Class 23 – MRM: Qualitative Regressors (Part II)

Pedram Jahangiry

JON M. HUNTSMAN
SCHOOL OF BUSINESS
UtahStateUniversity

A Binary dependent variable: the linear probability model

- Linear regression when the dependent variable is binary

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

Using Zero conditional mean assumption

$$\Rightarrow \quad E(y|\mathrm{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

If the dependent variable only takes on the values 1 and 0

$$E(y|\mathrm{x}) = 1 \cdot P(y=1|\mathrm{x}) + 0 \cdot P(y=0|\mathrm{x})$$

Linear probability model (LPM)

$$\Rightarrow \quad \boxed{P(y=1|\mathrm{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}$$

The multiple linear regression model with a binary dependent variable is called the linear probability model (LPM) because the response probability is linear in the parameters

$$\Rightarrow \quad \beta_j = \Delta P(y=1|\mathrm{x})/\Delta x_j$$

In the linear probability model, the coefficients describe the effect of the explanatory variables on the probability that y=1

	inlf	hours	kidslt6	kidsge6	age	educ	wage	repwage	hushrs	husage	huseduc	huswage	faminc	mtr	motheduc	fatheduc	unem	city	exper	nwifeinc	lwage	expersq
1	1	1610	1	0	32	12	3.3540	2.65	2708	34	12	4.0288	16310	0.7215	12	7	5.0	0	14	10.910060	1.21015370	196
2	1	1656	0	2	30	12	1.3889	2.65	2310	30	9	8.4416	21800	0.6615	7	7	11.0	1	5	19.499981	0.32851210	25
3	1	1980	1	3	35	12	4.5455	4.04	3072	40	12	3.5807	21040	0.6915	12	7	5.0	0	15	12.039910	1.51413774	225
4	1	456	0	3	34	12	1.0965	3.25	1920	53	10	3.5417	7300	0.7815	7	7	5.0	0	6	6.799996	0.09212332	36
5	1	1568	1	2	31	14	4.5918	3.60	2000	32	12	10.0000	27300	0.6215	12	14	9.5	1	7	20.100058	1.52427220	49
6	1	2032	0	0	54	12	4.7421	4.70	1040	57	11	6.7106	19495	0.6915	14	7	7.5	1	33	9.859054	1.55648005	1089
7	1	1440	0	2	37	16	8.3333	5.95	2670	37	12	3.4277	21152	0.6915	14	7	5.0	0	11	9.152048	2.12025952	121
8	1	1020	0	0	54	12	7.8431	9.98	4120	53	8	2.5485	18900	0.6915	3	3	5.0	0	35	10.900038	2.05963421	1225
9	1	1458	0	2	48	12	2.1262	0.00	1995	52	4	4.2206	20405	0.7515	7	7	3.0	0	24	17.305000	0.75433636	576
10	1	1600	0	2	39	12	4.6875	4.15	2100	43	12	5.7143	20425	0.6915	7	7	5.0	0	21	12.925000	1.54489934	441
⋮																						
741	0	0	1	1	31	12	NA	0.00	800	33	14	3.0000	4000	0.8015	12	7	9.5	1	10	4.000000	NA	100
742	0	0	0	1	44	12	NA	0.00	3022	46	12	10.5890	40500	0.5815	7	7	7.5	1	5	40.500000	NA	25
743	0	0	0	1	48	11	NA	0.00	1512	50	14	10.9130	21620	0.7215	10	7	7.5	1	7	21.620001	NA	49
744	0	0	0	1	53	12	NA	0.00	2677	53	12	5.6033	23426	0.7215	0	0	7.5	1	11	23.426001	NA	121
745	0	0	0	3	42	10	NA	2.75	3150	44	12	7.9365	26000	0.6615	3	3	11.0	1	14	26.000000	NA	196
746	0	0	2	6	39	12	NA	0.00	1430	34	12	2.9476	7840	0.9415	7	0	9.5	1	5	7.840000	NA	25
747	0	0	1	2	32	10	NA	0.00	3307	36	4	2.0562	6800	0.7915	7	3	7.5	0	2	6.800000	NA	4
748	0	0	0	2	36	12	NA	0.00	3120	39	12	1.3013	5330	0.7915	7	12	14.0	0	4	5.330000	NA	16
749	0	0	0	2	40	13	NA	0.00	3020	43	16	9.2715	28200	0.6215	10	10	9.5	1	5	28.200001	NA	25
750	0	0	2	3	31	12	NA	0.00	2056	33	12	4.8638	10000	0.7715	12	12	7.5	0	14	10.000000	NA	196
751	0	0	0	0	43	12	NA	0.00	2383	43	12	1.0898	9952	0.7515	10	3	7.5	0	4	9.952000	NA	16
752	0	0	0	0	60	12	NA	0.00	1705	55	8	12.4400	24984	0.6215	12	12	14.0	1	15	24.983999	NA	225
753	0	0	0	3	39	9	NA	0.00	3120	48	12	6.0897	28363	0.6915	7	7	11.0	1	12	28.363001	NA	144

Example: Labor force participation of married women

=1 if in labor force, =0 otherwise

Non-wife income (in thousand dollars per year)

$$\widehat{inlf} = \underset{(.154)}{.586} - \underset{(.0014)}{.0034}\ nwifeinc + \underset{(.007)}{.038}\ educ + \underset{(.006)}{.039}\ exper$$

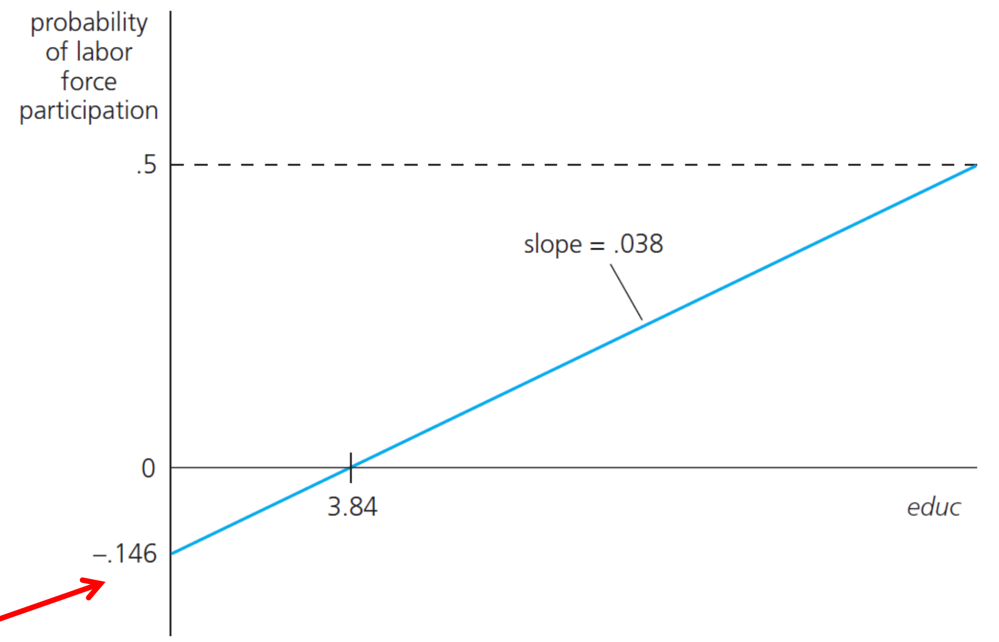
$$- \underset{(.00018)}{.00060}\ exper^2 - \underset{(.002)}{.016}\ age - \underset{(.034)}{.262}\ kidslt6$$

$$+ \underset{(.0132)}{.0130}\ kidsge6,\ n = 753, R^2 = .264$$

If the number of kids under six years increases by one, the probability that the woman works falls by 26.2%

Example: Female labor participation of married women (cont.)

FIGURE 7.3 Estimated relationship between the probability of being in the labor force and years of education, with other explanatory variables fixed.



Graph for **nwifeinc=50**, **exper=5**, **age=30**, **kindslt6=1**, and **kidsge6=0**

The maximum level of education in the sample is educ=17. For the given case, this leads to a predicted probability to be in the labor force of about **50%**.

Negative predicted probability but no problem because no woman in the sample has educ < 5.

Goodness-of-fit measure for binary dependent variables: Percent Correctly Predicted

Percent Correctly Predicted (y=1) = $\frac{count\ \widehat{inlf}=1}{count\ inlf=1} = \frac{8}{10} = 0.8$

Percent Correctly Predicted (y=0) = $\frac{count\ \widehat{inlf}=0}{count\ inlf=0} = \frac{6}{10} = 0.6$

Overall correct prediction =

$$\frac{count\ (\widehat{inlf}=1 \mid inlf=1)+count(\ \widehat{inlf}=0\ \mid inlf=0)}{total\ count\ inlf} = \frac{14}{20} = 0.7$$

obs	predicted_inlf	inlf_hat	inlf	kidslt6	kidsge6	age	educ	exper	nwifeinc
1	0.99	1	1	0	4	39	12	21	12.9
2	0.50	1	1	1	3	36	11	10	10.7
3	0.64	1	1	0	2	49	12	13	14.4
4	0.29	0	1	1	1	45	12	9	23.7
5	0.58	1	1	2	0	32	17	14	15.1
6	0.47	0	1	0	5	36	10	2	18.2
7	0.90	1	1	0	1	40	12	21	22.6
8	0.92	1	1	0	2	43	13	22	21.6
9	0.88	1	1	0	1	33	12	14	24.0
10	0.76	1	1	0	1	30	12	7	16.0
11	0.27	0	0	0	1	49	12	2	21.0
12	0.29	0	0	2	0	30	16	5	23.6
13	0.61	1	0	1	0	30	12	12	22.8
14	0.35	0	0	0	4	41	12	1	35.9
15	0.64	1	0	0	1	45	12	12	21.7
16	0.49	0	0	0	5	43	12	4	21.8
17	0.62	1	0	0	1	42	13	9	31.0
18	0.33	0	0	0	0	60	12	9	15.3
19	0.30	0	0	0	0	57	12	6	12.9
20	0.51	1	0	0	2	38	10	5	15.8

Advantages vs. Disadvantages of the linear probability model

Disadvantages

- Predicted probabilities may be larger than one or smaller than zero
  - Marginal probability effects sometimes logically impossible (having 4 kids under age 6 in previous example)
  - The linear probability model is necessarily heteroskedastic (Violation of MLR5)

$$Var(y|\mathbf{x}) = P(y=1|\mathbf{x})\,[1 - P(y=1|\mathbf{x})]$$ ← Variance of Bernoulli variable

  - Heteroskedasticity consistent standard errors need to be computed

Advantanges

- Easy estimation and interpretation
  - Estimated effects and predictions are often reasonably good in practice

More on policy analysis and program evaluation

Are nonwhite customers discriminated against (Discrimination in loan approval)?

Dummy indicating whether loan was approved

Race dummy

Credit rating

$$approved = \beta_0 + \beta_1 nonwhite + \beta_2 income + \beta_3 wealth + \beta_4 credrate + u$$

- ✓ It is important to control for other characteristics that may be important for loan approval (e.g. profession, unemployment)
- ✓ Omitting important characteristics that are correlated with the non-white dummy will produce spurious evidence for discrimination