

Lineare und logistische Regression mit SPSS

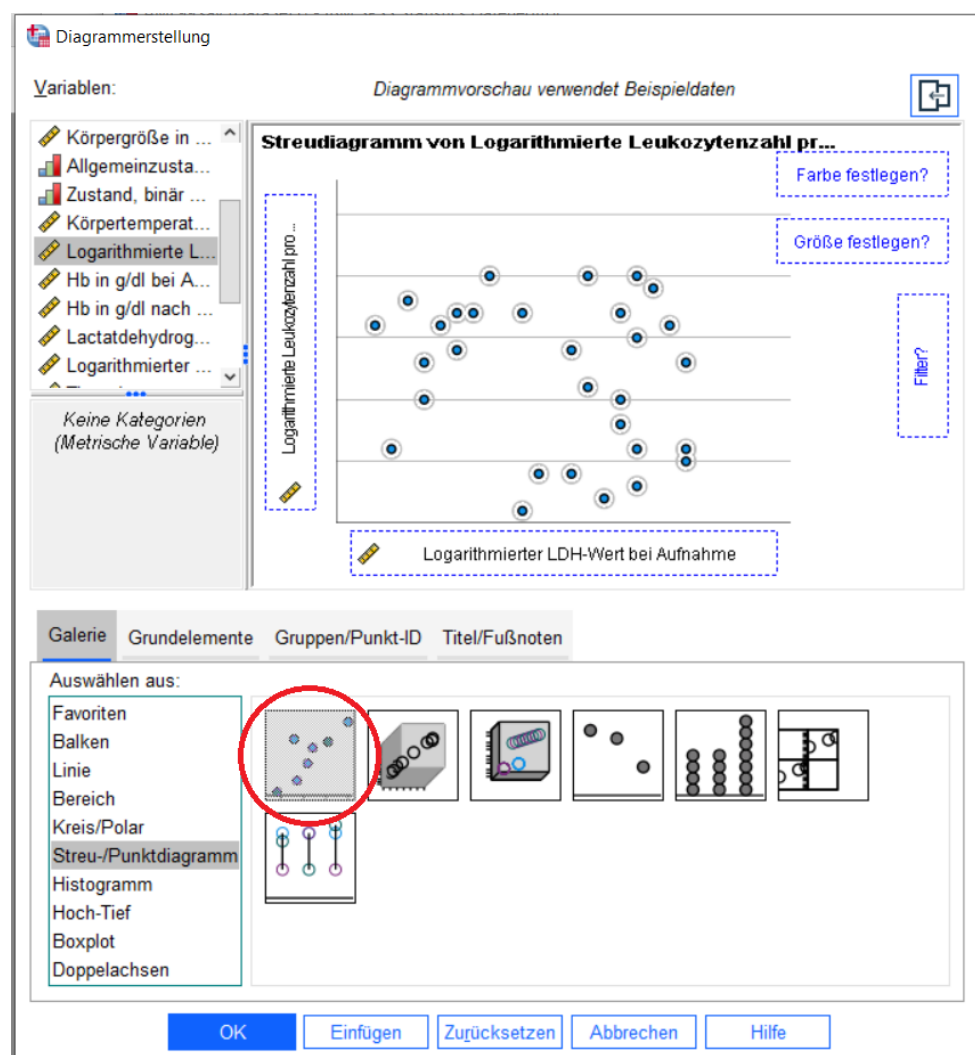
Alle Beispiele stammen aus dem Datensatz AML99.sav, der in der zweiten und dritten Praktikumseinheit verwendet wird.

1. Streudiagramm

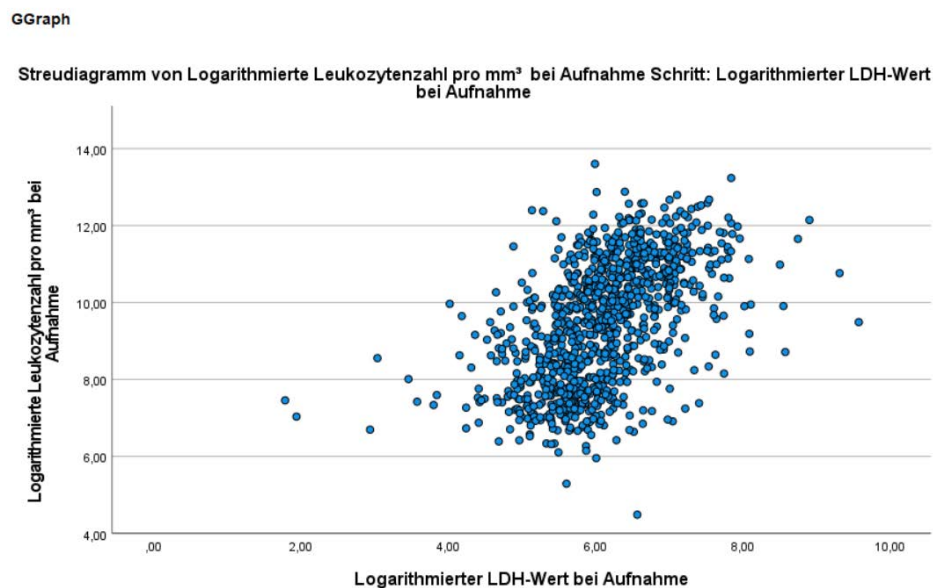
In einem Streudiagramm werden die Ausprägungen zweier quantitativer Merkmale graphisch gegeneinander abgetragen. Im folgenden Beispiel sind dies der logarithmierte LDH-Wert und der logarithmierte Leukozytenwert.

Das Streudiagramm wird über **Grafik > Diagrammerstellung** angefordert. Man wählt die Kategorie **Streu-/Punktdiagramm** aus. Nun kann per Drag & Drop in das Feld *Diagrammvorschau* oder per Doppelklick auf das Symbol für ein *einfaches Streudiagramm* die gewünschte Art der Darstellung ausgewählt werden. Die Auswahl erscheint im Feld *Diagrammvorschau*.

Die *X- und Y-Achse* werden definiert, indem die entsprechenden Variablen aus der Variablenliste auf das gestrichelt umrahmte Feld für die X- und Y-Achse gezogen werden.



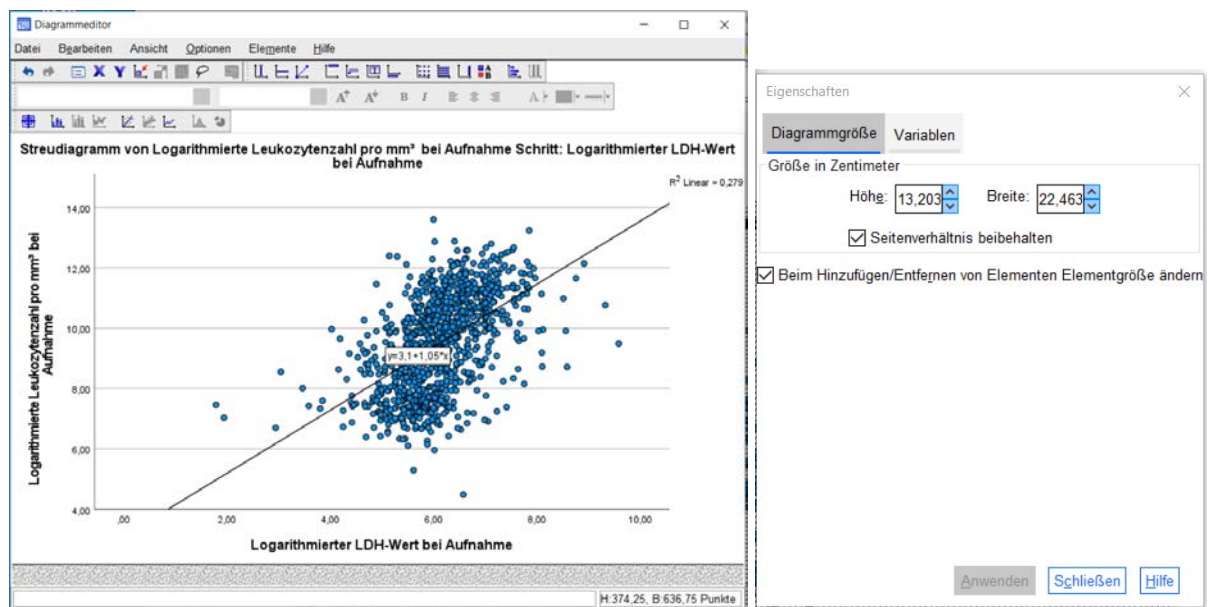
Nach Bestätigen mit **OK** wird folgende Abbildung erzeugt:



Die Abbildung gibt Hinweise auf den funktionalen Zusammenhang der Variablen. Der logarithmierte Leukozytenwert ist tendenziell hoch, wenn auch der logarithmierte LDH-Wert hoch ist, und umgekehrt. Es liegt ein positiver linearer Zusammenhang vor. Damit ist die lineare Regression geeignet, um den Zusammenhang der beiden Merkmale zu quantifizieren.

Hinweis:

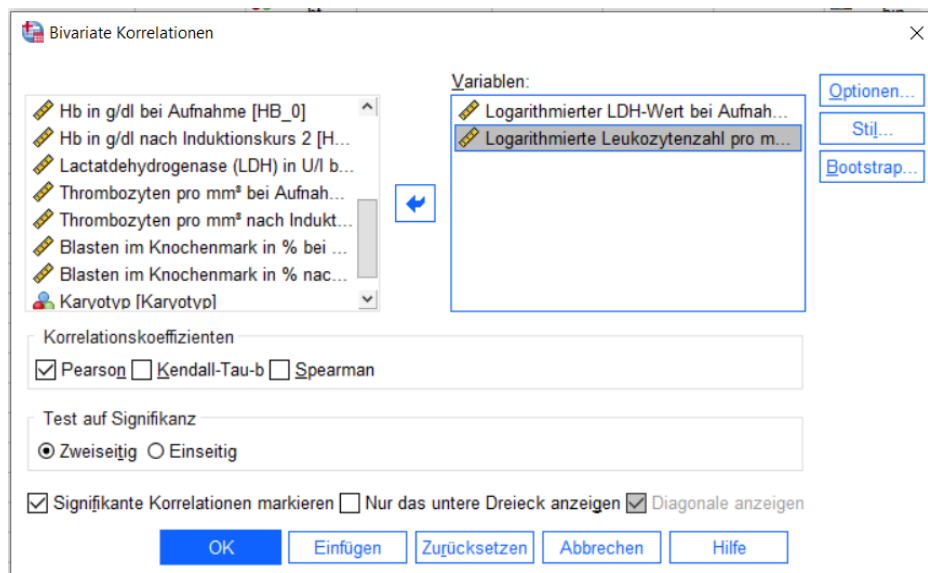
Durch **Doppelklick auf die Abbildung** im Ausgabefenster öffnet sich die Abbildung für eine mögliche Bearbeitung. So lassen sich beispielsweise die Größe der Abbildung oder die verwendeten Farben verändern oder eine Regressionsgerade einzeichnen.



2. Korrelationskoeffizient

Für die Berechnung des Korrelationskoeffizienten wählt man **Analysieren > Korrelation > Bivariat**. Die Variablen, für welche die Korrelation bestimmt werden soll (hier der logarithmierte LDH-Wert und der logarithmierte Leukozytenwert), werden aus der Variablenliste ausgewählt und in das rechte Fenster mit der Überschrift *Variablen* eingefügt.

Durch **Setzen eines Häkchens** können verschiedene Korrelationskoeffizienten angefordert werden. Für normalverteilte Daten eignet sich der Korrelationskoeffizient nach **Pearson**. Sind die Daten nicht normalverteilt, sollte der Korrelationskoeffizient nach **Spearman** verwendet werden. Um den geeigneten Korrelationskoeffizienten für die vorliegenden Variablen auswählen zu können, muss zuvor deren Verteilung mit geeigneten Mitteln (z.B. Histogramme, siehe Übungsstunde 1) untersucht werden. Dabei zeigt sich, dass beide Variablen als normalverteilt angesehen werden können.



Nach Bestätigen mit **OK** wird folgender Output erzeugt. Der Korrelationskoeffizient ist rot hervorgehoben.

Korrelationen

Korrelationen			
		Logarithmierter LDH-Wert bei Aufnahme	Logarithmierte Leukozytenzahl pro mm³ bei Aufnahme
Logarithmierter LDH-Wert bei Aufnahme	Pearson-Korrelation	1	,528**
	Sig. (2-seitig)		,000
	N	985	983
Logarithmierte Leukozytenzahl pro mm³ bei Aufnahme	Pearson-Korrelation	,528**	1
	Sig. (2-seitig)	,000	
	N	983	1001

** . Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

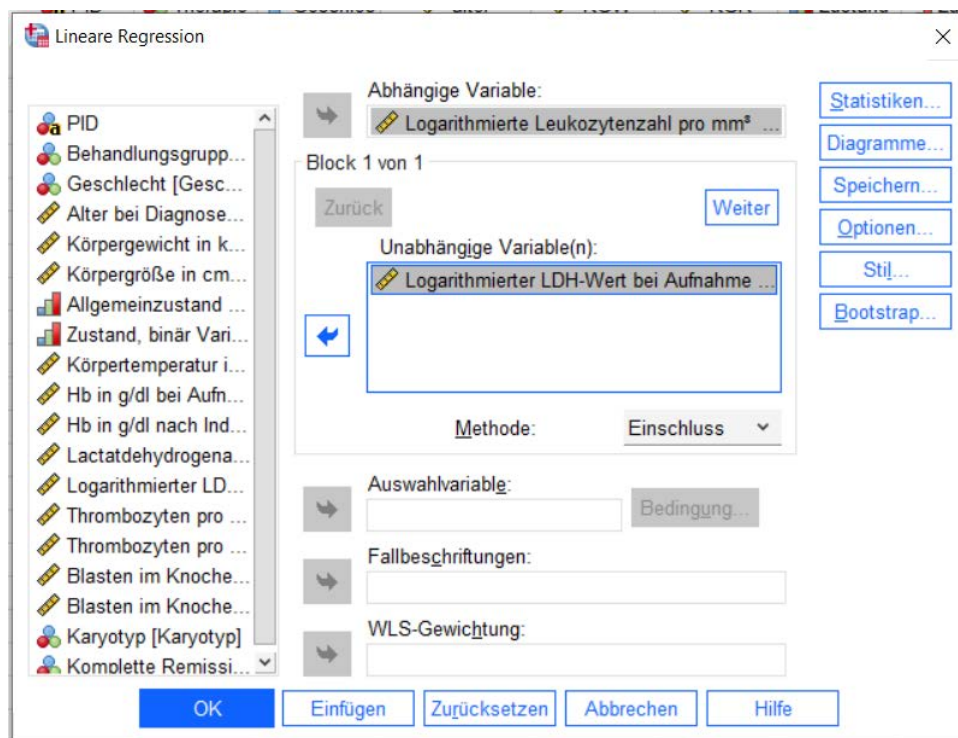
In der Tabelle kann der Korrelationskoeffizient nach Pearson (bzw. Spearman) abgelesen werden. Der Korrelationskoeffizient kann Werte zwischen -1 und 1 annehmen. Je größer der Absolutbetrag des Korrelationkoeffizienten, desto stärker ist der zugrunde liegende lineare (Spearman: monotone) Zusammenhang. Dabei zeigt ein positiver Korrelationskoeffizient einen positiven linearen Zusammenhang und ein negativer Korrelationskoeffizient einen negativen linearen Zusammenhang an.

Im vorliegenden Beispiel bestätigt der Korrelationskoeffizient den bereits im Streudiagramm erkennbaren positiven linearen Zusammenhang.

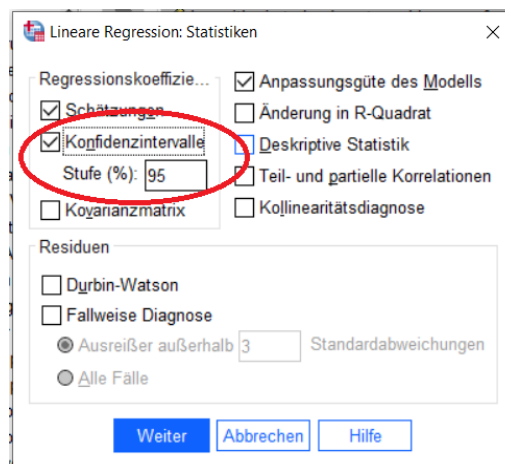
3. Linear Regression

Die lineare Regression eignet sich zur Bestimmung des Einflusses einer oder mehrerer Einflussgrößen auf eine stetige Zielgröße. Im folgenden Beispiel wird der Einfluss des logarithmierten LDH-Wertes auf den logarithmierten Leukozytenwert untersucht.

Um eine lineare Regressionsanalyse durchzuführen, wählt man **Analysieren > Regression > Linear**. Aus der Variablenliste werden die *abhängige Variable* und die *unabhängigen Variable(n)* ausgewählt und in die entsprechenden Felder eingetragen. Die abhängige Variable ist die Zielgröße (logarithmierter Leukozytenwert) und die unabhängige Variable die Einflussgröße (logarithmierter LDH-Wert).



Um zusätzlich zu den Regressionskoeffizienten des Regressionsmodells die zugehörigen Konfidenzintervalle ausgeben zu lassen, klickt man auf den Button **Statistiken** und setzt im folgenden Fenster ein Häkchen bei Konfidenzintervalle.



Nach Bestätigung mit **Weiter** und **OK** wird der folgende Output erzeugt, in dem die für diese Praktikumseinheit relevanten Kenngrößen rot hervorgehoben sind:

Regression

Aufgenommene/Entfernte Variablen^a

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	Logarithmierter LDH-Wert bei Aufnahme ^b		Einschluß

a. Abhängige Variable: Logarithmierte Leukozytenzahl pro mm³ bei Aufnahme

b. Alle gewünschten Variablen wurden eingegeben.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,528 ^a	,279	,278	1,36961

a. Einflußvariablen : (Konstante), Logarithmierter LDH-Wert bei Aufnahme

ANOVA^a

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	712,768	1	712,768	379,973	,000 ^b
	Nicht standardisierte Residuen	1840,200	981	1,876		
	Gesamt	2552,968	982			

a. Abhängige Variable: Logarithmierte Leukozytenzahl pro mm³ bei Aufnahme

b. Einflußvariablen : (Konstante), Logarithmierter LDH-Wert bei Aufnahme

Koeffizienten^a

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	95,0% Konfidenzintervalle für B	
		Regressionskoeffizient B	Std.-Fehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	3,096	,331		9,368	,000	2,448	3,745
	Logarithmierter LDH-Wert bei Aufnahme	1,046	,054	,528	19,493	,000	,940	1,151

a. Abhängige Variable: Logarithmierte Leukozytenzahl pro mm³ bei Aufnahme

In der Tabelle *Modellzusammenfassung* ist das Bestimmtheitsmaß R^2 angegeben, das beschreibt, welcher Anteil der Varianz der abhängigen Variable durch das Modell erklärt wird. Das Bestimmtheitsmaß entspricht dem quadrierten Korrelationskoeffizienten nach Pearson ($0,528^2 = 0,279$, siehe oben). Im Gegensatz zum Korrelationskoeffizienten gibt das Bestimmtheitsmaß nicht die Richtung des Zusammenhangs (d.h. positiv oder negativ) an.

In der Tabelle *Koeffizienten* werden die Regressionskoeffizienten angegeben, die zugehörigen p-Werte sowie die oberen und unteren Grenze der 95%-Konfidenzintervalle. Der geschätzte Regressionskoeffizient für den logarithmierten LDH-Wert beträgt 1,046, d.h. bei einer Erhöhung der logarithmierten LDH um eine Einheit steigt der logarithmierte Leukozytenwert durchschnittlich um 1,046 Einheiten. Dies entspricht der Steigung der Geraden, die in Abschnitt 1 in das Streudiagramm gezeichnet wurde. Der Regressionskoeffizient ist mit einem p-Wert von $p < 0,001$ deutlich unterschiedlich von 0 (95%-Konfidenzintervall: [0,940; 1,151]). Im Falle eines positiven linearen Zusammenhangs ist der Regressionskoeffizient positiv, bei einem negativen Zusammenhang ist er negativ. Anders als beim Bestimmtheitsmaß und dem Korrelationskoeffizienten ist der Wertebereich des Regressionskoeffizienten nicht auf ein bestimmtes Intervall eingeschränkt. Der Regressionskoeffizient kann theoretisch jede mögliche reelle Zahl annehmen.

Der Y-Achsenabschnitt beträgt 3,096, d.h. die Regressionsgerade schneidet die Y-Achse bei diesem Wert. Dies entspricht dem geschätzten Wert der abhängigen Variablen, wenn die unabhängige Variable den Wert 0 annimmt, d.h. $x = 0$. Inhaltlich ist der Y-Achsenabschnitt oft nicht interpretierbar.

Mit dem obigen Regressionskoeffizienten und Y-Achsenabschnitt lautet die geschätzte Modellgleichung des Regressionsmodells:

$$y = 3,096 + 1,046 \cdot x,$$

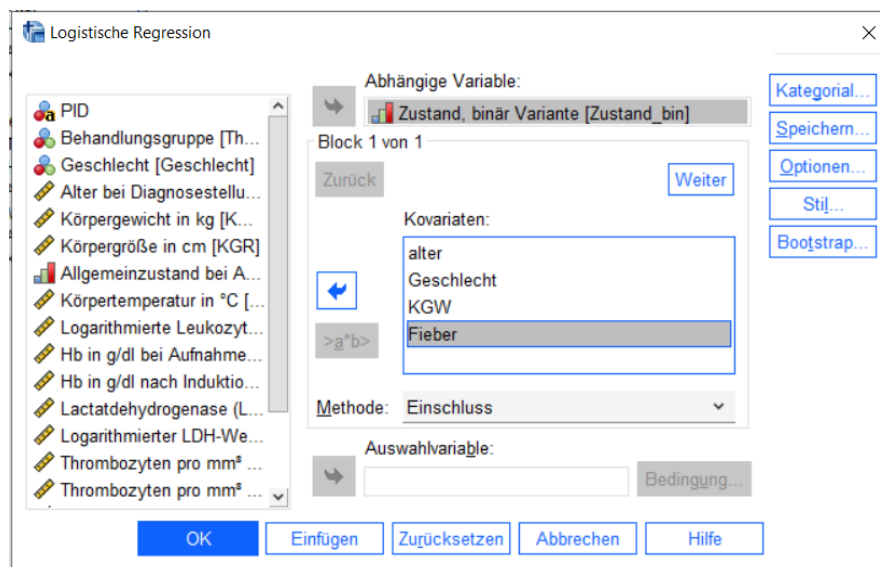
wobei die Variable y den logarithmierten Leukozytenwert angibt und die Variable x den logarithmierten LDH-Wert.

4. Logistische Regression

In einer Regressionsanalyse ergibt sich die zu verwendende Methode aus dem Messniveau der Zielgröße. Im Falle einer stetigen Zielgröße wurde im Abschnitt 3 eine lineare Regressionsanalyse durchgeführt. Ist die Zielgröße binär, so führt man eine logistische Regression durch. Im folgenden Beispiel soll der Einfluss des Alters, der Körpertemperatur, des Gewichts und des Geschlechts auf den Allgemeinzustand der Patienten bei Beginn der Studie untersucht werden. In der Variablen Zustand_bin ist angegeben, ob ein Patient nicht oder nur leicht beeinträchtigt (Zustand_bin = 0) oder stark beeinträchtigt (Zustand_bin = 1) ist. In einer logistischen Regressionsanalyse wird der Einfluss der genannten Variablen auf die Wahrscheinlichkeit (bzw. die Odds, d.h. die Chance) stark beeinträchtigt zu sein, modelliert.

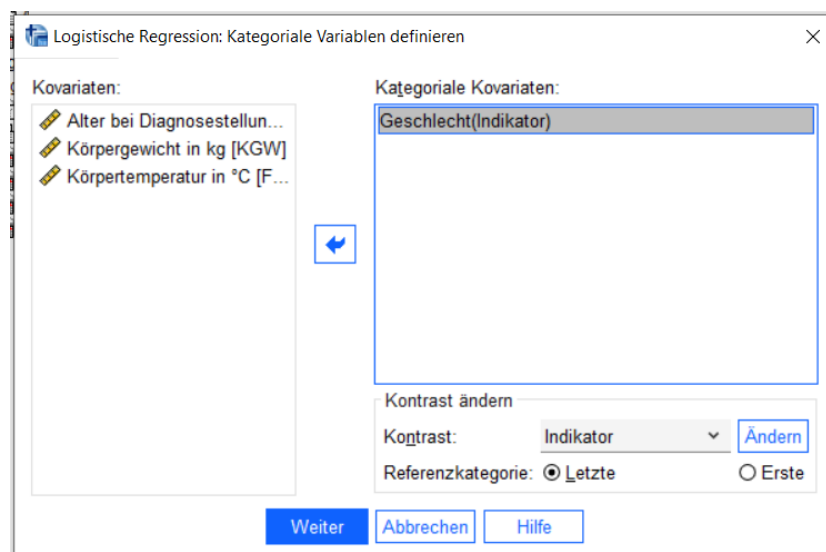
Um eine logistische Regressionsanalyse durchzuführen, wählt man **Analysieren > Regression > Binär logistisch**.

Als *abhängige Variable* wird die binäre Zielgröße aus der Variablenliste ausgewählt und eingetragen, die Einflussgrößen werden als *Kovariaten* eingetragen.



Kategoriale (d.h. nominale und ordinale) Merkmale müssen als solche gekennzeichnet werden. Dies geschieht durch Klicken auf den Button **Kategorial...**

Es öffnet sich ein neues Dialogfenster, in dem die kategorialen Variablen definiert werden können. Dazu werden ordinale und nominale Merkmale aus der angezeigten Liste der Kovariaten ausgewählt und in die Liste der *kategorialen Kovariaten* eingefügt. Für die einzelnen kategorialen Variablen kann anschließend die verwendete Codierung festgelegt werden, indem entweder die erste oder die letzte Kategorie der Variable als Referenz ausgewählt wird.

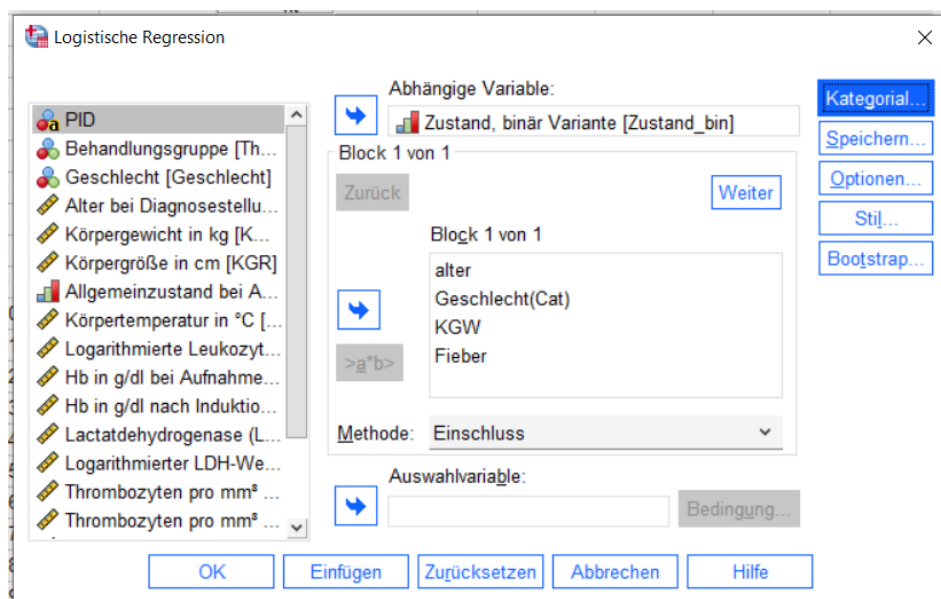


Die aktuelle Kodierung der Variable kann über einen Rechtsklick auf die Variable Geschlecht und **Variablenbeschreibung** angezeigt werden.



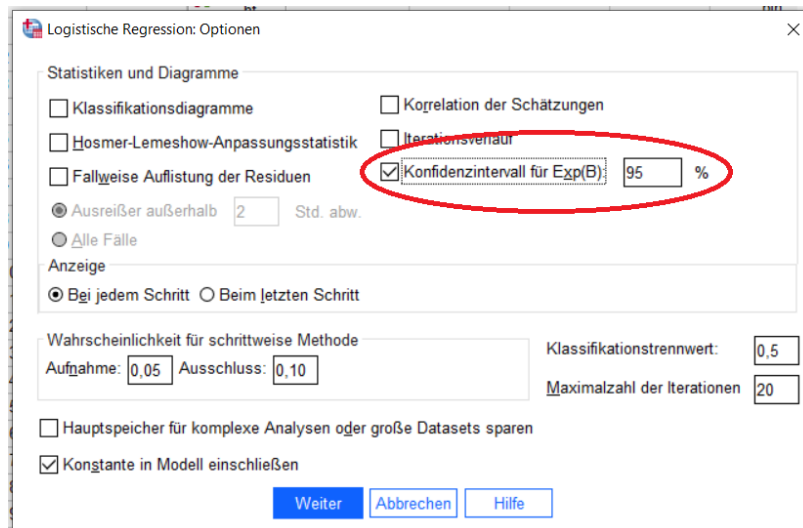
Per Voreinstellung wird die letzte Kategorie als Referenz gewählt. Im oben dargestellten Beispiel ist die Kategorie „Männlich“ mit 1 codiert und „Weiblich“ mit 2. Daher entspricht die Auswahl der letzten Kategorie des Merkmals Geschlecht, dass die Kategorie „Weiblich“ als Referenz verwendet wird. Soll die erste Kategorie als Referenz definiert werden, wählt man **Erste** bei *Referenzkategorie* und bestätigt dies über den Button **Ändern**.

Durch Bestätigen mit **OK** wird die Definition der kategorialen Variablen in das Modell übernommen. Dass die kategorialen Variablen erfolgreich definiert wurden, ist daran zu erkennen, dass sie in der Liste der Kovariaten anschließend durch die zusätzliche Angabe (*Cat*) gekennzeichnet sind.



Die interessierenden Kenngrößen in einer logistischen Regressionsanalyse sind die so genannten Odds Ratios. Sie geben den Faktor an, mit dem sich eine Einflussvariable auf die abhängige Variable auswirkt. Die zugehörigen Konfidenzintervalle zeigen die Präzision des entsprechenden Schätzwerts an und sollten in der Regel zusätzlich zum Odds Ratio angegeben werden. Enthält das Konfidenzintervall den Wert 1, so ist das zugehörige Odds Ratio nicht signifikant.

Um die Konfidenzintervalle der Odds Ratios ausgeben zu lassen, klickt man auf den Button **Optionen...** Anschließend setzt man ein **Häkchen** bei dem *Konfidenzintervall für Exp(B)* und bestätigt diese Auswahl mit **Weiter**.



Wenn alle notwendigen Einstellungen vorgenommen wurden, bestätigt man im Hauptdialogfenster mit **OK**, dass die logistische Regressionsanalyse durchgeführt werden soll.

Im Folgenden werden die wichtigsten Tabellen des SPSS-Outputs erläutert.

Logistische Regression

Zusammenfassung der Fallverarbeitung

Ungewichtete Fälle ^a		N	Prozent
Ausgewählte Fälle	Einbezogen in Analyse	847	84,1
	Fehlende Fälle	160	15,9
	Gesamt	1007	100,0
Nicht ausgewählte Fälle		0	,0
Gesamt		1007	100,0

a. Wenn die Gewichtung wirksam ist, finden Sie die Gesamtzahl der Fälle in der Klassifizierungstabelle.

Codierung abhängiger Variablen

Ursprünglicher Wert	Interner Wert
Normal oder leicht beeinträchtigt	0
Stark beeinträchtigt	1

Codierungen kategorialer Variablen

		Häufigkeit	Parameterkodierung (1)
Geschlecht	Männlich	465	1,000
	Weiblich	382	,000

Die erste Tabelle gibt an, wie viele Fälle in der Regressionsanalyse verwendet werden und wie viele auf Grund fehlender Werte in der abhängigen oder den unabhängigen Variablen nicht verwendet werden können. In der zweiten Tabelle wird die Codierung der abhängigen Variablen angegeben. Der Wert „0“ entspricht dem Referenzzustand und „1“ dem Alternativzustand.

Anschließend wird die Codierung der kategorialen Variablen angegeben. Im vorliegenden Beispiel wurde für das Merkmal „Geschlecht“ eine Codierung mit dem Wert „0“ für männlich und „1“ für weiblich gewählt. Die Angaben in der Tabelle „Codierungen kategorialer

Variablen“ sind wichtig für das Verständnis der zu den kategorialen Variablen gehörenden Odds Ratios (siehe unten).

Der als Omnibus-Test der Modellkoeffizienten bezeichnete Test ist ein Likelihood-Quotienten-Test, der die Gesamtanpassung des Modells überprüft. Ein p-Wert von $p \leq 0,05$ zeigt an, dass die gewählten Einflussvariablen bzw. das erstellte Modell grundsätzlich geeignet zur Beschreibung der abhängigen Variable ist.

Omnibus-Tests der Modellkoeffizienten

		Chi-Quadrat	df	Sig.
Schritt 1	Schritt	63,743	4	,000
	Block	63,743	4	,000
	Modell	63,743	4	,000

In der Tabelle *Variablen in der Gleichung* befinden sich die geschätzten Odds Ratios des angepassten Modells in der mit $Exp(B)$ überschriebenen Spalte. Die Bezeichnung beruht darauf, dass die Odds Ratios die exponierten Regressionskoeffizienten darstellen.

Variablen in der Gleichung									
		Regressions koeffizient B	Standardfehler	Wald	df	Sig.	Exp(B)	95% Konfidenzintervall für EXP (B)	
			r					Unterer Wert	Oberer Wert
Schritt 1 ^a	Alter bei Diagnosestellung (in Jahren)	,027	,012	5,563	1	,018	1,028	1,005	1,051
	Geschlecht(1)	-,223	,284	,616	1	,433	,801	,459	1,396
	Körpergewicht in kg	-,001	,009	,018	1	,892	,999	,981	1,017
	Körpertemperatur in °C	,867	,116	55,668	1	,000	2,379	1,895	2,987
	Konstante	-36,618	4,657	61,830	1	,000	,000		

a. In Schritt 1 eingegebene Variablen: Alter bei Diagnosestellung (in Jahren), Geschlecht, Körpergewicht in kg, Körpertemperatur in °C.

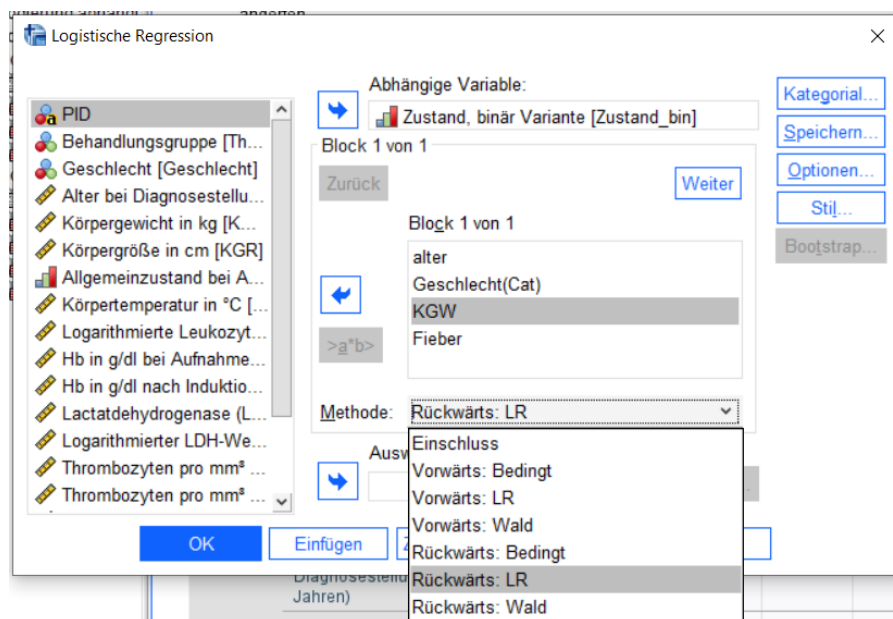
In den beiden Spalten hinter $Exp(B)$ werden die unteren und oberen Grenzen der entsprechenden 95%-Konfidenzintervalle angegeben. Außerdem werden die Regressionskoeffizienten mittels Wald-Tests überprüft. Die jeweiligen p-Werte werden in der Spalte *Sig.* angegeben.

Im vorliegenden Beispiel haben die Variablen „Alter“ und „Fieber“ einen Einfluss auf den Zustand der Patienten. Pro zusätzlichem Jahr im Lebensalter steigt die Chance (Odds) auf eine starke Beeinträchtigung um den Faktor 1,028. Dies entspricht einer Erhöhung um ca. 2,8%. Bei der Körpertemperatur folgt aus der Erhöhung um 1°C eine Erhöhung der Chance auf eine starke körperliche Beeinträchtigung um den Faktor 2,379. Dies entspricht einer Erhöhung um 137,9%.

Bei den kategorialen Variablen ist es wichtig die gewählte Codierung zu beachten, um die Odds Ratios richtig zu interpretieren. Aus der Tabelle *Codierung kategorialer Variablen* ist zu entnehmen, dass die Kategorie „weiblich“ als Referenz verwendet wird und „männlich“ den Alternativzustand darstellt. Damit bedeutet das Odds Ratio von 0,801 in der Tabelle *Variablen in der Gleichung*, dass für Männer ein verringertes Risiko einer starken Beeinträchtigung um den Faktor 0,801 besteht. Dieser Zusammenhang kann allerdings mit einem p-Wert von $p = 0,433$ nicht als bestätigt angesehen werden (95%-Konfidenzintervall: [0,459; 1,396]).

Die in der Tabelle aufgeführte *Konstante* hat keine Bedeutung für die angegebenen Odds Ratios.

Im Rahmen der so genannten Modellwahl trifft man die Entscheidung, welche Einflussvariablen in ein erstelltes Regressionsmodell aufgenommen werden. Es gibt viele verschiedene Möglichkeiten der Modellwahl. Die Auswahl der Einflussvariablen kann inhaltlich begründet sein oder auf univariablen logistischen Modellen beruhen. Ziel ist die Erstellung eines Modells, das die beobachteten Ausprägungen der Zielgröße möglichst gut beschreibt. Im Hauptdialogfenster der **logistischen Regression** kann im Feld *Methode* eine automatisierte Modellwahl angefordert werden. Häufig führt man eine Rückwärts-Selektion per Likelihood-Quotienten-Test (*Rückwärts: LR*, LR für Likelihood Ratio) durch:



Nach Bestätigen mit **OK** wird die angeforderte Variablenselektion durchgeführt.

Bei der Rückwärts-Selektion wird zunächst ein logistisches Regressionsmodell mit allen eingetragenen Kovariaten angepasst. Anschließend wird folgendes schrittweises Verfahren durchgeführt: In jedem Schritt wird für jede Variable überprüft, ob sich die Modellanpassung relevant verschlechtert (z.B. $p \leq 0,10$; dieser Wert kann unter **Optionen** geändert werden), wenn die jeweilige Variable aus dem Modell entfernt wird. Ist dies für keine Variable der Fall, so ist jede Variable offenbar wichtig und das schrittweise Verfahren stoppt. Andernfalls wird die Variable mit dem größten p-Wert aus dem Modell entfernt und es folgt der nächste Selektionsschritt.

Im Ausgabefenster wird die Modellbildung in mehreren Tabellen dargestellt. Die Tabelle *Variablen in der Gleichung* fasst die schrittweise angepassten Modelle, die Odds Ratios sowie die zugehörigen p-Werte und Konfidenzintervalle zusammen. Im letzten Schritt ist das endgültige Modell angegeben, das sich aus der Selektion ergibt. Die Tabelle *Modellieren, wenn Term entfernt* gibt die p-Werte des Likelihood-Quotienten-Tests an. Bei einer Rückwärts-Selektion wird anhand dieser p-Werte die Entscheidung getroffen, ob eine Variable in dem Model verbleibt oder entfernt wird.

Im vorliegenden Beispiel ergibt sich aus der Rückwärts-Selektion ein Modell mit dem Alter und der Körpertemperatur als relevante Einflussvariablen. Mit steigendem Alter steigt das Risiko einer starken Beeinträchtigung um den Faktor 1,028 (95%-Konfidenzintervall: [1,005; 1,051]). Bei der Körpertemperatur wirken sich ebenfalls höhere Werte negativ aus, d.h. aus einem Anstieg um 1°C folgt eine Erhöhung des Risikos einer starken Beeinträchtigung um den Faktor 2,377 (95%-Konfidenzintervall: [1,894; 2,984]).

Grundsätzlich sollten in einer Regressionsanalyse sowohl inhaltliche Überlegungen als auch automatisierte Methoden der Modellwahl angewendet werden, um ein geeignetes Modell zur Beschreibung der Zielgröße zu entwickeln.

Variablen in der Gleichung								95% Konfidenzintervall für EXP (B)	
		Regressionskoeffizient B	Standardfehler	Wald	df	Sig.	Exp(B)	Unterer Wert	Oberer Wert
Schritt 1 ^a	Alter bei Diagnosestellung (in Jahren)	,027	,012	5,563	1	,018	1,028	1,005	1,051
	Geschlecht(1)	-,223	,284	,616	1	,433	,801	,459	1,396
	Körpergewicht in kg	-,001	,009	,018	1	,892	,999	,981	1,017
	Körpertemperatur in °C	,867	,116	55,668	1	,000	2,379	1,895	2,987
	Konstante	-36,618	4,657	61,830	1	,000	,000		
Schritt 2 ^a	Alter bei Diagnosestellung (in Jahren)	,027	,011	5,699	1	,017	1,028	1,005	1,051
	Geschlecht(1)	-,236	,265	,797	1	,372	,790	,470	1,326
	Körpertemperatur in °C	,867	,116	55,731	1	,000	2,380	1,895	2,988
	Konstante	-36,728	4,587	64,109	1	,000	,000		
Schritt 3 ^a	Alter bei Diagnosestellung (in Jahren)	,028	,011	5,924	1	,015	1,028	1,005	1,051
	Körpertemperatur in °C	,866	,116	55,749	1	,000	2,377	1,894	2,984
	Konstante	-36,840	4,581	64,679	1	,000	,000		

a. In Schritt 1 eingegebene Variablen: Alter bei Diagnosestellung (in Jahren), Geschlecht, Körpergewicht in kg, Körpertemperatur in °C.

Modellieren, wenn Term entfernt

Variable		Log-Likelihood des Modells	Änderung der -2 Log-Likelihood	df	Signifikanz der Änderung
Schritt 1	Alter bei Diagnosestellung (in Jahren)	-210,346	6,168	1	,013
	Geschlecht	-207,570	,614	1	,433
	Körpergewicht in kg	-207,272	,018	1	,892
	Körpertemperatur in °C	-236,723	58,921	1	,000
Schritt 2	Alter bei Diagnosestellung (in Jahren)	-210,422	6,300	1	,012
	Geschlecht	-207,670	,797	1	,372
	Körpertemperatur in °C	-236,770	58,998	1	,000
Schritt 3	Alter bei Diagnosestellung (in Jahren)	-210,957	6,574	1	,010
	Körpertemperatur in °C	-237,152	58,964	1	,000

Variablen nicht in der Gleichung

			Wert	df	Sig.
Schritt 2 ^a	Variablen	Körpergewicht in kg	,018	1	,892
	Gesamtstatistik		,018	1	,892
Schritt 3 ^b	Variablen	Geschlecht(1)	,799	1	,371
		Körpergewicht in kg	,199	1	,655
	Gesamtstatistik		,818	2	,664

a. In Schritt 2 entfernte Variablen: Körpergewicht in kg.

b. In Schritt 3 entfernte Variablen: Geschlecht.