“How to Build a Regression Model”

Melinda K. Higgins, Ph.D.

14 November 2008

Outline

I. Basics
II. Assumptions
III. Linear Relationships
IV. Understanding Correlation
V. 3 Types of Linear Regression (theory development)
VI. Diagnostics of Model Fit
VII. Problems with Multicollinearity
VIII. Outliers
IX. References

Basics

- Dependent Variable
  - Continuous
  - Single
- Independent Variables
  - Continuous or Discrete (dummy coding)
  - Single or Multiple
- “Linear” Relationship – discussion following

Assumptions

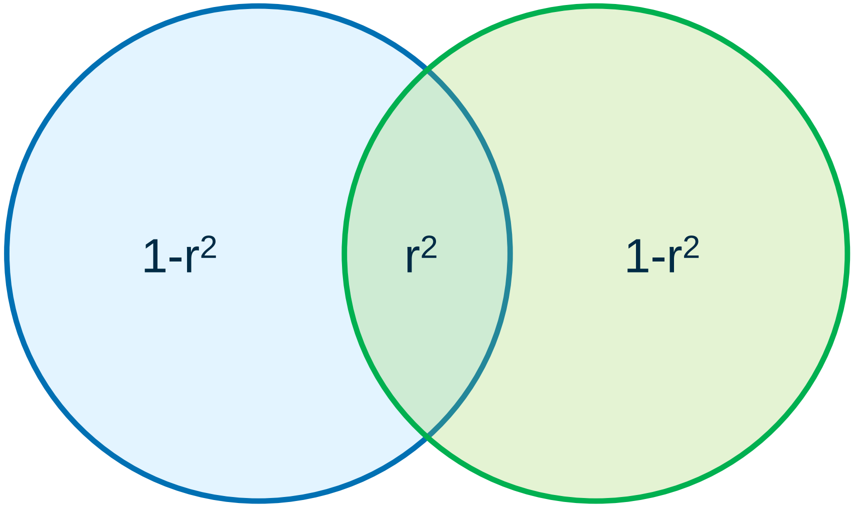
- The IVs (independent variables) have a relationship with the DV (dependent variable)
- That relationship is linear (either directly or through transformations)
- The independent variables are independent of one another (no multicollinearity) – although some "mild/minor" correlation may be tolerated.
- Any case which has "missing" data on any of the IVs or DV will be eliminated from the analysis
- The intercept and all coefficients for the IVs are "Fixed." ["Random Coefficient Models" to be discussed later.]

Linear Relationships

Which are Linear?

- ☑ $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
- ☑ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$
- ☑ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- ☑ $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2(\mathrm{sqrt}(X_2)) + \varepsilon$
- ? $Y = \beta_0 X_1^{\beta 1} X_2^{\beta 2} e^{\varepsilon}$ Take ln() natural log of both sides
- ☑ $[\ln Y = \ln\beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \varepsilon]$
- ☒ $Y = \beta_0 X_1^{\beta 1} X_2^{\beta 2} + \varepsilon$ *Intrinsically nonlinear*

Correlation (2 variables)



Suppose $r = 0.5$, then $r^2 = 0.25$ and $1-r^2 = 0.75$

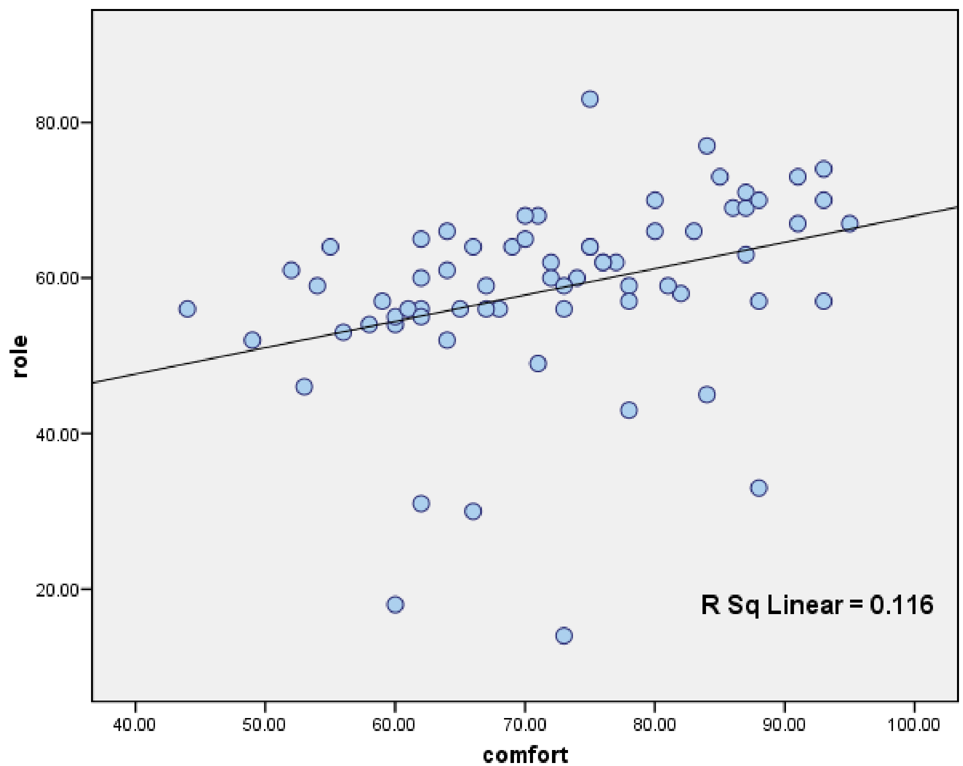
- **Venn Diagram ["ballantine"] - where each circle represents the variance of that variable**
- **The overlap in circles represents the degree of correlation (or r^2).**

•Just because we know r2 – MUST STILL MAKE PICTURES!!

Correlations

		comfort	role	involvement
comfort	Pearson Correlation	1	.341**	.162
	Sig. (2-tailed)		.004	.165
	N	76	68	75
role	Pearson Correlation	.341**	1	.381**
	Sig. (2-tailed)	.004		.001
	N	68	68	67
involvement	Pearson Correlation	.162	.381**	1
	Sig. (2-tailed)	.165	.001	
	N	75	67	75

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



$(0.341)^2 = .116$

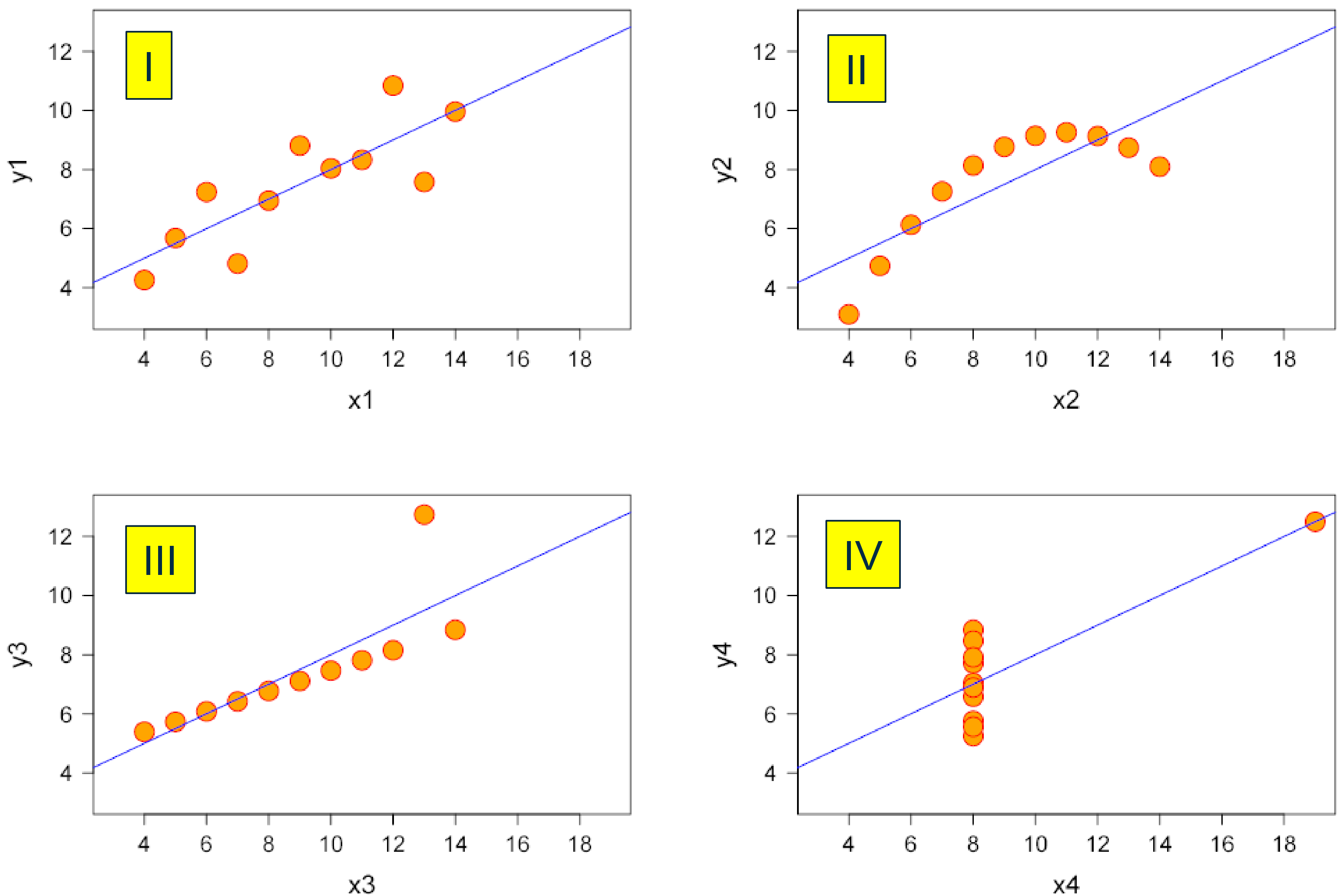
4 similar datasets – they have identical regression results – or do they?

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

For all four datasets:

Property	Value
Mean of each x variable	9.0
Variance of each x variable	10.0
Mean of each y variable	7.5
Variance of each y variable	3.75
Correlation between each x and y variable	0.816
Linear regression line	y = 3 + 0.5x

But what do they look like? →


I
II
III
IV
y1
y2
y3
y4
x1
x2
x3
x4
4
6
8
10
12
14
16
18

See http://en.wikipedia.org/wiki/Anscombe%27s_quartet

F.J. Anscombe, "Graphs in Statistical Analysis," American Statistician, 27 (February 1973), 17-21.

Anscombe's quartet - From Wikipedia, the free encyclopedia

Anscombe's quartet comprises four datasets which have identical simple statistical properties, yet which are revealed to be very different when inspected graphically. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician F.J. Anscombe to demonstrate the importance of graphing data before analyzing it, and of the effect of outliers on the statistical properties of a dataset.

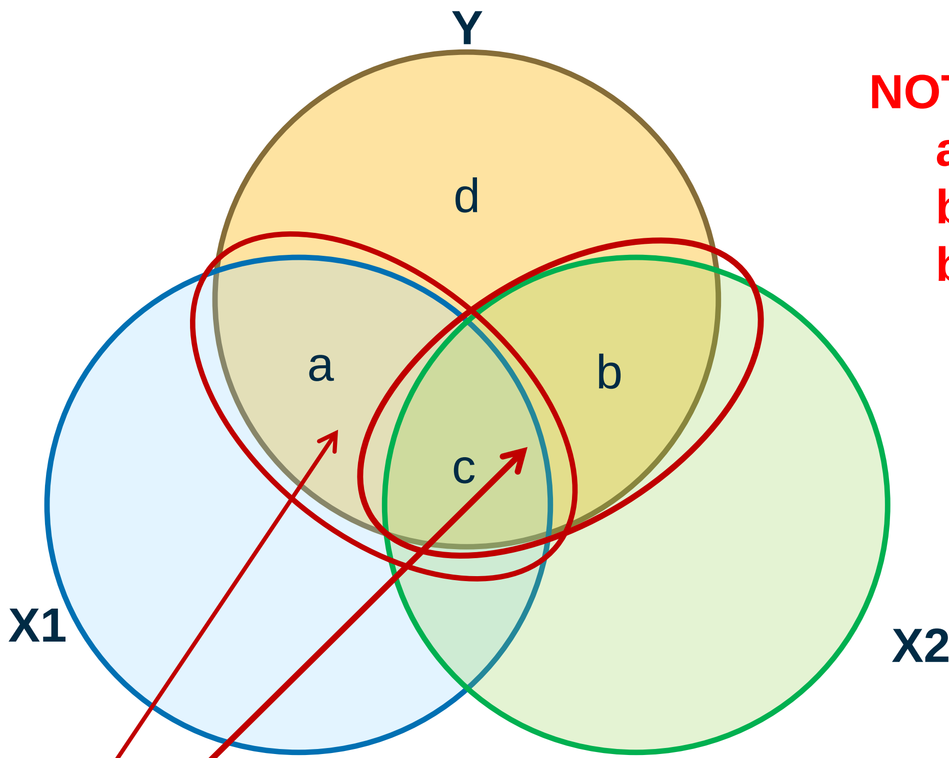
- The first one (top left) seems to be distributed normally, and corresponds to what one would expect when considering two variables correlated and following the assumption of normality.

- The second one (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant.

- In the third case (bottom left), the linear relationship is perfect, except for one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.81.

- Finally, the fourth example (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

Correlation (among DV and IVs)



NOTE:
a is always > 0
b is always > 0
but c can be + or -

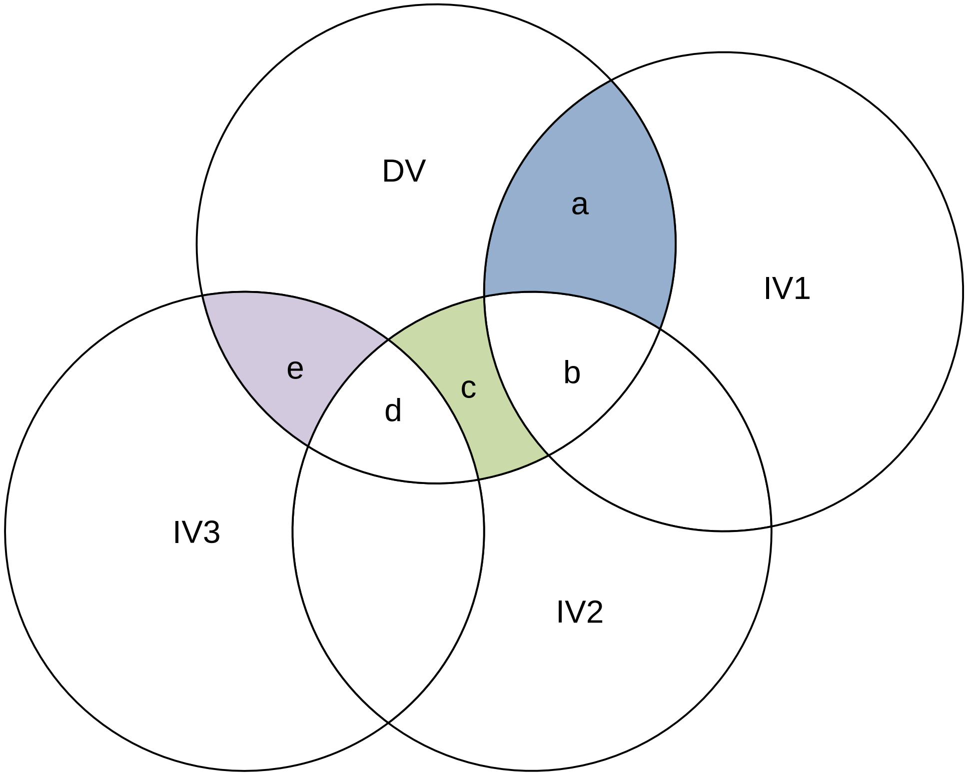
Y with X_1 -- r^2_{y1} = a+c
Y with X_2 -- r^2_{y2} = b+c
Y with X_1 and X_2 -- r^2_{y12} = a+b+c

[semi-partial, controlling for X_2]
$a = sr^2_1 = r^2_{y12} - r^2_{y2}$
[semi-partial, controlling for X_1]
$b = sr^2_2 = r^2_{y12} - r^2_{y1}$

Standard vs Sequential vs Stepwise

- Standard – all variables enter at one time [“ENTER”]
- Sequential (also sometimes called “Hierarchical Regression”) – variables enter in specific (sequential) order [“BLOCK1” “BLOCK2” etc]
- Stepwise (also called “Statistical Regression”) – variables are allowed to “compete statistically” [“Forward, Backward”]

Standard – X_1 X_2 and X_3 all at once



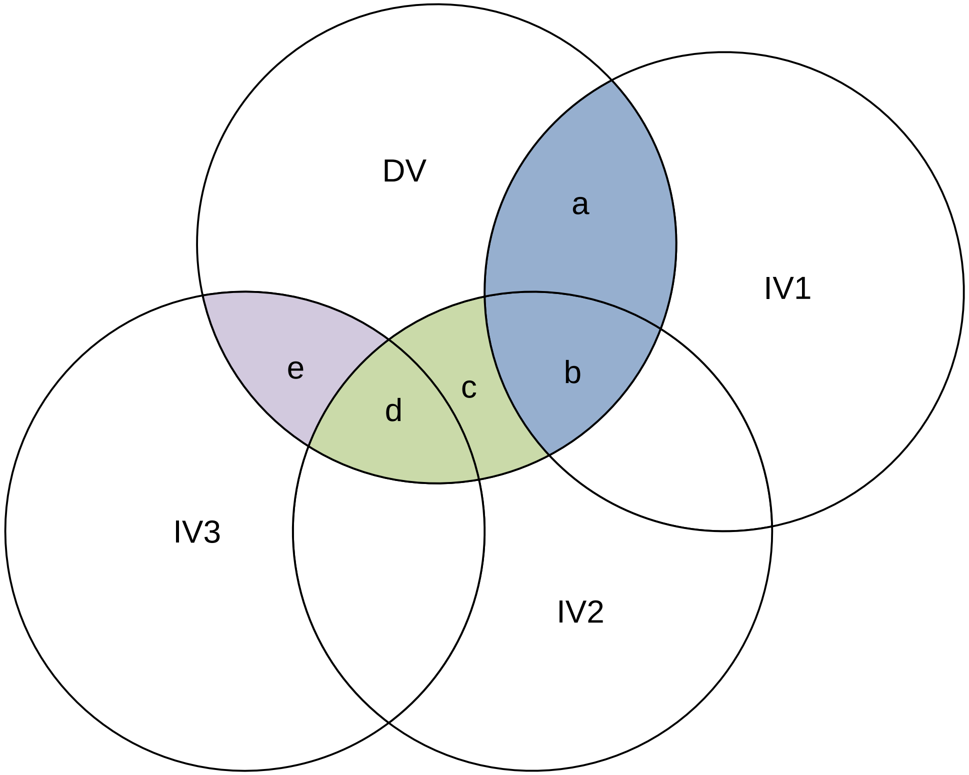
R^2 = a+b+c+d+e

But who gets credit for each piece?

IV_1 gets credit for a
IV_2 gets credit for c
IV_3 gets credit for e

b and c do not get "assigned" to any of the 3 IVs

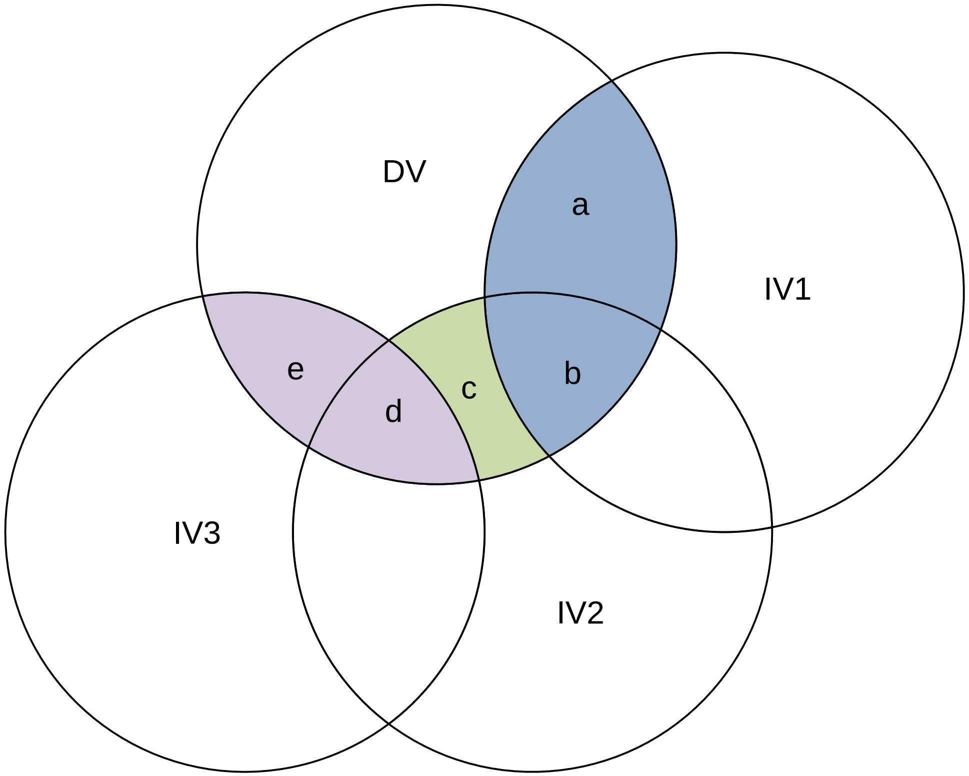
Sequential – X_1 then X_2 then X_3



R^2 = a+b+c+d+e

IV_1 gets credit for a + b
IV_2 gets credit for c + d
IV_3 gets credit for e

Stepwise (statistical) – X_1 X_2 and X_3 "compete"



R^2 = a+b+c+d+e

Assume the "amount of variance explained" is
- highest for X1
- followed by X3
- followed by X2

IV_1 gets credit for a+b
IV_2 gets credit for c
IV_3 gets credit for e+d

Exam Anxiety Example (A. Field Book)

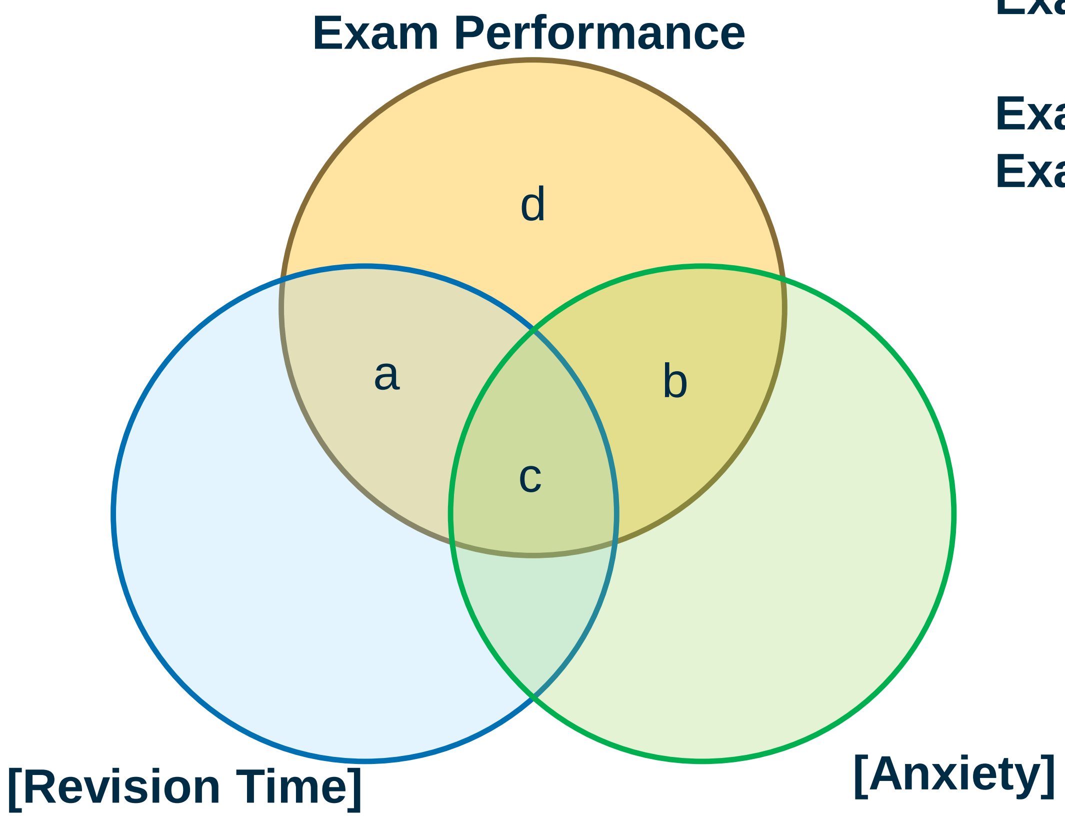
Compare Exam Performance (DV=Y) against Time Spent Revising Exam (IV1=X1) and Anxiety Level (IV2=X2) [Correlation Matrix]

Correlations

		exam Exam Performance (%)	anxiety Exam Anxiety	revise Time Spent Revising
exam Exam Performance (%)	Pearson Correlation	1	-.441**	.397**
	Sig. (2-tailed)		.000	.000
	Sum of Squares and Cross-products	68637.204	-20048.511	19061.592
	Covariance	672.914	-196.554	186.878
	N	103	103	103
anxiety Exam Anxiety	Pearson Correlation	-.441**	1	-.709**
	Sig. (2-tailed)	.000		.000
	Sum of Squares and Cross-products	-20048.511	30112.058	-22571.667
	Covariance	-196.554	295.216	-221.291
	N	103	103	103
revise Time Spent Revising	Pearson Correlation	.397**	-.709**	1
	Sig. (2-tailed)	.000	.000	
	Sum of Squares and Cross-products	19061.592	-22571.667	33634.816
	Covariance	186.878	-221.291	329.753
	N	103	103	103

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

Assigning Variance



Exam-[Anxiety+Revision] R^2 = a+b+c

Exam-[Anxiety] $r^2 = (-.441)^2 = .194$
Exam-[Revision] $r^2 = (.397)^2 = .157$

Partial Correlation

Correlations

Control Variables			exam Exam Performance (%)	anxiety Exam Anxiety	revise Time Spent Revising
-none-[a]	exam Exam Performance (%)	Correlation	1.000	-.441	.397
		Significance (2-tailed)	.	.000	.000
		df	0	101	101
	anxiety Exam Anxiety	Correlation	-.441	1.000	-.709
		Significance (2-tailed)	.000	.	.000
		df	101	0	101
	revise Time Spent Revising	Correlation	.397	-.709	1.000
		Significance (2-tailed)	.000	.000	.
		df	101	101	0
revise Time Spent Revising	exam Exam Performance (%)	Correlation	1.000	-.247	
		Significance (2-tailed)	.	.012	
		df	0	100	
	anxiety Exam Anxiety	Correlation	-.247	1.000	
		Significance (2-tailed)	.012	.	
		df	100	0	

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

When Revision is controlled for, the squared correlation between Exam Performance and Anxiety is reduced from $(-.441)^2 = .194$ down to only $(-.247)^2 = .061$.

Semi-Partial Correlation [obtained from Sequential Regression]

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.397[a]	.157	.149	23.92947	.157	18.865	1	101	.000
2	.457[b]	.209	.193	23.30573	.051	6.479	1	100	.012

a. Predictors: (Constant), revise Time Spent Revising

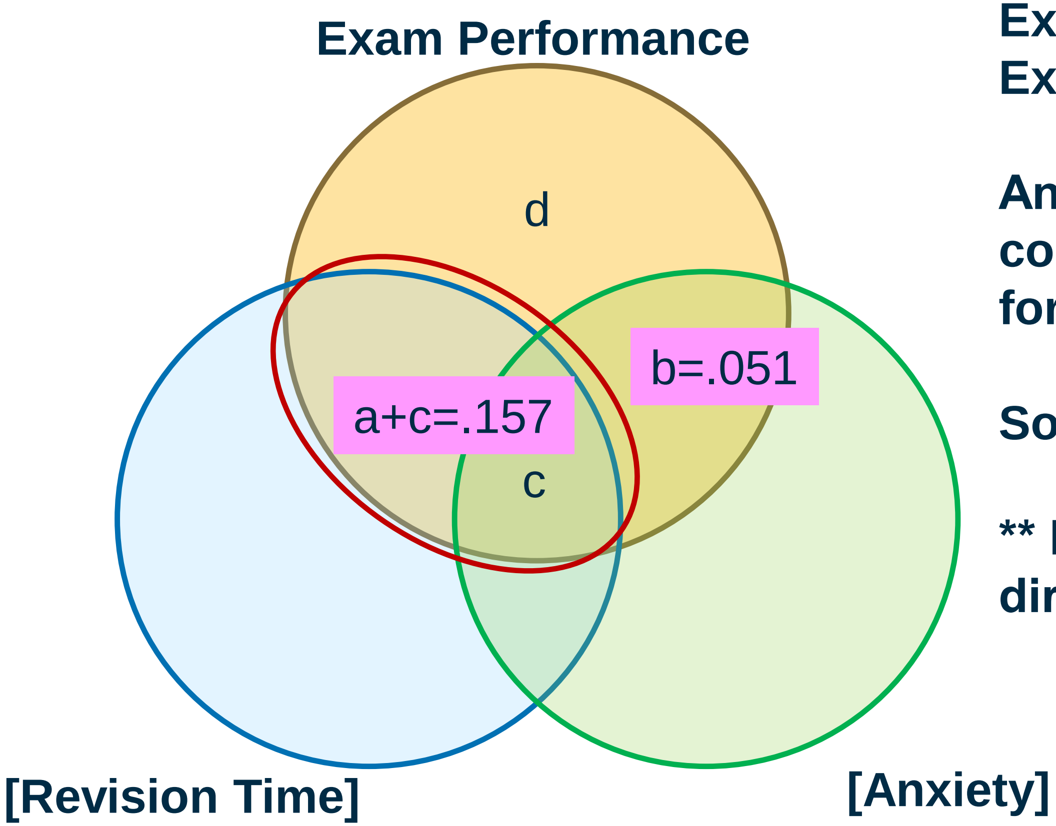
b. Predictors: (Constant), revise Time Spent Revising, anxiety Exam Anxiety

c. Dependent Variable: exam Exam Performance (%)

We knew this from the Initial Correlation Matrix Exam-[Revision] $r^2 = (.397)^2 = .157$

The amount of variance "left over" after "controlling for" Revision, for Anxiety is .051 (5.1%).

Back to the Venn Diagram



Exam-[Anxiety+Revision] R^2 = a+b+c

Exam-[Anxiety] $r^2 = (-.441)^2 = .194$
Exam-[Revision] $r^2 = (.397)^2 = .157$

And now we know the "semi-partial" correlation for Anxiety (controlling for Revision Time) = .051.

So, overall r^2=.209=.157+.051

** Remember we don't know c directly and we can't get d directly**

R^2 Breakdown (overall r^2_{x1x2}=0.209)

Standard Regression (X1 and X2 together)	= or ≠	Sequential Regression (X1 then X2)
$r^2_{X1} = (a+c)/(a+b+c+d) = 0.157$ (revision)	=	$r^2_{X1} = (a+c)/(a+b+c+d) = 0.157$
$r^2_{X2} = (b+c)/(a+b+c+d) = 0.194$ (anxiety)	=	$r^2_{X2} = (b+c)/(a+b+c+d) = 0.194$
$sr^2_{X1} = a/(a+b+c+d) = 0.014$ (“Part” in SPSS)	≠	$sr^2_{X1} = (a+c)/(a+b+c+d) = 0.157$ (r2 change)
$sr^2_{X2} = b/(a+b+c+d) = 0.051$	=	$sr^2_{X2} = b/(a+b+c+d) = 0.051$
$pr^2_{X1} = a/(a+d) = 0.018$ (“Partial” in SPSS)	≠	$pr^2_{X1} = (a+c)/(a+c+d) = 0.1576$ (“Partial” in SPSS)
$pr^2_{X2} = b/(b+d) = 0.061$	=	$pr^2_{X2} = b/(b+d) = 0.061$

From p.145 – Tabachnick – but notation based on previous Venn, where “c” represents the overlap of X1 and X2

Separate Regressions

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.441[a]	.194	.186	23.39691	.194	24.384	1	101	.000

a. Predictors: (Constant), anxiety Exam Anxiety

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.397[a]	.157	.149	23.92947	.157	18.865	1	101	.000

a. Predictors: (Constant), revise Time Spent Revising

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.457[a]	.209	.193	23.30573	.209	13.184	2	100	.000

a. Predictors: (Constant), revise Time Spent Revising, anxiety Exam Anxiety

b. Dependent Variable: exam Exam Performance (%)

Standard Regressions – “correlations”

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	87.833	17.047		5.152	.000	54.012	121.653
	anxiety Exam Anxiety	-.485	.191	-.321	-2.545	.012	-.863	-.107
	revise Time Spent Revising	.241	.180	.169	1.339	.184	-.116	.599

a. Dependent Variable: exam Exam Performance (%)

Correlations Zero-order	Correlations Partial (pr)	Correlations Part (sr)	Collinearity Statistics Tolerance	Collinearity Statistics VIF
-.441	-.247	-.226	.497	2.012
.397	.133	.119	.497	2.012

Take the squares of these to get pr^2 and sr^2 respectively.

Sequential Regression – "correlations" Revision Time goes in first

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change (sr^2)	F Change	df1	df2	Sig. F Change
1	.397[a]	.157	.149	23.92947	.157	18.865	1	101	.000
2	.457[b]	.209	.193	23.30573	.051	6.479	1	100	.012

a. Predictors: (Constant), revise Time Spent Revising

b. Predictors: (Constant), revise Time Spent Revising, anxiety Exam Anxiety

c. Dependent Variable: exam Exam Performance (%)

Model		Unstandardized Coefficients		Correlations			Collinearity Statistics	
		B	Std. Error	Zero-order	Partial (pr)	Part (sr)	Tolerance	VIF
1	(Constant)	45.321	3.503					
	revise Time Spent Revising	.567	.130	.397	.397	.397	1.000	1.000
2	(Constant)	87.833	17.047					
	revise Time Spent Revising	.241	.180	.397	.133	.119	.497	2.012
	anxiety Exam Anxiety	-.485	.191	-.441	-.247	-.226	.497	2.012

a. Dependent Variable: exam Exam Performance (%)

So $(-0.226)^2 = 0.051$ ☑

If Anxiety goes in first ….

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.441[a]	.194	.186	23.39691	.194	24.384	1	101	.000
2	.457[b]	.209	.193	23.30573	.014	1.792	1	100	.184

a. Predictors: (Constant), anxiety Exam Anxiety

b. Predictors: (Constant), anxiety Exam Anxiety, revise Time Spent Revising

c. Dependent Variable: exam Exam Performance (%)

R^2 Breakdown (overall r^2_{x1x2}=0.209)

Standard Regression (X1 and X2 together)	= or ≠	Sequential Regression (X1 then X2)
r^2_{X1} = (a+c)/(a+b+c+d) = 0.157 *(revision)*	=	r^2_{X1} = (a+c)/(a+b+c+d) = 0.157
r^2_{X2} = (b+c)/(a+b+c+d) = 0.194 *(anxiety)*	=	r^2_{X2} = (b+c)/(a+b+c+d) = 0.194
sr^2_{X1} = a/(a+b+c+d) = 0.014 ("Part" in SPSS)	≠	sr^2_{X1} = (a+c)/(a+b+c+d) = 0.157 (r2 change)
sr^2_{X2} = b/(a+b+c+d) = 0.051	=	sr^2_{X2} = b/(a+b+c+d) = 0.051
pr^2_{X1} = a/(a+d) = 0.018 ("Partial" in SPSS)	≠	pr^2_{X1} = (a+c)/(a+c+d) = 0.1576 ("Partial" in SPSS)
pr^2_{X2} = b/(b+d) = 0.061	=	pr^2_{X2} = b/(b+d) = 0.061

*"Unique" contribution to variance from each IV – i.e. Anxiety contributes 5.1% of the variance (controlling for revision time) – OR Revision Time contributed 1.4% of the variance (controlling for Anxiety Level) – although this was not significant (p-val=0.184)***

*** NOTE that sr^2_{x1} is very small 0.014 – and if we run a sequential regression putting Anxiety in first, this r^2change (0.014) for adding in Revision Time has a non-sig p-val (0.184).*

"Stepwise"/"Statistical" Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	anxiety Exam Anxiety	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: exam Exam Performance (%)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.441[a]	.194	.186	23.39691	.194	24.384	1	101	.000

a. Predictors: (Constant), anxiety Exam Anxiety

b. Dependent Variable: exam Exam Performance (%)

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	revise Time Spent Revising	.169[a]	1.339	.184	.133	.497	2.012	.497

a. Predictors in the Model: (Constant), anxiety Exam Anxiety

b. Dependent Variable: exam Exam Performance (%)

So what does this mean?

- **You need to know which variables to put into the model and when – ideally based on theory.**
- **If you have no idea – try combinations – use "stepwise"/"statistical" regression to "see which ones fall out of the model"**

Diagnostics

- Look at your residuals (should look normal with a mean of 0 and even scatter) [observed vs predicted]
- Look at normal probability plots
- Look at plots of residuals versus the order in which the data were collected – should look random.
- Look at residuals versus variables left out of the model – should look random, if not, you may want to consider including the variable in the model.

Figure 5.19 – A. Field Book - Slides

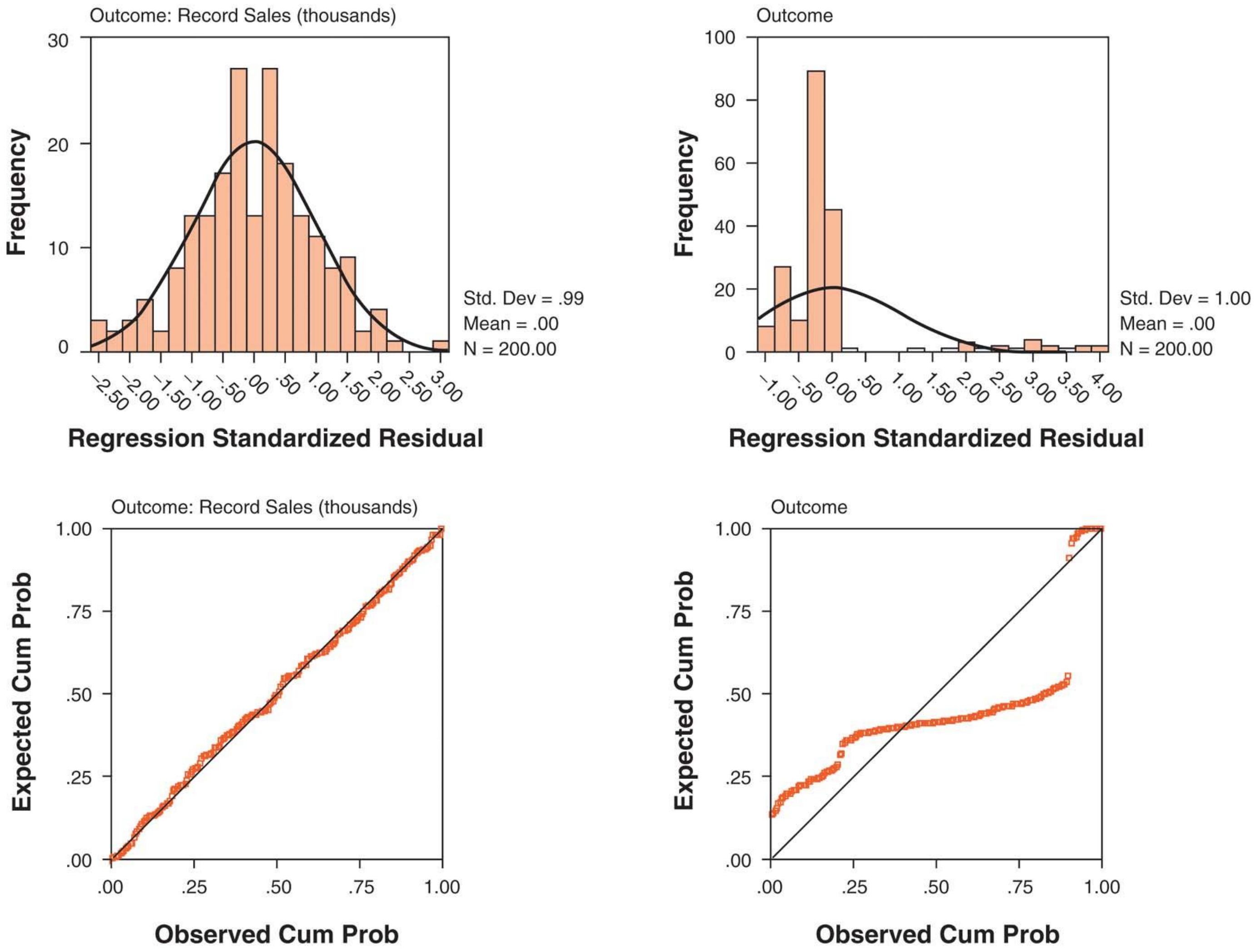
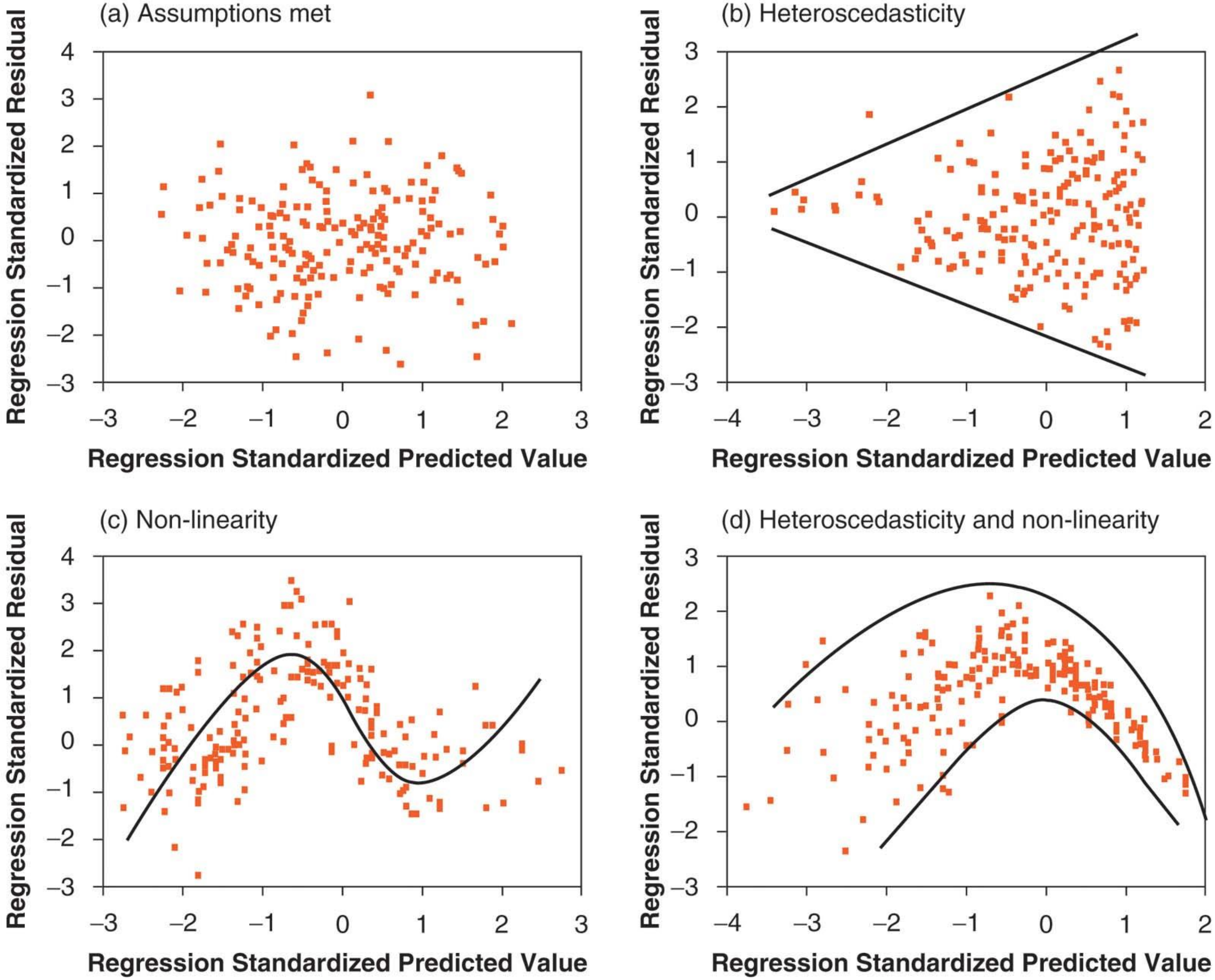


Figure 5.18 – A. Field Book - Slides



Multicollinearity – minor here

Coefficients[a]

Model		Collinearity Statistics Tolerance	VIF
1	(Constant)		
	revise Time Spent Revising	1.000	1.000
2	(Constant)		
	revise Time Spent Revising	.497	2.012
	anxiety Exam Anxiety	.497	2.012

a. Dependent Variable: exam Exam Performance (%)

Ideally want VIF close to 1 and Tolerance close to 1.

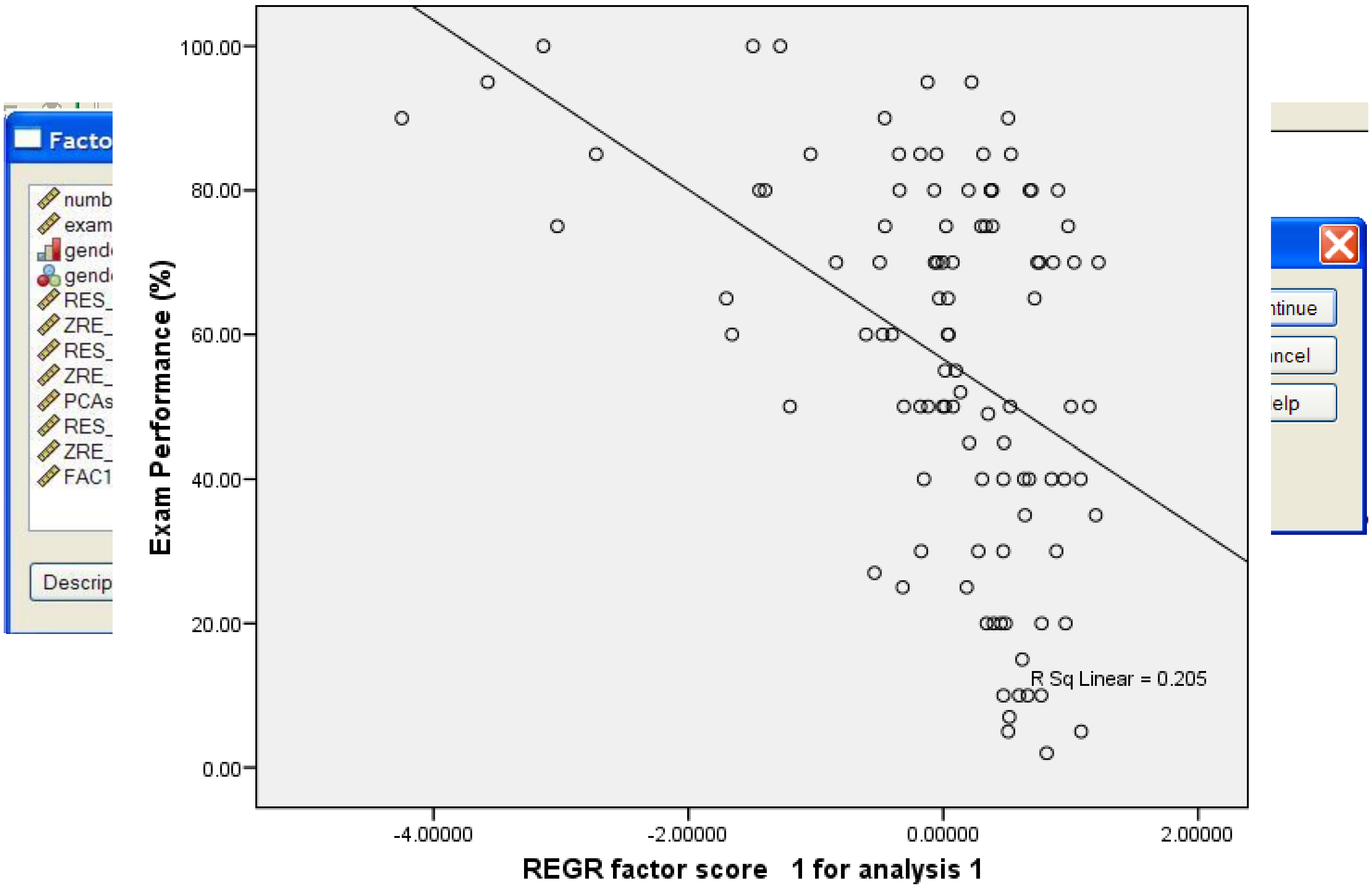
Tolerances < 0.2 (<0.4) and VIF's >10 (>2.5) or so are cause for concern.

Also review, Condition Index (>30 is a problem) and variance proportions (>0.5) for dimensions that correspond to higher condition indexes – if on same line.

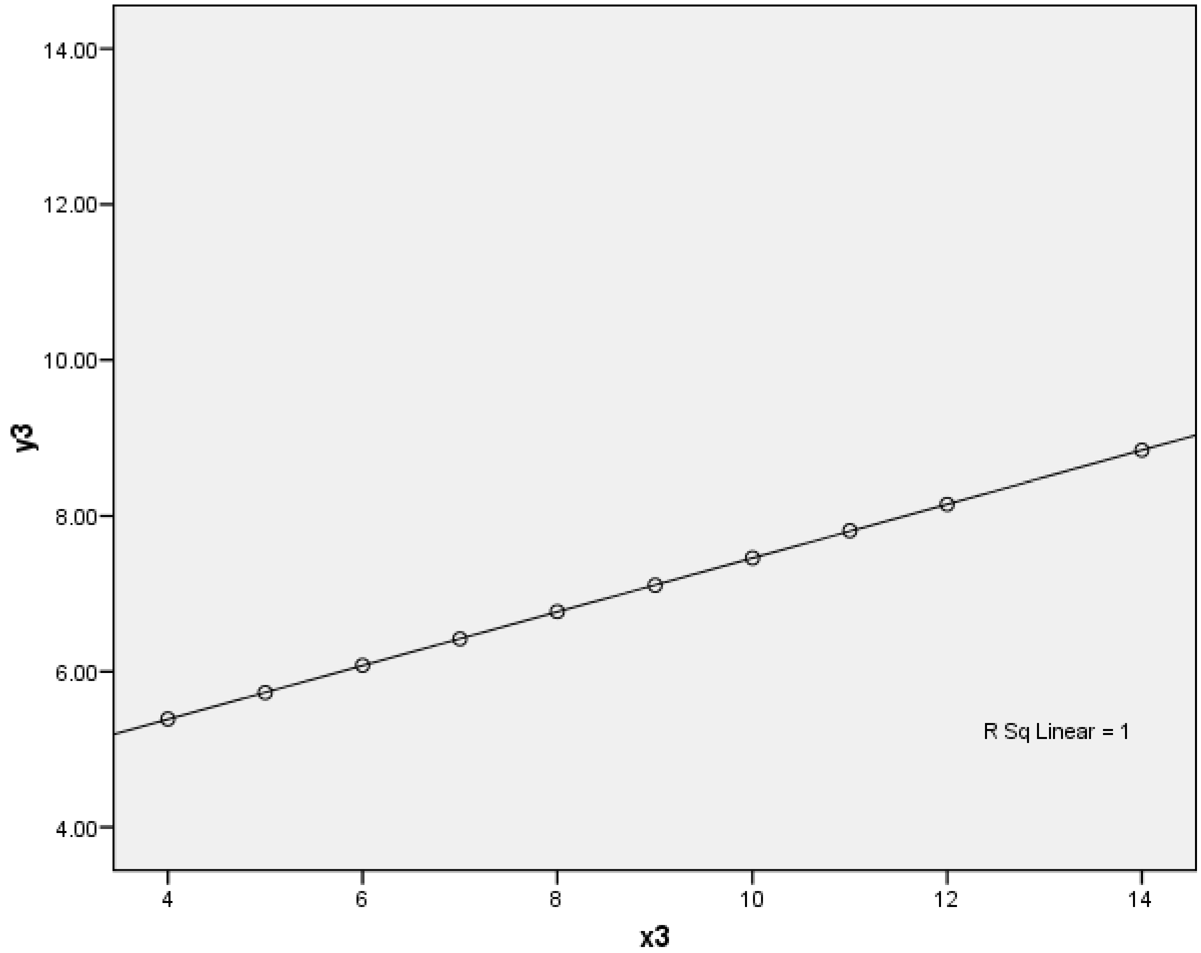
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	revise Time Spent Revising	anxiety Exam Anxiety
1	1	1.740	1.000	.13	.13	
	2	.260	2.584	.87	.87	
2	1	2.562	1.000	.00	.02	.00
	2	.428	2.447	.00	.37	.01
	3	.010	15.754	.99	.61	.98

a. Dependent Variable: exam Exam Performance (%)

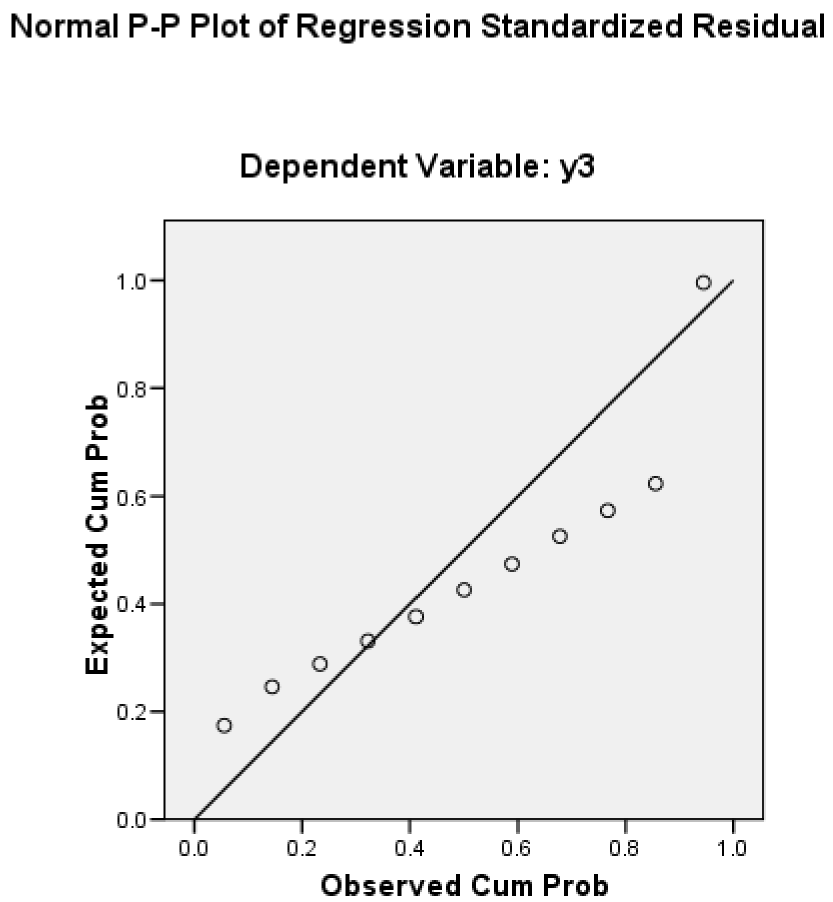
Could leave one out Or Could do PCA (principal components analysis)



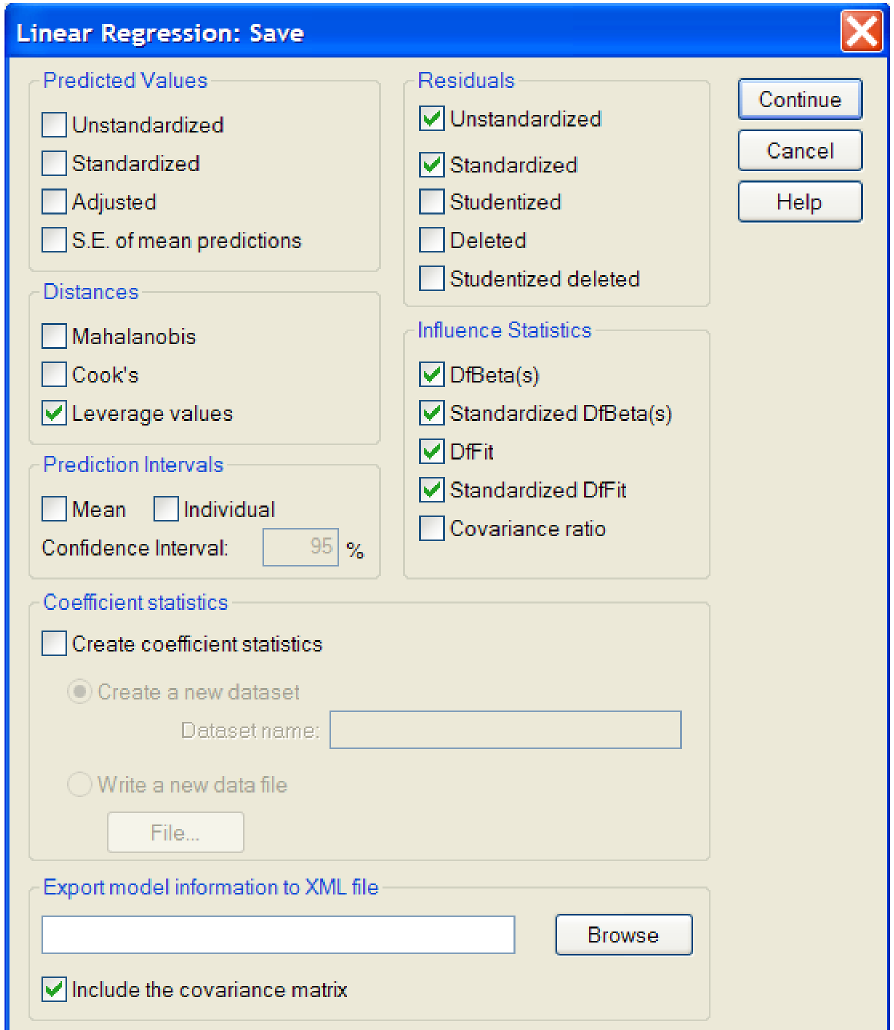
"Outliers?" - "Leverage"&"DFFit"&"DFBetas"



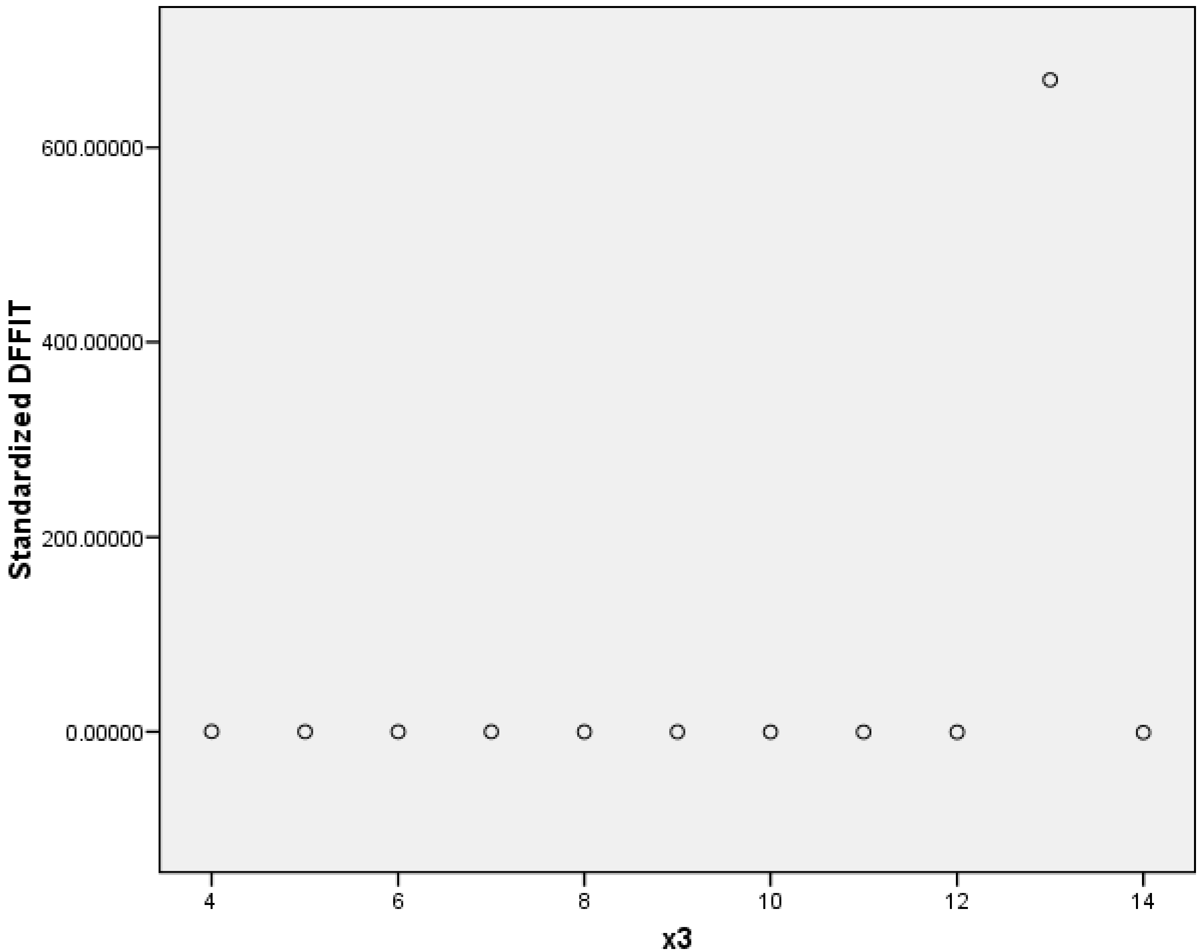
Residuals Plot – something's off



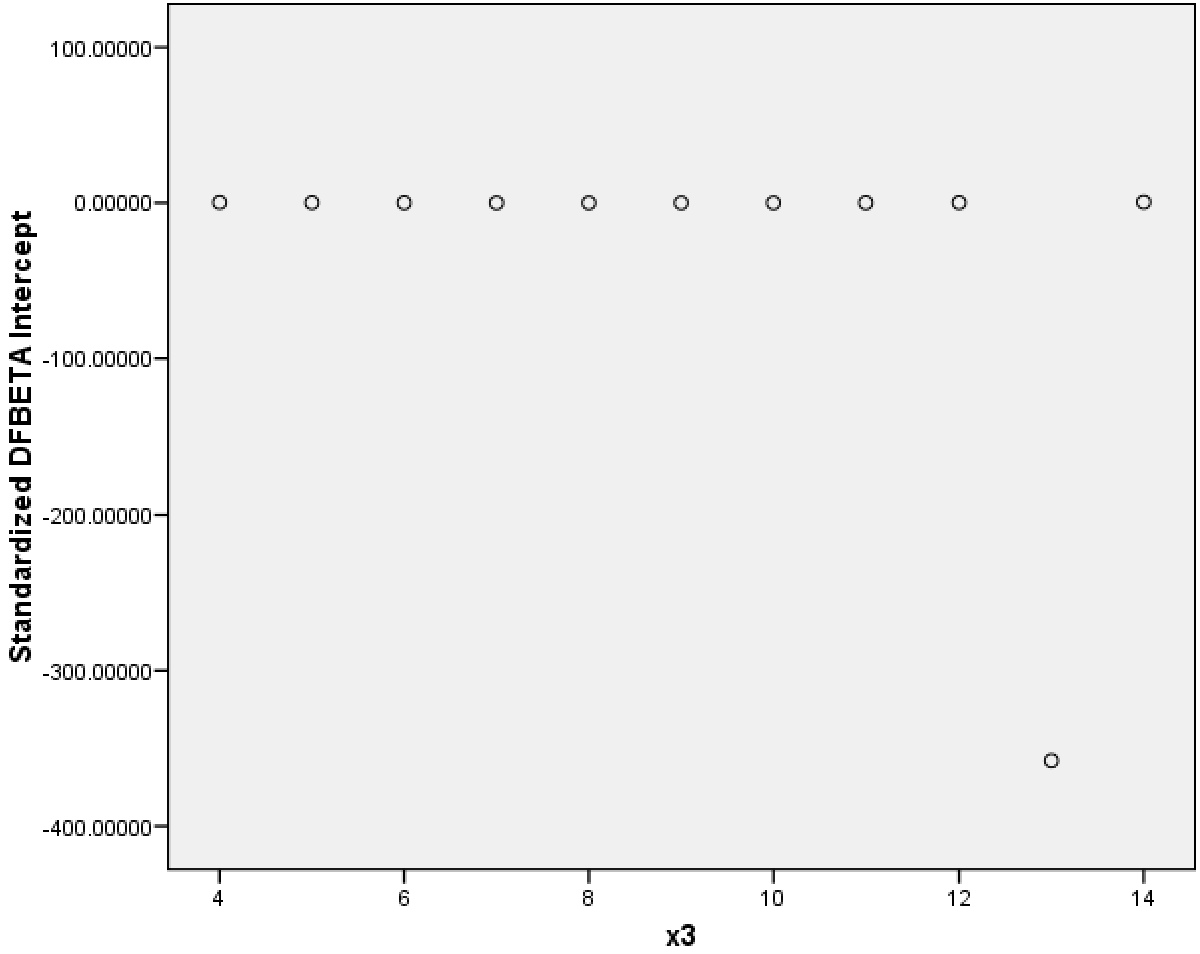
SPSS Linear Regression “Save” Menu



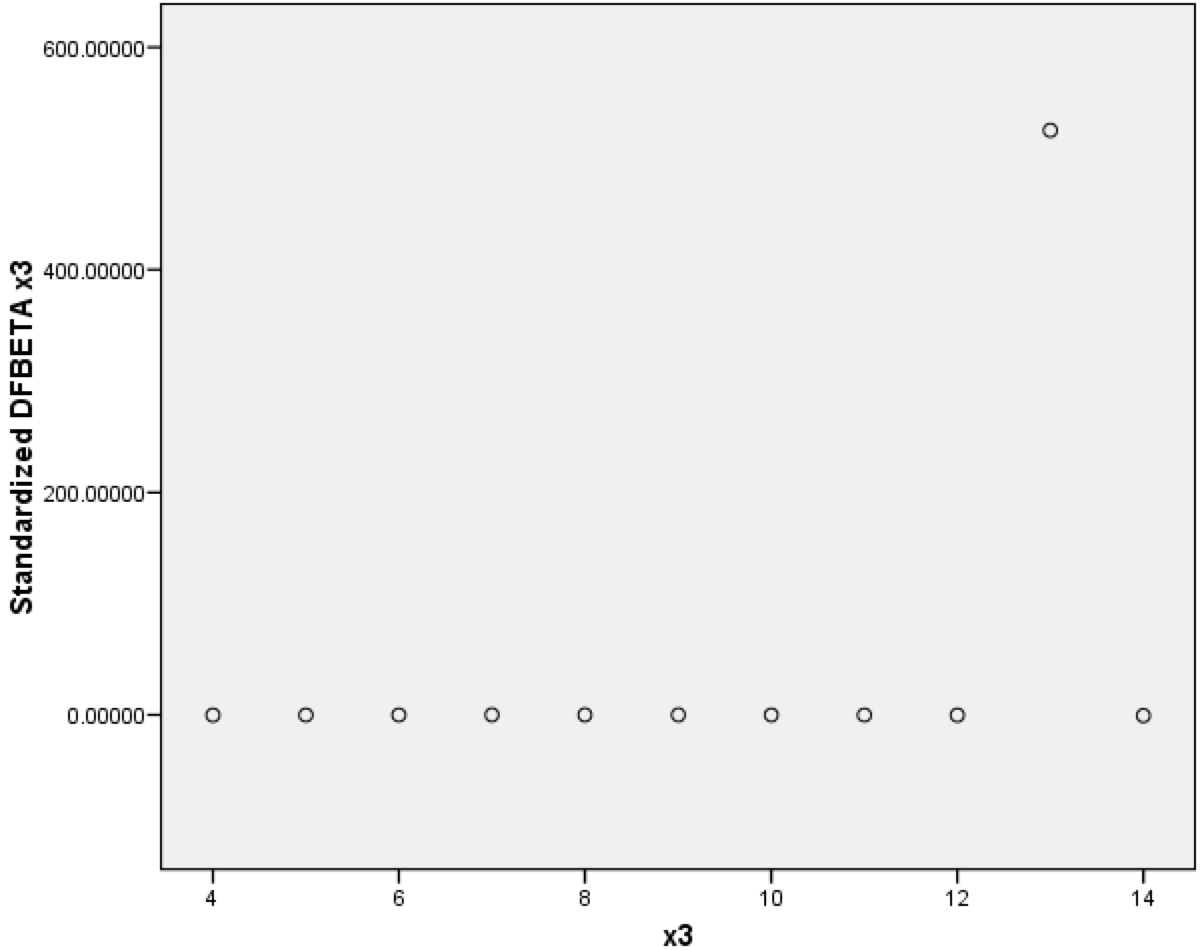
Plot x3 against Std DFfit



Plot x3 against Std DFBeta-intercept



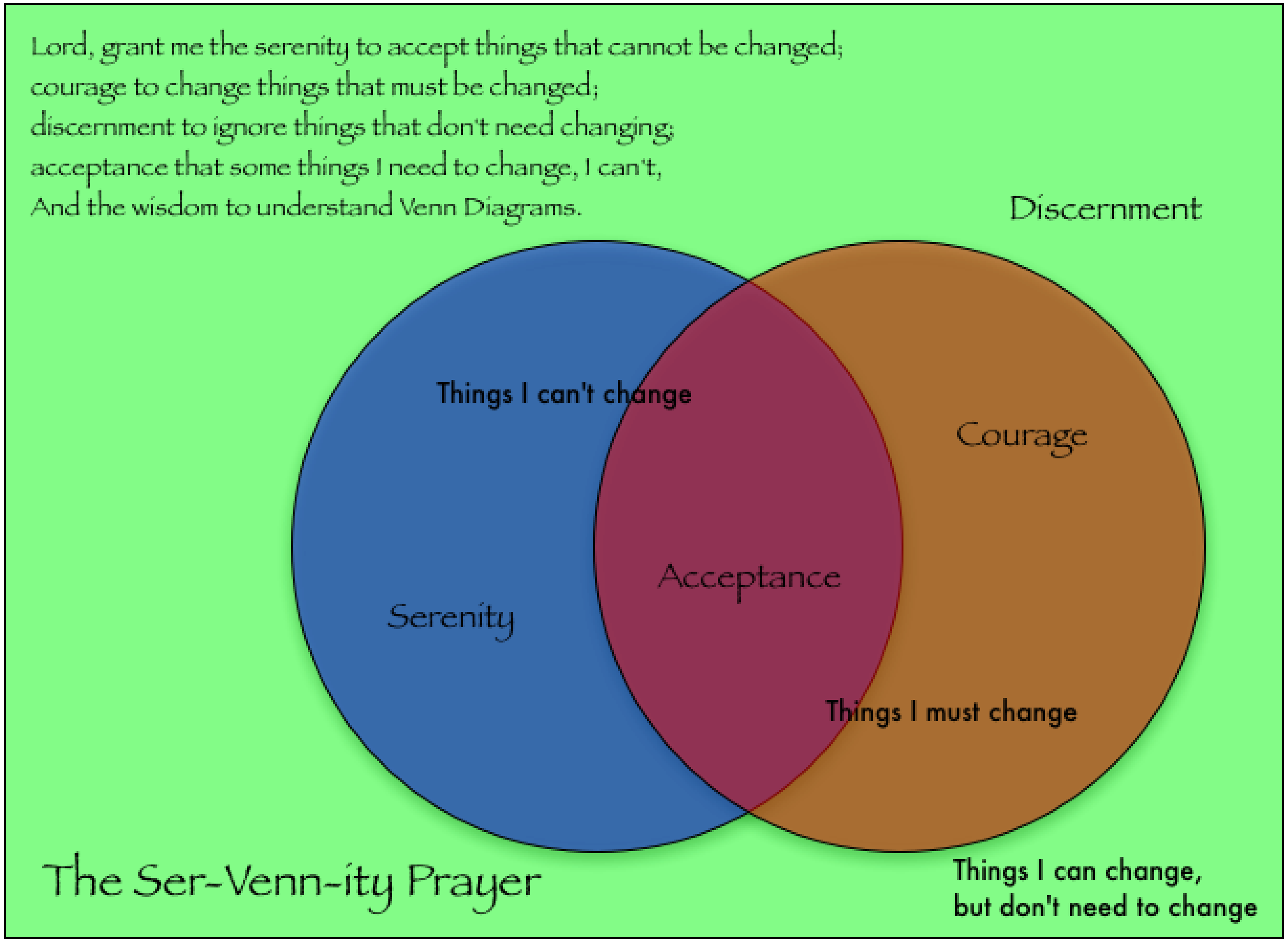
Plot x3 against DFBeta-x3



References

- Field, Andy. “Discovering Statistics Using SPSS” 2nd edition, SAGE Publications, 2005.
- Tabachnick, Barbara G.; Fidell, Linda S. “Using Multivariate Statistics” 5th edition, Pearson Eduation Inc., 2007.
- Cohen, Jacob; Cohen, Patricia; West, Stephen; Aiken, Leona “Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences” 3rd edition, Lawrence Erlbaum Associates Inc., 2003.

Parting Thoughts



VIII. Statistical Resources and Contact Info

SON S:\Shared\Statistics_MKHiggins\website2\index.htm

[updates in process]

- Working to include tip sheets (for SPSS, SAS, and other software), lectures (PPTs and handouts), datasets, other resources and references
- Statistics At Nursing Website: [website being updated] http://www.nursing.emory.edu/pulse/statistics/
- And Blackboard Site (in development) for "Organization: Statistics at School of Nursing"

Contact

Dr. Melinda Higgins

Melinda.higgins@emory.edu

Office: 404-727-5180 / Mobile: 404-434-1785