Grundbegriffe der medizinischen Biometrie und Epidemiologie (Sommer 2011)

Übungsblatt 8 am 23.6.2011

1 Aufgabe: Variablenselektion in der Diskriminanzanalyse

Die lineare Diskriminanzanalyse (LDA) setzt ein Mischmodel für p-dimensionale Daten $\boldsymbol{x}$ voraus

$$f(\boldsymbol{x}) = \sum_{j=1}^{K} \pi_j f(\boldsymbol{x}|j),$$

wobei jede der K Klassen durch eine multivariate Normalverteilung repräsentiert wird:

$$f(\boldsymbol{x}|k) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \times \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\}$$

mit klassenspezifischen Mittelwerten $\boldsymbol{\mu}_k$ und einer in allen Klassen gleichen Kovarianzmatrix $\boldsymbol{\Sigma}$. Die Wahrscheinlichkeit für die k-te Gruppe, gegeben die Daten $\boldsymbol{x}$ und die Mischungsgewichten π_j ergibt sich durch Anwendung des Bayes'schen Satzes als

$$\Pr(k|\boldsymbol{x}) = \frac{\pi_k f(\boldsymbol{x}|k)}{f(\boldsymbol{x})}.$$

Der LDA Klassifizierungsscore ergibt sich Logarithmus dieser Wahrscheinlichkeit. $d_k(\boldsymbol{x}) = \log\{\Pr(k|\boldsymbol{x})\}$, der sich zu

$$d_k^{\mathrm{LDA}}(\boldsymbol{x}) = \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k)\,. \tag{1}$$

vereinfacht. Hat man nun eine Stichprobe $\boldsymbol{x}$ gegeben wählt man unter allen Klassen, diejenige aus, welche den Klassifizierungsscore maximiert.

Im Fall zweier Klassen, also $K = 2$ ergibt die Differenz $\Delta^{\mathrm{LDA}}(\boldsymbol{x}) = d_1^{\mathrm{LDA}}(\boldsymbol{x}) - d_2^{\mathrm{LDA}}(\boldsymbol{x})$ zwischen den Klassifizierungsscores der zwei Klassen eine einfache Vorhersageregel:

Wenn $\Delta^{\mathrm{LDA}} \geq 0$ ist, wird die Stichprobe Klasse 1, zugewiesen,
sonst Klasse 2.

(a) Weisen Sie nach, dass nach Zerlegung der Kovarianzmatrix $\boldsymbol{\Sigma}$ in $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$ mit Korrelationen $\boldsymbol{P} = (\rho_{ij})$ und Varianzen $\boldsymbol{V} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$ gilt:

$$\Delta^{\mathrm{LDA}}(\boldsymbol{x}) = \boldsymbol{\omega}^T\boldsymbol{\delta}(\boldsymbol{x}) + \log(\frac{\pi_1}{\pi_2})$$

mit

$$\boldsymbol{\omega} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

und

$$\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}).$$

(b) Hat man eine sehr hohe Dimension der Daten, d.h. ist p sehr groß, will man oft Variablen selektieren, d.h. eine Teilmenge der p Variablen auswählen. Begründen Sie, dass die Einträge des Vektors $\boldsymbol{\omega}$ dafür geeignet sind. Welcher bekannten Teststatistik entsprechen diese Einträge im Fall $\boldsymbol{P} = I$?

2 Aufgabe: Lineare Einfachregression und Korrelation

(a) Leiten Sie die Beziehung zwischen dem Regressionskoeffizient und Korrelation für die Variablen X und Y her.

(b) Zeigen Sie, dass der Regressionskoeffizient von X auf Y unterschiedlich sein kann zu dem Regressionskoeffizient von Y auf X.

3 Aufgabe: Regressionsmodell-Output

Im folgenden wird die Wegstrecke untersucht, die ein Spielzeugauto zurückgelegt hat, nachdem man es in unterschiedlichen Winkeln eine Rampe herunterfahren ließ.

Der Datensatz enthält folgende Variablen:

- *distance*: Wegstrecke, den ein Auto von einer Rampe herab gefahren ist.
- *angle*: Winkel der Rampe.

Angle	Distance
1.3	0.37
4.0	0.92
2.7	0.64
2.2	0.70
3.6	0.89
4.9	1.30
0.9	0.38
1.1	0.43
3.1	0.69

Das Statistikprogramm `R` liefert folgende Ergebnisse:
`Coefficients:`

	`Wert`	`Standardfehler`	`t-Wert`	`Pr(>|t|)`	
β_0	`0.14811`	`0.06503`			
β_1	`0.20954`	`0.02203`			

(a) Testen Sie, ob die Regressionskoeffizienten signifikant von 0 verschieden sind. Interpretieren Sie diese Ergebnisse.

(b) Berechnen Sie auch das R^2 und interpretieren Sie es.

4 Aufgabe: Diagnosegraphiken für Regression

(a) Wiederholen Sie die Voraussetzungen für die lineare Regression. Wie können Sie diese Annahmen mittels graphischer Verfahren untersuchen?

(b) In den folgenden Graphiken finden Sie Diagnosegraphiken für verschiedene lineare Einfachregressionen. Interpretieren Sie, ob in diesen Graphiken Abweichungen von den Annahmen zu erkennen sind.

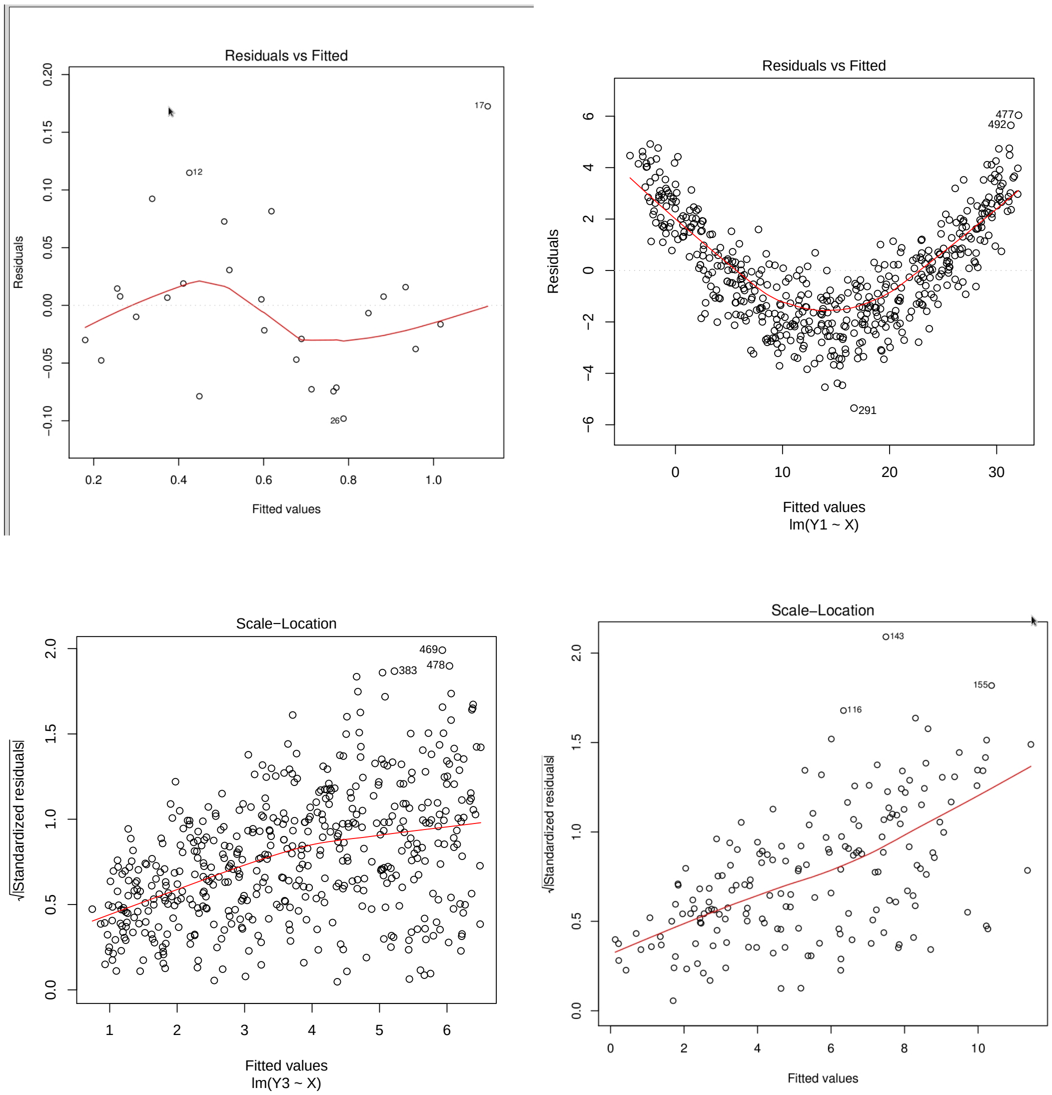

Abbildung 1: Diagnoseplots (a) Residuen gegen gefittete Werte (b) Standardisierte Residuen gegen gefittete Werte

Übungsleiter:
Bernd Klaus (Dipl. Wi-Math) Mail: *bernd.klaus@uni-leipzig.de*
Verena Zuber (M.Sc.) Mail: *vzuber@uni-leipzig.de*