(2) impliziert Zufalls-Stichprobe

1. $Y = X\beta + \varepsilon$ (Strukturannahme – ohne geht garnichts)
2. $E(\varepsilon_i) = 0$ 3. $Var(\varepsilon_i) = \sigma^2$ const. wobei KQ-Methode generell gar nicht auf parametrisches Modell angewiesen.
4. ε_i paarweise unabhängig
5. $\varepsilon_i \sim N(0, \sigma^2)$

d) Müssen alle aufgenommenen Kovariablen $x_1, \dots, x_j$ unkorreliert sein? Nein (aber es wäre ideal, wenn es so ist). Sie dürfen nur nicht perfekt oder näherungsweise perfekt linear abh. sein [cor( ) = +/− 1], denn dann ist X^TX nicht mehr (stabil) invertierbar.

a) Welche Annahmen 1.) - 5.) sind für die in den Zeilen angegebenen Schätzer, Eigenschaften bzw. Größen notwendig? **Multiples Regressionsmodell:** $Y_{gewicht,i} = \beta_0 + \beta_{größe,i} \cdot x_{größe,i} + \beta_{alter,i} \cdot x_{alter,i} + \varepsilon_i$

MLE weniger fehleranfällig, weil wir Annahmen (2) bis (5) gründlich vorher prüfen müssen

"wahrscheinlichste" Gerade	1.)	2.)	3.)	4.)	5.)
KQ-Schätzer 2D: Welche Gerade passt am besten zu Daten?	x	(2) bis (5) ✓ => $\hat\beta_{KQ} = \hat\beta_{MLE}$			
ML-Schätzer 2D: Welche Gerade passt am besten zu Daten UND Annahmen (2) bis (5)?	x	x	x	x	x
$E(\hat{\beta}) = \beta$	x	x			
$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$	x	x	x	x	(x)*
Konfidenintervalle (von Regressionskoeffizienten, Prädiktionen, ...)	x	x	x	x	(x)*
Prädiktionsintervalle (von Prädiktionen, ...)	x	x	x	x	x

Zusatz zu d) Korrelation der Prädiktor-Variablen untereinander
• Nicht so relevant, wenn wir nur Prädiktion wollen
Potentiell Problem, wenn wir erklären, also einzelne Koeff. interpretieren möchten! (ggf. muss man dann Variablen entfernen oder mehrere als Linearkomb. zu einer einzigen x-Variable kombinieren, wenn Korrelation zu stark, besonders ist der isolierte Effekt u. U. nicht sinnvoll interpretierbar)

Asymptotisch ist jeder MLE-Schätzer normalverteilt?! großes n kann falsche Normal-Verteilungsannahme (ε_i) kompensieren

*Diese Eigenschaften gelten auch für ausreichend großen Stichprobenumfang n approximativ. Für kleine Stichprobenumfänge ist die Normalverteilungsannahme jedoch zwingend notwendig!

↳ Verteilungsannahmen (ε_i) und Varianzannahmen (ε_i) können total falsch sein ⇝ struktureller Effekt (ggf. Modell) trotzdem unverzerrt schätzbar!

Hinweis: Besonders anzumerken ist, dass für den KQ-Schätzer nichts weiter benötigt wird als die Modellgleichung. Nimmt man zusätzlich noch an, dass der Erwartungswert der Residuen 0 ist, sind die Schätzer für β sogar erwartungstreu.

Um Konfidenz-Invervalle bilden zu können sind alle 5 Annahmen notwendig (oder 1.)-4.) und eine große Stichprobe). Für Prädiktions-Intervalle sind stets alle 5 Annahmen notwendig (insbesondere genügt hier auch eine ausreichend große Stichprobe NICHT).

Unverzerrte Trendschätzung (1-2 reichen) vs. Unverzerrte Trend-Extrapolation (auch 3-5 nötig)

b)
- Annahme 1.) $Y = X\beta + \epsilon$ (Linearität)

  Es wird angenommen, dass sich die Zielgröße (Körpergewicht) als eine Linearkombination aus den Einflussgrößen (Alter, Körpergröße) beschreiben lässt - also z.B. konkret dass sich das Gewicht von Grundschulkindern direkt nach der Formel in der Angabe berechnen lässt (und nicht z.B. $Gewicht = Alter^{\text{Größe}}$ oder sonst einer anderen Formel).

  Diese Annahme scheint als Laie beurteilt nicht vollkommen unplausibel und kann daher zunächst so angenommen werden.

  Hinweise: Diese Annahme ist vor allem von Wissenschaftlern aus dem jeweiligen Fachgebiet zu beurteilen. Es gibt jedoch Methoden der Modelldiagnostik (Residuen-vs.-Kovariablen-Plot), welche noch später im Kurs behandelt werden, mit der man diese Annahme im Nachhinein beurteilen kann. Auch gibt es die Möglichkeit Einflussvariablen nicht-linear ins Modell aufzunehmen (z.B. via Splines, welche ebenso später im Kurs behandelt werden).

- Annahme 2.) $E(\epsilon_i) = 0$ (Oft eine vernünftige Annahme)

  Die Residuen sind zufällige Fehler. Man nimmt an, dass diese Fehler im Erwartungswert 0 sind. Anders ausgedrückt: Die Modellgleichung $Y = X\beta$ spiegelt den Zusammen hang wieder. Es kommt keine systematische Störgröße ϵ mehr hinzu, von der man erwarten kann, dass sie im Mittel nicht 0 ist.

Exkurs zu Median (ε_i) $E(\varepsilon_i) = 0$ -> Per KQ-Design ist immer $MW(\hat{\varepsilon}_i) = 0$, da steckt keine Information drin

Schauen uns $Med(\hat{\varepsilon}_i)$ an, da kann man schiefe Verteilung der Fehler [ggf. $E(\varepsilon_i) \neq 0$] tatsächlich erkennen

Davon gehen wir aus.

(3) steckt inhärent in Likelihood drin

- Annahme 3.) $Var(\epsilon_i) = \sigma^2$: Homoskedastizität (Varianzhomogenität)

  Es wird angenommen, dass die Varianz für jede Beobachtung identisch ist.

  Diese Annahme könnte verletzt sein. Z.B. könnte bei größeren Kindern das Gewicht stärker streuen, als bei kleineren Kindern. Bei größeren Kindern könnte also eine höhere Varianz vorliegen.

  Hinweise: Auch dies können ggf. Wissenschaftler aus dem jeweiligen Fachbereich besser beurteilen.

  Prüfen kann man diese Annahme mittels später im Kurs vorgestellten Residuen-Plots (z.B. Residuen-vs.-Fitted values oder Residuen-vs.Kovariable-Plot).

  Die Varianz-Annahme kann auch aufgeweicht werden (z.B. via verallgemeinerte lineare Modelle, welche später im Kurs vorgestellt werden).

(4) nötig für ML: Bildung der Likelihood-Fkt $\prod_{i=1}^{n}$ .... erfordert Unabh. der Beobachtungen $|\varepsilon_i$ zum Auseinanderziehen

- Annahme 4.) ϵ_i paarweise unabhängig

  Es wird angenommen, dass die Beobachtungen der Zielgröße bei gegebenen Einflussgrößen unabhängig sind.

  Diese Annahme ist bei der Wahl einer einigen kleinen Grundschule unwahrscheinlich. Es könnten z.B. Kinder der selben Familie in die Schule gehen und vielleicht genetisch bedingt zu Fettleibigkeit neigen. Diese Beobachtungen wären also z.B. nicht unabhängig voneinander.

  Hinweise: Auch bei Verletzung dieser Annahme gibt es modellierungstechnische Möglichkeiten dies zu berücksichtigen. Z.B. mittels Gemischten Modellen (Mixed Models - z.B. für einen weiteren zufälligen Effekt für jede Familie - später in diesem Kurs) oder Korrelationen zwischen Beobachtungen (Generalized Estimation Equations (GEE) -> Vorlesung: Generalisisierte Regression).

(5) MLE braucht Verteilung mit konst. Varianz, aber es muss nicht unbedingt NV sein, denn jetzt Anwendung auf new data

- Annahme 5.) $\epsilon_i \sim N(0, \sigma^2)$ dasselbe gilt für Prädiktion. Man sollte sich um Verteilung der Fehler (egal welche) sehr sicher sein!

  Dass die übrige Reststreuung normalverteilt ist, ist durchaus denkbar. Störgrößen sind in der Natur häufig normalverteilt. Auch der Wertebereich der Zielgröße lässt nicht auf mögliche Verletzung der Normalverteilungsannahme hindeuten.

  Explizit aufpassen muss man, wenn z.B. offensichtlich ist, dass die Größe nicht normalverteilt sein kann. Dies ist z.B. der Fall wenn die Zielgröße binär ist, nur ganze Zahlen annehmen kann oder häufig Werte an den Grenzen des möglichen Wertebereichs hat.

  Hintergrund: Die Normalverteilung ist eine stetige, symmetrische Verteilung mit Werten von $-\infty$ bis ∞. In den gerade genannten Fällen können die Residuen (= tatsächlicher y-Wert - geschätzter y-Wert) nicht oder nur schwer einer Normalverteilung folgen.

  Überprüfen kann man die Normalverteilungsannahme z.B. mit dem später im Kurs vorgestelltem Q-Q-Plot.

c) Solange die Annahmen 1.) und 2.) erfüllt sind, können die Regressionsparameter problemlos interpretiert werden. Insbesondere für die Interpretation von Konfidenzintervallen, p-Werten etc. müssen aber entsprechende Annahmen erfüllt sein! (Es geht um die konkrete Bezifferung von Wahrscheinlichkeiten). Falls dies (teilweise) nicht der Fall ist, können diese wahrscheinlichkeitsbasierten Kennzahlen evtl. noch als Indikatioren für selbige verwendet werden, aber keinesfalls explizit mit Wahrscheinlichkeits-Interpretation!

e) Körpergröße und Alter sind korrelierte Einflussgrößen. Was folgt daraus für Interpretation der einzelnen Regr.koeffizienten, c.p.?

Interpretation zunächst normal wie üblich: z.B. "Wenn Alter um ein Jahr steigt, nimmt Gewicht im Mittel um 3kg zu"

Vorsicht! Modellschätzung besagt nicht, dass Kind pro Jahr im Schnitt 3kg zulegt. Kind wird auch wachsen u. insgesamt pro Jahr > 3kg an Gewicht zunehmen.

$\hat{\beta}_{Alter}$: Es geht nur um Effekt auf verschieden alte Kinder GLEICHER KÖRPERGRÖßE !! Kombination möglich] Wieviel Gewichtszunahme durchschn., wenn Kind halbes Jahr älter wird und 4 cm wächst? => $\frac{1}{2} \cdot \hat{\beta}_{Alter} + 4 \cdot \hat{\beta}_{Größe}$