Team 3 / WVS: Exploratory Linear Model Plots

Michael Friendly

June 27, 2017

Exploratory model-based plots

One data analysis strategy for an unknown data set is to produce a set of purely exploratory graphs of just data, and see what turns up.

An alternative is to use simple models, mainly as a smoothing technique to get a **higher-level summary**, with the understanding that your model may not be the best summary, or may be totally **wrong**. But, you can **tune** the model (non-linear terms, add interactions, ...) and then the plotting step(s) **remain the same** for the new model(s).

The caveats are:

- Numeric variables may actually be factors, not quantitative, but you forgot to use `as.factor()`
- You don't understand the direction of coding: is `1` the highest or lowest?
- You forgot to account for, or analyse missing data, which, annoyingly was not coded as `NA` in the `RData` file, but used negative values a la SAS or SPSS for various reasons why data was missing.

But, this can give a jump start over purely EDA approaches, and allows me to show some things from R packages that John Fox & I have developed: `car`, `effects`, `heplots`.

```
library(effects)
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:effects':
##
##     Prestige
```

```
library(heplots)
```

Load our semi-cleaned data file

```
load("WVS.Rdata")
head(WVS)
```

```
##   country code happiness health life_sat volition marital kids
## 1           12         2      1        8        7  single    0
## 2           12         2      2        5        6  single    0
## 3           12         2      2        4        6  single    0
## 4           12         2      1        8        6  single    0
## 5           12         1      3        8        6 married    3
## 6           12         2      1        7        4 married    2
##   financial_sat social_class income    sex age education country
## 1            10            4      5   male  21         7 Algeria
## 2            10            3      6 female  24         7 Algeria
## 3             6            4      6 female  26         5 Algeria
## 4             6            4      5 female  28         6 Algeria
## 5             4            3      7 female  35         3 Algeria
## 6             8            3      5   male  36         8 Algeria
```

Fit some univariate linear models

We chose three possible variables to treat as responses:

```
"V10",  # happiness
"V11",  # health
"V23",  # life satisfaction
```

The possible predictors kept changing from one discussion to the next. What I'm using here are just:

```
"V57",  # marital_status
"V58",  # kids
"V238", # social class
"V240", # sex
"V248"  # education
```

Let's fit a simple additive linear model to each of these. The strategy here is to:

- Fit `mod <- lm()`
- Run `summary(mod)` and/or `car::Anova(mod)` to see test statistics, R^2, etc.
- Run `effects::plot(allEffects(mod))` to see a visual summary of the model predicted values for **each** predictor, controlling for all other variables in the model.

Think of an *effect plot* as one kind of visual summary of a table of (partial regression) coefficients.

Happiness

```
wvs.mod1 <- lm(happiness ~ marital + kids + social_class + sex + education,
        data=WVS)
summary(wvs.mod1)
```

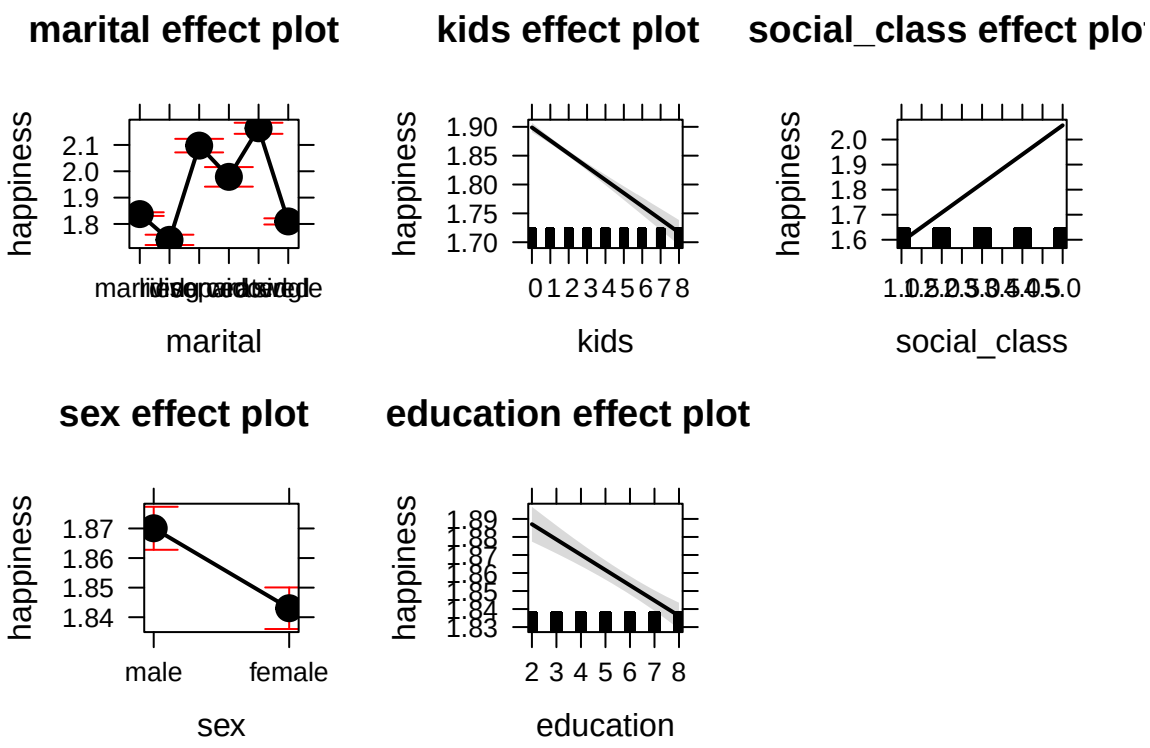
```
##
## Call:
## lm(formula = happiness ~ marital + kids + social_class + sex +
##     education, data = WVS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44307 -0.70635  0.08836  0.29633  2.58696
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.561923   0.014082 110.915  < 2e-16 ***
## maritalliving    -0.097753   0.010631  -9.195  < 2e-16 ***
## maritaldivorced   0.260003   0.013638  19.065  < 2e-16 ***
## maritalseparated  0.141375   0.019274   7.335 2.24e-13 ***
## maritalwidowed    0.325796   0.011172  29.162  < 2e-16 ***
## maritalsingle    -0.027771   0.007368  -3.769 0.000164 ***
## kids             -0.022750   0.001759 -12.937  < 2e-16 ***
## social_class      0.116155   0.002700  43.020  < 2e-16 ***
## sexfemale        -0.027024   0.005196  -5.201 1.99e-07 ***
## education        -0.008476   0.001156  -7.334 2.25e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7297 on 81274 degrees of freedom
```

```
## Multiple R-squared:  0.04609,    Adjusted R-squared:  0.04599
## F-statistic: 436.4 on 9 and 81274 DF,  p-value: < 2.2e-16
```

With such a huge sample size, everything is a significant effect. What we want to understand is roughly the magnitude and direction of the effects of each predictor on happiness in a relatively compact display.

Effect plot, using the short-hand default `plot(allEffects(mod))`. Details of the screen size are important here. There are many, many graphic options to improve presentation, none used here.

```
plot(allEffects(wvs.mod1))
```



NB:

- This plot (by default) can be misleading, because each panel is scaled *separately*, so small absolute effects for a given predictor can appearlarger than if a common scale was used for all panels.
- This "model" looks pretty good, until we notice that the $R^2 = 0.04$!!! But— hey, we *can* say with a straight face that $R^2 > 0$, however, We aren't using this for inference.

Health

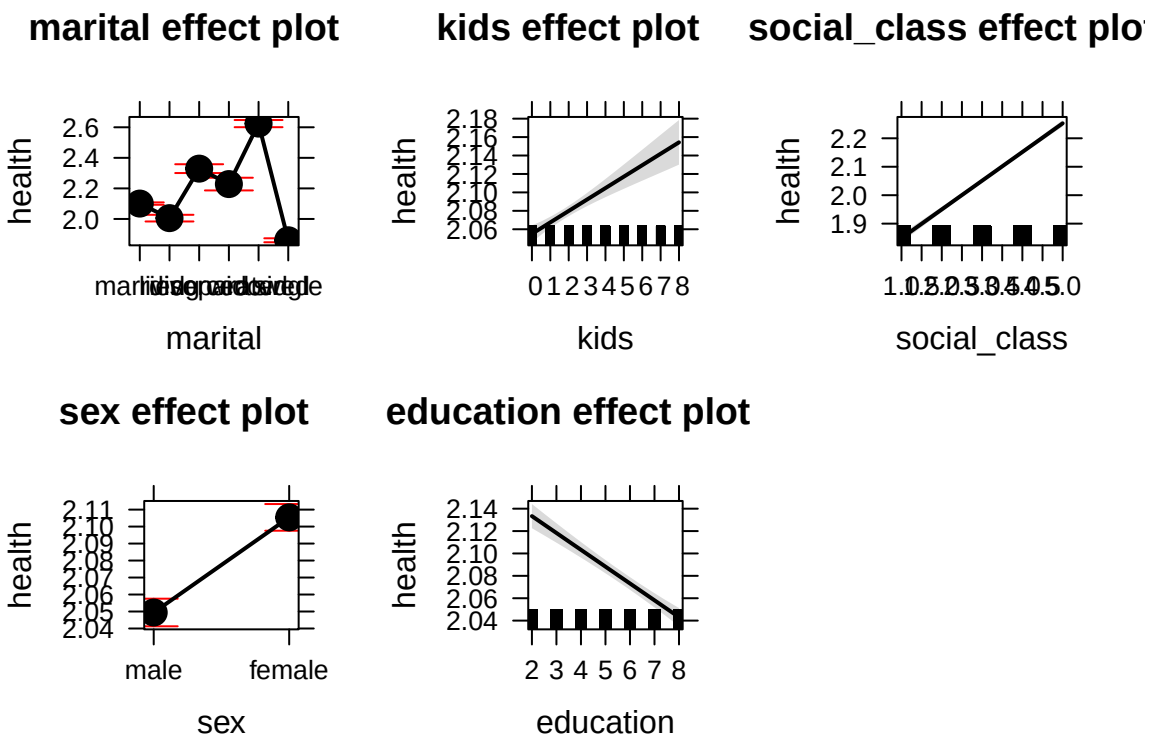
```
wvs.mod2 <- lm(health ~ marital + kids + social_class + sex + education, data=WVS)
summary(wvs.mod2)
```

```
##
## Call:
## lm(formula = health ~ marital + kids + social_class + sex + education,
##     data = WVS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.95628 -0.71971 -0.04306  0.65927  2.45645
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.803165   0.015731 114.627  < 2e-16 ***
## maritalliving     -0.095319   0.011876  -8.026 1.02e-15 ***
## maritaldivorced    0.228683   0.015235  15.011  < 2e-16 ***
## maritalseparated   0.127566   0.021530   5.925 3.14e-09 ***
## maritalwidowed     0.523408   0.012480  41.940  < 2e-16 ***
## maritalsingle     -0.240176   0.008230 -29.183  < 2e-16 ***
## kids               0.012412   0.001964   6.318 2.66e-10 ***
## social_class       0.100897   0.003016  33.452  < 2e-16 ***
## sexfemale          0.056008   0.005804   9.649  < 2e-16 ***
## education         -0.015041   0.001291 -11.651  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8152 on 81274 degrees of freedom
## Multiple R-squared:  0.0802, Adjusted R-squared:  0.0801
## F-statistic: 787.4 on 9 and 81274 DF,  p-value: < 2.2e-16
```

Effect plot:

```
plot(allEffects(wvs.mod2))
```



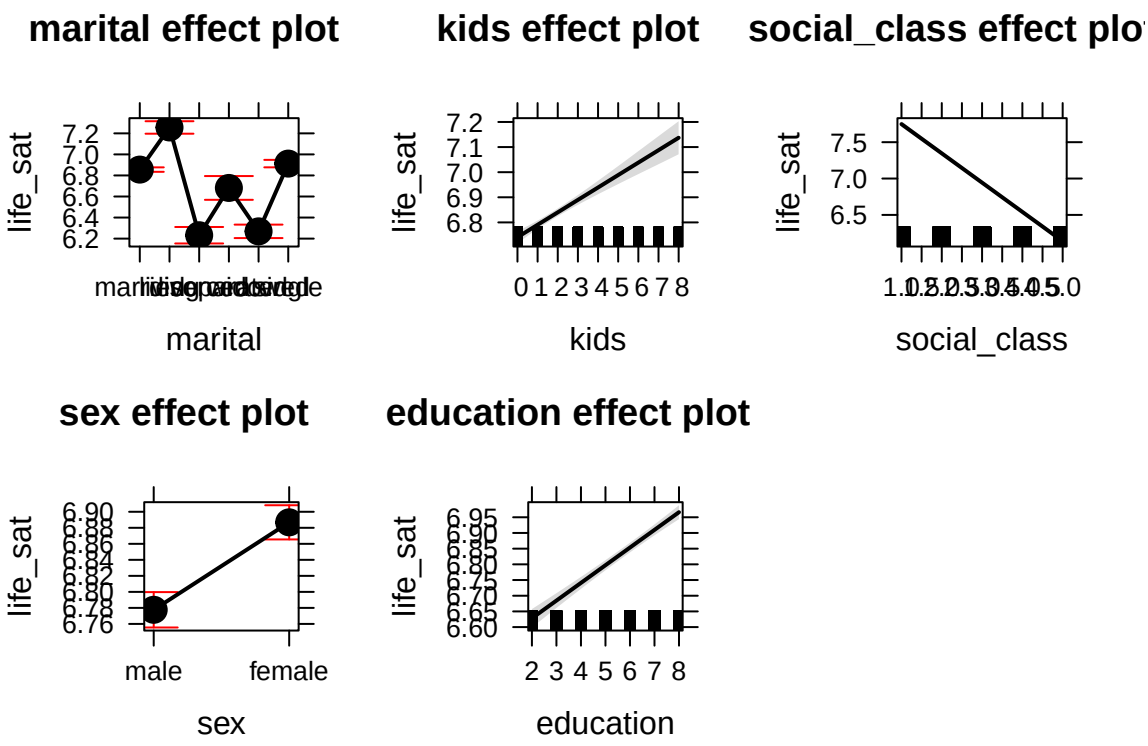
Life satisfaction

```
wvs.mod3 <- lm(life_sat ~ marital + kids + social_class + sex + education, data=WVS)
summary(wvs.mod3)
```

```
##
## Call:
## lm(formula = life_sat ~ marital + kids + social_class + sex +
##     education, data = WVS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2281 -1.4073  0.2817  1.5467  4.8746
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.704413   0.042776 180.110  < 2e-16 ***
## maritalliving     0.400188   0.032294  12.392  < 2e-16 ***
## maritaldivorced  -0.623535   0.041427 -15.051  < 2e-16 ***
## maritalseparated -0.174274   0.058547  -2.977  0.00292 **
## maritalwidowed   -0.586723   0.033936 -17.289  < 2e-16 ***
## maritalsingle     0.056866   0.022380   2.541  0.01106 *
## kids              0.049524   0.005342   9.271  < 2e-16 ***
## social_class     -0.402369   0.008202 -49.059  < 2e-16 ***
## sexfemale         0.109225   0.015784   6.920 4.55e-12 ***
## education         0.056341   0.003511  16.049  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.217 on 81274 degrees of freedom
## Multiple R-squared:  0.05043,    Adjusted R-squared:  0.05032
## F-statistic: 479.6 on 9 and 81274 DF,  p-value: < 2.2e-16
```

Effect plot:

```
plot(allEffects(wvs.mod3))
```


marital effect plot
life_sat
7.2
7.0
6.8
6.6
6.4
6.2
marital
kids effect plot
life_sat
7.2
7.1
7.0
6.9
6.8
0 1 2 3 4 5 6 7 8
kids
social_class effect plo
life_sat
7.5
7.0
6.5
social_class
sex effect plot
life_sat
male
female
sex
education effect plot
life_sat
6.95
6.65
6.60
2 3 4 5 6 7 8
education