Statistical Inference HW#4 programming sections

Student Name:
Pouya Haji Mohammadi Gohari

SID:810102113

Date of deadline
Monday 6th February, 2023

Dept. of Computer Engineering

University of Tehran

Contents

Attention: Note that we have fully solved 11 first questions in the previous submission and for last submission we solve the last 2 remaining questions.

1 Problem 1

1.1 Choose a sample size of 100 from population.

We have created a function that create sample from population(BMI with outcome 1) which generate 100 samples from our population. As you can see in the figure 1 histogram of generated data is:

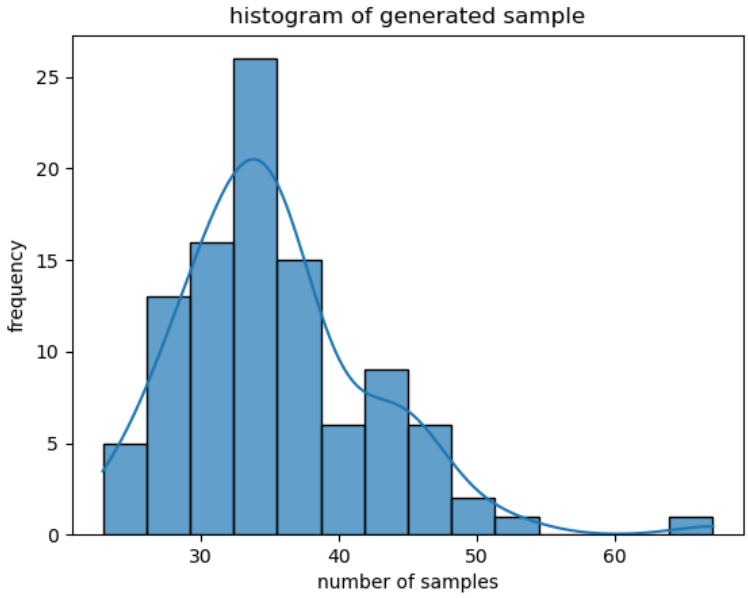

Figure 1: Histogram of the generated sample

1.2 Construct a 90% confidence interval for the mean using bootstrapping on this sample. (use the percentile method and set the number of repetitions equal to 1000)

We have define a function to use bootstrap method to calculate the means for each generated sample from our population therefore in each simulation we get 100 samples from population then calculate the mean. The mean statistic using bootstrap method is in figure: 2:

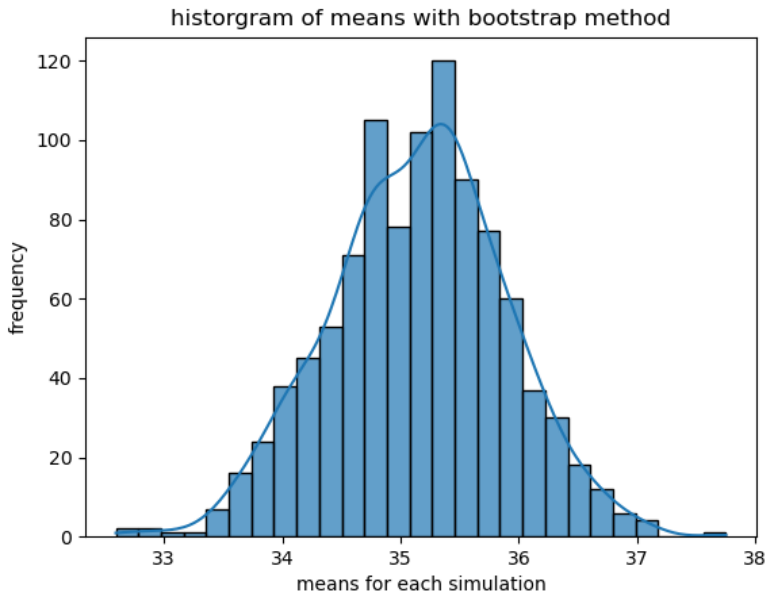

Figure 2: Statistic of mean using bootstrap method

In order to construct a 90% confidence interval for the mean with percentile method with bootstrap we must find the $(\alpha/2B)^{\text{th}}$ and $((1-\alpha/2)B)^{\text{th}}$ ordered of $\hat{\theta}^*$. So the procedure of this method is:

- First sort the means using bootstrap method.
- Find the $(\alpha/2B)^{\text{th}}$ of the means. which is : 33.92295
- Find the $((1-\alpha/2)B)^{\text{th}}$ of the means. which is : 36.3195

Therfore a 90% CI for the mean is: $(33.92295, 36.3195)$

1.3 Report bootstrap mean and population mean.

The bootstrap mean and population mean is in table 2:

Table 1: Mean of population and bootstrap

bootstrap mean	population mean
35.159435	35.14253731343284

1.4 Plot the histogram of the bootstrap distribution and show the vertical lines of the confidence interval on it.

We have already plot histogram of the data in figure 2 but for see the CI from previous section on this histogram, we refer you to the figure 3:

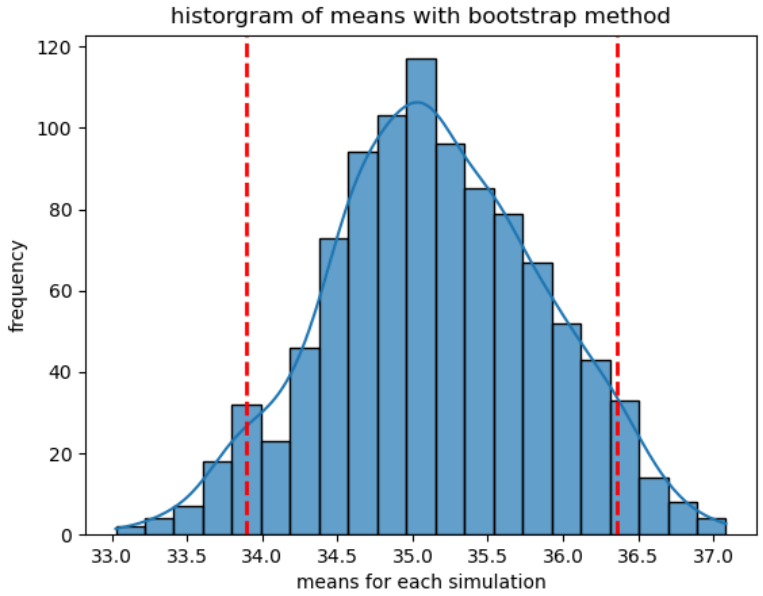

Figure 3: Histogram of bootstraping method and CI

1.5 Repeat steps 1-3 for a smaller sample size (10) and compare with the previous results.

We have repeated the following proceude with sample size of 10 the histogram of the generated sample size of 10 is in figure 4:

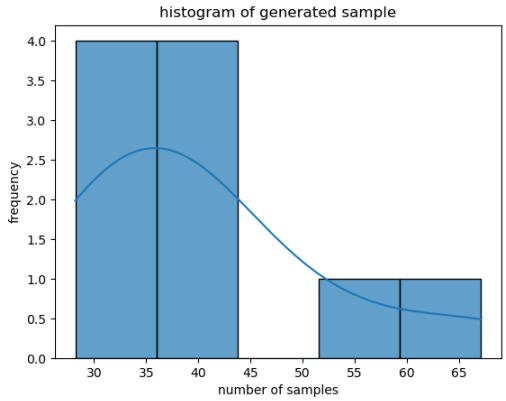

Figure 4: Histogram of the generated sample with size of 10

For conducting a 90% with 1000 simulations with bootstraping method and percentile we have $(31.4595, 39.15)$. The mean of population and bootstrap mean is in table ??:

Table 2: Mean of population and bootstrap

bootstrap mean	population mean
35.18583	35.14253731343284

Histogram of the bootstraping method of sample size of 10 for each simulation is in figure 5:

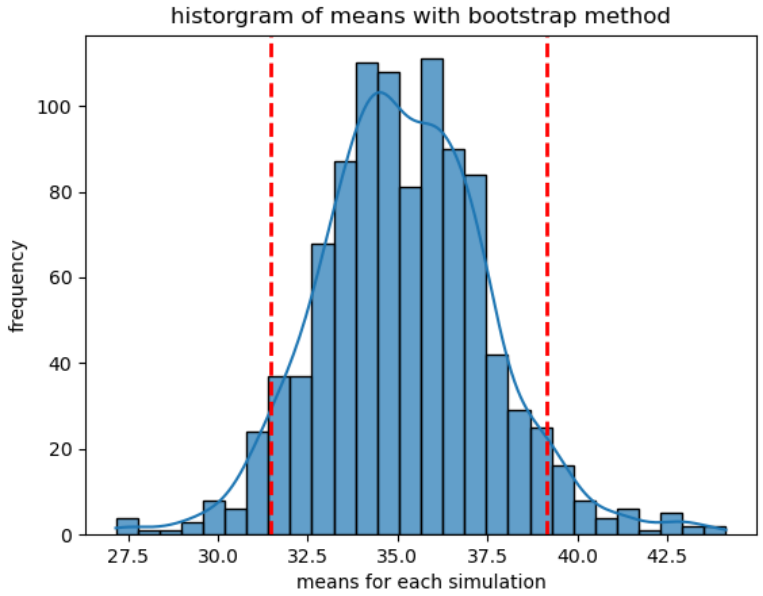

Figure 5: Histogram of bootstraping method alongside CI for 10 sample size

Advantages of each scenarios is:
For bigger sample size:

- The estimate of the mean is more accurate since the larger sample size are given.
- The variablity of the larger sample size is much lesser since we saw that in CI for each scenarios.
- It is indeed less bias since larger sample sizes are usually more represantive of the population.
- From CLT[1] indicates that distribution of the sample means is likely to be approximately normal.

For smaller sample size:

- The esimate is not accurate compared to the larger sample size.
- Has higer variablity since the conducted CI for this scenario has shown this.

[1]Centeral Limit Theorem

- It is more bias since we have less data for capturing population characteristics.
- From CLT is can be shown it does not have a good approximation.
- It requires less data which is very good for us when data collection is expensive.
- Computationally speeking is faster since we generate smaller sample size in each simulation.

1.6 Are the confidence intervals symmetric around the point estimate? If not, what might be the reason?

In both scenarios the confidence intervals around the mean estimate is symmetric since the CLT says they must be normal but for larger sample size it is more symmetric since of sample size is being large enough. And for smaller sizes of simulations the mean might not be well-approximated by a normal distribution and we can see that in the figures 5 and 3.

2 Problem 2

2.1 At the 5% significance level, does there appear to be a difference in mean lifetime among the four brands of batteries?

An naive approach is using t-test for pair of groups, but we acctually can use ANOVA with following procedure:

- The source of variability are: 1. within variability 2.between variability
- The statistic model is

  $$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

  Where ϵ_{ij} is normal with zero mean and std of σ(This assumtion is given by question which the data samples are indeed independent and also comes from normal populations with equal std)
- We can use the folloing identity:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i.})^2 + J\sum_{j=1}^{J}(\overline{Y}_{i.} - \overline{Y}_{..})^2 \tag{1}$$

  Where:

$$\overline{Y}_{i.} = \frac{1}{j}\sum_{j=1}^{J} Y_{ij}$$
$$\overline{Y}_{..} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} \tag{2}$$

  The first term in equation 1 SST[2] and second term is SSW and last one is SSB.
- We can use F-test for following hypothesis:

$$H_0 : \alpha_0 = \alpha_1 = \cdots = \alpha_I = 0$$
$$H_a : \exists i \in \{1, 2, \cdots, I\} : \alpha_i \neq 0$$

  The F-test is:

$$F = \frac{\frac{SSB}{I-1}}{\frac{SSW}{I(J-1)}} \tag{3}$$

- If the null hypothesis is true, the F statistic should be close to 1, whereas if it is false, the statistic should be larger.

[2] sum of square total

We can calculate mean of each group as in table 3

	Brand D	Brand C	Brand B	Brand A
	20	24	28	42
	32	36	36	30
	38	28	31	39
	28	28	32	28
	25	33	27	29
means	28.6	29.8	30.8	33.6

Table 3: Calculate the mean

The overall mean is $\overline{Y}_{..} = 30.7$, Next step is calculating the SS_{TOT},SS_w SS_b:

$$\begin{aligned} SS_{\text{TOT}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{..})^2 \\ &= 560.20 \\ SS_W &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i.})^2 \\ &= 492.0 \\ SS_B &= J\sum_{j=1}^{J}(\overline{Y}_{i.} - \overline{Y}_{..})^2 \\ &= 68.2 \end{aligned} \tag{4}$$

Calculate F statistic as follow:

$$\begin{aligned} F &= \frac{\frac{SSB}{I-1}}{\frac{SSW}{I(J-1)}} \\ &= \frac{\frac{68.2}{3}}{\frac{492}{4(4)}} \\ &= 0.739295393 \end{aligned} \tag{5}$$

Where the p-value for obtained F statistic is: 0.549 so the null hypothesis with 0.05 significant level will not be rejected. The ANOVA table is in table 4:

Source	df	Sum of Square	Mean of Square	F
Between Brands	3	68.2	22.733333333	0.739295393
Within Brands	16	492	30.75	
Total	15	560.2		

Table 4: ANOVA table

3 Problem 3

3.1 What are the hypotheses?

The hypotheses is where mean of each treatment is equal to each other more precisely:

$$H_0 : \alpha_0 = \alpha_1 = \cdots = \alpha_I = 0$$

3.2 What is the conclusion of the test? Use a 5% significance level.

We have obtain that p-value for F statistic is 0.0461 where is less than 0.05 significant level and therfore null hypotheses will not be rejected.

3.3 Conduct pairwise tests (bonferroni - t test) to determine which groups are different from each other.

For each pair we conduct t-test, Between treatment 1 and 2:

$$\begin{aligned} t &= \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ &= \frac{6.21 - 2.86}{\sqrt{10.806428571 + 4.503114286}} \\ &= \frac{3.35}{3.912741093} \\ &= 0.856177273 \end{aligned} \tag{6}$$

Between treatment 2 and 3:

$$\begin{aligned} t &= \frac{\mu_2 - \mu_3}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_3^2}{n_3}}} \\ &= \frac{2.86 + 3.21}{\sqrt{4.503114286 + 5.246064286}} \\ &= \frac{6.07}{3.122367463} \\ &= 1.944037681 \end{aligned} \tag{7}$$

Between treatment 1 and 3:

$$\begin{aligned} t &= \frac{\mu_1 - \mu_3}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_3^2}{n_3}}} \\ &= \frac{6.21 + 3.21}{\sqrt{10.806428571 + 5.246064286}} \\ &= \frac{9.42}{4.006556234} \\ &= 2.351146334 \end{aligned} \tag{8}$$

And degree of freedom is $I(J-1) = 13 * 3 = 39$ also for bonferroni correction the significant level will divided by 3(number of comparisons) so the t-score for two tail will be:

$$t_{39}(\alpha/3) = t_{39}(0.025/3) = \pm 2.501658$$

And non of the t-test above can exceed the t-score with bonferroni correction.

3.4 State one advantage and one disadvantage of the bonfer- roni correction method.

The bonferroni correction is a method where used to adjust the type I error when multiple hypothesis are testing.
Advantage: The primary advantage is decrease the chance of type I error which is incorrect rejection of a true null hypothesis.When multiple test are conducted, the likelihood of a type I error will increase and duty of bonferroni is to decrease that.
Disadvatage: One of the major of disadvantages of bonferroni correction is that however it will reduce type I error, it will increase the likelihood of Type II error since the correction is too conservative.

4 Problem 4

4.1 Use the Benjamini-Hochberg method, which is a method to control the FDR (false discovery rate), and determine the significant p-values. (Consider a control level of 5%)

In Benjamini-Hochberg procedure is:

- Sort the p-values: $P_{(1)} \leq P_{(2)} \leq \cdots P_{(n)}$
- Find the biggest $P_{(r)} \leq \frac{qr}{n}$ where q is significant level and n is the number of tests.
- Reject every null hypothesis of $H_0^{(1)}, H_0^{(2)}, \cdots, H_0^{(r)}$.

The given p-values are:

$$\text{P-values} = 0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019 \tag{9}$$

Sort the p-values:

$$\text{Sorted p-values} = 0.005, 0.009, 0.019, 0.022, 0.051, 0.101, 0.361, 0.387 \tag{10}$$

where n and q are $8, 0.05$.The $\frac{qr}{n}$ vector is:

$$\frac{qr}{n} = 0.00625, 0.0125, 0.01875, 0.025, 0.03125, 0.0375, 0.04375, 0.05 \tag{11}$$

So from equations 10, 11 we can obtain that biggest r where $P_{(r)} \leq \frac{qr}{n}$ is 0.022 so we will reject all null hypothesis where p-value is less than or equal to 0.022.

4.2 Plot the p-value chart according to their rank and show the cut off line.

The plot is in figure 6: The cutoff line is the green light and the 0.05 significant level is shown by red line and as

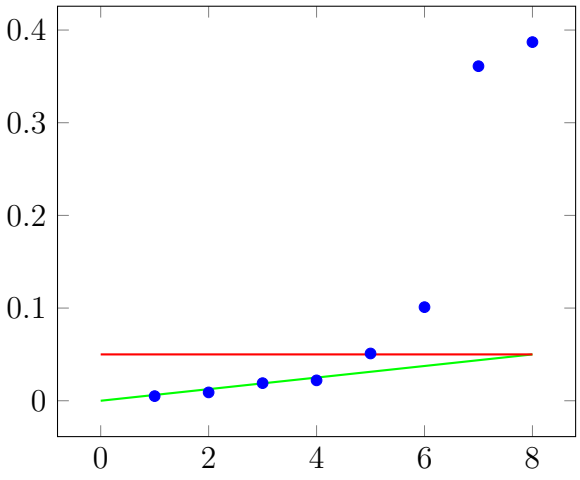

Figure 6: Plot for p-value chart

you can see in the figure 6 the 5-th of the ordered p-value is above the cutoff line (green line) so we will reject any null hypothesis where it's p-value is less or equal to 4-th p-value(0.022)!

4.3 Briefly explain the difference between FDR control meth- ods (such as Benjamini-Hochberg) and FWER (family wise error rate) control methods such as Bonferroni.

We can classify the controls for multiple testing hypothesis as 2 groups:

- FWER
- FDR

Where Controlling the FWER $P[v \geq 1]$ may be too conservative and greatly reduce our power to detect real effects, especially when n(the total number of tested hypothesis) is large. In moder "large-scale testing" applications, focus has shifted to the false dicovery proportion. Where the statistical test is thought of as providing a "definitive answer" for whether an effect is real, FWER control is still correct objective.In contrast, for applications where the statistical test identifies candidate effects that are likely to be real and which merit furhter study, it may be better to target FDR control.

5 Problem 5

5.1 If the number of groups increases, then the type 1 error increases in multiple comparisons tests, so the corrected significance level should increase.

If the number of groups increases, then we must perform such as FWER controlls(like bonferreni correction) since the comparisions are increasing, so the corrected significant level should decrease.

5.2 If the number of samples increases, the degree of freedom for the residuals also increases.

In ANOVA for SS_w we have $I(J-1)$ degree of freedom, if the number of samples increases in each gropu or level therefore the degree of freedom of residuals(SS_w) will also increase.

5.3 The F distribution is a symmetric distribution around the zero mean.

The F distribution with several freedom is in figure 7:

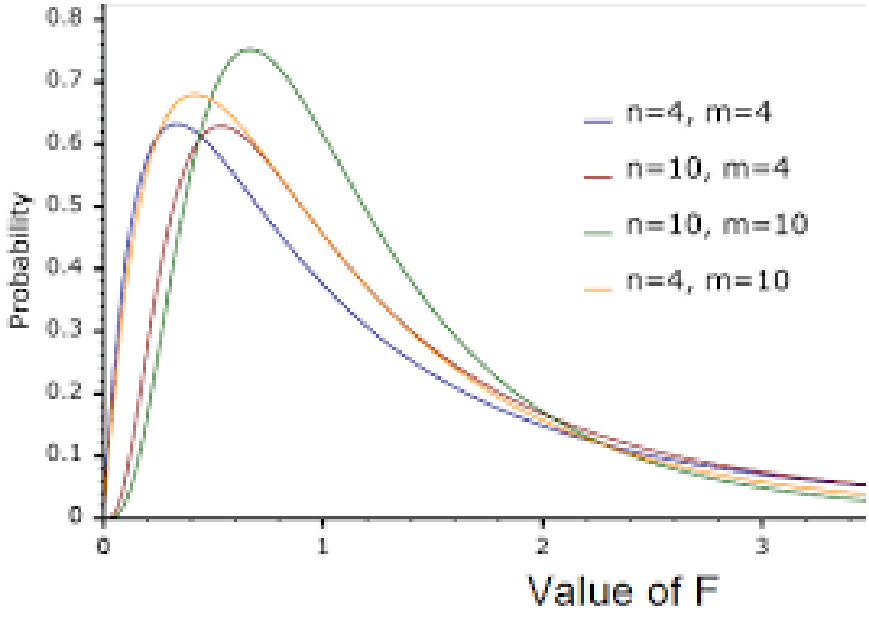

Figure 7: F distribution with several degree of freedoms

The fisher F distribution is not symmetric and it is right tailed.Thus this probability distribution is not symmetric at all.

5.4 Using ANOVA test, we can conclude that all means are different from each other.

It is not true since if ANOVA test would be rejected therefore we can say for some groups their means are not equal.For seeing if some of the groups are different in their means we must use some other methods like Tukey.

5.5 If the initial hypothesis is rejected in the ANOVA test, the standardized variability between groups is higher than the standardized variability within groups.

If the ANOVA test would be rejected therefore the F statistics will be large enough more formally:

$$\begin{aligned} F &= \frac{MSB}{MSE} \\ &= \frac{\frac{SSB}{I-1}}{\frac{SSW}{I(J-1)}} \end{aligned} \tag{12}$$

F-stats is depend on 4 term:

- SSB
- SSW
- I(number of gropus)
- J(number of samples in each group)

So if the F is high enough to reject the ANOVA test therefore we can not say if SSB is greater than SSW!.But we can dinfinitly say that standardized between variability (MSB) is greater than standardized within variability(MSE).So this statement is true.

6 Problem 6

6.1 Write the Null and Alternative hypotheses and conduct analysis using one-way ANOVA.(use the R or Python programming language to solve this part).

The null hypothesis is mean of each group with different aroma in that place is equal to each other so we can perform and conduct ANOVA test for following test.
From the python language we can add from scipy library to test oneway ANOVA where we can see the F statistic and p-value in table 5: Where the null hypothesis will be rejected since the F statistic is large enough and also

Table 5: Statisit and p-value of ANOVA test

F statistic	p-value
13.0	0.0019196756805123854

p-value is less than 0.01 where is less than significant level of 0.01.

6.2 Determine the significantly different pairs of means using the Tukey's method.(Use a 5% significance level).

We know that ANOVA will just say that mean of each group is equal to each other or not but it can not say that which group is different to another so we can use other methods like:

- Tukey
- Bonferreni

The procedure of Tukey's method is:

- First calculate the mean of each group
- Next step is calculate the mean difference between each group
- If absolute value of the differences is bigger than $q_{I,I(J-1)}(\alpha)\frac{S_p}{\sqrt{J}}$ Then the null hypothesis of mean of two groups are equall will be rejected!

Step1: Obtain the mean of each group(in table 6):

Table 6: Mean of each group

	Mean
Lemon	11
Flora	12
Fried Food	6
None	7

Now we are goin to calculate absolute value of mean differences of each group in table 7:

Table 7: Absolute value of differences between each mean

	Absoulte value of differences
Between 1,2	1
Between 1,3	5
Between 1,4	4
Between 2,3	6
Between 2,4	5
Between 3,4	1

At first we are going to calculate each of sample varinces s_i^2 of each group:

$$\begin{aligned} s_1^2 &= \frac{1}{J-1}\sum_{j=1}^{J}(x_j - \overline{x}_1)^2 \\ &= 1 \\ s_2^2 &= \frac{1}{J-1}\sum_{j=1}^{J}(x_j - \overline{x}_2)^2 \\ &= 3 \\ s_3^2 &= \frac{1}{J-1}\sum_{j=1}^{J}(x_j - \overline{x}_3)^2 \\ &= 3 \\ s_4^2 &= \frac{1}{J-1}\sum_{j=1}^{J}(x_j - \overline{x}_4)^2 \\ &= 1 \end{aligned} \tag{13}$$

Calcuate the S_p^2 as follow:

$$\begin{aligned} S_p^2 &= \frac{\sum_{i=1}^{I}(J-1)s_i^2}{I(J-1)} \\ &= \frac{2*(1+3+3+1)}{4*2} \\ &= \frac{16}{8} \\ &= 2 \end{aligned} \tag{14}$$

$q_{I,I(J-1)}(0.05)$ is studentized range distribution and the value is :

$$\begin{aligned} q_{I,I(J-1)}(0.05)\frac{S_p}{\sqrt{J}} &= q_{4,8}(0.05)\frac{2}{\sqrt{3}} \\ &= 5.229638738 \end{aligned} \tag{15}$$

So if the absolute value of differences between means are greater than 5.229638738 therfore we can say their means is not equal.So the mean of Flora aroma's group and mean of Fried food aroma's group are not equal.

6.3 State one limitation of Tukey's procedure.

The limitation of Tukey's method is when the number of groups is large and it can be too conservative.This conservativeness means that Tukey's method might have reduced power to detect actual differences between group means when many comparisions are made.

7 Problem 7

7.1 What type of study is this?

This is a randomized controlled study.

7.2 What are the experimental and control treatments in this study?

The participants in controlled group are treated placebo and experminal's group are treated by 25 grams chia seeds twice a day.

7.3 Has blocking been used in this study? If so, what is the blocking variable?

There has not been blocking usage in this study.There is no other variable than controll and experimental groups!

7.4 Has blinding been used in this study?

Yes, since random assinments are used to classify them into two groups therfore blinding has beed used in this study.

7.5 Has double-blinding been used in this study?

No since we know each participant's groups inforamtion.

7.6 Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.

From random sampling, we can generalize the conclusion to the population at large and from random assinments we can make a causal statement.
This study has random assinments but we don't know the recruits are random sampling or not therefore we can make causal statement since we have random assinments but we can not generalize the conclusion to the large population since we are not sure this study is based on random sampling or not and we can conclude to this sample!

8 Problem 8

8.1 Describe the relationship between height and weight.

Since given-scatterplot is showing us there is a positive correlation between two variables we can actually fit a line to these variables.

8.2 Write the equation of the regression line. Interpret the slope and intercept in context.

The regression line is:

$$\text{height} = -105.0113 + 1.0176\text{weight} \tag{16}$$

The intercept is -105.0113 and the slope is 1.0176.

8.3 Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion

We must conduct null hypothesis as follow:

$$\begin{aligned} H_0 &: B_1 = 0 \\ H_1 &: B_1 \neq 0 \end{aligned} \tag{17}$$

So we can perform t_{n-2} statistic is:

$$\begin{aligned} \frac{\hat{B}_1 - B_1}{SE(\hat{B}_1)} &= t_{n-2} \\ &= \frac{1.0176 - 0}{0.0440} \\ &= 23.127272727 \end{aligned} \tag{18}$$

So for the p-value for is approximately 0 so the null hypothesis will be rejected and therefore increasing of height is associated with increasing weight.

8.4 The correlation coefficient for height and weight is 0.72. Calculate R2 and interpret it in context.

The R^2 value in this context is 0.5184. This means that approximately 51.84% of the variability in the weight can be explained by the linear relationship with height among the 507 physically active individuals.(Note that the R2 coefficient of determination is the square of correlation coefficient!)

9 Problem 9

9.1 Using maximum likelihood estimator, calculate ...

From maximum likelihood estimator we can calculate the B_0, B_1, $var(B_0)$ and $var(B_1)$: B_0 and B_1 is the same for LSE's:

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ &= 48.142999999999994 \\ S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= 108.969 \\ S_{yy} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\ &= 21.561 \end{aligned} \tag{19}$$

Now we can estimate the B_0 and B_1 as follows:

$$\begin{aligned} \hat{B}_1 &= \frac{S_{xy}}{S_{xx}} \\ &= 0.44180455 \\ \hat{B}_0 &= \overline{y} - \hat{B}_1\overline{x} \\ &= 2.23 - 0.44180455 * 4.79 \\ &= 0.113756205 \end{aligned} \tag{20}$$

And from MLE we can find biased $\hat{\sigma}^2$ as follows:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{SSE}{n} \\ S^2_{\text{unbiased}} &= \frac{SSE}{n-2} \end{aligned} \tag{21}$$

For obtaining the S^2_{unbiased} we shall calculate SSE firts:

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \\ &= 0.29120355330415076 \end{aligned} \tag{22}$$

Eventually can calculate the S^2_{unbiased}:

$$\begin{aligned} S^2_{\text{unbiased}} &= \frac{16.868139699280267}{8} \\ &= 0.036400444 \end{aligned} \tag{23}$$

Variances of B_0 and B_1 are:

$$\begin{aligned}
\sigma^2_{B_0} &= s^2 \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} \\
&= 0.036400444 \frac{338.41}{10 * 108.969} \\
&= 0.011304384 \\
\sigma^2_{B_1} &= \frac{s^2}{S_{xx}} \\
&= \frac{0.036400444}{108.969} \\
&= 0.000334044
\end{aligned} \tag{24}$$

9.2 Test the following hypotheses at the level of significance 0.05:

- $$\begin{aligned} H_0 &: B_0 = 0.5 \\ H_a &: B_0 \neq 0.5 \end{aligned} \tag{25}$$

  We can perform the t-test with degree of freedom (n-2) and we can say:

  $$\begin{aligned}
  t_0 &= \frac{\hat{B}_0 - B_0}{SE(\hat{B}_0)} \\
  &= \frac{0.113756205 - 0.5}{\sqrt{0.011304384}} \\
  &= -3.632771357
  \end{aligned} \tag{26}$$

  The $t_8(\pm 0.025) = \pm 2.306$ since the t-score is -3.632771357 we will reject the null hypothesis.

- The regression line passes through the origin in the XY plane. This null hypothesis says that:

  $$\begin{aligned} H_0 &: B_0 = 0 \\ H_a &: B_0 \neq 0 \end{aligned} \tag{27}$$

  And we can calculate the t-score as follows:

  $$\begin{aligned}
  t_0 &= \frac{\hat{B}_0 - B_0}{SE(\hat{B}_0)} \\
  &= \frac{0.113756205}{\sqrt{0.011304384}} \\
  &= 1.069920834
  \end{aligned} \tag{28}$$

  The $t_8(\pm 0.025) = \pm 2.306$ since the t-score is 1.069920834 we will not reject the null hypothesis.

10 Problem 10

For particular x we can say that:

$$y = \hat{B}_0 + \hat{B}_1 x \tag{29}$$

We can conduct a confidence interval for y as follow:

$$\hat{y} \pm s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}} \tag{30}$$

The length of this CI is:

$$length = 2s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}} \tag{31}$$

So if we minimize the length we should minimize the under the square so we shall say:

$$x = \overline{x} \tag{32}$$

Therefore the length would be minimize if $x = \overline{x}$!

11 Problem 11

11.1 Show that the MLE (Maximum Likelihood Estimate) of p is $\overline{p} = X/n$.

The maximum likelihood estimate for binomial distribution is:

$$\begin{aligned}
L &= f(x_1, x_2, \cdots, x_n|n, p) \\
&= f(x_1|n, p)f(x_2|n, p)\cdots f(x_n|n, p) \\
&= \left(\prod_{i=1}^{n}\binom{n}{x_i}\right) p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n}(n-x_i)} \\
Ln(L) &= \left(\sum_{i=1}^{n} Ln(\binom{n}{x_i})\right) + (\sum_{i=1}^{n} x_i)Ln(p) + (\sum_{i=1}^{n}(n-x_i))Ln(1-p) \\
\frac{dLn(L)}{dp} &= \frac{\sum_{i=1}^{n} x_i}{p} + \frac{\sum_{i=1}^{n}(n-x_i)}{1-p} \\
&= \frac{\sum_{i=1}^{n} x_i}{p} = \frac{\sum_{i=1}^{n}(n-x_i)}{p-1} \\
&= \frac{1}{p} - 1 = \frac{\sum_{i=1}^{n}(n-x_i)}{\sum_{i=1}^{n} x_i} \\
&= \frac{1}{p} - 1 = \frac{n^2}{\sum_{i=1}^{n} x_i} - 1 \\
&= p = \frac{\sum_{i=1}^{n} x_i}{n^2} \\
&= p = \frac{\overline{x}}{n}
\end{aligned} \tag{33}$$

Therefore we have shown that estimate of maximum likelihood is $\frac{\overline{x}}{n}$.

11.2 Show that MLE of part (a) attains the Cramer-Rao lower bound.

From Cramer-Rao we will say:

$$MSE(\hat{\theta}) = V[\hat{\theta}] \geq \frac{1}{nI(\theta_0)} \tag{34}$$

We can use score function as follow:

$$\begin{aligned}
logf(x|p) &= log\left(\binom{n}{x}p^x(1-p)^{n-x}\right) \\
&= log\left(\binom{n}{x}\right) + xlogp + (n-x)log(1-p) \\
\frac{\partial log(f(x|p))}{\partial p} &= \frac{x}{p} - \frac{n-x}{1-p} \\
\frac{\partial^2 log(f(x|p))}{\partial p^2} &= -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}
\end{aligned} \tag{35}$$

The fisher information is:

$$\begin{aligned} I(p) &= -E_p[-\frac{x}{p^2} - \frac{n-x}{(1-p)^2}] \\ &= \frac{E_p[x]}{p^2} + \frac{n}{(1-p)^2} - \frac{E_p[x]}{(1-p)^2} \\ &= n(\frac{p(1-p)^2 + p^2 - p^3}{(p-p^2)^2}) \\ &= n(\frac{(p-p^2)}{(p-p^2)^2}) \\ &= \frac{n}{p-p^2} \end{aligned} \tag{36}$$

Therefore $Var[\hat{p}]$ is greater or equal to:

$$\frac{1}{nI(p)} = \frac{p(1-p)}{n^2} \tag{37}$$

12 Problem 12

From statsmodels we had fit a regression line and as you can see in the table 8: And we can see the paramters(B_0,

Table 8: Result of ordinary least square

Dep. Variable:	lpsa	R-squared:	0.539
Model:	OLS	Adj. R-squared:	0.535
Method:	Least Squares	F-statistic:	111.3
Date:	Fri, 02 Feb 2024	Prob (F-statistic):	1.12e-17
Time:	02:32:08	Log-Likelihood:	-113.45
No. Observations:	97	AIC:	230.9
Df Residuals:	95	BIC:	236.1
Df Model:	1		
Covariance Type:	nonrobust		

B_1) as you can see in table 9:

Table 9: Fited line with regression using least square method

	coef	std err	t	$P > \|t\|$	$[0.025, 0.975]$
const	1.5073	0.122	12.361	0.000	1.265 , 1.749
lcavol	0.7193	0.068	10.548	0.000	0.584 , 0.855

The residuals can be calucalted with the model.scale which it is 0.620155621631307. Cramer-Rao Lower Bounds for the variances of the estimators: $[0.0148686, 0.00465027]$.
Empirical standard errors of the estimates:0.121937. From the Cramer-Rao lower bound we know that we must calculate the fisher information for $\hat{\beta}_1$:
First we must calculate the Likelihood:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right) \tag{38}$$

Log Likelihood:

$$\ln(L(\beta_0, \beta_1)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2} \tag{39}$$

First deviation of Log Likelihood with respect to β_1:

$$\frac{\partial \ln(L)}{\partial \beta_1} = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)(-x_i)}{\sigma^2} \tag{40}$$

Second deviation:

$$\frac{\partial^2 \ln(L)}{\partial \beta_1^2} = \sum_{i=1}^{n} \frac{x_i^2}{\sigma^2} \tag{41}$$

Fisher information:

$$I(\beta_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})2}{\sigma^2} \tag{42}$$

From Cramer-Rao if the estimated parameter is unbiased(where we know that):

$$\text{Var}(\hat{\beta}_1) \geq \frac{1}{I(\beta_1)} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{43}$$

The Empirical value is 0.004650270803149708 and the Cramer-Rao using formula is 0.004650270803149714 where it is same from the code provided in first place.

13 Problem 13

13.1 Which explanatory variable do you guess is the more significant predictor and why?

For this purpose we can fit a line on this datas and see the p-value of each, The one have smaller p-value is generally considered to be the more statistically significant predictor. And as you can see in table 10:

Table 10: Fited line with regression using least square method

	coef	std err	t	$P > \|t\|$	[0.025 , 0.975]
const	2.2359	0.328	6.814	0.000	1.584 , 2.887
age	0.0163	0.005	3.150	0.002	0.006 , 0.027
lpsa	0.1430	0.033	4.296	0.000	0.077 , 0.209

Since the p-value for lpsa is smaller therefore it is considered to be more statistically significant predictor.

13.2 For each explanatory variable

13.2.1 Investigate the linearity of data points using scatter plot of residuals.

We can use our model in previous part and obtain the residuals of the fitted regression line with OLS and see each scatter plot of both explanatory variable (age, lpsa) and as you can see in figures 8 and 9:

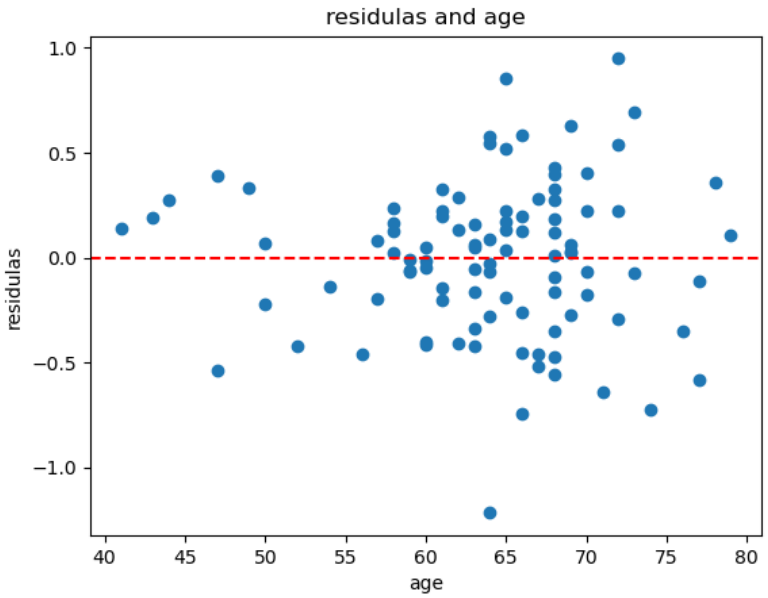

Figure 8: scatter plot for residuals and age

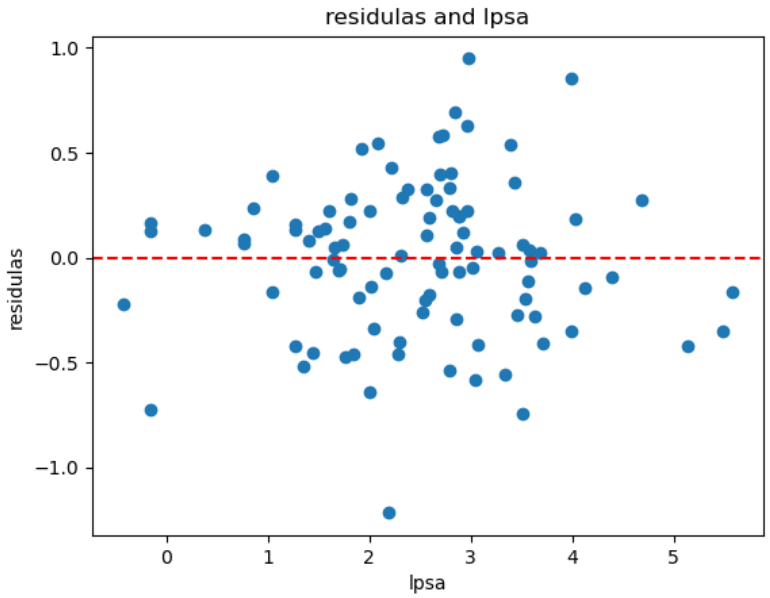

Figure 9: scatter plot for residuals and lpsa

In order to investigate the linearity of the relationship between the explanatory variables and the dependent variables using residuals we should consider the following aspcets:

- Randomness of residuals: Residuals must be scattered randomly around the horizontal line at zero.
- Pattern of symmetric structures: If there are clear patterns this indicates non-linearity.
- Homoscedasticity: The spread of the residuals should be consistent across all values of the explanatory variables.

So from both figures 8 and 9 we can see no pattern and also see a good scattering of randomness of residuals over horizontal line at zero this reasons indicates that the relationship might be linear.

13.2.2 Compute the least squares regression.

Approach one to see both explanatory variable in one equation: The general framework of least square regression is as follow:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, i = 1, 2, \cdots, n \tag{44}$$

We can rewrite it as matrix notation as follow:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

Now we can esimate with linear square method in order to estimate the parameters of fitted line:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{45}$$

The result of calculating this parameters using python is:

$$\hat{\beta} = \begin{bmatrix} 2.2358606 \\ 0.01626208 \\ 0.14303179 \end{bmatrix}$$

As you can see it is very close to OLS result in python.
Second approach see each explanatory variable once:
For this purpose we have fitted 2 regression lines with age and lpsa variable(using them one at a time): The result of OLS for age is in table11:

Table 11: Fitted line with regression using least square method for age and weight

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	2.3502	0.356	6.604	0.000	1.644 , 3.057
age	0.0200	0.006	3.618	0.000	0.009 , 0.031

The predictive line using same formula as before:

$$\text{weight} = 2.35015161 + 0.02002304\text{age} \tag{46}$$

The result of OLS for lpsa is in table ??:

Table 12: Fitted line with regression using least square method for lpsa and weight

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	3.2304	0.094	34.462	0.000	3.044 , 3.416
lpsa	0.1608	0.034	4.686	0.000	0.093 , 0.229

The predictive line is:

$$\text{weight} = 3.23036915 + 0.16081973\text{lpsa} \tag{47}$$

13.2.3 Write the predictive equation for the response variable and interpret its parameters.

From OLS the preditive equation is:

$$\text{weight} = 2.24 + 0.02\text{age} + 0.14\text{lpsa} \tag{48}$$

From my own calculation the predictive equation is:

$$\text{weight} = 2.2358 + 0.0162\text{age} + 0.143\text{lpsa} \tag{49}$$

The result are close to each other either it is from calculation or OLS(built-in function). The interpretation:

- Tntercept: This is the expected value of weight when both age and lpsa are zero.
- Slope of age: This parameter represents the change in the dependent variable weight for a one-unit change in the independent variable age.
- Slope of lpsa: This parameter represents the change in the dependent variable weight for a on-unit change in the independent variable lpsa.

Second approach fiting single line for each explanatory variables:
From OLS(age) results the line is:

$$\text{width} = 2.35 + 0.02\text{age} \tag{50}$$

From OLS(lpsa) results the line is:

$$\text{width} = 3.23 + 0.16 * lpsa \tag{51}$$

Both regression line from OLS results are very close to the calculations in the previous section!

13.2.4 Draw a scatter plot of the relation between these two variables overlaid with the least-squares fit as a dashed line.

Scatter plot of the relation between these two variables with the least-squares fit is in figures 10 and 11:

Figure 10: Age and fitted line

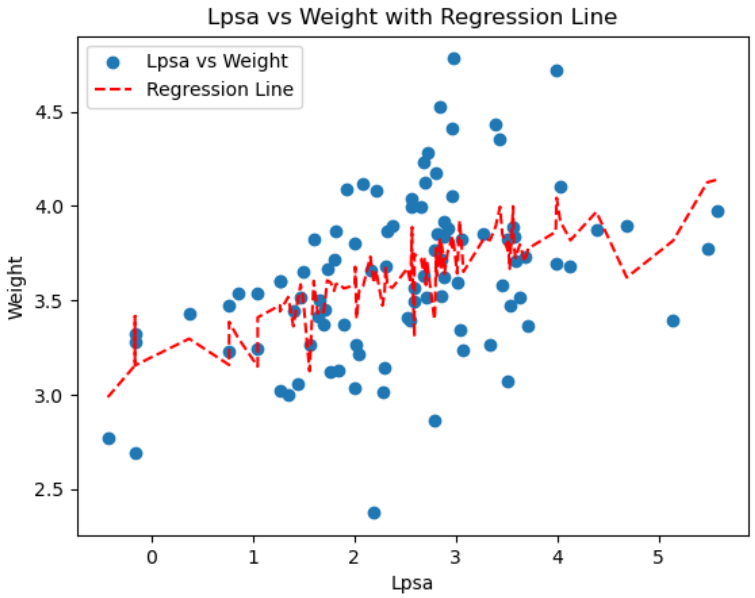

Figure 11: Lpsa and fitted line

As you can see in the figures above regression has beeen shown by red dashed line. In the first scatter plot the data points are fairly scattered around the regression line without a clear pattern of increasing or decreasing residuals which could indicate a linear relationship between age and weight.
From the second scatter plot(Lpsa and weight), we can conclude that the points do not follow the regression line like the first scatter plot.(There seems to be more variablity in weight for given lpsa values).
So we must consider the transformations of the lpsa variables.
Attention: We have seen the a lower p-value indicates a statistically significant relationship, but assesing linearity requires looking at the pattern of the residuals and the fit of the regression line to the data points.It's possibel for a variable to have a significant relationship with the dependent variable but not a prefectly linear one.From the scatter plots provided, if the regression line for lpsa vs weight appears ro be less straight or shows more variablity around the line compared to the plot for age vs weight, this suggest that while there is a significant relationship, it might not be prefectly linear, or it might be other factors influencing the variabltiy in weight that are not captured by lpsa alone. Second approach when for each explanatory variable we fit a regression line:
For age as a single explanatory variabl and weight the scatter plot between these two variables and and the regression line is in figure 13:

Figure 12: Age and fitted line

For lpsa and weight:

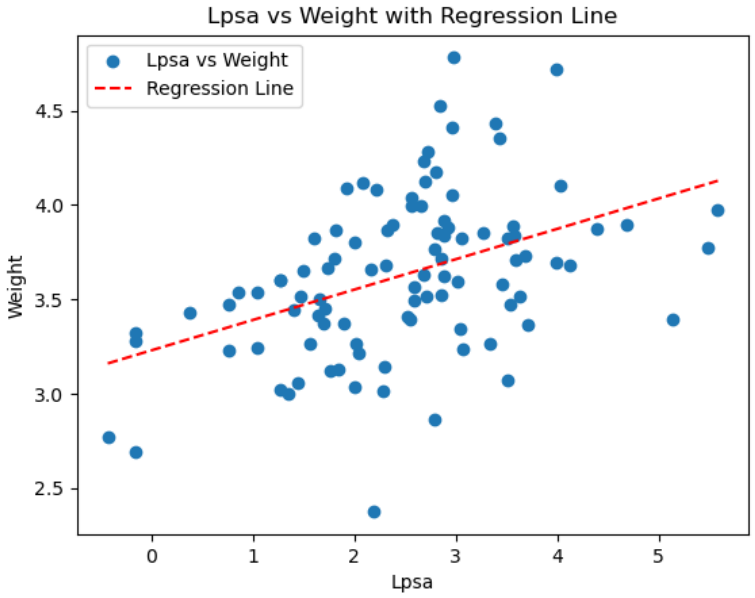

Figure 13: Lpsa and fitted line

13.3 Using the previous part results, try to explain which variable is the more significant predictor.

From table 10 since we have 0.14 coeffient for lpsa and lower p-value therefore lpsa is more significant predictive.

13.4 Choose a random sample of 50 data points from the dataset.

We have generated 50 sample from real data for following purposes:

13.4.1 By 90 percent of data, build two Linear Regression models and design hypothesis tests to see if these explanatory variables are a significant predictor of the response variable or not.

In this section we use split test train function to split our 50 data into 90% train and 10% test size. OLS regression results for age is in table 13:

Table 13: Fitted line with regression using least square method for age and weight

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	2.3804	0.595	4.000	0.000	1.180 , 3.580
age	0.0179	0.009	1.901	0.064	-0.001 , 0.037

The p-value for const is aprroximately zero and for age is 0.064(where is not significantly for 0.05 significant level) and the predictive equation is:

$$\text{weight} = 2.3804 + 0.0179\text{age} \tag{52}$$

The R-squared is 0.056 suggest that 5.6% of the variation of weight is accounted for by linear regression on age where the relationship between the two is weakly linear with a positive slope.
OLS regression results for lpsa is in table 14:

Table 14: Fitted line with regression using least square method for lpsa and weight

	coef	std err	t	P>|t|	[0.025 0.975]
const	3.1612	0.115	27.460	0.000	2.929 3.393
lpsa	0.1595	0.046	3.454	0.001	0.066 0.253

Since the p-value is of lpsa is 0.001 it is significant predictor(for 0.05 significant level) and predictive equation is:

$$\text{width} = 3.1612 + 0.1595\text{lpsa} \tag{53}$$

The R-squared is 0.217 suggest that 21.7% of the variation of weight is accounted for by linear regression on lpsa where the relationship between the two is not strongly linear with a positive slope. But compared to age it has more linear relationship with weight.
Result: lpsa has higher impact on weight since has higher R-squared but significant predictor is lpsa and not age since the p-value for age is higher than 0.05.

13.4.2 Calculate the 95% confidence interval for the slope of the relationship between response variable and explanatory variables. Interpret these CIs.

We have conduct a confidence interval for both slope of predicted equations where for age is :

$$CI = [-0.001, 0.03692720085038906] \tag{54}$$

The interpret of CI for slop corresponding to the age is we are 95% confindent that the real slop is in this range and since this CI include zero therefore weight might has not linear relationship with age.
confidence interval for slope of lpsa:

$$CI = [0.066, 0.2527] \tag{55}$$

The interpret of lpsa CI is we are 95% confindent that real slop is in this range and since this CI does not include zero therfore we can say that there is a linear relationship between lpsa and weight with positive slope.

13.4.3 Use your models to predict the values of the response variable for the remaining percent of samples.

We have predict remain data(0.1 percentage of sample data) where you can see the actual values for weight and predicted with different explanatory variables in table 15:

Actual values	Predict with age	Predict with lpsa
3.974998	3.598828	4.051898
3.764682	3.258379	3.607009
3.539509	3.222542	3.328320
4.524502	3.688420	3.614630
3.825375	3.616746	3.648827

Table 15: Actual values vs predicted values with age and lpsa

13.4.4 Compare the predicted values with actuals. Report the success rate.

In here we don't have labelize data where we can set an accuracy for our method, Instead we can use different metrics to see the success rate, generally R^2 and MSE are two common metrics used to evaluate the performance of regression models:

- R^2 score:The R^2 score is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. In simpler terms, it measures how well the regression predictions approximate the real data points. An R^2 score of 1 indicates that the regression predictions perfectly fit the data. Mathematically, R2R2 is defined as $1 - \frac{SSres}{SStot}$ where SSres is the sum of squares of residuals (the differences between the observed and predicted values) and SStot is the total sum of squares (the differences between the observed values and the mean of observed values). R^2 can sometimes be negative in models that do not intercept, indicating that the model is worse than simply predicting the mean of the dependent variable.

- MSE: Is a measure of the average squared difference between the actual and predicted values, giving insight into the error of a model. It's used to assess how close a fitted line is to the actual data points. The smaller the MSE, the closer the fit is to the data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

And we can see these metrics in table 16: where R^2 scores shows that the model has significant issues, possibly

Table 16: Some metrics for success rates

metrics	R^2	MSE
age samples	-1.2721269110345732	0.24817508603508687
lpsa samples	-0.7109715689039531	0.18688239387252475

due to overfitting or underfitting or using features that do not have a predictive relation with the target value.But the MSE value provides a measure of the average error magnitude, but without more context such as range or distribution of dependent variable, it is hard to assess whether this values are high or low.Neverthelss, obtained the negeative R^2 score, improving the model is neccessary.