Multiple lineare Regression: Ein vertiefter Einblick in die statistische Modellierung

Mariana Nold[1], und Florian Meinfelder[2]

[1]Institut für Soziologie, Universität Jena
[2]Institut für Statistik, Universität Bamberg

1. Oktober 2025

Inhaltsverzeichnis

1 Ziele des Readers

In der praktischen Anwendung zeigt sich häufig, dass zentrale Grundlagen der Statistik missverstanden oder verkürzt interpretiert werdenVan Calster u. a. (2021); Gelman und Loken (2014) – ein Umstand, der nicht zuletzt auf historisch gewachsene Routinen Kennedy-Shaffer (2019) und falsche Lehrtraditionen Carlin und Moreno-Betancur (2025) zurückzuführen ist. Ziel dieses Readers ist es, ausgewählte Aspekte der Regressionsanalyse kritisch zu beleuchten und ein besseres Verständnis für die zugrundeliegenden Annahmen, Grenzen und Fallstricke zu fördern. Dabei geht es weniger um eine technische Einführung, sondern um die Reflexion typischer Fehlinterpretationen – etwa im Umgang mit p-Werten oder Modellannahmen – sowie um die Frage, wie statistische Modelle sinnvoll eingesetzt werden können.

Somit richtet sich dieser Reader vor allem an Promovierende, fortgeschrittene Studierende und Anwender*innen, die keine formale Ausbildung in statistischer Inferenztheorie absolviert haben, aber Regressionsmodelle zur Analyse empirischer Daten verwenden. Er soll helfen, ein kritisches Bewusstsein für die Modellwahl zu entwickeln, die Anwendungspraxis zu hinterfragen und die Bedeutung der Reproduzierbarkeit und Transparenz von Entscheidungen im Prozess der statistischen Modellierung hervorzuheben. [1]

Im Abschnitt 2 wird zunächst die multiple lineare Regression anhand eines Beispiels rekapituliert. Abschnitt 3 vertieft zentrale Begriffe der statistischen Inferenz. Abschließend wird in Abschnitt 4 eine Simulation eingesetzt, um das Verständnis für häufig diskutierte Probleme im Umgang mit p-Werten zu schärfen. [2]

[1] Für Leser*innen, die noch nicht mit linearer Regression vertraut sind, gibt es eine Reihe von guten Statistiklehrbüchern, z. B. Scheid und Vogl (2021), Fahrmeir u. a. (2016) oder das Online-`R`-Buch Applied Statistics with `R` (Dalpiaz, 2021).

[2] Der im Reader verwendete `R`-Code ist auf der zugehörigen Webseite unter nold.info/RCodeReader verfügbar, ergänzende Informationen finden sich im Anhang.

2 Multiple Regression: Ein Überblick

Dieser Abschnitt beginnt mit einem fiktiven Beispiel das genutzt wird, um statistische Konzepte zu erläutern. In dem Beispiel geht es um den so genannten Gender Pay Gap (GPG), der die Einkommenslücke zwischen Frauen und Männern beschreibt. In der (ebenfalls fiktiven) Stichprobe werden nur in Vollzeit arbeitende Frauen und Männer berücksichtigt. Die Outcome-Variable des Regressionsmodells ist das monatliche Bruttoeinkommen. Ziel der Analyse ist die Darstellung des multidimensionalen Zusammenhangs zwischen monatlichem Bruttoeinkommen und der Verteilung anderer Variablen. Diese Variablen werden oft als Einflussfaktoren bezeichnet, wobei diese Bezeichnung irreführende Assoziationen bezüglich kausaler Zusammenhänge wecken kann. Berücksichtigt werden in unserem Beispiel die Variablen *Alter, (höchster erworbener) Schulabschluss* sowie die *Branchenzugehörigkeit* des ausgeübten Berufes. Dabei gibt es in dem fiktiven Datensatz zwei Branchen: In der einen arbeiten mehr als 70% Frauen, in der anderen weniger als 30%. Die erste Branche wird als *weiblich* bezeichnet. Der Schulabschluss ist in drei Kategorien erhoben, diese sind *mittlerer Schulabschluss*, *Abitur* und *Hochschulabschluss*.

Die deskriptive Forschungsfrage lautet: *Wie unterscheidet sich das mittlere Einkommen in den durch die Variablen definierten Subpopulationen?*

Der Datensatz umfasst 400 Personen, wobei jede Person einen Schulabschluss aufweist und alle Angaben vollständig sind. Der mittlere Schulabschluss wird als Referenzkategorie festgelegt. Die Tabelle 1 gibt einen Überblick über die verwendeten Variablen [3].

Im Abschnitt 2.1 wird zunächst das Standardmodell der multiplen Regression zusammengefasst und im Abschnitt 2.2 werden dann die Stichprobenverteilungen der Parameterschätzer angegeben und inhaltlich interpretiert. Abschnitt 2.3 enthält einen Überblick über Konfidenzintervalle und Hypothesentests sowie Erläuterungen zu den jeweiligen Güteeigenschaften.

[3] Eine kurze Beschreibung des Datensatzes findet sich im beiliegenden `R`-Code, dort findet sich auch ein weiteres Beispiel, welches auf einem realen Beispiel beruht. Nähre Informationen dazu finden sich im Anhang A.

Variable	Label der Variable	Beschreibung
Y	*Einkommen*	Monatliches Bruttoeinkommen
X_1	*(Alter)*	Alter in Jahren
X_2	*(Frau)*	Geschlecht, binär, 1 = weibl.
X_3	*(Abschluss)*	Abitur binär, 1 = ja
X_4	*(Abschluss)*	Hochschulabschluss, binär, 1 = ja
X_5	*(Frau_Branche)*	Branchenzugehörigkeit, binär, 1 = weibl.

Tabelle 1: Überblick über die Variablen im fiktiven Datensatz

Der Abschnitt 2.5 führt die Devianz, ein in der Statistik wichtiges Konzept, ein und stellt darauf beruhende Möglichkeiten der Modellselektion vor. Eine hohe Devianz weist hierbei auf ein Modell hin, das den Daten nicht gut gerecht wird. Im Abschnitt 2.4 wird ein klassischer Hypothesentest im multiplen Regressionsmodell vorgestellt. Im Abschnitt 2.6 erfolgt schließlich eine kurze Ergebniszusammenfassung für das Anwendungsbeispiel.

2.1 Multiple Regression und Methode der kleinsten Quadrate

Das Standardmodell der multiplen Regression ist gegeben durch (vgl. Fahrmeir u. a., 2016, Kap. 12)

Standardmodell

$$Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \ldots + \beta_p \cdot x_{p,i} + \epsilon_i, \quad i = 1, \ldots, n. \quad (1)$$

Dabei sind:

- $Y_1, \ldots Y_n$ beobachtbare metrische Zufallsvariablen,
- $\underline{x}_1, \underline{x}_2, \underline{x}_3$ gegebene deterministische Werte oder Realisierungen metrischer Zufallsvektoren $\underline{x}_1, \underline{x}_2, \underline{x}_3$, der Länge n.
- $\epsilon_1, \ldots \epsilon_n$ unbeobachtete Zufallsvariablen, die unabhängig und identisch verteilt sind mit $\mathbb{E}(\epsilon_i) = 0$ und $\mathbb{V}(\epsilon_i) = \sigma^2$.

Die Regressionskoeffizienten $\beta_0, \beta_1, \beta_2, \beta_3$ und die Varianz σ^2 sind **feste Größen**, die aus den Daten $(x_i, y_i), i = 1, \ldots, n$ zu schätzen sind.

In der Darstellung des Modells wird ein Unterstrich verwendet um einen Vektor darzustellen, so ist

$$\underline{x}_1 := \begin{pmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,i} \\ \vdots \\ x_{1,n} \end{pmatrix} \quad (2)$$

der Vektor der die Ausprägungen des Merkmals X_1 für alle n Individuen der Stichprobe enthält. Entsprechend kann man auch die Outcome-Variable als Vektor darstellen

$$\underline{Y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}. \tag{3}$$

Das Standardmodell kann auch in Matrixschreibweise dargestellt werden. Hierzu wird zunächst die Design-Matrix $\underline{\underline{x}}$ definiert als

$$\begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{n,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{n,2} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{n,n} \end{pmatrix}. \tag{4}$$

Analog kann man auch den Parametervektor und den Vektor der Residuen darstellen als

$$\underline{\beta} := \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} \tag{5}$$

und

$$\underline{\epsilon} := \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{6}$$

Damit kann die Gleichung (1) in Matrixnotation dargestellt werden

$$\underline{Y} = \underline{\underline{x}} \cdot \underline{\beta} + \underline{\epsilon}. \tag{7}$$

Die Methode der kleinsten Quadrate ist das verbreitetste Verfahren zur Schätzung der Parameter im linearen Regressionsmodell und wird nachfolgend ebenfalls als Grundlage für die Schätzung der Regressionskoeffizienten vorgestellt. Die Streuungszerlegung in der Regression teilt die Gesamtstreuung in die erklärte Streuung und die Reststreuung auf.

Es gilt

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{8}$$

wobei $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ die Gesamtstreuung ist, $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ die durch das Modell erklärte Streuung und $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ die (nicht erklärte) Reststreuung. Die Reststreuung ist der Abstand zwischen den durch das Modell vorhergesagten Punkten und den beobachteten Punkten. Die Schätzung der Regressionskoeffizienten erfolgt nun, indem man den quadrierten Term der Reststreuung minimiert, was zur Bezeichnung "Kleinste-Quadrate-Methode" führte. Je niedriger diese Reststreuung ist, desto besser ist die Modellanpassung. Die Schätzung der Parameter besteht konkret darin, einen Parametervektor $\underline{\beta}$ zu finden, so dass

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \ldots - \beta_p x_{i,p}\right)^2 \tag{9}$$

minimal wird. Die Stelle, an der die Gleichung (9) ein Minimum annimmt, wird dann mit $\underline{\hat{\beta}}$ bezeichnet. Es lässt sich zeigen, dass man die in Gleichung (9) dargestellte Aufgabe analytisch lösen kann, d. h. man kann eine Formel als Lösung dieser Gleichung angeben. Diese Formel lautet

$$\hat{\beta} = (\underline{\underline{x}}'\,\underline{\underline{x}})^{-1}\underline{\underline{x}}'\,\underline{Y}. \tag{10}$$

(vgl. z. B. Dalpiaz, 2021, Kapitel 17 , Kap. 7.2). Es bleibt noch der Parameter σ zu schätzen, dieser wird definiert als

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \tag{11}$$

Es lässt sich zeigen, dass der so definierte Schätzer die Eigenschaft hat erwartungstreu zu sein (vgl. z. B. Dalpiaz, 2021, Kapitel 17 , Kap. 7.2 und 8). Das bedeutet, dass dieser Schätzer den wahren unbekannten Parameter σ weder unterschätzt noch überschätzt.

Zusammenfassend: In der "klassischen" statistischen Sichtweise (auch als frequentistische Inferenz bezeichnet) sind die Parameter $\underline{\beta}$ und σ^2 feste Werte, die aus den Daten geschätzt werden. Ziel ist es, mit Blick auf frequentistische Gütekriterien Schätzverfahren zu nutzen, die auf lange Sicht (in der *long-run-frequency*) gut funktionieren. Diese Güteeigenschaften stehen in engem Zusammenhang mit der Stichprobenverteilung und werden im nächsten Abschnitt exemplarisch dargestellt. Um die Schätzer in den Gleichungen (10) und (11) als gut zu bewerten, muss gezeigt werden, dass sie tatsächlich den Güteeigenschaften genügen. Um sich zu vergewissern, dass dies der Fall ist braucht man die Stichprobenverteilung der entsprechenden Schätzer. Dies soll anhand des Anwendungsbeispiels erklärt werden.

2.2 Stichprobenverteilung der Parameterschätzer

Das Standardmodell der linearen Einfachregression erhält man aus dem Modell der multiplen Regression (vgl. Gleichung (1)) für den Spezialfall $p = 1$. Es handelt sich also um das Modell, das nur einen Prädiktor berücksichtigt. Es wird zunächst das Alter als einzige Einflussgröße betrachtet. In diesem Abschnitt wird für die Einfachregression dargestellt, dass Güteeigenschaften von Schätzern und statistischen Hypothesentests direkt aus den Eigenschaften der Stichprobenverteilung folgen.

Als Stichprobenverteilung bezeichnet man die Verteilung einer bestimmten Test- bzw. Schätzstatistik. Es gibt im Rahmen der Regression eine Reihe von Test- bzw. Schätzstatistiken. Als Formeln für die Schätzer der Einfachregression ergeben sich aus Gleichung (10) die folgenden Formeln:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x - x_i) \cdot (y - y_i)}{\sum_{i=1}^{n}(y - y_i)^2} \tag{12}$$

und

$$\hat{\beta}_0 = \bar{y} - \beta_1 \cdot \bar{x} \tag{13}$$

Die beiden Schätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ sind Kleinste-Quadrate-Schätzer (KQ-Schätzer). Durch die Berechnung der KQ-Schätzer können die Residuen geschätzt werden. Die geschätzten Residuen sind

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i,\ i\ \in 1, \ldots n. \tag{14}$$

Zu beachten ist, dass die wahren Residuen, also die Abstände der beobachteten Daten von der **wahren** Regressionsgerade unbekannt sind. Wir können lediglich die Abstände der beobachteten Daten zur **geschätzten** Regressionsgerade berechnen. Aus den geschätzten Residuen ergibt sich nach Gleichung (11) für die Einfachregression

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2. \tag{15}$$

Die in den Gleichungen (12), (13) und (15) gegebenen Schätzer haben stochastische Eigenschaften. Deren Herleitung finden sich in vielen Statistikbüchern z. B. in (Scheid und Vogl, 2021, Kapitel 12.1.2). Aus diesen Eigenschaften können der Erwartungswert und die Varianz der KQ-Schätzer berechnet werden.

Erwartungswerte und Varianzen der KQ-Schätzer

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \quad \mathbb{E}(\hat{\beta}_1) = \beta_1, \quad \mathbb{E}(\hat{\sigma}) = \sigma \tag{16}$$

$$\mathbb{V}(\hat{\beta}_0) = \sigma^2_{\hat{\beta}_0} = \sigma^2 \cdot \frac{\sum_{i=1}^{n} x_i^2}{n \cdot (x_i - \bar{x})^2} \tag{17}$$

$$\mathbb{V}(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \sigma^2 \cdot \frac{1}{n \cdot (x_i - \bar{x})^2} \tag{18}$$

Es lassen sich auch mit den Methoden der mathematischen Statistik die Verteilungen der Schätzer angeben. Die t-Verteilung mit k Freiheitsgraden wird mit t_k bezeichnet. Es gilt[4]:

Verteilungen der standardisierten KQ-Schätzer (Einfachregression)

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t_{n-2} \tag{19}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2} \tag{20}$$

Des weiteren ergibt sich für den Schätzer $\hat{\sigma}^2$ eine χ^2-Verteilung mit $n-2$ Freiheitsgraden, genauer:

Verteilungen des den Schätzers $\hat{\sigma}^2$ (Einfachregression)

$$\frac{(n-2) \cdot \hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \tag{21}$$

Doch, was bedeuten diese Gleichungen in einfachen Worten? Zunächst machen die Gleichungen (16) bis (21) offensichtlich, dass hinter einem Regressionsoutput, den ein statistisches Programmpaket ausgibt, eine ganze Reihe von Grundlagen der Mathematischen Statistik stehen. Wie kann man diese Ergebnisse allgemeinverständlich ausdrücken?

Die Gleichung (16) informiert darüber, dass die Schätzer erwartungstreu sind. Wenn man unter den selben Bedingungen wiederholt eine Zufallsstichprobe vom Umfang n zieht und für jede Stichprobe die drei Parameterschätzer berechnet, dann bekommt man für jede Stichprobe einen anderen Wert. Wenn man z. B. 100 solche Stichproben zieht und 100 Mal den Schätzer berechnet, dann hat jede dieser Stichproben ihren eigenen Schätzer

[4]Man beachte, dass für σ die Schätzung $\hat{\sigma}$ eingesetzt wird in die Gleichungen (17) und (18) um einen Schätzer für die Standardabweichung der Schätzer zu erhalten (vgl. z. B. Scheid und Vogl, 2021, Kap. 12).

$\hat{\beta}_0^{(k)}$, $\hat{\beta}_1^{(k)}$, und $\hat{\sigma}^{(k)}$ $k \in 1, \ldots, 100$. Der Schätzer hat also selbst eine Verteilung. Diese Verteilung nennt man die Stichprobenverteilung des Schätzers. Die Erwartungstreue sagt nun, dass der Erwartungswert der Stichprobenverteilung eines bestimmten Schätzers dem zu schätzenden wahren Parameter entspricht. Die Stichprobenverteilungen der KQ-Schätzer sind in den Gleichungen (19) und (20) gegeben.

Beruhend auf diesen Stichprobenverteilungen lassen sich auch statistische Tests und Konfidenzintervalle herleiten, diese haben Güteeigenschaften. Konkret hat das Verfahren zur Berechnung eines Konfidenzintervalls die Güteeigenschaft, dass mit vorgegebener Wahrscheinlichkeit ein Intervall berechnet wird, welches den wahren Parameter enthält. Der statistische Test hat die Güteeigenschaft, dass die Nullhypothese mit einer vorgegebenen Irrtumswahrscheinlichkeit fälschlicherweise abgelehnt wird. Gerade das ist die Leistung des Tests. Wenn man eine Entscheidung trifft, dann kann man immer einen Fehler machen. Die Irrtumswahrscheinlichkeit zu kennen, erlaubt es, die Unsicherheit der Testentscheidung zu quantifizieren.

Man kann die Verteilung der Parameterschätzer natürlich nicht nur für die Einfachregression herleiten, sondern allgemein für p Prädiktoren und es ergibt sich

Verteilungen der standardisierten KQ-Schätzer (multiple Regression)

$$T_j := \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-p-1} \tag{22}$$

$j \in \{1, \ldots, p\}$

Für die Varianz der Residuen ergibt sich der Schätzer

Verteilungen des den Schätzers $\hat{\sigma}^2$ (multiple Regression)

$$\frac{(n-p-1) \cdot \hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{23}$$

2.3 Konfidenzintervalle und statistische Hypothesentests

In diesem Abschnitt werden zunächst für die Einfachregression und dann für die multiple Regression Konfidenzintervalle und Hypothesentests vorgestellt. Aus den Stichprobenverteilungen der KQ-Schätzer, gegeben in den Gleichungen (19) und (20), lassen sich durch einfache arithmetische Umformungen Konfidenzintervalle und statistische Hypothesentests herleiten. Aus der Stichprobenverteilung ergibt sich unmittelbar

$$\mathbb{P}\left(\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1}\right) = 1 - \alpha \tag{24}$$

und damit das Konfidenzintervall für β_1

$$\left(\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1}\right). \tag{25}$$

Für den Achsenabschnitt β_0 ergibt sich entsprechend

$$\left(\hat{\beta}_0 - t_{n-2,1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{n-2,1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_0}\right). \tag{26}$$

Analog zur Einfachregression können Konfidenzintervalle auch für die multiple Regression berechnet werden, diese sind in Gleichung (27) gegeben.

Beruhend auf der Stichprobenverteilung lassen sich auch statistische Hypothesentests konstruieren. Diese stellen formale Entscheidungsregeln dar, die in einer Reihe von Schritten ablaufen (vgl. Fahrmeir u. a. (2016, Kapitel 10.2 , S. 381 ff): Im ersten Schritt eines Hypothesentests drückt man die quantifizierte inhaltliche Hypothese in Form einer statistischen Hypothese aus. Die Modellannahmen werden benannt, das Signifikanzniveau wird festgelegt und mit Bezug auf diese Irrtumswahrscheinlichkeit wird schließlich der Ablehnbereich angegeben. Im nächsten Schritt wird der Wert der Prüfgröße berechnet. Im letzten Schritt wird die Testentscheidung getroffen, indem man abgleicht, ob dieser Wert im Ablehnbereich liegt. Ist das der Fall, so wird die Nullhypothese verworfen. Es lassen sich drei Hypothesen formulieren:

Hypothesen für Regressionskoeffizienten (multiple Regression)

(a) $H_0 : \beta_j = \beta_{0,j} \quad \leftrightarrow \quad H_1 : \beta_j \neq \beta_{0,j}$

(b) $H_0 : \beta_j \geq \beta_{0,j} \quad \leftrightarrow \quad H_1 : \beta_j < \beta_{0,j}$

(c) $H_0 : \beta_j \leq \beta_{0,j} \quad \leftrightarrow \quad H_1 : \beta_j > \beta_{0,j}$

In Gleichung (22) wurde die Verteilung der standardisierten Schätzer definiert. Sie lautet

$$T_j := \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-p-1}.$$

Beruhend auf dieser Verteilung lassen sich Ablehnbereiche für die Schätzer konstruieren. Die Entscheidungsregel lautet dann:

Entscheidungsregel im Hypothesentest (multiple Regression)

(a) $H_0 : \beta_j = \beta_{0,j} \quad \leftrightarrow \quad H_1 : \beta_j \neq \beta_{0,j}$, H_0 ablehnen, falls $|T_j| > t_{1-\frac{\alpha}{2},n-p-1}$

(b) $H_0 : \beta_j \geq \beta_{0,j} \quad \leftrightarrow \quad H_1 : \beta_j < \beta_{0,j}$ H_0 ablehnen, falls $T_j < -t_{1-\frac{\alpha}{2},n-p-1}$

(c) $H_0 : \beta_j \leq \beta_{0,j} \quad \leftrightarrow \quad H_1 : \beta_j > \beta_{0,j}$ H_0 ablehnen, falls $T_j > t_{1-\frac{\alpha}{2},n-p-1}$

Für die multiple Regression kann analog wie für die Einfachregression ein Konfidenzintervall für den Regressionskoeffizienten β_j $j \in \{1, \ldots, p\}$ hergeleitet werden:

Konfidenzintervall für Regressionskoeffizienten β_j $j \in \{1, \ldots, p\}$ (multiple Regression)

$$\left(\hat{\beta}_j - t_{n-p-1,1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-p-1,1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_j}\right). \tag{27}$$

Wie das statistische Konfidenzintervall ist auch der statistische Hypothesentest ein Verfahren, das über lange Sicht verlässlich ist. Sofern die Modellannahmen erfüllt sind, ist abgesichert, dass das Signifikanzniveau die maximale Wahrscheinlichkeit ist, mit der die Nullhypothese verworfen wird, obwohl sie tatsächlich richtig ist. Sowohl beim Konfidenzintervall als auch beim statistischen Hypothesentest stellt man sich vor, dass wiederholt unter den selben Bedingungen eine Stichprobe gezogen wird. Was das konkret bedeutet, wird am Anwendungsbeispiel im Abschnitt 2.4 demonstriert.

2.4 Ein statistischer Hypothesentest im Regressionsmodell

In diesem Abschnitt geht es darum, eine statistische Hypothese mit Hilfe eines Regressionsmodells zu testen. Die Vorgehensweise bei der Konstruktion des Tests folgt den Schritten in Fahrmeir u. a. (2016, Kapitel 10.2 , S. 381 ff).

2.4.1 Schrittweises Vorgehen

Der erste Schritt besteht darin, dass inhaltliche Problem zu quantifizieren. Konkret bezogen auf die gegebene Fragestellung ist die Quantifizierung wie folgt: Die inhaltliche Frage lautet, ob weibliche Personen ein geringeres Einkommen haben als männliche Personen. Mit anderen Worten: Wenn man das Einkommen für jeweils eine weibliche und eine männliche Person vorhersagt und sich die beiden Personen nicht im Alter unterscheiden, in der gleichen Branche arbeiten und den gleichen Schulabschluss haben, dann ist die quantifizierte inhaltliche Hypothese, dass für die weibliche Person ein niedrigeres mittleres Einkommen vorhergesagt wird als für die männliche Person.

Im zweite Schritt müssen die Modellannahmen genannt werden. Hier konkret: Das in H_0 spezifizierte Modell stellt den wahren datengenerierenden Prozess dar. Es liegt ein Standardmodell der multiplen linearen Regression vor.

$$M : \textit{(Einkommen)} \sim \textit{(Alter)} + \textit{(Frau)} + \textit{(Abschluss)} + \textit{(Frau\_Branche)}$$

Der dritte Schritt besteht darin, dass im ersten Schritt quantifizierte Testproblem als statistisches Testproblem zu formulieren. In diesem dritten Schritt werden Aussagen über die Parameter des statistischen Modells in der Nullhypothese H_0 und der Alternative H_1 formuliert. Hier soll die Nullhypothese $\beta_{(Frau)} \geq 0$ getestet werden, gegen die Alternative $\beta_{(Frau)} < 0$. [5]

Die maximal akzeptierte Irrtumswahrscheinlichkeit wird im vierten Schritt festgelegt. Sie wird als Signifikanzniveau α bezeichnet, das im Rahmen von Hypothesentests häufig mit $\alpha = 0.05$ spezifiziert wird.[6]

Im fünften Schritt wird der Ablehnbereich festgelegt. Fällt der realisierte Wert der Teststatistik in den Ablehnbereich, so wird die Nullhypothese abgelehnt. Schließlich wird der realisierte Wert der Teststatistik im sechsten Schritt berechnet. Im letzten siebten Schritt wird dann die formale Entscheidung getroffen, entweder muss H_0 beibehalten werden oder es kann H_1 nachgewiesen werden.

2.4.2 Testentscheidung

Die Teststatistik und deren Vereitelung ist in Gleichung (22) gegeben, sie lautet hier

$$\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}} \sim t_{394}.$$

[5] Zu beachten ist dabei, dass sich diese Hypothese auf den Parameter $\beta_{(Frau)}$ bezieht und sich direkt auf die Vorhersage des Einkommens bei weiblichen und männlichen Personen überträgt.

[6] Der historische Ursprung hierfür ist eine eher beifällige Bemerkung von Sir R.A. Fisher aus dem Jahr 1925.

Beruhend auf dieser Teststatistik und ihrer Vereitelung kann man nun einen statistischen Test für die einseitige Hypothese

$$H_0 : \beta_{(Frau)} \geq 0$$

versus

$$H_1 : \beta_{(Frau)} < 0$$

ableiten.

Der Wert der Teststatistik beträgt -19.96. Da der Ablehnbereich $(-\infty, -1.65)$ beträgt, gilt die Alternative (bei einer Irrtumswahrscheinlichkeit von 5%) als nachgewiesen.

Im Fazit kann man für die Testentscheidung zusammenfassen: Die Nullhypothese wird abgelehnt. Dieser Test stellt eine formale Entscheidungsregel dar, bei der man mit kontrollierter Irrtumswahrscheinlichkeit schließen kann.[7] Wenn man die Logik des frequentistischen Denkens verstehen möchte, ist es sehr wichtig, sich vor Augen zu halten, dass die Güteeigenschaften der frequentistischen Verfahren in Eigenschaften der Stichprobenverteilung verankert sind. Die Stichprobenverteilung wiederum beruht auf der Vorstellung, die Stichprobe unter den gleichen Bedingungen wieder und wieder zu ziehen und geht davon aus, dass das genutzte Modell das wahre Modell ist, das den unbekannten datengenerierenden Prozess abbildet. Der statistische Hypothesentest in diesem Abschnitt ist auf einen dezisionistischen Zweck ausgerichtet. Es geht also darum, eine Entscheidung zu treffen, die auf lange Sicht selten zu Fehlern führt. Auf lange Sicht kann man bei dieser Entscheidungsregel darauf vertrauen, dass man nur in 5% der Situationen die Nullhypothese ablehnt, obwohl tatsächlich diese Nullhypothese dem wahren Sachverhalt entspricht.

[7] Die Väter dieses Ansatzes haben es selbst so formuliert: *We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis. But we may look at the purpose of test from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.* (Jerzy Neyman & Egon Pearson, On the problem of the most efficient tests of statistical hypotheses, 1933)

Man kann die Testentscheidung auch treffen, indem man den p-Wert nutzt. Dieser p-Wert ergibt sich hier zu $4.86 \cdot 10^{-62}$. Da die Realisation der Zufallsvariable p-Wert hier unter 0.05 liegt und diese Zufallsvariable hier gleichverteilt ist, wird H_0 abgelehnt.

Bisher stand die inferentielle Unsicherheit im Fokus – also jene Unsicherheit, die sich aus dem Schluss von Stichprobe auf Population ergibt. Diese Form der Unsicherheit betrifft vor allem die statistische Absicherung von Ergebnissen und die Frage, wie zuverlässig sich beobachtete Effekte auf die Grundgesamtheit übertragen lassen. Doch Unsicherheit begegnet uns in der Statistik auf verschiedenen Ebenen. Neben der Unsicherheit des Inferenzschlusses spielen auch andere Formen eine zentrale Rolle: etwa die Unsicherheit der Messung, die Unsicherheit der Vorhersage und die Unsicherheit, die mit der Wahl des Modells verbunden ist. Letztere – die Modellunsicherheit – beschreibt laut Anderson (2019) all jene Entscheidungen, die die Modellbildung betreffen. Im vorliegenden Hypothesentest wurde diese Dimension nicht weiter berücksichtigt, d. h. das Modell wurde ohne kritische Reflexion übernommen. Ein zentraler Aspekt im Umgang mit Modellunsicherheit besteht darin, sie stets im Kontext der jeweiligen Fragestellung zu betrachten. Nur so lässt sich die Rolle des Modells innerhalb der Analyse klar bestimmen. Erst eine präzise formulierte Forschungsfrage ermöglicht es, Unsicherheiten gezielt zu erkennen und die Analyse entsprechend auszurichten (Carlin und Moreno-Betancur, 2025).

Im nächsten Abschnitt werden die *Devianz* und das *Akaike Information Criterion* (AIC) eingeführt, um eine Herangehensweise vorzustellen, die sich besser eignet, um der Modellunsicherheit gerecht zu werden. Test und Konfidenzintervalle lassen sich auch auf Grundlage der sogenannten Likelihood-Ratio-Test-Statistik formulieren, die in einer engen Beziehung zur Devianz steht. Die Modellierung, also Modellanpassung und Modellselektion beruhend auf der Devianz, erlaubt eine höhere Transparenz im Umgang mit Modellunsicherheit und ist daher besser geeignet als der in diesem Absatz beschriebene Ansatz. Die Devianz ist ein Maß das angibt, wie gut bzw. wie schlecht ein Modell zu den Daten passt. Je höher die Devianz, desto schlechter die Modellanpassung. Die Devianz stellt eine Verallgemeinerung der Summe

der Residuen dar und wird vertiefend in Abschnitt 3 behandelt.

2.5 Modellanpassung und Modellwahl: Devianz und AIC

Ein Grundverständnis der Devianz erlaubt einen tieferen Einblick in die Logik des statistischen Denkens. Die Devianz liegt der Parameterschätzung bei den sogenannten *Generalized Linear Models* (GLMs) zu Grunde. Die GLMs sind eine Verallgemeinerung der multiplen linearen Regression. Mit anderen Worten: Das multiple lineare Regressionsmodell ist ein spezielles GLM. Aus der Devianz lassen sich Kriterien der Modellanpassung und Modellwahl ableiten. In diesem Abschnitt werden Eigenschaften genannt, die sie zu einer grundlegend wichtigen Größe in der statistischen Modellierung machen. Die statistischen Tests und Konfidenzintervalle, die sich auf die Devianz beziehen, werden angegeben. Wenn der Parameter σ bekannt wäre, also nicht aus den Daten geschätzt werden müsste, dann wären die Test- und Konfidenzintervalle, die auf der Devianz beruhen, identisch zu den oben angegebenen. Wenn σ aus den Daten geschätzt wird, kommt es zu minimal unterschiedlichen Ergebnissen bei den Konfidenzintervallen und Tests. Bei großen Stichproben stimmen die Konfidenzintervalle überein (siehe z. B. Fahrmeir u. a., 2009, B 4. 3.). Die Devianz[8] hat einige Eigenschaften, die sie attraktiv machen für den Prozess der Modellselektion und damit der Modellierung.

Die Auseinandersetzung mit Likelihood und Devianz hat noch einen weiteren sehr zentralen Vorteil: Man gewinnt dadurch weitere Gütekriterien, die man der Analyse zu Grunde legen kann, nämlich "likelihoodistische" Gütekriterien. Diese erlauben eine andere Interpretation des Konfidenzintervalls. Wenn man rein frequentistische Gütekriterien anwendet, dann unterstellt man bei einem 95%-Konfidenzintervall ein Verfahren, dass a priori dazu führt, dass ein Intervall geschätzt wird, das den wahren Parameter mit einer Wahrscheinlichkeit von 95% enthält. Beruhend auf einem likelihoodis-

[8]Mit Devianz wird hier die Log-Likelihood auf dem *deviance scale* bezeichnet. Der *deviance scale* wird üblicherweise in Informationskriterien, wie dem AIC, verwendet da es Berechnungen vereinfacht. Siehe Gleichung (29) in Abschnitt 3.

tischen Gütekriterium kann man das Intervall anders interpretieren: Jedes Intervall, das bestimmte Werte ausschließt, sollte auch die weniger durch die Daten gestützten Werte ausschließen, das heißt, die Werte mit geringerer Likelihood, basierend auf der Likelihoodfunktion (Barnard, 1967). Dann handelt es sich um ein likelhoodistisches Gütekriterium, das sich nicht auf frequentistische Fehlerwahrscheinlichkeiten bezieht, sondern auf die Likelihoodfunktion und eine Interpretation der Likelihood als komparatives Plausibilitätsmaß. Die Likelihood-Ratio Test-Statistik, die in enger Beziehung zur Devianz steht, spielt eine zentrale Rolle bei likelhoodistischen Gütekriterien. Wenn man der eigenen Analyse diese Gütekriterien zu Grunde legt, vertritt man einen moderat frequentistischen Standpunkt (Lin, 2024). Die Verwendung von likelhoodistischen Gütekriterien erlaubt einen flexibleren Umgang mit der Modellierung und bietet daher mehr Möglichkeiten der Modellunsicherheit gerecht zu werden. Der moderat frequentistische Standpunkt wird den Anforderungen der Forschungsfrage oft besser gerecht als der rein frequentistische oder radikal frequentistische Ansatz, der ausschließlich auf den Gütekriterien der Stichprobenverteilung beruht.

2.5.1 Eigenschaften der Devianz

Für die eingangs formulierte deskriptive Forschungsfrage ist es das konkrete Ziel der statistischen Modellierung ein Modell zu finden, das zum einen nicht zu komplex ist und zum anderen zentrale Eigenschaften der Daten, also die Datenstruktur, gut erfassen. Im Allgemeinen wird ein komplexeres Modell den Daten immer besser gerecht, als ein weniger komplexes Modell. Mit anderen Worten: Komplexere Modelle können sich dem Datensatz besser anpassen, weil für die Modellanpassung mehr Parameter zu Verfügung stehen. Bei einer sehr komplexen Modellierung besteht unter anderem das Risiko, dass das Modell zunehmend an Interpretierbarkeit verliert, numerische Instabilitäten auftreten und die darauf basierenden Vorhersagen wenig verlässlich sind. Besonders kritisch wird es, wenn das Modell beginnt, sich zu stark an die zufälligen Schwankungen der Daten – also das Rauschen – anzupassen, statt das zugrunde liegende Signal zu erfassen. Dieses soge-

nannte Overfitting führt dazu, dass scheinbar präzise Ergebnisse entstehen, die jedoch inhaltlich irreführend sind und bei neuen Daten versagen. Daher versucht man oft ein möglichst einfaches Modell zu finden, dass zentrale Eigenschaften der Daten ausreichend erfasst. Man ist in einem Zielkonflikt und versucht dass Modell so sparsam wie möglich und so differenziert wie nötig zu spezifizieren. Diese Idee wird hier kurz durch die Auswahl eines Modells illustriert, eine mathematische Darstellung der Grundlagen erfolgt in Abschnitt 3. Die mathematische Grundlage für den Vergleich von Modellen beruhend auf der Devianz liefert der Satz von Wilks (siehe Gleichung (31)).

Die Devianz hat die folgenden vorteilhaften Eigenschaften (vgl. z. B. Gelman und Hill, 2007, S. 100):

1. Die Devianz ist ein Maß für die Abweichung zwischen Modell und Daten, eine geringere Devianz bedeutet eine bessere Anpassung des Modells an die Daten.

2. Wenn ein Prädiktor, der lediglich zufälliges Rauschen ist, zu einem Modell hinzugefügt wird, erwartet man, dass die Devianz im Durchschnitt um 1 abnimmt.

3. Wenn ein informativer Prädiktor zu einem Modell hinzugefügt wird, erwartet man, dass die Devianz um mehr als 1 sinkt. Wenn k informative Prädiktoren zu einem Modell hinzugefügt werden, erwartet man dass die Abweichung um mehr als k abnimmt.

Die Beurteilung der Modellanpassung beruhend auf der Devianz soll anhand des eingangs beschriebenen Beispiels zum GPG vorgestellt werden: Im Anwendungsbeispiel ist ein erstes sinnvolles Modell gegeben durch

$$M_0 : \textit{(Einkommen)} \sim \textit{(Alter)} + \textit{(Frau)} + \textit{(Abschluss)} + \textit{(Frau\_Branche)}$$

Für dieses Modell berechnet sich die Devianz zu $D_{M_0} = 4526.809$[9]. Diese Zahl alleine ist nicht interpretierbar, erst wenn man die Devianz für ein weiteres

[9]Die entsprechende mathematische Formel ist im Abschnitt 3 zu finden. Im beigefügten `R`-Code findet man die Umsetzung in `R`.

Modell berechnet, kann man durch einen Vergleich der beiden Devianzen die Modellanpassung beurteilen. Wenn dem Modell ein Prädiktor hinzugefügt wird, der rein zufällig ist, dann erwartet man, dass die Devianz um 1 zurück geht. Das bedeutet: Wenn man eine Variable simuliert und diese Variable anschließend in den linearen Prädiktor aufnimmt, erwartet man eine Abnahme der Devianz um einen Punkt. Die Devianz hat also die Eigenschaft, dass ihr Erwartungswert immer zurückgeht, wenn man einen Prädiktor hinzufügt. Ist der Prädiktor tatsächlich im wahren Modell enthalten, dann lässt sich mit Sätzen der mathematischen Statistik zeigen, dass die Differenz der Devianzen M_1 und M_0 eine χ^2-Verteilung mit einem Freiheitsgrad folgt (siehe z. B. Fahrmeir u. a., 2009, Kap. 4.1).[10]

Wir können ein zweites Modell in Erwägung ziehen und die entsprechenden Devianzen vergleichen: Eine Möglichkeit ist es, dass Frauen die in der weiblichen Branche arbeiten ein systematisch anderes mittleres Einkommen haben als Frauen, die in der nicht-weiblichen Branche tätig sind. Das entsprechende Modell ist

$$M_1 : \textit{(Einkommen)} \sim \textit{(Alter)} + \textit{(Frau)} + \textit{(Abschluss)} + \textit{(Frau\_Branche)} + \textit{(Frau)}{:}\textit{(Frau\_Branche)}$$

Die Devianz ergibt sich zu $D_{M_1} = 4526.512$, der Unterschied zu $D_{M_1} - D_{M_0} = 0.297$, so dass M_1 eine kaum bessere Modellanpassung bietet als M_0. Im nächsten Schritt kann man zu M_0 z. .B einen quadratischen Term für das Alter ergänzen.

$$M_2 : \textit{(Einkommen)} \sim \textit{(Alter)} + \textit{(Alter)}^2 + \textit{(Frau)} + \textit{(Abschluss)} + \textit{(Frau\_Branche)}$$

Die Devianz ergibt sich zu $D_{M_2} = 4444.715$, diese Devianz ist deutlich niedriger als die von M_0. Die Differenz der Devianzen ist $D_{M_2} - D_{M_0} = 4526.809 - 4444.715 = 82.094$, und somit deutlich größer als 1.

[10] Die χ^2-Verteilung spielt in der Statistik eine sehr zentrale Rolle, sie wird z. B. in (Fahrmeir u. a., 2016, Kap. 6.3) vorgestellt.

Eine denkbare Modellerweiterung ist nun, dass der Schulabschluss sich auf das Einkommen von Frauen anders auswirkt als auf das Einkommen von Männern. Das entsprechende Modell ist

$$M_3 : \textit{(Einkommen)} \sim \textit{(Alter)} + \textit{(Alter)}^2 + \textit{(Frau)} + \textit{(Abschluss)} + \textit{(Frau\_Branche)} + \textit{(Frau)}{:}\textit{(Abschluss)}.$$

Die Devianz des Modells M_3 ist $D_{M_3} = 4443.553$, diese Devianz ist im Vergleich zu der Devianz von M_2 nahezu unverändert. Daher ist das Modell M_2 bisher das Modell mit der besten Modellanpassung.

2.5.2 Der AIC als Grundlage der Modellselektion

In Abschnitt 2.5.1 wurde bereits darauf eingegangen, dass man in der statistischen Modellierung in vielen Situation versucht ein Modell zu finden, dass die Parameter präzise schätzt, also mit niedriger Varianz und auch unverzerrt. Hier ergibt sich ein Zielkonflikt: komplexere Modelle reduzieren die Verzerrung und erhöhen gleichzeitig die Varianz. Um ein Modell auszuwählen, dass hier einen guten Kompromiss darstellt, gibt es in der Statistik unterschiedliche Herangehensweisen. Eine bekannte Größe in diesem Kontext ist der sogenannte AIC, also das Akaike Information Criterion.

Man erhält den AIC, indem zur Devianz eines Modells die doppelte Anzahl der Parameter des Modells addiert. Die Definition lautet

$$AIC_M := 2 \cdot p_M + D_M \tag{28}$$

wobei p_M die Anzahl der Parameter im Modell M bezeichnet. Eine hohe Devianz ist Ausdruck einer schlechten Modellanpassung, je mehr Parameter ins Modell aufgenommen werden, desto besser wird die Modellanpassung. Wenn das Ziel der Modellierung darin besteht, ein möglichst sparsames Modell mit guter Modellanpassung zu finden, dann ergibt es Sinn die Hinzunahme von Parametern anzurechnen. Das geschieht beim AIC indem die doppelte Anzahl der Parameter des Modells addiert wird. Wenn man die Modellwahl auf den AIC gründet, so wählt man das Modell mit dem niedrigsten AIC. Der

Modell	Devianz	AIC
M_0	4526.809	4540.809
M_1	4526.512	4542.512
M_2	4444.715	4460.715
M_3	4443.553	4463.553

Tabelle 2: Überblick über die vier Modelle und die jeweilige Devianz und AIC

AIC ist eine Schätzung für die Devianz, die zu erwarten wäre, wenn das angepasste Modell zur Vorhersage neuer Daten verwendet würde (vgl. Gelman u. a., 2020, S. 175).

Modell 2 ist das Modell mit dem niedrigsten AIC-Wert und wird daher als optimales Modell unter Berücksichtigung der Komplexität des Modells gesehen. Es wäre falsch daraus zu schließen, dass das Modell 2 das wahre Modell ist oder auch nur, dass man annehmen kann, dass der Zusammenhang zwischen Alter und Einkommen quadratisch ist.

2.6 Regressionsergebnisse im Anwendungsbeispiel

Im obigen Abschnitt wird bei dem schrittweisen Vorgehen vorausgesetzt, dass das Modell bekannt ist und sich die Unsicherheit auf einen bestimmten Modellparameter bezieht. Diese Situation tritt in der realen Forschungspraxis kaum auf. Die Modellwahl ist mit Unsicherheit verbunden – doch diese lässt sich nur sinnvoll adressieren, wenn die Funktion des Modells zuvor geklärt ist. Ohne eine klare Zielsetzung kann keine statistische Theorie angemessen angewendet werden. Im vorliegenden Fall besteht die Funktion darin, ein Modell zu identifizieren, das geeignet ist, zentrale Zusammenhänge in den Daten zu erfassen und zu beschreiben. Die folgenden Ergebnisse werden daher mit diesem Ziel dargestellt.

2.6.1 Modellanpassung und Modellwahl

Es gibt unterschiedliche Möglichkeiten ein geeignetes deskriptives Modell zu begründen. Der AIC ist ein bekanntes Maß, aber nicht das einzig denkbare, welches hier verwendet werden kann. Der Prozess des Modellierens führt schließlich zu einem finalen Modell. Die Ergebnisse dieses finalen Modells werden mit grafischen und numerischen Mitteln dargestellt. Die Modellierung und Ergebnisinterpretation ist oft auf eine inhaltliche Forschungsfrage ausgerichet. In diesem Beispiel geht es in der inhaltlichen Forschungsfrage um den Verdienstunterschied zwischen Frauen und Männern. Unter den infrage kommenden Modellen wird das Modell mit dem geringsten AIC gewählt.

Dieses Modell 2 lautet:

$$M_2 : \textit{(Einkommen)} \sim \textit{(Alter)} + \textit{(Alter)}^2 + \textit{(Frau)} + \textit{(Abschluss)} + \textit{(Frau\_Branche)}$$

Die Ergebnisse sind in Tabelle 3 gegeben. Beruhend auf dem Modell erkennt man, dass Frauen im Durchschnitt etwa 192 Euro weniger verdienen als Männer. Dabei wird ein Modell zu Grunde gelegt, das mit Bezug auf den AIC eine gute Modellanpassung hat. Mit Blick auf das Konfidenzintervall erkennt man, dass die Lohndifferenz zwischen 175 Euro und 210 Euro liegt. Dieses Konfidenzintervall enthält Werte der Lohnlücke, die mit den Daten kompatibel sind. Für jeden Wert im Konfidenzintervall wird die Nullhypothese, dass der Parameterschätzer von $\beta_{(Frau)}$ eben diesen Wert hat, beibehalten. Das Konfidenzintervall quantifiziert die Unsicherheit des Inferenzschlusses. Das ist die Form von Unsicherheit, die dadurch zustande kommt, dass man von einer Stichprobe auf die Grundgesamtheit schließt.

Bei der Interpretation dieser Regressionsergebnisse kann man nun auch den p-Wert mit anführen. Der p-Wert ist die Überschreitungswahrscheinlichkeit. Konkret bedeutet das folgendes: Wie oben bereits erwähnt, ist die Differenz zwischen zwei Devianzen χ^2-verteilt. Man kann also die Wahrscheinlichkeit berechnen, dass unter der Gültigkeit des kleineren Modells (das entspricht H_0, nämlich dass der Regressionskoeffizient $\beta_{(Frau)}$ gleich Null ist) die Differenz der Devianzen Werte annimmt, die noch höher sind als der gegebene

Parameter	Schätzer	KI unten	KI oben
(Intercept)	467.55	385.96	549.14
(Alter)	27.4	23.3	31.47
(Alter)^2	-0.23	-0.28	-0.19
(Frau)	-192.06	-209.25	-174.87
(Abschluss: Abi)	76.62	62.22	91.02
(Abschluss: Uni)	496.88	481.19	512.57
(Frau_Branche)	-59.12	-77.22	-41.0

Tabelle 3: Regressionsergebnisse des besten Modells

Wert dieser Differenz. Diese Wahrscheinlichkeit nennt man Überschreitungswahrscheinlichkeit oder p-Wert. Wenn das Konfidenzintervall die Null nicht enthält, dann ist der entsprechende p-Wert ≤ 0.05. [11]

2.6.2 Interpretation von Regressionskoeffizienten

Wie können die in Tabelle 3 angegebenen Regressionskoeffizienten korrekt interpretiert werden?

Man kann grob zwei Arten der Interpretation von Regressionskoeffizienten unterscheiden(vgl. Gelman u. a., 2020, Kap. 10.2): die *prädiktive Interpretation* und die *kontrafaktische Interpretation.* Bei der *prädiktiven Interpretation* kann man vergleichen, wie sich der Mittelwert der Outcome-Variable ändert, wenn man Prädiktoren auf bestimmte Werte setzt. So kann man beschreiben, wie stark die Outcome-Variable zwischen Subpopulationen schwankt. In der *kontrafaktischen Interpretation* interpretiert man Regressionskoeffizienten als mittlere Veränderungen der Outcome-Variable, wenn man einen bestimmten Prädiktor verändert. Man fragt hier: Was hätte ich beobachtet, wenn dieser Prädiktor einen anderen Wert hätte? Bei der *kontrafaktischen Interpretation* kann man daher wirklich von einem Effekt des entsprechenden Prädiktors sprechen, bei der *prädiktiven Interpretation* sollte dieser Begriff nicht genutzt werden. Es handelt sich hier nur um eine Charakterisierung

[11] Man sollte die Bedeutung des p-Werts hier nicht überinterpretieren.

von Subpopulationen, die (unter bestimmten Modellannahmen) verglichen werden.

Eine Art der Ergebnisdarstellung, die sich für nichtlineare Modelle oder Modelle mit Interaktionen für die *prädiktive Interpretation* eignet sind die sogenannten AMEs (average marginal effects) Arel-Bundock u. a. (2024). Wenn man z. B. den AME für den Prädiktor *(Alter)* berechnen möchte, geht man wie folgt vor: Man berechnet für jede Person im Datensatz die Steigung der Variable Alter. Diese Variable ist quadratisch in der Modellgleichung, daher ist ihre Ableitung also die Steigung, vom Wert der Variable abhängig. Im nächsten Schritt berechnet man die durchschnittliche Steigung, es ergibt sich für Modell M_2 ein Wert von 7.96 mit einem Standard Error von 5.60. Die mittlere Steigung kann auch gruppenspezifisch berechnet werden z. B. getrennt für Frauen und Männer. Für die Prädiktoren, die linear und ohne Interaktion im Modell vorkommen unterscheidet sich der AME nicht von dem in Tabelle 3 angegeben Regressionskoeffizienten. Insgesamt kann man die Ergebnisse wie folgt interpretieren: Das mittlere Einkommen unterschiedet sich deutlich in den durch die Prädiktoren beschriebenen Subpopulationen. Die Unterschiede sind inhaltlich relevant[12] und statistisch signifikant. Die Modellselektion beruht auf dem Vergleich von drei Modellen beruhend auf dem AIC. Wenn man zwei Modelle vergleicht, wählt man das Modell, dass die niedrigere Devianz hat. Der AIC integriert in diese Entscheidung auch die Anzahl der Parameter, so dass große Modelle eher nicht gewählt werden. So gelangt man zu einem Kompromiss zwischen Komplexität des Modells und mangelnder Anpassung an die Datenstruktur.

Um zu prüfen, ob das gewählte Modell die Daten tatsächlich adäquat zusammenfasst, sollte in einer realen Analyse eine Regressionsdiagnostik durchgeführt werden (siehe z. B. Gelman und Hill, 2007, Kap. 3.6). Ergänzend sind Robustheitsbetrachtungen sinnvoll, um die Auswirkungen alternativer Modellentscheidungen systematisch zu untersuchen.

Ein zentraler Aspekt der Modellselektion ist die Frage, wie viele Modelle man überhaupt vergleichen darf, ohne die Aussagekraft der Analyse zu

[12]Wie hoch der Unterschied sein muss, um als inhaltlich relevant zu gelten, sollte vor Beginn der Analyse festgelegt werden.

gefährden. Um diese Frage fundiert beantworten zu können, ist ein Grundverständnis für die Stichprobenverteilung von Differenzen in Devianzen notwendig. Denn durch den Modellvergleiche mittels Hypothesentests verändert sich mit der Anzahl der Vergleiche die Verteilungseigenschaft des Teststatistik und damit die Interpretation der Ergebnisse.

Ziel der Modellwahl ist es, ein Modell zu identifizieren, das die Struktur der Daten möglichst prägnant zusammenfasst. Dabei können durchaus mehrere Modelle in Frage kommen. Wenn diese zu inhaltlich vergleichbaren Schlussfolgerungen führen, ist die Zusammenfassung als gelungen zu betrachten – unabhängig davon, welches Modell im Detail gewählt wurde.

3 Vertiefende Betrachtung

In diesem Abschnitt wird beruhend auf der mathematischen Definition der Devianz das Verständnis für die Nutzung der Devianz zum Modellvergleich vertieft. Man kann Parameter schätzen, indem man die Devianz minimiert und damit die Modellanpassung maximiert. Dieses Prinzip der Parameterschätzung wird als Maximum-Likelihood-Schätzung (ML-Schätzung) bezeichnet. [13]

3.1 Likelihood und Devianz als Grundlage der frequentistischen Parameterschätzung

Die Log-Likelihood auf der deviance scale, kurz Devianz, für eine multiples lineares Regressionsmodell M ist definiert als

$$D_M(\hat{\beta}, \hat{\sigma}) := -2 \log(L_M\big(\hat{\beta}, \hat{\sigma}\big)), \tag{29}$$

wobei L_M die Likelihood des Analysemodells M bezeichnet. Die Devianz eines Modells kann als statistische Zusammenfassung gesehen werden, die die Modellanpassung quantifiziert, analog zu den Residuen in der multiplen Regression (vgl. Gelman und Hill, 2007, S. 100).

[13] Eine sehr verständliche Einführung in die Logik der ML-Schätzung für die Einfachregression findet sich im Online R-Buch Applied Statistics with R in den Kapiteln 7.5 und 7.6 oder in Fahrmeir u. a. (2016, Kapitel 9.3)

Die Likelihood des Analysemodells ist

$$L_M(\underline{\beta}, \sigma) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \cdot \exp\left(- \frac{\left(y_i - \sum_{i=1}^{p} \beta_j \underline{x}_{i,j}\right)^2}{2\sigma^2} \right).$$

Zur Schätzung der Regressionskoeffizienten wird der ML-Schätzer berechnet. Für den Vektor der Regressionskoeffizienten $\hat{\beta}$, der die Likelihood des Analysemodells maximiert, gilt

$$\hat{\beta} := \arg\max_{\underline{\beta}} L_M(\underline{\beta}, \sigma).$$

Es gilt

$$\hat{\beta} = \arg\min_{\underline{\beta}} \left(y_i - \sum_{i=1}^{p} \beta_j \underline{x}_{i,j} \right)^2$$

und somit entspricht der ML-Schätzer dem KQ-Schätzer. Für die Devianz $D_M(\hat{\beta}, \hat{\sigma})$ des Analysemodells M ergibt sich durch Gleichung (29)

$$D_M(\underline{\hat{\beta}}, \underline{y}) = -2 \cdot \log(L_M\big(\hat{\beta}, \hat{\sigma})\big) =$$

$$= -2 \cdot \left(- \frac{n}{2} \log(2\pi\hat{\sigma}^2) - \sum_{i=1}^{n} \frac{\left(y_i - \sum_{i=1}^{p} \underline{\hat{\beta}_j \underline{x}}_{i,j}\right)^2}{2\hat{\sigma}^2} \right) =$$

$$= n\big(\log(2\pi) + \log(\hat{\sigma}^2) + 1\big) \tag{30}$$

Zusammenfassend: Wenn zwei Modelle M_0 und M_1 gegeben sind, dann kann man die jeweiligen Devianzen bzw. die entsprechenden Differenzen dieser Devianzen nutzen, um die Anpassungsgüte zu vergleichen und man kann auch einen statistischen Test entwickeln, um eine Entscheidung für eines der beiden Modelle zu treffen. Diese beruhen auf dem Satz von Wilks. Dieser Satz sagt aus, dass die Differenzen von Devianzen einer χ^2-Verteilung folgen.

3.2 Modellvergleiche beruhen auf Devianz und AIC

Die Devianz kann für unterschiedliche Modelle berechnet werden. Oft interessiert man sich dafür, ob die Hinzunahme einer Variable zu einer deutlichen

Verbesserung der Modellanpassung führt. Um diese Frage mit Blick auf die Devianz beantworten zu können, sei das Modell M_0 ein Modell, welches in einem größeren Modell M_1 enthalten ist. Mit anderen Worten, man erhält M_0 indem man einige der Parameter die in M_1 enthalten sind herausnimmt. Mathematisch gesehen bedeutet das, dass man den Parametervektor des Modells M_0 erhält, indem man einige Komponenten des Parametervektors von M_1 gleich Null setzt.

Der Parametervektor des Modells M_1 wird mit $\underline{\beta}$ bezeichnet, dieser hat die Länge $p+1$. Der Parametervektor des Modells M_0 mit $\underline{\gamma}$ und hat die Länge $l+1$. Es gilt $p > l$. Die Devianz des Modells M_1 ist sicher geringer als die des Modells M_0. Das gilt, da das Modell M_1 alle Parameter des Modells M_0 enthält. Daher ist das Modell M_1 mit Sicherheit besser an die Daten angepasst als M_0. Die Diskrepanz $D_{M_0} - D_{M_1}$ hat also ein positives Vorzeichen. Wenn diese Diskrepanz einen hohen Betrag hat, bedeutet das inhaltlich, dass das Modell M_1 eine deutlich bessere Datenanpassung bietet als M_0. Man braucht eine Messlatte, um abzumessen, was eine „große“ Diskrepanz bedeutet. Insbesondere ist es in der frequentistischen Herangehensweise Ziel des Inferenzschlusses, mit kontrollierter Irrtumswahrscheinlichkeit von der Stichprobe auf den zu Grunde liegenden probabilistischen Zusammenhang zu schließen. Um das tun zu können brauchen wir die Stichprobenverteilung der Statistik $\tilde{D}_{1,0}(\underline{\hat{\beta}}, \underline{\hat{\gamma}}) := D_{M_0}(\underline{\hat{\beta}}) - D_{M_1}(\underline{\hat{\gamma}})$. Es lässt sich zeigen, dass asymptotisch

$$\tilde{D}_{1,0}(\underline{\hat{\beta}}, \underline{\hat{\gamma}}) \sim \chi^2(k) \tag{31}$$

gilt, sofern das Modell M_0 das wahre Modell ist. Dabei ist $\chi^2(k)$ die χ^2-Verteilung mit $k = p - l$ Freiheitsgraden (vgl. z. B. Pruscha, 2000, S. 288, 255, Beispiel (c)). Dieser zentrale Sachverhalt wird als Satz von Wilks bezeichnet. Der Satz von Wilks liefert, analog zum AIC, eine Grundlage für den Vergleich von zwei Modellen. Beruhend auf dem Satz von Wilks lassen sich sogar Hypothesentests und Konfidenzintervalle herleiten.

Durch Gleichung (31) wird ein mathematischer Sachverhalt ausgedrückt, doch was hat dieser Ausdruck mit der Realität zu tun? An dieser Stelle ist es wichtig, sich die Grundlogik des frequentistischen Denkens vor Augen zu

führen. Die Zufallsstichprobe $\underline{Y}$ entsteht durch die Ziehung einer Zufallsstichprobe.[14] Beruhend auf der Stichprobe $\underline{Y}$ kann man eine Reihe von sogenannten Statistiken berechnen. Eine Statistik ist definiert als Funktion die sich aus der Stichprobe $\underline{Y}$ berechnen lässt, wie z. B. der Mittelwert $\bar{Y}$ der Stichprobe. Auch der Median bzw. die Varianz sind solche Statistiken. Als Statistik bezeichnet man also allgemein eine Größe, die man aus der konkreten Stichprobe berechnen kann. Die Begriffe Teststatistik oder Schätzstatistik werden verwendet, wenn das Ziel der Herangehensweise darin besteht einen statistischen Test bzw. einen Schätzer zu konstruieren. Die Statistik $\tilde{D}_{1,0}(\underline{\hat{\beta}}, \underline{\hat{\theta}})$ ist hier unsere Teststatistik. Ziel ist es einen Test zu konstruieren, mit dem man die Hypothese

$$H_0 : M_0 \text{ ist das wahre Modell}$$
$$\text{versus}$$
$$H_1 : M_1 \text{ ist das wahre Modell}$$

testen kann. Die Gleichung (31) informiert uns über die (asymptotische) Verteilung der Teststatistik $\tilde{D}_{1,0}(\underline{\hat{\beta}}, \underline{\hat{\theta}})$ unter H_0. Diese Gleichung besagt inhaltlich folgendes: Wenn man immer wieder (unter den gleichen Bedingungen) eine Zufallsstichprobe $\underline{Y}$ zieht und jedes Mal den entsprechenden Wert der Teststatistik $\tilde{D}_{1,0}(\underline{\hat{\beta}}, \underline{\hat{\theta}})$ berechnet, dann ist unter H_0 die Verteilung die man daraus erhält eine χ^2-Verteilung mit k Freiheitsgraden. Es ist in der frequentistischen Herangehensweise üblich, die tatsächlich beobachtete Stichprobe zu vergleichen, mit allen anderen (nicht-beobachteten) Stichproben, die unter H_0 möglich gewesen wären. Um den Vergleich der tatsächlich beobachteten Stichprobe mit den unter H_0 möglichen Stichproben konkret durchzuführen, berechnet man den Wert der Teststatistik für die Stichprobe. Diesen berechneten Wert kann man dann mit den entsprechenden Werten, der unter H_0 möglichen Stichproben, vergleichen.

[14] Die Prädiktoren $x_1 \dots x_p$ werden als deterministisch betrachtet bzw. das Modell wird bedingt auf die Realisation der Prädiktoren formuliert.

3.3 Statistische Signifikanz: p-Werte, Devianz und Freiheitsgrade

Nachdem durch die Gleichung (31) eine Teststatistik gegeben ist, deren Verteilung unter H_0 bekannt ist, kann ein Signifikanztest hergeleitet werden. Ein statistischer Test stellt, wie in Abschnitt 2.4.2 bereits ausgeführt, eine Entscheidungsregel dar, die darauf ausgerichtet ist, mit kontrollierter Irrtumswahrscheinlichkeit auf die Grundgesamtheit bzw. allgemeiner auf einen probabilistischen Zusammenhang zu schließen. Es handelt sich also um induktives Verhalten, dass sich dadurch auszeichnet, dass die Irrtumswahrscheinlichkeit vorgegeben werden kann.

Die Vorgehensweise bei der Konstruktion des Tests folgt den Schritten die in Abschnitt 2.4.1 beschrieben wurden. Um die Konstruktion der Teststatistik in diesen Schritten zu zeigen, soll hier die Nullhypothese $\beta_p = 0$ getestet werden, d. h. $\underline{\gamma} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}, 0)$. Der Parametervektor des Modells M_1 ist $\underline{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}, \beta_p)$. Zur Wiederholung: Die maximal akzeptierte Irrtumswahrscheinlichkeit wird im vierten Schritt festgelegt. Sie wird als Signifikanzniveau α bezeichnet. Im fünften Schritt wird der Ablehnbereich festgelegt. Fällt der realisierte Wert der Teststatistik in den Ablehnbereich, so wird die Nullhypothese abgelehnt. Schließlich wird der realisierte Wert der Teststatistik im sechsten Schritt berechnet. Im letzten siebten Schritt wird dann die formale Entscheidung getroffen, entweder muss H_0 beibehalten werden oder es kann H_1 nachgewiesen werden.

Beruhend auf dem Satz von Wilks (siehe Gleichung (31)) wissen wir, dass unter Gültigkeit von H_0 die Teststatistik $\tilde{D}_{1,0}(\hat{\underline{\beta}}, \hat{\underline{\gamma}})$ einer χ^2-Verteilung mit einem Freiheitsgrad folgt. Das 95%-Quantil der $\chi^2(1)$-Verteilung ist 3.84. Wenn die Realisation der Teststatistik diesen Wert überschreitet, dann wird H_0 abgelehnt. Für die Testentscheidung kann alternativ der p-Wert verwendet werden. So wie die Teststatistik selbst, ist auch der p-Wert eine Zufallsvariable. Die Testentscheidung beruht darauf festzustellen, ob die Realisation einer Zufallsvariable im kritischen Bereich liegt oder nicht. Der kritische Bereich wird dadurch festgelegt, dass die Verteilung der Teststatistik unter H_0 bekannt ist. Im konkreten Fall ist der kritische Bereich $[3.84, \infty)$.

Man kann die Teststatistik $\tilde{D}_{1,0}(\hat{\underline{\beta}}, \hat{\underline{\gamma}})$ so transformieren, dass dabei eine gleichverteilte Zufallsvariable entsteht. Diese Zufallsvariable heißt p-Wert und ist definiert als $V := \mathbb{P}(T > t) = 1 - F(T)$, wobei F die Verteilungsfunktion der Teststatistik $\tilde{D}_{1,0}(\hat{\underline{\beta}}, \hat{\underline{\theta}})$ bezeichnet. Die so transformierte Zufallsvariable ist gleichverteilt (vgl. z. B. Murdoch u. a., 2008). Für diese gleichverteilte Zufallsvariable V ergibt sind der kritische Bereich zu $(0, 0.05]$. Die Nullhypothese wird also abgelehnt, wenn wenn $V \leq 0.05$ gilt. Man kann die Testentscheidung daher treffen, indem man entweder die Realisation der Teststatistik $\tilde{D}_{1,0}(\hat{\underline{\beta}}, \hat{\underline{\theta}})$ abgleicht mit dem kritischen Bereich $[3.84, \infty)$ oder indem man die Realisation von V abgleicht mit dem kritischen Bereich $(0, 0.05]$. Beide Vorgehensweisen sind mathematisch äquivalent und führen daher immer zu selben Ergebnis.

Festzuhalten ist unbedingt, dass die Zufallsvariable V eben eine Zufallsvariable ist und keine Wahrscheinlichkeit. Die Realisation dieser Zufallsvariable ist Grundlage der Testentscheidung. Wenn sie in den Ablehnbereich fällt, dann ist die Alternativhypothese signifikant. Das bedeutet konkret, dass man auf lange Sicht (in der long-run-frequency) in nur 5% der Fälle, H_0 fälschlicherweise ablehnt.

In vielen statistischen Analysen finden sich falsche Interpretationen von p-Werten, es gibt eine Reihe von etablierten Missverständnissen. Ein historischer Grund dafür ist, dass sich die Interpretationen des p-Wertes, die in der Praxis häufig verwendet werden, aus zwei unterschiedlichen und unvereinbaren statistischen Theorien zusammensetzen, nämlich der Herangehensweise von Neyman und Pearson auf der einen Seite und Fisher auf der anderen Seite (siehe z. B. Haig, 2018, Kap. 3). Im nächsten Abschnitt stehen nicht die historischen Ursachen für die Missverständnisse im Zusammenhang mit p-Werten im Fokus, sondern mögliche Auswirkungen solcher Missverständnisse [15]. Ein Prozess des Modellierens, der auf falschen Vorstellungen von statistischen

[15] Historische Entwicklungen der Statistik können in diesem Reader nicht umfassend behandelt werden, da ihre Komplexität den Rahmen eines kurzen Textes übersteigt. Für weiterführende Einblicke werden zwei Artikel empfohlen: Kennedy-Shaffer (2019) zur Entstehung des „Kults um den p-Wert“ und Kennedy-Shaffer (2024) zur historisch belasteten Entwicklung der Statistik.

Modellen und auf irreführenden p-Wert-Interpretationen beruht, kann zu einer Alchemie führen, die statistisches Rauschen in scheinbare wissenschaftliche Evidenz „verwandelt." Die Zufallsvariable p-Wert ist nur dann gleichverteilt, wenn eine bestimmte Vorgehensweise eingehalten wird. Wenn man von dieser Vorgehensweise abweicht, ändert sich die entsprechende Stichprobenverteilung und das führt zu irreführenden Schlüssen (siehe z. B. Meinfelder und Kluge, 2020, Kap. 3). [16]

4 Viele Entscheidungen: Garden of Forking Paths

In einem Artikel mit dem Titel *The Statistical Crisis in Science Data-dependent analysis— a „garden of forking paths" — explains why many statistically significant comparisons don't hold up* beschreiben Gelman und Loken (2014) einen „Garten der Weggabelungen" und meinen damit, dass im Prozess der Modellierung zu einem stark verzweigten Entscheidungsbaum führen. Im letzten Abschnitt dieses Readers wird mit Bezug auf das Anwendungsbeispiel ein kleiner Einblick in diesen „garden of forking paths" gegeben.

Die Statistik ist eine Wissenschaft, die sich systematisch mit der Frage nach dem Umgang mit unterschiedlichen Formen von Unsicherheit beschäftigt. Jeder Inferenzschluss ist mit Unsicherheit behaftet und die Leistung eines Inferenzkonzeptes besteht darin verlässliche Schlussverfahren zu begründen. Eine vertretbare statistische Analyse zeichnet sich dadurch aus, dass der Inferenzschluss nachvollziehbar und transparent ist. In der statistischen Fachwelt wird seit Jahrzehnten kritisiert, dass zahlreiche statistische Analysen an dieser Stelle mangelhaft sind.

Kritisiert wird u.a. ein Umgang mit statistischen Methoden der mehr am Erfolg der beteiligten Wissenschaftler*innen als an der Richtigkeit des Ergebnisses interessiert ist (Van Calster u. a., 2021). Ferner wird kritisiert, dass in

[16] Ein interessanter Artikel, der praktische Hinweise zur Interpretation des AIC ist Sutherland u. a. (2023). Dieser Artikel ist empfehlenswert, wenn man ein besseres Verständnis für den Zusammenhang von p-Werte, Devianz und AIC entwickeln möchte.

der Wissenschaft Methoden genutzt werden, die nicht vertrauenswürdig genug sind, da ihre tatsächlichen Anwendungspotenziale und Grenzen zu wenig untersucht wurden (Heinze u. a., 2024b) oder ein wissenschaftliches System, dass methodologische Kurzsichtigkeit mehr fördert als valide Herangehensweisen (van Ravenzwaaij u. a., 2023). Ähnliche Vorstöße gibt es auch von der American Statistical Association, die zu einem Reformprozess aufgerufen hat (Wasserstein u. a., 2019).

Über das vermutlich bekannteste Problem hat sogar die Presse berichtet:

Grobe Fehler in Statistik: Der "p-Wert"gilt als Goldstandard, doch er führt in die Irre. Er schadet damit seit Jahren der Wissenschaft.[17]

Der p-Wert: Alle kennen ihn, die meisten missbrauchen ihn, kaum einer versteht ihn. Statistiker fordern nun ein Umdenken. Denn: Forschende ohne Statistikkenntnisse schaden der Wissenschaft. [18]

oder

Signifikanter Unfug: Die statistische Signifikanz, gemessen mit dem sogenannten p-Wert, hat in der Wissenschaft eine geradezu götzenhafte Bedeutung erlangt. 800 Forscher beklagen Fehler und fordern ein Umdenken. [19]

Im nächsten Abschnitt folgt ein kleiner Einblick in das „p-Wert-Problem“ und die damit verbundenen Missverständnisse.

4.1 Statistische Alchemie: Wenn Rauschen zu Erkenntnis wird

Im Laufe des Prozesses der Modellierung muss man viele Entscheidungen treffen. Diese betreffen z. B. die Auswahl der Variablen, die ins Modell auf-

[17] `www.spektrum.de/news/statistik-wenn-forscher-durch-den-signifikanztest-fallen/1224727`

[18] www.nzz.ch/karriere/studentenleben/der-problem-wert-ld.133818

[19] https://www.sueddeutsche.de/wissen/statistik-p-wert-signifikanz-hypothese-nullhypothese-1.4375636

genommen werden, die funktionale Form oder die Frage ob, für einige Variablen Interaktionen berücksichtigt werden sollten. Oft beginnt man mit einem plausiblen Modell oder einer Liste von plausiblen Modellen und wählt ad-hoc ein finales Modell aus. Diese Entscheidung beruht nicht selten auf Kriterien, die keine belastbare statistische Grundlage haben. Am Ende lesen sich viele Publikationen so, als wären all die Entscheidungen im Prozess der Modellbildung kaum wichtig für die finale Ergebnisinterpretation. Bereits vor 25 Jahren wurde das Problem treffend zusammengefasst:

Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are. (Hoeting u. a., 1999)

Was meint die bekannte Statistikerin Jennifer Hoeting mit überzogenen Schlussfolgerungen und Entscheidungen, die risikoreicher sind, als man denkt? Eine fehlerhafte etablierte Herangehensweise die zu solchen falschen Schlüssen führt, wird in der Fachliteratur als "p-Hacking"bezeichnet.

p-Hacking kann, insbesondere im Zusammenspiel mit anderen Fehlern in der statistischen Inferenz, je nach Forschungskontext zu gravierenden Fehlinterpretationen und methodischen Problemen führen. In den Worten von Andrew Gelman: *statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an "un- certainty laundering" that begins with data and concludes with success as measured by statistical significance* (siehe Quelle). Wie kann eine solche Alchemie entstehen?

Um diesen Sachverhalt anschaulicher zu machen, werden dem Datensatz 25 Variablen hinzugefügt, die reines Rauschen sind, also nichts mit dem Modell zu tun haben. Diese Variablen werden mit V_1 bis V_{25} bezeichnet. Für jede dieser Variablen ist die Wahrscheinlichkeit, dass der entsprechende p-Wert signifikant ist 5%, denn p-Werte sind unter der Nullhypothese gleichverteilte Zufallsvariablen. Bei 20 Variablen ist daher die Wahrscheinlichkeit, dass mindestens eine signifikante Variable dabei ist

$$1 - (0.95)^{20} = 64.15\%.$$

Wenn man sogar 25 Variablen ausprobiert, kommt man auf 72.26%.

Das p-Hacking bietet also gerade in großen Datensätzen ein gewisses Potential signifikante Effekte zu finden, die keine sind. Gerade wenn große Datensätze, wie ALLBUS oder PISA, von vielen Forschenden genutzt werden, kann man sich gut vorstellen, dass diese Form des Overfitting auftritt. Die Problematik kann noch dadurch verschärft werden, dass man Hypothesen nachträglich ändert bzw. formuliert (man spricht von Harking (Kerr, 1998): „Hypothesizing After the Results are Known“) oder die Relevanz des Effekts (also die Effektstärke) mit der statistischen Signifikanz vermischt und die Begriffe „statistische Hypothese“ und „inhaltliche Hypothese“ fälschlicherweise gleichsetzt. Gelman und Geurts (2017) sprechen von einer „‘sociology’ of the process of research and publication,“ die erklärt, wie es dazu kommen kann, dass wohlmeinende Forschende Publikationsprozesse in einem Fachgebiet über Jahrzehnte aufrechterhalten werden können, selbst wenn es keine konsistenten zugrunde liegenden Effekte gibt.

Im Beispiel werden 25 Variablen simuliert, die keinen Bezug zur Outcome-Variable haben, also reines Rauschen sind. Die Variablen enthalten ganzzahlige Werte zwischen 0 und 20, so dass sie z. B. Likert-Skalen entsprechen. Zur Veranschaulichung werden auch diese Variablen inhaltlich belegt: Die Variablen $V1$ bis $V10$ beziehen sich auf Haltungen die Naturverbundenheit, Religiosität bzw. Spiritualität messen, die Variablen $V11$ bis $V20$ enthalten Einstellungen zu Gerechtigkeit im Job für Männer und Frauen, zu Familie, zu beruflichen Zielen und insbesondere zur Wichtigkeit der eigenen Karriere in der Lebensplanung usw. Die letzten fünf Variablen $V21$ bis $V25$ enthalten die Selbsteinschätzung zu Persönlichkeitseigenschaften, konkret Anstrengungsbereitschaft, Durchhaltevermögen, Motivation, Zielstrebigkeit und Selbstwirksamkeit.

Jede Variable wird einzeln ausprobiert und in dem Modell $M2$ aus Abschnitt 2.5.1 hinzugefügt. Die Variablen $V6$, $V17$, $V19$, und $V25$ haben p-Werte die höchstens 0.1 betragen und werden in das finale Modell aufgenommen, dieses ist damit

$$M_{final} : \textit{(Einkommen)} \sim \textit{(Alter)} + \textit{(Alter)}^2 + \textit{(Frau)} + \textit{(Abschluss)} +$$

(Frau_Branche) + (V6) + (V17) + (V19) + (V25)

Im finalen Modell haben schließlich alle Variablen außer $V6$ und $V17$ einen p-Wert der höchstens 0.05 beträgt. Der p-Wert von $V6$ und $V17$ liegt unter 0.1. Die Ergebnisse finden sich im beiliegenden `R`-Code. Hervorzuheben ist, dass dieses Ergebnis *statistisch signifikant* ist, was nicht unbedingt bedeutet, dass es *inhaltlich relevant* ist. Wenn die inhaltliche Relevanz nicht vor Beginn der Analyse definiert wurde (beispielsweise, dass man erst ab einer bestimmten Effektstärke von einem inhaltlich relevanten Ergebnis sprechen kann), kommt es leicht dazu, dass man die statistische Signifikanz und die inhaltliche Relevanz miteinander verwechselt. Die Signifikanz trifft jedoch nur eine Aussage darüber, ob ein Parameterschätzer systematisch von Null abweicht.[20]

Es wurden somit signifikante Effekte nachgewiesen, gemäß denen hohe berufliche Ziele und eine hohe Selbsteffizienz das Einkommen erhöhen. Außerdem wurde ein immerhin schwach signifikanter Hinweis darauf gefunden, dass Naturverbundenheit das Einkommen signifikant reduziert und eine positive Haltung zur Gleichberechtigung von Mann und Frau im Job das Einkommen signifikant erhöht. Jetzt kann nachträglich eine entsprechende Hypothese formuliert werden und schon wurde *Harking* betrieben. Die anderen Variablen im Modell bekommen dabei beispielsweise die Rolle „Kontrollvariable." Bei dieser irreführenden Schlussweise wird meist auch die inhaltliche Hypothese mit der statistischen Hypothese gleichgesetzt und es wird oft die kontrafaktische Interpretation der Regressionskoeffizienten verwendet, obwohl durch die Modellierung nur die prädiktive Interpretation gerechtfertigt ist [21]. [22]

In einem Nachtrag zu ihrem Artikel *The Statistical Crisis in Science-Data-dependent analysis— a „garden of forking paths" — explains why many*

[20]Die Wahrscheinlichkeit ein statistisch signifikantes Ergebnis zu erzielen steigt bei gleicher Effektstärke mit wachsendem Stichprobenumfang.

[21]Der Wissenschaftstheoretiker Haig weist auf die zentrale Bedeutung hin, die eine unbegründete Gleichsetzung der statistischen und inhaltliche Hypothese für irreführende Schlüsse in der Psychologie hat. (Haig, 2018, S. 57)

[22]Leser*innen, die sich vertiefend mit dem Thema auseinandersetzen möchten, wird die Lektüre von Meinfelder und Kluge (2020) empfohlen.

statistically significant comparisons don't hold up weisen Gelman und Loken darauf hin, dass in vielen Analysesituationen eine Zweckentfremdung statistischer Methoden nicht absichtlich geschehen muss (siehe Quelle). In diesem Artikel schreiben die Autoren *„In this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done implicitly. The researchers are not trying multiple tests to see which has the best p-value; rather, they are using their scientic common sense to formulate their hypotheses in reasonable way, given the data they have. The mistake is in thinking that, if the particular path that was chosen yields statistical significance, that this is strong evidence in favor of the hypothesis.“* Die Autoren stellen dann klar, dass es sich in vielen Fällen um eine Überinterpretation der Ergebnisse handelt. Die Ergebnisse sind oft, wenn man ihren eigentlichen statistischen Gehalt betrachtet, teilweise explorativ und fragil und sollten auch so kommuniziert werden. So fordern Gelman und Loken schließlich: *„As a result, the authors of such studies would have to recognize that the evidence in favor of their research hypotheses is much weaker than they had presumed.“*

4.2 Initial Data Analysis (IDA) – Ein guter erster Schritt

Fehler in der statistischen Analyse entstehen häufig nicht aus mangelnder Sorgfalt, sondern schleichen sich ein, weil im Vorfeld zentrale Aspekte der Datenstruktur und Modellierung nicht ausreichend bedacht wurden. Gerade in der Anwendungspraxis zeigt sich, dass Forschende oft zu früh in komplexe Analysen einsteigen – ohne ein fundiertes Verständnis der Eigenschaften, Zusammenhänge und potenziellen Fehlerquellen ihrer Daten. Dies kann zu Fehlinterpretationen und methodisch fragwürdigen Ergebnissen führen.

Ein vielversprechender Ansatz, um diesem Problem zu begegnen, ist die Initial Data Analysis (IDA). Sie stellt eine systematische Vorstufe zur eigentlichen Modellierung dar und zielt darauf ab, die Daten zunächst in ihrer Struktur, Qualität und Besonderheit zu verstehen. IDA hilft, die Angemessenheit der Modellstrategie zu prüfen, die Interpretation der Ergebnisse vorzubereiten und die Präsentation der Resultate zu strukturieren. Ein zentraler Grundsatz von IDA ist dabei, auf die Bewertung von Zusammenhängen zwi-

schen Outcome und Prädiktoren zu verzichten – um Verzerrungen in der späteren Inferenz zu vermeiden Heinze u. a. (2024a).

Damit IDA ihr volles Potenzial entfalten kann, sollte sie vorab geplant und als fester Bestandteil in den statistischen Analyseplan eines Projekts integriert werden. Eine gut dokumentierte IDA erhöht nicht nur die Reproduzierbarkeit, sondern schafft auch die Grundlage für eine robuste und transparente Modellierung. Empfehlungen zur Ausgestaltung eines IDA-Plans sowie ein Beispiel aus einem diagnostischen Modellierungsprojekt finden sich bei Heinze u. a. (2024a).

4.3 Gute Entscheidungen: Reasoning With Data

Der Satz *„Choosing inputs to a regression is often the most challenging step in the analysis "* bringt zum Ausdruck, dass eine zentrale Herausforderung im Umgang mit Modellunsicherheit die Variablenselektion betrifft. Er findet sich in (Gelman und Hill, 2007, S. 45) in einem Abschnitt zur Modelldiagnostik. Die Modell- bzw. Variablenselektion kann unterschiedlich begründet werden, z. B. durch inhaltliches Vorwissen oder durch statistische Gütekriterien wie dem AIC.

Die Einbindung einer Regression in eine statistische Argumentation ist eine anspruchsvolle Aufgabe, die von der konkreten Forschungsfrage und Analysesituation abhängt. Oft braucht man dafür nicht nur statistisches und inhaltliches Wissen über die Analysesituation, sondern auch Erfahrung. Statistisches Denken bedeutet aus einer Reihe von mathematischen Annahmen unter Nutzung empirischer Daten eine nachvollziehbare Argumentation aufzubauen. Statistik ist „reasoning with data" oder genauer gesagt, die Theorie die die Konzepte der Fachsprache, für das „reasoning with data" definiert und diskutiert (Frøslie und Røislien, 2019). [23] Nold und Heinze (2025) hebt hervor, wie herausfordernd es ist, Statistik auf unterschiedlichen Anspruchsebenen zu

[23]Ein zentraler Bestandteil dieser Fachsprache ist auch die Bayes-Inferenz, die alternative Wege des statistischen Schließens eröffnet und insbesondere bei komplexen Modellierungsfragen wertvolle Perspektiven bietet. Aufgrund des Umfangs kann sie in diesem Reader nicht vertieft behandelt werden. Für eine fundierte Einführung wird auf die Lehrbücher von Johnson u. a. (2022) und Kaplan (2023) verwiesen, die die Prinzipien der Bayes'schen

vermitteln – insbesondere im Hinblick auf die Fähigkeit, statistische Literatur kritisch zu lesen und die eigenen methodischen Grenzen zu erkennen. Wichtig sind hier sowohl die aktive als auch passive Methodenkompetenz: Aktiv bedeutet, selbst Analysen durchführen und die einzelnen Schritte begründen zu können, passiv bedeutet, veröffentlichte statistische Analysen kritisch zu lesen und einschätzen zu können, inwieweit deren methodische Entscheidungen nachvollziehbar sind oder wo eine eigene Bewertung nicht möglich ist. Dieser Reader möchte hierfür eine Grundlage schaffen – ohne Anspruch auf Vollständigkeit, aber mit gezielten Verweisen auf weiterführende Literatur zur Anwendung und Bewertung statistischer Verfahren.

4.4 Reproduzierbare Forschung

Die Vertrauenswürdigkeit von statistischen Ergebnissen kann im Zusammenhang mit deren Replizierbarkeit bzw. Reproduzierbarkeit diskutiert werden (Peels und Bouter, 2023). Die Vertrauenswürdigkeit der Ergebnisse hängt maßgeblich von der Reproduzierbarkeit des Studienprotokolls ab. Das Studienprotokoll beinhaltet alle Schritte, die während der Datenanalyse vorgenommen werden. Es ist eine vollständige Beschreibung, wie die Forschungsfrage beantwortet wird. Im Idealfall enthält es alle Details, die für eine Reproduktion der Studie erforderlich sind. Theoretisch legt das Studienprotokoll nicht nur fest, wie die Forschungsfrage lautet und wie die Daten erhoben werden, sondern auch, wie die Daten analysiert werden. In der Praxis ist die Anleitung zur Datenanalyse oft in einem separaten Dokument enthalten: dem Datenanalyseplan (Synonym: Analyseplan). Eine Analyse wurde erfolgreich reproduziert, wenn eine unabhängige Re-Analyse der Originaldaten unter Verwendung desselben Studienprotokolls zu den gleichen Ergebnissen führt (Nuijten, 2022). Die Transparenz des Protokolls ist daher eine wesentliche Voraussetzung für die Reproduzierbarkeit einer Studie.

Dieser Reader möchte nicht die Verwendung von bestimmten Herangehensweisen empfehlen, wie z. B. die Verwendung des AICs zur Modellselektion, sondern dafür sensibilisieren, dass jede statistische Analyse einem gut

Statistik praxisnah und verständlich vermitteln.

durchdachten und transparent kommunizierten Forschungsplan folgen sollte. Die einzelnen Schritte der Analyse sollten dabei klar definiert und kohärent verwendet werden [24]. Die Offenlegung des Studienprotokolls, ist wichtig um die Vertrauenswürdigkeit von statistischer Forschung zu gewährleisten. Peels und Bouter (2023) formulieren es so: *„The availability of a detailed study protocol and data-analysis plan, however, does enable a replication and thereby opens up the possibility to increase its trustworthiness. That being said, even if replication studies are not carried out, the requirement of replicability may actually provide a strong incentive to perform the study well. The demand to be replicable comes with a risk that flaws and errors will be identified by others. That awareness in itself may make the results of a study more trustworthy due to a more careful execution.“*

Im Prozess der Modellierung werden zahlreiche subjektive Entscheidungen getroffen, diese Entscheidungen beeinflussen maßgeblich die Ergebnisse. Daher sind Transparenz und Reproduzierbarkeit wichtig um die Aussagen einer statistischen Analyse richtig einschätzen zu können. Denn die Bedeutung der Zahlen und Tabellen in statistischen Analysen wird durch den Prozess der Modellierung definiert:

> The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning. [25]

[24] In diesem Reader wurde eine deskriptive Forschungsfrage behandelt. Hier geht es darum, einen guten Kompromiss zwischen der Anpassung an die Daten und der Komplexität des Modells zu finden. Bei anderen Arten von Forschungsfragen, z. B. kausalen Forschungsfragen, ist die Modellspezifikation an andere Ziele geknüpft. Es gilt jedoch für alle Arten von Forschungsfragen: Die Aussagekraft und die konkrete Interpretation eines statistischen Modells hängen also von der Kohärenz des Studienprotokolls ab.

[25] Zitiert nach Spiegelhalter (2019), Original (Silver, 2012).

A R-Code

Im Ordner `Source`, der auf der begleitenden Webseite unter nold.info/RCodeReaderverfügbar ist, befinden sich folgende Dateien:

- **gpg.RData** – Beispieldaten, die im Reader verwendet werden.
- **rpackages.R** – Eine Liste der in der Analyse verwendeten R-Packages.
- **functions.R** – Benutzerdefinierte R-Funktionen zur Datenverarbeitung.

Diese Dateien können zur Reproduktion der im Reader dargestellten Analysen verwendet werden. Der Ordner ist über die Homepage direkt zugänglich.

Die folgenden Seiten enthalten den vollständigen R-Code zu den im Reader und im Anhang beschriebenen Beispielen:

- **R-Code im Reader** – enthält den vollständigen Code zu den im Haupttext behandelten Auswertungen und Visualisierungen.
- **R-Code in Appendix A2** – dokumentiert ergänzende Analysen und weiterführende Berechnungen, die im Anhang A2 erläutert werden.
- **R-Code in Appendix A3** – beinhaltet zusätzliche Funktionen und Auswertungen, die im Anhang A3 beschrieben sind.

Diese Seiten ermöglichen eine transparente Nachvollziehbarkeit der im Reader dargestellten Ergebnisse und dienen als Grundlage für eigene Reproduktionen oder Erweiterungen.

`R`-Code zu dem Geburtsgewicht-Beispiel (A 2)

Im `R`-Code findet sich ein zweites Beispiel, das auf einem realen Datensatz beruht.

Der Geburtsgewicht-Datensatz enthält 10 Variablen. Die Daten wurden im Baystate Medical Center in Springfield, Massachusetts, im Jahr 1986 erhoben. Die Daten enthalten Risikofaktoren, die vermutlich mit einem niedrigen

Geburtsgewicht des Kindes in Zusammenhang stehen. In dem Beispiel soll untersucht werden, welche Indikatoren, insbesondere die Handlungen und der Gesundheitszustand der Mutter während der Schwangerschaft, das Geburtsgewicht des Kindes beeinflussen. Das Modell beinhaltet als potentielle Einflussgrößen, die Angaben der Mutter zu ihrer Hautfarbe, zu ihrem Alter, ihrem Gewicht zu Beginn der Schwangerschaft, zum Vorliegen von Hypertonie und zum Raucherstatus der Mutter.

Simuliert werden 25 potentielle weitere Einflussfaktoren, die im z. B. Fragen zu Gesundheitsthemen entsprechen könnten. Fünf dieser Variablen sind erneut ßignifikant".

R-Code zum Human-Computer-Interaction (HCI) Beispiel (A 3

In einem dritten Beispiel wird der engen Zusammenhang zwischen dem Zwei-Stichproben t-Test und der Einfachregression dargestellt.

Inhalt des Beispiels

Forschungsfrage: Kann ein empathischer Chatbot die negativen Folgen sozialer Ausgrenzung abmildern?

Forschungshypothese: Ein Gespräch mit einem empathischen Chatbot nach einer Erfahrung sozialer Ausgrenzung führt zu einer signifikanten Verbesserung der Stimmung im Vergleich zu einer unempathischen Kontrollinteraktion.

In einem fiktiven Experiment nehmen die Proband*innen an einer Gruppendiskussion zu einem aktuellen Thema teil, die in einem Chatforum stattfindet. Während der Diskussion werden sie konsequent ignoriert, was eine Erfahrung sozialer Ausgrenzung darstellt. Anschließend werden die Teilnehmenden zufällig in zwei gleich große Gruppen mit jeweils 100 Personen aufgeteilt: Gruppe 0 erhält lediglich ein höfliches Dankeschön für die Teilnahme, während Gruppe 1 ein Gespräch mit einem empathischen Chatbot führt.

Zur Erfassung des emotionalen Zustands werden 25 verschiedene Skalen

erhoben, die jeweils unterschiedliche inhaltliche Dimensionen der Stimmung abbilden. In den zugrunde liegenden simulierten Daten besteht tatsächlich kein systematischer Zusammenhang. Dennoch zeigt die Simulation, dass sich für zwei der 25 Skalen signifikante Effekte nachweisen lassen – entweder mittels einer einfachen linearen Regression oder äquivalent durch einen Zwei-Stichproben-t-Test.

Äquivalenz zwischen Zwei-Stichproben t-Test und Einfachregression

Man kann sich durch folgende Überlegungen vergegenwärtigen, dass die Einfachregression mit einem binären Prädiktor und der Zweistichproben-t-Test (Fahrmeir u. a., 2016)[vgl. z. B.][Kapitel 11] äquivalent sind:

Der Zweistichproben-t-Test vergleicht die Mittelwerte der beiden Gruppen, bezeichnet mit μ_0 und μ_1. Daher kann man die beiden Parameter der Einfachregression auf diese beiden Mittelwerte abbilden, nämlich

$$\beta_0 = \mu_0$$

und

$$\beta_1 = \mu_1 - \mu_0$$

Für die Regressionskoeffizienten stimmten der ML-Schätzer und die KQ-Schätzer überein. Es ergibt sich also in diesem Fall

$$\hat{\beta}_0 = \hat{\mu}_0 = \bar{Y}_0$$

und

$$\hat{\beta}_1 = \hat{\mu}_1 - \hat{\mu}_0 = \bar{Y}_1 - \bar{Y}_0,$$

wobei $\bar{Y}_j$, $j \in \{0, 1\}$ den Mittelwert der Outcome-Variable in der Gruppe $x = j$ bezeichnet.

Der Schätzer für $\hat{\sigma}$ ist[26] (vgl. (15))

[26]Dieser entspricht nicht dem ML-Schätzer, der ML-Schätzer ist mit $\frac{1}{n}$ normiert.

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\left(\sum_{k=1}^{n_0}(Y_{k,0} - \bar{Y}_0)^2 + \sum_{k=1}^{n_1}(Y_{k,1} - \bar{Y}_1)^2\right)$$

Es bleibt noch zu zeigen, dass die Teststatistik im Zweistichproben-t-Test mit der Teststatistik in der Einfachregression für β_1 identisch ist (vgl. (20)).

Durch Umformungen sieht man, dass

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n_0 \cdot n_1}{n}.$$

Damit ergibt sich, dass

$$\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{(\bar{Y}_1 - \bar{Y}_0)}{\hat{\sigma} \cdot \sqrt{\frac{n_0+n_1}{n_0 \cdot n_1}}},$$

wobei $\hat{\sigma} := \sqrt{\hat{\sigma}^2}$. Diese Teststatistik ist die Teststatistik des Zweistichproben-t-Tests. Sie ist t-verteilt mit $n-2$ Freiheitsgraden und äquivalent zur Teststatistik des Regressionskoeffizienten β_1 in der Einfachregression. Daher liefern z. B. beide Herangehensweisen, also der Zweistichproben-t-Test und die Einfachregression, die selben p-Werte.

Literatur

Anderson, Andrew A.: Assessing Statistical Results: Magnitude, Precision, and Model Uncertainty. In: *The American Statistician* 73 (2019), Nr. sup1, S. 118–121. – URL https://doi.org/10.1080/00031305.2018.1537889

Arel-Bundock, Vincent ; Greifer, Noah ; Heiss, Andrew: How to interpret statistical models using marginaleffects for R and Python. In: *Journal of Statistical Software* 111 (2024), S. 1–32

Barnard, George A.: The use of the likelihood function in statistical practice. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* Bd. 1 Univ of California Press (Veranst.), 1967, S. 27–40

Carlin, John B. ; Moreno-Betancur, Margarita: On the uses and abuses of regression models: a call for reform of statistical practice and teaching. In: *Statistics in Medicine* 44 (2025), Nr. 13-14, S. e10244

Dalpiaz, David: *Applied statistics with R.* 2021

Fahrmeir, Ludwig ; Heumann, Christian ; Künstler, Rita ; Pigeot, Iris ; Tutz, Gerhard: *Statistik: Der Weg zur Datenanalyse.* 8. Aufl. 2016. Berlin, Heidelberg : Springer Berlin Heidelberg, 2016 (Springer-Lehrbuch). – URL http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1576240. – ISBN 9783662503720

Fahrmeir, Ludwig ; Kneib, Thomas ; Lang, Stefan: *Regression - Modelle, Methoden und Anwendungen.* Berlin : Springer Verlag, 2009

Frøslie, Kathrine F. ; Røislien, Jo: Sprechen sie statistik? In: *Tidsskrift for Den norske legeforening* (2019)

Gelman, Andrew ; Geurts, Hilde M.: The statistical crisis in science: how is it relevant to clinical neuropsychology? In: *The Clinical Neuropsychologist* 31 (2017), Nr. 6-7, S. 1000–1014. – URL https://doi.org/10.1080/13854046.2016.1277557. – PMID: 28075223

Gelman, Andrew ; Hill, Jennifer: *Data analysis using regression and multilevel/hierarchical models*. Cambridge and New York : Cambridge University Press, 2007 (Analytical methods for social research). – ISBN 978-0521686891

Gelman, Andrew ; Hill, Jennifer ; Vehtari, Aki: *Regression and Other Stories*. Cambridge University Press, 2020 (Analytical Methods for Social Research)

Gelman, Andrew ; Loken, E.: The statistical crisis in science. In: *American Scientist* 102 (2014), S. 460

Haig, Brian D.: *The Philosophy of Quantitative Methods: Understanding Statistics*. Oup Usa, 2018

Heinze, Georg ; Baillie, Mark ; Lusa, Lara ; Sauerbrei, Willi ; Schmidt, Carsten O. ; Harrell, Frank E. ; Huebner, Marianne ; TG2 ; STRATOS initiative, TG3 of the: Regression without regrets–initial data analysis is a prerequisite for multivariable regression. In: *BMC Medical Research Methodology* 24 (2024), Nr. 1, S. 178

Heinze, Georg ; Boulesteix, Anne-Laure ; Kammer, Michael ; Morris, Tim P. ; White, Ian R. ; STRATOS initiative, Simulation P. of the: Phases of methodological research in biostatistics—building the evidence base for new methods. In: *Biometrical Journal* 66 (2024), Nr. 1, S. 2200222

Hoeting, Jennifer A. ; Madigan, David ; Raftery, Adrian E. ; Volinsky, Chris T.: Bayesian Model Averaging: A Tutorial. In: *Statistical Science* 14 (1999), Nr. 4, S. 382–401. – URL `http://www.jstor.org/stable/2676803`. – Zugriffsdatum: 2022-04-17. – ISSN 08834237

Johnson, Alicia A. ; Ott, Miles Q. ; Dogucu, Mine: *Bayes rules!: An introduction to applied Bayesian modeling*. Chapman and Hall/CRC, 2022

Kaplan, David: *Bayesian statistics for the social sciences*. Guilford Publications, 2023

Kennedy-Shaffer, Lee: Before p¡ 0.05 to beyond p¡ 0.05: using history to contextualize p-values and significance testing. In: *The American Statistician* 73 (2019), Nr. sup1, S. 82–90

Kennedy-Shaffer, Lee: Teaching the difficult past of statistics to improve the future. In: *Journal of Statistics and Data Science Education* 32 (2024), Nr. 1, S. 108–119

Kerr, Norbert L.: HARKing: Hypothesizing after the results are known. In: *Personality and social psychology review* 2 (1998), Nr. 3, S. 196–217

Lin, Hanti: To be a frequentist or Bayesian? Five positions in a spectrum. In: *Harvard Data Science Review* 6 (2024), Nr. 3

Meinfelder, Florian ; Kluge, Rebekka: *Bad Science: Die dunkle Seite der Statistik*. Vahlen, 01 2020. – ISBN 9783800660292

Murdoch, Duncan J. ; Tsai, Yu-Ling ; Adcock, James: P-Values are Random Variables. In: *The American Statistician* 62 (2008), Nr. 3, S. 242–245. – URL https://doi.org/10.1198/000313008X332421

Nold, Mariana ; Heinze, Georg: Commentary: Teaching Statistics as Minor Subject—Handing on Fire, Not Worshipping Ashes. In: *Statistics in Medicine* 44 (2025), Nr. 13-14, S. e10284

Nuijten, Michèle B: Assessing and improving robustness of psychological research findings in four steps. In: *Avoiding questionable research practices in applied psychology*. Springer, 2022, S. 379–400

Peels, Rik ; Bouter, Lex: Replication and trustworthiness. In: *Accountability in Research* 30 (2023), Nr. 2, S. 77–87. – URL https://doi.org/10.1080/08989621.2021.1963708. – PMID: 34346793

Pruscha, Helmut: *Vorlesungen über Mathematische Statistik*. Springer VS, 01 2000. – ISBN 978-3-519-02393-7

Scheid, S. ; Vogl, S.: *Data Science: Grundlagen, Methode und Modelle der Statistik*. Carl Hanser Verlag GmbH & Company KG, 2021. – URL https://books.google.de/books?id=qjA7EAAAQBAJ. – ISBN 9783446470019

Silver, Nate: *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012

Spiegelhalter, David: *The art of statistics: Learning from data*. Penguin UK, 2019

Sutherland, Chris ; Hare, Darragh ; Johnson, Paul J. ; Linden, Daniel W. ; Montgomery, Robert A. ; Droge, Egil: Practical advice on variable selection and reporting using Akaike information criterion. In: *Proceedings of the Royal Society B* 290 (2023), Nr. 2007, S. 20231261

Van Calster, Ben ; Wynants, Laure ; Riley, Richard D. ; van Smeden, Maarten ; Collins, Gary S.: Methodology over metrics: current scientific standards are a disservice to patients and society. In: *Journal of Clinical Epidemiology* 138 (2021), S. 219–226. – URL https://www.sciencedirect.com/science/article/pii/S0895435621001700. – ISSN 0895-4356

van Ravenzwaaij, D ; Bakker, M ; Heesen, R ; Romero, F ; van Dongen, N ; Crüwell, S ; Field, S M ; Held, L ; Munafò, M R ; Pittelkow, M M ; Tiokhin, L ; Traag, V A ; van den Akker, O R ; van 't Veer, A E ; Wagenmakers, E J: Perspectives on scientific error. In: *Royal Society Open Science* 10 (2023), Juli, Nr. 7. – ISSN 2054-5703

Wasserstein, Ronald L. ; Schirm, Allen L. ; Lazar, Nicole A.: Moving to a World Beyond „$p < 0{,}05$". In: *The American Statistician* 73 (2019), Nr. sup1, S. 1–19. – URL https://doi.org/10.1080/00031305.2019.1583913