

Einführung in die Logistische Regression mit SPSS

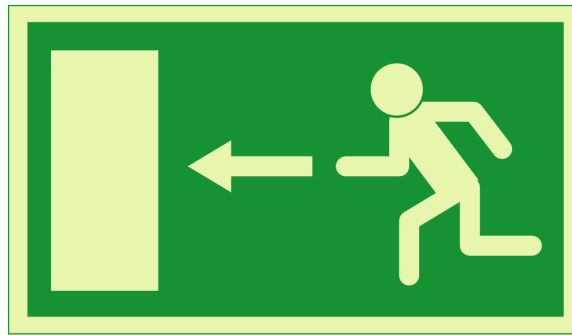
Felix Bittmann

V. 1.0

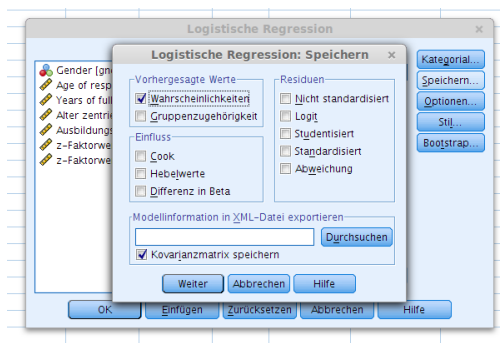
www.felix-bittmann.de

2015

Für Eilige



Daten herunterladen und vorbereiten: S. 6

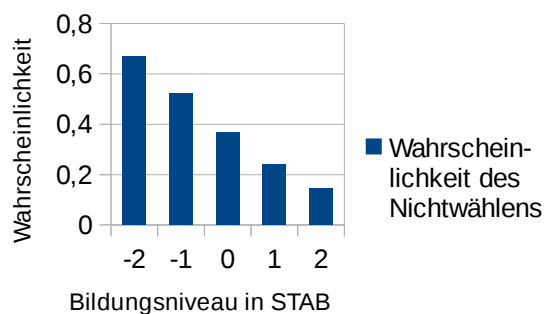


Durchführung in SPSS: S. 13

Interpretation: S. 15

Modellübersicht

Schritt	-2 Log-Likelihood	R-Quadrat nach Cox & Snell	R-Quadrat nach Nagelkerke
1	2265,744 ^a	,061	,100



Ergebnisdarstellung: S. 21

Inhaltsverzeichnis

Einleitung: wann braucht man Logit-Modelle?.....	1
Funktionsweise der Logit-Regression.....	2
Regressionsgleichungen.....	4
Es wird ernst: das Beispiel.....	6
Die Daten herunterladen.....	6
Unser Beispiel.....	6
Der grobe Überblick.....	8
Feintuning: Zentrieren und Standardisieren.....	10
Durchführung der logistischen Regressionsanalyse.....	13
Die Interpretation der Ergebnisse.....	15
Ergebnisdarstellung.....	21
Conditional Effect Plots.....	21
Säulendiagramme.....	24
Literatur und Quellen.....	27

Einleitung: wann braucht man Logit-Modelle?

In diesem Leitfaden wird allgemein vorausgesetzt, dass der Leser mit den Ideen und Methoden der normalen linearen Regression (OLS-Regression) in Grundzügen vertraut ist. Zur kurzen Wiederholung: oftmals ist zwischen bestimmten Beobachtungen ein Zusammenhang erkennbar, was sich beispielsweise durch einen **Korrelationskoeffizienten** ausdrücken lässt. Allerdings erlaubt ein Korrelationskoeffizient einerseits nur den Zusammenhang zwischen **zwei** Variablen zu messen. Andererseits kann alleine aus der Kenntnis des Koeffizienten noch keine **Vorhersage** getroffen werden. Zudem unterscheidet der Korrelationskoeffizient nicht zwischen abhängiger und unabhängiger Variable, es ist also offen, was Ursache und was Wirkung ist.

Durch Regressionsmodelle wird es möglich, Kausalzusammenhänge festzulegen und dadurch mehr Erkenntnisse zu gewinnen. Von großer Bedeutung ist auch die Möglichkeit, **Vorhersagen** treffen zu können, wenn eine oder mehrere unabhängige Variablen bekannt sind. Allerdings können auch Regressionsmodelle nicht entscheiden, welche Variablen Ursache und welche Wirkungen sind, dies muss immer vor der Untersuchung auf Basis von Theorien und Plausibilitätsargumenten durch den Anwender erfolgen. In der Regel sind diese Festsetzungen jedoch recht einfach möglich und können als gültig angesehen werden. So kann beispielsweise das Geschlecht einer Person deren Alkoholkonsum beeinflussen, umgekehrt funktioniert das nicht. Hat man eine solche Kausalbeziehung festgelegt, hat man eine **abhängige Variable** (AV, also die Variable, die vorhergesagt werden soll) und eine oder mehrere **unabhängige Variablen** (UV, also die Variablen, die zur Vorhersage der abhängigen Variable herangezogen werden sollen). Möchte man das obige Beispiel aufgreifen, könnten etwa die Variablen Geschlecht, Alter und Bildungslevel einer Person benutzt werden, um deren Alkoholkonsum vorherzusagen.

Diese Annahmen gelten grundsätzlich für alle Regressionsmodelle. Um zu verstehen, wann eine **Logit-Regression** sinnvoll ist, ist es wichtig, sich den Charakter der herangezogenen Variablen zu verdeutlichen. Grob gesagt werden drei verschiedene Arten unterschieden (so auch in SPSS). Zwar sind auch feinere Abstufungen möglich, jedoch an dieser Stelle nicht von Relevanz. Man unterscheidet genauer gesagt verschiedene **Skalenniveaus** der Variablen. Dieses Niveau gibt den Aussagegehalt einer Variable an und bestimmt, welche statistischen Rechenoperationen mit diesen Variablen durchgeführt werden können. Variablen können nominal, ordinal oder metrisch skaliert sein. Für die lineare Regression wird verlangt, dass die abhängige Variable metrisch skaliert ist. So könnte man beispielsweise das Einkommen in Euro, den Alkoholkonsum in Millilitern oder die Temperatur in Grad Celsius vorhersagen, allerdings nicht das Geschlecht, die Religion oder Schulausbildung einer Person, da diese Variablen nominal oder ordinal skaliert sind. Jedoch gibt es in der Realität bestimmte Variablen, die eine sehr spezielle Skalierung aufweisen, nämlich nur genau zwei verschiedene Werte annehmen können. Man spricht auch von **dichotomen** oder **binären** Variablen. Meistens werden diese Werte dann mit zwei Zahlen unterschieden, etwa 0 und 1. So kann eine Frau entweder schwanger sein (1) oder eben auch nicht (0). Zwischenstufen sind nicht möglich, da man nicht „etwas“ schwanger sein kann. Ebenso ist eine Person entweder HIV-positiv oder negativ, also krank oder gesund. Auch viele Entscheidungen können so beschrieben werden. Eine Person kann zur Bundestagswahl gehen oder auch nicht, aber eine Dritte Möglichkeit

ist nicht vorgesehen. Anzumerken ist jedoch, dass die Einteilung einer Variable durchaus auch von der **Fragestellung** abhängt. Für Sozialwissenschaften könnte es interessant sein zu prüfen, welche Variablen eine Aussage darüber machen, ob eine Person HIV-positiv oder negativ ist, beispielsweise das Geschlecht, der Bildungsstand oder die sexuelle Präferenz. In diesem Fall würde man die Variable „HIV“ als dichotom werten. Für eine medizinische Studie, die neue Medikamente gegen HIV testet, wäre es jedoch sinnvoller, von der dichotomen Variable abzurücken und vielmehr die Virenlast der Patienten zu messen. So wäre ein Medikament sehr wirksam, das die Virenlast etwa von 100 Viren pro ml auf 5 Viren pro ml senkt. HIV-positiv wäre die Personen natürlich weiterhin, dennoch wäre die Variable hier als metrisch anzunehmen.

Die Logit-Regression wird für uns immer dann interessant, wenn wir die abhängige Variable als dichotom annehmen. Zwar ist es letztlich so, dass es immer nur zwei verschiedene Möglichkeiten gibt, dennoch wird es möglich, **Wahrscheinlichkeiten** angeben zu können. So erlaubt uns die Logit-Regression beispielsweise anzugeben, wie wahrscheinlich es ist, dass eine bestimmte Person zur Bundestagswahl geht, wenn wir Geschlecht, Parteipräferenz und Alter der Person kennen.

Es ist wichtig, sich den Unterschied zu linearen Regression zu verdeutlichen. Diese erlaubt, den **Wert** einer metrischen Variable vorherzusagen. Ein Schüler wird x Punkte in einer Klausur erreichen, ein Medikament wird den Blutdruck um y mmHg senken, eine Person wird z Euro pro Monat verdienen. Die Logit-Regression hingegen gibt **Wahrscheinlichkeiten** an: eine Frau wird mit einer Wahrscheinlichkeit von $X\%$ schwanger sein, eine Person wird mit einer Wahrscheinlichkeit von $Y\%$ zur Wahl gehen, ein Sportler wird mit einer Wahrscheinlichkeit von $Z\%$ das Rennen gewinnen. Die unabhängigen Variablen, die wir dann zur Vorhersage heranziehen wollen, können hingegen **beliebig** skaliert sein.

Funktionsweise der Logit-Regression

An dieser Stelle wollen wir nicht in die mathematischen Details einsteigen, da wir diese den Mathematikern und Statistikern vorbehalten wollen. Jedoch ist es später zur Interpretation unserer Ergebnisse sehr wichtig, die verschiedenen Ebenen zu verstehen, die bei der Berechnung zur Anwendung kommen. Ohne dieses Wissen wird es uns nicht möglich sein, die SPSS-Tabellen zu verstehen und in die Alltagssprache übersetzen zu können. Es sei also angemerkt, dass nicht das Auswendiglernen von Formeln wichtig ist, sondern das grundsätzliche Verstehen, wie bestimmte Ebenen bestimmte Ergebnisse angeben.

Ebene 1: Wahrscheinlichkeiten. Diese Ebene ist uns im Alltag sehr vertraut, sodass wir später versuchen werden, alle Ergebnisse irgendwie wieder in dieser Ebene auszudrücken. Folgende Aussagen sind beispielsweise leicht für jedermann verständlich:

- Die Wahrscheinlichkeit, dass eine Münze Kopf zeigt, liegt bei 50 %.
- Die Wahrscheinlichkeit, dass ein katholischer Renter die CDU wählt, liegt bei 69 %.

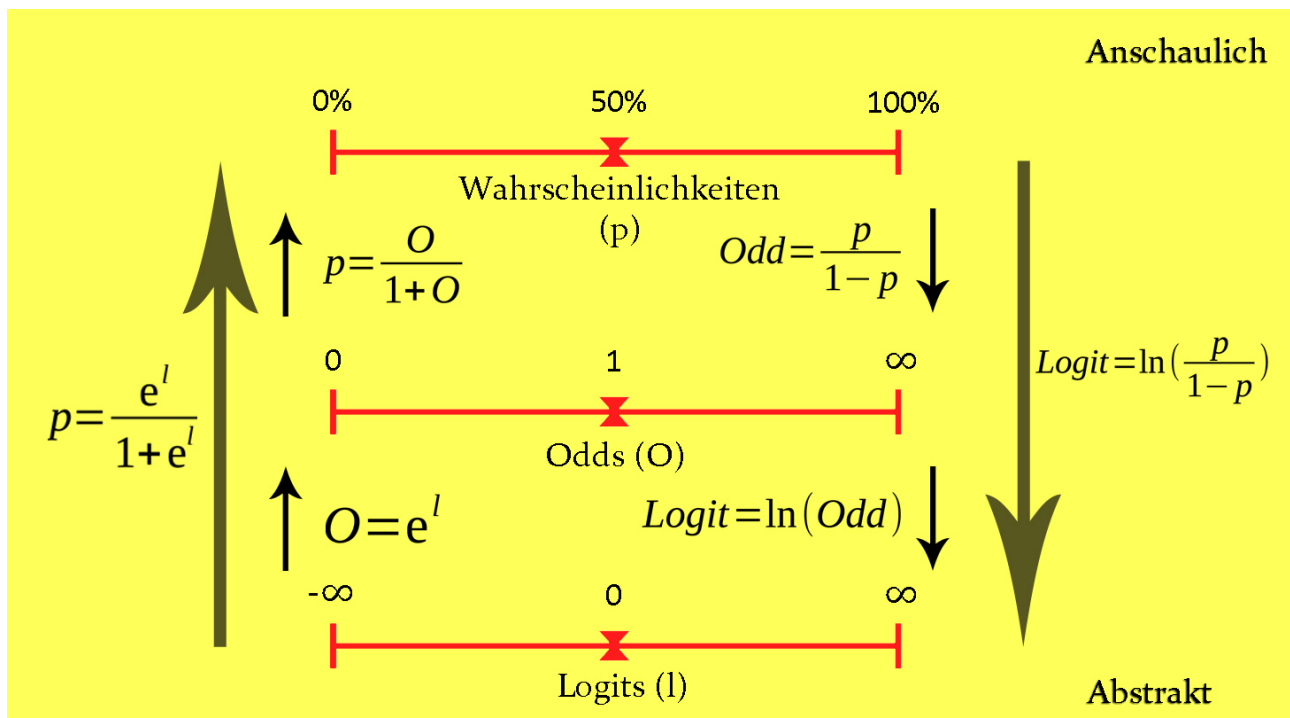
- Die Wahrscheinlichkeit, dass eine heterosexuelle Frau mit Universitätsabschluss HIV-positiv ist, liegt bei 0,1 %.

Allgemein kann eine Wahrscheinlichkeit stets nur zwischen 0 und 1 liegen, also zwischen 0 % und 100 %. Wahrscheinlichkeiten über 100 % sind genauso unmöglich wie solche unter 0. Eine Wahrscheinlichkeit von 100 % spricht für ein sicheres Ergebnis, eine Wahrscheinlichkeit von 0 % für ein unmögliches. In der Realität werden alle Wahrscheinlichkeiten zwischen diesen beiden Extremen liegen.

Ebene 2: Odds. Wahrscheinlichkeiten lassen sich durch eine einfache Formel in die sog. Odd-Ebene übertragen. Diese Formel lautet $Odd = \frac{p}{1-p}$ wobei p für die Wahrscheinlichkeit steht. Die

Wahrscheinlichkeit wird zwischen 0 und 1 angegeben. Betrachtet man die beiden Extremfälle, 0 und 1, wird auch die maximale Breite der Odd-Ebene deutlich. Eine Wahrscheinlichkeit von 0 entspricht einem Odd von 0, eine Wahrscheinlichkeit von 1 entspricht einem Odd von unendlich.

Eine Wahrscheinlichkeit von 50 % entspricht einem Odd von 1, weil: $Odd = \frac{0,5}{1-0,5} = 1$.



Ebene 3: Logits. Für die mathematische Handhabung ist es jedoch sinnvoller, statt der Odd-Ebene noch eine dritte Ebene einzuführen, nämlich die Logit-Ebene. Der einzige Unterschied besteht darin, dass alle Odds zur Basis e (Eulersche Zahl) logarithmiert werden. Kurz: $Logit = \ln(Odd)$. Folgende Tabelle gibt eine kurze Übersicht der drei Ebenen an. Auf der Logit-Ebene können alle Werte zwischen $-\infty$ und $+\infty$ angenommen werden. Die Mitte liegt daher bei 0.

Wahrscheinlichkeiten	→	Odds	→	Logits
p	$\frac{p}{1-p}$	o	$\ln(o) = \ln\left(\frac{p}{1-p}\right)$	l

Eine Rückrechnung erfolgt analog, es müssen nur die Formeln umgestellt werden:

Wahrscheinlichkeiten	←	Odds	←	Logits
p	$p = \left(\frac{o}{1+o}\right) = \frac{e^l}{1+e^l}$	o	$o = e^l$	l

Beispiel: eine Wahrscheinlichkeit von 0,35 (35 %) entspricht eine Odd von 0,538 und einem Logit von -0,619. Umgekehrt entspricht ein Logit von 3 einem Odd von 20,09 und einer Wahrscheinlichkeit von 0,9526 (95,26 %). Wie auch die Farbgrafik oben aufzeigt, lassen sich alle Werte ineinander umrechnen. Dies wird später für uns wichtig werden, weil SPSS diese Umrechnungen nur teilweise vornimmt und der Rest manuell getätigt werden muss.

Regressionsgleichungen

Das Ziel einer Regression ist es letztlich, die abhängige Variable vorhersagen zu können. Eine solche Vorhersage wird dann mit einer Gleichung möglich, in die bestimmte Werte einsetzen werden können, um bestimmte Ergebnisse zu erhalten. Wir wollen dieses Konzept an einem einfachen (fiktiven) Beispiel wiederholen.

Untersucht wurde, wie lange Kinder und Jugendliche benötigen, um eine festgelegte Distanz zu joggen. Als abhängige Variable wird die gemessene Zeit herangezogen, gemessen in Minuten, als unabhängige Variable wird das Alter der Kinder in Jahren verwendet. Ziel ist es nun eine Funktion zu finden, in die man ein beliebiges Alter einsetzt und am Ende eine Zeit in Minuten herauskommt. Dies kann man so formulieren: $y_i = a + b \cdot x_i$

Dabei ist y_i die Zeit in Minuten, die Person i laut Vorhersage benötigen wird. In der Formel ist a der Achsenabschnitt, eine Konstante, und b Steigungskoeffizient der Variable Alter. x_i ist dann das Alter der Person i in Jahren. Beispielsweise könnten die Daten folgende Gleichung ergeben:

$y_i = 33,6 - 1,64 \cdot x_i$ Dies würde bedeuten, dass eine Person mit einem Alter von 15 Jahren $33,6 - 24,6 = 9$ Minuten zur Bewältigung der Strecke benötigen wird. Zwei wichtige Erkenntnisse lassen sich aus dieser Gleichung ableiten: was passiert, wenn wir ein Alter von 0 in die Gleichung einsetzen und was passiert, wenn das Ergebnis negativ wird? Zunächst besagt die Gleichung, dass eine Person, die 0 Jahre alt ist, 33,6 Minuten für die Strecke benötigen wird. Dieses Ergebnis ist sinnlos, da jeder Mensch älter als 0 Jahre sein muss. Wir sehen, dass die Aussagekraft der Gleichung begrenzt ist. Mathematisch ist alles völlig korrekt, aber aufgrund von Erfahrungen müssen wir einschreiten, wenn Ergebnisse unsinnig werden. Dies kann nur durch **Mitdenken**

erreicht werden, auch ein Computer kann diese Probleme nicht alleine lösen. Wie wir dieses Problem abmildern können, werden wir weiter unten besprechen. Auch wird deutlich, dass Ergebnisse irgendwann negativ werden können, nämlich dann, wenn das Produkt aus Alter und Koeffizient größer als 33,6 werden. Eine Person mit einem Alter von 25 Jahren würde beispielsweise ein Ergebnis von -7,4 Minuten erreichen. Dies ist ebenfalls unmöglich, denn dann würde sie ankommen, bevor sie losgelaufen wäre. Wir sehen also, dass unsere Gleichung auf einen bestimmten Bereich beschränkt ist, in dem sie sinnvolle Ergebnisse liefert. Auch wenn unsere Daten sehr gut sind und Vorhersagen prinzipiell möglich sind, stoßen mathematische Modelle immer an Grenzen.

Jedes statistische Modell ist letztlich fehlspezifiziert. Unsere Aufgabe ist es, das am wenigsten schlechte Modell zu finden und dessen Aussagekraft immer kritisch zu beurteilen.

Damit ist gemeint, dass alle Modelle, die nicht reine Mathematik thematisieren, komplex sind und immer nur eine Annäherung an die Realität erlauben. Es mag sein, dass wir 1.000 Kinder untersucht haben und obige Gleichung gefunden habe. Deutlich wird aber auch, dass eine sehr gute Abbildung der Realität nicht möglich ist, da wir unendlich viele weitere Faktoren unberücksichtigt lassen. Kinder, die in einem Sportverein sind, werden besser abschneiden als untrainierte Kinder, Kinder mit einem gebrochenen Bein werden schlechter abschneiden als gesunde Kinder, etc... Da es insgesamt unmöglich ist, alle Faktoren zu berücksichtigen, muss es unser Ziel sein, die wichtigsten Faktoren ausfindig zu machen, sodass eine adäquate Beschreibung der Realität möglich wird. Klar sein muss aber auch, dass kein Regressionsmodell eine perfekte Vorhersage ermöglichen wird. Statistik ist immer mit Unsicherheit behaftet (wie das Leben an sich auch...).

Es wird ernst: das Beispiel

Die Daten herunterladen

Nach der Einführung in die Theorie der Regression wollen wir uns nun dem konkreten Beispiel nähern. Als Datengrundlage dient der **European Social Survey** (Stand: 2012), eine Umfrage, die regelmäßig in ganz Europa durchgeführt wird. Allerdings werden wir uns hier auf Daten aus Deutschland beschränken, um das Beispiel nicht zu komplex werden zu lassen. Um die Datensätze im SPSS Format (.sav) aufrufen zu können, folgt man folgendem Link:

<http://www.europeansocialsurvey.org/download.html?file=ESS6DE&c=DE&y=2012>

Dort kann man die Daten im richtigen Format herunterladen und anschließend entpacken. Vor dem Download muss man sich auf der Website registrieren und zustimmen, die Daten nur für bestimmte Zwecke zu nutzen. Für alle privaten und wissenschaftlichen Anwendungen sollte dies kein Problem sein. Der Download ist **kostenlos**.

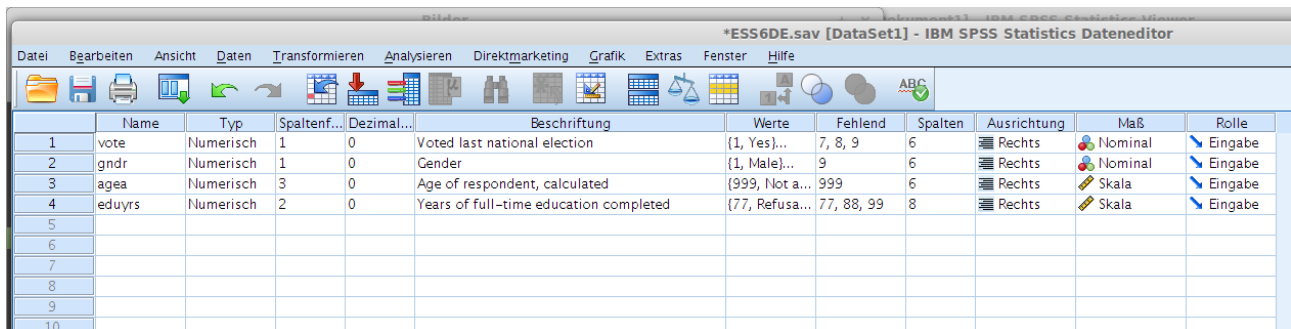
Nun können wir SPSS starten und den Datensatz laden. Ich verwende im Beispiel die Version 22 unter Linux. Bei älteren Versionen wird die Grafik möglicherweise leicht abweichen, auch einige spezielle Funktionen könnten fehlen. Im Grunde werden aber alle nötigen Komponenten auch enthalten sein, sodass eine Durchführung möglich wird. Wenn wir den Datensatz geöffnet haben, können wir uns sowohl die Daten als auch die Variablen ansehen (zwei Schaltflächen unten links). Wir sehen, dass der Datensatz extrem umfangreich ist und über 600 Variablen von mehreren tausend Befragten enthält. Grundsätzlich ist das klasse, da wir damit sehr gute und vollständige Daten haben, die eine hohe wissenschaftliche Aussagekraft und Güte besitzen. Andererseits kann uns diese Menge an Daten auch überfordern. Wir werden uns daher hier auf ein relativ einfaches Beispiel mit wenigen Variablen beschränken.

Unser Beispiel

Unsere Fragestellung für das Beispiel wird sein: **mit welcher Wahrscheinlichkeit wird eine bestimmte Person zur Wahl gehen** (z.B. zur Bundestagswahl)? Zur Beantwortung dieser Frage werden wir drei unabhängige Variablen heranziehen: das Alter der Person, die Ausbildungsdauer der Person sowie das Geschlecht der Person.

Typ der Variable	Aussage der Variable	Name der Variable im Datensatz	Skalierung
Abhängige Variable	Geht die Person zur Wahl?	vote	Dichotom
Unabhängige Variablen	Geschlecht der Person	gndr	Dichotom
	Ausbildungsniveau der Person	eduys	Metrisch
	Alter der Person	agea	Metrisch

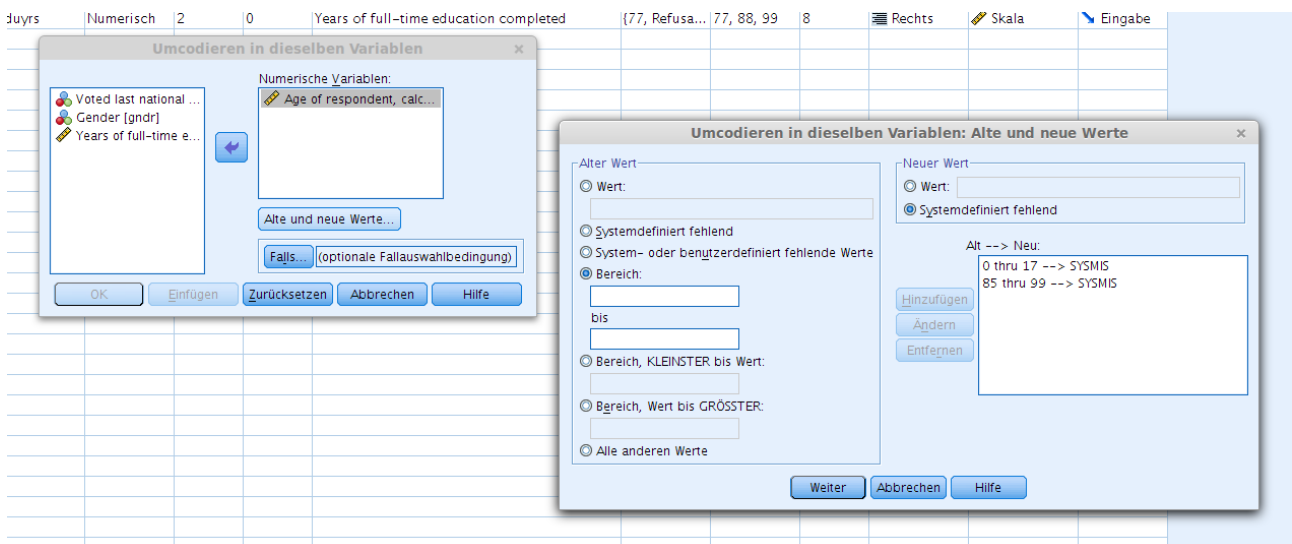
Um die Übersicht zu behalten, habe ich zunächst sämtliche anderen Variablen aus dem Datensatz gelöscht. Dann habe ich den Datensatz unter einem neuen Namen gespeichert, um die alten Daten nicht zu verlieren. Danach sollte SPSS in etwa so aussehen:



	Name	Typ	Spalten...	Dezimal...	Beschriftung	Werte	Fehlend	Spalten	Ausrichtung	Maß	Rolle
1	vote	Numerisch	1	0	Voted last national election	{1, Yes}...	7, 8, 9	6	Rechts	Nominal	Eingabe
2	gndr	Numerisch	1	0	Gender	{1, Male}...	9	6	Rechts	Nominal	Eingabe
3	agea	Numerisch	3	0	Age of respondent, calculated	{999, Not a... 999	999	6	Rechts	Skala	Eingabe
4	eduys	Numerisch	2	0	Years of full-time education completed	{77, Refusa... 77, 88, 99	77, 88, 99	8	Rechts	Skala	Eingabe
5											
6											
7											
8											
9											
10											

Zunächst müssen die Daten allerdings vorbereitet werden, damit wir aussagekräftige und korrekte Ergebnisse erhalten. Dabei kann es hilfreich sein, sich zuerst deskriptive Statistiken, Häufigkeitstabellen und Grafiken für jede Variable anzusehen. Dabei fällt beispielsweise auf, dass auch Personen im Datensatz enthalten sind, die jünger als 18 Jahre sind und daher noch gar nicht wählen dürfen. Auch einige sehr alte Personen sind enthalten, die möglicherweise zu alt sind und an Wahlen nicht mehr teilnehmen können. Wir werden daher alle Personen für die Untersuchung ausschließen, die jünger als 18 Jahre und älter als 84 Jahre sind.

Dazu gehen wir in SPSS auf **Transformieren** → **Umcodieren in dieselben Variablen**. Dort wählen wir die Variable *agea* aus, verschieben sie nach rechts in das Fenster und klicken auf **Alte und neue Werte**. Es öffnet sich ein Fenster, in dem wir bestimmte Bereiche definieren können. Wir geben jeweils die Grenzen ein und setzen sie auf **Systemdefiniert fehlend**.

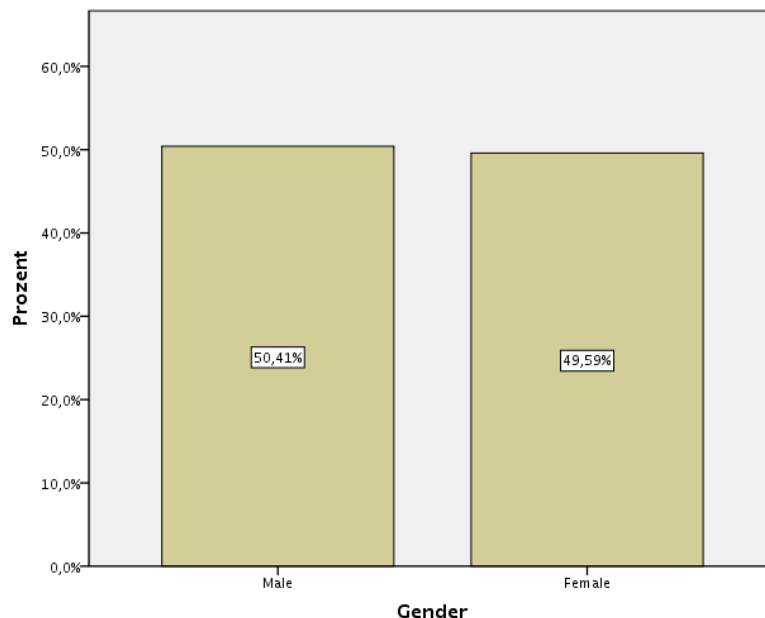


Danach klicken wir auf **Weiter** und dann auf **OK**. Lässt man sich nun eine Häufigkeitstabelle der Variable *agea* ausgeben wird man feststellen, dass nur noch Werte zwischen 18 und 84 auftauchen, die anderen Werte werden nun als fehlende Werte gezählt. Sehr ähnlich gehen wir bei der Variable *eduys* vor. Hier möchten wir alle Personen erfassen, die eine Ausbildungsdauer über 21 Jahren haben. Dies sind einige Ausreißer, die wohl keinen bestimmten Effekt mehr auf die

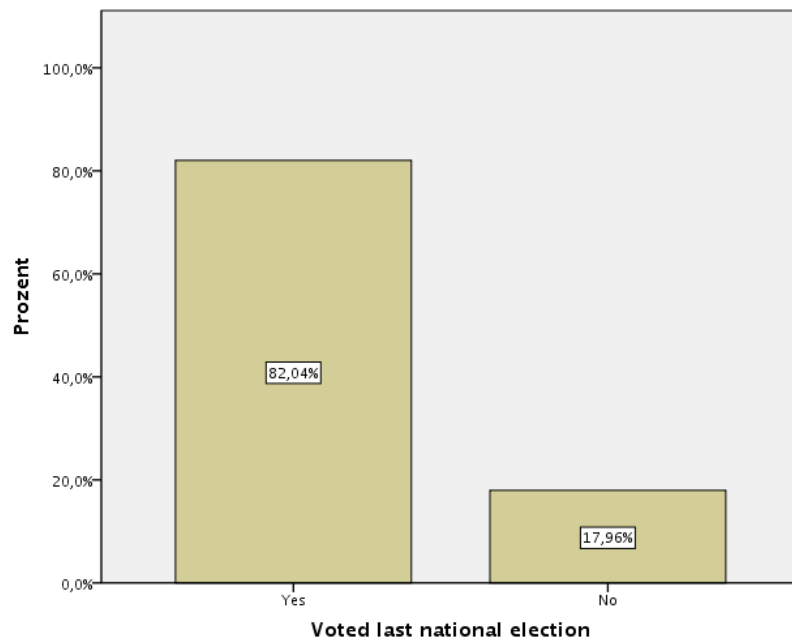
Wahlbereitschaft haben, da eine Ausbildungsdauer von 21 Jahren bereits extrem lange ist. Wir wählen wieder **Umcodieren in dieselben Variablen**, wählen dort die Variable *eduyrs* aus und setzen alle Werte zwischen 22 und 99 auf 21. Jemand, der also beispielsweise 30 Jahre Ausbildungszeit hat, wird ab nun mit 21 Jahre Ausbildungszeit geführt. Die Fälle werden also nicht ausgeschlossen, sie werden nur auf eine bestimmte Obergrenze normiert. Auch dies können wir mit einer Häufigkeitstabelle kontrollieren. Zuletzt setzen wir noch die Personen, die aus anderen Gründen nicht wahlberechtigt sind, ebenfalls auf Missing values. Dazu klicken wir in der Variablenansicht in der Zeile *vote* auf die Werte bei **fehlend** und geben als Bereich der fehlenden Werte 3 bis 11 ein. Damit ist unsere Vorbereitung der Daten vorerst abgeschlossen.

Der grobe Überblick

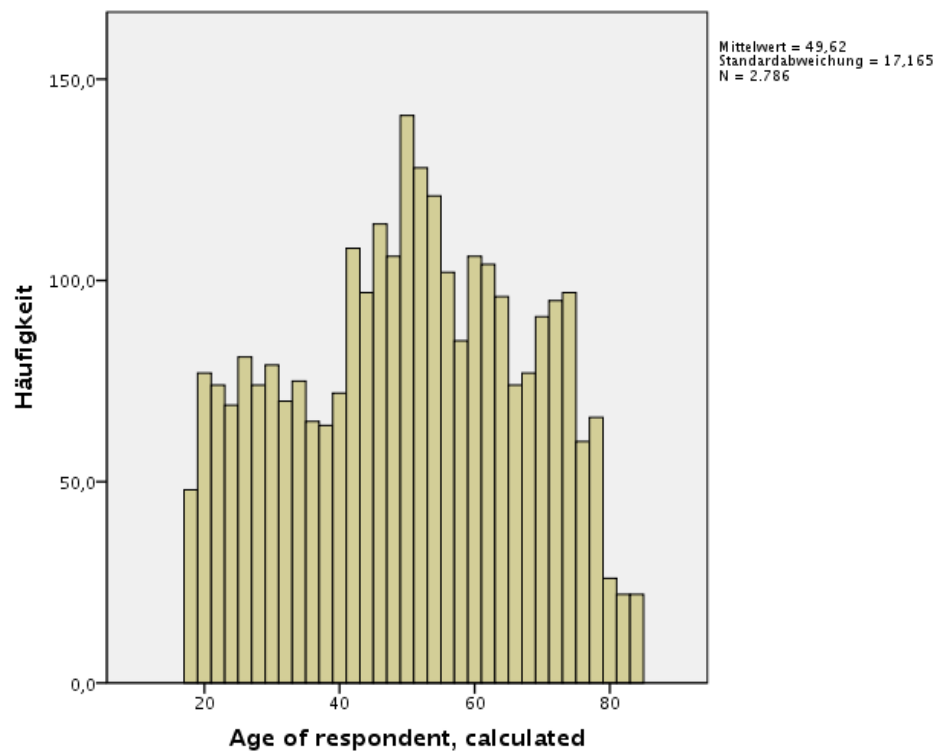
Bevor wir mit der eigentlichen Untersuchung und logistischen Regression beginnen, wollen wir uns die Daten kurz ansehen. Dazu verwenden wir einfache deskriptive Statistiken und vor allem Schaubilder. Ich habe für jede der vier Variablen eine schlichte aber informative Grafik erstellt. Diese sollten kurz besprochen werden:



Bei der Variable *gndr* fällt auf, fast gleich viele Männer und Frauen befragt wurden, die Zahlen unterscheiden sich nur minimal. Dies deutet auf eine hohe Datengüte hin. Wir können annehmen, dass ungefähr gleich viele Männer und Frauen in Deutschland leben. Würden die Daten extremen Unterschiede offenbaren, könnte dies auf eine Verzerrung der Daten hinweisen. Unsere Aussagen wären dann möglicherweise stark verfälscht. Als Wahl der Darstellung wurde ein Balkendiagramm gewählt.

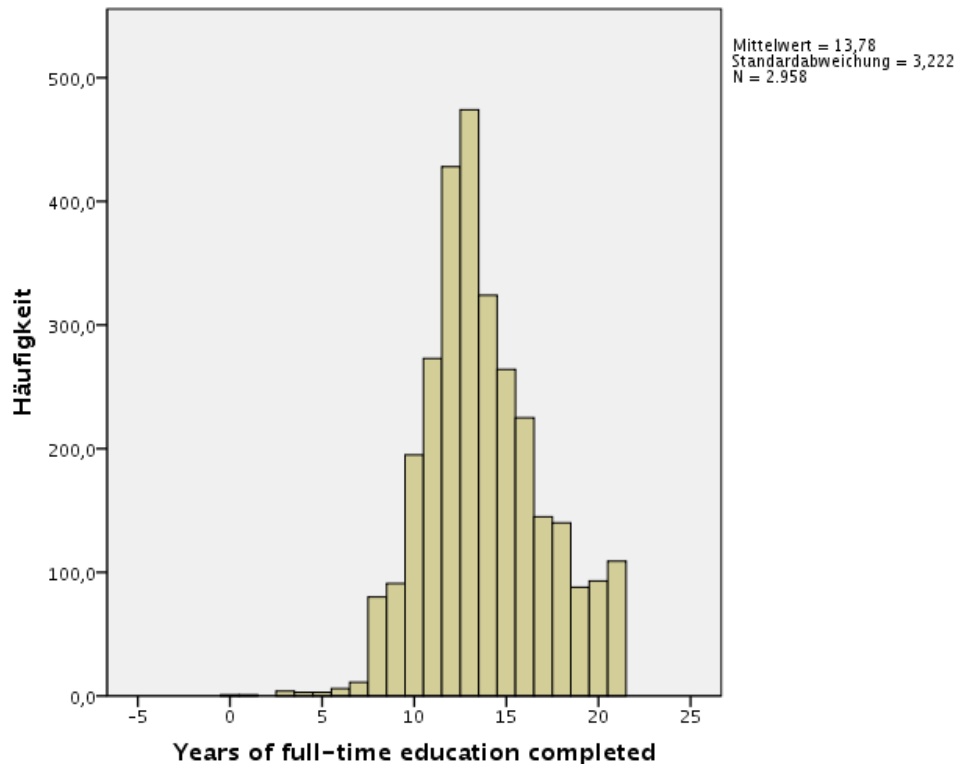


Die Variable *vote* ist hingegen sehr asymmetrisch verteilt. Über 80% aller Befragten haben bei den letzten Wahlen ihre Stimme abgegeben. Ob dies realistisch ist, können wir nicht nachprüfen, hier müssen wir uns auf die Daten verlassen. Eine so schiefe Verteilung der abhängigen Variable ist nicht unproblematisch und wird uns bei der Interpretation der Daten noch beschäftigen. Auch hier haben wir zur Veranschaulichung ein Balkendiagramm gewählt.



Bei der Variable *agea* haben wir hingegen ein Histogramm verwendet, um die Daten zu

veranschaulichen. Histogramme sind vor allem bei metrischen Variablen eine gute Form der Darstellung. Es fällt auf, dass keine Lücken vorhanden sind und Personen jeden Alters befragt wurden (bis auf die Fälle, die wir ausgeschlossen haben). Der Charakter einer Normalverteilung wird leicht deutlich. Wie der Mittelwert aussagt, ist die durchschnittliche Person fast 14 Jahre alt, was einem recht hohen Durchschnittsalter entspricht, aber in einer alternden Gesellschaft wie Deutschland normal ist. Zumal ist der Wert höher als im Bundesdurchschnitt, da wir ja alle Personen unter 18 ausgeschlossen haben.

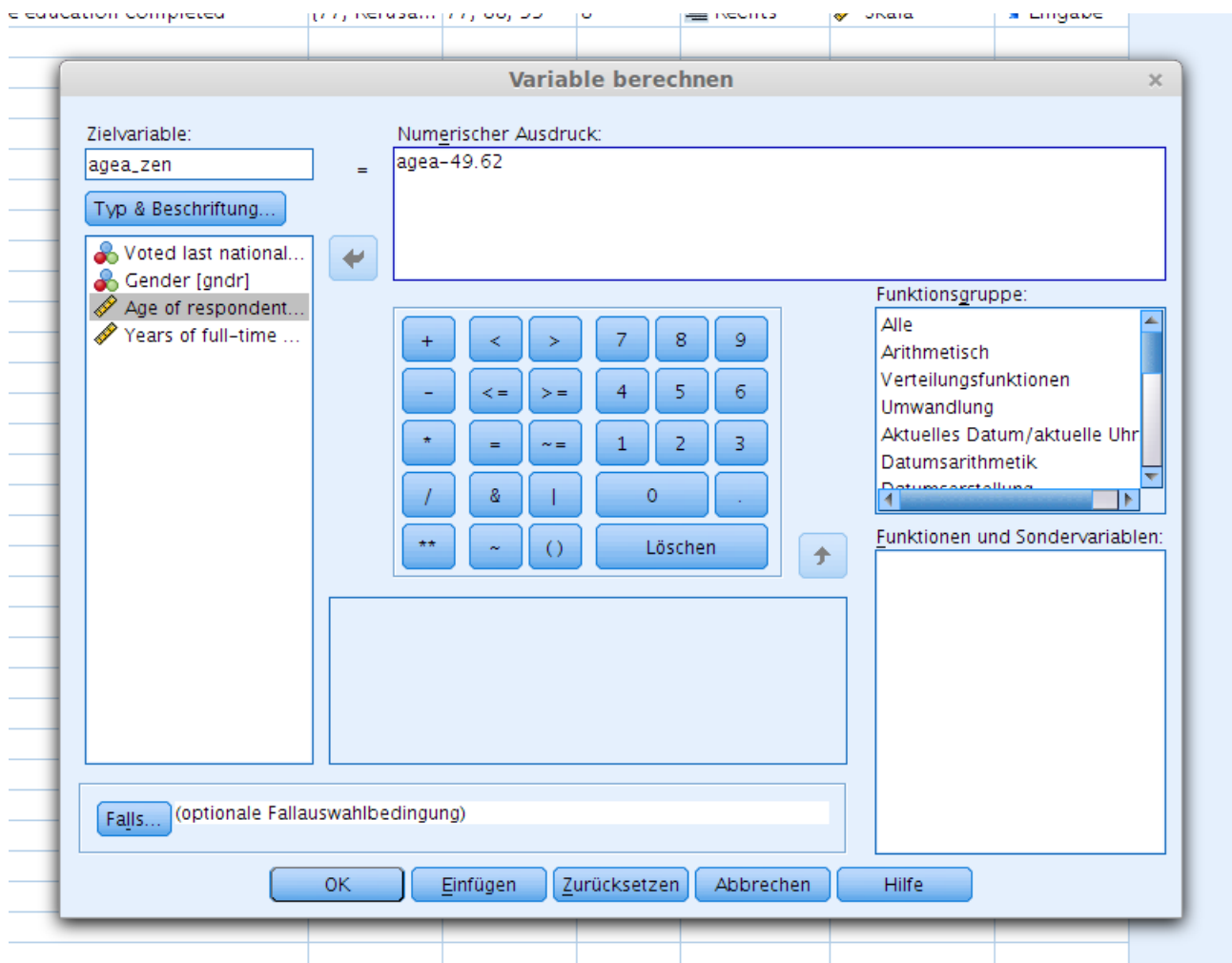


Bei der Variable *eduyrs* fällt auf, dass die Bildung insgesamt recht hoch ist. Sehr viele Menschen scheinen mehr als 13 Jahre zur Schule gegangen zu sein, was einem Gymnasialabschluss entsprechen dürfte. Die Anzahl der Personen mit einer sehr geringen Bildung (kleiner als 7 Jahre) ist äußerst gering. Auch hier erfolgt die Darstellung mit einem Histogramm.

Feintuning: Zentrieren und Standardisieren

Eigentlich könnten wir nun mit der Berechnung der logistischen Regression beginnen. Allerdings wollen wir die metrischen Variablen noch leicht anpassen, was eine spätere Vergleichbarkeit erleichtert und bei der Interpretation der Daten eine Rolle spielt. Die Ursache für diese Anpassung ist folgende: wie wir in der Theorie gesehen haben, gibt der Achsenabschnitt in der Regressionsgleichung die Wahrscheinlichkeit für die abhängige Variable an, wenn **alle unabhängigen** Variablen den Wert 0 annehmen. Dies kann zu Problemen führen, wie am Beispiel der negativen Laufzeit der Kinder verdeutlicht wurde (vgl. S. 5). Um dieses Problem zu umgehen, werden wir die metrischen Variablen **zentrieren**. Dadurch wird erreicht, dass in der späteren Regressionsgleichung nicht mehr über den Wert 0 eine Aussage getroffen wird, sondern über den

Mittelwert der Variable. Der Mittelwert der Variable *agea* ist 49,62, bei der Variable *edyrs* ist er 13,78. Um diese Zentrierung zu erreichen, klicken wir auf **Transformieren** → **Variable berechnen**. Wir erstellen eine neue Variable, die wir *agea_zen* nennen. Wir wählen die Variable *agea* aus und ziehen sie in das rechte Fenster. Dann schreiben wir einfach -49,62 dahinter. Das bedeutet, dass nun für jede Person dieser Wert vom Alter der Person subtrahiert wird und das Ergebnis als neue Variable gespeichert wird. Eine Person, die beispielsweise 70 Jahre alt ist, erreicht nun einen Wert von 20,38; eine Person, die 30 Jahre alt ist, einen Wert von -19,62. Dann klicken wir auf **OK**.

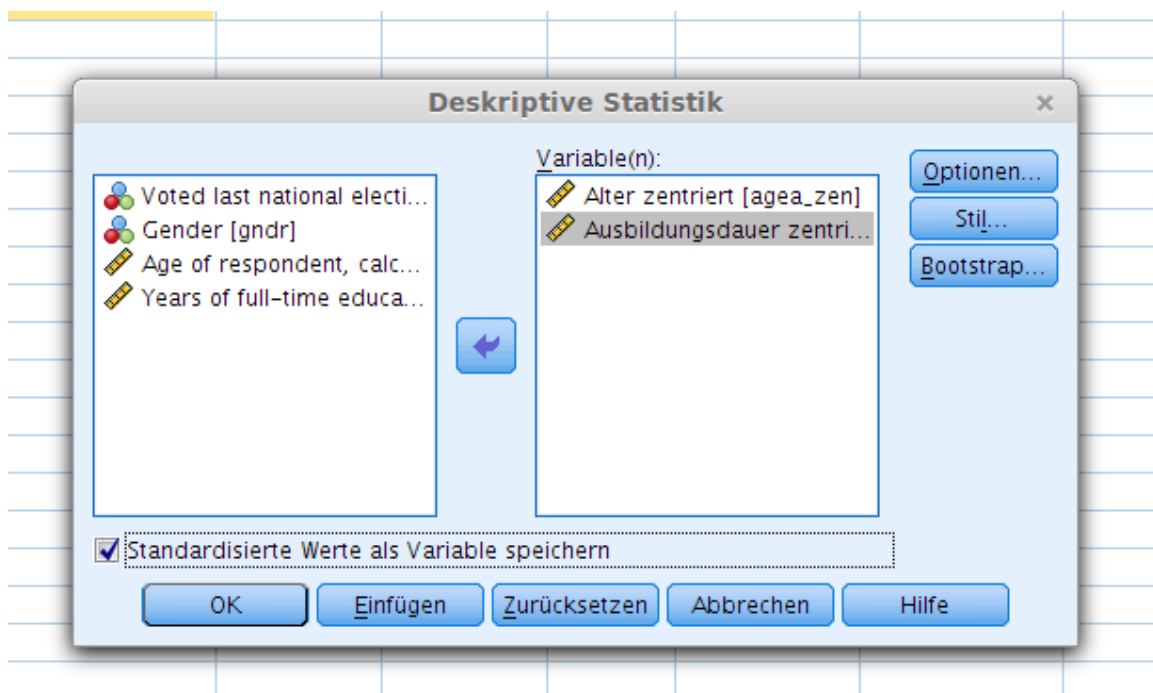


Ganz analog gehen wir für die Variable *edyrs* vor. Die neue Variable nennen wir *edyrs_zen*. Hier subtrahieren wir nun 13,78. Damit haben wir nun zwei neue Variablen erstellt, die wir für alle weiteren Anwendungen nutzen können.

Zuletzt wollen wir die beiden metrischen Variablen noch **standardisieren**. Grundgedanke dieser Operation ist folgender: später werden wir eine Regressionsgleichung mit mehrere Koeffizienten erhalten (jede unabhängige Variable wird später durch einen Koeffizienten dargestellt). Dann können wir die jeweiligen x-Werte variieren und sehen, wie sich die abhängige Variable verändern. Beispielsweise werden wir prüfen können, welchen Einfluss 10 Jahre mehr an Alter einer Person die Wahrscheinlichkeit verändern, ob diese wählen geht oder nicht. Allerdings ist es fast immer so, dass die Skalen der metrischen Variablen nicht übereinstimmen. Würden wir Monatseinkommen

und Alter einer Person als Variablen wählen, so wäre dies offensichtlich, denn die Maßeinheiten sind Euro und Jahre. Diese sind eindeutig nicht gleich und können nicht direkt miteinander verglichen werden. Auch in unserem Beispiel ist dies ein Problem: Lebensalter und Ausbildungsdauer werden zwar beide in „Jahren“ gemessen, doch entspricht eine Ausbildungsdauer von 10 Jahren auch 3650 Tagen? Eher nicht. Wir werden deshalb die Variablen standardisieren, indem wir für jede Variable über die deskriptiven Statistiken die **Standardabweichung** berechnen und anschließend jeden Wert durch diese dividieren. Wir erhalten dann eine neue Variable mit einer neuen Skalierung.

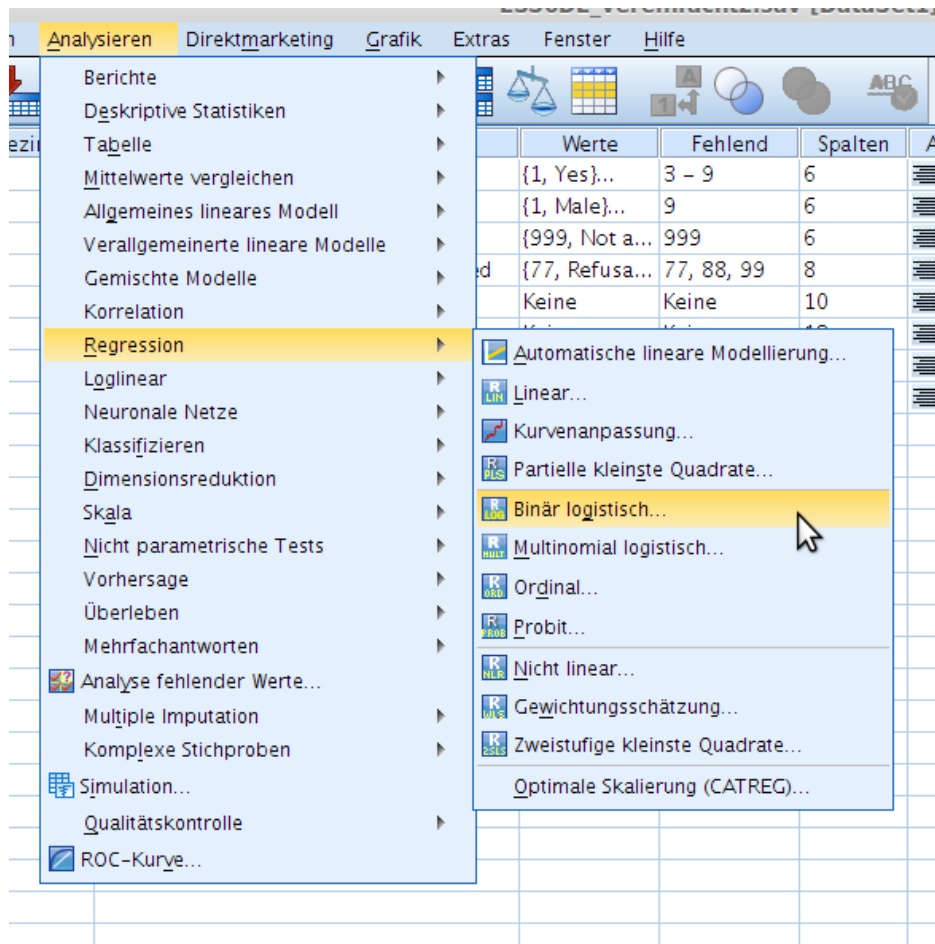
Beispiel: Angenommen, die Standardabweichung des Alters sei 7 Jahre. Wir ziehen die zentrierte Altersvariable *agea_zen* heran und stellen fest, dass eine Person mit einem durchschnittlichen Alter (ca. 50 Jahren) dort einen Wert von 0 erreicht. Eine Person von 57 Jahren würde nun auf der standardisierten Skala einen Wert von 1 erreichen, da sie 1 Standardabweichung vom Mittelwert entfernt ist. Eine Person, die 43 Jahre alt wäre, würde dementsprechend einen Wert von -1 erhalten. Somit interpretieren wir nicht in Jahren, sondern in Standardabweichungen. Dies ist prinzipiell nicht schwierig und kann auch anschaulich sein, sofern man die vorher berechnete Standardabweichung wieder in Jahren oder einer anderen bekannten Einheit ausdrücken kann.



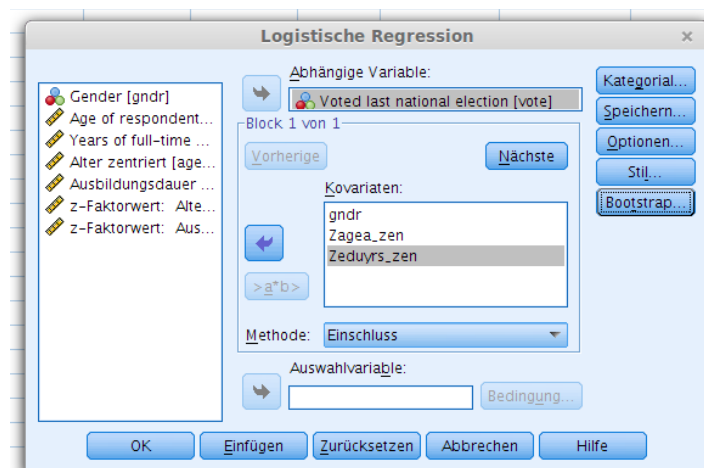
Glücklicherweise nimmt uns SPSS die Standardisierung ab. Dazu klicken wir auf **Analysieren** → **Deskriptive Statistiken** → **Deskriptive Statistik**. Wir ziehen die beiden zentrierten metrischen Variablen nach rechts in das Fenster (*agea_zen* und *eduyrs_zen*). Dann setzen wir unten links einen Haken bei **Standardisierte Werte als Variable speichern** und drücken auf **OK**. Damit sind wir fertig. In unserer Variablenansicht sehen wir, dass zwei neue Variablen erstellt wurden (*Zagea_zen* und *Zeduyrs_zen*). Diese können wir für die folgende logistische Regressionsanalyse verwenden.

Durchführung der logistischen Regressionsanalyse

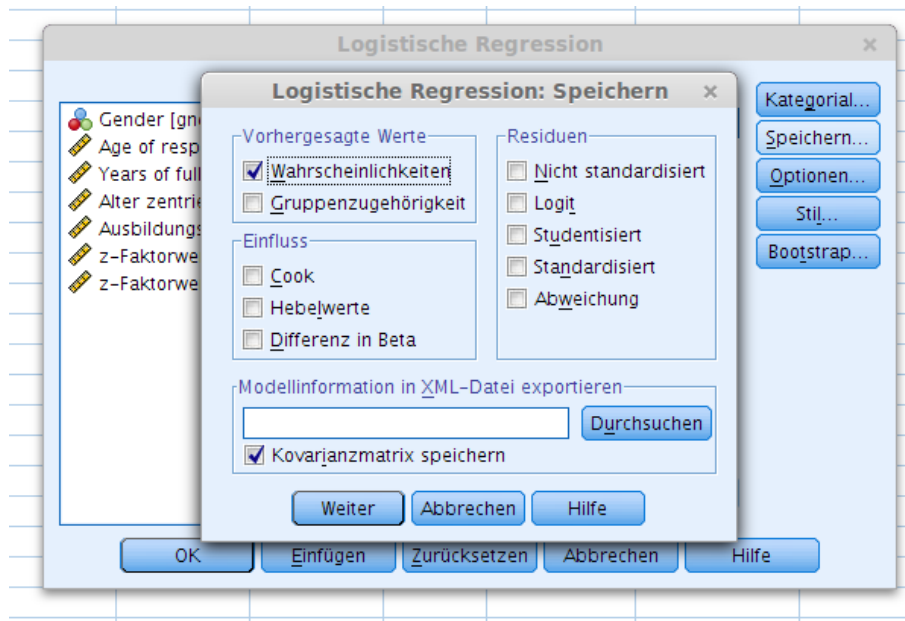
Nachdem wir nun unsere Daten gesichtet und vorbereitet haben, können wir endlich zum Kern der Sache vordringen. Wir öffnen das Dialogfenster, indem wir auf **Analysieren** → **Regression** → **Binär Logistisch** klicken.



In dem sich öffnenden Fenster wählen wir zuerst die abhängige Variable *vote* aus und fügen Sie bei „Abhängige Variable“ ein. Die drei unabhängigen Variablen *gndr*, *Zeduyrs_zen* und *Zagea_zen* fügen wir bei „Kovarianten“ ein. Als Methode wählen wir Einschluss. Das Fenster sollte nun so aussehen:



Nun können wir noch einige Optionen anpassen, die wir rechts über die blauen Schaltflächen erreichen können. Die Option „Kategorial“ ist für uns im Moment nicht von Bedeutung, da alle unabhängigen Variablen entweder metrisch oder dichotom sind. Hätten wir ordinale oder nominale Variablen mit mehr als zwei Ausprägungen, würden wir dort noch Einstellungen vornehmen. Dazu mehr in einem anderen Tutorial. Bei der Option „Speichern“ können wir noch etwas verändern. Dazu öffnen wir den Dialog. Wir setzen dann einen Haken bei „Wahrscheinlichkeiten“. Dies bedeutet, dass SPSS eine neue Variable anlegt, in dem für jede Person die errechnete Wahrscheinlichkeit gespeichert wird, dass diese Person zur Wahl geht. Dies kann später zur Erstellung von Grafiken nützlich sein.



Die anderen Optionen lassen wir unverändert. Unter „Einfluss“ wäre es möglich, durch bestimmte Optionen Ausreißer in den Daten auffindig zu machen, was für komplexere Analysen sinnvoll sein kann. Dadurch wird es möglich, einige sehr wenige extreme Fälle zu entfernen, wenn diese die Ergebnisse der Berechnungen signifikant verzerren. Bei „Residuen“ wären noch einige Informationen erhältlich, die uns Aussagen über die Güte unseres Modells machen könnten. Diese fortgeschrittenen Methoden wollen wir hier aber nicht vertiefen. Wir klicken daher auf **Weiter**. Damit sind wir auch fertig, da wir die drei restlichen Optionen nicht verändern. Dies sind alles fortgeschrittene Methoden und zur Durchführung unserer Analyse hier nicht von Interesse. Wir klicken daher im Dialogfenster auf **OK** und können anschließend mit der Interpretation beginnen.

Die Interpretation der Ergebnisse

Die eigentliche Durchführung der logistischen Regressionsanalyse ist in SPSS ein Kinderspiel, ausführlicher wollen wir die Interpretation der Ergebnisse besprechen. Hier wird deutlich, dass einem SPSS dort wenig Arbeit abnimmt, sondern es dem Nutzer überlassen bleibt, sich aus dem Wust der Zahlen einen Sinn zu konstruieren. Wir werden daher die Ergebnisse Tabelle für Tabelle durchgehen. Zunächst sollten aber die Ergebnisse gespeichert werden. Am besten lässt man sich die Ausgabe als PDF-Datei exportieren, sodass man sie später in eine Publikation einbauen kann.

1. Zusammenfassung der Fallverarbeitung

Zusammenfassung der Fallverarbeitung			
Ungewichtete Fälle ^a		H	Prozent
Ausgewählte Fälle	Einbezogen in Analyse	2579	87,2
	Fehlende Fälle	379	12,8
	Gesamtsumme	2958	100,0
Nicht ausgewählte Fälle		0	,0
Gesamtsumme		2958	100,0

Diese Tabelle gibt eine kurze Übersicht über die Daten und zeigt auf, wie viele der Fälle bzw. Personen auch tatsächlich für die Berechnung des Modells verwendet wurden. In unserem Beispiel wurden 379 Fälle ausgeschlossen, beispielsweise, weil die befragten Personen zu jung waren oder diese schlichtweg manche Antworten verweigert haben, sodass keine Daten vorliegen. Fehlende Fälle sind typisch für jede Untersuchung. Sollte die Zahl allerdings auffallend hoch sein, könnte man einen Fehler gemacht haben. Dies ist dann zu kontrollieren, beispielsweise durch Kreuztabelle und grafische Darstellungen.

2. Codierung abhängiger Variablen

Codierung abhängiger Variablen	
Ursprünglicher Wert	Interner Wert
Yes	0
No	1

Diese Tabelle ist für die Interpretation von erheblicher Bedeutung. Unsere abhängige Variable *vote* kann nur zwei Ausprägungen annehmen: ja (gewählt) und nein (nicht gewählt). Dies kann im Datensatz beliebig durch Zahlen kodiert werden (z.B. Ja = 101 und Nein = 202). Egal wie die Kodierung in den Daten auch sein mag, SPSS rechnet diese Zahlen für die Berechnung des Modells immer in 0 und 1 um. „Ja“ wurde in diesem Fall zu 0, „Nein“ zu 1. Wenn eine Person also einen Wert von 1 hat, dann hat sie mit einer Wahrscheinlichkeit von 100% nicht gewählt. **Achtung:** hat eine Person nun eine berechnete Wahrscheinlichkeit von 85 %, so bedeutet dies, dass sie eine Wahrscheinlichkeit von 85 % hat, **nicht** zur Wahl zu gehen. Die Wahrscheinlichkeit, dass diese Person wählen geht, ist also 15 %.

3. Klassifikationstabelle

Klassifikationstabelle^{a,b}

Beobachtet			Vorhersagewert		
			Voted last national election		Prozentsatz richtig
			Yes	No	
Schritt 0	Voted last national election	Yes	2116	0	100,0
		No	463	0	,0
Gesamtprozentsatz					82,0

a. Die Konstante ist im Modell enthalten.

b. Der Trennwert ist ,500

Diese Tabelle findet sich unter der Überschrift „Block 0“. Dies bedeutet, dass SPSS hier das Nullmodell rechnet, also eine Regression, in die keine unabhängigen Variablen eingehen. Dies mag sinnlos erscheinen, denn dadurch wird keine Aussage über unsere unabhängigen Variablen getroffen. Der Gedanke ist folgender: stellt sich am Ende heraus, dass das Nullmodelle ohne Variablen genau so gut ist, wie das Modell mit den Variablen, dann sind unsere Variablen äußerst schlechte Prädiktoren und sollten verworfen werden. Ob dies der Fall ist, muss weiter unten geklärt werden. Hier scheint ein richtiger Prozentsatz 82,0 % recht beeindruckend zu sein. Allerdings ist dieses Ergebnis irreführend. SPSS nimmt schlichtweg an, dass alle befragten Personen gewählt haben. Dies ist nicht der Fall, 463 Personen haben nicht gewählt. Von 2116+463=2579 Personen hat

SPSS also nur 2116 korrekt zugeordnet, dies entspricht den eben genannten 82 % ($\frac{2116}{2579}$).

Dieses Ergebnis ist auf die sehr ungleiche Verteilung der Variable *vote* zurückzuführen. Hätten beispielsweise nur 60 % aller Befragten gewählt, hätte SPSS deutlich mehr Personen falsch zugeteilt; der Prozentsatz wäre erheblich niedriger. Es bleibt zu hoffen, dass unsere Variablen das Nullmodell noch verbessern können. Der **Trennwert** von 0,500 (siehe bei b unter der Tabelle) gibt an, ab welchem Wert eine Person in eine Kategorie eingeteilt wird. Dies bedeutet, dass eine Person mit einer berechneten Wahlwahrscheinlichkeit von 49 % als Nichtwähler gezählt wird, eine Person mit einer Wahrscheinlichkeit von 51 % als Wähler. In fortgeschrittenen Analysen kann man diesen Wert auch je nach Fragestellung und Modell manuell anpassen.

4. Variablen in der Gleichung

Variablen in der Gleichung

	B	Standardfehler	Wald	df	Sig.	Exp(B)
Schritt 0 Konstante	-1,520	,051	877,160	1	,000	,219

SPSS berechnet uns für das Nullmodell eine „Regressionsgleichung“, was recht sinnlos ist, da ja noch keine Prädiktoren (also unabhängige Variablen) ins Modell eingegangen sind. Es bleibt daher nur die Angabe der Konstanten. Diesen Wert kennen wir aber schon! Der Logit-Wert wird bei SPSS unter **B** angezeigt, der Odd-Wert unter **Exp(B)**. Wenn wir unsere Formel von Seite 3 benutzen,

können wir auch die Wahrscheinlichkeit berechnen: $p = \left(\frac{O}{1+O} \right) = \left(\frac{0,219}{1+0,219} \right) = 0,1797$. Dies ist

exakt der Wert, den wir bereits über die Säulendiagramme erfahren haben (vgl. S. 9). Dies war die Wahrscheinlichkeit, dass eine Person nicht zur Wahl geht.

5. Nicht in der Gleichung vorhandene Variablen

Nicht in der Gleichung vorhandene Variablen			
Schritt 0	Variablen	Score	Sig.
	gndr	,958	1
	Zagea_zen	53,477	1
	Zeduyrs_zen	68,610	1
	Gesamtstatistik	152,042	3

Wie der Titel bereits aussagt, zeigt uns SPSS hier nur kurz an, welche Variablen noch nicht in die Gleichung aufgenommen werden. Diese Tabelle ist sinnvoller, wenn man ein Modell schrittweise aufbaut, also zuerst keine Variable und dann später immer jeweils eine Variable mehr ins Modell aufnimmt. Dies kann hilfreich sein, da dadurch Einflüsse einzelner Variablen stärker aufgezeigt werden. Dann ist diese Tabelle wichtig, damit man weiß, welche Variablen eigentlich noch fehlen. In unserem Beispiel ist das jedoch sehr einfach zu sehen, da schlichtweg noch alle fehlen. Im nächsten Schritt werden wir direkt **alle** Variablen gleichzeitig ins Modell aufnehmen, da wir keine andere Option eingestellt haben. Auffallend ist bereits hier, dass der Wert der Signifikanz bei der Variable Geschlecht (*gndr*) recht hoch ist (wir wollen hier alle Werte als hoch ansehen, die einen Signifikanzwert über 0,050 haben). Dies bedeutet, dass diese Variable vermutlich keinen Einfluss auf das Modell haben wird. Mehr dazu weiter unten.

6. Omnibustest der Modellkoeffizienten

Omnibustests der Modellkoeffizienten				
		Chi-Quadrat	df	Sig.
Schritt 1	Schritt	162,000	3	,000
	Block	162,000	3	,000
	Modell	162,000	3	,000

Man beachte, dass sich diese Tabelle unter der Überschrift „Block 1“ befindet. Hier fängt SPSS also an, alle ausgewählten unabhängigen Variablen mit ins Modell aufzunehmen. Die eigentliche Interpretation beginnt erst jetzt wirklich. Mit diesem ersten Test versucht SPSS festzustellen, ob unser Modell mit allen Variablen besser ist als das Nullmodell. Dazu zieht es den Chi-Quadrat-Test heran. Da alle angezeigten Signifikanzen sehr klein sind, bedeutet dies, dass unser Modell besser ist als das Nullmodell. Unsere Vorhersagekraft wird also signifikant besser, wenn wir unsere unabhängigen Variablen ins Modell aufnehmen. Dies ist ein gutes Ergebnis und bedeutet, dass die Auswahl unserer unabhängigen Variablen nicht schlecht war.

7. Modellübersicht

Modellübersicht

Schritt	-2 Log-Likelihood	R-Quadrat nach Cox & Snell	R-Quadrat nach Nagelkerke
1	2265,744 ^a	,061	,100

Dies ist eine sehr wichtige Tabelle und für fortgeschrittene Modelle von großer Bedeutung. Hier werden nach bestimmten Formeln Werte berechnet, die eine Aussage über die Güte des Modells machen. Grundsätzlich gilt, dass ein Modell besser wird, je kleiner der -2 Log-Likelihood-Wert und je größer die beiden R-Quadrat-Werte werden. Da wir kein anderes Modell zum Vergleich haben, sind die Werte hier insgesamt nicht wirklich interpretierbar. Würde man ein Modell schrittweise aufbauen, also Variablen nach und nach aufnehmen, könnte man sehen, bei welchem Modell die R-Quadrat-Werte am größten wären. Dies wäre demnach das Modell mit der höchsten Güte und sollte favorisiert werden. Zu beachten ist, dass der Wert nach Nagelkerke aufgrund der Berechnungsvorschrift immer größer sein wird als der Wert nach Cox&Snell. Zudem sollte man wissen, dass ein Vergleich über verschiedene Datensätze hinweg nicht möglich ist. Ein wichtiger Stichpunkt zu dem Thema ist **unbeobachtete Heterogenität** (siehe Literaturangaben).

Zur Einschätzung: lässt man bei der Berechnung der Regression zwei Variablen weg und verwendet als unabhängige Variable **nur** die Altersvariable, errechnet SPSS R-Quadratwerte von 0,021 bzw. 0,034. Diese sind kleiner als die hier berechneten. Es ist also sehr gut, dass wir alle drei Variablen herangezogen haben. **Wichtig:** man kann diese sog. Pseudo-R-Quadrat-Kennzahlen nur vergleichen, wenn man die gleichen Daten und die gleiche abhängige Variable betrachtet. Würde ein Kollege eine ähnliche Untersuchung durchführen, dabei aber einen anderen Datensatz verwenden, wären die von ihm berechneten Werte mit meinen Werten niemals vergleichbar. Sie dienen nur dazu um zu prüfen, welche Variablen wir ins Modell aufnehmen sollten und welche nicht.

8. Klassifikationstabelle

Klassifikationstabelle^a

Beobachtet			Vorhersagewert		
			Voted last national election		Prozentsatz richtig
			Yes	No	
Schritt 1	Voted last national election	Yes	2108	8	99,6
		No	453	10	2,2
	Gesamtprozentsatz				82,1

a. Der Trennwert ist ,500

Wieder gibt uns hier SPSS eine Klassifikationstabelle aus. Hierbei berücksichtigt SPSS nun die eingegangenen Variablen und versucht, seine Schätzung zu verbessern. Dies erkennt man daran, dass nun auch Werte in der Spalte „No“ stehen (8 und 10). Allerdings fällt auf, dass trotz der Prädiktoren die Schätzung nur minimal besser wird (hier: 82,1, oben: 82,0). Diese sehr geringe Verbesserung deutet darauf hin, dass der Einfluss unserer Prädiktoren nicht sehr groß ist. Wie gut

diese nun wirklich sind, ist sehr schwer zu beurteilen

9. Variablen in der Gleichung

Variablen in der Gleichung

	B	Standardfehler	Wald	df	Sig.	Exp(B)
Schritt 1 ^a						
gndr	,032	,107	,090	1	,764	1,032
Zagea_zen	-,537	,058	86,256	1	,000	,584
Zeduyrs_zen	-,620	,064	94,308	1	,000	,538
Konstante	-1,611	,170	90,039	1	,000	,200

a. In Schritt 1 eingegebene Variable(n): gndr, Zagea_zen, Zeduyrs_zen.

Wie so oft im Leben kommt auch in SPSS das Beste zum Schluss. Die vielen Tabellen geben uns wichtige Einblicke, aber das, was für eine Interpretation wirklich zählt, sind die Regressionskoeffizienten. Damit wird es uns dann auch möglich, die Regressionsgleichung aufzustellen. Zunächst die Signifikanzen. Hierbei fällt auf, dass die Geschlechtervariable *gndr* einen sehr hohen Wert aufweist (wir betrachten hier alle Werte über 0,050 als hoch). Dies bedeutet, dass der Einfluss des Geschlechts auf das Wahlverhalten mit einer hohen Wahrscheinlichkeit zufällig ist. Dies bedeutet, dass die Berücksichtigung des Geschlechts einer Person uns keinen Hinweis darauf gibt, ob sie wählen geht oder nicht. Wir werden daher diese Variable nicht in der Regressionsgleichung berücksichtigen. Alle anderen Variablen weisen sehr geringe Werte bei der Signifikanz auf, das ist ein gutes Ergebnis.

Wir erinnern uns, dass die Logit-Werte in SPSS in der Spalte **B** zu finden sind, die Odd-Werte in der Spalte **Exp(B)**. Welchen Wert wir letztlich benutzen ist irrelevant, sofern wir die korrekte Umrechnung in Wahrscheinlichkeiten vornehmen. Eine direkte Interpretation ist auch mit den Odd-Werten möglich, allerdings schwierig und eine Fehlerquelle, die selbst Fachleute ins Schleudern bringen kann. Die einschlägige Fachliteratur rät daher von der Interpretation der Odd oder Logit-Werte ab¹. Zwei Dinge lassen sich dennoch aus den Odd-Werten ablesen:

- **Richtung** des Zusammenhangs. Sowohl die Alters- als auch die Bildungsvariable weisen Werte von kleiner 1 auf. Wenn wir uns an die Umrechnungstabelle auf S. 3 erinnern, stellen wir fest, dass der Odd-Wert 1 eine wichtige Trennschwelle darstellt. Werte über 1 deuten auf einen positiven Zusammenhang hin, Werte unter 1 auf einen negativen. Dabei ist allerdings noch die **Kodierung** der abhängigen Variable zu beachten (vgl. Tabelle 2 weiter oben)! Aufgrund aller Informationen können wir daher nun folgende Schlüsse ziehen: **je älter und je gebildeter Personen sind, desto kleiner ist die Wahrscheinlichkeit, dass diese nicht zur Wahl gehen.** Anders (und zugespitzt) ausgedrückt: alte und gebildete Personen wählen fast immer.

1 Vgl. Wolf, Christof; Best, Henning: Logistische Regression, in: Handbuch der sozialwissenschaftlichen Datenanalyse, Wiesbaden 2010: S. 845.

- **Stärke** des Zusammenhangs: grob lässt sich noch aussagen: je weiter ein Odd-Werte von 1,0 entfernt ist, desto stärker ist sein Einfluss. Beispielsweise hätte eine Variable mit einem Koeffizienten von 0,3 einen stärkeren Einfluss als eine mit einem Koeffizienten von 0,9. Umgekehrt: eine Variable mit einem Odd-Wert von 7 hat einen stärkeren Einfluss als eine mit einem Wert von 2. Eine Variable mit einem Koeffizienten von 0,5 hat genau den gleichen Einfluss wie eine mit einem Wert von 2,0. Die Richtung des Einflusses ist jedoch entgegengesetzt! Warum ist das so? $\rightarrow |\ln(0,5)| = |\ln(2)| = 0,693$.

Den **Stellenwert** der einzelnen unabhängigen Variablen können wir jedoch nur dann angeben, wenn diese in der gleichen Skala (Einheit) messen. Wären die Einheiten beispielsweise Lebensalter und Einkommen, wäre ein Vergleich unmöglich. Da wir aber beide Variablen vor der Regression standardisiert haben, messen sie in der gleichen Einheit (Standardabweichungen). **Nur deshalb** können wir feststellen, welche Variable den größeren Einfluss hat: es ist die Bildungsvariable, wenn auch nur mit sehr knappem Vorsprung. Dies ist so, weil 0,538 weiter von 1 entfernt ist als 0,584.

Die **Regressionsgleichung** können wir nun auch angeben:

$$p(\text{Nichtwähler}) = \frac{e^{\text{const} + b_1 \cdot x_1 + b_2 \cdot x_2}}{1 + e^{\text{const} + b_1 \cdot x_1 + b_2 \cdot x_2}} = \frac{e^{-1,611 - 0,620 \cdot x_1 - 0,537 \cdot x_2}}{1 + e^{-1,611 - 0,620 \cdot x_1 - 0,537 \cdot x_2}}$$

Diese Darstellung ist einerseits ideal, weil sie alle Informationen in knapper Form zusammenfasst und wir nun beliebige Konstellationen aus Alter und Bildungsstand betrachten könnten. Sie ist aber gleichzeitig unmöglich, weil kein Leser Lust hat, diese Operationen durchzuführen. Gewonnene Informationen sollten **einfach verständlich**, wahrheitsgemäß und möglichst unverzerrt dargestellt werden. Dazu sind andere Formen der Darstellungen besser, die wir im nächsten Kapitel kennenlernen werden.

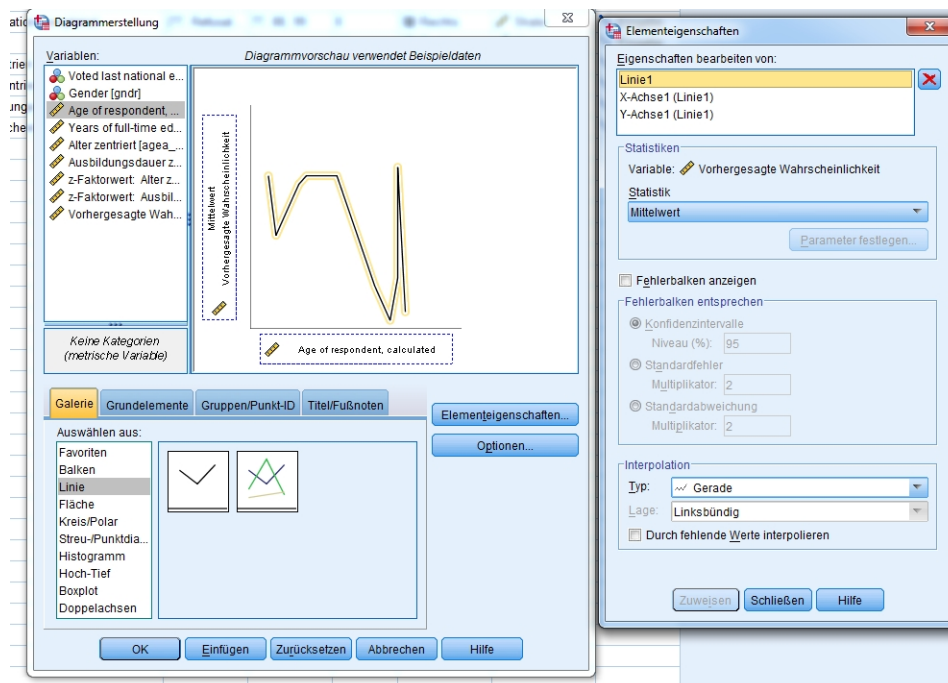
Ein Wort zur **Nichtlinearität**: im Gegensatz zur normalen Regression (OLS-Regression) sind Logit-Regressionen **grundsätzlich** nichtlinear modelliert. Während also in einer Grafik einer Einfachregression eine Gerade den Zusammenhang veranschaulichen kann, ist dies bei der Logit-Regression nicht möglich. Dies ist von Vorteil, weil viele reale Zusammenhänge so besser veranschaulicht werden können (Glück und Einkommen beispielsweise hängen so zusammen. Ab einem bestimmten Wert tritt eine Sättigung ein, die Kurve steigt dann nicht mehr. Menschen, die 50.000€ pro Monat verdienen sind nicht wirklich glücklicher als solche, nur „nur“ 45.000€ verdienen). Das macht jedoch die Interpretation schwieriger. Mathematisch gesehen ist für dieses multiplikative Verhalten der Logitregression die Exponentialfunktion e verantwortlich.

Ergebnisdarstellung

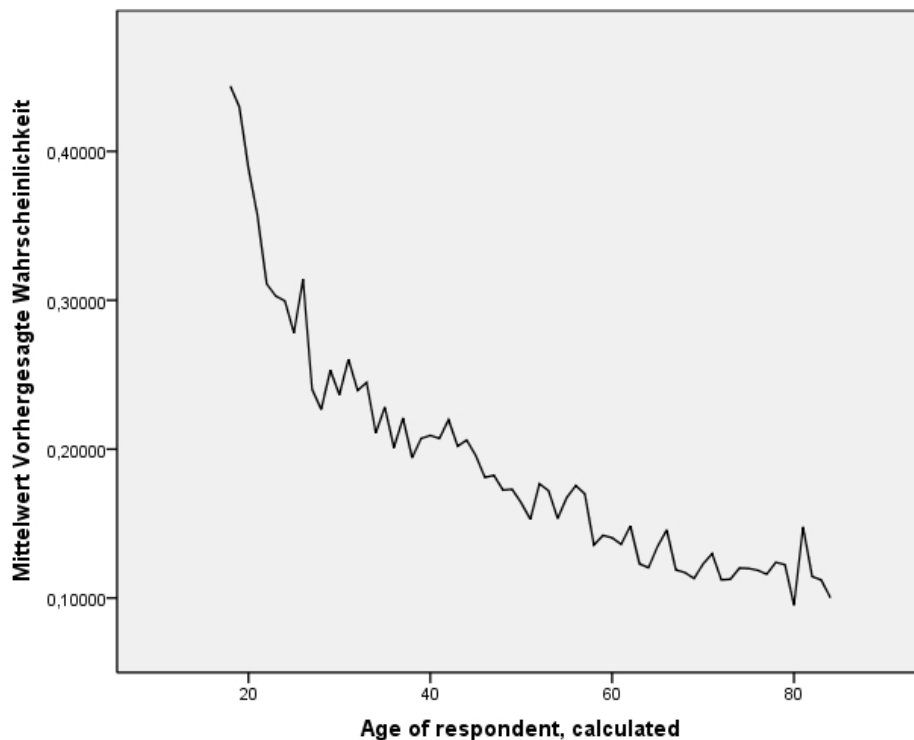
Nachdem wir nun in mühevoller Arbeit unsere Zusammenhänge und Daten berechnet haben, wollen wir diese nun allgemein verständlich und publikationsfähig darstellen. Dies ist eine anspruchsvolle Aufgabe und von entscheidender Bedeutung. Wissenschaft bedeutet nicht, möglichst viele und komplizierte Ergebnisse zu produzieren. Ebenso wichtig ist es, Ergebnisse auch so darstellbar zu machen, dass sie auch für Nichtwissenschaftler erfassbar sind. Einstein sagte dazu: *man soll alles so einfach machen wie möglich, aber nicht einfacher*. Dieser Herausforderung wollen wir uns nun stellen.

Conditional Effect Plots

Die Möglichkeit der Darstellung über solche Plots ist wichtig, allerdings in SPSS nicht wirklich gut umsetzbar. Andere Programme wie R oder STATA leisten hier mehr als SPSS, was uns aber nicht stören soll, da wir unser noch sehr einfaches Modell doch einigermaßen modellieren können. Zunächst sollte uns auffallen, dass SPSS bei der Berechnung der Regression eine neue Variable erzeugt hat (*PRE_1*). Darin hat SPSS für jede einzelne Person im Datensatz die **errechnete** Wahrscheinlichkeit gespeichert, dass diese Person ein Nichtwähler ist. Das mag sinnlos erscheinen, da wir ja bereits durch die Variable *vote* wissen, ob die Person tatsächlich gewählt hat oder nicht. Allerdings erleichtert uns dieser Trick die Auswertung. Fangen wir mit einem Liniendiagramm an, in das wir nur eine der beiden abhängigen Variablen aufnehmen. Ich habe dazu das Alter der Personen gewählt. Dazu klickt man auf **Diagramme** → **Diagrammerstellung**.

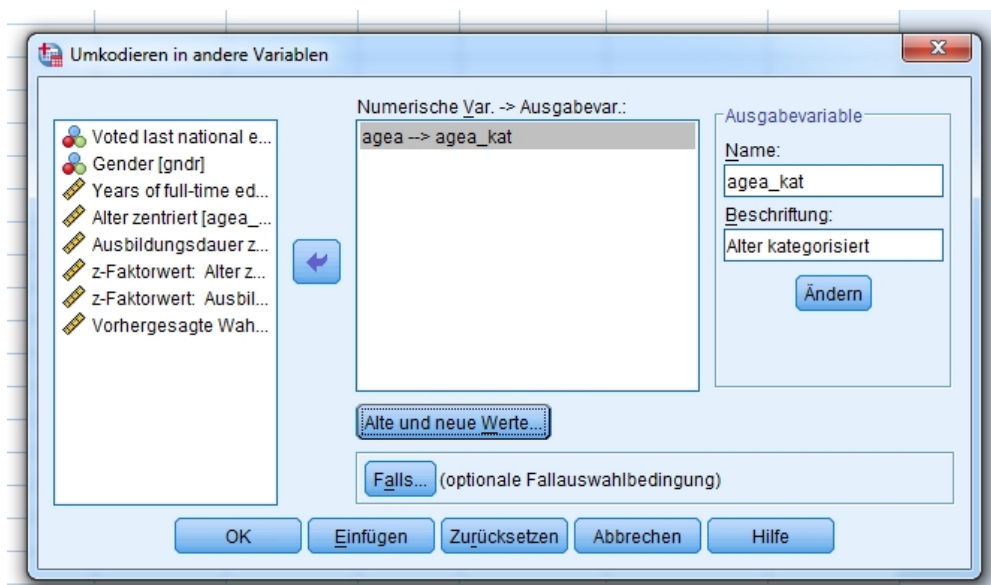


Folgende Einstellungen sollte man vornehmen. Wir wählen als Art Linie und setzen die Altersvariable (*agea*) auf die x-Achse. Als Variable für die y-Achse wählen wir die Variable mit den Wahrscheinlichkeiten (*PRE_1*). Den Rest lassen wir auf Standard. Wir erhalten folgende Ansicht:

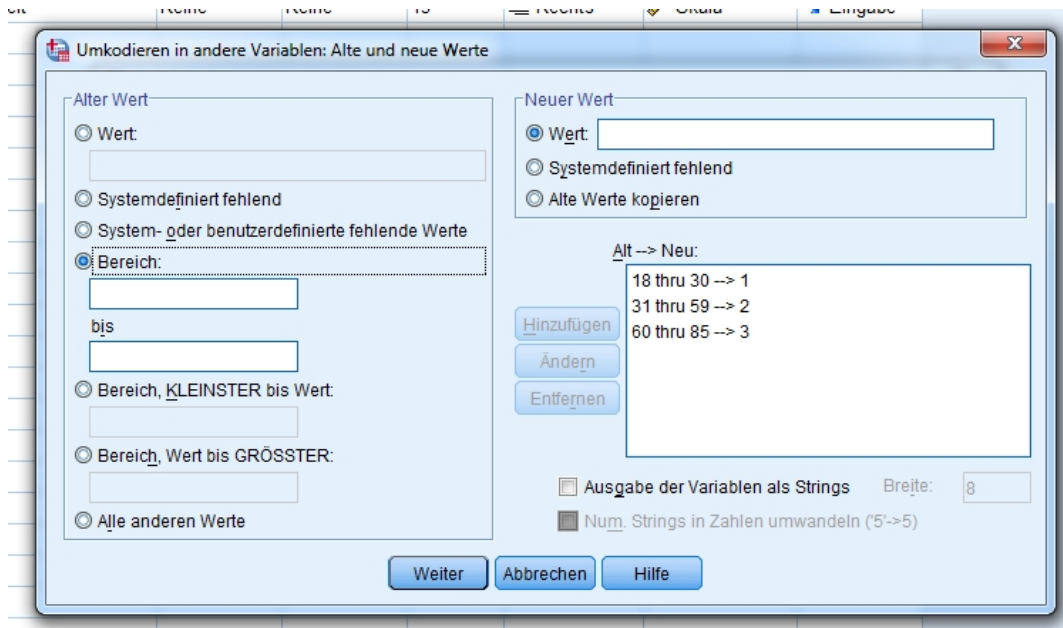


Wir haben in diesem Fall als Variable die nichtstandardisierte Altersvariable ausgewählt. Grundsätzlich ist es in diesem Fall egal, welche man wählt, da nichts wirklich berechnet wird. So machen wir uns aber die Interpretation einfacher. Wie man deutlich erkennen kann, nimmt die Wahrscheinlichkeit, Nichtwähler zu sein, mit steigendem Alter stark ab. Allerdings stören uns die vielen Zacken, zudem wird die Bildungsvariable überhaupt nicht einbezogen. Um dies zu erreichen, bedienen wir uns eines Tricks. Wir verwandeln dazu die Altersvariable in eine **kategoriale Variable** mit drei Kategorien. Wir unterscheiden dann junge, mittelalte und alte Personen und prüfen, ob sich Unterschiede ergeben.

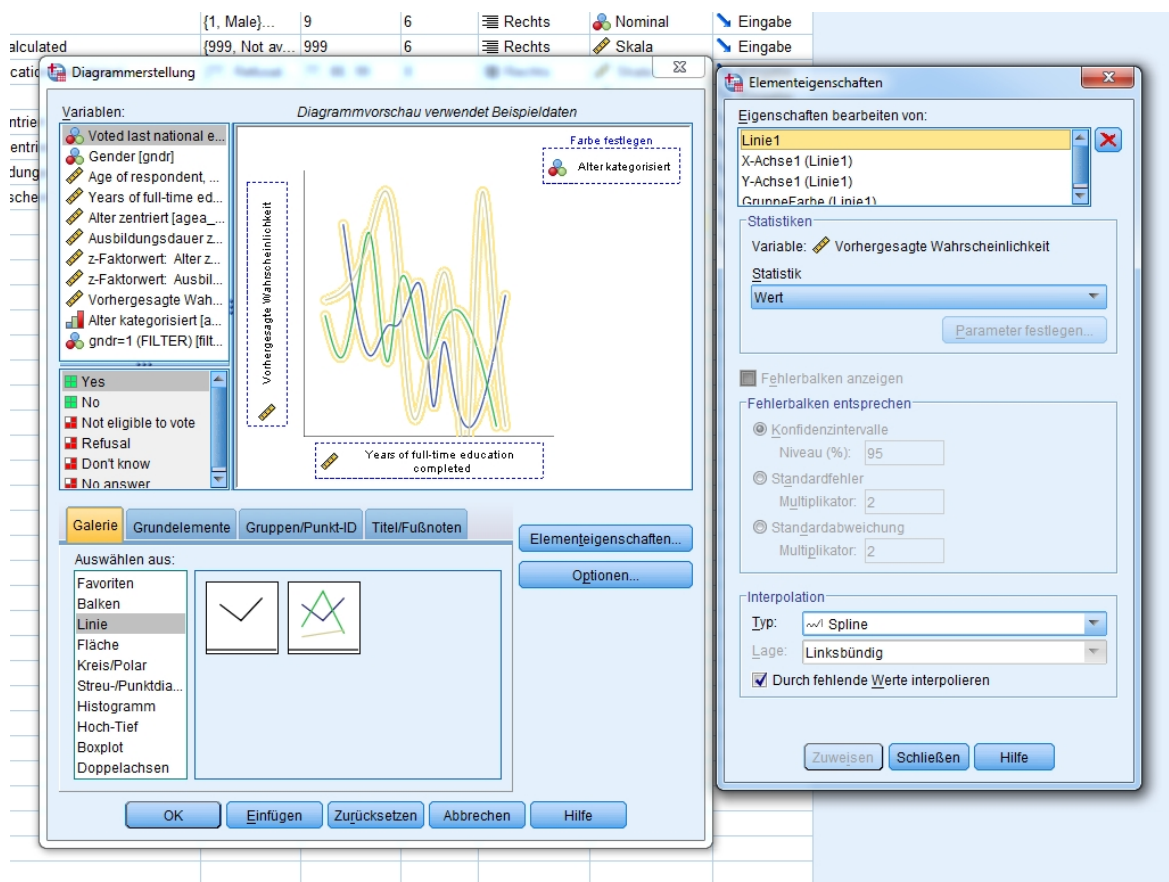
Dazu müssen wir eine neue Variable erstellen. Wir klicken auf **Daten** → **Umkodieren in andere Variablen**. Die neue Variable nennen wir *agea_kat*, als Ausgangsvariable nehmen wir *agea*.



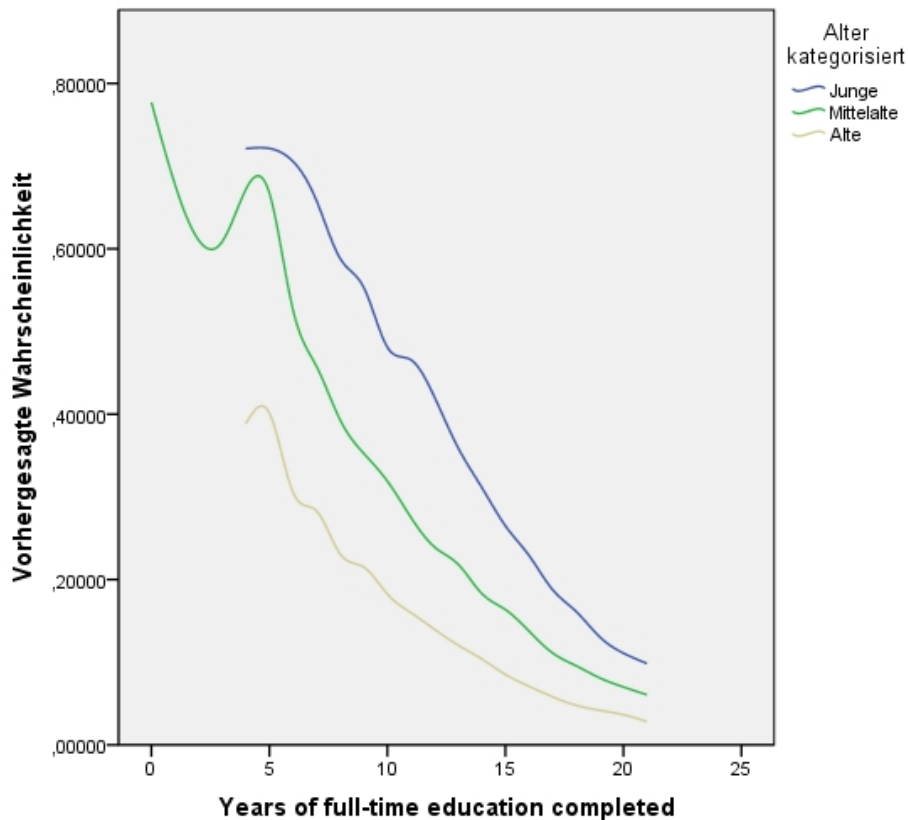
Dann klicken wir auf **Alte und neue Werte** und definieren dort die Grenzen für die Kategorien. „Jung“ sollen Personen zwischen 18 und 30 Jahren sein, „mittelalt“ zwischen 31 und 59 und „alt“ alle über 60.



Wir nummerieren die Kategorien hier einfach durch und benennen sie dann in der SPSS Variablenansicht bei „Wertelabels“ wie gewünscht. Wir klicken auf **Weiter** und dann auf **OK**. Jetzt können wir wieder das Diagrammmenü öffnen. Wir wählen wieder Linie, jedoch diesmal das rechte Bild mit den mehreren farbigen Linien.



Oben rechts bei „Farbe festlegen“ ziehen wir unsere neu gebildete Altersvariable *agea_kat* hinein. Auf die x-Achse ziehen wir diesmal unsere Bildungsvariable *eduysr*, auf die y-Achse wieder *PRE_1*. Um die Darstellung ansprechender zu gestalten, wählen wir im rechten Fenster („Elementeigenschaften“) bei Interpolation als Typ **Spline** und setzen den Haken bei „Durch fehlende Werte interpolieren“. Dann klicken wir auf **Zuweisen** und dann auf **OK**. Wir erhalten folgendes Diagramm:



Die Legende rechts oben erklärt, welche Farbe zu welcher Altersgruppe gehört. Hier haben wir nun alle Variablen in einer Grafik vereint. Die Interpretation gestaltet sich deutlich einfacher und intuitiv. Sofort wird klar, dass bei allen Altersgruppen mit steigender Bildung die Nichtwählwahrscheinlichkeit rapide sinkt. Auch wird deutlich, dass junge Personen generell deutlich weniger wahrscheinlich zur Wahl gehen als alte Personen. Junge Personen mit wenig Bildung gehen selten zur Wahl (Nichtwählwahrscheinlichkeit bei ca. 70%), gebildete Alte hingegen wählen fast immer (Nichtwählwahrscheinlichkeit bei unter 5%). All diese Infos können direkt aus dem Diagramm abgelesen werden. Deutlich wird aber auch hier, wo die Grenzen dieser Darstellung liegen. Hätten wir mehr als zwei signifikante Variablen in der Gleichung, müssten wir eine andere Form der Darstellung wählen oder mehrere Diagramme erstellen (beispielsweise jeweils eins für jedes Geschlecht).

Säulendiagramme

Indem man bestimmte Variablenkonstellationen ausrechnet und diese dann als Säulendiagramme darstellt, lassen sich ebenfalls aussagekräftige Grafiken erstellen. Dies kann man manuell oder mit einer Tabellenkalkulationssoftware wie Excel oder LibreOffice Calc erreichen. Betrachten wir dazu

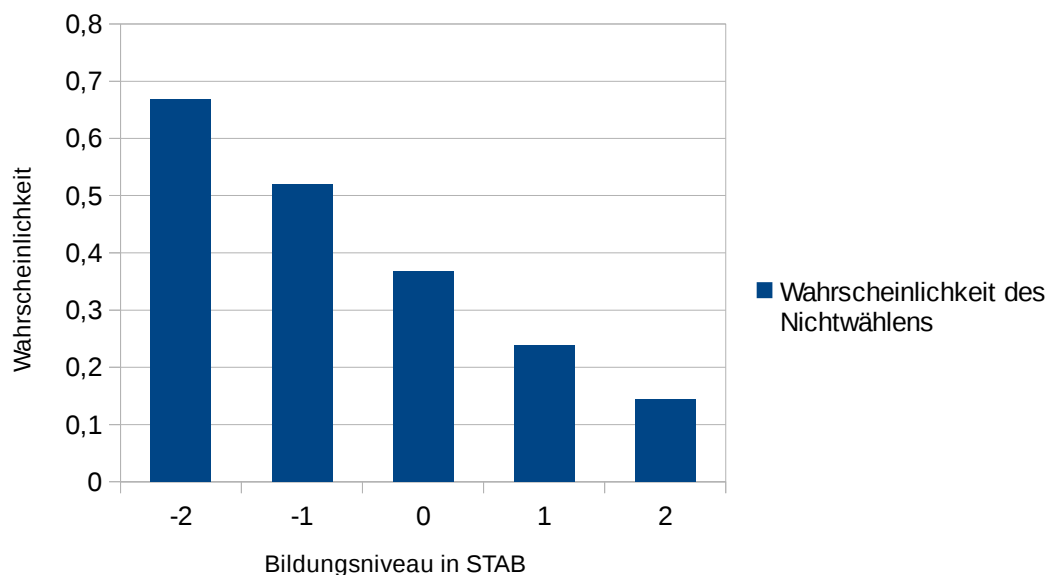
nochmal die Regressionsgleichung:

$$p(\text{Nichtwähler}) = \frac{e^{-1,611 - 0,620 \cdot x_1 - 0,537 \cdot x_2}}{1 + e^{-1,611 - 0,620 \cdot x_1 - 0,537 \cdot x_2}}$$

x_1 ist dabei die Bildungsvariable, x_2 die Altersvariable. Hier wollen wir nun beispielsweise das Alter fixen (also konstant halten) und zwar auf eher junge Menschen. Deshalb setzen wir für x_2 nun -2 fest und lassen die andere Variable von -2 bis +2 durchlaufen, immer in ganzzahligen Schritten. Wir erhalten somit fünf verschiedene Werte:

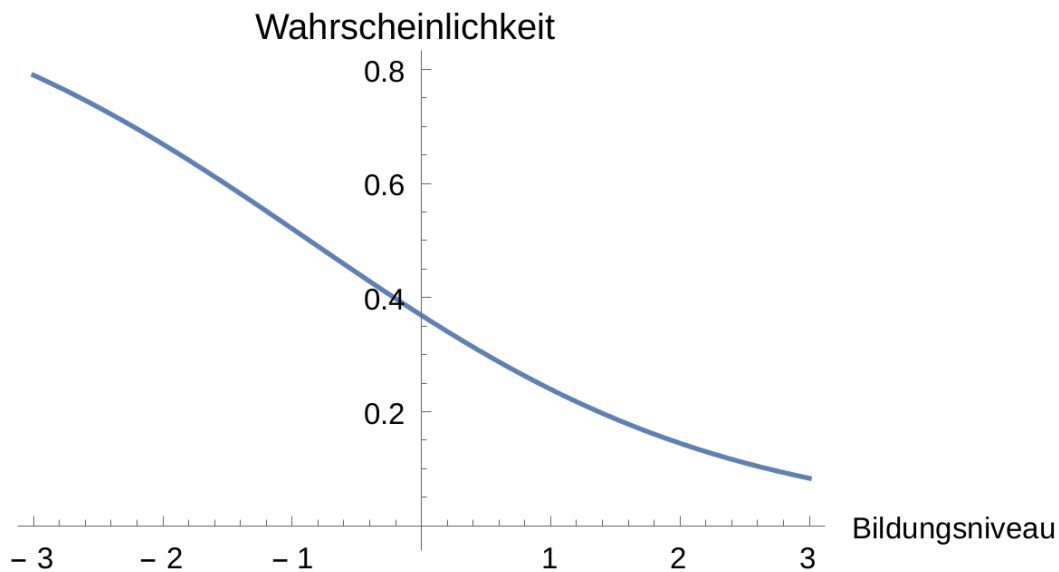
Wahrscheinlichkeit des Nichtwählens	Wert für x_1 (Bildung)	Wert für x_2 (Alter) = gefixte Variable
0,66885	-2	-2
0,52073	-1	-2
0,36888	0	-2
0,23921	1	-2
0,14467	2	-2

Es dürfte uns nicht überraschen, dass die Wahrscheinlichkeiten mit steigender Bildung wieder sinken. Hierbei betrachten wir jetzt allerdings nur jüngere Personen, die zwei Standardabweichungen vom Mittelwert (ca. 50 Jahre) abweichen. Daraus können wir nun ein Säulendiagramm erstellen:



Möchte man noch detaillierter Grafiken, muss man andere Programme zu Hilfe ziehen. Beispielsweise kann man die erhaltene Regressionsgleichung in einen Funktionsplotter nehmen und

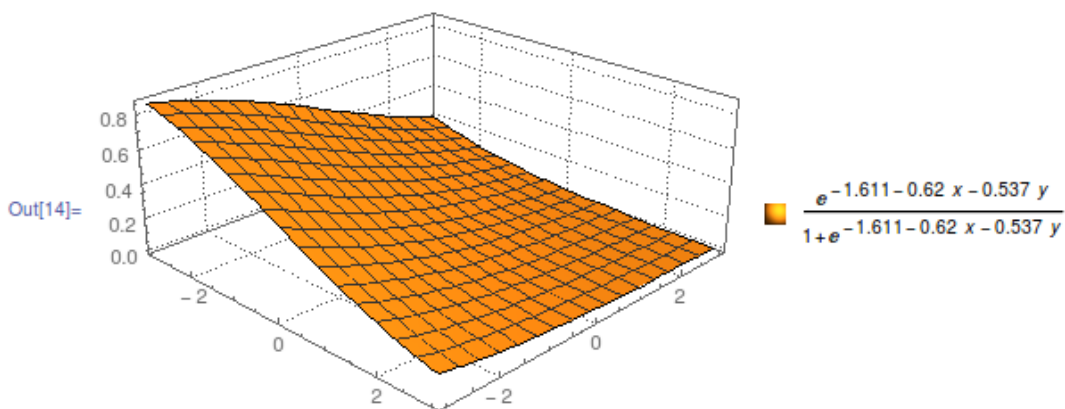
dort als Funktion darstellen lassen. Ich benutze dazu für das Beispiel **Wolfram Mathematica**². Damit lassen sich dann die gleichen Daten folgendermaßen darstellen:



Mathematica berechnet dabei nicht nur fünf Werte, sondern erstellt eine kontinuierliche Funktion. Deutlich wird hierbei auch, dass der Verlauf keine perfekte Gerade ist. Wie wir bereits weiter oben erfahren haben, ist dies in einer Logit-Regression auch gewünscht.

In Mathematica können wir sogar **3D-Plots** realisieren. Hilfreich sind diese aber meist nur bedingt und ohne Wert, wenn wir mehr als zwei Variablen einbeziehen wollen. Trotzdem kann man es auf jeden Fall einmal ausprobieren:

```
In[14]:= Plot3D[ $\frac{e^{-1.611-0.62 x-0.537 y}}{1+e^{-1.611-0.62 x-0.537 y}}$ , {x, -3, 3}, {y, -3, 3}, PlotTheme -> "Detailed"]
```



² Da Mathematica eher wenig verbreitet ist, lohnt es sich auf jeden Fall, einen Blick auf die kostenlose Online-Version zu werfen, die auch im Browser funktioniert: <http://www.wolframalpha.com/>

Literatur und Quellen

- Behnke, Joachim: Logistische Regressionsanalyse. Eine Einführung, Wiesbaden 2014.
- Mood, Carina: Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It, in: European Sociological Review 26 (1): S. 67-82.
- Urban, Dieter; Mayerl, Jochen: Regressionsanalyse: Theorie, Technik und Anwendung, Wiesbaden 2008.
- Wolf, Christof; Best, Henning (Hg.): Handbuch der Sozialwissenschaftlichen Datenanalyse, Wiesbaden 2010.

Bildquelle Titelseite 2: Copyright: CC BY-SA 3.0 (Fire exit.svg, User: Steinbach).
https://commons.wikimedia.org/wiki/File:Fire_exit.svg?uselang=de