University of Zurich^UZH

Makerere University

Data Analysis with R:

Day 7

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2017

Other topics

So far, we have covered some of the basic **tools** necessary for helping you to start to analyse your own study data.

Many things have been mentioned others have not been included - some are worth mentioning as they are particularly common in some areas of veterinary research.

Normally distributed response variables

So far, we have only considered analyses where the response variable is a **continuous** measurement.

We now have a brief look at two other common types of **response** variables.

The methods we use are largely similar to what we have looked at previously.

Data types in R

- numeric & integers

```
ToothGrowth$len[1:6]

## [1]  4.2 11.5  7.3  5.8  6.4 10.0

bacteria$week[1:6]

## [1]  0  2  4 11  0  2
```

Data types in R

- numeric & integers

```
ToothGrowth$len[1:6]

## [1]  4.2 11.5  7.3  5.8  6.4 10.0

bacteria$week[1:6]

## [1]  0  2  4 11  0  2
```

- unordered factor with 2 levels

```
lung$sex[1:6]

## [1] female male   female female male   female
## Levels: female male
```

- (un/ordered) factor (with more than 2 levels)

```
chickwts$feed[1:6]

## [1] horsebean horsebean horsebean horsebean horsebean horsebean
## Levels: casein horsebean linseed meatmeal soybean sunflower
```

Generalised Linear Modelling

- random component: $\mathbf{E}(\mathbf{Y}) = \mu$
- systematic component: covariates: $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ produce a linear predictor $\eta = \sum_1^p \mathbf{x}_j \beta_j$
- link between random and systematic components: $\mu = \eta$, e.g.
  - **Logistic Regression:**
    $\eta_i = g(\mu_i)$, $g(.)$ is logit function for binomial data
  - **Poisson Regression:**
    $\eta_i = g(\mu_i)$, $g(.)$ is exponential function for poisson data

Logistic Regression: unordered factor with 2 levels as response

E. g. a logistic regression with one covariate (sex) fitting a model to $\mathbb{P}(Y = TRUE) = \pi$, e. g. presence of respiratory symptoms (experienced by the child over the previous 12 months) , gives

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \mathbf{x}_1$$

```r
logreg.mod <- glm(respsymptoms ~ sex,
                  family = binomial, # description of the link function
                  data = lung)
summary(logreg.mod)
```

Logistic Regression: unordered factor with 2 levels as response

```
logreg.mod <- glm(respsymptoms ~ sex,
               family = binomial, # description of the link function
               data = lung)
summary(logreg.mod)
```

```
##
## Call:
## glm(formula = respsymptoms ~ sex, family = binomial, data = lung)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7440  -0.7440  -0.6915  -0.6915   1.7597
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1429     0.1276  -8.957   <2e-16 ***
## sexmale      -0.1663     0.1901  -0.875    0.382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 682.85  on 635  degrees of freedom
## Residual deviance: 682.08  on 634  degrees of freedom
## AIC: 686.08
##
## Number of Fisher Scoring iterations: 4
```

Logistic Regression: unordered factor with 2 levels as response

Interpretation of model coefficients

```
summary(logreg.mod)
```

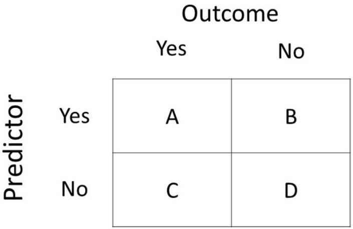
```
##
## Call:
## glm(formula = respsymptoms ~ sex, family = binomial, data = lung)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7440  -0.7440  -0.6915  -0.6915   1.7597
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1429     0.1276  -8.957   <2e-16 ***
## sexmale      -0.1663     0.1901  -0.875    0.382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 682.85  on 635  degrees of freedom
## Residual deviance: 682.08  on 634  degrees of freedom
## AIC: 686.08
##
## Number of Fisher Scoring iterations: 4
```

```
cbind(logodds = coef(logreg.mod), odds = exp(coef(logreg.mod)))
```

```
##                logodds      odds
## (Intercept) -1.1428851 0.3188976
## sexmale     -0.1662919 0.8467990
```

Odds ratios versus log odds ratios

- The logistic regression model in R outputs the log odds ratios.
- $\exp(\text{log odds ratio}) = \text{odds ratio}$
- An **odds ratio (OR)** is a measure of association between an exposure (here: `sex`) and an outcome (here: `respsymptoms`). The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.



		Outcome	
		Yes	No
Predictor	Yes	A	B
	No	C	D

$$\text{OR} = \frac{(A*D)}{(B*C)}$$

Logistic Regression: unordered factor with 2 levels as response

Interpretation of model coefficients (cont'd)

```
cbind(logodds = coef(logreg.mod), odds = exp(coef(logreg.mod)))
```

```
##                logodds      odds
## (Intercept) -1.1428851 0.3188976
## sexmale     -0.1662919 0.8467990
```

The coefficient β_{sex} is the estimated amount by which the log odds of `respsymptoms` would in-/decrease if `sex` is male. The log odds of `respsymptoms` when `sex` is female is just above in the first row ($\beta_{\text{(Intercept)}}$, reference level of `sex`).

Logistic Regression: probability of success as response

- A researcher is examining beetle mortality after 5 hours of exposure to carbon disulphide, at various levels of concentration of the gas.

- Beetles were exposed to gaseous carbon disulphide at various concentrations (in mg/L) for five hours (Bliss, 1935) and the number of beetles killed were noted.

```
library("investr")
data(beetle)
colnames(beetle) <- c("Dose", "Num.Beetles", "Num.Killed")
str(beetle)

## 'data.frame': 8 obs. of  3 variables:
##  $ Dose       : num  1.69 1.72 1.76 1.78 1.81 ...
##  $ Num.Beetles: num  59 60 62 56 63 59 62 60
##  $ Num.Killed : num  6 13 18 28 52 53 61 60

beetle$Num.Surv <- beetle$Num.Beetles - beetle$Num.Killed
```

Logistic Regression: probability of success as response

```
logreg.mod <- glm(cbind(Num.Killed, Num.Surv) ~ Dose,
               family = binomial, # description of the link function
               data = beetle)
summary(logreg.mod)
```

```
##
## Call:
## glm(formula = cbind(Num.Killed, Num.Surv) ~ Dose, family = binomial,
##     data = beetle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5941  -0.3944   0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## Dose          34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

Logistic Regression: probability of success as response

Interpretation of model coefficients

```
summary(logreg.mod)
```

```
##
## Call:
## glm(formula = cbind(Num.Killed, Num.Surv) ~ Dose, family = binomial,
##     data = beetle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5941  -0.3944   0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## Dose          34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

```
cbind(logodds = coef(logreg.mod), odds = exp(coef(logreg.mod)))
```

```
##               logodds         odds
## (Intercept) -60.71745 4.273114e-27
## Dose         34.27033 7.645631e+14
```

Logistic Regression: probability of success as response

Interpretation of model coefficients (cont'd)

```
cbind(logodds = coef(logreg.mod), odds = exp(coef(logreg.mod)))
```

```
##               logodds         odds
## (Intercept) -60.71745 4.273114e-27
## Dose         34.27033 7.645631e+14
```

The coefficient β_{Dose} is the estimated amount by which the log odds of `being killed` would in-/decrease if `Dose` would increase by one unit. The log odds of `being killed` when `Dose` is 0 is just above in the first row $(\beta_{\text{(Intercept)}})$.

Some things to remember

- Using generalised linear models (`glm`) is very similar to the previous models we have seen for normal data (`lm`) - things like residuals need to be checked for randomness, although this is more difficult with counts. Also the residuals being normally distributed is not relevant to these models.
- **As always explore the data first visually and with tables before doing any formal analyses.**

Short excursion to quantile regression

- When talking about regression, we almost exclusively model the conditional mean of a response given one or more explanatory variables.

- With the classical linear model we cannot skewed or otherwise non normal distribution as the corresponding quantiles from the linear model will be misleading. Therefore, we shift our attention to a completely distribution free approach that directly addresses conditional quantile modelling. (Text taken from Hothorn, Torsten, and Brian S. Everitt. A handbook of statistical analyses using R. CRC press, 2014.)

Short excursion to quantile regression

The data set `sambia.csv` (given on the switch drive) contains information on malnutrition of sambian children. The outcome variable is the `z-score`, which gives information on the severity of malnutrition.
The following covariates are considered:

`zscore`	Z-score of the child
`sex`	Sex of the child (0 = female, 1 = male)
`age.child`	Age of the child
`work`	Is the mother working (0 = no, 1 = yes)
`age.mother.birth`	Age of the mother at birth of child
`bmi`	BMI of the mother

A `z-score` indicates how many standard deviations an element is from the mean. A `z-score` can be calculated from the following formula. $z = (X - \mu)/\sigma$ where z is the `z-score`, X is the value of the element, μ is the population mean, and σ is the standard deviation.

Quantile regression in R

```r
sambia <- read.csv("~/201710_Makerere/data/sambia.csv", sep = ",")
str(sambia)
sambia$sex <- factor(sambia$sex, levels = c(0,1),
                     labels = c("female", "male"))
sambia$work <- factor(sambia$work, levels = c(0,1),
                      labels = c("no", "yes"))
```

```r
str(sambia)

## 'data.frame': 4421 obs. of  7 variables:
##  $ X                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ zscore           : int  -159 -205 -192 -146 69 -80 -34 -17 -225 468 ...
##  $ sex              : Factor w/ 2 levels "female","male": 1 1 2 2 1 1 1 2 1 2
##  $ age.child        : int  4 26 56 6 54 1 2 2 29 14 ...
##  $ work             : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 1 2 2 2 ...
##  $ age.mother.birth: num  24.7 22.8 15.3 21.5 17.5 ...
##  $ bmi              : num  21.8 21.8 20.4 22.3 22.3 ...
```

Quantile regression in R (cont’d)

```r
library("quantreg")

## Loading required package:  SparseM
##
## Attaching package:  ’SparseM’
## The following object is masked from ’package:base’:
##
##     backsolve

# tau = 2.5 %
rq.025 <- rq(zscore ~ sex + age.child + work + bmi + age.mother.birth,
             tau = 0.025,
             data = sambia)
# tau = 5 %
rq.05 <- rq(zscore ~ sex + age.child + work + bmi + age.mother.birth,
            tau = 0.05,
            data = sambia)
# tau = 50 %
rq.5 <- rq(zscore ~ sex + age.child + work + bmi + age.mother.birth,
           tau = 0.5,
           data = sambia)
```

Quantile regression in R (cont'd)

Comparison of the three models `rq.025`, `rq.05`, `rq.5`

```
rq_coef <- cbind(coef(rq.025), coef(rq.05), coef(rq.5))
rownames(rq_coef) <- rownames(summary(rq.5)$coefficients)
colnames(rq_coef) <- paste0("tau = ", c("0.025", "0.05", "0.5"))
round(rq_coef, 3)
```

```
##                  tau = 0.025 tau = 0.05 tau = 0.5
## (Intercept)         -523.808   -491.812  -242.188
## sexmale                1.511     -3.816   -13.197
## age.child             -2.235     -2.496    -1.903
## workyes              -13.947    -16.280    -6.489
## bmi                    6.947      7.884     4.966
## age.mother.birth       0.244      0.465     0.830
```

Quantile regression in R (cont'd)

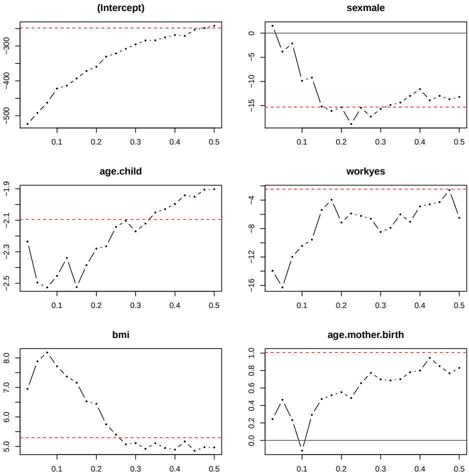
Estimate `rq` for a grid of quantiles

```r
# Estimate rq for a single tau
# rq.5 <- rq(zscore ~ sex + age.child + work + bmi + age.mother.birth,
#            tau = 0.5,
#            data = sambia)
# Estimate rq for a grid of tau
rq <- rq(zscore ~ sex + age.child + work + bmi + age.mother.birth,
         tau = seq(from = 0.025, to = 0.5, by = 0.025),
         data = sambia)
```

Quantile regression in R (cont'd)

Plot results

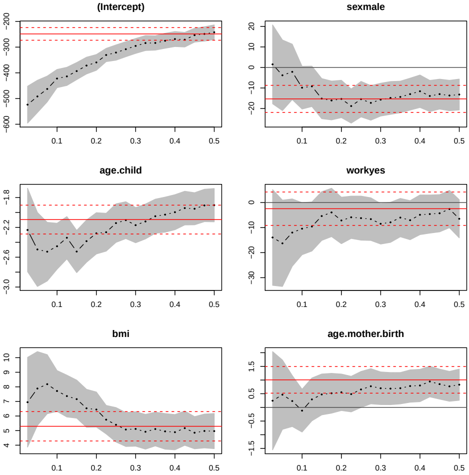
```r
plot(rq) # plot including line for the OLS coefficient (as estimated by lm)
```



Quantile regression in R (cont'd)

Plot results in more details

```r
# A sequence of coefficient estimates for quantile regressions with
# varying tau parameters is visualized along with associated confidence bands.
plot(summary(rq))
```


(Intercept)
sexmale
age.child
workyes
bmi
age.mother.birth