Neuere Methoden empirischer Wirtschaftsforschung

(Anwendung der Software „SPSS“)

Feedback bitte über die e-Mail Adresse auf der Homepage www.terragon.de.

(Vielen Dank an Dr. Stecking für diese hervorragende Vorlesung)
www.iksf.uni-bremen.de

Allgemein soll man lernen wie man wissen aus empirischen Daten extrahieren kann.

4 Wichtige Ziele:

- Wirkungsanalyse: Welcher Werbeansatz kann zu welchen Verkaufszahlen führen ?
- Prognose: Zeitreihenprognose, Aktien und Wechselkursprognose.
- Klassifikation: Kreditwürdigkeitsprüfung auf grund von ein paar Informationen d. Kreditnehmers.
- Segmentierung: In welche Teilgruppen lässt sich der Markt einteilen ?

Theoretische Methoden lernen wir in der Vorlesung.
Praktisch werden wir SPSS im PC Raum benutzen. Zur Zeit der Vorlesung.

Was ist Data-Mining ?

-DM entdeckt wertvolle Informationen aus sehr großen Datenbeständen: "Knowledge Discovery". Früher hatte Man Matrizenmäßige Auswertung, jetzt geht man eher über Algorithmen, weil die Matrizen bei millionen Einträgen zu umfangreich werden. Zum Beispiel Krankenkassen mit Millionen von Patienten mit jeweils hunderten von Einträgen, ABrechnungen, Krankheiten etc.
-Man hat kein Vorwissen von den Daten, sondern will einfach wissen was die Daten die ich habe bieten können.
-unter DM werden alte und neue Methoden zusammengefasst.
-Kommerzielle DM Software ist SPSS. "Clementine" gehört zu SPSS, wurde aber eingekauft von SPSS und ist vom Umgang anders, Datenaustausch ist aber ziemlich gut.

Methoden des Data-Mining :
- Statistische Verfahren : Lineare und Logistische Regression. Diese Verfahren funktionieren auf Knopfdruck und man braucht sie vor allem auch als Vergleichsverfahren.
- Clusteranalyse : Oberbegriff für verschiedene Methoden. Wir behandeln 3 bis 4. Die sind in SPSS.
- Entscheidungsbaumverfahren.
- Künstliche Neuronale Netze : Überbegriff über verschiedene Methoden. MLP, RBF-Netze. Man versucht das Gehirn nachzubilden seit den 40er jahren. Man will selbstorganisierte Merkmalskarten (Kohonen-Netze). Man kann damit ganz komplizierte Zusammenhänge abbilden. Was wirkt auf was ? Verteilung der Variablen. Wir machen das nur am Rande, es wäre eigentlich eine Veranstaltung an sich. es gibt Algorithmen, die die Organisation des Gehirns abbilden.

Daten-Preprocessing :

Bringt theoretisch wenig, ist in der Praxis aber relevant, weil man oft Messfehler etc hat die man vorher anpassen und korrigieren muss.

1. **Datenmatrix**
   Eine Matrix von 4 Variablen (Personnummer, Geschlecht, Größe) Siehe Tabelle!

2. **Skalenniveau**
   Skalenniveau der betrachteten Variablen bestimmt zum Teil die Methoden die wir anwenden können.

Nicht Metrisch:
-Nominal (Geschlecht zb. man kann schlecht mit ihnen rechnen, sondern vor allem Häufigkeiten feststellen. Auspägung A und B sind gleichwertig aber verschieden)
- Ordinal (Bildungsabschluss, man kann sagen welches Merkmal höher ist als das andere. Wie groß der Unterschied ist weiss man nicht. Man verwendet dafür vor allem Median und Quantile)
Metrisch :
- Intervall (Temperatur, Skalen mit gleichgroßen Abständen ohne natürlichen Nullpunkt. Man kann damit auch nur Addition und Substraktion verwenden.)
- Ratio-Skala. (Hier hat man dazu den Natürlichen Nullpunkt, und da kann man auch Multiplizieren und Dividieren).

Man kann aus höheren Variablen niedere machen, umgekehrt geht es nicht. Man macht es aber trotzdem um einheitliche Skalenniveaus zu erzeugen.

3. Kodierung, Skalierung und Transformation von Variablen.

Kodierung :
- Intervall und Ratioskala als Metrische verwenden. Es wird dann auch dividiert und so.
- Nominalskala : Man benutzt "Dummy Coding" heisst Nominale Skalen in Null und Eins umzuwandeln und macht sie somit metrisch. Zum Beispiel Mann und Frau Einteilung. Man sagt bei 0 und 1 gibt es ja nur einen Abstand von 1 dazwischen.
- Ordinalskala: Wenn sie nicht allzu schräg ist, wird sie auch als metrisch genommen. Zum Beipsiel Zufriedenheit von Konsumenten zwischen 0 und 10. Man will sie auch metrisieren.

Skalierung :
Bei verschiedenen Messniveaus (zum beispiel Meter oder cm bei Messungen) muss man über eine Transformation angleichen. Zb: Dow Jones und Dax auf einen Nenner bringen.

Z-Transformation:
   zi = (xi - Mittelwert) / Standardabweichung
Das verlangt, dass ein Mittelwert existriert und sie annähernd Normalverteilt sind !

T-Transformation:
   ti = (Xi - Xmin) / (Xmax - Xmin)
min = Minimale Ausprägung der Variablen.

Bei beiden muss man auf Ausreisser achten die sich aus Messfehlern ergeben.

Transformation :
- (Prognoseorientiert) Differenzenbildung
- (Prognoseorientiert) Lead / Lag Verschiebung (Wert von heute mit wert von gestern vergleichen)
- Konstruktion neuer Variablen (Aus 2 Var eine neue Machen)

Dummy Coding:
Man macht wenn man 5 Auspägungen hat, 5 Neue Variablen die jeweils so heissen wie eine Auspägung. Dann bekommen diese neuen Variablen jeweils eine 0 oder 1 zugeordnet, jeweils wenn es zutrifft oder nicht zutrifft. (Es reichen bei 5 auch 4 neue, weil der letzte ja auch überall Null haben kann. Das ist wichtig, wenn man Abhängige Variablen hat ?!) Man kann dann so tun als wären diese Variablen Metrisch.
Differenzenbildung / Lead / lag Verschiebungen :
Datum DowJones DJ/Diff DJ/LagDJ/Lag

Lineare Regression

1.Anwendunsgbereiche

Woher kommt der Begriff ?
Universales Regressionsgesetz : Jede Besonderheit eines Menschen wird an seine Nachkommen weitergegeben, aber im Schnitt in einem geringeren Grad.

Beispiel: Es wurde überprüft, wie groß die Söhne im Schnitt gegenüber ihren Vätern sind. In der Regel war es so, dass sie kleiner sind. Es trat also eine Regression ein !
Man hat da bei der ersten Anwendung einen Rückgang festgestellt. Deshalb reden wir von Regressionsanalyse und nicht von Progrogressionsanalyse.
Abhängige Variable ist hier die Größe des Sohnes.

Anwendungsbereiche
a) Ursachenanalyse : wie groß ist der Einfluss der abhängigen auf die unabhängige Var. ?
b) Wirkungsprognosen :
c) Zeitreihenanalyse :

2.Das Modell der linearen Regression

Wir haben eine abhängige Variable (Y) und eine oder mehrere Variablen (X1,X2).
Diese müssen alle metrisch sein !
Es kann nur eine abhängige Variable geben !

Modell besagt, dass Y von einer linearen Kombination der unabhängigen Variablen abhängt.
$y = b0 + b1*x1 + b2*x2 + ...$

Die Linearkoeffizienten b sollen so bestimmt werden, dass die Summe der aufsummierten Fehlerquadrate (differenz zwischen realwert y und dem geschätzten y) möglichst minimiert wird.
Bei zwei Variablen können wir die Werte noch von Hand berechnen, so wie bei Statistik II.

Vorraussetzungen dafür sind :
-Fehler haben eine **konstante Varianz**. Also egal welchen Abscnitt der daten man herausgreift, die selbe Varianz.
-Störgrößen sollen voneinander **unabhängig** sein. Im Beispiel von Zeitreihenanalyse darf der Wert von Zeitpunkt t1 nicht vom wert aus t0 abhängen. (Autokorrelation) Man kann dann nicht den Index an sich betrachten, sondern das Verhältnis von heute zu gestern. Dadurch kann man die Abhängigkeiten identifizieren.
-Zwischen den ganzen X darf **keine exakte lineare Abhängigkeit** bestehen. Es darf sich keine Variable aus den anderen errechnen lassen. Häufig hängen Variablen nicht exakt ab, sondern haben eine starke Korrelation. Dann gilt das Modell als ineffizient.(Keine exakte Multikollinearität)
-Die Störgrößen müssen **normalverteilt** sein. Bei Stichproben über 30 sind alle Variablen normalverteilt (nach dem zentralen Grenzwertsatz).

3.Prüfung des Gesamtmodells

Der erste Schritt nach der Koeffizientenbrechnung ist die Prüfung der Aussage !
Um das machen zu können, machen wir eine Streuungszerlegung :
(Streuung ist die Abweichung eines Variablenwertes von dem Arithmetrischen Mittels. Das aufsummiert ist die Gesamtstreuung)
Die Gesamtstreuung wird in erklärte Streuung + nicht erklärte Streuung (Fehlerwerte).
Dass diese Aufteilung so geht, nutzt man in verschiednen Prüfmaßen :

1. Bestimmtheitsmaß = erklärte Streuung / gesamte Streuung = R^2
   Je höher der Anteil der erklärten Streuung desto besser. Er liegt zwischen 0 und 1.

2. Korrigiertes Bestimmtheitsmaß : Wenn man Modelle vergleichen will die viele unabhängige Variablen haben. Bei 100 * 100 kann man auch $R^2=1$ bekommen, obwohl es einem garnichts erklärt.
   R^2KORR = Anzahl der Variablen (J) / Anzahl der Fälle (N) (?! Folie gucken)

3. Standartfehler der Schätzung
   Er sagt uns, wieviel wir im Schnitt daneben liegen.
   Wurzel aus (nicht erklärte Streuung / Anzahö Fälle - Anzahl unabhängiger Var. - 1)

4. Wesentlich ist jedoch : F-Statistik !

1-3 sind nur deskriptiv und sind schwerer zu beurteilen.
F-Statistik prüft, ob die Ergebnisse der Stichprobe für die Grundgesamtheit gelten.
On der Zusammenhang im Modell also auch auf die GG übertragbar ist.
F-Wert basiert auf der Streuungszerlegung.

F emp. = (erklärte Streuung / J) / (nicht erklärte Streuung / J-N-1)

J = Anzahl der Variablen
N = Anzahl der Fälle
Wir testen also eine Nullhypothese.
Sie lautet: H0 = Die Wahren Regressionskoeffizienten der GG sind = 0

4 Schritte der F-Statistik : (verweis auf den Backhaus, wo das alles steht)

a) Brechnung des empirischen F-Wertes mit der Formel von eben.

b) Signifikanzniveau vorgeben (Irrtumswahrscheinlichkeit Alpha)
Wie groß ist die Wahrscheinlichkeit, dass die Nullhypothese abgelehnt wird obwohl sie richtig ist ? Meistens nehmen wir 5% oder 1%.

c) Theoretischen F-Wert finden
Wenn alles reine Zufallsvariablen sind, folgt der F-Wert einer theoretisch F-Verteilung. Dieser Wert ist in einer Tabelle abzulesen. (F-Tabelle). Er hängt ab von der Anzahl der unabhängigen Variablen (J) und der Fälle (N). (Die nennt man 2 Freiheitsgrade)
Dazu nehmen wir die Irrtumswahrscheinlichkeit
und man weiss dann, dass in 5% der Fälle der F-Wert über 5,32 liegt. (95% sind kleiner)

d) Ist der empirische Wert größer als der theoretische ist, ist die Nullhypothese zu verwerfen.
Wenn man H0 verwirft, besteht ein signifikanter Zusammenhang.
Ist Tabellenwert > Empirischer Wert , dann besteht kein Zusammenhang !
Wir haben meistens Fälle in denen ein Zusammenhang besteht.

4.Die Prüfung einzelner Koeffizienten

Wenn im ersten Schritt von 3. H0 abgelehnt wurde,
dann Prüpfen wir jetzt, welcher der einzelnen Koeffizienten signifikant ist.

Beta-Werte

Beta j = empirischer Koeffizient j * (Standartabweichung von Xj / Standartabw. von Y)

Jetzt haben wir einen Vergleichsmaßstab zwischen den Koeffizienten. Die Koeffizienten die den größten Beta Wert haben, haben den größten Einfluss auf den gesamten Zusammenhang !

Multikollinearität

Die Untersuchung der linearen Abhängigkeit der Variablen Xj.
Schritte:

a) Korrelationsmatrix aufstellen. Wenn Werte über 0,9 sind (kleiner -0,9) haben wir einen Hinweis auf Multikoll.

b) Toleranz TOL = 1 - Rj^2

R ist wieder ein Bestimtheitsmaß. Für die Toleranz der Variablen j bauen wir ein Modell, in dem die Variable j einen bestimmten R^2 Wert bekommt. Je näher die Toleranz an 1 ist (also wenn R^2 nahe an Null ist) heisst es, dass die Variable überhaupt nicht abhängig ist.
Anzeichen dafür, wieviele Variablen in dem Modell voneinander abhängig sind.

c) Variance Inflating Factor (VIF) = 1 / TOL

Umkehrwert der Toleranz.

t-Statistik

Analog zum F-Wert. Hier werden die einzelnen Regressionskoeffizienten bj geprüft.
Wenn der F-Test sagt dass wir H0 ablehnen, bedeutet es dass nicht alle ungleich 0 sind.
Hier gehen wir nun in jede einzelne unabhängige Var vor, und testen ob sie signifikant sind.

H0 = Betaj = 0

t-Wert empirisch = bj / Standartfehler des Koeffizienten sbj

Diesen t-Wert muss man nun wieder mit dem tabellierten Wert vergleichen.

t-Wert theoretisch = Alpha (5%)

Jetzt wieder beide vergleichen : Wenn t-empirisch > t-theoretisch verwerfen wir H0,
und folgern, dass es einen Zusammenhang gibt zwischen bj und der abhängigen Variablen.

Vorgehensweise um eine lineare Regressionsanalyse durchzuführen:

- Um in SPSS lineare Regression aufzurufen, "Statistik" aus dem Menü wählen.
- Untermenü Regression. - linear.
- In dem Fenster, die Variable "Menge" als abhängige Variable wählen.
- unabhängige Variablen sollen dann die anderen 3 sein.

Hinter dem Button "Statistik..." kann man noch "Deskriptive Statistik" und "Kollineale..." Dann auf weiter klicken, und "Einfügen" klicken.
- Jetzt öffnet sich das "*Syntaxfenster*". (Das ist eine einfache Textdatei, die man immer wider ablaufen lassen kann. Dort sieht man auch die Struktur der SPSS Syntax)
In dem Fenster alles Markieren und in der Leiste auf "ausführen" klicken (Play-Button).
- Dann kommt man schon zum *Ausgabefenster*.

Es gibt Haupt und Unterbefehle:
- Hauptbefehle werden mit einem Punkt abgeschlossen.
- Unterbefehle werden durch einen "/" getrennt.

Wichtige Befehle sind :
"REGRESSION"
"DEPENDENT" = gibt an welche abhängig ist.
"METHOD=ENTER"
"LIST. /Menge" zum Anzeigen von etwas ?! Mal ausprobieren.

F-Test Erklärung:

Man hat 2 Urnen in die man nicht gucken kann.
- Urne A : In der Urne sind 1000 Kugeln von denen 500 weiss und 500 schwarz sind.
- Urne B : In der Urne sind 1000 schwarze Kugeln.

Man will jetzt herausfinden welche Urne man vor sich stehen hat, darf aber nur 5 mal reingreifen und dann die Farbe ansehen.
Man kriegt bei einer Stichprobe von 5 , 5 schwarze Kugeln !
Jetzt soll man entscheiden um welche Urne es sich handelt ?
Man kann sich also nur auf die Wahrscheinlichkeitsrechnung verlassen und die Wahrscheinlichere Urne wählen (B).
- Die Irrtumswahrscheinlichkeit ist die Wahrscheinlichkeit, dass man aus Urne A 5 schwarze zieht. Also in dem Fall 1/2 * 1/2 .... ist. = 1/32 = 3%.

Wir nehmen die Hypothese, dass wir aus A gezogen haben, lehnen diese aber ab , und nehmen die Wahrscheinlichkeit dass wir uns irren (3%) in Kauf.

In Bezug auf den F-Test, nennen wir Urne A jetzt "Nullhypothese" und haben es nicht mehr mit Kugeln, sondern mit zufällig verteilten F-Werten zu tun. Also hat man einen Beutel mit ganz vielen "kleinen Preisen von Plus" (8,31 - 0,4 - 1,23 - ....) die voneinander unabhängig sind.:-)
Man weiss, dass F-Werte meistens immer "klein" sind, wenn sie zufällig sind. (Meist irgendwas unter 3,0) Große Werte sind zwar möglich,aber sehr unwahrscheinlich.
Wenn man also einen großen F-Wert berechnet aus seinen empirischen Daten, dann ist es unwahrscheinlich, dass der Wert zufällig entstanden ist. !!!
Nulhypothese = Beta1 = Beta2 = Betaj = 0 , wird abgelehnt wenn der F-Wert über 5 ist.
Sie bedeutet, dass keine Variable irgendwas erklärt (=0)
In dem Beispiel *ist der F-Wert über 60. Das bedeutet, dass mindestens eine der Variablen etwas über die Absatzmenge aussagen muss*.

Wenn man das weiss, will man wissen, *welche der Variablen* signifikant ist ?
Der T-Wert gibt an, wie einflussreich eine Variable ist.
T - Wert = B / Standartfehler
Mann versucht die Nullhypothese, dass Beta=0 ist abzulehnen.

- Um zu sehen, welche der Variablen Signifikant sind, guckt man in die Spalte "**Signifikanz**", die angibt, wie groß die Irrtumswahrscheinlichkeit ist :
Die ist beim Preis 0,008 (0,8%) und bei den Ausgaben und Besuchen 0,000 .
Also sind alle 3 Variblanen Signifikant ! Sie tragen alle zur Erklärung der Absatzmenge bei, da sie alle unter 5% liegen.
Ausgaben = die Einflussreichste Variable, weil der T-Wert am höchsten ist (B / Standardf.) und weil der "Beta-wert" am höchsten ist.

- Das Vorzeichen sagt aus, ob es positiv oder negativ korrelliert. Positives Vorzeichen heisst, wenn man z.b. die Ausgaben erhöht, erhöht sich auch der Absatz. (Beim Preis umgekehrt)

Der Zusammenhang in dem Beispiel lässt sich so darstellen :
Menge = 763,650 Konstante - 45,177 Preis + 0,551 Ausgaben + 9,705 Besuche
Die Zahlen sind jeweils die B-Werte, die SPSS ausgibt.

In der Tabelle "Korrelationen" kann man sehen, wie die Variablen untereinander Korrelieren.

In der Datenmatrix jetzt nochmal lineare Regression anklicken und bei "Speichern..." oben die beiden "nicht standardisiert" anklicken.
Dadurch bekommt man dann 2 neue Variablen. "pre_1" und "res_1".

Logistische Regression (binär - keine multinominale)

2 Ausprägungen bei der abhängigen Variable. (Mann/Frau, Kredit/kein Kredit (!)...)
Bei Medizinern ist sie Tod/Lebendig, was dann von mehreren Einflüssen abhängt.
Abhängige Variable (Y) ist eine **Nominalskalierte Variable** ! Nicht wie bei lin. Regression.
Unabhängige Variablen sind auch metrisch. Hiervon kann es mehrere geben.

Modell : mit der e-Funktion

Abhängige Variable Yi nimmt nur die Werte 1 oder 0 an.
Der Gedankensprung ist, dass es beobachtbare Dinge in der Welt gibt. Wenn ein Kreditnehmer früher mal nicht gezahlt hat, bekommt er eine 0.
Hinter dieser Beobachtung steckt jedoch eine Wahrscheinlichkeitsüberlegung. Es gibt eine Wahrscheinlichkeit, mit der ein Kreditnehmer kreditwürdig ist, oder nicht. Dies ist latent (nicht beobachtbar)
Man nimmt also an , dass die 1 oder 0 mit einer bestimtmen Wahrscheinlichkeit auftritt.
Yi = 1 = 0,30
Yi = 0 = 1 - 0,30 = 0,70

Wir nehmen an, dass die Wahrscheinlichkeit von Yi = 1 abhängig ist , von:
konstante + b1*x1 + b2*x2+....
Man kann hier kein = P schreiben, weil ...

odd (Chance zwischen 0 und unendlich)
1. Wir konzentrieren uns nicht auf die Wahrscheinlichkeit, sondern auf die "Chance":
Chance = P / 1-P
Wenn man sagt die Chance ist 50/50, sagt man die Wahrscheinlichkeiten 0,5 / 0,5 = 1 z.b.
Verschiebt sich die Wahrscheinlichkeit, kann die Chance beliebig groß werden (nicht beliebig klein)
Man bläst den Wertebereich nun von 1 auf +unendlich auf.

Logit (zwischen -unendlich und +unendlich)
2. **logit i = ln (P / 1-P)**
Man nimmt also aus der Chance noch den Logarithmus.(Log wird bei Zahlen die kleiner als 1 sind, negative Werte an)
Je größer die Chance also wird, desto größer wird der Logarithmus.

Man hat nun aus einer Wahrscheinlichkeit zwischen 0 und 1 , einen Wert zwischen - und + unendlich gemacht.
Die logistische Regression modelliert diesen LOGIT wert.
(Formeln siehe Folie)

Die Gleichung der logistischen Regression ist die Aufgelöste Gleichung mit der e-Funktion :

Wir haben jetzt ein Yi, das 1 wird, wenn eine der latenten Variablen Yi* größer als 0 wird.
Sonst ist Yi = 0.
Die latente Variable ist wieder eine lineare kombination (b1*x1)
Summe aus : bo + (b1*x1) **+ ui**
=> Z + ui
ui = Residuum (Restgröße)
Z = Der ganze Term vor ui , kürzen wir mit Z ab.

Zi + ui = Yi*
Wenn das Yi* > 0 , dann wird Yi=1 (es wird zutreffen)
Dann ist gleichzeitig auch ui > -Zi.
also : P (ui > Zi) = P (ui <= Zi)
Bei der standartnormalverteilung haben ich nun links ein -Zi , und die Fläche rechts davon ist die Wahrscheinlichkeit, dass u größer ist als Zi.
Durch die symmetrie, ist die Fläche links von einem +Zi genausogroß. (Statistik N.V.)

Jetzt sagen wir, dass die Dichtefunktion (h) von ui symmatrisch ist.
Die Verteilungsfunktion (aufleitung) ist nun logistisch Verteilt.
F (Zi) = 1 / (1+e hoch -Zi)
Gezeichnet : S-Schwung von unten links nach oben rechts, zwischen 0 und 1.

Die logistische Funktion ist so ähnlich wie die Normalverteilungsfunktion.
Störterme ui sind annähernd Normalverteilt, weil sie exakt logistisch verteilt sind !
Das ist ein toller Trick, weil man sich nicht die Streuungen wie F-Wert ansehen müssen.

Wie komme ich aber nun aus die Modellparameter ? (Die Beta j , bj)

Dies macht man mit einer Likelihood Funktion. (Folie)
Das schauen wir uns aber nicht genau an.
Wenn gemessenes Yi = 1 ist, soll P auch = 1 sein.
Wenn Yi = 0 ist, soll auch 1-P = 1 sein.

Wenn ich die Parameter nun habe, muss ich beurteilen, was das taugt
Beurteilung :
Man berechnet "Pseudo R^2" Bestimmtheitsmaße.
Diese sind immer zwischen 0 und 1. (0 = kein Zusammenhang)

Diese kann man auf 30 verschiedenen Arten berechnen, wir betrachten jedoch 3 . (Folie)

McFadden´s R^2 :

Nullmodell kennt keine unabhängigen Variablen, sondern nur eine Konstante.
P (Yi=1) = 1 / (a + e hoch -b0)
Man benutzt es als Vergleichsmodell zu dem Vollständigen Modell.
McFadden rechnet nun einfach :
R^2 = 1 - (Vollständiges Modell / Nullmodell)
Das taugt nicht viel. Es liegt zwischen 0 und 1, aber man bekommt meistens kleine Werte.
Das Ergebnis ist schon okay, wenn es zwischen 0,2 und 0,4 liegt.

Der beste ist der Nagelkerkers R^2 ! (Folie)

Der Likelihood Ratio Test

Man testet nun einfach, ob das Vollständige und das Nullmodell signifikant zusammenhängen.
Diese Größe ist Chi² Verteilt.
Für die Freiheitsgrade berechnet man die Differenz der Paramter der beiden Modelle.
Diese ist meist die Anzahl der unabhängigen Variablen. !
Also ist die Zahl der unabhängigen Variablen = Anzahl Freiheitsgrade.

So ähnlich wie beim F-Test.
Methodik ähnelt jetzt der linearen Regression.

- Wenn ich nun wissen will welche Variable welchen Beitrag leistet,
kann man direkt die Beta Werte interpretieren. (Wald Test- entspricht in etwa dem t-Test)
(Signifikanz der bj feststellen)
- Wirkungsrichung ermitteln.

Abhängige Variable wird immer in 0 und 1 umgewandelt. (Internal Value)
SPSS berechnet immer die Wahrscheinlichkeiten für die Gruppe die intern als 1 gekennzeichnet ist!

-2 facher Log Likelihood = Der erste ist immer der des Nullmodells. Nur mit der Konstanten.
Danach erst werden die Variablen einbezogen.
Der zweite Log Likelihood den SPSS ausgibt ist dann der des Vollmodells !
Dann zeigt SPSS auch den Nagelkerke R^2.

Chi² Signifikanz-Test
Differenz der beiden LL ist ungefähr 7, und hat 2 Freiheitsgrade (wie unabh. Variablen).
Irrtuimswahrscheinlichkeit, dass wir H0 ablehnen obwohl sie richtig ist = 0,0351.
Da wir mit 0,05 rechnen, können wir die Nullhypothese ablehnen. Es besteht also ein Zusammenhang zwischen (x1,x2) und der Gruppe.

Weil man die Wahrscheinluchkeiten wissen will, interessiert einen, wieviel % der vorhergesagten Gruppenzugehörigkeiten denn auch wirklich richtig vorhergesagt wurden.
Das ist in diesem Fall wurden 7 Fälle von A auch richtig geschätzt. Man hat da die Tabelle "Observed" und "Predicted", wo man sieht wieviele übereinstimmen.

Jetzt muss man fragen, an welcher der Variablen die Signifikanz hängt:
----Variables in the Equatation
- Die "B" s sind die Koeffizienten der logistischen Regression ! also b1 = 0,3429.
- S.E. = Standartfehler des Koeffizienten (brauchen wir nicht, wäre für Walls-Statistik)
- Wald Statistik ist wichtig, H0 = das wahre B = 0, wir versuchen die H0 abzulehnen und schauen bei Sig.=Signifikanz = die Wahrscheinlichkeit mit der wir H0 ablehnen können. Hier ist x2 Signifikanz, weil die Signifikanz kleiner als 0,05 ist (0,0457).
R =
Exp(B) = brauchen wir beides nicht, wir nehmen nur bis zur Signifikanz !

Die Variable Gruppe hängt also nur signifikant von der Variablen x2 ab.
In welche Richtung, lässt sich durch das Vorzeichen von "B" erkennen.

Ausgabe in der Tabelle ; dort kommen 2 neue Spalten :
pre_1 = Wahrscheinlichkeit
pgr_1 = Aus der Wahrscheinlichkeit errechnete Gruppenzugehörigkeit
Weil wir immer P von der Zweiten Gruppe berechnen, haben wir Zugehörigkeit zu Gruppe 2, wenn die Wahrscheinlichkeit größer als 50% ist.

In der Chance stecken dann wieder die berechneten "B"s : e hoch (b0+b1x1....)

Clusteranalyse

Clusteranalyse hat nichts mit Abhängigkeiten und Signifikanzen zu tun,
ist aber ein aktuelles DataMinung verfahren, wenn man fast garnichts über die Daten weiss.
Es gibt nicht "das" Clusteranalyseverfahren, sondern es ist ein Oberbegriff von vielen Varianten.

Wesentlich ist, dass sie nicht auf Datenmatrizen aufsetzt, sondern auf Distanzenmatrix aufbaut.
Man kann von einer Daten- auf eine Distanzmatrix kommen.

Anwendungsbeispiele sind auswahl von Testmärkten mit verschiedenen Merkmalen und guckt dann, welche der Märkte ähnlich sind (ein Cluster bilden). Allgemein möchte man immer aus einer Gruppe von Dingen mit bestimmten Merkmalen kleinere Gruppen herausfinden, die in sich homogen sind, untereinander jedoch möglichst heterogen.

Es gibt also keine abhängigen Variablen, von der man vermutet, dass unabhängige Variablen einen Einfluss auf sie haben. Man unterteilt die Variablen also nicht.
Von diesen will man mindestens 2 gleichartige Gruppierungen herausfinden (Cluster).
Es wird mehr auf die Zeilen (Inhalte) als auf die Spalten (Variablen) geguckt. Es wird geguckt welche Zeilen gut zusammenpassen und bildet daraus dann Gruppen.

3 Merkmale zur Unterscheidung der Verfahren :

1- **Proximitätsmaß** = Unähnlichkeiten oder Ähnlichkeiten der Objekte. Welches Skalenniveau haben die Variablen ?
Für uns wesentlich ist die "**quadrierte euklidsche Distanz**".

2- **Art der Partition** = Wie sind die Gruppen beschaffen die ich herausbekomme ?
Am einfachsten ist die nichtüberlappende Partition, also jedes Miglied ist in genau einem Cluster.
Es kann aber auch sein, dass sie sich überlappen, dass also einer Mitglied in mehreren Clustern sein.
Fuzzy Partition heisst, dass man nur mit einer bestimmten Wahrscheinlichkeit zu einem Cluster gehört. Also z.b. mit 0,5 zu Cluster 1 und 0,5 zu Cluster 2.
Wir werden meistens die nichtüberlappende benutzen , weil diese einfach zu interpretieren sind.
Das ist aber nicht immer realitätsnah.

3- **Fusionierungsalgorithmus** =
hierarchisch-agglomertaive Verfahren .. Man geht davon aus, dass jedes Element der Matrix genau ein Cluster bildet. Dann werden schritt für schritt die Elemente vereinigt, bis am Ende dann ein einziges Cluster übrigbleibt. Dann ist die Frage, in welchem Schritt man den Algorithmus stoppt. (bottom-up) Mehr als 200 Datensätze sind nicht so sinnvoll.
iterativ-Minimaldistanz Verfahren .. Man geht davon aus, dass alle Elemente am Anfang ein großes Cluster sind, welche dann in mehrere kleine Cluster aufgespalten werden. Die Fragen ist auch hier die Frage nach der Anzahl der Cluster die man haben will. (top-down) K-MEANS. Hier kann man mehrere Millionen Datensätz verarbeiten. Das Ergebnis ist aber nicht so genau, da es von Startwerten abhängt, die man erstmal aussuchen muss (?)

Man kann hier nicht mehr gewiss sagen, welche Variable nun Einfluss hat und was nicht.
Tabellenzettel gucken, wie eine Distanzmatrix aussieht, auf die man aus der Datenmatrix

kommen muss.

Wie kommt man von der Datenmatrix auf die Distanzmatrix ?
Beispiel dazu auf den Folien (Margarine Marken).

quadrierte euklidsche Distanz

Man will die Distanz zwischen 2 Punkten berechnen.
P1 (X1, Y1)
P2 (X2, Y2) X und Y sind dann die Variablen.

Wenn man dann in einem Diagramm die beiden Punkte einzeichnet.
Die Distanz ist dann der direkte weg von P1 zu P2.
(Es gibt auch die City-block Methode, die erst nach unten, dann nach rechts geht)
Wenn man nun weiss, dass es dort einen rechten Winkel gibt, kann man den Satz des Pytagoras anwenden. Also muss man den Abstand $(X1 - X2)^2$ plus den Abstand $(Y1 - Y2)^2$ quadrieren um das Quadrat der Distanz zu bekommen.
Dies ist die quadrierte euklidsche Distanz.

Man rechnet also aus der Datenmatrix die jeweiligen Distanzen der Werte der Variablen und quadriert sie jeweils und summiert die Quadrate auf. Dies geht dann auch für mehr als 2.

Man kann bei dieser Methode ganz viele Merkmale haben, bei wenigen Objekte, wie zum Beispiel Bewertung der DAX Unternehmen anhand von hunderten von Indikatoren.

Schritte bei der Clusteranalyse

1. Clustervaerfahren festlegen
2. Variablen kodieren oder skalieren. Man muss also Variable Einkommen mit der Variable Anzahl der Kinder auf einen nenner bringen, da man ja bei einer Distanz von 500€ zu 1500€ mit der Distanz von 1 Kind und 2 Kinder viel extremere Werte bekommt, mit denen man nicht mehr rechnen kann. Sie müssen also auf einen Skalenbereich normiert werden.
Bei nominalskalierten Variablen treten Probleme auf, da man sie in geeignete numerische Werte überführen muss.
Dies wird bei Regressionsverfahren intern skaliert, hier jedoch nicht !
3. Clusternalyse durchführen. Auf den Knopf drücken.
4. Anzahl der Cluster bestimmen. Dies ist das schweste.
5. Cluster interpretieren. Man will ja wissen, wie die Cluster charakterisiert sind ? Was steckt nun hinter CLuster 1 oder 2 ?
Man kann zum Beispiel für jedes CLuster die mittelwerte einzelner Variablen berechnen, z.b. Einkommen, dann hat man in Cluster 1 geringes Einkommen.
(Deskriptive F-Statistik - große F werte = großer einfluss auf variable, das ist aber nicht sinnvoll, da es ja keine Zufallsvariable ist)
Darüber hinaus kann man auch wieder die logistische Regression anwenden !
Hierbei werden z.B. die Clustergrößen als abhängige Variable gestellt und mit unabhängigen Variablen in Verbindung gesetzt.

Man kann diese Cluster auch gut grafisch darstellen und damit herumschustern ☺

hierarchische Clusteranalyseverfahren (Single-Linkage Verfahren)

In SPSS heisst das "nächstgelegener Nachbar" ("nearest neigbour").
Im Output steht dann trotzdem wieder Single Linkage.
Distanzmatrix heisst in SPSS Näherungsmatrix.

Beispiel mit Margarine wieder (Arbeit ab jetzt anhand der Folien)

Proximitätsmaß ist die euklidsche Distanz.
Partition ist nicht überlappend.
Fusionierungsalgorithmus ist der, dass man in der Distanzmatrix guckt, was die geringste Distanz hat. Diese beiden bilden dann ein 2er Cluster und guckt dann als nächstes was die nächste Distanz hat.

Aus der Näherungsmatrix die SPSS liefert, werden nun die beiden geringsten Werte gesucht.
In dem Beispiel ist es die 1,099. Da sind dann die beiden eintsprechenden Objekte ein Cluster.
Zweiter Schritt wäre dann zu gucken, wo die nächst geringere Distanz ist.

In der Zuordnungsübersicht sieht man dann :
Schritt 1 : Cluster 1 (Objekt 3) und Cluster 2 (Objekt 9) mit dem kleinsten Abstand zu Objekt 3.
In den Schritten 1 bis 6 kommen immer neue Objekte zu dem Cluster hinzu.
In Schritt 7 beginnt nun ein neues echtes Cluster !!!
In Schritt 8 wird dann nochmal das erste Cluster nochmal um 7 erweitert.
Der letzte Schritt 10 ist immer die Zusammenführung von den eben gebildeten Clustern.

Weil 1 bereits in einem Cluster ist, ist der Koeffizient in Schritt 6 nicht 3,792 wie es von 1 auf 2 wäre, sondern ... Veranschaulichung in dem Baumdiagramm...

Man kann dann bei der Auswertung auch noch eines Ausschließen (Delikado) , wenn es starke Unterschiede in der Distanz gibt. Es also als Ausreißer betrachten.

- Wenn man die Anzahl der Cluster bestimmen will, muss man gucken, bei welchen Schritten die Distanz zum nächst höheren Cluster auffällig groß wird !!!
Man zieht also einen senkrechten Schnitt, dort wo die Distanzen groß sind, und zählt dann einfach die Anzahl der Linien , die noch gegen den senkrechten Schnitt gehen.
(Dann gibt es 3 Cluster, von denen ein Cluster Delicado ist, was ein Ausreißer ist.)
An diesen Strichen hängen dann nach links gesehen die ganzen anderen kleineren Cluster.
Hier kommt dann eine Art Willkür mit rein.

Man kann diesen Schnitt auch anhand der Koeffizienten sehen, man macht den Schnitt dann jeweils da, wo die Koeffizienten sprunghaft ansteigen.
Also da wo sich beim Koeffizienten nicht mehr viel ändert, hat man wahrscheinlich Cluster.

Befehl : Statistik - Klassifizieren - hierarchische Cluster
Alle Eigenschaftsvariablen aktivieren und in die Variablen übernehmen.
Übrig bleibt die Margarinenmarke, die man dann als Fallbeschschriftung benutzt.
Bei "Statistik" aktivieren wir "Distanzmatrix" und Zuordnungübersicht. Eiszapfen deaktivieren.
- Je mehr Fälle, desto größer wir die Matrrix. Bei 10 Fällen = 10 X 10 Matrix.
Unter "Methode" die "Nächsthelegener Nachbar" aktivieren.

Unter "Diagramm" muss Dendrogramm zur Ausgabe des Diagramm aktiviert werden.
Dazu den Quadrierten Euklidschen Abstand.

Das Ergebnis ist dann ein Diagramm wie in den Folien letztes mal .
Dieses Single Linkage Verfahren wird vor allem dazu genutzt, Ausreisser zu identifizieren.
Man kann dann einfach die Spalte "Delicado" löschen.

Besser als Single Linkage Verfahren, ist das

Ward Verfahren

Zielkriterium sind Fehlerquadrate. (Fehlerquadratsumme über alle Cluster)
Die Abfolge der Schritte ist genauso wie beim Single Linkage Verfahren.
Formel zum berechnen auf Folien.
Jetzt unter "Methode" einfach Ward-Methode auswählen.

1. Jedes einzelne Objekt = 1 Cluster. Hier ist die Fehlerquadratsumme = 0.

Bei der Ausgabe sind unter "Koeffizienten" die Fehlerquadrate aufgeführt.
Diese wird bei jedem Schritt größer.

Clusterzugehörigkeit ausgeben
1. Hierarchische Clusteranalyse
2. Speichern - Einzelne Lösung : 2 Cluster - weiter
3. Einfügen - Unterbefehl /SAVE CLUSTER(2).
4. Ergebnis ist, eine neue Spalte, die die Clusterzugehörigkeit angibt. In SPSS 8 ist diese Funktion fehlerhaft, in späteren sollte es klappen.

Mittelwerte bearbeiten
- Statuistik - Mittelwerte vergleichen - Mittelwerte - Alle Var. bis zur Natürlichkeit als abhängige Variablen übernehmen.
- Ward Methode als unabhängige Vaiable.
- Eine bleibt über, die ignorieren.
- Optionen : Mittelwert, Anzahl der Fälle und ANOVA-Tabelle aktivieren.
- Dann ausführen, und im Ergebnis aus dem Menü "Zeilen Vertauschen" wählen.
(Doppelklick auf ANOVA Tabelle, dann im Menü oben PIVOT - Zeilen vertauschen.

ANOVA Tabelle liefert über den F-Test, wie stark der Unterschied der einzelnen Variablen innerhalb der beiden Gruppen ist.
Welche Variablen tragen wesentlich zur Unterscheidung der beiden Cluster bei ? Dort wo der F-Wert in der Tabelle am größten ist, ist die Variable am einflussreichsten.

Tabelle - Bericht :
Streichfähigkeit mit 5,086 liegt bei Margarine (Cluster 1) überdurchschnittlich bewertet über dem gesamten Mittelwert von 4,7633.
In Cluster 2 ist die Streichfähigkeit mit 3,472 unter dem Durschnitt bewertet (4,7633 Mittelwert).

Klausurwiederholung

Lineare Regression

Frage :
Hängt die Variable X von den anderen Variablen ab ?

Antwort :
Modellzusammenfassung ist wesentlich.

Güte des Modells :
R und R^2. Gibt den Anteil der erklärten Streuung zur Gesamtstreuung an. Je größer der Wert (0-1) , desto mehr hängt es ab. 0,84 heisst, dass 84% der Variationen werden erklärt.

Signifikanz des Modells :
Nullhypothese ist, dass alle Koeffizienten = 0 sind, also dass kein Zusammenhang zwischen den abh. und unabh. Variablen besteht.
Wir messen das mit dem F-Test.
Wenn die Signifikanz größer als (z.B.) 0,05 ist, lehnen wir die Nullhypothese ab.
F-Wert wird berechnet, wenn man die Mittelwertquadrate durcheinander teilt.
F-Wertist nur in weniger als 1% der Fälle geringer als 1.
Wenn der empirische F-Wert größer ist als der theoretische, dann lehnen wie die Nullhypothese ab.
Dann ist die Wahrscheinlichkeit, dass wir sie ablehnen, obwohl sie richtig ist, beträgt 0,05.
SPSS gibt gibt als "Signifikanz" die Irrtumswahrscheinlich, mit der man die Nullhypothese ablehnt.
Wir akzeptieren in der Spalte alles was kleiner als 0,05 ist.
Das sagt uns dann, dass mindestens eine der Variablen Einfluss auf die abhängige hat.
(Mindestens ein Koeffizient ist ungleich Null)

Welche Variablen sind signifikant :
Hierfür haben wir den t-Test. Die Nullhypothese ist hier, dass es keinen Zusammenhang gibt, also dass die Koeffizienten wieder = 0 sind.
Entscheidungsregel ist auch wieder so wie beim F-Test.
SPSS gibt emirische T-Werte aus und vergleichen diese mit theoretischen T-Werten.
Faustregel : Wenn der T-Wert größer als 2 ist, ist es Signifikant.
Wir bekommen aber direkt die Signifikanz des T-Wertes für jede Variable in der nächsten Spalte angezeigt. Wenn die Signifikanz wieder kleiner als 0,05 ist, hat die Variable Einfluss.
- Den größten Einfluss hat dann die Variable mit dem größten Beta-Wert (Spalte davor).
An dem Vorzeichen des Beta Wertes erkennt man, ob ein positiver oder negativer Zusammenhang besteht.
- Toleranz ist 1-Bestimmtheitsmaß($1-R^2$). Das R^2 berechnet man, wenn man jeweile eine Regressionsanalyse mit den unabhängigen Variablen untereinader durchführt (ohne die abhängige).
Dadurch stellt man fest, ob sich die unabhängigen Variablen untereinander erklären.
Wenn die Toleranz viel mehr als 1 ist , kann miltikollinearität vorliegen.
- In der Spalte "B" hat man dann die direkten Werte zu den Variablen, also für die Konstante und die ganzen bs zu den Variablen.

Logistische Regression

unabhängige Variable = metrisch
abhängige Variable = Nominalskalierte Variable
Wir beschränken und auf Binäre Variablen die immer nur 2 Zustände einnimmt.
In Klausur kommt die neue Variante der Ausgabe (neue SPSS Version) :
Signifikanz des Modells :
Chi² : empirisch Ermittelt.
df : Anzahl der Freiheitsgrade.
Sig. : Der empirische Chi² Wert, wird nur in 0,003 der Fälle erreicht (Sig.-Wert). Die Irrtumswahrshceinlichkeit liegt also bei 0,3 %.
- Wie kommt man zu dem Chi² Wert ? -2 LogLikelihood des Nullmodells - -2 LogLikelihood des Vollmodells. Likelihood wissen, aber nicht berechnen oder so. Ungefähr verstehen.
Welche der Variablen tragen nun signifikant bei :
Dann wieder schauen welche der Variablen Sig. kleiner als 0,05 ist.
Wald Test : Je kleiner die Sig. , desto größer ist der Wald Wert.
Modell der logistischen Regression :
Y=1 = zum Beipsiel "ist leser". Er gibt an, welche Ausprägung 1 und 0 ist.
z = (Regressionskoeffizient B der Konstante) +/-
Summe : (Regressionskoeffizient B der Variable * Wert der Variable) -

P (Y=1) = 1 / (1+e hoch -z)

Wenn man jetzt die Vorzeichen beachtet, muss man sehen wie die negativen Koeffizienten B sich auf z auswirken, und dann auf die e-Funktion.
Wenn z gegen unendlich geht = P geht gegen 1.
Wenn z gegen minus unendlich geht = P gegen Null.

Wenn der Regressionskoeffizient also ein negatives Vorzeichen hat, gilt ein umgekehrter Zusammenhang. Das reicht auch für die Klausur zu wissen.

Clusteranalyse

Es ist keine Abhängigkeitsanalyse.
Quadrierte Euklidsche Distanz :
P - 25 - 40
J - 24 - 47

Distanz zwischen P und J : $(25-24)^2 + (40-47)^2 = 50$

Diese Distanzen fasst man nun zusammen mit der jeweils höheren und kommt dann am Ende dazu, dass alle ein riesen Cluster sind.
In der Clusterdarstellung wählt man dann eine Grenzlinie etc.
3 Merkmale zur Unterscheidung wissen von 2005-06-03.

K-Means : 2 Sachen wissen :
Clusterzentren : 3 Variablen und eine Clusterzuordnung. Wie kommt man auf ein Clusterzentrum ? Es gibt soviele Clusterzentren wie es Cluster gibt. Man berechnet die Durchschnittswerte der Variablen 1 im Cluster 1. Und das für jede Variable in jedem Cluster. - Rechnen mit Taschenrechner.
Distanzen : Distanz jedes Objektes zum Clusterzentrum berechnen.
Auf Folie gucken, wie man die 5 Schritte der K-Means Clusteranalyse macht.
Wichtigste Variable ist die mit dem höchsten F-Wert.
Letzte Tabelle zeigt an, welche Ausprägungen in welchem Cluster zusammengefasst sind.