Lineare Regressionsanalyse

Lineare Regressionsanalyse

Das Hauptanliegen der Regressionsanalyse ist die Vorhersage von Variablenwerten in der abhängigen Variablen Y auf Grund der Kenntnis der Variablenausprägungen in der bzw. den unabhängigen Variable(n) X bzw. X_1, X_2, ... (bspw. die Vorhersage von Verkaufszahlen auf Grund der Häufigkeit gesendeter Werbespots). Im bivariaten Fall wird von einer unabhängigen Variable X auf eine abhängige Variable Y geschlossen. Im multivariaten Fall wird die Ausprägung der abhängigen Variable Y durch die Linearkombination mehrerer unabhängiger Variablen X_1, X_2,... vorhergesagt. Die Kausalitätsrichtung (Bestimmung der Variablen als unabhängig, bzw. abhängig) wird theoretisch festgelegt und geht der Regressionsanalyse voraus. Desweiteren wird eine lineare Beziehung zwischen der bzw. den unabhängigen Variable(n) und der abhängigen Variable vorausgesetzt.

1 Bivariate Regression

Um die Ausprägung der abhängigen Variable (Y) auf Grund der Ausprägung der unabhängigen Variable (X) vorherzusagen (diese Schätzung wird im folgenden $\hat{Y}$ genannt), muß ein funktionaler Zusammenhang spezifiziert werden. Im einfachsten Fall wird ein solcher Zusammenhang durch eine lineare Funktion, also durch eine Gerade dargestellt:

$$\hat{Y} = aX + b.$$

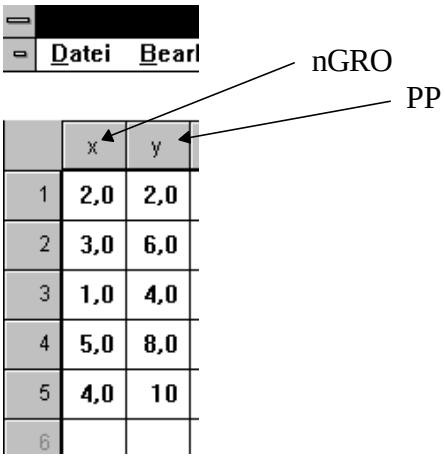
Der empirisch gegebene Zusammenhang zwischen den Variablen X und Y kann durch eine Punktwolke in einem Diagramm veranschaulicht werden, das Streudiagramm genannt wird. Die einzelnen Punkte werden anhand der in der Datenmatrix enthaltenen (X,Y)-Koordinaten dargestellt. Die Koeffizienten der Geradengleichung (a und b) werden nun so geschätzt, daß die Abweichungen der tatsächlichen Y-Werte von den durch die Regressionsgerade geschätzten $\hat{Y}$ - Werten minimal wird. Dies geschieht mit Hilfe der Methode der kleinsten Quadrate, auch OLS-Schätzung (Ordinary Least Squares) genannt (vgl. Urban, 1982, S. 41).

Für die bivariate Regression müssen folgende Voraussetzungen erfüllt sein:
- X, Y müssen intervallskaliert sein (metrische Variablen)
- Das lineare Modell muß die Beziehung zwischen X und Y adäquat abbilden (d.h. die Form der Beziehung zwischen X und Y ist linear.
- Um statistische Tests durchführen zu können, müssen zusätzliche Voraussetzungen erfüllt sein (siehe Kap. 2.2.).

Bevor die Regression berechnet wird, sollte überprüft werden, inwieweit der Zusammenhang zwischen X und Y durch eine Gerade (lineare Funktion) hinreichend genau approximiert werden kann (z.B. Sichtprüfung des Streudiagramms; vgl. Kap. 1.1.1.).

1.1 Beispiel: Übungsaufgabe I.1 (Skript, Anhang, S.1)

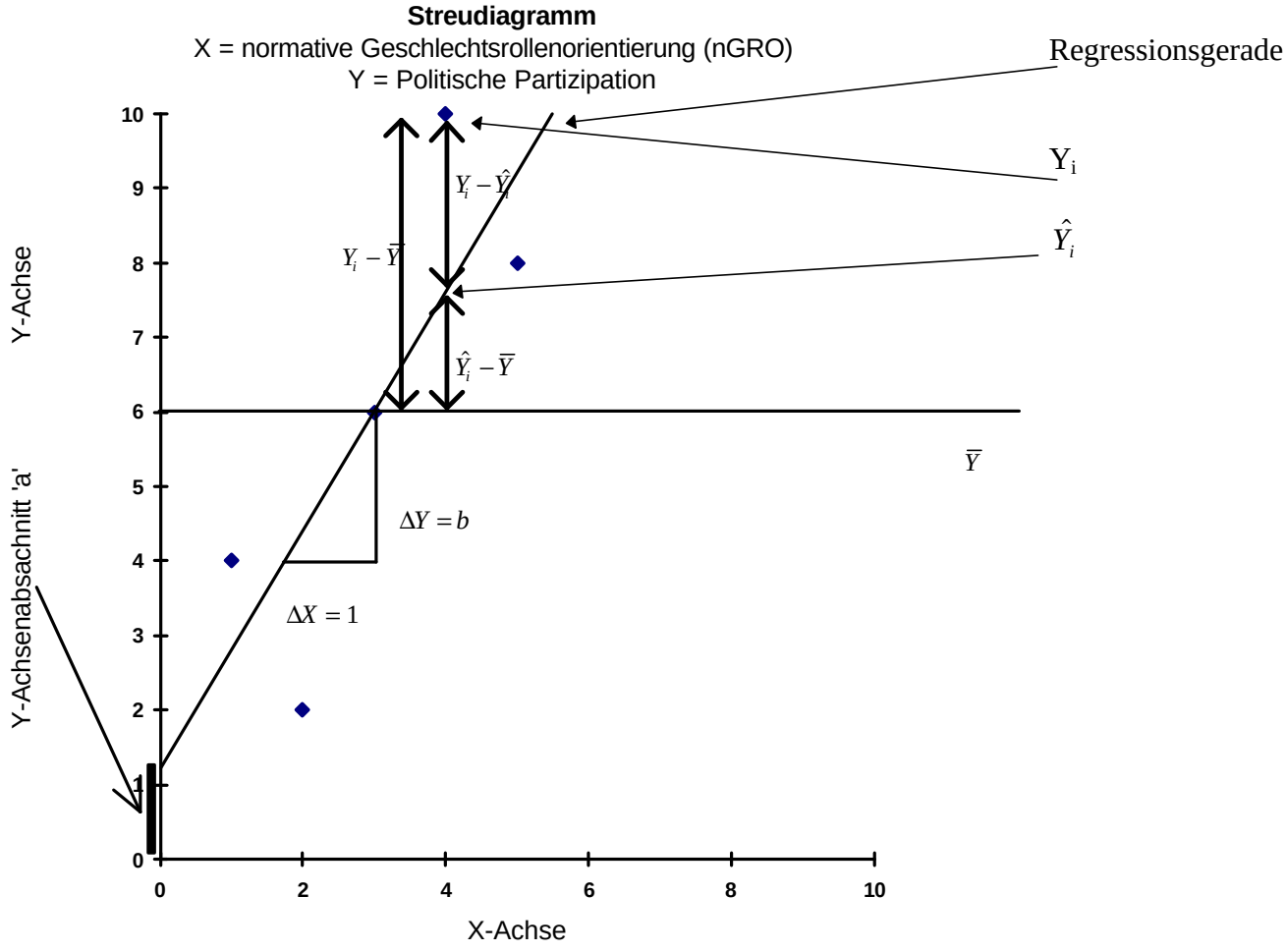
Erster Arbeitsschritt ist die Eingabe der Datentabelle, die diesem Beispiel zugrunde liegt.



	x	y
1	2,0	2,0
2	3,0	6,0
3	1,0	4,0
4	5,0	8,0
5	4,0	10
6		

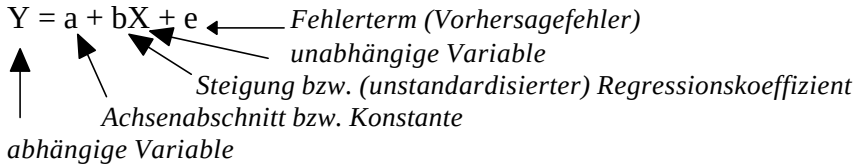
1.1.1 Darstellung der Regressionsgeraden im Streudiagramm

Das folgende Streudiagramm (anforderbar im SPSS-Menü GRAFIK; SCATTERPLOT) veranschaulicht die hypothetische Beziehung zwischen den Variablen X (normative Geschlechtsrollenorientierung) und Y (Politische Partizipation).



Wie man sieht, ist die Beziehung annähernd linear. Wir betrachten es daher als gerechtfertigt, von einer linearen Beziehung zwischen der nGRO und der PP auszugehen. Die lineare Beziehung zwischen den beiden Variablen kann durch eine Gerade dargestellt werden, die zusätzlich in das Streudiagramm eingezeichnet wurde (In SPSS existiert die Möglichkeit, erstellte Grafiken (Charts) anzuzeigen und mit der Option BEARBEITEN zu ergänzen).

Die Gleichung der Regressionsgeraden lautet
(Diese und die nachfolgenden Formeln sind im Skript zu finden, S. 40 ff.)



bzw.

$$\hat{Y} = a + bX$$

wobei $\hat{Y}$ den bei Kenntnis von X durch die Regressionsgleichung geschätzten Wert von Y darstellt. Die Abweichung $Y_i - \hat{Y}_i$ stellt den Fehlerterm (e) der Regressionsgleichung dar.

Um eine bestmögliche Anpassung der Punkteansammlung durch die Regressionsgerade zu erreichen, muß der Vorhersagefehlerder Regression minimiert werden. Eine Gerade, die diesen Anforderungen entspricht läßt sich mit der *Methode der kleinsten Quadrate* (OLS) berechnen. Hiermit ist gemeint, daß die Summe der quadrierten Abstände zwischen allen Punkten und der Geraden minimal wird (formal: Minimierung der Summe der Abweichungsquadrate [SAQ]).

$$\text{SAQ (Vorhersagefehler)} = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 \qquad \Rightarrow \min !$$

Die Regressionskonstante a gibt den Schnittpunkt mit der Y-Achse an (X=0). Der Regressionskoeffizient b gibt an, um wieviele Einheiten sich die abhängige Variable verändert, wenn sich die unabhängige Variable um eine Einheit verändert ($\Delta Y = b\Delta X$, vgl. das obige Schaubild).

1.1.2 Durchführung der Regression

Die folgende Syntax-Datei wird von SPSS für Windows „automatisch“ erzeugt, wenn man eine Regression mit Y als abhängiger und X als unabhängiger Variable anfordert.

Syntax-Datei

```
-> REGRESSION
->   /MISSING LISTWISE
->   /STATISTICS COEFF OUTS R ANOVA
->   /CRITERIA=PIN(.05) POUT(.10)
->   /NOORIGIN
->   /DEPENDENT y
->   /METHOD=ENTER x  .
```

Output-Datei

```
        * * * *  M U L T I P L E  R E G R E S S I O N  * * * *


Listwise Deletion of Missing Data
Equation Number 1    Dependent Variable..   Y   Politische Partizipation
Block Number  1.  Method:  Enter      X

Variable(s) Entered on Step Number
   1..    X         normative Geschlechtsrollenorientierung

Multiple R           ,80000
R Square             ,64000
Adjusted R Square    ,52000
Standard Error      2,19089

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           1            25,60000         25,60000
Residual             3            14,40000          4,80000

F =       5,33333       Signif F =  ,1041


------------------ Variables in the Equation ------------------
Variable              B        SE B       Beta         T  Sig T
X             1,600000     ,692820    ,800000     2,309  ,1041
(Constant)    1,200000    2,297825                ,522  ,6376
```

Erläuterung der SPSS-Ausgabedatei

Der <u>Multiple Korrelationskoeffizient</u> (Multiple R = .8) ist im Falle der bivariaten Regression mit dem einfachen Korrelationskoeffizienten (r_{XY}) identisch (für den Fall der multiplen Regression vgl. Kap. 2.1.).

<u>R-Square</u> ist das Quadrat des multiplen Korrelationskoeffizienten und damit ein Maß für den Anteil der Varianz von Y, der durch X erklärt wird (zur varianzanalytischen Interpretation von R-Square siehe unten).

<u>Adjusted R-Square</u> stellt das korrigierte *R-Square* dar, in das die Größe des Stichprobenumfanges und die Anzahl der unabhängigen Variablen eingehen. Bei steigendem Stichprobenumfang nähert sich das korrigierte R-Square an das unkorrigierte R-Square an. Eine steigende Anzahl unabhängiger Variablen wirkt sich gegenteilig aus.

Varianzanalytische Interpretation der Regression:
In der Rubrik *Analysis of Variance* werden die Summen der Abweichungsquadrate SAQ (*Sum of Squares*) mitgeteilt. Die Fehlervarianz [SAQ (Residual)] wird aufgrund der Differenzen zwischen den empirisch beobachteten Werten der abhängigen Variable (Y_i) und den durch die Regressionsgerade geschätzten Werten ($\hat{Y}_i$) berechnet. Die Residuen (Fehlerterme) lassen sich gut anhand des Streudiagramms als Streuung der Punkte um die Regressionsgerade veranschaulichen. Das *Sum of Squares* der Residuen (14,4) kann als Maß für den nicht erklärten Teil der Gesamtvariation aufgefasst werden.

$$SAQ(\mathrm{Re}\,sidual) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = 14{,}4$$

Genau diese Größe wurde durch die Methode der kleinsten Quadrate minimiert. Der erklärte Teil der Variation der abhängigen Variable (25,6) ist unter *Regression* zu finden.

$$SAQ(\mathrm{Re}\,gression) = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 = 25{,}6$$

Erklärte und nicht erklärte Variation ergeben die Gesamtvariation (vgl. auch Streudigramm).

$$SAQ(Gesamt) = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = 40{,}0$$

Damit ergibt sich der Anteil der erklärten Varianz (identisch mit dem Anteil der erklärten Variation an der Gesamtvariation):

$$R - Square = \frac{erklärte\ \ Variation}{Gesamt\ \mathrm{var}\,iation} = \frac{25{,}6}{25{,}6 + 14{,}4} = 0{,}64$$

Prüfwert 'F':
Wenn die entsprechenden Voraussetzungen für einen Signifikanztest erfüllt sind (vgl. Kap. 2.2), kann der F-Wert dazu benutzt werden, zu testen, ob die Varianzerklärung von Y signifikant ist

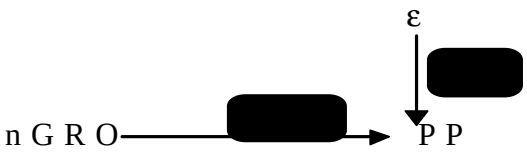
In der Rubrik *Variables in the Equation* werden in der Spalte *B* der Regressionskoeffizient für die unabhängige Variable und die Konstante - also die Koeffizienten b und a, die die Regressionsgerade definieren - ausgewiesen. Die Regressionskoeffizient b läßt sich dahingehend interpretieren, daß bei einer Änderung der unabhängigen (erklärenden) Variable um eine Einheit, die abhängige Variable um den Wert des Regressionskoeffizienten zunimmt (Steigung der Regressionsgeraden). Auf unser Beispiel bezogen läßt sich damit sagen, daß bei Änderung der Variable X (nGRO) um eine Einheit, die abhängige Variable Y (PP) um 1,6 Einheiten zunimmt. Die Konstante a=1,2 gibt den Schnittpunkt mit der der y-Achse an (vgl. Streudiagramm)

$$\hat{Y} = a + bX = 1{,}2 + 1{,}6X$$

In der Spalte '*SE B*' wird der Standardfehler des Regressionskoeffizienten wiedergegeben. Er bildet die Grundlage für einen Signifikanztest (vgl. Kap. 2.2).

Der *Beta*-Koeffizient ist der standardisierte Regressionkoeffizient. Für seine Berechnung werden die Werte der X- und der Y-Variablen zunächst standardisiert (z-transformiert; vgl. Skript S. 41 und ausführlicher Urban, S. 58-72). Beide Variablen haben nach der Transformation den Mittelwert '0' und die Standardabweichung '1', die Skalenausprägungen sind somit vergleichbar. Der unter dieser Voraussetzung berechnete standardisierte Regressionskoeffizient (Beta-Koeffizient) liegt in einem Wertebereich zwischen -1 und +1. Im bivariaten Fall ist der standardisierte Regressionskoeffizient mit der Produkt-Moment-Korrelation identisch ($b = r_{XY} = R$)

Bezogen auf eine pfadanalytische Interpretation des Regressionsmodells gilt, daß im bivariaten Fall der Pfadkoeffizient p_{yx} gleich dem standardisierten Regressionskoeffizienten b und damit gleich dem bivariaten Korrelationskoeffizienten r_{XY} ist (vgl. Skript S. 41).



[Pfaddiagramm]

$$y = p_{yx}\, x + p_{y\varepsilon}\varepsilon$$ [Strukturgleichung]

Der T-Wert (Spalte *T*) der Regressionskoeffizienten gibt den Quotienten aus Regressionskoefizienten (*B*) und Standardfehler (*SE B*) an und dient als Grundlage für einen Signifikanztest (t-Test) des Regressionskoeffizienten (vgl. 2.2).

1.2 Beispiel: Scherhorn-Studie

Hier soll die folgende Hypothese untersucht werden: Je höher die materielle Gütergebundenheit (GTGB), desto stärker die Positionalität (POSIT)

Die Daten stammen aus einer Studie von G. Scherhorn zur Kaufsucht (1991; vgl. Anhang zu Kap. 1 im Skript). Aus den Items, die materielle Gütergebundenheit und Positionalität messen, wurden zuvor jeweils eine Likert-Skala gebildet.

Syntax-Datei

```
-> REGRESSION
->   /MISSING LISTWISE
->   /STATISTICS COEFF OUTS R ANOVA
->   /CRITERIA=PIN(.05) POUT(.10)
->   /NOORIGIN
->   /DEPENDENT POSIT
->   /METHOD=ENTER GTGB  .
```

Output-Datei

```
            * * * *   M U L T I P L E   R E G R E S S I O N   * * * *


Listwise Deletion of Missing Data

Equation Number 1    Dependent Variable..   POSIT   POS
Block Number  1.  Method:  Enter      GTGB

Variable(s) Entered on Step Number
   1..    GTGB      GB

Multiple R           ,55103
R Square             ,30363
Adjusted R Square    ,30313
```

Standard Error	10,71849

```
Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           1         70429,41298      70429,41298
Residual          1406        161529,63177        114,88594

F =     613,03770       Signif F =  ,0000


------------------ Variables in the Equation ------------------

Variable             B        SE B       Beta         T  Sig T

GTGB            ,603760     ,024385    ,551025    24,760  ,0000
(Constant)    24,224261    1,075121               22,532  ,0000


End Block Number   1   All requested variables entered.
```

Interpretation der Ergebnisse:

Da es sich bei diesem Beispiel um eine bivariate Regression handelt, ist der standardisierte Reressionskoeffizient (Beta) mit dem Multiplen Korrelationskoeffizienten (Multiple R=r_{XY}) identisch (Beta = Multiple R = .551025). Der Zusammenhang beider Variablen ist als hoch zu bewerten und statistisch abgesichert, d.h. signifikant (Sig T = 0,0000; zur Siginifikanz vgl. Kap. 2.2). Immerhin erklärt die Variation der Werte in der unabhängigen Variablen (GTGB) 30,363% der Variation der Werte in der abhängigen Variablen (POSIT).

nicht erklärte Variation: $SAQ(\text{Re}\,sidual) = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = 161529{,}63$

erklärte Variation $SAQ(\text{Re}\,gression) = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = 70429{,}41$

Anteil der durch die Regression erklärten Variation (erklärte Varianz):

$$R - Square = \frac{erklärte\ Variation}{Gesamt\ \text{var}\,iation} = \frac{70429{,}41}{70429{,}41 + 161529{,}63} = 0{,}30363$$

[Pfaddiagramm]

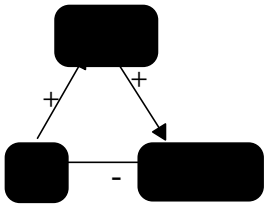
2 Multivariate Regression

Im Gegensatz zur bivariaten Regression, wo nur eine erklärende Variable existiert, werden bei der multivariaten Regression zwei oder mehr unabhängige Variablen (X_1, X_2, ...) in das Regressionsmodell integriert.
Damit entfällt die anschauliche Darstellung von abhängiger und unabhängiger Variable anhand eines zweidimensionalen Streudiagramms.Die Regressionsgleichung hat folgende formale mathematische Form (vgl. Skript S. 43 ff.):

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n \qquad \text{bzw.}$$
$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

2.1 Beispiel: Scherhorn-Studie

Wir veranschaulichen die Durchführung einer multivariaten Regression, indem wir unsere bivariate Hypothese „Je höher die materielle Gütergebundenheit (GTGB), desto stärker die Positionalität (POSIT)“ um eine zusätzliche Variable erweitern. Die Variable Geschlecht (f109) wird als antezedierende Variable berücksichtigt. Die weiteren Hypothesen lauten:
„Frauen sind weniger positional als Männer“
„Frauen sind gütergebundener als Männer“
In einem Kausalmodell können die hypothetischen Zusammenhänge folgendermaßen dargestellt werden.



bzw. $\widehat{POSIT} = a + b_1 GTGB + b_2\, f109$

Zunächst wird die postulierte bivariate Beziehung zwischen der unabhängigen (antezedierenden) Variablen Geschlecht (f109) und Gütergebundenheit (GTGB) überprüft.

Geschlecht als unabhängige *Dummy-Variable:*
Eine Besonderheit in diesem Regressionsmodell stellt die Variable f109 (Geschlecht) dar. Hier handelt es sich um eine Variable mit lediglich zwei Kategorien (Dichotomie). Damit erfüllt sie nicht die Voraussetzung für eine Regression, in der für die beteiligten Variablen Intervallskalenniveau gefordert wird. Jedoch können Dichotomien als unabhängige Variablen in das Regressionsmodell einbezogen werden, wenn man sie als sogenannte Dummy-Variablen behandelt. Hierbei werden den beiden Kategorien numerische Werte als Voraussetzung für statistische Auswertungen zugeordnet. In unserem Beispiel haben wir der Kategorie „männlich“ eine „1“ und der Kategorie „weiblich“ eine „2“ zugeordnet.

Syntax-Datei

```
-> REGRESSION
->   /MISSING LISTWISE
->   /STATISTICS COEFF OUTS R ANOVA
->   /CRITERIA=PIN(.05) POUT(.10)
->   /NOORIGIN
->   /DEPENDENT gtgb
->   /METHOD=ENTER f109  .
```

Output-Datei

```
          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Listwise Deletion of Missing Data
Equation Number 1    Dependent Variable..   GTGB   GB
Block Number  1.  Method:  Enter      F109

Variable(s) Entered on Step Number
   1..    F109      GESCHLECHT
Multiple R           ,16605
R Square             ,02757
Adjusted R Square    ,02691
Standard Error     11,55159

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           1          5569,34972       5569,34972
Residual          1472        196422,54987        133,43923

F =      41,73697       Signif F =  ,0000


------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

F109           3,894251     ,602786    ,166049     6,460  ,0000
(Constant)    36,460216     ,969628               37,602  ,0000


End Block Number   1   All requested variables entered.
```

<u>Interpretation des Regressionskoeffizienten B für die Dummy-Variable Geschlecht (f109):</u>
Grundsätzlich ist bei der Interpretation von Dummy-Variablen auf die Kodierung zu achten:
1 = männlich
2 = weiblich
Inhaltlich besagt der Regressionskoeffizient von B = 3.89, daß Frauen materiell gütergebundener sind als Männer. Der Anstieg der Dummy-Variablen um eine Einheit bedeutet einen Anstieg von 1 auf 2, was inhaltlich dem Wechsel in die andere Kategorie entspricht, nämlich dem Wechsel von „männlich“ zu „weiblich“. Die Frauen haben demnach auf der Skala für materielle Gütergebundenheit einen im Durchschnitt um 3.89 Einheiten höheren Wert, sind also gütergebundener als die Männer.

Insgesamt ist der Einfluß der unabhängigen Variablen Geschlecht (f109) auf die abhängige Variable Gütergebundenheit (GTGB) - im Sinne einer linearen Beziehung- eher als gering zu bezeichnen (Beta = .16) und erklärt lediglich 2,75% der Varianz in der abhängigen Variablen (vgl. R-Square). Dennoch ist die Beziehung, wie die F-Statistik der Varianzanalyse anzeigt, hochsignifikant (vgl. Kap. 2.2).
Zur Überprüfung der weiteren Hypothesen wird eine Regression gerechnet, in der Geschlecht (f109) und materielle Gütergebundenheit (GTGB) als unabhängige und Positionalität (POSIT) als abhängige Variablen angegeben sind.

Syntax-Datei

```
-> REGRESSION
->   /MISSING LISTWISE
->   /STATISTICS COEFF OUTS R ANOVA
->   /CRITERIA=PIN(.05) POUT(.10)
->   /NOORIGIN
->   /DEPENDENT POSIT
->   /METHOD=ENTER GTGB f109  .
```

Output-Datei

```
          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *
Listwise Deletion of Missing Data
Equation Number 1    Dependent Variable..   POSIT
Block Number  1.  Method:  Enter      F109     GTGB

Variable(s) Entered on Step Number
   1..    GTGB
   2..    F109      GESCHLECHT
Multiple R           ,59781
R Square             ,35737
Adjusted R Square    ,35646
Standard Error     10,30024

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           2         82895,60722      41447,80361
Residual          1405        149063,43752        106,09497

F =     390,66699       Signif F =  ,0000


------------------ Variables in the Equation ------------------

Variable             B        SE B       Beta         T  Sig T

F109          -6,051479     ,558267   -,235428   -10,840  ,0000
GTGB            ,648716     ,023798    ,592054    27,260  ,0000
(Constant)    31,545361    1,234339               25,556  ,0000

End Block Number   1   All requested variables entered.
```

Der Output kann nun, analog zur Beispielrechnung aus der bivariaten Regression interpretiert werden. R-Square kann im Falle der multiplen Regression wieder als Anteil der erklärten Varianz in der abhängigen Variablen interpretiert werden. Allerdings wird diese Varianz nun von den beiden unabhängigen Variablen gemeinsam erklärt (vgl. Skript S.44f). Mit R^2=0,35 für R-Square ist der

Erklärungsbeitrag des Gesamtmodells moderat, d.h. es bleibt hier noch Raum für weitere erklärende Faktoren.

$$SAQ(\text{Re}\,sidual) = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = 149063{,}43$$

$$SAQ(\text{Re}\,gression) = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = 82895{,}60$$

$$R - Square = \frac{erklärte\ Variation}{Gesamt\ \text{var}\,iation} = \frac{82895{,}60}{82895{,}60 + 149063{,}43} = 0{,}35$$

Damit erklären die beiden unabhängigen Variablen Geschlecht und materielle Gütergebundenheit zusammen 35% der Varianz der abhängigen Variable Positionalität.

Der standardisierte partielle Regressionskoeffizient fällt für die Variable GTGB mit einem Wert von 0,592 etwas höher aus als zuvor in der bivariaten Analyse. Der Erklärungsbeitrag der Variable Geschlecht (f109) (*Beta* = -0,23) fällt moderat aus. Der Einfluß der u.V. GTGB ist mehr als doppelt so groß wie der Einfluß durch das Geschlecht (vgl. die Beta-Koeffizienten, nicht die unstandardisierten Regressionskoeffizienten). Alle Beziehungen sind hochsignifikant, d.h. es kann davon ausgegangen werden, daß der in der Haupthypothese vermutete Zusammenhang eine Entsprechung in der Grundgesamtheit findet (vgl. genauer Kap. 2.2).

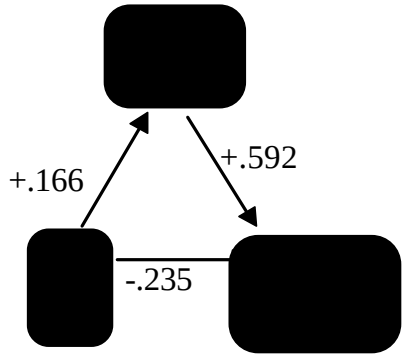
Zur Interpretation der Regressionskoeffizienten für die Dummy-Variable Geschlecht:
Inhaltlich besagt der Regressionskoeffizient von B= -6.05, daß Männer positionaler sind als Frauen. Frauen haben demnach auf der Positionalitätsskala im Durchschnitt einen um 6,05 Einheiten geringeren Wert, sind also weniger positional (wobei der Einfluß der unterschiedlichen Gütergebundenheit schon berücksichtigt wurde).

Interpretation des Kausalmodells:
Den Regressionsrechungen lagen folgende korrelative Zusammenhänge zugrunde:

	f109	**GTGB**	**POSIT**
f109	1,0000	,1660	-,1338
GTGB	,1660	1,0000	,5510
POSIT	-,1338	,5510	1,0000

Anm.: Korrekterweise müßte für die Korrelationen $r_{f109\text{-}POSIT}$ und $r_{f109\text{-}GTGB}$ der punktbiseriale Korrelationskeoffizent berechnet werden (Zusammenhangsmaß für die Beziehung zwischen einer nominalskalierten und einer intervallskalierten Variablen)

Insgesamt ergibt sich aufgrund der Regressionsanalysen folgendes Kausalmodell



Die Ausgangshypothese lautete: ‘je höher die materielle Gütergebundenheit, desto stärker ist die Positionlität ausgeprägt’. Als bivariater Zusammenhang wird die Hypothese mit r = .551 bestätigt

(entspricht auch dem β-Koeffizienten und damit dem Kausalkoeffizienten der bivariaten Regression). Durch Hinzunahme der Drittvariablen ‘Geschlecht’ als antezedierende Variable haben wir in der multivariaten Regressionsanaylse einen höheren kausalen Zusammenhang von GTGB und POSIT mit p=.592 festgestellt. Offensichtlich handelt es sich bei der vorliegenden Konstellation der Variablen um eine Verstärkung ($p_{YX} > r_{YX;}$ vgl. Skript S. 49). Die Variable Geschlecht hat einen gegengerichteten scheinkorrelativen Effekt auf die empirische Beziehung zwischen den Variablen GTGB und POSIT und vermindert die Kausalbeziehung von p=.592 auf eine korrelative Beziehung in Höhe von r=.551.

$$r_{YX} = p_{YX} + p_{ZX}\, p_{ZY}$$
$$= +.551 + (+.166) * (-.235) = .592$$

2.2 Voraussetzungen und Signifikanz-Tests für die Regressionsanalyse

2.2.1 Voraussetzungen

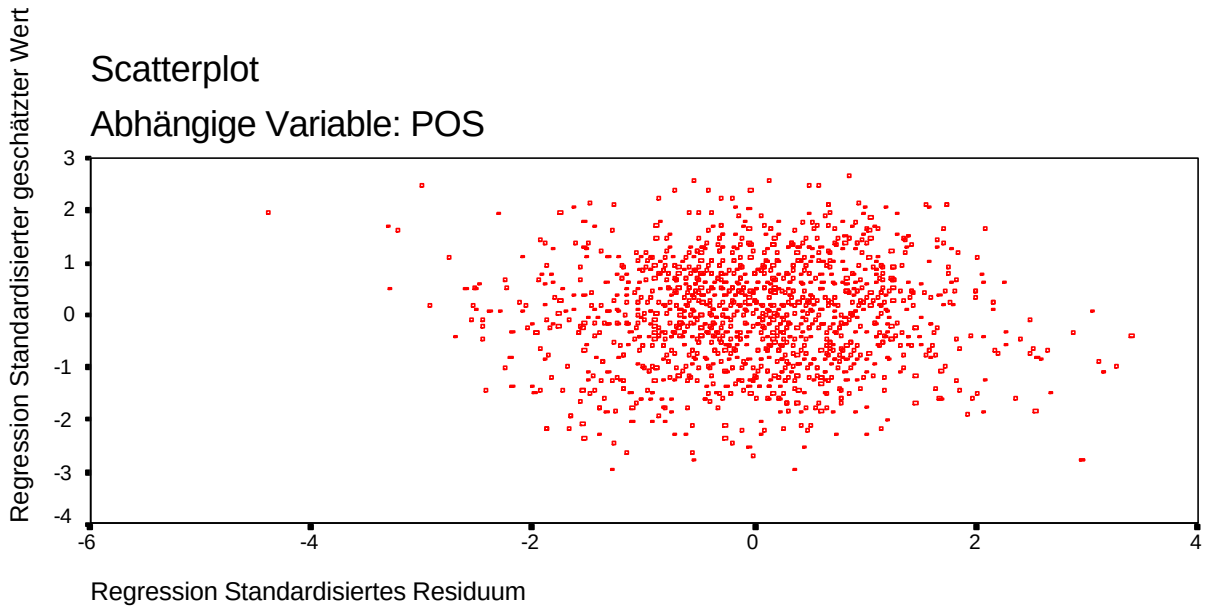
Die aufgrund der Stichprobe berechneten Regressionskoeffizienten (a, b) sind Schätzungen der wahren Koeffizienten, der entsprechenden Regressionsgleichung für die Grundgesamtheit. Es kommt nun darauf an, daß das Verfahren, mit dem sie geschätzt wurden, erwartungstreue, konsistente und effiziente Schätzungen liefert (vgl. Urban, S. 99-116). Für die OLS-Schätzung läßt sich dies nachweisen, wenn folgende Voraussetzungen erfüllt sind (vgl. Kähler, S. 370-379):

1. Für jede Wertekombination der unabhängigen Variablen ist das ermittelte Residuum ($Y_i - \hat{Y}_i$) eine Realisation einer normalverteilten Zufallsvariablen.

2. Sämtliche dieser Zufallsvariablen sind paarweise voneinander statistisch unabhängig (unkorreliert) und ihre Verteilungen haben alle den Mittelwert 0 und die gleiche Varianz (Homoskedastizität).

2.2.1.1 Prüfung der Linearitätsannahme

Anhand des Streudiagramms zwischen den standardisierten Vorhersagewerten ZPRED ($\hat{Y}_i$ [standardisiert]) und den standardisierten Residuuen ZRESID ($Y_i - \hat{Y}_i$ [standardisiert]) läßt sich die Linearität der Beziehung überprüfen. Es wird angefordert im Dialogfeld *Grafiken*, in dem man angibt, welche Variablen dargestellt werden sollen.
Für ZPRED wurde die Y-Achse und für ZRESID die X-Aches ausgewählt.



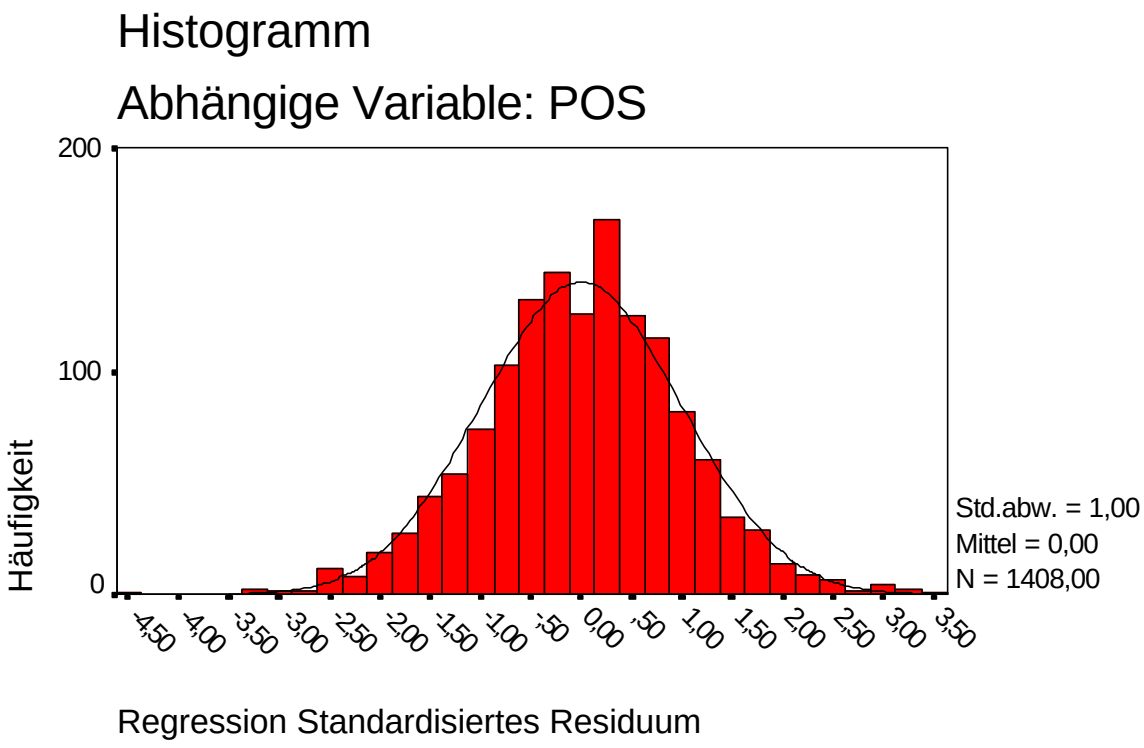
Es sollte möglichst ein horizontales Punkteband resultieren, das aus Punkten besteht, die zufällig um die Waagerechte (durch 0) verteilt sind. Oft treten auch ellipsenförmige oder kreisförmige Punktwolken auf. Sofern kein systematischer Kurvenverlauf vorliegt, liegt eine lineare Beziehung vor. Da in der dargestellten Punktwolke kein systematischer Zusammenhang erkennbar ist, gehen wir von einer linearen Beziehung zwischen den unabhängigen Variablen und der abhängigen Variablen bei der von uns durchgeführten Regression aus.

2.2.1.2 Prüfung der Honoskedastizität

Legt man durch die Punktwolke eine waagerechte Linie (Null-Linie), so ist erkennbar, daß die Breite der Punktwolke entlang dieser Linie weder wächst noch abnimmt. Zwar sind die Punkte im Zentrum dichter als in den Randbereichen verteilt, jedoch bleibt die Streuung der Punkte um die gedachte Null-Linie ungefähr gleich. Damit wird die Annahme gleicher Varianzen der Residuen (Homoskedastizität) bestätitgt.

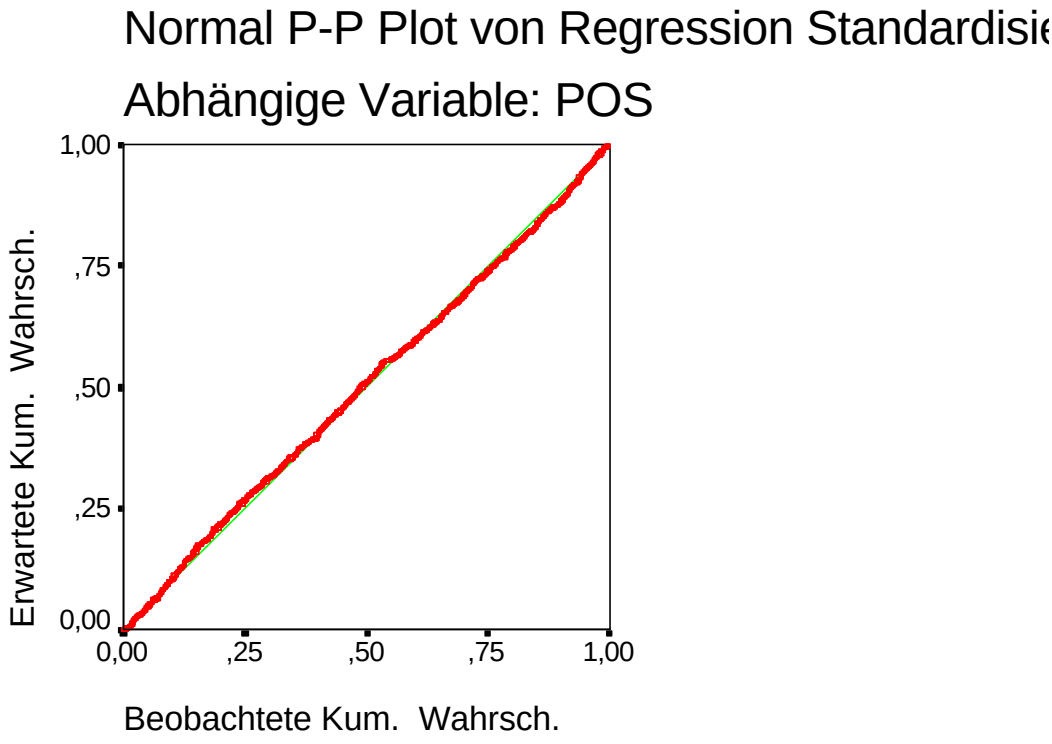
2.2.1.3 Prüfung der Residuen auf Normalverteilung

Eine Form der Überprüfung ist die graphische Veranschaulichung. Im Dialogfeld LINEARE REGRESSION kann unter GRAFIKEN... im Feld „Darstellung der standardisierten Residuen“ ein „Histogramm“ und ein „Vergleich mit der Normalverteilung“ angefordert werden. In unserem Beispiel ergibt sich folgendes Histogramm.



Für diese Grafik wurden die Residuen standardisiert (ZRESID), um sie mit der Standardnormalverteilung vergleichen zu können. Die Säulen des Histogramms entsprechen den empirischen Häufigkeiten der standardisierten Residuen; die glockenförmige Kurve gibt die entsprechende Normalverteilung wieder.
Es zeigt sich eine Verteilung der Residuen, die kaum von der Normalverteilung abweicht.
Dies zeigt auch der folgende Plot der kummulierten Wahrscheinlichkeiten.

Der Normalverteilungsplott stellt die kummulierte Häufikeitsverteilung der standadisierten Residuen (Punkte) sowie die kummulierte Normalverteilung (Gerade) dar. Bei Normalverteilung der Residuen müssen diese auf der Geraden liegen.
Auch hier erkennt man die nur geringen Abweichungen der Residuen von der Normalverteiling.



2.2.2 Signifikanztests für die multivariate Regressionsanalyse (siehe Kap. 2.1)

Signifikanztest des Regressionskoeffizienten (b):

Mit Hilfe des Standardfehlers (SE B) läßt sich ein Konfidenzintervall bestimmen, in dem der Koeffizient B mit einer gewissen Wahrscheinlichkeit zu finden ist. Geprüft wird, ob der geschätzte Regressionskoeffizient in der Grundgesamtheit einen von Null verschiedenen Wert hat. Die hierfür verwendete Verteilung ist die t-Verteilung.
Der t-Wert (T) des Regressionskoeffizienten für GTGB beträgt 27,200. Die Auftretenswahrscheinlichkeit für einen so hohen Wert unter der Bedingung, daß der Wert des Regressionskoeffizienten in der Grundgesamtheit gleich Null ist, ist äußerst gering (SIG T = 0.0000). Damit gilt, daß der standardisierte Regressionskoeffizient für GTGB (Beta = 0,5920) signifikant von Null verschieden ist und zwar sowohl bei einem Sicherheitsniveau von 95% wie auch von 99% (sogar von 99,9%). Ebenso ist der standardisierte Regressionskoeffizient für die Variable Geschlecht (f109) mit einem T-Wert von T=-10,840 hochsignifikant von 0 verschieden.

Signifikanztest des Gesamtmodells (F-Test):
Getestet wird die Signifikanz des erklärten Varianzanteils - relativ zum unerklärten Varianzanteil. Dabei wird die Gesamtschätzung des Modells zugrunde gelegt (vgl. die varianzanalytische Interpretation in Kap. 1.1.2.). Die Prüfgröße bildet der F-Wert. F-Werte folgen einer bestimmten Verteilung, die von der Stichprobengröße und der Anzahl der geschätzten Parameter (Anzahl der geschätzten Koeffizienten) abhängig ist. Der in unserem Beispiel ermittelte F-Wert beträgt $F_{(2,1405)}$=390,666. Er ist hochsignifikant (Signif F = 0,0000). Damit ist die Erklärungsleistung des Regressionsmodells kein Zufallsergebnis, d.h. wir können die Nullhypothese (die erklärte Varianz in der Grundgesamtheit ist gleich Null) für alle üblichen Signifikanzniveaus zurückweisen.