HOMEWORK WEEK3

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

A residual is the difference between the actual value of a dependent variable and the predicted value of that dependent variable.

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

The data is better approximated by a linear model for lower values than for higher values of independent variable.

This is because the values for the residuals are smaller for lower values of independent variables while the residual values are larger

For higher values of independent variable

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

The R2 called the coefficient of determination represents the percentage of variance in the dependent variable represented by the independent variable for a sample, while the adjusted R2 represent the coefficient of determination for the population

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

* Yes
* No

Yes

Q5. Justify the above answer.

 Durbin-Watson test which gives a value within the required range (1.7)

Q6. Do residual data show homoscedasticity?

* Yes
* No

Yes

Q7. Justify the above answer.

The plot of the residuals vs fitted model showed rough randomness in the spread of the residuals. There is no clear pattern

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

0.2352. It indicates that 23.5% of the variation in length of the baby can be explained by the height of the mother.

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

Yes. The relationship is statistically significant as the p-value for the F-statistic is 0.0011 which is less than 0.05 level of significance.

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

The null hypothesis states that there is no linear relationship between the mother’s height and baby length.

The alternative hypothesis states that there is a linear relationship between the mother’s height and baby length.

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

The b1 is the coefficient of the indepenedent(mother’s height) variable

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

It tells me that for every one unit increase in mother’s height, the value of baby’s length increase by b1

Q13. Could you claim the same for the mother’s height in the range between 140cm and 145cm and why?

No. this range falls outside of the range of mother’s height that we used for our model.

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

52.56cm

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

A linear regression established that mother’s height could statistically significantly predict baby’s length.

F(1,40)=12.3, p-value = 0.00113

Baby’s length = -0.944+(0.037\*mother’s height)

R2 =0.2161

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

No. Because the father’s age has no relationship with the length of the baby

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

Homogeneity of variance means that the variance of the two population that we are comparing are roughly equal. This essentially means that the two populations are similar. It allow us to test whether the difference in the two populations is only due to variable that we are concerned about.

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

* Yes
* No

Yes

Q19. Justify your choice.

p-value is 0.3175 which is greater than 0.05. this tells us that the variance is roughly equal(null hypothesis).

Q20. Do smokers have lighter babies? Justify your answer.

Yes. The mean for birthweight of smokers (3.187) is less than the mean of birthweight for non smokers. Furthermore p-value for the T-test is 0.045. Indicating that there is an evidence that there is difference in the mean of the two population with sample mean for non- smoker (3.50) higher than the mean for smokers( 3.134)

Q21. Do women over 35 have lighter babies? Justify your answer.

NO. no significant difference was found between the two groups with p-value = 0.49

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

Yes. The mean value for the amount of cholesterol before the diet was introduced(6.41) is higher than the mean value for the amount of cholesterol after 8 weeks(5.73) the diet was introduced. Furthermore, the dependent sample t-test that was carried out also confirmed that there is a very strong statistical evidence against the null hypothesis that there is no difference in cholesterol levels in the two variables. P-value is 3.279e-11

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

The null hypothesis is that the diet is not effective 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.

The diet is more effective after the first 4 weeks of use than the last 4 weeks. The reduction in mean cholesterol level after 4 weeks (6.41-5.842 = 0.568)(which is significant with p-value less than 0.05) is much more than the reduction in the mean cholesterol levels in the last 4 weeks of use(5.842-5.73= 0.112)

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 average adult population? Justify your answer.

The sample is significantly worse than the average adult. Reason being that the mean of the cholesterol level for the average adult(3) is less than for the sample (6.41) and the p-value for the one-sample t-test confirmed that there is a statistically significant difference between the two means with a p-value of 8.441e-10