Einfaches Lineares Regressionsmodell

Grob: y-Werte (mithilfe von X-Werten) bestmöglich im Mittel beschreiben
Interessierende Regr. koeffizient (linear) schätzt, wie sich der durchschnittliche Wert von y bei Δx verändert

Wollen Zusammenhang zwischen Zielgröße y und Einflussgrößen x modellieren.

Eine problematische Annahme: Wahres, richtiges Modell (bezüglich Struktur, also linear mit bestimmten Einflussvariablen) ist bereits bekannt.
Dies impliziert, dass wir datengenerierenden Prozess kennen! Selbst mit Fachwissen über den jeweiligen Bereich ist das niemals korrekt.
immer gut zwecks sinnvolle(r) Modellierung

Reality: All models are wrong, but some are useful (or at least more useful than other models)

Ziel: Zusammenhang beschreiben, Effekt (β-koeff.) schätzen und interpretieren, Prognostizieren / Prädiktion von neuen / zukünftigen Daten
Zusammenhänge erklären, Trends finden, Supervised Learning (Zusammenhänge zwischen Variablen aus Daten lernen)

Modell in Vektor-Schreibweise $i = 1, \dots, n$ Beobachtungen

Jede einzelne Beobachtung modelliert als ZV

$$\underbrace{y_i = \beta_0 + \beta_1}_{\text{unbekannte Parameter}} x_i + \underbrace{\varepsilon_i}_{\text{Zufallsfehler / Error-Term}}, \text{ mit } Var(\varepsilon_i) = \sigma^2 \text{ konst. unbekannter Par.}$$

ε_i bayesianisch: Alles, was wir nicht kennen

y: Zielvariable, dependant variable (default: numerisch stetige Variable)
x: numerische Einflussvariable, independant variable, regressor, covariates

Achtung! Nur Zielgröße $y_{(i)}$ sei Zufallsvariable. Indirekt steckt Zufall also nur in den Error-Termen ε_i
Willkürliche Annahme
also y_i aus Stichprobe, repräsentativ & zufällig (in Realität leider oft „convenient" samples aus verzerrtem Bruchteil d. Population)
$x_{(i)}$ ist demnach feste bekannte Einflussvariable, wird als deterministisch betrachtet [Messfehler damit auch ausgeschlossen]
In Realität kommen x-Daten jedoch meistens aus einer (Zufalls-) Stichprobe, ergo enthält auch $x_{(i)}$ eine Zufallskomponente.

Ausschließen bzw. weitgehend kontrollieren lässt sich Zufälligkeit von x-Werten durch kontrollierte Experimente. $x_{(i)}$ sind dann experimentell bis zu einem gewissen Grad festgelegt.
Randomisierung (zufällige Zuweisung von x-Werten innerhalb des Experiments, im Modell insgesamt aber nicht zufällig!) ist hier sehr mächtig.
↳ Kein Confounding bzw. Störfaktoren haben keine systematische Beziehung zu Einflussgröße, Störeffekte „mitteln sich weg"

Vorteil der Annahme fester x-Werte: Berechnung von OLS-Schätzern $\hat{\beta}$ wird mathematisch leichter
$\hat{\beta}$-Verteilung bedingt auf X (konditional auf X) ausdrückbar, ohne Verteilung von X kennen zu müssen
Konditionierung auf $X \Rightarrow$ Zufälligkeit von $\hat{\beta}$ beruht nur auf Zufälligkeit von ε
Nähme man Regressor(en) als zufällig an, müsste man ihre Verteilung mitmodellieren, $\hat{\beta}$-Verteilung wäre eine unbedingte. OLS-Schätzer wäre dennoch erwartungstreu und (asymptotisch) normalverteilt. β nicht als Effekt gegeben x, sondern als struktureller Parameter des gemeinsamen Modells für (x, y)
In vielen statistischen Modellen interessiert man sich für gesamte Unsicherheit. Gemeinsame Zufallsverteilung von (y, X) ist wesentlich realitätsnäher.

Linearität muss nur in Parametern gelten! Man kann auch komplexere x-Terme (Polynome, reciprocal $1/x$, $\log x$, $\sqrt{x}$ ...) mit linearen Parametern modellieren. Linearkombination der Kovariable(n)

Weitere Annahmen des einfachen Linearen Modells

$y = \underbrace{f(x)}_{\text{systemat. Komponente}} + \underbrace{\varepsilon}_{\text{Zufalls-komp.}} = \underbrace{E(y|X)}_{\beta_0 + \beta_1 x} + \varepsilon$, also $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (1) Mit Param. schätzern $\hat{\beta}_0, \hat{\beta}_1$: $\hat{=}$ konditionalen EW / MW von y gegeben X bestmöglich approx. bzw. vorhersagen

Beachte: $\mathbb{E}(\varepsilon) = 0$, aber $\varepsilon = 0$ kann man nicht annehmen!
unabh. identisch verteilte Beobachtungen (y_i, x_i) i.i.d., also auch Errorterme ε_i sind i.i.d. (2) hat nichts mit (3), (4) zu tun perse

$\mathbb{E}(\varepsilon_i) = 0$ (2)
↳ $E(\hat{\varepsilon}_i) = \frac{1}{n}\sum \hat{\varepsilon}_i = 0$ per KQ-Design mit β_0

$1/n \sum \varepsilon_i^2$, $E(\varepsilon_i) = 0$: $E(\varepsilon_i^2) = Var(\varepsilon_i) = \sigma^2$
$\mathbb{E}(\varepsilon) = 0$ in Vektor-Schreibweise

$Var(\varepsilon_i) = \sigma^2$ konstant (3)
Homoskedastizität / Varianz-Homogen.
(3) impliz. (4)

$\{\varepsilon_i \mid i = 1, \dots, n\}$ paarweise stochastisch unabhängige Fehler (4) „$\Rightarrow$" $cov(\varepsilon_i, \varepsilon_j) = 0$ $corr(\,) = 0$

Bei zufälligen Regressoren wäre es strenger:
$\mathbb{E}(\varepsilon_i | x_i) = 0$ trivial erfüllt bei festem X
Exogenität der Einflussvariable
impliziert $\varepsilon_i \perp x_i$ Unabh.
$Var(\varepsilon_i | x_i) = \sigma^2$ konst.
$\Rightarrow Var(y_i) = \sigma^2$ konst (im Kasten??) bezogen auf ZV y_i
$Var(y_i | X_i) \neq$ empirische $Var(y_i)$
$\neq$ berechnete Gesamt-Varianz $Var(\hat{y}_{(i)}) \in$ Varianz-Zerlegung $var(y) = var(\hat{y}) + var(\hat{\varepsilon})$ wobei $var(\hat{y})$ durch Modell, also X, erklärt wird

nach CLT auch $\sum_i \varepsilon_i \overset{\text{a.s.}}{\sim} N(0, \sigma^2 n)$ als Summe von i.i.d. sample-errors $\hat{\varepsilon}_i^{(n)}$
$\varepsilon_i \sim N(0, \sigma^2)$ normalverteilt (5) unwichtigste Annahme
Anders als oft in der Praxis angewendet, müssen y_i keineswegs normalv. sein!
(Und auch NV der Zufallsfehler ist zwar nützlich, aber nicht unbed. notwendig)

$\hat{\varepsilon} = y - \hat{y}$

$\hat{\varepsilon} = 0$, $\bar{\hat{\varepsilon}} = 0$ MW(Residuen) = Null(vektor) per KQ-Design!
auch gilt $\bar{\hat{y}} = \bar{y}$ demnach folgerichtig aber nur mit Intercept β_0!

Wie wichtig diese Annahme sind (insbes. (3) bis (5)) hängt stark davon ab, was wir mit dem linearen Modell bezwecken, und manches lässt sich auch korrigieren.

im multiplen linearen Modell Statt $y = \beta_0 + \beta_1 X + \gamma Z + u$ schätzen wir $y = \beta_0 + \beta_1 X + \varepsilon$

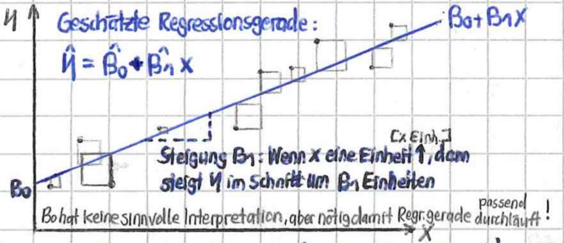
Confounding Variables Variablen, die nicht im Modell enthalten sind, mit einer oder mehreren X-Variablen zusammenhängen und zudem Einfluss auf Zielgröße y haben (nennen wir sie Z)

$\Rightarrow$ (Implizite bei festem x) Exogenitätsannahme $\mathbb{E}(\varepsilon_i | x_i) = 0$ ist verletzt, X korreliert mit Störfaktoren und damit folgt $cov(x_i, \varepsilon_i) \neq 0$ $corr(\,) \neq 0$
Teils des (den man aus X nicht perfekt vorhersagen kann ↑ Rest aus orthog. Projektion von Z auf X) $(\Rightarrow \varepsilon \not\perp X)$ $\mathbb{E}(\varepsilon | X) \neq 0$
→ Effekts der eigtl. relevanten Variable wird in den Errorterm geschoben, der systematisch wird ($E(\varepsilon_i) \neq 0$ für manche i, $Var(\varepsilon_i)$ evtl. nicht mehr konst.)

OMITTED VARIABLE BIAS

→ OLS-Schätzer $\hat{\beta}_p$ sind nicht erwartungstreu, sondern verzerrt $\mathbb{E}(\hat{\beta}_1) = \beta_1 + \gamma \cdot \frac{Cov(X, Z)}{Var(X)}$
Bias verschwindet nur, wenn $\gamma = 0$ (Z hat Null Effekt), $\boxed{Z \perp X} \Rightarrow Cov(X, Z) = 0$ $Var(\varepsilon | X) = Var(\gamma Z + u) = \gamma^2 var(Z) + Var(u)$
Ein Teil des Effekt von Z auf y wird durch eine / mehrere Regressoren (X) „aufgesaugt", $\hat{\beta}_1$ wird nach oben / unten verzerrt
Z geht in Errorterm ein, ist aber orthog. zu Regressoren aber $Var(\hat{\beta}_1)$ wird aufgebläht, also Ineffizienz (auch $R^2 \downarrow$)

Herleitung der (KQ) Schätzer im einfachen linearen Modell



B_0 hat keine sinnvolle Interpretation, aber nötig damit Regr.gerade passend durchläuft!

Intercept B_0: Erwartungswert von y bei $x=0$ (in vielen Fällen irrelevant)

OLS

Kleinste Quadrate bedeutet: Minimieren (Summe aller) quadratische Abweichung von Regressionsgeraden $B_0 + B_1 X$

Quadratische Abw. bestrafen hohe Abweichung & Ausreißer überproportional, sodass diese eine hohe Gewichtung im Minimierungsproblem bekommen (vs. least absolute deviations $|\ |$). Das ist heikel, aber mathematisch schöner u. weniger rechenintensiv.

σ: SD des Störterms ≙ durchschnittl. (grob) Abweichung der y-Werte von Modellgerade (genauer Wurzel d. quadr. Abw.)

$\sigma^2 = var(\varepsilon_i)$ $\neq Var(\varepsilon)$ gesamt, nicht sinnig

Ziel: Minimiere $\sum_{i=1}^{n}(\varepsilon_i)^2$, also minimale Summe für quadratische Abweichungen vom Vorhersagewert des Modells

$\left(\text{Pearson-Korrelation } r_{xy} = \frac{cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}\right)$

KQ-Schätzer $(\hat{B}_0, \hat{B}_1) = \arg\min_{B_0,B_1} \sum_{i=1}^{n} (y_i - B_0 - B_1 x_i)^2$

$\hat{\varepsilon}_i = \underbrace{(y_i - \hat{B}_0 - \hat{B}_1 x_i)}_{= y_i - \hat{y}}$ Residuen

$\hat{=} \widehat{corr}(x,y) \frac{S_y \cdot \sqrt{Var(y)}}{S_x \cdot \sqrt{Var(x)}}$

$g(B_0,B_1) = \sum_{i=1}^{n}(y_i - B_0 - B_1 x_i)^2$ Ableiten $\Longrightarrow$ LGS der zwei Normalgleichungen $\boxed{\begin{matrix}\sum \hat{\varepsilon}_i = 0 \\ \sum x_i \hat{\varepsilon}_i = 0\end{matrix}}$ immer V per KQ-Design! liefert $\hat{B}_1 = \frac{S_{xy}}{S_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = r_{xy}\frac{S_y}{S_x}$

KQ-Schätzer eindeutig, falls $\sum_{i=1}^{n}(x_i-\bar{x})^2 \neq 0$, also >0

$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$ ergibt Sinn

$\frac{dg(B_0,B_1)}{dB_0} = -2\underbrace{\sum_{i=1}^{n}(y_i - B_0 - B_1 x_i)}_{\sum \varepsilon_i = 0} \overset{!}{=} 0 \iff \sum_{i=1}^{n} y_i - nB_0 - B_1\sum_{i=1}^{n} x_i = 0 \iff nB_0 = n\bar{y} - n\bar{x}B_1 \iff \hat{B}_0 = \bar{y} - B_1\bar{x}$

$\frac{dg(B_0,B_1)}{dB_1} = -2\sum_{i=1}^{n} x_i(y_i - B_0 - B_1 x_i) \overset{!}{=} 0 \iff \sum_{i=1}^{n} x_i(y_i - \bar{y} + B_1\bar{x} - B_1 x_i) = 0 \ \ldots \iff B_1\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

$B_0 = (\bar{y} - B_1\bar{x})$ einsetzen

Reihen kann man nicht einfach kürzen

$\sum x_i\varepsilon_i = 0$

$\iff \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) - B_1\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0 \iff \hat{B}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \overset{\cdot 1/n}{=} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$\iff \sum_{i=1}^{n} x_i(y_i - \bar{y}) + B_1\sum_{i=1}^{n} x_i(\bar{x} - x_i) = 0$

$= \bar{x}\sum x_i - \sum x_i^2 = n\bar{x}^2 - \sum x_i^2 = -(\sum x_i^2 - n\bar{x}^2) = -\sum(x_i-\bar{x})^2$

Bonus $Var(\hat{B}_1) = Var(\ldots) = \frac{1}{(S_x^2)^2}\frac{1}{n^2} Var\left(\sum_{i=1}^{n} y_i(x_i-\bar{x})\right)$ (quadr. rausziehen)

x_i unabh. rausziehen aus Var(): $\frac{1}{n^2(S_x^2)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 Var(y_i)$

$= \frac{1}{n^2(S_x^2)^2} \cdot nS_x^2 \cdot \sigma^2 = \frac{\sigma^2}{nS_x^2}$

$\left(Var(y_i) = \sigma^2,\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = S_x^2\right)$

Betrachte: $S_x^2 = \sum(x_i-\bar{x})^2 = \sum x_i^2 - 2\sum x_i\bar{x} + \sum\bar{x}^2 = \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$

$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum x_i y_i - \sum x_i\bar{y} = \sum x_i y_i - n\bar{x}\bar{y} = \sum(x_i-\bar{x})(y_i-\bar{y})$

$= \sum x_i y_i - n\bar{x}\bar{y} = \sum y_i(x_i - \bar{x})$

Betrachte: $S_{xy} = \sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - \sum\bar{x}y_i - \sum x_i\bar{y} + \sum\bar{x}\bar{y} = \sum x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y}$

$\sum x_i y_i - 2n\bar{x}\bar{y} = \sum x_i y_i - \bar{x}\sum y_i = \sum x_i y_i - \bar{x} n \frac{1}{n}\sum y_i = \sum x_i y_i - n\bar{x}\bar{y}$

Beachte: $\sum_{i=1}^{n}(x_i - \bar{x}) = n\frac{1}{n}\sum x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$

Eigenschaften des KQ-Schätzers

Beachte für Beweise, dass x_i fest seien und wir nur ε_i als stochastische Komponente haben

- $(\hat{B}_0, \hat{B}_1)$ ist erwartungstreuer, also unverzerrter, Schätzer für (B_0, B_1): $\mathbb{E}(\hat{B}_0, \hat{B}_1) = (B_0, B_1)$

$E(\hat{B}_1) = E\left[\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n} E\left[(\underbrace{B_0 + B_1 x_i + \varepsilon_i}_{y_i} \underbrace{- B_0 - B_1\bar{x} - \bar{\varepsilon}}_{-\bar{y}})(x_i - \bar{x})\right] \overset{!}{=}$ (EW linear, nur ε stochastisch)

$\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}\left[(\cancel{B_0} + B_1 x_i + \underbrace{E(\varepsilon_i)}_{=0} - \cancel{B_0} - B_1\bar{x} - \underbrace{E(\bar{\varepsilon})}_{=0})(x_i - \bar{x})\right] = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n} B_1(x_i-\bar{x})^2 = B_1$

$E(\hat{B}_0) = E(\bar{y} - \hat{B}_1\bar{x}) = E(B_0 + B_1\bar{x} + \bar{\varepsilon}) - B_1\bar{x} = B_0$ ($E(\bar{\varepsilon}) = 0$)

- $Var(\hat{B}_1) = \frac{\sigma^2}{nS_x^2}$ (Erwartungswert) $\left(\text{int-Vert. } \frac{\hat{B}_1 - B_1}{\hat{SD}(\hat{B}_1)} \sim t_{n-2} \implies \hat{SD}(\hat{B}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

- $Var(\hat{B}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)$ (Erwartungswert)

beide unter Annahme $var(y_i) = \sigma^2$ $\quad var(y_i) = var(B_1 x_i + \varepsilon)$

vs. real mit zufälligen x_i: $Var(y_i) = B_1^2 var(x_i) + \sigma^2$

Mit determin. x_i anderer Zugang: $Var(y_i | x_i) = var(\varepsilon_i) = \sigma^2$

Exkurs: Wenn Modellannahmen (2) bis (5) korrekt sind, dann berechnet MLE-Methode selbe Schätzer für $(\hat{B}_0, \hat{B}_1)$

MLE "braucht" insbes. $\varepsilon_i \sim N(0, \sigma^2)$. Likelihood-Fkt. benötigt Verteilungsannahme → aus dieser Verteilung beziehen wir Likelihood

$\implies y_i \sim N(B_0 + B_1 x_i, \sigma^2)$ stamme aus Normalvert. mit Varianz σ^2

Likelihood von Beobachtungen (y_i, x_i): $L(y_1,\ldots,y_n, x_1,\ldots,x_n) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\varepsilon_i(B_0,B_1)^2}{2\sigma^2}\right)$, wobei $\varepsilon_i = y_i - B_0 - B_1 x_i$

Summe aus Produkt machen: $\log\left(\prod \sqrt{2\pi\sigma^2}^{-1}\right) = \sum -\log(2\pi\sigma^2)^{1/2}$ $\quad$ abh. von Parametern. $L(B_0, B_1, \sigma^2) = \ldots$

Ziehen Log-Likelihood $\log L(\ ) = -n\cdot\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2$

Da Parameter B_0 und B_1 nur in $\sum_{i=1}^{n}\varepsilon_i^2$ vorkommen, entspricht Maximierung von $\log L(y_i, x_i)$ der Minimierung von $\sum_{i=1}^{n}\varepsilon_i^2$ ≙ KQ-Methode $\implies$ selbe $(\hat{B}_0, \hat{B}_1)$

$\hat{\varepsilon}_i := \varepsilon_i(\hat{B}_0, \hat{B}_1)$ in Log-Likelihood eingesetzt → hängt nur noch ab von σ^2: $\underbrace{\log L(\sigma^2)}_{\ell(\sigma^2)} = \frac{-n}{2}\log(2\pi\sigma^2) - \underbrace{\frac{1}{2\sigma^2}}_{\frac{1}{2}(\sigma^2)^{-1}}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ $\quad \frac{1}{2}[\log(2\pi) + \log(\sigma^2)]$

$\frac{\partial\ell}{\partial\sigma^2} = -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2}\cdot\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \overset{!}{=} 0 \overset{\cdot 2\sigma^4}{\iff} -n\sigma^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = 0 \iff \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$

MLE-Schätzung von σ^2: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ nicht erwartungstreu, deshalb selten in Praxis verwendet bzw. Korrektur $\cdot\left(\frac{n}{n-p}\right)$ (und $\hat{\varepsilon}_i$ unterschiedlich $\forall i$) nicht nur von σ^2 (letztlich von X-Daten)

Berücksichtigt nicht, dass zuvor zwei Parameter geschätzt wurden (B_0, B_1), beanspruchen zwei FG

aber konsistent, also (approximativ) korrekt für große n

$\hat{\varepsilon}_i$ hängt von Modell/Parametern ab

Erwartungstreuer Schätzer von σ^2: $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ weil $\mathbb{E}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right) = (n-p)\sigma^2$ (geometr. hergeleitet (siehe multiple Regression))

So kann man nicht schätzen: $E(\hat{\varepsilon}_i) = 0$ aber nur unter MV.

$Var(\hat{\varepsilon}_i) = \mathbb{E}\left[(\hat{\varepsilon}_i - E(\hat{\varepsilon}_i))^2\right] = E(\hat{\varepsilon}_i^2)$