

Einige Merkpunkte zur Regression (Anova)

Lorenz Gygax
Abteilung für angewandte Mathematik
Uni-Zürich

April 17, 1996

1 Variablentransformation

Folgende Transformationen sollten (für Ziel- UND erklärende Variablen) immer durchgeführt werden, falls es keine triftigen Gründe dagegen gibt (“Tukey first aid transformations”):

Konzentrationen und Beträge LOG-transformieren

Zählraten WURZELtransformation

Anteile (Prozentzahlen/100) $\tilde{y} = \arcsin\sqrt{y}$

2 Modellwahl

Mit einem grossen Modell beginnen (alle Faktoren und Interaktionen), dann jeweils den Faktor mit dem grössten p-Wert weglassen (wichtig: einzelne Faktoren können nur weggelassen werden, wenn sie an keinen Interaktionen beteiligt sind). Anhand der Residuen entscheiden, ob ein Modell den Annahmen entspricht (dort sieht man auch, ob z. B. ein Faktor auch noch in einer höheren Potenz in das Modell einbezogen werden muss).

3 Prüfen der Annahmen: Residuenanalyse

Mit folgenden graphischen Hilfsmitteln sollen die Annahmen zur Verteilung und Unabhängigkeit der zufälligen Fehler geprüft werden:

Normal plot. Y-Achse: Residuen; X-Achse: Quantile der Normalverteilung. Sind die Residuen normalverteilt (keine Ausreisser, keine schiefe Verteilung, keine Langschwänzigkeit). (Ev. auch anhand eines Histogrammes.)

Tukey-Anscombe plot. Y-Achse: Residuen; X-Achse: Estimate. Ist die Varianz der Residuen konstant über den Bereich der Estimates? Positive und negative Residuen gleich häufig und gleich gross (ev. einen Glätter wie lowess oder loess einzeichnen).

Leverage plot. Y-Achse: Residuen; X-Achse: Leverages (oder ev. Mahalanobis-Distanz). Gibt es Einflussreiche Beobachtungen (Ausreisser: grosse Residuen und grosses Leverage).

Residuen versus erklärende Variablen und Zeit. Y-Achse: Residuen; X-Achse: erklärende Variablen und zeitliche Abfolge. Varianz und Erwartungswert der Residuen sollten über den Wertebereich der Variablen konstant sein (insbesondere keine zeitliche Abhängigkeit!) Ev. mehrdimensional (Residuen versus Kombination von zwei Variablen): gibt Aufschluss über Interaktionen.

Trotz allem: KEINE Überinterpretation! (Beobachtungen können immer mit den gefundenen Resultaten simuliert werden, um ein Gefühl für zufällige Abweichungen zu bekommen.)

4 Literatur

Stahel, Werner A. 1995. Statistische Datenanalyse; Eine Einführung für Naturwissenschaftler. vieweg, Braunschweig/Wiesbaden.