In the trenches of OLS

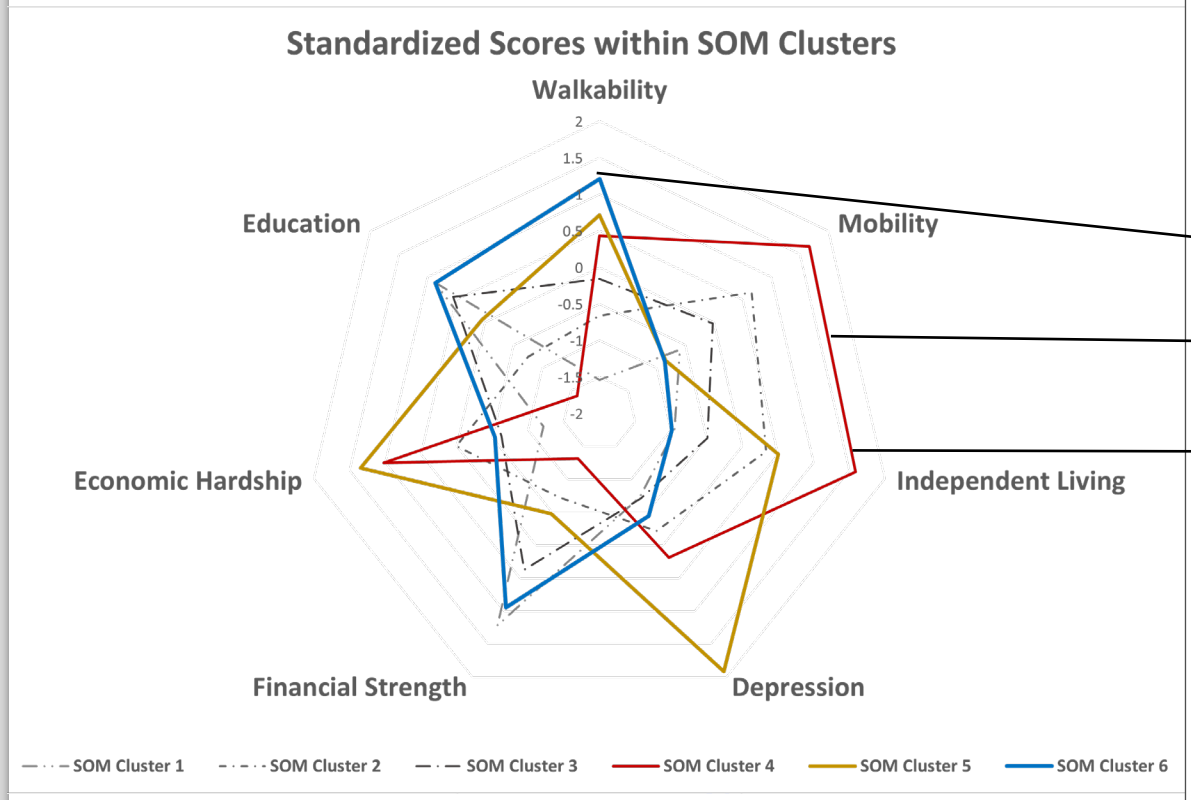
- Mechanics of SLR revisited
- Violation of Assumptions, Common Variance
- Introduction to Multiple Regression
- Examples and Intuition

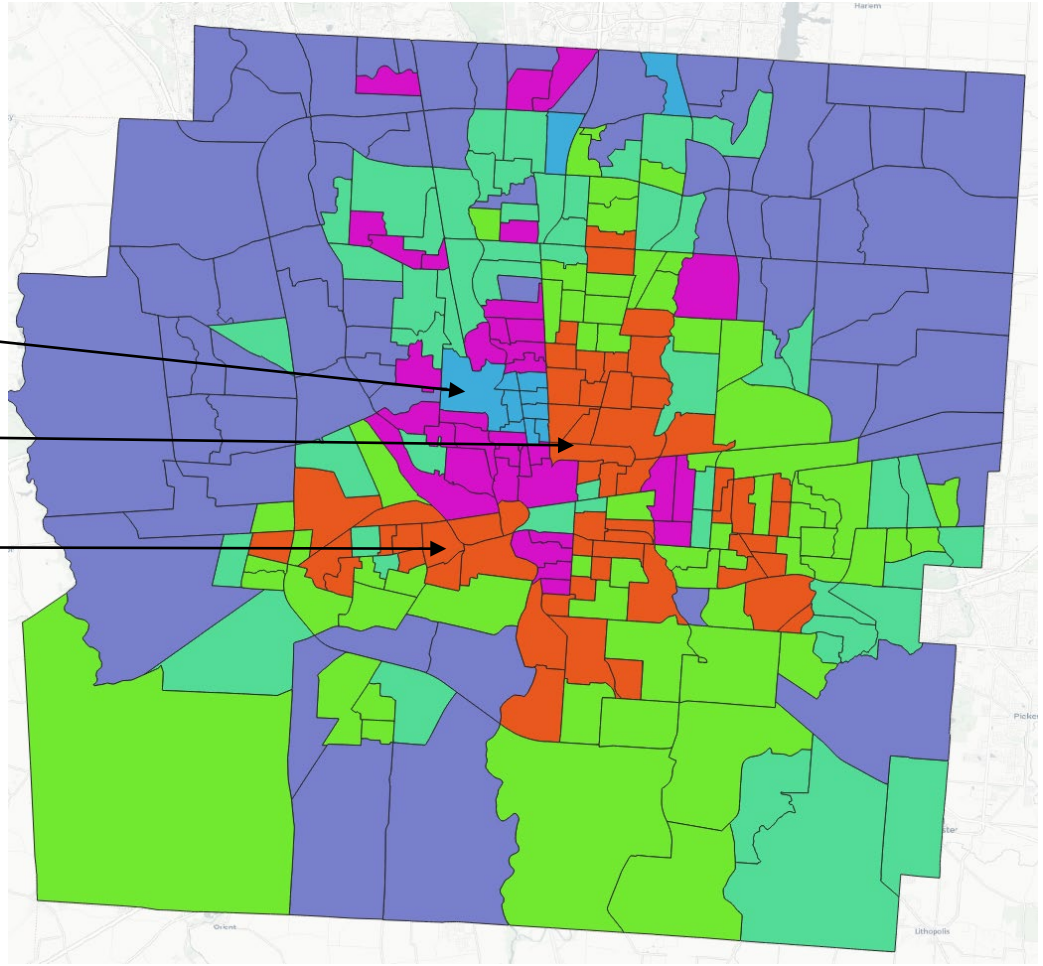
Today's (Next time's) Objectives

- Summarize the four assumptions of the linear regression model
- Understand the population variance σ^2 in the regression setting
- Know how to obtain the estimated *MSE* of the unknown population variance σ^2 from SPSS's fitted line plot and regression analysis output
- Know how to interpret the R^2 value; understand its limitations.
- Evaluate a SLR model

Review mechanics of SLR

- First order linear model
- Best fitting line
- Equations for population and sample regression line
- Interpretation of parameters
- Prediction
- Notation notes


Standardized Scores within SOM Clusters
Walkability
Mobility
Independent Living
Depression
Financial Strength
Economic Hardship
Education
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
SOM Cluster 1
SOM Cluster 2
SOM Cluster 3
SOM Cluster 4
SOM Cluster 5
SOM Cluster 6




NO
OUTLET
© Google


Google

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	indeplv[b]	.	Enter

a. Dependent Variable: deprssn

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.479[a]	.230	.227	2.49688

a. Predictors: (Constant), indeplv

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	524.196	1	524.196	84.081	<.001[b]
	Residual	1758.097	282	6.234		
	Total	2282.292	283			

a. Dependent Variable: deprssn

b. Predictors: (Constant), indeplv

Coefficients[a]

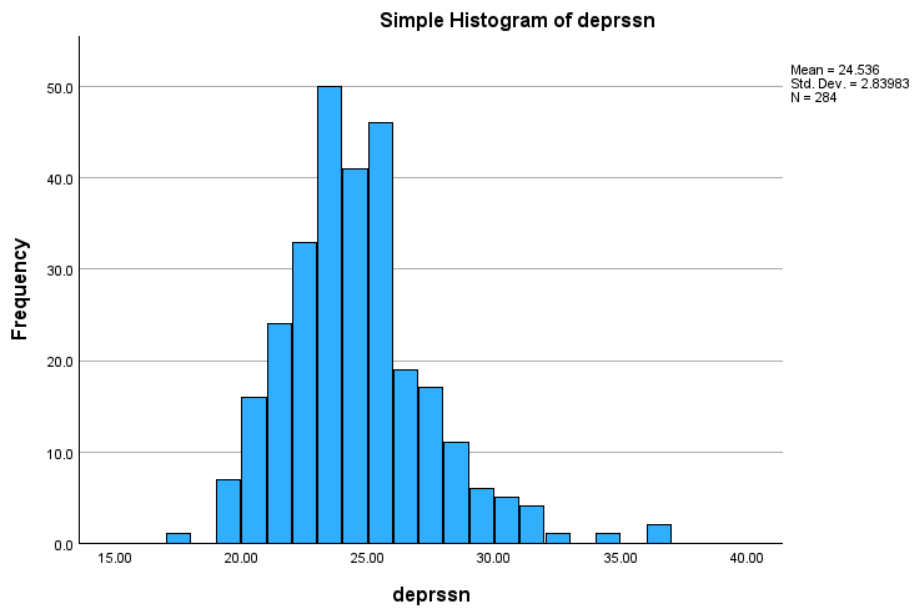
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	21.746	.338		64.254	<.001
	indeplv	.310	.034	.479	9.170	<.001

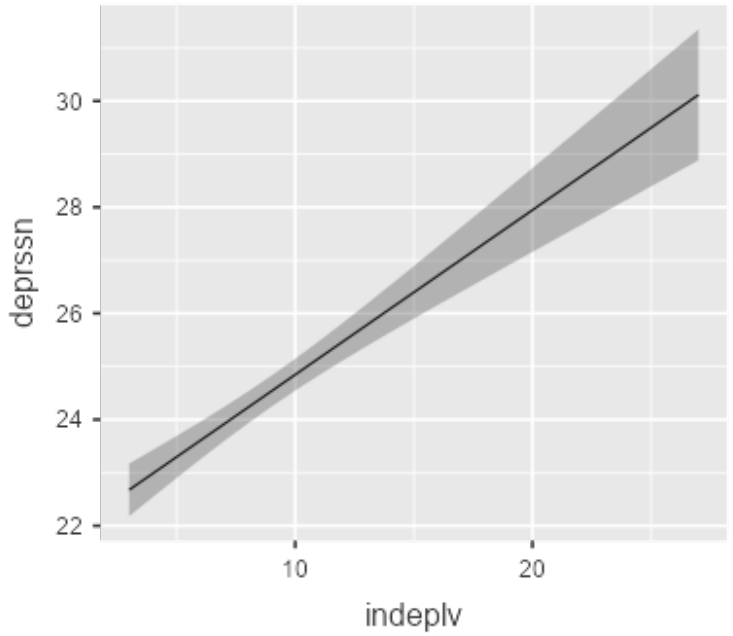
a. Dependent Variable: deprssn

Let's use these results to revisit the SLR mechanics....

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
indeplv	284	3.30	26.10	9.0007	4.39038
deprssn	284	17.80	36.90	24.5360	2.83983
Valid N (listwise)	284				





Estimated Marginal Means - indeplv

			95% Confidence Interval	
indeplv	Marginal Mean	SE	Lower	Upper
4.61 ⁻	23.18	0.21	22.76	23.59
9.00 ᵘ	24.54	0.15	24.24	24.83
13.39 ⁺	25.90	0.21	25.48	26.31

Note. ⁻ mean - 1SD, ᵘ mean, ⁺ mean + 1SD

Deterministic Relationship between 2 variables

Age	ACEs	Test Score
5	7	1400
10	1	200
20	3	600

A functional relation between two variables is expressed by a mathematical formula

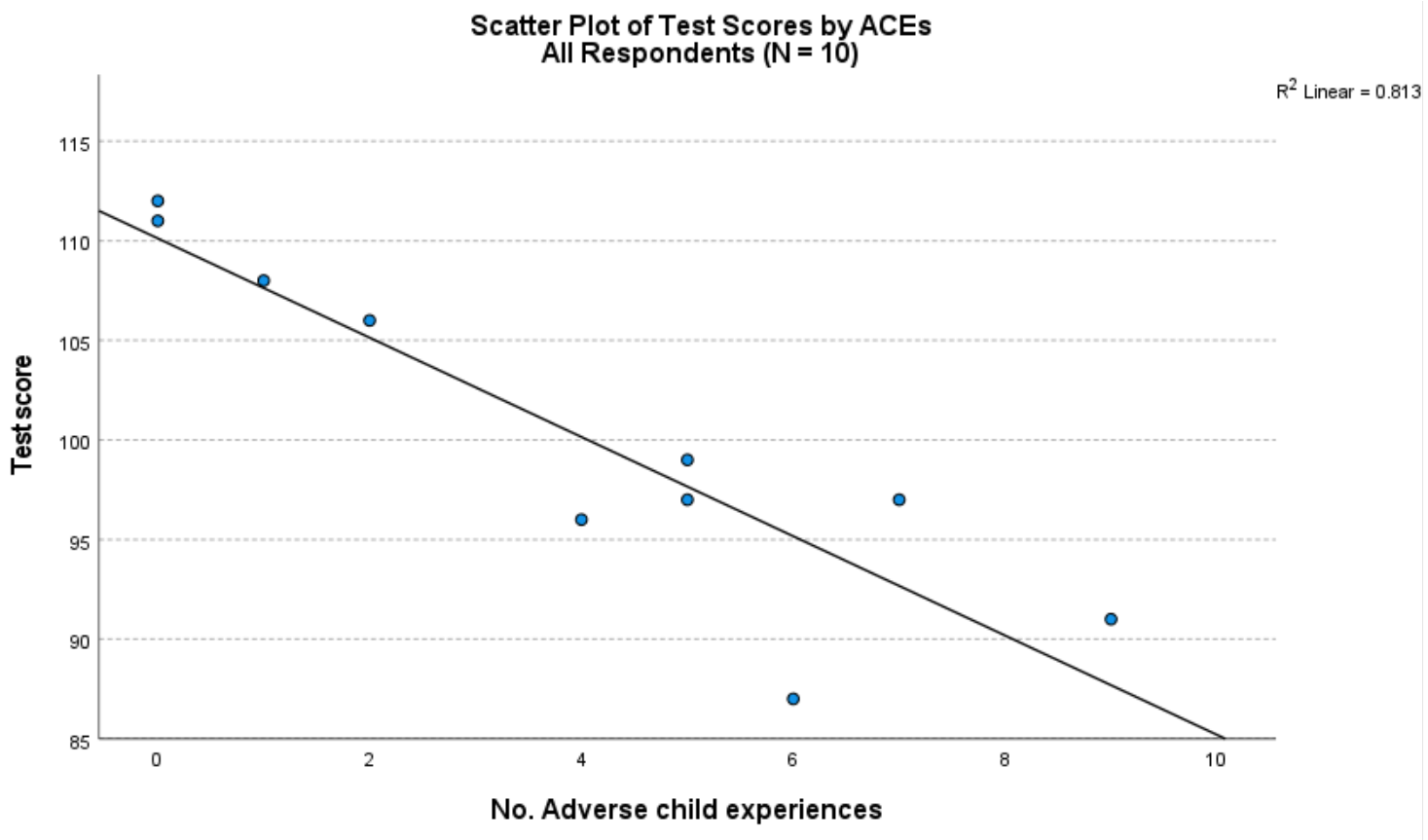
If X denotes the independent variable and Y the dependent variable, a functional relation is:

$$Y = f(X)$$

given a value of *X*, the function *f*(*X*) indicates the value of Y

Example: Y = 200X

Statistical Relationship between 2 Variables



- A statistical relation, unlike a functional relation, is not a perfect one
- The observations for a statistical relation do not fall directly on-the curve of relationship

Simple Linear Regression

- Assumed Probabilistic Linear Model: Assume a linear relationship between X & Y exists
- Observed Y values can be expressed as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The straight line (average)

Deviation from the line

- Use data to estimate that linear relationship with a straight line:

$$\hat{y} = b_0 + b_1 x$$

The Model

- Y_i denotes the observed response for the experimental unit *i* (a fancy way of saying it's the value of the dependent variable for observation *i*)
  - **Note:** again, *i* refers to the line number of the data, so for example if I have 10 rows of data, the number of observations range from *i* = 1 to 10. With *n* rows of data, the number of observations ranges from *i* = 1 … n.
- X_i denotes the value of the independent or predictor variable for experimental unit *i* (the one corresponding to the dependent variable)
- $\hat{y}_i$ is the predicted value (also called fitted value) for experimental unit *i*
- The equation for the best fitting line is given by $\hat{y}_i = b_0 + b_1 x_i$
  - **Note**: experimental unit = unit of analysis

Inference from Regression

- We want to use our regression equation to make predictions for our dependent variable based on the linear combination of independent variables. To do this we need assumptions on how the data is made!
- We imagined that in the population, the values of *Y* really are generated from the linear probabilistic model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

  where ε is an error term normally distributed with mean 0 and variance σ^2

- For our linear regression inference, we will always assume ε has a normal distribution, N(0, σ^2)

Regression Model Error

- The standard assumption is that the error term is normally distributed
- The model you are to use for SLR (which will be extended to the multivariate case) is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$\beta_1 > 0 \Rightarrow$ Positive Association

$\beta_1 < 0 \Rightarrow$ Negative Association

$\beta_1 = 0 \Rightarrow$ No Association

Y_i The observed response for the *i*th trial (or row of data)

X_i A known constant, the level of the predictor in the *i*th trial

β_0, β_1 Unknown regression parameters

ε_i Error term, independently and normally distributed, N(0, σ^2)

i The index for the total number of cases in your data, *i* = 1, … *N*

Interpretation of Regression Equation

$$\hat{y} = b_0 + b_1 x$$

- The magnitude of the slope b_1 tells how much Y is expected to increase per unit increase in X.

- Example: $\hat{y}$ = 3 + 2x
  - b_0 = 3, $b_1 = 2$
  - If *X* increases by 1, *Y* is expected to increase by 2
  - If *X* increases by 1, then the predicted value of *Y* increases by 2

Interpreting the Intercept

- The intercept b_0 tells the predicted value $\hat{y}$ when $X = 0$.
- Note: If zero is not in or near the range of values observed for X, then the intercept has no interpretation (i.e. it represents an extrapolation).
- Example: Predicting the selling price of a house (Y) based on square footage (X)
  - It would not make sense to have house with 0 square feet, in this case the intercept would not have an interpretation.

Sampling Distributions & Inference

- We can calculate the standard errors of the coefficients: b_0, b_1, in order to construct confidence intervals for them.
- We can use *t*-tests to determine whether the intercept or slope is significantly different from zero
  - This is only interesting for the case of the slope, since it amounts to a test for *whether Y is significantly linearly related to X*
- **Idea**: To check whether the data really is drawn from an Assumed Linear Model (so that we can make inferences) we must make sure the assumptions of the error term are satisfied
  - Fatal to the model: the variance is not constant, or the errors are not normally distributed

Method of least -Squares

- To find "good" estimators of the regression parameters β_0 and β_1, we employ the method of least squares.
- The method of least squares considers the deviation of *Y*i from its expected value
- In particular, the method of least squares requires that we consider the sum of the *n* squared deviations.
- This criterion is denoted by Q:

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \Rightarrow \quad Q = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Least Squares Estimation of β_0, β_1

- β_0 → Mean response when *X* = 0 (*y*-intercept)
- β_1 → Change in mean response when *X* increases by 1 unit (slope)
- β_0, β_1 are unknown parameters (like μ)
- $\beta_0+\beta_1 x$ → Mean response when explanatory variable takes on the value *x*
- Goal: Choose values (estimates) that minimize the sum of squared errors (*SSE*) of observed values to the straight-line:

$$Q = SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \big(y_i - (b_0 + b_1 x_i)\big)^2$$

Residuals of Linear Model

- A residual (AKA prediction error) is the difference between the observed value (y) and the predicted value $(\hat{y})$ for each obs.

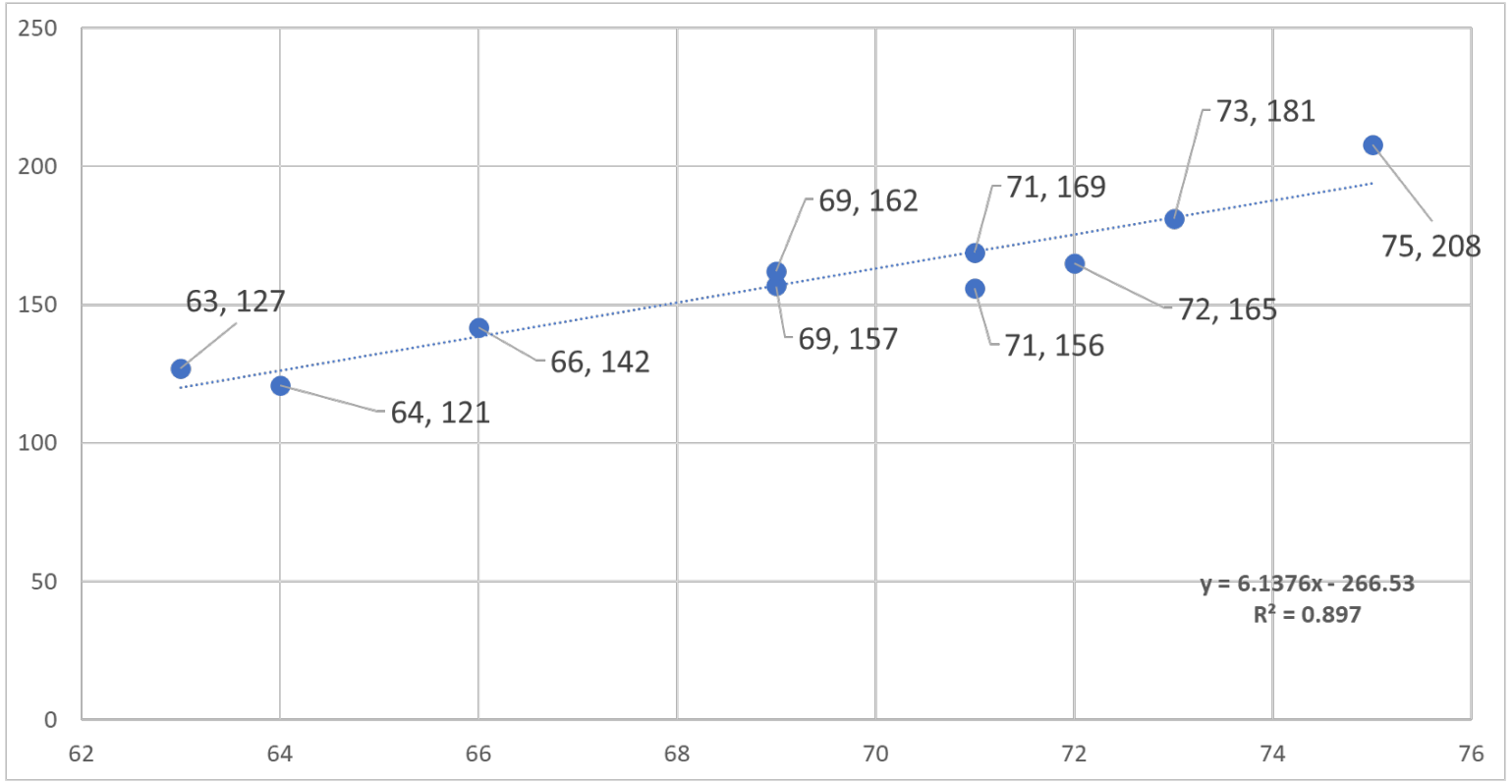
$$e_i = y_i - \hat{y}_i$$

- A small residual means that the response variable is close to what it was predicted to be under the fitted regression model
  - The point on the scatterplot is close to the fitted line
- A large residual means that the response variable was unexpectedly high or low when compared to the fitted regression model
- Intuitively we would like our model to generate residuals that are as small as possible.

More about the best fitting line: minimizing the sum of squared errors

If we did know the value for observation 5 the equation line would predict it to be 157. The actual value for observation 5 is 162, hence the prediction error is 162 – 157 = 5. **What is the general equation for the error?**

i	x_i	y_i	$\hat{y}_i$
1	63	127	120.1
2	64	121	126.3
3	66	142	138.5
4	69	157	157.0
5	69	162	157.0
6	71	156	169.2
7	71	169	169.2
8	72	165	175.4
9	73	181	181.5
10	75	208	193.8



Note: begin to practice visualizing data, you should be able to make this chart in excel, jasp or spss

Regression Model

- The deviation of an observation is calculated around its estimated mean, hence the deviations are the residuals

$$e_i = Y_i - \widehat{Y}_i$$

$$SSE = \sum (Y_i - \widehat{Y}_i)^2 = \sum e_i^{\,2}$$ Residual sum of squares

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum (Y_i - \widehat{Y}_i)^2}{n-2} = \frac{\sum e_i^{\,2}}{n-2}$$ Mean square error

Summary

- In general, when we use $\hat{y}_i = b_0 + b_1 x_i$ to predict the actual response y_i we make a prediction, or residual error, equal to

$$\epsilon_i = y_i - \hat{y}_i$$ Prediction error for data point *i*

- The best fitting line makes all prediction errors as small as possible
- The best fitting line minimizes the sum of the squared prediction errors → the least squares criterion
  - **Note**: why do we need to square the prediction errors?

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
127	120	6.9
121	126	-5.3
142	139	3.5
157	157	0
162	157	5
156	169	-13.2
169	169	-0.2
165	175	-10.4
181	182	-0.5
208	194	14.2
		Sum = 0

More about the best fitting line

- To minimize the SSE, we find the values for b_0, b_1 which make the sum of the squared prediction errors the smallest they can be. So, we find these values so that the following equation is minimized

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \Rightarrow \quad Q = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

- Since $\hat{y}_i = b_0 + b_1 x_i$

$$e_i = y_i - \hat{y}_i$$

Estimation of Error Terms Variance, σ^2

- The variance of the error terms needs to be estimated to obtain an indication of the variability of the probability distributions of *Y*
- We know the variance of a <u>population</u> is estimated by the sample variance, s^2

$$s^2 = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$$

Sum of squares

df

Mean square

- The logic of developing an estimator of the variance for the regression model is the same as for sampling from a single population that we've seen previously (i.e., standard error)

The variance of deviations

- The variance of the deviation (ε_i) is denoted σ^2 and is estimated using the average squared residual that we have seen before!

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SSE}{n-2} = MSE$$

Note: to make inferences about the model parameters (i.e., hypothesis test) we also need to assume that the deviations ε_i are normally distributed

Analysis of Variance for SLR

- A measure of ‘goodness of fit’
- SSE is the "error sum of squares" and quantifies how much the data points, y_i, vary around the estimated regression line, $\hat{y}_i$.
  - SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- SSR is the "regression sum of squares" and quantifies how far the estimated sloped regression line, $\hat{y}_i$, is from the horizontal "no relationship line," the sample mean or $\bar{y}$.
  - SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- SSTO is the "total sum of squares" and quantifies how much the data points, y_i, vary around their mean, $\bar{y}$.
  - SSTO = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Note: SST0 = SSR + SSE

Analysis of Variance in Regression

- **Goal**: partition the total variation in *Y* into variation "explained" by X and random variation
- The total variation in an observed response about its mean can be written as a sum of two parts – (a) its *deviation from the fitted value* plus (b) the deviation of the fitted value from the mean response

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$(y_i - \bar{y})^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2$$

SST SSE SSR

Note: go back to last week's lecture introducing SLR

The Estimated Coefficients

The regression equation that estimates the first order linear model is:

$$\hat{Y} = b_0 + b_1 X$$

To calculate the estimates of the line coefficients that minimize the differences between the data points and the line, use the formulas:

$$b_1 = \frac{\text{cov}(X,Y)}{s_X^2} \left(= \frac{s_{XY}}{s_X^2} \right)$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Last week: Recall the calculation of *r*

$$\hat{r} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerator of covariance

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerators of variance

Relation between *r* and b_1

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

$$r\frac{S_y}{S_x} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} \times \frac{S_y}{S_x}$$

Multiply the correlation coefficient by the standard deviation of *y divided by the standard deviation of x*

$$r \times \frac{S_y}{S_x} = b_1$$

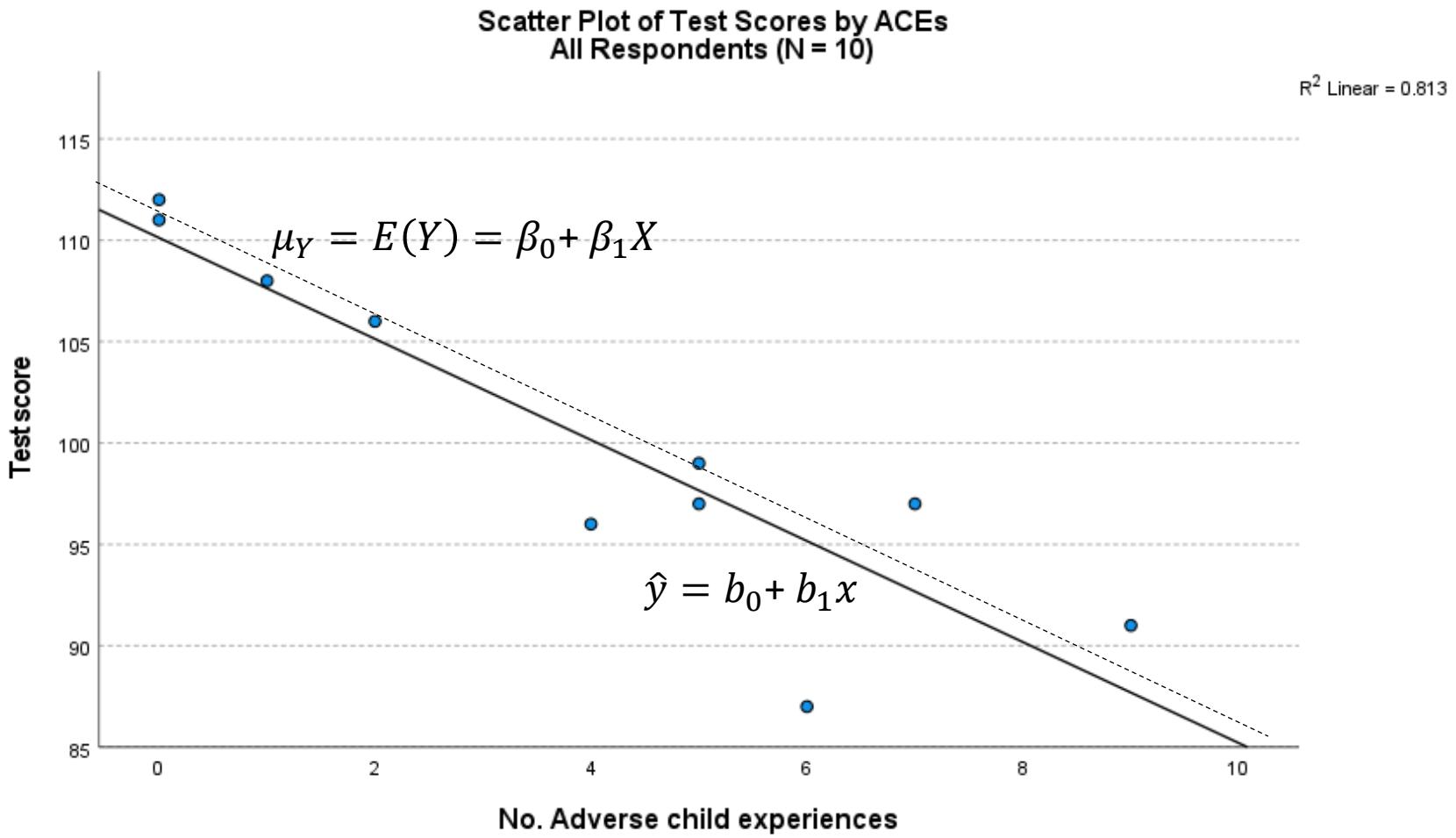
Do Lab Questions #1,2

- At home: replicate the LSD and math score example at the end of this lecture

What do the regression parameters estimate?

Example from simple-linear-regression-annotated-mechanics.docx

Before moving on make sure you understand the example



Solid line = sample

Dotted line = population

Linear Regression Mechanics

Assumptions

Seven key assumptions of linear regression

- Linear in the predictors: the mean response for *Y*, at each value of the predictor, x_i, is a linear function of the x_i
- Independent: the errors are independent
- Normally distributed: the errors at each value of the predictor x_i are normally distributed
- Equal variances (σ^2): the errors at each value of the predictor xi have equal variances
  - The notation for this is as follows: $\varepsilon \sim iid\ N(0, \sigma^2)$
- Model specification – the model should be properly specified (including all relevant variables, and excluding irrelevant variables)
- Fixed *X*
- Noncollinearity

Four testable model assumptions

$$y = \beta_0 + \beta_1 x + \varepsilon$$

1. Linearity: Linear association between the dependent variable and the independent variable
   - The mean values of response variables fall on a straight line as a function of the explanatory variable

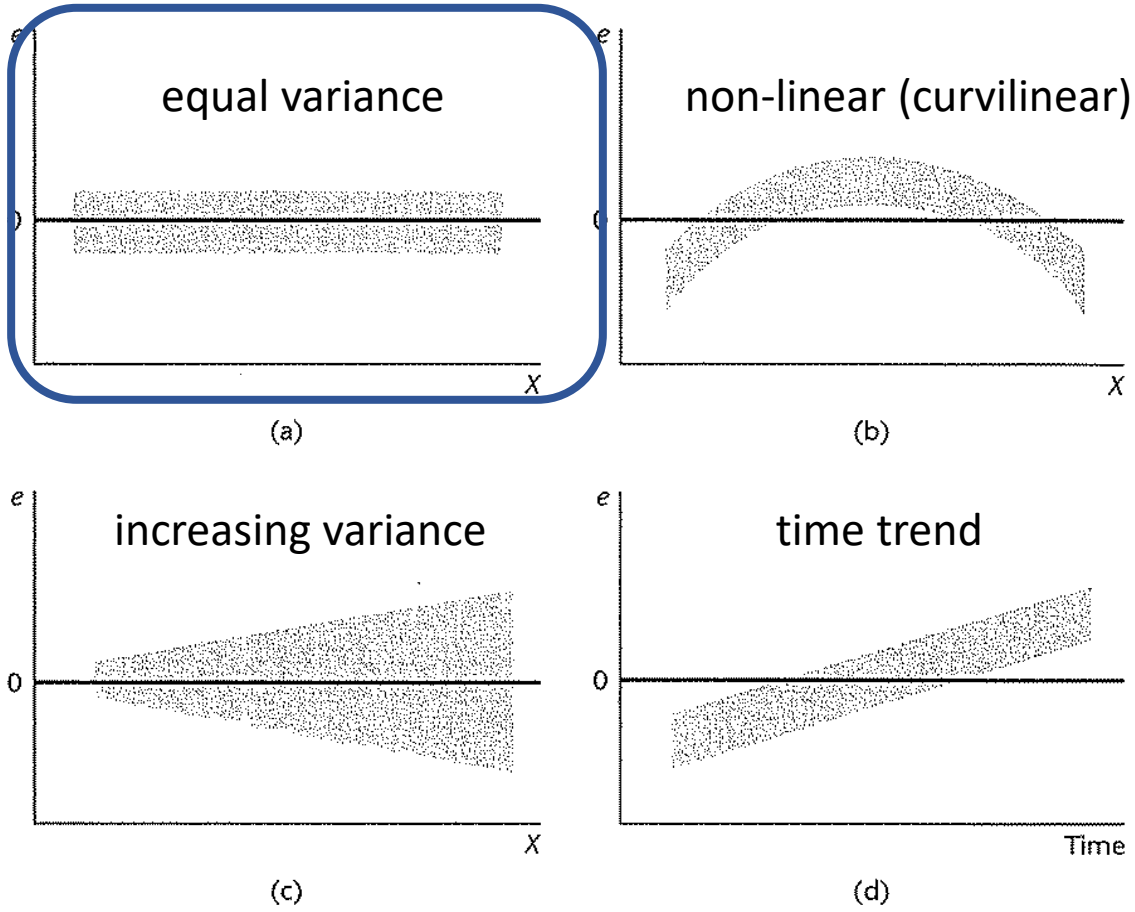
2. Normality: Normality of the distribution of errors i.e. normality of the residuals (pp plot, qqplot of residuals is diagonal)

3. Independence: Independence of the errors around the regression line between the actual and predicted values of the response variable

Four testable model assumptions

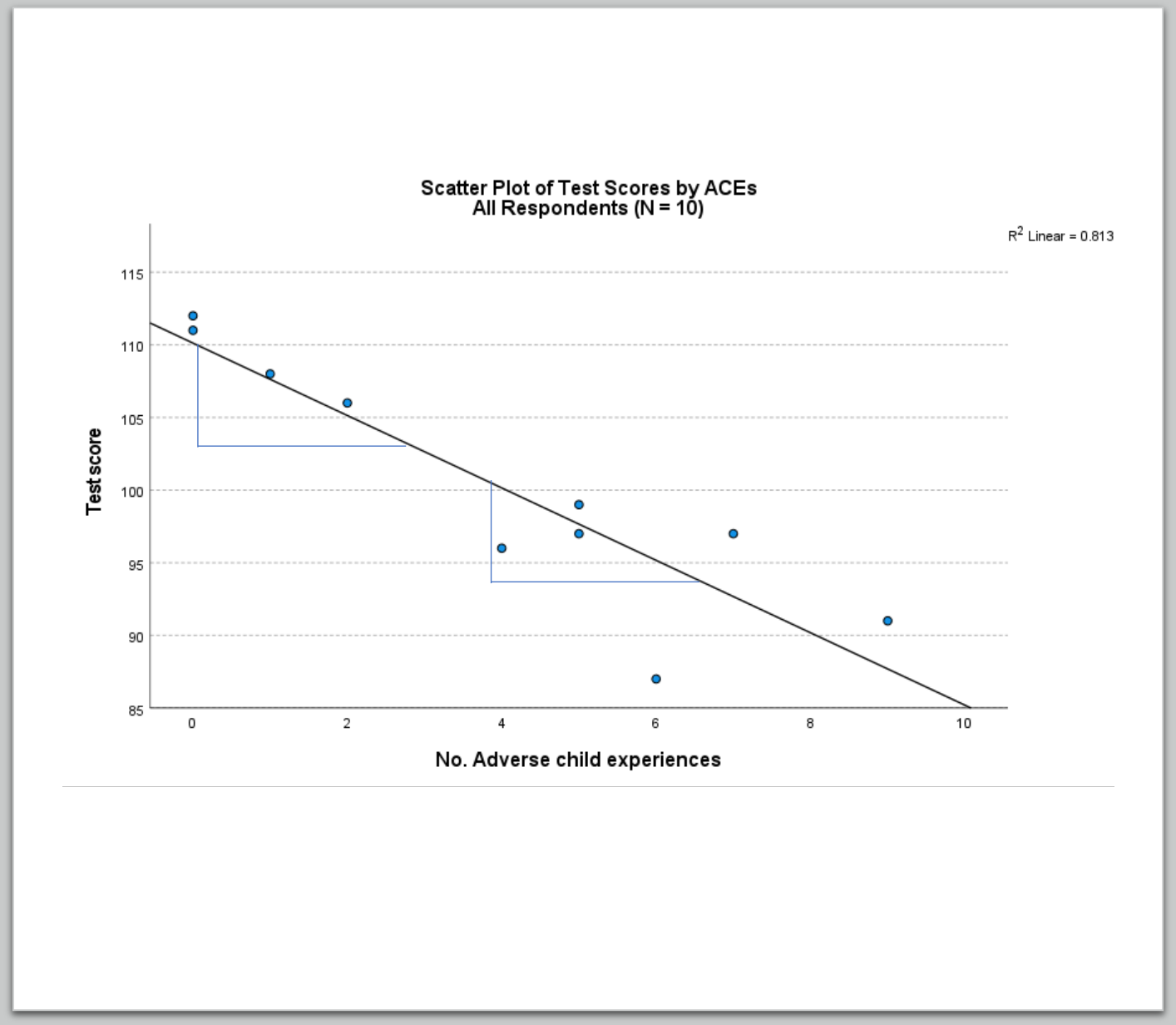
$$y = \beta_0 + \beta_1 x + \varepsilon$$

4. Constant variance: Equal variance of the distribution of the response variable for each level of the independent variable. (homoscedasticity, the violation is called heteroskedasticity)
   - The spread of the response around the straight line is same for all levels of explanatory variable
   - Errors have constant variance



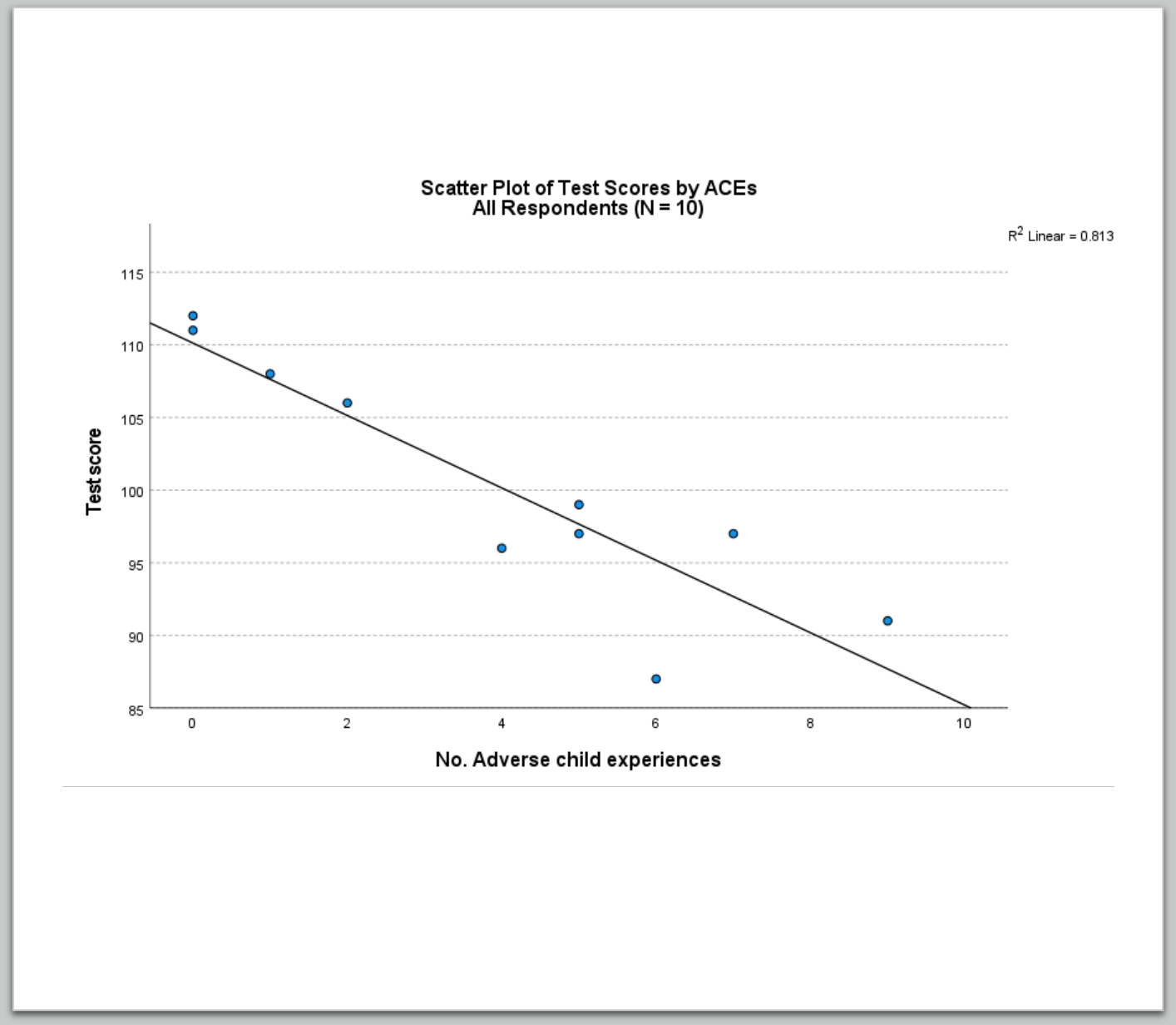
Linearity

- Notice that we used a line to summarize the relationship between ACEs and test score. *Is it reasonable to assume that the mean test scores are linearly related to the number of ACEs?* Why/why not?
  - Before answering let's be clear about what linearity means. It means that every additional ACEs has the same effect on test scores, an effect that is quantified by the slope of the regression
  - Quantified by the interpretation of the slope ?



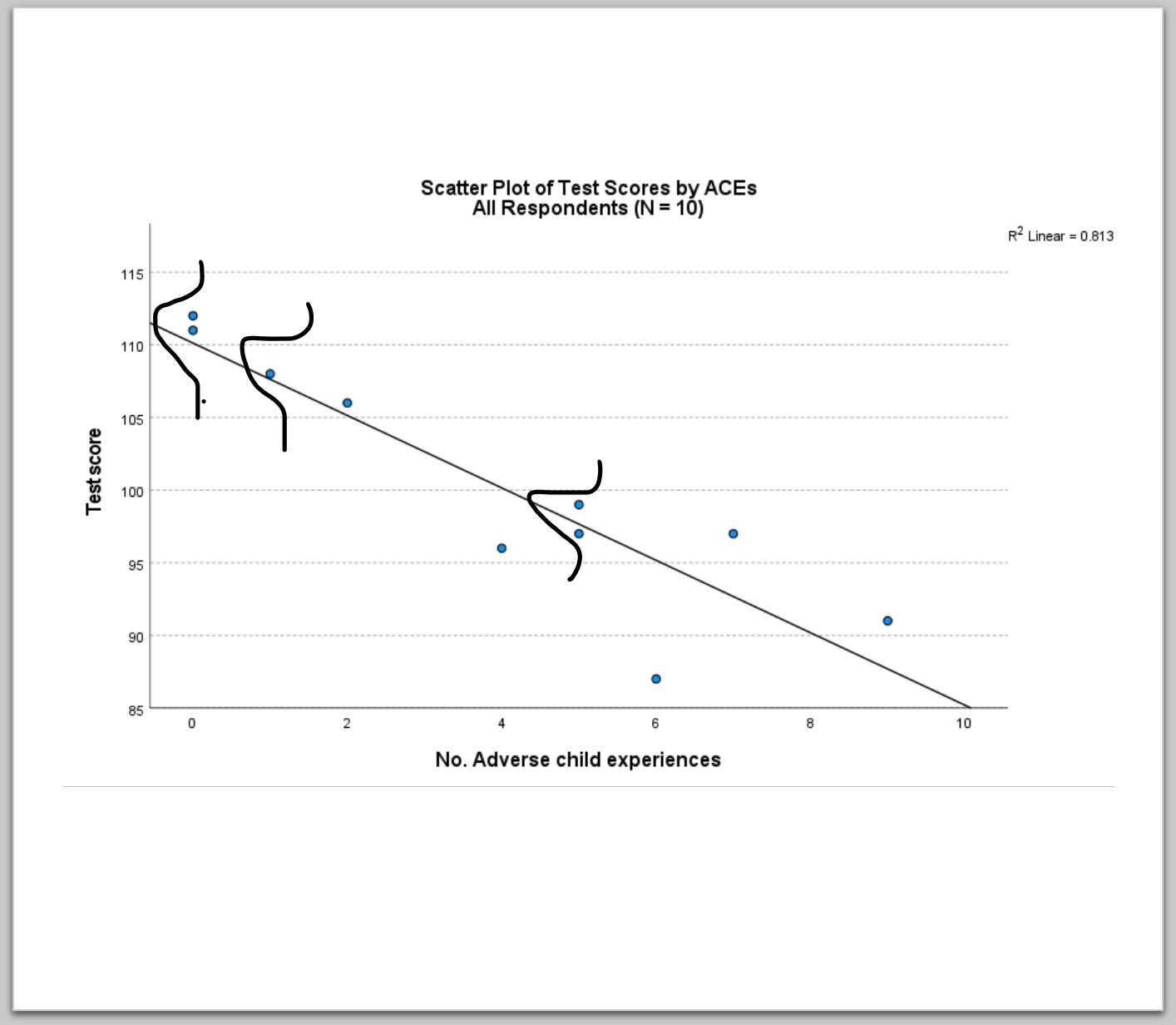
Normality of ***error*** term

- Try to visualize the spread of the errors in this figure. Does it seem reasonable to assume that the errors for each subpopulation are normally distributed? Why/Why not?
- Before answering: notice that some of the errors are very small and some are much larger, is the distribution equal across all possible combinations of ACEs and scores?



Independence

- Does it seem reasonable to assume that the error for one child's test score is independent of the error *for another child's test score*? Why/Why not?
- <u>Before answering:</u> think of situations when the assumption does not hold, i.e., when would one child's score be dependent on another child's score?

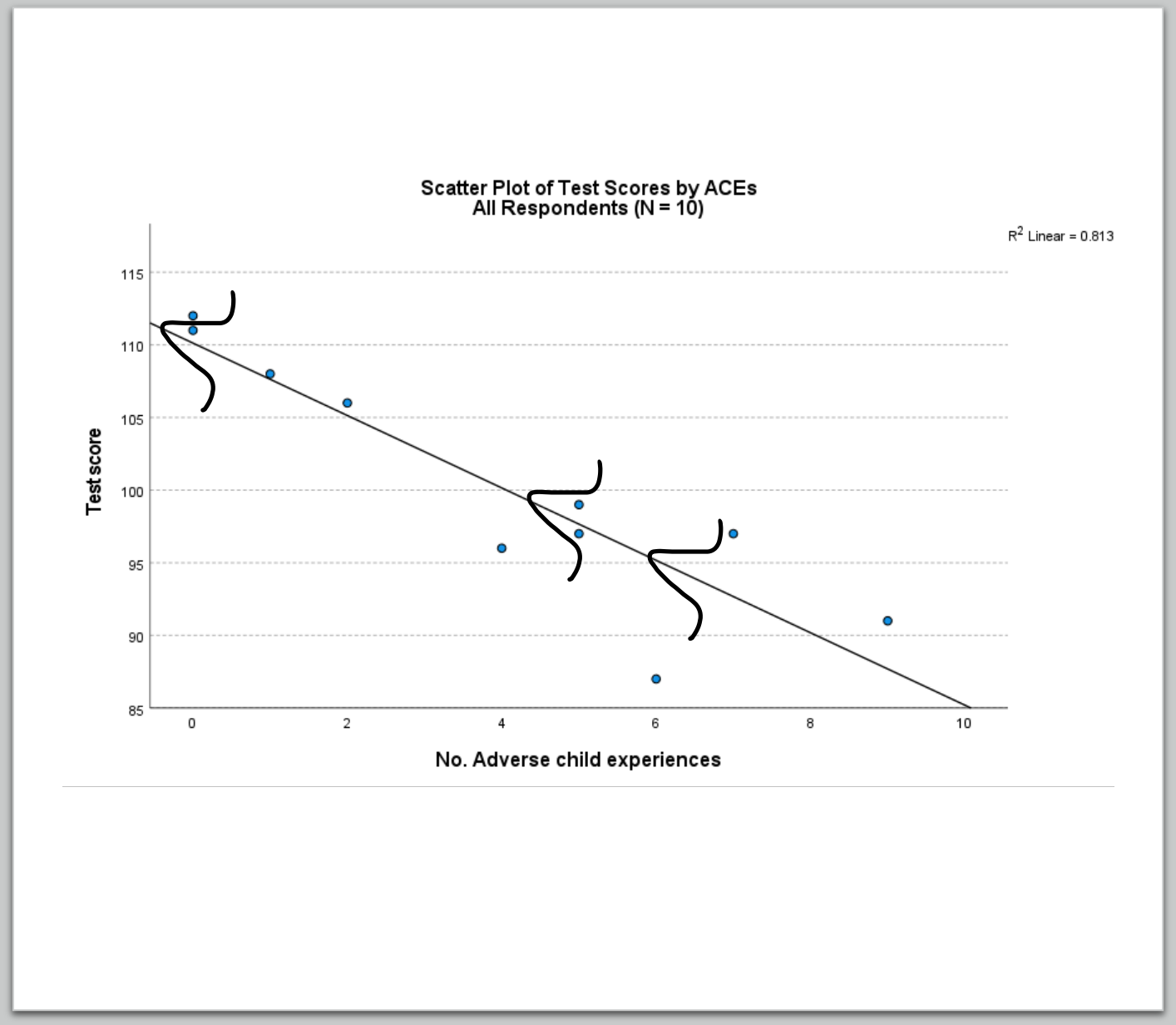


Independence

- Due to sampling scheme
- **Example**: I just reviewed a paper that was exploring the impact of county-level correlates and aggregate county-level perceptions of social workers on the outcomes of youth in FC
- The authors used OLS
- This is incorrect, why?

Questions to consider: homoskedasticity

- Try to visualize the spread of the errors in this figure. Does it seem reasonable to assume that the spread of the prediction errors for each subpopulation are equal no matter the number of ACEs? Why/Why not?
- <u>Before answering:</u> think of a situation where the variation in ACEs would be different at each value of ACEs.



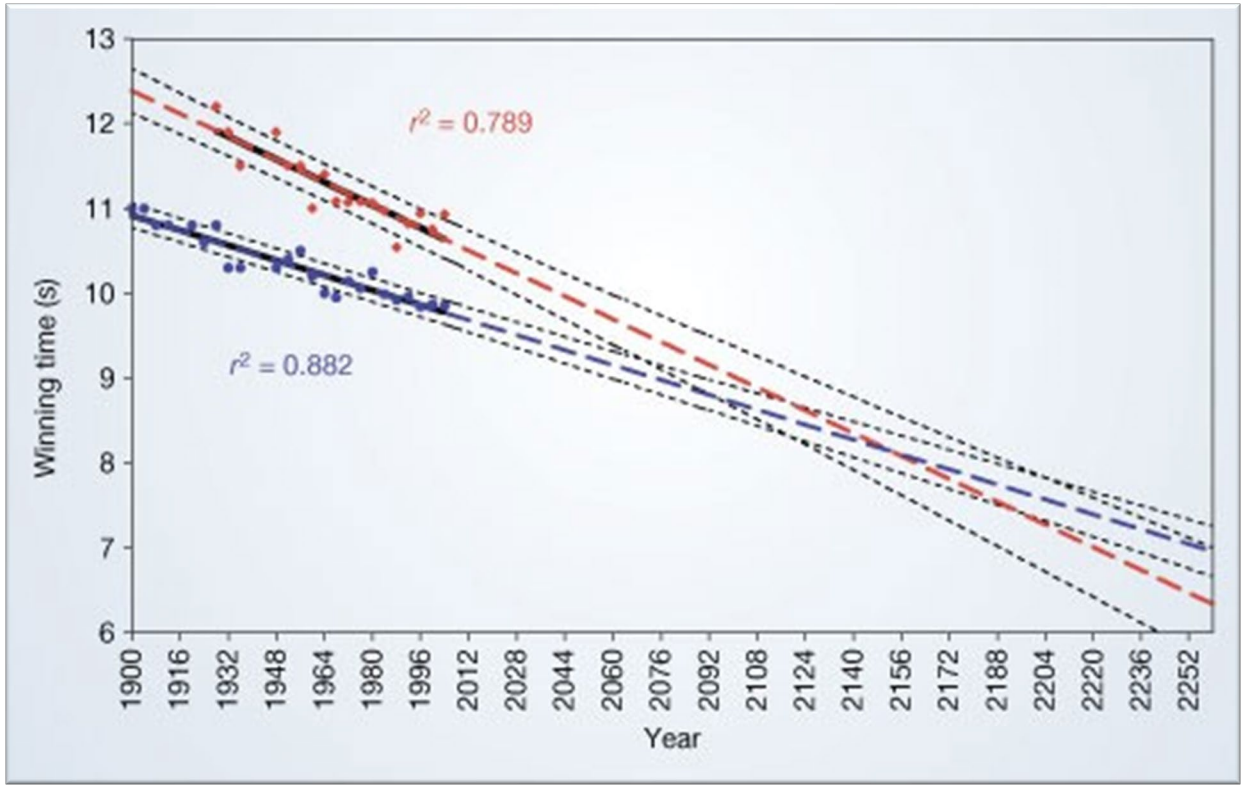
Non-testable assumptions

Use common sense

Fixed X

- Analyzed athletes' performance on the 100 m dash
- Found an interesting pattern: the time it took to run it decreased steadily, such that males and females were getting faster and faster over the years
- Around the year 2156, the lines crossed: sometime mid-century, they predicted, women would outrun men!
- Even though the lines LOOK convincing, they are not truly capturing the underlying pattern of the data… **why**?

The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.



Tatem, A. J., Guerra, C. A., Atkinson, P. M., & Hay, S. I. (2004). Momentous sprint at the 2156 Olympics?. Nature, 431(7008), 525-525.

Noncollinearity & model specification

- Less is more
- Always examine correlations – substantial overlap means one should be taken out of the model
- But be weary – I learned that the forward/backwards elimination is not a good strategy, why?
- Start with theory → test hypotheses → draw conclusions

The Coefficient of Determination R^2

- The question is: how much of the total variation in the response *Y* is due to the regression of *Y* on *X* as opposed to random error?
- **EXAMPLE**: SSE = 1708.5; SSR = 6679.3; SSTO = 8487.8
- We can see that most of the total variation (SSTO) is accounted for by the regression (SSR). That is, $\frac{SSR}{SSTO} = .799$

$$R^2 = \frac{SSR}{SSTO} \rightarrow 1 - \frac{SSE}{SSTO}$$

Explained Not Explained

Interpretation R^2

- Since R^2 is a proportion, it is always a number between 0 and 1.
  - If R^2 = 1, all of the data points fall perfectly on the regression line. The predictor *X* accounts for *all* of the variations in *Y*!
  - If R^2 = 0, the estimated regression line is perfectly horizontal. The predictor *X* accounts for *none* of the variations in *Y*!

R^2 × 100 percent of the variation in *Y* is reduced by taking into account predictor *X*

or

R^2 × 100 percent of the variation in *Y* is explained by *X*

Pearson Correlation, *r* & R^2!!

- So, we have another interpretation of *r* – the square root of the coefficient of determination
- The correlation coefficient, *r,* is directly related to the coefficient of determination R^2
  - If R^2 is represented in decimal form, *e.g.* 0.39 or 0.87, then all we have to do to obtain *r* is to take the square root of R^2 : $r = \pm\sqrt{R^2}$
- The sign of *r* depends on the sign of the estimated slope coefficient b_1 :
  - If b_1 is negative, then *r* takes a negative sign.
  - If b_1 is positive, then *r* takes a positive sign.
- That is, the estimated slope and the correlation coefficient *r* always share the same sign. Furthermore, because R^2 is always a number between 0 and 1, the correlation coefficient *r* is always a number between -1 and 1.

Note

- If you know R^2 and SST we can calculate

$$SSR = R^2 SST$$

and

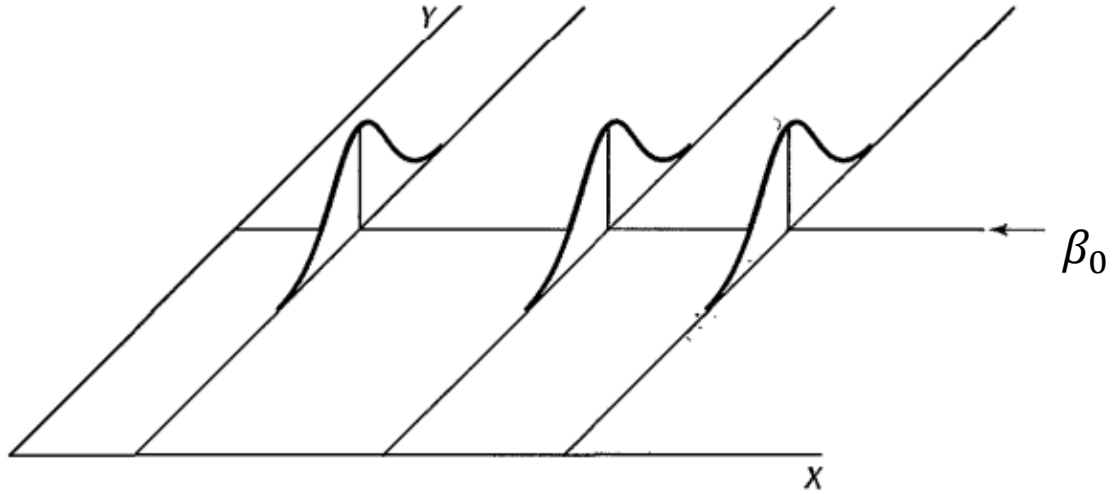
$$SSE = (1 - R^2) SST$$

Inferences for point estimates

- We are interested in drawing inferences about the slope of the regression line
- The most common hypothesis to test is:

$$H_0: \beta_1 = 0 \text{ vs. } H_0: \beta_1 \neq 0$$

- The reason for testing this is that when $\beta_1 = 0$ there is no linear association between *Y* and *X*



Mean(Y) $= \beta_0 + 0X = \beta_0$

Hypothesis testing of coefficients

- We follow standard hypothesis testing for the slope parameters in a linear regression model
- First, specify the null and alternative
  - H_o: $\beta_1 = 0$
  - H_1: $\beta_1 \neq 0$
- Second, calculate the test statistic
  - $t^* = \frac{b_1 - \beta_1}{\sqrt{MSE} \Big/ \sqrt{(x_i - \bar{x})^2}} = \frac{b_1 - \beta_1}{se(b_1)}$
- Calculate the p-value**
  - The p-value is determined by referring to a t-distribution with *n* – 2 *df***
- Decision: If *p-value* < α ⇒ reject H_o

Confidence intervals as hypothesis testing

- It is also useful to look at the confidence interval for β_1
  - As always, the CI is the sample estimate $\pm$ (t-multiplier x standard error*)*
  - $b_1 \pm t_{(\alpha/2,\, n-2)} \times se(b_1)$ → $b_1 \pm 1.96 \times se(b_1)$
- The CI provides the following information:
  - It gives us a range of values that is likely to contain the true unknown value of β_1 → we are 95% certain that the true population parameter lies within the confidence interval; and
  - It gives us the answer to whether there is a significant relationship between X and Y → if 0 lies in the confidence interval then we know to reject the null hypothesis. Why?

Critical: Inference, Error and Associations

We do not reject the null hypothesis, $H_o : \beta_1 = 0$

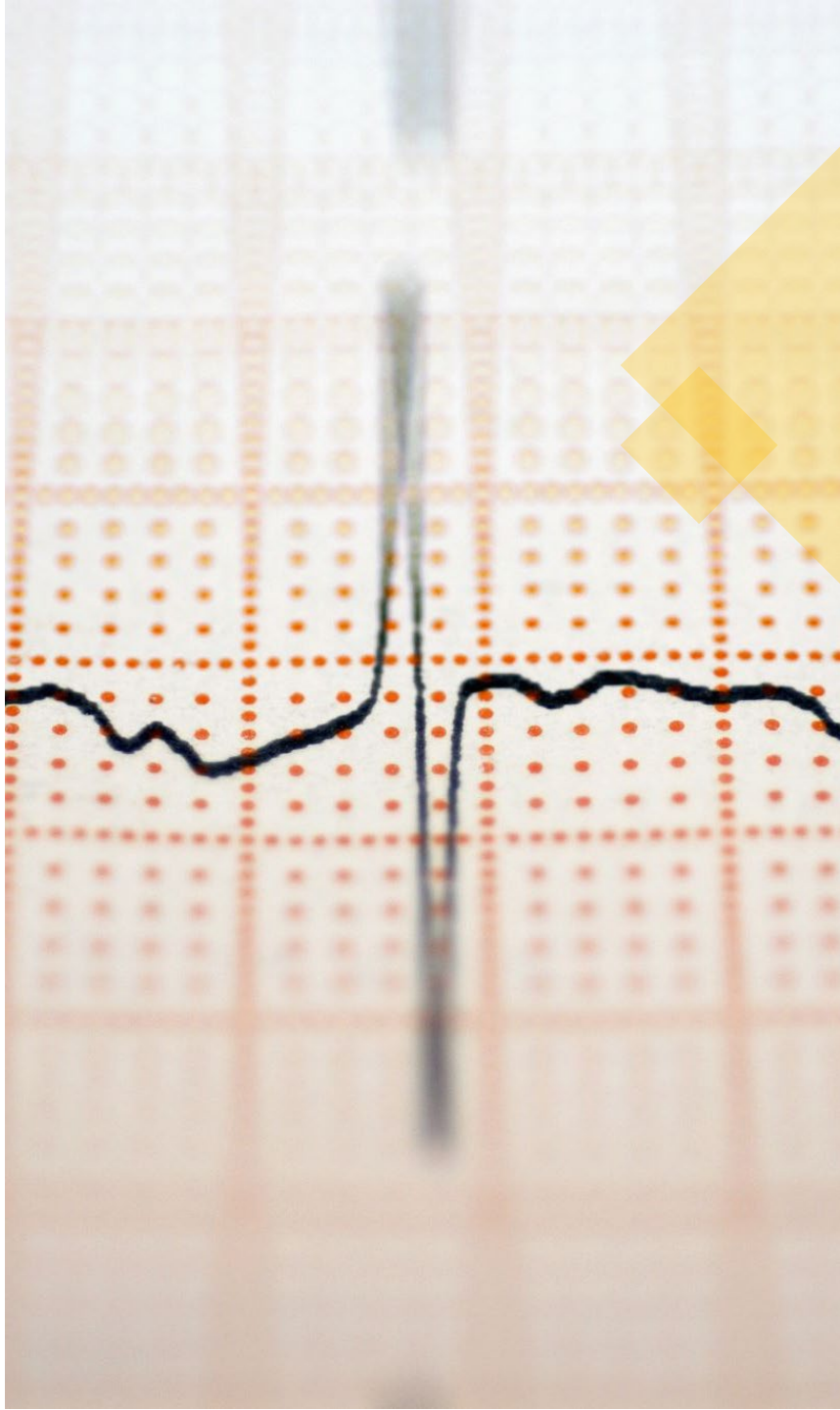
- We committed a Type II error
- There really is not much of a linear relationship between *x* and *y*
- There is a relationship between *x* and *y* — it is just not linear

We do reject the null hypothesis, $H_1 : \beta_1 \neq 0$

- We committed a Type I error
- The relationship between *x* and *y* is indeed linear.
- A linear function fits the data, okay, but a curved ("curvilinear") function would fit the data even better (i.e., the model is mis-specified)

Factors affecting the width of the CI:
Making the interval as narrow as possible

- As the confidence level decreases, the width of the interval decreases
  - Compare $b_1 \pm 1.96 \times se(b_1)$ with $b_1 \pm 1.56 \times se(b_1)$
  - Note: the wider interval makes it less likely we reject the null that $\beta_1 = 0$, i.e. that 0 will be contained in the CI
- As the regression error (MSE) decreases, the width of the interval decreases
  - Less error means we are more confident about rejecting the null
  - Look at formula as MSE increases, *t*-statistic *decreases*
- The more spread out the predictor x values, the narrower the interval
  - Look at formula, more variation means *t*-statistic *increases* because the denominator will be smaller
- As the sample size increases, the width of the interval decreases, i.e. it will be more narrow
  - Larger *n* means more variation



Model Evaluation

The four conditions tell us what can go wrong

- The population regression function is nonlinear
- The error term is not independent
- The error terms are not normally distributed
- The error terms do not have equal variance

Other major issues

- Presence of outliers
- Bias due to an omitted variable (i.e., omitted variable bias)

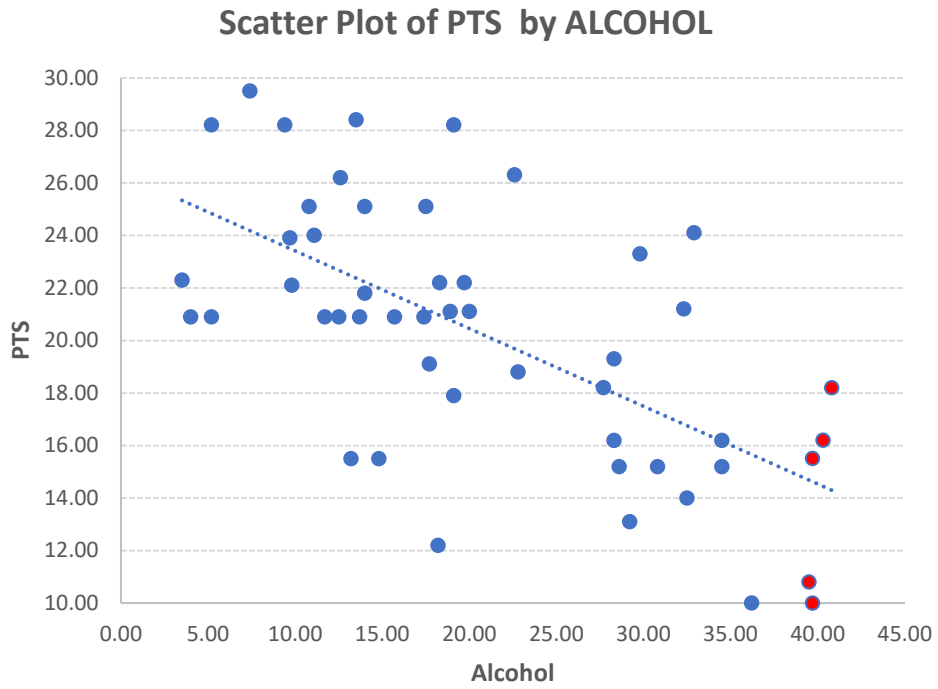
Reinforce the importance of these assumptions

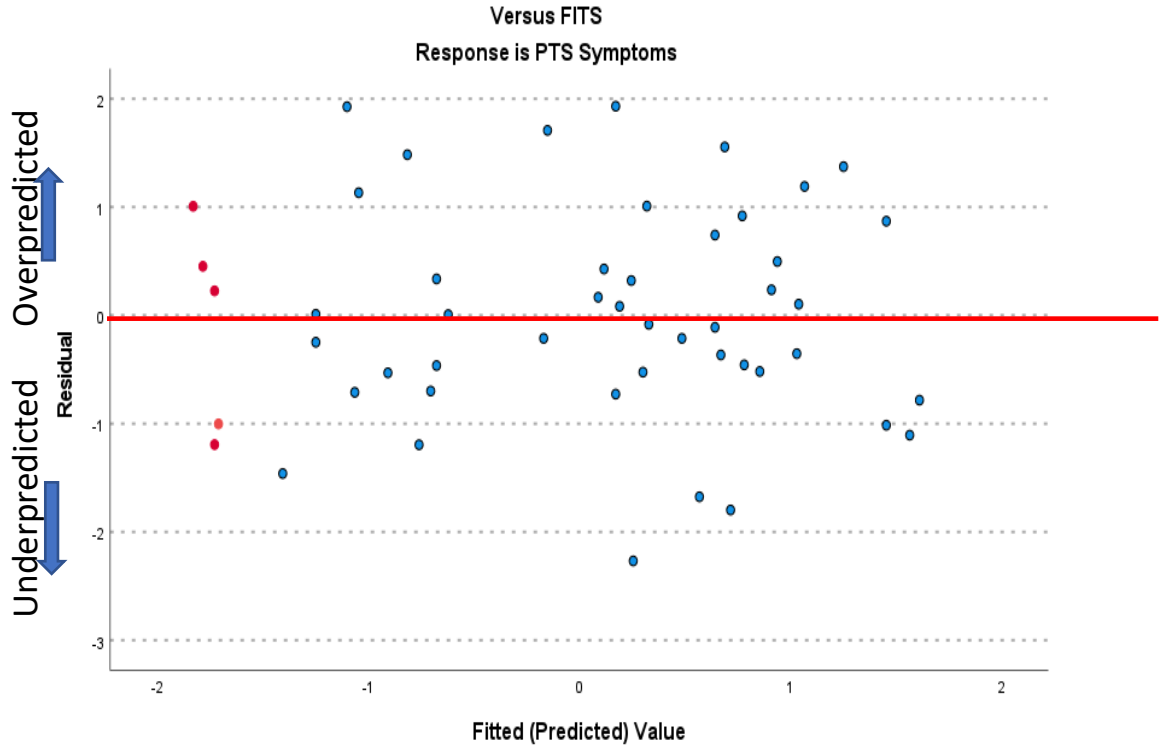
- In OLS the tests we perform are extremely sensitive to model assumptions

Residual v Fits Plot

- The bivariate plot of the predicted value against residuals can help us infer whether the relationships of the predictors to the outcome is linear and for homogeneity of variance.
- The plot consists of residuals on the y-axis and fitted values on the x-axis
- Here is what you are looking for:
  - The residuals "bounce randomly" around the residual = 0 line. This suggests that the assumption that the relationship is linear is reasonable.
  - The residuals roughly form a "horizontal band" around the residual = 0 line. This suggests that the variances of the error terms are equal.
  - No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.

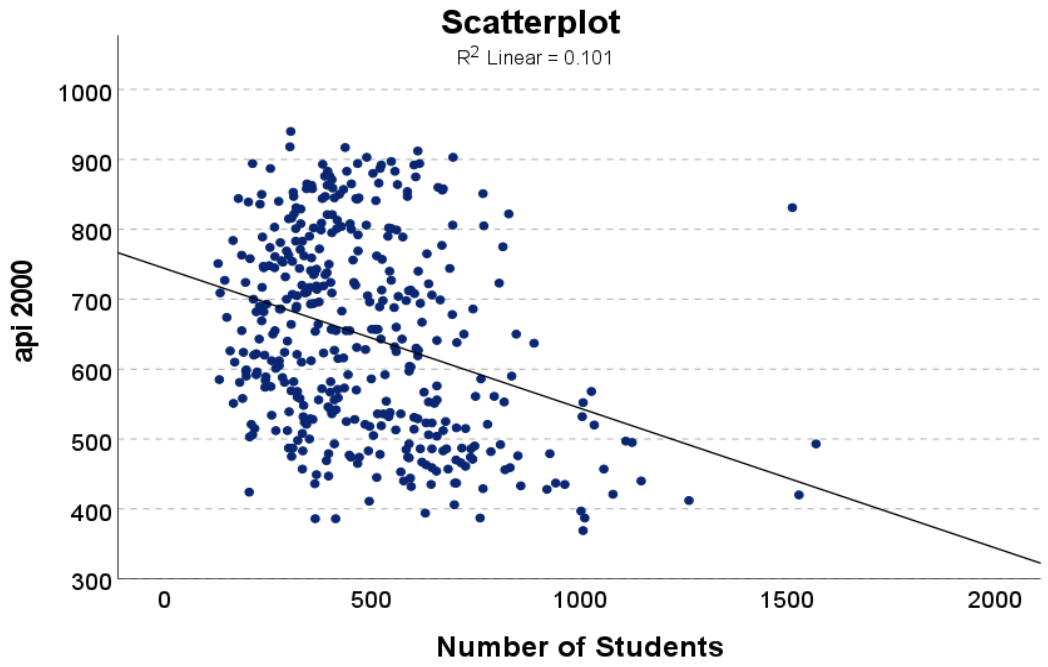
Example: Linearity

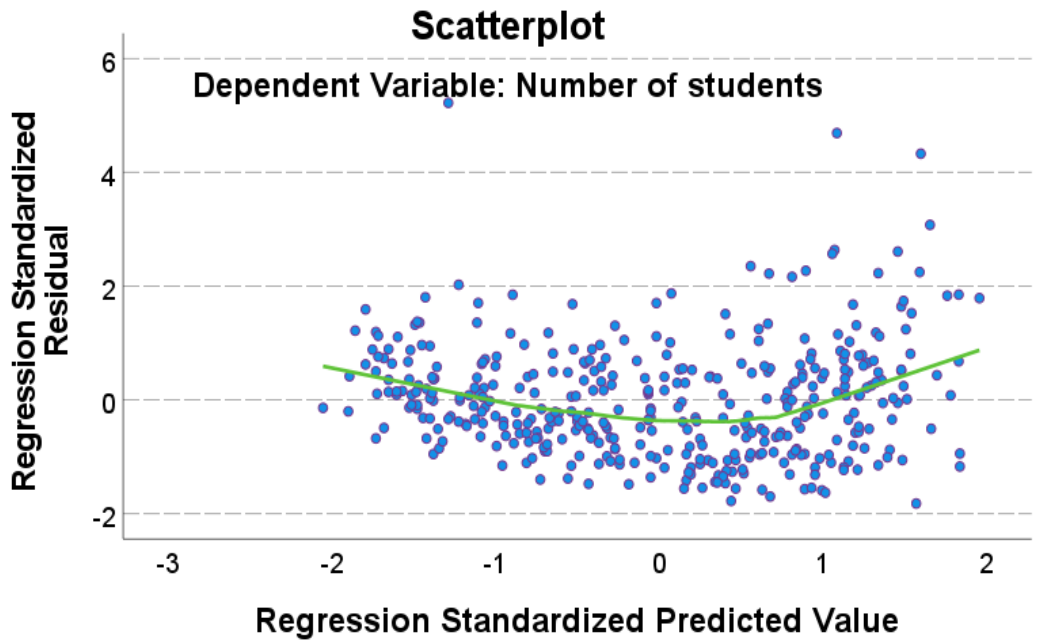




Next obvious question: What does a bad plot look like?

API2000 (academic performance) v Enroll (number of students)

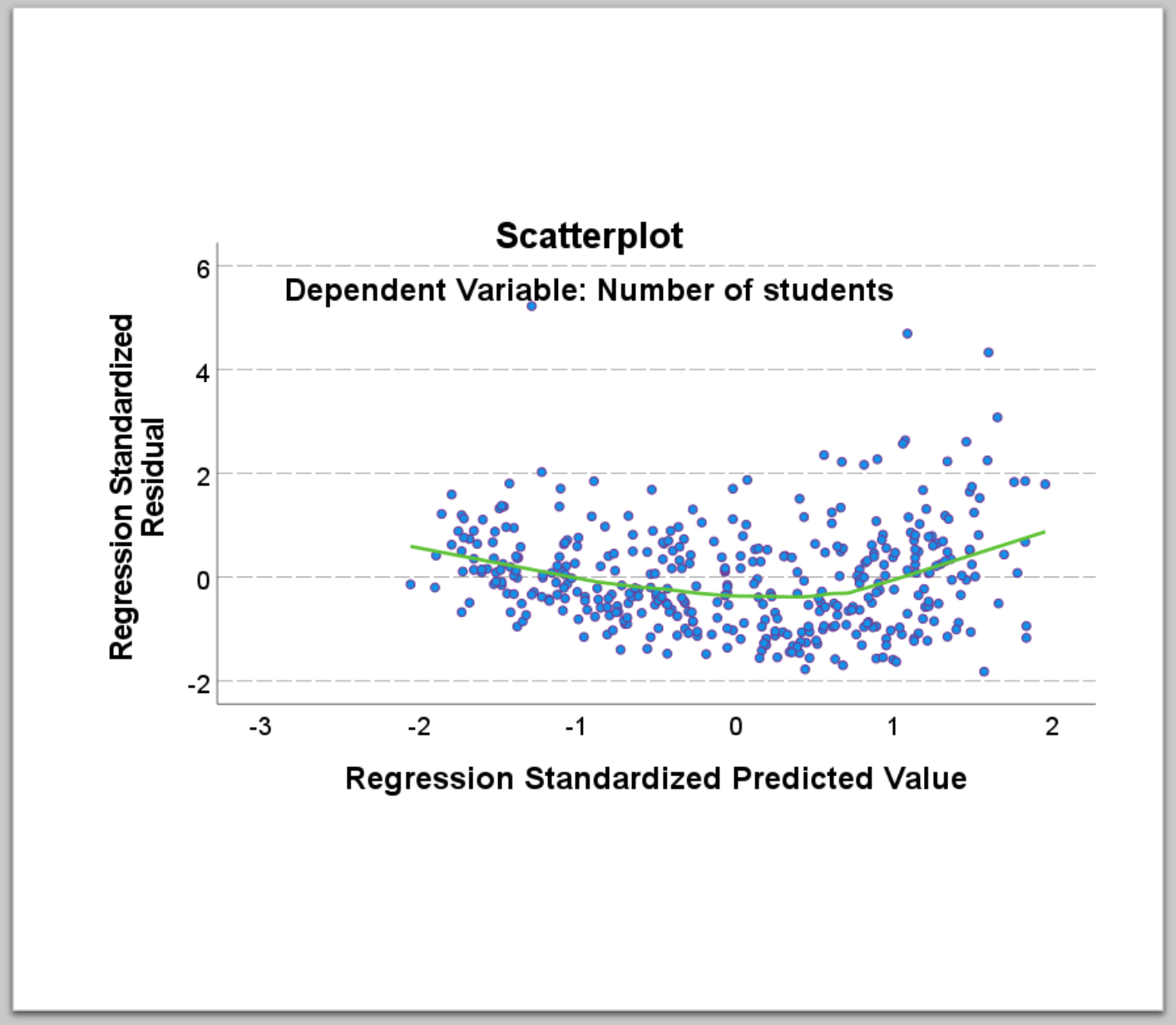




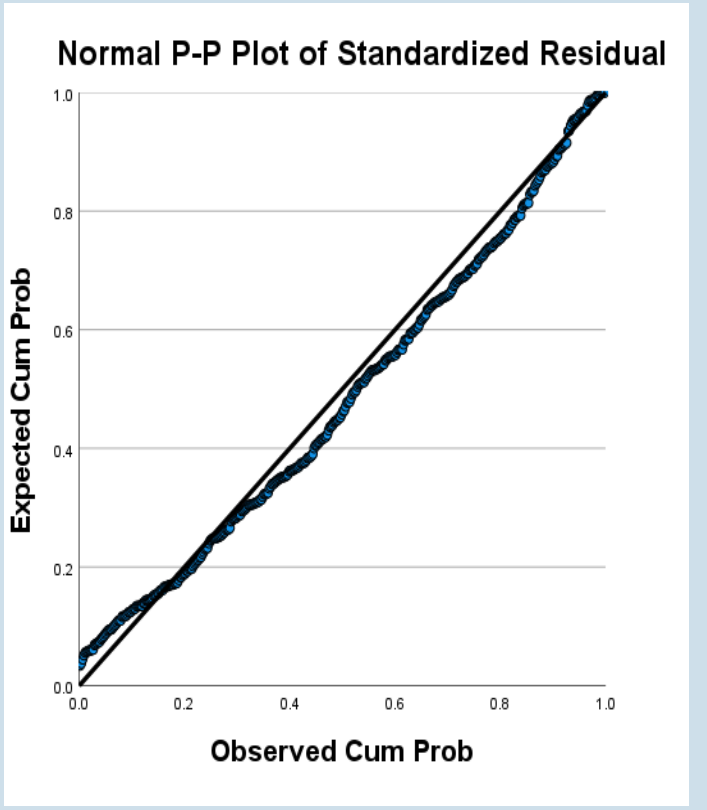
badplot_elemapi2v2.sav

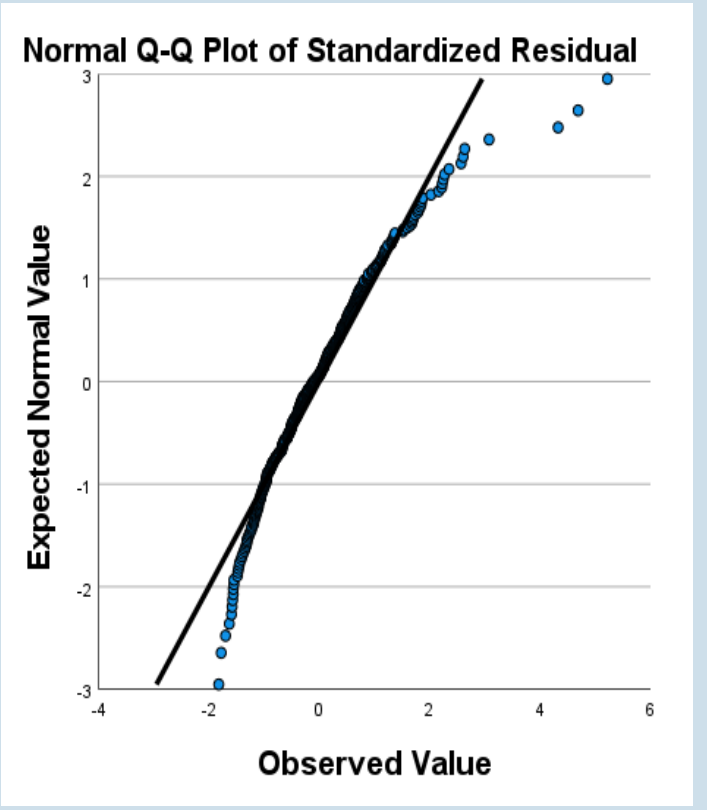
Homogeneity of Variance

- The variance of the residuals is homogeneous across levels of the predicted values, also known as homoscedasticity
- If the model is well-fitted, there should be **no pattern** to the residuals plotted against the fitted values
- If we look carefully at the plot, we see that the lower values of the standardized predicted values tend to have lower variance around zero.



Tests on Normality of Residuals

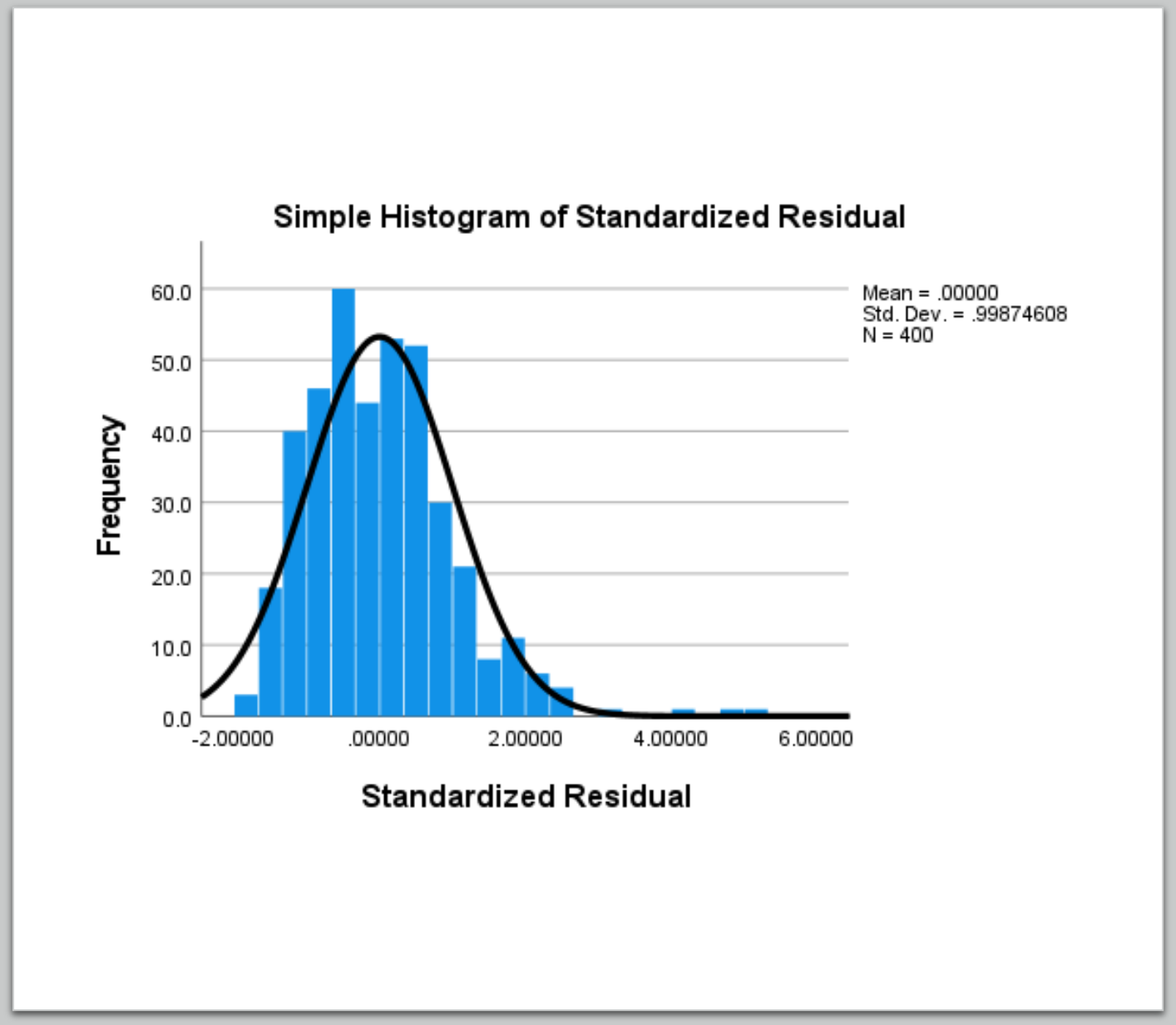




- In linear regression, a common misconception is that the outcome has to be normally distributed, but the assumption is actually that **the residuals are normally distributed**
- A normal probability plot (P-P plot) helps us assess normality of the errors
- A normal Q-Q performs the same function (both SPSS and JASP have this option)

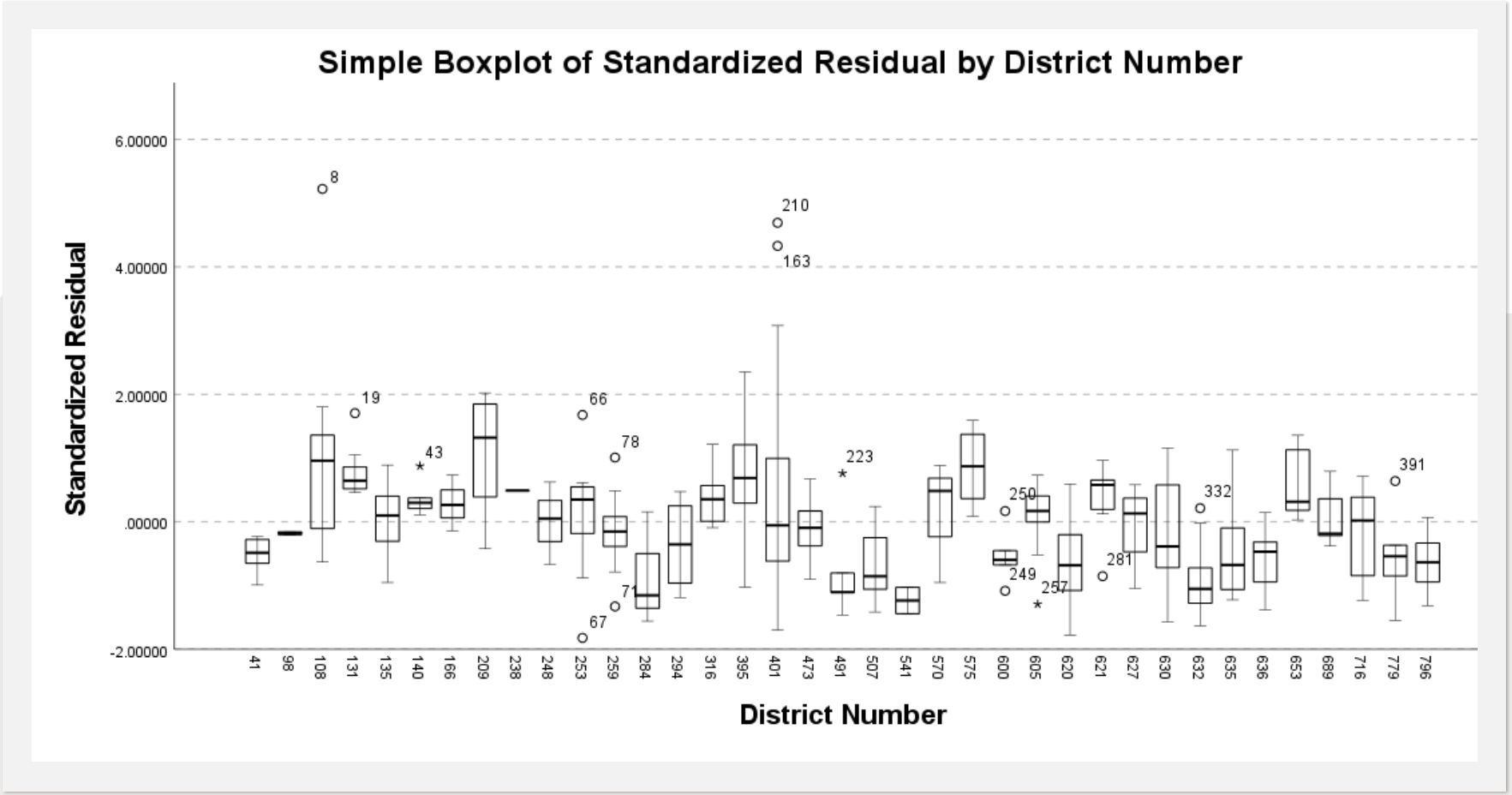
Histogram of Standardized Residuals

- You can also get a sense of normality of errors by saving the standardized residuals and then creating a histogram plot with a normal overlay
- This is the same result as the normal probability plot, showing a slightly right skew to the data



Note that these tests are model dependent meaning that this can change if we add more predictors.

Note also that two additional model fit measures, omitted variables (model specification) and multicollinearity, are applicable to multiple linear regression and we will cover these hopefully later today



Independence of Observations

Issues of Independence

- This is driven to a large degree by the method of data collection
- Therefore, violations can occur in a variety of situations
- Consider data on child welfare allegations in a particular state
  - Why might these data violate the independence assumption
- If you have clustered data, you need to incorporate the clustering into your model
- You can get a general sense about violations of independence via descriptive statistics

Do Review, Questions #3

Introduction to Multiple Linear Regression

1. Introduction to Multiple Linear Regression
2. Example of linear regression with two features
3. Multiple Regression in R
4. Simple Regression → Multiple Regression
5. Adjusted R^2
6. Multicollinearity
7. Examples

Key Questions

1. Is the overall model (regression equation) useful? i.e. are any of the fitted coeff. significant?
2. Which features are “significant” and what does that mean?
3. Are the various features positively or negatively related to the response? How do we interpret them?
4. How much of the variability in the response variable is explained by the model?
5. Do the assumptions of the OLS Regression Model hold?

Topic: Multiple Linear Regression

- **Data**: (Y_i, $X_{i,1}$, $X_{i,2}$,… ,$X_{i,p}$) i.e. there are now *p* variables for each observation, *i* remains the number of cases
- **Linear Model** is now:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_p + \varepsilon, \text{ where } \varepsilon \sim \text{Norm}(0,\sigma^2)$$

- **Goal**: To fit a Linear Regression Model with coefficients b_0, …, b_p:

$$\hat{Y} = b_0 + b_1 x_1, \ldots, + b_p x_p$$

- Note that the fitting is still done with Ordinary Least Squares (OLS) Method just like in simple linear regression.

$$Q = min_{b_0,\, b_1,\ldots,b_k} \sum_i (y_i - b_0 + b_1 x_1, \ldots, + b_p x_p)^2$$

Key differences between MLR and SLR

- F-test (ANOVA table): Tells us whether *any* indep. vars are significant
  - H_0: $\beta_i = 0$ for all *i* or H_0: $\beta_1 = \beta_2 = \beta_3 \ldots \beta_p = 0$ (None of the X variables have linear relationships with Y)
  - H_1: $\beta_i \neq 0$ for some *i* (At least one X variable has a linear relationship with Y)
- Intercept: β_0 is the expected value of Y when *all* predictors are zero
- Slope(s): β_i is the expected change in response (Y) for every 1 unit increase in X_i, *while holding all other predictors constant*.
  - Individual t-tests for each coefficient are reported in summary
- R^2: Proportion of variance in Y that is explained by *all* independent vars
  - Use **adjusted R-sq** because it adjusts for number of predictors (same interpretation) as opposed to regular R^2 which always increases in # predictors

Example: Regression with two independent/explanatory variables

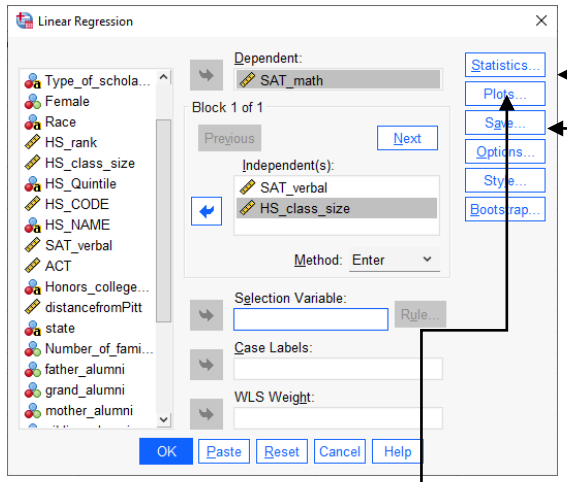
- **Question**: How is math SAT score related to verbal SAT score and class size?
  - "Regress math SAT score on verbal SAT score and class size"
  - Intuition: higher verbal scores should mean higher math scores because students who work hard at math probably work hard at other courses
  - Intuition: bigger class size means lower math SAT scores because students have less opportunity to interact with teachers, etc.
  - Intuition: Also though, verbal SAT scores should mean larger classes as well, for the same reason.
- Let's fit a linear model to find out the relation (if any!).

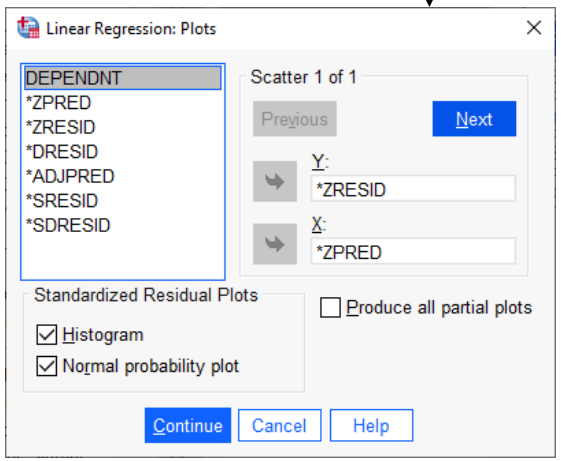
Variable definitions in the school admissions dataset: admissions.sav

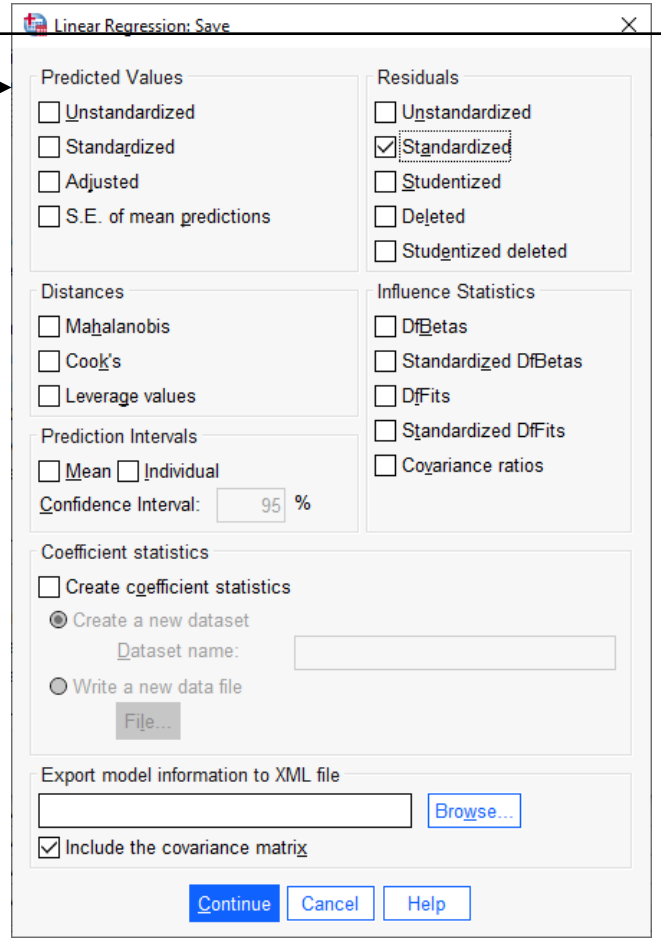
Column Name	Variable Definition
row_number	row number from original dataset
paiddeposit	1 if paid deposit, 0 if not
scholarship_yes_no	1 if student offered scholarship, 0 if not
Type_of_scholarship_offered	type of scholarship offered if applicable
Female	1 if female student, 0 if male student
Race	student race
HS_rank	high school rank
HS_class_size	high school class size
HS_Quintile	high school quintile
HS_CODE	high school code
HS_NAME	high school name
SAT_math	SAT math score
SAT_verbal	SAT verbal score

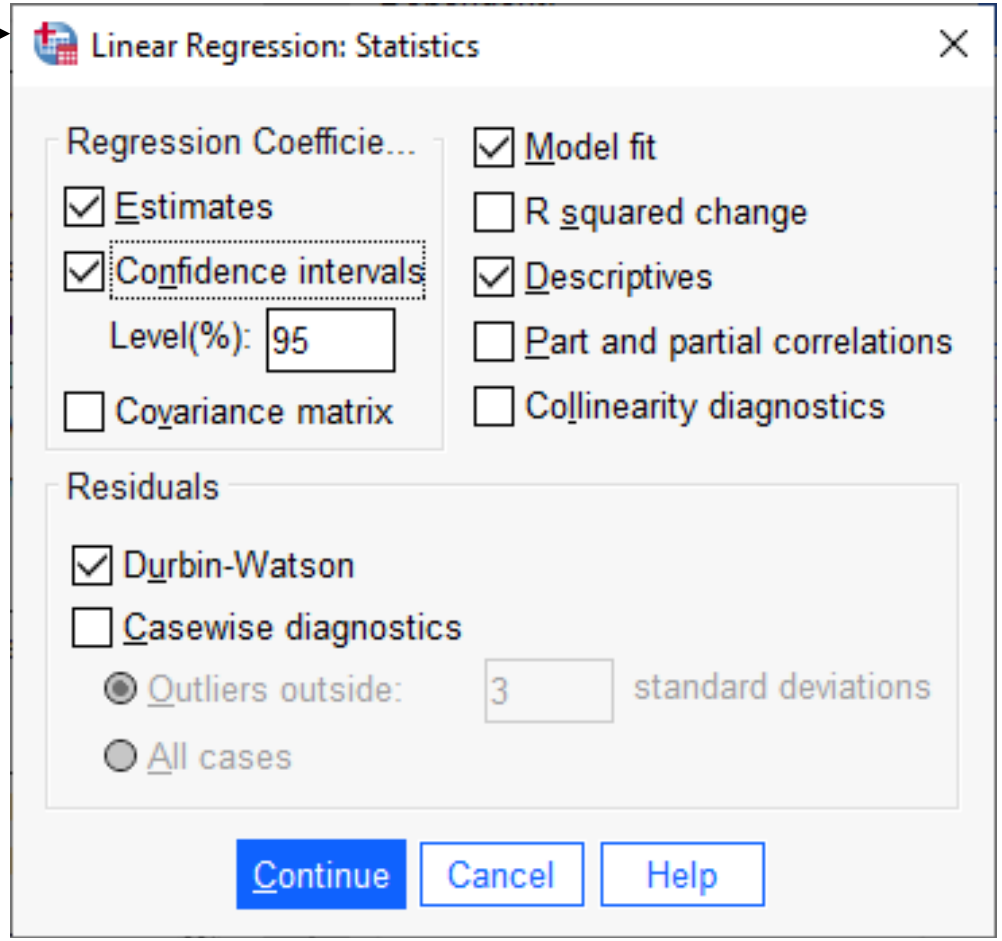
...

Analyze → Regression









What you already know…Descriptive Statistics & Relationships

Descriptive Statistics

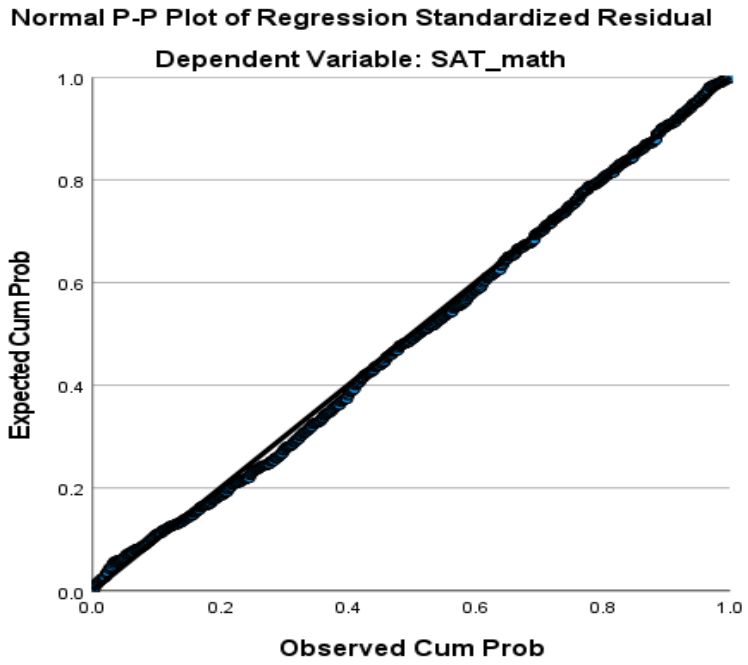
	Mean	Std. Deviation	N
SAT_math	585.007	68.7157	691
SAT_verbal	561.216	67.8124	691
HS_class_size	316.33	178.043	691

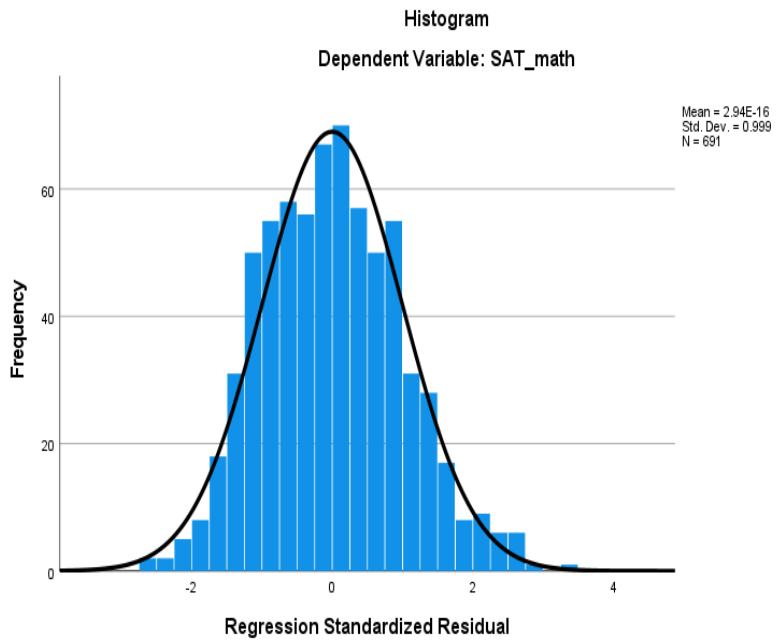
Note: the correlation between SAT math score and SAT verbal score is .450.

Correlations

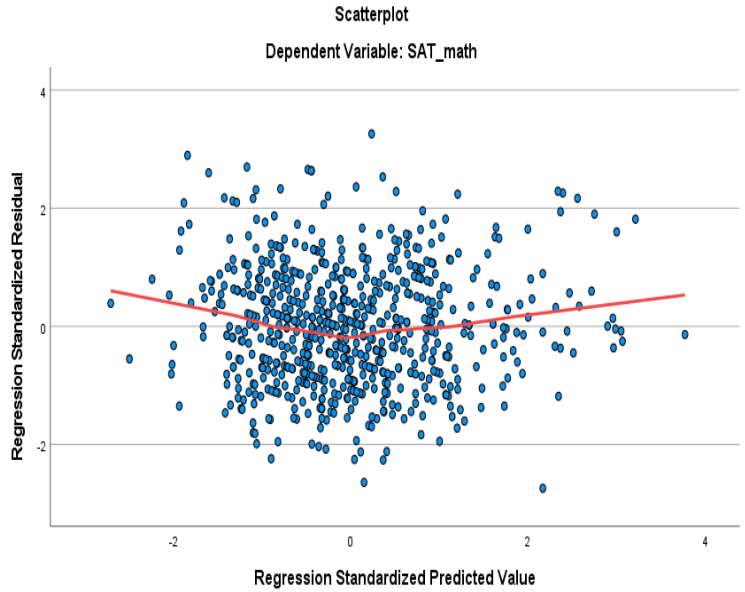
		SAT_math	SAT_verbal	HS_class_size
Pearson Correlation	SAT_math	1.000	.450	.176
	SAT_verbal	.450	1.000	.045
	HS_class_size	.176	.045	1.000
Sig. (1-tailed)	SAT_math	.	<.001	<.001
	SAT_verbal	.000	.	.120
	HS_class_size	.000	.120	.
N	SAT_math	691	691	691
	SAT_verbal	691	691	691
	HS_class_size	691	691	691

What you already know…Diagnostics





Linear + Constant Variance – Check Residuals v Fitted



Normality of Errors – qqplot of residuals

- Data is very close to line that represents a normal distribution

➔ Normality assumption satisfied

No funnel shape (increasing variance)
Not much of a curvilinear pattern
➔ Linearity and Equal Variance assumptions are reasonably satisfied

What you already know…

The independent variables are moderately to highly correlated with the dependent variable $r = .476$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.476[a]	.227	.224	60.5177	2.002

a. Predictors: (Constant), HS_class_size, SAT_verbal

b. Dependent Variable: SAT_math

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	738352.691	2	369176.346	100.802	<.001[b]
	Residual	2519722.272	688	3662.387		
	Total	3258074.964	690			

a. Dependent Variable: SAT_math

b. Predictors: (Constant), HS_class_size, SAT_verbal

22.7% of the variation in MATH SAT scores is explained by the independent variables (Class size and Verbal SAT)

This means about 87% remains unexplained (an indication that there are variables omitted from the model that are important to explain MATH SAT scores)

Small *p*-value → the model is significant

What's new …

Durbin-Watson = 2 meaning
the error term *may be* independent

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.476[a]	.227	.224	60.5177	2.002

a. Predictors: (Constant), HS_class_size, SAT_verbal

b. Dependent Variable: SAT_math

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	738352.691	2	369176.346	100.802	<.001[b]
	Residual	2519722.272	688	3662.387		
	Total	3258074.964	690			

a. Dependent Variable: SAT_math

b. Predictors: (Constant), HS_class_size, SAT_verbal

The more predictors you have in the model the lower the sum of squared residuals, the higher R^2, indicating the better prediction

The Adjusted R2 accounts for the number of parameters in the model (incurs a penalty for each additional variable)

Small *p*-value → the model is significant

H_o: β_1 and $\beta_2 = 0$

H_1: at least one β is not 0

We conclude that at least one of the slopes is not equal to 0

What is new…

This is the standardized beta coefficient. This statistic puts the coefficients on the same "footing" and hence it can be used to compare the strength of the coefficients, indicating which variable is more meaningfully related to the dependent variable.

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)	314.147	19.477		16.129	<.001	275.904	352.389
	SAT_verbal	.449	.034	.443	13.192	<.001	.382	.515
	HS_class_size	.060	.013	.156	4.656	<.001	.035	.086

a. Dependent Variable: SAT_math

Recall the model is: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$ Here, $k = 2 \rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

MATH SAT Score = 314.147 + .449VERBAL_SAT(X_1) + .060CLASS_SIZE(X_2)

Confidence intervals for the coefficients

What is new…

Note: the relationship (i.e., correlation) between SAT math score and SAT verbal score is .449 *after controlling for class size*. That is, after the effect of class size is removed from the relationship between SAT and MATH score.

Notice there are now multiple independent variables in the model. This means that the effect of one variable on Y needs to be interpreted slightly different. The slope of a variable now represents the effect after all other variables in the model are controlled.

Every 1 unit increase in SAT verbal score increases the MATH verbal score by .449 *holding class size constant at its mean*

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)	314.147	19.477		16.129	<.001	275.904	352.389
	SAT_verbal	.449	.034	.443	13.192	<.001	.382	.515
	HS_class_size	.060	.013	.156	4.656	<.001	.035	.086

a. Dependent Variable: SAT_math

Recall the model is: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$ Here, $k = 2 \rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

MATH SAT Score = 314.147 + .449VERBAL_SAT(X_1) + .060CLASS_SIZE(X_2)

Partial Correlation, i.e. ‘partialling out’ AKA ‘controlling for’… (this is the relationship that holds for all values of class size)

Correlations

Control Variables			SAT_math	SAT_verbal	HS_class_size
-none-[a]	SAT_math	Correlation	1.000	.450	.176
		Significance (2-tailed)	.	<.001	<.001
		df	0	689	689
	SAT_verbal	Correlation	.450	1.000	.045
		Significance (2-tailed)	<.001	.	.240
		df	689	0	689
	HS_class_size	Correlation	.176	.045	1.000
		Significance (2-tailed)	<.001	.240	.
		df	689	689	0
HS_class_size	SAT_math	Correlation	1.000	.449	
		Significance (2-tailed)	.	<.001	
		df	0	688	
	SAT_verbal	Correlation	.449	1.000	
		Significance (2-tailed)	<.001	.	
		df	688	0	

a. Cells contain zero-order (Pearson) correlations.

What does ‘controlling for’ mean?

- Let’s say you think education is related to income
- You regress income on education
- A feminist suggests your analysis is flawed…
  - Because income is driven exclusively by gender, he thinks there will be no relationship between education and income once gender is accounted for…?
- To test this, you include gender in the model
- Two scenarios:
  - (1) education significantly predicts income controlling for gender → this means that for all levels of gender, education ‘matters’
  - (2) education does not significantly predict income controlling for gender → (?)

Multicollinearity

- As the number of predictors increases, multiple regression modeling can become very complicated
- **Multicollinearity** is when the predictors in the regression equation are correlated with each other
- Example:
  - Suppose you use both educational attainment and job type as independent variables to predict income
  - Note educational attainment and job type are highly correlated with each other! Thus having both educational attainment and job type in your multiple regression equation may only slightly improve the R^2 over an equation with just educational attainment
  - Might conclude that job type is highly influential on income, while educational attainment is unimportant (or vice versa) which is not true!

- This happened in our paper for SSRW when we included an index of redlining, concentrated disadvantage and concentrated affluence in the model
- **Note** that while there is substantial overlap between variables theoretically, they are quite distinguishable
  - The opposite of concentrated disadvantage is not concentrated advantage
  - Use theory to guide the model building!!!!!

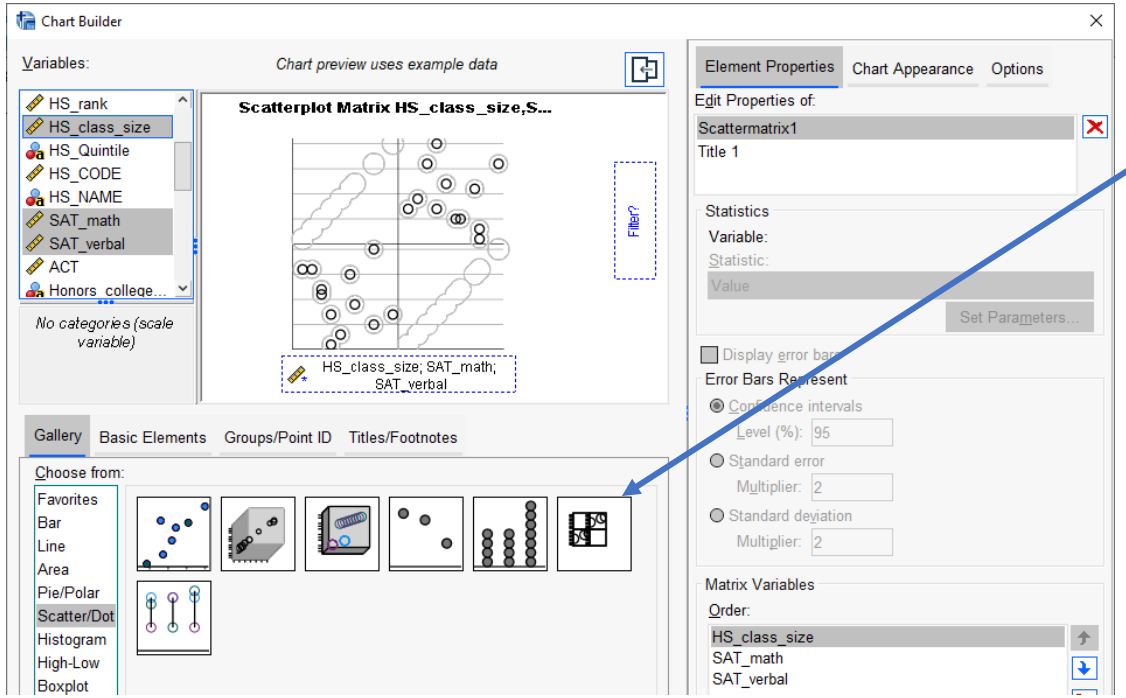
What to do if you find multicollinearity?

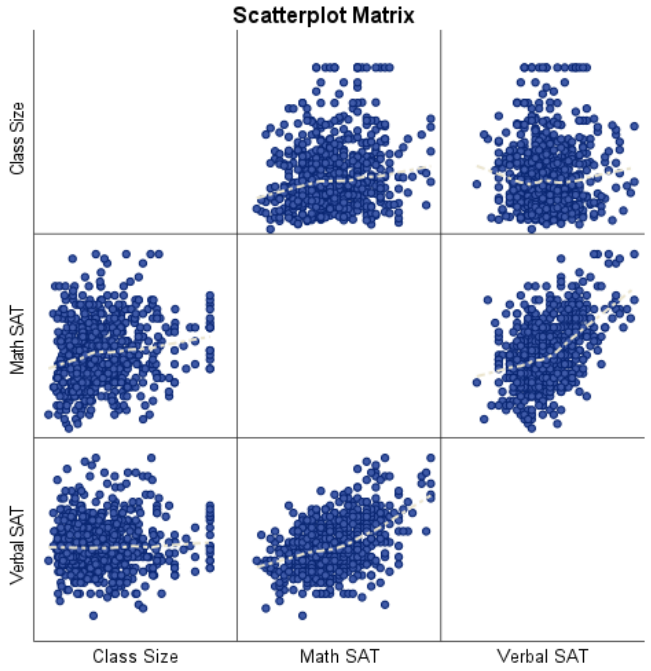
- If two variables are highly correlated, it is probably not a good idea to use both in the regression equation. Why?
  1. Many correlated variables hurts the Adjusted R^2
  2. We generally prefer smaller models
  3. Smaller models are much easier to interpret
  4. Multiple regression models describe the effect of one variable on another ***after partialling out the effects of all other variables*** in the model
     - Regarding MC, this means that if two variables are highly correlated, once one of them is partialled out, there is much less variation left in the other variable for the model to "explain"
     - Check the partial correlation coefficients and zero order coefficients before including variables!

GPA data.sav

How to check for multicollinearity?

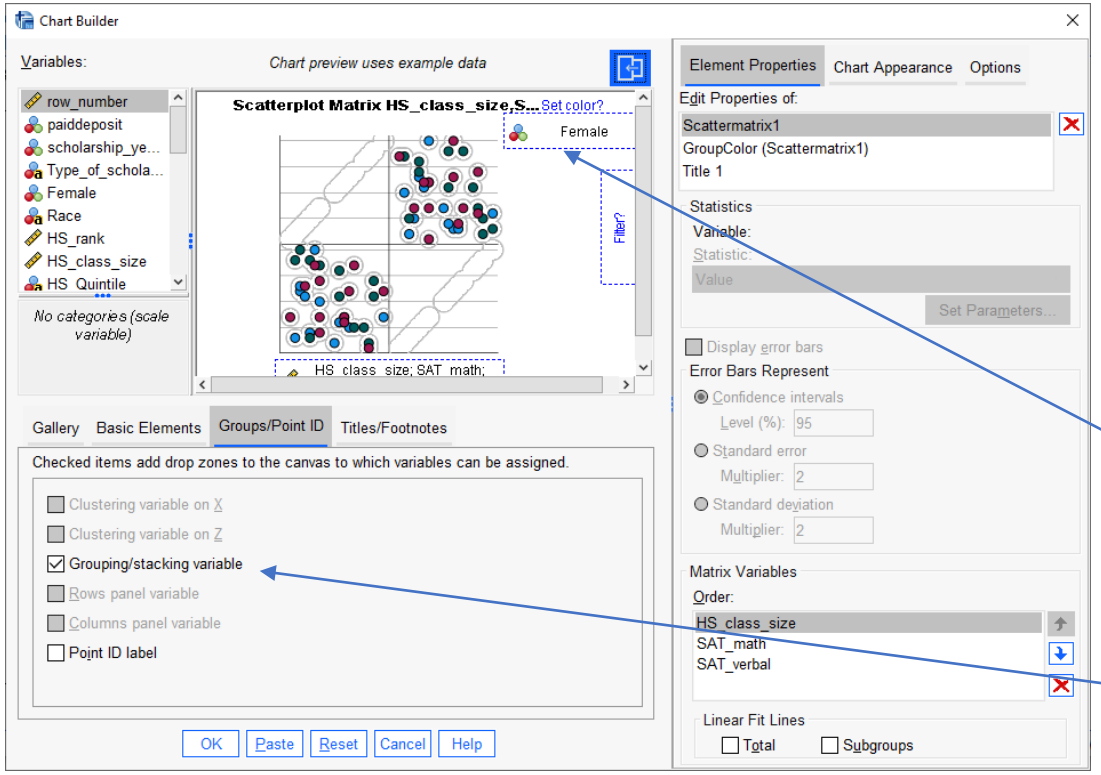
- Scatterplots, Scatterplot Matrix
  - A **scatterplot matrix** is a matrix of scatterplots, one plot for each pair of variables.
  - **Note**: To read it, pick a plot in the matrix. Look in the column for the X variable and the row for the Y variable
- Graph → Chart Builder → Scatter/Dot (drag variables into chart area)

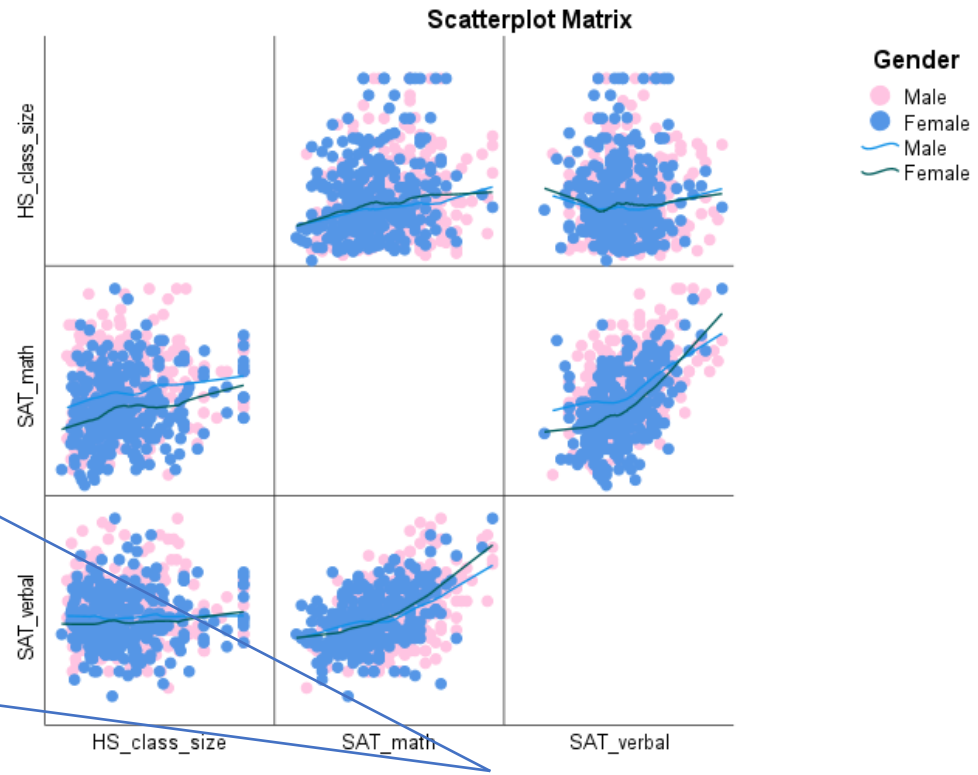




Scatterplots by Grouping Variable

- Can color code a scatterplot by a categorical variable to get a sense of how the distribution of variables is heterogeneous across groups



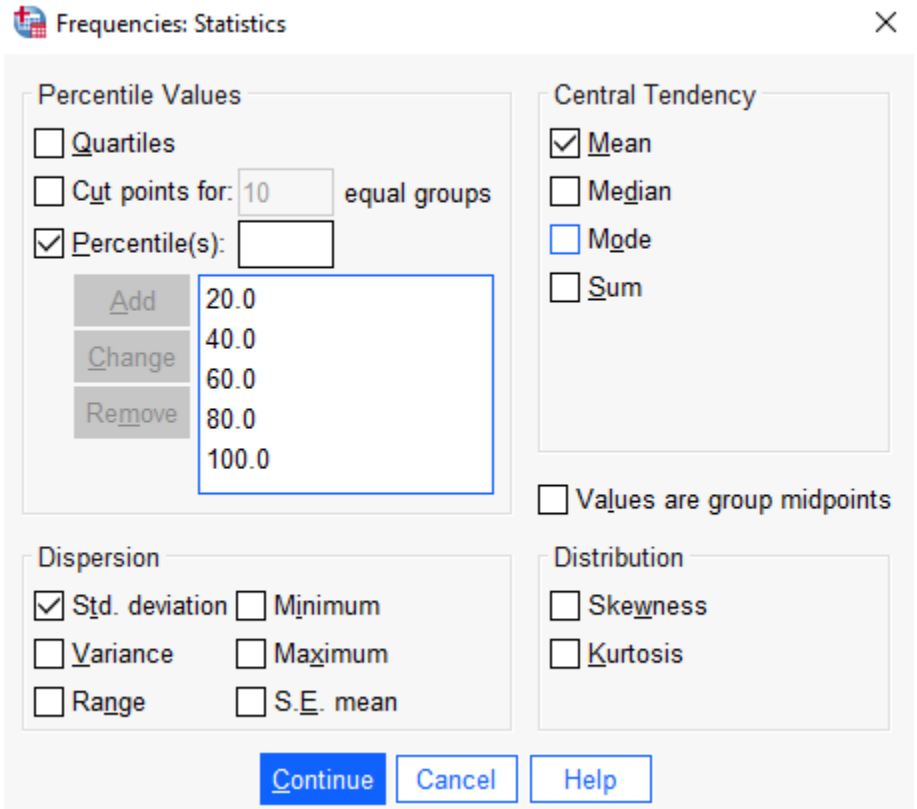


Your turn

- Can I predict your graduate GPA based on your undergraduate GPA & score on the GRE
- **gre_example.sav**
- Process
  - First, look at descriptive statistics
  - Second, calculate the zero order and partial correlations between the variables
  - Third, calculate b2 using formula
  - Fourth, run regression analysis
  - Fifth, interpret coefficients
  - Sixth, evaluate overall model for assumptions and fit

GRE-GPA Example Data

Student	GRE-Total (X_1)	Undergraduate GPA (X_2)	Graduate GPA(Y)
1	145	3.2	4.0
2	120	3.7	3.9
3	125	3.6	3.8
4	130	2.9	3.7
5	110	3.5	3.6
6	100	3.3	3.5
7	95	3.0	3.4
8	115	2.7	3.3
9	105	3.1	3.2
10	90	2.8	3.1
11	105	2.4	3.0



Statistics

		GRE score overall	GPA in undergraduate	GPA in graduate school
N	Valid	11	11	11
	Missing	0	0	0
Mean		112.7273	3.1091	3.5000
Std. Deviation		16.33457	.40113	.33166
Percentiles	20	97.0000	2.7400	3.1400
	40	105.0000	2.9800	3.3800
	60	116.0000	3.2200	3.6200
	80	128.0000	3.5600	3.8600
	100	145.0000	3.7000	4.0000

$$b_1 = \frac{(r_{Y.X1} - r_{Y.X2}r_{12})s_Y}{(1 - r_{12}^2)s_{X1}}$$

Y = GPA in graduate school
*X*1 = GRE score
*X*2 = Undergraduate GPA

$$b_1 = \frac{(r_{Y.X1} - r_{Y.X2}r_{12})s_Y}{(1 - r_{12}^2)s_{X1}} = \frac{(r_{Y.X1} - r_{Y.X2}r_{12}).3312}{(1 - r_{12}^2).4011}$$

Zero order (the Pearson you already know)

Partial correlation (the correlation after one variable has been partialed out)

Correlations

Control Variables			GPA in graduate school	GPA in undergraduate	GRE score overall
-none-[a]	GPA in graduate school	Correlation	1.000	.752	.784
		Significance (2-tailed)	.	.008	.004
		df	0	9	9
	GPA in undergraduate	Correlation	.752	1.000	.301
		Significance (2-tailed)	.008	.	.368
		df	9	0	9
	GRE score overall	Correlation	.784	.301	1.000
		Significance (2-tailed)	.004	.368	.
		df	9	9	0
GRE score overall	GPA in graduate school	Correlation	1.000	.872	
		Significance (2-tailed)	.	.001	
		df	0	8	
	GPA in undergraduate	Correlation	.872	1.000	
		Significance (2-tailed)	.001	.	
		df	8	0	

a. Cells contain zero-order (Pearson) correlations.

$r_{Y.X1} = .784$ The correlation between GRE and Graduate GPA

$r_{Y.X2} = .752$ The correlation between Graduate GPA and Undergrad GPA

$r_{12} = .301$ The correlation between GRE and Undergraduate GPA

$$b_1 = \frac{(r_{Y1} - r_{Y.X2}r_{12})s_Y}{(1 - r_{12}^2)s_{X1}} = \frac{[.784 - (.752)(.301)].332}{(1 - .301^2)16.33}$$

$$b_2 = \frac{(r_{Y.X2} - r_{Y.X1}r_{12})s_Y}{(1 - r_{12}^2)s_{X2}} = \frac{[.7516 - (.784)(.301)].332}{(1 - .301^2).401}$$

Sample Partial Slope and Intercept

$$b_1 = \frac{(r_{Y1} - r_{Y2}r_{12})s_Y}{(1 - r_{12}^2)s_{X1}} = \frac{[.7845 - (.7516)(.3011)].3317}{(1 - .3011^2)16.3346} = .0125$$

$$b_2 = \frac{(r_{Y.X2} - r_{Y.X1}r_{12})s_Y}{(1 - r_{12}^2)s_{X2}} = \frac{[.7516 - (.7845)(.3011)].3317}{(1 - .3011^2).4011} = .4687$$

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 3.5000 - (.0125)(112.7273) - (.4687)(3.1091) = .6337$$

Sample Multiple Linear Regression Model

$Y_i = b_0 + b_1 X_1 + b_2 X_2 + e_i$

$\widehat{Y}_i = .6337 + .0125\ X_1 + .4687\ X_2 + e_i$

- If your score on the GRETOT was 130 and your UGPA was 3.5, then your predicted score on the GGPA would be computed as:
  - $\widehat{Y}_i = .0125\ (130) + .4687\ (3.5000) + .6337 = 3.8992$

Run the MLR of grad school GPA on undergraduate GPA and GRE using gre_example.SAV

Overall *F* Test Statistic

$$F = \frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}} = \frac{0.9998/2}{.1002/8} = 39.9122$$

- The null: all slopes equal 0
- The critical value, at the .05 level of significance, is $_{.05}F_{2,8} = 4.46$
- Test statistic exceeds the critical value, so we reject H_0 and conclude that all of the partial slopes are not equal to zero at the .05 level of significance

Re-visiting Multicollinearity (summary)

- One of the assumptions of the linear regression model is that there is no exact linear relationship among the regressors
  - *Perfect collinearity*: A perfect linear relationship between the two variables exists.
  - *Imperfect collinearity*: The regressors are highly (but not perfectly) collinear
- Why care?
  - Even though some regression coefficients are statistically insignificant, the R2 value may be very high.
  - One may conclude (misleadingly) that the true values of these coefficients are not different from zero

Variance Inflation Factor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$VIF = \frac{1}{1 - r_{12}^2}$$

- VIF is a measure of the degree to which the variance of the OLS estimator is inflated because of collinearity.

Clues that MC is a problem

- High R^2 but few significant *t* ratios
- High pair-wise correlations among explanatory variables or regressors
- High partial correlation coefficients
- High Variance Inflation Factor (VIF) and low Tolerance Factor (TOL, the inverse of VIF)

What should we do if we detect multicollinearity?

- Nothing, for we often have no control over the data and our model is driven by theory
- Redefine the model by excluding variables may attenuate the problem, provided we do not omit relevant variables
- Principal components analysis (this is something covered in STATs II)

Advanced Topic: Heterogeneous Variable Types

1. Categorical variables in the Regression
   - Use dummy coding to encode the categories as 0 and 1
   - Shows up as additive corrections for the categorical value

2. Polynomial Terms in the Regression
   - We can fit the coefficients of a polynomial relationship via raising obs. to powers i.e.
   - To fit a quadratic: $\hat{y} = b_0 + b_1 x + b_2 x^2$

New variable, simply the square of the x

3. Interaction Terms in the Regressions
   - Effects are not always simply additive, can have multiplicative interactions i.e.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \rightarrow \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

   - Different lines for different levels of a categorical variable

New variable, product of x_1 and x_2

Categorical Predictors in Regression

Thus far we've seen regressions characterized by:

- Numeric response (Y) variable
  - Example: SAT Verbal Score
- Numeric explanatory (X) variable(s)
  - Example: SAT Math Score, ACT Score, Class Rank

- **Idea**: Explanatory variables are not just limited to numeric variables! We can incorporate categorical and higher order variables into the regression equation with (almost) no additional work.

Models with Dummy Variables

- Some models have both numeric and categorical explanatory variables (e.g., gender or race)
- If a categorical variable has *k* levels, need to create *k*-1 dummy variables that take on the values 1 if the level of interest is present, 0 otherwise.
- The baseline level of the categorical variable for which all *k*-1 dummy variables are set to 0
- The regression coefficient corresponding to a dummy variable is the difference between the mean for that level and the mean for baseline group, controlling for all other predictors

Interpreting Dummy Variables

- Can add categorical variables to regression model, but first need to create dummy codes
  - If k categories, you only need k-1 dummy variables
  - Other category acts as reference group; absorbed into intercept
- Each dummy variable acts as an indicator of whether the observation is in that category
- Each dummy coefficient can be interpreted as *the average difference* in Y between the dummy category and the reference group.
  - Specifically, the dummy coefficients are the linear correction in the model for that level, relative to the reference group.

Categorical Variables and Dummies

- **Example:** You have a variable coding years of experience. Years of experience is coded as
  - 1 = 10 years or less
  - 2 = 11 years or more
- What is *K;* how many dummies do you need?
- **Idea:** Use a binary (0/1) variable to identify different groups i.e. X_i = 1 if observation is of the category and X_i = 0 otherwise.
- The binary categorical variable acts as an additive correction term
  - Suppose x_2 is binary with coefficient b_2
  - Then the model is: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ which splits into:

$$\hat{y}|(x_2=0) = b_0 + b_1 x_1 \qquad \hat{y}|(x_2=1) = b_0 + b_1 x_1 + b_2 x_2$$

Additive Correction

Example: Recoding dummy variables

Use the school admissions data variable (gre_example.sav) for whether a student has a family member that is an alumni (Number_of_family_alumni) and recode it as a simple binary variable (Y/N).

- Label the groups with no alum family 0, else 1
- The “0” group is called the **reference group**, no additive correction present here
- We will use it in a multiple linear regression where the “slope” for the binary categorical variable we create will represent the ***difference in the intercepts between of the two groups***

Number_of_family_alumni

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	699	79.4	79.4	79.4
	1	138	15.7	15.7	95.1
	2	34	3.9	3.9	99.0
	3	6	.7	.7	99.7
	4	3	.3	.3	100.0
	Total	880	100.0	100.0	

alumniYN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Family is not an alum	699	79.4	79.4	79.4
	Family is an alum	181	20.6	20.6	100.0
	Total	880	100.0	100.0	

Example

- Run a multiple linear regression of HS class rank on Math SAT score and the dummy variable coding the effect of having a family member as an alum (=1) or not (=0)

With only two groups, the dummy variable represents the average difference in HS rank between students who have family members that are an alumni and those who do not.

Note: this is not really any different, an increase in X means going from 0 to 1

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	68.891	21.512		3.202	.001
	SAT_math	.007	.036	.008	.206	.837
	alumniYN	-2.604	5.985	-.017	-.435	.664

a. Dependent Variable: HS_rank

Regression Equation:
Rank = 68.89 + .007*SAT_math – 2.6*(IF ALUM)

Interpretation: the estimated difference in HS rank between students who have family that is an alumn and those who do not is -2.604. Students who have one or more family members who are alums have a HS rank that is 2.604 points *lower*, on average, *holding SAT score constant at its mean.*

The correction

- Then the model is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ which splits into:

$\hat{y}|(x_2=0) = \hat{\beta}_0 + \hat{\beta}_1 x_1$ $\qquad$ $\hat{y}|(x_2=1) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2$

Regression Equation:
Rank = 68.89 + .007*SAT_math – 2.6*(IF ALUM)

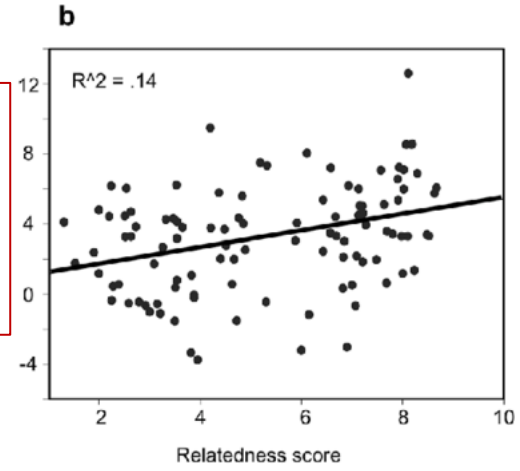
Regression Equation:
Rank|Alum = 0: 68.89 + .007*SAT_math

Regression Equation:
Rank|Alum = 1: = 68.89 + .007*SAT_math – 2.6*(1)

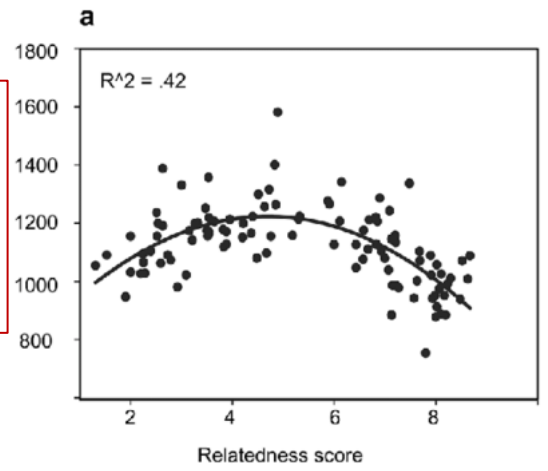
Quadratic Terms

- **Problem**: What if, before running our multiple linear regression we look at paired scatter plots and observe that the relationship between independent and response is not simply a straight line?

(Barely) Linear Scatterplot Rel.



Nonlinear Scatterplot Rel.



- <u>One Solution</u>: Use X^2 to model a curvilinear relationship:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 {x_1}^2$$

- x_1^2 is treated as another predictor, data is generated by simply squaring the features value for each observation!

Quadratic Terms

- A quadratic term is a the variable multiplied with itself (i.e., squared
  - Example: Age and crime?

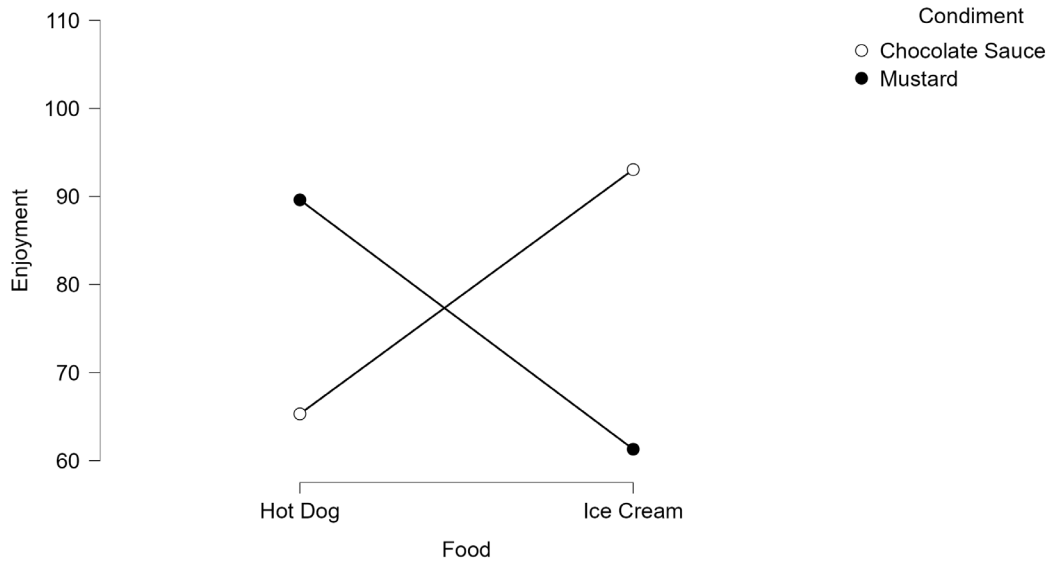
- **Warning**: Adding quadratic terms increases the number of variables in your model like anything else, and thus hurts the models R^2 unless this relationship is strongly present
  - Why?
  - It's also an easy way to overfit your data, can lead to bizarre predictions!
  - Don't do…

Interaction Terms

- More generally than just quadratic terms, sometimes the the effect of one variable depends on the value of another variable
  - Can include interactions of 2 or more variables
  - Usually one continuous and one categorical variable (but not always)
  - Model is of the form: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$
- Note these sorts of interactions are *hard to pick up from scatterplots*. In this class, they will come most often from THEORY
- Natural Interpretation when interactions include categorical variables!
  - Variables of the form x_1x_2 where x_1 is binary and x_2 is numeric, model linear corrections as opposed to additive corrections. That is they are corrections to the slope of the coefficient on the numeric variable.
  - Example: in the school data, the relationship between GRE score and GPA may depend in a *different linear* way conditional on the race and gender of the student. Why?

Interpreting Interaction Terms (1)

- The “it depends”
  - Do you enjoy chocolate sprinkles or ketchup on your food. Well, it depends on the type of food.



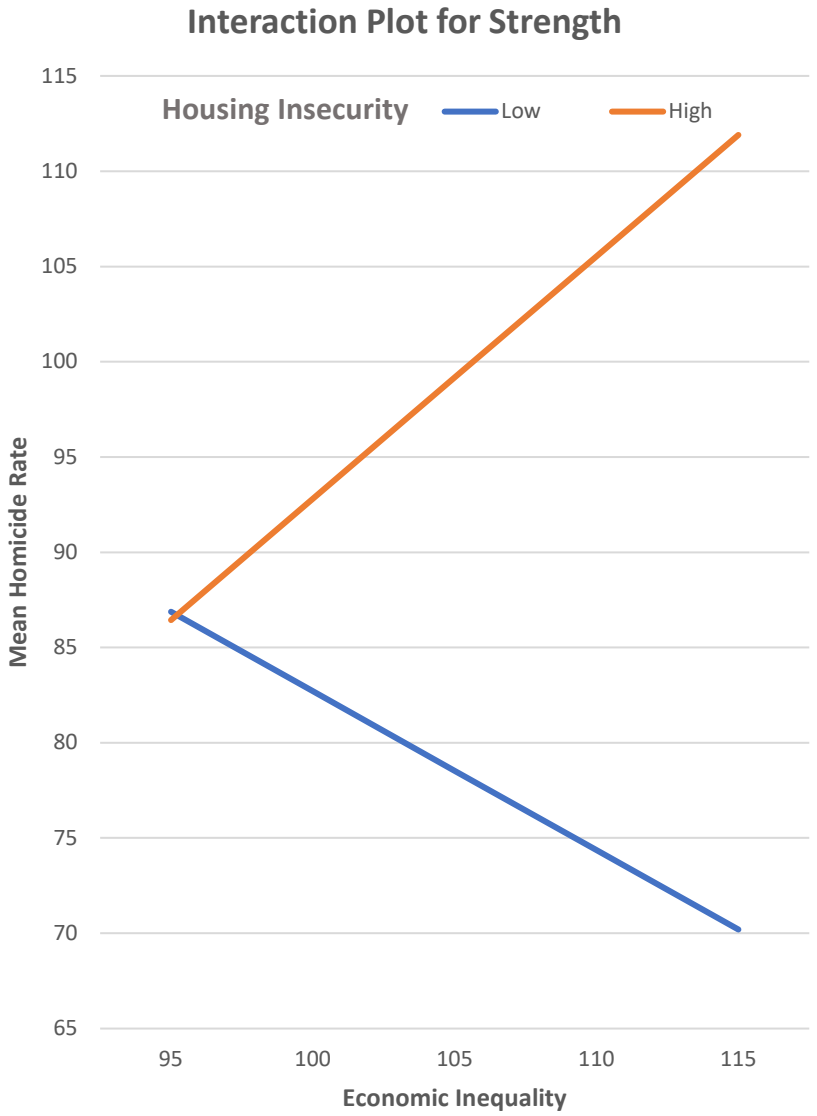
Enjoyment levels are higher for chocolate sauce when the food is ice cream but higher for mustard when the food is a hot dog.

If you put mustard on ice cream or chocolate sauce on hot dogs, you won’t be happy!

Note the measurement of the variables

Interpreting Interaction Terms (2)

- Say you have data on homicide rate, poverty, economic inequality and % of persons who are housing insecure in a neighborhood
- You think that these variables are important for explaining the homicide rate but you also believe there is an interaction between economic inequality and housing insecurity
  - The effect of economic inequality on homicide depends housing insecurity such that areas with *higher levels of economic inequality* will have higher homicide rates if housing insecurity is high but not when housing insecurity is low



Continuous-Interaction-Excel-Example.xlsx

Descriptive Statistics

	Mean	Std. Deviation	N
TR: Trauma: PTS T Score	50.46	11.230	1735
EV: Severe Violence Total # of Exposure	1.08	1.350	1735
Child OOH Situation (YN)	1.74	.441	1735
Child gender (chgendr)	.45	.498	1735
Child age in years (chAge_b)	11.21	2.159	1735
physab	.2646	.44122	1735

Correlations

		TR: Trauma: PTS T Score	EV: Severe Violence Total # of Exposure	Child OOH Situation (YN)	Child gender (chgendr)	Child age in years (chAge_b)	physab	sexab	nhw
Pearson Correlation	TR: Trauma: PTS T Score	1.000	.270	-.030	.029	-.054	.022	.047	-.020
	EV: Severe Violence Total # of Exposure	.270	1.000	-.157	-.047	.043	-.004	.031	-.045
	Child OOH Situation (YN)	-.030	-.157	1.000	-.006	-.083	.093	-.026	.051
	Child gender (chgendr)	.029	-.047	-.006	1.000	-.078	.045	-.208	.010
	Child age in years (chAge_b)	-.054	.043	-.083	-.078	1.000	.046	.071	.019
	physab	.022	-.004	.093	.045	.046	1.000	-.292	-.043
	sexab	.047	.031	-.026	-.208	.071	-.292	1.000	.019
	nhw	-.020	-.045	.051	.010	.019	-.043	.019	1.000
Sig. (1-tailed)	TR: Trauma: PTS T Score	.	<.001	.104	.114	.012	.182	.025	.199
	EV: Severe Violence Total # of Exposure	.000	.	.000	.026	.038	.436	.096	.031
	Child OOH Situation (YN)	.104	.000	.	.400	.000	.000	.137	.017
	Child gender (chgendr)	.114	.026	.400	.	.001	.029	.000	.339
	Child age in years (chAge_b)	.012	.038	.000	.001	.	.028	.002	.208
	physab	.182	.436	.000	.029	.028	.	.000	.037
	sexab	.025	.096	.137	.000	.002	.000	.	.216
	nhw	.199	.031	.017	.339	.208	.037	.216	.

Your turn

- This data is from NASCW on youth in the child welfare system
- The variables are as follows:
  - EV severe violence exposure
  - Child is in out-of-home care (dummy, 1 = Yes, 0 = No)
  - Child gender (0 = male, 1 = female)
  - Child Age in Years
  - Child experienced physical abuse (1 = yes, 0 = no)
  - Child experienced sexual abuse (1 = yes, 0 = no)
  - Child is Non-Hispanic White (1 = yes, 0 = no)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.288[a]	.083	.079	10.774	1.968

a. Predictors: (Constant), nhw, Child gender (chgendr), Child OOH Situation (YN), physab, Child age in years (chAge_b), EV: Severe Violence Total # of Exposure, sexab

b. Dependent Variable: TR: Trauma: PTS T Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	50.645	1.856		27.291	<.001	47.005	54.285
	EV: Severe Violence Total # of Exposure	2.282	.195	.274	11.727	<.001	1.900	2.663
	Child OOH Situation (YN)	.141	.599	.006	.235	.814	-1.035	1.316
	Child gender (chgendr)	1.089	.533	.048	2.043	.041	.044	2.135
	Child age in years (chAge_b)	-.353	.121	-.068	-2.911	.004	-.591	-.115
	physab	1.073	.618	.042	1.736	.083	-.139	2.286
	sexab	1.874	.704	.066	2.664	.008	.494	3.254
	nhw	-.156	.521	-.007	-.299	.765	-1.178	.866

a. Dependent Variable: TR: Trauma: PTS T Score

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18196.139	7	2599.448	22.394	<.001[b]
	Residual	200468.412	1727	116.079		
	Total	218664.551	1734			

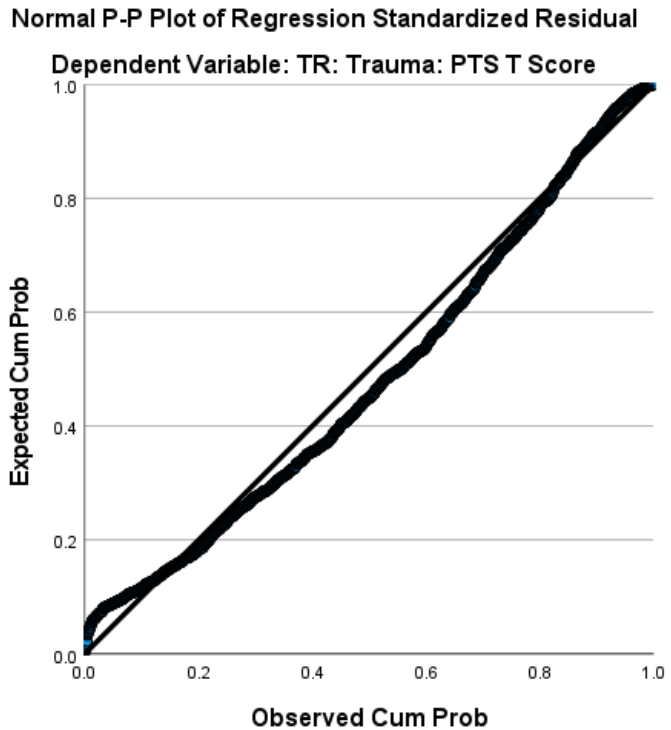
a. Dependent Variable: TR: Trauma: PTS T Score

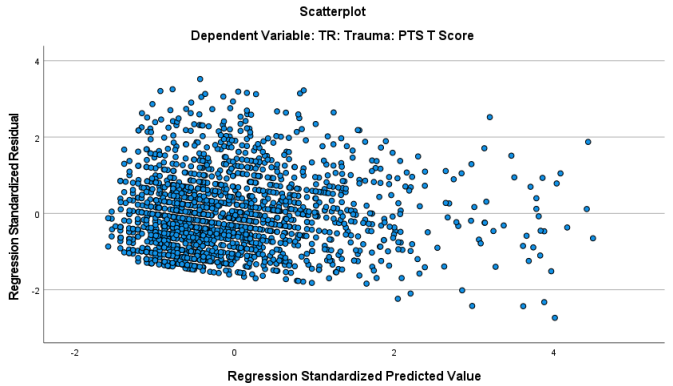
b. Predictors: (Constant), nhw, Child gender (chgendr), Child OOH Situation (YN), physab, Child age in years (chAge_b), EV: Severe Violence Total # of Exposure, sexab

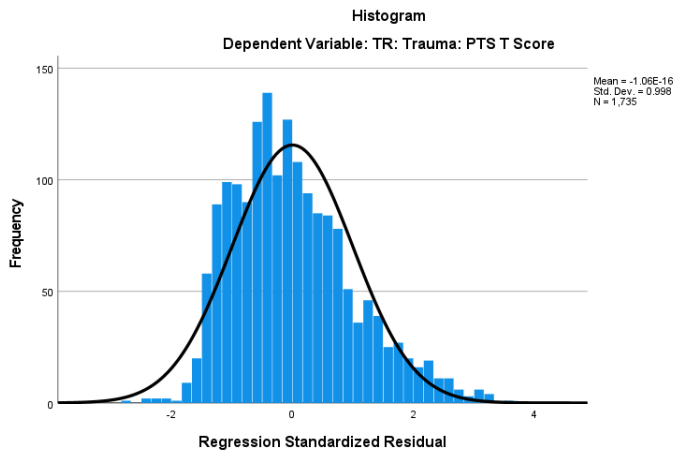
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	45.33	65.01	50.46	3.239	1735
Residual	-29.462	37.935	.000	10.752	1735
Std. Predicted Value	-1.581	4.492	.000	1.000	1735
Std. Residual	-2.735	3.521	.000	.998	1735

a. Dependent Variable: TR: Trauma: PTS T Score







Other hypotheses to test

- Older children with higher levels of violence exposure have more PTS symptoms compared to younger children
- Children who are sexually abused with higher levels of violence exposure have more PTS symptoms compared children who have experienced other forms of abuse
- Children who have more depressive symptoms and higher levels of violence exposure have more PTS symptoms compared to children with less depressive symptoms
  - Note: this is the same hypothesis as: Children who have higher levels of violence exposure and more depressive symptoms have more PTS symptoms compared to children with less violence exposure

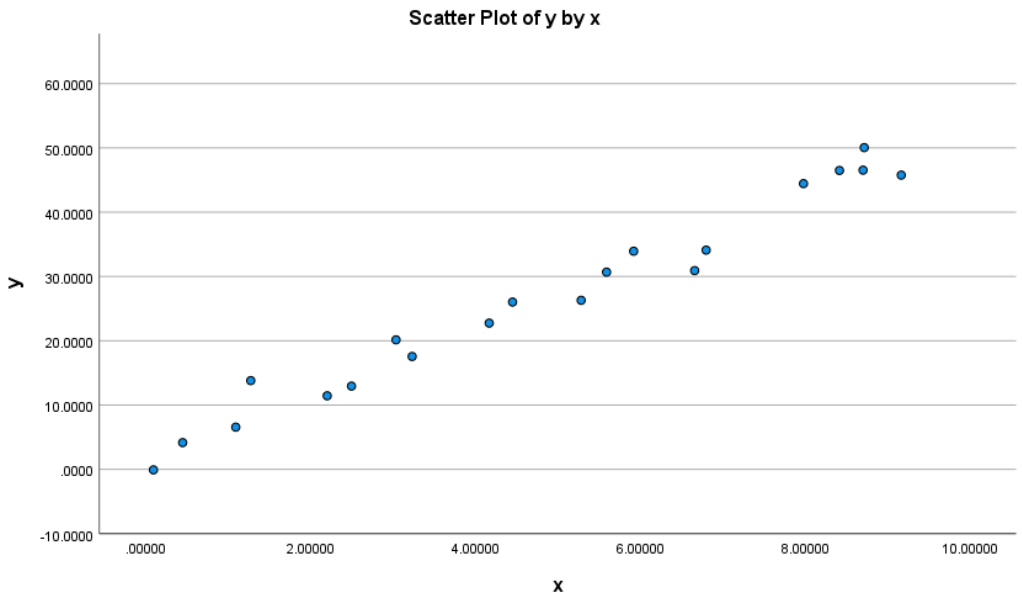
More diagnostics

- An outlier is a data point whose response y does not follow the general trend of the rest of the data
  - It is defined as a point that is 1.5(Q3-Q1) = 1.5*IQR*
- A data point has high leverage if it has "extreme" predictor *X* values
  - With a single predictor, an extreme x value is simply one that is particularly high or low.
  - With multiple predictors, extreme *X* values may be particularly high or low for one or more predictors
    - Example: r = +.90 for X_1, X_2 but a case has a high X_1 and a low value on X_2

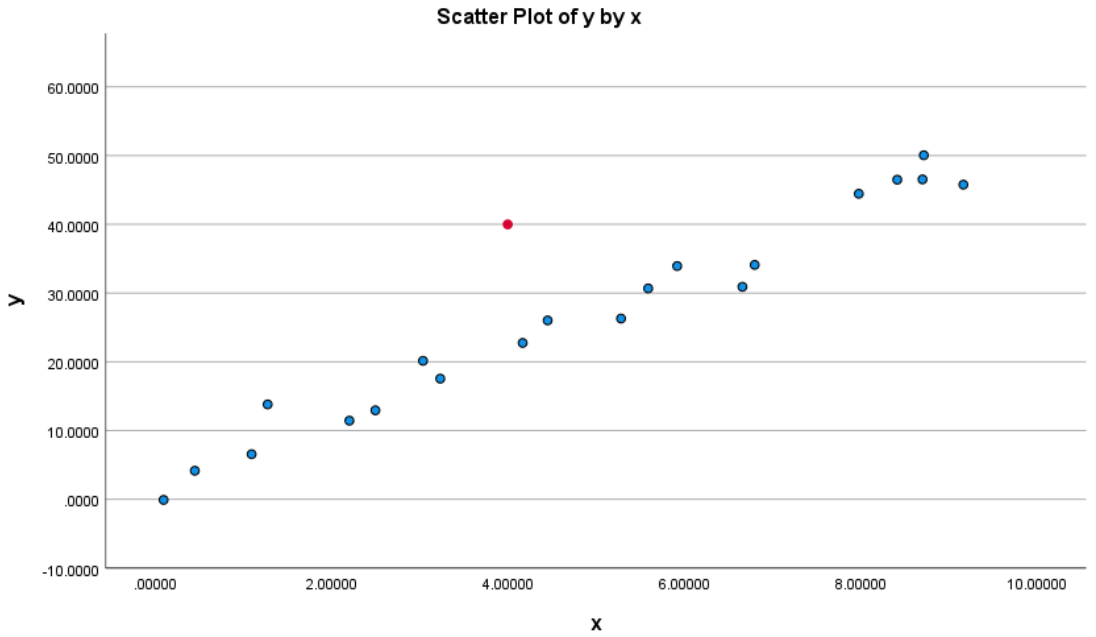
Outliers and unusual values

- Does anything look unusual or different about plot A?
- Does anything look unusual or different about plot B?

A



B



Regression output

- Compare the two regressions, any differences?

A

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.986[a]	.973	.972	2.5919877

a. Predictors: (Constant), x

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4386.068	1	4386.068	652.844	<.001[b]
	Residual	120.931	18	6.718		
	Total	4506.999	19			

a. Dependent Variable: y

b. Predictors: (Constant), x

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1.732	1.121		1.546	.140
	x	5.117	.200	.986	25.551	<.001

a. Dependent Variable: y

B

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.954[a]	.910	.905	4.7107501

a. Predictors: (Constant), x

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4265.823	1	4265.823	192.231	<.001[b]
	Residual	421.632	19	22.191		
	Total	4687.456	20			

a. Dependent Variable: y

b. Predictors: (Constant), x

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.958	2.009		1.472	.157
	x	5.037	.363	.954	13.865	<.001

a. Dependent Variable: y

Why is the standard error bigger?

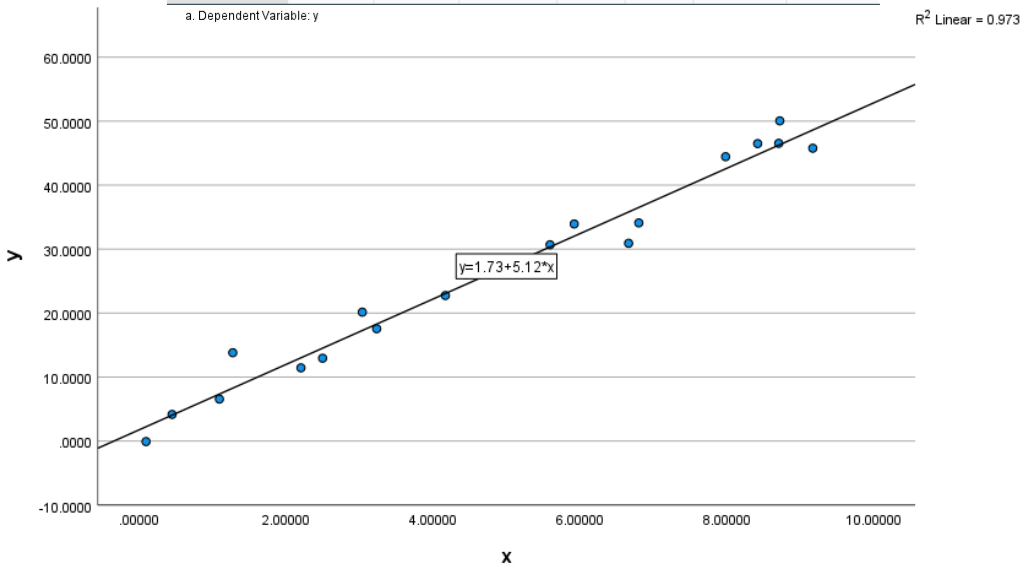
Compare & Intuit

- Lines are fairly similar BUT
  - More error & less confidence → smaller t-value, bigger standard errors, wider confidence interval

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)	1.732	1.121		1.546	.140	-.622	4.086
	x	5.117	.200	.986	25.551	<.001	4.696	5.538

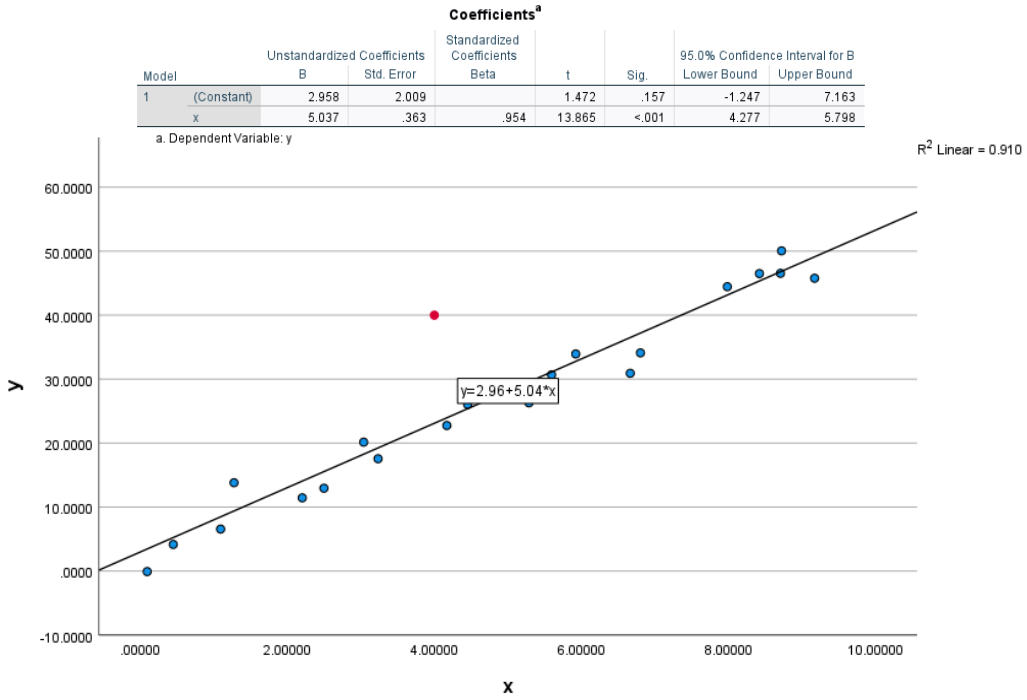
a. Dependent Variable: y



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)	2.958	2.009		1.472	.157	-1.247	7.163
	x	5.037	.363	.954	13.865	<.001	4.277	5.798

a. Dependent Variable: y



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.989[a]	.977	.976	2.7091121

a. Predictors: (Constant), x

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6028.817	1	6028.817	821.444	<.001[b]
	Residual	139.446	19	7.339		
	Total	6168.263	20			

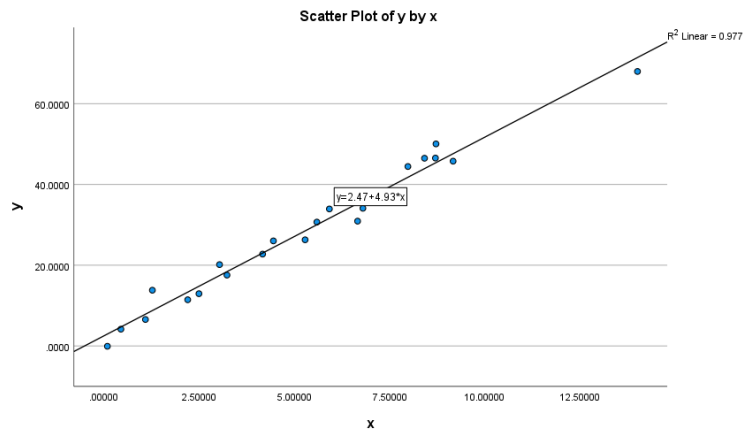
a. Dependent Variable: y

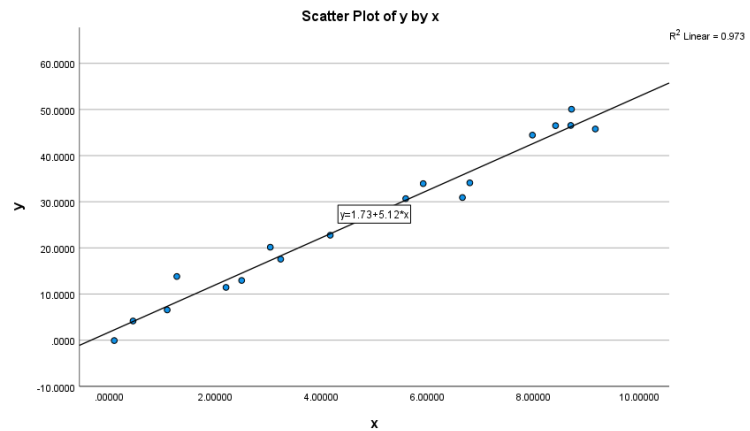
b. Predictors: (Constant), x

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.468	1.076		2.294	.033
	x	4.927	.172	.989	28.661	<.001

a. Dependent Variable: y





Unusual values and influential points

- Does anything look unusual or different about plot A?
- Does anything look unusual or different about plot B?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.743[a]	.552	.528	10.4459325

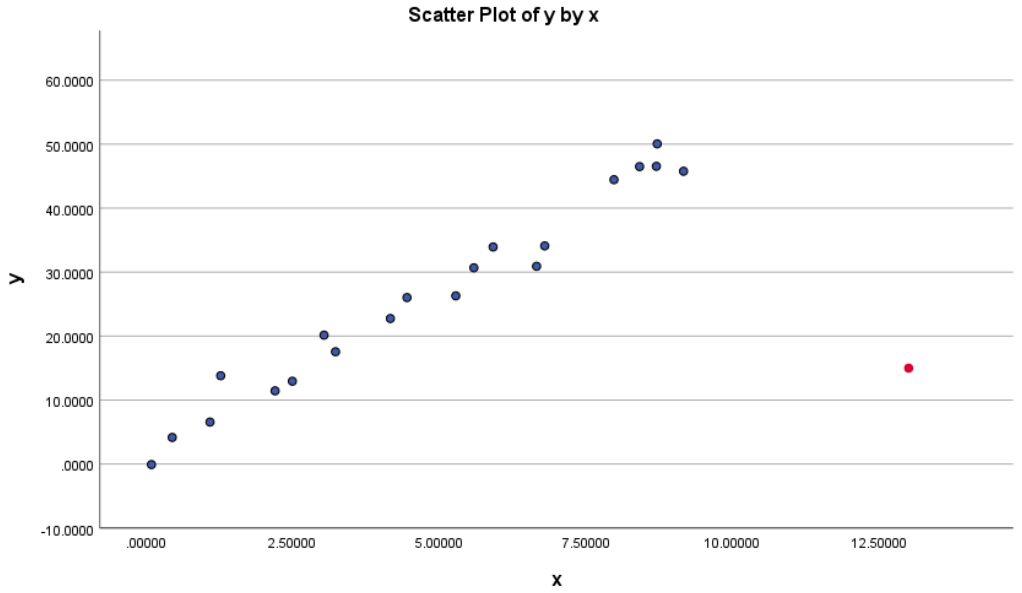
a. Predictors: (Constant), x

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)	8.505	4.222		2.014	.058	-.333	17.342
	x	3.320	.686	.743	4.838	<.001	1.884	4.756

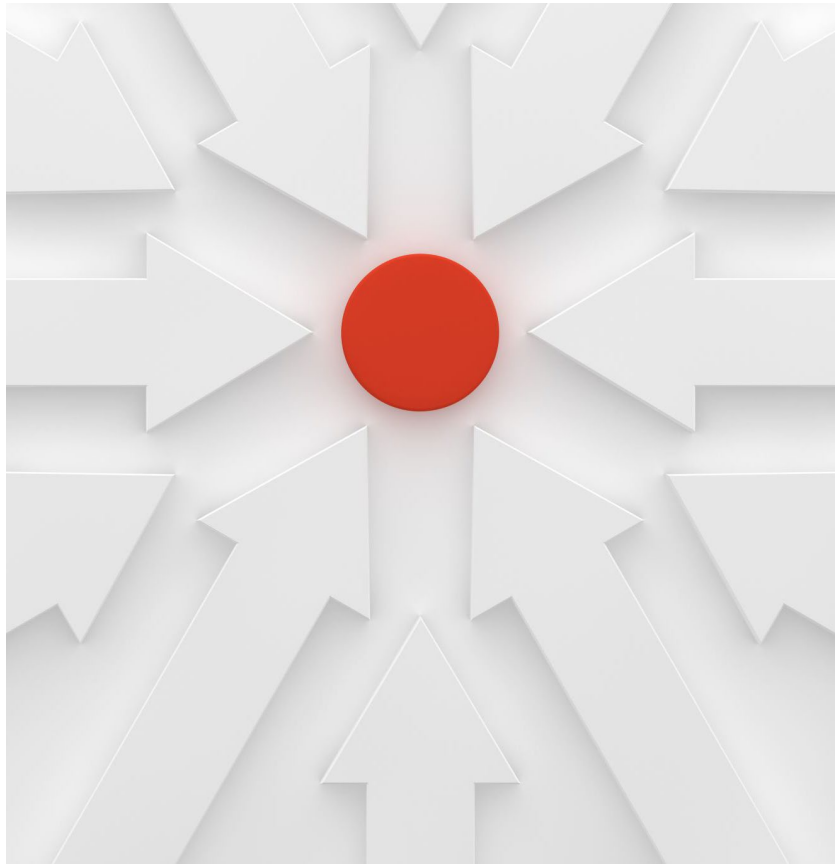
a. Dependent Variable: y

B



Identifying data points whose x values are extreme

- The leverage depends only on the predictor values
  - The leverage suggests only that a data point potentially exerts a strong influence on the regression analysis
  - Whether it is influential or not in actuality depends on the observed value of the response Y_i
- How to determine when leverage is large and worrisome?
  - Any observation whose leverage value, denoted h_{ii}, is > 3 times larger than the mean leverage value



Leverage

$$\bar{h} = \sum_{i=1}^{n} \frac{h_{ii}}{n} = \frac{p}{n}$$

p = # of parameters in the model and *n* = the number of observations

That is, if:

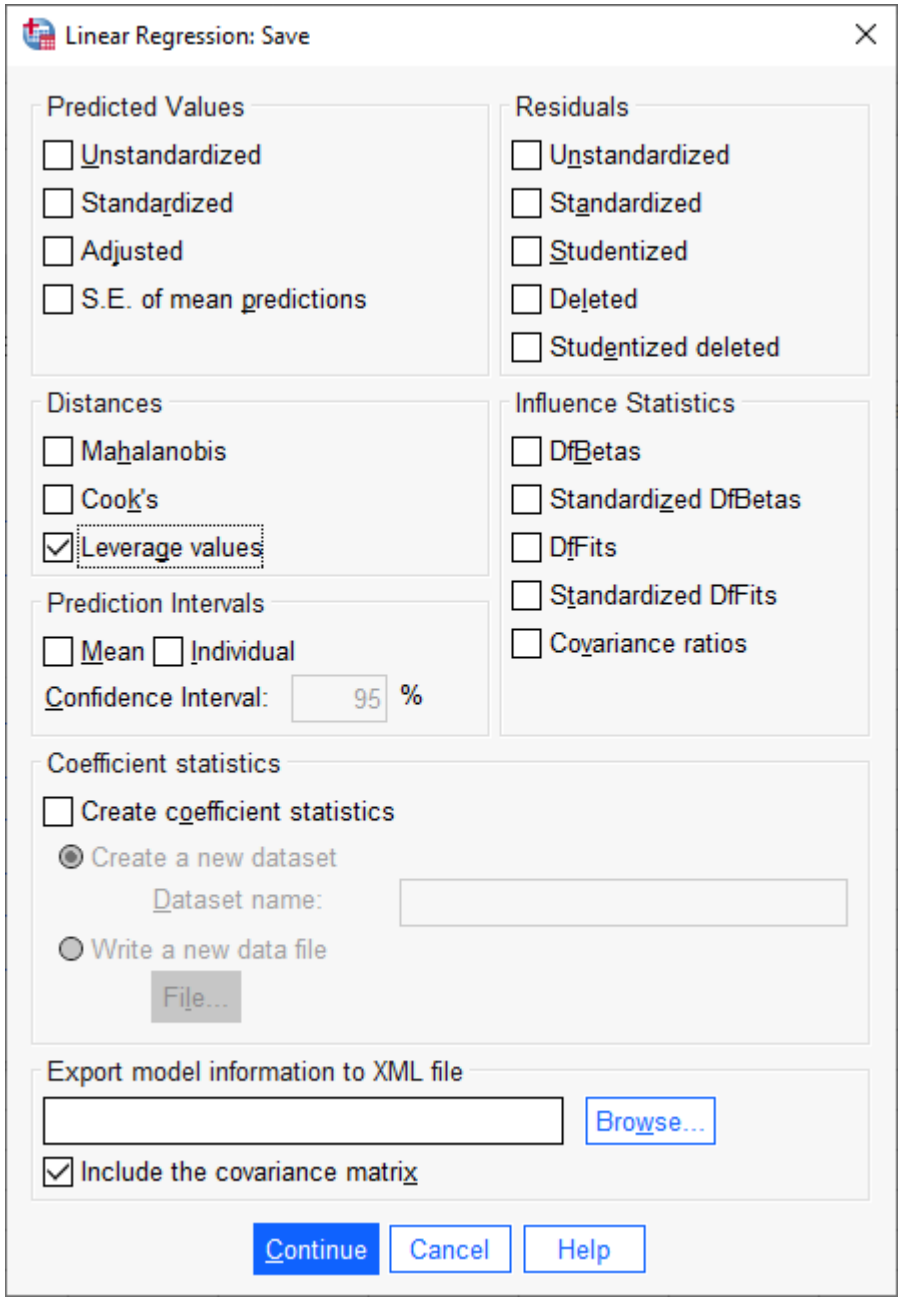
$$h_{ii} > 3\left(\frac{p}{n}\right)$$

Then flag the observations as unusual → X is an observation whose value gives it large *leverage* for the regression analysis

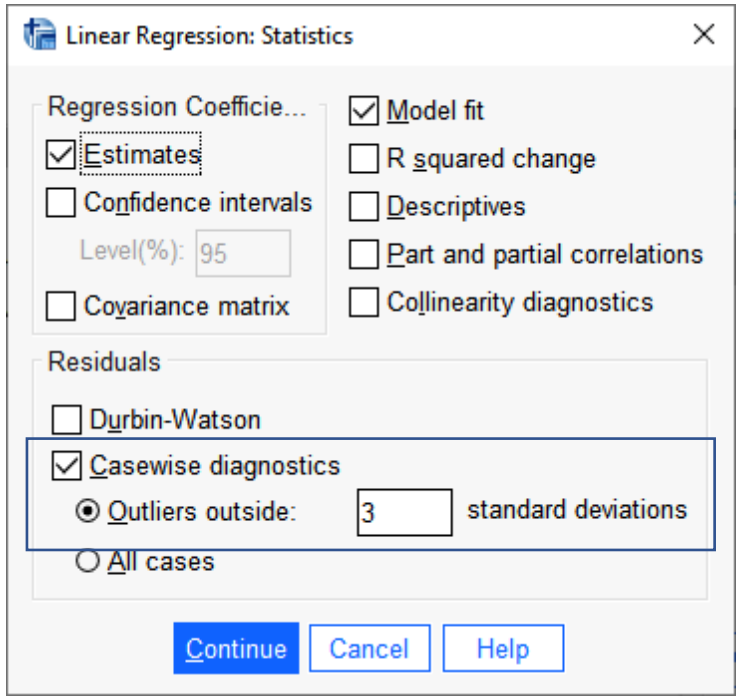
Example

- You perform a SLR with n = 21 cases. What is the leverage cutoff value that gives you some reason to be concerned?
- $p = 2$, $n = 21$

$h_{ii} > 3\left(\frac{2}{21}\right)$=.286



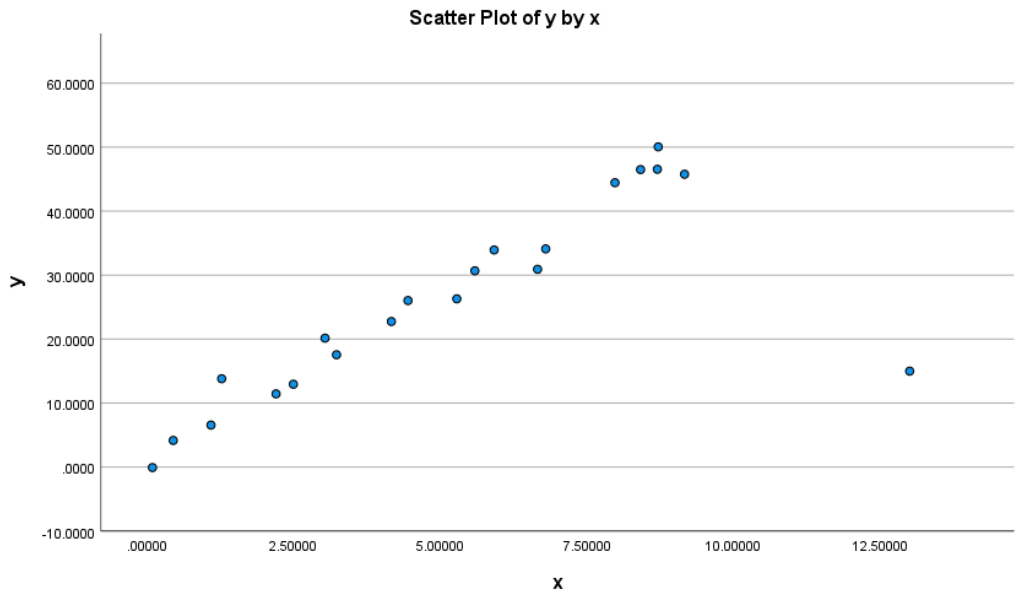
Can have SPSS provide outliers



Casewise Diagnostics[a]

Case Number	Std. Residual	y	Predicted Value	Residual
21	-3.510	15.0000	51.661922	-36.6619218

a. Dependent Variable: y



Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	8.836530	51.661922	25.699862	11.3003936	21
Std. Predicted Value	-1.492	2.297	.000	1.000	21
Standard Error of Predicted Value	2.281	5.830	3.117	.842	21
Adjusted Predicted Value	10.520250	68.251472	26.391249	13.1341963	21
Residual	-36.6619225	12.6166601	.0000000	10.1814356	21
Std. Residual	-3.510	1.208	.000	.975	21
Stud. Residual	-4.230	1.274	-.030	1.125	21
Deleted Residual	-53.2514763	14.0432806	-.6913867	13.6620603	21
Stud. Deleted Residual	-17.047	1.297	-.639	3.804	21
Mahal. Distance	.001	5.278	.952	1.185	21
Cook's Distance	.000	4.048	.213	.879	21
Centered Leverage Value	.000	.264	.048	.059	21

a. Dependent Variable: y

Recall the mean leverage is *p/n.* If all the observations have roughly equivalent influence on the estimated value of the coefficients, the leverages would be close to

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.743[a]	.552	.528	10.4459325

a. Predictors: (Constant), x

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2553.978	1	2553.978	23.406	<.001[b]
	Residual	2073.233	19	109.118		
	Total	4627.211	20			

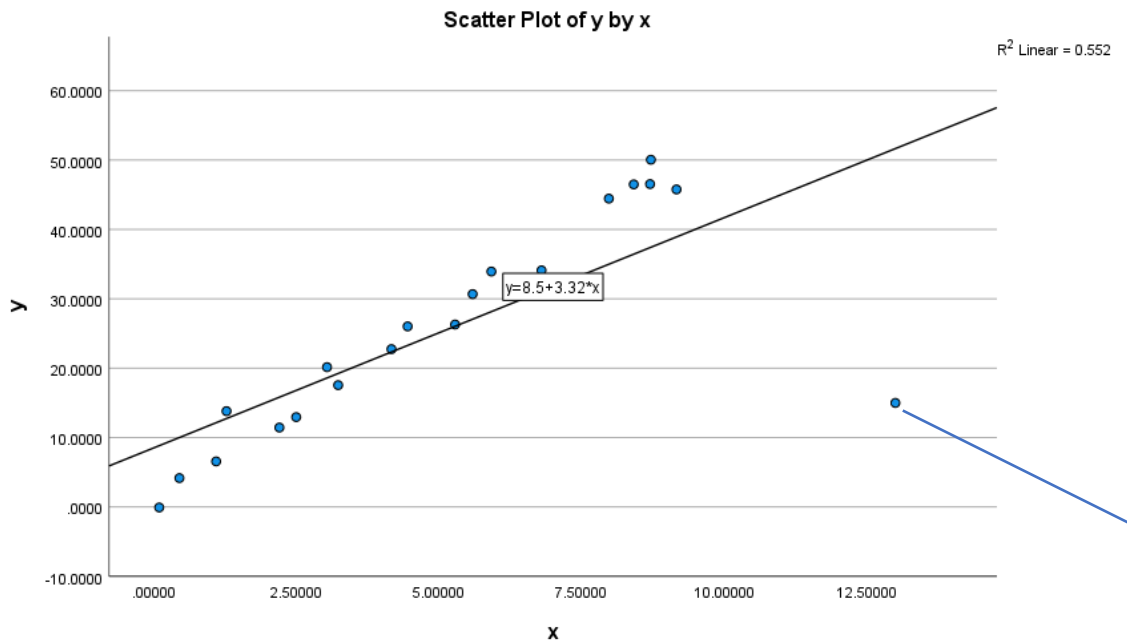
a. Dependent Variable: y

b. Predictors: (Constant), x

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8.505	4.222		2.014	.058
	x	3.320	.686	.743	4.838	<.001

a. Dependent Variable: y



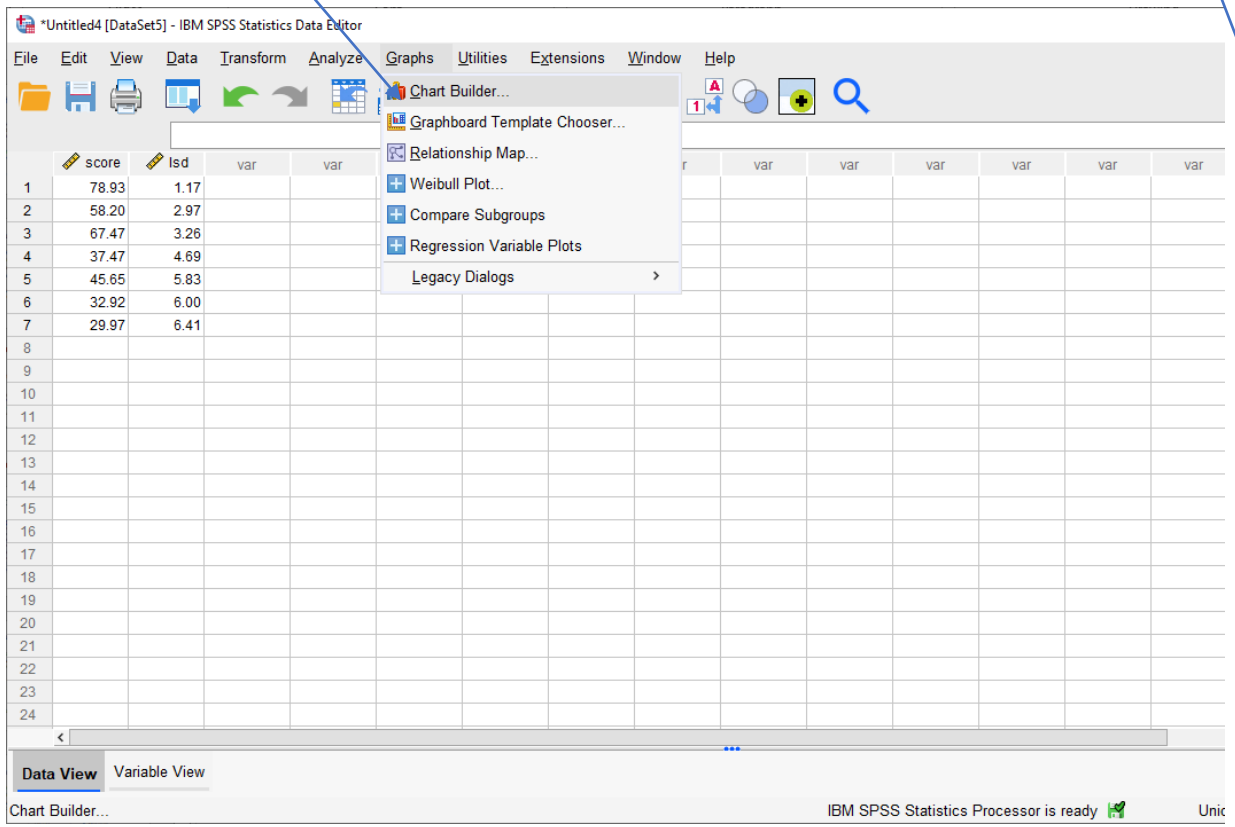
LEV_1
.11134
.09637
.07190
.06564
.03815
.03097
.01975
.01630
.00228
.00440
.00005
.00074
.00237
.00947
.01132
.03383
.04518
.05397
.05353
.06853
.26391

Example - Pharmacodynamics of LSD

- Response (*Y*) - Math score (mean among 5 volunteers)
- Predictor (*X*) - LSD tissue concentration (mean of 5 volunteers)
- Make scatterplot of data in SPSS
- USE LSD.SAV – a made up example of LSD concentration in blood and math score on a test

Math Score (y)	LSD Conc (x)
78.93	1.17
58.20	2.97
67.47	3.26
37.47	4.69
45.65	5.83
32.92	6.00
29.97	6.41

From SPSS


*Untitled4 [DataSet5] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Chart Builder...
Graphboard Template Chooser...
Relationship Map...
Weibull Plot...
Compare Subgroups
Regression Variable Plots
Legacy Dialogs
score lsd
1 78.93 1.17
2 58.20 2.97
3 67.47 3.26
4 37.47 4.69
5 45.65 5.83
6 32.92 6.00
7 29.97 6.41
Data View Variable View
Chart Builder...
IBM SPSS Statistics Processor is ready

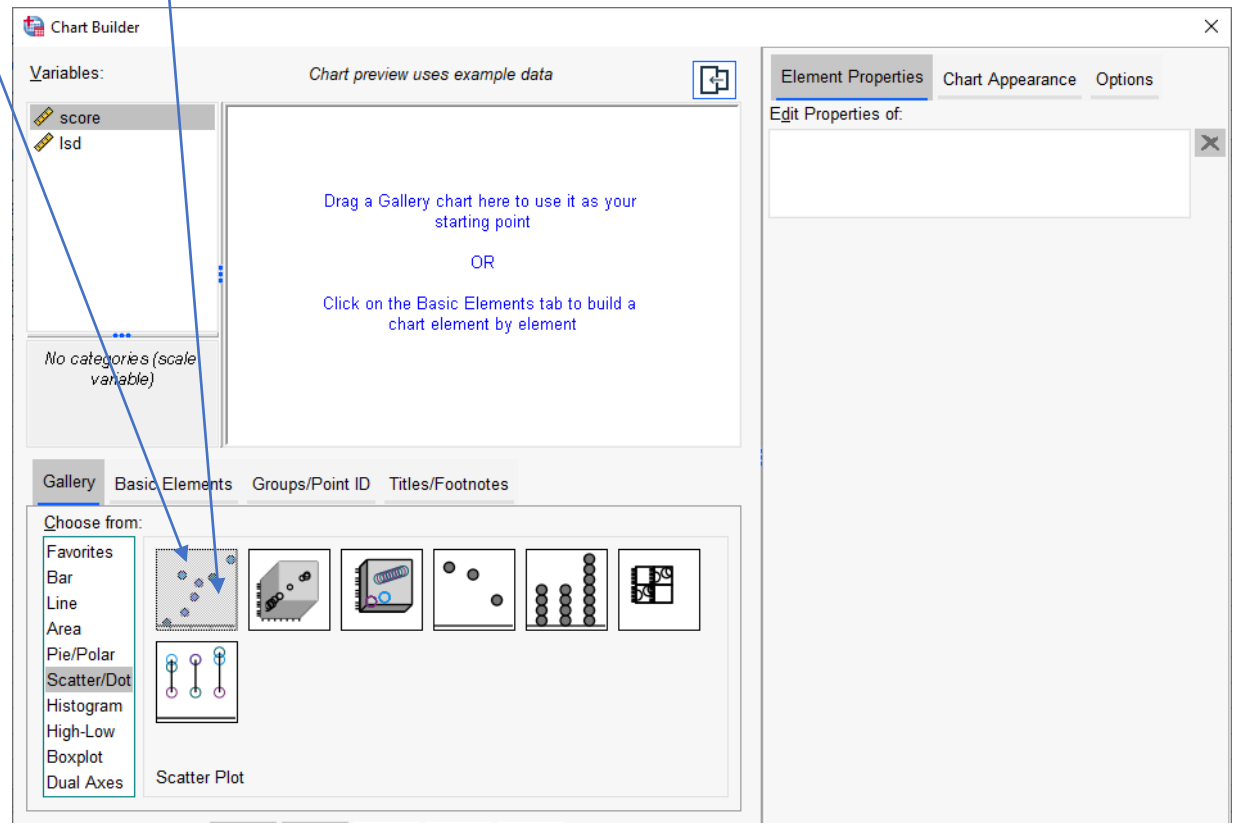
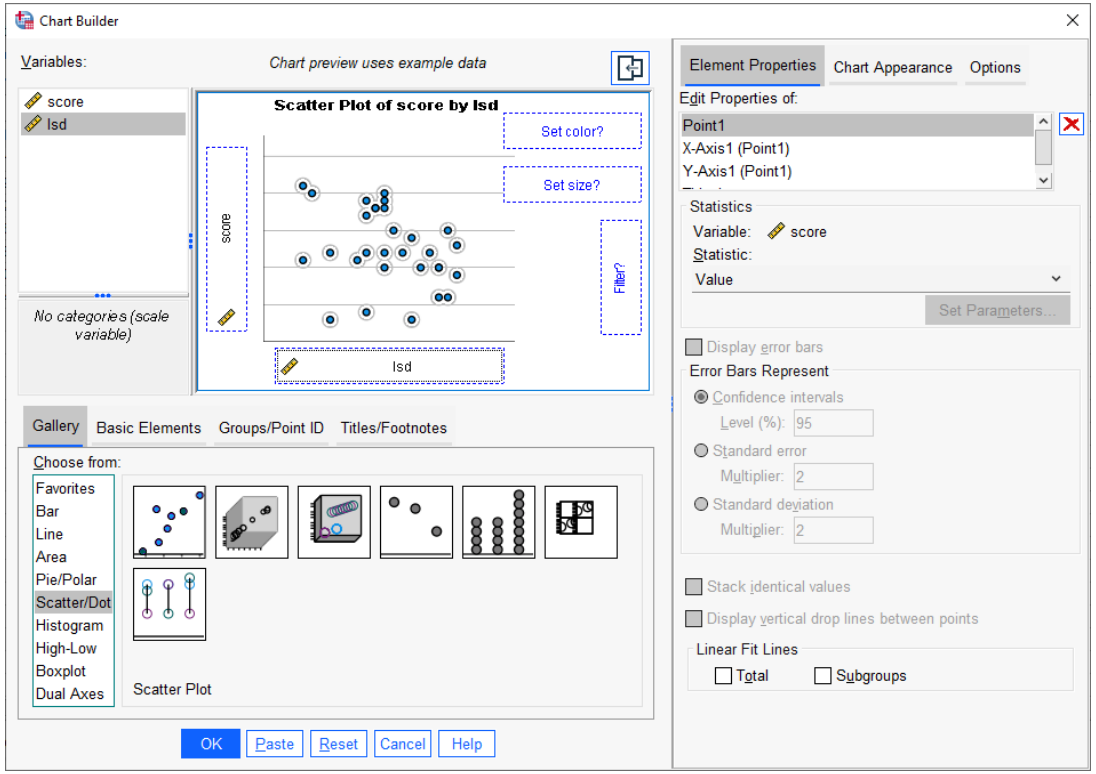
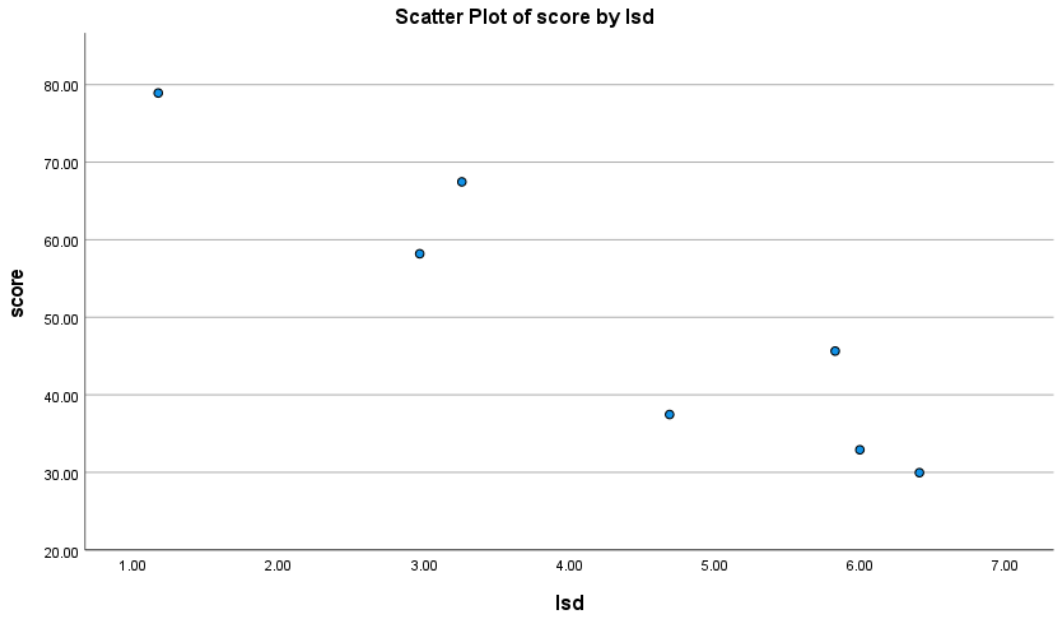


Chart Builder
Variables:
Chart preview uses example data
score
lsd
Drag a Gallery chart here to use it as your starting point
OR
Click on the Basic Elements tab to build a chart element by element
No categories (scale variable)
Element Properties
Chart Appearance
Options
Edit Properties of:
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
Scatter Plot

From SPSS

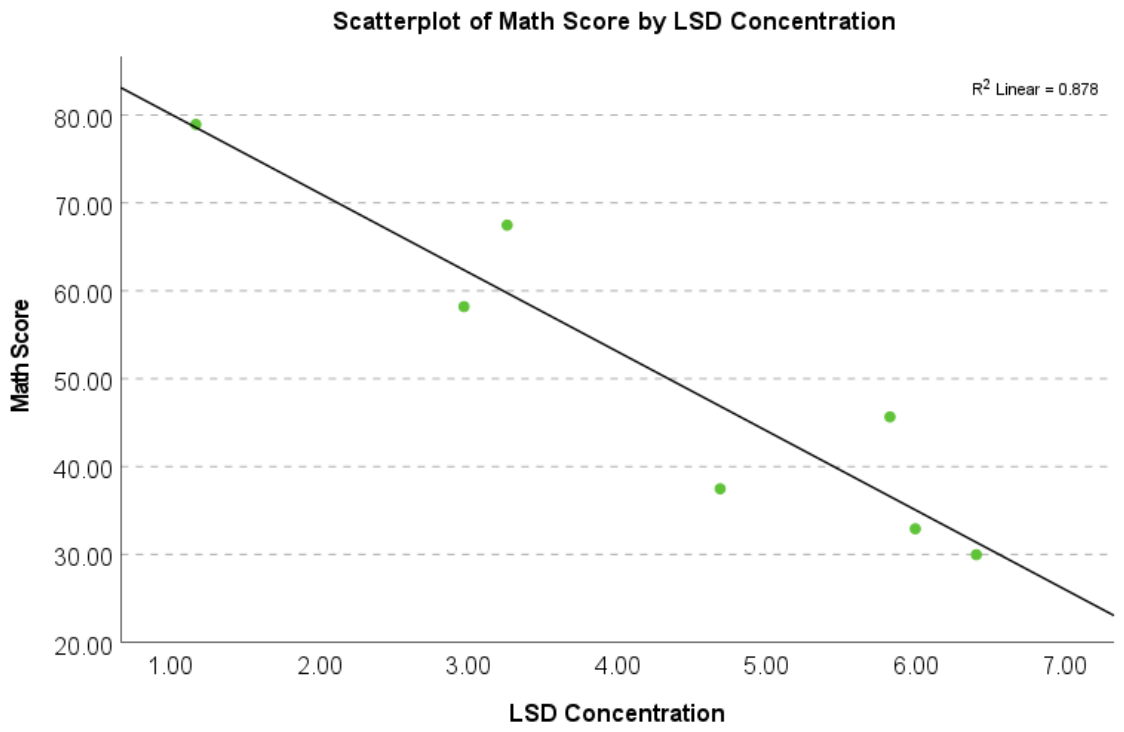
- Put score on Y axis and LSD on X axis
- Before we customize the plot it will look like this
- Let’s take some time making the chart look prettier





Revised Scatterplot

- As LSD concentration increases, math score decreases
- Looks like a strong, linear relationship
- Let's do a SLR to get the nature of the relationship



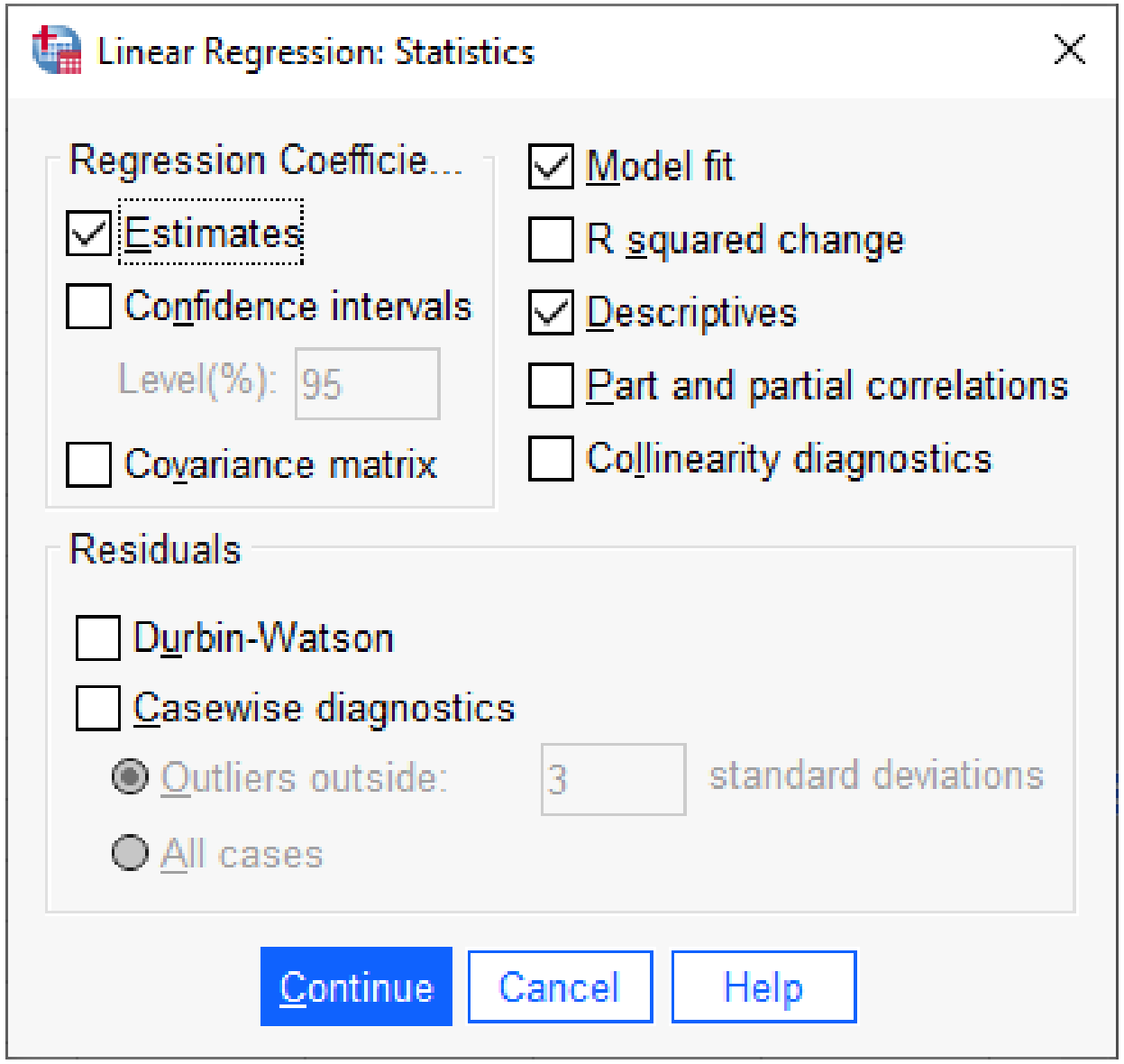
Least Squares Computations

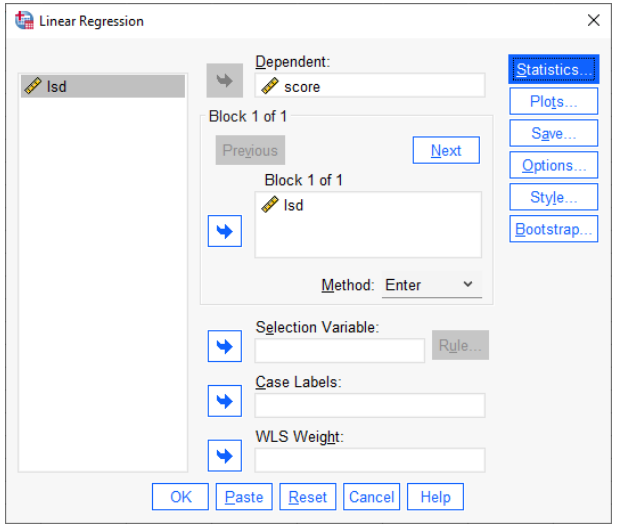
- $S_{xx} = \sum(X - \bar{X})^2$
- $S_{yy} = \sum(Y - \bar{Y})^2$
- $S_{xy} = \sum(X - \bar{X})(Y - \bar{Y})$
- $b_1 = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{S_{xy}}{S_{xx}}$
- $b_0 = \bar{Y} - b_1\bar{X}$
- $s^2 = \frac{\sum(Y-\hat{Y})^2}{n-2} = \frac{SSE}{n-2}$

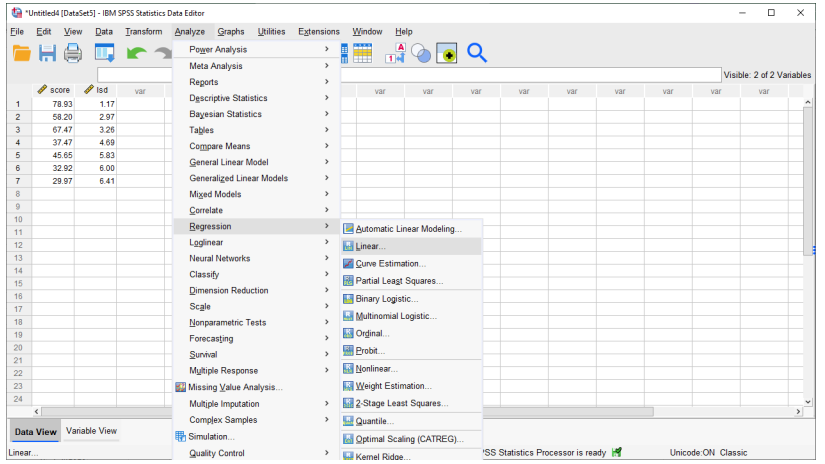
Example - Pharmacodynamics of LSD

- Column totals are in red
- Verify
  - $\bar{Y}$ = 50.087
  - $\bar{X}$ = 4.333
  - b_1= -202.49/ 22.47=-9.01
  - b_0 = 50.09 – (-9.01 x 4.33) = 89.10
  - $\hat{Y}$= 89.10-9.01X
  - s^2 = 50.72
- Let's do it in SPSS

Score (Y)	LSD Conc (X)	$X - \bar{X}$	$Y - \bar{Y}$	S_{xx}	S_{xy}	S_{yy}
78.93	1.17	-3.16	28.84	10.00	-91.23	831.92
58.20	2.97	-1.36	8.11	1.86	-11.06	65.82
67.47	3.26	-1.07	17.38	1.15	-18.65	302.17
37.47	4.69	0.36	-12.62	0.13	-4.50	159.19
45.65	5.83	1.50	-4.44	2.24	-6.64	19.69
32.92	6.00	1.67	-17.17	2.78	-28.62	294.71
29.97	6.41	2.08	-20.12	4.31	-41.78	404.69
350.61	**30.33**	**0.00**	**0.00**	**22.47**	**-202.49**	**2078.18**


Linear Regression: Statistics
Regression Coefficie...
Estimates
Confidence intervals
Level(%): 95
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside: 3 standard deviations
All cases
Continue
Cancel
Help


Linear Regression
lsd
Dependent:
score
Block 1 of 1
Previous
Next
Block 1 of 1
lsd
Method: Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help


File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Power Analysis
Meta Analysis
Regors
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Quantile...
Optimal Scaling (CATREG)...
Kernel Ridge...
Data View
Variable View

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.937[a]	.878	.853	7.12575

a. Predictors: (Constant), lsd

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1824.302	1	1824.302	35.928	.002[b]
	Residual	253.881	5	50.776		
	Total	2078.183	6			

a. Dependent Variable: score

b. Predictors: (Constant), lsd

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	89.124	7.048		12.646	<.001
	lsd	-9.009	1.503	-.937	-5.994	.002

a. Dependent Variable: score

$$Score = 89.125 - 9.009LSD$$

Example -Pharmacodynamics of LSD

H_o: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$

H_o: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1824.302	1	1824.302	35.928	.002[b]
	Residual	253.881	5	50.776		
	Total	2078.183	6			

a. Dependent Variable: score

b. Predictors: (Constant), lsd

n = 7; $b_1 = -9.01$; $s = \sqrt{50.72} = 7.12$; S*xx* – 22.475; se(b_1) = 7.12/22.475 = 1.50

Testing

t* = -9.01 / 1.50 = -6.01 → p = .002 → reject Null

Note, the *t* equals the *Z* only when N is large

Example -Pharmacodynamics of LSD

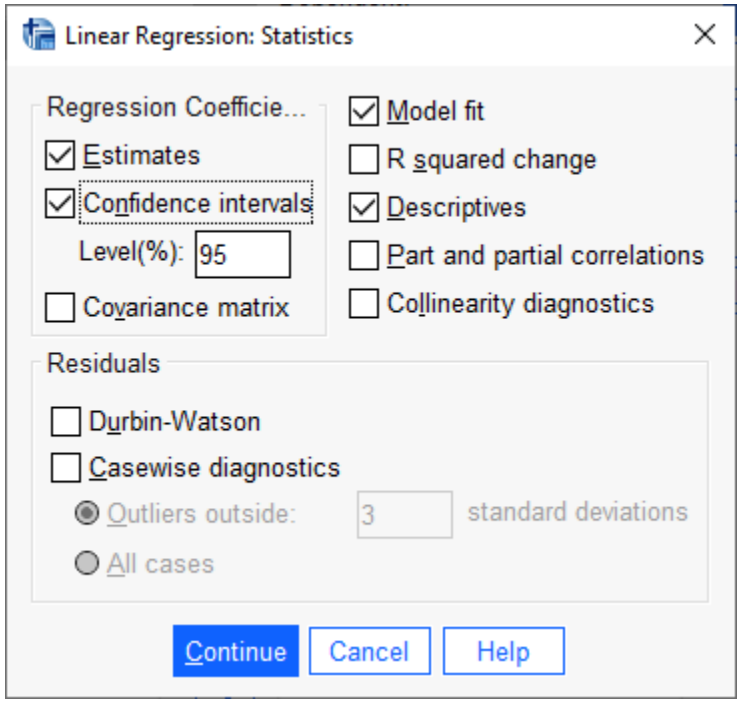
- n = 7
- b_1=-9.01
- se(b_1) = 7.12/22.475 = 1.50

- 95% CI
  - -9.01 $\pm$ 2.57(1.50) → -9.01$\pm$3.86 → (-12.87, -5.15)
  - From this confidence interval, do you reject the Null?

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)	89.124	7.048		12.646	<.001	71.008	107.240
	lsd	-9.009	1.503	-.937	-5.994	.002	-12.873	-5.146

a. Dependent Variable: score



Example

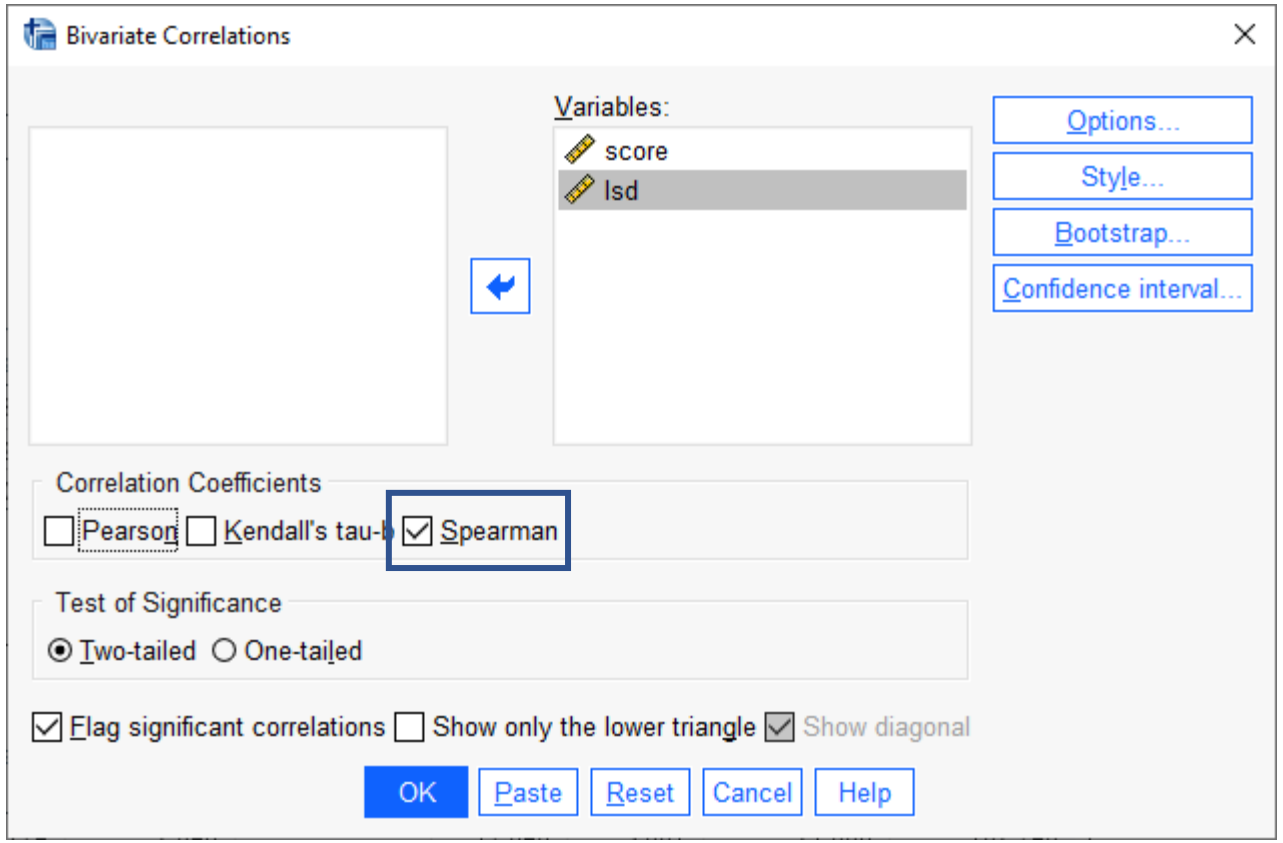
$$S_{xx} = 22.475 \quad S_{xy} = -202.487 \quad S_{yy} = 2078.183 \quad SSE = 253.89$$

$$r = \frac{-202.487}{\sqrt{(22.475)(2078.183)}} = -0.94$$

$$R^2 = \frac{2078.183 - 253.89}{2078.183} = 0.88 = (-0.94)^2$$

$R^2 \times 100 =$ % of variation in the dependent variable explained by the independent variable(s) in the mode

Example



Correlations

		score	lsd
score	Pearson Correlation	1	-.937**
	Sig. (2-tailed)		.002
	N	7	7
lsd	Pearson Correlation	-.937**	1
	Sig. (2-tailed)	.002	
	N	7	7

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			score	lsd
Spearman's rho	score	Correlation Coefficient	1.000	-.929**
		Sig. (2-tailed)	.	.003
		N	7	7
	lsd	Correlation Coefficient	-.929**	1.000
		Sig. (2-tailed)	.003	.
		N	7	7

**. Correlation is significant at the 0.01 level (2-tailed).

ANOVA F-test

- Analysis of Variance *F-test*
- Tests the significance of the overall model including ALL independent variables
- For SLR, the F-test is the same as the t-test for b1
- $F = t^2$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Model	SSR	1	MSR = SSR/1	F = MSR/ MSE
Error	SSE	n-2	MSE = SSE/(n-2)	
Total	S_{yy}	n-1		

ANOVA F-test for SLR

- Analysis of Variance *F-test*
  - H_o: $\beta_1 = 0$
  - H_1: $\beta_1 \neq 0$
- $F^* = \frac{MSR}{MSE} > F_{\alpha, df=1, n-2}$ → *reject null*

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Model	SSR	1	MSR = SSR/1	*F* = MSR/ MSE
Error	SSE	*n*-2	MSE = SSE/(n-2)	
Total	S_{yy}	*n*-1		

Example - Pharmacodynamics of LSD

- Total Sum of squares:

$$S_{yy} = \sum (y_i - \bar{y})^2 = 2078.183 \qquad df_{Total} = 7 - 1 = 6$$

- Error Sum of squares:

$$SSE = \sum (y_i - \hat{y}_i)^2 = 253.890 \qquad df_{Error} = 7 - 2 = 5$$

- Model Sum of Squares:

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = 2078.183 - 253.890 = 1824.293 \qquad df_{Model} = 1$$

Example - Pharmacodynamics of LSD

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Model	1824.293	1	1824.293	35.93
Error	253.890	5	50.778	
Total	**2078.183**	**6**		

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1824.302	1	1824.302	35.928	.002[b]
	Residual	253.881	5	50.776		
	Total	2078.183	6			

a. Dependent Variable: score

b. Predictors: (Constant), lsd

Analysis of Variance - *F*-test

H_0: all betas equal 0 vs H_A: at least one beta is 0

H_0: $\beta_1 = 0$ vs H_A: $\beta_1 \neq 0$

$$F = \frac{MSR}{MSE} = 35.928 > F_{\alpha, df=1, n-2} = 6.61 \rightarrow \text{Reject Null}$$

Fitted Values & Residuals

- Fitted values for the sample cases are obtained by substituting the appropriate X values into the estimated regression function.

$Score = 89.125 - 9.009(4.69)$=46.87279

- The residual is the difference between the observed value Yi and the fitted value $\hat{Y}_i$. The residual is denoted by $e_i$and is defined as

$e_i = Y_i - \hat{Y}_i$

- For the case above

$e_4 = Y_4 - \hat{Y}_4 = 37.47 - 46.87 = -9.4$

Score (Y)	LSD Conc (X)
78.93	1.17
58.20	2.97
67.47	3.26
37.47	4.69
45.65	5.83
32.92	6.00
29.97	6.41
350.61	**30.33**