§12 Regressionsanalyse pfad

▼

```
SetDirectory[
    "C:\Dokumente und Einstellungen\Administrator\Desktop\Stochastik mit Mathematica
      6.0\Statistik\Datenordner"];

<< StatisticalPlots`;
<< HypothesisTesting`;
<< ANOVA`;
<< LinearRegression`;

ScatterPlot[matrix_, options___] := Module[{ax, ay, bx, plot1, plot2},
    ax = Min[Part[matrix, All, 1]];
    ay = Min[Part[matrix, All, 2]];
    bx = Max[Part[matrix, All, 1]];
    plot1 = ListPlot[matrix, options];
    plot2 = Plot[Evaluate[Fit[matrix, {1, x}, {x}]], {x, ax, bx}, PlotStyle → {Thickness[0.01], Red}];
    Show[{plot1, plot2}]];

ScoreTest[stichprobe_, verteilungsgesetz_, options___] := Module[{z, n, p},
   z = Sort[stichprobe];
   n = Length[stichprobe];
   p = Table[{i, CDF[verteilungsgesetz, z[[i]]]}, {i, 1, n}]; ListPlot[p, options]]

RegressTest[daten_, model_, vars_, H_, ζ_] :=
  Module[{n, m, s, r, xdat, ydat, tran, mmod, xmat, inv, beta, sse, norm},
   n = Length[daten];
   m = Length[model];
   s = Length[vars];
   r = Length[H];
   xdat = Transpose[Most[Transpose[daten]]];
   ydat = Last[Transpose[daten]];
   tran = Table[vars[[i]] → xdat[[All, i]], {i, 1, s}];
   mmod = If[model[[1]] =!= 1, model, Prepend[Rest[model], Table[1, {n}]]];
   xmat = Transpose[mmod /. tran];
   inv = Inverse[Transpose[xmat].xmat];
   beta = inv.Transpose[xmat].ydat;
   sse = ydat.(IdentityMatrix[n] − xmat.inv.Transpose[xmat]).ydat;
   norm = (H.beta − ζ).Inverse[H.inv.Transpose[H]].(H.beta − ζ);
   If[Or[Det[Transpose[xmat].xmat] == 0, Det[H.inv.Transpose[H]] == 0],
     Print["Datenmaterial x und/oder Matrix H ungeeignet"],
     Print["PValue → ", 1 − CDF[FRatioDistribution[r, n − m], (norm (n − m)) / (sse r)] // N]]]
```

```
RegressDifferenceTest[daten1_, daten2_, model_, vars_, H_] := Module[{n1, n2, m, s, r, xdat1, xdat2, ydat1, ydat2,
    tran1, tran2, mmod1, mmod2, xmat1, xmat2, inv1, inv2, inv, beta1, beta2, sse1, sse2, norm},
  n1 = Length[daten1]; n2 = Length[daten2];
  m = Length[model];
  s = Length[vars];
  r = Length[H];
  xdat1 = Transpose[Most[Transpose[daten1]]]; xdat2 = Transpose[Most[Transpose[daten2]]];
  ydat1 = Last[Transpose[daten1]]; ydat2 = Last[Transpose[daten2]];
  tran1 = Table[vars[[i]] → xdat1[[All, i]], {i, 1, s}]; tran2 = Table[vars[[i]] → xdat2[[All, i]], {i, 1, s}];
  mmod1 = If[model[[1]] =!= 1, model, Prepend[Rest[model], Table[1, {n1}]]];
  mmod2 = If[model[[1]] =!= 1, model, Prepend[Rest[model], Table[1, {n2}]]];
  xmat1 = Transpose[mmod1 /. tran1]; xmat2 = Transpose[mmod2 /. tran2];
  inv1 = Inverse[Transpose[xmat1].xmat1]; inv2 = Inverse[Transpose[xmat2].xmat2];
  inv = inv1 + inv2;
  beta1 = inv1.Transpose[xmat1].ydat1; beta2 = inv2.Transpose[xmat2].ydat2;
  sse1 = ydat1.(IdentityMatrix[n1] − xmat1.inv1.Transpose[xmat1]).ydat1;
  sse2 = ydat2.(IdentityMatrix[n2] − xmat2.inv2.Transpose[xmat2]).ydat2;
  norm = H.(beta1 − beta2).Inverse[H.inv1.Transpose[H]].H.(beta1 − beta2);
  If[Or[Det[Transpose[xmat1].xmat1] == 0, Det[Transpose[xmat2].xmat2] == 0, Det[H.inv.Transpose[H]] == 0],
    Print["Datenmaterial 1x und/oder Datenmaterial 2x und/oder Matrix H ungeeignet"],
    Print["PValue → ", 1 − CDF[FRatioDistribution[r, n1 + n2 − 2 m], (norm (n1 + n2 − 2 m)) / ((sse1 + sse2) r)] //

GoldfeldQuandtTest[daten1_, daten2_, model_, vars_] := Module[{n1, n2, m, s, xdat1, xdat2, ydat1, ydat2,
    tran1, tran2, mmod1, mmod2, xmat1, xmat2, inv1, inv2, beta1, beta2, σ1sq, σ2sq, MLQ},
  n1 = Length[daten1]; n2 = Length[daten2];
  m = Length[model];
  s = Length[vars];
  xdat1 = Transpose[Most[Transpose[daten1]]]; xdat2 = Transpose[Most[Transpose[daten2]]];
  ydat1 = Last[Transpose[daten1]]; ydat2 = Last[Transpose[daten2]];
  tran1 = Table[vars[[i]] → xdat1[[All, i]], {i, 1, s}]; tran2 = Table[vars[[i]] → xdat2[[All, i]], {i, 1, s}];
  mmod1 = If[model[[1]] =!= 1, model, Prepend[Rest[model], Table[1, {n1}]]];
  mmod2 = If[model[[1]] =!= 1, model, Prepend[Rest[model], Table[1, {n2}]]];
  xmat1 = Transpose[mmod1 /. tran1]; xmat2 = Transpose[mmod2 /. tran2];
  inv1 = Inverse[Transpose[xmat1].xmat1]; inv2 = Inverse[Transpose[xmat2].xmat2];
  beta1 = inv1.Transpose[xmat1].ydat1; beta2 = inv2.Transpose[xmat2].ydat2;
  σ1sq = (ydat1 − xmat1.beta1).(ydat1 − xmat1.beta1) / n1; σ2sq = (ydat2 − xmat2.beta2).(ydat2 − xmat2.beta2) / n2
  MLQ = (σ1sq^(n1/2) σ2sq^(n2/2)) / ((n1 σ1sq + n2 σ2sq) / (n1 + n2))^((n1+n2)/2);
  If[Or[Det[Transpose[xmat1].xmat1] == 0, Det[Transpose[xmat2].xmat2] == 0],
    Print["Datenmaterial 1x und/oder Datenmaterial 2x ungeeignet"],
    Print["PValue → ", 1 − CDF[ChiSquareDistribution[1], −2 Log[MLQ]] // N]]]
```

Ähnlich wie bei der Varianzanalyse geht man bei der Regressionsanalyse davon aus, dass ein oder mehrere Faktoren eine Messung beeinflussen können. Im Unterschied zur Varianzanalyse handelt es sich bei der Regressionsanalyse aber in der Regel um quantitative Faktoren, wobei nicht nur untersucht wird, ob diese Faktoren einen Einfluss auf die Messung haben, sondern auch versucht wird, den Einfluss dieser Faktoren (im Rahmen der Regressionsanalyse spricht man von Einflussgrößen) auf diese Messung (der sogenannten Zielgröße) zu modellieren und quantitativ zu beschreiben.

Nach einer ausführlichen Darstellung der Aufgabenstellung der Regressionsanalyse befassen wir uns zuerst mit der Frage, wie sich die unbekannten Parameter der Regressionsanalyse schätzen lassen. Anschließend untersuchen wir einige Statistiken, welche im Rahmen der Regressionsanalyse eine herausragende Rolle spielen und wenden die dabei erzielten Erkenntnisse bei der Konstruktion von Tests von Hypothesen über diese Parameter an. Anschließend zeigen wir, wie sich die Varianzanalyse in die Regressionsanalyse einbetten lässt. Diese Möglichkeit ist vor allem für die Varianzanalyse von nicht-balanziertem Datenmaterial von großer Bedeutung. In einem eigenen Abschnitt befassen wir uns mit verschiedenen Diagnose-Tools. Eine Zusammenstellung von wichtigen Ergebnissen der Matrizenrechnung schließt dieses Kapitel ab.

12.1 Modellbildung und Aufgabenstellung

Den Ausgangspunkt bilden n Messergebnisse $y_1, y_2, \ldots, y_n$ einer Zielgröße, welche an den (nicht notwendig verschiedenen) Stellen $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^m$ mit $\vec{x}_i = \{x_{i\,1}, x_{i\,2}, \ldots, x_{i\,m}\}$ der m betrachteten Einflussgrößen ermittelt werden. Dabei nehmen wir an, dass zwar die Messergebnisse y_i dieser Zielgröße vom Zufall beeinflusst werden (Messfehler, Unkenntnis aller tatsächlich wirksamen Einflussgrößen, zufällige Einflüsse), die Stellen $\vec{x}_i$ der m Einflussgrößen aber keinerlei zufällige Charakter besitzen.

Ein typisches Beispiel soll diese Situation verdeutlichen: Bei einer klinischen Studie über den Einfluss von Blutdruck und Cholesteringehalt auf die Lebensdauer wurden von n jeweils 50-jährigen Personen der Blutdruck $x_{i\,1}$, der Cholesteringehalt $x_{i\,2}$, sowie die restliche Lebensdauer y_i dieser Personen ermittelt. Natürlich hängt die restliche Lebensdauer einer Person neben dem Blutdruck und dem Cholesteringehalt noch von zahlreichen weiteren Einflussgrößen (die nicht vollständig bekannt sind) sowie vom Zufall ab. Die Messergebnisse von Blutdruck und Cholesteringehalt dieser Person werden hingegen als nicht zufällig angesehen; es handelt sich dabei um Werte, die (neben einer Fülle von anderen Größen) den derzeitigen Gesundheitszustand dieser Person charakterisieren.

Fundamental für die weiteren Überlegungen ist die folgende Modellbildung:

12.1.1 Modell der linearen Regressionsanalyse: Die n Messergebnisse $y_1, y_2, \ldots, y_n$ der Zielgröße werden als Realisierungen von n Zufallsvariablen $Y_1, Y_2, \ldots, Y_n$ der Form

$$Y_i = \beta_0 + \beta_1\, x_{i\,1} + \beta_2\, x_{i\,2} + \ldots + \beta_m\, x_{i\,m} + E_i$$

interpretiert. Dabei sind $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ unbekannte Parameter, $\vec{x}_i = \{x_{i\,1}, x_{i\,2}, \ldots, x_{i\,m}\}$ die vorgegebenen Stellen der m Einflussgrößen und $E_1, E_2, \ldots, E_n$ vollständig unabhängige und $\mathcal{N}[0, \sigma]$-verteilte Zufallsvariable mit unbekannter (aber gleicher) Streuung σ. Mit den Abkürzungen

$$\vec{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}; \qquad \vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \qquad \boldsymbol{x} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1\,m} \\ 1 & x_{21} & \ldots & x_{2\,m} \\ \ldots & \ldots & \ddots & \ldots \\ 1 & x_{n\,1} & \ldots & x_{n\,m} \end{pmatrix}; \qquad \vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}; \qquad \vec{E} = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix}$$

lässt sich das Modell der linearen Regressionsanalyse in der kompakten Form

$$\vec{Y} = \boldsymbol{x} \cdot \vec{\beta} + \vec{E}$$

schreiben. Der Zufallsvektor $\vec{Y}$ ist damit offenbar $\mathcal{MN}[\boldsymbol{x} \cdot \vec{\beta}, \sigma^2\, \mathbf{E}_n]$-verteilt.

▼

Sowohl der Zufallsvektor $\vec{Y}$ als auch der Erwartungswertvektor $\boldsymbol{x} \cdot \vec{\beta}$ sind Spaltenvektoren; diese Tatsache ist bei der Verwendung von Transformationsformeln zu beachten.

Zu dieser Modellbildung sind einige Bemerkungen und Begriffsbildungen angebracht:

▪ Der Name lineare Regressionsanalyse bezieht sich auf die Tatsache, dass die Parameter $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ linear in das Modell eingehen. Man beachte aber, dass hinsichtlich der Einflussgrößen keine Linearität vorausgesetzt wird. Beispielsweise lässt sich die polynomiale Beziehung

$$Y_i = \beta_0 + \beta_1\, x_i + \beta_2\, x_i^{\,2} + \ldots + \beta_m\, x_i^{\,m} + E_i$$

zwischen der Zielgröße Y und einer einzigen Einflussgröße x durch die Transformation $x_i^{\,k} \to x_{i\,k}$ in das lineare Regressionsmodell

$$Y_i = \beta_0 + \beta_1\, x_{i\,1} + \beta_2\, x_{i\,2} + \ldots + \beta_m\, x_{i\,m} + E_i$$

mit den m Einflussgrößen $x_1, x_2, \ldots, x_m$ überführen.

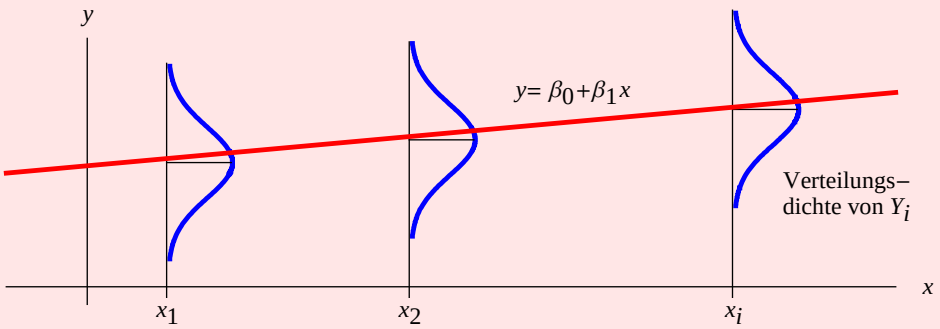
▪ Die Hyperebene

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

des $\mathbb{R}^{m+1}$ nennt man Regressionshyperebene. Im Fall $m = 1$ spricht man von einer Regressionsgeraden.

▪ Sind die n Stellen $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^m$ der m Einflussgrößen so gewählt, dass die Matrix $\boldsymbol{x}^t \cdot \boldsymbol{x} \in \mathbb{R}^{m+1}_{m+1}$ den Rang $m+1$ besitzt (und damit invertierbar ist), so spricht man von einer linearen Regressionsanalyse mit vollem Rang. Wir setzen in Zukunft stets stillschweigend voraus, dass wir es mit einer linearen Regressionsanalyse mit vollem Rang zu tun haben.

▪ Für $m = 1$ lässt sich das Modell der linearen Regressionsanalyse durch die folgende Zeichnung veranschaulichen:



Wir wollen nun die Aufgabe der linearen Regressionsanalyse formulieren:

12.1.2 Aufgabe der linearen Regressionsanalyse: Die Aufgabe der linearen Regressionsanalyse besteht darin, aufgrund der n Messwerte $\{\vec{x}_1, y_1\}, \{\vec{x}_2, y_2\}, \ldots, \{\vec{x}_n, y_n\} \in \mathbb{R}^{m+1}$

- die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ und σ^2 zu schätzen;
- zu beschreiben, wie gut sich die Zielgröße durch die Einflussgrößen beschreiben lässt;
- Hypothesen über die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ zu testen; die Frage, ob die i-te Einflussgröße tatsächlich einen Einfluss auf die Zielgröße hat, lässt sich beispielsweise beantworten, indem man die Hypothese $\beta_i = 0$ gegen die Alternative $\beta_i \neq 0$ testet.

12.2 Schätzer für die Parameter $\beta_0, \beta_1, \ldots, \beta_m$ und σ^2

Gegeben seien die n Messwerte $\{\vec{x}_1, y_1\}, \{\vec{x}_2, y_2\}, \ldots, \{\vec{x}_n, y_n\} \in \mathbb{R}^{m+1}$, wobei wir voraussetzen, dass die Matrix $\boldsymbol{x}^t \cdot \boldsymbol{x} \in \mathbb{R}^{m+1}_{m+1}$ den Rang $m+1$ besitzt, wir es also mit einer linearen Regressionsanalyse mit vollem Rang zu tun haben. Wir befassen uns in diesem Abschnitt damit, ausgehend von diesen Messwerten, die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ und σ des Modells

$$\vec{Y} = \boldsymbol{x} \cdot \vec{\beta} + \vec{E}$$

zu schätzen.

12.2.1 Methode der kleinsten Quadrate: Die Methode der kleinsten Quadrate besteht darin, die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ so zu bestimmen, dass die Regressionshyperebene

$$y = \beta_0 + \beta_1 x_1 + \beta_m x_m$$

den Punktschwarm $\{\vec{x}_1, y_1\}, \{\vec{x}_2, y_2\}, \ldots, \{\vec{x}_n, y_n\} \in \mathbb{R}^{m+1}$ möglichst gut approximiert. Man erhält auf diese Weise die Schätzwerte

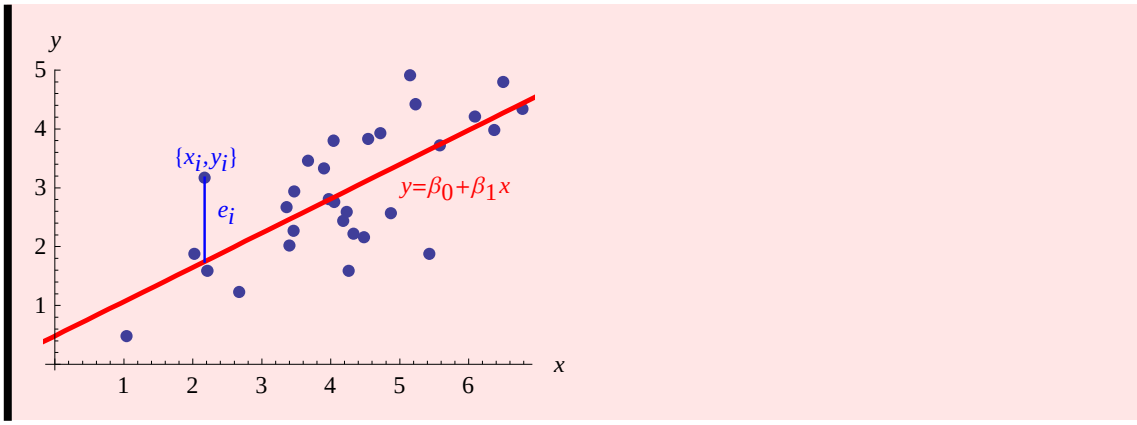
$$\hat{\vec{\beta}} = \{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m\}^t = \left(\boldsymbol{x}^t \cdot \boldsymbol{x}\right)^{-1} \cdot \boldsymbol{x}^t \cdot \vec{y}$$

▼

Beweis: Wir bestimmen die Parameter $\beta_0, \beta_1, \ldots, \beta_m$ so, dass die Summe der quadratischen Abweichungen

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i\,1} + \beta_m x_{i\,m})]^2 = (\vec{y} - \boldsymbol{x}\cdot\vec{\beta})^t \cdot (\vec{y} - \boldsymbol{x}\cdot\vec{\beta})$$

der Punkte $\{\vec{x}_i, y_i\}$ von der Regressionshyperebene $y = \beta_0 + \beta_1 x_1 + \beta_m x_m$ zu einem Minimum wird. Für $m = 1$ lässt sich diese Aufgabenstellung durch die folgende Zeichnung veranschaulichen (man beachte, dass es sich bei diesen Abweichungen nicht! um die kürzesten Abstände der Punkte $\{x_i, y_i\}$ von der Regressionsgeraden handelt):



Diese Extremwertaufgabe lösen wir in der üblichen Weise, indem wir die partiellen Ableitungen der Zielfunktion $(\vec{y} - \boldsymbol{x}\cdot\vec{\beta})^t \cdot (\vec{y} - \boldsymbol{x}\cdot\vec{\beta})$ nach den Variablen $\beta_0, \beta_1, \ldots, \beta_m$ gleich Null setzen:

$$\frac{\partial}{\partial \vec{\beta}} (\vec{y} - \boldsymbol{x}\cdot\vec{\beta})^t \cdot (\vec{y} - \boldsymbol{x}\cdot\vec{\beta}) = -2\,(\boldsymbol{x}^t \cdot \vec{y} - \boldsymbol{x}^t \cdot \boldsymbol{x} \cdot \vec{\beta}) = \vec{0}$$

Da wir vorausgesetzt haben, dass die Matrix $\boldsymbol{x}^t \cdot \boldsymbol{x}$ invertierbar ist, hat dies offenbar

$$\hat{\vec{\beta}} = (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{y}$$

zur Folge.

12.2.2 Maximum-Likelihood-Methode: Die Maximum-Likelihood-Methode besteht darin, die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ und σ so zu bestimmen, dass die Likelihood-Funktion

$$L[y_1, y_2, \ldots, y_n \,|\, \beta_0, \beta_1, \ldots, \beta_m, \sigma]$$

maximal wird. Man erhält auf diese Weise die Schätzwerte

$$\hat{\vec{\beta}} = \{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m\}^t = \left(\boldsymbol{x}^t \cdot \boldsymbol{x}\right)^{-1} \cdot \boldsymbol{x}^t \cdot \vec{y} \qquad \text{und} \qquad \hat{\sigma}^2 = \frac{1}{n} (\vec{y} - \boldsymbol{x}\cdot\hat{\vec{\beta}})^t \cdot (\vec{y} - \boldsymbol{x}\cdot\hat{\vec{\beta}})$$

▼

Beweis: Der Zufallsvektor $\vec{Y}$ ist bekanntlich $\mathcal{MN}[\boldsymbol{x}\cdot\vec{\beta}, \sigma^2\,\mathbf{E}_n]$-verteilt, also gilt für die Likelihood-Funktion

$$L[y_1, y_2, \ldots, y_n \,|\, \beta_0, \beta_1, \ldots, \beta_m, \sigma] = \frac{1}{(2\,\pi\,\sigma^2)^{n/2}}\, \mathrm{Exp}[-\frac{1}{2\,\sigma^2} (\vec{y} - \boldsymbol{x}\cdot\vec{\beta})^t \cdot (\vec{y} - \boldsymbol{x}\cdot\vec{\beta})]$$

Wir lösen diese Extremwertaufgabe in der üblichen Weise, indem wir die partiellen Ableitungen des Logarithmus der Likelihood-Funktion nach den Variablen $\beta_0, \beta_1, \ldots, \beta_m$ und σ gleich Null setzen:

$$\frac{\partial}{\partial \vec{\beta}} \mathrm{Log}[L[y_1, y_2, \ldots, y_n \,|\, \beta_0, \beta_1, \ldots, \beta_m, \sigma]] = \frac{1}{\sigma^2} (\boldsymbol{x}^t \cdot \vec{y} - \boldsymbol{x}^t \cdot \boldsymbol{x} \cdot \vec{\beta}) = \vec{0}$$

$$\frac{\partial}{\partial \sigma} \text{Log}[L[y_1, y_2, \ldots, y_n \mid \beta_0, \beta_1, \ldots, \beta_m, \sigma]] = -\frac{n}{\sigma} + \frac{1}{\sigma^3} (\vec{y} - \boldsymbol{x} \cdot \vec{\beta})^t \cdot (\vec{y} - \boldsymbol{x} \cdot \vec{\beta}) = 0$$

Da wir vorausgesetzt haben, dass die Matrix $\boldsymbol{x}^t \cdot \boldsymbol{x}$ invertierbar ist, gilt

$$\hat{\vec{\beta}} = (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{y}$$

und damit

$$\hat{\sigma}^2 = \frac{1}{n} (\vec{y} - \boldsymbol{x} \cdot \hat{\vec{\beta}})^t \cdot (\vec{y} - \boldsymbol{x} \cdot \hat{\vec{\beta}})$$

Man beachte, dass die Methode der kleinsten Quadrate und die Maximum-Likelihood-Methode auf die gleichen Schätzwerte für die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ führen, mit der Maximum-Likelihood-Methode aber auch noch der unbekannte Parameter σ^2 geschätzt werden kann!

12.2.3 Bemerkung: Im Fall $m = 1$ ergeben sich speziell die Schätzwerte

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n y_i)^2}$$

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

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Zur Berechnung der Schätzwerte $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ der unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ dient der Befehl Fit (soll auch der unbekannte Parameter σ geschätzt werden, so verwende man den weiter unten beschriebenen Befehl Regress). *Mathematica* unterscheidet dabei zwischen Variablen und Einflussgrößen (Funktionen der Variablen). Dies hat den Vorteil, dass (etwa beim polynomialen Modell) das einzugebende Datenmaterial oft wesentlich weniger umfangreich ist, da nur die Werte der s Variablen und nicht die Werte der m Einflussgrößen eingegeben werden müssen. Auch muss das Datenmaterial nicht jedes Mal geändert werden, wenn das Modell geändert wird.

■ Fit[*daten, model, vars*]

gibt für das Datenmaterial *daten* die Gleichung der Regressionshyperebene (und damit auch die Schätzwerte für die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$) aus.

Das Datenmaterial *daten* muss dabei die Form $\{\{x_{11}, x_{12}, \ldots, x_{1\,s}, y_1\}, \{x_{21}, x_{22}, \ldots, x_{2\,s}, y_2\}, \ldots\}$ besitzen. Mit der Liste *model* lässt sich steuern, welche Funktionen der Variablen als Einflussgrößen in das Modell aufgenommen werden sollen. In der Liste *vars* werden alle im Datenmaterial *daten* aufscheinenden Variablen aufgelistet.

Wir erläutern die Verwendung dieses Befehls an Hand von Beispielen:

12.2.4 Beispiel: Von $n = 548$ Stahlblechen wurde die Dicke, die Haspeltemperatur, der Mangangehalt, und die Zugfestigkeit gemessen (man vergleiche dazu das Datenmaterial zugfestigkeit). Aus langjähriger Erfahrung weiß man, dass zwischen der Zugfestigkeit, der Dicke, der Haspeltemperatur und dem Mangangehalt von Stahlblechen die lineare Beziehung

$$\text{Zugfestigkeit} = \beta_0 + \beta_1\,\text{Dicke} + \beta_2\,\text{Haspeltemperatur} + \beta_3\,\text{Mangangehalt} + E$$

besteht, wobei die den Zufall und weitere nicht bekannte Einflussgrößen charakterisierende Zufallsvariable E wie üblich als $\mathcal{N}[0, \sigma]$-verteilt angenommen wird. Man schätze die unbekannten Parameter β_0, β_1, β_2, β_3.

▼

Lösung: Wir ermitteln die unbekannten Parameter β_0, β_1, β_2, β_3 unter Verwendung von Fit

```
zugfestigkeit = Rest[<< zugfestigkeitfile];
Fit[zugfestigkeit, {1, Dicke, Haspeltemperatur, Mangangehalt}, {Dicke, Haspeltemperatur, Mangangehalt}]
Clear[zugfestigkeit]
```

```
532.316 - 3.48119 Dicke - 0.243508 Haspeltemperatur + 28.8377 Mangangehalt
```

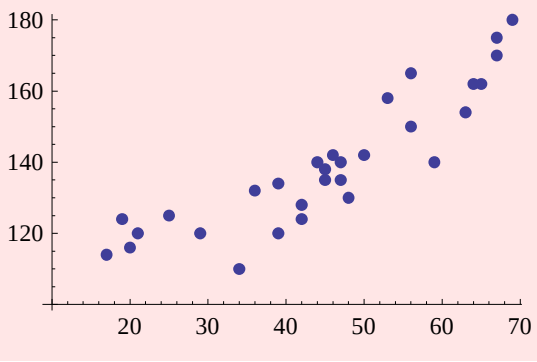
und erhalten damit die Schätzwerte $\hat{\beta}_0 = 532.316$, $\hat{\beta}_1 = -3.48119$, $\hat{\beta}_2 = -0.243508$ und $\hat{\beta}_3 = 28.8377$.

12.2.5 Beispiel: Von $n = 30$ zufällig ausgewählten Personen wurde das Alter und der systolische Blutdruck ermittelt (man vergleiche dazu das Datenmaterial blutdruck). Welche Art von funktionalem Zusammenhang besteht zwischen diesen beiden Größen?

▼

Lösung: Wir veranschaulichen das Datenmaterial zuerst graphisch mit Hilfe von ListPlot:

```
blutdruck = Rest[<< "blutdruckfile"];
ListPlot[blutdruck, PlotStyle → PointSize[0.025], AxesOrigin → {10, 100}, ImageSize → {200, 125}]
```



An Hand dieser Graphik erkennt man, dass der systolische Blutdruck mit dem Alter zunimmt, wobei sowohl das Modell $\text{Blutdruck} = \alpha_0 + \alpha_1\,\text{Alter}$ als auch das Modell $\text{Blutdruck} = \beta_0 + \beta_1\,\text{Alter} + \beta_2\,\text{Alter}^2$ naheliegend sind. Mit Hilfe von Fit schätzen wir nun die Parameter α_0 und α_1 bzw β_0, β_1 und β_2 dieser beiden Modelle:

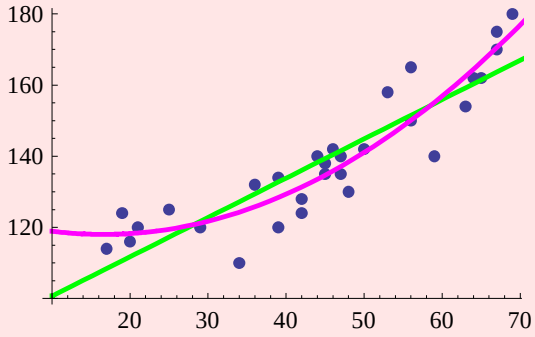
```
a = Fit[blutdruck, {1, Alter}, {Alter}]
b = Fit[blutdruck, {1, Alter, Alter^2}, {Alter}]
```

```
89.6724 + 1.10401 Alter

123.734 - 0.686878 Alter + 0.0206646 Alter^2
```

Zeichnet man zusätzlich zu den Messwerten auch die Graphen dieser beiden Regressions-Kurven in eine gemeinsame Zeichnung, so erkennt man, dass die lila Kurve (quadratische Beziehung) den Punktschwarm offenbar besser approximiert, als die grüne Kurve (lineare Beziehung).

```
p1 = ListPlot[blutdruck, PlotStyle → PointSize[0.025], AxesOrigin → {10, 100}];
p2 = Plot[a, {Alter, 10, 75}, PlotStyle → {Green, Thickness[0.01]}];
p3 = Plot[b, {Alter, 10, 75}, PlotStyle → {Magenta, Thickness[0.01]}];
Show[{p1, p2, p3}, ImageSize → {200, 125}]
Clear[blutdruck, a, b, p1, p2, p3]
```



12.2.6 Beispiel: Eine bestimmte Sorte Eiscreme wird in Behältern von je 100 cm^3 geliefert. Bei längerer Lagerung schrumpft das Volumen des in diesen Behältern aufbewahrten Eises. Für einige verschiedene Lagerzeiten wurde der mittlere Volumenverlust ermittelt (man vergleiche dazu das Datenmaterial eiscreme). Welche Art von funktionalem Zusammenhang besteht zwischen diesen beiden Größen?

▼

12.2.7 Beispiel: Von einem fallenden Körper mit unbekannter Anfangsgeschwindigkeit v wurden die bis zum Zeitpunkt t [sek] zurückgelegten Fallhöhen $h[t]$ [cm] ermittelt (man vergleiche dazu das Datenmaterial fallhöhe). Man verwende diese Daten zur Bestimmung der Erdbeschleunigung g.

▼

Lösung: Die Fallhöhe $h[t]$ eines Körpers bis zum Zeitpunkt t berechnet sich bekanntlich nach der Formel

$$h[t] = v\,t + \frac{g}{2}\,t^2$$

wobei v [cm/sek] die Anfangsgeschwindigkeit und g $\left[\mathrm{cm}/\mathrm{sek}^2\right]$ die Erdbeschleunigung bezeichnet. Wir schätzen diese beiden unbekannten Parameter v und $g/2$ mit Hilfe von Fit

```
Fit[Rest[<< fallhöhefile], {Zeit, Zeit^2}, {Zeit}]
```

```
54.8078 Zeit + 517.061 Zeit^2
```

und erhalten für die Erdbeschleunigung den Schätzwert $\hat{g} = 10.34\ m/\mathrm{sek}^2$.

Im folgenden Satz fassen wir die wichtigsten Eigenschaften der mit Hilfe der Methode der kleinsten Quadrate bzw der Maximum-Likelihood-Methode gewonnenen Schätzer

$$\hat{\vec{B}} = \{\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_m\}^t = (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{Y} \qquad \text{und} \qquad \hat{\Sigma}^2 = \frac{1}{n}\,(\vec{Y} - \boldsymbol{x} \cdot \hat{\vec{B}})^t \cdot (\vec{Y} - \boldsymbol{x} \cdot \hat{\vec{B}})$$

für die unbekannten Parameter $\vec{\beta} = \{\beta_0, \beta_1, \ldots, \beta_m\}^t$ und σ^2 zusammen (man beachte dabei den Unterschied zwischen den Schätzwerten $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ bzw $\hat{\sigma}^2$ und den Schätzern $\hat{\mathrm{B}}_0, \hat{\mathrm{B}}_1, \ldots, \hat{\mathrm{B}}_m$ bzw $\hat{\Sigma}^2$, der bekanntlich darin besteht, dass an Stelle der konkreten Messwerte $\vec{y}$ der Zufallsvektor $\vec{Y}$ verwendet wird):

12.2.8 Satz: Die beiden Schätzer $\hat{\vec{\mathrm{B}}}$ und $\hat{\Sigma}^2$ für die unbekannten Parameter $\vec{\beta} = \{\beta_0, \beta_1, \ldots, \beta_m\}^t$ und σ^2 besitzen die folgenden Eigenschaften

a) Der Zufallsvektor $\hat{\vec{\mathrm{B}}}$ ist $\mathcal{MN}[\vec{\beta}, \sigma^2 (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1}]$-verteilt;

b) Die Zufallsvariable $n \hat{\Sigma}^2 / \sigma^2$ ist $\mathcal{C}\mathrm{hi}[n - (m+1)]$-verteilt;

c) Der Zufallsvektor $\hat{\vec{\mathrm{B}}}$ und die Zufallsvariable $\hat{\Sigma}^2$ sind unabhängig.

▼

Beweis:

a) Der Zufallsvektor $\vec{Y}$ ist bekanntlich $\mathcal{MN}[\boldsymbol{x} \cdot \vec{\beta}, \sigma^2 \mathbf{E}]$-verteilt. Aus dem Satz über die affine Transformation folgt damit wegen

$$\mathbb{E}[\hat{\vec{\mathrm{B}}}] = \mathbb{E}[(\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{Y}] = (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \mathbb{E}[\vec{Y}] = (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \boldsymbol{x} \cdot \vec{\beta} = \vec{\beta}$$

($\hat{\vec{\mathrm{B}}}$ ist damit ein erwartungstreuer Schätzer für $\vec{\beta}$) und

$$(\mathbb{K}[\hat{\mathrm{B}}_i, \hat{\mathrm{B}}_k])_{i,k \in \{0,1,\ldots,m\}} = (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot (\sigma^2 \mathbf{E}) \cdot \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} = \sigma^2 (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1}$$

dass der Schätzer $\hat{\vec{\mathrm{B}}}$ multinormalverteilt ist mit dem Mittelwertsvektor $\vec{\beta}$ und der Kovarianzmatrix $\sigma^2 \left(\boldsymbol{x}^t \cdot \boldsymbol{x}\right)^{-1}$.

b) Mit dem offenbar $\mathcal{MN}[\vec{0}, \mathbf{E}]$-verteilten Zufallsvektor $\vec{Z} = (\vec{Y} - \boldsymbol{x} \cdot \vec{\beta}) / \sigma$ gilt

$$n \hat{\Sigma}^2 / \sigma^2 = \frac{1}{\sigma^2} (\vec{Y} - \boldsymbol{x} \cdot \hat{\vec{\mathrm{B}}})^t \cdot (\vec{Y} - \boldsymbol{x} \cdot \hat{\vec{\mathrm{B}}}) = \frac{1}{\sigma^2} (\vec{Y} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{Y})^t \cdot (\vec{Y} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{Y}) =$$

$$= \frac{1}{\sigma^2} (\vec{Y}^t \cdot \vec{Y} - \vec{Y}^t \cdot \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{Y}) = \frac{1}{\sigma^2} (\vec{Y}^t \cdot (\mathbf{E} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t) \cdot \vec{Y}) =$$

$$= \frac{1}{\sigma^2} ((\sigma \vec{Z} + \boldsymbol{x} \cdot \vec{\beta})^t \cdot (\mathbf{E} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t) \cdot (\sigma \vec{Z} + \boldsymbol{x} \cdot \vec{\beta})) = \vec{Z}^t \cdot (\mathbf{E} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t) \cdot \vec{Z}$$

i) Die symmetrische Matrix $\boldsymbol{\Gamma}_1 = \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t$ besitzt den Rang $m+1$: Wir haben vorausgesetzt, dass die Matrix $\boldsymbol{x}^t \cdot \boldsymbol{x}$ den Rang $m+1$ besitzt. Damit ist der Rang von $\boldsymbol{\Gamma}_1$ jedenfalls nicht größer als $m+1$. Wäre er aber kleiner als $m+1$, so wäre der Rang von $\boldsymbol{\Gamma}_1 \cdot \boldsymbol{x} = \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \boldsymbol{x} = \boldsymbol{x}$ und damit auch der Rang von $\boldsymbol{x}^t \cdot \boldsymbol{x}$ kleiner als $m+1$, was im Widerspruch zur Voraussetzung steht.

ii) Die symmetrische Matrix $\boldsymbol{\Gamma}_2 = \mathbf{E} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t = \mathbf{E} - \boldsymbol{\Gamma}_1$ besitzt den Rang $n - (m+1)$: Die Matrizen $\boldsymbol{\Gamma}_1$ und $\boldsymbol{\Gamma}_2$ sind offenbar idempotent, also folgt aus Hilfssatz 1

$$\mathrm{Rg}[\boldsymbol{\Gamma}_2] = \mathrm{Spur}[\boldsymbol{\Gamma}_2] = \mathrm{Spur}[\mathbf{E} - \boldsymbol{\Gamma}_1] = \mathrm{Spur}[\mathbf{E}] - \mathrm{Spur}[\boldsymbol{\Gamma}_1] = n - \mathrm{Rg}[\boldsymbol{\Gamma}_1] = n - (m+1)$$

Der Zufallsvektor $\vec{Z}$ und die Matrizen $\boldsymbol{\Gamma}_1$ und $\boldsymbol{\Gamma}_2$ erfüllen wegen i) und ii) die Voraussetzungen des Satzes von Cochran, also ist die Zufallsvariable

$$n \hat{\Sigma}^2 / \sigma^2 = \vec{Z}^t \cdot (\mathbf{E} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t) \cdot \vec{Z} = \vec{Z}^t \cdot \boldsymbol{\Gamma}_2 \cdot \vec{Z}$$

$\mathcal{C}\mathrm{hi}[n - (m+1)]$-verteilt.

c) Die Matrizen $\mathbf{A} = (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t$ und $\mathbf{B} = \mathbf{E} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t$ erfüllen zusammen mit dem Zufallsvektor $\vec{Z}$ die Voraussetzungen von Hilfssatz 2, also sind $(\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{Z}$ und $\vec{Z}^t \cdot (\mathbf{E} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t) \cdot \vec{Z}$ und damit auch

$$\hat{\vec{\mathrm{B}}} = \sigma\,(\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t \cdot \vec{Z} + \vec{\beta} \qquad \text{und} \qquad \hat{\Sigma}^2 = \frac{\sigma^2}{n}\,\vec{Z}^t \cdot (\mathbf{E} - \boldsymbol{x} \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}^t) \cdot \vec{Z}$$

unabhängig.

12.3 Wichtige Statistiken der Regressionsanalyse

Unter Verwendung der Statistiken

$$\hat{Y}_i = \hat{\mathrm{B}}_0 + \hat{\mathrm{B}}_1\, x_{i\,1} + \hat{\mathrm{B}}_2\, x_{i\,2} + \ldots + \hat{\mathrm{B}}_m\, x_{i\,m}$$

(da $\hat{\vec{\mathrm{B}}}$ bekanntlich ein erwartungstreuer Schätzer für $\vec{\beta}$ ist, handelt es sich bei $\hat{Y}_i$ um einen erwartungstreuen Schätzer für den unbekannten Mittelwert $\mathbb{E}[Y_i]$) sowie der arithmetischen Mittel

$$\overline{Y}_{\blacksquare} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{bzw} \qquad \overline{\hat{Y}}_{\blacksquare} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i$$

der Zufallsvektoren $\vec{Y} = \{Y_1, Y_2, \ldots, Y_n\}^t$ bzw $\vec{\hat{Y}} = \boldsymbol{x} \cdot \hat{\vec{\mathrm{B}}} = \{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n\}^t$ führen wir nun einige für die Regressionsanalyse fundamentale Statistiken ein:

12.3.1 Definition:

a) Die Statistik SSR (**S**um of **S**quares of **R**egression) mit

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y}_{\blacksquare})^2$$

ist ein Maß für den Anteil der Schwankung der Zufallsvariablen $Y_1, Y_2, \ldots, Y_n$ um den Gesamtmittelwert $\overline{Y}_{\blacksquare}$, welcher durch die Regression erklärt werden kann.

b) Die Statistik SSE (**S**um of **S**quares of **E**rror) mit

$$\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

ist ein Maß für den Anteil der Schwankung der Zufallsvariablen $Y_1, Y_2, \ldots, Y_n$ um den Gesamtmittelwert $\overline{Y}_{\blacksquare}$, welcher durch die Regression nicht erklärt werden kann (vergleiche dazu Bemerkung 12.3.3).

c) Die Statistik SStotal (total **S**um of **S**quares) mit

$$\mathrm{SStotal} = \sum_{i=1}^{n} (Y_i - \overline{Y}_{\blacksquare})^2$$

ist ein Maß für die gesamte Schwankung der Zufallsvariablen $Y_1, Y_2, \ldots, Y_n$ um den Gesamtmittelwert $\overline{Y}_{\blacksquare}$.

d) Die Statistik R^2 (**R**squared) mit

$$R^2 = \frac{\left(\sum_{i=1}^{n} (Y_i - \overline{Y}_{\blacksquare})\,(\hat{Y}_i - \overline{\hat{Y}}_{\blacksquare})\right)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y}_{\blacksquare})^2 \sum_{i=1}^{n} (\hat{Y}_i - \overline{\hat{Y}}_{\blacksquare})^2}$$

ist ein Maß dafür, wie stark die Zufallsvektoren $\vec{Y}$ und $\vec{\hat{Y}}$ korrelieren.

e) Die Statistik R^2_{ad} (adjusted **R**squared) mit

$$R^2_{\mathrm{ad}} = 1 - \frac{n-1}{n-(m+1)}\left(1 - R^2\right)$$

dient dazu, für verschiedene Regressionsmodelle die Korrelation der Zufallsvektoren $\vec{Y}$ und $\vec{\hat{Y}}$ miteinander vergleichen zu können.

Für weitere theoretische Untersuchungen aber auch für praktische Berechnungen am Computer sind die im folgenden Satz angeführten Formeln von Bedeutung. Dazu bezeichne **I** jene Matrix vom Format $n \times n$, deren sämtliche

Einträge gleich $1/n$ sind und $\mathbf{G}$ jene Matrix vom Format $m\times(m+1)$, welche man erhält, wenn man der $m\times m$ Einheitsmatrix eine Nullspalte voransetzt, also

$$\mathbf{I}=\frac{1}{n}\begin{pmatrix}1&1&\dots&1\\1&1&\dots&1\\\vdots&\vdots&\ddots&\vdots\\1&1&\dots&1\end{pmatrix}\in\mathbb{R}_n^n \qquad \text{und} \qquad \mathbf{G}=\begin{pmatrix}0&1&0&\dots&0\\0&0&1&\dots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\dots&1\end{pmatrix}\in\mathbb{R}_m^{m+1}$$

Man beachte, dass $\mathbf{I}\cdot\mathbf{I}=\mathbf{I}$ ist und $\mathbf{I}\cdot\vec{Y}=\{\overline{Y}_\bullet,\underbrace{\overline{Y}_\bullet,\dots,\overline{Y}_\bullet}_{n\text{ mal}}\}^t$ sowie $\mathbf{I}\cdot\boldsymbol{x}\cdot\hat{\vec{\mathrm{B}}}=\mathbf{I}\cdot\hat{\vec{Y}}=\{\overline{\hat{Y}}_\bullet,\underbrace{\overline{\hat{Y}}_\bullet,\dots,\overline{\hat{Y}}_\bullet}_{n\text{ mal}}\}^t$ und damit

$$\vec{Y}^t\cdot\mathbf{I}\cdot\vec{Y}=\vec{Y}^t\cdot\mathbf{I}\cdot\mathbf{I}\cdot\vec{Y}=n\,\overline{Y}_\bullet^{\,2} \qquad \text{sowie} \qquad \hat{\vec{\mathrm{B}}}^t\cdot\boldsymbol{x}^t\cdot\mathbf{I}\cdot\boldsymbol{x}\cdot\hat{\vec{\mathrm{B}}}=\hat{\vec{\mathrm{B}}}^t\cdot\boldsymbol{x}^t\cdot\mathbf{I}\cdot\mathbf{I}\cdot\boldsymbol{x}\cdot\hat{\vec{\mathrm{B}}}=n\,\overline{\hat{Y}}_\bullet^{\,2}$$

gilt.

12.3.2 Satz: Für die Statistiken SSR, SSE und SStotal und R^2 gelten die folgenden Formeln:

$$\text{SSR}=\vec{Y}^t\cdot\left(\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t-\mathbf{I}\right)\cdot\vec{Y}=\hat{\vec{\mathrm{B}}}^t\cdot\mathbf{G}^t\cdot(\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t)^{-1}\cdot\mathbf{G}\cdot\hat{\vec{\mathrm{B}}}$$
$$\text{SSE}=\vec{Y}^t\cdot(\mathbf{E}-\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t)\cdot\vec{Y}=n\,\hat{\Sigma}^2$$
$$\text{SStotal}=\vec{Y}^t\cdot(\mathbf{E}-\mathbf{I})\cdot\vec{Y}$$
$$R^2=\frac{\text{SSR}}{\text{SStotal}}$$

▼

Beweis: a) Berücksichtigt man, dass wegen $\boldsymbol{x}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}=\mathbf{E}$ natürlich $\boldsymbol{x}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t=\boldsymbol{x}^t$ ist, womit gezeigt ist, dass alle Spaltensummen der Matrix $\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t$ gleich eins sind (die erste Zeile der Matrix $\boldsymbol{x}^t$ besteht aus lauter Einsern) und damit $\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\mathbf{I}=\mathbf{I}\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t=\mathbf{I}$ gilt, so erhält man

$$\begin{aligned}\text{SSR}&=\textstyle\sum_{i=1}^n(\hat{Y}_i-\overline{Y}_\bullet)^2=(\boldsymbol{x}\cdot\hat{\vec{\mathrm{B}}}-\mathbf{I}\cdot\vec{Y})^t\cdot(\boldsymbol{x}\cdot\hat{\vec{\mathrm{B}}}-\mathbf{I}\cdot\vec{Y})=\\&=\vec{Y}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\vec{Y}-2\,\vec{Y}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\mathbf{I}\cdot\vec{Y}+\vec{Y}^t\cdot\mathbf{I}\cdot\mathbf{I}\cdot\vec{Y}=\\&=\vec{Y}^t\cdot\left(\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t-\mathbf{I}\right)\cdot\vec{Y}\end{aligned}$$

Bezeichnet nun $\tilde{\boldsymbol{x}}\in\mathbb{R}_n^m$ jene Matrix, welche aus der Matrix $\boldsymbol{x}$ durch Weglassen der ersten Spalte hervorgeht und $\vec{\tilde{\boldsymbol{x}}}_\bullet^{\,t}=\{\overline{x}_{1\,\bullet},\overline{x}_{2\,\bullet},\dots,\overline{x}_{m\,\bullet}\}\in\mathbb{R}^m$ jenen Zeilenvektor, dessen i-ter Eintrag $\overline{x}_{i\,\bullet}$ gleich dem Mittelwert der i-ten Spalte von $\tilde{\boldsymbol{x}}$ ist, so gilt mit $\mathbf{S}=\tilde{\boldsymbol{x}}^t\cdot\tilde{\boldsymbol{x}}-n\,\vec{\tilde{\boldsymbol{x}}}_\bullet\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet^{\,t}$ wegen Hilfssatz 3 einerseits (wir verwenden Blockmatrizen)

$$\begin{aligned}\text{SSR}&=\vec{Y}^t\cdot\left(\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t-\mathbf{I}\right)\cdot\vec{Y}=\vec{Y}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\vec{Y}-n\;\overline{Y}_\bullet^{\,2}=\\&=(\,n\,\overline{Y}_\bullet\;|\;\vec{Y}^t\cdot\tilde{\boldsymbol{x}}\,)\cdot\left(\begin{array}{c|c}1/n+\vec{\tilde{\boldsymbol{x}}}_\bullet^{\,t}\cdot\mathbf{S}^{-1}\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet & -\vec{\tilde{\boldsymbol{x}}}_\bullet^{\,t}\cdot\mathbf{S}^{-1}\\\hline -\mathbf{S}^{-1}\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet & \mathbf{S}^{-1}\end{array}\right)\cdot(\,n\,\overline{Y}_\bullet\;|\;\vec{Y}^t\cdot\tilde{\boldsymbol{x}}\,)^t=\\&=n^2\,\overline{Y}_\bullet^{\,2}\,\vec{\tilde{\boldsymbol{x}}}_\bullet^{\,t}\cdot\mathbf{S}^{-1}\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet-2\,n\,\overline{Y}_\bullet\,\vec{Y}^t\cdot\tilde{\boldsymbol{x}}\cdot\mathbf{S}^{-1}\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet+\vec{Y}^t\cdot\tilde{\boldsymbol{x}}\cdot\mathbf{S}^{-1}\cdot\tilde{\boldsymbol{x}}^t\cdot\vec{Y}\end{aligned}$$

und andererseits wegen

$$\begin{aligned}\mathbf{G}\cdot\vec{\mathrm{B}}&=\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\vec{Y}=(\,\mathbf{0}\;|\;\mathbf{E}\,)\cdot\left(\begin{array}{c|c}1/n+\vec{\tilde{\boldsymbol{x}}}_\bullet^{\,t}\cdot\mathbf{S}^{-1}\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet & -\vec{\tilde{\boldsymbol{x}}}_\bullet^{\,t}\cdot\mathbf{S}^{-1}\\\hline -\mathbf{S}^{-1}\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet & \mathbf{S}^{-1}\end{array}\right)\cdot(\,n\,\overline{Y}_\bullet\;|\;\vec{Y}^t\cdot\tilde{\boldsymbol{x}}\,)^t=\\&=(\,-\mathbf{S}^{-1}\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet\;|\;\mathbf{S}^{-1}\,)\cdot(\,n\,\overline{Y}_\bullet\;|\;\vec{Y}^t\cdot\tilde{\boldsymbol{x}}\,)^t=\mathbf{S}^{-1}\cdot(\tilde{\boldsymbol{x}}^t\cdot\vec{Y}-n\,\overline{Y}_\bullet\cdot\vec{\tilde{\boldsymbol{x}}}_\bullet)\end{aligned}$$

und

$$(\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t)^{-1} = ((\,\mathbf{0}\;\vert\;\mathbf{E}\,)\cdot\left(\begin{array}{c|c} 1/n+\bar{\vec{x}}_\blacksquare^{\,t}\cdot\mathbf{S}^{-1}\cdot\bar{\vec{x}}_\blacksquare & -\bar{\vec{x}}_\blacksquare^{\,t}\cdot\mathbf{S}^{-1} \\ \hline -\mathbf{S}^{-1}\cdot\bar{\vec{x}}_\blacksquare & \mathbf{S}^{-1}\end{array}\right)\cdot(\,\mathbf{0}\;\vert\;\mathbf{E}\,)^t)^{-1} = \mathbf{S}$$

offenbar

$$\hat{\bar{\mathrm{B}}}^t\cdot\mathbf{G}^t\cdot(\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t)^{-1}\cdot\mathbf{G}\cdot\hat{\bar{\mathrm{B}}} = (\tilde{\boldsymbol{x}}^t\cdot\bar{Y}-n\,\bar{Y}_\blacksquare\cdot\bar{\vec{x}}_\blacksquare)^t\cdot\mathbf{S}^{-1}\cdot\mathbf{S}\cdot\mathbf{S}^{-1}\cdot(\tilde{\boldsymbol{x}}^t\cdot\bar{Y}-n\,\bar{Y}_\blacksquare\cdot\bar{\vec{x}}_\blacksquare) =$$

$$= n^2\,\bar{Y}_\blacksquare^{\,2}\,\bar{\vec{x}}_\blacksquare^{\,t}\cdot\mathbf{S}^{-1}\cdot\bar{\vec{x}}_\blacksquare - 2\,n\,\bar{Y}_\blacksquare\,\bar{Y}^t\cdot\tilde{\boldsymbol{x}}\cdot\mathbf{S}^{-1}\cdot\bar{\vec{x}}_\blacksquare + \bar{Y}^t\cdot\tilde{\boldsymbol{x}}\cdot\mathbf{S}^{-1}\cdot\tilde{\boldsymbol{x}}^t\cdot\bar{Y}$$

b) Aus dem Beweis von Satz 12.2.8 folgt unmittelbar

$$\mathrm{SSE} = \sum_{i=1}^n (Y_i-\hat{Y}_i)^2 = (\bar{Y}-\boldsymbol{x}\cdot\hat{\bar{\mathrm{B}}})^t\cdot(\bar{Y}-\boldsymbol{x}\cdot\hat{\bar{\mathrm{B}}}) = \bar{Y}^t\cdot(\mathbf{E}-\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t)\cdot\bar{Y} = n\,\hat{\Sigma}^2$$

c) Durch einfaches Ausmultiplizieren ergibt sich

$$\mathrm{SStotal} = \sum_{i=1}^n (Y_i-\bar{Y}_\blacksquare)^2 = (\bar{Y}-\mathbf{I}\cdot\bar{Y})^t\cdot(\bar{Y}-\mathbf{I}\cdot\bar{Y}) = \bar{Y}^t\cdot(\mathbf{E}-\mathbf{I})\cdot\bar{Y}$$

d) Außerdem gilt wegen (man beachte, dass $\mathbf{I}\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t = \boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\mathbf{I} = \mathbf{I}$ gilt)

$$\sum_{i=1}^n (Y_i-\bar{Y}_\blacksquare)(\hat{Y}_i-\bar{\hat{Y}}_\blacksquare) = (\bar{Y}-\mathbf{I}\cdot\bar{Y})^t\cdot(\boldsymbol{x}\cdot\hat{\bar{\mathrm{B}}}-\mathbf{I}\cdot\boldsymbol{x}\cdot\hat{\bar{\mathrm{B}}}) =$$

$$= \bar{Y}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\bar{Y} - \bar{Y}^t\cdot\mathbf{I}\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\bar{Y} = \bar{Y}^t\cdot\left(\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t-\mathbf{I}\right)\cdot\bar{Y} = \mathrm{SSR}$$

und

$$\sum_{i=1}^n (\hat{Y}_i-\bar{\hat{Y}}_\blacksquare)^2 = (\boldsymbol{x}\cdot\hat{\bar{\mathrm{B}}}-\mathbf{I}\cdot\boldsymbol{x}\cdot\hat{\bar{\mathrm{B}}})^t\cdot(\boldsymbol{x}\cdot\hat{\bar{\mathrm{B}}}-\mathbf{I}\cdot\boldsymbol{x}\cdot\hat{\bar{\mathrm{B}}}) =$$

$$= \bar{Y}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\bar{Y} - \bar{Y}^t\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\mathbf{I}\cdot\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t\cdot\bar{Y} =$$

$$= \bar{Y}^t\cdot\left(\boldsymbol{x}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\boldsymbol{x}^t-\mathbf{I}\right)\cdot\bar{Y} = \mathrm{SSR}$$

schließlich

$$R^2 = \frac{(\sum_{i=1}^n (Y_i-\bar{Y}_\blacksquare)(\hat{Y}_i-\bar{\hat{Y}}_\blacksquare))^2}{\sum_{i=1}^n (Y_i-\bar{Y}_\blacksquare)^2\sum_{i=1}^n (\hat{Y}_i-\bar{\hat{Y}}_\blacksquare)^2} = \frac{\sum_{i=1}^n (Y_i-\bar{Y}_\blacksquare)(\hat{Y}_i-\bar{\hat{Y}}_\blacksquare)}{\sum_{i=1}^n (Y_i-\bar{Y}_\blacksquare)^2} = \frac{\mathrm{SSR}}{\mathrm{SStotal}}$$

Aus diesem Satz folgt unmittelbar

12.3.3 Bemerkung: Es gilt wieder die als Zerlegung der Varianz bekannte Formel

$$\mathrm{SStotal} = \mathrm{SSR} + \mathrm{SSE}$$

und damit $0 \le R^2 \le 1$. Die Statistik R^2 ist daher ein Maß für den relativen Anteil der Schwankung des Zufallsvektors $\{Y_1, Y_2, \ldots, Y_n\}$, welcher durch die Regression erklärt werden kann. Ist $R^2 = 0$, so besteht keinerlei linearer Zusammenhang zwischen den Einflussgrößen und der Zielgröße. Ist hingegen $R^2 = 1$, so besteht zwischen den Einflussgrößen und der Zielgröße ein perfekter linearer Zusammenhang; dieser Fall liegt genau dann vor, wenn alle Punkte $\{\vec{x}_1, y_1\}, \{\vec{x}_2, y_2\}, \ldots, \{\vec{x}_n, y_n\}$ auf der Regressionshyperebene liegen.

Die beiden folgenden Sätze und die anschließende Bemerkung bilden die Grundlage für die im nächsten Abschnitt zu behandelnden Tests von Hypothesen über die unbekannten Parameter $\beta_0\,\beta_1, \ldots\beta_m$:

12.3.4 Satz: Die Statistiken SSR und SSE besitzen die folgenden Eigenschaften:
a) Die beiden Statistiken SSR und SSE sind unabhängig;
b) Die Statistik SSE/σ^2 ist $\mathcal{C}\text{hi}[n-(m+1)]$-verteilt;
c) Im Fall $\beta_1 = \beta_2 = \ldots = \beta_m = 0$ genügt die Statistik SSR/σ^2 einer $\mathcal{C}\text{hi}[m]$-Verteilung.

▼

Beweis: Die Statistiken SSR bzw SSE sind Funktionen der Schätzer $\hat{\vec{B}} = \{\hat{B}_0, B_1, \ldots, B_1\}^t$ bzw $\hat{\Sigma}^2$ und somit wegen Satz 12.2.8 unabhängig. Die Tatsache, dass die Statistik SSE/σ^2 einer $\mathcal{C}\text{hi}[n-(m+1)]$-Verteilung genügt, folgt unmittelbar aus Satz 12.3.2 und Satz 12.2.8. Somit bleibt lediglich die Aussage c) zu zeigen:

Wegen Satz 12.2.8 ist der Zufallsvektor $\mathbf{G}\cdot\hat{\vec{B}}$ bekanntlich multinormalverteilt mit dem Mittelwertsvektor $\mathbf{G}\cdot\vec{\beta}$ und der Kovarianzmatrix $\sigma^2\,\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t$. Nun ist aber die Matrix $\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t$ symmetrisch und (als Kovarianzmatrix einer Multinormalverteilung) positiv definit. Also existiert eine orthogonale Matrix $\mathbf{P}$ mit

$$\mathbf{P}\cdot\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t\cdot\mathbf{P}^t = \text{Diag}[\lambda_1, \lambda_2, \ldots, \lambda_m]$$

wobei $\lambda_1, \lambda_2, \ldots, \lambda_m$ die Eigenwerte der Matrix $\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t$ bezeichnen (man beachte, dass eine positiv definite Matrix lauter positive Eigenwerte besitzt). Im Fall $\beta_1 = \beta_2 = \ldots = \beta_m = 0$ ist daher der Zufallsvektor

$$\vec{S} = \{S_1, S_2, \ldots, S_m\}^t = \frac{1}{\sigma}\,\text{Diag}\Big[\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \ldots, \frac{1}{\sqrt{\lambda_m}}\Big]\cdot\mathbf{P}\cdot\mathbf{G}\cdot\hat{\vec{B}}$$

multinormalverteilt mit dem Mittelwertsvektor

$$\frac{1}{\sigma}\,\text{Diag}\Big[\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \ldots, \frac{1}{\sqrt{\lambda_m}}\Big]\cdot\mathbf{P}\cdot\mathbf{G}\cdot\vec{\beta} = \vec{0}$$

und der Kovarianzmatrix

$$\text{Diag}\Big[\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \ldots, \frac{1}{\sqrt{\lambda_m}}\Big]\cdot\mathbf{P}\cdot\mathbf{G}\cdot(x^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t\cdot\mathbf{P}^t\cdot\text{Diag}\Big[\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \ldots, \frac{1}{\sqrt{\lambda_m}}\Big] = \mathbf{E}$$

Damit gilt

$$\text{SSR}/\sigma^2 = \frac{1}{\sigma^2}\,\hat{\vec{B}}^t\cdot\mathbf{G}^t\cdot(\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t)^{-1}\cdot\mathbf{G}\cdot\hat{\vec{B}} =$$

$$= \frac{1}{\sigma^2}\,\hat{\vec{B}}^t\cdot\mathbf{G}^t\cdot\mathbf{P}^t\cdot\text{Diag}[1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_m]\cdot\mathbf{P}\cdot\mathbf{G}\cdot\hat{\vec{B}} = \vec{S}^t\cdot\vec{S}$$

womit gezeigt ist, dass die Statistik SSR/σ^2 als Summe der Quadrate der vollständig unabhängigen, $\mathcal{N}[0, 1]$-verteilten Zufallsvariablen $S_1, S_2, \ldots, S_m$ einer $\mathcal{C}\text{hi}[m]$-Verteilung genügt.

Vollständig analog dazu beweist man die folgende Verallgemeinerung dieses Satzes:

12.3.5 Satz: Ist $\mathbf{H} \in \mathbb{R}_r^{m+1}$ eine beliebige Matrix mit der Eigenschaft, dass $\mathbf{H}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{H}^t$ invertierbar ist und bezeichnet $\mathrm{SSH} = \hat{\vec{\mathrm{B}}}^t\cdot\mathbf{H}^t\cdot\left(\mathbf{H}\cdot\left(\boldsymbol{x}^t\cdot\boldsymbol{x}\right)^{-1}\cdot\mathbf{H}^t\right)^{-1}\cdot\hat{\vec{\mathrm{B}}}$, so gilt:

a) Die beiden Statistiken SSH und SSE sind unabhängig;

b) Die Statistik SSE/σ^2 ist $\mathcal{C}\mathrm{hi}[n-(m+1)]$-verteilt;

c) Im Fall $\mathbf{H}\cdot\vec{\beta} = \vec{0}$ genügt die Statistik SSH/σ^2 einer $\mathcal{C}\mathrm{hi}[r]$-Verteilung.

Die folgende Bemerkung ist in gewissem Sinn ein Spezialfall von Satz 12.3.5:

12.3.6 Bemerkung: Ist $\vec{h} \in \mathbb{R}^{m+1}$ mit $\vec{h} \neq \vec{0}$ und bezeichnet $\mathrm{SSh} = \vec{h}\cdot\hat{\vec{\mathrm{B}}}/\sqrt{\vec{h}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\vec{h}^t}$, so gilt:

a) Die beiden Statistiken SSh und SSE sind unabhängig;

b) Die Statistik SSE/σ^2 ist $\mathcal{C}\mathrm{hi}[n-(m+1)]$-verteilt;

c) Im Fall $\vec{h}\cdot\vec{\beta} = 0$ genügt die Statistik SSh/σ einer $\mathcal{N}[0, 1]$-Verteilung.

12.4 Tests von Hypothesen über die Parameter $\beta_0, \beta_1, \ldots, \beta_m$

Gegeben seien die n Messwerte $\{\vec{x}_1, y_1\}, \{\vec{x}_2, y_2\}, \ldots, \{\vec{x}_n, y_n\} \in \mathbb{R}^{m+1}$, wobei wir voraussetzen, dass die Matrix $\boldsymbol{x}^t\cdot\boldsymbol{x} \in \mathbb{R}_{m+1}^{m+1}$ den Rang $m+1$ besitzt, wir es also mit einer linearen Regressionsanalyse mit vollem Rang zu tun haben. Wir befassen uns in diesem Abschnitt damit, ausgehend von diesen Messwerten, Hypothesen über die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ des Modells

$$\vec{Y} = \boldsymbol{x}\cdot\vec{\beta} + \vec{E}$$

zu testen.

Zu Beginn befassen wir uns mit der Frage, ob die m Einflussgrößen insgesamt einen Einfluss auf die Zielgröße besitzen. Diese Frage lässt sich mit einem Test für die Hypothese $\mathcal{H}_0 \ldots \{\beta_1, \beta_2, \ldots, \beta_m\} = \{0, 0, \ldots, 0\}$ gegen die Alternative $\mathcal{H}_1 \ldots \{\beta_1, \beta_2, \ldots, \beta_m\} \neq \{0, 0, \ldots, 0\}$ beantworten (der Einfluss der Konstanten wird dabei nicht berücksichtigt):

12.4.1 Der vollständige Regressionstest: Der vollständige Regressionstest wird verwendet, wenn getestet werden soll, ob die m Einflussgrößen insgesamt einen Einfluss auf die Zielgröße besitzen:

$\mathfrak{P}_{\vec{Y}}$	$\mathcal{H}_0$	$\mathcal{H}_1$	Ablehnungsbereich
$\{\mathcal{MN}[\boldsymbol{x}\cdot\vec{\beta}, \sigma^2\,\mathbf{E}] \mid \vec{\beta}\in\mathbb{R}_{m+1},\ \sigma>0\}$	$\{\beta_1, \beta_2, \ldots, \beta_m\} = \vec{0}$ $\sigma > 0$	$\{\beta_1, \beta_2, \ldots, \beta_m\} \neq \vec{0}$ $\sigma > 0$	$\frac{\mathrm{SSR}}{\mathrm{SSE}}\ \frac{n-(m+1)}{m} > f_{m,n-(m+1);1-\alpha}$

Dabei bezeichnet $f_{m,n;q}$ das q-Quantil der $\mathcal{F}[m, n]$-Verteilung.

▼

Beweis: Falls die Hypothese $\mathcal{H}_0$ zutrifft, also $\beta_1 = \beta_2 = \ldots = \beta_m = 0$ ist, so wird sich $\mathbf{G}\cdot\hat{\vec{\mathrm{B}}}$ vom Nullvektor $\vec{0}$ nur wenig unterscheiden und damit

$$\mathrm{SSR} = \hat{\vec{\mathrm{B}}}^t\cdot\mathbf{G}^t\cdot\left(\mathbf{G}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{G}^t\right)^{-1}\cdot\mathbf{G}\cdot\hat{\vec{\mathrm{B}}}$$

"klein" sein, wobei dieses "klein sein" in der üblichen Weise noch durch SSE zu relativieren ist. Dass es sich bei einem Test mit dem oben angeführten Ablehnungsbereich tatsächlich um einen Test für die Hypothese $\mathcal{H}_0$ gegen die Alternative $\mathcal{H}_1$ mit Signifikanzniveau α handelt ist offenbar, wenn man berücksichtigt, dass die Statistik

$$\frac{\text{SSR}}{\text{SSE}}\ \frac{n-(m+1)}{m}$$

wegen Satz 12.3.4 und Satz 23.5.1 bei Zutreffen der Hypothese $\mathcal{H}_0$ einer $\mathcal{F}[m, n-(m+1)]$-Verteilung genügt.

Viel wichtiger als die Frage, ob die m Einflussgrößen insgesamt einen Einfluss auf die Zielgröße besitzen, ist die Frage, ob die i-te Einflussgröße die Zielgröße beeinflusst. Diese Frage lässt sich mit einem Test für die Hypothese $\mathcal{H}_0 \dots \beta_i = 0$ gegen die Alternative $\mathcal{H}_1 \dots \beta_i \neq 0$ beantworten (im Fall $i = 0$ lässt sich damit auch prüfen, ob die Regressionshyperebene durch den Ursprung geht):

12.4.2 Der partielle Regressionstest: Der partielle Regressionstest wird verwendet, wenn getestet werden soll, ob die i-te Einflussgröße einen Einfluss auf die Zielgröße besitzt:

$\mathbb{P}_{\vec{Y}}$	$\mathcal{H}_0$	$\mathcal{H}_1$	Ablehnungsbereich
$\{\mathcal{MN}[\boldsymbol{x}\cdot\vec{\beta}, \sigma^2\,\mathbf{E}]\ \vert\ \vec{\beta}\in\mathbb{R}_{m+1},\ \sigma>0\}$	$\beta_i = 0$ $\sigma > 0$	$\beta_i \neq 0$ $\sigma > 0$	$\dfrac{\vert\hat{\text{B}}_i\vert}{\sqrt{\text{SSE}}\,\sqrt{\xi_i}}\ \sqrt{n-(m+1)} > t_{n-(m+1);1-\alpha/2}$

Dabei bezeichnet $t_{n;q}$ das q-Quantil der $\mathcal{T}[n]$-Verteilung und ξ_i den Eintrag der Matrix $(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\in\mathbb{R}^{m+1}_{m+1}$ am Schnittpunkt der i-ten Zeile mit der i-ten Spalte (i läuft dabei von 0 bis m).

▼

Beweis: Falls die Hypothese $\mathcal{H}_0$ zutrifft, also $\beta_i = 0$ ist, so wird $\hat{\text{B}}_i = \vec{h}\cdot\hat{\vec{\text{B}}}$ nur wenig von 0 abweichen (mit $\vec{h}$ bezeichnen wir dabei jenen Vektor aus $\mathbb{R}^{m+1}$, bei dem an der i-ten Stelle eine 1 steht und dessen übrigen Einträge gleich 0 sind). Dass es sich bei einem Test mit dem oben angeführten Ablehnungsbereich tatsächlich um einen Test für die Hypothese $\mathcal{H}_0$ gegen die Alternative $\mathcal{H}_1$ mit Signifikanzniveau α handelt wird offenbar, wenn man berücksichtigt, dass die Statistik

$$\frac{\hat{\text{B}}_i}{\sqrt{\text{SSE}}\,\sqrt{\xi_i}}\ \sqrt{n-(m+1)} = \frac{\text{SSh}}{\sqrt{\text{SSE}}}\ \sqrt{n-(m+1)}$$

auf Grund von Bemerkung 12.3.6 und Satz 23.4.1 bei Zutreffen der Hypothese $\mathcal{H}_0$ einer $\mathcal{T}[n-(m+1)]$-Verteilung genügt.

Sowohl beim vollständigen Regressionstest als auch beim partiellen Regressionstest handelt es sich um Spezialfälle des Regressionstests, bei dem die Hypothese $\mathcal{H}_0 \dots \mathbf{H}\cdot\vec{\beta} = \vec{\zeta}$ gegen die Alternative $\mathcal{H}_1 \dots \mathbf{H}\cdot\vec{\beta} \neq \vec{\zeta}$ getestet wird. Dabei ist $\mathbf{H}\in\mathbb{R}^{m+1}_r$ eine beliebige Matrix mit der Eigenschaft, dass $\mathbf{H}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{H}^t$ invertierbar ist und $\vec{\zeta}\in\mathbb{R}_r$ ein beliebiger Vektor:

12.4.3 Der allgemeine Regressionstest: Der allgemeine Regressionstest wird verwendet, wenn getestet werden soll, ob die unbekannten Parameter $\vec{\beta}$ der Beziehung $\mathbf{H}\cdot\vec{\beta} = \vec{\zeta}$ genügen. Dabei ist $\mathbf{H}\in\mathbb{R}^{m+1}_r$ eine beliebige Matrix mit der Eigenschaft, dass $\mathbf{H}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{H}^t$ invertierbar ist und $\vec{\zeta}\in\mathbb{R}_r$ ein beliebiger Vektor:

$\mathbb{P}_{\vec{Y}}$	$\mathcal{H}_0$	$\mathcal{H}_1$	Ablehnungsbereich
$\{\mathcal{MN}[\boldsymbol{x}\cdot\vec{\beta}, \sigma^2\,\mathbf{E}]\ \vert\ \vec{\beta}\in\mathbb{R}_{m+1},\ \sigma>0\}$	$\mathbf{H}\cdot\vec{\beta} = \vec{\zeta}$ $\sigma > 0$	$\mathbf{H}\cdot\vec{\beta} \neq \vec{\zeta}$ $\sigma > 0$	$\dfrac{\Vert\mathbf{H}\cdot\hat{\vec{\text{B}}} - \vec{\zeta}\Vert}{\text{SSE}}\ \dfrac{n-(m+1)}{r} > f_{r,n-(m+1);1-\alpha}$

Dabei bezeichnet $f_{m,n;q}$ das q-Quantil der $\mathcal{F}[m, n]$-Verteilung und $\Vert\mathbf{H}\cdot\hat{\vec{\text{B}}} - \vec{\zeta}_0\Vert$ die Statistik

$$\Vert\mathbf{H}\cdot\hat{\vec{\text{B}}} - \vec{\zeta}\Vert = (\mathbf{H}\cdot\hat{\vec{\text{B}}} - \vec{\zeta})^t\cdot(\mathbf{H}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{H}^t)^{-1}\cdot(\mathbf{H}\cdot\hat{\vec{\text{B}}} - \vec{\zeta})$$

▼

Beweis: Wir beschränken uns auf den Fall $\vec{\zeta} = \vec{0}$. Falls die Hypothese zutrifft, also $\mathbf{H}\cdot\vec{\beta} = \vec{0}$ ist, so wird $\mathbf{H}\cdot\hat{\vec{\text{B}}}$ nur

wenig von $\vec{0}$ abweichen und damit

$$||\mathbf{H}\cdot\hat{\vec{B}}|| = \hat{\vec{B}}^{t}\cdot\mathbf{H}^{t}\cdot(\mathbf{H}\cdot(\boldsymbol{x}^{t}\cdot\boldsymbol{x})^{-1}\cdot\mathbf{H}^{t})^{-1}\cdot\mathbf{H}\cdot\hat{\vec{B}}$$

"klein" sein, wobei dieses "klein sein" in der üblichen Weise wieder durch SSE zu relativieren ist. Dass es sich bei dem Test mit dem oben angeführten Ablehnungsbereich tatsächlich um einen Test für die Hypothese $\mathcal{H}_0$ gegen die Alternative $\mathcal{H}_1$ mit Signifikanzniveau α handelt wird offensichtlich, wenn man berücksichtigt, dass die Testgröße

$$\frac{||\mathbf{H}\cdot\hat{\vec{B}}||}{\mathrm{SSE}}\,\frac{n-(m+1)}{r} = \frac{\mathrm{SSH}}{\mathrm{SSE}}\,\frac{n-(m+1)}{r}$$

wegen Satz 12.3.5 und Satz 23.5.1 bei Zutreffen der Hypothese $\mathcal{H}_0$ einer $\mathcal{F}[r, n-(m+1)]$-Verteilung genügt.

Will man nicht nur die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$ schätzen (dazu dient der Befehl Fit) sondern auch Hypothesen über diese Parameter testen, so lade man zuerst das Paket LinearRegression` und verwende den Befehl Regress. Man beachte dabei, dass die Syntax des Befehls Regress mit der Syntax des Befehls Fit übereinstimmt (soll der Einfluss der Konstanten nicht berücksichtigt werden, so ist jedoch die Option IncludeConstant → False zu verwenden).

<< LinearRegression`

ladet das Paket LinearRegression`.

■ Regress[*daten*, *model*, *vars*]

führt für das Datenmaterial *daten* eine Regressionsanalyse durch.

Das Datenmaterial *daten* muss dabei die Form $\{\{x_{11}, x_{12}, \ldots, x_{1s}, y_1\}, \{x_{21}, x_{22}, \ldots, x_{2s}, y_2\}, \ldots\}$ besitzen. Mit der Liste *model* lässt sich steuern, welche Funktionen der Variablen als Einflussgrößen in das Modell aufgenommen werden sollen. In der Liste *vars* werden alle im Datenmaterial *daten* aufscheinenden Variablen aufgelistet.

Der von *Mathematica* gelieferte Output besteht aus drei Blöcken:

i) Im ersten Block werden die in *model* angeführten m Einflussgrößen, die Schätzwerte (Estimate) für die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$, die Schätzwerte der im partiellen Regressionstest verwendeten Statistiken $\sqrt{\mathrm{SSE}}\,\sqrt{\xi_i}$ (SE), die dort verwendeten Testgrößen (TStat) und die zugehörigen p-Werte (PValue) ausgegeben.

ii) Im zweiten Block findet man die Schätzwerte der Statistiken R^2 und R^2_{ad} sowie den Schätzwert der unbekannten Varianz σ^2.

iii) Im dritten Block werden die Freiheitsgrade (DF) und die Schätzwerte der Statistiken SSR, SSE und SStotal (SumOfSq) sowie SSR/m bzw $\mathrm{SSE}/(n-(m+1))$ (MeanSq) und die Testgröße für den vollständigen Regressionstest (FRatio) zusammen mit dem zugehörigen p-Wert (PValue) ausgegeben.

Wir demonstrieren die Regressionsanalyse und die Verwendung des Befehls Regress an konkreten Beispielen:

12.4.4 Beispiel: Von $n = 548$ Stahlblechen wurde die Dicke, die Haspeltemperatur, der Mangangehalt, und die Zugfestigkeit gemessen (man vergleiche dazu das Datenmaterial zugfestigkeit). Man prüfe, ob die Zugfestigkeit tatsächlich signifikant von der Dicke, der Haspeltemperatur und dem Mangangehalt abhängt und wie gut sich die Zugfestigkeit durch diese Einflussgrößen beschreiben lässt.

▼

Lösung: In Beispiel 12.2.4 haben wir unter Verwendung von Fit die unbekannten Parameter $\beta_0, \beta_1, \beta_2, \beta_3$ geschätzt. Wir wenden nun auf unser Datenmaterial den Befehl Regress an und erhalten neben den (bereits bekannten) Schätzwerten für die unbekannten Parameter $\beta_0, \beta_1, \beta_2, \beta_3$ einen Schätzwert für die unbekannte Varianz σ^2

sowie Aussagen darüber, ob die Einflussgrößen Dicke, Haspeltemperatur bzw Mangangehalt die Zielgröße Zugfestigkeit tatsächlich beeinflussen und wie gut sich die Zugfestigkeit durch diese Einflussgrößen beschreiben lässt:

```
zugfestigkeit = Rest[<< zugfestigkeitfile];
Regress[zugfestigkeit, {1, Dicke, Haspeltemperatur, Mangangehalt}, {Dicke, Haspeltemperatur, Mangangehalt}]
Clear[zugfestigkeit]
```

```
{ParameterTable →
```

	Estimate	SE	TStat	PValue
1	532.316	6.22622	85.4958	$1.494818580744021\times10^{-317}$
Dicke	−3.48119	0.277612	−12.5398	7.27965×10^{-32}
Haspeltemperatur	−0.243508	0.0093313	−26.0959	5.90498×10^{-98}
Mangangehalt	28.8377	3.96827	7.26707	1.28253×10^{-12}

```
,
RSquared → 0.731334, AdjustedRSquared → 0.729852, EstimatedVariance → 31.5278,
```

ANOVATable →		DF	SumOfSq	MeanSq	FRatio	PValue
	Model	3	46 687.	15 562.3	493.606	0.
	Error	544	17 151.1	31.5278		
	Total	547	63 838.1			

```
}
```

Die p-Werte aller Einflussgrößen sind extrem klein, also beeinflusst jede dieser Einflussgrößen die Zugfestigkeit sehr signifikant. Der Schätzwert $\hat{R}^2 = 0.731334$ der Statistik R^2 zeigt, dass sich die Zugfestigkeit durch diese Einflussgrößen bereits recht gut beschreiben lässt, dass es aber neben diesen Einflussgrößen noch weitere (uns unbekannte) Faktoren geben wird, welche die Zugfestigkeit beeinflussen. Die tatsächlichen Zugfestigkeiten streuen dabei wegen $\hat{\sigma}^2 = 31.5278$ in $\mathcal{N}[0, 5.61496]$-verteilter Weise um die Regressionshyperebene

$$\text{Zugfestigkeit} = 532.316 - 3.48119\ \text{Dicke} - 0.243508\ \text{Haspeltemperatur} + 28.8377\ \text{Mangangehalt}$$

Schließlich entnimmt man der ANOVA-Tabelle, dass die drei Einflussgrößen Dicke, Haspeltemperatur und Mangangehalt (der entsprechende p-Wert ist 0) insgesamt einen überaus deutlichen Einfluss auf die Zugfestigkeit besitzen.

12.4.5 Beispiel: Von $n = 30$ zufällig ausgewählten Personen wurde das Alter und der systolische Blutdruck ermittelt (man vergleiche dazu das Datenmaterial blutdruck). In Beispiel 12.2.5 haben wir für die beiden Modelle Blutdruck $= \alpha_0 + \alpha_1$ Alter bzw Blutdruck $= \beta_0 + \beta_1$ Alter $+ \beta_2$ Alter^2 die Parameter α_0 und α_1 bzw β_0, β_1 und β_2 geschätzt. Man prüfe nun, ob sowohl die Einflussgröße Alter als auch die Einflussgröße Alter^2 den Blutdruck tatsächlich signifikant beeinflussen und vergleiche außerdem die Güte dieser Modelle.

▼

Lösung: Wir analysieren dieses Datenmaterial auf mehrere Arten mit Hilfe von Regress:

a) Modell Blutdruck $= \alpha_0 + \alpha_1$ Alter: Das Alter hat einen sehr signifikanten Einfluss ($p = 4.39873\times10^{-11}$) auf den Blutdruck. Der Blutdruck lässt sich wegen $\hat{R}^2_{\text{ad}} = 0.785656$ recht gut durch das Alter allein beschreiben:

```
blutdruck = Rest[<< blutdruckfile];
Regress[blutdruck, {1, Alter}, {Alter}]
```

```
{ParameterTable → 
                  |Estimate  SE        TStat    PValue
              1   |89.6724   5.07031   17.6858  0.
              Alter|1.10401   0.106581  10.3584  4.39873×10^-11
 ,
 RSquared → 0.793047, AdjustedRSquared → 0.785656, EstimatedVariance → 77.0566,
                    DF   SumOfSq    MeanSq    FRatio    PValue
 ANOVATable → Model  1    8267.91   8267.91   107.297   4.39873×10^-11}
              Error  28   2157.59   77.0566
              Total  29   10 425.5
```

b) Modell $\text{Blutdruck} = \beta_0 + \beta_1\ \text{Alter} + \beta_2\ \text{Alter}^2$: Während die Einflussgröße Alter^2 einen sehr signifikanten Einfluss ($p = 0.000964314$) auf den Blutdruck hat, besitzt die Einflussgröße Alter nun aber keinen signifikanten Einfluss ($p = 0.173698$) auf den Blutdruck. Der Blutdruck lässt sich wegen $R^2_{\text{ad}} = 0.852599$ sehr gut durch die beiden Einflussgrößen Alter und Alter^2 beschreiben:

```
Regress[blutdruck, {1, Alter, Alter^2}, {Alter}]
```

```
{ParameterTable → 
                  |Estimate   SE          TStat     PValue
              1   |123.734    10.1127     12.2356   1.58851×10^-12
              Alter|-0.686878  0.491572    -1.39731  0.173698
              Alter^2|0.0206646 0.00557969  3.70354   0.000964314
 ,
 RSquared → 0.862764, AdjustedRSquared → 0.852599, EstimatedVariance → 52.9908,
                    DF   SumOfSq    MeanSq    FRatio    PValue
 ANOVATable → Model  2    8994.75   4497.37   84.8708   2.26885×10^-12}
              Error  27   1430.75   52.9908
              Total  29   10 425.5
```

c) Modell $\text{Blutdruck} = \gamma_0 + \gamma_1\ \text{Alter}^2$: Der Einfluss der Einflussgröße Alter^2 auf den Blutdruck ist sehr signifikant ($p = 3.5949 \times 10^{-13}$). Der Blutdruck lässt sich wegen $R^2_{\text{ad}} = 0.847584$ nahezu genau so gut allein durch die Einflussgröße Alter^2 beschreiben, wie durch die beiden Einflussgrößen Alter und Alter^2. Da man im Rahmen der Regressionsanalyse stets darauf achtet, die Anzahl der Einflussgrößen möglichst klein zu halten, ist dieses Modell daher zu bevorzugen:

```
Regress[blutdruck, {1, Alter^2}, {Alter}]
Clear[blutdruck]
```

```
{ParameterTable → 
                  |Estimate   SE          TStat    PValue
              1   |110.09     2.67519     41.1523  0.
              Alter^2|0.0129951 0.00102014 12.7385  3.5949×10^-13
 ,
 RSquared → 0.85284, AdjustedRSquared → 0.847584, EstimatedVariance → 54.7934,
                    DF   SumOfSq    MeanSq    FRatio    PValue
 ANOVATable → Model  1    8891.28   8891.28   162.269   3.5949×10^-13}
              Error  28   1534.22   54.7934
              Total  29   10 425.5
```

12.4.6 Beispiel: Von $n = 72$ zufällig der Produktion entnommenen Stahlblechen wurde unter anderem die Walzzeit, das Gewicht und die Länge ermittelt (man vergleiche dazu das Datenmaterial walzzeit). Die Abhängigkeit der Walzzeit von den beiden Einflussgrößen Gewicht und Länge ist zu analysieren.

▼

Lösung: Eine graphische Veranschaulichung mehrdimensionaler Daten ist nur bedingt möglich. Aber auch ohne graphische Veranschaulichung erkennt man unter Verwendung der Regressionsanalyse

```
walzzeit = Rest[<< walzzeitfile];
Regress[walzzeit[[All, {2, 3, 1}]], {1, Gewicht, Länge}, {Gewicht, Länge}]
Clear[walzzeit]
```

```
{ParameterTable →
                  |  Estimate     SE          TStat     PValue
 1                |  3.33922      0.191627    17.4256   0.
 Gewicht          | -0.205468     0.0185074   -11.102   5.25681×10^-17,
 Länge            | -0.00941399   0.00123885  -7.59898  1.08109×10^-10

 RSquared → 0.721915, AdjustedRSquared → 0.713855, EstimatedVariance → 0.0477123,

                     DF   SumOfSq    MeanSq       FRatio    PValue
 ANOVATable → Model  2    8.54651    4.27326      89.563    0.      }
              Error  69   3.29215    0.0477123
              Total  71   11.8387
```

dass sowohl das Gewicht als auch die Länge einen äußerst signifikanten Einfluss auf die Walzzeit haben, dass sich die Walzzeit wegen $\hat{R}^2 = 0.721915$ durch das Gewicht und die Länge allein relativ gut beschreiben lässt, es aber noch weitere Einflussgrößen geben muss, welche auf die Walzzeit einen Einfluss haben und dass wegen $\hat{\sigma}^2 = 0.0477123$ die Walzzeiten in $\mathcal{N}[0, 0.218431]$-verteilter Weise um die Regressionshyperebene

$$\text{Walzzeit} = 3.33922 - 0.205468\ \text{Gewicht} - 0.00941399\ \text{Länge}$$

streuen.

Der allgemeine Regressionstest ist in *Mathematica* aber nicht standardmäßig implementiert; er lässt sich aber mit dem folgenden Befehl aufrufen:

■ RegressTest[*daten*, *model*, *vars*, $\mathbf{H}$, $\vec{\zeta}$]
berechnet für das Datenmaterial *daten*, das Modell *model* und die Variablen *vars* den *p*-Wert des Regressionstests für die Hypothese $\mathcal{H}_0 \ldots \mathbf{H}\cdot\vec{\beta} = \vec{\zeta},\ \sigma > 0$ gegen die Alternative $\mathcal{H}_1 \ldots \mathbf{H}\cdot\vec{\beta} \neq \vec{\zeta},\ \sigma > 0$.

Das Datenmaterial *daten* muss dabei die Form $\{\{x_{11}, x_{12}, \ldots, x_{1\,s}, y_1\}, \{x_{21}, x_{22}, \ldots, x_{2\,s}, y_2\}, \ldots\}$ besitzen. Mit der Liste *model* lässt sich steuern, welche Funktionen der Variablen als Einflussgrößen in das Modell aufgenommen werden sollen. In der Liste *vars* werden alle im Datenmaterial *daten* aufscheinenden Variablen aufgelistet. Weiters ist $\mathbf{H} \in \mathbb{R}_r^{m+1}$ eine beliebige Matrix mit der Eigenschaft, dass $\mathbf{H}\cdot(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1}\cdot\mathbf{H}^t$ invertierbar ist und $\vec{\zeta} \in \mathbb{R}_r$ ein beliebiger Vektor.

Wir erläutern den Regressionstest an einem konkreten Beispiel:

12.4.7 Beispiel: Zwischen der Zielgröße y und den beiden Einflussgrößen x_1 und x_2 besteht auf Grund von theoretischen Überlegungen die Beziehung

$$y = 2.5\, x_1^{-0.7}\, x_2^{-0.2}$$

Bei einem Experiment wurden $n = 18$ Datensätze ermittelt (man vergleiche dazu das Datenmaterial theorie). Widersprechen diese experimentell gewonnenen Werte der Theorie?

▼

Lösung: Durch Logarithmieren geht die obige Beziehung zwischen der Zielgröße y und den beiden Einflussgrößen x_1 und x_2 in die lineare Beziehung

$$\text{Log}[Y] = \text{Log}[2.5] - 0.7\,\text{Log}[x_1] - 0.2\,\text{Log}[x_2] + E$$

über, wobei wir voraussetzen, dass zwischen den logarithmierten Werten der additive, normalverteilte Fehler E auftritt. Wir haben somit die Hypothese $\mathcal{H}_0 \ldots \{\beta_0, \beta_1, \beta_2\} = \{\text{Log}[2.5], -0.7, -0.2\}$ gegen die Alternative

$\mathcal{H}_0 \ldots \{\beta_0, \beta_1, \beta_2\} \neq \{\text{Log}[2.5], -0.7, -0.2\}$ zu überprüfen. Diese Frage lässt sich mit dem Regressionstest leicht beantworten:

```
theorie = Log[Rest[<< theoriefile]];
RegressTest[theorie, {1, x1, x2}, {x1, x2}, {{1, 0, 0}, {0, 1, 0}, {0, 0, 1}}, {Log[2.5], −0.7, −0.2}]
Clear[theorie]
```

```
PValue → 0.927043
```

Die Hypothese $\mathcal{H}_0$ wird wegen $p = 0.927043$ angenommen; die experimentell gewonnenen Daten widersprechen daher der Theorie nicht.

12.5 Tests für den Vergleich zweier Regressions-Modelle

In diesem Abschnitt befassen wir uns mit zwei Tests, mit denen sich die Parameter ${}_1\vec{\beta} = \{{}_1\beta_0, {}_1\beta_1, \ldots, {}_1\beta_m\}$ und ${}_2\vec{\beta} = \{{}_2\beta_0, {}_2\beta_1, \ldots, {}_2\beta_m\}$ bzw die Parameter ${}_1\sigma$ und ${}_2\sigma$ der beiden Regressions-Modelle (der links unten stehende Index bezieht sich auf das Modell)

$${}_1\vec{Y} = {}_1\boldsymbol{x}\cdot{}_1\vec{\beta} + {}_1\vec{E} \qquad \text{bzw} \qquad {}_2\vec{Y} = {}_2\boldsymbol{x}\cdot{}_2\vec{\beta} + {}_2\vec{E}$$

miteinander vergleichen lassen. Von den die Messfehler und die zufälligen Einflüsse beschreibenden Zufallsvariablen ${}_1E_1, {}_1E_2, \ldots, {}_1E_{{}_1n}$ bzw ${}_2E_1, {}_2E_2, \ldots, {}_2E_{{}_2n}$ setzten wir voraus, dass sie vollständig unabhängig und $\mathcal{N}[0, {}_1\sigma]$-verteilt bzw $\mathcal{N}[0, {}_2\sigma]$-verteilt sind.

Wir befassen uns zuerst mit einem Test, mit dem sich die Parameter ${}_1\vec{\beta}$ und ${}_2\vec{\beta}$ miteinander vergleichen lassen:

12.5.1 Der Regressionstest für zwei Grundgesamtheiten: Der Regressionstest für zwei Grundgesamtheiten wird dann verwendet, wenn überprüft werden soll, ob die unbekannten Parameter ${}_1\vec{\beta}$ und ${}_2\vec{\beta}$ der Beziehung $\mathbf{H}\cdot{}_1\vec{\beta} = \mathbf{H}\cdot{}_2\vec{\beta}$ genügen, wobei vorausgesetzt wird, dass $\mathbf{H} \in \mathbb{R}_r^{m+1}$ eine beliebige Matrix ist, für die $\mathbf{H}\cdot(({}_1\boldsymbol{x}^t\cdot{}_1\boldsymbol{x})^{-1} + ({}_2\boldsymbol{x}^t\cdot{}_2\boldsymbol{x})^{-1}))\cdot\mathbf{H}^t$ invertierbar ist und die Streuungen ${}_1\sigma$ und ${}_2\sigma$ der beiden Modelle übereinstimmen:

$\mathcal{P}_{{}_1\vec{Y}\times\mathcal{P}\,{}_2\vec{Y}}$	$\mathcal{H}_0$	$\mathcal{H}_1$	Ablehnungsbereich
$\{\{\mathcal{MN}[{}_1\boldsymbol{x}\cdot{}_1\vec{\beta}, \sigma^2\,\mathbf{E}], \mathcal{MN}[{}_2\boldsymbol{x}\cdot{}_2\vec{\beta}, \sigma^2\,\mathbf{E}]\} \mid {}_1\vec{\beta}, {}_2\vec{\beta} \in \mathbb{R}_{m+1}, \sigma > 0\}$	$\mathbf{H}\cdot{}_1\vec{\beta} = \mathbf{H}\cdot{}_2\vec{\beta}$ $\sigma > 0$	$\mathbf{H}\cdot{}_1\vec{\beta} \neq \mathbf{H}\cdot{}_2\vec{\beta}$ $\sigma > 0$	$\dfrac{\lVert\mathbf{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}})\rVert}{{}_1\text{SSE} + {}_2\text{SSE}} \quad \dfrac{n-2\,(m+1)}{r} > f_{r,n-2\,(m+1);1-\alpha}$

Dabei bezeichnet $f_{m,n;q}$ das q-Quantil der $\mathcal{F}[m, n]$-Verteilung, $\lVert\mathbf{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}})\rVert$ die Statistik

$$\lVert\mathbf{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}})\rVert = (\mathbf{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}})^t\cdot(\mathbf{H}\cdot(({}_1\boldsymbol{x}^t\cdot{}_1\boldsymbol{x})^{-1} + ({}_2\boldsymbol{x}^t\cdot{}_2\boldsymbol{x})^{-1}))\cdot\mathbf{H}^t)^{-1}\cdot\mathbf{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}})$$

und $n = {}_1n + {}_2n$ die Summe der Stichprobenumfänge ${}_1n$ und ${}_2n$ der beiden Modelle.

▼

Beweis: Falls die Hypothese $\mathcal{H}_0$ zutrifft, also $\mathbf{H}\cdot{}_1\vec{\beta} = \mathbf{H}\cdot{}_2\vec{\beta}$ ist, so wird $\mathbf{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}})$ nur wenig von $\vec{0}$ abweichen und damit

$$\lVert\boldsymbol{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}})\rVert = (\boldsymbol{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}}))^t\cdot[\mathbf{H}\cdot(({}_1\boldsymbol{x}^t\cdot{}_1\boldsymbol{x})^{-1} + ({}_2\boldsymbol{x}^t\cdot{}_2\boldsymbol{x})^{-1})\cdot\mathbf{H}^t]^{-1}\cdot\mathbf{H}\cdot({}_1\hat{\vec{B}} - {}_2\hat{\vec{B}})$$

"klein" sein, wobei dieses "klein sein" in der üblichen Weise wieder zu relativieren ist. Dass es sich bei dem Test mit dem oben angeführten Ablehnungsbereich tatsächlich um einen Test für die Hypothese $\mathcal{H}_0$ gegen die Alternative $\mathcal{H}_1$ mit Signifikanzniveau α handelt wird klar, wenn man berücksichtigt, dass einerseits wegen Satz 12.2.8

$$\boldsymbol{H}\cdot({}_1\hat{\vec{B}}-{}_2\hat{\vec{B}})\approx\mathcal{MN}[\vec{0},\,\sigma^2\,\boldsymbol{H}\cdot(({}_1\boldsymbol{x}^t\cdot{}_1\boldsymbol{x})^{-1}+({}_2\boldsymbol{x}^t\,{}_2\boldsymbol{x})^{-1})\cdot\boldsymbol{H}^t]$$

gilt und damit die Zufallsvariable

$$||\boldsymbol{H}\cdot({}_1\hat{\vec{B}}-{}_2\hat{\vec{B}})||/\sigma^2=(\boldsymbol{H}\cdot({}_1\hat{\vec{B}}-{}_2\hat{\vec{B}}))t\cdot[\mathbf{H}\cdot(({}_1\boldsymbol{x}^t\cdot{}_1\boldsymbol{x})^{-1}+({}_2\boldsymbol{x}^t\cdot{}_2\boldsymbol{x})^{-1})\cdot\mathbf{H}^t]^{-1}\cdot\mathbf{H}\cdot({}_1\hat{\vec{B}}-{}_2\hat{\vec{B}})/\sigma^2$$

auf Grund des Satzes über die affine Transformation sowie Satz 23.3.1 $\mathcal{C}\text{hi}[r]$-verteilt ist und andererseits die davon offenbar unabhängige Zufallsvariable $({}_1\text{SSE}+{}_2\text{SSE})/\sigma^2$ wegen Satz 12.2.8 und der bekannten Faltungsformel einer Chi-Quadrat Verteilung mit ${}_1n-(m+1)+{}_2n-(m+1)=n-2\,(m+1)$ Freiheitsgraden genügt.

Auch der Regressionstest für zwei Grundgesamtheiten lässt sich leicht in *Mathematica* implementieren:

■ RegressDifferenceTest[*daten*1, *daten*2, *model*, *vars*, **H**]

berechnet für das Datenmaterial *daten*1 und *daten*2, das Modell *model* und die Variablen *vars* den *p*-Wert des Regressionstests für zwei Grundgesamtheiten für die Hypothese $\mathcal{H}_0\ldots\mathbf{H}\cdot{}_1\vec{\beta}=\mathbf{H}\cdot{}_2\vec{\beta},\,\sigma>0$ gegen die Alternative $\mathcal{H}_1\ldots\mathbf{H}\cdot{}_1\vec{\beta}\neq\mathbf{H}\cdot{}_2\vec{\beta},\,\sigma>0$.

Die Daten müssen wieder die übliche Form $\{\{{}_1x_{11},\,{}_1x_{12},\,\ldots,\,{}_1x_{1\,s},\,{}_1y_1\},\{{}_1x_{21},\,{}_1x_{22},\,\ldots,\,{}_1x_{2\,s},\,{}_1y_2\},\ldots\}$ bzw $\{\{{}_2x_{11},\,{}_2x_{12},\,\ldots,\,{}_2x_{1\,s},\,{}_2y_1\},\{{}_2x_{21},\,{}_2x_{22},\,\ldots,\,{}_2x_{2\,s},\,{}_2y_2\},\ldots\}$ besitzen. Mit der Liste *model* lässt sich steuern, welche Funktionen der Variablen in das Modell aufgenommen werden. In der Liste *vars* werden alle in den Daten aufscheinenden Variablen aufgelistet. Weiters muss die Matrix $\mathbf{H}\in\mathbb{R}_r^{m+1}$ die Eigenschaft besitzen, dass $\mathbf{H}\cdot(({}_1\boldsymbol{x}^t\cdot{}_1\boldsymbol{x})^{-1}+({}_2\boldsymbol{x}^t\cdot{}_2\boldsymbol{x})^{-1}))\cdot\mathbf{H}^t$ invertierbar ist.

Wir demonstrieren diesen Regressionstest für zwei Grundgesamtheiten wieder an konkreten Beispielen:

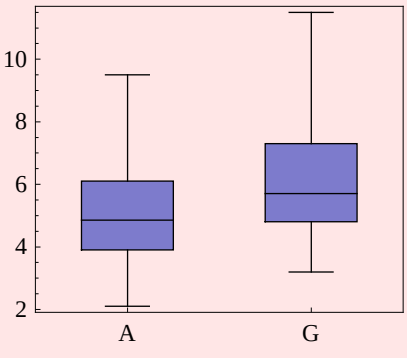
12.5.2 Beispiel: In einem Krankenhaus wurde das Geschlecht, das Alter und der Folatwert von Alzheimer-Patienten (Gruppe *A*) und von den auf Besuch weilenden, gesunden Angehörigen (Gruppe *G*) ermittelt (man vergleiche dazu das Datenmaterial alzheimer). Diese Daten sind zu analysieren

▼

Lösung: Wir demonstrieren mit diesem Beispiel, wie sich diese Datenanalyse tatsächlich abgespielt hat:

a) Mit Hilfe eines Box-Plots veranschaulichte sich der dieses Datenmaterial auswertende Arzt den Folatwert von Alzheimer-Patienten und deren Angehörigen:

```
alzheimer = Rest[<< "alzheimerfile"];
folatA = Select[alzheimer, #1[[2]] == A &][[All, 4]];
folatG = Select[alzheimer, #1[[2]] == G &][[All, 4]];
BoxWhiskerPlot[folatA, folatG, BoxLabels → {"A", "G"}, ImageSize → {150, 130}]
```



Dieses Box-Plot veranlasste den Arzt zu der Vermutung, dass Alzheimer-Patienten einen niedrigeren Folatwert

aufweisen, als gesunde Personen. Mit Hilfe eines t-Tests für zwei Grundgesamtheiten wurde diese Vermutung auch sehr signifikant bestätigt:

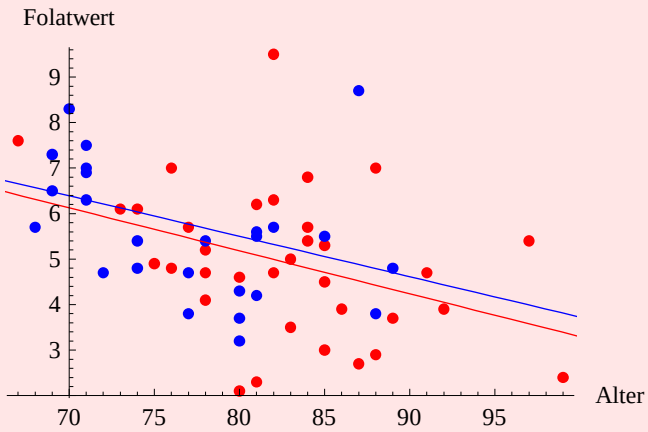
```
MeanDifferenceTest[folatA, folatG, 0, TwoSided → False]
```

```
OneSidedPValue → 0.00265945
```

Diese Analyse ließ den Arzt glauben, durch Verabreichung eines Folatmittels das Fortschreiten der Alzheimer-Krankheit verlangsamen zu können. Er wandte sich daher an einen Statistiker mit der Bitte, seine Analyse zu bestätigen.

b) Der Statistiker zeichnete ein Scatter-Plot, in dem er auf der x-Achse das Alter und auf der y-Achse den Folatwert sowohl der Alzheimer-Patienten (rot) als auch der gesunden Angehörigen (blau) zusammen mit den zugehörigen Regressionsgeraden einzeichnete:

```
alterfolatA = Select[alzheimer, #1[[2]] == A &][[All, {3, 4}]];
alterfolatG = Select[alzheimer, #1[[2]] == G &][[All, {3, 4}]];
plot1 = ListPlot[alterfolatA, PlotStyle → {PointSize[0.02], Red}];
line1 = Plot[Evaluate[Fit[alterfolatA, {1, x}, {x}]], {x, 50, 100}, PlotStyle → Red];
plot2 = ListPlot[alterfolatG, PlotStyle → {PointSize[0.02], Blue}];
line2 = Plot[Evaluate[Fit[alterfolatG, {1, x}, {x}]], {x, 50, 100}, PlotStyle → Blue];
Show[{plot1, plot2, line1, line2}, AxesLabel → {"Alter", "Folatwert"}]
Clear[plot1, plot2, line1, line2]
```



Diese Zeichnung zeigt deutlich, dass sowohl bei Alzheimer-Patienten als auch bei deren gesunden Angehörigen der Folatwert mit zunehmendem Alter abnimmt und somit der niedrigere Folatwert bei Alzheimer-Patienten auf deren höheres Alter zurückzuführen sein dürfte. Um letzte Zweifel zu beseitigen, zeigte der Statistiker schließlich noch mit Hilfe des Regressionstests für zwei Grundgesamtheiten

```
RegressDifferenceTest[alterfolatA, alterfolatG, {1, x}, {x}, {{1, 0}, {0, 1}}]
Clear[alzheimer, folatA, folatG, alterfolatA, alterfolatG]
```

```
PValue → 0.491417
```

dass sich die beiden Regressionsgeraden wegen $p = 0.491417$ nicht signifikant unterscheiden. Eine weitergehende Analyse (auf die wir hier aber nicht eingehen) zeigte außerdem keine geschlechtsspezifischen Unterschiede hinsichtlich des Zusammenhangs zwischen Alter und Folatwert.

12.5.3 Beispiel: Von $n = 548$ Stahlblechen wurde die Dicke, die Haspeltemperatur, der Mangangehalt, sowie die Zugfestigkeit gemessen (man vergleiche dazu das Datenmaterial zugfestigkeit sowie Beispiel 12.2.4). In Beispiel 12.4.4 haben wir uns zwar bereits mit der Frage befasst, wie gut sich die Zugfestigkeit durch die

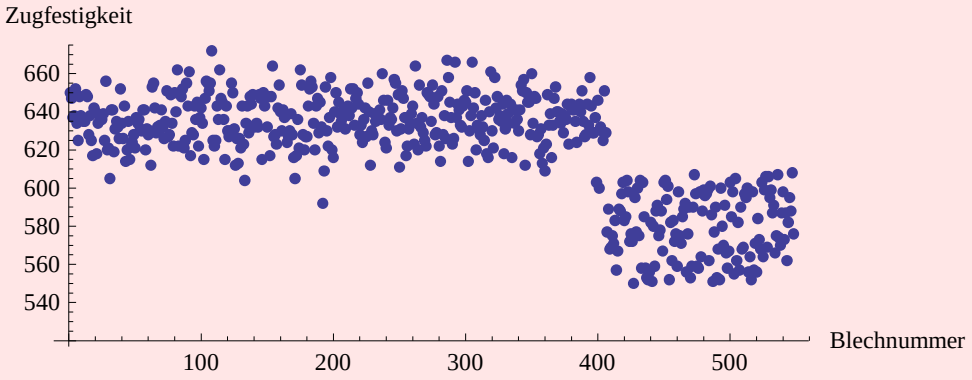
Dicke, die Haspeltemperatur und den Mangangehalt beschreiben lässt. Man analysiere dieses Datenmaterial nun aber genauer.

▼

Lösung: Auch mit diesem Beispiel zeigen wir, wie sich diese Datenanalyse einst tatsächlich abgespielt hat:

a) Es ist eine unter Statistikern wohlbekannte Erfahrungstatsache, dass jede statistische Datenanalyse damit beginnen sollte, das vorliegende Datenmaterial auf vielfältige Weise graphisch darzustellen. Diese Tatsache berücksichtigend, wurde mit Hilfe eines Plots, bei dem auf der x-Achse die laufende Blechnummer und auf der y-Achse die zugehörige Haspeltemperatur aufgetragen wurde, untersucht, ob die Haspeltemperatur von der Blechnummer abhängt. Diese - auf den ersten Blick sinnlose - Untersuchung lieferte das überraschende Ergebnis, dass ab dem Blech mit der laufenden Nummer $n_0 = 407$ die Haspeltemperatur in einem gänzlich anderen Bereich schwankt. Eine Rückfrage bei der Firma ergab, dass ab dieser Blechnummer bei der Ermittlung der Haspeltemperatur ein anderes Messverfahren verwendet wurde.

```
zugfestigkeit = Rest[<< "zugfestigkeitfile"];
ListPlot[zugfestigkeit[[All, 2]], AxesOrigin → {0, 520}, AspectRatio → 0.4, PlotStyle → PointSize[0.015],
  AxesLabel → {"Blechnummer", "Zugfestigkeit"}]
```



b) Die Frage, wie gut sich die Zugfestigkeit dieser Bleche durch die Dicke, die Haspeltemperatur und den Mangangehalt beschreiben lässt, läuft auf eine Regressionsanalyse hinaus. Da aber die Haspeltemperatur der einzelnen Bleche mit verschiedenen Messmethoden ermittelt wurde, erhebt sich die Frage, ob sowohl die ersten 406 Bleche als auch die letzten 142 Bleche dem gleichen Modell genügen. Diese Frage lässt sich mit dem Regressionstest für zwei Grundgesamtheiten leicht beantworten. Dabei zeigt sich, dass sich die für die ersten 406 Bleche zuständigen Parameter $\{{}_1\beta_1, {}_1\beta_2, {}_1\beta_3\}$ von den für die restlichen 142 Bleche zuständigen Parameter $\{{}_2\beta_1, {}_2\beta_2, {}_2\beta_3\}$ überaus deutlich unterscheiden:

```
daten1 = Take[zugfestigkeit, 406];
daten2 = Take[zugfestigkeit, −142];
RegressDifferenceTest[daten1, daten2, {1, Dicke, Haspeltemperatur, Mangangehalt},
  {Dicke, Haspeltemperatur, Mangangehalt}, {{0, 1, 0, 0}, {0, 0, 1, 0}, {0, 0, 0, 1}}]
```

```
PValue → 5.52525×10^-7
```

c) Die beiden Grundgesamtheiten sind daher getrennt zu analysieren, wobei sich nun aber (im Gegensatz zu dem in Beispiel 12.4.4 erzielten Resultat) zeigt, dass sich die Zugfestigkeit in beiden Fällen nur unzureichend durch die Dicke, die Haspeltemperatur und den Mangangehalt beschreiben lässt:

```
Regress[daten1, {1, Dicke, Haspeltemperatur, Mangangehalt}, {Dicke, Haspeltemperatur, Mangangehalt}][[2]]
Regress[daten2, {1, Dicke, Haspeltemperatur, Mangangehalt}, {Dicke, Haspeltemperatur, Mangangehalt}][[2]]
Clear[zugfestigkeit, daten1, daten2]
```

```
RSquared → 0.456219

RSquared → 0.284472
```

Wir befassen uns nun mit einem Test, mit dem sich prüfen lässt, ob die Streuungen ${}_1\sigma$ und ${}_2\sigma$ der beiden Modelle übereinstimmen:

12.5.4 Der Test von Goldfeld und Quandt: Der Test von Goldfeld und Quandt wird verwendet, wenn überprüft werden soll, ob die unbekannten Parameter ${}_1\sigma$ und ${}_2\sigma$ übereinstimmen:

$\mathcal{P}_{{}_1\vec{Y}} \times \mathcal{P}_{{}_2\vec{Y}}$	$\mathcal{H}_0$	$\mathcal{H}_1$	Ablehnungsbereich
$\{\{\mathcal{MN}[{}_1\boldsymbol{x}\cdot{}_1\vec{\beta},\ {}_1\sigma^2\,\mathbf{E}],\ \mathcal{MN}[{}_2\boldsymbol{x}\cdot{}_2\vec{\beta},\ {}_2\sigma^2\,\mathbf{E}]\} \mid {}_1\vec{\beta},\ {}_2\vec{\beta} \in \mathbb{R}_{m+1},\ {}_1\sigma,\ {}_2\sigma > 0\}$	${}_1\sigma = {}_2\sigma$	${}_1\sigma \neq {}_2\sigma$	$-2\,\mathrm{Log}[\mathrm{MLQ}] > c_{1;1-\alpha}$

Dabei bezeichnet $c_{1;q}$ das q-Quantil der Chi[1]-Verteilung und $-2\,\mathrm{Log}[\mathrm{MLQ}]$ die Statistik

$$-2\,\mathrm{LogMLQ} = ({}_1n + {}_2n)\,\mathrm{Log}\left[\frac{{}_1n\,{}_1\hat{\Sigma}^2 + {}_2n\,{}_2\hat{\Sigma}^2}{{}_1n + {}_2n}\right] - {}_1n\,\mathrm{Log}[{}_1\hat{\Sigma}^2] - {}_2n\,\mathrm{Log}[{}_2\hat{\Sigma}^2]$$

▼

Beweis: Wir konstruieren für dieses Testproblem den Maximum-Likelihood-Quotiententest und verwenden dazu die Bezeichnungen $\Theta = \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \times \left]0,\ \infty\right[\times \left]0,\ \infty\right[$ und $\Theta_0 = \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \times \{\{\sigma,\ \sigma\} \mid \sigma > 0\}$:

Mit Hilfe der Maximum-Likelihood-Methode ergeben sich die Schätzer

$$\hat{\vartheta}[{}_1\vec{Y},\ {}_2\vec{Y}] = \{{}_1\hat{\mathrm{B}},\ {}_2\hat{\mathrm{B}},\ {}_1\hat{\Sigma}^2,\ {}_2\hat{\Sigma}^2\} \qquad \mathrm{b} \qquad \mathrm{z} \qquad \mathrm{w}$$

$$\hat{\vartheta}_0[{}_1\vec{Y},\ {}_2\vec{Y}] = \{{}_1\hat{\mathrm{B}},\ {}_2\hat{\mathrm{B}},\ \frac{{}_1n\,{}_1\hat{\Sigma}^2 + {}_2n\,{}_2\hat{\Sigma}^2}{{}_1n + {}_2n},\ \frac{{}_1n\,{}_1\hat{\Sigma}^2 + {}_2n\,{}_2\hat{\Sigma}^2}{{}_1n + {}_2n}\}$$

mit

$${}_i\hat{\mathrm{B}} = ({}_i\boldsymbol{x}^t \cdot {}_i\boldsymbol{x})^{-1} \cdot {}_i\boldsymbol{x}^t \cdot {}_i\vec{Y} \qquad \text{und} \qquad {}_i\hat{\Sigma}^2 = \frac{1}{{}_in}\,({}_i\vec{Y} - {}_i\boldsymbol{x}\cdot{}_i\hat{\mathrm{B}})^t \cdot ({}_i\vec{Y} - {}_i\boldsymbol{x}\cdot{}_i\hat{\mathrm{B}})$$

für die unbekannten Parameter $\vec{\vartheta} = \{{}_1\vec{\beta},\ {}_2\vec{\beta},\ {}_1\sigma,\ {}_2\sigma\}$ bzw $\vec{\vartheta}_0 = \{{}_1\vec{\beta},\ {}_2\vec{\beta},\ \sigma,\ \sigma\}$. Für den Maximum-Likelihood-Quotient MLQ ergibt sich damit

$$\mathrm{MLQ} = \frac{L[{}_1\vec{Y},\ {}_2\vec{Y} \mid \hat{\vartheta}[{}_1\vec{Y},\ {}_2\vec{Y}]]}{L[{}_1\vec{Y},\ {}_2\vec{Y} \mid \hat{\vartheta}[{}_1\vec{Y},\ {}_2\vec{Y}]]} = \ldots = \frac{[{}_1\hat{\Sigma}^2]^{{}_1n/2}\,[{}_2\hat{\Sigma}^2]^{{}_2n/2}}{\left[\frac{{}_1n\,{}_1\hat{\Sigma}^2 + {}_2n\,{}_2\hat{\Sigma}^2}{{}_1n + {}_2n}\right]^{({}_1n+{}_2n)/2}}$$

Wir werden die Hypothese $\mathcal{H}_0$ somit im Fall

$$-2\,\mathrm{Log}[\mathrm{MLQ}] = ({}_1n + {}_2n)\,\mathrm{Log}\left[\frac{{}_1n\,{}_1\hat{\Sigma}^2 + {}_2n\,{}_2\hat{\Sigma}^2}{{}_1n + {}_2n}\right] - {}_1n\,\mathrm{Log}\left[{}_1\hat{\Sigma}^2\right] - {}_2n\,\mathrm{Log}\left[{}_1\hat{\Sigma}^2\right] > c_{1;1-\alpha}$$

ablehnen.

Auch dieser Test lässt sich in *Mathematica* leicht implementieren:

■ GoldfeldQuandtTest[*daten*1, *daten*2, *model*, *vars*]

berechnet für das Datenmaterial *daten*1 und *daten*2, das Modell *model* und die Variablen *vars* den *p*-Wert des

Tests von Goldfeld und Quandt für die Hypothese $\mathcal{H}_0 \ldots {}_1\sigma = {}_2\sigma$ gegen die Alternative $\mathcal{H}_1 \ldots {}_1\sigma \neq {}_2\sigma$.

Beide Daten müssen dabei die übliche Form $\{\{{}_1x_{11}, {}_1x_{12}, \ldots, {}_1x_{1s}, {}_1y_1\}, \{{}_1x_{21}, {}_1x_{22}, \ldots, {}_1x_{2s}, {}_1y_2\}, \ldots\}$ bzw $\{\{{}_2x_{11}, {}_2x_{12}, \ldots, {}_2x_{1s}, {}_2y_1\}, \{{}_2x_{21}, {}_2x_{22}, \ldots, {}_2x_{2s}, {}_2y_2\}, \ldots\}$ besitzen. Mit der Liste *model* lässt sich steuern, welche Funktionen der Variablen in das Modell aufgenommen werden. In der Liste *vars* werden alle in den Daten aufscheinenden Variablen aufgelistet.

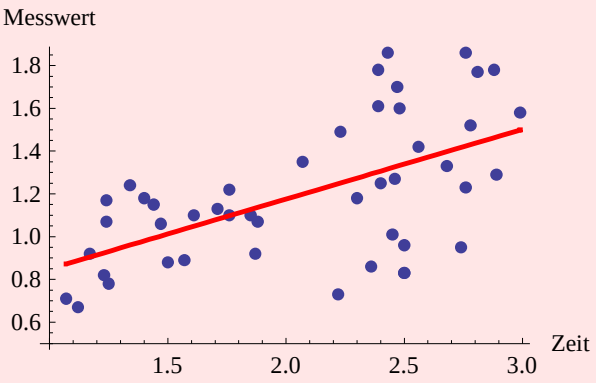
Wir illustrieren den Test von Goldfeld und Quandt an einem konkreten Beispiel:

12.5.5 Beispiel: Eine Messgröße, die mit der Zeit linear zunimmt, wurde während des Zeitintervalls [1, 3] insgesamt $n = 45$ mal gemessen. Die mit Messfehlern behafteten Messwerte sind im Datenmaterial zeitreihe aufgelistet. Man prüfe, ob die Streuung der Messfehler während des ganzen Zeitintervalls [1, 3] gleich bleibt.

▼

Lösung: Wir veranschaulichen dieses Datenmaterial zuerst graphisch durch ein Scatter-Plot

```
zeitreihe = Rest[<< zeitreihefile];
ScatterPlot[zeitreihe, PlotStyle → PointSize[0.025], AxesOrigin → {1, 0.5}, AxesLabel → {"Zeit", "Messwert"}]
```



und erkennen, dass die vor dem Zeitpunkt $t = 2$ ermittelten Werte weniger streuen dürften, als die nach dem Zeitpunkt $t = 2$ ermittelten Werte. Diese Vermutung soll nun mit Hilfe des Tests von Goldfeld und Quandt bestätigt werden:

```
daten1 = Select[zeitreihe, #[[1]] < 2 &];
daten2 = Select[zeitreihe, #[[1]] ≥ 2 &];
GoldfeldQuandtTest[daten1, daten2, {1, t}, {t}]
Clear[zeitreihe, daten1, daten2]
```

```
PValue → 0.000441566
```

Da dieser p-Wert sehr klein ist, wird die Hypothese $\mathcal{H}_0$ deutlich abgelehnt. Die vor dem Zeitpunkt $t = 2$ ermittelten Werte streuen also tatsächlich signifikant weniger, als die nach dem Zeitpunkt $t = 2$ ermittelten Werte.

12.6 Die ANOVA als Design-Modell der Regressionsanalyse

Die Varianzanalyse kann als Spezialfall der Regressionsanalyse aufgefasst werden. Dadurch lassen sich Fragestellungen der Varianzanalyse mit Methoden der Regressionsanalyse behandeln. Beispielsweise lassen sich auf diese Weise auch bei Vorliegen eines nicht-balanzierten Stichprobenplans Aussagen über den Einfluss der einzelnen Faktoren sowie ihre Wechselwirkung überprüfen.

Wir zeigen zuerst, wie sich die einfache Varianzanalyse in die Regressionsanalyse einbetten lässt:

12.6.1 Bemerkung: Die einfache Varianzanalyse als Design-Modell der Regressionsanalyse:

a) Ersetzt man den in s Stufen wirkenden Faktor S durch die $s-1$ Einflussgrößen $S_1, S_2, \ldots, S_{s-1}$ mit

$$S_i = \begin{cases} +1 & \text{falls der Faktor } S \text{ in der } i-\text{ten Stufe wirkt} \\ -1 & \text{falls der Faktor } S \text{ in der } s-\text{ten Stufe wirkt} \\ 0 & \text{sonst} \end{cases}$$

so gilt für die $n_\bullet$ Zufallsvariablen $X_{i\,k}$ der einfachen Varianzanalyse

$$X_{i\,k} = \mu + \sum_{i=1}^{s-1} \xi_i\, S_i + E_{i\,k}$$

wobei die $E_{i\,k}$ vollständig unabhängige, $\mathcal{N}[0, \sigma]$-verteilte Zufallsvariablen sind.

b) Unter Verwendung der Bezeichnungen

$$\vec{Y} = \{X_{11}, X_{12}, \ldots, X_{1\,n_1}, X_{21}, X_{22}, \ldots, X_{2\,n_2}, \ldots, X_{s\,1}, X_{22}, \ldots, X_{2\,n_2}\}^t$$
$$\vec{E} = \{E_{11}, E_{12}, \ldots, E_{1\,n_1}, E_{21}, E_{22}, \ldots, E_{2\,n_2}, \ldots, E_{s\,1}, E_{22}, \ldots, E_{2\,n_2}\}^t$$
$$\vec{\beta} = \{\mu, \xi_1, \xi_2, \ldots, \xi_{s-1}\}^t$$

sowie der Matrix

$$\boldsymbol{x} = \{\vec{1}, \vec{S}_1, \vec{S}_2, \ldots, \vec{S}_{s-1}\} \in \mathbb{R}^s_{n_\bullet}$$

wobei $\vec{1}$ den Spaltenvektor $\vec{1} = \{1, 1, \ldots, 1\}^t \in \mathbb{R}_{n_\bullet}$ und $\vec{S}_1, \vec{S}_2, \ldots, \vec{S}_{s-1} \in \mathbb{R}_{n_\bullet}$ Spaltenvektoren mit den Einträgen der Einflussgrößen $S_1, S_2, \ldots, S_{s-1}$ bezeichnet, lässt sich das Modell der einfachen Varianzanalyse in der von der Regressionsanalyse her bekannten Form

$$\vec{Y} = \boldsymbol{x} \cdot \vec{\beta} + \vec{E}$$

anschreiben.

c) Damit entspricht die Hypothese $\mathcal{H}_0$ der einfachen Varianzanalyse, dass nämlich der Faktor S keinen Einfluss auf die Mittelwerte der s Grundgesamtheiten hat, der Hypothese $\mathcal{H}_0 \ldots \{\xi_1, \ldots, \xi_{s-1}\} = \vec{0}$ der Regressionsanalyse, welche sich mit Hilfe des vollständigen Regressionstests prüfen lässt. Es gelten dabei offenbar die folgenden Entsprechungen:

einfache ANOVA	Regressionsanalyse
SSS	SSR
SSE	SSE
SStotal	SStotal

Als nächstes zeigen wir, wie sich die zweifache Varianzanalyse in die Regressionsanalyse einbetten lässt:

12.6.2 Bemerkung: Die zweifache Varianzanalyse als Design-Modell der Regressionsanalyse:

a) Ersetzt man den in s Stufen wirkenden Faktor S durch die $s-1$ Einflussgrößen $S_1, S_2, \ldots, S_{s-1}$ mit

$$S_i = \begin{cases} +1 & \text{falls der Faktor } S \text{ in der } i-\text{ten Stufe wirkt} \\ -1 & \text{falls der Faktor } S \text{ in der } s-\text{ten Stufe wirkt} \\ 0 & \text{sonst} \end{cases}$$

und den in t Stufen wirkenden Faktor T durch die $t-1$ Einflussgrößen $T_1, T_2, \ldots, T_{t-1}$ mit

$$T_j = \begin{cases} +1 & \text{falls der Faktor } T \text{ in der } j-\text{ten Stufe wirkt} \\ -1 & \text{falls der Faktor } T \text{ in der } t-\text{ten Stufe wirkt} \\ 0 & \text{sonst} \end{cases}$$

so gilt für die $n_{\bullet\bullet}$ Zufallsvariablen $X_{i\,j\,k}$ der zweifachen Varianzanalyse mit Wechselwirkung

$$X_{i\,j\,k} = \mu + \sum_{i=1}^{s-1} \xi_i\, S_i + \sum_{j=1}^{t-1} \eta_j\, T_j + \sum_{i=1}^{s-1} \sum_{j=1}^{t-1} \zeta_{i\,j}\, S_i\, T_j + E_{i\,j\,k}$$

und für die $n_{\bullet\bullet}$ Zufallsvariablen $X_{i\,j\,k}$ der zweifache Varianzanalyse ohne Wechselwirkung

$$X_{i\,j\,k} = \mu + \sum_{i=1}^{s-1} \xi_i\, S_i + \sum_{j=1}^{t-1} \eta_j\, T_j + E_{i\,j\,k}$$

wobei die $E_{i\,j\,k}$ vollständig unabhängige und $\mathcal{N}[0, \sigma]$-verteilte Zufallsvariablen sind.

b) Unter Verwendung der Bezeichnungen

$$\vec{Y} = \{X_{111}, X_{112}, \ldots, X_{11\,n_{11}}, X_{121}, X_{122}, \ldots, X_{12\,n_{12}}, \ldots, X_{s\,t\,1}, X_{s\,t\,2}, \ldots, X_{s\,t\,n_{s\,t}}\}^t$$
$$\vec{E} = \{E_{111}, E_{112}, \ldots, E_{11\,n_{11}}, E_{121}, E_{122}, \ldots, E_{12\,n_{12}}, \ldots, E_{s\,t\,1}, E_{s\,t\,2}, \ldots, E_{s\,t\,n_{s\,t}}\}^t$$
$$\vec{\beta} = \{\mu, \xi_1, \xi_2, \ldots, \xi_{s-1}, \eta_1, \eta_2, \ldots, \eta_{t-1}, \zeta_{11}, \zeta_{12}, \ldots, \zeta_{s-1\,t-1}\}^t$$
$$\vec{\beta}' = \{\mu, \xi_1, \xi_2, \ldots, \xi_{s-1}, \eta_1, \eta_2, \ldots, \eta_{t-1}\}^t$$

sowie der Matrizen

$$\boldsymbol{x} = \left\{\vec{1}, \vec{S}_1, \vec{S}_2, \ldots, \vec{S}_{s-1}, \vec{T}_1, \vec{T}_2, \ldots, \vec{T}_{t-1}, \vec{S}_1\,\vec{T}_1, \vec{S}_1\,\vec{T}_2, \ldots, \vec{S}_{s-1}\,\vec{T}_{t-1}\right\} \in \mathbb{R}^{s\,t}_{n_{\bullet\bullet}}$$

bzw

$$\boldsymbol{x}' = \left\{\vec{1}, \vec{S}_1, \vec{S}_2, \ldots, \vec{S}_{s-1}, \vec{T}_1, \vec{T}_2, \ldots, \vec{T}_{t-1}\right\} \in \mathbb{R}^{s+t-1}_{n_{\bullet\bullet}}$$

wobei $\vec{1}$ den Spaltenvektor $\vec{1} = \{1, 1, \ldots, 1\}^t \in \mathbb{R}_{n_{\bullet\bullet}}$ und $\vec{S}_1, \vec{S}_2, \ldots, \vec{S}_{s-1}, \vec{T}_1, \vec{T}_2, \ldots, \vec{T}_{t-1} \in \mathbb{R}_{n_{\bullet\bullet}}$ Spaltenvektoren mit den Einträgen der Einflussgrößen $S_1, S_2, S_{s-1}, T_1, T_2, \ldots, T_{t-1}$ bezeichnet, lässt sich das Modell der zweifachen Varianzanalyse mit Wechselwirkung bzw das Modell der zweifachen Varianzanalyse ohne Wechselwirkung in der von der Regressionsanalyse her bekannten Form

$$\vec{Y} = \boldsymbol{x} \cdot \vec{\beta} + \vec{E} \qquad \text{bzw} \qquad \vec{Y} = \boldsymbol{x}' \cdot \vec{\beta}' + \vec{E}$$

anschreiben.

c) Damit entsprechen die Hypothesen $\mathcal{H}_0^S$ bzw $\mathcal{H}_0^T$ bzw $\mathcal{H}_0^{ST}$ der zweifachen Varianzanalyse, dass nämlich der Faktor S bzw der Faktor T bzw die Wechselwirkung zwischen diesen Faktoren keinen Einfluss auf die Mittelwerte der $s\,t$ Grundgesamtheiten hat, den Hypothesen $\mathcal{H}_0 \ldots \{\xi_1, \ldots, \xi_{s-1}\} = \vec{0}$ bzw $\mathcal{H}_0 \ldots \{\eta_1, \ldots, \eta_{s-1}\} = \vec{0}$ bzw $\mathcal{H}_0 \ldots \{\zeta_{11}, \ldots, \zeta_{s-1\,t-1}\} = \vec{0}$ der Regressionsanalyse, welche sich mit Hilfe von geeigneten Regressionstests prüfen lassen.

d) Mit der (im Rahmen der einfachen ANOVA ermittelten) Größe SSE sowie den für die beiden Regressionsmodelle $\vec{Y} = \boldsymbol{x}' \cdot \vec{\beta}' + \vec{E}$ bzw $\vec{Y} = \boldsymbol{x} \cdot \vec{\beta} + \vec{E}$ ermittelten Größen SSE' bzw SSE gilt für die im Rahmen der zweifachen ANOVA für nichtbalanzierte Stichprobenpläne verwendeten Statistiken ${}^{S}\mathrm{SST}_{\mathrm{reg}}$ bzw ${}^{S,T}\mathrm{SS\,(ST)}_{\mathrm{reg}}$

$${}^{S}\mathrm{SST}_{\mathrm{reg}} = \mathrm{SSE} - \mathrm{SSE}' \qquad \text{bzw} \qquad {}^{S,T}\mathrm{SS\,(ST)}_{\mathrm{reg}} = \mathrm{SSE}' - \mathrm{SSE}$$

Man beachte in diesem Zusammenhang die folgenden Entsprechungen:

einfache ANOVA	zweifache ANOVA ohne Wechselwirkung	zweifache ANOVA mit Wechselwirkung	Regression $\vec{Y} = \boldsymbol{x}' \cdot \vec{\beta}' + \vec{E}$	Regression $\vec{Y} = \boldsymbol{x} \cdot \vec{\beta} + \vec{E}$
SSS	SSS	SSS		
	${}^{S}\mathrm{SST}_{\mathrm{reg}}$	${}^{S}\mathrm{SST}_{\mathrm{reg}}$		
		${}^{ST}\mathrm{SS}\,(\mathrm{ST})_{\mathrm{reg}}$		
		$\mathrm{SSE}_{\mathrm{reg}}$		SSE
	$\mathrm{SSE}'_{\mathrm{reg}}$	${}^{ST}\mathrm{SS}\,(\mathrm{ST})_{\mathrm{reg}} + \mathrm{SSE}_{\mathrm{reg}}$	SSE'	
SSE	${}^{S}\mathrm{SST}_{\mathrm{reg}} + \mathrm{SSE}'_{\mathrm{reg}}$	${}^{S}\mathrm{SST}_{\mathrm{reg}} + {}^{ST}\mathrm{SS}\,(\mathrm{ST})_{\mathrm{reg}} + \mathrm{SSE}_{\mathrm{reg}}$		
SStotal	SStotal	SStotal	SStotal	SStotal

Mit einem Beispiel soll diese Beziehung zwischen ANOVA und Regressionsanalyse verdeutlicht werden:

12.6.3 Beispiel: Man betrachte nochmals Beispiel 11.3.1, beantworte nun aber die dort gestellten Fragen unter Verwendung der Regressionsanalyse.

▼

Lösung: a) Beim Datenmaterial "gesundheit" wird der Einfluss der beiden Faktoren durch die beiden ersten Einträge beschrieben. So wirken etwa beim Datum {1, 1, 4.64} die beiden Faktoren S und T in der jeweils ersten Stufe. Wir müssen dieses Datenmaterial nun so modifizieren, dass die ersten $s - 1 = 2$ Einträge den beiden Einflussgrößen S_1 und S_2 und die nächsten $t - 1 = 2$ Einträge den beiden Einflussgrößen T_1 und T_2 entsprechen. In der folgenden Tabelle wird angegeben, welche Modifikationen dabei vorgenommen werden müssen:

Eintrag	Modifikation	Eintrag	Modifikation	Eintrag	Modifikation
1, 1	1, 0, 1, 0	2, 1	0, 1, 1, 0	3, 1	−1, −1, 1, 0
1, 2	1, 0, 0, 1	2, 2	0, 1, 0, 1	3, 2	−1, −1, 0, 1
1, 3	1, 0, −1, −1	2, 3	0, 1, −1, −1	3, 3	−1, −1, −1, −1

Dieses so modifizierte Datenmaterial findet sich bereits im Datenordner unter dem Namen gesundheitmodfile.

b) Der die ANOVA-Tabelle betreffende Output der Regressionsanalyse für die Einflussgrößen S_1 und S_2 stimmt mit dem Output der einfachen Varianzanalyse für den Faktor Substanz überein. Insbesondere lassen sich damit die drei Größen SSS, SSE und SStotal der einfachen ANOVA mit Hilfe der Regressionsanalyse ermitteln:

```
gesundheit = Rest[<< gesundheitfile];
gesundheitmod = Rest[<< gesundheitmodfile];
Regress[gesundheitmod, {S1, S2}, {S1, S2, T1, T2}][[5]]
ANOVA[gesundheit, {Substanz}, {Substanz, Altersklasse}, CellMeans → False]
```

```
ANOVATable →
            DF   SumOfSq    MeanSq     FRatio    PValue
Model       2    2.39113    1.19557    1.99077   0.142249
Error       95   57.0529    0.600556
Total       97   59.444

ANOVA →
            DF   SumOfSq    MeanSq     FRatio    PValue
Substanz    2    2.39113    1.19557    1.99077   0.142249
Error       95   57.0529    0.600556
Total       97   59.444
```

c) Die Zeile "Error" der ANOVA-Tabelle der Regressionsanalyse für die Einflussgrößen S_1, S_2, T_1, T_2 stimmt mit der Zeile "Error" der zweifachen Varianzanalyse ohne Wechselwirkung überein. Insbesondere lassen sich damit die Größen $\mathrm{SSE}'_{\mathrm{reg}}$ und ${}^{S}\mathrm{SST}_{\mathrm{reg}}$ der zweifachen ANOVA ohne Wechselwirkung mit Hilfe der Regressionsanalyse

ermitteln. Es gilt nämlich $SSE'_{reg} = SSE'$ und ${}^{S}SST_{reg} = SSE - SSE'$:

```
Regress[gesundheitmod, {S1, S2, T1, T2}, {S1, S2, T1, T2}][[5]]
ANOVA[gesundheit, {Substanz, Altersklasse}, {Substanz, Altersklasse}, CellMeans → False]
```

```
ANOVATable →
            DF   SumOfSq   MeanSq    FRatio    PValue
Model       4    9.33466   2.33367   4.33115   0.002947
Error       93   50.1093   0.53881
Total       97   59.444
```

```
ANOVA →
                DF   SumOfSq   MeanSq    FRatio    PValue
Substanz        2    2.39113   1.19557   2.2189    0.114453
Altersklasse    2    6.94353   3.47176   6.44339   0.00239506
Error           93   50.1093   0.53881
Total           97   59.444
```

d) Die Zeile "Error" der ANOVA-Tabelle der Regressionsanalyse für die Einflussgrößen S_1, S_2, T_1, T_2, $S_1\,T_1$, $S_1\,T_2$, $S_2\,T_1$, $S_2\,T_2$ stimmt mit der Zeile "Error" der zweifachen Varianzanalyse mit Wechselwirkung überein. Insbesondere lassen sich damit die Größen SSE_{reg} und ${}^{ST}SS\left(ST\right)_{reg}$ der zweifachen ANOVA mit Wechselwirkung mit Hilfe der Regressionsanalyse ermitteln. Es gilt nämlich $SSE_{reg} = SSE$ und ${}^{ST}SS\left(ST\right)_{reg} = SSE' - SSE$:

```
Regress[gesundheitmod, {S1, S2, T1, T2, S1 T1, S1 T2, S2 T1, S2 T2}, {S1, S2, T1, T2}][[5]]
ANOVA[gesundheit, {Substanz, Altersklasse, Substanz Altersklasse}, {Substanz, Altersklasse},
  CellMeans → False]
Clear[gesundheit, gesundheitmod]
```

```
ANOVATable →
            DF   SumOfSq   MeanSq     FRatio    PValue
Model       8    30.322    3.79025    11.5834   3.84371×10^-11
Error       89   29.122    0.327214
Total       97   59.444
```

```
ANOVA →
                          DF   SumOfSq   MeanSq     FRatio    PValue
Substanz                  2    2.39113   1.19557    3.65378   0.0298516
Altersklasse              2    6.94353   3.47176    10.6101   0.0000736611
Altersklasse Substanz     4    20.9873   5.24683    16.0349   6.37621×10^-10
Error                     89   29.122    0.327214
Total                     97   59.444
```

12.7 Diagnose-Tools

Wir befassen uns in diesem Abschnitt mit einigen Fragestellungen, welche im Zusammenhang mit der Regressionsanalyse oft auftreten und zeigen, wie sich diese Fragestellungen mit Hilfe von *Mathematica* behandeln lassen. Eine wesentliche Rolle kommt dabei der Option RegressionReport des Befehls Regress zu, mit der *Mathematica* angehalten werden kann, eine Fülle von interessanten Outputs zu liefern.

Ausgangspunkt unserer Überlegungen ist dabei stets das Modell $\vec{Y} = \boldsymbol{x}\cdot\vec{\beta} + \vec{E}$, das unter Verwendung der Daten $\{\{x_{11}, x_{12}, \ldots, x_{1\,s}, y_1\}, \{x_{21}, x_{22}, \ldots, x_{2\,s}, y_2\}, \ldots\}$ zur Ermittlung der Schätzwerte $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ und $\hat{\sigma}^2$ bzw der Schätzer $\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_m$ und $\hat{\Sigma}^2$ herangezogen wird.

■ Vorhersagewert für Y_*

Im Rahmen der Regressionsanalyse ist man oft an der Zufallsvariablen

$$Y_* = \beta_0 + \beta_1 x_{*1} + \beta_2 x_{*2} + \ldots + \beta_m x_{*m} + E_*$$

interessiert, wobei die m Einflussgrößen den Wert $\vec{x}_* = \{x_{*1}, x_{*2}, \ldots, x_{*m}\}$ annehmen.

12.7.1 Definition: Die Zahl

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_{*1} + \hat{\beta}_2 x_{*2} + \ldots + \hat{\beta}_m x_{*m}$$

nennt man den Vorhersagewert für Y_*.

▼

Man beachte:

- der Punkt $\{\vec{x}_*, \hat{y}_*\}$ liegt auf der Regressionshyperebene $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_m x_m$;
- die Zufallsvariable $\hat{Y}_* = \hat{B}_0 + \hat{B}_1 x_{*1} + \hat{B}_2 x_{*2} + \ldots + \hat{B}_m x_{*m}$ ist ein erwartungstreuer Schätzer für $\mathbb{E}[Y_*]$.

Der Vorhersagewert für Y_* lässt sich mit dem Befehl Fit leicht ermitteln:

12.7.2 Beispiel: Von $n = 30$ zufällig ausgewählten Personen wurde das Alter und der systolische Blutdruck ermittelt (man vergleiche dazu das Datenmaterial blutdruck sowie Beispiel 12.4.5). Mit welchem Blutdruck muss daher eine 60-jährige Person bei Verwendung des Modells Blutdruck = $\gamma_0 + \gamma_1$ Alter^2 rechnen?

▼

Lösung: Der voraussichtliche Blutdruck einer 60-jährigen Person entspricht dem Wert des Regressionspolynoms Blutdruck = $\gamma_0 + \gamma_1$ Alter^2 an der Stelle Alter = 60. Dieser Wert lässt sich folgendermaßen ermitteln:

```
blutdruck = Rest[<< blutdruckfile];
Fit[blutdruck, {1, Alter^2}, {Alter}] /. Alter -> 60
Clear[blutdruck]
```

```
156.873
```

Auf Grund des uns zur Verfügung stehenden Datenmaterials muss eine 60-jährige Person daher mit einem Blutdruck von 156.873 mmHg rechnen.

Möchte man für alle $i \in \{1, 2, \ldots, n\}$ die Vorhersagewerte für Y_i (also die zu $\vec{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ gehörenden Werte $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_m x_{im}$) bzw die Fehler $e_i = y_i - \hat{y}_i$ ermitteln, so verwende man bei der Option RegressionReport des Befehls Regress die Einstellungen PredictedResponse bzw FitResiduals:

■ Regress[*daten*, *model*, *vars*, RegressionReport → PredictedResponse]

ermittelt alle Vorhersagewerte $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_m x_{im}$.

■ FitResiduals /. Regress[*daten*, *model*, *vars*, RegressionReport → FitResiduals]

ermittelt alle Fehler $e_i = y_i - \hat{y}_i$.

12.7.3 Beispiel: Wir betrachten nochmals das Datenmaterial zugfestigkeit zusammen mit dem bereits in Beispiel 12.2.4 untersuchten Modell

$$\text{Zugfestigkeit} = \beta_0 + \beta_1 \text{Dicke} + \beta_2 \text{Haspeltemperatur} + \beta_3 \text{Mangangehalt} + E$$

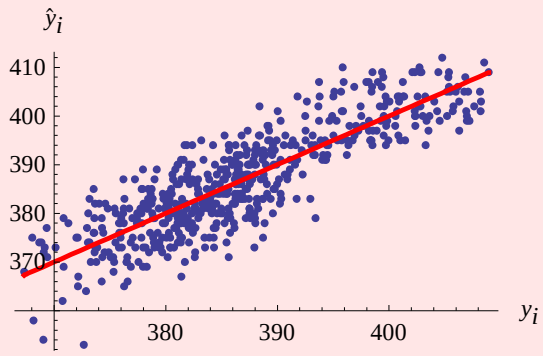
Man zeichne ein Scatter-Plot, bei dem auf der x-Achse die Vorhersagewerte $\hat{y}_i$ und auf der y-Achse, die

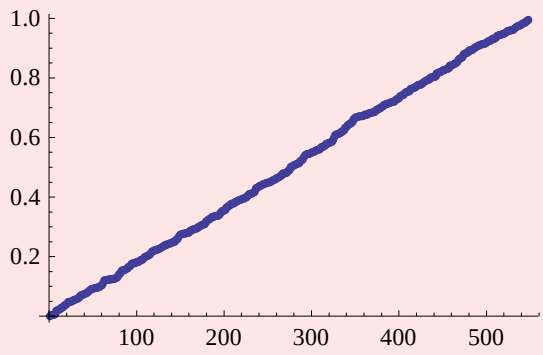
tatsächlich gemessenen Werte y_i der Zugfestigkeiten aufgetragen werden und prüfe unter Verwendung des Score Tests, ob die Fehler $e_i = y_i - \hat{y}_i$ tatsächlich normalverteilt sind.

▼

Lösung: Vom Datenmaterial zugfestigkeit ausgehend, erzeugen wir eine Liste (w) der tatsächlich gemessenen Werte y_i der Zugfestigkeiten sowie unter Verwendung der Option RegressionReport des Befehls Regress mit den Einstellungen PredictedResponse bzw FitResiduals Listen (v bzw f) der Vorhersagewerte $\hat{y}_i$ der Zugfestigkeiten bzw der Fehler $e_i = y_i - \hat{y}_i$. Anschließend erzeugen wir unter Verwendung von Thread eine Liste aller Paare der Form $\{\hat{y}_i, y_i\}$ und plotten diese Liste mit Hilfe von ScatterPlot. Schließlich wenden wir den Befehl ScoreTest auf die Liste der Fehler e_i an:

```
zugfestigkeit = Rest[<< zugfestigkeitfile];
w = zugfestigkeit[[All, 4]];
v = Regress[zugfestigkeit, {1, Dicke, Haspeltemperatur, Mangangehalt}, {Dicke, Haspeltemperatur, Mangangehalt},
      RegressionReport → PredictedResponse][[1, 2]];
f = Regress[zugfestigkeit, {1, Dicke, Haspeltemperatur, Mangangehalt}, {Dicke, Haspeltemperatur, Mangangehalt},
      RegressionReport → FitResiduals][[1, 2]];
ScatterPlot[Thread[{v, w}], AxesLabel → {"yi", "ŷi"}, ImageSize → {200, 130}]
ScoreTest[f, NormalDistribution[Mean[f], StandardDeviation[f]], ImageSize → {200, 130}]
Clear[zugfestigkeit, w, v, f]
```





Die erste Zeichnung zeigt, wie gut sich die einzelnen Zugfestigkeiten y_i durch die Vorhersagewerte $\hat{y}_i$ vorhersagen lassen. Die zweite Zeichnung zeigt, dass die Fehler e_i tatsächlich normalverteilt sind (die Punkte liegen annähernd auf einer Geraden).

12.7.4 Beispiel: Eine auf den Verleih von Fahrzeugen spezialisierte Firma ermittelte über einen Zeitraum von 20 Wochen für jeden Arbeitstag die Anzahl der jeweils bereits eine Woche vorher reservierten Autos, sowie die Anzahl der tatsächlich vermieteten Autos (man vergleiche dazu das Datenmaterial autoverleih). Gesucht ist eine Formel, mit der sich die Anzahl der tatsächlich vermieteten Autos vorhersagen lässt sowie ein Scatter-Plot, bei dem auf der x-Achse die Vorhersagewerte $\hat{y}_i$ und auf der y-Achse die Anzahl der tatsächlich

vermieteten Autos y_i aufgetragen werden.

▼

Lösung: Bei unserer Zielgröße (Vermietet) handelt es sich um die Anzahl der tatsächlich vermieteten Autos. Diese Zielgröße hängt von der Anzahl der für diesen Tag langfristig reservierten Autos (Reserviert) und vom Wochentag (Mo, Di, Mi, Do, Fr) ab. Um die Abhängigkeit vom Wochentag modellieren zu können, führen wir für jeden Wochentag eine eigene Indikatorvariable ein, welche für diesen Wochentag den Wert 1 und sonst den Wert 0 annimmt. Unser Modell hat damit die Gestalt (man beachte, dass diese Indikatorvariablen und die Konstante 1 linear abhängig sind; man darf daher nicht alle Indikatorvariablen in das Modell aufnehmen, da *Mathematica* sonst eine Fehlermeldung ausgibt):

$$\text{Vermietet} = \beta_0 + \beta_1 * \text{Mo} + \beta_2 * \text{Di} + \beta_3 * \text{Mi} + \beta_4 * \text{Do} + \beta_5 * \text{Reserviert} + E$$

a) Mit Hilfe von Regress analysieren wir zuerst dieses Modell (man beachte, dass die erste Spalte - also die Spalte mit den Wochentagen - nicht in die Regressionsanalyse aufgenommen werden darf, da es sich dabei um qualitative Daten handelt):

```
autoverleih = Rest[<< autoverleihfile][[All, {2, 3, 4, 5, 6, 7, 8}]];
Regress[autoverleih, {1, Mo, Di, Mi, Do, Reserviert}, {Mo, Di, Mi, Do, Fr, Reserviert}]
```

{ParameterTable →

	Estimate	SE	TStat	PValue
1	2.04979	4.01393	0.510668	0.610781
Mo	16.4365	3.294	4.98982	2.76558×10^{-6}
Di	30.9038	3.28153	9.4175	3.10862×10^{-15}
Mi	-2.44123	3.28122	-0.744001	0.458731
Do	-0.772512	3.28118	-0.235437	0.814382
Reserviert	2.72512	0.141795	19.2187	0.

, RSquared → 0.843675, AdjustedRSquared → 0.83536, EstimatedVariance → 107.66,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	5	54 617.	10 923.4	101.462	0.
Error	94	10 120.	107.66		
Total	99	64 737.			

}

Die p-Werte für die Variablen Mi und Do sind groß; diese Variablen haben keinen Einfluss auf unsere Zielgröße und können daher aus dem Modell entfernt werden, ohne dass sich die Güte des Modells wesentlich ändert:

```
Regress[autoverleih, {1, Mo, Di, Reserviert}, {Mo, Di, Mi, Do, Fr, Reserviert}]
```

{ParameterTable →

	Estimate	SE	TStat	PValue
1	0.958249	3.51013	0.272995	0.785444
Mo	17.5095	2.67449	6.54686	2.88811×10^{-9}
Di	31.9753	2.65955	12.0228	0.
Reserviert	2.726	0.140733	19.37	0.

, RSquared → 0.842713, AdjustedRSquared → 0.837798, EstimatedVariance → 106.065,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	3	54 554.7	18 184.9	171.45	0.
Error	96	10 182.3	106.065		
Total	99	64 737.			

}

Die Anzahl der tatsächlich vermieteten Fahrzeuge lässt sich somit mit der Formel

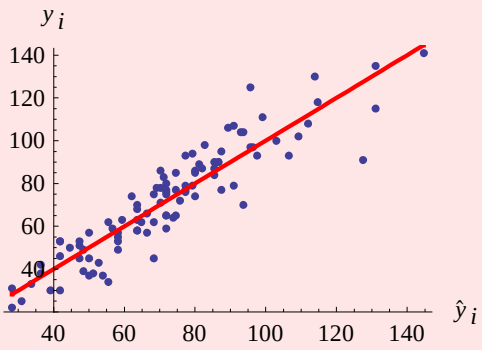
$$\text{Vermietet} = 0.958249 + 17.5095 * \text{Mo} + 31.9753 * \text{Di} + 2.726 * \text{Reserviert}$$

gut ($\hat{r}^2 = 0.842713$ und $\hat{\sigma}^2 = 106.065$ also $\hat{\sigma} = 10.2988$) vorhersagen.

b) Wir erzeugen eine Liste (w) der Anzahl der tatsächlich vermieteten Fahrzeuge y_i und ausgehend von obigem Modell unter Verwendung der Option RegressionReport des Befehls Regress mit der Einstellung PredictedRe-

sponse eine Liste (v) der entsprechenden Vorhersagewerte $\hat{y}_i$. Unter Verwendung von Thread konstruieren wir eine Liste aller Paare der Form $\{\hat{y}_i, y_i\}$ und plotten diese Liste mit Hilfe von ScatterPlot.

```
w = Rest[<< autoverleihfile][[All, 8]];
v = Regress[autoverleih, {1, Mo, Di, Reserviert}, {Mo, Di, Mi, Do, Fr, Reserviert},
      RegressionReport → PredictedResponse][[1, 2]];
ScatterPlot[Thread[List[v, w]], AxesLabel → {"ŷ i", "y i"}, ImageSize → {200, 130}]
Clear[autoverleih, w, v]
```



Die Zeichnung zeigt, dass sich die Anzahl der tatsächlich vermieteten Autos y_i durch die Vorhersagewerte $\hat{y}_i$ recht gut vorhersagen lässt. Punkte, welche weit von der Regressionsgeraden entfernt sind, entsprechen Tagen, welche durch unser Modell nicht gut beschrieben werden. Eine genauere Analyse des Datenmaterials zeigte, dass es sich dabei um Tage vor Feiertagen bzw Tage mit besonders schlechter Wetterprognose handelte. Würde man diese Tage aus unserem Datenmaterial entfernen, so würde die geschätzte Streuung $\hat{\sigma}$ etwas kleiner ausfallen.

■ Konfidenzintervall für $\mathbb{E}[Y_*]$ und Vorhersageintervall für Y_*

Im Rahmen der Regressionsanalyse ist man aber nicht nur am Vorhersagewert $\hat{y}_*$ der Größe

$$Y_* = \beta_0 + \beta_1 x_{*1} + \beta_2 x_{*2} + \ldots + \beta_m x_{*m} + E_*$$

für den Fall, dass die Einflussgrößen den Wert $\vec{x}_* = \{x_{*1}, x_{*2}, \ldots, x_{*m}\}$ annehmen, interessiert. Oft benötigt man Intervalle, welche den Erwartungswert $\mathbb{E}[Y_*]$ der Größe Y_* bzw diese Größe Y_* mit einer vorgegebenen Wahrscheinlichkeit enthalten. Wir setzen dazu im Folgenden voraus, dass die Zufallsvariable E_* von den in die Regressionsanalyse einfließenden Zufallsvariablen $E_1, E_2, \ldots, E_n$ unabhängig ist.

12.7.5 Definition:

a) Ein Intervall $[\underline{S}, \overline{S}]$, welches den Erwartungswert $\mathbb{E}[Y_*]$ der Größe Y_* mit der vorgegebenen Wahrscheinlichkeit β enthält, nennt man ein Konfidenzintervall für $\mathbb{E}[Y_*]$ mit Niveau β.

b) Ein Intervall $[\underline{T}, \overline{T}]$, welches die Größe Y_* mit der vorgegebenen Wahrscheinlichkeit β enthält, nennt man einen Vorhersageintervall für Y_* mit Niveau β.

12.7.6 Satz:

a) Das Intervall $[\underline{S}, \overline{S}]$ mit

$$\underline{S} = \vec{h}_* \cdot \hat{\vec{B}} - t_{n-(m+1),(1+\beta)/2} \sqrt{\vec{h}_* \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \vec{h}_*^t} \ \sqrt{\text{SSE}} \ / \sqrt{n-(m+1)}$$

$$\overline{S} = \vec{h}_* \cdot \hat{\vec{B}} + t_{n-(m+1),(1+\beta)/2} \sqrt{\vec{h}_* \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \vec{h}_*^t} \ \sqrt{\text{SSE}} \ / \sqrt{n-(m+1)}$$

ist ein Konfidenzintervall für $\mathbb{E}[Y_*]$ mit Niveau β.

b) Das Intervall $[\underline{T}, \overline{T}]$ mit

$$\underline{T} = \vec{h}_* \cdot \hat{\vec{B}} - t_{n-(m+1),(1+\beta)/2} \sqrt{1 + \vec{h}_* \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \vec{h}_*^t} \ \sqrt{\text{SSE}} \ / \sqrt{n-(m+1)}$$

$$\overline{T} = \vec{h}_* \cdot \hat{\vec{B}} + t_{n-(m+1),(1+\beta)/2} \sqrt{1 + \vec{h}_* \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \vec{h}_*^t} \ \sqrt{\text{SSE}} \ / \sqrt{n-(m+1)}$$

ist ein Vorhersageintervall für Y_* mit Niveau β.

Dabei bezeichnet $t_{n;q}$ das q-Quantil der $\mathcal{T}[n]$-Verteilung, $\vec{h}_*$ den Vektor $\vec{h}_* = \{1, x_{*1}, x_{*2}, \ldots, x_{*m}\} \in \mathbb{R}^{m+1}$ und SSE wie üblich die Statistik Sum of Squares of Error.

▼

Beweis: a) Für alle $\vec{\beta} \in \mathbb{R}_{m+1}$ und alle $\sigma > 0$ gilt wegen Bemerkung 12.3.6 und Satz 23.4.1

$$\mathbb{P}\left[\left\{\underline{S} \le \mathbb{E}[Y_*] \le \overline{S}\right\}\right] = \mathbb{P}[\{\frac{|\vec{h}_* \cdot \hat{\vec{B}} - \vec{h}_* \cdot \vec{\beta}| \sqrt{n-(m+1)}}{\sqrt{\text{SSE}/\sigma^2} \sqrt{\sigma^2 \vec{h}_* \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \vec{h}_*^t}} \le t_{n-(m+1),(1+\beta)/2}\}] = \beta$$

b) Berücksichtigt man, dass die $\mathcal{N}[0, \sigma]$-verteilte Zufallsvariable E_* von $E_1, E_2, \ldots, E_n$ und damit auch von $\hat{\vec{B}}$ und SSE unabhängig ist, so ergibt sich für alle $\vec{\beta} \in \mathbb{R}_{m+1}$ und alle $\sigma > 0$ analog zu a)

$$\mathbb{P}\left[\left\{\underline{T} \le Y_* \le \overline{T}\right\}\right] = \mathbb{P}[\{\frac{|\vec{h}_* \cdot \hat{\vec{B}} - \vec{h}_* \cdot \vec{\beta} - E_*| \sqrt{n-(m+1)}}{\sqrt{\text{SSE}/\sigma^2} \sqrt{\sigma^2 (1 + \vec{h}_* \cdot (\boldsymbol{x}^t \cdot \boldsymbol{x})^{-1} \cdot \vec{h}_*^t)}} \le t_{n-(m+1),(1+\beta)/2}\}] = \beta$$

Möchte man für alle $i \in \{1, 2, \ldots, n\}$ die Konfidenzintervalle für $\mathbb{E}[\tilde{Y}_i]$ bzw die Vorhersageintervalle für $\tilde{Y}_i$ mit

$$\tilde{Y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \tilde{E}_i$$

ermitteln, so verwende man bei der Option RegressionReport des Befehls Regress die Einstellungen MeanPredictionCITable bzw SinglePredictionCITable (man nimmt dabei an, dass die Zufallsvariablen $\tilde{E}_1, \tilde{E}_2, \ldots, \tilde{E}_n$ von den in die Regressionsanalyse einfließenden Zufallsvariablen $E_1, E_2, \ldots, E_n$ unabhängig sind):

■ Regress[*daten*, *model*, *vars*, RegressionReport → MeanPredictionCITable]

gibt alle tatsächlich beobachteten Werte y_i (Observed), alle Vorhersagewerte $\hat{y}_i$ (Predicted) sowie die gesuchten Konfidenzintervalle für $\mathbb{E}[\tilde{Y}_i]$ (CI) aus.

■ Regress[*daten*, *model*, *vars*, RegressionReport → SinglePredictionCITable]

gibt alle tatsächlich beobachteten Werte y_i (Observed), alle Vorhersagewerte $\hat{y}_i$ (Predicted) sowie die gesuchten Vorhersageintervall für $\tilde{Y}_i$ (CI) aus.

Als Konfidenzniveau wird dabei standardmäßig der Wert $\beta = 0.95$ verwendet. Mit der Option ConfidenceLevel lässt sich dieser Wert beliebig verändern.

Wir machen dazu wieder zwei Beispiele:

12.7.7 Beispiel: Für das Modell $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + E$ mit den Einflussgrößen x_1 und x_2 kennt man die konkreten Messungen

{{1, 1, 2}, {2, 1, 3}, {2, 2, 4}, {2, 3, 8}, {3, 2, 7}, {4, 2, 6}, {5, 3, 9}, {4, 3, 7}}

Gesucht sind für alle $i \in \{1, 2, \ldots, 8\}$ sowohl das Konfidenzintervall für $\mathbb{E}[\tilde{Y}_i]$ als auch das Vorhersageintervall für $\tilde{Y}_i$.

▼

Lösung: Wir ermitteln die gesuchten Konfidenzintervalle und Vorhersageintervalle unter Verwendung der Option RegressionReport des Befehls Regress mit den Einstellungen MeanPredictionCITable bzw SinglePredictionCITable:

```
daten = {{1, 1, 2}, {2, 1, 3}, {2, 2, 4}, {2, 3, 8}, {3, 2, 7}, {4, 2, 6}, {5, 3, 9}, {4, 3, 7}};
Regress[daten, {1, x1, x2}, {x1, x2}, RegressionReport -> MeanPredictionCITable][[1, 2]]
```

Observed	Predicted	SE	CI
2.	2.33904	0.704892	{0.527058, 4.15102}
3.	2.84932	0.663309	{1.14422, 4.55441}
4.	5.03082	0.473057	{3.81479, 6.24685}
8.	7.21233	0.888216	{4.9291, 9.49556}
7.	5.5411	0.389509	{4.53983, 6.54236}
6.	6.05137	0.61278	{4.47617, 7.62657}
9.	8.74315	0.726097	{6.87666, 10.6096}
7.	8.23288	0.564452	{6.78191, 9.68385}

```
Regress[daten, {1, x1, x2}, {x1, x2}, RegressionReport -> SinglePredictionCITable][[1, 2]]
Clear[daten]
```

Observed	Predicted	SE	CI
2.	2.33904	1.26665	{-0.916992, 5.59507}
3.	2.84932	1.24399	{-0.348466, 6.0471}
4.	5.03082	1.15383	{2.06482, 7.99683}
8.	7.21233	1.37712	{3.67233, 10.7523}
7.	5.5411	1.12216	{2.65648, 8.42571}
6.	6.05137	1.2178	{2.92092, 9.18182}
9.	8.74315	1.27857	{5.45647, 12.0298}
7.	8.23288	1.19421	{5.16306, 11.3027}

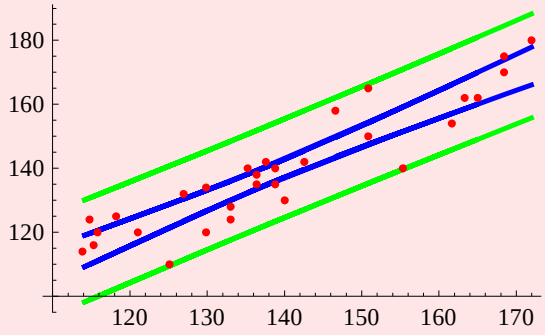
12.7.8 Beispiel: Von $n = 30$ zufällig ausgewählten Personen wurde das Alter und der systolische Blutdruck ermittelt (man vergleiche dazu das Datenmaterial blutdruck sowie Beispiel 12.4.5). Unter Verwendung des Modells $\text{Blutdruck} = \gamma_0 + \gamma_1 \text{Alter}^2$ zeichne man ein Plot, bei dem auf der x-Achse die Vorhersagewerte $\hat{y}_i$ und auf der y-Achse, die tatsächlich gemessenen Blutdruckwerte y_i sowie die zugehörigen Konfidenzintervalle und Vorhersageintervalle aufgetragen werden.

▼

Lösung: Wir ordnen zuerst unser Datenmaterial mit Hilfe von Sort hinsichtlich des Alters der untersuchten Personen (dies ist notwendig, um die Konfidenzbereiche bzw die Vorhersagebereiche "schön" zeichnen zu können). Anschließend erzeugen wir unter Verwendung der Option RegressionReport des Befehls Regress mit den Einstellun-

gen MeanPredictionCITable bzw SinglePredictionCITable sowie den Befehlen Flatten und Take Listen der tatsächlich beobachteten Blutdruckwerte (observed), der vorhergesagten Blutdruckwerte (predicted), sowie der Konfidenzintervalle (kon) für $\mathbb{E}[\tilde{Y}_i]$ und der Vorhersageintervalle für $\tilde{Y}_i$ (vor). Diese Listen zeichnen wir mit ListPlot in eine gemeinsame Graphik:

```
blutdruck = Sort[Rest[<< "blutdruckfile"], #1[[2]] < #2[[2]] &];
a = Regress[blutdruck, {1, Alter^2}, {Alter}, RegressionReport → MeanPredictionCITable];
b = Regress[blutdruck, {1, Alter^2}, {Alter}, RegressionReport → SinglePredictionCITable];
observed = Flatten[Take[a[[1, 2, 1]], All, {1}]];
predicted = Flatten[Take[a[[1, 2, 1]], All, {2, 2}]];
kon = Flatten[Take[a[[1, 2, 1]], All, {4}], 1];
vor = Flatten[Take[b[[1, 2, 1]], All, {4}], 1];
ListPlot[{Transpose[{predicted, observed}],
    Transpose[{predicted, First /@ kon}], Transpose[{predicted, Last /@ kon}],
    Transpose[{predicted, First /@ vor}], Transpose[{predicted, Last /@ vor}]},
  PlotStyle → {{Red}, {Blue, Thickness[0.01]}, {Blue, Thickness[0.01]}, {Green, Thickness[0.01]}, {Green, Thickness[
  Joined → {False, True, True, True, True}, AxesOrigin → {110, 100}, PlotRange → All, ImageSize → {200, 130}]
Clear[blutdruck, a, b, observed, predicted, kon, vor, p1, p2, p3, p4, p5]
```



Man erkennt, dass tatsächlich (fast) alle Punkte innerhalb des grünen Vorhersagebereiches liegen und die Regressionsgerade mit großer Wahrscheinlichkeit innerhalb des blauen Konfidenzbereiches liegen wird. Man beachte, dass sowohl der blaue als auch der grünen Bereich gegen den Rand hin geringfügig breiter wird.

■ Konfidenzintervalle für die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_m$

Bisher haben wir uns damit befasst, die unbekannten Parameter $\beta_0, \beta_1, \ldots, , \beta_m$ der Regressionsanalyse zu schätzen bzw Hypothesen über diese Parameter zu testen. Unter Verwendung des Dualitätsprinzips und dem partiellen Regressionstest sind wir in der Lage, für diese unbekannten Parameter auch Konfidenzintervalle anzugeben:

12.7.9 Konfidenzintervall mit Niveau β für den Parameter β_k der Regressionsanalyse:

$\mathfrak{P}_{\vec{Y}}$	Parameter	Konfidenzintervall
$\{\mathcal{MN}[\boldsymbol{x}\cdot\vec{\beta}, \sigma^2\,\mathbf{E}] \mid$ $\vec{\beta} = \{\beta_0, \beta_1, \ldots, \beta_m\} \in \mathbb{R}_{m+1}, \sigma > 0\}$	β_k	$\hat{\mathrm{B}}_k - t_{n-(m+1),(1+\beta)/2}\sqrt{\xi_{k\,k}}\ \sqrt{\mathrm{SSE}}\ /\sqrt{n-(m+1)} \le \beta_k$ $\beta_k \le \hat{\mathrm{B}}_k + t_{n-(m+1),(1+\beta)/2}\sqrt{\xi_{k\,k}}\ \sqrt{\mathrm{SSE}}\ /\sqrt{n-(m+1)}$

Dabei bezeichnet $t_{n;q}$ das q-Quantil der $\mathcal{T}[n]$-Verteilung, $\xi_{k\,k}$ den Eintrag der Matrix $(\boldsymbol{x}^t\cdot\boldsymbol{x})^{-1} \in \mathbb{R}^{m+1}_{m+1}$ am Schnittpunkt der k-ten Zeile mit der k-ten Spalte (k läuft dabei von 0 bis m) und SSE die Statistik Sum of Squares of Error.

Sollen diese Konfidenzintervalle mit Hilfe von *Mathematica* ermittelt werden, so verwende man bei der Option RegressionReport des Befehls Regress die Einstellung ParameterCITable:

■ Regress[*daten, model, vars,* RegressionReport → ParameterCITable]

ermittelt für alle Parameter $\beta_0, \beta_1, \ldots, \beta_m$ den Schätzwert (Estimate) sowie das zugehörige Konfidenzintervall (CI).

Als Konfidenzniveau wird dabei standardmäßig der Wert $\beta = 0.95$ verwendet. Mit der Option ConfidenceLevel lässt sich dieser Wert wieder beliebig verändern.

Wir machen dazu wieder ein Beispiel:

12.7.10 Beispiel: Wir betrachten nochmals das Datenmaterial zugfestigkeit zusammen mit dem in Beispiel 12.2.4 und Beispiel 12.7.3 untersuchten Modell

$$\text{Zugfestigkeit} = \beta_0 + \beta_1\ \text{Dicke} + \beta_2\ \text{Haspeltemperatur} + \beta_3\ \text{Mangangehalt} + E$$

Gesucht sind die Konfidenzintervalle für die Parameter $\beta_0, \beta_1, \beta_2, \beta_3$.

▼

Lösung: Wir ermitteln die gesuchten Konfidenzintervalle unter Verwendung der Option RegressionReport des Befehls Regress mit der Einstellung ParameterCITable:

```
zugfestigkeit = Rest[<< zugfestigkeitfile];
Regress[zugfestigkeit, {1, Dicke, Haspeltemperatur, Mangangehalt},
    {Dicke, Haspeltemperatur, Mangangehalt}, RegressionReport → ParameterCITable][[1, 2]]
Clear[zugfestigkeit]
```

	Estimate	SE	CI
1	532.316	6.22622	{520.086, 544.546}
Dicke	−3.48119	0.277612	{−4.02651, −2.93586}
Haspeltemperatur	−0.243508	0.0093313	{−0.261838, −0.225179}
Mangangehalt	28.8377	3.96827	{21.0427, 36.6327}

■ Leverages

Oft ist man am Einfluss interessiert, den der i-te Messwert $\{x_{i\,1}, x_{i\,2}, \ldots, x_{i\,m}, y_i\}$ auf die Regressionsanalyse hat. Eine Möglichkeit, diesen Einfluss zu beschreiben besteht darin, zu untersuchen, wie stark sich der Vorhersagewert

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\, x_{i\,1} + \hat{\beta}_2\, x_{i\,2} + \ldots + \hat{\beta}_m\, x_{i\,m}$$

der Größe $\tilde{Y}_i$ ändert, wenn man in der Regressionsanalyse den Wert y_i durch $y_i + 1$ ersetzt. Man spricht in diesem Zusammenhang vom Leverage des i-ten Messwerts. Die Leverages können dazu verwendet werden, die Messwerte auf Ausreißer hin zu überprüfen.

12.7.11 Bemerkung: Bezeichnet man den Vorhersagewert der Größe Y_i mit $\hat{y}_i$ bzw $\hat{\tilde{y}}_i$ je nachdem, ob man in der Regressionsanalyse mit dem Wert y_i bzw dem Wert $y_i + 1$ arbeitet, so gilt für den Leverage l_i des i-ten Messwertes

$$l_i = \hat{\tilde{y}}_i - \hat{y}_i = \{1, x_{i\,1}, \ldots, x_{i\,m}\} \cdot \left(\boldsymbol{x}^t \cdot \boldsymbol{x}\right)^{-1} \cdot \{1, x_{i\,1}, \ldots, x_{i\,m}\}^t$$

Man beachte, dass es sich dabei um das i-te Diagonalelement der Matrix $\boldsymbol{x} \cdot \left(\boldsymbol{x}^t \cdot \boldsymbol{x}\right)^{-1} \cdot \boldsymbol{x}^t$ handelt.

▼

Beweis: Aus der Beziehung

$$\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}^t = \boldsymbol{x}\cdot\hat{\vec{\beta}} = \boldsymbol{x}\cdot\left(\boldsymbol{x}^t\cdot\boldsymbol{x}\right)^{-1}\cdot\boldsymbol{x}^t\cdot\vec{y}$$

folgt unmittelbar

$$l_i = \hat{\hat{y}}_i - \hat{y}_i = \{1, x_{i\,1}, \ldots, x_{i\,m}\}\cdot\left(\boldsymbol{x}^t\cdot\boldsymbol{x}\right)^{-1}\cdot\boldsymbol{x}^t\cdot\left(\{y_1, \ldots, y_i + 1, \ldots, y_n\}^t - \{y_1, \ldots, y_i, \ldots, y_n\}^t\right) =$$

$$= \{1, x_{i\,1}, \ldots, x_{i\,m}\}\cdot\left(\boldsymbol{x}^t\cdot\boldsymbol{x}\right)^{-1}\cdot\{1, x_{i\,1}, \ldots, x_{i\,m}\}^t$$

Sollen die Leverages der einzelnen Messwerte mit Hilfe von *Mathematica* ermittelt werden, so verwende man bei der Option RegressionReport des Befehls Regress die Einstellung HatDiagonal:

■ Regress[*daten*, *model*, *vars*, RegressionReport → HatDiagonal]

ermittelt die Diagonalelemente der Matrix $\boldsymbol{x}\cdot\left(\boldsymbol{x}^t\cdot\boldsymbol{x}\right)^{-1}\cdot\boldsymbol{x}^t$ und damit die Leverages aller Messwerte.

12.7.12 Beispiel: Von $n = 30$ zufällig ausgewählten Personen wurde das Alter und der systolische Blutdruck ermittelt (man vergleiche dazu das Datenmaterial blutdruck sowie Beispiel 12.4.5 und Beispiel 12.7.2). Für das Modell Blutdruck = $\gamma_0 + \gamma_1$ Alter^2 stelle man das Datenmaterial zusammen mit der Regressionsparabel und dem Leverage des *i*-ten Messwertes in dynamischer Weise graphisch dar.

▼

Lösung: Wir ermitteln die gesuchten Leverages unter Verwendung der Option RegressionReport des Befehls Regress mit der Einstellung HatDiagonal und erzeugen mittels Manipulate die gesuchte dynamische Graphik:

```
Manipulate[blutdruck = Rest[<< blutdruckfile];
  leverages = Regress[blutdruck, {1, Alter^2}, {Alter}, RegressionReport → HatDiagonal][[1, 2]];
  plot1 = Plot[Evaluate[Fit[blutdruck, {1, Alter^2}, {Alter}]], {Alter, 15, 70}];
  plot2 = ListPlot[blutdruck, PlotStyle → {PointSize[0.025], Red}];
  plot3 = ListPlot[{blutdruck[[i]]}, PlotStyle → {PointSize[0.03], Blue}];
  plot4 = Graphics[Text["Leverage des i−ten Punktes:", {35, 175}]];
  plot5 = Graphics[Text[leverages[[i]], {35, 165}]];
  Show[{plot1, plot2, plot3, plot4, plot5}, AspectRatio → 0.5, PlotRange → All, AxesLabel → {Alter, Blutdruck},
    AxesOrigin → {15, 100}, ImageSize → {250, 150}],
  {i, 1, 30, 1, Appearance → "Labeled"}]
```

12.8 Hilfssätze der Matrizenrechnung

Die Matrizenrechnung ist ein Werkzeug, mit dem sich viele Aussagen, Sätze und Beweise der Regressionsanalyse in prägnanter und übersichtlicher Weise formulieren lassen. Dabei benötigt man neben den üblichen Ergebnissen der Matrizenrechnung einige spezielle Formeln und Eigenschaften, welche in diesem Abschnitt in Form von Hilfssätzen zusammengestellt werden.

Hilfssatz 1: Ist die Matrix $\mathbf{A} \in \mathbb{R}_n^n$ idempotent, gilt also $\mathbf{A}\cdot\mathbf{A} = \mathbf{A}$, so ist der Rang von **A** gleich der Spur von **A** (unter der Spur einer quadratischen Matrix versteht man die Summe der Elemente in der Hauptdiagonale).

▼

Beweis: Besitzt die Matrix **A** den Rang *r*, so existieren bekanntlich zwei invertierbare Matrizen **P** und **Q** mit der Eigenschaft

$$\mathbf{A} = \mathbf{P} \cdot \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{Q}$$

wobei $\mathbf{E}_r \in \mathbb{R}_r^r$ die $r \times r$ Einheitsmatrix bezeichnet und $\mathbf{0}$ Nullmatrizen von entsprechendem Format sind. Bezeichnen wir nun die Spaltenvektoren der Matrix $\mathbf{P}$ mit $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_n$ und die Zeilenvektoren der Matrix $\mathbf{Q}$ mit $\vec{q}_1{}^t, \vec{q}_2{}^t, \ldots, \vec{q}_n{}^t$, so folgt aus der Beziehung $\mathbf{A} \cdot \mathbf{A} = \mathbf{A}$, also

$$\mathbf{P} \cdot \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{Q} \cdot \mathbf{P} \cdot \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{Q} = \mathbf{P} \cdot \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{Q}$$

unmittelbar (die Matrizen $\mathbf{P}$ und $\mathbf{Q}$ sind invertierbar)

$$\begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{Q} \cdot \mathbf{P} \cdot \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Also gilt für alle $i, k \in \{1, 2, \ldots, r\}$ offenbar $\vec{q}_i{}^t \cdot \vec{p}_k = \delta_{i\,k}$ und damit

$$\text{Spur}[\mathbf{A}] = \text{Spur}[\mathbf{P} \cdot \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{Q}] = \text{Spur}[\mathbf{P} \cdot \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{Q}] =$$

$$= \text{Spur}[(\vec{p}_1, \ldots, \vec{p}_r, \vec{0}, \ldots, \vec{0}) \cdot (\vec{q}_1{}^t, \ldots, \vec{q}_r{}^t, \ldots, \vec{0}, \ldots, \vec{0})^t] =$$

$$= \text{Spur}[(\vec{q}_1{}^t, \ldots, \vec{q}_r{}^t, \ldots, \vec{0}, \ldots, \vec{0})^t \cdot (\vec{p}_1, \ldots, \vec{p}_r, \vec{0}, \ldots, \vec{0})] = \text{Spur}[\begin{pmatrix} \mathbf{E}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}] = r$$

Hilfssatz 2: Ist $\vec{Z} = \{Z_1, Z_2, \ldots, Z_n\}^t$ ein $\mathcal{MN}[\vec{0}, \mathbf{E}]$-verteilter Zufallsvektor und sind $\mathbf{A} \in \mathbb{R}_m^n$ und $\mathbf{B} \in \mathbb{R}_n^n$ Matrizen mit der Eigenschaft $\mathbf{B}^t = \mathbf{B}$, $\text{Rg}[\mathbf{B}] = r$ und $\mathbf{A} \cdot \mathbf{B} = \mathbf{0}$, so sind $\mathbf{A} \cdot \vec{Z}$ und $\vec{Z}^t \cdot \mathbf{B} \cdot \vec{Z}$ unabhängig.

▼

Beweis: Die Matrix $\mathbf{B}$ ist symmetrisch mit $\text{Rg}[B] = r$, also gibt es eine orthogonale Matrix $\mathbf{P}$ mit der Eigenschaft

$$\mathbf{P}^t \cdot \mathbf{B} \cdot \mathbf{P} = \begin{pmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

wobei $\mathbf{D}_r \in \mathbb{R}_r^r$ eine $r \times r$ Diagonalmatrix ist und $\mathbf{0}$ Nullmatrizen von entsprechendem Format sind. Mit dem ebenfalls $\mathcal{MN}[\vec{0}, \mathbf{E}]$-verteilten Zufallsvektor $\vec{U} = \{U_1, U_2, \ldots, U_n\}^t = \mathbf{P}^t \cdot \vec{Z}$ gilt damit einerseits

$$\vec{Z}^t \cdot \mathbf{B} \cdot \vec{Z} = \vec{Z}^t \cdot \mathbf{P} \cdot \begin{pmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \mathbf{P}^t \cdot \vec{Z} = \vec{U}^t \cdot \begin{pmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot \vec{U} = f[U_1, U_2, \ldots, U_r]$$

Berücksichtigt man, dass aus der Beziehung $\mathbf{A} \cdot \mathbf{B} = \mathbf{0}$ offenbar

$$\mathbf{A} \cdot \mathbf{P} \cdot \begin{pmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \mathbf{0}$$

folgt, was wiederum zur Folge hat, dass die ersten r Spalten der Matrix $\mathbf{A} \cdot \mathbf{P}$ gleich $\vec{0}$ sind, so gilt andererseits

$$\mathbf{A} \cdot \vec{Z} = \mathbf{A} \cdot \mathbf{P} \cdot \vec{U} = g[U_{r+1}, U_{r+2}, \ldots, U_n]$$

Berücksichtigt man nun, dass die Zufallsvariablen $U_1, U_2, \ldots, U_n$ vollständig unabhängig sind, so folgt die Aussage dieses Hilfssatzes unmittelbar aus der Familieneigenschaft.

Hilfssatz 3: Sind $\mathbf{A} \in \mathbb{R}_n^p$ und $\mathbf{B} \in \mathbb{R}_n^q$ Matrizen und besitzt die (offenbar symmetrische) Matrix

$$\boldsymbol{M} = \begin{pmatrix} \mathbf{A}^t \\ \hline \mathbf{B}^t \end{pmatrix} \cdot (\, \mathbf{A} \mid \mathbf{B} \,) = \left(\begin{array}{c|c} \mathbf{A}^t \cdot \mathbf{A} & \mathbf{A}^t \cdot \mathbf{B} \\ \hline \mathbf{B}^t \cdot \mathbf{A} & \mathbf{B}^t \cdot \mathbf{B} \end{array}\right) \in \mathbb{R}_{p+q}^{p+q}$$

vollen Rang, so gilt für die inverse Matrix $\mathbf{M}^{-1}$ von $\mathbf{M}$

$$\mathbf{M}^{-1} = \left(\begin{array}{c|c} \left(\mathbf{A}^t \cdot \mathbf{A}\right)^{-1} + \left(\mathbf{A}^t \cdot \mathbf{A}\right)^{-1} \mathbf{A}^t \cdot \mathbf{B} \cdot \mathbf{W}^{-1} \cdot \mathbf{B}^t \cdot \mathbf{A} \cdot \left(\mathbf{A}^t \cdot \mathbf{A}\right)^{-1} & -\left(\mathbf{A}^t \cdot \mathbf{A}\right)^{-1} \cdot \mathbf{A}^t \cdot \mathbf{B} \cdot \mathbf{W}^{-1} \\ \hline -\mathbf{W}^{-1} \cdot \mathbf{B}^t \cdot \mathbf{A} \cdot \left(\mathbf{A}^t \cdot \mathbf{A}\right)^{-1} & \mathbf{W}^{-1} \end{array}\right)$$

wobei wir die Abkürzung $\mathbf{W} = \left[\mathbf{B}^t \cdot \mathbf{B} - \mathbf{B}^t \cdot \mathbf{A} \cdot \left(\mathbf{A}^t \cdot \mathbf{A}\right)^{-1} \cdot \mathbf{A}^t \cdot \mathbf{B}\right]$ verwendet haben.

▼

Beweis: a) Die Matrix $\mathbf{A}^t \cdot \mathbf{A}$ ist invertierbar: Wäre $\mathbf{A}^t \cdot \mathbf{A}$ nämlich nicht invertierbar, so wäre der Rang der Matrix $\mathbf{A}^t \cdot \mathbf{A} \in \mathbb{R}_p^p$ und damit auch der Rang der Matrix $\mathbf{A}$ kleiner als p. Dies hätte aber zur Folge, dass der Rang der Matrix $(\, \mathbf{A} \mid \mathbf{B} \,) \in \mathbb{R}_n^{p+q}$ kleiner als $p + q$ wäre, was wiederum zur Folge hätte, dass die Matrix $\mathbf{M}$ nicht vollen Rang haben könnte.

b) Wir nehmen an, dass $\mathbf{M}^{-1}$ die oben angeführte Gestalt besitzt, wobei die Matrix $\mathbf{W} \in \mathbb{R}_q^q$ noch unbekannt ist und zeigen durch einfaches Ausmultiplizieren, dass dies $\mathbf{W} = \left[\mathbf{B}^t \cdot \mathbf{B} - \mathbf{B}^t \cdot \mathbf{A} \cdot \left(\mathbf{A}^t \cdot \mathbf{A}\right)^{-1} \cdot \mathbf{A}^t \cdot \mathbf{B}\right]$ zur Folge hat.