

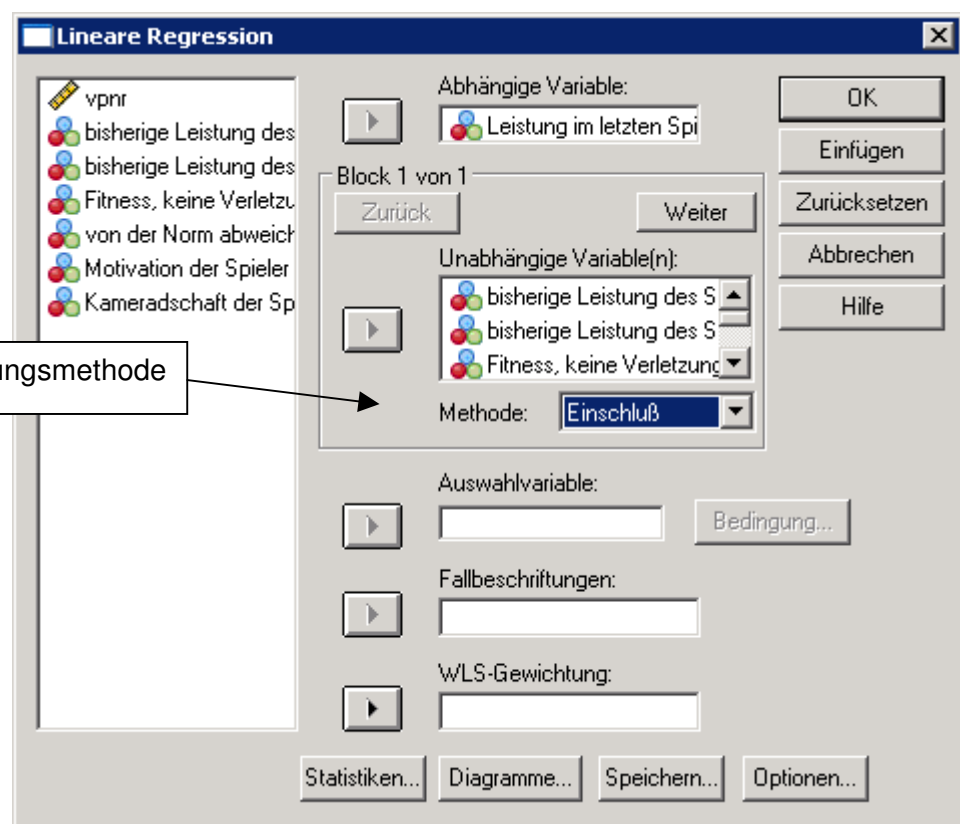
Computergestützte Datenanalyse (Modul M3)

Tobias Heine
Fabian Hölzenbein
Miriam Seel

Mo 10-12 & 12-14 Uhr
Do 12-14 Uhr
Di 12-14 Uhr

Multiple Regression

Die multiple Regression verwendet nicht nur eine Prädiktorvariable, sondern bildet Linearkombinationen aus mehreren Prädiktoren, um eine Kriteriumsvariable vorherzusagen. Bei der Berechnung einer multiplen Regression mit SPSS ist wie bei der einfachen Regression die Kriteriumsvariable Y unter *Abhängige Variable* festzulegen. Die k Prädiktorvariablen sind unter *Unabhängige Variable(n)* anzugeben. Über den Pfad *Analysieren / Regression / Linear* gelangt man zu folgendem Dialogfeld:

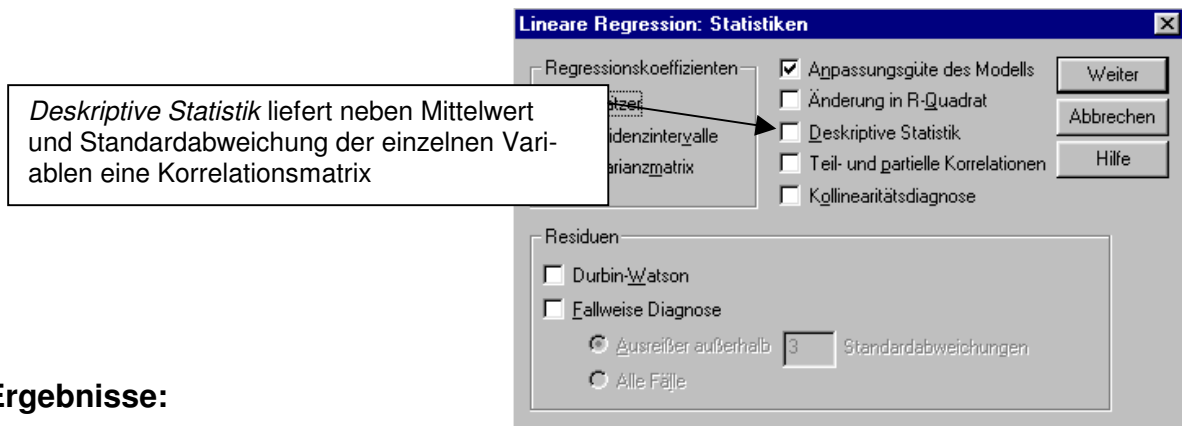


In unserem Beispiel soll anhand von unterschiedlichen Variablen die Leistung von Fußballspielern vorhergesagt werden (Datensatz: KREUZV1.SAV). Es gibt nun zwei Möglichkeiten die Regressionsgleichung zu berechnen. Mit der *Methode: Einschluss* werden alle Variablen, die man unter Unabhängige Variablen angibt in die Berechnung mit einbezogen. Will man hingegen nur Variablen in der Regressionsgleichung haben, die auch tatsächlich einen signifikanten Beitrag zur Vorhersage des Kriteriums liefern, ist als Methode *Schrittweise* zu wählen.

1. Multiple Regression mit der Prozedur Einschluss

Sind alle Prädiktoren, die in die Berechnung eingehen sollen, spezifiziert und ist als Methode Einschluss gewählt, kann über *Statistiken...* Veränderungen an der Ergebnisausgabe vorge-

nommen werden. So werden bei Anklicken von *Deskriptive Statistik* Mittelwerte und Standardabweichung angegeben sowie die Interkorrelationsmatrix aller beteiligten Variablen angegeben.



Ergebnisse:

Modellzusammenfassung

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|--------|-------------------|-----------|------------------------|------------------------------|
| 1 | ,477 ^a | ,228 | ,149 | 1,845 |

a. Einflußvariablen : (Konstante), Kameradschaft der Spieler, von der Norm abweichendes Alter, Fitness, keine Verletzungen, bisherige Leistung des Spielers (Selbsteinschätzung), bisherige Leistung des Spielers (Trainereinschätzung), Motivation der Spieler

Die Tabelle Modellzusammenfassung gibt Auskunft darüber, wie gut die Prädiktoren das Kriterium darstellen können. Die Modellgleichung korreliert zu $R = 0,477$ mit der Kriteriumsvariable und kann damit $R^2 = 22,8\%$ aufklären. Der erwartungstreue Wert für die Populationsschätzung liegt bei $R^2 = 14,9\%$

Die Vorhersage wird signifikant ($F = 2,899$, $p = 0,015$).

ANOVA^b

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Signifikanz |
|--------|------------|--------------|----|---------------------|-------|-------------------|
| 1 | Regression | 59,223 | 6 | 9,871 | 2,899 | ,015 ^a |
| | Residuen | 200,898 | 59 | 3,405 | | |
| | Gesamt | 260,121 | 65 | | | |

a. Einflußvariablen : (Konstante), Kameradschaft der Spieler, von der Norm abweichendes Alter, Fitness, keine Verletzungen, bisherige Leistung des Spielers (Selbsteinschätzung), bisherige Leistung des Spielers (Trainereinschätzung), Motivation der Spieler

b. Abhängige Variable: Leistung im letzten Spiel

| Koeffizienten ^a | | | | | | |
|----------------------------|---|-------------------------------------|----------------|-------------------------------|--------|-------------|
| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | T | Signifikanz |
| | | B | Standardfehler | Beta | | |
| 1 | (Konstante) | 6,109 | 2,553 | | 2,393 | ,020 |
| | bisherige Leistung des Spielers (Trainereinschätzung) | ,261 | ,112 | ,384 | 2,328 | ,023 |
| | bisherige Leistung des Spielers (Selbsteinschätzung) | ,151 | ,112 | ,216 | 1,339 | ,186 |
| | Fitness, keine Verletzungen | -6,20E-02 | ,025 | -,344 | -2,529 | ,014 |
| | von der Norm abweichendes Alter | -,203 | ,155 | -,468 | -1,309 | ,196 |
| | Motivation der Spieler | ,372 | ,244 | ,541 | 1,527 | ,132 |
| | Kameradschaft der Spieler | -1,88E-02 | ,139 | -,017 | -,135 | ,893 |

a. Abhängige Variable: Leistung im letzten Spiel

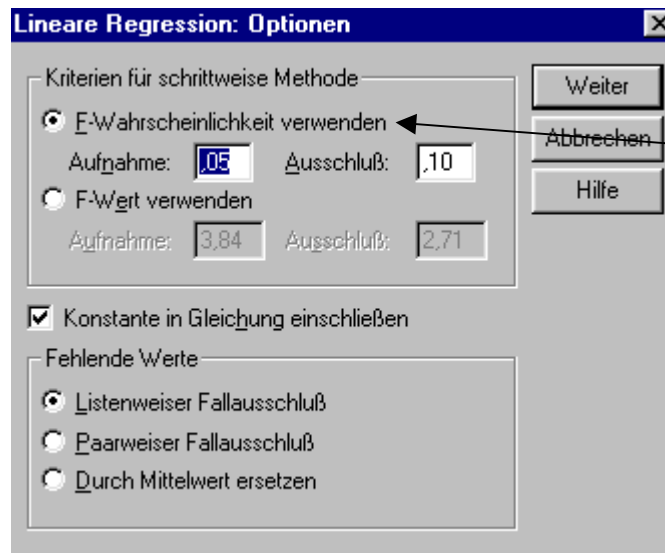
Die standardisierten Partialregressionskoeffizienten unter *Beta* geben Hinweise auf die relative Wichtigkeit der Prädiktorvariablen. Lediglich die bisherige Leistung (Trainereinschätzung) und die Fitness der Spieler liefern signifikante Beiträge zur Varianzaufklärung an dem Kriterium, dennoch müssen auf der Basis dieser Modellberechnung alle Prädiktoren in die Gleichung mit aufgenommen werden:

$$\hat{Y}_{\text{krit}} = 6,109 + 0,261 \cdot X_{\text{Trainereinschätzung}} + 0,151 \cdot X_{\text{Selbsteinschätzung}} - 0,062 \cdot X_{\text{Fitness}} - 0,203 \cdot X_{\text{Alter}} + 0,372 \cdot X_{\text{Motivation}} - 0,0188 \cdot X_{\text{Kameradschaft}}$$

2. Schrittweise Aufnahme der Prädiktoren

Wenn von vorneherein nur Variablen in die Regressionsgleichung aufgenommen werden sollen, die einen signifikanten Beitrag zur Vorhersage des Kriteriums leisten, kann auf die *Methode: Schrittweise* ausgewichen werden. Die Berechnung wird dann auf folgende Weise durchgeführt:

1. Diejenige Variable mit der höchsten Partialkorrelation wird in die Regressionsgleichung aufgenommen. Die Variable muss einen signifikanten Beitrag ($p < 0,05$ = Put-In-Kriterium) zur Vorhersage des Kriteriums leisten.
2. Wenn eine Variable in die Regressionsgleichung aufgenommen wird, verändern sich die β -Gewichte der anderen Variablen. Falls eine Variable in der Gleichung die Signifikanz verfehlt, ($p > 0,10$ = Put-Out-Kriterium) wird sie wieder aus der Gleichung herausgenommen.
3. Diese Schritte werden so lange wiederholt, bis keine Variable mehr in die Gleichung aufgenommen oder ausgeschlossen werden kann.



Möchte man sein eigenes P-IN- und P-OUT-Kriterium festlegen kann man das über den Button *Optionen...*

Die Methode Schrittweise hat den Vorteil, dass redundante Prädiktoren nicht in die Regressionsgleichung aufgenommen werden. Die Gleichung wird einfacher, kann aber trotzdem einen relativ hohen Varianzanteil erklären. Suppressorvariablen (Nachtigall & Wirtz, 2002, Teil 1 S.185) können ebenfalls entdeckt werden.

Dem stehen unter Umständen gravierende Nachteile gegenüber: Zum einen inflationiert bei der großen Zahl von Schritten (Iterationen) und Signifikanztests der α -Fehler. Zum anderen gibt es eine erhöhte capitalisation of chance, da systematisch Prädiktoren bevorzugt werden, die eine hohe Korrelation aufweisen. Je höher die Korrelation, desto höher die Fehlerkomponenten, welche beigesteuert werden. Die schrittweise Regression sollte daher nicht als hypothesentestendes Verfahren, sondern als ein hypothesengenerierendes Verfahren eingesetzt werden.

Ergebnisse:

Modellzusammenfassung

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|--------|-------------------|-----------|------------------------|------------------------------|
| 1 | ,343 ^a | ,118 | ,104 | 1,894 |
| 2 | ,430 ^b | ,185 | ,159 | 1,835 |

- a. Einflußvariablen : (Konstante), bisherige Leistung des Spielers (Trainereinschätzung)
- b. Einflußvariablen : (Konstante), bisherige Leistung des Spielers (Trainereinschätzung), Fitness, keine Verletzungen

Bei der Modellzusammenfassung ist immer das Modell mit der höchsten Varianzaufklärung zu wählen (hier Modell 2)

ANOVA^c

Die Varianzanalyse für Modell
zwei wird signifikant

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Signifikanz |
|--------|------------|--------------|----|---------------------|-------|-------------------|
| 1 | Regression | 30,610 | 1 | 30,610 | 8,536 | ,005 ^a |
| | Residuen | 229,512 | 64 | 3,586 | | |
| | Gesamt | 260,121 | 65 | | | |
| 2 | Regression | 48,053 | 2 | 24,026 | 7,138 | ,002 ^b |
| | Residuen | 212,068 | 63 | 3,366 | | |
| | Gesamt | 260,121 | 65 | | | |

a. Einflußvariablen : (Konstante), bisherige Leistung des Spielers (Trainereinschätzung)

b. Einflußvariablen : (Konstante), bisherige Leistung des Spielers (Trainereinschätzung), Fitness, keine Verletzungen

c. Abhängige Variable: Leistung im letzten Spiel

Koeffizienten^a

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | T | Signifikanz |
|--------|---|-------------------------------------|----------------|-------------------------------|--------|-------------|
| | | B | Standardfehler | Beta | | |
| 1 | (Konstante) | 4,005 | ,484 | | 8,282 | ,000 |
| | bisherige Leistung des Spielers (Trainereinschätzung) | ,233 | ,080 | ,343 | 2,922 | ,005 |
| 2 | (Konstante) | 3,371 | ,545 | | 6,186 | ,000 |
| | bisherige Leistung des Spielers (Trainereinschätzung) | ,333 | ,089 | ,490 | 3,746 | ,000 |
| | Fitness, keine Verletzungen | -5,36E-02 | ,024 | -,298 | -2,276 | ,026 |

a. Abhängige Variable: Leistung im letzten Spiel

Auch hier kann die Regressionsgleichung unter Modell 2 bestimmt werden.

$$\hat{Y}_{\text{krit}} = 3,371 + 0,333 \cdot X_{\text{Trainereinschätzung}} - 0,0536 \cdot X_{\text{Fitness}}$$

Allerdings werden aufgrund der Methode *Schrittweise* nur Variablen aufgenommen, die auch tatsächlich einen signifikanten Beitrag an der Varianzaufklärung liefern. Eine umfangreiche Tabelle mit den ausgeschlossenen Variablen wurde aus Platzgründen weggelassen, da sie nur unwichtige Informationen enthält.

Zum Abschluss die Syntax für die zwei Varianten Einschluss und Schrittweise:

| Einschluss | Schrittweise |
|--|---|
| REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT leist /METHOD=ENTER spqual squalsel fitness alter motiv kamerad . | REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT leist /METHOD=STEPWISE spqual squalsel fitness alter motiv kamerad . |

3. Übungsaufgabe:

Mit Hilfe einer Regressionsanalyse soll die Aussagekraft der Ergebnisse einer Abschlussprüfung einer Schule für Bürokräfte in bezug auf die spätere Berufslaufbahn untersucht werden. Die Variablen des Datensatzes Eign2.sav sind:

vp: Versuchspersonennummer

krit: Experteneinschätzung der Arbeitszeugnisse (Ratings von 1 –10)

masch: Leistung Maschinenschreiben

steno: Leistung beim Stenographietest

sozial: Verhalten und Mitarbeit im Unterricht

Rechnet für den Datensatz eine Regressionsanalyse mit der Methode *Einschluss* und *Schrittweise* und versucht jeweils folgende Fragen zu beantworten:

1. Wie viele Versuchspersonen gehen in die Regressionsanalyse ein?
2. Wie groß ist die multiple Korrelation zwischen dem Kriterium und den aufgenommenen Prädiktoren?
3. Wie viel Prozent der Varianz des Kriteriums wird durch alle (aufgenommenen) Prädiktoren erklärt?
4. Wie lautet die Regressionsgleichung?
5. Wie lautet die standardisierte Regressionsgleichung?
6. Um wie viel Prozent steigt der Anteil der erklärten Varianz am Kriterium durch Aufnahme der Variablen **stenographie**? (Nur für Schrittweise zu beantworten!)