Hochschule Heilbronn

Fakultät für Management und Vertrieb

Studiengang Business Analytics, Controlling & Consulting (MAC)

Handhabung ordinalskalierter exogener Variablen in linearen Regressionsmodellen

---

Propädeutik: Master-Thesis

im Sommersemester 2022 (3. Semester)

Prüfer: Prof. Dr. Oliver Schwarz und Prof. Dr. Danny Stadelmayer

Referent: Markus Köhnlein **Matrikelnummer:** 200067

Abgabe: Schwäbisch Hall, den 01.09.2022

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

$BLUE$	Best Linear Unbiased Estimators
df	Freiheitsgrade (*Degrees of Freedome*)
KQ	Kleinste-Quadrate-Methode
MAE	Mean Absolute Error
$MAPE$	Mean Absolute Percentage Error
MSE	Mean Square Error
$OMTC$	Ordinal-Metric-Transformation-Coefficient
$RMSE$	Root Mean Square Error
RSE	Residual Standard Error
SS	Sum of Squares
SSE	Explained Sum of Squares
SSR	Residual Sum of Squared
SST	Total Sum of Squares

Indize- und Parameterverzeichnis

a	Faktor
b	Konstante
i	Laufindex der einzelnen Beobachtungen
J	Anzahl an unabhängigen Variablen
k	Index der unterschiedlichen Merkmalsausprägungen von X
K	Anzahl an unterschiedlichen Merkmalsausprägungen von X
met	metrisch
n, N	Anzahl an Beobachtungen in der Stichprobe
nc	new coding
oc	old coding
ord	ordinalskaliert
$p(x_k)$	relative Häufigkeit der k-ten Merkmalsausprägung von X
$\hat{u}_i$	Residuum der i-ten Merkmalsausprägung
X	exogene Variable
x_i	Merkmalsausprägung von X der i-ten statistischen Einheit
Y	endogene Variable
$\hat{Y}$	Regressionsgerade
$\bar{y}_k$	arithmetisches Mittel von Y aller statistischen Einheiten mit der k-ten Merkmalsausprägung von X
y_i	Merkmalsausprägung von Y der i-ten statistischen Einheit
$\hat{y}_i$	Schätzwert von Y der i-ten statistischen Einheit
$y_{i,k}$	Merkmalsausprägung von Y der i-ten statistischen Einheit mit der k-ten Merkmalsausprägung von X
β_1	Achsenabschnitt (*Intercept*)
β_2	Steigungsparameter

1. Einleitung

1.1 Problemstellung

Lineare Regressionsmodelle gehören zu den wichtigsten Vertretern der klassischen linearen Modelle.[1] Unter den multivariaten Analyseverfahren gehören sie zu den strukturprüfenden Verfahren, da ein linearer Zusammenhang zwischen einer oder mehrerer exogener Variablen auf eine endogene Variable geprüft wird.[2] Lineare Regressionsanalysen eignen sich für die Beschreibung und Erklärung von Zusammenhängen, Wirkungsprognosen, sowie für Zeitreihenanalysen.[3] Bei ihrer Verwendung gelten klare Voraussetzung an das Skalenniveau der Variablen.[4] Sowohl die exogenen, als auch die endogene Variable müssen ein metrisches Skalenniveau aufweisen.[5] Während sowohl für die nominalskalierten als auch für die metrisch skalierten Variablen eine eindeutige Umgangsform in Bezug auf die Verwendung als exogene Variable in linearen Regressionsmodellen besteht, lassen ordinalskalierte Merkmale unterschiedliche Handhabungen zu. Diese sollen in dieser Arbeit thematisiert werden.[6] Im Allgemeinen können ordinalskalierte Merkmale entweder als nominalskalierte oder als metrische Variable angenommen werden. Beide Handhabungsformen haben ihre Vor- und Nachteile. Es wird aufgeführt, welche Vorgehensweise unter welchen Umständen zu empfehlen ist. Zudem wird eine eigene Umgangsform für ordinalskalierte Merkmale in linearen Regressionsmodellen erarbeitet.

1.2 Zielsetzung

Die hier gestellte Forschungsfrage lautet, welche Eigenschaften ordinalskalierte Merkmale aufweisen, damit sie sich für die eine oder andere Umgangsform in linearen Regressionsmodellen anbieten und welche Empfehlungen sich daraus ergeben. Zielsetzung dieser Arbeit ist es zudem die Gütekriterien der Regressionsmodelle, in denen ein ordinalskaliertes exogenes Merkmal enthalten ist, zu verbessern, indem die Kodierungen der Merkmalsausprägungen der exogenen Variablen angepasst werden. Die Umkodierung des ordinalskalierten Merkmals, wird verkürzt auch für eine

[1] Vgl. Backhaus, et al. (2021), S. 62.
[2] Vgl. ebd., S. 12 f.
[3] Vgl. Berekhoven, Eckert, Ellenrieder (2009), S. 198, vgl. Backhaus, et al. (2021), S .14, 64.
[4] Vgl. Backhaus, et al. (2021), S. 13.
[5] Vgl. ebd., S. 158.
[6] Vgl. ebd., S. 14.

multiple lineare Regression untersucht. Betrachtet man die Eignung oder Nichteignung eines ordinalskalierten Merkmals für die Verwendung in einem linearen Regressionsmodell, stellt sich die Frage, wie geeignet oder ungeeignet die betrachtete Variable ist. Dies wird in dieser Arbeit anhand einer eigens entwickelten Kennzahl dargestellt. Zum Ende dieser Arbeit soll deutlich geworden sein, wie mit der Problemstellung, ordinalskalierter Variablen in linearen Regressionsmodellen umzugehen ist.

1.3 Abgrenzung des Themas

In dieser Arbeit soll die Handhabung ordinalskalierter exogener Variablen in linearen Regressionsmodellen untersucht werden. Anderweitige Skalenniveaus wie nominalskalierte oder metrische Merkmale werden hierbei nicht weiter berücksichtigt. Zum Verständnis sollten beim Leser Grundkenntnisse über lineare Regressionsmodelle vorhanden sein, da auf die behandelte Rechenmethode der linearen Einfachregression, sowie der multiplen Regression nur verkürzt eingegangen wird. Zur Untersuchung des Einflusses eines ordinalskalierten Merkmals auf die endogene Variable wird auf einfache lineare Regressionsmodelle zurückgegriffen.[7] Zum Ende hin wird deren Einfluss in einer multiplen linearen Regression dargestellt.[8] Die nichtlinearen Regressionsmodelle sind kein Bestandteil dieser Ausarbeitung.[9] Auf alternative lineare Regressionsmodelle wie beispielsweise die logistische Regression wird in dieser Arbeit nicht eingegangen.[10] Zudem werden lediglich Regressionsmodelle mit metrischen endogenen Variablen betrachtet. Für nicht-metrische Skalenniveaus der endogenen Variablen stehen alternative Modelle zur Verfügung wie beispielsweise *Logit Modelle* oder die *Poisson Regression*, auf die in dieser Arbeit nicht eingegangen wird.[11] Zudem ist das hier behandelte Themengebiet nicht mit der ordinalen Regression zu verwechseln, die sich mit Regressionsmodellen mit einer ordinalskalierten endogenen Variable beschäftigt.[12]

[7] Vgl. Backhaus, et al. (2021), S. 64.
[8] Vgl. ebd., S. 81.
[9] Vgl. Backhaus, Erichson, Weiber (2015), S. 24.
[10] Vgl. Backhaus, et al. (2021), S. 290 ff.
[11] Vgl. Hedderich, Sachs (2020), S. 866, vgl. Backhaus, et al. (2021), S. 301 ff., vgl. Wollschläger (2020), S. 377.
[12] Vgl. Wollschläger (2020), S. 365.

1.4 Aufbau der Arbeit

Diese Arbeit ist folgendermaßen aufgebaut: In *Kapitel 2* werden die unterschiedlichen Skalenniveaus der exogenen Variablen thematisiert, sowie die unterschiedlichen Umgangsformen ordinalskalierter Merkmale in linearen Regressionsmodellen. Es wird der *Root Mean Square Error*, sowie die Eigenschaften der ordinalskalierten Variablen und die Vor- und Nachteile der unterschiedlichen Umgangsformen mit diesen thematisiert. In *Kapitel 3* wird ein allgemeingültiger Lösungsansatz, unter welchen Umständen welche Umgangsform zu empfehlen ist, dargestellt. Zudem wird aufgezeigt, wie eine Umkodierung des ordinalen Merkmals, die in *Kapitel 4.1* behandelt wird, zu einer besseren Modellanpassung führen kann. Darüber hinaus wird in *Kapitel 4.2* eine Kennzahl entwickelt, die die Eignung des ordinalen Merkmals für die Verwendung als metrische Variable in einem linearen Regressionsmodell abbildet. In *Kapitel 4.3* wird die Umkodierung bei einem multiplen Regressionsmodell angerissen, bevor es in *Kapitel 5* in eine Schlussbetrachtung des Themas geht.

Parallel zu dieser Arbeit wird ein Skript für die Verwendung in *R-Studio* angefertigt, das alle hier aufgezeigten Analysen und Modelle, sowie Tabellen und Darstellungen beinhaltet. Alle Inhalte dieser Arbeit lassen sich mit dem Skript reproduzieren. Dieses wird dieser Ausarbeitung beigelegt.[13] In den folgenden Ausführungen wird des Öfteren Bezug auf das *R*-Skript genommen, mit Verweisen auf die Codezeilen, auf denen sich die relevanten Befehle wiederfinden. Die hier verwendeten Datensätze entstehen jeweils durch randomisierte Werte direkt im *R*-Skript und können somit leicht reproduziert werden. Durch einen `set.seed`-Befehl wird gewährleistet, dass bei wiederholter Durchführung des *R*-Skripts exakt dieselben Zufallswerte entstehen.[14] Für die Ausführung des *R*-Skriptes sollten unten genannte Pakete installiert und ausgeführt werden.[15] Zusätzlich zu dem beigelegten *R*-Skript wurde ein Paket für *R*-Studio programmiert und über *GitHub* veröffentlicht. Die Nutzungsmöglichkeiten dieses Paketes, sowie dessen Funktionen werden im Anhang näher beschrieben.[16]

[13] Siehe: *Anhang*, S. VII.
[14] Vgl. Wollschläger (2020), S. 223.
[15] Pakete: *`mime`, `rmarkdown`, `tinytex`, `car`, `reshape`, `boot`, `devtools`, `ggplot2`, `ggfortify`*.
[16] Siehe: *Anhang*, S. VII.

2. Grundlagen der Arbeit

2.1 Skalenniveaus der Variablen in linearen Regressionsmodellen

Bei empirischen Untersuchungen mit Regressionsmodellen geht man von einer geteilten Variablenmenge aus.[17] Das bedeutet, dass die vorliegenden Variablen in eine oder mehrere unabhängige Variablen und meist eine abhängige Variable unterteilt werden können.[18] Als exogene Variablen bezeichnet man hierbei die unabhängigen bzw. die erklärenden Variablen des Datensatzes.[19] Mithilfe der exogenen Variablen sollen kausale Zusammenhänge zu den endogenen, bzw. abhängigen Variable analysiert werden.[20] Durch den Zusammenhang zwischen den x-Variablen und der y-Variable spricht man auch oft von Dependenzanalysen.[21] In den hier betrachteten linearen Regressionsmodellen wird ein linearer Zusammenhang zwischen der oder den exogenen und der endogenen Variablen unterstellt.[22] Die geteilte Variablenmenge ist in *Tabelle 1* mit ihren jeweiligen Begrifflichkeiten dargestellt:

Tabelle 1: Geteilte Variablenmenge (Dependenzanalyse)

Unabhängige Variablen	Abhängige Variable
$X_1, X_2, X_3, \dots$	Y
x-Variablen	y-Variable
erklärende Variablen	erklärte Variable
Prädiktor-Variablen	Response-Variable
Input-Variablen	Output-Variable
Regressoren	Regressand
Kovariablen	Prognosevariable

Quelle: Backhaus, et al. (2021), S. 6, 64.

In dieser Arbeit soll das Skalenniveau der exogenen Variablen thematisiert werden. Unter den exogenen Variablen, die zur Prognose einer endogenen Variablen verwendet werden, unterscheidet man zwischen folgenden Skalenniveaus:[23]

[17] Vgl. Backhaus, et al. (2021), S. 6.
[18] Vgl. ebd.
[19] Vgl. ebd.
[20] Vgl. ebd., S. 45.
[21] Vgl. ebd., S. 6.
[22] Vgl. ebd.
[23] Vgl. Meffert, Burmann, Kirchgeorg (2015), S. 142 ff., vgl. Backhaus, et al. (2021), S. 7 ff.

- Bei qualitativen Merkmalen kann nur die Gleichheit oder Ungleichheit verschiedener Merkmalsausprägungen unterschieden werden. Diese werden lediglich benannt und lassen sich nicht in eine logische Reihenfolge bringen. Sie werden daher auf einer Nominalskala abgebildet. Hierzu gehören Merkmale wie beispielsweise das Geschlecht, die Religion, der Beruf, Farben, aber auch quantitative Merkmale ohne definierte Reihenfolge wie die Postleitzahl.
- Liegt ein komparatives Merkmal vor, dann können die einzelnen Merkmalsausprägungen in eine logische Reihenfolge gebracht werden. Die Abstände zwischen den Merkmalsausprägungen lassen sich nicht berechnen oder interpretieren. Komparative Merkmale werden auf einer Ordinalskala abgebildet und werden daher auch als ordinalskalierte Merkmale bezeichnet. Die Ordinalskala ermöglicht die Aufstellung einer Rangordnung mithilfe von Rangwerten, die sich in der Kodierung des Merkmals widerspiegeln. Diese muss der Rangordnung des Merkmals entsprechen. Zu ordinalskalierten Variablen gehören beispielsweise Merkmale wie ein Rating-Urteil oder der Schulabschluss. Es können zudem auch quantitative Merkmale ordinalskaliert sein, wie beispielsweise Schulnoten.
- Quantitative bzw. kardinalskalierte Merkmale werden auf metrischen Skalen abgebildet. Man unterscheidet zwischen intervallskalierten und verhältnisskalierten Merkmalen. Eine Intervallskala lässt die Interpretation der Abstände einzelner Merkmalsausprägungen zu. Hierzu zählen Merkmale wie die Temperatur oder das Geburtsjahr. Da hierbei zwar Abstände interpretiert werden können, jedoch kein logischer Nullpunkt besteht, lassen sich keine Verhältnisse zwischen den Merkmalsausprägungen bilden. Handelt es sich um ein quantitatives Merkmal, dass zudem einen natürlichen Nullpunkt aufweist, spricht man von einem verhältnisskalierten Merkmal. Diese Variablen werden auf einer Ratio- oder auch Verhältnisskala abgebildet. Hierbei können nicht nur Abstände, sondern auch die Verhältnisse zwischen verschiedenen Merkmalsausprägungen betrachtet werden. Hierzu zählen Merkmale wie beispielsweise der Preis, Umsatz, Einkommen oder Alter.

Die unterschiedlichen Skalenniveaus der exogenen Variablen sind in *Tabelle 2* dargestellt.

Tabelle 2: Skalenniveau von Variablen

Skala		**Merkmale**	**Mögliche rechnerische Handhabung**
Nichtmetrisch Skalen (kategorial)	Nominalskala	Klassifizierung qualitativer Eigenschaftsausprägungen	Bildung von Häufigkeiten, Modus
	Ordinalskala	Rangwert mit Ordinalzahlen	Median, Quantile
Metrische Skalen (kardinal)	Intervallskala	Skala mit gleichgroßen Abschnitten ohne natürlichen Nullpunkt	Subtraktion, Mittelwert, Standardabweichung, Korrelation, t-Test, F-Test
	Ratioskala	Skala mit gleichgroßen Abschnitten und natürlichem Nullpunkt	Summe, Division, Multiplikation, geometrisches Mittel, harmonisches Mittel, Varianzkoeffizient

Quelle: Backhaus, et al. (2021), S. 8.

Durch das Skalenniveau der Variablen ergibt sich sowohl der Informationsgehalt der Daten als auch die Anwendbarkeit verschiedener Rechenoperationen.[24] In *Tabelle 2* sind hierzu die erlaubten Rechenoperationen zur statistischen Auswertung aufgelistet. Die unterschiedlichen Skalenniveaus stehen zueinander in einer Hierarchie. Somit sind ordinalskalierte Merkmale immer auch nominalskaliert, jedoch nicht umgekehrt. Genauso sind alle metrisch skalierten Merkmale auch ordinal- und nominalskaliert. Es besteht die Möglichkeit Daten von einem höheren Skalenniveau auf ein niedrigeres Skalenniveau zu transformieren, jedoch nicht auf ein höheres.[25] Dies geschieht häufig durch die Bildung von klassierten Daten, wie beispielsweise Einkommensklassen.[26] Umso höher das Skalenniveau des Merkmals, desto höher ist auch dessen Informationsgehalt und desto mehr Rechenoperationen lassen sich für die statistische Auswertung anwenden.[27] Zur Feststellung des Skalenniveaus einer Variablen, muss diese auf die oben beschriebenen Eigenschaften der einzelnen

[24] Vgl. Backhaus, et al. (2021), S. 8.
[25] Vgl. Berekhoven, Eckert, Ellenrieder (2009), S. 66.
[26] Vgl. Backhaus, et al. (2021), S. 10.
[27] Vgl. ebd., S. 10.

Skalenniveaus untersucht werden. Bei Befragungen in der Marktforschung kommen häufig sogenannte Ratingskalen zum Einsatz.[28] Sie dienen der Abfrage der Einschätzung der Befragten zu bestimmten Aussagen. Man unterscheidet hierbei zwischen Bewertungsskalen, Wichtigkeitsskalen, Intensitätsskalen und Zustimmungsskalen.[29] Hierzu gehören auch Sonderformen von Ratingskalen wie die *Likert-Skalen*, oder das *Semantische Differential*.[30] Eine Fragestellung einer Rating-Skala könnte beispielsweise wie in *Abbildung 1* aussehen.

Abbildung 1: Rating-Skala

Wie stark stimmen Sie der folgenden Aussage zu? „…"?			
☐	☐	☐	☐
Stimme voll und ganz zu	Stimme eher zu	Stimme eher nicht zu	Stimme gar nicht zu

Quelle: Berekhoven, Eckert, Ellenrieder (2009), S. 69.

Bei der hier dargestellten Rating-Skala handelt es sich um eine ordinalskalierte Variable, da die Antwortmöglichkeiten in eine logische Reihenfolge gebracht werden können, die Abstände zwischen den Merkmalsausprägungen jedoch nicht äquidistant (gleich groß) sind.[31] Hierbei werden den einzelnen Antwortkategorien in der Regel natürliche Zahlen zugeordnet, die dem Skalenwerten des Merkmals entsprechen.[32]

2.2 Umgangsformen ordinalskalierter exogener Variablen

Voraussetzung für die Verwendung eines Merkmals in einem linearen Regressionsmodell ist dessen metrisches Skalenniveau.[33] Nominalskalierte und ordinalskalierte Merkmale können daher nicht ohne differenzierte Betrachtung in linearen Regressionsmodellen verwendet werden. Kommen nominalskalierte Merkmale zum Einsatz, wird häufig auf sogenannte Dummy-Variablen zurückgegriffen.[34] Hierbei wird für jede Merkmalsausprägung des qualitativen Merkmals eine eigene dichotome

[28] Vgl. Backhaus, et al. (2021), S. 10.
[29] Vgl. ebd.
[30] Vgl. Berekhoven, Eckert, Ellenrieder (2009), S. 68, vgl. Hedderich, Sachs (2020), S. 845.
[31] Vgl. Backhaus, et al. (2021), S. 11.
[32] Vgl. Berekhoven, Eckert, Ellenrieder (2009), S. 64.
[33] Vgl. Backhaus, et al. (2021), S. 64.
[34] Vgl. Janssen, Laatz (2017), S. 439.

Variable erstellt.[35] Dichotome Variablen haben lediglich zwei Merkmalsausprägungen.[36] Liegt die jeweilige Merkmalsausprägung vor, dann ist der Variablenwert für die statistische Einheit 1. Für alle anderen ist er 0.[37] Bei der Kodierung mit den Zahlen 0 und 1 spricht man auch von binären Variablen oder Dummy-Variablen.[38] Wird eine Dummy-Variable in einem Regressionsmodell verwendet, kann auf eine der Merkmalsausprägungen verzichtet werden, deren Einfluss auf die endogene Variable in den Achsenabschnitt einfließt.[39] Die Umkodierung der einzelnen ordinalskalierten Merkmalsausprägungen kann beispielsweise wie in *Tabelle 3* erfolgen.

Tabelle 3: Dummy-Variablen des komparativen Merkmals: Schulabschluss

Statistische Einheit	Merkmal: Schulabschluss	Hauptschul-abschluss	Realschulab-schluss	Abitur
1	Hauptschulabschluss	1	0	0
2	Realschulabschluss	0	1	0
3	Realschulabschluss	0	1	0
4	Hauptschulabschluss	1	0	0
5	Hauptschulabschluss	1	0	0
6	Abitur	0	0	1
7	Realschulabschluss	0	1	0
8	Hauptschulabschluss	1	0	0
9	Realschulabschluss	0	1	0
10	Abitur	0	0	1
…	…	…	…	…

Quelle: Eigene Darstellung, in Anlehnung an Backhaus, et al. (2021), S. 153.

Die Umwandlung zu mehreren Dummy-Variablen sorgt dafür, dass das ordinalskalierte Merkmal wie eine metrische Variable in einem Regressionsmodell behandelt werden kann.[40] Alternativ kann die letzte Merkmalsausprägung auch mit dem Wert -1 kodiert werden. Dadurch kann der Einfluss jeder Merkmalsausprägung auf die

[35] Vgl. Backhaus, et al. (2021), S. 11.
[36] Vgl. Hedderich, Sachs (2020), S. 839.
[37] Vgl. Backhaus, et al. (2021), S. 149.
[38] Vgl. ebd., S. 11.
[39] Vgl. Janssen, Laatz (2017), S. 439.
[40] Vgl. Backhaus, et al. (2021), S. 11.

endogene Variable auf Größe und Signifikanz untersucht werden.[41] In dem beigelegten *R*-Skript werden vereinfacht die ordinalskalierten Merkmale zur `character`-Variable umgewandelt, um keine aufwendige Umkodierung vornehmen zu müssen.[42] Dies führt dazu, dass der Parameter der ersten Merkmalsausprägung nicht direkt ausgegeben wird, sondern lediglich der Achsenabschnitt des linearen Modells. Hierauf wird in *Kapitel 2.4* näher eingegangen. Alternativ hierzu kann das ordinalskalierte Merkmal als metrische Variable angenommen werden. In diesem Fall erhält man bei der linearen Einfachregression lediglich zwei Parameter; je einen für Steigung und Achsenabschnitt und nicht einen für jede Merkmalsausprägung der exogenen Variable.[43] Ordinalskalierte Merkmale stehen zwischen den nominalskalierten und den metrisch skalierten Variablen, da die Merkmalsausprägungen zwar in eine Reihenfolge gebracht werde können, wie es bei metrischen Merkmalen der Fall ist, diese jedoch nicht auf einer Kardinalskala abgebildet werden können, da die Abstände nicht berechnet werden können.[44] Liegt ein ordinalskaliertes Merkmal vor, bestehen die beiden Umgangsformen der nominalskalierten und der metrischen Variablen.[45] Zum einen kann das ordinalskalierte Merkmal zu mehreren Dummy-Variablen umgewandelt werden, wie man es bei qualitativen Merkmalen tut. Die dabei vollzogene Transformation auf ein niedrigeres Skalenniveau ist zulässig, verursacht jedoch einen Informationsverlust.[46] Zum anderen kann man die Variable wie ein metrisch skaliertes Merkmal in das Regressionsmodell einfließen lassen.[47] Dabei wird die Ordinalskala als quasi-metrische Skala angenommen, wobei man die Abstände zwischen den Merkmalsausprägungen als gleich groß annimmt, da diese auch auf dem Fragebogen wie in *Abbildung 1* gleich groß dargestellt werden.[48] Die hierbei vollzogene Transformation auf ein höheres Skalenniveau ist genaugenommen nicht zulässig, wird jedoch in der Praxis oft aus Vereinfachungsgründen vollzogen.[49] Das ordinalskalierte Merkmal weist in den meisten Fällen eine quantitative Kodierung auf, bei der

[41] Vgl. Schlittgen, Sattarhoff (2020), S. 44.
[42] Siehe: *R*-Skript, Zeile: 85.
[43] Vgl. Backhaus, et al. (2021), S. 65.
[44] Vgl. ebd., S. 9.
[45] Vgl. ebd.
[46] Vgl. ebd., S. 9 f.
[47] Vgl. Meffert, Burmann, Kirchgeorg (2015), S. 144.
[48] Vgl. Berekhoven, Eckert, Ellenrieder (2009), S. 68.
[49] Vgl. Backhaus, et al. (2021), S. 9.

auf die natürlichen Zahlen beginnend mit 1 zurückgegriffen wird. Bei dieser Kodierung muss die Reihenfolge der Rangwerte der Reihenfolge der Kodierung entsprechen.[50] Um das ordinalskalierte Merkmal als metrisch anzunehmen, wird die Kodierung des ordinalskalierten Merkmals als eine metrische Größe angenommen.

2.3 Root Mean Square Error als Kennzahl der Fehleranalyse

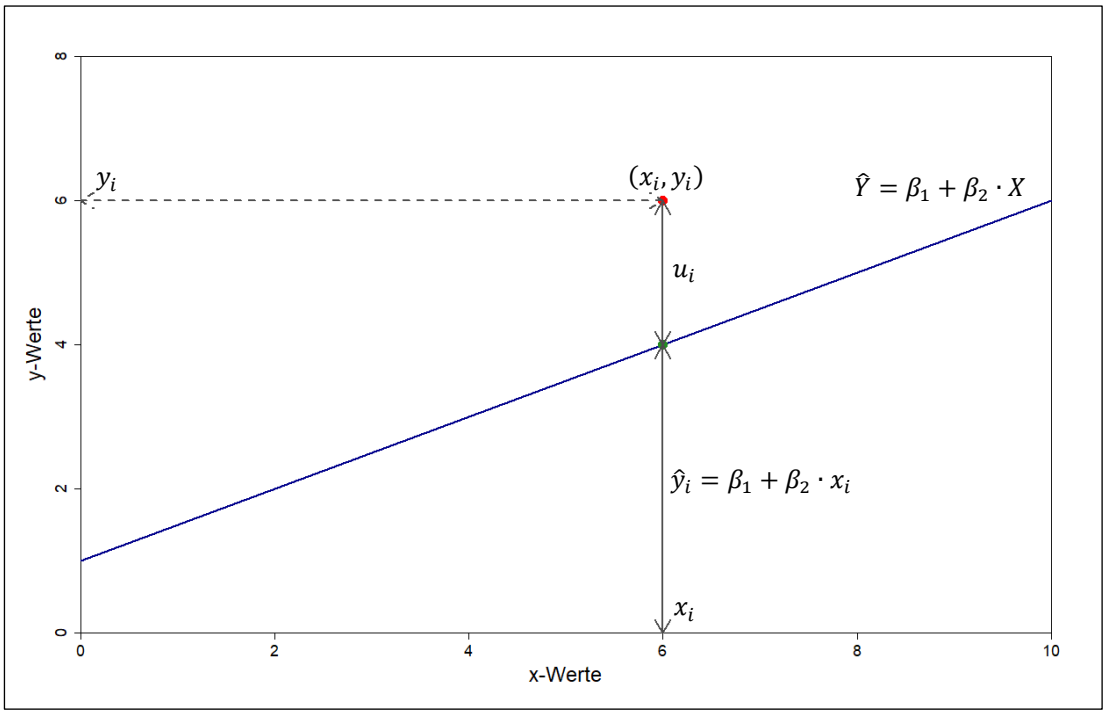
Der *Root Mean Square Error* (*RMSE*), oder auch *Standardfehler der Schätzung* (*SEE*) ist eine der wesentlichen Fehlerkennzahlen für lineare Regressionsmodelle.[51] Diese wird oftmals auch als *Root MSE* oder *Residual Standard Error* (*RSE*) bezeichnet.[52] Der *RMSE* berechnet sich folgendermaßen:

$$RMSE = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2} \qquad (2.1)^{53}$$

Als $\hat{u}_i$ werden dabei die Residuen bezeichnet. Sie berechnen sich aus der Differenz der Beobachtungswerte und der Schätzwerte des Regressionsmodells für die i-te statistische Einheit.[54] Diese ergeben sich aus Einflüssen auf die endogene Variable, die im Modell nicht berücksichtigt werden.[55] In *Abbildung 2* ist zu erkennen, wie sich das Residuum als vertikaler Abstand zwischen der Beobachtung in rot und dem Schätzwert in grün ergibt. Die Summe der Residuenquadrate in der hier dargestellten *Formel 2.1* wird als *Sum of Squares* bezeichnet.[56] Die Residuen werden quadriert, da sich positive und negative Residuen ansonsten exakt aufheben würden und in der Summe 0 ergäben.[57] Laut der Methode der kleinsten Quadrate (*KQ-Methode*) soll die Summe der quadrierten Residuen so klein wie möglich sein.[58]

[50] Vgl. Hedderich, Sachs (2020), S. 27.
[51] Vgl. Backhaus, et al. (2021), S. 85.
[52] Vgl. ebd.
[53] Quelle: ebd.
[54] Vgl. Backhaus, et al. (2021), S. 77.
[55] Vgl. ebd.
[56] Vgl. ebd., S. 78.
[57] Vgl. Kronthaler, Zöllner (2021), S. 90.
[58] Vgl. Hedderich, Sachs (2020), S. 136.

Abbildung 2: Regressionsgerade und Residuum



Quelle: In Anlehnung an Backhaus, et al. (2021), S. 78, siehe: *R*-Skript, Zeile: 46-62.

Das von *Carl Friedrich Gauß* entwickelte Verfahren zeigt auf, dass die Regressionsgerade die besten linearen unverzerrten Schätzer liefert.[59] Dies wird als *BLUE*-Eigenschaft bezeichnet (*Best Linear Unbiased Estimators*).[60] Das ermittelte Modell hat somit die kleinstmögliche Varianz und damit die höchste Präzision.[61] Werden die *Sum of Squares* durch $n - k - 1$ geteilt, erhält man den *Mean Square Error* (*MSE*).[62] Die Größe n beschreibt dabei die Anzahl an Beobachtungen und k ist die Anzahl der Regressoren, bzw. die Anzahl an zu schätzenden Parametern ohne den Achsenabschnitt.[63] Im Nenner steht somit die Anzahl der Freiheitsgrade der Schätzung.[64] Zieht man die Wurzel aus dem *Mean Square Error* erhält man den *Root Mean Square Error*.[65] Die Einheit des *RMSE* ist dieselbe wie die der endogenen Variable.[66]

[59] Vgl. Backhaus, et al. (2021), S. 79.
[60] Vgl. Hedderich, Sachs (2020), S. 825.
[61] Vgl. Backhaus, et al. (2021), S. 102.
[62] Vgl. ebd., S. 78.
[63] Vgl. ebd., S. 86.
[64] Vgl. ebd.
[65] Vgl. ebd., S. 85.
[66] Vgl. ebd., S. 86.

Teilt man den *RMSE* zusätzlich durch den Durchschnitt der endogenen Variablen, kann eine Aussage darüber getroffen werden, wie hoch der *RMSE* in Bezug auf die endogene Variable ausgeprägt ist.[67]

$$rel.RMSE = \frac{RMSE}{\bar{y}} \cdot 100\% \qquad (2.2)^{68}$$

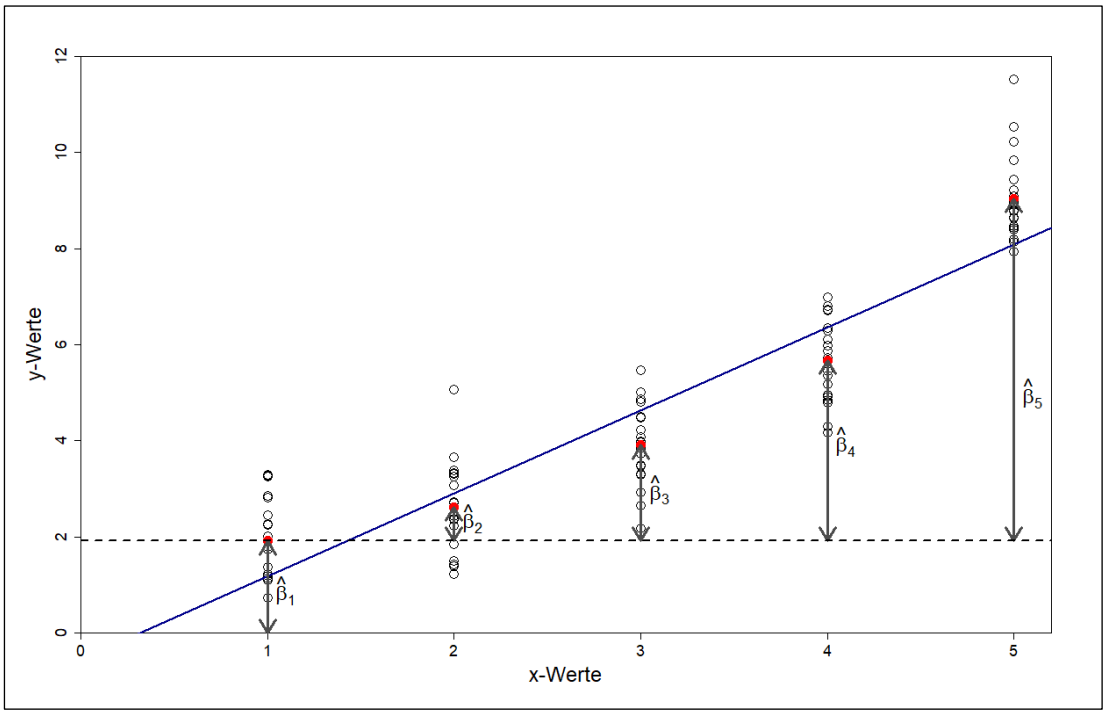
Bei der Betrachtung des *Root Mean Square Errors* muss die unterschiedliche Behandlung von Test- und Trainingsdaten beachtet werden. Für Trainingsdaten wird für die Berechnung des *RMSE* die *Sum of Squares* durch $n - k - 1$ geteilt.[69] Hingegen wird bei der Berechnung des *RMSE* für Testdaten lediglich durch n geteilt, da hierbei ein bestehendes Modell getestet wird und keine Parameter geschätzt werden müssen.[70] Dies gilt auch, wenn das Modell für Prognosen verwendet wird.[71] *R-Studio* nutzt hier standardmäßig den *RMSE* für Trainingsdaten.[72] Unterhalb des Bruchstriches stehen somit die Freiheitsgrade des Modells.[73] Im Folgenden werden beide Kennzahlen thematisiert.

2.4 Zusammenhang zwischen Koeffizienten und bedingten Mittelwerten

Um den Zusammenhang zwischen Koeffizienten und bedingten Mittelwerten herauszustellen, wird eine Stichprobe mit randomisierten Werten erstellt.[74] Die ordinalskalierte exogene Variable besteht aus fünf Merkmalsausprägungen. Die einzelnen Beobachtungen werden in *Abbildung 3* in schwarz dargestellt. Die Schätzwerte der Dummy-Variablen sind in rot dargestellt. Die Regressionsgerade in blau stellt das Ergebnis dar, wenn das ordinalskalierte Merkmal als ein metrisch skaliertes Merkmal angenommen wird. Die Kodierung des Merkmals entspricht in diesem Fall einer metrischen Größe. In diesem Beispiel wurden für die Kodierung die natürlichen Zahlen beginnend mit 1 gewählt.

[67] Vgl. Backhaus, et al. (2021), S. 86.
[68] Quelle: ebd.
[69] Vgl. ebd.
[70] Vgl. Janssen, Laatz (2017), S. 452.
[71] Vgl. ebd., S. 555.
[72] Vgl. Wollschläger (2020), S. 241.
[73] Vgl. Backhaus, et al. (2021), S. 23 f.
[74] Siehe: *R*-Skript, Zeile: 71-83.

Abbildung 3: Ordinalskaliertes Merkmal mit fünf Merkmalsausprägungen



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 65-122.

Die bedingten Mittelwerte der jeweiligen Merkmalsausprägung von X entsprechen den Koeffizienten der Regression, wenn das ordinalskalierte Merkmal zu Dummy-Variablen umgewandelt wird.[75] Hierbei ergibt sich der in *Tabelle 4* dargestellte Zusammenhang zwischen den Koeffizienten der Dummy-Variablen und den Mittelwerten der einzelnen Merkmalsausprägungen.

Tabelle 4: Zusammenhang zwischen bedingten Mittelwerten und Koeffizienten

x_k	$\bar{y}_k$	*Koeffizienten*	
1	$\bar{y}_1 = \beta_1 = Intercept = 1{,}919$	β_1	1,919
2	$\bar{y}_2 = \beta_1 + \beta_2 = 2{,}625$	β_2	0,706
3	$\bar{y}_3 = \beta_1 + \beta_3 = 3{,}916$	β_3	1,997
4	$\bar{y}_4 = \beta_1 + \beta_4 = 5{,}678$	β_4	3,759
5	$\bar{y}_5 = \beta_1 + \beta_5 = 9{,}041$	β_5	7,122

Quelle: Eigene Darstellung.

[75] Vgl. Hedderich, Sachs (2020), S. 841.

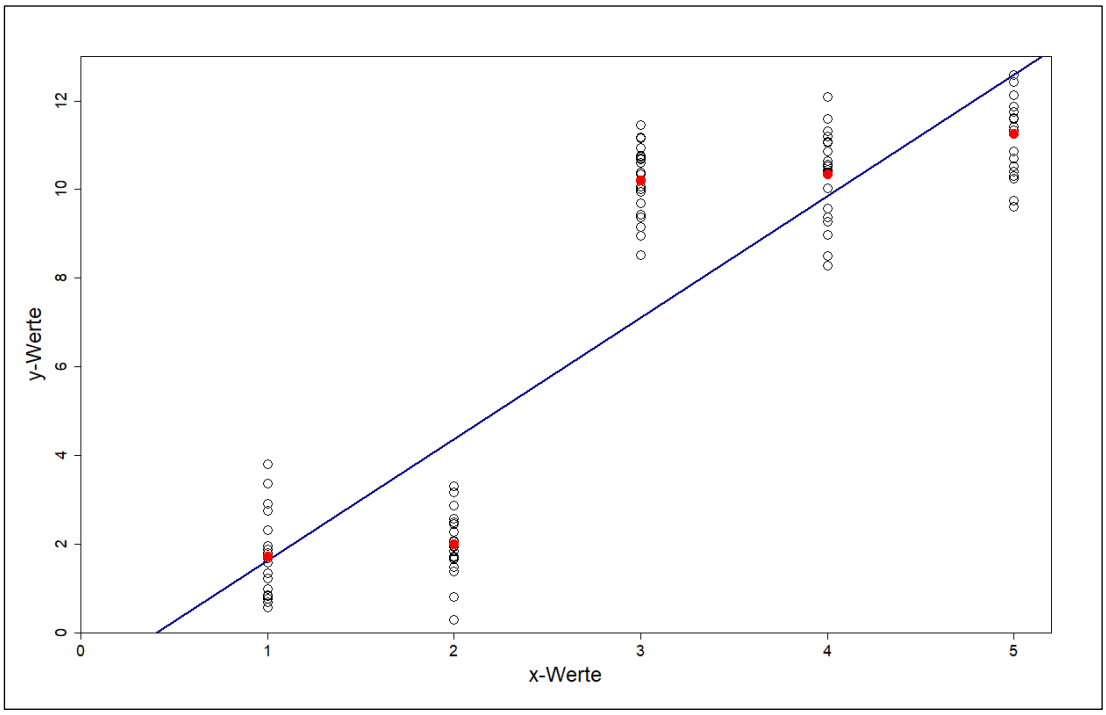
Es ist zu erkennen, dass die Dummy-Variablen eine genauere Schätzung der endogenen Variablen zulassen, da die einzelnen Sprünge zwischen den bedingten Mittelwerten exakt abgebildet werden können. In *Tabelle 4* ist zudem zu erkennen, dass das Regressionsmodell mit Dummy-Variablen die Schätzwerte nicht direkt ausgibt, sondern man den Achsenabschnitt (β_1) und den Koeffizienten der jeweiligen Merkmalsausprägung (β_i) addiert.[76] Dieser Zusammenhang wird in *Abbildung 3* durch die senkrechten Pfeile verdeutlicht.

2.5 Eigenschaften ordinalskalierter exogener Variablen

Da der *RMSE* minimiert werden soll, gilt es im Folgenden zu ermitteln, bei welcher Vorgehensweise der niedrigere *RMSE* erzielt werden kann.[77] In der Regel wird das bei der Umwandlung zu Dummy-Variablen der Fall sein, da diese die Sprünge zwischen den Beobachtungspunkten der einzelnen Merkmalsausprägungen von X abbilden können. In *Abbildung 4* ist ein Streudiagramm zu sehen, dass den Einfluss des ordinalskalierten Merkmals auf der x-Achse auf die endogene Variable auf der y-Achse abbildet.[78] Die schwarzen Kreise repräsentieren die einzelnen Beobachtungen. Die roten Punkte stehen für die Schätzwerte der linearen Einfachregression bei einer Umwandlung zu Dummy-Variablen. Die blaue Gerade ist die Regressionsgerade, die entsteht, wenn das ordinalskalierte Merkmal als metrisch skaliert angenommen wird.[79] Es ist zu erkennen, dass sich das ordinalskalierte Merkmal für die Umgangsform als metrisches Merkmal nicht eignet, da die Merkmalsausprägungen einen großen Abstand zur dazugehörigen Regressionsgeraden aufweisen. Vor allem bei der zweiten und dritten Merkmalausprägung von X werden äußerst schlechte Schätzergebnisse durch das als metrisch angenommene Regressionsmodell erzielt. Diese Variable ist daher in Dummy-Variablen umzuwandeln.

[76] Vgl. Janssen, Laatz (2017), S. 440, vgl. Hedderich, Sachs (2020), S. 842.
[77] Vgl. Backhaus, et al. (2021), S. 85.
[78] Vgl. ebd., S. 52.
[79] Vgl. ebd.

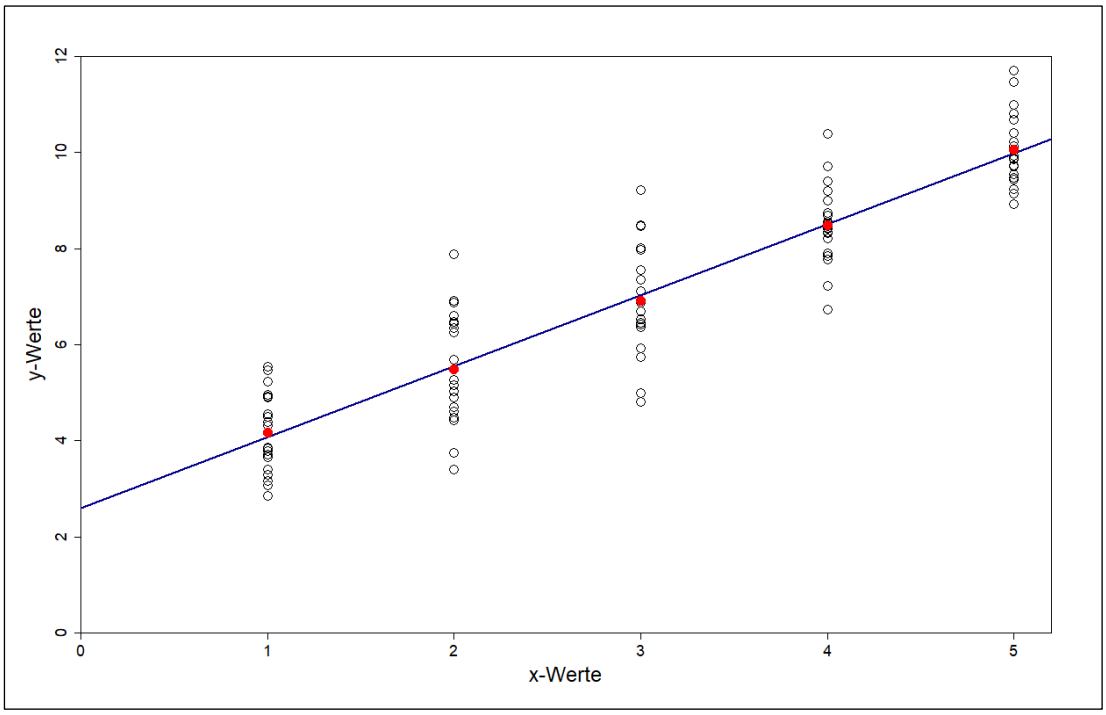
Abbildung 4: Ungeeignetes ordinalskaliertes Merkmal als metrische Variable



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 125-160.

Liegen hingegen alle Schätzwerte der Dummy-Variablen näherungsweise auf einer Geraden, wie es in *Abbildung 5* dargestellt ist, können diese Sprünge auch mit der Annahme als metrische Variable erzielt werden. In diesem Beispiel wurden die Merkmalausprägungen so festgelegt, dass die Schätzwerte der Dummy-Variablen möglichst exakt eine Gerade abbilden. Ob das ordinalskalierte Merkmal somit als metrische Variable oder als qualitative Variable angenommen wird, macht in diesem Beispiel im Hinblick auf den *RMSE* nahezu keinen Unterschied und ist somit variabel. In *Tabelle 5* sind die Fehlerkennzahlen *Sum of Squares* und *Root Mean Square Error* beider Vorgehensweisen für die Zahlenbespiele aus *Abbildung 4* und *Abbildung 5* dargestellt. In *Abbildung 4* gibt das Modell mit Dummy-Variablen den deutlich niedrigeren *RMSE* aus, da die Regressionsgerade die Sprünge zwischen den Beobachtungswerten der jeweiligen Merkmalsausprägung von X nicht abbilden kann. Bei den Fehlerkennzahlen zu *Abbildung 5* ist zu erkennen, dass die *Sum of Squares* sowohl bei der metrischen Annahme als auch bei den Dummy-Variablen nahezu identisch sind. Das metrische Modell gibt dabei den besseren *RMSE* aus, da hierbei weniger Freiheitsgrade verloren gehen als bei einer Schätzung des Parameters für alle Merkmalsausprägungen.

Abbildung 5: Geeignetes ordinalskaliertes Merkmal als metrische Variable



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 179-214.

Somit ist bewiesen, dass beide Vorgehensweisen ein besseres Ergebnis im Sinne des *RMSE* erzielen können. Die *Sum of Squares* sind bei dem metrischen Modell zwar höher, der *Root Mean Square Error* ist jedoch geringer, da hier die Zahl der Freiheitsgrade (*df*) mit einfließt.[80]

Tabelle 5: Fehlerkennzahlen der Modelle aus Abbildung 4 und Abbildung 5

	Abbildung 4		***Abbildung 5***	
	Dummy	***metric***	***Dummy***	***metric***
SS	76,0869	419,9001	87,7769	88,3725
RMSE	0,8949	2,0700	0,9612	0,9496
df	95	98	95	98

Quelle: Eigene Darstellung.

Es zeigt sich, dass für die Ermittlung des Modells mit dem niedrigeren *RMSE* die Zahl an Freiheitsgraden mitunter eine entscheidende Rolle spielen kann. Es lässt sich somit sagen, dass umso kleiner die Stichprobengröße n und umso größer die Anzahl an

[80] Hierbei wurde auf den *RMSE* für Trainingsdaten zurückgegriffen, mit $df = n - k - 1$ Freiheitsgraden.

Merkmalsausprägungen k ist, das metrische Modell ein besseres Ergebnis erzielt, wenn die bedingten Mittelwerte näherungsweise eine Gerade abbilden. Wird der *RMSE* für die Testdaten ermittelt, gilt dies nicht. Hier kann das metrische Regressionsmodell im besten Fall denselben *RMSE* wie das Modell mit Dummy-Variablen erzielen da bei der Berechnung des *RMSE* in beiden Fällen lediglich durch die Stichprobengröße n geteilt wird und diese Größe bei beiden Modellen gleich ist.

2.6 Vor- und Nachteile unterschiedlicher Umgangsformen

Beide hier behandelten Umgangsformen von ordinalskalierten Merkmalen können zu einem besseren Modell im Sinne des *RMSE* führen. Sie haben beide ihre Vor- und Nachteile. Als wesentlicher Vorteil einer Umwandlung zu Dummy-Variablen, sind die genaueren Schätzungen der endogenen Variablen zu nennen. Sie entstehen dadurch, dass keine Regressionsgerade, sondern der jeweilige bedingte Mittelwert der endogenen Variablen geschätzt wird. Dies ist der einzige Vorteil, den diese Umgangsform mit sich bringt. Jedoch ist dieser Vorteil unter Umständen marginal, wenn die bedingten Mittelwerte der jeweiligen Merkmalausprägung annähernd eine Gerade abbilden. Als Nachteil der Umwandlung zu Dummy-Variablen ist zu nennen, dass durch die Schätzung eines Koeffizienten pro Merkmalsausprägung dementsprechend viele Freiheitsgrade verloren gehen, da sich die Anzahl an Freiheitsgraden aus der Differenz zwischen der Anzahl an Beobachtungen und der Anzahl der geschätzten Parameter ergibt.[81] Hingegen gehen lediglich zwei Freiheitsgrade verloren, wenn man das ordinalskalierte Merkmal als metrisch skaliert annimmt. Lässt man das ordinalskalierte Merkmal als metrische Variable mit in die Regression einfließen, spart man sich dadurch $k-2$ Freiheitsgrade, wobei k die Anzahl an Merkmalsausprägungen angibt. Dies berechnet sich:

$$df_{met} = n - 2 \tag{2.3}$$

$$df_{ord} = n - k \tag{2.4}$$

$$df_{met} - df_{ord} = (n-2) - (n-k) = n - 2 - n + k = k - 2 \tag{2.5}$$

[81] Vgl. Hedderich, Sachs (2020), S. 845.

Darüber hinaus kann durch die Darstellung einzelner Koeffizienten, wie in der vierten Spalte von *Tabelle 4*, keine Information über die Erhöhung der endogenen Variablen, von einer Merkmalsausprägung von X zur anderen, extrahiert werden. Somit lassen sich keine Aussagen über die Steigung der Regressionsgeraden treffen, da für jede Merkmalsausprägung ein eigener Koeffizient ermittelt wird.

Die Information der Beobachtungswerte, die in der Reihenfolge der einzelnen Merkmalsausprägungen steckt, geht bei einer Umwandlung in Dummy-Variablen vollständig verloren. Das Merkmal wird dadurch lediglich wie ein nominalskaliertes Merkmal behandelt. Zudem sind bei Dummy-Variablen keine Inter- und Extrapolationen möglich. Bei der Aufstellung eines linearen Regressionsmodells ist es oftmals Ziel ein simples Modell zu erstellen, das nur die einflussreichen und signifikanten Variablen beinhaltet.[82] Bei einer Umwandlung zur Dummy-Variablen entstehen eine Vielzahl neuer Variablen, vor allem wenn mehrere ordinalskalierte oder nominalskalierte Merkmale vorliegen, die eine Umwandlung zu Dummy-Variablen benötigen.[83] Diese Vielzahl an Variablen ist als Nachteil von Dummy-Variablen zu nennen, da sie zu einem komplexen Regressionsmodell führen.[84] Man spricht auch von dem *Prinzip der Sparsamkeit*, dass besagt, dass das Modell so einfach wie möglich, jedoch so komplex wie nötig sein soll.[85] Zudem weist jede dieser Variablen ein unterschiedliches Signifikanzniveau auf. Somit kann nicht eindeutig gesagt werden, ob ein Merkmal einen signifikanten Einfluss auf die endogene Variable hat, oder nur auf einzelne Merkmalsausprägungen derselben. Zusammenfassend kann man sagen, dass es von Vorteil sein kann, ein ordinalskaliertes Merkmal als ein metrisch skaliertes Merkmal anzunehmen. Hierbei darf der Schätzfehler jedoch nicht zu groß sein. Der Schätzfehler des Modells soll in den folgenden Kapiteln detaillierter untersucht werden.

[82] Vgl. Backhaus, et al. (2021), S. 69.
[83] Vgl. ebd., S. 12.
[84] Vgl. Hedderich, Sachs (2020), S. 845.
[85] Vgl. Backhaus, et al. (2021), S. 69.

3. Ableitung von Handlungsempfehlungen

3.1 Eigenschaften der Regressionskoeffizienten

Ein ordinalskaliertes Merkmal kann als metrische Variable in linearen Regressionsmodellen verwendet werden, wenn es bestimmte Eigenschaften aufweist. Verwendet man das Merkmal in einem linearen Regressionsmodell, dann sollte dieser lineare Zusammenhang zwischen exogener und endogener Variablen möglichst exakt gegeben sein. Dies ist der Fall, wenn die bedingten Mittelwerte der endogenen Variablen für die jeweiligen Merkmalsausprägungen näherungsweise eine Gerade abbilden. Die Abweichungen zwischen der Regressionsgeraden der linearen Einfachregression und den Beobachtungen, kann durch zwei zu unterscheidende Sachverhalte entstehen:

- Es besteht kein linearer Zusammenhang zwischen der exogenen und endogenen Variablen.
- Das ordinalskalierte Merkmal ist willkürlich kodiert, weswegen die Abbildung einer Geraden nicht möglich ist.

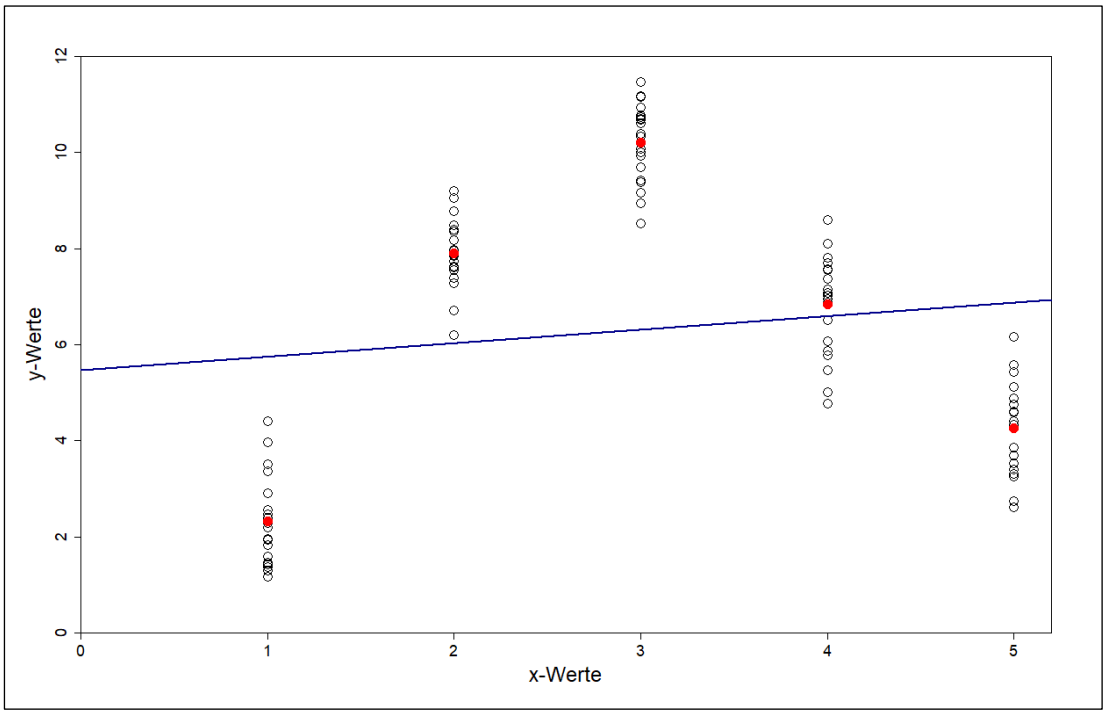
Die Kodierung eines ordinalskalierten Merkmals erfolgt in der Regel rein subjektiv und ohne Systematik. In den meisten Fällen werden die natürlichen Zahlen beginnend mit 1 entlang der Rangwertreihe des ordinalskalierten Merkmals verwendet. Solange die Reihenfolge der Kodierung beibehalten wird, kann diese geändert werden, sodass die bedingten Mittelwerte der endogenen Variablen für jede Merkmalsausprägung exakt auf einer Geraden liegen. Diese Thematik wird in *Kapitel 4.1* behandelt. Für die Untersuchung, ob sich ein ordinalskaliertes Merkmal für die Verwendung in einem linearen Regressionsmodell eignet, kann in erster Linie festgestellt werden, ob die Rangwertreihe der Ordinalskala der exogenen Variablen auch bei der endogenen Variablen vorliegt. Dies leitet sich folgendermaßen her:

$$\bar{y}_1 < \bar{y}_2 < \bar{y}_3 < \cdots < \bar{y}_K \vee \bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \cdots > \bar{y}_K \quad \forall\, k \in K \tag{3.1}$$

$$\bar{y}_k = \bar{y}_{[k]} \vee \bar{y}_k = \bar{y}_{[K-k+1]} \quad \forall\, k \in K \tag{3.2}$$

Ist diese Bedingung erfüllt, dann ist ein linearer Zusammenhang zwischen exogener und endogener Variablen möglich. Ein ordinalskaliertes Merkmal, bei dem dieser kausale Zusammenhang nicht besteht und das sich für die Verwendung in einem linearen Regressionsmodell somit nicht eignet, ist in *Abbildung 6* abgebildet.

Abbildung 6: Nichtkausales ordinalskaliertes Merkmal



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 233-268.

Das dargestellte Merkmal eignet sich nicht für eine lineare Regressionsfunktion, da ein höheres X in diesem Fall nicht zu einem höheren, bzw. niedrigeren Y führt. Hier wäre es angeraten die Variable nicht in das lineare Regressionsmodell einfließen zu lassen oder sie zu mehreren Dummy-Variablen umzuwandeln. In diesem Fall würde das Modell die bedingten Mittelwerte der endogenen Variablen (in *Abbildung 6* als rote Punkte dargestellt) schätzen. Alternativ könnte man dieses Merkmal mithilfe eines nichtlinearen Regressionsmodells abbilden.[86] Es ist darauf hinzuweisen, dass die hier beschriebene Kausalbeziehung zwischen X und Y nicht für jede einzelne Beobachtung einzuhalten ist, sondern lediglich für die bedingten Mittelwerte der einzelnen Merkmalsausprägungen von X. Ob die Rangwertreihe der exogenen Variablen auch der der endogenen Variablen entspricht, lässt sich im beigefügten *R*-Skript automatisch ausgeben. Hierbei wird ein Antwortsatz erstellt, der beschreibt, bei wie vielen der Rangwerte von X diese Bedingung erfüllt ist.[87]

[86] Vgl. Backhaus, Erichson, Weiber (2015), S. 24.
[87] Beispiel: "`Ordinal sequence observed in 5/5 cases.`", siehe: *R*-Skript, Zeile: 602.

3.2 Beurteilung der Güte der Regressionsmodelle

Um die Güte der unterschiedlichen Regressionsmodelle zu beurteilen, werden diese jeweils einmal mit der Verwendung von Dummy-Variablen für das ordinalskalierte Merkmal und einmal als metrisches Merkmal ausprobiert. Hierfür wird lediglich eine lineare Einfachregression verwendet, auch wenn das ordinalskalierte Merkmal Teil einer multiplen Regression sein soll. Für die Beurteilung des Merkmals kommen folgende Fehlerkennzahlen zum Einsatz:

- *R2*: R-Quadrat ist das sogenannte Bestimmtheitsmaß.[88] Es gibt den Anteil der Streuung von Y an, der durch das Modell erklärt wird.[89] Das Bestimmtheitsmaß des Regressionsmodells berechnet sich:

$$R^2 = r_{YX}^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST} = \frac{erklärte\ Streuung}{Gesamtstreuung} \qquad (3.3)^{90}$$

- *Adjusted R2*: Das korrigierte Bestimmtheitsmaß berücksichtigt zusätzlich zu dem Bestimmtheitsmaß den Stichprobenumfang, sowie die Anzahl an Parametern.[91] Diese Kennzahl zielt darauf ab, eine möglichst exakte Prognose durch das Modell zu erhalten, da durch das Hinzunehmen neuer Variablen zwar das Bestimmtheitsmaß verbessert wird, jedoch nicht die Prognosegüte („*overfitting*“).[92]

$$R_{adj}^2 = 1 - \frac{N-1}{N-J-1}(1-R^2) = 1 - \frac{\frac{SSR}{N-J-1}}{\frac{SST}{N-1}} = 1 - \frac{MSR}{MST} \qquad (3.4)^{93}$$

[88] Vgl. Backhaus, et al. (2021), S. 86.
[89] Vgl. ebd.
[90] Quelle: Backhaus, Erichson, Weiber (2015), S. 40.
[91] Vgl. Backhaus, et al. (2021), S. 93 f.
[92] Vgl. Backhaus, Erichson, Weiber (2015), S. 320.
[93] Quelle: Hedderich, Sachs (2020), S. 836.

- *SS*: Die *Sum of Squares* ist die Summe der quadrierten Residuen.[94] Die Aussagekraft dieser Kennzahl ist gering, da die Werte quadriert sind und die Größe von der Anzahl an Beobachtungen und der Einheit der endogenen Variablen abhängt.[95]

$$SS = \sum_{i=1}^{n} \hat{u}_i^2 \qquad (3.5)^{96}$$

- *MSE*: Der *Mean Square Error* berechnet sich, indem man die *Sum of Squares* durch die entsprechenden Freiheitsgrade teilt.

$$MSE = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2 \qquad (3.6)^{97}$$

- *RMSE*: Der *Root Mean Square Error* oder auch Standardfehler der Schätzung (englisch: *standard error of estimate*) bemisst, wie stark die Beobachtungen von ihrem Schätzwert abweichen.[98] Die Präzision des Modells ist hoch, wenn der *RMSE* gering ausgeprägt ist.[99] Der *RMSE* hat dieselbe Einheit wie die endogene Variable.[100] Er berechnet sich als Wurzel des *MSE*:

$$RMSE = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2} \qquad (3.7)^{101}$$

- *Rel. RMSE (in %)*: Teilt man den *RMSE* durch das arithmetische Mittel der endogenen Variablen, erhält man den *RMSE* als relative Größe.[102] Dieser gibt an

[94] Vgl. Backhaus, et al. (2021), S. 85.
[95] Vgl. ebd.
[96] Quelle: Hedderich, Sachs (2020), S. 137.
[97] Quelle: ebd., S. 414, 825.
[98] Vgl. Backhaus, Erichson, Weiber (2015), S. 41, vgl. Backhaus, et al. (2021), S. 85.
[99] Vgl. Backhaus, et al. (2021), S. 85.
[100] Vgl. ebd., S. 86.
[101] Vgl. Backhaus, Erichson, Weiber (2015), S. 41.
[102] Vgl. Backhaus, et al. (2021), S. 86.

um wieviel Prozent der Standardfehler der Schätzung von dem Mittelwert der endogenen Variablen abweicht.[103]

$$Rel.RMSE = \frac{RMSE}{\bar{y}} \cdot 100\% \qquad (3.8)^{104}$$

- *MAE*: Der *Mean Absolute Error* gibt die durchschnittliche absolute Abweichung zwischen Schätzung und Beobachtung an.[105] Diese Kennzahl hat den Vorteil, dass sie robuster ist als der *RMSE*, da sie weniger empfindlich gegenüber Ausreißern ist, die beim *RMSE* quadriert werden.[106]

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{u}_i| \qquad (3.9)^{107}$$

- *MAPE*: Der *Mean Absolute Percentage Error* gibt die durchschnittliche prozentuale Abweichung der Beobachtungen und der Schätzwerte an.[108]

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{u}_i}{y_i}\right| \qquad (3.10)^{109}$$

- *df*: Zudem werden die Freiheitsgrade des Modells berücksichtigt.[110] Diese berechnen sich aus der Differenz aus Beobachtungswerten und der Anzahl an Parametern, die geschätzt werden.[111] Bei der Modellierung sollten möglichst wenige Freiheitsgrade verbraucht werden, damit das Modell eine höhere Genauigkeit aufweist.[112]

Im *R*-Skript wird für das ordinalskalierte Merkmal ein Dashboard ausgegeben, das beide Modelle gegenüberstellt. Somit kann eine fundierte Entscheidung für die

[103] Vgl. Backhaus, et al. (2021), S. 86.
[104] Vgl. Backhaus, Erichson, Weiber (2015), S. 41.
[105] Vgl. Backhaus, et al. (2021), S. 79.
[106] Vgl. ebd., S. 80.
[107] Quelle: ebd., S. 154.
[108] Vgl. Kühne, Wenger (2011), S. 321.
[109] Quelle: ebd.
[110] Siehe: *Formel 2.3* und *Formel 2.4* auf S. 17.
[111] Vgl. Backhaus, et al. (2021), S. 24.
[112] Vgl. ebd.

Handhabung der ordinalskalierten Variable getroffen werden. Dieses Dashboard ist in *Tabelle 6* dargestellt. Alle hierbei berechneten Kennzahlen werden einmal für Trainingsdaten (bei denen man davon ausgeht, dass die benötigten Parameter geschätzt werden) sowie für Test- oder Validierungsdaten (bei denen die Stichprobengröße die Anzahl an Freiheitsgraden darstellt) ermittelt.[113] Dadurch kann das in *Tabelle 6* dargestellte Dashboard für beide Datensätze verwendet werden.

Tabelle 6: Dashboard der Fehlerkennzahlen der beiden Modelle

	Dummy	***metric***	***better model***
R2	0,9708	0,8771	*Dummy*
Adjusted R2	0,9696	0,8759	*Dummy*
SS	76,0869	320,2612	*Dummy*
MSE Train	0,8009	3,2680	*Dummy*
MSE Test	0,7609	3,2026	*Dummy*
RMSE Train	0,8949	1,8078	*Dummy*
RMSE Test	0,8723	1,7896	*Dummy*
rel. RMSE Train (in %)	7,7761	15,7075	*Dummy*
rel. RMSE Test (in %)	7,5792	15,5496	*Dummy*
MAE Train	0,7374	1,5323	*Dummy*
MAE Test	0,7006	1,5017	*Dummy*
MAPE Train (in %)	7,4707	15,3441	*Dummy*
MAPE Test (in %)	7,0971	15,0372	*Dummy*
df Train	95	98	*Dummy*
df Test	100	100	*Dummy*
Ordinal sequence	observed in 5/5 cases.		*metric*

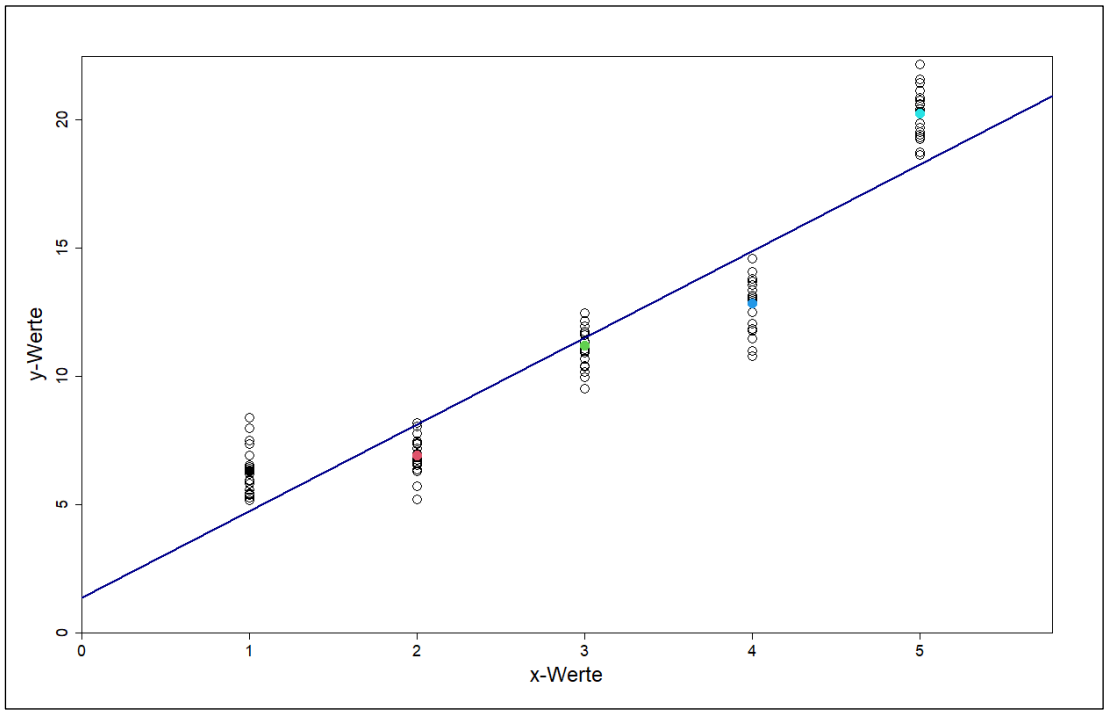
Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 398-602.

Durch das Dashboard kann ermittelt werden, welches Verfahren die besseren Werte der einzelnen Fehlerkennzahlen ausgibt. Für das Bestimmtheitsmaß *R2*, sowie das das korrigierte Bestimmtheitsmaß *Adjusted R2* sind Werte nahe 1 am besten, während die restlichen Fehlerkennzahlen möglichst niedrige Werte aufweisen sollen. Zudem ist es von Vorteil, wenn bei der Einfachregression mit dem ordinalskalierten Merkmal möglichst wenig Freiheitsgrade verloren gehen, um eine höhere

[113] Vgl. Hirschle (2021), S. 51 ff.

Genauigkeit zu erhalten. Welche Vorgehensweise das bessere Modell ergibt, wird für jedes Gütekriterium gesondert in der letzten Spalte angegeben. In der letzten Zeile von *Tabelle 6* wird ermittelt, ob die Rangwertreihe der exogenen Variablen auch für die bedingten Mittelwerte der endogenen Variablen gilt.[114] Ist dies nicht der Fall, wird das Modell mit Dummy-Variablen empfohlen. Ergänzend zu dem Dashboard wird ein Data-Plot (*Abbildung 7*) mit den Schätzwerten der Dummy-Variablen, sowie der Regressionsgerade des metrischen Modells ausgegeben, um eine optische Kontrolle durchführen zu können.

Abbildung 7: Data-Plot eines ordinalskalierten Merkmals



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 271-323.

Die schwarzen Kreise sind die Beobachtungswerte. Die farbigen Punkte stellen die Schätzwerte der Dummy-Variablen dar und die blaue Regressionsgerade die Schätzwerte des metrischen Modells. Durch die Grafik wird deutlich, dass die Beobachtungswerte näherungsweise eine Gerade abbilden. Jedoch würde die Dummy-Variable bessere Schätzwerte liefern als die metrische Regressionsgerade.

[114] Liegt die korrekte Reihenfolge vor, lautet der Satz „`Ordinal sequence observed in k/k cases.`"

3.3 Allgemeingültige Lösung

Bei der Entscheidung, ob ein ordinalskaliertes Merkmal als metrisches oder nominalskaliertes Merkmal in ein lineares Regressionsmodell eingearbeitet werden soll, ist prinzipiell zu empfehlen die Umgangsform zu wählen, die das bessere Modell im Sinne der präferierten Fehlerkennzahl ausgibt. Eine Reihe von Fehlerkennzahlen sind in dem in *Tabelle 6* dargestellten Dashboard aufgelistet und können für jedes ordinalskalierte Merkmal gesondert ausgegeben werden. Zudem kann das in *Tabelle 6* dargestellte Dashboard sowohl für Trainings-, als auch Test- und Validierungsdaten verwendet werden, was die Fehlerkennzahlen im Hinblick auf deren Freiheitsgrade unterscheidet. Aus bereits behandelten Gründen wird das genauere Modell im Regelfall das Modell mit Dummy-Variablen sein. Ist die Differenz der Fehlerkennzahlen für beide Fälle marginal, dann ist das metrische Regressionsmodell zu empfehlen, da es die in *Kapitel 2.6* behandelten Vorteile mit sich bringt. Die Entscheidung bleibt dennoch beim Analysten und ist eine Einzelfallbetrachtung. Von besonderem Interesse ist, ob für die einzelne Fehlerkennzahl, wie beispielsweise den *RMSE* oder das korrigierte Bestimmtheitsmaß, die höheren Freiheitsgrade des metrischen Modells oder die genauere Schätzung der Dummy-Variablen mehr ins Gewicht fallen. Um die Vorteile beider Umgangsformen zu vereinen, wird im folgenden Kapitel eine zusätzliche Vorgehensweise erarbeitet.

4. Umkodierung der ordinalskalierten Variable

4.1 Umkodierung bei der linearen Einfachregression

4.1.1 Grundidee der Umkodierung

Die Kodierung des ordinalskalierten exogenen Merkmals X erfolgt im Regelfall anhand der natürlichen Zahlen, beginnend mit 1 entlang der Ordinalskala. Diese mehr oder weniger willkürliche Kodierung wird im Folgenden kritisch betrachtet. Von der Kodierung des ordinalskalierten Merkmals ist die Güte des Modells sowie die der Schätzwerte abhängig. Für unterschiedliche Kodierungen kommt man somit zu unterschiedlichen Ergebnissen, die wiederrum verschiedene Gütekriterien für das lineare Regressionsmodell ergeben. Dies widerspricht dem Wissenschaftsgrundsatz, dass bei unabhängiger Durchführung eines Experimentes dieselben Ergebnisse erzielt

werden müssen („Auswertungsobjektivität).[115] Dasselbe gilt auch für die Analyse eines Datensatzes. Die Schätzwerte eines Modells sollten nicht davon abhängig sein, wie der Analyst die Merkmalsausprägungen des ordinalskalierten Merkmals kodiert. Dass unterschiedliche Kodierungen zu unterschiedlichen Modellen und unterschiedlichen Schätzergebnissen führen, wird in folgendem Anwendungsbeispiel aufgezeigt.

Tabelle 7: Rating-Skala mit unterschiedlichen Kodierungen

"Ich persönlich verwende Marke X…"	*Kodierung 1*	*Kodierung 2*
„…täglich mehrmals“	1	11
„…täglich einmal“	2	12
„…wöchentlich mehrmals“	3	21
„…wöchentlich einmal“	4	22
„…monatlich mehrmals“	5	31
„…monatlich einmal“	6	32
„…seltener“	7	40
„…so gut wie nie / nie“	8	50

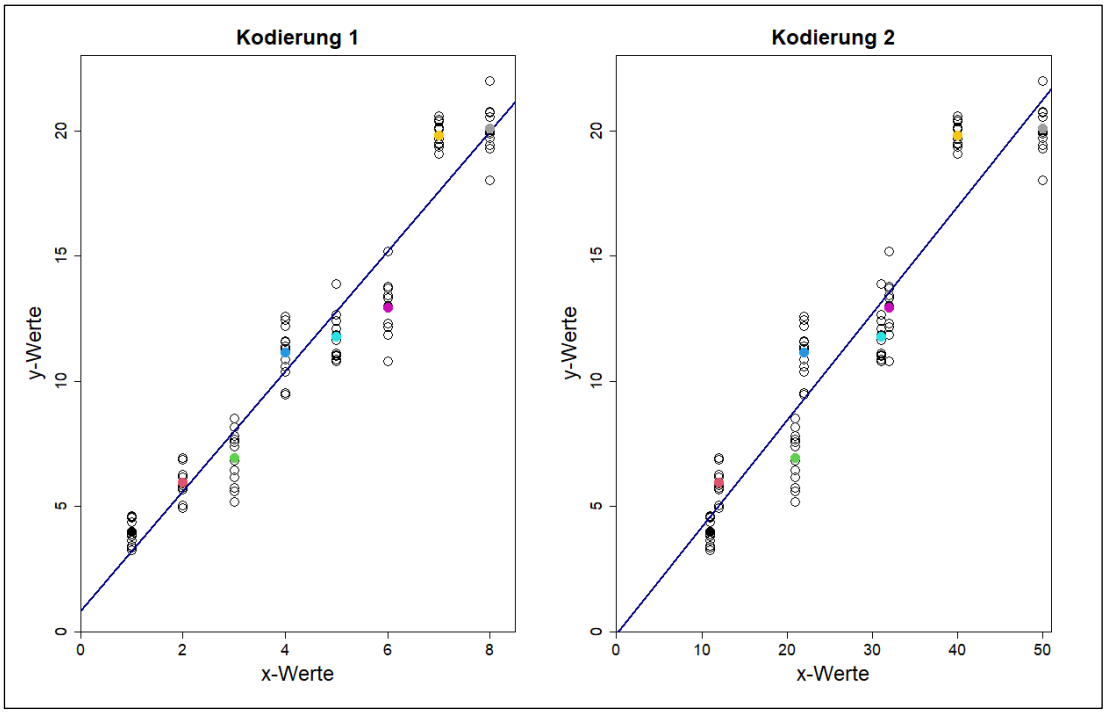
Quelle: Eigene Darstellung, in Anlehnung an: Berekhoven, Eckert, Ellenrieder (2009), S. 68.

In *Tabelle 7* sind zwei Kodierungen für ein ordinalskaliertes Merkmal aufgelistet. Beide Kodierungen wurden aufsteigend mit absteigender Häufigkeit der Nutzung kodiert. Würde man stattdessen absteigend kodieren würde man dasselbe Modell mit denselben Schätzwerten erhalten. Es würde sich lediglich das Vorzeichen des Steigungsparameters ändern.[116] Beide hier dargestellten Kodierungen orientieren sich an der Rangwertreihe, die durch die unterschiedlichen Antwortkategorien vorgegeben werden. Bei der ersten Kodierung werden die natürlichen Zahlen von 1 bis 8 verwendet. Bei der zweiten Kodierung wurde mehr Bezug auf die Antwortkategorien genommen. Hierbei steht die erste Stelle der Kodierung für „täglich“ (1), „wöchentlich“ (2), „monatlich“ (3) oder für die beiden übrigen Antwortkategorien $(4$ und $5)$. Die zweite Stelle der Kodierung steht für „mehrmals“ (1), „einmal“ (2) oder die übrigen Antwortkategorien (0). Beide Kodierungen sind korrekt und könnten in dieser Form in der Praxis Anwendung finden. In folgender *Abbildung 8* ist zu sehen, dass die beiden Kodierungen für dieselben Beobachtungswerte von Y zu unterschiedlichen Modellen

[115] Vgl. Berekhoven, Eckert, Ellenrieder (2009), S. 82.
[116] Siehe: *R*-Skript, Zeile: 384-395.

sowie unterschiedlichen Schätzwerten und damit zu abweichenden Gütekriterien führen.

Abbildung 8: Unterschiedliche Modelle in Abhängigkeit der Kodierung



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 326-382.

Die unterschiedlichen Schätzergebnisse sind am besten an der ersten Merkmalsausprägung von X zu erkennen, die in *Abbildung 8* als ausgefüllter schwarzer Kreis dargestellt wird. Während bei der ersten Kodierung der bedingte Mittelwert der ersten Merkmalsausprägung von X unterschätzt wird, wird dieser bei der zweiten Kodierung überschätzt. Die Schätzwerte der Modelle weichen somit voneinander ab, je nach subjektiv gewählter Kodierung. Dies zeigt sich auch in den Fehlerkennzahlen der Modelle. Während das lineare Modell der ersten Kodierung einen *Root Mean Square Error* von $1{,}565$ aufweist, liegt dieser bei dem Modell mit der zweiten Kodierung bei $1{,}814$. Die unterschiedlichen Modelle resultieren aus den unterschiedlichen Abständen zwischen den Merkmalsausprägungen der exogenen Variablen. Würde man beispielsweise statt der ersten Kodierung mit den natürlichen Zahlen von 1 bis 8, jeweils die doppelten Werte von 2 bis 16 wählen, würde man dasselbe Modell mit denselben Schätzwerten und denselben Fehlerkennzahlen erhalten. Lediglich die Steigung des

Modells würde sich dadurch halbieren.[117] Will man das hier beobachtete ordinalskalierte exogene Merkmal als metrische Variable mit in das Regressionsmodell aufnehmen, ist die Kodierung, was die Güte des resultierenden Modells betrifft, von entscheidender Bedeutung. Dies gilt nicht, wenn das Merkmal mithilfe von Dummy-Variablen in das Modell einfließt. Hierbei werden in jedem Fall die bedingten Mittelwerte der einzelnen Merkmalsausprägungen geschätzt. Die ursprüngliche Kodierung, wie sie in *Tabelle 7* dargestellt wird, spielt hierfür keine Rolle.

Statt der herkömmlichen Kodierung mit natürlichen Zahlen könnten die jeweiligen Merkmalsausprägungen von X so kodiert werden, dass die bedingten Mittelwerte der jeweiligen Merkmalsausprägungen exakt auf einer Geraden liegen. Diese Gerade entspricht nach der Umkodierung der Regressionsgeraden der linearen Einfachregression. Die Umkodierung des ordinalen Merkmals beruht auf folgender Annahme: Das ordinalskalierte Merkmal X wird in einem linearen Regressionsmodell als exogene Variable verwendet. Somit besteht die Vermutung, dass ein linearer Zusammenhang zwischen X und Y besteht.[118] Da die Kodierung rein subjektiv und daher willkürlich erfolgt, kann diese auch nach Belieben geändert werden, solange die Rangwertreihe nicht verletzt wird. Die Reihenfolge der Merkmalsausprägungen muss somit auch nach der Umkodierung zwingend beibehalten werden. Die Abstände zwischen den Merkmalsausprägungen von X lassen sich nicht interpretieren oder analysieren und können daher nach Belieben geändert werden. Wird ein ordinalskaliertes Merkmal in der Form umkodiert, dass die bedingten Mittelwerte und die Schätzwerte des Regressionsmodells übereinstimmen, dann kombiniert man die Vorteile von Dummy-Variablen und metrischen Regressionsmodellen. Zum einen können alle Sprünge zwischen den Beobachtungen aller Merkmalsausprägungen abgebildet werden, wie es nur die Dummy-Variablen können, da hierbei jeweils das arithmetische Mittel der endogenen Variablen geschätzt wird. Zum anderen entsteht durch diese Vorgehensweise nur eine Variable pro exogenem Merkmal. Dies hat den Vorteil, dass das Signifikanzniveau, sowie die Größe des Parameters für das ordinalskalierte Merkmal ausgegeben werde kann.

[117] Siehe: *R*-Skript, Zeile: 384-391.
[118] Vgl. Hedderich, Sachs (2020), S. 137.

4.1.2 Vorgehen bei der Umkodierung eines ordinalskalierten Merkmales

Eine bespielhafte Umkodierung eines ordinalskalierten Merkmals ist in *Tabelle 8* dargestellt. Hierbei wird in der ersten Spalte die ursprüngliche Kodierung des ordinalskalierten Merkmals mit den natürlichen Zahlen von 1 bis 5 dargestellt. In diesem Beispiel kommt jede Merkmalsausprägung von X gleich oft vor, wie die relativen Wahrscheinlichkeiten $p(x_k)$ zeigen. Die bedingten Mittelwerte pro Merkmalsausprägung von X werden als $\bar{y}_k$ bezeichnet.

Tabelle 8: Beispielhafte Umkodierung eines ordinalskalierten Merkmals

x_k	$p(x_k)$	$\bar{y}_k$	$\hat{y}_{met,oc}$	$\hat{y}_{Dummy}$	$x_{k,nc}$	$\hat{y}_{met,nc}$
1	0,2	6	4,6	$6 = \beta_1 = Int.$	1,412	6
2	0,2	7	8,0	$7 = \beta_1 + \beta_2$	1,706	7
3	0,2	11	11,4	$11 = \beta_1 + \beta_3$	2,882	11
4	0,2	13	14,8	$13 = \beta_1 + \beta_4$	3,471	13
5	0,2	20	18,2	$20 = \beta_1 + \beta_5$	5,529	20

Quelle: Eigene Darstellung.

Es ist zu erkennen, dass die Schätzwerte pro Merkmalsausprägung $\hat{y}_{met,oc}$ stark von den Beobachtungswerten abweichen.[119] Hingegen geben die Schätzwerte der Dummy-Variablen jeweils den exakten Mittelwert der jeweiligen Merkmalsausprägung wieder. Die umkodierte Variable wird in der Tabelle als $x_{k,nc}$ bezeichnet.[120] Wird für das umkodierte Merkmal X die lineare Regression als metrisches Modell durchgeführt, werden ebenfalls exakt die bedingten Mittelwerte geschätzt, wie es bei der Verwendung von Dummy-Variablen der Fall ist. Die neue Kodierung $x_{k,nc}$ berechnet sich folgendermaßen: In einem ersten Schritt wird die ursprüngliche Kodierung als metrische Größe angenommen, um hieraus die Gerade der linearen Einfachregression zu bilden. Die Regressionsgerade der linearen Einfachregression hat im Allgemeinen die Form:

$$\hat{y}_i = \beta_1 + \beta_2 \cdot x_i \qquad (4.1)^{121}$$

[119] Diese Bezeichnung wird auch im *R*-Skript verwendet und steht für „*old coding*“.
[120] Diese Bezeichnung wird auch im *R*-Skript verwendet und steht für „*new coding*“.
[121] Quelle: Backhaus, et al. (2021), S. 72.

Es werden die Bezeichnungen β_1 für den Achsenabschnitt und β_2 für die Steigung des metrischen Modells verwendet, damit die Indizes der Parameter mit dem *R*-Skript übereinstimmen.[122] In dem in *Tabelle 8* aufgezeigten Zahlenbeispiel ergibt sich die Regressionsgerade:

$$\hat{y}_i = 1{,}2 + 3{,}4 \cdot x_i \tag{4.2}$$

Diese Gerade soll für die folgende Vorgehensweise beibehalten werden.[123] Dies hat den Vorteil, dass die Kodierung von X nur minimal angeglichen werden muss. Die Geradengleichung aus der linearen Einfachregression, bei der die Kodierung als metrische Größe angenommen wird, wird anschließend nach x_k umgeformt, um die neue Kodierung zu ermitteln. Die Merkmalsausprägungen der exogenen Variablen sollen dabei so kodiert werden, dass ihr bedingter Mittelwert der endogenen Variablen auf der ursprünglichen Geraden liegt und dieser somit dem Schätzwert des Modells entspricht. Daraus ergibt sich die Voraussetzung:

$$\hat{y}_k = \bar{y}_k \tag{4.3}$$

Die neue Kodierung berechnet sich somit:

$$x_{k,nc} = \frac{\bar{y}_k - \beta_1}{\beta_2} \tag{4.4}$$[124]

Anhand dieser Formel ist zu erkennen, dass durch den Steigungsparameter im Nenner, die Steigung der Regressionsgerade, die angenommen wird, weder 0 noch unendlich sein darf. Anhand *Tabelle 8* ist zudem zu erkennen, dass die Behandlung des ordinalskalierten Merkmals als metrische Variable nach der Umkodierung exakt die gleichen Schätzwerte wie die Umwandlung zur Dummy-Variablen liefert. Wie sich die Umkodierung der einzelnen Merkmalsausprägungen grafisch abbildet, ist in

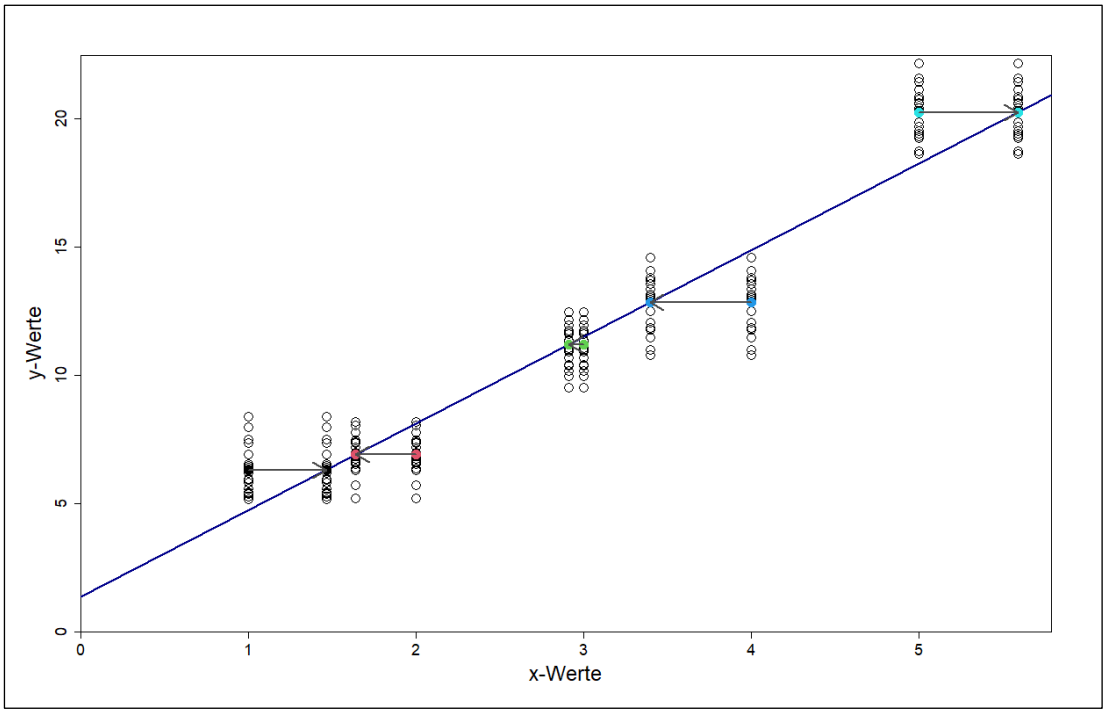
[122] Im *R*-Skript erhält man die Koeffizienten des Achsenabschnitts durch den Befehl: `coef(modell)[1]`, siehe: *R*-Skript, Zeile: 421.

[123] In *Kapitel 4.1.4* wird darauf eingegangen, dass auch jede andere Gerade zu denselben Schätzwerten führen würde, wenn die Kodierung in der Form angeglichen wird, dass die bedingten Mittelwerte eine Gerade abbilden.

[124] Es wird der Index k verwendet, da sowohl vor als auch nach der Umkodierung, k unterschiedliche Merkmalsausprägungen von X vorliegen. Die bei der Umkodierung vollzogenen Berechnung gleicht stark der Formel der inversen Prädiktion einer linearen Regression, bei der von einem y-Wert auf einen x-Wert rückgeschlossen wird. Siehe hierzu: Hedderich, Sachs (2020), S. 422.

Abbildung 9 dargestellt.[125] Diese Grafik wird auch bei der Umkodierung mit dem selbstprogrammierten *R*-Paket ausgegeben.[126]

Abbildung 9: Umkodierung des ordinalskalierten Merkmals



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 398-470.

In der hier dargestellten *Abbildung 9* ist jede Beobachtung zweimal gegeben. Einmal vor und einmal nach der Umkodierung. Anhand der farbigen Punkte ist zu erkennen, wie sich durch die Umkodierung die bedingten Mittelwerte auf der x-Achse verschieben und sich die Abstände der Beobachtungen zur Regressionsgerade damit deutlich verringern. Die Regressionsgerade ist für beide Kodierungen exakt dieselbe, da bei der Umkodierung die Koeffizienten des ursprünglichen Modells mit eingeflossen sind.[127] Mit Bezug auf die Umkodierung ist darauf hinzuweisen, dass ein ordinalskaliertes Merkmal aus mindestens drei Merkmalsausprägungen besteht. Liegen lediglich zwei Merkmalsausprägungen für eine Variable vor, spricht man von einer dichotomen Variablen. Diese zwei Merkmalsausprägungen lassen sich nicht in eine bestimmte Reihenfolge bringen. Hierbei ist die Umkodierung nicht möglich, da beide

[125] Durch die streuenden Beobachtungswerte ergeben sich hierbei minimal abweichende Parameter, verglichen zu den Werten aus *Tabelle 8*.
[126] Siehe: *Anhang*, S. VII.
[127] Siehe: *Formel 4.4* auf S. 31.

bedingten Mittelwerte der jeweiligen Merkmalsausprägung auf der Regressionsgeraden liegen. Sowohl das Modell mit Dummy-Variablen als auch das metrische Modell, vor und nach der Umkodierung, liefern hierbei dieselben Schätzwerte.

4.1.3 Fehlerkennzahlen der einzelnen Modelle

Durch die Umkodierung werden die Residuen, also die vertikalen Abstände zwischen der Regressionsgerade und den Beobachtungswerten, minimiert. Somit erfüllt die umkodierte ordinalskalierte Variable als metrisches Merkmal die Bedingung der *KQ-Methode* (*Methode der kleinsten Quadrate*).[128] Eine Folge der Schätzung mit der *Kleinsten-Quadrate-Methode* ist, dass die Regressionsgerade durch den Schwerpunkt, bzw. den Mittelpunkt der Daten verläuft.[129] Mithilfe der Umkodierung des ordinalskalierten Merkmals, geht die Regressionsgerade exakt durch alle Schwerpunkte für jede Merkmalsausprägung von X. Zudem ist das Bestimmtheitsmaß für das Modell mit Dummy-Variablen und dem metrischen Modell mit dem umkodierten ordinalskalierten Merkmal dasselbe, wie in *Tabelle 9* dargestellt wird. Auch die *Sum of Squares* sind für das Regressionsmodell mit Dummy-Variablen dieselben wie bei der neuen Kodierung mit metrischer Annahme, da die Modelle jeweils exakt dieselben Schätzwerte ausgeben. Wird der *RMSE* für Testdaten ermittelt, dann liefert das Modell mit der metrischen umkodierten Variablen exakt den gleichen Wert, wie die Umwandlung zu mehreren Dummy-Variablen, da in beiden Fällen das arithmetische Mittel der jeweiligen Merkmalsausprägung exakt abgebildet wird. Wird der *RMSE* für Trainingsdaten ermittelt, erzielt das Modell mit der metrischen umkodierten Variablen minimal schlechtere Ergebnisse als das Modell mit Dummy-Variablen, da für die Umkodierung des ordinalskalierten Merkmals mehr Freiheitsgrade verloren gehen.[130] Für die übrigen Fehlerkennzahlen ist ähnliches zu beobachten.

[128] Vgl. Backhaus, Erichson, Weiber (2015), S. 33.
[129] Vgl. Backhaus, et al. (2021), S. 74.
[130] Für die `summary` der Modelle darf der ausgegebene `Residual Standard Error` nicht berücksichtigt werden, da dieser mit einer falschen Anzahl an Freiheitsgraden berechnet wurde. Stattdessen sollten die Werte des Dashboards verwendet werden.

Tabelle 9: Dashboard der Fehlerkennzahlen mit dem umkodierten Modell

	Dummy	*metric oc*	*better model*	*metric nc*
R2	0,9708	0,8771	*Dummy*	0,9708
Adjusted R2	0,9696	0,8759	*Dummy*	0,9689
SS	76,0869	320,2612	*Dummy*	76,0869
MSE Train	0,8009	3,2680	*Dummy*	0,8181
MSE Test	0,7609	3,2026	*Dummy*	0,7609
RMSE Train	0,8949	1,8078	*Dummy*	0,9045
RMSE Test	0,8723	1,7896	*Dummy*	0,8723
rel, RMSE Train (in %)	7,7761	15,7075	*Dummy*	7,8592
rel, RMSE Test (in %)	7,5792	15,5496	*Dummy*	7,5792
MAE Train	0,7374	1,5323	*Dummy*	0,7533
MAE Test	0,7006	1,5017	*Dummy*	0,7006
MAPE Train (in %)	7,4707	15,3441	*Dummy*	7,6313
MAPE Test (in %)	7,0971	15,0372	*Dummy*	7,0971
df Train	95	98	*Dummy*	93
df Test	100	100	*Dummy*	100
Ordinal sequence	observed in 5/5 cases.		*metric*	

Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 398-602.

Formel 4.4 zeigt auf, dass es zwei Freiheitsgrade braucht, um die Steigung und den Achsenabschnitt des metrischen Modells zu schätzen und einen weiteren Freiheitsgrad für jede der k Kodierungen der exogenen Variablen. Für die drei Modelle ergeben sich die in *Tabelle 10* dargestellte Zahl der Freiheitsgrade für die lineare Einfachregression.

Tabelle 10: Freiheitsgrade der unterschiedlichen Modelle

Modell	**Zahl der Freiheitsgrade**
Dummy	$df_{Dummy} = n - k$
metric oc	$df_{metric,oc} = n - 2$
metric nc	$df_{metric,nc} = n - k - 2$

Quelle: Eigene Darstellung.

Die Zahl der Freiheitsgrade des umkodierten Modells lässt zwei gleich gültige Interpretationen zu. In jedem Fall liegen n Beobachtungen vor, woraus sich die Freiheitsgrade abzüglich aller zu schätzenden Parameter ergeben. Zum einen können die

Parameter für Steigung und Achsenabschnitt für das umkodierte Modell festgelegt werden.[131] Da diese Parameter nicht geschätzt werden, gehen dadurch auch keine Freiheitsgrade verloren. Für die Umkodierung der k Merkmalsausprägungen der ordinalskalierten endogenen Variable gehen k Freiheitsgrade verloren. Anschließend werden die Regressionsparameter für das lineare Modell anhand der endogenen Variablen und der umkodierten erklärenden Variablen erneut geschätzt, wodurch weitere zwei Freiheitsgrade verloren gehen.[132] Die herkömmlichere Vorgehensweise ist die in *Kapitel 4.1.2* beschriebene, bei der die Parameter der Regressionsgerade, die sich aus der ursprünglichen Kodierung ergibt, geschätzt und beibehalten werden.[133] Hierfür gehen zwei Freiheitsgrade verloren. Darüber hinaus gehen k Freiheitsgrade für die Umkodierung der k unterschiedlichen Merkmalsausprägungen verloren. Mit der neuen Kodierung muss die Regressionsgerade nicht erneut geschätzt werden, da diese zwingend der alten Geraden entspricht. Dies liegt daran, dass die Neukodierungen auf den Parametern des ursprünglichen Modells basieren.[134] Somit stehen in jedem Fall $df_{metric,nc} = n - k - 2$ Freiheitsgrade zur Verfügung.

Bei der Umkodierung muss kontrolliert werden, ob die Rangwertreihe, die durch die ursprüngliche Kodierung vorgegeben wird, auch nach der Umkodierung noch besteht. Dies ist nur dann der Fall, wenn die Kodierungen durch die Verschiebung auf der x-Achse nicht überkreuzt verschoben werden. Die Information über die Ordnung der Merkmalsausprägungen muss somit erhalten bleiben.[135] Da für die Umkodierung die bedingten Mittelwerte der endogenen Variablen der jeweiligen Merkmalsausprägung von X verwendet werden, müssen diese auch vor der Umkodierung zwingend der Größe nach aufsteigend oder absteigend sortiert sein. Ist dies nicht der Fall, gibt das *R*-Skript eine Fehlermeldung in der Form aus, dass nicht alle k Merkmalsausprägungen in aufsteigender oder absteigender Reihenfolge sortiert sind.[136] Eine wie hier beschriebene Umkodierung des ordinalskalierten Merkmals ist auch in dem *R*-Paket

[131] Beispielsweise $\beta_1 = 0$ und $\beta_1 = 1$, wie es in *Kapitel 4.1.4* beschrieben wird.
[132] Siehe: *Tabelle 10* auf S. 34.
[133] Siehe: Formel *4.4* auf S. 31.
[134] Siehe: ebd.
[135] Vgl. Hedderich, Sachs (2020), S. 27.
[136] Siehe: Letzte Zeile von *Tabelle 9* auf S. 34.

vorhanden.[137] Die relevanten Codezeilen sind in dem beigelegten *R*-Skript zu finden.[138]

Tabelle 11: Umkodierung des ordinalen Merkmals

k	x_{oc}	x_{nc}	y_{means}
1	1	1,4663	6,3235
2	2	1,6393	6,9084
3	3	2,9108	11,2074
4	4	3,3949	12,8442
5	5	5,5885	20,2609

Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 398-432.

Hierbei wird die Umkodierung verwendet, bei der die ursprüngliche Regressionsgerade des nicht umkodierten und als metrische Variable angenommenen Merkmals, bestehen bleibt. Es ändert sich lediglich die Kodierung der Merkmalsausprägungen, sodass die bedingten Mittelwerte der einzelnen Merkmalsausprägungen exakt auf dieser Geraden liegen. In *Tabelle 11* werden die neue und die alte Kodierung gegenübergestellt.[139]

4.1.4 Lineare Transformation der Kodierungen

Zusätzlich zu der hier ermittelten idealen Kodierung des ordinalskalierten Merkmals, bestehen auch andere ideale Kodierungen, bei denen alle bedingten Mittelwerte der jeweiligen Merkmalsausprägung eine Gerade abbilden. Dies kann bei jeder Geraden der Fall sein, solange deren Steigung nicht 0 oder unendlich beträgt. Um diese idealen Kodierungen zu ermitteln, sind alle Rechenoperationen der linearen Transformation zulässig. Hierzu gehört die Steigung der Regressionsgerade zu ändern, oder den Achsenabschnitt zu verschieben. Dies kann beispielsweise von Vorteil sein, wenn die Kodierung der ersten Merkmalsausprägung den Wert 1 aufweisen soll. Zudem kann es von Vorteil sein, die Kodierung so festzulegen, dass für das ordinalskalierte Merkmal

[137] Siehe: *Anhang*, S. VII.
[138] Siehe: *R*-Skript, Zeile: 398-443.
[139] Im beigelegten *R*-Skript können zudem die Diagnoseplots des Paketes `DescTools` ausgegeben werden. Diese haben keinerlei Auffälligkeiten ergeben, außer, dass die starke Abhängigkeit zwischen Y und den Residuen durch die Umkodierung stark reduziert werden konnte, siehe: *R*-Skript, Zeile: 604-618.

ein Achsenabschnitt von 0 entsteht. In *Tabelle 12* wird dargestellt, dass alle idealen Kodierungen des ordinalskalierten Merkmals, also alle Kodierungen, bei denen die bedingten Mittelwerte eine Gerade abbilden, dieselben Schätzwerte ausgeben. Dies hat den Vorteil, dass wenn für die Aufstellung des Regressionsmodells eine der möglichen idealen Kodierungen verwendet wird, man immer zum selben Schätzergebnis kommt. Wie in *Kapitel 4.1.1* beschrieben, ist ein lineares Regressionsmodell von der Kodierung des ordinalskalierten Merkmals abhängig. Wird hingegen eine Kodierung des exogenen Merkmals gewählt, bei der alle bedingten Mittelwerte eine Gerade abbilden, ist das Modell nicht mehr von der ursprünglichen Kodierung abhängig. In *Tabelle 12* sind zudem in den letzten zwei Zeilen die jeweiligen Achsenabschnitte (β_1) und Steigungen (β_2) abgebildet.

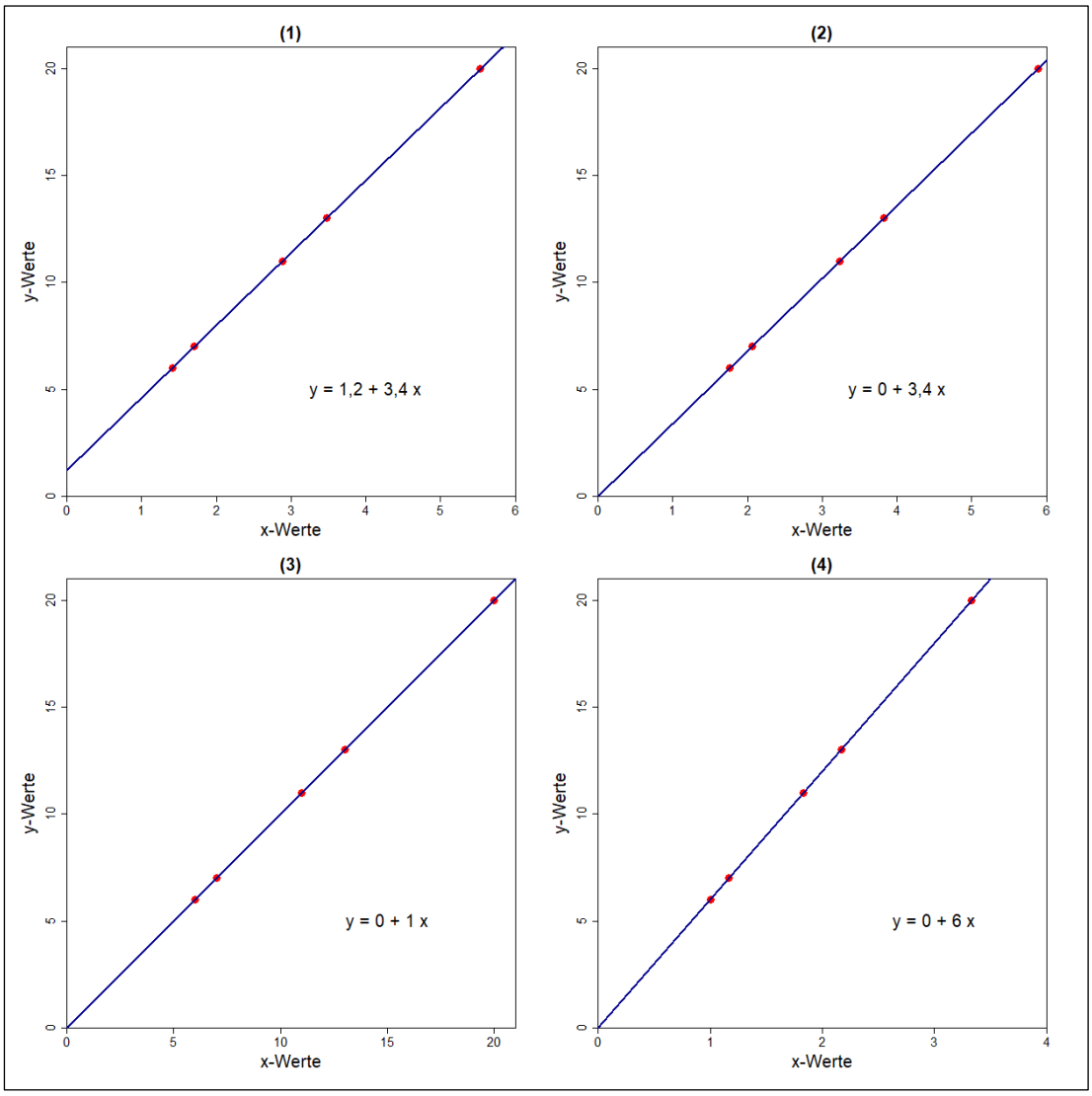
Tabelle 12: Unterschiedliche ideale Kodierungen des ordinalskalierten Merkmals

x_k	$\bar{y}_k$	$x_{nc(1)}$	$\hat{y}_{(1)}$	$x_{nc(2)}$	$\hat{y}_{(2)}$	$x_{nc(3)}$	$\hat{y}_{(3)}$	$x_{nc(4)}$	$\hat{y}_{(4)}$
1	6	1,412	6	1,765	6	6,000	6	1,000	6
2	7	1,706	7	2,059	7	7,000	7	1,167	7
3	11	2,882	11	3,235	11	11,000	11	1,833	11
4	13	3,471	13	3,824	13	13,000	13	2,167	13
5	20	5,529	20	5,882	20	20,000	20	3,333	20
$\beta_1 = 1{,}2$		$\beta_1 = 1{,}2$		$\beta_1 = 0$		$\beta_1 = 0$		$\beta_1 = 0$	
$\beta_2 = 3{,}4$		$\beta_2 = 3{,}4$		$\beta_2 = 3{,}4$		$\beta_2 = 1$		$\beta_2 = 6$	

Quelle: Eigene Darstellung.

Da für jede der hier dargestellten idealen Kodierungen dieselben Schätzwerte ausgegeben werden, sind auch alle Fehlerkennzahlen wie die *R-Squared*, *Adjusted R-Squared*, *SS*, *MSE* und *RMSE* dieselben, genauso wie die Signifikanz der Koeffizienten und des Modells. Grafisch stellen sich die einzelnen Kodierungen wie in *Abbildung 10* dar. Im Folgenden wird berechnet, wie sich die idealen Kodierungen ändern, wenn die Parameter der Regressionsgeraden geändert werden, sodass dennoch dieselben Schätzwerte entstehen.

Abbildung 10: Ideale Kodierungen des ordinalskalierten Merkmals



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 627-689.

Änderung der Steigung der Regressionsgerade um den Faktor a:

$$\hat{y}_i = \hat{y}_j \tag{4.5}$$

$$\beta_1 + \beta_2 \cdot x_i = \beta_1 + \beta_3 \cdot x_j$$

$$\beta_3 = \beta_2 \cdot a$$

$$\beta_1 + \beta_2 \cdot x_i = \beta_1 + \beta_2 \cdot a \cdot x_j$$

$$\beta_2 \cdot x_i = \beta_2 \cdot a \cdot x_j$$

$$x_i = a \cdot x_j$$

$$x_j = \frac{x_i}{a}$$

Änderung des Achsenabschnitts um den Wert b:

$$\hat{y}_i = \hat{y}_j \tag{4.6}$$
$$\beta_1 + \beta_2 \cdot x_i = \beta_3 + \beta_2 \cdot x_j$$
$$\beta_3 = \beta_1 + b$$
$$\beta_1 + \beta_2 \cdot x_i = \beta_1 + b + \beta_2 \cdot x_j$$
$$\beta_2 \cdot x_i = b + \beta_2 \cdot x_j$$
$$x_i = \frac{b}{\beta_2} + x_j$$
$$x_j = x_i - \frac{b}{\beta_2}$$

Wird die Steigung der Regressionsgeraden der linearen Einfachregression um den Faktor a erhöht, dann ändert sich die Kodierung um den Faktor $1/a$. Verschiebt sich der Achsenabschnitt der Regressionsgeraden um den Wert b, dann ändert sich die Kodierung um den Wert $-b/\beta_2$. In *Abbildung 10* ist in *Beispiel (1)* die Kodierung dargestellt, die entsteht, wenn die ursprüngliche Regressionsgerade beibehalten wird. Für eine einheitliche Umgangsform mit ordinalskalierten Variablen kann es sinnvoll sein, für jedes Merkmal die Regressionsgerade der linearen Einfachregression in der Form festzulegen, dass der Achsenabschnitt bei 0 und die Steigung bei 1 liegt, wie es in *Beispiel (3)* dargestellt wird. Die Kodierung der jeweiligen Merkmalsausprägung liegt hierbei exakt bei deren bedingtem Mittelwert der endogenen Variablen. Es lässt sich somit allgemeingültig sagen, dass ohne Rechenaufwand die Merkmalsausprägungen des ordinalskalierten Merkmals mit deren bedingten Mittelwerten der endogenen Variablen Y kodiert werden können, um eine der idealen Kodierungen abzubilden. In diesem Fall würde man sich die Schätzung der Regressionsparameter sparen. Lediglich die bedingten Mittelwerte der endogenen Variablen müssen hierfür berechnet werden.

$$x_{k,nc} = \bar{y}_k \tag{4.7}$$

Sollte man diese Vorgehensweise wählen, lassen sich in einem nächsten Schritt die Kodierungen mit den Mitteln der linearen Transformation wie in *Formel 4.5* und *4.6* nach Belieben anpassen. Soll beispielsweise die erste Merkmalsausprägung mit dem

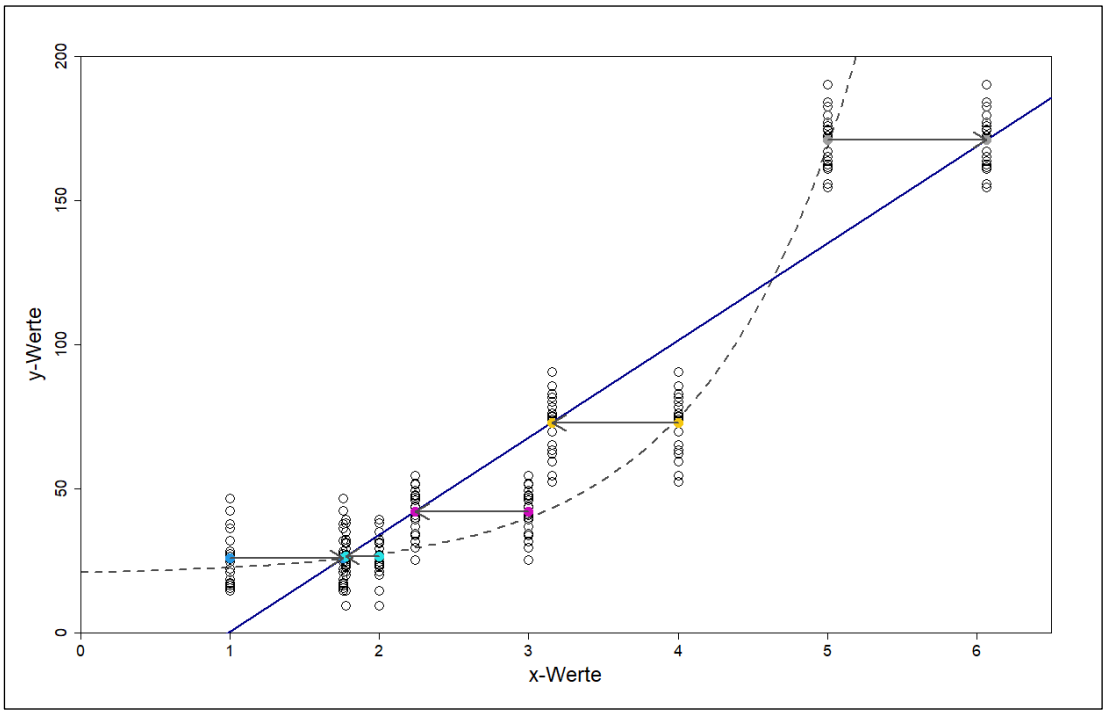
Wert 1 kodiert sein, dann muss die in *Beispiel (3)* dargestellte Kodierung lediglich durch den Divisor 6 geteilt werden, um auf die Kodierung von *Beispiel (4)* zu kommen. Die Steigung erhöht sich dadurch um den Faktor 6.

4.1.5 Umkodierung bei nichtlinearem Zusammenhang zwischen exogener und endogener Variablen

In den bisher besprochenen Beispielen wurde ein linearer Zusammenhang zwischen den ordinalskalierten exogenen Variablen und der endogenen Variablen vermutet. Besteht stattdessen ein nichtlinearer Zusammenhang, wie beispielsweise ein exponentieller Zusammenhang zwischen exogener und endogener Variablen, verstößt das untersuchte Merkmal gegen die Annahmen des linearen Regressionsmodells, dass ein linearer Zusammenhang bestehen muss.[140] Ist das der Fall, kann die hier beschriebene Umkodierung beim Aufbau eines linearen Zusammenhangs helfen. Dies soll anhand des folgenden Beispiels veranschaulicht werden. Wie in *Abbildung 11* dargestellt, besteht zwischen X und Y näherungsweise ein exponentieller Zusammenhang.[141] Die Variable X ist ordinalskaliert. Da der exponentielle Verlauf in einem linearen Regressionsmodell nicht dargestellt werden kann, wird die Variable umkodiert, wodurch ein linearer Zusammenhang geschaffen wird. Dies funktioniert nur für ordinalskalierte Variablen und darf nicht bei kardinalskalierten Variablen durchgeführt werden, da hinter deren metrischer Größe ein tatsächlich beobachtbarer Wert steht und nicht nur eine subjektiv gewählte Kodierung. Die Neukodierung ist in *Abbildung 11* dargestellt. Anhand der grauen Pfeile ist zu erkennen, wie die Beobachtungen entlang der x-Achse verschoben werden, damit die Regressionsgerade die bedingten Mittelwerte des endogenen Merkmals exakt abbildet.

[140] Vgl. Backhaus, et al. (2021), S. 102, vgl. Hedderich, Sachs (2020), S. 144 ff.
[141] Die hier näherungsweise dargestellte exponentielle Funktion lautet: $y = e^x + 20$.

Abbildung 11: Exponentieller Einfluss des ordinalskalierten Merkmals



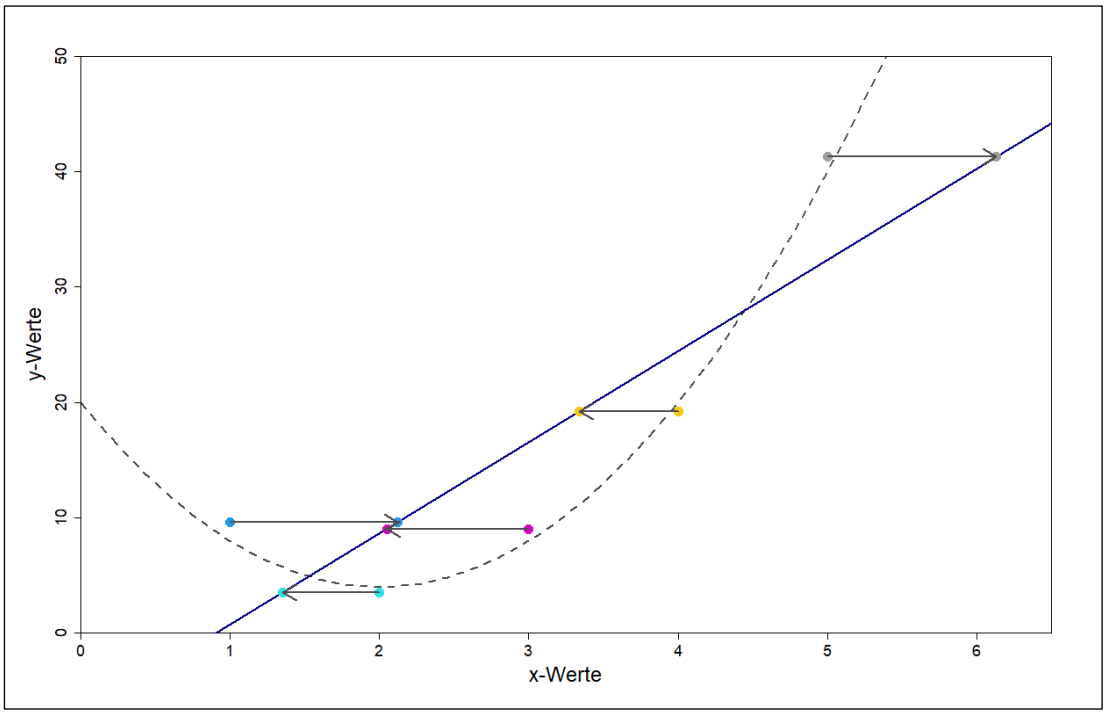
Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 692-764.

Statt des linearen Zusammenhangs muss hierfür mindestens ein kausaler Zusammenhang zwischen den beiden Variablen bestehen. Das heißt, dass ein höheres X in jedem Fall zu einem höheren Y oder in jedem Fall zu einem niedrigeren Y führe muss. Dies gilt nicht für alle einzelnen Beobachtungen, sondern für die bedingten Mittelwerte von Y. Ein u-förmiger Verlauf ist somit nicht zulässig. Diese Bedingung lässt sich mit dem Test aus *Kapitel 3.1* überprüfen.

Besteht hingegen ein quadratischer Zusammenhang zwischen X und Y, dann würde man bei der Umkodierung die Merkmalsausprägungen der exogenen Variablen überkreuzt umkodieren. Dies wird in *Abbildung 12* abgebildet.[142] Zur übersichtlicheren Darstellung wurden hierbei die Beobachtungswerte weggelassen und lediglich die bedingten Mittelwerte der endogenen Variablen vor und nach der Umkodierung dargestellt. Der hier betrachtete quadratische Zusammenhang wäre für die Umkodierung kein Problem, wenn lediglich die rechte oder linke Hälfte jenseits des Tiefpunktes durch die Beobachtungspunkte abgebildet werden würde.

[142] Die hier näherungsweise dargestellte quadratische Funktion lautet: $y = 4x^2 - 16x + 20$.

Abbildung 12: Quadratischer Einfluss des ordinalskalierten Merkmals



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 767-839.

Das Problem liegt somit im Wesentlichen in dem u-förmigen Funktionsverlauf, der eine kausale Beziehung zwischen X und Y widerlegt. Dies führt, wie in *Abbildung 12* dargestellt zu einer Überkreuzumkodierung, wodurch die Rangwertreihe des ordinalskalierten Merkmals verletzt wird. In diesem Fall sollte man das ordinalskalierte Merkmal entweder als Dummy-Variable in das lineare Regressionsmodell mit einfließen lassen, oder man muss die Parameter der quadratischen Funktion durch ein nichtlineares Regressionsmodell schätzen. Es zeigt sich, dass jeder zumindest kausale Zusammenhang zwischen der exogenen und der endogenen Variablen in einen linearen Zusammenhang umkodiert werden kann. Dies hat den Vorteil, dass die Parameter des nichtlinearen Modells nicht geschätzt werden müssen und dass mehrere exogene Variablen in einer multiplen Regression verwendet werden können, auch wenn kein linearer Zusammenhang zur endogenen Variablen besteht. Die hier beschriebene Umkodierung stellt somit eine Alternative zu den deutlich komplizierteren nichtlinearen Transformationen der exogenen Variablen dar.[143]

[143] Vgl. Backhaus, et al. (2021), S. 103 f.

4.1.6 Validierung des neuen Regressionsmodells

Nachdem durch die Umkodierung die optimierte Zuverlässigkeit (Reliabilität) des Modells unter Beweis gestellt wurde, gilt es im Folgenden die Gültigkeit (Validität) des optimierten Modells zu prüfen. Die Validierung des Umkodierungsverfahrens teilt sich in folgende Punkte auf:

- Zur Prüfung der Modelle wird eine fünffach Kreuzvalidierung durchgeführt, bei der der *Root Mean Square Error* für jedes Modell berechnet wird.
- Ob das umkodierte Modell besser als das Modell mit ursprünglicher Kodierung ist, wird anhand eines *t-Tests* über einen Vergleich zweier Mittelwerte auf Signifikanz überprüft.
- Es wird ein Konfidenzintervall der Neukodierungen x_{nc} aufgestellt, um zu eruieren, ob Soll- und Ist-Kodierung signifikant voneinander abweichen.

Kreuzvalidierung

Um ein Modell zu validieren, ist es üblich, den verwendeten Datensatz in zwei Teile zu unterteilen.[144] Die Trainingsdaten stellen dabei den Teil der Daten dar, der zur Modellierung genutzt werden.[145] Das hierbei aufgestellte Modell wird anschließend anhand der Testdaten auf dessen Prognosegüte überprüft.[146] Ein Modell kann eine gute Anpassung an die Trainingsdaten darstellen, jedoch eine schlechte Prognosefähigkeit aufweisen.[147] Die Güte des Modells ist somit anhand von Datensätzen zu kontrollieren, die nicht mit in die Modellerstellung eingeflossen sind.[148] Üblich wäre es den gegebenen Datensatz in Trainingsdaten (ca. 80%), Validierungsdaten (ca. 10%) und Testdaten (ca. 10%) zu unterteilen.[149] Die Trainingsdaten werden dafür verwendet, das Modell zu erstellen. Anschließend werden unterschiedliche Modelle, bzw. unterschiedliche Methoden der Modellerstellung anhand der Validierungsdaten miteinander verglichen und somit eine Modellauswahl, bzw. eine Auswahl an relevanten Parametern getroffen.[150] Das ausgewählte Modell wird anschließend anhand der

[144] Vgl. Backhaus, Erichson, Weiber (2015), S. 321.
[145] Vgl. ebd.
[146] Vgl. Backhaus, et al. (2021), S. 154.
[147] Vgl. ebd.
[148] Vgl. Backhaus, Erichson, Weiber (2015), S. 321.
[149] Vgl. ebd., S. 322.
[150] Vgl. Backhaus, Erichson, Weiber (2015), S. 321.

Testdaten auf dessen Güte überprüft.[151] Durch den Validierungsdatensatz wird verhindert, dass das Modell an den Testdaten abgestimmt wird und dadurch ein an die Testdaten adaptiertes Modell erstellt wird.[152] Der Validierungsdatensatz dient somit lediglich der Beurteilung der Güte des Lernprozesses.[153] Darüber hinaus ist es üblich statt der einmaligen Unterteilung in Trainings- und Validierungsdaten, eine Kreuzvalidierung durchzuführen.[154] In dem hier betrachteten Fall ist eine Modellauswahl, im Hinblick auf eine Wahl zwischen verschiedenen Modellierungsmethoden oder der Auswahl relevanter Modellparameter, nicht von Bedeutung, da sich lediglich mit der linearen Einfachregression und der damit einhergehenden Umkodierung des ordinalskalierten exogenen Merkmals beschäftigt wird. Dadurch kann auf einen Validierungsdatensatz verzichtet werden. Somit wird der Datensatz lediglich in Test- und Trainingsdaten unterteilt. Um den Einfluss einer besonders guten oder besonders schlecht geeigneten Stichprobe der Testdaten zu eliminieren, wie es der Fall wäre, würde man den Datensatz nur einmal in Trainings- und Testdaten unterteilen, wird stattdessen eine fünffach Kreuzvalidierung durchgeführt.[155] Dadurch befindet sich jeder Beobachtungswert einmal in den Testdaten. In diesem Fall wird anhand der Trainingsdaten das Modell, sowie die Neukodierung ermittelt und anschließend deren Einfluss auf die Testdaten dargestellt.[156] Hierbei gilt, dass in jedem Fall der Datensatz, mit dem die Neukodierung des ordinalskalierten Merkmals vorgenommen wird und der Datensatz, mit dem die Parameter geschätzt werden, exakt übereinstimmen sollten. Anderenfalls würde man nicht die Voraussetzung erfüllen, dass die bedingten Mittelwerte der endogenen Variablen auf der Regressionsgeraden liegen.

[151] Vgl. Backhaus, Erichson, Weiber (2015), S. 321.
[152] Vgl. Hirschle (2021), S. 53.
[153] Vgl. Backhaus, Erichson, Weiber (2015), S. 321.
[154] Vgl. Hirschle (2021), S. 53.
[155] Vgl. ebd.
[156] Eine Kreuzvalidierung mit den in *R* verfügbaren Paketen ist in diesem Fall nicht möglich, da diese nach einem fertigen Modell zur Validierung verlangen. Stattdessen werden mit den Trainingsdaten die ursprünglichen Parameter geschätzt, sowie die bedingten Mittelwerte aller Merkmalsausprägungen von X und die daraus resultierende Umkodierung. Dies lässt sich mit den in *R* verfügbaren Paketen nicht abbilden, weswegen für die Kreuzvalidierung ein eigenes Skript angefertigt wurde. Siehe: *R*-Skript, Zeile: 842-1082.

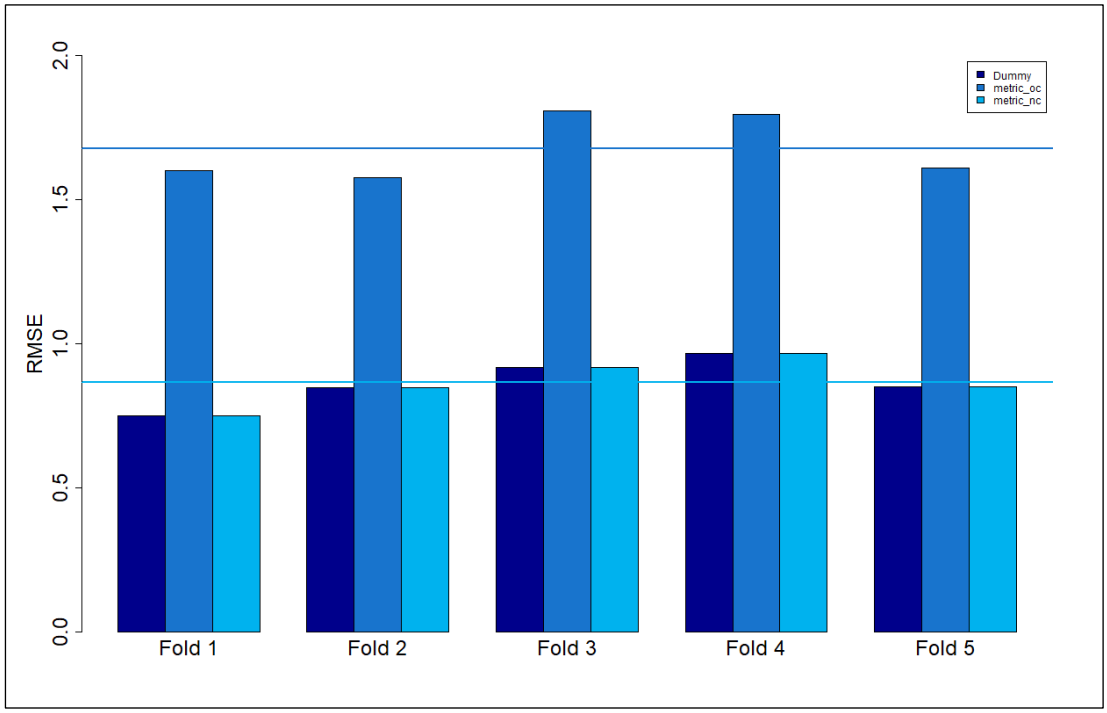
Tabelle 13: Fünffach Kreuzvalidierung der drei Regressionsmodelle

	Dummy	*metric_oc*	*metric_nc*
Fold 1	0,7488	1,5996	0,7488
Fold 2	0,8477	1,5780	0,8477
Fold 3	0,9184	1,8085	0,9184
Fold 4	0,9681	1,7948	0,9681
Fold 5	0,8507	1,6104	0,8507
mean RMSE	0,8667	1,6783	0,8667

Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 842-1082.

Bei der fünffach Kreuzvalidierung wird der Datensatz zufällig in fünf gleichgroße Teile unterteilt.[157] Mit jeweils vier der fünf Teile wird das Regressionsmodell, sowie die neuen Kodierungen der exogenen Variablen berechnet. Der *Root Mean Square Error* wird anschließend für die 20% der Daten berechnet, die nicht bei der Berechnung des Modells mit eingeflossen sind.

Abbildung 13: Fünffach Kreuzvalidierung der drei Regressionsmodelle



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 842-1120.

[157] Vgl. Hirschle (2021), S. 53.

Um einen Überblick über die Validität der Modelle zu erhalten, wird der Durchschnitt der *Root Mean Square Errors* zwischen den fünf *Folds* für jedes der drei Modelle gebildet. Die Ergebnisse sind in *Tabelle 13* dargestellt. Zudem wird für die Kreuzvalidierung *Abbildung 13* ausgegeben, die die Ergebnisse aus *Tabelle 13* visualisiert. Hier ist der jeweilige durchschnittliche *Root Mean Square Error* für Testdaten als vertikale Linie dargestellt. Es zeigt sich, dass der *RMSE* für Testdaten bei dem Modell mit Dummy-Variablen und dem metrischen umkodierten Modell exakt gleich groß ist. Der Abstand zwischen den beiden horizontalen Linien repräsentiert die Verbesserung des *RMSE* durch die Umkodierung. Des Weiteren zeigt sich eine hohe Stabilität des Modells, dadurch dass der *RMSE* bei allen fünf *Folds* nahezu gleich hoch ausgeprägt ist und auch bei dem Vergleich zwischen dem *RMSE* von Trainings- und Testdaten ähnlich hoch ist. Im ersten *Fold* ist der *RMSE* für Testdaten sogar geringer als der für Trainingsdaten.[158] Wird der vorliegende Datensatz in Trainings- und Testdaten unterteilt, so ist es für die Umkodierung des ordinalskalierten Merkmals essenziell, dass jede Merkmalsausprägung von X mindestens einmal in dem Trainingsdatensatz vorkommt. Ansonsten kann durch den fehlenden bedingten Mittelwert dieser Merkmalsausprägung keine Umkodierung vorgenommen werden. In so einem Fall empfiehlt es sich, entweder die Stichprobe der Trainingsdaten neu zu ziehen oder bewusst einen Beobachtungswert den Trainingsdaten zuzuordnen.

t-Test über Vergleich zweier Mittelwerte

Im Folgenden soll untersucht werden, ob das umkodierte Modell signifikant besser ist als das Modell mit der alten Kodierung. Die hierfür gängigen Tests wie beispielsweise die Untersuchung des *Akaike-Informationskriteriums* oder die Durchführung einer *ANOVA* zur Analyse der erklärten Varianz der Modelle kommt in diesem Fall nicht in Frage, da all diese *R*-Funktionen nicht die korrekte Anzahl an Freiheitsgraden für das umkodierte Modell berücksichtigen. Stattdessen wird ein *t*-Test über den Vergleich zweier Mittelwerte durchgeführt, indem der *Mean Absolute Error* beider Modelle für Testdaten verglichen wird. Hierbei wird getestet, ob der durchschnittliche absolute Fehler des umkodierten Modells signifikant kleiner ist als der des ursprünglichen Modells. Für die korrekte Spezifikation der Funktion ist darauf hinzuweisen,

[158] Siehe: *R*-Skript, Zeile: 1084-1102.

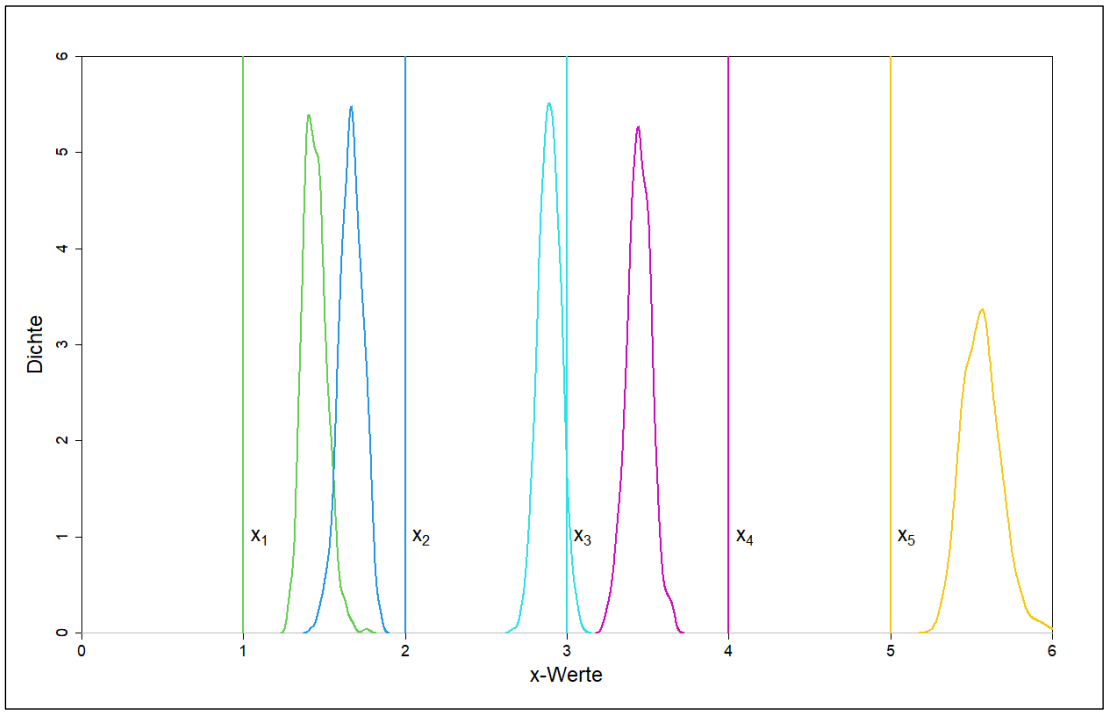
dass bei den beiden Modellen nicht davon ausgegangen werden kann, dass die Standardabweichung der absoluten Fehler gleich groß ist.[159] Dieser Test wird für alle fünf Gruppen der Testdaten durchgeführt. In all diesen Gruppen entsteht auf dem 95%-Niveau ein signifikantes Ergebnis. Bei vier von fünf sogar auf dem 99%-Niveau. Sollte das Ergebnis signifikant sein, kann man sicher sein, dass das umkodierte Modell signifikant besser ist, da mit dem *t*-Test überprüft wird, ob der *MAE* für das umkodierten Modell signifikant kleiner ist.

Konfidenzintervall der Neukodierungen

Die Neukodierung des ordinalskalierten exogenen Merkmals ist von der alten Kodierung, sowie den Parametern des ursprünglichen metrischen Modells abhängig.[160] Die Parameter werden geschätzt und sind somit Zufallsvariablen. Sie hängen damit von der Stichprobe ab, auf Basis derer das Modell erstellt wird. Die neue Kodierung variiert je nach Stichprobe, die gezogen wird. Um den Einfluss der Stichprobe zu untersuchen, wird für jede Neukodierung ein Konfidenzintervall ermittelt, in dem die Kodierung liegt.[161] In *Abbildung 14* werden die Dichtefunktionen der einzelnen Umkodierungen dargestellt, wie auch die ursprüngliche Kodierung in derselben Farbe. Es ist zu erkennen, dass die Soll- und Ist-Kodierung in jedem Fall signifikant voneinander abweichen, außer bei der Kodierung der dritten Merkmalsausprägung (x_3).[162] Die in *Abbildung 14* dargestellten Dichtefunktionen der Konfidenzintervalle werden mit der Funktion `boot.ci()` erstellt.[163] Das Bootstrap-Verfahren zieht hierfür eine Stichprobe mit Zurücklegen.[164] Hierbei stehen unterschiedliche Methoden zur Ermittlung der Breite des Konfidenzintervalls zur Verfügung.[165] In diesem Fall wird die Methode `basic` verwendet, die als klassisches Bootstrap-Intervall die beobachteten Neukodierungen von 95% der Stichproben beinhaltet.[166]

[159] Siehe: *R*-Skript, Zeile: 1122-1155.
[160] Siehe: *Formel 4.4* auf S. 31.
[161] Hierfür wird das Paket `boot` verwendet, siehe: *R*-Skript, Zeile: 1197-1198.
[162] Es ist darauf hinzuweisen, dass die hier als Dichtefunktion dargestellte Soll-Kodierung lediglich dann gilt, wenn die Parameter des Modells mit der alten Kodierung exakt beibehalten werden.
[163] Siehe: *R*-Skript, Zeile: 1226-1231.
[164] Vgl. Hedderich, Sachs (2020), S. 67.
[165] Vgl. Wollschläger (2020), S. 481.
[166] Vgl. Hedderich, Sachs (2020), S. 403 ff., vgl. Wollschläger (2020), S. 481.

Abbildung 14: Dichtefunktion der neuen Kodierungen



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1158-1271.

Laut dem Testen von Hypothesen über Mittelwerte unterscheidet sich der Erwartungswert für die Neukodierung von der bestehenden Kodierung, wenn diese außerhalb des Konfidenzintervalls der Neukodierung liegt.[167] In dem beigefügten *R*-Skript wird hierfür eine binäre Variable `Significant` ausgegeben, die testet, ob jede Neukodierung von der bestehenden Kodierung signifikant abweicht.[168] Zudem wird ein Antwortsatz zu den besagten Tests ausgegeben. Dieser lautet beispielsweise „`Ordinal sequence observed in 5/5 cases, significant deviation from the old coding in 4/5 cases.`"[169] Damit wird darauf hingewiesen, dass die Reihenfolge zwar für alle fünf Merkmalsausprägungen stimmt, die Neukodierung sich jedoch nur in vier von fünf Fällen von der bestehenden Kodierung unterscheidet. Eine Umkodierung des ordinalskalierten Merkmals ist somit zu empfehlen, da sie in den meisten Fällen von der ermittelten Soll-Kodierung signifikant abweicht.

[167] Sowohl bei dem Testen von Hypothesen über Mittelwerte als auch bei der Aufstellung eines Konfidenzintervalls bei einem heterograden Fall, muss die Abweichung mindestens:
$\delta = \bar{x} - \mu_0 = t(n-1,\alpha) \cdot \frac{s}{\sqrt{n}}$ betragen, damit eine signifikante Abweichung vorliegt.

[168] Siehe: *R*-Skript, Zeile: 1236.

[169] Siehe: *R*-Skript, Zeile: 1246.

4.1.7 Anwendungsfall der Umkodierung

In den vorhergehenden Kapiteln wurde aufgezeigt, dass die Umkodierung des ordinalskalierten Merkmals zu einem besseren Modell im Hinblick auf Validität und Reliabilität führt. Im Folgenden soll aufgezeigt werden, wie die Neukodierung des ordinalskalierten Merkmals interpretiert werden kann und welche praktischen Anwendungsfälle daraus resultieren. Für das hier betrachtete ordinalskalierte Merkmal wird folgende Rating-Skala verwendet (*Abbildung 15*), die schon in ähnlicher Form in *Kapitel 2.1* vorgestellt wurde. Hierbei wurde der Fehler gemacht, jede der möglichen Antwortkategorien zu benennen, wodurch eine Annahme als metrisches Merkmal schwierig wird, da nicht davon ausgegangen werden kann, dass die Abstände zwischen den Merkmalsausprägungen äquidistant sind. Die Umkodierung dieser Variable kann dafür Abhilfe schaffen.

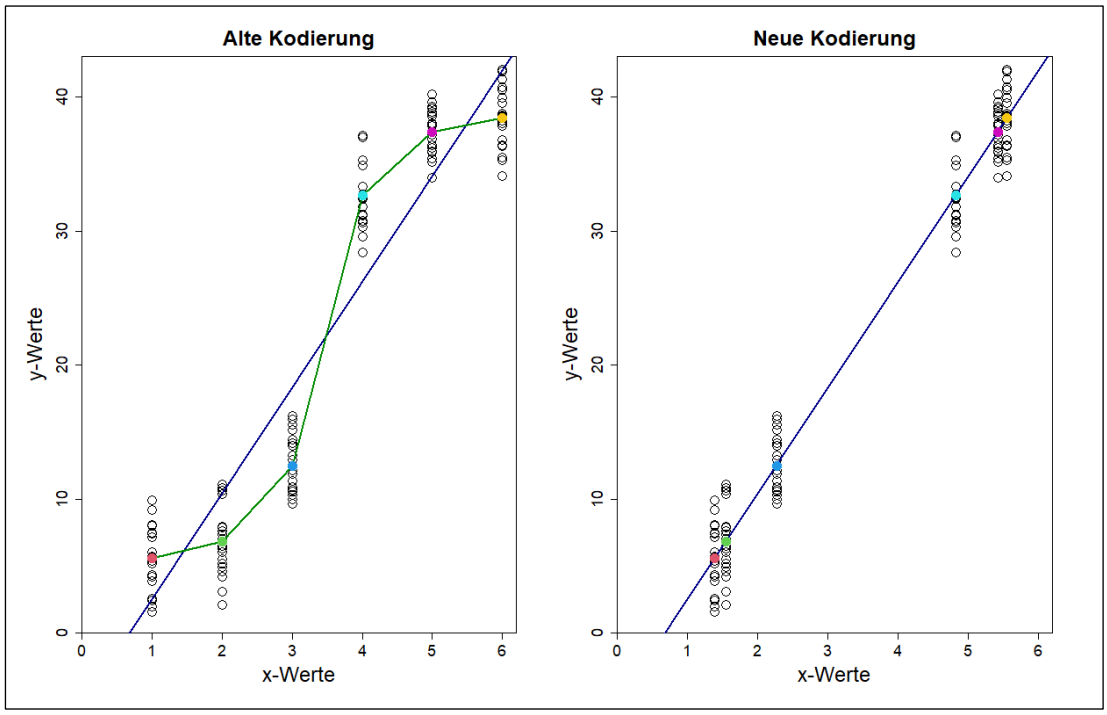
Abbildung 15: Beispiel einer Zustimmungsskala

Wie stark stimmen Sie der folgenden Aussage zu? „…“?					
☐	☐	☐	☐	☐	☐
stimme überhaupt nicht zu	stimme nicht zu	stimme eher nicht zu	stimme eher zu	stimme zu	stimme voll und ganz zu
(1)	(2)	(3)	(4)	(5)	(6)

Quelle: Eigene Darstellung, in Anlehnung an Backhaus, et al. (2021), S. 10.

Anhand *Abbildung 16* ist zu sehen, dass ein positiver Zusammenhang zwischen dem ordinalskalierten exogenen Merkmal und der endogenen Variablen besteht. Umso mehr die Befragten der Aussage zustimmen, desto höher ist die endogene Variable ausgeprägt. Es zeigt sich, dass die Abstände der bedingten Mittelwerte nicht konstant sind. Anhand der Benennung der einzelnen Merkmalsausprägungen ist zu vermuten, dass der Abstand zwischen „stimme überhaupt nicht zu" und "stimme nicht zu“ nicht so groß ist, wie zwischen „stimme eher nicht zu“ und „stimme eher zu“. Dies bestätigt sich auch bei der Betrachtung der bedingten Mittelwerte, die in *Abbildung 16* durch die farbigen Punkte dargestellt werden.

Abbildung 16: Glättung der Rating-Skala



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1274-1365.

Wie schon in den vorhergehenden Kapiteln beschrieben, zeigt sich in *Abbildung 16*, dass die Regressionsgerade vor und nach der Umkodierung exakt dieselbe ist. Voraussetzung für die hier beschriebenen Umkodierung der Befragungsergebnisse anhand der Rating-Skala ist, dass ein linearer Zusammenhang zwischen den beiden Merkmalen zu vermuten ist. Es zeigt sich, dass die Umkodierung nicht zu einer Verletzung der Ordinalskala führt. Ansonsten würde man die Beobachtungswerte überkreuzt verschieben. Die neue und alte Kodierung ist in *Tabelle 14* dargestellt.

Tabelle 14: Interpretation der Umkodierung

k	x_{oc}	x_{nc}	$\bar{y}_k$	x_{diff}	$\bar{y}_{k,diff}$	x_{ratio}	y_{ratio}
1	1	1,388	5,568	–	–	–	–
2	2	1,551	6,848	0,162	1,280	1,000	1,000
3	3	2,265	12,489	0,715	5,640	4,405	4,405
4	4	4,820	32,645	2,554	20,156	15,741	15,741
5	5	5,419	37,376	0,600	4,731	3,695	3,695
6	6	5,556	38,458	0,137	1,083	0,846	0,846

Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1274-1388.

Die Umkodierung bringt im Hinblick auf die Interpretation der vorliegenden Daten zwei entscheidende Vorteile mit sich:

- Zum einen lässt sich der Steigungsparameter des Regressionsmodells besser interpretieren als bei der ursprünglichen Kodierung.
- Zum anderen können die Abstände zwischen den Merkmalsausprägungen der exogenen Variablen interpretiert werden.

Geht man anhand der neuen Kodierung auf der x-Achse um eine Einheit weiter erhöht sich die endogene Variable um den Steigungsparameter $(7{,}891)$. Für die ursprüngliche Kodierung lässt sich diese Aussage nicht treffen, da, wie man anhand *Abbildung 16* sehen kann, für jede Merkmalsausprägung von X, mehr oder weniger von der ermittelten Regressionsgerade abgewichen wird. Problem hierbei ist, dass zwischen den Neukodierungen der exogenen Variablen nicht mehr länger die Schrittweite 1 liegt. Berechnet man die Abstände zwischen den Neukodierungen erhält man die in *Tabelle 14* in Spalte x_{diff} dargestellten Ergebnisse. Die hierbei ermittelten Abstände zwischen den Neukodierungen lassen sich besonders gut interpretieren, indem man sie zusätzlich durch deren ersten Wert teilt. Hieraus ergibt sich die x_{ratio}. Hiermit lässt sich sagen, dass der Abstand zwischen den Merkmalsausprägungen (*3*) und (*4*) ca. 15-mal so groß ist, wie zwischen den Merkmalsausprägungen (*1*) und (*2*). Es zeigt sich zudem, dass der Abstand zwischen den Merkmalsausprägungen (*1*) und (*2*) und der zwischen (*5*) und (*6*) miteinander vergleichbar ist, genauso wie die Abstände zwischen (*2*) und (*3*), und zwischen (*4*) und (*5*). Das Verhältnis zwischen den Abständen der Neukodierung (x_{ratio}) ist auch zwischen den bedingten Mittelwerten der endogenen Variablen (y_{ratio}) zu beobachten. Durch die Umkodierung lassen sich somit die Abstände zwischen den gegebenen Antwortkategorien „stimme überhaupt nicht zu" und "stimme nicht zu“ ins Verhältnis setzen zu dem Abstand zwischen „stimme eher nicht zu“ und „stimme eher zu“.[170]

[170] Vgl. Hedderich, Sachs (2020), S. 845.

4.2 Kennzahl für die Güte eines ordinalskalierten Merkmals

4.2.1 Grundidee der Kennzahl

Betrachtet man die Eignung oder Nichteignung eines ordinalskalierten Merkmals für die Verwendung in einem linearen Regressionsmodell, stellt sich die Frage, wie geeignet oder ungeeignet die betrachtete Variable ist. Diese Thematik soll im folgenden Kapitel untersucht werden. Die Idee der Umkodierung des ordinalskalierten Merkmals wird dabei weitergeführt, um eine Kennzahl zu entwickeln, die die Eignung eines ordinalen Merkmals für eine Annahme als metrisches Merkmal ohne Umkodierung abbildet. Zur Aufstellung der Kennzahl wird untersucht, wie stark die Soll-Kodierung von der bestehenden Ist-Kodierung abweicht. Sie stellt somit die alte und die neue Kodierung des Merkmals gegenüber. Die zu ermittelnde Kennzahl soll folgende Eigenschaften aufweisen. Der Wertebereich der Kennzahl soll zwischen 0 und 1 liegen. Sie soll 0 sein, wenn das ordinalskalierte Merkmal für die Verwendung als metrische Variable in einem linearen Regressionsmodell ideal geeignet ist. Dies ist der Fall, wenn alle bedingten arithmetischen Mittel eine Gerade bilden. Umso mehr von dieser Bedingung abgewichen wird, desto näher soll der Wert gegen 1 tendieren. Die in *Kapitel 4.2.3* aufgestellte Fehlerkennzahl wird anschließend in *Kapitel 4.2.4* anhand verschiedener Zahlenbeispiele getestet.

4.2.2 Zerlegung der Streuung der endogenen Variablen

Schaut man sich die Werte der exogenen und der endogenen Variablen in einem Streudiagramm an, ist zu erkennen, dass die Beobachtungswerte stark streuen. Die Abweichungen zwischen den Beobachtungspunkten und dem Mittelwert von Y lassen sich in einen erklärten Teil und einen nicht erklärten Teil der Abweichung unterteilen.[171] Der nicht erklärte Teil der Streuung wird als Residuum bezeichnet und ergibt sich aus der Abweichung zwischen Beobachtungswert und Schätzwert des Modells.[172] Der erklärte Teil der Streuung ist die Abweichung zwischen dem Schätzwert und dem Mittelwert der endogenen Variablen.[173]

[171] Vgl. Backhaus, et al. (2021), S. 87.
[172] Vgl. ebd.
[173] Vgl. ebd.

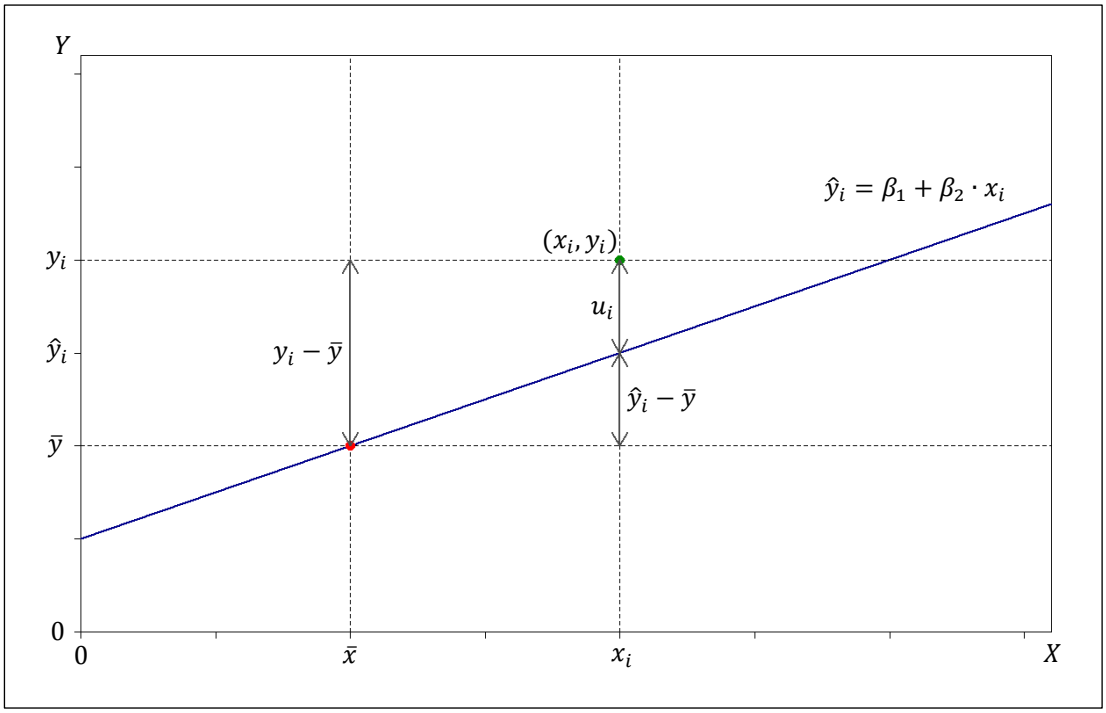
Zusammen ergibt sie die Gesamtabweichung.[174] Somit besteht folgender Zusammenhang:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \quad (4.8)^{175}$$

$$Gesamtstreuung = erklärte\ Streuung + nicht\ erklärte\ Streuung$$

Die hier rechnerisch dargestellten Komponenten der Abweichung stellen sich grafisch wie in *Abbildung 17* dar. Hierbei ist die Regressionsgerade in blau, der Mittelwert der beiden Merkmale in rot und der Beobachtungswert in grün dargestellt.

Abbildung 17: Zersetzung der Abweichung vom Mittelwert



Quelle: Backhaus, et al. (2021), S. 87, siehe: *R*-Skript, Zeile: 1391-1412.

Um die Streuung zu ermitteln, werden die Abweichungen jeder Beobachtung quadriert und aufsummiert.[176] Daraus ergibt sich das Prinzip der Zerlegung der Gesamtstreuung der endogenen Variablen.[177]

[174] Vgl. Backhaus, et al. (2021), S. 87.
[175] Quelle: ebd.
[176] Vgl. ebd.
[177] Vgl. ebd.

$$SST = SSE + SSR \quad (4.9)^{178}$$

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

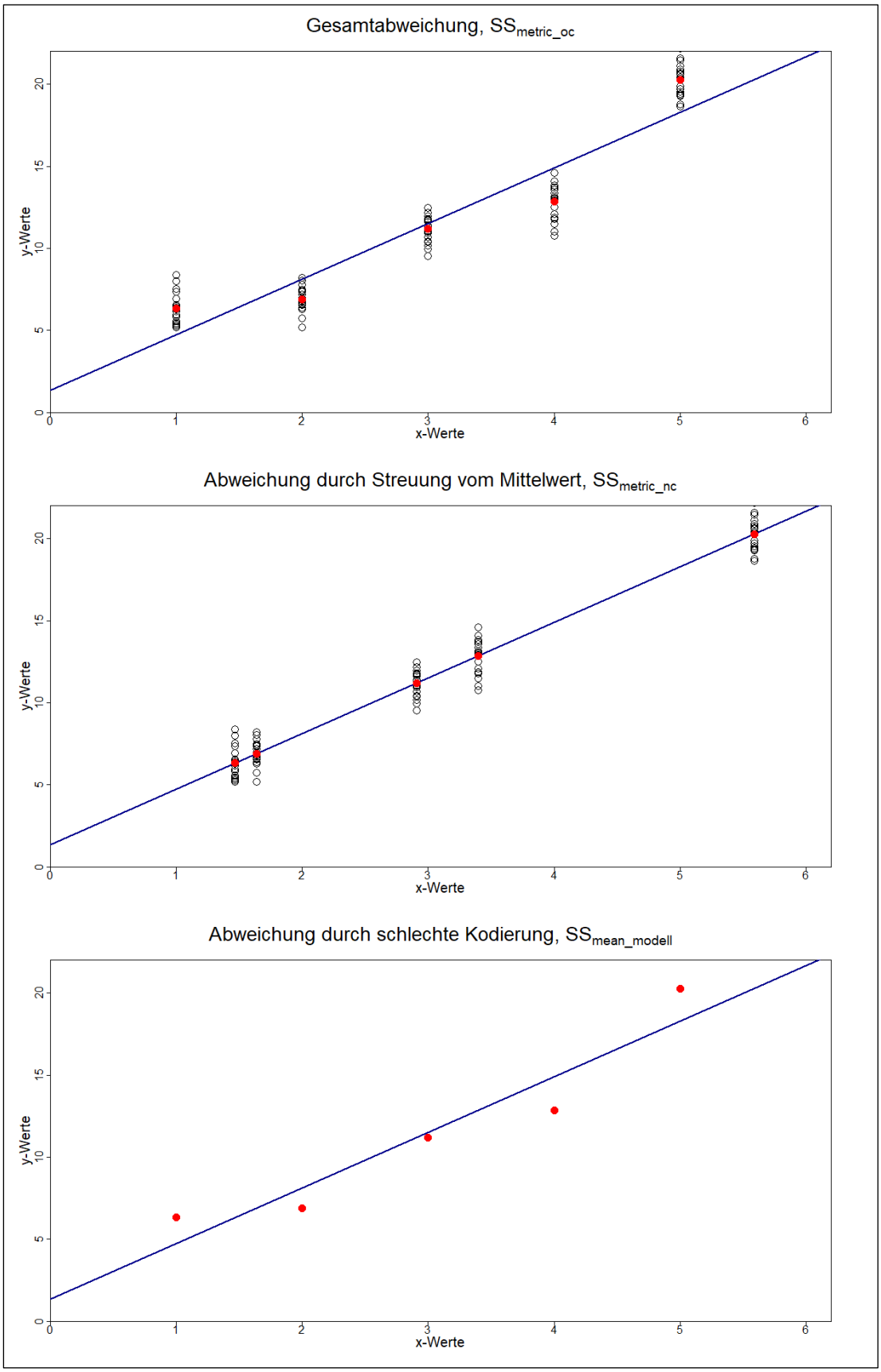
$$Gesamtstreuung = erklärte\ Streuung + nicht\ erklärte\ Streuung$$

Die *Total Sum of Squares* werden hierbei als SST beschrieben und stellen die Gesamtstreuung der Daten dar.[179] Durch die Gegenüberstellung der Schätzwerte des Regressionsmodells und des Mittelwertes von Y ergibt sich die erklärte Streuung bzw. die *Explained Sum of Squares*, die in der Formel als SSE beschrieben wird.[180] Auf dieselbe Art und Weise errechnet sich die residuale Streuung SSR (*Residual Sum of Squares*).[181] Die in den vorhergehenden Kapiteln beschriebenen SS (*Sum of Squares*), die beispielsweise in dem Dashboard als Fehlerkennzahl zum Einsatz kommen, beziehen sich auf die in *Formel 4.9* hergeleiteten *Residual Sum of Squares*.[182]

Auf dasselbe Prinzip wie bei der Zerlegung der Streuung, wird bei der Berechnung der hier beschriebenen Kennzahl zurückgegriffen. Zur Zerlegung der Streuung in einzelne Komponenten wird ein weiteres Modell erstellt. Dieses besteht aus den x-Werten vor der Umkodierung und aus den $\bar{y}_k$-Werten für die jeweilige Merkmalsausprägung von X. In diesem Modell bestehen daher keine Abweichungen zwischen den Datenpunkten und den bedingten Mittelwerten der jeweiligen Merkmalsausprägung von X. Die Parameter des linearen Regressionsmodells zu diesen Daten entsprechen exakt denen, die auch bei dem Regressionsmodell mit den Ursprungsdaten entstehen. Im Folgenden wie auch im beigelegten *R*-Skript wird dieses Modell als *meanmodell* bezeichnet.[183] Dieses Modell bezieht sich auf die Abweichung zwischen Soll- und Ist-Kodierung und die daraus entstehende Streuung der bedingten Mittelwerte von der ermittelten Regressionsgeraden. Es besteht genauso wie die metrischen Modelle vor und nach der Umkodierung aus n Beobachtungen.

[178] Quelle: Backhaus, et al. (2021), S. 88, Janssen, Laatz (2017), S. 408, Hedderich, Sachs (2020), S. 822.
[179] Vgl. Backhaus, et al. (2021), S. 88.
[180] Vgl. ebd.
[181] Vgl. ebd.
[182] Siehe *Kapitel 3.2*.
[183] Siehe: *R*-Skript, Zeile: 1474-1475.

Abbildung 18: Komponenten der Streuung bei ordinalskalierten Merkmalen



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1415-1529.

In *Abbildung 18* sind die drei Regressionsmodelle zu erkennen, die zur Zerlegung der Streuung benötigt werden. Alle drei Modelle basieren auf den exakt gleichen Parametern. Zwischen den drei Modellen besteht folgender Zusammenhang der *Sum of Squares*:

$$SS_{metric\_oc} = SS_{metric\_nc} + SS_{mean\_modell} \quad (4.10)^{184}$$

$$(y_{i,k} - \hat{y}_{i,k}) = (y_{i,k} - \bar{y}_k) + (\bar{y}_k - \hat{y}_{i,k})$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\left(y_{i,k} - \hat{y}_{i,k}\right)^2 = \sum_{i=1}^{N}\sum_{k=1}^{K}\left(y_{i,k} - \bar{y}_k\right)^2 + \sum_{i=1}^{N}\sum_{k=1}^{K}\left(\bar{y}_k - \hat{y}_{i,k}\right)^2$$

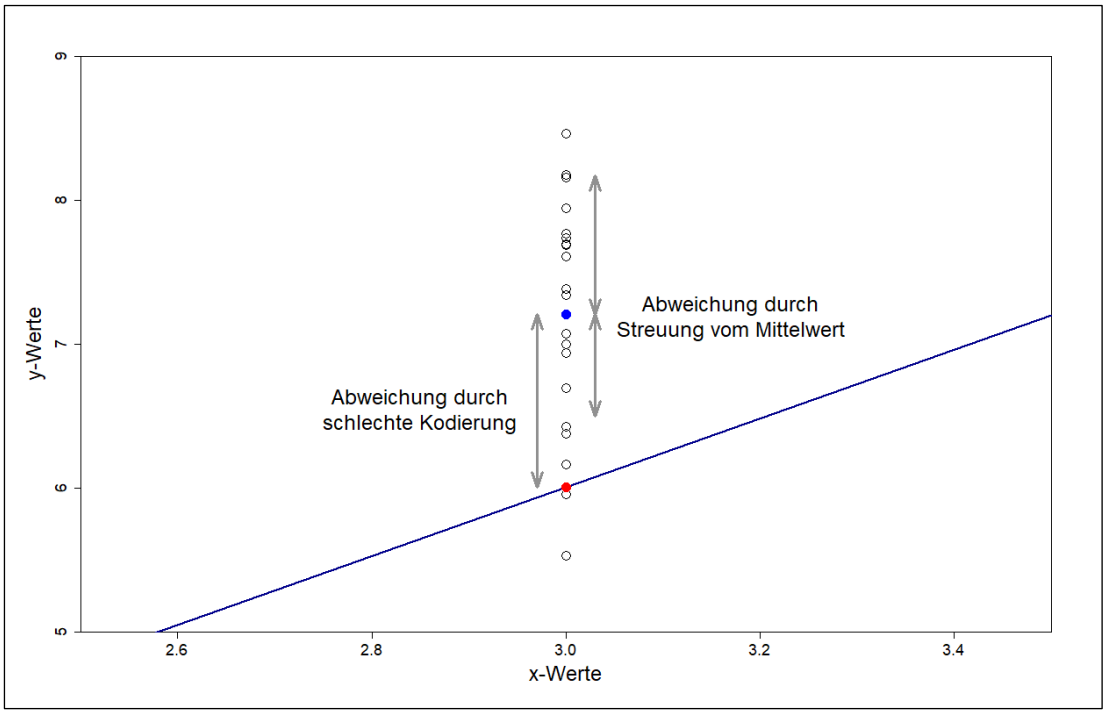
$$\begin{matrix}\textit{Nicht erklärte}\\ \textit{Streuung}\ (SSR)\end{matrix} = \begin{matrix}\textit{Abweichung}\\ \textit{durch Streuung}\\ \textit{vom Mittelwert}\end{matrix} + \begin{matrix}\textit{Abweichung}\\ \textit{durch schlechte}\\ \textit{Kodierung}\end{matrix}$$

Die gesamten *Residual Sum of Squares* des Modells, bei dem das ordinalskalierte Merkmal als metrisch skaliert angenommen wird, stehen links von dem Gleichheitszeichen. Diese setzen sich zusammen aus der Summe der *Sum of Squares* des metrischen Modells, nachdem das ordinalskalierte Merkmal umkodiert wurde und den *Sum of Squares* des *mean-modell*. Der nicht erklärte Teil der Streuung des metrischen Modells (SSR) mit alter Kodierung wird somit weiter aufgeschlüsselt in zwei Teile. Zum einen entsteht die residuale Streuung durch eine schlechte Kodierung des ordinalskalierten Merkmals. Zum anderen besteht die residuale Streuung aus der Abweichung der Beobachtungen von ihrem bedingten Mittelwert. Die Abweichung zwischen Beobachtung und Schätzung des metrischen Modells mit unveränderter Kodierung – hier als *Sum of Squares* dargestellt – ist somit auf zwei voneinander zu unterscheidende Sachverhalte zurückzuführen. Zum einen weichen die Beobachtungen von ihrem bedingten Mittelwert ab. Diese Abweichung wird hier durch die *Sum of Squares* des metrischen Modells nach der Umkodierung dargestellt. Zum anderen entsteht ein Schätzfehler durch die schlechte Kodierung, der in dieser Formel durch die *Sum of Squares* des *mean-modell* dargestellt wird. Dies führt dazu, dass die

[184] Die hier dargestellte Formel und die Aufteilung der Streuung in zwei unterschiedliche Modelle stellt eine Eigenleistung des Verfassers dar. Dennoch ist darauf hinzuweisen, dass diese Vorgehensweise stark der „Prüfung der Linearität einer Regression“ ähnelt (vgl. Hedderich, Sachs (2020), S. 805 f.). Alternativ zu der im Folgenden hergeleiteten Kennzahl kann auch der in der angegebenen Quelle dargestellte *F*-Test verwendet werden, um die Eignung eines ordinalskalierten Merkmals als metrische Variable in einem linearen Regressionsmodell zu prüfen.

bedingten Mittelwerte der jeweiligen Merkmalsausprägungen von dem Schätzwert der Regressionsgeraden bei metrischer Annahme abweichen, wenn das ordinalskalierte Merkmal nicht passend umkodiert wird. In *Abbildung 19* werden die beiden Fehlerquellen der Schätzung grafisch dargestellt.

Abbildung 19: Darstellung der zwei Fehlerquellen der Schätzung



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1532-1577.

Für die Ermittlung der Kennzahl müssen die beiden Fehlerquellen unterschieden werden. Dies liegt daran, dass lediglich die Abweichung zwischen dem bedingten Mittelwert und der Regressionsgeraden durch eine Umkodierung eliminiert werden kann. Dieser Teil des Fehlers ist auf eine unpassende Kodierung des ordinalskalierten Merkmals zurückzuführen. Die Abweichungen durch eine Streuung vom Mittelwert lässt sich bei einer multiplen Regression durch andere Merkmale erklären, oder ist auf einen nichtbeobachtbaren Restterm zurückzuführen, der im Modell nicht berücksichtigt werden kann.[185] Dieser Teil des Fehlers bleibt bei der linearen Einfachregression in jedem Fall bestehen, da er durch die betrachtete ordinalskalierte Variable nicht erklärt werden kann. Durch die Umkodierung soll der jeweilige bedingte Mittelwert geschätzt werden, wie es auch bei einer Verwendung von Dummy-Variablen der Fall

[185] Vgl. Backhaus, et al. (2021), S. 78.

ist. Um die Kennzahl daran auszurichten, wie stark die Fehlerkennzahlen durch eine Umkodierung verbessert werden könnten, wird lediglich der Fehler durch die schlechte Kodierung berücksichtigt. Verwendet man *Formel 4.10* und teilt auf beiden Seiten durch SS_{metric_oc} erhält man:

$$\begin{aligned} 1 &= \frac{SS_{metric\_nc}}{SS_{metric\_oc}} + \frac{SS_{mean\_modell}}{SS_{metric\_oc}} \\ 1 &= \frac{76{,}0869}{320{,}2612} + \frac{244{,}1743}{320{,}2612} \\ 100\% &= 23{,}76\% + 76{,}24\% \\ \textit{Nicht erklärte Streuung} &= \textit{Abweichung durch Streuung vom Mittelwert} + \textit{Abweichung durch schlechte Kodierung} \end{aligned} \qquad (4.11)$$

Auf diese Weise kann berechnet werden, welcher Anteil der nicht erklärten Streuung auf die fehlerhafte Kodierung und welcher Teil auf die Streuung vom jeweiligen bedingten Mittelwert zurückzuführen ist. Die relativen Anteile der Zerlegung der Streuung sind in *Abbildung 20* grafisch dargestellt. Die grün markierte Fläche stellt den Anteil der Gesamtstreuung dar, der auch vor der Umkodierung durch das metrische Regressionsmodell erklärt werden kann. Die orange markierte Fläche repräsentiert den Teil der Streuung, der durch die schlechte Kodierung resultiert, so wie die rot markierte Fläche den Teil der Streuung darstellt, der durch die Abweichung vom bedingten Mittelwert entsteht. Die rote und die orange Fläche ergeben zusammen die nicht erklärte Streuung, die sich auch mit folgender Formel berechnen lässt:

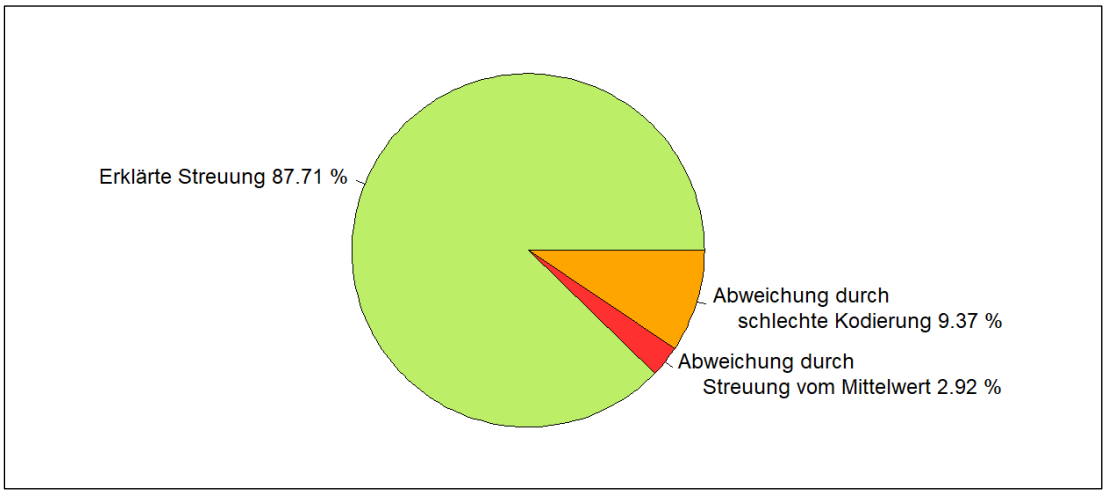
$$Anteil\ der\ nicht\ erklärte\ Streuung = 1 - R^2 = \frac{SSR}{SST} \qquad (4.12)^{186}$$

$$1 - R^2 = 1 - 0{,}8771 = 12{,}29\%$$

Bezieht man die beiden relativen Anteile der Komponenten der nicht erklärten Streuung auf die gesamte nicht erklärte Streuung ergeben sich die in *Abbildung 20* dargestellten Werte.

[186] Quelle: Backhaus, Erichson, Weiber (2015), S. 40.

Abbildung 20: Anteile an der Gesamtstreuung



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1580-1692.

Durch die in *Kapitel 4.1* thematisierte Umkodierung des ordinalskalierten exogenen Merkmals kann der hier orange markierte Teil der Streuung durch das neue Modell erklärt werden. Statt einer $87{,}71$ prozentigen Streuungserklärung erreicht man somit eine $97{,}08$ prozentige Streuungserklärung. Die Zerlegung der Streuung liefert somit interessante Erkenntnisse darüber, wie hoch die Streuungserklärung durch das ursprüngliche Modell bereits ist, und welchen zusätzlichen Anteil von der Gesamtstreuung durch eine Umkodierung erklärt werden kann. Die hieraus gezogen Erkenntnisse werden im Folgenden für die Herleitung der Kennzahl genutzt.

4.2.3 Herleitung der Kennzahl

Die hier zu ermittelnde Kennzahl soll von dem Abstand zwischen dem Schätzwert des metrischen Regressionsmodells ohne Umkodierung und dem arithmetischen Mittel der jeweiligen Merkmalsausprägung abhängen. Der bedingte Mittelwert entspricht dem Schätzwert des Modells mit Dummy-Variablen und dem des umkodierten metrischen Modells. Diese Abstände beschreiben die Größe des Fehlers, der durch eine schlechte Kodierung entsteht, auch wenn in diesem Fall nicht betrachtet wird, wie stark die Kodierungen für X geändert werden. Der hier betrachtete Abstand ist der Teil des Fehlers, der durch eine schlechte Kodierung resultiert und der durch eine Umkodierung zu einer der idealen Kodierungen vollständig eliminiert wird. Der Restterm, der zu einer Abweichung der einzelnen Beobachtung von deren bedingtem arithmetischen Mittel führt, ist durch das Merkmal nicht zu eliminieren und bleibt

somit bestehen. Diese Vorgehensweise hat den Vorteil, dass die Fehlerkennzahl für alle idealen Kodierungen des ordinalskalierten Merkmals denselben Wert ergibt. Der hier betrachtete Abstand hat die Einheit der endogenen Variablen des Regressionsmodells. Der absolute oder relative Abstand zwischen Soll- und Ist-Kodierung von X wird somit nicht direkt betrachtet, da es unendlich viele ideale Kodierungen gibt, bei denen die metrische Regressionsgerade die bedingten Mittelwerte exakt abbildet. Stattdessen wird der Teil des Schätzfehlers ermittelt, der durch die Differenz aus Soll- und Ist-Kodierung resultiert.[187]

Die Abweichung zwischen der Regressionsgeraden und den jeweiligen Mittelwerten wird im Folgenden als systematischer Fehler bezeichnet. Systematische Fehler liegen vor, wenn es bei wiederholter Messung zu einer Über- oder Unterschätzung des wahren Wertes kommt.[188] Dieser Fehler resultiert aus Mängeln bei der Messung, wie sie in diesem Fall durch die schlechte Kodierung des ordinalskalierten Merkmals hervorgerufen werden.[189] Die Abweichung der Beobachtungen zu ihrem bedingten Mittelwert wird als unsystematischer Fehler bezeichnet. Dieser zufällige Fehler ändert sich unvorhersehbar zwischen den einzelnen Beobachtungen und streut um den wahren Mittelwert.[190] Laut dem zentralen Grenzwertsatz folgt er oftmals einer Normalverteilung.[191] Zufallsfehler sind nicht vermeidbar, können aber durch eine Erhöhung der Stichprobe verringert werden.[192] Systematische Fehler können bei sorgsamer Durchführung einer empirischen Erhebung vermieden werden, wie es hier bei der Umkodierung der ordinalskalierten Variable der Fall ist.[193]

Die zwei zu unterscheidenden Quellen der Schätzfehler führen dazu, dass die beiden Fehlerkennzahlen, wie beispielsweise der *RMSE*, für die Annahme als metrische Variable und als Dummy-Variable nicht einfach gegenübergestellt werden können, da nicht bekannt ist, wie groß der Anteil des unsystematischen Fehlers anhand des gesamten Fehlers ist. Die hier aufgestellte Kennzahl orientiert sich an den

[187] Siehe: *Abbildung 19*, S. 57.
[188] Vgl. Backhaus, et al. (2021), S. 30.
[189] Vgl. ebd.
[190] Vgl. ebd.
[191] Vgl. ebd., S. 102.
[192] Vgl. ebd., S. 30.
[193] Vgl. ebd.

Abweichungen zwischen dem bedingten Mittelwert und der Regressionsgeraden. Ziel ist es somit eine Varianzkennzahl zu entwickeln, die lediglich den Schätzfehler für die schlechte Kodierung abbildet ohne die Varianz, die durch die Abweichung zum jeweiligen Mittelwert entsteht. Die Normierung der Kennzahl hat den Vorteil, dass sie bei 1 liegen kann, auch wenn eine starke Abweichung zum Mittelwert der jeweiligen Merkmalsausprägung vorliegt. So muss beispielsweise das Bestimmtheitsmaß immer im Verhältnis zum maximal möglichen Bestimmtheitsmaß betrachtet werden, das dem des Modells mit Dummy-Variablen entspricht. In diesem Fall ist das Bestimmtheitsmaß der Dummy-Variablen und der umkodierten metrischen Variablen immer dasselbe. Vergleicht man lediglich das Bestimmtheitsmaß bei beiden Methoden, dann kann dieses auch bei der Umwandlung zur Dummy-Variablen sehr schlecht sein, wenn eine hohe Heterogenität der Werte von Y bei den unterschiedlichen Merkmalsausprägungen von X vorliegt. Die Kennzahl bietet somit einen klaren Vorteil zum einfachen Vergleich der Fehlerkennzahlen beider Modelle.

Die hier hergeleitete Kennzahl wird als „*Ordinal-Metric-Transformation-Coefficient*“ (*OMTC*) bezeichnet, da sie die Eignung eines ordinalskalierten Merkmals für die Verwendung als metrische Variable in einem linearen Regressionsmodell abbilden soll. Der Koeffizient beschreibt dabei die Eignung der Variable für die dafür notwenige Transformation auf ein höheres Skalenniveau.

$$OMTC = \frac{SS_{mean\_modell}}{SS_{oc}} \tag{4.13}$$

Der *OMTC* beschreibt die *Sum of Squares* des *mean-modell* im Verhältnis zu den gesamten *Residual Sum of Squares* des metrischen Modells ohne Umkodierung. Dadurch ist die Kennzahl automatisch normiert auf Werte zwischen 0 und 1.

$$0 \leq OMTC \leq 1 \tag{4.14}$$

Der $OMTC$ ist 0, wenn das ordinalskalierte Merkmal ideal als metrische Variable geeignet ist, also wenn keinerlei Abweichungen zwischen den bedingten Mittelwerten und der Regressionsgeraden bestehen ($SS_{mean\_modell} = 0$). Hingegen ist der $OMTC$ gleich 1, wenn keinerlei Streuung zwischen den bedingten Mittelwerten der

jeweiligen Merkmalsausprägung und den einzelnen Beobachtungen besteht ($SS_{mean\_modell} = SS_{oc}$). Zwischen den drei Werten der *Sum of Squares* besteht folgender Zusammenhang:

$$SS_{oc} = SS_{nc} + SS_{mean\_modell} \tag{4.15}$$

$$SS_{mean\_modell} = SS_{oc} - SS_{nc}$$

$$OMTC = \frac{SS_{oc} - SS_{nc}}{SS_{oc}} \tag{4.16}$$

Der *OMTC* kann daher alternativ aus der normierten Differenz der beiden *Sum of Squares* vor und nach der Umkodierung berechnet werden. Die im Zähler stehende Differenz ist die absolute Verbesserung der *Sum of Squares* durch die Umkodierung. Wird der *OMTC* für das Zahlenbeispiels aus *Kapitel 4.2.2* berechnet, ergibt sich:

$$OMTC = \frac{SS_{oc} - SS_{nc}}{SS_{oc}} = \frac{320{,}2612 - 76{,}0869}{320{,}2612} = 0{,}7624 \tag{4.17}$$

Es ist zu erkennen, dass der *OMTC* exakt den relativen Anteil der schlechten Kodierung an der nicht erklärten Gesamtstreuung beschreibt. Somit lässt der *OMTC* neben der normierten Verbesserung der *Sum of Squares* durch die Umkodierung folgende Interpretation zu: Liegt der *OMTC* bei $0{,}7624$, dann sind $76{,}24\%$ der nicht erklärten Streuung auf die nicht lineare Kodierung des ordinalskalierten Merkmals zurückzuführen. Dieser Anteil wird somit durch die Umkodierung eliminiert, bzw. zusätzlich zum ursprünglichen Modell erklärt. Da die *Sum of Squares* des Modells mit Dummy-Variablen und des umkodierten Regressionsmodells übereinstimmen, kann der *OMTC* auch folgendermaßen berechnet werden:

$$SS_{Dummy} = SS_{nc} \tag{4.18}$$

$$OMTC = \frac{SS_{oc} - SS_{Dummy}}{SS_{oc}} \tag{4.19}$$

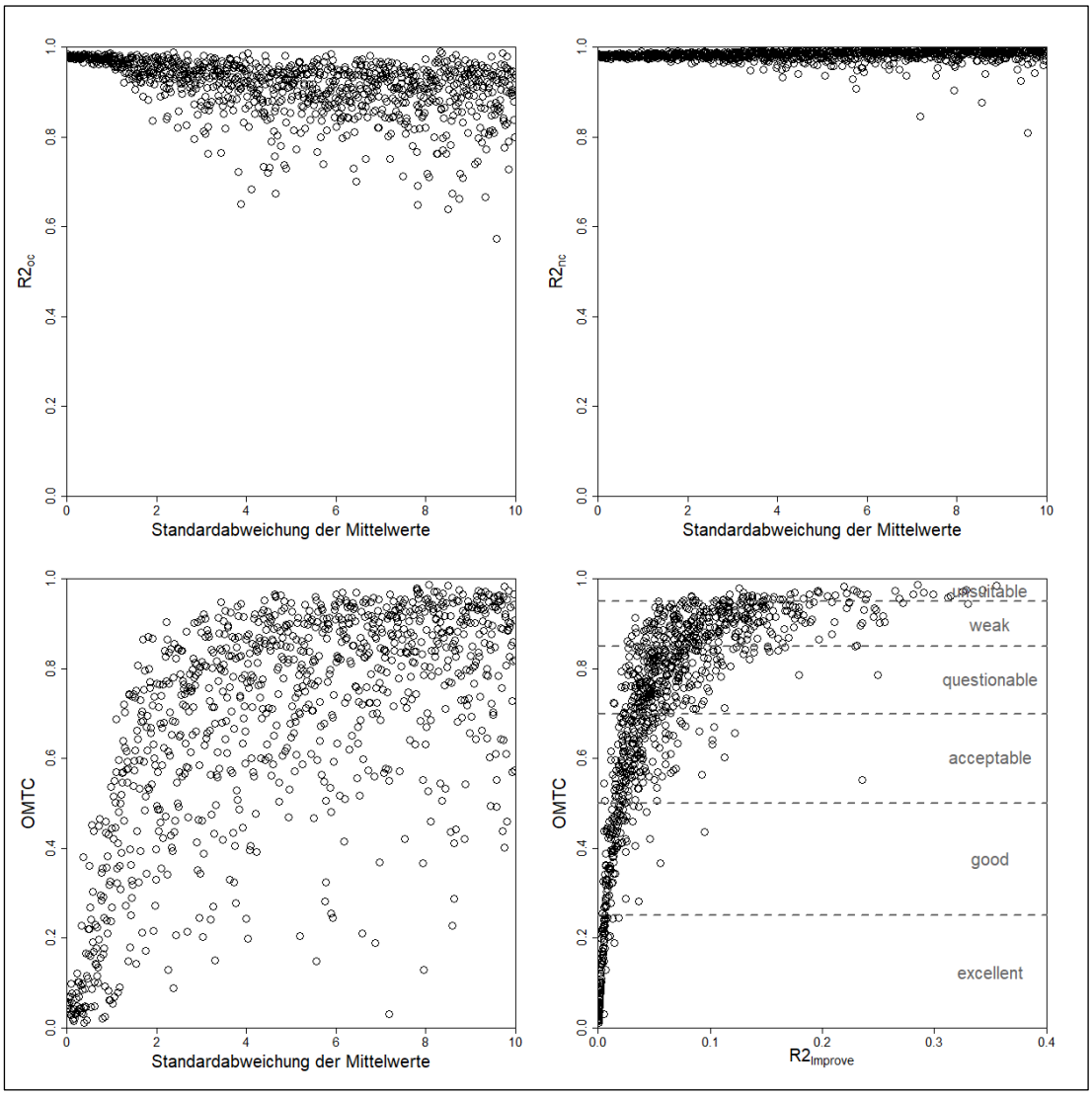
Die hier gewählte Verwendung der *Sum of Squares* bringt eine Reihe von Vorteilen mit sich. Zum einen sind die Freiheitsgrade für die hier hergeleitete Kennzahl nicht von Belang. Somit kann der *OMTC* unabhängig von Test- und Trainingsdaten verwendet werden. Dadurch, dass die *Sum of Squares* verwendet werden und nicht nur

einzelne Wertepaare für Kodierung der Merkmalsausprägung von X und deren bedingten Mittelwert, berücksichtigt der *OMTC*, dass manche Merkmalsausprägungen häufiger vorkommen als andere. Somit werden von der Kennzahl unterschiedliche Werte für n, k und $p(k)$ berücksichtigt. Alternativ zu der normierten Verbesserung der *Sum of Squares* könnte die Kennzahl auch als normierte Verbesserung des *Root Mean Square Errors* hergeleitet werden. Die Verwendung der *Sum of Squares* bietet jedoch den Vorteil, dass hierbei die vorliegenden Freiheitsgrade keine Berücksichtigung finden. Somit werden keine zwei unterschiedlichen Kennzahlen für Trainings- und Testdaten benötigt. Zudem besteht bei dem *OMTC* eine eindeutige Interpretierbarkeit des Wertes, der darstellt, welcher relative Anteil der nicht erklärten Streuung durch eine Umkodierung eliminiert werden könnte.

4.2.4 Test der Kennzahl

Im Folgenden soll der *OMTC* für unterschiedliche Zahlenbeispiele getestet werden. Hierfür werden 1000 Datensätze erstellt und dazu der *OMTC*, sowie das Bestimmtheitsmaß berechnet. Beim ersten Datensatz bilden die bedingten Mittelwerte nahezu exakt eine Gerade ab. Der *OMTC* liegt somit näherungsweise bei 0. Über die 1000 Datensätze hinweg steigt die Standardabweichung der bedingten Mittelwerte von ihrem ursprünglich vorgegebenen Wert. Die Standardabweichung der einzelnen 100 Beobachtungspunkte von ihrem bedingten Mittelwert bleibt dabei konstant. In der Grafik oben links in *Abbildung 21* ist zu erkennen, dass sich das Bestimmtheitsmaß vor der Umkodierung verschlechtert, wenn die bedingten Mittelwerte stärker von einer Geraden abweichen. Wie die Grafik oben rechts zeigt, ist dies für das Bestimmtheitsmaß des umkodierten Modells nicht der Fall. Anhand der Grafik unten links ist zu sehen, dass der *OMTC* von der Abweichung der bedingten Mittelwerte zur Regressionsgeraden abhängig ist. Es zeigt sich, dass bei einer größeren Abweichung der *OMTC* schnell ansteigt. Zu interpretieren ist diese Grafik am oberen Rand der Datenpunkte. Dies liegt daran, dass die Abweichungen der Mittelwerte von einer Geraden durch Zufallszahlen mit vorgegebener Standardabweichung erstellt wurden. Diese Abweichung wird bei Zufallszahlen nicht in jedem Fall vollständig ausgereizt, weswegen viele Beobachtungspunkte unterhalb der zu interpretierenden Kurve liegen.

Abbildung 21: OMTC mit steigender Streuung der bedingten Mittelwerte



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1707-1796.

In der Grafik unten rechts ist zu sehen, dass der *OMTC* einen klaren Zusammenhang mit der Verbesserung des Bestimmtheitsmaßes aufweist. Der hier dargestellte *R2-Improve* ist die Differenz der Bestimmtheitsmaße vor und nach der Umkodierung und wird in Prozentpunkten angegeben. Dadurch, dass bei der Berechnung des *OMTC* quadrierte Werte anhand der *Sum of Squares* verwendet werden, klumpt diese Kennzahl am oberen Rand des Wertebereichs. Dies ist für die Interpretation des *OMTC* von entscheidender Bedeutung. Aus den zahlreichen Beobachtungspunkten der Kennzahl bei unterschiedlichen Zahlenbeispielen ergeben sich die in *Tabelle 15* dargestellte Beurteilungen der Eignung eines ordinalskalierten Merkmals für die Verwendung in

einem linearen Regressionsmodell als metrische Variable, die auch in *Abbildung 21* zu finden sind.

Tabelle 15: Beurteilung der Kennzahl

Kennzahl	Empfehlung
$0 - 0{,}25$	Exzellent
$0{,}25 - 0{,}5$	Gut
$0{,}5 - 0{,}7$	Akzeptabel
$0{,}7 - 0{,}85$	Fragwürdig
$0{,}85 - 0{,}95$	Schwach
$0{,}95 - 1$	Ungeeignet

Quelle: Eigene Darstellung.

Es ist darauf hinzuweisen, dass die Kennzahl nur für ein einziges Merkmal Geltung findet und nicht für einen ganzen Datensatz, auch wenn dieser nur aus ordinalskalierten Merkmalen bestehen sollte. Es werden lediglich die *Sum of Squares* der Einfachregression dargestellt und nicht die der multiplen Regression. Sollten mehrere ordinalskalierte Merkmale in einem Datensatz vorhanden sein, sollten alle Variablen separat auf ihre Eignung untersucht werden. Zusätzlich zu der Höhe der hier beschriebenen Kennzahl ist in jedem Fall zu testen, ob das betrachtete ordinalskalierte Merkmal einen kausalen Zusammenhang zu der endogenen Variablen aufweist. In dem programmierten *R*-Paket kann für jedes ordinalskalierte Merkmal eine Varianzanalyse, sowie die hergeleitete Kennzahl ausgegeben werden (*Tabelle 16*).

Tabelle 16: Varianzanalyse und Kennzahl

	metric_oc	***metric_nc***	***mean_modell***
abs, Residual SS	320,2612	76,0869	244,1743
rel, Residual SS in %	100	23,7578	76,2422
R2	0,8771	0,9708	
OMTC	0,7624	*questionable*	

Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1799-1867.

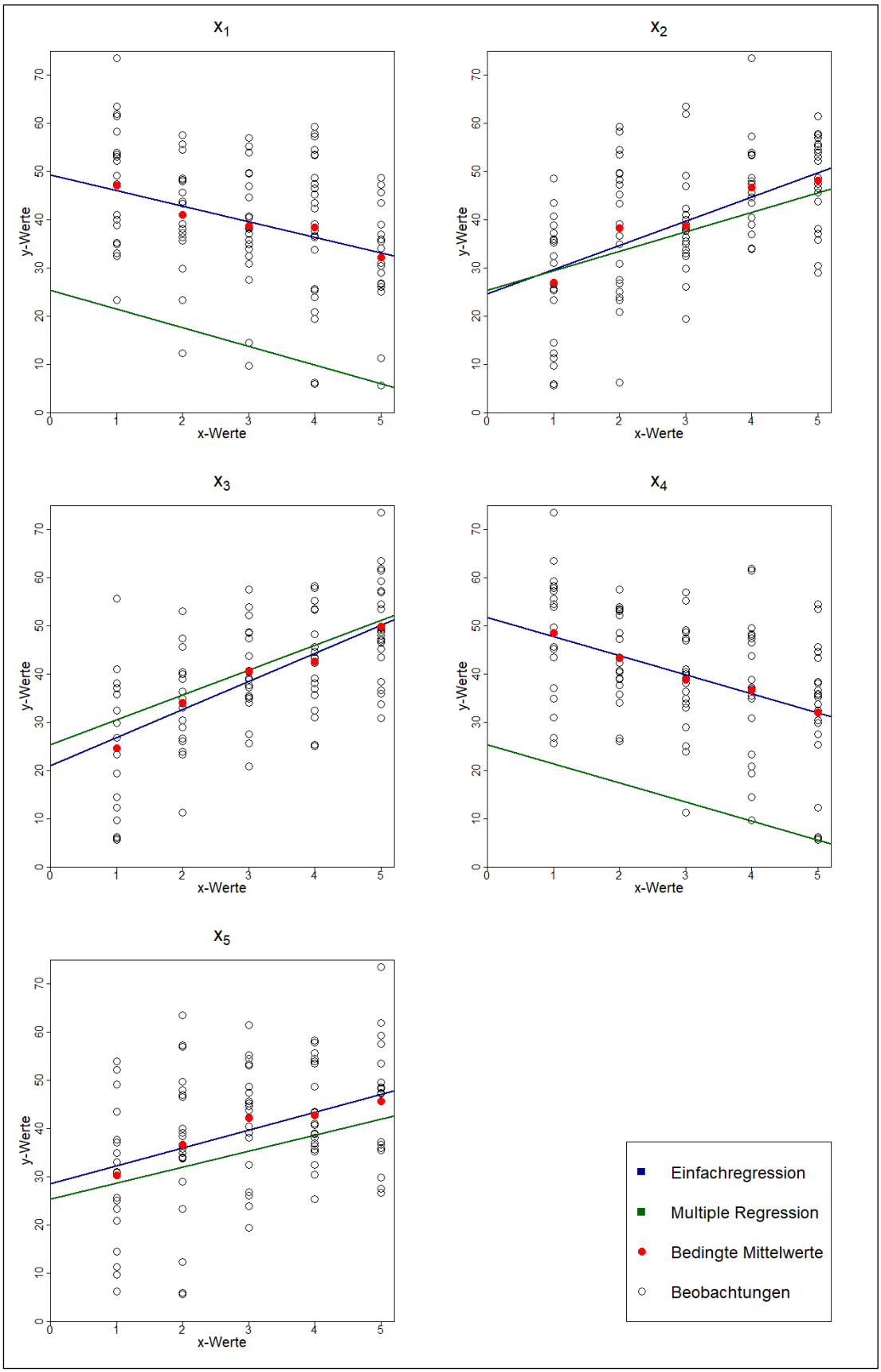
Es ist zu erkennen, dass die *Sum of Squares* durch eine Umkodierung deutlich verbessert werden können. Die Eindrücke der *Sum of Squares* sollten dennoch nicht trügen. Die Varianzerklärung des umkodierten Modells beträgt nicht ein Vielfaches des ursprünglichen Modells, wie die *Sum of Squares* suggerieren. Die Anteile der Streuungserklärung sind besser durch einen Vergleich der Bestimmtheitsmaße zu interpretieren. Von besonderem Interesse, in Bezug auf die Umkodierung des ordinalskalierten Merkmals, ist, wie sich die Anteile der nicht erklärte Streuung auf die beiden Fehlerquellen aufteilen. Anhand *Tabelle 16* ist zu erkennen, dass in dem hier gewählten Zahlenbeispiel ca. 76% der nicht erklärten Streuung, auf die Kodierung des ordinalskalierten Merkmals zurückzuführen sind. Lediglich ca. 24% der nicht erklärten Streuung sind auf die Abweichungen der Beobachtungen von deren bedingten Mittelwert zurückzuführen. Der *OMTC* beträgt somit $0{,}7624$. Dieser hohe Wert deutet auf eine fragwürdige Verwendung des ordinalskalierten Merkmals als metrische Variable in einem linearen Regressionsmodell hin.

4.3 Umkodierung in einem multiplen Regressionsmodell

Der zuvor festgestellte positive Einfluss auf die Schätzwerte einer linearen Einfachregression soll im Folgenden auch für multiple Regressionsmodelle geprüft werden. Hierbei werden mehrere Einflussgrößen auf die endogene Variable berücksichtigt.[194] Die Größe J beschreibt die Anzahl an unabhängigen Variablen, die in das multiple Regressionsmodell einfließen.[195] Bei der multiplen Regression wird nur ein *Intercept* geschätzt.[196] Somit wird durch die ordinalskalierte Variable nur ein weiterer Koeffizient geschätzt und es geht nur ein Freiheitsgrad verloren, wenn das jeweilige Merkmal als metrisch angenommen wird. Um die Einflüsse der Umkodierung auf die multiple Regression zu prüfen, wird ein Datensatz mit 100 Beobachtungswerten randomisiert erstellt.[197] Der Datensatz besteht aus einer endogenen metrischen Variablen Y, sowie fünf exogenen ordinalskalierten Merkmalen mit jeweils fünf Merkmalsausprägungen, beispielsweise anhand einer Bewertung mit Rating-Skalen.

[194] Vgl. Hedderich, Sachs (2020), S. 824.
[195] Vgl. Backhaus, et al. (2021), S. 81.
[196] Vgl. ebd.
[197] Siehe: *R*-Skript, Zeile: 1872-1921.

Abbildung 22: Ergebnisse der multiplen Regression



Quelle: Eigene Darstellung, siehe: *R*-Skript, Zeile: 1872-2131.

Es ist darauf hinzuweisen, dass der Datensatz in der Form erstellt wurde, dass die bedingten Mittelwerte für jedes exogene Merkmal aufsteigend oder absteigend sortiert sind.[198] Anschließend wird für jedes Merkmal eine Einfachregression durchgeführt, genauso wie eine multiple Regression mit allen fünf erklärenden Merkmalen.[199] Die Ergebnisse hiervon sind in *Abbildung 22* dargestellt. Es zeigt sich, dass die Einfachregression sowie die multiple Regression einen sehr ähnlichen Steigungsparameter, jedoch teilweise deutlich unterschiedliche Achsenabschnitte aufweisen. Die ordinalskalierten Merkmale werden umkodiert, sodass die bedingten Mittelwerte, die in *Abbildung 22* als rote Punkte dargestellt werden, exakt eine Gerade abbilden. Der Steigungsparameter, sowie der Achsenabschnitt, die für die Umkodierung verwendet werden, stammen aus der linearen Einfachregression des jeweiligen Merkmals.[200] Wie es in *Kapitel 4.1.4* bereits beschrieben wurde, ist es irrelevant welche Kodierung ein ordinalskaliertes Merkmal aufweist, solange die bedingten Mittelwerte eine Gerade abbilden. Somit wäre es auch möglich das ordinalskalierte Merkmal anhand der Parameter der multiplen Regression umzukodieren, genauso wie jede Kombination hiervon (Steigung der Einfachregression und Achsenabschnitt der multiplen Regression oder andersrum) möglich wäre. Jedes dieser Modelle unterscheidet sich zwar im Hinblick auf die ermittelten Parameter, gibt jedoch exakt dieselben Gütemaße im Hinblick auf die Schätzergebnisse aus. Dies belegen die in *Kapitel 4.1.4* dargestellten Ergebnisse, dass die Parameter auf Basis derer eine Umkodierung vorgenommen werden, irrelevant sind, solange die bedingten Mittelwerte eine Gerade abbilden. Interessanterweise ist zu beobachten, dass die Umkodierung des ordinalskalierten Merkmals für die Einfachregression in jedem Fall zu einer Besserung der Schätzergebnisse führt, für die multiple Regression jedoch schlechtere Gütekriterien aufweist.[201] Ziel der Umkodierung ist es, die bedingten Mittelwerte der jeweiligen Merkmalsausprägung zu schätzen. Dies würde das multiple Regressionsmodell verbessern, wie die Ergebnisse mit Dummy-Variablen unter Beweis stellen.[202] Dieses Modell gibt genauere Schätzwerte aus, jedoch nur für Fehlerkennzahlen, die die 21

[198] Siehe: *R*-Skript, Zeile: 1872-2014.
[199] Siehe: *R*-Skript, Zeile: 2016-2032.
[200] Siehe: *R*-Skript, Zeile: 2140-2144, 2162-2166.
[201] Siehe: *R*-Skript, Zeile: 2221-2269.
[202] Siehe: *R*-Skript, Zeile: 2207.

verbrauchten Freiheitsgrade nicht berücksichtigen. Wird nur eines der fünf exogenen Merkmale umkodiert, verschlechtert das ebenfalls die Güte des Modells.[203] Dass die Umkodierung dennoch korrekt durchgeführt wurde, ist daran zu erkennen, dass für die lineare Einfachregression jede umkodierte Variable ein genaueres Modell ergibt als mit den ursprünglichen Kodierungen.[204] Zudem wird das Modell der multiplen Regression meistens verbessert, wenn nur zwei der fünf Merkmale in das Modell mit einfließen und diese beiden umkodiert werden.[205] Werden hingegen drei Merkmale berücksichtigt, wird seltener eine Verbesserung erzielt.[206] Verwendet man vier Merkmale in der multiplen Regression, ist nur dann eine Verbesserung zu erzielen, wenn die dritte Variable ausgelassen wird.[207]

Die schlechteren Schätzwerte der Umkodierung bei einer multiplen Regression können auf zwei Sachverhalte zurückzuführen sein. Zum einen ist zu vermuten, dass der negative Einfluss der Umkodierung auf die Multikollinearität zurückzuführen ist. Diese liegt vor, wenn zwischen einer der unabhängigen Größen und den anderen Einflussgrößen eine hohe multiple Korrelation vorliegt.[208] Ein gewisses Maß an Multikollinearität ist für empirische Daten üblich, da diese oftmals ähnliche Informationen hinsichtlich der endogenen Variablen beschreiben.[209] In diesem Fall wird jedoch die Multikollinearität durch die Umkodierung teilweise verstärkt. Durch die Umkodierung der fünf ordinalskalierten Merkmale steigt zwar die Korrelation zwischen den exogenen und der endogenen Variablen in allen fünf Fällen an, jedoch erhöht sich dadurch teilweise auch die Korrelation zwischen den exogenen Variablen untereinander. Stellt man die Korrelationsmatrix der exogenen Variablen vor und nach der Umkodierung gegenüber, fällt auf, dass in exakt der Hälfte aller Fälle die Multikollinearität durch die Umkodierung erhöht und in der anderen Hälfte verringert wurde.[210] Somit ist zu vermuten, dass die Multikollinearität nicht der entscheidende Faktor der schlechteren Schätzwerte ist. Untersucht man zusätzlich die Korrelationen mit den

[203] Siehe: *R*-Skript, Zeile: 2185-2198.
[204] Siehe: *R*-Skript, Zeile: 2221-2269.
[205] Siehe: *R*-Skript, Zeile: 2277-2298.
[206] Siehe: *R*-Skript, Zeile: 2300-2308.
[207] Siehe: *R*-Skript, Zeile: 2310-2337.
[208] Vgl. Hedderich, Sachs (2020), S. 828.
[209] Vgl. Backhaus, et al. (2021), S. 121.
[210] Siehe: *R*-Skript, Zeile: 2339-2358.

Schätzwerten, bzw. mit den Residuen der beiden Modelle, fällt folgendes auf: Durch die Umkodierung wurde die Korrelation zwischen der endogenen Variablen Y und den Residuen des umkodierten Modells stark erhöht.[211] Ziel sollte es jedoch sein, dass die Beobachtungswerte möglichst stark mit den Schätzwerten und möglichst schwach mit den Residuen korrelieren. Es scheint so, dass die in *Kapitel 4.1.1* aufgestellte Bedingung, die ordinalskalierte Variable in der Form umzukodieren, dass die bedingten Mittelwerte eine Gerade abbilden, bei der multiplen Regression keine Geltung hat. Stattdessen sind bei der multiplen Regression die Kodierungen zu suchen, die die Korrelation zwischen Beobachtungs- und Schätzwerten maximieren und damit die Korrelation zwischen Residuen und Beobachtungswerten minimieren. Damit wird die Bedingung der Normalverteilung der Störgrößen erfüllt.[212] Bei der Einfachregression wurde durch die Umkodierung diese Bedingung auch erfüllt. Durch die Verschiebung der Merkmalsausprägungen von X wurde die Korrelation zwischen den Beobachtungs- und den Schätzwerten und damit auch das Bestimmtheitsmaß maximiert. Eine mögliche Lösung dieses Problems könnte es sein, bei der multiplen Regression bei J unabhängigen Merkmalen mit $J-1$ Merkmalen die Regression durchzuführen und die Umkodierung mit den Residuen zu berechnen, um den Einfluss der restlichen Variablen auf die endogene Variable zu eliminieren. Dies sollte für alle ordinalskalierten Variablen des Datensatzes wiederholt werden.[213]

$$x_{k,nc} = \frac{\bar{u}_k - \beta_1}{\beta_2} \tag{4.20}$$

Sobald eine ordinalskalierte exogene Variable umkodiert wurde, fließt diese in das Modell mit ein, um die Residuen für die Neukodierung der nächsten Variablen zu ermitteln. Die ursprüngliche Kodierung wird somit nicht mehr verwendet. Nachdem alle unabhängigen ordinalskalierten Variablen in dieser Form umkodiert wurden, kann durch einen weiteren Durchgang die Kodierung weiter verfeinert werden, da sich durch die Umkodierungen auch das Modell und somit die Residuen verändern. Interessanterweise wird die Güte des Modells durch einen weiteren Durchgang der

[211] Siehe: *R*-Skript, Zeile: 2277-2293.
[212] Vgl. Backhaus, et al. (2021), S. 119.
[213] Siehe: *R*-Skript, Zeile: 2429-2565.

Umkodierung verschlechtert.[214] Dieser Ansatz der korrelationsmaximierenden Kodierung zwischen Beobachtungs- und Schätzwerten, wird in dieser Arbeit nicht mehr weiter ausgeführt.

5. Schlussbetrachtung

5.1 Zusammenfassung der Ergebnisse

In dieser Arbeit wurde die Handhabung ordinalskalierter exogener Variablen in linearen Regressionsmodellen ausführlich aufgearbeitet. So wurden in *Kapitel 2* unterschiedliche Herangehensweisen dargelegt und deren Vor- und Nachteile diskutiert. In *Kapitel 3* wurde ein Dashboard mit diversen Fehlerkennzahlen zur Beurteilung der einzelnen Vorgehensweisen vorgestellt, sowie eine allgemeingültige Handlungsempfehlung hierzu abgegeben. Hierbei gibt das Dashboard die jeweils bessere Umgangsform im Hinblick auf die einzelnen Fehlerkennzahlen aus. Allgemeingültig sollte das Modell gewählt werden, das die besseren Schätzwerte nach der präferierten Fehlerkennzahl ausgibt. Mit der hier beschriebenen Umkodierung aus *Kapitel 4.1* wird eine eigene Umgangsform für das Skalenniveau der komparativen Variablen in linearen Regressionsmodellen vorgestellt. Sie hat den wesentlichen Vorteil, dass man bei der Aufstellung des Regressionsmodells mit einer der idealen Kodierungen, immer zum selben Schätzergebnis kommen wird und somit nicht mehr länger von der rein subjektiv gewählten Kodierung abhängig ist. Kodiert man das ordinalskalierte Merkmal willkürlich, kommt man auf unterschiedliche Schätzwerte des Modells. Die hier vorgestellte Vorgehensweise könnte als Standardmethode dazu dienen, eine einheitliche Umgangsform mit ordinalskalierten Merkmalen zu etablieren. Damit wurde das in der Literatur nur sehr schwach behandelte Thema detaillierter betrachtet. In den meisten Fällen findet man für die Handhabung von ordinalskalierten Merkmalen in Regressionsmodellen lediglich die Transformation auf ein höheres oder niedrigeres Skalenniveau. Dadurch, dass die Regressionsgerade durch alle bedingten Mittelwerte der einzelnen Merkmalsausprägungen von X verläuft, wird durch die Umkodierung

[214] Siehe: *R*-Skript, Zeile: 2568-2648. Diese Codezeilen können beliebig oft wiederholt werden. Hierbei ist zu erkennen, dass sich mit jeder weiteren Durchführung sowohl die *Sum of Squares*, als auch das Bestimmtheitsmaß verschlechtern. Somit sollte man bei einem Durchgang der Umkodierung bleiben und diese anschließend nicht weiter verfeinern, indem die neuen Residuen für die Umkodierung verwendet werden.

die Bedingung erfüllt, dass die Regressionsgerade durch den Schwerpunkt der Beobachtungen verläuft.[215] Dies erfüllt somit den Wissenschaftsgrundsatz, dass bei unabhängiger Durchführung eines Experimentes dieselben Ergebnisse erzielt werden müssen.[216] Dasselbe gilt auch für die Analyse eines Datensatzes. Die Schätzwerte eines Modells sollten nicht davon abhängig sein, wie der Analyst die Merkmalsausprägungen des ordinalskalierten Merkmals kodiert. Kritisch ist hierbei darauf hinzuweisen, dass die Vorteile, die sich durch eine Umkodierung ergeben, wie beispielsweise die genaueren Schätzwerte, auch für eine Umwandlung zu Dummy-Variablen gegeben sind. Das Modell mit dem umkodierten Merkmal verbraucht zudem zwei Freiheitsgrade mehr als das Modell mit Dummy-Variablen und erreicht somit nicht dieselben Ergebnisse für den *RMSE* für Trainingsdaten, auch wenn die Schätzwerte in beiden Fällen dieselben sind. Liegen nur wenige Beobachtungen vor, kann das metrische Modell, dass deutlich weniger Freiheitsgrade wie das Modell mit Dummy-Variablen verbraucht, bessere Werte aufweisen, wenn sich die *Sum of Squares* nur geringfügig unterscheiden. Ob die höheren *Sum of Squares* oder die niedrigere Anzahl an Freiheitsgraden für das metrische Modell mehr ins Gewicht fallen, kann mit dem Dashboard ermittelt werden.

Eine Extrapolation des Modells ist auch mit der Umkodierung nicht möglich. Diese würde, ob mit oder ohne Umkodierung, von der Kodierung der neu dazukommenden Merkmalsausprägung abhängen, für die wiederum ein bedingter Mittelwert der endogenen Variablen benötigt würde. Hierfür kann das in *Kapitel 4* beschriebene Verfahren somit keine Abhilfe schaffen. Eine Interpretation des Achsenabschnitts ist zudem nicht möglich, da dieser auch bei der Umkodierung von der ursprünglichen Kodierung des ordinalskalierten Merkmals abhängig ist. Die Vorteile der Umkodierung liegen darin, dass nur ein Parameter pro Merkmal verwendet werden muss und man damit dennoch die besseren Schätzwerte der Dummy-Variablen erzielt. Für diesen einen Parameter lässt sich zudem das Signifikanzniveau des ordinalskalierten Merkmals untersuchen, was bei Dummy-Variablen in der Form nicht möglich ist, da jede Merkmalsausprägung ein separates Signifikanzniveau aufweist.

[215] Vgl. Backhaus, et al. (2021), S. 74.
[216] Vgl. Berekhoven, Eckert, Ellenrieder (2009), S. 80.

In *Kapitel 4.1.4* wurde aufgezeigt, dass jede Kodierung für das ordinalskalierte Merkmal zu den Schätzergebnissen der Dummy-Variablen führen, wenn die bedingten Mittelwerte eine Gerade abbilden. Dabei ist es nicht relevant, welche Parameter diese Gerade aufweist, weswegen unendlich viele ideale Kodierungen existieren. Darüber hinaus wurde in *Kapitel 4.1.5* der Nutzen der Umkodierung bei einem nichtlinearen Zusammenhang zwischen exogener und endogener Variablen aufgezeigt. Hierbei kann aus jedem mindestens kausalen Zusammenhang ein lineares Modell gebildet werden. Das umkodierte lineare Regressionsmodell wurde in *Kapitel 4.1.6* validiert und somit auf dessen Gültigkeit überprüft. Des Weiteren lassen sich durch die Umkodierung Rückschlüsse auf die Abstände bzw. Verhältnisse der Merkmalsausprägung des ordinalskalierten Merkmals ziehen, wie es in *Kapitel 4.1.7* aufgezeigt wurde. Zudem konnte in *Kapitel 4.2* mit dem *Ordinal-Metric-Transformation-Coefficient* eine Fehlerkennzahl hergeleitet werden, die die Eignung eines ordinalskalierten Merkmals für eine Transformation auf das metrische Skalenniveau abbildet. Kritisch sind die teilweise enttäuschenden Ergebnisse der Umkodierung bei einer multiplen Regression aus *Kapitel 4.3* zu nennen, die die ursprüngliche Annahme, dass die bedingten Mittelwerte aller Merkmalsausprägungen eine Gerade abbilden sollten, widerlegt. Stattdessen kann hierbei die neue Hypothese aufgestellt werden, dass die ideale Kodierung eines ordinalskalierten Merkmals die Korrelation zwischen exogener und endogener Variablen maximiert.

5.2 Ausblick

In dieser Arbeit wurde das Augenmerk nur auf einen kleinen Ausschnitt der weiten Welt der multivariaten Analyseverfahren gelegt, indem sich hauptsächlich mit der linearen Einfachregression und der Verwendung ordinalskalierter Merkmale in dieser beschäftigt wurde. Weitergehend können ähnliche Untersuchungen zu der Umkodierung ordinalskalierter Merkmale in Regressionsverfahren durchgeführt werden, die hier nicht behandelt wurden. Hierzu gehören beispielsweise die logistische Regression oder nicht-lineare Regressionsmodelle. Auf welche Weise die Umkodierung bei einer multiplen Regression durchgeführt werden muss, damit diese zu einem besseren Modell führt, bleibt der Inhalt weiterer Forschungen. Hierbei ist die in *Kapitel 4.3* thematisierte Vermutung der Maximierung der Korrelation zwischen exogener und

endogener Variablen zu nennen, mit der das Bestimmtheitsmaß und somit die Anpassung des Modells optimiert werden kann. Die Umkodierung mit den Residuen statt den Beobachtungswerten könnte hierfür ein geeigneter Ansatz sein. Darüber hinaus können ähnliche Untersuchungen, insbesondere die Relevanz der Kodierung von ordinalskalierten Variablen, für anderweitige multivariate Analyseverfahren durchgeführt werden. Von besonderem Interesse ist es hierbei, die Kodierung eines ordinalskalierten Merkmals nicht als fixe Größe, sondern als Zufallsvariable zu sehen, die an die Beobachtungsdaten angeglichen werden sollte, damit sie keinen subjektiven Einflüssen des Analysten unterliegt.

Anhang

Der elektronischen Fassung dieser Arbeit auf dem beigelegten USB-Stick sind folgende Dokumente beigefügt:

- *R-Skript, Master-Thesis, Markus Köhnlein, Handhabung ordinalskalierter exogener Variablen in linearen Regressionsmodellen.R*

Zusätzlich zu dem dieser Arbeit beigelegten *R*-Skript wurde ein eigenes Paket mit vier Funktionen programmiert und über *GitHub* publiziert.[217] Um die Funktionen des Paketes namens *OMTC* zu nutzen, ist ein Beispielskript in dem beigelegten *R*-Skript zu finden.[218] Zu Beginn müssen die Pakete *devtools* und *reshape* installiert und ausgeführt werden, genauso wie das hier betrachtete Paket *OMTC*. Es wird der Datensatz *mtcars* verwendet, um einen Basisdatensatz zu erstellen. Die Werte für den Fahrzeugverbrauch in mpg werden hierbei als metrische endogene Variable verwendet. Die Zylinderanzahl cyl wird als ordinalskalierte exogene Variable angenommen.[219] Mit dem Paket *OMTC* lassen sich folgende vier Funktionen durchführen:

- Für den Datensatz kann mit der Funktion *Coding()* die Neukodierung des Datensatzes berechnet werden. Hierbei wird eine neue Variable x_nc erstellt.[220]
- Mit der Funktion *Dashboard()* lässt sich das aus *Kapitel 4.1.3* bekannte Dashboard der einzelnen Fehlerkennzahlen ausgeben.[221]
- Mithilfe der Funktion *Grafic()* lässt sich die zuvor durchgeführte Umkodierung wie in *Abbildung 9* darstellen.[222]
- Des Weiteren kann mit der Funktion *Variance_Analysis()* ein weiteres Dashboard ausgegeben werden, um die Bestandteile der Streuungserklärung, sowie den *OMTC* wie in *Tabelle 16* darzustellen.[223]

---

[217] Zu finden unter: https://github.com/MarkusKoehnlein/OMTC

[218] Siehe: *R*-Skript, Zeile: 1-30. Soll das *R*-Paket für andere Datensätze verwendet werden, ist sicherzustellen, dass der Datensatz als „Daten“ und die Variablen als „x“ und „y“ bezeichnet werden.

[219] In dem dargelegten Zahlenbeispiel ist es besonders interessant zu sehen, dass das ordinalskalierte Merkmal so gut als metrische Variable geeignet ist, dass das metrische Modell für manche Fehlerkennzahlen die besseren Werte ausgibt. Diese herausragende Eignung wird auch durch den *OMTC* dargestellt.

[220] Siehe: *R*-Skript, Zeile: 2651-2668.

[221] Siehe: *R*-Skript, Zeile: 2670-2838.

[222] Siehe: *R*-Skript, Zeile: 2840-2890.

[223] Siehe: *R*-Skript, Zeile: 2892-2938.

Literaturverzeichnis

Backhaus, Klaus / Erichson, Bernd / Weiber, Rolf (2015), *Fortgeschrittene Multivariate Analysemethoden, Eine anwendungsorientierte Einführung*, 3. Auflage, Wiesbaden, Springer Gabler.

Backhaus, Klaus / Erichson, Bernd / Gensler, Sonja / Weiber, Rolf / Weiber, Thomas (2021), *Multivariate Analysemethoden, Eine anwendungsorientierte Einführung*, 16. Auflage, Wiesbaden, Springer Gabler.

Berekhoven, Ludwig / Eckert, Werner / Ellenrieder, Peter (2009), *Marktforschung, Methodische Grundlagen und praktische Anwendungen*, 12. Auflage, Wiesbaden, Springer Gabler.

Canty, Angelo (2021): *Package ‚boot'*, Functions and datasets for bootstrapping from the book ``Bootstrap Methods and Their Application'' by A. C. Davison and D. V. Hinkley (1997, CUP), originally written by Angelo Canty, Version 1.3-28, https://cran.r-project.org/web/packages/boot/boot.pdf.

Fox, John, et al. (2022): *Package ‚car'*, Companion to Applied Regression, Version 3.0-13, https://cran.r-project.org/web/packages/car/car.pdf.

Hedderich, Jürgen / Sachs, Lothar (2020): *Angewandte Statistik, Methodensammlung mit R*, 17. Auflage, Berlin, Springer Spektrum.

Hirschle, Jochen (2021): *Machine Learning für Zeitreihen, Einstieg in Regressions-, ARIMA- und Deep-Learning-Verfahren mit Python*, 1. Auflage, Carl Hanser Verlag München.

Horikosh, Masaaki (2022), *Package 'ggfortify'*, Data Visualization Tools for Statistical Analysis Results, Version 0.4.14, https://cran.r-project.org/web/packages/ggfortify/ggfortify.pdf.

Janssen, Jürgen / Laatz, Wilfried (2017): *Statistische Datenanalyse mit SPSS, Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests*, 9. Auflage, Wiesbaden, Springer Gabler.

Kühne, Annegret / Wenger, Wolf (2011): *Mengenprognose mit dem Holt-Winters-Verfahren am Beispiel des monatlichen Energiebedarfs von Industrieunternehmen*, in: Wolf Wenger, Martin Josef Geiger, Andreas Kleine (Hrsg.), Business Excellence in Produktion und Logistik, Wiesbaden, Springer Gabler.

Kronthaler, Franz / Zöllner, Silke (2021): *Data Analysis with RStudio, An Easygoing Interduction*, 1. Auflage, Berlin, Springer Spektrum.

Meffert, Heribert / Burmann, Christoph / Kirchgeorg, Manfred (2015): *Marketing, Grundlagen marktorientierter Unternehmensführung, Konzepte – Instrumente – Praxisbeispiele*, 12. Auflage, Wiesbaden, Springer Gabler.

Schlittgen, Rainer / Sattarhoff, Cristina (2020): *Angewandte Zeitreihenanalyse mit R*, 4. Auflage, Walter de Gruyter GmbH, Berlin/Boston.

Wickham, Hadley, et al. (2016): *Package 'ggplot2',* Create Elegant Data Visualisations Using the Grammar of Graphics, Version 3.3.6, https://cloud.r-project.org/web/packages/ggplot2/ggplot2.pdf.

Wickham, Hadley, et al. (2021): *Package 'devtools', Tools to Make Developing R Packages Easier*, Version 2.4.3, https://cran.r-project.org/web/packages/devtools/devtools.pdf.

Wickham, Hadley (2022): *Package 'reshape'*, Flexibly Reshape Data, Version 0.8.9, https://cran.r-project.org/web/packages/reshape/reshape.pdf.

Wollschläger, Daniel (2020): *Grundlagen der Datenanalyse mit R – Eine anwendungsorientierte Einführung*, 5. Auflage, Heidelberg, Springer.

Xie, Yihui, et al. (2021): *Package 'mime'*, Map Filenames to MIME Types, Version 0.12, https://cran.r-project.org/web/packages/mime/mime.pdf.

Xie, Yihui, et al. (2022): *Package 'rmarkdown'*, Dynamic Documents for R, Version 2.14, https://cran.r-project.org/web/packages/rmarkdown/rmarkdown.pdf.

Xie, Yihui, et al. (2022): *Package 'tinytex'*, Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents, Version 7.1.2, https://cran.r-project.org/web/packages/tinytex/tinytex.pdf.

Ehrenwörtliche Versicherung

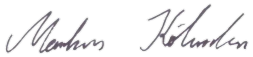
Hiermit versichere ich, Markus Köhnlein, Matrikel-Nr. 200067, ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe angefertigt und mich anderer als der in den beigefügten Verzeichnissen angegebenen Hilfsmittel nicht bedient habe.

Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Alle Quellen, die dem World Wide Web entnommen oder in einer sonstigen digitalen Form verwendet wurden, sind der Arbeit beigelegt.

Der Durchführung einer elektronischen Plagiatsprüfung stimme ich hiermit zu. Die eingereichte elektronische Fassung der Arbeit entspricht der eingereichten schriftlichen Fassung exakt.

Ich bin mir bewusst, dass eine unwahre Erklärung rechtliche Folgen haben wird.

Schwäbisch Hall, 11.08.2022
(Ort, Datum)


(Unterschrift: Markus Köhnlein)