Herleitungen im multiplen linearen Modell

$y = X\beta + \varepsilon$

KQ-Schätzer: Minimiere $\boxed{\varepsilon^T\varepsilon}$ (Skalarprod. der Fehlervektoren $\triangleq \|y - X\beta\|^2$) $= (y-X\beta)^T(y-X\beta) = y^Ty - y^TX\beta - (X\beta)^Ty + (X\beta)^TX\beta = y^Ty - \underbrace{y^TX\beta}_{\text{Skalar}} - \underbrace{\beta^TX^Ty}_{\text{Skalar}} + \beta^TX^TX\beta$

$\underset{1\times n\;\; n\times(p+1)\;\;(p+1)\times 1}{} \quad 1\times(p+1)$

$\hat\beta = \arg\min_\beta (y - X\beta)^T(y-X\beta) \qquad \sum_{i=1}^n (y_i - x_i^T\beta)^2, \; \hat\varepsilon_i := (y_i - x_i^T\beta)^2$

$= y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta \qquad y^TX\beta = (y^TX\beta)^T = \beta^TX^Ty$

$\frac{\partial b^Ta}{\partial b} = a \qquad \frac{\partial b^TAb}{\partial b} = 2Ab$ für A symmetr.

Ableiten nach β: $\frac{\partial\, \varepsilon^T\varepsilon}{\partial \beta} = -2X^Ty + 2X^TX\beta \overset{!}{=} 0$

$\iff X^TX\beta = X^Ty \overset{(X^TX)^{-1}\text{ links}}{\iff} \hat\beta = (X^TX)^{-1}X^Ty \Rightarrow \hat y = X\hat\beta = \underbrace{X(X^TX)^{-1}X^T}_{\text{Hat Matrix } P} y = Py$

KQ-Schätzer erfüllt Normalgleichung(en) $X^T(y - X\hat\beta) = X^T\hat\varepsilon = 0$

$\hat\varepsilon = y - \hat y = \underbrace{(I-P)}_{\text{Residual Matrix } Q} y = Qy = (I-P)\varepsilon = Q\varepsilon$

$y = X\beta + \varepsilon, \; (I-P)(X\beta + \varepsilon)$ wobei $(I-P)X\beta = 0$ weil $P(X\beta) = X\beta$

KQ-Schätzer ist erwartungstreu: $\mathbb{E}(\hat\beta) = \beta$

$\mathbb{E}(\hat\beta) = \mathbb{E}((X^TX)^{-1}X^Ty) \overset{X \text{ nicht zufällig}}{=} (X^TX)^{-1}X^T\mathbb{E}(y) = (X^TX)^{-1}X^T\mathbb{E}(X\beta + \varepsilon) \overset{E(\varepsilon)=0,\, X \text{ ausz.}}{=} (X^TX)^{-1}X^T X\mathbb{E}(\beta) \overset{E(\beta)=\beta}{=} \underbrace{\cancel{(X^TX)^{-1}X^TX}}_{I}\beta = \beta$ □

Varianz des KQ-Schätzers: $\mathbb{V}(\hat\beta) = \mathbb{V}((X^TX)^{-1}X^Ty) = (X^TX)^{-1}X^T\,\mathrm{Var}(y)\,((X^TX)^{-1}X^T)^T = (X^TX)^{-1}X^T\,\mathbb{V}(y)\,X(X^TX)^{-1} \overset{\mathrm{Var}(y) = \mathrm{Var}(\varepsilon) = \sigma^2 I}{=} (X^TX)^{-1}X^T\cdot\sigma^2 I \cdot X(X^TX)^{-1}$

mehrere Parameter $\Rightarrow$ Kov. matrix $\qquad \mathrm{Var}(Ay) = A\,\mathrm{Var}(y)A^T \qquad X((X^TX)^{-1})^T = X((X^TX)^T)^{-1} = X(X^TX)^{-1}$

$= \sigma^2 I (X^TX)^{-1}\underbrace{\cancel{X^TX(X^TX)^{-1}}}_{I} = \sigma^2(X^TX)^{-1}$

Homoskedastizität & unkorrelierte Fehler

Wenn $\varepsilon \sim N(0, \sigma^2 I)$ erfüllt ist $\Longrightarrow \hat\beta$ ist normalverteilt als lineare Transf. einer normalverteilten ZV ε, $\hat\beta \sim N_p(\beta, \sigma^2(X^TX)^{-1})$

$\hat\varepsilon = (I-P)\varepsilon$

Beweis: $\hat\beta = (X^TX)^{-1}X^Ty \overset{y = X\beta+\varepsilon}{=} (X^TX)^{-1}X^T(X\beta + \varepsilon) = \underbrace{\cancel{(X^TX)^{-1}X^TX}}_{I_p}\beta + (X^TX)^{-1}X^T\varepsilon = \beta + (X^TX)^{-1}X^T\cdot\varepsilon$ $\hat\beta$ **abh. von ε aber unabh. von $\hat\varepsilon$**

also $\hat\beta$ ist affin-lineare Transformation des Fehlervektors ε

ε folgt multivariater NV $\Rightarrow \hat\beta$ als lineare Transf. dieser multivariaten NV ist selbst normalverteilt!

Es gilt: $\hat\beta \perp \hat\sigma^2$

σ^2 kann man auch als fixen unbekannten Parameter interpretieren

Erwartungstreue Schätzung von σ^2: $\hat\sigma^2 = \frac{1}{n-(p+1)}\hat\varepsilon^T\hat\varepsilon = \frac{1}{n-(p+1)}\sum_{i=1}^n \hat\varepsilon_i^2$ $\qquad$ erfordert $\varepsilon \sim N(0, \sigma^2 I)$ Verteilung: $\frac{\hat\sigma^2}{\sigma^2}(n-p') \sim \chi^2(n-p')$

erfordert alle Modellannahmen bis auf NV von ε $\qquad$ FG-Grade-Korrektur: Bereits $(p+1)$ Parameter geschätzt

Exkurs: Maximum Likelihood Unter Annahme $y \sim N(X\beta, \sigma^2)$ ist $f(y) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\underbrace{(y-X\beta)^T(y-X\beta)}_{\varepsilon^T\varepsilon}\right)$

$L(y_1,\dots,y_n,x_1,\dots,x_n) \Rightarrow \max L = (2\pi)^{-n/2}\hat\sigma^{-n}e^{-\frac{n}{2}\frac{\hat\sigma^2}{\hat\sigma^2}}$ $\qquad \sum\hat\varepsilon_i^2 = n\hat\sigma^2_{MLE}$

$\max L = (2\pi)^{-n/2}\hat\sigma^{-n}\exp(-n/2)$

$\Rightarrow$ Log-Likelihood $\operatorname{Log}L := \ell(\beta, \sigma^2) = -\frac{n}{2}\cdot\underbrace{\operatorname{Log}(2\pi\sigma^2)}_{\log(2\pi) + \log(\sigma^2)} - \frac{1}{2\sigma^2}\underbrace{(y-X\beta)^T(y-X\beta)}_{\varepsilon^T\varepsilon}$

Kontrolle auf Max: $\frac{\partial^2\ell}{\partial\beta\partial\beta^T} = -\frac{1}{\sigma^2}X^TX$ negativ definit ✓

(Halten σ^2 fest)

MLE für β: $\frac{\partial\ell}{\partial\beta} = -\frac{1}{2\sigma^2}\cdot 2(-X^T)(y-X\beta) = \frac{1}{\sigma^2}X^T(y-X\beta) \overset{!}{=} 0 \Rightarrow X^TX\beta = X^Ty \iff \underline{\hat\beta_{MLE}\,(X^TX)^{-1}X^Ty}$ identisch zu KQ-Schätzer

↓

Verzerrte MLE-Schätzung für σ^2: $\frac{\partial\ell}{\partial\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\cdot\frac{1}{(\sigma^2)^2}\varepsilon^T\varepsilon \overset{!}{=} 0 \overset{\cdot 2\sigma^4}{\Rightarrow} -n\sigma^2 + \varepsilon^T\varepsilon = 0 \iff \hat\sigma^2_{MLE}(\beta) = \frac{1}{n}(y-X\hat\beta)^T(y-X\hat\beta) = \frac{1}{n}\sum_{i=1}^n\hat\varepsilon_i^2$

Effektplots: Isolierung von individuellen Effekten einzelner Regressoren, siehe Ex-02_wdi

↳ geht auch für Prognose: Entsprechend $x_{n+1,j}$ Effekt auf $\hat y_{n+1}$ mit zugehörigen Konfidenzintervallen

a) Konditionale Darstellung: Meist Betrachtung von $\mathbb{E}(y_i \mid x_{i1}, \dots, x_{ij}, \dots, x_{ip})$ für verschiedene Werte von x_{ij} der interessierenden Kovariable j unter MW der anderen Einflussgrößen [allgemein konstant gehalten] $\rightarrow$ neue Zeilen in Designmatrix $\tilde x_i = (\bar x_1, \dots, x_{ij}, \dots, \bar x_p)$

Unsicherheit der Schätzung $\tilde y_i$ dann geg. durch $\tilde y_i \pm \hat\sigma_{\tilde y_i} t_{1-\alpha/2}(n-p')$ also $\tilde y_i$ ist t-verteilt

Steigung im Plot ist exakt β_k, andere Modellkomp(Intercept, weitere Kovariablen) verschieben nur "Offset"

Im multiplen Modell ohne non-lineare x-Terme u. ohne Interaktionen ist kond. Effektplot von Prädiktor im Grunde nur grafische Darstellung des Regr. koeff. in 2D

Bei Interaktionstermen und/oder Polynomen/Splines zeigt Effektplot mehr, und zwar modellierte (potenzielle komplexe) Beziehung zwischen isolierter x-Variable und y

b) Kontrast-Darstellung: Idee ↝ Wie ändert sich $\hat y_i$, wenn sich x_{ij} in Bezug zu Referenzwert der isolierten Kovariable x_j^* (meist $MW(x)$) ändert?

$\hat y_i - \hat y_i^* = \hat\beta_0 + \hat\beta_1 x_{i1} + \dots + \hat\beta_j x_{ij} + \dots + \hat\beta_p x_{ip} - (\hat\beta_0 + \hat\beta_1 x_{i1} + \dots + \hat\beta_j x_j^* + \dots + \hat\beta_p x_{ip}) = (x_{ij} - x_j^*)\hat\beta_j$

Unsicherheit der Schätzung ergibt sich durch $\mathrm{Var}(\hat y_i - \hat y_i^*) = \mathrm{Var}((x_{ij} - x_j^*)\hat\beta_j) = (x_{ij} - x_j^*)^2\,\mathrm{Var}(\hat\beta_j)$

Kontrastbasierter Effektplot ermöglicht isolierte Betrachtung e. einzelnen Kovariable x_j (berücksichtigt bei Unsicherheitsschätzung allerdings nur Schätzung v. $\hat\beta_j$)

Eng verwandt mit Analyse von bedingten Mittelwerten bei diskreten/gruppierten Einflussgrößen $\rightarrow$ geschätzter MW von Gruppe B bedingt auf MW von Referenzgruppe A

Im simplen linearen Modell sind beide Methoden nur mathematisch unterschiedliche Repräsentationen desselben linearen Zusammenhangs.

Konditionaler Effektplot ist besser, wenn wir untersuchen wollen, wie sich Outcome über Range von Prädiktor hinweg verändert.

Konfidenzintervall für Parameter β_k

Sattelfest formuliert: Konf. intervall überdeckt in $(1-\alpha)\%$ der Fälle den wahren Parameter
↳ macht deutlich, dass Intervall zufällig auf sample-Basis und Koeffizient fix!!

$\hat{\beta}$-Vektor ist normalverteilt $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2 (X^T X)^{-1})$ (multivariat (p))   Beachte, dass $\hat{\beta}$ und $\hat{\sigma}^2$ unabhängig sind

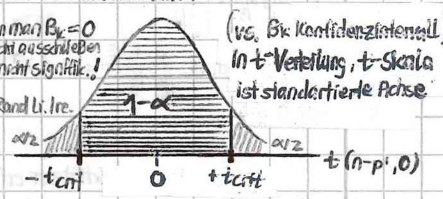
Einzelne Koeffizienten β_k jedoch t-Verteilt: $\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_k}}} \sim t(n-p', 0)$ $\hat{\sigma}^2_{\hat{\beta}_k} := (\hat{\Sigma}_{\hat{\beta}})_{kk}$ Diagonaleintrag in geschätzter Kovarianzmatrix von $\hat{\beta}$

hypothet. wdh. samples ⇒ Modell   Symmetrisch!

In $(1-\alpha)\%$ der Fälle liegt wahres β_k innerhalb von Konf. intervall um unseren Schätzer $\hat{\beta}_k$

Konfidenzintervall zum Niveau $1-\alpha$: $[\hat{\beta}_k - \hat{\sigma}_{\hat{\beta}_k} t_{1-\alpha/2}(n-p'); \hat{\beta}_k + \hat{\sigma}_{\hat{\beta}_k} t_{1-\alpha/2}(n-p')]$ (Quantil)

Wenn $0 \in$ intervall, so kann man $\beta_k = 0$ nicht ausschließen ↳ nicht signifik.!
"$t(\hat{\beta}_k)$" $\in$ Rand li. /re.



(vs. β_k Konfidenzintervall)
In t-Verteilung, t-Skala ist standardisierte Achse

Darauf basierend zweiseitiger t-Test auf $P(>|t|)$ siehe R-Output, für Signifikanz $P \leq \alpha$ gesucht
Testwert $\frac{\hat{\beta}_k - 0}{\hat{\sigma}_{\hat{\beta}_k}}$ (also $\beta_k = 0$) zum Niveau $1-\alpha$

⇒ Wenn t-Wert mit Wkeit $\leq \alpha$ außerhalb $|t_{crit}|$ liegt, so ist β_k bzw. Effekt von Kovariable signifikant zum Niveau $1-\alpha$ [typisch 95%]

Berechnung in R: Estimate / Std. Error, FG$(n-p')$ unten aufgeführt bei Residual standard error   Nullhypothese wird dann verworfen ⇒ $\beta_k \neq 0$ ✓

Interpretation: in $(1-\alpha)\%$ der Fälle liegt geschätzte Gerade / Geometrie innerhalb der Konfidenzbänder

Darüber hinaus gibt es auch Konfidenzintervall für wahre Regressionsgerade (bzw. wahre Modellgeometrie, aber nur in 2D-Modell grafisch sinnvoll zu interpret.)

Konfidenzbänder: Wollen EW einfangen $E(y_i|x_i) = \hat{y}_i = \beta_0 + \beta_1$ $[\hat{y}_i - \sqrt{Var(\hat{y}_i)}\, t_{1-\alpha/2}(n-p'); \hat{y}_i + \sqrt{Var(\hat{y}_i)}\, t_{1-\alpha/2}(n-p')]$ mit $V(\hat{y}_i) = V(\beta_0 + \beta_1 x_i)$

Berücksichtigen nur Streuung durch Parameter $\beta_0, \dots, \beta_p$ (deren Unsicherheit bzw. Varianz) $= \dots \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{n s_x^2}\right)$

2D: Ober/untere Grenze für $E(y_i)$ wird für alle Beob. i gebildet. Lokal symmetrisch um $\hat{y}_i$, aber nicht überall gleich breit. Trichterförmig
↳ typisch mehr Streuung an Rändern (Leverage Points, Entfernung von MW_x. ggf. inkl. $Var(\beta_0)$ richtig rein hier)

quasi Konfidenzintervall für neue Beobachtung, neue Zeilen $x_{n+1}^T \to y_{n+1}$ $x_{n+1}^T = (1\ x_{n+1,1} \dots x_{n+1,p})$

Prädiktionsintervalle zu Niveau $1-\alpha$: $[\hat{y}_{n+1} - \hat{\sigma}_{\hat{y}_{n+1}} \cdot t_{1-\alpha/2}(n-p'); \hat{y}_{n+1} + \hat{\sigma}_{\hat{y}_{n+1}} \cdot t_{1-\alpha/2}(n-p')]$

Es gilt $E(\hat{y}_{n+1} - y_{n+1}) = 0$ mit Prognosefehler (Unsicherheit von) $Var(\hat{y}_{n+1} - y_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = Var(\varepsilon_{n+1}) + Var(\hat{y}_{n+1})$ (muss man mit $\hat{\sigma}^2$ schätzen)

$E(\hat{y}_{n+1} - y_{n+1}) = x_{n+1}^T E(\hat{\beta}) - x_{n+1}^T \beta - \underbrace{E(\varepsilon_{n+1})}_{0} = 0$
$= E(x_{n+1}^T \hat{\beta} - x_{n+1}^T \beta - \varepsilon_{n+1})$

allgemein: $Var(\hat{y}_{n+1} - y_{n+1}) = Var(x_{n+1}^T \hat{\beta} - \underbrace{x_{n+1}^T \beta}_{=0\ \text{(determin.)}} - \varepsilon_{n+1})$ *
Kov. terme fallen weg, weil $\hat{\beta} \perp \varepsilon_{n+1}$!

2D, allgemein $\hat{\sigma}^2_{\hat{y}_{n+1}} = \hat{\sigma}^2(1 + x_{n+1}^T (X^T X)^{-1} x_{n+1})$
*) $Var(x_{n+1}^T \hat{\beta}) = x_{n+1}^T \hat{\sigma}^2 (X^T X)^{-1} x_{n+1}$ (x_{n+1}^T quadr. rausziehen)

- Deutlich breiter als Konfidenzbänder, weil zusätzlich Streuung von ε_{n+1} berücksichtigt wird
- nicht parallel zu Regr. gerade aber punktsymmetrisch um jeweilige Vorhersagewerte auf der Geraden

* diese Symmetrie fällt ganz weg, wenn wir statt verteilungsbasiert (mit t-Verteilung der Parameter) mit bootstrapping arbeiten
Verteilung schätzen mit Verteilung von Sample

Gauss-Markov-Theorem:

$y = X\beta + \varepsilon$ mit $rank(X) = p'$
$E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2 I$ also Homoskedastizität & unkorrelierte Fehler

⇒ Unter allen erwartungstreuen linearen Schätzern ist KQ-Schätzer derjenige mit der kleinsten Varianz
BLUE-Schätzer ≙ "best linear unbiased estimator"

Also: Für jeden anderen unverzerrten linearen Schätzer $\tilde{\beta}$ gilt $Var(\hat{\beta}) < Var(\tilde{\beta}) \Leftrightarrow Var(\tilde{\beta}) = Var(\hat{\beta}) + M$, M positiv semidefinit

Was machen wir, wenn Fehler nicht normalverteilt, sondern nur homoskedastisch mit endlicher Varianz $\sigma^2 I$?

Konsistenz des KQ-Schätzers Betrachten Modell mit steigendem Stichprobenumfang n   Achtung: Für Prognose(intervall) brauchen wir aber immer Verteilungsannahme für ε

X_n: Designmatrix der ersten n Beobachtungen, Folge X_n
$\hat{\beta}^{(n)}$: KQ-Schätzer aus diesen n-Beobachtungen   Voraussetzung: X_n hat vollen Rang für alle $n > p'$ sodass $X_n^T X_n$ immer invertierbar

ähnlich zu $\lim \frac{1}{n}(X^T X) = V$: $(X^T X)^{-1} = n^{-1}(\frac{1}{n} X^T X)^{-1}$, $\frac{1}{n} X^T X \to V$, V^{-1} ex. $(\frac{1}{n} X^T X)^{-1} \to V^{-1}$ und $\frac{1}{n}(\frac{1}{n} X^T X)^{-1} \to V^{-1}/n \xrightarrow{n \to \infty} 0$

$\lim_{n \to \infty} (X_n^T X_n)^{-1} = 0$ "information in den Daten wächst mit n"

(Schätzer wird besser / genauer mit n↑)   denn $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$, also Unsicherheit von $\hat{\beta}$ verschwindet mit n↑↑

⇒ Dann nähert sich $\hat{\beta}$ mit steigendem n wahrem β an   i) $\hat{\beta}^{(n)} \xrightarrow{P} \beta$ *asymptotisch folgt Erwartungstreue, mit $n \to \infty$ geht Bias gegen 0
auch für $\hat{\sigma}^2$ gilt dies $\hat{\sigma}^{2(n)} \longrightarrow \sigma^2$ ii) $\hat{\beta}^{(n)} \xrightarrow{f.s.} \beta$ falls zudem ε_i identisch verteilt

vs. Erwartungstreue: $\frac{1}{s}\sum_{m=1}^{s} \hat{\beta}^{(m)} \longrightarrow \beta$ Mittel vieler samples geht gegen wahres β (also jeweils 1x $\hat{\beta}$-Berechnung) ⇍ Konsistenz: $\lim_{n \to \infty} \hat{\beta}^{(n)} = \beta$ (⇐ asymptotisch ✓)
EINE Stichprobe konvergiert mit $n \to \infty$ gegen wahres β

Konsistenz ist relativ robuste Eigenschaft. Gegenbsp recht künstlich z.B. extrem schiefe Verteilung der Fehler / fat tails mit var( ) ∞
oder Endogenität $E[\varepsilon_i | x_i] \neq 0$, was aber meist Annahme (2)-(4) verletzt

Superrobust ist asymptotische Normalverteilung von KQ-Schätzer   Konsistenz von $\hat{\sigma}^2$ und $\hat{\beta}$ gilt weiterhin [Wenden CLT mit Kniff an für NV-Approx.]
"keine dauerhaften High Leverage Points, keine einzelne Beobachtung dominiert"   auch ohne $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$

Stärkere Voraussetzung: $\lim \frac{1}{n} X^T X = V$ mit V positiv definit   ⇒ $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} \mathcal{N}_p(0, \sigma^2 V^{-1})$
↳ $\hat{\beta} \overset{a}{\sim} \mathcal{N}_p(\beta, \hat{\sigma}^2 \cdot (X^T X)^{-1}/n)$

⟹ Inferenz möglich, wenn n groß genug   statt t-Verteil.
Konf. intervalle basieren dann aber auf approx. Standard-NV