Multiple Regression

Fortgeschrittene statistische Methoden

SS2020

Multiple Regression

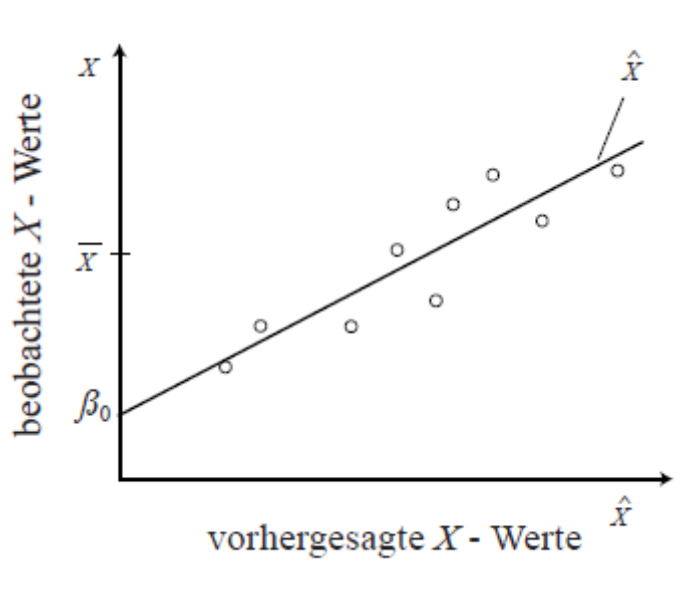
Ziele

- Schätzung eines Kriteriums aus einer Linearkombination von Prädiktoren
- Mit konkreten Meßvariablen
- Schätzung des Zusammenhanges in der Grundgesamtheit
- Information über die Wichtigkeit einzelner Einflußgrößen im Kontext von Multikollinearität
- Ermittlung suzessiver Kriteriumsaufklärung durch unabhängige Prädiktoranteile

Allgemeine Schätzgleichung

$$\widehat{X}_{0i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

hierin:	$\widehat{X}_{0i}$	Schätzung des Wertes der i-ten Person auf der Kriteriumsvariable X_0
	β_j	Regressionsgewicht des j-ten Prädiktors
	X_{ji}	Wert der i-ten Person auf der j-ten Prädiktorvariable



Vorgehen

1. Schätzung der Regressionsgewichte
2. Prüfung der Gleichung in der Grundgesamtheit (statistische Absicherung)
3. Interpretation der Koeffizienten, Rolle der Prädiktoren in der Vorhersagegleichung

Modellformulierung

Schätzfehler: $e_i = X_{0i} - \hat{X}_{0i}$

Optimierungskriterium für die Regressionsgewichte:

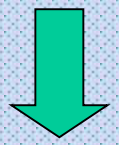
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(X_{0i} - \hat{X}_{0i} \right)^2 \longrightarrow \min$$

wobei

$$\hat{X}_{0i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

Modellformulierung, standardisiert

$$\widehat{z}_{0i} = b_1 z_{1i} + b_2 z_{2i} + \ldots + b_k z_{ki}$$



die additive Konstante fällt weg, nur Gewichte für Prädiktoren

Lösung: Normalgleichungen

Für

$$\widehat{z}_{0i} = b_1 z_{1i} + b_2 z_{2i} + \ldots + b_k z_{ki}$$

Multipliziere nacheinander mit jedem Prädiktor, summiere über Fälle und teile durch *n*, führt auf:

$$\begin{array}{ccccccccc}
b_1 & + & b_2 r_{12} & + & \cdots & + & b_k r_{1k} & = & r_{10} \\
b_1 r_{21} & + & b_2 & + & \cdots & + & b_k r_{2k} & = & r_{20} \\
b_1 r_{31} & + & b_2 r_{32} & + & \cdots & + & b_k r_{3k} & = & r_{3k} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
b_1 r_{k1} & + & b_2 r_{k2} & + & \cdots & + & b_k & = & r_{k0}
\end{array}$$

Normalgleichungen in Matrix Notation

System:

$$\mathbf{Rb} = \mathbf{r}_{k0}$$

wobei:

$\mathbf{R}$	$k \times k$ Matrix der Prädiktorinterkorrelationen
$\mathbf{b}$	gesuchter $k \times 1$ Vektor der Regressionskoeffizienten
$\mathbf{r}_{k0}$	$k \times 1$ Vektor der Kriteriumskorrelationen

Lösung: mit Inverser Prädiktorkorrelationsmatrix vormultiplizieren

$$\begin{aligned} \mathbf{R}^{-1}\mathbf{Rb} &= \mathbf{R}^{-1}\mathbf{r}_{k0} \\ \mathbf{b} &= \mathbf{R}^{-1}\mathbf{r}_{k0} \end{aligned}$$

Multipler Korrelationskoeffizient

$$R_{0\cdot 12\cdots k} = \sqrt{\sum_{j=1}^{k} b_j \cdot r_{j0}}$$

Bedeutung aus Varianzzerlegung:

$$\begin{array}{ccccc} \text{Gesamt-Streuung} & = & \text{Erklärte Streuung} & + & \text{Nicht Erklärte Streuung} \\ \frac{1}{n}\sum_{i=1}^{n}\left(X_{0i}-\overline{X}_{0i}\right)^2 & = & \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{X}_{0i}-\overline{X}_{0i}\right)^2 & + & \frac{1}{n}\sum_{i=1}^{n}\left(X_{0i}-\widehat{X}_{0i}\right)^2 \end{array}$$

$$R^2 = \frac{\text{Erklärte Streuung}}{\text{Gesamt-Streuung}}$$

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{X}_{0i}-\overline{X}_{0i}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_{0i}-\overline{X}_{0i}\right)^2}$$

1. Schätzung der Regressionsgewichte

Statistische Absicherung

1. Ist der Zusammenhang zwischen dem Kriterium und den Prädiktoren statistisch signifikant? (Läßt sich die gefundene multiple Korrelation mit der Annahme vereinbaren, daß die ’wahre’ multiple Korrelation zwischen dem Kriterium und den Prädiktoren gleich Null ist ?)
2. Welche Regressionskoeffizienten leisten einen statistisch bedeutsamen Beitrag zur Schätzung des Kriteriums (ist ihr „wahrer“ Regressionskoeffizient ungleich Null?

Statistische Absicherung

1. Muliple Korrelation mit F- Test absichern:

$$F = \frac{R^2\,(n-k-1)}{(1-R^2)\cdot k}$$

mit $df_1 = k$
$df_2 = n\text{-}k\text{-}1$

2. Regressionskoeffizienten mit t- Test prüfen:

$$t = \frac{b_j}{\sqrt{\frac{r^{jj}(1-R^2)}{n-k-1}}}$$

mit $df = n\text{-}k\text{-}1$

Hierin ist r^{jj} das jj-te Element der invertierten Korrelationsmatrix

Interpretation der Lösung

1. Unabhängige Prädiktoren

$$\begin{aligned} \mathbf{b} &= \mathbf{I}\mathbf{r}_{k0} \\ \mathbf{b} &= \mathbf{r}_{k0} \end{aligned}$$

Bei rein unabhängigen Prädiktoren sind die b- Gewichte identisch mit den Kriteriumskorrelationen

$$R^2_{0 \cdot 12 \cdots k} = \sum_{j=1}^{k} r^2_{j0}$$

Der multiple Determinationskoeffizient ist dann einfach die Summe der quadrierten Kriteriumskorrelationen (=erklärter Varianzanteil)

Interpretation: abhängige Prädiktoren

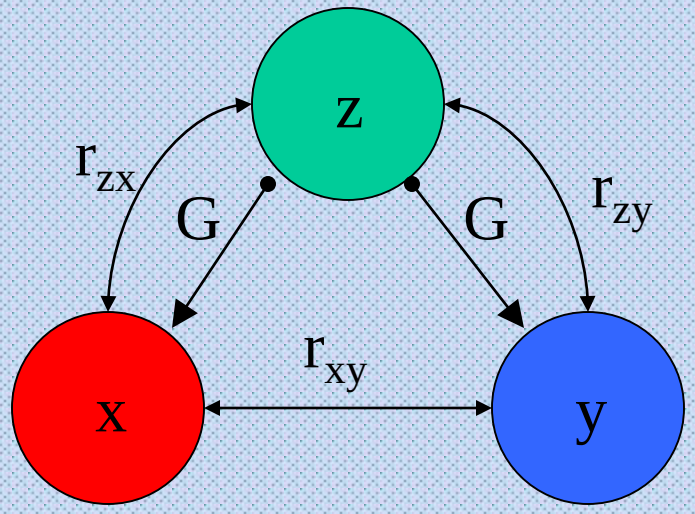
Untersuchung der Abhängigkeit im Kontext von Multikollinearität

1. Bedeutet die Abhängigkeit Redundanz, d.h. messen die vielen Variablen Aspekte gemeinsam, so daß man prinzipiell weniger (latente) Variablen benötigt? (unerwünschter Aspekt)
2. Erfassen die Abhängigkeiten Teile der *Kontamination* der Variablen und wirken so optimierend auf die gesamte Schätzgleichung (Suppressionseffekt, erwünscht)?

Partialkorrelation

Die Korrelation zweier Variablen, die vom Effekt anderer (spezifizierter) Variablen bereinigt wurden.

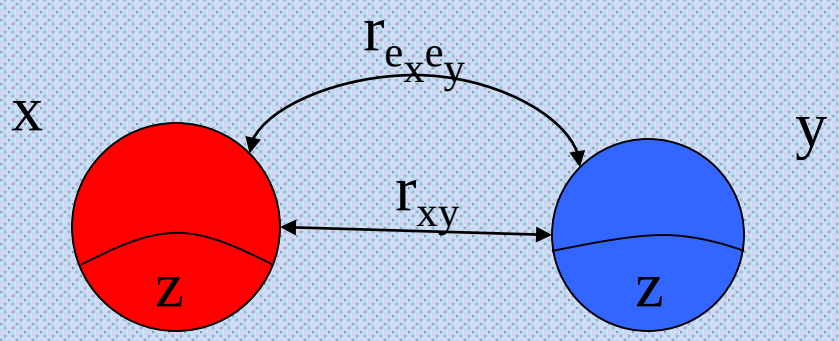
Prüfung einer Kausalvermutung: r_{xy} komme dadurch zustande, daß z ursächlich auf x und y einwirkt:



Partialkorrelation

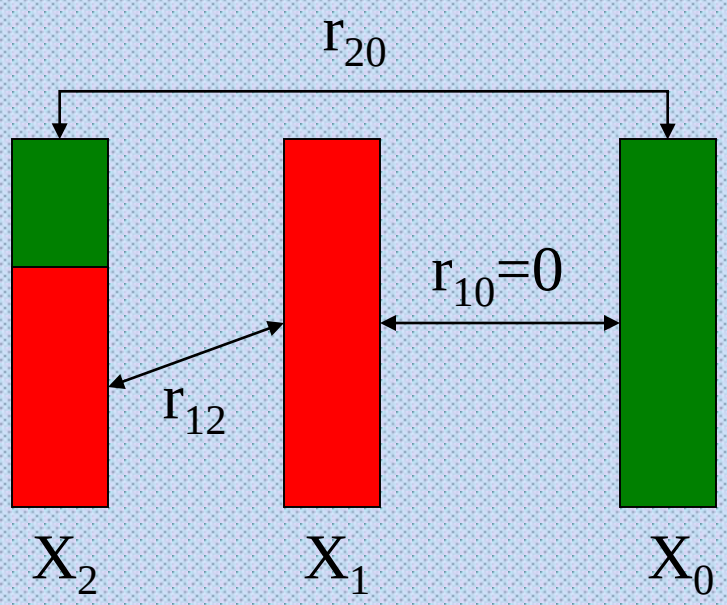
Prüfung:

1. Sage x aus z voraus und berechne Residuen e_x
2. Sage y aus z voraus und berechne Residuen e_y
3. Berechne die Korrelation $r_{e_x e_y}$



Ist Partialkorrelation Korrelation $r_{e_x e_y}$ Null, so beruht die Korrelation r_{xy} tatsächlich nur auf der Einwirkung von z.

Suppression



X_1 „bindet“ irrelevante Kriteriumsinformation

Partialkorrelation $r_{20.1}$ ist erheblich größer als r_{20}

Identifikation von Suppressionseffekten

Über die *Nützlichkeit:*

Variable X_j ist Suppressor, wenn gilt

$$U_j > r_{j0}^2$$

wobei

$$U_j = +\Delta R^2 \left(j \right)$$

U_j meint den Betrag, um den der multiple Determinations-Koeffizient zunimmt durch Aufnahme von X_j in die Gleichung

Semipartialkorrelation

Reihe von *Semipartialkorrelationen*:

$$R^2 = r^2_{01} + r^2_{0(2\cdot1)} + r^2_{0(3\cdot21)} + r^2_{0(4\cdot321)} + \dots$$

Die aufgeklärte Kriteriumsvarianz wird zerlegt in eine Reihe von *Semipartialkorrelationen.* Jeder neu hinzukommende Prädiktor bindet einen neuen Aspekt des Kriteriums, der nicht in den vorherigen Prädiktoren enthalten ist.

Quadrierte Semipartialkorrelation: Anteil des Prädiktors an der gesamten Kriteriumsvarianz:

$$r^2_{0(j\cdot12\cdots k)} = \frac{Varianz(X_{j\cdot12\cdots k})}{Gesamtvarianz}$$

Voraussetzungen

1. Normalverteilung der Residuen

2. Unkorreliertheit der Residuen

3. Varianzhomogenität der Fehlerkomponenten, die durch alle *k*-1 Prädiktoren im Kriterium eingeführt werden

Q-Q Plots der Residuen und Plot der Residuen gegen die vorhergesagten Werte zur Feststellung von Ausreißern