Problem Set #1

Danny Edgel
Econ 717: Applied Econometrics
Spring 2022

February 15, 2022

The attached file, edgel_ps1_log.log, includes all console output for this problem set. The coefficients and standard errors of all requested regressions are reported in table 1 below. The label at the top of each column corresponds to the question for which the regression was run (e.g., "Q2" corresponds to the regression from question 2). For Question 6, which requests two regressions, the columns are labeled with the name of the specification.

Table 1: Regression Results

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Q2	Q3	Q5	Logit	Probit
Treated	0.0444	0.0444	0.0520	0.334	0.184
	(0.0337)	(0.0324)	(0.00356)	(0.252)	(0.138)
Client_Age	0.000440	0.000440	-0.000657	0.00328	0.00215
	(0.00188)	(0.00184)	(0.000240)	(0.0138)	(0.00763)
Client_Married	0.0299	0.0299	0.0371	0.233	0.133
	(0.0482)	(0.0456)	(0.00708)	(0.366)	(0.200)
Client_Education	-0.00347	-0.00347	-0.00134	-0.0248	-0.0133
	(0.00380)	(0.00374)	(0.000441)	(0.0274)	(0.0153)
HH_Income	3.36e-06	3.36e-06	2.66e-06	2.30e-05	1.32e-05
	(3.57e-06)	(3.72e-06)	(4.80e-07)	(2.43e-05)	(1.39e-05)
muslim	-0.0184	-0.0184	-0.0371	-0.133	-0.0713
	(0.0357)	(0.0352)	(0.00377)	(0.259)	(0.144)
Hindu_SC_Kat	-0.0289	-0.0289	-0.0577	-0.216	-0.115
	(0.0505)	(0.0492)	(0.00672)	(0.375)	(0.206)
Constant	0.106	0.106	0.133	-2.084	-1.249
	(0.0866)	(0.0758)	(0.0116)	(0.653)	(0.364)

Standard errors in parentheses

The coefficients for the linear probability model (LPM) with OLS standard errors, displayed in column one, suggest that the treatment is not a statistically significant determinant of loan take-up, nor is the coefficient of any other covariate. However, there are many issues with this specification. Client age and

education, for example, should be modeled either as saturated variables or in buckets, as it is highly unlikely that there is a single, constant marginal effect for one year of age or education, and in the data, neither of these variables are continuous.[1]

In comparing the first two specifications (for questions two and three), we are obviously only interested in the standard errors, as the coefficients are exactly equal. Surprisingly, the robust standard errors are smaller for every coefficient other than the household income coefficient. This is evidence against meaningful heteroskedasticity in the data generating process (DGP).

Table 2 displays summary statistics for the predicted probabilities of every model used in this assignment. As you can see, all predicted probabilities for the LPM lie between 0 and 1. In face, the summary statistics of the (non-quartic) LPM align closely with those of the probit model.

Table 2: Predicted Probabilities

	LPM	Quartic LPM	Logit	Probit
Mean	0.168	0.168	0.168	0.168
Std. Dev.	0.033	0.052	0.033	0.033
Min	0.077	0.001	0.091	0.088
p5	0.112	0.090	0.117	0.115
p25	0.145	0.138	0.143	0.144
Median	0.170	0.170	0.167	0.168
p75	0.190	0.195	0.189	0.190
p95	0.219	0.234	0.224	0.223
Max	0.287	0.848	0.313	0.312

Column three of Table 2 displays the results of the variance-weighted least squares regression.[2] By and large, the coefficients for this specification do not meaninfully differ from those of the OLS LPM, with the notable exception of the coefficient for client age, which changes sign. However, the standard errors differ substantially, across all covariates. Namely, they get much smaller, resulting in every coefficient being statistically significant.

Columns four and five of Table 1 display the results of the logit and probit regressions, respectively. The coefficient estimates for these specifications are nothing like the coefficients for the LPM specifications, nor should they be. The coefficients of an LPM are estimates of a constant marginal effect of the covariate on the probability of the dependent variable. The coefficient of a generalized linear model (GLM) such as logit or probit represents the constant marginal effect of the covariate on the model's latent scoring variable.[3]

[1] Granted, one could argue that age and education are continuous in the data generating process.

[2] The weights used in the variance-weighted least squares (VWLS) are the standard error of the OLS estimation.

[3] In the case of the logit model, a variable's coefficient measures the (constant) marginal effect of that variable on the log odds ratio of the dependent variable being equal to one, relative to being equal to 0.

Table 3 displays the mean partial effect of client age on the probability of loan uptake in each of the four (including the quartic LPM) specifications. In the LPM, this represents the increase in probability of loan uptake caused by a one-year increase in client age. For the GLM models, this represents the average observed increase in the probability of loan uptake caused by a one-year increase in client age.

Table 3: Mean Partial Effects

	LPM	Logit	Probit	Quartic LPM
Mean Partial Effect	0.000440	0.000454	-	-
7a)	-	-	0.000535	-
7b)	-	-	0.000534	-
7c)	-	-	0.000570	0.000560
7d)	-	-	0.000534	-

The mean partial derivative estimates for LPM and logit are very similar, but the probit estimates are much higher. The four methods of calculating the probit estimates are very similar, though the numerical estimate is a bit higher than the rest. The numeric estimate of the mean partial derivative for the quartic LPM is similar to the estimate for the probit model, though, as Table 2 shows, the predicted probabilities of the quartic LPM model are much more wildly distributed than those of the other three models.

Calculating the LRI for the probit model by hand yields a value of 0.00863 . In practice, this is generally viewed as as a measure of what share of the variance in loan uptake probability is explained by the model. Technically speaking, this value is simply the ratio of the log-likelihood of a probit model with just an intercept to that of our probit model, subtracted from one. Since this value is bounded by zero, at the bottom, at one, at the top (asymptotically), this very low value suggests that our model does a poor job of explaining the variation in loan uptake.

Table 4 displays the correct prediction rate for for each of the models, using both $\hat{p} = 0.168$ (the sample take-up rate) and 0.5 as the threshold for predicting loan take-up. The first two columns display the rates for each model using all observations to estimate the model, whereas the last two columns display the rates for the model that was estimates with only $imidlineid < 1400$.

Table 4: Prediction Rates

	In-Sample		Out-of-Sample	
	≥ 0.5	$\geq \hat{p}$	≥ 0.5	$\geq \hat{p}$
LPM	0.832	0.519	0.832	0.496
Quartic LPM	0.834	0.504	0.832	0.529
Logit	0.832	0.542	0.832	0.517
Probit	0.832	0.533	0.832	0.515

As the table shows, prediction rates vary little across the four models, particularly using 0.5 as a threshold, and the out-of-sample prediction rate is similar to the in-sample rate for each model. The slight decrease in prediction rates for

out-of-sample predictions is consistent with expectations. The model is much more accurate if you use 0.5 as a threshold for determining predictability.[4] The method of using sample take-up rates to determine prediction accuracy is better for assessing phenomena that occur at a rate that is very different from 50%. I personally prefer this measure, though think that, in settings such as this with subpopulations that have clear and observable differences in take-up rates (such as Muslim borrowers), prediction rates should be estimated within sub-populations.

Table 5 displays the results of the baseline probit from question six (in column one) and the results of a probit estimation that includes an interaction between the dummy variables for whether a potential borrow is Muslim and/or married. The LRI increases slightly when the interaction term is included, but the model remains poorly-fitted, and the change in the magnitude and statistical significance of coefficients (other than those that are interacted) is minimal. Futhermore, the interaction term itself is not significant, nor do the coefficients of the relevant dummy variables become significant. All of this suggests that the interaction term is not likely a critical determinant of loan take-up.

Table 5: Probit Model Comparison

	(1)	(2)	(3)	(4)
VARIABLES	Baseline	Interaction Effect	HetProbit ;	lnsigma
Treated	0.184	0.190	0.570	
	(0.138)	(0.139)	(0.473)	
Client_Age	0.00215	0.00269	-0.0233	0.0137
	(0.00763)	(0.00772)	(0.0345)	(0.0132)
Client_Education	-0.0133	-0.0139	-0.164	0.0648
	(0.0153)	(0.0153)	(0.166)	(0.0532)
HH_Income	1.32e-05	1.25e-05	3.94e-05	
	(1.39e-05)	(1.39e-05)	(4.12e-05)	
Hindu_SC_Kat	-0.115	-0.118	-0.177	
	(0.206)	(0.206)	(0.470)	
1.Client_Married	0.133	0.218	0.449	
	(0.200)	(0.251)	(0.576)	
1.muslim	-0.0713	0.140	-0.185	
	(0.144)	(0.392)	(0.334)	
1.Client_Married#1.muslim		-0.243		
		(0.419)		
Constant	-1.249	-1.341	-1.554	
	(0.364)	(0.400)	(0.688)	
LRI	0.00863	0.00929		

Standard errors in parentheses

The mean finite difference for the interaction term is -0.057 . The marginal effect for each case (e.g., Muslim but not married, married and Muslim), using a model with and without an interaction term, is displayed in Table 6. Including

[4] In this example, this is because, if the DGP results in rates on either extreme, a model that assigns a probability of 0 or 1 to all observations can appear to have an extremely high prediction rate. A threshold that is equal to the population mean requires the model to predict better than randomization.

an interaction term results in Muslim potential borrowers being more likely to take out a loan, *ceteris paribus*, than non-Muslim potential borrowers. This further supports suggests that either the model is mis-specified or the sample is not representative of the population, since usury is more taboo among Muslim communities than others in the United States. However, the model without the interaction term does result in the expected effect, with unmarried Muslims being less likely to borrow than unmarried non-Muslims, and married Muslims being more likely to borrow than unmarried Muslims.

Table 6: Mean finite differences

		Mean Finite Difference	
$\mathbb{1}\{Client\_Married\}$	$\mathbb{1}\{Muslim\}$	w/o interaction	w/ interaction
1	0	0.032	0.051
0	1	-0.016	0.032
1	1	0.017	0.026

The standard deviation of the interaction effect between the married and Muslim dummies is 0.0072, or 12.7% of the mean , which is not especially low, considering the small number of married Muslims in the sample—141, or 25.1% of the sample.

Table 7 displays the results of regressing the squared residuals of the (non-quartic) LPM on the covariates of the LPM. The only statistically significant coefficient is the one for client age, suggesting that there is some heteroskedasticity in the client age variable (likely due to misspecification), but not in any other variable.

Table 7: Regression of squared errors on covariates

VARIABLES	(1)
Treated	0.0297
	(0.0222)
Client_Age	0.000444
	(0.00124)
Client_Married	0.0225
	(0.0317)
Client_Education	-0.00197
	(0.00250)
HH_Income	2.26e-06
	(2.35e-06)
muslim	-0.0108
	(0.0235)
Hindu_SC_Kat	-0.0176
	(0.0332)
Constant	0.0872
	(0.0570)
Standard errors in parentheses	

Columns three and four of Table 5 display the coefficients from the heteroskedasticity robust probit model and the heteroskedasticity parameters for client age and education, respectively. The coefficients on treatment and client age differ substantially from those in the baseline model,[5], but the other coefficients do not differ much. However, neither of the heteroskedasticity parameters are statistically significant. Taken as a whole, this is weak evidence in favor of heteroskedasticity that would likely be strengthened in a larger sample.

[5]Given our results from the heteroskedasticity analysis in the last question, the result for client age is unsurprising