Prediction and Time Series Computer Exercise Week 1

Name: Nguyen Xuan Binh   Student ID: 887799

A) Formulate a linear regression model, where the variable ILL is explained with the variable CONSUMPTION. Include a constant term in your model

```
tobacco_data<- read.table('tobacco.txt', header=T, sep = '\t')
lm_ill <- lm(ILL ~ CONSUMPTION, data = tobacco_data)
summary <- summary(lm_ill)
summary

## 
## Call:
## lm(formula = ILL ~ CONSUMPTION, data = tobacco_data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -169.016  -32.813    0.004   45.804  136.914 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.74886   48.95871   1.343  0.21217   
## CONSUMPTION  0.22912    0.06921   3.310  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 84.13 on 9 degrees of freedom
## Multiple R-squared:  0.549,  Adjusted R-squared:  0.4989 
## F-statistic: 10.96 on 1 and 9 DF,  p-value: 0.009081
```

lm_ill is the linear regression model of response variable ILL from explanatory variable CONSUMPTION. Results above is summary of the linear regression model Intercept of the coefficients is the constant term in the regression model

B) Estimate the regression coefficients of the model by using the least squares method and give interpretations for the estimated regression coefficients.

```
consumption <- tobacco_data$CONSUMPTION
X = as.matrix(cbind(rep(1,nrow(tobacco_data)), consumption))
b_coefficient <- solve(t(X) %*% X) %*% t(X)%*% tobacco_data$ILL
b_coefficient

##                   [,1]
##             65.7488570
## consumption  0.2291153
```

Least squares method: b = (X^T * X)^-1 * X^T * y, where X = [vector(1) vector(xi)… vector(xk)]

=> regression coefficients of the model: b1 = 0.2291153 Interpretation: b1 = 0.2291153 is an unbiased estimator for the real β coefficient of linear regression model of ILL-CONSUMPTION y = Xβ + Є

The intercept/constant term for the least squares method is 65.7488570

C) What is the coefficient of determination of the model?

```
summary$r.squared

## [1] 0.54904
```

The coefficient of determination of the model is R^2 = 0.54905

D) Is the model statistically significant according to the F-test? Use 1% as the level of significance

```
summary

## 
## Call:
## lm(formula = ILL ~ CONSUMPTION, data = tobacco_data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -169.016  -32.813    0.004   45.804  136.914 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.74886   48.95871   1.343  0.21217   
## CONSUMPTION  0.22912    0.06921   3.310  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 84.13 on 9 degrees of freedom
## Multiple R-squared:  0.549,  Adjusted R-squared:  0.4989 
## F-statistic: 10.96 on 1 and 9 DF,  p-value: 0.009081
```

To know whether a model is statistically significant in F test, we need to compare F test's p value with the given significant level that we want to achieve. In the calculation above, we find that 0.009081 is smaller than the significant level of 0.01

=> the model is statistically significant according to the F-test

=> null hypothesis is unlikely

E) Is the variable CONSUMPTION statistically significant according to the t-test? Compare the p-value with the p-value obtained in part (d) and explain the connection between them

```
p_value_consumption <- summary$coefficients[,4][2] # p-value of CONSUMPTION
p_value_consumption

## CONSUMPTION
## 0.009081016

p_value_consumption < 0.01 # Level of significance

## CONSUMPTION
##        TRUE
```

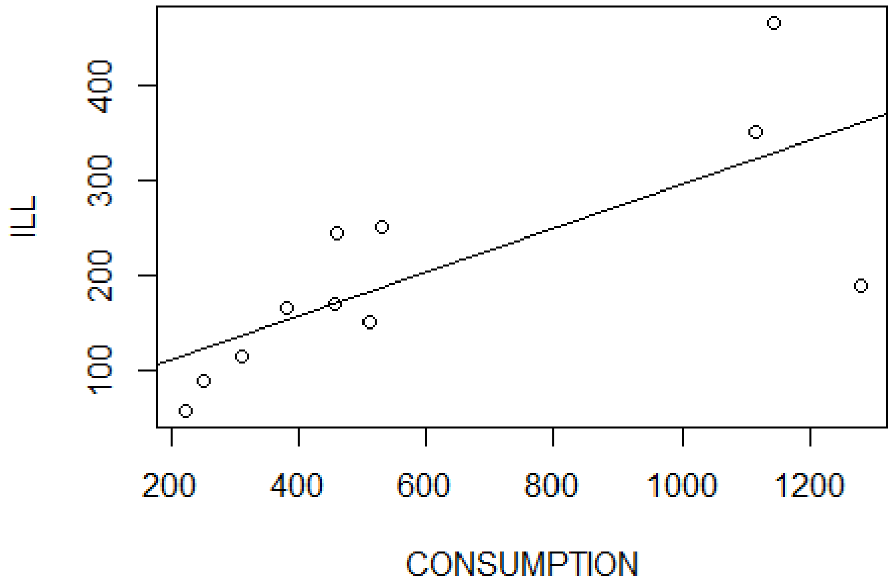
The p-value of the CONSUMPTION variable (Pr(>|t|)) is smaller than the level of significance, which implies than the model of ILL-CONSUMPTION is statistically significant The p value of CONSUMPTION and of F-test found in (d) are equal/almost equal

=> p value of F test is directly derived from CONSUMPTION variable which is the only explanatory variable

=> we can safely believe that there is correlations between ILL and CONSUMPTION and thus the null hypothesis is incorrect for this model

F) Plot the estimated regression line together with the data The plot is as follows

```
plot(tobacco_data$CONSUMPTION, tobacco_data$ILL, xlab = "CONSUMPTION", ylab =
"ILL")
abline(lm_ill)
```



G) Explain the concept of confidence interval

The confidence interval (CI) in statistics is a range of values, within some degree of confidence, that will contain the actual concerned statistical value. It is denoted in percentage % where the true value actually lies between the upper and lower interval. For example, a 95% confidence interval means if random intervals are calculated 100 times, 95 of them are likely to include the true value

H) Suppose that the normality assumptions holds. Form a 95% confidence interval for the slope of the regression line. Give also the 99% confidence interval. Note that the confidence interval is notably wide for the constant term. Can you give some explanation for this?

```
confint(lm_ill, level = 0.95)
```

```
##                     2.5 %      97.5 %
## (Intercept) -45.00344053 176.5011546
## CONSUMPTION   0.07254024   0.3856904
```

```
confint(lm_ill, level = 0.99)
```

```
##                      0.5 %      99.5 %
## (Intercept) -93.358900306 224.8566143
## CONSUMPTION   0.004178126   0.4540525
```

95% and 99% confidence interval for the slope of the regression line has been calculated which are shown in the table above.

The confidence interval for the constant term of 95% confidence interval is

```
176.5011546 - (-45.00344053) = 221.5045951
```

and 99% confidence interval is

```
224.8566143 - (-93.358900306) = 318.215514606
```

These intervals are particluarly wide because the wider the interval, the more likely the true parameter will fall inside it. In this case it is 95% and 99% confidence interval, which accounts for these wide gaps.

I) Compute 95% confidence intervals for the regression coefficients with bootstrapping (2000 repetitions). Compare to part (h)

```
n_boot <- 2000
fit_helper_i <- function(X,y) {
  # Sample with replacement
  inds <- sample(1:nrow(X), replace = TRUE)
  X <- X[inds,]
  y <- y[inds]
  # Bootstrap LS estimate
  solve((t(X) %*% X)) %*% t(X) %*% y
}
```

```
y <- tobacco_data$ILL
boot_samples <- replicate(n_boot, fit_helper_i(X,y), simplify = 'matrix')
boot_samples <- cbind(boot_samples, b_coefficient[2])
alpha = 0.05
bootstrap_confint <- t(apply(boot_samples, 1, function(x) quantile(x, probs =
c(alpha/2, 1- alpha/2, 1-alpha/2)))[-3,])
bootstrap_confint
```

```
##              2.5%       97.5%
## [1,] -29.4246055 150.6441567
## [2,]   0.0674636   0.4376498
```

In the table above, [,1] row is the Intercept and row [2,] column is CONSUMPTION. We can see that with bootstrap methods, the values of the confidence interval varies but not too significantly from the values found in the original 95% confidence interval in (h)

=> Bootstrap method still truthfully reflects the original confidence interval

J) What advantages bootstrapping has when compared to the conventional way of determining confidence intervals?

The advantages bootstrapping offers in determining confidence intervals:

+ The bootstrapping approach can be used in all data models because it does not assume distribution of the data => Normality assumption is not necessary

+ An outlier in the dataset distort the actual mean and the standard error of the model => bootstrapping will help flatten out the effects of outlier

+ In reality, getting new sets of data is considered very difficult, impractical or even impossible => Bootstrapping randomizes the existing data set and helps us avoid the cost of repeating the experiment to get other groups of sampled data.