Multiple Linear Regression and an Application on Language Attrition

Gulsen Yilmaz

Rijksuniversiteit Groningen,
April 2009

Outline

- Introduction to multiple linear regression
  -Method of least squares
  -Methods of regression
  -Outliers/residuals
  -Assumptions
- How to run and interpret regression analysis
- The study

Introduction to multiple linear regression

- Investigates relationships between variables using several independent variables and predicts numerical variable
- Effect of each variable can be estimated separately
- Used in econometrics, policy making and also ‘linguistics’
- Difference from correlation: predictive power
- example: income dependent on education, experience, school performance,…

Simple and multiple linear regression

Simple regression

- **D** : use of L2 Dutch (IV)
- **A** : attrition in L1 Turkish (DV)

A = a + bD + e

- a = constant (attrition with no Dutch use)
- b = ‘coefficient’ of D
  effect of an additional unit of Dutch use on attrition
- e = other factors that influence attrition (error, deviation)

→ mean of the outcome depends on one variable

Multiple regression

- **P**: positive attitude towards Dutch culture

A = a + bD + yP + e

- b: estimated effect of additional use of Dutch on attrition, holding positive attitude constant
- y: estimated effect of positive attitude on attrition, holding Dutch use constant

→ mean of the outcome depends on two variables

Multiple linear regression model

- $\text{Outcome}_i = \text{Model}_i + \text{error}_i$
- $y_i = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n + \epsilon_i$

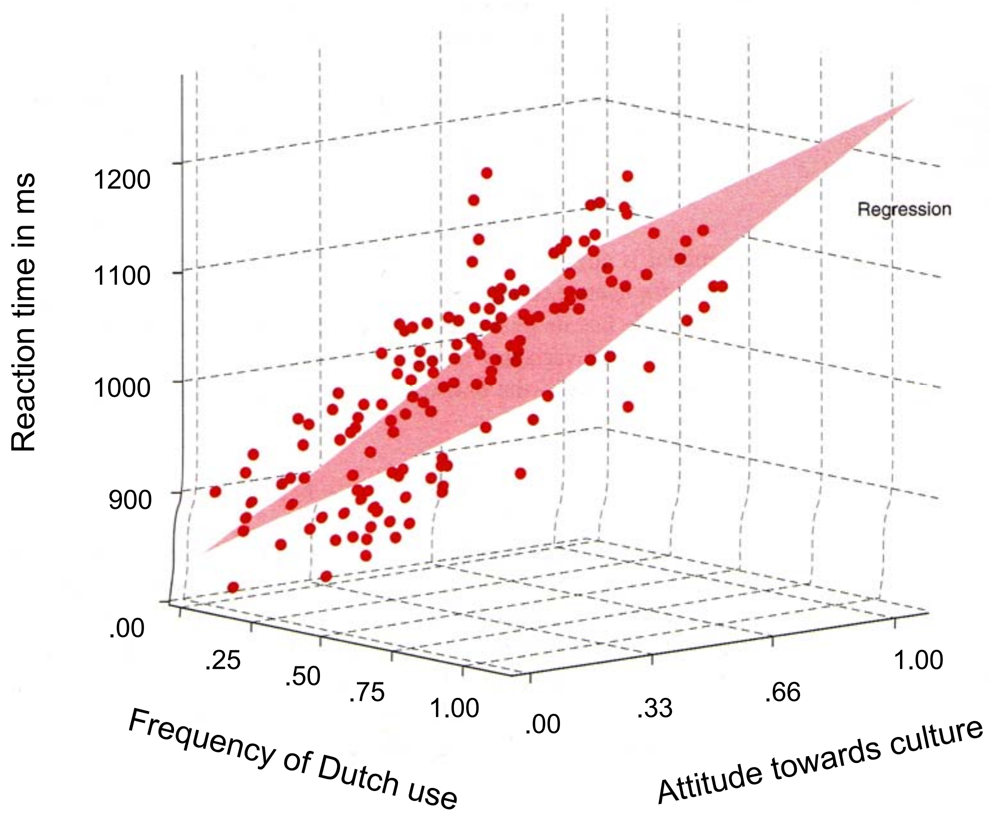
  y_i : outcome
  b_1 : coefficient of the first predictor x
  b_2 : coefficient of the second predictor x and so on
  ϵ_i : **deviations**, independent and normally distributed

Method of least squares

- Deviation = $\sum(\text{observed- model})^2$
- Line of best fit: the line that best describes the data
- The best fit if we have more variables
- Multiple regression: selects a plane so that the sum of squared errors is at a minimum

Scatterplot of the relationship between reaction time in L1, Dutch L2 use and positive attitude towards L2 culture

(hypothetical values)



Where does R^2 come from

- SS_T,Total Sum of Squares:
  observed data – mean of outcome
- SS_R, Residual Sum of Squares:
  observed data – regression line
- SS_M, Model Sum of Squares:
  mean of outcome – regression line
- $R^2 = SS_M / SS_T$
  example: $R^2 = 0.81$
  81% of the variability in the outcome is captured by the predictors in the equation
  19% residual
- Smaller the residual, the better the quality of the model

Methods of regression

stepwise methods for complex models:

- Enter: all predictors at once, builds the complex model all at once
- Forward: one predictor at a time,the best predictor, then the second best predictor
- Backward: builds the complex model, drops the least good predictor, then the second least good one

Which method to choose

- Not too many predictors

  i.e. principal component analysis

  correct children

  regret if they forget L1 → importance of L1 for children

  saturday classes

  etc.

- Past research
- Supression: Supressor effects occur when a predictor has a significant effect only when another variable is held constant.
- Forward selection → type 2 error due to supressor effects

Outliers and residuals (regression diagnostics)

- Outlier: very different from the rest of the data
- Influential: case with a large influence on our model
- See both outliers and influentials to assess your model
- But, no justification for data removal to have significant results

Some tips for regression diagnostics

Case summaries on the output:

- standardized: no more than 5% of cases > above 2
no more than 1% > above 2.5
any case > 3 could be an outlier
- Cook's distance: any value above 1, concern
- leverage: values 0-1, big values concern
- Mahalanobis distance: values above 25 (N=500, 5 predictors), and values above 15(N=100, 3 predictors), concern
- DFBeta: greater than 1, concern
- CVR (covariance ratio): if close to 1, ok

Assumptions of multiple regression

- **Variables**:

  Predictor:quantitative or categorical(with two categories)

  Outcome:quantitative, continious, unbounded
- **Nonzero variance**:Predictors should have some variation in value
- Predictors should be **uncorrelated** with **external** variables

Assumptions cont.

- **No perfect collinearity:** no perfect linear relationship between two or more of the predictors

→otherwise multicollinearity:

1. weak exlanatory power
2. difficult to assess the importance of individual factors
3. unstable predictor equations

→check: **VIF** (variance inflation factor)

tolerance statistic (1/VIF)

-largest VIF>10,concern

-average VIF>1, regression maybe biased

-tolerance< .1, serious problem

-tolerance< .2, a potential problem

Assumptions cont.

- **Homoscedasticity**: Residuals at each level of the predictors should have the same variance.
  →check by visual inspection of the residual scatter plot

- **Independent errors**: Errors should be uncorrelated
  →check **Durbin-Watson** test
  -If 2: residuals are uncorrelated, fine
  -concern: values <1 and values >3

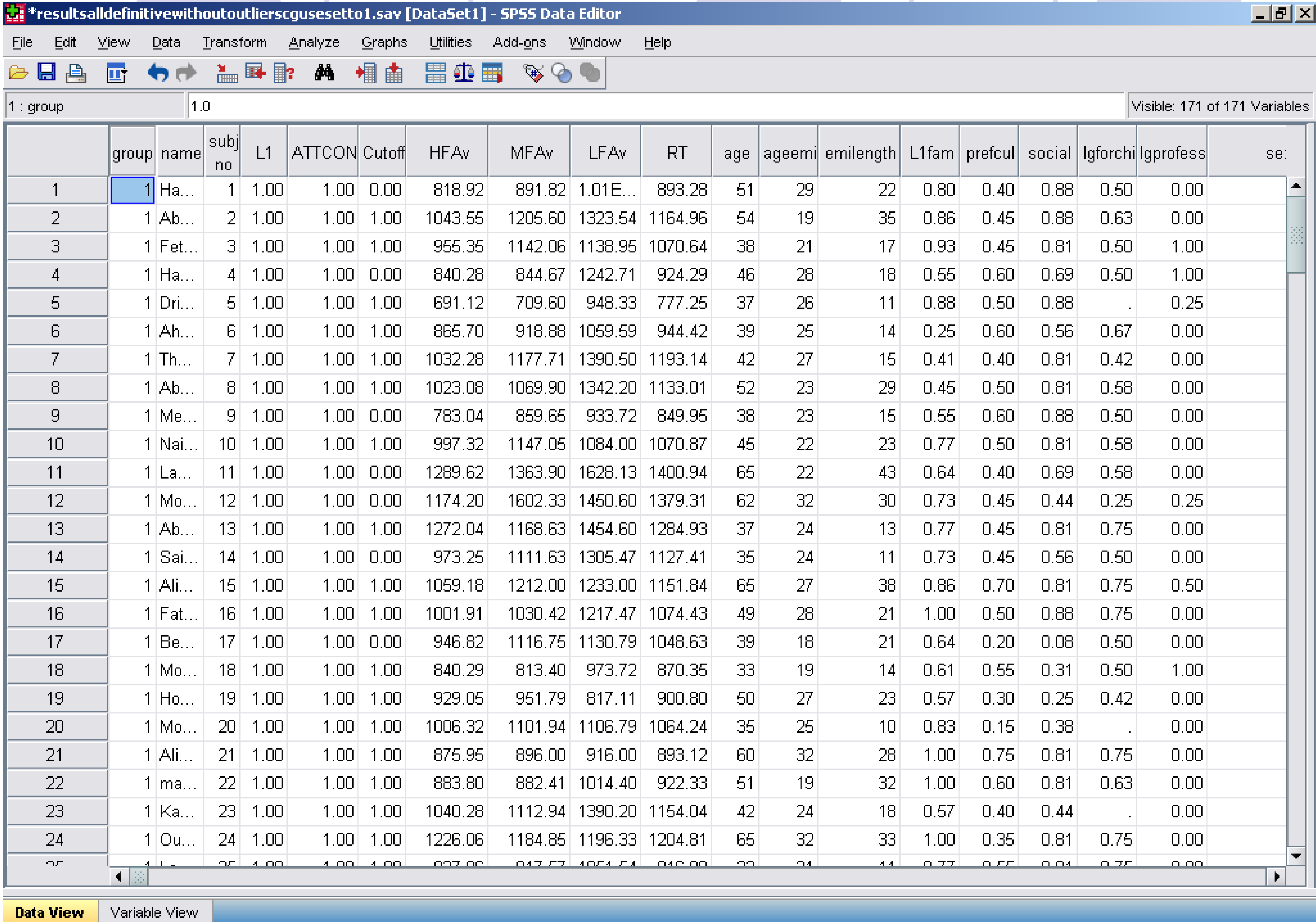
Assumptions cont.

- **Normally distibuted errors:** Residuals should be normally distributed with a mean of zero
- **Independence**: Each value of outcome variable should come from a separate entity
- **Linearity**: The mean values of the outcome variable for each increment of the predictors lie along a straight line

How to do multiple regression?

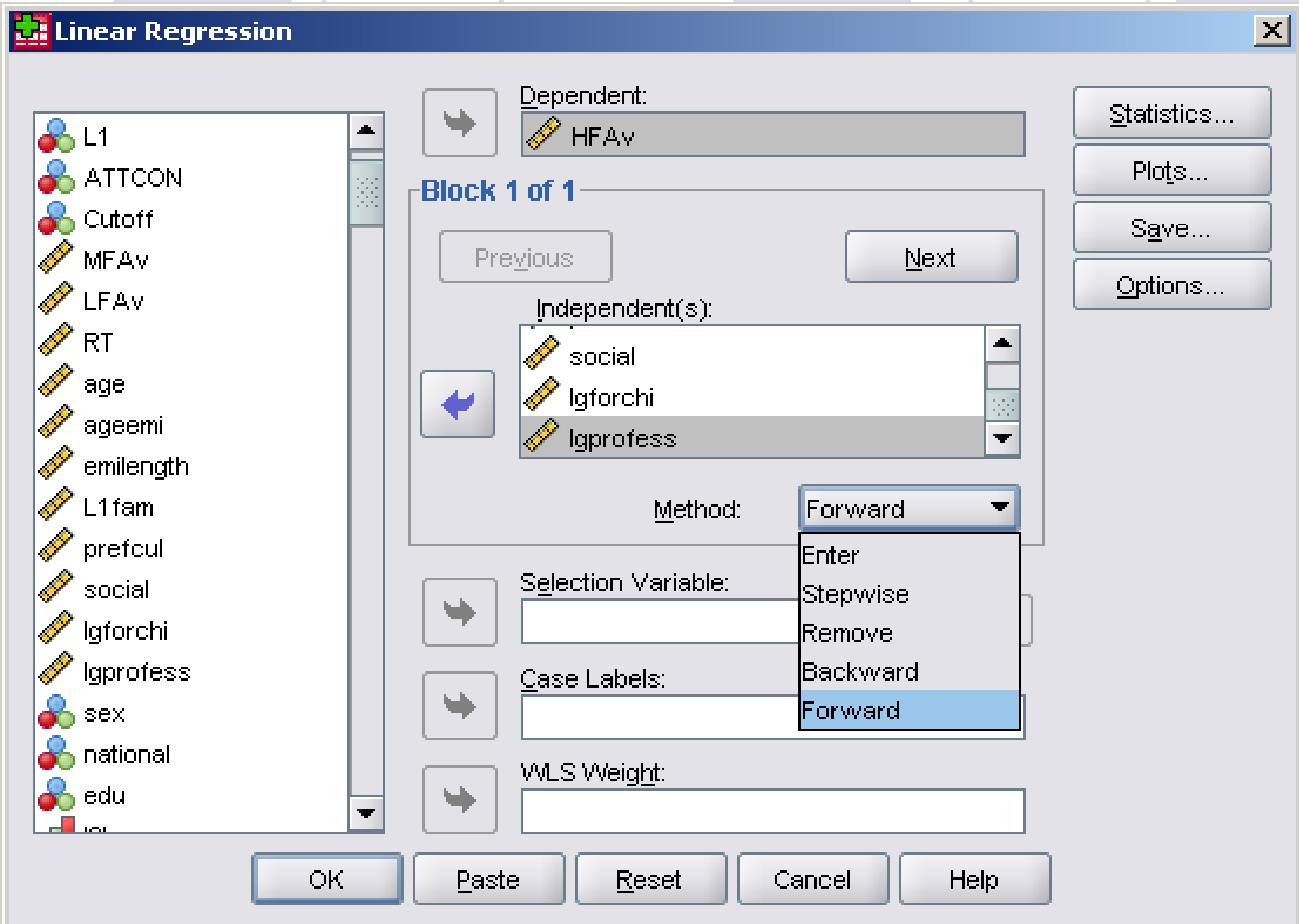
How to interpret the output?

This is how the data looks like on SPSS:

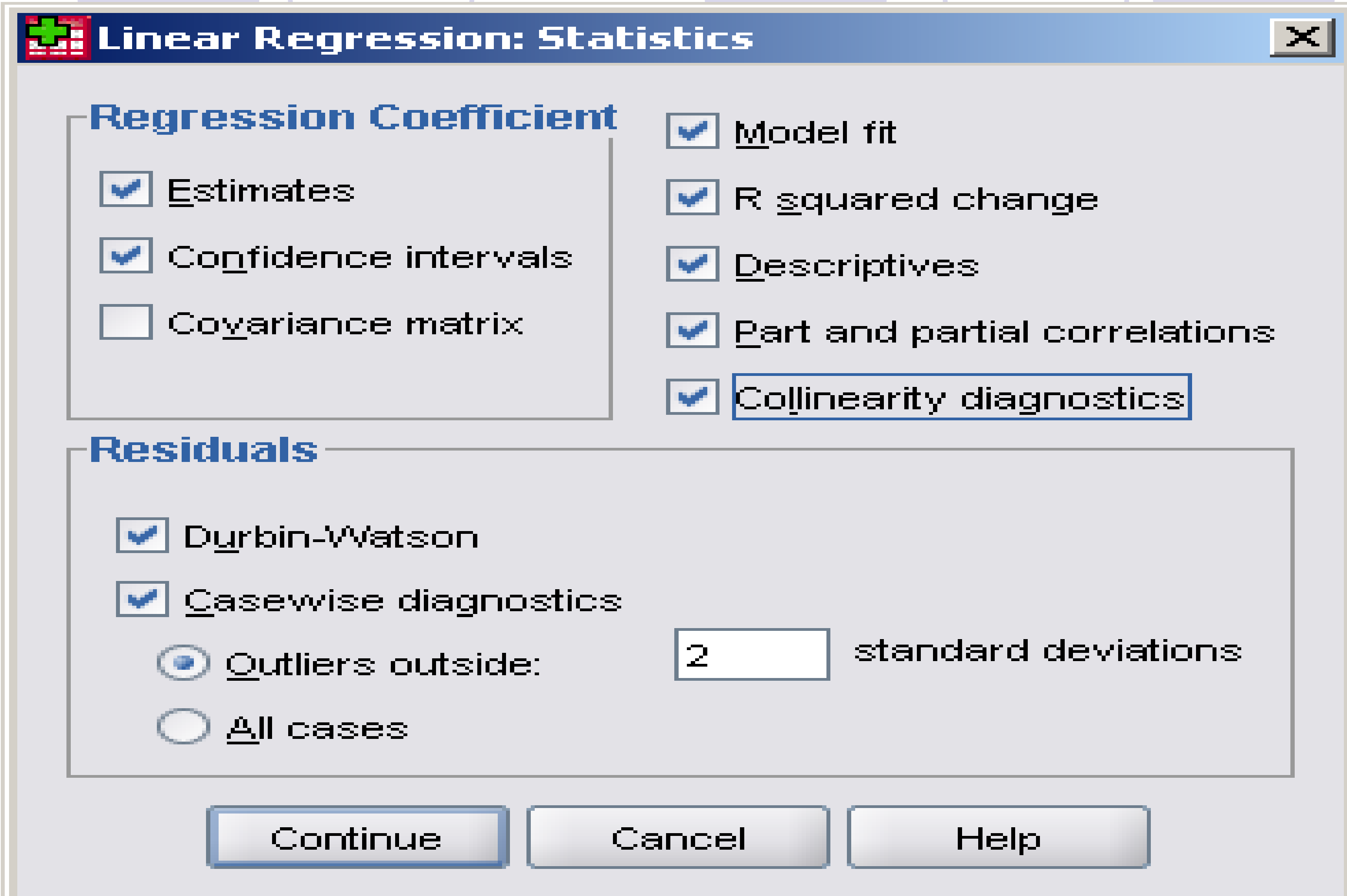


	group	name	subj no	L1	ATTCON	Cutoff	HFAv	MFAv	LFAv	RT	age	ageemi	emilength	L1fam	prefcul	social	lgforchi	lgprofess	se:
1	1	Ha...	1	1.00	1.00	0.00	818.92	891.82	1.01E...	893.28	51	29	22	0.80	0.40	0.88	0.50	0.00	
2	1	Ab...	2	1.00	1.00	1.00	1043.55	1205.60	1323.54	1164.96	54	19	35	0.86	0.45	0.88	0.63	0.00	
3	1	Fet...	3	1.00	1.00	1.00	955.35	1142.06	1138.95	1070.64	38	21	17	0.93	0.45	0.81	0.50	1.00	
4	1	Ha...	4	1.00	1.00	0.00	840.28	844.67	1242.71	924.29	46	28	18	0.55	0.60	0.69	0.50	1.00	
5	1	Dri...	5	1.00	1.00	1.00	691.12	709.60	948.33	777.25	37	26	11	0.88	0.50	0.88	.	0.25	
6	1	Ah...	6	1.00	1.00	1.00	865.70	918.88	1059.59	944.42	39	25	14	0.25	0.60	0.56	0.67	0.00	
7	1	Th...	7	1.00	1.00	1.00	1032.28	1177.71	1390.50	1193.14	42	27	15	0.41	0.40	0.81	0.42	0.00	
8	1	Ab...	8	1.00	1.00	1.00	1023.08	1069.90	1342.20	1133.01	52	23	29	0.45	0.50	0.81	0.58	0.00	
9	1	Me...	9	1.00	1.00	0.00	783.04	859.65	933.72	849.95	38	23	15	0.55	0.60	0.88	0.50	0.00	
10	1	Nai...	10	1.00	1.00	1.00	997.32	1147.05	1084.00	1070.87	45	22	23	0.77	0.50	0.81	0.58	0.00	
11	1	La...	11	1.00	1.00	0.00	1289.62	1363.90	1628.13	1400.94	65	22	43	0.64	0.40	0.69	0.58	0.00	
12	1	Mo...	12	1.00	1.00	0.00	1174.20	1602.33	1450.60	1379.31	62	32	30	0.73	0.45	0.44	0.25	0.25	
13	1	Ab...	13	1.00	1.00	1.00	1272.04	1168.63	1454.60	1284.93	37	24	13	0.77	0.45	0.81	0.75	0.00	
14	1	Sai...	14	1.00	1.00	0.00	973.25	1111.63	1305.47	1127.41	35	24	11	0.73	0.45	0.56	0.50	0.00	
15	1	Ali...	15	1.00	1.00	1.00	1059.18	1212.00	1233.00	1151.84	65	27	38	0.86	0.70	0.81	0.75	0.50	
16	1	Fat...	16	1.00	1.00	1.00	1001.91	1030.42	1217.47	1074.43	49	28	21	1.00	0.50	0.88	0.75	0.00	
17	1	Be...	17	1.00	1.00	0.00	946.82	1116.75	1130.79	1048.63	39	18	21	0.64	0.20	0.08	0.50	0.00	
18	1	Mo...	18	1.00	1.00	1.00	840.29	813.40	973.72	870.35	33	19	14	0.61	0.55	0.31	0.50	1.00	
19	1	Ho...	19	1.00	1.00	1.00	929.05	951.79	817.11	900.80	50	27	23	0.57	0.30	0.25	0.42	0.00	
20	1	Mo...	20	1.00	1.00	1.00	1006.32	1101.94	1106.79	1064.24	35	25	10	0.83	0.15	0.38	.	0.00	
21	1	Ali...	21	1.00	1.00	1.00	875.95	896.00	916.00	893.12	60	32	28	1.00	0.75	0.81	0.75	0.00	
22	1	ma...	22	1.00	1.00	1.00	883.80	882.41	1014.40	922.33	51	19	32	1.00	0.60	0.81	0.63	0.00	
23	1	Ka...	23	1.00	1.00	1.00	1040.28	1112.94	1390.20	1154.04	42	24	18	0.57	0.40	0.44	.	0.00	
24	1	Ou...	24	1.00	1.00	1.00	1226.06	1184.85	1196.33	1204.81	65	32	33	1.00	0.35	0.81	0.75	0.00	

Steps to run a multiple linear regression:



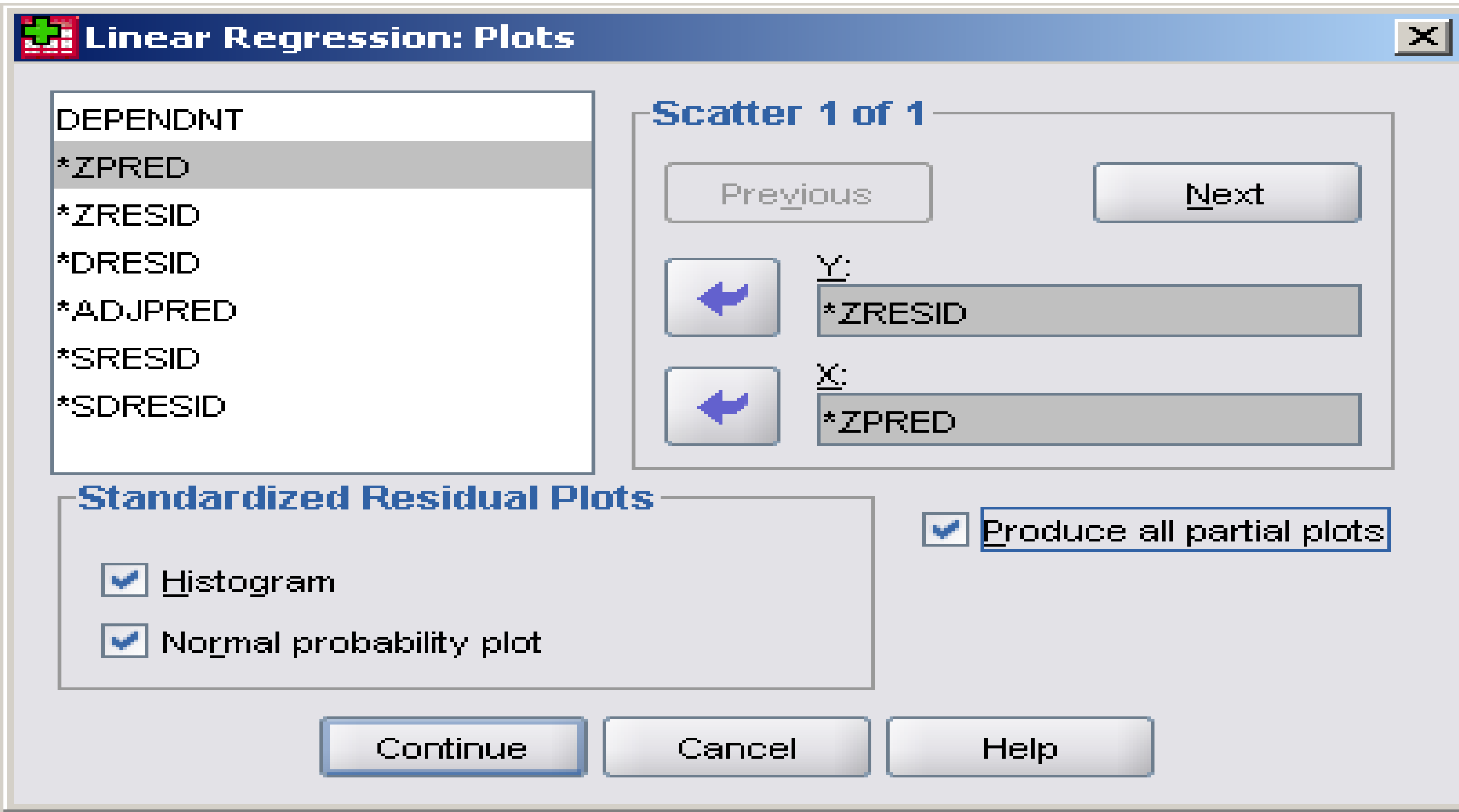
What the statistics options mean:



Regression plot is a good way to check the assumptions of random errors and homoscedasticity

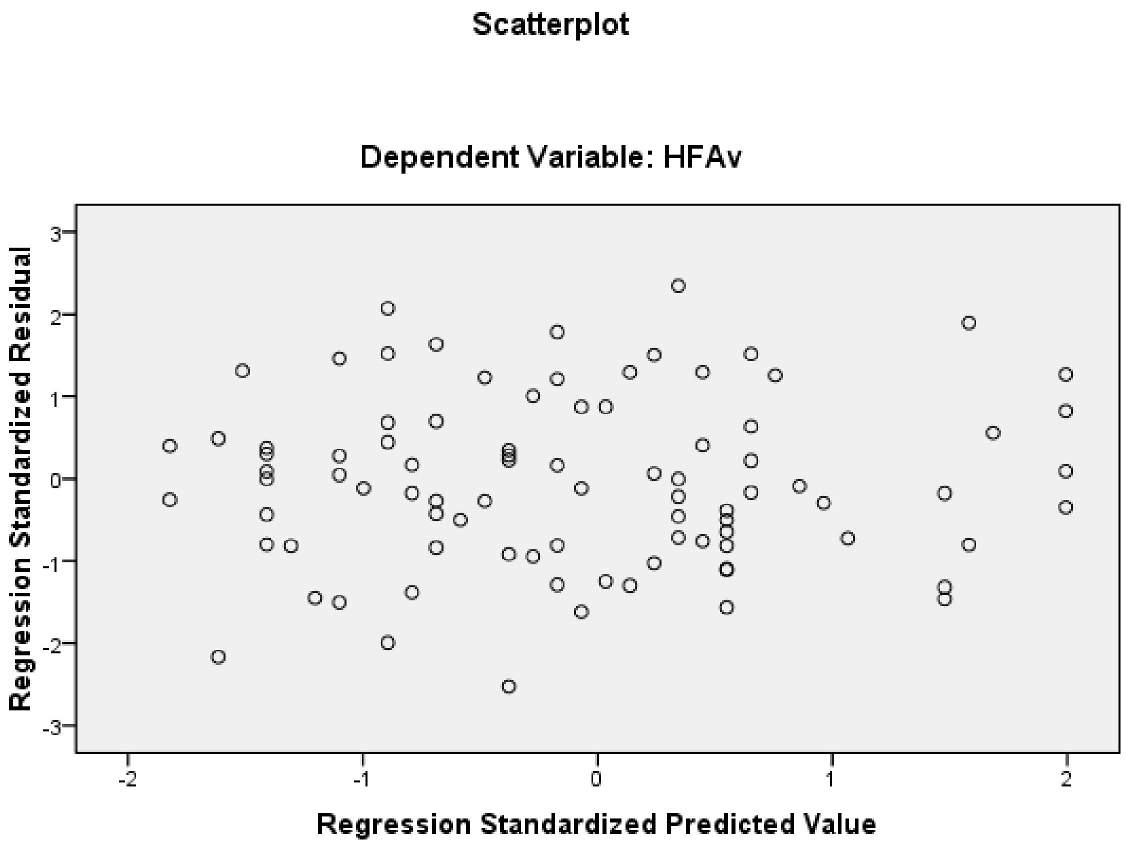
*ZRESID(standardized residuals, errors)

*ZPRED(standardized predicted values of DV based on the model



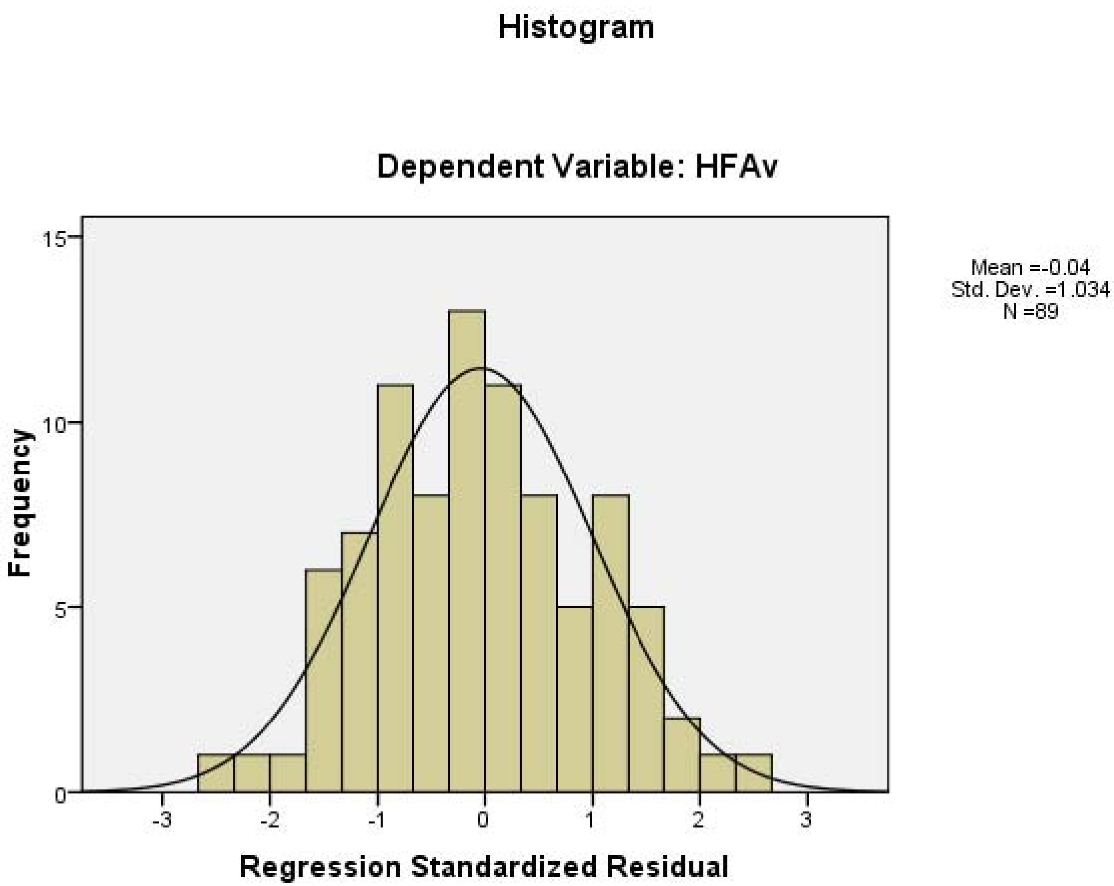
Plot of *ZRESID against *ZPRED

- assumptions of linearity and homoscedasticity met?
- yes, because points are random, widely dispersed, no sign of trend



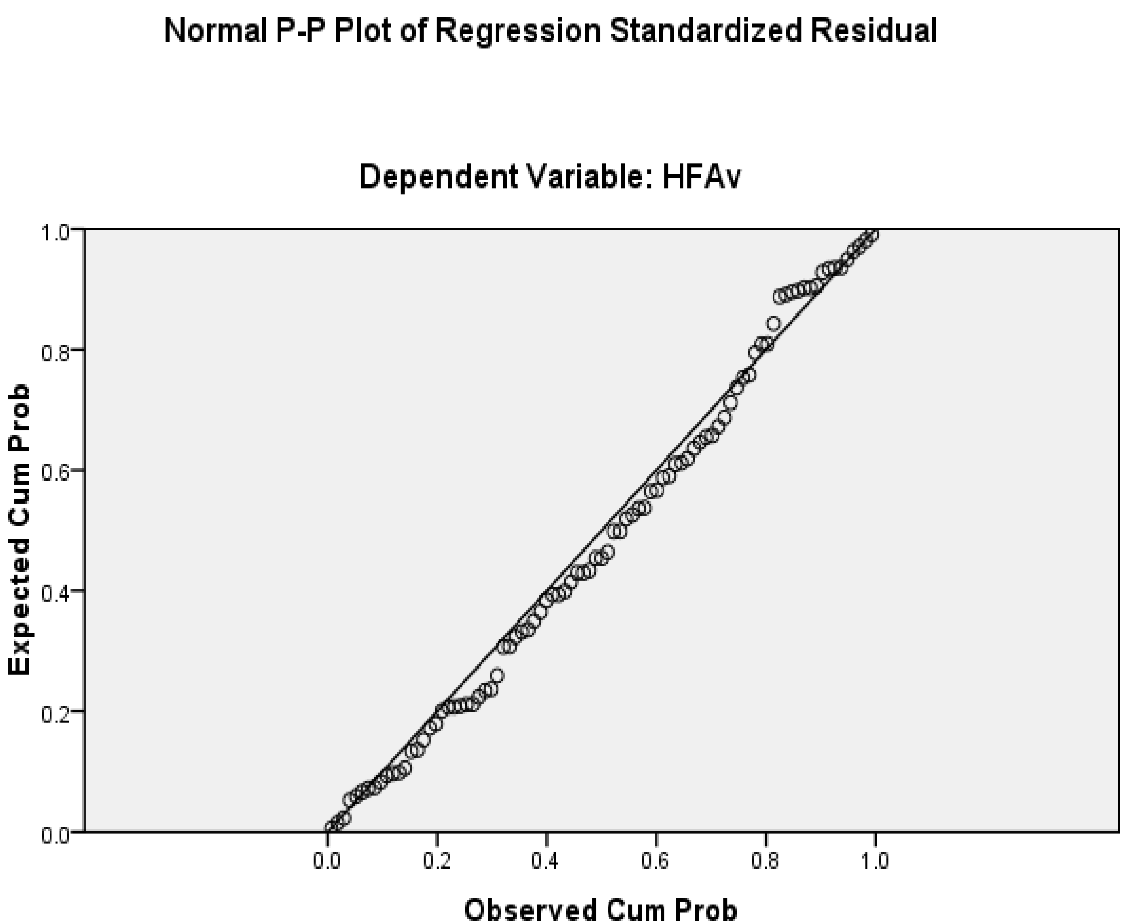
Histogram of residuals

- assumption of normal distibution of errors met?
- yes, a bell shaped curve means normal distribution



Plot of residuals

- assumption met?
- yes, straight line represents normal distribution



How to interpret multiple regression Anova

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.308[a]	,095	.083	142.81332	.095	7,657	1	73	,007	1.693

a. Predictors: (Constant), age

b. Dependent Variable: HFAv

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	156173.656	1	156173.656	7.657	.007[a]
	Residual	1488881.987	73	20395.644		
	Total	1645055.643	74			

a. Predictors: (Constant), age

b. Dependent Variable: HFAv

Interpretation

- Look at **F-ratio** and **significance** and **R^2**
- For this data F ratio is 7,657 and significant at $p < .01$
- Regression model predicts the outcome well
- R^2 = ,095
  age accounts for about 10% variation in the reaction time
- Durbin-Watson is close to 2, so fine

How to interpret multiple regression Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	800.573	79.913		10,018	,000	641.307	959.839					
	age	4.738	1.712	.308	2.767	.007	1.326	8.151	.308	.308	.308	1.000	1.000

a. Dependent Variable: HFAv

Interpretation

- Look at **t-ratio** and **significance**
- t- statistics: If a variable significantly predicts the outcome, it should have a coefficient significantly different from zero
- For this data t- ratio is 10.018, significant at $p<.001$
- Age is a good predictor

Introduction to the study

- **Aim**: Investigate L1 *attrition* among Turks and Moroccans in the Netherlands

- **Attrition:** “a linguistic system in disuse will be vying for memory space with the other linguistic system(s) occupying the same brain, […] not being kept ‘fresh’ and ‘strong’ through constant use will somehow weaken it, and […] it will therefore suffer in some way.” (Schmid, 2006:74)

L1 proficiency in a migrant context

- Limited exposure to L1 and less opportunities to use it
- Attitudes towards L1/ L2 and L1/L2 culture
- Factors that enhance L1 maintenance:

  a large community size, symbolic value of language, cultural and linguistic dissimilarity
- Yet, stability of the native language cannot be guaranteed

Activation Threshold Hypothesis(ATH): an account for attrition

ATH: Language disuse → higher thresholds → attrition

- First affects lexical items
  - -Word finding/retrieval problems
  - -Decreased lexical diversity
  - -Disfluency in speech
- Word retrieval: 2-5 words/second

  Conceptualization → Formulation → Articulation
- Bilingual disadvantage

Predictions of the study

- Lexical access problems: Slower Reaction Times (RTs)
- Despite

  - dominant L1 use

  - strong attachment to L1 and L1 culture

The Study

- Informants: first generation Moroccans (n = 35) and Turks (n = 54)
- Degree of bilingualism: various
- Age at arrival: 14 – 42 (mean: 22.00)
- Age: 28 – 65 (mean: 44.73)
- Length of residence: 10 – 43 years (mean: 22.37)
- Control groups: collected, matched

  (age: 25-62, mean: 43.45)

Research Design

I. Picture Naming Task

- 78 pictures (26 high, 26 mid, 26 low fam.)
- no cognates, no ambiguous pictures
- timed: 3000 ms
- accuracy and reaction time measured
- E-prime software

2. Sociolinguistic questionnaire

- L1 and L2 use, social networks, linguistic/cultural affiliation, attitudes towards language learning

3. Free speech

Variables in multiple regression

Predictors

(Independent Variables)

- L1 use in the family
- L1 social use
- Preferred culture
- Importance of L1 for children
- L1 professional use

Outcome

(Dependent Variable)

- RT on the PNT task

Outliers

- If half or more than half of the participants couldn’t name an object, item excluded
- If the response was below 250 ms, response excluded
- Cutoff point: those subjects with more than 25% invalid responses get a 0, those with less get a 1

Recode between 0 and 1

Example: Do you consider yourself a bilingual?

1= NL better, 2=bilingual, 3=TR better

original 1=NL better, recoded as 0

original 2=bilingual, recoded as 0.5

original 3= TR better, recoded as 1

Check reliability of subscales

Example:

L1 use in family : nationality of partner, language with partner, with children, with grandchildren

Reliability goes up when grandchildren are omitted

Compute mean for predictors and reaction time

- Example: Preferred culture is L1 or L2 culture

  COMPUTE prefcul = MEAN(mosque,culture,L1friend,L1club,L1media)

- RT measured in miliseconds

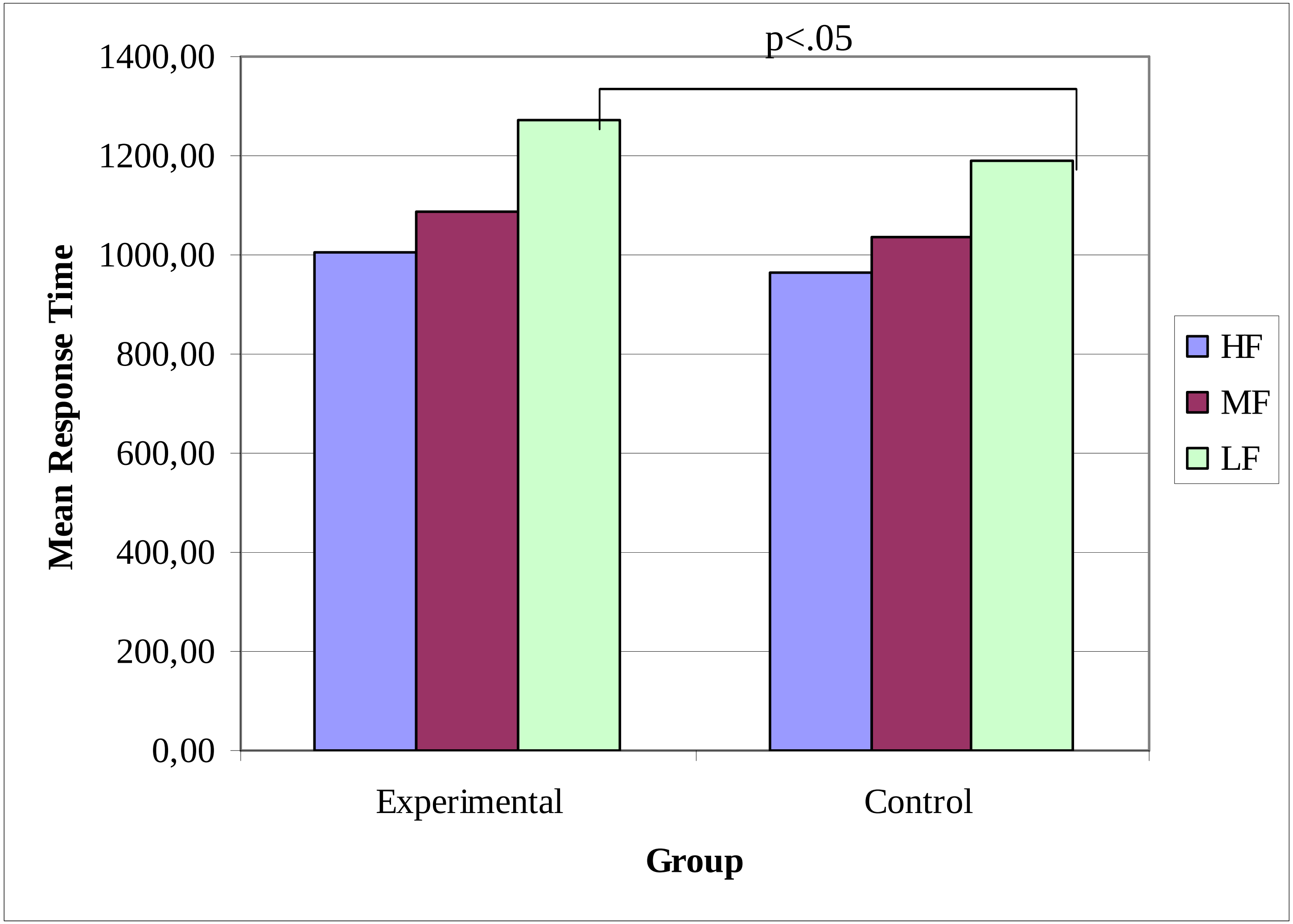
  Total RT (78 items)

  High Fam RT (26 items)

  Medium Fam RT (26 items)

  Low Fam RT (26 items)

Picture Naming Task: Reaction Time Results



Results

- Slower RTs in the experimental group compared to controls
- LF significant, HF and MF approaching significance
- So, sign of lexical retrieval difficulties

Predicting performance on the PNT on the basis of L1 use/attitudes

Multiple linear regression

- Attrition not related to variables in question except age

T-tests

- Attrition in only MA group
- TR: maintainers
- MA controls faster than TR controls

Discussion: Why Moroccans differ from Turks

Group level differences:

- MA: early multilinguals (Berber and/or French)
- Turks: no other languages before coming to NL
- Moroccans more open to Dutch language and culture

Individual level factors/predictors:

- Total languages, attitudes not related to attrition
- L2 proficiency may be a potential factor

Discussion: Multiple linear regression

- Why the other predictors turned out to be weak?
- Possible correlation between the predictors?

  i.e. if they prefer L1 culture they would automatically use L1 more
- Enough number of participants?
- Small range of variation in reaction time?

  i.e. only 80 ms yields to significant difference
- What other potential predictors can account for the outcome? i.e. Dutch proficiency, language specific factors in TR and MA

Future of the Study

- Data collection in L2 from the same speakers
- Analysis of spoken data in L1 and L2
- Effects of multicompetence on lexical access in L1 and L2
- Signs of lexical attrition in free speech
- Effects of attrition in other domains

THANK YOU!