Multiple linear regression

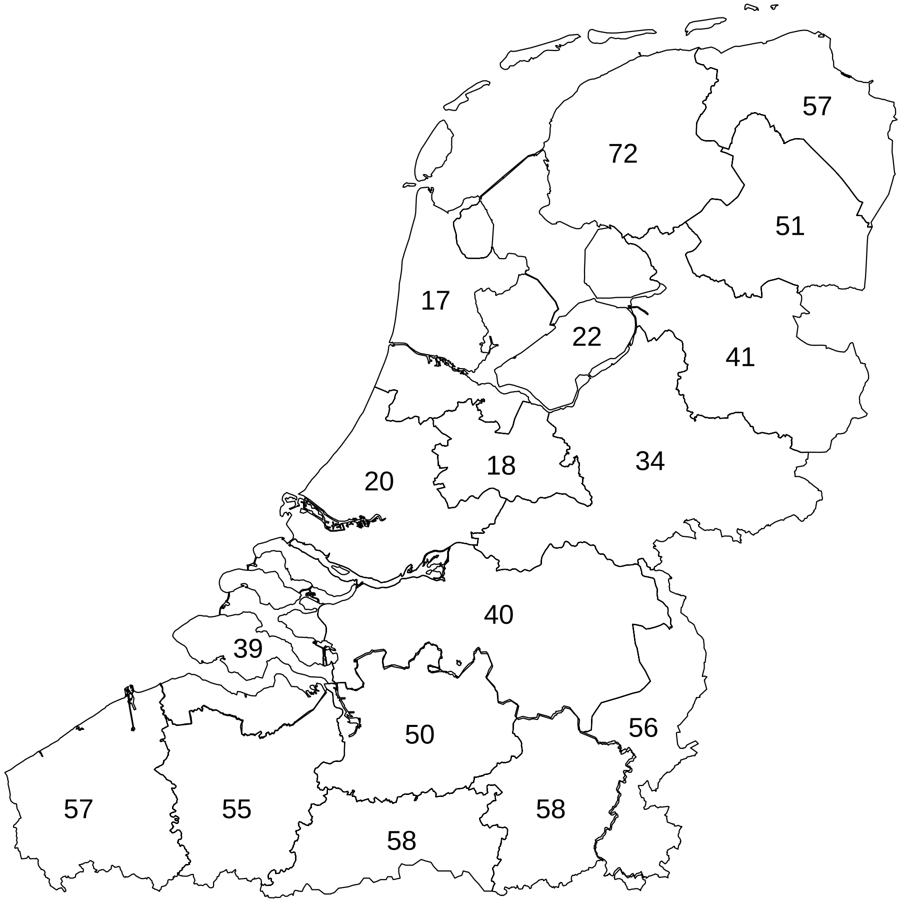
Statistics II (LIX002X05)

University of Groningen, Faculty of Arts, Information Science

Wilbert Heeringa

Variables

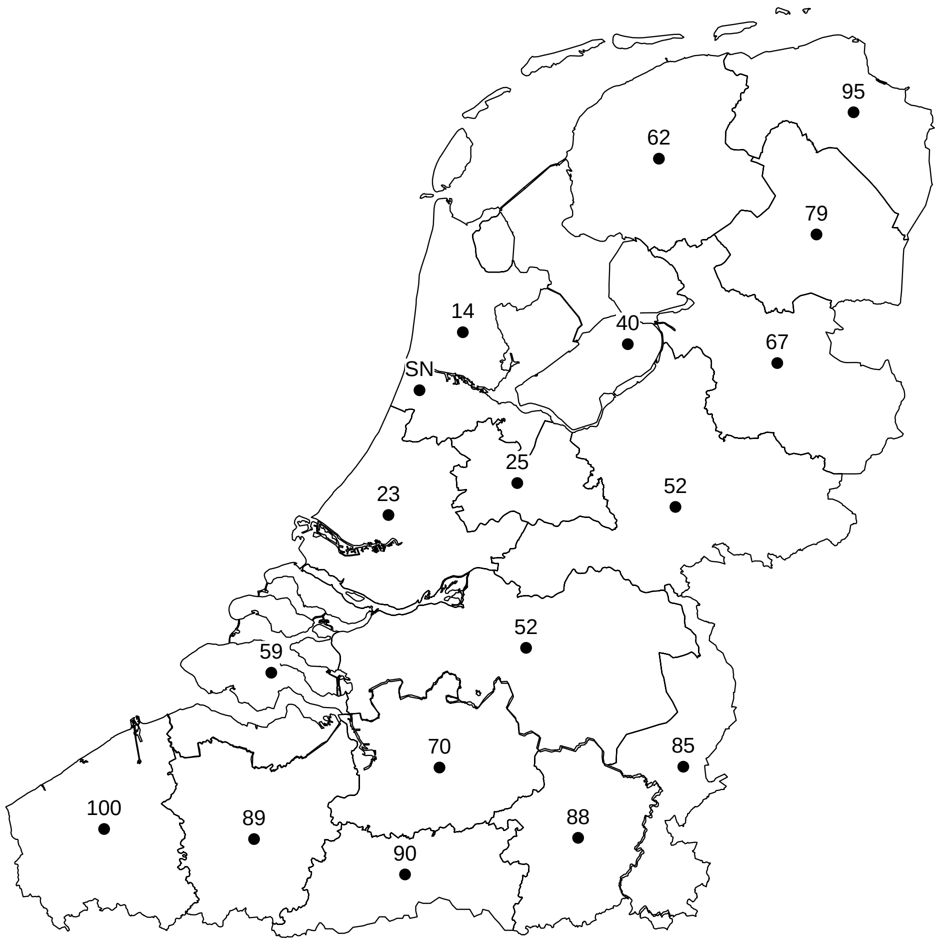
- Experiment Van Bezooijen & Heeringa (2006): measure intuitions of non-linguists about dialects in the Netherlands and Flanders.
- Task: rate the dialect distance compared to standard Dutch per province in a map: 0=no distance, 100=maximal distance.
- 140 Dutch subjects were involved in the experiment.



Average intuitive linguistic distances per province compared to standard Dutch.

Variables

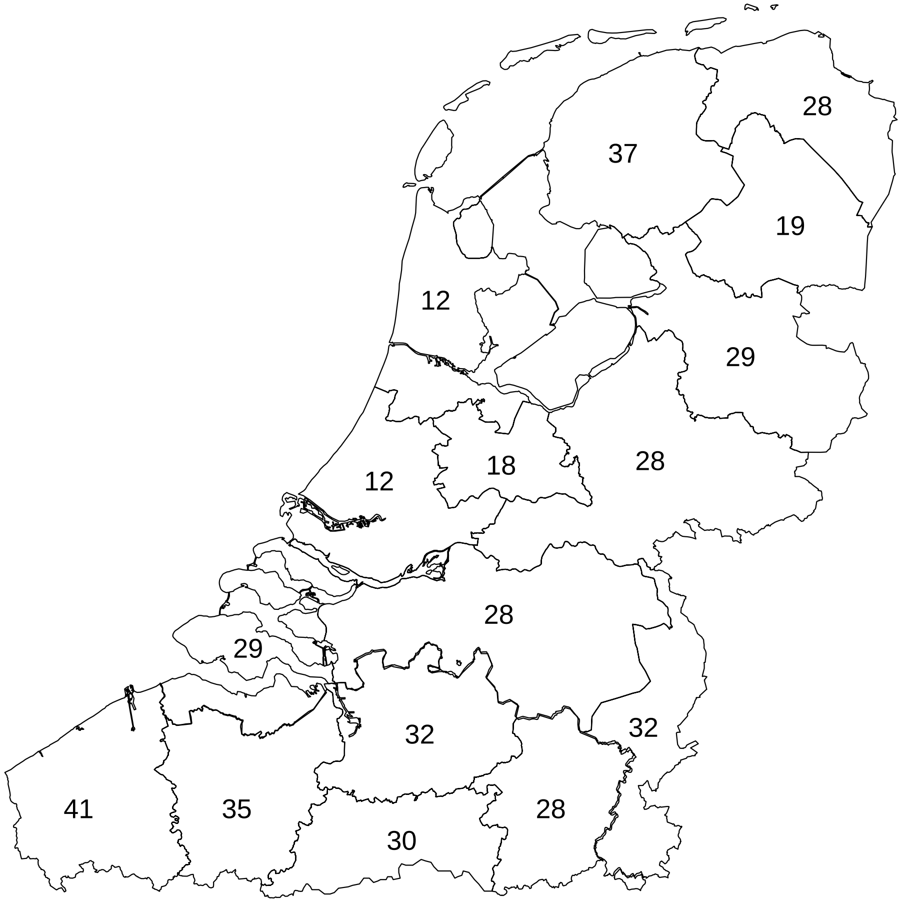
- Geographic distances between provinces and standard Dutch can also be measured.
- We locate standard Dutch at the position of Haarlem (situated west of Amsterdam).
- We measure as-the-crow-flies distances between the geographic centers of the provinces and Haarlem in millimeters scaled between 0 and 100.



Geographic distances between the geographic centers of the provinces and standard Dutch (located at the position of Haarlem).

Example

- In the previous lecture we found that $geography$ is an explanatory variable explaining $intuition$ (76%).
- Do subjects use linguistic knowledge when judging the distance of a province compared to standard Dutch?
- Per province we measured the extent to which the pronunciation of words differs from the standard Dutch pronunciation.
- For each province we measured the **percentage** of speech segments in the words which are pronounced differently compared to the standard language.



Average pronunciation differences per province compared to the standard language.

Example

- We use **multiple** linear regression analysis with:
  - Response variable:
    $intuition$
  - Explanatory variable:
    $geography$
  - A **second** explanatory variable:
    $pronunciation$

Model

- We describe the intuitive distance μ_y as population mean reponse as a function of the explanatory variables $geography$ and $pronunciation$.
- The response variable y depends on p explanatory variables (in our example p=2). These variables are referred to as:

$$x_1, x_2, \ldots, x_p$$

- The mean response is a linear function of the explanatory variables:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- For any fixed $x_1..x_p$ (i.e. a combination of p values, for each x_i one value), the responses y follow a normal distribution with standard deviation σ.

Model

- DATA = FIT + RESIDUAL
- Assume n cases (or: individuals), and i being the index of the cases ($1 \leq i \leq n$). For the ith case the model for the sample data is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

  where $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ is the population mean response.
- In our example: $n = 16$ provinces. There are predictors x_1 (geographic distance) and x_2 (pronunciation distance).
- The deviations ϵ_i are supposed to be independent and normally distributed with mean 0 and standard deviation σ. I.e. they are a SRS from a population with $N(0, \sigma)$ distribution.
- The parameters of the model are: $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ en σ.

Model

- When using simple linear regression, the least-squares line is used as a basis for inference regarding a population which our sample of observations is taken from.
- Multiple regression analysis works basically the same, but the details are more complicated.
- The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ and σ are estimated by $b_0, b_1, \ldots, b_n$ and s.
- For the ith observation the predicted response is:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

Estimate of the parameters

- The least-squares method chooses values for b's which minimize the residues (differences $y_i - \hat{y}_i$).
- Calculation of the b's in multiple regression is more complicated.
- Standard deviation: measures the variation of y around the population regression line:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p - 1}}$$

$n - p - 1$ is the number of degrees of freedom: n is the number of individuals, and $p + 1$ is the number of β's to be estimated $\rightarrow$ n - $(p + 1) = n - p - 1$.

Confidence intervals

- A level C confidence interval for β_j is:

$$b_j \pm t^* SE_{b_j}$$

- t^* is the value for the $t(n-p-1)$ distribution with surface C between $-t^*$ and t^*.

- The calculation of the SE_{b_j}'s is complicated.

Significance tests

- Hypotheses of significance test for β_j:

  $H_0 : \beta_j = 0$ (mean of y does not vary with x)

  $H_a : \beta_j > 0$; the p-value is $P(T \geq t)$
  $H_a : \beta_j < 0$; the p-value is $P(T \leq t)$
  $H_a : \beta_j \neq 0$; the p-value is $2P(T \geq |t|)$

  where T is a stochastic variable with the $t(n-p-1)$ distribution.

- In our example we have β_1 ($\beta_{geography}$) and β_2 ($\beta_{pronunciation}$)

Significance tests

- The test statistic t is:
$$t = \frac{b_j}{SE_{b_j}}$$
- The number of degrees of freedom is $n - p - 1$.

Assumptions

- 1. Linearity
  The residual plot should not show any obvious pattern. If you find a curve or another pattern there is no linearity.
- When performing a regression analyis in SPSS make a plot: *ZPRED (X) versus *ZRESID (Y).
- 2. No perfect multicollinearity
  When two or more predictor variables are highly correlated, meaning that one can be linearly predicted from the others with a non-trivial degree of accuracy, we call this $multicollinearity$.
- Multicollinearity does not reduce the predictive power or reliability of the model as a whole.
- The model still indicates how well the entire bundle of explanatory variables predicts the response variable.

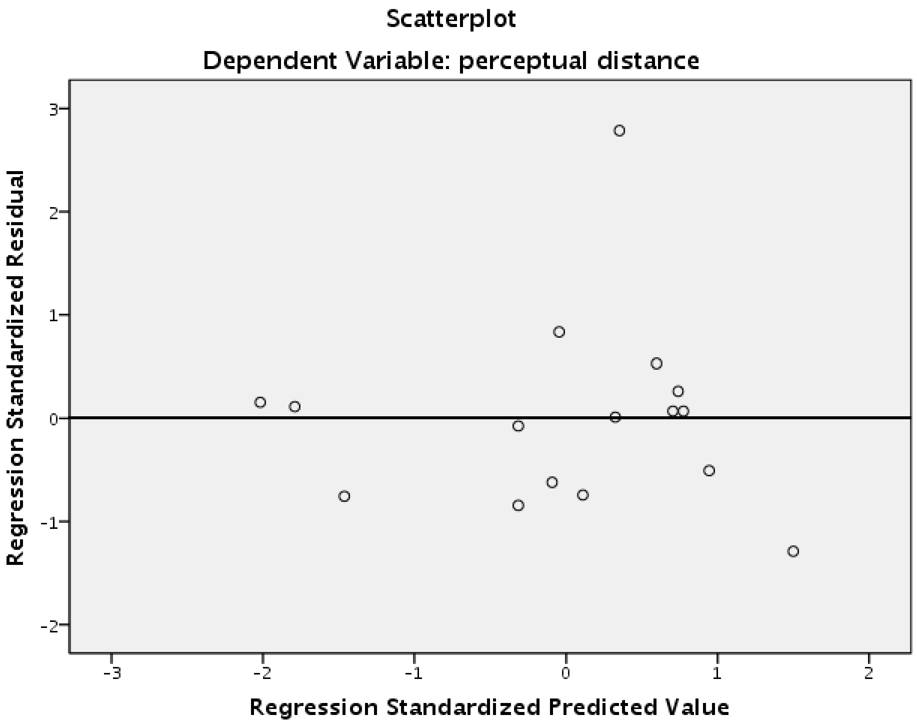
Assumptions

- But it may cause invalid results about any individual predictor, or about which predictors are redundant with respect to others.
- Make scatterplots and calculate correlation coefficients for each pair of predictors. The r's should be lower than 0.9.
- 3. Homoskedasticity
  The variability of the data should be approximately equal across the range of the predicted values. At each level of the predictors the variance of the residuals should be constant.
- The residuals need to roughly have a similar amount of deviation from the predicted values. A good residual plot essentially looks blob-like.
- 4. Normality of residuals
  This assumption is the least important and sometimes even not mentioned.
- Perform a Shapiro-Wilk test on the residuals and make a normal quantile plot of the residuals.

Assumptions

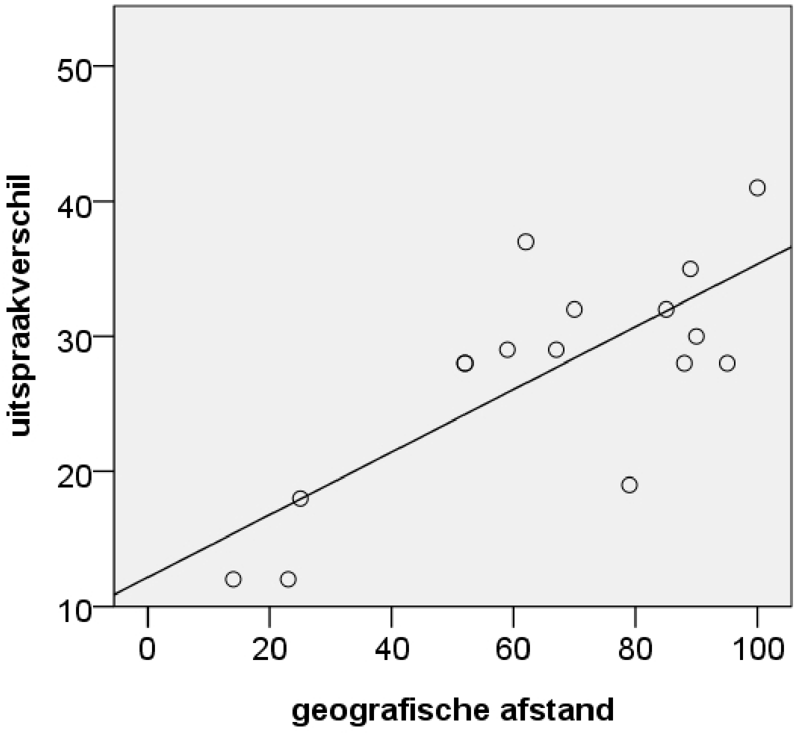
- 5. Absence of influential datapoints
  Calculate Cook's distance. A rule-of-thumb is to be concerned about Cook's distance greater than one.
- The Cook's values are generated when doing the regression analysis. Simply make a bar plot of the values.
- Do not automatically remove outliers and influential points! Check the data, look for errors or explanations for the outlier. You should always have a substantive reason to remove outliers.
- 6. Independence
  All the values of the dependent variable (outcome variable) are independent of each other.

1. Linearity / 3. Homoskedasticity



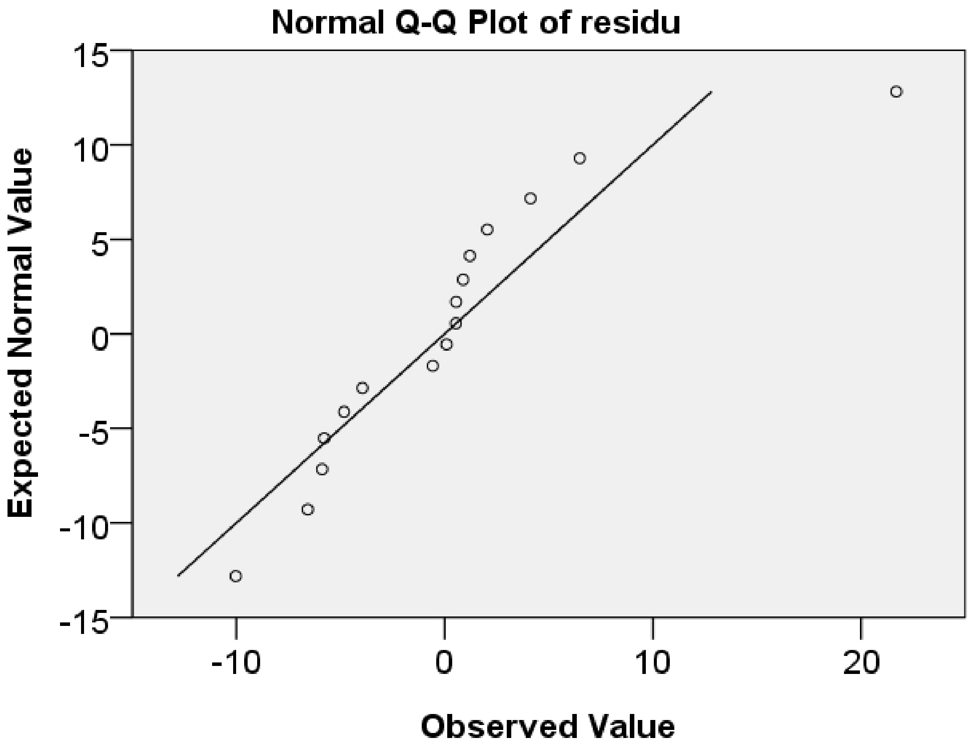
Residual plot: residues drawn against the predicted intuitive linguistic distances ($\hat{y}$).

2. No perfect multicollinearity



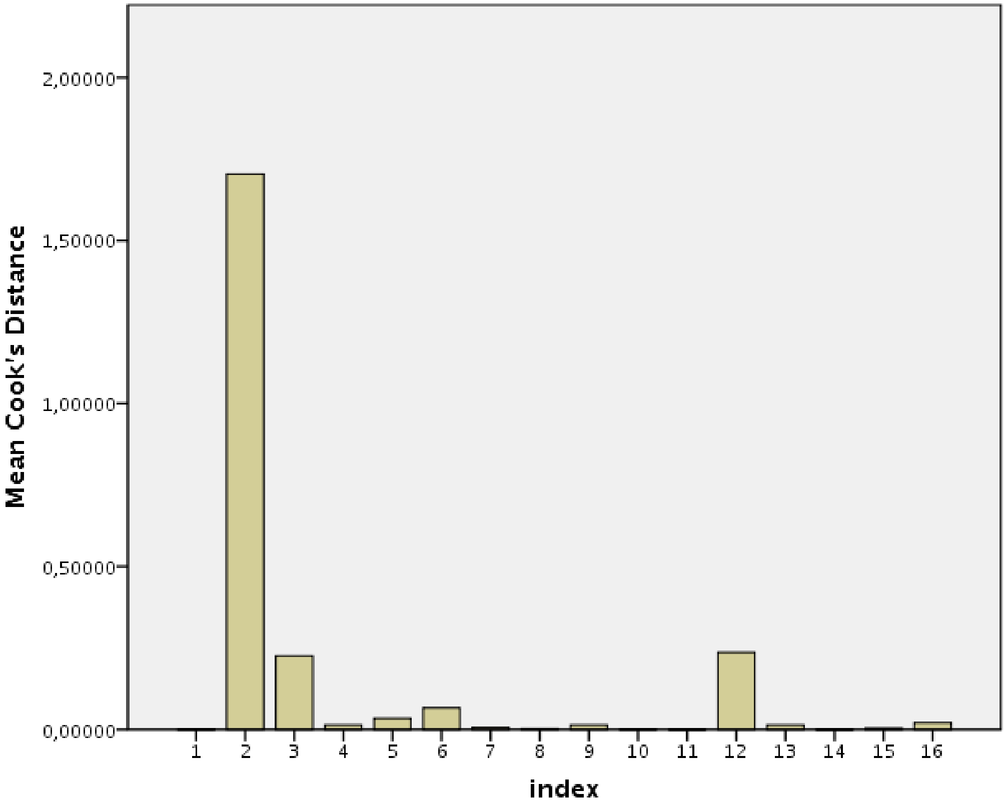
Correlation r=0.754 ($p = 0.001$), and r^2=0.569.

4. Normality of residuals



Normal quantile plot of the residues. The Shapiro-Wilk test gives $p = 0.012$

5. Absence of influential datapoints



Plot with Cook's values. Friesland has a Cook's distance $>$ 1: 1.70382

6. Independence

- All the values of the dependent variable *intuitive linguistic distance* (henceforth *perceptieve afstand*) are independent of each other.

Selection of predictors

- A simpler model should be preferred to a more complex model if they have the same explanatory power (R^2).
- Over-fitting:
  Having to many variables in the model that hardly contribute to predicting the outcome.
- Under-fitting:
  Important predictors are left out.

Selection of predictors

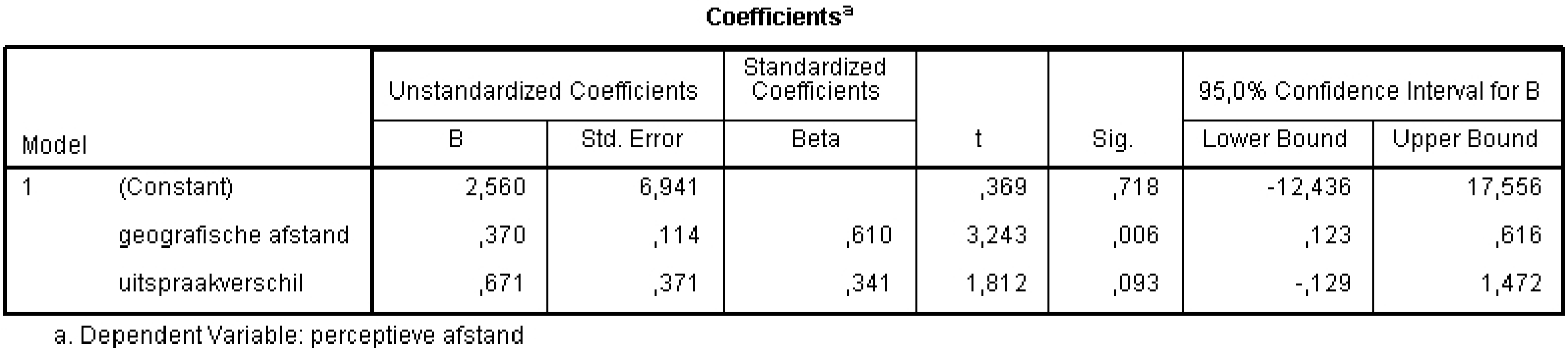
- All-at-once regression:
  All of the explanatory variables are entered in the model at once.
- Forward selection:
  Start with the most important explanatory variable, subsequently add the other explanatory variables the one after the other in order of importance.
- Backward elimination:
  Start with all explanatory variables, remove the other explanatory variables the one after the other in order of unimportance.
- Stepwise regression:
  Start with the most important explanatory variable. In each next step; add a variable if this improves the model, remove a variable if this makes the model worse.

Adding and/or removing is repeated until further improvement of the model is no longer possible.

Selection of predictors

- All-at-once (or: forced entry) is preferred.
- Stepwise techniques are influenced by random variation in the data, they rarely give the same results when the analysis is repeated on other data (Studenmund & Cassidy 1987).
- The usefulness of one variable is assessed on the basis of other variables that have been selected, which may be risky when potential explanatory variables are correlated (Levshina 2015).
- When using backward elimination, the risk of missing out important variables is smaller.

Results



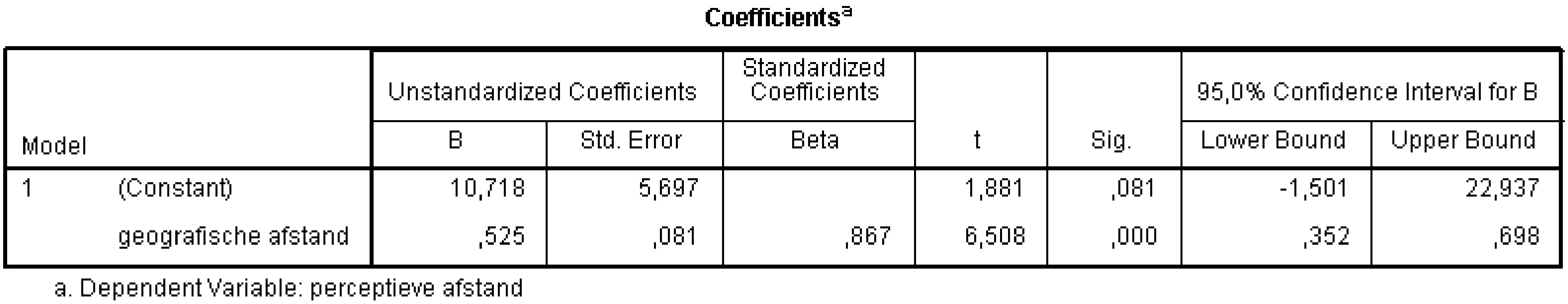
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2,560	6,941		,369	,718	-12,436	17,556
	geografische afstand	,370	,114	,610	3,243	,006	,123	,616
	uitspraakverschil	,671	,371	,341	1,812	,093	-,129	1,472

a. Dependent Variable: perceptieve afstand

Output in SPSS of all-at-once analysis. b_0=2.560, b_1=0.370 and b_2=0.671.

Results



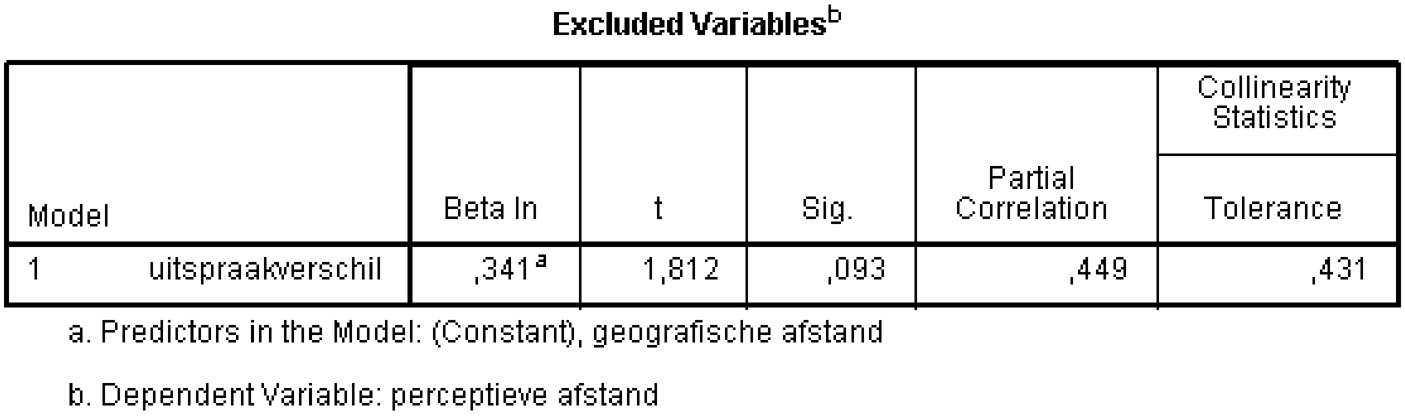
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	10,718	5,697		1,881	,081	-1,501	22,937
	geografische afstand	,525	,081	,867	6,508	,000	,352	,698

a. Dependent Variable: perceptieve afstand

Output in SPSS of stepwise linear regression analysis: included variable(s).

Results



Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	uitspraakverschil	,341[a]	1,812	,093	,449	,431

a. Predictors in the Model: (Constant), geografische afstand

b. Dependent Variable: perceptieve afstand

Output in SPSS of stepwise linear regression analysis: excluded variable(s).

Suppression

- Assume two explanatory variables x_1 and x_2 and a response variable y.
- x_1 has a reasonable positive correlation with y, x_2 has little or no correlation with y.
- Yet when x_2 is included in the regression model, the fit of the model improves and the coefficient b_1 for x_1 becomes higher.
- When x_2 is not included, it $suppresses$ the relationship between x_1 and y, because x_1 and x_2 are related in some way.

ANOVA

- The factors in ANOVA are comparable to the explanatory variables in multiple regression analysis.
- The factors in ANOVA are categorical. When there is just one factor, each value of the factor defines a group. Example: sex: 1=male, 2=female.
- The number of values of a factor is restricted: two, three, four, but more are possible.
- The factor (=explanatory variable) in a regression analysis may also be continuous. For example age (being an explanatory factor of $length$). Just as in ANOVA each value of the factor defines a group.
- The number of values of a factor is large, e.g. for age: 2, 18, 19, 25, 40, 65, 83, etc.

ANOVA

- When there are i cases and explanatory variables x_j, each unique combination of x_{ij}'s defines a group.
- We have 16 provinces, therefore there are 16 combinations of a value for geography and a value for pronunciation. Each unique combination defines a group.
- The values y_i of the response variable y are the observations. The $\hat{y}_i$'s predicted by the regression model are the group means, and $\overline{y}$ is the global mean of the observations y_i.
- y_i's are intuitive linguistic distances obtained with the experiment; $\hat{y}_i$'s are intuitive linguistic distances predicted by the regression line on the basis of geography and pronunciation.

ANOVA

- DATA = FIT + RESIDUAL
  - FIT:
SSM = Sum of Squares Model
  - RESIDU:
SSE = Sum of Squares Error
  - DATA:
SST = Sum of Squares Total

FIT

- ANOVA:
  SSM measures the variation of the group means around the global mean.
- Regression:
  SSM measures the variation of the mean responses $\hat{y}_i$ compared to the global mean $\overline{y}$:

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

- Mean Sum of Squares Model:

$$MSM = SSM/DFM$$

RESIDUAL

- ANOVA:
  SSE measures variation of individual observations compared to their group means.
- Regression:
  SSE measures variation of the individual observations y_i compared to their mean $\hat{y}_i$:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Mean Sum of Squares Error:

$$MSE = SSE/DFE = s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p - 1}}$$

DATA

- ANOVA:
  SST measures variation of individual observations compared to the global mean.
- Regression:
  SST measures variation of individual observations y_i compared to the global mean $\overline{y}$:

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- Mean Sum of Squares Total:

$$MST = SST/DFT$$

ANOVA

- F test statistic:

$$F = \frac{MSM}{MSE}$$

- p-value:
  the probability that a stochastic variable with distribution $F(p, n-p-1)$ is larger than or equal to the value of the test statistic F while assuming that H_0 is true.

ANOVA table

Source	Sum of squares	Degrees of freedom	Mean sum of squares	F
Model	$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	p	SSM/DFM=MSM	MSM/MSE
Error	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-p-1$	SSE/DFE=MSE	
Totaal	$\sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$	SST/DFT=MST	

ANOVA

- Hypotheses:
  - ○ H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$;

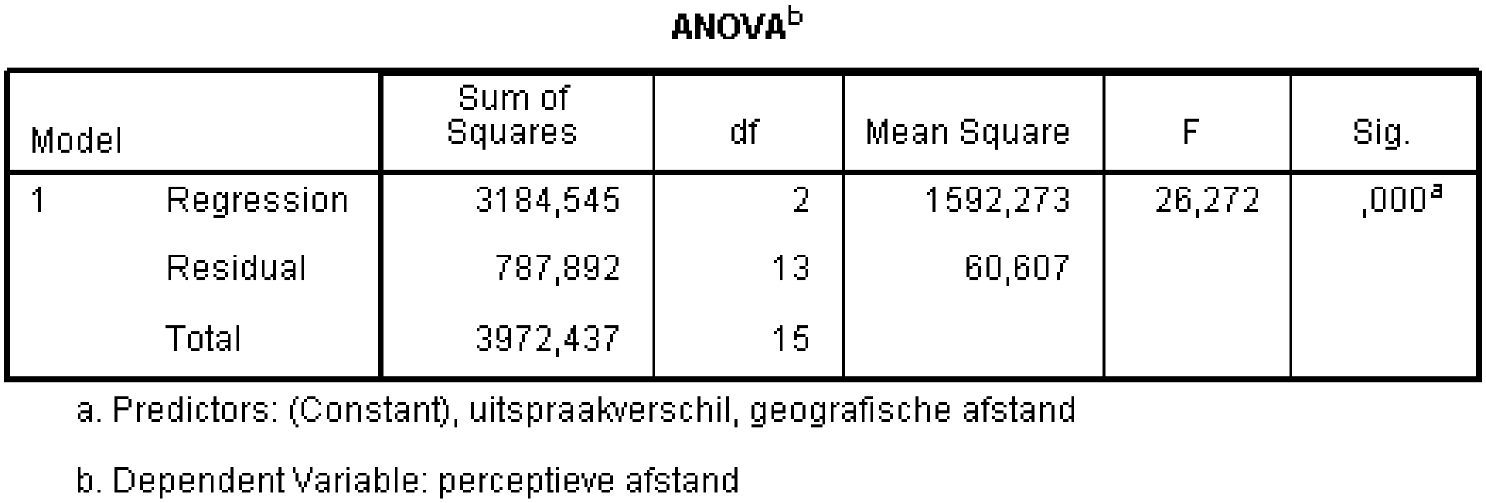
    all y's are taken from the same population with mean μ_y;

    none of the explanatory variables is a predictor of the response variable in the regression equation.
  - ○ H_a: at least one of the β_j's is not equal to 0.

    not all y's are taken from the same population;

    at least one of the explanatory variables has a linear relationship with the response variable.

ANOVA



ANOVA[b]

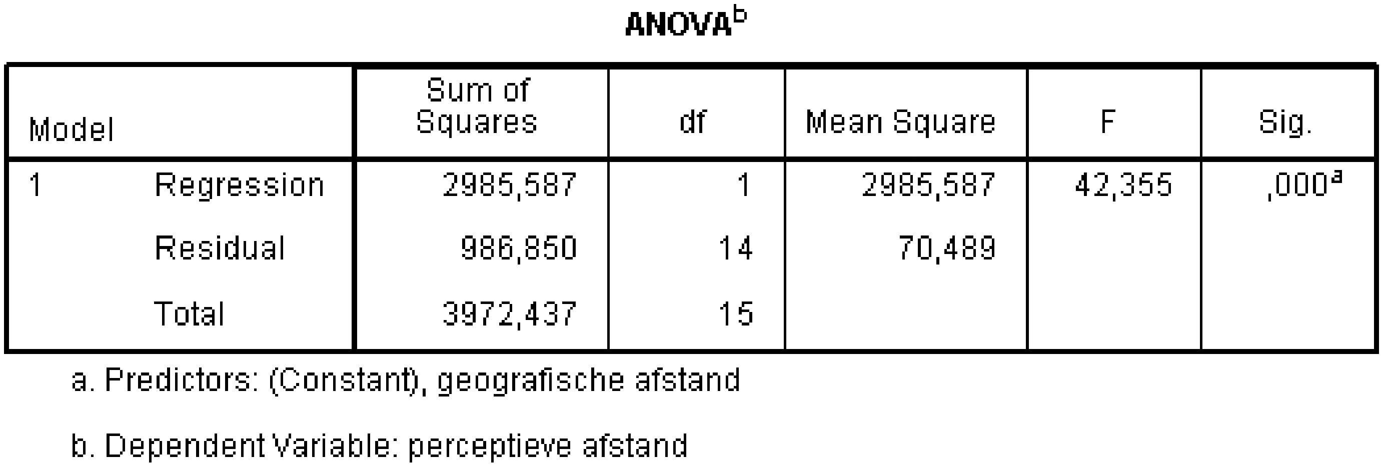
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3184,545	2	1592,273	26,272	,000[a]
	Residual	787,892	13	60,607		
	Total	3972,437	15			

a. Predictors: (Constant), uitspraakverschil, geografische afstand

b. Dependent Variable: perceptieve afstand

The ANOVA table in SPSS ('enter').

ANOVA



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2985,587	1	2985,587	42,355	,000[a]
	Residual	986,850	14	70,489		
	Total	3972,437	15			

a. Predictors: (Constant), geografische afstand

b. Dependent Variable: perceptieve afstand

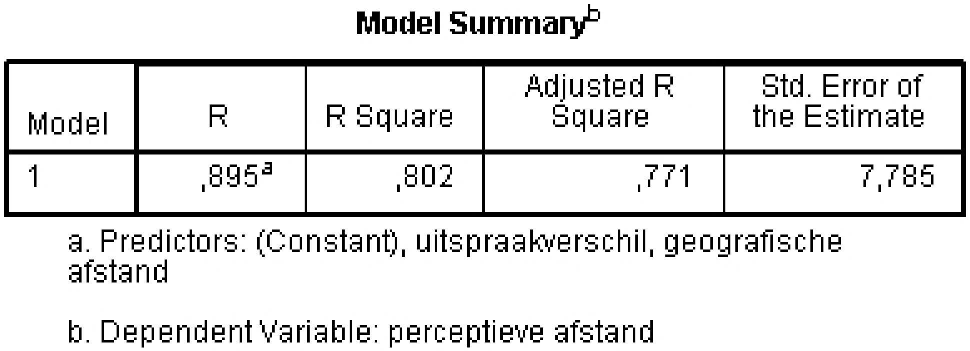
The ANOVA table in SPSS ('stepwise'). The procedure excludes *pronunciation distance*.

Effect size

- In a simple regression analysis the response variable y is correlated with the explanatory variable x.
- The squared correlation gives the amount of variance in y explained by x.
- The correlation between the explanatory variable x and the predicted values $\hat{y}$ always is 1.
- Therefore: correlation y with x gives the same result as correlating y with $\hat{y}$ in a simple linear regression analysis.
- In a multiple regression model the determination coefficient R^2 is the proportion of variance in the response variable y explained by the explanatory variables $x_1, x_2, \ldots, x_p$.
- R^2 is calculated as:

$$R^2 = \frac{variance\ of\ predicted\ values\ \hat{y}}{variance\ of\ observed\ values\ y} = \frac{SSM}{SST}$$

Multiple correlation

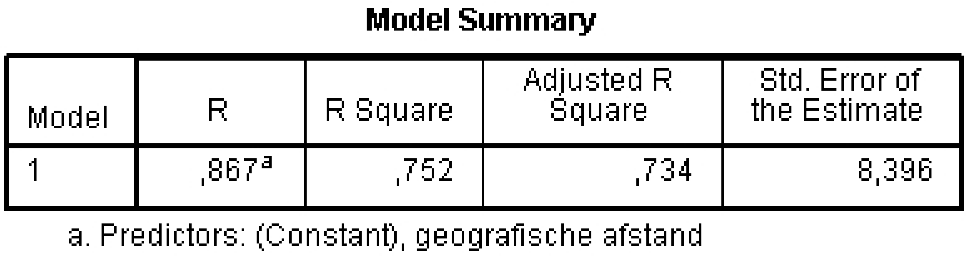


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,895[a]	,802	,771	7,785

a. Predictors: (Constant), uitspraakverschil, geografische afstand

b. Dependent Variable: perceptieve afstand



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,867[a]	,752	,734	8,396

a. Predictors: (Constant), geografische afstand

The (squared) correlation coefficients of the all-at-once (upper) and stepwise (lower) regression analysis in the output of SPSS. In the all-at-once analysis only we find a multiple correlation.