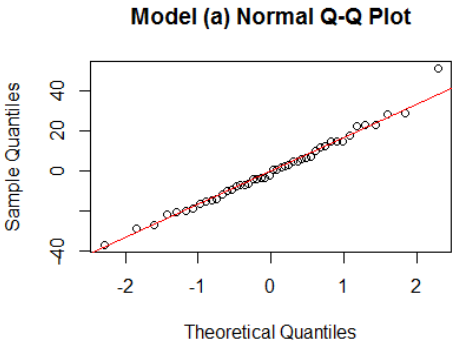
   Moreover, in model (b), $\hat{\alpha} = \bar{Y}$.

7. Based on the previous estimated values, it can be calculated that the corrhosis mortality rate is 151.9673. Model (a) and (b) have the same predicted value.

   It is hard to say this is a good predictor. The R-squared for both models is aournd 0.45, which is not too good. The liquor consumption of nearly all training data (except one) is below 100 ounces, so we doubt whether the linear relationship is still true when the liquor consumption is as large as 180 ounces.

8. Since the plots for model (a) and model (b) are the same except for the scales, so we only include the plots generated by model (a).





From the first plot, we can see there is no pattern of residuals. That is, residual is not a function of liquor consumption. From the second plot, we can see that the residuals belong to a normal distribution, since almost all points are on the line. Thus we think it is reasonable to assume that errors are $i.i.d.$ normal.

9.
```
> confint(fit)
                2.5 %     97.5 %
(Intercept) 7.485087 36.4448077
df$liquor   0.486901  0.9575696
```

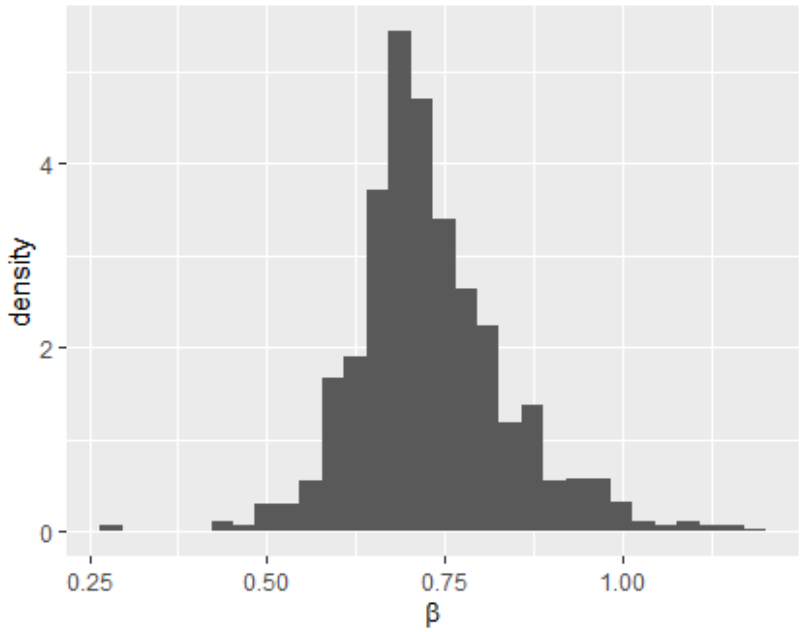
Therefore the exact 95% CI for β is [0.486901, 0.9575696].

If the error is not normal, the estimator is still asymptotically normally distributed. In point 5, we have already got that the standard error of β is 0.1168. We also know that $z_{0.975} = 1.96$ and $\hat{\beta} = 0.7222$. Therefore the asymptotic 95% CI is [0.7222-1.96*0.1168, 0.7222+1.96*0.1168]. That is, [0.493272, 0.951128].

When the errors are normal, the exact CI should be used. It is wider than the asymptotic CI, so it is a more conservative and safer way of inference.

The asymptotic CI also works when the errors are not normal and the sample size is very large.

10. The confidence interval of bootstrapping method is [0.5511782, 0.9791503], which is very similar to the CIs obtained above. Generally speaking, the bootstrapped CI is not stable.



11. Part of the code and output is as follows.

```
> cor(df$cirrhosis, df$liquor)
[1] 0.6819694
```

```
> quantile(diff)
           0%           25%           50%           75%          100%
-6.621551e-02 -3.911770e-03  8.479921e-05  5.205320e-03  2.090183e-02
```

As we can see the $\hat{\rho} = 0.6819694$. As for $\hat{\rho}_{(i)} - \hat{\rho}$, the minimum difference is around -0.066, and the maximum difference is around 0.021, so we do not think there are any observations that are particularly influential in the analysis.

FYI, the minimum difference is caused by $(y_{36}, x_{36}) = (149, 122)$; the maximum difference is caused by $(y_{43}, x_{43}) = (59, 28)$