Martin Kerntopf
Email: martin.kerntopf@uni-greifswald.de
25. Januar 2022
Seminar: Data Science für Geistes- und Sozialwissenschaftler

REGRESSIONSANALYSE

V 0.1

Übersicht

1 Einleitung

In diesem Script finden Sie eine Zusammenfassung und weiterführende Erläuterung der im Seminar benutzten begriffe und Verfahren. Zentral hierbei ist die Auflistung der wichtigsten Begrifflichkeiten und Konzepte im Zusammenhang mit dem Thema der Regressionsanalysen. Dies wird ergänzt durch eine Kurzbeschreibung der zentralen Annahmen im Zusammenhang mit der Durchführung einer Regressionsanalyse. Folgend werden zentrale Begriff im Bereich der Ergebnisauswertung und -überprüfung aufgeführt.

1.1 Werte

1.2 Koeffizienten

- a = Der Schnittpunkt mit der y-Achse, wenn x = 0 ist
- b = Die Steigung der Geraden, wenn x um 1 erhöht wird

1.3 Fehlerterm

- ϵ = unbeobachtbare Zufallsvariablen, die den vertikalen Abstand zwischen Beobachtungspunkt und wahrer Gerade messen. Für sie nimmt man für gewöhnlich an, dass sie unkorreliert sind,

einen Erwartungswert von Null und eine homogene Varianz aufweisen. Sie beinhalten unbeobachtete Faktoren, die sich auf die abhängige Variable auswirken.

1.4 Residuen

- Residuen messen den vertikalen Abstand zwischen Beobachtungspunkt und der geschätzten Regressionsgerade. Diese Abstände sind für die Ermittlung der kleinsten Quadrate (Ordinary Least Squares – OLS) relevant

1.5 Grundannahmen

1.5.1 Der Erwartungswert der Residuen ist Null

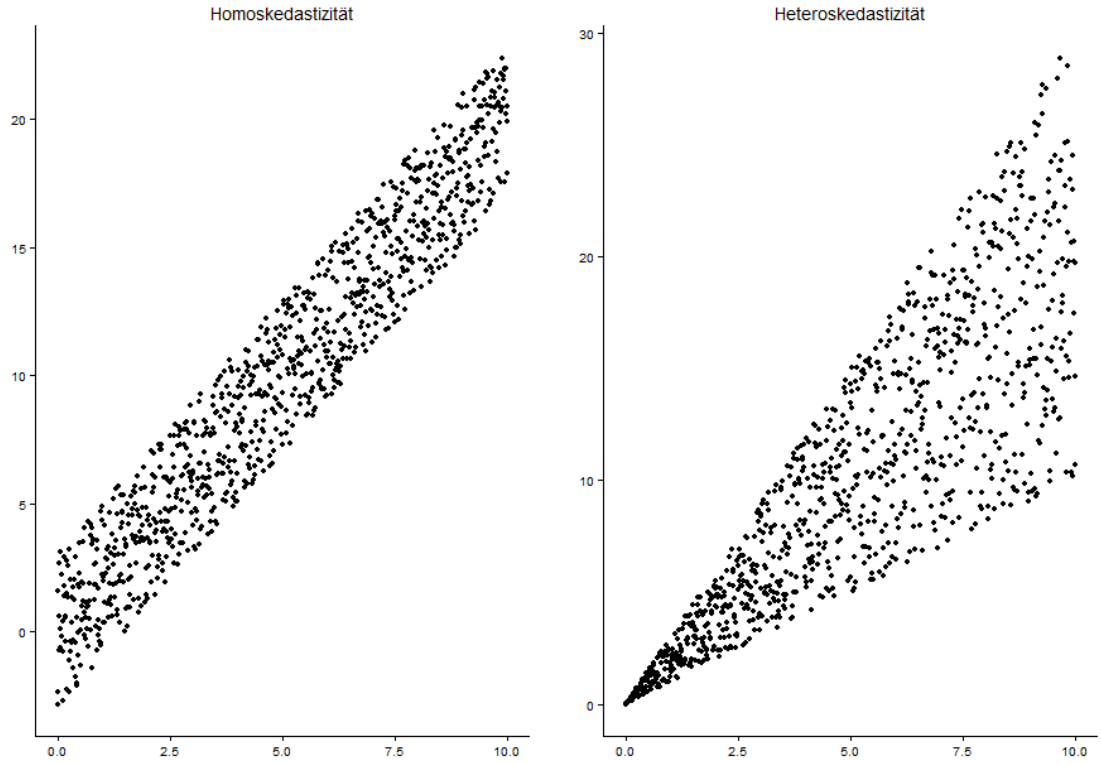
Die erste Annahme geht davon aus, dass die Fehlerterme im Durchschnitt Null ergeben (genannt Erwartungswert). Es gibt sowohl Abweichungen gegen unten als auch gegen oben. Im Schnitt heben sich diese jedoch auf. Deshalb werden auch die Quadratterme der Residuen minimiert und nicht die Residuen selber, da bei Quadratwerten sich negative und positive Abweichungen nicht aufheben.

1.5.2 Fehlerterme sind unkorreliert

Die zweite Annahme ist eine sehr wichtige Annahme. Sie besagt, dass die Residuen nicht miteinander korrelieren. Dies tun sie dann, wenn wichtige Variablen nicht in das Modell aufgenommen wurden.

1.5.3 Homo- und Heteroskedastizität

Die Homoskedastizität besagt, dass die Varianz der Fehlerterme gleichmässig verteilt ist. Ist dies nicht der Fall spricht man von Heteroskedastizität. Liegt Heteroskedastizität vor, ist OLS nicht mehr effizient und es müssen andere Verfahren angewandt werden. Kann optisch oder via Breush Pagan Test ermittelt werden (bei signifikante p-Wert im BPT liegt Heteroskedastizität vor).



1.5.4 Fehlerterme sind normalverteilt

Die Normalverteilung (auch Gauss-Verteilung genannt) ist eine stetige Verteilung. Viele Fehler lassen sich durch diese symmetrische Verteilung erklären.

2 Ergebnis auswerten und überprüfen

2.1 p-Wert – Statistische Signifikanz

Wird ein statistisches Ergebnis als signifikant bezeichnet, so drückt dies aus, dass die Irrtumswahrscheinlichkeit, eine angenommene Hypothese treffe auch auf die Grundgesamtheit zu, nicht über einem festgelegten Niveau liegt. Anders ausgedrückt: in gemessener Zusammenhang zwischen zwei Variablen tritt in der Stichprobe nicht einfach zufällig auf, sondern trifft auch für die Grundgesamtheit zu. Auf Signifikanz geprüft werden können nur Hypothesen, nicht das Ergebnis von Einzelmerkmalen.

- Der p-Wert bewegt sich zwischen 0 und 1, je kleiner er ist, desto signifikanter das Ergebnis
- $p \leq 0.05$ bedeutet das die Nullhypothese zurückgewiesen wird
- $p > 0.05$ heißt das die Alternativhypothese nicht anzunehmen ist. Damit wird die Nullhypothese angenommen

2.2 R^2 – Goodness of Fit

Nach dem eine Regression durchgeführt wurde bleibt meist noch die Frage, wie gut das Regressionsmodell denn die Daten nun tatsächlich erklären kann, d.h. wie gut die Regressionsgerade zu den Daten “passt.” Um diese Frage zu beantworten wird der Coefficient of Determination verwendet,

der auch als R-squared bzw. R^2-Wert bezeichnet wird. Der R^2-Wert ermittelt, welcher Anteil der Gesamtvarianz in den Daten durch die Regressionsgerade erklärt werden kann.

- R^2 repräsentiert den Anteil an der Gesamtvariabilität in y an, der mit dem Regressionsmodell mittels der unabhängigen Variable x erklärt werden kann
- R^2 ist ein Wert zwischen 0 und 1, je näher der R^2-Wert an 1 ist, desto besser passt die Regressionsgerade zu den Daten
- Ein R^2-Wert von z.B. 0.24 bedeutet, dass 24% der Variabilität der abhängigen Variable y mit dem Regressionmodell durch die unabhängige Variable x erklärt werden kann
- Der Adjusted R^2-Wert wird genutzt wenn das Modell mehrere unabhängige Variablen aufweist, da der Multiple R^2-Wert mit zunehmender Anzahl unabhängiger Variablen automatisch steigt

2.3 Multikolliniarität

- Liegt dann vor, wenn mehrere unabhängige Variablen in einer Regressionsanalyse (stark) miteinander korrelieren
  → Variance Inflation Factor (VIF) Test

```
install.packages('regclass') # Das Package das den Test bereitstellt
VIF(modell_name)
```

  – Generell wird bei einem Ergebniswert von 5 bis 10 angenommen, das Multikolliniarität vorliegt

2.4 Heteroskedastizität

- Breush Pagan Test

3 Weiterführende Literatur

- Jürgen Hedderich and Lothar Sachs (2016). *Angewandte Statistik: Methodensammlung mit R*. Heidelberg: Springer
- Christof Wolf and Henning Best, eds. (2010). *Handbuch der sozialwissenschaftlichen Datenanalyse*. Wiesbaden: VS Verlag für Sozialwissenschaften
- Giuseppe Ciaburro (2018). *Regression Analysis with R*. Birmingham: Packt Publishing
- Iain Pardoe (2020). *Applied Regression Modeling*. John Wiley & Sons