מודלים סטטיסטיים ויישומיהם 52518 תשע"ח – פתרון בוחן בית 3

להגשה עד 24.1.18 בשעה 23:55

1. הרעיון של ANOVA דו-כווני ניתן להכללה ל-ANOVA רב כווני. למשל, מודל ANOVA תלת-כווני נתון לפי

$$Y_{ijkl} = \mu_{ijk} + \epsilon_{ijkl}$$

כאשר $\epsilon_{ijkl} \underset{iid}{\sim} N(0,1)$ ויכול להרשם בצורה

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \zeta_{ik} + \eta_{jk} + \theta_{ijk} + \epsilon_{ijkl}$$

עם המשקלות $\{\pi_i\},\{\tau_j\},\{\nu_k\}$. תהי H_0 ההשערה כי כל גורמי האינטראקציות הדו-כווניות והתלת-כווניות הן 0.

א. הוכיחו כי אם H_0 מתקיימת עבור סט משקלות מסוים, היא מתקיימת לכל סט משקלות.

ב. כתבו תכנית R המקבלת נתונים ומפיקה p-Value עבור ההשערה H_0. תוכלו להשתמש בפעולות מובנות על וקטורים\מטריצות אך אסור לכם להשתמש בפונקציות סטטיסטיות מובנות. כתבו הסבר (באורך 1-2 דפים) המסביר את אופן הפעולה של התכנית שכתבתם. הריצו את התכנית על הנתונים בקובץ popcorn.csv (הסבר מצוי בקובץ popcorn.txt).

א. יהיו משקלות בלשהן $\{\pi_i\},\{\tau_j\},\{\nu_k\}$. מתקיים:

$$\bar{\mu}_{i..} = \sum_j \sum_k \tau_j \nu_k \mu_{ijk}\,, \qquad \bar{\mu}_{.j.} = \sum_i \sum_k \pi_i \nu_k \mu_{ijk}\,, \qquad \bar{\mu}_{..k} = \sum_i \sum_j \pi_i \tau_j \mu_{ijk}\,, \qquad \bar{\mu}_{ij.} = \sum_k \nu_k \mu_{ijk}\,,$$

$$\bar{\mu}_{i.k} = \sum_j \tau_j \mu_{ijk}\,, \qquad \bar{\mu}_{.jk} = \sum_i \pi_i \mu_{ijk}\,, \qquad \mu = \sum_i \sum_j \sum_k \pi_i \tau_j \nu_k \mu_{ijk}$$

$$\alpha_i = \bar{\mu}_{i..} - \mu, \qquad \beta_j = \bar{\mu}_{.j.} - \mu, \qquad \gamma_k = \bar{\mu}_{..k} - \mu$$

$$\delta_{ij} = \bar{\mu}_{ij.} - (\mu + \alpha_i + \beta_j), \qquad \zeta_{ik} = \bar{\mu}_{i.k} - (\mu + \alpha_i + \gamma_k), \qquad \eta_{jk} = \bar{\mu}_{.jk} - (\mu + \beta_j + \gamma_k),$$

$$\theta_{ijk} = \mu_{ijk} - (\mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \zeta_{ik} + \eta_{jk})$$

לפי השערת האפס, מתקיים $\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k$ עבור המשקלות $\{\pi_i\},\{\tau_j\},\{\nu_k\}$. נראה בעת כי זה מתקיים גם עבור המשקלות $\{\pi_i^*\},\{\tau_j^*\},\{\nu_k^*\}$:

$$\mu^* = \sum_i \sum_j \sum_k \pi_i^* \tau_j^* \nu_k^* \mu_{ijk} = \sum_i \sum_j \sum_k \pi_i^* \tau_j^* \nu_k^* (\mu + \alpha_i + \beta_j + \gamma_k) = \mu + \bar{\alpha}_.^* + \bar{\beta}_.^* + \bar{\gamma}_.^*$$

כאשר

$$\bar{\alpha}_.^* = \sum_i \pi_i^* \alpha_i\,, \qquad \bar{\beta}_.^* = \sum_j \tau_j^* \beta_j\,, \qquad \bar{\gamma}_.^* = \sum_k \nu_k^* \gamma_k$$

באופן דומה,

$$\bar{\mu}_{i..}^* = \mu + \alpha_i + \bar{\beta}_.^* + \bar{\gamma}_.^*, \qquad \bar{\mu}_{.j.}^* = \mu + \bar{\alpha}_.^* + \beta_j + \bar{\gamma}_.^*, \qquad \bar{\mu}_{..k}^* = \mu + \bar{\alpha}_.^* + \bar{\beta}_.^* + \gamma_k$$

ומקבלים:

$$\alpha_i^* = \alpha_i - \bar{\alpha}_.^*, \qquad \beta_j^* = \beta_j - \bar{\beta}_.^*, \qquad \gamma_k^* = \gamma_k - \bar{\gamma}_.^*$$

בעת נסמן:

$$\bar{\mu}_{ij.}^* = \sum_k \nu_k^* \mu_{ijk} = \mu + \alpha_i + \beta_j + \bar{\gamma}_.^*$$

$$\bar{\mu}_{i.k}^* = \sum_k \tau_j^* \mu_{ijk} = \mu + \alpha_i + \bar{\beta}_.^* + \gamma_k$$

$$\bar{\mu}_{.jk}^* = \sum_k \pi_i^* \mu_{ijk} = \mu + \bar{\alpha}_.^* + \beta_j + \gamma_k$$

ולכן:

$$\delta_{ij}^* = \bar{\mu}_{ij.}^* - (\mu^* + \alpha_i^* + \beta_j^*) = (\mu + \alpha_i + \beta_j + \bar{\gamma}_.^*) - (\mu + \bar{\alpha}_.^* + \bar{\beta}_.^* + \bar{\gamma}_.^* + \alpha_i - \bar{\alpha}_.^* + \beta_j - \bar{\beta}_.^*)$$
$$= (\mu + \alpha_i + \beta_j + \bar{\gamma}_.^*) - (\mu + \bar{\gamma}_.^* + \alpha_i + \beta_j) = 0$$

$$\zeta_{ik}^* = \bar{\mu}_{i.k}^* - (\mu^* + \alpha_i^* + \gamma_k^*) = (\mu + \alpha_i + \bar{\beta}_.^* + \gamma_k) - (\mu + \bar{\alpha}_.^* + \bar{\beta}_.^* + \bar{\gamma}_.^* + \alpha_i - \bar{\alpha}_.^* + \gamma_k - \bar{\gamma}_.^*)$$
$$= (\mu + \alpha_i + \bar{\beta}_.^* + \gamma_k) - (\mu + \bar{\beta}_.^* + \alpha_i + \gamma_k) = 0$$

$$\eta^*_{jk} = \bar{\mu}^*_{\cdot jk} - (\mu^* + \beta^*_j + \gamma^*_k) = (\mu + \bar{\alpha}^*_{\cdot} + \beta_j + \gamma_k) - (\mu + \bar{\alpha}^*_{\cdot} + \bar{\beta}^*_{\cdot} + \bar{\gamma}^*_{\cdot} + \beta_j - \bar{\beta}^*_{\cdot} + \gamma_k - \bar{\gamma}^*_{\cdot})$$
$$= (\mu + \bar{\alpha}^*_{\cdot} + \beta_j + \gamma_k) - (\mu + \bar{\alpha}^*_{\cdot} + \beta_j + \gamma_k) = 0$$

כלומר גם תחת המשקלות $\{\pi^*_i\},\{\tau^*_j\},\{\nu^*_k\}$ גורמי האינטראקציה הזוגית מתאפסים (ובהתאם גם האינטראקציה המשולשת).

ב. הפונקציה מקבלת data frame וארבעה אינדקסים – המציינים את מספרי העמודות של כ"א מהגורמים וכן את מספר העמודה של המשתנה התלוי. כיוון שהוכחנו בסעיף הקודם כי חוסר תלות אינה קשורה לבחירת משקלות, נשתמש בסט משקלות אחידים (אשר ערכיהם במטריצה X יצטמצמו לבדי ± 1).

קוד R:

```
threeway <- function(data, A_idx, B_idx, C_idx, y_idx){
# יצירת וקטורי הגורמים והמשתנה התלוי
 A <- factor(data[, A_idx])
 B <- factor(data[, B_idx])
 C <- factor(data[, C_idx])
 y <- as.numeric(data[, y_idx])
 #I,J,K,N מציאת הערכים
 I <- nlevels(A)
 J <- nlevels(B)
 K <- nlevels(C)
 N <- nrow(data)
# יצירת תתי מטריצות אשר ירכיבו יחד את המודל המלא
 Xm <- matrix(1, ncol = 1, nrow = N)
 XA <- matrix(0, ncol = I - 1, nrow = N)
 XB <- matrix(0, ncol = J - 1, nrow = N)
 XC <- matrix(0, ncol = K - 1, nrow = N)
 XAB <- matrix(0, ncol = (I - 1) * (J - 1), nrow = N)
 XAC <- matrix(0, ncol = (I - 1) * (K - 1), nrow = N)
 XBC <- matrix(0, ncol = (J - 1) * (K - 1), nrow = N)
 XABC <- matrix(0, ncol = (I - 1) * (J - 1) * (K - 1), nrow = N)
# עבור כל שורה, שומרים את הרמה של כל גורם
 for(n in 1:N){
  i <- as.numeric(A[n])
  j <- as.numeric(B[n])
  k <- as.numeric(C[n])
  #טיפוול גורמים בודדים
  if(i <= I - 1){ #Y_i
   XA[n,i] <- 1
  }
  else{ #Y_I
   XA[n,] <- rep(-1, ncol(XA))
  }
  if(j <= J - 1){ #Y_j
   XB[n,j] <- 1
  }
  else{ #Y_J
   XB[n,] <- rep(-1, ncol(XB))
  }
  if(k <= K - 1){ #Y_k
   XC[n,k] <- 1
  }
  else{ #Y_K
   XC[n,] <- rep(-1, ncol(XC))
  }
  #טיפול באינטראקציות זוגיות
  if(i <= I - 1){
   if(j <= J - 1){ #Y_ij
```

```
        XAB[n, i + j - 1] <- 1
      }
      else{ #Y_iJ
        XAB[n, ] <- rep(-1, ncol(XAB))
      }
      if(k <= K - 1){ #Y_ik
        XAC[n, i + k - 1] <- 1
      }
      else{ #Y_iK
        XAC[n, ] <- rep(-1, ncol(XAC))
      }
    }
    else{
      if(j <= J - 1){ #Y_Ij
        XAB[n, ] <- rep(-1, ncol(XAB))
      }
      else{ #Y_IJ
        XAB[n, ] <- rep(1, ncol(XAB))
      }
      if(k <= K - 1){ #Y_Ik
        XAC[n, ] <- rep(-1, ncol(XAC))
      }
      else{ #Y_IK
        XAC[n, ] <- rep(1, ncol(XAC))
      }
    }
    if(j <= J - 1){
      if(k <= K - 1){ #Y_jk
        XBC[n, j + k - 1] <- 1
      }
      else{ #Y_jK
        XBC[n, ] <- rep(-1, ncol(XBC))
      }
    }
    else{
      if(k <= K - 1){ #Y_Jk
        XBC[n, ] <- rep(-1, ncol(XBC))
      }
      else{ #Y_JK
        XBC[n, ] <- rep(1, ncol(XBC))
      }
    }
    #טיפול באינטראקציה משולשת
    if((i <= I - 1) && (j <= J - 1) && (k <= K - 1)){ #Y_ijk
      XABC[n, i + j + k - 2] <- 1
    }
    else{
# נשים לב כי אם רק אחד מהגורמים בערך מקסימלי אז שמים 1-, אם שני גורמים אז 1 ואם שלושת הגורמים שוב 1- ולכן נוכל לשים פשוט 1- בחזקת מספר הגורמים שערכם מקסימלי
      t <- ((i == I) + (j == J) + (k == K))
      XABC[n,] <- rep((-1)^t, ncol(XABC))
    }
  }
# בניית המטריצות למודל החלקי והמלא
  X0 <- cbind(Xm, XA, XB, XC)
  X <- cbind(X0, XAB, XAC, XBC, XABC)
#חישוב האומדים למודלים
  tau_hat_0 <- solve(t(X0) %*% X0) %*% t(X0) %*% y
  tau_hat <- solve(t(X) %*% X) %*% t(X) %*% y
#חישוב דרגות החופש
```

```
  df <- N - length(tau_hat)
  d <- length(tau_hat) - length(tau_hat_0)
#חישוב וקטורי התחזיות
  y_hat <- X %*% tau_hat
  y_hat_0 <- X0 %*% tau_hat_0
#חישוב וקטורי השגיאות
  e <- y - y_hat
  e_0 <- y_hat - y_hat_0
# חישוב המונה והמכנה של סטטיסטיF
  MSE_0 <- sum(e_0^2) / d
  MSE <- sum(e^2) / df

# חישוב הסטטיסטי ומציאת p-val
  F_stat <- MSE_0 / MSE
  pval <- pf(q = F_stat, df1 = d, df2 = df, lower.tail = F)
  return(pval)
}
```

```
> D <- read.csv("popcorn.csv")
> threeway(data = D, A_idx = 1, B_idx = 2, C_idx = 3, y_idx = 4)
[1] 2.12123e-15
```

כלומר נדחה את השערת חוסר האינטראקציה בנתוני הקובץ popcorn.

2. במודל לוג-ליניארי עם 4 משתנים A, B, C, D, בחנו את המודל המכיל את גורמי λ הבאים: A, B, C, AB, AC, AD, BC, BD, CD, ABC,BCD. הגדירו

$$\psi_{jk}^{AD}(i,i',l,l') = \frac{\pi_{ijkl}\pi_{i'jkl'}}{\pi_{ijkl'}\pi_{i'jkl}}$$

הוכיחו כי עבור i,i',l,l' נתונים, $\psi_{jk}^{AD}(i,i',l,l')$ אינו תלוי ב-j או k.

$$\theta_{ijkl} = \bar{\theta}_{....} + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{il}^{AD} + \lambda_{jk}^{BC} + \lambda_{jl}^{BD} + \lambda_{kl}^{CD} + \lambda_{ijk}^{ABC} + \lambda_{jkl}^{BCD}, \qquad \pi_{ijkl} = e^{\theta_{ijkl}}$$

$$\psi_{jk}^{AD}(i,i',l,l') = \frac{\pi_{ijkl}\pi_{i'jkl'}}{\pi_{ijkl'}\pi_{i'jkl}} = \frac{e^{\theta_{ijkl}}e^{\theta_{i'jkl'}}}{e^{\theta_{ijkl'}}e^{\theta_{i'jkl}}} = \frac{e^{\theta_{ijkl}+\theta_{i'jkl'}}}{e^{\theta_{ijkl'}+\theta_{i'jkl}}} = e^{\theta_{ijkl}+\theta_{i'jkl'}-\theta_{ijkl'}-\theta_{i'jkl}}$$

$$\begin{aligned}
\theta_{ijkl} + \theta_{i'jkl'} - \theta_{ijkl'} - \theta_{i'jkl} \\
&= \left(\bar{\theta}_{....} + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{il}^{AD} + \lambda_{jk}^{BC} + \lambda_{jl}^{BD} + \lambda_{kl}^{CD} + \lambda_{ijk}^{ABC} + \lambda_{jkl}^{BCD}\right) \\
&+ \left(\bar{\theta}_{....} + \lambda_{i'}^A + \lambda_j^B + \lambda_k^C + \lambda_{l'}^D + \lambda_{i'j}^{AB} + \lambda_{i'k}^{AC} + \lambda_{i'l'}^{AD} + \lambda_{jk}^{BC} + \lambda_{jl'}^{BD} + \lambda_{kl'}^{CD} + \lambda_{i'jk}^{ABC} + \lambda_{jkl'}^{BCD}\right) \\
&- \left(\bar{\theta}_{....} + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{l'}^D + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{il'}^{AD} + \lambda_{jk}^{BC} + \lambda_{jl'}^{BD} + \lambda_{kl'}^{CD} + \lambda_{ijk}^{ABC} + \lambda_{jkl'}^{BCD}\right) \\
&- \left(\bar{\theta}_{....} + \lambda_{i'}^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{i'j}^{AB} + \lambda_{i'k}^{AC} + \lambda_{i'l}^{AD} + \lambda_{jk}^{BC} + \lambda_{jl}^{BD} + \lambda_{kl}^{CD} + \lambda_{i'jk}^{ABC} + \lambda_{jkl}^{BCD}\right)
\end{aligned}$$

נשים לב כי האברים הבאים מצטמצמים:

$$\bar{\theta}_{....}, \lambda_i^A, \lambda_{i'}^A, \lambda_j^B, \lambda_k^C, \lambda_l^D, \lambda_{l'}^D \lambda_{ij}^{AB}, \lambda_{i'j}^{AB}, \lambda_{ik}^{AC}, \lambda_{i'k}^{AC}, \lambda_{jk}^{BC}, \lambda_{jl}^{BD}, \lambda_{jl'}^{BD}, \lambda_{kl}^{CD}, \lambda_{kl'}^{CD}, \lambda_{ijk}^{ABC}, \lambda_{i'jk}^{ABC}, \lambda_{jkl}^{BCD}, \lambda_{jkl'}^{BCD}$$

ונקבל:

$$\theta_{ijkl} + \theta_{i'jkl'} - \theta_{ijkl'} - \theta_{i'jkl} = \left(\lambda_{il}^{AD} + \lambda_{i'l'}^{AD} - \lambda_{il'}^{AD} - \lambda_{i'l}^{AD}\right)$$

כלומר מתקיים:

$$\psi_{jk}^{AD}(i,i',l,l') = e^{\theta_{ijkl}+\theta_{i'jkl'}-\theta_{ijkl'}-\theta_{i'jkl}} = e^{\lambda_{il}^{AD}+\lambda_{i'l'}^{AD}-\lambda_{il'}^{AD}-\lambda_{i'l}^{AD}}$$

וזה ביטוי שאינו תלוי ב-j,k, כנדרש.

3. בקובץ titanic.csv נמצאים פרטי הנוסעים על ספינת הטיטאניק, כפי שמוסבר בקובץ titanic.txt. משתנה התגובה הוא Survived ויתר המשתנים הם המנבאים.

א. קראו את הנתונים ב-R והסירו עמודות בהן לא תשתמשו. הסבירו את בחירתכם. לאחר הסרת העמודות הללו, הסירו שורות עם ערכים חסרים.

<u>אזהרה</u>: בחיים האמיתיים, מחיקת רשומות עם מידע חסר אינה גישה מקובלת וישנן שיטות להתמודדות עם מצבים כאלו, אולם הן אינן חלק מהחומר בקורס זה.

ב. בנו מודל רגרסיה לוגיסטית מתאים. מהם המשתנים בעלי רמת המובהקות הגבוהה יותר?

ג. הריצו את כלי בדיקת המודל בהם השתמשנו בקורס ונסו לשפר את המודל ע"י הוספת טרנספורמציה על המשתנה age או גורם לא-לינארי של המשתנה age. הוספה של גורם\טרנספורמציה אחת תספיק, זכרו לכלול את כל הגרפים הרלוונטיים.

ד. הוסיפו עמודה חדשה לנתונים בשם child, משתנה בינארי המציין האם הנוסע צעיר מ-15 שנים או לא. הסירו את העמודה age ואת העמודות שהוספתם בסעיף ג'.

ה. בנו מודל רגרסיה לוגיסטית תחת אוסף הנתונים המעודכן. האם ניכר שינוי (בהקשר של מובהקות משתנים)?

כאשר דנים ברגרסיה לינארית ובANOVA, שיטה נפוצה להשוואה בין מודלים היא מבחן F. במקרה של רגרסיה לוגיסטית, איננו יכולים להשתמש בה ביוון שמתקיים $y_i - \hat{y}_i = e_i \in \{0,1\}$. במקומו, נשתמש בכלי הקרוי Analysis of Deviance (ניתוח סטיה, AoD), שהוא למעשה מבחן יחס נראות. פונקצית R הרלוונטית משווה בין שני מודלים באמצעות מבחן נתון, עליכם לקרוא לה באופן

$$\boldsymbol{anova}(reduced, full, test = "Chisq")$$

כאשר reduced ו-full הם שני אובייקטי glm המייצגים את שני המודלים. סטטיסטי המבחן הוא ההפרש בין פונקציות הנראות של המודלים (מוכפל ב-2) וההתפלגות המקורבת שלו, לפי משפט Wilks, היא χ^2 עם מספר דרגות חופש השווה להפרש במספר הפרמטרים בין שני המודלים.

ו. נרצה לבחון את ההשערה הפופולרית 'נשים וילדים קודם'. נסחו את H_0, H_1 בהתאם.

ז. בנו שני מודלי רגרסיה לוגיסטית אשר ייצגו את השערותיכם ובחנו את ה-p-Value של בדיקת ההשערות הנ"ל באמצעות הפונקציה לביצוע AoD.

3

a + b

Passenger ID and name mean nothing for the prediction, as well as the ticket and cabin codes (which is a sparse vector). We shall also remove rows with missing data.

```
D <- read.csv("titanic.csv")
D <- D[,-c(1,4,9,11)]
D <- D[-which(is.na(D$Age) | D$Embarked == ""),]
l <- glm(data = D, Survived ~ . , family = binomial(link = "logit"))
summary(l)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = D)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7233  -0.6447  -0.3799   0.6326   2.4457
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.637407   0.634550   8.884  < 2e-16 ***
## Pclass      -1.199251   0.164619  -7.285 3.22e-13 ***
## Sexmale     -2.638476   0.222256 -11.871  < 2e-16 ***
## Age         -0.043350   0.008232  -5.266 1.39e-07 ***
## SibSp       -0.363208   0.129017  -2.815  0.00487 **
## Parch       -0.060270   0.123900  -0.486  0.62666
## Fare         0.001432   0.002531   0.566  0.57165
## EmbarkedQ   -0.823545   0.600229  -1.372  0.17005
## EmbarkedS   -0.401213   0.270283  -1.484  0.13770
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 632.34  on 703  degrees of freedom
## AIC: 650.34
##
## Number of Fisher Scoring iterations: 5
```

We can see that the more significant predictors are a passenger's class, sex and age. The number of siblings / spouses aboard ($SibSp$) is of lesser significance level.

c

The file "logcheck.R" contains the r function specified in "logcheck1.pdf", with modifications to handle empty bins and print the plots instead of creating pdf files.

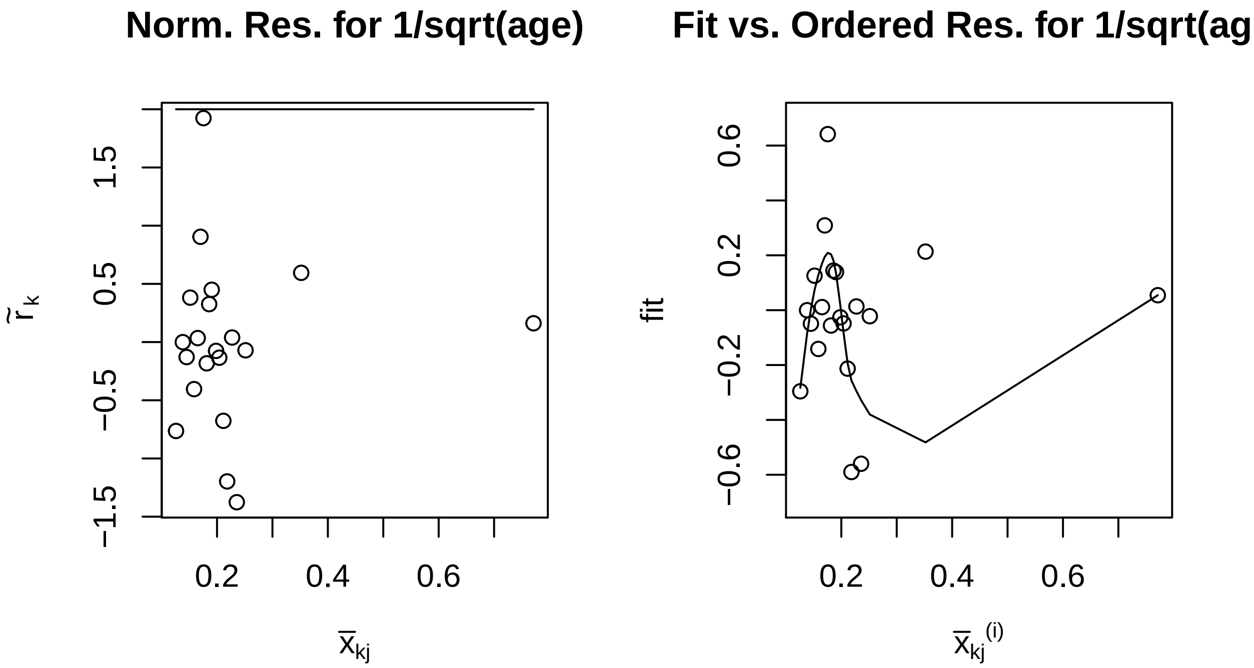
```
source("logcheck.R")
par(mfrow=c(1,2))
resid_anal(D$Age, D$Survived, l$fitted, 20, "Age")
```

We can see an apparent outlier. Adding $\frac{1}{\sqrt{age}}$ solves this:

```
par(mfrow=c(1,2))
D$k <- (D$Age)^-0.5
l2 <- glm(data = D, Survived ~ . , family = binomial(link = "logit"))
summary(l2)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = D)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0586  -0.6428  -0.3867   0.5991   2.3634
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.945119   0.767011   5.143  2.7e-07 ***
## Pclass      -1.108503   0.167237  -6.628  3.4e-11 ***
## Sexmale     -2.755470   0.228354 -12.067  < 2e-16 ***
## Age         -0.018152   0.010442  -1.738 0.082160 .
## SibSp       -0.452604   0.136767  -3.309 0.000935 ***
## Parch       -0.165629   0.130908  -1.265 0.205788
## Fare         0.002594   0.002672   0.971 0.331538
## EmbarkedQ   -0.746019   0.602144  -1.239 0.215369
## EmbarkedS   -0.347314   0.277438  -1.252 0.210620
## k            3.685336   1.119622   3.292 0.000996 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 617.51  on 702  degrees of freedom
## AIC: 637.51
##
## Number of Fisher Scoring iterations: 5
```

```
resid_anal(D$k, D$Survived, l2$fitted, 20, "1/sqrt(age)")
```



d + e

```r
D$child <- 0 + (D$Age < 15)
D <- D[,-c(4,9)]
l2 <- glm(data = D, Survived ~ . , family = binomial(link = "logit"))
summary(l2)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = D)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8293  -0.6952  -0.4498   0.6528   2.3601
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.603481   0.470370   7.661 1.85e-14 ***
## Pclass      -0.931146   0.149618  -6.223 4.86e-10 ***
## Sexmale     -2.724419   0.223967 -12.164  < 2e-16 ***
## SibSp       -0.483550   0.140726  -3.436  0.00059 ***
## Parch       -0.197221   0.127801  -1.543  0.12279
## Fare         0.003529   0.002736   1.290  0.19716
## EmbarkedQ   -0.665405   0.575758  -1.156  0.24780
## EmbarkedS   -0.350747   0.271973  -1.290  0.19718
## child        2.005549   0.398121   5.038 4.72e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 634.12  on 703  degrees of freedom
## AIC: 652.12
##
## Number of Fisher Scoring iterations: 5
```

Changing the *Age* predictor into a Boolean one resulted in the same significance level for itself and a higher significance level for *SibSp*.

f + g

For the hypothesis 'Women and children first', the proper hypotheses are $H_0 : \hat{\beta}_{sex} = \hat{\beta}_{child} = 0; \quad H_1 : (\hat{\beta}_{sex} \neq 0) \vee (\hat{\beta}_{child} \neq 0)$.

```
reduced <- glm(data = D, Survived ~ Pclass + SibSp + Parch +
               Fare + Embarked, family = binomial(link = "logit"))
full <- glm(data = D, Survived ~ Pclass + SibSp + Parch +
               Fare + Embarked + Sex + child, family = binomial(link = "logit"))
anova(reduced, full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Pclass + SibSp + Parch + Fare + Embarked
## Model 2: Survived ~ Pclass + SibSp + Parch + Fare + Embarked + Sex + child
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       705     850.22
## 2       703     634.12  2    216.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value of the AoD result is $< 2.2 \cdot 10^{-16}$, we'll reject the null hypothesis for practically any given significance level.