Sommersemester 2009
Statistik mit SPSS

Überblick

1. Korrelation vs. Regression
2. Ziele der Regressionsanalyse
3. Syntax für den Regressionsbefehl
4. Interpretation
   - 4.a. Modellzusammenfassung: R, R^2
   - 4.b. Anova-Tabelle: Varianzzerlegung
   - 4.c. Koeffizientenblock

1. Korrelation vs. Regression

a) Korrelation: symmentrisches Maß

„Größe“  „Gewicht“

b) Regression (Einfachregression): asymmetrisches Maß

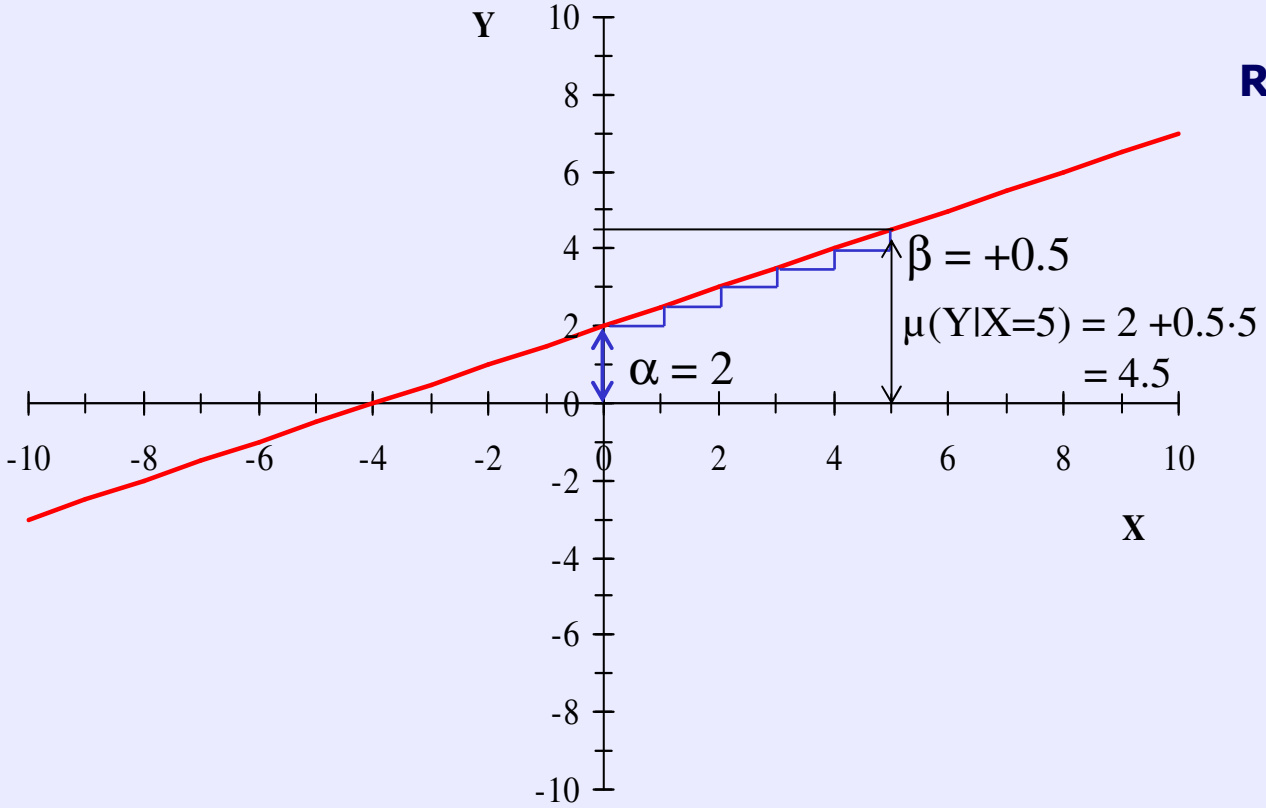
„ Größe“ (UV)  „Gewicht“ (AV)

2. Ziele der Regressionsanalyse

Durch die Regressionsanalyse wird eine Variable (AV) durch eine oder mehrere andere Variablen (UV) **erklärt** oder **prognostiziert**.

Die AV und die UV sind **metrische Variablen**. Ausnahmsweise werden als UV Dummy-Variablen zugelassen (0-1 codiert). Nominale oder ordinale Variablen können in Dummy-Variablen rekodiert werden und als UV analysiert werden.

Grundvoraussetzung ist die **Linearitätsannahme**, nach der das lineare Regressionsmodell in der Population für alle Ausprägungen der UV gilt.



Regressionsgerade mit

Regressionskonstante $\alpha = 2$

Regressionsgewicht $\beta = 0.5$

$\mu(Y|X) = 2 + 0.5 \cdot X.$

Wenn X = 0, Y = 2.

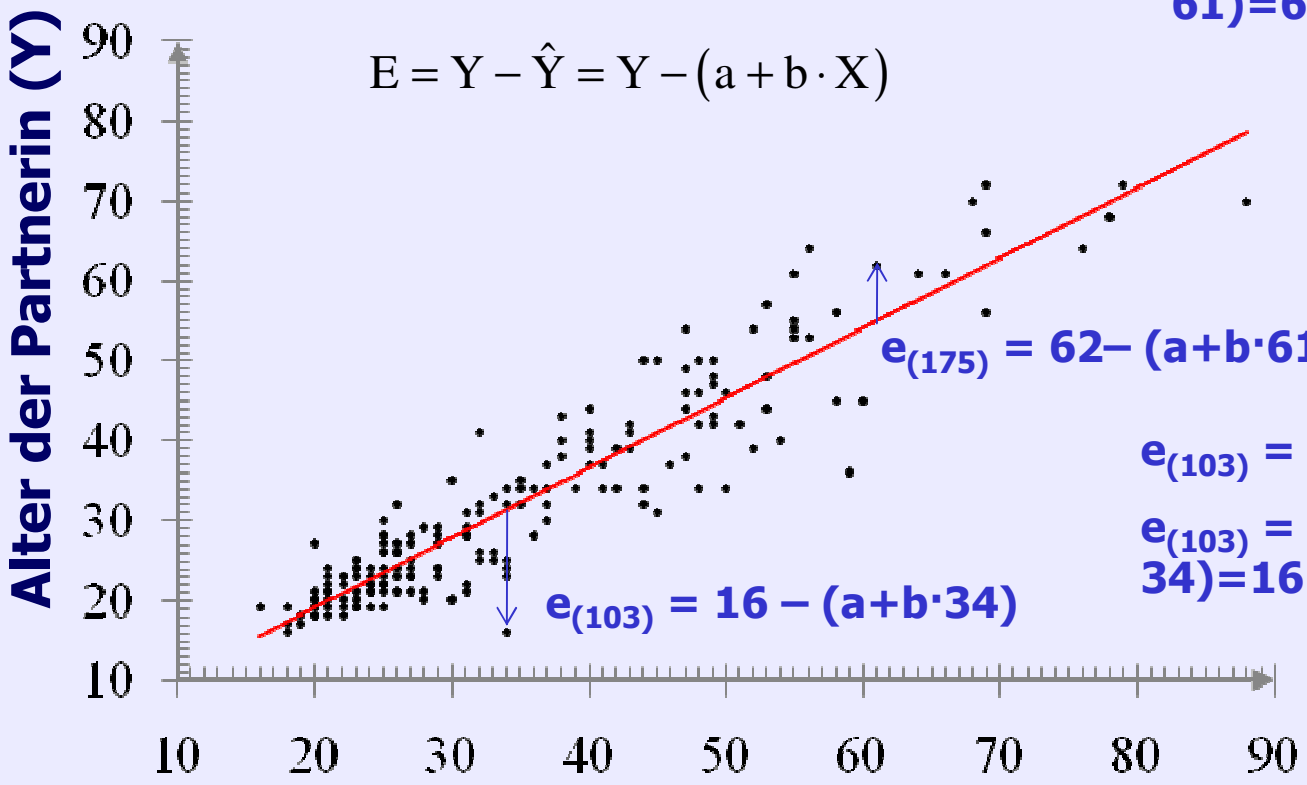
Wenn X um +1 Einheit ansteigt, steigt Y um +0.5 Einheiten an.

2. Ziele der Regressionsanalyse

Es wird angenommen, dass sich die bedingten Populationsmittelwerte der abhängigen Variable durch eine lineare Funktion der Ausprägungen der erklärenden Variable beschreiben lassen.

$\hat{Y}_i=f(X_i)=a+bX_i$	**Vorhersage/ Prognose**
a	**ist die *Regressionskonstante* und gibt den Mittelwert von Y an, wenn X=0**
b	**ist das *Regressionsgewicht* und gibt die Steigung der Geraden an**
$Y_i=a+bX_i+ e_i$	**Tatsächliche Werte**
$e_i = Y_i-\hat{Y}_i$	**Residuen oder nicht erklärte Abweichung**

$$E = Y - \hat{Y} = Y - (a + b \cdot X)$$



$e_{(175)} = 62 - (a+b \cdot 61)$

$e_{(175)} = 62 - (1.623 + 0.876 \cdot 61) = 62 - 55{,}06 = 6{,}94$

$e_{(175)} = 62 - (a+b \cdot 61)$

$e_{(103)} = 16 - (a+b \cdot 34)$

$e_{(103)} = 16 - (a+b \cdot 34)$

$e_{(103)} = 16 - (1.623 + 0.876 \cdot 34) = 16 - 31{,}41 = -15{,}41$

2. Ziele der Regressionsanalyse

Die Regressionsanalyse ermöglicht Aussage zu folgenden Fragestellungen:

- Wie stark ist der Einfluss einer einzelnen unabhängigen Variablen auf die abhängige Variable?
- Wie ändert sich die abhängige Variable bei einer Veränderung der unabhängigen Variablen (Prognose)?
- Bei multipler Regression: Wie groß ist die Erklärungskraft aller unabhängigen Variablen zusammen?

3. Syntax für den Regressionsbefehl

```
REGRESSION [MATRIX=[IN(file)] [OUT(file)]]
  [/VARIABLES={varlist    }]
              {(COLLECT)**} Command Syntax
              {ALL        }
  [/DESCRIPTIVES=[DEFAULTS][MEAN][STDDEV][CORR][COV]
                 [VARIANCE][XPROD][SIG][N][BADCORR]
                 [ALL][NONE**]]
  [/SELECT={varname relation value}
  [/MISSING=[{LISTWISE**      }] [INCLUDE]]
            {PAIRWISE         }
            {MEANSUBSTITUTION}
  [/REGWGT=varname]
  [/STATISTICS=[DEFAULTS**][R**][COEFF**]
               [ANOVA**][OUTS**]
               [ZPP][CHA][CI][F][BCOV]
               [SES][XTX][COLLIN]
               [TOL][SELECTION][ALL]]
  …
  /DEPENDENT=varlist
  [/METHOD=]{STEPWISE [varlist]       }
            {FORWARD [varlist]        }
            {BACKWARD [varlist]       }
            {ENTER [varlist]          }
            {REMOVE varlist           }
            {TEST(varlist)(varlist)...}
  [/RESIDUALS=[DEFAULTS][ID(varname
```

4. Interpretation

Einfachregression mit SPSS

➔ Regression des Gewichtes (AV) auf die Körpergröße (UV).

4.a. Ausgabe Modellzusammenfassung: R

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,513[a]	,263	,263	13,288

a. Einflußvariablen : (Konstante), v305 KOERPERGROESSE IN CM, BEFRAGTE<R>

R entspricht im bivariaten Modell dem Korrelationskoeffizienten (Pearson's r)

4.a. Ausgabe Modellzusammenfassung: R^2

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,513[a]	,263	,263	13,288

a. Einflußvariablen : (Konstante), v305 KOERPERGROESSE IN CM, BEFRAGTE<R>

R^2: Determinationskoeffizient (Bestimmtheitsmaß)

- Er gibt Auskunft über die „Güte" des Modells (Maß für die Erklärungskraft des Modells insgesamt), d.h.
- R^2 gibt an, wie viel Varianz der abhängigen Variablen durch die unabhängige Variable erklärt wird,
- hier wird 26,3 % der Varianz des Gewichtes (AV) durch die Varianz der Körpergröße (UV) „erklärt".

4.a. Ausgabe Modellzusammenfassung: R^2 als PRE-Maß

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,513[a]	,263	,263	13,288

a. Einflußvariablen : (Konstante), v305 KOERPERGROESSE IN CM, BEFRAGTE<R>

R^2 als PRE-Maß:

Bei Kenntnis der Var. „Körpergröße" (UV) werden die Vorhersagefehler im Vergleich zur Vorhersage des Gewichtes (AV) ohne Kenntnis dieser Var. um 26,3% reduziert.

Standardfehler: bestimmt die Genauigkeit der Schätzung.

4.b. Ausgabe Anova-Tabelle: Varianzzerlegung

Wie gut beschreibt das Regressionsmodell die beobachteten Daten?

- Die Güte oder Erklärungskraft des Regressionsmodells (R^2) kann mithilfe der Varianzanalyse geschätzt werden.

- In der Varianzanalyse wird die Gesamtvarianz der abhängigen Variable zerlegt in einen „erklären" Teil und in einen „unerklärten" Teil der Varianz bzw. der Variation.

- Die erklärte Varianz ist der Teil der Varianz, der durch Kenntnis der Werte der unabhängigen Variable ‚erklärt' wird.

- D.h.: je größer der Anteil der erklärten Varianz, desto größer ist R^2 und damit die Erklärungskraft des Regressionsmodells.

4.b. Ausgabe Anova-Tabelle: Varianzzerlegung

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	181534,306	1	181534,306	1028,092	,000[a]
	Residuen	509239,227	2884	176,574		
	Gesamt	690773,533	2885			

a. Einflußvariablen : (Konstante), v305 KOERPERGROESSE IN CM, BEFRAGTE<R>

b. Abhängige Variable: v307 GEWICHT IN KG, BEFRAGTE<R>

Gesamt: Gesamtstreuung

Regression: Anteil der Gesamtstreuung, der durch das Regressionsmodell erklärt wird

Residuen: Anteil der Gesamtstreuung, der durch das Regressionsmodell nicht erklärt wird

4.b. Ausgabe Anova-Tabelle: Varianzzerlegung

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	181534,306	1	181534,306	1028,092	,000[a]
	Residuen	509239,227	2884	176,574		
	Gesamt	690773,533	2885			

a. Einflußvariablen : (Konstante), v305 KOERPERGROESSE IN CM, BEFRAGTE<R>

b. Abhängige Variable: v307 GEWICHT IN KG, BEFRAGTE<R>

F-Test:

Das Regressionsmodell ist hochsignifikant, d.h. der Determinations-koeffizient ist in der Grundgesamtheit mit einer Irrtumswahrscheinlichkeit von weniger als 0,1% ($p < 0,001$) von null verschieden.

Irrtumswahrscheinlichkeit

Irrtumswahrscheinlichkeit		Bedeutung	Symbol
P> 0,05	(>5%)	nicht signifikant	n.s.
P<= 0,05	(<= 5%)	signifikant	*
P<= 0,01	(<= 1%)	sehr signifikant	**
P<= 0,001	(<= 0,1%)	höchst signifikant	***

4.c. Ausgabe Koeffizientenblock

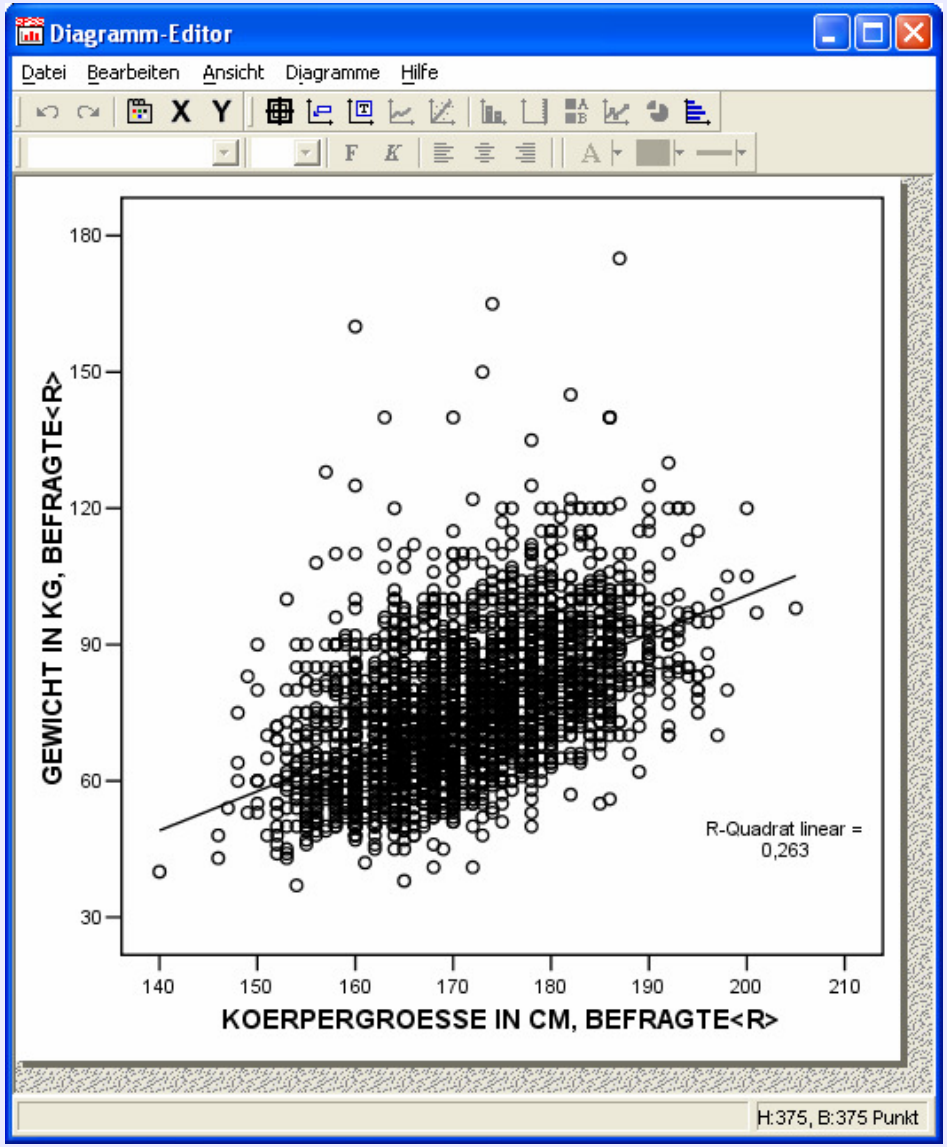
Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	-71,617	4,613		-15,526	,000
	v305 KOERPERGROESSE IN CM, BEFRAGTE<R>	,862	,027	,513	32,064	,000

a. Abhängige Variable: v307 GEWICHT IN KG, BEFRAGTE<R>

Regressionskonstante:

Schnittpunkt der Regressionsgerade mit der y-Achse, wenn x=0 (s. Streudiagramm).



Konstante:

Schnittpunkt der Regressionsgerade mit der y-Achse, wenn x=0 (s. Streudiagramm).

4.c. Ausgabe Koeffizientenblock

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	-71,617	4,613		-15,526	,000
	v305 KOERPERGROESSE IN CM, BEFRAGTE<R>	,862	,027	,513	32,064	,000

a. Abhängige Variable: v307 GEWICHT IN KG, BEFRAGTE<R>

Unstandardisierter Regressionskoeffizient (Regressionsgewicht/Steigung):

Wenn X (UV) um eine Einheit ansteigt, dann steigt Y (AV) um 0,86 Einheiten an.

D.h. hier: Wenn die Körpergröße um einen Zentimeter ansteigt (cm als Einheit), dann steigt das Körpergewicht im Durchschnitt um 0,86 kg an.

4.c. Ausgabe Koeffizientenblock

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	-71,617	4,613		-15,526	,000
	v305 KOERPERGROESSE IN CM, BEFRAGTE<R>	,862	,027	,513	32,064	,000

a. Abhängige Variable: v307 GEWICHT IN KG, BEFRAGTE<R>

T-Test: Dient der Signifikanzprüfung des Regressionskoeffizienten.

Signifikanz: empirische Signifikanz, d.h. mit einer Irrtumswahrscheinlichkeit von <0,1% ($p < 0,001$) ist der Regressionskoeffizient von Null verschieden (oder: der Koeffizient ist höchst signifikant).

Zusammenfassung Einfachregression

R^2 :	**Anteil erklärter Varianz, PRE-Maß**
F-Test und Sign.:	**testet Nullhypothese, dass der Determinationskoeffizient in der GG null ist; Signifikanz des Modells.**
Konstante:	**Schnittpunkt der Regressionsgerade mit der y-Achse, wenn x=0**
B:	**Stärke des Effektes von X (UV) auf Y (AV),**
T-Test und Sign.:	**testet Nullhypothese, dass der Regressions-koeffizient in der GG null ist; Signifikanz der Regressionskoeffizienten**

Weitere Beispiele mit dem Allbus 2004

Morgen: Klausurvorbereitung

In der Übung behandelte SPSS-Befehle (jeweils inkl. Unterbefehle)

DATENTRANSFORMATION

FREQUENCIES
VARIABLE LABELS
VALUE LABELS
MISSING VALUES
COMMENT
RECODE
EXAMINE
EXECUTE
COMPUTE
IF
COUNT
DISPLAY DICTIONARY
RENAME

DATENMANAGEMENT

SELECT IF
SORT CASES
SPLIT FILE
WEIGHT
TEMPORARY

DATENANALYSE

GRAPH
T-TEST
CROSSTABS
CORRELATIONS
REGRESSION