**Questions class 3**

**Please download from CANVAS the file “jobtraining.csv”. The file contains information about a randomised experiment with workers in the US (back in the 80s). Workers were randomly assigned to a job training program and the earnings afterwards tracked (our outcome variable).**

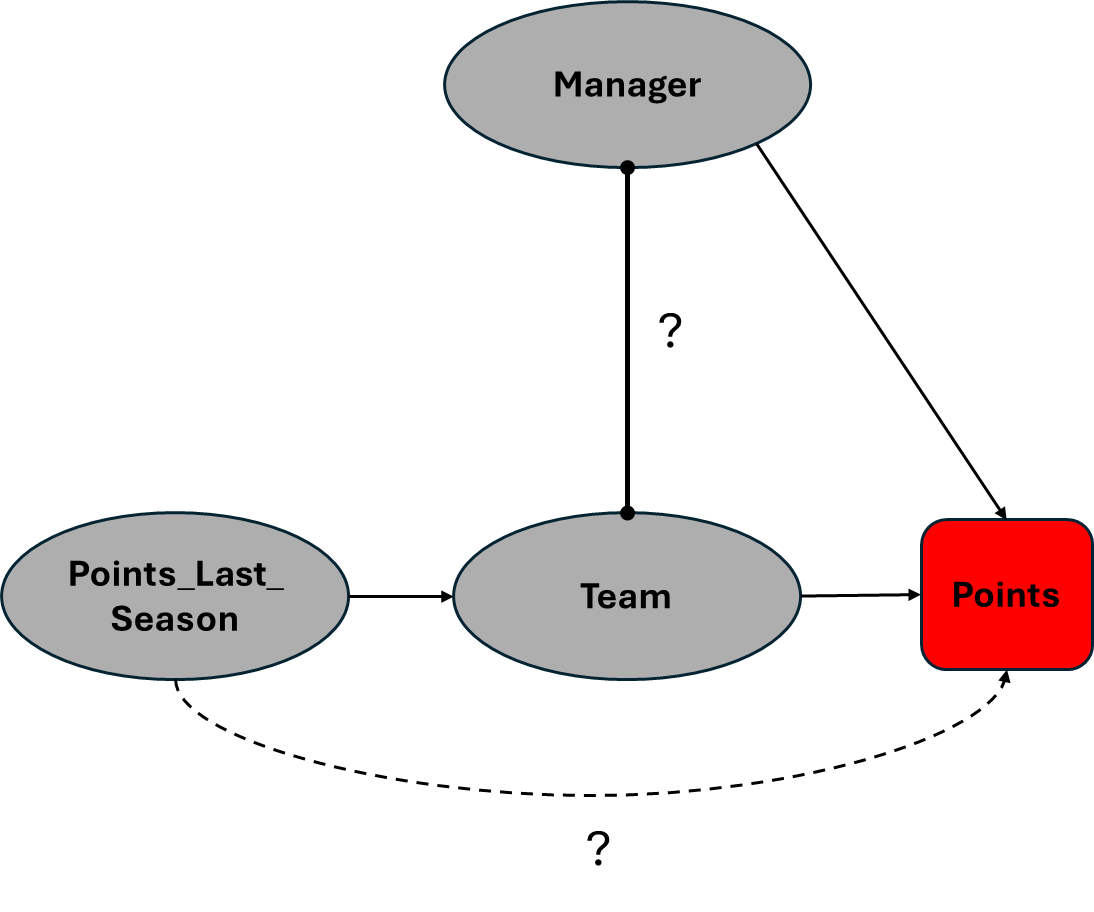
|  |  |
| --- | --- |
| **Variable name** | **Description** |
| earnings | Salary in thousands of USD the year after the possible job training |
| treatment | Dummy variable whether person underwent job training |
| age | measured in years |
| education | measured in years |
| married | indicating marital status (1 if married, 0 otherwise); |
| unemployment | Dummy variable whether someone was unemployed the year before (1=yes, 0 otherwise) |
| black | Dummy variable indicating race (1 if black, 0 otherwise) |
| hisp | Dummy variable indicating race (1 if Hispanic, 0 otherwise); |

1. **Let’s start with analysing the balance of the design and covariate balance.**
   1. **Check the frequency distribution of people in treatment and control groups.**
   2. **Check the frequency distribution across all treatment levels of all covariates (use the mean for continuous variables and frequency tables for discrete/ categorical covariates).**
   3. **Do you see potential issues with multi-collinearity for all covariates that are numeric (include any variables that are already dummy coded)?**
2. **Let’s analyse the impact of the treatment levels.**
   1. **Check whether you think the outcome variable should be log-transformed.**
   2. **Carry out a t-test. Are the means significantly different?**
   3. **Use linear regression and only include the treatment. What is the causal effect of job training (on earnings)?**
   4. **Use linear regression to examine the treatment effect and include all covariates as control variables. What is the causal effect of job training (on earnings) now?**
   5. **Rerun your regression from d) with log-transformed continuous input variables and examine using AIC whether this model better fits the data.**
   6. **Check whether you can rely on the standard errors of your best model (choosing between the specification from d and e).**
   7. **Compare the model from f with one where neither the outcome variable nor input variables are no log-transformed and check whether you can rely on the standard errors of this model.**
   8. **What is the effect of job training on earnings according to the model in g?**

**Please download from CANVAS the file “online\_store\_promo.csv”. The file contains information about a completely randomised experiment in which a marketing manager at an online fashion store analysed the effect of online promotions on sales**. **The products are randomly assigned to one of three variants of promotions and whether the promotion was mentioned in the newsletter. The outcome variable is the number of sold units during the promotion.**

1. **Let’s perform the following analyses.**
   1. **Double check the balance of all treatment levels and check whether you may want to log-transform the outcome variable.**
   2. **Plot boxplots for each individual treatment and their levels (2 boxplots, one for promotion and one for newsletter)**
   3. **Perform a one-way ANOVA on the impact of the promotion variant on sales.**
   4. **Perform a one-way ANOVA on the impact of newsletters on sales.**
   5. **Perform a two-way ANOVA on the impact of the main effects of newsletters and promotion on sales.**
   6. **Perform a two-way ANOVA on the impact of the main effects and two-way interaction of newsletters and promotion on sales (include a graph for the interaction).**
   7. **Perform pairwise comparisons for your ANOVA in f) using Tukey HSD.**
   8. **Run a linear regression for the same model in f).**
   9. **Investigate whether you can trust the standard errors of your model in h).**
2. **Let’s return to our football data set from the lecture in week 1. We want to revisit the question as to which football manager to hire (our treatment), focussing on the aggregate data per season and points as the outcome of interest (keep the raw points, no log transformation). We will again use all data (2007-2018), which you can access in the file**

**football\_managers\_all.csv. We use the same R code as from the lecture in week 1 to create the data we need, with the exception that we now use the R library *lubridate* to format date columns (an alternative to the as.Date command of base R). We again create a new variable *manager\_name* for the cases, where we have multiple managers per season. In addition, for our exercise today, we create a variable called ‘duration’ which captures the time a manager stayed with a club (based on the provided dates). Reminder: keep all variables when generating the aggregate data set.**

****

* 1. **Now let’s first consider the diagram above. Let’s test whether there is indeed no direct link from ‘Points last season’ to ‘Points’ and adjust the diagram if necessary. What manager to hire according to the model based on the diagram after your analysis (based on the potentially adjusted model)?**
  2. **What happens with ‘Points last season’ if we remove ‘Team’ from the model and only include ‘Points last season’ as control variable?**
  3. **We currently have a barbell link (indicating just some association with unknown causal direction) between variables Team and Manager. Let’s improve this. Which way do you think is the causal effect? Consider what the Team fixed effects (which capture information across all years) measure versus the variable ‘points last season’. Do we need to add another link to the diagram if we believe a manager is influenced by the performance of a team in the past?**

**Note: Just reason theoretically (no analysis needed) and update the causal diagram accordingly.**

* 1. **Use both a two-way ANOVA (with Team and Manager as factors) and one-way ANOVA (only factor is Manager) to test whether there is any possible association between the variable Manager and outcome variable Points (NOTE: since this data is not from an experiment, we refer to associations for these ANOVAs). Which ANOVA would be more appropriate to answer your question of any potential total impact of managers on Points? If necessary, update the causal diagram based on the findings of your ANOVAs (remove the link if you find no association).**
  2. **Now let’s examine whether our new variable Duration fully mediates the effect of Manager on Team and has indeed no direct effect on points. We need 3 steps: first, test whether there is an association between Manager and Duration (managers resulting in different durations at team). Second, test whether Duration may have an association with points above and beyond the Team fixed effects. Thirdly, check what happens if you omit Team (is Duration now affecting points in addition to Manager?). Update the causal diagram accordingly.**
  3. **Based on your final updated diagram, which effect has a stronger relative impact on points, ‘points last season’ or ‘duration’? *Tip: consider how we can compare two numeric variables in a fair way.***
  4. **Which variables would you include in your final analysis to determine which manager to hire? Check the robustness of your inference for this final model (normality and homoscedasticity of your residuals). Which manager would you hire according to this analysis?**
  5. **Given your final model from g), how much does the impact of Manager on Points matter relative to other variables and what variance does Manager explain?**

**While we have not covered this in the lecture, note that we can include a numeric variable to an ANOVA as a kind of control variable too (just like in a linear regression). This is then called ANCOVA (C for covariate).**

**So just run the ANOVA command for your final model from g) to estimate ANCOVA. After obtaining your ANCOVA result, you just need to consider how you obtain eta squared for the variance and how you can compare the effect of your factor (Manager) versus other variables in a fair (standardised way). *Tip: Think of a standardised effect measure we covered at the end of the lecture (you can use the R library effectsize)***

* 1. **Before making a final recommendation to your club about which manager to hire, do you think there are omitted confounders? Discuss carefully**