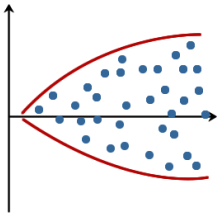
Q1. In your own words, describe what a residual is in linear regression.
Ans: Residuals are the errors of the regression line. In other words, it is the difference between the predicted value (by regression model) and the actual observed value.

residual = $y_{actual} - y_{predicted}$

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?



Ans: Lower values of independent variable.

Here y-axis => residuals, x-axis => independent variable.

The residuals, which are the differences between actual and predicted values, are smaller / closer to 0 in the lefthand side of the x-axis indicating lower values of the independent variable. Smaller residuals indicate accurate predictions by the model.

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

Ans: Both values are used represent the goodness of fit for a regression model.

R^2 value represents the proportion of variance in the dependent variable that can be explained by the independent variable. This is based on the **'sample'.** If we use a single variable to predict the dependent variable, **R^2** gives a good measure of the fit.

Adjusted R^2 gives the value that would be expected in the '**population'.** Also, when we use multiple independent variables to predict the dependent variable, this value adjusts and accounts for the complexity of the model, that is the presence of multiple predictors.

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

- Yes
- No

Ans: Yes

Q5. Justify the above answer.

Ans:

```
> durbinWatsonTest(model)
 lag Autocorrelation D-W Statistic p-value
   1     0.08619051      1.724487   0.376
 Alternative hypothesis: rho != 0
```

The test value is 1.72 which is within the range (1.5-2.5) for independence of observations. And the p-value is 0.376.

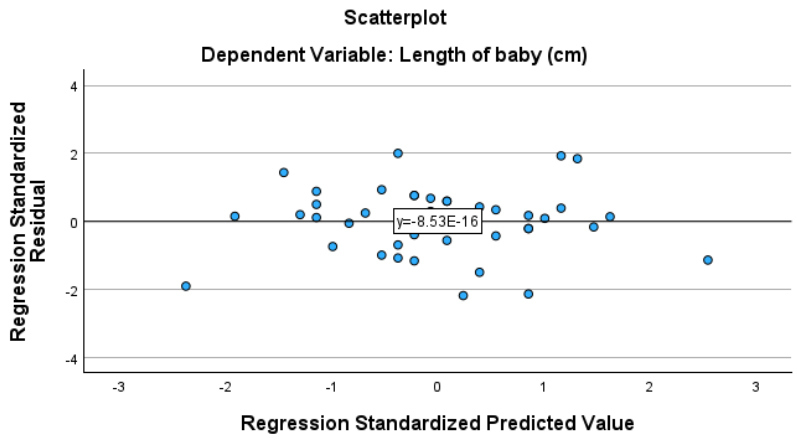
From SPSS – this test value is 2.316.

Q6. Do residual data show homoscedasticity?

- Yes
- No

Ans: Yes

Q7. Justify the above answer.



Ans: Residual data is homoscedastic. As visible from the residual plot and also from the results of the Breusch-Pagan test.

studentized Breusch-Pagan test
data: model
BP = 0.026314, df = 1, p-value = 0.8711

p-value of the Breusch-Pagan test > 0.05.
Hence we reject the null hypothesis that data is heteroscedastic. There is no evidence of heteroscedasticity, meaning the residuals are homoscedastic.

Q8. What is the value of R^2 and what does this tell you?
Ans: $R^2 = 0.235$.
This means that 23.5 % of the variation in Length of babies can be explained by the variable mother's height.

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

From the ANOVA table, the p-value for this model is 0.001
P value needs to be <.05 for the model to be statistically significant and for the relationship to be considered as statistically significant linear relationship.

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

Null Hypothesis: The model using 'mother's height' does not explain a statistically significant portion of the variability in the 'Length' of the baby
Alternate Hypothesis: The model using 'mother's height' explains more variability in the 'Length' of the baby than using the mean of the 'Length' column.

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

Ans: b_1 is the slope coefficient. (Slope of the regression line). It represents the how much the predicted value increases or decreases, for 1 unit of change in the predictor value.
In this example, b1 = 0.219. For every 1 cm increase in mother's height, length of baby increases by 0.219.

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

Ans: b_1 gives the rate of change of the dependent variable for each unit of change in the independent variable. There are two aspects – direction and strength.
Direction: If b_1 is **positive**, then the dependent variable **increases** with increase in independent variable. If b_1 is **negative**, then the dependent variable **decreases** with increase in independent variable.
Strength: The magnitude of increase/decrease is given by the value of b_1. The absolute value of $\mathbf{b_1}$ tells you the **strength** of the relationship between the independent and dependent variables. A larger value of $\mathbf{b_1}$ means that the dependent variable is more sensitive to changes in the independent variable.
In this example, b1 = 0.219. For every 1 cm increase in mother's height, length of baby increases by 0.219 for mother's height in the range 149-181 cm.

Q13. Could you claim the same for the mother's height in the range between 140cm and 145cm and why?
Ans: No.
Prediction Equation => Length = 15.334 + 0.219 * mheight => is valid for 149 <= mother's height <=181
The given values are out of range.

Q14. According to this model, what is the prediction of baby length for mother's height of 170cm?
Ans: Prediction Equation =>

Length = 15.334 + 0.219 * mheight
Length = 15.334 + 0.219 (170)
= 52.56 cms with 95%CI [52.47, 52.91]

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

Ans:

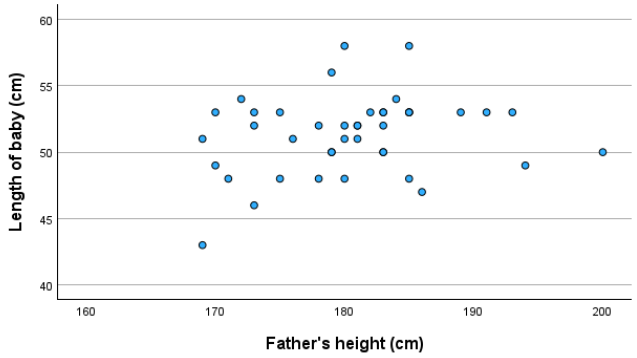
A linear regression established that mother's height could statistically significantly predict baby's length. F(1, 40) = 12.302, p = .001 and mother's height accounted for 23.5% (R2 = 0.235) of the explained variability in baby length.

The regression equation is as follows:

predicted baby length = 15.334 + 0.219(mother's height)

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

Ans: No



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15.339	1	15.339	1.815	.185[b]
	Residual	337.994	40	8.450		
	Total	353.333	41			

a. Dependent Variable: Length of baby (cm)

b. Predictors: (Constant), Father's height (cm)

Correlations

		Length of baby (cm)	Father's age
Length of baby (cm)	Pearson Correlation	1	.137
	Sig. (2-tailed)		.386
	N	42	42
Father's age	Pearson Correlation	.137	1
	Sig. (2-tailed)	.386	
	N	42	42

Correlations

		Length of baby (cm)	Father's age
Length of baby (cm)	Pearson Correlation	1	.137
	Sig. (2-tailed)		.386
	N	42	42
Father's age	Pearson Correlation	.137	1
	Sig. (2-tailed)	.386	
	N	42	42

1. No linear relationship between father's age and baby's length (from scatter plot).
2. No statistically significant correlation between the variables. (both correlation, partial correlation scores)
3. ANOVA table gives p-value = 0.185. p-value should be less than 0.05 for the linear relationship to be statistically significant.

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

Ans: Homogeneity of variance is an important assumption in independent t-test – when comparing the means of two groups. It refers to the assumption that the variance (or spread) of the scores in each group being compared is roughly equal. In simple terms it means that the variance of the two groups is very similar and any difference in the groups that we find, comes not from the variance, but from the difference in values.

Homogeneity of variance is crucial for the independent-measures t-test because it ensures that the populations being compared have similar variability, allowing for a fair and accurate comparison of their means. If variances are significantly different between groups, the t-test may produce misleading results, potentially leading to incorrect conclusions about the true difference between the groups.

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

Ans: Yes

Q19. Justify your choice.
Ans: p-value of Levene's test = 0.368. If p-value is > 0.05 then there is homogeneity of variance.

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

Ans: Yes. We use an independent means t-test to justify the answer by first checking the assumptions.
1. Birthweight for babies of smoker / non-smoker mother are normally distributed, continuous data and are from different groups of people accordingly.
2. Lavene's test statistic has p-value > 0.05 (so there is homogeneity of variance)
3. p-value for independent means t-test is 0.043 which is < 0.05. We have t = 2.093. Which means the groups are different.
4. Mean baby birthweight (smoker) = 3.1341, SD = 0.631
   Mean baby birthweight (non-smoker) = 3.5095, SD = 0.518
   Babies of smoker mothers are lighter.

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

Ans: No

There is no sufficient evidence to support this statistically.

1. Create a new variable that indicates whether mother's age is above 35 or not.
2. The two groups of data have normal distribution of birthweight
3. p-value of Lavene's test = 0.791 which indicates homogeneity of variance.
4. However, the t-test for independent means gives us a p-value of 0.492 which is > 0.05. This means that there is no enough evidence to conclude that the means of the two groups are different. ie, the difference between the groups are not statistically significant.

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

Ans: Yes.

The difference in pre/post values follow normal distribution. Hence we apply dependent means t-test to find if our answer is statistically significant.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Before	6.4078	18	1.19109	.28074
	After 8 weeks	5.7789	18	1.10191	.25972

Paired Samples Test

		Paired Differences					t	df	Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			One-Sided p	Two-Sided p
Pair 1	Before - After 8 weeks	.62889	.17852	.04208	.54011	.71766	14.946	17	<.001	<.001

The p-value for the paired sample t-test <.001. We reject the null hypothesis that the means of the groups are not different.

That is, the test results show that the means of the two groups are different.

(After 8 weeks) 5.7 < 6.4 (Before)

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

Ans: Null Hypothesis: The average cholesterol of the people before the diet and after 8 weeks of the diet are the same.

Alternate Hypothesis: The average cholesterol of the people before the diet and after 8 weeks of the diet are different.

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

Ans: Diet was more effective in the first 4 weeks.

The values to be compared are (Before-After4weeks) and (After4weeks-After8weeks)

Mean of B_4wks = 0.5661 and SD = 0.156

Mean of 4_8wks = 0.0628 and SD = 0.07

We use a dependent means t-test to compare this. We get a t(17) = 13.128 with p-value < 0.001 and a mean difference = 0.50 with 95% CI [0.42, 0.58]. We reject the null hypothesis. This means that there is a significant difference between the effectiveness of diet in the first 4 weeks and last 4 weeks. From the values. diet was effective in the first 4 weeks

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	diff_B_4	.5661	18	.15557	.03667
	diff_4_8	.0628	18	.07044	.01660

Paired Samples Test

		Paired Differences					t	df	Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			One-Sided p	Two-Sided p
Pair 1	diff_B_4 - diff_4_8	.50333	.16266	.03834	.42244	.58422	13.128	17	<.001	<.001

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.

Ans:

Mean cholesterol for the sample = 6.4. This means that the cholesterol of the sample is worse than average adult population. To support statistically, let's do the one sample t-test. The p-value < .001. This means that we reject the null hypothesis that the sample mean equals to the population mean. This means that the average cholesterol of the given sample is significantly different (worse) than the population average.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Before	18	6.4078	1.19109	.28074

One-Sample Test

Test Value = 3

	t	df	Significance		Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p		Lower	Upper
Before	12.138	17	<.001	<.001	3.40778	2.8155	4.0001

.