Univ.-Prof. Dr. Robinson Kruse-Becher
Dr. Pascal Goemans

32491
Angewandte Datenanalyse
Leseprobe

Einheit 1: Lineare Regression

Fakultät für
Wirtschafts-
wissenschaft

FernUniversität in Hagen

Der Inhalt dieses Studienbriefs wird gedruckt auf Recyclingpapier (80 g/m2, weiß), hergestellt aus 100 % Altpapier.

Angewandte Datenanalyse - Modul 32491

Einführung

Prof. Dr. Robinson Kruse-Becher
Dr. Pascal Goemans

FernUniversität in Hagen
Fakultät für Wirtschaftswissenschaft
Lehrstuhl für Angewandte Statistik

Gegenstand des Kurses und Lernziele

- Die Modellierung und Prognose relevanter Variablen aus verschiedenen Bereichen der Wirtschaftswissenschaften steht im Vordergrund.
- Sie erlernen
  - mit Hilfe geeigneter Methoden Prognosen für stetige und qualitative Variablen zu bilden und die Prognosequalität zu beurteilen.
  - wie Sie wirtschaftswissenschaftliche Fragestellungen sachgerecht in ein ökonometrisches Modell überführen.
  - die geeigneten Daten auszuwählen.
  - die statistischen Schätz- sowie Testverfahren hinsichtlich ihrer Angemessenheit für die jeweilige Fragestellung zu beurteilen.
  - Analysen sowie Ergebnisse adressatengerecht aufzubereiten.
  - empirische Befunde kritisch zu reflektieren.
- Vorkenntnisse in Mathematik und Statistik auf dem Niveau des Moduls 32741 "Vertiefung der Wirtschaftsmathematik und Statistik" sind erforderlich.

Thematik

- Behandelte Themen:
  - Statistisches Lernen
  - Überwachtes statistisches Lernen: Lineare Regressionsmodelle, Klassifikationsmodelle, Kreuzvalidierung, Modellauswahl
  - Unüberwachtes statistisches Lernen: Hauptkomponentenanalyse und Clustering
- Ausblick auf Erweiterungen; mögliche Themen für Seminar- und Abschlussarbeiten
- Der Kurs behandelt Themen und ist Grundlage für komplexere Methoden aus den Bereichen: Statistisches Lernen, Maschinelles Lernen, Datenwissenschaft (Data Science), Multivariate Verfahren
- Im Studiengang Master Wirtschaftswissenschaft baut der Kurs auf den Inhalten des Moduls "Vertiefung der Wirtschaftsmathematik und Statistik" auf und bereitet auf die Module "Angewandte Ökonometrie" und "Zeitreihenökonometrie" vor.

Ansatz und Konzeption

- Es ist wichtig, die Ideen hinter den verschiedenen Methoden zu verstehen, um zu wissen, wie und wann sie eingesetzt werden können.
- Es ist wichtig, zuerst die einfachen Methoden zu verstehen, um die komplizierteren zu beherrschen.
- Es ist wichtig, die Leistungsfähigkeit einer Methode einschätzen zu können (einfachere Methoden sind häufig ähnlich gut wie ausgefeiltere).
- Die Darstellung der Methoden ist so formal wie nötig. Im Vordergrund steht die empirische Modellierung.
- Es handelt sich um ein aktuelles Gebiet mit Anwendungen in verschiedenen Bereichen der Wirtschaftswissenschaften wie der Bank- und Finanzwirtschaft, dem Marketing, dem Controlling, der Wirtschaftsprüfung, der Wirtschaftspsychologie sowie der Mikro- und Makroökonomie.

Lehr- und Lernformate

- Lehr- und Lernformat: Lehrtext als Grundlage, Videovorlesungen (asynchron) und synchrone Übungsveranstaltungen
- Grundlage des Kurses ist der Lehrtext "An Introduction to Statistical Learning: with Applications in R" (Second Edition, Springer Texts in Statistics) von Gareth James, Daniela Witten, Trevor Hastie und Robert Tibshirani. Den Lehrtext stellen die Autoren auf `https://www.statlearning.com/` kostenfrei zur Verfügung.
- Es werden zu den relevanten Kapiteln R-Labs zur Verfügung gestellt, in denen Methoden anhand von Beispieldaten angewandt und verglichen werden.
- R ist eine führende open-source Statistiksoftware, die mit RStudio (IDE) kombinierbar ist.

Inhaltsverzeichnis

Die Kapitelangaben beziehen sich auf die zugehörigen Kapitel in "An Introduction to Statistical Learning: with Applications in R" (Second Edition, Springer Texts in Statistics).

Überwachtes Statistisches Lernen

- Messwert für abhängige Variable Y (auch Ergebnis oder Zielvariable)
- Vektor von Messwerten für die p Prädiktoren X (auch erklärende oder unabhängige Variablen, Regressoren, Features)
- Beim *Regressionsproblem* ist Y quantitativ (z. B. Preis oder Einkommen in €).
- Beim *Klassifizierungsproblem* nimmt Y qualitative Werte in einer endlichen ungeordneten oder geordneten Menge an (z. B. Kredit wurde oder wurde nicht zurückgezahlt, es handelt sich um ein Entwicklungs-, Schwellen- oder Industrieland).
- Wir verfügen über Trainingsdaten $(x_1, y_1), \ldots, (x_N, y_N)$. Dies sind Beobachtungen für die Messungen von X und Y.

Ziele des überwachten statistischen Lernens

Auf Basis der Trainingsdaten wollen wir Modelle trainieren, um

- genaue Prognosen für Testdaten zu erstellen, die das Modell vorher nicht gesehen hat.
- zu verstehen, welche Prädiktoren die abhängige Variable beeinflussen und wie.
- die Qualität der Prognosen und Schlussfolgerungen zu bewerten.

Unüberwachtes statistisches Lernen und Ziele

- Wir verfügen über keine abhängige Variable, sondern lediglich über eine Reihe von Prädiktoren (Features), die anhand einer Stichprobe ermittelt wurden.
- Das Ziel ist es, relevante Strukturen in den Daten zu entdecken und valide Schlüsse zu ziehen:
  - Welche informativen Möglichkeiten gibt es, die Daten zu visualisieren?
  - Können wir Abhängigkeiten und Cluster in den Beobachtungen entdecken?
  - Inwiefern lassen sich Informationen in hochdimensionalen Datenstrukturen auf das Wesentliche reduzieren?
- Wir behandeln zwei Methoden:
  - *Hauptkomponentenanalyse*: Ein Werkzeug für die Datenvisualisierung oder Daten-Vorverarbeitung bevor Methoden aus dem überwachten statistischen Lernen angewendet werden
  - *Clustering*: Eine breite Klasse von Methoden zur Entdeckng unbekannter Untergruppen in den Daten

Angewandte Datenanalyse - Modul 32491

Lineare Regression

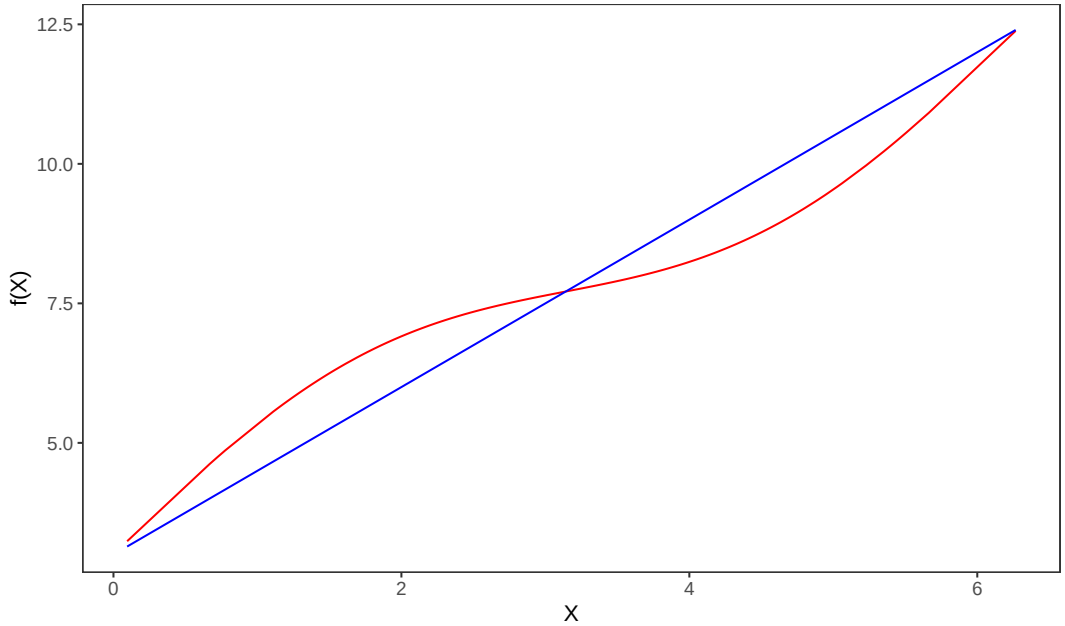
Prof. Dr. Robinson Kruse-Becher
Dr. Pascal Goemans

FernUniversität in Hagen
Fakultät für Wirtschaftswissenschaft
Lehrstuhl für Angewandte Statistik

Wir bedanken uns bei Herrn Gabriel Preuß für vielfältige Unterstützung.

Lineare Regression

- Die lineare Regression ist ein vereinfachender Ansatz für das überwachte Lernen. Dabei nimmt man an, dass der Zielwert Y ausschließlich linear von den Eingabewerten $X_1, X_2, \ldots, X_p$ abhängt.
- Aber: Wahre Regressionsfunktionen sind nie linear!



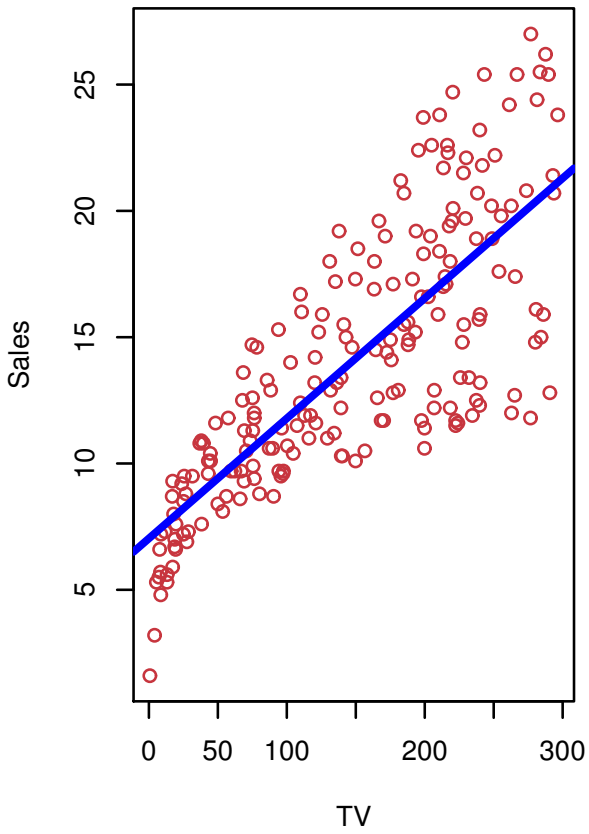
- Auch wenn es stark vereinfachend wirkt, ist die lineare Regression extrem nützlich, sowohl in konzeptueller als auch in praktischer Hinsicht.

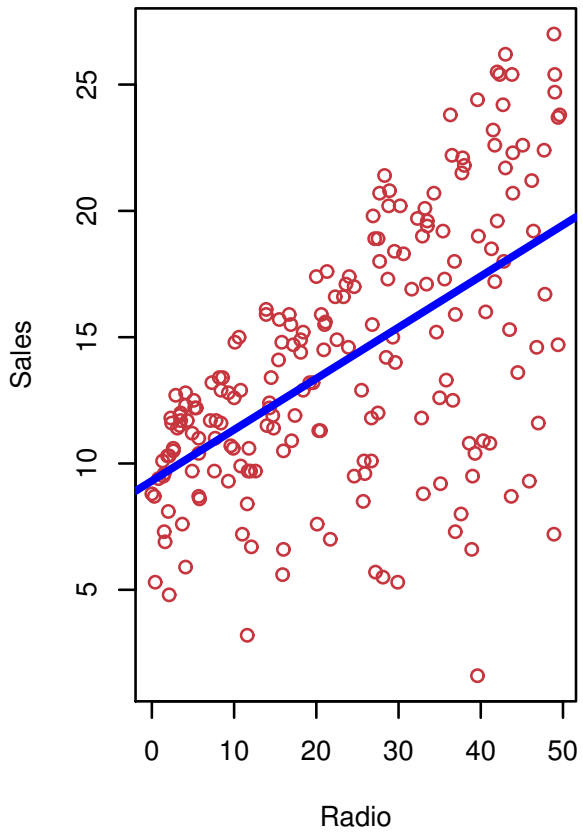
Lineare Regression mit Werbedaten

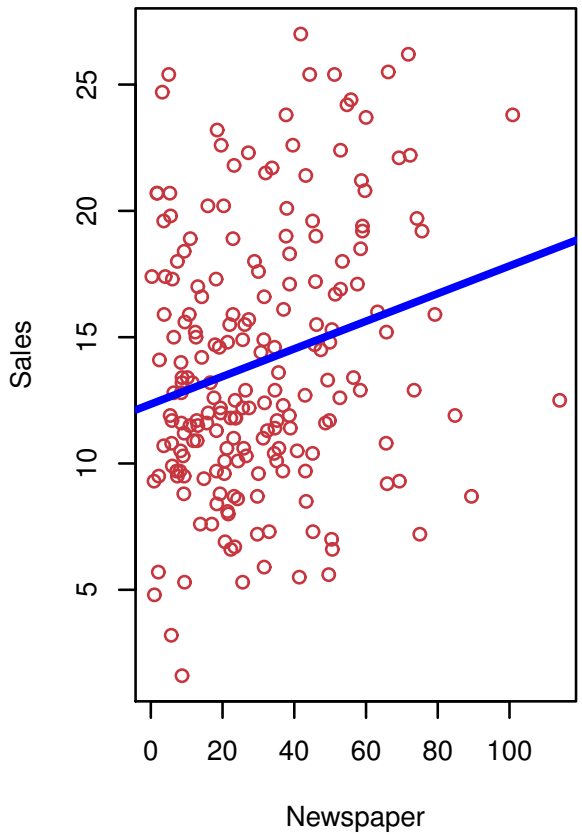
Bezüglich der Werbedaten auf der nächsten Folie könnten wir folgende Fragestellungen untersuchen:

- Gibt es einen Zusammenhang zwischen dem Umsatz und dem Werbebudget?
- Falls ja, wie stark ist dieser?
- Welches Medium wirkt sich am stärksten auf den Umsatz aus?
- Wie genau kann man den zukünftigen Umsatz vorhersagen?
- Gibt es Synergieeffekte zwischen den einzelnen Medien?

Werbedaten


Sales
5
10
15
20
25
0
50
100
200
300
TV


Sales
5
10
15
20
25
0
10
20
30
40
50
Radio


Sales
5
10
15
20
25
0
20
40
60
80
100
Newspaper

Lineare Regression mit einem einzelnen Prädiktor X

Wir unterstellen ein Modell der Form

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Sprachregelungen:

Y	X
abhängige Variable	unabhängige Variable
erklärte Variable	erklärende Variable
Regressand	Regressor
Regressand	Prädiktor
Effekt	Ursache
Zielgröße	Kontrollvariable

Hierin sind der *Intercept/Achsenabschnitt* β_0 und der *Steigungsparameter* β_1 zwei unbekannte Konstanten, die allgemeiner auch *Koeffizienten* oder *Parameter* genannt werden, und ε ist der unbekannte *Fehlerterm*.

Schätzung der Parameter mithilfe der KQM

- Haben wir zwei Schätzer $\widehat{\beta}_0$ und $\widehat{\beta}_1$ für unsere Parameter, können wir zukünftige Werte durch

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

voraussagen, wobei $\widehat{y}$ die Prognose von Y unter der Bedingung $X = x$ angibt. Mit dem Symbol $\widehat{\ }$ werden geschätzte Werte dargestellt.
- Sei $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ die Vorhersage von Y basierend auf dem i-ten Wert von X. Dann ist $e_i = y - \widehat{y}$ die i-te Abweichung (*Residuum*).
- Wir definieren die *Residuenquadratsumme* (RSS) als

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

oder äquivalent dazu

$$RSS = (y_1 - \widehat{\beta}_0 - \widehat{\beta}_1 x_1)^2 + (y_2 - \widehat{\beta}_0 - \widehat{\beta}_1 x_2)^2 + \ldots + (y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n)^2.$$

- Bei der *Kleinst-Quadrate-Methode (KQM)* sucht man nun diejenigen Parameter, welche die Residuenquadratsumme RSS minimieren.

Schätzung der Parameter mithilfe der KQM

Es kann gezeigt werden, dass dies der Fall ist für

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

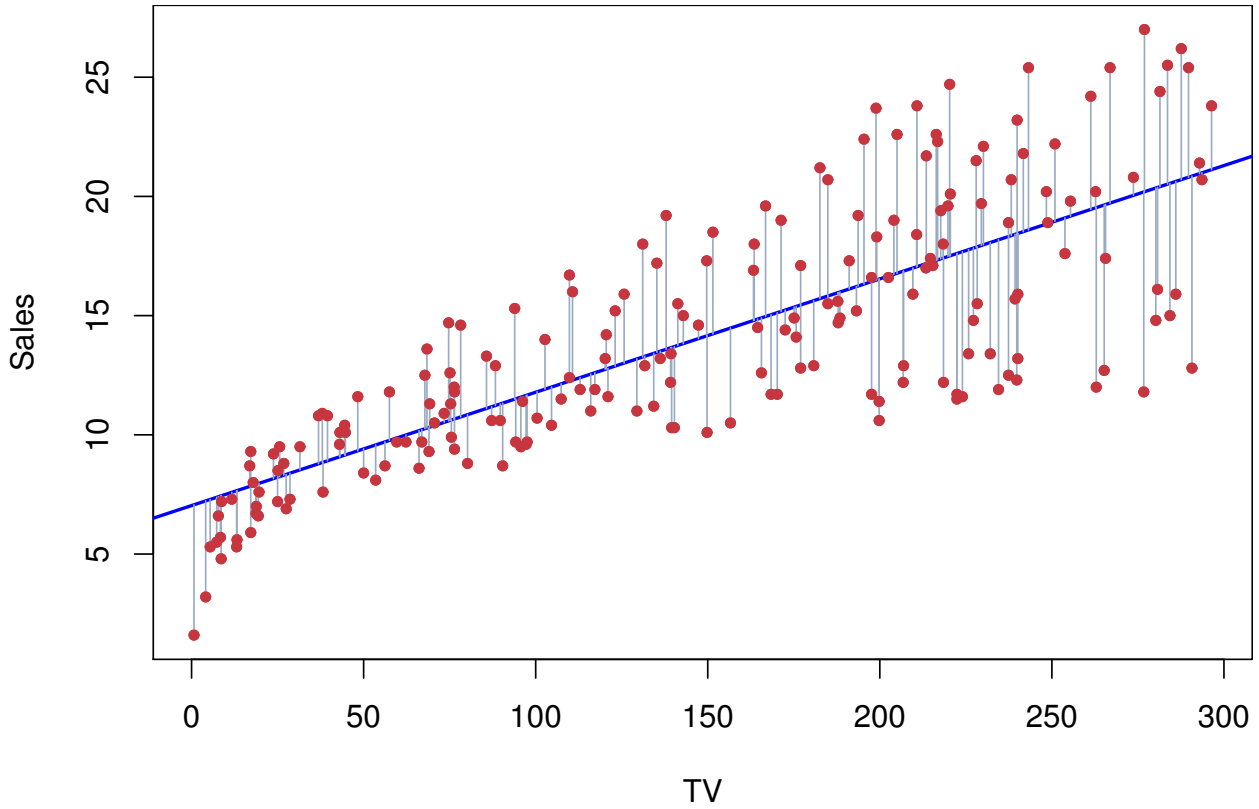
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

mit den jeweiligen arithmetischen Mittelwerten

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ und}$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Beispiel Werbedaten



Methode der kleinsten Quadrate für die Regression von Umsatz auf Fernsehen. In diesem Fall erfasst ein lineares Modell den wesentlichen Zusammenhang, obwohl es auf der linken Seite des Diagramms etwas abweicht.

Genauigkeit der Koeffizientenschätzung

Der *Standardfehler (SE, standard error)* eines Schätzers ist ein Maß dafür, wie stark dieser bei wiederholten Stichprobenziehungen schwankt:

$$\mathsf{SE}(\widehat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]$$

$$\mathsf{SE}(\widehat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\text{mit } \sigma^2 = \mathsf{Var}(\varepsilon)$$

Der Standardfehler kann verwendet werden, um *Konfidenzintervalle* zu berechnen.

Konfidenzintervalle

Ein 95%-Konfidenzintervall wird als ein Bereich von Werten definiert, sodass dieses Intervall einer Wahrscheinlichkeit von 95% den wahren unbekannten Wert des Parameters enthält. Es hat die Form

$$\widehat{\beta}_1 \pm 2 \cdot \mathsf{SE}(\widehat{\beta}_1).$$

Das bedeutet, dass das Intervall

$$\left[\widehat{\beta}_1 - 2 \cdot \mathsf{SE}(\widehat{\beta}_1), \widehat{\beta}_1 + 2 \cdot \mathsf{SE}(\widehat{\beta}_1)\right]$$

den wahren Wert von β_1 mit einer 95%igen Wahrscheinlichkeit enthält (unter der Annahme, einer wiederholten Stichprobenziehung).
Für die Werbedaten erhalten wir für β_1 das 95%-Konfidenzintervall $[0.042; 0.053]$

Hypothesentests

Standardfehler können ebenfalls verwendet werden, um *Hypothesentests* zu den Koeffizienten durchzuführen.
Der häufigste Hypothesentest prüft die *Nullhypothese*

$$H_0\text{: Es gibt keinen Zusammenhang zwischen } X \text{ und } Y$$

gegen die *Alternativhypothese*

$$H_A\text{: Es gibt einen Zusammenhang zwischen } X \text{ und } Y.$$

Mathematisch entspricht dies der Überprüfung von

$$H_0\text{: } \beta_1 = 0$$

gegen die Alternativhypothese

$$H_A\text{: } \beta_1 \neq 0$$

da sich das Modell mit $\beta_1 = 0$ zu $Y = \beta_0 + \varepsilon$ vereinfacht und X dann keinen Einfluss auf Y mehr hat.

Hypothesentests

Um die Nullhypothese zu testen, berechnen wir die *t-Statistik* als

$$t = \frac{\widehat{\beta}_1 - 0}{\mathsf{SE}(\widehat{\beta}_1)}.$$

Diese ist unter der Annahme $\beta_1 = 0$ t-verteilt mit $n - 2$ Freiheitsgeraden.

Mit Hilfe von Statistikprogrammen lässt sich leicht die Wahrscheinlichkeit berechnen, dass ein zufälliger Wert größer oder gleich $|t|$ ist. Diese Wahrscheinlichkeit nennen wir den *p–Wert*.

Lineare Regression mit Werbedaten

	Koeffizient	Std.Abw.	t-Statistik	p-Wert
Intercept	7.0325	0.4578	15.36	0.0000
Fernsehen	0.0475	0.0027	17.67	0.0000

- Der Intercept zeigt an, dass der erwartete Umsatz auch ohne Werbeausgaben im Fernsehen positiv und signfikant von Null verschieden wären (siehe t-Statistik und zugehöriger p-Wert).
- Zusätzliche Werbeausgaben im Fernsehen haben jedoch einen positiven Effekt auf den erwarteten Umsatz. Dieser Effekt ist zudem statistisch signifikant (siehe t-Statistik und zugehöriger p-Wert).

Anpassungsgüte des Modells

- Zuerst standardisieren wir die Residuenquadratsumme RSS und erhalten den Abweichungsfehler RSE (Residual Standard Error)

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}.$$

- R^2 oder das *Bestimmtheitsmaß*:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \text{ mit } TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Anpassungsgüte des Modells

- Im Fall einer linearen Einfachregression (ein Prädiktor) stimmt der Wert von R^2 mit der quadrierten Korrelation von X und Y überein, d.h. es gilt $R^2 = r^2$ mit

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Multivariate Lineare Regression

Nun betrachten wir ein Modell der Form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon.$$

Wir interpretieren die jeweiligen β_j als die *durchschnittliche Auswirkung* auf Y bei einem Anstieg von X_j um eine Einheit, wobei *alle anderen Regressoren gleich bleiben* (ceteris paribus Annahme).

Beispielsweise könnte ein multivariates Regressionsmodell mit Werbedaten wie folgt lauten

$$\text{Umsatz} = \beta_0 + \beta_1 \cdot \text{Fernsehen} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{Zeitung} + \varepsilon.$$

Interpretation der Regressionskoeffizienten

Das beste Szenario wäre, wenn die Prädiktoren nicht miteinander korreliert sind.

Ein ausgewogenes Design:

- Jeder Koeffizient kann separat geschätzt und getestet werden.
- Interpretationen der Form '*eine Einheit Veränderung in X_j ist mit einer β_j-Veränderung in Y verbunden, während alle anderen Variablen konstant bleiben*', sind möglich.

Interpretation der Regressionskoeffizienten

Korrelationen zwischen den Prädiktoren verursachen Probleme:

- Die Varianz aller Koeffizienten neigt dazu, manchmal dramatisch, zu steigen.
- Interpretationen werden riskant – wenn sich X_j ändert, ändert sich alles andere ebenfalls.

Behauptungen über Kausalität sollten in Studien mit Beobachtungsdaten vermieden werden.

Probleme der Regressionskoeffizienten

"Data Analysis and Regression" Mosteller and Tukey 1977

- *Die Annahme, dass alle anderen Variablen konstant bleiben, ist normalerweise nicht erfüllt.* Prädiktoren ändern sich normalerweise gemeinsam!
- Beispiel: Geld im Portmonee (Y) in Abhängigkeit von der Anzahl der Münzen (X_1) und Anzahl 1€ Münzen (X_2). Isoliert betrachtet wird der Koeffizient von X_2 größer als Null sein, aber was ist, wenn X_1 auch Teil des Modells ist?
- Y ist der Stundenlohn einer Person in €; E und A sind die Berufserfahrung und das Alter in Jahren. Das angepasste Regressionsmodell lautet

$$\widehat{Y} = \beta_0 + 0.3 \cdot E - 0.1 \cdot A.$$

Wie interpretieren Sie $\widehat{\beta}_2 < 0$?

Zwei Zitate berühmter Statistiker

"Essentially, all models are wrong, but some are useful"
George Box

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"
Fred Mosteller and John Tukey, paraphrasing George Box

Schätzungen und Prognosen der multivariaten linearen Regression

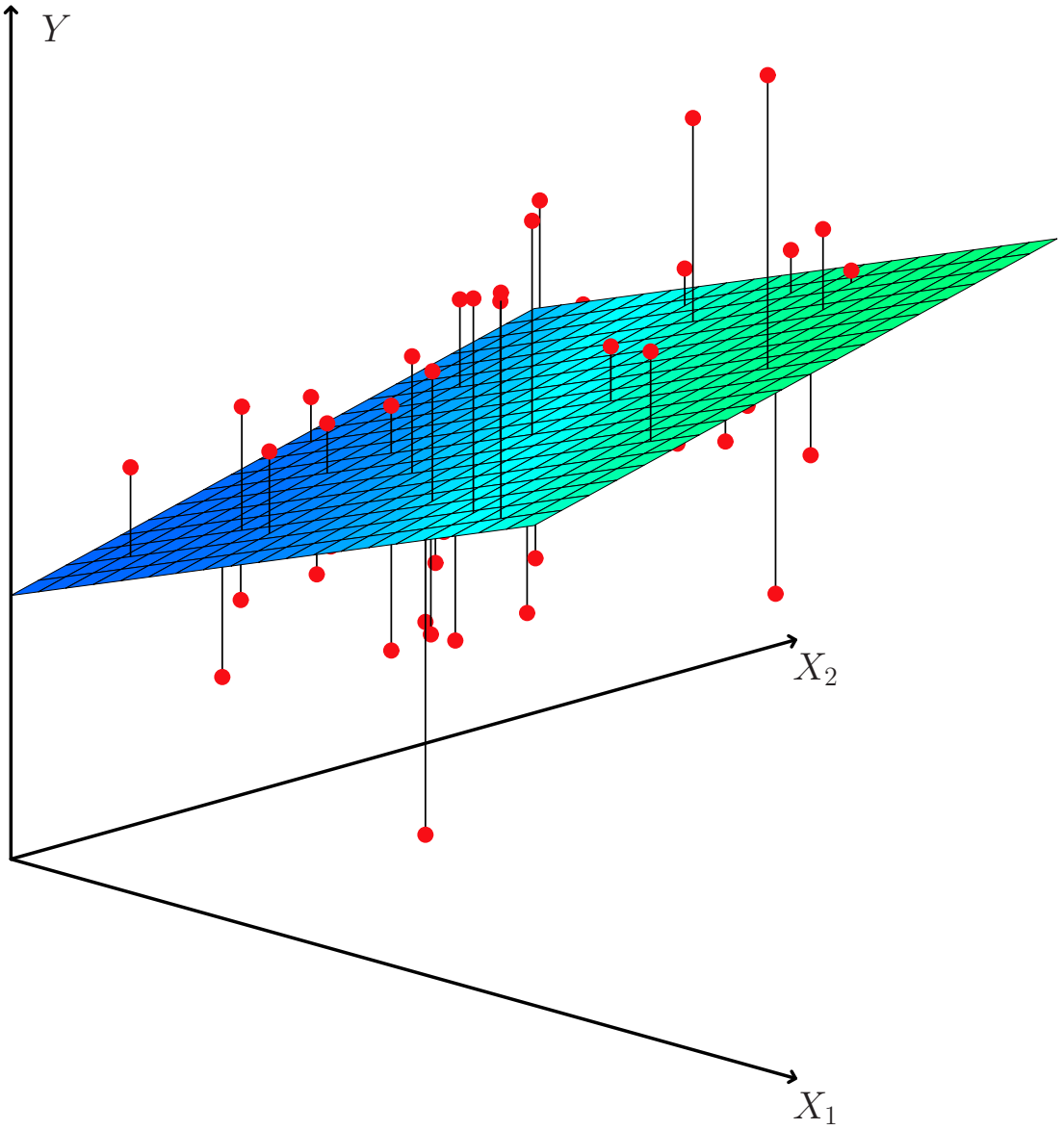
Für gegebene Schätzungen $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ können wir die folgende Gleichung für Prognosen verwenden

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p \ .$$

Wir schätzen $\beta_0, \beta_1, \ldots, \beta_p$ wiederum, indem wir die Residuenquadratsumme RSS minimieren:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip})^2. \end{aligned}$$

Dies macht man normalerweise mithilfe von Statistik-Programmen.


Y
X2
X1

Ergebnisse einer linearen Regression mit den Werbedaten

	Koeffizient	Std.Abw.	t-Statistik	p-Wert
Intercept	2.939	0.3119	9.42	0.0000
Fernsehen	0.046	0.0014	32.81	0.0000
Radio	0.189	0.0086	21.89	0.0000
Zeitung	-0.001	0.0059	-0.18	0.8599

Korrelationen:

	Fernsehen	Radio	Zeitung	Umsatz
Fernsehen	1.0000	0.0548	0.0567	0.7822
Radio		1.0000	0.3541	0.5762
Zeitung			1.0000	0.2283
Umsatz				1.0000

Einige wichtige Fragestellungen

1. Ist mindestens einer der Prädiktoren $X_1, X_2, \ldots, X_p$ nützlich für die Prognose von Y?
2. Tragen alle Prädiktoren dazu bei, Y zu erklären, oder ist nur eine Teilmenge der Prädiktoren nützlich?
3. Wie gut passt das Modell zu den Daten?
4. Wie lautet die optimale Prognose für die abhängige Variable Y für gegebene Werte der Prädiktoren $X_1, X_2, \ldots, X_p$ und wie genau ist diese Prognose?

Ist mindestens einer der Prädiktoren nützlich?

Hypothese: $H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0$
Unter H_0 gilt: $Y = \beta_0 + \varepsilon$, so dass $\widehat{Y} = \overline{Y}$ gilt.

Damit ergibt sich als *F-Statistik*

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F(p, n-p-1)$$
$$= \frac{n-p-1}{p} \frac{R^2}{1-R^2}.$$

Eine große Korrelation zwischen der Modellprognose $\widehat{Y}$ und den realisierten Werten Y (hohes R^2) führt also zu einem großen Wert der *F-Statistik* (Ablehnung von H_0).

ANOVA-Tafel

Dies lässt sich auch mit Hilfe der *ANOVA-Tafel* darstellen:

Varianz-Quelle	SQ		FG	F-Statistik
Hypothese	ESS	$(\widehat{Y}-\overline{Y})'(\widehat{Y}-\overline{Y})$	p	$F = \frac{\text{ESS}/p}{\text{RSS}/(n-p-1)}$
Residuen	RSS	$(Y-\widehat{Y})'(Y-\widehat{Y})$	$n-p-1$	$= \frac{n-p-1}{p}\frac{R^2}{1-R^2}$
Total	TSS	$(Y-\overline{Y})'(Y-\overline{Y})$	$n-1$	$\sim F(p, n-p-1)$

Variablenauswahl

- Der direkteste Ansatz wird als *"all subsets"* oder *"best subsets"* Regression bezeichnet: Hierbei wird die Methode der kleinsten Quadrate für alle möglichen Teilmengen berechnet, und dann wird anhand eines Kriteriums, welches Trainingsfehler und Modellgröße berücksichtigt, die "optimale" Teilmengen gewählt.
- Allerdings ist es oft nicht möglich, alle möglichen Modelle zu untersuchen, da es 2^p Modelle gibt. Für $p = 40$ gibt es beispielsweise schon über eine Milliarde Modelle!
- Stattdessen benötigen wir einen automatisierten Ansatz, der durch eine Teilmenge der Modelle sucht. Im Folgenden werden zwei häufig verwendete Ansätze diskutiert.

Schrittweise Vorwärtsauswahl

1. Beginne mit dem Nullmodell - einem Modell, das einen Intercept, aber keine Prädiktoren enthält.
2. Passe p einfache lineare Regressionen an und füge dem Nullmodell die Variable hinzu, die zu der niedrigsten Residuenquadratsumme (RSS) führt.
3. Füge diesem Modell die Variable hinzu, die zu der niedrigsten RSS unter allen Zwei-Variablen-Modellen führt.
4. Fahre fort, bis eine Abbruchbedingung erfüllt ist, zum Beispiel wenn alle verbleibenden Variablen einen p-Wert über einem bestimmten Schwellenwert haben.

Schrittweise Rückwärtsauswahl

1. Beginne mit allen Variablen im Modell.
2. Entferne die Variable mit dem größten p-Wert, das heißt die Variable, die statistisch am wenigsten signifikant ist.
3. Das neue $(p-1)$-Variablen-Modell wird angepasst, und die Variable mit dem größten p-Wert wird entfernt.
4. Fahre fort, bis eine Abbruchbedingung erreicht ist. Zum Beispiel kann man aufhören, wenn alle verbleibenden Variablen einen signifikanten p-Wert haben, der durch einen bestimmten Signifikanz-Schwellenwert definiert ist.

Modellauswahl - Fortsetzung

Später werden wir systematische Kriterien für die Auswahl eines optimalen Modells aus mehreren Modellspezifikationen kennenlernen.
Dazu gehören:

- *Mallow's C_p*
- das *Akaike-Informationskriterium (AIC)*
- das *Bayessche Informationskriterium (BIC)*
- das *adjustierte Bestimmtheitsmaß R^2*
- die *Kreuzvalidierung (CV)*.

Wir schauen uns jedoch bereits in diesem Kapitel das adjustierte R^2 genauer an.

Das Bestimmtheitsmaß R^2 und das adjustierte R^2

Das Bestimmtheitsmaß misst den Anteil der Stichprobenvarianz von Y der durch das Regressionsmodell erklärt bzw. prognostiziert wird

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Ein Nachteil ist jedoch, dass R^2 mit jedem zusätzlichen Regressor steigt, es sei denn der Regressionskoeffizient des hinzugefügten Regressors ist exakt 0. Aus diesem Grund steigt das R^2 automatisch mit der Modellgröße. Eine Möglichkeit ist es, das Bestimmtheitsmaß mit einen Strafterm zu korrigieren

$$R^2_{adj} = 1 - \frac{n-1}{n-(p+1)} \cdot \frac{RSS}{TSS}.$$

Das adjustierte Bestimmtheitsmaß R^2_{adj} steigt im Gegensatz zu R^2 nur, wenn der Absolutbetrag des t-Wertes für den hinzugefügten Koeffizienten über 1 liegt. Im Gegensatz zu R^2 kann R^2_{adj} auch negativ werden.