EC 587 HW2

Erik Andersen

October 2023

Question 1

a)

Below are the coefficient estimates, and standard errors for age of the client and treatment status without robust standard errors. Since we are assuming homogeneous effects in this model, the partial effects are just the two coefficients. We can thus interpret the effect of age on taking up a loan as being a year older causes the percent someone takes a new lone by by about 0.05%. The effect of being treated is about 4%. Again because this is a linear probability model, the effects are constant for all values of x and treatment.

Variable	Coefficient	SE
Client_Age	0.00055941	0.00188300
Treated	0.04509724	0.03369204

b)

Below are the results from the same regression, but with heteroskedasticity robust standard errors. The coefficients, and interpretations are the same as in part a. The standard errors are smaller for with the robust standard errors.

Variable	Coefficient	SE
Client_Age	0.00055941	0.00183202
Treated	0.04509724	0.03240097

c)

The maximum predicted value is 0.32, and the minimum is 0.035. None are outside of the [0,1] interval.

d)

Below are the new coefficient estimates and standard errors from the regression estimated using variance weighted least squares.

Variable	Coefficient	SE
Client_Age	0.00146514	0.00178386
Treated	0.04473244	0.03263838

The estimates for both coefficients and standard errors are very close. The treated coefficient and standard errors are almost identical with the point estimate slightly lower. The coefficient for age is larger by a more significant amount, but still none of the coefficients are significant, so this is probably just noise.

e)

Because we are using a LPM, the effects are constant across x and treatment status, so the average partial effect is just the coefficient on the variable of interest. For parts a and b, the average partial effect for being treated is 0.045; for part d it is 0.044. Likewise the average partial effect for increasing age is 0.00056 in a an b; it is 0.001 in d.

f)

The coefficients on age, treatment, and the interaction term are shown below.

Variable	Coefficient	SE
Client_Age	-0.00081737	0.00228733
Treated	0.04475369	0.03368982
Client_Age*Muslim	0.00417254	0.00393633

The average partial effect is now the weighted sum of the coefficient on age times the probability that Muslim is 0 plus the probability they're not Muslim times the sum of the age coefficient and the interaction coefficient.

$$APE = P(Muslim = 0) * \beta_{age} + P(Muslim = 1) * (\beta_{age} + \beta_{interact})$$

Calculating this from the values in the data set gives us the average partial effect is about 0.0004; a very similar value to the previous questions.

Question 2

a)

Below are the coefficient estimates for age across the LPM, logit, and probit models.

LPM	Logit	Probit
0.00055941	0.0004059	0.0002679

The values are fairly close. They are all the same order of magnitude. This is a bit of a coincidence. Unlike the LPM model, the coefficients are not the average marginal average partial effects. We will calculate the partial effects in the next few parts, which we'll see have pretty much the same magnitude as in the LMP.

b)

Below are the correlations between the predictions between all the models. The predictions are almost perfectly correlated.

	LPM	Logit	Probit
LPM	1	0.9915219	0.9932232
Logit	0.9915219	1	0.9988347
Probit	0.9932232	0.9988347	1

c)

We will calculate the partial effect of age on loan take up at the mean of the data. The closed form formula is $\frac{\partial g(\bar{X}\hat{\beta})}{\partial X} = g'(\bar{X}\hat{\beta})\hat{\beta}$ where $g()$ is the pdf of a normal distribution. The partial effect is of age on loan take up is 0.000424746 which is again very similar to all the previous effects we've found.

d)

Calculating the mean partial derivative of using the margins command gives us the value 0.0006637 which is again similar and the same order of magnitude.

e)

Now we'll calculate the mean of the partial effects. The formula is almost the same as part c, but instead of $\bar{X}$, we use the formula for each X then sum over all the effects. The effect is 0.00003 which is much smaller than the previous answers.

f)

Calculating the partial effect manually gives us the value 0.00009. Which is similar to the effect from part e.

g)

Below are the effects from each calculation method.

Part	Partial Effect
c)	0.000424746
d)	0.0006637
e)	0.00003
f)	0.00009

The estimates are qualitatively similar. Parts e and f are about an order of magnitude smaller, but none of the estimates are significant, so this is not a huge difference. We expect the calculation from part be a biased estimator of the effect, but the others we expect to be close to the true value, so they should all roughly agree.

h)

We calculated the average partial effect in part 1a as 0.0005, which is very close to the estimates from each of the above estimates. The results should be similar because they are all calculating the same partial effect.

Question 3

a)

The correct prediction percentage when using 0.5 as the cutoff is 0.8324421. This is quite high, but note that the maximum fitted value of the model was about 0.35, so no one was rounded to 1, so the model predicts that no one takes up a loan. This correct prediction percentage is thus just the percentage in the sample who didn't take up a loan. If we change the cutoff to the mean of the y variable, the correct prediction percentage drops to 0.5187166. This is much worse. So changing to the mean of the y variable is worse than the null estimator.

b)

Using the probit model, we get similar results. Using 0.5 as the cutoff, we get the same prediction rate of 0.8324421 which is again just the null estimator of prediction no one take a loan. Using the other cutoff, the prediction rate improves to 0.543672, which is still not great.

c)

Restricting the data set has only a small effect on the prediction rate. Using 0.5 as the cutoff we get the same result. Using the other cutoff, we improve very very slightly to 0.5464286, which is not a qualitative improvement.

Question 4

a)

Regressing the covariates on the on the residuals from the LPM model gives us no significant coefficients. This is expected. This regression tells us that the error term is conditionally mean independent from the covariates. Since this was a randomized study that should be true.

b)

Regressing the covariates on the squared residuals again gives us no significant coefficients. This is unexpected. This regression is telling us the variance of the errors is conditionally independent from all the covariates. This should not be true in a LPM which should always have heteroskedasticity. I think this is an effect of the small sample size of only $\tilde{5}00$ observations. We may just not have enough power to pick up the heteroskedasticity.

c)

The hetglm command spits out a Breusch-Pagan test result which tests for heteroskedasticity. The null hypothesis is that there is homoskedasticity. The p-value for the BP test is 0.3991. So we cannot reject the null of homoskedasticity.