Utrecht University

Tutorials
Week 8

Utrecht University

Pdf file on Blackboard	Dataset on Blackboard	Papers related to the datasets	Description
C.13.3	Kielmc.dta	K.A. Kiel and K.T. McClain (1995): House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor Through Operation, Journal of Environmental Economics and Management 28, 241-255.	Dif-in-dif estimator
C.7.13	apple.dta	Van Ravenswaay, E.O and Blend, J.R. (1998): Consumer Demand for Ecolabeled apples: Survey Methods and Descriptive Results, AgEconsearch, 98-20, 1-45. 10.22004/ag.econ.11645	Estimation of Linear Probability Model (LPM).
C.7.8	loanapp.dta	W.C. Hunter and M.B. Walker (1996): The Cultural Affinity Hypothesis and Mortgage Lending Decisions, Journal of Real Estate Finance and Economics 13, 57-70.	Estimation with probit model, calculation of marginal effects, comparison logit model.
C.17.2	loanapp.dta	W.C. Hunter and M.B. Walker (1996): The Cultural Affinity Hypothesis and Mortgage Lending Decisions, Journal of Real Estate Finance and Economics 13, 57-70.	Estimation the restricted model, estimation with logit, probit fitted values, LPM.

C3 Use the data in KIELMC.RAW for this exercise.

(i) The variable *dist* is the distance from each home to the incinerator site, in feet. Consider the model

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 \log(dist) + \delta_1 y81 \cdot \log(dist) + u.$$

If building the incinerator reduces the value of homes closer to the site, what is the sign of δ_1? What does it mean if $\beta_1 > 0$?

(ii) Estimate the model from part (i) and report the results in the usual form. Interpret the coefficient on $y81 \cdot \log(dist)$. What do you conclude?

(iii) Add *age*, age^2, *rooms*, *baths*, $\log(intst)$, $\log(land)$, and $\log(area)$ to the equation. Now, what do you conclude about the effect of the incinerator on housing values?

(iv) Why is the coefficient on $\log(dist)$ positive and statistically significant in part (ii) but not in part (iii)? What does this say about the controls used in part (iii)?

(i) Model Specification: $\log(price) = \beta_0 + \delta_0 y_{81} + \beta_1 \log(dist) + \delta_1 y_{81} * \log(dist) + u$

```
. reg lprice y81 ldist y81ldist
```

Source	SS	df	MS
Model	24.3172548	3	8.10575159
Residual	37.1217306	317	.117103251
Total	61.4389853	320	.191996829

Number of obs	=	321
F(3, 317)	=	69.22
Prob > F	=	0.0000
R-squared	=	0.3958
Adj R-squared	=	0.3901
Root MSE	=	.3422

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	-.0113101	.8050622	-0.01	0.989	-1.59525	1.57263
ldist	.316689	.0515323	6.15	0.000	.2153005	.4180775
y81ldist	.0481862	.0817929	0.59	0.556	-.1127394	.2091117
_cons	8.058468	.5084358	15.85	0.000	7.058133	9.058803

Other things equal, homes farther from the incinerator should be worth more, so $\delta_1 > 0$. If $\beta_1 > 0$, then the incinerator was located farther away from more expensive homes.

Additional: interpret the coefficient on log(dist) $\widehat{\beta_1}$ and its statistical significance. Is $\widehat{\delta_1}$ it statistically different from zero?

(ii) The estimated equation is:

$$\begin{aligned}\log(\textit{price}) &= \underset{(0.51)}{8.06} - \underset{(.805)}{.011}\, y81 + \underset{(.052)}{.317} \log(\textit{dist}) + \underset{(.082)}{.048}\, y81 \cdot \log(\textit{dist})\\ n &= 321, \quad R^2 = .396, \quad \bar{R}^2 = .390.\end{aligned}$$

While $\hat{\delta}_1 = .048$ is the expected sign, it is not statistically significant (t statistic $\approx$ .59).

- For houses sold in 1981, a 1% increase in distance is associated with an additional 0.048% increase in the price of the house compared to houses sold in other years.
- However, due to its lack of statistical significance (p-value 0.5560.556), we cannot confidently assert that this interaction has a meaningful impact on house prices.

Utrecht University

(iii)

```
. reg lprice y81 ldist y81ldist age agesq rooms baths lintst lland larea
```

Source	SS	df	MS
Model	48.353762	10	4.8353762
Residual	13.0852234	310	.042210398
Total	61.4389853	320	.191996829

Number of obs	=	321
F(10, 310)	=	114.55
Prob > F	=	0.0000
R-squared	=	0.7870
Adj R-squared	=	0.7802
Root MSE	=	.20545

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	-.2254466	.4946914	-0.46	0.649	-1.198824	.7479309
ldist	.0009226	.0446168	0.02	0.984	-.0868674	.0887125
y81ldist	.0624668	.0502788	1.24	0.215	-.036464	.1613976
age	-.0080075	.0014173	-5.65	0.000	-.0107962	-.0052187
agesq	.0000357	8.71e-06	4.10	0.000	.0000186	.0000528
rooms	.0461389	.0173442	2.66	0.008	.0120117	.0802662
baths	.1010478	.0278224	3.63	0.000	.0463032	.1557924
lintst	-.0599757	.0317217	-1.89	0.060	-.1223929	.0024414
lland	.0953425	.0247252	3.86	0.000	.046692	.143993
larea	.3507429	.0519485	6.75	0.000	.2485266	.4529592
_cons	7.673854	.5015718	15.30	0.000	6.686938	8.660769

- When we add the list of housing characteristics to the regression, the coefficient on *y81*log(dist)* becomes 0.062 (se=0.050). The estimated effect is larger – the elasticity of price with respect to *dist* is 0.062 after the incinerator site was chosen - but its t-statistics is only 1.24.
- One could conclude that there is no evidence in favor of a positive effect. → **Show H_0, H_1, rejection area, etc**.

(iv) After including further home specifications, the variable distance is not significant anymore (p-value: 0.98 vs. 0.00). The controls used in part (iii) signalized that the housing characteristics capture differences between houses close and far away from the incinerator.

Utrecht University

C13 Use the data in APPLE.RAW to answer this question.

(i) Define a binary variable as $ecobuy = 1$ if $ecolbs > 0$ and $ecobuy = 0$ if $ecolbs = 0$. In other words, *ecobuy* indicates whether, at the prices given, a family would buy any ecologically friendly apples. What fraction of families claim they would buy ecolabeled apples?

(ii) Estimate the linear probability model

$$\begin{aligned} ecobuy = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc \\ + \beta_4 hhsize + \beta_5 educ + \beta_6 age + u, \end{aligned}$$

and report the results in the usual form. Carefully interpret the coefficients on the price variables.

(iii) Are the nonprice variables jointly significant in the LPM? (Use the usual F statistic, even though it is not valid when there is heteroskedasticity.) Which explanatory variable other than the price variables seems to have the most important effect on the decision to buy ecolabeled apples? Does this make sense to you?

(iv) In the model from part (ii), replace *faminc* with $\log(faminc)$. Which model fits the data better, using *faminc* or $\log(faminc)$? Interpret the coefficient on $\log(faminc)$.

(v) In the estimation in part (iv), how many estimated probabilities are negative? How many are bigger than one? Should you be concerned?

(vi) For the estimation in part (iv), compute the percent correctly predicted for each outcome, $ecobuy = 0$ and $ecobuy = 1$. Which outcome is best predicted by the model?

Utrecht University

i) First, generate the dummy variable:

ecobuy = 1 if ecolbs > 0 → if at prices given, a family would buy ecological apples.
ecobuy = 0, if ecolbs = 0

ecolbs: quantity eco-labeled apples, lbs

```
. gen ecobuy= ecolbs>0
```

Create a table of frequency:

```
. tab ecobuy
```

ecobuy	Freq.	Percent	Cum.
0	248	37.58	37.58
1	412	62.42	100.00
Total	660	100.00	

62.42% of the families claim they would buy ecolabelled apples.

(ii) What are the price variables?

Ecoprc : price of eco-labeled apples
Regprc: price of regular apples

```
. reg ecobuy ecoprc regprc faminc hhsize educ age, rob

Linear regression                               Number of obs =      660
                                                F(  6,    653) =    14.93
                                                Prob > F       =   0.0000
                                                R-squared      =   0.1098
                                                Root MSE       =   .45939
```

ecobuy	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
ecoprc	-.8026219	.1056678	-7.60	0.000	-1.010112	-.5951321
regprc	.7192675	.1302317	5.52	0.000	.463544	.9749911
faminc	.0005518	.0005245	1.05	0.293	-.0004781	.0015817
hhsize	.0238227	.0124672	1.91	0.056	-.0006579	.0483033
educ	.0247849	.0084565	2.93	0.003	.0081796	.0413901
age	-.0005008	.0012655	-0.40	0.692	-.0029858	.0019842
_cons	.4236865	.1677529	2.53	0.012	.0942864	.7530867

- If *ecoprc* increases by, say, 10 cents (.10), then the probability of buying eco-labeled apples falls by about .080, c.p.
- If *regprc* increases by 10 cents, the probability of buying eco-labeled apples increases by about .072, c.p.
- We can assume that both sorts of apples are substitutes. – cross-price elasticity.

(iii) There are four non price variables: hhsize, faminc, educ and age

```
. reg ecobuy ecoprc regprc faminc hhsize educ age, rob

Linear regression                                      Number of obs =     660

                                                       F(  6,   653) =   14.93
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1098
                                                       Root MSE      =  .45939

------------------------------------------------------------------------------
             |               Robust
      ecobuy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ecoprc |  -.8026219   .1056678    -7.60   0.000    -1.010112   -.5951321
      regprc |   .7192675   .1302317     5.52   0.000      .463544    .9749911
      faminc |   .0005518   .0005245     1.05   0.293    -.0004781    .0015817
      hhsize |   .0238227   .0124672     1.91   0.056    -.0006579    .0483033
        educ |   .0247849   .0084565     2.93   0.003     .0081796    .0413901
         age |  -.0005008   .0012655    -0.40   0.692    -.0029858    .0019842
       _cons |   .4236865   .1677529     2.53   0.012     .0942864    .7530867
------------------------------------------------------------------------------
```

```
. test faminc hhsize educ age

 ( 1)  faminc = 0
 ( 2)  hhsize = 0
 ( 3)  educ = 0
 ( 4)  age = 0

       F(  4,   653) =    4.24
            Prob > F =    0.0021
```

hhsize: household size
faminc: family income, thousands
educ: years schooling
age: in years

They are jointly significant at 1% level of significance.
For the exam, you have to write all the statistical steps.

- The *F* test, with 4 and 653 *df*, is 4.23, with *p*-value = .0021. Thus, based on the usual *F*-test, the four non-price variables are jointly very significant. Of the four variables, *educ* appears to have the most important effect.
- For example, a difference of four years of education implies an increase of .025(4) = .10 in the estimated probability of buy ing eco-labeled apples. This suggests that more highly educated people are more open to buying products that is environmentally friendly, which is perhaps expected. Household size (*hhsize*) also has an effect.
- Comparing a couple with two children to one with no children – other factors equal – the couple with two children has a .048 higher probability of buying eco-labeled apples.

Utrecht University

(iv) Compare both models and decide which fits better. Interpret log(faminc)

```
. gen lfaminc=ln(faminc)

. reg ecobuy ecoprc regprc lfaminc hhsize educ age, rob

Linear regression                               Number of obs =       660
                                                F(  6,   653) =     15.24
                                                Prob > F      =    0.0000
                                                R-squared     =    0.1116
                                                Root MSE      =    .45893

------------------------------------------------------------------------------
             |               Robust
      ecobuy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ecoprc |  -.8006664   .1055406    -7.59   0.000    -1.007906   -.5934265
      regprc |    .721377   .1298925     5.55   0.000     .4663197    .9764343
     lfaminc |   .0445162   .0292792     1.52   0.129    -.0129766     .102009
      hhsize |   .0227002   .0124989     1.82   0.070    -.0018426    .0472429
        educ |    .023093   .0085234     2.71   0.007     .0063564    .0398296
         age |  -.0003865   .0012645    -0.31   0.760    -.0028695    .0020964
       _cons |   .3037519   .1817885     1.67   0.095    -.0532087    .6607125
------------------------------------------------------------------------------
```

```
. reg ecobuy ecoprc regprc faminc hhsize educ age, rob

Linear regression                               Number of obs =       660

                                                F(  6,   653) =     14.93
                                                Prob > F      =    0.0000
                                                R-squared     =    0.1098
                                                Root MSE      =    .45939

------------------------------------------------------------------------------
             |               Robust
      ecobuy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ecoprc |  -.8026219   .1056678    -7.60   0.000    -1.010112   -.5951321
      regprc |   .7192675   .1302317     5.52   0.000      .463544    .9749911
      faminc |   .0005518   .0005245     1.05   0.293    -.0004781    .0015817
      hhsize |   .0238227   .0124672     1.91   0.056    -.0006579    .0483033
        educ |   .0247849   .0084565     2.93   0.003     .0081796    .0413901
         age |  -.0005008   .0012655    -0.40   0.692    -.0029858    .0019842
       _cons |   .4236865   .1677529     2.53   0.012     .0942864    .7530867
------------------------------------------------------------------------------
```

- The model with log(*faminc*) fits the data slightly better: the *R*-squared increases to about .112. (We would not expect a large increase in *R*-squared from a simple change in the functional form.)
- The coefficient on log(*faminc*) is about .045 (t = 1.52).
- **Level-log interpretation**: holding other factors fixed, one percentage increase in family income leads to 0.00045 higher probability of families buying eco-labelled apples.
- If log(*faminc*) increases by .10, which means roughly a 10% increase in *faminc*, then P(*ecobuy* = 1) is estimated to increase by about .0045, a pretty small effect.

Utrecht University

(v) To know how many probabilities are negative, and how to deal with that:

```
. predict yhat
(option xb assumed; fitted values)

. gen above=yhat>1

. gen below=yhat<0

. tab above

      above |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        658       99.70       99.70
          1 |          2        0.30      100.00
------------+-----------------------------------
      Total |        660      100.00
```

```
. tab below

      below |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        660      100.00      100.00
------------+-----------------------------------
      Total |        660      100.00
```

- The fitted probabilities range from about .185 to 1.051, so none are negative.
- There are two fitted probabilities above 1, which is not a source of concern with 660 observations.
- Probabilities can be above. However, we can assume (in this example) that the above 1 probabilities represent 1.
- None of the estimated probabilities are below 0.

Utrecht University

(vi) We need to choose a cutoff value to convert the estimated probabilities into a dummy variable. Without additional information, we choose 50% as the cutoff value. If the estimated probability of purchasing eco-labeled apples is above 50%, our model predicts that the household will buy eco-labeled apples.

Cross-tabulate ecobuy=1 and ecobuy=0

```
. gen estecobuy=yhat>.5

. tab estecobuy ecobuy

           |        ecobuy
 estecobuy |         0          1 |     Total
-----------+----------------------+----------
         0 |       102         72 |       174
         1 |       146        340 |       486
-----------+----------------------+----------
     Total |       248        412 |       660
```

- Using the standard prediction rule – predict one when $ecobuy_i \geq 0.5$ and zero otherwise –gives the fraction correctly predicted for *ecobuy* = 0 as 102/248 ≈ .411, so about 41.1%.
- The model correctly predicts 340/412 = 0.825, that is 82.5% of the cases when a family bought eco-labeled apples.
- With the usual prediction rule, the model performs better for families that buy ecolabeled apples.
- The model correctly predicts 442 cases out of 660, which is 67%. This is a pseudo R^2 and describes the overall fit of the model.

Utrecht University

C8 Use the data in LOANAPP.RAW 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 *white*, a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic.

To test for discrimination in the mortgage loan market, a linear probability model can be used:

$$approve = \beta_0 + \beta_1 white + other\ factors.$$

(i) If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of β_1?

(ii) Regress *approve* on *white* and report the results in the usual form. Interpret the coefficient on *white*. Is it statistically significant? Is it practically large?

(iii) As controls, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, and *vr*. What happens to the coefficient on *white*? Is there still evidence of discrimination against nonwhites?

(iv) Now, allow the effect of race to interact with the variable measuring other obligations as a percentage of income (*obrat*). Is the interaction term significant?

(v) Using the model from part (iv), what is the effect of being white on the probability of approval when *obrat* = 32, which is roughly the mean value in the sample? Obtain a 95% confidence interval for this effect.

(i) If the appropriate factors have been controlled for, $\beta_1 > 0$ signals discrimination against minorities: a white person has a greater chance of having a loan approved, other relevant factors fixed.

(ii) The simple regression results are

$$\begin{aligned} approve &= \underset{(.018)}{.708} + \underset{(.020)}{.201}\, white \\ n &= 1{,}989, \quad R^2 = .049. \end{aligned}$$

```
. reg approve white, rob

Linear regression                               Number of obs =     1989
                                                F(  1,  1987) =    55.75
                                                Prob > F      =   0.0000
                                                R-squared     =   0.0489
                                                Root MSE      =    .3201

------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .2005957   .0268651     7.47   0.000      .147909    .2532824
       _cons |   .7077922   .0259264    27.30   0.000     .6569465     .758638
------------------------------------------------------------------------------
```

- The coefficient on *white* means that, in the sample of 1,989 loan applications, an application submitted by a white applicant was 20 percentage points more likely to be approved than that of a nonwhite applicant.
- It is statistically significant 1% , and the t statistic is 7.47.
- 20% more chance of getting the application approved is a significant difference.

(iii) Is there still evidence for discrimination against non-whites?

```
. reg approve white hrat obrat loanprc unem male married dep sch cosign chist pubrec
mortlat1 mortlat2
>  vr, rob

Linear regression                                      Number of obs =    1971
                                                       F( 15,  1955) =   14.98
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1656
                                                       Root MSE      =  .30208

------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .1288196   .0258693     4.98   0.000     .0780852     .179554
        hrat |    .001833    .001467     1.25   0.212    -.0010441    .0047101
       obrat |  -.0054318    .001331    -4.08   0.000    -.0080421   -.0028215
     loanprc |  -.1473001   .0378351    -3.89   0.000    -.2215013   -.0730988
        unem |  -.0072989   .0037122    -1.97   0.049    -.0145792   -.0000187
        male |  -.0041441   .0193044    -0.21   0.830    -.0420035    .0337152
     married |   .0458241   .0172374     2.66   0.008     .0120186    .0796296
         dep |  -.0068274   .0069038    -0.99   0.323    -.0203669    .0067122
         sch |   .0017525    .017146     0.10   0.919    -.0318739    .0353789
      cosign |   .0097722   .0395825     0.25   0.805    -.0678561    .0874005
       chist |   .1330267   .0246202     5.40   0.000     .0847421    .1813114
      pubrec |  -.2419268   .0427922    -5.65   0.000    -.3258498   -.1580037
    mortlat1 |  -.0572511   .0662234    -0.86   0.387    -.1871269    .0726247
    mortlat2 |  -.1137234   .0910697    -1.25   0.212    -.2923274    .0648806
          vr |  -.0314408   .0144855    -2.17   0.030    -.0598493   -.0030322
       _cons |   .9367312   .0593886    15.77   0.000     .8202595    1.053203
------------------------------------------------------------------------------
```

```
. des hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
hrat            float  %9.0g                  housing exp, % total inc
obrat           float  %9.0g                  other oblgs, % total inc
loanprc         float  %9.0g                  amt/price
unem            float  %9.0g                  unemployment rate by industry
male            byte   %9.0g                  =1 if applicant male
married         byte   %9.0g                  =1 if applicant married
dep             byte   %9.0g                  number of dependents
sch             byte   %9.0g                  =1 if > 12 years schooling
cosign          byte   %9.0g                  is there a cosigner
chist           byte   %9.0g                  =0 if accnts deliq. >= 60 days
pubrec          byte   %9.0g                  =1 if filed bankruptcy
mortlat1        byte   %9.0g                  one or two late payments
mortlat2        byte   %9.0g                  > 2 late payments
vr              byte   %9.0g                  =1 if tract vac rte > MSA med
```

- Yes, there is still evidence for discrimination against non-whites. It is represented by an increase of 12.9 percentage points in the probability of mortgage loan approval for white individuals compared to non-white individuals.
- The coefficient has fallen by some margin because we are now controlling for factors that should affect loan approval rates, and some of these differ by race.
- The race effect is still strong and very significant (*t* statistic = 4.98).

Utrecht University

(iv) Now, create an interaction term: (other obligations * white)

```
. gen white_obrat=white*obrat

. reg approve white hrat obrat white_obrat loanprc unem male married dep sch cosign
chist pubrec mortl
> at1 mortlat2 vr, rob

Linear regression                                      Number of obs =    1971
                                                       F( 16,  1954) =   14.41
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1709
                                                       Root MSE      =  .30119

------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |  -.1459751   .1050932    -1.39   0.165    -.3520816    .0601314
        hrat |   .0017897   .0014702     1.22   0.224    -.0010938    .0046731
       obrat |  -.0122262   .0030209    -4.05   0.000    -.0181507   -.0063017
 white_obrat |   .0080879   .0031094     2.60   0.009     .0019897    .0141861
     loanprc |  -.1525356   .0381022    -4.00   0.000    -.2272607   -.0778105
        unem |  -.0075281   .0036972    -2.04   0.042    -.0147789   -.0002772
        male |  -.0060154   .0191269    -0.31   0.753    -.0435267    .0314958
     married |   .0455358   .0172009     2.65   0.008     .0118018    .0792699
         dep |    -.00763   .0068808    -1.11   0.268    -.0211245    .0058646
         sch |   .0017766   .0171474     0.10   0.917    -.0318526    .0354058
      cosign |   .0177091   .0386821     0.46   0.647    -.0581535    .0935716
       chist |   .1298548   .0245869     5.28   0.000     .0816354    .1780742
      pubrec |   -.240325   .0429733    -5.59   0.000    -.3246034   -.1560467
    mortlat1 |  -.0627819   .0653656    -0.96   0.337    -.1909755    .0654116
    mortlat2 |  -.1268446   .0903701    -1.40   0.161    -.3040764    .0503872
          vr |  -.0305396   .0144395    -2.12   0.035    -.0588579   -.0022212
       _cons |   1.180648   .1106498    10.67   0.000     .9636445    1.397652
------------------------------------------------------------------------------
```

- The white coefficient becomes statistically insignificant, while the interaction variable yields a significant, positive coefficient.
- The interactive effect suggests that the percentage of other obligations mattered less in the approval of mortgage requests by whites than by non-whites.

(v)

```
. nlcom _b[white]+_b[white_obrat]*32

       _nl_1:  _b[white]+_b[white_obrat]*32

------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .1128382   .0255754     4.41   0.000     .0627114     .162965
------------------------------------------------------------------------------
```

- Replace *white* obrat* with *white** (*obrat* – 32); the coefficient on *white* is now the race differential when *obrat* = 32.
- We obtain about .113 and se = .025. So, the 95% confidence interval is about 0.113 ∓ 1.96(0.025) or about 0.063 to 0.162. This interval excludes zero, so at the average *obrat* there is evidence of discrimination (or, at least loan approval rates that differ by race for some other reason that is not captured by the control variables).
- The effect of being white on the probability of successful application is estimated at 11.3% for people with 32% other obligations.

Utrecht University

C2 Use the data in LOANAPP.RAW for this exercise; see also Computer Exercise C8 in Chapter 7.

(i) Estimate a probit model of *approve* on *white*. Find the estimated probability of loan approval for both whites and nonwhites. How do these compare with the linear probability estimates?

(ii) Now, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, and *vr* to the probit model. Is there statistically significant evidence of discrimination against nonwhites?

(iii) Estimate the model from part (ii) by logit. Compare the coefficient on *white* to the probit estimate.

(iv) Use equation (17.17) to estimate the sizes of the discrimination effects for probit and logit.

(i) Estimate the effect of white on approval in a Probit model.

The probit model predicts the probability of loan approval as: $P(Y = 1|X) = \phi(X\beta)$

```
. probit approve white

Iteration 0:  Log likelihood = -740.34659
Iteration 1:  Log likelihood = -701.33221
Iteration 2:  Log likelihood = -700.87747
Iteration 3:  Log likelihood = -700.87744

Probit regression                                    Number of obs =  1,989
                                                     LR chi2(1)    =  78.94
                                                     Prob > chi2   = 0.0000
Log likelihood = -700.87744                          Pseudo R2     = 0.0533

-------------------------------------------------------------------------------
     approve | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+-----------------------------------------------------------------
       white |   .7839465   .0867118     9.04   0.000     .6139946    .9538985
       _cons |   .5469463    .075435     7.25   0.000     .3990964    .6947962
-------------------------------------------------------------------------------

. mfx

Marginal effects after probit
      y  = Pr(approve) (predict)
         =   .8867641
------------------------------------------------------------------------------
variable |      dy/dx    Std. err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  white* |   .2005957      .02685    7.47   0.000   .147968  .253224   .845148
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

```
. predict lin_pred, xb

. list lin_pred in 1/10

     +----------+
     | lin_pred |
     |----------|
  1. | 1.330893 |
  2. | 1.330893 |
  3. | 1.330893 |
  4. | 1.330893 |
  5. | 1.330893 |
     |----------|
  6. | 1.330893 |
  7. | 1.330893 |
  8. | 1.330893 |
  9. | 1.330893 |
 10. | 1.330893 |
     +----------+
```

```
. display normal(1.3308928)
.90838786
```

- As there is only one explanatory variable that takes on just two values, there are only two different predicted values: the estimated probabilities of loan approval for white and nonwhite applicants.

- Rounded to three decimal places, these are .708 for nonwhites and .908 for whites.

- Without rounding errors, these are *identical* to the fitted values from the linear probability model.

- This is the case when the independent variables in a binary response model are mutually exclusive and exhaustive binary variables.

- Then, the predicted probabilities, whether we use the LPM, probit, or logit models, are simply the cell frequencies (in this case, how many loans were approved vs denied for the independent variable: white)

- In other words, 0.708 is the proportion of loans approved for nonwhites and .908 is the proportion approved for whites.

Utrecht University

Additional: We can estimate the model with logit and compare it to the previous one.

```
. logit approve white

Iteration 0:   log likelihood = -740.34659
Iteration 1:   log likelihood =  -709.1878
Iteration 2:   log likelihood =  -700.9007
Iteration 3:   log likelihood = -700.87744
Iteration 4:   log likelihood = -700.87744

Logistic regression                               Number of obs   =       1989
                                                  LR chi2(1)      =      78.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -700.87744                       Pseudo R2       =     0.0533

------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   1.409422   .1511511     9.32   0.000     1.113172    1.705673
       _cons |   .8846854   .1252927     7.06   0.000     .6391162    1.130255
------------------------------------------------------------------------------

. mfx

Marginal effects after logit
      y  = Pr(approve) (predict)
         =   .8885343
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
   white*|   .2005957      .02685    7.47   0.000   .147968  .253224   .845148
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Utrecht University

(ii)

```
. probit approve white hrat obrat loanprc unem male married dep sch cosign chist pubrec mor
> tlat1 mortlat2 vr

Iteration 0:   Log likelihood = -737.97933
Iteration 1:   Log likelihood =  -603.5925
Iteration 2:   Log likelihood = -600.27774
Iteration 3:   Log likelihood = -600.27099
Iteration 4:   Log likelihood = -600.27099

Probit regression                                       Number of obs =  1,971
                                                        LR chi2(15)   = 275.42
                                                        Prob > chi2   = 0.0000
Log likelihood = -600.27099                             Pseudo R2     = 0.1866
```

approve	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
white	.5202525	.0969588	5.37	0.000	.3302168	.7102883
hrat	.0078763	.0069616	1.13	0.258	-.0057682	.0215209
obrat	-.0276924	.0060493	-4.58	0.000	-.0395488	-.015836
loanprc	-1.011969	.2372396	-4.27	0.000	-1.47695	-.5469881
unem	-.0366849	.0174807	-2.10	0.036	-.0709464	-.0024234
male	-.0370014	.1099273	-0.34	0.736	-.2524549	.1784521
married	.2657469	.0942523	2.82	0.005	.0810159	.4504779
dep	-.0495756	.0390573	-1.27	0.204	-.1261266	.0269753
sch	.0146496	.0958421	0.15	0.879	-.1731974	.2024967
cosign	.0860713	.2457509	0.35	0.726	-.3955917	.5677343
chist	.5852812	.0959715	6.10	0.000	.3971805	.7733818
pubrec	-.7787405	.12632	-6.16	0.000	-1.026323	-.5311578
mortlat1	-.1876237	.2531127	-0.74	0.459	-.6837153	.308468
mortlat2	-.4943562	.3265563	-1.51	0.130	-1.134395	.1456823
vr	-.2010621	.0814934	-2.47	0.014	-.3607862	-.041338
_cons	2.062327	.3131763	6.59	0.000	1.448512	2.676141

```
. mfx

Marginal effects after probit
      y  = Pr(approve) (predict)
         =  .91065604
```

variable	dy/dx	Std. err.	z	P>|z|	[95% C.I.	]	X
white*	.105747	.02386	4.43	0.000	.058988	.152506	.846271
hrat	.0012721	.00113	1.13	0.258	-.000933	.003477	24.8001
obrat	-.0044726	.00098	-4.58	0.000	-.006387	-.002558	32.3898
loanprc	-.1634429	.03772	-4.33	0.000	-.237367	-.089519	.770431
unem	-.005925	.00282	-2.10	0.036	-.011456	-.000394	3.88853
male*	-.0058835	.0172	-0.34	0.732	-.039599	.027832	.813293
married*	.045491	.01701	2.68	0.007	.012161	.078821	.659564
dep	-.0080069	.0063	-1.27	0.204	-.020354	.00434	.771689
sch*	.0023787	.01564	0.15	0.879	-.028284	.033042	.770167
cosign*	.0131566	.03547	0.37	0.711	-.056364	.082677	.028919
chist*	.1213625	.0242	5.02	0.000	.073937	.168788	.836631
pubrec*	-.1867903	.04019	-4.65	0.000	-.265569	-.108012	.068493
mortlat1*	-.0341006	.05129	-0.66	0.506	-.134632	.066431	.01928
mortlat2*	-.1075809	.08988	-1.20	0.231	-.283752	.06859	.010654
vr*	-.0333289	.01381	-2.41	0.016	-.06039	-.006268	.407915

(*) dy/dx is for discrete change of dummy variable from 0 to 1

- With the set of controls added, the probit estimate on *white* becomes about .520 (se = .097). Therefore, there is still very strong evidence of discrimination against nonwhites.
- The effect of white is about 10.5 p.p. when calculated around the average approval rate.

(iii) When we use logit instead of probit, the coefficient (standard error) on *white* becomes 0.938 (0.173).

```
. logit approve white hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr, rob

Iteration 0:  Log pseudolikelihood = -737.97933
Iteration 1:  Log pseudolikelihood = -634.97536
Iteration 2:  Log pseudolikelihood = -601.41194
Iteration 3:  Log pseudolikelihood = -600.49724
Iteration 4:  Log pseudolikelihood = -600.49616
Iteration 5:  Log pseudolikelihood = -600.49616

Logistic regression                                    Number of obs =  1,971
                                                       Wald chi2(15) = 210.98
                                                       Prob > chi2   = 0.0000
Log pseudolikelihood = -600.49616                      Pseudo R2     = 0.1863
```

approve	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
white	.9377643	.1747271	5.37	0.000	.5953054	1.280223
hrat	.0132631	.0135997	0.98	0.329	-.0133918	.039918
obrat	-.0530338	.0124078	-4.27	0.000	-.0773526	-.028715
loanprc	-1.904951	.508246	-3.75	0.000	-2.901095	-.9088075
unem	-.0665789	.0353345	-1.88	0.060	-.1358332	.0026755
male	-.0663852	.2068806	-0.32	0.748	-.4718638	.3390934
married	.5032817	.1838261	2.74	0.006	.1429891	.8635743
dep	-.0907336	.0739676	-1.23	0.220	-.2357075	.0542403
sch	.0412287	.1762664	0.23	0.815	-.3042471	.3867046
cosign	.132059	.3918293	0.34	0.736	-.6359124	.9000304
chist	1.066577	.1699995	6.27	0.000	.7333838	1.39977
pubrec	-1.340665	.227446	-5.89	0.000	-1.786451	-.8948791
mortlat1	-.3098821	.5171693	-0.60	0.549	-1.323515	.703751
mortlat2	-.8946755	.5675692	-1.58	0.115	-2.007091	.2177397
vr	-.3498279	.154458	-2.26	0.024	-.6525601	-.0470958
_cons	3.80171	.6333556	6.00	0.000	2.560356	5.043064

```
. mfx

Marginal effects after logit
      y  = Pr(approve) (predict)
         =  .91417919
```

variable	dy/dx	Std. err.	z	P>|z|	[95%	C.I.]	X
white*	.0967431	.02275	4.25	0.000	.052145	.141341	.846271
hrat	.0010406	.00107	0.97	0.330	-.001055	.003136	24.8001
obrat	-.0041608	.00095	-4.38	0.000	-.006021	-.0023	32.3898
loanprc	-.1494541	.03921	-3.81	0.000	-.226303	-.072605	.770431
unem	-.0052235	.00278	-1.88	0.060	-.010667	.00022	3.88853
male*	-.0051197	.01568	-0.33	0.744	-.035861	.025622	.813293
married*	.0423998	.01655	2.56	0.010	.009963	.074837	.659564
dep	-.0071186	.0058	-1.23	0.220	-.01849	.004253	.771689
sch*	.0032647	.01408	0.23	0.817	-.024335	.030865	.770167
cosign*	.0098414	.02772	0.36	0.723	-.044483	.064166	.028919
chist*	.1133208	.02299	4.93	0.000	.068255	.158386	.836631
pubrec*	-.1676967	.04081	-4.11	0.000	-.247682	-.087712	.068493
mortlat1*	-.0275065	.0516	-0.53	0.594	-.128634	.073621	.01928
mortlat2*	-.1002576	.08511	-1.18	0.239	-.26707	.066555	.010654
vr*	-.02826	.01296	-2.18	0.029	-.053654	-.002866	.407915

(*) dy/dx is for discrete change of dummy variable from 0 to 1

With a logit model, we obtain a bit lower estimate for the effect of white (9.7 p.p.), but it is still large, positive, and statistically significant.

(iv)

- Recall that, to make probit and logit estimates roughly comparable, we can multiply the logit estimates by 0.625.
- The scaled logit coefficient becomes .625(.938) = .586, which is reasonably close to the probit estimate.
- A better comparison would be to compare the predicted probabilities by setting the other controls at interesting values, such as their average values in the sample.