Multivariable regression examples

Brian Caffo, Jeff Leek and Roger Peng

May 19, 2016

Data set for discussion

```
require(datasets); data(swiss); ?swiss
```

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].

- [,1] Fertility a common standardized fertility measure
- [,2] Agriculture % of males involved in agriculture as occupation
- [,3] Examination % draftees receiving highest mark on army examination
- [,4] Education % education beyond primary school for draftees
- [,5] Catholic % catholic (as opposed to protestant)
- [,6] Infant.Mortality live births who live less than 1 year

All variables but Fertility give proportions of the population.

Calling `lm`

```
summary(lm(Fertility ~ .   , data = swiss))
```

Example interpretation

- Agriculture is expressed in percentages (0 - 100)
- Estimate is -0.1721.
- Our models estimates an expected 0.17 decrease in standardized fertility for every 1% increase in percentage of males involved in agriculture in holding the remaining variables constant.
- The t-test for $H_0 : \beta_{Agri} = 0$ versus $H_a : \beta_{Agri} \neq 0$ is significant.
- Interestingly, the unadjusted estimate is

```
summary(lm(Fertility ~ Agriculture, data = swiss))$coeffic
```

How can adjustment reverse the sign of an effect? Let's try a simulation.

```
n <- 100; x2 <- 1 : n; x1 <- .01 * x2 + runif(n, -.1, .1);
summary(lm(y ~ x1))$coef
summary(lm(y ~ x1 + x2))$coef
```

Back to this data set

- The sign reverses itself with the inclusion of Examination and Education.
- The percent of males in the province working in agriculture is negatively related to educational attainment (correlation of -0.6395225) and Education and Examination (correlation of 0.6984153) are obviously measuring similar things.
- Is the positive marginal an artifact for not having accounted for, say, Education level? (Education does have a stronger effect, by the way.)
- At the minimum, anyone claiming that provinces that are more agricultural have higher fertility rates would immediately be open to criticism.

What if we include an unnecessary variable?

z adds no new linear information, since it's a linear combination of variables already included. R just drops terms that are linear combinations of other terms.

```
z <- swiss$Agriculture + swiss$Education
lm(Fertility ~ . + z, data = swiss)
```

Dummy variables are smart

- Consider the linear model

$$Y_i = \beta_0 + X_{i1}\beta_1 + \epsilon_i$$

where each X_{i1} is binary so that it is a 1 if measurement i is in a group and 0 otherwise. (Treated versus not in a clinical trial, for example.)
- Then for people in the group $E[Y_i] = \beta_0 + \beta_1$
- And for people not in the group $E[Y_i] = \beta_0$
- The LS fits work out to be $\hat\beta_0 + \hat\beta_1$ is the mean for those in the group and $\hat\beta_0$ is the mean for those not in the group.
- β_1 is interpretted as the increase or decrease in the mean comparing those in the group to those not.
- Note including a binary variable that is 1 for those not in the group would be redundant. It would create three parameters to describe two means.

More than 2 levels

- Consider a multilevel factor level. For didactic reasons, let's say a three level factor (example, US political party affiliation: Republican, Democrat, Independent)
- $Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i$.
- X_{i1} is 1 for Republicans and 0 otherwise.
- X_{i2} is 1 for Democrats and 0 otherwise.
- If i is Republican $E[Y_i] = \beta_0 + \beta_1$
- If i is Democrat $E[Y_i] = \beta_0 + \beta_2$.
- If i is Independent $E[Y_i] = \beta_0$.
- β_1 compares Republicans to Independents.
- β_2 compares Democrats to Independents.
- $\beta_1 - \beta_2$ compares Republicans to Democrats.
- (Choice of reference category changes the interpretation.)

Insect Sprays

Linear model fit, group A is the reference

```
summary(lm(count ~ spray, data = InsectSprays))$coef
```

Hard coding the dummy variables

```
summary(lm(count ~
             I(1 * (spray == 'B')) + I(1 * (spray == 'C'))
             I(1 * (spray == 'D')) + I(1 * (spray == 'E'))
             I(1 * (spray == 'F'))
           , data = InsectSprays))$coef
```

What if we include all 6?

```
summary(lm(count ~
    I(1 * (spray == 'B')) + I(1 * (spray == 'C')) +
    I(1 * (spray == 'D')) + I(1 * (spray == 'E')) +
    I(1 * (spray == 'F')) + I(1 * (spray == 'A')), data = In
```

What if we omit the intercept?

```r
summary(lm(count ~ spray - 1, data = InsectSprays))$coef
library(dplyr)
summarise(group_by(InsectSprays, spray), mn = mean(count))
```

Reordering the levels

```
spray2 <- relevel(InsectSprays$spray, "C")
summary(lm(count ~ spray2, data = InsectSprays))$coef
```

Summary

- If we treat Spray as a factor, R includes an intercept and omits the alphabetically first level of the factor.
- All t-tests are for comparisons of Sprays versus Spray A.
- Emprirical mean for A is the intercept.
- Other group means are the itc plus their coefficient.
- If we omit an intercept, then it includes terms for all levels of the factor.
- Group means are the coefficients.
- Tests are tests of whether the groups are different than zero. (Are the expected counts zero for that spray.)
- If we want comparisons between, Spray B and C, say we could refit the model with C (or B) as the reference level.

Other thoughts on this data

- Counts are bounded from below by 0, violates the assumption of normality of the errors.
- Also there are counts near zero, so both the actual assumption and the intent of the assumption are violated.
- Variance does not appear to be constant.
- Perhaps taking logs of the counts would help.
- There are 0 counts, so maybe $\log(\text{Count} + 1)$
- Also, we'll cover Poisson GLMs for fitting count data.

Recall the swiss data set

```r
library(datasets); data(swiss)
head(swiss)
```

Create a binary variable

```r
library(dplyr);
swiss = mutate(swiss, CatholicBin = 1 * (Catholic > 50))
```

Plot the data

No effect of religion

```r
summary(lm(Fertility ~ Agriculture, data = swiss))$coef
```

The associated fitted line

Parallel lines

```
summary(lm(Fertility ~ Agriculture + factor(CatholicBin), c
```

Fitted lines

Lines with different slopes and intercepts

```
summary(lm(Fertility ~ Agriculture * factor(CatholicBin), d
```

Fitted lines

Just to show you it can be done

```
summary(lm(Fertility ~ Agriculture + Agriculture : factor((
```