Question 1

a. **Describe the smoking patterns in the dataset based on the variable *perday* (number of cigarettes smoked per day). What is the mean, minimum and maximum number of cigarettes smoked per day by the individuals in the dataset? Draw a histogram of the variable *perday.* How do you account for the "spikes" of the distribution? Based on your initial statistics, do you have any evidence of overdispersion for this variable?**

The mean is 14.07 cigarettes per day
The maximum is 60 cigarettes per day
The minimum is 1 cigarette per day

```
. sum perday, detail

                Number of sticks smoked per day.
-------------------------------------------------------------
      Percentiles      Smallest
 1%            2              1
 5%            3              1
10%            4              2       Obs                 779
25%            7              2       Sum of wgt.         779

50%           10                      Mean           14.07317
                      Largest       Std. dev.      9.773977
75%           20             60
90%           20             60       Variance       95.53063
95%           30             60       Skewness       1.557834
99%           42             60       Kurtosis       6.617315

. histogram perday
(bin=27, start=1, width=2.1851852)

.
```

There seems to be evidence for overdispersion as the variance seems to be fairly high – a log transformation could help with the overdispersion.

b. **Many believe that heavy smoking is more prevalent among men than women. What is the observed difference in the number of cigarettes smoked per day (*perday*) among men and women? Do you find statistical support for the aforementioned assumption?**

c. **Estimate a count model for the number of cigarettes smoked per day (*perday*) explained by gender (*female*), whether individuals are older than 45 years old (*age45*), whether individuals are so addicted that they admit to need help to quit smoking (*so_addicted*), and whether they declare that they want to quit smoking sometime in their life (*want_quit*). What type of count data model do you prefer to fit to this data? Is it a satisfactory model for this data? Explain why and provide a brief interpretation of the results.**

d. **As an alternative model, estimate a linear regression for the number of cigarettes smoked per day in logarithms (*lnperday*), using *female, age45, so_addicted* and *want_quit* as regressors. Are these regressors able to explain a significant part of smoking variations in this data? Interpret the coefficient of *want_quit* in this regression.**
e. **Experts on the topic consider if women and men differ on average in terms of their ability to quit smoking. Specifically, their claim is that among individuals who are motivated for quitting, men are less effective in actually implementing this in their smoking practices. To investigate this, extend your model from question d) with an interaction term between *female* and *want_quit.* Produce a plot where you illustrate the differences in the predicted number of cigarettes smoked per day for the four groups (i.e. motivated men, motivated women, non-motivated men, non-motivated women). Do you find support for the experts' claim?**

Question 2

A reviewer for your analysis of smoking patterns argues that *want_quit* may be endogenous since it may correlate with many unobserved characteristics of the respondent. In particular, the reviewer argues that self-discipline or general optimism about the future could be potential sources of bias. To address their concerns empirically, we will continue with the linear model specification from Question 1d).

a. **Explain how this concern could be a problem for a causal interpretation of the estimated coefficient of *want_quit* in your regression from Question 1d).**
b. **b) The reviewer suggests to use *avoid* (whether the respondent avoids situations that make him or her want to smoke) and *quit_6mos* (whether the respondent specifically wishes to quit within 6 months) as instruments for *want_quit.* Follow the reviewer's suggestion and implement this IV estimation. What is the new estimate for the coefficient of *want_quit* in this case?**
c. **c) Evaluate the choice of these instruments. Are they relevant and valid instruments? Provide arguments and explicit tests to support your conclusions.**

Question 3

a. **What do we learn about the potential effectiveness of this intervention from this comparison? What could be the potential concerns here? Explain**

b. **Run a logit of takeup on female, age, age2, perday, smellsmoke, want_quit for individuals who were offered the treatment (cares=1). Calculate the "marginal" effect of want_quit on takeup. Specify which type of marginal effect you are calculating here. Comment on the significance of the predictors of takeup.**

c. **Using propensity score matching, estimate the ATET of the CARES-deposit treatment (takeup) on the probability to pass the final urine test (passedtest). Use the following set of variables to estimate the propensity score of taking up the test: female, age, age2, perday, smellsmoke, want_quit. What do you conclude?**

d. **Overlaping**

e. **Balancing**

f. **Do you trust the estimate of the CARES-deposit treatment effect from the PSM as an estimate of the causal effect? What could be the potential concerns here? Explain.**

Question 4

a. **What is the percentage of the individuals offered the CARES program (cares = 1) who have missing values for passedtest because they did not take the test?**

b. **Why could these missing values be a concern in this case? What implications could this have in the estimation of the effect of takeup on passedtest? If there is a potential bias, can you say something about the sign of the potential bias in this case? Provide a cross-tabulation from the data to back up your conclusion**

c. **Some individuals gave their phone number in the initial interview. The reviewer suggests this as an exclusion restriction to be included in the estimation of a Heckman model. We provide the output of this model below. What is the rationale behind using phone_nr as an exclusion restriction here? Based on the output provided, are the missing values for those who did not take the final test a concern in this case? Justify.**