**Hands-on-Training in Data Analysis by Kreativstorm. Homework-W3**

Q1. In your own words, describe what a residual is in linear regression.

Residual in linear regression is the difference between the observed values of the dependent variable (the actual data points) and the values predicted by the linear regression model. In other words, it quantifies the discrepancy between what the model predicts and the actual outcomes.

Q2. If you know that your residual data follow the below pattern, are your data better approximated with a linear model for the lower values of independent variable or higher values of independent variable and why?Chart, scatter chart

Description automatically generated

Yes. The residual data exhibits varying levels of variance along the range of the independent variable. Specifically, the spread of the residual may widen or narrow for different segments of the independent variable, which suggests that the linear model's assumptions are violated. In this case, the linear model may perform better for either the lower or higher values of the independent variable, depending on the specific pattern of heteroscedasticity.

Q3. What is the difference between *R2* and adjusted *R2*?

*R2* value represents the proportion of variance in the dependent variable that can be explained by the independent variable, while Adjusted *R2* provides a value that would be expected in the population. The difference is that *R2* measures the proportion of variance explained by the model without considering model complexity, adjusted *R2* adjusts for the number of predictors and provides a more realistic assessment of model fit, helping to prevent overfitting by penalizing the inclusion of unnecessary variables. Researchers often prefer adjusted R-squared when comparing models with different numbers of predictors or when trying to strike a balance between model simplicity and explanatory power.

Q4. Is there independence of observations if you are trying to predict baby length with mother’s height? **Yes**

* Yes
* No

Q5. Justify the above answer.

The independence of observations can be confirmed using the Durbin-Watson statistics (the value should be between 1.5-2.5). In this case, the independence of observations is 1.724.

Q6. Do residual data show homoscedasticity? **Yes**

* Yes
* No

Q7. Justify the above answer.

In a homoscedastic dataset, you would see a scatterplot with points dispersed evenly around a horizontal line, without any apparent funnel-like or cone-shaped pattern.

Q8. What is the value of *R2* and what does this tell you?

The value of *R2* is 0.24, and it’s relatively low. This suggests that there is 24% variation in mother’s height by baby length. Further indication is that the independent variable can account for a portion of the observed variation in the dependent variable. This means that the variables used for the prediction are not important since it cannot explain much or account for a significant portion of the observed variation in the dependent variables.

Q9. Can you consider the relationship between mother’s height and baby length a statistically significant linear relationship and why?

Yes, the relationship shows a linear regression on the graph plot, and secondly, the significance value in the ANOVA table is 0.001 which is less than 5%, indicate that the relationship is statistically significant.

P= 0.001

P<0.05

0.001< 0.05

Q10. Having the ANOVA table for the linear regression in mind, what is the null and alternative hypothesis in this case?

Define the Hypothesis:

Null hypothesis: There is no statistically significant linear relationship between mother's height (independent variable) and baby length (dependent variable).

Alternative hypothesis: There is a statistically significant linear relationship between mother's height (independent variable) and baby length (dependent variable).

Since the p-value in ANOVA is less than the significance level of 5%: 0.001< 0.05.

We therefore reject the null hypothesis and conclude that, “there is a statistically significant linear relationship between mother's height and baby length”.

Q11. In your own words, describe what the b1 is.

b1 represents the slope of the regression line or the rate of change in the dependent variable for each unit change in the independent variable. It indicates the strength and direction of the linear relationship between the two variables. If b1 is positive, it suggests a positive linear relationship (an increase in the independent variable leads to an increase in the dependent variable), and if b1 is negative, it suggests a negative linear relationship (an increase in the independent variable leads to a decrease in the dependent variable).

Q12. What does the value of b1 tell you in practical terms?

The value of b1 helps a researcher or data analyst to understand how changes in the independent variable are associated with changes in the dependent variable, which can be valuable for making predictions, drawing conclusions, and understanding the impact of various factors in the dataset or analysis.

Q13. Could you claim the same for the mother’s height in the range between 140cm and 145cm and why? Yes, because b1 is positive and statistically significant. This suggests that, on average, for each one-centimetre increase in the mother's height within the range of 140cm to 145 cm, it would increase the baby's length by 0.2cm.

Q14. According to this model, what is the prediction of baby length for mother’s height of 170cm? The prediction of baby length is 52.47 to 52.91 centimeters.

Q15. Report on your findings for predicting baby length with mother’s height.

**Introduction:**

The purpose of the analysis is to predict baby length with mother’s height.

**Data and Methodology:**

The dataset used is the baby length (dependent variable) and the mother’s height (independent variable). A linear regression model was used to model the relationship between mother's height and baby length.

**Regression Model:**

The linear regression model used in this analysis is represented as follows:

* + - Baby Length = β₀ + β₁ \* Mother's Height
    - where β₀ is the intercept, β₁ is the coefficient for Mother's Height

**Findings:**

The coefficient estimate for Mother's Height (β₁) was found to be 0.219.

Where:

P-value in ANOVA =0.001

The coefficient estimate for Mother's Height was statistically significant (p < 0.05), indicating that Mother's Height has a significant effect on predicting Baby Length.

Prediction for Mother's Height of 170cm:

BL =15.33 +(0.219 \*170) = 52.56cm

Based on the regression model, for a mother's height of 170cm, the predicted baby length is 52.56cm.

**Confidence Interval for Prediction:**

52.56-0.093(Lower bound) = 52.47cm

52.56 +0.345(Upper bound) = 52.91cm

The 95% confidence interval for the predicted baby length for a mother's height of 170cm

is 52.47 to 52.91 centimeters.

Conclusion:

The prediction of baby length for mother’s height of 170cm is 52.47 to 52.91 centimeters.

Q16. Can you predict baby length with father’s age? Why? **Yes**

Define the Hypothesis:

Null hypothesis: There is no statistically significant linear relationship between father's age (independent variable) and baby length (dependent variable)

Alternative hypothesis: There is a statistically significant linear relationship between father's age (independent variable) and baby length (dependent variable).

since the p-value in ANOVA in the distribution is 0.386 greater than the significant level of 5%, we fail to reject the null hypothesis, which suggest that there is no statistically significant linear relationship between father's age (independent variable) and baby length (dependent variable) and hence we cannot predict baby length with father’s age.

Q17. What does homogeneity of variance mean and why its important assumption of an independent t-test?

Homogeneity of variance, also known as homoscedasticity, is an important assumption when conducting an independent t-test. It refers to the condition in which the variances (spread or dispersion) of the two groups being compared in the t-test are approximately equal. In other words, it means that the variability in the dependent variable is roughly the same for both groups. It is important because:

1. Validity of Statistical Tests: Many statistical tests, including t-tests, analysis of variance (ANOVA), and linear regression, assume that the variances of the groups being compared or analysed are roughly equal. Violating this assumption can lead to incorrect results and conclusions.
2. Type I Error Control: When variances are not homogeneous across groups, it can affect the performance of statistical tests and lead to an increased risk of Type I errors. Type I errors occur when a test incorrectly detects a significant difference or relationship when none exists. Ensuring homogeneity of variance helps control the Type I error rate.
3. Interpretability: Homogeneity of variance simplifies the interpretation of statistical results. For example, in a t-test, when variances are roughly equal, the t-value has a straightforward interpretation in terms of the standard error of the difference.

between means. When variances are not equal, interpreting the results becomes more complex.

1. Confidence Intervals: Homoscedasticity is important when constructing confidence intervals around group means or regression coefficients. Violating this assumption can lead to biased or inaccurate confidence intervals.
2. Model Assumptions: In linear regression and many other modeling techniques, homoscedasticity is an important assumption. Violations of this assumption can affect the validity of regression coefficients and predictions.

Q18. Is there homogeneity of variance between head circumference for babies of smoking mothers and head circumference for babies of non-smoking mothers? **Yes**.

* Yes
* No

Q19. Justify your choice.

Define the Hypothesis:

* Null Hypothesis (H0): There is homogeneity of variance between head circumference for babies of smoking mothers and head circumference for babies of non-smoking mothers.
* Alternative Hypothesis (Ha): There is no homogeneity of variance between head circumference for babies of smoking mothers and head circumference for babies of non-smoking mothers.

The p-value from Levene's test is 0.368, which is greater than the significance level of 0.05.

0.368 > 0.05, therefore, we fail to reject the null hypothesis, and conclude that there is homogeneity of variance between head circumference for babies of smoking mothers and head circumference for babies of non-smoking mothers.

Q20. Do smokers have lighter babies? Justify your answer. **No**

Define the Hypothesis:

* Null Hypothesis (H0): Smoking during pregnancy has no significant effect on birth weight.
* Alternative Hypothesis (Ha): Smoking during pregnancy has significant effect on birth weight.

The p-value from Levene's test is 0.584, which is greater than the significance level of 0.05.

0.584 > 0.05, therefore, we fail to reject the null hypothesis, and conclude that smoking during pregnancy has no significant effect on birth weight.

Q21. Do women over 35 have lighter babies? Justify your answer. No

Define the Hypothesis:

Null Hypothesis (H0): Women over 35 has no significant effect on birth weight.

Alternative Hypothesis (Ha): Women over 35 has significant effect on birth weight.

The p-value from Levene's test 0.166, is greater than the significance level of 0.05.

0.166 > 0.05, therefore, we fail to reject the null hypothesis, and conclude that women over 35 has no significant effect on birth weight.

Q22. Using the cholesterol dataset, was the diet effective in lowering cholesterol concentration after 8 weeks of use? Justify your answer. Yes

The p-value of the paired samples test table showed a significant value at 0.0001.

p< 0.05

In addition, the paired sample statistics table showed that the mean value Before the diet is 6.4078 and After 8weeks is 5.7789, a Mean difference of 0.62.

Using the T-TEST formular for the dependence sample to get the impact of the difference:

d= M/SRT 2 \* SD

M= mean

SD= standard deviation

D = difference

d =0.62/ 1.41 \*0.18

d= 2.5

The difference of 2.5 units on the t-test result indicates that, on average, cholesterol concentration decreased by 2.5 mg/dL after 8 weeks of using the diet. This difference is considered statistically significant.

Q23. For the above case, what is the null and alternative hypothesis?

Define the Hypothesis:

* Null Hypothesis (H0): The diet has no significant effect in lowering cholesterol concentration after 8 weeks of use.
* Alternative Hypothesis (Ha): The diet has significant effect in lowering cholesterol concentration after 8 weeks of use.

From the above analysis, since p < 0.05, we reject the null hypothesis, and conclude that diet has significant effect in lowering cholesterol concentration after 8 weeks of use.

Q24. Was the diet more effective in the first 4 weeks of use or the last 4 weeks of use? Justify your answer.

Define the Hypothesis:

Null Hypothesis (H0): There is no significant difference in the effectiveness of the diet between the first 4 weeks and the last 4 weeks of use.

Alternative Hypothesis (Ha): The diet is more effective in the first 4 weeks of use compared to the last 4 weeks of use.

The p-value of the paired samples test table showed a significant value at 0.0001.

p< 0.05

In addition, the paired sample statistics table showed that the mean value Before the diet is 6.4078 and After 8weeks is 5.8417, a Mean difference of 0.57.

Using the T-TEST formular for the dependence sample to get the impact of the difference:

d= M/SRT 2 \* SD

M= mean

SD= standard deviation

D = difference

d =0.57/ 1.41 \*0.156

d= 2.6

The difference of 2.6units on the t-test result indicates that, on average, cholesterol concentration decreased by 2.6 mg/dL in the first 4 weeks of use compared to the last 4 weeks of use. This difference is considered statistically significant.

In summary, since p<0.05, we reject the null hypothesis and conclude that the diet is more effective in the first 4 weeks of use compared to the last 4 weeks of use.

Q25. If you know that the average cholesterol concentration in healthy adults is 3 mmol/L, would you consider your sample (N=18) significantly better or worse than the average adult population? Justify your answer.

Define the Hypothesis:

Null Hypothesis (H0): The average cholesterol concentration in the sample is not significantly different from the average cholesterol concentration in healthy adults.

Alternative Hypothesis (Ha): The average cholesterol concentration in the sample is significantly different from the average cholesterol concentration in healthy adults.

After confirming the normality test, the p-value of the Sample Test table indicate a significance value of 0.0001, which shows that p< 0.05, we reject the null hypothesis. This indicates that the sample's average cholesterol concentration is significantly different from the average cholesterol concentration in healthy adults.