Math 455 Class 14 February 21

Pitfalls of Wilkinson - Rogers Notation in R.

Wilkinson - Rogers notation is used to specify a linear regression, e.g.

```
lmod <- lm(Y ~ X)
```

- convenient in the sense that specification is very short.
- annoying because symbols have unexpected meanings / effects.

Today : some simple examples of this

Next page : a bit of the R manual from which I took these examples

- Ways to deal with this :
  - Stack Exchange / Chat GPT / manual.
  - experiment with small data
  - know it already.

Examples

Before giving a formal specification, a few examples may usefully set the picture.

Suppose `y`, `x`, `x0`, `x1`, `x2`, ... are numeric variables, `X` is a matrix and `A`, `B`, `C`, ... are factors. The following formulae on the left side below specify statistical models as described on the right.

`y ~ x`
`y ~ 1 + x`

Both imply the same simple linear regression model of *y* on *x*. The first has an implicit intercept term, and the second an explicit one.

`y ~ 0 + x`
`y ~ -1 + x`
`y ~ x - 1`

Simple linear regression of *y* on *x* through the origin (that is, without an intercept term).

`log(y) ~ x1 + x2`

Multiple regression of the transformed variable,log(y), on *x1* and *x2* (with an implicit intercept term).

`y ~ poly(x,2)`
`y ~ 1 + x + I(x^2)`

Polynomial regression of *y* on *x* of degree 2. The first form uses orthogonal polynomials, and the second uses explicit powers, as basis.

`y ~ X + poly(x,2)`

Multiple regression *y* with model matrix consisting of the matrix *X* as well as polynomial terms in *x* to degree 2.

`y ~ A`

Single classification analysis of variance model of *y*, with classes determined by *A*.

`y ~ A + x`

Single classification analysis of covariance model of *y*, with classes determined by *A*, and with covariate *x*.

`y ~ A*B`
`y ~ A + B + A:B`
`y ~ B %in% A`
`y ~ A/B`

Two factor non-additive model of *y* on *A* and *B*. The first two specify the same crossed classification and the second two specify the same nested classification. In abstract terms all four specify the same model subspace.

`y ~ (A + B + C)^2`
`y ~ A*B*C - A:B:C`

Three factor experiment but with a model containing main effects and two

Recall that a regression coefficient is only defined in the context of the regression in which it is computed.

If we change the specification of the regression, the coefficient may change sign and/or significance.

Sometimes the effect produces unexpected coefficients.

Remember the example

Shark Attacks ~ Ice-Cream Sales.

Once we include another variable for weather, the coefficient of Ice-Cream Sales becomes insignificant.

This phenomenon is sometimes called "Omitted Variable Bias": the coefficient of Ice-Cream sales has greater magnitude and significance than it "should" have because it is acting as a proxy for something that "should" be present (weather) but is not.

Another example: 3 vars W, X, Y.

~~the~~ data on X explain Y well.

~~the~~ But X is expensive to obtain.

W is highly correlated to X.

We think W will explain Y well.

But it doesn't!

How is this possible?

Recall that in a 1-variable regression $Y \sim X$.

$R^2 = \text{cor}(Y, X)^2$

What we are saying ~~the~~ is: it's possible that $\text{cor}(Y, X)$ is "large" and $\text{cor}(X, W)$ is large but $\text{cor}(Y, W)$ is NOT.

<u>Example</u>: it's possible that

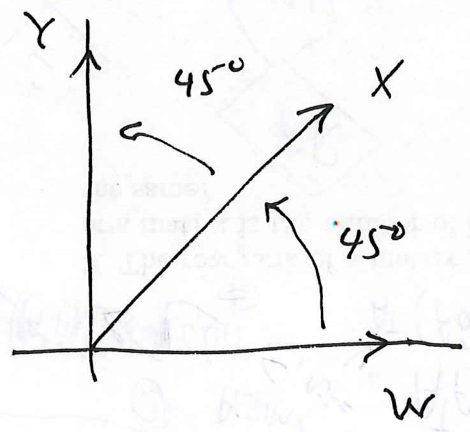
$\text{cor}(Y, X) \approx 0.7$ $\quad\quad$ $\text{Cor}(X, W) \approx 0.7$

but $\quad \text{Cor}(Y, W) \approx 0$.

How do we know this is possible?

Remember that correlation is like the cosine of an angle and.

$$\cos(45^\circ) = \frac{\sqrt{2}}{2} \approx 0.707\ldots$$



$\cos(45^\circ) \approx 0.7$

$\cos(90^\circ) = 0$.

Note that bounds can be produced with careful study of $\cos^{-1}$ function.

To ~~gen~~ get nonzero lower bound on cor(Y, W) we need higher values of cor(Y, X) & cor(X, W).

These bounds come from:

angle between (Y, W) ≤ angle between (Y, X) + angle between (X, W).

MIDTERM Monday March 3.