

## ECO240: Statistics II

### HONOR CODE STATEMENT


In Department of Economics at Hacettepe University, we take the academic integrity on homework assignments very seriously and expect you to take it seriously too. The aim of this document is to make our expectations as clear as possible to prevent possible academic integrity violations. The following rules will be in force for homework assignments in ECO240:

- You are NOT allowed to discuss with anyone, including your classmates, family members, friends... in finding strategies for the homework questions.
- All work on assignments must be done by you. You may not copy code in whole or in part from someone else. You may not share your code. You may not discuss your solution with others.
- If you need detailed help, ask your instructor. Rather than your instructor, you should not receive help on your solution from individuals. Likewise, you should not hire a tutor to get your homework done.
- You should not show your code to others as a means of helping them. Sometimes, very good students provide their code (sometimes "just a peek" at their code) or detailed help to struggling students for helping them. However, such good intentions often turn into the violation of the academic integrity. If you have such struggling friends, direct them to the instructor.
- You may not leave your code in publicly accessible areas.
- You may not share computers in any way. If you work in a public lab, delete all files related to the assignment when you leave; do not forget to empty the recycle bin.
- You may not use others' storage devices (others' CDs, USBs, etc) to save your work.
- You may not leave any printouts lying around anywhere and you may not dispose them in public trash cans until your assignment has been graded.
- You may not rely on the assumptions regarding to this issue (experience of your friends or your experience in other courses). If you have any questions, ask them to the instructor. Similarly, if any of these rules is not clear for you, ask what they mean to the instructor.

We check similar pairs of assignments ourselves carefully and make our own judgment. *Students caught cheating on assignments will be subject to disciplinary action.* Do not forget that it is much easier to explain a poor grade to others than to explain a cheating conviction.

***As a Hacettepe student registered to ECO240, we assume that you have read this document and fully understand your responsibilities.***

***"By signing below, I confirm that I have read this document and fully understood my responsibilities. I hereby declare that I am not violating any academic integrity."***

ID	Name	Signature
21941688	Büşra Gülerüz	

# “ORANGE” DATASET

The dataset that is going to be analyzed is called "[Orange](#)". It is a data frame that has 35 rows and 3 columns of records of the growth of orange trees. 3 variables are: "tree", "age", "circumference". The aim of this data set is to show whether the circumference of a given tree gets larger as the tree gets older. "tree" variable consists of 5 different numbers: 1,2,3,4,5. These are the numbers given to recognize each tree, meaning that 5 trees were observed in order to make this research. Each tree's circumference levels are given with their age.

My research questions are,

1. To find whether a tree's circumference gets larger as it gets older.
2. If so, which tree gets more larger (than others do) as it gets older?

## VARIABLES

- Tree: independent variable. Number of a tree is basically its ID.
- Age: independent variable. It indicates the age (in days) of a given tree.
- Circumference: I am trying to prove that this variable dependent; and I am going to do my analyses assuming that it is dependent and interpret the results to prove. The measurement is mm.

## NUMERICAL AND GRAPHICAL SUMMARIES

```
> cor(Orange$age,Orange$circumference)
```

```
[1] 0.9135189
```

0.91 is a very high number for a correlation level. This result shows us that, circumference of a tree, indeed, is related to its age.

```
> cov(Orange$age,Orange$circumference)
```

```
[1] 25831.02
```

Also, covariance value is positive. This also shows us that the age and circumference variables are related positively.

Now I am excluding the first column “tree” because I am going to take the summary() of the data which summarizes it numerically, and tree variable is more of a categorical variable rather than numerical.

```
> orange_notree <- Orange[,-1]
> summary(orange_notree)
      age      circumference
Min.   :118.0    Min.   : 30.0
1st Qu.:484.0    1st Qu.: 65.5
Median :1004.0   Median :115.0
Mean   :922.1    Mean   :115.9
3rd Qu.:1372.0   3rd Qu.:161.5
Max.   :1582.0   Max.   :214.0
```

The youngest tree in search is 118 days old and the oldest one is 1582 days old. The smallest circumference is 30 mm where largest is 214 mm. The variables are given in quite wide ranges.

In order to detect the tree with the *highest rate of growth*, I am going to create new data tables for each tree and summarize them. (It is important to note that there are 7 observations for each tree.)

```
> library(tidyverse)

> orange1 <- Orange %>%
+   filter(Tree==1)

> orange2 <- Orange %>%
+   filter(Tree==2)

> orange3 <- Orange %>%
+   filter(Tree==3)

> orange4 <- Orange %>%
+   filter(Tree==4)

> orange5 <- Orange %>%
+   filter(Tree==5)
```

```
> summary(orange1)
```

Tree	age	circumference
3:0	Min. : 118.0	Min. : 30.00
1:7	1st Qu.: 574.0	1st Qu.: 72.50
5:0	Median :1004.0	Median :115.00
2:0	Mean : 922.1	Mean : 99.57
4:0	3rd Qu.:1301.5	3rd Qu.:131.00
	Max. :1582.0	Max. :145.00

```
> summary(orange2)
```

Tree	age	circumference
3:0	Min. : 118.0	Min. : 33.0
1:0	1st Qu.: 574.0	1st Qu.: 90.0
5:0	Median :1004.0	Median :156.0
2:7	Mean : 922.1	Mean :135.3
4:0	3rd Qu.:1301.5	3rd Qu.:187.5
	Max. :1582.0	Max. :203.0

```
> summary(orange3)
```

Tree	age	circumference
3:7	Min. : 118.0	Min. : 30
1:0	1st Qu.: 574.0	1st Qu.: 63
5:0	Median :1004.0	Median :108
2:0	Mean : 922.1	Mean : 94
4:0	3rd Qu.:1301.5	3rd Qu.:127
	Max. :1582.0	Max. :140

```
> summary(orange4)
```

Tree	age	circumference
3:0	Min. : 118.0	Min. : 32.0
1:0	1st Qu.: 574.0	1st Qu.: 87.0
5:0	Median :1004.0	Median :167.0
2:0	Mean : 922.1	Mean :139.3
4:7	3rd Qu.:1301.5	3rd Qu.:194.0
	Max. :1582.0	Max. :214.0

```
> summary(orange5)
```

Tree	age	circumference
3:0	Min. : 118.0	Min. : 30.0
1:0	1st Qu.: 574.0	1st Qu.: 65.0
5:7	Median :1004.0	Median :125.0
2:0	Mean : 922.1	Mean :111.1
4:0	3rd Qu.:1301.5	3rd Qu.:158.0
	Max. :1582.0	Max. :177.0

Starting from these numerical summaries, we can say that

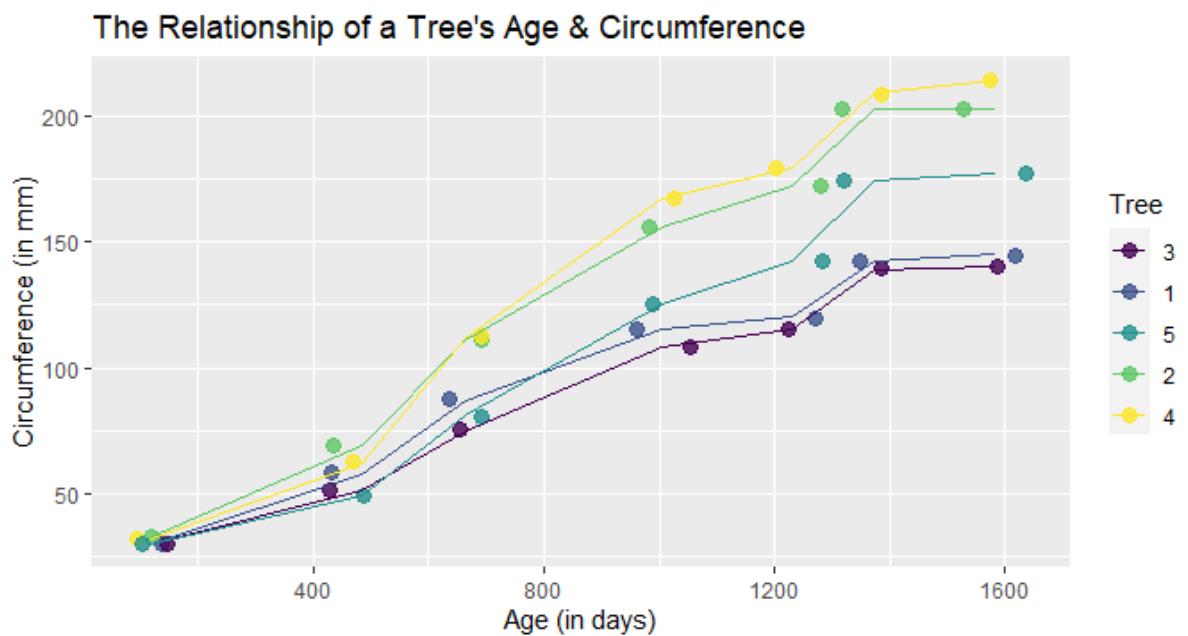
1. Starting ages of trees are the same. Minimum sizes of trees 1, 3 and 5 are 30 while it is 33 for tree 2 and 32 for tree 4.
2. Their final ages are also the same, with 1582 days; making it easier to compare their growth speed.
3. The maximum sizes of trees are as follows (from lowest to highest):
  1. Tree 3: 140 mms
  2. Tree 1: 145 mms
  3. Tree 5: 177 mms
  4. Tree 2: 203 mms
  5. Tree 4: 214 mms

Here now I am going to plot a scatterplot and add a linear line to dictate the positive relationship of age and circumference and I am also going to group the trees by coloring them individually to compare their growth rate.

```
> library(ggplot2)

> plot1 <- ggplot(Orange, aes(age, circumference, col=Tree)) +
  geom_jitter(alpha=.8,size=3) + geom_line() + labs(x="Age (in days)", y = "Circumference (in
  mm)") + ggtitle("The Relationship of a Tree's Age & Circumference")

> plot1
```



Here, we prove our hypothesis visually. As can be seen above, the relationship between age and circumference is positive. Also, Tree 4 has the steepest line where Tree 3 has the flattest one. meaning that Tree 4 has the highest rate of growth and Tree 3 has the lowest. The order is also made in the legend from lowest to highest.

## REGRESSION ANALYSIS

**Explanatory variable:** Age.

**Response variable:** Circumference.

- I am creating a model to use for regression:

```
> model_agevscirc <- lm(circumference~age,data=Orange)
> model_agevscirc
```

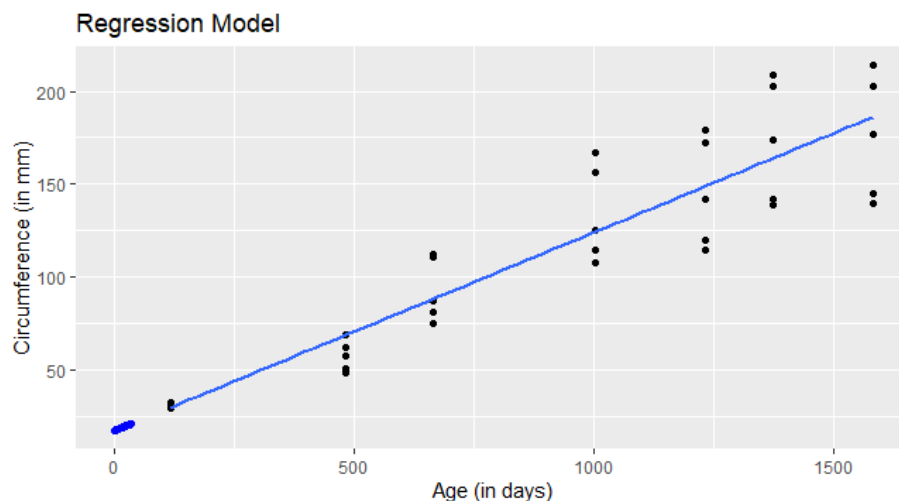
Call:

```
lm(formula = circumference ~ age, data = Orange)
```

Coefficients:

(Intercept)	age
17.3997	0.1068

```
> ggplot(Orange, aes(age, circumference)) + geom_point() + geom_smooth(method =
"lm" , se =FALSE) + geom_point( data = prediction_data, color ="blue")+labs(x="Age (in
days)",y="Circumference (in mm)") + ggtitle("Regression Model")
```



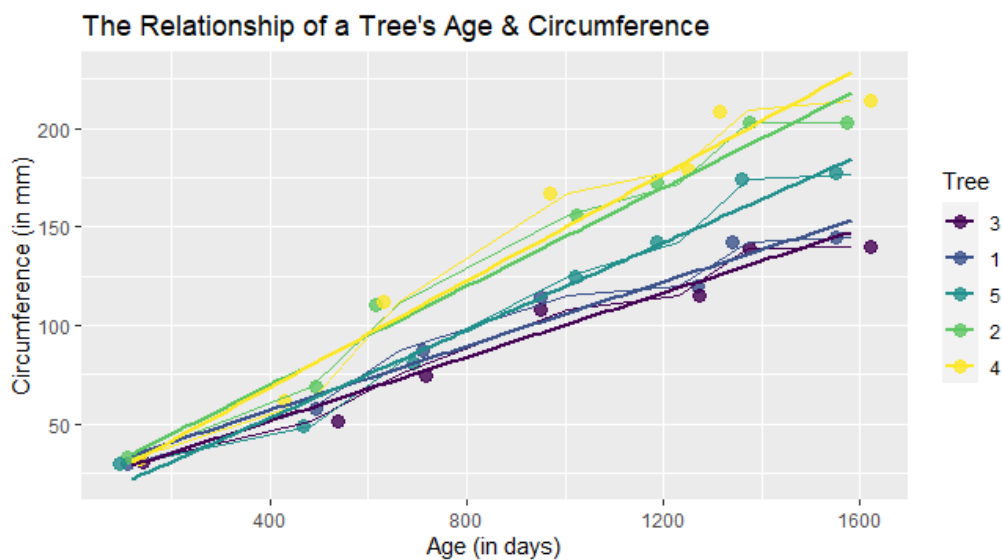
This linear model proves that our model has worked correctly. Circumference gets larger as the tree is older.

```
> explanatory_data <- tibble(age=1:35)

> predict(model_agevsirc,explanatory_data)
```

1	2	3	4	5
17.50642	17.61319	17.71996	17.82673	17.9335
6	7	8	9	10
18.04027	18.14704	18.25381	18.36058	18.46735
11	12	13	14	15
18.57412	18.68089	18.78766	18.89443	19.00121
16	17	18	19	20
19.10798	19.21475	19.32152	19.42829	19.53506
21	22	23	24	25
19.64183	19.7486	19.85537	19.96214	20.06891
26	27	28	29	30
20.17568	20.28245	20.38922	20.49599	20.60276
31	32	33	34	35
20.70953	20.8163	20.92307	21.02984	21.13661

```
> plot2 <- plot1 + geom_smooth(method = "lm" , se=FALSE )
> plot2
```



This adjustment also shows us that this linear model is true as well, not only the circumference gets larger as the tree is older; but also proves that Tree 4 has the highest growth rate while Tree 3 has the lowest.

```
> summary(model_agevscirc)
```

Call:

```
lm(formula = circumference ~ age, data = Orange)
```

Residuals:

Min	1Q	Median	3Q	Max
-46.310	-14.946	-0.076	19.697	45.111

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	17.399650	8.622660	2.018	0.0518 .
age	0.106770	0.008277	12.900	1.93e-14 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.74 on 33 degrees of freedom

Multiple R-squared: 0.8345, Adjusted R-squared: 0.8295

F-statistic: 166.4 on 1 and 33 DF, p-value: 1.931e-14

- Residuals are the difference between the actual observed response values and the response values that the model predicts. Response value is circumference and minimum level for it is 30 and maximum is 214.

The residual data is not exactly symmetric, but it is very close to be symmetric – so our model worked well.

- Coefficients are two unknown constants that represent the *intercept* and *slope* terms in the linear model.

The coefficient estimate contains two rows; the first one is the intercept.

**The Intercept** (often labeled the constant) is the expected mean value of circumference when all trees' ages equal to zero. It is 17.4 in this example. The second row in the coefficients is the slope.

**Slope** is 0.106. The slope is numerically very small because age can go up to thousands (because it is measured in days) and circumference only goes up to 214. But we shouldn't let it confuse us, because as calculated before, the correlation coefficient is 0.91, which proves the strong positive relationship between the variables.

- The coefficient **Standard Error** measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We want this to



be small for the consistency of our results. In this example, it is 0.008, which is pretty small of a number.

- The coefficient **t-value** is a measure of how many standard deviations our coefficient estimate is far away from 0. We want the t-value to be far from zero as this would indicate that we could reject the null hypothesis, meaning that we could actually declare a relationship between age and circumference. In our example, t-values are far from zero. t-values are also used in calculating the p-values.
- The **Pr(>t) acronym** found in the model relates to the probability of observing any value equal or larger than t. A small p-value tells us that it is unlikely to observe a relationship between the predictor (age) and response (circumference) variables. A small p-value for the intercept and the slope proves that we can reject the null hypothesis, and this allows us to declare a relationship between age and circumference.
- **Residual standard error** is measure of the quality of a linear regression fit. Every linear model is assumed to have an error term “E”. because of the existence of this error term, we are not capable of perfectly predicting our response variable (circumference) from the predictor (age). The residual standard error is 23.74 on 33 degrees of freedom.

## CONCLUSION

Consequently, all of my analyses show that the circumference of a tree gets larger as the tree gets older. This is a biological fact; indeed, the age of a tree can be calculated from its trunk size. Each year, the tree's trunk gets larger by 1 ring. So that as the tree gets 1 year old, its trunk expands with 1 ring added. Its age is calculated by counting at the number of rings in the trunk of the tree. Trees expand ring-to-ring within the time period they spend. This is a complete



database of rings. The historical identity of the tree lies within these rings. These rings are called growth rings in the scientific world. These rings are all about the seasons. Like the first spring, when nature blooms, trees sprout, the growth rate of the rings is

at the maximum level. This rapid growth causes the color of the tree rings to become lighter. But after entering the summer period, this growth slows down considerably. Here, too, the colors of the Rings begin to get darker and darker. Since the development process of the Rings depends entirely on the seasons, we can say that it takes 1 year to complete the ring. So, the sum of these lines tells us how old the tree is. <sup>1</sup>

---

<sup>1</sup> Ağaçların Yaşı Nasıl Hesaplanır, <https://www.bilgeyik.com/agaclarin-yasi-nasil-hesaplanir-360>.