Modellannahmen der linearen Regression

Zur Durchführung einer Regressionsanalyse werden eine Reihe von Annahmen gemacht, die das zugrunde gelegte stochastische Modell betreffen.

Prämisse	Prämissenverletzung	Konsequenzen	Prüfung	Maßnahmen
ℵ Linearität in den Parametern	Nichtlinearität	Verzerrung der Schätzwerte	Betrachten des Punktediagramms	nichtlineare Transformationen, Einführung einer Dummyvariablen bei Strukturbrüchen
ℑ Erwartungswert der Störgröße gleich Null	Störgröße enthält einen systematischen Effekt	Verzerrung der Schätzwerte	Messverfahren prüfen	
ℜ Berücksichtigung aller relevanten Variablen	Unvollständigkeit zu viele Variablen	Verzerrung der Schätzwerte Ineffizienz der Schätzwerte (die Varianz ist nicht mehr minimal)	keine Korrelation zwischen den erklärenden Variablen und der Störgröße	
℘ Homoskedastizität der Störgrößen	Heteroskedastizität (meist auch ein Problem von Nichtlinearität)	Ineffizienz der Schätzung	visuelle Kontrolle, Goldfeld/Quandt-Test oder Verfahren nach Glesjer	Transformation der abhängigen Variablen oder der Regressionsgleichung
⊗ Unabhängigkeit der Störgrößen	Autokorrelation (meist auch ein Problem von Nichtlinearität)	Ineffizienz der Schätzung	visuelle Kontrolle, Durbin/Watson-Test	
⊕ keine lineare Abhängigkeit zwischen den unabhängigen Variablen	Multikollinearität	verminderte Präzision der Schätzwerte	Korrelationskoeffizienten der unabhängigen Variablen prüfen (Toleranz)	unwichtige Variablen entfernen oder Transformation
∅ Störgrößen sind normalverteilt	keine Normalverteilung	Ungültigkeit des Signifikanztests		

Unter den Annahmen ℵ bis ⊕ liefert die Methode der kleinsten Quadrate unverzerrte und effiziente lineare Schätzfunktionen für die Regressionsparameter, zudem ist Annahme ∅ für die Durchführung von Signifikanztests von Bedeutung.

ℵ **Das Modell ist linear in den Parametern und enthält alle relevanten erklärenden Variablen**

- nichtlineare Beziehungen können durch Wachstums- oder Sättigungsprobleme bedingt sein
- die Folge von nicht entdeckter Linearität ist eine Verzerrung der Schätzwerte der Parameter, die geschätzten Werte für den Regressionskoeffizient streben mit wachsendem Stichprobenumfang nicht mehr gegen die wahren Werte
- ob Linearität vorliegt lässt sich durch Betrachten des Punktediagramms entdecken, in vielen Fällen ist es jedoch möglich, eine nichtlineare Beziehung durch Transformation der Variablen in eine lineare Beziehung zu überführen
- eine weitere Verletzung der Annahme sind Strukturbrüche, die man häufig bei Zeitreihen findet, wenn z.B. die durch Änderung der Rahmenbedingungen eine Änderung in der zeitlichen Entwicklung der betrachteten abhängigen Variablen bewirkt wird

ℑ **Störgröße hat den Erwartungswert Null, somit gleichen sich Schwankungen im Mittel aus**

- die Störgröße umfasst nur zufällige Effekte, die Abweichungen zwischen den beobachteten und geschätzten Werten verursachen
- eine Verletzung dieser Annahme ergibt sich bei einem systematischen Messfehler, wenn die Werte von Y mit einem konstanten Fehler zu hoch oder zu niedrig gemessen werden, ergibt sich für die Störgröße ein systematischer Effekt
- der systematische Messfehler verzerrt den Schätzwert des konstanten Gliedes

ℜ **Vollständigkeit des Regressionsmodells**

- da eine empirische Variable nie vollständig durch eine begrenzte Anzahl von beobachtbaren Variablen erklärt werden kann, bleibt das Modell unvollständig und die mögliche Folge ist die Verzerrung der Schätzwerte
- dies kann man unter der Annahme umgehen, dass keine Korrelation zwischen den im Modell enthaltenen erklärenden Variablen und der Störgröße, die die nicht berücksichtigten Variablen enthält, besteht
- neben der Vernachlässigung relevanter Variablen kann es auch vorkommen, dass ein Modell zu viele erklärende Variablen enthält, was zu ineffizienten Schätzwerten führt
- bei einer großen Anzahl von Variablen kann es sowohl vorkommen, dass sich eine irrelevante Variable als statistisch signifikant erweist, als auch, dass ein relevanter Einflussfaktor nicht signifikant erscheint

℘ **Störgrößen haben eine konstante Varianz (Homoskedastzität)**

- Homogenität der Varianz der Fehlervariablen, das heißt, die Störgröße darf nicht von den unabhängigen Variablen und von der Reihenfolge der Beobachtungen abhängig sein, z.B. bei zunehmender Störgröße in einer Reihe von Beobachtungen aufgrund von Messfehlern durch nachlassende Aufmerksamkeit
- Heteroskedastizität führt zu ineffizienten Schätzungen und verfälscht den Standardfehler des Regressionskoeffizienten, damit wird auch die Schätzung des Konfidenzintervalls ungenau
- zur graphischen Aufdeckung kann man die Störgrößen gegen die geschätzten Werte von Y plotten, bei Vorliegen von Heteroskedastizität ergibt sich dabei meist ein Dreiecksmuster
- Testverfahren: Goldfeld/Quandt-Test, bei dem die Stichprobenvarianzen der Störgrößen in zwei Unterstichproben verglichen und ins Verhältnis gesetzt werden oder das Verfahren nach Glesjer, bei dem eine Regression der absoluten Störgrößen auf die Regressoren durchgeführt wird

⊗ **Unabhängigkeit der Störgrößen**

- Störgrößen sind unkorreliert, Autokorrelation tritt dagegen vor allem bei Zeitreihen auf, die Abweichungen von der Regressionsgeraden sind nicht mehr zufällig, sondern in ihrer Richtung von den Abweichungen, z.B. des vorangegangenen Beobachtungswertes, abhängig
- das führt zu Verzerrungen bei der Ermittlung des Standardfehlers der Regressionskoeffizienten und ihrer Konfidenzintervalle
- zur graphischen Aufdeckung kann man die Störgrößen gegen die geschätzten Werte von Y plotten, bei positiver Autokorrelation liegen aufeinanderfolgende Werte der Störgrößen dicht beieinander, bei negativer Autokorrelation schwanken sie stark
- Testverfahren: Durbin/Watson-Test, der die Reihenfolge der Störgrößen der Beobachtungswerte untersucht

⊕ **keine lineare Abhängigkeit zwischen den erklärenden Variablen**

- eine erklärende Variable darf sich nicht als lineare Funktion der anderen erklärenden Variablen darstellen lassen, perfekte Multikollinearität wird selten vorkommen und wenn, dann meist infolge des Fehlers, dass man dieselbe Einflussgröße zweimal als unabhängige Variable in das Regressionsmodell aufgenommen hat
- ein gewisser Grad an Multikollinearität wird bei empirischen Daten immer bestehen, aber mit zunehmender Multikollinearität werden die Schätzungen der Regressionsparameter unzuverlässiger
- bei Multikollinearität überschneiden sich die Streuungen der unabhängigen Variablen, das weist auf Redundanz in den Daten und damit auf weniger Aussagekraft hin, zudem bedeutet es, dass sich die vorhandene Information nicht mehr eindeutig den Variablen zuordnen lässt
- die Folge kann sein, dass das Bestimmtheitsmaß der Regressionsfunktion signifikant ist, obwohl keiner der Koeffizienten in der Funktion signifikant ist oder die Regressionskoeffizienten ändern sich erheblich, wenn eine weitere Variable in die Funktion einbezogen oder eine enthaltene Variable entfernt wird
- erste Anhaltspunkte liefert die Korrelationsmatrix, zur Aufdeckung von Multikollinearität empfiehlt es sich zudem eine Regression jeder unabhängigen Variablen auf die anderen unabhängigen Variablen durchzuführen um die zugehörigen Korrelationskoeffizienten zu ermitteln, ein ähnliches Maß zur Prüfung ist die Toleranz
- weniger wichtige Variablen könnten aus der Regressionsgleichung entfernt werden, eine andere Möglichkeit wäre es, die Variablen zu transformieren oder durch Faktoren zu ersetzen

∅ **Normalverteilung der Störgrößen**

- dass die Störgrößen normalverteilt sind, ist Voraussetzung für die Durchführung statistischer Testverfahren, wenn diese Annahme erfüllt ist, sind auch die Y-Werte und folglich die Regressionsparameter normalverteilt
- sollten die zu testenden Schätzwerte der Regressionsparameter nicht normalverteilt sein, wären die Tests nicht gültig