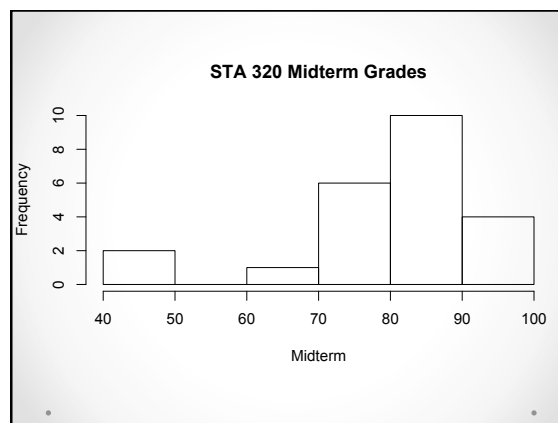


Comments on Midterm Comments on HW4 Final Project Regression Example Sensitivity Analysis? Quiz

STA 320
Design and Analysis of Causal Studies
Dr. Kari Lock Morgan and Dr. Fan Li
Department of Statistical Science
Duke University



Midterm Comments

- 2: why gain scores?
- 3: Probabilistic
- 4a: covariates must be pre-treatment
- 4b: matching best when $n_T \ll n_C$
- 4c: estimand, not estimate, ATT
- 5b: Estimation or test depends on question
- 5b: Neyman relies on CLT and large n
- 6c: If balance good, no more subclasses
- 6f: average of subclass estimates / original SE
- 7b: $\sqrt{\sum_i (1/5)^2 SE_i^2}$
- 7e: $Y_i^{obs} = \alpha + \beta' X_i + \hat{\tau} W_i + \varepsilon_i$

Hypothesis Test Conclusions (HW4)

- Generic decision:
 - in terms of null
 - Reject or do not reject null
- Interpret in context:
 - in terms of alternative
 - have or do not have evidence for alternative

Not statistically significant

- Do not reject H_0
- We do not reject the null that there is no effect of increasing the minimum wage on employment
- We do not have evidence that increasing the minimum wage effects employment

Statistically significant

- Reject H_0
- We reject the null that the job training program has no effect on wages
- We have evidence that the job training program increases wages

Final Project

- Details [here](#)
- Due in class on the last day of class, 4/23
- 30% of final grade
- Topic can be anything related to causal inference
- Good link for finding data: <http://guides.library.duke.edu/stat101>
- Maximum length: 10 pages
- Individual – no communication or help

Why Regression (without thinking) is dangerous?

Fan Li
March 26, 2014

A Toy Example

- Population: patients with heart attack.
- Treatment (W): 1 surgery; 0 medication
- Outcome (Y): prognosis after 1 month
- One covariate (X) severity: 1 severe; 0 mild-average.
- All other covariates are matched between groups.
- Sample: n patients admitted in a hospital.
- Goal: (1) estimate the effect of surgery comparing to medication; (2) predict the prognosis of a new patient.
- It happens to be: in the observed sample, all patients with X=1 get W=1, and all patients with X=0 get W=0.

Regression

- An idiosyncratic way is to write down the following regression model for the *observed data*:
$$Y \sim a + b \cdot W + c \cdot X$$
- For goal (1): fit the model to the sample, and the coefficient “b” is the “treatment effect”.
- For goal (2): for a new patient, plug in W and X to get a predicted Y.
- **Question: what if the new patient is with (W=0, X=1), or (W=1, X=0)?**
- What is odd here?

What is odd?

- There is no interaction $W \cdot X$ in the model.
- No interaction: effectively (implicitly) assuming the effect of W is additive.
- In fact, in all observed data, $W \cdot X = 0$ (complete imbalance in X between W groups). Therefore even if there is an interaction term, there is no information in the data to estimate the coefficient.
- What does the regression model here really assume?
- Clear if we adopt a potential outcome approach.

Potential Outcome Approach

- One regression for each potential outcome:
$$Y(0) \sim a_0 + c_0 \cdot X$$
$$Y(1) \sim a_1 + c_1 \cdot X$$
- Remember the “fancy” way of writing the observed outcome Y in terms of Y(0) and Y(1):
$$Y = Y(1) \cdot W + Y(0) \cdot (1 - W)$$
- Plug in the two potential outcome regressions to this form, we have a regression for Y:
$$Y = (a_1 + c_1 \cdot X) \cdot W + (a_0 + c_0 \cdot X) \cdot (1 - W)$$
$$= a_0 + (a_1 - a_0) \cdot W + c_0 \cdot X + (c_1 - c_0) \cdot X \cdot W$$
- Clearly the interaction should be there.
- Without interaction, one implicitly assumes $c_1 = c_0$.

- But by the implicit additivity (or equivalently homogenous effect or linear) assumption, the previous model can get a point prediction.
- This is not magic, but extrapolation based on an untestable assumption.
- Regression (or any model) comes with a package, you need to know and acknowledge what assumptions---explicit or implicit---come with the model.

Take home message

- Potential outcome approach forces you to look at the data, think hard and be honest about the assumptions. In this case: (1) overlap, (2) linearity (or additivity).
- Immediately resorting to OLS regression misses all of these.
- Causal inference is a way of thinking, not a particular model or procedure (like OLS regression).
- This doesn't say you can't use regressions for causal inference, in fact, they are routinely used. But the point of causality is free of what particular model you use, it is more about how to formulate a problem.