Lineares Modell: Was ändert sich durch Transformation?

Lineare Transf. von Zielvariable und Einflussvariable im einfachen Modell

$x_i \to t_i = a_0 + a_1 x_i \ (a_1 \neq 0) \qquad y_i \to u_i = b_0 + b_1 y_i \ (b_1 \neq 0)$

Vorsicht: $\frac{1}{x_i}$, $\log(x_i)$, x_i^2 etc. sind alles KEINE linearen Transf.

KQ-Schätzer für (linear) transformiertes Modell $\underline{u_i = \delta_0 + \delta_1 t_i + \varepsilon_i}$

$$\hat{\delta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(t_i-\bar{t})(u_i-\bar{u})}{\frac{1}{n}\sum_{i=1}^{n}(t_i-\bar{t})^2} = \ldots \quad t_i-\bar{t} = \cancel{a_0}+a_1x_i-\cancel{a_0}-a_1\bar{x} = a_1(x_i-\bar{x}), \quad u_i-\bar{u} = \cancel{b_0}+b_1y_i-\cancel{b_0}-b_1\bar{y} = b_1(y_i-\bar{y})$$

& Faktoren rausziehen

$$= \frac{a_1 b_1}{a_1^2} \cdot \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \underline{\frac{b_1}{a_1}\hat{\beta}_1}$$

$$\hat{\delta}_0 = \bar{u} - \hat{\delta}_1\bar{t} = b_0 + b_1\bar{y} - \frac{b_1}{a_1}\hat{\beta}_1(a_0+a_1\bar{x}) = b_0+b_1\bar{y}-b_1\hat{\beta}_1\bar{x} - \frac{b_1}{a_1}\hat{\beta}_1 a_0 = b_0 + b_1\underbrace{(\bar{y}-\hat{\beta}_1\bar{x})}_{\hat{\beta}_0} - \frac{b_1}{a_1}a_0\hat{\beta}_1$$

$$= \underline{b_0 + b_1\hat{\beta}_0 - \frac{b_1}{a_1}a_0\hat{\beta}_1}$$

- Parameter-Schätzer und deren absoluter Standardfehler ändert sich

  Deren Konfidenz (t-Wert) bleibt hingegen identisch zumindest für entsch. $\hat{\beta}_1$ Konf-Intervall nur absolute Unterschiede, in Relationen gleichbleibend

- R^2 ändert sich logischerweise nicht

Beispiel] i) Zentrierung von X $(a_0 = -\bar{x}, a_1 = 1)$, Y bleibt gleich $(b_0 = 0, b_1 = 1)$: $\hat{\delta}_1 = \hat{\beta}_1$
$x_i^c = x_i - \bar{x} \Rightarrow \bar{x^c} = 0$ $\hat{\delta}_0 = \hat{\beta}_0 \overset{(--)}{+} \bar{x}\hat{\beta}_1$

ii) Standardisierung, also Zentrierung u. Teilen durch Standardabweichung, von beiden Variablen sodass ihre MW = 0 und SD = 1

$$\tilde{x}_i = \frac{x_i - \bar{x}}{s_x - \sqrt{Var(x)}} \ (s_x^2) \text{ mit } Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad \text{analog } \tilde{y}_i = \frac{y_i - \bar{y}}{\sqrt{Var(y)}}$$

Standardisierung ist keine lineare Transformation mehr

Teilen durch Konstant, also $Var(\ldots)$, die vom gesamten Datensatz abhängt

$$\tilde{\beta}_1 = \hat{corr}(x,y)\frac{s_y}{s_x} \overset{s_x,s_y=1}{=} \hat{corr}(x,y), \quad \tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1\bar{x} \overset{\bar{x}=\bar{y}=0}{=} 0$$

Wozu (non-)lineare Transformationen von Regressoren und/oder Zielvariable?

Extreme Verteilungen (z.B. weite Range mit vielen Lücken, irre fette tails, super schief) eignen sich nicht für lineare Regression

$\Rightarrow$ Transformation. Klassisch dämpfender Effekt von $\log$, $\sqrt{\ }$ aber auch $X/10000$ kann sinnvoll sein

$\log 10$ besonders beliebt wegen einfacher Interpretation, $\log 10(x+1)$ bei linksseitigen Variablen, die Wert 0 enthalten

Vorsicht: Interpretation der Parameter von transformiertem Modell ändert sich!

Beispiel Y und X mit $\log 10$ transformiert: $\log 10(y_i) = \beta_0 + \beta_1 \log 10(x_i) + \varepsilon_i$

$\beta_0 \triangleq$ erwartete Logarithmierte Y wenn $X = 0$, nicht sinnvoll interpretierbar

$\hat{\beta}_1 \triangleq$ steigt $\log X$ um eine Einheit, so erhöhen sich $\log Y$ ceteris paribus durchschnittlich um $\hat{\beta}_1$ bzw. steigt $\log X$ um eine Einheit erhöhen sich $EW(Y)$ c.p. um $10^{\hat{\beta}_1}$ $10^{\log 10(x)} = x$

bzw. "steigt" X um Faktor a, so steigt $EW(Y)$ c.p. um $a^{\hat{\beta}_1} = 10^{\hat{\beta}_1 \log 10(a)} = 10^{\log 10(a^{\hat{\beta}_1})}$

Siehe Übung 03: $\log 10(Popul_i) = \beta_0 + \beta_1 \log 10(Area_i)$, gegeben Konfidenzintervall $(\alpha = 0.01)$ für $\beta_1 \ldots$ grob $[0.56, 0.81]$

Frage] Besteht zum Signifikanzniveau $\alpha = 0.01$ ein linearer Zusammenhang zwischen Bevölkerungsdichte und Fläche, basierend auf geschätztem Log-Modell, also unlogarithmiert?

$$\log 10(Popul_i) = \beta_0 + \beta_1 \log 10(Area_i) \Leftrightarrow \log 10(Popul_i) = \beta_0 + \log 10(Area_i^{\beta_1}) \Leftrightarrow Popul_i = 10^{\beta_0 + \log_{10}(Area_i^{\beta_1})} = 10^{\beta_0} \cdot Area_i^{\beta_1}$$

Nun auf Bevölkerungsdichte umrechnen: $Bev.dichte_i = \frac{Popul_i}{Area_i} = \frac{10^{\beta_0} \cdot Area_i^{\beta_1}}{Area_i} = 10^{\beta_0} \cdot Area_i^{\beta_1 - 1} \Rightarrow$ Ja. Fläche hat sign. Einfluss auf Bev. dichte, wenn 1 nicht innerhalb des Konf. Intervalls von β_1 liegt

(bei $\beta_1 = 1$ ist $Area_i^{\beta_1 - 1} = 1$ also ohne Einfluss als Faktor)

Was man nicht durch Transformationen fixen kann

Kategoriale Variablen "trennen" Datensatz in schwer vergleichbare Teile, tauchen aber im Modell nicht auf

Fix i) Eigenes Modell für jede Kategorie bzw. ein Modell für häufigste Kategorie

Fix ii) Kategoriale Var. als Dummy ins Modell aufnehmen

Fix iii) Extreme Ausreißer streichen, wenn inhaltlich sinnvoll (z.B. exotische Länder)