

Stat 215b (Spring 2004): Comments on Lab 3

B. M. Bolstad
bolstad@stat.berkeley.edu

Feb 19, 2004

As we are around the halfway point for this class, everyone should know how to write a lab report for this class. If you have not already, please read the instructions on the web (linked off the webpage). This particular lab assignment was fairly difficult and in general there was a lot of room for improvement. This document has a few comments on some common mistakes.

Is checking agreement in distribution or mean enough?

Many people just checked that the distribution of the active and passive samplers agreed. Others did a two-sample t-test to check agreements in mean. This is not enough. Since each pair of measurements should correspond to each other you should check that they agree in some manner, or are related. Not completely unsensible approaches would be to test paired sample t-test or to consider a model

$$\text{Active} = \beta_0 + \beta_1 \text{Passive} + \epsilon$$

and carry out tests on β_1 or both β_0 and β_1 .

To see why testing agreement in distribution is not enough to say that two estimates (samples) are measuring the same variable which might be changing over time (Ozone Level) consider this artificial example. Suppose X_1 is randomly distributed with mean μ and some distribution F . Similarly suppose that X_2 is random distributed with mean μ and some distribution F . Suppose we take n independent samples for each of X_1 and X_2 . There should be no relationship between X_1 and X_2 , yet their means are the same as are their distributions.

```
x1 <- rexp(1000,10)
x2 <- rexp(1000,10)

plot(sort(x1),sort(x2)) # a qq plot

t.stat <- (mean(x1) - mean(x2))/sqrt(var(x1)/100 + var(x2)/100)

2*(1-pt(abs(t.stat),df=1000-2)) # should not be significant

plot(x1,x2) # no relationship
```

Variables in model selection procedure

You should think carefully about what terms you are going to include or exclude from the model selection process. Does it make sense to include something like SID or Date as a covariate variable (rather than a factor)? Probably not, since I don't think there is any reason to believe that there would be a linear change in the ozone by changing sid from 1 to 2 to 3 to etc. Do the cross-product terms really make sense?

Prediction

A sensible thing to do might be to compare two models developed using model selection procedures by seeing how well they do at predicting new data. A sensible metric might be something like average squared prediction error or average absolute prediction error. In either case the best model would have the smallest value. Several people looked at average prediction error. This is not a sensible approach because a worse method could still have a better average (closer to zero). example

```
truth <- c(0,1,2,3,4)
method1 <- c(-0.5,1.5,2.1,2.75,4.25)
method2 <- c(-1,2,2,2.5,4.5)

# an incorrect approach
mean(method1 - truth) # this is larger than
mean(method2 - truth) # this one

# a better approach
mean((method1-truth)^2) # now this is smaller
mean((method2 - truth)^2) # then this
```