Activity 7.14 **[S2009/S2009-29]**

We're going to use the ten-mile-race data to explore the idea of redundancy: Why redundancy is a problem and what we can do about it.

Read in the data:

```
> run = ISMdata("ten-mile-race.csv")
```

The data includes information about the runner's age and sex, as well as the time it took to run the race.

I'm interested in how computer and cell-phone use as a child may have affected the runner's ability. I don't have any information about computer use, but as a rough proxy, I'm going to use the runner's year of birth. The assumption is that runners who were born in the 1950s, 60s, and 70s, didn't have much chance to use computers as children.

Add in a new variable: yob. We'll approximate this as the runner's age subtracted from the year in which the race was run: 2005. That might be off by a year for any given person, but it will be pretty good.

```
> run$yob = 2005 - run$age
```

Each of the following models has two terms.

```
mod1 = lm( net ~ age + yob - 1, data=run)
mod2 = lm( net ~ 1 + age, data = run)
mod3 = lm( net ~ 1 + yob, data=run )
```

- Fit each of the the models and interpret the coefficients in terms of the relationship between age and year of birth and running time. Then look at the R^2 and the sum of square residuals in order to decide which is the better model.

Using special software that you don't have, I have fitted a model — I'll call it mod4 — with all three terms: the intercept, age, and year of birth. The model coefficients are:

My Fantastic Model: mod4

Intercept	age	yob
-20050	20.831891052	12.642004612

My conclusion, based on the mod4 coefficients, is that people slow down by 20.8 seconds for every year they age. Making up for this, however, is the fact that people who were born earlier in the last century tend to run slower by 12.6 seconds for every year later they were born. Presumably this is because those born earlier had less opportunity to use computers and cell phones and therefore went out and did healthful, energetic, physical play rather than typing.

- Using these coefficients, calculate the model values. The statement will look like this:

```
mod4vals = -20050 + 20.831891052*run$age +
    12.642004612*run$yob
```

- Calculate the residuals from mod4 by substracting the model values from the response variable (net running time). Compare the size of the residuals using a sum of squares or a standard deviation or a variance to the size of the residuals from models 1 through 3. Judging from this, which is the better model?

I needed special software to find the coefficients in mod4 because R won't do it. See what happens when you try the models with three terms, like this:

```
lm( net ~ 1 + age + yob, data=run )
lm( net ~ 1 + yob + age, data=run )
```

- Can you get three coefficients from the R software?

I'm very pleased with mod4 and the special methods I used to find the coefficients.

Unfortunately, my statistical arch-enemy, Prof. Nalpak Ynnad, has proposed another model. He claims that computer and cell-phone use is helpful. According to his twisted theory, people actually run faster as they get older. Impossible! But look at his model coefficients.

Ynnad's Evil Model: mod5

Intercept	age	yob
60150	-19.16810895	-27.35799539

Ynnad's ridiculous explanation is that the natural process of aging (that you run faster as you age), is masked by the beneficial effects of exposure to computers and cell phones as a child. That is, today's kids would be even slower (because they are young) except for the fact that they use computers and cell phones so much. Presumably, when they grow up, they will be super fast, benefiting both from their advanced age and from the head start they got as children from their exposure to computers and cell phones.

- Looking at Ynnad's model in terms of the R^2 or size of residuals, how does it compare to my model? Which one should you believe?

- Give an explanation of why both my model and Ynnad's model are bogus. See if you can also explain why we shouldn't take the coefficients in mod1 seriously at face value.