

Multiple Regression

Bedingte Koeffizienten, Modellvergleiche und gemeinsame Prädiktorinformation

Ratiomera Statistics

Inhaltsverzeichnis

Zusammenfassung der Kursseite	2
Eine Gleichung, mehrere Rollen	2
Tests und Modellvergleich	2
Interaktionen und Diagnostik	3
Erweiterte Referenz	4
Zweck und Grundlagen	4
Zentrale Ideen	4
Formelleitfaden	5
Die erklärende Abbildung lesen	8
Checkliste zur Interpretation	9
Verbindung zu anderen Themen	9

Zusammenfassung der Kursseite

Die multiple Regression modelliert eine quantitative Ergebnisvariable gleichzeitig aus mehreren Prädiktortermen. Jeder Koeffizient ist bedingt: Er beschreibt eine angepasste Änderung für einen Prädiktor, während die anderen Terme in der Gleichung festgehalten werden.

Eine Gleichung, mehrere Rollen

Für zwei quantitative Prädiktoren gilt

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

Tabelle 1: Das Modell kombiniert Prädiktorbeiträge, aber jeder Term behält eine bestimmte bedingte Bedeutung.

Bestandteil	Interpretation
Achsenabschnitt	Angepasster Mittelwert, wenn alle quantitativen Prädiktoren null sind und alle kategorialen Prädiktoren auf ihrer Referenzstufe liegen
Quantitativer Koeffizient	Angepasste Ergebnisänderung für eine zusätzliche Prädiktoreinheit, bedingt auf die anderen Terme
Dummy-Koeffizient	Bedingter Unterschied zur benannten Referenzkategorie bei den quantitativen Referenzwerten
Interaktionskoeffizient	Änderung einer Steigung oder eines Gruppenunterschieds, wenn sich eine andere Variable verändert
Anpassungswert und Residuum	Kombinierte Modellvorhersage und beobachtete Abweichung für einen Fall

Beim Zentrieren wird von einem quantitativen Prädiktor ein sinnvoller Referenzwert, etwa sein Stichprobenmittelwert, abgezogen. Dies verändert den Achsenabschnitt und die an einer Interaktion beteiligten Terme niedrigerer Ordnung, aber nicht die Anpassungswerte oder die Gesamtanpassung. Zentrieren verbessert die Interpretation. Es entfernt keine Prädiktorüberlappung.

Tests und Modellvergleich

- Ein t -Test für einen einzelnen Koeffizienten fragt, ob ein Populationskoeffizient bedingt auf die anderen enthaltenen Terme null ist.
- Der globale F -Test fragt, ob alle geprüften Steigungen gemeinsam null sind.
- Ein F -Test für verschachtelte Modelle fragt, ob ein festgelegter Block zusätzlicher Terme die Anpassung gegenüber einem kleineren Modell verbessert.
- R^2 kann beim Hinzufügen von Prädiktoren nicht kleiner werden. Das korrigierte R^2 enthält eine Komplexitätsstrafe und kann abnehmen.
- AIC vergleicht die relative Anpassungs-Komplexitäts-Balance der angegebenen Kandidatenmodelle. Es identifiziert kein wahres oder kausales Modell.

Prädiktorüberlappung kann einzelne Schätzungen unpräziser machen. Ein Varianzinflationsfaktor fasst zusammen, wie stark ein Prädiktor durch die übrigen Prädiktoren linear erklärt wird. Er ist eine Diagnose für Koeffizientenunsicherheit, keine allgemeine Bestehensgrenze.

Interaktionen und Diagnostik

Wenn eine Interaktion vorhanden ist, interpretiere bedingte Steigungen oder Anpassungswerte bei sinnvollen Kombinationen. Interpretiere einen Haupteffektkoeffizienten nicht als einen universellen Effekt über alle Stufen der interagierenden Variable.

Verwende Residuen-gegen-Anpassungswerte- und Normal-Q-Q-Diagramme, um Modellform, Streuung und grobe Fehlerform zu prüfen. Ein Hebelwert beschreibt ungewöhnliche Prädiktorkombinationen. Die Cook-Distanz verbindet Prädiktorposition und Residualinformation, um Sensitivitätsprüfungen zu priorisieren. Gültige Fälle werden nicht nur wegen einer grossen Diagnosegrösse gelöscht.

Die Entwicklung von Thema 4 bis 7 ist eine zusammenhängende Geschichte: Korrelation beschreibt gepaarte Bewegung, einfache Regression gibt ihr Richtung und Einheiten, partielle Korrelation macht Bereinigung sichtbar und multiple Regression setzt mehrere bedingte Beziehungen in eine angepasste Gleichung. Thema 8 zeigt, dass kategoriale Gruppenvergleiche zur selben allgemeinen Familie linearer Modelle gehören.

Erweiterte Referenz

Zweck und Grundlagen

Die multiple lineare Regression modelliert den bedingten Mittelwert einer numerischen Zielvariable mithilfe von zwei oder mehr Prädiktoren. Sie erweitert die Idee der angepassten Geraden aus der einfachen Regression. Statt sich entlang einer einzigen Prädiktorachse zu bewegen, kann sich der angepasste Wert der Zielvariable über mehrere Prädiktordimensionen verändern. Jeder Koeffizient beschreibt die mit einem Prädiktor verbundene angepasste Veränderung, während die anderen Prädiktoren im Modell konstant gehalten werden.

Diese letzte Formulierung ist eine Vergleichsregel und keine physische Handlung. Betrachte ein Modell, das den Punktwert im statistischen Denken aus Übungsstunden und Vorwissenswert vorhersagt. Der Übungskoeffizient vergleicht Fälle, die sich um eine Übungsstunde unterscheiden, aber denselben modellierten Vorwissenswert besitzen. Der Vorwissenskoeffizient vergleicht Fälle, die sich um eine Vorwissenseinheit unterscheiden, aber dieselben modellierten Übungsstunden besitzen. Ob diese Vergleiche durch die Daten gut gestützt sind, hängt von den beobachteten Kombinationen der Prädiktoren und von der Angemessenheit des linearen Modells ab.

Prädiktoren können gemeinsame Informationen enthalten. Übungsstunden und Vorwissen können beide mit denselben Teilen der Zielvariable und zugleich miteinander verbunden sein. Eine Steigung aus einer einfachen Regression enthält die gesamte Variation der Zielvariable, die auf ihren einzigen Prädiktor ausgerichtet ist. Ein multipler Regressionskoeffizient isoliert dagegen die bedingte lineare Komponente dieses Prädiktors unter Berücksichtigung der anderen. Eine Veränderung zwischen dem einfachen und dem bedingten Koeffizienten ist zu erwarten und aufschlussreich, erfordert aber eine inhaltliche Interpretation.

Modellbestandteil	Bedeutung	Zu stellende Frage
Achsenabschnitt b_0	Angepasster Wert der Zielvariable, wenn alle numerischen Prädiktoren null sind und kategoriale Prädiktoren auf ihren Referenzstufen liegen	Sind diese Referenzwerte sinnvoll und in den Daten vertreten?
Numerischer Koeffizient b_j	Angepasste Differenz der Zielvariable bei einer Zunahme des Prädiktors um eine Einheit, während die anderen konstant bleiben	Welche Variablen werden in welchen Einheiten konstant gehalten?
Indikatorkoeffizient	Angepasste Differenz zu einer angegebenen Referenzkategorie	Welche Kategorie ist die Referenz?
Interaktionskoeffizient	Veränderung der Steigung eines Prädiktors über Werte oder Gruppen eines anderen Prädiktors	Welche bedingte Steigung wird verändert?

Zentrale Ideen

Beginne mit einem Modell, das aus der Forschungsfrage hervorgeht, und nicht mit einer mechanischen Suche über jede verfügbare Variable. Ein Prädiktor kann eine interessierende Einflussgrösse, eine geplante Kontrolle, einen Gruppenvergleich oder einen für die funktionale Form benötigten Term darstellen. Erkläre jede Rolle. Das Hinzufügen einer Variable verändert die Frage, die jeder bedingte Koeffizient beantwortet. Zwei Modelle mit verschiedenen Prädiktormengen sind deshalb keine austauschbaren Beschreibungen.

Kategoriale Prädiktoren gelangen über Indikatorvariablen in das Modell. Bei drei Tutoriumsformaten wird eine Kategorie zur Referenz. Zwei Indikatoren vergleichen die anderen Formate mit ihr.

Das Wechseln der Referenzkategorie verändert den ausgegebenen Achsenabschnitt und die Vergleiche, aber nicht die angepassten Werte. Die Referenz muss in Tabellen und im Text genannt werden.

Eine Interaktion bedeutet, dass sich der bedingte Zusammenhang eines Prädiktors über Werte eines anderen Prädiktors unterscheidet. Bei einer Interaktion zwischen Übung und Format gibt es keine einzelne Übungssteigung für alle Formate. Der Hauptkoeffizient der Übung ist die Steigung innerhalb des Referenzformats. Jeder Interaktionskoeffizient zeigt, wie sich die Steigung eines anderen Formats davon unterscheidet. Interpretiere die zusammengehörigen Koeffizienten gemeinsam und zeige angepasste Geraden oder vorhergesagte Werte.

Beurteilungsebene	Nützliche Grösse oder Darstellung	Beitrag zur Beurteilung
Einzelner Koeffizient	Schätzung, Standardfehler, Intervall, t -Test	Bedingte Richtung, Grösse und Unsicherheit
Hinzugefügter Prädiktorblock	F -Test verschachtelter Modelle und Veränderung von R^2	Ob der Block zusätzliche modellierte Variation der Zielvariable beiträgt
Gesamtmodell	R^2 , korrigiertes R^2 , Gesamt- F -Test	Anpassung in der Stichprobe und gemeinsame Evidenz für die Prädiktormenge
Modellangemessenheit	Residual-, Quantil-, Hebelwirkungs- und Einflussgrafiken	Ob die angepasste Form und die Annahmen zur Unsicherheit glaubwürdig sind

R^2 kann beim Hinzufügen von Prädiktoren nicht abnehmen, auch wenn diese wenig nützliche Information beitragen. Das korrigierte R^2 enthält eine Berücksichtigung der Prädiktoranzahl und kann abnehmen. Informationskriterien wie AIC gleichen ebenfalls Anpassung und Modellkomplexität aus. Vergleiche sind jedoch nur zwischen Modellen sinnvoll, die an dieselbe Zielvariable und dieselben Beobachtungen angepasst wurden. Kein einzelner Anpassungskennwert ersetzt Residualdiagnosen oder inhaltliches Urteilsvermögen.

Eine starke Prädiktorüberlappung bedeutet, dass Prädiktoren stark überlappende lineare Informationen enthalten. Sie kann die Standardfehler der Koeffizienten vergrössern und einzelne Schätzungen instabil machen, während angepasste Werte weiterhin nützlich bleiben. In einem ansonsten geeigneten Modell erzeugt sie nicht von selbst eine Verzerrung. Untersuche die Beziehungen zwischen den Prädiktoren, die Unsicherheit der Koeffizienten und das Design. Entferne keine konzeptuell notwendige Variable mit dem einzigen Ziel, einen anderen Koeffizienten signifikant zu machen. Wende ebenso wenig einen allgemeingültigen numerischen Grenzwert an, den das registrierte Kursmaterial nicht stützt.

Die Residualannahmen erweitern jene der einfachen Regression: ein geeigneter linearer bedingter Mittelwert, unabhängige Fehler, eine geeignete Varianz über die angepassten Werte und eine für die beabsichtigte Inferenz angemessene Residualverteilung. Einflussreiche Fälle können mehrere Koeffizienten verändern. Extrapolation kann zudem bei Kombinationen von Prädiktoren auftreten, selbst wenn jeder einzelne Wert innerhalb seines beobachteten Bereichs liegt.

Formelleitfaden

Für p Prädiktoren lautet das Populationsmodell:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

Der angepasste Stichprobenwert verwendet geschätzte Koeffizienten. Das Residuum bleibt beobachteter minus angepasster Wert:

$$\hat{y}_i = b_0 + \sum_{j=1}^p b_j x_{ji}, \quad e_i = y_i - \hat{y}_i$$

Bei zwei quantitativen Prädiktoren lassen sich die bedingten Steigungen mithilfe der drei paarweisen Korrelationen und der Standardabweichungen der Variablen ausdrücken:

$$b_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2} \frac{s_Y}{s_{X_1}}, \quad b_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{1 - r_{12}^2} \frac{s_Y}{s_{X_2}}.$$

Die Subtraktion stellt jene Korrelationsinformation dar, die mit dem anderen Prädiktor geteilt wird. Das Verhältnis der Standardabweichungen übersetzt das Ergebnis zurück in Einheiten der Zielvariable pro Prädiktoreinheit. Falls $|r_{12}| = 1$ gilt, ist der Nenner null. Die beiden getrennten Steigungen können dann aus diesem Modell nicht geschätzt werden.

Der Residualstandardfehler beschreibt die typische nicht erklärte Streuung in Einheiten der Zielvariable:

$$s_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n - p - 1}}.$$

Dabei zählt p die Prädiktorparameter ohne Achsenabschnitt. Ein kategorialer Prädiktor kann mehr als einen Parameter benötigen.

Für einen quantitativen Prädiktor lautet der standardisierte Koeffizient:

$$\hat{\beta}_j = b_j \frac{s_{X_j}}{s_Y}.$$

Er beschreibt die bedingte angepasste Veränderung in Standardabweichungen der Zielvariable bei einer Prädiktordifferenz von einer Standardabweichung. Anders als eine bivariate Korrelation ist er von den anderen Modelltermen abhängig und nicht auf das Intervall von -1 bis $+1$ beschränkt.

Das Bestimmtheitsmass vergleicht die Residualquadratsumme mit der Gesamtquadratsumme:

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}}$$

Das korrigierte R^2 berücksichtigt die Stichprobengröße n und die Anzahl Prädiktoren p :

$$R_{\text{adjusted}}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

Die Gesamt- F -Statistik vergleicht das Modellmittel der Quadrate mit dem Residualmittel der Quadrate:

$$F = \frac{SS_{\text{model}}/p}{SS_{\text{residual}}/(n - p - 1)}$$

Die globale Nullhypothese lautet $H_0 : \beta_1 = \dots = \beta_p = 0$. Ein signifikantes Ergebnis besagt unter dem Modell, dass sich mindestens ein Populationskoeffizient ausserhalb des Achsenabschnitts von null unterscheidet. Es zeigt jedoch nicht, welcher Koeffizient dies ist. Für einen einzelnen Koeffizienten gilt:

$$t = \frac{b_j}{SE(b_j)}, \quad df = n - p - 1.$$

Dieser Test betrifft Koeffizient j unter der Bedingung der genau festgelegten anderen Modellterme. Sein Standardfehler ist nicht der Residualstandardfehler.

Bei zwei verschachtelten Modellen kann der zusätzliche Beitrag von q neuen Prädiktoren getestet werden, indem ihre Verringerung der Residualquadratsumme mit dem Residualmittel der Quadrate des grösseren Modells verglichen wird:

$$F = \frac{(SS_{\text{residual, reduced}} - SS_{\text{residual, full}})/q}{SS_{\text{residual, full}}/(n - p - 1)}$$

Das reduzierte Modell muss daraus hervorgehen, dass die hinzugefügten Koeffizienten des vollständigen Modells auf null gesetzt werden. Beide Modelle müssen dieselbe Zielvariable und dieselben analysierten Fälle verwenden. In dieser Formel bezeichnet p die Anzahl Prädiktorparameter ohne Achsenabschnitt im vollständigen Modell. Der Nenner verwendet daher die Residualfreiheitsgrade dieses vollständigen Modells. Bei einem hinzugefügten Prädiktor liefert die semipartielle Korrelation denselben Anpassungszuwachs:

$$sr_j^2 = R_{\text{larger}}^2 - R_{\text{smaller}}^2 = \Delta R^2.$$

Bei einer semipartiellen Korrelation wird nur der neu betrachtete Prädiktor residualisiert. Die partielle Korrelation aus Thema 6 residualisiert dagegen beide interessierenden Variablen.

Kategoriale Prädiktoren benötigen Indikatorvariablen. Bei einem Achsenabschnitt und k Kategorien werden $k - 1$ Indikatoren verwendet. Für einen quantitativen Prädiktor X und einen binären Indikator D lautet ein additives Modell:

$$\hat{Y} = b_0 + b_1X + b_2D.$$

Bei $D = 0$ lautet die angepasste Gerade $b_0 + b_1X$. Bei $D = 1$ lautet sie $(b_0 + b_2) + b_1X$. Die Geraden sind parallel, und b_2 ist die angepasste Gruppendifferenz beim selben Wert von X .

Eine Interaktion erlaubt unterschiedliche Steigungen:

$$\hat{Y} = b_0 + b_1X + b_2D + b_3XD.$$

Die Steigung der Referenzgruppe ist b_1 , jene der Vergleichsgruppe ist $b_1 + b_3$, und b_3 ist die Differenz zwischen den Steigungen. Der Koeffizient b_2 ist die Gruppendifferenz bei $X = 0$. Das Zentrieren von X kann diesem Vergleich deshalb einen nützlicheren Bezugspunkt geben.

Das Akaike-Informationskriterium für den Vergleich von Kandidatenmodellen lautet:

$$AIC = -2\log(L) + 2k,$$

wobei L die angepasste Likelihood und k die Anzahl geschätzter Likelihood-Parameter bezeichnet. Ein kleineres AIC weist nur unter Modellen, die an dieselbe Zielvariable und dieselben Fälle angepasst wurden, auf ein besseres relatives Gleichgewicht zwischen Anpassung und Komplexität hin. Es beweist nicht, dass das ausgewählte Modell wahr oder kausal ist oder neue Daten genau vorhersagt.

Grösse	Frage	Wesentliche Begrenzung
R^2	Wie viel Stichprobenvariation der Zielvariable stellt dieses angepasste Modell dar?	Kann beim Hinzufügen von Termen zum selben OLS-Modell nicht abnehmen

Grösse	Frage	Wesentliche Begrenzung
Korrigiertes R^2	Wiegt die zusätzliche Anpassung in der Stichprobe die Berücksichtigung weiterer Parameter auf?	Ist keine Validierung an neuen Fällen
Verschachtelter F -Test	Verbessern die hinzugefügten Koeffizienten gemeinsam die Anpassung?	Erfordert tatsächlich verschachtelte Modelle und dieselben Fälle
AIC	Welches der angegebenen Kandidatenmodelle besitzt das beste relative Gleichgewicht zwischen Anpassung und Komplexität?	Besitzt keinen allgemeingültigen Grenzwert

Diese Formeln quantifizieren Stichprobenanpassung und Unsicherheit. Sie bestimmen nicht, welche Prädiktoren wissenschaftlich sinnvoll sind oder ob ein bedingter Koeffizient kausal interpretiert werden darf.

Die erklärende Abbildung lesen

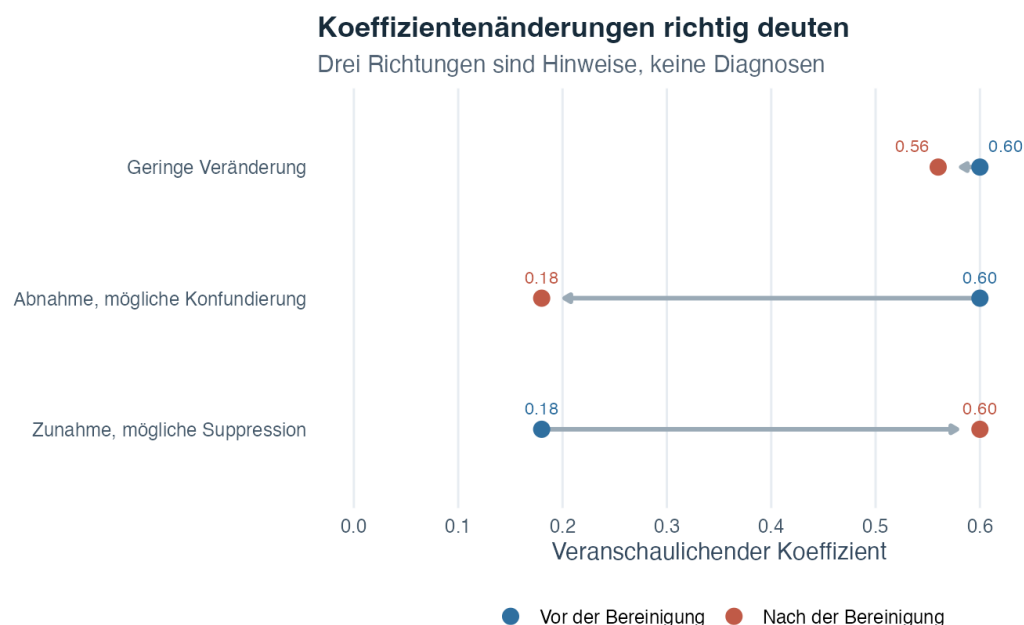


Abbildung 1: Drei horizontale Koeffizientenvergleiche zeigen geringe Veränderung, Verkleinerung und Vergrößerung zwischen Werten vor und nach der Bereinigung in einer multiplen Regression.

Jede Zeile vergleicht einen blauen Koeffizienten vor der Bereinigung mit einem orangefarbenen Koeffizienten, nachdem weitere Prädiktoren in das Modell aufgenommen wurden. In der oberen Zeile verändert sich 0.60 zu 0.56. Das bedingte Ergebnis ist dem unbereinigten Ergebnis ähnlich. Die hinzugefügten Prädiktoren haben diesen bestimmten Koeffizienten also wenig verändert. Dies beweist nicht, dass die Prädiktoren bedeutungslos sind. Sie können die Vorhersage verbessern oder für andere Koeffizienten wichtig sein.

In der mittleren Zeile verkleinert sich 0.60 auf 0.18. Der interessierende Prädiktor teilte erhebliche zielvariablenbezogene Information mit den hinzugefügten Variablen. Konfundierung ist eine mögliche inhaltliche Erklärung. Die Grafik kennzeichnet sie bewusst nur als möglich, weil die Bewegung eines Koeffizienten allein keine kausale Rolle bestimmen kann. Auch Messüberlappung, Auswahl, funktionale Form oder Stichprobenvariation können eine Rolle spielen.

In der unteren Zeile wächst 0.18 auf 0.60. Die Bereinigung hat einen stärkeren bedingten Zusammenhang sichtbar gemacht. Ein solches Muster wird oft als mögliche Suppression beschrieben. Auch hier ist die Bezeichnung ein Hinweis und keine Schlussfolgerung. Untersuche die Beziehungen zwischen den Prädiktoren, die Koeffizientenintervalle, das Design und die Modelldiagnosen. Der horizontale Abstand veranschaulicht die numerische Bewegung, zeigt aber keine Unsicherheit. Eine vollständige Analyse benötigt deshalb zusätzlich Konfidenzintervalle.

Checkliste zur Interpretation

Benenne Zielvariable, alle Prädiktoren, Einheiten, Codierung und Referenzkategorien. Erkläre, weshalb jeder Prädiktor enthalten ist und ob Interaktionen geplant waren. Untersuche Verteilungen, Prädiktorbeziehungen, fehlende Daten und die unterstützten Kombinationen. Übersetze jeden Koeffizienten als bedingte angepasste Differenz und benenne dabei, was konstant gehalten wird. Berichte bei Interaktionen bedingte Steigungen oder vorhergesagte Werte, statt einen einzelnen Term isoliert zu interpretieren.

Vergleiche verschachtelte Modelle nur, wenn sie dieselben Beobachtungen und dieselbe Zielvariable verwenden. Berichte Koeffizientenschätzungen und Intervalle, R^2 , korrigiertes R^2 , relevante Modellvergleiche und Diagnosen. Prüfe Residualform, veränderliche Varianz, Hebelwirkung, Einfluss und Prädiktorüberlappung. Halte Vorhersage, Zusammenhang und Kausalität auseinander. Wenn sich Koeffizienten zwischen Modellen verändern, beschreibe die Veränderung und untersuche ihre Quelle, statt ihr automatisch eine kausale Bezeichnung zuzuweisen.

Verbindung zu anderen Themen

Die multiple Regression vereint die vorangegangenen Werkzeuge für Zusammenhänge in einem Rahmen. Kovarianz und Korrelation führten die gemeinsame lineare Variation ein. Die einfache Regression verwandelte dieses Muster in eine gerichtete angepasste Gleichung mit Residuen. Die partielle Korrelation zeigte, dass «eine Variable konstant halten» durch die Residualisierung der interessierenden Variablen verstanden werden kann. Ein multipler Regressionskoeffizient wendet dieselbe bedingte Logik an, behält dabei die Einheit der Zielvariable und erlaubt die gemeinsame Beurteilung mehrerer Prädiktoren.

Die Varianzanalyse ist der nächste Ausdruck dieses Rahmens. Gruppenzugehörigkeit kann durch Indikatorprädiktoren dargestellt werden. Dadurch wird der Vergleich von Gruppenmittelwerten zu einem Regressionsmodell mit kategorialer Information. Der ANOVA- F -Test fragt, ob die Gruppenterme gemeinsam Variation der Zielvariable darstellen, die über die Residualvariation hinausgeht. Was zunächst wie ein getrenntes Verfahren erscheint, ist somit eine weitere Sicht auf dieselbe allgemeine Logik linearer Modelle.

Dokument-ID: topic-07-multiple-regression-summary-de

Notizen 1 von 3

Multiple Regression | Thema 7

Notizen 2 von 3

Multiple Regression | Thema 7

Notizen 3 von 3

Multiple Regression | Thema 7