Data analysis with R software

Data-analyysi R-ohjelmistolla

Tommi Härkänen

National Institute for Health and Welfare (THL), Helsinki
E-mail: `tommi.harkanen@helsinki.fi`

University of Helsinki, February 12, 2013

Contents

Categorical covariates

Interaction of a categorical and a continuous covariate

Interaction of two categorical covariates

Level of measurement

Variables have been categorized into 4 categories[1]:

Categorical variables Qualitative data.

- Nominal No meaningful ordering, e.g. marital status. Possible to estimate **point probabilities** (prevalences), **mode**.
- Ordinal Values are ordered but differences are not meaningful, e.g. education: basic, middle, high. Possible to estimate also **median** or other **quantiles**.

Continuous variables Quantitative data.

- Interval Differences are meaningful, e.g. temperature in Celsius or Fahrenheit. Possible to estimate also **means** and **standard deviations**.
- Ratio "Zero" exists, thus possible to present relative differences. E.g. geographical distances, age, height and weight.

[1] Stevens, S.S (June 7, 1946). "On the Theory of Scales of Measurement". Science 103 (2684): 677–680.

Qualitative and quantitative data in R

Categorical variables are of type `factor`

Nominal E.g.,

```
factor(c(9, 12, 17, 9, 17, 17), levels = c(9, 12, 17),
    labels = c("basic", "middle", "high"))

## [1] basic  middle high   basic  high   high
## Levels: basic middle high
```

Ordinal Function `ordered` is used, e.g.

```
ordered(c(9, 12, 17, 9, 17, 17), levels = c(9, 12,
    17), labels = c("basic", "middle", "high"))

## [1] basic  middle high   basic  high   high
## Levels: basic < middle < high
```

Continuous variables are numerical variables.

Categorical covariate in a regression model

Subset "Ever had any pain in chest" of the NHANES data: weight, "get chest pain when walk uphill or hurry" and age

```
prop.table(table(nhanes[, "haf2"]))

##
##                 Yes          No (HAF9) Never uphill/hurry
##              0.2852             0.6620             0.0528
```

Research question: “Are there differences in average weight between chest pain groups?”
Note that the age distributions differ between chest pain groups:

```
summary(lm(hsageir ~ haf2, data = nhanes))[["coefficients"]]

##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 50.35      0.492  102.38 0.00e+00
## haf2No (HAF9)               -2.63      0.588   -4.48 7.59e-06
## haf2Never uphill/hurry      19.03      1.261   15.09 1.87e-50
```

Categorical covariate in a regression model

Adjusting for confounders age and height

```
summary(lm(ham6s_kg ~ haf2 + hsageir + ham5s_m, data = nhanes))

##
## Call:
## lm(formula = ham6s_kg ~ haf2 + hsageir + ham5s_m, data = nhanes)
##
## Residuals:
##     Min      1Q Median      3Q     Max
##   -56.9   -10.9   -2.4     8.0   115.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -58.1105     3.6552  -15.90  < 2e-16 ***
## haf2No (HAF9)             -1.8680     0.4949   -3.77  0.00016 ***
## haf2Never uphill/hurry    -3.5172     1.1097   -3.17  0.00154 **
## hsageir                    0.0186     0.0112    1.66  0.09696 .
## ham5s_m                   79.8349     2.1418   37.27  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Categorical covariate in a regression model

Change the reference level of chest pain variable

Usually the reference level is chosen to be the group with **lowest risk** or **largest size**.
Here the group `haf2=="No"` is the largest, so choose that using `relevel()`:

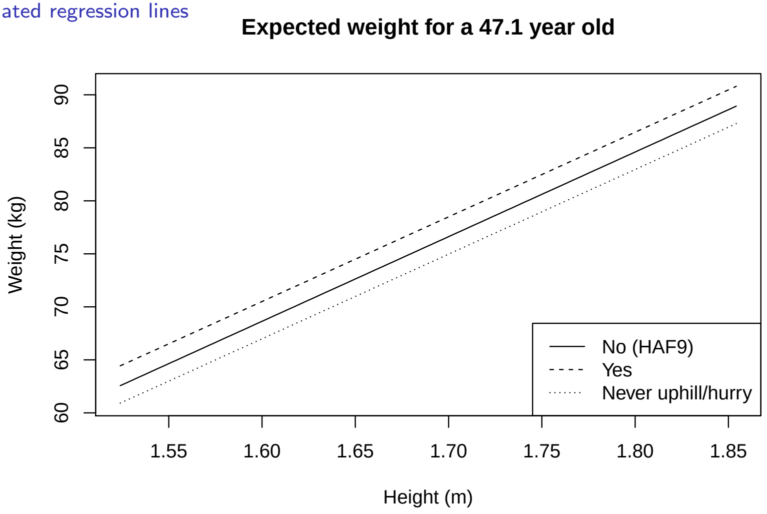
```
nhanes[, "haf2"] <- relevel(nhanes[, "haf2"], "No (HAF9)")
summary(lm(ham6s_kg ~ haf2 + hsageir + ham5s_m, data = nhanes))[["coefficien

##                          Estimate Std. Error t value  Pr(>|t|)
## (Intercept)              -59.9785     3.6768  -16.31  2.16e-58
## haf2Yes                    1.8680     0.4949    3.77  1.62e-04
## haf2Never uphill/hurry    -1.6492     1.0722   -1.54  1.24e-01
## hsageir                    0.0186     0.0112    1.66  9.70e-02
## ham5s_m                   79.8349     2.1418   37.27 5.48e-270
```

Note that the `haf2No` line has changed.
The regression coefficients correspond now to the differences

- `haf2=="No"` vs. `haf2=="Yes"` and
- `haf2=="No"` vs. `haf2=="Never uphill/hurry"`

Categorical covariate in a regression model

Estimated regression lines



Regression coefficients

Interaction of continuous and categorical covariates

Imaginary example in R: `lm(y ~ age + gender + age*gender)`

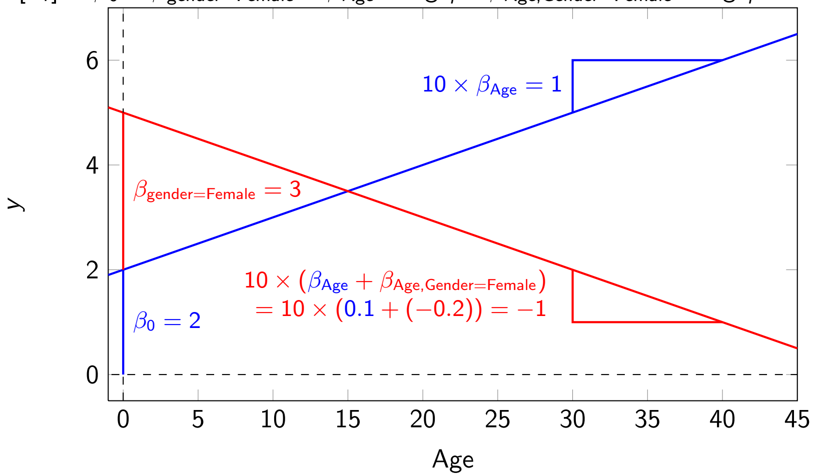
```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.0         ...
age                      0.1         ...
genderFemale             3.0         ...
age:genderFemale        -0.2         ...
```

Age	Gender	Linear predictor	Prediction
0	Male	$2.0 + 0 \times 0.1 + 0 \times 3.0 + (-0.2) \times 0 \times 0 =$	2.0
0	Female	$2.0 + 0 \times 0.1 + 1 \times 3.0 + (-0.2) \times 0 \times 0 =$	5.0
40	Male	$2.0 + 40 \times 0.1 + 0 \times 3.0 + (-0.2) \times 40 \times 0 =$	6.0
40	Female	$2.0 + 40 \times 0.1 + 1 \times 3.0 + (-0.2) \times 40 \times 1 =$	1.0

Regression coefficients

Interaction of continuous and categorical covariates

$$\mathbb{E}[Y_i] = \beta_0 + \beta_{\text{gender=Female}} + \beta_{\text{Age}} \times \text{Age}_i + \beta_{\text{Age,Gender=Female}} \times \text{Age}_i$$



Example of interaction of two categorical covariates

Using Nhanes data. Regress weight on gender, smoking (`har1`, “Have you smoked 100+ cigarettes in life”) and their interaction.

```
fit1 <- with(nhanes, lm(ham6s_kg ~ hssex + har1 + hssex * har1))
round(summary(fit1)$coefficients, d = 2)

##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     80.63       0.21  382.43     0.00
## hssexFemale                    -11.84       0.33  -35.73     0.00
## har1No (HAR14)                  -0.96       0.34   -2.82     0.00
## hssexFemale:har1No (HAR14)       0.17       0.47    0.36     0.72
```

Gender	Smoking	Linear predictor	Prediction
Male	Yes	80.6 + 0×-11.8 + 0×-0.96 + 0×0.17 =	80.6
Female	Yes	80.6 + 1×-11.8 + 0×-0.96 + 0×0.17 =	68.8
Male	No (HAR14)	80.6 + 0×-11.8 + 1×-0.96 + 0×0.17 =	79.7
Female	No (HAR14)	80.6 + 1×-11.8 + 1×-0.96 + 1×0.17 =	68