Adding Categorical Predictors

Math 430, Winter 2017

The data

`UN11.csv` contains national health, welfare, and education statistics for 210 places, mostly UN members, but also other areas like Hong Kong that are not independent countries.

Variable	Description
`region`	region of the world
`group`	`oecd`, `africa`, or `other`
`fertility`	number of children per woman
`ppgdp`	per capita gross domestic product (US$)
`lifeExpF`	female life expectancy (years)
`pctUrban`	% urban

The data

```
UN11 <- read.csv("https://github.com/math430-lu/data/raw/master/UN11.csv")
head(UN11)
```

```
##        region  group fertility   ppgdp lifeExpF pctUrban
## 1        Asia  other     5.968   499.0    49.49       23
## 2      Europe  other     1.525  3677.2    80.40       53
## 3      Africa africa     2.142  4473.0    75.00       67
## 4      Africa africa     5.135  4321.9    53.17       59
## 5   Caribbean  other     2.000 13750.1    81.10      100
## 6 Latin Amer  other     2.172  9162.1    79.89       93
```

Problem overview

Primary research question

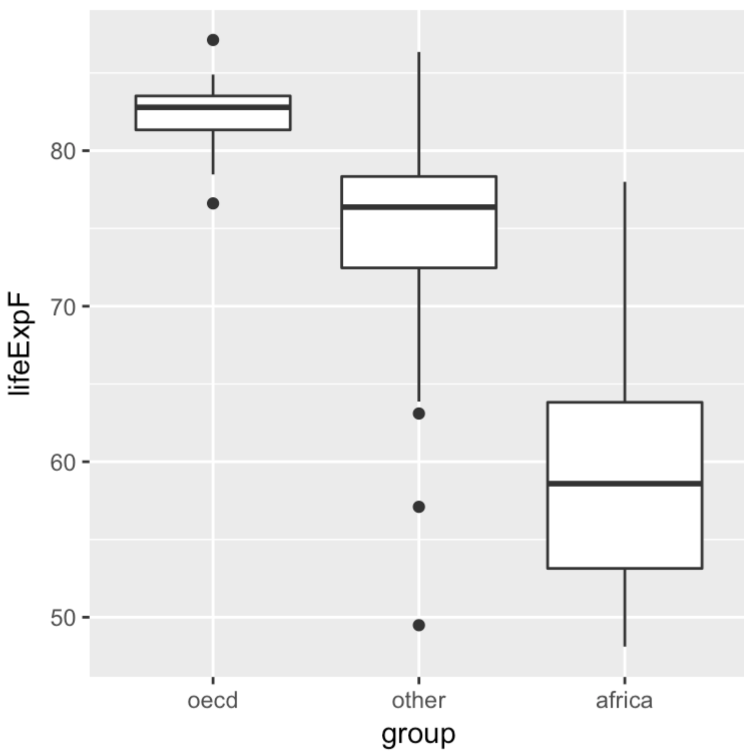
- How does the expected life span of women (`lifeExpF`) differ between the three groups of countries?

Analysis

- What is the response variable?
- What should we use for the predictor?

Life expectancy by group

```
UN11$group <- factor(UN11$group, levels = c("oecd", "other", "africa"))
ggplot(data = UN11, aes(x = group, y = lifeExpF)) +
  geom_boxplot()
```



Categorical predictors

- We use **dummy variables** (i.e. indicator variables) to put categories in a mathematical formula
- General idea:
- Coding `group` into dummy variables

Fitting the no-intercept model in R

Model formula:

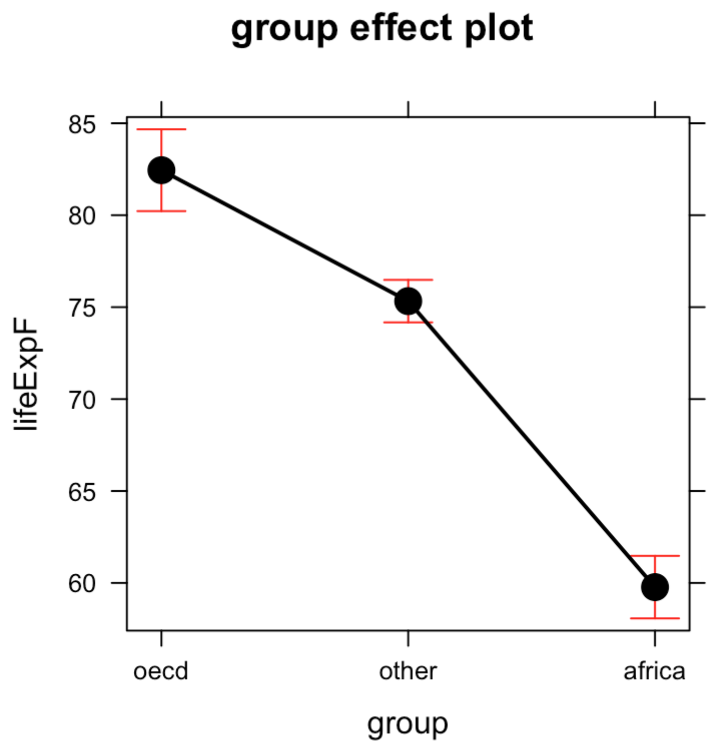
```
lexp_grp <- lm(lifeExpF ~ group - 1, data = UN11)
summary(lexp_grp)
```

```
##
## Call:
## lm(formula = lifeExpF ~ group - 1, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8367  -3.3045   0.3635   2.7183  18.2277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## groupoecd    82.4465     1.1279   73.09   <2e-16 ***
## groupother   75.3267     0.5856  128.63   <2e-16 ***
## groupafrica  59.7723     0.8626   69.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.28 on 196 degrees of freedom
## Multiple R-squared:  0.9927, Adjusted R-squared:  0.9926
## F-statistic:  8896 on 3 and 196 DF,  p-value: < 2.2e-16
```

Interpreting the coefficients

Interpreting the coefficients

```
plot(Effect("group", mod = lexp_grp))
```



Fitting the model with an intercept in R

```
lexp_grp <- lm(lifeExpF ~ group, data = UN11)
summary(lexp_grp)
```

```
##
## Call:
## lm(formula = lifeExpF ~ group, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8367  -3.3045   0.3635   2.7183  18.2277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.446      1.128  73.095  < 2e-16 ***
## groupother    -7.120      1.271  -5.602  7.1e-08 ***
## groupafrica  -22.674      1.420 -15.968  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.28 on 196 degrees of freedom
## Multiple R-squared:  0.6191, Adjusted R-squared:  0.6152
## F-statistic: 159.3 on 2 and 196 DF,  p-value: < 2.2e-16
```

Interpreting the coefficients

Comparing level means

Comparing level means

```
vcov(lexp_grp)
```

```
##             (Intercept) groupother groupafrica
## (Intercept)    1.272249  -1.272249   -1.272249
## groupother    -1.272249   1.615204    1.272249
## groupafrica   -1.272249   1.272249    2.016395
```

Adding a continuous predictor

Research question

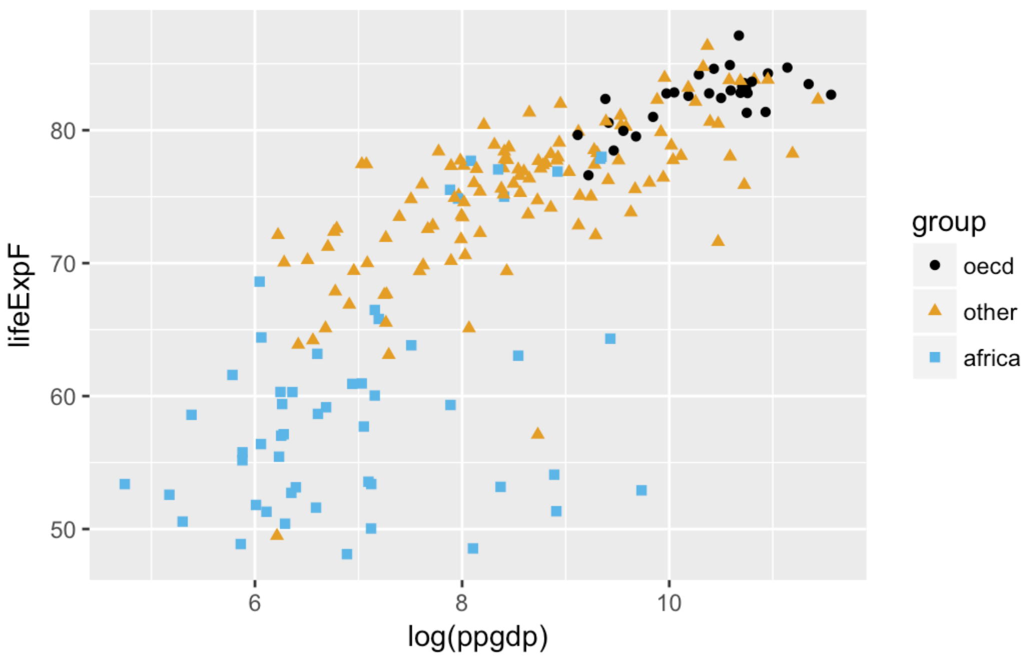
- Can we find a model that better explains expected life span by incorporating other predictors?

Analysis

- MLR model with `group` and other predictors

Adding a continuous predictor

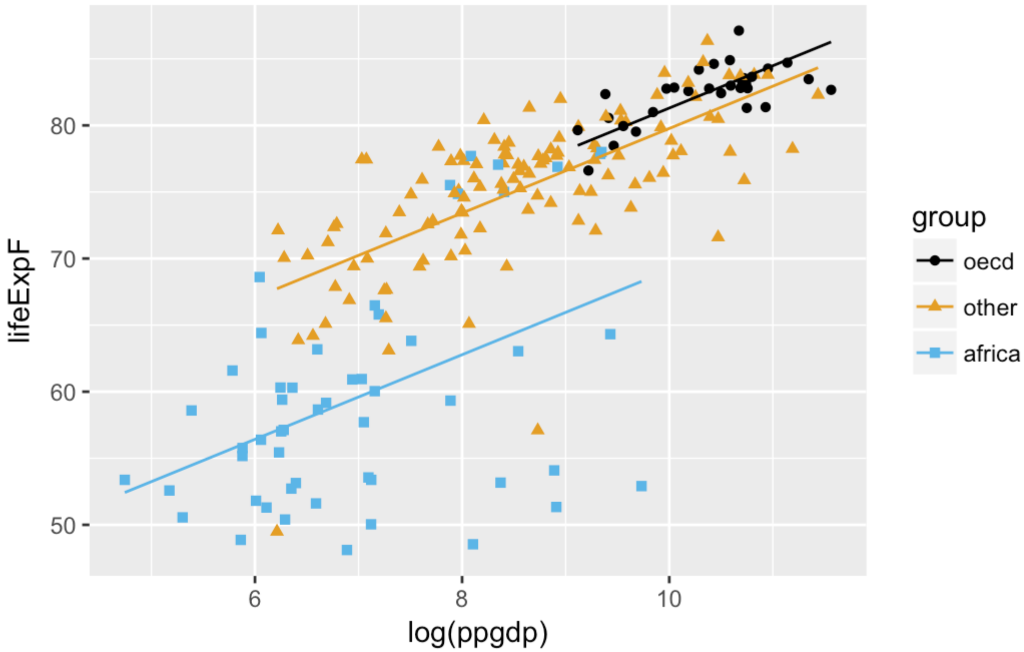
```
ggplot(data = UN11, aes(x = log(ppgdp), y = lifeExpF, color = group, shape = group)) +
  geom_point() +
  scale_color_colorblind()
```



How can we build this model?

Parallel lines model

$$Y = \beta_0 + \beta_1 I_{other} + \beta_2 I_{africa} + \beta_3 x + e$$



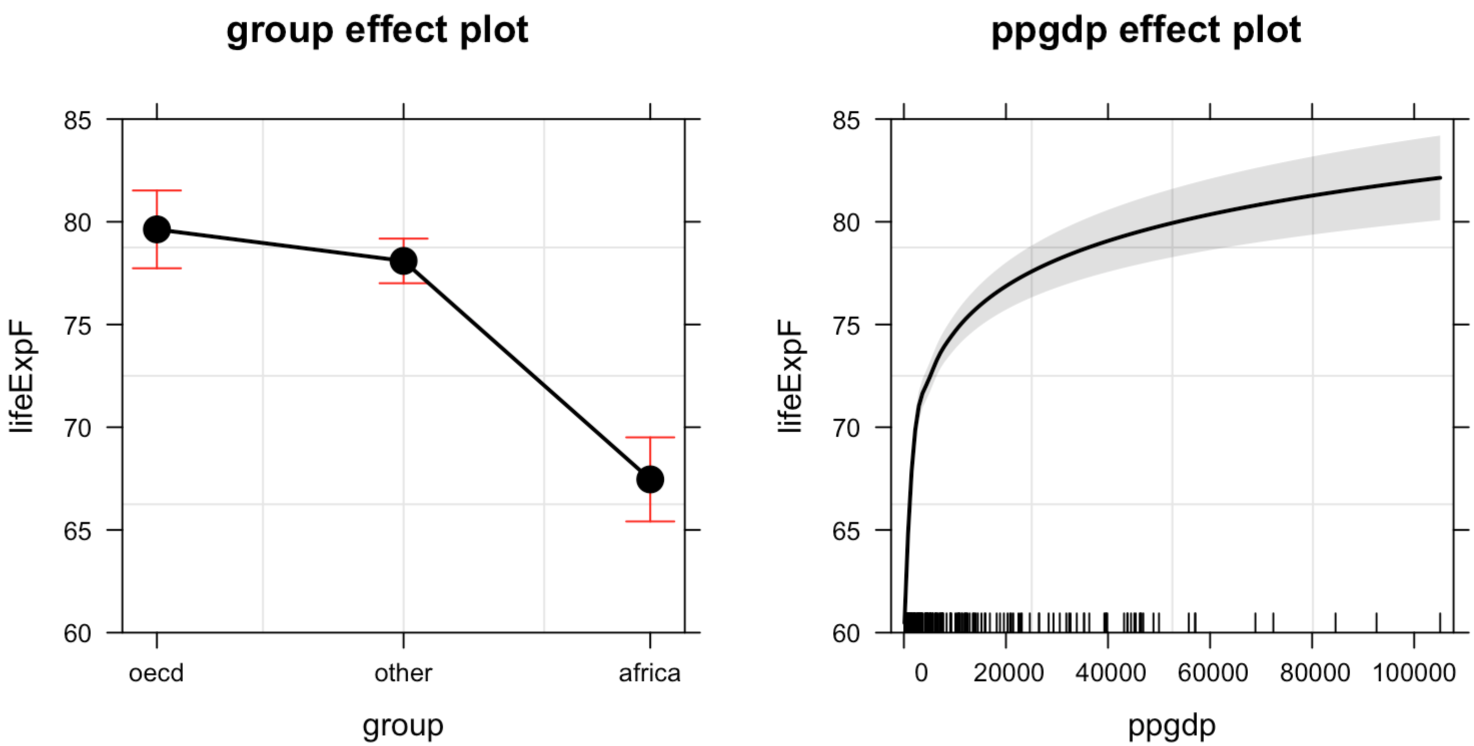
Fitting the model in R

```
parallel_mod <- lm(lifeExpF ~ group + log(ppgdp), data = UN11)
summary(parallel_mod)
```

```
##
## Call:
## lm(formula = lifeExpF ~ group + log(ppgdp), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6348  -2.1741   0.2441   2.3537  14.6539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   49.529      3.400  14.569  < 2e-16 ***
## groupother    -1.535      1.174  -1.308    0.193
## groupafrica  -12.170      1.557  -7.814 3.35e-13 ***
## log(ppgdp)     3.177      0.316  10.056  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.109 on 195 degrees of freedom
## Multiple R-squared:  0.7492, Adjusted R-squared:  0.7453
## F-statistic: 194.1 on 3 and 195 DF,  p-value: < 2.2e-16
```

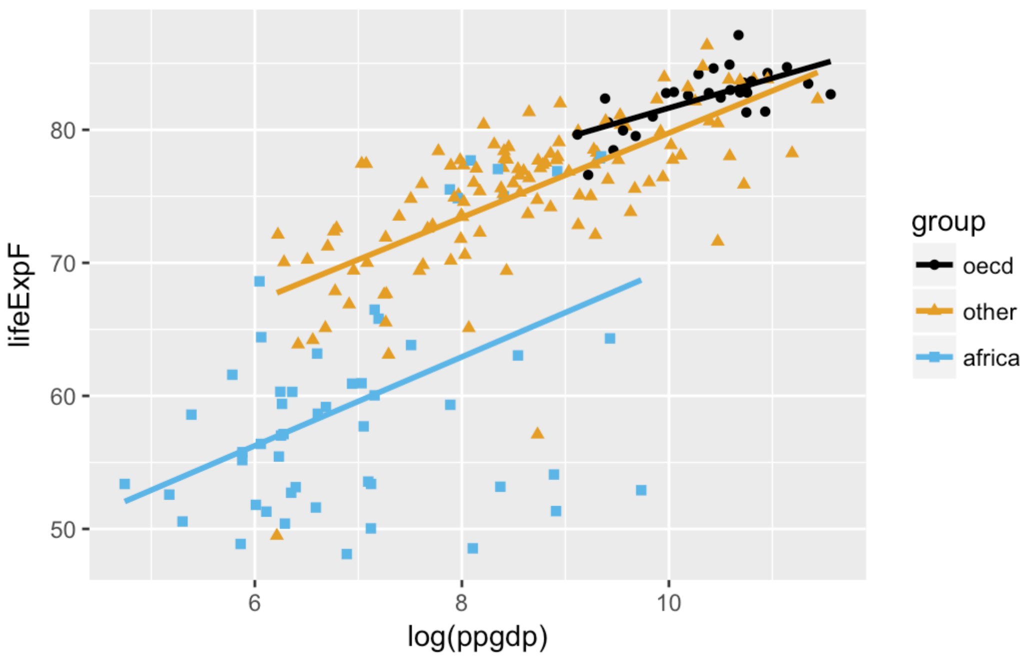
Interpreting the coefficients

```
plot(allEffects(parallel_mod, default.levels=50), ylim=c(60, 85),
     grid=TRUE, multiline=TRUE)
```



Unrelated lines model

$$Y = \beta_0 + \beta_1 I_{other} + \beta_2 I_{africa} + \beta_3 x + \beta_4 I_{other} x + \beta_5 I_{africa} x + e$$



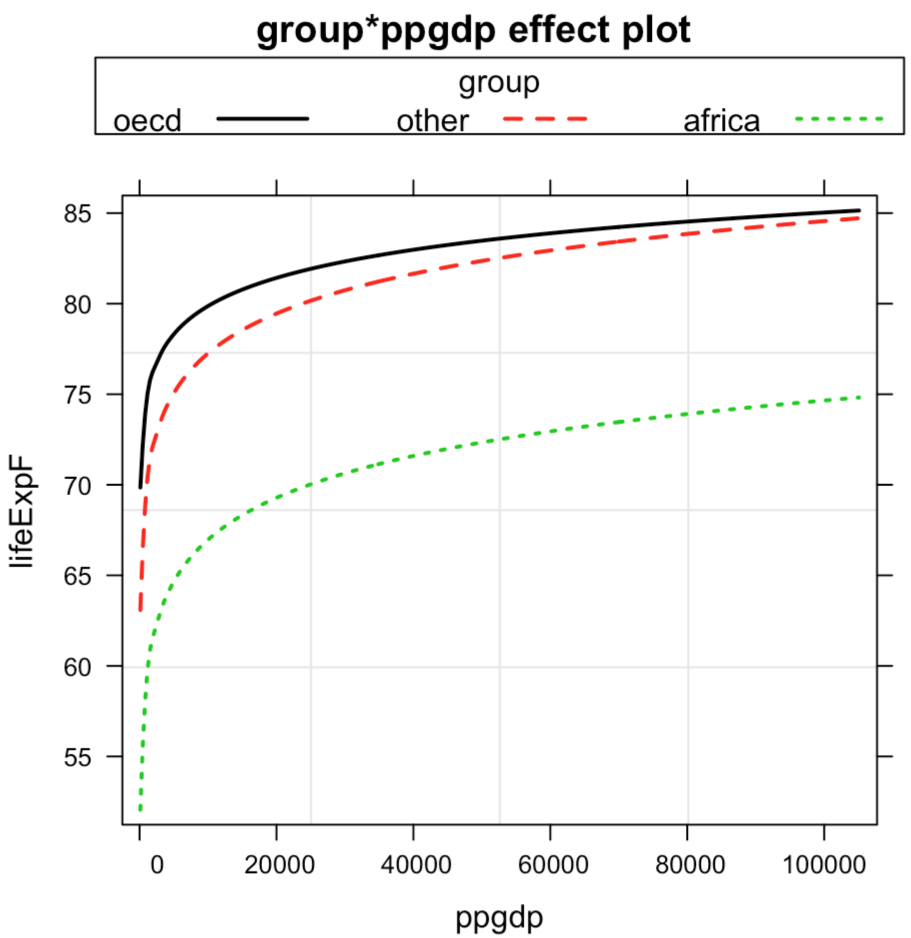
Fitting the model in R

```
unrelated_mod <- lm(lifeExpF ~ group * log(ppgdp), data = UN11)
summary(unrelated_mod)
```

```
##
## Call:
## lm(formula = lifeExpF ~ group * log(ppgdp), data = UN11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.634  -2.089   0.301   2.255  14.489
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            59.2137    15.2203   3.890 0.000138 ***
## groupother            -11.1731    15.5948  -0.716 0.474572
## groupafrica           -22.9848    15.7838  -1.456 0.146954
## log(ppgdp)              2.2425     1.4664   1.529 0.127844
## groupother:log(ppgdp)   0.9294     1.5177   0.612 0.540986
## groupafrica:log(ppgdp)  1.0950     1.5785   0.694 0.488703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.129 on 193 degrees of freedom
## Multiple R-squared:  0.7498, Adjusted R-squared:  0.7433
## F-statistic: 115.7 on 5 and 193 DF,  p-value: < 2.2e-16
```

Interpreting the coefficients

```
plot(Effect(c("group", "ppgdp"), unrelated_mod, default.levels=100),
     rug=FALSE, grid=TRUE, multiline=TRUE)
```



Comparing models

How can we determine which model we should prefer?

Partial F-tests

Partial F-tests in R

```
anova(parallel_mod, unrelated_mod)
```

```
## Analysis of Variance Table
##
## Model 1: lifeExpF ~ group + log(ppgdp)
## Model 2: lifeExpF ~ group * log(ppgdp)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    195 5090.4
## 2    193 5077.7  2    12.675 0.2409 0.7862
```