Stat 265 - Exam Solutions/Comments (Winter 2010)

1. **Basics of causal inference**

   (a) $Y(0)$ is the outcome that would be observed if a unit received the control treatment and $Y(1)$ is the outcome that would be observed if the same unit received the active treatment. These are known as the potential outcomes for the unit; both can not be observed for the same unit.

   (b) The causal effect of treatment for unit 1 is $Y_1(1) - Y_1(0) = 93 - 81 = 12$.

   (c) The SUTVA assumption has two components. First is that the potential outcomes for unit i do not depend on the treatment assigned to other units. Thus $Y_i(1)$ is the same regardless of what happens to other units. It is a form of independence but we don't always think of $Y_i(1)$ as a random variable so you need to be careful when discussing independence. The second aspect of the assumption concerns the requirement that there be only one form of the treatment so that $Y_i(1)$ is well defined and always means the same thing.

   (d) Note that the question did not ask for an "estimate". Here you can find the actual average treatment effect, $\tau = \frac{1}{6}\sum_i (Y_i(1) - Y_i(0)) = (12 + 11 + 5 + 8 + 9 + 7)/6 = 8.67$.

   (e) The most straightforward way to make the argument is to note that $Y_i^{obs} = W_i Y_i(1) + (1 - W_i) Y_i(0)$ which means that Y_i^{obs} is clearly dependent on W_i (unless $Y_i(1) = Y_i(0)$). Several people noted that the stronger the treatment effect is, the stronger is the dependence between W_i and Y_i^{obs}.

2. **Regression -** This turned out to be much trickier to grade than I would have imagined. Your grade may have changed multiple times as I tried to figure out a reasonable scale. One critical point is to avoid ambiguous terms like "the estimate" – you should specify which estimate (naive, regression, etc.) that you are talking about.

   (a) The aim of this question was to address why in a regression analysis it does not matter whether X is included or not. Many people indicated that since it is a randomized experiment X is balanced and we don't need to use it to estimate the treatment effect. This is true and a good answer although it doesn't connect back to the regression model that way that I hoped people would. Other people said that the OLS estimate of τ will be the same with or without X in the model. This is NOT TRUE as written. You can show that $\hat{\tau} = \bar{Y}_t - \bar{Y}_c - \hat{\beta}_1(\bar{X}_t - \bar{X}_c)$; it follows that the expected value of $\hat{\tau}$ (over randomizations) or the limit in large samples of $\hat{\tau}$ is unaffected by X (because $\bar{X}_t = \bar{X}_c$ on average or in the limit). It is not true in a single finite sample. If you talked about regression then you needed to make this point.

   (b) The short answer that I was looking for here is that regression CAN yield valid causal estimates if there is a regular assignment mechanism as long as the model is correct. The regular assignment mechanism means we can get valid causal inference by conditioning on X but it doesn't tell us how. Regression is one approach but makes fairly strong assumptions about the nature of the dependence on X. Matching and propensity score estimation make weaker assumptions about the nature of the dependence.

3. **Matching**

   (a) Note that the distance measure gives a penalty of 100 for two restaurants of different chains. This guarantees (given the number of initial employees is always less than 40) that we will only match two restaurants from the same chain. Then for observation 1 we find $D(1,4) = 4; D(1,5) = 2.5; D(1,6) = 9; D(1,7) = 10$ and observation 5 is the closest match. For observation 2 we find that $D(2,3) = 0.2; D(2,8) = 7.0$ and observation 3 is the closest match.

   (b) The natural matching estimate of the ATT is $\hat{\tau}_{match,att} = \frac{1}{2}((30 - 19.5) + (12.5 - 17.0)) = 3$. A few people used the gain in employment score $(Y_i^{obs} - X_{i2})$ in place of Y_i^{obs} in the above. This is a reasonable alternative to use here.

   (c) I believe you can make an argument for either ATT or ATE here. One might argue that since the treatment is only applied in New Jersey we only care about the effect on restaurants of the type that we would find there. On the other hand one might argue that restaurants in PA are just like those in NJ and thus we will get a more robust conclusion by including more restaurants. If there were a difference between the types of restaurants we would find in the two states, then it would not be appropriate to use the ATE. A couple of comments about your answers. Several people spoke specifically about what would happen with these 8 restaurants – e.g., ATE doesn't make sense because there would not be a good match for restaurant 7. That's a fine statement but these 8 were just a small sample of the full dataset. Also, several people made strong statements like we always want the ATT – I don't think that is true.

4. **Propensity scores** – This question is modified from one that appears in Gelman and Hill's Chapter 9. I like it very much.

(a) A randomized experiment means that W is independent of $Y(0)$ and $Y(1)$. This terminology allows a randomized experiment to have W depend on X (as long as the functional form is known). Most people use the term completely randomized if W is independent of X as well as the potential outcomes. One important point is that randomized does not have to mean $P(W = 1) = 0.5$; you can have a randomized experiment with uneven probability assignment. My question should really have asked if you believe the study was completely randomized. I was hoping people would note that when $X = 0$ the treatment is assigned to half the units (50/100) whereas when $X = 1$ the treatment is assigned to three-fourths of the units. This means treatment assignment is not completely randomized. You can do a similar comparison for $Y(0)$ (or $Y(1)$). When $Y(0) = 4$ the treatment is given to 60% of the units whereas when $Y(0) = 10$ the treatment is given to 65% of the units. Thus they are not independent (although one could wonder if this difference just occurred by chance).

(b) The treatment assignment is unconfounded given X if W is independent of $Y(0)$ and $Y(1)$ conditional on X.

(c) In this case (which is made up) you can show that the definition from (b) is satisfied. Let's restrict attention to the case when $X = 0$. Now if $Y(0) = 4$ the treatment is given to 30/60 subjects and if $Y(0) = 10$ the treatment is given to 20/40 subjects thus the treatment assignment process is independent of the level of $Y(0)$. Same is true for $Y(1)$ and when we consider $X = 1$ cases. You didn't need to show all of these but did need to show that you knew what to check and how to check it.

(d) The propensity score $e(X)$ is the average probability of receiving the treatment for units with covariate value X. In an observational study with regular treatment assignment this is equal to the probability of receiving treatment as a function of$X$.

(e) Back in (a) we noticed that $P(W = 1) = 0.5$ for $X = 0$ and $P(W = 1) = 0.75$ for $X = 1$. These are the propensity scores: $e(0) = 0.50$ and $e(1) = 0.75$.

(f) The very first thing to note here is that from the table we know what the answer is supposed to be. The average treatment effect is 2 because the treatment effect is 2 for every single person! There are two natural approaches. First is "subclassification" on the propensity score (equivalent to classifying on X). For the $X = 0$ cases we find the estimated treatment effect is $(30 * 6 + 20 * 12)/50 - (30 * 4 + 20 * 10)/50 = 8.4 - 6.4 = 2.0$ and for the $X = 1$ case we find the estimated treatment effect is $(30 * 6 + 45 * 12)/75 - (10 * 4 + 15 * 10)/25 = 9.6 - 7.6 = 2.0$. Then the estimaetd average treatment effect is the weighted average $\hat{\tau}_{subclass} = (100/200) * 2.0 + (100/200) * 2.0 = 2.0$. A second approach is weighting by the propensity score which yields $\hat{\tau}_{weight} = \sum_i W_i Y_i/e(X_i)/(\sum_i W_i/e(X_i)) - \sum_i (1 - W_i)Y_i/(1 - e(X_i))/(\sum_i (1 - W_i)/(1 - e(X_i)))$. The first term is $(30 * 6/.5 + 30 * 6/.75 + 20 * 12/.5 + 45 * 12/.75)/(30/.5 + 30/.75 + 20/.5 + 45/.75) = (360 + 240 + 480 + 720)/200 = 9.0$ The second term is $(30 * 4/.5 + 10 * 4/.25 + 20 * 10/.5 + 15 * 10/.25)/(30/.5 + 10/.25 + 20/.5 + 15/.25) = (240 + 160 + 400 + 600)/200 = 7.0$ which yields $\hat{\tau}_{weight} = 9.0 - 7.0 = 2.0$. The two most common mistakes in the weighting estimator were: (1) forgetting to include the number of people in each category and (2) forgetting to change to $(1 - e(X))$ in the second term.