**Questions week 2**

**Please download from CANVAS the file “firemen.csv”. The file contains information about fire fighters in the US city of Springfield.**

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| damage | Price of property damage in $ (2018 adjusted) – our outcome variable |
| attended | How many firemen attend a fire |
| dispatch | How many firefighters were dispatched |
| grade | The severity of a fire (from 1-5, 5 being most serious) – this is the assessment of the on-call officer at emergency offices after asking questions (how long the fire has been burning, how large it is, etc). |
| Trucks | How many trucks drove to the fire |

**Now imagine the mayor of Springfield is considering budget changes for the next year and proposed cutting funding for the fire department of the city. His reasoning is that he investigated the role of firefighters on reducing property damage and apparently found that firefighters did not help reduce property damage (of course, he did not mention that firemen also save lives). Now you were hired as an analyst by the fire department to examine whether it’s true that firemen did not help prevent property damage.**

1. **Let’s start with some exploratory analysis and check all correlations of numeric variables.**
   1. **Which input variable has the greatest/ weakest correlation with damage?**
   2. **Do you see potential issues with multi-collinearity?**
2. **Now perform a linear regression in which you only include *attended* as input variable and and *damage* as outcome variable (without any transformations).** 
   1. **What is the coefficient sign and p-value of this variable?**
   2. **Are the findings on your coefficients as expected (do they make theoretically sense in your view)? Why or why not?**
3. **Because of your findings from B, you next perform a linear regression in which you only include ‘attended’ and ‘grade’ as input variables and damage as outcome variable.** 
   1. **What is the coefficient of *attended* if we include *grade* as a numeric variable.**
   2. **What is the adjusted R2 /AIC for this model?**
   3. **What is the coefficient of *attended* if we include *grade* as a factor variable.**
   4. **What is the adjusted R2 / AIC for this model?**
   5. **Are the findings on your coefficients as expected (do they make theoretically sense in your view)? Why or why not?**

**Given your findings so far, you wonder what is going on. Could including *trucks* and *dispatch* further help your analysis? Given the high correlations with each other (and *attended*), your initial thought was that these variables capture the same information (dispatched firefighters drive in trucks to the location to attend the fire). However, you request further information on the variables and process from the manager of fire services. The manager explains the following:**

***“On the basis of this grading and other factors, the dispatcher sends out a certain number of trucks and a certain number of firefighters. Obviously more trucks typically means more firefighters but it is not a matter of just putting 5 people on each truck; for some fires there might be extra trucks required because the fire needs to be attacked from several directions, but for others feet on the ground might be more important.”***

***“Regarding differences between dispatch and attended numbers, people are on-call and supposed to be available but sometimes they are ill or for some reason can’t come. They are allowed to decline a certain number per year. Even if they reach their limit, we only dock them their on-call fee for that day. It’s like they take a day of annual leave and only get charged for it if we call them. So, it’s not unusual for several to not attend.”***

1. **You decide to draw a causal diagram with all the variables. Check the diagram below.** 
   1. **Regarding the arrow connecting *trucks* and *damage*, what direction should it be?**
   2. **Regarding the arrow connecting *grade* and *damage*, what direction should it be?**
   3. **Complete the causal diagram, adding the arrows between trucks and damage, and grade and damage.**

**A diagram of a truck

Description automatically generated with medium confidence**

1. **Write down all theoretical paths and state which variables you need to include in your regression model to close all backdoors if you were interested** **in the total causal effect of ‘attended’ on ‘damage’.**
2. **Perform a regression model and add all control variables (****we basically assume now that all observed variables are not colliders or complete mediators and may have a direct effect on the outcome).**
   1. **Are the findings on your coefficients as expected (do they make sense)?**
   2. **Would you have expected this result? What could be the reason?**
   3. **Update the causal diagram purely based on the new findings of this regression (tip: you need to change the links of two input variables to ‘damage’).**
   4. **Based on your revised causal diagram in c), what is the total causal effect of ‘dispatch’ on ‘damage’?**
   5. **Based on your revised causal diagram in c),** **calculate the omitted variable bias for the causal effect of attended on damage if you don’t have ‘*dispatch’* in your model.**
3. **You source data on the size of buildings from the fire events in your database. You then create a rating from 1 to 5 to capture the size, with 5 being the largest building in your classification. Include ‘building size’ along with all other variables (including trucks) in your analysis examining the causal drivers of property damage of fires *(tip: you need to decide how to best include ‘building size’ according to fit*). Are the findings on your coefficients as expected (do they make theoretically sense in your view)? Why or why not?**
4. **Examine the relative contribution of all effects property damage in our multiple regression model with all input variables (without transformations or interactions). *Tip: use your code from class 1 for this question. Treat all variables as numeric for this exercise (include any dummies, if there are any).***
   1. **Provide a bar chart comparing the coefficients for each variable.**
   2. **Which of the all the input variables is the most influential one in term of relative impact (in absolute terms)?**
5. **Since you cannot access additional data, you decide to further test different functional forms for your model.** 
   1. **Rerun your regression from F with log-transformed outcome variable damage. Does anything change in terms of your coefficients sign and significance?**
   2. **Independently from your reasoning in a, would you say your outcome variable should be log-transformed?**
   3. **Following your decision on the best model in section b), use that model and include an interaction for grade and dispatch. Do the results make sense?**
   4. **Following your decision on the best model in section b), log-transform continuous input variables. Do the results improve in terms of fit or sensitivity of findings?**
   5. **Pick a model from c and d, do some final check whether you can trust the model’s standard errors and report the effects of *attended* and *trucks* on damage.**
6. **How much trust do you place in your estimated causal effects? What suggestions do you have to improve your analysis (provide examples)?**