

Chapters 13 and 22 Homework

Do Exercise D13.1 for the multiple regression of the total fertility rate on GDP per capita, female illiteracy, contraception in the U.N. data. Repeat the exercise adding dummy regressors for the factor region to the model.

Do Exercise 22.1 (*note*: NOT D22.1) (a), (b), and (c).

To generate the data in part (a),

```
set.seed(xxx)  
Dataset <- as.data.frame(matrix(rnorm(500*101), 500, 101))
```

Replace ~~xxx~~ by a three-digit random number, which you can generate by entering `sample(999, 1)` at the R > command prompt; e.g.,

```
> sample(999, 1)  
[1] 901
```

Of course, you'll likely get a number different from 901 (which is the basic idea). Note that this is not part of the R script in your .Rmd file. Using an explicit, known seed for R's random number generator will make your randomly generated data reproducible (i.e., the same each time you compile your .Rmd file into a .html file), but different from other students' randomly generated data.

Then you can perform the regression as

```
mod.1 <- lm(V1 ~ ., data=Dataset)  
summary(mod.1)
```

The variables are automatically named V1 through V101; the model formula `V1 ~ .` specifies a regression of V1 on all of the other variables in the `Dataset` data frame.

For part (c), you can use the `step` function in R to perform a stepwise regression by backward elimination for these variables:

```
step(mod.1, k=1000)
```

The argument `k=1000` is intended to force `step` to run to completion: see `?step` for details.

Examine the (voluminous) output from `step` to locate the selected model with three explanatory variables; this will be near the end of the output. Don't feel that it is necessary to include *all* of the `step` output in your printed homework – in fact, please don't do it: You can edit the .html file produced by RStudio (simply open the file in the RStudio editor) to elide most of the `step` output before you print your homework; retain the output from, say, the first one or two steps and the final three or four steps, and delete the rest.

Notes: `step` will take a while to run to completion (about 30 seconds on my very fast Windows desktop), and, when you run the command in the R console, the output will overflow the RStudio output buffer, so some of the earlier output will be lost. That's OK, since you're only interested in the model with three explanatory variables, which is near the end of the output. The `.html` file that you produce will, however, include all of the `step` output – hence the necessity to edit it.

In parts (b) and (c), you're asked to fit models with three of the explanatory variables, determined respectively by the t -statistics from the full regression and by the stepwise regression. For example, if the explanatory variables with the largest t -statistics were V5, V15, and V28, you would proceed as follows in part (b):

```
mod.2 <- lm(V1 ~ V5 + V15 + V28, data=Dataset)
summary(mod.2)
```