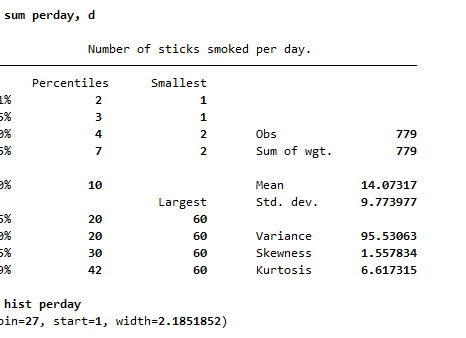
a)

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?

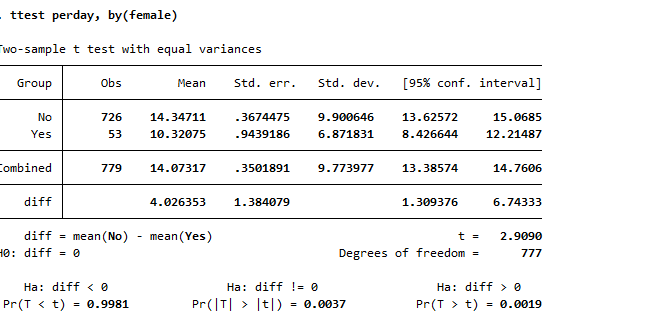




Variable seems to be overdispersed: The variance (95.5) exceeds the mean (14.1), which is around 7 times higher.

From the histogram, we can see that there are spikes in the data between 10 and 20. This means that most of the people smoke between 10 and 20 cigarettes per day.

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?



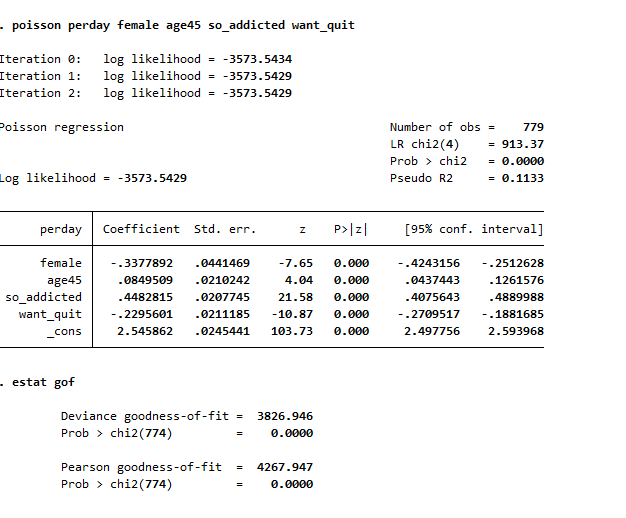
Men, on average, appear to smoke approximately 14 cigarettes per day, while women, on average smoke 10 cigarettes. The difference is statistically significant on the univariate level, which means that we find statistical support for our assumption.

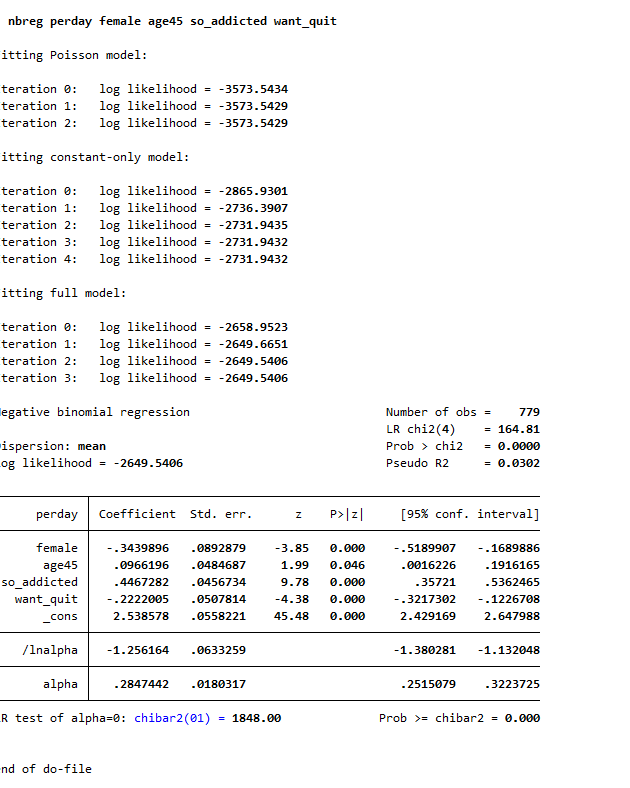
c) Estimate a count model for the number of cigarettes smoked per day (*perday*) explained by gender (*female*), whether individuals are older than 45 years old (*age45*), whether individuals are so addicted that they admit to need help to quit smoking (*so\_addicted*), and whether they declare that 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.

We would prefer to use the negative binomial model since it takes into account the over-dispersion (variance exceeding the mean). By comparison, the Poisson model would not be a good fit if our model is over-dispersed.

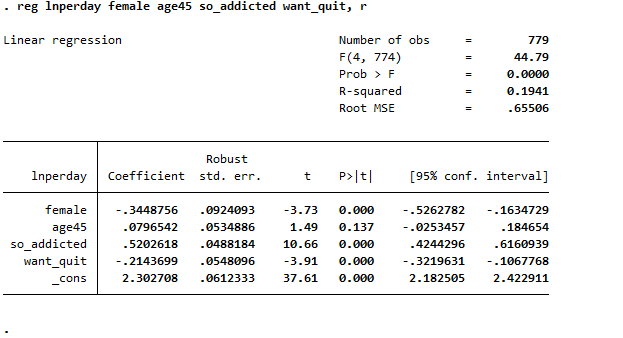
Indeed, the test suggests that our data does not fit the model very well.

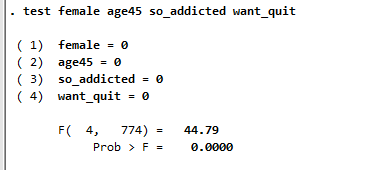
In the negative binomial, we see that alpha is greater than zero, which suggests that over-dispersion is indeed a problem.





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.



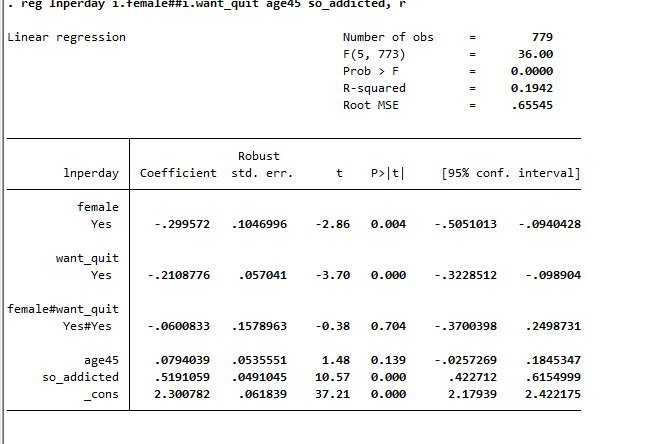


The r-squared is 0.19, which suggests that around a fith of the variation is explained by our model. The f-test suggests that regressors are jointly significant determinants of the numbers of cigarettes smoked per day.

Want-quit is a binary variable that equals 1 if a certain individual wants to quit smoking at a certain point in time. The estimated coefficient is -0.214, suggesting that they smoke (exp(-.2143699)-1)\*100= -19.295 percentage less.

Qe)

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?





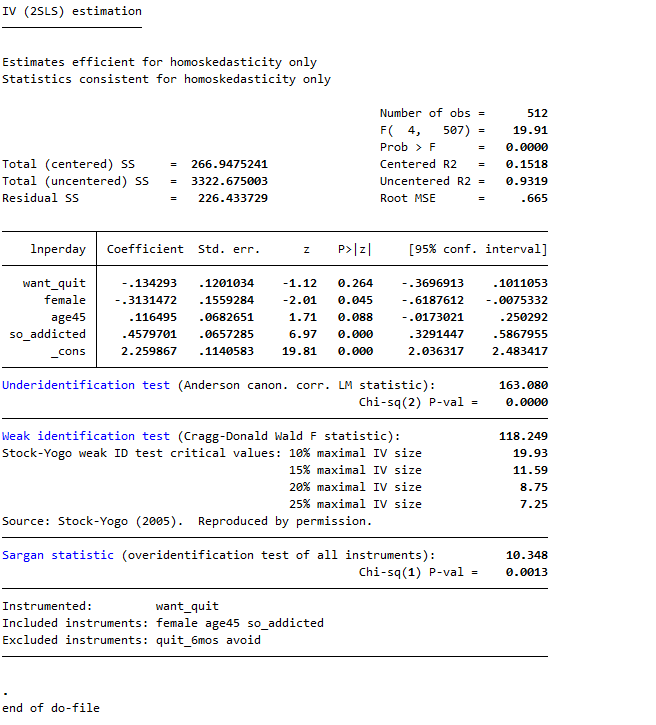
The MEs suggests that on average female clients, smoke 0.34 less cigarettes per day. We can see that the interaction effect is insignificant in the main model and therefore the claim made by the experts is not supported.

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).

If there are unobserved factors, correlated with our error term and our X’s, then the OLS estimator of 𝛽1 (and 𝛽2) is biased. Optimism could yield an upwards bias in the OLS estimate.

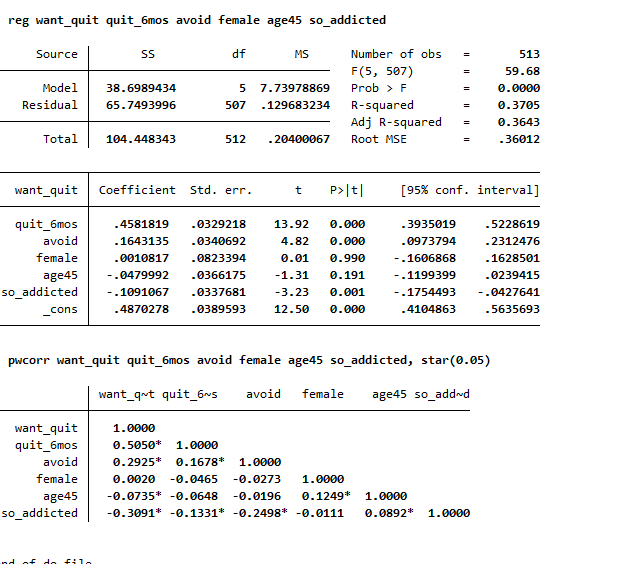
b) The reviewer suggests to use *avoid* (whether the respondent avoids situations that make him or her want to smoke) and *quit\_6mos* (whether the respondent specifically wishes to quit within 6 months) as instruments for *want\_quit*. Follow the reviewer’s suggestion and implement this IV estimation. What is the new estimate for the coefficient of *want\_quit* in this case?

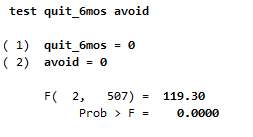


The new coefficient estimate is -.134293. The estimate is now, however, statistically insignificant.

**c) Evaluate the choice of these instruments. Are they relevant and valid instruments? Provide arguments and explicit tests to support your conclusions**.

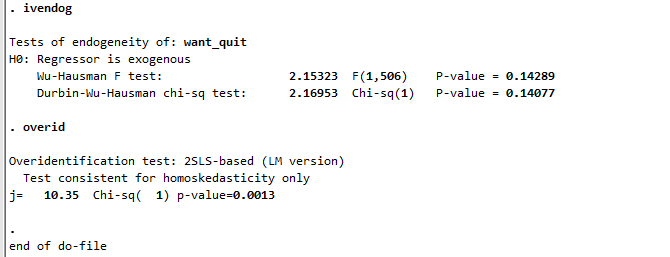
Relevance





The relevance condition of our instruments seems to be fulfilled. Pairwise correlations are significant. The F-test suggests that our instruments are jointly significant in the reduced form equation.

Validity

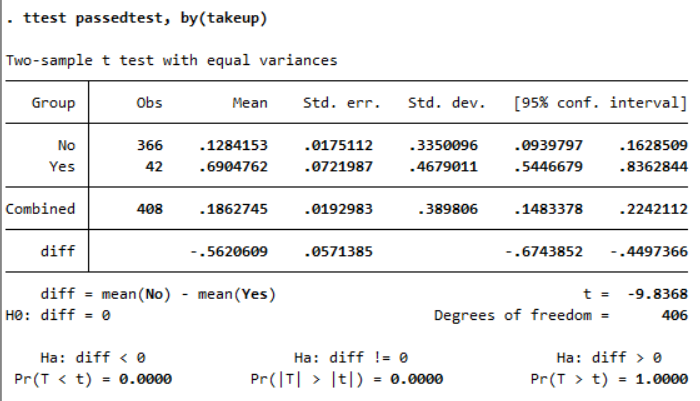


However, when running the over-identification test, we can see that at least one of the instruments is not valid.

Note: when using only one instrument the IV is uncorrelated with u. This suggests that we could drop one of the variables and obtain a valid model.

**Question 4**

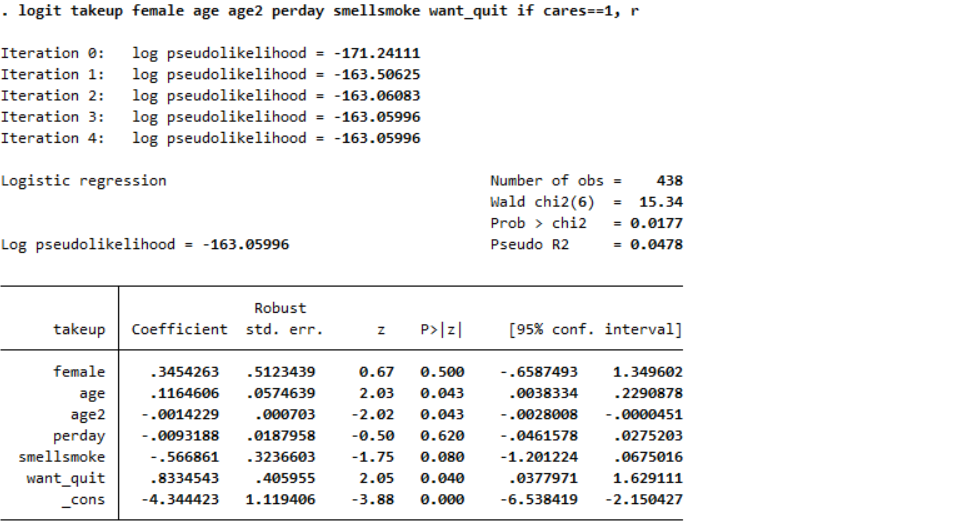
1. What do we learn about the potential effectiveness of this intervention from this comparison? What could be the potential concerns here? Explain

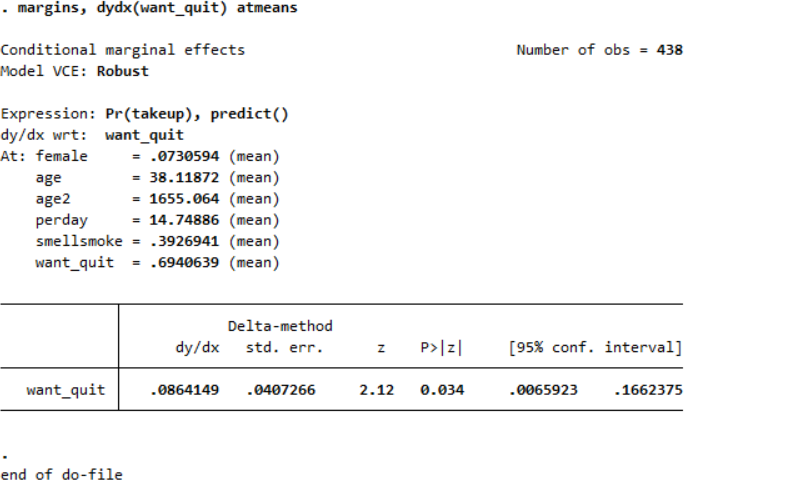


We can see that 13% of those who did not participate in CARES passed the test, while 69% of those participated in CARES did pass.

Potential concerns could be that participants self-selected themselves into treatment (only 14% of those offered actually agreed to part-take). They might be fundamentally different with respect to X variables.

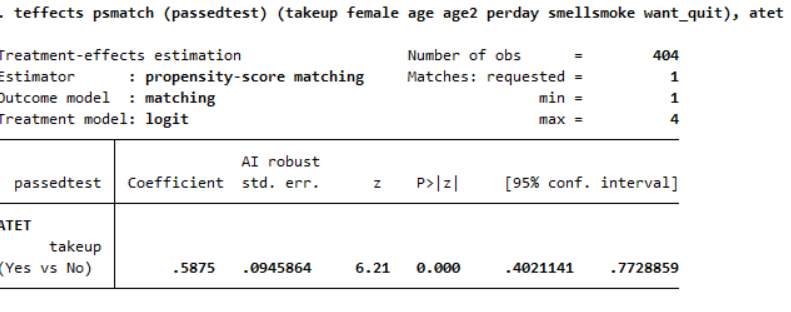
1. Run a logit of takeup on female, age, age2, perday, smellsmoke, want\_quit for individuals who were offered the treatment (cares=1). Calculate the “marginal” effect of want\_quit on takeup. Specify which type of marginal effect you are calculating here. Comment on the significance of the predictors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





Most predictors are significant in determining “Takeup” (except gender and perday). We evaluate the marginal effect of want-quit on takeup at its mean. Want-quitters appear to be significantly more likely to part-take in the CARE program.

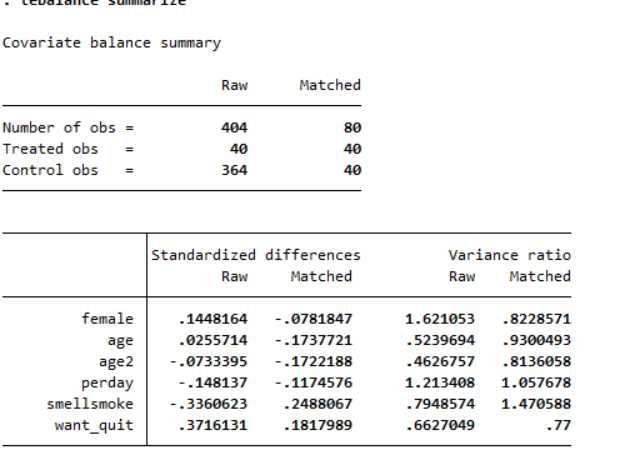
1. Using propensity score matching, estimate the ATET of the CARES-deposit treatment (takeup) on the probability to pass the final urine test (passedtest). Use the following set of variables to estimate the propensity score of taking up the test: female, age, age2, perday, smellsmoke, want\_quit. What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



The ATET is significant and positive, suggesting that the participation in the CARE program increases the chances of passing the urine test.

d)

Balancing condition



Improved level of balance for some variables (e.g. perday, want\_quit), however, the quality of the PSM is actually worse for our age variables (compared with the raw data). Ideally, the covariate would have a standardized difference of 0 and variance of 1. This is not case with our data and therefore the balancing condition is violated.

Overlap condition



The overlap assumption is violated since the density of the graphs is not concentrated in regions where they overlap.

**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.**

Given that our balancing and overlap conditions are violated, we cannot draw valid inferences on the CARES-deposit effect. Hence, we cannot trust our estimate.

**Question 4**

1. 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?

7.69% have missing values 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

Y is only observed for individuals taking the test, not all do. Potential non-random (i.e., selected) sample. Decision to take test is likely related to unobserved factors that also influence takeup.

OLS assumes random sampling. If the variables are indeed missing by random there won’t be a problem. However, if they are not missing at random then the estimate could be biased.

It is unclear what sign of the potential bias is. It could be that those people who did not take the test knew that they were going to fail, leading to an over-estimation of our OLS estimate. But it is also possible that those who did not take the test, knew they did not smoke leading to an underestimate.

The cross-tabulations do not make clear in which direction the bias would be going, since those who took test and those who did not have similar underlying characteristics.

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.

Strong exclusion restrictions are necessary for two-step heckman models. The reason for using phone-number could be that those who gave out their mobile numbers, were the ones more likely to quit smoking. In principle, Phone\_NR should be a significant predictor in the selection equation (which is the case). At the same time, Phone\_NR should be insignificant in the main outcome equation (which we would have to test).

The rho gives the correlation between the error terms of the 2 equations. Insignificant rho suggests that unobserved factors that make individuals more likely to take the test, do not significantly affect the main outcome (passed test). Therefore, selection bias does not appear to be a concern in the estimation and OLS would be a consistent estimator.