Einführung in die Logistische Regression mit Stata

Felix Bittmann

v.1.0

www.felix-bittmann.de

2018

Der Artikel kann folgendermaßen zitiert werden:

Bittmann, Felix (2018): Einführung in die Logistische Regression mit Stata. Abrufbar unter: http://felix-bittmann.de/downloads/artikel/einfuehrung_logit_regression_mit_Stata.pdf. Abgerufen am: [DATUM].

Inhaltsverzeichnis

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. 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 Aussagehalt 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

Achtung: dieses Kapitel ist etwas technisch und zum Anwenden nicht unbedingt notwendig. Für ein besseres Verständnis wird eine Lektüre allerdings empfohlen!

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 Ergebnisses sehr wichtig, die verschiedenen Ebenen zu verstehen, die bei der Berechnung zur Anwendung kommen. Ohne dieses Wissen wird es uns nicht möglich sein, die Stata-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.

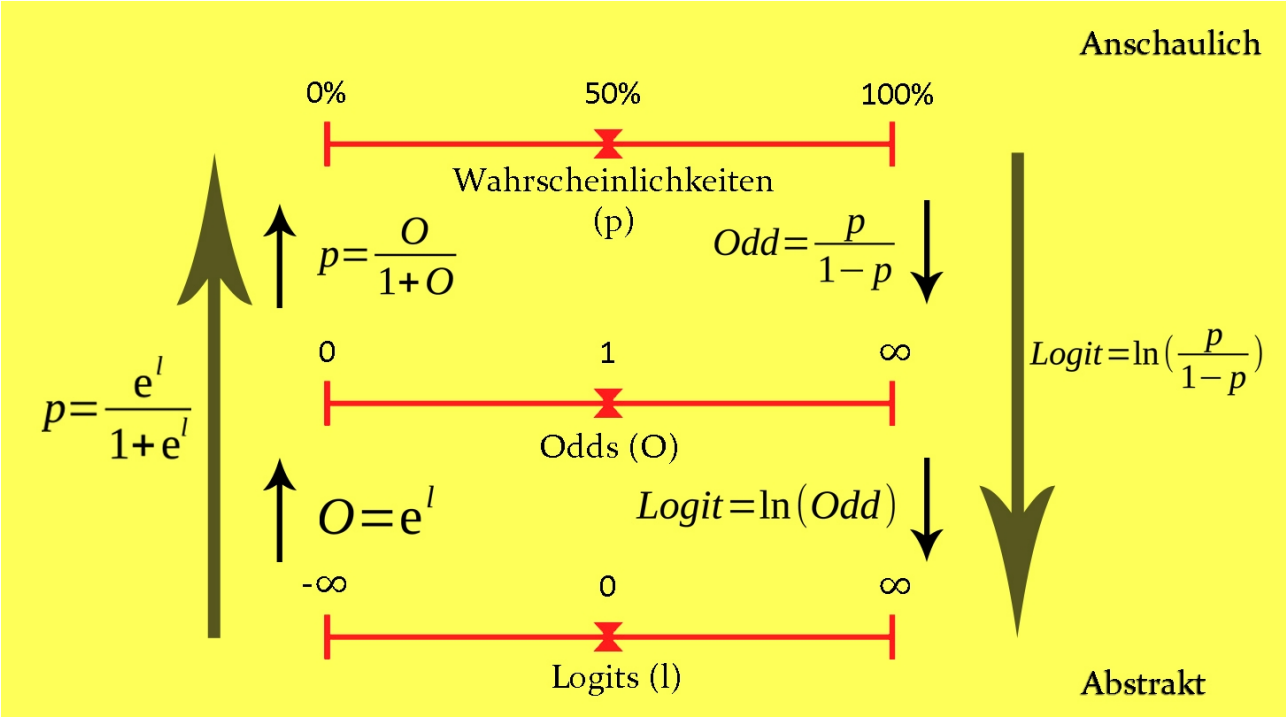
Ebene 1: Wahrscheinlichkeiten. Diese Ebene ist uns im Alltag sehr vertraut, sodass wir später versuchen werden, alle Ergebnisse irgendwie wieder in dieser Ebene auszudrücken. Folgende Aussagen sind beispielsweise leicht für jedermann verständlich:

- 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-positv 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	$\ln(o)=\ln(\frac{p}{1-p})$	l

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

Wahrscheinlichkeiten	←	**Odds**	←	**Logits**
p	$p=(\frac{o}{1+o})=\frac{e^l}{1+e^l}$	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. Stata übernimmt glücklicherweise hier den Großteil der Arbeit für uns.

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 abschneidet 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...).

Fragestellung und Datensatz

Die Fragestellung des Beispiels ist es, wie sich Alter und Geschlecht auf die Wahrscheinlichkeit auswirken, einen Herzanfall zu erleiden. Dazu werden wir Daten des National Health and Nutrition Examination Survey (NHANES-II) verwenden, die kostenlos zum Download bereitstehen. Wir können die Daten entweder direkt in Stata laden oder sie manuell herunterladen[1] und dann in Stata öffnen. Da erstere Option deutlich bequemer ist, werden wir dies so handhaben.

```
webuse nhanes2, clear
```

Die Option *clear* bedeutet, dass alle möglicherweise noch geöffneten Daten im Speicher gelöscht werden. Anschließen wollen wir uns zunächst die Variablen im Datensatz ansehen. Dazu tippen wir

```
describe
```

```
Contains data from http://www.stata-press.com/data/r8/nhanes2.dta
  obs:        10,351
 vars:            57                          3 Sep 2002 12:25
 size:     1,066,153
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
sampl           long    %9.0g                 unique case identifier
strata          byte    %9.0g                 stratum identifier, 1-32
psu             byte    %9.0g                 primary sampling unit, 1 or 2
region          byte    %9.0g      region     1=NE, 2=MW, 3=S, 4=W
smsa            byte    %9.0g                 1=SMSAcity,2=SMSA~city,4=~SMSA
location        byte    %9.0g                 stand number, 1-64
houssiz         byte    %9.0g                 # persons in household, 1-14
sex             byte    %9.0g      sex        1=male, 2=female
race            byte    %9.0g      race       1=white, 2=black, 3=other
age             byte    %9.0g                 age in years
height          float   %9.0g                 height (cm)
weight          float   %9.0g                 weight (kg)
bpsystol        int     %9.0g                 systolic blood pressure
bpdiast         int     %9.0g                 diastolic blood pressure
tcresult        int     %9.0g                 serum cholesterol (mg/dL)
tgresult        int     %9.0g                 serum triglycerides (mg/dL)
hdresult        int     %9.0g                 high density lipids (mg/dL)
hgb             float   %9.0g                 hemoglobin (g/dL)
hct             float   %9.0g                 hematocrit (%)
tibc            int     %9.0g                 total iron bind. cap. (mcg/dL)
iron            int     %9.0g                 serum iron (mcg/dL)
hlthstat        byte    %9.0g                 1=excellent,..., 5=poor
heartatk        byte    %9.0g                 heart attack, 1=yes, 0=no
diabetes        byte    %9.0g                 diabetes, 1=yes, 0=no
sizplace        byte    %9.0g                 1=urban,..., 8=rural
finalwgt        long    %9.0g                 sampling weight (except lead)
leadwt          long    %9.0g                 sampling weight for lead
corpuscl        float   %9.0g                 mean corpuscular volume (fL)
```

Wir sehen eine Liste mit allen Variablen. Von zentraler Bedeutung für uns sind die Variablen heartatk, sex und age. Zunächst verschaffen wir uns einen Überblick über diese Variablen. Dazu nutzen wir den Befehl *tabulate*.

1 http://www.stata-press.com/data/r8/nhanes2.dta (2018-08-20)

```
. tabulate sex
```

1=male, 2=female	Freq.	Percent	Cum.
Male	4,915	47.48	47.48
Female	5,436	52.52	100.00
Total	10,351	100.00	

```
. tabulate heartatk
```

heart attack, 1=yes, 0=no	Freq.	Percent	Cum.
0	9,873	95.40	95.40
1	476	4.60	100.00
Total	10,349	100.00	

Der Output zum Alter ist aus Platzgründen nicht dargestellt. Wir sehen, dass 4,6 % aller Personen im Datensatz eine Herzattacke erlitten haben. 52,2 % aller Personen sind weiblich. Da unsere abhängige Variable (die, die erklärt werden soll) in diesem Fall binär ist (entweder man hatte eine Attacke oder nicht), benutzen wir eine logistische Regression, um diesen Sachverhalt zu erklären.

Das Modell einzugeben ist einfach. Zunächst nennt man das gewünschte Modell (logit), dann die abhängige Variable (heartatk). Danach folgen direkt alle erklärenden Variablen in beliebiger Reihenfolge. Es ist wichtig, Statas Faktor-Variablen-Notation zu benutzen, sofern man mit Variablen arbeitet, die nicht metrisch skaliert sind (also beispielsweise binär, nominal oder ordinal skaliert sind). Alle diese Variablen erhalten das Präfix *i.*, alle metrisch skalierten Variablen das Präfix *c.*

```
logit heartatk i.sex c.age
```

```
Iteration 0:   log likelihood = -1930.5936
Iteration 1:   log likelihood = -1699.6524
Iteration 2:   log likelihood = -1628.8241
Iteration 3:   log likelihood = -1626.8074
Iteration 4:   log likelihood = -1626.8003
Iteration 5:   log likelihood = -1626.8003

Logistic regression                             Number of obs     =     10,349
                                                LR chi2(2)        =     607.59
                                                Prob > chi2       =     0.0000
Log likelihood = -1626.8003                     Pseudo R2         =     0.1574
```

heartatk	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex						
Female	-.9106814	.1019417	-8.93	0.000	-1.110483	-.7108794
age	.0855818	.0048761	17.55	0.000	.0760248	.0951388
_cons	-7.499824	.3119363	-24.04	0.000	-8.111208	-6.88844

Das Modell sollte relativ schnell berechnet werden. Zunächst sehen wir oben links den Iterationsblock, den wir nicht weiter interpretieren brauchen. Rechts zeigt Stata die Anzahl der Fälle an, die im Modell verwendet wurden (10349). Da Stata eine *listwise deletion* durchführt, werden alle Fälle nicht ins Modell aufgenommen, die auch nur einen fehlenden Wert bei irgendeiner Variable im Modell aufweisen. Fehlt beispielsweise bei einer Person die Altersangabe, wird sie nicht ins Modell aufgenommen. **LR chi2(2)** und **Prob > chi2** gehören zusammen. Diese Werte zeigen an, ob das Modell *insgesamt* Varianz der abhängigen Variable erklären kann. Da chi2(2) mit

607 relativ hoch ausfällt und der zweite Wert hochsignifikant ist (der Wert ist kleiner als 0,05) wissen wir, dass dies der Fall ist. Würde dies nicht der Fall sein kämen wir zu der Schlussfolgerung, dass alle erklärenden Variablen in Modell zusammengenommen uns nicht helfen, die abhängige Variable zu erklären. In diesem Fall wären unsere theoretischen Modelle entweder sehr falsch oder wir hätten massive Kodierungsfehler in unseren Daten. **Pseudo R2** gibt an, wie viel Varianz unsere erklärenden Variablen insgesamt erklären können, also etwa 15,7 %. Je höher dieser Wert ist, desto besser können wir Vorhersagen machen. Falls wir aber vor allem kausale Hypothesen testen wollen ist es nicht zwingend notwendig, dass dieser Wert sehr hoch ausfällt. Log likelihood ist ebenfalls eine Statistik, die eine Bewertung des Modells im Vergleich zu anderen Modellen erlaubt. Absolut betrachtet ist diese Zahl für uns sinnlos.

Es folgt der Koeffizientenblock. Wir sehen für jede erklärende Variable im Modell den Koeffizienten (logit-Wert), den Standardfehler des Koeffizienten, die z-Statistik, den p-Wert und ein 95 % Konfidentintervall für den Koeffizienten. Da wir standardmäßig logits betrachten, sollten wir für diese **ausschließlich das Vorzeichen** interpretieren (Mood 2010). Dies mag langweilig klingen, aber wir werden später sehen, wie wir bessere Veranschaulichungen produzieren können. Am Ende sehen wir noch die Konstante, die für uns momentan irrelevant ist.

Der Koeffizient für female ist negativ (-0,91). Dies bedeutet, dass Frauen im Vergleich zu Männern eine geringere Wahrscheinlichkeit haben, einen Herzanfall zu erleiden (unter Kontrolle des Alters). Warum werden hier Männer und Frauen vergleichen? Dazu müssen wir einen Blick auf die Kodierung von sex werden, was wir oben getan haben. Dort sehen wir, dass Männer den Wert 1 haben, Frauen den Wert 2. Da Stata automatisch stets den niedrigsten numerischen Wert als Referenzkategorie wählt, wissen wir, dass Frauen mit Männern verglichen werden. Wenn wir wollen können wir die Referenzkategorie auch stets explizit einblenden. Dazu tippen wir

```
set showbaselevels on
```

Da der p-Wert für female hochsignifikant ist (kleiner als 0.05), wissen wir, dass dieser Effekt statistisch signifikant ist.

Der Koeffizient für age ist positiv. Daher wissen wir, dass mit steigendem Alter die Wahrscheinlichkeit ansteigt, einen Herzanfall zu erleiden. Wenn wir wollen, können wir statt logits auch odds-ratios anzeigen lassen. Dazu lassen wir das Model erneut laufen und ergänzen die Option *or*.

```
logit heartatk i.sex c.age, or
```

```
Iteration 0:   log likelihood = -1930.5936
Iteration 1:   log likelihood = -1699.6524
Iteration 2:   log likelihood = -1628.8241
Iteration 3:   log likelihood = -1626.8074
Iteration 4:   log likelihood = -1626.8003
Iteration 5:   log likelihood = -1626.8003

Logistic regression                             Number of obs     =     10,349
                                                LR chi2(2)        =     607.59
                                                Prob > chi2       =     0.0000
Log likelihood = -1626.8003                     Pseudo R2         =     0.1574

------------------------------------------------------------------------------
    heartatk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |          1  (base)
      Female |      .40225    .041006    -8.93   0.000     .3293997    .4912121
             |
         age |    1.089351   .0053118    17.55   0.000     1.078989    1.099811
       _cons |    .0005532   .0001726   -24.04   0.000     .0003002    .0010195
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Wie man sieht haben sich die Koeffizienten geändert. Die dargestellten Effekte sind natürlich identisch, wir haben nur die Darstellungsform verändert. Die Interpretation ist folgendermaßen: Frauen haben, im Vergleich zu Männern, ein um den Faktor 0.40 geringeres Chancenverhältnis, einen Herzanfall zu erleiden. Bedenke, wie bereits oben erläutert: odds-ratios kleiner als 1 bedeuten eine geringere Wahrscheinlichkeit, während odds-ratios größer als 1 eine höhere Wahrscheinlichkeit meinen. Ist ein odds-ratio genau 1, so ist die Wahrscheinlichkeit in beiden Gruppen identisch.

In der Vergangenheit wurde gezeigt, dass sich logits und odds-ratios, trotz ihrer weiten Verbreitung und Tradition, in manchen Fällen nicht mehr als robuste Interpretationsmöglichkeiten erweisen. Heutzutage zieht man **Average Marginal Effects** (AMEs) vor. Um diese Werte zu erhalten, lässt man zunächst das gewünschte Modell laufen und tippt dann:

```
margins, dydx(*)
```

```
Average marginal effects                        Number of obs     =     10,349
Model VCE    : OIM

Expression   : Pr(heartatk), predict()
dy/dx w.r.t. : 2.sex age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |          0  (base)
      Female |   -.036853    .004074    -9.05   0.000    -.0448378   -.0288682
             |
         age |   .0035083   .0002284    15.36   0.000     .0030606     .003956
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Dieser Code bedeutet, dass man marginale Effekte anfordert. Die Option dydx(*) heißt, dass man AMEs möchte, der Stern in der Klammer bedeutet dabei „für alle erklärenden Variablen im Modell“. Das Ergebnis für Frauen ist -0,03685. Die Interpretation ist folgendermaßen: Frauen haben, im Vergleich zu Männern, eine um 3,69 Prozentpunkte (nicht Prozent!) geringere Wahrscheinlichkeit, einen Herzanfall zu erleiden (unter Kontrolle des Alters). Für Alter ist der Effekt, dass die Wahrscheinlichkeit, einen Herzanfall zu erleiden, mit jedem Jahr um 0,35 Prozentpunkte ansteigt (unter Kontrolle des Geschlechts). Beide Effekte sind hochsignifikant.

Wie wird ein AME berechnet? Die Idee dabei ist, alle Informationen in den Daten zu nutzen. Zunächst setzt Stata intern alle Geschlechter auf 1 (Mann), also alle Frauen im Modell werden so behandelt, als wären sie Männer. Die anderen Informationen, z.B. das Alter, bleiben unverändert. Dann berechnet Stata für jeden Fall die individuelle Wahrscheinlichkeit, einen Herzanfall zu erleiden und gibt dann die durchschnittliche Wahrscheinlichkeit aus. Anschließend wird der gesamte Prozess wiederholt, wobei nun das Geschlecht auf 2 (Frau) gesetzt wird. Am Ende werden die beiden kontrafaktischen Wahrscheinlichkeiten voneinander abgezogen. Die Differenz ist der AME für das Geschlecht.

Nun wollen wir diese Effekte über vorhergesagte Wahrscheinlichkeiten veranschaulichen. Eine sehr einfache Weise ist es, den allgemeinen Durchschnittswert für das Sample zu verwenden. Dazu tippt man einfach

```
margins
```

Das Ergebnis ist 0,04599. Die Interpretation ist, dass die durchschnittliche Wahrscheinlichkeit, dass eine Person aus dem Sample einen Herzanfall erleidet, etwa 4,6 % beträgt. Dieser Wert ist jedoch problematisch, da er stark von der Samplezusammensetzung abhängt. Haben wir beispielsweise sehr viele alte Personen im Sample wird dieser Wert wohl recht hoch ausfallen, obwohl die Wahrscheinlichkeit, einen Herzanfall zu erleiden, für junge Personen niedrig sein wird. Um dieses Problem zu umgehen, können wir Wahrscheinlichkeiten für bestimmte Werte anfordern, etwa in Bezug auf Alter und Geschlecht:

```
margins sex, at(age = (20 40 60 80))
```

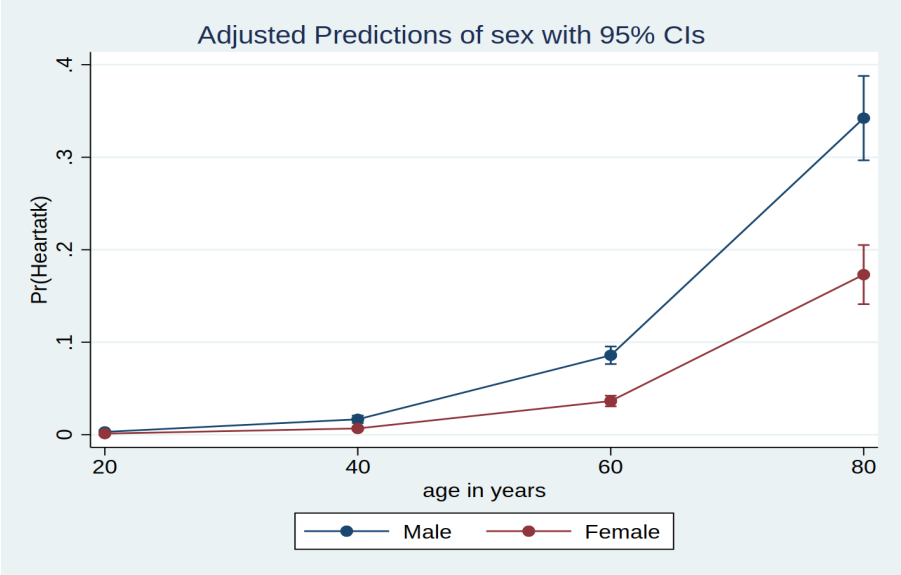
```
Adjusted predictions                            Number of obs     =     10,349
Model VCE    : OIM

Expression   : Pr(heartatk), predict()
1._at        : age             =          20
2._at        : age             =          40
3._at        : age             =          60
4._at        : age             =          80
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at#sex						
1#Male	.0030542	.0006609	4.62	0.000	.0017588	.0043496
1#Female	.0012308	.0002812	4.38	0.000	.0006796	.001782
2#Male	.0166828	.0020719	8.05	0.000	.0126221	.0207436
2#Female	.0067783	.0009568	7.08	0.000	.0049029	.0086536
3#Male	.0858875	.0048437	17.73	0.000	.0763941	.095381
3#Female	.0364179	.0029538	12.33	0.000	.0306287	.0422072
4#Male	.3422511	.0232669	14.71	0.000	.2966488	.3878534
4#Female	.1730791	.0163132	10.61	0.000	.1411058	.2050525

Dies heißt, wir möchten für alle Werte der Variable sex (also für Männer und Frauen) die Wahrscheinlichkeiten für die Altersstufen 20, 40, 60 und 80 haben. Nun können wir bestimmte Werte ablesen. Die Wahrscheinlichkeit für eine Frau, die 20 Jahre alt ist, beträgt 0,123 %. Die Wahrscheinlichkeit für einen Mann, der 80 Jahre alt ist, beträgt 34,2 %. Diese Darstellung ist sehr nützlich, wenn man bestimmte Wahrscheinlichkeiten in einer numerischen Form benötigt. Zur Interpretation eignet sich eine Grafik besser. Dazu tippt man im Anschluss

```
marginsplot
```

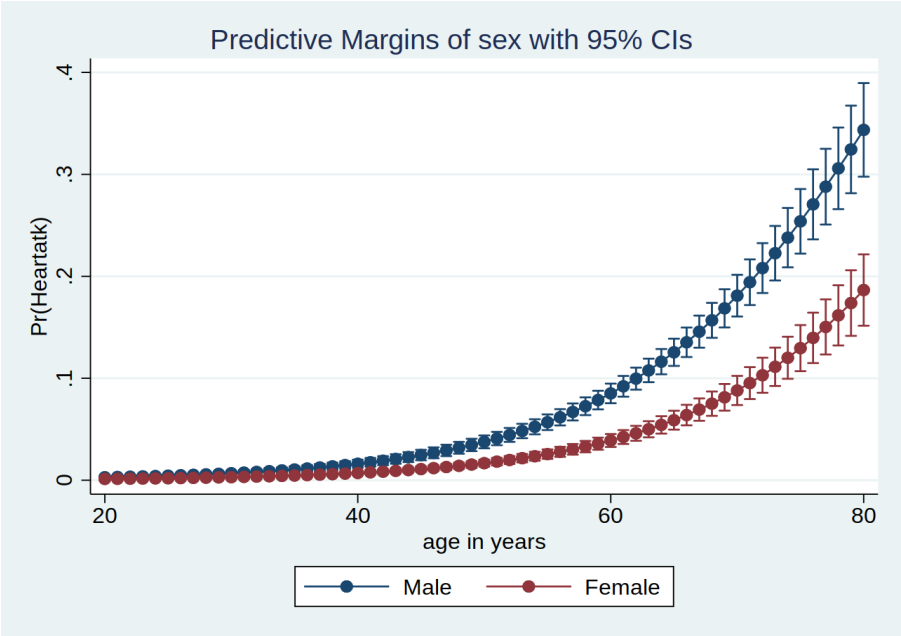


Wir nennen dies einen **Conditional Effects Plot**. Hier sehen wir deutlich, wie die Wahrscheinlichkeit mit dem Alter stark ansteigt und zudem für Männer noch viel stärker als für Frauen. Dabei sollte man jedoch bedenken, dass das aktuelle Modell keine weiteren Kontrollvariablen beinhaltet, also beispielsweise Scheinkorrelationen vorhanden sein können. Zudem sind keine Interaktionen berücksichtigt.[2] Für ein besseres Modell nehmen wir noch folgenden Kontrollvariablen ins Modell auf: BMI, Region, Ethnie und Vitamin-C-Spiegel. Danach berechnen wir die gezeigten Grafiken erneut mit mehr Detailstufen.

```
logit heartatk i.sex c.age i.region i.race c.vitaminc
quietly margins sex, at(age = (20(1)80)))[3]
marginsplot
```

2 Jedenfalls nicht explizit. Hier sind logistische Modelle im Vergleich zu linearen Regressionen flexibler, da diese bereits implizit alle erklärenden Variablen interagieren lassen. Wir sehen dies schon daran, dass der Alterseffekt für Männer und Frauen unterschiedlich ist.

3 Die Option *quietly* sorgt dafür, dass der Output nicht gezeigt wird, sondern die Ergebnisse nur intern gespeichert werden.



Wir sehen deutlich, dass die früheren Trends auch mit Kontrollvariablen weiterhin vorhanden sind. Indem Werte für jedes Alter berechnet werden, sind die Linien insgesamt weicher und runder. Wenn man möchte, kann man die Darstellungsform der Punkte und Konfidenzintveralle noch verbessern. Informationen finden sich dazu im Handbuch (*help marginsplot* eingeben).

Interaktionseffekte

Eine Interaktion liegt dann vor, wenn man vermutet, dass zwei Variablen miteinander interagieren, also der Effekt einer Variable nicht für alle Personen identisch ausfällt. Ein fiktives Beispiel: man könnte annehmen, dass ein hoher BMI für Männer nachteilig ist, während er für Frauen vorteilhaft wäre, oder anders ausgedrückt: dicke Männer haben höhere Chancen auf einen Herzinfarkt als dünne Männer, während dicke Frauen eine geringere Wahrscheinlichkeit haben als dünne Frauen. In diesem Fall läge eine Interaktion zwischen BMI und Geschlecht vor. Jedoch erscheint diese Annahme in der Realität unsinnig und zeigt auf, dass man Interaktionseffekte nur dann ins Modell aufnehmen sollte, wenn man theoretische Fundierungen dafür hat.

Für unser Beispiel wollen wir testen, ob ein Interaktionseffekt zwischen Diabetes und Geschlecht vorliegt. Es ist eine binäre Variable im Datensatz vorhanden, die angibt, ob jemand Diabetes hat oder nicht (*tabulate diabetes*). Grundsätzlich ist das Vorgehen folgendermaßen. Man rechnet drei Modelle, wobei das erste Modell nur die zentrale erklärende Variable enthält. Das zweite Modell enthält zusätzlich noch alle Kontrollvariablen sowie die Interaktionsvariable, aber nicht den Interaktionseffekt. Das dritte Modell enthält dann zusätzlich noch den Interaktionseffekt. Wer gewohnt ist, Interaktionsvariablen manuell zu erstellen, kann nun diese Technik ablegen. In Stata wird mittlerweile ein anderer Weg bevorzugt, der zudem deutlich bequemer ist. Da wir unser Modell einfach halten wollen, nehmen wir als einzige Kontrollvariable das Alter auf. Unsere Modelle würden demnach folgendermaßen aussehen:

```
logit heartatk i.sex                                    //Modell 1
estat ic
logit heartatk i.sex i.diabetes c.age                   //Modell 2
estat ic
logit heartatk i.sex##i.diabetes c.age                  //Modell 3
estat ic
```

Der Befehl *estat ic* berechnet zusätzlich für jedes Modell den AIC, der dazu dient, verschachtelte Modelle miteinander zu vergleichen. Grundsätzlich sollte das Modell am besten sein (in Bezug auf den Fit), das den *geringsten* AIC Wert aufweist. Die zwei Rauten (##) bedeuten, dass wir einen Interaktionseffekt zwischen sex und diabetes anfordern. Wichtig ist es dabei stets die Faktorvariablennotation zu verwenden. Tippt man nur ein Rautenzeichen, so bedeutet dies, dass man *nur* den Interaktionseffekt anfordert, jedoch nicht die beiden Haupteffekte. Das ist in den wenigsten Fälle gewünscht.

```
Logistic regression                               Number of obs     =     10,349
                                                  LR chi2(4)        =     622.82
                                                  Prob > chi2       =     0.0000
Log likelihood = -1619.1843                       Pseudo R2         =     0.1613
```

heartatk	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex						
Female	-.9319271	.1096803	-8.50	0.000	-1.146897	-.7169577
1.diabetes	.5628876	.1918745	2.93	0.003	.1868204	.9389548
sex#diabetes						
Female#1	.1195121	.3026386	0.39	0.693	-.4736486	.7126729
age	.0836348	.0048961	17.08	0.000	.0740385	.093231
_cons	-7.432045	.3117398	-23.84	0.000	-8.043043	-6.821046

```
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	10,349	-1930.594	-1619.184	5	3248.369	3284.592

Note: N=Obs used in calculating BIC; see [R] BIC note.

Wir sehen, dass der Interaktionseffekt im letzten Modell nicht signifikant ist (p=0,693). Auch sehen wir, dass der AIC im zweiten Modell (3246,524) niedriger ist, als im Modell 3 (3248,369). Wir kommen demnach zu dem Schluss, dass kein Interaktionseffekt vorliegt.

Allerdings soll dennoch kurz gezeigt werden, wie eine Interpretation aussehen kann, wenn im eigenen Modell ein signifikanter Effekt gefunden wird. Dazu können wir wieder AMEs benutzen.

```
margins, dydx(diabetes) at(sex = (1 2))
```

Dieser Befehl bedeutet, dass man den Effekt einer Diabeteserkrankung auf das Herzinfarktrisiko getrennt für beide Geschlechter berechnet. Das Ergebnis ist folgendermaßen:

```
Average marginal effects                        Number of obs     =     10,349
Model VCE    : OIM

Expression   : Pr(heartatk), predict()
dy/dx w.r.t. : 1.diabetes
1._at        : sex             =           1
2._at        : sex             =           2
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
0.diabetes	(base outcome)					
1.diabetes _at						
1	**.0378713**	**.0150591**	**2.51**	**0.012**	**.0083561**	**.0673866**
2	**.0234208**	**.0100501**	**2.33**	**0.020**	**.003723**	**.0431187**

Note: dy/dx for factor levels is the discrete change from the base level.

Haben Männer Diabetes, so erhöht sich ihr Risiko, einen Herzinfarkt zu erleiden, um 3,79 Prozentpunkte. Für Frauen beträgt die Erhöhung nur 2,34 Prozentpunkte. Beide Effekte sind statistisch von 0 verschieden, da die Werte signifikant sind. Wir wissen somit, dass Diabetes das Herzinfarktrisiko erhöht. Allerdings sehen wir auch, dass sich diese beiden Effekte *voneinander* nicht signifikant unterscheiden, da sich die Konfidenzintervalle deutlich überlappen. Wir können daher nicht aussagen, dass Männer oder Frauen stärker unter Diabetes leiden als das jeweils andere Geschlecht.

Diagnostik

Insgesamt sind logistische Regressionen in der Regel robuster als lineare (OLS) Regressionen. Dennoch sollte man gewisse Aspekte überprüfen und natürlich mit einer gründlichen theoretischen Ausarbeitung des Modells beginnen. Als Grundlage für alle gezeigte Diagnostik ist folgendes Modell:

```
logit heartatk i.sex c.age i.region i.race c.vitaminc
```

Fehlspezifikation des Modells

Ein einfacher Test um zu prüfen, ob wichtige Variablen fehlen, ist der linktest. Dazu führt man zuerst das eigentlich Modell aus und gib dann folgenden Befehl ein

```
linktest
```

```
Iteration 0:   log likelihood = -1885.4259
Iteration 1:   log likelihood =  -1755.241
Iteration 2:   log likelihood = -1583.4662
Iteration 3:   log likelihood = -1570.4021
Iteration 4:   log likelihood = -1569.2842
Iteration 5:   log likelihood = -1569.2778
Iteration 6:   log likelihood = -1569.2778

Logistic regression                             Number of obs     =      9,971
                                                LR chi2(2)        =     632.30
                                                Prob > chi2       =     0.0000
Log likelihood = -1569.2778                     Pseudo R2         =     0.1677

------------------------------------------------------------------------------
    heartatk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .1783177   .2475509     0.72   0.471    -.3068733    .6635086
      _hatsq |  -.1432992   .0435348    -3.29   0.001    -.2286258   -.0579725
       _cons |  -1.026573   .3287029    -3.12   0.002    -1.670819    -.382327
------------------------------------------------------------------------------
```

Interessant sind für uns die Variablen _hat und _hatsq. Dabei sollte die erste Variable _hat ein signifikantes Ergebnis zeigen (p kleiner als 0,05), während die zweite Variable _hatsq nicht signifikant sein sollte. In unserem Beispiel ist es genau umgekehrt, was aufzeigt, dass unser Modell Probleme hat. Wir sollten also überlegen, ob wichtige Prädiktorvariablen fehlen oder ob unnötige Variablen im Modell sind. Oftmals kann es helfen, Interaktionsvariablen aufzunehmen. Jedoch sollte man diesen Test nicht zu ernst nehmen, da er keine inhaltlichen Aussagen machen kann. Wenn wir davon überzeugt sind, dass unsere Theorie korrekt ist und wir alle wichtigen Variablen im Modell haben, sollte die Theorie dem linktest vorgezogen werden.

Fallzahlen

Eine logistische Regression sollte mindestens 100 Fälle einbeziehen und damit deutlich mehr als eine normale Regression (dort reichen etwa 30 Fälle, je nach Anzahl der Variablen im Modell). Wird diese Fallzahl nicht erreicht, kann man stattdessen eine exakte logistische Regression durchzuführen (*help exlogistic*). Auch ist es in einer logistischen Regression wichtig, dass keine Nullzellen vorkommen, es also „leere“ Kategorien gibt. Dies kann besonders dann Vorkommen, wenn man kategoriale Prädiktoren einbezieht, bei denen manchen Kategorien nur schwach besetzt sind. Grundsätzlich wird das Problem entschärft, je mehr Fälle im Modell sind. Unser aktuelles Modell hat über 10.000 Fälle, was sehr gut ist.

Multikollinearität

Wie auch bei einer linearen Regression kann es problematisch sein, wenn eine hohe Korrelation zwischen erklärenden Variablen besteht. Dies kann man testen, wozu man ein Ado benötigt. Dieses kann direkt in Stata heruntergeladen werden. Dazu tippt man

```
search collin
```

und sucht im sich öffnenden Fenster nach dem Paket von Philip B. Ender. Hat man dieses gefunden, kann es direkt installiert werden. Ist dies geschehen, kann man den Test durchführen. Dazu gibt man den Namen des Befehls ein und listet danach alle erklärenden Variablen auf, also

```
collin sex age region race vitaminc
```

```
 Collinearity Diagnostics

                      SQRT                   R-
 Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
      sex      1.04    1.02    0.9655      0.0345
      age      1.01    1.01    0.9892      0.0108
   region      1.01    1.01    0.9876      0.0124
     race      1.02    1.01    0.9803      0.0197
 vitaminc      1.05    1.03    0.9490      0.0510
----------------------------------------------------
 Mean VIF      1.03

                        Cond
       Eigenval         Index
---------------------------------
    1     5.4128          1.0000
    2     0.2270          4.8831
    3     0.1272          6.5229
    4     0.1137          6.8986
    5     0.0946          7.5653
    6     0.0246         14.8271
---------------------------------
 Condition Number        14.8271
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.9364
```

Die Ergebnisse sehen gut aus, da der VIF (Variance Inflation Factor) für alle Variablen deutlich unter 10 liegt. Ist er höher, liegt wohl eine Multikollinearität vor. In diesem Fall ist es sinnvoll eine Variable mit einem sehr hohen VIF aus dem Modell zu entfernen und dann zu prüfen, ob das Ergebnis besser wird. Dann kann man davon ausgehen, dass bereits eine andere Variable die gleichen Informationen enthält wie die entfernte Variable. Ausnahmen sind abgeleitete Variablen, also beispielsweise Interaktionen oder quadrierte Terme, für die eine neue Variable erstellt wurde. In diesem Fall ist es normal, dass die abgeleitet Variable mit der Ursprungsvariable korreliert.

Einflussreiche Fälle

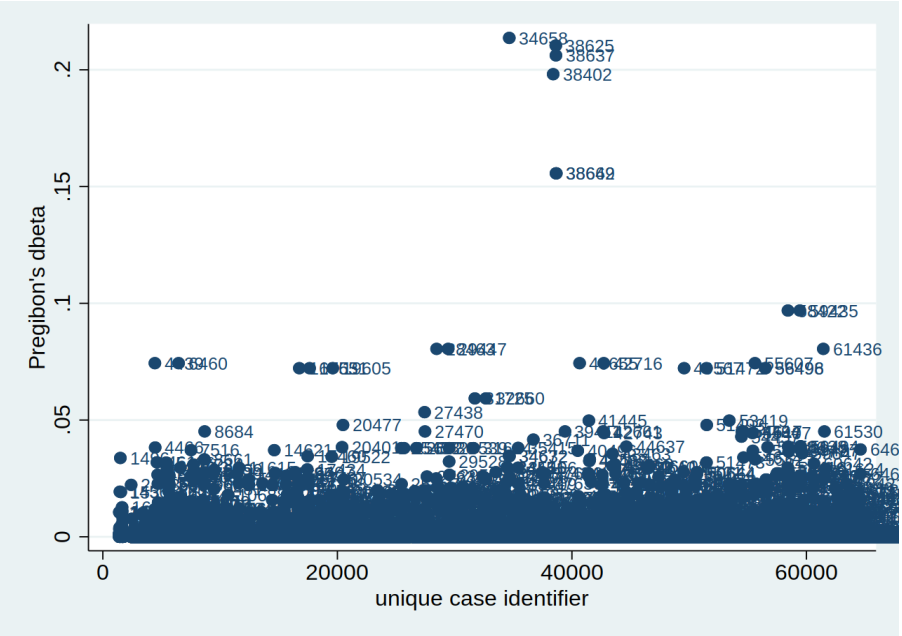
Die letzte Diagnostik ist nicht unbedingt notwendig, kann jedoch nützlich sein. Die Idee ist zu prüfen, ob einige wenige, sehr absonderliche Fälle das Modell insgesamt stark verzerren. Dies kann etwa sein, wenn Fälle falsch kodiert wurden oder aber Muster dabei sind, die extreme Ausreißer darstellen. Letztlich muss jeder selbst entscheiden, ob man solche Fälle im Modell belassen will oder nicht. Um diese zu findet kann man **Pregibons Beta** verwenden. Diese Variable fasst

zusammen, wie stark ein Fall das Modell beeinflusst. Zunächst wird eine neue Variable erstellt, die für jeden Fall angibt, wie stark dieser das Modell beeinflusst:

```
quietly logit heartatk i.sex c.age i.region i.race c.vitaminc
predict beta, dbeta
```

beta ist dabei der Name der neu erstellten Variable und die Option *dbeta* gibt an, dass wir Pregibons Beta anfordern. Nun kann man einen Scatterplot nutzen, um Ausreißer zu entdecken.

```
scatter beta sampl, mlabel(sampl)
```



Scatter bedeutet, dass wird einen Scatterplot anfordern. Dann wird als erstes die Varible der y-Achse genannt, dann die Variable der x-Achse. Dazu benutzen wir einfach die ID für jeden Fall. Die Option *mlabel(sampl)* gibt an, dass wir jeden Datenpunkt mit der ID des Falls labeln. So können wir Fälle bequem zuordnen.

Wir sehen, dass es einige wenige Fälle gibt, die Ausreißer darstellen, etwa der Fall 38402. Wir können diesen Fall entweder genauer auf Kodierfehler untersuchen, ihn löschen oder einfach von der Untersuchung ausschließen.

```
list heartatk sampl sex age region race vitaminc if beta > 0.07
list heartatk sex age region race vitaminc if sampl == 38402
drop if sampl == 38402   //Datensatz vorher abspeichern!
logit heartatk i.sex c.age i.region i.race c.vitaminc if sampl != 38402
```

So können wir mit allen Fällen verfahren, die uns problematisch erscheinen. Letztlich gibt es keinen Goldstandard der besagt, wie mit solchen Fällen umgegangen werden soll. Wichtig ist nur dass man in seinem Forschungsbericht beschreibt, wie man die Daten prozessiert hat, damit andere Forscher nachvollziehen können, was wir gemacht haben.

Quellenverzeichnis

Behnke, Joachim (2014): Logistische Regressionsanalyse. Eine Einführung, Wiesbaden.

Long, Scott; Freese, Jeremy (2014): Regression Models for Categorical Dependent Variables Using Stata, College Station, Texas.

Mood, Carina (2010): 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 (2008): Regressionsanalyse: Theorie, Technik und Anwendung, Wiesbaden.

Wolf, Christof; Best, Henning (Hg.) (2010): Handbuch der Sozialwissenschaftlichen Datenanalyse, Wiesbaden.