

# Multiple Regression

## Statistik II

Wiederholung  
Multivariate Zusammenhänge  
Multiple Regression  
Zusammenfassung

## Übersicht

Wiederholung

Literatur

Regression

Multivariate Zusammenhänge

Assoziation und Kausalität

Statistische Kontrolle

Multivariate Beziehungen

Inferenz

Multiple Regression

Das Multivariate Modell

Beispiel: Bildung und Verbrechen

Fit

Zusammenfassung

## Literatur für heute

- ▶ Agresti ch. 10
- ▶ Zur Nach- und Vorbereitung:
- ▶ Agresti ch. 11

## Literatur für nächste Woche

- ▶ Mason/Wolfinger: Cohort Analysis.
- ▶ Kish: Weighting: Why, When, and How?

## Daten/Kommandos für heute

- ▶ `net get from`  
`http://www.kai-arzheimer.com/Statistik-II/stata/`
- ▶ `net describe floridacrime`
- ▶ `net get floridacrime`

## Was ist Regression?

- ▶ Modellierung konditionaler Verteilung (Mittelwert und Streuung)
- ▶ Beschreibung vs. Inferenz
- ▶ Daten vs. Modellannahmen
- ▶ Bekanntestes und einfachstes Modell: lineare (Einfach)-Regression

## Was ist lineare Einfachregression?

- ▶ Der konditionale Mittelwert einer abhängigen Variablen  $y$  wird modelliert als
- ▶ lineare Funktion einer unabhängigen Variablen  $x$  und einer Konstanten
- ▶ Gemeinsame Verteilung von  $x$  und  $y$  als Punktwolke (Fehlervarianz) um eine gerade Linie
- ▶ Allgemeines Muster für viele andere statistische Modelle
- ▶ Bestimmung der Parameter durch OLS (Minimale quadrierte Abweichung in  $y$ -Richtung)

## Wie funktioniert OLS (ohne Mathematik)

- ▶ SAQ = Funktion(Daten, Parameterschätzungen)
- ▶ Daten sind gegeben
- ▶ Welche Parameterschätzungen machen SAQ möglichst klein (guter Fit, gute Schätzung)?
- ▶ Minimum der SAQ-Funktion suchen → 1. Ableitung auf null setzen, nach Konstante und Steigung auflösen
  1. Formeln aus Formelsammlung
  2. Alternativ: kompakte Matrixalgebra

## Was ist Kausalität?

- ▶  $X \rightarrow Y$
- ▶ Hypothetisch-kontrafaktisches Konzept von Kausalität
  - ▶  $X$  und  $Y$  an *einem Fall* messen
  - ▶ Realität für *diesen Fall* mit anderem Wert von  $X$  „wiederholen“ – Änderung von  $Y$ ?
  - ▶ In der Praxis nicht durchführbar, nur Annäherung an dieses Ideal
- ▶ Experiment:
  - ▶ Viele Objekte
  - ▶  $X$  von Forscherin variiert, zeitliche Reihenfolge klar
  - ▶ Vergleichbar bezüglich anderer Eigenschaften wg. zufälliger Aufteilung auf Experimental-/Kontrollgruppe
- ▶ Beobachtung/Befragung (ex post facto)
  - ▶ Viele Objekte
  - ▶ Keine Kontrolle über  $X$  (zeitliche Reihenfolge), keine zufällige Aufteilung

▶ Andere Eigenschaften nur „statistisch kontrollierbar“

Statistik II

Multiple Regression (8/33)

- ▶ Statistische Verfahren kein Ersatz für gutes Design

## Was setzt Kausalität voraus?

- ▶  $X \rightarrow Y$
1. (Theorie)
  2. Statistische Assoziation (Übergang deterministische/probabilistische Aussagen!)
  3. Richtige zeitliche Reihenfolge – in ex post facto Designs fast nicht zu prüfen
  4. **Ausschluß von Drittvariablen**

## Beispiel: Körpergröße und Mathematik-Leistung

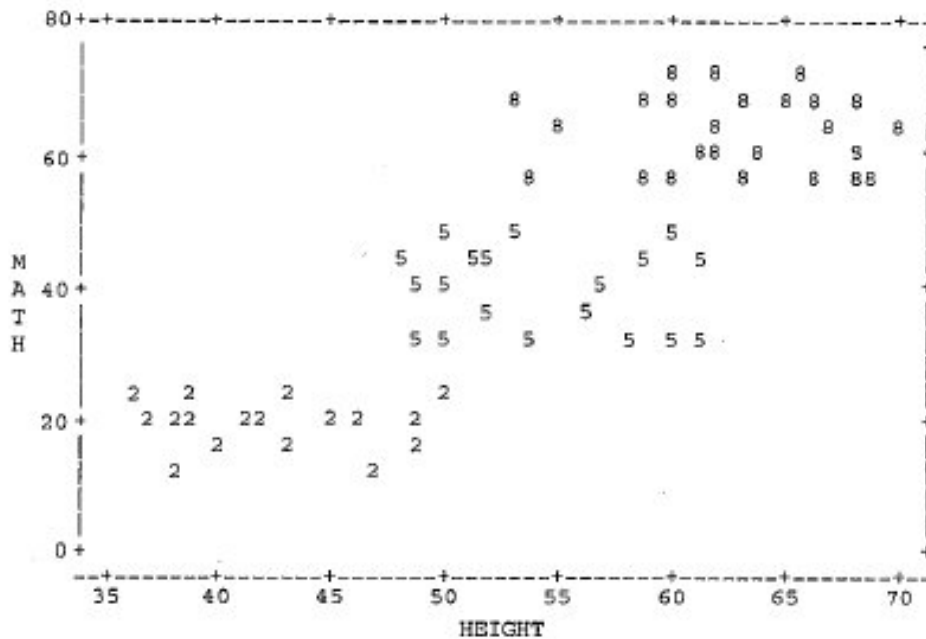


FIGURE 10.1: Printout Showing Relationship between Height and Math Achievement Test Score, with Observations Labeled by Grade Level. Students at a particular grade level have about the same age.

## Beispiel: Pfadfinder und Delinquenz

TABLE 10.1: Contingency Table Relating Scouting and Delinquency

		Delinquency		Total
		Yes	No	
Boy Scout	Yes	36 (9%)	364 (91%)	400
	No	60 (15%)	340 (85%)	400

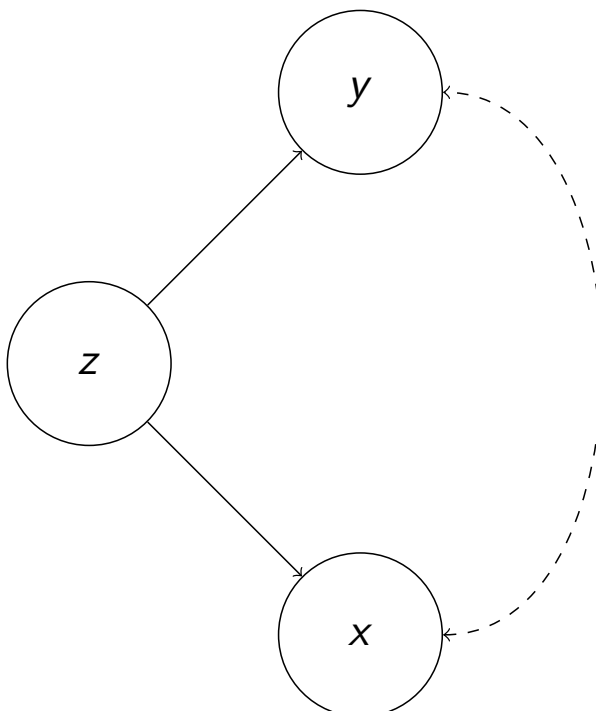
TABLE 10.2: Contingency Table Relating Scouting and Delinquency, Controlling for Church Attendance

Delinquency		Church Attendance					
		Low		Medium		High	
		Yes	No	Yes	No	Yes	No
Scout	Yes	10 (20%)	40 (80%)	18 (12%)	132 (88%)	8 (4%)	192 (96%)
	No	40 (20%)	160 (80%)	18 (12%)	132 (88%)	2 (4%)	48 (96%)

## Welche Beziehungen können zwischen drei Variablen bestehen?

1. „Scheinkorrelation“ / „scheinbare Non-Korrelation“
2. Mediatorvariable
3. Multiple Verursachung
4. Interaktion
5. ...

### 1. „Scheinkorrelation“



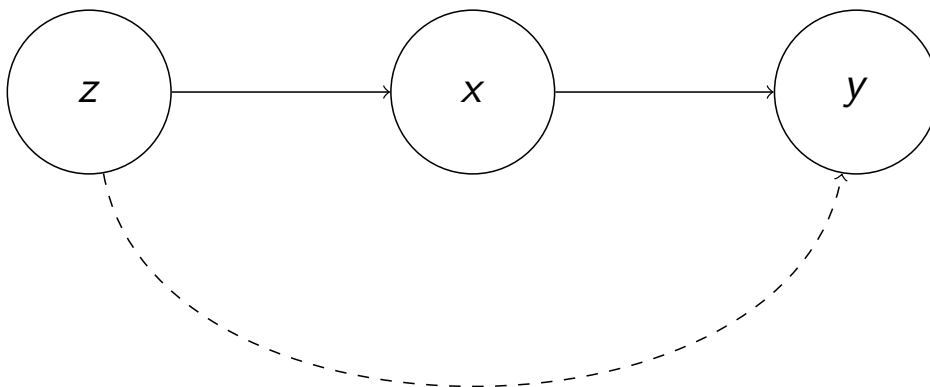
## „Scheinbare Non-Korrelation“ (Suppression)

- Kein Zusammenhang zwischen Bildung und Einkommen?

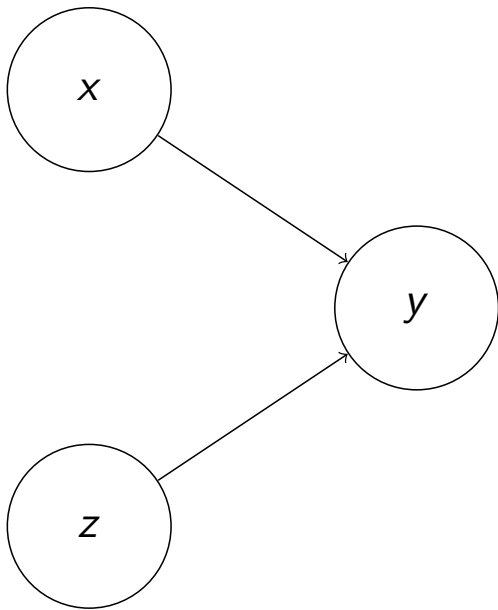
TABLE 10.4: Bivariate Tables Relating Education, Income, and Age

Education	Income		Age	Income		Age	Education	
	High	Low		High	Low		High	Low
High	250	250	High	350	150	High	150	350
Low	250	250	Low	150	350	Low	350	150

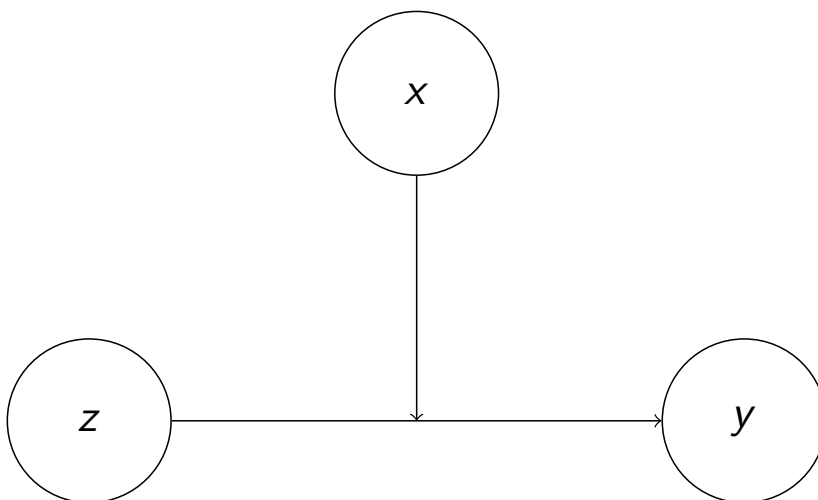
## 2. „Mediatorvariable“



### 3. „Multiple Verursachung“



### 4. „Interaktion“



## Schluß auf die Grundgesamtheit?

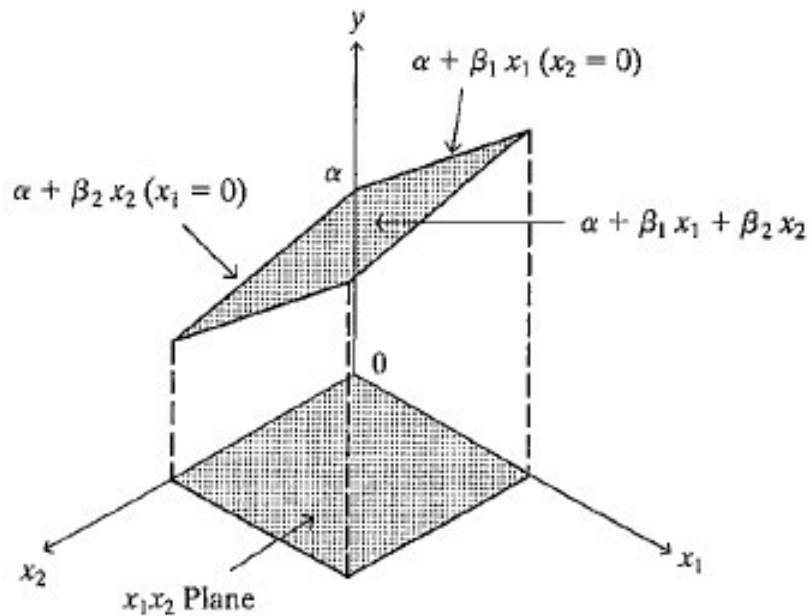
- ▶ Kontrolle für multivariate Beziehungen durch multivariate Modelle
- ▶ Inferenzen verfügbar

## Modell (zwei unabhängige Variablen)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- ▶ Wert von  $y$  von *beiden* unabhängigen Variablen beeinflusst
- ▶ Effekte linear (proportional) und additiv
- ▶ Effekte unabhängig voneinander
  - ▶  $\beta_1$  Effekt von  $x_1$  während  $x_2$  konstant gehalten wird (vgl. Pfadfinder-Tabelle)
  - ▶  $\beta_2$  Effekt von  $x_2$  während  $x_1$  konstant gehalten wird
- ▶ D. h. wechselseitige statistische Kontrolle

## Graphische Darstellung (zwei unabhängige Variablen)



## Beispiel: Bildung und Verbrechen

- ▶ 67 counties in Florida
- ▶ Sind counties mit höherem Niveau von formaler Bildung (% high school Absolventen) krimineller (mehr Verbrechen pro Einwohner)?

```
. reg c hs
```

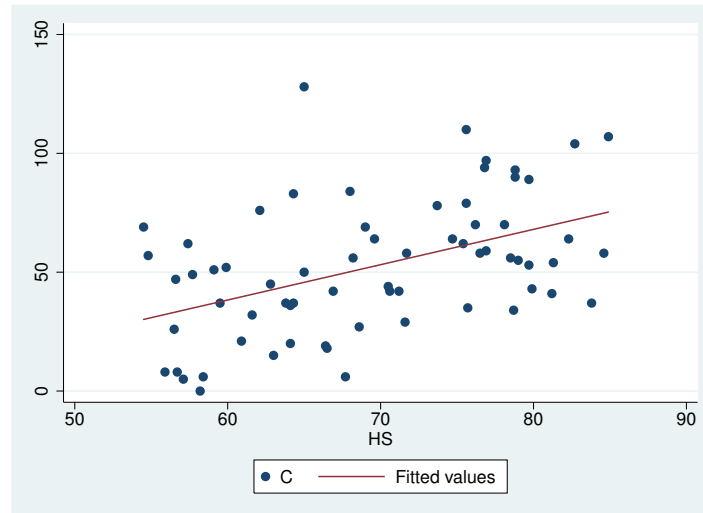
Source	SS	df	MS			
Model	11437.0945	1	11437.0945	Number of obs =	67	
Residual	41025.0249	65	631.154229	F( 1, 65) =	18.12	
Total	52462.1194	66	794.880597	Prob > F =	0.0001	
				R-squared =	0.2180	
				Adj R-squared =	0.2060	
				Root MSE =	25.123	

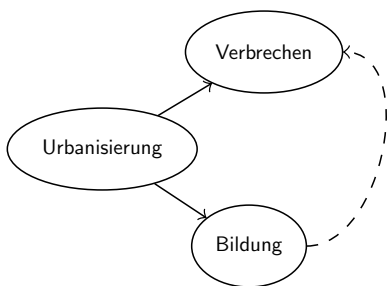
c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hs	1.485977	.3490777	4.26	0.000	.788821	2.183133
_cons	-50.85691	24.45065	-2.08	0.041	-99.68823	-2.025583

## Scatterplot + Regression

```
graph twoway (scatter c hs) (lfit c hs)
```



## „Scheinkorrelation“?



- ▶ Kontrolle: multiple Regression
- ▶  $\text{Verbrechen} = \alpha + \beta_1 \text{Bildung} + \beta_2 \text{Urbanisierung}$
- ▶ Effekt von Bildung für jedes denkbare Niveau von Urbanisierung
- ▶ Effekt von Urbanisierung für jedes denkbare Niveau von Bildung

## Regression in Stata

- ▶ Grundbefehl (reg)ress y x1 x2 ...
- ▶ Variablennamen können abgekürzt werden
- ▶ Jokerzeichen oder Bereiche für Variablen
- ▶ Ergebnis der letzten Regression → reg
- ▶ (Optionen mit Komma abtrennen)
- ▶ Postestimation (z. B. predict)

## In Stata...

```
. reg c hs u
```

Source	SS	df	MS			
Model	24731.6571	2	12365.8286	Number of obs =	67	
Residual	27730.4623	64	433.288473	F( 2, 64) =	28.54	
Total	52462.1194	66	794.880597	Prob > F =	0.0000	
				R-squared =	0.4714	
				Adj R-squared =	0.4549	
				Root MSE =	20.816	

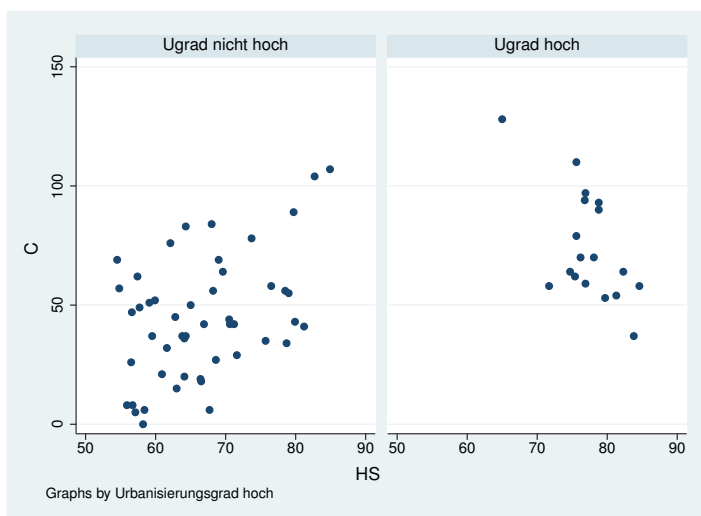
c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hs	-.5833773	.4724591	-1.23	0.221	-1.527223	.3604684
u	.6825014	.1232126	5.54	0.000	.436356	.9286469
_cons	59.11806	28.36531	2.08	0.041	2.45184	115.7843

- ▶ Urbanisierung hat einen starken positiven Effekt
- ▶ Bildung hat *negativen* Effekt
- ▶ *Partieller Effekt* (vs. bivariater Effekt)

## Partielle Effekte

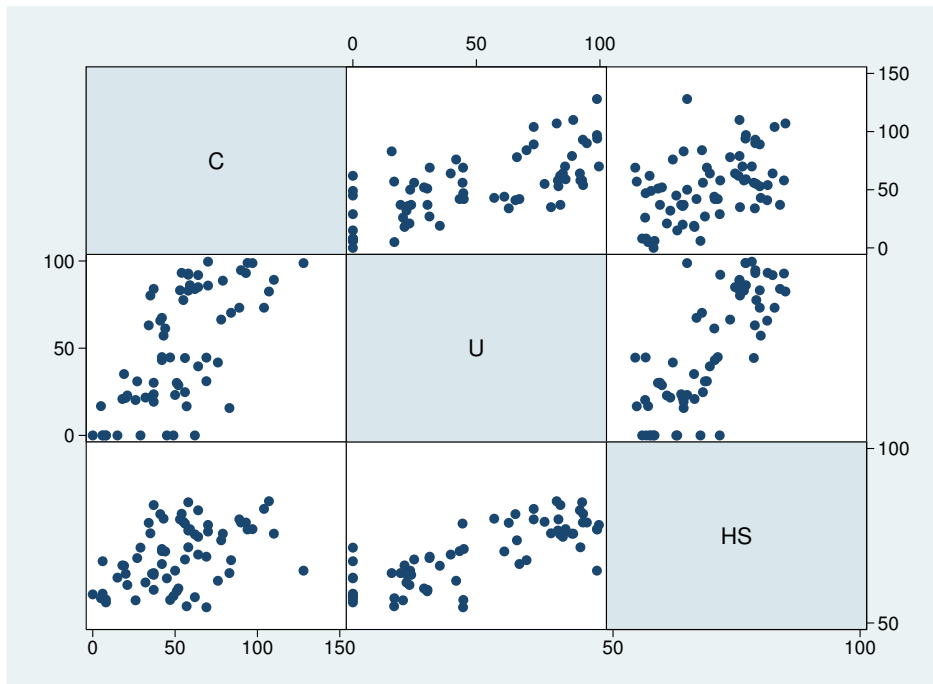
- ▶ Partielle Effekte = Effekt von Bildung
- ▶ Innerhalb einer Gruppe von counties mit identischem (aber beliebigem) Urbanisierungsgrad
- ▶ Schätzung über alle Niveaus von Urbanisierungsgrad
- ▶ Und umgekehrt
- ▶ Analog zur Betrachtung von Subgruppen im Pfadfinder-Beispiel
- ▶ Partielle Koeffizienten  $\neq$  Bivariate Koeffizienten wg. Korrelation zwischen unabhängigen Variablen
- ▶ Nicht beim Experiment

## Warum ist der partielle Effekt negativ?

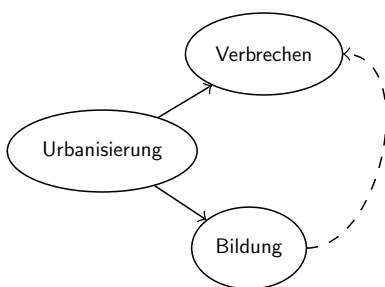


- ▶ Urbanisierungsgrad wird „konstant gehalten“

## Visualisierung: Matrixplot



## Partielle Regressionsplots



- ▶ (Added Variable Plot)
- ▶ Residuum: Differenz zwischen beobachtetem und geschätztem Wert
- ▶  $\text{Verbrechen} = \alpha_1 + \beta_1 \text{Urbanisierung} \rightarrow \text{Residuum} = \text{Verbrechen} \text{ abzüglich Effekt von Urbanisierung}$
- ▶  $\text{Bildung} = \alpha_2 + \beta_2 \text{Urbanisierung} \rightarrow \text{Residuum} = \text{Bildung} \text{ abzüglich Effekt von Urbanisierung}$
- ▶ Regression von Residuum 1 auf Residuum 2  $\rightarrow$  identisch mit partiellem Regressionskoeffizienten

## Root Mean Squared Error

- ▶ Verbrechen: 0 – 128; Residuum = Vorhersagefehler
- ▶ Residuum quadrieren und aufsummieren → SAQ
- ▶ SAQ/n= Mittlerer quadrierter Fehler
- ▶ Wurzel → RMSE
- ▶ Wie bei Einfachregression

```
. predict abweichung, resid
. gen aq=abweichung *abweichung
. sum aq
```

Variable	Obs	Mean	Std. Dev.	Min	Max
aq	67	413.8875	495.9546	.8899312	2568.242

```
. displ sqrt(414)
20.34699
```

## $R^2$

- ▶ Analog zur Einfachregression
  - ▶  $R$  = Korrelation zwischen vorhergesagten/beobachteten Wert bzw.
  - ▶ Gesamt SAQ (TSS) = Modell-SAQ (MSS) + Residuale SAQ (RSS)
  - ▶ (PRE-Interpretation)

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{MSS}{TSS} = \frac{\sum (y - \bar{y})^2 - \sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

- ▶ Kollinearität zwischen unabhängigen Variablen
- ▶ Adjusted  $R^2$

## $R/R^2$ von Hand ausrechnen

```
. predict yhat  
(option xb assumed; fitted values)  
. corr yhat c  
(obs=67)
```

	yhat	c
yhat	1.0000	
c	0.6866	1.0000

```
. display .69^2  
.4761
```

## Zusammenfassung

- ▶ Korrelation  $\neq$  Kausalität
- ▶ Multiple Regression
  - ▶ Keine Kontrolle über unabhängige Variable(n) (vs. Experiment)
  - ▶ (Schwacher) Ersatz für Randomisierung (Drittvariablenkontrolle)
- ▶ Partielle vs. bivariate Effekte
- ▶ Fit analog zu Einfachregression
- ▶ Inferenz für Koeffizienten?
- ▶ Nächste Woche: Agresti ch. 11