



We're done!

Week 13: Advanced stats



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
-

Hypothesis: 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

- interpret the interaction
- is the hypothesis supported?
- to the extent it is NOT supported *Why?*
- what else is interesting here?

Useful: Block testing coefficients (The null model)

Model Summary - tra1

Model	R	R ²	Adjusted R ²	RMSE
H ₀	0.000	0.000	0.000	11.230
H ₁	0.288	0.083	0.079	10.774

The null model ANOVA table

ANOVA						
Model		Sum of Squares	df	Mean Square	F	p
H ₁	Regression	18196.139	7	2599.448	22.394	< .001
	Residual	200468.412	1727	116.079		
	Total	218664.551	1734			

Note. The intercept model is omitted, as no meaningful information can be shown. *Why?*

Coefficients of the null model

Coefficients						
Model		Unstandardize d	Standard Error	Standardized ^a	t	p
H ₀	(Intercept)	50.457	0.270		187.157	< .001
H ₁	(Intercept)	50.786	1.551		32.739	< .001
	ageY	-0.353	0.121	-0.068	-2.911	0.004
	GENDER (1)	1.089	0.533		2.043	0.041
	OOH (2)	0.141	0.599		0.235	0.814
	nhw (1)	-0.156	0.521		-0.299	0.765
	physab (1)	1.073	0.618		1.736	0.083
	sexab (1)	1.874	0.704		2.664	0.008
	EV2	2.282	0.195	0.274	11.727	< .001

^a Standardized coefficients can only be computed for continuous predictors.

Adding covariates to the null model

HOW MUCH MORE OF THE VARIANCE DOES VIOLENCE EXPOSURE EXPLAIN
OF PTSS COMPARED TO NON-VIOLENCE EXPOSURE?

VERY USEFUL (ASK YUJEONG TO EXPLAIN HER PROJECT)

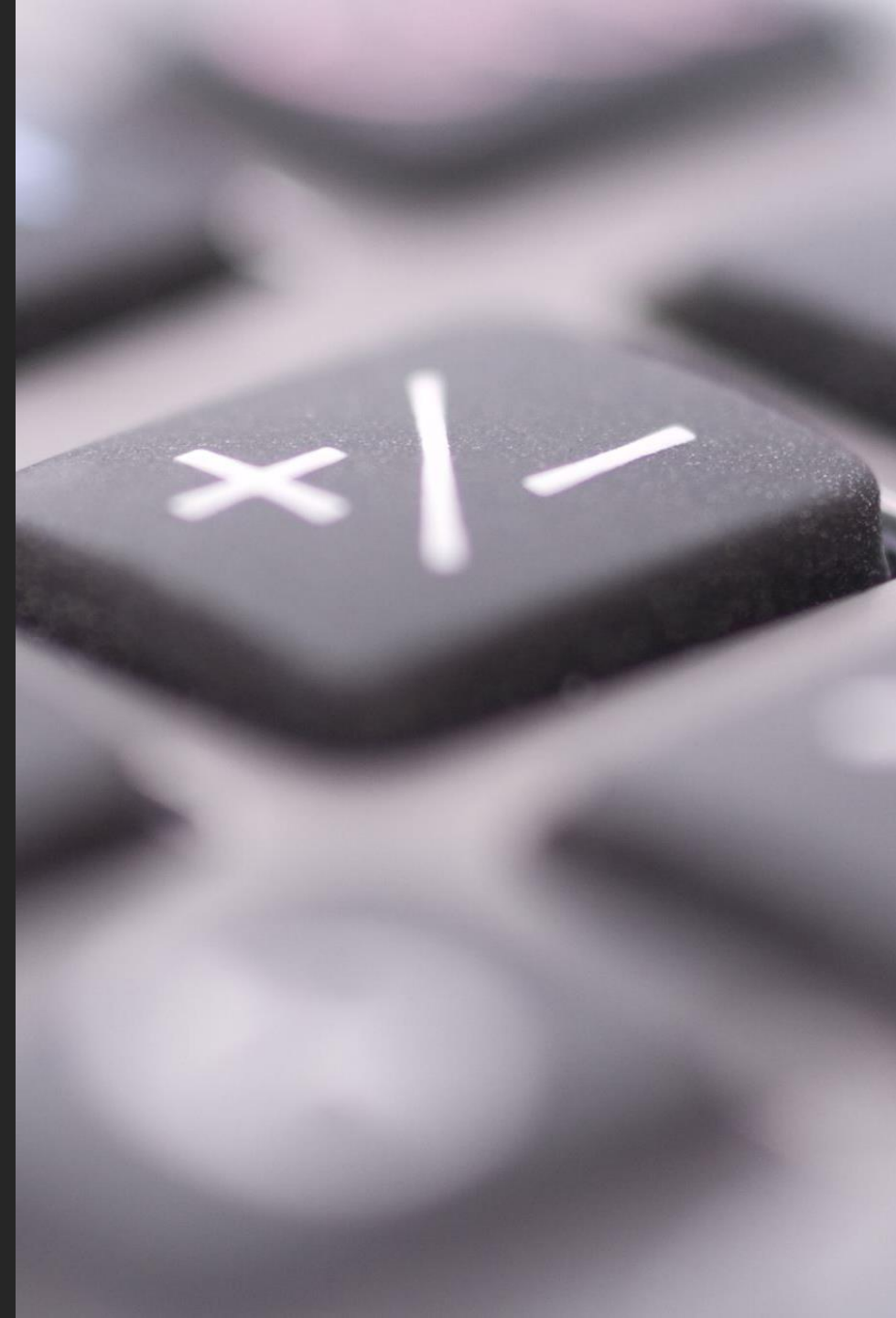
In Jasp – under model tab, “Add to null model”

Click on the variables you want to test

Here I will test EV2, physab and sexab – click on “Add to null model” and add all variables EXCEPT the violence exposure variables above

That essentially means that I am asking how much more of the variance in PTSS is explained by violence exposure above and beyond demographic characteristics (*why?*)

Open the file nscaw_sample.JASP to examine the results



More diagnostics

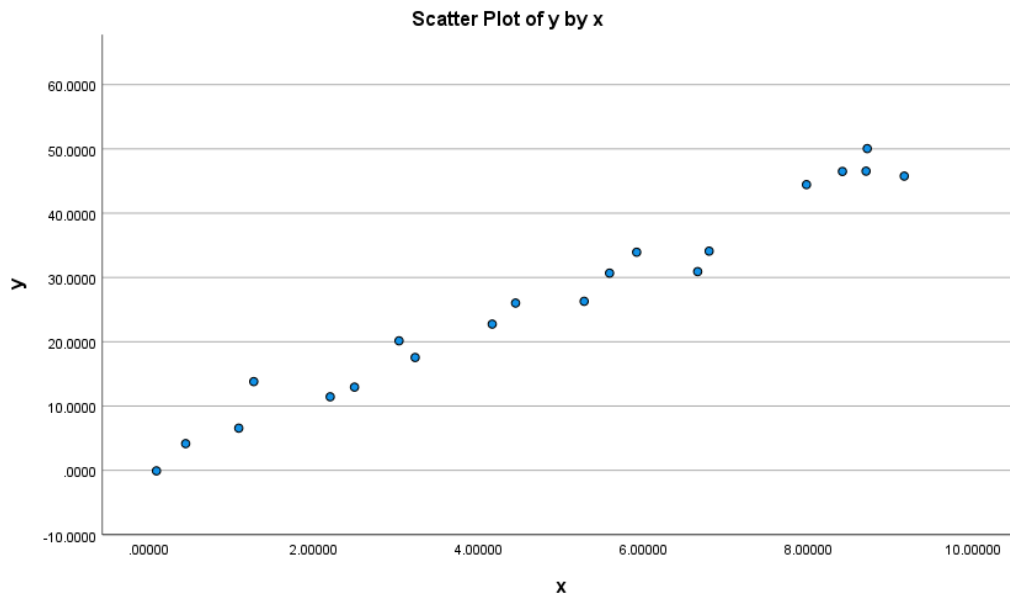
- An outlier is a data point whose response y **does not follow the general trend** of the rest of the data
 - It is defined as a point that is $1.5(Q3-Q1) = 1.5IQR$
- A data point has high leverage if it has "**extreme**" X values
 - With a single predictor, an extreme X value is simply one that is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



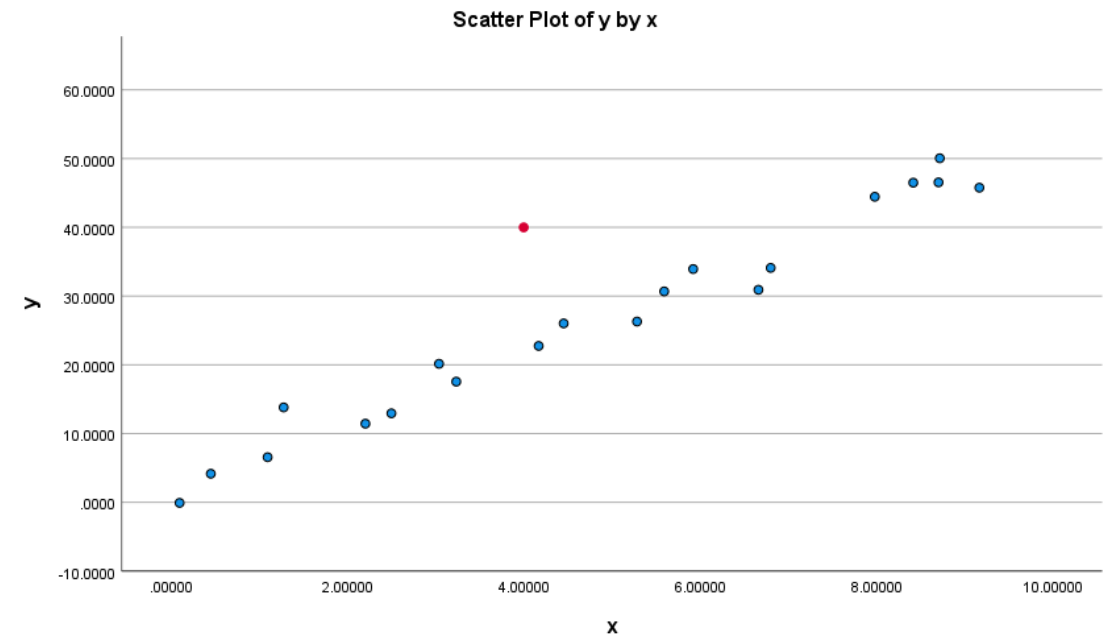
Outliers and unusual values

- Does anything look unusual or different about plot A?
- Does anything look unusual or different about plot B?
- Is the red dot in B a high leverage point, i.e. extreme?

A



B



Regression output

- Compare the two regressions, any differences?
- The odd variable has no effect on the regression

A does not include odd point

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

a. Dependent Variable: y

B does include odd point

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

a. Dependent Variable: y

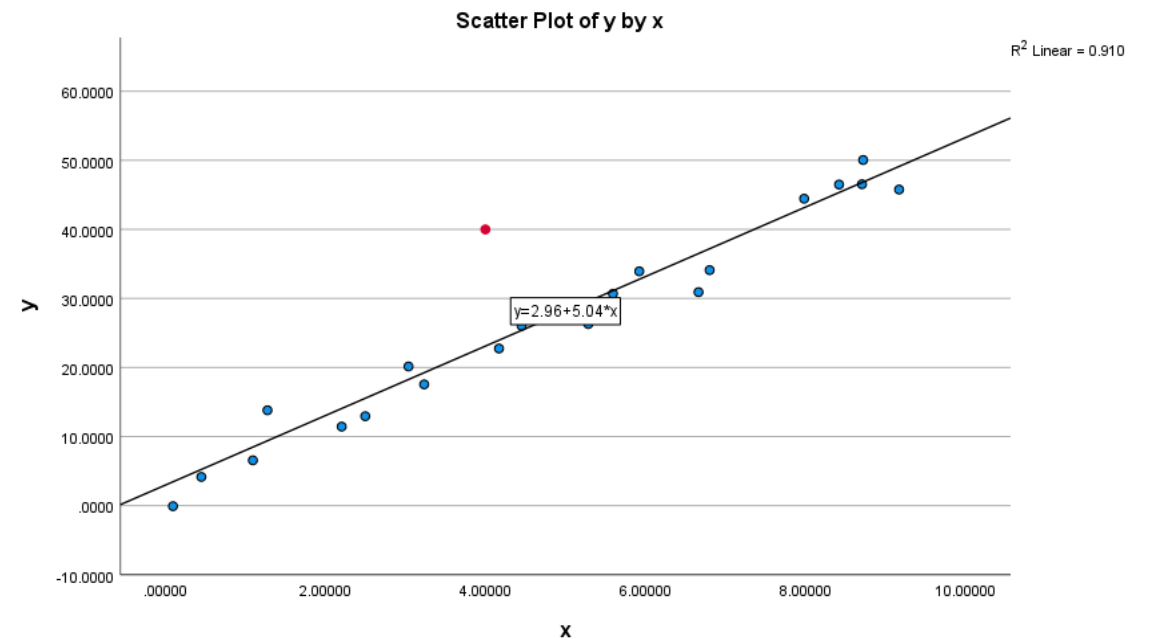
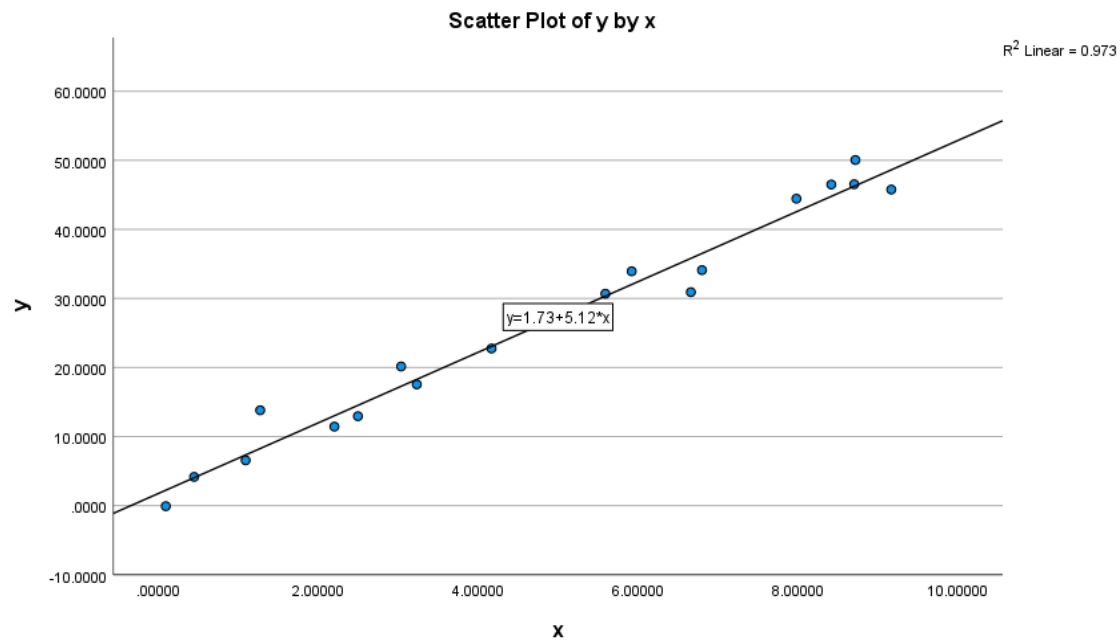
Why is the standard error bigger?

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)	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

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)	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



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

Compare & Intuit

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			

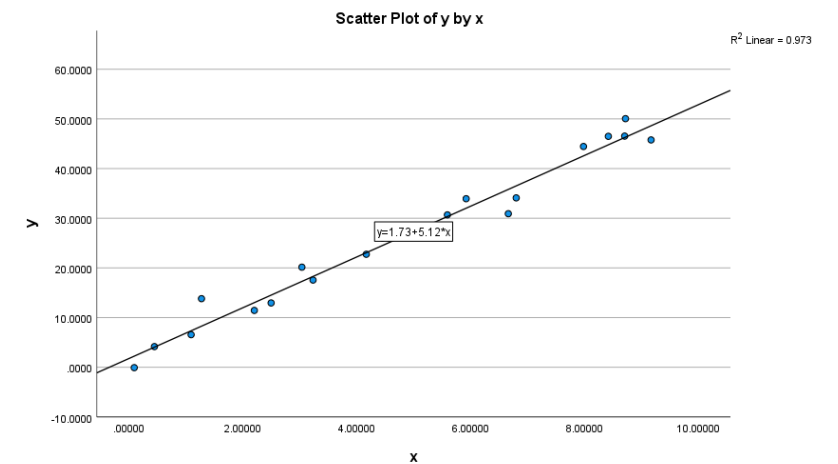
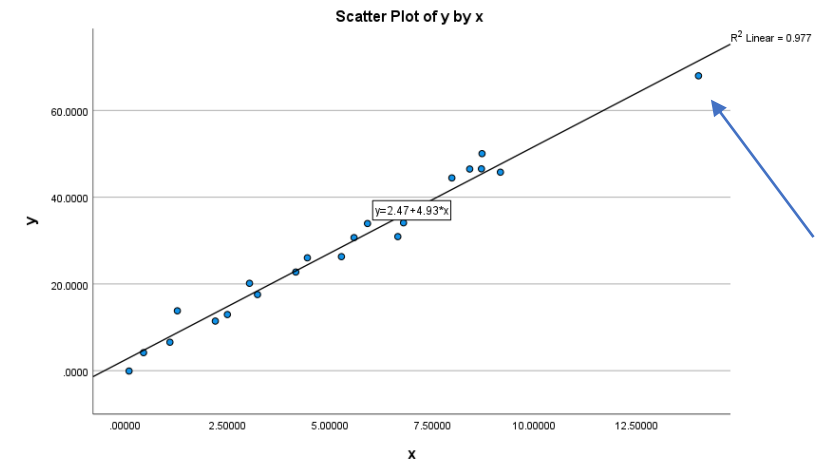
a. Dependent Variable: y

b. Predictors: (Constant), x

Coefficients^a

Model		Unstandardized Coefficients B	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

Model Summary

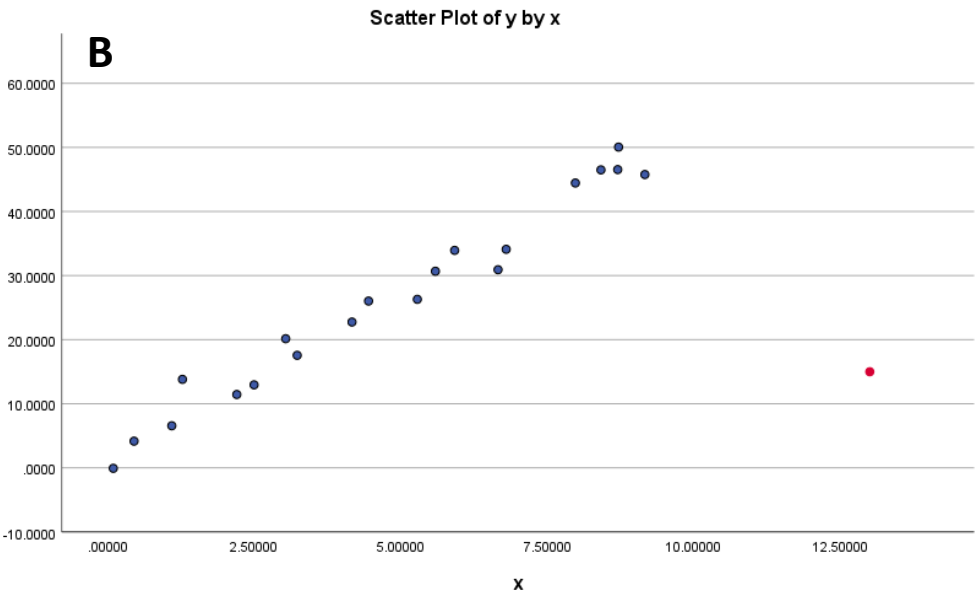
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.743 ^a	.552	.528	10.4459325

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

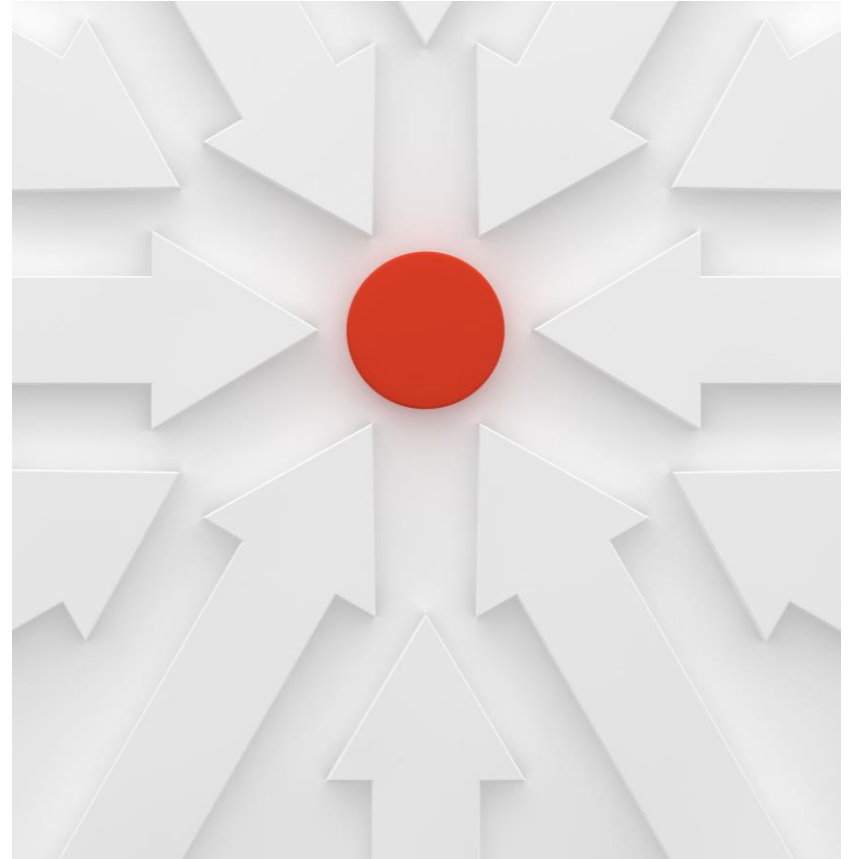
a. Dependent Variable: y



- Does anything look unusual or different about plot B?
- May want to consider nonlinear function of x (like what?)

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:

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

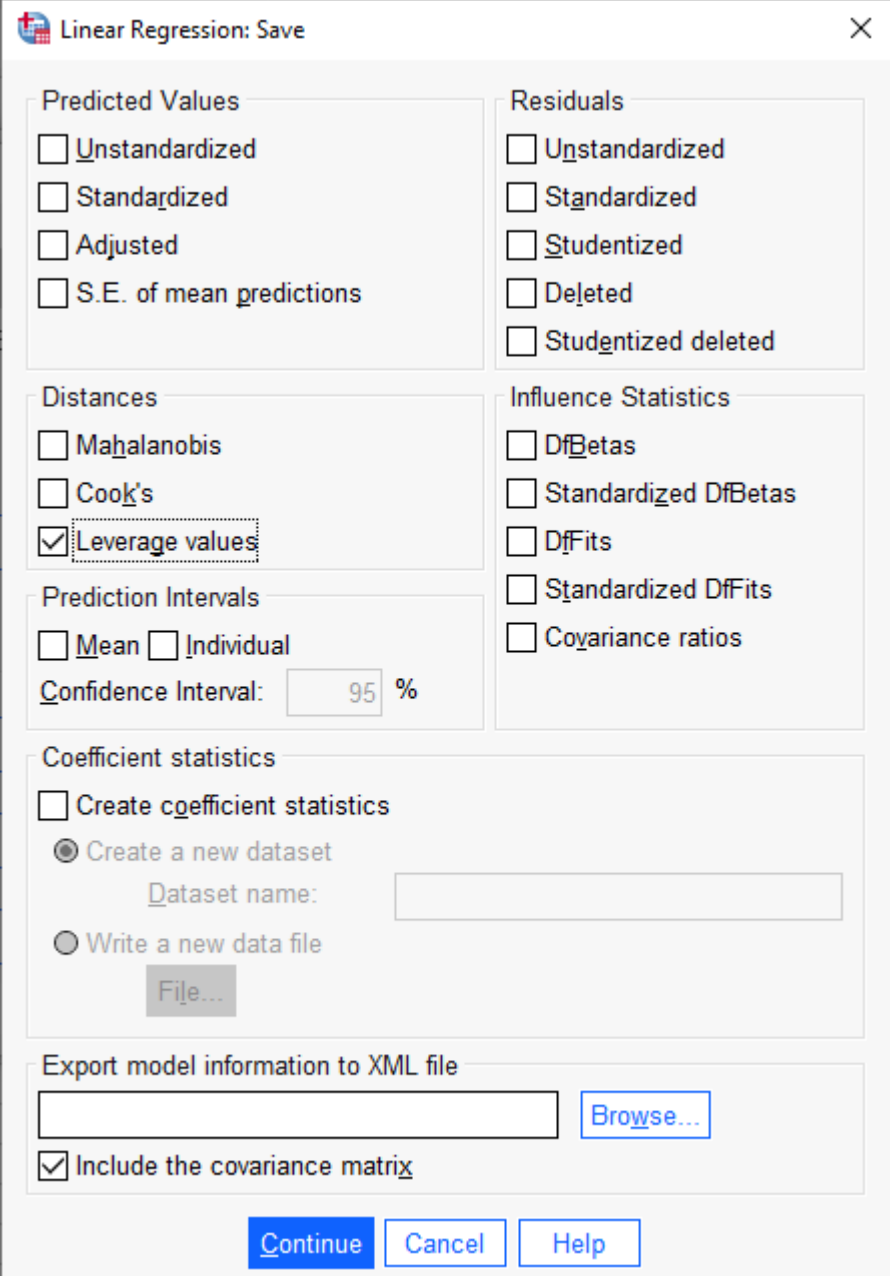
Then flag the observations as unusual \rightarrow X is an observation whose value gives it large *leverage* for the regression analysis

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 the 'Linear Regression: Save' dialog box in SPSS. The 'Predicted Values' section has four options: 'Unstandardized', 'Standardized', 'Adjusted', and 'S.E. of mean predictions', all of which are unchecked. The 'Residuals' section has five options: 'Unstandardized', 'Standardized', 'Studentized', 'Deleted', and 'Studentized deleted', all unchecked. The 'Distances' section has three options: 'Mahalanobis', 'Cook's', and 'Leverage values', with 'Leverage values' checked. The 'Prediction Intervals' section has two options: 'Mean' and 'Individual', both unchecked, and a 'Confidence Interval' of 95%. The 'Influence Statistics' section has five options: 'DfBetas', 'Standardized DfBetas', 'DfFits', 'Standardized DfFits', and 'Covariance ratios', all unchecked. The 'Coefficient statistics' section has two options: 'Create coefficient statistics' (unchecked) and 'Create a new dataset' (selected). The 'Create a new dataset' option has a 'Dataset name' field. The 'Write a new data file' option has a 'File...' button. The 'Export model information to XML file' section has a text field and a 'Browse...' button. The 'Include the covariance matrix' checkbox is checked. At the bottom are 'Continue', 'Cancel', and 'Help' buttons.

Linear Regression: Save

Predicted Values

- ☐ Unstandardized
- ☐ Standardized
- ☐ Adjusted
- ☐ S.E. of mean predictions

Residuals

- ☐ Unstandardized
- ☐ Standardized
- ☐ Studentized
- ☐ Deleted
- ☐ Studentized deleted

Distances

- ☐ Mahalanobis
- ☐ Cook's
- ☒ Leverage values

Prediction Intervals

- ☐ Mean ☐ Individual
- Confidence Interval: 95 %

Influence Statistics

- ☐ DfBetas
- ☐ Standardized DfBetas
- ☐ DfFits
- ☐ Standardized DfFits
- ☐ Covariance ratios

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

Can have SPSS
provide outliers

Linear Regression: Statistics

Regression Coefficients

☒ 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

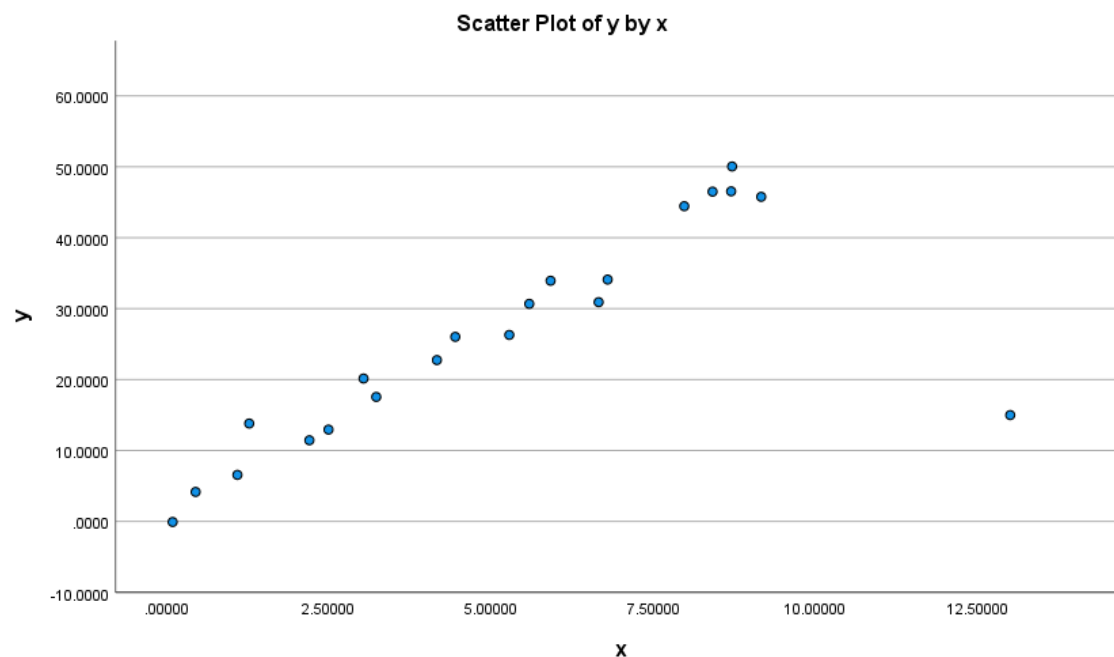
☐ All cases

Continue Cancel Help

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			

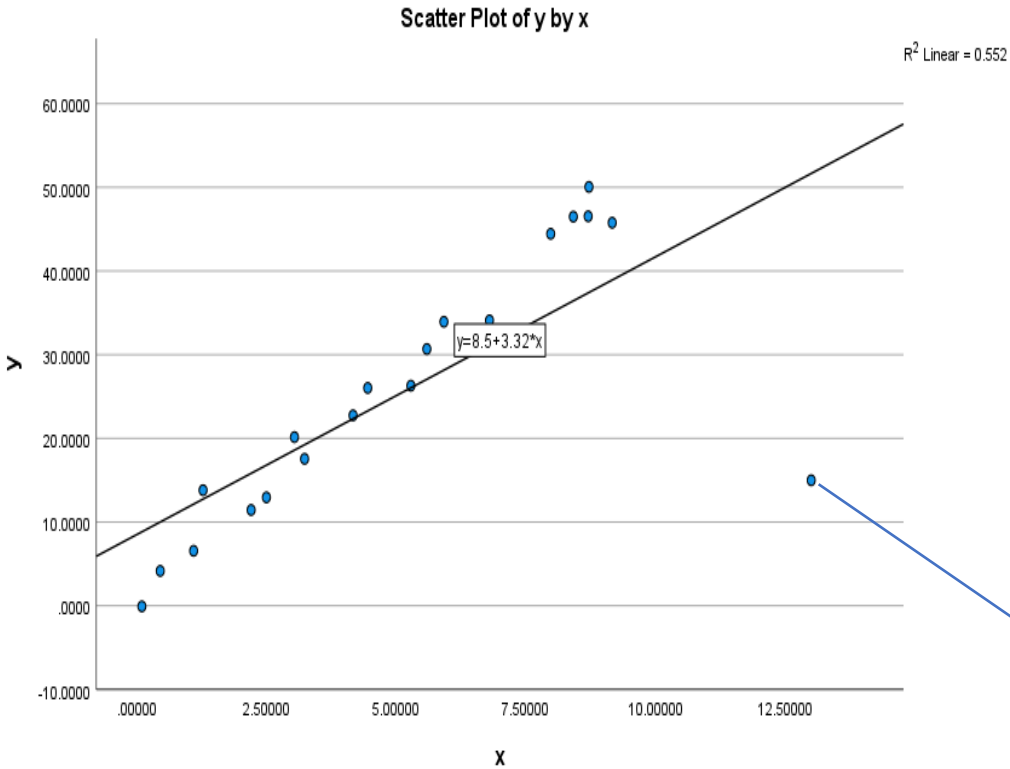
a. Dependent Variable: y

b. Predictors: (Constant), x

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
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

Introduction to Logistic Regression

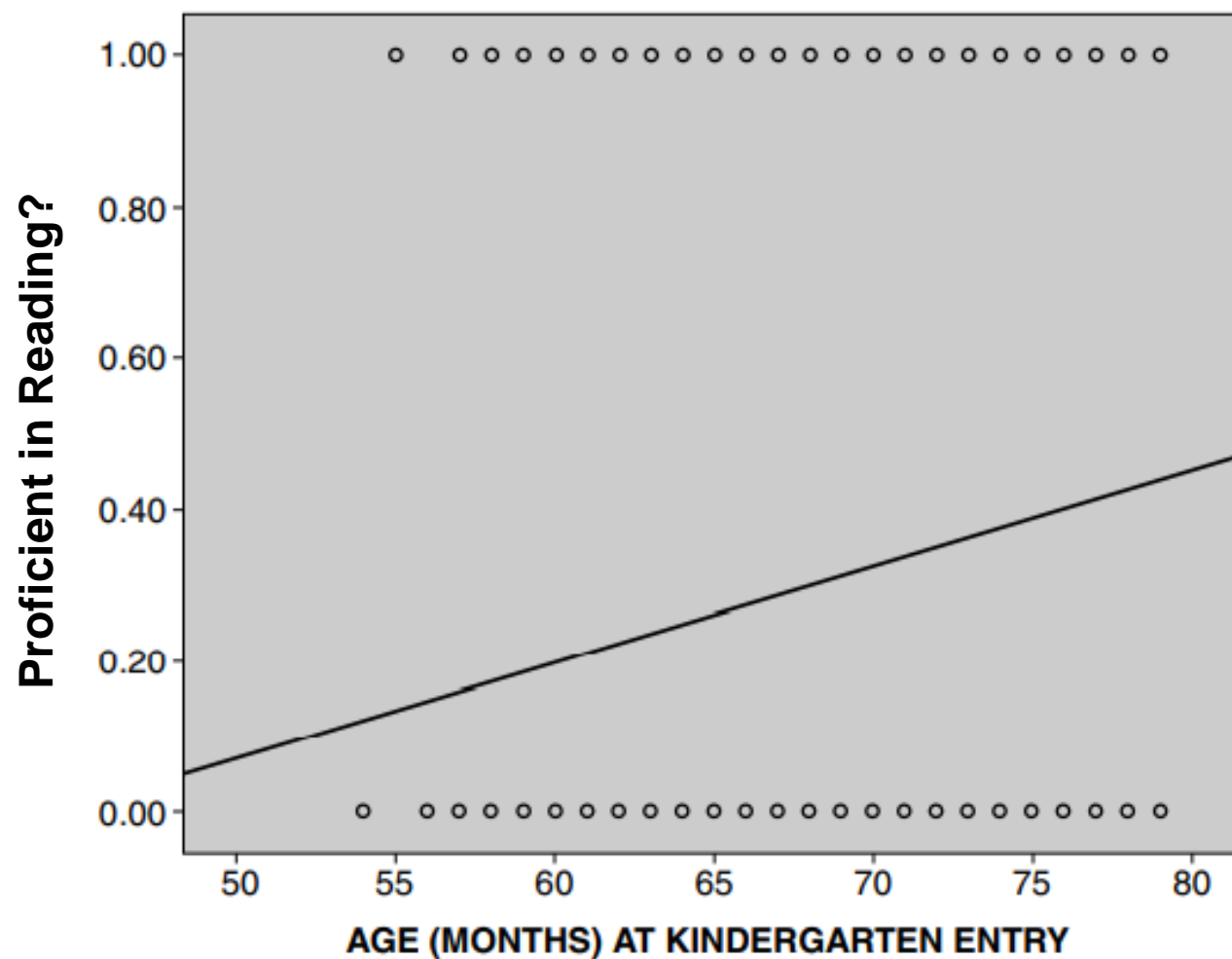
Outcome is **categorical** and allows modeling **prediction**

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 of logistic regression

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

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 beta
 - **Problem:** nobody knows how to interpret a log odds!
- 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

Significance tests: Model fit

- Overall logistic regression model: how well do the predicted values “fit” with the empirical observations (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

Model fit

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 likelihood function (from MLE) 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
- Is used as an indicator of model fit
- *LL* values range from zero to negative infinity (select smaller, i.e., more negative, values)

Change in Log Likelihood

Test of overall model fit

- Useful when **comparing two nested models (give example of a nested model)**
- 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

$$H_0: \beta_1 = \beta_2 \dots \beta_m = 0 \quad H_1: \text{at least one } \beta \neq 0$$

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 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 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, β_k

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

The test statistic is called a Wald test computed as

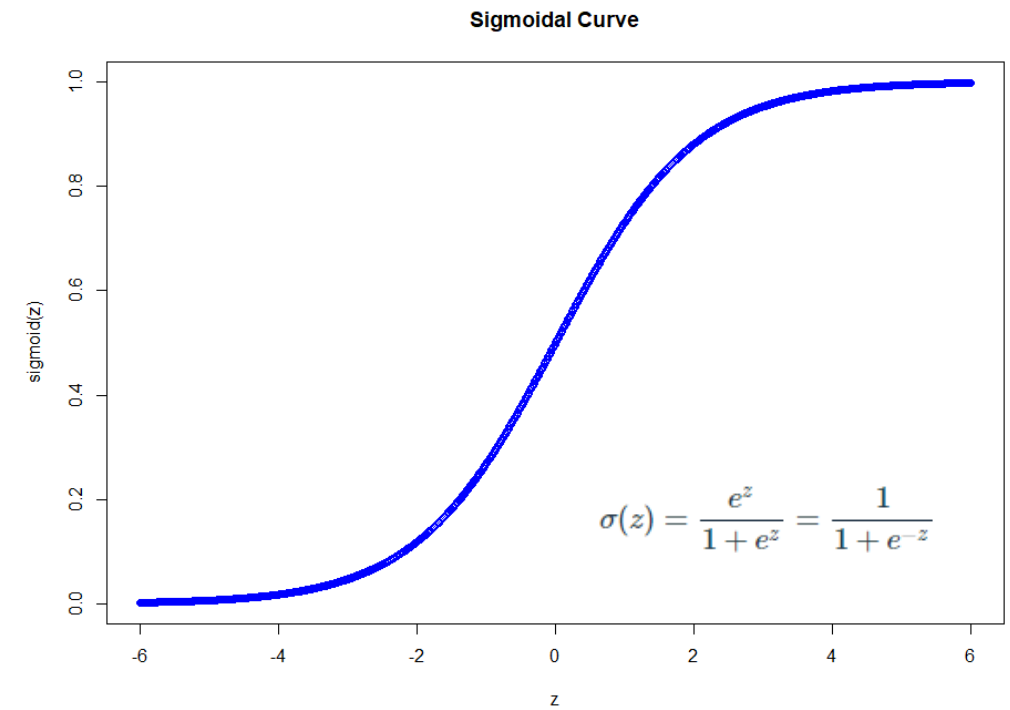
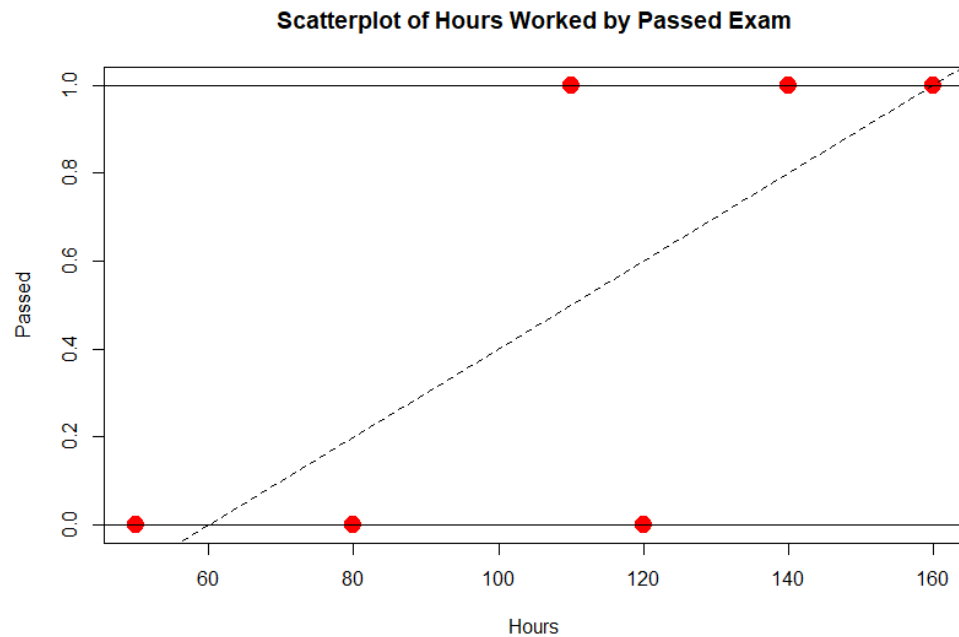
$$W = \frac{\beta_k^2}{SE_{\beta_k^2}} \sim \chi_2$$

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

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

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

$$e^0 = 1$$

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

Mathematical Snapshot

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

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

Assume one independent variable

Snapshot

Exponentiate the logit and convert back to odds

$$\text{Odds}(Y = 1) =$$

$$e^{\text{logit}(Y)} =$$

$$e^{\ln[\text{Odds}(Y=1)]} =$$

$$e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m} =$$

$$(e^{\beta_0})(e^{\beta_1 X_1})(e^{\beta_2 X_2}) \dots (e^{\beta_m X_m})$$

- Why this matters? 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).
- Coefficients greater than 1 increase the odds, and coefficients less than 1 decrease the odds.
- In addition, the odds will change *nonlinearly* 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 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`

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 two slides show how the classification table is interpreted and how we can add the probabilities and classifications to our dataset for additional intuition

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

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

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
Observed		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

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

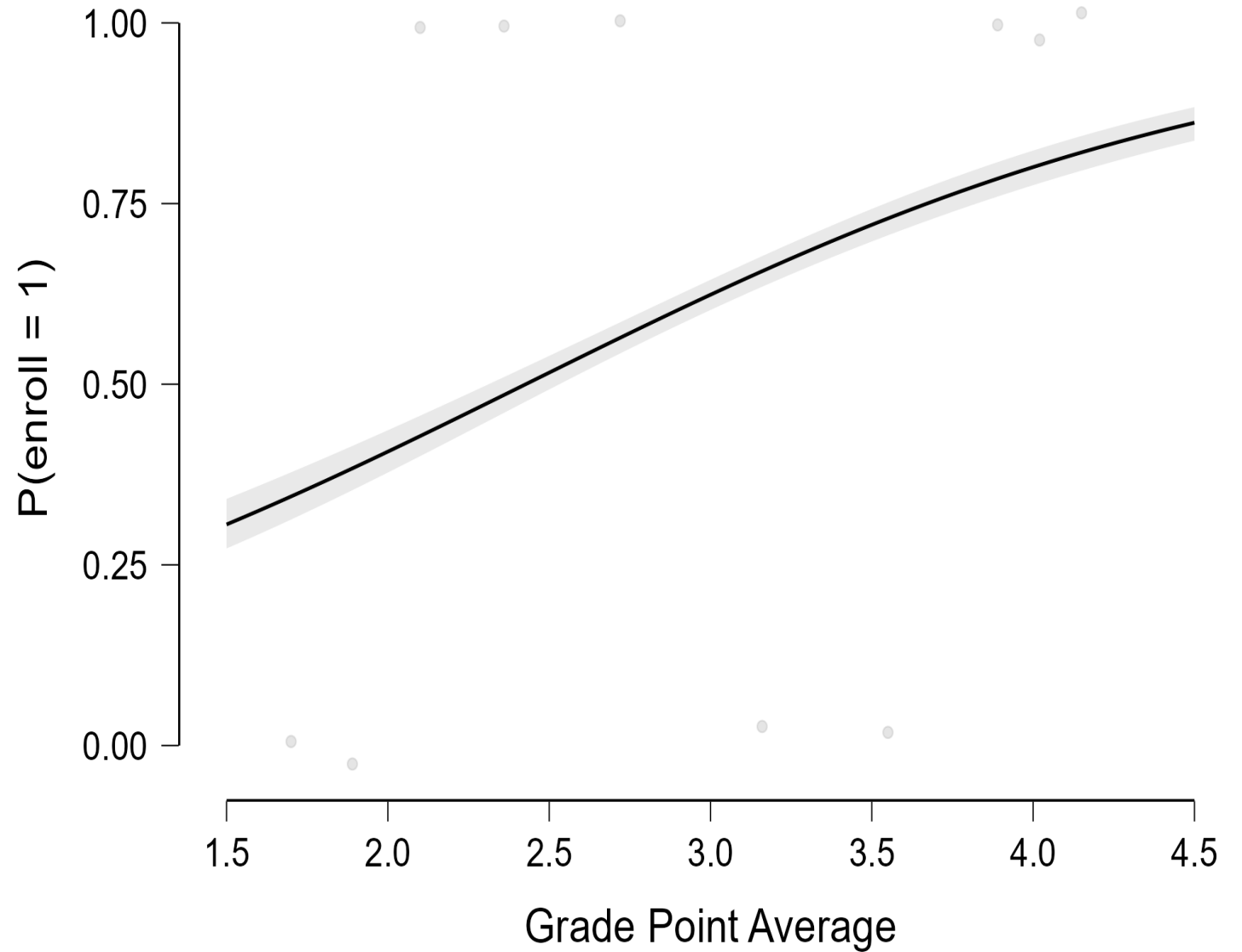
In this case, the results are no better than chance

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



Your turn

- Use the jasp file gpa-college-enroll.JASP
- 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 enrolled on gpa and gender
 - Interpret coefficients
 - Does adding gender improve the model fit?

You can use SPSS its just MUCH more difficult

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

- Use the jasp file gpa-college-enroll.JASP
- 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: $\exp(.082) = 1.085$

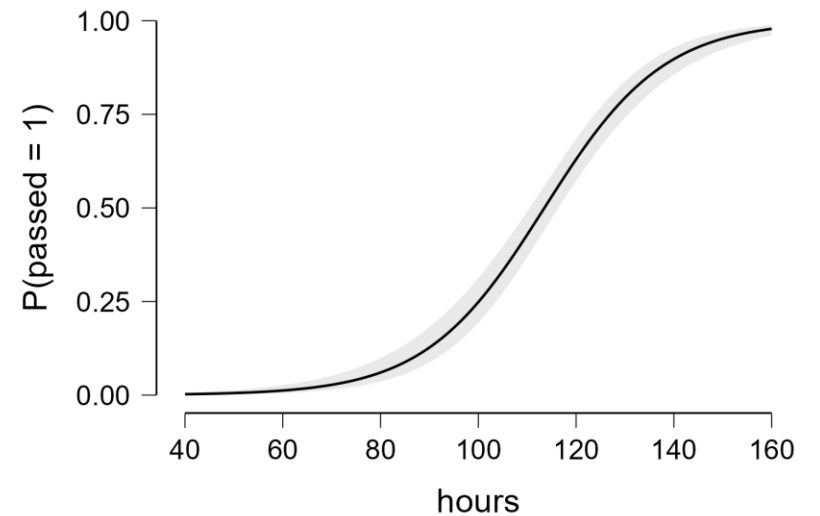


Table 1. Predicted probability of Passing Conditional on Hours Studying

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

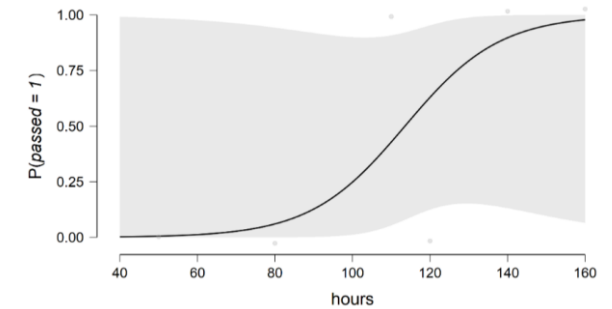
Model Summary - passed ▼

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.

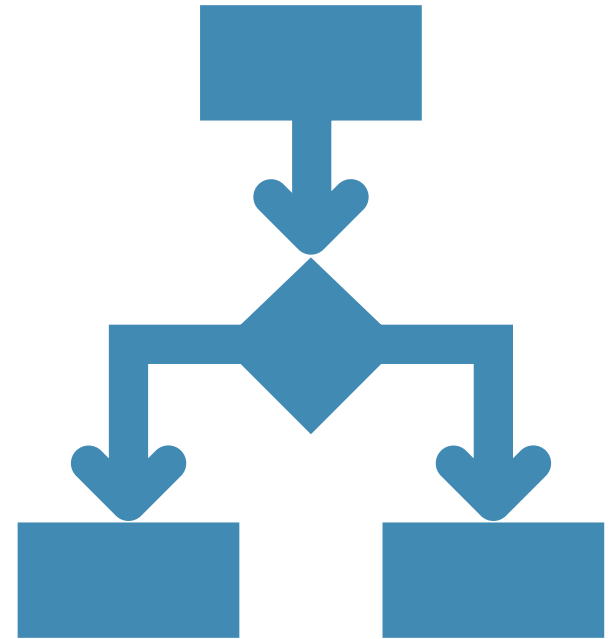
The odds of passing for males are $[\exp(-.654)-1 * 100]$ percent lower compared to females

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

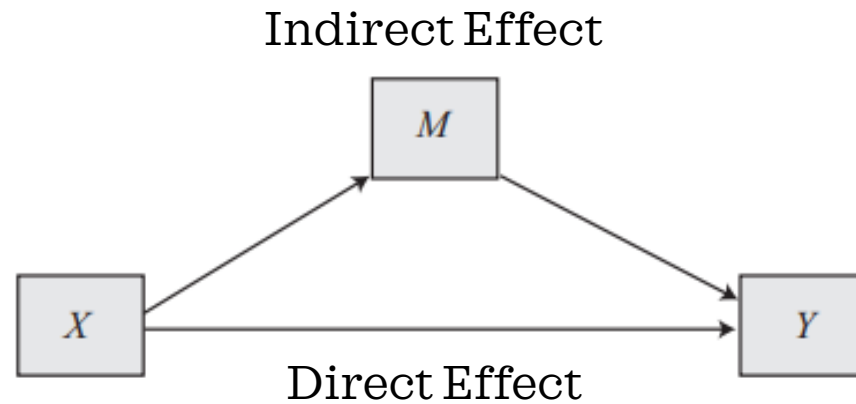
The odds of passing are .5100 ($\exp(-.654)$) times lower for males compared to females

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

MEDIATION, MODERATION & CONDITIONAL PROCESS ANALYSIS



Introduction



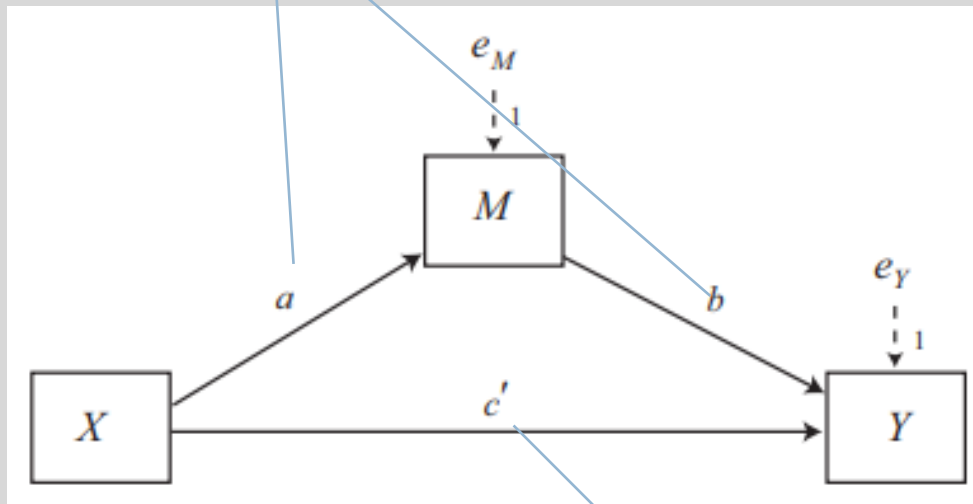
- Mediation analysis is a statistical method used to evaluate evidence from studies designed to test hypotheses about how some causal antecedent variable X transmits its effect on a consequent variable Y
- RQ: What is the mechanism, be it emotional, cognitive, biological, or otherwise, by which X influences Y ?

Conceptualizing a Mediation Process

- Mediation is ultimately a causal explanation hence the *assumptions are that*
 - The relationships in the system are causal
 - M is causally located between X and Y
 - X causes M, which in turn causes Y
→ M is located causally between X and Y
- Note about cross-sectional data

The Statistical Model

The indirect effect of X on Y



The direct effect of X on Y

- The diagram represents two equations

$$M = i_M + aX + e_M$$

$$Y = i_Y + c'X + bM + e_Y$$

- The analytical goal is to estimate these coefficients, piece them together, and interpret
- In this scheme, a , b and c' are unknown parameters, i.e., regression coefficients
- We are predicting M and Y so there should be ***two separate equations***

The Direct effect of X on Y

- c' estimates the direct effect of X on Y
- **Interpretation:** two cases that differ by one unit on X **but are equal on M** differ by c' units on Y, or
$$c' = [\hat{Y}|(X = x, M = m)] - [\hat{Y}|(X = x - 1, M = m)]$$
- **Note:** in the special case where X is dichotomous, c' is the difference between the two groups' means holding M constant
 - This is equivalent to an adjusted mean differences in ANOVA

The Indirect Effect of X on Y

- a quantifies how much two cases that differ by one unit on X are estimated to differ on M

$$a = [\hat{M}|(X = x)] - [\hat{M}|(X = x - 1)]$$

- b quantifies how much two cases that differ by one unit on M but that are equal on X are estimated to differ by b units on Y

$$b = [\hat{Y}|(M = m, X = x)] - [\hat{Y}|(M = m - 1, X = x)]$$

- The indirect effect of X on Y through M is the product of a and b
 - The indirect effect tells us that two cases that differ by one unit on X are estimated to differ by **ab units on Y** as a result of the effect of X on M (which, in turn, affects Y)
 - **Note:** you must consider the signs of a and b when interpreting the indirect effect

The Total Effect of X on Y

- The total effect = the direct effect + the indirect effect

$$c = c' + ab \rightarrow ab = c - c'$$

- The total effect c quantifies how much two cases that differ by one unit on X are estimated to differ on Y
- c represents the change in Y for every 1-unit increase in X

$$c = [\hat{Y}|X = x] - [\hat{Y}|X = x - 1]$$

- The difference between the total effect and the direct effect is that the direct effect controls for M

The Process Macro

- To run these models you need to download and install the Process macro written by Andrew Hayes
- [Download - The PROCESS macro for SPSS, SAS, and R](#)
- Unzip the file
- From SPSS go to Extensions → Utilities → Install Custom Dialog Builder
- Browse to the unzipped files and open the folder that contains the SPSS application (C:\...\Downloads\processv40\PROCESS v4.0 for SPSS\Custom dialog builder file\process)
- The dialog files will be located in Analyze → Regression → process.spd
- The latest version is 4.0

Example 1: Dichotomous Predictor, Simple Mediation

The PMI study

- The participants in this study (43 male and 80 female) read one of two newspaper articles describing an economic crisis that may affect the price and supply of sugar in Israel
- Approximately half of the participants (n = 58) were given an article they were told would be appearing on the front page of a major Israeli newspaper (i.e., the 'front-page' condition).
- The remaining participants (n = 65) were given the same article but were told it would appear in the middle of an economic supplement of this newspaper (i.e., the interior page condition)
- The participants were asked about their reactions to the story including their intention to buy sugar (REACTION)
- They were also asked questions about how much they believed others would buy sugar in the community (*presumed media influence* measured by PMI))
- The research hypothesis was that individuals who read the article on the front page would be more likely to ***think others would promptly buy sugar*** which in turn would lead them to buy sugar

COND: front (1) or interior (0) page of the newspaper = interior

Descriptive Statistics^a

	N	Minimum	Maximum	Mean	Std. Deviation
PMI: presumed media influence	65	1.50	7.00	5.3769	1.33765
IMPORT: article is on an important topic	65	1.00	7.00	3.9077	1.65570
REACTION: sugar purchase	65	1.00	6.75	3.2500	1.60809
GENDER: female (0) or male (1)	65	.00	1.00	.2923	.45836
AGE: age	65	18.00	60.00	24.4923	5.26523
Valid N (listwise)	65				

a. COND: front (1) or interior (0) page of the newspaper = interior

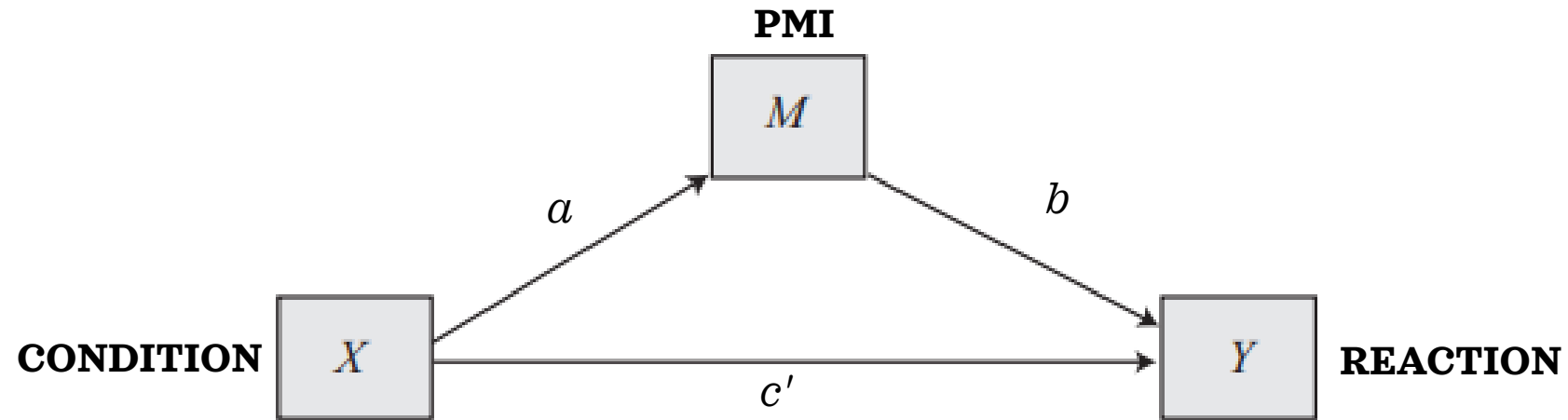
COND: front (1) or interior (0) page of the newspaper = front

Descriptive Statistics^a

	N	Minimum	Maximum	Mean	Std. Deviation
PMI: presumed media influence	58	1.00	7.00	5.8534	1.26702
IMPORT: article is on an important topic	58	1.00	7.00	4.5345	1.77917
REACTION: sugar purchase	58	1.25	7.00	3.7457	1.45208
GENDER: female (0) or male (1)	58	.00	1.00	.4138	.49681
AGE: age	58	19.00	61.00	24.7845	6.39155
Valid N (listwise)	58				

a. COND: front (1) or interior (0) page of the newspaper = front

DESCRIPTIVE STATS



SIMPLE MEDIATION MODEL FOR PMI STUDY

Steps for Mediation, Moderation and Mediated Moderation

1

Select the model
from the appendix

2

Specify the X , Y , M
and W variables

3

Incorporate
covariates/controls
for M and Y

4

Open the process
macro and run the
analysis

PROCESS_v4.0

Variables:

- IMPORT: article is on an important to...
- GENDER: female (0) or male (1) [gen...]
- AGE: age [age]

Y variable:

REACTION: sugar ...

X variable:

COND: front (1) or i...

Mediator(s) M:

PMI: presumed me...

Covariate(s):

Model number:

4

Confidence intervals

95

Number of bootstrap samples

5000

☐ Save bootstrap estimates

☐ Bootstrap inference for model coefficients

Do not use PASTE button

OK Paste Reset Cancel Help

About Options Multicategorical Long variable names

PROCESS options

☐ Show covariance matrix of regression coefficients

☐ Generate code for visualizing interactions

☒ Show total effect model (only models 4, 6, 80, 81, 82)

☐ Pairwise contrasts of indirect effects

☐ Standardized effects (mediation-only models)

☐ Test for X by M interaction(s)

☐ Residual correlations

Heteroscedasticity-consistent inference

None

Decimal places in output

4

Mean center for construction of products

☒ No centering

☐ All variables that define products

☐ Only continuous variables that define products

Moderation and conditioning

Probe interactions...

if $p < .10$

Conditioning values

☒ 16th, 50th, 84th percentiles

☐ -1SD, Mean, +1SD

☐ Johnson-Neyman output

Many options available in PROCESS through command syntax are not available through this dialog box. See Appendices A and B of <http://www.guilford.com/p/hayes3>

Continue Cancel

Does the condition (i.e., placement) affect PMI (i.e. the mediator)?

OUTCOME VARIABLE:

pmi - this is the mediator variable

Model Summary

R	R-sq	MSE	F	df1	df2	p
.1808	.0327	1.7026	4.0878	1.0000	121.0000	.0454

Model

	coeff	se	t	p	LLCI	ULCI
constant	5.3769	.1618	33.2222	.0000	5.0565	5.6973
cond	.4765(a)	.2357	2.0218	.0454	.0099	.9431

$$\hat{M} = 5.377 + .477X \rightarrow$$

People who were given the front-page article condition are .477 units higher on PMI, on average, meaning they were more likely to believe others would buy sugar (PMI)

Is PMI and article condition associated with intention to buy (reaction)?

OUTCOME VARIABLE:

reaction → the dependent variable (intention to buy)

Model Summary

R	R-sq	MSE	F	df1	df2	p
.4538	.2059	1.9404	15.5571	2.0000	120.0000	.0000

Model

	coeff	se	t	p	LLCI	ULCI
constant	.5269	.5497	.9585	.3397	-.5615	1.6152
cond	.2544 (C')	.2558	.9943	<u>.3221</u>	-.2522	.7609
pmi	.5064 (b)	.0970	5.2185	<u>.0000</u>	.3143	.6986

$$\hat{Y} = .527 + .254X + .506M \rightarrow$$

Every one-unit increase in presumed media influence leads to a .5064 increase in intention to buy sugar regardless of whether participants read the front-page condition or interior condition

Is there a mediation effect?

Direct effect of X on Y

Effect	se	t	p	LLCI	ULCI
.2544	.2558	.9943	.3221	<u>-.2522</u>	<u>.7609</u>

Indirect effect(s) of X on Y:

	Effect	BootSE	BootLLCI	BootULCI
pmi	.2413	.1291	<u>.0017</u>	<u>.5098</u>

$c = ab \rightarrow .4765 (.5064) = .2413$ **Note:** check that my labels associated with the coefficients for a and b are equal to the indirect effect when multiplied

Does the effect of article condition on intention to buy *operate through presumed media influence*?

The Bootstrap confidence interval does not contain 0 and hence the result is significant

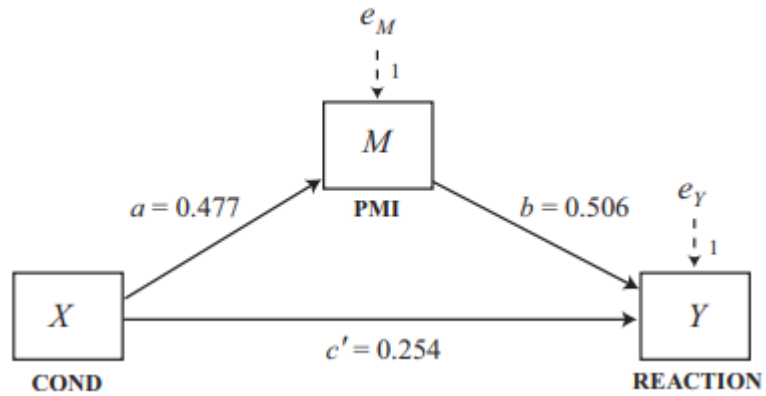


Table. Model Coefficients for Mediation Analysis

		Consequent						
		M(PMI)			Y(REACTION)			
Antecedent		Coeff.	SE	p				
X(COND)	a	.477	.236	.045	c'	.254	.256	.322
M(PMI)	--	--	--	--	b	.506	.097	<.001
Constant	iM	5.377	.162	<.001	iY	.527	.550	.340
		$R^2=.033$ $F(1,121) = 4.088, p = .045$				$R^2=.206$ $F(2,120)=15.557, p < .001$		