Preparing data [MCQ] (Version : 0)

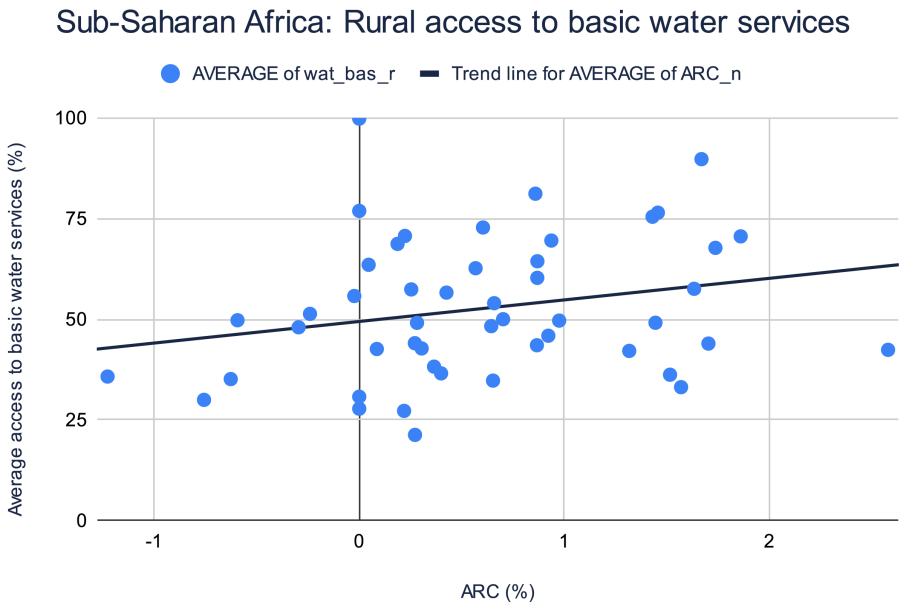
TEST

Correct Answer

Answered in 10.983333333333 Minutes

Question 1/10

True or false? The expected correlation coefficient of the relationship in the visualisation is closer to +0.5 than -1.



- [x] True
- [ ] False

Explanation:

The variables increase in the same direction. In other words, as the independent variable (ARC) increases the dependent variable (Average access to basic water services) increases. If the correlation coefficient had to be calculated it would be 0.234, so closer to +0.5 than -1.

Question 2/10

Considering a relationship between two variables with an R2 equal to 0.01, which of the following statements are true?

a. Almost all of the variability is attributed to factors not related to the variables under consideration.
b. The independent variable is not a strong predictor of the dependent variable.
c. The independent variable is a strong predictor of the dependent variable.
d. The line of best fit describes the relationship between the two variables well.

- [ ] Only a and d
- [ ] Only c and d
- [ ] Only a and c
- [x] Only a and b

Explanation:

Option a is true since only 1% of the variability in the dependent variable can be explained by the independent variable, while the other 99% of variability is attributed to other factors. Option b is also true since a low R2 suggests that the independent variable may not be a strong predictor of the dependent variable; as a result, option c, which is the opposite, is incorrect. Option d is incorrect because the explanations on options a and b indicate that there is too much variability to be able to say that the line described the relationship well.

Question 3/10

Which of the following statements are true considering the following line of best fit equation?

$y = 1.28x - 30.2$

a. The independent variable increases as the dependent variable decreases.
b. The independent variable increases as the dependent variable increases.
c. The line of best fit describes the relationship between the two variables well.
d. The line of best fit does not describe the relationship between the two variables well.

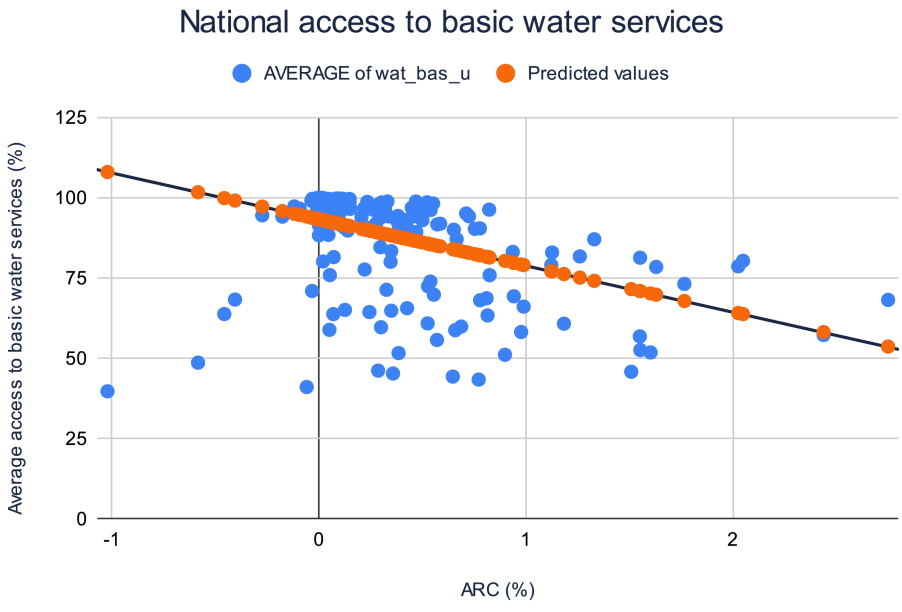
- [ ] Only a and d
- [ ] Only b and c
- [x] Only b
- [ ] Only a

Explanation:

Option a is false since the statement indicates a negative relationship and therefore the slope value m would've been negative. Option b is true since the slope value (m=1.28) is positive which represents a positive relationship between x and y. Both options c and d are false because we cannot only use the equation to assess the goodness of fit.

Question 4/10

If we predict the national average access to basic water services based on the Annual Rates of Change (ARC), as in the visualisation, and we've calculated the accuracy measures as:



MAE = 9.67
MSE = 187.03
RMSE = 13.68

Which of the following statements is true?

- [ ] The MAE and RMSE are relatively low, indicating that the predicted values are generally accurate and the greater variation below the line of best fit is simply overemphasised by the MSE.

- [x] The measures suggest that the predicted values are generally inaccurate, with an average error of approximately 9.67 percentage points from the actual values, as indicated by the MAE.

- [ ] The measures do not provide enough information to assess the accuracy of the predicted values, as they are not expressed as percentages and do not account for the specific units of the target variable.

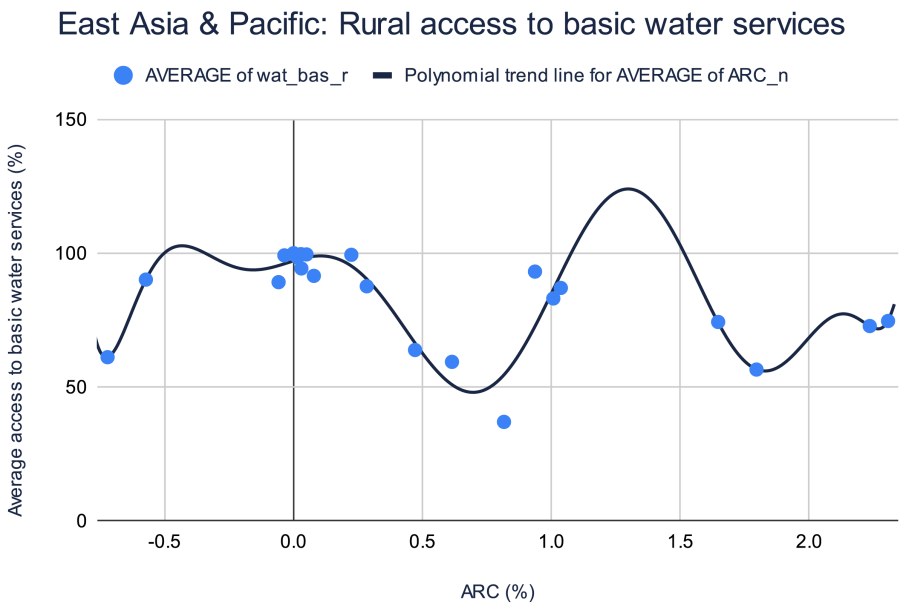
- [ ] None of the above.

Explanation:

The MAE, MSE, and RMSE are relatively high which suggests that the predicted values deviate from the actual values and the model's predictions therefore have a moderate level of uncertainty and may not be accurate.

Question 5/10

Considering the visualisation, which of the following statements is true for the line of best fit?



- [x] It is over-fitted but a linear line will also not be able to describe the relationship because the data are non-linear.
- [ ] It is under-fitted but a linear line will also not be able to describe the relationship because the data are non-linear.
- [ ] It is over-fitted but a linear line will be able to describe the relationship better.
- [ ] It is under-fitted because the line doesn't pass through all of the data points.

Explanation:

The line captures most of the noise or variability in the data, which indicates over-fitting. The data have a high variance, so if we were to add another data point, the line would change completely. We also see that there is no obvious linear relationship between the variables, and it is unlikely that a linear line of best fit will model the data well.

Question 6/10

Considering only the following table of correlation coefficients for the average access to basic water services and the Annual Rates of Change (ARC) in rural and urban areas, which of the following statements is true?

	Rural	Urban
All regions	-0.2115	-0.1425
East Asia & Pacific	-0.4351	-0.4342
Europe & Central Asia	-0.7532	-0.3965
Latin America & Caribbean	-0.0661	-0.1316
North America	-0.7168	0.6217
South Asia	-0.9306	-0.2133
Sub-Saharan Africa	0.2376	0.1450

- [ ] The correlation between the average rural access to basic water services and the Annual Rates of Change in Sub-Saharan Africa is greater than the correlation observed in any other region, both rural and urban.

- [x] The correlation between the average rural access to basic water services and the Annual Rates of Change in South Asia is less than the correlation observed in any other region, both rural and urban.

- [ ] The correlation between the average urban access to basic water services and the Annual Rates of Change in North America is greater than the correlation observed in any other region, both rural and urban.

- [ ] The correlation between the average rural access to basic water services and the Annual Rates of Change in South Asia is greater than the correlation observed in any other region, both rural and urban.

Explanation:

Since the strength of correlation is determined by the absolute value of the correlation coefficient, South Asia has the highest correlation coefficient for rural at -0.9306. The positive or negative sign of the correlation coefficient doesn't influence the

strength of the relationship, only the direction of it.

Question 7/10

Which of the following statements best represents a null hypothesis?

- [ ] None of the above.
- [x] The average access to basic water services on a national level for a specific country is equal to the average access in rural and urban areas for the same specific country.
- [ ] The average access to basic water services in rural areas is significantly lower than the average access in urban areas.
- [ ] Higher-income regions have on average higher access to basic water services compared to lower-income areas.

Explanation:

A null hypothesis is a statement of no effect, no difference, and no relationship. As a result, statements that include “lower” and “higher” are not null hypotheses but rather alternative or simply research hypotheses. The statement that includes “equal” is a null hypothesis because it establishes a baseline “default” assumption that can be tested using hypothesis testing.

Question 8/10

Consider the following null and alternative hypotheses:

Null hypothesis: The average access to basic water services of people in rural and urban areas is equal.

Alternative hypothesis: The average access to basic water services of people in rural areas is lower than that of people in urban areas.

If the calculated test statistic is equal to -1.7756 and the critical value is 1.782 at a 5% level of significance, which of the following statements is true?

- [ ] We cannot reject or fail to reject the null hypothesis based only on the information provided.
- [x] We fail to reject the null hypothesis, suggesting that we don't have enough evidence to say that the difference in the average access to basic water services between rural and urban areas is statistically significant.
- [ ] We fail to reject the null hypothesis, suggesting that the average access to basic water services of people in rural and urban areas is equal.
- [ ] We reject the null hypothesis, suggesting that the average access to basic water services of people in rural areas is lower than that of people in urban areas.

Explanation:

Since the absolute value of the test statistic (1.7756) is smaller than the critical value (1.782) we fail to reject the null hypothesis. However, failing to reject the null doesn't mean that we accept the null, but rather that we do not have enough evidence to statistically accept it as the truth.

Question 9/10

Which of the following hypothesis tests would be most appropriate to use, considering the following null hypothesis and sample details?

Null hypothesis: The average access to basic water services of people in rural and urban areas is equal.

Sample size: 7
Unknown population standard deviation
Both samples are normally distributed

- [ ] Any parametric test
- [ ] Any non-parametric test

- [x] A two-sample t-test
- [ ] A two-sample z-test

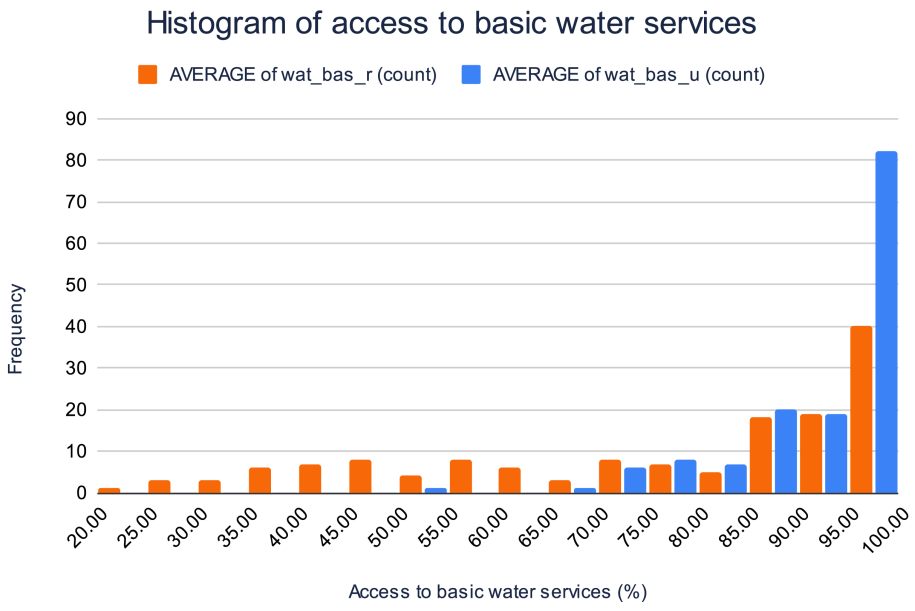
Explanation:

Since we know our data are normally distributed, we know we can use a parametric test. Since the sample size is small and we don't have the population standard deviation, we already know that we can use a type of t-test.

Question 10/10

Which of the following hypothesis tests would be most appropriate to use, considering the following null hypothesis and histogram of the two samples?

Null hypothesis: The distributions of access to basic water services in rural and urban areas are the same.



- [ ] The paired two-sample z-test
- [ ] None of the above
- [ ] The independent two-sample z-test
- [x] The Kolmogorov-Smirnov test

Explanation:

There are two indications in the question that the Kolmogorov-Smirnov test is appropriate: 1) the null hypothesis refers to *distributions*, and 2) the distribution in the visualisation indicates non-normality, which means we cannot use a parametric test such as a z-test.