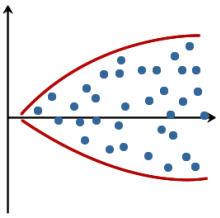
Q1. describe what a residual is in linear regression.

Residual is the vertical distance between the data point and the fit line ,where the line should tell us the predicted value so the residual should be the error of that prediction

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?

Lower values ,as we can see the horizontal line as the regression line the point are the residuals where the are more close to the line on the low values then spread away from the line as the values get higher



Q3. What is the difference between R^2 and adjusted R^2?

R^2 represents the effect of the independent variable on the dependent variable on the sample data, while adjusted R^2 represents the same definition but on the population

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

Yes

Q5. Justify the above answer.

Measuring the relation between the two variables we can see the Durban Watson value 2.145 which indicates independence of observations

Q6. Do residual data show homoscedasticity?

Yes

Q7. Justify the above answer.

Looking to the scatter plot of the two variables we see data points are randomly scattered across the regression line and follow no specific pattern which indicates homoscedasticity.

Q8. What is the value of R^2 and what does this tell you?

0.235 it indicates on scale from 0 to 1 how much the independent variable affects the dependent variable, in other words mother's height affect the length of the baby by 23.5%

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

Yes, because the p-value of the ANOVA test is 0.001 which is less than 0.05, so we can say that the relation is statistically significant linear relation

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

H_0: the null hypothesis states that the relationship between mother's height and baby length is statistically significant linear relationship
H_a: the alternative hypothesis states that the relationship between mother's height and baby length is not statistically significant linear relationship

Q11. In your own words, describe what the b_1 is.

It describes the slope of the regression line mathematically, but from the perspective of the study it represent the predicted value of change to the dependent variable based on 1 measure unit of change for the independent variable

Q12. What does the value of b_1 tell you in practical terms?

The change of 1 cm in the mother's height will result in b1 change of the baby's length

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

No, these numbers are outside the range of the sample which the study is carried on, so we can't generalize the rule.

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

52.564

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

A linear regression established that the mothers' height could statistically significantly predict their babies length, $F(1, 40) = 12.302$, $p < .001$ and the mother's height accounted for 23.5% of the explained variability in baby's length. The regression equation was: predicted baby's length (cm.) = 15.334 + 0.219 x (mother's height (cm.)).

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

No, the linear regression relationship shows no significance as p value is 0.386, $R^2 = 0.019$ which is 1.9% which is negligible with Durbin Watson test value = 1.15

Answer: No because there negligible correlation between them and the scatter plot shows no linearity and the significant value given by ANOVA table is 0.386 showing that the relationship is statistically non significant.

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

Homogeneity of variance is the condition where the data points of specific sample are normally distributed on their fit line, meaning that they are spread randomly without concentration or deviation regions and they remain consistent across the fit line.
Since T-test and ANOVA assume that the data of observation is normally distributed or little skewed, we have to ensure the HOV, otherwise the results will be biased.

Answer: The independent samples t-test and ANOVA both assume homogeneity of variance, which states that all comparison groups have the same variance. The t and F statistics are used in the independent samples t-test and ANOVA, which are both generally robust to assumption violations as long as group sizes are equal.

It's important because statistical tests like ANOVA and the Student's t-test require it. If the data sets have equal sample sizes, the unequal variance has little effect on ANOVA.

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

- Yes

Q19. Justify your choice.

Both showed normally distribution through the normality test with significance value of 0.372 and 0.085 respectively, plus the levene's test value is 0.368 above 0.05

Answer: The answer is yes because the significance is 0.368 more than 0.05.

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

Yes, the t-test results show significant difference with value 0.043 and effect size of 0.58 which is moderate, on top of the difference between the mean of both groups which is 0.375 K.grams.

Answer: Yes the smokers have lighter babies because comparing the mean for the two groups for smoker we get 3.11 and for non-smoker we obtain 3.50 and this is also explained by the value of the effect and the significance.

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

Yes, based on the mean difference.
although the t-test results show no significant difference with value 0.492 and effect size of 0.608 which is moderate, on top of the difference between the mean of both groups which is 0.221 grams.

Answer: Yes mothers other 35 have lighter babies because the mean gives 3,11 but 3,33 for mother under 35. The value of the effect and the significance also attest the fact.

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

Yes, the t-test results show significant difference with value <0.001 and effect size of 0.17852 which is medium, on top of the difference between the mean of both groups which is 0.629

Answer: Yes, it does because of the effect value that is 0.1785

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

H_0: the null hypothesis states that the margarine brand isn't effective in lowering cholesterol concentration after 8 weeks of use
H_a: the alternative hypothesis states that the margarine brand is effective in lowering cholesterol concentration after 8 weeks of use

Q24. Was the margarine diet more effective after 4 weeks of use or after 8 weeks of use? Justify your answer.

Effective after 4 weeks of usage, if we measured the dif. between (after 4 weeks, after 8 weeks) we can find a significant dif. Of value 0.001 and mean dif. Equals 0.06278 with effect size of 0.070
Meanwhile the dif. between (before, after 4 weeks) we can find a significant dif. Of value <0.001 and mean dif. Equals 0.56611 with effect size of 0.156

With the tests being done we can see that the effect of the diet is more effective after 4 weeks ,although it is still decreasing the level of cholesterol after 8 weeks but its not that much as after 4 weeks

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.

significantly worse, because their average cholesterol concentration after 8 weeks is 5.78 mmol/L

Answer: The significant for the one sample t test is below 0.001 and the mean for the sample is 6.4078 at least two times 3 mmol/L so the people in the sample have already high level of cholesterol and therefore this sample is not significantly better than average adult population.