2.4 Residualkomponente und Korrelogramm

Residualkomponente:
Restkomponente u_t einer Zeitreihe (y_t) nach Ausschaltung der systematischen Komponenten

Wenn die Residualkomponente keine Systematik mehr enthält, d.h. nur noch zufällig schwankt, müssen die Autokorrelationen (Reihenkorrelationen) approximativ gleich 0 sein. Unter **Autokorrelation** der Residuen u_1, u_2, ..., u_n versteht man die Korrelation der aktuellen Residuen mit den um eine Periode verzögerten Residuen (Autokorrelation 1. Ordnung), zwei Perioden verzögerten Zeitreihe (Autokorrelation 2. Ordnung), usw. Das Ausmaß der noch in den Residuen enthaltenen systematischen Schwankungen wird durch den **Autokorrelationskoeffizienten** gemessen.

Der **empirische Autokorrelationskoeffizient** ist analog zu dem aus der deskriptiven Statistik bekannten Korrelationskoeffizienten nach Bravais und Pearson, r_{xy}, definiert, der durch das Verhältnis der Kovarianz s_{xy} zum Produkt der Standardabweichungen s_x und s_y der beiden Merkmale X und Y gegeben ist: $r = s_{xy} / (s_x \cdot s_y)$.

Korrelationskoeffizient
Nach Bravais und Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Empirischer Autokorrelationskoeffizienten für eine zeitabhängige Variable X

gegeben: Zeitreihe $(x_1, x_2, \ldots, x_n)$

Autokorrelationskoeffizient 1. Ordnung:
Korrelation zwischen den n-1 Wertepaaren $(x_2, x_1), (x_3, x_2), \ldots, (x_n, x_{n-1})$

$$r_1 = \frac{\sum_{t=2}^{n}(x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \qquad (2.65)$$

Autokorrelationskoeffizient 2. Ordnung:
Korrelation zwischen den n-2 Wertepaaren $(x_3, x_1), (x_4, x_2), \ldots, (x_n, x_{n-2})$

$$r_2 = \frac{\sum_{t=3}^{n}(x_t - \bar{x})(x_{t-2} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \qquad (2.66)$$

Allgemein: Autokorrelationskoeffizient k-ter Ordnung:
Korrelation zwischen den n-k Wertepaaren $(x_{k+1}, x_1), (x_{k+2}, x_2), \ldots, (x_n, x_{n-k})$

(2.67)
$$r_k = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

Beispiel:
Wir berechnen beispielhaft die Autokorrelationskoeffizienten 1. und 2. Ordnung für die Zeitreihe des Kfz-Bestands.

t	x_t	$x_t - \bar{x}$	x_{t-1}	$x_{t-1} - \bar{x}$	$(x_t - \bar{x})^2$	$(x_t - \bar{x})(x_{t-1} - \bar{x})$
1	27116	-3033	-	-	9199089	-
2	27858	-2291	27116	-3033	5248681	6948603
3	28452	-1697	27858	-2291	2879809	3887827
4	29122	-1027	28452	-1697	1054729	1742819
5	29905	-244	29122	-1027	59536	250588
6	30618	469	29905	-244	219961	-114436
7	31748	1599	30618	469	2556801	749931
8	32762	2613	31748	1599	6827769	4178187
9	33764	3615	32762	2613	13068225	9445995
45	271345				41114600	27089514

$$r_1 = \frac{\sum_{t=2}^{n}(x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} = \frac{27089514}{41114600} = 0{,}659$$

t	x_t	$x_t - \bar{x}$	x_{t-2}	$x_{t-2} - \bar{x}$	$(x_t - \bar{x})^2$	$(x_t - \bar{x})(x_{t-1} - \bar{x})$
1	27116	-3033	-	-	9199089	-
2	27858	-2291	-	-	5248681	-
3	28452	-1697	27116	-3033	2879809	5147001
4	29122	-1027	27858	-2291	1054729	2352857
5	29905	-244	28452	-1697	59536	414068
6	30618	469	29122	-1027	219961	-481663
7	31748	1599	29905	-244	2556801	-390156
8	32762	2613	30618	469	6827769	1225497
9	33764	3615	31748	1599	13068225	5780385
45	271345				41114600	14047989

$$r_2 = \frac{\sum_{t=3}^{n}(x_t - \bar{x})(x_{t-2} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} = \frac{14047989}{41114600} = 0{,}342$$

Definition des Autokorrelationskoeffizienten 1. Ordnung der Residuen

Um den Autokorrelationskoeffizienten 1. Ordnung der Residuen u_t, t=1,2,…,n zu bestimmen, betrachten wird die Korrelation zwischen den n-1 Wertepaaren

$$(u_2, u_1), (u_3, u_2), \ldots, (u_n, u_{n-1}).$$

mit

$$\bar{u}_1 = \frac{1}{n-1}\sum_{t=2}^{n} u_t \qquad \text{und} \qquad \bar{u}_2 = \frac{1}{n-1}\sum_{t=1}^{n-1} u_t$$

erhält man für den **Autokorrelationskoeffizienten 1. Ordnung** den Ausdruck

(2.68)
$$r_1 = \frac{\sum_{t=2}^{n}(u_t - \bar{u}_1)(u_{t-1} - \bar{u}_2)}{\sqrt{\sum_{t=2}^{n}(u_t - \bar{u}_1)^2 \sum_{t=1}^{n-1}(u_t - \bar{u}_2)^2}} \quad .$$

Für ein großes n kann bei Mittelwertstationarität $E(u_t)=0$ von den Approximationen

$$n \approx n\text{-}1 \text{ und } \bar{u}_1 \approx \bar{u}_2 \approx \bar{u} \text{ mit } \bar{u} = \frac{1}{n}\sum_{t=1}^{n} u_t \approx 0$$

Gebrauch gemacht werden. Hiermit modifiziert sich (2.68) zu

(2.69)
$$r_1 = \frac{\sum_{t=2}^{n} u_t \cdot u_{t-1}}{\sum_{t=1}^{n} u_t^{\,2}} \quad .$$

Der **Autokorrelationskoeffizient 1. Ordnung** (2.69) ist durch das Verhältnis der Autokovarianz 1. Ordnung, c_1, zur Varianz der Zeitreihe der Residuen, c_0, gegeben:

(2.70) $r_1 = c_1 / c_0$.

Autokovarianz 1. Ordnung: (2.71) $c_1 = \frac{1}{n} \sum_{t=2}^{n} u_t \cdot u_{t-1}$

Varianz von u_t: (2.72) $c_0 = \frac{1}{n} \sum_{t=1}^{n} u_t^2$

Allgemein erhält man die Folge der **Autokorrelationskoeffizienten k-ter Ordnung**, (r_k), aus der Folge der Autokovarianzen, (c_k), für k=0,1,2,... gemäß

(2.73) $r_k = c_k / c_0$

mit (2.74) $c_k = \frac{1}{n} \sum_{t=k+1}^{n} u_t \cdot u_{t-k}$

Beispiel:
Wir überprüfen die Residuen des linearen Trendmodells der Zeitreihe des Kfz-Bestandes auf Autokorrelation 1. und 2. Ordnung.

t	y_t	$\hat{y}_t$	u_t	u_t^2	u_{t-1}	$u_t \cdot u_{t-1}$
1	27116	26856,6	259,4	67288,36	-	-
2	27858	27679,8	178,2	31755,24	259,4	46225,08
3	28452	28503,0	-51,0	2601,00	178,2	-9088,2
4	29122	29326,2	-204,2	41697,64	-51,0	10414,2
5	29905	30149,4	-244,4	59731,36	-204,2	49906,48
6	30618	30972,6	-354,6	125741,16	-244,4	86664,24
7	31748	31795,8	-47,8	2284,84	-354,6	16949,88
8	32762	32619,0	143,0	20449,00	-47,8	-6835,4
9	33764	33442,2	321,8	103555,24	143,0	46017,4
	271345	271345,0	0,4≈0	455103,84		240253,68

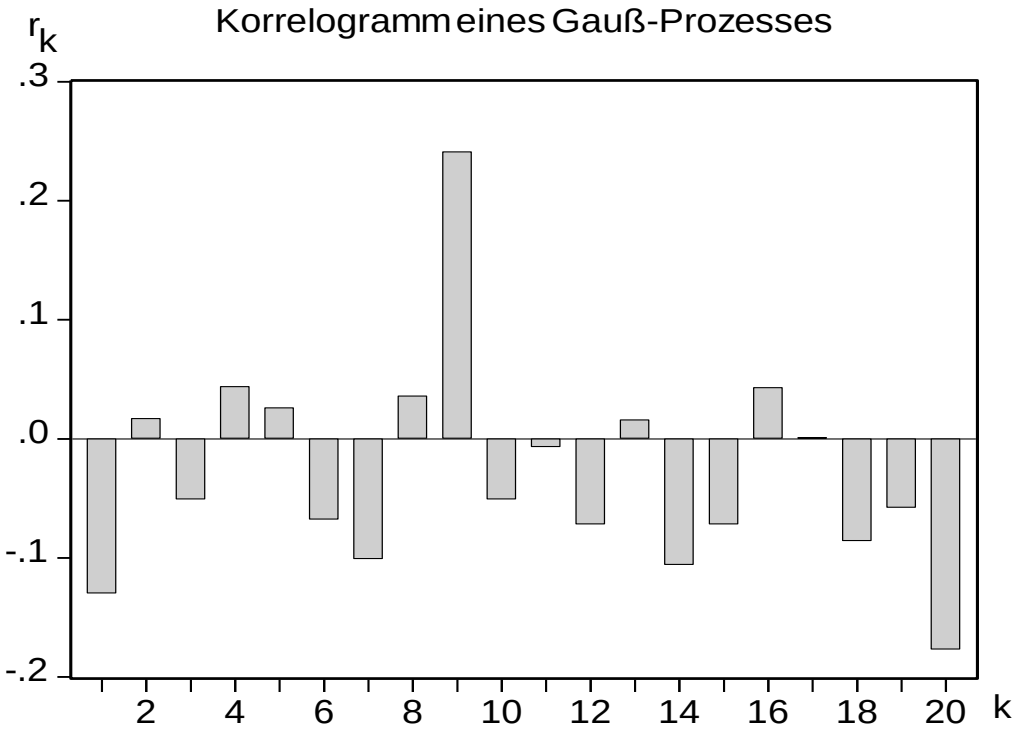
$$r_1 = \sum_{t=2}^{n} u_t \cdot u_{t-1} / \sum_{t=1}^{n} u_t^2 = \frac{240253{,}68}{455103{,}84} = 0{,}528$$

t	y_t	$\hat{y}_t$	u_t	u_t^2	u_{t-2}	$u_t \cdot u_{t-2}$
1	27116	26856,6	259,4	67288,36	-	-
2	27858	27679,8	178,2	31755,24	-	-
3	28452	28503,0	-51,0	2601,00	259,4	-13229,40
4	29122	29326,2	-204,2	41697,64	178,2	-36388,44
5	29905	30149,4	-244,4	59731,36	-51,0	12464,40
6	30618	30972,6	-354,6	125741,16	-204,2	72409,32
7	31748	31795,8	-47,8	2284,84	-244,4	11682,32
8	32762	32619,0	143,0	20449,00	-354,6	-50707,80
9	33764	33442,2	321,8	103555,24	-47,8	-15382,04
	271345	271345,0	0,4≈0	455103,84		-19151,64

$$r_2 = \sum_{t=3}^{n} u_t \cdot u_{t-2} / \sum_{t=1}^{n} u_t^2 = \frac{-19151,64}{455103,84} = -0,042$$

Korrelogramm

Die grafische Darstellung der Folge der Autokorrelationskoeffizienten (r_k) in Abhängigkeit von den Lags k in einem Koordinatensystem heißt **Korrelogramm**. Mit Hilfe eines Korrelogramms lassen sich die Abhängigkeitsstrukturen einer Zeitreihe aufdecken.



Gauß-Prozess:
Unabhängige normalverteilte Zufallsvariablen

Interpretation des Korrelogramms

Bei einem **reinen Zufallsprozess** wie z.B. dem Gauß-Prozess sind die theoretischen Autokorrelationskoeffizienten für alle Lags k gleich 0, was bedeutet, dass keinerlei zeitlichen Abhängigkeiten der Zufallsvariablen vorliegen. Bei großem n ist daher zu erwarten, dass die empirischen Autokorrelationskoeffizienten r_k in der Nähe von 0 liegen:

$$r_k \approx 0.$$

Autokorrelationen, die bei einem Signifikanzniveau von 5% nicht signifikant von 0 abweichen, liegen näherungsweise innerhalb eines Bereichs von $\pm$ 2 Standardabweichungen um 0, da die Größen r_k approximativ normalverteilt sind mit dem Erwartungswert

(2.75) $$E(r_k) = 1/n \approx 0.$$

Mit dem Standardfehler von r_k,

(2.76) $$\sigma(r_k) = \sqrt{1/n},$$

erhält man das approx. 95%-Konfidenzintervall

$$[- 2\sqrt{1/n}, 2\sqrt{1/n}]$$

für die empirischen Autokorrelationskoeffizienten.

Bei der Interpretation ist zu beachten, dass die Hypothese eines reinen Zufallsprozesses der Residualkomponente noch nicht verworfen werden muss, wenn einer von 20 Autokorrelationskoefizienten außerhalb des 95%-Konfidenzintervalls liegt, wie es in unserem Beispiel der Fall ist. Denn dies ist bei einem Signifikanzniveau von 5% durchaus zu erwarten. Hierdurch können Schwierigkeiten bei der Identifikation des datenerzeugenden Prozesses auftreten.

Beispiel:
Es wird überprüft, ob die Residuen des linearen Trendmodells der Zeitreihe des Kfz-Bestands als Realisationen eines reinen Zufallsprozesses aufgefasst werden können.

Approx. 95%-Konfidenzintervall für r_1 und r_2:

$[- 2\sqrt{1/n}, 2\sqrt{1/n}] = [-2 \sqrt{1/9}, 2 \sqrt{1/9}] = [-0{,}667; 0{,}667]$

Autokorrelationskoeffizient 1. Ordnung der Residuen:

$r_1 = 0{,}528 \in [-0{,}667; 0{,}667]$

Autokorrelationskoeffizient 2. Ordnung der Residuen:

$r_2 = -0{,}042 \in [-0{,}667; 0{,}667]$

Da r_1 und r_2 im approx. 95%-Konfidenzintervall liegen, kann die Nullhypothese eines reinen Zufallsprozesses der Residuen des linearen Trendmodells nicht abgelehnt werden.

<u>Hinweis:</u> Die Überprüfung hat illustrativen Charakter, da sich eine Ablehnung der Nullhypothese angesichts des geringen Stichprobenumfangs trotz des hohen r_1-Wertes als nicht möglich erweist.

Portmanteau-Tests

Das Korrelogramm wird im Allgemeinen durch einen Portmanteau-Test ergänzt, der die in den Autokorrelationskoeffizienten bis zu einem Lag k enthaltenen Informationen in komprimierter Form auswertet. Der **Box-Pierce-Test** basiert auf der Teststatistik

(2.77) $$Q(k) = n\sum_{j=1}^{k} r_j^2,$$

die bei großem n approximativ χ^2-verteilt ist mit k-m Freiheitsgraden. m ist hierbei die Anzahl der geschätzten Modellparameter. Wenn die Residuen u_t Realisationen eines reinen Zufallsprozesses sind, werden die empirischen Autokorrelationskoeffizienten r_k niedrige Werte annehmen, was ebenfalls für die Prüfgröße Q zutrifft. Die Nullhypothese eines reinen Zufallsprozesses (White-Noise-Prozess) wird abgelehnt, wenn die Box-Pierce-Statistik Q bei einem Signifikanzniveau α das (1- α)-Quantil der χ^2-Verteilung mit k-m Freiheitsgraden überschreitet:

$$Q(k) > \chi^2_{k-m;1-\alpha} \Rightarrow H_0 \text{ ablehnen.}$$

In diesem Fall muss davon ausgegangen werden, dass in den Residuen noch eine Systematik vorhanden ist. Das zugrunde gelegte Zeitreihenmodell erfasst nicht alle systematischen Komponenten der Zeitreihe.

Wie sich gezeigt hat, verbessert sich die Verteilungsapproximation beim Box-Pierce-Test nur langsam mit wachsendem n. Der Test ist extrem konservativ, d.h. die Nullhypothese eines reinen Zufallsprozesses wird im Allgemeinen zu lange beibehalten.

Ljung und Box haben eine Modifikation des Box-Pierce-Tests vorgeschlagen, die eine verbesserte Approximation der Verteilung der Teststatistik an die χ^2-Verteilung führt. Anstelle der Residualkorrelationen r_k verwenden sie die asymptotisch äquivalenten Größen

$$\sqrt{\frac{n+2}{n-j}} \cdot r_j$$

womit sich die Teststatistik

$$Q^*(k) = n(n+2)\sum_{j=1}^{k} \frac{r_j^2}{n-j} \qquad (2.78)$$

des **Ljung-Box-Tests** ergibt. Q* ist ebenso wie Q approximativ χ^2-verteilt mit k-m Freiheitsgraden. Mit dem Ljung-Box-Test lassen sich unter Umständen inadäquate Zeitreihenmodelle aufdecken, die bei Anwendung des Box-Pierce-Testes möglicherweise übersehen werden.

Beispiel:
Die beiden Portmanteau-Tests seien anhand der Zeitreihe des Kfz-Bestands illustriert.

- **Box-Pierce-Test**

Teststatistik: $Q(2) = n\sum_{j=1}^{2} r_j^2 = 9\cdot(0{,}659^2 + 0{,}342^2) = 9\cdot 0{,}551245 = 4{,}961$

Kritischer Wert (α=0,05): $\chi^2_{2;0{,}95} = 5{,}99$

Testentscheidung: $Q(2) = 4{,}961 < \chi^2_{2;0{,}95} = 5{,}99 \Rightarrow$ H_0 annehmen

- **Ljung-Box-Test**

Teststatistik: $Q^*(2) = n(n+2)\sum_{j=1}^{2}\frac{r_j^2}{n-j} = 9\cdot(9+2)\cdot\left(\frac{0{,}659^2}{9-1} + \frac{0{,}342^2}{9-2}\right)$

$= 99\cdot(0{,}054285 + 0{,}016709) = 99\cdot 0{,}070994 = 7{,}028$

Kritischer Wert (α=0,05): $\chi^2_{2;0{,}95} = 5{,}99$

Testentscheidung: $Q^*(2) = 7{,}028 > \chi^2_{2;0{,}95} = 5{,}99 \Rightarrow$ H_0 ablehnen

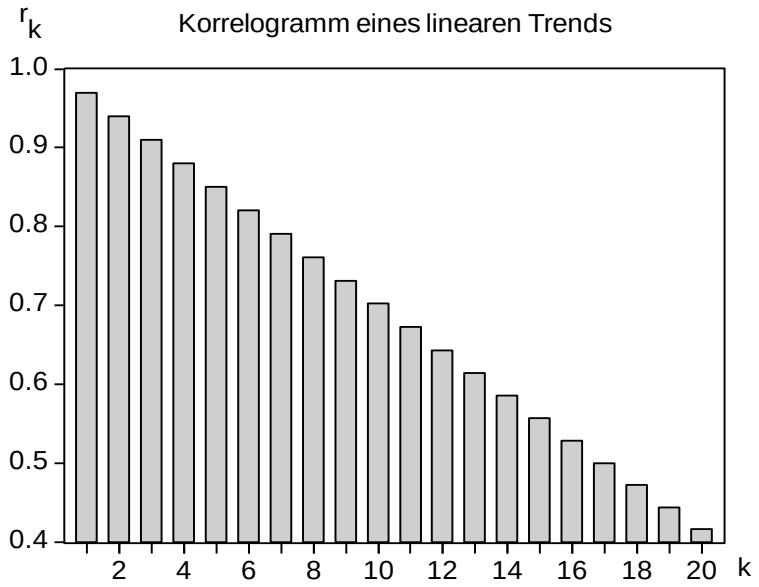
Korrelogramm und Ljung-Box-Test eines reinen Zufallsprozesses

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
.*| . |	.*| . |	1	-0.130	-0.130	1.7284	0.189
. | . |	. | . |	2	0.017	0.001	1.7601	0.415
. | . |	. | . |	3	-0.051	-0.050	2.0345	0.565
. | . |	. | . |	4	0.044	0.032	2.2420	0.691
. | . |	. | . |	5	0.026	0.037	2.3131	0.804
.*| . |	.*| . |	6	-0.068	-0.064	2.8146	0.832
.*| . |	.*| . |	7	-0.101	-0.117	3.9225	0.789
. | . |	. | . |	8	0.036	0.010	4.0629	0.851
. |** |	. |** |	9	0.241	0.250	10.594	0.305
. | . |	. | . |	10	-0.051	0.008	10.894	0.366
. | . |	. | . |	11	-0.007	-0.013	10.900	0.452
.*| . |	.*| . |	12	-0.072	-0.063	11.502	0.486
. | . |	. | . |	13	0.016	-0.039	11.531	0.566
.*| . |	.*| . |	14	-0.106	-0.135	12.863	0.537
.*| . |	.*| . |	15	-0.072	-0.071	13.490	0.565
. | . |	. |*. |	16	0.043	0.097	13.719	0.620
. | . |	. | . |	17	0.001	0.000	13.719	0.687
.*| . |	.*| . |	18	-0.086	-0.180	14.637	0.687
.*| . |	.*| . |	19	-0.058	-0.102	15.057	0.719
.*| . |	**| . |	20	-0.177	-0.231	19.039	0.519

Korrelogramme bei unterschiedlichen Zeitreihenmustern

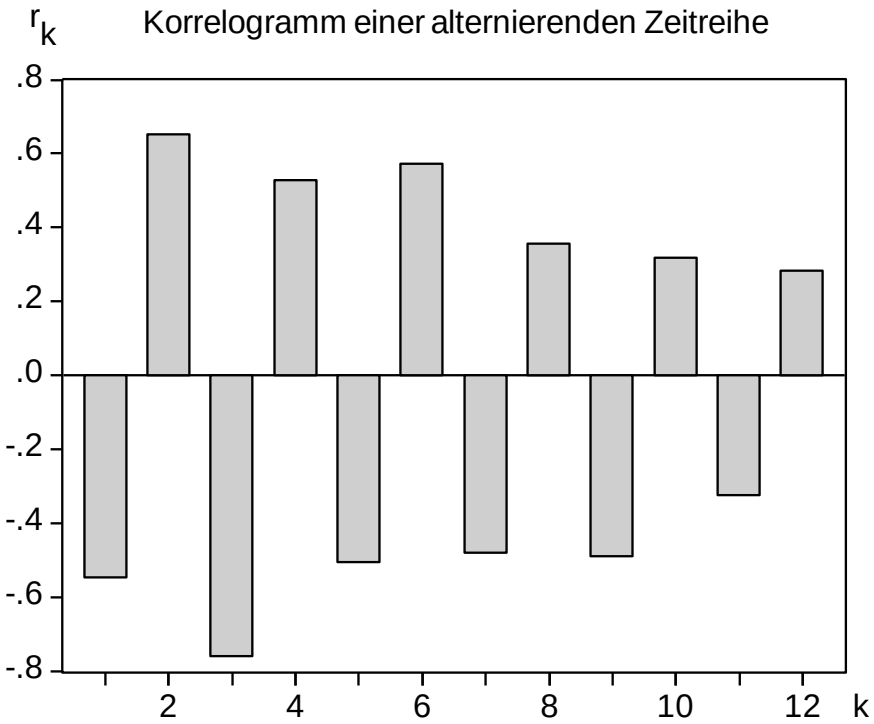
Trendbehaftete Zeitreihen

Wenn die Zeitreihe einen Trend enthält, verschwinden die empirischen Autokorrelationen r_k im Allgemeinen erst für sehr große Lags. Einer Beobachtung auf einer Seite des Gesamtmittels, d.h. einer unter- oder übedurchschnittlichen Beobachtung, folgt dann nämlich aufgrund es Trends eine große Anzahl von Zeitreihenwerten auf der gleichen Seite. Da der Trend alle anderen Eigenschaften überdeckt, bietet ein Korrelogramm bei trendbehafteten Zeitreihen nur wenig Informationen. Aus diesem Grund sollte in diesem Fall eine Trendbereinigung vorgeschaltet werden.



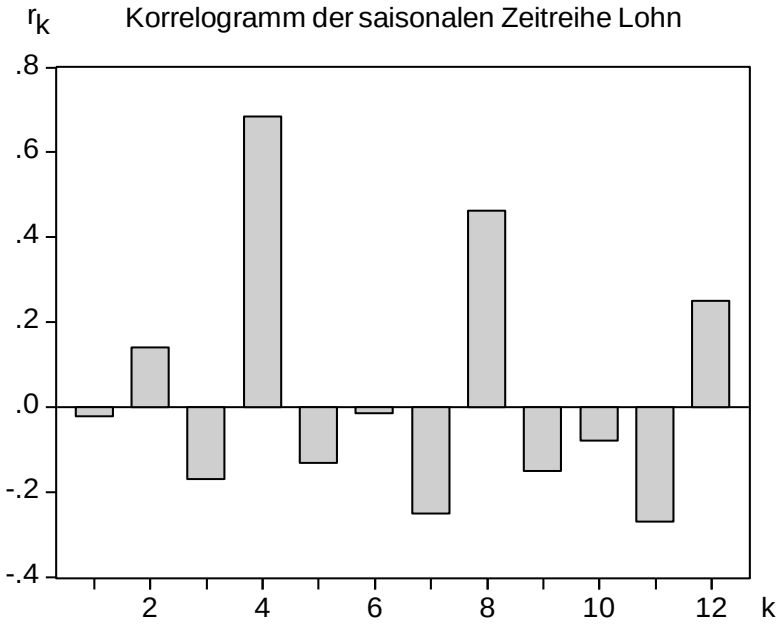
Alternierende Zeitreihen

Bei alternierenden Zeitreihen liegen aufeinanderfolgende Werte auf verschiedenen Seiten des Gesamtmittelwerts. Bei einem ungeraden Lag ist der Autokorrelationskoeffizient negativ, da die Beobachtungen auf verschiedenen Seiten des Mittelwerts liegen. Entsprechend tendieren die Beobachtungen bei einem geraden Lag auf der gleichen Seite des Mittelwerts zu liegen, so dass der Autokorrelationskoeffizient positiv ist.



Saisonschwankungen

Saisonschwankungen findet man im Korrelogramm mit der gleichen Frequenz wieder. Bei Quartalsdaten ist r_4 groß und positiv, während dies bei Monatsdaten für r_{12} der Fall ist.



Das Korrelogramm vermittelt bei saisonalen Daten wenig zusätzliche Informationen, da Saisonschwankungen bereits in dem Zeitreihenplot gut erkennbar sind. Sehr nützliche Informationen liefert dagegen ein Korrelogramm der saisonbereinigten Zeitreihen

Kurzzeit-Korrelationen

Bei nicht-trendbehafteten (oder trendbereinigten) Zeitreihen findet man häufig einen großen Wert für r_1, dem zwei oder drei signifikant von 0 verschiedene Autokorrelationskoeffizienten folgen. Diese sind aber sukzessive kleiner. Die übrigen Autokorrelationen, d.h. die r_k-Werte für größere Lags k, liegen dagegen nahe bei 0.

In der zugrunde liegenden Zeitreihe besteht die Tendenz, dass einer überdurchschnittlich (unterdurchschnittlich) großen Beobachtung einige wenige, d.h. etwa eine, zwei, drei überdurchschnittlich (unterdurchschnittlich) große Beobachtungen folgen. Ein solches Verhalten der Zeitreihenwerte wird als **Kurzzeit-Korrelation** bezeichnet.