Admin

- Small Group Exercise.
  - Hand in at your tutorial
  - You will be asked to provide feedback on your peers' exercise and return these 'marked' exercises in the tute next week.
- Review Quiz grades will be posted on Blackboard.
- Assignment 1
  - Use student ID and password to get data
  - Look at the items that could comprise scales (in the questionnaire).
  - Use common sense and any practical experience to think about questions of interest about the relationship between a criterion variable and a set of predictor variables (up to five predictors).

General process for doing Assignment 1

- Focus on the research question
  - Initial checks on data (run frequencies - missing data)
  - Run regression analysis
  - Interpret regression
  - Determine if the results makes sense and are trustworthy (check assumptions, modify data if necessary)
  - Rerun analyses
  - Has substantive interpretation changed? (Play with the data)

NOTE: Don’t over-interpret data and assumption checks. They are important but need to be summarised briefly in write-up.

Major questions answered by multiple regression

Question 1: Is there an overall relationship between the two predictors and the criterion?

Question 2: Is there a relationship between each *individual* predictor and the criterion? What is the relative importance of each predictor?

Question 1: Is there an overall relationship between the two predictors and the criterion?

- **Question 1A**: Strength of relationship

  How strong is the relationship between $X_1 \ldots X_p$ and Y?

  Consider the correlation between Y' and Y (i.e., the multiple correlation between the predictors and the criterion).

  $R^2 =$ the squared multiple correlation between the predictors and criterion.

  $\frac{Var_{regression}}{Var_{total}} = R^2$ represents the proportion of variance of the criterion that can be predicted by knowing $X_1 \ldots X_p$

  $1 - R^2$ represents the *lack of fit* of the model to the data.

Question 1: Is there an overall relationship between the two predictors and the criterion?

- **Question 1B**: Statistical significance

  So we've described the *strength* of the overall relationship with R^2 but is the strength of this overall relationship significantly different than no overall relationship at all?

  The null hypothesis is: $H_0 : R^2 = 0$

  ...which indicates that there is no predictable variance. What we know about one variable tells us nothing about the others.

  We test the null hypothesis using an F statistic:

$$F = \frac{MS_{regression}}{MS_{residual}} \ (df = p, N - p + 1)$$

$N =$ Sample size

$p =$ Number of potential independent variables: $X_1, X_2, \ldots X_p$

Question 2: Is there a relationship between each *individual* predictor and the criterion? What is the relative importance of each predictor?

1. Simple correlations
2. Standardised regression weights
   - significance testing
   - confidence intervals
3. Semipartial correlations
4. Partial correlations
5. Relative weights

All measures of the importance of individual predictors

	Simple Correlation	Standardised Regression Weight	Semipartial Correlation	Usefulness	Partial Correlation	Relative Weights
	r	β	sr	sr^2	pr	RW
MOTIV (X_1)	0.59	0.35	0.33	10.75%	0.50	30.93%
GRADE(X_2)	0.75	0.62	0.57	32.71%	0.71	69.07%

Standard multiple regression

- The purpose involves mainly explaining the nature of the relationship between the predictors and the criterion.
- All predictors enter into the regression equation at once.
- Each predictor is assessed as if it had entered the regression equation after all the other predictors had entered.

The research question

- In the T&F Women's Health study, a question of interest is whether the number of times a woman visits the doctor is related to their physical health, mental health and stress levels.

$$timedrs \leftarrow phyheal \;\; menheal \;\; stress$$

- See Section 5.7 of Chapter 5, page 161: Complete Examples of Regression Analysis, and Section B.1 of Appendix B: Women's Health and Drug Study.

Aims of interpreting the output

- Find the answer to each question addressed by the analysis (strength; relative importance).
  - Task: Convert the raw output into a format appropriate for a results section.
- Determine if the analysis makes sense and is trustworthy. Are the interpretations sound?
  - Task: Check that assumptions/limitations/other issues have been addressed

Simple checks:
Means and standard deviations

Descriptive Statistics

	N	Mean	Std. Deviation
LTIMEDRS	465	.7413	.41525
MENHEAL	465	6.12	4.194
LPHYHEAL	465	.6484	.20620
SSTRESS	465	13.3995	4.97217

Simple checks:
Correlation Matrix

Correlations

	LTIMEDRS	MENHEAL	LPHYHEAL	SSTRESS
LTIMEDRS	1	.355	.586	.359
MENHEAL	.355	1	.511	.383
LPHYHEAL	.586	.511	1	.317
SSTRESS	.359	.383	.317	1

Strength of the overall relationship: R and R^2

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SSTRESS, LPHYHEAL, MENHEAL[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: LTIMEDRS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.614[a]	.377	.373	.32888

a. Predictors: (Constant), SSTRESS, LPHYHEAL, MENHEAL

Statistical significance of the overall relationship:

ANOVA

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	30.146	3	10.049	92.901	.000[a]
	Residual	49.864	461	.108		
	Total	80.010	464			

a. Predictors: (Constant), SSTRESS, LPHYHEAL, MENHEAL

b. Dependent Variable: LTIMEDRS

4.889258778314550E-47[a]

but

$p < .0001$

will suffice

This is a test of $H_0 : R^2_{pop} = 0$

Calculated $R^2 = .377$

Importance of predictors:

B weights, t-values and statistical significance

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	-.155	.058	-2.661	.008
	MENHEAL	1.884E-03	.004	.428	.669
	LPHYHEAL	1.040	.087	11.928	.000
	SSTRESS	1.571E-02	.003	4.671	.000

a. Dependent Variable: LTIMEDRS

Again - check to see what these values are, and $p < .0001$ will suffice

Importance of predictors:

β, pr and sr

Coefficients[a]

Model		Standardized Coefficients		Correlations		
		Beta	Std. Error	Zero-order	Partial	Part
1	MENHEAL	.019	.044	.355	.020	.016
	LPHYHEAL	.516	.043	.586	.486	.439
	SSTRESS	.188	.040	.359	.213	.172

a. Dependent Variable: LTIMEDRS

Square these to get sr^2

Calculating 95% Confidence Intervals for β weights

- provides a range of values which are likely to include the “true” population value of a regression weight.
- a wider confidence interval indicates a less trustworthy estimate of the weight.
- a confidence interval which spans zero indicates a non-significant weight.

Calculating 95% Confidence Intervals for β weights

A Confidence Interval =

Sample Value $\pm$ Critical Value $\times$ Sample Standard Error

A $(1-\alpha)\%$ Confidence Interval for $\beta_{MENHEAL}$ =

$$\beta_{MENHEAL} \pm t_{N-p-1,\alpha/2} \times SE_{MENHEAL}$$

$$.019 \pm 1.96 \times .044$$

$$= -0.068 \leftrightarrow 0.106$$

Coefficients[a]

Model		Standardized Coefficients		Correlations		
		Beta	Std. Error	Zero-order	Partial	Part
1	MENHEAL	.019	.044	.355	.020	.016
	LPHYHEAL	.516	.043	.586	.486	.439
	SSTRESS	.188	.040	.359	.213	.172

a. Dependent Variable: LTIMEDRS

Calculating 95% Confidence Intervals for β weights

Variable	Beta	Lower	Upper
MENHEAL	0.019	-0.068	0.106
LPHYHEAL	0.516	0.432	0.601
SSTRESS	0.188	0.109	0.267

Coefficients[a]

Model		Standardized Coefficients		Correlations		
		Beta	Std. Error	Zero-order	Partial	Part
1	MENHEAL	.019	.044	.355	.020	.016
	LPHYHEAL	.516	.043	.586	.486	.439
	SSTRESS	.188	.040	.359	.213	.172

a. Dependent Variable: LTIMEDRS

Calculating 95% Confidence Intervals for β weights

Variable	Beta	Lower	Upper
MENHEAL	0.019	-0.068	0.106
LPHYHEAL	0.516	0.432	0.601
SSTRESS	0.188	0.109	0.267

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	-.155	.058	-2.661	.008
	MENHEAL	1.884E-03	.004	.428	.669
	LPHYHEAL	1.040	.087	11.928	.000
	SSTRESS	1.571E-02	.003	4.671	.000

a. Dependent Variable: LTIMEDRS

	Part
.020	.016
.486	.439
.213	.172

Relative Weights and ‘usefulness’ or uni

Coefficients[a]

Model		Standardized Coefficients		Correlations		
		Beta	Std. Error	Zero-order	Partial	Part
1	MENHEAL	.019	.044	.355	.020	.016
	LPHYHEAL	.516	.043	.586	.486	.439
	SSTRESS	.188	.040	.359	.213	.172

a. Dependent Variable: LTIMEDRS

Variable		
MENHEAL	$\frac{.019 \times .355}{.377} = 1.79\%$	$.016^2 = .02\%$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.614[a]	.377	.373	.32888

a. Predictors: (Constant), SSTRESS, LPHYHEAL, MENHEAL

$$RW_i = \frac{\beta_i r_{yi}}{R^2}$$

Relative Weights and 'usefulness' or unique contribution

Variable	RW%	$\mathbf{sr^2}\%$
MENHEAL	$\frac{.019 \times .355}{.377} = 1.79\%$	$.016^2 = .02\%$
LPHYHEAL	$\frac{.516 \times .586}{.377} = 80.28\%$	$.439^2 = 19.24\%$
SSTRESS	$\frac{.188 \times .359}{.377} = 17.93\%$	$.172^2 = 2.95\%$

$$\sum RW\% = 100\% \qquad \sum sr^2\% = 22.21\%$$

$$RW_i = \frac{\beta_i r_{yi}}{R^2}$$

$$R^2 - \sum sr^2 = 15.49\%$$

Relative Weights and 'usefulness' or unique contribution

Variable	RW%	$\mathbf{sr^2\%}$
MENHEAL	$\frac{.019 \times .355}{\quad} = 1.79\%$	$.016^2 = .02\%$
	%	$.439^2 = 19.24\%$
		$.172^2 = 2.95\%$

$$\sum sr^2\% = 22.21\%$$

$$R^2 - \sum sr^2 = 15.49\%$$

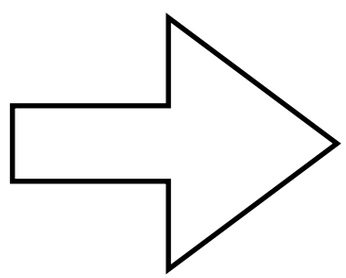
For standard regression, T&F suggest that the total variance accounted for can be separated into *unique* and *shared* variability. *Unique* variance is the sum of the sr^2 values, and *shared* variance is (R^2 minus the sum of the sr^2 values). Algebraically, there is no reason why the former should ever be less than the latter. So it doesn't tend to be a useful concept.

$$RW_i = \frac{\beta_i r_{yi}}{R^2}$$

A word on computing RW, Confidence Intervals, etc. by hand

Be as precise as you can in your calculations, or your result may not be accurate (e.g., RW may not sum to 100%). As a general rule: if you're planning on reporting your results to three decimal places, include values to four decimal places in your computations.

Standardized Beta
.019
.516
.188



Standardized Beta
.01902178
.51640667
.18810541

The prediction equation

- SPSS gives B weights and the intercept (Constant).
- B weights indicate the amount the criterion variable would change if the predictor changed by one (and all the other predictors were held constant).
- The intercept indicates the score on the criterion variable if all predictor variables were zero.

- Using B weights and the intercept, we can calculate a predicted score on the criterion for any set of scores on the predictor.

$$Y = -0.155 + 0.002(X_1) + 1.040(X_2) + 0.016(X_3)$$

Model		Unstandardized B
1	(Constant)	-.155
	MENHEAL	1.884E-03
	LPHYHEAL	1.040
	SSTRESS	1.571E-02

Presentation of results

- The SPSS output needs to be reformatted and presented either in tables or in the text.
- Statistics should be reported for all variables, not just the ‘statistically significant’ ones.
- The statement of results is the ‘story’ the results are implying. These are linked to the meaning of the variables to provide an interpretation of the results. Of course, how far the researcher goes with this interpretation (and how far the reader trusts the interpretation) depends on how robust it is to violations of the assumptions of multiple regression.

Regression diagnostics

- General strategy
- Data checking
- Importance of residuals
- Independence, linearity, normality
- Outliers and influential data points
- Homoscedasticity, multicollinearity, and singularity

Overview

- statistical analyses make certain assumptions about the nature of the data being analysed
  - we need to check that the data meet these assumptions
- the majority of analyses are based on correlation
  - the main aim of data checking is to identify adverse influences on correlations

Purpose of Diagnostics

- To engender trust in the presented results
  - To reassure the reader about the trustworthiness of the results
- To detect fishiness
- To check how robust the results are.
  - to check that the results aren't due to a few spurious observations
- To better estimate the "true" nature of relationships in the data

Getting to know your data

You have to have a basic understanding of your data and the relationship between the variables. This can help a lot when it comes to interpreting your results.

You also have to ensure that your data meet all the requirements for a multivariate analysis. Missing data, outliers, etc. are difficult to assess when you're dealing with several variables.

General strategy for handling violation of assumptions

- Violations of assumptions due to problem data may or may not influence the results of the multiple regression.
- Run the analysis with and without correcting the problem data, and compare the results and interpretations.
- Correcting for problem data may or may not change the interpretation.

If the interpretation *doesn't* change, then

- report the results from the original data and report that even with the problem data corrected, the interpretation is the same.

General strategy for handling violation of assumptions

If the interpretation *does* change, then

- for variables now *included* in the interpretation:
  - be wary, perhaps the finding is not robust.
  - report that the variable was not significant (not important) with original data.
  - consider how important the variable is theoretically.

- for variables now *dropped* from the interpretation:
  - be wary, perhaps the finding is not robust.
  - report that the variable was significant (important) with original data.
  - consider how important the variable is theoretically.

Checking Data

It is essential to know:

- The type and nature of respondent, subject or experimental unit. That is, the unit of analysis.
- The procedure for data collecting.
- The unit of measurement for each variable.
- A plausible range of values and a typical value for each variable.
- Expected relationships between the focal variables of your study – so you know what to expect when you do your analyses.
- *Hint*: Keep a codebook and document the design decisions, data collection and data analysis strategies.

Checking Data

Types of problem data:

- Recording or data entry errors
- Missing data
- Outliers
- Non-normal distributions
- Non-linear relationships

Summary statistics for problem data:

- need a measure of whether problem data is really a problem.
- a statistic divided by its standard error is approximately a Z-score.
- examine the Z-scores to see whether they “fit” with the rest of the data.
- Z-scores < -3 or > +3 are beyond what would be expected.

Univariate checks (single variable)

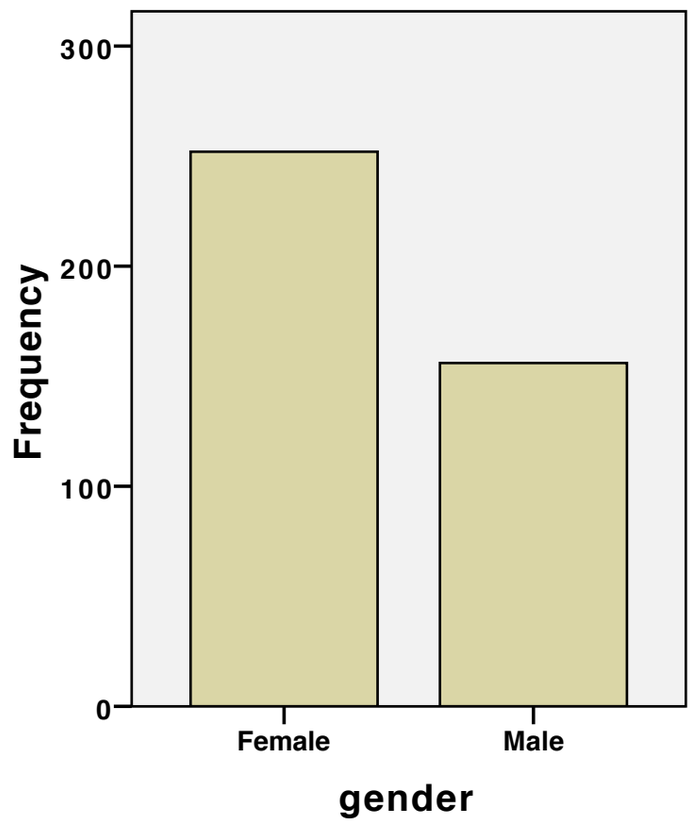
- This is done by calculating various descriptive statistics via SPSS DESCRIPTIVES or EXPLORE procedure for each variable. The checks are different for categorical and continuous variables.

Categorical variables:

- Check the proportion of cases in each category.
- T&F suggest splits in the frequencies for dichotomous variables larger than 90% lead to difficulties when correlated with other variables. A dichotomous variable that is symmetrically distributed has a 50/50 split.

gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	252	61.8	61.8	61.8
	Male	156	38.2	38.2	100.0
	Total	408	100.0	100.0	



Univariate checks (single variable)

The descriptive statistics for continuous variables are:

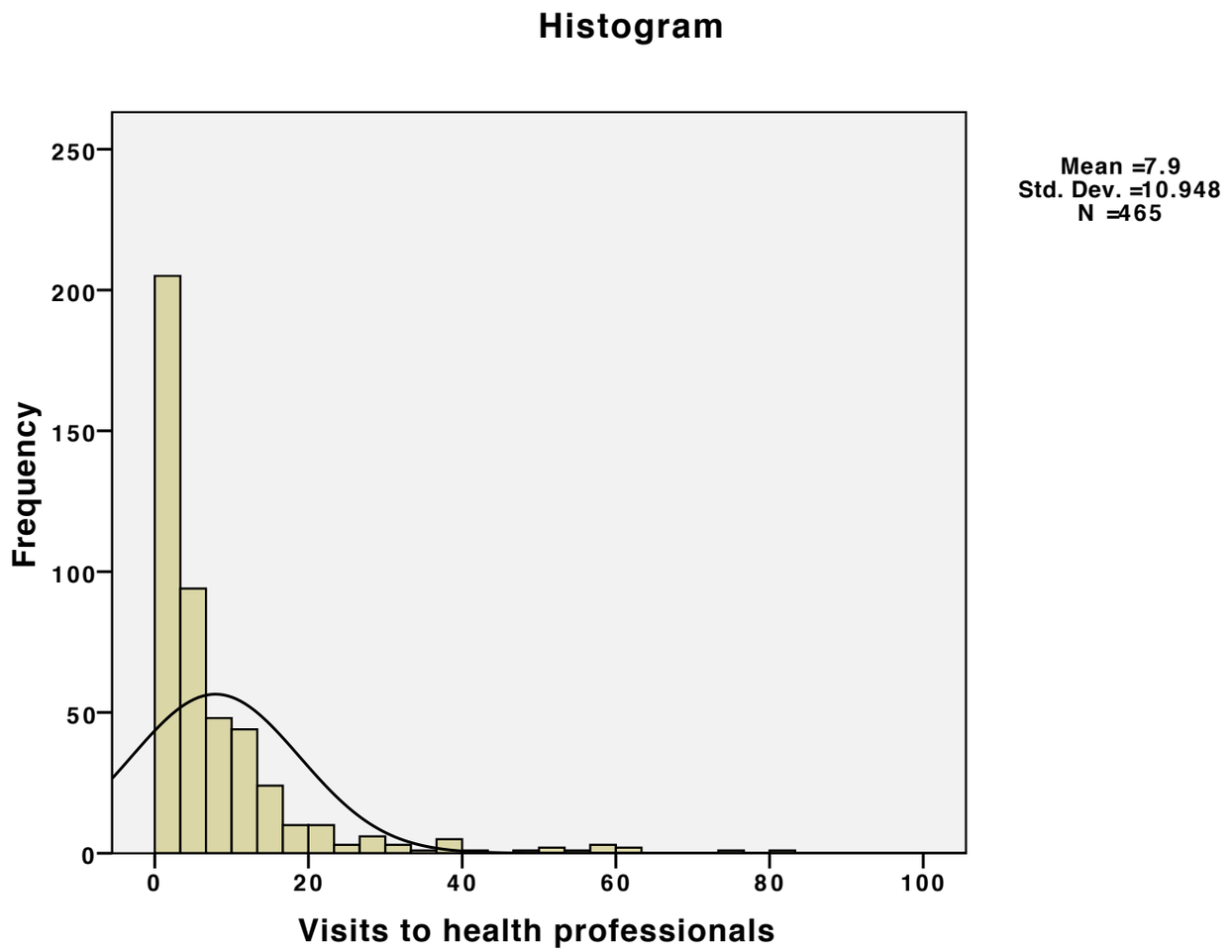
- Measures of central tendency; mean, median, mode.
- Measures of the shape of the distribution;
- its spread, range, minimum, maximum, variance, standard deviation.
- its symmetry, skewness. This is an important characteristic and strongly influences correlations among variables.
- its ‘peakiness’, kurtosis.
- Ways of compactly displaying some of these characteristics are to use a box plot, stem and leaf plot and/or a histogram.
- Checking for bi-modality is detected by visual inspection. If it occurs it may indicate that the data from two distinct groups have been mixed. However, there are no tests for its presence.

Univariate checks (single variable)

The descriptive statistics for continuous variables are:

timedrs

N	Valid	465.000
	Missing	.000
	Mean	7.901
	Median	4.000
	Mode	2.000
	Std. Deviation	10.948
	Skewness	3.248
	Std. Error of Skewness	.113
	Kurtosis	13.101
	Std. Error of Kurtosis	.226
	Minimum	.000
	Maximum	81.000



Univariate checks (single variable)

- Checking for normality
  - A major assumption for statistical inference in multivariate statistics is that the data are normally distributed. (Normal distributions are symmetrical). Symmetry is perhaps the most important aspect of the shape of the distribution.
  - non-normality can be “fixed” by transforming the variable.

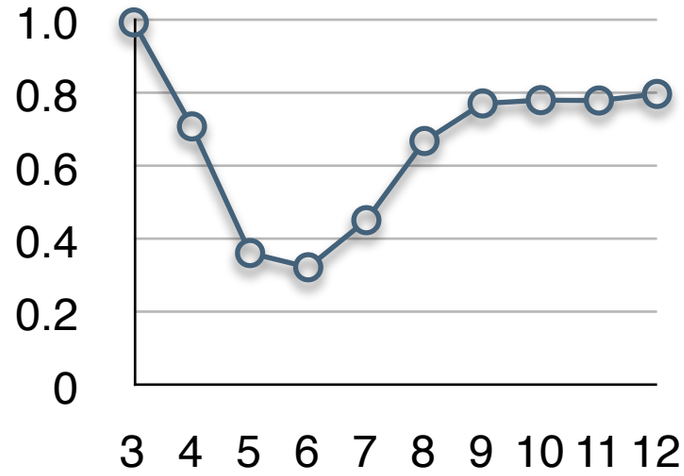
- Checking for univariate outliers
  - Check the frequency distribution for implausible, extremely small or large values. A standardised (z) score of > ± 3.0 may be used as a cutoff to define extreme values, (remember it is only a ‘rule-of-thumb’).
  - Remedies: This depends on whether the ‘outlying’ cases are part of the process that you intend to sample. BEWARE! Be as conservative as possible! Do not delete outliers willy nilly. You must always justify your decision in your results section. If some alternative systematic process was operating (e.g., your subject fell asleep) then you can probably justify their elimination.

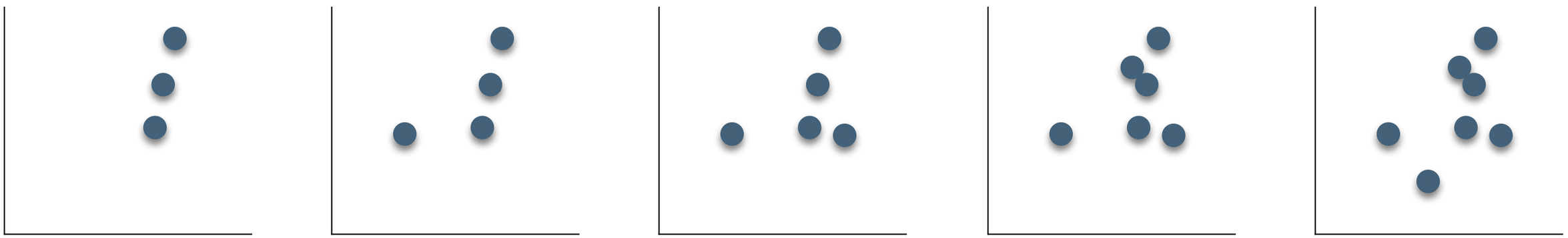
Bivariate and multivariate checks (2 or more variables)

- Checking for bivariate normality
  - univariate normality, scatterplots
  - If each of the variables are normally distributed then, although this does not guarantee bivariate normality, the chances of bivariate normality are increased.
- Check for linearity of relationships
  - Scatterplots of the two variables should be linear. This can be checked by visual inspection.
- Check for bivariate outliers
  - several measures covered later

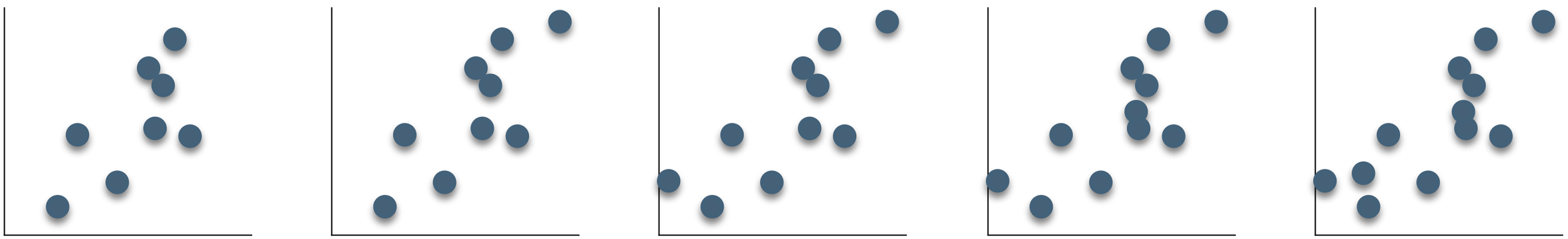
Factors influencing the correlation coefficient

Stability and sample size





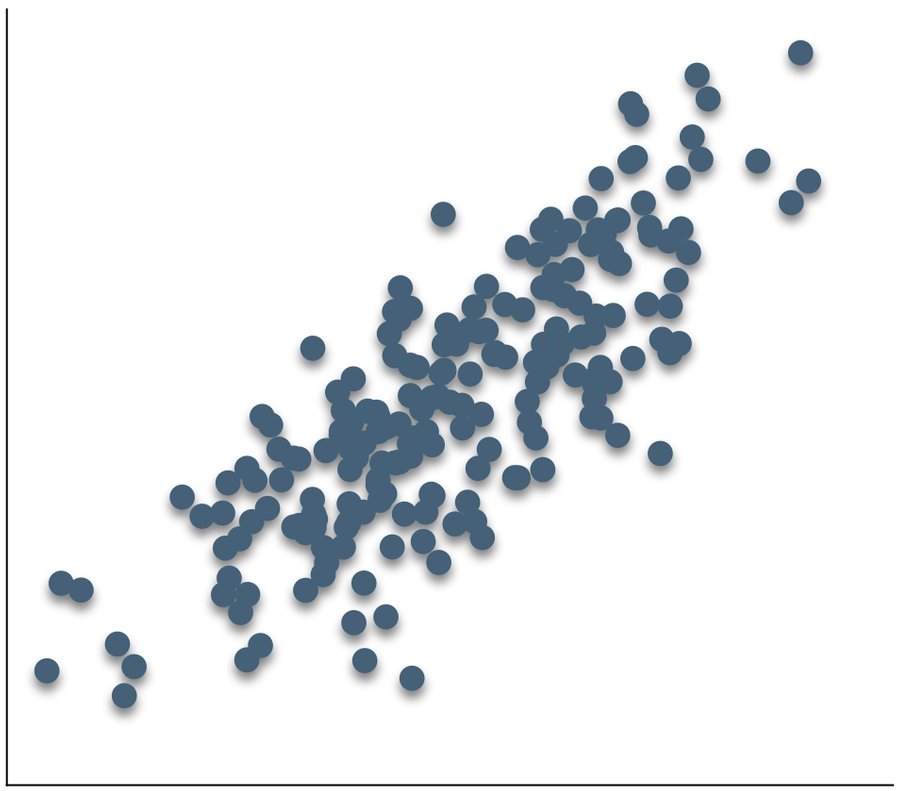
r = 0.996 | r = 0.711 | r = 0.364 | r = 0.325 | r = 0.454



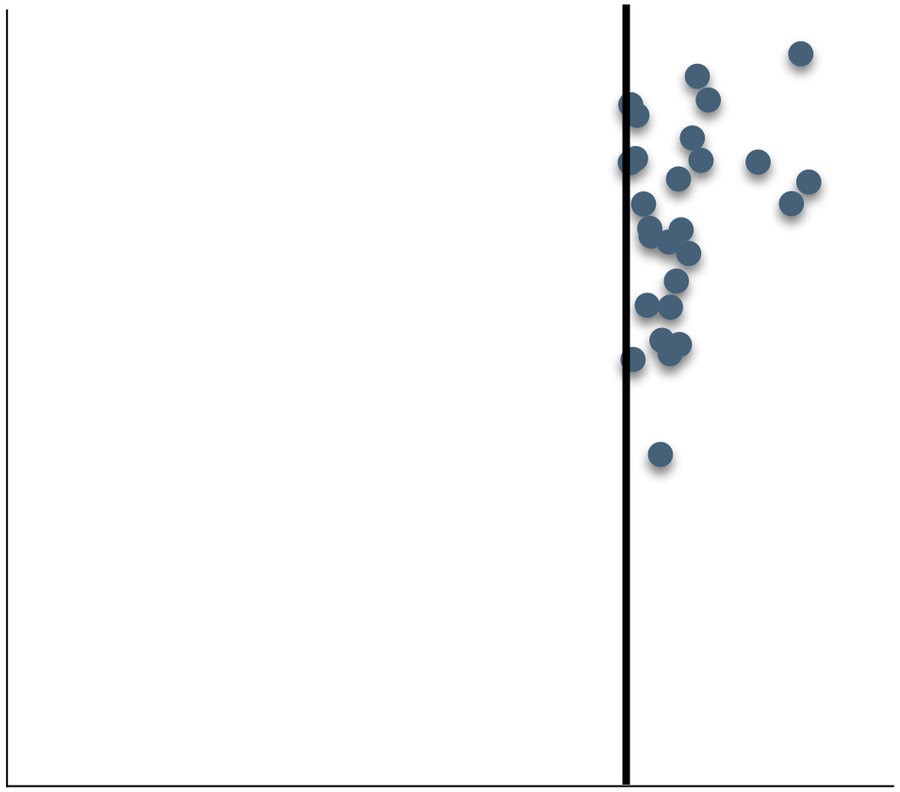
r = 0.671 | r = 0.774 | r = 0.783 | r = 0.782 | r = 0.800

Factors influencing the correlation coefficient

Restriction of range



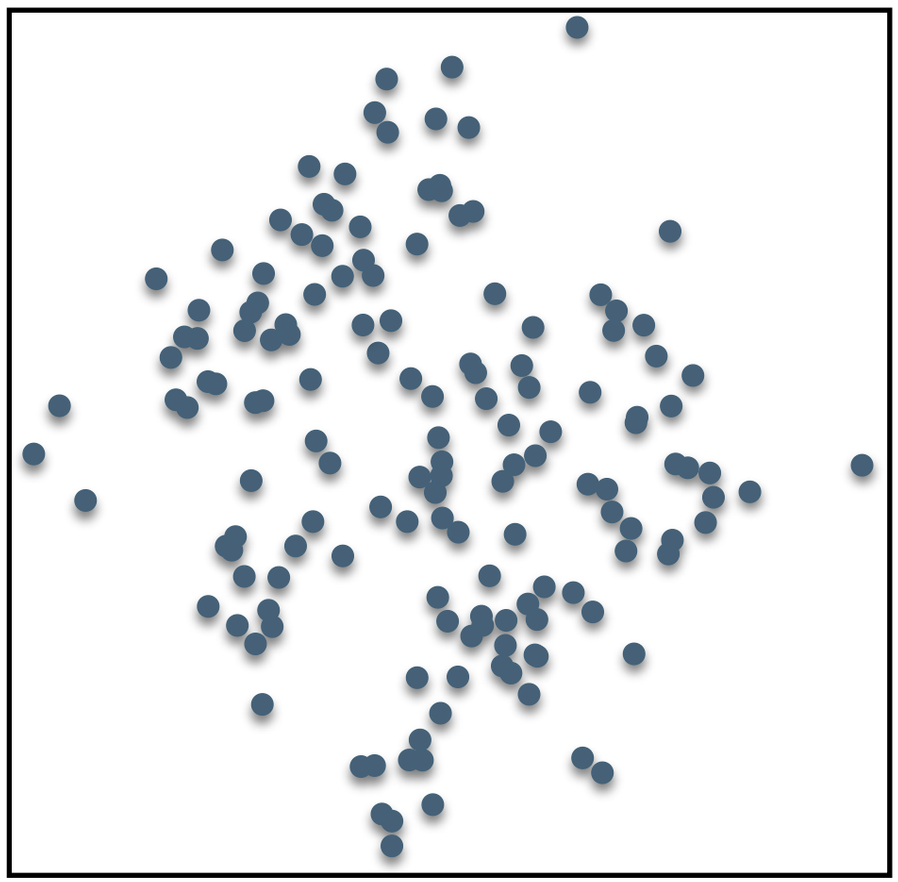
r = 0.800

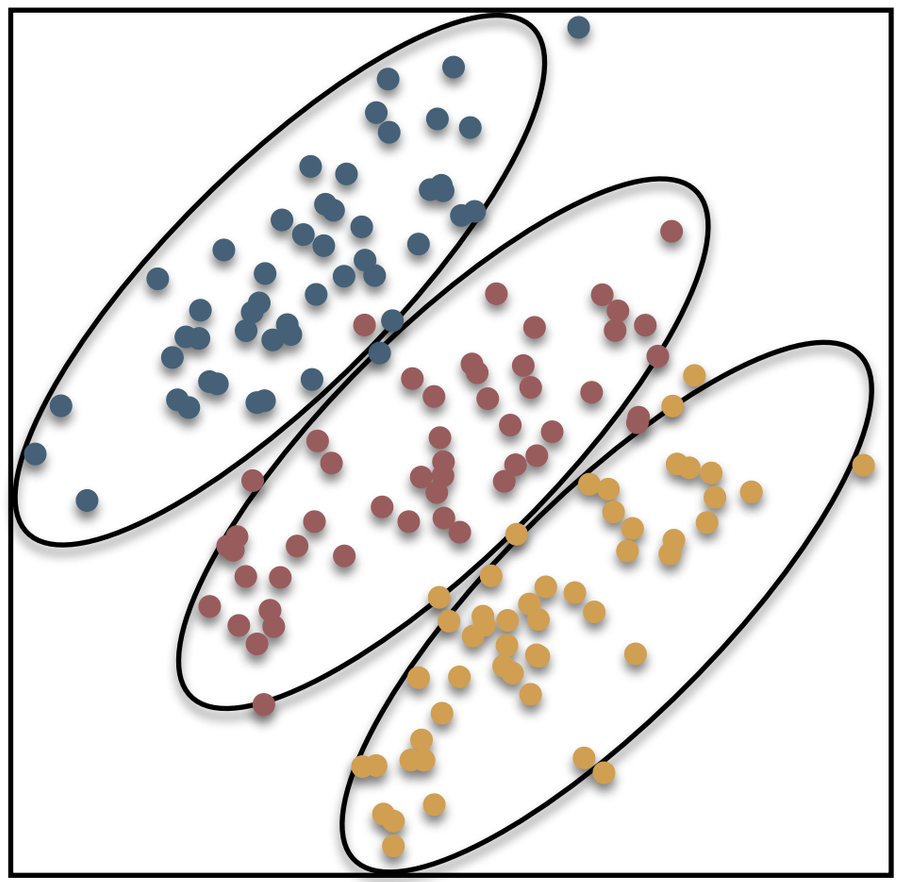


r = 0.323

Factors influencing the correlation coefficient

Combining several different populations

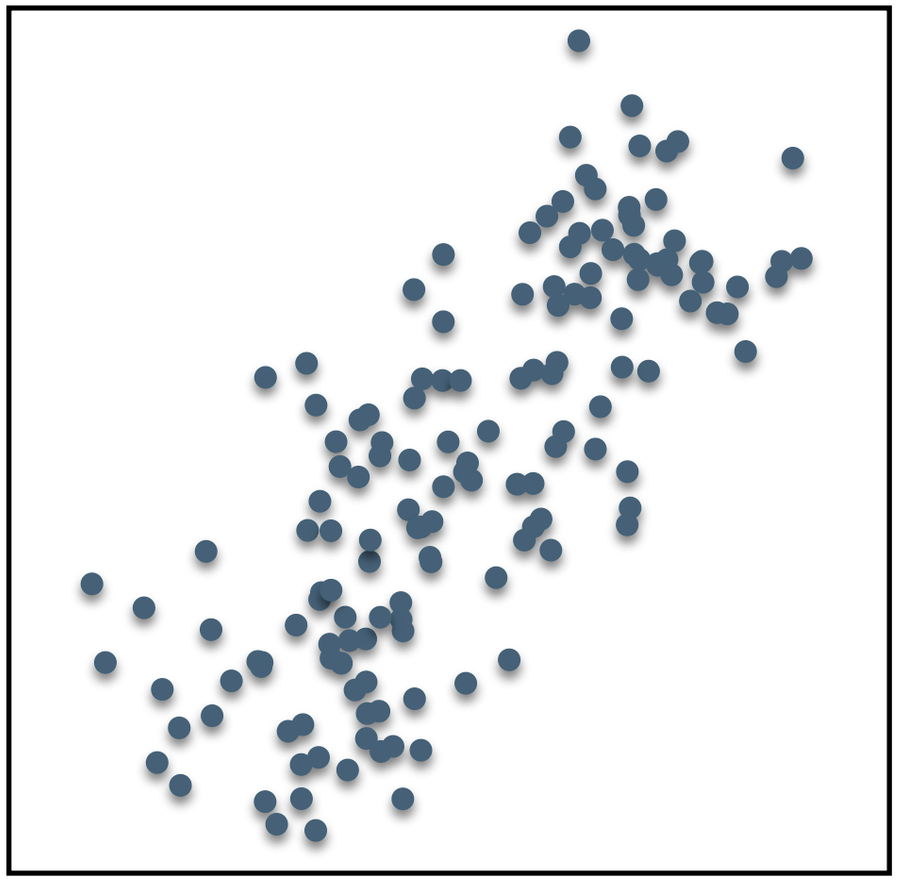


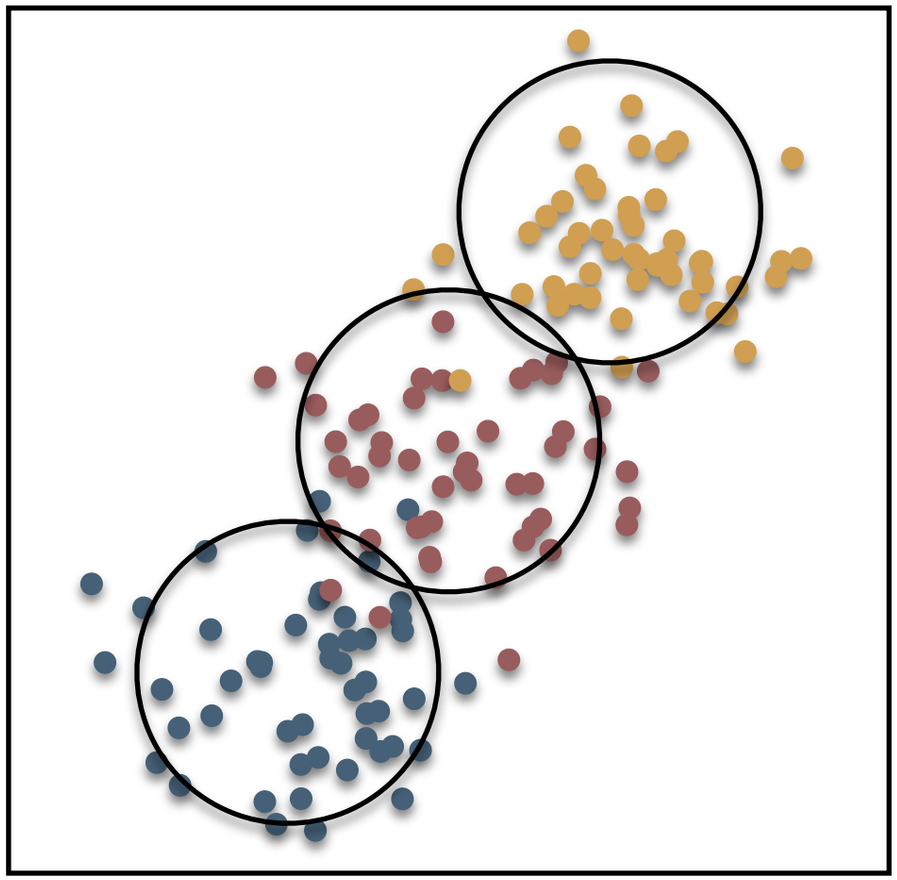


In each sample the correlation coefficient is relatively high and positive, but the correlation coefficient for the combined samples is close to zero.

Factors influencing the correlation coefficient

Combining several different populations

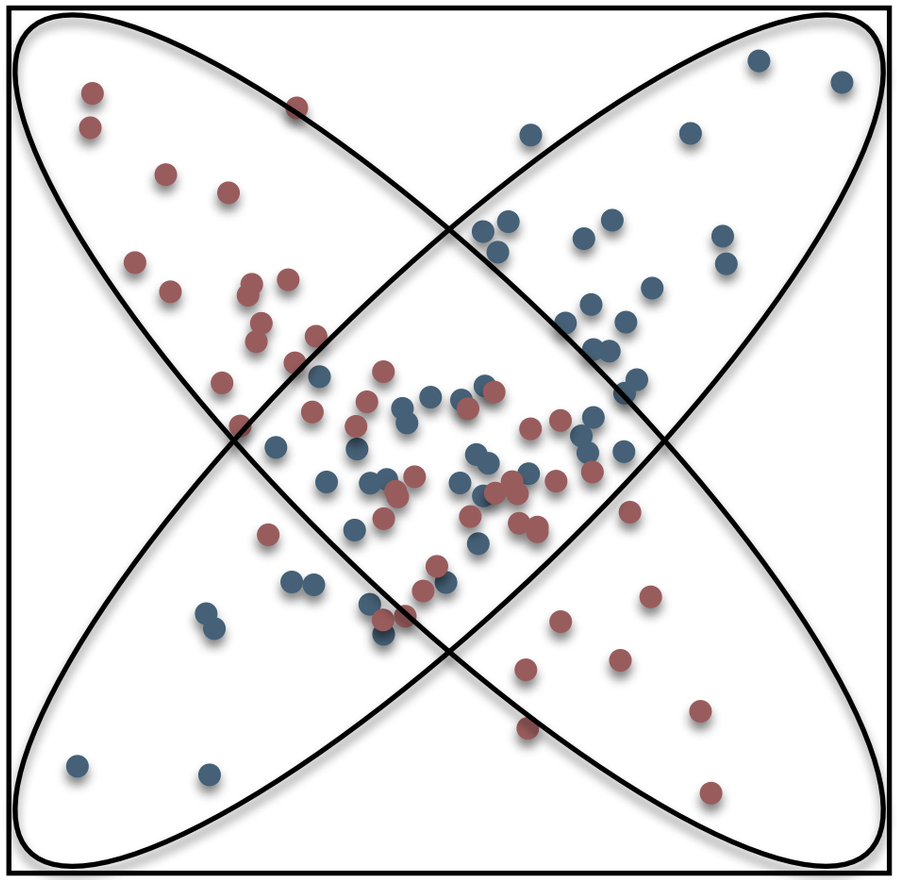




In each sample the correlation coefficient is close to zero. The correlation for the combined samples is moderately high and positive.

Factors influencing the correlation coefficient

Combining several different populations



In one sample the correlation is high and positive and in the other it is high and negative. The correlation for the combined samples is zero.

General strategy for data checking

- identify problem data (cases/variables)
- check whether problem data are influential
  - run analyses with the problem corrected
  - run analyses without correcting the problem
  - compare results of the analyses
- if results are substantively different, *decide* which results are more trustworthy

Back to the multiple regression example...

T&F example

Importance of residuals

Remember:

DATA = MODEL + RESIDUAL

or

RESIDUAL = DATA – MODEL

...it's what is left after the model has been fitted.

For the linear model, one of the most useful diagnostics is the residual.

Violation of regression assumptions will influence the properties of the residual scores.

Therefore, various analyses of residuals are useful for determining whether assumptions have been violated.

Importance of residuals

The most appropriate type of residual to work with is the Studentised deleted residual, SDRESID. This the residual obtained from an analysis in which an observation is deleted from the analysis and its predicted value determined by the other observations. This residual is divided by an estimate of its standard error and so becomes a t-value. It can then be tested for statistical significance.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT ltimedrs
  /METHOD=ENTER menheal lphyheal sstress
  /SCATTERPLOT=(*ZRESID ,*ZPRED )
  /RESIDUALS DURBIN HIST(ZRESID) NORM(ZRESID) OUTLIERS (SDRESID MAHAL COOK).
```

If the assumptions are satisfied, then patterns in the different types of residuals will be similar. As potential problems arise, SDRESID will make suspicious observations more obvious.

Some Assumptions:

Independence of Observations

- Regression assumes that each subject's data has no influence on any other subject's data. That is, each observation is independent of the others.
- Check via the Durbin-Watson statistic:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.614[a]	.377	.373	.32888	2.064

a. Predictors: (Constant), SSTRESS, LPHYHEAL, Mental health symptoms

b. Dependent Variable: LTIMEDRS

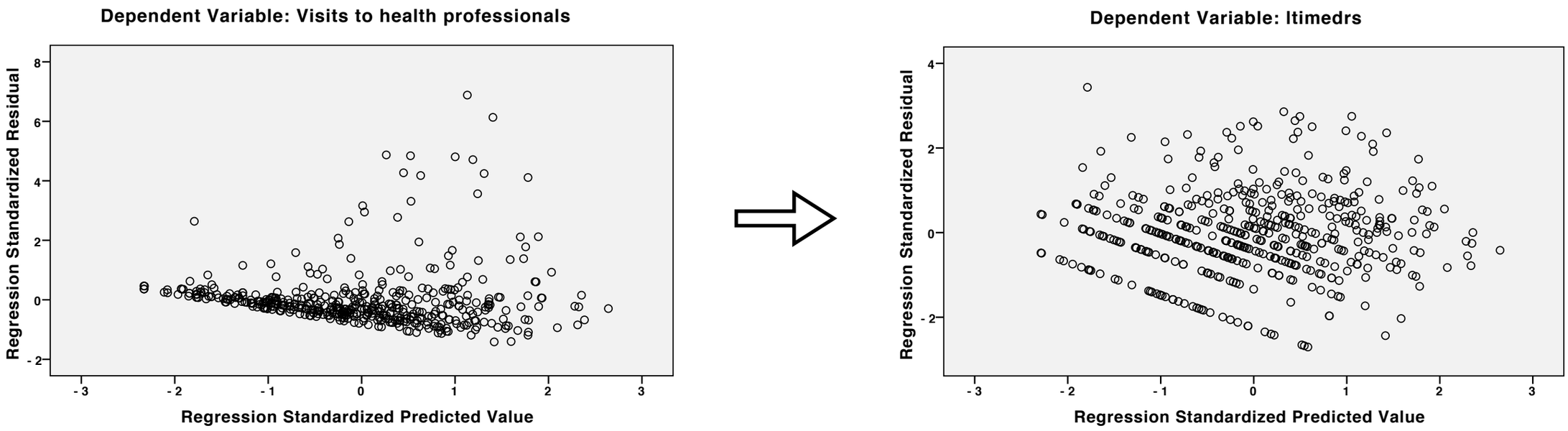
  - The Durbin-Watson statistic always lies between 0 and 4, the expected value of the Durbin-Watson statistic is 2. Values less than 2 indicate a positive correlation, values greater than 2 indicate a negative correlation.

- Remedies: discard dependent data.

Some Assumptions:

Linearity

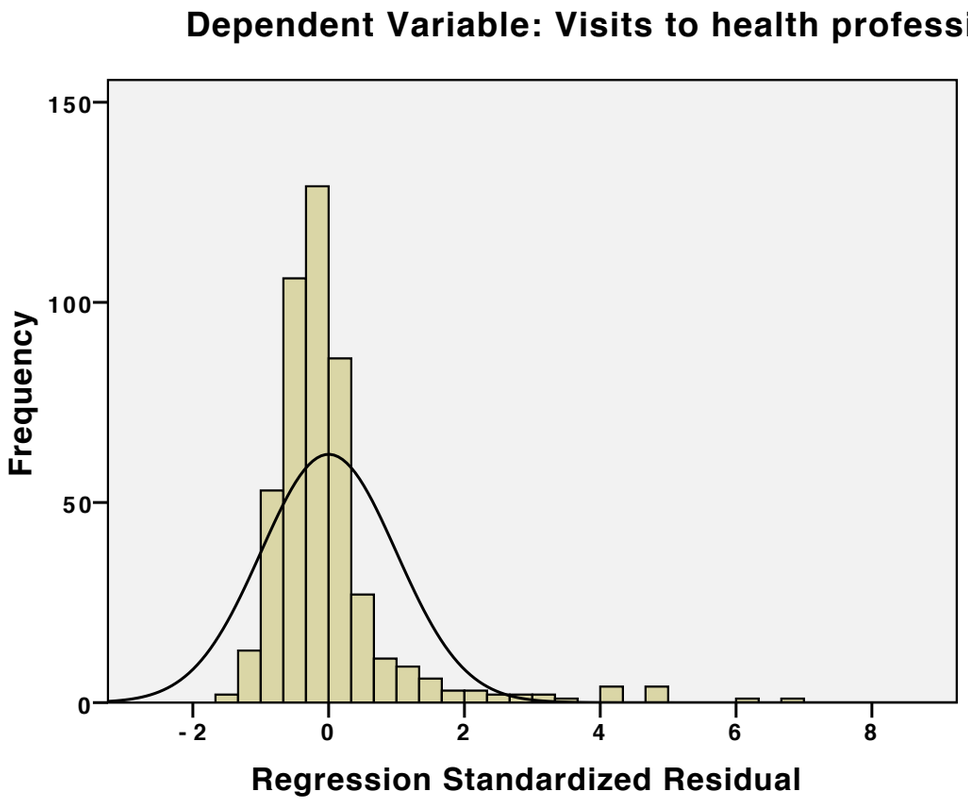
- Regression assumes the relationships between predictors and criterion are linear.
- Scatterplots between each pair of variables should be linear.
- The residuals plot of RESID and PRED should show linearity. That is, no pattern should be evident.
- Remedies: transformations

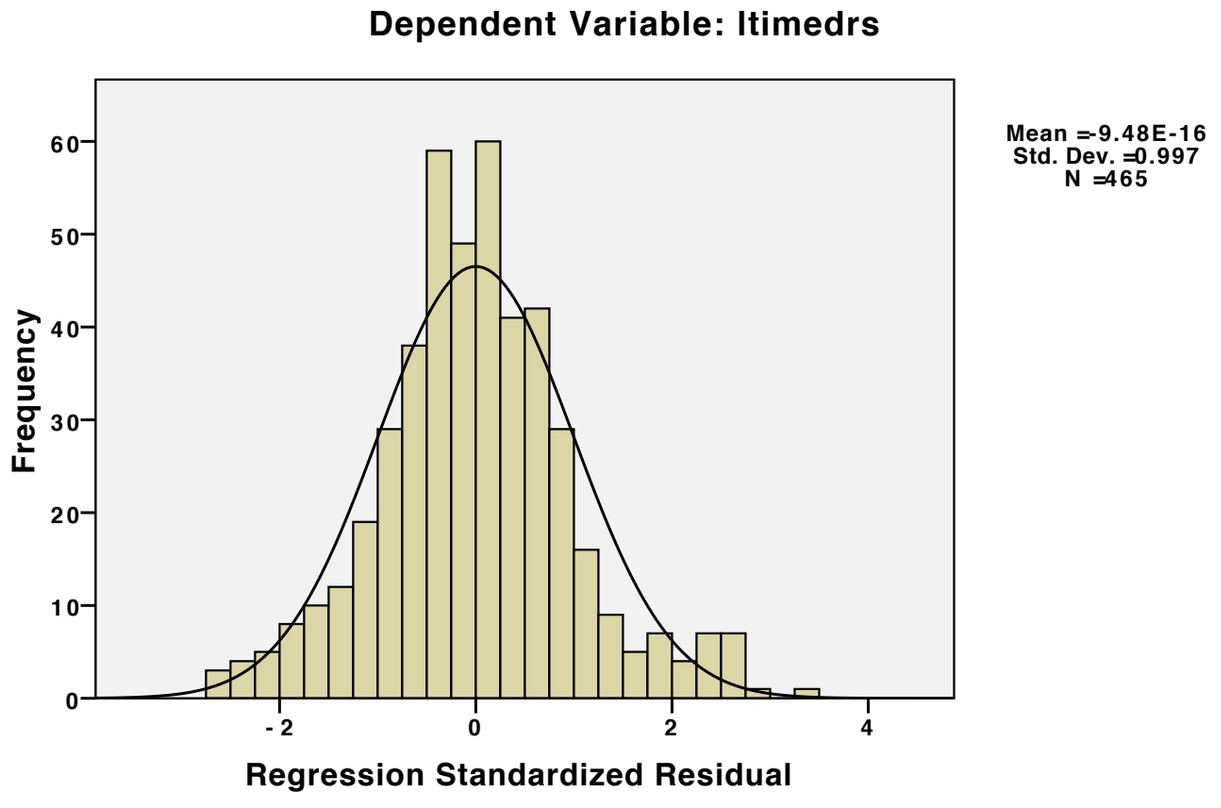


Some Assumptions:

Normality

- Regression assumes that all variables are normally distributed.
- Each pair of variables should show bivariate normality.
- The residuals should be normally distributed.
- Remedies: transformations, removal of outliers.

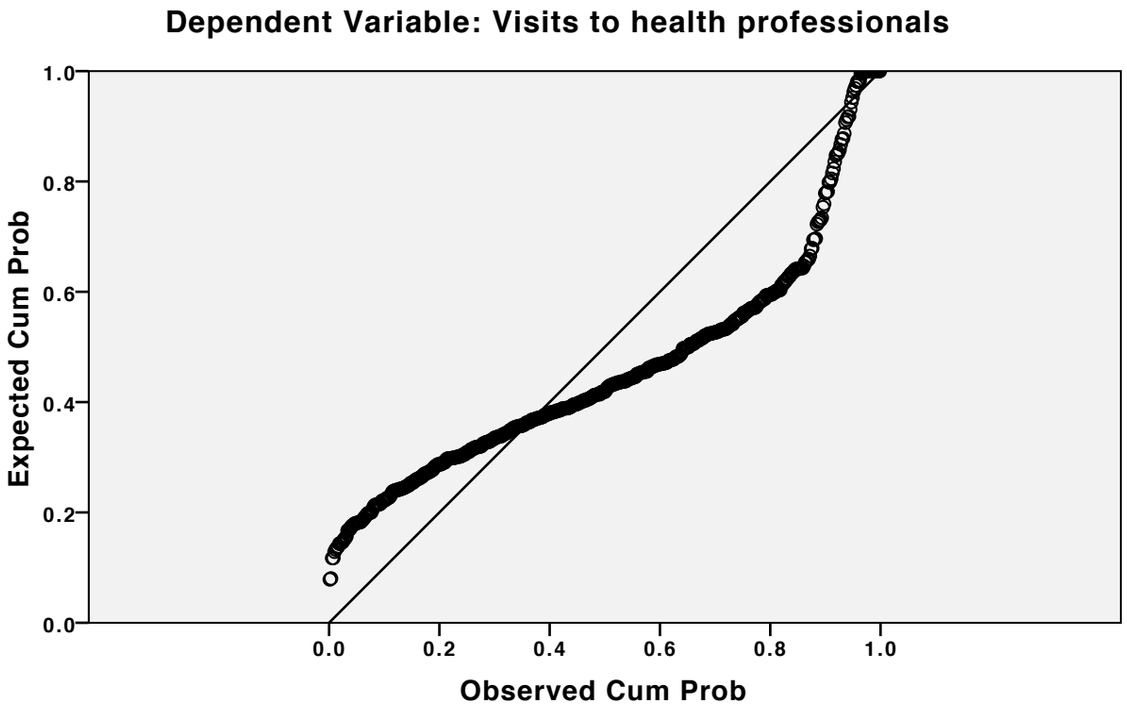


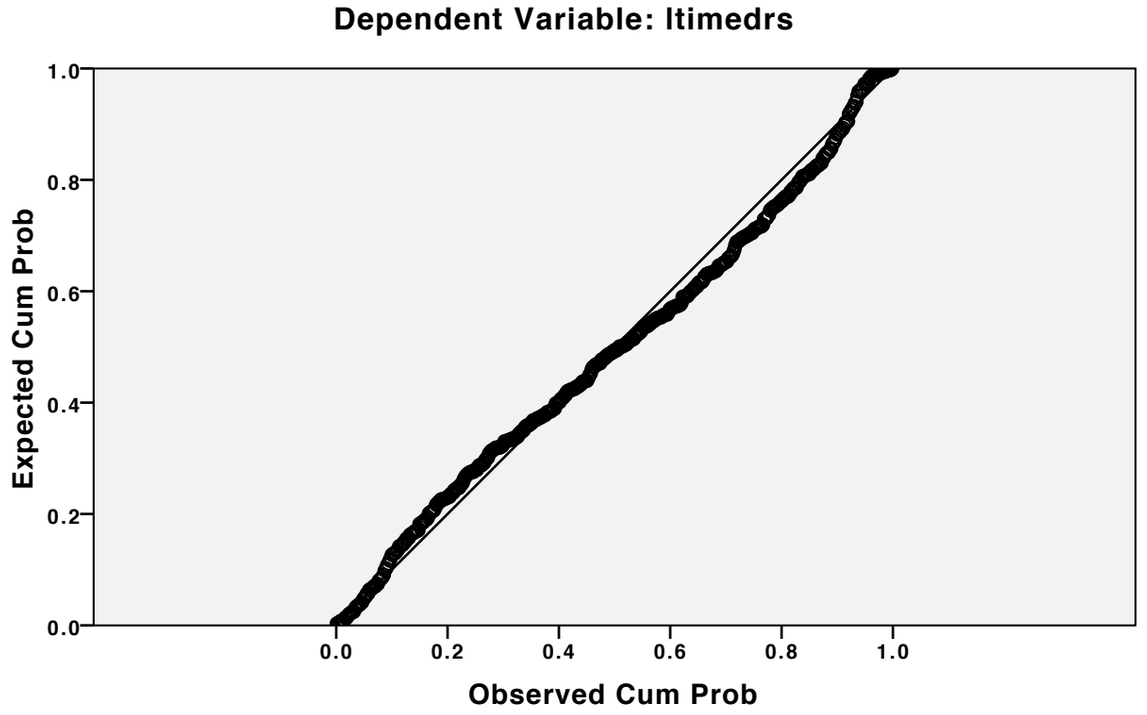


Some Assumptions:

Normality

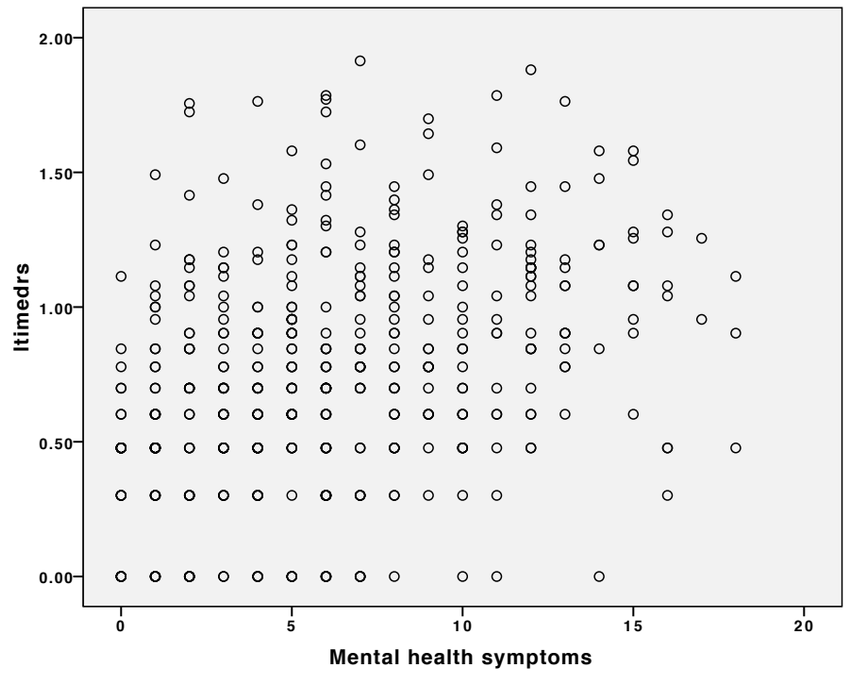
- Regression assumes that all variables are normally distributed.
- Each pair of variables should show bivariate normality.
- The residuals should be normally distributed.
- Remedies: transformations, removal of outliers.





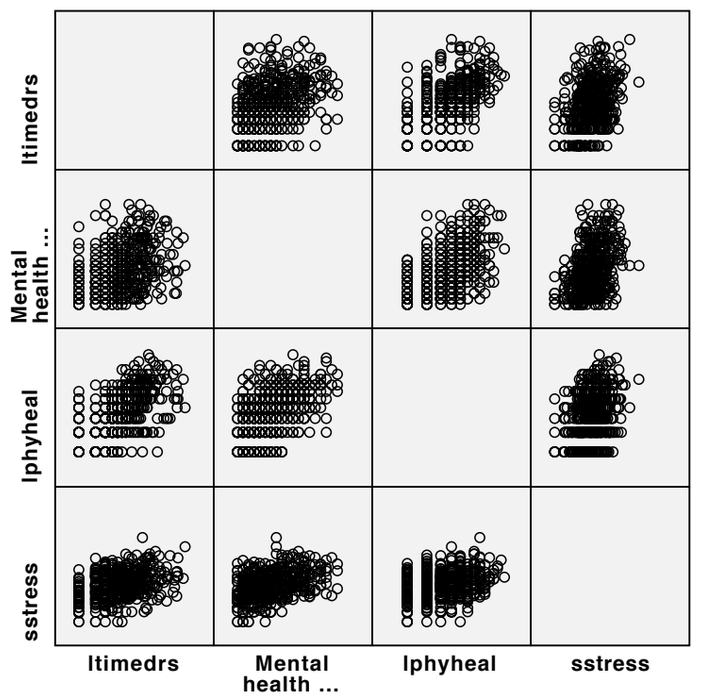
Outliers and influential data points

- Scatterplots of the two variables may show points that are outside the shape of the bivariate distribution.



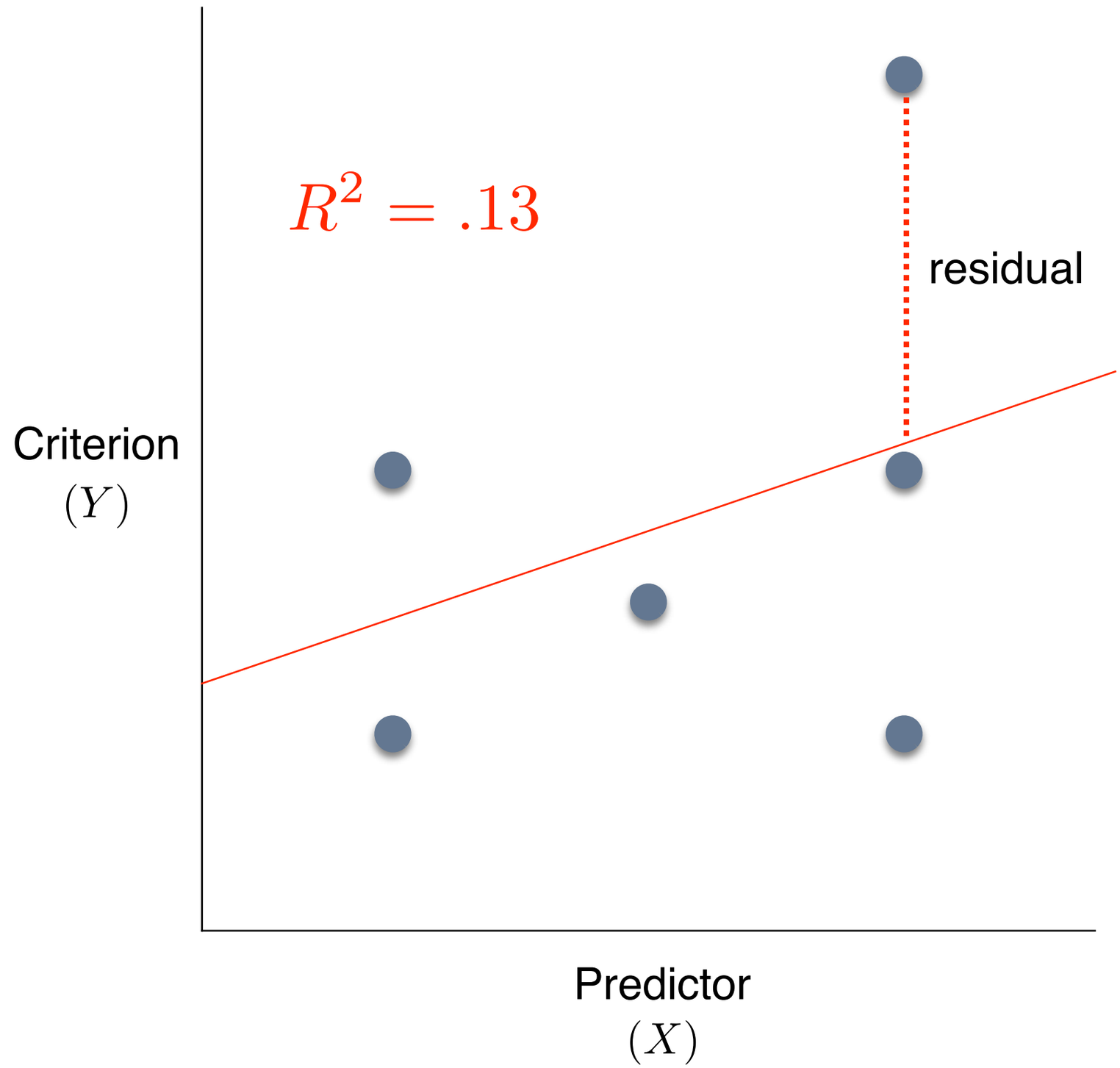
- In the multivariate case, scatterplots are not as convenient because of the potentially large number of scatterplots. The use of the matrix of scatterplots in SPSS can be useful.

```
GRAPH
  /SCATTERPLOT(MATRIX)=ltimedrs menheal lphyheal sstress
```



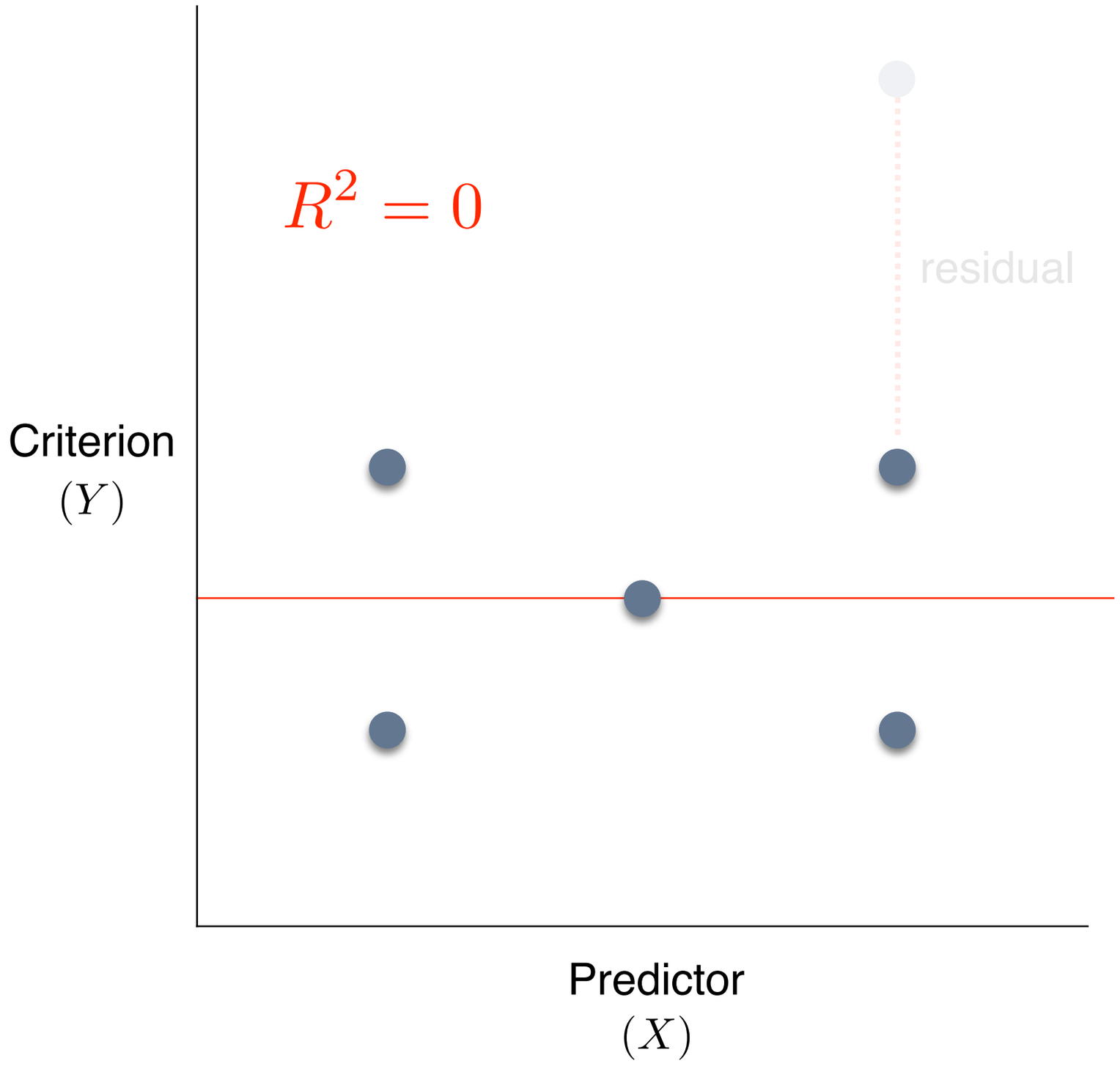
Outliers and influential data points

Outlier on Criterion



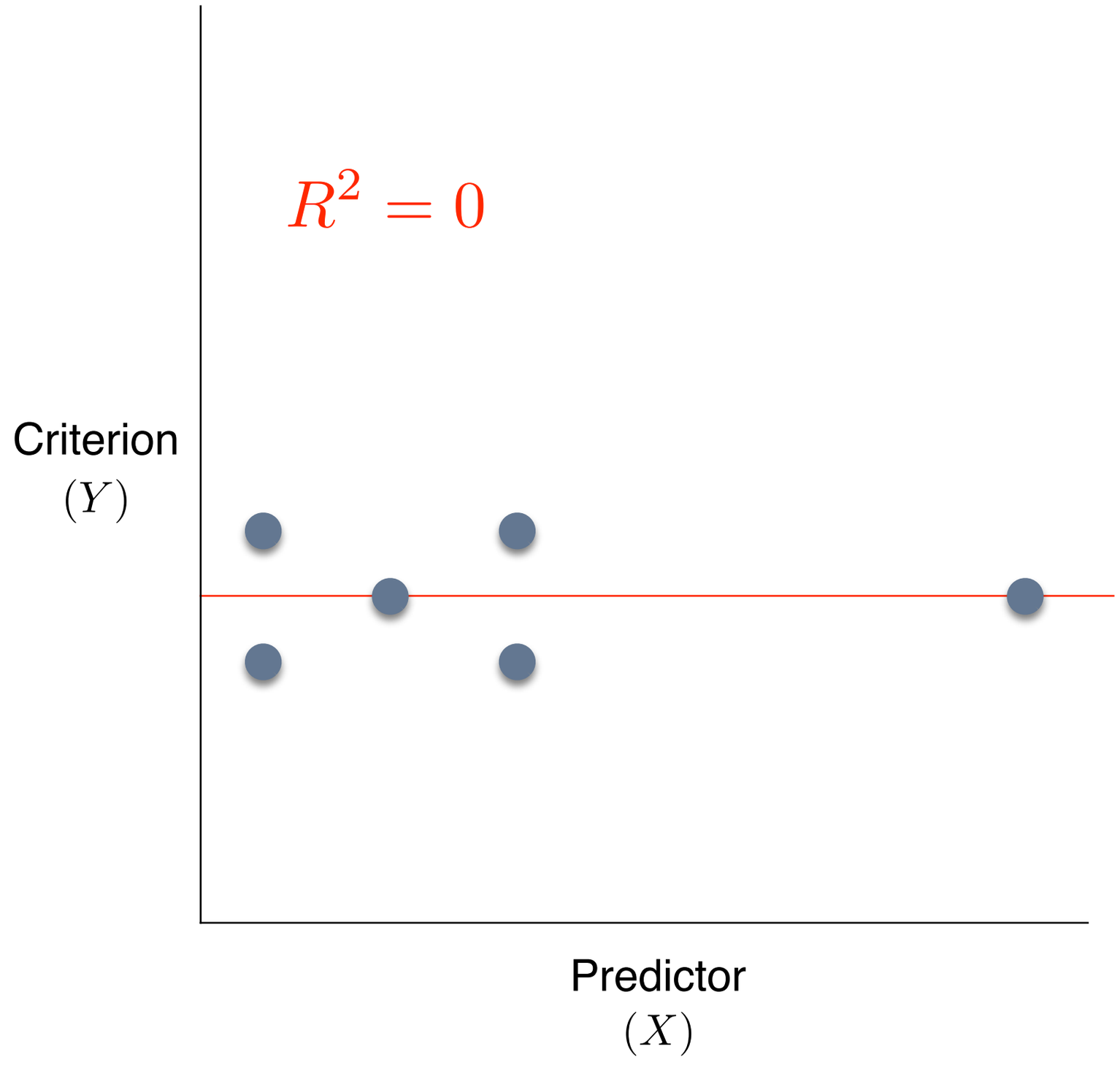
Outliers and influential data points

Outlier on Criterion



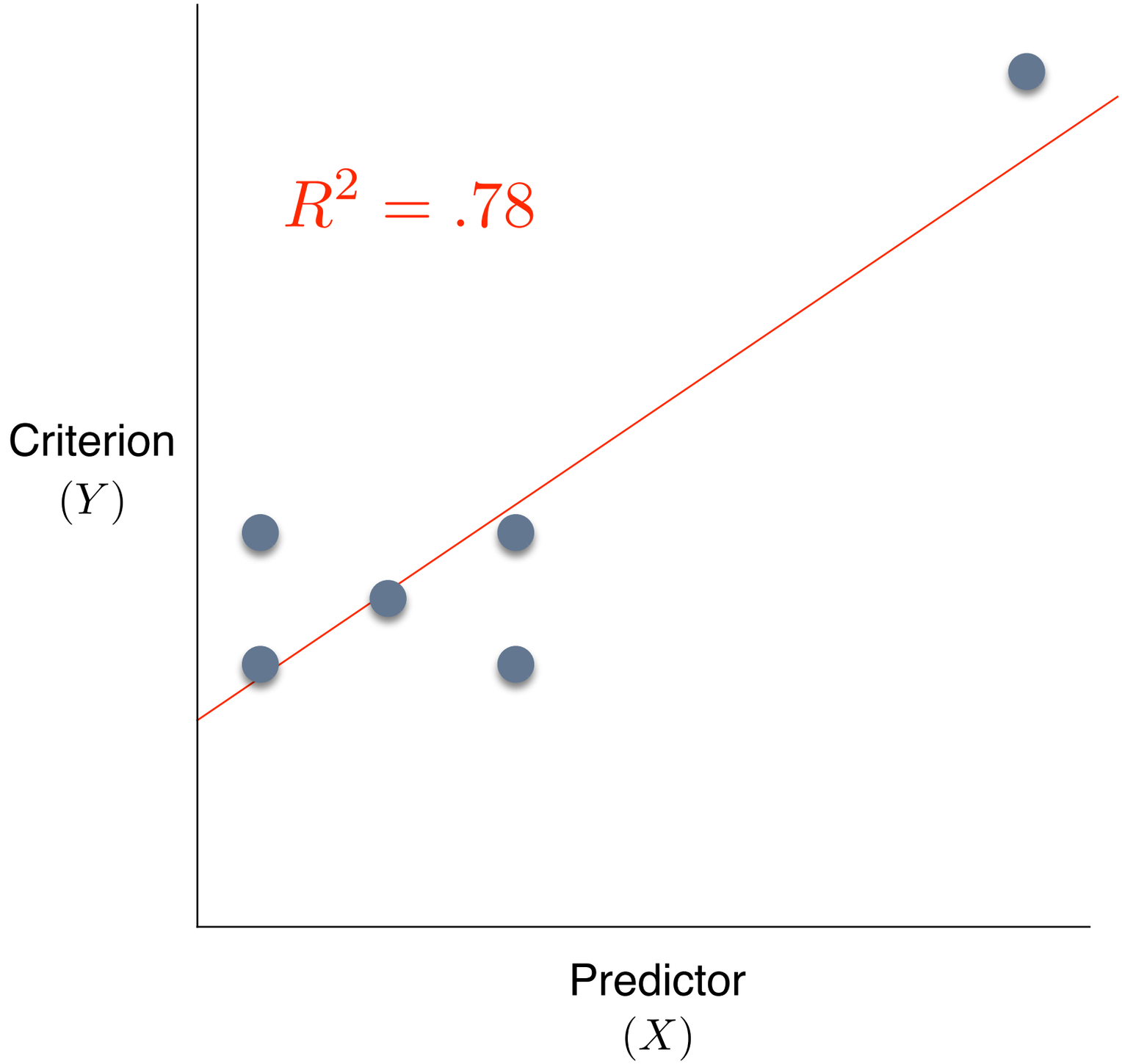
Outliers and influential data points

Outlier on Predictor



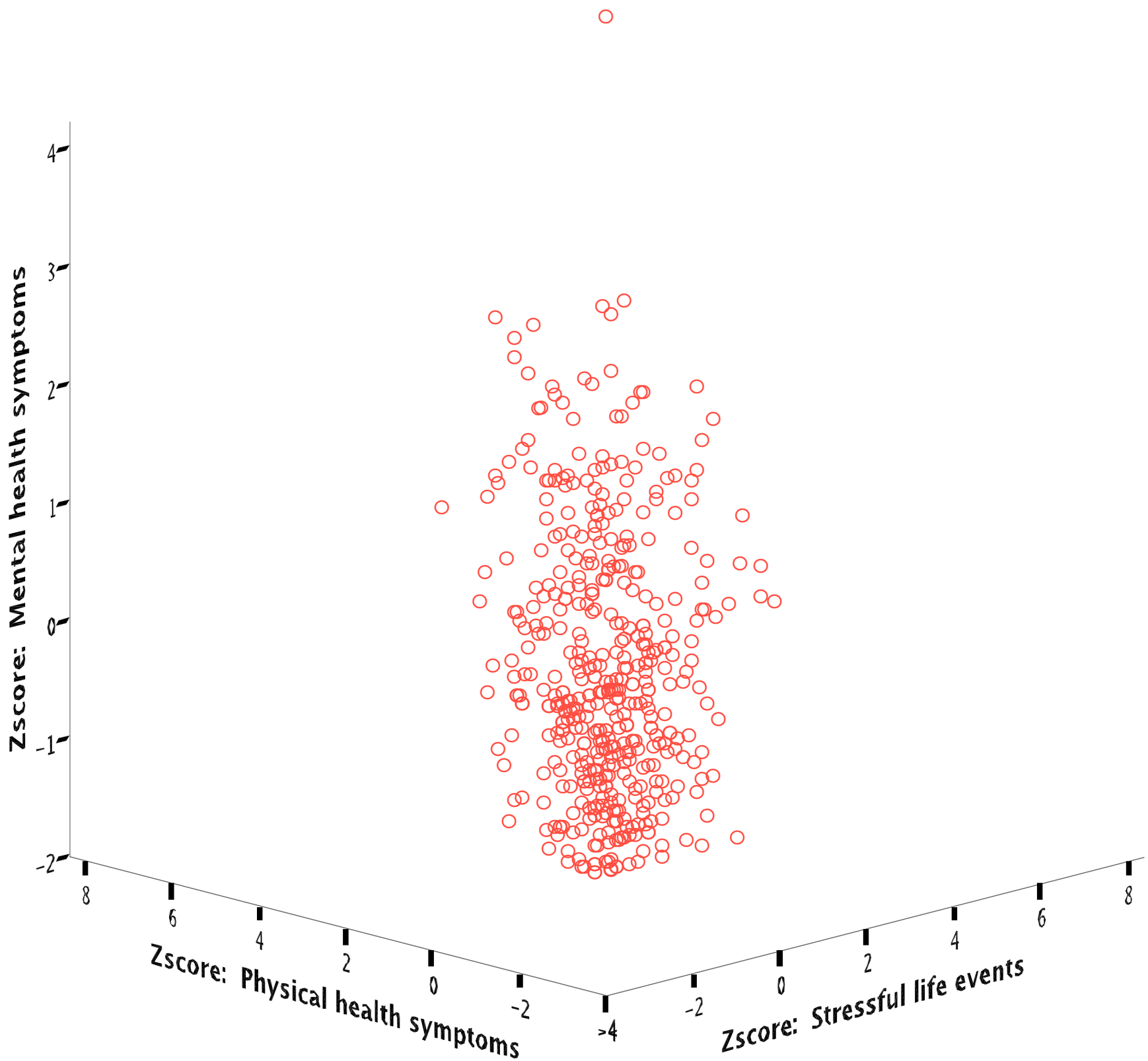
Outliers and influential data points

Influential Data Point



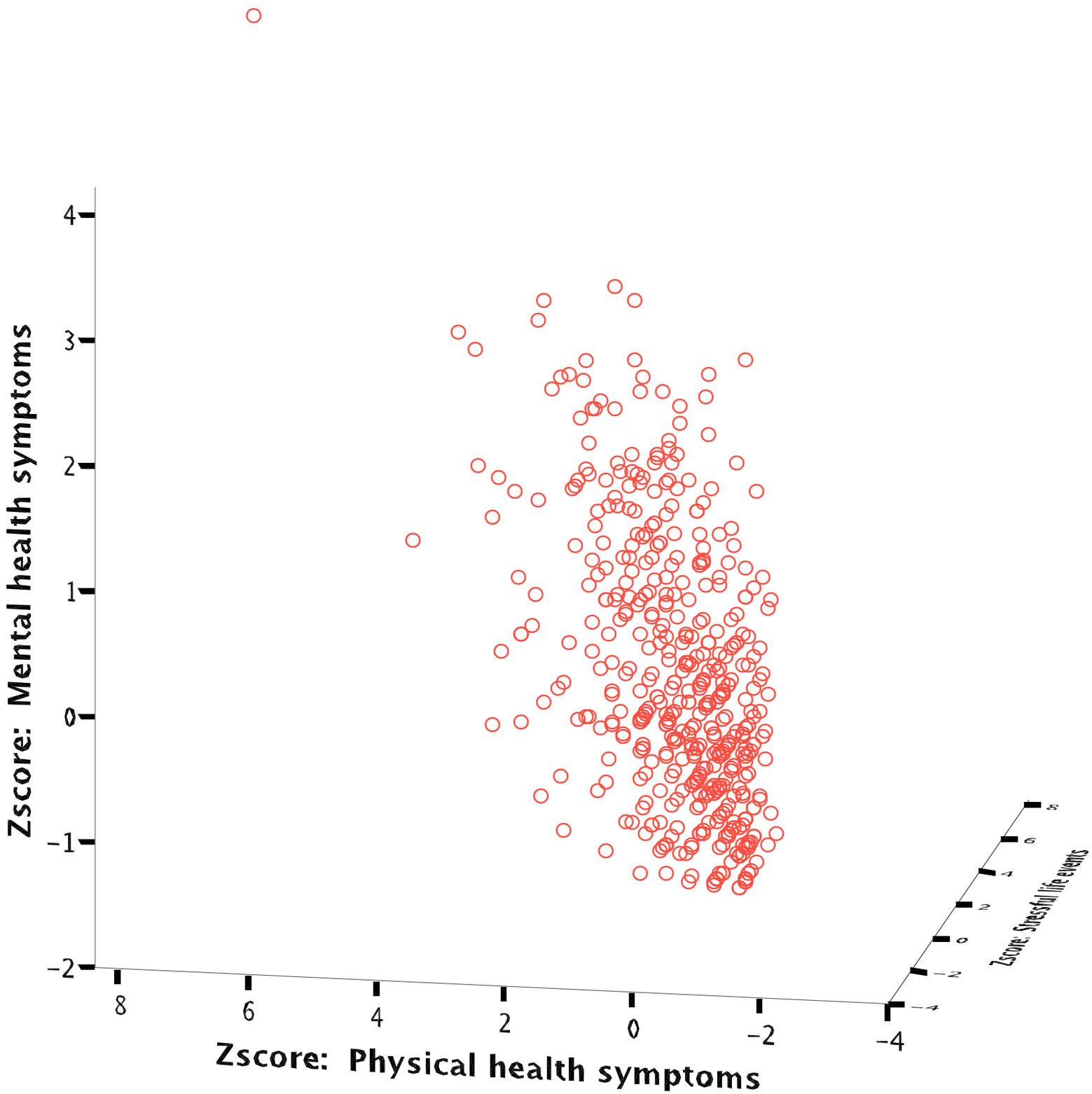
Outliers and influential data points

Influential Data Point: A data point added to the T&F data to illustrate outliers in the space of the predictors



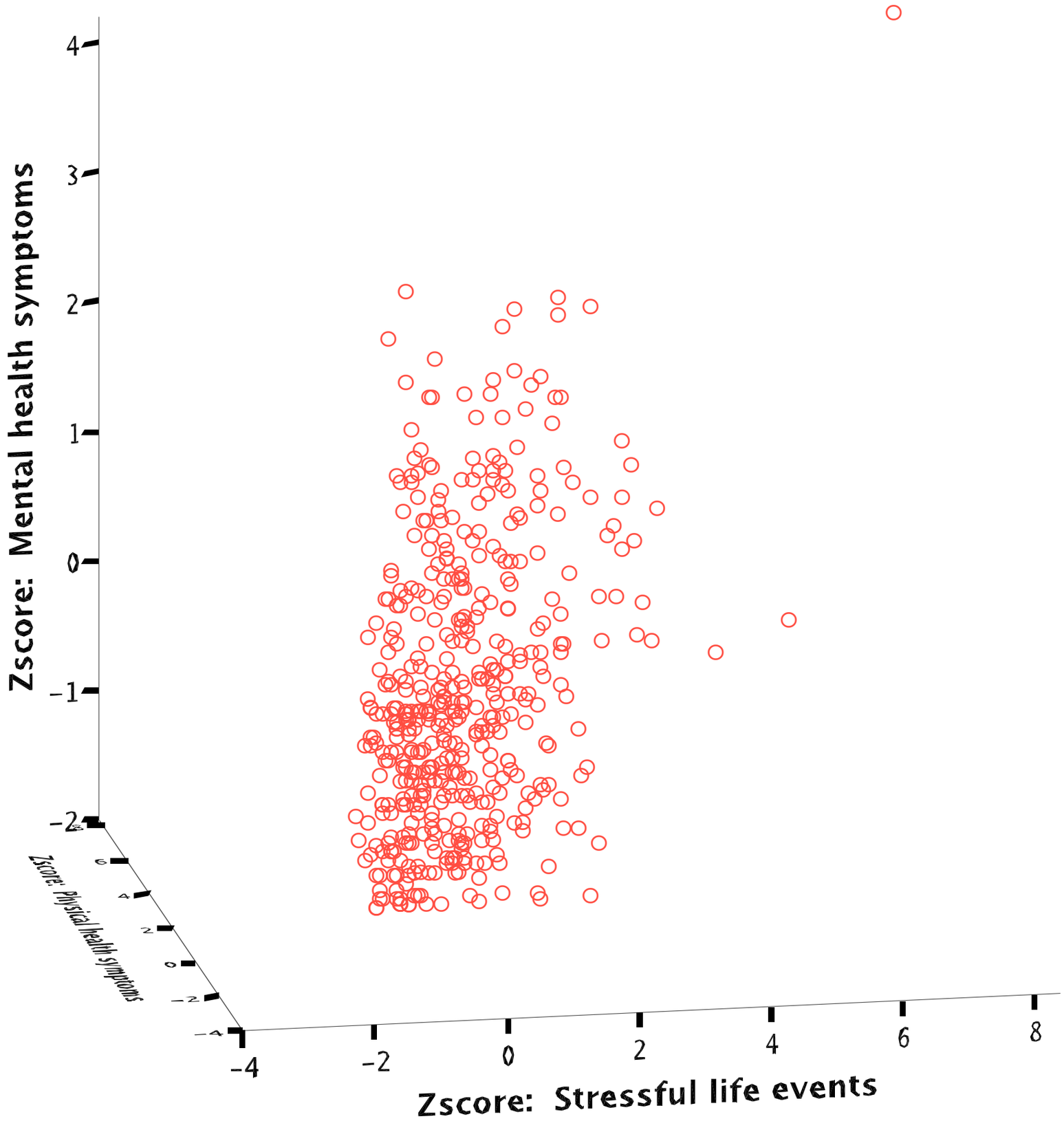
Outliers and influential data points

Influential Data Point: A data point added to the T&F data to illustrate outliers in the space of the predictors



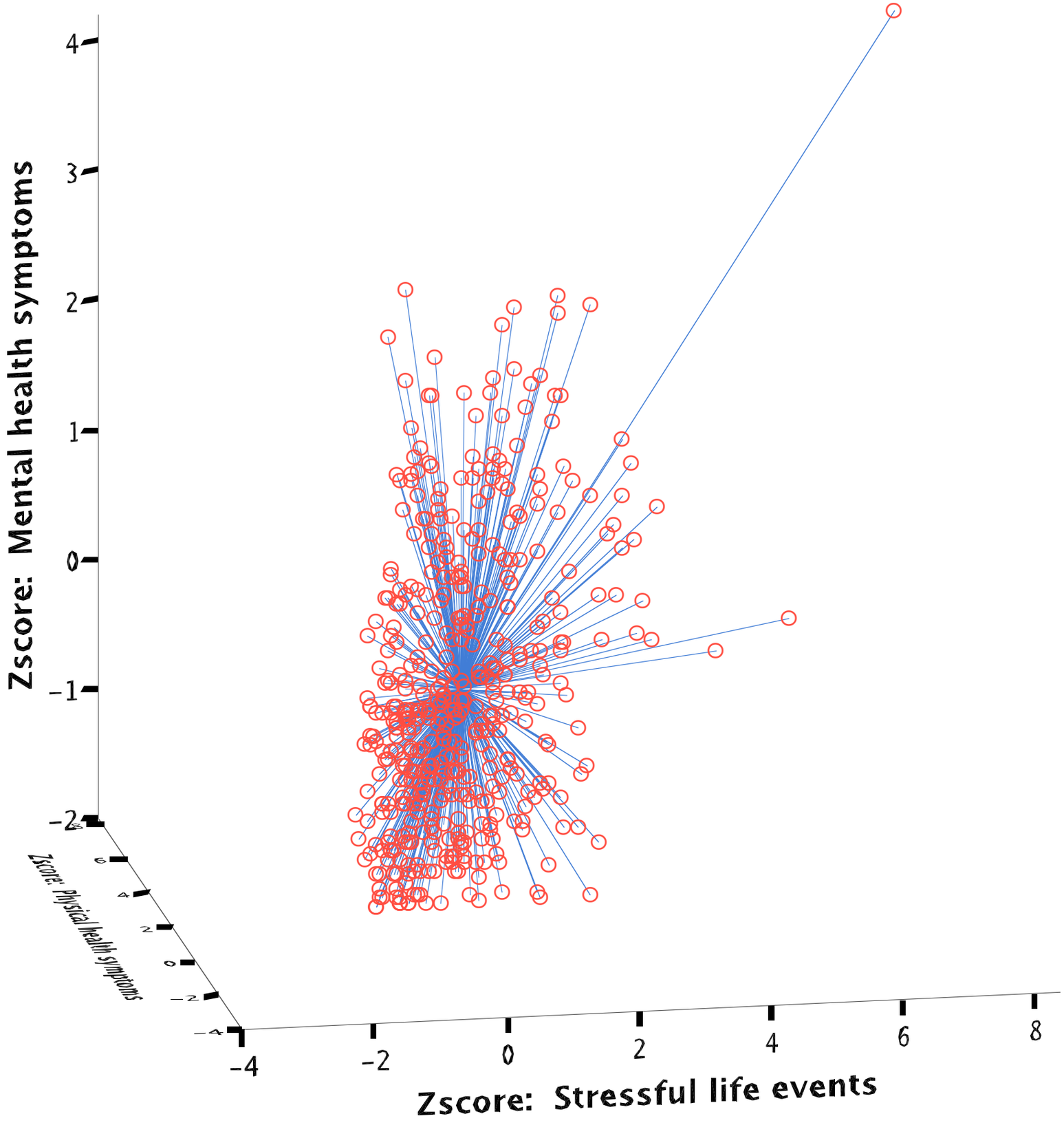
Outliers and influential data points

Influential Data Point: A data point added to the T&F data to illustrate outliers in the space of the predictors



Outliers and influential data points

Influential Data Point: A data point added to the T&F data to illustrate outliers in the space of the predictors



Outliers and influential data points

Outliers on Criterion

Look at the largest SDRESIDs.

These can be tested for significance as outliers. They follow a Student's t with (N-p-2) degrees of freedom, so they can be tested for significance. Critical values need to be looked up in tables, use p < .001.

		Case Number	Subject number	Statistic
Stud. Deleted Residual	1	345	479	3.492
	2	275	370	2.900
	3	405	548	2.798
	4	249	330	2.772
	5	143	184	-2.730
	6	43	51	-2.704
	7	71	86	-2.680
	8	213	291	2.674
	9	67	82	2.643
	10	286	386	2.549

For SDRESID the critical t values with 460 df are:

$p = .05$	$t = 1.96$
$p = .01$	$t = 2.57$
$p = .001$	$t = 3.29$

Outliers and influential data points

Outliers in the space of the predictors

Look at the largest Mahalanobis (MAHAL) distances for each case.

This is the distance of an observation, in the space of the predictors, from the centroid, defined by the means of the variables, of the space. The MAHAL values are distributed as Chi-square with p degrees of freedom, where p is number of predictors, if N-p > 50. Thus, the MAHAL values can be tested for significance, use p < 0.001.

		Case Number	Subject number	Statistic
Mahal. Distance	1	403	540	14.135
	2	125	145	11.649
	3	198	267	10.569
	4	52	67	10.548
	5	446	685	10.225
	6	159	226	9.351
	7	33	37	8.628
	8	280	380	8.587
	9	405	548	8.431
	10	113	133	8.353

For MAHAL the critical χ^2 values with 3 df are:

$p = .05$	$\chi^2 = 7.81$
$p = .01$	$\chi^2 = 11.34$
$p = .001$	$\chi^2 = 16.27$

Prasanta Chandra Mahalanobis
1893–1972

Outliers and influential data points

Look at Cook's Distance (COOK D) values

This is a measure of the change in the regression coefficients that would occur if this observation were deleted. It indicates which observations are most influential in affecting the regression equation. If these exceed 1, then the observation is considered influential. They are tested for significance in SPSS.

Influential data points

		Case Number	Subject number	Statistic
Cook's Distance	1	405	548	.040
	2	275	370	.028
	3	345	479	.027
	4	286	386	.023
	5	24	28	.020
	6	279	379	.020
	7	37	45	.020
	8	368	502	.018
	9	113	133	.016
	10	170	237	.015

While case 405 is an outlier on the criterion variable, (from SDRESID), Cook's distance indicates that it is not an influential data point, (Cook's distance < 1).

*Disregard the significance test provided by SPSS.

Outliers and influential data points

Suggested Remedies

Dropping cases from an analysis.

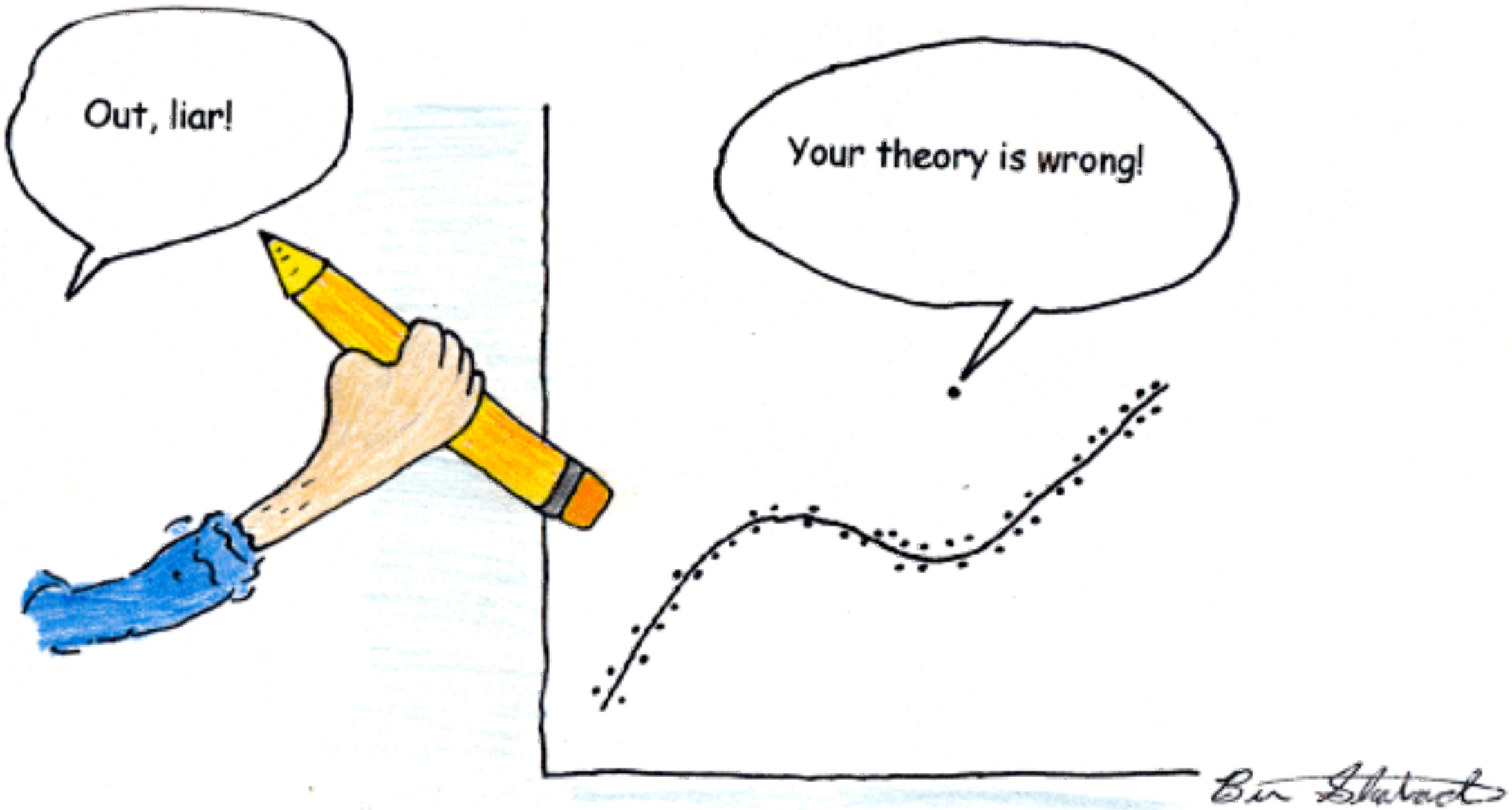
Potential outliers need to be identified by their unique ID such as subject number. The ID keyword in the residuals subcommand in the regression procedure does this. In a subsequent run, a case or cases can be dropped by inserting the following commands above the regression procedure. For example, suppose the cases with a subject numbers 479 and 236 and with a variable name of ‘subjno’ have been identified.

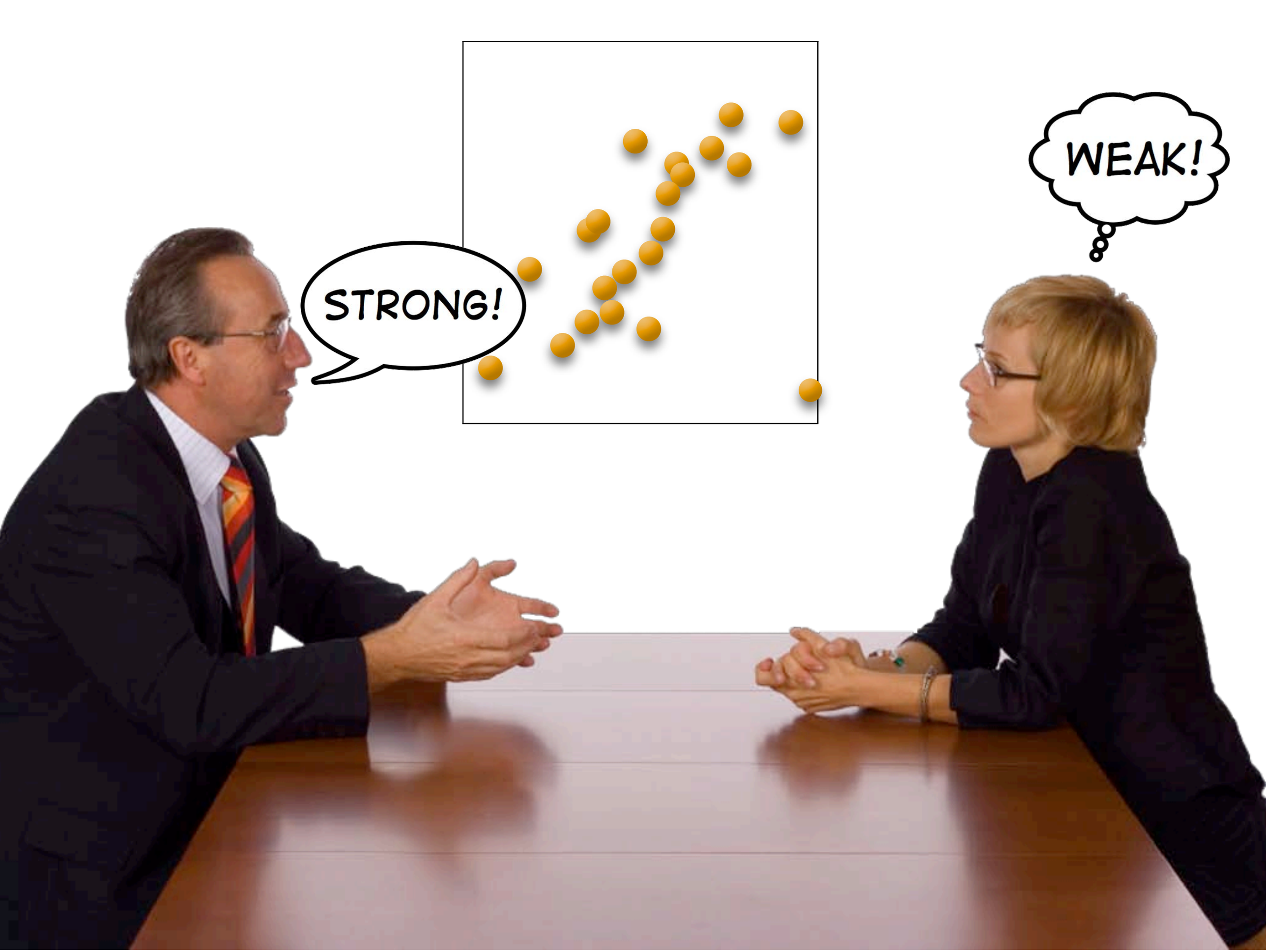
```
select if (subjno ~= 479)
select if (subjno ~= 236)
```

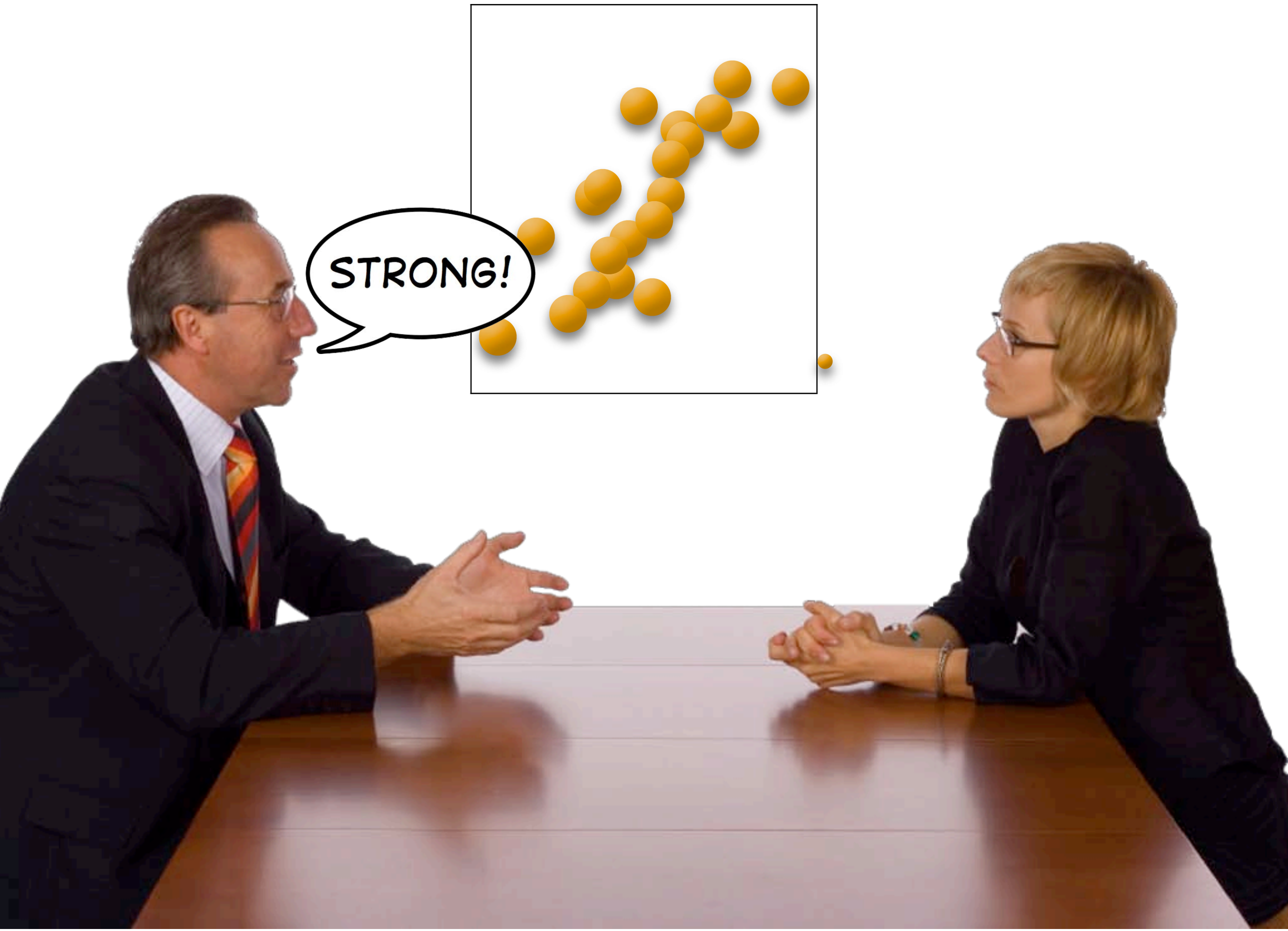
Transformations

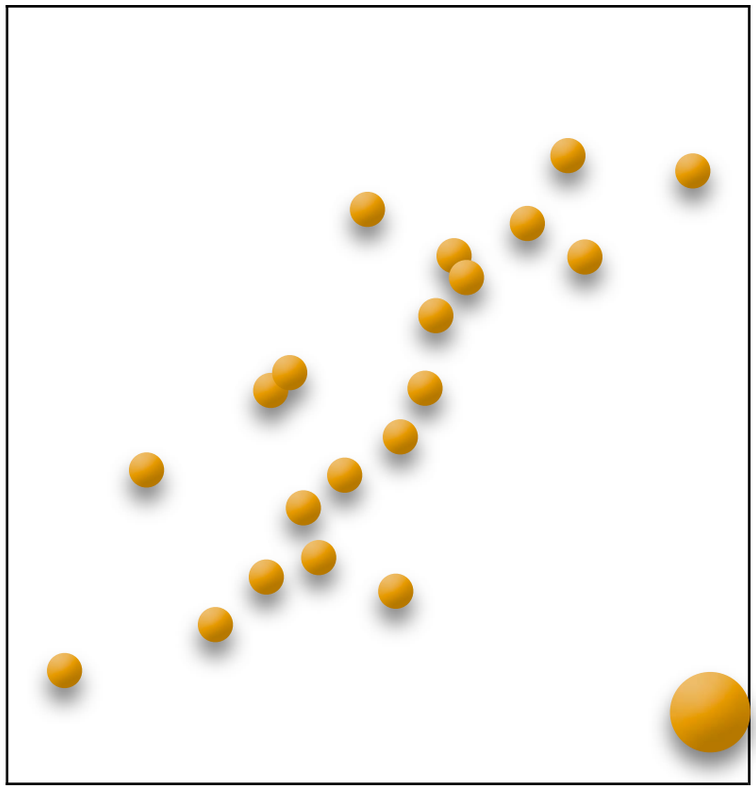
Tabachnick and Fidell (Chapter 4) detail the use of transformations for reducing the effects of skewness and for reducing the effects of ‘outlying’ data points.

A word of warning is in order here, for it is obvious that there is room for misuse of the above procedures. High-influence data points could conceivably be removed solely to effect a desired change in a particular estimated coefficient, i.e. t value, or some other regression output. While this danger exists, it is an unavoidable consequence of a procedure that successfully highlights such points. The benefits obtained from information in influential points far outweighs any dangers.

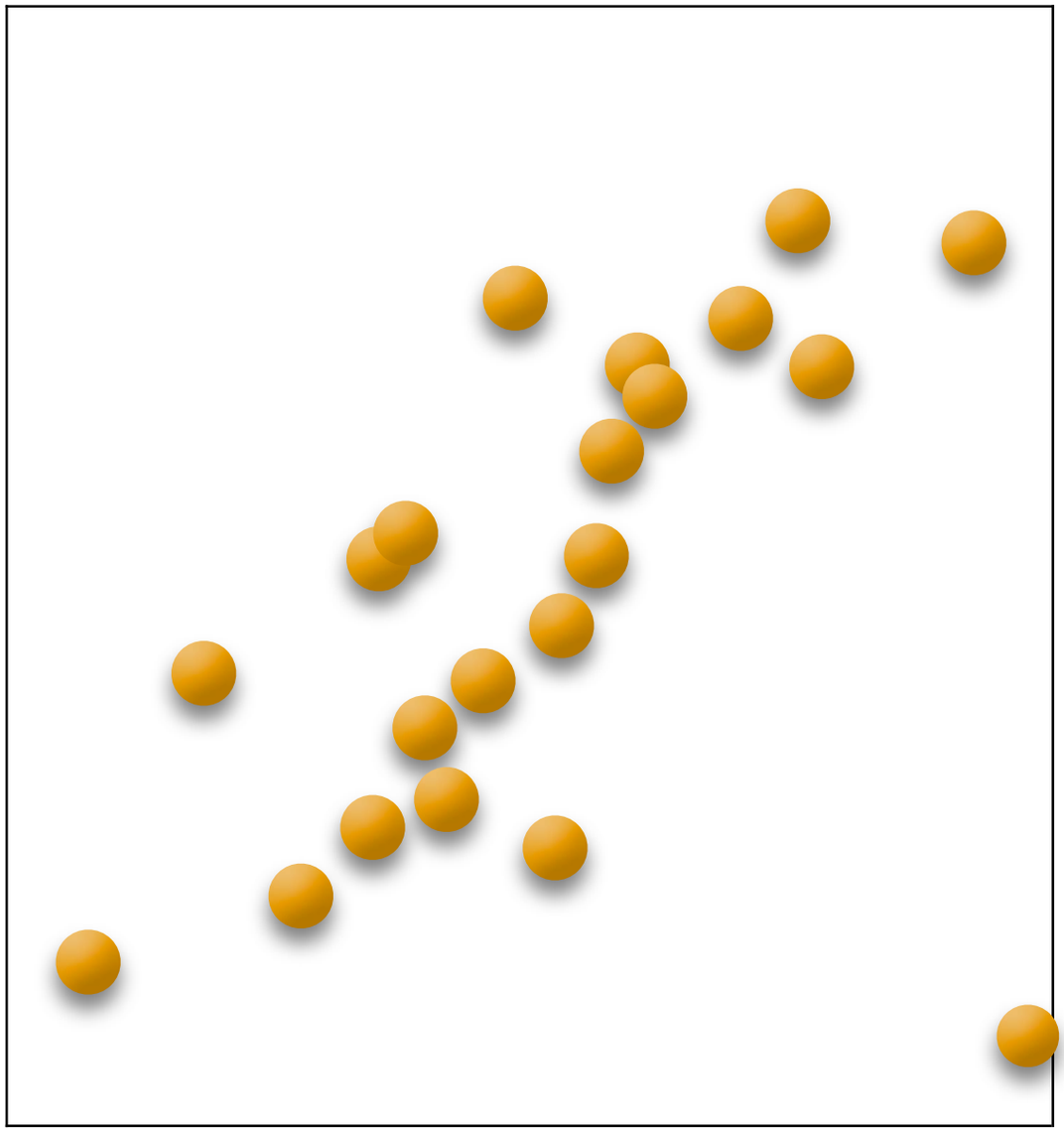


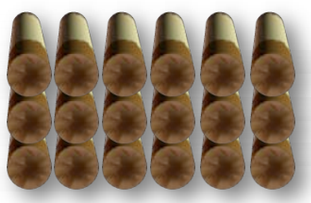

STRONG!
WEAK!

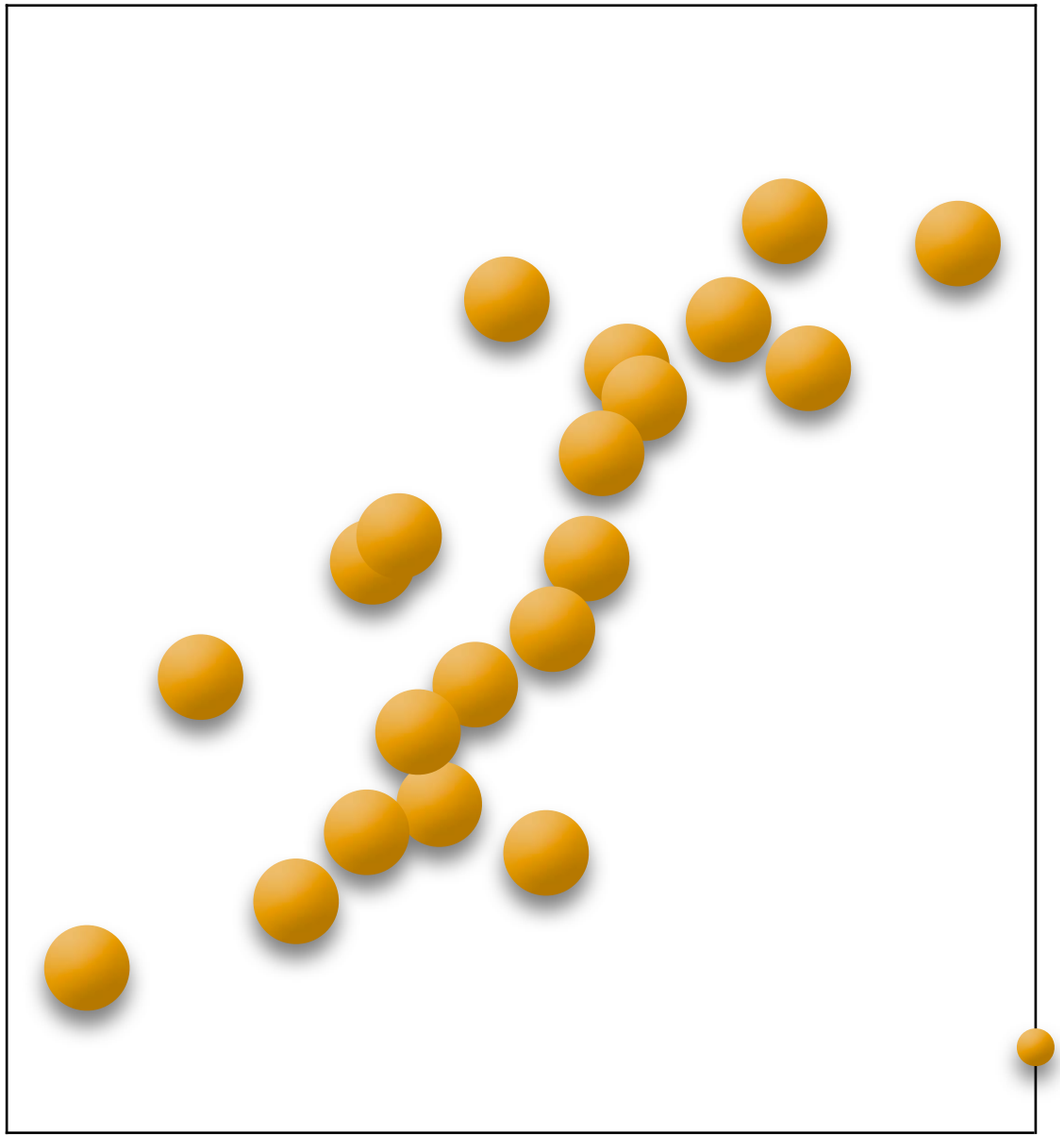

STRONG!

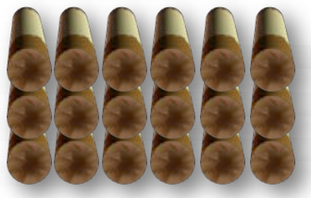


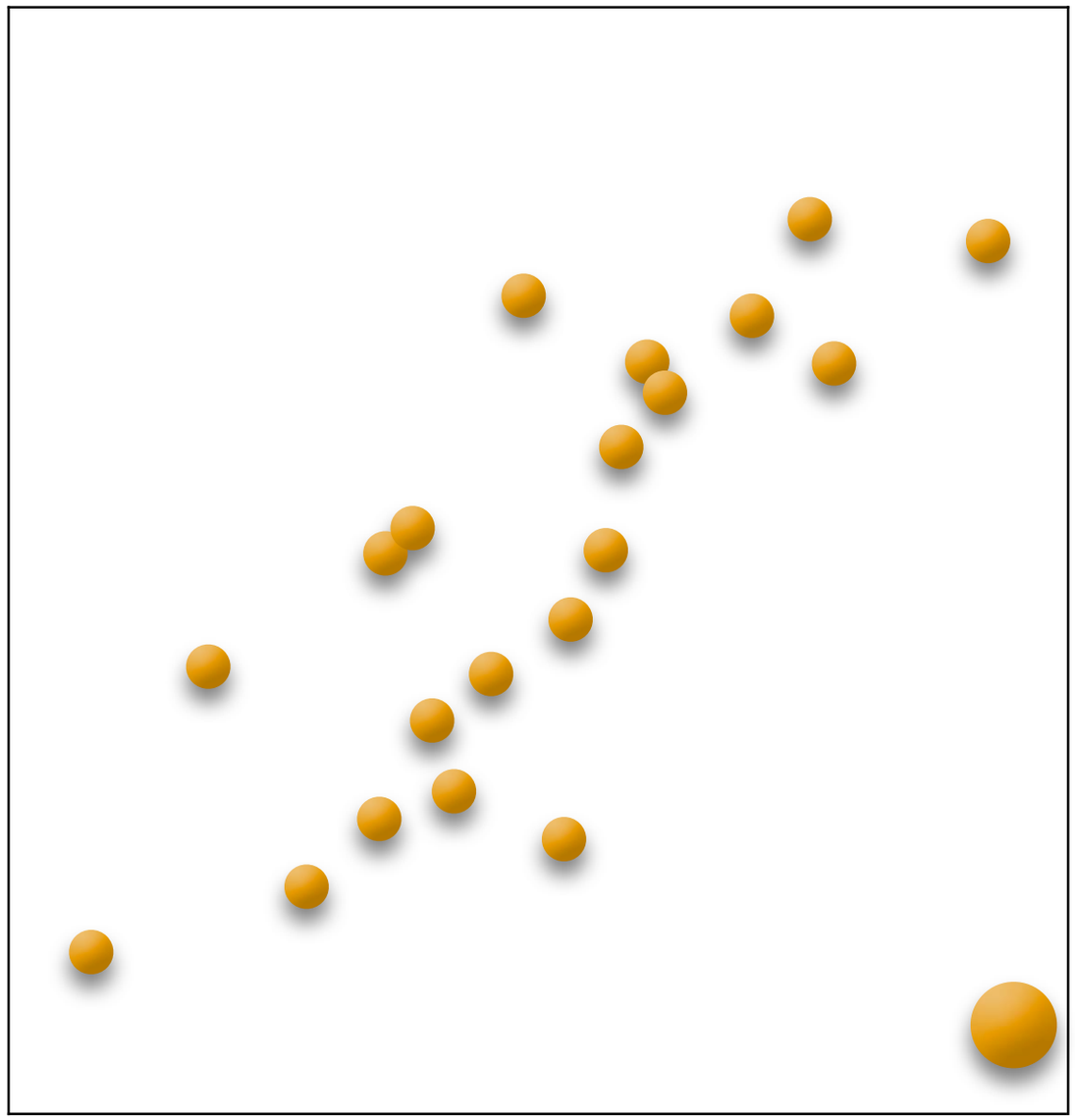
WEAK!

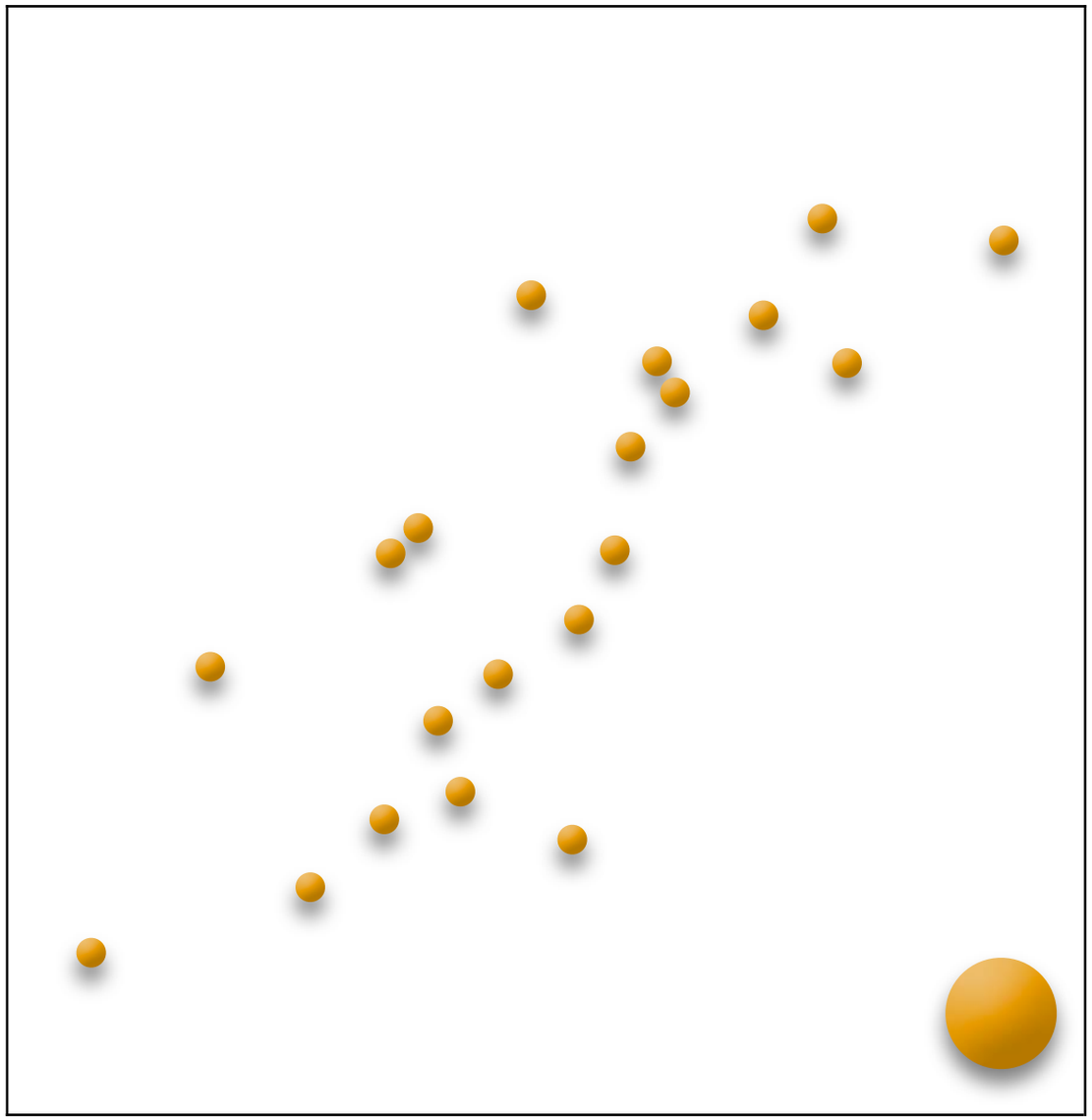


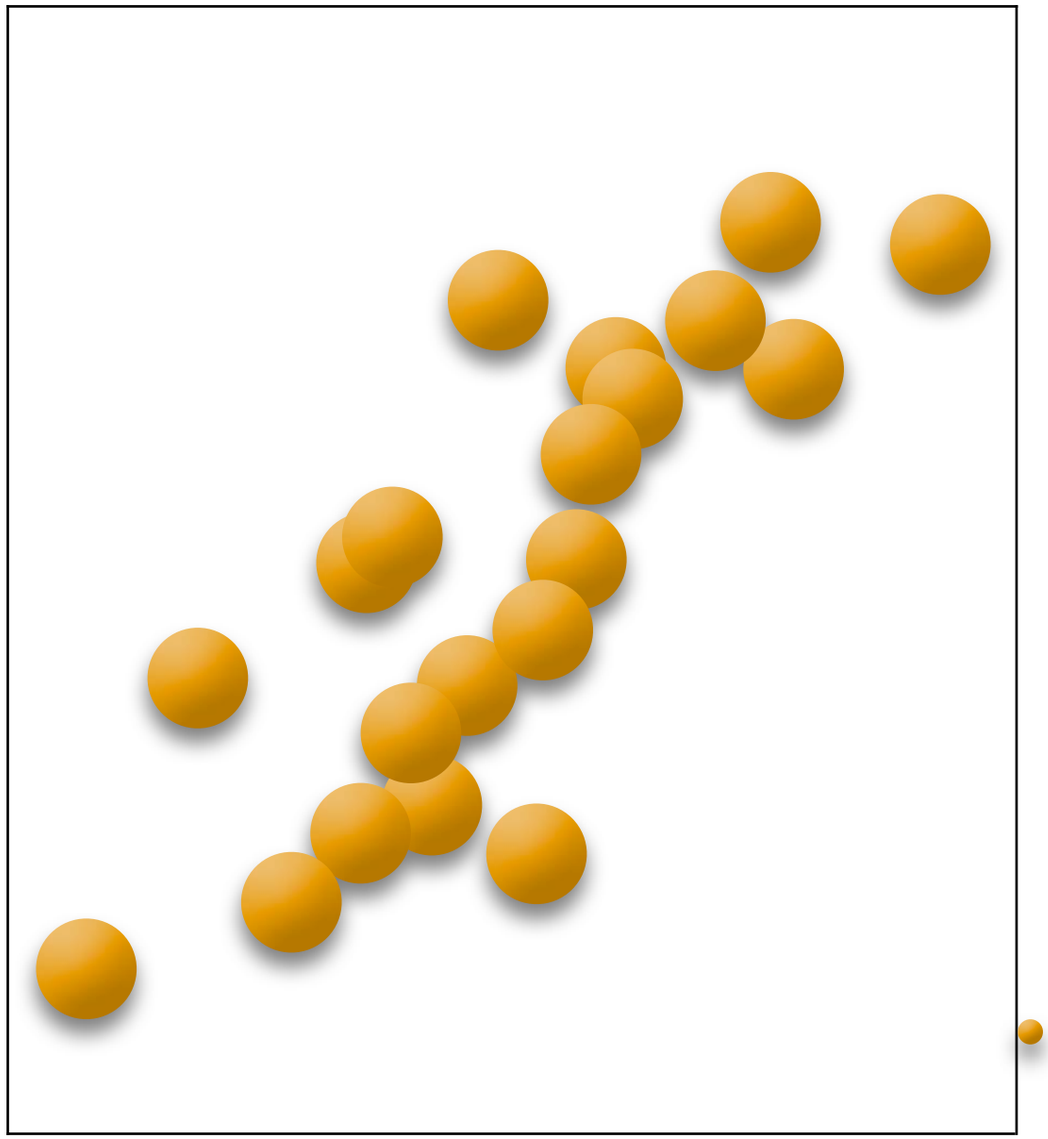


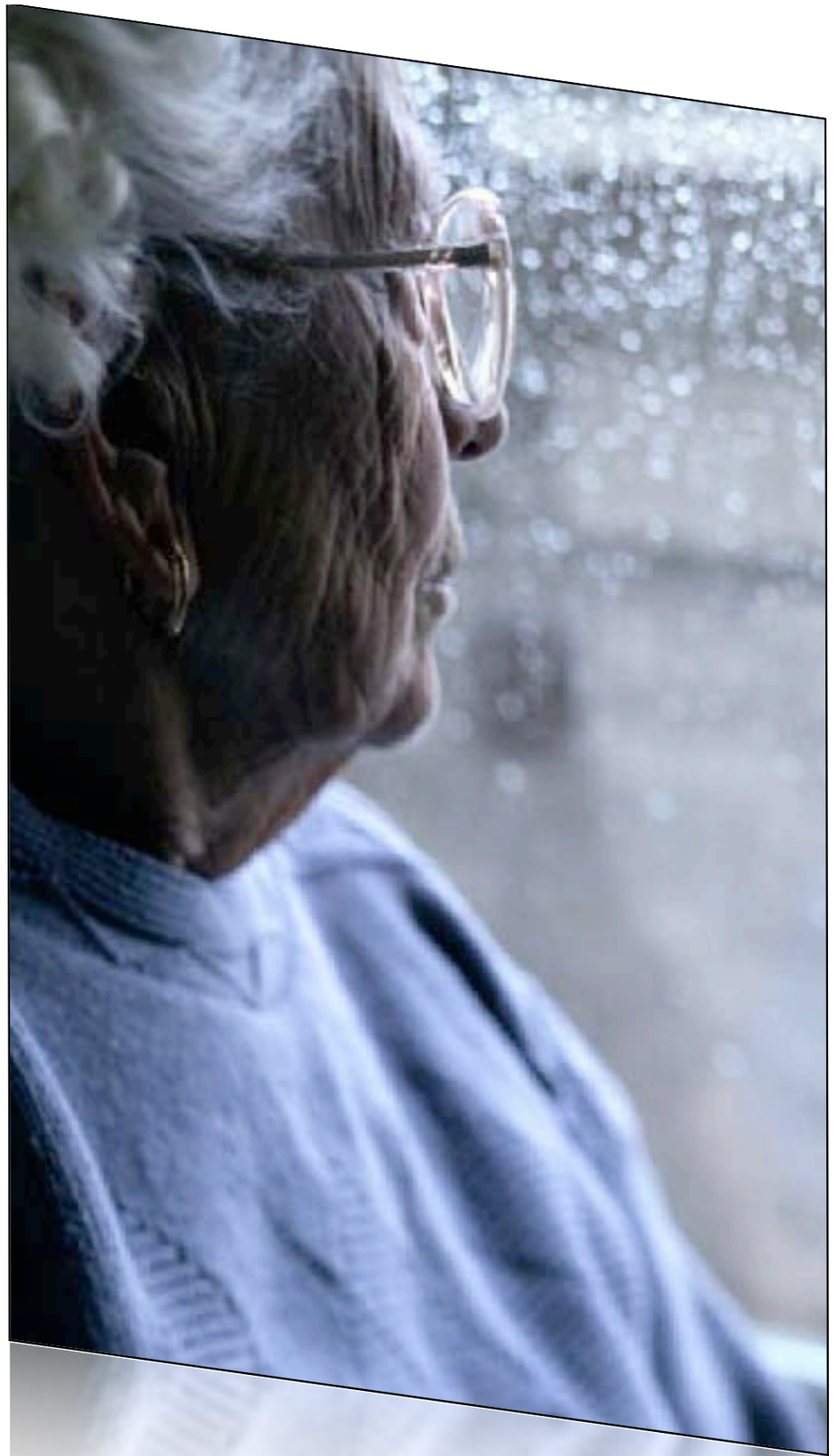










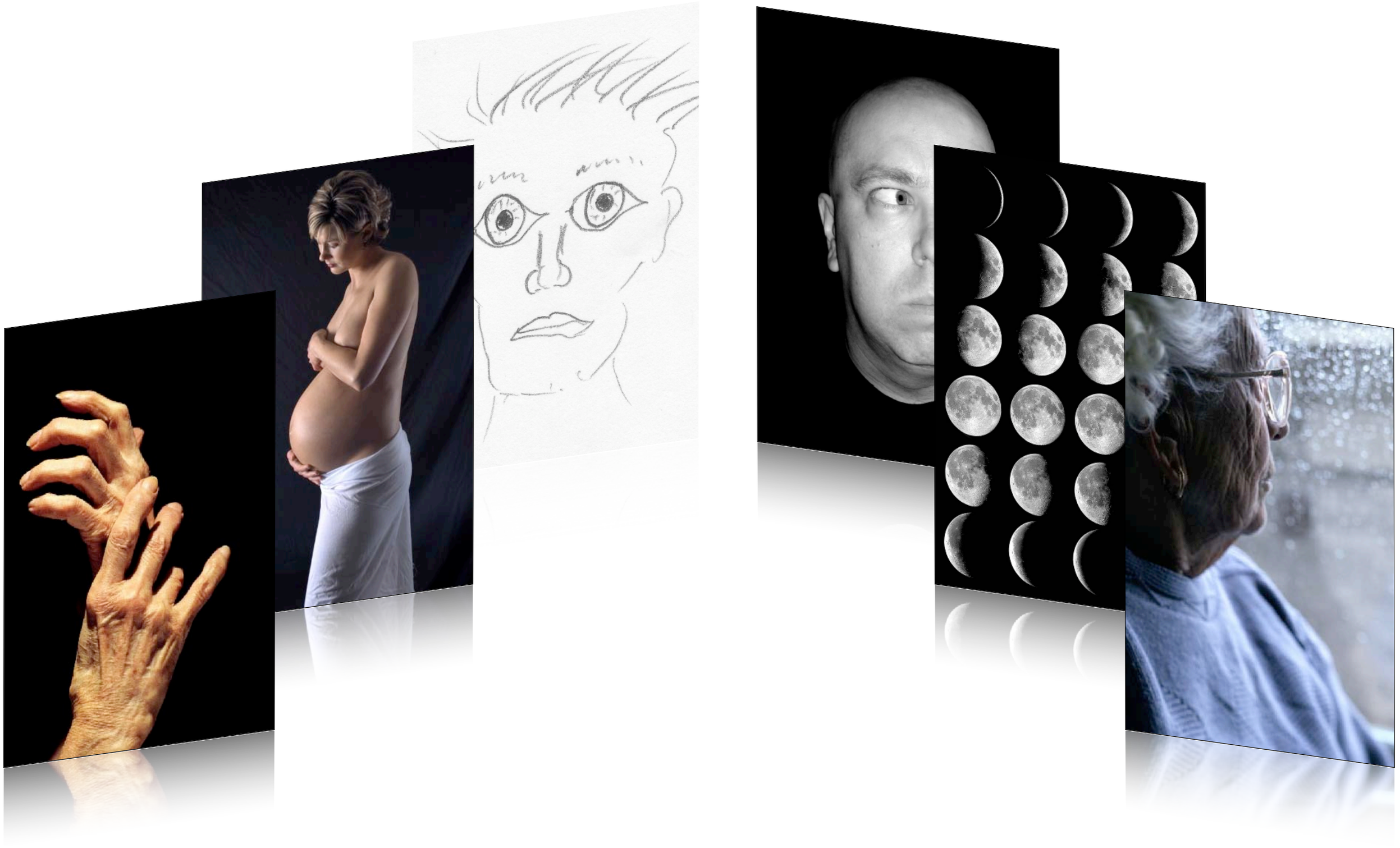


Redelmeier & Tversky (1996)

Abell & Greenspan (1979)

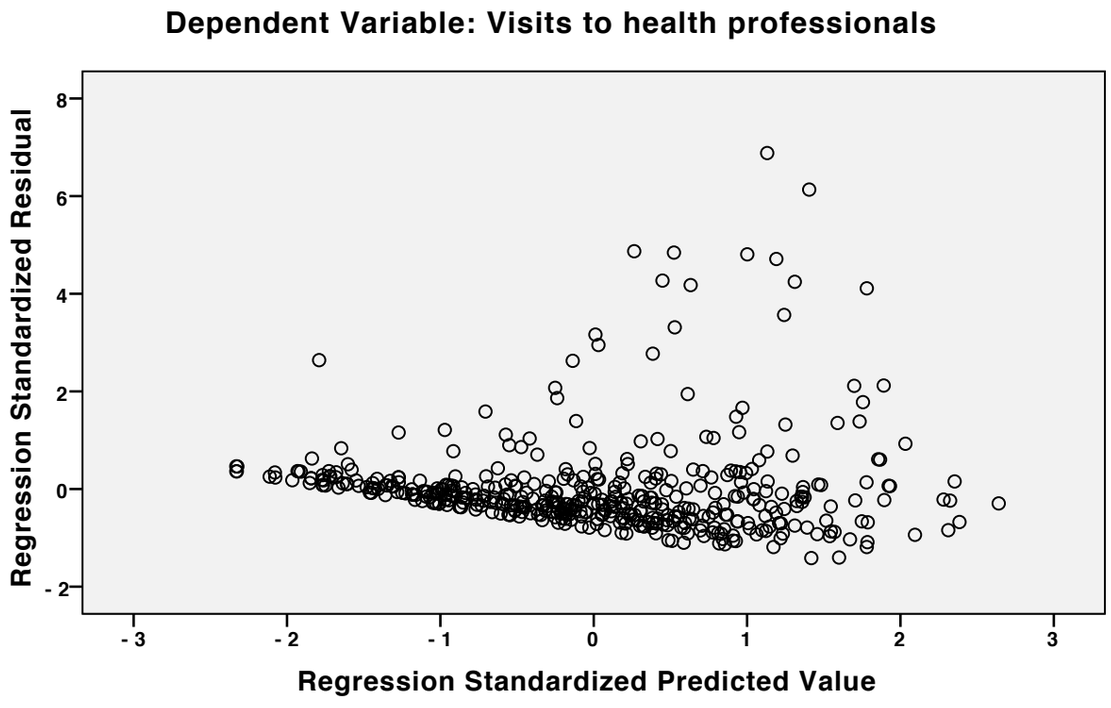
Chapman & Chapman (1969)

Hastorf & Cantril (1954)



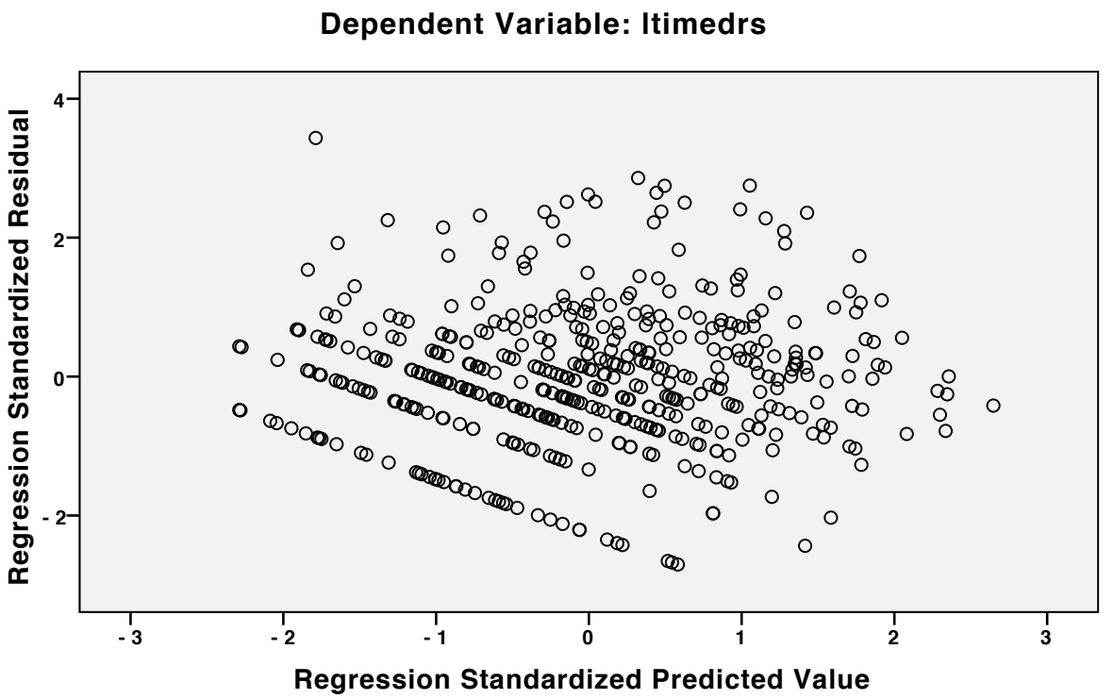
Homoscedasticity

The variability in scores in one variable is the same at all values of the other variable.



Residuals plot of RESID against PRED is shaped like a fan indicates heteroscedascity. If so, the scatterplot of the variables is pear shaped.

Remedy: Transformations to reduce skew (see T&F Ch 4).



Multicollinearity/singularity

There is no reason to be concerned if the predictor variables are moderately to highly correlated since multiple regression was ‘designed’ to analyse such data. Only if predictors are very highly intercorrelated (near perfect) do they cause difficulties with the estimation of the regression coefficients.

Inspection of the correlation matrix of predictors for very high (> 0.9) correlations. To test for combinations of variables being highly correlated with another variable, (i.e. multi-collinearity) the TOLERENCE is calculated. These are (1 - squared multiple correlation) and are calculated from a multiple regression in which each predictor variable is predicted from all other predictors in turn.

Remedy: If the tolerence is very low (approx < .01), then dropping a variable may be necessary.

T&F results:

Nutshell version, Original data

- Overall relationship significant.

$$R^2 = 21.9\%, F(3, 461) = 43.0, p < .001$$

- Two significant predictors.

$$phyheal, \beta = .39, p < .001; sr^2 = 11.1\%$$

$$stress, \beta = .17, p < .001; sr^2 = 2.4\%$$

- Story in words...
  - Is it a sound interpretation?
  - Do perturbations in data influence the story?

T&F results:

Nutshell version, Transformed data

- Overall relationship significant.

$$R^2 = 37.7\%, F(3, 461) = 92.9, p < .001$$

- Two significant predictors.

$$phyheal, \beta = .52, p < .001; sr^2 = 19.2\%$$

$$stress, \beta = .19, p < .001; sr^2 = 2.95\%$$

- Largish increase in R^2. Same predictors important.
  - Why? Large residuals reduced through transformations
  - Did it change the story?

Summary

- Appropriate remedies for “fixing” problems are sample specific.
  - That is, in another sample from the same population, the same distributional problems or outliers may not be found. Not only that, deleting cases and/or transforming variables makes the results and interpretation of an analysis even more sample specific.
- The topic is not as clear as T&F suggest.
- Data analysis is a dynamic process demanding the exercise of professional judgement.