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.

# Einfache Regression

## Übersicht

Das Verfahren der einfachen Regression im Assistenten passt lineare und quadratische Modelle mit einem stetigen Prädiktor (x) und einer stetigen Antwortvariablen (y) durch Schätzung nach der Methode der kleinsten Quadrate an. Der Benutzer kann den Modelltyp wählen oder dem Assistenten die Auswahl des am besten angepassten Modells überlassen. 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 und der Koeffizienten ausgewertet. Um schließlich sicherzustellen, dass das Modell sinnvoll ist, muss der ausgewählte Modelltyp die Beziehung zwischen x und y präzise widerspiegeln.

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
- Modellanpassung

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 einem stetigen Prädiktor und einer stetigen Antwortvariablen angepasst. Es können zwei Modelltypen angepasst werden:

- Linear: $F(x) = \beta_0 + \beta_1 X$
- Quadratisch: $F(x) = \beta_0 + \beta_1 X + \beta_2 X^2$

Der Benutzer kann den Modelltyp vor der Analyse auswählen oder dem Assistenten die Auswahl des Modells überlassen. Zum Ermitteln, welches Modell am besten für die Daten geeignet ist, können mehrere Methoden verwendet werden. Um sicherzustellen, dass das Modell sinnvoll ist, muss der ausgewählte Modelltyp die Beziehung zwischen x und y präzise widerspiegeln.

### 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 Methoden untersucht, die normalerweise für die Modellauswahl verwendet werden (Neter et al., 1996). Die erste Methode identifiziert das Modell, in dem der Term der höchsten Ordnung signifikant ist. Die zweite Methode wählt das Modell mit dem höchsten Wert von $R^2_{kor}$ aus. Die dritte Methode wählt das Modell aus, in dem der F-Gesamttest signifikant ist. Weitere Informationen finden Sie in Anhang A.

Um den Ansatz im Assistenten festzulegen, wurden die Methoden untersucht und ihre Berechnungen miteinander verglichen. Darüber hinaus wurde Feedback von Experten in der Qualitätsanalyse eingeholt.

### Ergebnisse

Auf der Grundlage unserer Untersuchungen haben wir beschlossen, die Methode zu verwenden, die das Modell basierend auf der statistischen Signifikanz des Terms höchster Ordnung im Modell auswählt. Der Assistent untersucht zunächst das quadratische Modell und testet, ob der quadratische Term im Modell ($\beta_2$) statistisch signifikant ist. Wenn dieser Term nicht signifikant ist, wird der quadratische Term aus dem Modell verworfen, und der lineare Term ($\beta_1$) wird getestet. Das mit Hilfe dieses Ansatzes ausgewählte Modell wird im Modellauswahlbericht aufgeführt. Wenn der Benutzer ein anderes Modell als der Assistent ausgewählt hat, wird dies im Modellauswahlbericht und in der Auswertung angegeben.

Ein Grund für die Auswahl dieser Methode lag im Feedback der Qualitätsexperten, die angaben, in der Regel einfachere Modelle zu bevorzugen, bei denen nicht signifikante Terme ausgeschlossen werden. Außerdem ergab der Vergleich der Methoden, dass die Verwendung der statistischen Signifikanz des Terms höchster Ordnung im Modell stringenter ist als die

Verwendung der Methode, die das Modell auf Basis des höchsten Werts von $R^2_{kor}$ auswählt. Weitere Informationen finden Sie in Anhang A.

Obwohl zur Auswahl des Modells die statistische Signifikanz des Terms höchster Ordnung im Modell verwendet wird, werden im Modellauswahlbericht auch der Wert von $R^2_{kor}$ und der F-Gesamttest für das Modell angegeben. Informationen zu den angezeigten Statusindikatoren in der Auswertung finden Sie unter „Modellanpassung“ im Abschnitt „Datenprüfungen“ weiter unten.

# 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 $R^2_{kor}$, 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 $R^2_{kor}$ finden Sie in Anhang A.

### Methode

Um die Trennschärfe des F-Gesamttests zu untersuchen, wurden Trennschärfeberechnungen für einen Bereich von Werten von $R^2_{kor}$ und Stichprobenumfängen durchgeführt. Zur Untersuchung der Genauigkeit von $R^2_{kor}$ wurde die Verteilung von $R^2_{kor}$ für verschiedene Werte von $R^2$ korrigiert nach Grundgesamtheit ($\rho^2_{kor}$) und unterschiedliche Stichprobenumfänge simuliert. Untersucht wurde die Streuung in den Werten von $R^2_{kor}$, um zu ermitteln, wie groß die Stichprobe sein muss, damit $R^2_{kor}$ nahe bei $\rho^2_{kor}$ liegt. Weitere Informationen zu den Berechnungen und Simulationen finden Sie in Anhang B.

### Ergebnisse

Wir haben festgestellt, dass die Regression für mittelgroße Stichproben eine gute Trennschärfe aufweist, um Beziehungen zwischen x und y zu erkennen, selbst wenn die Beziehungen nicht stark genug sind, um von praktischem Interesse zu sein. Die genaueren Ergebnisse lauten wie folgt:

- Bei einem Stichprobenumfang von 15 und einer starken Beziehung zwischen x und y ($\rho^2_{kor}$ = 0,65) beträgt die Wahrscheinlichkeit, dass eine statistisch signifikante lineare Beziehung erkannt wird, 0,9969. Wenn der Test daher bei 15 oder mehr Datenpunkten keine statistisch signifikante Beziehung erkennt, ist es wahrscheinlich, dass die tatsächliche Beziehung nicht sehr stark ist ($\rho^2_{kor}$ < 0,65).
- Bei einem Stichprobenumfang von 40 und einer mäßig schwachen Beziehung zwischen x und y ($\rho^2_{kor}$ = 0,25) beträgt die Wahrscheinlichkeit, dass eine statistisch signifikante lineare Beziehung erkannt wird, 0,9398. Bei 40 Datenpunkten ist es daher

wahrscheinlich, dass der F-Test eine Beziehung zwischen x und y erkennt, selbst wenn die Beziehung mäßig schwach ist.

Die Regression kann Beziehungen zwischen x und y relativ leicht erkennen. Wenn Sie daher eine statistisch signifikante Beziehung erkennen, sollten Sie auch mit Hilfe von $R^2_{kor}$ die Stärke der Beziehung auswerten. Es hat sich gezeigt, dass $R^2_{kor}$ bei nicht ausreichenden Stichprobenumfängen nicht sehr zuverlässig ist und zwischen den Stichproben stark variieren kann. Bei einem Stichprobenumfang von mindestens 40 haben wir jedoch festgestellt, dass die Werte von $R^2_{kor}$ stabiler und zuverlässiger sind. Bei einem Stichprobenumfang von 40 können Sie zu 90 % sicher sein, dass der beobachtete Wert von $R^2_{kor}$innerhalb von 0,20 von $\rho^2_{kor}$ liegt, unabhängig vom tatsächlichen Wert und Modelltyp (linear oder quadratisch). Weitere Informationen zu den Ergebnissen der Simulationen finden Sie in Anhang B.

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

| Status | Bedingung |
|---|---|
|  | **Stichprobenumfang < 40**<br>Der Stichprobenumfang ist nicht groß genug, um eine sehr genaue Schätzung der Stärke der Beziehung zu liefern. Messungen der Stärke der Beziehung wie R-Quadrat und R-Quadrat (korrigiert) können stark variieren. Um eine genauere Schätzung zu erhalten, müssen größere Stichproben (typischerweise 40 oder mehr) verwendet werden.<br>**Stichprobenumfang ≥ 40**<br>Ihre Stichprobe ist groß genug um eine genaue Schätzung der Stärke der Beziehung zu erhalten. |

# 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 Modelltyps 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 liegen keine ungewöhnlichen Datenpunkte vor. Ungewöhnliche Datenpunkte können einen starken Einfluss auf die Ergebnisse ausüben. |
|  | Es liegt mindestens ein großes standardisiertes Residuum oder mindestens ein hoher Hebelwirkungswert vor.<br>Sie können mit der Maus auf einen Punkt zeigen oder die Markierungsfunktion verwenden, um die Zeilen des Arbeitsblatts zu identifizieren. Ungewöhnliche Daten können einen erheblichen Einfluss auf die Ergebnisse haben. Versuchen Sie daher, die Ursache für diese Daten zu ermitteln. Korrigieren Sie sämtliche Dateneingabe- und Messfehler. Erwägen Sie, Daten zu entfernen, die auf Ausnahmebedingungen zurückzuführen sind, und die Analyse zu wiederholen. |

# Vorliegen einer Normalverteilung

Ein typische Annahme bei der Regression lautet, dass die Zufallsfehler ($\varepsilon$) normalverteilt sind. Die Annahme der Normalverteilung ist wichtig, wenn Hypothesentests der Schätzungen der Koeffizienten ($\beta$) 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 des 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 für das lineare und das quadratische Modell mit einem Stichprobenumfang von 15 durchgeführt. Untersucht wurden der F-Gesamttest und der Test des Terms höchster Ordnung im Modell.

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 0,05 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

Sowohl beim F-Gesamttest als auch beim Test des Terms höchster Ordnung im Modell unterscheiden sich die Wahrscheinlichkeiten, statistisch signifikante Ergebnisse zu finden, für die jeweiligen Nicht-Normalverteilungen nicht erheblich. Die Wahrscheinlichkeiten eines Fehlers 1. Art liegen alle zwischen 0,038 und 0,0529 und somit sehr nahe am Soll-Signifikanzniveau von 0,05.

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. |

# Modellanpassung

Sie können den Modelltyp vor der Durchführung der Analyse auswählen oder dem Assistenten die Auswahl des Modells überlassen. Zur Auswahl eines angemessenen Modells können mehrere Methoden verwendet werden.

## Zielstellung

Es sollten die verschiedenen Methoden zur Modellauswahl untersucht werden, um zu ermitteln, welcher Ansatz im Assistenten verwendet werden soll.

## Methode

Es wurden drei Methoden untersucht, die normalerweise für die Modellauswahl verwendet werden. Die erste Methode identifiziert das Modell, in dem der Term der höchsten Ordnung signifikant ist. Die zweite Methode wählt das Modell mit dem höchsten Wert von $R^2_{kor}$ aus. Die dritte Methode wählt das Modell aus, in dem der F-Gesamttest signifikant ist. Weitere Informationen finden Sie in Anhang A.

Um den Ansatz im Assistenten festzulegen, wurden die Methoden untersucht und ihre Berechnungen miteinander verglichen. Darüber hinaus wurde Feedback von Experten in der Qualitätsanalyse eingeholt.

## Ergebnisse

Wir haben beschlossen, die Methode zu verwenden, die das Modell basierend auf der statistischen Signifikanz des Terms höchster Ordnung im Modell auswählt. Der Assistent untersucht zunächst das quadratische Modell und testet, ob der quadratische Term im Modell ($\beta_3$) statistisch signifikant ist. Wenn dieser Term nicht signifikant ist, wird der lineare Term ($\beta_1$) im linearen Modell getestet. Das mit Hilfe dieses Ansatzes ausgewählte Modell wird im Modellauswahlbericht aufgeführt. Wenn darüber hinaus der Benutzer ein anderes Modell als der Assistent ausgewählt hat, wird dies im Modellauswahlbericht und in der Auswertung angegeben. Weitere Informationen finden Sie im Abschnitt „Regressionsmethoden“ weiter oben.

Auf der Grundlage unserer Ergebnisse wird in der Auswertung des Assistenten der folgende Statusindikator angezeigt:

| Status | Bedingung |
|---|---|
|  | **Wenn das Modell des Benutzers mit dem am besten angepassten Modell des Assistenten übereinstimmt**<br>Sie sollten die Daten und die Modellanpassung im Hinblick auf Ihre Zielsetzungen auswerten. Betrachten Sie die Darstellungen der Anpassungslinien, um sich hinsichtlich folgender Aspekte zu vergewissern:<br>• Die Stichprobe deckt den Bereich der x-Werte ausreichend ab.<br>• Das Modell ist gut an eine ggf. vorhandene Krümmung in den Daten angepasst (vermeiden Sie eine übermäßige Anpassung).<br>• Die Linie ist in den Bereichen, die von besonderem Interesse sind, gut angepasst.<br><br>**Wenn das Modell des Benutzers nicht mit dem am besten angepassten Modell des Assistenten übereinstimmt**<br>Im Modellauswahlbericht wird ein alternatives Modell angezeigt, das möglicherweise besser geeignet ist. |

# Literaturhinweise

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

# Anhang A: Modellauswahl

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(x) |
|---|---|
| **Linear** | $\beta_0 + \beta_1 X$ |
| **Quadratisch** | $\beta_0 + \beta_1 X + \beta_2 X^2$ |

Die Werte der Koeffizienten $\beta$ 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:

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

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.

Um die im Assistenten verwendete Methode zur Auswahl des Modelltyps festzulegen, wurden drei Optionen ausgewertet:

- Signifikanz des Terms höchster Ordnung im Modell
- F-Gesamttest des Modells
- Wert des korrigierten $R^2$ ($R^2_{kor}$)

## Signifikanz des Terms höchster Ordnung im Modell

Bei diesem Ansatz beginnt der Assistent mit dem quadratischen Modell. Der Assistent testet die Hypothesen für den quadratischen Term im quadratischen Modell:

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

Wenn diese Nullhypothese zurückgewiesen wird, schlussfolgert der Assistent, dass der Koeffizient des quadratischen Terms ungleich null ist und wählt das quadratische Modell aus. Andernfalls testet der Assistent die Hypothesen auf das lineare Modell:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

# F-Gesamttest

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

$H_0: f(X)\ ist\ konstant$

$H_1: f(X)\ ist\ nicht\ konstant$

# Korrigiertes $R^2$

Das korrigierte $R^2$ ($R^2_{kor}$) 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

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

Dabei gilt Folgendes:

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

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 $R^2_{kor}$ erfolgt für die Anzahl der Koeffizienten (p) im vollständigen Modell, wobei n – p Freiheitsgrade zur Schätzung der Streuung von $\varepsilon$ verbleiben. $R^2$ nimmt nicht ab, wenn dem Modell weitere Koeffizienten hinzugefügt werden. Aufgrund der Korrektur kann $R^2_{kor}$ 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 $R^2_{kor}$ ab und weist so darauf hin, dass der zusätzliche Term nicht nützlich ist. Daher sollte das korrigierte Maß verwendet werden, um das lineare und das quadratische Modell zu vergleichen.

# Beziehung zwischen den Methoden zur Modellauswahl

Es sollte untersucht werden, welche Beziehung zwischen den drei Methoden zur Modellauswahl besteht, wie diese berechnet werden und welche Auswirkungen sie aufeinander haben.

Zunächst wurde die Beziehung zwischen der Berechnung des F-Gesamttests und von $R^2_{kor}$ untersucht. Die F-Statistik für den Test des Gesamtmodells kann mit Hilfe von SSE und SSTO ausgedrückt werden, die auch bei der Berechnung von $R^2_{adj}$ verwendet werden:

$$F = \frac{(SSTO - SSE)/(p-1)}{SSE/(n-p)}$$

$$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R^2_{kor}}{1-R^2_{kor}}.$$

Die oben aufgeführten Formeln zeigen, dass die F-Statistik eine ansteigende Funktion von $R^2_{kor}$ ist. Somit weist der Test $H_0$ ausschließlich in dem Fall zurück, in dem $R^2_{kor}$ einen bestimmten Wert überschreitet, der durch das Signifikanzniveau ($\alpha$) des Tests vorgegeben wird. Zur Veranschaulichung wurde das Minimum von $R^2_{kor}$ berechnet, das erforderlich ist, um eine statistische Signifikanz des quadratischen Modells bei $\alpha = 0{,}05$ für verschiedene Stichprobenumfänge zu erzielen, wie in Tabelle 1 weiter unten gezeigt. Bei n = 15 muss der Wert von $R^2_{kor}$ für das Modell beispielsweise mindestens 0,291877 betragen, damit der F-Gesamttest statistisch signifikant ist.

**Tabelle 1** Minimum von $R^2_{kor}$ für einen signifikanten F-Gesamttest für das quadratische Modell bei $\alpha = 0{,}05$ mit verschiedenen Stichprobenumfängen

| Stichprobenumfang | Minimum von $R^2_{kor}$ |
|---|---|
| 4 | 0,992500 |
| 5 | 0,900000 |
| 6 | 0,773799 |
| 7 | 0,664590 |
| 8 | 0,577608 |
| 9 | 0,508796 |
| 10 | 0,453712 |
| 11 | 0,408911 |
| 12 | 0,371895 |
| 13 | 0,340864 |
| 14 | 0,314512 |
| 15 | 0,291877 |
| 16 | 0,272238 |
| 17 | 0,255044 |
| 18 | 0,239872 |
| 19 | 0,226387 |

| Stichprobenumfang | Minimum von $R^2_{kor}$ |
|---|---|
| **20** | 0,214326 |
| **21** | 0,203476 |
| **22** | 0,193666 |
| **23** | 0,184752 |
| **24** | 0,176619 |
| **25** | 0,169168 |
| **26** | 0,162318 |
| **27** | 0,155999 |
| **28** | 0,150152 |
| **29** | 0,144726 |
| **30** | 0,139677 |
| **31** | 0,134967 |
| **32** | 0,130564 |
| **33** | 0,126439 |
| **34** | 0,122565 |
| **35** | 0,118922 |
| **36** | 0,115488 |
| **37** | 0,112246 |
| **38** | 0,109182 |
| **39** | 0,106280 |
| **40** | 0,103528 |
| **41** | 0,100914 |
| **42** | 0,098429 |
| **43** | 0,096064 |
| **44** | 0,093809 |
| **45** | 0,091658 |
| **46** | 0,089603 |
| **47** | 0,087637 |

| Stichprobenumfang | Minimum von $R^2_{kor}$ |
|---|---|
| **48** | 0,085757 |
| **49** | 0,083955 |
| **50** | 0,082227 |

Dann wurde die Beziehung zwischen dem Hypothesentest des Terms höchster Ordnung in einem Modell und $R^2_{kor}$ untersucht. Der Test für den Term höchster Ordnung, z. B. den quadratischen Term in einem quadratischen Modell, kann mit Hilfe der Summen der Quadrate oder mit Hilfe von $R^2_{kor}$ des vollständigen Modells (z. B. quadratisch) und von $R^2_{kor}$ des reduzierten Modells (z. B. linear) ausgedrückt werden:

$$F = \frac{SSE(Reduziert) - SSE(Voll)}{SSE(Voll)/(n-p)}$$

$$= 1 + \frac{(n-p+1)\left(R^2_{kor}(Voll) - R^2_{kor}(Reduziert)\right)}{1 - R^2_{kor}(Voll)}.$$

Die Formeln zeigen, dass die F-Statistik für einen festen Wert von $R^2_{kor}(Reduziert)$ eine ansteigende Funktion von $R^2_{kor}(Voll)$. ist und wie die Teststatistik von der Differenz zwischen den beiden Werten von $R^2_{kor}$ abhängt. Insbesondere muss der Wert für das vollständige Modell größer als der Wert für das reduzierte Modell sein, um einen F-Wert zu erhalten, der groß genug ist, um statistisch signifikant zu sein. Daher ist die Methode, die die Signifikanz des Terms höchster Ordnung zur Auswahl des besten Modells verwendet, stringenter als die Methode, die das Modell mit dem höchsten Wert von $R^2_{kor}$ auswählt. Die Methode mit dem Term höchster Ordnung entspricht auch dem Wunsch vieler Benutzer, einfachere Modelle zu verwenden. Daher haben wir beschlossen, die statistische Signifikanz des Terms höchster Ordnung zu verwenden, um das Modell im Assistenten auszuwählen.

Einige Benutzer neigen dazu, das Modell auszuwählen, das am besten an die Daten angepasst ist, d. h. das Modell mit dem höchsten Wert von $R^2_{kor}$. Der Assistent stellt diese Werte im Modellauswahlbericht und in der Auswertung bereit.

# 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 $R^2_{kor}$ (die geschätzte Stärke des Modells) auswirkt.

Um die Stärke der Beziehung zu quantifizieren, wird die neue Größe $\rho^2_{kor}$ für die Grundgesamtheit als Gegenstück zur Stichprobenstatistik $R^2_{kor}$ eingeführt. Zur Erinnerung:

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

Daher wird Folgendes definiert:

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

Der Operator E(·|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:

$$\frac{E(SSE|X)}{n-p} = \sigma^2 = Var(\varepsilon)$$

$$\frac{E(SSTO|X)}{n-1} = \sum_{i=1}^{n} \frac{(f(X_i) - \bar{f})^2}{(n-1)+\sigma^2} + \sigma^2 \sum_{i=1}^{n} \frac{(f(X_i) - \bar{f})^2}{(n-1)+\sigma^2}$$

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

Daher:

$$\rho^2_{adj} = \frac{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1)}{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1) + \sigma^2}$$

## 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:

$$\theta = \sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/\sigma^2$$

$$= \frac{(n-1)\rho^2_{kor}}{1-\rho^2_{kor}}$$

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

Mit Hilfe der oben genannten Formel wurde die Trennschärfe des F-Gesamttests für einen Bereich von Werten von $\rho^2_{kor}$ mit n = 15 für das lineare und das quadratische Modell berechnet. Die Ergebnisse finden Sie in Tabelle 2.

**Tabelle 2** Trennschärfe für lineare und quadratische Modelle mit unterschiedlichen Werten von $\rho^2_{kor}$ bei n=15

| $\rho^2_{kor}$ | θ | Trennschärfe von F Linear | Trennschärfe von F Quadratisch |
|---|---|---|---|
| **0,05** | **0,737** | 0,12523 | 0,09615 |
| **0,10** | **1,556** | 0,21175 | 0,15239 |
| **0,15** | **2,471** | 0,30766 | 0,21896 |
| **0,20** | **3,500** | 0,41024 | 0,29560 |
| **0,25** | **4,667** | 0,51590 | 0,38139 |
| **0,30** | **6,000** | 0,62033 | 0,47448 |
| **0,35** | **7,538** | 0,71868 | 0,57196 |
| **0,40** | **9,333** | 0,80606 | 0,66973 |
| **0,45** | **11,455** | 0,87819 | 0,76259 |
| **0,50** | **14,000** | 0,93237 | 0,84476 |
| **0,55** | **17,111** | 0,96823 | 0,91084 |
| **0,60** | **21,000** | 0,98820 | 0,95737 |
| **0,65** | **26,000** | 0,99688 | 0,98443 |
| **0,70** | **32,667** | 0,99951 | 0,99625 |
| **0,75** | **42,000** | 0,99997 | 0,99954 |
| **0,80** | **56,000** | 1,00000 | 0,99998 |
| **0,85** | **79,333** | 1,00000 | 1,00000 |
| **0,90** | **126,000** | 1,00000 | 1,00000 |
| **0,95** | **266,000** | 1,00000 | 1,00000 |

Insgesamt haben wir festgestellt, dass der Test eine hohe Trennschärfe aufweist, wenn die Beziehung zwischen x und y stark ist und der Stichprobenumfang mindestens 15 beträgt. Wenn zum Beispiel $\rho^2_{kor}$ = 0,65, zeigt Tabelle 2, dass die Wahrscheinlichkeit, eine statistisch signifikante lineare Beziehung bei $\alpha = 0{,}05$ zu erkennen, 0,99688 beträgt. In weniger als 0,5 % der Stichproben würde der F-Test eine solche starke Beziehung nicht erkennen. Selbst bei einem quadratischen Modell würde die Beziehung mit dem F-Test in weniger als 2 % der

Stichproben nicht erkannt werden. Wenn der Test daher bei 15 oder mehr Beobachtungen keine statistisch signifikante Beziehung erkennt, ist dies ein gutes Anzeichen dafür, dass die tatsächliche Beziehung (sofern vorhanden), einen Wert von $\rho_{kor}^2$ aufweist, der kleiner als 0,65 ist. Beachten Sie, dass $\rho_{kor}^2$ auch kleiner als 0,65 sein kann, um in der Praxis von Interesse zu sein.

Außerdem wollten wir die Trennschärfe des F-Gesamttests untersuchen, wenn der Stichprobenumfang größer war (n = 40). Wir haben festgestellt, dass der Stichprobenumfang n = 40 ein wichtiger Schwellenwert für die Genauigkeit von $R_{kor}^2$ ist (siehe Abschnitt „Stärke der Beziehung" weiter unten), und wollten die Trennschärfewerte für den Stichprobenumfang auswerten. Die Trennschärfe der F-Gesamttests wurde für einen Bereich von Werten von $\rho_{kor}^2$ mit n = 40 für das lineare und das quadratische Modell berechnet. Die Ergebnisse finden Sie in Tabelle 3.

**Tabelle 3** Trennschärfe für lineare und quadratische Modelle mit unterschiedlichen Werten von $\rho_{kor}^2$ bei n = 40

| $\rho_{kor}^2$ | θ | Trennschärfe von F Linear | Trennschärfe von F Quadratisch |
|---|---|---|---|
| **0,05** | **2,0526** | 0,28698 | 0,21541 |
| **0,10** | **4,3333** | 0,52752 | 0,41502 |
| **0,15** | **6,8824** | 0,72464 | 0,60957 |
| **0,20** | **9,7500** | 0,86053 | 0,76981 |
| **0,25** | **13,0000** | 0,93980 | 0,88237 |
| **0,30** | **16,7143** | 0,97846 | 0,94925 |
| **0,35** | **21,0000** | 0,99386 | 0,98217 |
| **0,40** | **26,0000** | 0,99868 | 0,99515 |
| **0,45** | **31,9091** | 0,99980 | 0,99905 |
| **0,50** | **39,0000** | 0,99998 | 0,99988 |
| **0,55** | **47,6667** | 1,00000 | 0,99999 |
| **0,60** | **58,5000** | 1,00000 | 1,00000 |
| **0,65** | **72,4286** | 1,00000 | 1,00000 |

Wir haben festgestellt, dass die Trennschärfe auch dann hoch war, wenn die Beziehung zwischen x und y relativ schwach war. Beispiel: Sogar wenn $\rho_{kor}^2$ = 0,25, zeigt Tabelle 3, dass die Wahrscheinlichkeit, eine statistisch signifikante lineare Beziehung bei $\alpha = 0{,}05$ zu erkennen, 0,93980 beträgt. Bei 40 Beobachtungen ist es unwahrscheinlich, dass der F-Test eine Beziehung zwischen x und y nicht erkennt, selbst wenn diese Beziehung relativ schwach ist.

# Stärke der Beziehung

Wie bereits 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 $R^2_{kor}$, um zu erfahren, wie stark die Beziehung tatsächlich ist. Wenn $R^2_{kor}$ als Schätzung von $\rho^2_{kor}$ betrachtet wird, sollte sichergestellt sein, dass die Schätzung relativ nah am wahren Wert von $\rho^2_{kor}$ liegt.

Zur Veranschaulichung der Beziehung zwischen $R^2_{kor}$ und $\rho^2_{kor}$ wurde die Verteilung von $R^2_{kor}$ für verschiedene Werte von $\rho^2_{kor}$ simuliert, um zu ermitteln, welche Streuung $R^2_{kor}$ für verschiedene Werte von n aufweist. Die Grafiken in den folgenden Abbildungen 1–4 zeigen Histogramme von 10.000 simulierten Werten von $R^2_{kor}$. In jedem Paar von Histogrammen ist der Wert von $\rho^2_{kor}$ identisch, so dass die Streuung von $R^2_{kor}$ für Stichproben des Umfangs 15 bis zu Stichproben des Umfangs 40 verglichen werden kann. Getestet wurde $\rho^2_{kor}$ bei Werten von 0,0, 0,30, 0,60 und 0,90. Alle Simulationen wurden mit dem linearen Modell durchgeführt.

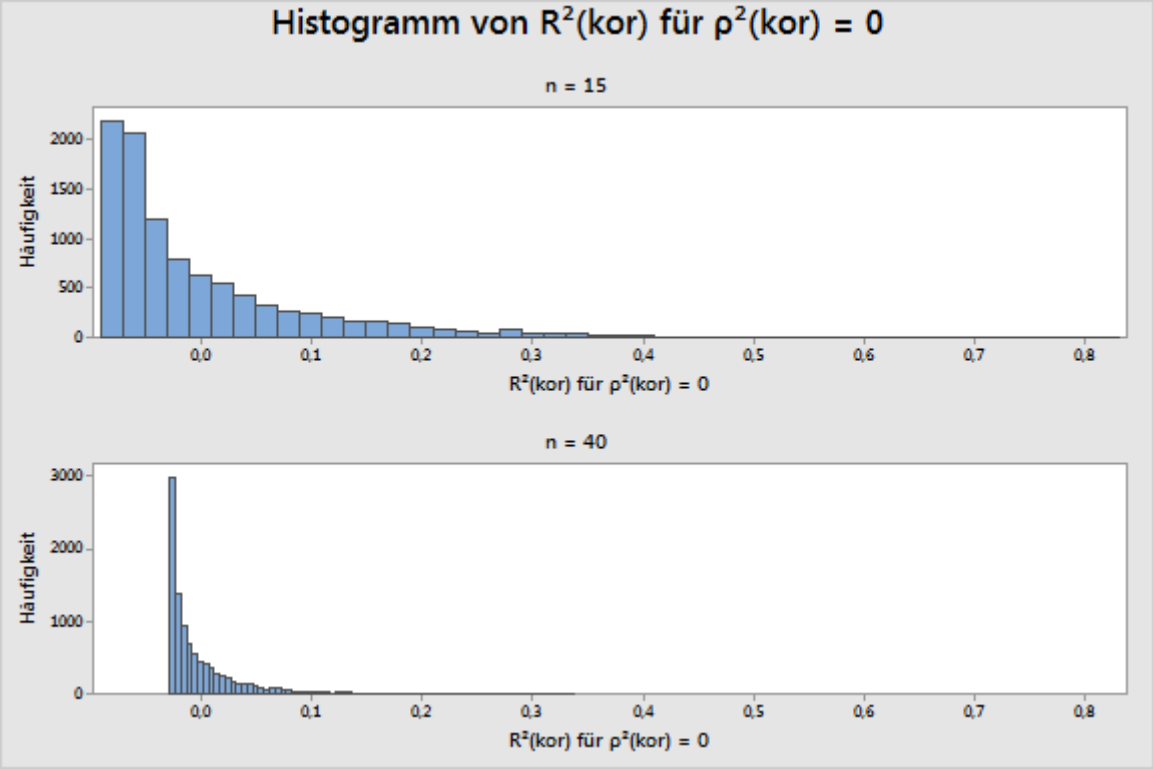


**Abbildung 1** Simulierte Werte von $R^2_{kor}$ für $\rho^2_{kor}$ = 0,0 bei n=15 und n=40

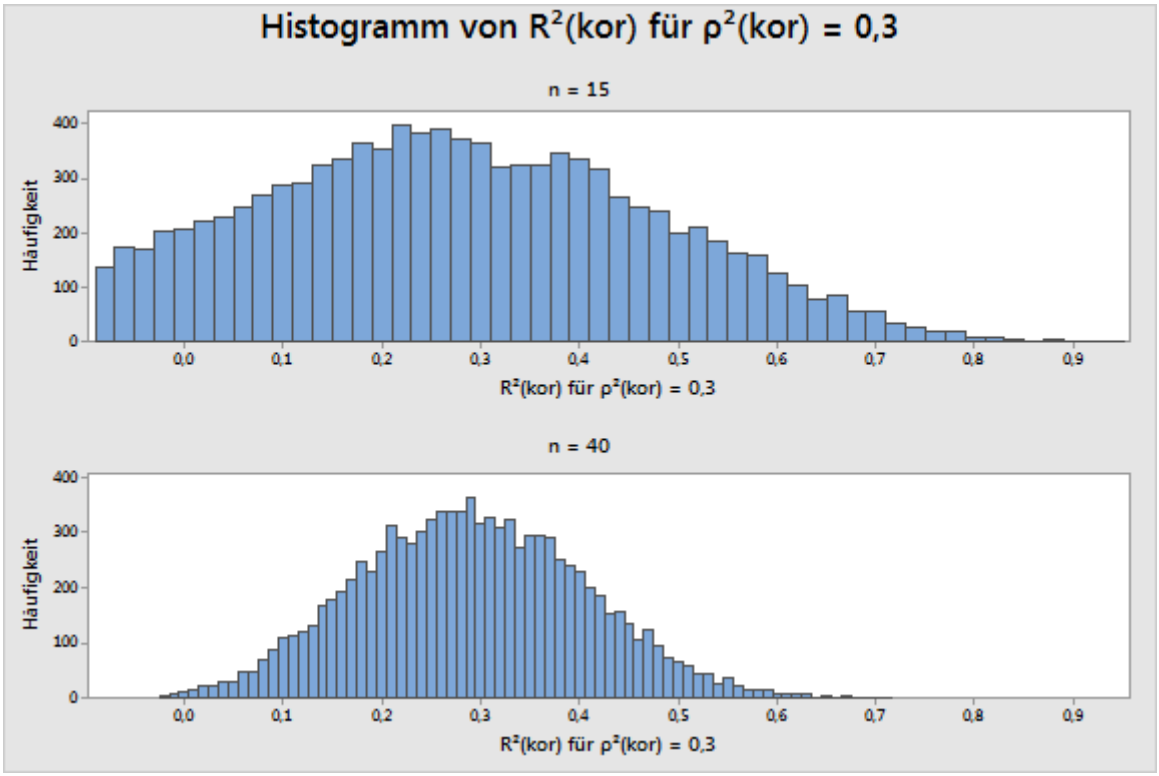


**Abbildung 2** Simulierte Werte von $R^2_{kor}$ für $\rho^2_{kor}$ = 0,30 bei n=15 und n=40

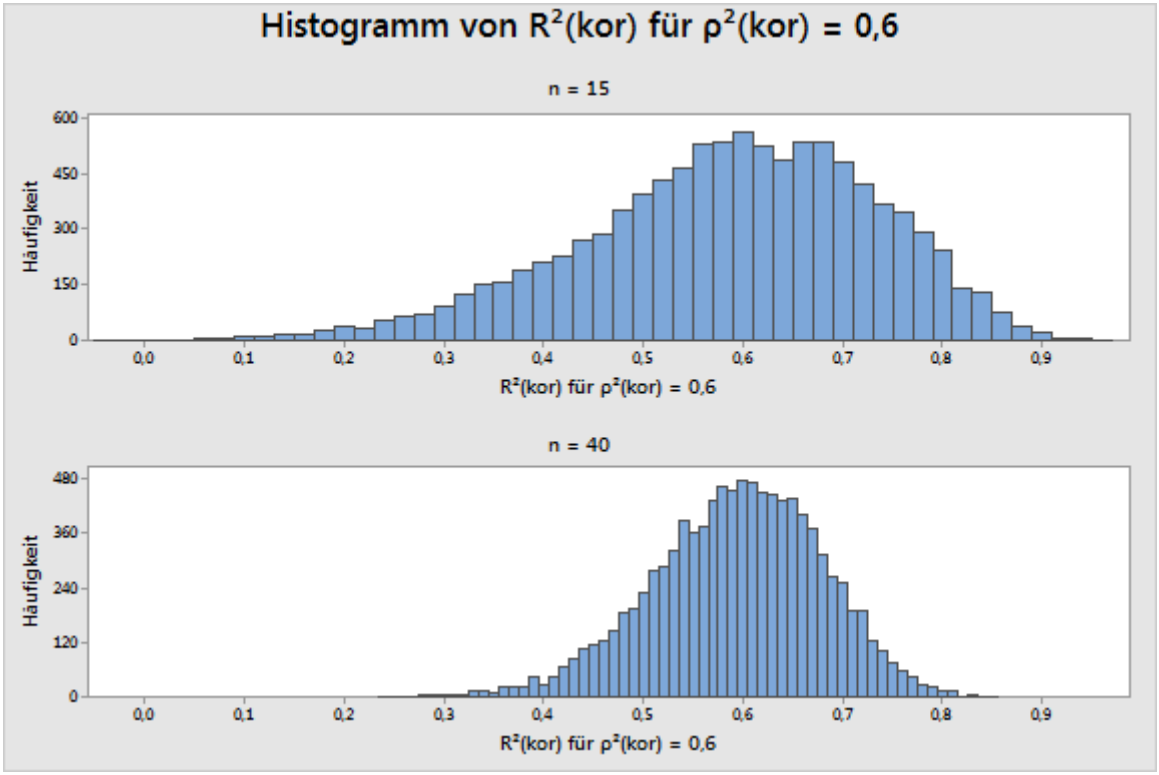


**Abbildung 3** Simulierte Werte von $R^2_{kor}$ für $\rho^2_{kor}$ = 0,60 bei n=15 und n=40

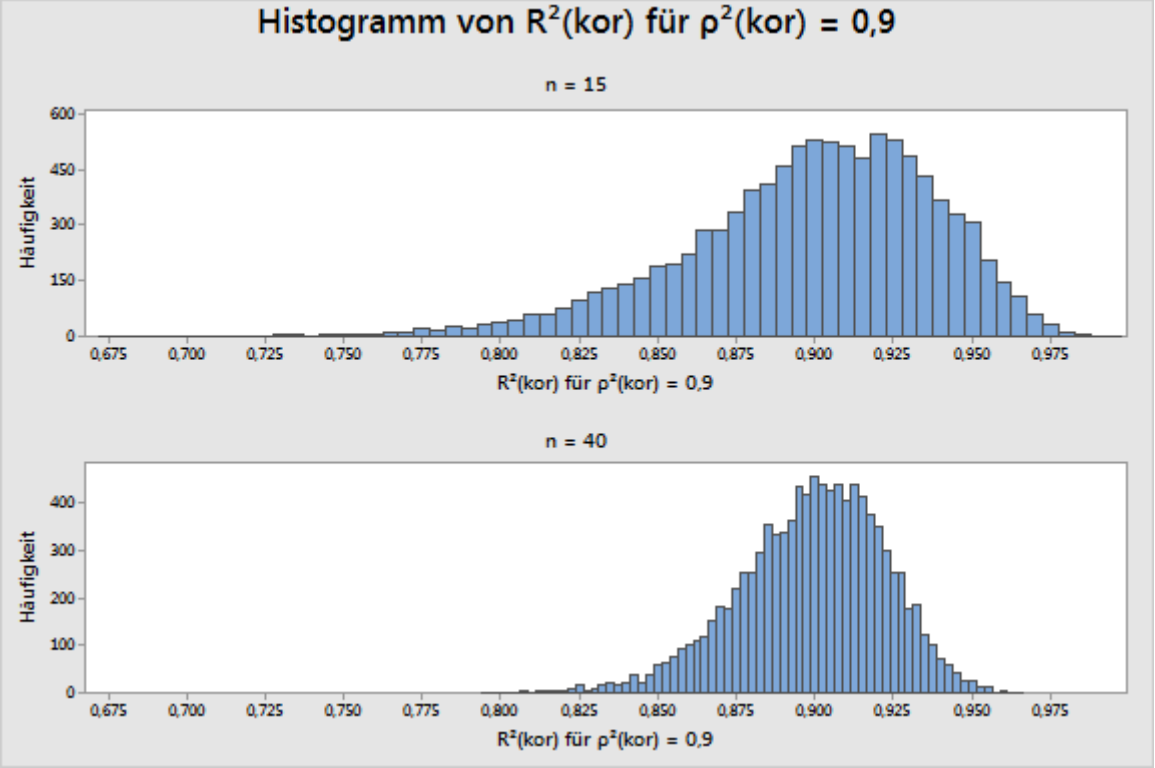


**Abbildung 4** Simulierte Werte von $R^2_{kor}$ für $\rho^2_{kor}$ = 0,90 bei n=15 und n=40

Insgesamt zeigen die Simulationen, dass die tatsächliche Stärke der Beziehung ($\rho^2_{kor}$) und die in den Daten beobachtete Beziehung ($R^2_{kor}$) eine erhebliche Differenz aufweisen können. Durch Erhöhen des Stichprobenumfangs von 15 auf 40 wird die wahrscheinliche Höhe der Differenz deutlich reduziert. Es wurde ermittelt, dass 40 Beobachtungen einen angemessenen Schwellenwert darstellen, indem der Minimalwert für n identifiziert wurde, bei dem absolute Differenzen $|R^2_{kor} - \rho^2_{kor}|$ größer als 0,20 mit einer Wahrscheinlichkeit von höchstens 10 % auftreten. Dies gilt unabhängig vom wahren Wert von $\rho^2_{kor}$ in den betrachteten Modellen. Beim linearen Modell war der schwierigste Fall $\rho^2_{kor}$ = 0,31, wobei n = 36 erforderlich war. Der schwierigste Fall beim quadratischen Modell war $\rho^2_{kor}$ = 0,30, wobei n = 38 erforderlich war. Bei einem Stichprobenumfang von 40 können Sie zu 90 % sicher sein, dass der beobachtete Wert von $R^2_{kor}$ innerhalb von 0,20 von $\rho^2_{kor}$ liegt, unabhängig davon, wie der Wert lautet und ob Sie das lineare oder das quadratische Modell verwenden.

# Anhang C: Vorliegen einer Normalverteilung

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

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

Die typische Annahme hinsichtlich der Zufallsterme $\varepsilon$ lautet, dass sie unabhängig sind und es sich um identisch verteilte normalverteilte Zufallsvariablen mit einem Mittelwert von null und der gemeinsamen Varianz $\sigma^2$ handelt. Die Schätzungen der kleinsten Quadrate der $\beta$-Parameter stellen noch immer die besten linearen erwartungstreuen Schätzwerte dar, selbst wenn die Annahme aufgegeben wird, dass die $\varepsilon$ 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 4 wird der Anteil von 10.000 Simulationen gezeigt, bei dem der F-Gesamttest bei $\alpha = 0{,}05$ für verschiedene Verteilungen von ε für das lineare und das quadratische Modell signifikant war. In diesen Simulationen war die Nullhypothese, die angibt, dass keine Beziehung zwischen x und y vorhanden ist, wahr. Die x-Werte waren regelmäßig über ein Intervall verteilt. Für alle Tests wurde der Stichprobenumfang n=15 verwendet.

**Tabelle 4** Wahrscheinlichkeiten eines Fehlers 1. Art für F-Gesamttests für lineare und quadratische Modelle bei n = 15 für Nicht-Normalverteilungen

| Verteilung | Linear signifikant | Quadratisch signifikant |
|---|---|---|
| **Normal** | 0,04770 | 0,05060 |
| **t(3)** | 0,04670 | 0,05150 |
| **t(5)** | 0,04980 | 0,04540 |
| **Laplace** | 0,04800 | 0,04720 |
| **Gleichverteilung** | 0,05140 | 0,04450 |
| **Beta(3; 3)** | 0,05100 | 0,05090 |
| **Exponential** | 0,04380 | 0,04880 |
| **Chi(3)** | 0,04860 | 0,05210 |
| **Chi(5)** | 0,04900 | 0,05260 |
| **Chi(10)** | 0,04970 | 0,05000 |

| Verteilung | Linear signifikant | Quadratisch signifikant |
|---|---|---|
| **Beta(8; 1)** | 0,04780 | 0,04710 |

Dann wurde der Test des Terms höchster Ordnung zur Auswahl des besten Modells untersucht. Für jede Simulation wurde ermittelt, ob der quadratische Term signifikant war. In Fällen, in denen der quadratische Term nicht signifikant war, wurde ermittelt, ob der lineare Term signifikant war. In diesen Simulationen war die Nullhypothese wahr, das Soll-$\alpha = 0,05$ und n = 15.

**Tabelle 5** Wahrscheinlichkeiten eines Fehlers 1. Art für Tests des Terms höchster Ordnung für lineare oder quadratische Modelle bei n = 15 für Nicht-Normalverteilungen

| Verteilung | Quadratisch | Linear |
|---|---|---|
| **Normal** | 0,05050 | 0,04630 |
| **t(3)** | 0,05120 | 0,04300 |
| **t(5)** | 0,04710 | 0,04820 |
| **Laplace** | 0,04770 | 0,04660 |
| **Gleichverteilung** | 0,04670 | 0,04900 |
| **Beta(3; 3)** | 0,05000 | 0,04860 |
| **Exponential** | 0,04600 | 0,03800 |
| **Chi(3)** | 0,05110 | 0,04290 |
| **Chi(5)** | 0,05290 | 0,04490 |
| **Chi(10)** | 0,04970 | 0,04610 |
| **Beta(8; 1)** | 0,04770 | 0,04380 |

Die Simulationsergebnisse zeigen, dass die Wahrscheinlichkeiten, statistisch signifikante Ergebnisse zu finden, sowohl beim F-Gesamttest als auch beim Test des Terms höchster Ordnung für die jeweiligen Fehlerverteilungen nicht erheblich voneinander abweichen. Die Wahrscheinlichkeiten eines Fehlers 1. Art liegen alle zwischen 0,038 und 0,0529.