Applied Statistics and Econometrics

Binary dependent variable

Giuseppe Ragusa

15 apr 2016

The problem

So far the dependent variable (Y) has been continuous:

- district-wide average test score
- traffic fatality rate

What if Y is binary?

- Y = get into college, or not; X = high school grade
- Y = person smokes, or not; X = income
- Y = mortgage application is accepted, or not; X = income, house characteristics, marital status, race

The Boston Fed HMDA data

Individual applications for single-family mortgages made in 1990 in the greater Boston area

- 2379 observations, collected under Home Mortgage Disclosure Act (HMDA)

Variable

Dependent variable:

- Is the mortgage denied or accepted? (`deny`)

Independent variables:

- demographic characteristics of applicants and other loan and property characteristics

Linear Probability Model

A natural starting point is the linear regression model with a single regressor:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

But:

- What does β_1 mean when Y is binary?
- What does $\beta_0 + \beta_1$ X_i mean when Y is binary?
- What does the predicted value $\hat{Y}$ mean when Y is binary?

The linear probability model, ctd.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Recall assumption #1 : E(u_i|X_i)=0, so

$$E(Y_i|X_i) = E(\beta_0 + \beta_1 X_i + u_i|X_i) = \beta_0 + \beta_1 X_i$$

When Y is binary,

$$E(Y_i|X_i) = 1 \times \Pr(Y = 1|X_i) + 0 \times \Pr(Y = 0|X_i) = \Pr(Y_i = 1|X_i)$$

Important

$$E(Y_i|X_i) = \Pr(Y_i = 1|X_i)$$

The linear probability model, ctd.

When Y is binary, the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

is called the **linear probability model**.

- The predicted value is a **probability**
  - $E(Y|X = x) = Pr(Y = 1|X = x) =$ prob. that $Y = 1$ given $X = x$
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i =$ the predicted probability that $Y_i = 1$, given $X = x$
- $\hat{\beta}_1 =$ change in probability that $Y = 1$ for a given X

$$\beta_1 = \Pr(Y_i = 1|X_i = x + 1) - \Pr(Y_i|X_i = x)$$

HMDA Act

The Home Mortgage Disclosure Act (HMDA) was enacted by Congress in 1975.

The Home Mortgage Disclosure Act was enacted to monitor minority and low-income access to the mortgage market. The data collected for this purpose show that minorities are more than twice as likely to be denied a mortgage as whites.

HMDA Data (I)

```
- deny: Factor. Was the mortgage denied?
- pirat: Payments to income ratio.
- hirat: Housing expense to income ratio.
- lvrat: Loan to value ratio.
- chist: Factor. Credit history: consumer payments.
- mhist: Factor. Credit history: mortgage payments.
- phist: Factor. Public bad credit record?
- unemp: 1989 Massachusetts unemployment rate in applicant's
  industry.
- selfemp: Factor. Is the individual self-employed?
- insurance: Factor. Was the individual denied mortgage
  insurance?
- condomin: Factor. Is the unit a condominium?
- afam: Factor. Is the individual African-American?
- single: Factor. Is the individual single?
- hschool: Factor. Does the individual have a high-school
  diploma?
```

HMDA Data (II)

```r
library(ase)
data(hmda)
```

```
##   deny          pirat             hirat             lvrat            chist
##  no :2095   Min.   :0.0000   Min.   :0.0000   Min.   :0.0200   1:1352
##  yes: 284   1st Qu.:0.2800   1st Qu.:0.2140   1st Qu.:0.6530   2: 441
##             Median :0.3300   Median :0.2600   Median :0.7797   3: 126
##             Mean   :0.3297   Mean   :0.2542   Mean   :0.7378   4:  77
##             3rd Qu.:0.3700   3rd Qu.:0.2984   3rd Qu.:0.8685   5: 182
##             Max.   :1.4200   Max.   :1.1000   Max.   :1.9500   6: 201
##  mhist    phist      selfemp     condomin     afam
##  1: 747   no :2204   no :2103   no :1693   no :2040
##  2:1571   yes: 175   yes: 276   yes: 686   yes: 339
##  3:  40
##  4:  21
##
##
```

Example: Linear probability model, HMDA data

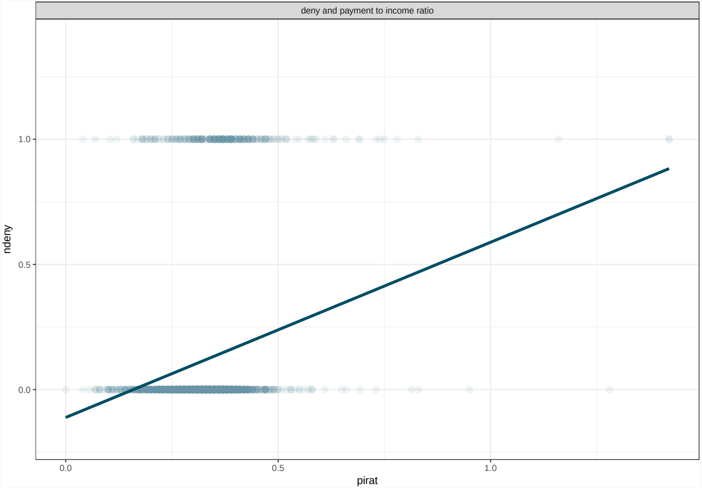


Figure 1: Mortgage denial v. ratio of debt payments to income. Source: HMDA

Example: Linear probability model: HMDA data, ctd.

```r
lm1 <- lm(ndeny ~ pirat, data = hmda)
summary_rob(lm1)
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.11138    0.02949  -3.777 0.000159
## pirat        0.69993    0.09191   7.615 2.63e-14
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.3179 on 2377 degrees of freedom
## Multiple R-squared:  0.03965, Adjusted R-squared:  0.03925
## F-statistic: 57.99 on 1 and Inf DF,  p-value: 2.628e-14
```

Prediction

To obtain (in sample) predicted probabilities

```
lm1 <- lm(ndeny ~ hirat, data = hmda)
denyhat <- predict(lm1)
summary(denyhat)  ## Summarize the probabilities
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.01075  0.09880  0.12240  0.11940  0.14200  0.55240
```

```
head(cbind(actual = hmda$ndeny, predicted = ifelse(denyhat >
    0.5, 1, 0)), 10)
```

```
##   actual predicted
## 1      0         0
## 2      0         0
## 3      0         0
## 4      0         0
## 5      0         0
## 6      0         0
## 7      0         0
```

The linear probability model

Models $\Pr(Y = 1|X)$ as a linear function of X

- Advantages:
  - simple to estimate and to interpret
  - inference is the same as for multiple regression (need heteroskedasticity-robust standard errors)
- Disadvantages:
  - Does it make sense that the probability should be linear in X
  - Predicted probabilities can be < 0 or > 1

Solution

These disadvantages can be solved by using a nonlinear probability model:

- **probit** regression
- **logit** regression

Probit Regression

The problem with the linear probability model is that it models the probability of Y=1 as being linear:

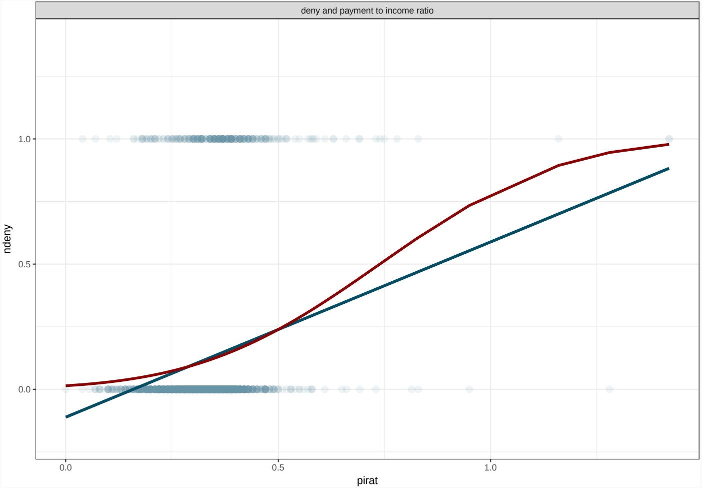
$$\Pr(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i$$

Instead, we want:

- $0 \leq Pr(Y = 1|X) \leq 1$ for all X
- $Pr(Y_i = 1|X_i)$ to be increasing in X (for $\beta_1 > 0$)

This requires a nonlinear functional form for the probability. How about an "S-curve"...

Graphical Intuition



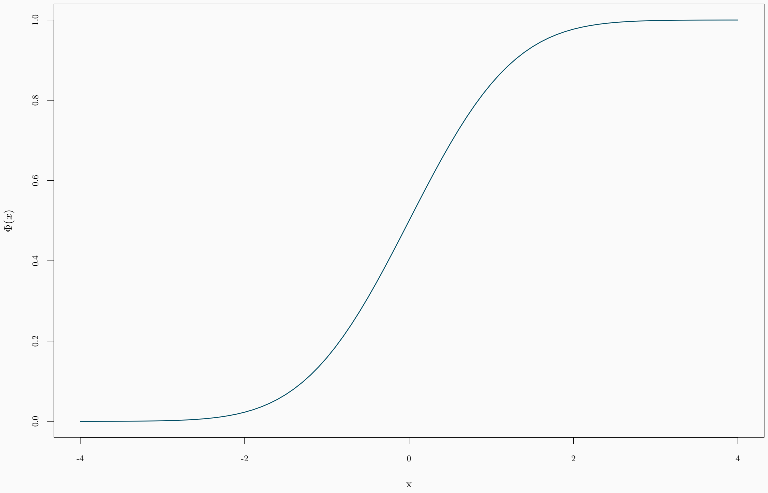
Probit Regression

The probit Regression models the probability that $Y = 1$ using the cumulative standard normal distribution function, evaluated at $\beta_0 + \beta_1 X$:

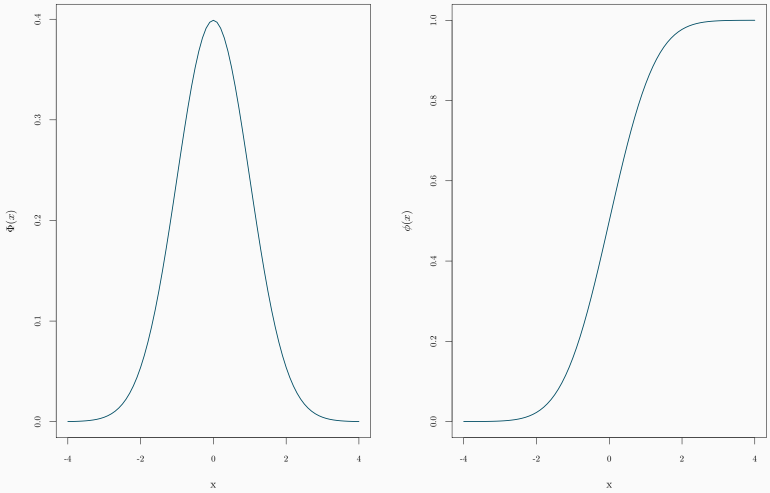
$$\Pr(Y_i = 1 | X_i) = \Phi(\beta_0 + \beta_1 X_i)$$

- $\Phi()$ is the cumulative normal distribution function
- $z = \beta_0 + \beta_1 X$ is the "z-score" of the probit model

The normal cumulative distribution



The normal cumulative distribution



Probit regression

Suppose $\beta_0 = -2$, $\beta_1 = 3$, and $X = .4$, so

$$\Pr(Y = 1|X = .4) = \Phi(-2 + 3 \times .4) = \Phi(-.8)$$

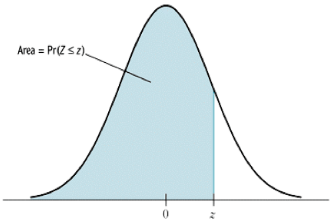
Pr(Y=1|X=.4) is the area under the standard normal density to left of $z = -.8$, which is given. . .

```r
# R command to calculate the c.d.f. of the standard normal
# distribution evaluated at -.8
pnorm(-0.8)
```

. . . by 0.2118554

Probit Regression

TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z)$ = Pr(Z ″ z)



z	Second Decimal Value of z									
	0	1	2	3	4	5	6	7	8	9
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121

Probit regression, ctd.

Why use the cumulative normal probability distribution?

- The "S-shape" gives us what we want:
  - $0 \leq Pr(Y = 1|X) \leq 1$ for all X
  - $Pr(Y = 1|X)$ is increasing in X (for $\beta_1 > 0$)
- Easy to use — the probabilities are tabulated in the cumulative normal tables
- Relatively straightforward interpretation:
  - z-score $= \beta_0 + \beta_1 X_i$
  - $\hat{\beta}_0 + \hat{\beta}_1 X_i$ is the predicted z-score, given X
  - β_1 is the change in the zscore for a unit change in X

Probit Regression: Example: HMDA data

```r
glm(deny ~ pirat, data = hmda, family = binomial(probit))
```

```
##
## Call:  glm(formula = deny ~ pirat, family = binomial(probit), data = hmda)
##
## Coefficients:
## (Intercept)        pirat
##      -2.194        2.968
##
## Degrees of Freedom: 2378 Total (i.e. Null);  2377 Residual
## Null Deviance:       1740
## Residual Deviance: 1664  AIC: 1668
```

$$\widehat{\Pr}(deny_i = 1|pirat) = \Phi(\underset{(.16)}{-2.19} + \underset{(.46)}{2.97} \times pirat)$$

Probit regression: HMDA data, ctd.

$$\widehat{\Pr}(deny_i = 1 | pirat) = \Phi(\underset{(.16)}{-2.19} + \underset{(.46)}{2.97} \times pirat)$$

- Positive coefficient: does this make sense?
- Standard errors have the usual interpretation
- Predicted probabilities:

$$\begin{aligned}
\widehat{\Pr}(deny_i = 1 | pirat = .3) &= \Phi(-2.19 + 2.97 \times .3) \\
&= \Phi(-1.30) = .097 \\
\widehat{\Pr}(deny_i = 1 | pirat = .4) &= \Phi(-2.19 + 2.97 \times .4) \\
&\approx \Phi(-1.0) = .159
\end{aligned}$$

Probit regression

Effect of a change in payment income ratio from .3 to .4 is

$$\Pr(deny_i \widehat{= 1|p}irat = .4) - \Pr(deny_i \widehat{= 1|p}irat = .3)$$
$$= .159 - .097$$
$$= 0.062$$

Predicted probability of denial rises from .097 to .159 (an increase of 6.2 percentage point).

Probit regression with multiple regressors

$$Pr(Y = 1|X_1, \ldots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

- Φ is the cumulative normal distribution function
- $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ is the "z-score" of the probit model
- β_1 is the effect on the **z-score** of a unit change in X_1, holding costant $X_2, \ldots, X_k$

Note: the z-score does not have anything to do with the z-value reported in the third columns of the output in R.

Probit Regression: R Example - HMDA data

```r
summary_rob(glm(deny ~ pirat, data = hmda, family = binomial(probit)))
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1942     0.1891 -11.605  < 2e-16
## pirat         2.9681     0.5372   5.525 3.29e-08
## ---
## Heteroskadasticity robust standard errors used
##
## Multiple R-squared:   , Adjusted R-squared:
## F-statistic: 30.53 on 1 and Inf DF,  p-value: 3.292e-08
```

Probit Regression

Interpretation of the coefficient

We want to estimate (when there is only one X):

$$\frac{\partial \Pr(Y_i = 1|X_i)}{\partial X_i}$$

that is, the effect on the probability of increasing X_i.

For the probit regression, this effect is equal to:

$$\frac{\partial \Phi(\beta_0 + \beta_1 X_i)}{\partial X_i} = \phi(\beta_0 + \beta_1 X_i)\beta_1$$

Since $\phi(u) > 0$ (the probability density function of a normal), β_1 only identifies the **sign** of the effect, but not its magnitude.

Marginal effects

$$\frac{\partial \Phi(\beta_0 + \beta_1 X_i)}{\partial X_i} = \phi(\beta_0 + \beta_1 X_i)\beta_1$$

The probit regression model is a non linear model — the effect of X_i on $\Pr(Y_i = 1|X_i)$ depends on the value of X_i

Estimating the effect

There are *two* approaches to estimate the marginal effect

1. Set $X_i = \bar{X}$ and calculate
$$\phi(\hat{\beta}_0 + \hat{\beta}_1 \bar{X})\hat{\beta}_1$$
2. Calculate the average marginal effect
$$\frac{1}{n}\sum_{i=1}^{n} \phi(\hat{\beta}_0 + \hat{\beta}_1 X_i)\hat{\beta}_1$$

Marginal effects

`R` has a package `mfx` that automatically calculate the marginal effects

- The command `probitmfx(obj, data, atmean=TRUE)` calcuate the marginal effect at the mean value of X_i
- The command `probitmfx(obj, data, atmean=FALSE)` calcuate the averga marginal effects

Maarginal effect at the mean

```r
library(mfx)
probitmfx(glm1, data = hmda, atmean = TRUE)
```

```
## Call:
## probitmfx(formula = glm1, data = hmda, atmean = TRUE)
##
## Marginal Effects:
##          dF/dx Std. Err.      z     P>|z|
## pirat 0.565551  0.073019 7.7453 9.539e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Average Marginal effect

```r
probitmfx(glm1, data = hmda, atmean = FALSE)
```

```
## Call:
## probitmfx(formula = glm1, data = hmda, atmean = FALSE)
##
## Marginal Effects:
##          dF/dx Std. Err.      z     P>|z|
## pirat 0.566748  0.073145 7.7483 9.312e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What about discrimination

```
glm1 <- glm(deny ~ pirat + afam, data = hmda, family = binomial(probit))
summary_rob(glm1)
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.25877    0.17660 -12.790  < 2e-16
## pirat        2.74174    0.49765   5.509  3.6e-08
## afam         0.70816    0.08309   8.523  < 2e-16
## ---
## Heteroskadasticity robust standard errors used
##
## Multiple R-squared:   , Adjusted R-squared:
## F-statistic: 111.3 on 2 and Inf DF,  p-value: < 2.2e-16
```

What about discrimination

```r
probitmfx(glm1, atmean = FALSE, data = hmda)
```

```
## Call:
## probitmfx(formula = glm1, data = hmda, atmean = FALSE)
##
## Marginal Effects:
##          dF/dx Std. Err.      z     P>|z|
## pirat 0.501624  0.069043 7.2654 3.720e-13 ***
## afam  0.169664  0.024254 6.9954 2.644e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "afam"
```

Logit Regression

Logit regression models the probability of Y=1 as the cumulative standard logistic distribution function, evaluated at $\beta_0 + \beta_1 X$:

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

- F is the cumulative logistic distribution function:

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Logit Regression ctd.

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

where

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Example

Suppose $\beta_0 = -3$, $\beta_1 = 2$, $X = .4$. So,

$$\beta_0 + \beta_1 X = -3 + 2 \times .4 = -2.2,$$

and

$$\Pr(Y = 1|X) = 1/(1 + e^{-(-2.2)}) = 0.998$$

.

Logit Regression

Why bother with logit if we have probit?

- Historically, logit is more convenient computationally
- In practice, logit and probit are very similar

Logit is very convinient

$$P(Y = 1|X_1, \dots, X_k) = \frac{1}{1 + \exp \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}$$

implies

$$P(Y = 0|X_1, \dots, X_k) = \frac{\exp \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}{1 + \exp \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}$$

implies

$$\frac{P(Y = 0|X_1, \dots, X_k)}{P(Y = 1|X_1, \dots, X_k)} = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$

Logit Regression: partial effect

```r
logitmfx(glm(deny ~ pirat + afam, data = hmda, family = binomial(logit)),
    data = hmda, atmean = FALSE)
```

```
## Call:
## logitmfx(formula = glm(deny ~ pirat + afam, data = hmda, family = binomial(logit)),
##     data = hmda, atmean = FALSE)
##
## Marginal Effects:
##          dF/dx Std. Err.      z     P>|z|
## pirat 0.518495  0.078853 6.5755 4.850e-11 ***
## afam  0.166472  0.023876 6.9725 3.114e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "afam"
```

Probit

```r
probitmfx(glm(deny ~ pirat + afam, data = hmda, family = binomial(probit)),
    data = hmda, atmean = FALSE)
```

```
## Call:
## probitmfx(formula = glm(deny ~ pirat + afam, data = hmda, family = binomial(probit)),
##     data = hmda, atmean = FALSE)
##
## Marginal Effects:
##          dF/dx Std. Err.      z     P>|z|
## pirat 0.501624  0.069043 7.2654 3.720e-13 ***
## afam  0.169664  0.024254 6.9954 2.644e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "afam"
```

LPM

```r
summary_rob(lm(ndeny ~ pirat + afam, data = hmda))
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.11575    0.02901  -3.990 6.60e-05
## pirat        0.63744    0.09067   7.030 2.06e-12
## afam         0.17520    0.02491   7.033 2.02e-12
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.312 on 2376 degrees of freedom
## Multiple R-squared:  0.07501, Adjusted R-squared:  0.07423
## F-statistic: 108.2 on 2 and Inf DF,  p-value: < 2.2e-16
```

Example: Characterizing the Background of Hezbollah Militants

Source: Alan Krueger and Jitka Maleckova, “Education, Poverty and Terrorism: Is There a Causal Connection?” Journal of Economic Perspectives, Fall 2003, 119-144

Example: Characterizing the Background of Hezbollah Militants

Table 4

Characteristics of Hezbollah Militants and Lebanese Population of Similar Age

Characteristic	*Deceased Hezbollah Militants*	*Lebanese Population Age 15–38*
< *Poverty*	28%	33%
Education		
Illiterate	0%	6%
Read and write	22%	7%
Primary	17%	23%
Preparatory	14%	26%
Secondary	33%	23%
University	13%	14%
High Studies	1%	1%
Age		
Mean	22.17	25.57
[std.dev.]	(3.99)	(6.78)
15–17	2%	15%
18–20	41%	14%
21–25	42%	23%
26–30	10%	20%
31–38	5%	28%
Hezbollah Education System	21%	NA
Region of Residence		
Beirut	42%	13%
Mount Lebanon	0%	36%
Bekaa	26%	13%
Nabatieh	2%	6%
South	30%	10%
North	0%	22%
Marital Status		
Divorced	1%	NA
Engaged	5%	NA
Married	39%	NA
Single	55%	NA

Notes: Sample size for Lebanese population sample is 120,796. Sample size for Hezbollah is 50 for poverty status, 78 for education, 81 for age (measured at death), 129 for education in Hezbollah system, 116 for region of residence and 75 for marital status.

Figure 2:

Example: Characterizing the Background of Hezbollah Militants

Table 5

Logistic Estimates of Participation in Hezbollah

(*dependent variable is 1 if individual is a deceased Hezbollah militant, and 0 otherwise; standard errors shown in parentheses*)

	All of Lebanon:				*Heavily Shiite Regions:*	
	Unweighted Estimates		*Weighted Estimates*		*Weighted Estimates*	
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Intercept	−4.886	−5.910	−5.965	−6.991	−4.658	−5.009
	(0.365)	(0.391)	(0.230)	(0.255)	(0.232)	(0.261)
Attended Secondary	0.281	0.171	0.281	0.170	0.220	0.279
School or Higher (1 = yes)	(0.191)	(0.193)	(0.159)	(0.164)	(0.159)	(0.167)
Poverty (1 = yes)	−0.335	−0.167	−0.335	−0.167	−0.467	−0.500
	(0.221)	(0.223)	(0.158)	(0.162)	(0.159)	(0.166)
Age	−0.083	−0.083	−0.083	−0.083	−0.083	−0.082
	(0.015)	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)
Beirut (1 = yes)	—	2.199	—	2.200	—	0.168
		(0.219)		(0.209)		(0.222)
South Lebanon (1 = yes)	—	2.187	—	2.187	—	1.091
		(0.232)		(0.221)		(0.221)
Pseudo R-Square	0.020	0.091	0.018	0.080	0.021	0.033
Sample Size	120,925	120,925	120,925	120,925	34,826	34,826

Notes: Sample pools together observations on 129 deceased Hezbollah fighters and the general Lebanese population from 1996 PHS. Weights used in columns 3 and 4 are the relative share of Hezbollah militants in the population to their share in the sample and relative share of PHS respondents in the sample to their share in the population. Weight is 0.273 for Hezbollah sample and .093 for PHS sample.

Figure 3:

Example: Characterizing the Background of Hezbollah Militants

$$
\begin{aligned}
&\Pr(Y = 1 | \textit{secondary} = 1, \textit{poverty} = 0, \textit{age} = 20) \\
&\quad - \Pr(Y = 1 | \textit{secondary} = 1, \textit{poverty} = 0, \textit{age} = 20) \\
&= .000646 - .000488 = .000158
\end{aligned}
$$

Both these statements are true:

- The probability of being a Hezbollah militant increases by 0.0158 percentage point, if secondary school is attended.
- The probability of being a Hezbollah militant increases by 32%, if secondary school is attended ($.000158/.000488 = .32$).

Logit and Probit: Estimation and Inference

- Logit and Probit coefficients are estimated through Maximum likelihood
- Once the coefficients are estimated, R gives you all the information to carry out inference on the parameters (confidence intervals, testing, etc.)
- What happens if the X of the probit model is expressed in logarithm? And if X is a dummy variable?