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?

```
sum perday, d
```

Number of sticks smoked per day.

	Percentiles	Smallest		
L%	2	1		
5%	3	1		
0%	4	2	Obs	779
5%	7	2	Sum of wgt.	779
0%	10		Mean	14.07317
		Largest	Std. dev.	9.773977
5%	20	60		
0%	20	60	Variance	95.53063
5%	30	60	Skewness	1.557834
0%	42	60	Kurtosis	6.617315

```
hist perday
bin=27, start=1, width=2.1851852)
```

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?

```
. ttest perday, by(female)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
No	726	14.34711	.3674475	9.900646	13.62572	15.0685
Yes	53	10.32075	.9439186	6.871831	8.426644	12.21487
Combined	779	14.07317	.3501891	9.773977	13.38574	14.7606
diff		4.026353	1.384079		1.309376	6.74333

```
    diff = mean(No) - mean(Yes)                                   t =   2.9090
H0: diff = 0                                     Degrees of freedom =      777

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9981         Pr(|T| > |t|) = 0.0037          Pr(T > t) = 0.0019
```

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.

```
. poisson perday female age45 so_addicted want_quit

Iteration 0:   log likelihood = -3573.5434
Iteration 1:   log likelihood = -3573.5429
Iteration 2:   log likelihood = -3573.5429

Poisson regression                                      Number of obs =    779
                                                        LR chi2(4)    = 913.37
                                                        Prob > chi2   = 0.0000
Log likelihood = -3573.5429                             Pseudo R2     = 0.1133
```

perday	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
female	-.3377892	.0441469	-7.65	0.000	-.4243156	-.2512628
age45	.0849509	.0210242	4.04	0.000	.0437443	.1261576
so_addicted	.4482815	.0207745	21.58	0.000	.4075643	.4889988
want_quit	-.2295601	.0211185	-10.87	0.000	-.2709517	-.1881685
_cons	2.545862	.0245441	103.73	0.000	2.497756	2.593968

```
. estat gof

        Deviance goodness-of-fit =  3826.946
        Prob > chi2(774)         =    0.0000

        Pearson goodness-of-fit  =  4267.947
        Prob > chi2(774)         =    0.0000
```

```
 nbreg perday female age45 so_addicted want_quit

itting Poisson model:

teration 0:   log likelihood = -3573.5434
teration 1:   log likelihood = -3573.5429
teration 2:   log likelihood = -3573.5429

itting constant-only model:

teration 0:   log likelihood = -2865.9301
teration 1:   log likelihood = -2736.3907
teration 2:   log likelihood = -2731.9435
teration 3:   log likelihood = -2731.9432
teration 4:   log likelihood = -2731.9432

itting full model:

teration 0:   log likelihood = -2658.9523
teration 1:   log likelihood = -2649.6651
teration 2:   log likelihood = -2649.5406
teration 3:   log likelihood = -2649.5406

egative binomial regression                             Number of obs =    779
                                                        LR chi2(4)    = 164.81
ispersion: mean                                         Prob > chi2   = 0.0000
og likelihood = -2649.5406                              Pseudo R2     = 0.0302
```

perday	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
female	-.3439896	.0892879	-3.85	0.000	-.5189907	-.1689886
age45	.0966196	.0484687	1.99	0.046	.0016226	.1916165
so_addicted	.4467282	.0456734	9.78	0.000	.35721	.5362465
want_quit	-.2222005	.0507814	-4.38	0.000	-.3217302	-.1226708
_cons	2.538578	.0558221	45.48	0.000	2.429169	2.647988
/lnalpha	-1.256164	.0633259			-1.380281	-1.132048
alpha	.2847442	.0180317			.2515079	.3223725

```
R test of alpha=0: chibar2(01) = 1848.00                 Prob >= chibar2 = 0.000


nd of do-file
```

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.

```
. reg lnperday female age45 so_addicted want_quit, r

Linear regression                               Number of obs     =        779
                                                F(4, 774)         =      44.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1941
                                                Root MSE          =     .65506

------------------------------------------------------------------------------
             |               Robust
    lnperday |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3448756   .0924093    -3.73   0.000    -.5262782   -.1634729
       age45 |   .0796542   .0534886     1.49   0.137    -.0253457    .184654
  so_addicted |   .5202618   .0488184    10.66   0.000     .4244296    .6160939
   want_quit |  -.2143699   .0548096    -3.91   0.000    -.3219631   -.1067768
       _cons |   2.302708   .0612333    37.61   0.000     2.182505    2.422911
------------------------------------------------------------------------------
```

.

```
. test female age45 so_addicted want_quit

 ( 1)  female = 0
 ( 2)  age45 = 0
 ( 3)  so_addicted = 0
 ( 4)  want_quit = 0

       F(  4,   774) =   44.79
            Prob > F =    0.0000
```

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?

```
. reg lnperday 1.female##1.want_quit age45 so_addicted, r

Linear regression                               Number of obs     =        779
                                                F(5, 773)         =      36.00
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1942
                                                Root MSE          =     .65545
```

lnperday	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
female						
Yes	-.299572	.1046996	-2.86	0.004	-.5051013	-.0940428
want_quit						
Yes	-.2108776	.057041	-3.70	0.000	-.3228512	-.098904
female#want_quit						
Yes#Yes	-.0600833	.1578963	-0.38	0.704	-.3700398	.2498731
age45	.0794039	.0535551	1.48	0.139	-.0257269	.1845347
so_addicted	.5191059	.0491045	10.57	0.000	.422712	.6154999
_cons	2.300782	.061839	37.21	0.000	2.17939	2.422175

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 $\beta 1$ (and $\beta 2$) 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?

```
IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =       512
                                                      F(  4,   507) =     19.91
                                                      Prob > F      =    0.0000
Total (centered) SS     =  266.9475241                Centered R2   =    0.1518
Total (uncentered) SS   =  3322.675003                Uncentered R2 =    0.9319
Residual SS             =   226.433729                Root MSE      =      .665

------------------------------------------------------------------------------
     lnperday |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
    want_quit |   -.134293   .1201034    -1.12   0.264    -.3696913    .1011053
       female |  -.3131472   .1559284    -2.01   0.045    -.6187612   -.0075332
        age45 |    .116495   .0682651     1.71   0.088    -.0173021    .250292
  so_addicted |   .4579701   .0657285     6.97   0.000     .3291447    .5867955
        _cons |   2.259867   .1140583    19.81   0.000     2.036317    2.483417
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):         163.080
                                                      Chi-sq(2) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):              118.249
Stock-Yogo weak ID test critical values: 10% maximal IV size             19.93
                                         15% maximal IV size             11.59
                                         20% maximal IV size              8.75
                                         25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):          10.348
                                                      Chi-sq(1) P-val =    0.0013
------------------------------------------------------------------------------
Instrumented:         want_quit
Included instruments: female age45 so_addicted
Excluded instruments: quit_6mos avoid
------------------------------------------------------------------------------

.
end of do-file
```

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

```
reg want_quit quit_6mos avoid female age45 so_addicted
```

Source	SS	df	MS	Number of obs	=	513
				F(5, 507)	=	59.68
Model	38.6989434	5	7.73978869	Prob > F	=	0.0000
Residual	65.7493996	507	.129683234	R-squared	=	0.3705
				Adj R-squared	=	0.3643
Total	104.448343	512	.20400067	Root MSE	=	.36012

want_quit	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
quit_6mos	.4581819	.0329218	13.92	0.000	.3935019	.5228619
avoid	.1643135	.0340692	4.82	0.000	.0973794	.2312476
female	.0010817	.0823394	0.01	0.990	-.1606868	.1628501
age45	-.0479992	.0366175	-1.31	0.191	-.1199399	.0239415
so_addicted	-.1091067	.0337681	-3.23	0.001	-.1754493	-.0427641
_cons	.4870278	.0389593	12.50	0.000	.4104863	.5635693

```
pwcorr want_quit quit_6mos avoid female age45 so_addicted, star(0.05)
```

	want_q~t	quit_6~s	avoid	female	age45	so_add~d
want_quit	1.0000					
quit_6mos	0.5050*	1.0000				
avoid	0.2925*	0.1678*	1.0000			
female	0.0020	-0.0465	-0.0273	1.0000		
age45	-0.0735*	-0.0648	-0.0196	0.1249*	1.0000	
so_addicted	-0.3091*	-0.1331*	-0.2498*	-0.0111	0.0892*	1.0000

```
end of do-file
. test quit_6mos avoid

( 1)  quit_6mos = 0
( 2)  avoid = 0

       F(  2,   507) =  119.30
            Prob > F =    0.0000
```

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

```
. ivendog

Tests of endogeneity of: want_quit
H0: Regressor is exogenous
    Wu-Hausman F test:                 2.15323  F(1,506)    P-value = 0.14289
    Durbin-Wu-Hausman chi-sq test:     2.16953  Chi-sq(1)   P-value = 0.14077

. overid

Overidentification test: 2SLS-based (LM version)
  Test consistent for homoskedasticity only
j=   10.35  Chi-sq(  1) p-value=0.0013

.
end of do-file
```

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

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

```
. ttest passedtest, by(takeup)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
No	366	.1284153	.0175112	.3350096	.0939797	.1628509
Yes	42	.6904762	.0721987	.4679011	.5446679	.8362844
Combined	408	.1862745	.0192983	.389806	.1483378	.2242112
diff		-.5620609	.0571385		-.6743852	-.4497366

```
    diff = mean(No) - mean(Yes)                                   t =  -9.8368
H0: diff = 0                                     Degrees of freedom =      406

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

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.

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.

```
. logit takeup female age age2 perday smellsmoke want_quit if cares==1, r

Iteration 0:   log pseudolikelihood = -171.24111
Iteration 1:   log pseudolikelihood = -163.50625
Iteration 2:   log pseudolikelihood = -163.06083
Iteration 3:   log pseudolikelihood = -163.05996
Iteration 4:   log pseudolikelihood = -163.05996

Logistic regression                                     Number of obs =    438
                                                        Wald chi2(6)  =  15.34
                                                        Prob > chi2   = 0.0177
Log pseudolikelihood = -163.05996                       Pseudo R2     = 0.0478
```

takeup	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
female	.3454263	.5123439	0.67	0.500	-.6587493	1.349602
age	.1164606	.0574639	2.03	0.043	.0038334	.2290878
age2	-.0014229	.000703	-2.02	0.043	-.0028008	-.0000451
perday	-.0093188	.0187958	-0.50	0.620	-.0461578	.0275203
smellsmoke	-.566861	.3236603	-1.75	0.080	-1.201224	.0675016
want_quit	.8334543	.405955	2.05	0.040	.0377971	1.629111
_cons	-4.344423	1.119406	-3.88	0.000	-6.538419	-2.150427

```
. margins, dydx(want_quit) atmeans

Conditional marginal effects                                    Number of obs = 438
Model VCE: Robust

Expression: Pr(takeup), predict()
dy/dx wrt:  want_quit
At: female     = .0730594 (mean)
    age        = 38.11872 (mean)
    age2       = 1655.064 (mean)
    perday     = 14.74886 (mean)
    smellsmoke = .3926941 (mean)
    want_quit  = .6940639 (mean)
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
want_quit	.0864149	.0407266	2.12	0.034	.0065923	.1662375

```
.
end of do-file
```

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.

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?

```
. teffects psmatch (passedtest) (takeup female age age2 perday smellsmoke want_quit), atet

Treatment-effects estimation                   Number of obs      =        404
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Treatment model: logit                                     max =          4
```

passedtest	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATET						
takeup (Yes vs No)	.5875	.0945864	6.21	0.000	.4021141	.7728859

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

. tebalance summarize

Covariate balance summary

	Raw	Matched
Number of obs =	404	80
Treated obs =	40	40
Control obs =	364	40

	Standardized differences		Variance ratio	
	Raw	Matched	Raw	Matched
female	.1448164	-.0781847	1.621053	.8228571
age	.0255714	-.1737721	.5239694	.9300493
age2	-.0733395	-.1722188	.4626757	.8136058
perday	-.148137	-.1174576	1.213408	1.057678
smellsmoke	-.3360623	.2488067	.7948574	1.470588
want_quit	.3716131	.1817989	.6627049	.77

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

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?

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.