ETH Zürich
Statistik II (BIOL, HST) **Serie 9 - 01**

1. Bei 50 Personen soll eine Zielgrösse (y) durch eine erklärende Variable (x) und die Gruppenzugehörigkeit (g) erklärt werden.

   Die Daten sind in folgendem csv-File gespeichert: ueb387177.csv.

   Beachten Sie für folgende Fragen auch das Streudiagramm in der Abbildung unten.

   Welche der folgenden Aussagen sind korrekt?

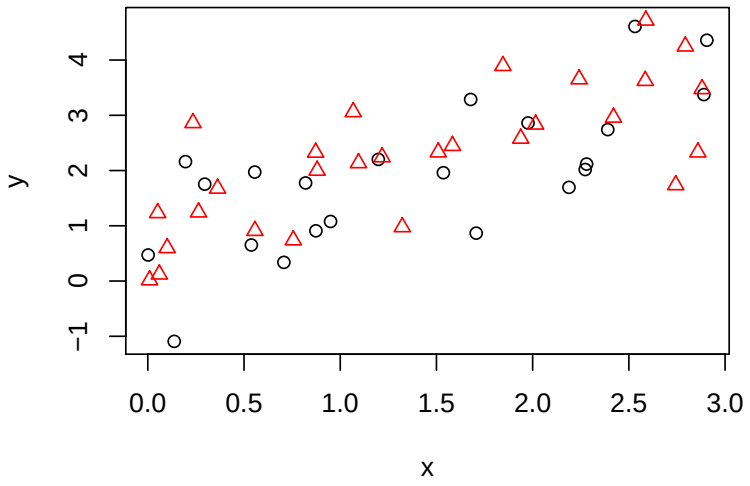

Figure 1: Streudiagramm.

   (a) Laden Sie die Daten. Folgender Wert (gerundet) steht in der Zeile 37 (Spaltennamen zählen nicht als Zeile) und dritten Spalte: 5.0837.

   (b) Der Effekt der Gruppe ist signifikant.

   (c) Die Wechselwirkung von Gruppe und `x` ist signifikant.

   (d) Gemäss Streudiagramm sollte man ein Modell ohne Wechselwirkung verwenden.

   (e) Für eine neue Person mit $x = 2.755$ und $g = W$ ist die Vorhersage für y laut dem linearen Modell (mit Wechselwirkung): 3.8096

2. Es soll eine Zielgrösse (y) durch mehrere erklärende Variablen (x) vorhergesagt werden. Es wird vermutet, dass nicht alle erklärenden Variablen für eine gute Vorhersage nötig sind. Das Ziel dieser Aufgabe ist es, ein Subset der erklärenden Variablen zu finden, welches eine möglichst gute Vorhersage zulässt.

   Die Daten sind in folgendem rda-File gespeichert: *ueb*302411.*rda*. Laden Sie dieses File mit dem Befehl `load` oder über das Menü in RStudio. Das rda-File enthält zwei data frames: `dat` enthält Trainingsdaten mit einer Zielgrösse `y` und mehreren erklärenden Variablen (`x1`, `x2`, etc.). `datTest` ist ein Testdatensatz mit einer Beobachtung, bei dem der Wert für die Variable `y` fehlt (`NA`).

(a) Laden Sie die Daten. Im Trainingsdatensatz steht in Zeile 76 der Daten der Variable x5 der Wert 0.826 .

(b) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit best subset selection). Dieses Modell besteht aus den Variablen x2 x8 x9 .

(c) Nehmen Sie an, dass das beste Modell ausschliesslich die Variablen x1 x3 (und einen Achsenabschnitt) enthält. Dieses Modell sagt auf dem Testdatensatz den Wert 0.557 voraus.

(d) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit forward selection). Dieses Modell besteht aus den Variablen x2 x8 x9 .

(e) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit backward selection). Dieses Modell besteht aus den Variablen x2 x7 x8 x9 .

3. Bei einer Gruppe von Personen soll eine binäre Zielgrösse ($y = 0$ oder $y = 1$) durch eine erklärende Variable (x) und das Geschlecht ($g =$ "M" oder $g =$ "W") erklärt werden.

   Die Daten sind im data frame `dat` in folgendem rda-File gespeichert: ueb107593.rda.

   Passen Sie ein logistisches Regressionsmodell mit Wechselwirkung an.

   Welche der folgenden Aussagen sind korrekt?

   (a) Im data frame `dat` sind 441 Datenpunkte enthalten.

   (b) Der Effekt der Gruppe ist signifikant.

   (c) Die Wechselwirkung von g und x ist signifikant.

   (d) Verwenden Sie nun die Logistische Regression mit der Formel `y ~ g + x` (unabhängig von Ihren vorherigen Resultaten). Wenn man von der Gruppe der Frauen in die Gruppe der Männer wechselt, verändern sich die odds für $y = 1$ um den Faktor 0.425.

   (e) Verwenden Sie wieder die Logistische Regression mit der Formel `y ~ g + x` (unabhängig von Ihren vorherigen Resultaten). Die log-odds für $y = 1$ für eine Frau mit $x = -0.8809$ werden als 0.6946 vorhergesagt.

4. Eine grosse Fast-Food Kette hat auf der ganzen Welt Filialen. Wenn ein neues Restaurant eröffnet wird, steigt die Besucherzahl in den ersten Tagen auf Grund von Marketingmassnahmen in etwa linear an. Wir untersuchen diese anfängliche Besucherzunahme bei einigen Restaurants.

   Die Daten sind im csv-File *ueb*712232*.csv* gespeichert. Laden Sie die Daten und passen Sie daran ein RIRS (random intercept random slope) Modell an.

   Welche der folgenden Aussagen sind korrekt?

   (a) Gemäss dem Datensatz waren am 4-ten Tag im Restaurant Nummer 11 genau 183 Besucher anwesend.

   (b) Im Mittel über alle Restaurants gab es am Eröffnungstag etwa 113 Besucher

   (c) Ein 95%-Vertrauensintervall für die mittlere Zunahme der Gäste pro Tag während der Anfangsphase des Restaurants ist etwa von 11.9 bis 24.3.

   (d) Der Besucheranstieg pro Tag auf Grund der Marketingmassnahme war in den Restaurants nicht überall gleich. Die Schwankung der Besucheranstiege ist ca. 5.2 Besucher pro Tag.

   (e) Man kann erkennen, dass Restaurants mit einer überdurchschnittlichen Besucherzahl am Eröffnungstag einen unterduchschnittlichen Besucherzuwachs in den Folgetagen hatten.

   (f) Achsenabschnitt und Steigung der Geradengleichung für Restaurant Nummer 1 lauten: 130.9 und 22.1

5. Bei 115 Personen soll eine Zielgrösse (y) durch die Gruppenzugehörigkeit (g) erklärt werden.

   Die Daten sind in folgendem rda-File gespeichert: ueb704805.rda.

   Welche der folgenden Aussagen sind korrekt?

   (a) Machen Sie eine 1-weg ANOVA. Der p-Wert ist $1.691e-09$.

   (b) Wir führen paarweise Vergleiche mit dem Tukey Honest Significant Difference Test durch. Das (korrigierte) 95%-Vertrauensintervall für die Differenz E-C geht von -0.9896 bis 0.3883.

   (c) Wir betrachten nun in dieser und der nächsten Teilaufgaben Kontraste und zugehörige korrigierte p-Werte. Erstellen Sie ein Set von 2 Kontrasten: der 1. Kontrast für den Vergleich von Gruppe (C,D) mit der Gruppe (A,B,E) und der 2. Kontrast für den Vergleich innerhalb der Gruppe (C,D). Der p-Wert für den 1. Kontrast ist $2e-04$. (Tipp: Benutzen Sie die Funktion `glht` aus dem Paket `multcomp`)

   (d) Der p-Wert für den 2. Kontrast ist 0.9539.

6. Diese Aufgabe besteht aus zwei Teilen: Zunächst sollen Sie anhand eines Interaction Plots beurteilen ob Wechselwirkungen vorhanden sind. Anschliessend sollen Sie Daten laden und mit einer Varianzanalyse analysieren. Die Daten und der Interaction Plot haben *keinen* Zusammenhang.

   Die Daten sind in folgendem rda-File gespeichert: *ueb*329346.*rda*.

   Es wurde untersucht, wie ein Medikament bzw. Placebo (Variable `m`) bei Männern bzw. Frauen (Variable `g`) wirkt. Die Zielgrösse ist in der Variable `y` gespeichert. Passen Sie ein 2-weg ANOVA Modell mit Interaktion an die Daten an.

   Welche der folgenden Aussagen sind korrekt?

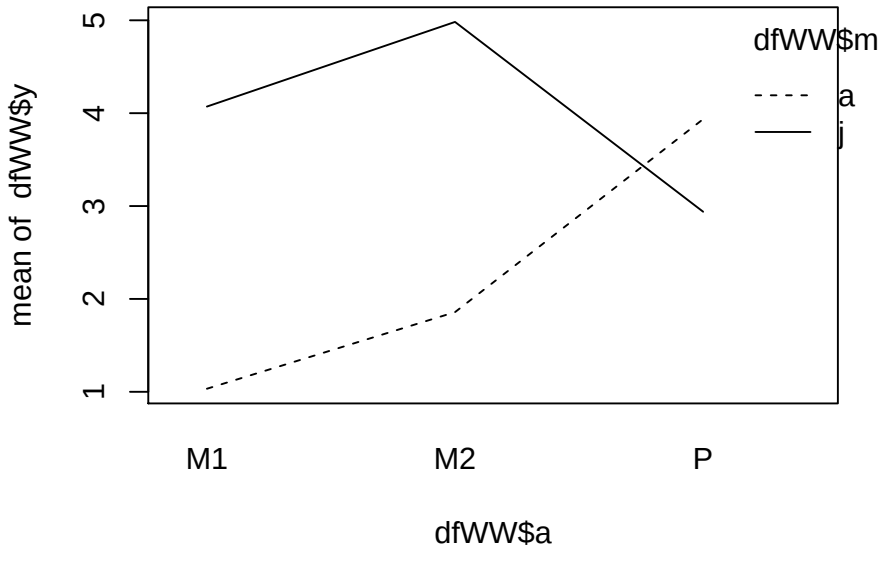

Figure 2: Interaction plot (nur für die erste Teilaufgabe).

(a) Ausgehend vom WW-Plot ist wahrscheinlich eine Wechselwirkung vorhanden.

(b) In Zeile 156 des Datensatzes stehen die Daten von einer Frau, die mit dem Medikament behandelt wurde.

(c) Der p-Wert für den Medikamenteneffekt ist ungefähr 0.937 .

(d) Der Medikamenteneffekt ist nicht geschlechterspezifisch (p-Wert ist nicht signifikant auf 5%-Niveau).

(e) Ein (adjustiertes) 95%-Vertrauensintervall für den mittleren Unterschied zwischen der Gruppe der Frauen mit Medikament und der Gruppe der Männer mit Medikament ist von -0.674 bis 0.12 .

7. In einer klinischen Studie mit 193 Patienten wurde entweder die Standardbehandlung oder ein neues Medikament angewendet. Nach einer Woche wurde festgehalten, bei welchen Patienten das Medikament gewirkt hat.

   Die Daten sind in folgendem rda-File gespeichert: ueb550317.rda.

   Untersuchen Sie die Daten im Folgenden mit einem zweiseitigen Fisher-Test.

   Welche der folgenden Aussagen sind korrekt?

   (a) Stellen Sie die Daten in einer Tabelle dar. 92 Personen haben das neue Medikament erhalten.

   (b) Untersuchen Sie, ob es einen signifikanten Zusammenhang zwischen Heilung und Art des Medikaments gibt. Der p-Wert ist 0.1817.

   (c) Das 95%-Vertrauensintervall für die odds ratio geht von 0.8093 bis 2.844.

8. Welche der folgenden Aussagen sind korrekt (Genauigkeit der angegebenen Macht plus/minus 5%)?

   (a) Ein neues Medikament soll getestet werden. Bei 14 (kranken) Patienten wird das Medikament angewendet. Nach einer Woche wird festgestellt, welche Patienten geheilt wurden. Das Medikament ist wirtschaftlich interessant, wenn die Wirkwahrscheinlichkeit grösser als 0.15 ist (einseitiger Binomialtest, Signifikanzniveau 0.01). Die Macht in dieser Studie für die konkrete Alternative pA=0.35 ist ca. 0.06.

   (b) In einer Getränkefabrik sollen 0.53 Liter abgefüllt werden. Nach einem Stromausfall soll getestet werden, ob die Einstellung geändert wurde (zweiseitiger ein-Stichproben t-Test, Sign.niveau 0.05). Der Hersteller der Abfüllmaschine gibt an, dass die einzelnen Abfüllungen eine Std.abw. von 0.03 haben. Es sollen 9 Flaschen bzgl. ihrer Abfüllmenge untersucht werden. Eine Einstellung der Abfüllmenge von 0.48 kann mit Wahrscheinlichkeit 0.45 erkannt werden.

   (c) Wir wollen untersuchen, ob untrainierte Personen nach einem neuartigen Trainingsprogramm eine bessere Ausdauerleistung erbringen als nach einem herkömmlichen Trainingsprogramm. Die Messung der Ausdauerleistung wird nach dem Programm in beiden Gruppen mit dem gleichen Verfahren durchgeführt (kontinuierliche Skala von 0 bis 100; wir nehmen in jeder Gruppe eine Standardabweichung von 20 an). Anschliessend wollen wir einen ungepaarten, zweiseitigen zwei-Stichproben t-Test mit dem Sign.niveau 0.01 durchführen (H0: Beide Trainingsmethoden sind gleich gut). Angenommen, wir interessieren uns für die konkrete Alternative: muHerkömmlich = 40, muNeu = 55. Wenn wir pro Gruppe 9 Testpersonen einsetzen, haben wir eine Macht von 0.01.

   (d) Wir vergleichen 4 Bakterienstämme auf Resistenz gegen ein Antibiotikum. Pro Bakterienstamm werden 21 Petrischalen mit der gleichen Bakterienmenge angesetzt. Anschliessend wird überall die gleiche Menge vom Antibiotikum beigegeben. Nach einer vorgegebenen Zeit wird die Menge der Bakterien pro Petrischale bestimmt (skalierte

Einheit zw. 0 und 100). Die Standardabweichung der Bakterienmengen pro Bakterienstamm nehmen wir als 40 an. Wir prüfen nun mit einer 1-weg ANOVA die Nullhypothese, dass die mittlere Bakterienmenge am Ende des Experiments für alle Stämme gleich ist (Sign.niveau 0.05). Angenommen, wir interessieren uns für die konkrete Alternative, bei der alle Bakterienstämme am Ende des Experiments den mittleren Wert 33 haben, nur ein Bakterienstamm hat den mittleren Wert 65 (resistent). Die Macht für diese Alternative ist 0.9.

(e) Wir haben ein Verfahren entwickelt, mit dem wir den Blutzuckergehalt in einer Speichelprobe bestimmen wollen. Um nachzuweisen, dass zwischen dem Marker im Speichel (x) und dem Blutzuckergehalt (y) wirklich ein Zusammenhang besteht, untersuchen wir bei 34 Personen sowohl eine Speichel- als auch eine Blutprobe mit einer einfachen linearen Regression; prüfe dazu ob die Steigung auf Sign.niveau 0.05 von Null verschieden ist. Die Standardabweichung für den Fehler in der Linearen Regression nehmen wir als 2.3 an; die Menge des Markers im Speichel kann so genau gemessen werden, dass sie als exakt angenommen werden kann. Die Macht für die konkrete Alternative y = 1.8 + 0.9 *x ist 0.24 .

(f) Ein neues Medikemant (Wirkungswa. pM) wird mit einem Placebo (Wirkungswa. pP) verglichen. 77 Patienten bekommen das Placebo, 36 Patienten bekommen das neue Medikament. Anschliessend halten wir fest, welche Patienten nach einer Woche gesund geworden sind (zweiseitiger Fisher-Test, Sign.niveau 0.01). Angenommen, wir interessieren uns für die konkrete Alternative pM=0.46 und pP=0.3, dann ist die Macht für diese Alternative 0.13.