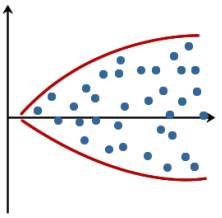
Q1. In your own words, describe what a residual is in linear regression.

A residual is the difference between 2 values (observed value – predicted value), this difference is the residual for one data point. The predicted value is going to be on the line.
Residual data means all information about how accurate or wrong we are given a model, in this case the linear regression.
It's also important to mention that with the linear regression we are trying to regress/approximate the relationship between variables. Hence we find the regression line that best fits the data, the regression line gives us the predicted values for each observation.

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?



The data is going to be better approximated with a linear model for the lower values of independent variable. Because the residual data of this linear regression follows a pattern, the lower values of x are closer to the regression line, hence the lower values will be more accurate and closer to the regression line. In contrast to the higher values, these are not close to the regression line, hence the model is not accurate for the higher values.
This make me conclude that our residual data would be better approximated with the linear model for the lower values of independent variable (x). It's also important to mention that this current model is inconsistent in terms of accuracy.

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

The difference between R^2 and adjusted R^2 is that R^2 is based on the sample, and adjusted R^2 is based on the population.

The 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.

In comparison, the adjusted R^2 provides a value that would be expected in the population. This is what we use.

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.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin–Watson
1	.485[a]	.235	.216	2.599	1.724

a. Predictors: (Constant), Maternal height (cm)

b. Dependent Variable: Length of baby (cm)

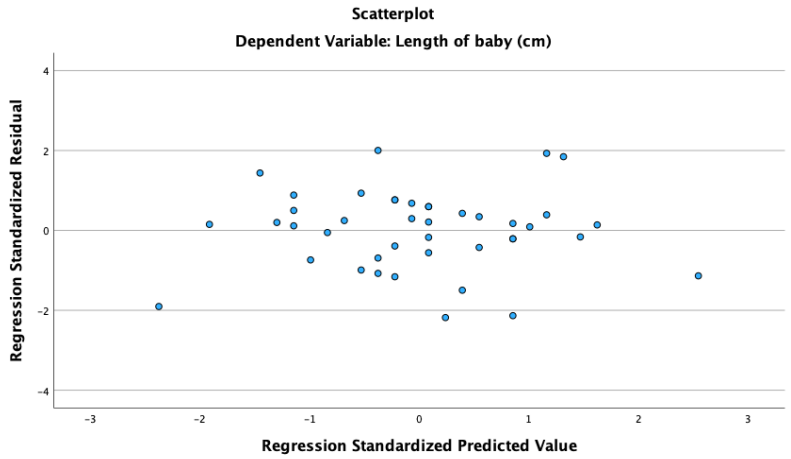
The Durbin-Watson value is 1.72, meaning that there is independence of observations for baby length (dependent/outcome) and maternal height (independent/predictor).
In order to confirm the independence of observations the Durbin-Watson value should be around 1.5 - 2.5).

Q6. Do residual data show homoscedasticity?

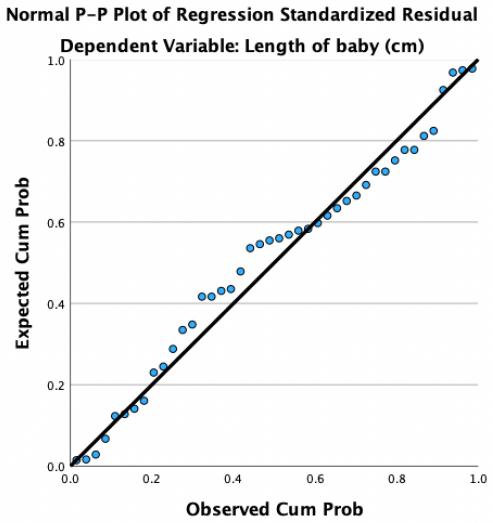
- **Yes**
- **No**

Yes.

Q7. Justify the above answer.



I see that the residual data shows homoscedasticity, because it's randomly scattered on the plot and doesn't follow any pattern.



The normality also looks fine, there is a bit of deviation between 0.25 and 0.55.

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

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.485[a]	.235	.216	2.599	1.724

a. Predictors: (Constant), Maternal height (cm)
b. Dependent Variable: Length of baby (cm)

The R^2 value is 0.23, this indicates that 23% of the variation in length of baby can be explained by maternal height.

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

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	83.110	1	83.110	12.302	.001[b]
	Residual	270.223	40	6.756		
	Total	353.333	41			

a. Dependent Variable: Length of baby (cm)
b. Predictors: (Constant), Maternal height (cm)

The relationship between mother's height and baby length is statistically significant linear, because the significance is lower than 0.05. The value obtained in the significance is 0.001

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

The null hypothesis is that the mother's height doesn't affect the baby length.

The alternative hypothesis is that the mother's height affects the baby length, in this case we reject the null hypothesis and accept the alternative hypothesis, because we find statistically significant linear relationship between the mother's height and the baby length.

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

b_1 is the Slope coefficient, that represents the change in the dependent variable for a unit of change in the independent variable.

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	15.334	10.271		1.493	.143	-5.425	36.093
	Maternal height (cm)	.219	.062	.485	3.507	.001	.093	.345

a. Dependent Variable: Length of baby (cm)

The slope coefficient is 0.219, which means that for every extra one cm of mother's height, the baby's length will increase for 0.219 cm.

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

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Maternal height (cm)	42	149	181	164.45	6.504
Valid N (listwise)	42				

I cannot claim the same for mother's height in the range between 140 cm and 145 cm, because there is no mother in the dataset with height values between 140 cm and 145 cm.
The minimum height is 149 cm and the maximum height is 181 cm.
Hence, the slope coefficient doesn't apply to the values within this range (140 cm - 145 cm).
The slope coefficient only works for the range of values between the minimum and maximum value of the independent variable (maternal height).

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

First, I confirm that I can do the prediction for 170 cm because the value is within the range.
The prediction of baby length for mother's height of 170 cm is 52.56 cm of baby length.

15.33 + (0.219 * 170) = 52.56 cm

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 length, F (1, 40) = 12.302, p-value = 0.001 and mother's height accounted for 23% (R^2

= 0.23) of the explained variability in baby length. The regression equation: predicted baby length = 15.33 + (0.219 * mother height value)

Both variables are related, there is statistically significant linear relationship. There is independence of observations for baby length (dependent/outcome) and maternal height (independent/predictor).
For mother height of 170 cm we expect baby length of 52.56 cm with 95% confidence interval that is between 52.467 and 52.905, this is our prediction.

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

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.650	1	6.650	.767	.386[b]
	Residual	346.684	40	8.667		
	Total	353.333	41			

a. Dependent Variable: Length of baby (cm)
b. Predictors: (Constant), Father's age

I cannot predict the baby length with father's age, given the fact the significance value is 0.386 (higher than 0.05). There is no statistically significant linear relationship between these 2 variables, hence we accept the null hypothesis.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin–Watson
1	.137[a]	.019	-.006	2.944	1.736

a. Predictors: (Constant), Father's age
b. Dependent Variable: Length of baby (cm)

I also see that there is independence of observations for length of baby and father's age, the Durbin-Watson value is 1.73 (within the range of 1.5 - 2.5).

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

Homogeneity of variance is the assumption that the spread of scores is roughly equal in different groups of cases, or more generally that the spread of scores is roughly equal at different points on the predictor variable.
The homogeneity of variance is one of the assumptions of Independent means t-test.
The variances in the populations have to be roughly equal – homogeneity
of variance (Levene's test > 0.05). If Levene's test value is below 0.05, this means that there is no homogeneity of variance, and we cannot proceed with the Independent means t-test.

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, there is homogeneity of variance between head circumference for babies of smoking mothers and head circumference for babies of non-smoking mothers.

Q19. Justify your choice.

1. First of all, I check that the sampling distribution is normally distributed. Both groups are normally distributed. Both significance values are larger than 0.05

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	smoker	Statistic	df	Sig.	Statistic	df	Sig.
Head circumference (cm)	Non-smoker	.192	20	.051	.917	20	.085
	Smoker	.128	22	.200*	.954	22	.372

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

2. I confirm that Head circumference is a continuous variable.
3. They are 2 independent groups, smoking(1) and non smoking(0).

Independent Samples Test

		Levene's Test for Equality of Variances			
		F	Sig.	t	
Head circumference (cm)	Equal variances assumed	.828	.368	1.176	
	Equal variances not assumed			1.189	3

4. The significance value obtained in Levene's test is 0.36, this means that there is homogeneity of variance (because the value is larger than 0.05)

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

Yes, smoking mothers have lighter babies

I conduct an Independent means t-test to answer this question.

1. First of all, I check that the sampling distribution is normally distributed. Both groups are normally distributed. Both significance values are larger than 0.05

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	smoker	Statistic	df	Sig.	Statistic	df	Sig.
Birthweight (kg)	Non-smoker	.128	20	.200*	.967	20	.696
	Smoker	.095	22	.200*	.982	22	.949

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

2. Birth weight is a continuous variable

3. They are 2 independent groups, smoking(1) and non smoking(0).

		Levene's Test for Equality of Variances	
		F	Sig.
Birthweight (kg)	Equal variances assumed	.305	.584
	Equal variances not assumed		

4. The significance value obtained in Levene's test is 0.58, this means that there is homogeneity of variance (because the value is larger than 0.05)

Since there is homogeneity I look at the significance value of the first upper row, the significance value is 0.043. The p-value is below 0.05, this means that we find statistically significance difference between the 2 groups.
The mean difference is 0.38, which is 380 grams, hence the mean difference between the 2 groups is 380 grams in birth weight.

	t-test for		
	Significance		Mean
df	One-Sided p	Two-Sided p	Difference
40	.021	.043	.37541
39.618	.020	.041	.37541

I also check the mean for both groups and this value also confirms that smoking mothers have lighter babys (mean = 3.13) than non-smoking mothers (mean = 3.51).
The effect size is medium, 0.58.

Group Statistics

	smoker	N	Mean	Std. Deviation	Std. Error Mean
Birthweight (kg)	Non-smoker	20	3.5095	.51849	.11594
	Smoker	22	3.1341	.63125	.13458

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

No, women over 35 don't have lighter babies.

I conduct an Independent means t-test to answer this question.

1. First of all, I check that the sampling distribution is normally distributed. Both groups are normally distributed. Both significance values are larger than 0.05

Tests of Normality

	Mother aged over 35	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Birthweight (kg)	Aged < 35	.072	38	.200*	.988	38	.955
	Aged 35+	.187	4	.	.980	4	.900

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

2. Birth weight is a continuous variable
3. They are 2 independent groups, mother over 35, 0 = No and 1 = yes

4. The significance value obtained in Levene's test is 0.166, meaning there is homogeneity of variance (because the value is larger than 0.05).

Indepen

		Levene's Test for Equality of Variances F	Sig.	Significance One-Sided p	Two-Sided p
Birthweight (kg)	Equal variances assumed	1.988	.166	.246	.492
	Equal variances not assumed			.330	.660

Given the fact the p-value is 0.49 (above 0.05), this means that we cannot reject the null hypothesis, there is no statistically significance difference between the 2 groups.
This means that both groups are similar, hence I conclude that I didn't find enough evidence to affirm that women over 35 have lighter babies.

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

Yes, the diet was effective in lowering cholesterol concentration after 8 weeks of use.

First of all, I check the Dependent means t-test assumptions:

1. The sampling distribution of the differences* between scores should be (approximately) normally distributed.
   The variable of the differences (Cholesterol after 8 weeks on the diet (mmol/L) - Cholesterol before the diet (mmol/L)) is normally distributed (significance value = 0.98)

Tests of Normality

	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
diff_8	.114	18	.200*	.985	18	.987

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

2. Both variables are continuous

Paired Samples Test

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

3. I conduct the Dependent means t-test

The significance value is very low, <.001 (below 0.05), we reject the null hypothesis. I can confirm that there is difference between the 2 groups (Cholesterol after 8 weeks on the diet (mmol/L) and Cholesterol before the diet (mmol/L)).

The mean cholesterol concentration difference is 0.63.
The mean of cholesterol concentration before the diet was 6.40, and after 8 weeks of diet, the mean cholesterol concentration is lower 5.77.
The cholesterol went down, and the difference is significant. The diet is effective in lowering cholesterol. The effect is small, 0.17.

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 Effect Sizes

			Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
Pair 1	Before - After 8 weeks	Cohen's d	.17852	3.523	2.255	4.776
		Hedges' correction	.18691	3.365	2.153	4.562

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference.
Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

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

The null hypothesis is that the 2 groups are not different. Hence the diet is not effective in lowering cholesterol.
The alternative hypothesis is that the 2 groups are different, meaning that the diet is effective in lowering cholesterol.

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

The margarine diet was more effective after 8 weeks of use.
The mean cholesterol concentration difference is 0.63 for the variable after 8 weeks. In contrast, the mean cholesterol concentration difference is 0.56 for the variable after 4 weeks.
After 4 weeks of diet, the mean cholesterol concentration is 5.84 lower, and

after 8 weeks of diet, the mean cholesterol concentration is 5.77 lower. The mean for the cholesterol concentration is lower after 8 weeks of diet. The effect is also higher after 8 weeks of diet (0.17) than after 4 weeks of diet (0.15).
These values confirm that the margarine diet was more effective after 8 weeks of diet.

I've followed these steps to calculate if the diet was effective in lowering cholesterol concentration after 4 weeks of use.

First of all, I check the Dependent means t-test assumptions:

1. The sampling distribution of the differences* between scores should be (approximately) normally distributed.
   The variable of the differences (Cholesterol after 4 weeks on the diet (mmol/L) - Cholesterol before the diet (mmol/L)) is normally distributed (significance value = 0.92)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
diff_4	.142	18	.200*	.977	18	.920

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

2. Both variables are continuous

Paired Samples Test

		Paired Differences					t	df	Significance	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			One-Sided p	Two-Sided p
Pair 1	Before - After 4 weeks	.56611	.15557	.03667	.48875	.64347	15.439	17	<.001	<.001

3. I conduct the Dependent means t-test

The significance value is very low, <.001 (below 0.05). We reject the null hypothesis.
I can confirm that there is difference between the 2 groups (Cholesterol after 4 weeks on the diet (mmol/L) and Cholesterol before the diet (mmol/L)).
The mean cholesterol concentration difference is 0.56

The mean cholesterol concentration before the diet was 6.40, and after 4 weeks of diet, the mean cholesterol concentration is lower 5.84.
The cholesterol went down, and the difference is significant. The diet is effective in lowering cholesterol. The effect is small, 0.15

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Before	6.4078	18	1.19109	.28074
	After 4 weeks	5.8417	18	1.12335	.26478

Paired Samples Effect Sizes

			Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
Pair 1	Before - After 4 weeks	Cohen's d	.15557	3.639	2.335	4.929
		Hedges' correction	.16288	3.476	2.230	4.707

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference.
Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

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 people in our sample are significantly worse than the average adult population.
I've conducted a One-Sample Test, and I found significance, which means that our variable before (Cholesterol before the diet (mmol/L)) is significantly higher in cholesterol than the average healthy (3 mmol/L).

To obtain the values, I've followed these steps:

1. First of all, I check the normality of the variable before. The variable is normally distributed.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Before	.115	18	.200*	.982	18	.967

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

2. The variable is continuous
3. I conduct the One-Sample Test

One-Sample Test

Test Value = 3

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

We have significance (<.001), meaning that the values we had before are significantly different than 3.
The mean for the variable before is 6.40, and the mean in healthy adults is 3. The variable before is significantly higher in cholesterol than the average healthy (3), the people from our sample are not healthy; they were unhealthy when they started the diet.

One-Sample Statistics

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