

Week 12

Stats I

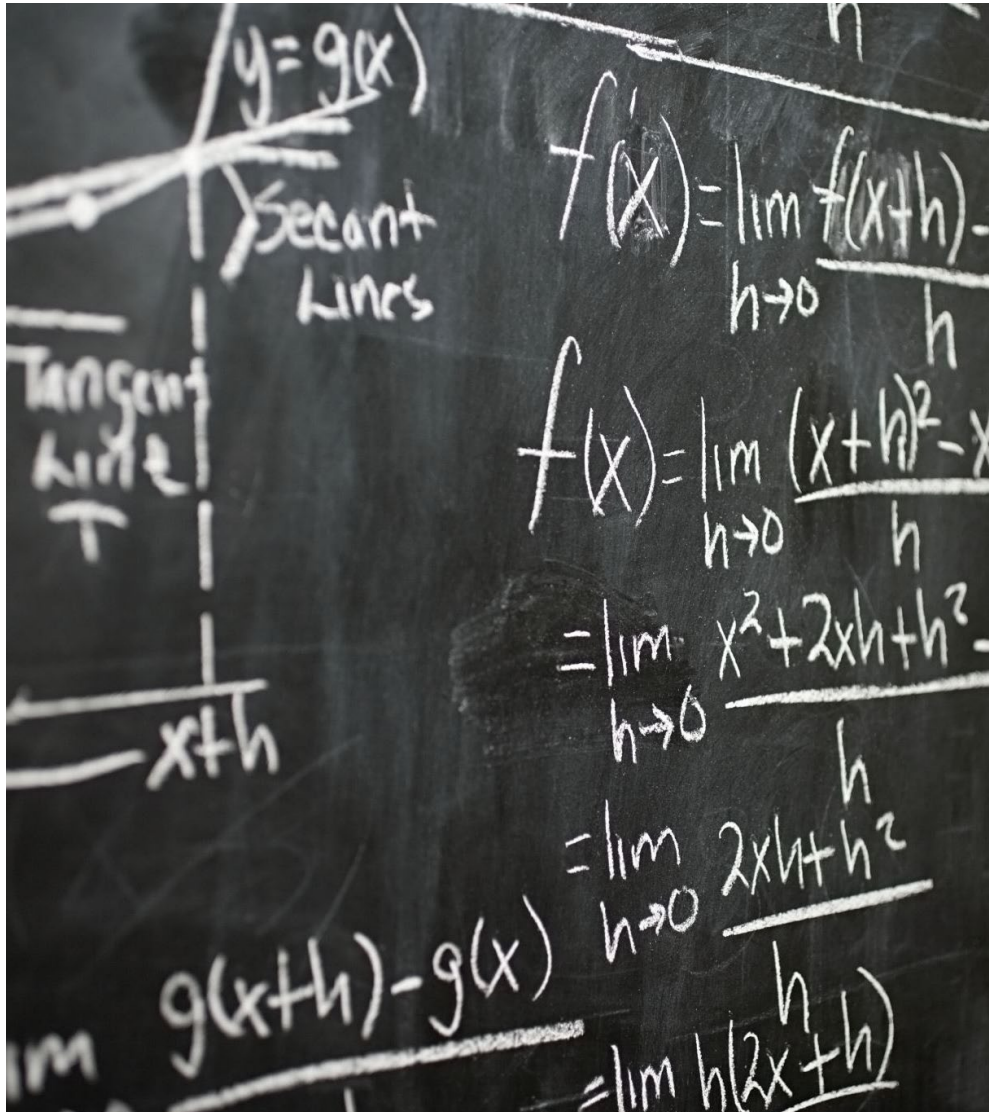
Multiple Linear Regression

Intro to Logistic Regression

Introduction to Multiple Linear Regression

1. Introduction to Multiple Linear Regression
2. Application & Example
3. Simple Regression → Multiple Regression
4. Adjusted R^2
5. Multicollinearity & model assumptions
6. More 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. Are the OLS assumptions satisfactory?

Multiple Linear Regression

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

- **Linear Model** is now:

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

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

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

- Example: $p = 3$
- Note that the fitting is still done with Ordinary Least Squares (OLS) method just like in simple linear regression.

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

Similarities/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 $H_0: \beta_1 = \beta_2 = \beta_3 \dots \beta_p = 0$ (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” What is your intuition about the nature of the relationships in this model and *why*?
 - 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 relationships (if any!).

Variable definitions in the school admissions dataset: **admissions.sav**

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

■ ■ ■

Analyze → Regression

Linear Regression

Dependent: SAT_math

Block 1 of 1

Independent(s): SAT_verbal, HS_class_size

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

OK Paste Reset Cancel Help

Linear Regression: Save

Predicted Values

☐ Unstandardized

☐ Standardized

☐ Adjusted

☐ S.E. of mean predictions

Residuals

☐ Unstandardized

☒ Standardized

☐ Studentized

☐ Deleted

☐ Studentized deleted

Influence Statistics

☐ DfBetas

☐ Standardized DfBetas

☐ DfFits

☐ Standardized DfFits

☐ Covariance ratios

Distances

☐ Mahalanobis

☐ Cook's

☐ Leverage values

Prediction Intervals

☐ Mean ☐ Individual

Confidence Interval: 95 %

Coefficient statistics

☐ Create coefficient statistics

☒ Create a new dataset

Dataset name:

☐ Write a new data file

File...

Export model information to XML file

Browse...

☒ Include the covariance matrix

Continue Cancel Help

Linear Regression: Statistics

Regression Coefficient...

☒ 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: Plots

DEPENDENT

*ZPRED

*ZRESID

*DRESID

*ADJPRED

*SRESID

*SDRESID

Scatter 1 of 1

Y: *ZRESID

X: *ZPRED

Standardized Residual Plots

☒ Histogram

☒ Normal probability plot

☐ Produce all partial plots

Continue Cancel Help

What you already know...Descriptive Statistics & Relationships

Descriptive Statistics

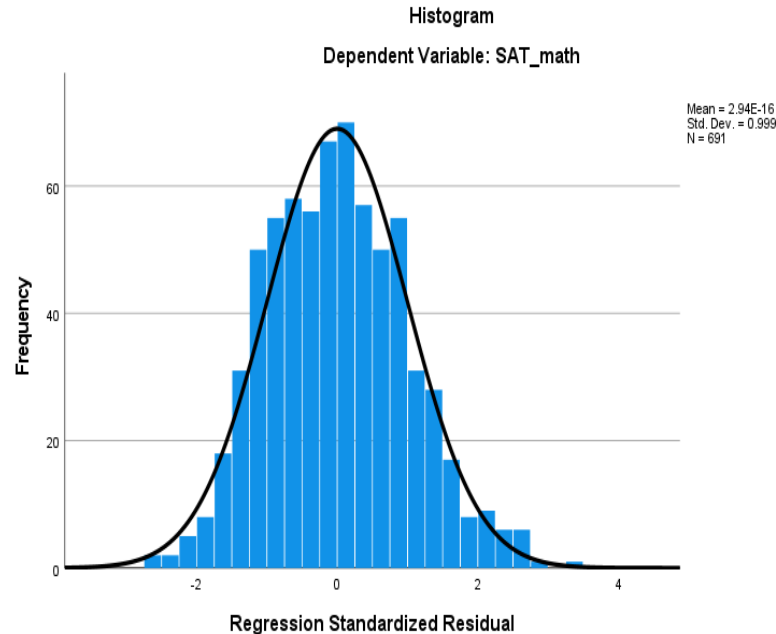
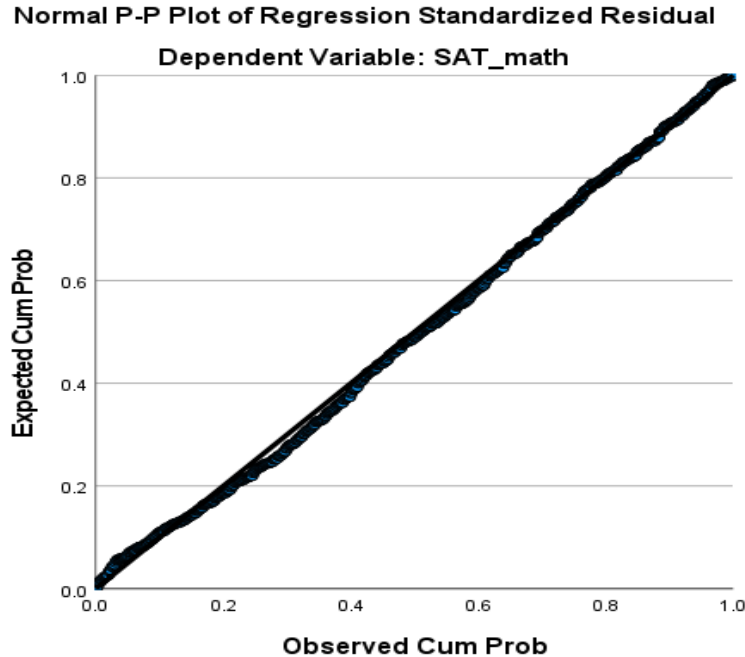
	Mean	Std. Deviation	N
SAT_math	585.007	68.7157	691
SAT_verbal	561.216	67.8124	691
HS_class_size	316.33	178.043	691

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

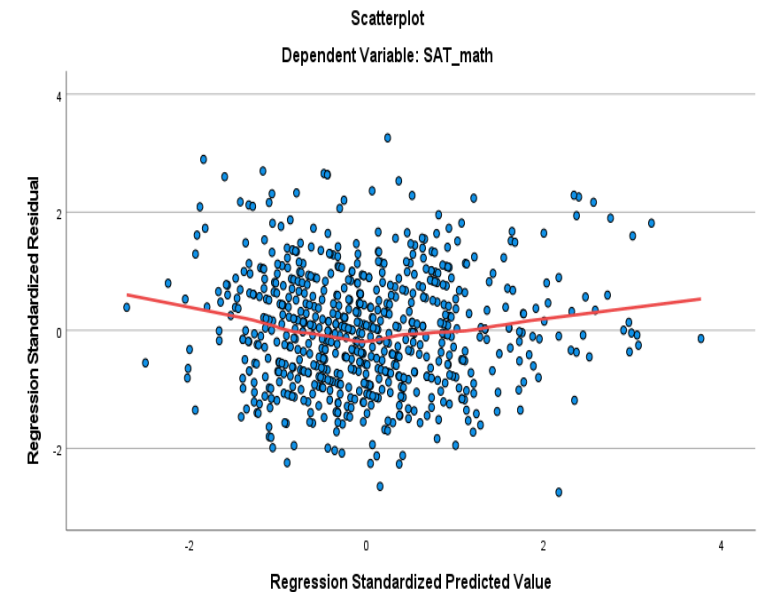
Correlations

		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

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

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

Small p -value \rightarrow the model is
significant

H_0 : 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								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		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: $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$ Here, $k = 2 \rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

MATH SAT Score = $314.147 + .449 \text{VERBAL_SAT}(X_1) + .060 \text{CLASS_SIZE}(X_2)$

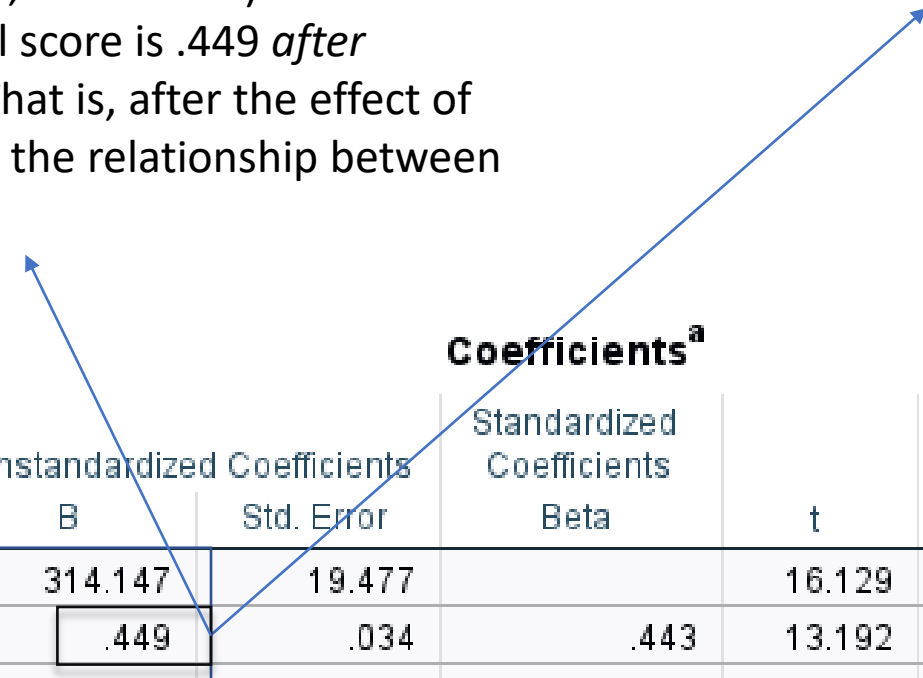
Confidence intervals for the coefficients

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



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			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: $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$ Here, $k = 2 \rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

MATH SAT Score = $314.147 + .449 \text{VERBAL_SAT}(X_1) + .060 \text{CLASS_SIZE}(X_2)$

(aside): part and partial correlation

- Partial correlation: the correlation between an independent variable and a dependent variable after controlling for the influence of other variables on **both** the independent and dependent variable
 - **Example:** regress MATH SAT score on VERBAL SAT score and class size – the partial correlation the influence of class size on both MATH and VERBAL SAT score is taken into consideration. This means that the partial correlation between VERBAL SAT score and MATH SAT score takes into account the impact of class size on both VERBAL and MATH SAT score
 - This is the idea of ‘partialing’ out in regression
- Part correlation: the correlation between an independent and dependent variable after controlling for the influence of other variables **only on the independent variable**
 - **Example:** the part correlation above only considers the impact of class size on VERBAL SAT score
 - This is nice if you want to assess how much unique variance CLASS SIZE explains in relation to the total variance in MATH SAT score

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

Correlations			SAT_math	SAT_verbal	HS_class_size
Control Variables					
-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. To test this, 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, under what scenarios is the feminist 'correct'?
- 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 believed to be the problem of too much correlation between independent variables...
- 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 compared to the model with just educational attainment
 - Might conclude that job type is highly influential on income, while educational attainment is unimportant (or vice versa) which may not be true!

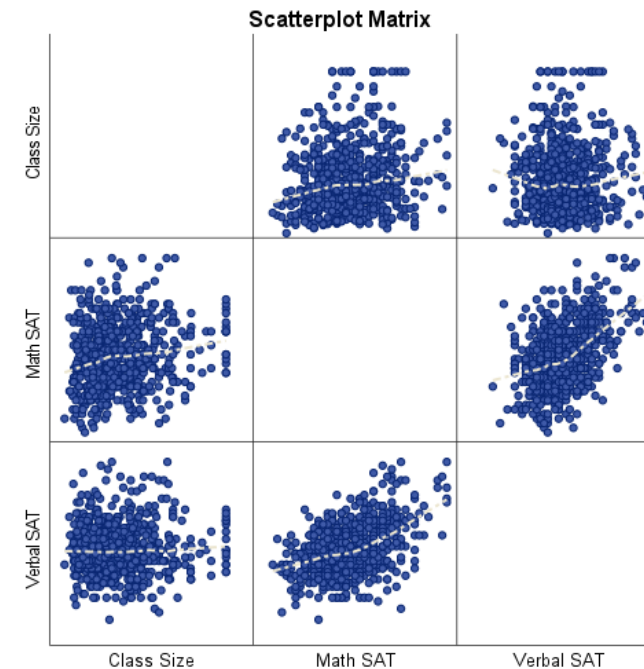
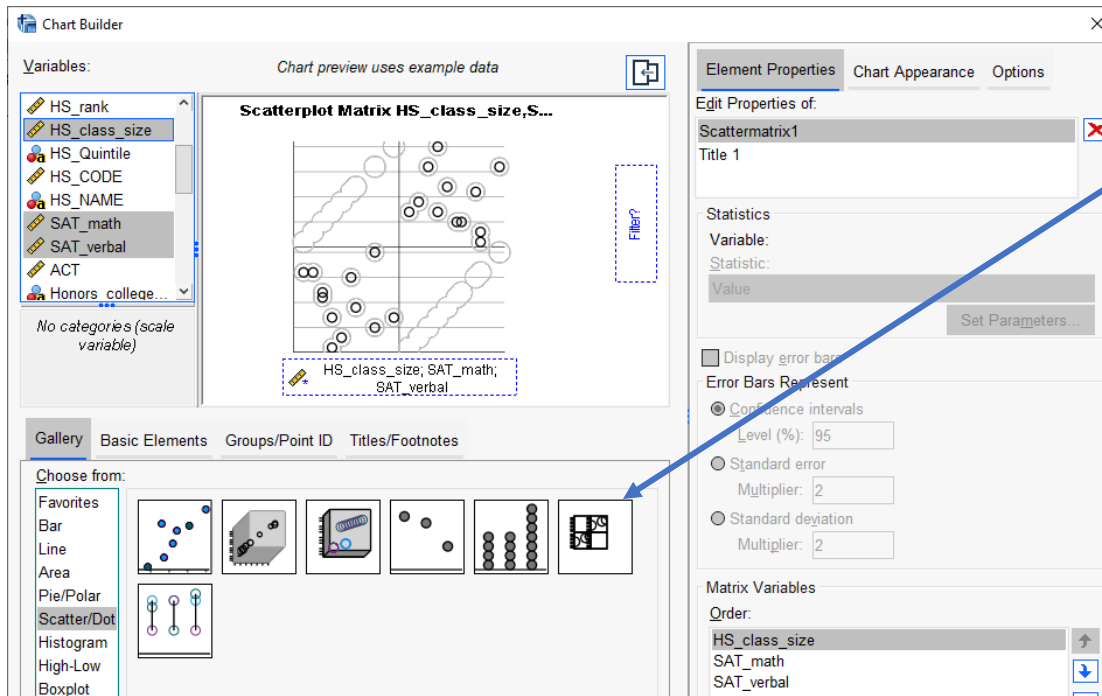
- This happened in our paper for SSRW when we included an index of redlining, concentrated disadvantage and concentrated affluence in the model
- **Note** that while there is substantial overlap between variables theoretically, they are quite distinguishable
 - The opposite of concentrated disadvantage 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 may not be a good idea to use both in the regression equation. Why?
 1. Many correlated variables hurts the 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 partialing 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!

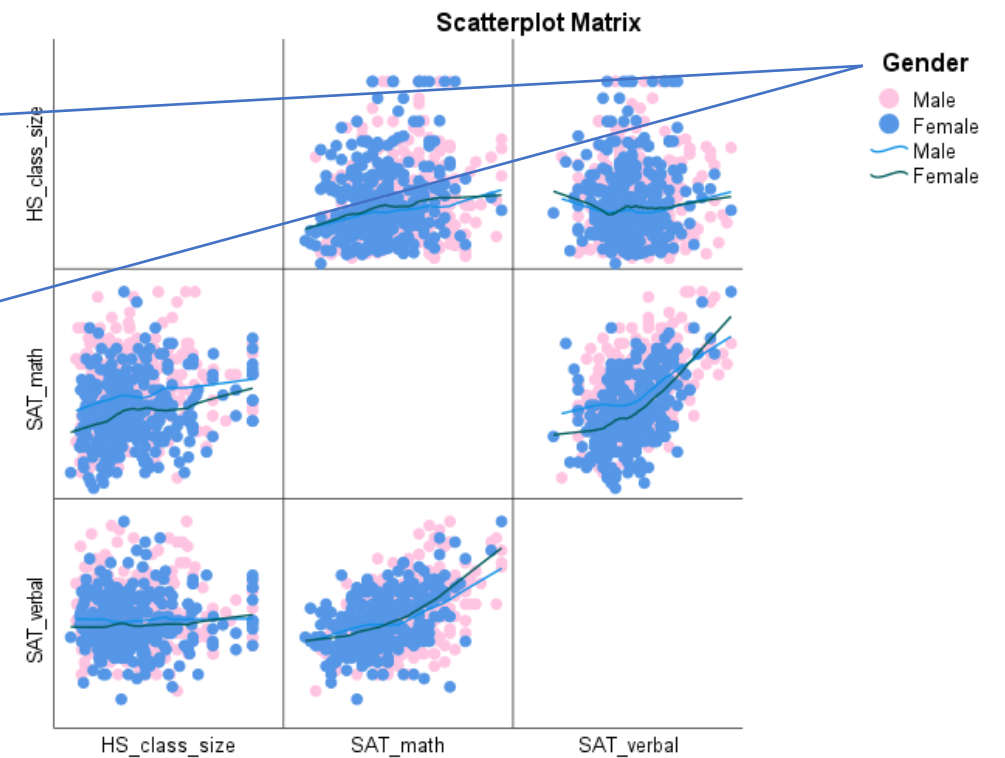
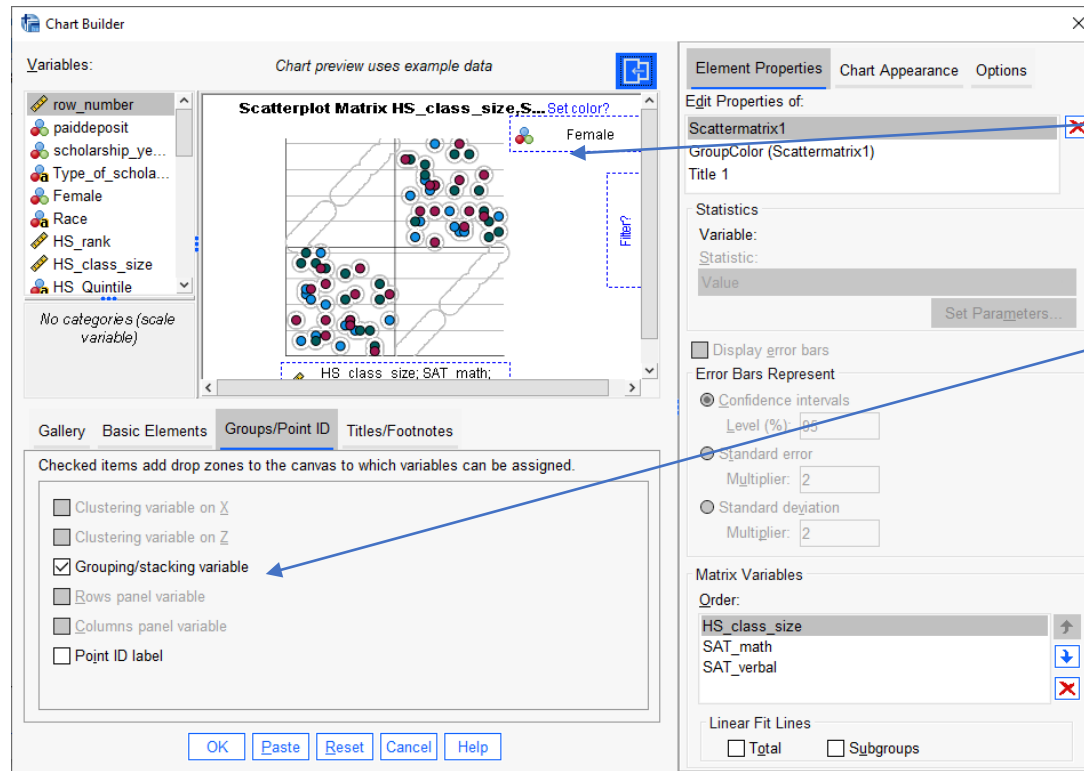
How to check for multicollinearity?

- 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
 - Look at descriptive statistics
 - Calculate the zero order and partial correlations between the variables
 - Run regression analysis
 - Interpret coefficients
 - 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

Percentile Values

☐ Quartiles

☐ Cut points for: 10 equal groups

☒ Percentile(s):

Add
Change
Remove

20.0
40.0
60.0
80.0
100.0

Central Tendency

☒ Mean

☐ Median

☐ Mode

☐ Sum

☐ Values are group midpoints

Dispersion

☒ Std. deviation ☐ Minimum

☐ Variance ☐ Maximum

☐ Range ☐ S.E. mean

Distribution

☐ Skewness

☐ Kurtosis

Continue
Cancel
Help

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

X1 = GRE score

X2 = 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)

		Correlations		
		GRE Score	undergraduate GPA	graduate school GPA
GRE Score	Pearson Correlation	1	.301	.784**
	Sig. (2-tailed)		.368	.004
	N	11	11	11
undergraduate GPA	Pearson Correlation	.301	1	.752**
	Sig. (2-tailed)	.368		.008
	N	11	11	11
graduate school GPA	Pearson Correlation	.784**	.752**	1
	Sig. (2-tailed)	.004	.008	
	N	11	11	11

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

Y = GPA in graduate school
X1 = GRE score
X2 = Undergraduate GPA

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

Correlations						
Variable	Variable2	Correlation	Count	Statistic		Notes
				Lower C.I.	Upper C.I.	
gradgpa	undergpa	.752	11	.276	.932	
	gradgpa	1.000	11	--	--	
	gre	.784	11	.349	.941	
gre	undergpa	.301	11	-.365	.763	
	gradgpa	.784	11	.349	.941	
	gre	1.000	11	--	--	
undergpa	undergpa	1.000	11	--	--	
	gradgpa	.752	11	.276	.932	
	gre	.301	11	-.365	.763	

Missing value handling: PAIRWISE, EXCLUDE. C.I. Level: 95.0

$r_{Y.X1} = .784$ 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

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

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

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

$$Y_i = b_0 + b_1X_1 + b_2X_2 + e_i$$

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

- 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:
 - $\hat{Y}_i = .0125 (130) + .4687 (3.5000) + .6337 = 3.8992$

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 not all the partial slopes are equal to zero at the .05 level of significance, i.e., at least one is significant

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 R^2 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
- Redefining 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: $\hat{y} = b_0 + b_1x + b_2x^2$

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.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 \rightarrow \hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$

New variable, product
of x_1 and x_2

- Different lines for different levels of a categorical variable

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:

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

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

Additive
Correction

Example: Recoding dummy variables

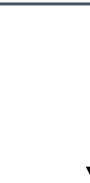
Use the school admissions data variable (admissions.sav) for whether a student has a family member that is an alumni (Number_of_family_alumni) and recode it as a simple binary variable (Y/N).

Label the groups with no family alum = 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		Standardized Coefficients	Sig.
		B	Std. Error	Beta	
1	(Constant)	68.891	21.512		.001
	SAT_math	.007	.036	.008	.837
	alumniYN	-2.604	5.985	-.017	.664

a. Dependent Variable: HS_rank

Regression Equation:
 $\text{Rank} = 68.89 + .007 * \text{SAT_math} - 2.6 * (\text{IF ALUM})$

Interpretation: the estimated difference in HS rank between students who have family that is an alum 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:

$$\text{Rank} = 68.89 + .007 * \text{SAT_math} - 2.6 * (\text{IF ALUM})$$

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

Regression Equation:

$$\text{Rank} | \text{Alum} = 0: 68.89 + .007 * \text{SAT_math}$$

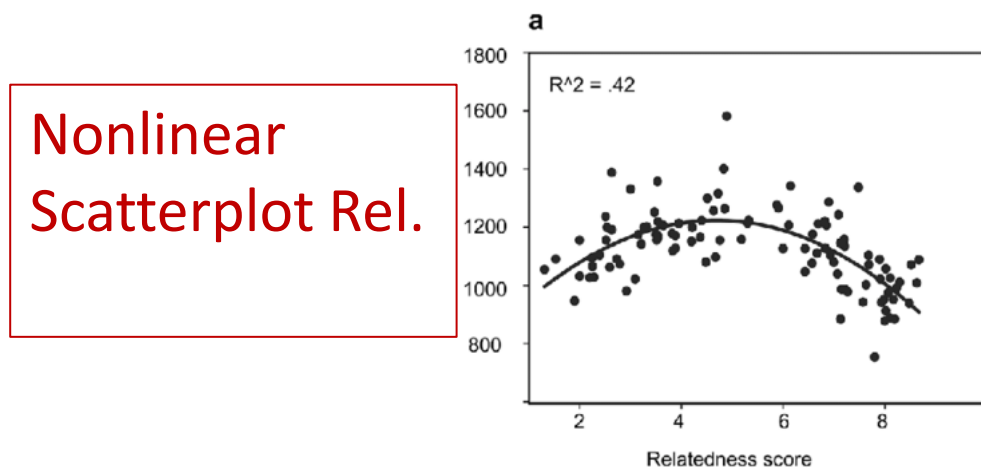
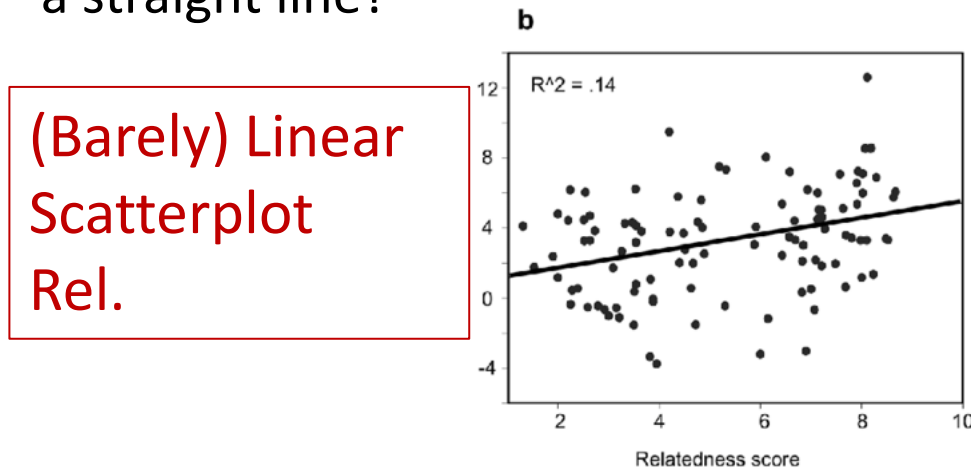
$$\hat{y}|(x_2=1) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2$$

Regression Equation:

$$\text{Rank} | \text{Alum} = 1: = 68.89 + .007 * \text{SAT_math} - 2.6 * (1)$$

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?



- One Solution: 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 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 model's 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...

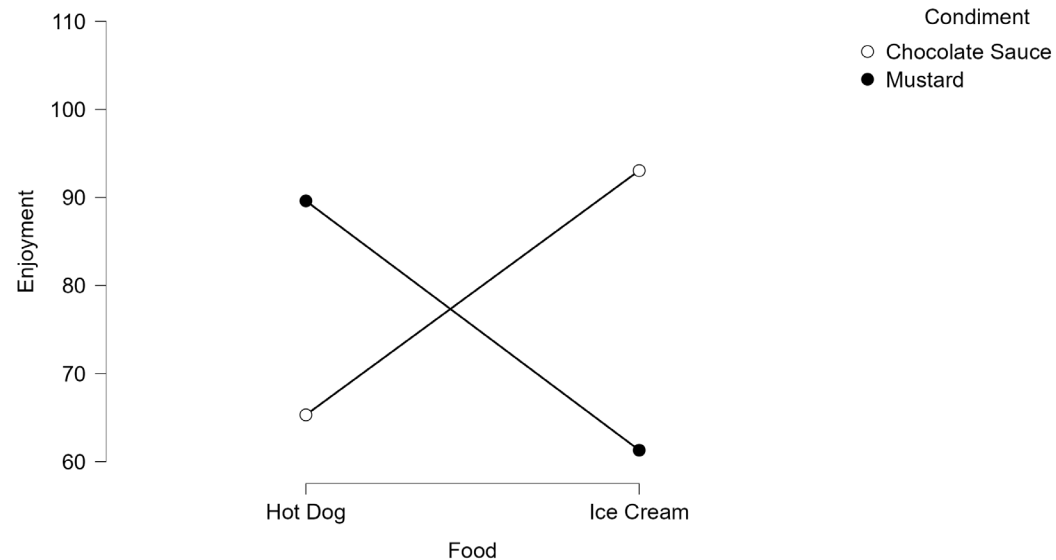
Interaction Terms

- 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: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \hat{\beta}_3x_1x_2$
- 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, models 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.



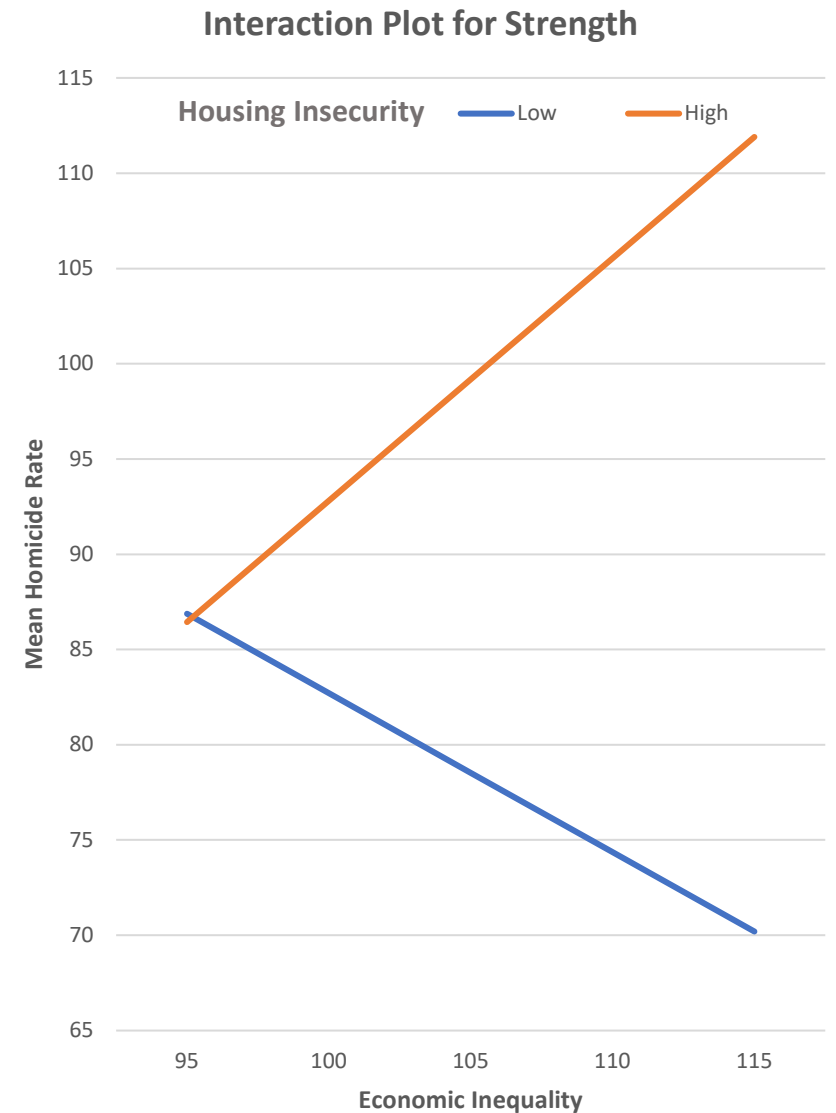
Enjoyment levels are higher for chocolate sauce when the food is ice cream but higher for mustard when the food is a hot dog.

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



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

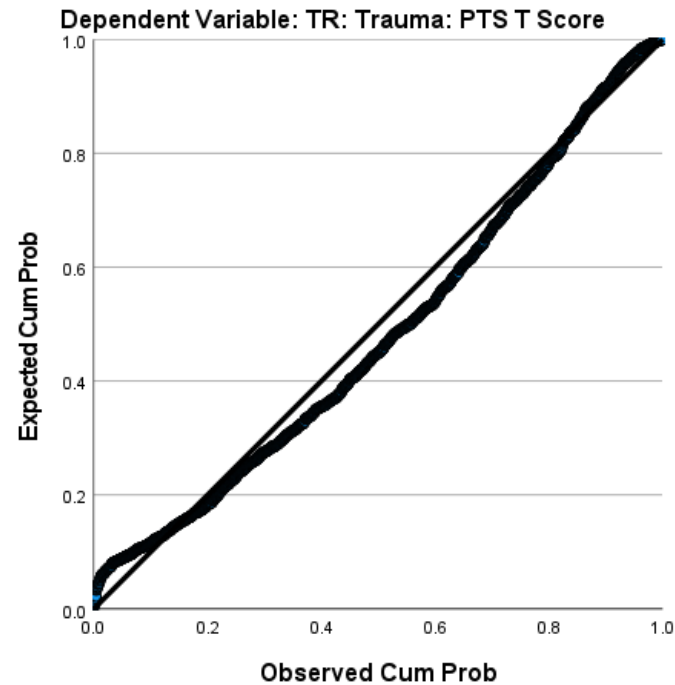
ANOVA^a

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

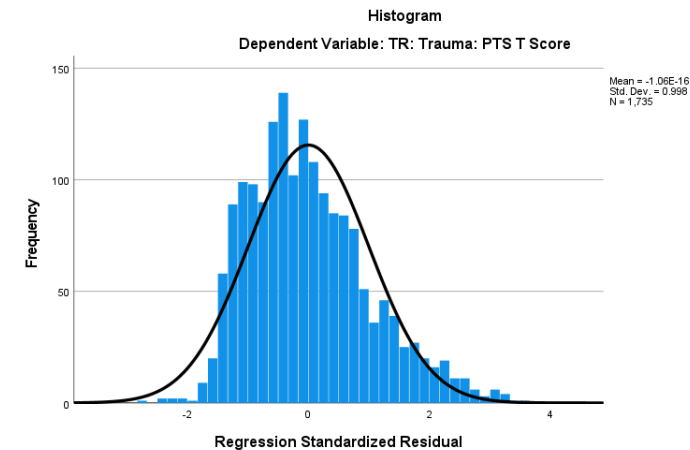
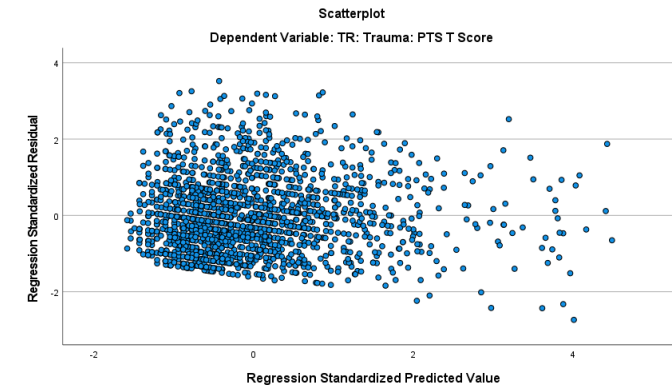
Normal P-P Plot of Regression Standardized Residual



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

Older children with higher levels of violence exposure have more PTS symptoms compared to younger children

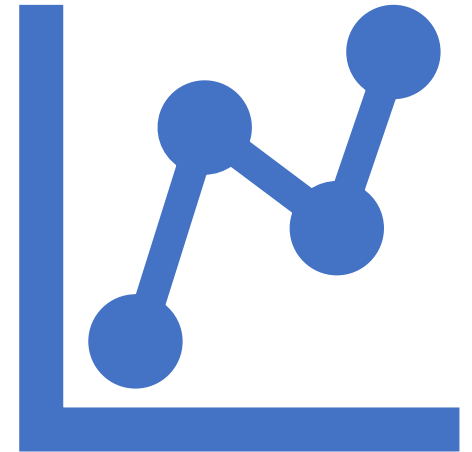
- Use nscaw_sample.sav

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	50.358	2.110		23.869	<.001
	Child age in years (chAge_b)	-.326	.154	-.063	-2.109	.035
	Child gender (chgendr)	1.080	.534	.048	2.022	.043
	Child OOH Situation (YN)	.131	.601	.005	.218	.828
	nhw	-.151	.521	-.007	-.291	.771
	physab	1.070	.619	.042	1.730	.084
	sexab	1.871	.704	.066	2.658	.008
	EV: Severe Violence Total # of Exposure	2.556	.977	.307	2.617	.009
	ageXEV	-.024	.084	-.034	-.287	.774

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

More diagnostics

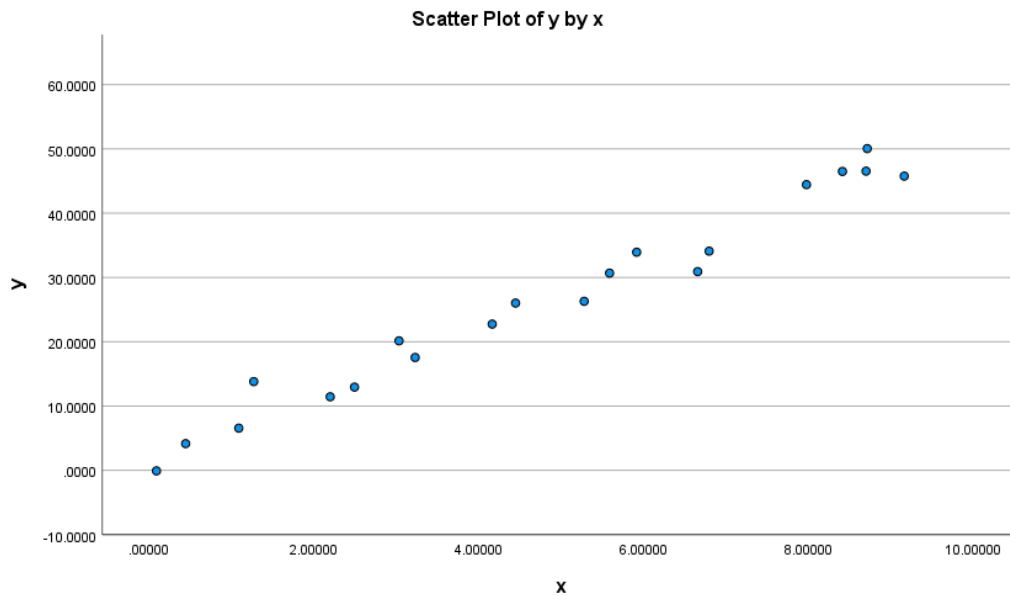
- An outlier is a data point whose response y does not follow the general trend of the rest of the data
 - It is defined as a point that is $1.5(Q3-Q1) = 1.5IQR$
- A data point has high leverage if it has "extreme" X values
 - With a single predictor, an extreme X value is simply one that is particularly 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



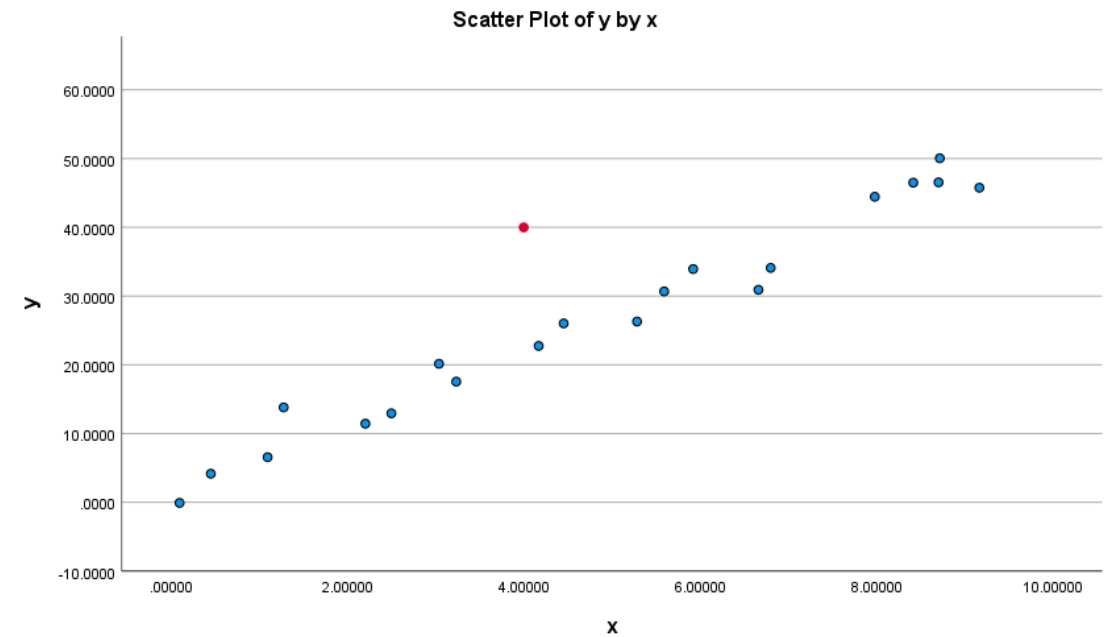
Outliers and unusual values

- 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	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	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

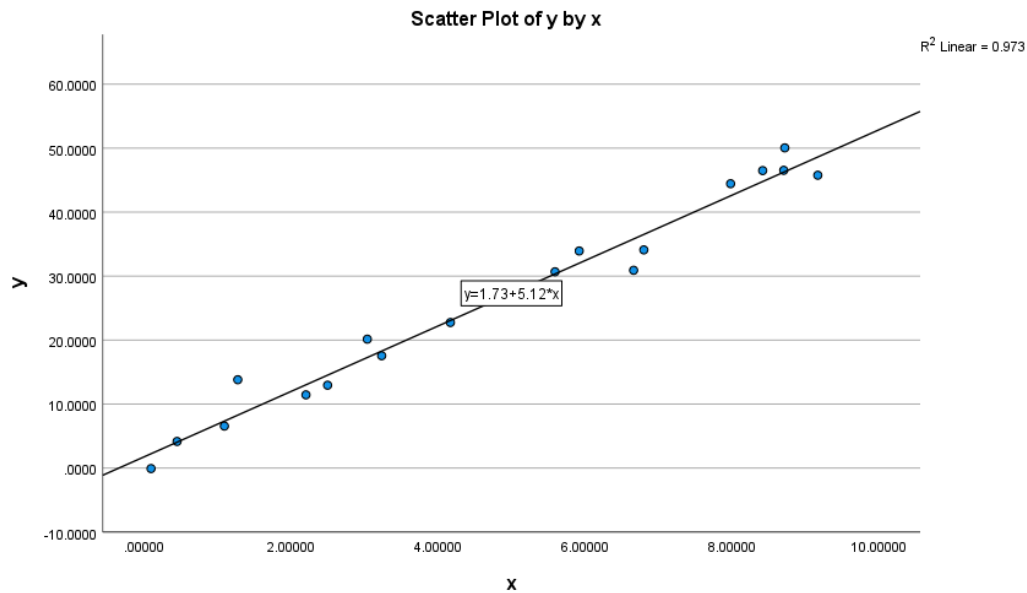
Why is the standard error bigger?

Compare & Intuit

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

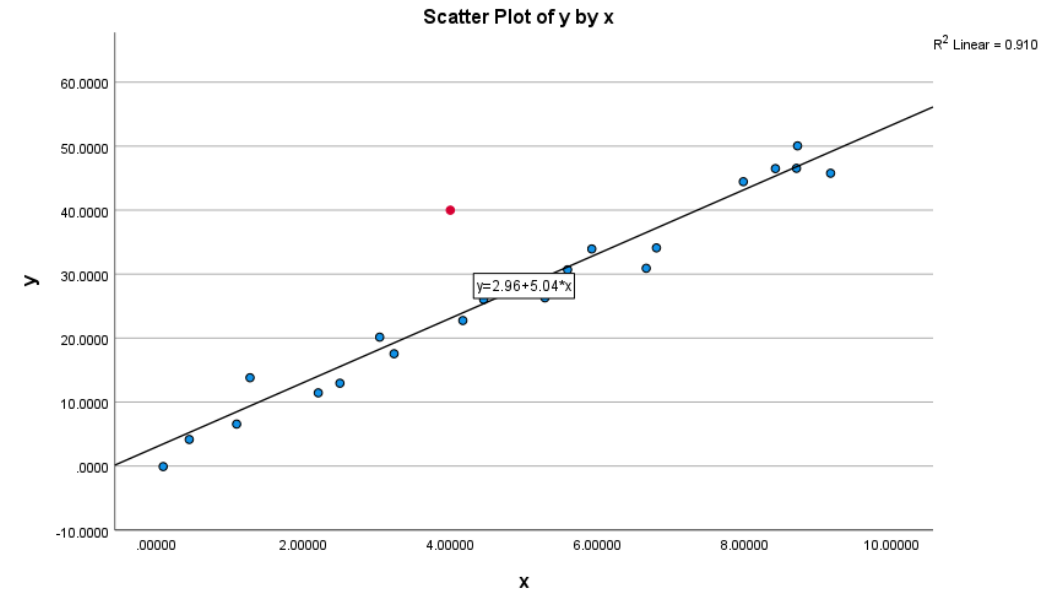
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.732	1.121	1.546	.140	-.622	4.086
	x	5.117	.200	25.551	<.001	4.696	5.538

a. Dependent Variable: y



Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2.958	2.009	1.472	.157	-1.247	7.163
	x	5.037	.363	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

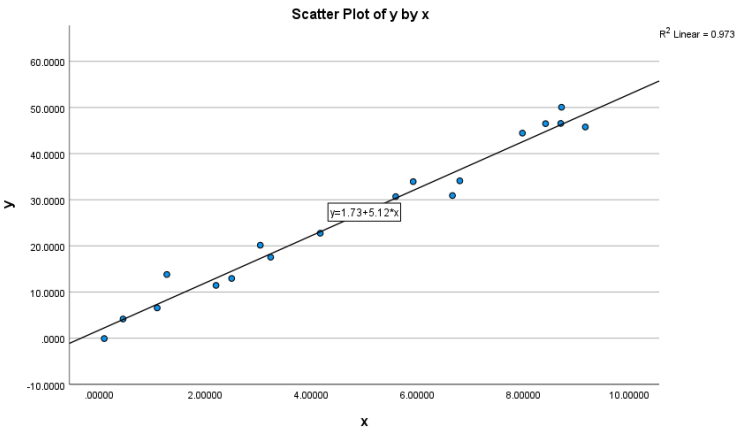
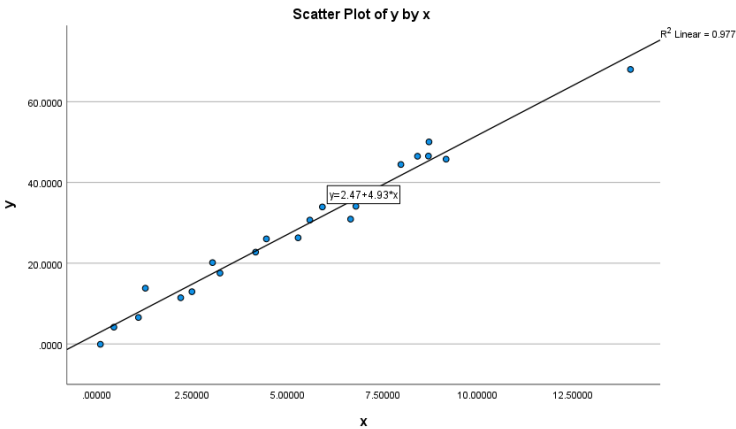
ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6028.817	1	6028.817	821.444	<.001 ^b
	Residual	139.446	19	7.339		
	Total	6168.263	20			

a. Dependent Variable: y

b. Predictors: (Constant), x

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

a. Dependent Variable: y



Unusual values and influential points

- 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

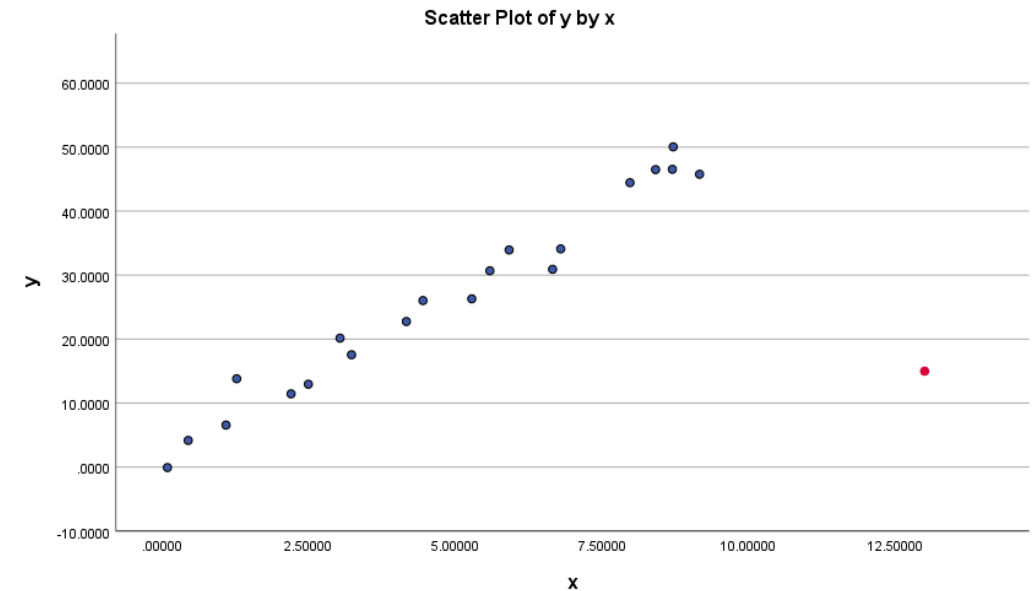
a. Predictors: (Constant), x

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error				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

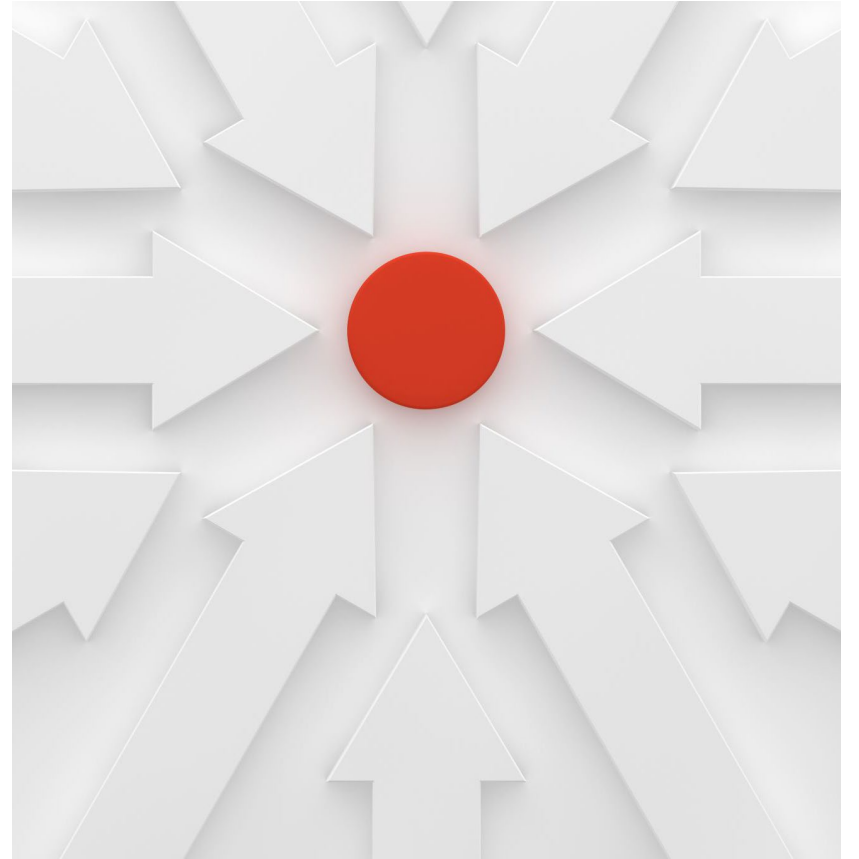
a. Dependent Variable: y

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:

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

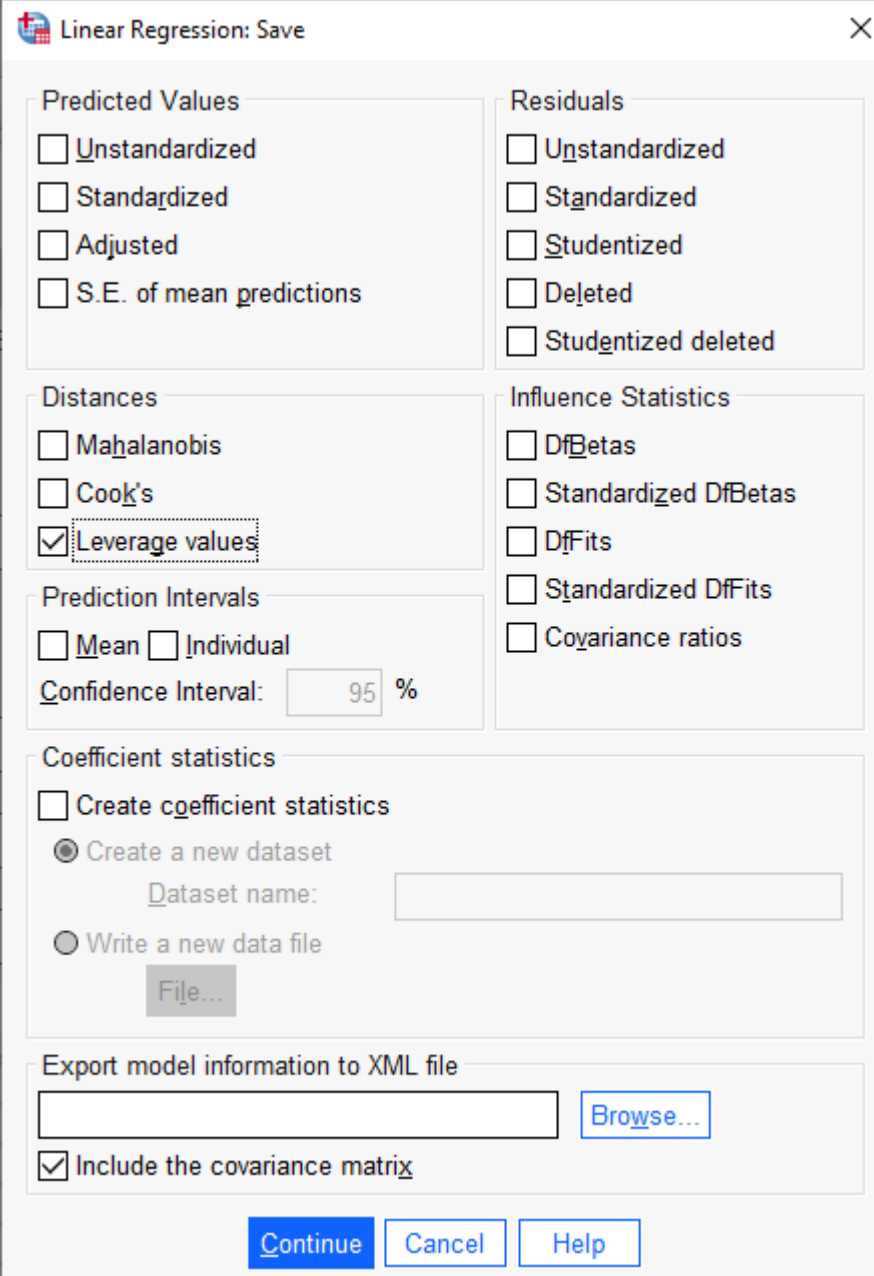
Then flag the observations as
unusual \rightarrow X is an observation
whose value gives it large *leverage*
for the regression analysis

Example

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

- $p = 2, n = 21$

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

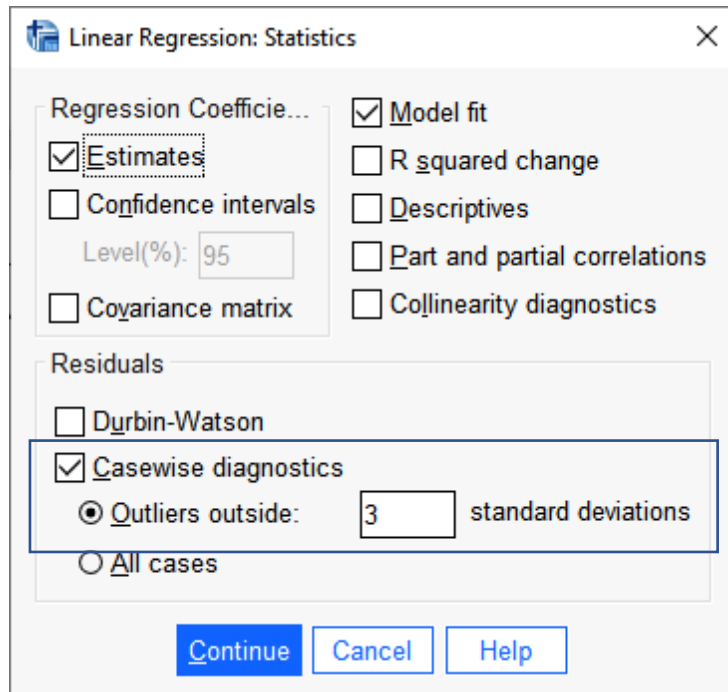


The image shows a 'Linear Regression: Save' dialog box with the following sections and options:

- Predicted Values:**
 - ☐ Unstandardized
 - ☐ Standardized
 - ☐ Adjusted
 - ☐ S.E. of mean predictions
- Residuals:**
 - ☐ Unstandardized
 - ☐ Standardized
 - ☐ Studentized
 - ☐ Deleted
 - ☐ Studentized deleted
- Distances:**
 - ☐ Mahalanobis
 - ☐ Cook's
 - ☒ Leverage values
- Influence Statistics:**
 - ☐ DfBetas
 - ☐ Standardized DfBetas
 - ☐ DfFits
 - ☐ Standardized DfFits
 - ☐ Covariance ratios
- Prediction Intervals:**
 - ☐ Mean ☐ Individual
 - Confidence Interval: %
- Coefficient statistics:**
 - ☐ Create coefficient statistics
 - ☒ Create a new dataset
 - Dataset name:
 - ☐ Write a new data file
 - File...
- Export model information to XML file:**
 -
 -
 - ☒ Include the covariance matrix

Buttons at the bottom:

Can have SPSS provide outliers



The image shows the 'Linear Regression: Statistics' dialog box in SPSS. The 'Regression Coefficients' section has 'Estimates' checked. The 'Residuals' section has 'Casewise diagnostics' checked, with 'Outliers outside: 3 standard deviations' selected. The 'Continue' button is highlighted.

Linear Regression: Statistics

Regression Coefficients

- ☒ Estimates
- ☐ Confidence intervals
- Level(%): 95
- ☐ Covariance matrix

Model fit

- ☒ Model fit
- ☐ R squared change
- ☐ Descriptives
- ☐ Part and partial correlations
- ☐ Collinearity diagnostics

Residuals

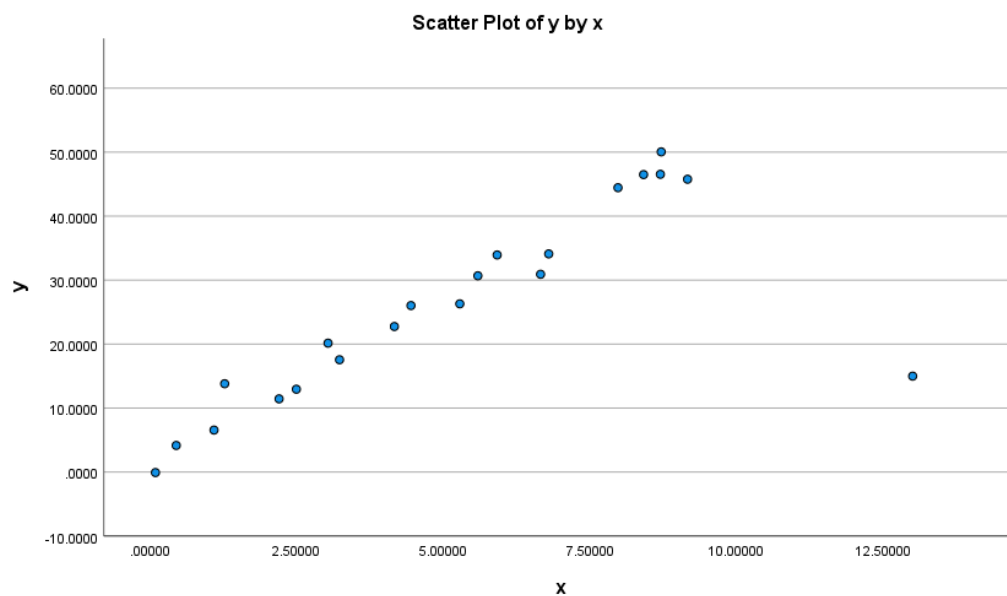
- ☐ Durbin-Watson
- ☒ Casewise diagnostics
 - ☒ Outliers outside: 3 standard deviations
 - ☐ All cases

Continue Cancel Help

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

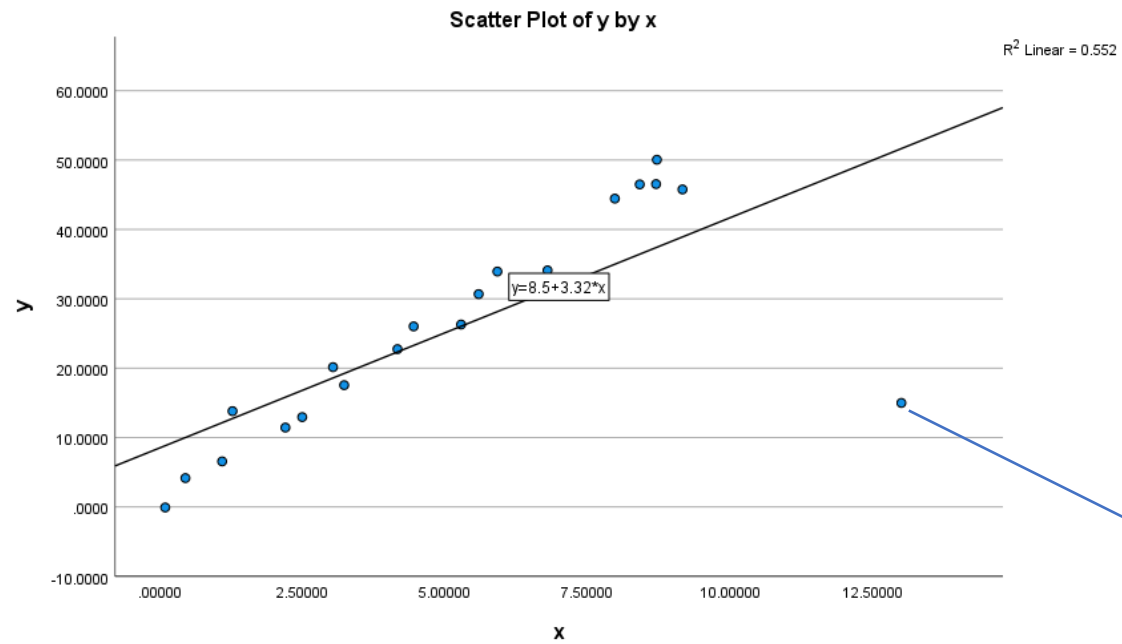
ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2553.978	1	2553.978	23.406	<.001 ^b
	Residual	2073.233	19	109.118		
	Total	4627.211	20			

a. Dependent Variable: y

b. Predictors: (Constant), x

Coefficients ^a						
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	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

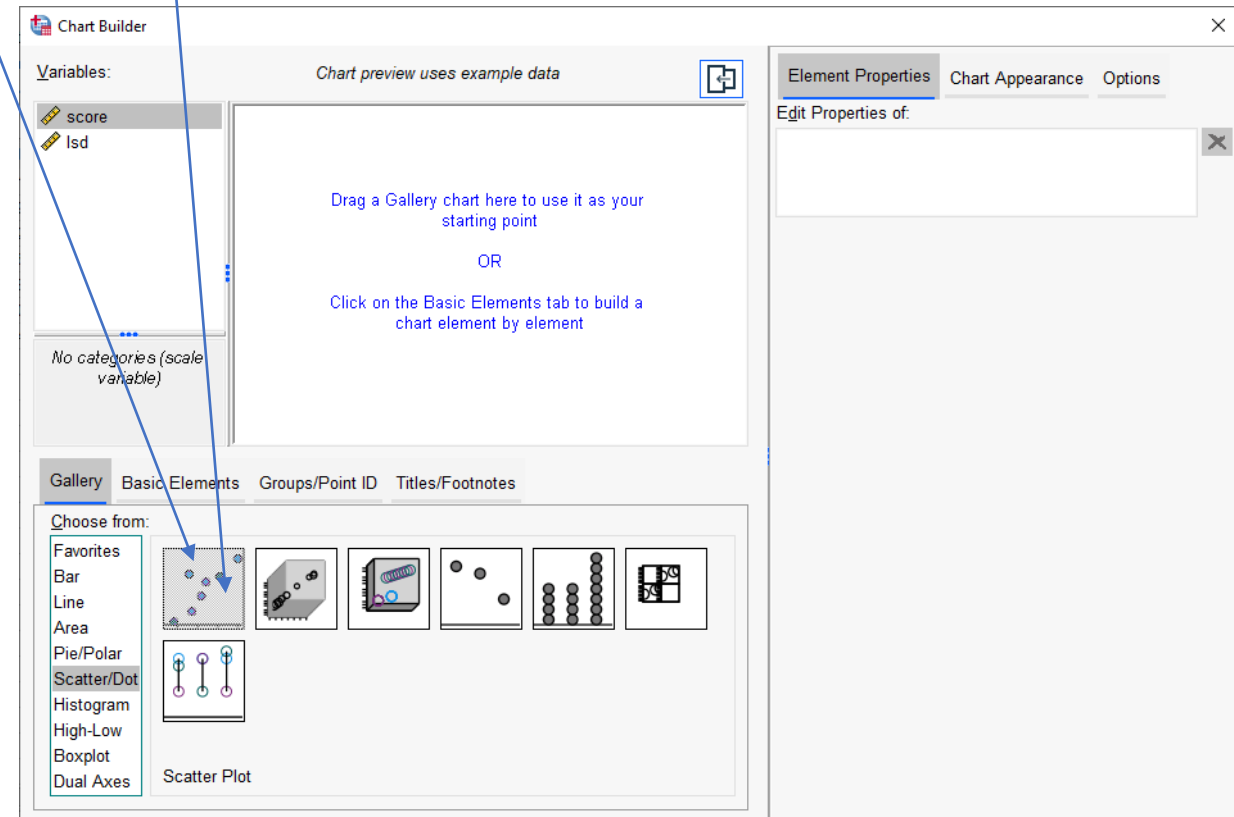
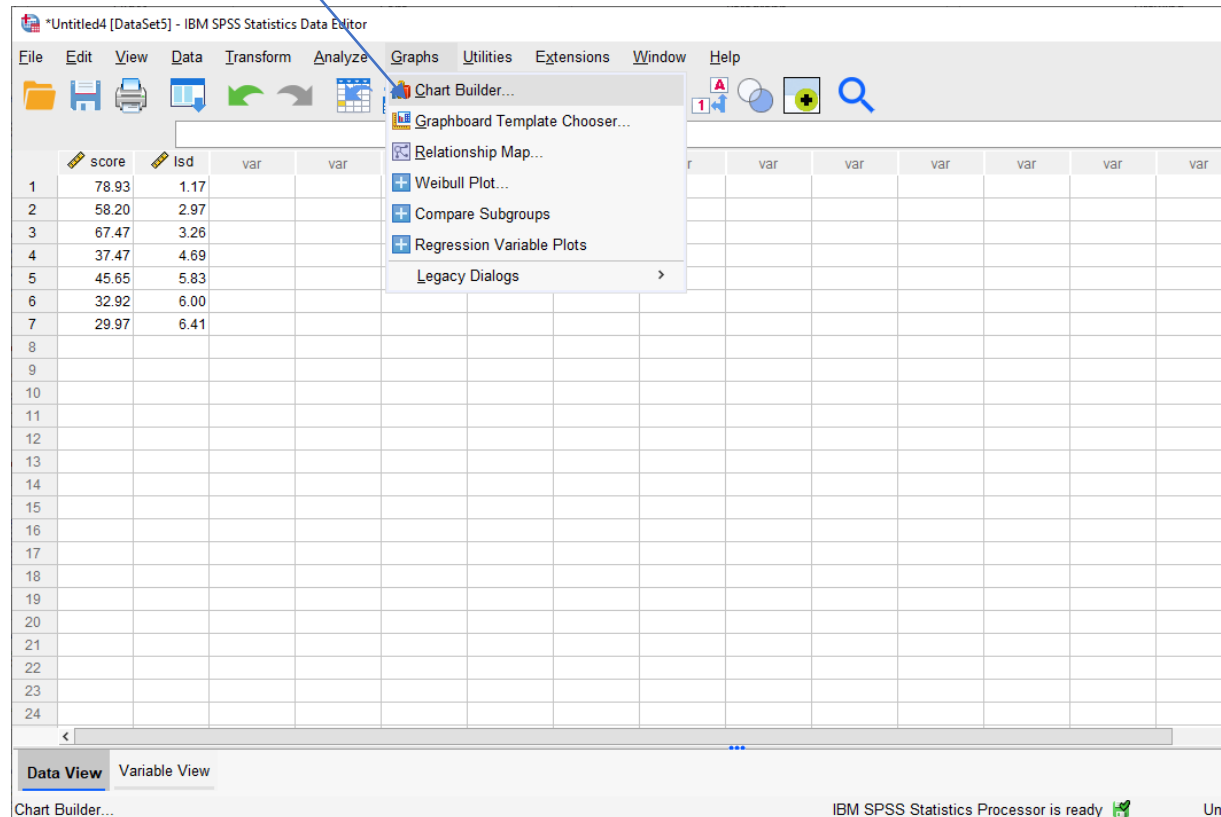
- Response (Y) - Math score (mean among 5 volunteers)
- Predictor (X) - LSD tissue concentration (mean of 5 volunteers)
- Make scatterplot of data in SPSS
- USE LSD.SAV – a made up example of LSD concentration in blood and math score on a test

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

Homework

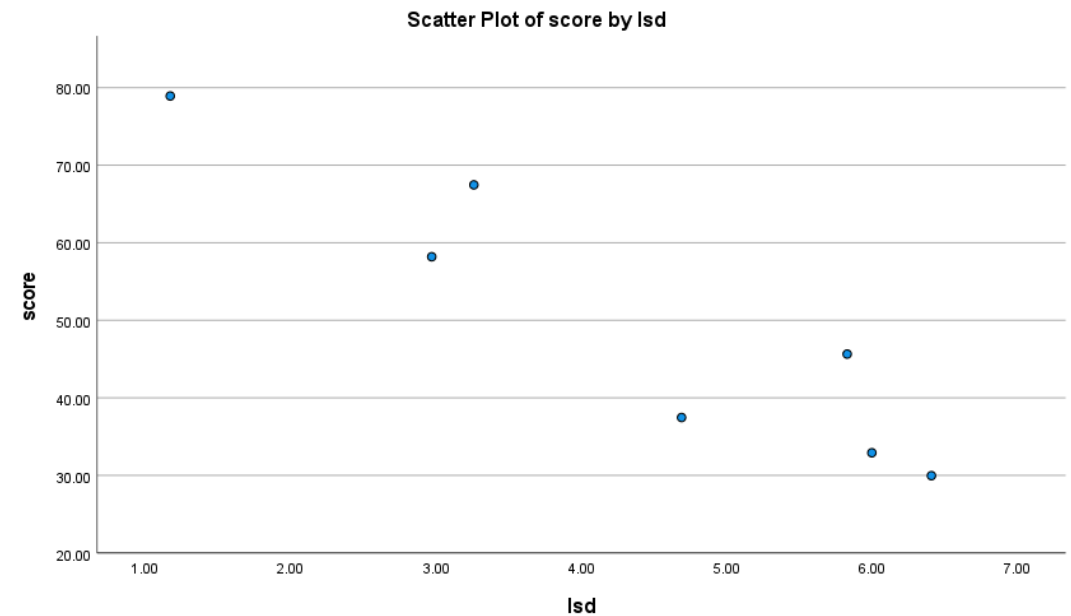
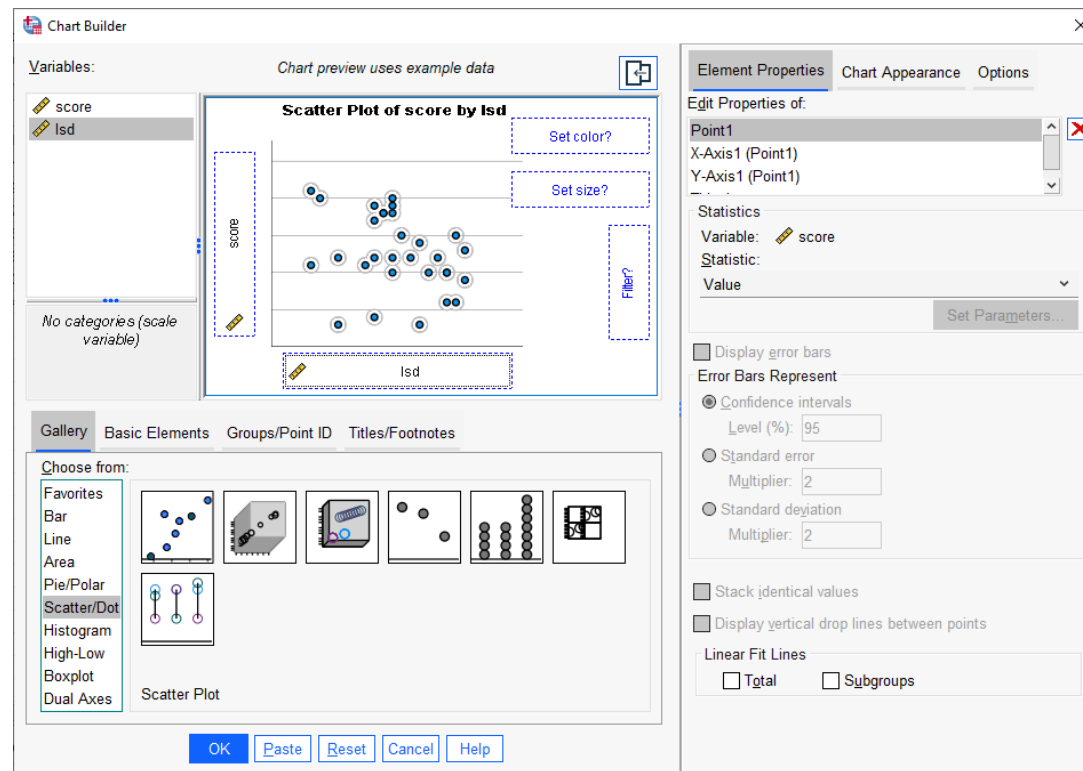
- Do practice problems in folder for this week and submit to me by next Friday on canvas
- Run through the next example yourself and make sure you can replicate the results (LSD.sav, LSD.jasp)

From SPSS



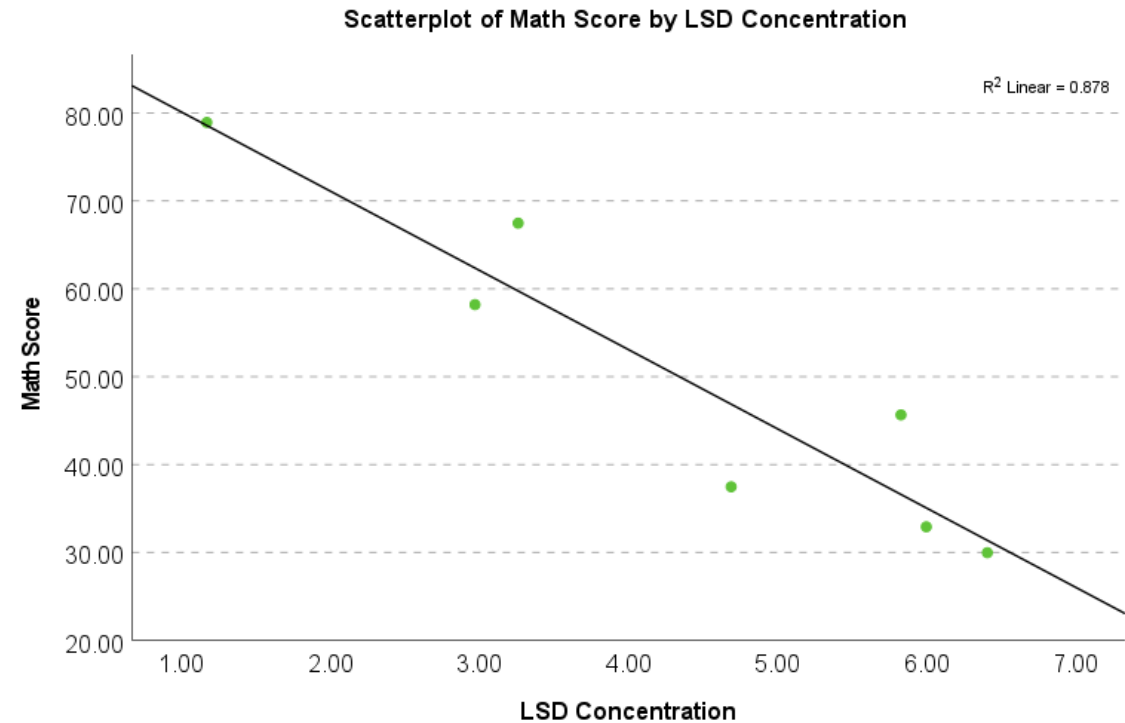
From SPSS

- 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
 - $\bar{Y} = 50.087$
 - $\bar{X} = 4.333$
 - $b_1 = -202.49 / 22.47 = -9.01$
 - $b_0 = 50.09 - (-9.01 \times 4.33) = 89.10$
 - $\hat{Y} = 89.10 - 9.01X$
 - $s^2 = 50.72$
- 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 Coefficient...

☒ Estimates

☐ Confidence intervals

Level(%):

☐ Covariance matrix

☒ Model fit

☐ R squared change

☒ Descriptives

☐ Part and partial correlations

☐ Collinearity diagnostics

Residuals


☐ Durbin-Watson

☐ Casewise diagnostics

☒ Otliers outside:

standard deviations

☐ All cases



Linear Regression

✕

Isd

➡

Dependent:
score

Block 1 of 1

PreviousNext

Block 1 of 1

Isd

Method: Enter

Selection Variable:

Rule...

Case Labels:

WLS Weight:

OK

Paste

Reset

Cancel

Help

Statistics...

Plots...

Save...

Options...

Style...

Bootstrap...

Unlabeled [DataSet1] - IBM SPSS Statistics Data Editor

FileEditViewDataTransformAnalyzeGraphsUtilitiesExtensionsWindowHelp

Pager Analysis

Meta Analysis

Reports

Descriptive Statistics

Bayesian Statistics

Tables

Copiers 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...

Complex Samples

Simulation...

Quality Control

Visible: 2 of 2 Variables

varvarvarvarvarvarvarvarvarvar

Linear

Variable View

SS Statistics Processor is ready

Unicode ON Classic

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

$$H_o: \beta_1 = 0 \text{ vs } H_1: \beta_1 \neq 0$$

$$H_o: \beta_1 = 0 \text{ vs } H_1: \beta_1 \neq 0$$

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

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

Testing

$$t^* = -9.01 / 1.50 = -6.01 \rightarrow p = .002 \rightarrow \text{reject Null}$$

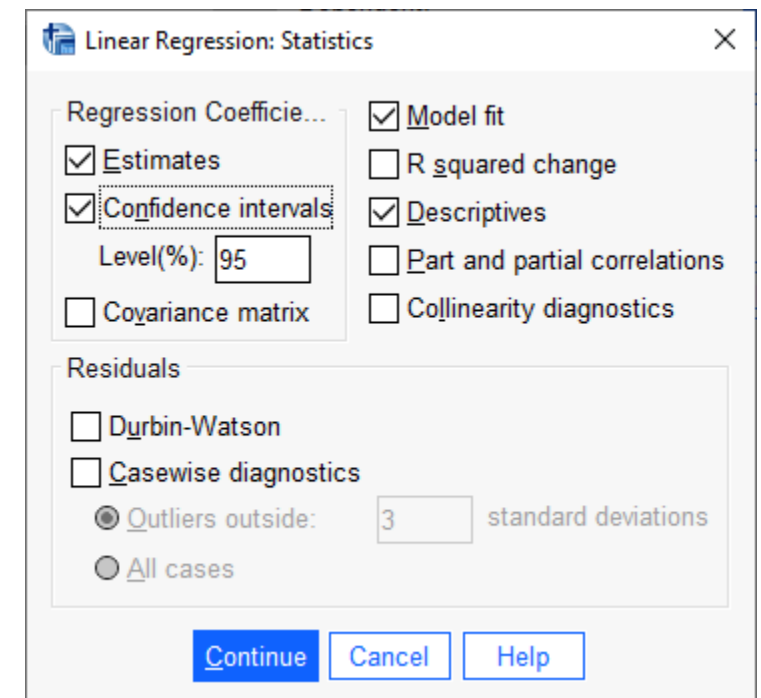
Note, the t equals the Z only when N is large

Example -Pharmacodynamics of LSD

- $n = 7$
- $b_1 = -9.01$
- $se(b_1) = 7.12/22.475 = 1.50$
- 95% CI
 - $-9.01 \pm 2.57(1.50) \rightarrow -9.01 \pm 3.86 \rightarrow (-12.87, -5.15)$
 - From this confidence interval, do you reject the Null?

Coefficients ^a								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		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



The image shows the 'Linear Regression: Statistics' dialog box in SPSS. The 'Regression Coefficients' section has 'Estimates', 'Confidence intervals', and 'Model fit' checked. The 'Level(%)' is set to 95. The 'Residuals' section has 'Durbin-Watson', 'Casewise diagnostics', and 'Outliers outside: 3 standard deviations' selected. The 'Continue', 'Cancel', and 'Help' buttons are at the bottom.

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 model

Example

Bivariate Correlations

Variables:

score
Isd

Options...
Style...
Bootstrap...
Confidence interval...

Correlation Coefficients

☐ Pearson ☐ Kendall's tau-b ☒ Spearman

Test of Significance

☒ Two-tailed ☐ One-tailed

☒ Flag significant correlations ☐ Show only the lower triangle ☒ Show diagonal

OK Paste Reset Cancel Help

Correlations

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

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

Correlations

			score	Isd
Spearman's rho	score	Correlation Coefficient	1.000	-.929**
		Sig. (2-tailed)	.	.003
		N	7	7
	Isd	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 b_1
- $F = t^2$

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

ANOVA F-test for SLR

- Analysis of Variance *F*-test
 - $H_0: \beta_1 = 0$
 - $H_1: \beta_1 \neq 0$
- $F^* = \frac{MSR}{MSE} > F_{\alpha, df=1, n-2} \rightarrow$
reject null

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

Example - Pharmacodynamics of LSD

- Total Sum of squares:

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

- Error Sum of squares:

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

- Model Sum of Squares:

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

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		

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 $H_A: \beta_1 \neq 0$

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

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

Fitted Values & Residuals

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

$$\text{Score} = 89.125 - 9.009(4.69) = 46.87279$$

- The residual is the difference between the observed value Y_i 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

Introduction to Logistic Regression

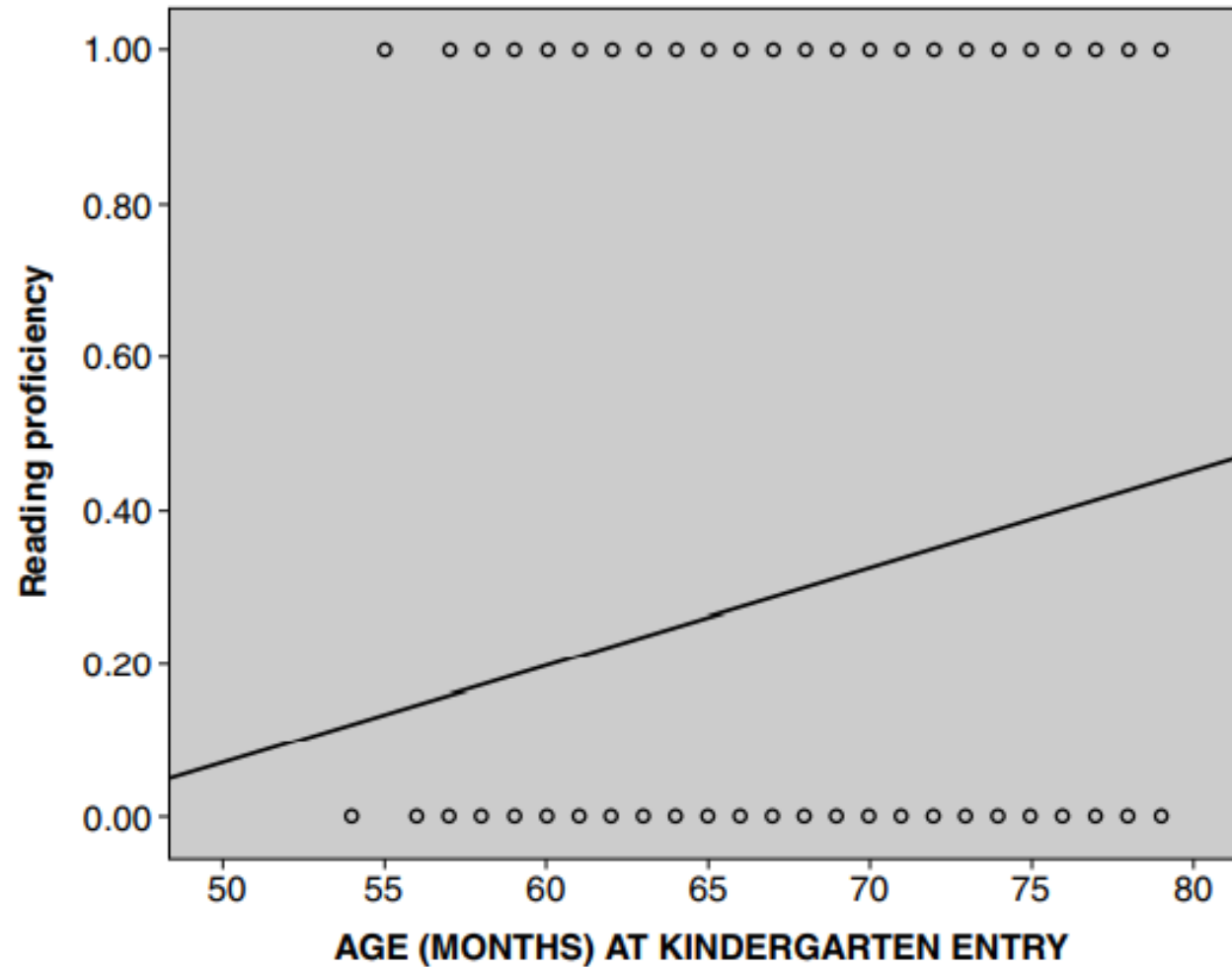
Outcome is **categorical** and allows modeling **prediction**

The equation is similar to OLS *however* in logistic regression

- The **dependent variable**, which is measured as binary (0/1), is **transformed into a logit variable** (i.e., natural log of the odds of the dependent variable occurring or not occurring) and the parameters are estimated using maximum likelihood
- We are estimating **the odds of an event occurring** (not the precise numerical value as in OLS)

Logistic regression equation allows us to compute a *probability*

- *Probability* that the dependent variable will occur
- Logistic regression equation generates predicted probabilities between values of 0 and 1



OLS is inappropriate when the dependent variable is binary

Introduction to Logistic Regression

A regression where the dependent variable is measured 0/1 (pass/fail; vote/didn't vote, etc.)

In ordinary regression the model predicts the *mean* Y for any combination of predictors.

Goal of logistic regression: Predict the “true” proportion of success, p , at any value of the predictor

p = Proportion of “Success”

What's the “mean” of a 0/1 indicator variable?

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\# \text{ of } 1's}{\# \text{ of trials}} = \text{Proportion of "success"}$$

Equivalent forms of the logistic regression model

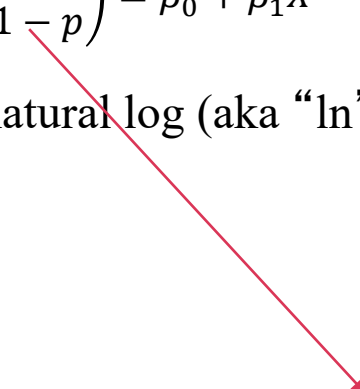
Y = Binary response

X = Quantitative predictor

Logit form

$$y = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

This is natural log (aka “ln”)



Log Odds

Probability form

$$p = \Pr(Y = 1) = \frac{\exp^{\beta_0 + \beta_1 X_1}}{1 + \exp^{\beta_0 + \beta_1 X_1}}$$

p = proportion of 1's (yes, success) at any X



Probability

What Logistic Regression Is and How It Works: Characteristics

Odds and logit (or log odds)

Odds = the ratio of the probability of the dependent variable's two outcomes

- Taking the log odds of Y creates a linear relationship between X and the probability of Y (Pampel, 2000)

$$\ln \frac{P(Y = 1)}{1 - P(Y = 1)} = \text{Logit}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$$

$$\ln \frac{P(Y = 1)}{1 - P(Y = 1)} = \text{Logit}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$$

What Logistic Regression Is and How It Works: Characteristics

- Interpretation: For each one-unit change in the independent variable, the **log odds** of Y increase or decrease by the beta coefficient
 - **Problem:** nobody knows what log odds are!
- Procedure
 - Interpret the independent variables as affecting the odds (rather than log odds) of the outcome
 - Exponentiate the coefficients (i.e., the outcome of the logistic regression equation), to convert it back to the odds
 - Convert odds to probability using our formula $p = \text{odds} / (1 + \text{odds})$
 - Probability values close to one indicate increased likelihood of occurrence
 - Much more intuitive

What Logistic Regression Is and How It Works: Characteristics

Estimation and model fit

Maximum likelihood estimation (MLE) is applied to the model and estimates the odds of occurrence after transformation into the logit

- MLE is a method that determines values for the parameters of a model by maximizing their likelihood
- Contrast OLS

The log of the likelihood function that results from ML estimation reflects the likelihood of observing the sample statistics given the population parameters

- Log likelihood provides an index of how much has not been explained in the model after the parameters have been estimated and can be used as an indicator of model fit
- *LL* values range from zero to negative infinity (select smaller, i.e., more negative, values)

What Logistic Regression Is and How It Works: Characteristics

Significance tests: Model fit

- Overall logistic regression model: determines the extent to which the predicted values accurately represent the observed values (Xie, Pendergast, & Clarke, 2008)
 - **Change in log likelihood**
 - Hosmer and Lemeshow goodness-of-fit test
 - Pseudo-variance explained
 - **Predicted group membership**

Significance tests: Beta coefficients

- Each logistic regression coefficient determines if the individual coefficients are statistically significantly different from zero

Change in Log Likelihood

Test of overall model fit

- Useful when **comparing two nested models**
- The test is based on the change in the log likelihood function from the smaller model (with fewer variables) to a larger model (with the same variables in the smaller model and one or more additional variables)
- Note: this includes the intercept-only model with no predictors

The test is all regression coefficients are 0 $\rightarrow H_0: \beta_1 = \beta_2 \dots \beta_m = 0$ $H_1: \text{at least one } \beta \neq 0$

The test statistic computed as $-2 * LL_{\text{diff}}$ is distributed as χ^2 with $df = df[\text{larger model}] - df[\text{smaller model}]$

The larger the difference, the better the model fit for the alternative model

Predicted Group Membership

If the predicted probability is above a cutoff (usually .5) assign 1, otherwise 0

A crosstab table of predicted to observed probabilities provides the frequency and percentage of cases correctly classified

A perfect model produces 100% correctly classified cases

A model that classifies no better than chance would provide 50% correctly classified cases.

Terminology

Sensitivity is the probability that a case coded as **1 for the dependent variable** (aka 'positive') is **classified correctly**

- the percentage of correct predictions of the cases that are coded as 1 for the dependent variable

Specificity is the probability that a case coded as **0 for the dependent variable** (aka 'negative') is **classified correctly**

- the percentage of correct predictions of the cases that are coded as 0 for the dependent variable

False positive rate is the probability that a case coded as 0 for the dependent variable (aka 'negative') is classified incorrectly.

- In other words, this is the percentage of cases in error where the dependent variable is predicted to be 1, but in fact the observed value is 0

False negative rate is the probability that a case coded as 1 for the dependent variable (aka 'positive') is classified incorrectly.

- In other words, this is the percentage of cases in error where the dependent variable is predicted to be 0, but in fact the observed value is 1

Test of Significance of the Logistic Regression Coefficients

The second test in logistic regression is the test of the statistical significance of each regression coefficient, b_k

$$H_0: \beta_k = 0 \qquad H_1: \beta_k \neq 0$$

The test statistic is called a Wald test computed as

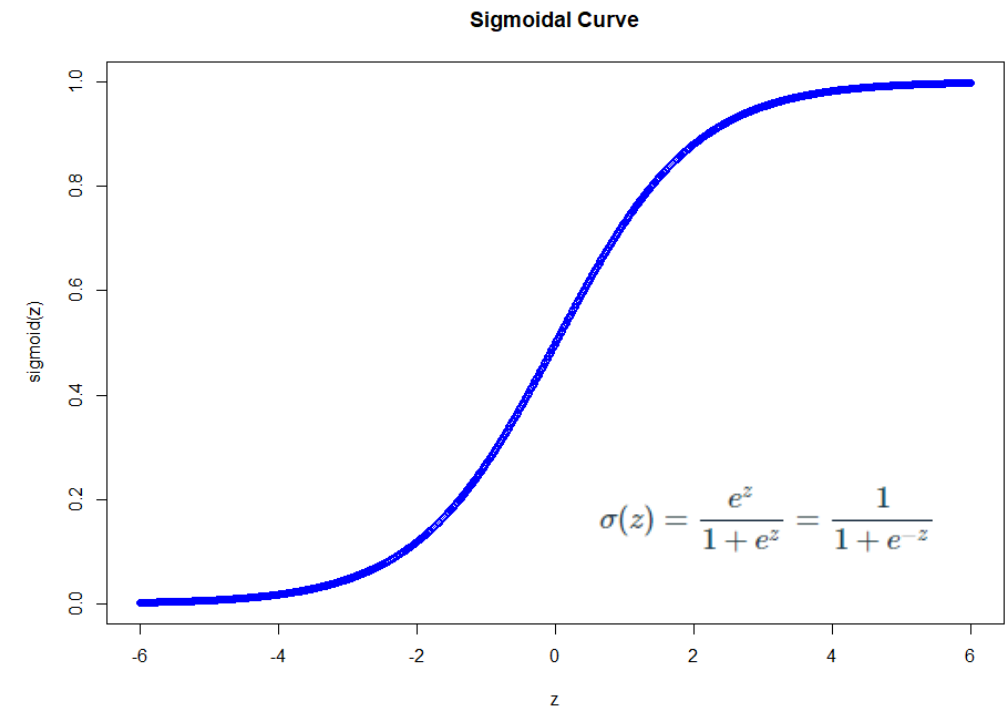
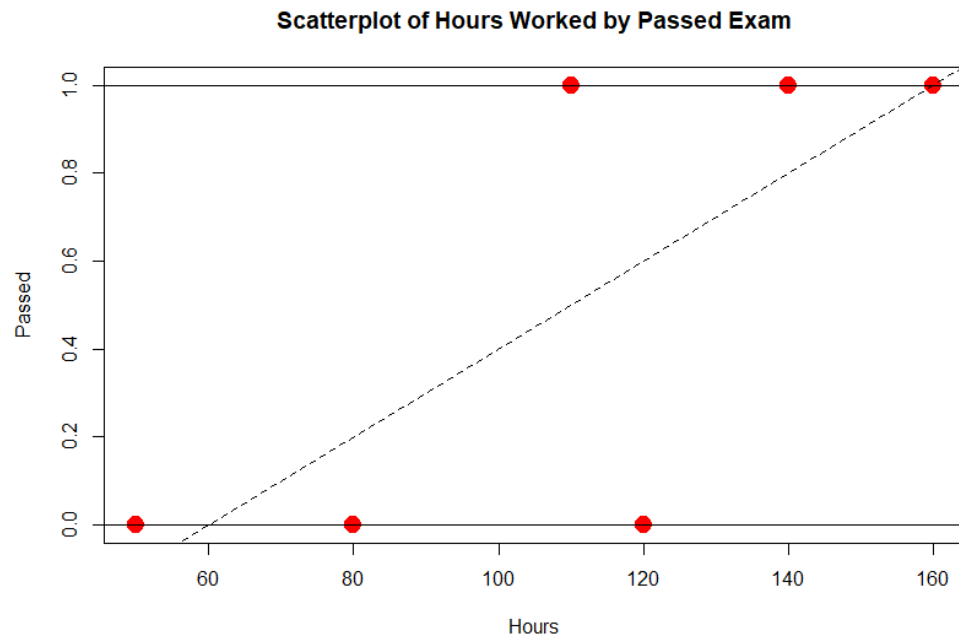
$$W = \frac{\beta_k^2}{SE_{\beta_k^2}} \sim \chi_2$$

What Logistic Regression Is and How It Works: Sample Size

Logistic regression is a large sample procedure

Samples of size 100 or greater are needed to accurately conduct tests of significance for logistic regression coefficients ([Long, 1997](#))

Linear vs. Logistic Regression



Recall the model

$$Y = \begin{cases} 0 & \text{If the condition is false} \\ 1 & \text{If the condition is true} \end{cases}$$

$$p(X) = P[Y = 1 | X = x] = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

Our goal is to estimate the unknown parameters, $\hat{\beta}_0, \hat{\beta}_1$

Review Algebra

Properties of exponents

$$e^x * e^y = e^x + e^y$$

$$e^{x+y} = e^x + e^y$$

Properties of logarithms

$$\ln(e^x) = x$$

$$e^{\ln(x)} = x$$

$$e^0 = 1$$

Mathematical Snapshot

$$\text{Logit}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$$

$$\text{Logit}(Y) = \beta_0 + \beta_1 X_1$$

β_0, β_1 is interpreted as “log odds of Y”

$$\exp^{\text{Logit}(Y)} = \exp^{\beta_0} + \exp^{\beta_1 X_1}$$

Take the exponent of each side

$$\exp^{\text{Logit}(Y)} = \exp^{\beta_0 + \beta_1 X_1}$$

This follows from properties

$\exp(\beta_0), \exp(\beta_1)$ is interpreted as the “odds of Y”

$$\Pr(Y = 1) = \frac{\exp^{\beta_0 + \beta_1 X_1}}{1 + \exp^{\beta_0 + \beta_1 X_1}}$$

This follows because $\frac{\text{Odds}}{1 + \text{Odds}}$ is defined as the probability

$$\Pr(Y = 1 | X = x) = \frac{\exp^{\beta_0 + \beta_1 X_1}}{1 + \exp^{\beta_0 + \beta_1 X_1}}$$

This is interpreted as a ‘conditional’ probability

Snapshot

Exponentiate the logit and convert back to odds

$$Odds(Y = 1) = e^{logit(Y)} = e^{\ln[Odds(Y=1)]} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m} = (e^{\beta_0})(e^{\beta_1 X_1})(e^{\beta_2 X_2}) \dots (e^{\beta_m X_m})$$

Exponentiation creates a **multiplicative rather than additive equation**, and this then changes the interpretation of the exponentiated coefficients. In OLS, when the product of the regression coefficient and its predictor is 0, that variable adds nothing to the prediction of the dependent variable. Not true here.

Here, when the coefficient value is 0, the odds are one (no difference). But, coefficients greater than 1 increase the odds, and coefficients less than 1 decrease the odds. In addition, the odds will change more the greater the distance the value is from 1

Converting the results to probabilities

Convert odds back to probability

$$Pr(Y = 1|X = x) = \frac{Odds(Y = 1|X = x)}{1 + Odds(Y = 1|X = x)} = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m}}$$

Binary Logistic Regression in SPSS

Use the file `gpa-college-enroll.sav`

Analyze → Regression → Binary Logistic

In this example we have

Y = college enrollment

X_1 = undergraduate GPA

Focus on the results first

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	gpa	.883	.821	1.157	1	.282	2.419
	Constant	-2.144	2.416	.788	1	.375	.117

a. Variable(s) entered on step 1: gpa.

$$\text{logit}(Y) = -2.144 + .833GPA$$

$$p(X) = \text{Pr}[Y = 1|X = x] = \frac{\exp(-2.144 + .883X)}{1 + \exp(-2.144 + .883X)}$$

Probability of enrolling in college 'conditional on' values of one's grade point average

Focus on the results first

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	gpa	.883	.821	1.157	1	.282	2.419
	Constant	-2.144	2.416	.788	1	.375	.117

a. Variable(s) entered on step 1: gpa.

The Wald test is the test statistic for the logistic regression model

The sig. is the *p*-value

EXP(B) is the odds ratio – the interpretation is: the odds of enrolling in college are about 2.5 times greater for each one unit increase in undergraduate gpa, but the result is not statistically significant

Computing probabilities

The most useful part of logistic regression is the ability to predict the conditional probability of Y given values of the independent variables

For example: what is the probability of college enrollment given that your gpa is 1.0, 2.0, 3.0 and 4.0?

$$\Pr(Y = 1) = p = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

$$\Pr(Y = 1|X = 1) = \frac{e^{-2.144 + .883(1)}}{1 + e^{-2.144 + .883(1)}} = \frac{e^{-2.144 + .883}}{1 + e^{-2.144 + .883}} = \frac{e^{-1.261}}{1 + e^{-1.261}} = \frac{.283371}{1 + .283371} = .221$$

$$\Pr(Y = 1|X = 2) = \frac{e^{-2.144 + .883(2)}}{1 + e^{-2.144 + .883(2)}} = \frac{e^{-2.144 + 1.766}}{1 + e^{-2.144 + 1.766}} = \frac{e^{-.378}}{1 + e^{-.378}} = \frac{.283371}{1 + .283371} = .407$$

Table of predicted probabilities

GPA (X)	β_0	β_1	$e^{\beta_0+\beta_1X_1}$	$1+e^{\beta_0+\beta_1X_1}$	$\frac{e^{\beta_0+\beta_1X_1}}{1+e^{\beta_0+\beta_1X_1}}$
1	-2.144	0.883	0.283370514	1.2833705	0.220801796
2	-2.144	0.883	0.685230501	1.6852305	0.406609363
3	-2.144	0.883	1.65698552	2.6569855	0.623633628
4	-2.144	0.883	4.006828377	5.0068284	0.800272763

See excel file `gpa_college_enrollment_example.xlsx`

Model Summary & Classification

The classification table is provided. Here we correctly classify 60% of enrollments, better than chance (not much better)

- Using the 50% cutoff, when the observed Y was the student did not enroll in college, we predicted enrollment correctly 1/2 of the time
- Similarly, using the same cutoff, when the student did enroll, we were correct 4 of 6 times (2/3)

Classification Table^a

Observed			Predicted		Percentage Correct
			.00	1.00	
Step 1	enroll	.00	2	2	50.0
		1.00	2	4	66.7
	Overall Percentage				60.0

a. The cut value is .500

Specificity

Sensitivity

$(2 + 4) / (2 + 2 + 2 + 4) = 60\%$

More on classifications and correctly predicted observations

Your observed outcome in logistic regression can ONLY be 0 or 1

The predicted probabilities from the model can take on all possible values between 0 and 1

So, for a given observation, the predicted probability from the model may have been 0.51 (51% probability of success), but your observation was actually a 0 (not a success)

By default, SPSS uses a classification of 50/50 → if the probability is .51 or greater, the case is coded as “1” (i.e., a success) and if it is $\leq .50$ it's coded as “0” (i.e., a fail)

The next slide shows how we can add the probabilities and classifications to our dataset for additional intuition

gpa
gender
Predicted probability [PRE_1]
Predicted group [PGR_1]

Dependent:



enroll

Block 1 of 1

Previous

Next

Categorical...

Save...

Options...

Style...

Bootstrap...

Logistic Regression: Save

Predicted Values

☒ Probabilities

☒ Group membership

Influence

☐ Cook's

☐ Leverage values

☐ DfBeta(s)

Residuals

☐ Unstandardized

☐ Logit

☐ Studentized

☐ Standardized

☐ Deviance

Export model information to XML file

[Browse](#)

☒ Include the covariance matrix

[Continue](#) [Cancel](#) [Help](#)

Method: Enter

Selection Variable:



Rule...






OK

Paste

Reset

Cancel

Help

 gpa	 gender	 enroll	 PRE_1	 PGR_1
1.89	.00	.00	.38350	.00
2.10	.00	1.00	.42819	.00
2.36	.00	1.00	.48511	.00
1.70	.00	.00	.34467	.00
4.15	.00	1.00	.82078	1.00
2.72	1.00	1.00	.56425	1.00
3.16	1.00	.00	.65636	1.00
3.89	1.00	1.00	.78448	1.00
4.02	1.00	1.00	.80326	1.00
3.55	1.00	.00	.72941	1.00

Improving classification

We can change the default value from .5 to something else

Example: Let's change the default to .4 and see how this 'improves' our classification








Classification Table^a

		Predicted		Percentage Correct
		enroll .00	1.00	
Step 1	enroll	.00	2	50.0
		1.00	0	100.0
	Overall Percentage			80.0

a. The cut value is .400

Default = .5

Default = .4

 gpa	 gender	 enroll	 PRE_1	 PGR_1	 PRE_2	 PGR_2
1.89	.00	.00	.38350	.00	.38350	.00
1.70	.00	.00	.34467	.00	.34467	.00
4.15	.00	1.00	.82078	1.00	.82078	1.00
2.72	1.00	1.00	.56425	1.00	.56425	1.00
3.16	1.00	.00	.65636	1.00	.65636	1.00
3.89	1.00	1.00	.78448	1.00	.78448	1.00
4.02	1.00	1.00	.80326	1.00	.80326	1.00
2.10	.00	1.00	.42819	.00	.42819	1.00
2.36	.00	1.00	.48511	.00	.48511	1.00
3.55	1.00	.00	.72941	1.00	.72941	1.00

When we changed the default value from .5 to .4, what changed?

You may want to report...

Is the 60% correctly predicted rate better than chance (i.e., 50%)

You can report Press's Q, which is distributed as

$$Q = \frac{[N - (nK)]^2}{N(K-1)},$$

N is the total sample size,
 n represents the number of cases that were correctly classified
 K equals the number of groups

$$Q = \frac{[10 - (6 \cdot 2)]^2}{10(2-1)} = .4 < \chi_2^2(1) = 3.841$$

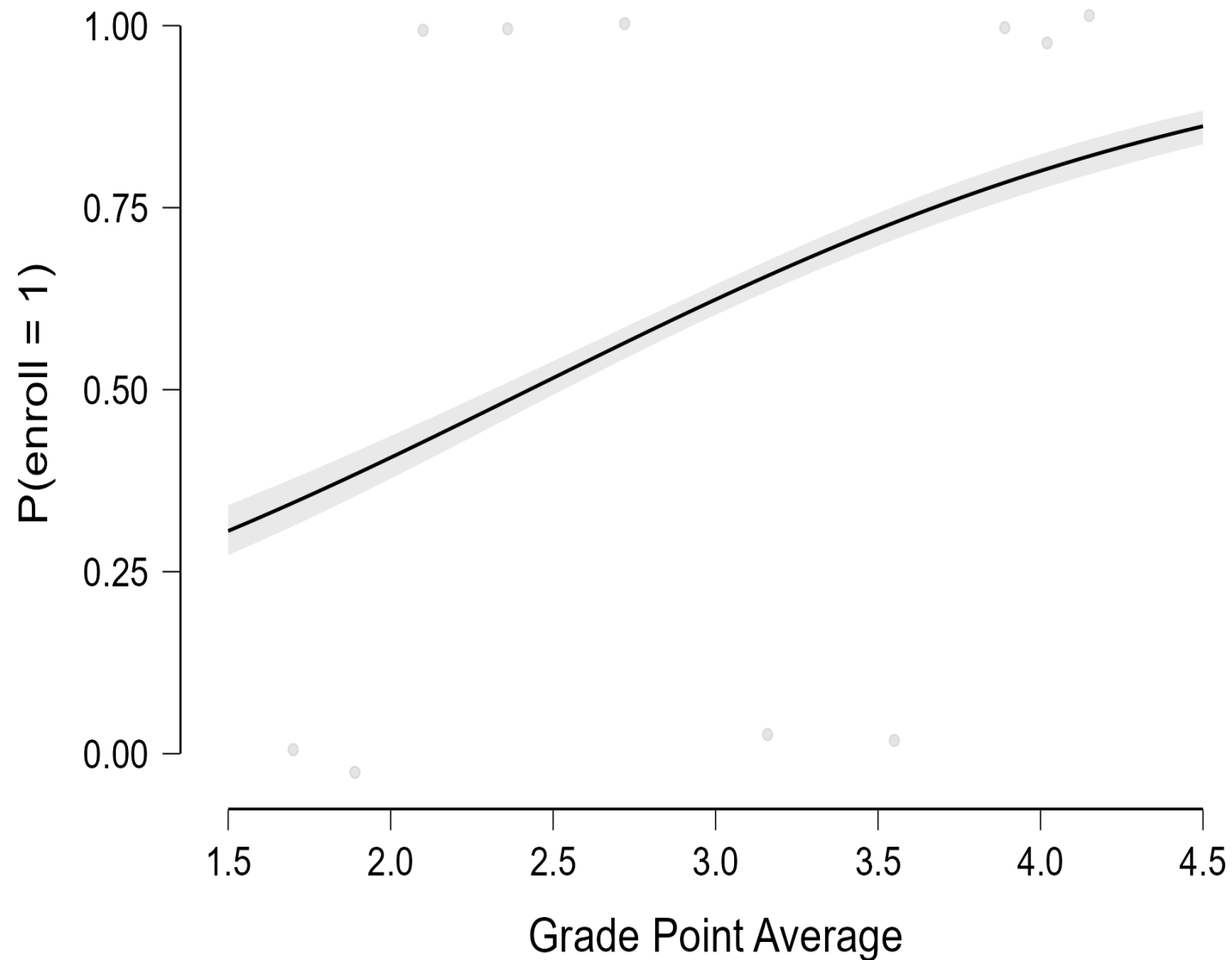
In this case, the results are no better than chance

Let's do this in JASP

Probability of enrolling in college conditional on high school gpa

As gpa increases the probability of enrolling increases

The probability is about .30 for a gpa = 1.5 but increases to about .9 for a gpa = 4.5



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	12.169 ^a	.121	.164

a. Estimation terminated at iteration number 4
because parameter estimates changed by less
than .001.

Log-Likelihood

The model summary information is provided only as
a method of comparison

Let's compare this model with a model that includes
student gender (next slide)

Does the addition of gender improve the model fit?

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	12.169 ^a	.121	.164

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	10.936 ^a	.223	.302

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

$$\chi^2(1) = 2(LL_{m2} - LL_{m1}) = -(10.936 - 12.169) = 1.233$$

m2 = model with additional predictors, it is also called the unrestricted model

m1 = model with fewer predictors, it is nested in m2, it is also called the restricted model

There is one parameter difference between the models, and one degree of freedom

The critical value is greater than the calculated statistic and hence there is no statistically significant difference between these models!

Your turn

- Open the SPSS file passing.sav
- The data has three made-up variables:
 - Passed a statistics test
 - Hours of study
 - Gender (1 = male, 0 = female)
- Run a binary logistic regression of passed on hours
 - Write out the regression equation
 - Predict the probability of passing for 0, 50, 100, 150, 200 and 250 hours of studying
 - Add gender into the model and interpret the coefficient in terms of odds and percent change in odds
 - Does adding gender improve the model fit?
 - Compute the probability of passing for males and females for 100 study hours

Model & Interpretation

$$\text{logit}(Y) = -9.3 + .082X$$

$$p(X) = P[Y = 1|X = x] = \frac{\exp(-9.3 + .082X)}{1 + \exp(-9.3 + .082X)}$$

Check: What is the Probability of passing when hours = X

Note: $\exp(.082) = 1.085$

Table 1. Predicted probability of Passing Conditional on Hours Studying

X	0	50	100	150	200	250
$P(Y x=X)$	0.0055	0.0603	0.4286	0.6298	0.8975	0.9783

Each additional hour spent working increases the odds of passing by $[\exp(.082)-1 * 100]$ percent

Each additional hour spent working increases the odds of passing by about 8.5 percent

The odds of passing are 1.085 times higher for each additional hour spent studying

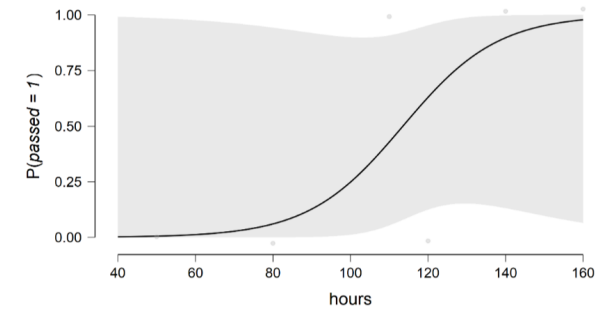
Model Summary - passed ▼

Model	Deviance	AIC	BIC	df	X ²	p	McFadden R ²	Nagelkerke R ²	Tjur R ²	Cox & Snell R ²
H ₀	8.318	10.318	10.110	5						
H ₁	4.078	8.078	7.661	4	4.240	0.039	0.510	0.676	0.536	0.507

Coefficients ▼

	Estimate	Standard Error	Odds Ratio	z	Wald Test		
					Wald Statistic	df	p
(Intercept)	-9.299	8.242	9.149e-5	-1.128	1.273	1	0.259
hours	0.082	0.070	1.085	1.169	1.367	1	0.242

Note. passed level '1' coded as class 1.



Jasp Output

Let's add a binary variable, gender, to the equation

Coefficients

	Estimate	Standard Error	Odds Ratio	z	Wald Test		
					Wald Statistic	df	p
(Intercept)	-8.924	8.487	1.332e -4	-1.051	1.106	1	0.293
hours	0.079	0.071	1.083	1.126	1.267	1	0.260
gender (1)	-0.654	3.764	0.520	-0.174	0.030	1	0.862

Note. passed level '1' coded as class 1.

The odds of passing for males are $[\exp(-.654)-1 * 100]$ percent lower compared to females

The odds of passing for males are 48.01% lower compared to females

The odds of passing are .5100 ($\exp(-.654)$) times lower for males compared to females

The file probability of passing by gender.xlsx has the probability of passing for males and females given 100 hours of study time