**STM4PSD ASSIGNMENT 4**

In submitting my work, you are consenting that it may be copied and transmitted by the University for the detection of plagiarism. Submission is your guarantee that the below statement of originality is correct.

*“This is my own work. I have not copied any of it from anyone else.”*

NAME: **Maninderpreet Singh Puri**

STUDENT NUMBER: **20494381**

2. Given in the question, the weight loss programs X and Y are using samples from the same population applied. So we can use Paired t-Test to compare the efficacy of two weight loss programs.

Let be the averages of the weight loss of identical twins in programs X, Y respectively and Is the difference in the averages of the weight loss of identical twins in programs X, Y.

So the null Hypothesis is:

= 0

And Alternate Hypothesis is:

≠ 0

t =

S.E. (Standard error) =

S.D. (Standard Deviation) =

Where, , n=12, = mean of the difference of all observations, = observation.

= = -1.385

= ( =2.4108

S.D. (Standard Deviation) = = 0.46815

S.E. (Standard error) = = = 0.13514723

t = = -10.248

1. The formula for the confidence Interval for paired t test is

C.I = t \*

Where t= 2.201 using the T table

= -1.385 = -1.385 2.201 \* 0.13514723 = -1.385 0.2974 = - 1.6824, -1.0876

[- 1.682, -1.088] is the 95 % Confidence Interval

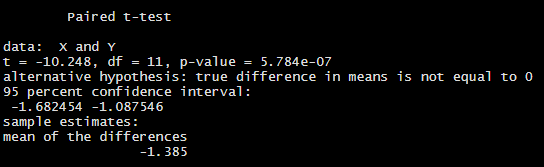
1. Using R command :

X <- c(10.97, 8.53, 7.61, 9.19, 11.22, 12.12, 5.80, 8.62, 8.41, 7.91, 7.32, 10.48)

Y <- c(11.82, 10.05, 8.54, 10.45, 12.62, 14.26, 6.83, 10.21, 9.52, 8.72, 9.25, 12.53)

t.test(X, Y, paired = TRUE)

Output:

 We can see that p- value is 5.784e-07 which is 5.784 \* 10^-07 or 0.0000005784. As the value is much smaller than (for two tailed test) and the value of null hypothesis is not in Confidence Interval, so we reject the null hypothesis that = 0.

Using R command :

X <- c(10.97, 8.53, 7.61, 9.19, 11.22, 12.12, 5.80, 8.62, 8.41, 7.91, 7.32, 10.48)

Y <- c(11.82, 10.05, 8.54, 10.45, 12.62, 14.26, 6.83, 10.21, 9.52, 8.72, 9.25, 12.53)

D <- X-Y

shapiro.test(D)

Output:

Shapiro-Wilk normality test

data: D

W = 0.92435, p-value = 0.3241

For a two tailed test we take = 0.025

p-value = 0.3241

#Null Hypothesis is that data is normally distributed.

#Alternate Hypothesis is that data is not normally distributed.

As the p-value is greater than we can say the null hypothesis is not rejected and data is normally distributed.



Using R command :

w= 386

W=402

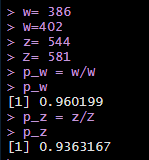
z= 544

Z= 581

= w/W

= z/Z

Output:



Or we can also find using

R input:

prop.test(c(386, 544), c(402, 581))

The last line says

prop 1 () = 0.960

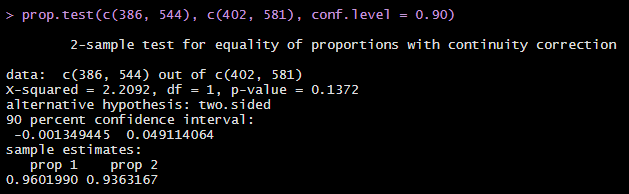
prop 2 () = 0.936



R code:

prop.test(c(386, 544), c(402, 581), conf.level = 0.90)

Output:





p-value = 0.137

b.

No, the p-value is greater than α = 0.1 so we cannot reject that p1 and p2 are equal at this level of significance.

c.

Yes we expect that one hotel performs better than the other with respect to the proportion of satisfied customers since zero is within the interval and cannot reject. We conclude that no statistical difference is significance between the satisfied customers between the two hotels.

d.

The proportion at Hotel W () = 0.960 is better as compare to Hotel Z () = 0.936

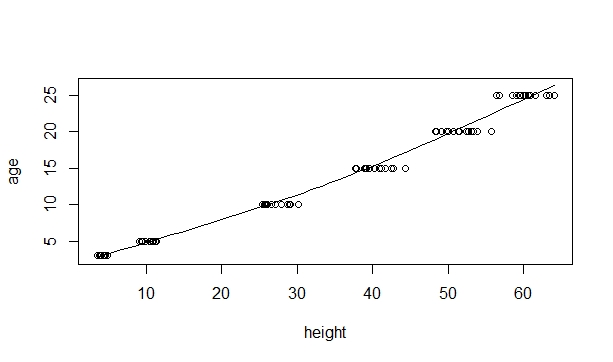
. But the difference of this proportion was not found to be satisfied at C.I 0.90. So we conclude that there is no significance difference in the proportion of satisfied customers between the two hotels. The data collected did not provide enough evidence for us to confidently claim that the one hotel performs better than the other with respect to the proportion of satisfied customers.



R input:

scatter.smooth(Loblolly)

output:





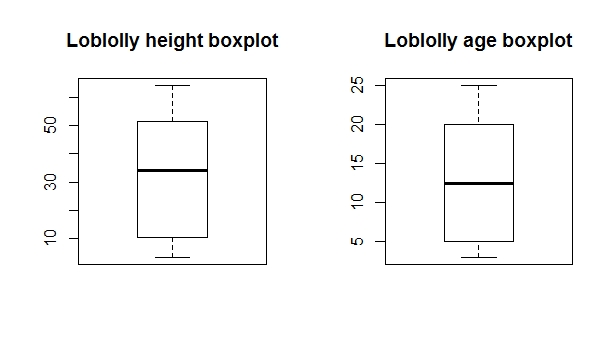
Input code:

par(mfrow= c(1, 2))

boxplot(Loblolly$height, main="Loblolly height boxplot")

boxplot(Loblolly$age, main="Loblolly age boxplot")

Output:



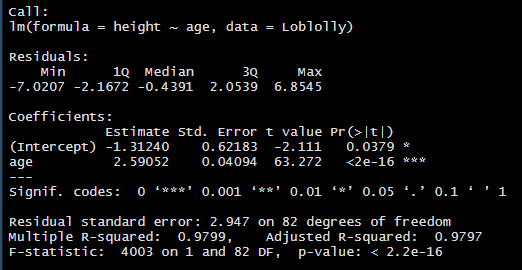


Input code:

loblolly\_lm <- lm(height ~ age, data = Loblolly)

summary(loblolly\_lm)

Output:



i.

We can estimate on average that the height will be increased by 2.59 every year from the Loblolly data set. The intercept coefficient needs to be considered in context so that the height of Loblolly (age=0) is not -1.312 ft. Hence the intercept coefficient can only make sense for ages greater than zero. At zero the height is -1.312ft as age will never be 0.

ii.

Yes. We have R2 or Mutiple R-sqaured = 0.979 which suggests that more than 97.9% of the variation in height can be explained by the estimated model.

iii.

R input:

loblolly\_lm <- lm(height ~ age, data = Loblolly)

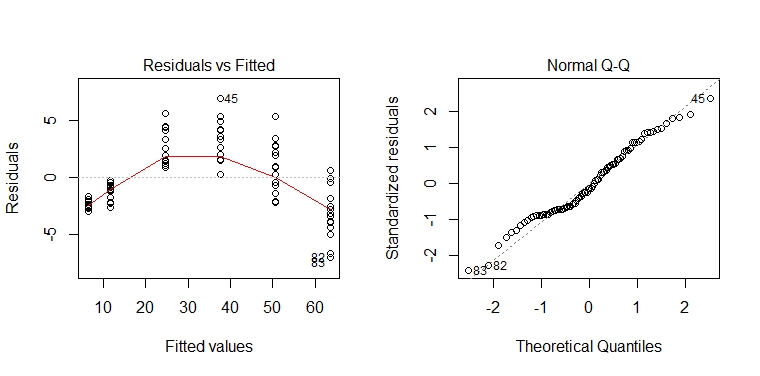
windows(width = 8, height = 4)

par(mfrow = c(1, 2))

plot(loblolly\_lm, which = c(1, 2))

Output:

The residuals versus fits plot and the Q-Q plot of the residuals are below.



There are no patterns or ‘fanning’ in the residuals vs fits plot suggesting independence between residuals and fits and constant error variance. The Q-Q plot of residuals is approximately linear so there are no clear violations of normality. The normality in Q-Q plot appears approximately linear with no clear violation of normality. Although for the standardized residuals of large magnitude there is some deviation from the normal. But the residual vs fitted plot has some concerning features. For example, for the high and low fitted values of the residuals are negatives. While for medium fitted values the residuals are positive. For a model which satisfies the condition of normality we can expect that there are no patterns in the plotted points in the residual vs fitted plot with a random scattered of points centered on 0. So we have some reason to suspect that the condition of normality is not satisfied for this case. In this case linear mixed effect model can be used to model the data.



Since the p-value associated with the coefficient β1 for this test is approximately (2.2e-16) and is smaller than α = 0.05, then we reject the null hypothesis that β1= 0. We conclude that the β1 not equal to 0.



Since the p-value value associated with the coefficient β0 for this test is approximately 0.0379 and is smaller than α = 0.05, we reject the null hypothesis that β0 = 0. We therefore have found evidence of a linear association between height and age and we conclude that β0 not equal 0.



Using

ci.b0 <- sort(c(-1.31240 - qt(0.025, 82) \* 0.62183, -1.31240 + qt(0.025, 82)\* 0.62183))ci.b0

we get:

[-2.549, -0.075]

-2.54941796 -0.07538204

b1.ci <- sort(c(2.59052 - qt(0.025, 82) \* 0.04094, 2.59052 + qt(0.025, 82)\* 0.04094))b1.ci

we get:

[2.509, 2.671]

2.509077 2.671963



R code:

new.data <- data.frame(age= 9)

new.data

predict (loblolly\_lm, new.data)

Output:

22.00231

The height is approx. 22ft.



Input code:

predict (loblolly\_lm, new.data, interval = “confidence)

Output:

predict(loblolly\_lm, new.data, interval = "confidence")

fit lwr upr

1 22.00231 21.28452 22.72011

GG<- predict(loblolly\_lm, data.frame(age = 9), interval = "prediction")

GG[2:3]

[1] 16.09651 27.90811

So the 95% C.I for the tree is (21.284, 22.720) when 1, whereas corresponding P.I (16.096, 27.908). The C.I is narrower than the P.I because the C.I is used when we consider the average height of a 9 year old Loblolly pine tree. On the other hand the P.I can be used in the context of estimating of height of an individual tree, not the average height. This is because the error variance while considering individual trees is greater than the average tree.



According to the above analysis there is a statistical evidence to consider that Loblolly pine tree increase in height as the age. This can be seen in the regression model output with the estimated coefficients β0 and β1 which are both statistically significant. When we tested that it is statistically different from zero whereas the residual plots suggested linear model which may not be the best choice for this data set. The above results confirm with other findings for this data.