Introductory Econometrics
Binary dependent variable
Suggested Solution

by Hieu Nguyen

Fall 2024

1.

Use the data in `loanapp_b.gdt` for this exercise. The binary variable to be explained is `approve`, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is `whiteskin`, a dummy variable equal to one if the applicant has light skin. The other applicants in the data set are `darkskin` and `Hispanic`. To test for discrimination in the mortgage loan market, a LPM can be used:

$$\text{approve} = \beta_0 + \beta_1 \text{whiteskin} + \text{other factors.}$$

(a) Regress `approve` on `whiteskin` and report the results in the usual form. Interpret the estimated coefficient on `whiteskin`. Is it significant? Is it practically large?

(b) As controls, add the variables `hrat`, `obrat`, `loanprc`, `unem`, `male`, `married`, `dep`, `sch`, `cosign`, `chist`, `pubrec`, `mortlat1`, `mortlat2`, and `vr`. What happens to the estimated coefficient on `whiteskin`? Is there still statistically significant evidence of discrimination against non-white skin individuals?

(c) Estimate the equation in part (b) computing the White heteroskedasticity-consistent robust standard errors. Compare the 95% confidence interval on $\beta_{\text{whiteskin}}$ with the non-robust confidence interval.

(d) Obtain the fitted values from the regression in part (c). Are any of them less than zero? Are any of them greater than one?

(e) Estimate a Probit model of `approve` on `whiteskin`. Check the direction of the effect and the statistical significance of `whiteskin`. Find the estimated probability of loan approval for both `whiteskin` and non-white skin individuals. How do these compare with the LPM estimates?

(f) Now, add the variables `hrat`, `obrat`, `loanprc`, `unem`, `male`, `married`, `dep`, `sch`, `cosign`, `chist`, `pubrec`, `mortlat1`, `mortlat2`, and `vr`. Is there still statistically significant evidence of discrimination against non-white skin people? Interpret also other information from the Gretl output.

(g) Estimate the model from part (f) by Logit. Compare the estimated coefficient on `whiteskin` to the Probit model.

(h) Estimate the sizes of the discrimination effects for Probit and Logit.

Solution:

(a) The benchmark LPM is:

```
Model 1: OLS, using observations 1{1989
Dependent variable: approve

Coefficient     Std. Error    t-ratio    p-value
------------------------------------------------
```

```
const           0.707792      0.0182393  38.8060   0.0000
whiteskin       0.200596      0.0198400  10.1107   0.0000

Mean dependent var  0.877325    S.D. dependent var  0.328146
Sum squared resid   203.5930    S.E. of regression  0.320098
R^2                0.048930    Adjusted R^2       0.048451
F (1, 1987)        102.2261    P-value(F )        1.81e{23
Log-likelihood     555.5405   Akaike criterion    1115.081
Schwarz criterion  1126.272    Hannan{Quinn        1119.191
```

Based on the standard t-test and the estimated coefficient, `whiteskin` is statistically significant at all standard levels. This elementary linear probability model (LPM) may suffer from various econometric problems (e.g., many omitted variables and inherent heteroskedasticity). However, it suggests a statistically as well as economically (practically large) significant effect of `whiteskin`: by 20 percentage points higher probability of obtaining a mortgage loan.

(b) The model now becomes:

```
Model 2: OLS, using observations 1{1989 (n = 1971)
Missing or incomplete observations dropped: 18

Dependent variable: approve

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           0.936731      0.0527354  17.7629   0.0000
whiteskin       0.128820      0.0197317  6.5286    0.0000
hrat            0.00183299    0.00126320  1.4511    0.1469
obrat          0.00543180    0.00110178 4.9300    0.0000
loanprc        0.147300      0.0375159 3.9263    0.0001
unem           0.00729893    0.00319799 2.2824    0.0226
male           0.00414414    0.0188644  0.2197    0.8261
married         0.0458241     0.0163077   2.8100    0.0050
dep            0.00682737    0.00670134 1.0188    0.3084
sch             0.00175251    0.0166498   0.1053    0.9162
cosign          0.00977222    0.0411394   0.2375    0.8123
chist           0.133027      0.0192627   6.9059    0.0000
pubrec         0.241927      0.0282274  8.5706    0.0000
mortlat1       0.0572511     0.0500120  1.1447    0.2525
mortlat2       0.113723      0.0669838  1.6978    0.0897
vr             0.0314408     0.0140313  2.2408    0.0252

Mean dependent var  0.876205    S.D. dependent var  0.329431
Sum squared resid   178.3935    S.E. of regression  0.302076
R^2                0.165582    Adjusted R^2       0.159180
F (15, 1955)       25.86339    P-value(F )        1.84e{66
Log-likelihood     429.2569   Akaike criterion    890.5139
Schwarz criterion  979.8946    Hannan{Quinn        923.3569
```

The discrimination in the mortgage loan market remains statistically and economically significant. Even after controlling for many other effects, `whiteskin` is associated with almost 13 percentage points statistically significantly higher probability of obtaining a mortgage loan.

(c) Inherent heteroskedasticity (because of the nature of the binary dependent variable, check, e.g., the White test) is remedied using the White heteroskedasticity-consistent robust standard errors:

```
Model 3: OLS, using observations 1{1989 (n = 1971)
```

```
Missing or incomplete observations dropped: 18

Dependent variable: approve
Heteroskedasticity-robust standard errors, variant HC1

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const            0.936731      0.0593886  15.7729   0.0000
whiteskin        0.128820      0.0258693  4.9796    0.0000
hrat             0.00183299    0.00146703  1.2495    0.2116
obrat           0.00543180    0.00133099 4.0810    0.0000
loanprc         0.147300      0.0378351 3.8932    0.0001
unem            0.00729893    0.00371219 1.9662    0.0494
male            0.00414414    0.0193044  0.2147    0.8300
married          0.0458241     0.0172374   2.6584    0.0079
dep             0.00682737    0.00690380 0.9889    0.3228
sch              0.00175251    0.0171460   0.1022    0.9186
cosign           0.00977222    0.0395825   0.2469    0.8050
chist            0.133027      0.0246202   5.4031    0.0000
pubrec          0.241927      0.0427922  5.6535    0.0000
mortlat1        0.0572511     0.0662234  0.8645    0.3874
mortlat2        0.113723      0.0910697  1.2488    0.2119
vr              0.0314408     0.0144855  2.1705    0.0301

Mean dependent var  0.876205    S.D. dependent var  0.329431
Sum squared resid   178.3935    S.E. of regression  0.302076
R^2                 0.165582    Adjusted R^2        0.159180
F (15, 1955)        14.97726    P-value(F )         4.04e{37
Log-likelihood      429.2569   Akaike criterion    890.5139
Schwarz criterion  979.8946    Hannan{Quinn        923.3569
```

Provided we are sure how to compute confidence intervals manually, we can obtain them directly in the Gretl from Model 3 menu: Analysis—Confidence intervals. For Model 3 with robust SEs, we get:

$$t(1955, 0.025) = 1.961$$

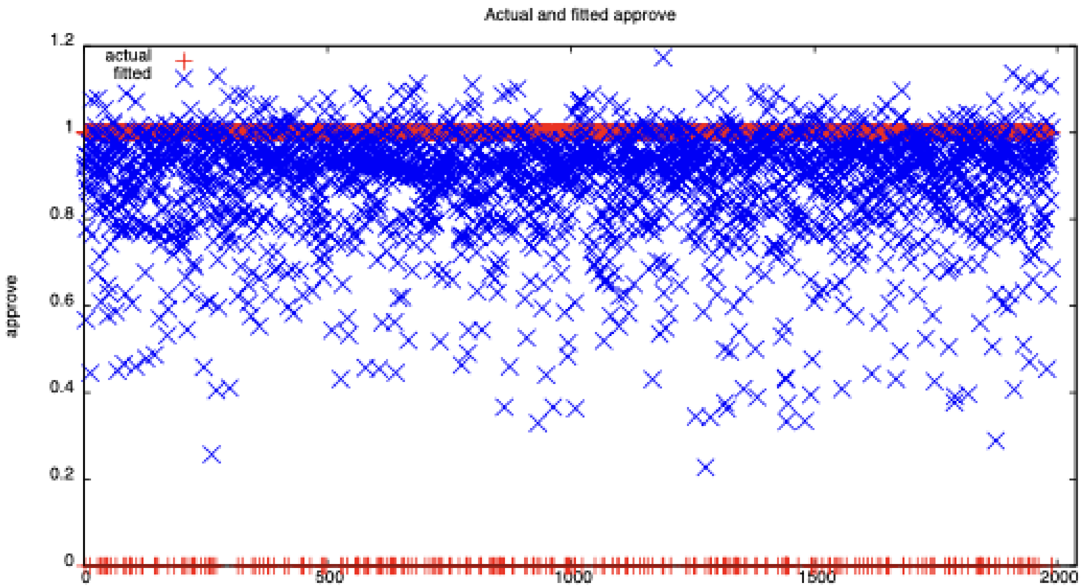
VARIABLE	COEFFICIENT	95% CONFIDENCE INTERVAL
whiteskin	0.128820	(0.0780852, 0.179554)

Compared to Model 2 (with smaller non-robust SEs):

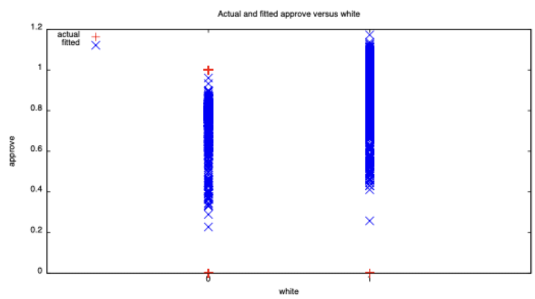
$$t(1955, 0.025) = 1.961$$

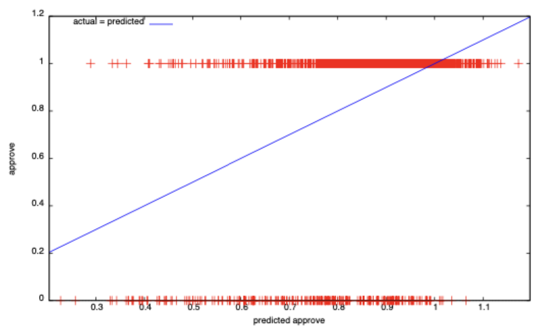
VARIABLE	COEFFICIENT	95% CONFIDENCE INTERVAL
whiteskin	0.128820	(0.0901223, 0.167517)

(d) In the Gretl Model 3 menu follow Graphs—Fitted, actual plot—By obs. number:



We observe that there are many fitted values above 1 (but only for `whiteskin`); however, none below 0 (this could have been, to some extent, expected from the mean of `approve` = 0.88). Two other optional graphical depictions of fitted values follow.





(e) The benchmark Probit model is (in the Gretl menu follow Model—Limited dependent variable—Probit—Binary... and tick Show p-values):

```
Model 4: Probit, using observations 1{1989
Dependent variable: approve

Standard errors based on Hessian

Coefficient      Std. Error     z           p-value
-----------------------------------------------------
const            0.546946       0.0754350  7.2506    0.0000
whiteskin        0.783946       0.0867118  9.0408    0.0000

Mean dependent var   0.877325    S.D. dependent var   0.328146
McFadden R2          0.053312    Adjusted R2          0.050610
Log-likelihood     700.8774     Akaike criterion     1405.755
Schwarz criterion 1416.946     Hannan{Quinn         1409.865
```

Directions of effects of individual explanatory variables, as well as their statistical significance, can be interpreted from the output directly in a similar way as for the OLS output. However, for the magnitude of the effects or fitted/predicted values, it is important to consider also the standard normal CDF:

- Estimated/fitted probability for `whiteskin` individuals:

$$\hat{p}_i = F(\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}) = \Phi(\beta_0+\beta_1 x_{i,\text{whiteskin}=1}) = \Phi(0.547+0.784\cdot 1) = \Phi(1.331) \approx 0.9082 \approx 91\%;$$

- Estimated/fitted probability for non-white skin individuals:

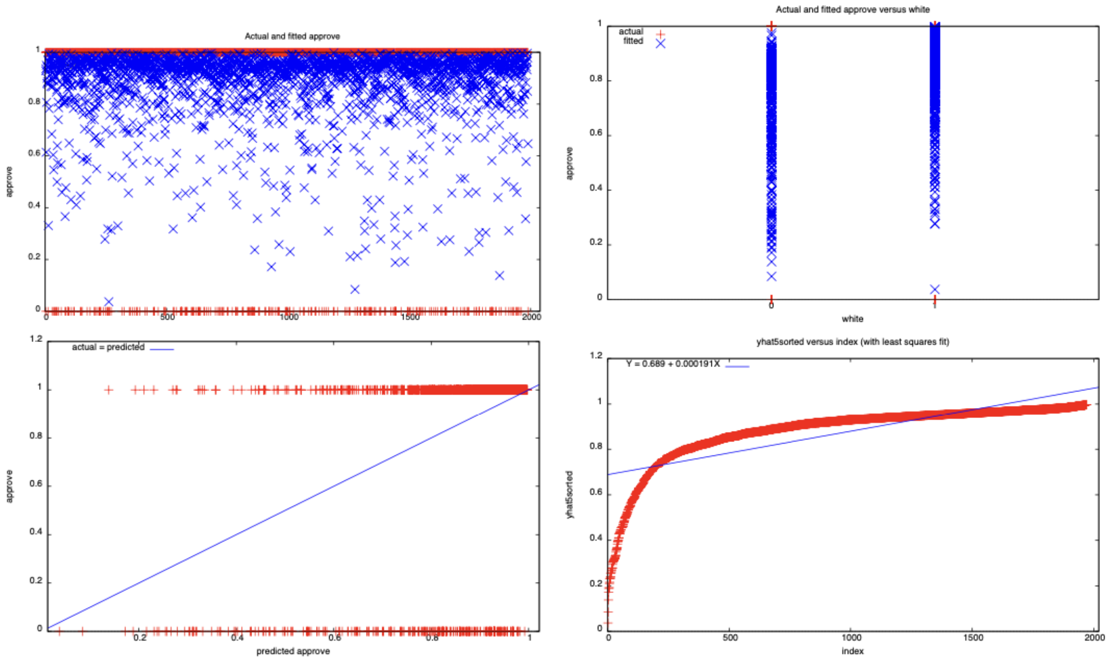
$$\hat{p}_i = F(\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}) = \Phi(\beta_0+\beta_1 x_{i,\text{whiteskin}=0}) = \Phi(0.547+0.784\cdot 0) = \Phi(0.547) \approx 0.7088 \approx 71\%.$$

(f) The model now becomes:

```
Model 5: Probit, using observations 2{1989 (n = 1971)
Missing or incomplete observations dropped: 17

Dependent variable: approve
Standard errors based on Hessian

Coefficient     Std. Error    z          p-value
-------------------------------------------------
const            2.06233       0.313176   6.5852    0.0000
whiteskin        0.520253      0.0969588  5.3657    0.0000
hrat             0.00787633    0.00696162 1.1314    0.2579
obrat           0.0276924     0.00604930 4.5778    0.0000
loanprc         1.01197       0.237240   4.2656    0.0000
unem            0.0366849     0.0174807  2.0986    0.0359
male            0.0370014     0.109927   0.3366    0.7364
married          0.265747      0.0942523  2.8195    0.0048
dep             0.0495756     0.0390573  1.2693    0.2043
sch              0.0146497     0.0958421  0.1529    0.8785
cosign           0.0860713     0.245751   0.3502    0.7262
chist            0.585281      0.0959715  6.0985    0.0000
pubrec          0.778741      0.126320  6.1648    0.0000
mortlat1        0.187624      0.253113  0.7413    0.4585
mortlat2        0.494356      0.326556  1.5138    0.1301
vr              0.201062      0.0814934 2.4672    0.0136

Mean dependent var  0.876205    S.D. dependent var  0.329431
McFadden R2        0.186602    Adjusted R2         0.164921
Log-likelihood    600.2710    Akaike criterion    1232.542
Schwarz criterion 1321.923    Hannan{Quinn         1265.385
```

Based on the Z-test on `whiteskin`, we still observe statistically significant discrimination on the mortgage loan market. Interpretation of the Probit output (McFadden R2, Log-likelihood, percent correctly predicted, LR test) was discussed in detail during the seminar.

Comparable graphical depictions (+ one extra) of fitted values for Probit (do observe differences to the LPM fitted values):



(g) Estimated Logit model:

Model 6: Logit, using observations 2{1989 (n = 1971)
Missing or incomplete observations dropped: 17

Dependent variable: approve
Standard errors based on Hessian

	Coefficient	Std. Error	z	p-value
const	3.80171	0.594707	6.3926	0.0000
whiteskin	0.937764	0.172904	5.4236	0.0000
hrat	0.0132631	0.0128802	1.0297	0.3031
obrat	0.0530338	0.0112803	4.7015	0.0000
loanprc	1.90495	0.460443	4.1372	0.0000
unem	0.0665789	0.0328086	2.0293	0.0424
male	0.0663851	0.206429	0.3216	0.7478
married	0.503282	0.177998	2.8275	0.0047
dep	0.0907335	0.0733342	1.2373	0.2160
sch	0.0412288	0.178404	0.2311	0.8172
cosign	0.132059	0.446094	0.2960	0.7672
chist	1.06658	0.171212	6.2296	0.0000
pubrec	1.34067	0.217366	6.1678	0.0000
mortlat1	0.309882	0.463520	0.6685	0.5038
mortlat2	0.894675	0.568581	1.5735	0.1156
vr	0.349828	0.153725	2.2757	0.0229

Mean dependent var	0.876205	S.D. dependent var	0.329431
McFadden R2	0.186297	Adjusted R2	0.164616
Log-likelihood	600.4962	Akaike criterion	1232.992
Schwarz criterion	1322.373	Hannan{Quinn	1265.835

Comparison Logit vs Probit vs LPM (to compute average marginal effects for Logit and Probit, tick Show slopes at mean):

```
Model 6: Logit, using observations 2{1989 (n = 1971)
Missing or incomplete observations dropped: 17

Dependent variable: approve
Standard errors based on Hessian

Coefficient     Std. Error    z          Slope
-------------------------------------------------
const           3.80171       0.594707   6.3926
whiteskin       0.937764      0.172904   5.4236   0.0967431
hrat            0.0132631     0.0128802  1.0297   0.00104057
. . .
```

```
Model 5: Probit, using observations 2{1989 (n = 1971)
Missing or incomplete observations dropped: 17

Dependent variable: approve
Standard errors based on Hessian

Coefficient     Std. Error    z          Slope
-------------------------------------------------
const           2.06233       0.313176   6.5852
whiteskin       0.520253      0.0969588  5.3657   0.105747
hrat            0.00787633    0.00696162 1.1314   0.00127210
. . .
```

Evaluated at the mean

```
Model 3: OLS, using observations 1{1989 (n = 1971)
Missing or incomplete observations dropped: 18

Dependent variable: approve
Heteroskedasticity-robust standard errors, variant HC1

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           0.936731      0.0593886  15.7729   0.0000
whiteskin       0.128820      0.0258693  4.9796    0.0000
hrat            0.00183299    0.00146703  1.2495    0.2116
. . .
```

Estimated coefficients are generally different and cannot be interpreted directly. However, we may use several rules of thumb to quickly and roughly compare the Logit, Probit, and LPM estimates:

- We can multiply the Probit estimates by $0.4/0.25 = 1.6$, or we can multiply the Logit estimates by $0.25/0.4 = 0.625$ to make them roughly comparable;
- We can multiply Probit estimates by 0.4 and Logit estimates by 0.25 to make them roughly comparable to the LPM estimates.

(h) Compare average marginal effects (computed at the means of all explanatory variables) between Logit and Probit ('Slopes' from the Gretl output) in the previous exercise. They are largely similar but slightly: $\beta_{1,\text{whiteskin}} > \text{Slope}_{\text{whiteskin,Probit}} > \text{Slope}_{\text{whiteskin,Logit}}$. This could have been expected given the different shapes of the linear vs standard normal CDF vs logistic CDF (because of the fatter tails of the logistic PDF, logistic CDF is positioned a bit below standard normal CDF for $x > 0$).