Data-analyysi R-ohjelmistolla

Tommi Härkänen

Terveyden ja hyvinvoinnin laitos (THL), Helsinki
E-mail: tommi.harkanen@helsinki.fi

Helsingin yliopisto, 9.4.2014

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
##              0.3246              0.6173              0.0582
```

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)                51.581       1.56  33.171 3.98e-131
## haf2No (HAF9)              -0.995       1.92  -0.518  6.04e-01
## haf2Never uphill/hurry     17.074       4.09   4.171  3.55e-05
```


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
## -31.92 -10.77  -2.80   7.47  75.65
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -32.77112   12.29927   -2.66    0.008 **
## haf2No (HAF9)            -1.81151    1.63987   -1.10    0.270
## haf2Never uphill/hurry   -3.55547    3.44924   -1.03    0.303
## hsageir                  -0.00469    0.03744   -0.13    0.900
## ham5s_m                  66.26697    7.28601    9.10   <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()`:

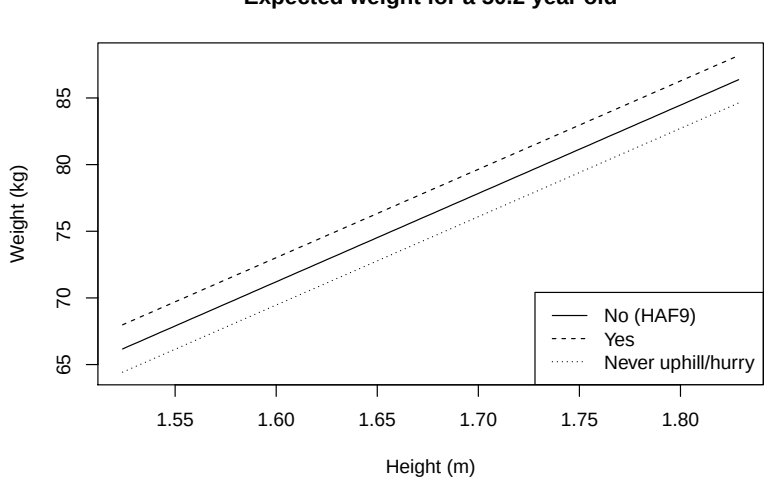
```
nhanes[, "haf2"] <- relevel(nhanes[, "haf2"], "No (HAF9)")
summary(lm(ham6s_kg ~ haf2 + hsageir + ham5s_m, data = nhanes))[["coefficien

##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -34.58263    12.3106  -2.809 5.18e-03
## haf2Yes                  1.81151     1.6399   1.105 2.70e-01
## haf2Never uphill/hurry  -1.74396     3.3326  -0.523 6.01e-01
## hsageir                 -0.00469     0.0374  -0.125 9.00e-01
## ham5s_m                 66.26697     7.2860   9.095 2.89e-18
```

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)`

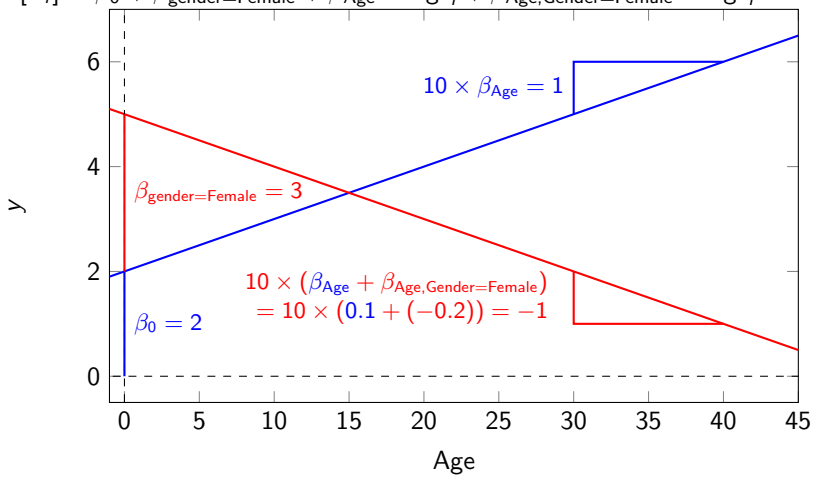
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0	...		
age	0.1	...		
genderFemale	3.0	...		
age:genderFemale	-0.2	...		

Age	Gender	Linear predictor	Prediction
0	Male	$2.0 + 0\times0.1 + 0\times3.0 + (\text{-}0.2)\times0\times0 =$	2.0
0	Female	$2.0 + 0\times0.1 + 1\times3.0 + (\text{-}0.2)\times0\times0 =$	5.0
40	Male	$2.0 + 40\times0.1 + 0\times3.0 + (\text{-}0.2)\times40\times0 =$	6.0
40	Female	$2.0 + 40\times0.1 + 1\times3.0 + (\text{-}0.2)\times40\times1 =$	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)                     79.27       0.64  124.74     0.00
## hssexFemale                     -8.05       1.10   -7.35     0.00
## har1No (HAR14)                  -1.94       1.06   -1.83     0.07
## hssexFemale:har1No (HAR14)      -1.62       1.51   -1.07     0.28
```

Gender	Smoking	Linear predictor	Prediction
Male	Yes	$79.3 + 0\times\text{-}8.05 + 0\times\text{-}1.94 + 0\times\text{-}1.62 =$	79.3
Female	Yes	$79.3 + 1\times\text{-}8.05 + 0\times\text{-}1.94 + 0\times\text{-}1.62 =$	71.2
Male	No (HAR14)	$79.3 + 0\times\text{-}8.05 + 1\times\text{-}1.94 + 0\times\text{-}1.62 =$	77.3
Female	No (HAR14)	$79.3 + 1\times\text{-}8.05 + 1\times\text{-}1.94 + 1\times\text{-}1.62 =$	67.7