Regression II

Statistik I

Sommersemester 2009

Zum Nachlesen

- Agresti: 9.1-9.4
- Gehring/Weins: 8
- Schumann: 8.1-8.2

And now for something completely different ...

- Fehler in den Folien zur Berechnung von V
- R ist hier das Minimum der Spalten/Zeilen in der zugrundeliegenden Kreuztabelle

Wozu Modelle?

Modelle . . .

- Ermöglichen kompakte *Beschreibung* und
- Erleichtern *Verständnis* der Zusammenhänge
- Ermöglichen ggf. *Schluß auf eine Grundgesamtheit*
- Gestatten *Prognosen* für *zukünftige* und *hypothetische* Fälle
- Wenn Modell korrekt spezifiziert und Annahmen realistisch
- Lineare Einfachregression = einfachstes aller Modelle; Muster

Was sind die Bestandteile des linearen Regressionsmodells?

- Systematischer Teil
  - Konstante
  - Unabhängige Variable(n) mit Regressionsgewicht(en) („Steigung“)
- Stochastische Komponente (zufällige Einflüsse)
- Konditionale Varianz um konditionale Mittelwerte (Modell vs. Beobachtungen)

Wie wird das Regressionsmodell interpretiert?

- Welcher y Wert ist für einen bestimmten x-Wert zu erwarten?
- Vergleich mit realen Werten
- Prognose für zukünftige und
- Hypothetische Werte

Wie rechnet man das?

- Tabelle konstruieren mit:
  1. x- und y-Werten $\longrightarrow$ Mittelwerte berechnen
  2. Einfachen Abweichungen x und y
  3. Quadrierten Abweichungen x und y
  4. Abweichungsprodukten
- Dann in Formeln einsetzen

Frauenerwerbsquote und Anteil Frauen im Parlament

Country	Women Parl. y	Fem. Labour Force (% of male) x
Austria	32	75
Germany	31	76
Denmark	37	84
Sweden	45	87

Was tun?

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32					
	76	31					
	84	37					
	87	45					
Σ							
$\bar{x};\bar{y}$							

1. Summen/Mittelwerte

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32					
	76	31					
	84	37					
	87	45					
Σ	322	145					
$\bar{x};\bar{y}$	80.5	36.3					

2. einfache Abweichungen

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32	-5.5	-4.3			
	76	31	-4.5	-5.3			
	84	37	3.5	0.8			
	87	45	6.5	8.8			
Σ	322	145	0	0			
$\bar{x};\bar{y}$	80.5	36.3					

3. quadrierte Abweichungen/SAQ

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32	-5.5	-4.3	30.2	18.1	
	76	31	-4.5	-5.3	20.2	27.6	
	84	37	3.5	0.8	12.2	0.6	
	87	45	6.5	8.8	42.2	76.6	
Σ	322	145	0	0	106	123.9	
$\bar{x};\bar{y}$	80.5	36.3					

4. Abweichungsprodukte/SAP

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32	-5.5	-4.3	30.2	18.1	23.4
	76	31	-4.5	-5.3	20.2	27.6	23.6
	84	37	3.5	0.8	12.2	0.6	2.6
	87	45	6.5	8.8	42.2	76.6	56.9
Σ	322	145	0	0	106	123.9	106.5
$\bar{x};\bar{y}$	80.5	36.3					

5. Berechnung von b

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32	-5.5	-4.3	30.2	18.1	23.4
	76	31	-4.5	-5.3	20.2	27.6	23.6
	84	37	3.5	0.8	12.2	0.6	2.6
	87	45	6.5	8.8	42.2	76.6	56.9
Σ	322	145	0	0	106	123.9	106.5
$\bar{x};\bar{y}$	80.5	36.3					

$$b = \frac{SAP}{SAQ_x} = \frac{106.5}{106} \approx 1$$

6. Berechnung von a

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32	-5.5	-4.3	30.2	18.1	23.4
	76	31	-4.5	-5.3	20.2	27.6	23.6
	84	37	3.5	0.8	12.2	0.6	2.6
	87	45	6.5	8.8	42.2	76.6	56.9
Σ	322	145	0	0	106	123.9	106.5
$\bar{x};\bar{y}$	80.5	36.3					

$b = \frac{SAP}{SAQ_x} = \frac{106.5}{106} \approx 1$
$a = \bar{y} - b \times \bar{x} = 36.5 - 1 \times 80.5 = -44$

7. Finale

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32	-5.5	-4.3	30.2	18.1	23.4
	76	31	-4.5	-5.3	20.2	27.6	23.6
	84	37	3.5	0.8	12.2	0.6	2.6
	87	45	6.5	8.8	42.2	76.6	56.9
Σ	322	145	0	0	106	123.9	106.5
$\bar{x};\bar{y}$	80.5	36.3					

$$b = \frac{SAP}{SAQ_x} = \frac{106.5}{106} \approx 1$$
$$a = \bar{y} - b \times \bar{x} = 36.5 - 1 \times 80.5 = -44$$
$$\hat{y} = -44 + 1 \times x$$

8. Schätzungen

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})\times(y-\bar{y})$
	75	32	-5.5	-4.3	30.2	18.1	23.4
	76	31	-4.5	-5.3	20.2	27.6	23.6
	84	37	3.5	0.8	12.2	0.6	2.6
	87	45	6.5	8.8	42.2	76.6	56.9
Σ	322	145	0	0	106	123.9	106.5
$\bar{x};\bar{y}$	80.5	36.3					

$b = \frac{SAP}{SAQ_x} = \frac{106.5}{106} \approx 1$
$a = \bar{y} - b \times \bar{x} = 36.5 - 1 \times 80.5 = -44$
$\hat{y} = -44 + 1 \times x$

Austria: 31; Germany: 32; Denmark: 40; Sweden: 43

Wie gut paßt das Modell zu den Daten?

- Qualitätskriterium: SAQ_y der vom Modell geschätzten Werte
- Sollte möglichst klein sein
- In Relation zu den SAQ_y insgesamt
- PRE-Maß:
  - Wie stark reduziert die Kenntnis von x den Vorhersagefehler gegenüber einer naiven Vorhersage
  - $\bar{y}$
- $R^2 = r^2 = \eta^2$

Wie berechnet man R^2?

- Möglichkeit 1: Zerlegung der quadrierten Abweichung zwischen beobachtetem Wert und Mittelwert in
  - Quadrierte Abweichung zwischen Mittelwert und vorhergesagtem Wert („erklärte Abweichung“) und
  - quadrierte Abweichung zwischen vorgesagtem und beobachtetem Wert („unerklärte Abweichung“)
  - Summe einfacher Abweichungen $= 0$; statt dessen Zerlegung der Varianzen
- Möglichkeit 2: r berechnen und quadrieren

Möglichkeit 1

- Entweder die Abweichungen zwischen vorhergesagten Werten und Mittelwert (erklärte Abweichung)
- Oder Abweichung zwischen vorhergesagten und beobachteten (unerklärte Abweichung) Werten berechnen und quadrieren

	y_i	$\hat{y}$	$(y_i - \hat{y})$	$(\hat{y} - \bar{y})$	$(y_i - \bar{y})$	$(y_i - \hat{y})^2$	$(\hat{y} - \bar{y})^2$	$(y_i - \bar{y})^2$
	32	31	1	-5.2	-4.2	1	27.6	18.1
	31	32	-1	-4.2	-5.2	1	18.1	27.6
	37	40	-3	3.8	0.8	9	14.1	0.6
	45	43	2	6.8	8.8	4	45.6	76.6
Σ						15 (unerkl.)	105.3 (erkl.)	122.8 (ges.)

Möglichkeit 1

R^2

$$R^2 = \frac{\text{erklärte SAQ}}{gesamteSAQ} = 1 - \frac{\text{unerklärte SAQ}}{\text{gesamte SAQ}}$$
$$= \frac{105.3}{122.8} = 1 - \frac{15}{122.8} = 0.86$$

Möglichkeit II

- Nach der (hoffentlich bekannten) Formel Pearson's r berechnen:
- $r = \frac{SAP}{SAQ_x \times SAQ_y} = \frac{106.5}{\sqrt{106 \times 123.9}} = \frac{106.5}{114.6} = 0.93$
- $r^2 = R^2 = 0.93^2 = 0.86$

Was ist die Bedeutung von R^2

- R^2 häufig als Kriterium für absolute Qualität mißverstanden („Wettbewerb um höheres R^2“)
- R^2: relative Bedeutung von zufälligen vs. systematischen Einflüssen
- R^2 hängt ab von Art der Daten/Anwendungsgebiet (Aggregatdaten)
- Streuung von x kann R^2 massiv beeinflussen
- Hohes R^2 durch „Scheinkorrelation“?
- R^2 nicht über Anwendungsgebiete vergleichbar. Vergleich über Datensätze?

Was ist die Bedeutung von R^2

- R^2 häufig als Kriterium für absolute Qualität mißverstanden („Wettbewerb um höheres R^2“)
- R^2: relative Bedeutung von zufälligen vs. systematischen Einflüssen
- R^2 hängt ab von Art der Daten/Anwendungsgebiet (Aggregatdaten)
- Streuung von x kann R^2 massiv beeinflussen
- Hohes R^2 durch „Scheinkorrelation“?
- R^2 nicht über Anwendungsgebiete vergleichbar. Vergleich über Datensätze?
- R^2: relevant für Vergleich konkurrierender Modelle

Was ist der „Root Mean Squared Error“

- Standard Error of the Estimate
- Abweichungen von $\hat{y}$ (Vorhersagefehler) haben Mittelwert von null, aber Varianz $\rightarrow$ „unerklärte SAQ“; $\Sigma(y_i - \hat{y})^2$
- Je mehr Varianz, desto unsicherer die Vorhersage
- Varianz der Vorhersage: $\frac{\Sigma(y_i - \hat{y})^2}{n}$
- Wurzel daraus: Standardabweichung der Vorhersage: $\sqrt{\frac{\Sigma(y_i - \hat{y})^2}{n}}$

Möglichkeit 1

R^2

$$R^2 = \frac{\text{erklärte SAQ}}{gesamteSAQ} = 1 - \frac{\text{unerklärte SAQ}}{\text{gesamte SAQ}}$$
$$= \frac{105.3}{122.8} = 1 - \frac{15}{122.8} = 0.86$$

Wie berechnet man den Root Mean Squared Error?

- Unerklärte SAQ (15)
- Durch Zahl der Fälle (Zahl der Fälle minus d.f. wenn Schätzung für Grundgesamtheit)
- Wurzel ziehen
- $\sqrt{\frac{15}{4}} = \sqrt{3.75} = 1.94$

Wie berechnet man den Root Mean Squared Error?

- Unerklärte SAQ (15)
- Durch Zahl der Fälle (Zahl der Fälle minus d.f. wenn Schätzung für Grundgesamtheit)
- Wurzel ziehen
- $\sqrt{\frac{15}{4}} = \sqrt{3.75} = 1.94$
- Interpretation: Bei unser Prognose machen wir stets Fehler
  - Fehler haben Mittelwert von null
  - Aber eine Standardabweichung von fast 2 Prozentpunkten
  - Wenn Fehler näherungsweise normalverteilt in 95% aller Prognosen
  - Fehler im Bereich von ± 4 Prozentpunkten
  - Prognosen relativ ungenau

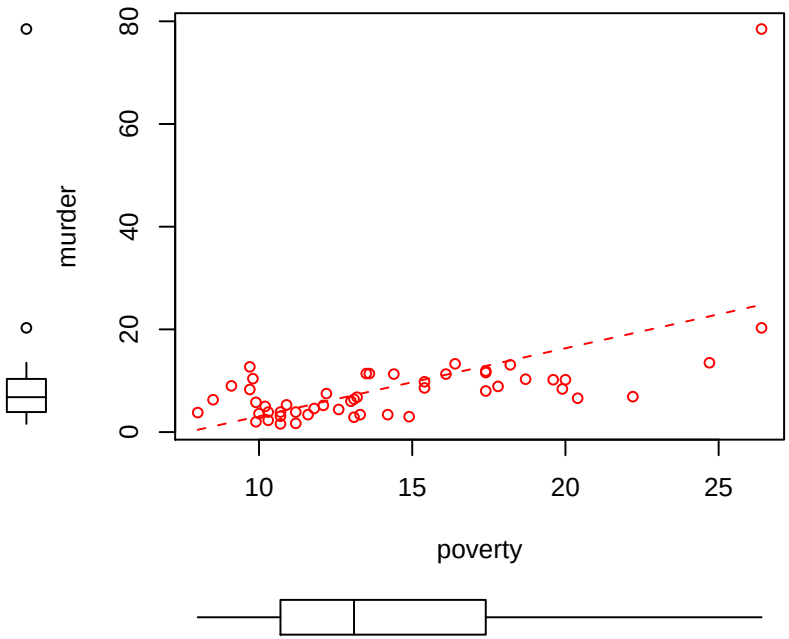
Was bedeutet das zusammengenommen?

- Relative starker linearer Zusammenhang, d. h. *relativ* geringe Streuung um Regressionslinie
- R^2 bzw. r^2 standardisieren Zusammenhang auf Wertebereich ± 1
- Trotzdem inhaltlich relevante *absolute* Abweichungen
- RMSE/SEE haben Wertebereich/Einheit von y

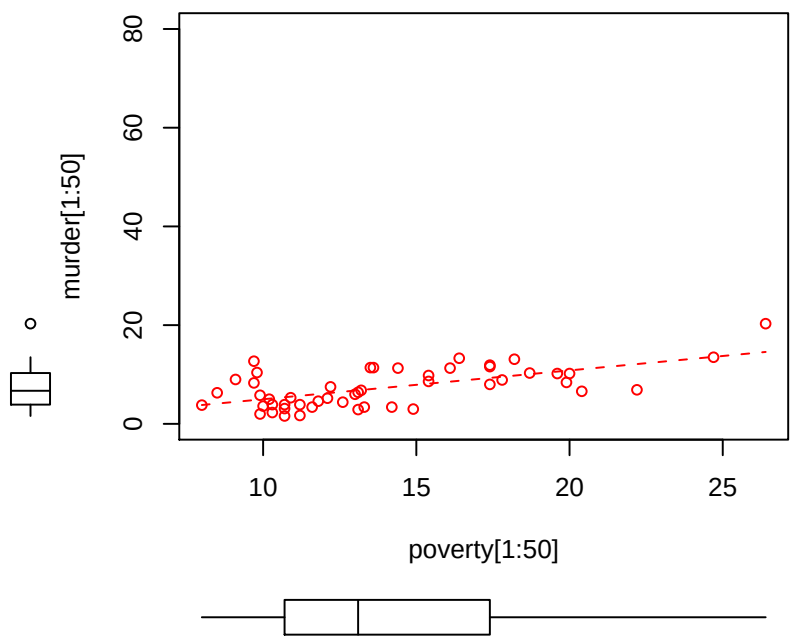
Was bedeutet das zusammengenommen?

- Relative starker linearer Zusammenhang, d. h. *relativ* geringe Streuung um Regressionslinie
- R^2 bzw. r^2 standardisieren Zusammenhang auf Wertebereich ± 1
- Trotzdem inhaltlich relevante *absolute* Abweichungen
- RMSE/SEE haben Wertebereich/Einheit von y
- Beide Maße angeben – vollständigeres Bild davon, wie gut Modell paßt

Modell mit/ohne Washington D. C.



$$\hat{y} = -10.14 + 1.32 \times x$$



$$\hat{y} = -0.86 + 0.58 \times x$$

Gibt es Probleme mit dem Regressionsmodell?

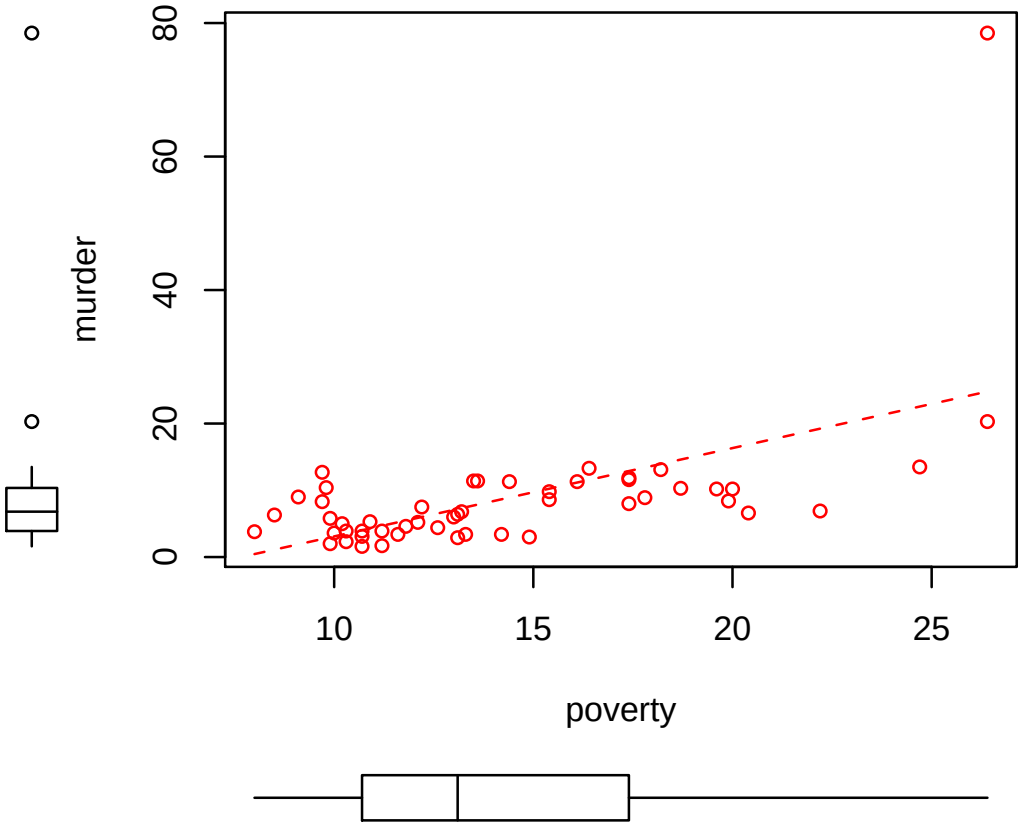
1. Gibt es Fälle, für die das Modell sehr schlecht paßt?
2. Haben einzelne Fälle sehr viel Einfluß auf das Modell?

- Ein weites Feld

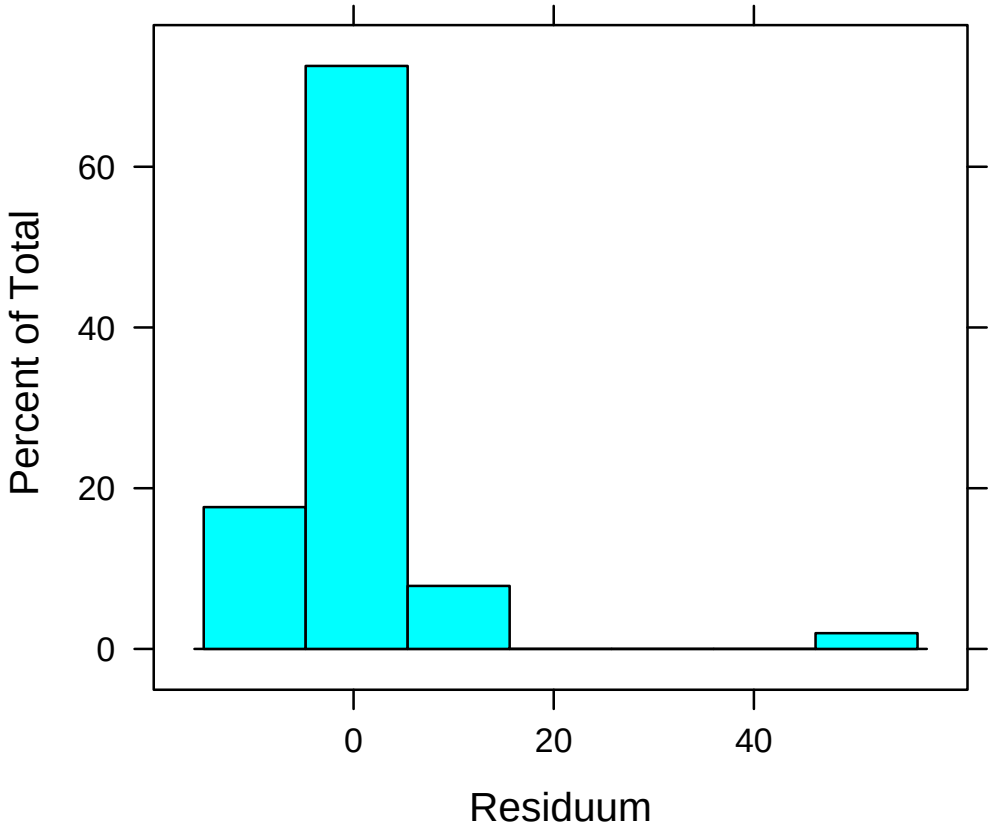
Residuen

- Residuum = Differenz zwischen gemessenem y und $\hat{y}$
- Annahmen über Verteilung der Residuen $\longrightarrow$ ungewöhnlich große Residuen
- Standardisierte bzw. „studentisierte“ Residuen
- Im Beispiel: Washington D. C.
  - Residuum 53.7
  - Nächstes Residuum 12.3
- Nach jedem Maßstab außergewöhnlich

Armut/Mord



Verteilung der Residuen



Welche Fälle sind „einflußreich“

- Um (im negativen Sinn) einflußreich zu sein, muß Fall ungewöhnlich sein (Ausreißer)
- Wie würde Modellschätzung ohne diesen Fall aussehen?
- Drei Konstellationen denkbar
  1. Ungewöhnlich, aber mit geringem Einfluß: wenig y, durchschnittliches x
  2. Fall ungewöhnlich, aber auf Regressionsgeraden (viel x, viel y, aber auf Linie)
  3. Ungewöhnlich und einflußreich: viel x, überproportional viel oder wenig y
- In Fällen 2 und 3 kann Extrapolation gefährlich sein

Welche Fälle sind einflußreich?

- Einfluß auf die Koeffizienten von zwei Faktoren bestimmt:

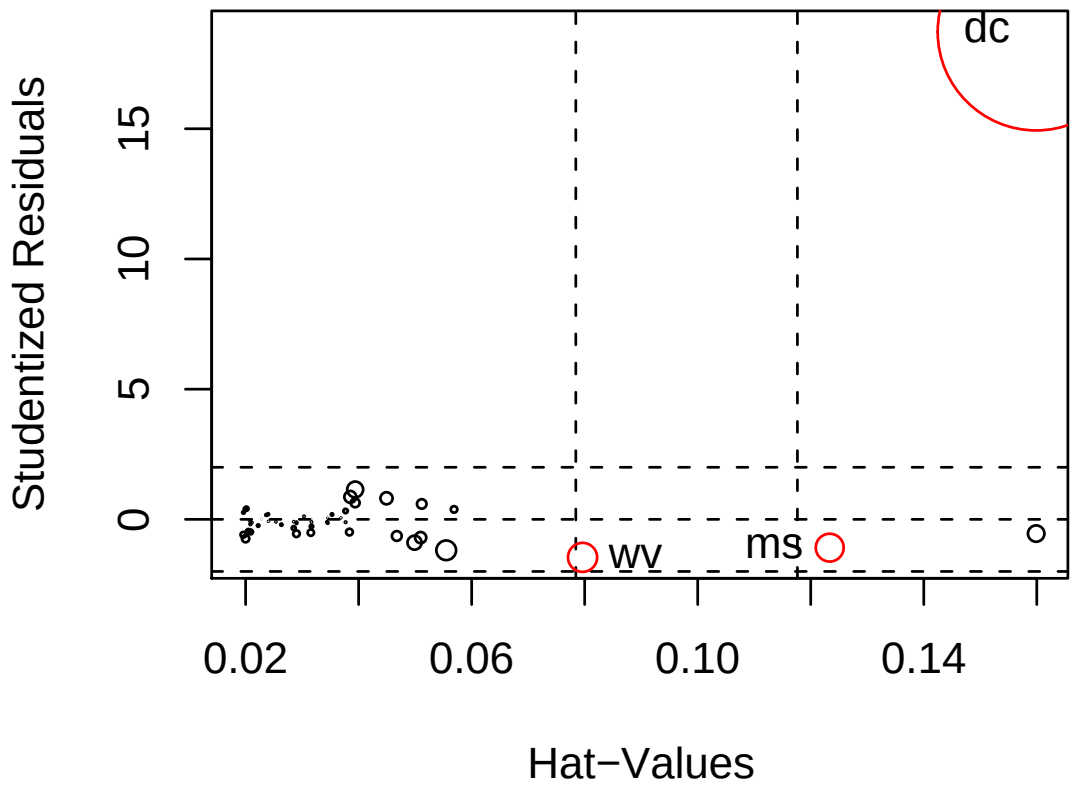
1. „Hebelwirkung“ (leverage) – hat Fall das Potential, Koeffizienten zu beeinflussen, ist er weit vom Zentrum der Verteilung entfernt?
2. „Ungewöhnlichkeit“ (discrepancy, outlyingness) – ist Fall „in einer Linie“ mit anderen oder nicht?

- Einfluß auf Koeffizienten = Produkt beider Faktoren
- Reihe von Maßzahlen und Faustregeln, um Einfluß abzuschätzen

Hat-Values, Residuen und Cook's D

- (Studentisierte) Residuen – bereits bekannt
- Hat-Value
  - Wieviel tragen x-Wert(e) eines Falles zu den prognostizierten y-Werten bei
  - Bei linearer Einfachregression Maß für Entfernung vom Mittelwert von x
  - Multiple Regression: komplizierter
- Cook's D – äquivalent zum Produkt aus „Hebel“ und „Ungewöhnlichkeit“

Armut/Morde: Ausreißer und Hebel



Was bedeutet das? Konsequenzen?

- Modell ohne Washington D. C. unterscheidet sich massiv von Modell mit
- Washington: einflußreicher Ausreißer
- Fragen/Konsequenzen:
  - Meßfehler?
  - Gleicher Mechanismus für alle Staaten $\longrightarrow$ Modell mit Washington schätzen
  - Anderer Mechanismus $\longrightarrow$ ausschließen
  - Fehlende Variablen???

Warum mehr als eine unabhängige Variable?

- Modell – radikale Vereinfachung
- 1:1 Beziehung x – y zu starke Vereinfachung?
- (Fast) alle sozialen Prozesse multikausal

Warum mehr als eine unabhängige Variable?

- Modell – radikale Vereinfachung
- 1:1 Beziehung $x - y$ zu starke Vereinfachung?
- (Fast) alle sozialen Prozesse multikausal
- Zweite mögliche Erklärung für Mordquote: ethnische Homogenität
  - Spannungen zwischen ethnischen Gruppen
  - Höhere Kriminalität innerhalb von Minderheiten
  - Höhere Wahrscheinlichkeit für schwarze Verdächtige verurteilt zu werden

Wie sieht ein Modell mit zwei unabhängigen Variablen aus?

Multivariate Regression

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

- y als Funktion von x_1 (z. B. Armut) und x_2 (z. B. Homogenität)
- Beide Variablen haben unabhängig voneinander Einfluß
- Additives Zusammenwirken
- Nicht gegenseitige Verstärkung von Armut und ethnischen Konflikten

Was sind die Ergebnisse?

Multivariates Modell

$$\text{„Morde“} = 36.2 + 0.8 \times \text{„Armut“} - 0.46 \times \text{„Prozent Weiße“}$$

Was sind die Ergebnisse?

Multivariates Modell

$$\text{„Morde“} = 36.2 + 0.8 \times \text{„Armut“} - 0.46 \times \text{„Prozent Weiße“}$$

- Für einen Bundesstaat ohne weiße Bevölkerung und ohne Armut werden 36 Morde / 100 000 Einwohner erwartet
- („Prozent Weiße“ ein guter Indikator für Homogenität?)
- Für jeden Prozentpunkt mehr Armut werden 0.8 Morde mehr erwartet
- Für jeden Prozentpunkt mehr weiße Bevölkerung werden 0.46 Morde *weniger* erwartet

Was sind die Ergebnisse?

Multivariates Modell

$$\text{„Morde“} = 36.2 + 0.8 \times \text{„Armut“} - 0.46 \times \text{„Prozent Weiße“}$$

- Für einen Staat mit mittlerer Armut (13%) und mittlerem Anteil von Weißen (87%) werden ca. 33 Morde erwartet
- Für einen extrem armen (26.4%) Staat mit einer extrem gemischten (32%) Bevölkerung wie Washington D. C. werden ca. 43 Morde erwartet
- (Immer noch knapp die Hälfte weniger als beobachtet)