Lecture 8

Empirical Research Methods
IN4304

Quantitative Data analysis II – Relations

Previous lecture

Idea behind hypothesis testing

- Is the difference between two samples a reflection of the difference of two different populations or simply caused by sampling error?

Conceptual models with specific statistical tests

- Different tests depending on: level, number of independent and dependent variables, between or within subjects set up

Statistical tests in SPSS

- **See the recording of this lecture**

Today

- Pearson product moment correlation
- Spearman Correlation
- Partial Correlation
- Linear Regression analysis
- Multiple regression

Learning outcomes of this lecture

After today's lecture you should be able :

- to explain the idea behind (partial) correlation
- to link conceptual models with regression model
- to understand output of regression analysis
- to perform regression analysis in SPSS

Statistical test and Scales of measurement

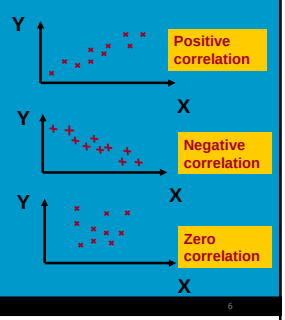
Aim	Samples	Data Level	Tests
Find difference	Single sample	Nominal	Binominal test, χ^2 goodness-of-fit
		Ordinal	Wilcoxon signed-ranks test
		Interval / Ratio	*z*-test, One-Sample *t*-test
	Independent	Nominal	Fisher-Exact test, χ^2
		Ordinal	Mann-Whitney *U* test
		Interval / Ratio	*z*-test, two independent sample *t*-test, ANOVA, MANOVA
	Dependent	Nominal	McNemar
		Ordinal	Sign Test, Wilcoxon matched-pairs signed-rank test, Friedman Test
		Interval / Ratio	Paired-Sample *t*-test, repeated measures ANOVA, MANOVA
Find relation		Nominal	Cramér's V, phi
		Ordinal	Kendall, Spearman correlation
		Interval / Ratio	Pearson product moment correlation, regression analysis

Correlation

Correlation is the degree of relationship between two variables.

We are interested in two properties of the relationship:

1. Its **direction**
2. Its **strength**

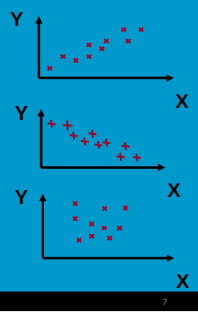


Correlation - Direction

Positive correlation: An increase in X is associated with an increase in Y.

Negative correlation: an increase in X is associated with a decrease in Y.

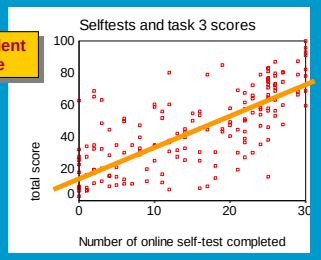
Zero correlation: an increase or decrease in X is neither associated with an increase or decrease in Y.



TUDelft

Scatter diagram – strength of correlation

Dependent Variable



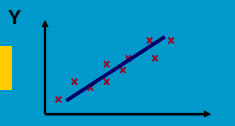
Independent Variable

TUDelft

Correlation – Strength of correlation

The **strength** of a correlation can be seen as an expression of how well data points **fit** on a line that represents the relationship between two variables.

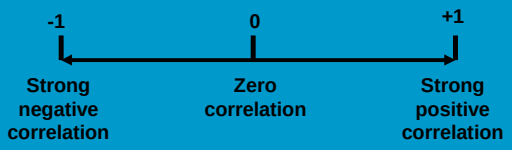
"Stronger" correlation



"Weaker" correlation

TUDelft

Correlation coefficient (r)



Correlation coefficient is a numeric measure of the strength and direction of a linear relationship between two variables.

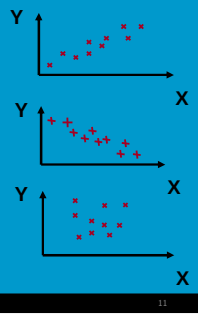
TUDelft

Correlation coefficient (r)

Positive correlation: $0 < r \leq +1$

Negative correlation: $-1 \leq r < 0$

Zero correlation: $r \approx 0$



TUDelft

Correlation coefficient (r)

Pearson Product-moment correlation coefficient (r)

$$r = \frac{\sum (z_x \times z_y)}{n-1}$$

Standard score, z-score, is the deviation of a data point from the mean in standard deviation units

Transforming data into z-scores makes it possible to compare them, because they are expressed in the same unit.

$$z = \frac{x - \bar{x}}{s}$$

TUDelft

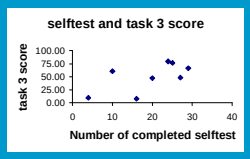
Correlation coefficient (r)

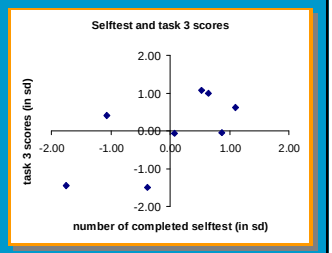
	x: Selftest	y: task 3 score	z_x	z_y
	16	7.50	-0.38	-1.50
	24	79.00	0.53	1.07
	20	47.30	0.07	-0.07
	27	47.68	0.87	-0.06
	29	66.25	1.10	0.61
	10	60.50	-1.07	0.40
	4	9.00	-1.75	-1.45
	25	76.84	0.64	0.99
Sum	**155**	**394.07**		
Mean	19.38	49.26		
SD	8.78	27.85		
n	8			

$$z = \frac{x - \bar{x}}{s} = \frac{16 - 19.38}{8.78} = -0.38$$

TUDelft

Correlation coefficient (r)

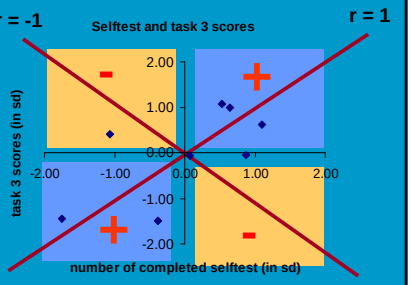




TUDelft

Correlation coefficient (r)

$$r = \frac{\sum (z_x \times z_y)}{n-1}$$



TUDelft

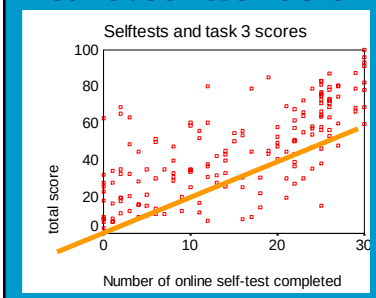
Correlation coefficient (r)

$$r = \frac{\sum (z_x \times z_y)}{n-1} = \frac{4.49}{8-1} = 0.64$$

	x: Selftest	y: task 3 score	z_x	z_y	z_x*z_y
	16	7.50	-0.38	-1.50	0.58
	24	79.00	0.53	1.07	0.56
	20	47.30	0.07	-0.07	-0.01
	27	47.68	0.87	-0.06	-0.05
	29	66.25	1.10	0.61	0.67
	10	60.50	-1.07	0.40	-0.43
	4	9.00	-1.75	-1.45	2.53
	25	76.84	0.64	0.99	0.63
Sum	**155**	**394.07**			**4.49**
Mean	19.38	49.26			

TUDelft

Correlation coefficient



r = 0.71

This means that an increase in the number of self-tests completed is associated with an increase in the total score for task 3

TUDelft

Correlation coefficient (r)

$$r = \frac{\sum (z_x \times z_y)}{n-1} = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

Social Science
r = 0.1 -> small effect
r = 0.3 -> medium size effect
r = 0.5 -> Large effect
(Cohen, 1992)

TUDelft

Pearson Correlation – Conceptual model

- **Conceptual model**
- Example
- Assumptions
- Interpretation outcome
- Demo

Attitude towards mobile phone design → Index score

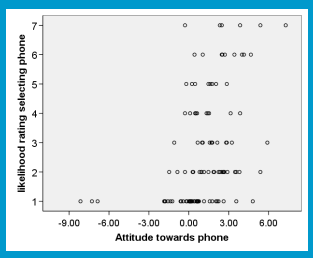
Intention to select phone → Rating from 1 to 7

Pearson Correlation - example

- Conceptual model
- **Example**
- Assumptions
- Interpretation outcome
- Demo

Pearson Correlation - example

- Conceptual model
- **Example**
- Assumptions
- Interpretation outcome
- Demo



Pearson Correlation - assumptions

- Conceptual model
- Example
- **Assumptions**
- Interpretation outcome
- Demo

1. Independence of observations
2. Interval data
3. Linear relationship
4. homoscedasticity (the same level of relationship throughout the range of the independent variable)

Pearson Correlation– interpretation outcomes

- Conceptual model
- Example
- Assumptions
- **Interpretation outcome**
- Demo

Correlations

		Attitude towards phone	likelihood rating selecting phone
Attitude towards phone	Pearson Correlation	1	.420**
	Sig. (2-tailed)		.000
	N	114	112
likelihood rating selecting phone	Pearson Correlation	.420**	1
	Sig. (2-tailed)	.000	
	N	112	147

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

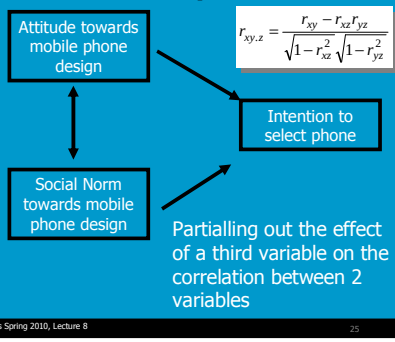
Pearson Correlation – SPSS Demo

- Conceptual model
- Example
- Assumptions
- Interpretation outcome
- **Demo**

Partial Correlation – Conceptual model

- **Conceptual model**
- Example
- Assumptions
- Interpretation outcome
- Demo

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$



Partialling out the effect of a third variable on the correlation between 2 variables

Partial Correlation - example

- Conceptual model
- **Example**
- Assumptions
- Interpretation outcome
- Demo

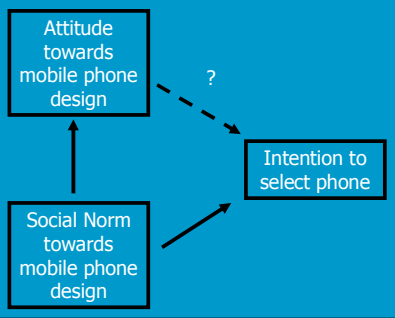
Correlations

		Attitude towards phone	Social Norm towards phone	likelihood rating selecting phone
Attitude towards phone	Pearson Correlation	1	.856**	.420**
	Sig. (2-tailed)		.000	.000
	N	114	108	112
Social Norm towards phone	Pearson Correlation	.856**	1	.391**
	Sig. (2-tailed)	.000		.000
	N	108	111	109
likelihood rating selecting phone	Pearson Correlation	.420**	.391**	1
	Sig. (2-tailed)	.000	.000	
	N	112	109	147

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

Partial Correlation - example

- Conceptual model
- **Example**
- Assumptions
- Interpretation outcome
- Demo



Partial Correlation - assumptions

- Conceptual model
- Example
- **Assumptions**
- Interpretation outcome
- Demo

1. See assumptions Pearson Correlation

Partial Correlation– interpretation outcomes

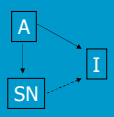
- Conceptual model
- Example
- Assumptions
- **Interpretation outcome**
- Demo

Correlations

Control Variables			Attitude towards phone	likelihood rating selecting phone
Social Norm towards phone	Attitude towards phone	Correlation	1.000	.188
		Significance (2-tailed)	.	.055
		df	0	103
	likelihood rating selecting phone	Correlation	.188	1.000
		Significance (2-tailed)	.055	.
		df	103	0

Partial Correlation – SPSS Demo

- Conceptual model
- Example
- Assumptions
- Interpretation outcome
- **Demo**



Spearman rank Correlation coefficient (r) for ordinal data

Usability rating, x	Frequency of use, y	Ranking of x	Ranking of y	Difference between ranking, d	d^2
Very low	5	1	2.5	-1.5	2.25
Low	2	2.5	1	1.5	2.25
Low	5	2.5	2.5	0	0
average	8	4	4	0	0
high	12	6	5	1	1
high	13	6	6	0	0
high	20	6	8	-2	4
very high	15	8	7	1	1
Total					**10.5**

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 10.5}{8(8^2-1)} = 0.875$$

Spearman Correlation – Conceptual model

- **Conceptual model**
- Example
- Assumptions
- Interpretation outcome
- Demo

Beauty of UI —— Usability of UI

Spearman Correlation - example

- Conceptual model
- **Example**
- Assumptions
- Interpretation outcome
- Demo

beauty of the interface * Usability of the interface Crosstabulation

Count

		Usability of the interface				Total
		very low	low	average	very high	
beauty of the interface	very ugly	0	1	0	0	1
	ugly	1	0	0	0	1
	avarage	0	1	0	1	2
	beautiful	0	0	1	1	2
	very beautiful	0	1	1	1	3
Total		1	3	2	3	9

Spearman Correlation - assumptions

- Conceptual model
- Example
- **Assumptions**
- Interpretation outcome
- Demo

1. Independence of observations
2. At least ordinal level

Spearman Correlation– interpretation outcomes

- Conceptual model
- Example
- Assumptions
- **Interpretation outcome**
- Demo

Correlations

			beauty of the interface	Usability of the interface
Spearman's rho	beauty of the interface	Correlation Coefficient	1.000	.440
		Sig. (2-tailed)	.	.236
		N	9	9
	Usability of the interface	Correlation Coefficient	.440	1.000
		Sig. (2-tailed)	.236	.
		N	9	9

Spearman Correlation – SPSS Demo

- Conceptual model
- Example
- Assumptions
- Interpretation outcome
- **Demo**

Why do we look at regression models?

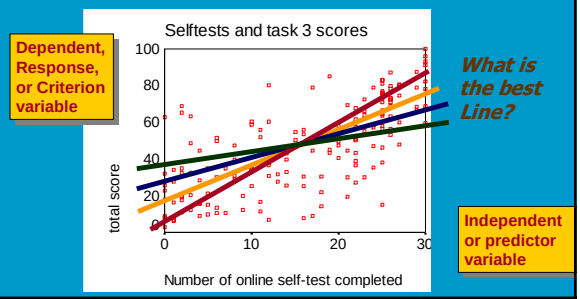
It helps us to model and examine the nature of the relation between two or more variables.

For example,

- What is the network speed when 10 or 50 PCs are connected to it?
- How much of the variation in the users' Acceptance of Information Technology can be explained by the Usability of an application, the Users training, and the level of IT support?
- What Webct mark for task 3 can I expect when I have completed 25 of the online self-tests?

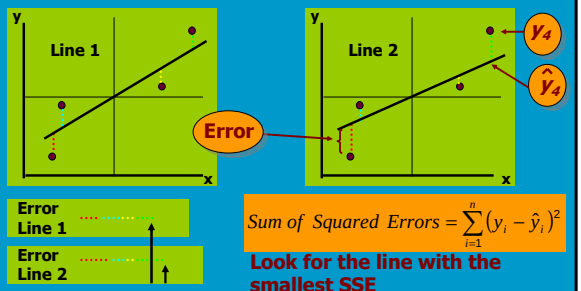
TUDelft

Fitting a straight line to represent the data



TUDelft

Least Squares Method

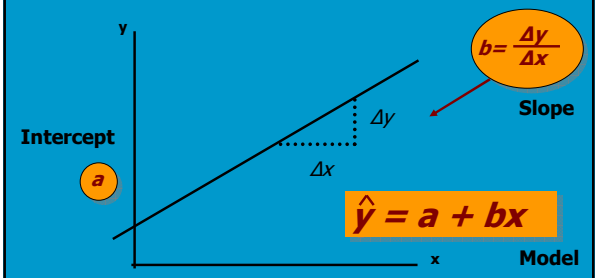


$$Sum\ of\ Squared\ Errors = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Look for the line with the smallest SSE

TUDelft

Simple Regression Model



TUDelft

Simple Regression Model

Table: Self-tests completed and overall score task 3

X: Selftest	Y: task 3 score	xy	x^2
16	7.50	120.00	256.00
24	79.00	1896.00	576.00
20	47.30	946.00	400.00
27	47.68	1287.36	729.00
29	66.25	1921.25	841.00
10	60.50	605.00	100.00
4	9.00	36.00	16.00
25	76.84	1921.00	625.00
155	**394.07**	**8732.61**	**3543.00**

Slope

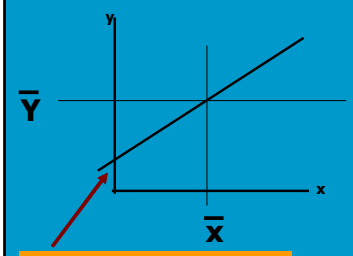
$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{8 \times 8732.61 - 155 \times 394.07}{8 \times 3543 - 155^2}$$

$$= 2.033$$

TUDelft

Simple Regression Model



The regression line always passes through the point ($\bar{x}$,$\bar{y}$)

$$\bar{y} = a + b\bar{x}$$

$$\equiv$$

$$a = \bar{y} - b\bar{x}$$

Intercept (a)

TUDelft

Simple Regression Model

Table: Self-test completed and overall score task 3

X: Selftest	Y: task 3 score	xy	x^2
16	7.50	120.00	256.00
24	79.00	1896.00	576.00
20	47.30	946.00	400.00
27	47.68	1287.36	729.00
29	66.25	1921.25	841.00
10	60.50	605.00	100.00
4	9.00	36.00	16.00
25	76.84	1921.00	625.00
155	**394.07**	**8732.61**	**3543.00**

Intercept

$$a = \frac{\sum y}{n} - b \times \frac{\sum x}{n}$$

$$= \frac{394.07}{8} - 2.033 \frac{155}{8}$$

$$= 9.87$$

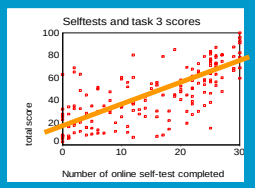
TU Delft

Simple Regression Model

$\hat{y} = 9.87 + 2.03x$

TU Delft

Class question



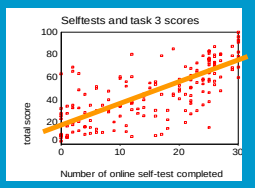
$\hat{y} = 18.20 + 1.76x$

*Predicted total score = 18.20 + 1.76*number of online self-test completed*

What is the predicted total score for task 3 for a student that completed 10 online self-tests?

TU Delft

Class question - Answer



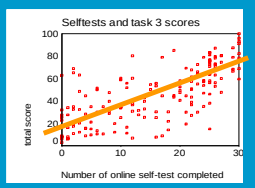
$\hat{y} = 18.20 + 1.76x$

*Predicted total score = 18.20 + 1.76*number of online self-test completed*

*the predicted total score = 18.20 + 1.76*10 = 35.8*

TU Delft

Class question



$\hat{y} = 18.20 + 1.76x$

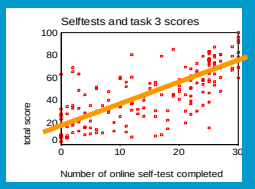
*Predicted total score = 18.20 + 1.76*number of online self-test completed*

1) In general, what is the impact of completing an online self-test on the predicted total score of task ?

2) What is the expected total task 3 score for a student who did not complete any self-tests ?

TU Delft

Class question - Answer



$\hat{y} = 18.20 + 1.76x$

*Predicted total score = 18.20 + 1.76*number of online self-test completed*

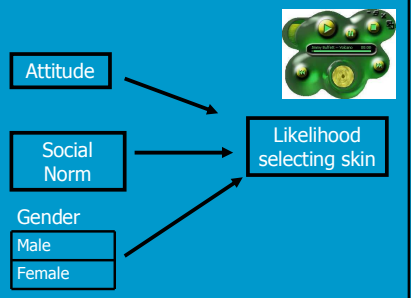
1) Completing 1 self-test will increase the task 3 total score with 1.76 points

2) If x=0 than predicted y = a = 18.20

TU Delft

Multiple Linear Regression – Conceptual model

- **Conceptual model**
- Assumptions
- Interpretation outcome
- Model selection
- Demo



Multiple Linear Regression - assumptions

- Conceptual model
- **Assumptions**
- Interpretation outcome
- Model selection
- Demo

1. Relation between independent variable(s) and dependent variable is linear
2. Errors are independent
3. Error is normal distributed
4. Error is distributed with constant variance

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + e_i$$

$$Sum\ of\ Squared\ Errors = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Multiple Linear Regression – interpretation outcomes – Model summary

- Conceptual model
- Assumptions
- **Interpretation outcome**
- Model selection
- Demo

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.459[a]	.210	.195	1.383

a. Predictors: (Constant), Social Norm towards skin, Attitude towards skin

$$s_{est} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{df}}$$

where by

$df = n - 1 - \#independent\ variables$

R - correlation between predicted and observed
R² - Coefficient of Determination shows the proportion of variation in the dependent variable that can be explained by the model
Adj. R² - Corrected for the number of Independent variables
Std Error of the Estimate – the accuracy of the prediction

Multiple Linear Regression – interpretation outcomes – ANOVA

- Conceptual model
- Assumptions
- **Interpretation outcome**
- Model selection
- Demo

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	54.033	2	27.016	14.119	.000[a]
	Residual	202.830	106	1.913		
	Total	256.862	108			

a. Predictors: (Constant), Social Norm towards skin, Attitude towards skin
b. Dependent Variable: likelihood to select skin

Multiple Linear Regression – interpretation outcomes – ANOVA

- Conceptual model
- Assumptions
- **Interpretation outcome**
- Model selection
- Demo

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	54.033	2	27.016	14.119	.000[a]
	Residual	202.830	106	1.913		
	Total	256.862	108			

a. Predictors: (Constant), Social Norm towards skin, Attitude towards skin
b. Dependent Variable: likelihood to select skin

Multiple Linear Regression – interpretation outcomes – Coefficient

- Conceptual model
- Assumptions
- **Interpretation outcome**
- Model selection
- Demo

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.358	.134		17.638	.000
	Attitude towards skin	.268	.117	.403	2.283	.024
	Social Norm towards skin	.052	.147	.062	.353	.725

a. Dependent Variable: likelihood to select skin

$$likelihood.select.skin_i = 2.36 + 0.27\,Attitude + 0.05\,SN + e_i$$

Beta – standardised coefficients, e.g. 1 sd rise in attitude there would be a 0.4 sd rise in skin selection likelihood

Multiple Linear Regression – interpretation outcomes – Collenearity

- Conceptual model
- Assumptions
- **Interpretation outcome**
- Model Selection
- Demo

Model		Unsta Coe	Sig.	Collinearity Statistics	
		B		Tolerance	VIF
1	(Constant)	2.35	.00		
	Attitude towards skin	.26	.02	.239	4.189
	Social Norm towards skin	.05	.72	.239	4.189

a. Dependent Variable: likelihood to select s

Collinearity -Correlation between Independent variables – limits the size of R, difficult to assess predictors, unstable predictor equation (B).
No strict rules, but problems when:
•Tolerance=1/VIF < 0.2
•Variance Inflation Factor > 10

Multiple Linear Regression– Model Selection

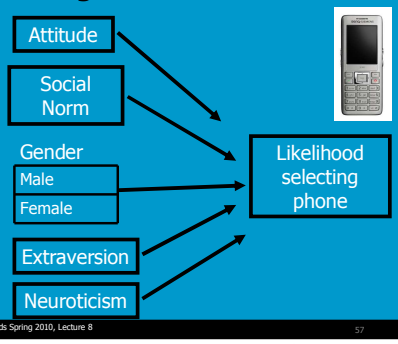
- Conceptual model
- Assumptions
- Interpretation outcome
- **Model selection**
- Demo

Enter – force all predictors into the model

Stepwise – computer selects least number of predictors to make an effective model. Continue to build model until new predictors will not make a significant contribution

Multiple Linear Regression– SPSS Demo

- Conceptual model
- Assumptions
- Interpretation outcome
- Model selection
- **Demo**



Summary

Correlation coefficient
Express strength and direction of relation
Pearson (interval level)
Spearman (ordinal level)

Partial correlation
Partialling out the effect of a third variable on the correlation between 2 variables

Regression analysis
relation between interval level predictor variables and an interval level dependent variable

- R^2
- Collinearity
- Std Error of the Estimate

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + e_i$$

This week in practicum

No Practicum

Next time

Judgement measure in Visual Perceptions

- **Threshold and just-noticeable difference**
- **Rank order method**
- **Paired comparison method**

References

- Cohen, J., (1992). A power primer. *Psychological bulletin*, 112(1), 155-159.