Comparing Odds Ratios and Marginal Effects from Logistic Regression and Linear Probability Models

Matt Bogard
Comparing Odds Ratios and Marginal Effects from Logistic Regression and Linear Probability Models in SAS and R

Models of binary dependent variables often are estimated using logistic regression or probit models, but the estimated coefficients (or exponentiated coefficients expressed as odds ratios) are often difficult to interpret from a practical standpoint. Empirical economic research often reports ‘marginal effects’, which are more intuitive but often more difficult to obtain from popular statistical software. The most straightforward way to obtain marginal effects is from estimation of linear probability models. This paper uses a toy data set to demonstrate the calculation of odds ratios and marginal effects from logistic regression using SAS and R, while comparing them to the results from a standard linear probability model.

Suppose we have a data set that looks at program participation (for some program or product or service of interest) by age and we want to know the influence of age on the decision to participate. Our data may look something like that below:

<table>
<thead>
<tr>
<th>Obs</th>
<th>PARTICIPANT</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>34</td>
</tr>
</tbody>
</table>

And we can see that there are differences or a decrease in average participation rates as age increases.

<table>
<thead>
<tr>
<th>Analysis Variable : PARTICIPANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGECAT</td>
</tr>
<tr>
<td>1: &lt;30</td>
</tr>
<tr>
<td>2: 30-50</td>
</tr>
<tr>
<td>3: 51+</td>
</tr>
</tbody>
</table>

Theoretically, this might call for logistic regression for modeling a dichotomous outcome like participant, so we could use SAS or R to get the following results:
Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 5.92972    | 2.34258 | 2.531    | 0.0114 * |
| age       | -0.14099   | 0.05656 | -2.493   | 0.0127 * |
| OR        | 2.5 %      | 97.5 %  |          |          |

While the estimated coefficients from logistic regression are not easily interpretable (they represent the change in the log of odds of participation for a given change in age) odds ratios might provide a better summary of the effects of age on participation (odds ratios are derived from exponentiation of the estimated coefficients from logistic regression -see also: The Calculation and Interpretation of Odds Ratios) and may be somewhat more meaningful. We can see the odds ratio associated with age is .8685 which implies that for every year increase in age the odds of participation are about (1-.865)*100 = 13.15% less. You tell me what this means if this is the way you think about the likelihood of outcomes in everyday life!

Marginal effects are an alternative metric that can be used to describe the impact of age on participation. Marginal effects can be described as the change in outcome as a function of the change in the treatment (or independent variable of interest) holding all other variables in the model constant. In linear regression, the estimated regression coefficients are marginal effects and are more easily interpreted (more on this later). Marginal effects can be output easily from STATA, however they are not directly available in SAS or R. However there are some adhoc ways of getting them which I will demonstrate here. (there are some packages in R available to assist with this as well). I am basing most of this directly on two very good blog posts on the topic:

https://statcompute.wordpress.com/2012/09/30/marginal-effects-on-binary-outcome/


One approach is to use PROC QLIM and request output of marginal effects. This computes a marginal effect for each observation’s value of x in the data set (because marginal effects may not be constant across the range of explanatory variables).
PROC QLIM DATA = DAT1;
  MODEL PARTICIPANT = AGE / DISCRETE(D = LOGIT);
  OUTPUT OUT = OUT1 MARGINAL;
RUN;

*GET AVG OF SAMPLE MARGINAL EFFECTS;
PROC MEANS DATA = OUT1 MEAN;
  VAR MEFF_P2_AGE;
RUN;

Taking the average of this result gives an estimated ‘sample average estimate of marginal effect’:

<table>
<thead>
<tr>
<th>Analysis Variable :</th>
<th>Meff_P2_AGE</th>
<th>Marginal effect of AGE on the probability of PARTICIPANT=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.025833</td>
<td></td>
</tr>
</tbody>
</table>

This tells us that for every year increase in age the probability of participation decreases on average by 2.5%. For most people, for practical purposes, this is probably a more useful interpretation of the relationship between age and participation compared to odds ratios. We can calculate this more directly (following the code from the blog post by WenSui Liu) using output from logistic regression and the data step in SAS:

PROC LOGISTIC DATA = DAT1 DESC;
  MODEL PARTICIPANT = AGE;
  OUTPUT OUT = OUT2 XBETA = XB;
RUN;

*DEFINE MARGINAL EFFECT USING THE FORMULATION: MF = EXP(XB) / ((1 + EXP(XB)) ^ 2) * beta;
DATA OUT2;
  SET OUT2;
  MARGIN_AGE = EXP(XB) / ((1 + EXP(XB)) ** 2) * (-0.1410);
RUN;

PROC MEANS DATA = OUT2 MEAN;  VAR MARGIN_AGE;RUN;

We get similar results.
We can run the same analysis in R, either replicating the results from the data step above, or using the mfx function defined by Alan Fernihough referenced in the diffuseprior blog post mentioned above or the paper referenced below:

```r
mylogit <- glm(participate ~ age, data = dat1, family = "binomial")
mfx(mylogit)
```

The paper notes that this function gives similar results to the mfx function in STATA. And we get almost the same results we got from SAS above but additionally provides bootstrapped standard errors:

<table>
<thead>
<tr>
<th>marginal.effects</th>
<th>standard.error</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0258330</td>
<td>0.6687069</td>
</tr>
</tbody>
</table>

**Marginal Effects from Linear Probability Models**

Earlier I mentioned that you could estimate marginal effects directly from the estimated coefficients from a linear probability model. While in some circles LPMs are not viewed favorably, they have a strong following among applied econometricians (see references for more on this). As Angrist and Pischke state in their very popular book *Mostly Harmless Econometrics*:

"While a nonlinear model may fit the CEF (population conditional expectation function) for LDVs (limited dependent variables) more closely than a linear model, when it comes to marginal effects, this probably matters little"

Using SAS or R we can get the following results from estimating a LPM for this data:

```
Coefficients:

            Estimate Std. Error   t value  Pr(>|t|)
(Intercept)   1.700260   0.378572     4.491 0.000111 ***
dat1$age    -0.028699   0.009362    -3.065 0.004775 **
```

---

**Analysis Variable**

<table>
<thead>
<tr>
<th>MARGIN_AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>-0.0258373</td>
</tr>
</tbody>
</table>
You can see that the estimate from the linear probability model above gives us a marginal effect almost identical to the previous estimates derived from logistic regression, as is often the case, and as indicated by Angrist and Pischke.

For a multivariable scenario, I used the binary.csv data set used by the UCLA statistical computing example *R Data Analysis Examples: Logit Regression* estimating the model admit = gre + gpa. Nearly results for marginal effects were obtained using all of the previously mentioned approaches using SAS and R:

### Odds Ratios:

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE</td>
<td>1.003</td>
<td>1.001 - 1.005</td>
</tr>
<tr>
<td>GPA</td>
<td>2.127</td>
<td>1.137 - 3.979</td>
</tr>
</tbody>
</table>

### Marginal Effects:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARGIN_GRE</td>
<td>0.00055504</td>
</tr>
<tr>
<td>MARGIN_GPA</td>
<td>0.15566</td>
</tr>
</tbody>
</table>

### Linear Probability Model:

| Variable | DF    | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|-------|--------------------|----------------|---------|------|---|
| Intercept| 1     | -0.52793           | 0.20873        | -2.53   | 0.0118 |
| GRE      | 1     | 0.000549           | 0.000214       | 2.56    | 0.0107 |
| GPA      | 1     | 0.15424            | 0.06498        | 2.37    | 0.0181 |

### Summary

Using a toy data set relating program participation to age, results from logistic regression (odds ratios) were compared to marginal effects calculated in a number of ways including linear probability models. Marginal effects derived from logistic regression and linear probability models were practically identical, and may often provide a more useful interpretation of the relationships between the dependent and independent variables than direct interpretation of logistic regression coefficients or their exponentiated form (odds ratios).
Note: Sample average marginal effects vs marginal effects at the mean. In the SAS ETS example cited in the references below, a distinction is made between calculating sample average marginal effects (which were discussed above) vs. calculating marginal effects at the mean:

“To evaluate the "average" or "overall" marginal effect, two approaches are frequently used. One approach is to compute the marginal effect at the sample means of the data. The other approach is to compute marginal effect at each observation and then to calculate the sample average of individual marginal effects to obtain the overall marginal effect. For large sample sizes, both the approaches yield similar results. However for smaller samples, averaging the individual marginal effects is preferred (Greene 1997, p. 876)”

The SAS reference implements both approaches. In the R code that follows the reference section below you will find an example of marginal effects at the mean for this data.

References:

http://support.sas.com/rnd/app/examples/ets/margeff/

Linear Regression and Analysis of Variance with a Binary Dependent Variable (from EconomicSense, by Matt Bogard).


Probit better than LPM? http://www.mostlyharmlesseconometrics.com/2012/07/probit-better-than-lpm/

Love It or Logit. By Marc Bellemare. marcfbellemare.com/wordpress/9024


SAS CODE:

`*------------------------------------------------------------------
| PROGRAM NAME: MEFF vs ORs
| DATE: 3/3/16
| CREATED BY: MATT BOGARD
| PROJECT FILE:
*------------------------------------------------------------------
| PURPOSE: GENERATE MARGINAL EFFECTS FOR LOGISTIC REGRESSION AND COMPARE TO:
| ODDS RATIOS / RESULTS FROM R
| REFERENCES: https://statcompute.wordpress.com/2012/09/30/marginal-effects-on-binary-outcome/
| UCD CENTRE FOR ECONOMIC RESEARCH
| WORKING PAPER SERIES
| 2011`
```
/*CREATE TOY DATA SET;
DATA DAT1;
   INPUT PARTICIPANT AGE;
CARDS;
1 25
1 26
1 27
1 28
1 29
1 30
0 31
1 32
1 33
0 34
1 35
1 36
1 37
1 38
0 39
0 40
0 41
1 42
1 43
0 44
0 45
1 46
1 47
0 48
0 49
0 50
0 51
0 52
1 53
0 54
; RUN;
*/

*----------------------------------------------------------------------;
| EXPLORE DATA
*----------------------------------------------------------------------;

PROC PRINT DATA=DAT1 (OBS=10); RUN;

PROC MEANS DATA=DAT1 N MIN Q1 MEDIAN MEAN Q3 MAX;
VAR AGE;
RUN;

PROC FREQ DATA=DAT1;
   TABLES PARTICIPANT;
RUN;

DATA DAT1;
   SET DAT1;
   IF AGE < 30 THEN AGECAT = "1: <30" ;
   IF (30 LE AGE LE 50) THEN AGECAT = "2: 30-50" ;
   IF AGE > 50 THEN AGECAT = "3: 51 +" ;
RUN;

PROC MEANS DATA=DAT1 N MIN MEAN MEDIAN MAX;
CLASS AGECAT;
```
```
VAR AGE;
RUN;

PROC MEANS DATA=DAT1 N MEAN;
   CLASS AGECAT;
   VAR PARTICIPANT;
RUN;

*------------------------------------------------------------------------
| LOGISTIC REGRESSION AND MARGINAL EFFECTS FROM PROC QLIM
|------------------------------------------------------------------------

*LOGISTIC REGRESSION;
PROC QLIM DATA = DAT1;
   MODEL PARTICIPANT = AGE/ DISCRETE(D = LOGIT);
   OUTPUT OUT = OUT1 MARGINAL;
RUN;

*GET AVG OF SAMPLE MARGINAL EFFECTS;
PROC MEANS DATA = OUT1 MEAN;
   VAR MEFF_P2_AGE;
RUN;

*------------------------------------------------------------------------
| FORMULA BASED 'AVERAGE OF THE SAMPLE MARGINAL EFFECTS'
| CALCULATION USING RESULTS FROM PROC LOGISTIC
|------------------------------------------------------------------------

*ADHOC CALCULATION;
PROC LOGISTIC DATA = DAT1 DESC;
   MODEL PARTICIPANT = AGE;
   OUTPUT OUT = OUT2 XBETA = XB;
RUN;

*DEFINE MARGINAL EFFECT USING THE FORMULATION: MF = EXP(XB) / ((1 + EXP(XB)) ^ 2) * beta;
DATA OUT2;
   SET OUT2;
   MARGIN_AGE = EXP(XB) / ((1 + EXP(XB)) ** 2) * (-0.1410);
RUN;

PROC MEANS DATA = OUT2 MEAN; VAR MARGIN_AGE;
RUN;

*------------------------------------------------------------------------
| COMPARISON TO LINEAR PROBABILITY MODEL
|------------------------------------------------------------------------

PROC REG DATA=DAT1;
   MODEL PARTICIPANT = AGE;
RUN;

*------------------------------------------------------------------------
| |
| |
| |
| *MULTIVARIABLE CASE
| |
| |
*------------------------------------------------------------------------

DATA DAT2;
   INPUT ADMIT GRE GPA RANK;
CARDS;
  0  380  3.61  3
  1  660  3.67  3
  1  800  4     1
  1  640  3.19  4
  0  520  2.93  4
  1  760  3     2
More lines......;
```
*LOGISTIC REGRESSION AND MARGINAL EFFECTS FROM PROC QLIM

*logistic regression;
proc qlim data = dat2;
  model admit = gre gpa / discrete(d = logit);
  output out = out1 marginal;
run;

*get avg of sample marginal effects;
proc means data = out1 mean;
  var meff_p2_gre meff_p2_gpa;
run;

*ad hoc calculation;
proc logistic data = dat2 desc;
  model admit = gre gpa;
  output out = out3 xbeta = xb;
run;

data out3;
  set out3;
  margin_gre = exp(xb) / ((1 + exp(xb)) ** 2) * (0.002691);
  margin_gpa = exp(xb) / ((1 + exp(xb)) ** 2) * (0.754687);
run;
proc means data = out3 mean;
  var margin_gre margin_gpa;
run;

*comparison to linear probability model
proc reg data = dat2;
  model admit = gre gpa;
run;

R code:

#--------------------------
# program name: meff and odds ratios
# date: 3/3/16
# created by: matt bogard
# project file:
#--------------------------
# purpose: generate marginal effects for logistic regression and compare to:
# odds ratios / results from r
# #
# references: https://statcompute.wordpress.com/2012/09/30/marginal-effects-on-binary-outcome/
#
# UCD CENTRE FOR ECONOMIC RESEARCH
# WORKING PAPER SERIES
# 2011
# Simple Logit and Probit Marginal Effects in R
# Alan Fernihough, University College Dublin
# WP11/22
# October 2011
#----------------------------------------------------------;

# generate data for continuous explanatory variable
#----------------------------------------------------------

Input = ("participate age
1 25
1 26
1 27
1 28
1 29
1 30
0 31
1 32
1 33
0 34
1 35
1 36
1 37
1 38
0 39
0 40
0 41
1 42
1 43
0 44
0 45
1 46
1 47
0 48
0 49
0 50
0 51
0 52
1 53
0 54
")
```r
dat1 <- read.table(textConnection(Input),header=TRUE)
summary(dat1) # summary stats

### run logistic regression model

mylogit <- glm(participate ~ age, data = dat1, family = "binomial")
summary(mylogit)

exp(cbind(OR = coef(mylogit), confint(mylogit))) # get odds ratios

# marginal effects calculations

# mfx function for marginal effects from a glm model
#
# from: https://diffuseprior.wordpress.com/2012/04/23/probitlogit-
marginal-effects-in-r-2/
# based on:
# UCD CENTRE FOR ECONOMIC RESEARCH
# WORKING PAPER SERIES
# 2011
# Simple Logit and Probit Marginal Effects in R
# Alan Fernihough, University College Dublin
# WP11/22
# October 2011

mfx <- function(x,sims=1000){
  set.seed(1984)
  pdf <- ifelse(as.character(x$call)[3]=="binomial(link = "probit\")","",
               mean(dnorm(predict(x, type = "link"))),
               mean(dlogis(predict(x, type = "link"))))
  pdfsd <- ifelse(as.character(x$call)[3]=="binomial(link = "probit\")","",
                  sd(dnorm(predict(x, type = "link"))),
                  sd(dlogis(predict(x, type = "link"))))
  marginal.effects <- pdf*coef(x)
  sim <- matrix(rep(NA,sims*length(coef(x))), nrow=sims)
  for(i in 1:length(coef(x))){
    sim[,i] <- rnorm(sims, coef(x)[i], diag(vcov(x)^0.5)[i])
  }
  pdfsim <- rnorm(sims,pdf,pdfsd)
}
```

By: Matt Bogard
sim.se <- pdfsim*sim
res <- cbind(marginal.effects, sd(sim.se))
colnames(res)[2] <- "standard.error"
ifelse(names(x$coefficients[1]) == '(Intercept)', return(res[2:nrow(res),]), return(res))

# marginal effects from logit
mfx(mylogit)

### code it yourself for marginal effects at the mean

summary(dat1)

b0 <- 5.92972   # estimated intercept from logit
b1 <- -0.14099  # estimated b from logit

xvar <- 39.5    # reference value (i.e. mean) for explanatory variable
d <- .0001      # incremental change in x

xbi <- (xvar + d)*b1 + b0
xbj <- (xvar - d)*b1 + b0
meff <- ((exp(xbi)/(1+exp(xbi)))-(exp(xbj)/(1+exp(xbj))))/(d*2) ; print(meff)

### a different perhaps easier formulation for me at the mean

XB <- xvar*b1 + b0  # this could be expanded for multiple b's or x's
meffx <- (exp(XB)/((1+exp(XB))^2))*b1
print(meffx)

### averaging the meff for the whole data set

dat1$XB <- dat1$age*b1 + b0

meffx <- (exp(dat1$XB)/((1+exp(dat1$XB))^2))*b1
summary(meffx) # get mean

### marginal effects from linear model

lpm <- lm(dat1$participate~dat1$age)
summary(lpm)

#---------------------------------------------------------------
#
# multivariable case
#
#
```r
#---------------------------------------------------------------

dat2 <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")

head(dat2)

summary(dat2) # summary stats

### run logistic regression model

mylogit <- glm(admit ~ gre + gpa, data = dat2, family = "binomial")

summary(mylogit)

exp(cbind(OR = coef(mylogit), confint(mylogit))) # get odds ratios

# marginal effects from logit
mfx(mylogit)

### code it yourself for marginal effects at the mean

summary(dat1)

b0 <- -4.949378    # estimated intercept from logit
b1 <- 0.002691     # estimated b for gre
b2 <- 0.754687    # estimated b for gpa

x1 <- 587        # reference value (i.e. mean) for gre
x2 <- 3.39       # reference value (i.e. mean) for gpa

# incremental change in x

d <- .0001

# meff at means for gre

xbi <- (x1 + d)*b1 + b2*x2 + b0

xbj <- (x1 - d)*b1 + b2*x2 + b0

meff <- ((exp(xbi)/(1+exp(xbi)))-(exp(xbj)/(1+exp(xbj))))/(d*2)

print(meff)

# meff at means for gpa

xbi <- (x2 + d)*b2 + b1*x1 + b0

xbj <- (x2 - d)*b2 + b1*x1 + b0

meff <- ((exp(xbi)/(1+exp(xbi)))-(exp(xbj)/(1+exp(xbj))))/(d*2)

print(meff)

### a different perhaps easier formulation for me at the mean

XB <- x1*b1 +x2*b2 + b0 # this could be expanded for multiple b's or x's

# meff at means for gre

meffx <- (exp(XB)/((1+exp(XB))^2))*b1

print(meffx)
```
# meff at means for gpa
meffx <- \( \frac{\exp(XB)}{(1+\exp(XB))^2} \)*b2
print(meffx)

### averaging the meff for the whole data set

dat2$XB <- dat2$gre*b1 + dat2$gpa*b2 + b0

# sample avg meff for gre
meffx <- \( \frac{\exp(dat2$XB)}{(1+\exp(dat2$XB))^2} \)*b1
summary(meffx)  # get mean

# sample avg meff for gpa
meffx <- \( \frac{\exp(dat2$XB)}{(1+\exp(dat2$XB))^2} \)*b2
summary(meffx)  # get mean

#### marginal effects from linear model

lpm <- lm(admit ~ gre + gpa, data = dat2)
summary(lpm)