6.3 Nominale Einflussgrößen in Regressionsmodellen, Varianzanalyse

6.3.1 Dichotome Kovariablen

Bisher wurden $Y, X_1, X_2, \ldots, X_p$ als metrisch vorausgesetzt. Ähnlich wie für Korrelationskoeffizienten können dichotome Variablen, sofern sie mit 0 und 1 (wichtig!) kodiert sind, ebenfalls als Einflussgrößen zugelassen werden können.

Die zugehörigen Koeffizienten geben dann an, um wieviel sich Y – ceteris paribus – erhöht, wenn die entsprechende Kovariable den Wert 1 statt 0 hat.

Beispiel: Einfluss von Arbeitszeit und Geschlecht auf das Einkommen.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \\ \text{mit} \qquad X_1 &= \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich} \end{cases} \\ X_2 &= \text{(vertragliche) Arbeitszeit} \\ Y &= \text{Einkommen} \end{aligned}$$

Interpretation:

Würde man ansetzen $X_1 = \begin{cases} 1 & \text{weiblich} \\ 0 & \text{männlich} \end{cases}$,

so ergäben sich dieselben Schätzungen für β_0 und β_1, die Schätzung für β_2 wäre betragsmäßig gleich, aber mit umgekehrten Vorzeichen. (z.B. positiver Männereffekt $\Longleftrightarrow$ negativer Fraueneffekt)

6.3.2 Interaktionseffekte

Wechselwirkung zwischen Kovariablen lassen sich durch den Einbezug des Produkts als zusätzliche Kovariable modellieren

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3\, x_{1i} \cdot x_{2i} + \varepsilon_i$$

β_3 gibt den Interaktions- oder Wechselwirkungseffekt an. Dieser lässt sich insbesondere bei dichotomen Kovariablen einfach interpretieren:

Fortsetzung des Beispiels:

Die geschätzte Regressionsgerade hat bei den Männern die Form

$$\begin{aligned} \hat{y}_i &= \\ &= \end{aligned}$$

und bei den Frauen die Form

$$\begin{aligned} \hat{y}_i &= \\ &= \end{aligned}$$

6.3.3 Dummykodierung

Betrachten wir nun ein nominales Merkmal X mit q Kategorien, z.B. Parteipräferenz

Man beachte, dass man umbedingt $q-1$ und nicht q Dummyvariablen verwendet, da sonst die Schätzwerte völlig willkürlich und unsinnig werden.

$$X = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 2 & \text{SPD oder Grüne} \\ 3 & \text{Sonstige} \end{cases}$$

Man darf X nicht einfach mit Werten 1 bis 3 besetzen, da es sich um ein nominales Merkmal handelt.

Idee:

$$X_1 = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 0 & \text{andere} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{SPD oder Grüne} \\ 0 & \text{andere} \end{cases}$$

Beispiel zur Interpretation:

Y: Score auf Autoritarismusskala

X bzw. X_1, X_2: Parteienpräferenz

X_3: Einkommen

6.3.4 Varianzanalyse

Ist ein nominales Merkmal X mit insgesamt k verschiedenen Ausprägungen die einzige unabhängige Variable, so führt die Regressionsanalyse mit den entsprechenden $k-1$ Dummyvariablen auf die sogenannte (einfaktorielle) Varianzanalyse, die insbesondere in der Psychologie als Auswertungsmethode sehr verbreitet ist.

Als Schätzwert $\hat{y}_i$ ergibt sich für jede Einheit i genau der Mittelwert aller Werte y_i, die zu Einheiten l gehören, die dieselben Ausprägungen bei dem Merkmal X, also den zugehörigen Dummyvariablen $X_1, \ldots, X_{k-1}$, haben. Man bildet also k Gruppen bezüglich X, und $\hat{y}_i$ ist der Mittelwert der Gruppe, zu der i gehört.

Beispiel:

Y Autoritarismusscore

X Parteienpräferenz

X_1 CDU/CSU oder FDP, X_2 SPD oder Grüne, X_3 Sonstiges

Die Streuungszerlegung

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 \quad = \quad \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

der linearen Regression vereinfacht sich in diesem Fall und hat eine ganz charakteristische Form:

Indiziert man die Beobachtungen um und betrachtet die k Gruppen, so hat man in der j-ten Gruppe n_j Beobachtungen $y_{1j}, y_{2j}, \ldots, y_{n_j j}$ und den Gruppenmittelwert $\bar{y}_j$. Damit erhält man:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2=\sum_{j=1}^{k} n_j\cdot(\bar{y}_j-\bar{y})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2.$$

Dies ist genau die Streuungszerlegung aus Kapitel 3!

Das zugehörige Bestimmtheitsmaß wird üblicherweise mit η^2 bezeichnet:

$$\eta^2 = \frac{SQE}{SQT} = \frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}.$$

η^2 und $\eta = \sqrt{\eta^2}$ werden auch als Maße für den Zusammenhang zwischen einer metrischen Variable und einer nominalen Variable verwendet.

Mehr dazu am Ende von Statistik II...

6.4 Korrelation und „Kausalität“

Tücken bei der Interpretation von Zusammenhängen und Regressionsmodellen

Alle Zusammenhangsmaße messen ausschließlich die statistische Koinzidenz von Variablenwerten. Ob tatsächlich eine echte wirkende Beziehung vorliegt, kann – bestenfalls – aufgrund substanzwissenschaftlicher Überlegungen entschieden werden. In einem strengen Sinn bedürfen Kausalaussagen ohnehin eines experimentellen Designs.

- erstes (kleineres) Problem:
  Viele Zusammenhangsmaße sind symmetrisch, Kausalität ist eine gerichtete Beziehung.

- zweites, sehr schwerwiegendes Problem:
  Die falsche Beurteilung von Zusammenhängen entsteht insbesondere dadurch, dass entscheidende Variablen nicht in die Analyse miteinbezogen werden.

klassisches (fiktives) Beispiel:
Erhebung aus den 60er Jahren von Gemeinden:

X Anzahl der Störche

Y Anzahl der neugeborenen Kinder

X und Y sind hochkorreliert.
$\Rightarrow$ Störche bringen die Kinder...?

Weiteres Beispiel aus Gemeindestudie:

X Alter

Y Anzahl ausgeliehener Bücher in Bibliothek

X und Y stark negativ korreliert
$\Rightarrow$ Angebot für alte Gemeindemitglieder schlecht?

Rechnerischer Ausweg