

Lösung zu Kapitel 9: Beispiel 4

Im Datenfile `mba.dat` sind folgende Variablen enthalten:

<code>mba_gpa</code>	Punktedurchschnitt im MBA-Programm
<code>undergpa</code>	Punktedurchschnitt im Undergraduate-Kurs
<code>gmat</code>	Punktezahl im Zulassungstest
<code>work</code>	Berufserfahrung in Jahren

Die Leiterin eines MBA-Programms, das vor 20 Jahren gegründet wurde, will analysieren, welche Faktoren die Leistungen der Kursteilnehmer beeinflussen und bestimmen. Die Leistung wird durch den Punktedurchschnitt im MBA-Programm (GPA, grade point average) gemessen, als Einflussfaktoren werden der Punktedurchschnitt im Undergraduate-Kurs, die Punktezahl im Zulassungstest und die Berufserfahrung bei Eintritt in das MBA-Programm untersucht. Von 100 zufällig bestimmten MBA-Kursteilnehmern werden die entsprechenden Daten gesammelt.

- Finden Sie ein passendes Modell zur Prognose der Leistung im MBA-Kurs.

Zuerst lesen wir die Daten ein:

R

```
> mba <- read.table("mba.dat", header = TRUE)
> attach(mba)
```

Mit Streudiagrammen der erklärenden Variablen gegen die Responsevariable `mba_gpa` gewinnen wir etwas Einblick in den Zusammenhang dieser Variablen (linear oder nicht linear) und können auch etwaige Ausreißer lokalisieren.

R

```
> par(mfrow = c(1, 3))
> plot(undergpa, mba_gpa)
> plot(gmat, mba_gpa)
> plot(work, mba_gpa)
> par(mfrow = c(1, 1))
```

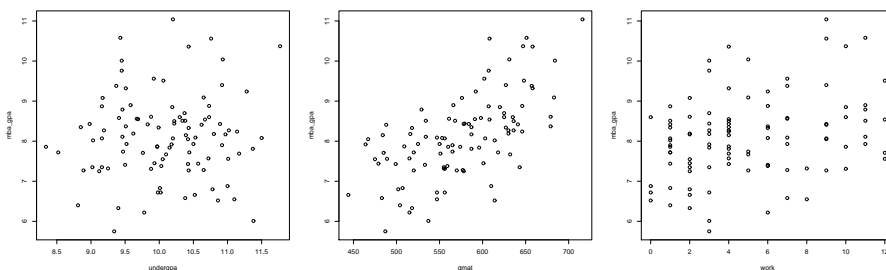


Abbildung 1: Streudiagramme von erklärenden und Responsevariablen.

Von den drei Streudiagrammen zeigt das mit `undergpa` nur einen sehr schwachen, das mit `gmat` einen sehr starken und das mit `work` einen deutlichen Zusammenhang an. Dies wird auch bestätigt, wenn wir ein Modell mit allen drei Variablen gemeinsam als erklärende Variablen schätzen.

R

```
> mba_3vars <- lm(mba_gpa ~ undergrad + gmat + work)
> summary(mba_3vars)
```

Call:

```
lm(formula = mba_gpa ~ undergrad + gmat + work)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.61105	-0.50200	-0.03703	0.49637	1.66796

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.530454	1.323232	0.401	0.689
undergpa	0.082364	0.105269	0.782	0.436
gmat	0.010922	0.001284	8.505	2.40e-13
work	0.092754	0.021956	4.225	5.45e-05

Residual standard error: 0.7541 on 96 degrees of freedom

Multiple R-squared: 0.4881, Adjusted R-squared: 0.4721

F-statistic: 30.52 on 3 and 96 DF, p-value: 6.092e-14

Wir eliminieren wir die nicht signifikante Variable `undergpa`

R

```
> mba_2vars <- lm(mba_gpa ~ gmat + work)
> summary(mba_2vars)
```

Call:

```
lm(formula = mba_gpa ~ gmat + work)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.53708	-0.54600	-0.05509	0.49719	1.71808

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.38380	0.74778	1.851	0.0673
gmat	0.01087	0.00128	8.492	2.39e-13
work	0.09445	0.02181	4.332	3.61e-05

Residual standard error: 0.7526 on 97 degrees of freedom

Multiple R-squared: 0.4849, Adjusted R-squared: 0.4742

F-statistic: 45.65 on 2 and 97 DF, p-value: 1.067e-14

und kommen zu einem Modell mit gutem Erklärungswert für die Daten (das Bestimmtheitsmaß $R^2 = 0.48$). Somit können das lineare Modell mit `gmat` und `work` als erklärende Variablen für `mba_gpa` vorschlagen.

Zum Schluss lösen wir die Verbindung zum Datensatz wieder auf:

R

```
> detach(mba)
```