

# 1. Sitzung: Einführung, Messniveau, Skalierung

## Qualitative und quantitative Forschung:

In den Sozialwissenschaften sind eindeutige Messungen schwieriger.

- Oft muss mit “qualitativen” (oder **kategorischen**) Variablen vorlieb genommen werden.
- Oft liegt nur eine kleine Anzahl Fälle (z.B. <10) vor, was die statistische Auswertung behindert

Es werden aber auch fundamentale Unterschiede betont:

- *quantitative* Forschung z.B. möchte die Effekte von Veränderungen oder Eingriffen **erklären** (effects-of-causes);
- *qualitative* Forschung vielmehr die Gründe von Phänomenen **verstehen** (causes-of-effects)

## Mess- bzw. Skalenniveau

*Zivilstand (nominales Messniveau)*: ledig, verheiratet, geschieden, verwitwet

*Intelligenztest (Intervallskala)*: Werte von 40-120

*EU-Staaten gemäss Finanzausgleich (ordinales Messniveau)*: Nettoempfänger, ausgeglichen, Nettozahler

*Wähleranteil einer Partei (Ratioskala)*: 0.1% bis 32%

## Messungen, Indikatoren, Indizes und Skalen

- Theorien -> Hypothesen
- Konzepte -> Dimensionen
- Operationalisierung -> Messungen
- Messvalidität und Reliabilität erfordern bei schwer zugänglichen Konzepten oft, dass mehrere Variablen zur Messung derselben Dimension erhoben werden. Diese müssen dann irgendwie in einer neuen Variablen zusammengefasst werden, welche dann das Konzept adäquat misst.

**Indizes** werden durch eine beliebige Kombination verschiedener Variablen gebildet und messen i.d.R. (aber nicht zwingend) mehrdimensionale Konzepte.

**Skalen** werden ebenfalls durch Kombination einzelner Variablen gebildet, dürfen aber zwingend nur eine einzige Dimension messen.

## Indexkonstruktion

Einen Index bilden bedeutet oft auch den Merkmalsraum reduzieren. Ein Index kann aus kategorischen und kontinuierlichen Variablen konstruiert werden.

**Kategorielle Variablen** als Basis (nominale und ordinale Skalenniveaus):

Kombination: *additiv* (hinreichende Bedingung), gewichtet additiv, multiplikativ (notwendige Bedingung)

**Kontinuierliche Variablen** als Basis (Intervall- und ratioskalen):

- Durchschnittsbildung
- “Standardisierungen”: Vergleichbarmachung durch Bezug auf gemeinsame Basis
- Quotienten von Variablen

## **Zusammenfassung Vorlesung 2 HS08**

### **Beschreibende Statistik**

#### **Begriffe**

Skalierung = Zusammenfassung mehrerer Indikatoren zur Bildung einer eindimensionalen Skala

Items = Einstellungsfragen

Likert Skala = Methode des summierten Rankings    Antworten sollten dieselbe Dimension messen  
(Kontrolle mithilfe des Trennschärfekoeffizienten)

#### **Interpretation und Prüfung der Eindimensionalität**

Der Korrelationskoeffizient misst, wie stark zwei kontinuierliche Variablen miteinander korrelieren. Je höher die Zahl, desto grösser die Korrelation (maximum 1).

#### **Häufigkeitsauszählung bei kategoriellen Variablen**

- Absolute und relative Häufigkeit der Kategorien

#### **Häufigkeitsauszählung bei kontinuierlichen Variablen**

- Da bei kontinuierlichen Variablen i.d.R. jeder Wert nur einmal vorkommt, müssen Kategorien definiert werden.

#### **Fehlende Werte**

- ausschliessen / nicht berücksichtigen  
- umkodieren in fehlende Werte

#### **Grafische Darstellung und Messniveau**

Kategorielle Variablen:

- Säulendiagramm  
- Sektorendiagramm

Kontinuierliche Variablen:

- Histogramm  
- Boxplot (Median, oberes und unteres Quartil, kleinste und grösste Werte, Ausreisser)  
- (Sektordiagramm)

#### **Zentrale Tendenz**

Nominal: Modus/Modalwert (häufigster vorkommender Wert)

Ordinal: Median (in der Mitte liegende Beobachtung bei geordneten Werten, immun gegen Ausreisser)

Intervall/Ratio: Mittelwert (Durchschnitt, anfällig auf Ausreisser)

#### **Streuung**

Ordinal: Interquartilsabstand (75. Quantil – 25. Quantil)

Intervall/Ratio: Standardabweichung (durchschnittliche Abweichung vom Mittelwert)

### Kreuztabellen: Zusammenhangsmasse (3. Vorlesung)

Beschreibende Statistiken: Univariate Analyse

Kreuztabellen: bivariate Analyse

	1. Kat. Unabhängige	2. Kat. Unabhängige	
1. Kat. Abhängige	1. Zelle	2. Zelle	TotalZeile
2. Kat. Abhängige	3. Zelle	4. Zelle	Total Zeile
	Total Spalte	Total Spalte	Total

Mit der Kreuztabelle (Kontingenztafel) werden kausale Zusammenhänge überprüft. Dabei sind die wenigen Variablen nominal oder ordinal skaliert. Die zu erklärende/abhängige Variable wird in den Zeilen dargestellt und die erklärende/unabhängige Variable in den Spalten.

In den Zellen werden neben den absoluten Häufigkeiten, die nur wenig aussagekräftig sind, auch die Spaltenprozentage (d.h. Alle Zellen einer Spalte = 100%) angegeben. Diese Spaltenprozentage können interpretiert werden, wobei ein positiver/negativer Effekt erst ab einem ordinalen Messniveau der Variable möglich ist, da nominale Variablen nicht gewertet werden dürfen (z.B. Mitglied nicht besser als Nicht-Mitglied).

Die Prozentsatzdifferenz (= Differenz 1. Zelle zu 2. Zelle bzw. 3. Zelle zu 4. Zelle) gibt Aufschluss über den Zusammenhang. Das heisst je grösser die Prozentsatzdifferenz desto deutlicher der Zusammenhang. Das Vorzeichen kann dabei nur bei ordinalen Messniveau interpretiert werden (positiver/negativer Effekt).

#### Kreuzproduktverhältnis:

Odds Ratio der 1. Kat. Unabhängige = 1. Zelle / 3. Zelle

Odds Ratio der 2. Kat. Unabhängige = 2. Zelle / 4. Zelle

Das Verhältnis dieser beiden Odds ergibt das Kreuzproduktverhältnis (= Odds Ratio):

Kreuzprodukt =  $(1. \text{Zelle} * 4. \text{Zelle}) / (3. \text{Zelle} * 2. \text{Zelle})$

Wenn das Kreuzproduktverhältnis gleich 1 ist, dann besteht kein Zusammenhang. Je weiter das Kreuzprodukt von 1 abweicht, desto stärker ist der Zusammenhang.

Kreuzprodukt > 1: mehr Häufigkeiten auf Diagonale 1. Zelle zu 4. Zelle

Kreuzprodukt < 1: mehr Häufigkeiten auf Diagonale 3. Zelle zu 2. Zelle

Die Zellenprozentage einer Kreuztabelle entsprechen den gemeinsamen Wahrscheinlichkeiten. Das heisst wir können aussagen wie gross die Eintretenswahrscheinlichkeit eines Merkmales ist. Wenn wir die Variablen unabhängig voneinander interpretieren wollen, bilden wir die Spaltenprozentage (= bedingte Wahrscheinlichkeit). Dabei müssen die bedingten Wahrscheinlichkeiten den Randwahrscheinlichkeiten entsprechen, damit die Unabhängigkeit nicht verletzt wird.

Unabhängigkeit bedeutet, dass kein Zusammenhang zwischen den Merkmalen besteht. Somit ist  $1. \text{Zellenprozent} = 2. \text{Zellenprozent} = \text{Randwahrscheinlichkeit}$ . Diese Prozentangaben entsprechen den erwarteten Häufigkeiten, wenn kein Zusammenhang besteht.

#### Exkurs

Variablen aus fehleranfälligen Beobachtungen sollten als Zufallsvariablen interpretiert werden. Das heisst jeder Beobachtung kann eine Eintretenswahrscheinlichkeit zugeordnet werden (Eintretenswahrscheinlichkeit = relative Häufigkeit eines Wertes bei unendlich grosser Stichprobe). Eine bedingte Wahrscheinlichkeit gibt die Wahrscheinlichkeit an, dass eine bestimmte Ausprägung von y eintritt, wenn die Unabhängige eine bestimmte Ausprägung hat (z.B. Ausprägung von y=4, wenn Ausprägung von x bereits=2 ist)

#### $\chi^2$

Der Zusammenhang zwischen den beiden Variablen wird umso grösser sein, desto höher die Differenz zwischen den erwarteten und den beobachteten Häufigkeiten ist. Das  $\chi^2$  summiert diese

quadrierte Differenz und teilt sie durch die erwarteten Häufigkeiten. Quadriert wird, damit die sich die Abweichungen nicht gleich null ergeben und die grossen Abweichungen stärker gewichtet werden.

Die Freiheitsgrade definierten wie viele zellhäufigkeiten variieren können ohne dass die Radverteilungen sich ändern. Generell weisst  $\chi^2$  (Anzahl Zeile-1)\*(Anzahl Spalte-1) Freiheitsgrade auf.

Das  $\chi^2$  nimmt Werte von 0 bis positiv unendlich an. Je grösser der Werte von  $\chi^2$ , desto stärker der Zusammenhang. Doch der Wert hängt unter anderem von der Anzahl Beobachtungen ab. Der Kontingenzkoeffizient (C) berücksichtigt die Anzahl Beobachtungen. Dabei ist der maximale Wert von C abhängig vom der kleiner Dimension(=Anzahl Spalten/Zeile) der Tabelle. Cramer's V basiert direkt auf der kleineren Dimension und ist deshalb geeignet für Tabellen mit ungleichen Anzahl Zeilen und Spalten. Cramer's V nimmt Werte zwischen 0 und 1 an, wobei grössere Werte für einen stärkeren Zusammenhang stehen.

### **Kreuztabellen: Inferenz, PRE, Kontrollvariable (4. Vorlesung)**

#### Exkurse

Wenn wir eine Stichprobe aus einer Grundgesamtheit ziehen, dann beobachten wir die Häufigkeitsverteilung der Ausprägungen eines Merkmals. Diese Häufigkeitsverteilung entspricht der Grundgesamtheit genauer, wenn die Stichprobe gross ist. Somit kann die Wahrscheinlichkeitsverteilung beschrieben werden (Achtung: Zufallsvariable ist das Merkmal). Wir nehmen an, dass die beobachteten Zellhäufigkeiten normal-verteilte Zufallsvariablen sind, das heisst wir finden bei vielen Stichproben aus derselben Grundgesamtheit ähnliche Zellhäufigkeiten. Weiter kann angenommen werden, dass bestimmte Kennwerte(Mittelwerte, Median, Standardabweichung) bei vielen Stichproben um den wahren Wert streuen und normalverteilt sind. Achtung: Bei dieser Annahme ist nun der Kennwert die Zufallsvariable.

Der  $\chi^2$ -Koeffizient wird aus der Stichprobe berechnet und fällt somit je nach Stichprobe unterschiedlich aus. Der  $\chi^2$ -Koeffizient drückt aus, wie die beobachteten Werte von den erwarteten (=kein Zusammenhang) Werten abweichen. Die Signifikanz des  $\chi^2$ -Koeffizient gibt an, ob sicher von einer Abweichung ausgegangen werden kann.

Die  $\chi^2$ -Verteilung ist die Verteilung von verschiedenen  $\chi^2$ -Koeffizienten aus vielen Stichproben. Unter der Annahme, dass für die Grundgesamtheit kein Zusammenhang existiert, sollte der  $\chi^2$ -Koeffizient gleich 0 sein. Dies ist aber selten der Fall, da Stichproben von der Grundgesamtheit abweichen und bei mehr Freiheitsgrade eher Werte über 0 beobachtet werden.

Die  $\chi^2$ -Verteilung nimmt nur positive Werte an und ist somit stark abhängig von den Freiheitsgraden (Mittelwert=Anzahl Freiheitsgrade, Varianz=Wurzel von 2\*Freiheitsgrade). Die  $\chi^2$ -Verteilung ist glockenförmiger bei vielen Freiheitsgraden. Die Fläche unter der Kurve der  $\chi^2$ -Verteilung entspricht der Wahrscheinlichkeit mit der ein entsprechendes Wertintervall beobachtet wird. Dieses  $\chi^2$ -Wertintervalle kann aufgrund der  $\chi^2$ -Verteilung und der Freiheitsgrade berechnet werden. Der rechte Endpunkt des Intervall bezeichnet den kritischen Wert.  $\chi^2$ -Werte, welche über dem kritischen Wert liegen, können nur zu 5% beobachtet werden. In diesen Fällen besteht ein Zusammenhang in der Grundgesamtheit auf dem 5%-Signifikanzniveau.

#### PRE-Zusammenhangsmasse

Dieses Zusammenhangsmass bringt zum Ausdruck um wie viel die Unabhängige die Abhängige besser erklärt als wenn sie nicht berücksichtigt wird. Die Reduktion der falschen Vorhersagen gilt als bessere Erklärung.

Bei Nichtberücksichtigung der Unabhängigen wird der Modalwert der Abhängigen benutzt um die beste Vorhersage zu erzielen (Summe der Abweichungen vom Modalwert=Fehler1).

Bei Berücksichtigung der Unabhängigen wird die Modalkategorien der Kreuztabelle benutzt um die

beste Vorhersage zu erzielen (Summe der Abweichungen von der Modalkat.=Fehler 2).

$$\lambda = (\text{Fehler1} - \text{Fehler2}) / (\text{Fehler1})$$

$\lambda = 0$  ist auch möglich, wenn ein Zusammenhang in der Tabelle festgestellt wurde. Dies ist der Fall, wenn für alle Kategorien der erklärenden Variablen jeweils dieselbe Kategorie die Moralkategorie ist.

### Konkordanzmasse

Für Variablen auf ordinalem Messniveau kann mit den Konkordanzmassen die Richtung des Effekts heraus gelesen werden können. Konkordanzmasse basieren auf dem Paarvergleich:

-Konkordante Paare: Paare von Subjekten, bei denen ein Subjekt bei beiden Variablen jeweils eine höhere Kategorie aufweist als das andere Subjekt ( $x_i > x_j$  und  $y_i > y_j$ ).

-Diskordante Paare: Paare von Subjekten, bei denen ein Subjekt bei einer Variable jeweils eine höhere Kategorie aufweist als das andere Subjekt ( $x_i > x_j$  und  $y_i < y_j$ ).

$$\text{gamma} = (\text{Konkordante} - \text{Diskordante Paare}) / (\text{Konkordante} + \text{Diskordante Paare})$$

gamma liegt zwischen -1 und +1. gamma=0=kein Zusammenhang

### Kendall's tau

Ta: Die Differenz in der Anzahl konkordanter und diskordanter Paare wird auf die Gesamtzahl der durchzuführenden Paarvergleiche bezogen.

Tb: Sollte nur auf quadratische Kontingenztafeln angewandt werden, da sonst der Wert +1 bzw. -1 nicht erreicht werden kann.

Das Vorzeichen kann interpretiert werden, da die Variablen ordinal skaliert sind.

		Unabhängige Variable	
		Nominal	Ordinal
Abhängige Variable	Nominal	Kontingenzkoeffizient C, Cramer's V, $\lambda$	Kontingenzkoeffizient C, Cramer's V, $\lambda$
	Ordinal	Kontingenzkoeffizient C, Cramer's V, $\lambda$	Ta, Tb

### Beziehungen zwischen kontinuierlichen Variablen

Die besprochenen Zusammenhangsmasse können nicht für kontinuierliche Variable genutzt werden. Lösung: - kontinuierliche Variable in Kategorien zusammengefasst (doch Informationsverlust)

- Korrelationskoeffizient nach Person (nicht für nicht-lineare Zusammenhänge)

Kovarianz: gemeinsame Streuung zweier kontinuierlichen Variablen. Die Kovarianz verschiedener Variablenpaare kann schlecht verglichen werden, da sie massstababhängig ist. Aus diesem Grund wird die Kovarianz durch das Produkt der Standardabweichung geteilt. Das heisst r ist zwischen -1 und +1. Dabei ist r um so grösser, je klarer die Beobachtungen eine Gerade bilden. Dabei sagt r aber **nichts** über die Steigung und die Form (linear, konvex, konkav) der Gerade aus.

### Prüfen einer Skala auf Eindimensionalität

Zur Berechnung des Trennschärfekoeffizienten (zur Prüfen eine Likert-Skala auf Eindimensionalität) greifen wir auf den Korrelationskoeffizient. Dies können wir aber unter der Annahme, dass das Item intervall-skaliert ist. Wie gross der Trennschärfekoeffizient ist, ist nicht definiert. Es sollten aber alle Werte etwa gleich gross sein, ansonsten muss das abweichende Item ausgeschlossen werden.

## Zusammenfassung Methoden Vorlesungen 5+6

### 5. Sitzung: Grundlagen Signifikanztest

(Es wird empfohlen, die Zusammenfassung mit den Folien anzuschauen, da es viele Gleichungen und Graphiken gibt, die nicht hier abgebildet wurden)

Basierend auf einer Zufallsstichprobe ist die Wahrscheinlichkeit einer bestimmten Ausprägung eines Merkmals die relative Häufigkeit dieser Ausprägung für eine sehr grosse Anzahl von Beobachtungen. Die Wahrscheinlichkeitsfunktion bestimmt die Wahrscheinlichkeit der einzelnen Ausprägungen. Mittels der Verteilungsfunktion kann die Wahrscheinlichkeit bestimmter Intervalle von möglichen Ausprägungen bestimmt werden. Es macht Sinn, Wahrscheinlichkeitsverteilungen in diskrete und kontinuierliche Zufallsvariablen zu unterteilen.

Diskret ist eine Zufallsvariable, wenn sie nur abzählbare Werte annimmt (wenn die Ausprägungen von 1 bis n durchnummerierbar sind). Die Verteilungsfunktion kann mit einem Histogramm dargestellt werden (die Fläche der Säulen entspricht der Wahrscheinlichkeit des Auftretens der entsprechenden Ausprägung).

Die Wahrscheinlichkeitsverteilung einer Zufallsvariable lässt sich durch den Erwartungswert (Durchschnitt zu erwartender Werte; **x mitem grade strich obe**) und die Varianz (Streuung der einzelnen Ausprägungen des Merkmals um den Erwartungswert des Merkmals; s-Quadrat) beschreiben.

Die Binomialverteilung (diskret) beschreibt den wahrscheinlichen Ausgang einer Folge von gleichartigen Versuchen, die jeweils nur zwei mögliche Ergebnisse haben.

Eine Zufallsvariable ist stetig, wenn sie alle möglichen Werte eines Kontinuums annimmt.

Die Verteilungsfunktion nicht mehr durch Summierung gebildet werden, sondern nur noch durch Integration der Wahrscheinlichkeitsdichtefunktion (wird auch Dichtefunktion genannt, weil sie die „Dichte“ der stetigen Verteilung an einem bestimmten Punkt angibt).

Erwartungswert: durchschnittlich zu erwartender Wert (gewichtetes arithmetisches Mittel) auf Kontinuum;  $\mu$  (mü), Varianz: Streuung der einzelnen Ausprägungen des Merkmals um den Erwartungswert des Merkmals;  $\sigma$  (sigma) –Quadrat.

Die Normalverteilung (stetig), die glockenförmig ist und bei der kleine Abweichungen vom Erwartungswert häufig, grosse eher selten sind, ist in jedem Fall symmetrisch, eingipflig und nähert sich asymptotisch der x-Achse an. Die Wendepunkte der Verteilung liegen bei  $\mu$  plus/minus  $\sigma$ . Dies gilt auf beiden Seiten.

Der Abstand einer bestimmten Ausprägung von  $\mu$  wird als  $z\sigma$  ausgedrückt (die Formel zur Berechnung der z-Werte ist auf der Seite 11). Die z-Werte für sich gesehen geben dementsprechend den Abstand vom Erwartungswert der Standardnormalverteilung (z-Transformation) an.

Manche Wahrscheinlichkeitsfunktionen sind wichtig, weil sie sich der Verteilung von Daten in der realen Welt gut annähern. Manche sind auch wichtig, weil sie Eigenschaften haben, die sie für die statistische Inferenz besonders geeignet machen. Die Normalverteilung kann auch dann eingesetzt werden, wenn die Daten der Stichprobe nicht glockenförmig verteilt sind.

Es kann auch ein Kennwert (z. B. relative Häufigkeit der Ausprägungen, auch Schätzer genannt) als Zufallsvariable verstanden werden. Für diese kann auch eine

Wahrscheinlichkeitsverteilung aufgestellt werden, wenn man sich vorstellt, dass viele gleich grosse Stichproben aus der Grundgesamtheit genommen und für diese Stichproben jeweils einzeln der Schätzer berechnet wird. Diese Schätzer selbst folgen einer Wahrscheinlichkeitsverteilung, der „sampling distribution“.

Die sampling distribution eines Schätzers (z.B. des Mittelwerts) die auf n Beobachtungen beruht, gibt die relativen Häufigkeiten an, mit denen mögliche Werte dieses Schätzers

auftreten, wenn man den Schätzer je für  $n$  unabhängige Stichproben aus der Grundgesamtheit berechnet (kann empirisch konstruiert oder theoretisch hergeleitet werden). Die Eigenschaften dieser sampling distribution hängen allerdings vom Messniveau der Variablen, von der Verteilung der Grundgesamtheit und von der Stichprobengrösse ab (zu den theoretischen Eigenschaften der sampling distribution siehe Seite 17).

Seiten 21-35 zeigen das oben Erwähnte, das Konfidenzintervall (bei (un-)bekannter Varianz) und den Signifikanztest (in 5 Schritten) anhand eines Beispiels. WICHTIG!

## 6. Sitzung: Bivariate Regressionsanalyse: Einführung

Die klassische lineare Regression erlaubt es, eine Untersuchung der Zusammenhänge zwischen intervall- bzw. ratioskalierten Variablen ohne grossen Informationsverlust (bei der Kreuztabelle problematisch). Der vermutete lineare Zusammenhang kann genau beschrieben werden. Die multiple Regressionsanalyse erlaubt eine effiziente Kontrolle der Störvariablen. Der Korrelationskoeffizient ist ein Zusammenhangsmass für mind. Intervallskalierte Variablen. Unterteilt man eine zweidimensionale Darstellung der Messwertpaare durch die Mittelwerte in Quadranten, so misst das Vorzeichen, ob der Hauptteil der Messwertpaare im 1. und 3. Quadranten (positiv) oder im 2. und 4. Quadranten (negativ) liegt. Die Grösse des Koeffizienten sagt etwas darüber aus, wie stark sich die Messwertpaare (die Punktwolke) einer Geraden angleicht (die Steigung der Geraden wird nicht beschrieben).

Wenn eine lineare Funktion zwischen den Variablen vermutet wird, so bedeutet dies, dass für jeden möglichen Punkt auf der Skala der Erklärenden ein entsprechender Punkt auf der Skala der Abhängigen existiert, der mittels einer linearen Funktion berechnet werden kann.

“Linearität” bedeutet, dass wenn sich die erklärende Variable um eine Einheit vergrössert oder verkleinert, dann hat das immer denselben Effekt auf die abhängige Variable, unabhängig vom Niveau der erklärenden Variable.

Eine lineare Funktion wird durch einen Achsenabschnitt und einen Steigungsparameter beschrieben: Achsenabschnitt  $\alpha$ : wo schneidet die Gerade die y-Achse? Steigung (Beta): um wie viele Einheiten verändert sich die Abhängige wenn die Unabhängige um eine Einheit verändert wird?

Alpha und Beta sind nicht bekannt und müssen deshalb aus den Stichprobendaten geschätzt werden (dazu beachte die Stichprobengleichungen auf Seite 9 und die geschätzten Alpha-/Beta-Dach).

Zunächst wissen wir nichts über die lineare Beziehung. Soll die Gerade aber jene lineare Beziehung beschreiben, die für die Messpunkte besteht, so muss die lineare Funktion in diese Punkte “eingepasst” werden.

Dieses “Einpassen” basiert auf den Abweichungen der beobachteten Messwerte der Abhängigen ( $y_i$ ) von den berechneten bzw. “vorhergesagten” Werten ( $\hat{y}_i$ ), also der Abstand der beobachteten Werte von  $y$  vom entsprechenden Punkt auf der Regressionslinie. Diese Abweichungen nennt man auch Residuen.

Jede Beobachtung hat ein Residuum und intuitiv ist die Schätzung umso besser, je kleiner die Residuen gesamtheitlich sind. Die Grösse der Gesamtheit der Residuen wir beschreiben durch die Summe der quadrierten Abweichungen (SSE: Sum of squared errors; Gleichung auf der Seite 12). Die weiter oben genannten Schätzer Alpha- und Beta-Dach garantieren, dass die Abweichungen kleinstmöglich sind, die Abweichungen zu 0 addieren und die Regressionslinie durch die Mittelwerte der Variablen geht.

Es gibt die bedingte Wahrscheinlichkeitsverteilung die beschreibt, welche Werte  $y$  für welche Werte von  $x$  wahrscheinlich sind, da  $y$  durch  $x$  nie vollständig determiniert wird und deshalb für denselben Wert von  $x$  mehrere Werte von  $y$  wahrscheinlich sind. Die lineare Regressionsfunktion (Gleichung auf der Seite 13) bringt nun die Werte von  $x$  mit dem

Mittelwert dieser bedingten Wahrscheinlichkeitsverteilung für  $y$  in Verbindung. Alpha und Beta werden Regressionskoeffizienten genannt.

Die bedingte Wahrscheinlichkeitsverteilung von  $y$  wird einerseits durch den Mittelwert  $E(y|x)$  beschrieben und durch eine bedingte Standardabweichung (Annahmen für die Kleinst-Quadrate-Regressionsanalyse auf der Seite 4; Normalverteilung, Homoskedastizität etc.).

Eine alternative Schreibweise geht nicht von den Mittelwerten der bedingten Wahrscheinlichkeitsverteilungen aus, sondern von den einzelnen (Populations-)Werten von  $y$ . Diese werden als das Resultat der Addition von Mittelwerten der bedingten Wahrscheinlichkeitsverteilungen und den entsprechenden Abweichungen von der Regressionslinie interpretiert.  $e_i$  (nicht das normale  $E$ ) wird der Fehlerterm genannt und misst die Fehler, die theoretisch zu erwarten sind, weil wir "nur" mit einem Modell der Wirklichkeit operieren (nicht zu verwechseln mit den Residuen).

Die Fehler sind entsprechend die Abweichungen zwischen den Populationswerten und der Regressionslinie.

Weil die Varianz für alle bedingten Wahrscheinlichkeitsverteilungen als konstant angenommen wird, entspricht sie dem bedingten Erwartungswert (dem Mittelwert) der quadrierten Fehler  $e_i$ .  $e_i$  können aber nicht beobachtet werden. Beobachtbar sind nur die Abweichungen der Stichprobenwerte  $\hat{e}_i$ .

Die geschätzte Streuung der Fehler um die Regressionslinie wird in der Regel durch den Standardfehler der Regression ausgedrückt, da dieser in den Einheiten der Abhängigen gemessen wird. Bezüglich der Freiheitsgrade wird dabei oft  $k$  für die Anzahl geschätzter Steigungsparameter geschrieben ( $k > 1$ : multiple Regression). Zusammen mit dem ebenfalls geschätzten Achsenabschnitt sind das dann  $n - k - 1$  Freiheitsgrade für den Schätzer der Varianz bzw. des Standardfehlers. Im bivariaten Fall entspricht das  $n - 2$ . Je kleiner die Streuung der Residuen, desto besser wurde die Linie durch die Punktwolke gelegt. Der Standardfehler der Regression ist demnach auch ein Mass für die Modellgüte.

Ein weiteres Mass für die Modellgüte entspricht dem PRE-Zusammenhangsmass für die Kreuztabelle. Es drückt aus, wie gut wir  $y$  durch Zuhilfenahme von  $x$  "erklären" können (Formel auf der Seite 21).

Beta-Dach darf nie unabhängig von der Masseinheit der involvierten Variablen interpretiert werden. Obwohl die Stärke der Beziehung dieselbe bleibt ändert sich der Steigungsparameter (der Koeffizient Beta-Dach) je nach Masseinheit der verwendeten Variablen.

## VL 7

$\alpha$   $\beta$

Bisher wurde die Varianz nur zur Berechnung des Standardfehlers verwendet. Sie ist aber auch nötig, um die statistische Inferenz zu berechnen.

Für die Überprüfung der statistischen Inferenz der Parameter aus der Regressionsgeraden muss die sampling distribution (Streuung der vielen geschätzten  $\beta$ -Dach um den wahren Wert von  $\beta$ ) der Koeffizienten berechnet werden, denn das geschätzte  $\beta$ -Dach ist für jede Stichprobe ein anderes.

Die sampling distribution berechnet sich aus dem Erwartungswert und der Varianz von  $\beta$ .

Die Varianz entspricht dem Erwartungswert der quadrierten Differenz zwischen geschätztem  $\beta$ -Dach und dem wahren  $\beta$  anhand vieler verschiedener Stichproben.  $\beta$ -Dach setzt sich also aus dem wahren  $\beta$  und einem linearen Fehlerterm zusammen.  $\Rightarrow$  Die Varianz von  $\beta$  ist die Varianz der Fehler geteilt durch die Variation der Unabhängigen.

Die wahre Varianz ist nicht beobachtbar, deshalb wird die geschätzte Varianz verwendet. So kann man die sampling distribution von  $\beta$ -Dach um das wahre  $\beta$  anhand einer einzigen Stichprobe schätzen.

Je grösser die Variation der Unabhängigen, desto kleiner ist die Varianz von  $\beta$ -Dach, desto genauer lässt sich das wahre  $\beta$  also schätzen.

Signifikanztest der Regressionskoeffizienten bedeutet Test auf statistische Unabhängigkeit zwischen  $x$  und  $y$ . Wenn Unabhängigkeit bestünde, wäre  $\beta = 0$ .

Man prüft also, ob  $\beta$  signifikant von 0 abweicht. Dafür nimmt man den Standardfehler von  $\beta$  für einen Hypothesentest. Die Nullhypothese ist also  $H_0: \beta=0$ .

Der Standardfehler hängt linear von der Varianz der Fehler ab. Man geht von normalverteilten Fehlern aus.

Bei einer bekannten Varianz würde die z-Statistik verwendet werden, um eine Standardnormalverteilung für den Abstand zwischen  $\beta$ -Dach und wahren  $\beta$  zu erhalten, indem durch den Standardfehler von  $\beta$  dividiert wird. (Formel: Folie 10) Ein Wert der z-Statistik von mehr als 1.96 würde eine signifikante Abweichung von  $\beta$ -Dach von 0 bedeuten.

Da die Varianz unbekannt ist, wird die t-Statistik verwendet! Die Varianz wird geschätzt. Die sampling distribution der t-Statistik ist entsprechend die t-Verteilung (Vgl. VL 5). Der t-Wert berechnet sich so:  $t = \text{Wert des Koeffizienten, z.B. } \beta / \text{Standardfehler}$ .  $t$  sagt aus, um wie viele Standardfehler der Koeffizient von der  $H_0$  abweicht. Ist ein best. kritischer Wert überschritten, bedeutet das eine signifikante Abweichung von  $H_0$ . Der kritische Wert kann in Tabellen abgelesen werden. Achtung: zweiseitiger Signifikanztest, da  $t$  und  $\beta$  auf beiden Seiten von 0 abweichen können. Also kritische t-Werte für links und rechts in der Kurve je 0.025 beachten für  $\alpha = 0.05$  (sh. Folie 14).

- Die t-Statistik ist symmetrisch um 0 verteilt.
- Ab 8 Freiheitsgraden entspricht sie der Standardnormalverteilung, vorher ist sie etwas gedehnter, der kritische  $\alpha$ -Wert liegt also etwas weiter von 0 entfernt.
- Freiheitsgrade =  $n-k-1$
- Normalverteilte Fehler in der Population müssen angenommen werden.

Konfidenzintervall von geschätztem  $\beta$ -Dach sagt aus, wie weit der 95%-Geltungsbereich von  $\beta$ -Dach von 0 entfernt ist. => Mehr Information als bloss durch Signifikanztest. Wird berechnet durch Wert des  $\beta$ -Dach-Koeffizienten +/- (kritischer t-Wert) \* geschätzte Varianz von  $\beta$ . [Achtung: nicht den t-Wert des Koeffizienten verwenden, sondern den kritischen t-Wert für die entsprechende Anzahl Freiheitsgrade und das verwendete  $\alpha$ .] Man sieht nun, wie weit 0 vom Konfidenzintervall entfernt liegt, ob die festgestellte Signifikanz also gar nicht so eindrücklich ist und evtl. nur durch eine sehr grosse Zahl von Stichproben bedingt wird.

Bei Hypothesentests entspricht das gewählte  $\alpha$  der Wahrscheinlichkeit des Fehlers 1. Art. Je kleiner  $\alpha$  gewählt wird, desto eher kann es zum Fehler 2. Art kommen.

Die ‚Mächtigkeit‘ des Hypothesentests: Power = 1 – p(Fehler 2. Art)

H<sub>0</sub> wird abgelehnt, wenn der Test die H<sub>1</sub> betätigt. Wenn H<sub>1</sub> nicht bestätigt wird, dann wird H<sub>0</sub> „nicht abgelehnt“. H<sub>0</sub> wird nie angenommen!

Immer öfter wird nur der p-Wert kommuniziert, um die Signifikanz auszudrücken. Jedoch wird der p-Wert durch den schrumpfenden Standardfehler bei einer wachsenden Anzahl Stichproben immer kleiner. Daher kann eine Signifikanz des p-Wertes täuschen, indem auch schon sehr kleine Abweichungen von H<sub>0</sub> als signifikant gelten (Vgl. Begründung Konfidenzintervall). => „Praktische“ Signifikanz sollte auch gegeben sein.

## VL8

Der lineare Regressionsmodell  $\alpha + \beta x_i + \varepsilon = y_i$  soll ein linearer Zusammenhang charakterisieren. Dies geschieht grundsätzlich durch die Interpretation des Achsenabschnittes  $\alpha$  und der Steigung  $\beta$ .

- Die Varianz:
  1. Gibt an, wie gut die erzeugte Gerade an den beobachteten Daten angepasst werden konnte.
  2. Welches ist der *Standardfehler* der *sampling distribution* der Regressionskoeffizienten (diese Koeffizienten entsprechen die Steigung und dem Achsenabschnitt)
    - *Exkurs*: Unter *sampling distribution* (Stichprobenverteilung) versteht man die Verteilung einer Schätzgröße bzw. einer Teststatistik unter hypothetischer Wiederholung der Stichprobennahme. Die Verteilung der Schätzgröße, die als Stichprobe aus einer Grundgesamtheit entnommen wird - wenn in Bezug gesetzt zur Verteilung dieser Schätzgröße in der Grundgesamtheit -, dient der Gewinnung von Aussagen über die Ermittlung dieser Schätzgrößen in der Grundgesamtheit aufgrund von Stichproben aus der Grundgesamtheit. Die Fragestellung lautet also: wie kann man von Stichproben auf die Grundgesamtheit zurück schließen. Die Beantwortung dieser Frage zur Genauigkeit/Zuverlässigkeit des Rückschlusses von einer Stichprobe auf die Grundgesamtheit geht über die Stichprobenverteilung und wird über Konfidenzintervalle bzw. p-Werte bemäht
    - Der *Standardfehler* liefert so eine Aussage über die Güte des ermittelten Mittelwertes.  
Wenn  $n$  die Größe der Stichprobe ist und  $\sigma^2$  die Varianz der Grundgesamtheit, so

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

ist der Standardfehler durch folgende Formel gegeben:

Um die lineare Regressionsanalyse durchführen zu können müssen gewisse Annahmen getroffen werden. Werden diese Annahmen nicht getroffen, müsste das lineare Regressionsmodell erweitert werden oder sogar ein neues Modell eingeführt werden (z.B logit Regression)

Die **Annahmen** zur Durchführung einer linearen Regression:

1. **Das Regressionsmodell muss mindestens in den Parameter linear sein.**  
Die Variablen  $x$  und  $y$  müssen nicht linear sein, da man sie z.B. durch Log erweitern kann.
2. **die UV  $x$  und der Fehler  $\varepsilon$  dürfen nicht korrelieren.**
3. **der bedingte Erwartungswert der Fehler muss 0 sein  $E(x_i/\varepsilon)=0$**   
auch der unbedingte Erwartungswert der Fehler muss 0 sein  $E(\varepsilon)=0$

falls der Erwartungswert der Fehler nicht = 0, ist dann existiert eine Verzerrung des Achsenabschnitt.  $\rightarrow E(x_i/\varepsilon)=w \rightarrow \alpha + \beta x_i + w = y_i \rightarrow$  Das heisst dass alle Messwerte systematisch um den selben Wert verzerrt sind.

4. **Homoskedastizität:**  
Residuen-Varianzhomogenität. Das heisst dass die Varianz der Residuen im vergleich zu den Residuen-Varianz der anderen Variablen nicht signifikant unterschiedlich sind.  
Kurz: Die bedingte Verteilung der Fehler ist immer gleich.
  - Ist jedoch die Varianz der Residuen der Variablen unterschiedlich so hat es unterschiedliche Streuung innerhalb einer Datenmessung. Man spricht dann von heteroskedastizität.

**5. Autokorrelation der Fehler muss = 0 sein.**

Das heisst dass z.B die Fehler 2 benachbarten Werte  $x_i$  und  $x_j$  deren Fehler nicht korrelieren, weil sonst  $y_i$  von 2 unterschiedlichen Fehler  $\varepsilon_i$  und  $\varepsilon_j$  beeinflusst wird.

- **Exkurs:** Grundsätzlich spricht man von einer Korrelation, wenn zwischen zwei Variablen ein Zusammenhang besteht. Wird bei Ausprägungen nur eines Merkmals im Zeitablauf ein Zusammenhang der Ergebniswerte beobachtet, spricht man von einer Autokorrelation.

**Bsp:** Die Arbeitslosenstatistik weist im März 3,56 Millionen Arbeitslose aus, im April sind es 3,48 Millionen, im Mai 3,42 Millionen. Diese Werte stehen in einem Zusammenhang. Eine große Anzahl der Arbeitslosen aus dem März ist auch noch im April ohne Arbeit und genauso im Mai – es sind nicht plötzlich 3,x Millionen andere Personen arbeitslos. Daher hängt die Anzahl der Arbeitslosen eines Monats immer mit der Anzahl des Vormonats zusammen – es besteht eine Autokorrelation.

**6. Keine Korrelation zwischen Unabhängiger und Fehlerterm.**

Die Fehler dürfen nicht mit der erklärenden Variablen korreliert sein. Falls die Fehler  $\varepsilon$  systematisch bei steigendem  $x$  auch zunehmen, wird diese Regel verletzt. Wenn ein unbeobachteter Wert  $x_2$  mit  $x$  positiv korreliert und auch  $y$  positiv beeinflusst, dann sind die Fehler im bivariaten Modell für kleine  $x$  zu tief und bei grossen  $x$  zu hoch  $\rightarrow$   $\beta$ -dach würde überschätzt

**OLS Schätzer: Sorry!!! Bitte in der 8.VL auf den Seiten 17-23 nachlesen.**

**7. Die Beobachtungen müssen genügend variieren.**

Also kleine und grosse  $x$  miteinbeziehen und diese sollen auch in  $y$  verschiedene Werte ergeben. Ist dies nicht der Fall, müssen wir davon ausgehen, dass der Schätzer sehr unpräzise ist (Standardsfehler gross).

**8. n muss gross genug sein.** Grosses  $n$  wirkt sich vorteilhaft auf die Schätzer aus.

**9. Keine Fehlspezifikation**

Es sollen irrelevante Unabhängige nicht berücksichtigt werden. Werden nicht-relevante Unabhängige ins Modell aufgenommen, würde in jedem Fall der Standardfehler des Schätzers falsch geschätzt, demnach der Hypothesentest ungültig. Die Konsequenzen einer Verletzung dieser Annahme wird in VL12 genauer angeschaut.

**10. Keine Multikollinearität**

Bei der multiplen Regression sollen die Unabhängigen nicht miteinander korrelieren (vgl. VL12). In der Praxis handelt es sich bei der Multikollinearität meistens um Probleme in der Stichprobe. Entweder zu kleines  $n$  oder Stichprobe ist aus „irgendwelchen Gründen“ verzerrt.

## 9 Sitzung.

### Einführung multiple Regression.

Mit einer multiplen Regression wird eine kausale Beziehung der von mehreren unabhängigen Variablen zur abhängigen Variablen untersucht. Eine Korrelation alleine (anhand von einem Zusammenhangsmass) reicht nicht aus, um eine kausale Beziehung zu begründen. Dazu müssten folgende Bedingungen erfüllt werden.

- Zusammenhangsmass weicht klar von 0 ab, ist statistisch signifikant.
- Ursache muss Wirkung zeitlich voraus gehen (oder zumindest theoretisch).
- Alternative Ursachen (kontrollvariablen) müssen aus dem Zusammenhang eliminiert werden. Dadurch soll die Variation der abhängigen möglichst nur noch durch die Variation der Unabhängigen erklärt werden. (Diekmann: ex post statistische Kontrolle)

Es gibt verschiedene Arten der Korrelation.

- Scheinkorrelation: man vermutet, dass X Y beeinflusst, aber beide Variablen werden vor allem von einer drittvariablen (Z) beeinflusst.
- Intervenierende Variable: X beeinflusst Z, Z beeinflusst Y.
- Gleichzeitig indirekte und direkte Kausalität: in Sozialwissenschaften häufig.
- Interaktion: die Stärke des Zusammenhangs zwischen X und Y hängt vom Wert einer Drittvariablen Z ab. Im Regressionsmodell bildet man einen Interaktionsterm (X mal Z)

Matritzen: eine Art der Darstellung einer Korrelation von Variablen (Vektoren), als Herleitung der Formel für das Regressionsmodell. Matritzen können nur addiert werden, wenn sie dieselben Dimensionen besitzen.

In der Formel:

$$Y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \epsilon_i$$

$\alpha$ : die konstante (im modell y-achsen Abschnitt)

$\beta$ : koeffizient (im modell die Steigung) jeder unabhängigen Variablen zum Y.

$\epsilon_i$ : Fehler.

Anhand dieses Modells kann man vorhergesagte Werte der abhängigen Variablen schätzen. Wichtig für ein multiples Regressionsmodell ist es, dass die unabhängigen Variablen untereinander nicht stark korrelieren, dass die Varianz der Residuen konstant ist (der Mittelwert der Residuen muss nach dieser Annahme 0 betragen) und dass es keine Ausreisser im Modell gibt, welche die Ergebnisse verzerren.

Das Problem des Einflusses verschiedener Variablen und der Multikollinearität kann anhand eines Venn-Diagramms dargestellt werden. Die Mengen (gesamte Variation der jeweiligen Variablen) weisen verschiedene Schnittmengen auf, welche einen Zusammenhang der Variation zweier oder mehrerer Variablen aufzeigen. Es interessiert der Anteil der Variation von Y, welche von X oder Z erklärt wird. Im Gesamtmodell zeigt das R-Quadrat im Grunde genommen, wie gross diese Schnittmengen sind. Aussage R-Quadrat 0.2: 20% der Varianz von Y wird durch die Varianz der Unabhängigen (durch das Modell erklärt). Ein sehr hohes R-Quadrat ist verdächtig und bedeutet oft, dass es eine multikollinearität gibt. Dies kann man

durch alternative Modelle überprüfen, bei denen die übrigen Variablen ganz ausgelassen werden.

Bei einer multiplen Regression wendet man grundsätzlich das standardisierte R-Quadrat an, dieses ist angepasst auf die Anzahl Variablen im Modell.

Die Formel für das gewöhnliche R-Quadrat beträgt:

$$R\text{-Quadrat} = \frac{TSS - SSE}{TSS}$$

Bei mehreren unabhängigen Variablen bleibt der TSS gleich, SSE sinkt hingegen mit der Anzahl Unabhängigen. Dies hat zur Folge, dass das R-Quadrat mit jeder Unabhängigen zunimmt, weshalb es korrigiert werden muss, da man ja ansonsten möglichst viele beliebige Variablen in das Modell aufnehmen könnte, um die Erklärungskraft zu erhöhen (Data-Mining). Das angepasste R-Quadrat ist mit Vorsicht zu interpretieren. Es kann anders als das gewöhnliche R-Quadrat sogar negativ ausfallen. Bei einem Modell mit vielen Beobachtungen und wenig Regressoren ist der Unterschied vom normalen zum angepassten R-Quadrat jedoch klein.

Das R-Quadrat (oder das angepasste) sagt also aus, wie hoch die Erklärungskraft des Modells für die beobachteten Werte ist. Aber Achtung: Die Jagd nach dem höchsten R-Quadrat ist gefährlich! Andere Kenngrößen geben Aufschluss über wichtigere Kriterien, wie die Signifikanz. Ein Vergleich von verschiedenen R-Quadrat-Werten macht nur Sinn, wenn es sich um dieselbe Variable handelt und derselben Stichprobe, aber um „beliebige“ Kombinationen verschiedener unabhängiger Variablen. Nur so macht der Vergleich der verschiedenen Aussagestärken der Modelle Sinn.

In der multiplen Regressionsanalyse wird der Effekt der jeweiligen unabhängigen isoliert von den anderen unabhängigen betrachtet. Der Koeffizient Beta weist also den Zusammenhang zwischen X und Y ceteris paribus auf. In der Rechnung werden die anderen unabhängigen konstant gehalten. Generell sollte die Auswahl der erklärenden und der Kontrollvariablen immer theoretisch begründet werden.

## 10 Sitzung. Multiple Regression.

### 1: Inferenz

### 2: Quadratische Terme

### 3: Interaktion.

1: Inferenz. Signifikanztests.

Es sind für multiple Regressionsmodelle verschiedene Arten von Signifikanztests möglich:

- Für die Signifikanz einzelner Variablen (bzw. deren Unabhängigkeit in der Population) wendet man den **t-test** an (siehe bivariate Regression)
- Der **F-Test** überprüft für die Signifikanz des Gesamtmodells, bzw. die Gemeinsame Unabhängigkeit der Regressionskoeffizienten.
- Der **F-Test** kann auch testen, ob zwei oder mehrere Koeffizienten dieselben sind. (es werden noch weitere genannt, doch wird hier nur auf diese Tests eingegangen.)

Formel für F-Wert:

$$F = \frac{R^2 / k}{\frac{1-R^2}{n-k-1}}$$

$n-k-1$  sind die Freiheitsgrade des Modells.

Man testet mit dem F-test die Nullhypothese in der Population.

$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$  oder:  $R^2 = 0$

Mindestens ein  $\beta$  muss nicht gleich null sein. Die Signifikanz ist ab einem bestimmten Schwellenwert von F bei einer bestimmten Anzahl Freiheitsgrade erreicht. Wenn der F-Wert den Schwellenwert deutlich überschreitet, ist der p-Wert entsprechend tief.

Der F-Test kann auch für Gruppen von geschätzten Koeffizienten überprüfen, ob sie gemeinsam Null sind. Dies macht man vor allem bei dichotomen unabhängigen Variablen, welche zusammen eine kategorielle Variable bilden (dummys) oder Variablen, welche eine hohe Multikollinearität aufweisen (oder wo man das vermutet)

Man vergleicht dazu das ursprüngliche Modell der AV und UVs mit einem restringiertes Modell, bei dem man gewisse Unabhängige entfernt hat. Es werden im Grunde genommen die F-Werte für beide Modelle berechnet und verglichen.

### Quadratische Terme

Quadratische Terme wendet man an, wenn man einen nicht-linearen effekt vermutet. Das heisst also, dass der Effekt einer UV mit steigendem Wert der UV abnimmt und an einer bestimmten Schwelle negativ wird. Bei einer Regression mit quadratischem Term müssen immer beide Terme in der Gleichung enthalten sein also der einfache Term und der quadrierte. Wenn keine Signifikanz für den quadratischen Term vorliegt, kann man von einem einfachen linearen Zusammenhang ausgehen.

## **Interaktionen**

Der Effekt der Unabhängigen auf die Variable (beta-Koeffizient) hängt vom Wert einer anderen Variable ab. Man bildet zusätzlich einen Interaktionsterm.

$$Y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

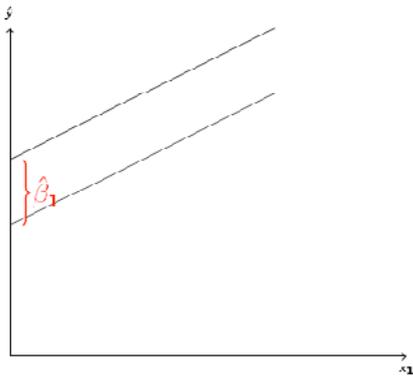
Bei einer Interaktion mit dichotomer unabhängiger Variable bedeutet dies, dass es zwei Regressionsgeraden gibt, eine für die Ausprägung 0 und für die Ausprägung 1. Bei kontinuierlichen unabhängigen ist die Interpretation etwas komplizierter, man muss hier die Änderung des Koeffizienten pro Zunahme der UV um eine Einheit messen. Die Koeffizienten sind nicht wie bei OLS-Regression als unmittelbarer Effekt, sondern als konditionale Effekte zu verstehen (dh bei gewisser Ausprägung der intervenierenden Variablen)

Für Interaktionsterme und Quadratische Terme müssen immer theoretische Argumente vorhanden sein.

## Vorlesung 11: „Dummies“, Modellspezifikation, Residuen

Indikatorvariable oder Dummy-Variable: hat nur die dichotome Ausprägung 0 und 1 (z.B. Geschlecht) → wird auf nominalem Skalenniveau gemessen

Wird die Dummyvariable in einem (bivariaten) Regressionsmodell verwendet, lässt sich das Ergebnis über zwei parallele Regressionsgeraden darstellen, wobei die eine um  $\beta_1$  höher liegt als die andere:



Will eine nominale Variable in einer multiplen Regression verwendet werden, so muss sie in mehrere Dummies aufgeteilt werden: Für jede Kategorie wird eine Variable generiert, falls eine bestimmte Beobachtung auftritt, liegt der Wert bei 1, falls sie nicht auftritt bei 0.

Diese neugenerierten Dummy-Variablen weisen *perfekte Multikollinearität* auf (weil sie in einer direkten linearen Beziehung stehen). So können sie nicht alle zusammen in einem Regressionsmodell verwendet werden → es wird eine Referenzkategorie definiert, welche weggelassen wird. Die Konstante entspricht dann dem Effekt der Weggelassenen, *aber nur wenn sonst keine Variablen ausser den Kategorie-Dummies im Modell sind. (ANOVA Modell)*

Die Koeffizienten der Verwendeten dürfen nur *relativ* zur Referenzkategorie interpretiert werden: Im Vergleich zur Referenzkategorie hat eine verwendete Kategorie *im Mittel* einen höheren bzw. tieferen Wert.

Ziel ANOVA: Mittelwerte von Gruppen vergleichen und einen Signifikanztest bezüglich deren Unterschiedlichkeit bereitstellen. Falls in das Regressionsmodell auch noch eine kontinuierliche Variable miteinbezogen wird, nennt sich dies ANCOVA.

Mögliche Fehler in der Modellspezifikation:

- *Relevante Variable* wird nicht aufgenommen: Falls  $x_1$  mit  $x_2$  korreliert und  $x_2$  wird nicht miteinbezogen, so wird  $\beta_1$  verzerrt und inkonsistent sein.

- *Irrelevante Variable* wird aufgenommen: Irrelevant=Variable hat wahren Koeffizienten von 0. OLS-Schätzer bleibt unverzerrt und konsistent, aber die Varianz des Schätzers wird grösser und so ist er weniger effizient.

- Eine *falsch funktionale Form* wird verwendet: z.B. sollte ein quadratischer Term aufgenommen werden.

- Variablen (AV od. UV) weisen *Messfehler* auf: AV: OLS-Schätzer bleiben unverzerrt, falls Messfehler nicht mit UVs korrelieren und solange sie sich im Durchschnitt ausgleichen. Varianz wird aber wiederum grösser. UV: Annahme der Nicht-Korrelation zwischen Fehlertermen und UV wird verletzt → OLS-Schätzer verzerrt und inkonsistent.

#### Mögliche Probleme bei Regressionsanalysen:

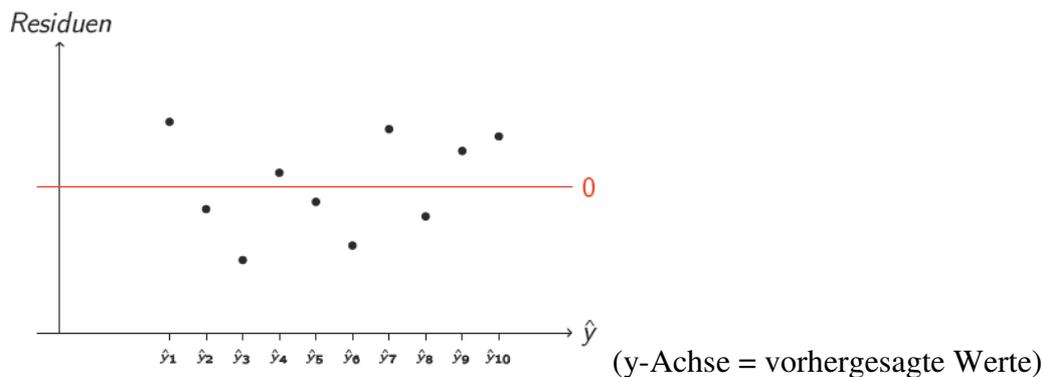
- *Ausreisser*: Beobachtungen, mit einem für den Wert der UV ungewöhnlichen Wert für die AV. Beobachtungen mit grossen Residuen.
- *Einflussreiche Beobachtungen*: liegen weit weg von den übrigen, beeinflussen so die Lage der Regressionslinie. Können grosse Residuen haben und /oder grossen Abstand vom Mittelwert der UV.
- *(Multi-)Kollinearität*: zwei oder mehrere Variablen korrelieren (zu) stark.
- *Heteroskedasizität*: Fehlervarianz bleibt nicht dieselbe über die Werte der UV hinweg.
- *Autokorrelation*: siehe Vorlesung 11, FS 09.

Erwartete (Normal)Verteilung der Residuen: Residuen streuen um 0, Streuung ist glockenförmig, Streuung für div. Werte der UV gleichmässig, Residuen folgen keinem Muster (→ nicht untereinander abhängig), Residuen korrelieren nicht mit der UV.

Die Verteilung kann in einem Histogramm erkannt werden. Mit den *standardisierten* Residuen kann der Verdacht auf Ausreisser grösser werden; Beobachtungen mit einem standardisierten Residuum von grösser 2 sind verdächtig. (Residuen ÷ Standardabweichung = Standard-Normalverteilung.)

## Vorlesung 12: Regressionsdiagnostik

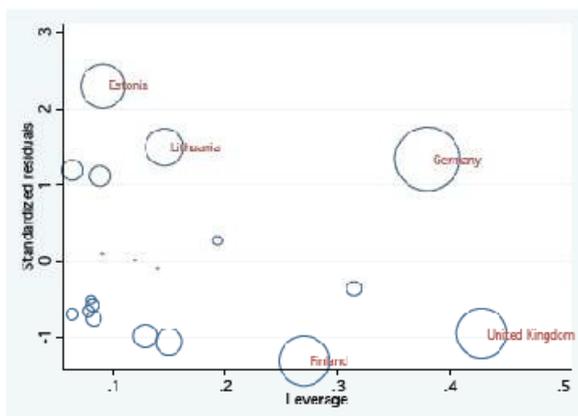
Residuenplots: anstelle von Histogrammen kann man die Residuen auch in Residuenplots darstellen, so sieht man, wie die Residuen um 0 variieren. Einflussreiche Beobachtungen können so aber nicht erkannt werden.



Einflussreiche Beobachtungen: Entscheidend ist der Begriff „leverage“, was soviel wie Hebelwirkung bedeutet, welche eine Beobachtung auf das Resultat der Regression hat. Der Einfluss ergibt sich aus der Grösse des Residuums und der Grösse des leverages.

Cook's D(istance): beschreibt den Einfluss einer Beobachtung auf alle  $\hat{y}$ . Distance meint den Unterschied zwischen den vorhergesagten Werten mit und ohne Beobachtung.

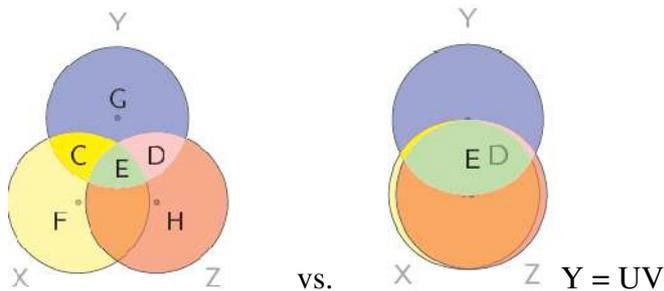
Zur Veranschaulichung:



Kreise stellen Cook's D dar. Es wird erkannt, dass auch um 0 herum (keine Ausreisser) die Kreise grösser sein können.

Lösungsansätze: die Beobachtungen können einfach ausgeschlossen werden, es soll aber vor allem herausgefunden werden, warum eine Beobachtung ein Ausreisser darstellt (Messfehler? Modell schlecht spezifiziert, z.B. nicht-linearer Zusammenhang?). Falls Ausreisser im Modell behalten werden, muss dies bei der Interpretation berücksichtigt werden.

Multikollinearität (MK): Je stärker die UVs miteinander korrelieren, desto stärker nähern sie sich der Multikollinearität.



im ersten Diagramm erklären die einen Variablen andere nur zum Teil (C, E, D). G, H, F bleiben unerklärt, dies deutet auf keine MK. Im zweiten Diagramm hingegen korreliert Z sehr stark mit X. X erklärt Y zudem nur sehr gering „eindeutig“ (Z erklärt quasi mit). Einflüsse auf Y können nicht eindeutig auf X oder Z zurückgeführt werden → Multikollinearität.

Wie wird MK erkannt?

- Hohes  $R^2$  bei nur wenig signifikanten t-Tests ist verdächtig (Variablen mit nicht-signifikanten sollten untersucht werden)
- starke bivariate Korrelationen zwischen Variablen ( $r > 0.8$ )
- sehr hohes  $R^2$  (höher als das eigentliche Regressionsmodell) weist ebenfalls auf MK hin.

Konsequenzen MK: perfekte MK lässt keine Berechnung des Regressionsmodells zu, bei nahezu perfekter MK sind die OLS-Schätzer immer noch unverzerrt und sogar BLUE (linear, unverzerrt, effizient), aber die Standardfehler sind, im Gegensatz zur Grundgesamtheit (Vermutung keine MK), zu gross. Obwohl  $R^2$  relativ hoch ist, sind die Koeffizienten meist nicht signifikant, kleine Veränderungen der Daten können grosse Auswirkungen auf Koeffizienten haben.

Lösungsansätze MK: aus den multikollinearen Variablen wird ein gemeinsamer Index konstruiert (z.B. BIP – Ausgaben = Sparquote) oder die Stichprobe wird vergrössert um mehr Variation zu erlangen. Ausschluss einer korrelierenden Variablen ist schlecht, da bewusste Verzerrung in Kauf genommen wird. Falls die Berechnung des Modells möglich ist, muss der t-Test durch einen gemeinsamen F-Test für die betroffenen Variablen ersetzt werden.

Heteroskedastizität: Ist gegeben, wenn die Residuen einer Stichprobe möglichst zufällig um die 0-Linie variieren (kann nur bei grossem n festgestellt werden). Für jedes  $x_i$  wird eine

eigene Varianz der Fehlerterme erwartet. Heteroskedastizität hat in der Regel einen Grund, die Varianzen folgen einem Muster (mit den Variablen an sich zu begründen, oder grosse Ausreisser, wahre Beziehung zwischen den Variablen ist nicht linear, UV hat eine sehr schiefe Verteilung, wichtige Variablen fehlen in der Regressionsgleichung)

Konsequenzen Heteroskedastizität: Schätzer sind nicht mehr BLUE (zwar unverzerrt und konsistent, aber ineffizient), Standardfehler wird falsch geschätzt, t- und F-Tests des OLS-Schätzers sind unzuverlässig.

Zur Überprüfung der Heteroskedastizität wird generell der White-Test durchgeführt. Er misst die Beziehung zwischen quadrierten Residuen und allen UVs

Lösungsansätze Heteroskedastizität:

- log-Transformation des Regressionsmodells oder nur der AV.
- Gründe versuchen zu antizipieren (Literatur)
- Gewichtete Regression (WLS): Werte der AV und der UV durch die Standardabweichung teilen, also *gewichten*. So wird die Varianz der Fehlerterme zur Konstanten transformiert und homoskedastisch.
- Berechnung robuster Standardfehler: falls die erwähnten Strategien nicht möglich sind, werden die robusten Standardfehler benutzt. Das Problem wird mathematisch umgangen. Das Statistikprogramm baut *neben den normalen Residuen der ursprünglichen Regression auch auf die Residuen von Hilfsregressionen zwischen den UVs*. Diese robusten Standardfehler sind zwar grösser als die ohne Heteroskedastizität, aber nicht unbedingt kleiner als ohne Korrektur.

## Vorlesung 13 HS08: Multiple Regression: Transformation

**Verschiedene Arten von Transformationen:** Logarithmische Transformation des ganzen Modells, nur T. der AV oder UV, Reziproke, Quadratisches und Kubisches Modell

### **Gründe für Transformationen:**

Verringern der Heteroskedastizität, Erleichterte Interpretation der Koeffizienten, Umsetzung eines theoretischen Modells

**ABER:** Transformationen verändern:

- die Interpretationen der Koeffizienten
- evtl. die funktionale Form und die Annahmen des Regressionsmodells.

### **1. Logarithmische Transformation des gesamten Modells „log-lineares Modell“**

Regressionsgleichung:  $\ln(y_i) = \ln(\alpha) + \beta \ln(x_i) + \epsilon_i$

- Modell ist nur sinnvoll, wenn das ursprüngliche Modell ein *exponentielles Regressionsmodell* ist:  $y_i = \alpha x_i^\beta * e^{\epsilon_i}$ , dieses wird durch die Logarithmierung linearisiert.

#### **Interpretation des Modells:**

- Steigungskoeffizient  $\beta$  misst Elastizität (= *proportionale* Veränderung von y aufgrund einer *proportionalen* Veränderung von x) von y bezüglich x. Diese wird als konstant über alle Werte von x angenommen.
- Koeffizient positiv: Kurve steigt an (in beiden Modellen), Koeffizient  $> 1$ : Anstieg ist im linearisierten Modell unterproportional (Gerade ist flacher als 45-Grad)
- Der Logarithmus von 0 ist nicht definiert, daher muss Skala immer um einen kleinen Wert erhöht werden.

#### **Probleme des Modells:**

- nur möglich für Werte grösser 0
- Annahmen bezüglich der Fehler des exp. Modells ändern sich je nach Spezifizierung des Fehlerterms
- die Annahmen des klassischen linearen Regressionsmodells gelten für die linearisierte Regressionsgleichung nur, wenn der Fehlerterm ( $\epsilon_i$ ) exponiert und multiplikativ ins Modell integriert wird.

### **2. Logarithmierung der AV „log-lin Modell“**

Regressionsgleichung:  $\ln(y_i) = \alpha + \beta x_i + \epsilon_i$

- verwendet um Ausreisser oder Heteroskedastizität in den Griff zu bekommen
- Modell lässt sich aus der Theorie ableiten

#### **Interpretation des Modells:**

- Steigungskoeffizient  $\beta$  misst die *konstante proportionale* oder die *relative* Veränderung von y für eine *absolute* Veränderung von x
- Koeffizient positiv: Kurve steigt

### **3. Logarithmierung einer UV „lin-log Modell“**

Regressionsgleichung:  $y_i = \alpha + \beta \ln(x_i) + \epsilon_i$

- um Ausreissen in den Griff zu bekommen oder wenn es theoretische Gründe für dieses Modell gibt

#### **Interpretation des Modells:**

- Steigungskoeffizient  $\beta$  misst die *absolute* Veränderung der AV aufgrund einer *proportionalen* Veränderung der UV

#### **Probleme von lin-log und log-lin:**

- Transformation nur möglich für Werte grösser 0
- erschwerte Interpretation der Koeffizienten

### **4. Reziprokes Modell (Kehrwert einer UV)**

Regressionsgleichung:  $y_i = \alpha + \beta * 1/x_i + \epsilon_i$

- AV nähert sich dem Wert des Achsenabschnitts an für sehr grosse und sehr kleine Werte der UV

- sinnvoll für Zusammenhänge für die ein best. Wert der AV nicht unterschritten werden kann,  
theoretische Überlegungen

# Zusammenfassung Methoden VL 1 FS09

## Repetition Regressionsanalyse (RA)

- § Ziel der RA: Kausalen Effekt\* einer UV auf die AV zu messen (mit einer Zufallsstichprobe) und zu wissen wie sicher\*\* es ist, dass der Effekt in der Grundgesamtheit gilt
- § Also interessiert: die Richtung\*, Stärke\* und statistische Inferenz\*\* des Effekts
- § Es wird eine Linie in die Beobachtungswerte „eingepasst“.
- § So werden die Parameter beta und alpha geschätzt/vorhergesagt
- § E steht hier für die Residuen aller Beobachtungswerte

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{\epsilon}_i$$

- § Wobei die Schätzung um so besser ist, je kleiner die SSE (Summe der quadrierten Abweichungen) sind

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\epsilon}_i^2$$

§ Ziel: diesen Wert minimieren! (y<sub>i</sub> ist der Beobachtungswert, y<sub>i</sub>-Dach der vorhergesagte Wert auf der RegLinie)

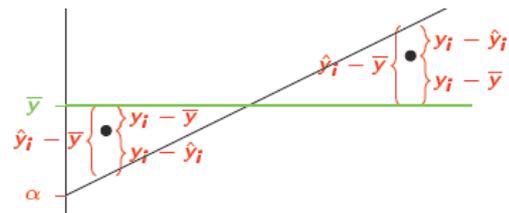
- § Das Modell nimmt eine lineare Beziehung der UV und der AV an

§ Lineare Regressionsfunktion:  $y_i = \alpha + \beta x_i + \epsilon_i$

- § Annahme: die Varianz bzw die Verteilung der Fehler ist konstant

$$R^2 = \frac{TSS - SSE}{TSS}$$

§ die grüne Linie bedeutet keinen Zusammenhang. Es wird jeweils der Abstand des Beobachtungswerts zur grünen Linie sowie zum vorhergesagten Wert (RegGerade.) quadriert,



anschliessend anhand der obigen Formel berechnet. R<sup>2</sup> sagt aus wie viel das Modell erklären kann im Vergleich zur Nullhypothese (kein Zusammenhang), d.h. R<sup>2</sup> von 0.4 bedeutet, dass die UV 40% der Veränderung der AV erklärt. Hypothetisch: Falls die SSE = 0 sind wäre gemäss Formel R<sup>2</sup> = 1 (TSS/TSS) d.h. alle Beobachtungen lägen auf der Reg.Gerade und wären somit korrekt vorhergesagt

- § Damit Inferenz (auf Grundgesamtheit schliessen): Viele Zufallsstichproben ziehen gäbe versch. Parameter (z.B.Beta für die Steigung), die aber entsprechend einer Sampling Distribution verteilt sind, welche normalverteilt ist (da Annahme, dass Fehler auch normalverteilt, bei wenigen N: t-verteilt)
- § Zweiseitiger Signifikanztest für den Steigungsparameter, Testen ob H<sub>0</sub> gilt:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} = \frac{0.0903}{0.0258} = -3.5$$

T berechnet sich folgendermassen: also t= Koeffizient/Standardfehler wobei  $\sigma^2$ = Varianz der Residuen

Wenn dieser Wert (-3.5) nun ausserhalb des kritischen Werts (Tabelle nachsehen) liegt für ein gewähltes Signifikanzniveau  $H_0$  verwerfen

§ Der Konfidenzintervall definiert sich für die Freiheitsgrade  $n-k-1$  und z.B.  $\alpha=0.5$  folgendermassen:

$$\hat{\beta} \pm t_{(0.025, n-k-1)} * \hat{\sigma}_{\hat{\beta}}$$

§ Statistische Signifikanz = Der geschätzte Effekt gilt mit einer bestimmten Wahrscheinlichkeit für die Grundgesamtheit

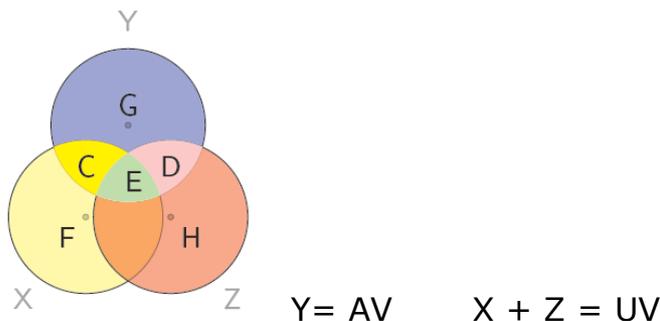
§ substantielle Signifikanz: Der Effekt ist genug stark um als substantiell (bedeutend) zu gelten Aufpassen: Was entspricht einer Einheit (Mio. Fr. oder Rp.?)

§ Der Koeffizient misst immer den Effekt auf die AV für eine Veränderung der UV um eine Einheit

Bsp. Parameter Beta=0.5 Wenn die UV um 1 Einheit steigt, dann steigt die AV um 0.5 Einheiten

§ Multiple Regression: Einfluss mehrerer UV's auf die AV messen ist möglich mit der RA

### Venn-Diagramm



D: Z erklärt Y eindeutig

E: Wird von UV's erklärt aber unsicher von welcher

FGH: Tragen nicht zur Erklärung von Y bei

§ Bei Vergleich von Regressionsmodellen beachten:

- Anzahl Fälle gleich?
- Koeffizienten nur vergleichen wenn zur gleichen Skala gehören (beides bspw. Dummies oder intervallskaliert)
- Einbezug von zusätzlichen Variablen muss theoretisch begründet sein
- Qualität beurteilen:  $R^2$ , N, Inferenz möglich? (Zufallsauswahl?)

## Interaktionsterme

Interaktionsmodelle werden benützt, wenn aus der Theorie ein konditionaler Zusammenhang vorhergesagt wird, das heisst, wenn die Stärke des Effektes einer Unabhängigen auf die Abhängige von einer zweiten Unabhängigen beeinflusst wird.

Interaktionsmodell:  $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$

### ***X<sub>2</sub> als Dummy-Variable***

Wenn vereinfachend angenommen wird, dass  $x_2$  eine dichotome Dummyvariable abbildet, die nur die Werte 0 oder 1 annehmen kann, so ergeben sich aus dem Modell eigentlich zwei Regressionsgeraden.

1. Für  $x_2 = 0$ : ist  $(\beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}) = 0$ , also  $y = \alpha + \beta_1 x_1$ 
  - a. Die Steigung ist also durch  $\beta_1$  ausgedrückt
  - b. Und der Achsenabschnitt liegt bei  $\alpha$
2. Für  $x_2 = 1$ : ist  $(\beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}) = (\beta_2 + \beta_3 x_1)$ , also  $y = \alpha + \beta_2 + (\beta_1 + \beta_3) x_1$ 
  - a. Die Steigung ist nun gleich  $\beta_1 + \beta_3$
  - b. Und der Achsenabschnitt bei  $\alpha + \beta_2$

Daraus folgt eine konditionale Interpretation der Koeffizienten:

$\alpha$  : Niveau der Abhängigen wenn  $x_1$  und  $x_2$  beide gleich 0 sind

$\beta_1$  : Effekt von  $x_1$  auf  $y$  wenn  $x_2 = 0$

$\beta_2$  : Effekt von  $x_2$  auf  $y$  wenn  $x_1 = 0$

$\beta_3$  : Veränderung des Effektes von  $x_1$  auf  $y$  wenn  $x_2$  von 0 auf 1 wechselt

$\alpha + \beta_2$  : Niveau der Abhängigen wenn  $x_1 = 0$  und  $x_2 = 1$

$\beta_1 + \beta_3$  : Effekt von  $x_1$  auf  $y$  wenn  $x_2 = 1$

### **Vorsicht:**

- An der Signifikanz des Interaktionsterms kann nicht abgelesen werden, ob das Interaktionsmodell sinnvoll ist
- Auch die **Standardfehler sind konditional** zu interpretieren (genau wie die Koeffizienten)

Aus der Signifikanz des Koeffizienten  $\beta_1$  lässt sich die Signifikanz des Zusammenhangs zwischen  $x_1$  und  $y$  ablesen, unter der Voraussetzung, dass  $x_2 = 0$  (=NEIN) ist.

Nun sollte auch noch überprüft werden, ob dieser Zusammenhang auch signifikant existiert, wenn  $x_2$  nicht (NEIN) ist. Dazu kann  $x_2$  einfach rekodiert werden, so dass 0 und 1 vertauscht werden, sodass 0 = JA und 1 = NEIN o.Ä.. Wenn dann die Regression noch einmal durchgeführt wird, gilt der Koeffizient  $\beta_1$  und ebenso seine Signifikanz für die Voraussetzung, dass  $x_2 = JA$  ist.

### ***X<sub>2</sub> als kontinuierliche Variable***

Nun wird die vereinfachende Annahme, dass  $x_2$  nur die Werte 0 und 1 annehmen kann, weggelassen.

Wenn angenommen wird, dass entweder

- Die Unabhängige merkwürdig verteilt ist und mit vielen extremen Ausreissern zu rechnen ist
- Oder die Grösse des Effektes von  $x_2$  auch  $y$  mit grösserem  $x_2$  abnimmt, also ein **exponentieller Zusammenhang** besteht

Dann kann die Unabhängige logarithmiert werden, so dass der Koeffizient die *absolute Veränderung von y aufgrund einer proportionalen Veränderung von  $x_2$*  misst (lin-log-Modell)

Eine Veränderung von  $\log(x_2)$  und eine Einheit wird also eine Veränderung von  $y$  um  $\beta_2$  nach sich ziehen. Verändert sich aber  $x_2$  um z.B. 1%, verändert dies durch die Logarithmierung  $y$  nur um  $\beta_2/100$

Der Koeffizient  $\beta_1$  wird nun nicht mehr konditional zu  $x_2$  interpretiert, sondern nun zu  $\log(x_2)$ . Er zeigt also den Effekt von  $x_1$  für den Fall das  $\log(x_2)=0$ , was  $x_2=1$  entspricht.  $\beta_2$  zeigt den Effekt einer Veränderung von  $\log(x_2)$  um eine Einheit.

## Zentrierung

Da jeder Koeffizient nur für den Fall, dass die zweite Unabhängige = 0 ist, überhaupt interpretiert werden kann, macht es keinen Sinn Modelle zu interpretieren, wo eine Unabhängige gar nicht 0 sein kann. In einem solchen Fall muss die Kodifizierung der Unabhängigen zuerst so verschoben werden, dass der Nullpunkt an einem interpretierbaren Ort zu liegen kommt, etwa beim Mittelwert. Dies geschieht durch einfache Subtraktion des Mittelwertes von allen Werten. Der Koeffizient stellt nun den Effekt der Unabhängigen dar, unter der Voraussetzung, dass die 2. Unabhängige gemittelt ist.

## Marginale Effekt im Interaktionsmodell

Der Marginale Effekt einer Unabhängigen  $x_1$  wird im normalen Regressionsmodell durch ihre Steigung  $\beta_1$  ausgedrückt. Im Interaktionsmodell entspricht er nun  $(\beta_1 + \beta_2 x_2)$  ist also von der zweiten Unabhängigen  $x_2$  abhängig.

In einem Interaktionsplot lässt sich die Veränderung des Marginalen Effektes in Abhängigkeit der zweiten Unabhängigen darstellen.

Auch der Standardfehler des marginalen Effektes ist im normalen Regressionsmodell einfach gleich dem Standardfehler der Steigung. Im Interaktionsmodell entspricht er einer komplizierteren Formel, die hoffentlich niemand von euch je brauchen wird.

And now for something completely different

## Logistische Regression

In den bisher gesehenen Regressionen war die Abhängige Variable jeweils metrisch skaliert. Oft hat man es aber mit ordinal oder nominal skalierten Variablen zu tun, etwa beim dichotomen Abstimmungsentscheid Ja/Nein oder bei nominalen Wahllisten.

Verwendet man dazu aber einfach eine gewöhnliche OLS-Regression, was man dann linear probability model (LPM) nennt, ergeben sich daraus einige Probleme.

Da es Unsinn ist, zu sagen, bei einer Veränderung der Unabhängigen um 1 verändert sich die Abhängige um  $\beta$  Einheiten (diese kann ja nur 0 oder 1 sein), wird  $\alpha + \beta x$  nun als **Wahrscheinlichkeit** dafür interpretiert, dass die binäre Abhängige den Wert 1 annimmt. Ist diese über 0.5 so nimmt man an, dass der Wert 1 eingetreten ist.

Die Punktwolke besteht nun aus Beobachtungen die immer entweder den Wert 0 oder 1 haben, die Regressionsgerade wird so gut wie möglich dazwischen gelegt, doch ist die Fehlervarianz dabei heteroskedastisch, die Fehler sind nicht normalverteilt und die Regressionsgerade läuft aus dem Wertebereich der Abhängigen hinaus.

Der beste Weg um diesen Problemen zu begegnen ist es, ein alternatives Schätzverfahren zu bemühen.

Ist die manifeste Variable **binär** wird die **Binomial-** bzw. die **Bernoulli-Verteilung** verwendet.

Die Bernoulli-Verteilung basiert auf dem Gedankenspiel, dass bei einem Würfelwurf das Ereignis ‚6-Augen‘ im Vergleich zum Gegenereignis ‚<6 Augen‘ interessiert. Für jeden Wurf wird nun notiert, ob das Ereignis ‚6-Augen‘ eingetroffen ist, oder sein Gegenereignis.

Die **Wahrscheinlichkeitsfunktion** ist dann:

$$f_B(y_i | n; p) = \frac{n!}{y_i!(n - y_i)!} p^{y_i} (1 - p)^{n - y_i}$$

n= Anzahl Versuche also =1, da pro Beobachtung nur ein Wurf gemacht wird;  
p=WS für ‚6-Augen‘ und y<sub>i</sub>=Anzahl ‚6-Augen‘ in n Versuchen

Ist n=1 so heisst die Verteilung Bernoulli-Verteilung, ist n grösser, handelt es sich um eine Binomialverteilung.

Dasselbe kann nun z.B. auf das Abstimmungsverhalten angewandt werden, und notiert werden, ob eine Person abstimmt oder nicht.

Die Eigenschaften der Bernoulli-Verteilung gelten dann auch hier.

$$E(y | p) = p$$

$$Var(y | p) = p(1-p)$$

Die manifeste Variable folgt einer Bernoulli-Verteilung und somit verletzt sie Grundannahmen der OLS: Ihre Fehler folgen auch einer Bernoulli-Verteilung und sind somit nicht normalverteilt, und die Fehler hängen von x<sub>i</sub> ab => Heteroskedastizität

Das Problem ist nun, dass die WS für das Eintreten des gefragten Ereignisses nicht wie bei den Würfeln bei allen gleich ist, sondern dass jeder Stimmbürger seine eigene WS dafür hat, abzustimmen. Deshalb brauch es eine **Link-Funktion**, die die Binomialverteilung reparametrisiert, also wird die WS für das Ereignis als Abhängige gesehen, die auf die Erklärende zurückgeführt werden kann.

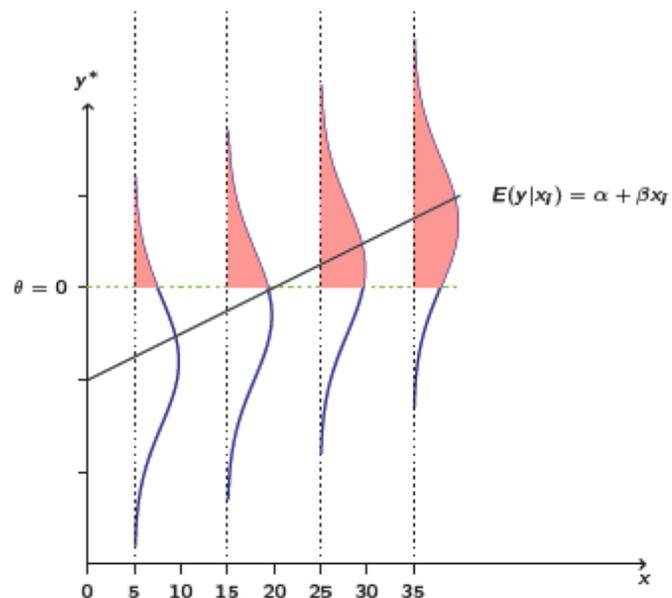
Eine Möglichkeit ist dafür das LPM:  $E(y_i | p_i) = p_i = \alpha + \beta x_i$ ; aber eben nicht die Beste.

Um die Interpretation von  $y^*$  als die WS für das Eintreten von y=1 zu rechtfertigen, wird eine der manifesten Variable zugrunde liegende latente, also nicht direkt messbare Variable angenommen. Ist die latente Variable grösser als 0 wird die manifeste als 1 angenommen und umgekehrt als 0, wenn die latente kleiner als 0 ist.

Diese **latente Variable** wäre eigentlich kontinuierlich messbar, linear und könnte ohne weiteres in einem normalen OLS-Modell verwendet werden.

Die gesuchte Wahrscheinlichkeit für y=1 der manifesten Variable

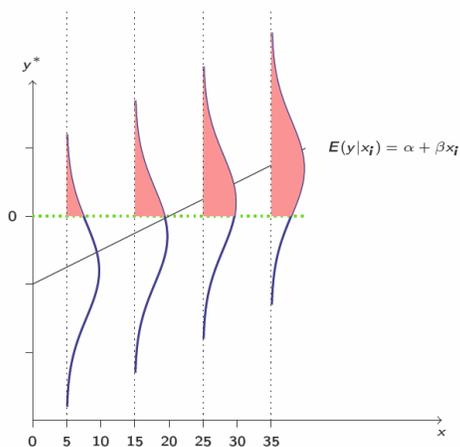
entspricht der Wahrscheinlichkeit dafür, dass die latente grösser als 0 ist. Dies entspricht wiederum der Fläche unter der Kurve der bedingten Wahrscheinlichkeitsverteilung im Bereich  $y^* > 0$ . Diese Fläche steigt für tiefe x zuerst stark an, der Anstieg wird aber immer schwächer.



#### 4. Sitzung FS 09: Probit- und Logit-Modelle; Maximum-Likelihood-Schätzung

##### Geeignete funktionale Formen für $\pi_i$ :

- Meistens wird eine kumulative Verteilungsfunktion (cumulative density function, cdf) als Identitätsfunktion benutzt. Diese lässt sich möglicherweise sogar explizit aus dem zugrunde liegenden theoretischen Entscheidungsmodell herleiten.
- Die Wahrscheinlichkeitsverteilungsfunktionen liegen immer im Intervall  $[0,1]$
- Kontinuierliche Wahrscheinlichkeitsverteilungsfunktionen haben eine Wahrscheinlichkeitsdichtefunktion (probability density function, pdf). Diese pdf zeigt sich oft als „Glockenkurve“. Die Fläche unter der Kurve für einen bestimmten Wertebereich ist die Wahrscheinlichkeit für das Eintreten dieser Werte.
- Die cdf hingegen ist eine Funktion für eben diese Fläche unter der pdf und misst deshalb die Wahrscheinlichkeit für alle Werte bis zu einem bestimmten Punkt.

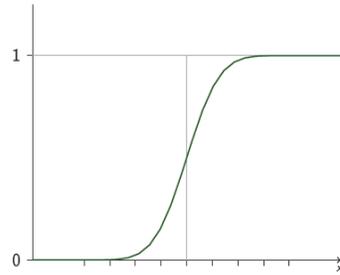


- Die cdf wird hier als Standardnormalverteilung aufgezeigt. Diese dient auch als Link-Funktion
- Die aufsteigende Gerade ist die Linkfunktion
- Die (rot) eingefärbte Fläche oberhalb der Link-Funktion zeigt hier im Beispiel die W'keit wählen zu gehen, die Fläche unterhalb die W'keit nicht wählen zu gehen.

##### Das Probit-Regressionsmodell:

- Die Link-Funktion dient zur Berechnung der Wahrscheinlichkeiten
- Als Link-Funktion bietet sich die Verteilung der Fehler aus dem linearen Modell für die latente abhängige Variable an.
- Es wird angenommen, dass diese Fehler normalverteilt sind mit bedingtem Mittelwert von jeweils Null und mit Varianz 1. Entspricht der Standard-Normalverteilung.

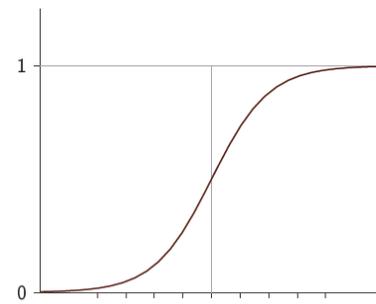
- Wenn man die Standardnormalverteilung als Link-Funktion einsetzt, präsentiert sie sich folgendermassen:



$$p_i = \int_{-\infty}^{\alpha + \beta x} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$$

### Das Logit-Regressionsmodell:

- Die am weitesten verbreitete Link-Funktion ist die cdf der Standardlogistischen Verteilung
- Sie wird folgendermassen geschrieben:
- Die Logit-Link-Funktion hat sich infolge beschränkter Computerleistungen durchgesetzt. Deshalb spricht man auch von einer logistischen Regression
- (Folie 10 einsetzen).



Die standardlogistische Verteilungsfunktion als Link-Funktion:

$$p_i = \frac{1}{1 + \exp^{-(\alpha + \beta x_i)}}$$

- Da sich das Logit-Modell auch alternative herleiten. Diese Herleitung geht von dem Problem aus, dass im linearen Wahrscheinlichkeitsmodell der Wertebereich der vorhergesagten Werte über das Intervall  $[0, 1]$  hinaus geht. Deshalb wird durch eine Transformation der binären abhängigen Variable dieser Wertebereich auf das Intervall  $[0, 1]$  zurückgebunden werden.
- Diese Transformation geschieht durch 3 Schritte: 1. Die Bildung von „odds“ (Chance der Beobachtung von 1, bzw. das Verhältnis von Wahrscheinlichkeit zu Gegenwahrscheinlichkeit). 2. Damit das Limit von 1 eingehalten wird, werden die „odds“ logarithmiert. Somit begrenzt sich der Wertebereich zwischen 0 und 1. 3. Formt man die Abhängige wieder auf die ursprüngliche Skala zurück, d.h. in eine Wahrscheinlichkeit zurückwandeln.

### Maximum-Likelihood-Schätzverfahren

- Mit dem MLS sollen jene Parameter des Modells gefunden werden, welche die Wahrscheinlichkeit, dass die Werte der manifesten Variablen beobachtet werden, maximieren.
- Der Ausgangspunkt einer MLS ist die Wahrscheinlichkeitsverteilung der Abhängigen. (Bei einer dichotomen Abhängigen die Bernoulli-Verteilung).
- Da alle Beobachtungen in einer Stichprobe berücksichtigt werden, wird die Funktion der gemeinsamen Verteilung der Beobachtungen der manifesten Variablen, die Likelihood-Funktion, verwendet. Diese ist das Produkt der Wahrscheinlichkeiten für das Auftreten der Ausprägung 1 für alle Beobachtungen.

- Likelihood-Funktion: 
$$LF = \prod \{ p_i^{y_i} \times (1 - p_i)^{1-y_i} \}$$

- Für das Logit-Modell gilt die Link-Funktion: 
$$p_i = \frac{1}{1 + e^{-(\alpha + \beta x_i)}}$$

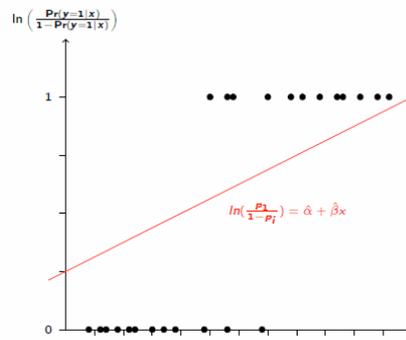
- Weil die Werte der Likelihood-Funktionen sehr klein werden, wird die Likelihood-Funktion logarithmiert. Somit geht der Wertebereich von minus unendlich bis 0. Die Werte sind abhängig davon, wie viele erklärende Variablen das Modell enthält und wie viele Beobachtungen in der Stichprobe sind. Je näher die Likelihood-Funktion bei 1 liegt, umso näher liegt die Log-Likelihood bei 0 und umso wahrscheinlicher ist es, dass die geschätzten Parameter die beobachteten Werte der manifesten Variable produzieren.

## 5. Sitzung FS 09: Interpretation der Effekte

### Interpretation der Koeffizienten

$$\ln\left(\frac{p_i}{1-p_i}\right) = \hat{\alpha} + \hat{\beta}x_i$$

- ▶  $\hat{\alpha}$  und  $\hat{\beta}$  stellen die lineare Verbindung her zwischen der Unabhängigen und dem "Logit" der Wahrscheinlichkeit die Ausprägung 1 für die Abhängige zu beobachten.
  - ▶ Die "rohen" Koeffizienten lassen sich also interpretieren wie im linearen Regressionsmodell, nur dass die abhängige Variable eine **intuitiv schwer nachvollziehbare Transformation** der eigentlich interessierenden Größe ist: der Wahrscheinlichkeit für  $y=1$ .
- Durch teilweise Auflösung des Logit erhält man auf der Seite der Unabhängigen die „odds“. Deren Interpretation ist leichter nachvollziehbar. Die klarste Interpretation der Ergebnisse erhält man, wenn die Wirkung einer Veränderung einer Unabhängigen auf die Wahrscheinlichkeit für  $y = 1$  beschrieben wird.



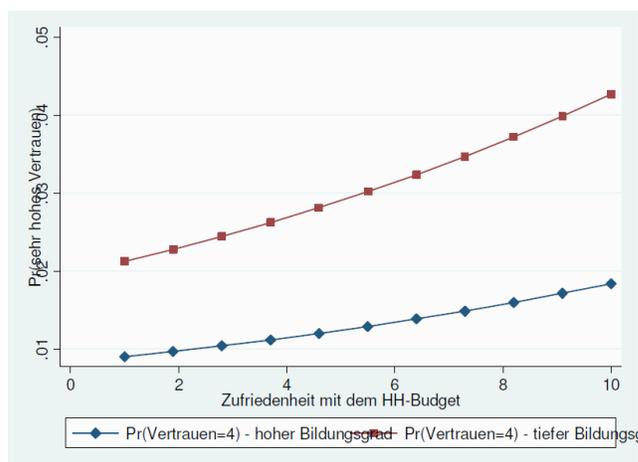
- Die exponentierten Koeffizienten messen die „odds ratio“, also die Veränderung der Odds durch die Veränderung der Erklärenden um eine Einheit. Sie liegen immer zwischen 0 und unendlich.
- Die Berechnung der vorhergesagten Wahrscheinlichkeit an einem bestimmten Punkt der unabhängigen Variablen garantiert die natürlichste Interpretation des Einflusses der Unabhängigen auf die Abhängige. Diese Berechnung ist aber nicht ganz trivial, da die Beziehung zwischen der Unabhängigen und der Wahrscheinlichkeit für  $y = 1$  nicht-linear ist und die übrigen Unabhängigen nicht-additiv ins Modell eingehen, weshalb deren Niveau auf den Effekt der einen Unabhängigen an einem bestimmten Punkt mitbestimmt

## Marginale Effekte

- Der marginale Effekt misst den Einfluss einer unabhängigen Variablen auf die Wahrscheinlichkeit dafür, dass  $y$  gleich 1 an einem bestimmten Punkt der unabhängigen Variable ist.
- Es wird also die Steigung der nicht-linearen Funktion für die Wahrscheinlichkeit an einer bestimmten Stelle der Unabhängigen gesucht. Diese Steigung verändert sich, je nachdem an welcher Stelle der unabhängigen Variable die Steigung gemessen wird
- Bei einer Dummy-Variablen macht es keinen Sinn einen marginalen Effekt zu berechnen. Hier interessiert vielmehr der Effekt auf die Wahrscheinlichkeit für den Sprung von 0 auf 1.

## Vorhergesagte Wahrscheinlichkeiten

- Die letzte Möglichkeit zur Interpretation der Ergebnisse eines Modells ist die Darstellung der vorhergesagten Wahrscheinlichkeiten
- Diese spielen bei der Berechnung der marginalen Effekte auch eine Rolle. Man kann sich aber darauf konzentrieren, die Werte der vorhergesagten Wahrscheinlichkeiten für unterschiedliche Werte der unabhängigen speziell darzustellen, um die Effekte anschaulicher zu machen.



hier ein Beispiel:

- je grösser die Zufriedenheit mit dem Haushaltsbudget (0= sehr unzufrieden, 10=sehr zufrieden), desto höher wird die W'keit für sehr hohes Vertrauen in die Regierung.
- Dies aber weniger stark bei einem hohen Bildungsgrad (untere Kurve)

6 Sitzung FS09; Lernziele:

Signifikanz der Koeffizienten – Korrekte und inkorrekte Vorhersagen – Likelihood Ratio Test – Pseudo  $R^2$  – Residuen

### Signifikanz der Koeffizienten

Die Interpretation [der Koeffizienten] deckt sich mit den Angaben in den entsprechenden Spalten bei der linearen Regression. (Kohler und Kreuter, 273) Für die Signifikanz (der einzelnen Koeffizienten) werden also in Erster Linie der z-Wert (entspricht dem t-Wert der linearen Regression) und der p-Wert benützt. Um zu schauen, ob der z-Wert Signifikant ist, muss nur der Wert betrachtet werden. Z-Werte, welche grösser als 1.64 (5%-Niveau) und kleiner als -1.64 (5%-Niveau) sind, sind Signifikant. Daraus folgt das Werte zwischen -1.64 & 1.64 nicht genügend Signifikant sind. Für den p-Wert gilt, je näher an 0 desto besser. Die Signifikanz der geschätzten Koeffizienten wird bei der logistischen Regression jedoch üblicherweise mit einem Likelihood-Ratio-Test ermittelt. (Kohler und Kreuter, 273)

### Korrekte und inkorrekte Vorhersagen & PRE-Mass

Das "schöne" an der logistischen Regression ist, dass berechnet werden kann wie gut unser Modell ist. Mittels einer Kreuztabelle (tab A B if e(sample), cell) kann erkannt werden, wie viele Beobachtungen richtig Vorhergesagt wurden. Bei dieser Methode werden diejenigen Prozenste zusammengezählt, welche für dieselben Ausprägungen stehen (z.B. für 0/0 **21%**, 1/1 **50%**). Den erhaltenen Wert kann in einem nächsten Schritt zur Berechnung des PRE-Masses dienen. Mit dem PRE-Mass können wir den Anteil der richtigen Vorhergesagten aus unserem Modell verglichen mit dem Anteil korrekt Vorhergesagten aus dem Null-Modell. Den Anteil korrekt Vorhergesagten aus dem Null-Modell kriegen wir, indem die Häufigkeiten der AV betrachtet werden. Dabei wird einfach für jedes Individuum das vorausgesagt, was am häufigsten Auftritt, denn dann liegen wir in mind. 50% der Fälle richtig. Sagen z.B. **70%** Ja, so nehmen wir für alle Individuen Ja an und liegen somit in 70% der Fälle richtig. Das PRE-Mass berechnet sich nun wie folgt:  $(0.21+0.5)-0.7/1-0.7 = 0.03$ . Diese Zahl sagt nun, dass unser Modell aus der logistischen Regression die Fehler um ca. 3% reduziert verglichen mit dem Nullmodell.

### Likelihood-Ratio-Test (LR-Test)

Generell lässt sich beim LR Test sagen, je grösser und näher der Wert an Null ist desto besser. Allerdings hängt dieser Wert mit der Fallzahl zusammen, weshalb der erhaltene Wert für jedes Modell beurteilt werden muss. Es gibt keinen kritischen Wert. Der LR-Test ist ein Test, in dem die LR-Test-Werte eines Anfangsmodells ( ohne eine oder mehrere Variable(n)) mit den LR-Werten des Endmodells verglichen werden (mit allen Variablen). Dadurch kann mit dem LR-Test auch getestet werden, ob die Hinzunahme oder Wegnahme einer Variable das Modell verbessert. Den LR  $\chi^2$  (X) Wert den Stata uns nach einer log-regression angibt, wurde mit -2 multipliziert, weshalb der Wert positiv ist. Durch diese Multiplikation wurde der Wert  $\chi^2$ -Verteilt (Welche Vorteile das bringt? Keine Ahnung, sorry). Auf jeden Fall gibt dieser Wert von Stata die Differenz des vollen Modells zum Nullmodell (hier: je kleiner und näher an null desto besser). Dies entspricht analog dem F-Test beim OLS-Modell.

## Pseudo R<sup>2</sup>

Die meisten der gebräuchlichen R<sup>2</sup> (einige Freaks entwickelten ihr eigenes) für die log-regression basieren auf dem Vergleich der Likelihood Werte für ein volles und ein Null-Modell. Solche Masse sind mit grosser Vorsicht zu interpretieren. Sie können auf keinen Fall mit den R<sup>2</sup>-Werten der OLS-Regression verglichen werden. Schulz schlägt vor, dass höchstens solche Pseudo R<sup>2</sup>-Werte für unterschiedliche Modelle verglichen werden.

## Residuen

Mittels Cook's D oder ähnlichen Massen, lassen sich Hinweise finden, welche Beobachtungen grosse Residuen und eine grosse Hebelwirkung auf die Regressionslinie aufweisen. Werden solche Beobachtungen gefunden, so sollte geprüft werden, ob diese Gemeinsamkeiten aufweisen. Anhand dieser Prüfung könnte anschliessend eine Kontrollvariable eingefügt oder sonst eine Anpassung des Modells die Konsequenz sein. (Vorschlag von Schulz)

## 7. Sitzung FS09; Lernziele: Anwendungsbeispiele (Erlach; Wird man in Vereinen politisch sozialisiert?)

Neu kommt in dieser Vorlesung die Möglichkeit Dummies für kategoriell unabhängige Variablen zu bilden. Indem man in Stata vor der Variable ein i. hinstellt, teilt sich die Variable in die Verschiedenen Ausprägungen aus. Gleichzeitig wird eine Kategorie unterdrückt (automatisch die Erste, wenn nichts anderes programmiert wird), welche dann als Referenzkategorie dient. Anhand dieser Referenzkategorie kann für jede einzelne Ausprägung (natürlich derselben Variable) beobachtet werden, welche Ausprägung einen positiven oder negativen Effekt auf die AV hat in Bezug auf die Referenzkategorie. Allerdings ist es schwierig die Koeffizienten zu interpretieren, weshalb hauptsächlich die Vorzeichen interessieren. Je nach grösse der Koeffizienten kann natürlich die Stärke der Wahrscheinlichkeit abgeschätzt werden.

Hier ein Ausschnitt aus der Vorlesungsfolie 7 S.15. Umkreist ist der i. Befehl und die Auswirkungen. Die Variable Konfliktfähigkeit wird in 4 Ausprägungen aufgeteilt, wobei die Kategorie 1 fehlt, da sie als

### Logit-Regression

```
xi: logit v12 skills_o i.avoidingconflict age male edu skills_w married ///
      integration_w i.socialisation

Logistic regression              Number of obs =      404
                                LR chi2(13)      =      97.25
                                Prob > chi2      =      0.0000
                                Pseudo R2       =      0.1798

Log likelihood = -221.7443

-----+-----
```

	v12	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	skills_o	.3354312	.2022732	1.66	0.097	-.061017 .7318793
	i.avoiding`2	.6219241	.3340039	-1.86	0.063	-1.27656 .0327116
	i.avoiding`3	-.959656	.3329812	-2.88	0.004	-1.612287 -.3070247
	i.avoiding`4	-.955472	.4630142	-2.06	0.039	-1.863038 -.048056
	age	.041363	.0121876	3.39	0.001	.0174758 .0652502
	male	.1045498	.2471451	0.42	0.672	-.3798457 .5889454
	edu	.4562227	.2746553	1.66	0.097	-.0820918 .9945372
	skills_w	.1804842	.1514733	1.19	0.233	-.1163981 .4773664
	married	-.2687854	.2634606	-1.02	0.308	-.7851588 .2475879
	integratio`w	.1669247	.1821937	0.92	0.360	-.1901684 .5240178
	i.socialis`1	-.6622193	.5415029	-1.22	0.221	-1.723546 .399107
	i.socialis`3	.4386501	.295321	1.49	0.137	-.1401685 1.017469
	i.socialis`4	1.617407	.2985163	5.42	0.000	1.032325 2.202488
	_cons	-3.517166	.8960673	-3.93	0.000	-5.273426 -1.760907

Referenzkategorie ausgeschlossen wurde. Die Koeffizienten mit deren Vorzeichen geben die Richtung der Wahrscheinlichkeit an, dass ein Ereignis eintritt. In diesem Beispiel sinkt die Wahrscheinlichkeit an politischen Diskussionen teilzunehmen mit sinkender Streitkultur.

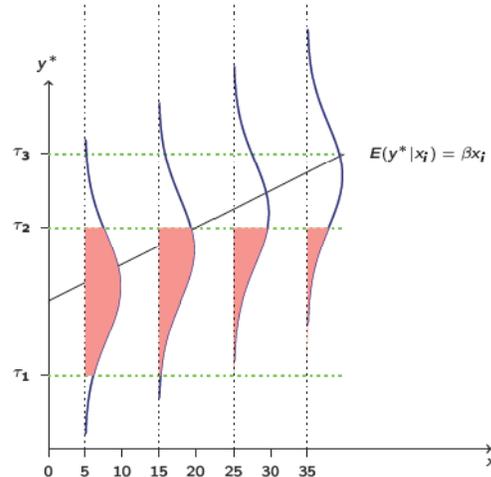
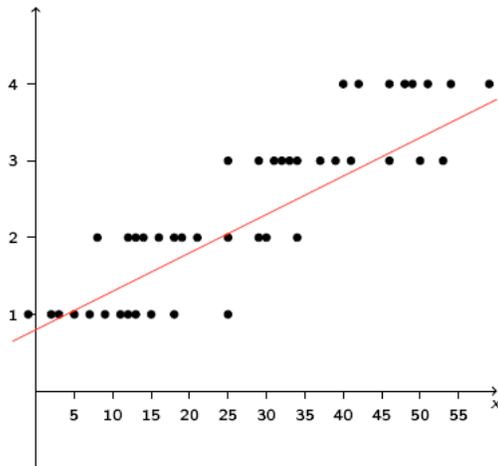
Nach der Logit-Regression wurden die Vorhergesagten Wahrscheinlichkeiten berechnet. Der Vorteil bei dieser Methode ist, dass die Wahrscheinlichkeit für ein Eintreten (oder Nichteintreten) eines Ereignisses anhand bestimmter Kriterien berechnet werden können. So kann z.B

betrachtet werden wie sich die Wahrscheinlichkeit von Männern zu Frauen wechselt. Somit können Ausprägungen ausfindig gemacht werden, welche stark auf ein Ereignis einwirken können. Solche Ausprägungen werden natürlich schon in der Logit-Regression ersichtlich, bei welcher allerdings eben keine Wahrscheinlichkeiten für ein Eintreten angezeigt werden.

## 9.4.2009, Vorlesung 8

### Ordered Logit und Anwendungsbeispiele

- Erweiterung des binären Logit-/Probit-Regressionsmodells, wenn abhängige Variable ordinal skaliert
- Problem des linearen Regressionsmodell, vgl. rote Linie
  - Regressionslinie reicht über Wertebereich der Abhängigen hinaus
  - Interpretation der Steigung der Regressionslinie ist unklar, da Werte auf der Linie, die von der Abhängigen gar nicht gemessen werden (Zwischenräume zwischen zwei Kategorien)



- Latente Variable, die kontinuierlich gemessen wird als Link zur ordinalen Abhängigen
  - ⇒ mehr Schwellenwerte ( $t_0, t_1, t_2$  etc.) für Übergang von einer Antwortkategorie zur nächsten (Anzahl Schwellenwert: Antwortkategorien - 1)
  - ⇒ Berechnung von Wahrscheinlichkeit für z.B.  $y=1$  oder  $y=2$  (vgl. rote Fläche, bei  $x=5, 15$  etc.) für latente Variable
- Wahrscheinlichkeit für eine Kategorie von  $y$  unabhängig von Werten des Achsenabschnitts ( $\alpha$ ) und Schwellenwerten, d.h. für verschiedene Werte des Achsenabschnitts können entsprechende Werte der Schwellenwerte ( $t$ ) eingesetzt werden
  - ⇒  $\alpha$  oder  $t$  muss = 0 gesetzt werden zur Identifikation des Modells
- **Logit:** Berechnung der logarithmierten Chance (Wahrscheinlichkeit/Gegenwahrscheinlichkeit), dass man Werte bis zu einem bestimmten  $y$  wählt, z.B.  $y=2$  als Grenze: Wahrscheinlichkeit darüber und darunter
  - ⇒ Dichotomisierung für jede Kategorie von  $y$
- **Probit:** „Paralleles Regressionsmodell“
  - Unterscheidung der kumulativen Wahrscheinlichkeiten (Summe der roten Flächen) für Werte auf der latenten Skala bis und mit zum nächsten Schwellenwert ( $t$ ) nur durch den Achsenabschnitt, nicht durch Steigung
  - ⇒ **für alle Antwortkategorien derselbe Koeffizient**
  - ⇒ z.B. eine abhängige Variable mit 4 Antwortkategorien hat 3 verschiedene kumulative Wahrscheinlichkeitsfunktionen mit derselben Steigung, aber jeweils nach rechts verschoben, sie sind also parallel
- Logit-Modell wird auch „Proportional Odds“ genannt, da „odds“ der beiden Modelle (z.B. logarithmierte Chancen für  $y=2$  und  $y=3$ ) proportional zueinander sind
  - ⇒ Logit-Koeffizient unabhängig von Antwortkategorie
- **Interpretation der rohen Logit-Koeffizienten:**
  - Logarithmierte Chance, dass die höheren Antwortkategorien anstatt bis und mit der interessierenden Kategorie gewählt wird:

Beschreibung der Veränderung der logarithmierten Chance, die Antwortkategorien 2, 3 oder 4 anstatt die 1 zu wählen, oder 3, 4 anstatt 1,2 oder 4 anstatt 1, 2, 3 (logarithmierte Chance für 4 ist um 0.25 grösser/kleiner (bei -0.25) anstatt für 2,3,4)

- **Signifikanz**

Wie bei OLS Vergleich von Koeffizient und Standardfehler, z- bzw. p-Wert

Cut-Points: Schwellenwerte der latenten kontinuierlich gemessenen Variable, Signifikanz nicht sehr wichtig, allenfalls Rekodierung der Schwellenwerte

- **Modellgüte:** Anteil korrekt vorhergesagter Werte/Beobachtungen anhand der Cut-Points

- **PRE-Mass:** vgl. binäres logit-Modell, jedoch eher Median basiertes PRE-Mass, da bessere Vorhersage mit Modus als bei vollständigem Modell

- **Vorhergesagte Wahrscheinlichkeit:**

für  $y=2$   $\Pr(\text{kumulierte Fläche für } t_2) - \Pr(\text{kumulierte Fläche für } t_1)$

für  $y=3$   $\Pr(\text{kumulierte Fläche für } t_3) - \Pr(\text{kumulierte Fläche für } t_2 \text{ und } t_1)$

## 23.4.2009, Vorlesung 9

### Multinomial Logit

- Gleiche Punktwolke wie im ordinalen Logit-Modell (vgl. Abbildung vorher), jedoch gilt Annahme der „proportional odds“ nicht mehr
- ⇒ Für jede Kategorie ein andere Steigung  $\beta$  und Achsenabschnitt  $\alpha$ ,  $\Pr(y=j|x_i)=\alpha_j+\beta_j x_i$ , lineares Modell kein guter link, da Werte negativ und grösser 1
- ⇒ lineare Term auf rechter Seite exponentieren, damit Wahrscheinlichkeit nicht mehr negativ:  $e^{\alpha_j+\beta_j x_i}$
- ⇒ Normalisierung: Dividieren durch Summe der  $e^{\alpha_j+\beta_j x_i}$  für jede Kategorie, damit Summierung der Wahrscheinlichkeiten =1
- ⇒ Identifikation: eigentlich für jede Ausprägung von  $y$  ein eigenes binäres Logit-Regressionsmodell, wobei logarithmierte Chance einer Kategorie vs. eine der anderen untersucht wird, daher Bildung einer Referenzkategorie, Koeffizienten  $\alpha$  und  $\beta =0$  für erste Kategorie
- **Interpretation der logit-Koeffizienten**  
„Wenn die Unabhängige um eine Einheit erhöht wird, dann verändert sich die logarithmierte Chance (der Logit) der Kategorie  $j$  vs. der ersten Kategorie(Referenzkategorie) der abhängigen Variable um  $\beta_j$ .“  
Nicht nur eine Unabhängige, sondern mehrere: für jede Kategorie (ausser Referenzkategorie) und jede Unabhängige ein eigener Koeffizient  $\beta$
- **Besonderheit:** auch Differenz der beiden Logit-Koeffizienten kann interpretiert werden
- ⇒ **Kontraste:**  $\beta_{kj} - \beta_{kn}$  „Eine Erhöhung der Unabhängigen  $k$  um eine Einheit verändert den Logit, logarithmierte Chance, der Kategorie  $j$  vs. der Kategorie  $n$  der abhängigen Variable um  $\beta_{kj} - \beta_{kn}$ .“
- **Signifikanz:** genügt nicht jeder Koeffizient einzeln anzuschauen, sondern Signifikanz des gemeinsamen Effekts all der zu einer Variable gehörenden Koeffizienten testen
- ⇒ **Loglikelihood-Ratio-Test (LR-Test):** Vergleich der Likelihood-Funktionen eines vollständigen und eines restringierten Modells (ohne zu untersuchende Variable); Angabe, ob Unterschied signifikant (Differenz  $\chi^2$  verteilt: Vergleich mit kritischem Wert bei gegebener Anzahl Freiheitsgraden)
- Kombination von Alternativen: wenn die Odds der Kategorie  $j$  vs. die Kategorie  $n$  der abhängigen Variable durch kein der unabhängigen Variablen signifikant beeinflusst wird (d.h. wenn alle  $k$  Koeffizienten die zu diesen Odds gehören nicht signifikant sind), dann geht man davon aus, dass die Alternativen/Kategorien  $j$  und  $n$  nicht voneinander unterschieden werden können
- ⇒ LR-Test, um Nullhypothese (keine Signifikanz, kein Einfluss) zu testen, dass Alternativen nicht voneinander unterschieden werden können
- Herleitung des multinomial Logit-Modell aus der „discrete Choice“-Theorie:  
Individuen sind Nutzenmaximierer, jede Kategorie der abhängigen Variable mit individuellem Nutzen, jedoch *fehlerbehaftet*:  $u_i = \mu_i + \varepsilon_i$  (Fehler)
- ⇒ Die Wahrscheinlichkeit eine bestimmte Kategorie/Alternative zu wählen, hängt davon ab, ob der Nutzen aus dieser Kategorie grösser ist als der Nutzen einer beliebig anderen Kategorie
- ⇒ Der Nutzen ist die latente Variable
- ⇒ Multinomial Logit-Modell resultiert, wenn die Fehler  $\varepsilon_i$  identisch und unabhängig verteilt sind
- ⇒ Odds/Chancen aus zwei Alternativen müssen immer dieselben bleiben, egal welche andere Alternativen sonst noch zur Wahl stehen, d.h. die zur Wahl stehenden Alternativen müssen möglichst plausibel unterscheidbar sein und durch die Subjekte unabhängig voneinander ausgewählt werden können
- ⇒ **Annahme der Unabhängigkeit von irrelevanten Alternativen (IIA)**

- ⇒ Oft problematische Annahme, v.a. bei Analyse von Umfragedaten zur Wahl politischer Parteien (wenn z.B. zwei Parteien ähnliche Ideologien haben, wird eine der beiden gewählt auf Kosten der anderen)
- ⇒ Zwei Arten von Tests für IIA-Annahme:
  - Hausmann-Test: Ausschließung einzelner Alternativen und testen, ob die Koeffizienten und ihre Varianz sich derart verändern, dass dies als eine signifikante Verletzung der IIA-Annahme verstanden werden muss, wenn  $\chi^2$  negativ, Nullhypothese nicht verworfbar
  - Small-Hsiao-Test: zufällige Aufteilung der Stichprobe in 2 Hälften und dem Vergleich der Likelihood-Werte aus geschätzten Modellen basierend auf diesen 2 Stichproben sowie einer weiteren, aus welcher Beobachtungen mit der fraglichen Antwortkategorie entfernt wurden
- Alternativen zum multinomial Logit-Modell
  - Multinomial Probit:
    - Alternative, wenn IIA-Annahme verletzt,
    - Unabhängigkeit der Alternativen nicht mehr zwingend, da Fehler der Nutzenfunktion einer multivariaten Normalverteilung folgt und Fehler korreliert sein können,
    - Höhere Flexibilität, aber schwierigere Berechnung der Parameter
    - Probit oder logit? Theoretische Überlegung
  - Conditional Logit:
    - Multinomial Logit Analyse nur, wenn alle Unabhängigen Unterschiede zwischen Entscheidungssubjekten messen, jedoch auch Variable berücksichtigen, wie sich Entscheidungsgegenstände für ein Subjekt unterscheiden, d.h. wie weit ein Subjekt von Gegenstand entfernt (z.B. ideologischer Abstand von verschiedenen Kandidaten)
      - ⇒ Für jedes Subjekt so viele Beobachtungen wie Antwortkategorien, wobei diese Werte verschieden sind für ein und dasselbe Subjekt
      - ⇒ Eigenschaften der Alternative, die für jedes Subjekt unterschiedlich wahrgenommen wird, interessiert
      - ⇒ Einfluss dieser Eigenschaft auf die Entscheidung zwischen Alternativen unabhängig von den Alternativen selbst, d.h. ein Subjekt ist von jedem Kandidaten unterschiedlich weit entfernt, aber diese Entfernung hat denselben Einfluss auf die Entscheidung unabhängig vom Kandidaten
  - **STATA**: abhängige hat z.B. 5 Ausprägungen, eine davon wird als Referenzkategorie verwendet, für jede Ausprägung der Abhängigen eigene List im Stata, Interpretation der Koeffizienten als Effekt (Erhöhung/Reduzierung) auf die logarithmierten Chance der einen Kategorie/Ausprägung vs. Referenzkategorie