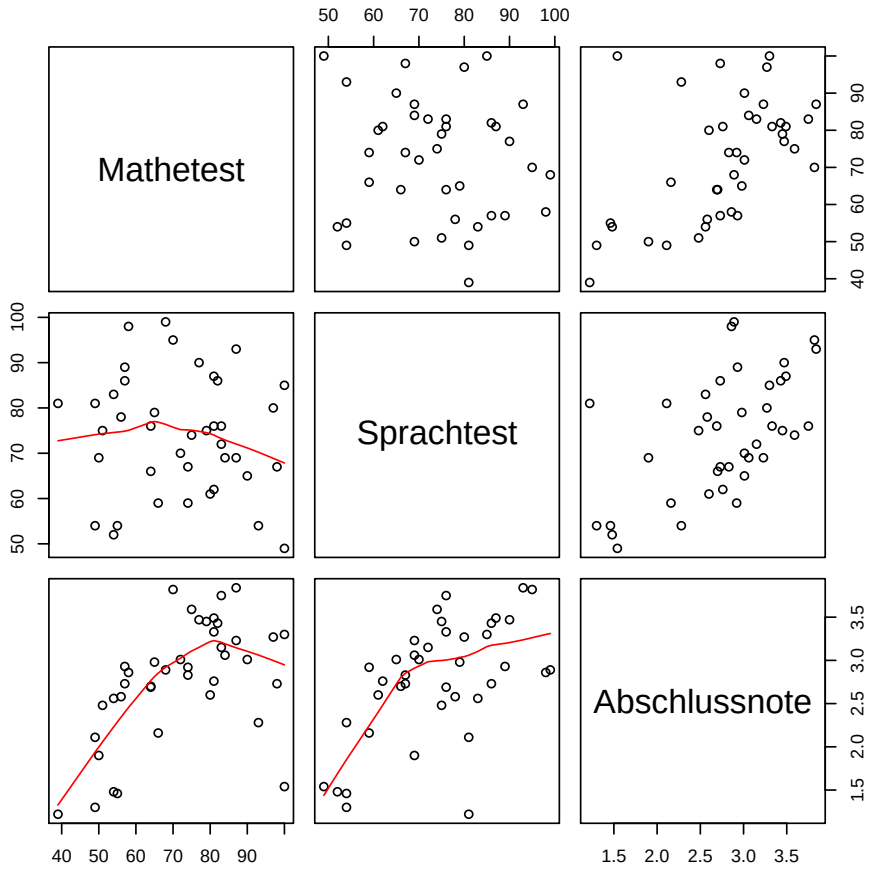
Der Datensatz zu Eingangstests und Abschlussnoten eines Colleges entstammen dem Buch Statistics and Data Analysis von Ajit C. Tamahane und Dorothy D. Dunlop, Prentice Hall, 2000.

Es soll ein Modell entwickelt werden, um den Notendurchschnitt von College Studenten auf der Grundlage der Ergebnisse der Eingangstests zur Feststellung der Mathematik- und Sprachkenntnisse vorherzusagen. Die Testergebnisse sind als Quantilswerte der erreichten Punktzahl angegeben, der Notendurchschnitt bezieht sich auf das amerikanische Punktesystem: hoher Durchschnitt = gute Note Zur Übersicht betrachten wir die Zusammenhänge zwischen den Testergebnissen

Daten im Überblick

```
> plot(Eingangstest, lower.panel = panel.smooth)
```



Einfaches multiples Modell

Die Abhängigkeit der Abschlussnote von den beiden Eingangstestergebnissen legt zunächst ein lineares Modell mit den Variablen Mathetest und Sprachtest als EinflussgröSSen nahe

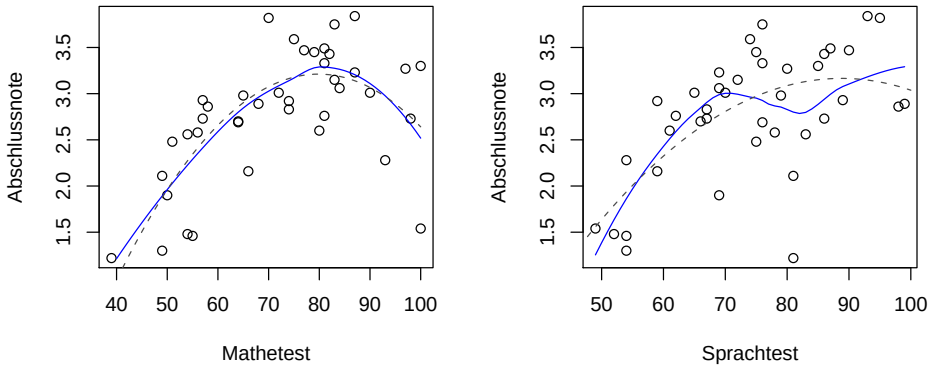
```
> fit <- lm(Abschlussnote ~ Mathetest + Sprachtest, data = Eingangstest)
```

Beide EinflussgröSSen haben signifikante relativ ähnliche Effekte auf die ZielgröSSe. Für den Fehlerterm erhalten wir $\widehat{\sigma}$ =0.4, 68 % der Varianz werden durch das Modell erklärt.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−1.5705	0.4937	−3.18	0.0030
Mathetest	0.0257	0.0040	6.40	0.0000
Sprachtest	0.0336	0.0049	6.82	0.0000

Erweiterung des Modells

Wir können fragen, ob sich die Prognosegüte des Modells noch verbessern lässt. Geht man davon aus, dass sich Leistungen nicht beliebig steigern lassen, könnte man einen quadratischen Einfluss der Vortests vermuten. Diesen Ansatz legen auch die geglätteten Ausgleichskurven nahe.



Eine so genannte Quadratische Oberfläche als Modellfunktion erhalten wir im folgenden Modell. Es gilt zu beachten, dass die Funktion immer noch linear in den Parametern ist, wir also bei einem Linearen Modell bleiben.

```
> fit.quad <- update(fit, . ~ . + I(Mathetest^2) + I(Sprachtest^2) +
+     I(Mathetest * Sprachtest))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−9.9168	1.3544	−7.32	0.0000
Mathetest	0.1668	0.0212	7.85	0.0000
Sprachtest	0.1376	0.0267	5.15	0.0000
I(Mathetest^2)	−0.0011	0.0001	−9.45	0.0000
I(Sprachtest^2)	−0.0008	0.0002	−5.29	0.0000
I(Mathetest * Sprachtest)	0.0002	0.0001	1.67	0.1032

Tatsächlich hat sich die Güte der Anpassung deutlich verbessert: Der Fehlerterm wird als $\widehat{\sigma}$ =0.19 geschätzt und wir erhalten ein BestimmteheitsmaSS von 94%.

Vergleich von linearem und quadratischem Modell

Das quadratische Modell können wir als volles Modell, das lineare als Spezialfall des vollen Modells als reduziertes Modell auffassen. Ein Vergleich der Fehlerquadratsummen ergibt dann einen F-Test, der überprüft, ob die Erweiterung des Modells signifikant ist. Wir berechnen

```
> model.compare <- anova(fit, fit.quad)
```

und erhalten

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37	5.99				
2	34	1.19	3	4.80	45.65	0.0000

den Nachweis, dass der quadratische Anteil einen signifikanten Beitrag zur Modellierung der Daten liefert.

Die geschätzte Modellfunktion

Um die Abhängigkeiten leichter erkennen zu können, betrachten wir die Niveaulinien der Regressionsfunktion. AuSSerdem bestimmen wir das Maximum der Funktion. Im Maximum ist der Gradient der Funktion gleich Null. Für eine quadratische Funktion

$$f(m,s) = \beta_0 + \beta_m\ m + \beta_s\ s + \beta_{ms}\ ms + \beta_{mm}\ m^2 + \beta_{ss}\ s^2$$

gilt

$$grad(f) = (\beta_m + 2\beta_{mm}\ m + \beta_{ms}\ s, \beta_s + \beta_{ms}\ m + 2\beta_{ss}\ s)$$

Setzen wir den Gradienten gleich Null, ergibt sich ein lineares Gleichungssystem

$$\begin{pmatrix} 2\beta_{mm} & \beta_{ms} \\ \beta_{ms} & 2\beta_{ss} \end{pmatrix} \begin{pmatrix} m \\ s \end{pmatrix} = - \begin{pmatrix} \beta_m \\ \beta_s \end{pmatrix}$$

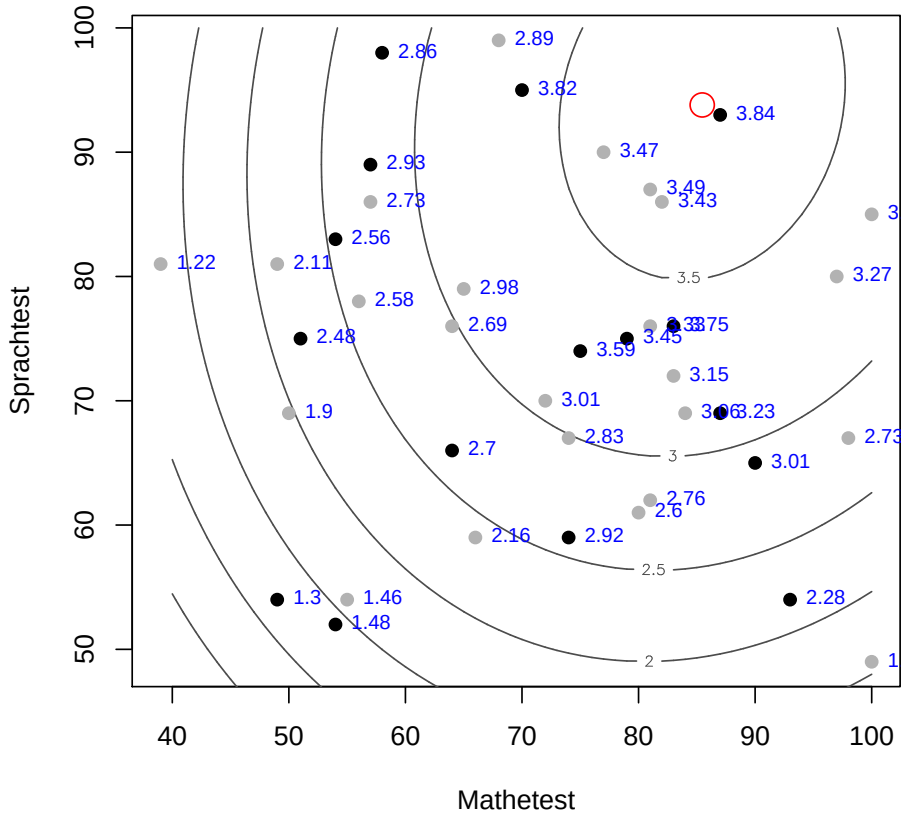
Wir erhalten die Lösung durch die folgenden R-Befehle:

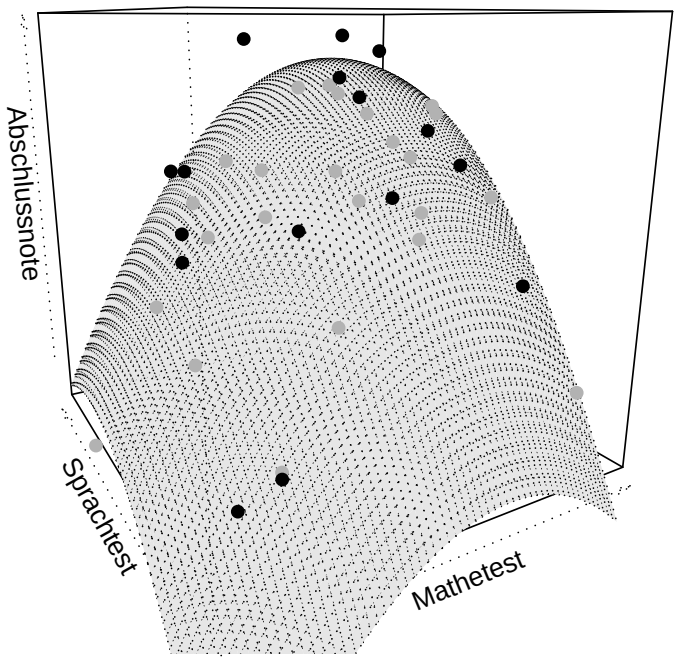
```
> beta <- coef(fit.quad)
> ms <- solve(matrix(c(2 * beta[4], beta[6], beta[6], 2 * beta[5]),
+     2, 2, byrow = T), -beta[2:3])
> round(ms, 1)

 Mathetest Sprachtest
      85.5       93.8

> fmax <- predict(fit.quad, newdata = data.frame(Mathetest = ms[1],
+     Sprachtest = ms[2]))
```

Der geschätzte Wert im Maximum beträgt 3.66.

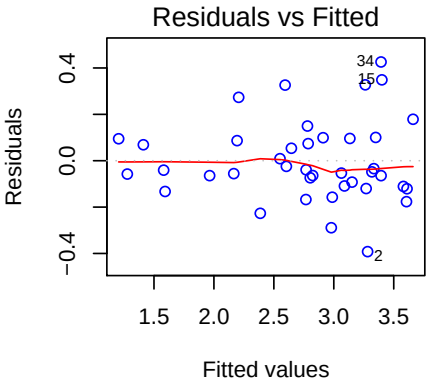

Sprachtest
Mathetest
100
90
80
70
60
50
40
50
60
70
80
90
100
2.89
2.86
3.82
3.84
3.47
2.93
3.49
3.43
2.73
2.56
1.22
2.11
3.27
3.5
2.98
2.58
2.69
2.48
3.59
3.45
3.15
3.01
1.9
3.23
2.83
2.73
2.7
3
3.01
2.76
2.6
2.16
2.92
2.5
1.3
1.46
2.28
1.48
2

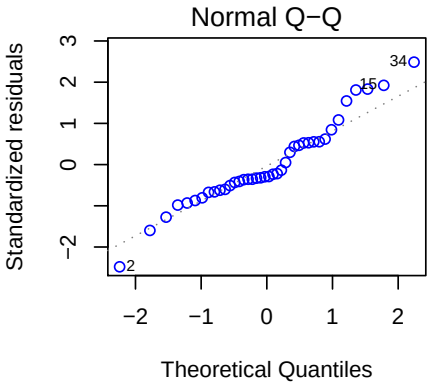


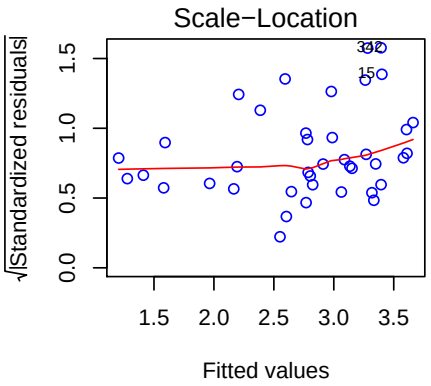
Check der Voraussetzungen

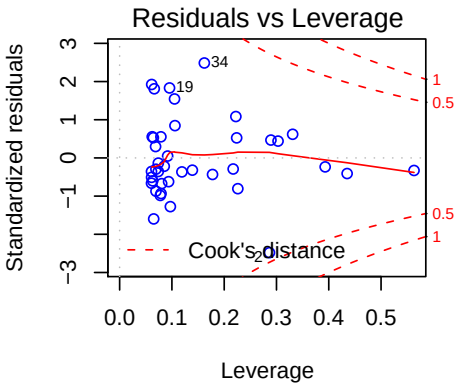
Zum Check der Modellvoraussetzungen betrachten wir die von **R** bereitgestellten diagnostischen Plots:

```
> opar <- par(mfrow = c(2, 2))
> plot(fit.quad, col = 4)
> par(opar)
```


Residuals vs Fitted
Residuals
Fitted values
34
15
2


Normal Q−Q
Standardized residuals
Theoretical Quantiles
34
15
2


Scale−Location
√|Standardized residuals|
Fitted values
34
15


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
34
19
2
1
0.5
0.5
1