CHAPTER 8 EXERCISES

8.1. To study the effectiveness of price discount on a six-pack of soft drink, a sample of 5500 consumers was randomly assigned to eleven discount categories as shown in (Table 8.9).

Table 8.9 The number of coupons redeemed and the price discount.

Price Discount (cents)	Sample size	Number of coupons redeemed
5	500	100
7	500	122
9	500	147
11	500	176
13	500	211
15	500	244
17	500	277
19	500	310
21	500	343
23	500	372
25	500	391

(*a*) Treating the redemption rate as the dependent variable and price discount as the regressor, see if the logit model fits the data.

Results using logit (weighted least squares):

```
. glogit redeemed ssize discount

Weighted LS logistic regression for grouped data

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =22943.74
       Model |  7.07263073     1  7.07263073           Prob > F      =  0.0000
    Residual |  .002774338     9   .00030826           R-squared     =  0.9996
-------------+------------------------------           Adj R-squared =  0.9996
       Total |  7.07540507    10  .707540507           Root MSE      =  .01756

------------------------------------------------------------------------------
    redeemed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |   .1357406   .0008961   151.47   0.000     .1337134    .1377678
       _cons |  -2.084928   .0145341  -143.45   0.000    -2.117807    -2.05205
------------------------------------------------------------------------------
```

Results using logit (maximum likelihood) are very similar:

```
. blogit redeemed ssize discount

Logistic regression for grouped data              Number of obs   =       5500
                                                  LR chi2(1)      =     870.93
                                                  Prob > chi2     =     0.0000
```

```
Log likelihood = -3375.6653                              Pseudo R2       =     0.1143

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |   .1357274   .0049571    27.38   0.000     .1260117     .145443
       _cons |  -2.084754   .0803976   -25.93   0.000    -2.242331   -1.927178
------------------------------------------------------------------------------
```

(*b*) See if the probit model does as well as the logit model.

Grouped probit (weighted least squares) gives the following:

```
. gprobit redeemed ssize discount

Weighted LS probit regression for grouped data

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =13260.16
       Model |  2.81240776     1  2.81240776           Prob > F      =  0.0000
    Residual |  .001908851     9  .000212095           R-squared     =  0.9993
-------------+------------------------------           Adj R-squared =  0.9992
       Total |  2.81431662    10  .281431662           Root MSE      =  .01456

------------------------------------------------------------------------------
    redeemed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |   .0832431   .0007229   115.15   0.000     .0816078    .0848784
       _cons |  -1.278202   .0117437  -108.84   0.000    -1.304768   -1.251636
------------------------------------------------------------------------------
```

Maximum likelihood results are similar:

```
. bprobit redeemed ssize discount;

Probit regression for grouped data                Number of obs   =       5500
                                                  LR chi2(1)      =     870.67
                                                  Prob > chi2     =     0.0000
Log likelihood =  -3375.794                       Pseudo R2       =     0.1142

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |   .0832308    .002921    28.49   0.000     .0775058    .0889558
       _cons |  -1.278027   .0474529   -26.93   0.000    -1.371033   -1.185021
------------------------------------------------------------------------------
```

(*c*) Fit the LPM model to these data.

```
. reg rrate discount

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) = 3112.95
       Model |   .41469561     1   .41469561           Prob > F      =  0.0000
    Residual |  .001198946     9  .000133216           R-squared     =  0.9971
-------------+------------------------------           Adj R-squared =  0.9968
       Total |  .415894556    10  .041589456           Root MSE      =  .01154

------------------------------------------------------------------------------
       rrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    discount |      .0307   .0005502    55.79   0.000     .0294553    .0319447
       _cons |   .0291364   .0089573     3.25   0.010     .0088736    .0493991
------------------------------------------------------------------------------
```

(*d*) Compare the results of the three models. Note that the coeffi cients of LPM and Logit models are related as follows:

Slope coeffi cient of LPM = 0.25* Slope coeffi cient of Logit

Intercept of LPM = 0.25* slope coeffi cient of Logit + 0.5.

The results are very similar. Since LPM = 0.25*Logit, we have 0.25*0.1357406 = 0.0339, similar to the LPM value of 0.0307 that we obtain. We expect the logit coefficient to be approximately equal to 1.81 multiplied by the probit coefficient: 1.81*0.832431 = 0.1507, which is somewhat comparable to the logit value we obtain.

8.2. Table 8.10 (available on the companion website) gives data on 78 homebuyers on their choice between adjustable and fixed rate mortgages and related data bearing on the choice. The variables are defi ned as follows:

***Adjust* = 1 if an adjustable mortgage is chosen, 0 otherwise.**
***Fixed rate* = fi xed interest rate**
***Margin* = (variable rate – fixed rate)**
***Yield* = the 10-year Treasury rate less 1-year rate**
***Points* = ratio of points on adjustable mortgage to those paid on a fixed rate mortgage**
***Networth* = borrower's net worth**

(*a*) Estimate an LPM of adjustable rate mortgage choice.

```
. reg adjust fixrate margin maturity networth points yield

      Source |       SS       df       MS              Number of obs =      78
-------------+------------------------------           F(  6,    71) =    5.45
       Model |  5.94768128     6  .991280213           Prob > F      =  0.0001
    Residual |  12.9241136    71  .182029769           R-squared     =  0.3152
-------------+------------------------------           Adj R-squared =  0.2573
       Total |  18.8717949    77  .245088245           Root MSE      =  .42665

------------------------------------------------------------------------------
      adjust |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     fixrate |   .1603915   .0822031     1.95   0.055    -.0035167    .3242998
      margin |  -.1318021    .049831    -2.64   0.010    -.2311623    -.032442
    maturity |  -.0341354   .1907662    -0.18   0.858    -.4145124    .3462417
    networth |   .0288939   .0117867     2.45   0.017     .0053917     .052396
      points |  -.0887104   .0711305    -1.25   0.216    -.2305405    .0531197
       yield |  -.7932019   .3234705    -2.45   0.017    -1.438184     -.14822
       _cons |  -.0707747   1.287665    -0.05   0.956    -2.638306    2.496757
------------------------------------------------------------------------------
```

(*b*) Estimate the adjustable rate mortgage choice using logit.

```
. logit adjust fixrate margin maturity networth points yield

Iteration 0:   log likelihood = -52.802235
Iteration 1:   log likelihood = -39.614778
Iteration 2:   log likelihood = -39.046815
Iteration 3:   log likelihood = -39.035313
Iteration 4:   log likelihood = -39.035305

Logistic regression                               Number of obs   =         78
                                                  LR chi2(6)      =      27.53
                                                  Prob > chi2     =     0.0001
Log likelihood = -39.035305                       Pseudo R2       =     0.2607
```

```
------------------------------------------------------------------------------
      adjust |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     fixrate |   .8957191   .4859245     1.84   0.065    -.0566754    1.848114
      margin |  -.7077102   .3035058    -2.33   0.020    -1.302571   -.1128497
    maturity |  -.2370469   1.039279    -0.23   0.820    -2.273997    1.799903
    networth |   .1504304   .0787145     1.91   0.056    -.0038473     .304708
      points |   -.521043   .4263876    -1.22   0.222    -1.356747    .3146614
       yield |  -4.105524   1.902219    -2.16   0.031    -7.833805   -.3772429
       _cons |  -3.647767   7.249959    -0.50   0.615    -17.85742    10.56189
------------------------------------------------------------------------------
```

(*c*) Repeat (*b*) using the probit model.

```
. probit adjust fixrate margin maturity networth points yield

Iteration 0:   log likelihood = -52.802235
Iteration 1:   log likelihood = -39.570168
Iteration 2:   log likelihood = -39.208823
Iteration 3:   log likelihood = -39.207128
Iteration 4:   log likelihood = -39.207128

Probit regression                                 Number of obs   =         78
                                                  LR chi2(6)      =      27.19
                                                  Prob > chi2     =     0.0001
Log likelihood = -39.207128                       Pseudo R2       =     0.2575

------------------------------------------------------------------------------
      adjust |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     fixrate |   .4987284   .2624758     1.90   0.057    -.0157148    1.013172
      margin |  -.4309509   .1739101    -2.48   0.013    -.7718083   -.0900934
    maturity |  -.0591854   .6225826    -0.10   0.924    -1.279425    1.161054
    networth |   .0838286    .037854     2.21   0.027     .0096361    .1580211
      points |  -.2999138   .2413875    -1.24   0.214    -.7730246    .1731971
       yield |  -2.383964   1.083047    -2.20   0.028    -4.506698   -.2612297
       _cons |  -1.877266   4.120677    -0.46   0.649    -9.953644    6.199112
------------------------------------------------------------------------------
```

(*d*) Compare the performance of the three models and decide which is a better model.

All three models yield results which are comparable, yet results for logit and probit are more similar. Since we have a dichotomous dependent variable, we should probably opt for the probit or the logit model rather than the LPM model. Since the pseudo R^2 for logit is slightly higher, we may be tempted to choose logit over probit in this case.

(*e*) Calculate the marginal impact of Margin on the probability of choosing the adjustable rate mortgage for the three models.

The marginal effects at the mean are very similar across all three models, with the results for logit and probit almost identical:

```
. *Marginal effect at mean for "margin" (LPM)
. mfx, var(margin)

Marginal effects after regress
      y  = Fitted values (predict)
         =  .41025641
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  margin |  -.1318021      .04983   -2.64   0.008  -.229469 -.034135   2.29192
------------------------------------------------------------------------------
. *Marginal effect at mean for "margin" (logit)
. mfx, var(margin)
```

```
Marginal effects after logit
      y  = Pr(adjust) (predict)
         =  .37718898
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  margin |  -.1662535      .07152   -2.32   0.020  -.306432 -.026075   2.29192
------------------------------------------------------------------------------
. *Marginal effect at mean for "margin" (probit)
. mfx, var(margin)

Marginal effects after probit
      y  = Pr(adjust) (predict)
         =  .38021288
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  margin |  -.1641149      .06634   -2.47   0.013  -.294146 -.034083   2.29192
------------------------------------------------------------------------------
```

8.3. For the smoker data discussed in the chapter, estimate the count R^2.

The count R^2 is equal to the number of correct predictions divided by the total number of observations, where the number of correct predictions is calculated by summing up observations for which the predicted probability is within 0.5 of the actual dichotomous value for "smoker" (0,1). In other words, probabilities of 0.5 or greater were interpreted as "1" and probabilities of less than 0.5 were interpreted as "0" and compared with actual "smoker" values. By this definition, the count R^2 is 730 out of 1196, or 0.6104.

8.4. Divide the smoker data into 20 groups. For each group compute p_i, the probability of smoking. For each group compute the average values of the regressors and estimate the grouped logit model using these average values. Compare your results with the ML estimates of smoker logit discussed in the chapter. How would you obtain the heteroscedasticity-corrected standard errors for the grouped logit?

Results are:

```
. glogit smoke samp age educ income pcigs79

Weighted LS logistic regression for grouped data

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  4,    15) =    0.35
       Model |  .125649254     4  .031412313           Prob > F      =  0.8426
    Residual |  1.36133328    15  .090755552           R-squared     =  0.0845
-------------+------------------------------           Adj R-squared = -0.1596
       Total |  1.48698254    19  .078262239           Root MSE      =  .30126

------------------------------------------------------------------------------
       smoke |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0301689    .032537    -0.93   0.368      -.09952    .0391821
        educ |   .0782033   .1822493     0.43   0.674    -.3102518    .4666584
      income |  -2.01e-06   .0000564    -0.04   0.972    -.0001221    .0001181
     pcigs79 |   .0009161   .0218406     0.04   0.967     -.045636    .0474681
       _cons |  -.2333753   2.901647    -0.08   0.937     -6.41809    5.951339
------------------------------------------------------------------------------
```

And for ML method:

```
. blogit smoke samp age educ income pcigs79

Logistic regression for grouped data              Number of obs   =       1200
```

```
                                                  LR chi2(4)      =       1.85
                                                  Prob > chi2     =     0.7629
Log likelihood = -792.43701                       Pseudo R2       =     0.0012

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0311982    .028837    -1.08   0.279    -.0877177    .0253213
        educ |   .0777483   .1615039     0.48   0.630    -.2387936    .3942902
      income |  -2.34e-06   .0000499    -0.05   0.963    -.0001002    .0000955
     pcigs79 |   .0006726   .0193552     0.03   0.972    -.0372628     .038608
       _cons |  -.1569088   2.571068    -0.06   0.951    -5.196109    4.882292
------------------------------------------------------------------------------
```

Results are comparable to non-grouped results, although standard errors likely need to be adjusted for heteroscedasticity using the *robust* option in Stata.

8.5. Table 8.11 on the companion website gives hypothetical data on admission to graduate school. The variables are defined as follows:

Admit = 1, if admitted to graduate school, 0 otherwise
GRE = graduate record examination score
GPA= grade point average
Rank of the graduating school, 1, 2, 3, 4; 1 is the best and 4 is the worst

(_a_) Develop a suitable logit model for admission to graduate school and estimate the parameters of the model.

Results are as follows:

```
. logit admit gre gpa rank

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -230.08375
Iteration 2:   log likelihood = -229.72097
Iteration 3:   log likelihood = -229.72088
Iteration 4:   log likelihood = -229.72088

Logistic regression                               Number of obs   =        400
                                                  LR chi2(3)      =      40.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -229.72088                       Pseudo R2       =     0.0811

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |    .002294   .0010918     2.10   0.036      .000154    .0044339
         gpa |   .7770137   .3274839     2.37   0.018     .1351571     1.41887
        rank |  -.5600314    .127137    -4.40   0.000    -.8092153   -.3108475
       _cons |  -3.449549   1.132846    -3.05   0.002    -5.669886   -1.229211
------------------------------------------------------------------------------
```

(_b_) How would you interpret the various coefficients, especially of the rank variable?

We can see that the higher the GRE score, the higher the GPA, and the higher the rank (denoted as a lower numerical value), the higher the predicted probability that a person is admitted to graduate school. Yet it is more useful to interpret the numerical values of the marginal effects at the means:

```
. mfx [Note: gives same result as following command: margins, dydx(*) atmeans]
```

```
Marginal effects after logit
      y  = Pr(admit) (predict)
         =  .29753409
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     gre |   .0004795      .00023    2.11   0.035   .000034  .000925     587.7
     gpa |   .1624017      .06811    2.38   0.017   .028906  .295897    3.3899
    rank |  -.1170508       .0261   -4.49   0.000  -.168197 -.065904     2.485
------------------------------------------------------------------------------
```

Here we see that, as the GPA goes up by one point, the predicted probability of being admitted to graduate school goes up by 16.24 percentage points, *ceteris paribus*.

(*c*) Obtain the various odds ratios.

The odds ratios are:

```
. logit admit gre gpa rank, or

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -230.08375
Iteration 2:   log likelihood = -229.72097
Iteration 3:   log likelihood = -229.72088
Iteration 4:   log likelihood = -229.72088

Logistic regression                               Number of obs   =        400
                                                  LR chi2(3)      =      40.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -229.72088                       Pseudo R2       =     0.0811

------------------------------------------------------------------------------
       admit | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   1.002297   .0010943     2.10   0.036     1.000154    1.004444
         gpa |   2.174967   .7122668     2.37   0.018     1.144717    4.132449
        rank |   .5711911   .0726195    -4.40   0.000     .4452073    .7328256
------------------------------------------------------------------------------
```

(*d*) Repeat your analysis using the probit model.

In Stata, one can obtain the marginal effects right away using the “dprobit” command:

```
. dprobit admit gre gpa rank

Iteration 0:   log likelihood = -249.98826
Iteration 1:   log likelihood = -229.93029
Iteration 2:   log likelihood = -229.74047
Iteration 3:   log likelihood =  -229.7404

Probit regression, reporting marginal effects           Number of obs =    400
                                                        LR chi2(3)    =  40.50
                                                        Prob > chi2   = 0.0000
Log likelihood =  -229.7404                             Pseudo R2     = 0.0810

------------------------------------------------------------------------------
   admit |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
     gre |   .0004873   .0002252     2.16   0.031     587.7   .000046  .000929
     gpa |   .1618311   .0673824     2.40   0.017    3.3899   .029764  .293898
    rank |  -.1156027    .025725    -4.47   0.000     2.485  -.166023 -.065183
---------+--------------------------------------------------------------------
  obs. P |      .3175
 pred. P |    .301553  (at x-bar)
```

```
------------------------------------------------------------------------------
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Similarly to the logit model, these marginal effects tell us, for example, that the predicted probability of being admitted to graduate school goes up by 16.18 percentage points, *ceteris paribus*, as GPA goes up by one point.

8.6 . Table 8.12 on the companion website provides data on heart attack within 48 hours of myocardial infarction onset. This is a large data set consisting of 4,483 observations. The variables used in the analysis are as follows:

***death* = 1, if within 48 hours of myocardial infarction onset, 0 otherwise.**
***anterior* =1 , anterior infarction**
***anterior* = 0, inferior infarction**
***hcabg* = 1 history of CABG (history of having had a cardiac bypass surgery)**
***hcabg* = no history of CABG**
***kk3* = killip class 3**
***kk4* = killip class 4**

(*a*) Estimate a probit model for death, obtaining the usual statistics.

The marginal effects are (note *kk1* and *age1* are dropped to avoid the dummy variable trap):

```
. dprobit  death anterior hcabg kk2 kk3 kk4 age2 age3 age4

Iteration 0:   log likelihood = -742.31027
Iteration 1:   log likelihood = -642.02785
Iteration 2:   log likelihood = -634.39268
Iteration 3:   log likelihood = -634.31308
Iteration 4:   log likelihood = -634.31304

Probit regression, reporting marginal effects           Number of obs =   4483
                                                        LR chi2(8)    = 215.99
                                                        Prob > chi2   = 0.0000
Log likelihood = -634.31304                             Pseudo R2     = 0.1455

------------------------------------------------------------------------------
   death |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
anterior*|    .017684   .0046975     3.92   0.000   .451483   .008477  .026891
   hcabg*|   .0272408   .0181741     1.97   0.049   .031229   -.00838  .062861
     kk2*|   .0268356   .0074588     4.40   0.000   .197859   .012217  .041455
     kk3*|   .0333861   .0142946     3.15   0.002   .051528   .005369  .061403
     kk4*|   .2636457   .0657135     7.26   0.000   .010707    .13485  .392442
    age2*|   .0113497   .0084633     1.45   0.148   .261209 -.005238  .027937
    age3*|   .0514412   .0109224     5.88   0.000   .258309   .030034  .072849
    age4*|    .118808   .0209923     8.61   0.000   .120678   .077664  .159952
---------+--------------------------------------------------------------------
  obs. P |   .0392594
 pred. P |   .0240731  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

(*b*) Obtain the odds ratios and interpret them.

The odds ratios after a **logit** model are:

```
. logit  death anterior hcabg kk2 kk3 kk4 age2 age3 age4, or
```

```
Iteration 0:   log likelihood = -742.31027
Iteration 1:   log likelihood = -667.44279
Iteration 2:   log likelihood = -637.67555
Iteration 3:   log likelihood = -636.62802
Iteration 4:   log likelihood = -636.62553
Iteration 5:   log likelihood = -636.62553

Logistic regression                               Number of obs   =       4483
                                                  LR chi2(8)      =     211.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -636.62553                       Pseudo R2       =     0.1424

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    anterior |   1.901333   .3185757     3.83   0.000     1.369103    2.640464
       hcabg |   2.105275   .7430694     2.11   0.035     1.054076    4.204801
         kk2 |   2.251732   .4064423     4.50   0.000     1.580786    3.207453
         kk3 |   2.172105    .584427     2.88   0.004     1.281907    3.680487
         kk4 |   14.29137   5.087654     7.47   0.000     7.112964    28.71423
        age2 |    1.63726   .5078582     1.59   0.112     .8914261    3.007115
        age3 |   4.532029   1.206534     5.68   0.000     2.689568    7.636647
        age4 |   8.893222    2.41752     8.04   0.000     5.219991    15.15125
------------------------------------------------------------------------------
```

These results suggest that the odds of death within 48 hours of myocardial infarction onset are 1.90 times larger for those with an anterior infarction than those with an inferior infarction, *ceteris paribus*. Moreover, the odds of death are 2.11 times larger for those with a history of HCABG, *ceteris paribus*. Those who are older and at more risk also have higher odds of death.

(*c*) Obtain the probability of death for each observation. (You may use *Stata's* command: predict mu).

This was done in Stata, with the means shown as follows:

```
. predict mu
(option pr assumed; Pr(death))
(905 missing values generated)

. su death mu

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |      5388    .0449146    .2071359          0          1
          mu |      4483    .0392594      .05449   .0063554   .6071695

. su death mu if mu!=.

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |      4483    .0392594    .1942332          0          1
          mu |      4483    .0392594      .05449   .0063554   .6071695
```

8.7 Direct marketing for financial products (DMF): Table 8.13 on the companion website gives data on the response of customers of a commercial bank to direct marketing campaign for a new financial product. The variables are as follows:

> ***Response* =1 if customer invests in the new product, 0 otherwise**
> ***Invest* = amount of money invested by the customer in the new product ('00 Dutch guilders)**
> ***Gender* = 1 for males, 0 for females**

***Activity* = activity indicator, 1 if customer already invests in other products of the bank, 0 otherwise**
***Age* = age of customer, in years**

(*a*) Develop an appropriate logit or probit model for the *Response* variable and interpret the results.

The following probit marginal effects are obtained (note that we cannot include the variable *invest* since there is only a value for this if individuals have invested in the product—we will use this variable in Exercise 11.4):

```
dprobit response invest gender activity age

outcome = invest > 0 predicts data perfectly

. dprobit response gender activity age

Iteration 0:   log likelihood = -641.03952
Iteration 1:   log likelihood = -604.07414
Iteration 2:   log likelihood = -603.96753
Iteration 3:   log likelihood = -603.96753

Probit regression, reporting marginal effects           Number of obs =    925
                                                        LR chi2(3)    =  74.14
                                                        Prob > chi2   = 0.0000
Log likelihood = -603.96753                             Pseudo R2     = 0.0578

------------------------------------------------------------------------------
response |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
  gender*|   .2383015   .0357268     6.36   0.000   .725405   .168278  .308325
activity*|   .2215124   .0403452     5.15   0.000   .188108   .142437  .300587
     age |   -.000291   .0012572    -0.23   0.817   50.6811  -.002755  .002173
---------+--------------------------------------------------------------------
  obs. P |   .5081081
 pred. P |   .5084628  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

The results suggest that the predicted probability of investing in the new product for males is 23.83 percentage points higher than that for females, *ceteris paribus*. Moreover, those who invest in other products (activity is higher) and those who are younger (although this is not significant) are more likely to invest in the new product.

(*b*) Since the data are cross-sectional, how would you handle the problem of heteroscedasticity?

I would address this by obtaining robust standard errors:

```
. dprobit response gender activity age, robust

Iteration 0:   log pseudolikelihood = -641.03952
Iteration 1:   log pseudolikelihood = -604.07414
Iteration 2:   log pseudolikelihood = -603.96753
Iteration 3:   log pseudolikelihood = -603.96753

Probit regression, reporting marginal effects           Number of obs =    925
                                                        Wald chi2(3)  =  70.48
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -603.96753                       Pseudo R2     = 0.0578

------------------------------------------------------------------------------
         |               Robust
response |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
```

```
  gender*|   .2383015    .035683     6.37   0.000   .725405   .168364  .308239
activity*|   .2215124   .0404862     5.13   0.000   .188108   .142161  .300864
     age |   -.000291    .001271    -0.23   0.819   50.6811  -.002782    .0022
---------+--------------------------------------------------------------------
  obs. P |   .5081081
 pred. P |   .5084628  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

(*c*) Instead of coding the gender variable 1 for male and 0 for female, how would the result change if female were coded as 1 and male as 0? Do you have to reestimate your model? Explain why or why not?

The coefficient on gender would simply be the opposite sign, so no, one does not have to reestimate the model. If we did, the results would be:

```
. g female=(gender==0)

. dprobit response female activity age

Iteration 0:   log likelihood = -641.03952
Iteration 1:   log likelihood = -604.07414
Iteration 2:   log likelihood = -603.96753
Iteration 3:   log likelihood = -603.96753

Probit regression, reporting marginal effects           Number of obs =    925
                                                        LR chi2(3)    =  74.14
                                                        Prob > chi2   = 0.0000
Log likelihood = -603.96753                             Pseudo R2     = 0.0578

------------------------------------------------------------------------------
response |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
  female*|  -.2383015   .0357268    -6.36   0.000   .274595  -.308325 -.168278
activity*|   .2215124   .0403452     5.15   0.000   .188108   .142437  .300587
     age |   -.000291   .0012572    -0.23   0.817   50.6811  -.002755  .002173
---------+--------------------------------------------------------------------
  obs. P |   .5081081
 pred. P |   .5084628  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

These results show that the sign has simply been flipped; the interpretation is exactly the same.

(*d*) Suppose you add a new variable to the model, *Gender* x *Age*, that is the interaction between the explanatory variables *Gender* and *sex*. Reestimate your model and comment on the results.

```
. g gender_age = gender*age
(75 missing values generated)

. dprobit response gender activity age gender_age

Iteration 0:   log likelihood = -641.03952
Iteration 1:   log likelihood = -604.04015
Iteration 2:   log likelihood = -603.93244
Iteration 3:   log likelihood = -603.93243

Probit regression, reporting marginal effects           Number of obs =    925
                                                        LR chi2(4)    =  74.21
                                                        Prob > chi2   = 0.0000
Log likelihood = -603.93243                             Pseudo R2     = 0.0579

------------------------------------------------------------------------------
```

```
response |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
  gender*|    .271608   .1291087     1.98   0.048   .725405    .01856  .524656
activity*|   .2213966   .0403534     5.15   0.000   .188108   .142305  .300488
     age |   .0001867   .0021973     0.08   0.932   50.6811   -.00412  .004493
gender~e |  -.0007099   .0026791    -0.26   0.791   36.7362  -.005961  .004541
---------+--------------------------------------------------------------------
  obs. P |   .5081081
 pred. P |    .508468  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

The interaction term is not statistically significant.

8.8 To find out if adolescents (ages 15 and 16) ever had sexual intercourse (yes/no), Morgan and Teachman studied a sample of 342 adolescents from the *National Survey of Children*, 134 white males, 149 white females, 23 black males and 36 black females and obtained the following results from a logistic regression: The underlying model is:

$$\ln \frac{P_i}{1-P_i} = B_1 + B_2 White_i + B_3 Female_i + u_i$$ **, where P_i = probability of sexual intercourse**

Variable	**Slope coefficient**	**se of slope coefficient**	***p* value**
White	**-1.314**	**0.226**	**0.000**
Female	**-0.648**	**0.225**	**0.004**
Constant	**0.192**	**0.226**	**0.365**
LR statistic 37.459, df = 2			

***Note*: All the regressor are dummy variables. The base or comparison categories are blacks and males, which takes values of 0.**

(*a*) How would you interpret the various coefficients?

Coefficient on white: The average logit value, or the log of the odds in favor of having sexual intercourse, for whites is 1.314 units lower, *ceteris paribus*.
Coefficient on female: The average logit value, or the log of the odds in favor of having sexual intercourse, for females is 0.648 units lower, *ceteris paribus*.

(*b*) Are the estimated slope coefficients individually statistically significant? How can you tell?

Yes, both coefficients on white and female are individually statistically significant at the 1% level, since the p-values (at 0.000 and 0.004, respectively) are both lower than 0.01.

(*c*) Can you compute the odds ratios from the estimated slopes? Show the necessary calculations.

The odds ratios are:
For white: $e^{-1.314}$ = 0.269. For female: $e^{-0.648}$ = 0.523.

(*d*) How would you interpret the odds ratios obtained in (*c*)?

For ***white***: The odds of having sexual intercourse are 3.717 (=1/0.269) larger for blacks than for whites, *ceteris paribus*. For ***female***: The odds of having sexual intercourse are 1.912 (=1/0.523) larger for males than for females, *ceteris paribus*.

(*e*) Suppose you change the assignments of the dummies, letting blacks and male take the value 1 instead of 0. Do you have to repeat the analysis or can you get this information from the results presented above? (Hint: Change the sign).

No, you would not have to repeat the above analysis. The slope coefficients would simply be the opposite sign. The slope coefficient for *black* would be 1.314, and the slope coefficient for *male* would be 0.648. The odds ratio, therefore, for *black* would be $e^{1.314}$ = 3.72, and the odds ratio for *male* is $e^{0.648}$ = 1.91; these are the odds ratios we obtain in part d when interpreting the odds ratios.

8.9 *President Clinton's Impeachment Trial*: On January 7, 1999, The U.S. House of Representatives impeached President Clinton on two counts, called Article 1 and Article 2. Article 1 was perjury to grand jury and Article 2 was obstruction of justice. By law, it is the duty of the U.S Senate to conduct a trial on these two counts, which was held on February 12, 1999. On Article1, the vote for 45 yes and 55 no, and on Article 2 the vote was 50 yes and 50 no. However, to remove the President from office, two-third votes are needed, which meant an affirmative vote of 67 senators in a body of 100 senators. Table 8.14 on the companion website provides some interesting data on the impeachment vote, Yes or No, such as the party affiliation of the senators, political ideology of individual senator, number of impeachment votes cast (maximum of 2) cast by the senator, whether a first term senator, the percent of vote Clinton received in 1996 in each senator's state, and the next election of the senator. A U.S. Senator is elected for a term of 6 years, at the end of which the senator may choose to run again.

(*a*) Estimate a probit model of the vote on Article 1 of impeachment in relation to the regressors and discuss your results. The dependent variable is either Yes or No.

The results are as follows:

```
. probit art1vote firstterm ideology nextele partyaff pctvote

note: partyaff != 1 predicts failure perfectly
      partyaff dropped and 45 obs not used

Iteration 0:   log likelihood = -26.077662
Iteration 1:   log likelihood = -17.527199
Iteration 2:   log likelihood =   -17.4581
Iteration 3:   log likelihood = -17.457906
Iteration 4:   log likelihood = -17.457906

Probit regression                                 Number of obs   =         55
                                                  LR chi2(4)      =      17.24
                                                  Prob > chi2     =     0.0017
Log likelihood = -17.457906                       Pseudo R2       =     0.3305

------------------------------------------------------------------------------
    art1vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   firstterm |   .2019032   .5187073     0.39   0.697    -.8147444    1.218551
    ideology |   .0424977   .0148557     2.86   0.004      .013381    .0716143
     nextele |  -.0489844   .1564929    -0.31   0.754     -.355705    .2577361
    partyaff |  (omitted)
     pctvote |  -.0369057   .0419618    -0.88   0.379    -.1191494     .045338
       _cons |   97.48117    313.326     0.31   0.756    -516.6266    711.5889
------------------------------------------------------------------------------
```

Note that the variable “party affiliation” (equal to 1 if the senator’s party affiliation is Republican, 0 if Democrat) has been dropped since all Democrats voted “no” for Article 1 (perjury to grand jury). The results suggest that first-term senators were more likely to vote “yes” for Article 1, as were those with higher political ideology. Those whose next election was in a later year were less likely to vote yes, and those senators in states where the percent of vote Clinton received in 1996 was higher were also less likely to vote yes.

(*b*) Estimate a probit model of vote on Article 2 of impeachment, using the same regressors as in (*a*) and discuss your results. Again, the dependent variable is either Yes or No.

Running the regression including ideology does not work since those with political ideology > 48 all voted “yes” to Article 2 (obstruction of justice). The results excluding this variable are as follows:

```
. probit art2vote firstterm ideology nextele partyaff pctvote

outcome = ideology > 48 predicts data perfectly

. probit art2vote firstterm nextele partyaff pctvote

note: partyaff != 1 predicts failure perfectly
      partyaff dropped and 45 obs not used

Iteration 0:   log likelihood = -16.754985
Iteration 1:   log likelihood = -10.956209
Iteration 2:   log likelihood = -9.1360912
Iteration 3:   log likelihood =  -9.074461
Iteration 4:   log likelihood = -9.0740673
Iteration 5:   log likelihood = -9.0740673

Probit regression                                 Number of obs   =         55
                                                  LR chi2(3)      =      15.36
                                                  Prob > chi2     =     0.0015
Log likelihood = -9.0740673                       Pseudo R2       =     0.4584

------------------------------------------------------------------------------
    art2vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   firstterm |    .375053   .7027446     0.53   0.594    -1.002301    1.752407
     nextele |  -.0347792   .2041046    -0.17   0.865    -.4348169    .3652585
    partyaff |  (omitted)
     pctvote |  -.3267312   .1437732    -2.27   0.023    -.6085216   -.0449409
       _cons |   86.94196   410.3336     0.21   0.832    -717.2972    891.1811
------------------------------------------------------------------------------
Note: 0 failures and 7 successes completely determined.
```

These results again suggest that *firstterm* is associated with a greater probability of voting yes to Article 2, while *nextele* and *pctvote* are both associated with a lower probability of voting yes. Party affiliation has again been dropped.

(*c*) Since a senator’s vote on the impeachment on the two counts are probably going to be the same because of political ideology and party politics, it may be possible to estimate a bivariate probit model to take into account the interdependence of the two votes. Using the bivariate probit procedures in Stata and Eviews, estimate a bivariate probit model of the impeachment trial. What do the results show?

The results are as follows:

```
. biprobit art1vote art2vote firstterm nextele pctvote

Fitting comparison equation 1:

Iteration 0:   log likelihood = -68.813881
Iteration 1:   log likelihood = -55.983011
Iteration 2:   log likelihood = -55.877103
Iteration 3:   log likelihood = -55.876966
Iteration 4:   log likelihood = -55.876966

Fitting comparison equation 2:

Iteration 0:   log likelihood = -69.314718
Iteration 1:   log likelihood = -54.607996
Iteration 2:   log likelihood = -54.543999
Iteration 3:   log likelihood = -54.543961
Iteration 4:   log likelihood = -54.543961

Comparison:    log likelihood = -110.42093

Fitting full model:

Iteration 0:   log likelihood = -110.42093
Iteration 1:   log likelihood = -73.765837
Iteration 2:   log likelihood = -70.597551
Iteration 3:   log likelihood = -70.080737
Iteration 4:   log likelihood = -70.011091
Iteration 5:   log likelihood = -69.997018
Iteration 6:   log likelihood = -69.994577
Iteration 7:   log likelihood = -69.993695
Iteration 8:   log likelihood =  -69.99352
Iteration 9:   log likelihood = -69.993501
Iteration 10:  log likelihood = -69.993499

Bivariate probit regression                     Number of obs   =        100
                                                Wald chi2(6)    =      24.07
Log likelihood = -69.993499                     Prob > chi2     =     0.0005

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
art1vote     |
   firstterm |   .6159445   .2879016     2.14   0.032     .0516678    1.180221
     nextele |  -.0538821   .0903506    -0.60   0.551     -.230966    .1232019
     pctvote |  -.0994973   .0257681    -3.86   0.000    -.1500017   -.0489928
       _cons |   112.1952   181.1102     0.62   0.536    -242.7742    467.1646
-------------+----------------------------------------------------------------
art2vote     |
   firstterm |    .565959   .2923499     1.94   0.053    -.0070363    1.138954
     nextele |  -.0784972   .0907713    -0.86   0.387    -.2564057    .0994114
     pctvote |  -.1163107    .027241    -4.27   0.000    -.1697021   -.0629193
       _cons |   162.4363   182.0133     0.89   0.372    -194.3032    519.1758
-------------+----------------------------------------------------------------
     /athrho |   7.819008   227.4331     0.03   0.973    -437.9417    453.5798
-------------+----------------------------------------------------------------
         rho |   .9999997    .000147                           -1           1
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0:     chi2(1) =  80.8549    Prob > chi2 = 0.0000
```

These results show the expected signs, with a very high and significant value for rho (the estimate of the correlation of the errors) of almost 1, suggesting that unobserved factors that make it more likely to vote “yes” for Article 1 also make it more likely to vote “yes” for Article 2.