Statistik II - Sitzung 9

Lena Masch

Institut für Politikwissenschaft

2. Dezember 2024

Statistik II - Sitzung 9

Wiederholung: Multivariate Regression

- Die multivariate Regression

$$y = \hat{y} + e = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ... + e$$

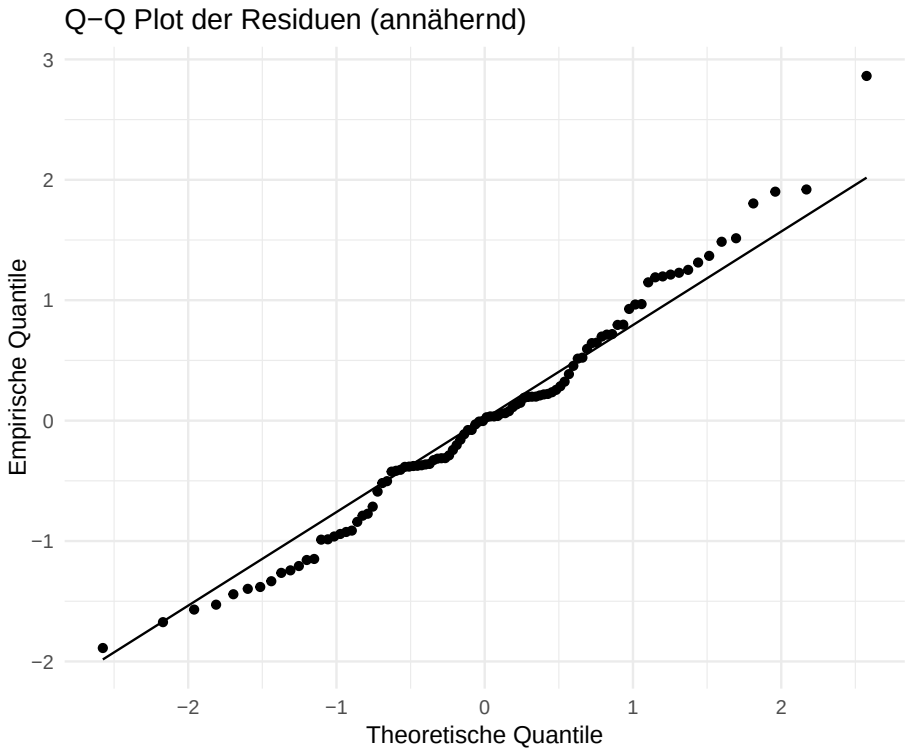
- die Regressionsgleichung kann genutzt werden, um Werte vorherzusagen $\hat{y}$.
- die Residuen im Fehlerterm (e) zeigen die Abweichung zwischen den vorhergesagten und den beobachteten Werten
- nach dem Schätzen der Regression können die Residuen visuell und anhand statistischer Tests inspiziert werden

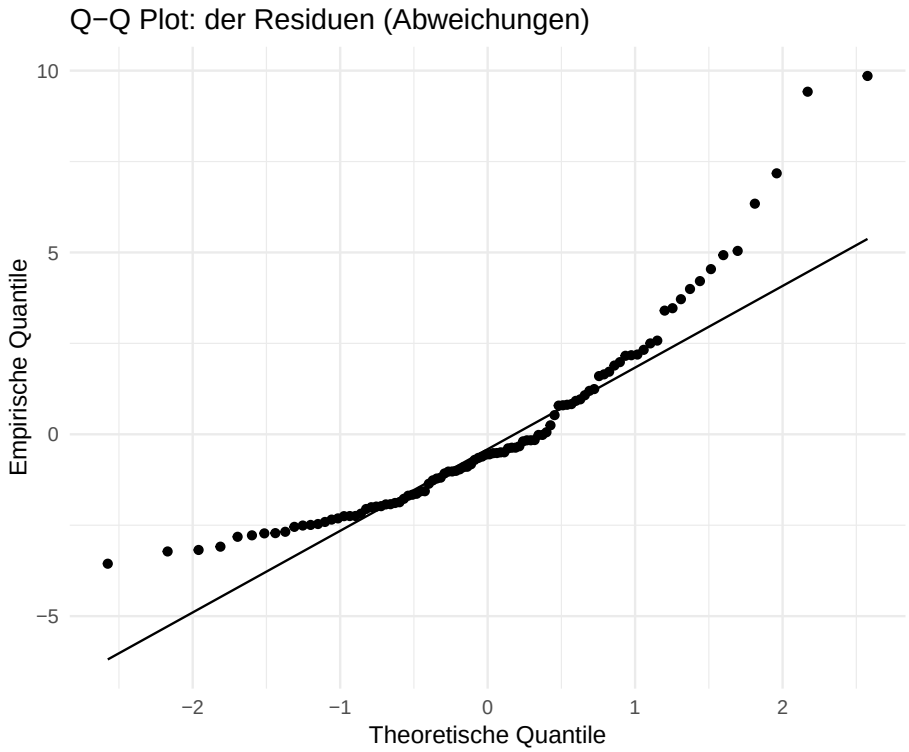
Wiederholung der Annahmen bzgl. Residuen

- Die Residuen zeigen, ob Annahmen der Regression verletzt werden, z.B. Linearität
- Die Residuen sollten in der Regel gewisse Eigenschaften benutzen (siehe Sitzung 8)
- Dazu gehören u.a.
  - Mittelwert der Residuen $= 0$
  - Normalverteilung
  - Homoskedastizität
- Die Residen werden i.d.R. mit Bezug auf die vorhergesagten Werte oder die unabhängige Variable dargestellt

Normalverteilung

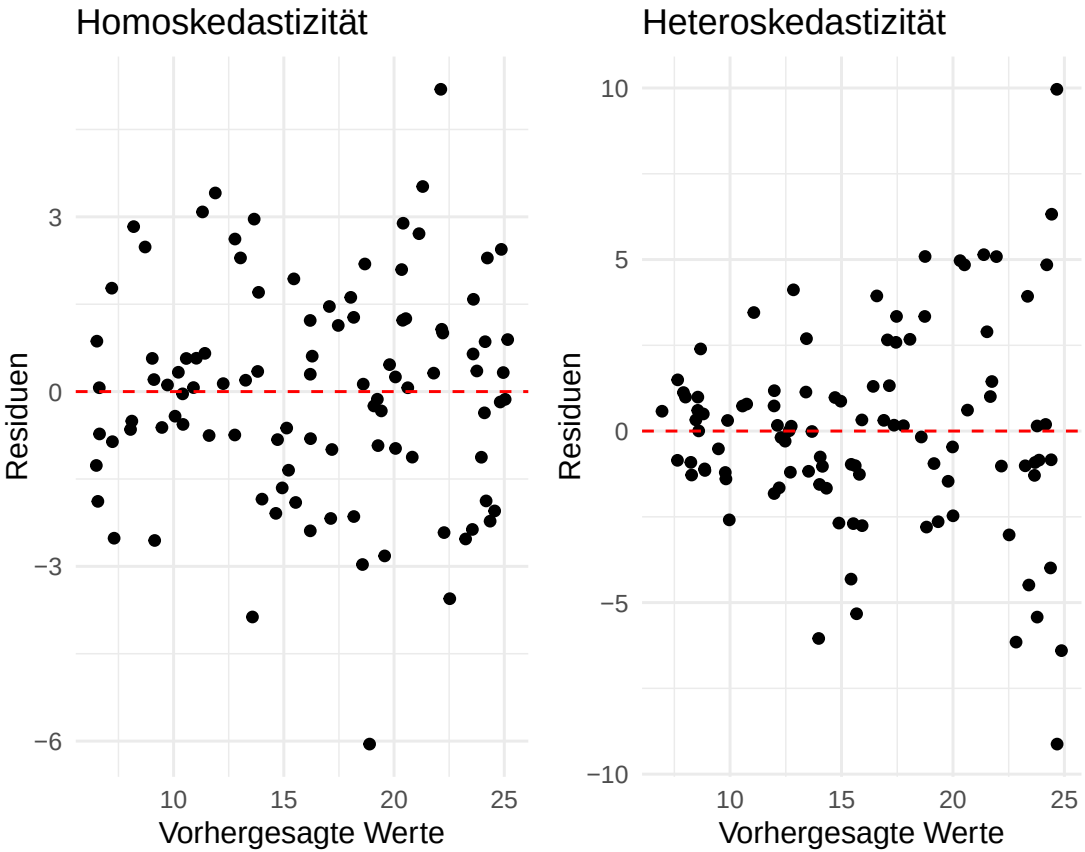
- Ein Q-Q-Plot gibt erste Hinweise auf die Normalverteilung
- Die Werte sollten auf oder nah an der Diagonale liegen (kleine Abweichungen)





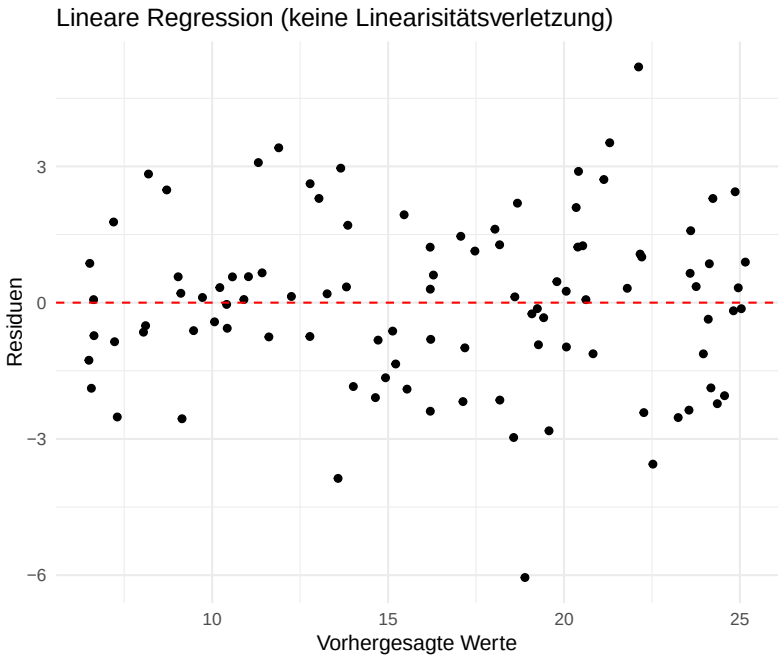
Homoskedastizität

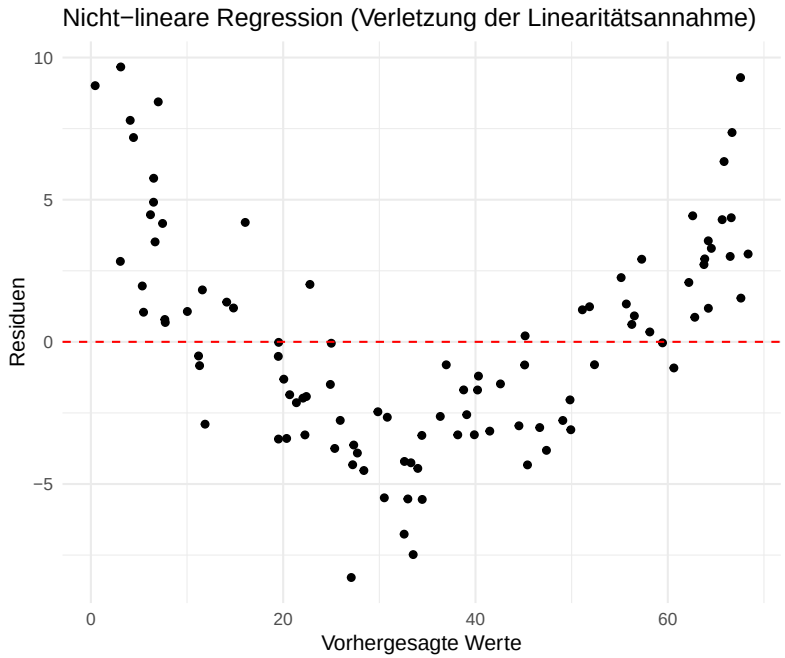
- Ein Streudiagramm der Residuen sollte keine Muster
- die Residuen sollten gleichermaßen um die Regressionsgerade streuen



Linearität

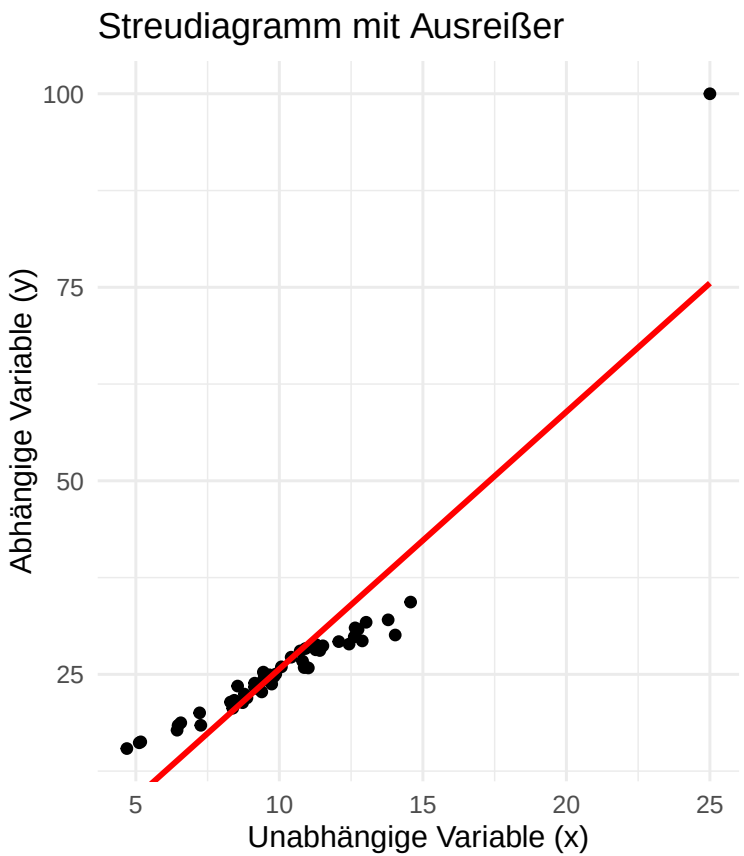
- Ein Streudiagramm der Residuen sollte keine Muster aufzeigen

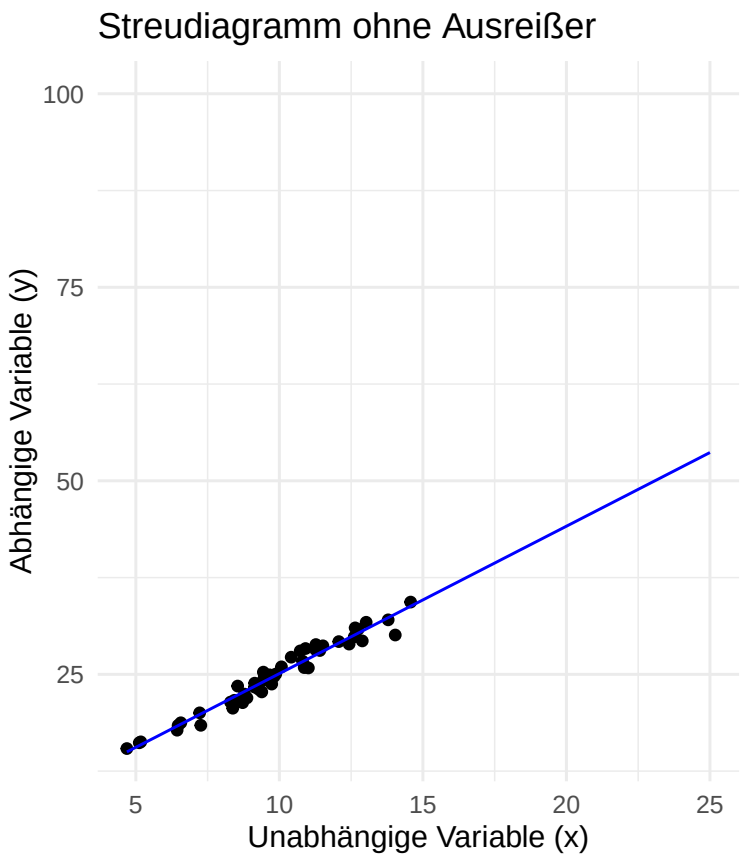




Einfluss von Ausreißern

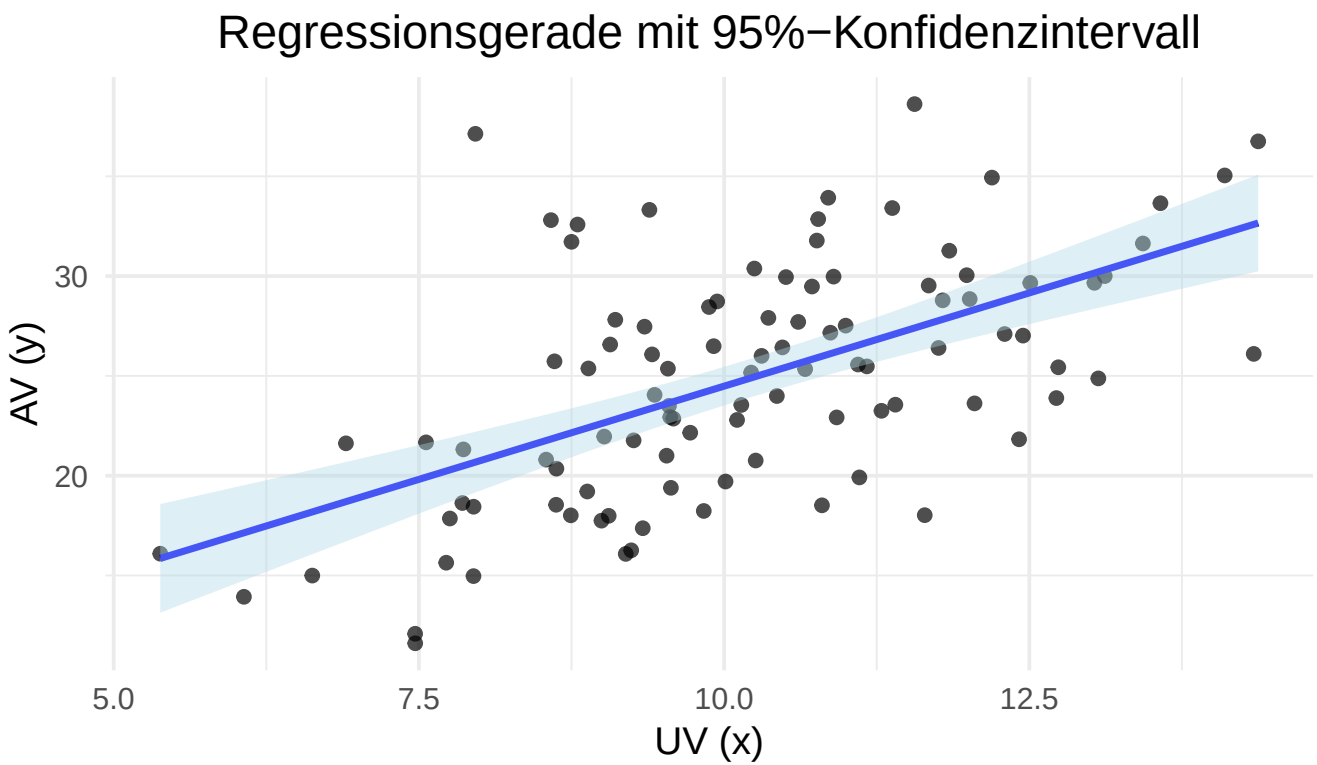
- Ausreißer können als einflussreiche Fälle den Verlauf der Regressionsgeraden überproportional beeinflussen
- Einfluss auf die Regressionsgerade kann getestet werden, z.B. Cook's Distance, Ausreißer ggf. entfernen





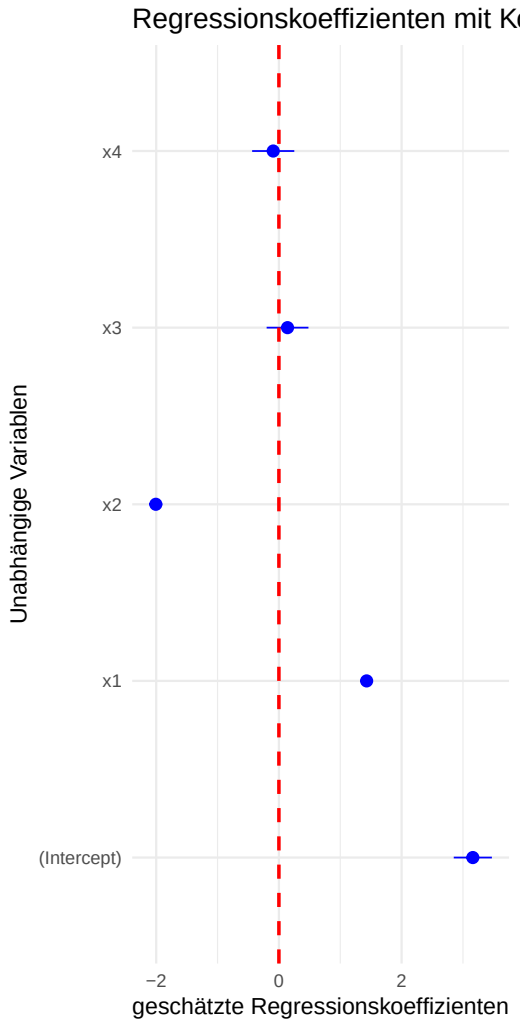
Konfidenzintervall

- für die Regressionskoeffizienten wird ein Konfidenzintervall geschätzt (üblicherweise 95%)
-



Konfidenzintervall

- Die Koeffizienten (+ 95% KI) werden häufig grafisch dargestellt



Konfidenzintervall für Regressionskoeffizienten

- Ein Konfidenzintervall (KI) gibt den Bereich an, in dem der wahre Wert des Regressionskoeffizienten mit einer bestimmten Wahrscheinlichkeit (z.B. 95%) liegt.
- In 95% aller Stichproben deckt das Konfidenzintervall den wahren Wert des Koeffizienten ab.
- Die Formel für das Konfidenzintervall eines Regressionskoeffizienten lautet:

$$\hat{\beta}_j \pm t_{\alpha/2} \cdot \text{SE}(\hat{\beta}_j)$$

- $\hat{\beta}_j$ ist der geschätzte Regressionskoeffizient.
- $t_{\alpha/2}$ ist der kritische Wert der t-Verteilung, z.B. 1,96.
- $\text{SE}(\hat{\beta}_j)$ ist der Standardfehler des geschätzten Koeffizienten.

Interpretation des Konfidenzintervalls

- Wenn das Konfidenzintervall für einen Regressionskoeffizienten den Wert 0 nicht enthält (einschließt), bedeutet dies, dass der Regressionskoeffizient statistisch signifikant von 0 verschieden ist = es gibt einen Effekt (positiv oder negativ).
- Wenn 0 im Konfidenzintervall enthalten ist, wird nicht davon ausgegangen, dass der Koeffizient sich von 0 unterscheidet (bzw. einen Einfluss auf die AV hat).
- das Konfidenzintervall kann Hinweise auf die Stärke des Effekts geben

Ausblick

- die logistische Regression, z.B. Wahl der AfD?
- Welche Variablen sollten wir berücksichtigen?
- Wieso erscheint eine lineare Regresssion ungeeignet?
- und nun Zeit für ein Quiz

partici.fi/04027684

1