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. Erinnerung: $x_i,\ i = 1, \cdots, n$ binär, $\bar{x}$ sinnvoll interpretierbar als relative Häufigkeit der Einser Ist Y binär, werden andere Modelle benötigt: Logit- bzw. Probitregression.
Die Logit-Regression lässt sich interpretieren als lineare Regression der erwarteten Odds von Y bei der jeweiligen Ausprägung der Kovariablen.
Odds$(x_{i1}, x_{i2}, \cdots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ mit Odds $(x_{i1}, x_{i2}, \cdots, x_{ip})$ als erwarteten Anteil der Einheiten mit $Y = 1$ zu denen mit $Y = 0$ unter alleen Einheiten mit Kovariablenwerten $x_{i1}, \cdots, x_{ip}$. Siehe für eine kurze Einführung z.B. Fahrmeir et al. (2012^7, Kapitel 12.3)

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

Bsp. 6.23. *Einfluss von Arbeitszeit und Geschlecht*

auf das Einkommen.

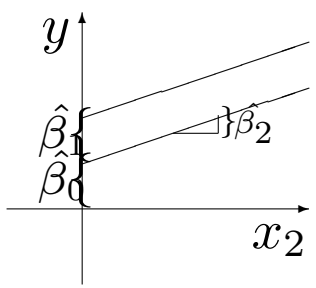
$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \\
\text{mit} \qquad X_1 &= \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich} \end{cases} \\
X_2 &= \text{(vertragliche) Arbeitszeit} \\
Y &= \text{Einkommen}
\end{aligned}
$$

Interpretation: Gleichung genau anschauen und überlegen. Die geschätzte Gerade für die Männer lautet:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot x_{2i}.$$

Für die Frauen hingegen gilt

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot x_{2i} \\ &= \hat{\beta}_0 + \hat{\beta}_2 \cdot x_{2i} \end{aligned}$$



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

so ergäben sich dieselben Schätzungen für $\hat{\beta}_0$ und $\hat{\beta}_2$, die Schätzung für $\hat{\beta}_1$ wäre betragsmäßig gleich, aber mit umgekehrten Vorzeichen. (also: positiver Männereffekt $\Longleftrightarrow$ negativer Fraueneffekt)

Natürlich würde man in der Praxis auch noch weitere Variablen mit einführen, z.B. die Berufserfahrung (metrisch) oder die formale Bildung (hier als binär vorstellen: Abitur ja/nein). Berücksichtigt man nur die formale Bildung, dann erhält man vier Geraden.

Dieser Modellierungsansatz bedeutet, dass sich Männer und Frauen nur im Einstiegsgehalt unterscheiden. Die angenommene gleiche Steigung impliziert gleichen Stundenlohn. Will man zudem eine vom Geschlecht abhängige Steigung zulassen, so ist dies wie folgt möglich:

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:

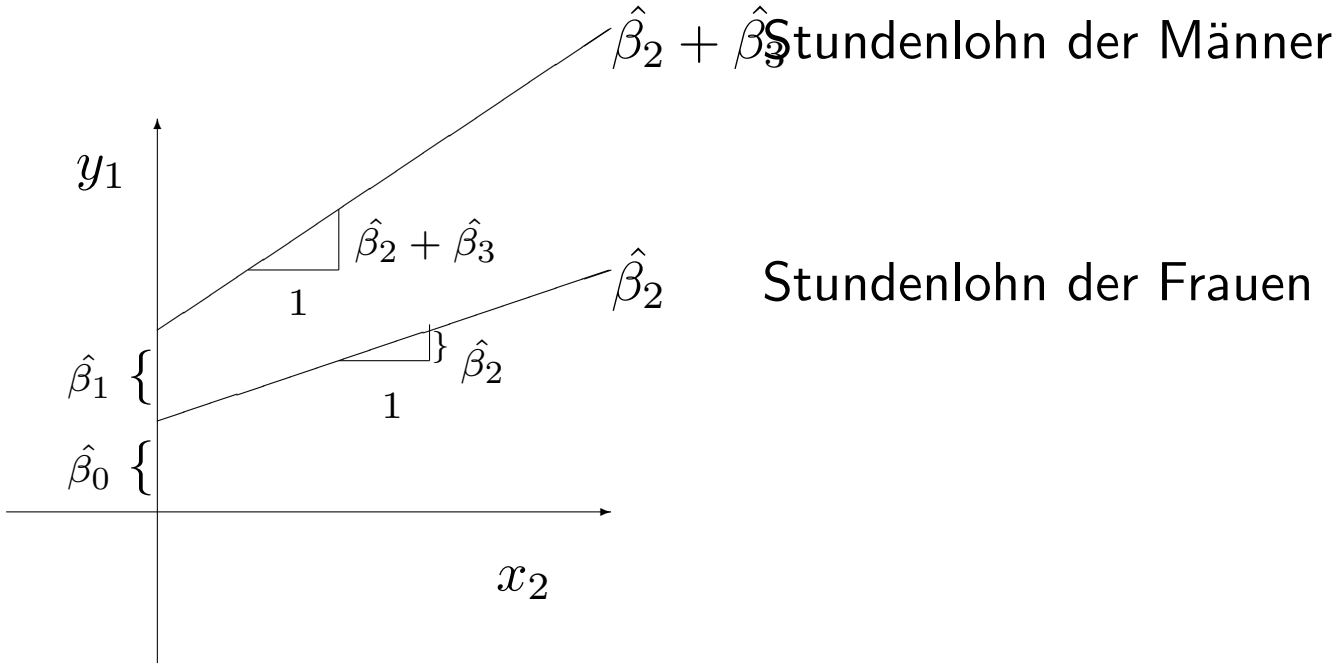
Bsp. 6.24. *Fortsetzung des Beispiels*

Die geschätzte Regressionsgerade hat bei den Männern die Form

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot x_{2i} + \hat{\beta}_3 \cdot 1 \cdot x_{2i} \\ &= \hat{\beta}_0 + \hat{\beta}_1 + (\hat{\beta}_2 + \hat{\beta}_3) \cdot x_{2i} \end{aligned}$$

und bei den Frauen die Form

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot x_{2i} + \hat{\beta}_3 \cdot 0 \cdot x_{2i} \\ &= \hat{\beta}_0 + \hat{\beta}_2 \cdot x_{2i}. \end{aligned}$$



$\hat{\beta}_1$ ist der Unterschied im Grundlevel, $\hat{\beta}_3$ der Unterschied in der Steigung.

6.3.3 Dummykodierung

Betrachten wir nun ein nominales Merkmal X mit $q > 2$ Kategorien in einer multiplen Regression[8], z.B. Parteipräferenz

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

Man darf X nicht einfach mit Werten 1 bis 3 besetzen, da es sich um ein nominales Merkmal handelt. Man könnte ja die Info äquivalent durch die Ausprägungen $-10, 0, 10$ darstellen, würde aber komplett andere $\hat{\beta}$'s erhalten

Idee: Mache aus der einen Variable mit k (hier 3) Ausprägungen $k-1$ (hier 2) Variablen

[8] Ist X die einzige Kovariable, so führt dies auf die Varianzanalyse, siehe Kapitel 6.3.4

mit den Ausprägungen ja/nein ($\hat{=}0/1$). Diese Dummyvariablen dürfen dann in der Regression verwendet werden.

$$X_1 = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 0 & \text{andere (SPD, Grüne oder Sonstige)} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{SPD oder Grüne} \\ 0 & \text{andere (CDU/CSU oder Sonstige)} \end{cases}$$

Beachte, durch die Ausprägungen von X_1 und X_2 sind alle möglichen Ausprägungen von X vollständig beschrieben:

X	Ausprägung von X	X_1	X_2
1	CDU/CSU, FDP	1	0
2	SPD, Grüne	0	1
3	Sonstige	0	0

Bsp. 6.25.

Beispiel zur Interpretation:

Y: Score auf Autoritarismusskala

X bzw. X_1, X_2: Parteienpräferenz

X_3: Einkommen

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

β_0: Grundniveau

β_1: ceteris paribus Effekt (Erhöhung des Grundniveaus) von CDU/CSU oder FDP

β_2: ceteris paribus Effekt (Erhöhung des Grundniveaus) von SPD oder Grünen

β_3: ceteris paribus Effekt des Einkommens

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

Unsinn. Das Beste was einem noch passieren kann, ist, dass das Programm „abstürzt“. (Die entsprechenden Parameterwerte sind „nicht identifiziert“, man muss einen Effekt von z.B. 0.5 additiv auf zwei $\beta' s$ verteilen, dazu gibt es unendlich viele Möglichkeiten, die $\beta' s$ sind also völlig sinnlos. Im Prinzip teilt man durch 0 (Inversion einer singulären Matrix). Im Computer steht aber meist nicht exakt 0, so dass irgendeinen Wert herauskommt.)
Will man Interaktionseffekte einführen, so kann man wie vorne vorgehen.