# 1 Einführung

In dieser Übung wenden wir die Lineare Regression auf den sogenannten *Auto-Datensatz* an. Dabei handelt es sich um die technischen Daten von 392 PKW-Modellen, die in den USA in den Jahren 1970 bis 1982 üblich waren. Die Daten sind der Webseite des Buches [1] entnommen[1].

Wir betrachten im Folgenden den Verbrauch als die Ausgangsvariable $y$ und die anderen Daten, nämlich Anzahl der Zylinder, Hubraum, Leistung, Gewicht, Beschleunigung und Baujahr als die Eingangsvariablen $x_i$. Der Datensatz enthält zusätzlich noch Informationen über die Herkunft der Autos, codiert als 1: USA, 2: Europa, 3: Asien. Da es sich bei diesen Daten um eine qualitative Variable handelt (da die 3 Werte willkürlich vergeben wurden und deshalb nicht geordnet sind, sprechen wir von einer Nominalskala), können wir sie nicht direkt in die Regression einbeziehen.

Ziel dieser Übung ist es, den Zusammenhang zwischen den Eingangsvariablen und dem Verbrauch der Autos zu ermitteln. Dazu ermitteln wir im Folgenden unterschiedliche Regressionsmodelle. Mit diesen Modellen versuchen wir dann Fragen zu beantworten, wie:

- Welche Eingangsvariable hat den größten Einfluss auf den Verbrauch? Leistung, Gewicht oder Beschleunigung?
- Mit welchem Modell lässt sich der Zusammenhang zwischen den Variablen am besten beschreiben?
- Wie lässt sich beurteilen, wie gut das Modell ist?
- Wie lässt sich mit dem Modell der Verbrauch eines Modells, dessen Verbrauch wir nicht kennen, vorhersagen?

# 2 Lineare Regression mit einer Eingangsvariable

Als Modell wird hier eine sogenannte *Regressionsgerade* verwendet, die für den Zusammenhang zwischen der Ausgangsvariable $y$ und der Eingangsvariable $x$ den folgenden Schätzwert verwendet:

$$\hat{y} = h(x) = \beta_0 + \beta_1 x \tag{1}$$

Wir nennen diesen Schätzwert auch die Hypothese $h(x)$. Die beiden Parameter des Modells sind der *Achsenabschnitt* (Englisch: *intercept*) $\beta_0$ und das *Regressionsgewicht* bzw. die *Steigung* (Englisch: *slope*) $\beta_1$. Diese beiden Parameter können mit Hilfe der *Methode der kleinsten Fehlerquadrate* aus den Daten bestimmt werden (überwachtes Lernen). Dazu wird die Kostenfunktion

$$C(\beta_0, \beta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left[ y^{(i)} - h(x^{(i)}) \right]^2 \tag{2}$$

bezüglich der Parameter $\beta_0$ und $\beta_1$ minimiert. Die Parameterwerte $\hat{\beta}_0$ und $\hat{\beta}_1$, welche zum Minimum der Kostenfunktion gehören, sind die gesuchten Schätzwerte für die Parameter. Das lässt sich folgendermaßen ausdrücken

$$\left( \hat{\beta}_0, \hat{\beta}_1 \right) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \; C(\beta_0, \beta_1) \tag{3}$$

Für dieses Optimierungsproblem ergeben sich die folgenden geschlossenen Lösungen

$$\hat{\beta}_0 = \overline{y} - \beta_1 \overline{x} \qquad\qquad \hat{\beta}_1 = \frac{\operatorname{cov}(y, x)}{var(x)} \tag{4}$$

[1] Ursprünglich stammen die Daten aus der StatLib Library, die von der Carnegie Mellon University in Pittsburgh betrieben wird. Die Daten wurden 1983 für die *American Statistical Association Exposition* verwendet.

**Durchführung**

### Aufgabe D1: Auto-Daten mit einer Eingangsvariable

In dieser Aufgabe greifen wir uns einige ausgewählte Eingangsvariablen heraus und führen dann jeweils eine lineare Regression mit einer Eingangsvariable durch.

a) Laden Sie die Daten aus dem Excel-File `Autos_DE.xlsx` in den MATLAB-Workspace. Dazu können Sie den Befehl `xlsread` benutzen.

b) Extrahieren Sie aus den Daten die Ausgangsvariable $y$ sowie die Eingangsvariablen $x_3$ (Leistung), $x_4$ (Gewicht) und $x_5$ (Beschleunigung).

c) Stellen Sie Streudiagramme (scatter plots) der Ausgangsvariable $y$ über der Eingangsvariable $x$ dar. Nehmen Sie für $x$ nacheinander die drei oben erwähnten Eingangsvariablen $x_3$, $x_4$ und $x_5$ (so erhalten wir 3 Streudiagramme).

d) Führen Sie für jede der drei oben genannten Eingangsvariablen eine lineare Regression mit einer Eingangsvariablen durch. Berechnen Sie dazu jeweils nach der Gleichung (4) die Parameter $\hat{\beta}_0$ und $\hat{\beta}_1$. Dadurch erhalten wir 3 Parametersätze: für jede der drei Eingangsvariablen einen Parametersatz. Sie können für die Berechnung der Kovarianz zwischen den Variablen $x$ und $y$ den Befehl **`cov`** verwenden und für die Berechnung der Varianz den Befehl `var`. Beachten Sie jedoch, dass **`cov`** nicht einfach die Kovarianz, sondern die $2 \times 2$-Kovarianzmatrix von $x$ und $y$ berechnet. Diese Matrix enthält auch den gesuchten Wert, den wir zur Berechnung von (4) brauchen.

e) Verifizieren Sie, ob Sie die Parameter einen plausiblen Eindruck machen, indem Sie jeweils die Regressionsgerade in die Streudiagramme einzeichnen.

f) Jetzt wollen wir etwas darüber nachdenken, was die Parameterwerte bedeuten. Was sagen uns die Werte $\hat{\beta}_0$ und $\hat{\beta}_1$ der Parameter über den Zusammenhang von Verbrauch und Leistung bzw. Verbrauch und Gewicht bzw. Verbrauch und Beschleunigung. Können wir aus den Werten herauslesen, welche der drei Eingangsvariablen $x_3$, $x_4$ und $x_5$ den größten Einfluss auf den Verbrauch hat?

g) Um zu Beurteilen, welche der drei Eingangsvariablen $x_3$, $x_4$ und $x_5$ besser zur Vorhersage der Ausgangsvariable $y$ geeignet ist, berechnen wir für jeder der drei Parametersätze $(\hat{\beta}_0, \hat{\beta}_1)$ die Kostenfunktion $C(\hat{\beta}_0, \hat{\beta}_1)$.

h) Jetzt greifen wir uns zufällig einen PKW-Typen heraus, z.B. Zeile 104 unserer Daten: Ford Maverick. Mit Hilfe der gefundenen Parameter prädizieren wir jetzt aus der entsprechenden Eingangsvariable die Ausgangsvariable (also den Verbrauch) nach Gleichung (1). Mit welcher der drei Eingangsvariablen $x_3$, $x_4$ und $x_5$ erhalten wir die beste Vorhersage? Liegt der tatsächliche Verbrauch jeweils höher oder niedriger als der vorhergesagte Verbrauch?

## 3 Lineare Regression mit mehreren Eingangsvariablen

Wenn wir mehr als eine Eingangsvariable benutzen ist es sinnvoll, alle Eingangsvariablen $x_1$ bis $x_n$ zu einen Vektor $\boldsymbol{x}$ zusammenzufassen. Um die Schreibweise zu vereinfachen ergänzen wir am Anfang des Vektors noch die Hilfsvariable $x_0 = 1$. Damit erhalten wir die folgende Definition des Vektors der Eingangsvariablen

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Auch die Parameter $\beta_0$ bis $\beta_n$ fassen wir zum sogenannten Parametervektor $\boldsymbol{\beta}$ zusammen, der wie folgt definiert ist

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

Jetzt können wir die Hypothese $h(\boldsymbol{x})$, als den Schätzwert für die Ausgangsvariable $y$ folgendermaßen schreiben:

$$\hat{y} = h(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \sum_{k=0}^{n} \beta_k x_k = \boldsymbol{\beta}^T \boldsymbol{x} \tag{5}$$

wobei $^T$ die Matrixtransponierung ausdrückt. Da (5) einen linearen Zusammenhang beschreibt, sprechen wir auch von einem linearen Modell.

Zur Bestimmung des Parametervektors $\boldsymbol{\beta}$ aus einem Satz von Trainingsdaten setzen wir wie bereits im Fall einer Eingangsvariable auch hier die *Methode der kleinsten Fehlerquadrate* ein. Die Trainingsdaten, die aus $m$ Datenpaaren für $\boldsymbol{x}$ und $y$ bestehen (also aus $m$ Punkten im $n+1$-dimensionalen Raum) lassen sich wie folgt darstellen

$$\begin{matrix} \left(\boldsymbol{x}^{(1)}, y^{(1)}\right) \\ \left(\boldsymbol{x}^{(2)}, y^{(2)}\right) \\ \vdots \\ \left(\boldsymbol{x}^{(m)}, y^{(m)}\right) \end{matrix} \tag{6}$$

wobei $\boldsymbol{x}^{(i)}$ und $y^{(i)}$ der Eingangsvektor und die Ausgangsvariable des $i$-ten Datenpunkts sind. Die Eingangvektoren aller $m$ Datenpunkte fassen wir zur $m \times (n+1)$ Datenmatrix $\boldsymbol{X}$ zusammen

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}^{(1)\,T} \\ \boldsymbol{x}^{(2)\,T} \\ \vdots \\ \boldsymbol{x}^{(m)\,T} \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{pmatrix} \tag{7}$$

und die Ausgangsvariablen zum Ausgangsvektor $\boldsymbol{y}$

$$\boldsymbol{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} \tag{8}$$

**Achtung**: Beachten Sie, dass $\boldsymbol{y}$ eine andere Dimension und eine andere Bedeutung hat als $\boldsymbol{x}$. Während $\boldsymbol{y}$ alle $m$ Beobachtungen der Ausgangsvariable $y$ zu einem Vektor zusammenfasst, fasst $\boldsymbol{x}$ alle $n+1$ Eingangsvariablen $x_0$ bis $x_n$ zu einem Vektor zusammen.

Die Abweichung zwischen $h(\boldsymbol{x})$ und $y$ wird als Fehlerterm bzw. Residuum $\varepsilon$ bezeichnet

$$\varepsilon = y - h(\boldsymbol{x}) \tag{9}$$

Wenn wir die Fehlerterme aller $m$ Beobachtungen zu einem Vektor $\boldsymbol{\varepsilon}$ zusammenfassen

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(m)} \end{pmatrix} \tag{10}$$

können wir die Kostenfunktion $C(\boldsymbol{\beta})$ folgendermaßen formulieren

$$C(\boldsymbol{\beta}) = \frac{1}{2m}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \frac{1}{2m}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \tag{11}$$

Die Minimierung von $C(\boldsymbol{\beta})$ bezüglich $\boldsymbol{\beta}$ liefert den gesuchten Schätzwert $\hat{\boldsymbol{\beta}}$ für den Parametervektor

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ C(\boldsymbol{\beta}) \tag{12}$$

Als geschlossene Lösung ergibt sich folgende Gleichung

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{13}$$

die auch als Normalengleichung bezeichnet wird. Hierbei ist $\boldsymbol{X}^T\boldsymbol{X}$ die empirische Korrelationsmatrix der Eingangsvariablen mit der Dimension $(n+1)\times(n+1)$ und $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ ist die zugehörige Moore-Penrose Pseudoinverse.

Um zu beurteilen, wie gut das Modell (5) mit den Parametern nach (13) zu den Daten passt, könnten wir einfach die Kostenfunktion (11) heranziehen. Je kleiner der Wert der Kostenfunktion ist, desto besser passt das Modell. Allerdings ist der absolute Wert, der sich für die Kostenfunktion $C(\hat{\boldsymbol{\beta}})$ ergibt nur schwer interpretierbar. Ein leichter zu interpretierender Wert ist das sogenannte *Bestimmtheitsmaß* $R^2$ (siehe z.B. [2, 3]), welches folgendermaßen definiert ist

$$R^2 = 1 - \frac{\operatorname{var}(\varepsilon)}{\operatorname{var}(y)} = \frac{\operatorname{var}(h(\boldsymbol{x}))}{\operatorname{var}(y)} \tag{14}$$

und nur Werte zwischen 0 und 1 annehmen kann

$$0 \leq R^2 \leq 1 \tag{15}$$

$R^2$ ist ein Maß dafür, welcher Anteil der Variation von $y$ durch das Regressionsmodell erklärt werden kann. Wenn $R^2 = 0$ ist, ist die Hypothese $h(\boldsymbol{x})$ keine bessere Vorhersage für $y$ als der Mittelwert $\bar{y}$. Die lineare Regression würde in diesem Fall also keinen Vorteil gegenüber eines gewöhnlichen Mittelwerts bringen. Wenn $R^2 = 1$ ist, liegen alle Datenpunkte exakt auf der Regressionsgerade. Die Hypothese stimmt in diesem Fall genau mit den tatsächlichen Daten überein. Die Ausgangsvariable $y$ lässt sich perfekt durch die Hypothese $h(\boldsymbol{x})$ vorhersagen. In einem solchen Fall, muss man sich aber die Frage stellen, ob eine Regressionsanalyse überhaupt notwendig ist. Wenn z. B. $R^2 = 0,85$ ist, dann lassen sich 85 % der Variation von $y$ durch die Hypothese nachbilden $h(\boldsymbol{x})$, 15 % bleiben unerklärt als Variation der Residuen zurück.

**Durchführung**

### Aufgabe D2: Auto-Daten mit mehreren Eingangsvariablen

In dieser Aufgabe führen wir die lineare Regression mit mehreren Eingangsvariablen am Beispiel der Auto-Daten durch. Vergleiche mit der linearen Regression mit einer Eingangsvariablen (Aufgabe D1) führen zu interessanten Einblicken.

a) Wie in der vorherigen Aufgabe benutzen wir die Daten aus dem Excel-File `Autos_DE.xlsx`.

b) Bilden Sie aus diesen Daten den Ausgangsvektor $\boldsymbol{y}$ sowie die Datenmatrix $\boldsymbol{X}$. Benutzen Sie zur Bildung von $\boldsymbol{X}$ zunächst nur die Variablen $x_0 = 1$ (Hilfsvariable), $x_3$ (Leistung), und $x_6$ (Baujahr). Damit ergibt sich für $\boldsymbol{X}$ die Dimension $392 \times 3$.

c) Führen Sie mit $\boldsymbol{y}$ und $\boldsymbol{X}$ eine lineare Regression durch. Berechnen Sie dazu den Schätzwert $\hat{\boldsymbol{\beta}}$ des Parametervektors nach Gleichung (13).

d) Verifizieren Sie, ob die Parameter einen plausiblen Eindruck machen, indem Sie jeweils die Regressionsgerade in das Streudiagramm einzeichnen.
**Hinweis**: Bei zwei Eingangsvariablen $x_3$ und $x_6$ ergeben sich eigentlich keine Regressionsgeraden, sondern eine Regressionsebene im dreidimensionalen Raum mit den Koordinaten $y, x_3, x_6$. Die grafi-

sche Darstellung wird deutlich einfacher, wenn wir statt des dreidimensionalen $yx_3x_6$-Raumes lieber dessen Projektionen auf die $yx_3$-Ebene bzw. auf die $yx_6$-Ebene zeichnen. In diesen Projektionen ergeben sich dann wieder Regressionsgeraden. Überlegen Sie, welchen Wert der Achsenabschnitt bei den Projektionen jeweils besitzt.

e) Vergleichen Sie das Regressionsgewicht, das sich in der Aufgabe D1 für die Eingangsvariable $x_3$ ergeben hat mit dem oben ermittelten Regressionsgewicht. Wie lässt sich die Abweichung erklären?

f) Bilden Sie jetzt die Datenmatrix $\boldsymbol{X}$ mit Hilfe aller vorhandenen Eingangsvariablen $x_0$ bis $x_6$. Dadurch ergibt sich jetzt eine Dimension von $392 \times 7$. Führen Sie dann die lineare Regression mit den gesamten Daten durch und interpretieren Sie die Regressionsgewichte. Wie ist es möglich dass das Regressionsgewicht für den Hubraum negativ wird?

g) Wir führen jetzt nacheinander die lineare Regression mit unterschiedlichen Variablenkombinationen durch und bestimmen jeweils das Bestimmtheitsmaß $R^2$. Beginnen Sie beispielsweise nur mit den beiden Variablen $x_0$ und $x_1$, nehmen Sie dann die Variable $x_2$ dazu und erhöhen Sie dann schrittweise die einbezogenen Variablen, bis alle Eingangsvariablen von $x_0$ bis $x_6$ am Regressionsmodell beteiligt sind. Wie ändert sich dabei $R^2$?

h) Führen Sie die lineare Regression nacheinander mit allen Einzelvariablen $x_1$ bis $x_6$ (jeweils unter Einbeziehung der Hilfsvariable $x_0 = 1$) durch und bestimmen Sie jeweils das Bestimmtheitsmaß $R^2$. Welche der Eingangsvariablen ist demnach am besten für die Vorhersage des Verbrauchs geeignet?

# 4 Gradientenverfahren

Das Gradientenverfahren (Englisch: gradient descent, method of steepest descent) ist eine iterative Methode zur Lösung von Optimierungsproblemen. Es ist vor allem für Probleme interessant, die sich nicht geschlossen lösen lassen, oder deren geschlossene Lösung sehr sehr viel Rechenaufwand verursacht. Man kann sich das Gradientenverfahren so vorstellen, dass man von einem beliebigen Startpunkt aus schrittweise immer in Richtung des steilsten Abstiegs geht, bis man den tiefsten Punkt einer Kostenfunktion erreicht hat.

Die Minimierung der Kostenfunktion für die lineare Regression lässt sich geschlossen lösen. Allerdings wird die geschlossene Lösung sehr aufwändig, wenn sehr viele Eingangsvariablen (sehr viele Merkmale) $x_i$ mit $i = 1 \ldots n$ verwendet werden. Wenn $n$ in die Größenordnung von $10^4$ kommt, wird die Invertierung der $(n+1) \times (n+1)$ Korrelationsmatrix $\boldsymbol{R} = \frac{1}{m}\boldsymbol{X}^T\boldsymbol{X}$ und damit die Berechnung der Normalengleichung (13) schwierig. Ab dann ist es besser iterative Verfahren, wie z..B. das Gradientenverfahren, einzusetzen, um den Parametervektor $\hat{\boldsymbol{\beta}}$ zu bestimmen.

Der Algorithmus für das Gradientenverfahren lautet:

**Algorithmus: Gradientenvervahren**

Setze $\boldsymbol{\beta}$ auf einen geeigneten Startwert, z.B. $\boldsymbol{\beta} = \mathbf{0}$.

Wiederhole bis zur Konvergenz

$$\boldsymbol{\beta} = \boldsymbol{\beta} - \alpha \nabla C$$

Der Wert von $\boldsymbol{\beta}$, der sich bei der Konvergenz des Verfahrens einstellt, ist der gesuchte Schätzwert $\hat{\boldsymbol{\beta}}$. Um die Konvergenz sicherzustellen, muss die Schrittweite $\alpha$ geeignet gewählt werden.

Bei der linearen Regression ergibt sich der Gradientenvektor $\nabla C$ wie folgt

$$\nabla C = \frac{1}{m}\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}^T\boldsymbol{y}\right) \tag{16}$$

wobei $\frac{1}{m}\boldsymbol{X}^T\boldsymbol{X}$ die Korrelationsmatrix der Eingangsvariablen und $\frac{1}{m}\boldsymbol{X}^T\boldsymbol{y}$ den Kreuzkorrelationsvektor von Ausgangsvariable und Eingangsvariablen darstellt.

Geeignete Werte für die Schrittweite $\alpha$ hängen von den Eingangsvariablen ab. Um Konvergenz des Verfahrens sicherzustellen muss $\alpha$ folgendermaßen gewählt werden

$$0 \leq \alpha \leq \frac{2}{\lambda_{\max}} \tag{17}$$

wobei $\lambda_{\max}$ der größte Eigenwert der Korrelationsmatrix $\boldsymbol{R} = \frac{1}{m}\boldsymbol{X}^T\boldsymbol{X}$ ist. Wenn wir ausreichend viele Datenpunkte zur Verfügung haben, ist

$$\alpha = \frac{1}{\lambda_{\max}} \tag{18}$$

eine gute Wahl. Wird $\alpha$ sehr klein gewählt, konvergiert das Verfahren zwar sicher, aber sehr langsam. Wird $\alpha$ relativ groß gewählt, aber noch unterhalb der oberen Grenze, dann oszillieren die Werte von $\boldsymbol{\beta}$ um den optimalen Wert und das Verfahren konvergiert nur langsam. Wird $\alpha$ größer als die obere Grenze gewählt, divergiert das Verfahren.

Im Allgemeinen ist die Konvergenzgeschwindigkeit des Gradientenverfahrens stark von der Korrelationsmatrix $\boldsymbol{R}$ abhängig. Wenn die Eigenwerte von $\boldsymbol{R}$ sehr unterschiedlich sind, konvergiert das Verfahren nur langsam. Wir sprechen in diesem Fall auch von einer großen Eigenwertstreuung (Englisch: eigen value spread). Bei geringer Eigenwertstreuung konvergiert das Verfahren relativ schnell.

Eine hohe Eigenwertstreuung ergibt sich, wenn

- die Eingangsvariablen deutlich unterschiedliche Wertebereiche aufweisen. Das ist bei unseren Daten der Fall. Während die Anzahl der Zylinder lediglich zwischen 3 und 8 schwankt, schwankt z.B. das Gewicht zwischen 732 kg und 2332 kg.
- die Eingangsvariablen stark korreliert sind. Auch das ist bei unserem Beispiel der Fall. So ist z.B. die Leistung und der Hubraum stark korreliert.

Eine einfache Maßnahme zur Reduzierung der Eigenwertstreuung ist die Skalierung der Eingangsvariablen, so dass die Wertebereiche aller Eingangsvariablen ähnlich sind. Es hat sich in der Praxis bewährt, jede Eingangsvariable $x_i$ folgendermaßen zu skalieren

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{\mathrm{std}(x_i)} \tag{19}$$

Die skalierte Variable $\tilde{x}_i$ ergibt sich also wenn wir von der ursprünglichen Eingangsvariable $x_i$ deren Mittelwert $\bar{x}_i$ abziehen und das Ergebnis durch die Standardabweichung $\mathrm{std}(x_i)$ teilen. Dadurch erreichen wir, dass die skalierte Variable $\tilde{x}_i$ einen Mittelwert von 0 und eine Standardabweichung von 1 hat.

Wenn wir mit den normierten Eingangsvariablen die lineare Regression durchführen, erhalten wir den Parametervektor $\tilde{\boldsymbol{\beta}}$, der zu den normierten Variablen gehört. Um daraus die Parameter für die ursprünglichen Eingangsvariablen zu erhalten, muss man folgende Umrechnungen durchführen

$$\beta_0 = \tilde{\beta}_0 - \sum_{i=1}^{n} \tilde{\beta}_i \, \frac{\bar{x}_i}{\mathrm{std}(x_i)} \tag{20}$$

$$\beta_i = \frac{\tilde{\beta}_i}{\mathrm{std}(x_i)} \qquad \forall i = 1 \ldots n \tag{21}$$

**Durchführung**

### Aufgabe D3: Gradientenverfahren

In dieser Aufgabe lösen wir das Optimierungsproblem, welches der Regressionsanalyse zu Grunde liegt, mit Hilfe des Gradientenverfahrens.

a) Wie in der vorherigen Aufgabe benutzen wir die Daten aus dem Excel-File `Autos_DE.xlsx`.

b) Bilden Sie aus diesen Daten den Ausgangsvektor $\boldsymbol{y}$ sowie die Datenmatrix $\boldsymbol{X}$. Benutzen Sie zur

Bildung von $\boldsymbol{X}$ zunächst nur die Variablen $x_0$ und $x_3$ (Leistung).

c) Bestimmen Sie die Eigenwerte der zugehörigen Korrelationsmatrix $\boldsymbol{R} = \frac{1}{m}\boldsymbol{X}^T\boldsymbol{X}$. Dazu können Sie den Befehl `eig` benutzen. Ist es sinnvoll, direkt mit diesen Daten das Gradientenverfahren durchzuführen?

d) Skalieren Sie jetzt die Eingangsvariable $x_3$ gemäß der Gleichung (19) und bilden Sie mit der skalierten Eingangsvariable $\tilde{x}_3$ eine skalierte Datenmatrix $\tilde{\boldsymbol{X}}$. Bestimmen Sie dann die Eigenwerte der zugehörigen Korrelationsmatrix $\boldsymbol{R} = \frac{1}{m}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$. Sind die skalierten Daten besser für das Gradientenverfahren geeignet?

e) Implementieren Sie jetzt den Algorithmus des Gradientenverfahrens. Wählen Sie eine geeignete Schrittweite $\alpha$ und führen Sie den Algorithmus für die skalierten Daten aus.

f) Berechnen Sie aus den Parametern der skalierten Daten mit Hilfe der Gleichungen (20) und (21) die Parameter der ursprünglichen Daten. Vergleichen Sie die Parameter mit denen, die Sie in der vorherigen Aufgabe mit Hilfe der Normalengleichung für die gleichen Daten berechnet haben.

g) Bilden Sie jetzt aus allen Eingangsvariablen $x_0$ bis $x_6$ die Datenmatrix $\boldsymbol{X}$ und führen Sie, wie in den vorherigen Teilaufgaben beschrieben, eine lineare Regression mit dem Gradientenverfahren durch. Vergleichen Sie die Ergebnisse mit denen der Normalengleichung.

# 5 Polynomiale Regression

Häufig sind lineare Modelle nicht ausreichend, um die Abhängigkeit der Ausgangsvariable $y$ von der Eingangsvariable $x$ mit ausreichender Genauigkeit zu beschreiben. In solchen Fällen kann es hilfreich sein, von einem linearen Modell auf ein polynomiales Modell zu wechseln. Durch eine geschickte Definition von abgeleiteten Eingangsvariablen ist es möglich, die polynomiale Regression wie eine lineare Regression zu berechnen. Dazu definieren wir aus der ursprünglichen Eingangsvariable $x$ die folgenden abgeleiteten Eingangsvariablen

$$\tilde{x}_i = x^i \qquad \forall i = 1 \ldots a \tag{22}$$

wobei $a$ der Grad des Polynoms ist, mit dem wir den Zusammenhang zwischen $x$ und $y$ modellieren wollen. Mit den abgeleiteten Eingangsvariablen $x_i$ können wir jetzt eine ganz normale lineare Regression mit mehreren Eingangsvariablen wie oben beschrieben durchführen.

Je höher wir den Grad $a$ des Polynoms wählen, desto besser kann das Modell die Trainingsdaten beschreiben. Leider bedeutet das nicht zwangsläufig, dass das auch zu besseren Ergebnissen bei neuen (nicht in den Trainingsdaten enthaltenen) Daten führt. Wenn wir $a$ zu groß wählen, passt das Modell zwar sehr gut zu den Trainingsdaten, es liefert aber in der Regel keine gute Vorhersagen für neue Daten. Wir sprechen in diesem Fall auch von *Overfitting*. Deshalb ist es in der Praxis äußerst wichtig, einen geeigneten Wert für $a$, der weder zu groß noch zu klein sein darf, zu ermitteln. In der folgenden Aufgabe sehen wir uns dazu eine Vorgehensweise an, wie sich das mit Hilfe eines sogenannten *Development-Datensatzes* erreichen lässt.

**Durchführung**

### Aufgabe D4: Polynomiale Regression und Overfitting

In dieser Aufgabe setzen wir die Regressionsanalyse ein, um aus relativ wenig Trainingsdaten ein Modell für den Zusammenhang zwischen der Lebenserwartung und dem BMI (Body Mass Index) zu ermitteln. Da der Zusammenhang nicht linear zu sein scheint, ist es sinnvoll, hier mit einem polynomialen Modell zu arbeiten.

a) Wir benutzen hier die Daten aus den beiden Files `Lebenserwartung_Training.mat` und `Lebenserwartung_Development.mat`.[a] Die $y$-Variable stellt hier die Lebenserwartung dar,

die $x$-Variable den BMI. Zur Unterscheidung von den Trainingsdaten, heißen die Ausgangs- und die Eingangsvariable bei den Development-Daten `yD` und `xD`.

b) Erstellen Sie Streudiagramme von Trainings- und den Testdaten. Wenn Sie unterschiedliche Farben für die Datenpunkte verwenden, können Sie beide Streudiagramme in ein Bild einzeichnen. Dann werden die Unterschiede zwischen den Datensätze besonders deutlich.

c) Führen Sie mit Hilfe der Trainingsdaten eine polynomiale Regression wie oben beschrieben durch. Benutzen Sie dafür folgende Grade $a = 1, 2, 3, 4, 6, 8$. Stellen Sie die jeweiligen Regressionskurven zusammen mit den Streudiagrammen in einem Bild dar. Welche Regressionskurve passt am besten?

d) Berechnen Sie für jede Regressionskurven das Bestimmtheitsmaß $R^2$. Das Bestimmtheitsmaß kann man als einen Wert ansehen, der aussagt, wie gut eine Regressionskurve passt.

e) Berechnen Sie jetzt für jede Regressionskurve (die aus den Trainingsdaten ermittelt wurde) das Bestimmtheitsmaß für die Development-Daten. Diese Werten zeigen, wie gut die Regressionskurven zu den Development-Daten passen. Was lässt sich aus diesen Werten für den optimalen Grad $a$ folgern?

f) Vermutlich haben Sie ab einem Grad von $a = 4$ Warnungen von MATLAB der Form `Warning: Matrix is close to singular or badly scaled. Results may be inaccurate` erhalten. Wie lassen sich diese Warnungen erklären? Was könnten wir machen, um die numerischen Probleme bei der Invertierung der Korrelationsmatrix zu reduzieren?

[a] Bei den Daten handelt es sich um simulierte Daten, die nicht von einer tatsächlichen Studie stammen. Ziehen Sie deshalb bitte keine medizinischen Schlüsse aus der Analyse in dieser Aufgabe.

## Weiterführende Literatur


[1] G. James, D. Witten, T. Hastie, and R.Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[2] M.H. Kutner, C.J. Nachtsheim, and J. Neter. *Applied Linear Statistical Models*. McGraw-Hill Irwin, 2004.

[3] Wikipedia. Bestimmtheitsmaß — wikipedia, die freie enzyklopädie, 2017. [Stand 4. April 2017].

[4] T. Hastie, R.Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[5] K. Schmidt and G. Trenkler. *Einführung in die moderne Matrix-Algebra*. Springer-Gabler, 2015.

[6] K.B. Peterson and M.S. Pedersen. *The Matrix Cookbook*. 2012.

[7] G. Golub and C. van Loan. *Matrix Computations*. The John Hopkins University Press, 2013.

[8] S. Haykin. *Adaptive Filter Theory*. Pearson, 2014.

[9] M. Eid, M. Gollwitzer, and M. Schmitt. *Statistik und Forschungsmethoden*. Beltz, 2015.

[10] A. Rooch. *Statistik für Ingenieure*. Springer, 2014.