**Exercise 8**

The file **JTRAIN2.dta** contains data on a job training experiment for a group of men. Men could enter the program starting in January 1976 through about mid-1977. The program ended in December 1977. The idea is to test whether participation in the job training program had an effect on unemployment probabilities and earnings in 1978.

1. The variable *train* is the job training indicator. How many men in the sample participated in the job training program? What was the highest number of months a man actually participated in the program?

**smpl train –restrict**

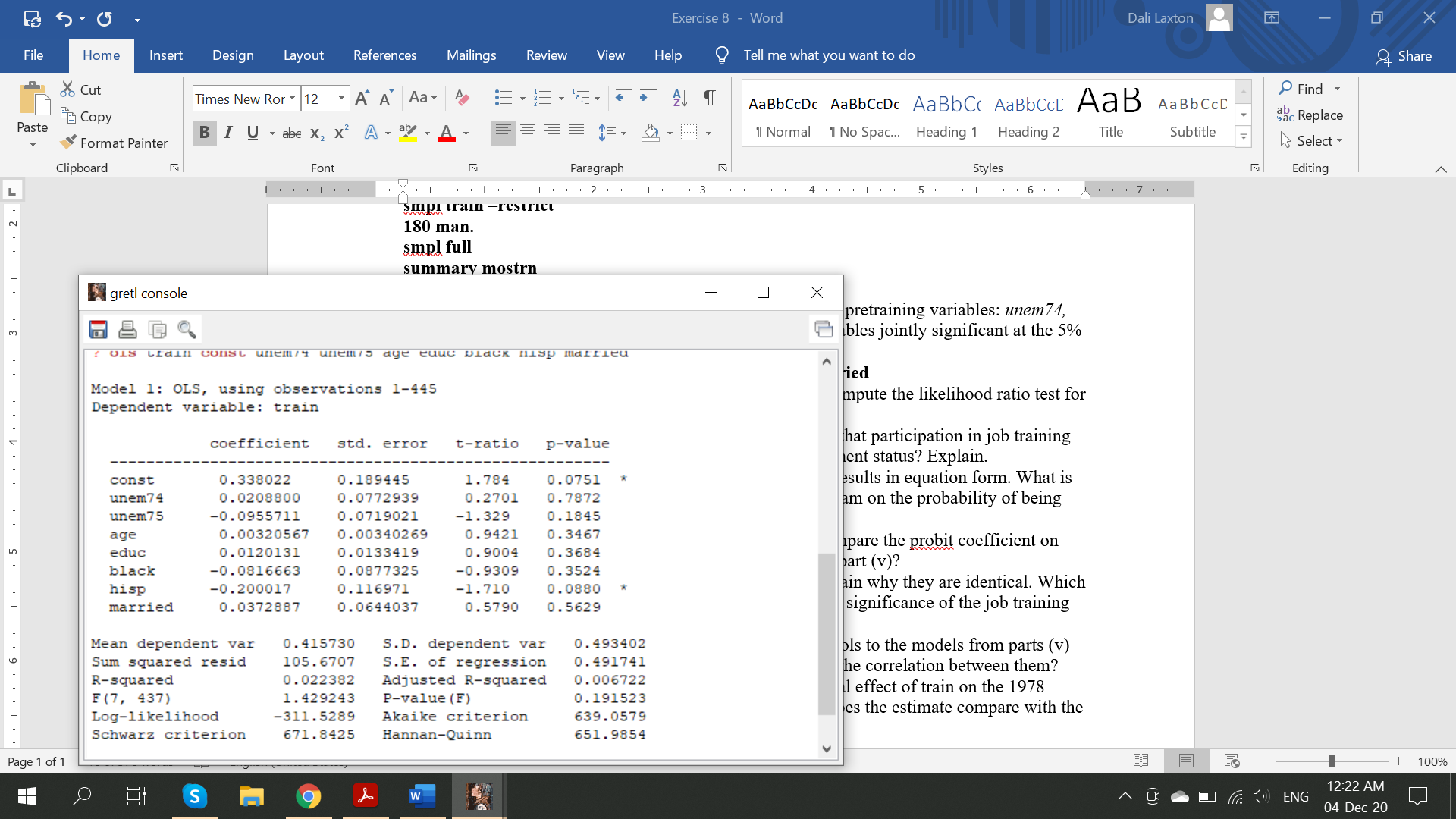
**smpl full**

**summary mostrn**

**185 out of 445 participated in the job training program. The longest time in the experiment was 24 months.**

1. Run a linear regression of *train* on several demographic and pretraining variables: *unem74, unem75, age, educ, black, hisp,* and *married*. Are these variables jointly significant at the 5% level?

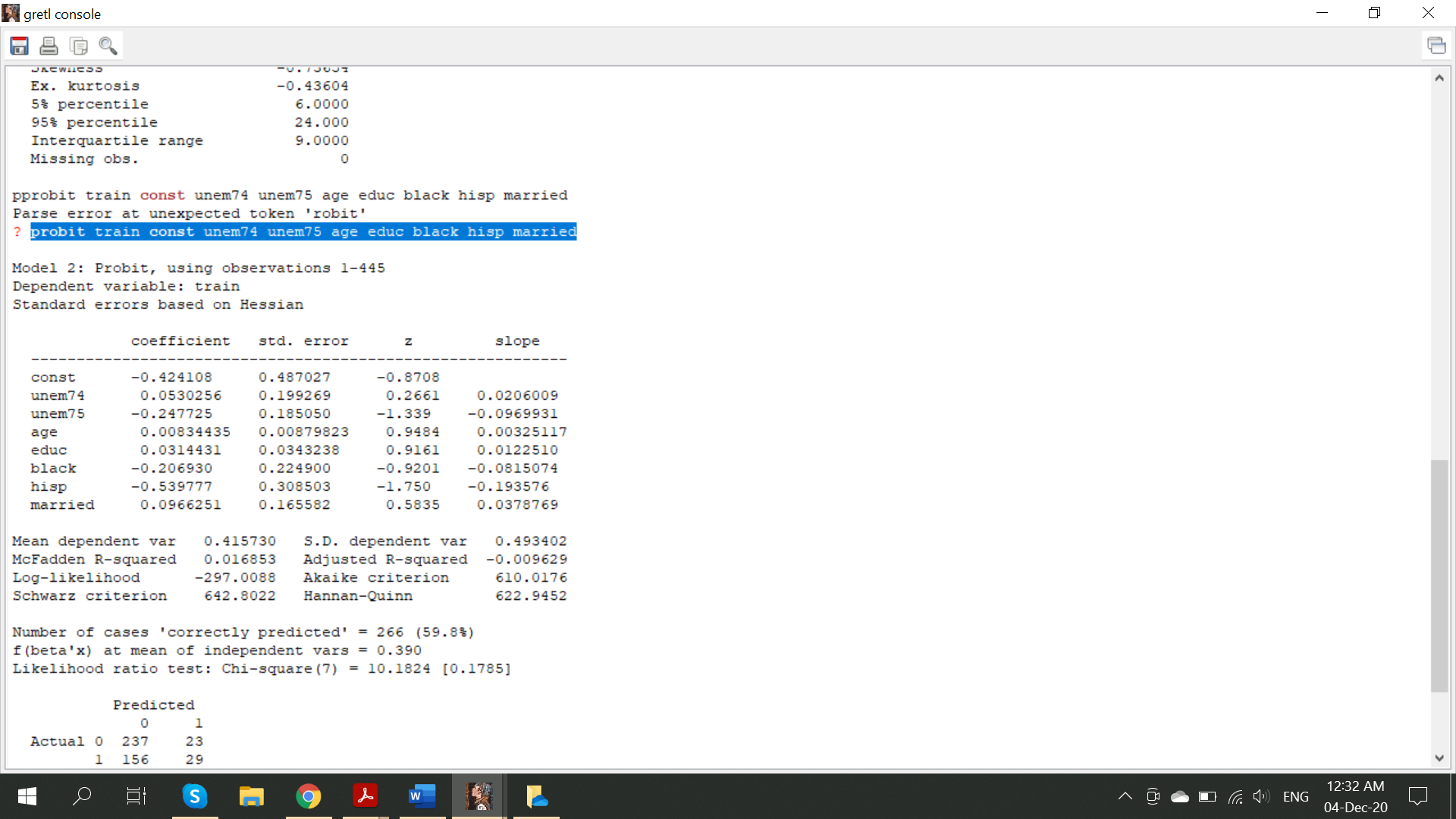
**ols train const unem74 unem75 age educ black hisp married**



**The *F* statistic for joint significance of the explanatory variables is *F*(7,437) = 1.43 with *p*-value = .19. Therefore, they are jointly insignificant at even the 15% level. Note that, even though we have estimated a linear probability model, the null hypothesis we are testing is that all slope coefficients are zero, and so there is no heteroskedasticity under H0. This means that the usual *F* statistic is asymptotically valid**

1. Estimate a probit version of the linear model in part (ii). Compute the likelihood ratio test for joint significance of all variables. What do you conclude?

**probit train const unem74 unem75 age educ black hisp married**



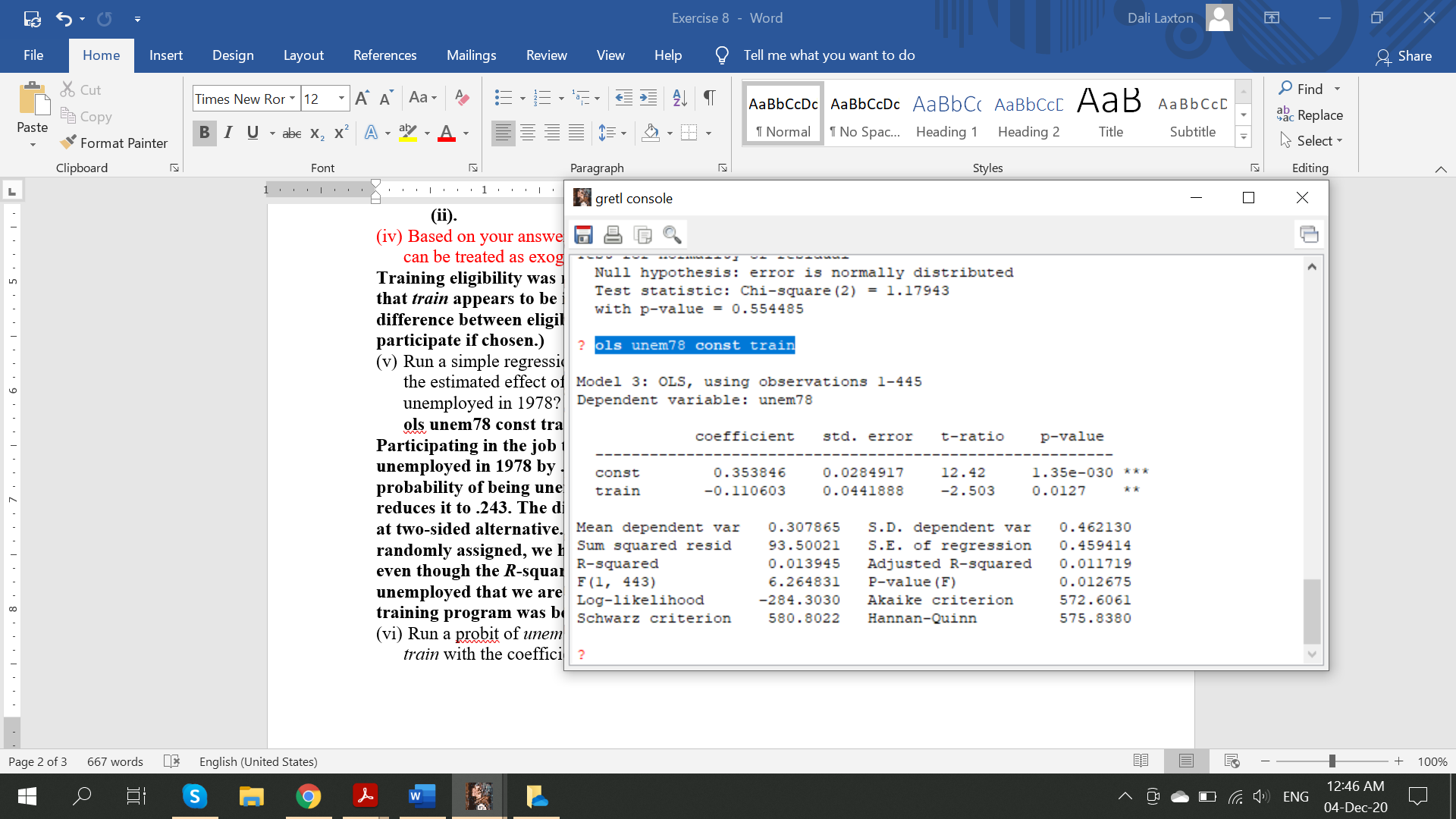
**After estimating the model P(train=1|X)= Φ( by probit maximum likelihood, the likelihood ratio test for joint significance is 10.18. In a distribution this gives P-value =0.18, which is very similar to that obtained in the LPM in part (ii).**

1. Based on your answers to parts (ii) and (iii), does it appear that participation in job training can be treated as exogenous for explaining 1978 unemployment status? Explain.

**Training eligibility was randomly assigned among the participants, so it is not surprising that *train* appears to be independent of other observed factors. Therefore, running a regression of train on unem78 would not suffer with the endogeneity issue (However, there can be a difference between eligibility and actual participation, as men can always refuse to participate if chosen.)**

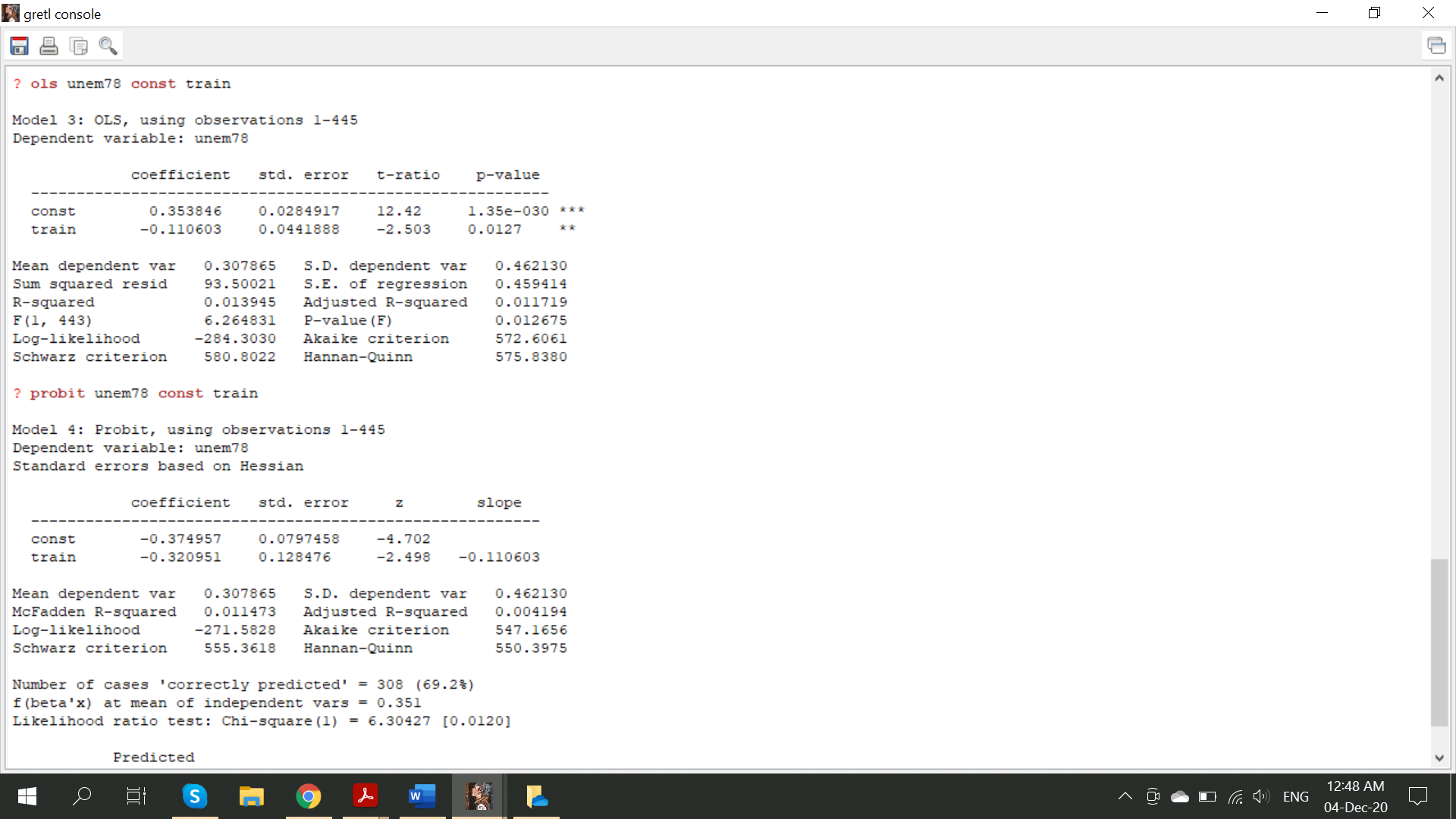
1. Run a simple regression of *unem78* on *train* and report the results in equation form. What is the estimated effect of participating in the job training program on the probability of being unemployed in 1978? Is it statistically significant?

**ols unem78 const train**



**Participating in the job training program lowers the estimated probability of being unemployed in 1978 by .111, or 11.1 percentage points. This is a large effect: the probability of being unemployed without participation is .354, and the training program reduces it to .243. The differences is statistically significant at almost the 1% level against at two-sided alternative. (Note that this is another case where, because training was randomly assigned, we have confidence that OLS is consistently estimating a causal effect, even though the *R*-squared from the regression is very small. There is much about being unemployed that we are not explaining, but we can be pretty confident that this job training program was beneficial.)**

1. Run a probit of *unem78* on *train*. Does it make sense to compare the probit coefficient on *train* with the coefficient obtained from the linear model in part (v)?



**It does not make sense to compare the coefficient on *train* for the probit, −.321, with the LPM estimate. The probabilities have different functional forms. However, note that the probit and LPM *t* statistics are essentially the same (although the LPM standard errors should be made robust to heteroskedasticity).**

1. Find the fitted probabilities from parts (v) and (vi). Explain why they are identical. Which approach would you use to measure the effect and statistical significance of the job training program?

**There are only two fitted values in each case, and they are the same: .354 when *train* =**

**0 and .243 when *train* = 1. This has to be the case, because any method simply delivers the cell frequencies as the estimated probabilities. The LPM estimates are easier to interpret because they do not involve the transformation by Φ(⋅), but it does not matter which is used provided the probability differences are calculated.**

1. Add all of the variables from part (ii) as additional controls to the models from parts (v) and (vi). Are the fitted probabilities now identical? What is the correlation between them?

**ols unem78 const train unem74 unem75 age educ black hisp married**

**series yhat=$yhat**

**probit unem78 const train unem74 unem75 age educ black hisp married**

**series yhat2=$yhat**

**corr yhat yhat2**

**The fitted values are no longer identical because the model is not saturated, that is, the explanatory variables are not an exhaustive, mutually exclusive set of dummy variables. But, because the other explanatory variables are insignificant, the fitted values are highly correlated: the LPM and probit fitted values have a correlation of about .993**