STA 302/1001

Note Title 12/6/2011

Reminder

Office Hours: Today Andriy 11-1 UC 244
Tomorrow me 11:15-1 SS 5016A
Thursday me 2-4 SS 5016A

(Cody* will be in NC aid centre Thursday 12-3)
* priority to STA 250

Remark requests? By Friday Dec 9 (firm.)

Please take your marked Assignment 3 before you go today.

Experiments versus Observational Studies

- experiment - a treatment imposed on experimental units
  eg. meadowfoam
  Advantages - strength of conclusion
  - If properly randomized can make cause-and-effect conclusions.
  eg can say that increasing light intensity causes a decrease in # of flowers/plant on average.
- observational study - data measured without intervention

eg. house prices, CFCs, assignments

- can say that there is an association, but can't say X causes Y
- perhaps they are both related to a 3rd variable

eg. Old Faithful data: i can't say longer duration causes there to be a longer interval — perhaps both related to some other geological phenomena

Multiple Regression: t-tests and the ANOVA F-test

$H_0: \beta_j = 0$ (all other X's in model)

= test for overall model

$H_0: \beta_1 = \cdots = \beta_p = 0$

It is possible to have

"significant" ⇒ small p-values

"n.s." = non significant ⇒ large p-value

F-test significant	All or some of t-tests significant	Model has useful variables for explaining Y
F-test n.s.	All t-tests n.s.	No explanatory variables useful for explaining Y
F-test significant	All t-tests n.s.	Multicollinearity

Individuals X's don't contribute to Y over and above other X's.

F-test n.s

Some t-test's significant

2 possible ways this can happen:

① model has no predictive value but got some Type I errors (falsely reject H_0) in the t-tests
Always a risk with multiple tests

② Predictors were chosen poorly
If one useful predictor among many

not useful predictors, the contribution from the useful one may not be enough for the F-test to be significant

Problems with Regression Model and Their Consequences

Wrong Model:

- Relationship is non-linear
  - biased $\hat{\beta}$'s
  - meaningless t & F-tests / CIs
- Omit important predictor variable
  - biased $\hat{\beta}$'s

- meaningless tests & CIs

Actual model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Fitted model: $Y = b_0 + b_1 X_1$

$E(b_1) \neq \beta_1$

- Include unimportant predictor variable
  - $\hat{\beta}$'s aren't biased
  - but variance unnecessarily large (inefficient)

Gauss-Markov Conditions

- Non-constant variance
  - $\hat{\beta}$'s not biased
  - s^2 is biased estimate

⇒ invalid tests / CIs

- Correlated errors
  - $\hat{\beta}$'s not biased
  - s^2 is not estimating right thing
    ⇒ invalid tests / CIs

Non-normal errors
- $\hat{\beta}$'s not biased
- s^2 unbiased est for σ^2
- but t / F-tests not valid except for large sample sizes and never for prediction intervals

Multicollinearity - consequence - high Var($\hat{\beta}$) (inefficient)

Predictor variables measured with error

Consequences - biased $\hat{\beta}$'s,
biased s^2,
invalid tests

OK if var in X's is small.

To fix: need multiple measurements
or Structural Equation Models