



MINITAB-ASSISTENT – WHITE PAPER

Dieses White Paper ist Teil einer Reihe von Veröffentlichungen, welche die Forschungsarbeiten der Minitab-Statistiker erläutern, in deren Rahmen die im Assistenten der Minitab Statistical Software verwendeten Methoden und Datenprüfungen entwickelt wurden.

Multiple Regression

Übersicht

Das Verfahren der multiplen Regression im Assistenten passt lineare und quadratische Modelle mit bis zu fünf Prädiktoren (x) und einer stetigen Antwortvariablen (y) durch Schätzung nach der Methode der kleinsten Quadrate an. Der Benutzer wählt den Modelltyp und der Assistent die Modellterme aus. In diesem White Paper werden die Kriterien erläutert, anhand derer der Assistent das Regressionsmodell auswählt.

Außerdem werden mehrere Faktoren untersucht, die wichtig sind, um ein gültiges Regressionsmodell zu erhalten. Zunächst muss die Stichprobe umfassend genug sein, um eine ausreichende Trennschärfe des Tests und eine ausreichende Genauigkeit für die geschätzte Stärke der Beziehung zwischen x und y zu ermöglichen. Dann ist es wichtig, ungewöhnliche Daten zu identifizieren, die sich auf die Ergebnisse der Analyse auswirken könnten. Ebenfalls besprochen wird die Annahme, dass der Fehlerterm einer Normalverteilung folgt, und es werden die Auswirkungen einer Nicht-Normalverteilung auf die Hypothesentests des Gesamtmodells ausgewertet.

Auf der Grundlage dieser Faktoren führt der Assistent automatisch die folgenden Prüfungen für die Daten durch und gibt die Ergebnisse in der Auswertung aus:

- Umfang der Daten
- Ungewöhnliche Daten
- Vorliegen einer Normalverteilung

In diesem White Paper wird untersucht, wie sich diese Faktoren in der Praxis auf die Regressionsanalyse auswirken. Außerdem wird erläutert, wie die Richtlinien für die Prüfung dieser Faktoren im Assistenten festgelegt wurden.

Regressionsmethoden

Modellauswahl

Bei der Regressionsanalyse im Assistenten wird ein Modell mit einer stetigen Antwortvariablen und zwei bis fünf Prädiktoren angepasst. Bei einem der Prädiktoren kann es sich um einen kategorialen Prädiktor handeln. Es stehen zwei Modelltypen zur Auswahl:

- Linear: $F(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$
- Quadratisch: $F(x) = \beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i < j} \beta_{ij} X_i X_j$

Der Assistent wählt die Modellterme aus dem vollständig linearen oder quadratischen Modell aus.

Zielstellung

Die verschiedenen Methoden, die bei der Modellauswahl angewendet werden können, sollten untersucht werden, um zu ermitteln, welche Methode im Assistenten verwendet werden soll.

Methode

Es wurden drei verschiedenen Typen der Modellauswahl untersucht: Rückwärtsauswahl, Vorwärtsauswahl und schrittweise Auswahl. Diese Typen der Modellauswahl umfassen mehrere Optionen, die ebenfalls untersucht wurden, darunter:

- Die Kriterien, die zum Aufnehmen oder Entfernen von Termen für das Modell verwendet werden.
- Die Festlegung, ob bestimmte Terme erzwungenermaßen in das Modell aufgenommen oder bestimmte Terme bereits in das Anfangsmodell aufgenommen werden sollen.
- Die Hierarchie der Modelle.
- Standardisierung der x-Variablen im Modell.

Diese Optionen wurden geprüft, ihre Auswirkungen auf das Ergebnis des Verfahrens untersucht, und es wurde berücksichtigt, welche Methoden von Fachleuten bevorzugt werden.

Ergebnisse

Das folgende Verfahren wird im Assistenten zur Auswahl der Modellterme verwendet:

- Die schrittweise Modellauswahl wird verwendet. Oft ist eine Menge an potenziellen x-Variablen korreliert, so dass die Auswirkungen eines Terms davon abhängen, welche weiteren Terme sich ebenfalls im Modell befinden. Die schrittweise Auswahl ist unter diesen Bedingungen wahrscheinlich der beste Ansatz, da es hierbei möglich ist, Terme in einem Schritt aufzunehmen und später wieder zu entfernen, abhängig davon, welche weiteren Terme im Modell enthalten sind.

- Die Hierarchie des Modells wird bei jedem Schritt beibehalten, und in jedem Schritt können mehrere Terme in das Modell aufgenommen werden. Wenn der signifikanteste Term zum Beispiel X_1^2 ist, wird dieser zusammen mit X_1 aufgenommen, unabhängig davon, ob X_1 signifikant ist. Die Hierarchie ist empfehlenswert, da sie die Übersetzung des Modells von standardisierte in nicht standardisierte Einheiten ermöglicht. Und da die Hierarchie es ermöglicht, in jedem Schritt mehrere Terme aufzunehmen, ist es möglich, einen wichtigen quadratischen Term oder Wechselwirkungsterm zu erkennen, selbst wenn der zugeordnete lineare Term keine stark Beziehung mit der Antwortvariablen aufweist.
- Terme werden basierend auf $\alpha = 0,10$ in das Modell aufgenommen oder aus diesem entfernt. Durch die Verwendung von $\alpha = 0,10$ wird das Verfahren selektiver als das reguläre schrittweise Verfahren in Minitab, bei dem $\alpha = 0,15$ verwendet wird.
- Für die Auswahl der Modellterme werden Prädiktoren durch Subtraktion der Mittelwerte und Division durch die Standardabweichung standardisiert. Das endgültige Modell wird in Einheiten der nicht standardisierten x-Werte angezeigt. Durch die Standardisierung der x-Werte wird der Großteil der Korrelation zwischen linearen und quadratischen Termen beseitigt, wodurch sich die Wahrscheinlichkeit verringert, dass Terme höherer Ordnung unnötigerweise hinzugefügt werden.

Datenprüfungen

Umfang der Daten

Die Trennschärfe ist ein Maß dafür, mit welcher Wahrscheinlichkeit die Nullhypothese aufgrund eines Hypothesentests zurückgewiesen wird, wenn diese falsch ist. Für die Regression gibt die Nullhypothese an, dass keine Beziehung zwischen x und y vorhanden ist. Wenn der Datensatz zu klein ist, reicht die Trennschärfe des Tests möglicherweise nicht aus, um eine Beziehung zwischen x und y zu erkennen, die tatsächlich vorhanden ist. Daher sollte der Datensatz groß genug sein, um eine in der Praxis wichtige Beziehung mit hoher Wahrscheinlichkeit zu erkennen.

Zielstellung

Es sollte festgestellt werden, wie sich der Umfang der Daten auf die Trennschärfe des F-Gesamttests für die Beziehung zwischen x und y sowie auf die Genauigkeit von R_{kor}^2 , der geschätzten Stärke der Beziehung zwischen x und y , auswirkt. Diese Information ist entscheidend, um zu ermitteln, ob der Datensatz groß genug ist, um sicherzustellen, dass die in den Daten beobachtete Stärke der Beziehung ein zuverlässiger Indikator der wahren zugrunde liegenden Stärke der Beziehung ist. Weitere Informationen zu R_{kor}^2 finden Sie in Anhang A.

Methode


Für die Ermittlung des empfohlenen Stichprobenumfangs wurde ein ähnlicher Ansatz wie bei der einfachen Regression verwendet. Untersucht wurde die Streuung in den Werten von R_{kor}^2 , um zu ermitteln, wie groß die Stichprobe sein muss, damit R_{kor}^2 nahe bei ρ_{kor}^2 liegt. Außerdem haben wir bestätigt, dass der empfohlene Stichprobenumfang selbst dann eine ausreichende Trennschärfe bietet, wenn die Stärke der Beziehung zwischen den y - und x -Variablen relativ schwach ist. Weitere Informationen zu den Berechnungen finden Sie in Anhang B.

Ergebnisse

Wie bei der einfachen Regression wird eine Stichprobe empfohlen, die ausreichend umfassend ist, um zu 90 % sicher zu sein, dass der beobachtete Wert von R_{kor}^2 innerhalb von 0,20 von ρ_{kor}^2 liegt. Wir haben festgestellt, dass sich der erforderliche Stichprobenumfang erhöht, wenn weitere Terme zum Modell hinzugefügt werden. Daher wurde der für jede Modellgröße erforderliche Stichprobenumfang berechnet. Der empfohlene Umfang wird auf das nächste Vielfache von 5 gerundet. Wenn das Modell zum Beispiel zusätzlich zur Konstanten acht Koeffizienten aufweist, z. B. vier lineare Terme, drei Wechselwirkungsterme und einen quadratischen Term, beträgt der mindestens erforderliche Stichprobenumfang, um das Kriterium zu erfüllen, $n = 49$. Der Assistent rundet diesen Wert auf den empfohlenen Stichprobenumfang $n = 50$ auf. Weitere Informationen zu spezifischen Empfehlungen für den Stichprobenumfang auf Basis der Anzahl der Terme finden Sie in Anhang B.

Es wurde außerdem überprüft, ob die empfohlenen Stichprobenumfänge eine ausreichende Trennschärfe bieten. Wir haben festgestellt, dass die Trennschärfe bei relativ schwachen Beziehungen, $\rho_{kor}^2 = 0,25$, normalerweise bei 80 % oder mehr liegt. Wenn Sie daher den Empfehlungen des Assistenten zu Stichprobenumfängen folgen, können Sie beim Schätzen der Stärke der Beziehung eine relativ gute Trennschärfe und Genauigkeit sicherstellen.

Basierend auf diesen Ergebnissen zeigt der Assistent für die Prüfung des Umfangs der Daten die folgenden Informationen in der Auswertung an:

Status	Bedingung
	<p>Stichprobenumfang < empfohlen</p> <p>Der Stichprobenumfang ist nicht groß genug, um eine sehr präzise Schätzung der Stärke der Beziehung zu ermöglichen. Maße der Stärke der Beziehung wie R-Quadrat und R-Quadrat (korrigiert) können stark variieren. Um eine genaue Schätzung zu erhalten, sollten für ein Modell dieser Größe umfassendere Stichproben verwendet werden.</p> <p>Stichprobenumfang >= empfohlen</p> <p>Die Stichprobe ist umfassend genug, um eine präzise Schätzung der Stärke der Beziehung zu erzielen.</p>

Ungewöhnliche Daten

Im Rahmen des Regressionsverfahrens im Assistenten haben wir ungewöhnliche Daten als Beobachtungen mit großen standardisierten Residuen oder großen Hebelwirkungswerten definiert. Diese Maße werden normalerweise verwendet, um ungewöhnliche Daten in der Regressionsanalyse zu erkennen (Neter et al., 1996). Da ungewöhnliche Daten erhebliche Auswirkungen auf die Ergebnisse haben können, müssen Sie die Daten möglicherweise korrigieren, damit die Analyse gültig ist. Ungewöhnliche Daten können jedoch auch als Folge der natürlichen Streuung des Prozesses auftreten. Daher ist es wichtig, die Ursache des ungewöhnlichen Verhaltens festzustellen, um zu ermitteln, wie derartige Datenpunkte behandelt werden sollen.

Zielstellung

Es sollte ermittelt werden, wie groß die standardisierten Residuen und die Hebelwirkungswerte sein müssen, damit signalisiert wird, dass ein Datenpunkt ungewöhnlich ist.

Methode

Wir haben die Richtlinien zum Identifizieren ungewöhnlicher Beobachtungen auf der Grundlage des regulären Regressionsverfahrens in Minitab (**Statistik > Regression > Regression**) entwickelt.

Ergebnisse

STANDARDISIERTES RESIDUUM


Das standardisierte Residuum entspricht dem Wert eines Residuums, e_i , dividiert durch einen Schätzwert von dessen Standardabweichung. Im Allgemeinen gilt eine Beobachtung als ungewöhnlich, wenn der Absolutwert des standardisierten Residuums größer als 2 ist. Diese Richtlinie ist jedoch relativ konservativ. Erwartungsgemäß könnten ungefähr 5 % aller Beobachtungen zufällig dieses Kriterium erfüllen (wenn die Fehler normalverteilt sind). Daher ist es wichtig, die Ursache des ungewöhnlichen Verhaltens festzustellen, um zu ermitteln, ob eine Beobachtung tatsächlich ungewöhnlich ist.

HEBELWIRKUNGSWERT

Hebelwirkungswerte beziehen sich nur auf den x-Wert einer Beobachtung und hängen nicht vom y-Wert ab. Eine Beobachtung wird als ungewöhnlich betrachtet, wenn der Hebelwirkungswert mehr als 3 Mal so groß wie die Anzahl der Modellkoeffizienten (p) dividiert durch die Anzahl der Beobachtungen (n) ist. Auch hierbei handelt es sich um einen gängigen Trennwert, obwohl in einigen Lehrbüchern $\frac{2 \times p}{n}$ verwendet wird (Neter et al., 1996).

Wenn die Daten hohe Hebelwirkungspunkte enthalten, sollten Sie überlegen, ob sie sich unangemessen auf die Auswahl des Modells zur Anpassung an die Daten auswirken. Beispielsweise könnte ein einzelner extremer x-Wert dazu führen, dass anstelle eines linearen ein quadratisches Modell ausgewählt wird. Sie sollten feststellen, ob die beobachtete Krümmung im quadratischen Modell mit Ihrer Analyse des Prozesses übereinstimmt. Wenn dies nicht der Fall ist, sollten Sie ein einfacheres Modell an die Daten anpassen oder zusätzliche Daten erfassen, um den Prozess gründlicher zu untersuchen.

Für die Prüfung auf ungewöhnliche Daten werden in der Auswertung des Assistenten die folgenden Statusindikatoren angezeigt:

Status	Bedingung
	Es gibt keine ungewöhnlichen Datenpunkte.
	Es liegt mindestens ein großes standardisiertes Residuum oder mindestens ein hoher Hebelwirkungspunkt vor.

Vorliegen einer Normalverteilung

Ein typische Annahme bei der Regression lautet, dass die Zufallsfehler (ε) normalverteilt sind. Die Annahme der Normalverteilung ist wichtig, wenn Hypothesentests der Schätzungen der Koeffizienten (β) durchgeführt werden. Doch selbst wenn die Zufallsfehler nicht normalverteilt sind, sind die Testergebnisse in der Regel zuverlässig, sofern die Stichprobe umfassend genug ist.

Zielstellung

Es sollte festgestellt werden, wie umfassend die Stichprobe sein muss, um zuverlässige Ergebnisse auf Basis der Normalverteilung zu erhalten. Dabei sollte ermittelt werden, in welchem Maß die tatsächlichen Testergebnisse mit dem Soll-Signifikanzniveau (Alpha oder Wahrscheinlichkeit eines Fehlers 1. Art) für den Test übereinstimmen, d. h., ob der Test die Nullhypothese häufiger oder seltener fälschlicherweise verwirft, als für verschiedene Nicht-Normalverteilungen erwartet wird.

Methode



Um die Wahrscheinlichkeit eines Fehlers 1. Art zu schätzen, wurden mehrere Simulationen mit schiefen Verteilungen sowie Verteilungen mit stärker und schwächer besetzten Randbereichen durchgeführt, die erheblich von der Normalverteilung abweichen. Es wurden Simulationen mit einem Stichprobenumfang von 15 durchgeführt. Der F-Gesamtttest für mehrere Modelle wurde untersucht.

Für jede Bedingung wurden 10.000 Tests durchgeführt. Dabei wurden Zufallszahlen generiert, so dass die Nullhypothese für jeden Test wahr ist. Dann wurden die Tests mit einem Soll-Signifikanzniveau von 0,10 durchgeführt. Es wurde gezählt, wie häufig die Nullhypothese in 10.000 Tests tatsächlich zurückgewiesen wurde, und dieser Anteil wurde mit dem Soll-Signifikanzniveau verglichen. Wenn der Test eine gute Leistung zeigt, sollten die Wahrscheinlichkeiten eines Fehlers 1. Art sehr nahe am Soll-Signifikanzniveau liegen. Weitere Informationen zu den Simulationen finden Sie in Anhang C.

Ergebnisse

Bei beiden F-Gesamttests unterscheiden sich die Wahrscheinlichkeiten, statistisch signifikante Ergebnisse zu finden, bei keiner der Nicht-Normalverteilungen erheblich voneinander. Die Wahrscheinlichkeiten eines Fehlers 1. Art liegen alle zwischen 0,08820 und 0,11850 und somit relativ nahe am Soll-Signifikanzniveau von 0,10.

Da diese Tests bei relativ kleinen Stichproben eine gute Leistung zeigen, testet der Assistent die Daten nicht auf eine Normalverteilung. Stattdessen überprüft der Assistent den Stichprobenumfang und weist auf Stichproben hin, die kleiner als 15 sind. In der Auswertung des Assistenten werden für die Regression die folgenden Statusindikatoren angezeigt:

Status	Bedingung
	Der Stichprobenumfang beträgt mindestens 15, daher ist es kein Problem, wenn keine Normalverteilung vorliegt.
	Da der Stichprobenumfang kleiner als 15 ist, könnte es ein Problem sein, wenn keine Normalverteilung vorliegt. Sie sollten den p-Wert mit Vorsicht interpretieren. Bei kleinen Stichproben ist der p-Wert empfindlich gegenüber Residuenfehlern aufgrund einer fehlenden Normalverteilung.

Literaturhinweise

Neter, J., Kutner, M.H., Nachtsheim, C.J. und Wasserman, W. (1996). *Applied linear statistical Models*. Chicago: Irwin.

Anhang A: Modell und Statistiken

Ein Regressionsmodell, das eine Beziehung zwischen einem Prädiktor x mit einer Antwortvariablen y herstellt, weist die folgende Form auf:

$$Y = f(X) + \varepsilon$$

Hierbei stellt die Funktion $f(x)$ den erwarteten Wert (Mittelwert) von y bei einem gegebenen x dar.

Im Assistenten gibt es zwei Auswahlmöglichkeiten für die Form der Funktion $f(x)$:

Modelltyp	$f(\mathbf{x})$
Linear	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$
Quadratisch	$\beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i < j} \beta_{ij} X_i X_j$

Die Werte der Koeffizienten β sind unbekannt und müssen anhand der Daten geschätzt werden. Die Schätzmethode ist die Methode der kleinsten Quadrate, bei der die Summe der quadrierten Residuen in der Stichprobe minimiert wird:

$$\min \sum_{i=1}^n (Y_i - \hat{f}(X_i))^2.$$

Ein Residuum ist die Differenz zwischen dem beobachteten Wert der Antwortvariablen Y_i und dem angepassten Wert $\hat{f}(X_i)$ auf der Grundlage der geschätzten Koeffizienten. Der minimierte Wert der Summe dieser Quadrate ist die SSE (Summe der quadrierten Fehler) für ein gegebenes Modell.

F-Gesamttest

Diese Methode stellt einen Test des Gesamtmodells dar (linear oder quadratisch). Für die ausgewählte Form der Regressionsfunktion $f(x)$ wird Folgendes getestet:

$$H_0: f(X) \text{ ist konstant}$$

$$H_1: f(X) \text{ ist nicht konstant}$$

Korrigiertes R^2

Das korrigierte R^2 (R_{kor}^2) gibt an, in welchem Ausmaß das Modell die Streuung in der Antwortvariablen auf x zurückführt. Es gibt zwei gängige Verfahren, um die Stärke der beobachteten Beziehung zwischen x und y zu messen:

$$R^2 = 1 - \frac{SSE}{SSTO}$$

Und

$$R_{kor}^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

Dabei gilt Folgendes:

$$SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

SSTO ist die Gesamtsumme der Quadrate, mit der Streuung der Antwortvariablen um den Gesamtmittelwert \bar{Y} dargestellt wird. Mit SSE wird deren Streuung um die Regressionsfunktion $f(x)$ dargestellt. Die Korrektur bei R_{kor}^2 erfolgt für die Anzahl der Koeffizienten (p) im vollständigen Modell, wobei $n - p$ Freiheitsgrade zur Schätzung der Streuung von ε verbleiben. R^2 nimmt nicht ab, wenn dem Modell weitere Koeffizienten hinzugefügt werden. Aufgrund der Korrektur kann R_{kor}^2 jedoch sinken, wenn die zusätzlichen Koeffizienten das Modell nicht verbessern. Wenn ggf. vorhandene zusätzliche Streuung nicht durch Hinzufügen eines weiteren Terms zum Modell erklärt werden kann, nimmt R_{kor}^2 ab und weist so darauf hin, dass der zusätzliche Term nicht nützlich ist. Daher sollte das korrigierte Maß verwendet werden, um Modelle unterschiedlicher Größe zu vergleichen.

Beziehung zwischen F-Test und R_{kor}^2

Die F-Statistik für den Test des Gesamtmodells kann mit Hilfe von SSE und SSTO ausgedrückt werden, die auch bei der Berechnung von R_{kor}^2 verwendet werden:

$$F = \frac{(SSTO - SSE)/(p-1)}{SSE/(n-p)}$$
$$= 1 + \left(\frac{n-1}{p-1} \right) \frac{R_{kor}^2}{1 - R_{kor}^2}$$

Die oben aufgeführten Formeln zeigen, dass die F-Statistik eine ansteigende Funktion von R_{kor}^2 ist. Somit weist der Test H_0 ausschließlich in dem Fall zurück, in dem R_{kor}^2 einen bestimmten Wert überschreitet, der durch das Signifikanzniveau (α) des Tests vorgegeben wird.

Anhang B: Umfang der Daten

In diesem Abschnitt wird erläutert, wie sich n (die Anzahl der Beobachtungen) auf die Trennschärfe des Modellgesamttests und die Genauigkeit von R_{kor}^2 (die geschätzte Stärke des Modells) auswirkt.

Um die Stärke der Beziehung zu quantifizieren, wird die neue Größe ρ_{kor}^2 für die Grundgesamtheit als Gegenstück zur Stichprobenstatistik R_{kor}^2 eingeführt. Zur Erinnerung:

$$R_{kor}^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

Daher wird Folgendes definiert:

$$\rho_{kor}^2 = 1 - \frac{E(SSE|X)/(n-p)}{E(SSTO|X)/(n-1)}$$

Der Operator $E(\cdot|X)$ stellt den erwarteten Wert, also den Mittelwert einer Zufallsvariablen bei einem gegebenen Wert von x dar. Unter der Annahme, dass das korrekte Modell $Y = f(X) + \varepsilon$ mit unabhängig identisch verteilten ε ist, ergibt sich:

$$\begin{aligned} \frac{E(SSE|X)}{n-p} &= \sigma^2 = \text{Var}(\varepsilon) \\ \frac{E(SSTO|X)}{n-1} &= \sum_{i=1}^n \frac{(f(X_i) - \bar{f})^2}{(n-1)} + \sigma^2 \end{aligned}$$

wobei $\bar{f} = \frac{1}{n} \sum_{i=1}^n f(X_i)$.

Daher:

$$\rho_{kor}^2 = \frac{\sum_{i=1}^n (f(X_i) - \bar{f})^2 / (n-1)}{\sum_{i=1}^n (f(X_i) - \bar{f})^2 / (n-1) + \sigma^2}$$

Signifikanz des Gesamtmodells

Beim Testen der statistischen Signifikanz des Gesamtmodells wird angenommen, dass die Zufallsfehler ε unabhängig und normalverteilt sind. Unter der Nullhypothese, dass der Mittelwert von y konstant ($f(X) = \beta_0$) ist, weist die F-Teststatistik eine Verteilung vom Typ $F(p-1, n-p)$ auf. Unter der Alternativhypothese weist die F-Statistik eine nicht zentrale Verteilung vom Typ $F(p-1, n-p, \theta)$ mit folgendem Nichtzentralitätsparameter auf:

$$\begin{aligned} \theta &= \sum_{i=1}^n (f(X_i) - \bar{f})^2 / \sigma^2 \\ &= \frac{(n-1)\rho_{kor}^2}{1 - \rho_{kor}^2} \end{aligned}$$

Die Wahrscheinlichkeit, dass H_0 zurückgewiesen wird, erhöht sich mit dem Nichtzentralitätsparameter, der ansteigende Werte bei sowohl n als auch ρ_{kor}^2 aufweist.

Stärke der Beziehung

Wie für die einfache Regression gezeigt, weist eine statistisch signifikante Beziehung in den Daten nicht zwangsläufig auf eine starke zugrunde liegende Beziehung zwischen x und y hin. Aus diesem Grund richten sich viele Benutzer nach Indikatoren wie R_{kor}^2 , um zu erfahren, wie stark die Beziehung tatsächlich ist. Wenn R_{kor}^2 als Schätzung von ρ_{kor}^2 betrachtet wird, sollte sichergestellt sein, dass die Schätzung relativ nah am wahren Wert von ρ_{kor}^2 liegt.

Für jede mögliche Modellgröße wurde ein angemessener Schwellenwert für einen akzeptablen Stichprobenumfang ermittelt, indem der Minimalwert für n identifiziert wurde, bei dem absolute Differenzen $|R_{kor}^2 - \rho_{kor}^2|$ größer als 0,20 mit einer Wahrscheinlichkeit von höchstens 10 % auftreten. Dies gilt unabhängig vom wahren Wert von ρ_{kor}^2 . Die empfohlenen Stichprobenumfänge $n(T)$ werden in der folgenden Tabelle zusammengefasst, wobei T die Anzahl der Koeffizienten im Modell zusätzlich zum konstanten Term darstellt.

T	n(T)
1-3	40
4-6	45
7-8	50
9-11	55
12-14	60
15-18	65
19-21	70
22-24	75
25-27	80
28-31	85
32-34	90
35-38	95
39-41	100
42-45	105
46-48	110
49-52	115
53-56	120
57-59	125
60-63	130

T	n(T)
64–67	135
68–70	140
71–73	145

Die Trennschärfe des F-Gesamttests des Modells wurde für einen relativ schwachen Wert von $\rho_{kor}^2 = 0,25$ ausgewertet, um zu bestätigen, dass die Trennschärfe bei den empfohlenen Stichprobenumfängen ausreicht. Die in der folgenden Tabelle aufgeführten Modellgrößen stellen den ungünstigsten Fall für jeden Wert von n(T) dar. Kleinere Modelle mit demselben Wert von n(T) weisen eine höhere Trennschärfe auf.

T	n(T)	Trennschärfe bei $\rho_{kor}^2 = 0,25$
3	40	0,902791
6	45	0,854611
8	50	0,850675
11	55	0,831818
14	60	0,820592
18	65	0,798003
21	70	0,796425
24	75	0,796911
27	80	0,798856
31	85	0,789861
34	90	0,794367
38	95	0,788625
41	100	0,794511
45	105	0,790864
48	110	0,797487
52	115	0,795250
56	120	0,793698
59	125	0,800982
63	130	0,800230

T	n(T)	Trennschärfe bei $\rho_{kor}^2 = 0,25$
67	135	0,799906
69	140	0,814664

Anhang C: Vorliegen einer Normalverteilung

Die im Assistenten verwendeten Regressionsmodelle weisen alle die folgenden Form auf:

$$Y = f(X) + \varepsilon$$

Die typische Annahme hinsichtlich der Zufallsterme ε lautet, dass sie unabhängig sind und es sich um identisch verteilte normalverteilte Zufallsvariablen mit einem Mittelwert von null und der gemeinsamen Varianz σ^2 handelt. Die Schätzungen der kleinsten Quadrate der β -Parameter stellen noch immer die besten linearen erwartungstreuen Schätzwerte dar, selbst wenn die Annahme aufgegeben wird, dass die ε normalverteilt sind. Die Annahme einer Normalverteilung wird erst dann wichtig, wenn versucht wird, diesen Schätzungen Wahrscheinlichkeiten zuzuordnen, wie in den Hypothesentests zu $f(x)$.

Es sollte ermittelt werden, wie groß n sein muss, damit die Ergebnisse einer Regressionsanalyse unter Annahme der Normalverteilung zuverlässig sind. Es wurden Simulationen durchgeführt, um die Wahrscheinlichkeiten eines Fehlers 1. Art der Hypothesentests bei einer Reihe von nicht normalen Fehlerverteilungen zu untersuchen.

In der nachfolgenden Tabelle 1 wird der Anteil von 10.000 Simulationen gezeigt, bei dem der F-Gesamttest bei $\alpha = 0,10$ für verschiedene Verteilungen von ε für drei verschiedene Modelle signifikant war. In diesen Simulationen war die Nullhypothese, die angibt, dass keine Beziehung zwischen x und y vorhanden ist, wahr. Die x -Werte wurden durch den Minitab-Befehl RANDOM als multivariate normalverteilte Variablen generiert. Für alle Tests wurde der Stichprobenumfang $n = 15$ verwendet. Alle Modelle umfassten fünf stetige Prädiktoren. Das erste Modell war das lineare Modell mit allen fünf x -Variablen. Das zweite Modell wies alle linearen und quadratischen Terme auf. Das dritte Modell wies alle linearen Terme und sieben der 2-Faktor-Wechselwirkungen auf.

Tabelle 1 Wahrscheinlichkeiten eines Fehlers 1. Art für F-Gesamttests bei $n = 15$ für Nicht-Normalverteilungen

Verteilung	Linear	Linear + quadratisch	Linear + 7 Wechselwirkungen
Normal	0,09910	0,10270	0,10060
t(3)	0,09840	0,11850	0,11800
t(5)	0,09980	0,10010	0,10430
Laplace	0,09260	0,09400	0,09650
Gleichverteilung	0,10630	0,10080	0,09480
Beta(3; 3)	0,09980	0,10120	0,10020
Exponential	0,08820	0,09500	0,09960

Verteilung	Linear	Linear + quadratisch	Linear + 7 Wechselwirkungen
Chi(3)	0,09890	0,11400	0,10970
Chi(5)	0,09730	0,10590	0,10330
Chi(10)	0,10150	0,09930	0,10360
Beta(8; 1)	0,09870	0,10230	0,10490

Die Simulationsergebnisse zeigen, dass die Wahrscheinlichkeiten, statistisch signifikante Ergebnisse zu finden, für die jeweiligen Fehlerverteilungen nicht erheblich vom Nominalwert von 0,10 voneinander abweichen. Die beobachteten Wahrscheinlichkeiten eines Fehlers 1. Art liegen alle zwischen 0,08820 und 0,11850.