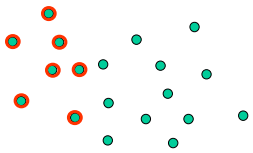
Repräsentativität und Unabhängigkeit

Ziel: Bestmögliche Erfassung der Eigenschaften der Grundgesamtheit

Problem: Beurteilung der Repräsentativität ist nur durch umfassende **Information** über die Grundgesamtheit möglich

Ansatz: Vergrößerung des Stichprobenumfangs



aber: "mehr Daten = mehr Information" gilt nicht automatisch!

=> je höher die neuen Daten mit den alten Daten korreliert sind, desto geringer ist der Informationsgehalt (= "Grad der Überraschung", "Grad der Unvorhersagbarkeit")

Korrelation, Regression

Korrelation: ***Enge*** des (linearen) Zusammenhangs

$$r = \frac{1}{s_x \cdot s_y} \cdot \frac{\sum_{i=1}^{n} (x - \bar{x}) \cdot (y - \bar{y})}{n}$$

Regression: ***Art*** des (linearen) Zusammenhangs

$$\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots$$

Zeitliche und Räumliche Autokorrelation

Generelle Beobachtung: Eng benachbarte (zeitlich, räumlich) Messungen liefern tendenziell ähnliche Werte

Autokorrelation: 1-dimensional (Zeitreihenanalyse)
Variogramm: 2(n)-dimensional (Geostatistik)

=> Dieselben Prinzipien, aus historischen Gründen allerdings unterschiedliche Begriffe

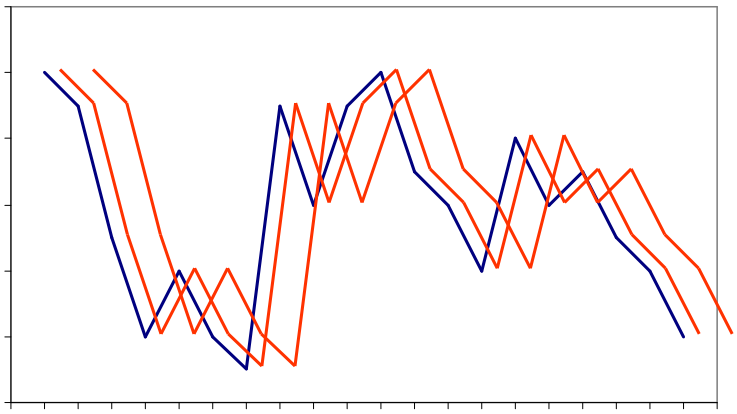
Zeitliche Korrelation

Korrelation: $$r = \frac{1}{s_x \cdot s_y} \cdot \frac{\sum_{i=1}^{n} (x - \bar{x}) \cdot (y - \bar{y})}{n}$$

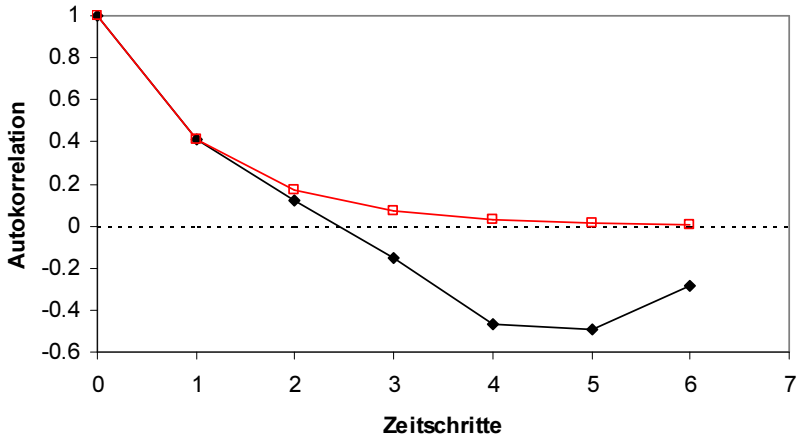
Autokorrelation: $$r_{ac} = \frac{1}{s_x \cdot s_x} \cdot \frac{\sum_{i=1}^{n} (f(x_i) - \overline{f(x)}) \cdot ([f(x_i + t)] - \overline{f(x)})}{n}$$

Kreuzkorrelation: $$r_{cc} = \frac{1}{s_x \cdot s_x} \cdot \frac{\sum_{i=1}^{n} (f(x_i) - \overline{f(x)}) \cdot ([g(x_i + t)] - \overline{g(x)})}{n}$$

Autokorrelation



Autokorrelationsfunktion



(i.d.R. äquidistante Zeitschritte Δt)

Räumliche Abhängigkeit: Variogramm

→ Erweiterung des Begriffs der Autokorrelation:

- ***n***-dimensionale Zusammenhänge (meist: ***n*** = 2)
- Äquidistanz der Datenpunkte nicht erforderlich
- inverse Darstellung: Zunahme der Varianz (Unabhängigkeit) als Funktion der Entfernung

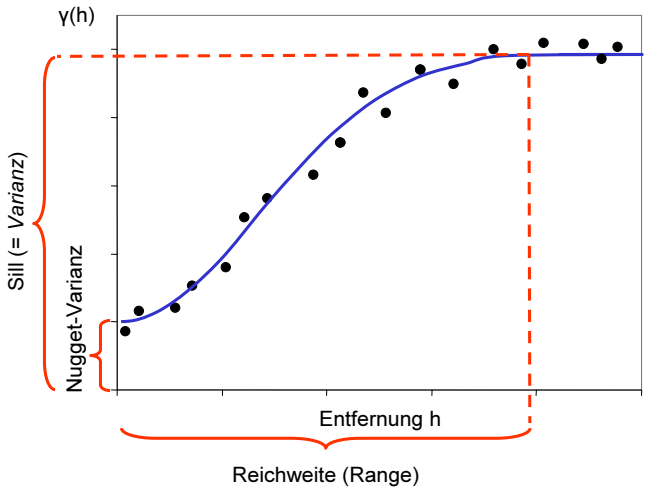
(Semi-)Variogramm-Funktion

Varianz: $\text{var} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left((f(x_i) - \overline{f(x)} \right)^2$

Autokorrelation: $r_t = \frac{1}{s_x \cdot s_x} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} (f(x_i) - \overline{f(x)}) \cdot ([f(x_i + t)] - \overline{f(x)})$

Semi-Varianz: $\gamma(h) = \frac{1}{2} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} [(f(x_i + h) - f(x_i)]^2$

(Semi-)Variogramm



Kreuz-Variogramm

Variogramm:

$$\gamma(h) = \frac{1}{2} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} [f(x_i + h) - f(x_i)]^2$$

$$= \frac{1}{2} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} [\{(f(x_i + h) - f(x_i)\} \cdot \{(f(x_i + h) - f(x_i)\}]$$

Kreuz-Variogramm:

$$\gamma_{f,g}(h) = \frac{1}{2} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} [\{f(x_i + h) - f(x_i)\} \cdot \{g(x_i + h) - g(x_n)\}]$$

Vorhersage-Modelle (Schätzer)

	zeitlich	**räumlich**
unabhängig von weiteren Variablen:		
- Modell:	Autoregression	Kriging
- Parametrisierung:	Autokorrelation	Semi-Variogramm
abhängig von weiteren Variablen:		
- Modell:	lin. Transferfunkt.	Co-Kriging
- Parametrisierung:	Kreuz-Korrelation	Kreuz-Variogramm

Multivariate Verfahren

	Lineare Regression	**Hauptkomponentenanalyse**	**Korrespondenzanalyse**	**Clusteranalyse**	**Diskriminanzanalyse**
Zweck:					
Vorhersage	x				
Dimensionsreduktion		x	x		
Klassifizierung				x	x
Eigenschaften:					
nicht-linear					
verteilungsfrei			x		
nominal skalierte Var.			x		x

Daten-Transformation

für die einzelnen Verfahren

Verfahren	Box-Cox-Transf.*	z-Transformation
multiple Regression	ja	nein
Hauptkomponentenanalyse	ja	ja
Korrespondenzanalyse	nein	ja (und min > 0)
Clusteranalyse	ja	ja
Diskriminanzanalyse	ja	ja
Selbstorganisierende Karte	nein	ja

**: falls nicht normalverteilt*

Regression

Regressand
(Kriteriumsvariable)

Regressoren
(Prädiktoren)

$$\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots$$

Regressionskoeffizienten

Regressionskoeffizient b_i: $b_{yx} = \frac{\mathrm{cov}(x,y)}{s_x^2}$

Bestimmung der Regressionsfunktion

Minimierung der Abweichungsquadrate (***least squares***):

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

Standardisierte Koeffizienten

β-Koeffizient: $\beta_i = \frac{s_{x_i}}{s_y} \cdot b_i$

= b_i-Koeffizienten der z-transformierten Variablen

=> Vergleichbarkeit der Koeffizienten verschiedener Regressoren

Güte des Modells

Wie gut ist das Modell?

- Bestimmtheitsmaß, korrigiertes r^2, totaler F-Test
- Konfidenzintervall des Regressanden
- Unkorreliertheit der Residuen (Durbin-Watson-Test)

Wie wichtig sind einzelne Regressoren?

- partieller F-Test
- Signifikanz / Konfidenzintervalle der Regressionskoeffizienten

Güte der Vorhersage

Bestimmtheitsmaß: $B = r^2$ (für lineare Regression)

= erklärte Varianz/Gesamtvarianz

= 1 - (nicht-erklärte Varianz/Gesamtvarianz)

Gesamtvarianz = erklärte Varianz + nicht-erklärte Varianz

$$\frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

korrigiertes (adjustiertes) Bestimmtheitsmaß:

$$r^2_{korr} = r^2 - \frac{k\cdot(1-r^2)}{n-k-1}$$

n: Anzahl der Werte
k: Anzahl der Regressoren

Standardschätzfehler des Regressanden

$$SE(\hat{y}) = \sqrt{\frac{\sum_{i=1}^{n}(y_{(i)} - \hat{y}_i)^2}{n}}$$

Konfidenzintervall für den Regressanden (*k = 1*) ($df = n - 2$):

$$\hat{y}_i - t_{(\alpha/2)} \cdot SE(\hat{y}) \cdot \sqrt{\frac{1}{n} \cdot \frac{(x_i - \bar{x})^2}{n \cdot s_x^2}} \leq \hat{y}_i^* \leq \hat{y}_i + t_{(\alpha/2)} \cdot SE(\hat{y}) \cdot \sqrt{\frac{1}{n} \cdot \frac{(x_i - \bar{x})^2}{n \cdot s_x^2}}$$

Totaler F-Test

- Nullhypothese: kein Zusammenhang zwischen Regressoren und Regressanden
- H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$; H_1: $\beta_i \neq 0$ für mindestens ein *i*
- Testgröße: $F = \frac{r^2}{k} \Big/ \frac{1 - r^2}{n - k - 1}$
- H_0 ablehnen, falls $F > F_{(1-\alpha,k,n-k-1)}$ (i.d.R. $\alpha = 0.05$ oder $\alpha = 0.01$)

Partieller F-Test (t-Test)

- Nullhypothese: der Regressor x_i übt keinen Einfluss auf den Regressanden aus, der nicht bereits in anderen Regressoren enthalten wäre:

  H_0: ß$_i$ = 0, H_1: ß$_i$ ≠ 0

- Testgröße: $F = t^2 = \frac{b_i^2}{SE(b_i)^2}$

- H_0 ablehnen, falls $F > F_{(1-\alpha,1,n-k-1)}$ bzw. $t > t_{(1-\alpha,n-k-1)}$ (i.d.R. α = 0.05 oder α = 0.01)

Standardschätzfehler des Regressionskoeffizienten: univariat

$$SE(b) = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\frac{1}{n-k-1}\cdot\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

=> Konfidenzintervall für den Regressionskoeffizienten:

$$b_i - t \cdot SE(b_i) \le b_{GG,i} \le b_i + t \cdot SE(b_i) \quad \text{mit} \quad df = n - k - 1$$

Standardschätzfehler des Regressionskoeffizienten: multivariat

- $\mathbf{X} = (n \times k)$ -Matrix der Werte der k Regressoren
- Kovarianz-Matrix der Regressionskoefizienten: $\mathbf{V} = SE(\hat{y})^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$
- Diagonalelemente: $a_{ii} = \left|(\mathbf{X}'\mathbf{X})^{-1}\right|_{ii}$
- Stnadardschätzfehler des Koeffizienten: $SE(b_i) = SE(\hat{y}) \cdot \sqrt{a_{ii}}$

Überprüfung der Autokorrelation der Residuen

- => wichtig für Simulation von Zeitreihen
- Durbin-Watson-Test
- H_0: die ***Residuen*** $e_i = y_i - \hat{y}_i$ sind nicht autokorreliert (sind unabhängig)

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (0 \le d \le 4)$$

- Ablehnung der H_0 für $\boldsymbol{d} < d^*$ (positive Autokorrelation)
  bzw. $\boldsymbol{d} > (4\text{-}d^*)$ (negative Autokorrelation)

Cooks Distanz (Cooks D)

= Maß für den Einfluss einzelner Fälle auf das Ergebnis der Regression → ***Ausreißer***

= Differenz der vorhergesagten Werte zwischen dem

- Regressionsmodell, dass ***alle Fälle*** berücksichtigt, im Vergleich zu dem
- Regressionsmodell, dass ***den jeweiligen Fall nicht*** berücksichtigt

- starker Einfluss für hohe Werte für D
- Faustregeln: Fälle weglassen, für die $D > 1$, bzw. $D > 4/(n - k - 1)$

Multikollinearinität

= Ausmaß der Korrelationen zwischen verschiedenen Regressoren

Probleme:

1. Verringerung der Genauigkeit der Bestimmung der Regressionskoeffizienten (Vergrößerung der Standardfehler)
2. Signifikantes Bestimmtheitsmaß trotz nicht-signifikanter Koeffizienten
3. Die bestimmten Regressionskoeffizienten sind nicht stabil

Maße der Multikollinearinität

Konditionsindex:

- Wurzel aus dem Verhältnis des größten Eigenwerts zu jedem einzelnen Eigenwert, d_l/d_j
- 30 bis 100 ist ein Indikator für mäßige bis starke Kollinearität

Toleranz (TOL), Varianzinflation (VIF):

- TOL = 1/VIF = $1 - r_i^2$ (r_i^2: Bestimmtheitsmaß für den Regressor x_i mit allen anderen Regressoren)
- TOL-Werte < 0.1, VIF-Werte >10 werden i.d.R. als problematisch angesehen

Voraussetzung: Mehrdimensionale Normalverteilung

- nicht erforderlich für Bestimmtheitsmaß
- erforderlich für Bestimmung der Konfidenzintervalle und Signifikanztests **für kleine *n* und große *k***

Aufgabe

- Berechnen Sie für einzelne ausgewählte Regressanden multiple lineare Regresssionen mit allen übrigen Variablen mit den Verfahren: Standdard, schrittweise vorwärts, schrittweise rückwärts.
- Inwieweit sind die bestimmten Regressionsfunktionen plausibel bzw. sinnvoll?
- Wie beurteilen Sie das Problem der Multikollinearität?
- Wie sehr ändert sich das Ergebnis, wenn die Fälle hoher Cooks-Distanz nicht berücksichtigt werden?
- Bestimmen Sie das "optimale" Modell.