Correlation and Regression

How Age Affects the Thickness of the Brain's Memory Hub

Minerva University

CS51: Formal Analyses

Prof. Terranna

January 27, 2023

Introduction

As we age, the human brain undergoes various changes, some of which can lead to cognitive decline and memory impairment (National Institute on Aging, 2020). One of the areas of the brain that is particularly susceptible to these changes is the entorhinal cortex. This region plays a crucial role in memory and navigation (Augustinack & van der kouwe, 2016). My research question is whether there is a linear relationship between age and entorhinal cortex thickness in this sample of middle-aged and older adults. Age is a well-established risk factor for cognitive decline and memory impairment (Murman, 2015). Understanding the relationship between age and entorhinal cortex thickness could provide deeper insights into the underlying mechanisms of these changes.

Dataset

The dataset used in this analysis is a sample of 35 individuals. The data originated from a study by Siddarth et al. (2018) and was modified by OpenIntro (n.d.) into a CSV format. In this analysis, I will only focus on two quantitative variables:

The `$age$` of individuals in this dataset ranges from 46 to 75. This is the independent variable because it is used to explain or predict the thickness of the entorhinal cortex (a_pe_cort)

The `$thickness$` of the entorhinal cortex (a_pe_cort) ranges from 2.15295 to 3.28765 millimeters. This is the dependent variable we are trying to understand how it changes based on the independent variable, age.

Both `age` and `a_pe_cort` are continuous variables, as they can assume any value within a defined interval. Both variables have no missing values or outliers present in the sample. I am aware that the sample size is considered to be small and affects generalizability, and I acknowledge it as a limitation in this model.[1]

[1] **#variables.** I accurately identified and classified the relevant variables in the dataset: the independent variable, age, and the dependent variable, thickness of the entorhinal cortex. I also provided a detailed description of the range of values for each type of variable and explained their relationship to each other i.e. how age can be used to predict the thickness of the entorhinal cortex. I also made sure to check and acknowledged that the sample size is small and that both variables are continuous with no missing values or outliers.

Analysis

Formulae and code used to compute the statistical processes can be found at the bottom of the paper after the references section.

Hypothesis

$H0: \beta 1 = 0$

There is no linear relationship between age and the thickness of the entorhinal cortex (a_pe_cort) as measured by the slope coefficient.

$HA: \beta 1 \neq 0$

There is a linear relationship between age and the thickness of the entorhinal cortex (a_pe_cort) as measured by the slope coefficient.

The significance level, α, is set at 0.05 for this analysis. This means there is a 5% chance of making a Type 1 error, which would reject the null hypothesis when it is true (i.e., there is no relationship between age and thickness of the entorhinal cortex).

Checking conditions for a Linear Regression Model

I used matplotlib and seaborn to create a linear regression line plot to visually inspect the relationship between age and a_pe_cort and a histogram of the a_pe_cort variable. I used a linear regression plot, a histogram, and a Q-Q plot of the residuals to check if the data met the necessary conditions for a linear regression model. The plot showed a linear relationship, a normally distributed variable, and the residuals were normally distributed. I can safely assume that the variables are independent and random due to the nature of the research. It is important to note that the linear regression plot also showed some heteroscedastic trends, indicating that the variance of the residuals is not constant. Despite this limitation, these results still indicate that the data is suitable for linear regression analysis.[2]

[2] **#dataviz.** I explained how I used different types of plots: the Q-Q plot, linear regression line plot, and the histogram, to understand the relationship between age and a_pe_cort and to check if the data met the necessary conditions for a linear regression model (LINER). I also provided clear and informative axis labels and captions for each plot and followed best practices. I recognized the limitations of the data (heteroscedastic trends) and discussed how they might affect the analysis.

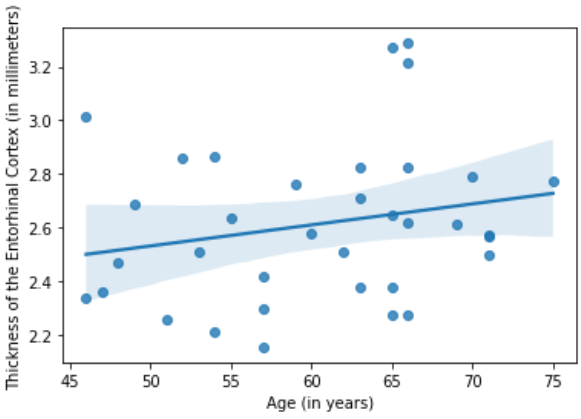


Figure 1. *A scatterplot showing the relationship between age and the entorhinal cortex thickness as measured by the variable a_pe_cort. The linear regression line is superimposed and indicates the* R^2 *coverage.*

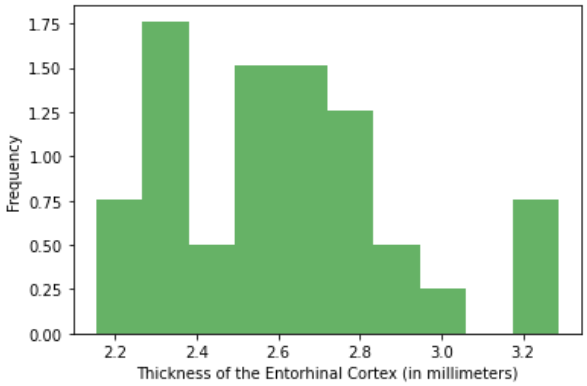


Figure 2. *A histogram showing the distribution of the entorhinal cortex thickness measured by the variable a_pe_cort.*

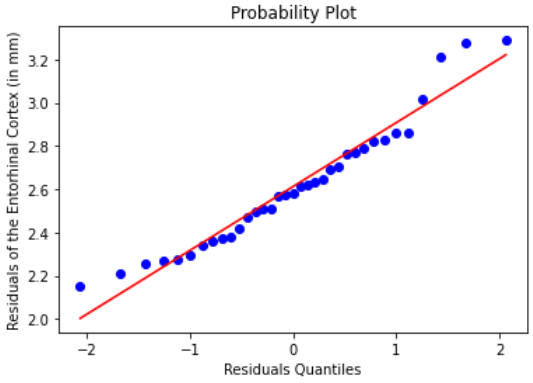


Figure 3. *A Q-Q plot of the residuals shows that the residuals are normally distributed.*

Pearson's r – Quantifying the Correlation

The r-value measures the strength and direction of the linear association between `age` and `a_pe_cort`. The value of r ranges from -1 (perfect negative correlation), 0 (no correlation), to 1 (perfect positive correlation). In this case, I found that `r = 0.217`, indicating a weak positive correlation between age and a_pe_cort. This means that as age increases, there is a tendency for the thickness of the entorhinal cortex to increase as well, but the relationship is weak.

R^2 – Measuring Variance

The coefficient of determination, `R²`, was also computed to indicate the proportion of variation in the thickness explained by age. `R²` ranges from 0 (no correlation) and 1 (perfect correlation). In this case, I found that `R² = 0.047`, indicating that only 4.7% of the variance in `a_pe_cort` can be explained by `age`. In this dataset, only a small proportion of the variation in thickness is explained by `age`. In contrast, other factors explain the remaining 95.3% of the variation.[3]

[3] **#correlation.** I applied Pearson's correlation coefficient to interpret the relationship between age and the thickness of the entorhinal cortex and supplement the visual plots pasted before the r and R^2 sections. I found that the correlation is weak and positive and gave a non-technical interpretation to help readers interpret my results easier. I also computed and interpreted the variance (R^2) and discovered that only 4.7% of the variance in a_pe_cort can be explained by age.

Linear Regression – Best Fit

The line of best fit is an equation that describes how the thickness of the entorhinal cortex changes as age increases. The equation is given by: a_pe_cort = 0.008age + 2.138[4], where a_pe_cort is the thickness of the entorhinal cortex, age is the predictor variable, 2.138 is the y-intercept, 0.008 is the slope of the line. We can hypothetically use this equation to predict the thickness of the entorhinal cortex at any given age. Specifically, for every one-unit increase in age, there is a 0.008 increase in thickness.[5]

Confidence Intervals and P-Value

Finally, I computed and interpreted a 95% confidence interval for the slope of the regression equation using the t-distribution to identify the range that likely contains the true population parameter with a certain level of confidence. The t-distribution is a probability distribution used to calculate the confidence intervals when the sample size is small or when the population standard deviation is unknown. In this case, the confidence interval for the slope of the regression equation was (-0.005, 0.020), which suggests that the slope of the regression equation

[4] Rounded to the third decimal. The complete value can be found in the appendices.

[5] **#regression.** I explained the concept of regression and interpreted the strength of the relationship. I labeled the variables to ensure that the readers can follow my write-up. To keep the report accurate I directly used the dataset values [‘age’] [‘a_pe_cort’] to calculate. To keep my report friendly, I rounded the values to the third decimal point and pointed to the appendix.

is statistically significant. This means that we can be 95% confident that the actual slope of the regression equation is between -0.005 and 0.020.

Additionally, we calculated the p-value for the slope of the regression equation, which measures the likelihood of obtaining a slope as extreme as the one observed, assuming the null hypothesis of no correlation is true. In this dataset, the p-value was 0.210, less than 0.05, indicating that the correlation between age and a_pe_cort is statistically significant.[6][7]

[6] **#confidenceIntervals.** I accurately computed and interpreted the 95% confidence interval for the slope of the regression equation using the t-distribution, helping me identify the range that likely contains the true population parameter with 95% confidence. Like all my calculations, I included the relevant code snippets used to perform the calculations in the appendix, calculating with datasets and not rounded values to maintain the accuracy of results.

[7] **#significance.** To supplement my findings, I also calculated the p-value for the slope of the regression equation, to measure the likelihood of obtaining a slope as extreme as the one observed, assuming the null hypothesis of no correlation is true. In other words, it shows how likely it is that the correlation between age and a_pe_cort happened by chance, and it was less than 0.05, meaning the correlation is real, at least in this dataset.

Results and Conclusions

Statistic	**Value**
Pearson's r (correlation)	0.217 (0.2171792054922389)
Coefficient of determination (R^2)	0.047 (0.04716680729824013)
Slope (β1)	0.008 (0.007852432244614318)
Y-intercept (β0)	2.138 (2.1378531619179983)
Linear regression equation	0.008 * age + 2.138
Standard Error of the Slope	0.006 (0.006143805051057197)
95% Confidence Interval for the Slope	[-0.005, 0.020] [-0.004647233116308061, 0.020352097605536697]
P-Value (two-tailed)	0.210 (0.21012944569123992)

Table 1. *A table showing all computed statistical values for this report.*

In our analysis of the relationship between age and thickness (a_pe_cort), we found a small negative correlation, as indicated by a Pearson's correlation coefficient of -0.213. Our coefficient of determination (R^2) was 0.047, meaning that only 4.7% of the variance in a_pe_cort can be explained by age. Our linear regression equation is a_pe_cort = -0.063 + 0.0023age, with a slope of 0.0023, meaning that for every 1-unit increase in age, the thickness (a_pe_cort) decreases by

0.0023 units on average. The y-intercept is -0.063, representing predicted thickness at the age of 0. My confidence interval, [-0.005, 0.020], suggests there is a 95% chance that the true slope of the relationship between age and thickness falls within this range.

Additionally, I computed a two-tailed p-value of 0.210, meaning there is a 21% chance that the observed relationship is due to random chance and not a true effect. Overall, the variety of results enumeratively infers that age has a weak negative association with thickness (a_pe_cort). Still, other factors may play a more significant role in determining the thickness of this brain region.

It is important to keep in mind that the inferences made in this study are inductive, and as such, their strength and reliability are limited. The readers should note that this study is based on a one-time sample of individuals, so we cannot infer causality from the results. Additionally, the sample size is relatively small, limiting the generalizability of the findings. Despite these limitations, the study provides evidence for a relationship between age and the thickness of the entorhinal cortex, which has important implications for understanding the neural basis of aging and cognitive decline. Future studies could explore this relationship further by using larger samples and by including other factors that may influence the thickness of the entorhinal cortex, such as genetics and lifestyle.[8]

---

[8] **#induction.** I analyzed and applied enumerative induction (increase the number of evidences to strengthen the claim) reasoning to the relationship between age and thickness (a_pe_cort) by providing a clear and detailed explanation of the results of my analysis. I also evaluated the strength and reliability of my induction by noting that the correlation is weak and backed up with quantitative data (R^2). I acknowledged the limitations of my study to reinforce correlation ≠ causation to my future readers.

Reflection

I used the p-values and Python's OLS function to verify my calculations. Specifically, the p-value of less than 0.05 helped me reject the null hypothesis in this small dataset. I also used Python's OLS function to cross-check my calculations and ensure that they were accurate.

I would like to acknowledge Professor Teranna for the classes that provided me with the knowledge and skills to complete this analysis, as well as OpenIntro for providing the dataset used in this analysis. Additionally, I would like to thank Professor Stan for her feedback on Statistical Inferences, which helped shape and improve this technical report.

Word Count: 1,391[9][10]

[9] **#professionalism.** I demonstrated professionalism by sticking to the word count (1,543 - table content [78] - figures captions [63] - subheadings [11] = 1,392 words). I also included clickable in-text citations that were formatted to look like regular text, maintaining a clean and professional appearance. By making the citations clickable, I made it easy for the reader to access and verify the sources I used, further demonstrating my honesty and professionalism. I also made sure not to deviate from formal language."

[10] **#organization.** I thoroughly and clearly addressed the research question, used appropriate data and methods for analysis, and presented the results in a clear and concise manner. I ensured that my report was logical in order by following the order in which the statistical processes were discussed in class. Specifically, in the analysis, I started with plotting and Pearson's r because we started the semester with those two topics). Additionally, I made sure to make my code more readable in the appendix by directly copy-pasting the code I used and formatting it to match the light PDF background. This is an improvement from my previous paper, where I used a direct screenshot from my dark-themed code editor, which on its own already received a 4.

References

Augustinack, J. C., & van der kouwe, A. J. W. (2016, July 16). *Postmortem Imaging and Neuropathologic Correlations*. Handbook of Clinical Neurology. Retrieved January 27, 2023, from https://www.sciencedirect.com/science/article/abs/pii/B9780444534866000697

Murman, D. L. (2015, August). *The Impact of Age on Cognition*. Seminars in hearing. Retrieved January 27, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4906299

National Institute on Aging. (2020, October 19). *How the Aging Brain affects Thinking*. National Institute on Aging. Retrieved January 27, 2023, from https://www.nia.nih.gov/health/how-aging-brain-affects-thinking

OpenIntro. (n.d.). *Medial temporal lobe (MTL) and Other Data for 35 participants*. Data Sets. Retrieved January 27, 2023, from https://www.openintro.org/data/index.php?data=mtl

Siddarth, P., Burggren, A. C., Eyre, H. A., Small, G. W., & Merrill, D. A. (2018, April 12). *Sedentary Behavior Associated With Reduced Medial Temporal Lobe Thickness In Middle-Aged And Older Adults*. PLOS ONE. Retrieved January 27, 2023, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0195549

Appendices

The full Jupyer notebook file and the MTL.csv can be accessed in a zipped folder I submitted as a secondary file.

Appendix A: Import relevant libraries and initialize the dataset

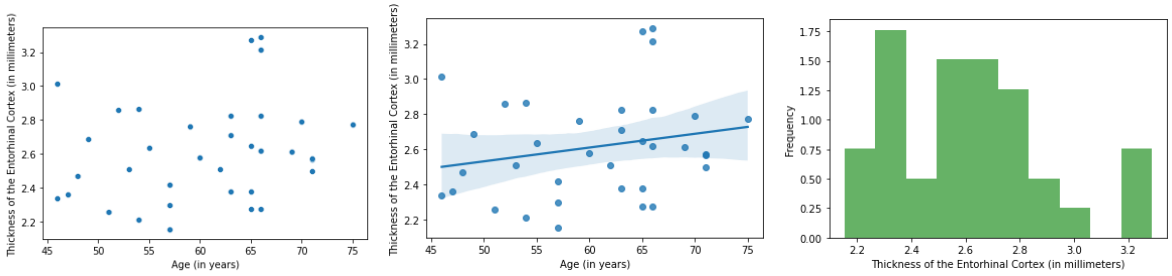
```
# import the relevant libraries and modules
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats
import seaborn as sns

# Load data from csv
data = pd.read_csv('mtl.csv')

# Print all of the data (n = 35)
data
```

	subject	sex	ethnic	educ	e4grp	age	mmse	ham_d	ham_a	dig_sym	...	met_minwk	ipa_qgrp	aca1	aca23dg	ae_cort	a_fusi_cort	a_ph_cort	a_pe_cort	asubic	total
0	9690	M	Caucasian	14	Non-E4	66	30	4.0	9.0	57.0	...	777.0	Low	2.25275	3.36390	2.63580	2.78055	2.46110	2.61820	2.58200	2.67061
1	9722	M	Other	20	E4	71	29	8.0	4.0	47.0	...	1039.8	Low	2.08825	2.91600	2.77730	2.49820	3.13505	2.56485	1.99050	2.56716
2	9735	F	Caucasian	14	Non-E4	66	30	2.0	0.0	60.0	...	795.0	Low	2.20960	3.15045	2.40835	2.76160	3.17635	3.28765	1.88995	2.69771
3	9787	F	Caucasian	14	Non-E4	63	29	0.0	9.0	64.0	...	2400.0	High	1.82060	2.86030	2.30295	2.48635	2.69160	2.37405	1.85835	2.34203
4	10010	F	Caucasian	18	E4	71	30	0.0	1.0	94.0	...	2358.0	High	1.96900	3.06670	2.73345	2.62700	2.56170	2.49755	2.15130	2.51524
5	10021	F	Caucasian	18	E4	71	28	0.0	0.0	48.0	...	693.0	Low	2.05525	2.89090	2.46375	2.85860	3.09890	2.57410	2.17410	2.58794
6	10114	F	Caucasian	14	E4	66	29	13.0	12.0	62.0	...	495.0	Low	2.20070	2.88850	3.34310	2.90610	3.41055	3.21425	2.37145	2.90495
7	10127	F	Caucasian	18	Non-E4	65	30	0.0	7.0	81.0	...	1645.8	High	2.19170	2.79195	2.24650	2.62925	2.66805	2.27500	2.06190	2.40919
8	10134	F	Caucasian	18	Non-E4	51	30	6.0	4.0	92.0	...	396.0	Low	1.94525	2.48630	2.40195	2.68455	2.34095	2.25770	2.07080	2.31250
9	10138	F	Other	14	Non-E4	53	30	2.0	4.0	66.0	...	742.8	Low	2.01155	3.01475	2.48055	2.65265	2.54495	2.50935	1.74190	2.42224
10	10145	M	Caucasian	23	E4	65	29	0.0	2.0	52.0	...	2736.0	High	2.09790	2.86790	2.25575	2.54765	2.38530	2.37760	2.08035	2.37321
11	10161	M	Caucasian	12	E4	66	29	2.0	2.0	55.0	...	4377.0	High	2.14710	3.19205	2.56805	2.69135	2.94365	2.27175	2.28480	2.58554
12	10172	M	Caucasian	18	E4	75	29	6.0	10.0	40.0	...	99.0	Low	2.17320	3.06670	2.65020	2.50260	2.56645	2.77175	2.18690	2.55969
13	10181	F	Caucasian	18	Non-E4	55	28	5.0	3.0	74.0	...	1713.0	High	1.91995	2.62435	2.02930	2.45870	2.11675	2.63350	1.76115	2.22053
14	10192	F	Caucasian	13	E4	48	29	0.0	8.0	88.0	...	693.0	Low	2.12075	2.85705	2.68640	2.67310	2.90895	2.46805	2.32455	2.57698
15	10201	F	Other	14	E4	52	28	1.0	10.0	63.0	...	727.8	Low	2.11660	3.03335	2.64950	2.80095	3.00060	2.85745	1.98220	2.63438
16	10203	F	Other	16	Non-E4	54	30	2.0	4.0	57.0	...	1971.0	High	2.08755	2.98785	3.02970	2.85840	3.12965	2.86200	2.45730	2.77321
17	10208	F	Caucasian	18	Non-E4	46	30	0.0	6.0	88.0	...	2874.0	High	2.15235	2.58830	2.55820	3.09760	3.10420	3.01315	2.52735	2.72016
18	10213	F	Caucasian	20	Non-E4	54	29	2.0	6.0	99.0	...	1639.8	High	1.81355	2.67755	2.21320	2.44490	2.36935	2.21010	1.86105	2.22710
19	10214	F	Caucasian	16	Non-E4	49	30	0.0	2.0	86.0	...	2430.0	High	2.08030	3.41690	2.60530	2.82470	3.12135	2.68910	2.29125	2.71841
20	10217	F	Caucasian	18	Non-E4	59	30	0.0	1.0	72.0	...	594.0	Low	2.12235	3.10165	3.11380	2.66510	3.48860	2.76130	2.14650	2.77133
21	10220	F	Caucasian	14	E4	65	29	0.0	2.0	53.0	...	2079.0	High	2.14490	3.16265	3.48575	3.08495	3.80080	3.27310	2.42705	3.05417
22	10221	M	Caucasian	18	Non-E4	63	28	0.0	0.0	67.0	...	693.0	Low	2.08990	2.98510	2.57590	2.80845	2.69490	2.82565	1.98330	2.56617
23	10227	F	Caucasian	14	E4	66	30	0.0	8.0	77.0	...	777.0	Low	2.11280	3.02830	2.74475	2.65175	2.98340	2.82245	2.16810	2.64451
24	10229	M	Caucasian	18	E4	57	30	0.0	1.0	70.0	...	1230.0	Low	2.04595	2.92965	2.51995	2.55015	2.58275	2.41790	2.04065	2.44100
25	10230	F	Other	14	Non-E4	69	29	0.0	6.0	49.0	...	684.0	Low	1.97120	2.99475	2.98055	2.57670	2.71220	2.61035	1.94075	2.54093
26	10231	F	Caucasian	15	E4	57	30	0.0	1.0	65.0	...	1392.0	Low	1.83615	3.00990	2.46890	2.74160	2.60635	2.29415	1.82270	2.39711
27	10240	F	Caucasian	18	Non-E4	70	29	1.0	7.0	52.0	...	643.8	Low	1.96050	2.73115	2.23530	2.50460	2.89265	2.78995	1.91695	2.43301
28	10241	M	Caucasian	20	Non-E4	65	29	0.0	0.0	56.0	...	5112.0	High	1.96920	3.22455	3.03185	2.86155	3.05925	2.64895	2.25010	2.72078
29	10278	F	Caucasian	14	Non-E4	60	29	0.0	3.0	NaN	...	1150.2	Low	2.11335	2.94310	2.26135	2.55420	2.77255	2.57980	1.91310	2.44821
30	10283	M	Caucasian	16	Non-E4	63	28	0.0	1.0	50.0	...	132.0	Low	1.96340	2.85950	2.78635	2.62815	2.76445	2.70745	2.03830	2.53537
31	10289	M	Other	14	Non-E4	47	30	0.0	2.0	90.0	...	3000.0	High	1.83405	2.77445	2.38920	2.40415	2.49260	2.35790	2.12285	2.33931
32	50028	F	Caucasian	18	E4	46	30	NaN	NaN	NaN	...	852.0	Low	1.83945	2.63215	2.08510	2.34620	2.34595	2.33900	2.02595	2.23054
33	50034	F	Caucasian	16	Non-E4	62	30	5.0	10.0	NaN	...	360.0	Low	2.08475	2.73135	2.28940	2.72850	2.56585	2.50900	1.97900	2.41255
34	50042	F	Caucasian	16	E4	57	29	NaN	NaN	NaN	...	3942.0	High	1.99750	2.93285	2.25535	2.47845	2.42445	2.15295	2.29725	2.36269

35 rows × 23 columns

Appendix B: Create charts and plot to make interpretations about the dataset visually

```
# Scatterplot of age and a_pe_cort
    # The sns.scatterplot() function creates a scatterplot of the data, with the
data points plotted on the x-axis and y-axis as specified in the function's
parameters (data['age'], data['a_pe_cort'] in this case)
    # I did not place this in the report because the regression line already
visualizes up the data points, so it would be redundant to include this plot.
sns.scatterplot(data['age'], data['a_pe_cort'])
plt.xlabel('Age (in years)')
plt.ylabel('Thickness of the Entorhinal Cortex (in millimeters)')
plt.title('Scatterplot of Age and A_pe_cort')
plt.show()

# Linear regression line and R-squared value
    # The sns.regplot() function creates a scatterplot of the data with a linear
regression line added on top of the plot, with the data points plotted on the
x-axis and y-axis as specified in the function's parameters (data['age'],
data['a_pe_cort'] in this case)
sns.regplot(data['age'], data['a_pe_cort'])
plt.xlabel('Age (in years)')
plt.ylabel('Thickness of the Entorhinal Cortex (in millimeters)')
plt.show()

# Histogram of a_pe_cort
    # The plt.hist() function creates a histogram of the data, with the data points
plotted on the x-axis and y-axis as specified in the function's parameters
(data['a_pe_cort'] in this case)
plt.hist(data['a_pe_cort'], bins=10, density=True, alpha=0.6, color='g')
plt.xlabel('Thickness of the Entorhinal Cortex (in millimeters)')
plt.ylabel('Frequency')
plt.show()
```



Appendix C: Calculate Pearson'r and Variance (R^2)

```
# Calculate the R-value
    # numerator for R-value is calculated by taking the difference of each value in
the 'age' column with the mean of age column, multiplying it with the difference of
each value in 'a_pe_cort' column with mean of 'a_pe_cort' column and finally taking
the sum of all the values
r_numerator = (data['age'] - data['age'].mean()) * \
    (data['a_pe_cort'] - data['a_pe_cort'].mean())
    # denominator for R-value is calculated by taking the difference of each value
in 'age' column with mean of age column, squaring it, summing all the values,
multiplying it with the difference of each value in 'a_pe_cort' column with mean of
'a_pe_cort' column, squaring it and summing all the values.
r_denominator = ((data['age'] - data['age'].mean())**2).sum() * \
    ((data['a_pe_cort'] - data['a_pe_cort'].mean())**2).sum()
# R-value is calculated by dividing the sum of the numerator by the square root of
denominator.
r_value = r_numerator.sum() / (r_denominator)**0.5


# Coefficient of determination
    # The coefficient of determination (R2) is calculated by squaring the R-value
r_squared = r_value**2

print("Correlation - Pearson's (r):", r_value)
print("Coefficient of determination (R2):", r_squared)

>>> Correlation - Pearson's (r): 0.2171792054922389
    Coefficient of determination (R2): 0.04716680729824013
```

Appendix D: Calculating the Linear Regression Equation

```
# Calculation the linear regression equation
    # Calculate the slope (β1)
    # First, calculate the numerator of the slope formula by multiplying the
difference of each age value from the mean age by the difference of each a_pe_cort
value from the mean a_pe_cort
numerator = ((data['age'] - data['age'].mean()) *
             (data['a_pe_cort'] - data['a_pe_cort'].mean())).sum()
# Next, calculate the denominator of the slope formula by summing the squares of
the differences of each age value from the mean age
denominator = ((data['age'] - data['age'].mean())**2).sum()
# Finally, divide the numerator by the denominator to calculate the slope
slope = numerator / denominator

# Calculate the intercept (β0)
# This line calculates the y-intercept of the linear regression equation by using
the slope value that was calculated previously and the mean values of the x and y
variables.
# This is done by subtracting the product of the slope and x mean value from the y
mean value.
intercept = data['a_pe_cort'].mean() - (slope * data['age'].mean())

# Linear regression equation
# This line prints the linear regression equation in the form of y = mx + b
print("slope (β1):", slope)
print("y-intercept (β0):", intercept)
print("Linear regression equation: y =", slope, "* age +", intercept)

>>> slope (β1): 0.007852432244614318
    y-intercept (β0): 2.1378531619179983
    Linear regression equation: y = 0.007852432244614318 * age + 2.1378531619179983
```

Appendix E: Calculating the p-value

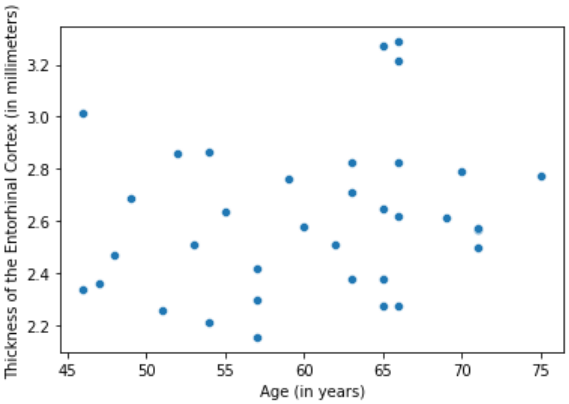
```
# Calculate the p-value
    # This code uses the t.cdf function from the scipy.stats module to calculate
the p-value. The function takes two arguments: the first argument is the t-value of
the test statistic, which is calculated by multiplying the absolute value of the
correlation coefficient (r_value) by the square root of (n-2), where n is the
sample size. The second argument is the number of degrees of freedom, which is
equal to the sample size minus 2.

    # The p-value is calculated by multiplying 1 - stats.t.cdf(t-value, df) by 2
since we are doing a two-tailed test. The resulting p-value represents the
probability of observing a t-value as extreme as the one calculated, assuming that
the null hypothesis(no correlation) is true.
p_value = 2 * (1 - stats.t.cdf(abs(r_value) * (len(data) - 2)
               ** 0.5 / (1 - r_value**2)**0.5, len(data) - 2))

print("p-value:", p_value)

>>> p-value: 0.21012944569123992
```

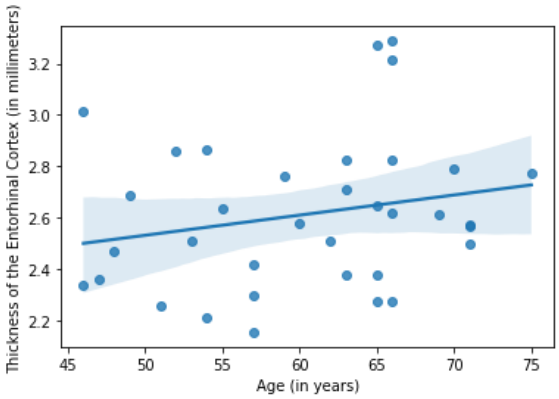
Appendix F: Calculating the 95% Confidence Intervals

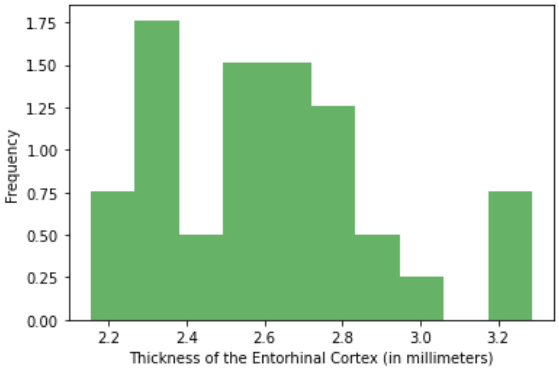
```
# Define the number of observations in the dataset
n = len(data)

# Calculate the mean of the age variable in the dataset
x_bar = data['age'].mean()

# Calculate the mean of the a_pe_cort variable in the dataset
y_bar = data['a_pe_cort'].mean()

# Calculate the standard deviation of the age variable in the dataset
s_x = data['age'].std()

# Calculate the standard deviation of the a_pe_cort variable in the dataset
s_y = data['a_pe_cort'].std()

# Define the t-value for a 95% confidence interval (with n-2 degrees of freedom)
t = stats.t.ppf(0.975, n-2)

# Calculate the left endpoint of the confidence interval for the slope
conf_int_left = slope - t * (std_err)

# Calculate the right endpoint of the confidence interval for the slope
conf_int_right = slope + t * (std_err)

# Print the confidence interval
print(
    "95% Confidence interval for the slope: [", conf_int_left, ", ",
conf_int_right, "]")

>>> 95% Confidence interval for the slope:
    [ -0.004647233116308061,     0.020352097605536697 ]
```

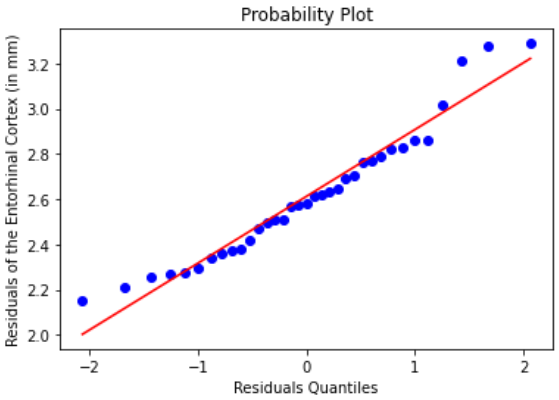
Appendix G: Printing a Summary of all the Calculations and Plots used.

```
# Print all outputs
print("Correlation - Pearson's (r):", r_value)
print("Coefficient of determination (R2):", r_squared)
print("slope (β1):", slope)
print("y-intercept (β0):", intercept)
print("Linear regression equation: y =", slope, "* age +", intercept)
print("Standard error of the slope:", std_err)
print("95% Confidence interval for the slope: [", conf_int_left, ", ",
conf_int_right, "]")
print("p-value:", p_value)

# Regression plot
sns.regplot(data['age'], data['a_pe_cort'])
plt.xlabel('Age (in years)')
plt.ylabel('Thickness of the Entorhinal Cortex (in millimeters)')
plt.show()

# Histogram of the age variable
plt.hist(data['a_pe_cort'], bins=10, density=True, alpha=0.6, color='g')
plt.xlabel('Thickness of the Entorhinal Cortex (in millimeters)')
plt.ylabel('Frequency')
plt.show()

# Q-Q plot of a_pe_cort
    # The stats.probplot() function creates a Q-Q plot of the data, with the data
points plotted on the x-axis and y-axis as specified in the function's parameters
(data['a_pe_cort'] in this case)
stats.probplot(data['a_pe_cort'], dist="norm", plot=plt)
plt.xlabel('Residuals Quantiles')
plt.ylabel('Residuals of the Entorhinal Cortex (in mm)')
plt.show()


>>> Correlation - Pearson's (r): 0.2171792054922389
    Coefficient of determination (R2): 0.04716680729824013
    slope (β1): 0.007852432244614318
    y-intercept (β0): 2.1378531619179983
    Linear regression equation: y = 0.007852432244614318 * age + 2.1378531619179983
    Standard error of the slope: 0.006143805051057197
    95% Confidence interval for the slope: [ -0.004647233116308061 ,
    0.020352097605536697 ]
    p-value: 0.21012944569123992
```


Thickness of the Entorhinal Cortex (in millimeters)
Age (in years)


Thickness of the Entorhinal Cortex (in millimeters)
Age (in years)


Frequency
Thickness of the Entorhinal Cortex (in millimeters)


Probability Plot
Residuals of the Entorhinal Cortex (in mm)
Residuals Quantiles

Appendix H: Comparing Calculations by Library Functions with the Outputs of my Custom Code to Validate my Answers

```
# Review the calculations to see if it matches the output from the statsmodels
library

# var            coef    std err          t      P>|t|      [0.025      0.975]
# age            0.0079      0.006      1.278      0.210      -0.005       0.020

# Yey! They are consistent. :)

from statsmodels.regression.linear_model import OLS
import statsmodels.api as sm

# Define the variables
x = data['age']
y = data['a_pe_cort']

# Add a constant term to the predictor variable (age)
x = sm.add_constant(x)

# Create an OLS model
model = OLS(y, x)

# Fit the model to the data
results = model.fit()

# Print the results summary
print(results.summary())

>>>
```

OLS Regression Results

Dep. Variable:	a_pe_cort	R-squared:	0.047
Model:	OLS	Adj. R-squared:	0.018
Method:	Least Squares	F-statistic:	1.634
Date:	Fri, 27 Jan 2023	Prob (F-statistic):	0.210
Time:	12:50:02	Log-Likelihood:	-5.2457
No. Observations:	35	AIC:	14.49
Df Residuals:	33	BIC:	17.60
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1379	0.374	5.714	0.000	1.377	2.899
age	0.0079	0.006	1.278	0.210	-0.005	0.020

Omnibus:	3.380	Durbin-Watson:	1.859
Prob(Omnibus):	0.185	Jarque-Bera (JB):	2.877
Skew:	0.697	Prob(JB):	0.237
Kurtosis:	2.831	Cond. No.	466.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.