Übungsblatt zum Referat

Multiple Regression in der Anwendung mit SPSS

In unserem **fiktiven** Beispiel soll ermittelt werden, wie sich die Häufigkeiten von Suizidversuchen aus verschiedenen Variablen vorhersagen lassen. Unsere unabhängigen Variablen sind Alter, Alkoholkonsum, Intelligenz, Einwohnerdichte am Wohnort und die durchschnittliche Sonneneinstrahlung.

Noch ein Hinweis: Wenn im Text von Subdialogen die Rede ist, dann bezieht sich das immer auf den Dialog der multiplen Regression (**Statistik→Regression→Linear** (**Statistics→Regression→linear**)).

AUFGABE 1: AGGREGIEREN

Die Intelligenz wurde in einem Test mit 5 Subskalen erfasst. Aus praktischen Gründen empfiehlt es sich, nicht alle 5 Skalenwerte einzeln in die Regression aufzunehmen. Darum sollen die Werte für IQ1 bis IQ5 aggregiert werden.

Dabei soll gewährleistet sein, daß mindestens 3 der Variablen in die neue Variable eingehen, daß also bei mehr als zwei fehlenden Werten auch in der neuen Variable ein fehlender Wert eingetragen wird.

Die aggregierte Variable soll IQ heißen.

HINWEIS:

Verwenden Sie entweder den Compute-Befehl (wenn Sie ihn von Hand eingeben wollen) oder nutzen Sie den Dialog, der über **Transformieren→Berechnen** (**Transform→Compute**) erreichbar ist.

Der erforderliche Befehl lautet **mean.*k*(*v1, v2, ..., vn*)**

AUFGABE 2: AUSREIßER

Die Methode der kleinsten Quadrate ist sehr anfällig gegenüber Ausreißern. Testen Sie, ob es Extremwerte gibt, die das Ergebnis der Regression verfälschen könnten.

HINWEIS:

Prinzipiell gibt es zwei sinnvolle Möglichkeiten, auf Ausreißer zu testen, eine grafische und eine numerische.

AUFGABE 3: NORMALVERTEILUNGSANNAHME

Eine mathematische Voraussetzung der multiplen Regression ist die Annahme, daß die Residuen normalverteilt sind. Prüfen Sie also, ob diese Voraussetzung erfüllt ist.

HINWEIS:

Verwenden Sie die Schaubilder **Histogramm** (**Histogramm**) und/oder **Normalverteilungs-Diagramm** (**Normal probability plot**) im Subdialog **Diagramme** (**Plots→Standardized residual plots**).

AUFGABE 4A: MODELLSELEKTION FORWARD

Die Verwendung der schrittweisen Verfahren ist bei der multiplen Regression mit einigen Unklarheiten verbunden. Besonders wenn zwischen den Prädiktoren hohe Korrelationen vorliegen, ergeben die verschiedenen Verfahren unterschiedliche Ergebnisse.

Rechnen Sie in einem ersten Schritt eine multiple Regression mit der Methode **Forward**.

HINWEIS:

In dem Feld **Methode** (**Method**) geben Sie **Forward** an.

AUFGABE 4B: MODELLSELEKTION BACKWARD

Wiederholen Sie die Regression aus Aufgabe 4a mit nur einer kleinen Änderung. In dem Feld **Methode** (**Method**) geben Sie statt der **Forward** die **Backward**-Methode an. Anschließend vergleichen Sie die Regressionsmodelle, die SPSS aufgestellt hat.

LÖSUNG ZU 1:

Öffnen Sie den Dialog **Transformieren→Berechnen** (**Transform→Compute**). Unter **Zielvariable** (**Target Variable**) wird der Name der neuen Variable eingegeben. Unter **Numerischer Ausdruck** (**Numeric Expression**) wird der Befehl wie folgt angegeben: **mean.3(IQ1,IQ2,IQ3,IQ4,IQ5)**

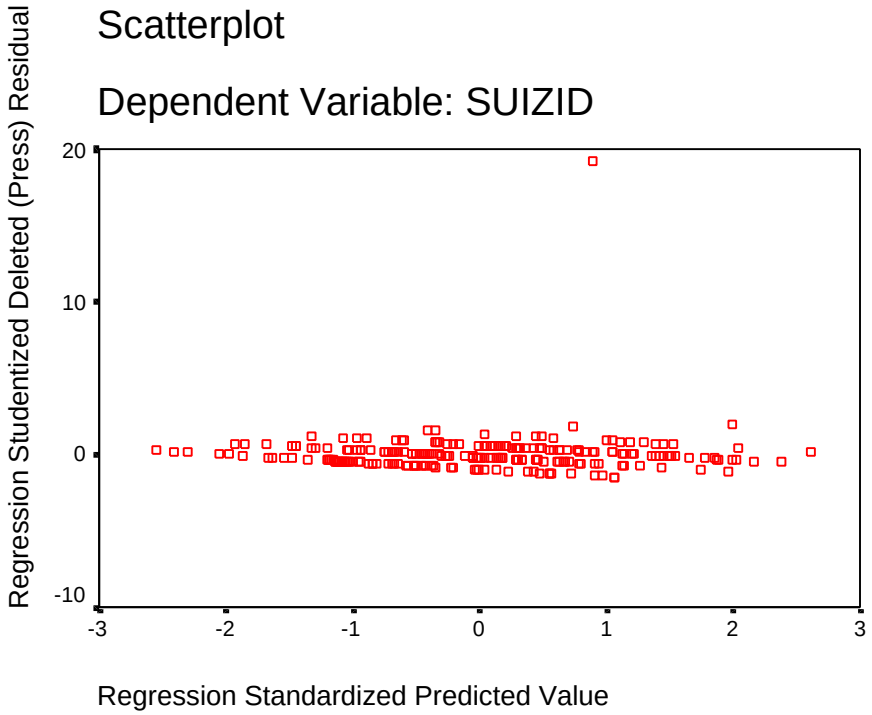
Die Zahl hinter dem Punkt gibt an, wie viele Werte zur Berechnung des aggregierten Wertes mindestens benutzt werden sollen.

LÖSUNG ZU 2:

Grafisch:

Man öffnet den Sub-Dialog **Diagramme** (**Plots)**. Auf der X-Achse wird der standardisierte vorhergesagte Wert benutzt (SPSS-Variable ***ZPRED**). Auf der Y-Achse das studentisierte (=standardisierte) ausgeschlossene Residuum (SPSS-Variable ***SDRESID**). Nicht vergessen, die abhängige und die unabhängigen Variablen zu definieren.

SPSS gibt dann mit der Regression ein **Streudiagramm von *sdresid über *zpred** (***sdresid by *zpred Scatterplot**) aus. Dort sucht man nach Werten ungefähr größer 3,5. Diese geben einen Hinweis auf Ausreißer.



Hier finden wir nun einen Ausreißer, bei dem es nun zu überlegen gilt, ob eine Streichung inhaltlich vertretbar ist. Dazu ist es hilfreich, die Nummer des Probanden zu kennen.

Dazu doppelklicken Sie auf das Schaubild, und es öffnet sich der **SPSS-Diagramm-Editor** (**SPSS Chart Editor**). Dort stellt man unter **Diagramme→ Optionen** (**Chart→ Options**) das Feld **Fallbeschriftungen** (**Case Labels**) auf **Ein** (**on**). Nun kann man im Schaubild die Nummer der Versuchsperson ablesen und sie sich im Datenblatt genauer ansehen. Die Versuchsperson 129 ist zu eliminieren.

Numerisch:

Eine zweite Möglichkeit bietet die Inspektion von Extremwerten mittels Cook-Distanzen. Im Sub-Dialog **Speichern (Save)** die Option **nach Cook** (**Cook's**) berechnet eine neue Spalte mit Werten im Datenblatt. In dieser neuen Variable müssen nur noch Werte größer *Eins* gesucht werden. Keine Cook-Distanz ist größer als Eins, auch nicht die der Versuchsperson 129.

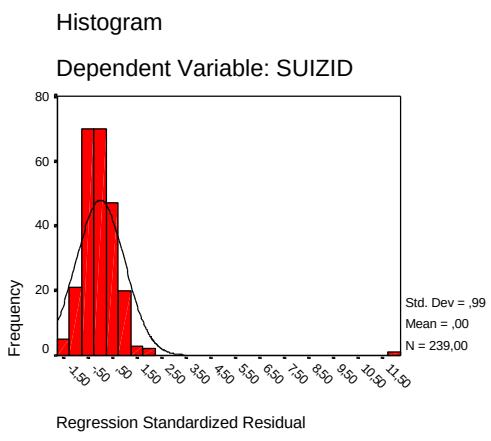
Man sieht, die Ergebnisse dieser beiden Methoden sind nicht immer identisch!

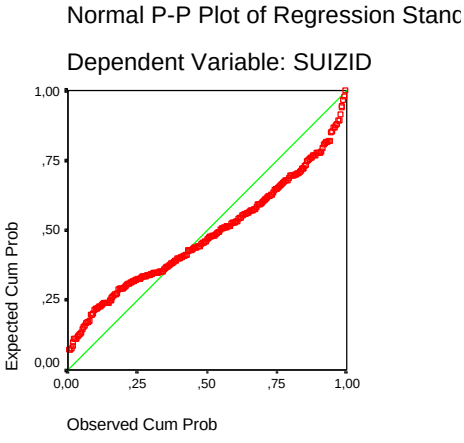
LÖSUNG ZU 3:

Ob die Residuen auch normalverteilt sind, läßt sich ebenfalls anhand eines Schaubildes überprüfen. Dieses läßt sich parallel zur Regression erstellen über den Sub-Dialog **Diagramme→ Diagramme der standardisierten Residuen** (**Plots→Standardized residual plots**). Die beiden Optionen **Histogramm** (**Histogramm**) und **Normalverteilungsdiagramm** (**Normal probability plot**) resultieren in jeweils einem Schaubild.

Im **P-P-Diagramm** (**P-P-Plot**) müssen die Punkte möglichst nahe an der 1. Winkelhalbierenden liegen, somit kann man die Normalverteilungsannahme bezüglich der Residuen als gegeben betrachten.

In diesem unserem Fall können wir die Normalverteilung der Residuen annehmen.





LÖSUNG ZU 4 INSGESAMT:

Vor dem Berechnen der Regressionen, ist zu beachten, daß Proband Nummer 129 eliminiert wurde (s. Aufgabe 2).

Wie man sieht, sind die Ergebnisse der beiden Selektionsverfahren in diesem Fall tatsächlich verschieden. Wenn Sie die Signifikanzberechnungen der β-Koeffizienten betrachten (Tabelle **Koeffizienten** (**Coefficients**)), dann werden Sie feststellen, daß bei der Backward-Methode eine Variable in dem Modell verbleibt, obwohl deren Siginifikanzniveau über 5% liegt. Das liegt an den standardmäßigen Einstellungen von SPSS (Subdialog **Optionen** (**Options**)). Für die Aufnahme einer Variablen in das Modell ist hier ein kritisches Niveau von 5% angegeben, zum Ausschluß einer Variablen dagegen eines von 10%.

In unserem Fall sind diese verschiedenen Ergebnisse also kein Resultat allzu hoher Korrelationen zwischen den Prädiktoren.

LÖSUNG ZU 4A:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1,470	,111		13,230	,000		
	Bevölkerungsdichte am Wohnort	,567	,079	,426	7,214	,000	1,000	1,000
2	(Constant)	-5,76E-03	,230		-,025	,980		
	Bevölkerungsdichte am Wohnort	,514	,072	,386	7,159	,000	,989	1,011
	Sonneneinstrahlung	1,713E-02	,002	,385	7,145	,000	,989	1,011

a. Dependent Variable: Anzahl der Selbstmordversuche

LÖSUNG ZU 4B:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-,666	,557		-1,196	,233		
	ALTER	-4,74E-03	,003	-,073	-1,359	,175	,987	1,013
	Bevölkerungsdichte am Wohnort	,532	,073	,399	7,249	,000	,937	1,067
	Sonneneinstrahlung	1,658E-02	,002	,373	6,859	,000	,963	1,038
	Durchschnittlicher Alkoholkonsum	4,854E-02	,066	,040	,736	,462	,985	1,015
	IQ_G	8,162E-03	,005	,095	1,729	,085	,943	1,060
2	(Constant)	-,582	,544		-1,069	,286		
	ALTER	-4,65E-03	,003	-,072	-1,336	,183	,988	1,012
	Bevölkerungsdichte am Wohnort	,529	,073	,397	7,224	,000	,941	1,063
	Sonneneinstrahlung	1,645E-02	,002	,370	6,831	,000	,968	1,033
	IQ_G	7,962E-03	,005	,093	1,691	,092	,946	1,057
3	(Constant)	-,796	,521		-1,527	,128		
	Bevölkerungsdichte am Wohnort	,539	,073	,404	7,379	,000	,950	1,053
	Sonneneinstrahlung	1,655E-02	,002	,372	6,860	,000	,969	1,032
	IQ_G	7,960E-03	,005	,093	1,688	,093	,946	1,057

a. Dependent Variable: Anzahl der Selbstmordversuche