Categorical Data Analyses:

Logistic Regression & Loglinear Analysis

Overview

- **Logistic regression**
  - Binary and multinomial (Ch. 8)
  - Examples:
    - Predicting life satisfaction and lifestyle

- **Categorical analysis**
  - Chi-square & Loglinear analyses (Ch. 18)
  - Examples:
    - Sex, job satisfaction, and lifestyle (no DV)

Logistic regression

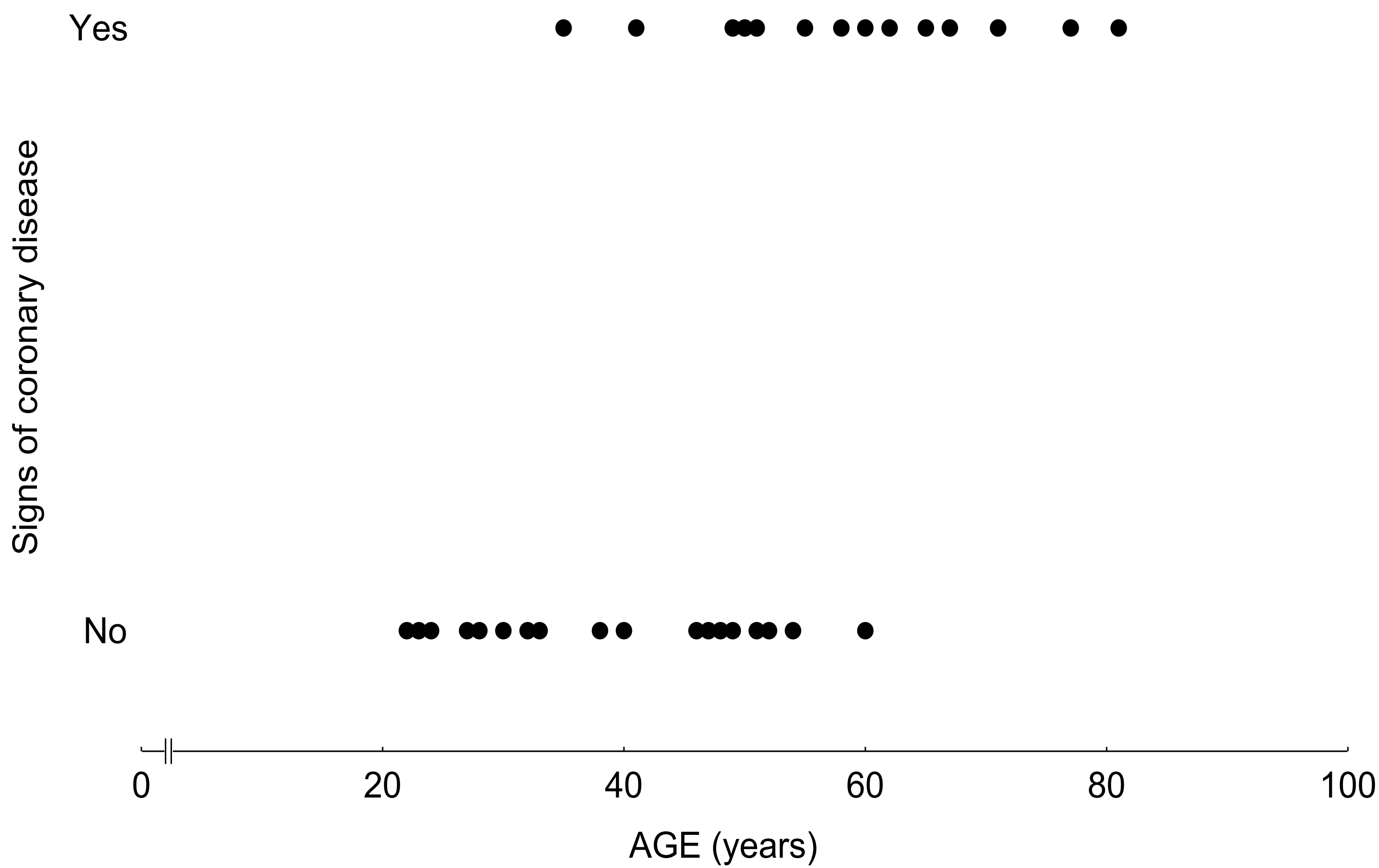
Binary and multinomial

- **When & why to use logistic regression**
  - Categorical DVs do not tend to have a linear relationship with predictors
- **How to assess**
  - …the model (LLR and deviance)
  - …the predictors (OR and CIs)
- **Assumptions and trouble-shooting**
  - Assumptions should be checked, and problems should be dealt with

Nonlinear association between age and CD

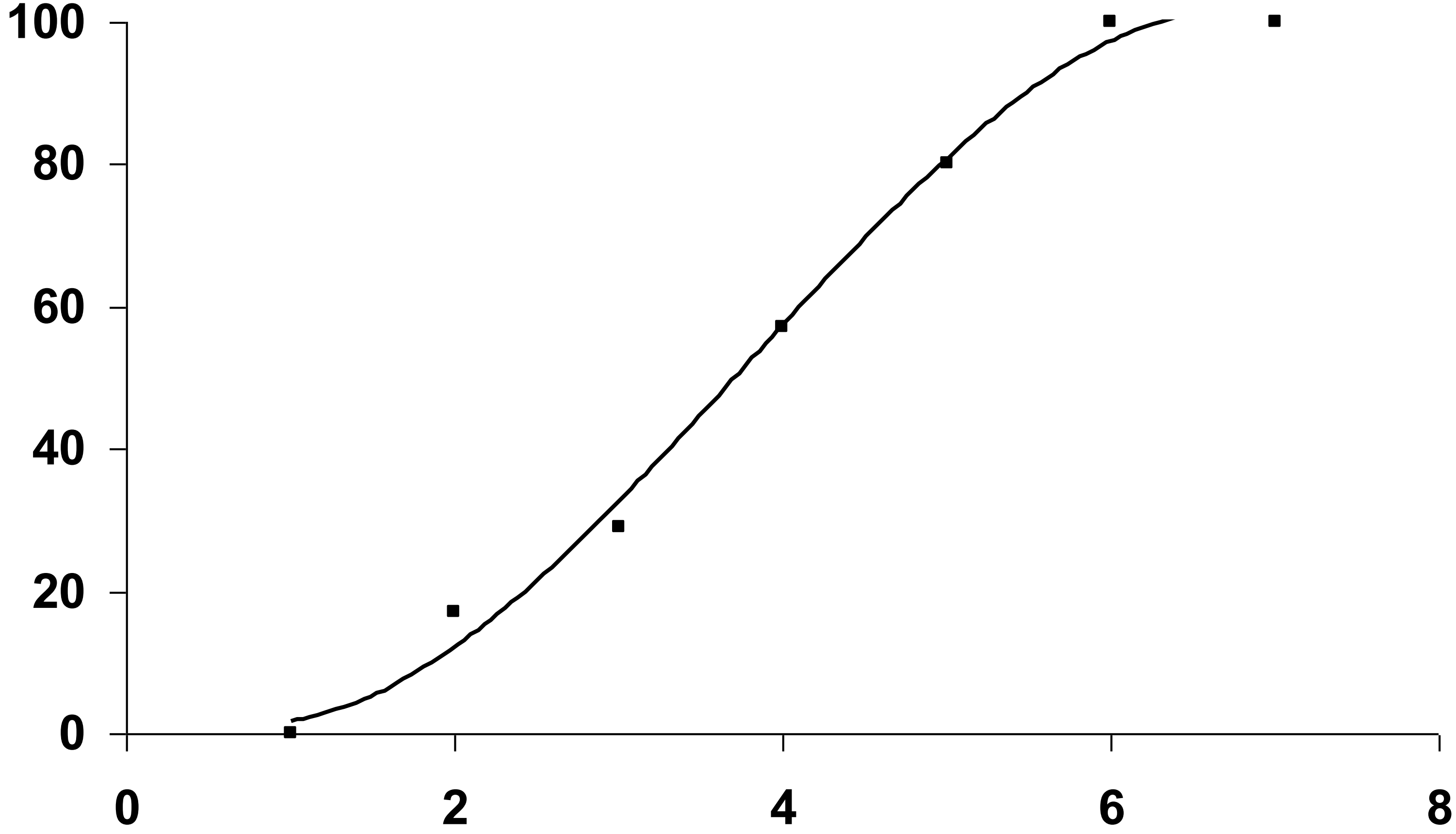
Toon's Show

Table 2 Age and signs of coronary heart disease (CD)



Nonlinear association between age and CD
Toon's Show

Table 3 Age and signs of coronary heart disease (CD)



Equation with One Predictor

Toon and Andy Agree?

Toon: $$P(y|x) = \frac{e^{a+bx}}{1+e^{a+bx}}$$

Andy: $$\boxed{P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+\varepsilon_i)}}}$$

- **Outcome**
  - Predict the *probability* of the outcome occurring
- ***a and bx (b_0 and b_1)***
  - Intercept of entire model; and slope (gradient) associated with individual predictors

Equation with Multiple Predictors

Toon and Andy Agree?

Toon: $$\ln\left(\frac{P}{1-P}\right) = a + b_1x_1 + b_2x_2 + \ldots b_ix_i$$

Andy: $$\boxed{P(Y) = \frac{1}{1+e^{-(b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon_i)}}}$$

- **Outcome**
  - We still predict the *probability* of the outcome occurring
- **Differences**
  - This part of the equation expands to accommodate additional predictors

Assessing a Model

The log-likelihood statistic

Toon: $$L(\mathrm{B}) = \ln\left[l(\mathrm{B})\right] = \sum_{i=1}^{n}\left\{y_i \ln\left[\pi(x_i)\right] + (1-y_i)\ln\left[1-\pi(x_i)\right]\right\}$$

Andy: $$\log-\text{likelihood} = \sum_{\mathrm{i}=1}^{\mathrm{N}}\left[Y_i \ln\left(P(Y_i)\right) + \left(1-Y_i\right)\ln\left(1-P(Y_i)\right)\right]$$

- Analogous to the residual sum of squares
- It is an **indicator of how much unexplained information** there is after the model has been fitted.
- Large values indicate poorly fitting statistical models, but values are not comparable.
  - **Chi-square, AIC, BIC**

Assessing Changes in Models

Deviance (a test of parsimony)

Toon:

Andy:

$$\chi^2 = 2\left[LL(\text{new}) - LL(\text{baseline})\right]$$
$$\left(df = k_{\text{new}} - k_{\text{baseline}}\right)$$

- It is possible to compare the difference (deviance) between the log-likelihoods of nested models.
  - null (intercepts only), main effects, higher-order interactions, saturated, etc.

Assessing Predictors

The Wald Statistic

Toon:

$$\text{Wald} = \left(\frac{b}{SE_b}\right)^2 \qquad (df = 1)$$

Andy:

$$\boxed{Wald = \frac{b}{SE_b}}$$

- Toon's version tested using chi-square; Andy's version tested using z-distribution
- Tests the null hypothesis that *b* = 0.
- **Wald statistic is biased when *b* is large.**
  - Better to look at odds ratios.

Assessing Predictors:
Odds Ratio

Toon:

Andy:

$$\text{odds ratio} = \frac{\text{odds after a unit change in the predictor}}{\text{odds before a unit change in the predictor}}$$

- **Indicates the change in odds resulting from a unit change in the predictor.**
  - OR > 1: Predictor ↑, Probability of outcome occurring ↑.
  - OR < 1: Predictor ↑, Probability of outcome occurring ↓.

Assumptions

Logistic regression

- **Linearity (in the logit)**
  - There should be a linear relationship between any <u>continuous predictor</u> and the logit of the outcome
    - This can be tested by examining the interaction between the continuous predictor and its log transformation
- **Independence of errors (observations)**
  - Durbin-Watson
- **Absence of multicollinearity, multivariate outliers and influential cases**
  - VIF, standardized residuals, DFBeta, leverage

R U ready?

Does lifestyle predict job satisfaction similarly for males and females?

- Job satisfaction (high/low), lifestyle (dull, routine, and exciting), and sex (female/male)…sound familiar?
  - 1500 participants, lots of missing values.

Set, Import, and Inspect

setwd("C:/…/Analyzing in R_Stats II_2015")

logex <-
read.spss("logreg.sav“,to.data.frame=T)

head(logex)

```
  id age sex satjob educ hours life income
1  1  43   1      2   11    35    2     17
2  2  44   1      2   16    21    3     18
3  3  43   2      1   16    45    3     18
4  4  45   2      2   15    20    0     22
5  5  78   2     NA   17    -1    3      0
6  6  83   1     NA   11    -1    2      0
```

Set, Import, and Inspect

…with value labels

```
logex <-
  read.spss("logreg.sav“,to.data.frame=T,
  use.value.labels = T)

head(logex)
```

```
> head(logex)
  id age    sex             satjob educ hours     life       income
1  1  43   Male Not very satisfied   11    35  Routine $35000-39999
2  2  44   Male Not very satisfied   16    21 Exciting $40000-49999
3  3  43 Female     Very satisfied   16    45 Exciting $40000-49999
4  4  45 Female Not very satisfied   15    20     <NA>         <NA>
5  5  78 Female               <NA>   17    NA Exciting         <NA>
6  6  83   Male               <NA>   11    NA  Routine         <NA>
```

VIM

Visualization and Imputation of Missing values

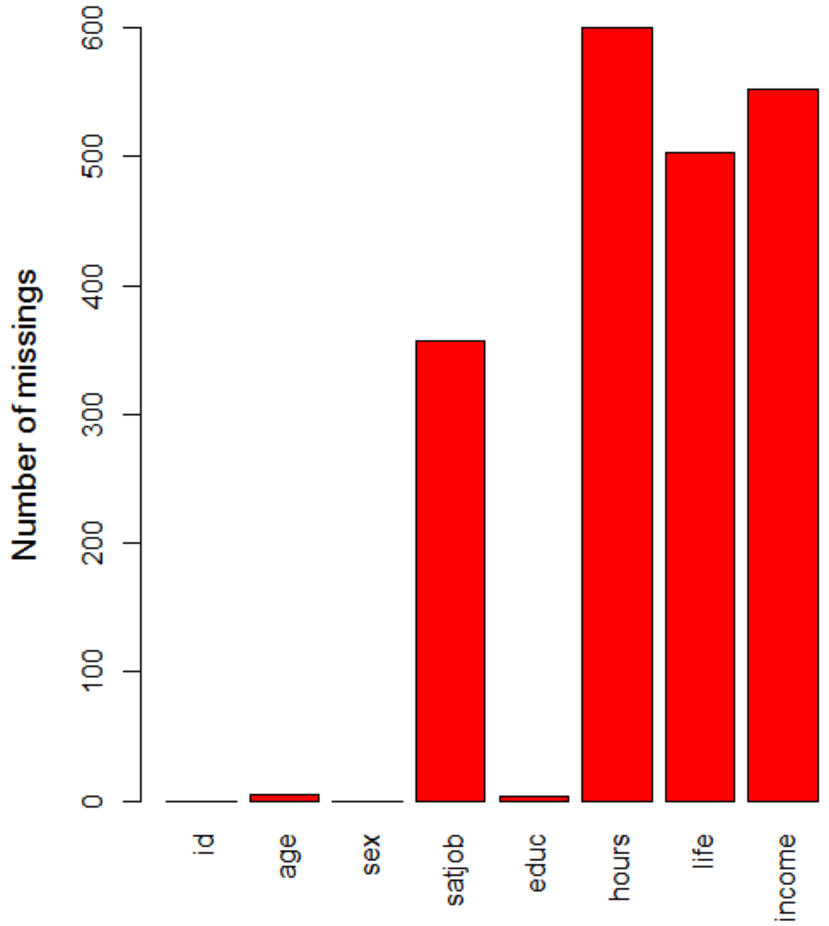
- **Install and load**

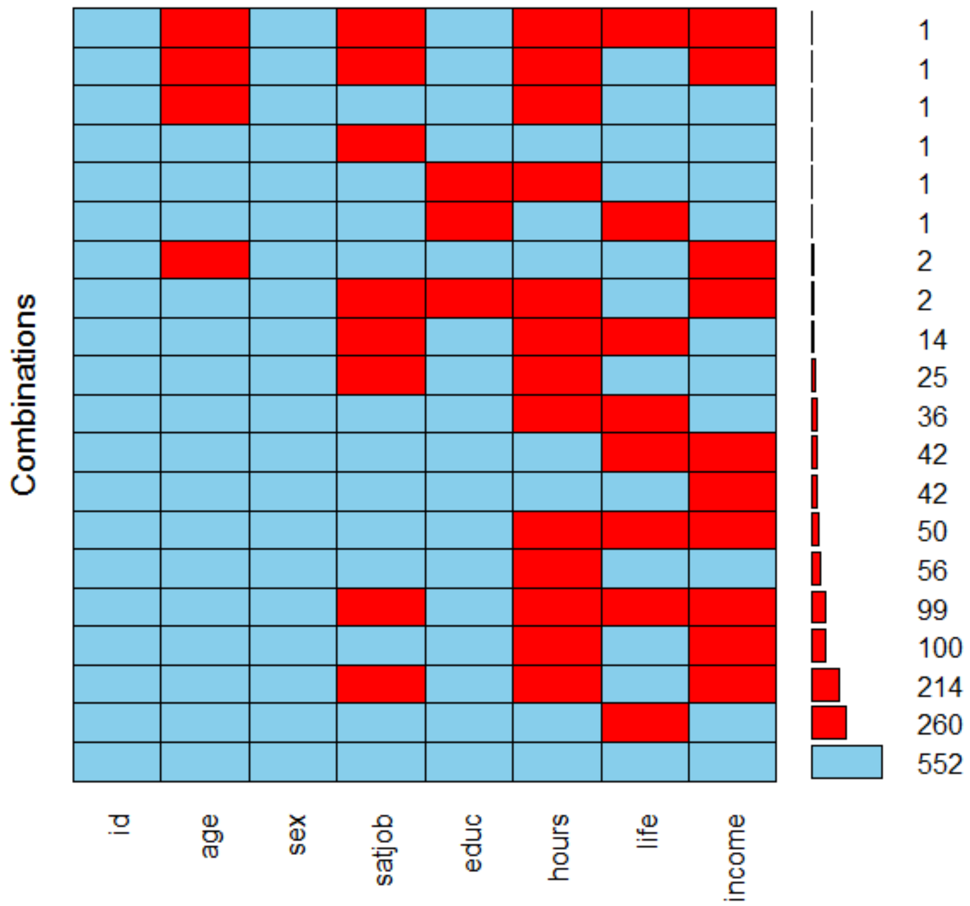
install.packages(“VIM")

library(VIM)

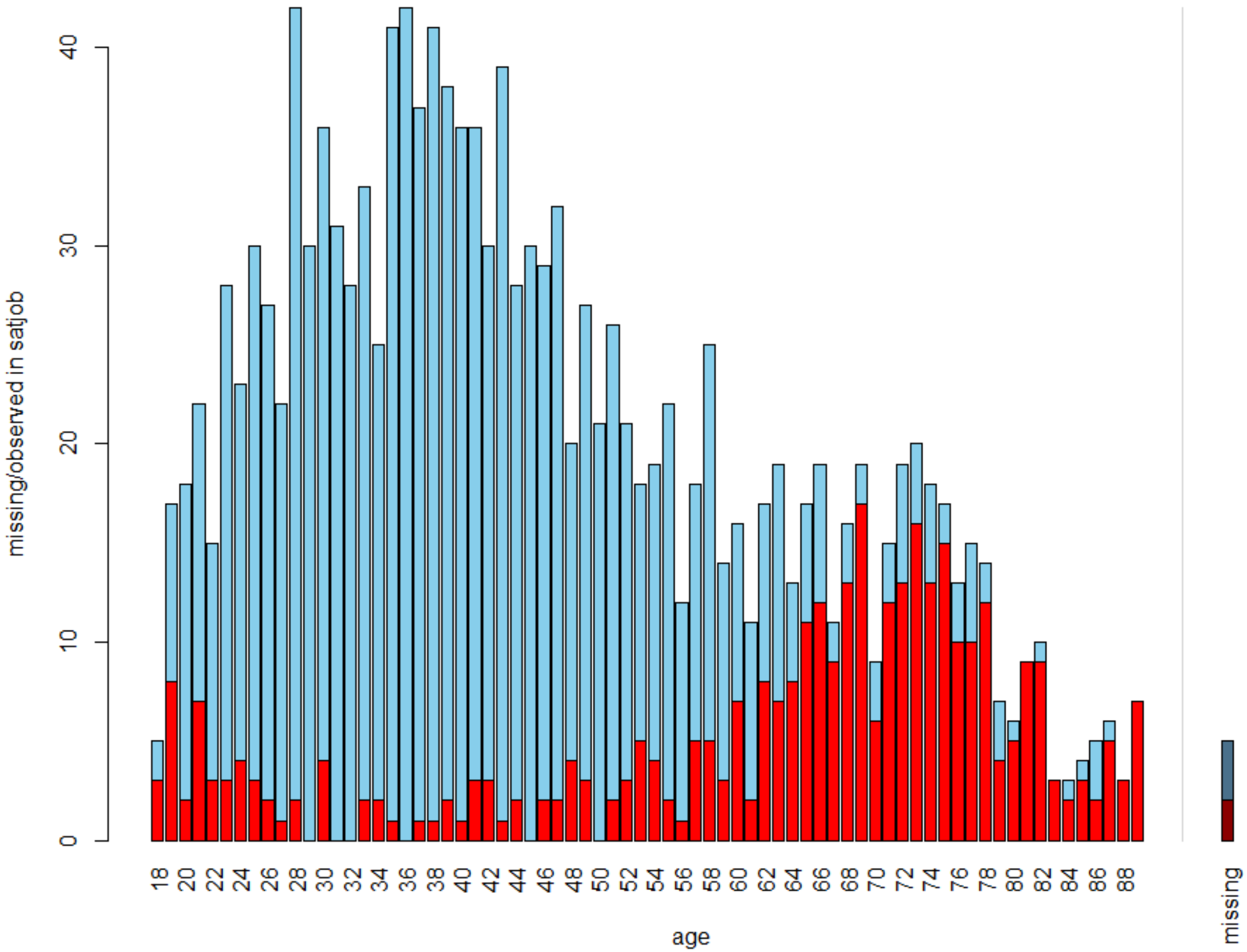
- **Import data**
  - Ensure that categorical measures are listed as factors

miss<- aggr(logex, numbers = T, prop = F) miss





miss2<- logex[,c(“age”,”satjob”)]
histMiss(miss2)



R U really ready?

Does lifestyle predict job satisfaction similarly for males and females?

- Job satisfaction (high/low), lifestyle (dull, routine, and exciting), and sex (female/male)…sound familiar?
  - 1500 participants, lots of missing values.
- I limit the analyses to the 754 participants with all three measures.
  - **Alternative strategies to dealing with missing values will be addressed in the SEM course.**

Extract
subset()

- **Select valid cases and relevant variables**

logex1sub<-subset(logex1, satjob != FALSE & life != FALSE, select=c("id", "sex", "satjob", "life"))

- **Check**

head(logex1sub)

nrow(logex1sub)

```
> head(logex1sub)
   id    sex               satjob     life
1   1   Male Not very satisfied  Routine
2   2   Male Not very satisfied Exciting
3   3 Female     Very satisfied Exciting
7   7 Female     Very satisfied  Routine
9   9   Male     Very satisfied Exciting
10 10 Female     Very satisfied Exciting
> nrow(logex1sub)
[1] 754
```

Extract

complete.cases & na.omit

- **Select relevant variables**

```
logex1 <- logex[,c(1,3,4,7)]
```

- **Use complete.cases()**

```
logex1sub2 <- logex1[complete.cases(logex1),]
```

- **Use na.omit()**

```
logex1sub3 <- na.omit(logex1)
```

Contrast settings?

More on this next week...

- **Default is dummy coding**
- What are the reference groups?

```
> contrasts(logex$sex)
       Female
Male        0
Female      1

> contrasts(logex$satjob)
                   Not very satisfied
Very satisfied                      0
Not very satisfied                  1
> contrasts(logex$life)

         Routine Exciting
Dull           0        0
Routine        1        0
Exciting       0        1
```

Perform logistic regression

Sex and lifestyle as predictors of job satisfaction

logmodel1 <- glm(satjob ~ sex + life, data = logexsub, family = binomial())

- **Examine the model diagnostics**
  - **Not the output (yet)!**
- **Note: The default of glm() is listwise deletion, so I could have also simply used the original data file, but…**

Diagnostic tests

in the car () package

- **Independence of errors (Durbin-Watson)**

```
dwt(logmodel1)
 lag Autocorrelation D-W Statistic p-value
   1     -0.01427466      2.027084   0.742
```

- **Multicollinearity (VIF)**

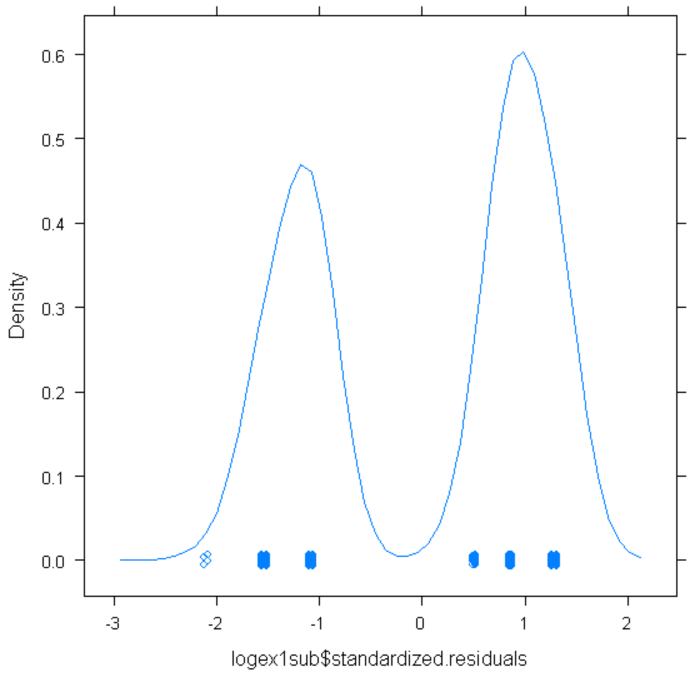
```
vif(logmodel1)
          GVIF    Df    GVIF^(1/(2*Df))
sex     1.003113  1         1.001555
life    1.003113  2         1.000777
```

Model Diagnostics

pp 338-341 in FMF

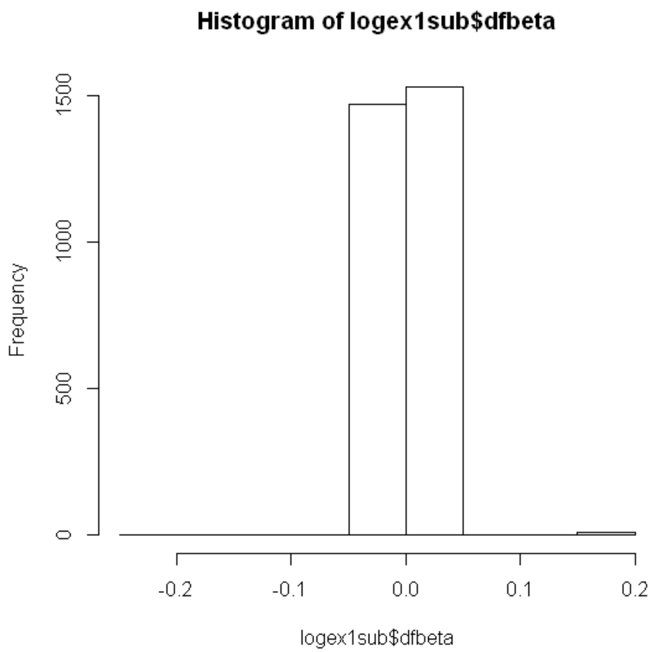
- **Examine standardized residuals (outliers)**
  - 5% of cases should have residuals > 2, 1% should have residuals > 2.5, and none > 3.
- **Examine DFBetas (influential cases)**
  - effect on coefficients after deleting observation
  - No values should be above 1
- **Calculate leverage (influential cases)**
  - (# predictors + 1) / sample size
  - Are any values larger than 2x or 3x this value?

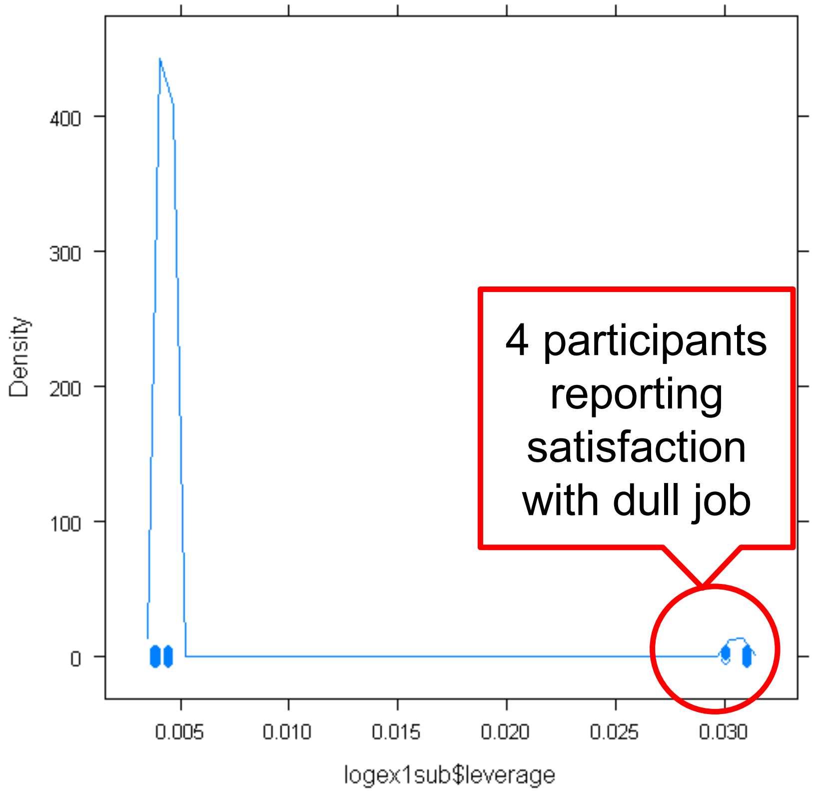
Diagnostic plots



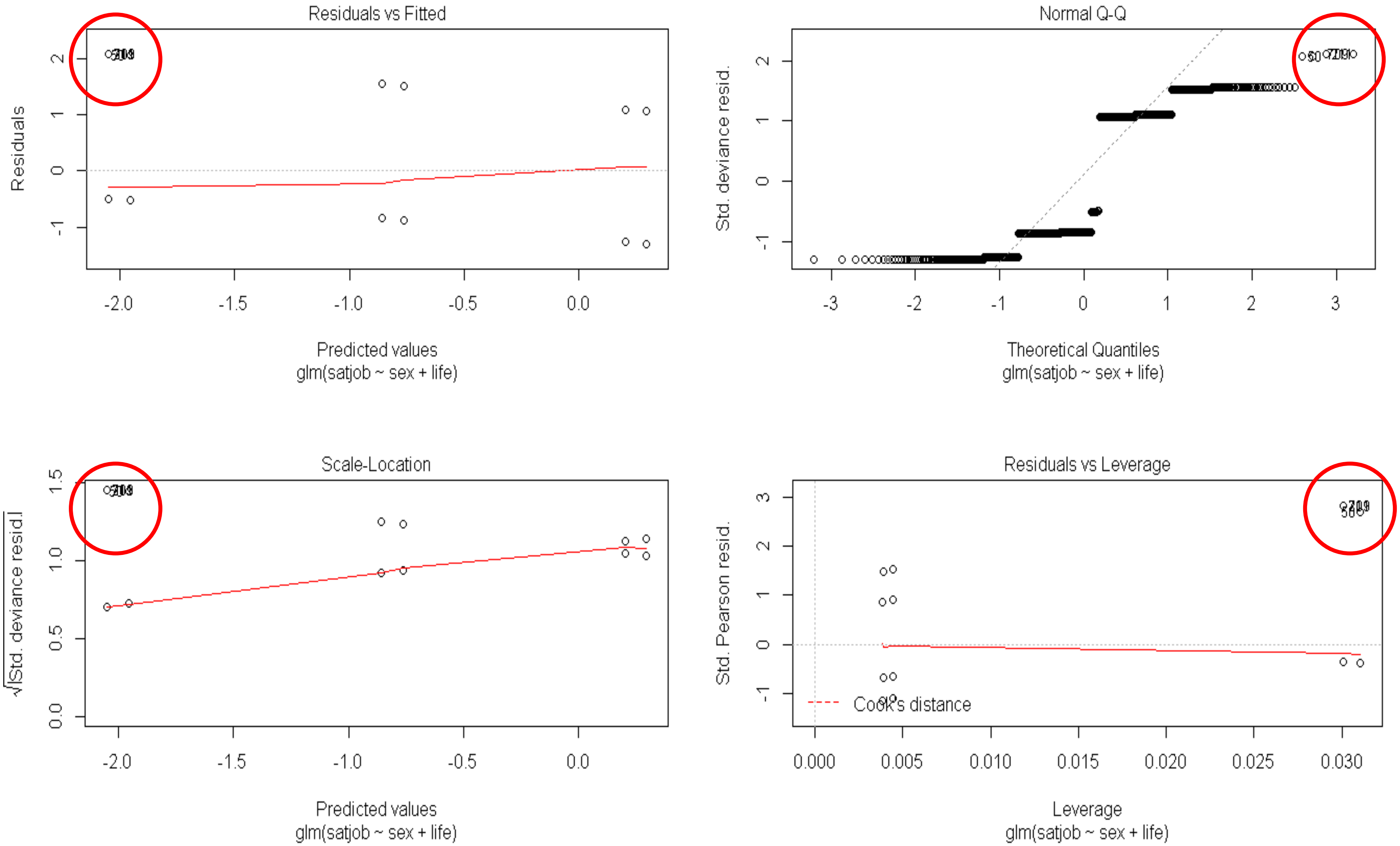
Leverage -> 3/754

[1] 0.00397878





plot(logmodel1)



summary(logmodel1)

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)     2.05027    0.54564   3.758 0.000172 ***
sexFemale      -0.09433    0.15495  -0.609 0.542670    
lifeRoutine    -1.19360    0.54662  -2.184 0.028993 *  
lifeExciting   -2.25033    0.54430  -4.134 3.56e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1029.16  on 753  degrees of freedom
Residual deviance:  965.96  on 750  degrees of freedom
AIC: 973.96
```

Where is R-square?

Assess model fit: χ^2 & R^2

```
#Computing chi-square
logChi <- logmodel1$null.deviance-logmodel1$deviance
chidf <- logmodel1$df.null - logmodel1$df.residual
chisq.prob <- 1 - pchisq(logChi, chidf)
logChi; chidf; chisq.prob
[1] 63.20087
[1] 3
[1] 1.216804e-13

#Compute R-square (WRITE A FUNCTION, p. 334, FMF!)
R2.hl<-logChi/logmodel1$null.deviance
R.cs<-1-exp(
(logmodel1$deviance-logmodel1$null.deviance)/754)
R.n<-R.cs/(1-(exp(-(logmodel1$null.deviance/754))))
R2.hl; R.cs; R.n
[1] 0.0614101
[1] 0.08040395
[1] 0.1079824
```

Assess predictors: ORs & CIs

```
#Compute odds ratios and 95% confidence intervals
exp(logmodel1$coefficients)
(Intercept)    sexfemale  liferoutine lifeexciting
  7.7700061    0.9099800    0.3031289    0.1053647
```

1/.303 = 3.30   1/.105 = 9.52

```
exp(confint(logmodel1))
                  2.5 %     97.5 %
(Intercept)  2.97096398 26.6791248
sexfemale    0.67117679  1.2325086
liferoutine  0.08818527  0.7952406
lifeexciting 0.03075298  0.2747995
```

Do individuals with routine & exciting lifestyles differ on job satisfaction?

- ***relevel***

logex$life <- relevel(logex$life, ref = 2)

- **Perform analysis w/ alternative reference group**

logmodel1alt <- glm(satjob ~ sex + life, data = logex, family = binomial())

- **The reference group is now routine lifestyle**

summary (logmodel1alt)

summary(logmodel1alt)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)   0.85667    0.14577   5.877 4.18e-09 ***
sexfemale    -0.09433    0.15495  -0.609    0.543    
lifedull      1.19360    0.54662   2.184    0.029 *  
lifeexciting -1.05673    0.15616  -6.767 1.32e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1029.16  on 753  degrees of freedom
Residual deviance:  965.96  on 750  degrees of freedom

AIC: 973.96
```

Perform hierarchical regression

- **<u>Main effects model (previous)</u>**

```
logmodel1 <- glm(satjob ~ sex + life, data =
   logex, family = binomial())
```

- **<u>Interactional model</u>**

```
logmodel2 <- glm(satjob ~ sex * life, data =
   logex, family = binomial())
```

```
summary(logmodel2)
```

summary(logmodel2)

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 0.9072     0.1786   5.079 3.79e-07 ***
sexfemale                  -0.1828     0.2355  -0.776    0.438    
lifedull                    0.3456     0.8214   0.421    0.674    
lifeexciting               -1.1223     0.2366  -4.743 2.11e-06 ***
sexfemale:lifedull          1.3279     1.1150   1.191    0.234    
sexfemale:lifeexciting      0.1158     0.3151   0.368    0.713    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1029.16  on 753  degrees of freedom
Residual deviance:  964.55  on 748  degrees of freedom
  AIC: 976.55
```

Model Comparison

Parsimony rules!

```
#Compare two models
anova(logmodel1,logmodel2)

Analysis of Deviance Table

Model 1: satjob ~ sex + life
Model 2: satjob ~ sex * life
  Resid. Df Resid. Dev Df Deviance
1       750     965.96
2       748     964.55  2   1.4092
```

```
> logChi1; chidf1; chisq.prob1
[1] 1.409211
[1] 2
[1] 0.4943036
```

ns, interactions did not significantly contribute

Multinomial logistic regression

Multinomial logistic regression

A simple extension?

- Multinomial logistic regression can be used when the DV has 3 or more **unordered** categories.
- Has same assumptions as binary logistic regression, plus…
  - IIA (Independence of Irrelevant Alternatives)
    - The odds of choosing A over B should not depend on whether some other alternative C is present or absent.
    - Hausman-McFadden test (1984) Small-Hsiao test (1985)

Multinomial logistic regression

Restructure and perform

- Several packages can be used to perform multinomial logistic regression models (multinom, mlogit, mnlogit, vgam)
- mlogit() is most flexible, but requires data to be in “its own” long format.
  - mlogit.data creates new data frame referencing the outcome variable

newDataFrame<-mlogit.data(oldDataFrame, choice = “outcome”, shape = “wide”)

Install and Load

install.packages("mlogit")

necessary for multinomial logistic regression

install.packages("car")

“Companion for Applied Regression”

library(car)

library(mlogit)

Long format

```
mlog1 <-mlogit.data(logex, choice
="life", shape ="wide")

head(mlog1)
    id sex satjob  life chid alt
1.1  1   1      0 FALSE    1   1
1.2  1   1      0  TRUE    1   2
1.3  1   1      0 FALSE    1   3
2.1  2   1      0 FALSE    2   1
2.2  2   1      0 FALSE    2   2
2.3  2   1      0  TRUE    2   3
```

Perform multinomial logistic regression

Sex and job satisfaction as predictors of lifestyle

mlogmodel1<-mlogit(life ~ 1 | sex + satjob, data = mlog1, reflevel = 1)

- **Diagnostics not straightforward☹**
  - Best option? Examine/evaluate diagnostics from all possible binary logistic regression models
    - In this case, 3 binary logistic regressions

…examine the output

summary(mlogmodel1)

summary(mlogmodel1)

```
Coefficients :
                                           Estimate Std. Error t-value  Pr(>|t|)
  exciting:(intercept)                      3.83945    0.71467  5.3723 7.772e-08
  routine:(intercept)                       3.09104    0.72300  4.2753 1.909e-05
  exciting:sexfemale                        0.23807    1.00957  0.2358   0.81358
  routine:sexfemale                         0.35893    1.01912  0.3522   0.72469
  exciting:satjobnot satisfied             -1.46787    0.81667 -1.7974   0.07227
  routine:satjobnot satisfied              -0.34560    0.82144 -0.4207   0.67395
  exciting:sexfemale:satjobnot satisfied -1.21205    1.11000 -1.0919   0.27486
  routine:sexfemale:satjobnot satisfied  -1.32787    1.11525 -1.1907   0.23379

  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Log-Likelihood: -600.42
  McFadden R^2:  0.054003
  Likelihood ratio test : chisq = 68.552 (p.value = 8.0999e-13)
```

Perform regression again with alternative reflevel

```
mlogmodel2<-mlogit(life ~ 1 | sex + satjob,
   data = mlog1, reflevel = 2)
```

- **Examine the output**

```
summary(mlogmodel2)
```

summary(mlogmodel2)

```
Coefficients :
                                          Estimate Std. Error t-value  Pr(>|t|)
dull:(intercept)                          -3.83945    0.71467 -5.3723 7.772e-08 ***
routine:(intercept)                       -0.74841    0.18298 -4.0902 4.310e-05 ***
dull:sexfemale                            -0.23807    1.00957 -0.2358   0.81358
routine:sexfemale                          0.12086    0.24047  0.5026   0.61525
dull:satjobnot satisfied                   1.46787    0.81667  1.7974   0.07227 .
routine:satjobnot satisfied                1.12227    0.23662  4.7430 2.106e-06 ***
dull:sexfemale:satjobnot satisfied         1.21205    1.11000  1.0919   0.27486
routine:sexfemale:satjobnot satisfied     -0.11582    0.31506 -0.3676   0.71316
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -600.42
McFadden R^2:  0.054003
Likelihood ratio test : chisq = 68.552 (p.value = 8.0999e-13)
>
```

OR = 3.07

Summary

Logistic regression

- **Evaluating models & estimates**
  - Chi-square determines model fit; LLR (deviance) used for model comparisons; ORs and CIs used to assess individual predictors
- **Assumptions/Diagnostics**
  - Fewer than MLR (multiple linear regression), but it is still an important issue to check *a-priori* and to potentially deal with *post-hoc*
    - *Multinomial logistic regression = piecewise*
- Some new concepts (R functions), but a high degree of consistency with MLR

Chi-square and Loglinear Analyses

No DV? No Problem!

- **Chi-square (bivariate)**
- **Loglinear analyses (multivariate)**
  - Assumptions, contingency tables, model comparisons (backward elimination)
- **Example**
  - Job satisfaction, lifestyle, and sex

Pearson’s Chi-Square Test

Andy:

$$\chi^2 = \sum \frac{\left(Observed_{ij} - Model_{ij}\right)^2}{Model_{ij}}$$

$$\text{Model}_{ij} = E_{ij} = \frac{\text{Row Total}_i \times \text{Column Total}_j}{n}$$

- **The ‘model’ is based on ‘expected frequencies’.**
  - Calculated for each of the cells in the contingency table.
  - *n* is the total number of observations.
- **Test statistic**
  - Checked against a distribution with (*r* − 1)(*c* − 1) degrees of freedom.
  - The test distribution is approximate so in small samples use *Fisher’s exact test*.

Likelihood Ratio Statistic

Andy:

$$L\chi^2 = 2\sum Observed_{ij} \ln\left(\frac{Observed_{ij}}{Model_{ij}}\right)$$

- **An alternative to Pearson's chi-square,**
  - based on maximum-likelihood theory.
  - Creates an "optimal" model (probability of obtaining the observed set of data is maximized), and this model is compared to the probability of obtaining those data under the null hypothesis
- **Test statistic**
  - Has a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom.
  - Preferred to the Pearson's chi-square when samples are small.

Statistical Assumptions

- **No distributional assumptions, but there are a few things to watch out for…**

- **Independence of observations:**
  - All cases contribute equally (once): *N* = # of cases
  - Same as MLinR and MLogR

- **Ratio of cases to variables**
  - 5x the # of cases to cells in the design

- **Adequacy of expected values**
  - All expected values are >1 & 80% >5

R U ready?

How are lifestyle, job satisfaction, and sex associated?

- Job satisfaction (high, low, lifestyle (dull, routine, and exciting), and sex (female, male)…sound familiar?
  - 1500 participants, lots of missing values.
- I limit the analyses to the 754 with all three measures.

Install and Load

```
install.packages("gmodels")
install.packages("MASS")

library(gmodels)
library(MASS)
```

χ^2 = CrossTable()

Options, options, options…

- **Default *(X^2)***

CrossTable(logex$satjob, logex$life, chisq = TRUE)

- **Add and suppress statistics (sresid)**

CrossTable(logex$satjob, logex$life, chisq = TRUE, expected = TRUE, prop.c = FALSE, prop.r = FALSE, prop.t = FALSE, prop.chisq = FALSE, sresid = TRUE, format = "SPSS")

- **Add and suppress statistics (asresid)**

CrossTable(logex$satjob, logex$life, chisq = TRUE, expected = TRUE, prop.c = FALSE, prop.r = FALSE, prop.t = FALSE, prop.chisq = FALSE, asresid = TRUE, format = "SPSS")

CrossTable

```
Total Observations in Table:  754

                  | logex1sub$life
logex1sub$satjob  |  routine   |       dull |  exciting  | Row Total |
------------------|------------|------------|------------|-----------|
        satisfied |       107  |          4 |       211  |      322  |
                  |   147.761  |     14.093 |   160.146  |           |
                  |    -6.022  |     -3.632 |     7.488  |           |
------------------|------------|------------|------------|-----------|
    not satisfied |       239  |         29 |       164  |      432  |
                  |   198.239  |     18.907 |   214.854  |           |
                  |     6.022  |      3.632 |    -7.488  |           |
------------------|------------|------------|------------|-----------|
     Column Total |       346  |         33 |       375  |      754  |
------------------|------------|------------|------------|-----------|

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  60.42679     d.f. =  2     p =  7.559437e-14

Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided  p =  2.09576e-14

       Minimum expected frequency: 14.09284
```

Hierarchical loglinear analysis

- Start with saturated model, and eliminate nonsignificant parameters until the most parsimonious model is identified.
  - Sometimes easier said than done…
- **Requires contingency table**
  - xtabs()
- **Requires model comparisons**
  - anova()
- **Requires subsequent post-hoc analyses**
  - CrossTable()

Step 1: Create Contingency table

- xtabs(~ classifying variables, data.frame)

llinx<-xtabs(~ satjob + life + sex, data = logex1sub)

Step 2: Specify Models

```
#Saturated
loglin1_sat<-loglm(~ satjob*life*sex, data=llinx,fit=TRUE)

#no3wayinteraction
no3way<-
loglm(~satjob+life+sex+satjob:life+satjob:sex+life:sex,data=
llinx,fit=TRUE)

#no2wayinteractions
nosexlife<-loglm(~satjob+life+sex + satjob:life + satjob:sex
data = llinx,fit, fit = TRUE)
nojobsex<-loglm(~satjob+life+sex + satjob:life + life:sex,
data = llinx, fit = TRUE)
nojoblife<-loglm(~satjob+life+sex + satjob:sex + life:sex,
data = llinx, fit = TRUE)
nolifeint<-loglm(~satjob+life+sex + satjob:sex,
data= llinx,fit=TRUE)
nojobint<-loglm(~satjob+life+sex + life:sex,
data = llinx, fit = TRUE)
nosexint<-loglm(~satjob+life+sex + satjob:life,
data = llinx, fit = TRUE)

#only main effects
mainonly<-loglm(~ satjob + life + sex,
data = llinx, fit = TRUE)
```

Step 3: Compare models

anova ()…for nested models

```
anova (loglin1_sat,no3way)
#ns, so saturated model not better than model without 3-way interaction

anova (no3way, nosexlife)
anova (no3way, nojobsex)
anova (no3way, nojoblife)
#sign. satjob:life improves fit

anova (nosexlife,nosexint)
#ns, satjob:sex does not improve

anova (nojobsex,nosexint)
#ns, life:sex does not improve

anova (nosexlife, nolifeint)
#sign. satjob:life improves fit
```

Most parsimonious model:

satjob:life + main effects

Step 4: Post-hoc contingency tables

- **Create subsets**

```
table(logex1sub$satjob, logex1sub$life, logex$sex)
xtabs(~sex + satjob + life, data = logex)
justmales = subset(logex1sub, sex ==“male")
justfemales = subset(logex1sub, sex ==“female")
```

- **Perform chi-square**

```
CrossTable(justmales$satjob, justmales$life, asresid
   = TRUE, prop.t=FALSE, prop.r=FALSE,
   prop.c=FALSE, prop.chisq=FALSE, format =
   "SPSS")
```

Life:Satjob

Males

Total Observations in Table: 330

justmales2$satjob	justmales2$life dull	routine	exciting	Row Total
not satisfied	7 5.209 1.226	109 88.555 4.571	75 97.236 -4.959	191
satisfied	2 3.791 -1.226	44 64.445 -4.571	93 70.764 4.959	139
Column Total	9	153	168	330

Pearson's Chi-squared test

Chi^2 = 24.74111   d.f. = 2   p = 4.241656e-06
Minimum expected frequency: 3.790909
Cells with Expected Frequency < 5: 1 of 6 (16.66667%)

Females

Total Observations in Table: 424

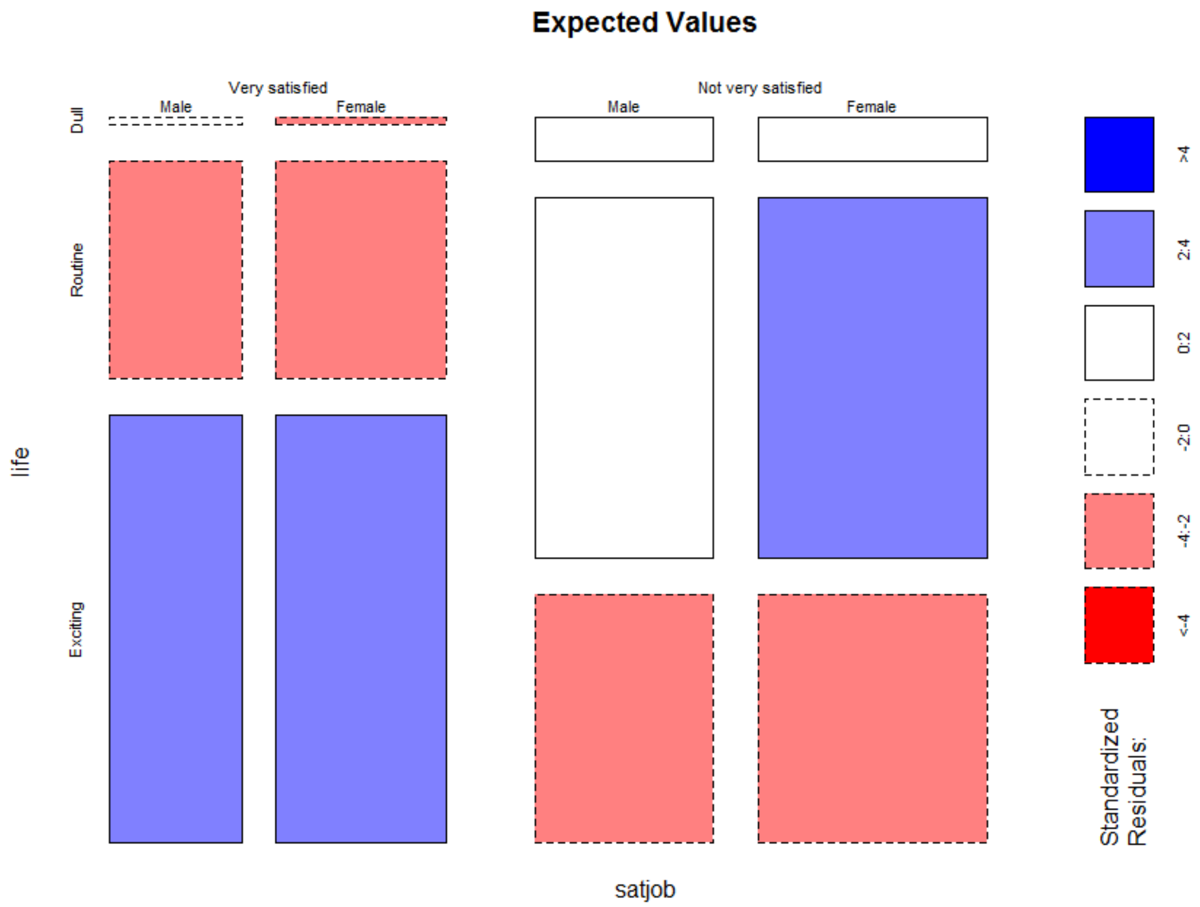
justfemales2$satjob	justfemales2$life dull	routine	exciting	Row Total
not satisfied	22 13.642 3.547	130 109.700 3.997	89 117.658 -5.621	241
satisfied	2 10.358 -3.547	63 83.300 -3.997	118 89.342 5.621	183
Column Total	24	193	207	424

Pearson's Chi-squared test

Chi^2 = 36.7421   d.f. = 2   p = 1.050884e-08
Minimum expected frequency: 10.35849

Mosaic plot

mosaicplot(nosexint$fit, shade = TRUE, main = "Expected Values")



Red is negative (less likely); Blue is positive (more likely)

Summary

Loglinear analyses

- **Hierarchical analyses require several steps**
  - Objective: Identify most parsimonious model
- **Post-hoc analyses**
  - Chi-squares and mosaic plots
- **Raw data and contingency tables**
  - The latter requires a bit more “Runderstanding”