UNIVERSITY OF BERGEN

Correlation and regression analysis

Sebastian Jentschke

UNIVERSITY OF BERGEN

Agenda

- introduction
- typical research questions, IV characteristics and limitations
- assumptions and requirements
- fundamental equations: do-it-yourself
- major types
- some important issues

Categorical vs. continuous vars.

- categorical variables contain a limited number of steps (e.g., male – female, experimentally manipulated or not)
- continuous variables have a (theoretically unlimited) number of steps (e.g., body height, weight, IQ)
- ANOVA (next session) is for categorical predictors, Correlation and regression analyses (this session) is for continuous predictors

Categorical vs. continuous vars.

		Dependent variable	
		Categorical	Continuous
Independent variable	Categorical	X^2 test (chi-squared)	***t-test*** **ANOVA**
	Continuous	Logistic regression	***Correlation*** **Linear regression**

Relation vs. difference hypotheses

- relation hypotheses explore whether there is a relation between one (or more) independent and a dependent variable
- difference hypotheses explore whether there is a difference between the steps of one (or more) independent and a dependent variable
- the distinction between IV and DV is blurred for relation hypotheses
  → causality can only be inferred if the independent variable was experimentally manipulated

Correlation and regression

- correlation: measure **size and direction of a linear relationship** of two variables (with the squared correlation as strength of association – explained variance)
- regression: **predict** one variable from one (or many) other (minimizing the squared distance between data points and a regression line)

$Y' = A/B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k$ $(y' = a + bx)$

$R = r_{YY'}$ (r_{xy})

Correlation and regression

- when calculating correlation (r) and regression coefficients (B), both use the covariance between IV and DV as **numerator**; but the correlation uses the variance of both IV and DV, the regression only the variance of the IV as **denominator**

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \qquad B = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

Questions?
Comments?

Correlation and regression

regression techniques:

- standard, sequential (hierarchical), statistical (stepwise)

typical research questions for using regression analysis:

- investigate a relationship between a DV and several IV
- investigate a relationship between one DV and some IVs with the effect of other IVs statistically eliminated
- compare the ability of several competing sets of IVs to predict a DV
- (ANOVA as a special case with dichotomous IVs; Ch. 5.6.5)

Correlation and regression

changing IVs:

- squaring IVs (or raising to higher power) to explore curvilinear relationships
- creating a cross-product of two (or more) IVs to explore interaction effects

predicting scores for members of a new sample:

- regression coefficients (B) can be applied to new samples
- generalizability should be checked with cross-validation (e.g., 50/50, 80/20 or boot-strapping)

Correlation and regression

limitations:

- implied causality
- theoretical assumptions (or lack of) regd. inclusion of variables
  theoretical: if the goal is the manipulation of a DV, include some IVs that can be manipulated as well as some who can't
  practical: include «cheaply obtained» IVs (SSB)
  statistical: IVs should correlate strongly with the DV but weak with other IVs (goal: predict the DV with as few as possible IVs); remove IVs that degrade prediction (check residuals)
  chose IVs with a high reliability

Correlation and regression

ratio of cases to IVs (m = IVs):

$N \geq 50 + 8m$ for multiple correlation (standard / hierarchical)
$N \geq 40m$ for multiple correlation (stepwise)
$N \geq 104 + m$ for individual predictors
(assuming $\alpha = .05$, $\beta = .20$ and medium effect size;
higher numbers if DV is skewed, small effect size is anticipated or substantial measurement error is expected)
$N \geq (8 / f^2) + (m - 1)$ (f = .02, .15, .35 for small, medium, large eff.)

strategies for insufficient N: exclude IVs, create composite meas.

Questions?
Comments?

Conditions for parametric tests

absence of multicollinearity and singularity:

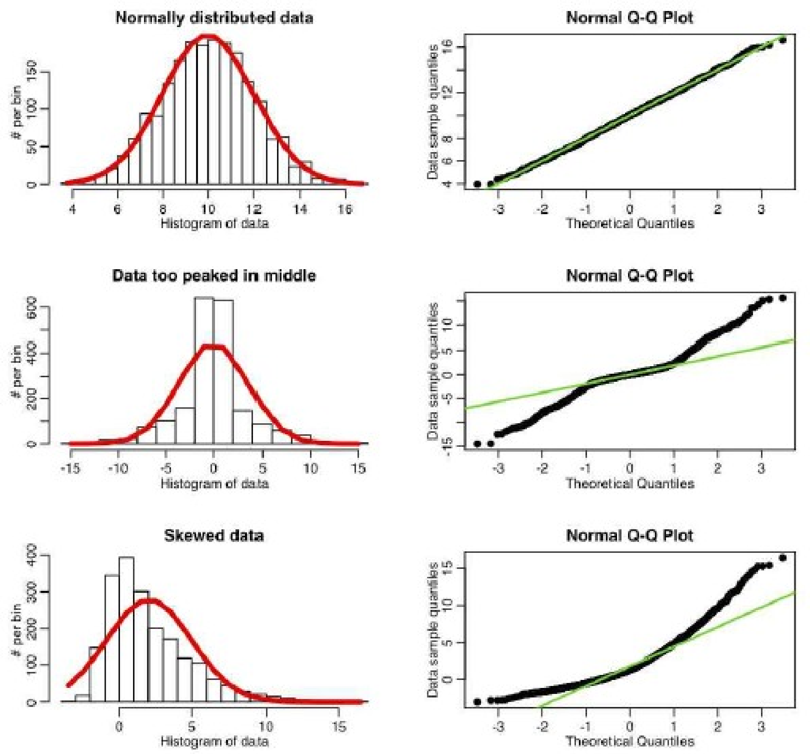
- regression is impossible if IVs are singular (i.e., a linear combination of other IVs) or unstable if they are multicollinear
- screening through detection of high R^2s when IVs are (in turn) predicted using other IVs
- variable removal should consider reliability and cost of acquisition

Conditions for parametric tests

- conditions for using parametric tests (such as correlation, regression, t-test, ANOVA)
- if one of these conditions is violated, non-parametric tests have to be used
- robustness against violation of certain assumptions (relatively robust against deviation from **normality**; deviations from **linearity** and **homoscedacity** do not invalidate an analysis but weaken it)

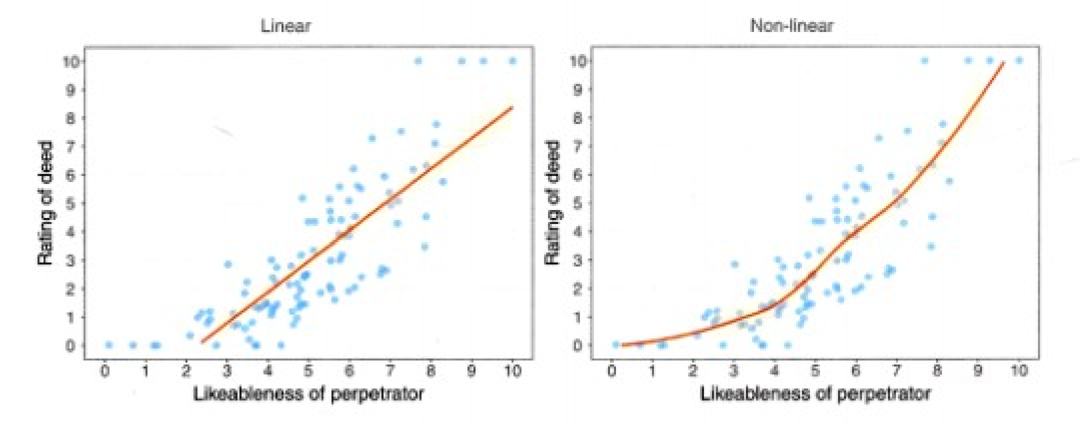
Conditions for parametric tests

- normality and possible causes for normality violations



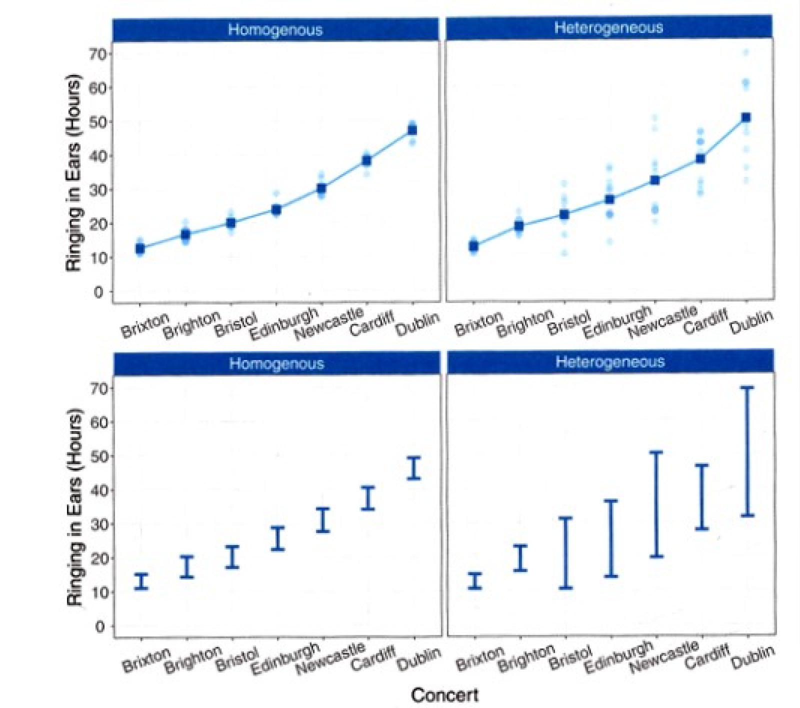
Conditions for parametric tests

- linearity (non-linear models are available, but not introduced here)

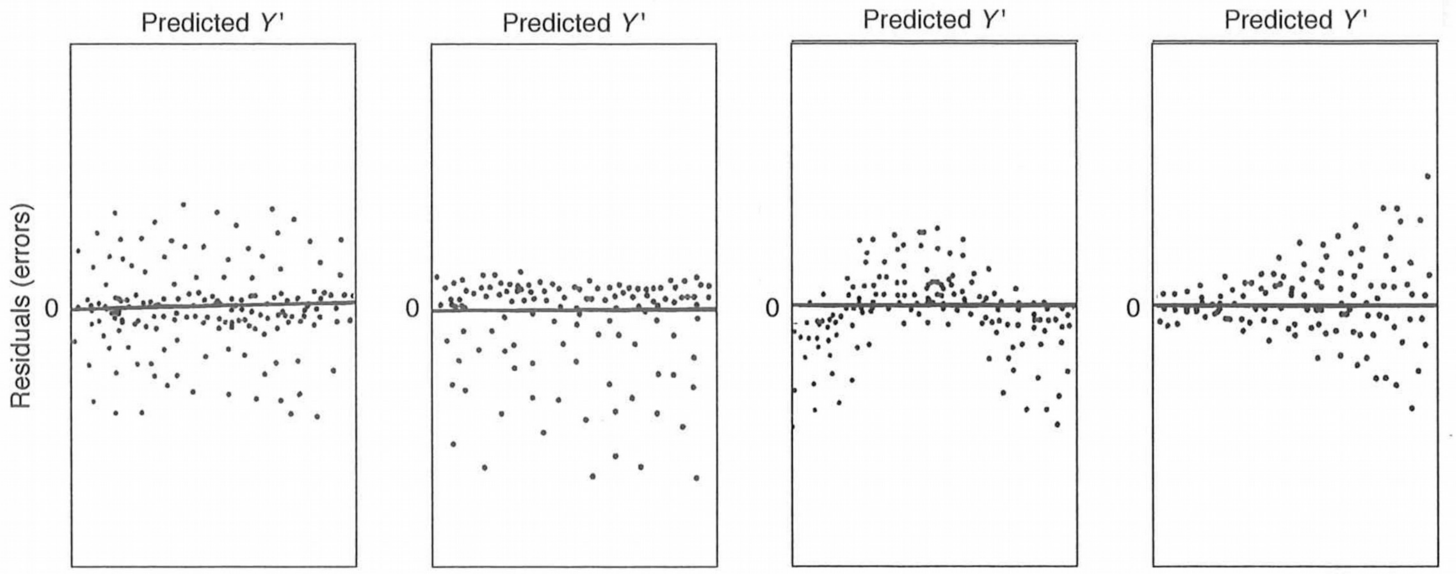


Conditions for parametric tests

- homogeneity of variance = homoscedasticity (heteroscedacity can be counteracted by using generalized least square regression where the DV is weighed by the IV that produces the hetero-scedacity)



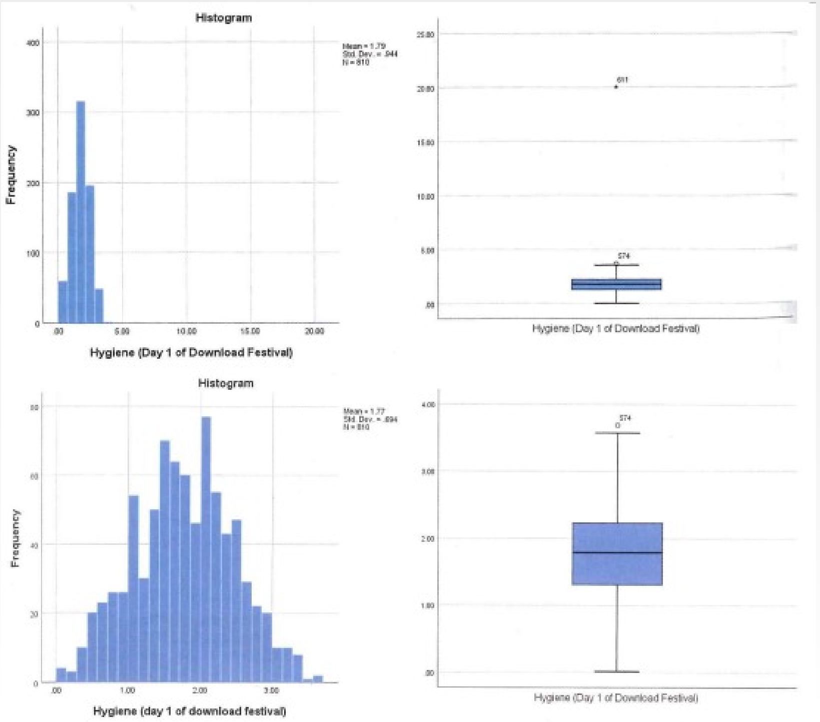
Conditions for parametric tests



assumptions met | failure of normality | non-linearity | heteroscedacity

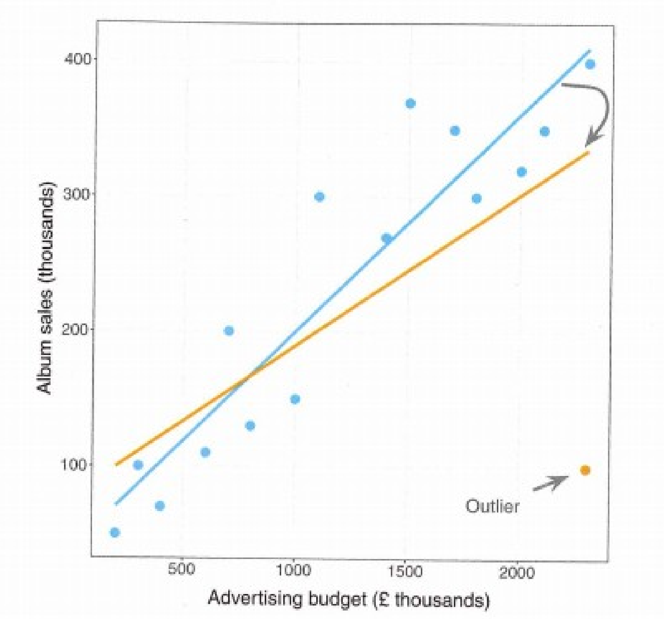
Conditions for parametric tests

- consequences of not removing out-liers on the skew-ness (and in con-sequence the nor-mality) of a distri-bution



Conditions for parametric tests

- consequences of not removing out-liers on the slope of a correlation / regression



Conditions for parametric tests

strategies for removing outliers:

- univariate – SPSS FREQUENCIES (box plots; for $N < 1000 \rightarrow p = .001 \rightarrow z = \pm 3.3$)
- multivariate: SPSS REGRESSION (Save → Distances → Mahalanobis; calculate "SIG.CHISQ(MAH_1,3)" and exclude $p < .001$)

Questions?
Comments?

General linear model

- Parameter estimation: Minimize the squared error
- $y = b_0 + b_1 \cdot x_1 + \ldots + b_n \cdot x_n + e$

  $Y = BX + E$

Y, y	=	dependent variable
X, $[x_1 \ldots x_n]$	=	predictor variable
B, $[b_0 \ldots b_n]$	=	predictor weights
		(b_0: intercept; $b_1 \ldots b_n$: slope)
E, [e]	=	error term

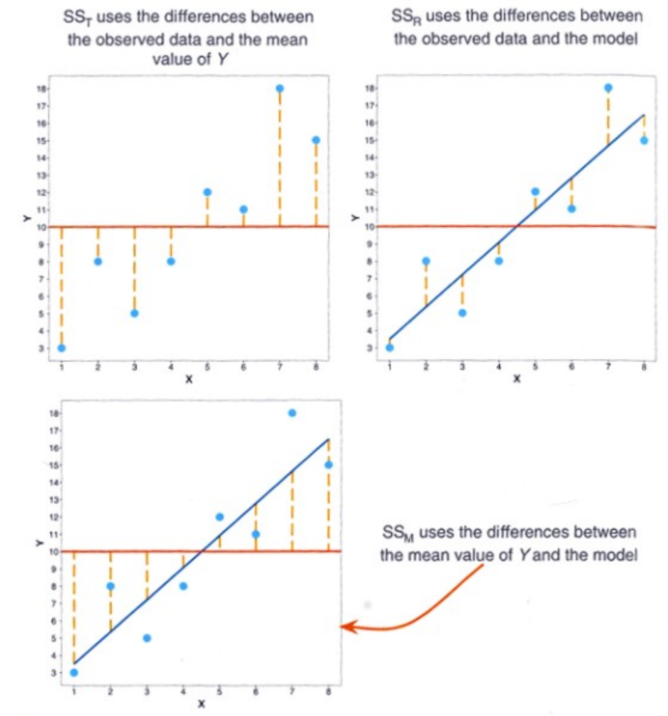


Figure 9.5 Diagram showing from where the sums of squares derive

General linear model

- $y = b_0 + b_1 \cdot x_1 + \ldots + b_n \cdot x_n + e$

  $Y = BX + E$

  Y, y = dependent variable

  X, $[x_1 \ldots x_n]$ = predictor variable [0, 1]

  B, $[b_0 \ldots b_n]$ = predictor weights

  [group mean - sample mean]

  E, [e] = error term

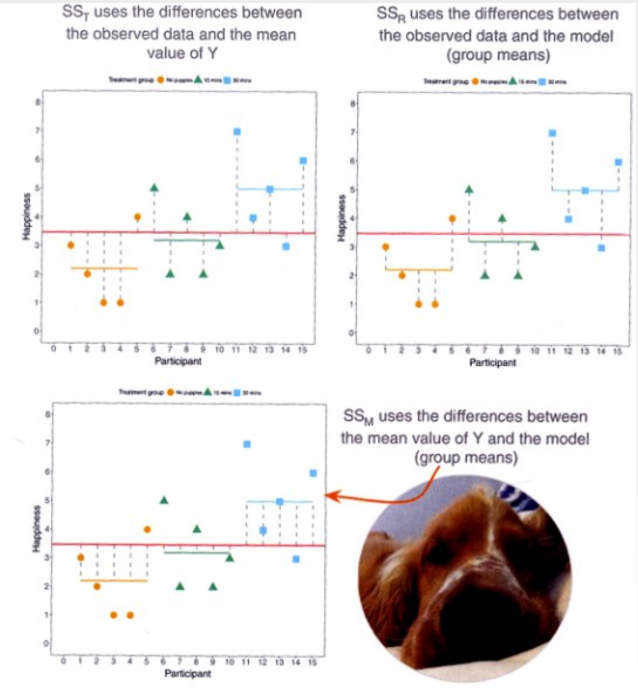


Figure 12.4 Graphical representation of the different sums of squares when comparing several means using a linear model. Also a picture of Ramsey as a puppy. Tufte would call him chartjunk, but I call him my adorable, crazy, spaniel

Fundamental equations/calculations

- PREREQUISITES FOR COMPARING TWO VARIABLES?
- WHAT WOULD LEAD TO AN PERFECT POSITIVE CORRELATION (r = 1.00) AND WHAT WOULD LEAD TO A PERFECT NEGATIVE CORRELATION (r = -1.00)?

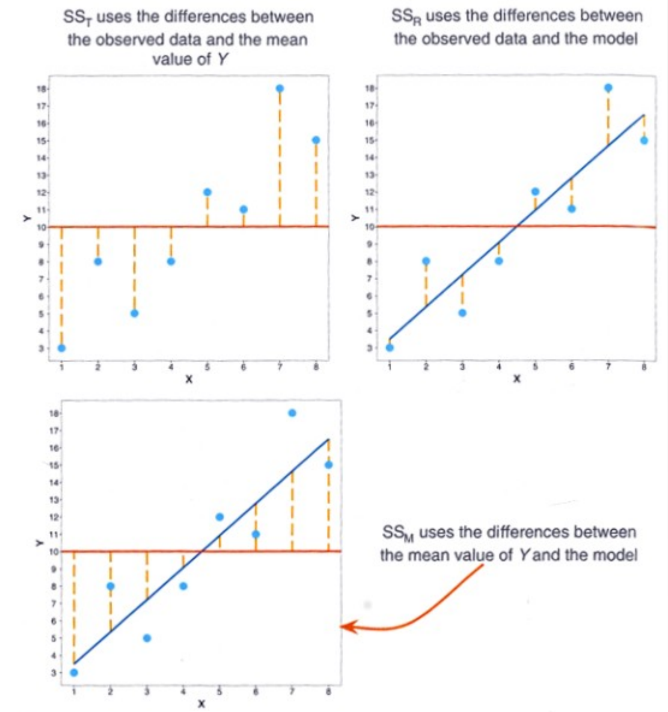


Figure 9.5 Diagram showing from where the sums of squares derive

Fundamental equations/calculations

- Correlation: hands-on
- z-standardize both variables (use popul. std. dev [STDEV.P])
- for each participant multiply these z-standardized values
- average these individual multi-plication products

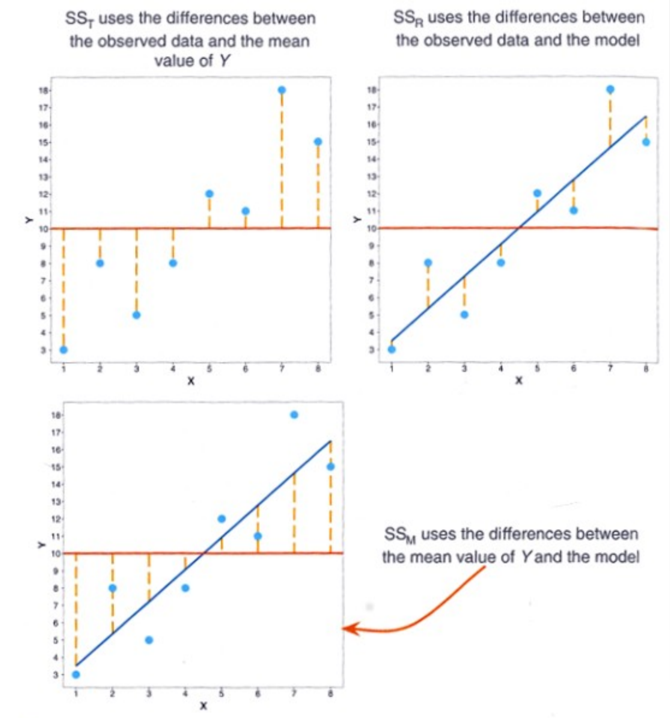


Figure 9.5 Diagram showing from where the sums of squares derive

Fundamental equations/calculations

using the example in Ch. 5.4 (pp. 165-172) in Octave / MATLAB:

```
% define independent and dependent variables and calculate correlations among them
IV = [14, 19, 19; 11, 11, 8; 8, 10, 14; 13, 5, 10; 10, 9, 8; 10, 7, 9]
DV = [18; 9; 8; 8; 5; 12]
R = corrcoef([IV, DV])
RII = R(1:3, 1:3)
RID = R(1:3, 4)
% determine the standaridized B-weights multiple correlation
BS = inv(RII) * RID
R2 = RID' * BS
% determine the unstandardized regression coefficients
BU = diag(BS * (std(DV) ./ std(IV)))
A = mean(DV) - mean(IV) * BU
% calculate the predicted DVs
DVP = IV * BU + A
% display your results
plot(DV, DVP, "r*"); xlim([0, 20]); ylim([0, 20]); line([0, 20], [0, 20]);
plot(DVP, DV - DVP, "b*"); xlim([0, 20]); ylim([-10, 10]); line([0, 20], [0, 0]);
% create an "artifical" new student and use this for prediction
[12, 14, 15] * BU + A
```

$$\mathbf{B}_i = \mathbf{R}_{ii}^{-1}\mathbf{R}_{iy}$$

$$B_i = \beta_i\left(\frac{S_Y}{S_i}\right)$$

Fundamental equations/calculations

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT COMPR
  /METHOD=ENTER QUAL GRADE MOTIV
  /SAVE MAHAL.
```

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	MOTIV, GRADE, QUAL[b]	.	Enter

a. Dependent Variable: COMPR

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,838[a]	,702	,256	3,896

a. Predictors: (Constant), MOTIV, GRADE, QUAL

b. Dependent Variable: COMPR

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71,640	3	23,880	1,573	,411[b]
	Residual	30,360	2	15,180		
	Total	102,000	5			

a. Dependent Variable: COMPR

b. Predictors: (Constant), MOTIV, GRADE, QUAL

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4,722	9,066		-,521	,654
	QUAL	,272	,589	,291	,462	,690
	GRADE	,416	,646	,402	,644	,586
	MOTIV	,658	,872	,319	,755	,529

a. Dependent Variable: COMPR

using the example in Ch. 5.4 (p. 165-172) in SPSS:

Questions?
Comments?

Major types of multiple regression

three analytic strategies:

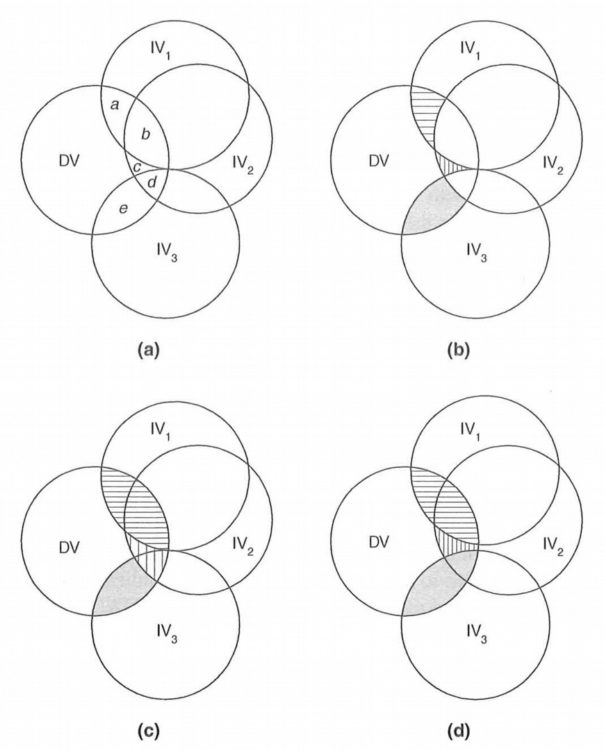
- standard
- sequential / hierarchical
- statistical / stepwise

differ in how the IVs contribution to the prediction is weighed

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-4,722	9,066		-,521	,654
	QUAL	,272	,589	,291	,462	,690
	GRADE	,416	,646	,402	,644	,586
	MOTIV	,658	,872	,319	,755	,529

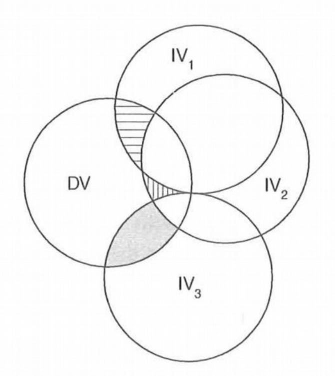
a. Dependent Variable: COMPR



Major types of multiple regression

standard regression:

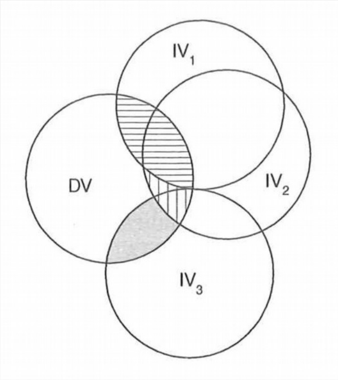
- enters all IVs at once in the equation
- only unique contributions are considered (may make the contribution of a variable look unimportant due to the correlation with other IVs, e.g., IV_2)



Major types of multiple regression

sequential / hierarchical regression:

- enters IVs in an order specified
  can be entered separately or in blocks according to logical or theoretical conside-rations, e.g. experimentally manipulated variables before nuisance variables, the other way round, or comparing different sets

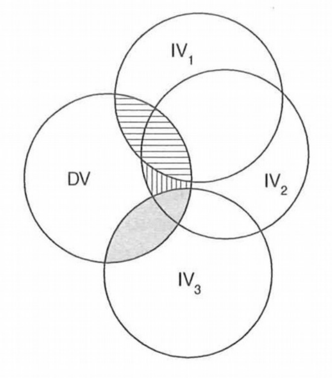


- additional contribution of each IV is considered

Major types of multiple regression

statistical / stepwise regression:

- controversial; order of entry (or possibly removal) specified by statistical criteria
- three versions: forward selection, backward deletion, stepwise regression
- tendency for overfitting → requires large and representative sample; should be cross-validated (R^2 discrepancies indicate lack of generalizability)



Major types of multiple regression

choosing regression strategies:

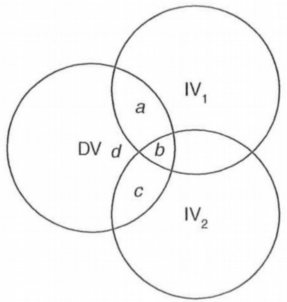
- ***standard***: simply assess relationships (atheoretical)
  what is the size of the overall relationship between IVs and DV?
- ***sequential***: testing theoretical assumptions or explicit hypotheses (IVs can be weighted by importance)
  how much does each variable uniquely contribute?
- ***statistical***: model-building (explorative, generating hypotheses) rather than model-testing
  can be very misleading unless based on large, representative samples
  can be helpful for identifiying multicollinear / singular vars.
  what is the best linear combination of variables / best prediction?

Questions?
Comments?

Important issues

Variable contribution / importance

- straightforward if IVs are uncorrel.
- relationship between correlation, partial and semipartial correlation (*SPSS Regression – Statistics – Part and partial corr.*)
- sum of semipartial corr. is smaller than R^2 if IVs are correlated



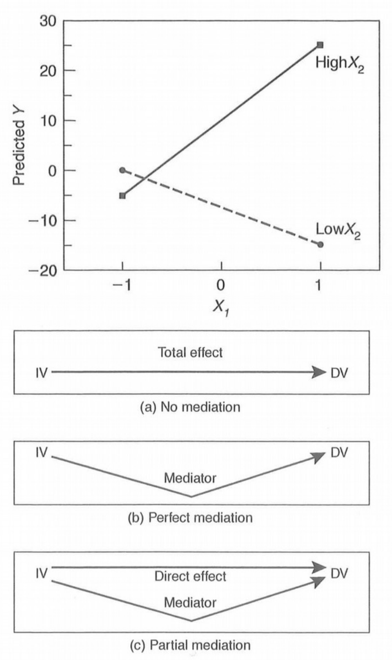
		Standard Multiple	Sequential
r_i^2	IV_1	$(a+b)/(a+b+c+d)$	$(a+b)/(a+b+c+d)$
	IV_2	$(c+b)/(a+b+c+d)$	$(c+b)/(a+b+c+d)$
sr_i^2	IV_1	$a/(a+b+c+d)$	$(a+b)/(a+b+c+d)$
	IV_2	$c/(a+b+c+d)$	$c/(a+b+c+d)$
pr_i^2	IV_1	$a/(a+d)$	$(a+b)/(a+b+d)$
	IV_2	$c/(c+d)$	$c/(c+d)$

Important issues

suppressor variables:

- IV that suppresses irrelevant variance by virtue of its correlation with other IVs
  (e.g., a questionaire and a measure of test-taking ability; the questionaire confounds the actual construct with test-taking skills and test-taking ability removes this [irrelevant] confundation)
- can be identified by the patterns of regr. coeffic. β and the correlations between IVs and DV:
  (1) $\beta \neq 0$; (2) $abs(r_{IV\text{-}DV}) < \beta$ *or* $sign(r_{IV\text{-}DV}) \neq sign(\beta)$

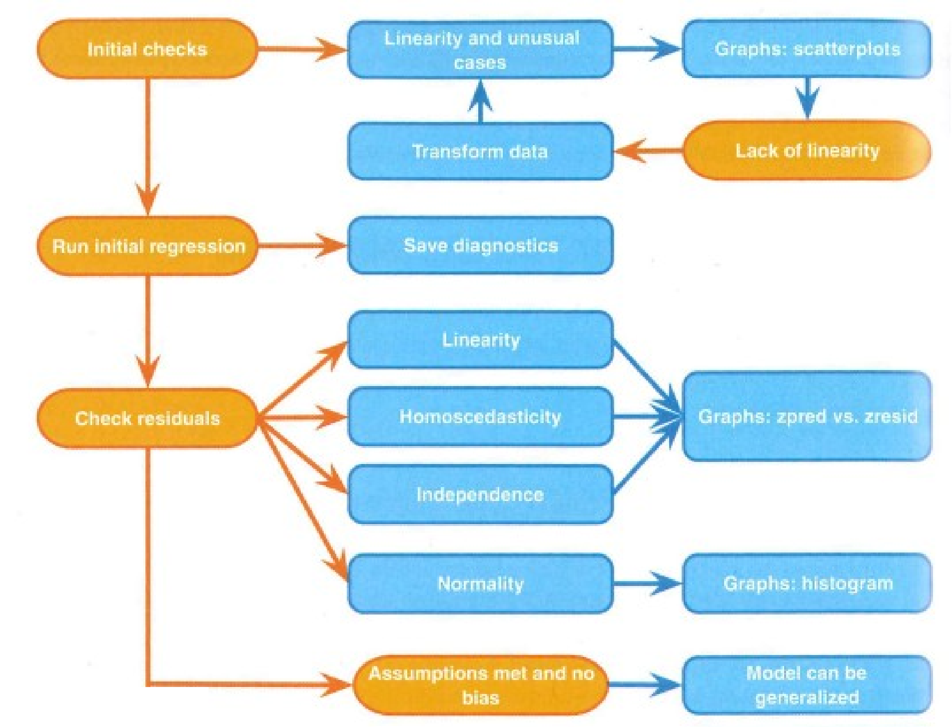
Important issues



mediation:

- causal **sequence** of three / more vars.
  (e.g., a relation between gender and visits to health care professionals mediated / driven by a personality aspect [«caused» by gender])
- variable is a mediatior if: sign. relat. IV ↔ DV and IV ↔ Md, Md (IV partialed out) ↔ DV, if mediator incl.: IV ↔ DV diminished
- decompose direct and mediation effects

Important issues



Summary

- typical research questions
- assumptions and requirements
- fundamental equations: do-it-yourself
- regression types and when to use them
- issues to keep in mind

Literature

Tabachnik, B. G., Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). New York, NY: Pearson. (Ch. 5)

Field, A. (2017). *Discovering Statistics Using IBM SPSS Statistics*. London, UK: Sage Publications Ltd.

Thank you for your interest!

UNIVERSITAS BERGENSIS

UNIVERSITY OF BERGEN

Eco-Lighthouse®