Statistics/Data Analysis

```
      name:  <unnamed>
       log:  C:\Users\mahdi\are256b-w24\week2.smcl
  log type:  smcl
 opened on:  19 Jan 2024, 09:54:28

1 . *---------------------------------------------------
2 .
3 .
4 . *open a .dta (Stata) file
5 . *we use clear to reaplce the new dataset with the former one
6 . use "data\EAWE01.dta", clear

7 .
8 . *---------------------------------------------------
9 . *linear model
10. *---------------------------------------------------
11. *let us work with some linear probability models
12. *P(Y_i=1|X_i) = \beta X_i + \epsilon_i
13. *Prob of finishing a bachelor's degree vs composite cognitive ability test
14.
15. reg EDUCBA  ASVABC, robust

Linear regression                               Number of obs     =        500
                                                F(1, 498)         =      87.02
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1185
                                                Root MSE          =     .42946

------------------------------------------------------------------------------
             |               Robust
      EDUCBA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1746469   .0187219     9.33   0.000     .1378633    .2114305
       _cons |   .2566462    .017902    14.34   0.000     .2214734     .291819
------------------------------------------------------------------------------

16.
17. * calcualting the \hat{Y}_i = \hat{\beta}X_i for some values of X_i
18.
19. sum ASVABC, detail

                            ASVABC
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -2.218827      -3.053471
 5%    -1.464511      -2.505618
10%    -.9897522      -2.462728       Obs                 500
25%    -.2895314      -2.348369       Sum of Wgt.         500

50%      .334199                      Mean           .2253335
                        Largest       Std. Dev.      .9005283
75%     .8584125       2.000001
90%     1.320404       2.003199       Variance       .8109511
95%     1.619688       2.049761       Skewness      -.4774249
99%     1.971871       2.319522       Kurtosis       3.170781

20.
21. display 0.2566+0.1746*0.3341
  .31493386
```

```
22. display 0.2566+0.1746*1.9718
  .60087628

23. display 0.2566+0.1746*(-2.2188)
  -.13080248

24.
25. * alternative way to calculate the predicted probability
26. display _b[_cons]+_b[ASVABC]*0.3341
  .31499574

27. display _b[_cons]+_b[ASVABC]*1.9718
  .60101499

28. display _b[_cons]+_b[ASVABC]*(-2.2188)
  -.13086036

29.
30. *Does the last predicted probability make sense?
31. * No, it yields a negative probability
32.
33. *let's find the fitted values for all the observations
34. *\hat{Y}_i = \hat{\beta}X_i
35. *command predict yields the fitted values for all the observations
36. * based on the "latest" model ran
37. help predict // like ? predict in R

38. predict EDUCBA_hat, xb

39.
40. browse EDUCBA EDUCBA_hat

41.
42. count if EDUCBA_hat>1
    0

43. count if EDUCBA_hat<0
    24

44. count if missing(EDUCBA_hat)
    0

45.
46. *Show the predicted probability graphically
47. twoway scatter EDUCBA_hat ASVABC

48. graph export outputs/linear.png, replace
  (file outputs/linear.png written in PNG format)

49.
50. *-----------------------------------------------------
51. *nonlinear model
52. *-----------------------------------------------------
53.
54. *let us move to non-linear probability models
55. *Non-linear probability models map the dependent variables using a function
56. *whose range lies between zero and one.
57.
58. *PROBIT: The function used for mapping is the cumulative distribution
59. *of a normal.
```

```
60. *P(Y_i=1|X_i) = \Phi(\beta X_i + \epsilon_i)
61.
62. probit EDUCBA  ASVABC, robust

  Iteration 0:   log pseudolikelihood = -303.71846
  Iteration 1:   log pseudolikelihood = -270.33421
  Iteration 2:   log pseudolikelihood = -269.96199
  Iteration 3:   log pseudolikelihood = -269.96172
  Iteration 4:   log pseudolikelihood = -269.96172

  Probit regression                                 Number of obs     =        500
                                                    Wald chi2(1)      =      59.05
                                                    Prob > chi2       =     0.0000
  Log pseudolikelihood = -269.96172                 Pseudo R2         =     0.1111

------------------------------------------------------------------------------
             |               Robust
      EDUCBA |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .6190642   .0805616     7.68   0.000     .4611664    .7769619
       _cons |  -.7621472   .0707757   -10.77   0.000    -.9008652   -.6234293
------------------------------------------------------------------------------

63. *Compare results with Linear Probability Model
64.
65. *--------------------------------------------------
66.
67. *Computing marginal effects
68. *Look at Slides 68-69 for definitions
69. *Average Margnal Effect
70. margins, dydx(ASVABC)

  Average marginal effects                          Number of obs     =        500
  Model VCE    : Robust

  Expression   : Pr(EDUCBA), predict()
  dy/dx w.r.t. : ASVABC

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1888049   .0206424     9.15   0.000     .1483465    .2292632
------------------------------------------------------------------------------

71.
72. *Marginal effects evaluated at the mean
73. margins, dydx(ASVABC) atmeans

  Conditional marginal effects                      Number of obs     =        500
  Model VCE    : Robust

  Expression   : Pr(EDUCBA), predict()
  dy/dx w.r.t. : ASVABC
  at           : ASVABC          =     .2253335 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .2034506   .0257674     7.90   0.000     .1529474    .2539537
------------------------------------------------------------------------------
```

```
74. // alternative: mfx compute, dydx
75.
76. *Marginal effects evaluated at a different point
77. margins, dydx(ASVABC) at(ASVABC=0.1)

  Conditional marginal effects                    Number of obs     =        500
  Model VCE    : Robust

  Expression   : Pr(EDUCBA), predict()
  dy/dx w.r.t. : ASVABC
  at           : ASVABC          =          .1
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ASVABC	.1932726	.0232853	8.30	0.000	.1476342	.2389111

```
78. margins, dydx(ASVABC) at(ASVABC=0.6)

  Conditional marginal effects                    Number of obs     =        500
  Model VCE    : Robust

  Expression   : Pr(EDUCBA), predict()
  dy/dx w.r.t. : ASVABC
  at           : ASVABC          =          .6
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ASVABC	.2288218	.0315965	7.24	0.000	.1668938	.2907499

```
79.
80.
81. // what does margins alone do?
82.
83. *Predict Probability
84. *\hat{Y}_i = \Phi{\hat{\beta}X_i}
85.
86. *calculating the predicted probability
87. h nlcom

88. h norm

89. *At 75 percentile
90. nlcom norm(_b[ASVABC]*0.8584 + _b[ _cons])

       _nl_1:  norm(_b[ASVABC]*0.8584 + _b[ _cons])
```

EDUCBA	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.4087574	.0278879	14.66	0.000	.3540982	.4634166

```
91. *At 1 percentile
92. nlcom norm(_b[ASVABC]*-2.2188 + _b[ _cons])

       _nl_1:  norm(_b[ASVABC]*-2.2188 + _b[ _cons])
```

EDUCBA	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.0163508	.0090219	1.81	0.070	-.0013318	.0340335

```
93.
94. *Generate variable that predicts for every observation
95. predict EDUCBA_probit_hat
  (option pr assumed; Pr(EDUCBA))

96. browse EDUCBA EDUCBA_hat EDUCBA_probit_hat

97. twoway (scatter EDUCBA_probit_hat ASVABC)

98.
99. *--------------------------------------------------
100 *model comparison based on rmse
101 *--------------------------------------------------
102 *How do the models compare? (linear vs probit)
103 twoway (scatter EDUCBA_probit_hat ASVABC) ///
  >        (scatter EDUCBA_hat ASVABC) ///
  >        (scatter EDUCBA ASVABC)

104
105 *We use root mean squared error (rmse) concept to compare
106 *rmse = sqrt{((1/n)*(\Sigma{(Y_i - \hat{Y_i})^2})}
107 * look at slide 65
108
109 gen sqerror        = (EDUCBA - EDUCBA_hat)^2

110 gen sqerror_probit = (EDUCBA - EDUCBA_probit_hat)^2

111
112
113 qui summarize sqerror

114 di r(mean)^0.5
  .42860029

115
116 qui summarize sqerror_probit

117 di r(mean)^0.5
  .42795955

118
119 *--------------------------------------------------
120 *Cenosred data and the Tobit model
121 *--------------------------------------------------
122
123 *Censored Data Generation (Monte Carlo Method):
124 clear all

125 set obs 50
  number of observations (_N) was 0, now 50

126 gen X=_n+10

127 gen U=rnormal(0,10)

128 gen Ystar=-40+1.2*X+U

129 gen Y= Ystar*(Ystar>0)

130
```

```
131 scatter Y X if Y>0 || lfit Y X if Y>0|| lfit Ystar X, ///
  > legend(label(1 "Y")  ///
  > label(2 "Truncated Regression")  ///
  > label(3 "True Regression Relationship") )

132
133 regress Y X if Y>0
```

Source	SS	df	MS		
Model	**1835.40834**	**1**	**1835.40834**	Number of obs	= **28**
Residual	**1354.44646**	**26**	**52.0940947**	F(1, 26)	= **35.23**
				Prob > F	= **0.0000**
				R-squared	= **0.5754**
				Adj R-squared	= **0.5591**
Total	**3189.8548**	**27**	**118.14277**	Root MSE	= **7.2176**

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
X	**.8690103**	**.146404**	**5.94**	**0.000**	**.5680726**	**1.169948**
_cons	**-24.52125**	**6.820086**	**-3.60**	**0.001**	**-38.54013**	**-10.50236**

```
134
135 *the truncated regression slope is biased (slide 77)
136 *one soltion is Tobit model
137 *ll() argument isleft-censoring limit i.e.
138 * We only observe Y_i > 0 in this regression.
139 tobit Y X, ll(0)  robust
```

Tobit regression

Number of obs	=	**50**
F(**1**, **49**)	=	**87.74**
Prob > F	=	**0.0000**
Pseudo R2	=	**0.2536**

Log pseudolikelihood = **-101.72532**

Y	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
X	**1.099829**	**.1174126**	**9.37**	**0.000**	**.8638792**	**1.335778**
_cons	**-36.71956**	**5.270491**	**-6.97**	**0.000**	**-47.31101**	**-26.12811**
/sigma	**7.493086**	**1.059238**			**5.364468**	**9.621704**

22 left-censored observations at Y <= **0**
28 uncensored observations
0 right-censored observations

```
140
141
142 *----------------------------------------------------
143 * presentation: exporting tables
144 *----------------------------------------------------
145 /*
  > *use estout to generate nice tables
  > ssc install estout, replace
  >
  > *To create nice LATEX/Doc tables we can use this command
  > *If you do not want/need Latex output, just erase the commands.
  > eststo clear
  > eststo model_l: quietly regress EDUCBA  ASVABC, robust
  > eststo model_p: quietly probit EDUCBA  ASVABC, robust
  >
  > esttab model_l
  >
  >
  > esttab model_l using outputs/model_l.rtf, replace ///
  > se onecell width(\hsize) ///
  > addnote() ///
  > label title(Estimation Result of Linear Model)
  > */
```

```
146 *---------------------------------------------------
147
148 log close // Close the log, end the file
       name:  <unnamed>
        log:  C:\Users\mahdi\are256b-w24\week2.smcl
   log type:  smcl
  closed on:  19 Jan 2024, 09:54:38
```