Modul Statistische Aspekte der Analyse molekularbiologischer und genetischer Daten

Übungsblatt 3: Statistische Konzepte in der Genetik

Janne Pott

WS 2021/22

Sie können Ihre Lösungen zu Aufgabe 3 & 4 als PDF in Moodle hochladen (Frist: 29.11.2021).

Aufgabe 1: Lineare Regression

Von fünf Personen sind Größe und Gewicht bekannt:

Größe (cm)	180	175	160	170	190
Gewicht (kg)	80	80	58	60	85

Die Gleichung für die Residuenquadratsumme (RSS) ist:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [(\beta_0 + \beta_1 x_i) - y_i]^2$$

a) Bestimmen Sie den Schätzer $\hat{\beta_0}$ und $\hat{\beta_1}$, indem Sie RSS minimieren! Hinweis: Partielle Ableitung bzgl. β_0 und β_1, Nullsetzen und geeignet umformen.
b) Schätzen Sie die Koeeffizienten β_0 und β_1 für den obigen Datensatz!
c) Welches Gewicht können Sie für eine 176 cm große Person vorhersagen?

Aufgabe 2: Hauptkomponentenanalyse

Betrachten Sie die Matrix A:

$$A = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 2 & 0 \\ -2 & 0 & 4 \end{pmatrix}$$

a) Berechnen Sie das charakteristische Polynom $det(A - \lambda E)$ und geben Sie die daraus resultierenden Eigenwerte $\lambda_i, i \in 1, 2, 3$, an.
b) Bestimmen Sie die zugehörigen Eigenvektoren w_i mit $(A - \lambda_i E) * w_i = 0$
c) Prüfen Sie die Eigenvektoren auf Orthogonalität und normieren Sie auf Länge 1.
d) Überprüfen Sie Ihre Ergebnisse, indem Sie $Q * \Lambda * Q^{-1}$ ausrechnen, wobei Λ die Diagonalmatrix der Eigenwerte und Q die normalisierte Matrix aus den Eigenvektoren ist. Stimmt Ihr Ergebnis mit A überein?

Aufgabe 3: Hardy-Weinberg-Gleichgewicht

Für einen SNP mit 2 Allelen (A, B) wird folgende Genotypverteilung beobachtet:

Genotyp	AA	AB	BB	Missings
Anzahl	824	1326	463	87

a) Bestimmen Sie die Allelfrequenzen p und q von A und B.
b) Stellen Sie die Hardy-Weinberg-Gleichung auf und berechnen Sie mittels der beobachteten Allelfrequenzen die zu erwarteten Genotypfrequenzen.
c) Testen Sie H_0: Die beobachteten Haufigkeiten der Genotypen sind im HWE (Signifikanzniveau von 5%).
d) Warum ist das HWE in der genetischen Statistik relevant?

Aufgabe 4: Linkage disequilibrium

Die Haplotypen zweier SNPs werden miteinander verglichen. Dabei entsteht die Matrix $t = \begin{pmatrix} u & v \\ v & u \end{pmatrix}$ bzw. eine Vierfeldertafel:

	SNP 1 Allel A	SNP 1 Allel a
SNP 2 Allel B	u	v
SNP 2 Allel b	v	u

a) Geben Sie die Randverteilung an und interpretieren Sie die Tafel! Welche Aussagen kann man über die Allelfrequenzen treffen?
b) Zeigen Sie, dass für t gilt: $D'(t) = r(t) = Y(t)$
c) Ab welchem u würde der LD-Threshold von 0.5 überschritten?

Betrachten Sie nun die Vierfeldertafeln der zwei SNPs aus der ersten Übung, rs8176747 und rs8176719, für europäische und afrikanische Samples getrennt:

EUR	rs8176747-C	rs8176747-G
rs8176719-T	609	0
rs8176719-TC	312	85

AFR	rs8176747-C	rs8176747-G
rs8176719-T	891	42
rs8176719-TC	207	182

d) Geben Sie die Randverteilung an und wandeln Sie die absoluten in relative Häufigkeiten um.
e) Berechnen Sie D' und r^2 und interpretieren Sie die Ergebnisse.