📘 WEEK 7 – Multiple Regression, Confounding & Prediction

🔷 1. What Is Multiple Linear Regression?

It models the relationship between **one continuous outcome** and **multiple predictors** (continuous or categorical).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

- **Y** = dependent variable (e.g. weight, crime rate)
- **X_1, X_2...** = independent variables (e.g. exercise, diet, age)
- **β_1, β_2...** = **partial regression coefficients** (change in Y for 1-unit change in X, holding others constant)
- **ε** = error/residual

✅ SPSS Steps to Run Multiple Linear Regression:

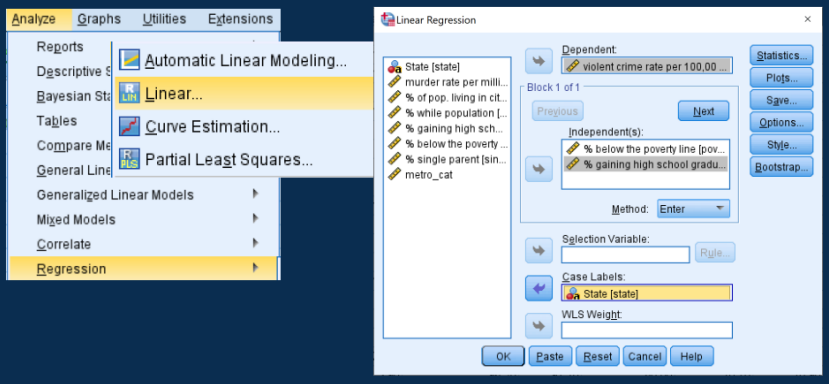
1. `Analyze > Regression > Linear`
2. Dependent: your outcome variable (e.g., `weight`)
3. Independent(s): your predictors (e.g., `exercise`, `diet`)
4. Click `Statistics` → Check **Estimates**, **Confidence Intervals**
5. Optional: `Save` → tick **Predicted values** or **Residuals**
6. Optional: `Plots` → ZRESID (Y) vs ZPRED (X) for residual plot

SPSS Slide: 'how to'

Researchers believe, in the population from which our data came, the % below the poverty line and % gaining a high school graduation have and effect on the Violent Crime rate

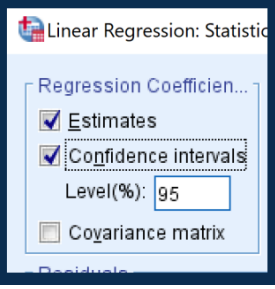
Step 1) Computing a multiple linear regression model for dependent variable '**crime**' and independent variables '**poverty**' and '**education**'

Use **Analyse -> Regression -> Linear**



Put '**crime**' in 'dependent', and '**poverty**' and '**education**' in 'independent'.

Click **Statistics**, select '**Confidence intervals**'.



Output and Interpretation Slide

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	345.852	1026.638		.337	.738	-1719.478	2411.181
	% below the poverty line	23.927	14.763	.347	1.621	.112	-5.774	53.627
	% gaining high school graduation	-1.502	11.239	-.029	-.134	.894	-24.112	21.109

a. Dependent Variable: violent crime rate per 100,000 population

β_0 β_1 β_2

$$y = 345.852 + 23.927\,x_1 - 1.502\,x_2$$

The intercept (β_0), is the extrapolated Violent Crime Rate at 0% below the poverty line and 0% of high school education

The estimated slope coefficient (β_1), suggests a 1% increase in poverty is associated with a 23.927 increase in Violent crime rate per 100 000 holding % of education constant (or adjusting for % of education).

The estimated slope coefficient (β_2), suggests a 1% increase in education is associated with a 1.502 decrease in Violent crime rate per 100 000 holding % poverty constant (or adjusting for % of poverty).

Output and Interpretation Slide

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)	345.852	1026.638		.337	.738	-1719.478	2411.181
	% below the poverty line	23.927	14.763	.347	1.621	.112	-5.774	53.627
	% gaining high school graduation	-1.502	11.239	-.029	-.134	.894	-24.112	21.109

a. Dependent Variable: violent crime rate per 100,000 population

Based on the multiple regression, poverty (x_1) has a partial regression coefficient β_1 of 23.927, with a 95% CI [-5.774, 53.627]

Given the hypothesis test for β_1: $H_0: \beta_1=0$, $H_a: \beta_1\neq0$ gives p=0.112 the result is not significant.

We conclude that poverty is not statistically significantly associated with crime when the poverty-crime relationship is adjusted for education (or education is held constant). We cannot generalise that poverty is associated with crime in the population (β_1= 23.927, t=1.621, p=0.112, 95%CI (-5.774, 53.627))

🔷 2. Interpreting Coefficients

Example:

$$\text{Weight} = 72 - 4 \times \text{Exercise} - 2 \times \text{VegIntake}$$

- **β₁ (exercise)** = -4 → Each extra session decreases weight by 4kg (holding vegetables constant)
- **β₂ (vegetables)** = -2 → Each additional veg/day lowers weight by 2kg (holding exercise constant)

🔷 3. Confounding and Adjustment

🔁 **Confounding happens when:**

A third variable **influences both** the independent and dependent variable, giving a **distorted relationship**.

> Example:
> Exercise appears to reduce weight. But **free time** might increase both exercise *and* weight control — making it a **confounder**.

Confounding Variables

Confounding: A situation in which the association between an explanatory variable (e.g exercise x_1) and outcome (e.g. weight y) is distorted by the presence of another variable (e.g. hours of free time $\mathbf{x_2}$).

Theory:

hours of free time causes **more exercise**
hours of free time also affects **weight**

i.e. hours of free time is a **common cause** of the exposure (exercise) and outcome (weight) of interest.

What happens if we only test the relationship between exercise (x_1) and weight (y) in a simple linear regression model with only exercise as an independent variable?

Confounding Problem

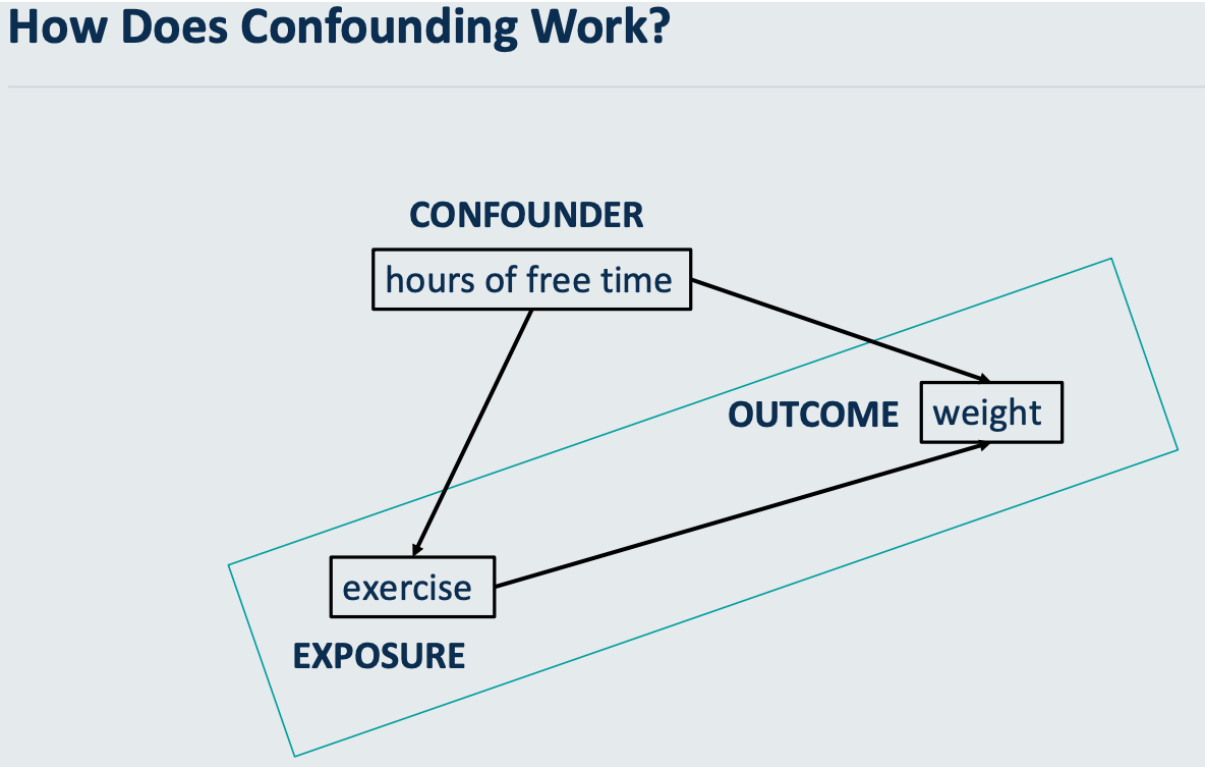
It appears:

- Weight might go down with more free time.
- Those with more free time might exercise more.

So even if more exercise did not cause any weight loss, we might still observe an association between exercise and weight.

In the presence of such a common cause, we **cannot attribute all the observed association** between the exposure and the outcome **to the exposure causing the outcome**.

This is known as **confounding**; or in other words free time is a **confounder** of the effect of exercise on weight.



✅ **To adjust for confounders:**

Include them as **additional variables** in the regression model.

SPSS Command:

- Just add the confounder to the list of independent variables

🔷 4. Prediction Using the Regression Model

To predict:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Plug in values for X_1 and X_2
- SPSS can **save predicted values** via:
  `Regression > Save > Predicted values`

🔷 5. R^2 and Adjusted R^2

Metric	Meaning
R^2	Proportion of variance in Y explained by the model
Adjusted R^2	Corrects for number of predictors (preferred for model comparison)

> R^2 ranges from 0 to 1
> R^2 = 0.6 → model explains 60% of the variation in Y

R^2 – The Coefficient of Determination

- The **coefficient of determination**, denoted R^2 and pronounced R-Squared, is a statistical measure of how well the regression line/hyperplane approximates the real data points.
- It is also known as a measure of **goodness of fit:** The goodness of fit of any statistical model describes how well it fits a set of observations.
- $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$ where SS = sum of squares, **res** = residuals (or errors) and **tot** = total
- R^2 ranges from 0 to 1.
  - ➢ R^2 of 0 indicates poor fit; the regression line would be perfectly horizontal.
  - ➢ R^2 of 1 indicates perfect fit; the regression line/hyperplane fit exactly to all data points.

R^2 – continued

- R^2 measures the fit of the model both in simple and multiple linear regression.
- In **simple linear regression $R^2 = r^2$**, where r is the Pearson correlation.
- In a context of regression where we are assessing **associations between variables**, R^2 is often interpreted as the proportion of the variance in the dependent variable that is "explained" by the independent variables in the model.
  - ➢ In our earlier example, this would be the proportion of variance in the weight that is explained by frequency of exercise and hours of free time.
  - ➢ R^2 of 0 indicates that none of the variance in y is explained.
  - ➢ R^2 of 1 indicates that 100% of the variance in y is explained.
- In a context of **prediction analysis**, R^2 is often interpreted as how well the model will be able to predict values of Y based on observed values for the independent variables x_i; with higher values of R^2 indicating better prediction.

Adjusted R^2 as a Measure for Model Selection

Adjusted R^2 (denoted R^2_{adj}) is a modified version of R^2 that adjusts for the **number of independent variables *p*** in the model:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

R^2_{adj} takes account of the phenomenon whereby R^2 **increases** every time an **extra independent variable** is added regardless of whether this added variable adds substantially to the explanation of dependent variable variance.

R^2_{adj} increases only when the increase in R^2 (due to the inclusion of a new independent variable) is more than one would expect to see by chance.

$R^2_{\text{adj}}$is considered to be a **better indicator for model selection:** between different models, the one with **higher** R^2_{adj} is the one that better fits the data, and should be selected.

SPSS Interpretation Slide

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	345.852	1026.638		.337	.738	-1719.478	2411.181
	% below the poverty line	23.927	14.763	.347	1.621	.112	-5.774	53.627
	% gaining high school graduation	-1.502	11.239	-.029	-.134	.894	-24.112	21.109

a. Dependent Variable: violent crime rate per 100,000 population

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.369[a]	.136	.100	280.763

a. Predictors: (Constant), % gaining high school graduation, % below the poverty line

The linear multiple regression model has an R^2_{adj} of 0.100. Poverty and education explained 10.0% of the variance in violent crime.

Knowledge Check - R^2

The Psychosis department at the IoPPN is investigating whether quality of life in people diagnosed with schizophrenia depends on a series of demographic and clinical variables. They have asked us to help them choose among different models.

Q4: Which one should they keep as the best model?

Dependent variable:
Quality of Life (QoL) measured with QOLS scale (ranging from 16 to 112)
Independent variables:
Severity of illness, age, gender (1=female), marital status (1=married)

Model	y	β_0	Severity β_1 (p-value)	Age β_2 (p-value)	Gender β_3 (p-value)	Marital Status β_4 (p-value)	R^2_{adj}
I	QOLS	50	-3.4 (0.01)	-2.1 (0.10)	Not included	5.1 (0.001)	0.73
II	QOLS	47	Not included	-1.8 (0.07)	1.03 (0.13)	6.2 (0.002)	0.51
III	QOLS	56	-3.1 (0.02)	Not included	Not included	5.3 (0.001)	0.85

Knowledge Check Solutions – R^2

Q4: Which one should they keep as the best model and why?

The best model is the model III with Severity of illness and status as the independent variables. This is because we see from the adjusted R^2 that it explains 85% of the variability in quality of life. If we compare to model I we can see that adding age decreased the adjusted R^2 – this makes sense in combination with the fact that age doesn't seem to be a significant predictor of quality of life. Model II has a lower adjusted R^2 because it is missing the important severity predictor.

Model	y	β_0	Severity β_1 (p val)	Age β_2 (p val)	Gender β_3 (p val)	Marital Status β_4 (p val)	R^2_{adj}
I	QOLS	50	-3.4 (0.01)	-2.1 (0.10)	Not included	5.1 (0.001)	0.73
II	QOLS	47	Not included	-1.8 (0.07)	1.03 (0.13)	6.2 (0.002)	0.51
III	**QOLS**	**56**	**-3.1 (0.02)**	**Not included**	**Not included**	**5.3 (0.001)**	**0.85**

🔷 6. Regression Assumptions

You **must check** these to trust your regression results:

Assumption	How to Check in SPSS
Linearity	Partial residual plots (`Plots > Produce all partial plots`)
Normality of errors	Histogram + P-P Plot of residuals
Homoscedasticity	ZRESID vs ZPRED scatterplot
Independence	Depends on design (not testable in SPSS)

✅ SPSS for Assumption Checks:

1. `Analyze > Regression > Linear`
2. Click `Plots`:
    - ZRESID in Y-axis, ZPRED in X-axis (for homoscedasticity)
    - Tick **Normal probability plot** and **Histogram**
    - Tick **Produce all partial plots** (for linearity)

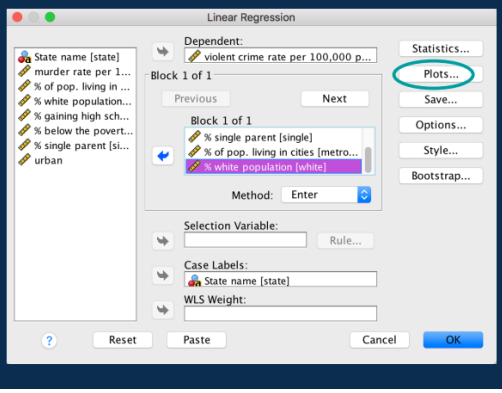
SPSS Slide: 'how to'

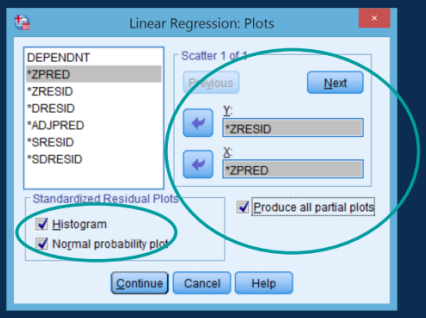
Assessing assumptions to make inference from a multiple linear regression model for crime rate from Lecture_7_data.sav data.

Use **Analyse -> Regression -> Linear**

Put '**crime**' in **dependent**, and **poverty**, **education**, **single**, **metropol** and **white** in '**independent**'.

Click '**Plots**', select 'Histogram', 'Normal probability plot', 'Produce all partial plots', put ZRESID (standardised residuals) in Y and ZPRED (standardised predicted values) in X.

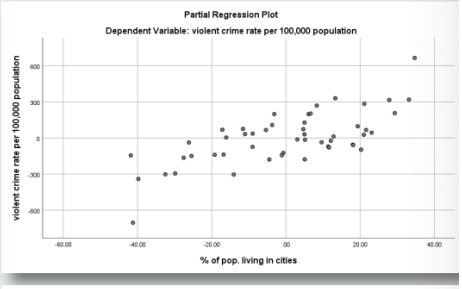




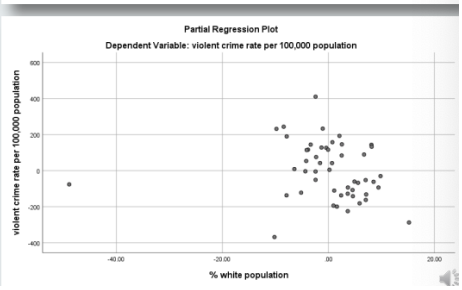
Output and Interpretation Slide – Assessing Linearity

Assumption #1

Partial residual plots from the regression of **crime** on **poverty, edu, single, metropol**, and **white –** note only two of the five partial plots are shown:



Top plot suggests a linear relationship between the independent variable **metropol** and the outcome variable **crime** – the **linearity assumption is met** for the **metropol** variable

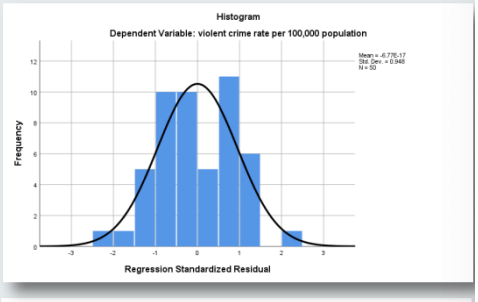


Bottom plot suggests there is not a linear relationship between the independent variable **white** and the outcome variable **crime** – **the linearity assumption is not met** for the **white** variable

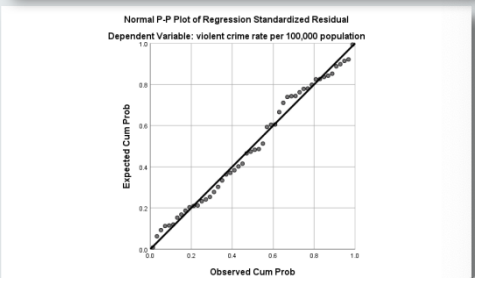
SPSS Interpretation Slide – Plots of Residuals for Assessing Normality

Assumption #2

Histogram – a gap at the right and possibly somewhat skewed, but errors/residuals look more or less normally distributed.



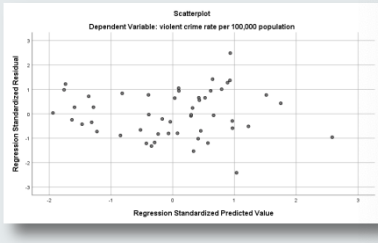
Normal P-P plot – gives similar information to the histogram; here we want to see that the points lie more or less close to the diagonal reference line, which they do.



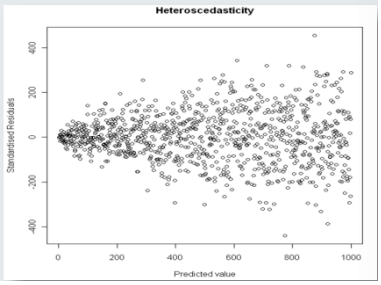
SPSS Interpretation Slide – Assessing Variance Homogeneity

Assumption #3

Top plot = **homoscedastic** = meets assumption.
There is no obvious trend, the residuals **scatter randomly** above and below zero. The scatter around the horizontal zero line is roughly constant, suggesting a constant variance. We can assume homogeneity and make inferences from our regression model.



Bottom plot = **heteroscedastic** = does not meet assumption. Here the residual variance **increases** with the size of the predicted value. Homoscedasticity cannot be assumed, we cannot make inferences from our model, and would need to use approaches beyond the scope of this course. This is an example of heteroscedasticity.



Quiz:

The model $y = 3 + 5x - 2z + \epsilon$ indicates?

Select one:

- a. For every one unit increase in *x*, the predicted value for *y* increases by 5 when *z* is held constant.
- b. *x* and *z* are significant predictors of *y*.
- c. *x*, *y* and *z* are significantly associated.
- d. A change of 5 units in *x* leads to a 2 unit decrease in *y*. ×

Your answer is incorrect.

A change of 1 unit of x, associates with an increase of 5 units in y, keeping all other variables constant.

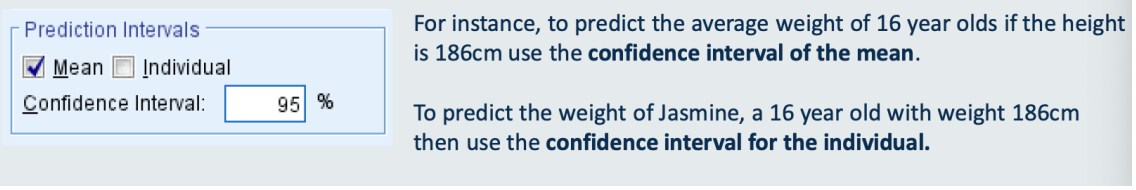
The correct answer is: For every one unit increase in *x*, the predicted value for *y* increases by 5 when *z* is held constant.

Knowledge Quiz:

If necessary help the journalist to correct his report. Use the best linear regression model to predict the field goals, given that the player got 6 total points in average per play, and his weight is 225. Please also give a 95% confidence interval for the prediction.

"*Kevin White, player of the Chicago Bears, played an amazing season, being a key role member for his team. The player scored 6 points in average per play, and he had a* [44.47] ✓ *percent of successful field goals. Kevin White, informing for ITV news, Chicago*"

The 95% confidence interval for the predicted value is [[41.54] × , [53.73] ✓]



When predicting a value-individual

Average value-mean!!

✅ Summary

Component	Your Answer	Correct?	Notes
Predicted value	44.47%	✅	Accurate
Confidence interval	42.62–46.31%	❌	You used CI for the mean
Correct interval	41.54–53.73%	✅	Use CI for **individual** prediction

In SPSS, redo:

1. Go to: `Analyze > Regression > Linear`
2. Click **"Save"**
3. Under **"Prediction Intervals"**, tick ✅ **"Confidence interval for individual prediction"**
4. Then re-run the regression with the player's inputs.

Check the inference assumptions for the linear model derived in Q6 and fill the gaps.

1) We already know from Q3 that there is a [linear] ✓ relationship between 'goals' and 'points'.

2) We plot the [scatterplot] ✗ of the errors and see they are [normally] ✓ distributed.

3) The error terms have [different variance] ✗ irrespective of the values of x. We can inspect this by plotting a [scatterplot] ✓ of the [independent] ✗ values. The plot showed [the same variance] ✗ .

In summary conditions were [met] ✓ , which [allows] ✓ us to make inferences from the model.

Your answer is partially correct.

You have correctly selected 5.
The correct answer is:
Check the inference assumptions for the linear model derived in Q6 and fill the gaps.

1) We already know from Q3 that there is a [linear] relationship between 'goals' and 'points'.

2) We plot the [histogram] of the errors and see they are [normally] distributed.

3) The error terms have [the same variance] irrespective of the values of x. We can inspect this by plotting a [scatterplot] of the [standardised predicted versus standardised residual] values. The plot showed [no pattern].

In summary conditions were [met], which [allows] us to make inferences from the model.

🧠 Quick Summary of Inference Assumptions:

Assumption	Test in SPSS	Visual/Tool
Linearity	Visual inspection of scatterplots	Scatterplot of Y vs X
Normality of residuals	Histogram or Q-Q Plot	Histogram of residuals
Homoscedasticity	Constant variance in errors	Scatterplot: predicted vs residuals
Independence of errors	Usually checked in time series (Durbin-Watson)	Not relevant here

📊 2. Why use a *histogram* of residuals?

You use a **histogram of residuals** to check whether the **residuals are normally distributed**, which is one of the key assumptions in linear regression.
This is **Assumption 2: Normality of residuals**.

- If the residuals are **normally distributed**, it supports valid inference (like CIs and p-values).
- If the histogram looks symmetric and bell-shaped, that's good.

✅ This was the **second blank** in your question — the correct answer is **histogram**.

🔍 1. What are *error terms* in regression?

In a regression model, **error terms (or residuals)** represent the difference between the **observed values** and the **predicted values** from your regression equation.

- **Mathematically**:
  $\text{Error} = Y_{\text{observed}} - Y_{\text{predicted}}$

These errors help us understand **how well our model fits the data** and are used to assess key assumptions in regression.

📈 3. Why use a *scatterplot* of residuals vs predicted values?

This is used to check the assumption of **homoscedasticity**, i.e., that the residuals have **constant variance** across all levels of predicted values.

This is **Assumption 3: Homogeneity of variance** (a.k.a. **no pattern** in residuals).

- A **random scatter of points** = good.
- A **funnel shape** = problematic (shows heteroscedasticity).

In your quiz, they were asking for this when they said:

> "...plotting a scatterplot of the standardised predicted vs standardised residual values..."

✅ Correct terms: **scatterplot**, **standardised predicted vs standardised residual**, and **no pattern**.

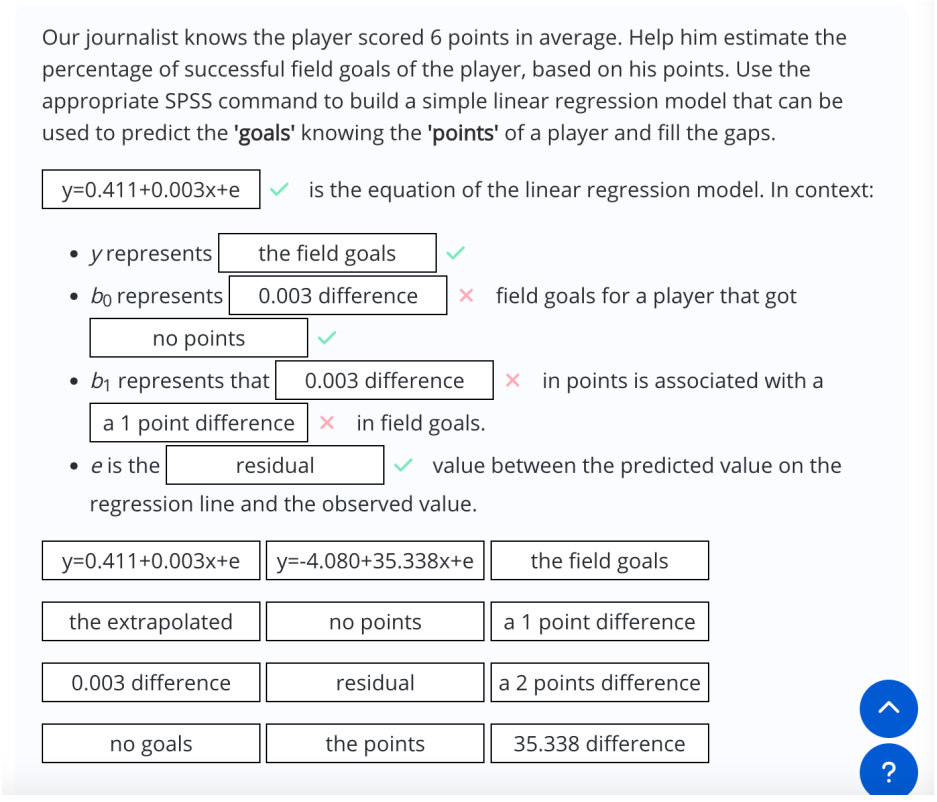
📘 4. Where is this in the lecture slides?

From the file you uploaded earlier (`W7ALL.pdf`), here's where the content is usually covered:

- **Slide titled: "Checking model assumptions"** or **"Regression assumptions"**
- Look for slides that describe:
  - **Normality** (e.g., "Residuals should be normally distributed" → Histogram)
  - **Homoscedasticity** (e.g., "Plot of residuals vs predicted values should have no pattern" → Scatterplot)
  - **Linearity** (e.g., The relationship between predictors and outcome must be linear)

These slides usually have visual examples of:

- Histogram of residuals (bell-shaped or not)
- Scatterplots (no pattern vs funnel shape)



Your answer is partially correct.

You have correctly selected 4.
The correct answer is:
Our journalist knows the player scored 6 points in average. Help him estimate the percentage of successful field goals of the player, based on his points. Use the appropriate SPSS command to build a simple linear regression model that can be used to predict the **'goals'** knowing the **'points'** of a player and fill the gaps.

[y=0.411+0.003x+e] is the equation of the linear regression model. In context:

- *y* represents [the field goals]
- b_0 represents [the extrapolated] field goals for a player that got [no points]
- b_1 represents that [a 1 point difference] in points is associated with a [0.003 difference] in field goals.
- *e* is the [residual] value between the predicted value on the regression line and the observed value.

🧠 Key Components of the Equation

- **y** = the **predicted outcome** — in this case, it's the **percentage of successful field goals**.
- **x** = the **predictor** (independent variable) — in this case, it's the **average points scored per game** by the player.
- **0.411** = **intercept (b_0)** — this represents the estimated **field goal percentage for a player who scores 0 points per game**.
- **0.003** = **slope (b_1)** — for every **1-point increase in average points scored**, the model **predicts a 0.003 increase** in the percentage of successful field goals.
- **e** = **residual** — this is the **difference between the actual observed value and the predicted value** from the model.

🛠️ How to interpret each term in context

- **Intercept ($b_0 = 0.411$)**
  This is **extrapolated** because a player who scores 0 points isn't realistic, but it's still the model's best estimate when x = 0.
- **Slope ($b_1 = 0.003$)**
  This tells us that if a player scores **1 more point per game**, their **predicted field goal percentage increases by 0.003** (or 0.3%).
- **Residual (e)**
  This captures **how far off our prediction is** from the actual data for each player.

✅ Correct Answers (from feedback):

- **y** = the field goals
- b_0 = the **extrapolated** field goals for a player that got **no points**
- b_1 = a **1-point difference** in points is associated with a **0.003 difference** in field goals
- **e** = the **residual** (difference between observed and predicted value)

Topic 7 Knowledge Check Quiz

Q1: What is multiple linear regression?

Method of studying the relationship between one dependent variable and two or more independent variables,

Q2:What is a confounder?

A confounder is a third variable that influences both the independent variable of interest (exposure) and the dependent variable (outcome)

Confounding occurs when an apparently causal relationship between an exposure (e.g. a *treatment*) and an *outcome* is, in reality, distorted by the effect of a third variable (the *confounder*). By definition, confounding factors must fulfil three criteria

1. they must be related to *both* exposure (i.e. risk factor, intervention, or treatment) *and* outcome;
2. they must be distributed *unequally* between study groups; and
3. they must not be an intermediary step in a causal pathway between exposure and outcome.

- Q3: R^2 is a coefficient for what?
- Measuring how well the regression line/hyperplane approximates the real data points.
- R^2 gives a measure of the amount of variation in the dependent variable being explained by the independent variables

Q4: What does the model $y = 3 + 5x - 2z$ indicate?

A change of 1 unit of x, associates with an increase of 5 units in y if z is held constant

Q5: Model A showed an Adjusted R^2 of 0.58. Model B showed an Adjusted R^2 of 0.78

Model B is better as it accounts for more of the variability in the dependent variable by the independent variable accounting for additional predictors in the model

Bonus:

Confounding variables can be thought of as unrecognized independent variables.
Indicate whether this statement is True or False and give a reason why.

What answers did you come up with?