Berlin University of Applied Sciences and Technology

Fachbereich II Mathematik - Physik - Chemie

Bachelor - Arbeit

von

Okhtay Wahid Far

zur Erlangung
des akademischen Grades

Bachelor of Science (B.Sc.)

im Studiengang
Mathematik

Thema:

Multivariate Regressionsanalyse exogener Variablen zur Prognose von Ertragszahlen im ÖPNV

Betreuer: Prof. Dr. Thomas Winter
Gutachter: Prof. Dr. Timothy Downie

Eingereicht: 01. Juli 2022

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Berlin, den 01. Juli 2022 

_____

Datum und Unterschrift (Okhtay Wahid Far)

Inhaltsverzeichnis

1. Erläuterung der Forschungsfrage 1

1.1 Abstract 1

1.2 Zusammenfassung 1

1.3 Einleitung 2

2. Erarbeitung in die Themen multivariate Regression und Prognose für Zeitreihendaten 3

2.1 Lineares Modell 4

2.2 Lineare Regression 7

2.2.1 V_1 – Linearität 10

2.2.2 V_2 – Homoskedastizität 11

2.2.3 V_3 – Normalverteilung 13

2.3 Multiple lineare Regression 16

2.3.1 V_4 – Keine Autokorrelation 18

2.3.2 V_5 – Keine perfekte Multikollinearität 22

2.4 Güte der Regression 26

2.4.1 G_1 : P-Wert 27

2.4.2 G_2 : Standardfehler 29

2.4.3 G_3 : Bestimmtheitsmaß 31

2.4.4 G_4 : Korrigiertes Bestimmtheitsmaß 35

2.4.5 G_5 : F-Statistik 36

2.5 T-Test der Regressionskoeffizienten 37

2.6 Ausreißer 39

3. Kurze Darstellung der Regressionsverfahren 42

3.1 Mehrdimensionale/multivariate lineare Regression 42

3.2 Ridge-Regression 43

3.3 LASSO-Regression 44

3.4 LARS-Regression 45

4. Kurze Darstellung der Prognoseverfahren 46

4.1 Einführung Zeitreihe 46

4.2 Güte einer Prognose 48

4.3 Einführung ARIMA-Modell 50

4.4 Nicht-stationäres ARIMA-Modell 52

4.5 Saisonales ARIMA-Modell 52

4.6 ARMAX-Modell 53
5. Beschreibung der Daten und der konkreten Aufgaben-/Problemstellung für den Berliner ÖPNV 53
6. Aufbereitung der Daten (Verkaufszahlen, Wetterdaten, Kalenderdaten) 55
7. Schrittweise Untersuchung und Anwendung verschiedener Regressionsverfahren auf ausgewählte exogene Variable in R 58
8. Vergleich und Bewertung der Güte der Regressionsverfahren auf Basis der erzielten Ergebnisse 60
8.1 Bewertung der Güte-Werte 60
8.2 Vergleich der Regressionsmethoden 76
8.3 Prognose durch Zeitreihenanalyse 80
9. Zusammenfassung und Schlussfolgerungen 83
9.1 Zusammenfassung 83
9.2 Schlussfolgerungen 85

Abbildungsverzeichnis I

Tabellenverzeichnis II

Literaturverzeichnis III

1 Erläuterung der Forschungsfrage

1.1 Abstract

Within this thesis it is to be proven whether and to what extent the multivariate regression analysis of exogenous variables can also lead to a forecast of the yield figures of Berlin's local public transport. The exogenous variables include the population development in Berlin (overall and in age groups, pupils, students), the tourism figures, the influence of the labor market (employed people and commuters), the weather and weather conditions and the price development in local public transport (for individual product groups). The methods used are multivariate regression, ridge regression, LASSO regression and LARS regression in the R language. The results shall detect the suitability of the multivariate regression methods such as the influence of the exogenous variables on the yield figures and forecasts.

1.2 Zusammenfassung

In der vorliegenden Arbeit soll die Hypothese geprüft werden, ob und inwiefern die multivariate Regressionsanalyse exogener Variablen auch zu einer Prognose der Ertragszahlen des öffentlichen Personennahverkehrs (kurz: ÖPNV) in Berlin führen kann. Die exogenen Variablen beinhalten die Bevölkerungsentwicklung in Berlin (insgesamt und in Altersgruppen, Schüler, Studierende), die Tourismuszahlen, die Einflüsse vom Arbeitsmarkt (Erwerbstätige und Pendler), die Wetter-und Witterungsdaten und die Preisentwicklung im ÖPNV (für einzelne Produktgruppen).

Als Methoden werden die multivariate Regression, die Ridge-Regression, die LASSO-Regression und die LARS-Regression in der Sprache R genutzt. Durch die Ergebnisse der Untersuchung soll u. a. eine entsprechende Eignung multivariater Regressionsmodelle im Bereich des öffentlichen Nahverkehrs festgestellt sowie mögliche Einflüsse exogener Variablen auf dazugehörige Ertragszahlen und Prognosen ermittelt werden.

1.3 Einleitung

Diese Bachelorarbeit untersucht die vom Drittmittelprojekt ReComMeND bereitgestellten Datensätze („Treiber_ab_2005.xls“ und „Ertrag_ohne_Schüler.xls“), wobei der Treiber-Datensatz („Treiber_ab_2005.xls) zuvor aufbereitet wurde und die originalen Ertragsdaten von der BVG zur Verfügung gestellt wurden. Mithilfe von multivariaten Regressionsmethoden, um den Einfluss von exogenen Variablen auf die Ertragszahlen im Öffentlichen-Personennahverkehr (ÖPNV) zu ermitteln. Dadurch soll die Hypothese belegt werden, in welchem Maße die exogenen Variablen die Ertragszahlen im Berliner-ÖPNV prägen. Daraus folgernd sollen Prognosen erschlossen und im Hinblick zu Zeitreihen verbunden werden. Die verwendeten Regressionsmethoden sind die multivariate Regression, die Ridge-Regression, die LASSO-Regression und die LARS-Regression. Konkret gilt es auch zu untersuchen, ob eine Aussage nach der Methodendurchführung in der Sprache R bezüglich der jeweiligen Ergebnisqualität getroffen werden kann. Als exogene Variablen gelten die Bevölkerungsentwicklung in Berlin (insgesamt und in Altersgruppen, Schüler, Studierende), die Tourismuszahlen, die Einflüsse vom Arbeitsmarkt (Erwerbstätige und Pendler), die Wetter-und Witterungsdaten und die Preisentwicklung im ÖPNV (für einzelne Produktgruppen).

Aufgrund der stetig steigenden Fahrgastzahl sind Prognosen für die Berliner Verkehrsbetriebe unabdingbar.[1] So wurde der rechtliche Rahmen für Investitionen im Berliner-ÖPNV durch das Mobilitätsgesetz geebnet, welches den Ausbau des Verkehrsnetzes und die zusätzliche Anschaffung von (klimafreundlichen) Fahrzeugen einleitet. In dieses Vorhaben investiert das Land-Berlin jährlich 800 Millionen Euro.[2]

Es ist beabsichtigt, bis 2030 noch mehr Fahrzeuge sowie Bahnen in Berlin einzusetzen, um eine kürzere Taktung und eine größere Verkehrsdichte erzielt werden soll. In diesem Zusammenhang wurden bereits 1,2 Milliarden Euro investiert.[3]

---

[1] Vgl. Süß, 2022, URL: https://unternehmen.bvg.de/news/zahlenspiegel-fuer-2022-ist-da/ [zuletzt zugegriffen am 17.06.2022]
[2] Vgl. Berlin-Webseite, URL: https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/oeffentlicher-personennahverkehr/ausbau/ [zuletzt zugegriffen am 17.06.2022]
[3] Vgl. Neumann, 2022: URL: https://www.berliner-zeitung.de/mensch-metropole/aus-berlin-fuer-berlin-die-ersten-neuen-u-bahnen-kommen-2022-ins-rollen-li.225426 [zuletzt zugegriffen am 17.06.2022]

Die vorliegende Bachelorarbeit widmet sich zunächst den Grundlagen der linearen sowie der multiplen linearen Regression. Zudem werden sowohl deren Eigenschaften als auch deren Anpassungsgüte vorgestellt. Im Anschluss sollen in Kapitel 3 die für die Beantwortung der Forschungsfrage relevanten Methoden definiert werden, wie beispielsweise die multivariate Regression, die Ridge-, die LASSO- und die LARS-Regression.

In Kapitel 4 werden zusätzlich noch die Prognoseverfahren vorgestellt und im Zuge dessen eine Einführung zu den Zeitreihen offeriert.

In den nächsten beiden Kapiteln wird dann die Struktur der Datensätze, deren Verwendung in den Regressionsmethoden und die nutzbaren -modelle beschrieben.

Kapitel 7 widmet sich der sogenannten *Code-Praxis*, wo die Regressionsmethoden in der Sprache R ausgeführt und implementiert werden. Im achten Kapitel werden die Ergebnisse aus dem R-Code bezüglich jeder Regressionsmethode zusammengetragen, miteinander verglichen und abschließend bewertet. In Kapitel 9 sollen die Resultate interpretiert werden, sowie eine Qualitätsbewertung stattfinden und ein kurzer Einblick in die Zeitreihen erfolgen. Im letzten Kapitel werden die relevanten Erkenntnisse zur Beantwortung der eingangs formulierten Hypothese kurz zusammengefasst und eine abschließende Schlussfolgerung diesbezüglich verfasst.

2 Erarbeitung in die Themen multivariate Regression und Prognose für Zeitreihendaten

Dieses Kapitel widmet sich den Grundlagen für die lineare bzw. multiple lineare Regression und deren Eigenschaften. Dadurch kann sich ein Grundverständnis für die Prognose von Zeitreihendaten entwickeln, das in späteren Kapiteln wieder aufgegriffen bzw. als Basis für die Interpretation der gewonnenen Daten genutzt werden soll. Zudem werden die Gütekriterien der Regression im Allgemeinen vorgestellt.

2.1 Lineares Modell

Das lineare Modell versucht ein reduziertes Abbild der tatsächlichen Daten zu schaffen und diese entsprechend schematisch zu vereinfachen. Das heißt, es setzt einen linearen Zusammenhang zwischen den gegebenen Daten der erklärenden Variablen $X_1, \dots, X_i$ und den gegebenen Daten für die Zielvariable Y voraus, wobei $i \in \{1, \dots, \mathrm{n}\}$ mit $n \in \mathbb{N}$ gilt. Konkret soll dadurch eine lineare Funktion gebildet werden, die die Ursache-Wirkung Beziehung zwischen den erklärenden Variablen X_i und der Zielvariable Y sinnvoll umsetzt.[4][5][6]
Im Gegensatz dazu wird beim allgemeinen linearen Modell für die Zielvariable Y keine Normalverteilung vorausgesetzt. Somit stellt also das allgemeine lineare Modell eine Erweiterung des linearen Modells dar.[7]
Insbesondere sind sogar nicht-lineare Zusammenhänge mit dem linearen Modell umsetzbar.[8] Bei metrisch erklärenden Variablen wird hierbei eine Variablen- oder Polynomtransformation durchgeführt. So wird bei der Variablentransformation die erklärende Variable als Nenner eines Bruches mit dem Zähler 1 umformuliert.[9] Bei der Polynomtransformation wird das lineare Modell höchstens zu einem Polynom dritten Grades erweitert, wodurch dieselbe erklärende Variable dem Modell zweimal hinzugefügt wird.[10] Bei beiden Transformationen muss aber der funktionale Zusammenhang bekannt sein.[11]

Insgesamt wird also beabsichtigt, die Zielvariable Y (abhängige Variable) durch die erklärenden Variablen $X_1, \dots, X_i$,auch unabhängige Variablen genannt, in einem linearen Modell auszudrücken. Hierbei stellen die unabhängigen Variablen $X_1, \dots, X_i$ den Einfluss auf die Zielvariable Y dar.[12]

[4] Vgl. Sauer, 2019, S. 246 f.
[5] Vgl. Chambers, 1983, S. 245.
[6] Vgl. Winke, 2020, S. 6.
[7] Vgl. Kabacoff, 2015, S. 302.
[8] Vgl. Fahrmeir, 2007, S. 60.
[9] Vgl. Fahrmeir, 2007, S. 72f.
[10] Vgl. Fahrmeir, 2007, S. 75.
[11] Vgl. Fahrmeir, 2007, S. 64.
[12] Vgl. Chambers, 1983, S. 245.

Sei die Zielvariable y_i mit $i \in \{1, \dots, n\}$ und $n \in \mathbb{N}\backslash\{0\}$ gegeben, wobei n die die Beobachtungen der Zielvariable wiedergibt. Seien zudem die erklärenden Variablen x_{ik} mit $i \in \{1, \dots, n\}$ und $n \in \mathbb{N}\backslash\{0\}$ gegeben, wobei n die Werte der $k-$ten erklärenden Variablen mit $k \in \{1, \dots, p\}$ und $p \in \mathbb{N}\backslash\{0\}$ darlegt. Seien die unbekannten Regressionskoeffizienten β_k und der zufällige Fehler ε_i gegeben. Dann gilt folgendes [13][14]:

$$Formel^{15}:\ y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i$$

$$\beta_0{}^{16}: Konstante\ bzw.\ Intercept$$

Diese zufälligen Fehler ε_i repräsentieren alle Schwankungen.[17] Sie werden auch als Residuen bezeichnet, die insbesondere durch unbeachtete unabhängige Variablen, Messfehler oder zufällige Umstände entstehen können. Konkret stellen die Residuen die Differenz zwischen beobachteten Wert y_i und geschätzten Wert $\hat{y}_i$ dar.[18] Die Singularform von Residuen ist Residuum.[19]

Seien die Residuen ε_i gegeben. Dann gilt für die Residuen folgendes[20][21] :

$$1)\ Residuen:\ \varepsilon_i = y_i - \hat{y}_i$$

$$2) Zusammenhang\ zum\ linearen\ Modell:\ y_i = \hat{y}_i + \varepsilon_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i$$

[13] Vgl. Chambers, 1983, S. 245.
[14] Vgl. Fahrmeir, 2007, S. 60.
[15] Chambers, 1983, S. 245.
[16] Vgl. Fahrmeir, 2007, S.60.
[17] Vgl. Chambers, 1983, S. 245.
[18] Vgl. Winke, 2020, S.10f.
[19] Vgl. Kuckartz, 2010, S. 234.
[20] Vgl. Winke, 2020, S. 11.
[21] Vgl. Eid, 2017, S. 596.

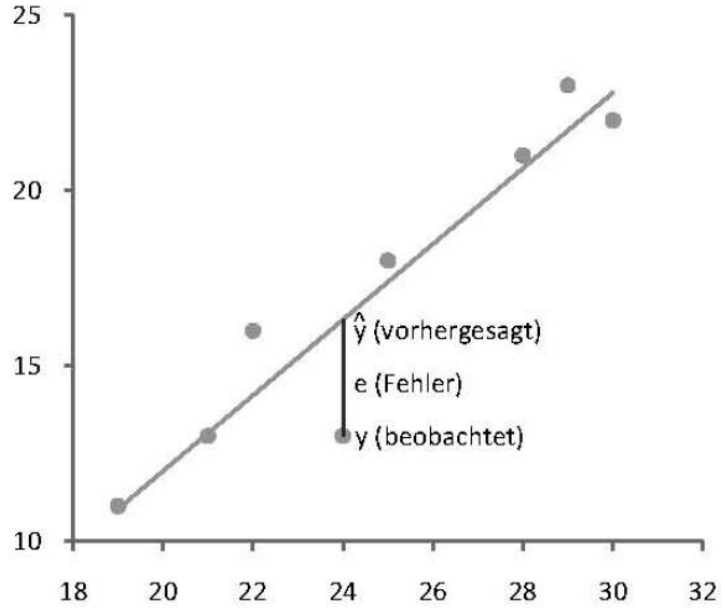

Abbildung 1 - Diagramm: Residuum und Schätzwert (Kuckartz,2010,S.235)

Die Abbildung 1 zeigt einen typischen Zusammenhang zwischen den beobachteten Werten y_i , die als Punkte illustriert sind, und den geschätzten Werten $\hat{y}_i$, die auf der linearen Funktion liegen.[22]

Da die gemessenen Werte eines Datensatzes meistens fehlerbehaftet sind, muss ein Schätzverfahren verwendet werden.[23] Um eine eindeutige Schätzung der Koeffizienten β_k in einem linearen Modell umzusetzen, ist die Zusammenfassung aller erklärender Variablen in einer Designmatrix $\boldsymbol{X}$ mit vollem Spaltenrang hilfreich. Das bedeutet wiederum, dass die Spalten in der Designmatrix $\boldsymbol{X}$ linear unabhängig sind und die Anzahl der Beobachtungen größer ist als die Zahl der Regressionskoeffizienten.[24]

Seien die Anzahl der Beobachtungen n und die Zahl der Regressionskoeffizienten p gegeben. Sei die Designmatrix X mit vollem Spaltenrang $rg(X) = k + 1 = p$ *und mit* $n > p$ *gegeben. Dann gilt für das lineare Modell mit der Designmatrix X folgendes*[25]:

[22] Vgl. Kuckartz, 2010, S. 234.
[23] Vgl. Urban, 2011, S. 40f.
[24] Vgl. Fahrmeir, 2007, S.61.
[25] Fahrmeir, 2007, S. 60f.

$$y_i = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \varepsilon_i = \begin{pmatrix} \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} = \begin{pmatrix} x_1{}' \\ \vdots \\ x_n{}' \end{pmatrix}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \Rightarrow \mathrm{y} = \mathbf{X}\beta + \varepsilon$$

2.2 Lineare Regression

Generell ist das Ziel eines Regressionsmodells, präzise Schätzungen bezüglich der unbekannten Koeffizienten β_0 und β_k durchzuführen.[26] Zusätzlich soll auch ein Näherungsmodell für die tatsächlichen Werte gebildet werden.[27] Des Weiteren definieren β_0 und β_k die genaue Lage einer Geradenfunktion.[28]

Seien die Residuen ε_i *mit* $E(\varepsilon_i) = 0$ *und* $Var(\varepsilon_i) = \sigma^2$ *unabhängig und identisch* $(i.i.d.)$ *verteilt. Seien zudem die metrischen Variablen* x *und* y $mit (y_i, x_i)$ *und* $i \in \mathbb{N}\backslash\{0\}$ *gegeben. Dann gilt für das lineare Regressionsmodell und für dessen geschätzte Regressionsgerade folgendes*[29]*:*

$$1) Lineare\ Regression\colon\ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$2) Geschätzte\ Regressionsgerade\colon\ \hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

In diesem Zusammenhang wird die Schätzfunktion $\hat{y}$ auch als Prognose bezeichnet.[30] Insbesondere wird die Regressionsschätzung gewöhnlich mit der Kleinst-Quadrate-

[26] Vgl. Winke, 2020, S. 9.
[27] Vgl. Kabacoff, 2015, S. 171.
[28] Vgl. Urban, 2018, S. 31.
[29] Vgl. Fahrmeir, 2007, S. 22.
[30] Vgl. Fahrmeir, 2007, S. 22.

Methode durchgeführt, welches auch als OLS-Schätzung bezeichnet wird.[31] Hierbei entstammt die Abkürzung OLS aus dem Englischen (*Ordinary Least Squares*) und wird mit gewöhnliche Methode der kleinsten Quadrate übersetzt.[32]

Aus praktischen Gründen berufen wir uns bei der Regressionsschätzung stets auf eine Zufallsstichprobe, da die wahrhaftigen Werte aus der Grundgesamtheit nicht messbar sind. Um präzise Schätzwerte für die Regressionskoeffizienten β_0 und β_1 ermitteln zu können, bedarf es gewisser Voraussetzungen. Demzufolge nehmen wir also an, dass die Regressionskoeffizienten β_0 und β_1 unverzerrt (engl. „unbiased“) sind. So besagt die Unverzerrtheit, dass der Erwartungswert des Schätzwertes mit dem tatsächlichen Wert aus der Grundgesamtheit übereinstimmt. Das heißt, der Schätzwert ist demnach nicht identisch mit dem tatsächlichen Wert aus der Grundgesamtheit. Zusammenfassend repräsentiert der Erwartungswert $E(X)$ den Durchschnittswert einer diskreten Zufallsvariable, wo die Anzahl der Messungen bezüglich der Zufallsvariable gegen unendlich strebt.[33]

Sei X eine diskrete Zufallsvariable und sei $i, k \in \mathbb{N}\backslash\{0\}$ *gegeben. Dann gilt für den Erwartungswert einer diskreten Zufallsvariable folgendes*[34] :

$$\mu = E(X) = x_1 p_1 + \ldots + x_k p_k + \ldots = \sum_{i \geq 1} x_i \, p_i$$

$$\mu : Erwartungswert$$
$$x_1, \ldots, x_k, \ldots : Werte\ der\ diskreten\ Zufallsvariable\ X$$
$$p_1, \ldots, p_k, \ldots : Werte\ der\ Wahrscheinlichkeitsverteilung$$

Zusätzlich muss neben der Unverzerrtheit noch die Varianz betrachtet werden, damit die Schätzung so exakt wie möglich wird. Man spricht hierbei von der Effizienz

[31] Vgl. Urban, 2011, S. 46.
[32] Vgl. Backhaus, 2021, S. 79.
[33] Vgl. Urban, 2011, S. 115ff.
[34] Vgl. Fahrmeir, 2016, S. 226.

einer Schätzung.[35] Die Varianz umfasst die Streuung von Daten um deren Mittelwert und eignet sich daher nur für metrische Variablen.[36]

$$Sei\ eine\ diskrete\ Zufallsvariable\ X\ gegeben. Sei\ k \in \mathbb{N}\backslash\{0\}\ und\ die\ relativen\ Häufigkeiten\ f_1, \dots, f_k, \dots\ gegeben.\ Dann\ gilt\ für\ die\ Varianz\ und\ Standardabweichung\ folgendes[37]:$$

$$1)\ Varianz:\ \sigma^2 = Var(X) = \sum_{i \geq 1} (x_i - \mu)^2 f(x_i)$$

$$2)\ Standardabweichung:\ \sigma = +\sqrt{Var(X)}$$

Das letzte Kriterium zur Bewertung der Schätzungen befasst sich mit der Konsistenz. Wenn sich der Stichprobenumfang vergrößert, so gelten die Schätzwerte als konsistent. Hierfür muss eine Reduzierung der Verzerrung und der Varianz der Schätzwerte vorherrschen. Mit den Grundbedingungen der Unverzerrtheit, Effizienz und Konsistenz können wir nun die Voraussetzungen für eine lohnenswerte Regressionsschätzung formulieren.[38]

$$Seien\ \varepsilon_i\ mit\ E(\varepsilon_i) = 0\ \ unabhängig\ und\ identisch\ verteilt\ (i.d.d.)$$
$$Dann\ gilt\ für\ die\ Annahmen\ der\ linearen\ Regression\ folgendes[39][40]:$$

$$V_1 - Linearität[41]:$$
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$V_2 - Homoskedastizität\ mit\ (i.d.d.):$$
$$i)\ Var(\varepsilon_i) = \sigma^2$$

[35] Vgl. Urban, 2011, S. 118.
[36] Vgl. Fahrmeir, 2016, S. 64.
[37] Vgl. Fahrmeir, 2016, S. 231.
[38] Vgl. Urban, 2011, S. 119f.
[39] Vgl. Urban, 2018, S. 29ff.
[40] Vgl. Fahrmeir, 2007, S. 21.
[41] Vgl. Urban, 2018, S. 29ff.

$$V_3 - Normalverteilung:$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

2.2.1 V_1 – Linearität

Die Linearitätsannahme vermittelt den Aspekt, dass eine Linearität zwischen den unabhängigen Variablen X_i und der abhängigen Variable Y bestehen muss. Anders gesagt, wenn sich in einem linearen Regressionsmodell der Wert in X_i erhöht oder verringert, so erhöht oder verringert sich auch stets der Wert in Y linear.[42]

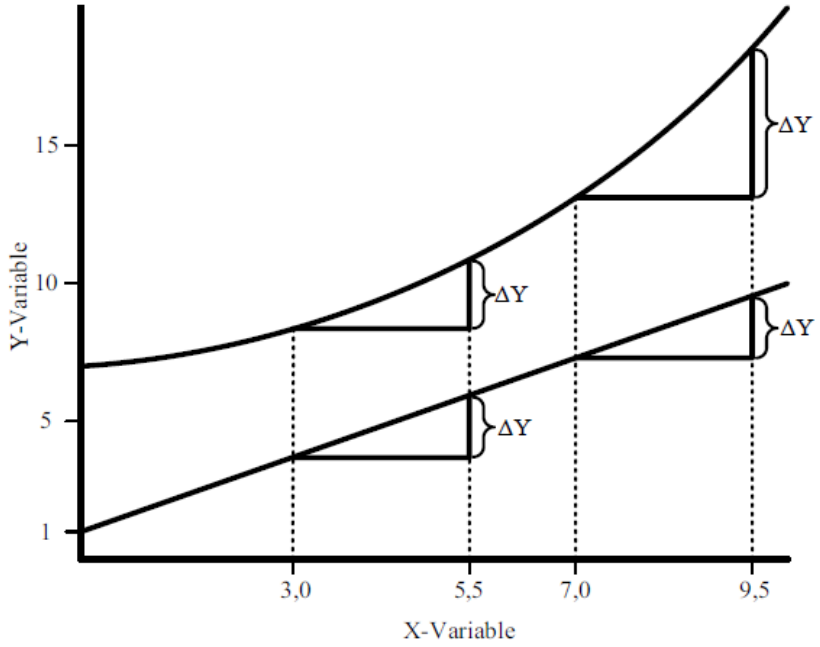

Abbildung 2 - Diagramm: Linearität und Nicht-Linearität (Urban, 2018, S.30)

Anhand der Abbildung 2 erkennen wir beim unteren Graf die Illustration eines linearen Zusammenhangs zwischen einer X- Variable und dessen Y-Variable. Im oberen Graf gilt die Linearitätsannahme nicht, da die Wertänderung der X- Variable nicht mit jener der Y-Variable übereinstimmt.[43]

[42] Vgl. Urban, 2018, S. 29f.
[43] Vgl. Urban, 2018, S. 29f.

In der Praxis mit R nutzt man das Residuendiagramm, um die Linearitätsannahme visuell festzustellen.[44] Hierbei wird ein Residuendiagramm erstellt, welches über die geschätzten Werte und die Residuen dargestellt wird.[45]
Insbesondere wird hier die zufällige und regelmäßige Verteilung der Residuen um die Null-Achse betrachtet. Konkret geht man also von einer Nicht-Linearität des Regressionsmodells aus, wenn die Residuenverteilung einem Muster folgt und nicht zufällig verläuft.[46]

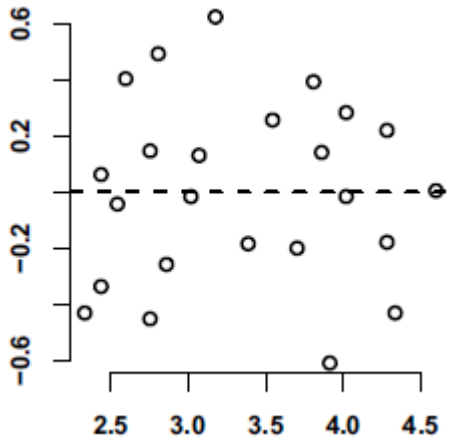

Abbildung 3 - Diagramm: Residuendiagramm zur Linearitätsprüfung (Hedderich,2020,S.821)

In R nutzen wir hierzu die Befehle $residuals()$, $fitted()$ und $abline()$.[47]
Generell führt eine Verletzung der Linearitätsannahme zur Verzerrung der Regressionskoeffizienten.[48]

2.2.2 V_2 - Homoskedastizität

Die Homoskedastizität, auch konstante Varianz genannt, beschreibt die gleichmäßige Varianz der Y-Werte um die Regressionsgerade herum.[49] Die Verletzung der Homoskedastizität wird als Heteroskedastizität bezeichnet.[50] Um die

[44] Vgl. Urban, 2011, S. 204.
[45] Vgl. Handl, 2017, S. 245.
[46] Vgl. Urban, 2011, S.205.
[47] Vgl. Handl, 2017, S. 245.
[48] Vgl. Backhaus, 2021, S. 102.
[49] Vgl. Fahrmeir, 2007, S. 64f.
[50] Vgl. Backhaus, 2021, S. 115.

Homoskedastizität visuell über eine Programmiersprache, wie zum Beispiel R, darzustellen, verwendet man ein Residuendiagramm.[51]

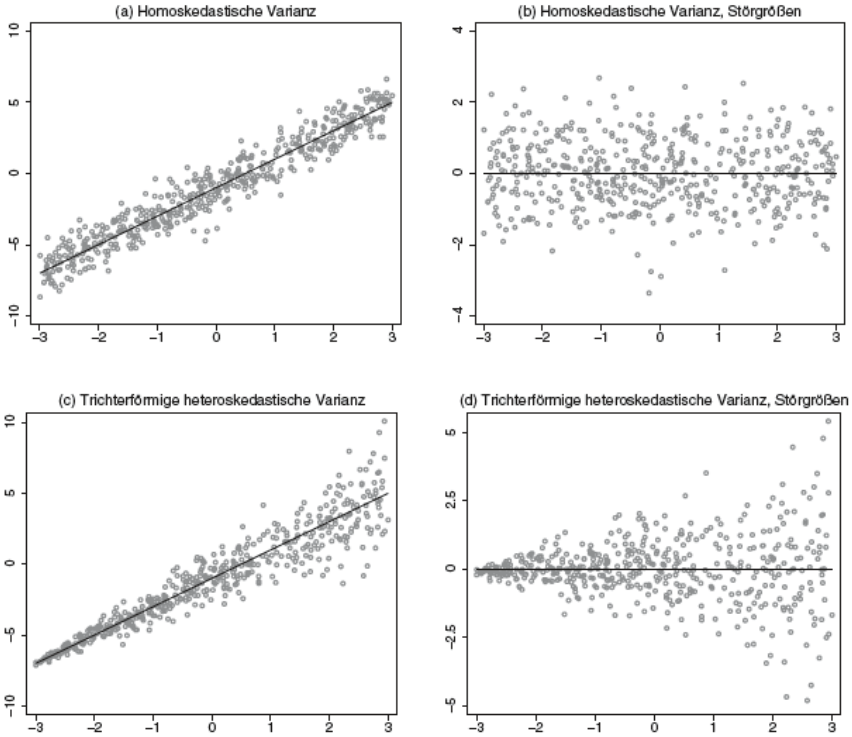

Abbildung 4 - Diagramm: Homoskedastizität und Heteroskedastizität (Fahrmeir,2007, S.65)

Die Diagramme in der ersten Zeile geben eine typische Homoskedastizität wieder, wo die Residuen relativ gleichmäßig und konstant um die Regressionsgerade $\hat{y}$ existieren. Die Diagramme in der zweiten Zeile beschreiben die typische Trichterform bei der Heteroskedastizität, wo die Streuung bei stetigem Zuwachs der X-Variable zunimmt.[52]

Zusätzlich ist auch die Erstellung eines Residuendiagramms möglich, wo die geschätzten Werte und die Residuen betrachtet werden. Hierbei sollte eine gleichmäßige Residuenverteilung um die Null-Achse gegeben sein, um von einer Homoskedastizität sprechen zu können.[53] Generell kann man die Prüfung auf Heteroskedastizität auch als Prüfung der Nicht-Linearität betrachten, welches man aber noch zusätzlich untersuchen sollte.[54]

[51] Vgl. Sauer, 2019, S. 330.
[52] Vgl. Fahrmeir, 2007, S. 64f.
[53] Vgl. Wollschläger, 2014, S. 200.
[54] Vgl. Backhaus, 2021, S. 117.

Bei Verletzung der Homoskedastizität werden zwar die geschätzten Regressionskoeffizienten nicht verzerrt. Trotz allem leidet aber die Genauigkeit der OLS-Schätzung.[55]

„Heteroskedastizität führt nicht zu verzerrten Schätzern, aber die Präzision der Schätzung mit der KQ-Methode wird verringert. Und auch die Standardfehler der Regressionskoeffizienten, ihre p-Werte und die Schätzung der Konfidenzintervalle werden ungenauer"(Backhaus,2021,S.115).

2.2.3 V_3 - Normalverteilung

Die Normalverteilung ist eine spezielle Verteilung, die auch als Gauß-Verteilung bezeichnet wird.[56]

Sei X eine Zufallsvariable und seien die Parameter μ *und* σ^2 *mit* $X \sim N(\mu, \sigma^2)$ *gegeben. Dann ist X normalverteilt, wenn für die Dichtefunktion* $f(x)$ *folgendes gilt*[57]:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$exp(x) := e^x$$

Bei $x = \mu$ existiert das Maximum der Normalverteilung, wobei diese glockenförmig gegen Null strebt.[58] Dadurch wird die Hauptlage der Verteilung festgelegt [59]
Die Normalverteilung $N(0, \sigma^2)$ der Daten besagt, dass sich der Mittelwert μ beim Nullpunkt sammelt. Insbesondere beeinflusst die Varianz σ^2 die Normalverteilung soweit, dass je größer sie ist, desto flacher wird die Kurve. Zusätzlich erhöht sich auch dementsprechend die Streuung.[60]

[55] Vgl. Backhaus, 2021, S. 115.
[56] Vgl. Fahrmeir, 2016, S. 83.
[57] Vgl. Schlittgen, 2012, S. 250.
[58] Vgl. Fahrmeir, 2016, S. 84.
[59] Vgl. Schlittgen, 2012, S. 251.
[60] Vgl. Schlittgen, 2012, S. 251.

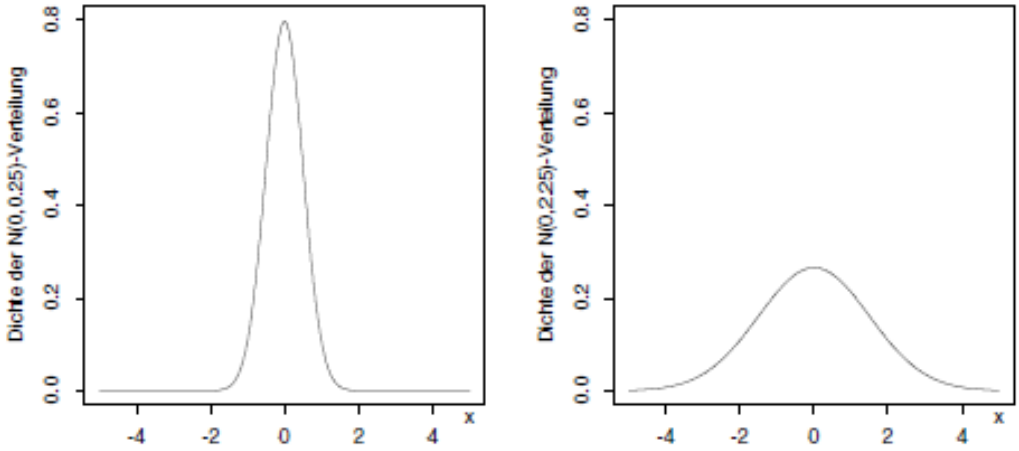

Abbildung 5 - Diagramm: Normalverteilung (Fahrmeir,2016,S.85)

In der Abbildung 5 erkennt man welchen Einfluss die Standardabweichung σ auf die Normalverteilung nimmt. In diesem Zusammenhang ist auf dem rechten Diagramm zu beobachten, dass aufgrund der Standardabweichung $\sigma = 2.25$ sowohl die Verteilung abflacht als auch μ kleiner wird. Folglich ist die Streuung im linken Diagramm schmaler, da für die Standardabweichung $\sigma = 0.25$ gilt.[61]

Mithilfe der R-Funktionen $qqplot()$ und $qqline()$ ist es unter anderem möglich, die Normalverteilung der Residuen anhand dessen Anordnung an der Regressionsgerade zu erkennen.[62]

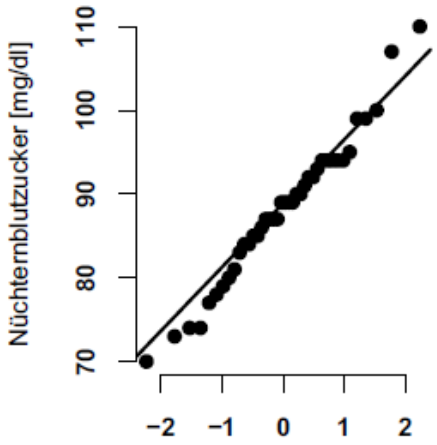

Abbildung 6 - Diagramm: QQ-Plot der Normalverteilung (Hedderich,2020,S.487)

[61] Vgl. Fahrmeir, 2016, S. 84f.
[62] Vgl. Hedderich, 2020, S. 487f.

In Abbildung 6 ist ein QQ-Plot mit einem Stichprobenumfang von $n = 40$ illustriert, wo man von einer Normalverteilung ausgeht.[63] Denn es ist zu erkennen, dass eine zufällig nahe Streuung der Residuen um die Regressionsgerade existiert. Abweichungen am Anfang und am Ende der Regressionsgerade haben keinen Einfluss auf die Normalität und sind in der Realität üblich.[64]

Insbesondere sind normalverteilte Residuen vorteilhaft für weitere Teststatistiken und die Verwendung von Konfidenzintervallen.[65] Aufgrund des zentralen Grenzwertsatzes werden bei einer Stichprobe von $n > 40$ die Schätzungen normalverteilt sein.[66] Das heißt, bei identisch verteilten und unabhängigen Zufallsvariablen wird bei circa 40 Summanden eine annähernde Normalverteilung vorherrschen.[67]

$$Seien\ X_1, X_2, \dots, X_n\ identisch\ verteilte\ und\ unabhängige\ (i.d.d.)$$
$$Zufallsvariablen\ gegeben.\ Zudem\ sei\ E(X_i) = \mu\ und\ Var(X_i) = \sigma^2 > 0$$
$$mit\ i \in \{1, \dots n\}\ und\ n \in \mathbb{N}\backslash\{0\}\ \ gegeben.\ Dann\ gilt\ für\ den$$
$$zentralen\ Grenzwertsatz\ folgendes^{68}:$$

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi\,(z)$$

$$Z_n: standardisierte\ Zufallsvariable$$
$$\Phi(z): standardisierte\ Normalverteilung$$

[63] Vgl. Hedderich, 2020, S. 487.
[64] Vgl. Backhaus, 2021, S. 120.
[65] Vgl. Fahrmeir, 2007, S. 21.
[66] Vgl. Backhaus, 2021, S. 121.
[67] Vgl. Schlittgen, 2012, S. 259.
[68] Vgl. Schlittgen, 2012, S. 258f.

2.3 Multiple lineare Regression

Das Regressionsmodell der multiplen linearen Regression basiert darauf, dass mehrere Regressoren X_i die Zielvariable Y gleichzeitig beeinflussen.[69] Hierbei werden nur die relevanten Regressoren betrachtet, die einen unentbehrlichen Einfluss auf die Zielvariable Y nehmen. Somit verhindert man ein Underfitting des multiplen linearen Regressionsmodells und die Verzerrung der Schätzer.[70] Gleichzeitig entsteht bei der Nutzung von zu vielen Regressoren ein Overfitting, wodurch die Qualität des Regressionsmodells gemindert wird.[71] Insgesamt stellt die richtige Wahl der essentiellen unabhängigen Variablen X_i eine große Herausforderung für jeden Anwender dar.[72]

„Modellierung ist ein Balanceakt zwischen Einfachheit und Komplexität oder zwischen underfitting und overfitting.“(Backhaus,2021,S.93)

$$Seien\ die\ Residuen\ \varepsilon_1, \dots, \varepsilon_n\ mit\ E(\varepsilon_i) = 0\ und\ Var(\varepsilon_i) = \sigma^2\ unabhängig\ und\ identisch\ verteilt\ (i.i.d.).\ Seien\ die\ metrischen\ Variablen\ y\ und\ die\ metrischen\ Regressoren\ x_1, \dots, x_k\ mit\ (y_i, x_{i1}, \dots, x_{ik})\ und\ i,k \in \mathbb{N}\backslash\{0\}\ gegeben.\ Dann\ gilt\ für\ das\ multiple\ lineare\ Regressionsmodell\ und\ dessen\ Schätzfunktion\ folgendes^{73}:$$

1) $Multiple\ lineares\ Regressionsmodell: y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$

2) $Geschätzte\ Regressionsgerade: \hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_k} x_k$

Wichtig sei aber anzumerken, dass die Annahmen für das lineare und multiple lineare Regressionsmodell lediglich notwendig und nicht hinreichend sind.[74]

[69] Vgl. Schlittgen, 2012, S. 442.
[70] Vgl. Backhaus, 2021, S. 106.
[71] Vgl. Backhaus, 2021, S. 93.
[72] Vgl. Backhaus, 2021, S. 108.
[73] Vgl. Fahrmeir, 2007, S. 29.
[74] Vgl. Backhaus, 2021, S. 102.

„Um eine ausreichende Präzision der Schätzungen zu erreichen, sind auch eine ausreichende Variation (Streuung) der unabhängigen Variablen, eine ausreichend große Stichprobengröße und eine geringe Multikollinearität erforderlich.“ (Backhaus,2021,S.102f)

Für das multiple lineare Regressionsmodell gelten dieselben Voraussetzungen wie für das lineare Regressionsmodell. Es gelten also die Annahmen V_1 (Linearität), V_2(Homoskedastizität) und V_3(Normalverteilung).[75]

$$Sei\ V_1\ , V_2\ und\ V_3\ gegeben.\ Sei\ zudem\ j \in \{1, \dots, p\}$$
$$mit\ k, p \in \mathbb{N}\backslash\{0\}\ \ und\ k \neq j\ gegeben.\ Sei\ p\ die\ Anzahl\ der\ unbekannten$$
$$Parameter\ und\ sei\ zudem\ X_0 \equiv 1.\ Dann\ gilt\ f\ddot{u}r\ die\ Annahmen$$
$$der\ multiplen\ linearen\ Regression\ folgendes[76][77][78]:$$

$$V_4 - Keine\ Autokorrelation[79]:$$
$$Cov(\varepsilon_i\ , \varepsilon_{i+r}) = 0\ \ mit\ i \in \{1, \dots, n-r\}\ und\ n, r \in \mathbb{N}\backslash\{0\}$$

$$V_5 - Keine\ perfekte\ Multikollinearit\ddot{a}t[80]:$$
$$Keine\ Variable\ X_j\ darf\ sich\ als\ Linearkombination\ der\ restlichen\ Variablen\ X_k$$
$$darstellen\ lasssen.\ Das\ hei\ss t, es\ darf\ f\ddot{u}r\ kein\ j$$
$$X_j = \sum_{k \neq j} a_k X_k + b$$
$$gelten.\ Insbesondere\ darf\ also\ nicht\ eine\ erkl\ddot{a}rende\ Variable\ X_j\ aus\ einer$$
$$anderen, etwa\ X_k\ , durch\ Lineartransformation\ hervorgehen.$$

[75] Vgl. Winke, 2020, S. 24.
[76] Vgl. Winke, 2020, S. 24.
[77] Vgl. Backhaus, 2021, S. 118.
[78] Vgl. Fahrmeir, 2016, S. 455.
[79] Vgl. Backhaus, 2021, S. 118.
[80] Vgl. Fahrmeir, 2016, S. 455.

2.3.1 V_4 – Keine Autokorrelation

Die Autokorrelation beinhaltet, dass eine positive oder negative Korrelation zwischen den Residuen ε_i und ε_{i+r} in einem Regressionsmodell vorherrscht. Das heißt, man spricht von der Unabhängigkeit der Residuen, wenn keine Autokorrelation existiert.[81]Autokorrelation ist besonders bei Zeitreihen anzutreffen, wo über einem längeren Zeitintervall Daten gemessen wurden.[82]

Denn aufgrund der zeitlichen Nähe, Häufigkeit und gegenseitigen Abhängigkeit der gemessenen Werte sind Korrelationen zwischen den Residuen in einer Zeitreihe sehr wahrscheinlich.[83] Die Idee der Residuen-Unabhängigkeit ist es festzustellen, ob die Entstehung der Beobachtungen (zeitlich) unabhängig voneinander und zufällig entstanden ist.[84]
Konkret misst die Kovarianz den linearen Zusammenhang zwischen zwei Zufallsvariablen.[85]

$$Seien\ X_i\ und\ X_j\ Zufallsvariablen\ mit\ i,j \in \mathbb{N}\backslash\{0\}\ gegeben.$$
$$Seien\ zudem\ \mu_i = E(X_i)\ und\ \mu_j = E(X_j)\ gegeben.\ Dann\ gilt\ f\ddot{u}r$$
$$die\ Kovarianz\ folgendes[86]:$$

$$Cov(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j)$$
$$\sigma_{ij} : Symbol\ f\ddot{u}r\ die\ Kovarianz$$

Die Daten der Kovarianz σ_{ij} kann man in einer quadratischen Matrix A zusammenfassen. Diese Matrix wird dann als Varianz-Kovarianz-Matrix oder auch kurz als Kovarianz-Matrix bezeichnet.[87]

[81] Vgl. Winke, 2020, S. 33.
[82] Vgl. Urban, 2011, S. 260.
[83] Vgl. Schlittgen, 2004, S. 78f.
[84] Vgl. Winke, 2020, S. 33.
[85] Vgl. Everitt, 2011, S. 12.
[86] Everitt, 2011, S. 12.
[87] Vgl. Everitt, 2011, S. 12f.

Sei eine quadratische Matrix A, also mit identischer Zeilen − und Spaltenanzahl, gegeben. Sei zudem $q \in \mathbb{N}\backslash\{0\}$ gegeben. Dann gilt für die Varianz − Kovarianz − Matrix folgendes[88]:

$$A = \begin{pmatrix} {\sigma_{11}}^2 & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{21} & {\sigma_{22}}^2 & \cdots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \cdots & {\sigma_{qq}}^2 \end{pmatrix}$$

Da die Kovarianz als Maß noch ungeeignet ist, nutzen wir den Korrelationskoeffizienten nach Bravais-Pearson, um die Werte der Kovarianz zu standardisieren.[89]

Seien eine Zufallsvariable X, der Erwartungswert $\mu = E(X)$ und die Standardabweichung $\sigma = Var(X)$ gegeben. Dann gilt für die Standardisierung einer Zufallsvariable X folgendes[90]:

$$Z = \frac{X - \mu}{\sigma} \; mit \; E(Z) = 0 \; und \; Var(Z) = 1$$

Sei der Korrelationskoeffizient (von Bravais − Pearson) r_{XY} der Variablen X und Y gegeben, die aus den Werten (x_v, y_v) die berechnete Maßzahl darstellt. Sei zudem $v \in \{1, \dots, n\}$ mit $n \in \mathbb{N}\backslash\{0\}$ gegeben. Dann gilt für den Korrelationskoeffizient folgendes[91]:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\frac{1}{n}\sum_{v=1}^{n}(x_v - \bar{x})(y_v - \bar{y})}{\sqrt{\frac{1}{n}\sum_{v=1}^{n}(x_v - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{v=1}^{n}(y_v - \bar{y})^2}}$$

r_{XY} : *Korrelationskoeffizient der x − und y − Werte*

[88] Vgl. Everitt, 2011, S. 12f.
[89] Vgl. Schlittgen, 2012, S. 95.
[90] Vgl. Fahrmeir, 2016, S. 269.
[91] Vgl. Schlittgen, 2012, S. 95.

Der lineare Zusammenhang von r_{XY} wird anhand der Maßzahl interpretiert.[92]

$$Sei\ der\ Korrelationskoeffizient\ (von\ Bravais-Pearson)\ r_{xy}\ gegeben.$$
$$Dann\ gilt\ f\ddot{u}r\ die\ Interpretation\ von\ r_{XY}\ folgendes^{93}:$$

$Korrelationskoeffizient\ \lvert r\rvert$	$Interpretation$
0	$keine\ Korrelation$
$0-0.5$	$schwache\ Korrelation$
$0.5-0.8$	$mittlere\ Korrelation$
$0.8-1$	$starke\ Korrelation$
1	$perfekte\ Korrelation$

Durch die Standardisierung der Werte erstrecken sich die Korrelationswerte von -1 bis +1. Ein negativer Korrelationskoeffizient bedeutet, dass die Variablen X und Y in die entgegen gesetzte Richtung korrelieren. Das heißt, wenn X steigt, dann fällt Y. Anders sieht es bei einem positiven Korrelationskoeffizienten aus, wo X und Y in die identische Richtung korrelieren. Das heißt, wenn X steigt, so steigt auch Y.[94]

Infolgedessen verwendet man für die Prüfung der Autokorrelation den Durbin-Watson-Test.[95]

$$Seien\ \varepsilon_i\ die\ Residuen\ mit\ i,n \in \mathbb{N}\backslash\{0\}\ gegeben.\ Dann\ gilt$$
$$f\ddot{u}r\ den\ Durbin-Watson-Test\ folgendes\ ^{9697}:$$

[92] Vgl. Schlittgen, 2012, S. 97.
[93] Vgl. Schlittgen, 2012, S. 97.
[94] Vgl. Kessler, 2007, S. 16.
[95] Vgl. Hedderich, 2020, S. 808.
[96] Vgl. Hedderich, 2020, S. 808.
[97] Vgl. Winke, 2020, S. 35.

$$d = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}{\varepsilon_i}^2}$$

$$d < 2 : Positive\ Autokorrelation$$
$$d = 2 : Keine\ Autokorrelation$$
$$d > 2 : Negative\ Autokorrelation$$

Anhand des folgenden Diagramms kann man das Interpretationsspektrum der Autokorrelation nachvollziehen. Hierbei wird d_O als Obergrenze und d_U als Untergrenze bezeichnet.[98]

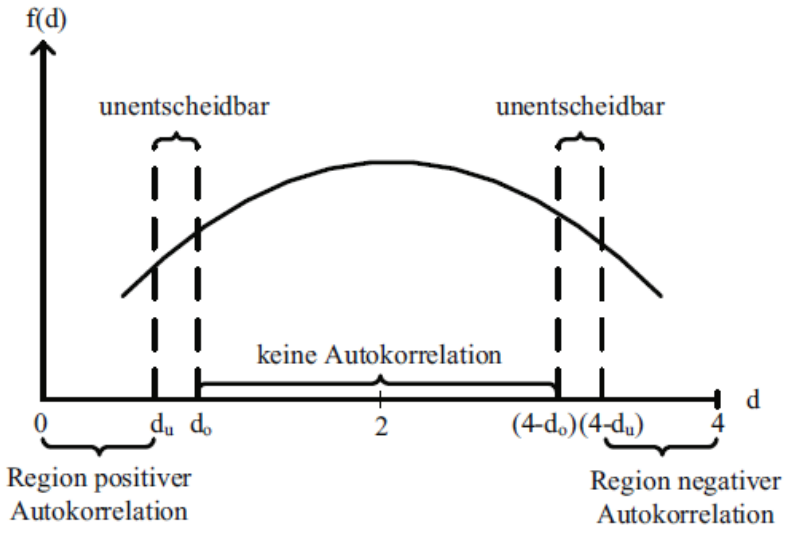

Abbildung 7 - Diagramm: Autokorrelation (Urban,2011,S.266)

Um die Autokorrelation zu überprüfen, nutzt man in der Anwendung folgendes Schema[99] :

$$i)\ d < 1 \Rightarrow Starke\ Autokorrelation$$

$$ii)\ d > 3 \Rightarrow Starke\ Autokorrelation$$

$$iii)\ d \in [1.5\ , 2.5] \Rightarrow Keine\ Autokorrelation$$

[98] Vgl. Urban, 2011, S. 265.
[99] Vgl. Urban, 2011, S. 266.

Da die Prüfung der Autokorrelation über ein Streudiagramm nur bedingt feststellbar ist[100], behilft man sich zusätzlich mit dem Durbin-Watson-Test, der auch in R mit der Funktion $dwtest()$ Verwendung findet.[101]
Allgemein ist festzuhalten, dass Autokorrelation keine verzerrten Schätzer hervorruft, wobei es aber gleichzeitig die Genauigkeit der OLS-Schätzung negativ beeinflusst. Zusätzlich entstehen noch andere negative Nebeneffekte.[102]

„Die Standardfehler der Regressionskoeffizienten, ihre p-Werte und die Schätzung der Konfidenzintervalle werden ungenauer."(Backhaus,2021,S.117)

Zudem verstärkt eine identifizierte Autokorrelation den Verdacht auf eine existierende Nicht-Linearität.[103] Insgesamt gilt für Zeitreihen, dass die Unabhängigkeit der Residuen sehr unwahrscheinlich ist.[104]

2.3.2 V_5 – Keine perfekte Multikollinearität

Diese Annahme besagt, dass in einer multiplen Regression keine starke lineare Beziehung zwischen allen unabhängigen Variablen X_i untereinander existieren darf. Grundsätzlich soll nämlich jeder Regressor einen eigenständigen Einfluss auf die Zielvariable Y darstellen. Der Einfluss der Regressoren untereinander sollte soweit auf das Minimum gehalten werden, so dass ein Regressor nicht als lineare Funktion aus mehreren anderen Regressoren bestimmt werden kann.[105]

Erfahrungsgemäß ist aber eine latente lineare Beziehung zwischen den Regressoren vorhanden, da diese inhaltlich miteinander zusammenhängen. Eine perfekte Multikollinearität ist aber in der Realität sehr selten anzutreffen. In der Praxis spricht man also eher von einer hohen Multikollinearität statt einer perfekten Multikollinearität.[106]

[100] Vgl. Urban, 2011, S. 264.
[101] Vgl. Hedderich, 2020, S. 808.
[102] Vgl. Backhaus, 2021, S. 117.
[103] Vgl. Backhaus, 2021, S. 119.
[104] Vgl. Fahrmeir, 2007, S. 66ff.
[105] Vgl. Urban, 2011, S. 225.
[106] Vgl. Backhaus, 2021, S. 121.

Kollinearität wird durch die Nutzung des Korrelationskoeffizienten geprüft, die aber nur zwei unabhängige Variablen gleichzeitig untersuchen kann. Zwar kann eine starke Korrelation zwischen unabhängigen Variablen den Verdacht auf Multikollinearität erhärten. Trotz allem reicht das nicht aus, um dadurch die Multikollinearität für das gesamte multiple lineare Regressionsmodell festzustellen oder zu entkräften. In diesem Zusammenhang kann also trotz niedriger paarweiser Korrelation eine Multikollinearität vorherrschen.[107]

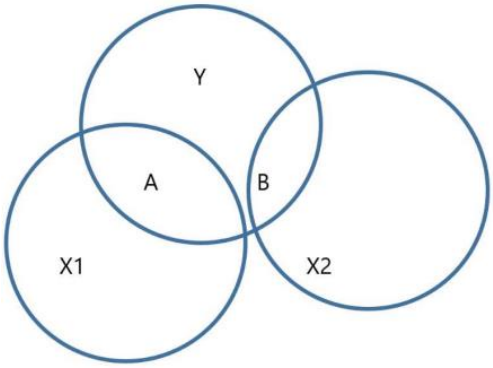

Abbildung 8 – Venn-Diagramm: Keine Multikollinearität (Winke,2020,S.37)

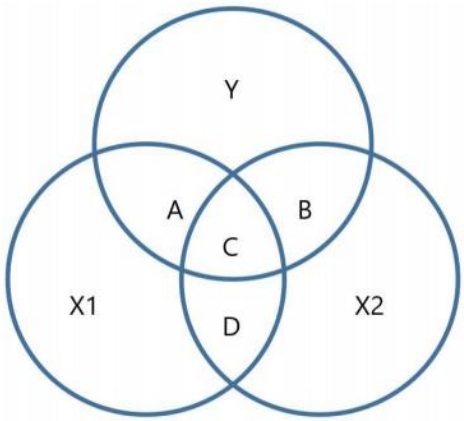

Abbildung 9 - Venn-Diagramm: Geringe Multikollinearität (Winke,2020,S.36)

[107] Vgl. Backhaus, 2021, S. 122f.

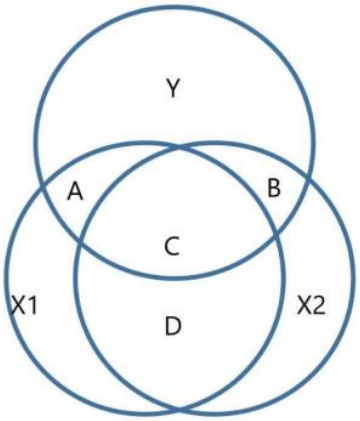

Abbildung 10 - Venn-Diagramm: Hohe Multikollinearität (Winke,2020,S.37)

In der Abbildung 8 sieht man den Fall, dass keine Multikollinearität zwischen den unabhängigen Variablen X_1 und X_2 existiert. Somit existiert also keine Überlappung der jeweiligen Streuung von X_1 und X_2. Die Werte der jeweiligen Regressoren sind klar zuzuordnen. Hierbei spricht man auch von einer Orthogonalität.[108]

In der Abbildung 9 existiert eine geringe Multikollinearität zwischen den Regressoren X_1 und X_2 , die generell bei jeder multiplen linearen Regression vorhanden ist.[109]

In der Abbildung 10 ist eine hohe Multikollinearität zwischen den Regressoren X_1 und X_2 und somit die Tatsache gegeben, dass die Werte der jeweiligen Regressoren schwer zuzuordnen sind. Die jeweilige Streuung der Regressoren überlappt sich so stark, dass die Informationswerte mehrfach vorhanden sind.[110]

Hierbei werden Residuendiagramme zur Feststellung von Multikollinearität nicht verwendet, da die linearen Zusammenhänge zwischen den Regressoren so nicht darstellbar sind.[111] Aus diesem Grund wird zu jedem Regressor X_j ein multiples lineares Regressionsmodell erstellt, wobei der betroffene Regressor als abhängige Variable Y festgelegt wird und die restlichen Regressoren Einfluss auf Y nehmen.[112]

[108] Vgl. Urban, 2011, S. 225.
[109] Vgl. Winke, 2020, S. 36.
[110] Vgl. Winke, 2020, S. 36f.
[111] Vgl. Urban, 2011, S. 230.
[112] Vgl. Urban, 2011, S. 231.

Als Maß für den Grad der Multikollinearität verwendet man den Varianzinflationsfaktor VIF_j.[113]

$$Sei\ T_j = 1 - {R_j}^2\ die\ Toleranz\ der\ Regressoren\ mit\ j \in \mathbb{N}\backslash\{0\}\ gegeben.$$
$$Sei\ zudem\ der\ quadrierte\ multiple\ Korrelationskoeffizient\ {R_j}^2\ gegeben.$$
$$Dann\ gilt\ für\ den\ Varianzinflationsfaktor\ folgendes\ ^{114115}:$$

$$VIF_j = \frac{1}{1 - {R_j}^2}$$

$$VIF_j = 1 : Keine\ Multikollinearität^{116}$$
$$VIF_j > 10 : Schwerwiegende\ Multikollinearität^{117}$$

In R nutzt man den Befehl $vif()$ aus der Bibliothek car, um den Varianzinflationsfaktor VIF_j zu bestimmen.[118]

Zusammenfassend ist festzuhalten, dass im Allgemeinen die Multikollinearität mit der Anzahl der Regressoren wächst. Das heißt, je mehr Regressoren in einem multiplen linearen Regressionsmodell genutzt werden, desto stärker ist eine potenzielle Multikollinearität vorhanden. Generell bleibt aber die Unverzerrtheit bei der OLS-Schätzung trotz Multikollinearität bestehen.[119]
Hohe Multikollinearität verursacht aber sowohl eine hohe Varianz als auch unzuverlässige Schätzungen der Regressionskoeffizienten. Denn aufgrund der hohen absoluten Varianz werden die geschätzten Regressionskoeffizienten instabiler.[120]

[113] Vgl. Backhaus, 2021, S. 123.
[114] Vgl. Backhaus, 2021, S. 123.
[115] Vgl. Winke, 2020, S. 38.
[116] Vgl. Winke, 2020, S. 38.
[117] Vgl. Winke, 2020, S. 38.
[118] Vgl. Wollschläger, 2014, S. 202.
[119] Vgl. Urban, 2011, S. 227.
[120] Vgl. Urban, 2011, S. 228.

„Je größer die Varianz, desto größer ist seine Instabilität und umso größer ist seine Sensibilität hinsichtlich minimalster Veränderungen in den Ausgangsbedingungen der Regressionsanalyse.“(Urban,2011,S.228)

Des Weiteren entstehen dadurch auch unzuverlässige Ergebnisse beim t-Test.[121] Gleichzeitig wird aber der F-Test durch die Multikollinearität nicht negativ beeinflusst.[122]

2.4 Güte der Regression

In diesem Abschnitt widmen wir uns der Qualität, auch Güte genannt, unserer geschätzten Regressionsfunktion. Das heißt, es soll ermittelt werden, wie gut die Schätzungsgerade eine Anpassung zu den tatsächlichen Werten erreicht hat. Die Beurteilung davon erfolgt über Gütekriterien.[123]

Sei eine geschätzte Regressionsfunktion $\hat{y}$ gegeben. Dann gilt für die Gütekriterien folgendes[124] :

$$G_1 : P - Wert$$

$$G_2 : Standardfehler$$

$$G_3 : Bestimmtheitsmaß$$

$$G_4 : Korrigiertes\ Bestimmtheitsmaß$$

$$G_5 : F - Statistik$$

[121] Vgl. Urban, 2011, S. 229.
[122] Vgl. Urban, 2011, S. 230.
[123] Vgl. Backhaus, 2021, S. 85.
[124] Vgl. Backhaus, 2021, S. 85.

2.4.1 G_1 : P -Wert

Die allgemeine Intention beinhaltet, dass man durch die Werte der gegebenen Zufallsstichprobe auf die unbekannten Werte der Grundgesamtheit (Population) schließen kann. Das heißt, Hypothesentests, inklusive dem p-Wert, gelten als Werkzeuge um auf die Werte der Grundgesamtheit schließen zu können.[125]

$$Sei\ H_0\ die\ Nullhypothese, die\ die\ Standardbehauptung\ wiedergibt.$$
$$Sei\ H_1\ die\ Alternativhypothese, die\ die\ Gegenbehauptung\ darstellt.$$
$$Sei\ \theta\ der\ wahre\ Populationsparameter\ und\ sei\ \hat{\theta}\ der\ geschätzte\ Parameter.$$
$$So\ gilt\ ohne\ Beschränkung\ der\ Allgemeinheit\ (o.B.d.A.)\ für\ den\ Hypothesentest\ folgendes[126]:$$

$$Es\ wird\ folgende\ Nullhypothese\ behauptet: H_0: \hat{\theta} = \theta$$
$$Es\ wird\ folgende\ Alternativhyptohese\ behauptet: H_1: \hat{\theta} \neq \theta$$

Mithilfe der Prüfgröße T und den Ablehnungsbereich K_α prüft man, welche Hypothese der Wahrheit entspricht.[127] Das Signifikanzniveau α gibt hierbei die Irrtumswahrscheinlichkeit wieder.[128] In der Regel legt man dafür das Signifikanzniveau $\alpha = 0.05$ also 5% fest, welches als generelle Konvention gilt.[129]

Tabelle 1 - Fehlerarten beim Hypothesentest (Hedderich,2020,S.465)

Entscheidung des Tests	Wirklichkeit	
	H_0 wahr	H_0 falsch
H_0 abgelehnt (H_A angenommen)	$P(T \in K_\alpha \mid H_0) \leq \alpha$ Fehler 1. Art	$P(T \in K_\alpha \mid H_A)$ richtige Entscheidung
H_0 beibehalten (H_A abgelehnt)	$P(T \notin K_\alpha \mid H_0) \geq 1 - \alpha$ richtige Entscheidung	$P(T \notin K_\alpha \mid H_A)$ Fehler 2. Art

[125] Vgl. Sauer, 2019, S. 268f.
[126] Vgl. Shardt, 2021, S. 65.
[127] Vgl. Hedderich, 2020, S. 465f.
[128] Vgl. Shardt, 2021, S. 66.
[129] Vgl. Sauer, 2019, S. 272.

Das heißt, wenn durch den Signifikanztest die Nullhypothese H_0 abgelehnt wurde, aber diese in der Realität wahr ist, so entsteht ein Fehler 1. Art. Wenn durch den Signifikanztest die Nullhypothese H_0 angenommen wurde, aber diese in der Realität falsch ist, so entsteht ein Fehler 2. Art.[130] H_1 wird auch als H_A bezeichnet.[131]

Um aber nun tatsächlich einen Signifikanztest durchzuführen, nutzt man den p-Wert, wobei der Hypothesentest als Entscheidungshilfe dient.[132] Infolgedessen stellt der p-Wert eine Maßzahl dar und gibt in der Anwendung die Wahrscheinlichkeit wieder, wie effektiv die Daten zu den Hypothesen passen. Das heißt, man testet und vergleicht mithilfe des Siginifikanztests den p-Wert mit dem Signifikanzniveau $\alpha = 0.05$. Für das Werteintervall des p-Wertes gilt $P \in [0,1]$. Zusätzlich repräsentiert H_1 die Hypothese, die man bewahrheitet sehen möchte.[133]

Sei ein rechtsseitiger Ablehnungsbereich gegeben. Seien der $P-Wert$ mit P, die Prüfgröße mit X und die Verteilungsfunktion von X bei Gültigkeit der Nullhypothese mit $F(X|H_0)$ gegeben. Dann gilt für den $p-Wert$ folgendes[134]:

$$P(X) = 1 - F(X|H_0)$$

Insgesamt gilt für die Handhabung des Signifikanztests folgendes[135] :

$$Fall\ 1)\ \ p-Wert \leq 0.05\ \Rightarrow Ablehnung\ von\ H_0\ und\ Annahme\ von\ H_1$$

$$Fall\ 2)\ p-Wert > 0.05 \Rightarrow Ablehnung\ von\ H_1\ und\ Annahme\ von\ H_0$$

„Je kleiner p, desto schlechter passen die Daten zur getesteten Hypothese, der sog. Nullhypothese“ (Sauer,2019,S.270).

[130] Vgl. Shardt, 2021, S. 66.
[131] Vgl. Hedderich, 2020, S. 460.
[132] Vgl. Hedderich, 2020, S. 461.
[133] Vgl. Sauer, 2019, S. 270.
[134] Vgl. Hedderich, 2020, S. 461.
[135] Vgl. Hedderich, 2020, S. 461.

Das heißt, wenn der erste Fall zutrifft, so spricht man davon, dass eine statistische Signifikanz vorherrscht. Dieses wird durch Schranken und eine Sternsymbolik weiter spezialisiert. Der zweite Fall repräsentiert eine nicht-statistische Signifikanz.[136]

„ $Für\ P \leq 0{,}05\ gibt\ man\ anhand\ der\ kritischen\ 5\% - ,1\% - und\ 0{,}1\% - Schranken\ an, zwischen\ welchen\ Grenzen\ P\ liegt\ und\ kennzeichnet\ statistisch\ signifikante\ Befunde\ durch\ die\ dreistufige\ Sternsymbolik$:
$[*]0{,}05 \geq P > 0{,}01 \quad [**]0{,}01 \geq P > 0{,}001 \quad [***]\ P \leq 0{,}001$ "
(Hedderich,2020,S.461)

2.4.2 G_2 : Standardfehler

Als Standardfehler bezeichnet man die Streuung der untersuchenden Zufallsstichprobe.[137] Dieser vermittelt den durchschnittlichen Fehler der Zufallsstichprobe zur geschätzten Regressionsfunktion.[138] Zudem dient es als Hilfsmittel für die statistische Präzision.[139] Insbesondere erfüllt der Standardfehler eine essentielle Rolle zur Berechnung des p-Wertes.[140]

$Seien\ die\ nicht\ erklärte\ Streuung\ SSR, die\ Anzahl\ der\ Beobachtungen\ N, die\ Anzahl\ der\ Regressoren\ J\ und\ die\ Anzahl\ der\ Freiheitsgrade\ N - J - 1\ , abgekürzt\ df, gegeben.$
$Sei\ zudem\ i, N, J \in \mathbb{N}\backslash\{0\}\ \ gegeben.\ Dann\ gilt\ für\ den\ Standardfehler\ SE$
$(englisch:\ standard\ error\ of\ the\ estimate)\ folgendes$ [141] [142] :

$$SE = \sqrt{\frac{SSR}{N-J-1}} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-J-1}}$$

[136] Vgl. Hedderich, 2020, S. 461.
[137] Vgl. Sauer, 2019, S. 273.
[138] Vgl. Winke, 2020, S. 20.
[139] Vgl. Backhaus, 2021, S. 85.
[140] Vgl. Sauer, 2019, S. 273.
[141] Vgl. Backhaus, 2021, S. 85f.
[142] Vgl. Backhaus, 2021, S. 88.

Die geschätzte Regressionsfunktion gilt am Idealsten, je kleiner SE ist.[143] Dementsprechend erhöht sich dann auch die Präzision.[144] Kurzum gilt, dass wenn alle Werte der Zufallsstichprobe auf der geschätzten Regressionsfunktion lokalisiert wären, dann würde folgendes gelten: $SE = 0$.[145] Zudem gilt auch, dass je größer die Zufallsstichprobe ist, desto kleiner wird der Standardfehler.[146]
Wichtig sei noch festzuhalten, dass SSR auch als die Kleinst-Quadrate-Methode (auch bekannt als die Methode der kleinsten Quadrate) bzw. als die Summe der quadrierten Residuen (englisch: *sum of squared residuals*) verstanden wird. Zudem gelten die Residuen ε_i unterhalb der geschätzten Regressionsfunktion als negative Residuen.[147]

$$Sei\ i, N \in \mathbb{N}\backslash\{0\}\ gegeben. Dann\ gilt\ für\ die\ Kleinst-\ Quadrate-Methode\ folgendes$$ [148][149] :

$$SSR = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} {\varepsilon_i}^2 \rightarrow min\,!$$

Das heißt, je kleiner die Residuen sind, desto besser ist die geschätzte Regressionsfunktion. Ebenso wird das Quadrieren der Residuen damit begründet, um der Aufhebung der positiven und negativen Residuen entgegenzuwirken.[150]

Abschließend ist festzuhalten, dass der Grund für große Standardfehler unter anderem in der Kollinearität bzw. Multikollinearität der unabhängigen Variablen liegen kann.[151]

[143] Vgl. Winke, 2020, S. 20f.
[144] Vgl. Backhaus, 2021, S. 85.
[145] Vgl. Stoetzer, 2017, S. 43.
[146] Vgl. Urban, 2011, S. 153.
[147] Vgl. Backhaus, 2021, S. 78.
[148] Vgl. Backhaus, 2021, S. 78.
[149] Vgl. Backhaus, 2021, S. 88.
[150] Vgl. Toutenburg, 2008, S. 174.
[151] Vgl. Urban, 2011, S. 108.

2.4.3 G_3 : Bestimmtheitsmaß

Das Bestimmtheitsmaß, welches als R^2 bezeichnet wird, basiert auf der Korrelation der Zufallsstichprobe Y und der geschätzten Regressionsfunktion $\hat{Y}$.[152] Zudem ist das Bestimmtheitsmaß auch als Determinationskoeffizient bekannt.[153]
Das heißt, R^2 beschreibt den Anteil der Gesamtstreuung, der durch die geschätzte Regressionsfunktion wiedergegeben wird. Der Wertebereich von R^2 liegt bei $R^2 \in [0,1]$.[154] Insgesamt setzt R^2 die Gesamtstreuung, die erklärte Streuung und die nicht erklärte Streuung ins Verhältnis.[155]

i) Die Gesamtstreuung, SST (englisch: *total sum of squares*) abgekürzt, gibt den Abstand der beobachteten Werte zum Mittelwert der beobachteten Werte wieder.[156][157]

$Für\ die\ Gesamtstreuung\ SST\ mit\ N, i \in \mathbb{N}\backslash\{0\}\ \ gilt\ folgendes$ [158]:

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

[152] Vgl. Backhaus, 2021, S. 86.
[153] Vgl. Sauer, 2019, S. 325.
[154] Vgl. Stoetzer, 2017, S. 40.
[155] Vgl. Backhaus, 2021, S. 88.
[156] Vgl. Backhaus, 2021, S. 88.
[157] Vgl. Field, 2012, S. 250f.
[158] Vgl. Backhaus, 2021, S. 88.

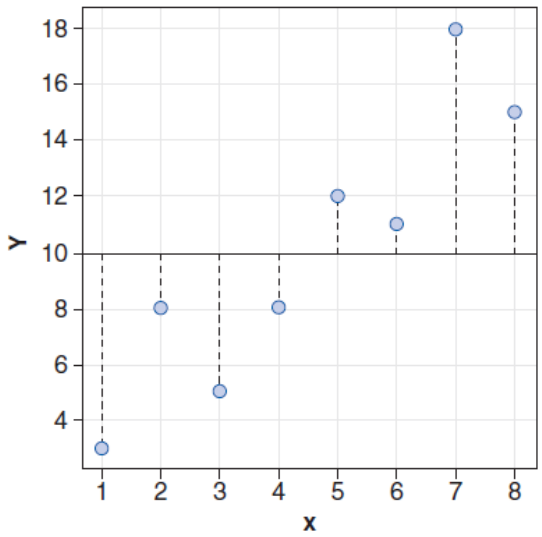

Abbildung 11 - Diagramm: Gesamtstreuung SST (Field,2012, S.251)

ii) Die erklärte Streuung, SSE (englisch: *explained sum of squares*) abgekürzt, beschreibt den Abstand der geschätzten Werte zum Mittelwert der beobachteten Werte. Es wird auch als SSM (englisch: *model sum of squares*) bezeichnet.[159][160]

$$Für\ die\ erklärte\ Streuung\ SSE\ mit\ N, i \in \mathbb{N}\backslash\{0\}\ gilt\ folgendes^{161}:$$

$$SSE = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

[159] Vgl. Field, 2012, S. 251.
[160] Vgl. Backhaus, 2021, S. 88.
[161] Vgl. Backhaus, 2021, S. 88.

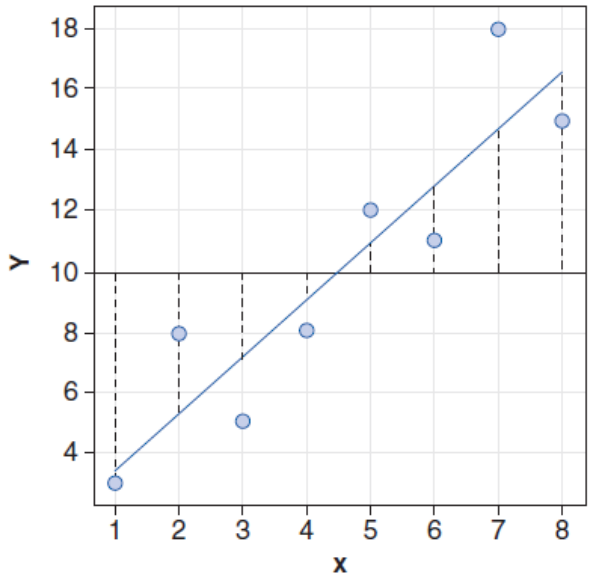

Abbildung 12 - Diagramm: Erklärte Streuung SSE (Field,2012,S.251)

iii) Die nicht-erklärte Streuung SSR gibt den Abstand der beobachteten Werte zu den geschätzten Werten wieder.[162] (Siehe Definition in Kapitel 2.4.2)

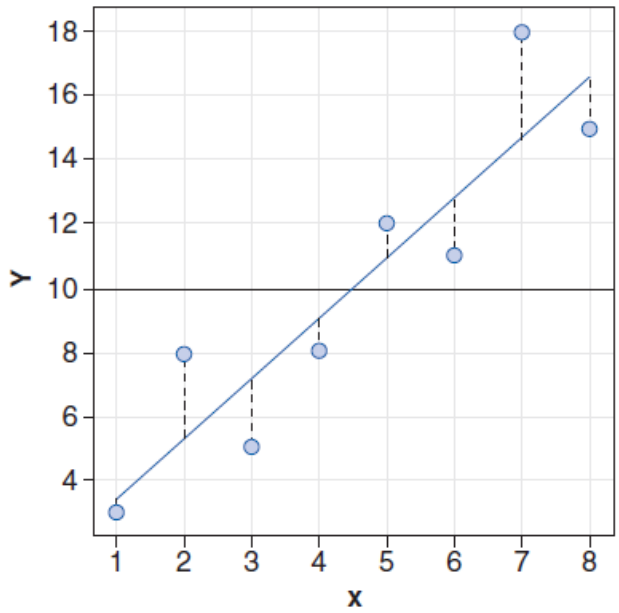

Abbildung 13 - Diagramm: Nicht erklärte Streuung SSR (Field, 2012,S.251)

[162] Vgl. Field, 2012, S. 251.

$$Sei\ das\ Bestimmtheitsmaß\ R^2\ mit\ R^2 \in [0,1]\ \ gegeben.$$

$$Sei\ zudem\ \ SST \neq 0\ gegeben, dann\ gilt\ für\ R^2\ folgendes^{163}:$$

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{Erklärte\ Streuung}{Gesamtstreuung}$$

Je näher sich R^2 an der eins befindet, desto besser ist das Regressionsmodell. Demnach ist das Regressionsmodell umso realitätsferner, je näher sich R^2 an der Null befindet.[164] Der R^2-Wert muss stets mit 100 multipliziert werden, um den prozentualen Anteil der Anpassung zu erhalten.[165]

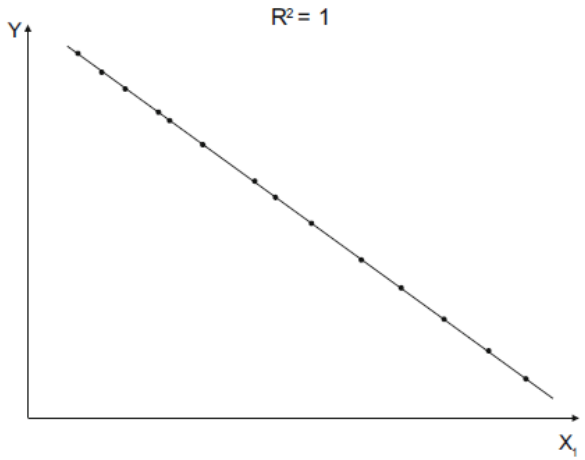

Abbildung 14 - Diagramm: Bestimmtheitsmaß mit dem Wert 1 (Stoetzer,2017,S.41)

In der Abbildung 14 kommt die perfekte Erklärung der Zielvariable Y mit dem Wert $R^2 = 1$ zum Ausdruck. Das heißt, alle beobachteten Werte können hundertprozentig durch die Regressoren erklärt werden.[166]

Zusätzlich besitzt das Bestimmtheitsmaß die negative Eigenschaft, dass durch stetiges Hinzufügen von Regressoren auch das Bestimmtheitsmaß dementsprechend mitwächst und nicht abnimmt.[167] Durch den Anstieg von R^2 wird aber die Qualität

[163] Vgl. Backhaus, 2021, S. 88.
[164] Vgl. Stoetzer, 2017, S. 40.
[165] Vgl. Field, 2012, S. 250.
[166] Vgl. Stoetzer, 2017, S. 40.
[167] Vgl. Winke, 2020, S. 18.

des Regressionsmodells nicht zwangsweise verbessert. Zudem kann es auch zu verzerrten Schätzungen aufgrund von Multikollinearität führen.[168]
Außerdem ist festzuhalten, dass das Bestimmtheitsmaß nur eine allgemeine Aussage über die Realitätsnähe des Regressionsmodells wiedergibt. Üblicherweise ist bei Zeitreihen ein sehr hohes Bestimmtheitsmaß festzustellen, was auf die zeitlichen Trends zurückzuführen ist. Insgesamt sollte man daher dem Bestimmtheitsmaß nicht zu viel Vertrauen schenken, da es im Endeffekt keine zuverlässige Aussage zur Qualität eines Regressionsmodells wiedergibt.[169] Daher werden in der Praxis stets der Standardfehler und der F-Test zusätzlich zum Bestimmtheitsmaß mitbestimmt.[170]

2.4.4 G_4 : Korrigiertes Bestimmtheitsmaß

Das korrigierte Bestimmtheitsmaß $R_{korr}{}^2$,welches auch als adjusted R^2 bezeichnet wird, gleicht die negative Eigenschaft des Bestimmtheitsmaßes R^2 wieder aus. [171][172]

„Im Gegensatz zum unkorrigierten Bestimmtheitsmaß, ermöglicht es den Vergleich von Ergebnissen, selbst bei unterschiedlicher Variablenzahl oder Stichprobengröße. Eine sehr große Differenz zwischen R^2 und $R_{korr}{}^2$ ist ein Indiz für Overfitting.“ (Winke,2020,S.18)

Somit wächst das korrigierte Bestimmtheitsmaß trotz stetiger Erweiterung von Regressoren nicht an. Es wird sogar kleiner.[173]

Sei das Bestimmtheitsmaß R^2 und die Anzahl der Regressoren J gegeben. Dann gilt für das korrigierte Bestimmtheitsmaß $R_{korr}{}^2$ folgendes[174] :

$$R_{korr}{}^2 = 1 - \frac{N-1}{N-J-1}(1-R^2)$$

[168] Vgl. Backhaus, 2021, S. 93.
[169] Vgl. Stoetzer, 2017, S. 42.
[170] Vgl. Urban, 2011, S. 65.
[171] Vgl. Winke, 2020, S. 18.
[172] Vgl. Stoetzer, 2017, S. 42.
[173] Vgl. Backhaus, 2021, S. 94.
[174] Backhaus, 2021, S. 94.

2.4.5 G_5 : F-Statistik

Das generelle Ziel ist es durch die Zufallsstichprobe auf die tatsächlichen Werte in der Grundgesamtheit zu schließen. Dementsprechend repräsentieren die geschätzten Regressionskoeffizienten eine Annäherung zu den wahren Werten in der Grundgesamtheit. Die Wahrscheinlichkeit bezüglich dieser Annäherung wird mit dem F-Test bestimmt.[175]

„Der F-Test prüft also, ob das Modell einen Beitrag zur Erklärung der AV leistet." (Winke,2020,S.19)

Hierbei wird der Hypothesentest benutzt, um die Daseinsberechtigung eines Regressionsmodells zu kontrollieren. Wenn die Nullhypothese abgelehnt wird, dann ist mindestens ein Regressor in der Lage die abhängige Variable Y zu erklären.[176] Für die Nullhypothese gilt daher folgendes[177] :

$$H_0 : Alle\ geschätzten\ Regressionskoeffizienten\ sind\ gleich\ Null.$$
$$H_1 : Mindestens\ ein\ Regressionskoeffizient\ ist\ ungleich\ Null.$$

Der empirische F-Wert wird diesbezüglich berechnet, um einen p-Wert zu bestimmen. Somit kann dann ein Signifikanztest durchgeführt werden.[178]

$$Seien\ das\ Bestimmtheitsmaß\ R^2\ und\ der\ empirische\ F-Wert\ als\ F_{emp}\ gegeben.\ Dann\ gilt\ für\ F_{emp}\ folgendes[179]:$$

$$F_{emp} = \frac{R^2/J}{(1-R^2)/(N-J-1)}$$

[175] Vgl. Stoetzer, 2017, S. 43.
[176] Vgl. Backhaus, 2021, S. 90.
[177] Vgl. Backhaus, 2021, S. 90.
[178] Vgl. Backhaus, 2021, S. 91.
[179] Backhaus, 2021, S. 91.

Insbesondere gilt hier die Regel, dass je größer F_{emp} ist desto kleiner wird der p-Wert. Das heißt, wenn $P < \alpha$ gilt, so wird H_0 abgelehnt. Somit ist die geschätzte Regressionsgerade statistisch signifikant.[180]

2.5 T-Test der Regressionskoeffizienten

Die Grundlage für den T-Test der Regressionskoeffizienten liegt in der statistischen Signifikanz der geschätzten Regressionsfunktion, die durch den F-Test global festgestellt wurde.[181] Dementsprechend gilt es nun zu ermitteln, welchen Erklärungsbeitrag die einzelnen Regressionskoeffizienten für die abhängige Variable Y liefern. Hierzu werden die Regressionskoeffizienten einzeln überprüft.[182]

Der t-Test beinhaltet dasselbe Vorgehen wie der F-Test, wobei ein Signifikanztest wieder vorgenommen wird. Das Ziel ist es, jene Regressoren aus dem Regressionsmodell zu entfernen, die keinen unterstützenden Einfluss auf die abhängige Variable nehmen.[183] Das heißt, wenn der P-Wert des jeweiligen Regressionskoeffizienten kleiner als das Signifikanzniveau $\alpha = 0.05$ ist, so gilt statistische Signifikanz. Somit leistet dieser Regressionskoeffizient einen Erklärungsbeitrag zum Regressionsmodell.[184]

$$Sei\ j \in \mathbb{N}\backslash\{0\}\ gegeben.\ Dann\ gilt\ f\text{ür}\ den\ Signifikanztest\ beim\ t-Test\ der\ Regressionskoeffizienten\ folgendes^{185}:$$

$$H_0 : Der\ einzelne\ Regressionskoeffizient\ \beta_j\ ist\ gleich\ Null.$$
$$H_1 : Der\ einzelne\ Regressionskoeffizient\ \beta_j\ ist\ ungleich\ Null.$$

[180] Vgl. Backhaus, 2021, S. 91f.
[181] Vgl. Backhaus, 2021, S. 95.
[182] Vgl. Winke, 2020, S. 21.
[183] Vgl. Winke, 2020, S. 21.
[184] Vgl. Fahrmeir, 2016, S. 460.
[185] Vgl. Backhaus, 2021, S. 96.

Der T-Test wird verwendet, um das Testen von einzelnen Regressionskoeffizienten durchzuführen.[186] Hierbei wird ein zweiseitiger T-Test mit $\alpha = 0.05$ genutzt.[187]

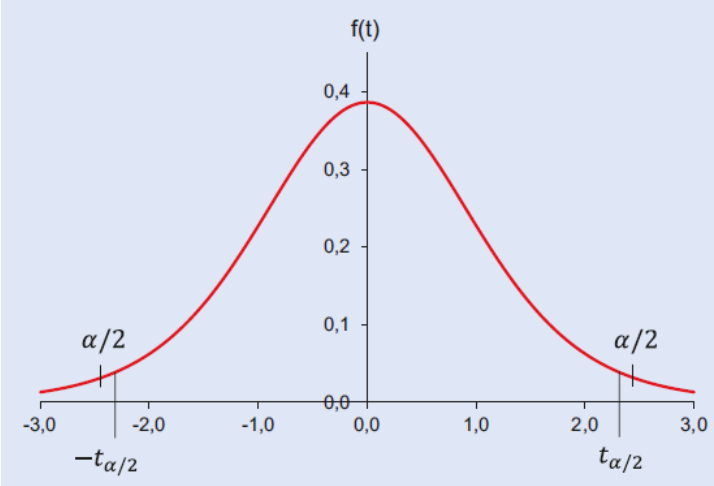

Abbildung 15 -Diagramm: t-Verteilung für zweiseitigen t-Test (Backhaus,2021,S.97)

Sei ein geschätzter Regressionskoeffizient b_j mit $j \in \{0, \dots, n\}$ mit $n \in \mathbb{N}$ gegeben. Sei zudem der Standardfehler eines geschätzten Regressionskoeffizienten in einer multiplen Regression mit $SE(b_j)$ gegeben. Seien der Standardfehler SE und die Standardabweichung des Regressors $s(x_j)$ gegeben. Sei zudem das multiple Bestimmtheitsmaß ${R_j}^2$ gegeben. Dann gilt für den empirischen $t-Wert$ also t_{emp} folgendes[188]:

$$t_{emp} = \frac{b_j}{SE(b_j)} \; mit \; SE(b_j) = \frac{SE}{s(x_j)\sqrt{N-1}\,\sqrt{1-{R_j}^2}}$$

Im Gegensatz der hier aufgeführten Punktschätzung existiert noch eine Intervallschätzung. Mit dieser ist es zusätzlich noch möglich, den Bereich des wahren Regressionskoeffizienten mithilfe eines Konfidenzintervalls und einer Konfidenzwahrscheinlichkeit einzugrenzen.[189]

[186] Vgl. Backhaus, 2021, S. 96f.
[187] Vgl. Backhaus, 2021, S. 97.
[188] Vgl. Backhaus, 2021, S. 96f.
[189] Vgl. Backhaus, 2021, S. 99.

2.6 Ausreißer

Im Allgemeinen werden Werte als Ausreißer bezeichnet, die sich stark von den anderen Beobachtungen unterscheiden. Durch Ausreißer kann die Regressionsschätzung stark verzerrt werden, was darin begründet liegt, dass die Kleinst-Quadrate-Methode in der OLS-Schätzung genutzt wird.[190]

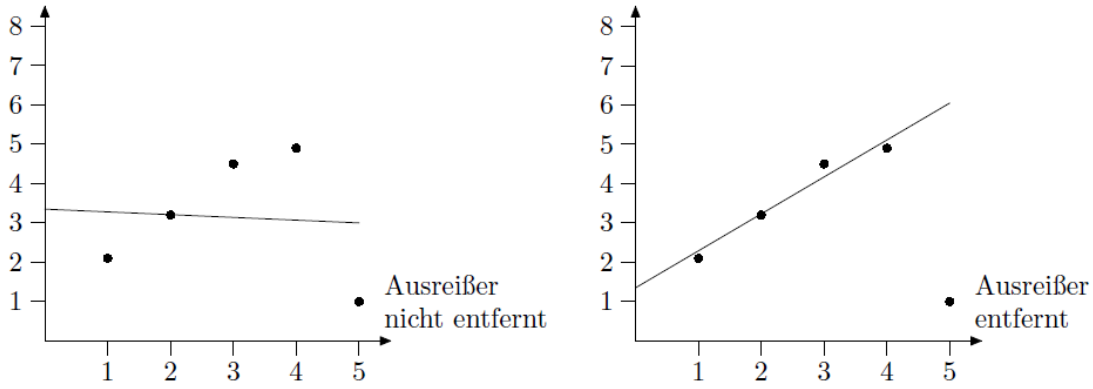

Abbildung 16 - Diagramm: OLS-Schätzung und Ausreißer (Toutenburg,2008,S.173)

Ausreißer können durch Messfehler, zufällige Fehler oder ungewöhnliche Ereignisse entstehen. Zudem ist anzumerken, dass nicht jeder Ausreißer einen stark verzerrenden Einfluss auf die OLS-Schätzung ausübt.[191] Durch Streudiagramme oder numerische Vorgehensweisen ist es möglich, Ausreißer leichter zu erkennen.[192] Generell wird durch die Hebelwirkung eines Ausreißers die OLS-Schätzung verzerrt, welches vor allem durch die Lage des Ausreißers hervorgerufen wird.[193]
Für die Ermittlung der Hebelwirkung einer Beobachtung i wird der Hut-Wert in der Regressionsanalyse verwendet, der mit $h_i \equiv h_{ii}$ bezeichnet wird und wo $i \in \{1, \dots, N\}$ gilt.[194]

$$Sei\ der\ Hut-Wert\ h_i \equiv h_{ii}\ mit\ i, N \in \mathbb{N}\backslash\{0\}\ gegeben.\ Seien\ zudem\ {}^{1}/_{N} \leq h_i \leq 1\ , die\ Standardabweichung\ des\ Regressors\ s_x\ und\ die\ quadrierte\ Entferung\ (x_i - \bar{x})^2\ gegeben.\ Dann\ gilt\ f\ddot{u}r\ die$$

[190] Vgl. Winke, 2020, S. 40.
[191] Vgl. Backhaus, 2021, S. 125.
[192] Vgl. Backhaus, 2021, S. 127.
[193] Vgl. Backhaus, 2021, S. 129.
[194] Vgl. Backhaus, 2021, S. 130.

$Hebelwirkung\ in\ einer\ linearen\ Regression\ folgendes$[195]:

$$h_i = \frac{1}{N} + \frac{1}{N-1}\,\frac{(x_i - \bar{x})^2}{s_x}$$

„Je weiter eine Beobachtung auf der x-Achse vom Mittelwert $\bar{x}$ entfernt ist, desto größer ist ihr Einfluss auf die Steigung der Regressionsgeraden. Dieser Effekt wird als Leverage (Hebelwirkung) bezeichnet.“(Backhaus,2021,S.129)

Somit haben jene Ausreißer eine große Hebelwirkung, welche die größte Distanz zum Mittelwert $\bar{x}$ aufweisen. Somit gilt die Hebelwirkung auch als Funktion der x-Werte. Zusätzlich wird der Einfluss eines Ausreißers auch durch dessen Größe also den y-Wert geprägt.[196] Im Allgemeinen gilt für den Einfluss eines Ausreißers folgendes[197] :

$$Einfluss = Größe \times Hebelwirkung$$

Konkret sind Ausreißer mit einem großen Einfluss schädlich für die OLS-Schätzung.[198] In diesem Zusammenhang nimmt die Hebelwirkung stets ab, je größer die Zufallsstichprobe wird.[199] In der Praxis mit R wird ein großer Hebelwert identifiziert, indem dieser numerisch das Doppelte bis Dreifache des Mittelwertes aller Hebelwerte darstellt. Dazu wird unter anderem die Funktion $hatvalues()$ verwendet.[200]
Des Weiteren nutzt man noch die Cook’sche Distanz, welches das gebräuchlichste Maß für den Einfluss der Ausreißer darstellt.[201]

„Die Cook-Distanzen sind die repräsentativste Einzelfallstatistik für den „Overall Fit“. Es ist ein Maß für den Einfluss den ein einzelner Fall auf das gesamte Modell

[195] Backhaus, 2021, S. 130.
[196] Vgl. Backhaus, 2021, S. 129.
[197] Backhaus, 2021, S. 129.
[198] Vgl. Backhaus, 2021, S. 129.
[199] Vgl. Backhaus, 2021, S. 130.
[200] Vgl. Wollschläger, 2014, S. 197.
[201] Vgl. Backhaus, 2021, S. 133.

nimmt. Es misst, wie stark sich die Gerade ändern würde, wenn der Fall ausgeschlossen wird.“(Winke,2020,S.41)

$$Seien\ i \in \mathbb{N}\backslash\{0\}\ und\ die\ studentisierten\ Residuen\ t_i\ gegeben.$$
$$Dann\ gilt\ f\ddot{u}r\ die\ Cook'scheDistanz\ D_i\ folgendes^{202}:$$

$$D_i = \frac{{t_i}^2}{J+1}\,\frac{h_i}{1-h_i}$$

Wenn $D_i > 1$ gilt, dann handelt es sich um einen schädlichen Ausreißer.[203] In R wird hierzu die Funktion $influence.measures()$ verwendet, um die Cook’sche Distanz zu ermitteln.[204]
Auch wenn festgestellt wurde, dass ein Ausreißer schädlich ist, darf dieser nicht automatisch gelöscht werden. Dazu zählen auch Ausreißer mit einem zufälligen Fehler. Denn nur mit einem konkreten Grund oder Beleg ist die Löschung erlaubt. In diesem Zusammenhang wäre ein Eingabefehler ein konkreter Grund.[205]

„Auch wenn wir Beweise dafür finden, dass der Ausreißer durch ein ungewöhnliches Ereignis außerhalb des Forschungskontextes verursacht wurde (z. B. ein Streik der Gewerkschaft oder ein Stromausfall), sollten wir die Beobachtung eliminieren.“ (Backhaus,2021, S.135)

Falls man doch schädliche Ausreißer löschen möchte, so sollte dies dokumentiert werden. Hierzu sollte man die Ergebnisse der OLS-Schätzung mit und ohne Ausreißer wiedergeben. Insgesamt kann nämlich die automatische Löschung von Ausreißern die Regressionsschätzung verfälschen.[206]

[202] Backhaus, 2021, S. 132f.
[203] Vgl. Backhaus, 2021, S. 133.
[204] Vgl. Wollschläger, 2014, S. 198.
[205] Vgl. Backhaus, 2021, S. 134f.
[206] Vgl. Backhaus, 2021, S. 134.

3 Kurze Darstellung der Regressionsverfahren

In diesem Kapitel werden die Regressionsverfahren vorgestellt, die in der Praxis mit R verwendet werden.

3.1 Mehrdimensionale/multivariate lineare Regression

Unter der mehrdimensionalen bzw. multivariaten Regression versteht man eine Regression, die mehrere abhängige Variablen besitzt. Hierbei ist $Y = (Y_1, \dots, Y_q)$ ein q-dimensionaler Zufallsvektor. Zudem werden die abhängigen Variablen Y_j durch mehrere Regressoren $X_1, \dots, X_p$ erklärt, wobei $q, p, j, i \in \mathbb{N}\backslash\{0\}$ gilt.[207] Die Regressionsbeziehung zwischen X_i und Y_j wird durch die Koeffizientenkombination $(p+1) \times q$ durchgeführt und in der Matrix $\boldsymbol{B}$ zusammengefasst.[208]

Seien die Matrix $\boldsymbol{B}$, *die Koeffizienten* X *und der Störterm* $\boldsymbol{E} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_i)'$ *gegeben. Seien zudem* $i, k \in \mathbb{N}\backslash\{0\}$ *und* $i \neq k$ *gegeben. Sei außerdem* $\boldsymbol{\varepsilon}_i$ *die* $i-te$ *Zeile von* $\boldsymbol{E}$. *Dann gilt für das multivariate Regressionsmodell und für die Eigenschaften des Störterms* $\boldsymbol{E}$ *folgendes*[209] :

$$Y = X\,\boldsymbol{B} + \boldsymbol{E}$$

$$M_1 : E(\boldsymbol{E}) = 0$$
$$M_2 : Cov(\boldsymbol{\varepsilon_i}) = D$$
$$M_3 : Cov(\boldsymbol{\varepsilon}_i\,, \boldsymbol{\varepsilon}_k) = 0$$

Für die Kovarianzen-Matrizen gilt folgendes[210]:

$$Cov(\boldsymbol{\varepsilon}_i) = D = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1q} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \dots & \sigma_{qq} \end{pmatrix}\ ,\ Cov(\boldsymbol{\varepsilon}_i\,, \boldsymbol{\varepsilon}_k) = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

[207] Vgl. Schlittgen, 2013, S. 59.
[208] Vgl. Schlittgen, 2013, S. 59f.
[209] Vgl. Schlittgen, 2013, S. 60.
[210] Vgl. Schlittgen, 2013, S. 60.

In der Praxis mit R wird wie bei einer univariaten Regression vorgegangen. Konkret wird also für jede abhängige Variable jeweils eine multiple lineare Regression mit den gleichen Regressoren durchgeführt.[211] Das heißt, man nutzt die Funktion $lm(\,)$, um die multiplen linearen Regressionsmodelle zu definieren.[212]

3.2 Ridge-Regression

Die Ridge-Regression gehört zu den Shrinkage-Methoden.[213] Das Ziel einer Shrinkage-Methode ist es, die Schätzungen der Koeffizienten gegen Null zu mindern. Dadurch verkleinern sich die Varianz und die Residuen, da die Abstände zwischen den tatsächlichen und geschätzten Werten gesenkt werden. Das Regressionsmodell passt somit besser zu den Beobachtungen.[214]
Bei der Ridge-Regression werden die Regressionskoeffizienten durch eine Beschränkung, auch Penalty genannt, bezüglich derer Größen eingegrenzt. Konkret wird eine Minimierung der Summe aller quadratischen Residuen durch diese Koeffizienten gegen Null bezweckt.[215]

$$Seien\ \lambda \geq 0\ und\ t,i,j,N,p \in \mathbb{N}\backslash\{0\}\ gegeben.\ Dann\ gilt$$
$$f\ddot{u}r\ die\ Ridge-Regression\ folgendes^{216}:$$

$$\hat{\beta}^{ridge} = argmin_{\beta}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2\right\}$$

Durch $\lambda \geq 0$ wird der Umfang der Schrumpfung reguliert. Das heißt, je größer λ ist, desto höher wird der Schrumpfungsumfang. Bei korrelierten Variablen, die eine hohe Varianz aufweisen können, werden λ und t als Größenbeschränkung zur Abmilderung dieses Problems verwendet.[217]

[211] Vgl. Zelterman, 2015, S. 238.
[212] Vgl. Wollschläger, 2014, S. 448.
[213] Vgl. James, 2013, S. 215.
[214] Vgl. James, 2013, S. 214f.
[215] Vgl. Hastie, 2009, S. 61ff.
[216] Hastie, 2009, S. 63.
[217] Vgl. Hastie, 2009, S. 63.

Generell wird die Ridge-Regression bei hoher Anzahl von Regressoren und bei hoher Multikollinearität verwendet.[218] In R wird die Ridge-Regression unter anderem mit der Funktion $lm.ridge()$ aus der Bibliothek $MASS$ durchgeführt.[219]

3.3 LASSO-Regression

Die LASSO-Methode gehört auch zu den Shrinkage-Methoden.[220] Hierbei entstammt die Abkürzung LASSO aus der englischen Formulierung *Least Absolute Shrinkage and Selection Operator*.[221]

$$Seien\ \lambda \geq 0\ und\ t,i,j,N,p \in \mathbb{N}\backslash\{0\}\ gegeben.\ Dann\ gilt$$
$$f\ddot{u}r\ die\ LASSO-Methode\ folgendes\ ^{222}:$$

$$\hat{\beta}^{lasso} = argmin_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

Die LASSO-Methode schrumpft die Regressionskoeffizienten und t gegen Null, wobei diese niemals Null werden.[223] [224]
Zudem werden nur die relevantesten Regressoren in einem Modell betrachtet, da durch die Beschränkung vieler Regressoren die Interpretation erschwert wird. Im Gegensatz dazu nimmt die Ridge-Regression auf alle Regressoren Bezug. Aus diesem Grund sind die Regressionsmodelle der LASSO-Methode besser interpretierbar.[225]
In R kann man unter anderem die Funktionen $cv.glmnet()$ und $glmnet()$ aus der Bibliothek $glmnet$ dafür nutzen.[226]

[218] Vgl. Wollschläger, 2014, S. 204.
[219] Vgl. Wollschläger, 2014, S. 205.
[220] Vgl. Hastie, 2009, S. 68.
[221] Vgl. Izenman, 2008, S. 150.
[222] Hastie, 2009, S. 68.
[223] Vgl. James, 2013, S. 219.
[224] Vgl. Hastie, 2009, S. 69.
[225] Vgl. James, 2013, S. 219.
[226] Vgl. Wollschläger, 2014, S. 205.

3.4 LARS-Regression

Die Abkürzung LARS steht für die englische Formulierung least angle regression. Unter anderem wird mithilfe der LASSO-Methode gearbeitet. Hierbei werden alle möglichen Schätzer durch eine bestimmte Anordnung ermittelt und zusätzlich weniger Rechenzeit benötigt.[227]

$$Seien\ \boldsymbol{X} = \left(X_{\boldsymbol{ij}}\right)\ und\ \boldsymbol{Y} = (Y_1, \dots, Y_n)^T\ eine\ (n \times r) - Matrix\ gegeben.$$

$$Seien\ zudem\ die\ Regressoren\ standardisiert\ und\ es\ gilt\ \mu = 0.$$

$$Sei\ f\ddot{u}r\ die\ Regressoren\ \sum_{i=1}^{n} X_{ij} = 0\ und\ \sum_{i=1}^{n} {X_{ij}}^2 = 1\ mit$$

$$n, r, j, i \in \mathbb{N}\backslash\{0\}\ gegeben.\ Sei\ f\ddot{u}r\ die\ abh\ddot{a}ngige\ Variable\ \sum_{i=1}^{n} Y_i = 0\ \ gegeben.$$

$$Sei\ f\ddot{u}r\ die\ Regressionsgesch\ddot{a}tzung\ \widehat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\beta}\ gegeben, wobei$$

$$\boldsymbol{X}_j = \left(X_{1j}, \dots, X_{nj}\right)^T die\ j - te\ Spalte\ von\ \boldsymbol{X} = (\boldsymbol{X_1}, \dots, \boldsymbol{X_r})\ darstellt.$$

$$Sei\ der\ kovariate\ Vektor\ \boldsymbol{X}_j\ gegeben\ .$$

$$Seien\ zudem\ \boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{\mu}}\ und\ \boldsymbol{X}^T\boldsymbol{r} = \hat{\boldsymbol{c}} = (\hat{c}_1, \dots, \hat{c}_{\boldsymbol{r}})^T gegeben.$$

$$Dann\ gilt\ f\ddot{u}r\ den\ LARS - Algorithmus\ folgendes^{228}:$$

$$1) Deklariere\ \hat{\beta} = 0, so\ dass\ \widehat{\boldsymbol{\mu}} = 0\ und\ \boldsymbol{r} = \boldsymbol{Y}\ gilt. Beginne\ mit\ dem\ aktiven$$

$$Set\ \boldsymbol{A}\ , welches\ eine\ leere\ Teilmenge\ von\ dem\ Set\ \{1,2, \dots, r\}\ ist.$$

$$2) Finde\ den\ Kovariaten\ Vektor\ \boldsymbol{X}_{j1}, der\ am\ Meisten\ mit\ \boldsymbol{r}\ korreliert\ und\ wo$$

$$j_1 = argmax_j\left|\hat{c}_j\right|\ gilt. Das\ neue\ aktive\ Set\ lautet\ \boldsymbol{A} \leftarrow \boldsymbol{A} \cup \{\, j_1 \,\}\ und\ X_{j1}$$

$$wurde\ dem\ Regressionsmodell\ hinzugef\ddot{u}gt.$$

$$3) Verschiebe\ \hat{\beta}_{j1}\ in\ Richtung\ sign\left(\hat{c}_{j1}\right)\ mit\ \delta_{j1} = \epsilon\ sign\left(\hat{c}_{j1}\right)\ bis\ ein\ anderer$$

$$kovariater\ Vektor\ \boldsymbol{X}_{j2}\ dieselbe\ Korrelation\ hat, genauso\ wie\ \boldsymbol{X}_{j1}\ mit\ \boldsymbol{r}\ korreliert.$$

$$Das\ neue\ aktive\ Set\ lautet\ \boldsymbol{A}\ \leftarrow \boldsymbol{A} \cup \{\, j_2 \,\}\ und\ X_{j2}\ wird\ dem$$

$$Regressionsmodell\ hinzugef\ddot{u}gt.$$

[227] Vgl. Efron, 2004, S. 407.
[228] Vgl. Izenman, 2008, S. 152f.

4) $Aktualisiere\ \boldsymbol{r}\ und\ verschiebe\ (\hat{\beta}_{j1}, \hat{\beta}_{j2})\ in\ Richtung\ der\ zusammenwirkenden\ OLS-Ausrichtung\ f\ddot{u}r\ die\ Regression\ von\ \boldsymbol{r}\ bis\ der\ dritte\ kovariate\ Vektor\ \boldsymbol{X}_{j3}\ ungef\ddot{a}hr\ mit\ \boldsymbol{r}\ genauso\ korreliert\ wie\ \boldsymbol{r}\ mit\ den\ ersten\ beiden\ Variablen\ korreliert.\ Das\ neue\ aktive\ Set\ lautet\ \boldsymbol{A} \leftarrow \boldsymbol{A} \cup \{ j_3 \}\ und\ X_{j3}\ wird\ dem\ Regressionsmodell\ hinzugef\ddot{u}gt.$

5) $Nach\ k-LARS\ Schritten\ gilt\ der\ aktuelle\ LARS-Sch\ddot{a}tzer\ \widehat{\boldsymbol{\mu}}_{\boldsymbol{A}},\ wobei\ \boldsymbol{A} = \{j_1, j_2, \dots, j_k\}\ und\ die\ k-gesch\ddot{a}tzten\ Koeffizienten\ \hat{\beta}_{j1}, \hat{\beta}_{j2}, \dots, \hat{\beta}_{jk}\ ungleich\ Null\ sind.\ Das\ Regressionsmodell\ wird\ durch\ X_{j1}, X_{j2}, \dots, X_{jk}\ gebildet\ und\ der\ aktuelle\ Korrelationsvektor\ lautet\ \hat{\boldsymbol{c}} = \boldsymbol{X}^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}).$

6) $Setze\ solange\ die\ Prozedur\ fort\ bis\ alle\ r\ Kovariaten\ dem\ Regressionsmodell\ hinzugef\ddot{u}gt\ wurden\ \ und\ gleichzeitig\ \hat{\boldsymbol{c}} = 0\ gilt.\ Daraus\ folgt\ dann\ die\ OLS-L\ddot{o}sung.$

4 Kurze Darstellung der Prognoseverfahren

In diesem Kapitel wird auf die Zeitreihe im Allgemeinen und dessen Grundlagen Bezug genommen. Zusätzlich werden noch nicht-stationäre Modelle vorgestellt.

4.1 Einführung Zeitreihe

Mithilfe von Zeitreihendaten ist es möglich, zukünftige Prognosen zu erstellen und Entwicklungen vorherzusehen.[229] Zeitreihen beinhalten Beobachtungen, die zeitlich strukturiert sind und die gleiche Realität betrachten.[230]

$Seien\ Zeitpunkte\ t\ mit\ den\ gleichen\ Abst\ddot{a}nden\ und\ t \in \{1, \dots, N\}\ mit\ t, N \in \mathbb{N}\{0\}\ gegeben.\ Die\ Zufallsvariablen\ Y_1, \dots, Y_N\ stellen\ Realisationen$

[229] Vgl. Hirschle, 2021, S. 1.
[230] Vgl. Schlittgen, 2013, S. 309.

$$der\ Beobachtungen\ dar.\ Es\ gilt\ der\ stochastische\ Prozess\ Y_t, wobei\ die\ Zeitreihe\ einen\ Ausschnitt\ von\ Y_t\ darstellt.\ Es\ gilt\ also\ f\ddot{u}r\ die\ Folge\ einer\ Zeitreihe\ folgendes^{231}:$$

$$y_1, \dots, y_N\ bzw.\ auch\ Y_1, \dots, Y_N$$

Insbesondere bezeichnet man Zeitreihen als stationär, wenn die allgemeine Übersicht der Daten keine stetigen Veränderungen wiedergeben.[232]

$$Seien\ ein\ stochastischer\ Prozess\ Y_t, die\ Zeit\ t\ und\ der\ Lag\ \tau\ mit\ t, \tau \in \mathbb{N}\backslash\{0\}\ gegeben.\ Sei\ der\ gleiche\ Erwartungswert\ \mu\ und\ die\ gleiche\ Varianz\ \sigma^2\ f\ddot{u}r\ alle\ Y_t\ gegeben.\ Sei\ zudem\ die\ vom\ Zeitpunkt\ t\ abh\ddot{a}ngige\ Korrelation\ Corr(Y_t, Y_{t+\tau})\ gegeben.\ Dann\ gilt\ Y_t\ als\ (schwach)\ station\ddot{a}r,\ wenn\ folgende\ drei\ Bedingungen\ gelten^{233}:$$

$$1) mittelwertstation\ddot{a}r: E(Y_t) = \mu$$
$$2) varianzstation\ddot{a}r: Var(Y_t) = \sigma^2$$
$$3) kovarianzstation\ddot{a}r: Cov(Y_{t+\tau}, Y_t)$$

Häufig sind Zeitreihen aber nicht stationär, da sie durch regelmäßige Veränderungen beeinflusst werden. Konkret existieren verschiedene Arten von Einflüssen, die in dem klassischen Komponentenmodell zusammengefasst werden.[234]

„(i) … Trend, das ist eine langfristige systematische Veränderung des mittleren Niveaus der Zeitreihe;
(ii) … Konjunkturkomponente, die eine mehrjährige, nicht notwendig regelmäßige Schwankung darstellt;
(iii) … Saison, das ist eine jahreszeitlich bedingte Schwankungskomponente, die sich relativ unverändert jedes Jahr wiederholt;
(iv) … Restkomponente, die die nicht zu erklärenden Einflüsse oder Störungen

[231] Vgl. Schlittgen, 2013, S. 309.
[232] Vgl. Schlittgen, 2001, S. 3.
[233] Vgl. Schlittgen, 2020, S. 16.
[234] Vgl. Schlittgen, 2001, S. 9.

zusammenfaßt."(Schlittgen,2001, S.9)

$$Sei\ die\ Zeitreihe\ x_t\ gegeben.\ Seien\ zudem\ der\ Trend\ m_t\ ,die\ Konjunktur\ k_t\ ,$$
$$die\ Saison\ s_t\ und\ der\ Rest\ u_t\ gegeben.\ Dann\ gilt\ f\text{ür}\ die\ Komponenten$$
$$und\ das\ klassische\ Komponentenmodell\ folgendes^{235}:$$

$$1)Additives\ Modell:$$
$$x_t = m_t + k_t + s_t + u_t \iff Reihe = Trend + Konjunktur + Saison + Rest$$

$$Glatte\ Komponente: g_t = m_t + k_t$$
$$Zyklische\ Komponente: z_t = k_t + s_t$$

$$2)Multiplikatives\ Modell:$$
$$x_t = g_t \cdot s_t \cdot u_t$$

Nur durch die Anwendung eines Verfahrens kann das Komponentenmodell einen effektiven Beitrag leisten. Hierzu wird für die komplette Zeitreihe ein lineares Regressionsmodell verwendet, wobei die Kleinst-Quadrate-Methode dessen Regressionskoeffizienten schätzt. Daraus ergibt sich der Begriff eines globalen Komponentenmodells. Insbesondere gilt es eine Prognose zu erstellen, um eine dauerhafte Entwicklung der Werte vorauszusagen. Hierzu nutzt man den Trend einer Zeitreihe.[236]

4.2 Güte einer Prognose

Um die prognostizierten mit den beobachteten Werten vergleichen zu können, nutzt man das Fehlermaß MSE (englisch für *mean squared error* also mittlerer quadratischer Fehler) und MAPE (englisch für *mean absolute percentage error* also der mittlere absolute prozentuale Fehler). Hierbei geben die Fehlermaße die durchschnittlichen Änderungen zwischen der Prognose und der Gegenwart wieder.[237]

[235] Vgl. Schlittgen, 2001, S. 9ff.
[236] Vgl. Schlittgen, 2001, S. 12.
[237] Vgl. Vogel, 2015, S. 15f.

Seien $n, h, t \in \mathbb{N}\backslash\{0\}$, die Beobachtungswerte in einer zeitlichen Abfolge $y_1, y_2, \dots, y_n$ und die daraus resultierenden Prognosewerte $\hat{y}_{n+1}, \dots, \hat{y}_{n+h}$ gegeben, wobei n die Gegenwart und h die nächsten Zunkunftswerte wiedergeben. Dann gelten für die Güte einer Prognose folgende Fehlermaße[238]:

$$MSE = \frac{1}{h} \sum_{t=n+1}^{n+h} (\hat{y}_t - y_t)^2$$

$$MAPE = \frac{100\%}{h} \sum_{t=n+1}^{n+h} \left| \frac{\hat{y}_t - y_t}{y_t} \right| \; mit \; y_t \neq 0$$

Der Vergleich zwischen den Beobachtungs- und Realitätsdaten erfolgt nur, wenn die Daten aus der Realität bereits vorhanden sind. Somit wäre eine bessere Prognose gegeben, wenn man die Güte einer Prognose bereits vorm Eintreffen der Zukunft bewerten könnte. Hierzu nutzt man einen Anteil der letzten Beobachtungswerte und vergleicht diese in einer Testphase mit den späteren Werten.[239]

Insbesondere wird MSE auch dazu verwendet, die Güte einer Schätzung zu beurteilen, wobei es die Varianz diesbezüglich mitbetrachtet. Hierbei kann man zwei Schätzungen miteinander vergleichen, wobei jene Schätzung mit einem kleineren MSE-Wert eine bessere Anpassung der Daten wiedergibt.[240]

Das MAPE-Fehlermaß ist vor allem für große Werte geeignet und erzielt diesbezüglich die besten Resultate.[241] Das heißt, MAPE gibt den prozentualen Prognosefehler wieder, der folgendermaßen interpretiert wird[242]:

[238] Vogel, 2015, S. 16.
[239] Vgl. Vogel, 2015, S. 16.
[240] Vgl. Fahrmeir, 2016, S. 346f.
[241] Vgl. Wikipedia-Webseite, 2022: URL: https://en.wikipedia.org/wiki/Mean_absolute_percentage_error [zuletzt zugegriffen am 30.06.2022]
[242] Vgl. Moreno, 2013, S. 501.

$MAPE - Wert$	$Interpretation$
< 10	$Hohe\ Prognosegenauigkeit$
$10 - 20$	$Gute\ Prognosegenauigkeit$
$20 - 50$	$Akzeptable\ Prognosegenauigkeit$
> 50	$Ungenaue\ Prognosegenauigkeit$

Zusätzlich zählt das RMSE (*Root Mean Squared Error*) als weitere Prognosegüte.[243] Insbesondere gilt es auch als übliche Prognosegüte.[244]

$Sei\ die\ Prognosegüte\ MSE\ gegeben.\ Dann\ gilt\ für\ die\ Prognosegüte\ RMSE\ folgendes$[245][246]:

$$RMSE = \sqrt{MSE}$$

4.3 Einführung ARIMA-Modell

Im Allgemeinen steht die Abkürzung ARIMA für die englische Formulierung *Auto Regressive Integrated Moving Average Model.*[247]

$Sei\ die\ Prognosevariable\ Y\ , die\ Regressionskoeffizienten\ b_0, b_1, \dots, b_p$
$und\ der\ Fehlerterm\ \varepsilon_t\ gegeben.\ Seien\ zudem\ X_1, X_2, X_3, .., X_p$
$mit\ X_1 = Y_{t-1}, X_2 = Y_{t-2}, X_3 = Y_{t-3}, X_p = Y_{t-p}\ gegeben.$
$Sei\ zudem\ t, p \in \mathbb{N}\backslash\{0\}.\ Dann\ gilt\ für\ das$
$Autoregressive\ Model\ und\ Moving\ Average\ Model\ folgendes$ [248] :

$$1) Autoregressive\ Model: Y_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + \cdots + b_p Y_{t-p} + \varepsilon_t$$

[243] Vgl. Mertens, 2005, S. 254f.
[244] Vgl. Hyndman, 2014, S. 46.
[245] Mertens, 2005, S. 255.
[246] Vgl. Mertens, 2005, S. 376.
[247] Vgl. Hyndman, 2014, S. 225f.
[248] Vgl. Makridakis, 1998, S. 335f.

$$2) Moving\ Average\ Model: Y_t = b_0 + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots + b_q Y_{t-q} + \varepsilon_t$$

Um das Autoregressive Moving Average Model (ARMA) zu definieren, fusioniert man das Autoregressive Model (AR) und das Moving Average Model (MA) miteinander. Das ARMA Modell wird aber nur für stationäre Daten verwendet. Infolgedessen existiert das ARIMA Modell, welches für nicht-stationäre Daten geeignet ist. Zudem ist hier die Differenzierung der Zeitreihe zugelassen.[249]

$$Für\ das\ allgemeine\ nicht-saisonale\ ARIMA\ Modell\ also\ ARIMA(p,d,q)\ gilt\ folgendes^{250}:$$

$$AR: p = Ordnung\ des\ Autoregressive\ Bereichs$$
$$I: d = Grad\ der\ ersten\ Differenz$$
$$MA: q = Ordnung\ des\ Moving\ Average\ Bereichs$$

Die Ermittlung des Grades der Differenz d ist für das ARIMA-Modell notwendig. Hierbei wird für den MA-Bereich eine Ordnung festgelegt, die mindestens so groß ist wie der Grad der Differenz d. Zudem wird durch jede Differenz d eine Einheitswurzel im MA-Bereich gebildet.[251]

$$Sei\ der\ konstante\ Term\ c, der\ j-te\ Autoregressive-Parameter\ \phi_j\ und\ der\ Fehlerterm\ zur\ Zeit\ t\ also\ \varepsilon_t\ gegeben. Dann\ gilt\ für\ ein\ ARIMA-Modell\ mit\ p-ter\ Ordnung\ also\ ARIMA(p,0,0)\ bzw. AR(p)\ folgendes^{252}:$$

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

$$i)\ Für\ p = 1: -1 < \phi_1 < 1$$
$$ii)\ Für\ p = 2: -1 < \phi_2 < 1\ und\ \phi_2 + \phi_1 < 1\ und\ \phi_2 - \phi_1 < 1$$

[249] Vgl. Makridakis, 1998, S. 336.
[250] Makridakis, 1998, S. 336.
[251] Vgl. Schlittgen, 2020, S. 101.
[252] Vgl. Makridakis, 1998, S. 339f.

$$Sei\ der\ konstante\ Term\ c, der\ j-te\ Moving-Average-Parameter\ \theta_j\ und$$
$$der\ Fehlerterm\ zur\ Zeit\ t-k\ also\ \varepsilon_{t-k}\ gegeben. Dann\ gilt\ f\ddot{u}r\ das$$
$$ARIMA-Modell\ der\ q-ten\ Ordnung\ also\ ARIMA(0,0,q)\ bzw. MA(q)$$
$$folgendes^{253}:$$

$$Y_t = c + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}$$

$$Es\ gelten\ analog\ dieselben\ Restriktionen\ f\ddot{u}r\ q = 1\ und\ q = 2\ wie$$
$$bei\ ARIMA(p,0,0).$$

4.4 Nicht-stationäres ARIMA-Modell

$$Sei\ der\ Backshift-Operator\ BY_t = Y_{t-1}\ und\ t \in \mathbb{N}\backslash\{0\}\ gegeben. Dann\ gilt$$
$$f\ddot{u}r\ das\ nicht-station\ddot{a}re\ ARIMA-Modell$$
$$folgendes^{254}:$$

$$(1-\theta_1 B)\,(1-B)Y_t = c + (1-\theta_1 B)e_t$$

$$BY_t: Verschiebung\ der\ Daten\ um\ einen\ Zeitraum\ bzw. Periode$$

4.5 Saisonales ARIMA-Modell

$$Seien\ die\ Anzahl\ der\ Perioden\ pro\ Saison\ s, der$$
$$nicht-saisonale\ Abschnitt\ eines\ ARIMA-Modells\ (p,d,q)$$
$$und\ der\ saisonale\ Abschnitt\ eines\ ARIMA-Modells\ (P,D,Q)_s\ gegeben.$$
$$Sei\ zudem\ s,p,d,q,P,D,Q \in \mathbb{N}\ gegeben. Dann\ gilt\ f\ddot{u}r\ das$$
$$saisonale\ ARIMA-Modell\ also\ ARIMA(1,1,1)(1,1,1)_4$$
$$folgendes^{255}:$$

[253] Vgl. Makridakis, 1998, S. 342.
[254] Makridakis, 1998, S. 345.
[255] Makridakis, 1998, S. 346.

$$(1-\phi_1 B)\,(1-\Phi_1 B^4)\,(1-B)\,(1-B^4)Y_t = (1-\theta_1 B)\,(1-\theta_1 B^4)\varepsilon_t$$

$$(1-\phi_1 B) : Nicht-saisonales\ AR(1)$$
$$(1-\Phi_1 B^4) : Saisonales\ AR(1)$$
$$(1-B) : Nicht-saisonale\ Differnez$$
$$(1-B^4) : Saisonale\ Differenz$$
$$(1-\theta_1 B) : Nicht-saisonales\ MA(1)$$
$$(1-\theta_1 B^4) : Saisonales\ MA(1)$$

4.6 ARMAX-Modell

$$Sei\ ein\ weißes\ Rauschen\ z_t\ mit\ \mu = 0\ und\ i.i.d.\ gegeben.\ Sei\ zudem\ die$$
$$Kovariate\ x_t\ ,die\ Zeit\ t\ und\ der\ Regressionskoeffizient\ \beta\ gegeben.$$
$$Seien\ \phi(B) = 1-\phi_1 B-\cdots-\phi_p B^p\ und\ \theta(B) = 1-\theta_1 B-\cdots-\theta_q B^q\ gegeben.$$
$$Sei\ f\ddot{u}r\ nicht-station\ddot{a}re\ Daten\ \phi(B) = (1-B)^d\phi(B)\ gegeben.$$
$$Dann\ gilt\ f\ddot{u}r\ das\ ARMAX-Modell\ folgendes\ [256]:$$

$$\phi(B)y_t = \beta x_t + \theta(B)z_t$$

5 Beschreibung der Daten und der konkreten Aufgaben-/Problemstellung für den Berliner ÖPNV

Die konkrete Aufgabenstellung besagt, dass anhand der bereitgestellten Datensätze und exogener Variablen, eine Prognose zu den Ertragszahlen im Berliner ÖPNV erstellt werden soll. Hierzu sollen die Methoden aus Kapitel drei verwendet werden, um die Einflüsse der Regressoren auf die Fahrscheine zu ermitteln. Insbesondere sollen die Ergebnisse der Regressionsmethoden untereinander verglichen und am

[256] Hyndman, 2010, URL: https://robjhyndman.com/hyndsight/arimax/ [zuletzt zugegriffen am 07.06.2022]

Ende auch festgestellt werden, welche Regressionsmethode gegebenenfalls am besten für die Problemstellung geeignet ist.

Hierbei wird quantitative Forschung betrieben.[257] Zusätzlich wird die deduktive Methode verwendet, um vom Allgemeinen zum Spezifischen schließen zu können.[258]

Das heißt, man schließt vom Allgemeinen also der Forschungsfrage *(„Multivariate Regressionsanalyse exogener Variablen zur Prognose von Ertragszahlen im ÖPNV“)* zum Speziellen also der Hypothese. Die Hypothese beinhaltet, dass der Einfluss exogener Variablen auf die Ertragszahlen im Berliner-ÖPNV mithilfe von Multivariaten Regressionsanalysen ermittelbar ist. Infolgedessen sollen durch die Regressionsmethoden Prognosen für die Zukunft möglich sein. Die Daten diesbezüglich wurden unter anderem vom Drittmittelprojekt ReComMeND bereitgestellt, wobei es sich um einen Datensatz mit den Einflussvariablen *(„Treiber_ab_2005.xls“)* und um einen Datensatz mit den Erträgen der Fahrscheinarten *(„Ertrag_ohne_Schüler.xls“)* handelt. Die Auswertung der Daten erfolgt mit den multivariaten Regressionsmethoden (Multivariate Regression, Ridge-Regression, LASSO, LARS) aus Kapitel drei. Daraufhin wird beurteilt, ob die ermittelten Aussagen die Hypothese bestätigen oder nicht. Insbesondere wird hierbei noch der Zusammenhang zu Zeitreihen betrachtet.[259]

Die exogenen Variablen lauten: *Bevölkerungsentwicklung in Berlin (insgesamt und in Altersgruppen, Schüler, Studierende), Tourismuszahlen, Einflüsse vom Arbeitsmarkt (Erwerbstätige und Pendler), Wetter-Witterungsdaten, Preisentwicklung im ÖPNV (für einzelne Produktgruppen).*

Allgemein ist festzuhalten, dass die exogenen Variablen außerhalb vom Modell festgelegt werden und zusätzlich auf das Modell einwirken.[260]

[257] Vgl. Werth, 2021, S. 38.
[258] Vgl. Werth, 2021, S. 43.
[259] Vgl. Werth, 2021, S. 43f.
[260] Vgl. Mankiw, 1993, S. 7f.

6 Aufbereitung der Daten (Verkaufszahlen, Wetterdaten, Kalenderdaten)

In beiden Datensätzen existieren Kalenderdaten, die in monatlichen Zeitschritten vom 01.01.2005 bis zum 01.12.2023 verlaufen. Vorwiegend sind nicht für jeden Monat und jede Variable identisch viele Daten vorhanden, so dass zu einigen Beobachtungen keine Werte gegeben sind.

Man bezeichnet diese fehlenden Werte auch als *missing values*.[261] Infolgedessen könnten dadurch die Ergebnisse der multivariaten Regressionsmethoden verzerrt werden, was es gilt zu vermeiden.[262] Aus diesem Grund werden fehlende Werte aus dem Datensatz listenweise ausgeschlossen, wodurch die gesamte Beobachtung eliminiert wird.[263]

Folglich werden daher nur die Daten (einschließlich) vom 01.01.2012 bis (einschließlich) zum 01.12.2020 mithilfe der multivariaten Regressionsmethoden untersucht. Somit sind alle ausgewählten Variablen aus beiden Datensätzen vergleichbar.

Im Datensatz der Einflussvariablen *(„Treiber_ab_2005.xls“)* existieren Wetterdaten, die die Niederschlags- und Schneeintensität wiedergeben. In der Untersuchung werden aber nur die Variablen *„Tage mit mäßigem Regen“* und *„Tage mit Schnee“* betrachtet.
Die Verkaufszahlen sind nur im Datensatz der Zielvariablen *(„Ertrag_ohne_Schüler.xls“)* vorhanden und geben die Ertragszahlen jeweils zu den Fahrausweisen wieder. Hierbei werden nur folgende Variablen untersucht: *Einzelfahrscheine Berlin ABC, Tageskarten Berlin ABC, Monatskarten Berlin ABC, Abo, Firmenticket, Semester-/ Hochschulticket, Berlin Ticket S, Gesamt vor EAVs.* .

[261] Vgl. Backhaus, 2021, S. 53.
[262] Vgl. Backhaus, 2021, S. 54.
[263] Vgl. Backhaus, 2021, S. 55ff.

Mit den ABC-Tickets werden tatsächlich nur die ABC-Tickets mit den Berliner ÖPNV-Bereichen (A-, B- und C-Bereich) gleichzeitig ausgedrückt. Zudem steht *Gesamt vor EAVs* für die Einnahmeaufteilung bzw. die Fahrgeldeinnahmen.

In der Praxis mit R bilden die Y-Variablen mit allen X-Variablen jeweils ein multiples lineares Regressionsmodell.[264] Hierzu wird bei den exogenen Variablen zwei und vier (Tourismuszahlen, Wetter-und Witterungsdaten) als Regressoren für die Regressionsmodelle alle Monate außer der August verwendet. Dadurch möchte man gegebenenfalls auftauchende Trends beobachten, wobei der August als Referenzmonat gilt.
Konkret werden bezüglich der exogenen Variablen folgende Regressionsmodelle nach dem Schema $X_1, X_2, \dots, X_n \rightarrow Y$, wobei $n \in \mathbb{N}$ gilt, untersucht:

1)Exogene Variable: Bevölkerungsentwicklung in Berlin (insgesamt und in Altersgruppen, Schüler, Studierende)

X-Variablen für Ex. 1:

$1) Arbeitslose,\ 2) Bevölkerungsab - und\ zunahme$

Regressionsmodelle zu Ex.1:

$Modell\ 1.1) X - Variablen\ für\ Ex.1 \rightarrow Monatskarten\ Berlin\ ABC$

$Modell\ 1.2) X - Variablen\ für\ Ex.1 \rightarrow Abo$

$Modell\ 1.3) X - Variablen\ für\ Ex.1 \rightarrow Firmenticket$

$Modell\ 1.4) X - Variablen\ für\ Ex.1 \rightarrow Gesamt\ vor\ EAVs$

Spezielles Regressionsmodell zu Ex. 1:

$Modell\ 1.5) Studierende \rightarrow Semester -/ Hochschulticket$

2)Exogene Variable: Tourismuszahlen

X-Variablen für Ex. 2:

$1) Übernachtungen\ , 2) Samstage, 3) Sonn - und\ Feiertage, 4) Januar,$

[264] Vgl. Zelterman, 2015, S. 238.

5)$Februar$, 6)$März$, 7)$April$, 8)Mai, 9)$Juni$, 10)$Juli$, 11)$September$, 12)$Oktober$,
13)$November$, 14)$Dezember$

Regressionsmodelle zu Ex. 2:

$Modell\ 2.1) X - Variablen\ für\ Ex.2 \rightarrow\ Einzelfahrscheine\ Berlin\ ABC$

$Modell\ 2.2) X - Variablen\ für\ Ex.2 \rightarrow\ Tageskarten\ Berlin\ ABC$

$Modell\ 2.3) X - Variablen\ für\ Ex.2 \rightarrow\ Gesamt\ vor\ EAVs$

<u>3)Exogene Variable: Einflüsse vom Arbeitsmarkt (Erwerbstätige und Pendler)</u>

X-Variablen für Ex. 3:

1)$Arbeitslose$, 2)$Ferientage$, 3)$Arbeitstage$,
4)$Einpendler\ insgesamt$

Regressionsmodelle zu Ex. 3:

$Modell\ 3.1) X - Variablen\ für\ Ex.3 \rightarrow Einzelfahrscheine\ Berlin\ ABC$

$Modell\ 3.2) X - Variablen\ für\ Ex.3 \rightarrow Tageskarten\ Berlin\ ABC$

$Modell\ 3.3) X - Variablen\ für\ Ex.3\ (Ohne\ Arbeitstage) \rightarrow Firmenticket$

$Modell\ 3.4) X - Variablen\ für\ Ex.3 \rightarrow Gesamt\ vor\ EAVs$

<u>4)Exogene Variable: Wetter-/Witterungsdaten</u>

X-Variablen für Ex. 4:

1)$Tage\ mit\ mäßigem\ Regen$, 2)$Tage\ mit\ Schnee$, 3)$Januar$, 4)$Februar$, 5)$März$,
6)$April$, 7)Mai, 8)$Juni$, 9)$Juli$, 10)$September$, 11)$Oktober$, 12)$November$, 13)$Dezember$

Regressionsmodelle zu Ex. 4:

$Modell\ 4.1) X - Variablen\ für\ Ex.4 \rightarrow Einzelfahrscheine\ Berlin\ ABC$

$Modell\ 4.2) X - Variablen\ für\ Ex.4 \rightarrow Tageskarten\ Berlin\ ABC$

$Modell\ 4.3) X - Variablen\ für\ Ex.4 \rightarrow Gesamt\ vor\ EAVs$

5)Exogene Variable: Preisentwicklung im ÖPNV (für einzelne Produktgruppen)

X-Variablen für Ex. 5:

$1)Arbeitslose, 2)Superbenzin, 3)Stau$

Regressionsmodelle zu Ex. 5:

$Modell\ 5.1) X - Variablen\ für\ Ex. 5 \rightarrow Einzelfahrscheine\ Berlin\ ABC$

$Modell\ 5.2) X - Variablen\ für\ Ex. 5 \rightarrow Abo$

$Modell\ 5.3) X - Variablen\ für\ Ex. 5 \rightarrow Firmenticket$

$Modell\ 5.4) X - Variablen\ für\ Ex. 5 \rightarrow Berlin\ Ticket\ S$

Im nachfolgenden Text werden für folgende Zielvariablen Abkürzungen verwendet:

$Einzelfahrscheine := Einzelfahrscheine\ Berlin\ ABC$

$Tageskarten := Tageskarten\ Berlin\ ABC$

$Monatskarten := Monatskarten\ Berlin\ ABC$

$Hochschulticket := Semester -/ Hochschulticket$

7 Schrittweise Untersuchung und Anwendung verschiedener Regressionsverfahren auf ausgewählte exogene Variable in R

Es gilt folgende Empfehlung zur Vorgehensweise [265][266]:

$$1)Korrelation\ zwischen\ betrachteten\ Variablen\ prüfen$$

$$2)Regressionsmodell\ erstellen$$

$$3)Regressionsfunktion\ schätzen$$

$$4)Statistische\ Signifikanz\ durch\ F - Statistik\ bzw.\ p - Wert\ prüfen$$

$$5)Betrachtung\ des\ korrigierten\ Bestimmtheitsmaßes, Standardfehlers\ und\ der\ Regressionskoeffizienten$$

$$6)P - Werte\ der\ Regresssionskoeffizienten\ aus\ dem\ T - Test\ betrachten$$

[265] Vgl. Winke, 2020, S. 2.
[266] Vgl. Backhaus, 2021, S. 158.

$7) Prüfung\ der\ Regressionsvoraussetzungen$

$8) Ausreißer\ prüfen$

Zur Umsetzung der Regressionsmodelle und zur Prognosenbestimmung wurden folgende Pakete in R verwendet:

1) „readxl“: Es wurde zum Importieren der CSV-Datensätze genutzt.
2) „olsrr“: Es wurde zur Erstellung von Diagrammen für die Linearität, Homoskedastizität und Normalverteilung genutzt.
3) „car“: Es wurde zur Überprüfung von Autokorrelation und Multikollinearität genutzt.
4) „ModelMetrics“: Es wurde zur Ermittlung des MSE für die Ridge-und LASSO-Regression genutzt.
5) „glmnet“: Es wurde zur Erstellung der Regressionsmodelle für die Ridge-und LASSO-Regression genutzt, wo $alpha = 0$ für Ridge-Regression und $alpha = 1$ für LASSO-Regression steht.
6) „lars“: Es wurde zur Erstellung der Regressionsmodelle für die LARS-Regression genutzt.
7) „dplyr“, „tidyverse“, „tsibble“, „fable“: Diese Pakete wurden zur Erstellung der Prognose bezüglich der Zielvariablen und zur Ermittlung des MAPE-Wertes genutzt.

Der Code für diese Bachelorarbeit wird als Anhang abgegeben. Ein kleiner Einblick in den Code für das folgende Regressionsmodell ($X_1, X_2 \rightarrow Y$) würde folgendermaßen aussehen:

$$Arbeitslose, Bevölkerungsab - und\ zunahme\ \rightarrow Monatskarten$$

Multivariate Regression (R-Code):

Ex1_Monatskarten←lm(dataFrame_Ertrag_Y$Monatskarten_ABC ~ dataFrame_Treiber_X$data_Treiber.Arbeitslose +dataFrame_Treiber_X$data_Treiber..Auszubildende.Insgesamt.+dataFrame_Treiber_X$data_Treiber..Bevölkerungfsab..zunahme.)

Ridge-Regression (R-Code):

Ex1_Ridge_Monatskarten←glmnet(Ex1_Ridge_XMatrix,

dataFrame_Ertrag_Y$Monatskarten_ABC,alpha=0)

LASSO-Regression (R-Code):
Ex1_LASSO_Monatskarten←glmnet(Ex1_LASSO_XMatrix,
dataFrame_Ertrag_Y$Monatskarten_ABC,alpha=1)

LARS-Regression (R-Code):
Ex1_LARS_Monatskarten←lars(Ex1_LARS_XMatrix,
dataFrame_Ertrag_Y$Monatskarten_ABC,type="lar")

8 Vergleich und Bewertung der Güte der Regressionsverfahren auf Basis der erzielten Ergebnisse

Dieses Kapitel sich den Ergebnissen aus dem R-Code und deren Bewertung.

8.1 Bewertung der Güte-Werte

In der Praxis durch den R-Code war festzustellen, dass die meisten Schritte zur Vorgehensweise von Regressionsmethoden nur durch die multivariate Regression erfüllt werden konnten. Daher orientieren sich die Bewertung und Ergebnisse der Güte an die multivariate Regression, die demnach auch auf die anderen Regressionsmethoden übertragbar sind. Es gilt somit folgende Bewertung:

1)$Korrelation\ zwischen\ betrachteten\ Variablen\ prüfen$
Alle Regressionsmodelle, die in Kapitel sechs aufgelistet sind, wurden im R-Code betrachtet und zuvor auf Korrelation überprüft. Damit soll vorab Multikollinearität und Kollinearität verhindert werden. (Siehe Kapitel sieben zur Vorgehensweise)

2)und 3) $Regressionsmodellerstellung\ und\ Schätzung$
Wenn keine starke Kollinearität und Multikollinearität im Regressionsmodell vorherrscht, wird es im R-Code mit dem entsprechenden Paket, wie $(lm(), glmnet(), lars())$, erstellt und geschätzt. Das Regressionsmodell

$Studierende \rightarrow Hochschulticket$ war nur in der multivariaten Regression und in der LARS-Regression schätzbar. Bei den anderen Regressionsmethoden war es nicht möglich dieses umzusetzen, da die genutzten Funktionen mindestens zwei Regressoren für ein Regressionsmodell gefordert haben.

4)$Statistische\ Signifikanz\ durch\ F-Statistik\ bzw.\ p-Wert\ prüfen$

In der Regel gab der Befehl $summary( )$ den p-Wert der F-Statistik aus, wo durch einen Signifikanztest die statistische Signifikanz des jeweiligen Regressionsmodells festgestellt wurde (Siehe Kapitel 2.4.1).

$Modell\ 1.1) X-Variablen\ für\ Ex.1 \rightarrow Monatskarten$

p-Wert = 4.097e-12

Da $p-Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 1.2) X-Variablen\ für\ Ex.1 \rightarrow Abo$

p-Wert = < 2.2e-16

Da $p-Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 1.3) X-Variablen\ für\ Ex.1 \rightarrow Firmenticket$

p-Wert = 1.849e-08

Da $p-Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 1.4) X-Variablen\ für\ Ex.1 \rightarrow Gesamt\ vor\ EAVs$

p-Wert = < 2.2e-16

Da $p-Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 1.5) Studierende \rightarrow Hochschulticket$

p-Wert = < 2.2e-16

Da $p-Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 2.1) X-Variablen\ für\ Ex.2 \rightarrow Einzelfahrscheine$

p-Wert = < 2.2e-16

Da $p-Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 2.2) X - Variablen\ für\ Ex. 2 \rightarrow Tageskarten$

p-Wert = < 2.2e-16

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 2.3) X - Variablen\ für\ Ex. 2 \rightarrow\ Gesamt\ vor\ EAVs$

p-Wert = 5.002e-09

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 3.1) X - Variablen\ für\ Ex. 3 \rightarrow Einzelfahrscheine$

p-Wert = < 2.2e-16

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 3.2) X - Variablen\ für\ Ex. 3 \rightarrow Tageskarten$

p-Wert = < 2.2e-16

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 3.3) X - Variablen\ für\ Ex. 3\ \ (Ohne\ Arbeitstage)\ \rightarrow Firmenticket$

p-Wert = < 2.2e-16

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 3.4) X - Variablen\ für\ Ex. 3\ \rightarrow Gesamt\ vor\ EAVs$

p-Wert = < 2.2e-16

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 4.1) X - Variablen\ für\ Ex. 4 \rightarrow Einzelfahrscheine$

p-Wert = 0.8082

Da $p - Wert > 0.05$ gilt, ist das Regressionsmodell nicht statistisch signifikant.

$Modell\ 4.2) X - Variablen\ für\ Ex. 4 \rightarrow Tageskarten$

p-Wert = 0.08814

Da $p - Wert > 0.05$ gilt, ist das Regressionsmodell nicht statistisch signifikant.

$Modell\ 4.3) X - Variablen\ für\ Ex. 4 \rightarrow Gesamt\ vor\ EAVs$

p-Wert = 0.8907

Da $p - Wert > 0.05$ gilt, ist das Regressionsmodell nicht statistisch signifikant.

$Modell\ 5.1) X - Variablen\ für\ Ex. 5 \rightarrow Einzelfahrscheine$

p-Wert = 2.052e-12

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 5.2) X - Variablen\ für\ Ex. 5 \rightarrow Abo$

p-Wert = < 2.2e-16

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 5.3) X - Variablen\ für\ Ex. 5 \rightarrow Firmenticket$

p-Wert = 7.028e-15

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

$Modell\ 5.4) X - Variablen\ für\ Ex. 5 \rightarrow Berlin\ Ticket\ S$

p-Wert = 4.599e-06

Da $p - Wert < 0.05$ gilt, ist das Regressionsmodell statistisch signifikant.

Hierbei fiel auf, dass alle Regressionsmodelle statistisch signifikant waren. Die Ausnahmen beliefen sich auf die Regressionsmodelle 4.1, 4.2 und 4.3. Diese gelten als statistisch nicht signifikant. In der weiteren Untersuchung werden daher die Regressionsmodelle 4.1, 4.2 sowie 4.3 und demnach die exogene Variable Wetter- und Witterungsdaten nicht mehr weiter betrachtet.

5) *und* 6) $Betrachtung\ des\ korrigierten\ Bestimmtheitsmaßes, Standardfehlers,$ $der\ Regressionskoeffizienten\ und\ der\ p - Werte\ aus\ dem\ T - Test\ der$ $Regressionskoeffizienten$

Das korrigierte Bestimmtheitsmaß hat den Vorteil, dass durch das Anwachsen der Regressorenanzahl kein Overfitting des Modells entsteht (Siehe Kapitel 2.4.4). Somit ist besser festzustellen, welches Regressionsmodell eine bessere Erklärung durch die Regressoren darstellt.

Zudem gibt der Estimate-Wert eines Regressors dessen geschätzten Regressionskoeffizienten wieder. Dieser vermittelt bei der Interpretation des Regressionsmodells wichtige Hinweise über die Entwicklung der Regressionsgerade. Wenn der Regressorwert eines positiven Regressionskoeffizienten um eine Einheit steigt, so wächst auch die Regressionsgerade. Wenn der Regressorwert eines negativen Regressionskoeffizienten um eine Einheit steigt, so fällt die Regressionsgerade.[267] Im Folgenden werden nur die relevanten Regressionskoeffizienten interpretiert.
Es gelten somit folgende Ergebnisse für die exogenen Variablen und die Regressionskoeffizienten der jeweiligen Regressionsmodelle (Siehe Kapitel 2.3, Kapitel 2.4.4, Kapitel 2.5 und Kapitel 6 zur Einsicht der gewählten Regressoren für die jeweiligen exogenen Variablen):

1)Exogene Variable: Bevölkerungsentwicklung in Berlin (insgesamt und in Altersgruppen, Schüler, Studierende)

Regressionsmodelle zu Ex.1:

$Modell - 1.1) X - Variablen\ für\ Ex.1 \rightarrow Monatskarten$

Korrigiertes R-Quadrat: 0.3816 ⇒ 38,16% mittelmäßige Erklärung durch die Regressoren

Standardfehler = 966600 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 - Arbeitslose: Estimate-Wert = -2.464e+01 , p-Wert = 9.35e-09 ***

Regressor 2 - Bevölkerungsab-und zunahme: Estimate-Wert = 2.309e+02, p-Wert = 1.89e-08 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben beide Regressoren einen statistisch signifikanten Einfluss auf die Monatskarten aus.
Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, so fallen die Ertragszahlen der Monatskarten um 2.464e+01. Wenn die Werte des Regressors Bevölkerungsab-und zunahme um eine Einheit wachsen, dann steigen die Ertragszahlen der Monatskarten um 2.309e+02.

[267] Vgl. Statistikguru-Webseite, 2022 , URL: https://statistikguru.de/spss/multiple-lineare-regression/regressionskoeffizienten-interpretieren.html [zuletzt zugegriffen am 01.07.2022]

$Modell - 1.2) X - Variablen\ für\ Ex. 1 \rightarrow Abo$

Korrigiertes R-Quadrat: 0.6959 ⇒ 69,59% gute bis sehr gute Erklärung durch die Regressoren

Standardfehler = 2282000 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 - Arbeitslose: Estimate-Wert = -1.255e+02 , p-Wert = < 2e-16 ***

Regressor 2 - Bevölkerungsab-und zunahme: Estimate-Wert = -5.782e+02, p-Wert = 3.36e-09 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben beide Regressoren einen statistisch signifikanten Einfluss auf das Abo aus. Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, so fallen die Ertragszahlen des Abos um 1.255e+02. Wenn die Werte des Regressors Bevölkerungsab-und zunahme um eine Einheit wachsen, dann fallen die Ertragszahlen des Abos um 5.782e+02.

$Modell - 1.3) X - Variablen\ für\ Ex. 1 \rightarrow Firmenticket$

Korrigiertes R-Quadrat: 0.2741 ⇒ 27,41% geringe Erklärung durch die Regressoren

Standardfehler = 564600 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 - Arbeitslose: Estimate-Wert = -2.367e+00, p-Wert = 0.307

Regressor 2 - Bevölkerungsab-und zunahme: Estimate-Wert = -1.388e+02 , p-Wert = 8.37e-09 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% übt nur die Bevölkerungsab-und zunahme einen statistisch signifikanten Einfluss auf das Firmenticket aus. Wenn die Werte des Regressors Bevölkerungsab-und zunahme um eine Einheit wachsen, dann fallen die Ertragszahlen des Firmentickets um 1.388e+02.

$Modell - 1.4) X - Variablen\ für\ Ex. 1 \rightarrow Gesamt\ vor\ EAVs$

Korrigiertes R-Quadrat: 0.8083 ⇒ 80,83% sehr gute Erklärung durch die Regressoren

Standardfehler = 2926000 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 - Arbeitslose: Estimate-Wert = -2.538e+02, p-Wert = < 2e-16 ***

Regressor 2 - Bevölkerungsab-und zunahme: Estimate-Wert = 4.595e+02, p-Wert = 0.000117

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% übt nur die Arbeitslosenzahl einen statistisch signifikanten Einfluss auf die gesamten Fahrgeldeinnahmen aus. Die Bevölkerungsab-und zunahme übt nur einen geringen Einfluss auf die Zielvariable aus, auch wenn diese statistisch signifikant ist. Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, dann fallen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 2.538e+02.

Spezielles Regressionsmodell zu Ex. 1:

$Modell - 1.5) Studierende \rightarrow Hochschulticket$

Korrigiertes R-Quadrat: 0.5402 ⇒ 54,02% durchschnittliche Erklärung durch den Regressor

Standardfehler = 384600 ⇒ stark erhöhter Wert

T-Test des Regressionskoeffizienten:

Regressor 1- Studierende: Estimate-Wert = 3.046e+01, p-Wert = < 2e-16 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% übt nur die Arbeitslosenzahl einen statistisch signifikanten Einfluss auf die gesamten Fahrgeldeinnahmen aus. Wenn die Werte des Regressors Studierende um eine Einheit wachsen, dann steigen die Ertragszahlen des Hochschultickets um < 2e-16.

2)Exogene Variable: Tourismuszahlen

Regressionsmodelle zu Ex. 2:

$Modell - 2.1) X - Variablen\ für\ Ex. 2 \rightarrow\ Einzelfahrscheine$

Korrigiertes R-Quadrat: 0.6286 ⇒ 62,86% überdurchschnittliche bis gute Erklärung durch die Regressoren

Standardfehler = 787500 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 - Übernachtungen: Estimate-Wert = 3.660e+00 , p-Wert = < 2e-16 ***

Regressor 2 - Samstage: Estimate-Wert = 1.189e+05 , p-Wert = 0.46959
Regressor 3 – Sonn-und Feiertage:Estimate-Wert = -3.712e+05,p-Wert = 0.00456 **
Regressor 4 - Januar: Estimate-Wert = 5.605e+06 , p-Wert = < 2e-16 ***
Regressor 5 - Februar: Estimate-Wert = 2.637e+06 , p-Wert = 1.84e-09 ***
Regressor 6 - März: Estimate-Wert = 2.384e+06 , p-Wert = 2.78e-08 ***
Regressor 7 - April: Estimate-Wert = 1.420e+06 , p-Wert = 0.00166 **
Regressor 8 - Mai: Estimate-Wert = 1.848e+06 , p-Wert = 0.00029 ***
Regressor 9 - Juni: Estimate-Wert = 1.094e+06 , p-Wert = 0.00483 **
Regressor 10 - Juli: Estimate-Wert = 1.288e+05 , p-Wert = 0.73261
Regressor 11 - September: Estimate-Wert = 1.180e+06 , p-Wert = 0.00219 **
Regressor 12 - Oktober: Estimate-Wert = 1.772e+06 , p-Wert = 2.93e-05 ***
Regressor 13 - November: Estimate-Wert = 2.884e+06 , p-Wert = 3.95e-11 ***
Regressor 14 - Dezember: Estimate-Wert = 3.571e+06 , p-Wert = 3.03e-11 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur folgende Regressoren einen statistisch signifikanten Einfluss auf die Einzelfahrscheine aus: Übernachtungen, Januar, Februar, März, Mai,, Oktober, November, Dezember.
Mit einer Irrtumswahrscheinlichkeit von 0,1% bis 1% üben folgende Regressoren einen statistisch signifikanten Einfluss auf die Zielvariable aus: Sonn-und Feiertage, April, Juni, September.
Wenn die Werte des Regressors Übernachtungen um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 3.660e+00. Wenn die Werte des Regressors Sonn-und Feiertage um eine Einheit wachsen, dann fallen die Ertragszahlen der Einzelfahrscheine um 3.712e+05. Wenn die Werte des Regressors Januar um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 5.605e+06. Wenn die Werte des Regressors Februar um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 2.637e+06. Wenn die Werte des Regressors März um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 2.384e+06. Wenn die Werte des Regressors April um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 1.420e+06. Wenn die Werte des Regressors Mai um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 1.848e+06. Wenn die Werte des Regressors Juni um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 1.288e+05. Wenn die Werte des

Regressors September um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 1.180e+06. Wenn die Werte des Regressors Oktober um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 1.772e+06. Wenn die Werte des Regressors November um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 2.884e+06. Wenn die Werte des Regressors Dezember um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 3.571e+06.

$Modell - \ 2.2) X - Variablen\ für\ Ex.2 \rightarrow \ Tageskarten$

Korrigiertes R-Quadrat: 0.6745 ⇒ 67,45% gute Erklärung durch die Regressoren
Standardfehler = 169100 ⇒ stark erhöhter Wert
T-Test der Regressionskoeffizienten:
Regressor 1 - Übernachtungen: Estimate-Wert = 1.639e+00, p-Wert = < 2e-16 ***
Regressor 2 - Samstage: Estimate-Wert = -1.448e+03, p-Wert = 0.967249
Regressor 3 – Sonn-und Feiertage: Estimate-Wert = 1.059e+05, p-Wert = 0.000206 ***
Regressor 4 - Januar: Estimate-Wert = 1.858e+05, p-Wert = 0.036915 *
Regressor 5 - Februar: Estimate-Wert = -1.656e+04, p-Wert = 0.845860
Regressor 6 - März: Estimate-Wert = -1.047e+05, p-Wert = 0.217632
Regressor 7 - April: Estimate-Wert = -9.837e+04, p-Wert = 0.298765
Regressor 8 - Mai: Estimate-Wert = -2.029e+05, p-Wert = 0.057234 .
Regressor 9 - Juni: Estimate-Wert = -9.211e+04, p-Wert = 0.260785
Regressor 10 - Juli: Estimate-Wert = -2.036e+05, p-Wert = 0.013320 *
Regressor 11 - September: Estimate-Wert = -1.940e+05, p-Wert = 0.017850 *
Regressor 12 - Oktober: Estimate-Wert = -1.011e+05, p-Wert = 0.246127
Regressor 13 - November: Estimate-Wert = 1.957e+04, p-Wert = 0.813583
Regressor 14 - Dezember: Estimate-Wert = 2.964e+05, p-Wert = 0.004452 **

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur folgende Regressoren einen statistisch signifikanten Einfluss auf die Tageskarten aus: Übernachtungen, Sonn-und Feiertage. Mit einer Irrtumswahrscheinlichkeit von 0,1% bis 1% übt der Dezember einen statistisch signifikanten Einfluss auf die Zielvariable aus. Mit einer Irrtumswahrscheinlichkeit von 1% bis 5% üben folgende Regressoren

einen statistisch signifikanten Einfluss auf die Zielvariable aus: Januar, Juli, September.

Wenn die Werte des Regressors Übernachtungen um eine Einheit wachsen, dann steigen die Ertragszahlen der Tageskarten um 1.639e+00. Wenn die Werte des Regressors Sonn-und Feiertage um eine Einheit wachsen, dann steigen die Ertragszahlen der Tageskarten um 1.059e+05. Wenn die Werte des Regressors Januar um eine Einheit wachsen, dann steigen die Ertragszahlen der Tageskarten um 1.858e+05. Wenn die Werte des Regressors Juli um eine Einheit wachsen, dann fallen die Ertragszahlen der Tageskarten um 2.036e+05. Wenn die Werte des Regressors September um eine Einheit wachsen, dann fallen die Ertragszahlen der Tageskarten um 1.940e+05. Wenn die Werte des Regressors Dezember um eine Einheit wachsen, dann steigen die Ertragszahlen der Tageskarten um 2.964e+05.

$Modell-\ 2.3) X-Variablen$ für $Ex.2 \rightarrow\ Gesamt\ vor\ EAVs$

Korrigiertes R-Quadrat: 0.4211 ⇒ 42,11% mittelmäßige Erklärung durch die Regressoren

Standardfehler = 5084000 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 - Übernachtungen: Estimate-Wert = 7.667e+00, p-Wert = 1.29e-14 ***

Regressor 2 - Samstage: Estimate-Wert = -3.477e+05, p-Wert = 0.74288

Regressor 3 – Sonn-und Feiertage: Estimate-Wert = -1.335e+05, p-Wert = 0.87163

Regressor 4 - Januar: Estimate-Wert = 7.645e+06 , p-Wert = 0.00467 **

Regressor 5 - Februar: Estimate-Wert = 5.174e+06 , p-Wert = 0.04564 *

Regressor 6 - März: Estimate-Wert = 5.178e+06 , p-Wert = 0.04404 *

Regressor 7 - April: Estimate-Wert = 2.867e+06 , p-Wert = 0.31361

Regressor 8 - Mai: Estimate-Wert = 2.560e+06 , p-Wert = 0.42091

Regressor 9 - Juni: Estimate-Wert = 2.477e+06 , p-Wert = 0.31398

Regressor 10 - Juli: Estimate-Wert = 5.119e+05, p-Wert = 0.83330

Regressor 11 - September: Estimate-Wert = 3.562e+06 , p-Wert = 0.14419

Regressor 12 - Oktober: Estimate-Wert = 4.605e+06 , p-Wert = 0.08015 .

Regressor 13 - November: Estimate-Wert = 5.865e+06 , p-Wert = 0.02046 *

Regressor 14 - Dezember: Estimate-Wert = 6.996e+06 , p-Wert = 0.02433 *

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur die Übernachtungen einen statistisch signifikanten Einfluss auf die gesamten Fahrgeldeinnahmen aus. Mit einer Irrtumswahrscheinlichkeit von 0,1% bis 1% übt der Januar einen statistisch signifikanten Einfluss auf die Zielvariable aus. Mit einer Irrtumswahrscheinlichkeit von 1% bis 5% üben folgende Regressoren einen statistisch signifikanten Einfluss auf die Zielvariable aus: Februar, März, November, Dezember.
Wenn die Werte des Regressors Übernachtungen um eine Einheit wachsen, dann steigen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 7.667e+00. Wenn die Werte des Regressors Januar um eine Einheit wachsen, dann steigen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 7.645e+06. Wenn die Werte des Regressors Februar um eine Einheit wachsen, dann steigen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 5.174e+06. Wenn die Werte des Regressors März um eine Einheit wachsen, dann steigen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 5.178e+06. Wenn die Werte des Regressors November um eine Einheit wachsen, dann steigen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 5.865e+06. Wenn die Werte des Regressors Dezember um eine Einheit wachsen, dann steigen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 6.996e+06.

<u>3)Exogene Variable: Einflüsse vom Arbeitsmarkt (Erwerbstätige und Pendler)</u>
Regressionsmodelle zu Ex. 3:
$Modell-\ 3.1) X-Variablen\ für\ Ex.3 \rightarrow Einzelfahrscheine$
Korrigiertes R-Quadrat: 0.6286 ⇒ 62,86% überdurchschnittliche bis gute Erklärung durch die Regressoren
Standardfehler = 1475000 ⇒ stark erhöhter Wert
T-Test der Regressionskoeffizienten:
Regressor 1 – Arbeitslose: Estimate-Wert = -1.097e+02 , p-Wert = < 2e-16 ***
Regressor 2 - Ferientage: Estimate-Wert = -1.795e+04 , p-Wert = 0.32156
Regressor 3 - Arbeitstage: Estimate-Wert = 4.107e+05 , p-Wert = 0.00153 **
Regressor 4 - Einpendler insgesamt: Estimate-Wert = -8.619e+01 , p-Wert = < 2e-16 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur folgende Regressoren einen statistisch signifikanten Einfluss auf die Einzelfahrscheine aus: Arbeitslose, Einpendler insgesamt. Mit einer Irrtumswahrscheinlichkeit von 0,1% bis 1% üben die Arbeitstage einen statistisch signifikanten Einfluss auf die Zielvariable aus.

Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, dann fallen die Ertragszahlen der Einzelfahrscheine um 1.097e+02. Wenn die Werte des Regressors Arbeitstage um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 4.107e+05. Wenn die Werte des Regressors Einpendler insgesamt um eine Einheit wachsen, dann fallen die Ertragszahlen der Einzelfahrscheine um 8.619e+01.

$Modell - 3.2) X - Variablen$ für $Ex. 3 \rightarrow Tageskarten$

Korrigiertes R-Quadrat: 0.6745⇒ 67,45% gute Erklärung durch die Regressoren

Standardfehler = 622500 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 – Arbeitslose: Estimate-Wert = -5.199e+01 , p-Wert = < 2e-16 ***

Regressor 2 - Ferientage: Estimate-Wert = 3.229e+04 , p-Wert = 4.78e-05 ***

Regressor 3 - Arbeitstage: Estimate-Wert = 2.561e+04, p-Wert = 0.631

Regressor 4 - Einpendler insgesamt: Estimate-Wert = -3.627e+01, p-Wert = < 2e-16 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur folgende Regressoren einen statistisch signifikanten Einfluss auf die Tageskarten aus: Arbeitslose, Ferientage, Einpendler insgesamt.

Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, dann fallen die Ertragszahlen der Tageskarten um 5.199e+01. Wenn die Werte des Regressors Ferientage um eine Einheit wachsen, dann steigen die Ertragszahlen der Tageskarten um 3.229e+04. Wenn die Werte des Regressors Einpendler insgesamt um eine Einheit wachsen, dann fallen die Ertragszahlen der Tageskarten um 3.627e+01.

$Modell - 3.3) X - Variablen\ für\ Ex.3\ (Ohne\ Arbeitstage) \rightarrow Firmenticket$

Korrigiertes R-Quadrat: 0.8122 ⇒ 81,22% sehr gute Erklärung durch die Regressoren

Standardfehler = 287200 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 – Arbeitslose: Estimate-Wert = 2.107e+01 , p-Wert = < 2e-16 ***

Regressor 2 - Ferientage: Estimate-Wert = -4.597e+03 , p-Wert = 0.179

Regressor 3 - Einpendler insgesamt: Estimate-Wert = 3.067e+01 , p-Wert = < 2e-16 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur folgende Regressoren einen statistisch signifikanten Einfluss auf das Firmenticket aus: Arbeitslose, Einpendler insgesamt.

Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, dann steigen die Ertragszahlen des Firmentickets um 2.107e+01. Wenn die Werte des Regressors Einpendler insgesamt um eine Einheit wachsen, dann steigen die Ertragszahlen des Firmentickets um 3.067e+01.

$Modell - 3.4) X - Variablen\ für\ Ex.3 \rightarrow Gesamt\ vor\ EAVs$

Korrigiertes R-Quadrat: 0.7959⇒ 79,59% sehr gute Erklärung durch die Regressoren

Standardfehler = 3019000 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 – Arbeitslose: Estimate-Wert = -2.784e+02 , p-Wert = < 2e-16 ***

Regressor 2 - Ferientage: Estimate-Wert = -8.717e+03 , p-Wert = 0.8136

Regressor 3 - Arbeitstage: Estimate-Wert = 6.155e+05 , p-Wert = 0.0190 *

Regressor 4 - Einpendler insgesamt: Estimate-Wert = -3.505e+01,p-Wert = 0.0227 *

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur die Arbeitslosenzahlen einen statistisch signifikanten Einfluss auf die gesamten Fahrgeldeinnahmen aus. Mit einer Irrtumswahrscheinlichkeit von 1% bis 5% üben folgende Regressoren einen statistisch signifikanten Einfluss auf die Zielvariable aus: Arbeitstage, Einpendler insgesamt.

Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, dann fallen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 2.784e+02. Wenn die Werte

des Regressors Arbeitstage um eine Einheit wachsen, dann steigen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 6.155e+05. Wenn die Werte des Regressors Einpendler insgesamt um eine Einheit wachsen, dann fallen die Ertragszahlen der gesamten Fahrgeldeinnahmen um 3.505e+01.

5)Exogene Variable: Preisentwicklung im ÖPNV (für einzelne Produktgruppen)

Regressionsmodelle zu Ex. 5:

$Modell - 5.1) X - Variablen\ für\ Ex. 5 \rightarrow Einzelfahrscheine$

Korrigiertes R-Quadrat: 0.4064 ⇒ 40,64% mittelmäßige Erklärung durch die Regressoren

Standardfehler = 1865000 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 – Arbeitslose: Estimate-Wert = -5.470e+01 , p-Wert = 1.10e-09 ***

Regressor 2 - Superbenzin: Estimate-Wert = 8.432e+04 , p-Wert = 2.22e-07 ***

Regressor 3 - Stau: Estimate-Wert = 4.958e+04 , p-Wert = 5.03e-07 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben alle Regressoren einen statistisch signifikanten Einfluss auf die Einzelfahrscheine aus.

Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, dann fallen die Ertragszahlen der Einzelfahrscheine um 5.470e+01. Wenn die Werte des Regressors Superbenzin um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 8.432e+04. Wenn die Werte des Regressors Stau um eine Einheit wachsen, dann steigen die Ertragszahlen der Einzelfahrscheine um 4.958e+04.

$Modell - 5.2) X - Variablen\ für\ Ex. 5 \rightarrow Abo$

Korrigiertes R-Quadrat: 0.7913 ⇒ 79,13% sehr gute Erklärung durch die Regressoren

Standardfehler = 1891000 ⇒ stark erhöhter Wert

T-Test der Regressionskoeffizienten:

Regressor 1 – Arbeitslose: Estimate-Wert = -1.020e+02 , p-Wert = < 2e-16 ***

Regressor 2 - Superbenzin: Estimate-Wert = -1.534e+05 , p-Wert = < 2e-16 ***

Regressor 3 - Stau: Estimate-Wert = -4.647e+04 , p-Wert = 2.81e-06 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben alle Regressoren einen statistisch signifikanten Einfluss auf das Abo aus.
Wenn die Werte des Regressors Arbeitslose um eine Einheit wachsen, dann fallen die Ertragszahlen des Abos um 1.020e+02. Wenn die Werte des Regressors Superbenzin um eine Einheit wachsen, dann fallen die Ertragszahlen des Abos um 1.534e+05.Wenn die Werte des Regressors Stau um eine Einheit wachsen, dann fallen die Ertragszahlen des Abos um 4.647e+04.

$Modell - 5.3) X - Variablen\ für\ Ex. 5 \rightarrow Firmenticket$

Korrigiertes R-Quadrat: 0.4684 ⇒ 46,84% mittelmäßige Erklärung durch die Regressoren
Standardfehler = 483200 ⇒ stark erhöhter Wert
T-Test der Regressionskoeffizienten:
Regressor 1 – Arbeitslose: Estimate-Wert = 2.770e+00 , p-Wert = 0.193
Regressor 2 - Superbenzin: Estimate-Wert = -3.448e+04 , p-Wert = 4.04e-14 ***
Regressor 3 - Stau: Estimate-Wert = -1.318e+04 , p-Wert = 2.73e-07 ***
Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur folgende Regressoren einen statistisch signifikanten Einfluss auf das Firmenticket aus: Superbenzin, Stau.
Wenn die Werte des Regressors Superbenzin um eine Einheit wachsen, dann fallen die Ertragszahlen des Firmentickets um 3.448e+04. Wenn die Werte des Regressors Stau um eine Einheit wachsen, dann fallen die Ertragszahlen des Firmentickets um 1.318e+04.

$Modell - 5.4) X - Variablen\ für\ Ex. 5 \rightarrow Berlin\ Ticket\ S$

Korrigiertes R-Quadrat: 0.2093⇒ 20,93% geringe Erklärung durch die Regressoren
Standardfehler = 528000 ⇒ stark erhöhter Wert
T-Test der Regressionskoeffizienten:
Regressor 1 – Arbeitslose: Estimate-Wert = -3.779e+00 , p-Wert = 0.1053
Regressor 2 - Superbenzin: Estimate-Wert = 8.015e+03 , p-Wert = 0.0654 .
Regressor 3 - Stau: Estimate-Wert = 1.417e+04 , p-Wert = 4.09e-07 ***

Das heißt, mit einer Irrtumswahrscheinlichkeit von unter 0,1% üben nur die Stau-Werte einen statistisch signifikanten Einfluss auf das Berlin-Ticket-S aus.
Wenn die Werte des Regressors Stau um eine Einheit wachsen, dann steigen die Ertragszahlen des Firmentickets um 1.417e+04.

7)*Prüfung der Regressionsvoraussetzungen*

Die Regressionsvoraussetzungen waren in der Regel nur für die multivariate Regression überprüfbar. Diese Ergebnisse können wir aber auf alle Regressionsmethoden übertragen. Zudem war bei der Untersuchung auffällig, dass für alle Regressionsmodelle dieselben Regressionsvoraussetzungen erfüllt sind. Hierbei habe ich mich auf die Kapitel 2.2 und 2.3 berufen. Es gelten somit für alle Regressionsmodelle folgende Voraussetzungen:

1)Nicht-Linearität
2)Homoskedastizität
3)Normalverteilung
4)Autokorrelation
5)Keine perfekte Multikollinearität

Die Nicht-Linearität ist bei realen Datensätze üblich und ist hier auch nicht verwunderlich.[268] Zudem existiert eine ausreichende Streuung durch die Homoskedastizität. Da die Stichprobe groß genug ist ($n > 40$) ,sind die Schätzungen aufgrund des zentralen Grenzwertsatzes normalverteilt (Siehe Kapitel 2.2.3). Die Autokorrelation ist für Zeitreihen sehr wahrscheinlich, da sich die Werte durch die enge zeitliche Messung gegenseitig beeinflussen und daher auch eine starke Korrelation mit $d < 1$ durchweg festgestellt wurde. Insbesondere sind aufgrund der Autokorrelation und der Nicht-Linearität die Standardfehler stark verzerrt und daher extrem hoch. Einer Multikollinearität konnte zu Anfang bereits durch die Korrelationsüberprüfungen der Regressoren entgegen gewirkt werden, die auch in einem nicht schädlichen Rahmen vorhanden ist. Das heißt, der VIF-Wert liegt stets unter zehn und ist fast immer unter drei.

[268] Vgl. Backhaus, 2021, S. 103.

Durch die Erfüllung der Homoskedastizität, der nicht schädlichen Multikollinearität und der ausreichenden Stichprobengröße wird eine zufriedenstellende Genauigkeit der Schätzungen bereitgestellt (Siehe Kapitel 2.3).

8)$Ausreißer\ prüfen$

Die Ausreißer-Prüfung konnte nur bei der multivariaten Regression durchgeführt werden, dessen Ergebnisse demnach wieder für alle Regressionsmethoden übertragen werden können. Hierbei wurde festgestellt, dass nach den Methoden in Kapitel 2.6 keine einflussreichen Ausreißer ermitteln werden konnten. Somit wurden keine Werte aus den Daten der Regressionsmodelle eliminiert.

8.2 Vergleich der Regressionsmethoden

Die betrachteten Regressionsmodelle wurden jeweils mithilfe der multivariaten Regression, Ridge-Regression, LASSO-Regression und LARS-Regression geschätzt. Der Vergleich behandelt die Frage, welche Regressionsmethode eine bessere Schätzung zu den betrachteten Regressionsmodellen ermöglichen kann. Hierbei wurde für jedes Regressionsmodell jeweils alle Regressionsmethoden ausgeführt und auch jeweils das Bestimmtheitsmaß bestimmt. welches als Vergleichswert verwendet wird.

Als alternativer Vergleichswert gilt der MSE (oder RMSE), der die Güte eines Gesamtmodells wiedergibt. Das MSE (oder RMSE) konnte für die LARS-Regression nicht in R bestimmt werden, da keine adäquate Funktion diesbezüglich existiert und es daher als Vergleichswert ausgeschlossen wurde. Es gilt daher für den Vergleich folgendes:

Regressionsmodelle zu Ex.1:

$Modell - 1.1) X - Variablen\ für\ Ex. 1 \rightarrow Monatskarten$

Multivariate Regression: R-Quadrat = 0.3931

Ridge-Regression: R-Quadrat = 0.3923086

LASSO-Regression: R-Quadrat = 0.3931135

LARS-Regression: R-Quadrat = 0.393

$Modell - 1.2) X - Variablen\ für\ Ex. 1 \rightarrow Abo$

Multivariate Regression: R-Quadrat = 0.7016

Ridge-Regression: R-Quadrat = 0.6985696

LASSO-Regression: R-Quadrat = 0.7014409

LARS-Regression: R-Quadrat = 0.702

$Modell - 1.3) X - Variablen\ für\ Ex. 1 \rightarrow Firmenticket$

Multivariate Regression: R-Quadrat = 0.2876

Ridge-Regression: R-Quadrat = 0.2838293

LASSO-Regression: R-Quadrat = 0.2876204

LARS-Regression: R-Quadrat = 0.288

$Modell - 1.4) X - Variablen\ für\ Ex. 1 \rightarrow Gesamt\ vor\ EAVs$

Multivariate Regression: R-Quadrat = 0.8118

Ridge-Regression: R-Quadrat = 0.8059744

LASSO-Regression: R-Quadrat = 0.8118021

LARS-Regression: R-Quadrat = 0.812

$Modell - 1.5) Studierende \rightarrow Hochschulticket$

Multivariate Regression: R-Quadrat = 0.5445

Ridge-Regression: R-Quadrat = nicht ermittelbar

LASSO-Regression: R-Quadrat = nicht ermittelbar

LARS-Regression: R-Quadrat = 0.545

Regressionsmodelle zu Ex.2:

$Modell - 2.1) X - Variablen\ für\ Ex. 2 \rightarrow\ Einzelfahrscheine$

Multivariate Regression: R-Quadrat = 0.908

Ridge-Regression: R-Quadrat = 0.8814305

LASSO-Regression: R-Quadrat = 0.9078979

LARS-Regression: R-Quadrat = 0.908

$Modell -\ 2.2) X - Variablen\ für\ Ex. 2 \rightarrow\ Tageskarten$

Multivariate Regression: R-Quadrat = 0.9791

Ridge-Regression: R-Quadrat = 0.9684188

LASSO-Regression: R-Quadrat = 0.9775896

LARS-Regression: R-Quadrat = 0.979

$Modell-\ 2.3) X-Variablen\ f\ddot{u}r\ Ex.2 \rightarrow\ Gesamt\ vor\ EAVs$

Multivariate Regression: R-Quadrat = 0.4968

Ridge-Regression: R-Quadrat = 0.4869708

LASSO-Regression: R-Quadrat = 0.3855611

LARS-Regression: R-Quadrat = 0.497

Regressionsmodelle zu Ex.3:

$Modell-\ 3.1) X-Variablen\ f\ddot{u}r\ Ex.3 \rightarrow Einzelfahrscheine$

Multivariate Regression: R-Quadrat = 0.6424

Ridge-Regression: R-Quadrat = 0.6341508

LASSO-Regression: R-Quadrat = 0.642412

LARS-Regression: R-Quadrat = 0.642

$Modell-\ 3.2) X-Variablen\ f\ddot{u}r\ Ex.3 \rightarrow Tageskarten$

Multivariate Regression: R-Quadrat = 0.6867

Ridge-Regression: R-Quadrat = 0.6739603

LASSO-Regression: R-Quadrat = 0.6844835

LARS-Regression: R-Quadrat = 0.687

$Modell-\ 3.3) X-Variablen\ f\ddot{u}r\ Ex.3\ (Ohne\ Arbeitstage) \rightarrow Firmenticket$

Multivariate Regression: R-Quadrat = 0.8175

Ridge-Regression: R-Quadrat = 0.7928274

LASSO-Regression: R-Quadrat = 0.8174764

LARS-Regression: R-Quadrat = 0.818

$Modell-\ 3.4) X-Variablen\ f\ddot{u}r\ Ex.3\ \rightarrow Gesamt\ vor\ EAVs$

Multivariate Regression: R-Quadrat = 0.8035

Ridge-Regression: R-Quadrat = 0.7915416

LASSO-Regression: R-Quadrat = 0.7967226

LARS-Regression: R-Quadrat = 0.804

Regressionsmodelle zu Ex. 5:

$Modell - 5.1) X - Variablen\ für\ Ex. 5 \rightarrow Einzelfahrscheine$

Multivariate Regression: R-Quadrat = 0.423

Ridge-Regression: R-Quadrat = 0.4216922

LASSO-Regression: R-Quadrat = 0.4230087

LARS-Regression: R-Quadrat = 0.423

$Modell - 5.2) X - Variablen\ für\ Ex. 5 \rightarrow Abo$

Multivariate Regression: R-Quadrat = 0.7971

Ridge-Regression: R-Quadrat = 0.7945193

LASSO-Regression: R-Quadrat = 0.7970937

LARS-Regression: R-Quadrat = 0.797

$Modell - 5.3) X - Variablen\ für\ Ex. 5 \rightarrow Firmenticket$

Multivariate Regression: R-Quadrat = 0.4833

Ridge-Regression: R-Quadrat = 0.480881

LASSO-Regression: R-Quadrat = 0.4833129

LARS-Regression: R-Quadrat = 0.483

$Modell - 5.4) X - Variablen\ für\ Ex. 5 \rightarrow Berlin\ Ticket\ S$

Multivariate Regression: R-Quadrat = 0.2315

Ridge-Regression: R-Quadrat = 0.2214869

LASSO-Regression: R-Quadrat = 0.2314806

LARS-Regression: R-Quadrat = 0.231

Insgesamt ist festzustellen, dass alle Bestimmtheitsmaße der jeweiligen Regressionsmethoden sehr ähnlich sind. Die Unterschiede liegen gelegentlich im Hundertstel-, Tausendstel- oder Zehntausendstel-Bereich. Demnach kann man keine wirklich allgemeingültige Aussage treffen, welche Regressionsmethode nun eine tatsächlich bessere Schätzung wiedergibt. Zudem könnten sich die Güte-Werte beim korrigierten Bestimmtheitsmaß wieder ausgleichen, da bekanntlich die Werte des Bestimmtheitsmaßes durch die Regressorenanzahl mitwachsen. Generell ist aber auffällig, dass die höchsten Bestimmtheitsmaßwerte bei der multivariaten und LARS-Regression zu finden sind, wobei diese meistens fast identisch sind. Das

Bestimmtheitsmaß der Ridge-Regression weist in der Regel den kleinsten Wert auf. Generell ist bei der LASSO-Regression das Bestimmtheitsmaß vergleichsweise kleiner als bei der multivariaten und Ridge-Regression.

8.3 Prognose durch Zeitreihenanalyse

Die Prognosen erfolgen durch verschiedene Pakete über den R-Code (Siehe Kapitel 7), wobei sich der Prognosezeitraum vom 01.01.2021 bis zum 01.12.2021 erstreckt. Das heißt, für alle Regressionsmodelle wird jeweils eine Prognose erstellt und der jeweilige MAPE-Wert ermittelt. Diese MAPE-Werte geben dann Auskunft über die Genauigkeit der Prognose (Siehe Kapitel 4.2).

$Prognose\ Modell - 1.1) X - Variablen\ f\ddot{u}r\ Ex.1 \rightarrow Monatskarten$

ARIMA-Modell: ARIMA(0,1,1)(2,0,0)[12]

Prognose: MAPE-Wert = 5.15

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell - 1.2) X - Variablen\ f\ddot{u}r\ Ex.1 \rightarrow Abo$

ARIMA-Modell: ARIMA(3,1,0)(1,0,0)[12]

Prognose: MAPE-Wert = 0.307

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell - 1.3) X - Variablen\ f\ddot{u}r\ Ex.1 \rightarrow Firmenticket$

ARIMA-Modell: ARIMA(1,1,2)(0,1,1)[12]

Prognose: MAPE-Wert = 1.19

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell - 1.4) X - Variablen\ f\ddot{u}r\ Ex.1 \rightarrow Gesamt\ vor\ EAVs$

ARIMA-Modell: ARIMA(2,0,0)(1,0,0)[12]

Prognose: MAPE-Wert = 2.79

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell - 1.5) Studierende \rightarrow Hochschulticket$

ARIMA-Modell: ARIMA(1,0,1)(0,1,0)[12]

Prognose: MAPE-Wert = 3.79

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell - 2.1) X - Variablen\ f\ddot{u}r\ Ex.2 \rightarrow\ Einzelfahrscheine$

ARIMA-Modell: ARIMA(3,1,0)(1,0,0)[12]

Prognose: MAPE-Wert = 3.74

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell -\ 2.2) X - Variablen\ f\ddot{u}r\ Ex.2 \rightarrow\ Tageskarten$

ARIMA-Modell: ARIMA(1,0,2)

Prognose: MAPE-Wert = 4.65

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell -\ 2.3) X - Variablen\ f\ddot{u}r\ Ex.2 \rightarrow\ Gesamt\ vor\ EAVs$

ARIMA-Modell: ARIMA(0,1,1)(0,0,1)[12]

Prognose: MAPE-Wert = 1.69

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell -\ 3.1) X - Variablen\ f\ddot{u}r\ Ex.3 \rightarrow Einzelfahrscheine$

ARIMA-Modell: ARIMA(2,0,0)(1,0,0)[12]

Prognose: MAPE-Wert = 7.08

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell-\ 3.2) X-Variablen\ f\ddot{u}r\ Ex.3 \rightarrow Tageskarten$

ARIMA-Modell: ARIMA(2,0,2)(2,0,0)[12]

Prognose: MAPE-Wert = 13.2

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine gute Prognosegenauigkeit auf.

$Prognose\ Modell-\ 3.3) X-Variablen\ f\ddot{u}r\ Ex.3\ (Ohne\ Arbeitstage) \rightarrow Firmenticket$

ARIMA-Modell: ARIMA(0,0,5)(0,0,1)[12]

Prognose: MAPE-Wert = 2.42

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell-\ 3.4) X-Variablen\ f\ddot{u}r\ Ex.3 \rightarrow Gesamt\ vor\ EAVs$

ARIMA-Modell: ARIMA(1,0,0)(1,0,0)[12]

Prognose: MAPE-Wert = 2.86

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell-5.1) X-Variablen\ f\ddot{u}r\ Ex.5 \rightarrow Einzelfahrscheine$

ARIMA-Modell: ARIMA(0,1,3)(1,0,0)[12]

Prognose: MAPE-Wert = 7.59

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell-5.2) X-Variablen\ f\ddot{u}r\ Ex.5 \rightarrow Abo$

ARIMA-Modell: ARIMA(1,1,0)(0,1,2)[12]

Prognose: MAPE-Wert = 0.282

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell - 5.3) X - Variablen\ für\ Ex. 5 \rightarrow Firmenticket$

ARIMA-Modell: ARIMA(2,1,2)

Prognose: MAPE-Wert = 1.15

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

$Prognose\ Modell - 5.4) X - Variablen\ für\ Ex. 5 \rightarrow Berlin\ Ticket\ S$

ARIMA-Modell: ARIMA(0,1,0)(0,1,0)[12]

Prognose: MAPE-Wert = 4.07

Das heißt, die Prognose weist bezüglich diesem Regressionsmodell eine hohe Prognosegenauigkeit auf.

Insgesamt ist festzustellen, dass alle Prognosen eine hohe Prognosegenauigkeit wiedergeben. Eine Ausnahme existiert bei der Prognose des Modells 3.2, wo nur eine gute Prognosegenauigkeit vorausgesagt wurde.

9 Zusammenfassung und Schlussfolgerungen

Dieses Kapitel trägt die Auswertungen der Ergebnisse kurz zusammen und formuliert eine Schlussfolgerung diesbezüglich.

9.1 Zusammenfassung

Es wird der Datensatz für die Einflussvariablen ("Treiber_ab_2005.xls") und der Datensatz für die Zielvariablen ("Ertrag_ohne_Schüler.xls") für die auszuführenden Regressionsmethoden (multivariate Regression, Ridge-Regression, LASSO-Regression und LARS-Regression) verwendet. Durch die Aufbereitung der Daten erstreckt sich der Untersuchungszeitraum einschließlich vom 01.01.2012 bis einschließlich zum 01.12.2020. Es werden die Einflüsse der Ertragszahlen der Zielvariablen Einzelfahrscheine, Tageskarten, Monatskarten, Abo, Firmenticket, Hochschulticket, Berlin Ticket S und Gesamt vor EAVs untersucht.

Die beste Prüfung der Regressionsvoraussetzungen und der Güte war mit der multivariaten Regression möglich. In R waren hierzu alle Funktionen und Dokumentationen ausführlich vorhanden. Für die Ridge- und LASSO-Regression konnten mit den bereitgestellten Funktionen und Paketen in R nur die Regressionsmodellschätzung, das R-Quadrat und einige Prognosegütewerte umgesetzt werden. Funktionen für die LARS-Regression sind am wenigsten in R vorhanden, weswegen auch nur die Regressionsmodellschätzung und das R-Quadrat für die LARS-Regression ermittelt werden konnte. Das Modell 1.5 konnte nur mithilfe der multivariaten Regression und der LARS-Regression durchgeführt werden, da die anderen Regressionsmethoden für das Regressionsmodell jeweils mindestens zwei Regressoren forderten. Der gesamte R-Code wird dieser Bachelorarbeit als Anhang beigefügt werden.

Somit wurden alle Ergebnisse aus den Ausführungen der multivariaten Regression gewonnen. Es wurde festgestellt, dass alle Regressionsmodelle, außer die Modelle 4.1, 4.2 und 4.3, statistisch signifikant sind. Im Allgemeinen gelten für alle statistisch signifikanten Regressionsmodelle folgende Regressionsvoraussetzungen:

1)Nicht-Linearität
2)Homoskedastizität
3)Normalverteilung
4)Autokorrelation
5)Keine perfekte Multikollinearität

Da die Datensätze eine ausreichende Stichprobengröße, Streuung und keine schädliche Multikollinearität aufweisen, sind die Schätzungen angemessen vertrauenswürdig (Siehe Kapitel 2.3).
Zudem wurden keine schädlichen Ausreißer entdeckt, weswegen keine Daten aus den Datensätzen eliminiert wurden. Der Vergleich der Regressionsmethoden wurde durch das Bestimmtheitsmaß als Vergleichsgüte umgesetzt, wo keine eindeutige Aussage über die beste Regressionsmethode getroffen werden konnte. Konkret kann man aber festhalten, dass die praktische Umsetzung der Regressionsuntersuchung mit der multivariaten Regression am angenehmsten war. In der Regel wurde für die

meisten Regressionsmodelle eine mittelmäßige bis sehr gute Erklärung der Zielvariablen durch Regressoren beobachtet.
Insbesondere wurde durch den MAPE-Wert eine hohe Prognosegenauigkeit jeweils für alle Regressionsmodelle identifiziert. Nur das Modell 3.2 wies eine gute Prognosegenauigkeit auf.

9.2 Schlussfolgerungen

Für die Regressionsmodelle im Einzelnen kann man aber folgende Rückschlüsse ziehen:

$Modell - 1.1) X - Variablen\ für\ Ex. 1 \rightarrow Monatskarten\ ABC$

Das Regressionsmodell leistet einen mittelmäßigen Erklärungsbeitrag (Korrigiertes R-Quadrat: 38,16%) durch die Regressoren. Die Regressoren Arbeitslose und Bevölkerungsab-und zunahme sind statistisch signifikant (Irrtumswahrscheinlichkeit von unter 0,1%). Zudem ist festzustellen, dass wenn die Werte des Regressors Arbeitslose und die Werte vom Regressor Bevölkerungsab-und zunahme wachsen würden, so würden die Ertragszahlen der Monatskarten steigen.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 5.15, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann. Für dieses Regressionsmodell wäre das gegebenenfalls weniger geeignet.

$Modell - 1.2) X - Variablen\ für\ Ex. 1 \rightarrow Abo$

Das Regressionsmodell leistet einen guten Erklärungsbeitrag (Korrigiertes R-Quadrat: 69,59%) durch die Regressoren. Beide Regressoren sind statistisch signifikant (Irrtumswahrscheinlichkeit von unter 0,1%). Zudem ist festzustellen, dass wenn die Werte des Regressors Arbeitslose und die Werte vom Regressor Bevölkerungsab-und zunahme wachsen würden, so würden die Ertragszahlen des Abos steigen.
Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Regressionsmodells) ein starker Einfluss durch die Regressoren Arbeitslose und Bevölkerungsab-und zunahme auf das Abo existiert.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 0.307, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell - 1.3) X - Variablen\ für\ Ex. 1 \rightarrow Firmenticket$

Das Regressionsmodell leistet nur einen geringen Erklärungsbeitrag (Korrigiertes R-Quadrat: 27,41%) durch die Regressoren. Nur der Regressor Bevölkerungsab-und zunahme ist statistisch signifikant (Irrtumswahrscheinlichkeit von unter 0,1%). Wenn die Werte des Regressors Bevölkerungsab-und zunahme wachsen, dann fallen die Ertragszahlen des Firmentickets.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 1.19, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann. Für dieses Regressionsmodell wäre das gegebenenfalls weniger geeignet.

$Modell - 1.4) X - Variablen\ für\ Ex. 1 \rightarrow Gesamt\ vor\ EAVs$

Das Regressionsmodell leistet einen sehr guten Erklärungsbeitrag (Korrigiertes R-Quadrat: 80,83%) durch die Regressoren. Beide Regressoren sind statistisch signifikant, wobei der Regressor Arbeitslose eine Irrtumswahrscheinlichkeit von unter 0,1% vorweist. Wenn die Werte des Regressors Arbeitslose wachsen würden, dann würden die Ertragszahlen der gesamten Fahrgeldeinnahmen fallen.

Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Regressionsmodells) ein starker Einfluss vom Regressor Arbeitslose auf die gesamten Fahrgeldeinnahmen existiert. Gegebenenfalls könnte man unter anderem die Fahrscheinpreise für Arbeitslose mindern, um dem entgegen zu wirken.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 2.79, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell - 1.5) Studierende \rightarrow Hochschulticket$

Das Regressionsmodell leistet einen durchschnittlichen Erklärungsbeitrag (Korrigiertes R-Quadrat: 54,02%) durch die Regressoren. Der Regressor Studierende ist mit einer Irrtumswahrscheinlichkeit von unter 0,1% statistisch signifikant. Wenn die Werte des Regressors Studierende wachsen würden, so würden die Ertragszahlen des Hochschultickets steigen. Das heißt, je mehr Studenten es gibt, desto mehr Erträge würde man durch die Hochschultickets erzielen.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 3.79, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell - 2.1) X - Variablen\ für\ Ex. 2 \rightarrow\ Einzelfahrscheine$

Das Regressionsmodell leistet einen überdurchschnittlichen bis guten Erklärungsbeitrag (Korrigiertes R-Quadrat:62,86%) durch die Regressoren.
Alle Regressoren außer Samstage und Juli sind statistisch signifikant. Die meisten Regressoren weisen eine Irrtumswahrscheinlichkeit von unter 0,1% oder von 0,1% bis 1% auf.
Wenn die Werte der folgenden Regressoren wachsen würden, dann würden die Ertragszahlen der Einzelfahrscheine steigen: Übernachtungen, Sonn-und Feiertage, Januar, Februar, März, April, Mai, Juni, September, Oktober, November, Dezember. Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Regressionsmodells) fast alle Monate des Jahres und die Übernachtungen einen starken Einfluss auf die Ertragszahlen der Einzelfahrscheine nehmen.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 3.74, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell - 2.2) X - Variablen\ für\ Ex. 2 \rightarrow\ Tageskarten$

Das Regressionsmodell leistet einen überdurchschnittlichen Erklärungsbeitrag (Korrigiertes R-Quadrat:67,45%) durch die Regressoren.

Die Regressoren Übernachtungen (Irrtumswahrscheinlichkeit von unter 0,1%), Dezember (Irrtumswahrscheinlichkeit von 0,1% bis 1%), Januar, Juli, September(alle drei mit einer Irrtumswahrscheinlichkeit von 1% bis 5%) sind statistisch signifikant.

Wenn die Werte der Regressoren Übernachtungen, Januar und Dezember wachsen würden, so würden die Ertragszahlen der Tageskarten steigen. Wenn aber die Werte der Regressoren Juli und September wachsen würden, so würden die Ertragszahlen der Tageskarten fallen. Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Modells) die Regressoren Übernachtungen, Januar und Dezember einen starken Einfluss auf die Fahrgeldeinnahmen der Tageskarten nehmen.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 4.65, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell - \ 2.3) X - Variablen \ für \ Ex. 2 \rightarrow \ Gesamt \ vor \ EAVs$

Das Regressionsmodell leistet einen mittelmäßigen Erklärungsbeitrag (Korrigiertes R-Quadrat:42,11%) durch die Regressoren.
Die Regressoren Übernachtungen (Irrtumswahrscheinlichkeit von unter 0,1%), Januar (Irrtumswahrscheinlichkeit von 0,1% bis 1%), Februar, März, November, Dezember(alle vier mit einer Irrtumswahrscheinlichkeit von 1% bis 5%) sind statistisch signifikant.
Wenn die Werte der Regressoren Übernachtungen, Januar, Februar, März, November und Dezember wachsen würden, so würden die Ertragszahlen der gesamten Fahrgeldeinnahmen steigen.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 1.69, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen kann. Für dieses Regressionsmodell wäre das gegebenenfalls weniger geeignet.

$Modell-\ 3.1) X-Variablen\ für\ Ex.3 \rightarrow Einzelfahrscheine$

Das Regressionsmodell leistet einen überdurchschnittlichen bis guten Erklärungsbeitrag (Korrigiertes R-Quadrat:62,86%) durch die Regressoren. Die Regressoren Arbeitslose, Arbeitstage und Einpendler insgesamt sind statistisch signifikant (alle mit der Irrtumswahrscheinlichkeit von unter 0,1%).

Wenn die Werte der Regressoren Arbeitslose und Einpendler insgesamt wachsen würden, so würden die Ertragszahlen der Einzelfahrscheine fallen. Wenn aber die Werte der Regressors Arbeitstage wachsen würde, so würden die Ertragszahlen der Einzelfahrscheine steigen.
Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Modells) die Regressoren Arbeitslose und Einpendler insgesamt einen negativen Einfluss auf die Ertragszahlen der Einzelfahrscheine nehmen. Gegebenenfalls könnte man die Preise der Einzelfahrscheine für Arbeitslose und Einpendler mindern, um dem entgegen zu wirken.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 7.08 , womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell-\ 3.2) X-Variablen\ für\ Ex.3 \rightarrow Tageskarten$

Das Regressionsmodell leistet einen guten Erklärungsbeitrag (Korrigiertes R-Quadrat:67,45%) durch die Regressoren. Die Regressoren Arbeitslose, Ferientage und Einpendler insgesamt sind statistisch signifikant (alle mit der Irrtumswahrscheinlichkeit von unter 0,1%).
Wenn die Werte der Regressoren Arbeitslose und Einpendler insgesamt wachsen würden, so würden die Ertragszahlen der Tageskarten fallen. Wenn aber die Werte der Regressors Ferientage wachsen würden, so würden die Ertragszahlen der Tageskarten steigen.
Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Modells) die Regressoren Arbeitslose und Einpendler insgesamt einen negativen Einfluss auf die Fahrgeldeinnahmen der Tageskarten nehmen. Gegebenenfalls könnte man die Preise der Tageskarten für Arbeitslose und Einpendlern mindern, um dem entgegen zu wirken.

Des Weiteren existiert eine gute Prognosegenauigkeit mit MAPE = 13.2, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen kann.

$Modell - \ 3.3) X - Variablen \ für \ Ex. 3 \ (Ohne \ Arbeitstage) \ \rightarrow Firmenticket$

Das Regressionsmodell leistet einen sehr guten Erklärungsbeitrag (Korrigiertes R-Quadrat:81,22%) durch die Regressoren. Die Regressoren Arbeitslose und Einpendler insgesamt sind mit einer Irrtumswahrscheinlichkeit von unter 0,1% statistisch signifikant.
Wenn die Werte der Regressoren Arbeitslose und Einpendler insgesamt wachsen würden, so würden die Ertragszahlen des Firmentickets steigen. Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Modells) die Regressoren Arbeitslose und Einpendler insgesamt einen positiven Einfluss auf die Fahrgeldeinnahmen des Firmentickets (unabhängig von den Arbeitstagen) nehmen.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 2.42, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell - \ 3.4) X - Variablen \ für \ Ex. 3 \ \rightarrow Gesamt \ vor \ EAVs$

Das Regressionsmodell leistet einen sehr guten Erklärungsbeitrag (Korrigiertes R-Quadrat:79,59%) durch die Regressoren. Die Regressoren Arbeitslose (mit einer Irrtumswahrscheinlichkeit von unter 0,1%), Arbeitstage und Einpendler insgesamt (beide mit einer Irrtumswahrscheinlichkeit von 1% bis 5%) sind statistisch signifikant.
Wenn die Werte der Regressoren Arbeitslose und Einpendler insgesamt um einen Wert wachsen würden, so würden die gesamten Fahrgeldeinnahmen fallen. Wenn aber die Werte des Regressors Arbeitstage wachsen würden, so würden die gesamten Fahrgeldeinnahmen steigen.
Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Modells) die Regressoren Arbeitslose und Einpendler insgesamt einen negativen Einfluss auf die gesamten Fahrgeldeinnahmen nehmen.

Gegebenenfalls könnte man allgemein die Preise für Arbeitslose und Einpendlern mindern oder anpassen, um dem entgegen zu wirken.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 2.86, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell - 5.1) X - Variablen\ für\ Ex. 5 \rightarrow Einzelfahrscheine$

Das Regressionsmodell leistet einen mittelmäßigen Erklärungsbeitrag (Korrigiertes R-Quadrat:40,64%) durch die Regressoren. Die Regressoren Arbeitslose, Superbenzin und Stau sind mit einer Irrtumswahrscheinlichkeit von unter 0,1% statistisch signifikant.
Wenn die Werte des Regressors Arbeitslose wachsen würden, so würden die Ertragszahlen der Einzelfahrscheine fallen. Wenn aber die Werte des Regressors Superbenzin und Stau wachsen würden, so würden die Ertragszahlen der Einzelfahrscheine steigen.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 7.59, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann. Für dieses Regressionsmodell wäre das gegebenenfalls weniger geeignet.

$Modell - 5.2) X - Variablen\ für\ Ex. 5 \rightarrow Abo$

Das Regressionsmodell leistet einen sehr guten Erklärungsbeitrag (Korrigiertes R-Quadrat:79,13%) durch die Regressoren. Die Regressoren Arbeitslose, Superbenzin und Stau sind mit einer Irrtumswahrscheinlichkeit von unter 0,1% statistisch signifikant.
Wenn die Werte der Regressoren Arbeitslose, Superbenzin und Stau wachsen würden, so würden die Ertragszahlen des Abos fallen. Man kann also davon ausgehen, dass (unter anderem aufgrund des Erklärungsbeitrags des Modells) die Regressoren Arbeitslose, Superbenzin und Stau einen negativen Einfluss auf die Ertragszahlen des Abos nehmen. Gegebenenfalls könnte man unter anderem die Abo-Preise für Arbeitslose mindern, um dem entgegen zu wirken.

Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 0.282, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann.

$Modell - 5.3) X - Variablen\ für\ Ex. 5 \rightarrow Firmenticket$

Das Regressionsmodell leistet einen durchschnittlichen Erklärungsbeitrag (Korrigiertes R-Quadrat:46,84%) durch die Regressoren. Die Regressoren Superbenzin und Stau sind mit einer Irrtumswahrscheinlichkeit von unter 0,1% statistisch signifikant.
Wenn die Werte der Regressoren Superbenzin und Stau wachsen würden, so würden die Ertragszahlen des Firmentickets fallen.
Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 1.15, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann. Für dieses Regressionsmodell wäre das gegebenenfalls weniger geeignet.

$Modell - 5.4) X - Variablen\ für\ Ex. 5 \rightarrow Berlin\ Ticket\ S$

Das Regressionsmodell leistet einen geringen Erklärungsbeitrag (Korrigiertes R-Quadrat:20,93%) durch die Regressoren. Nur der Regressor Stau ist mit einer Irrtumswahrscheinlichkeit von unter 0,1% statistisch signifikant.
Wenn die Werte des Regressors Stau wachsen würden, so würden die Ertragszahlen des Berlin-Ticket S steigen.
Des Weiteren existiert eine hohe Prognosegenauigkeit mit MAPE = 4.07, womit man vertrauenswürdige Vorhersagen bezüglich der Ertragszahlen ermitteln kann. Für dieses Regressionsmodell wäre das gegebenenfalls weniger geeignet.

Insgesamt wurde also festgestellt, dass durch die multivariate Regressionsanalyse exogener Variablen, außer für die Wetter- und Witterungsdaten, größtenteils gute Schätzungen erzielt werden können. Die meisten geschätzten Regressionsmodelle leisten einen ausreichenden Erklärungsbeitrag der jeweiligen Zielvariablen. Außerdem weisen alle Regressionsmodelle, außer Modell 3.2, eine hohe Prognosegenauigkeit für weiterfolgende Vorhersagen auf.

Abbildungsverzeichnis

Abbildung 1 - Diagramm: Residuum 6

Abbildung 2 - Diagramm: Linearität und Nicht-Linearität 10

Abbildung 1 - Diagramm: Residuendiagramm zur Linearitätsprüfung 11

Abbildung 4 - Diagramm: Homoskedastizität und Heteroskedastizität 12

Abbildung 5 - Diagramm: Normalverteilung 14

Abbildung 6 - Diagramm: QQ-Plot der Normalverteilung 14

Abbildung 7 - Diagramm: Autokorrelation 21

Abbildung 8 - Venn-Diagramm: Keine Multikollinearität 23

Abbildung 9 - Venn-Diagramm: Geringe Multikollinearität 23

Abbildung 10 - Venn-Diagramm: Hohe Multikollinearität 24

Abbildung 11 - Diagramm: Gesamtstreuung SST 32

Abbildung 12 - Diagramm: Erklärte Streuung SSE 33

Abbildung 13 - Diagramm: Nicht erklärte Streuung SSR 33

Abbildung 14 - Diagramm: Bestimmtheitsmaß mit dem Wert 1 34

Abbildung 15 - Diagramm: T -Verteilung für zweiseitigen T -Test 38

Abbildung 16 - Diagramm: OLS-Schätzung und Ausreißer 39

Tabellenverzeichnis

Tabelle 1 - Fehlerarten beim Hypothesentest……27

Literaturverzeichnis

(Backhaus,2021): Backhaus, Klaus, et al. *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*, 16. Vollständig überarbeitete und erweiterte Auflage, 2021, Springer Gabler

(Berlin-Webseite, unbekannt): URL: https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/oeffentlicher-personennahverkehr/ausbau/ [zuletzt zugegriffen am 17.06.2022]

(Chambers, 1983): Chambers, John M, et al. G*raphical Methods for Data Analysis., Bell Laboratories*, CRC Press, First published 1983, Reissued 2018

(Deistler,2018): Deistler, Manfred, et al. *Modelle der Zeitreihenanalyse*, 2018, Birkhäuser Springer International Publishing AG

(Efron,2004): Efron, Bradley, et al. *Least Angle Regression,The Annals of Statistics*, 2004, Vol. 32, No.2, 407-499

(Eid, 2017): Eid, Michael , et al. *Statistik und Forschungsmethoden: Mit Onlinematerialien*, 5. korrigierte Auflage, 2017, Beltz, Originalausgabe Edition

(Everitt, 2011): Everitt, Brian, et al. *An Introduction to Applied Multivariate Analysis with R*, 2011, Springer

(Fahrmeir,2007): Fahrmeir,Ludwig, et al. *Regression: Modelle,Methoden und Anwendungen*, 2007, Springer-Verlag Berlin Heidelberg

(Fahrmeir,2016): Fahrmeir, Ludwig, et al. *Statistik: Der Weg zur Datenanalyse*, 8. überarbeitete und ergänzte Auflage, 2016, Springer Verlag Berlin Heidelberg

(Field,2012): Field, Andy, et al. *Discovering Statistics Using R*, First published, 2012, SAGE Publications Ltd

(Handl,2017): Handl, Andreas, et al. *Multivariate Analysemethoden:Theorie und Praxis mit R*, 3. wesentl. überarbt. Aufl. 2017,Springer-Verlag GmbH Deutschland 2017

(Hastie,2009): Hastie, Trevor, et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, 2009, Springer Science+Business Media

(Hedderich, 2020): Hedderich, Jürgen, et al. *Angewandte Statistik: Methodensammlung mit R*, 17. überarbeitete und ergänzte Auflage, 2020, Springer Spektrum

(Hirschle,2021): Hirschle, Jochen. *Machine Learning für Zeitreihen: Einstieg in Regressions-, ARIMA- und Deep-Learning-Verfahren mit Python*, 2021, Carl Hanser Verlag München

(Hyndman,2010): Hyndman, Rob J. URL: https://robjhyndman.com/hyndsight/arimax/ [zuletzt zugegriffen am 07.06.2022]

(Hyndman,2014): Hyndman, Rob J, et al. *Forecasting: Principles and Practice*, 2014, Published by Otexts.com

(Izenman,2008): Izenman, Alan Julian. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, 2008, Springer Science+Business Media

(James,2013): James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*, 2013, Springer Science+Business Media New York

(Kabacoff,2015): Kabacoff, Robert *I. R in Action: Data analysis and graphics with R.*, Second Edition, 2015, Manning Publications & Co.

(Kessler,2007): Kessler, Waltraud. *Multivariate Datenanalyse: für die Pharma-, Bio- und Prozessanalytik*, 2007, Wiley-VCH Verlag GmbH & Co.KGaA

(Kuckartz,2010): Kuckartz, Udo , et al. *Statistik: Eine verständliche Einführung*, 2010, 1.Auflage 2010, VS Verlag für Sozialwissenschaften, Springer Fachmedien Wiesbaden GmbH 2010

(Makridakis,1998): Makridakis, Spyros, et al. *Forecasting: Methods and Applications*, Third Edition, 1998, John Wiley & Sons, Inc.

(Mankiw,1993): Mankiw, Gregory N. *Makroökonomik*, 1993, Betriebswirtschaftlicher Verlag Dr. Th. Gabler GmbH

(Mertens,2005): Mertens, Peter, et al. *Prognoserechnung*, Sechste völlig neu bearbeitete und erweiterte Auflage,2005,Physica-Verlag Heidelberg

(Moreno,2013): Moreno, Juan Jose Montano, et al. *Using the R-MAPE index as a resistant measure of forecast accuracy*, Psicothema 2013, Vol. 25, No. 4, 500-506

(Neumann,2022): URL: https://www.berliner-zeitung.de/mensch-metropole/aus-berlin-fuer-berlin-die-ersten-neuen-u-bahnen-kommen-2022-ins-rollen-li.225426 [zuletzt zugegriffen am 17.06.2022]

(Sauer, 2019): Sauer, Sebastian. *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren, modellieren und kommunizieren.*, FOM Edition, FOM Hochschule für Ökonomie & Management, Springer Gabler, 2019

(Schlittgen, 2004): Schlittgen, Rainer. *Statistische Auswertungen mit R: Standardmethoden und Alternativen mit ihrer Durchführung in R*, 2004, Oldenbourg Wissenschaftsverlag GmbH

(Schlittgen, 2012): Schlittgen, Rainer. *Einführung in die Statistik: Analyse und Modellierung von Daten*, 12. korrigierte Auflage, 2012, Oldenbourg Verlag München

(Schlittgen,2001):Schlittgen, Rainer, et al. *Zeitreihenanalyse*, 9. unwesentlich veränderte Auflage, 2001, Oldenbourg Wissenschaftsverlag GmbH

(Schlittgen,2013): Schlittgen, Rainer. *Regressionsanalysen mit R*, 2013, Oldenbourg Wissenschaftsverlag GmbH

(Schlittgen,2020):Schlittgen, Rainer, et al. *Angewandte Zeitreihenanalyse mit R*, 4.erweiterte und aktualisierte Auflage,2020,Walter de Gruyter GmbH

(Shardt, 2021): Shardt, Yuri A.W., et al. *Methoden der Statistik und Prozessanalyse: Eine anwendungsorientierte Einführung*, 2021, Springer-Verlag GmbH Deutschland

(Statistikguru-Webseite, 2022): URL: https://statistikguru.de/spss/multiple-lineare-regression/regressionskoeffizienten-interpretieren.html [zuletzt zugegriffen am 01.07.2022]

(Stoetzer,2017): Stoetzer, Matthias-W. *Regressionsanalyse in der empirischen Wirtschafts-und Sozialforschung Band 1:Eine nichtmathematische Einführung mit SPSS und Stata*, 2017, Springer Verlag GmbH Deutschland

(Süß,2022): URL: https://unternehmen.bvg.de/news/zahlenspiegel-fuer-2022-ist-da/ [zuletzt zugegriffen am 17.06.2022]

(Toutenburg,2008): Toutenburg, Helge, et al. *Deskriptive Statistik: Eine Einführung in Methoden und Anwendungen mit R und SPSS*, 6. Aktualisierte und erweiterte Auflage, Springer-Verlag Berlin Heidelberg

(Urban, 2011): Urban, Dieter, et al. *Regressionsanalyse: Theorie, Technik und Anwendung(Studienskripten zur Soziologie)*, 4.überarbeitete und erweiterte Auflage, 2011, VS Verlag für Sozialwissenschaften, Springer Fachmedien Wiesbaden GmbH

(Urban, 2018): Urban, Dieter, et al. *Angewandte Regressionsanalyse: Theorie, Technik und Praxis*, 5. Überarbeitete Auflage, 2018, Springer VS

(Vogel,2015): Vogel, Jürgen. *Prognose von Zeitreihen: Eine Einführung für Wirtschaftswissenschaftler*, 2015, Springer Fachmedien Wiesbaden 2015

(Werth,2021):Werth, Sofia. *Stressfrei zum Bachelor: Effektiv, strukturiert und schnell zur erfolgreichen wissenschaftlichen Arbeit im Studium*, 1.Auflage, 2021

(Wikipedia-Webseite,2022): URL:https://en.wikipedia.org/wiki/Mean_absolute_percentage_error [zuletzt zugegriffen am 30.06.2022]

(Winke, 2020): Winke, J. , et al. *Multiple Regression: Die Regressionsanalyse mit SPSS einfach erklärt*, 2020, Independently published

(Wollschläger, 2014): Wollschläger, Daniel. *Grundlagen der Datenanalyse mit R:Eine anwendungsorientierte Einführung*, 2014, dritte überarbeitete und erweiterte Auflage, Springer-Verlag Berlin Heidelberg

(Zelterman,2015): Zelterman, Daniel. *Applied Multivariate Statistics with R*, 2015, Springer International Publishing Switzerland