Methodik der multiplen linearen Regression

Sibel Aydemir

Statistisches Amt, Direktorium
Landeshauptstadt München

Regressionsanalyse: Schritt für Schritt

Schritt 1 Modellbildung
Auswahl der erklärenden Variablen

Schritt 2 Schätzung der Regressionsgerade

Schritt 3 Wie gut fittet die Regressionsgerade?

Schritt 4 Ist das Gesamtmodell brauchbar?

Schritt 5 Ist der Einfluss der erklärenden Variablen statistisch signifikant?

Schritt 6 Welche Variablen sind zur Erklärung der Zielvariable tatsächlich erforderlich?
Variablenselektion

Schritt 1:

Modellbildung

Schritt 1: Modellbildung

SB-Nr.	Wahlbeteiligung gesamt	Alleinerziehende in %	HH-Größe	RK %	EV %	Sonst. %	Arbeitslose in % 18-65	Ausländer in %	Eingebürgerte in %	Aussiedler in %	mit Migr.-hint. in %	ohne Migr.-hint. in %	SGB II Bedarfsgemeinsch. in % der HH	SGB II Personen in % der Einw.
111	66,9%	2,6%	1,73	44,1%	15,6%	40,2%	4,7%	17,2%	5,2%	10,3%	32,7%	67,3%	3,6%	3,3%
121	71,6%	2,8%	1,57	42,9%	14,8%	42,3%	4,7%	19,8%	5,4%	5,8%	31,0%	69,0%	6,1%	5,3%
122	69,0%	2,0%	1,49	49,0%	11,6%	39,4%	2,1%	22,1%	5,9%	4,7%	32,7%	67,3%	3,2%	2,6%
131	78,8%	3,6%	2,14	48,5%	17,6%	33,9%	3,5%	18,6%	4,0%	11,3%	33,9%	66,1%	9,9%	6,3%
132	52,7%	4,2%	2,28	50,1%	17,3%	32,6%	4,6%	20,2%	5,8%	11,3%	37,3%	62,7%	9,9%	6,5%
141	80,2%	2,0%	2,11	59,0%	14,8%	26,2%	2,1%	2,8%	3,7%	4,7%	11,2%	88,8%	2,3%	1,3%

- **Spezifikation der abhängigen Variable *y***
- **Auswahl der erklärenden Variablen $x_1,\ldots,x_n$ aufgrund theoretischer Vorüberlegungen**
- **Einschränkungen in der Auswahl, da nicht immer alle potentiellen Einflussvariablen verfügbar bzw. messbar sind**
- **Beachte: Ergebnis der Regressionsanalyse hängt von der Auswahl der unabhängigen Variablen ab**

Beispiel: Variable „mit Migrationshintergrund“ wird alternativ zu den 3 Variablen „Ausländer“, „Eingebürgerte“ und „Aussiedler“ verwendet

	Modell 1:	**Modell 2:**
abhängige Variable	Wahlbeteiligung	Wahlbeteiligung
unabhängige Variablen	Alleinerziehende Haushaltsgröße Römisch-katholisch Evangelisch Arbeitslose SGBII Bedarfsgemeinschaften SGBII Personen Mit Migrationshintergrund	Alleinerziehende Haushaltsgröße Römisch-katholisch Evangelisch Arbeitslose SGBII Bedarfsgemeinschaften SGBII Personen Ausländer Eingebürgerte Aussiedler
signifikante Variablen	SGBII Bedarfsgemeinschaften Mit Migrationshintergrund Römisch-katholisch	SGBII Bedarfsgemeinschaften Ausländer Aussiedler Evangelisch

Schritt 2:

Schätzung der Regressionsgerade

Schätzung der Regressionsgerade

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + e$$

- Schätzung der Regressionskoeffizienten $\beta_0, \beta_1, \ldots, \beta_n$ über KQ-Methode
- Interpretation des Regressionskoeffizienten β_i im multiplen Regressionsmodell: β_i gibt den Einfluss der Variablen x_i bei Konstanthaltung des Einflusses aller anderen erklärenden Variablen wieder

Beispiel: Interpretation der Regressions-koeffizienten im multiplen linearen Modell

	Modell 1	β-Koeffizient
abhängige Variable	Wahlbeteiligung	
unabhängige Variablen	Alleinerziehende	
	Haushaltsgröße	
	Römisch-katholisch	
	Evangelisch	
	Arbeitslose	
	SGBII Bedarfsgemeinschaften	
	SGBII Personen	
	Mit Migrationshintergrund	
signifikante Variablen	Römisch-katholisch	- 0,26
	SGBII-Bedarfsgemeinschaften	- 0,83
	Mit Migrationshintergrund	- 0,64

Erhöhung des Anteils der Personen mit Migrations-hintergrund um 1% führt zu einem Rückgang der Wahl-beteiligung um 0,6% bei Konstanthaltung der Anteile der Variablen Röm.-kath. und SGBII-BG

Schritt 3:

Wie gut fittet die Regressionsgerade?

Maße für die Modellgüte

- **Bestimmtheitsmaß R^2**
- **Akaikes Informationskriterium *AIC***
- **Schwarz‘sche Bayes Kriterium (*SBC* oder *BIC*)**
- **Mallows Cp**

Das Bestimmtheitsmaß R^2

$$\text{Bestimmtheitsmaß } R^2 = \frac{\text{erklärte Varianz}}{\text{Gesamtvarianz}}$$

- **R^2 = Anteil der durch das Regressionsmodell erklärten Varianz an der Gesamtvarianz**
- **R^2 nimmt Werte zwischen 0 und 1 an**
- **Je näher R^2 an 1 liegt, desto besser „passt“ die Regressionsgerade**

Beispiel: $R^2 = 0{,}7$

D.h. 70% der Variation der abhängigen Variable *y* sind auf die erklärende Variable *x* zurückzuführen.

Die nicht erklärte Varianz von 30% resultiert u.a. durch nicht berücksichtigte Variablen.

Zur Interpretation des Bestimmtheitsmaßes R^2

- **Hohes R^2 sagt nichts über die Erklärungskraft der einzelnen Koeffizienten aus**
- **Höhe des R^2 hängt stark von den Daten ab**
- **R^2 wächst mit zunehmender Anzahl von erklärenden Variablen**
- **Abhilfe: Korrigiertes Bestimmtheitsmaß (R^2 adjusted)**

Beispiel

Abhängigkeit des Bestimmtheitsmaßes von der Anzahl der Variablen

	Modell 1:	R^2	R^2 **korrigiert**
abhängige Variable	Wahlbeteiligung		
unabhängige Variablen	Alleinerziehende Haushaltsgröße Römisch-katholisch Evangelisch Arbeitslose SGBII Bedarfsgemeinschaften SGBII Personen Mit Migrationshintergrund	**0,880**	**0,871**
signifikante Variablen	SGBII Bedarfsgemeinschaften Mit Migrationshintergrund Römisch-katholisch	**0,875**	**0,872**

Hohes Bestimmtheitsmaß R^2 nur Zufall?

Deskriptive Statistik → **Induktive Statistik**

Schritt 4:

Ist das Gesamtmodell brauchbar?

Ist das Gesamtmodell brauchbar?

Spezifizierte Regressionsgleichung ist unbrauchbar

Gesamttest auf Signifikanz
Prüfverfahren: F-Test

- H_0 : $\beta_i = 0$, für alle i=1,…n
- D.h. keine der berücksichtigten unabhängigen Variablen $x_1,\ldots,x_n$ besitzt einen Einfluss auf die abhängige Variable *y*
- Wird H_0 abgelehnt, so hat mindestens eine der erklärenden Variablen $x_1,\ldots,x_n$ einen Einfluss auf *y*
- Faustregel: H_0 wird abgelehnt, falls *F*-Wert > 10

Schritt 5:

Ist der Einfluss der erklärenden Variablen statistisch signifikant?

Welche Variablen sind statistisch signifikant?

Prüfung der Regressionskoeffizienten $\beta_1,\ldots,\beta_n$ auf statistische Signifikanz

Prüfverfahren: t-Test

- H_0 : $\beta_i = 0$, i=1,…,n
- D.h. die Variable x_i besitzt keinen Einfluss auf die abhängige Variable *y*
- Faustregel: H_0 wird abgelehnt, falls *t*-Wert > |2|
- Signifikanzniveau (Irrtumswahrscheinlichkeit) < 0.05

Nicht-signifikante Variablen

- **Vorsicht bei der Interpretation: Ist eine erklärende Variable nicht signifikant, heißt das nicht unbedingt, dass sie keinen Einfluss auf die abhängige Variable *y* besitzt**

- **Korrelieren zwei (oder mehrere) unabhängige Variablen, so kann es sein, dass in der multiplen Regression eine Variable sich nicht durchsetzen kann, da sie keine zusätzliche Information zur Regression beiträgt**

- **Variablen, die in der multiplen Regression nicht-signifikant sind, können in der einfachen linearen Regression durchaus einen signifikanten Einfluss zeigen**

Beispiel

Multiple lineare Regression vs. einfache lineare Regression

	Modell 1:	Multiple lin. Regression	Einfache lin. Regression
abhängige Variable	Wahlbeteiligung		
unabhängige Variablen	Alleinerziehende	--	signifikant
	Haushaltsgröße	--	--
	Römisch-katholisch	signifikant	signifikant
	Evangelisch	--	--
	Arbeitslose	--	signifikant
	SGBII Bedarfsgemeinschaften	signifikant	signifikant
	SGBII Personen	--	signifikant
	Mit Migrationshintergrund	signifikant	signifikant

„Wichtigkeit“ einer erklärenden Variable

- **Vorsicht: Das Signifikanzniveau ist nicht ausreichend, um Aussagen über die „Wichtigkeit“ einer erklärenden Variable machen zu können**
- **Die „Wichtigkeit“ einer Variable lässt sich an der (standardisierten) Koeffizientenschätzung erkennen**
- **Denn: Eine im Vergleich „weniger signifikante“ Variable, kann evtl. mehr zur Erklärung/Vorhersage von *y* beitragen (→ Beispiel)**

Beispiel: „Wichtigkeit“ der erklärenden Variablen

	Modell 2:	p-Wert	β	β stand.
abhängige Variable	Wahlbeteiligung			
unabhängige Variablen	Alleinerziehende			
	Haushaltsgröße			
	Römisch-katholisch			
	Evangelisch			
	Arbeitslose			
	SGBII Bedarfsgemeinschaften			
	SGBII Personen			
	Ausländer			
	Eingebürgerte			
	Aussiedler			
signifikante Variablen	SGBII Bedarfsgemeinschaften	.029	- 0.62	- 0.22
	Ausländer	.007	- 0.28	- 0.19
	Aussiedler	.000	- 0.75	- 0.47
	Evangelisch	.001	0.37	0.12

P(Evangelisch) < P(SGBII-BG)

Aber: Variable SGBII-BG hat einen nahezu doppelt so großen Einfluss auf die Wahlbeteiligung als Variable Evangelisch

Angenommen, die Regressionsschätzung zeigt:

Einige Variablen besitzen einen statistisch signifikanten Einfluss auf y, andere Variablen besitzen keinen Einfluss.

Wie geht es weiter?

1.Möglichkeit: Schätzung eines Endmodells nur mit den signifikanten erklärenden Variablen

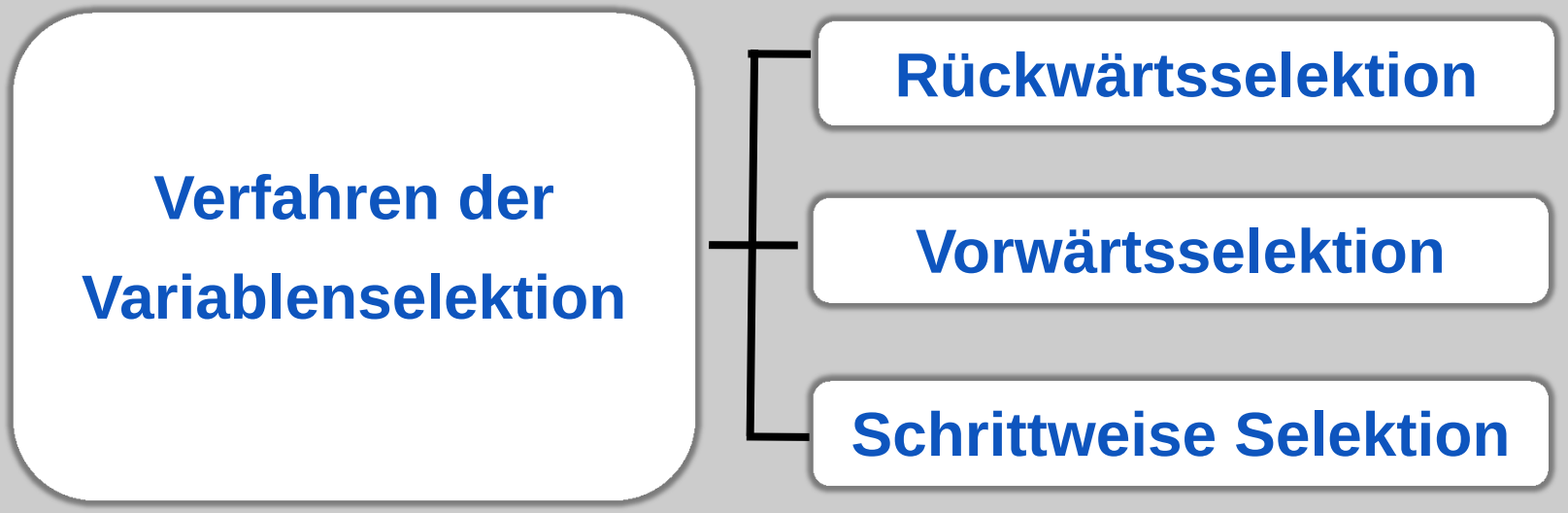
2.Möglichkeit: Schätzung eines Endmodells mittels Variablenselektion

Vorteil: Multikollinearität wird berücksichtigt

Schritt 6:

Welche Variablen sind zur Erklärung der Zielvariable tatsächlich erforderlich?

Variablenselektion



Rückwärtsselektion

- **Start: vollständiges Modell mit allen unabhängigen Variablen**
- **Sukzessive werden diejenigen Variablen entfernt, die zum geringsten Rückgang des Bestimmtheitsmaßes R^2 führen würden.**
- **Stopp: Verfahren bricht ab, falls sich beim Entfernen einer (bzw. der nächsten) Variable das Bestimmtheitsmaß R^2 signifikant verkleinert.**

Beispiel: Rückwärtsselektion (RS)

	Modell 1	p-Wert	RS	p (Endmodell)
abhängige Variable	Wahlbeteiligung			
unabhängige Variablen	Alleinerziehende	0,296	3	
	Haushaltsgröße	0,384	2	
	Römisch-katholisch	0,002		0,001
	Evangelisch	0,635	1	
	Arbeitslose	0,171	5	
	SGBII Bedarfsgemeinschaften	0,081		0,000
	SGBII Personen	0,362	4	
	Mit Migrationshintergrund	0,000		0,000

signifikante Variablen	Römisch-katholisch Mit Migrationshintergrund	⇄	Römisch-katholisch SGBII Bedarfsgemeinschaften Mit Migrationshintergrund

Vorwärtsselektion

- **Start:** Modell ohne unabhängige Variablen, also $y = \beta_0$
- Bestimme diejenige erklärende Variable, die mit y am stärksten korreliert ist und berechne das Bestimmtheitsmaß R^2 .
- Ist R^2 signifikant, wird diese Variable in das Modell aufgenommen.
- In weiteren Schritten werden sukzessive die Variablen ins Modell aufgenommen, die zum größten Anstieg von R^2 führen.
- **Stopp:** Verfahren bricht ab, falls sich bei Hinzunahme einer weiteren Variable das Bestimmtheitsmaß R^2 nicht signifikant vergrößert.

Beispiel: Vorwärtsselektion (VS)

	Modell 2	R^2	VS
abhängige Variable	Wahlbeteiligung		
unabhängige Variablen	Alleinerziehende		
	Haushaltsgröße		
	Römisch-katholisch		
	Evangelisch	0,877	4
	Arbeitslose		
	SGBII Bedarfsgemeinschaften	0,765	1
	SGBII Personen		
	Ausländer	0,866	3
	Eingebürgerte		
	Aussiedler	0,850	2

Signifikante Variablen bei VS

Lineares Regressionsmodell ohne Variablenselektion

Signifikante Variable: Aussiedler

Schrittweise Selektion

- **Kombination aus Vorwärts- und Rückwärtsselektion**

- **Es wird eine Vorwärtsselektion durchgeführt, bei der nach jedem Schritt untersucht wird, ob bei Entfernen einer der bisher aufgenommenen Variablen das Bestimmtheitsmaß signifikant abnehmen würde (=Rückwärtsselektion).**

Variablenselektion

Vorwärtsselektion, Rückwärtsselektion und schrittweise Selektion führen (meist) zum selben Ergebnis.

Überblick

Regressionsmodelle

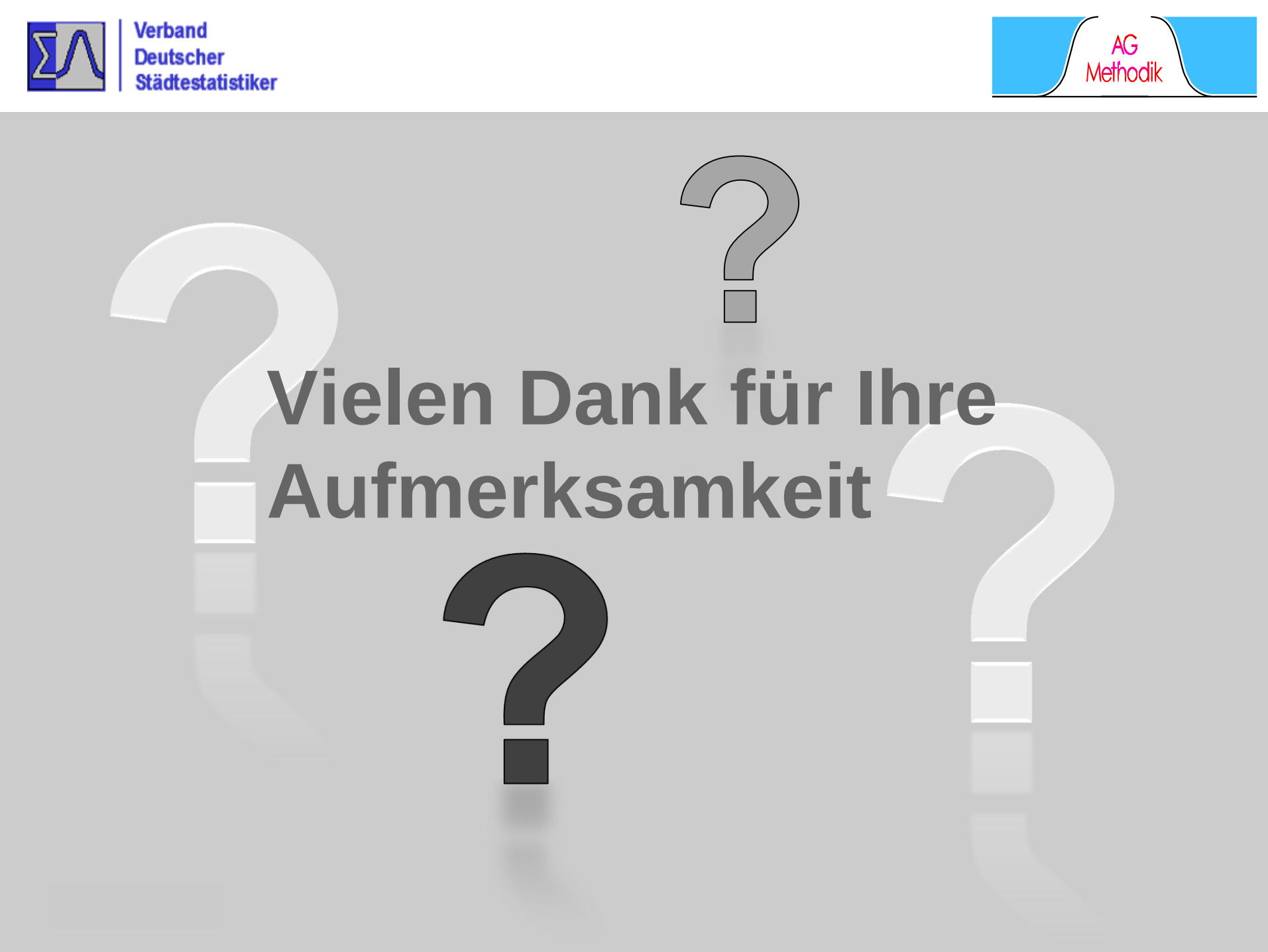
Regressionsmodelle

Ziel

Untersuchung des Zusammenhanges zwischen einer abhängigen Variable und mehreren unabhängigen Variablen

Die bekanntesten Regressionsmodelle

- **Lineare Regression (einfach / multipel)**
  Abhängige Variable *y* = metrisch
- **Logistische Regression**
  Abhängige Variable *y* = binär (0/1-Kodierung)
- **Cox-Regression**
  Abhängige Variable *y* = Zeitdauer
  (z.B. Wohndauer, Ehedauer)



Vielen Dank für Ihre Aufmerksamkeit

Schritt 3: Wie gut fittet die Regressionsgerade?

Bestimmtheitsmaß R^2

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{erklärteVarianz}}{\text{Gesamtvarianz}}$$