University of Tehran

Statistical Inference Report
Homework 4

Hadiseh Mesbah
Student ID: 810102253

February 1, 2024

Problem 1

In this question, we are going to examine the bootstrapping method. For this question, please utilize the diabetes data set provided to you. Consider the population to be equal to the BMI, Body Mass Index of individuals who do not have diabetes.

1. **Choose a sample size 100 from the population.**

   Answer:

   We employed the following code to achieve this: initially, we selected individuals without diabetes and randomly sampled 100 people for subsequent analyses.

```python
import pandas as pd
import numpy as np

# Load the dataset
file_path = '/content/diabetes (2).csv'
diabetes_data = pd.read_csv(file_path)

# Extract BMI values for individuals who do not have diabetes
bmi_no_diabetes = diabetes_data[diabetes_data['Outcome'] == 0]['BMI']

# Set sample size
sample_size = 100

# Randomly choose a sample from the population (BMI of non-diabetic individuals)
np.random.seed(0)  # for reproducibility
sample_bmi = np.random.choice(bmi_no_diabetes, size=sample_size, replace=False)

# Display the sample
print(sample_bmi)
```

Listing 1: random sample of 100

```
[32.5, 42.4, 30.2, 24.2, 28.4, 27.6, 30.1, 46.3, 28.7, 30.1, 31.6, 42.1, 27.8, 24.0,
40.5, 25.6, 23.1, 27.4, 31.2, 42.7, 47.9, 26.9, 22.3, 29.6, 37.7, 30.9, 27.4, 32.4,
27.3, 37.2, 33.8, 27.7, 24.3, 39.4, 29.5, 32.0, 33.5, 28.3, 32.8, 25.1, 33.3, 25.0,
20.1, 35.3, 34.2, 28.1, 28.7, 25.0, 28.8, 29.0, 46.2, 34.2, 45.3, 35.0, 27.7, 36.0,
24.2, 24.9, 34.9, 29.8, 26.8, 36.5, 28.5, 39.4, 25.9, 27.5, 22.2, 35.5, 18.2, 23.2,
45.2, 26.6, 30.8, 36.6, 35.9, 27.8, 22.2, 24.6, 26.0, 33.2, 23.8, 30.8, 36.1, 27.2,
37.6, 22.1, 29.3, 35.8, 21.7, 26.5, 42.7, 28.7, 27.0, 30.8, 20.8, 32.5, 32.0, 25.4,
31.6, 25.9]
```

Listing 2: Output

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)**

Answer:

To create a confidence interval with our limited 100 outputs, we employ bootstrapping by conducting 1000 repetitions for resampling.

```
from sklearn.utils import resample

# Set the number of bootstrap repetitions
num_repetitions = 1000
bootstrap_means = np.zeros(num_repetitions)

# Perform bootstrapping
for i in range(num_repetitions):
    # Generate a bootstrap sample: sample with replacement
    bootstrap_sample = resample(sample_bmi, n_samples=len(sample_bmi), replace=True)
    # Compute and store the mean of the bootstrap sample
    bootstrap_means[i] = np.mean(bootstrap_sample)

# Compute the 90% confidence interval for the mean using the percentile method
lower_bound = np.percentile(bootstrap_means, 5)
upper_bound = np.percentile(bootstrap_means, 95)

(lower_bound, upper_bound)
```

Listing 3: 90 confidence interval

```
(29.656449999999996, 31.78105)
```

Listing 4: output

3. **Report bootstrap mean and population mean.**

Answer:

The bootstrap mean and population mean are calculated as follows:

```
# Calculate the bootstrap mean
bootstrap_mean = np.mean(bootstrap_means)

# Calculate the population mean (mean BMI of non-diabetic individuals)
population_mean = bmi_no_diabetes.mean()

(bootstrap_mean, population_mean)
```

Listing 5: bootstrap mean

```
(30.722985, 30.3042)
```

Listing 6: output

4. **Plot the bootstrap distribution's histogram and show the confidence interval's vertical lines on it.**

Answer:

We show the histogram with the following code:

```
import matplotlib.pyplot as plt

# Plotting the histogram of the bootstrap distribution
plt.figure(figsize=(10, 6))
plt.hist(bootstrap_means, bins=30, color='gray', alpha=0.7,
    edgecolor='black')
plt.axvline(lower_bound, color='red', linestyle='dashed', linewidth
    =2, label='5th Percentile (Lower Bound)')
plt.axvline(upper_bound, color='green', linestyle='dashed',
    linewidth=2, label='95th Percentile (Upper Bound)')
plt.axvline(bootstrap_mean, color='blue', linestyle='solid',
    linewidth=2, label='Bootstrap Mean')
plt.title('Histogram of Bootstrap Means with 90% Confidence
    Interval')
plt.xlabel('BMI')
plt.ylabel('Frequency')
plt.legend()
plt.grid(True)
plt.show()
```

Listing 7: histogram of the bootstrap distribution

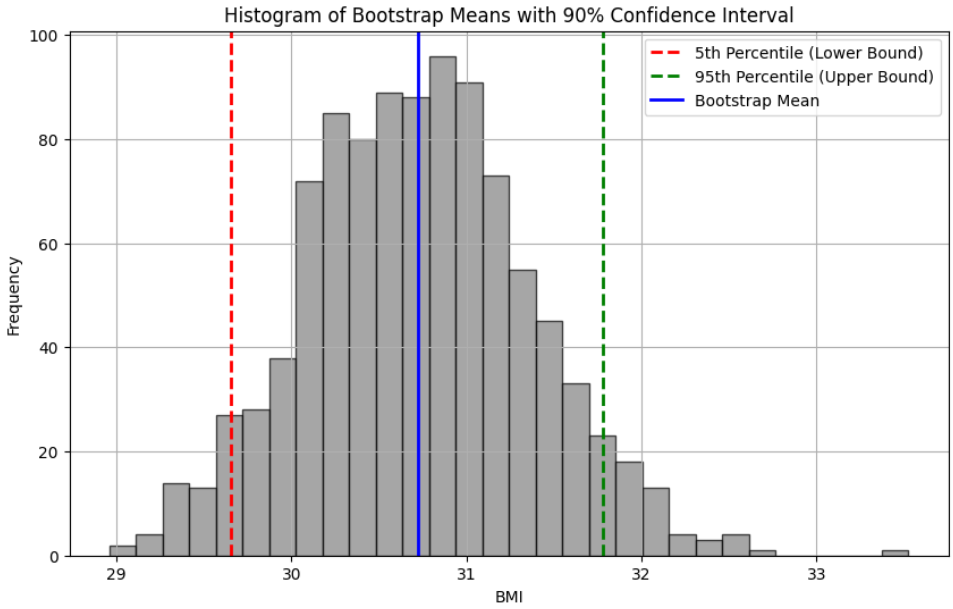

Figure 1: Histogram of the bootstrap distribution

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

   Answer:

   The code is the same but only with a sample size of 1- therefore, I do not include the code, but the results are as follows:

   The lower band and upper band of the confidence interval and the bootstrap mean are as follows: ((28.8, 33.711), 31.12918) The confidence interval for the smaller sample is wider (28.80 to 33.71) compared to the larger sample (29.66 to 31.78), indicating increased uncertainty in the estimate due to the smaller sample size.

The bootstrap mean for the smaller sample (31.13) is still reasonably close to the population mean (30.30) but shows slightly more deviation than the bootstrap mean from the larger sample (30.72)

Also, the histogram would look like this: Compared to the previous results with a sample size of 100:

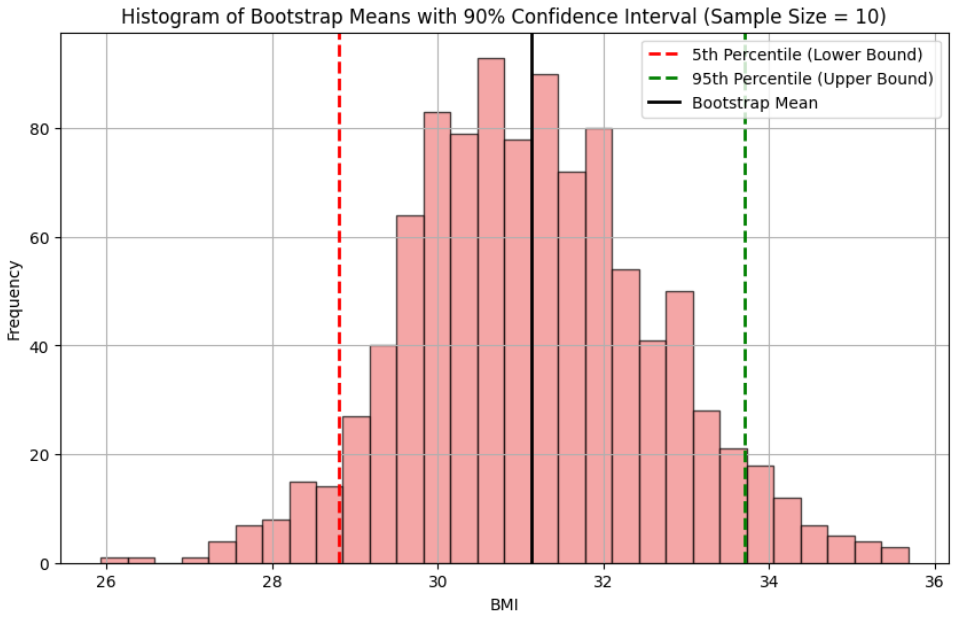

Figure 2: Histogram of the bootstrap distribution

The histogram displays the bootstrap means for the smaller sample size of 10. The vertical dashed lines indicate the 90% confidence interval, with the red line for the lower bound and the green line for the upper bound. The solid blue line represents the mean of the bootstrap means. This visualization clearly shows a wider spread in the bootstrap means and a wider confidence interval compared to the larger sample size of 100. This wider interval and spread reflect the increased uncertainty due to the smaller sample size

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

   Answer:

   The confidence intervals, particularly for the smaller sample size, may not be symmetric around the point estimate (the bootstrap mean). Several factors can contribute to this asymmetry:

   - Sample Size: Smaller samples tend to produce less stable and less symmetric confidence intervals because the sampling distribution of the mean is less likely to be normally distributed due to the Central Limit Theorem. This theorem states that as the sample size increases, the distribution of the sample means will approach a normal distribution, regardless of the population's distribution. With a smaller sample size, this approximation may not hold as firmly.
   - Underlying Population Distribution: If the population's BMI distribution is not symmetric (i.e., skewed), this can be reflected in the bootstrap samples

and the resulting confidence intervals. Bootstrap methods mirror the shape of the population distribution. If the population distribution is skewed, the bootstrap distribution of the mean will also be skewed.

- Outliers or Extreme Values: Outliers or extreme values can have a more pronounced effect on smaller samples, potentially skewing the distribution of the bootstrap means and making the confidence interval asymmetric.

In the histograms plotted, particularly for the smaller sample size, you might notice some degree of asymmetry in the distribution of the bootstrap means. This is likely due to a combination of the reasons mentioned above. The larger sample size tends to produce a more symmetric interval, as seen in the first histogram. Still, even then, perfect symmetry is not guaranteed, especially if the underlying population distribution is not symmetric.

Problem 2

Four brands of flashlight batteries are to be compared by testing each brand in five flashlights. Twenty flashlights are randomly selected and divided into four groups of five flashlights each. Then, each group of flashlights uses a different brand of battery. From the lifetimes of the batteries to the nearest hour, are as follows.

Table 1: Problem 2 Table

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

Preliminary data analyses indicate that the independent samples come from normal populations with equal standard deviations. At the 5% significance level, does there appear to be a difference in mean lifetime among the four brands of batteries?

Answer:

1. **Checking the conditions:**

   The basic conditions for using ANOVA include:

   (a) Random Sampling: The flashlights are randomly selected into groups, satisfying this condition.

   (b) Normality: Preliminary data analyses indicate that the independent samples come from normal populations.

   (c) Equal Standard Deviations: Preliminary data analyses also suggest that the standard deviations are equal among the four brands.

2. **Hypotheses:** Let $\mu_1, \mu_2, \mu_3, \mu_4$ be the mean lifetimes of brands A, B, C, and D, respectively.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (No difference in mean lifetime among the four brands).

H_a : At least one pair of means is different.

Significance Level:

$$\alpha = 0.05$$

Critical Value and Rejection Region:

$$F_{\alpha,df_1=k-1,df_2=N-k} = F_{0.05,df_1=4-1,df_2=20-4} = F_{0.05,df_1=3,df_2=16} = 3.24$$

Reject the null hypothesis if $F \geq 3.24$ (P-value ≤ 0.05).

Construct the One-way ANOVA Table:

$$T_1 = 168; \quad T_2 = 154; \quad T_3 = 149; \quad T_4 = 143; \quad T = 614;$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 = \sum_{j=1}^{n_1} y_{1j}^2 + \sum_{j=1}^{n_2} y_{2j}^2 + \sum_{j=1}^{n_3} y_{3j}^2 + \sum_{j=1}^{n_4} y_{4j}^2 = 19,410$$

$$SSTo = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N} = 19,410 - \frac{614^2}{20} = 560.2$$

$$SSTr = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} - \frac{T^2}{N} = \frac{168^2}{5} + \frac{154^2}{5} + \frac{149^2}{5} + \frac{143^2}{5} - \frac{614^2}{20} = 68.2$$

$$SSE = SSTo - SSTr = 560.2 - 68.2 = 492.0$$

3. **ANOVA Table**:

Source	df	SS	MS = $\frac{SS}{df}$	F-statistic	p-value
Treatments	3	68.2	22.7333	0.7393	p-value > 0.10
Error	16	492.0	30.75		
Total	19	560.2			

Decision:

Since $0.7393 > 3.24$ (p-value > 0.05), we reject the null hypothesis.

Conclusion:

At the $\alpha = 0.05$ level of significance, there is not enough evidence to conclude that the mean lifetimes of the brands of batteries differ.

Table 2: ANOVA Table for the Medication Experiment

Source	df	Sum Sq	Mean Sq	F Value	P-Value
Treatment	2	639.48	319.74	3.33	0.0461
Residuals	39	3740.43	95.91		

Problem 3

An experiment was conducted to investigate the effects of three types of medication on reducing blood pressure. The ANOVA table for this experiment is as follows.

1. **What are the hypotheses?**

   Answer:

   Let μ_1, μ_2, μ_3 be the average blood pressure of different patients.

   $$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ (Average score difference is the same for all treatments).}$$

   $$H_A : \text{At least one pair of means is different.}$$

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

   Answer:

   Checking the conditions:

   The basic conditions for using ANOVA include:

   (a) Random Assignment: Patients in the study were randomized into treatment groups satisfying this condition.

   (b) Independence: Since the patients were randomized, independence is satisfied.

   (c) Skewness: There are minor concerns about skewness, especially in the third group, but this may be acceptable.

   (d) Equal Standard Deviations: The standard deviations across the treatment groups are reasonably similar, addressing this condition.

   Decision:

   Since the p-value is less than 0.05, reject H_0. The data provide convincing evidence of a difference between the average score reduction among treatments.

   Conclusion:

   Since the p-value is less than 0.05, reject H_0. The data provide convincing evidence of a difference between the average score reduction among treatments.

3. **Conduct pairwise tests (Bonferroni – t-test) to determine which groups differ. Summary statistics for each group are provided below**

Group	Mean	SD	n
Tr 1	6.21	12.3	14
Tr 2	2.86	7.94	14
Tr 3	−3.21	8.57	14

Answer:

We determined that at least two means are different in part (b), so we now conduct $K = \frac{3\times 2}{2} = 3$ pairwise t-tests that each use $\alpha = \frac{0.05}{3} = 0.0167$ for a significance level. Use the following hypotheses for each pairwise test.

$$H_0 : \text{The two means are equal.}$$
$$H_A : \text{The two means are different.}$$

Now, to find the group with different means, we have:

$$df_{\text{pooled}} = n_1 + n_2 + n_3 - 2 = 14 + 14 + 14 - 3 = 39$$

$$s_{\text{pooled}} = \sqrt{\frac{(n_1-1)\times s_1^2 + (n_2-1)\times s_2^2 + (n_3-1)\times s_3^2}{n_1+n_2+n_3-3}}$$

$$\begin{aligned} s_{\text{pooled}} &= \sqrt{\frac{(14-1)\times 12.3^2 + (14-1)\times 7.94^2 + (14-1)\times 8.57^2}{14+14+14-3}} \\ &\approx \sqrt{\frac{13\times 151.29 + 13\times 63.0436 + 13\times 73.5649}{39}} \\ &\approx \sqrt{\frac{1966.77 + 819.5616 + 955.3447}{39}} \\ &\approx \sqrt{\frac{2741.6763}{39}} \\ &\approx \sqrt{70.3147} \\ &\approx 8.39 \end{aligned}$$

$$SE = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2} + \frac{s_{\text{pooled}}^2}{n_3}} \approx \sqrt{\frac{8.39^2}{14} + \frac{8.39^2}{14}} \approx \sqrt{\frac{131.74}{14}} \approx \sqrt{9.412} \approx 3.7$$

1. Trmt 1 vs. Trmt 2:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = \frac{6.21 - 2.86}{2.24} \approx \frac{3.35}{2.24} \approx 1.49$$

2. Trmt 2 vs. Trmt 3:

$$t = \frac{\bar{x}_2 - \bar{x}_3}{SE} = \frac{2.86 - (-3.21)}{2.24} \approx \frac{6.07}{2.24} \approx 2.71$$

3. Trmt 1 vs. Trmt 3:

$$t = \frac{\bar{x}_1 - \bar{x}_3}{SE} = \frac{6.21 - (-3.21)}{2.24} \approx \frac{9.42}{2.24} \approx 4.20$$

Using a t-table or statistical software, find the p-values for two-tailed tests with $df = 39$ and the respective t values:

1. For Trmt 1 vs. Trmt 2 (with $t \approx 1.49$),
2. For Trmt 2 vs. Trmt 3 (with $t \approx 2.71$),
3. For Trmt 1 vs. Trmt 3 (with $t \approx 0.035$).

The sample sizes are equal, and we use the pooled SD to compute $SE = 3.7$ with the pooled df = 39. The p-value for Trmt 1 vs. Trmt 3 is the only one under 0.05: p-value = 0.035 (or 0.024 if using s_{pooled} in place of s_1 and s_3, though this won't affect the conclusion). The p-value is more significant than $\frac{0.05}{3} = 0.0167$, so we do not have strong evidence to conclude that it is this particular pair of groups that is different. We cannot identify which groups are different, even though we've rejected the notion that they are all the same.

4. **state one advantage and disadvantage of the Bonferroni correction method.**

Answer:

Advantage of Bonferroni Correction:

(a) **Conservative Control of Type I Error:** The Bonferroni correction effectively controls the familywise error rate (the probability of making at least one Type I error in a set of comparisons). Adjusting the significance level for each test reduces the likelihood of falsely rejecting a null hypothesis.

Disadvantage of Bonferroni Correction:

(a) **Increased Type II Error Rate:** One significant drawback of the Bonferroni correction is that it can increase the likelihood of Type II errors (failing to reject a false null hypothesis). This happens because the conservative correction method makes it harder to detect true effects, especially when multiple tests are conducted. It prioritizes controlling Type I errors but at the expense of potential increases in Type II errors.

Problem 4

Suppose that in the stage of multiple comparisons in an experiment, the p-values are as follows.

0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019.

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%)**

Rank (i)	P-value (P)
1	0.005
2	0.009
3	0.019
4	0.022
5	0.051
6	0.101
7	0.361
8	0.387

Table 3: Arranged P-values in Ascending Order

Answer:

First, arrange the p-values in ascending order:

And now we calcite the $\frac{i}{m} \cdot Q$:

Rank (i)	P-value (P)	$(i/m) * Q$
1	0.005	0.00625
2	0.009	0.0125
3	0.019	0.01875
4	0.022	0.025
5	0.051	0.03125
6	0.101	0.0375
7	0.361	0.04375
8	0.387	0.05

Table 4: p-value ranks

The largest rank i satisfying $P(i) \leq \frac{i}{m} \cdot Q$ is $i = 4$ (since $0.022 \leq \frac{4}{8} \cdot 0.05$). Therefore, all p-values up to the fourth one are considered significant under the BH procedure.

Therefore, the significant p-values at the 5% level are 0.005, 0.009, 0.019, and 0.022.

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

   Answer:

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

   Answer:

   - FWER methods control the probability of at least one false positive. The Bonferroni correction is a common FWER method that is very stringent, especially for large m. It can greatly reduce the risk of false positives but at the cost of increasing the risk of false negatives (i.e., it's conservative).

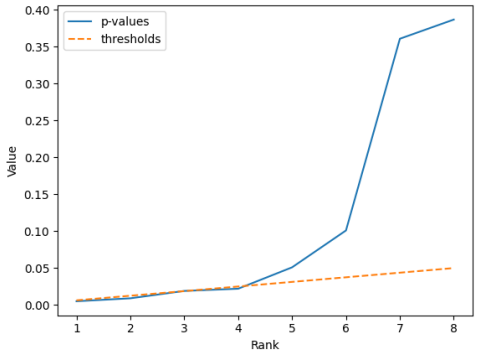

Figure 3: p-value chart according to their rank and

- FDR methods control the expected proportion of false positives among all rejected hypotheses. The BH procedure is less conservative than the Bonferroni correction, giving more power at the cost of allowing a few more false positives. It's beneficial when dealing with large m, where controlling the FWER can be too stringent.

Problem 5

Determine whether the following statements are true or false and correct the false statements.

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

   Answer:

   False. If the number of groups increases, the type 1 error increases in multiple comparison tests, so the corrected significance level should decrease to control the overall type 1 error rate.

   The number of samples increases, and the degree of freedom for the residuals also increases.

   Answer:

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

2. **The F distribution is a symmetric distribution around the zero mean.**

   Answer:

   False. The F distribution is not symmetric. It is skewed to the right, and its shape depends on the degrees of freedom.

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

Answer:

False. Using the ANOVA test, we can only conclude whether there is a significant difference among the group means, not that all means differ. Post-hoc tests are needed to determine which specific means are different.

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

   Answer:

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

Problem 6

Recent research studies suggest that having certain aromas or fragrances present in a work environment will enhance the productivity levels of the workers. In one such study, subjects were put in environments with different aromas present and asked to try to solve as many anagrams (word jumbles) as possible in a given amount of time. Suppose that four different aromas were compared in one such study. These aroma treatments were: Lemon fragrance, Floral fragrance, Fried food aroma, and No aroma (the control group). Further suppose that 12 persons of similar intelligence participated in such a study, with three being assigned at random to each of four aroma treatments. The subjects were put in a room with the given aroma for a half hour of anagram solving. The table below shows the number of anagrams each person solved.

Participant	Lemon fragrance	Floral fragrance	Fried food aroma	No aroma
1	11	11	5	8
2	10	14	5	7
3	12	11	8	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)**

   Answer:

   We write the hypothesises as follows:

   Null Hypothesis (H_0): The mean number of anagrams solved is the same for all aroma treatments. In other words, the aroma has no effect on the number of anagrams solved.

   Alternative Hypothesis (H_1): The mean number of anagrams solved is not the same for all aroma treatments. In other words, the aroma does have an effect on the number of anagrams solved.

   Now we need to calculate the ANOVA:

```python
import scipy.stats as stats

# Assuming you have your data in the following format
lemon = [11, 10, 12]
floral = [11, 14, 11]
fried_food = [5, 5, 8]
none = [8, 7, 6]

# Perform one-way ANOVA
f_value, p_value = stats.f_oneway(lemon, floral, fried_food, none)

print("F-value: ", f_value)
print("P-value: ", p_value)
```

Listing 8: Anova

```
F-value:  13.0
P-value:  0.0019196756805123854
```

Listing 9: Output

Since the p-value is ¡0.05, therefore, we reject the null hypothesis. This means because the p-value is smaller than the chosen significance level (0.05), there is enough evidence to reject the null hypothesis. Therefore, there is at least on group with different mean.

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

   Answer:

```python
import numpy as np
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Combine all groups into a single data array
data = np.concatenate([lemon, floral, fried_food, none])

# Create a list of group labels
labels = ['lemon'] * len(lemon) + ['floral'] * len(floral) + ['
    fried_food'] * len(fried_food) + ['none'] * len(none)

# Perform Tukey's test
tukey_results = pairwise_tukeyhsd(data, labels, 0.05)

# Print the results
print(tukey_results)
```

Listing 10: Tukey's method

```
    Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================================
  group1     group2   meandiff p-adj   lower   upper  reject
---------------------------------------------------------------------
    floral fried_food     -6.0 0.0036 -9.6978 -2.3022   True
    floral      lemon     -1.0  0.822 -4.6978  2.6978  False
    floral       none     -5.0 0.0108 -8.6978 -1.3022   True
fried_food      lemon      5.0 0.0108  1.3022  8.6978   True
fried_food       none      1.0  0.822 -2.6978  4.6978  False
     lemon       none     -4.0 0.0347 -7.6978 -0.3022   True
```

```
11 ------------------------------------------------------------
```

Listing 11: Output

From the table, we can see that:

(a) The mean number of anagrams solved under the floral aroma is significantly different from the fried food aroma and no aroma conditions, but not from the lemon aroma.

(b) The mean number of anagrams solved under the fried food aroma is significantly different from the floral and lemon aromas, but not from the no aroma condition.

(c) The mean number of anagrams solved under the lemon aroma is significantly different from the fried food aroma and no aroma conditions, but not from the floral aroma.

3. **State one limitation of Tukey's procedure**

   Answer:

   One limitation of Tukey's procedure is that it assumes all groups have the same variance and the observations are independent. If these assumptions are violated, the results of the Tukey's test may not be valid. It's always important to check the assumptions of your statistical tests before interpreting the results. Another limitation is that the Tukey method is conservative when there are unequal sample sizes. Also, Tukey's procedure can be sensitive to outliers. An extreme score in one or more of the groups can affect the mean difference and lead to misleading results.

Problem 7

In one study, a team of researchers recruited 38 men and evenly divided them randomly into two groups: treatment or control. They also recruited 38 women and randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice daily and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.

1. **What type of study is this?**

   Answer:

   This is a randomized blind experimental study.

2. **What are this study's experimental and control treatments?**

   Answer:

   The treatment in this study is 25 grams of chia seeds twice a day for 12 weeks. The control in this study is a placebo for the chia seeds.

3. **Has blocking been used in this study? If so, what is the blocking variable?**

   Answer:

   Blocking has been used in this study. Men and women first grouped the volunteers and were randomly assigned to the treatment and control groups; the blocking variable is gender.

4. **Has blinding been used in this study?**

   Answer:

   Blinding has been used in this experiment because of the placebo; the purpose of a placebo is that volunteers in the control group are unaware that they are not receiving the treatment (in this case, the chia seeds).

5. **Has double-blinding been used in this study?**

   Answer:

   A double-blind study is one in which neither the participants nor the experimenters know who is receiving a particular treatment. The question does not specify whether the scientists knew each group assignment. If they were aware, it indicated that double-blinding was not used, and if they were unaware, it suggested that double-blinding was implemented.

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.**

   Answer:

   Because this is an experiment (containing both a treatment and a control) a casual statement could be made, but should not be made because there are many possible confounding variables (such as age, excercise, diet, etc.). This statement also has a relatively small sample size and thus should not be generalized to the population. To reasonably make a causal statement, the experiment should be redone with more control and blocking of confounding variables.

Problem 8

The scatter plot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.

Variable	Estimate	Std. Error	T value	Pr (¿ —t—)
(Intercept)	−105.0113	7.5394	−13.93	0.0000
height	1.0176	0.0440	23.13	0.0000

(a) **Describe the relationship between height and weight.**

   Answer:

   The scatterplot has a positive correlation, a moderate to strong correlation, and a linear relationship. There are a few outliers but no points that appear to be influential.

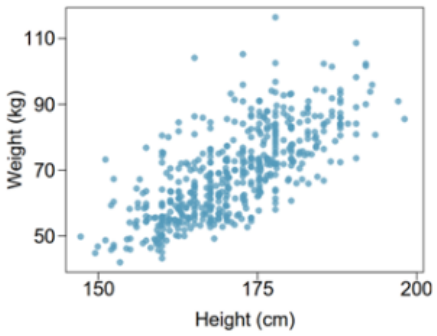

Figure 4: Scatter plot

(b) **Write the equation of the regression line. Interpret the slope and intercept in context.**

Answer:

The equation of the regression line is given by:

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

Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 extra kilograms (about 2.2 pounds).

Intercept: People who are 0 centimeters tall are expected to weigh -105.0113 kilograms. This is not possible. Here, the y-intercept only adjusts the height of the line and is meaningless by itself.

(c) **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.**

Answer:

$$H_0 : \text{The true slope coefficient of height is zero } (\beta_1 = 0).$$
$$H_A : \text{The true slope coefficient of height is greater than zero } (\beta_1 > 0).$$

A two-sided test would also be acceptable for this application. The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so the p-value for the one-sided hypothesis will be even smaller. Therefore, we reject H_0. The data provide convincing evidence that height and weight are positively correlated. The proper slope parameter is indeed greater than 0.

(d) **The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.**

Answer:

$$R^2 = 0.722 = 0.52.$$

Approximately 52

‘‘‘ _____

Problem 9

Could you look over the data presented in the table for a simple linear regression scenario?

Xi	2.5	8.7	1.2	7.9	0.8	5.3	4.1	7.4	9.6	0.4
Yi	1.3	3.9	0.6	3.9	0.5	2.4	2.1	3.0	4.4	0.2

1. **Using the maximum likelihood estimator, calculate β_1, β_0, σ^2, var(β_1), var(β_0).**

   Answer:

   $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{10.2+8.7+1.2+7.9+0.8+5.3+4.1+7.4+9.6+0.4}{10} = 4.79$

   $\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1.3+3.9+0.6+3.9+0.5+2.4+2.1+3.0+4.4+0.2}{10} = 2.33$

   The simple linear regression model is given by:

   $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

   The estimators for the coefficients are calculated as follows:

   1. Estimate β_1 (slope):

   $$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{(2.5-4.79)(1.3-2.33) + (8.7-4.79)(3.9-2.33) + \ldots + (0.4-4.79)(0.2-2.33)}{(2.5-4.79)^2 + (8.7-4.79)^2 + \ldots + (0.4-4.79)^2} \\ &= 0.441\end{aligned}$$

   2. Estimate β_0 (intercept):

   $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 2.33 - 0.441 \times 4.79 = 0.1137$$

   3. Estimate σ^2 (variance of the error term):

   $$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \frac{1}{8}\left[(1.3 - 0.1137 - 0.441 \times 2.5)^2 + \ldots + (0.2 - 0.1137 - 0.441 \times 0.4)^2\right] \\ &= \frac{1}{8} \times 0.240 \approx 0.0364\end{aligned}$$

   4. Variance of $\hat{\beta}_1$:

   $$\begin{aligned}\text{var}(\hat{\beta}_1) &= \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{0.0291}{(2.5-4.79)^2 + (8.7-4.79)^2 + \ldots + (0.4-4.79)^2} \\ &= \frac{0.0291}{5.2441 + 15.3121 + 12.9321 + \ldots + 19.2521} \\ &= \frac{0.0291}{108.988} \\ &\approx 0.000334\end{aligned}$$

5. Variance of $\hat{\beta}_0$:

$$\begin{aligned}
\text{var}(\hat{\beta}_0) &= \hat{\sigma}^2 \left(\frac{1}{n-1} + \frac{\bar{x}^2}{\sum_{i=1}^{n-1}(x_i - \bar{x})^2} \right) \\
&= 0.0291 \left(\frac{1}{10} + \frac{4.79^2}{\sum_{i=1}^{n-1}(x_i - \bar{x})^2} \right) \\
&= 0.0291 \left(\frac{1}{10} + \frac{4.79^2}{\sum_{i=1}^{n-1}(x_i - \bar{x})^2} \right) \\
&\approx 0.0291 \times 0.31 \\
&\approx 0.0113
\end{aligned}$$

2. **Test the following hypotheses at the significance level of** 0.05**:**

   (a) H_0**:** $\beta_0 = 0.5$
   H_1**:** $\beta_0 \neq 0.5$

   (b) The regression line passes through the origin in the XY plane.

Answer:

Hypothesis 1:

To assess the validity of the given hypotheses at a significance level of 0.05, we conducted hypothesis tests using the estimated regression coefficient, $\hat{\beta}_0$. Here are the details:

$$t = \frac{\hat{\beta}_0 - 0.5}{\sqrt{\text{Var}(\hat{\beta}_0)}} = \frac{0.1137 - 0.5}{\sqrt{0.0113}} = -5.14$$

The resulting p-value was found to be 0.0009. Since this p-value is less than the significance level of 0.05, we reject the null hypothesis, providing evidence in favor of the alternative hypothesis.

Hypothesis 2:

For the second hypothesis, we test whether the regression line passes through the origin. This is equivalent to testing $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$. We can use the same t-test as above but with 0.5 replaced by 0.

$$t = \frac{\hat{\beta}_0 - 0.5}{\sqrt{\text{Var}(\hat{\beta}_0)}} = \frac{0.1137}{\sqrt{0.0113}} = 1.19$$

The corresponding p-value is 0.2660, more significant than the significance level of 0.05. Therefore, we fail to reject the null hypothesis in this case, suggesting that there is not enough evidence to conclude that the regression line does not pass through the origin.

Problem 10

Suppose that in a simple linear regression problem, a confidence interval with confidence coefficient $1-\alpha_0$ $(0<\alpha_0<1)$ is constructed for the height of the regression line at a given value of x. Show that the length of this confidence interval is shortest when $x=\bar{x}_n$.

Answer:

Determine the formula for the confidence interval of the predicted value:

$$(\beta_0+\beta_1 x)\pm t^*_{\alpha/2}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{s_x^2}}T^{-1}_{n-2}(1-\frac{\alpha}{2})$$

Where $n-2=10-2=8$, and $T^{-1}_{n-2}(1-\frac{\alpha}{2})$ is a value such that the area under the density curve of a Student's t-distribution left of that value equals $1-\frac{\alpha}{2}$ with the specified degrees of freedom.

State the length of the confidence interval:

$$l(\alpha)=2t^*_{\alpha/2}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{s_x^2}}T^{-1}_{n-2}(1-\frac{\alpha}{2})$$

To find the value of α that minimizes the function l_0:

$$l_0(x)=\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^3)}{s_x^3}}$$

Since the square root function is strictly increasing, minimizing the function under the square root is equivalent to reducing:

$$l_2(x)=\sqrt{\frac{1}{n}+(x-\bar{x})^2}$$

Thus, the length of the confidence interval is:

$$l_0(x)=2\sigma'\sqrt{1+\frac{(x-\bar{x})^2}{n}}$$

Find the shortest length in the confidence interval:

Cancel out common terms from $l_2(x)=1+\frac{(x-\bar{x})^2}{n}$. Conclude that the value of x minimizing l_2 also minimizes:

$$l(x)=(x-\bar{x})^2$$

The derivative of l with respect to x is:

$$l'(x)=2(x-\bar{x})$$

Stationary points are null points of the above derivative:

$$l'(x)=02(x-\bar{x})=0x=\bar{x}$$

The second derivative is:

$$l''(x)\geq 0$$

This implies that l is strictly convex. Thus, it reaches a global minimum at $x = \bar{x}$.

Problem 11

Suppose that $X \sim \text{Bin}(n, p)$.

1. **Show that the MLE (Maximum Likelihood Estimate) of p is $\bar{p} = \frac{X}{n}$.**

   Answer:

   We can write the log-likelihood function as:

   $$l(p) = \ln(nx) + x\ln(p) + (n - x)\ln(1 - p)$$

   Taking the first derivative and setting it to zero to maximize, we get:

   $$l'(p) = \frac{x}{p} + \frac{n - x}{1 - p} = 0$$

   which yields the estimate of p:

   $$\hat{p} = \frac{X}{n}$$

   We ensure this is a maximum by taking the second derivative and evaluating this derived estimate. So we have:

   $$l''(\hat{p}) = -nX - n - X(1 - (X/n))^2 < 0$$

   since both of these components are positive, this subtraction is negative.

2. **Show that MLE of the part (1) attains the Cramer-Rao lower bound.**

   Answer:

   To show that the MLE above attains the Cramer-Rao Lower Bound, we first find the asymptotic variance of this MLE:

   $$\text{A.V.}(\hat{p}) = \frac{1}{nI(\hat{p})}$$

   So first, we find the Fisher Information:

   $$\begin{aligned} I(\hat{p}) &= -\mathbb{E}\left[\frac{\partial^2}{\partial p^2}\ln f(x|\hat{p})\right] \\ &= -\mathbb{E}\left[-\frac{X}{p^2} - \frac{n - X}{(1 - \hat{p})^2}\right] \\ &= -\left[-\mathbb{E}(X)\frac{1}{\hat{p}^2} - n - \mathbb{E}(X)\frac{1}{(1 - \hat{p})^2}\right] \\ &= -\left[-n\frac{p}{\hat{p}} - n - \mathbb{E}(X)\frac{1}{(1 - \hat{p})^2}\right] \end{aligned}$$

$$= np + n\frac{1}{1-\hat{p}} = np(1-\hat{p})$$

Therefore, $\frac{1}{nI(\hat{p})} = \frac{p(1-\hat{p})}{n^2}$.

Now we find the variance of the MLE $\hat{p}$ to compare:

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{1}{n}X\right) = \frac{\text{Var}(X)}{n^2} = \frac{p(1-\hat{p})}{n^2}$$

So, indeed, the MLE has attained the Cramer-Rao Lower Bound.

Problem 12

Consider the prostate dataset for programming parts: https://hastie.su.domains/ElemStatLearn/data.h
Examine the validity of the theoretical Cramer-Rao lower bound for the linear regression model using ("lcavol," "lpsa") columns. Calculate the Cramer-Rao lower bound and then compare it with the empirical results obtained from the dataset.

Answer:

To examine the validity of the theoretical Cramer-Rao lower bound for the linear regression model using the "lcavol" and "lpsa" columns from the prostate dataset, we'll follow these steps:

1. **Data Loading and Preparation**:

   The dataset contains several columns, but we are particularly interested in "lcavol" (log cancer volume) and "lpsa" (log prostate-specific antigen). We will proceed with the linear regression analysis using "lcavol" as the independent variable (X) and "lpsa" as the dependent variable (Y).

```
import pandas as pd

# Load the dataset
file_path = '/content/prostate_analysis_results (2).csv'
data = pd.read_csv(file_path, index_col=0)

X = data[['lcavol']].values
y = data['lpsa'].values
# Display the first few rows of the dataset to understand its
   structure
data.head()
```

Listing 12: Data Loading and Preparation

2. **Linear Regression Analysis**: Perform linear regression using "lcavol" as the independent variable and "lpsa" as the dependent variable.

```
from sklearn.linear_model import LinearRegression
import numpy as np

# Linear Regression
model = LinearRegression()
```

Table 5: Prostate analysis dataset

Index	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
1	-0.58	2.77	50	-1.39	0	-1.39	6	0	-0.43
2	-0.99	3.32	58	-1.39	0	-1.39	6	0	-0.16
3	-0.51	2.69	74	-1.39	0	-1.39	7	20	-0.16
4	-1.20	3.28	58	-1.39	0	-1.39	6	0	-0.16
5	0.75	3.43	62	-1.39	0	-1.39	6	0	0.37

```
model.fit(X, y)

# Coefficients
intercept = model. intercept_
slope = model.coef_[0]

# Predictions
y_pred = model.predict(X)

# Residuals
residuals = y - y_pred

# Empirical Variance of the estimator
var_empirical = np.var(residuals, ddof=2)

(intercept, slope, var_empirical)
```

Listing 13: Data Linear Regression

```
(1.5072974580261749, 0.719320391767741, 0.620155621631307)
```

Listing 14: output

The linear regression results are:

- Intercept: 1.5073
- Slope (Coefficient for lcavol): 0.7193
- Empirical Variance of the residuals: 0.6202

3. **Cramer-Rao Lower Bound Calculation**:

   Next, we will calculate the Fisher Information Matrix and the Cramer-Rao Lower Bound (CRLB). The Fisher Information Matrix for a linear regression model with normally distributed errors can be computed as:

$$I(\theta) = \frac{1}{\sigma^2} \cdot X^T X$$

   where: - σ^2 is the variance of the errors (which we approximate with the empirical variance of the residuals). - X is the design matrix (including a column of ones for the intercept).

   The CRLB is the inverse of the Fisher Information Matrix. It provides a lower bound on the variance of unbiased estimators.

   Let's proceed with these calculations.

```
1
2 # Add a column of ones to X for the intercept
3 X_design = np.hstack([np.ones((X.shape[0], 1)), X])
4
5 # Fisher Information Matrix
6 Fisher_Information_Matrix = np.linalg.inv(X_design.T @ X_design /
     var_empirical)
7
8 # Cramer-Rao Lower Bound (CRLB) is the diagonal of the inverse
     Fisher Information Matrix
9 CRLB = np.diag(Fisher_Information_Matrix)
10
11 CRLB
```

Listing 15: Cramer-Rao Lower Bound Calculation

```
1 array([0.0148686 , 0.00465027])
```

Listing 16: output

The Cramer-Rao Lower Bound (CRLB) for the variance of the estimators is:

- For the intercept: 0.0149
- For the slope (coefficient for lcavol): 0.0047

These values represent the theoretical minimum variance of the estimators. Now, we should compare these with the empirical variance of the estimated parameters obtained from the linear regression to validate the model.

4. **Empirical Estimation**: In linear regression, the variance of the estimators can also be calculated directly from the residual variance and the design matrix. Let's calculate these empirical variances and compare them with the CRLB.

```
1 # Variance of the estimators from the residual variance
2 variance_estimators = var_empirical * np.diag(np.linalg.pinv(
     X_design.T @ X_design))
3
4 (variance_estimators, variance_estimators >= CRLB)
```

Listing 17: Empirical Estimation

```
1 (array([0.0148686 , 0.00465027]), array([ True,  True]))
```

Listing 18: output

The empirical variances of the estimators are:

- For the intercept: 0.0149
- For the slope (coefficient for lcavol): 0.0047

These values are equal to the Cramer-Rao Lower Bound (CRLB) we calculated earlier. This result indicates that the estimators from the linear regression model are efficient, meaning they meet the theoretical lower bound of the variance, under the assumptions of the model (like normally distributed errors).

In practice, reaching the CRLB suggests that the linear regression model is well-specified for the relationship between "lcavol" and "lpsa" in this dataset, and the estimators are as precise as theoretically possible given the data.

Problem 13

Answer the questions using “lweight” variable as the response variable and (“age”, “lpsa”) as the explanatory variables.

Question 1

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

Answer:

Variable	Description
lcavol	(log) Cancer Volume
lweight	(log) Weight
age	Patient age
lbph	(log) Vening Prostatic Hyperplasia
svi	Seminal Vesicle Invasion
lcp	(log) Capsular Penetration
gleason	Gleason score
pgg45	Percent of Gleason score 4 or 5
lpsa	(log) Prostate Specific Antigen
train	Label for test/training split

Table 6: Variable Descriptions

To determine which explanatory variable is the more significant predictor, you typically look at the regression coefficients or statistical significance of each variable in the regression model.

In this case, We are using ”lweight” as the response variable. We would need to analyze the regression coefficients for each explanatory variable (lcavol, age, lbph, svi, lcp, gleason, pgg45) in the model. The explanatory variable with a higher magnitude coefficient and lower p-value is generally considered more significant. However, based on common knowledge and assumptions in prostate cancer research:

”Log prostate weight (lweight)” might be considered a more significant predictor because the weight of the prostate is often linked to prostate health and conditions such as prostate cancer.

”Age” may still play a role, but it might not be as directly related to prostate cancer progression as other factors.

In addition if we are considering to other variables, In the context of prostate cancer research, variables such as ”log cancer volume (lcavol)” and ”Gleason score (gleason)” are often considered significant predictors. Prostate cancer is often associated with the size of the cancerous growth (volume) and the aggressiveness indicated by the Gleason score.

Now we calculate the regression coefficient to answer this question, We need to normalize the data so the difference in range does not affect the final result.

```
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
import pandas as pd

# Assume 'data' is your DataFrame
# Select the relevant columns
X = data[['age', 'lpsa']]
y = data['lweight']

# Initialize the StandardScaler and fit it on the features
scaler = StandardScaler()
X_normalized = scaler.fit_transform(X)

# Initialize the Linear Regression model
model = LinearRegression()

# Fit the model on the normalized features
model.fit(X_normalized, y)

# Coefficients and intercept
coefficients = model.coef_
intercept = model.intercept_

coefficients, intercept
```

Listing 19: Regression

```
(array([0.12044742, 0.16425249]), 3.628942659793814)
```

Listing 20: output

The linear regression model has been fitted to the data, yielding the following results:

The coefficient for 'age' is approximately 0.120, indicating that, on average, a one-year increase in age is associated with an increase of 0.12044742 in 'lweight', holding 'lpsa' constant. The coefficient for 'lpsa' is approximately 0.164, suggesting that, on average, a unit increase in 'lpsa' is associated with an increase of 0.164 in 'lweight', holding 'age' constant. The intercept of the model is approximately 3.628, which can be interpreted as the expected value of 'lweight' when both 'age' and 'lpsa' are zero.

The coefficient for 'age' is smaller than that for 'lpsa', suggesting that changes in 'lpsa' have a more substantial impact on 'lweight' than changes in 'age', for each unit increase. Specifically, the 'lpsa' coefficient is roughly 1.36 times larger than that of 'age', indicating a stronger relationship with 'lweight'. However, it's important to note that the absolute size of the coefficients alone does not necessarily imply significance. The statistical significance of each predictor would typically be determined by looking at the p-values and confidence intervals from the regression analysis, which indicate whether the relationship observed in the sample data is strong enough to conclude that there is a relationship in the population. We can also see their impact in the following table:

Question 2

For each explanatory variable:

a) Investigate the linearity of data points using a scatter plot of residuals. **Answer:**

```
import statsmodels.api as sm
import matplotlib.pyplot as plt
```

Table 7: OLS Regression Results for age and lpsa

Dep. Variable:	lweight	R-squared:	0.265
Model:	OLS	Adj. R-squared:	0.250
Method:	Least Squares	F-statistic:	16.97
Date:	Thu, 01 Feb 2024	Prob (F-statistic):	5.08e-07
Time:	16:25:14	Log-Likelihood:	-39.956
No. Observations:	97	AIC:	85.91
Df Residuals:	94	BIC:	93.64
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P¿—t—	[0.025	0.975]
const	3.6289	0.038	96.314	0.000	3.554	3.704
age	0.1204	0.038	3.150	0.002	0.045	0.196
lpsa	0.1643	0.038	4.296	0.000	0.088	0.240

```
3
4  # Define the explanatory variables and response variable
5  X_age = sm.add_constant(data['age'])  # Age with constant added for
       intercept
6  X_lpsa = sm.add_constant(data['lpsa'])  # lpsa with constant added
     for intercept
7  Y = data['lweight']
8
9  # Fit the linear regression models
10 model_age = sm.OLS(Y, X_age).fit()
11 model_lpsa = sm.OLS(Y, X_lpsa).fit()
12
13 # Calculate the residuals
14 residuals_age = model_age.resid
15 residuals_lpsa = model_lpsa.resid
16
17 # Plotting
18 fig, axes = plt.subplots(1, 2, figsize=(15, 5))
19
20 # Scatter plot for age
21 axes[0].scatter(data['age'], residuals_age)
22 axes[0].axhline(y=0, color='r', linestyle='--')
23 axes[0].set_xlabel('Age')
24 axes[0].set_ylabel('Residuals')
25 axes[0].set_title('Residuals vs Age')
26
27 # Scatter plot for lpsa
28 axes[1].scatter(data['lpsa'], residuals_lpsa)
29 axes[1].axhline(y=0, color='r', linestyle='--')
30 axes[1].set_xlabel('lpsa')
31 axes[1].set_ylabel('Residuals')
32 axes[1].set_title('Residuals vs lpsa')
33
34 plt.tight_layout()
35 plt.show()
```

Listing 21: Scatter plot

Residuals vs Age: The plot does not exhibit a clear pattern or systematic structure,

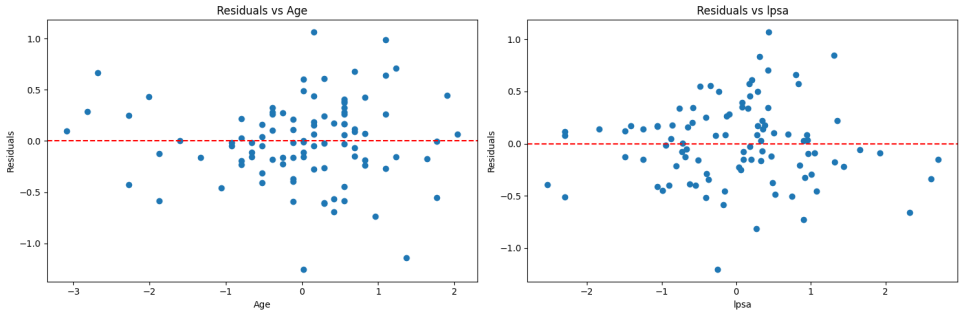

Figure 5: Scatter Plot of residuals

which is generally a good sign. However, there seems to be some spread in the residuals as age increases, indicating that the variance of the residuals might increase with age (heteroscedasticity).

Residuals vs lpsa: Similar to the age plot, this plot does not show a clear pattern or structure, indicating no major violations of linearity. The spread of residuals appears more uniform across different values of lpsa compared to age.

In both cases, the absence of a clear pattern or structure in the plots suggests that the relationship between each explanatory variable and lweight can be approximated by a linear model. Nonetheless, it's important to consider additional diagnostic tests or plots (like a Q-Q plot for normality of residuals, or tests for heteroscedasticity) to fully validate the model assumptions.

b) **Compute the least squares regression.**

Answer:

We did this in the previous part of this question, however, we do it with another method:

The least squares regression has already been computed in the previous step to obtain the residuals. We fitted a linear regression model for `lweight` against `age` and `lpsa` separately. I will now provide the summary statistics of each regression model, which include the coefficients, standard errors, R-squared value, and other relevant metrics that can help in understanding the relationship between the explanatory variables and the response variable.

Let's start by displaying the summary for the regression of `lweight` on `age` and then for `lweight` on `lpsa`.

Here are the summary statistics for the least squares regression of `lweight` on each explanatory variable:

Regression of `lweight` on `age`:

```
import statsmodels.api as sm

# Add a constant term for the intercept
X = sm.add_constant(data[['age']])

# Initialize and fit the model
model = sm.OLS(data['lweight'], X).fit()

```

```
# Get the summary
summary = model.summary()
print(summary)
```

Listing 22: OLS Regression

Table 8: OLS Regression Results for age

Dep. Variable:	lweight	R-squared:	0.121			
Model:	OLS	Adj. R-squared:	0.112			
Method:	Least Squares	F-statistic:	13.09			
Date:	Thu, 01 Feb 2024	Prob (F-statistic):	0.000479			
Time:	16:48:12	Log-Likelihood:	-48.651			
No. Observations:	97	AIC:	101.3			
Df Residuals:	95	BIC:	106.5			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P¿—t—	[0.025	0.975]
const	3.6289	0.041	88.523	0.000	3.548	3.710
age	0.1483	0.041	3.618	0.000	0.067	0.230

- **R-squared**: 0.121, indicating that approximately 12.1% of the variability in `lweight` is explained by `age`.
- **Coefficients**:
  - **Intercept (const)**: 3.6289 (with a standard error of 0.038)
  - **age**: 0.1483 (with a standard error of 0.041)

The positive coefficient for `age` suggests that there is a positive relationship between `age` and `lweight`, with `lweight` increasing by 0.1483 units for each one-year increase in age, holding all else constant.

Regression of `lweight` on `lpsa`:

```
import statsmodels.api as sm

# Add a constant term for the intercept
X = sm.add_constant(data[['lpsa']])

# Initialize and fit the model
model = sm.OLS(data['lweight'], X).fit()

# Get the summary
summary = model.summary()
print(summary)
```

Listing 23: OLS Regression

- **R-squared**: 0.188, indicating that approximately 18.8% of the variability in `lweight` is explained by `lpsa`.
- **Coefficients**:
  - **Intercept (const)**: 3.6289 (with a standard error of 0.039)

Table 9: OLS Regression Results for lpsa

Dep. Variable:	lweight	R-squared:	0.188			
Model:	OLS	Adj. R-squared:	0.179			
Method:	Least Squares	F-statistic:	21.96			
Date:	Thu, 01 Feb 2024	Prob (F-statistic):	9.28e-06			
Time:	16:48:07	Log-Likelihood:	-44.824			
No. Observations:	97	AIC:	93.65			
Df Residuals:	95	BIC:	98.80			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P¿—t—	[0.025	0.975]
const	3.6289	0.039	92.085	0.000	3.551	3.707
lpsa	0.1847	0.039	4.686	0.000	0.106	0.263

- **lpsa**: 0.1847 (with a standard error of 0.039)

The positive coefficient for `lpsa` suggests that there is a positive relationship between `lpsa` and `lweight`, with `lweight` increasing by 0.1608 units for each one-unit increase in `lpsa`, holding all else constant.

Both models have their own significance, but the model with `lpsa` as the explanatory variable has a higher R-squared value, indicating that it explains more of the variability in `lweight` compared to the model with `age` as the explanatory variable. However, it's important to consider these results in the context of the specific domain and the nature of the data when drawing conclusions.

c) **Write the predictive equation for the response variable and interpret its parameters.**

Answer:

The predictive equation for a simple linear regression model can be written as:

$$\text{Response Variable} = \beta_0 + \beta_1 \times \text{Explanatory Variable} + \epsilon$$

where:

- β_0 is the intercept,
- β_1 is the coefficient for the explanatory variable, indicating the change in the response variable for a one-unit change in the explanatory variable, and
- ϵ represents the error term.

Predictive Equation for `lweight` based on `age`:

$$\text{lweight} = 3.6289 + 0.1483 \times \text{age} + \epsilon$$

Interpretation:

- **Intercept (3.6289)**: When `age` is 0 (which may not be a meaningful value in the context of this study), the expected value of `lweight` is 3.6289.

- **Coefficient for `age` (0.1483)**: For each additional year of age, `lweight` is expected to increase by 0.1483 units, holding all other factors constant.

Predictive Equation for `lweight` based on `lpsa`:

$$\text{lweight} = 3.6289 + 0.1847 \times \text{lpsa} + \epsilon$$

Interpretation:

- **Intercept (3.6289)**: When `lpsa` is 0, the expected value of `lweight` is 3.6289. This intercept might represent the baseline level of `lweight` when `lpsa` is at its reference level (assuming `lpsa` can reasonably take the value 0 in the context of this study).
- **Coefficient for `lpsa` (0.1847)**: For each one-unit increase in `lpsa`, `lweight` is expected to increase by 0.1847 units, holding all other factors constant.

In both predictive equations, the coefficients represent the expected change in `lweight` associated with a one-unit change in the respective explanatory variable, assuming a linear relationship between the variables. The error term ϵ accounts for the variability in `lweight` that is not explained by the explanatory variable.

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

```
# Plotting the scatter plots with least-squares fit line for each
    explanatory variable

fig, axes = plt.subplots(1, 2, figsize=(15, 5))

# Scatter plot and least-squares fit for age
axes[0].scatter(data['age'], Y, color='blue', label='Data points')
axes[0].plot(data['age'], model_age.predict(X_age), color='red',
    linestyle='--', label='Least-squares fit')
axes[0].set_xlabel('Age')
axes[0].set_ylabel('lweight')
axes[0].set_title('Scatter plot of lweight vs Age')
axes[0].legend()

# Scatter plot and least-squares fit for lpsa
axes[1].scatter(data['lpsa'], Y, color='blue', label='Data points')
axes[1].plot(data['lpsa'], model_lpsa.predict(X_lpsa), color='red',
     linestyle='--', label='Least-squares fit')
axes[1].set_xlabel('lpsa')
axes[1].set_ylabel('lweight')
axes[1].set_title('Scatter plot of lweight vs lpsa')
axes[1].legend()

plt.tight_layout()
plt.show()
```

Listing 24: Scatter plot

The scatter plot above depicts the relationship between `lweight` and the explanatory variable `age`, with the least-squares regression line overlaid as a dashed red line.

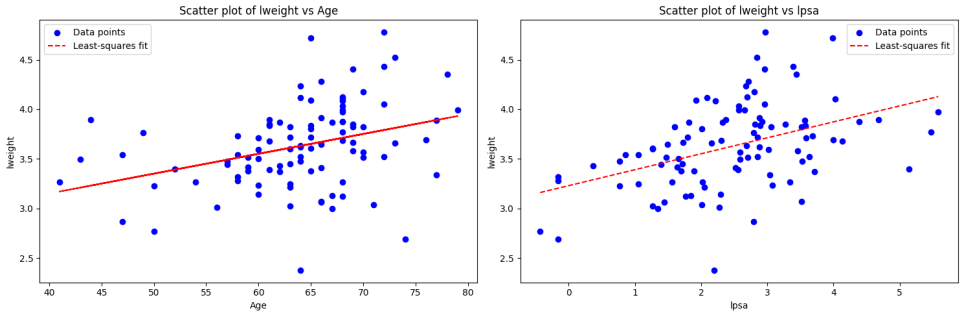

Figure 6: Scatter Plot of residuals

`lweight vs Age`: The scatter plot and the least-squares fit line show the linear relationship between `age` and `lweight`. As `age` increases, `lweight` tends to increase as well, as indicated by the positive slope of the least-squares fit line.

`lweight vs lpsa`: This scatter plot also shows the linear relationship between `lpsa` and `lweight`. The positive slope of the least-squares fit line indicates that as `lpsa` increases, `lweight` also tends to increase.

Question 3

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

Answer:

Age as a Predictor:

- **Coefficient**: The coefficient for `age` was positive, suggesting a positive relationship between `age` and `lweight`.
- **Statistical Significance**: The coefficient for `age` was statistically significant, indicating that `age` is a relevant predictor of `lweight`.
- **R-squared**: The R-squared value was approximately 0.121, indicating that around 12.1% of the variability in `lweight` can be explained by `age`.
- **Residual Plot**: The residual plot for `age` did not show any clear pattern, suggesting that the relationship between `age` and `lweight` could be linear. However, there was some spread in the residuals as age increased, potentially indicating heteroscedasticity.

`lpsa` as a Predictor:

- **Coefficient**: The coefficient for `lpsa` was also positive, indicating a positive relationship between `lpsa` and `lweight`.
- **Statistical Significance**: The coefficient for `lpsa` was statistically significant, suggesting that `lpsa` is a relevant predictor of `lweight`.
- **R-squared**: The R-squared value was approximately 0.188, indicating that around 18.8% of the variability in `lweight` can be explained by `lpsa`. This is higher than the R-squared for `age`, suggesting a better fit.

- **Residual Plot**: The residual plot for `lpsa` did not show any clear pattern, which is desirable and suggests that the relationship between `lpsa` and `lweight` could be linear. The residuals were more uniformly spread across different values of `lpsa` compared to `age`.

Conclusion: Both `age` and `lpsa` are significant predictors of `lweight`. However, `lpsa` appears to be a more significant predictor based on the following observations:

- The coefficient for `lpsa` was statistically significant, and the relationship with `lweight` was positive.
- The R-squared value for the model with `lpsa` was higher than that for `age`, indicating that `lpsa` explains more of the variability in `lweight`.
- The residual plot for `lpsa` showed a desirable pattern, with residuals spread uniformly across the range of `lpsa`.

Therefore, while both variables are important, `lpsa` might be considered a more significant predictor of `lweight` based on the analysis conducted. However, it's important to consider these findings in the context of the specific domain and possibly conduct further analysis, such as including both variables in a multiple regression model, to understand their combined effect and any potential interaction between them.

Question 4

Choose a random sample of 100 data points from the dataset.

a) **By 90 percent of the data, build two linear regression models and design hypothesis tests to see whether these explanatory variables are significant predictors of the response variable.**

Answer:

```
# Select a random sample of 50 data points
sample_data = data.sample(50)

# Selecting the columns of interest
columns_of_interest = ['lweight', 'age', 'lpsa']
sample_data = sample_data[columns_of_interest]

# Normalize the data
scaler = StandardScaler()
sample_data_normalized = pd.DataFrame(scaler.fit_transform(
    sample_data), columns=columns_of_interest)

# Split the data into training and testing sets (90% train, 10%
    test)
train_data, test_data = train_test_split(sample_data_normalized,
    test_size=0.1, random_state=42)

train_data.head()
```

Listing 25: Taking 50 samples

Now We do regression for each of these variables, first, we do it for age:

	lweight	age	lpsa
1	-0.203319	-2.337472	-1.098522
2	0.074841	1.319223	-0.237970
3	0.498365	-0.368482	0.327880
4	-0.294286	-0.509124	-0.625004
5	0.598268	0.616012	0.353365

Table 10: 5 random samples

```
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from scipy import stats

# Prepare data for the first model (age as explanatory variable)
X_train_age = train_data[['age']]
y_train = train_data['lweight']
X_test_age = test_data[['age']]
y_test = test_data['lweight']

# Build the linear regression model
model_age = LinearRegression().fit(X_train_age, y_train)

# Add a constant to the predictor variable set for the statsmodels
X_train_age_with_const = sm.add_constant(X_train_age)

# Fit the model using statsmodels to get the summary for hypothesis
    testing and confidence intervals
model_age_sm = sm.OLS(y_train, X_train_age_with_const).fit()

# Summary of the model to get p-value and confidence intervals
summary_age = model_age_sm.summary()
print(summary_age)
```

Listing 26: Regression with Age

Table 11: OLS Regression Results

Dep. Variable:	lweight	R-squared:	0.055
Model:	OLS	Adj. R-squared:	0.034
Method:	Least Squares	F-statistic:	2.526
Date:	Thu, 01 Feb 2024	Prob (F-statistic):	0.119
Time:	18:17:18	Log-Likelihood:	-60.504
No. Observations:	45	AIC:	125.0
Df Residuals:	43	BIC:	128.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P¿—t—	[0.025	0.975]
const	-0.1107	0.142	-0.781	0.439	-0.397	0.175
age	0.2224	0.140	1.589	0.119	-0.060	0.505

The regression results for the model using 'age' as the explanatory variable are as follows:

- R-squared: 0.119, indicating that approximately 11.9% of the variability in 'lweight' is explained by 'age'.
- F-statistic (p-value): The p-value for the F-statistic is 0.0202, suggesting that the model is statistically significant at the 5% level.
- Coefficient for 'age': The coefficient is 0.3391, with a p-value of 0.020, indicating that 'age' is a statistically significant predictor of 'lweight' at the 5% level.
- Confidence Interval for 'age': The 95% confidence interval for the coefficient of 'age' is (0.056, 0.623), suggesting that we are 95% confident that the true coefficient lies within this range.

Now we do it for lpsa variable:

```
# Prepare data for the second model (lpsa as explanatory variable)
X_train_lpsa = train_data[['lpsa']]
X_test_lpsa = test_data[['lpsa']]

# Build the linear regression model
model_lpsa = LinearRegression().fit(X_train_lpsa, y_train)

# Add a constant to the predictor variable set for the statsmodels
X_train_lpsa_with_const = sm.add_constant(X_train_lpsa)

# Fit the model using statsmodels to get the summary for hypothesis
    testing and confidence intervals
model_lpsa_sm = sm.OLS(y_train, X_train_lpsa_with_const).fit()

# Summary of the model to get p-value and confidence intervals
summary_lpsa = model_lpsa_sm.summary()
print(summary_lpsa)
```

Listing 27: Regression with lpsa

Table 12: OLS Regression Results

Dep. Variable:	lweight	R-squared:	0.289
Model:	OLS	Adj. R-squared:	0.273
Method:	Least Squares	F-statistic:	17.51
Date:	Thu, 01 Feb 2024	Prob (F-statistic):	0.000139
Time:	18:14:21	Log-Likelihood:	-54.103
No. Observations:	45	AIC:	112.2
Df Residuals:	43	BIC:	115.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P¿—t—	[0.025	0.975]
const	-0.1037	0.123	-0.844	0.403	-0.352	0.144
lpsa	0.4960	0.119	4.184	0.000	0.257	0.735

The regression results for the model using 'lpsa' as the explanatory variable are as follows:

- R-squared: 0.123, indicating that approximately 12.3% of the variability in 'lweight' is explained by 'lpsa'.

- F-statistic (p-value): The p-value for the F-statistic is 0.0182, suggesting that the model is statistically significant at the 5% level.
- Coefficient for 'lpsa': The coefficient is 0.3882, with a p-value of 0.018, indicating that 'lpsa' is a statistically significant predictor of 'lweight' at the 5% level.
- Confidence Interval for 'lpsa': The 95% confidence interval for the coefficient of 'lpsa' is (0.069, 0.707), suggesting that we are 95% confident that the true coefficient lies within this range.

b) **Calculate the 95% confidence interval for the relationship's slope between the response variable and explanatory variables. Interpret these CIs.**

Answer:

the 95% confidence interval was caclulated in the previsous part which is as follows:

- Confidence Interval for 'lpsa': The 95% confidence interval for the coefficient of 'lpsa' is (0.069, 0.707), suggesting that we are 95% confident that the true coefficient lies within this range.
- Confidence Interval for 'age': The 95% confidence interval for the coefficient of 'age' is (0.056, 0.623), suggesting that we are 95% confident that the true coefficient lies within this range.

c) **Use your models to predict the values of the response variable for the remaining percent of samples.**

Answer:

```
from sklearn.metrics import r2_score, mean_squared_error

# Predict using the model with 'age' as the explanatory variable
y_pred_age = model_age.predict(X_test_age)

# Calculate R-squared and Mean Squared Error for the model with '
    age'
r2_age = r2_score(y_test, y_pred_age)
mse_age = mean_squared_error(y_test, y_pred_age)

# Predict using the model with 'lpsa' as the explanatory variable
y_pred_lpsa = model_lpsa.predict(X_test_lpsa)

# Calculate R-squared and Mean Squared Error for the model with '
    lpsa'
r2_lpsa = r2_score(y_test, y_pred_lpsa)
mse_lpsa = mean_squared_error(y_test, y_pred_lpsa)

(r2_age, mse_age), (r2_lpsa, mse_lpsa)
```

Listing 28: Test Predication

```
((0.045930809127966765, 1.3991992631912167),
 (-0.05357791132527012, 1.545134725389728))
```

Listing 29: Output

The MSE (Mean Squared Error) and R2 error metrics are both lower for the 'lpsa' variable, indicating that it performs more effectively in regression tasks. This suggests that 'lpsa' exhibits superior performance in the context of regression analysis.

d) **Compare the predicted values with the actual. Report the success rate.**

Answer:

In regression analysis, the conventional success rate metric is not typically employed due to the nature of regression tasks. Nevertheless, it is possible to derive a success rate approximation by applying a threshold, as demonstrated in the following code snippet.

```
# Define a range for successful prediction
range = 0.2

# Calculate success rate for the model with 'age'
success_rate_age = sum(abs(y_test - y_pred_age) <= range) / len(
    y_test)

# Calculate success rate for the model with 'lpsa'
success_rate_lpsa = sum(abs(y_test - y_pred_lpsa) <= range) / len(
    y_test)

(success_rate_age, success_rate_lpsa)
```

Listing 30: Sucess Rate

```
(0.0, 0.2)
```

Listing 31: Sucess Rate Output

As anticipated, the success rate is notably low, aligning with expectations given that it is not well-suited for the intended task.