THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

ECON2300 - Introductory Econometrics

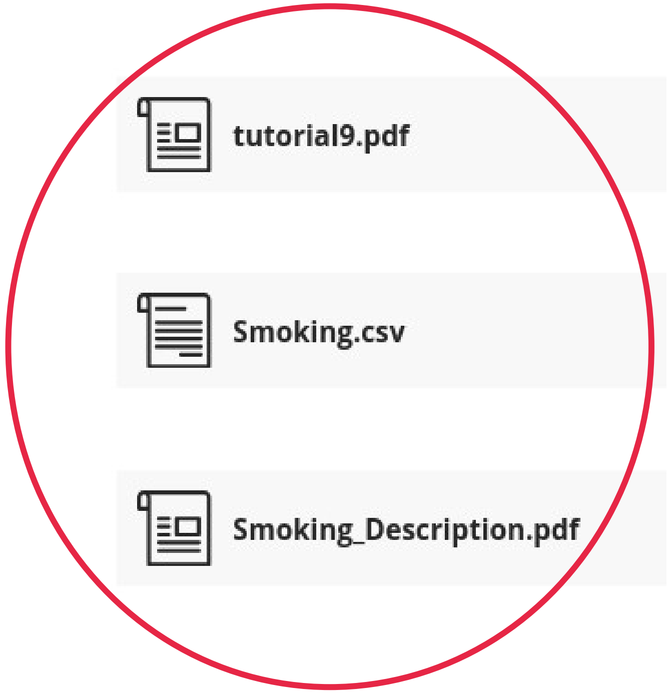
Tutorial 9: Regression with a Binary Dependent Variable

Tutor: Francisco Tavares Garcia

Quiz 7 is now available under the Assessment folder.

The due date is Thursday, 9th September, 16:00.

- Download the files for tutorial 09 from Blackboard,
- save them into a folder for this tutorial.



tutorial9.pdf

Smoking.csv

Smoking_Description.pdf

Now, let’s download the script for the tutorial.

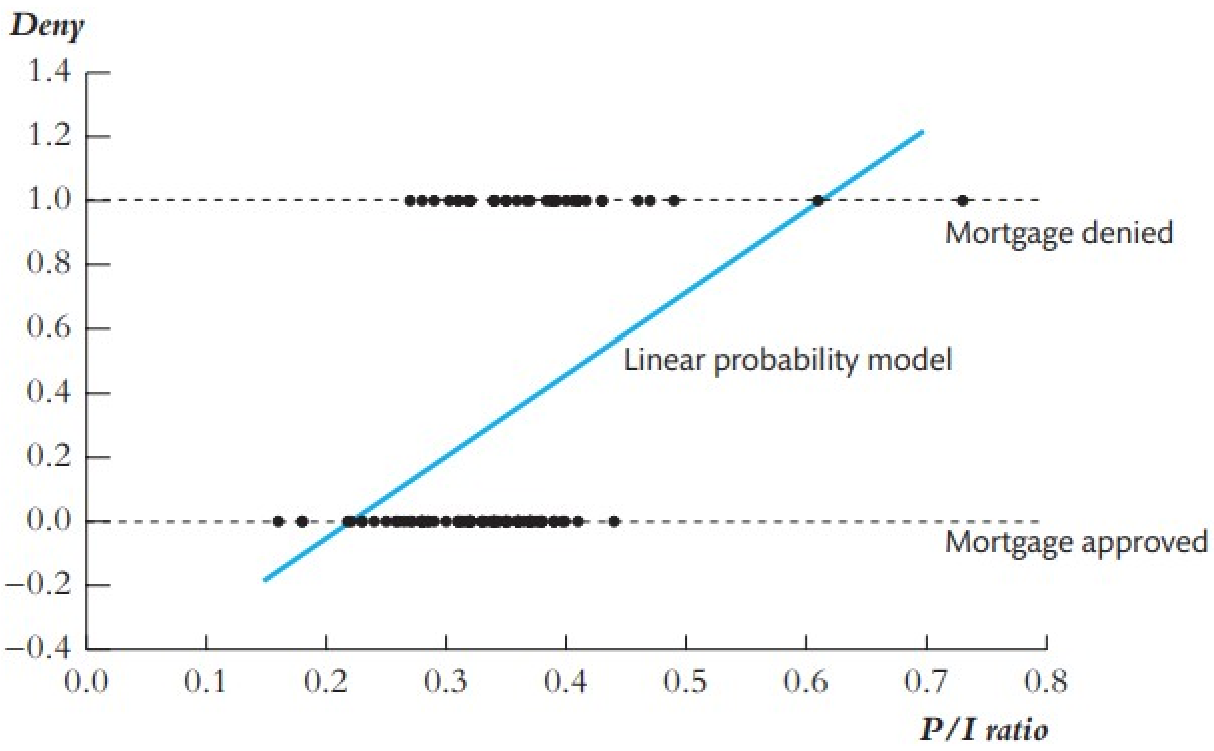
- Copy the code from Github,
  - https://github.com/tavaresgarcia/teaching
- Save the scripts in the same folder as the data.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

$$\underset{}{\widehat{deny}} = \underset{(0.032)}{-0.080} + \underset{(0.098)}{0.604}\ P/I\ ratio. \qquad (11.1)$$

FIGURE 11.1 Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

Mortgage applicants with a high ratio of debt payments to income (*P/I ratio*) are more likely to have their application denied (*deny* = 1 if denied; *deny* = 0 if approved). The linear probability model uses a straight line to model the probability of denial, conditional on the *P/I ratio*.



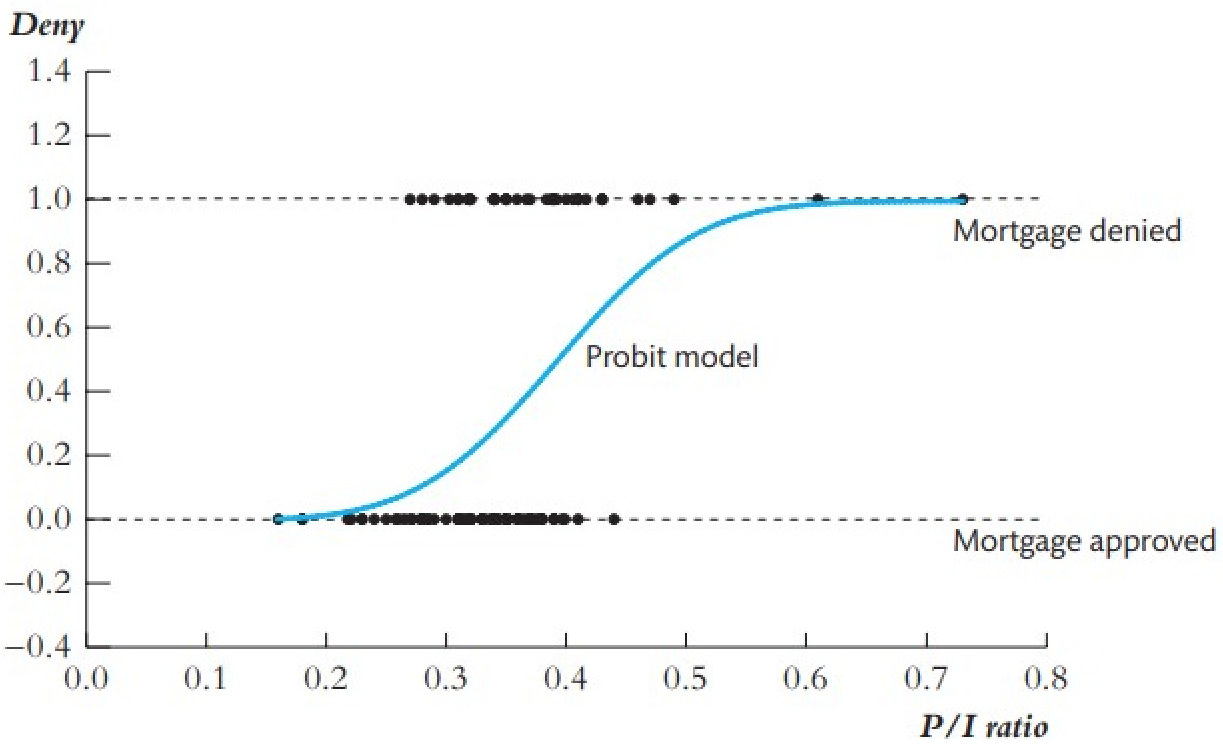
Source: Stock, J. H., & Watson, M. W. (2019). Introduction to econometrics (Fourth edition, global edition.). Pearson Education Limited.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

$$\widehat{\Pr(deny = 1 \mid P/I\ ratio)} = \underset{(0.16)}{\Phi(-2.19} + \underset{(0.47)}{2.97} P/I\ ratio). \qquad (11.7)$$

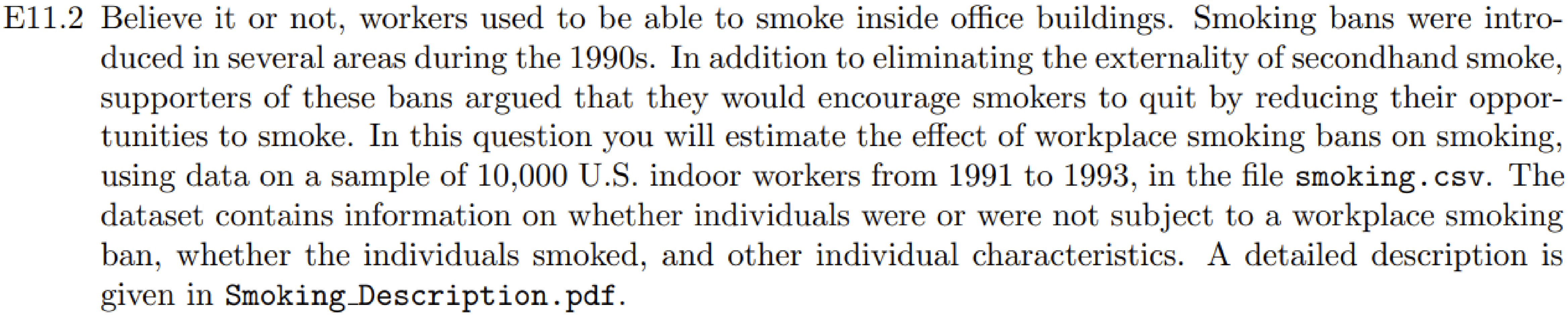
FIGURE 11.2 Probit Model of the Probability of Denial Given *P/I Ratio*

The probit model uses the cumulative normal distribution function to model the probability of denial given the payment-to-income ratio or, more generally, to model $Pr(Y = 1 | X)$. Unlike the linear probability model, the probit conditional probabilities are always between 0 and 1.



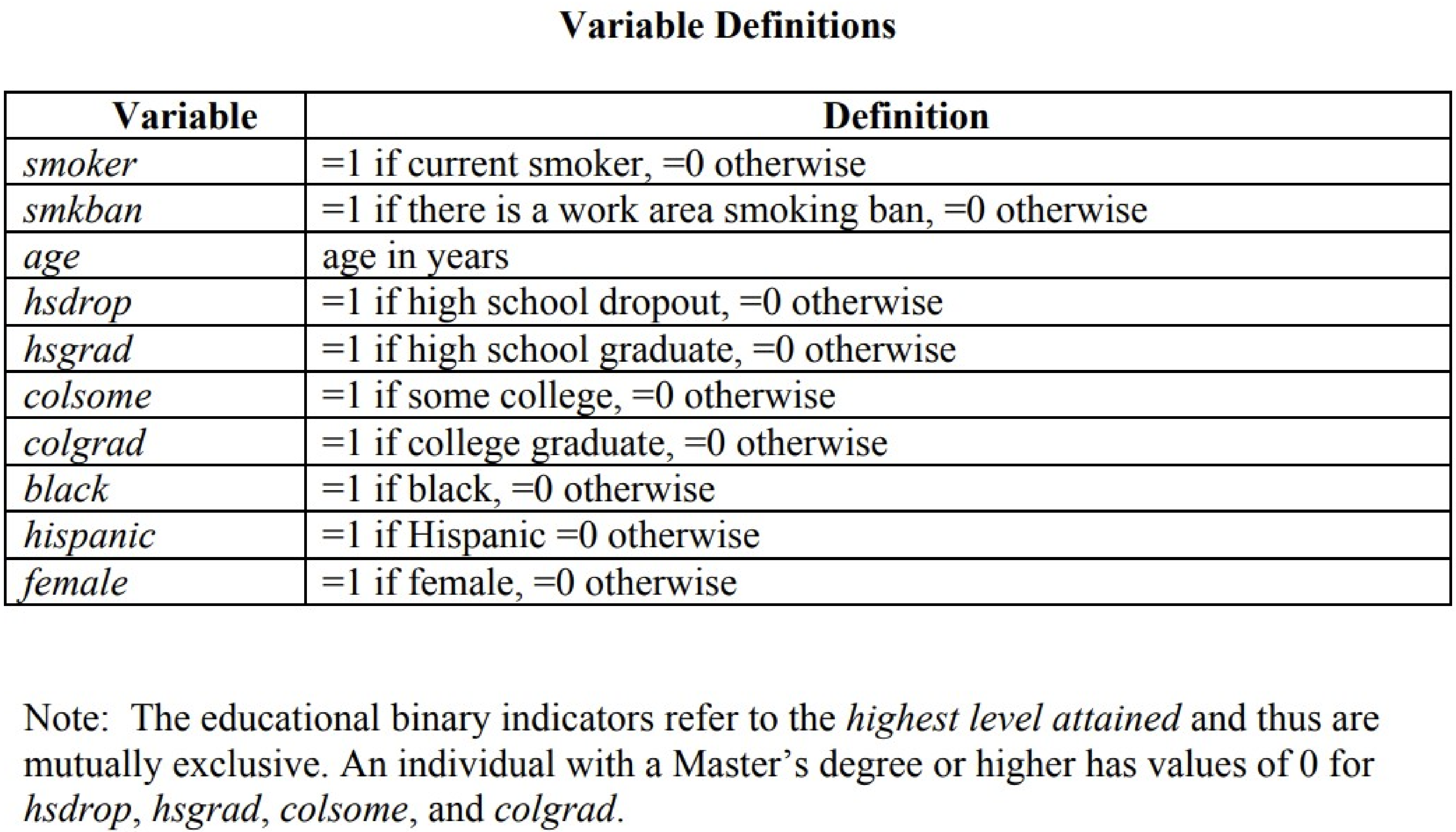
Source: Stock, J. H., & Watson, M. W. (2019). Introduction to econometrics (Fourth edition, global edition.). Pearson Education Limited.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h



E11.2 Believe it or not, workers used to be able to smoke inside office buildings. Smoking bans were introduced in several areas during the 1990s. In addition to eliminating the externality of secondhand smoke, supporters of these bans argued that they would encourage smokers to quit by reducing their opportunities to smoke. In this question you will estimate the effect of workplace smoking bans on smoking, using data on a sample of 10,000 U.S. indoor workers from 1991 to 1993, in the file `smoking.csv`. The dataset contains information on whether individuals were or were not subject to a workplace smoking ban, whether the individuals smoked, and other individual characteristics. A detailed description is given in `Smoking_Description.pdf`.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

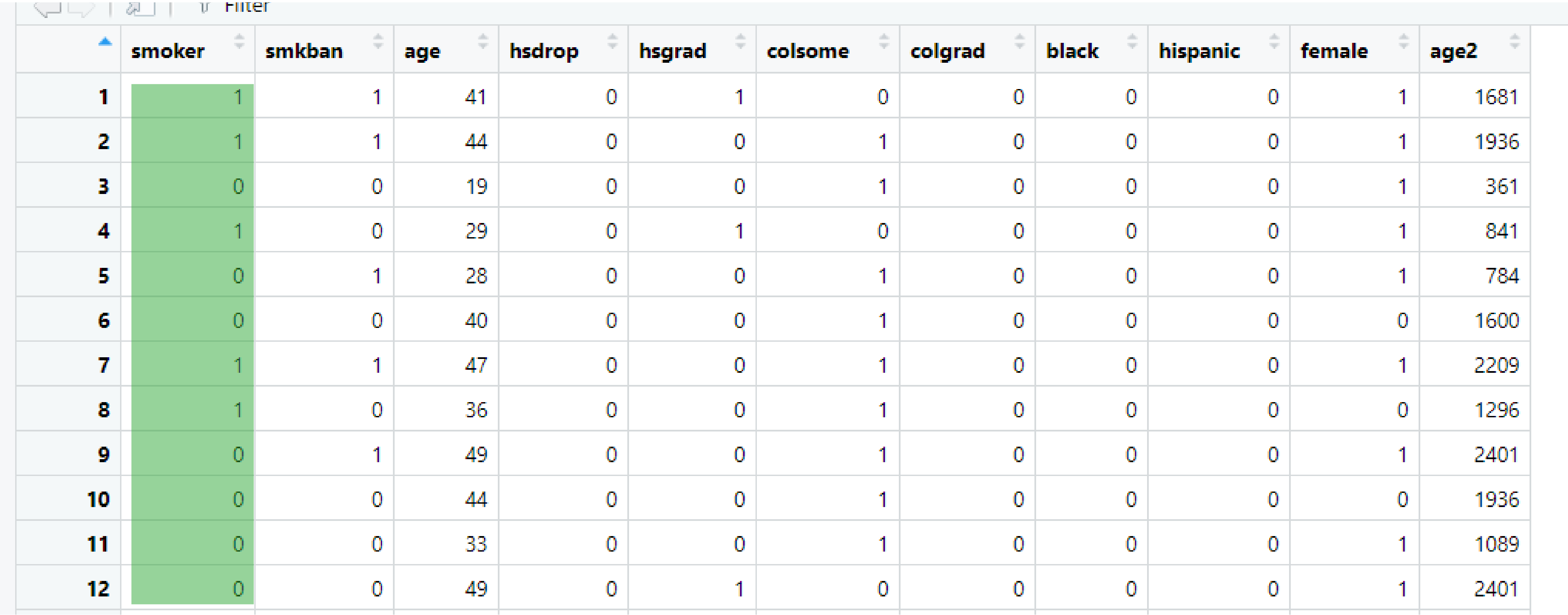


Variable Definitions

Variable	Definition
smoker	=1 if current smoker, =0 otherwise
smkban	=1 if there is a work area smoking ban, =0 otherwise
age	age in years
hsdrop	=1 if high school dropout, =0 otherwise
hsgrad	=1 if high school graduate, =0 otherwise
colsome	=1 if some college, =0 otherwise
colgrad	=1 if college graduate, =0 otherwise
black	=1 if black, =0 otherwise
hispanic	=1 if Hispanic =0 otherwise
female	=1 if female, =0 otherwise

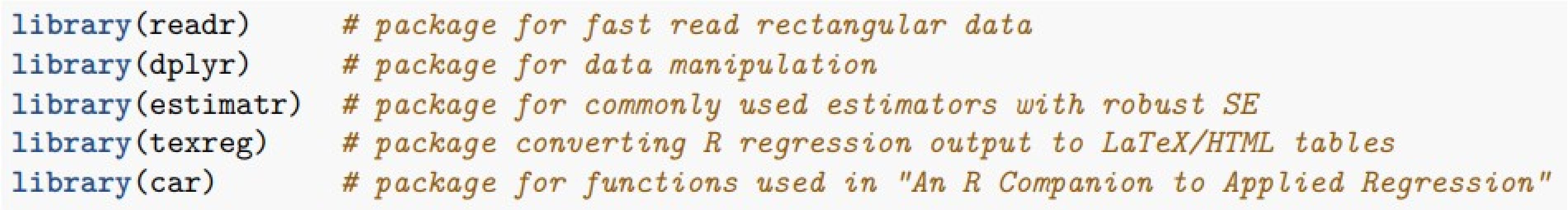
Note: The educational binary indicators refer to the *highest level attained* and thus are mutually exclusive. An individual with a Master's degree or higher has values of 0 for *hsdrop*, *hsgrad*, *colsome*, and *colgrad*.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

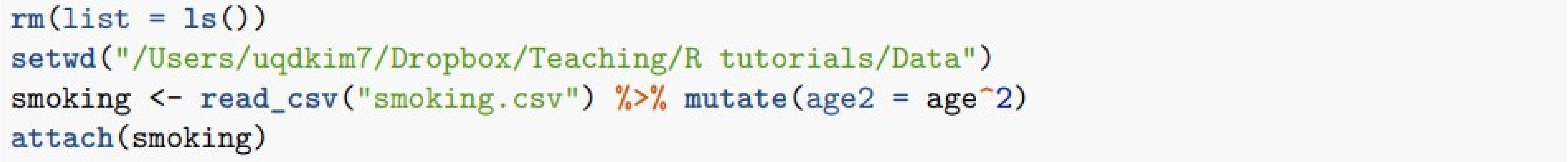


▲	smoker	smkban	age	hsdrop	hsgrad	colsome	colgrad	black	hispanic	female	age2
1	1	1	41	0	1	0	0	0	0	1	1681
2	1	1	44	0	0	1	0	0	0	1	1936
3	0	0	19	0	0	1	0	0	0	1	361
4	1	0	29	0	1	0	0	0	0	1	841
5	0	1	28	0	0	1	0	0	0	1	784
6	0	0	40	0	0	1	0	0	0	0	1600
7	1	1	47	0	0	1	0	0	0	1	2209
8	1	0	36	0	0	1	0	0	0	0	1296
9	0	1	49	0	0	1	0	0	0	1	2401
10	0	0	44	0	0	1	0	0	0	0	1936
11	0	0	33	0	0	1	0	0	0	1	1089
12	0	0	49	0	1	0	0	0	0	1	2401

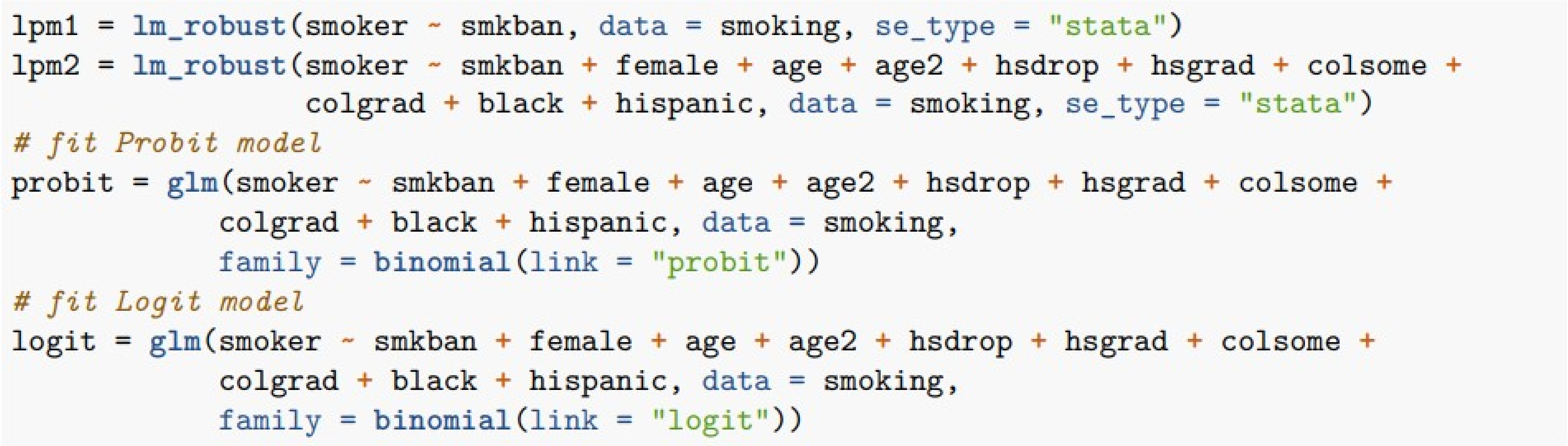
Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h



```r
library(readr)      # package for fast read rectangular data
library(dplyr)      # package for data manipulation
library(estimatr)   # package for commonly used estimators with robust SE
library(texreg)     # package converting R regression output to LaTeX/HTML tables
library(car)        # package for functions used in "An R Companion to Applied Regression"
```



```r
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Data")
smoking <- read_csv("smoking.csv") %>% mutate(age2 = age^2)
attach(smoking)
```



```r
lpm1 = lm_robust(smoker ~ smkban, data = smoking, se_type = "stata")
lpm2 = lm_robust(smoker ~ smkban + female + age + age2 + hsdrop + hsgrad + colsome +
                 colgrad + black + hispanic, data = smoking, se_type = "stata")
# fit Probit model
probit = glm(smoker ~ smkban + female + age + age2 + hsdrop + hsgrad + colsome +
             colgrad + black + hispanic, data = smoking,
             family = binomial(link = "probit"))
# fit Logit model
logit = glm(smoker ~ smkban + female + age + age2 + hsdrop + hsgrad + colsome +
            colgrad + black + hispanic, data = smoking,
            family = binomial(link = "logit"))
```

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

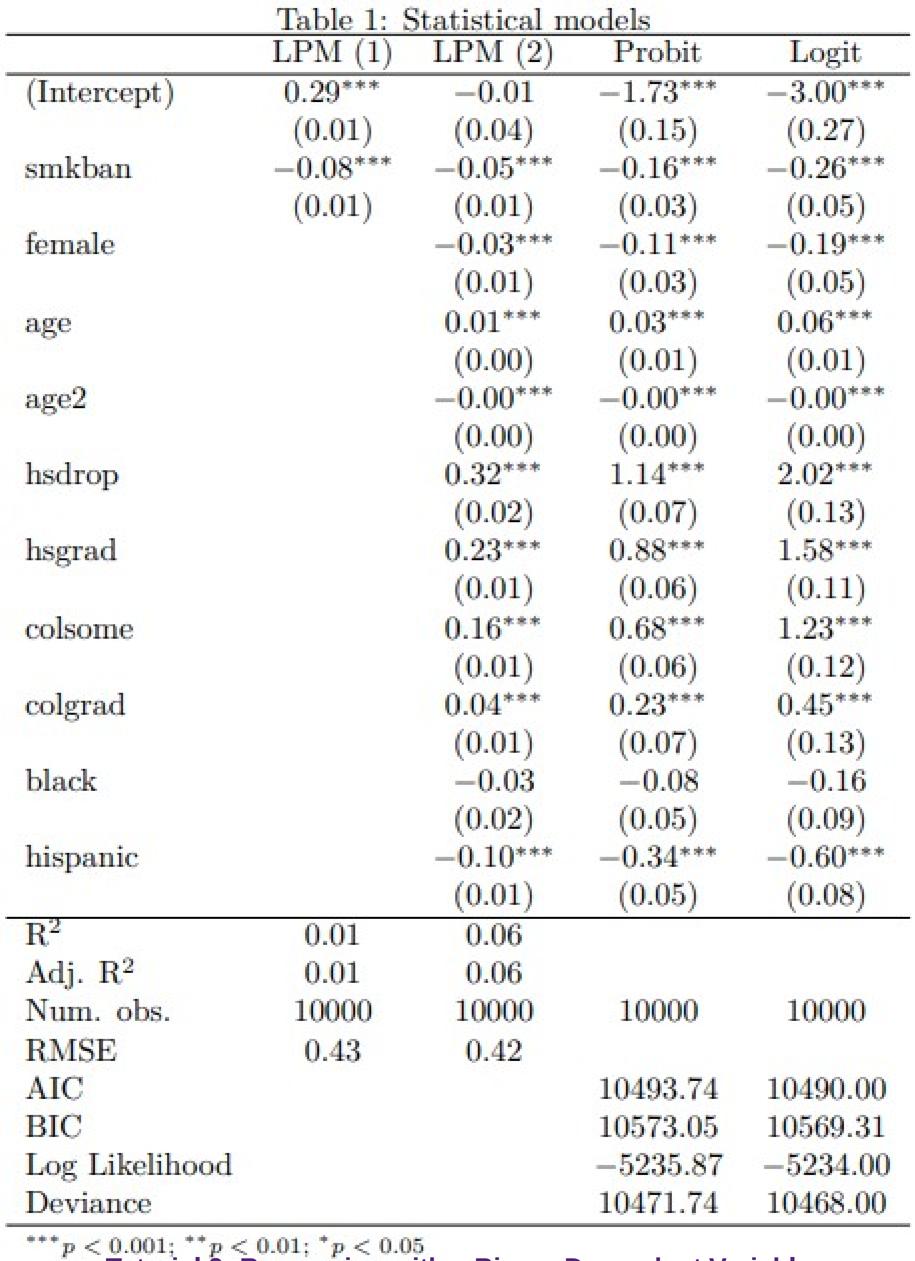


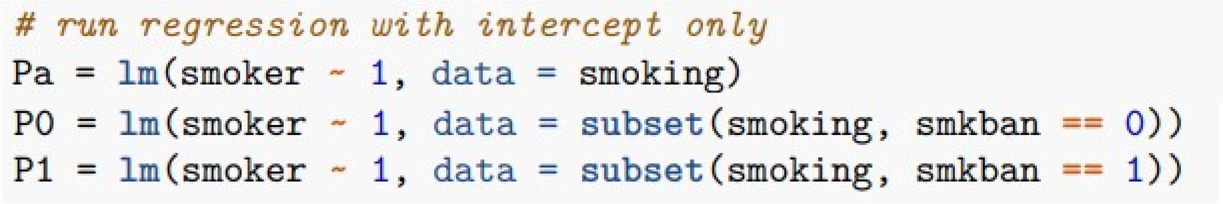
Table 1: Statistical models

	LPM (1)	LPM (2)	Probit	Logit
(Intercept)	0.29***	−0.01	−1.73***	−3.00***
	(0.01)	(0.04)	(0.15)	(0.27)
smkban	−0.08***	−0.05***	−0.16***	−0.26***
	(0.01)	(0.01)	(0.03)	(0.05)
female		−0.03***	−0.11***	−0.19***
		(0.01)	(0.03)	(0.05)
age		0.01***	0.03***	0.06***
		(0.00)	(0.01)	(0.01)
age2		−0.00***	−0.00***	−0.00***
		(0.00)	(0.00)	(0.00)
hsdrop		0.32***	1.14***	2.02***
		(0.02)	(0.07)	(0.13)
hsgrad		0.23***	0.88***	1.58***
		(0.01)	(0.06)	(0.11)
colsome		0.16***	0.68***	1.23***
		(0.01)	(0.06)	(0.12)
colgrad		0.04***	0.23***	0.45***
		(0.01)	(0.07)	(0.13)
black		−0.03	−0.08	−0.16
		(0.02)	(0.05)	(0.09)
hispanic		−0.10***	−0.34***	−0.60***
		(0.01)	(0.05)	(0.08)
R^2	0.01	0.06		
Adj. R^2	0.01	0.06		
Num. obs.	10000	10000	10000	10000
RMSE	0.43	0.42		
AIC			10493.74	10490.00
BIC			10573.05	10569.31
Log Likelihood			−5235.87	−5234.00
Deviance			10471.74	10468.00

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

(a) Estimate the probability of smoking for (i) all workers, (ii) workers affected by workplace smoking bans, and (iii) workers not affected by workplace smoking bans.



```
# run regression with intercept only
Pa = lm(smoker ~ 1, data = smoking)
P0 = lm(smoker ~ 1, data = subset(smoking, smkban == 0))
P1 = lm(smoker ~ 1, data = subset(smoking, smkban == 1))
```

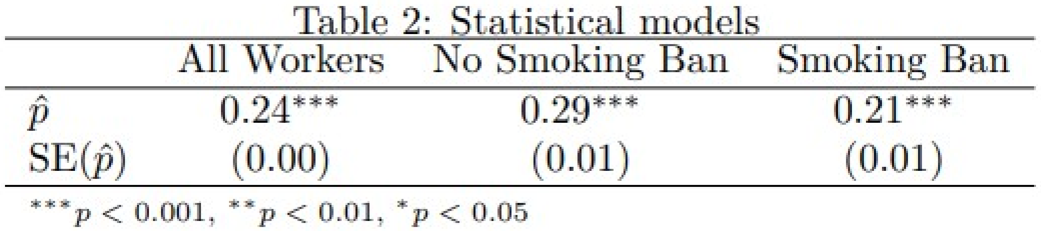


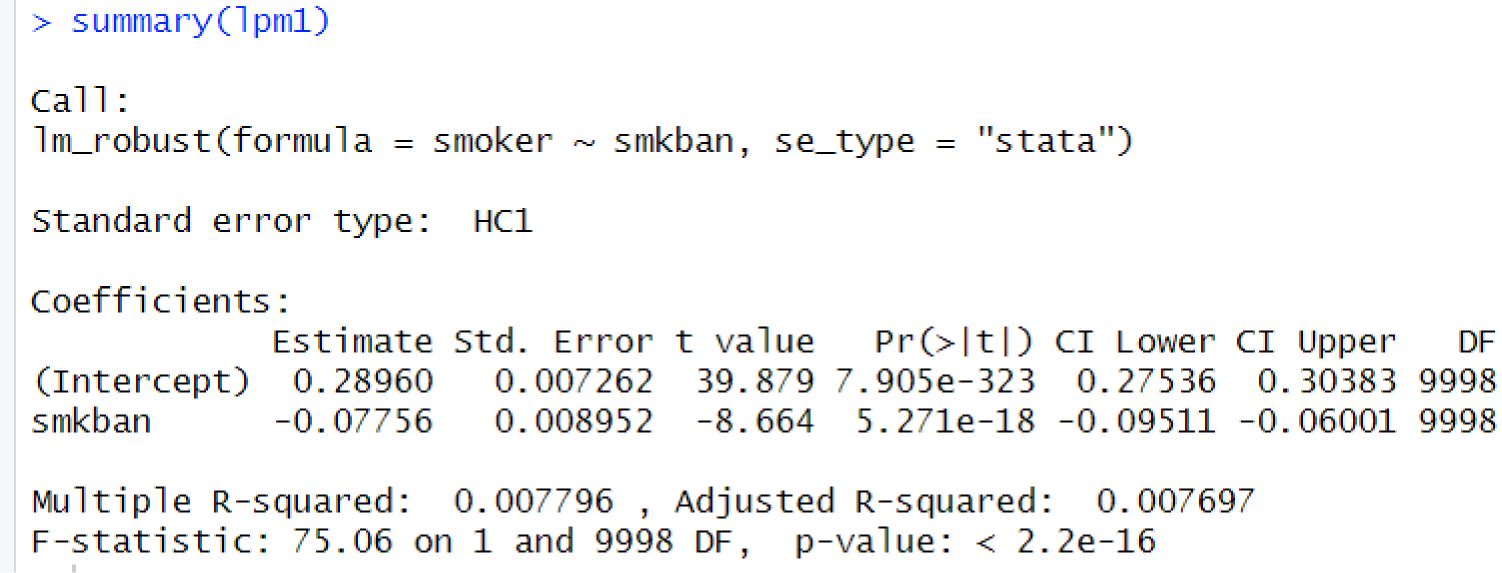
Table 2: Statistical models

	All Workers	No Smoking Ban	Smoking Ban
$\hat{p}$	0.24^{***}	0.29^{***}	0.21^{***}
SE($\hat{p}$)	(0.00)	(0.01)	(0.01)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

(b) What is the difference in the probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban? Use a linear probability model to determine whether this difference is statistically significant.



```
> summary(lpm1)

Call:
lm_robust(formula = smoker ~ smkban, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper   DF
(Intercept)  0.28960   0.007262  39.879 7.905e-323  0.27536  0.30383 9998
smkban      -0.07756   0.008952  -8.664  5.271e-18 -0.09511 -0.06001 9998

Multiple R-squared:  0.007796 , Adjusted R-squared:  0.007697
F-statistic: 75.06 on 1 and 9998 DF,  p-value: < 2.2e-16
```

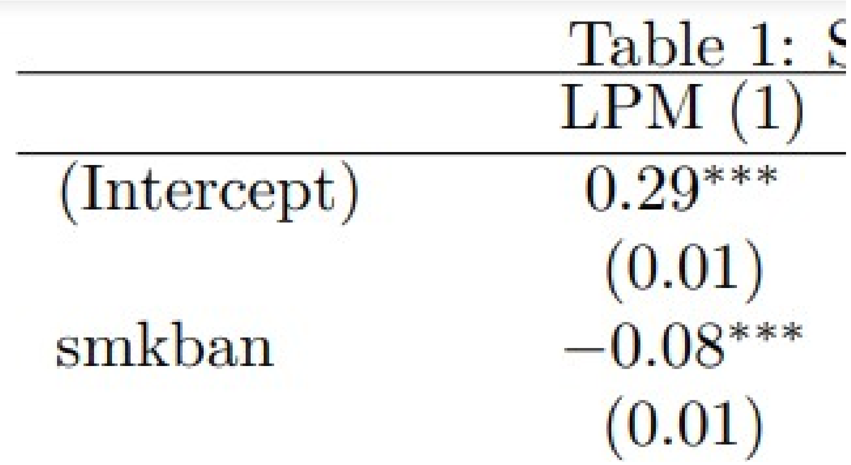


Table 1: S

	LPM (1)
(Intercept)	0.29^{***}
	(0.01)
smkban	-0.08^{***}
	(0.01)

From LPM(1), the difference is -0.08 with a standard error of 0.01.The resulting t-statistic is -8, so the coefficient is statistically significant.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

(c) Estimate a linear probability model with **smoker** as the dependent variable and the following regressors: **smkban**, **female**, **age**, **age2**, **hsdrop**, **hsgrad**, **colsome**, **colgrad**, **black**, and **hispanic**. Compare the estimated effect of a smoking ban from this regression with your answer from (b). Suggest a reason, based on the substance of this regression, explaining the change in the estimated effect of a smoking ban between (b) and (c).

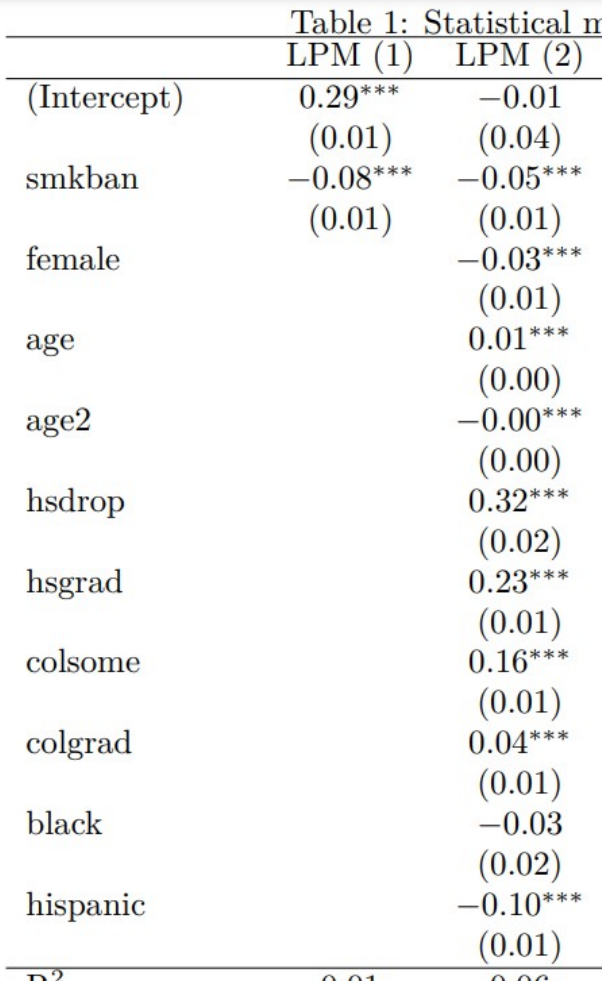
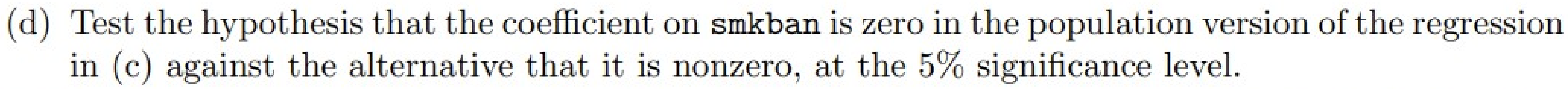


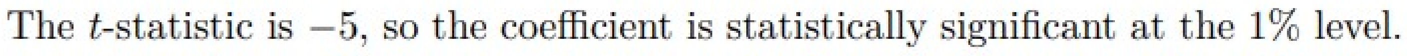
Table 1: Statistical m

	LPM (1)	LPM (2)
(Intercept)	0.29^{***}	-0.01
	(0.01)	(0.04)
smkban	-0.08^{***}	-0.05^{***}
	(0.01)	(0.01)
female		-0.03^{***}
		(0.01)
age		0.01^{***}
		(0.00)
age2		-0.00^{***}
		(0.00)
hsdrop		0.32^{***}
		(0.02)
hsgrad		0.23^{***}
		(0.01)
colsome		0.16^{***}
		(0.01)
colgrad		0.04^{***}
		(0.01)
black		-0.03
		(0.02)
hispanic		-0.10^{***}
		(0.01)
R^2	0.01	0.06

From LPM(2) the estimated difference is −0.05, smaller than the effect in LPM(1). Evidently (1) suffers from omitted variable bias. That is, **smkban** may be correlated with the education/race/gender or with age. For example, workers with a college degree are more likely to work in an office with a smoking ban than high-school dropouts, and college graduates are less likely to smoke than high-school dropouts.



(d) Test the hypothesis that the coefficient on smkban is zero in the population version of the regression in (c) against the alternative that it is nonzero, at the 5% significance level.



The t-statistic is -5, so the coefficient is statistically significant at the 1% level.

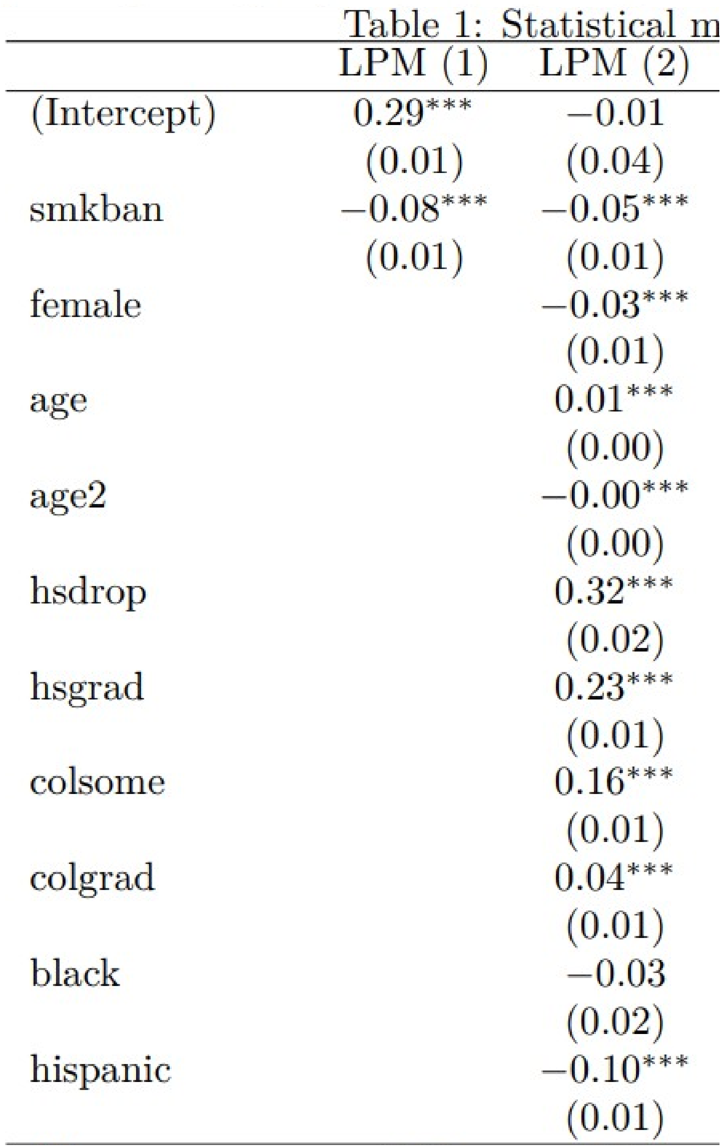
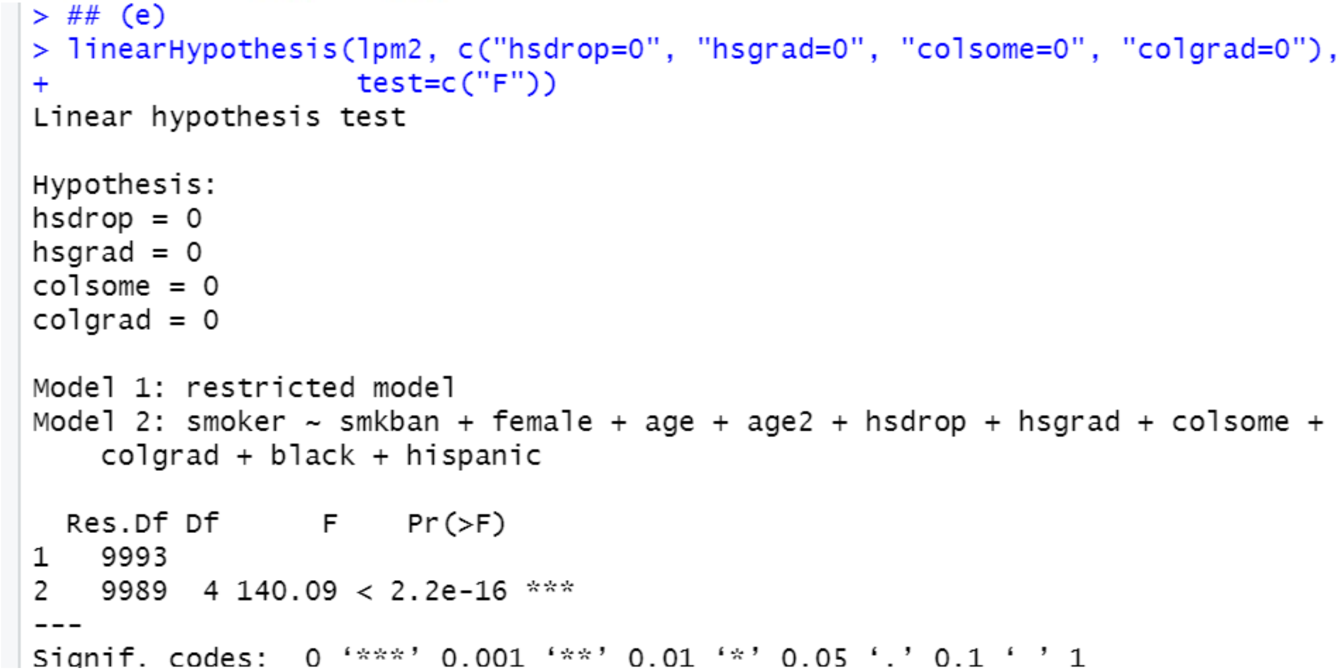


Table 1: Statistical m

	LPM (1)	LPM (2)
(Intercept)	0.29^{***}	-0.01
	(0.01)	(0.04)
smkban	-0.08^{***}	-0.05^{***}
	(0.01)	(0.01)
female		-0.03^{***}
		(0.01)
age		0.01^{***}
		(0.00)
age2		-0.00^{***}
		(0.00)
hsdrop		0.32^{***}
		(0.02)
hsgrad		0.23^{***}
		(0.01)
colsome		0.16^{***}
		(0.01)
colgrad		0.04^{***}
		(0.01)
black		-0.03
		(0.02)
hispanic		-0.10^{***}
		(0.01)

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

(e) Test the hypothesis that the probability of smoking does not depend on the level of education in the regression in (c). Does the probability of smoking increase or decrease with the level of education?



```
> ## (e)
> linearHypothesis(lpm2, c("hsdrop=0", "hsgrad=0", "colsome=0", "colgrad=0"),
+                  test=c("F"))
Linear hypothesis test

Hypothesis:
hsdrop = 0
hsgrad = 0
colsome = 0
colgrad = 0

Model 1: restricted model
Model 2: smoker ~ smkban + female + age + age2 + hsdrop + hsgrad + colsome +
    colgrad + black + hispanic

  Res.Df Df      F    Pr(>F)
1   9993
2   9989  4 140.09 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-statistic has a p-value of 0.00, so the coefficients are significant. The omitted education status is “Masters degree or higher.” Thus the coefficients show the increase in probability relative to someone with a postgraduate degree. For example, the coefficient on `colgrad` is 0.045, so the probability of smoking for a college graduate is 0.04 (4%) higher than for someone with a postgraduate degree. Similarly, the coefficient on `hsdrop` is 0.32, so the probability of smoking for a high school dropout is 0.32 (32%) higher than for someone with a postgraduate degree. Because the coefficients are all positive and get smaller as educational attainment increases, the probability of smoking falls as educational attainment increases.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

(f) Repeat (c)–(e) using a probit model.

Probit Regression, continued

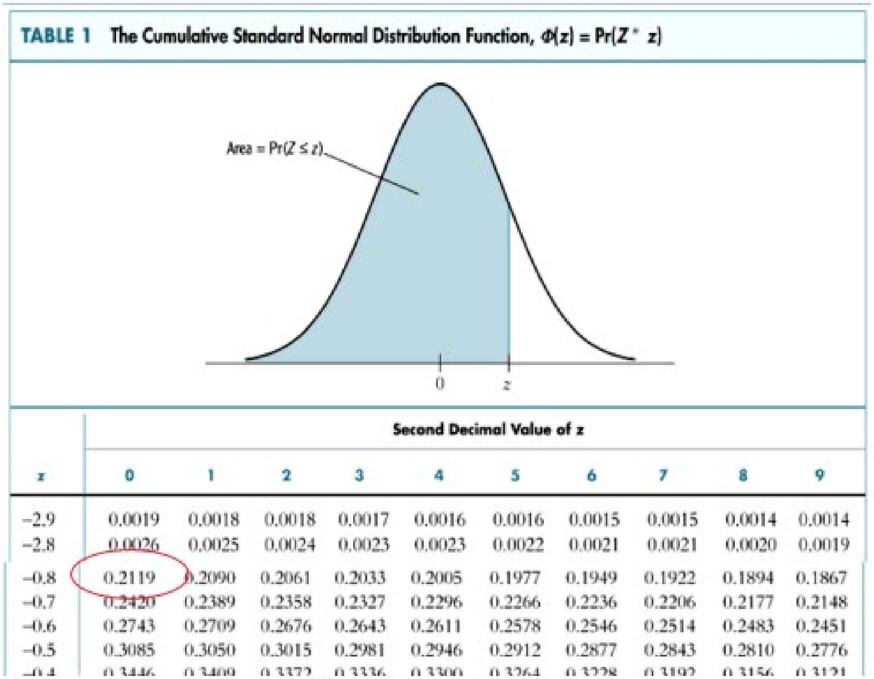


TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z) = \Pr(Z \leq z)$

Area = Pr(Z ≤ z)
0
z

	Second Decimal Value of z									
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

So, $\Pr(Y_i = 1 | X_i = 0.4) = \Phi(-0.8) = 0.2119$.

Table 1: Statistical models

	LPM (1)	LPM (2)	Probit	Logit
(Intercept)	0.29***	−0.01	−1.73***	−3.00***
	(0.01)	(0.04)	(0.15)	(0.27)
smkban	−0.08***	−0.05***	−0.16***	−0.26***
	(0.01)	(0.01)	(0.03)	(0.05)
female		−0.03***	−0.11***	−0.19***
		(0.01)	(0.03)	(0.05)
age		0.01***	0.03***	0.06***
		(0.00)	(0.01)	(0.01)
age2		−0.00***	−0.00***	−0.00***
		(0.00)	(0.00)	(0.00)
hsdrop		0.32***	1.14***	2.02***
		(0.02)	(0.07)	(0.13)
hsgrad		0.23***	0.88***	1.58***
		(0.01)	(0.06)	(0.11)
colsome		0.16***	0.68***	1.23***
		(0.01)	(0.06)	(0.12)
colgrad		0.04***	0.23***	0.45***
		(0.01)	(0.07)	(0.13)
black		−0.03	−0.08	−0.16
		(0.02)	(0.05)	(0.09)
hispanic		−0.10***	−0.34***	−0.60***
		(0.01)	(0.05)	(0.08)

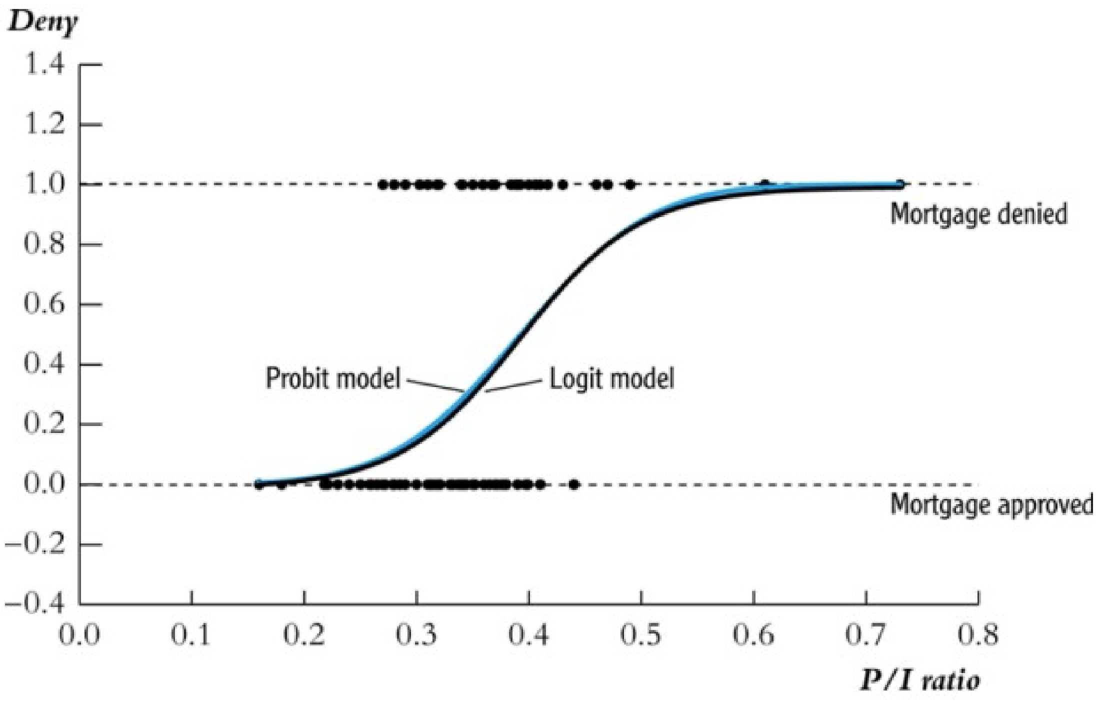
The estimated effect of the smoking ban in the probit model depends on the values of the other variable included in the regression. The estimated effects for various values of these regressors is given question (h). The t-statistic for the coefficient on smkban is −5.33, very similar to the value for the linear probability and probit models. The F-statistic is significant at the 1% level, as in the linear probability model.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

(g) Repeat (c)–(e) using a logit model.

Comparison: probit vs logit

The predicted probabilities from the probit and logit models are very close in these HMDA regressions:



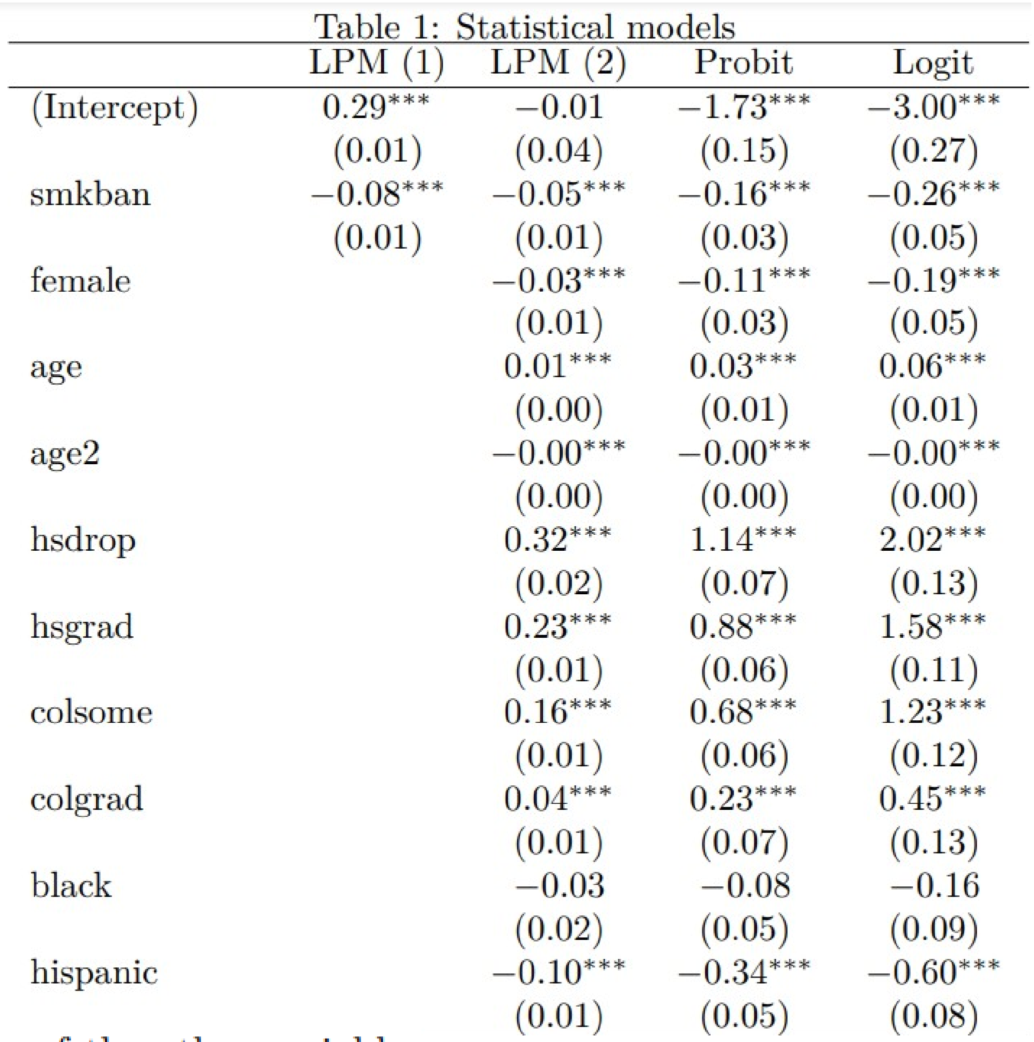
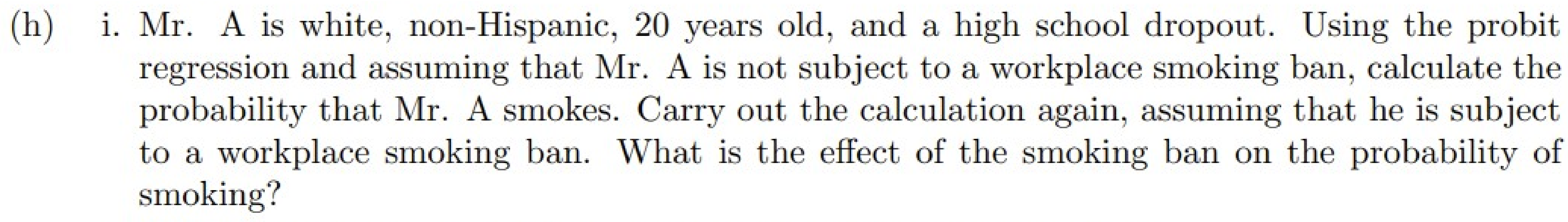


Table 1: Statistical models

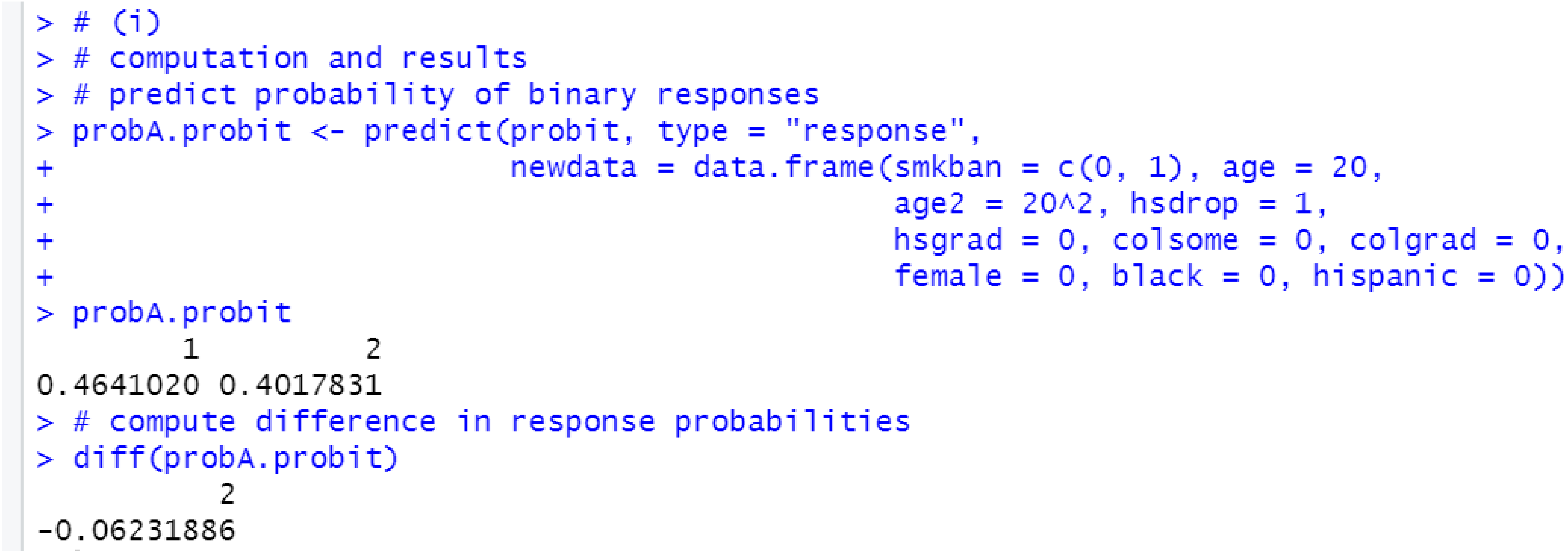
	LPM (1)	LPM (2)	Probit	Logit
(Intercept)	0.29***	−0.01	−1.73***	−3.00***
	(0.01)	(0.04)	(0.15)	(0.27)
smkban	−0.08***	−0.05***	−0.16***	−0.26***
	(0.01)	(0.01)	(0.03)	(0.05)
female		−0.03***	−0.11***	−0.19***
		(0.01)	(0.03)	(0.05)
age		0.01***	0.03***	0.06***
		(0.00)	(0.01)	(0.01)
age2		−0.00***	−0.00***	−0.00***
		(0.00)	(0.00)	(0.00)
hsdrop		0.32***	1.14***	2.02***
		(0.02)	(0.07)	(0.13)
hsgrad		0.23***	0.88***	1.58***
		(0.01)	(0.06)	(0.11)
colsome		0.16***	0.68***	1.23***
		(0.01)	(0.06)	(0.12)
colgrad		0.04***	0.23***	0.45***
		(0.01)	(0.07)	(0.13)
black		−0.03	−0.08	−0.16
		(0.02)	(0.05)	(0.09)
hispanic		−0.10***	−0.34***	−0.60***
		(0.01)	(0.05)	(0.08)

The estimated effect of the smoking ban in the logit model depends on the values of the other variable included in the regression. The estimated effects for various values of these regressors is given question (h). The t-statistic for the coefficient on `smkban` is -5.2, very similar to the value for the linear probability and probit models. The F-statistic is significant at the 1% level, as in the linear probability model.

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

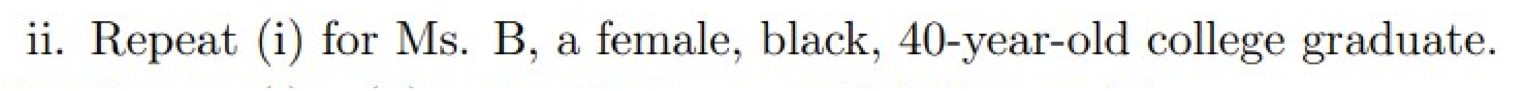


(h) i. Mr. A is white, non-Hispanic, 20 years old, and a high school dropout. Using the probit regression and assuming that Mr. A is not subject to a workplace smoking ban, calculate the probability that Mr. A smokes. Carry out the calculation again, assuming that he is subject to a workplace smoking ban. What is the effect of the smoking ban on the probability of smoking?

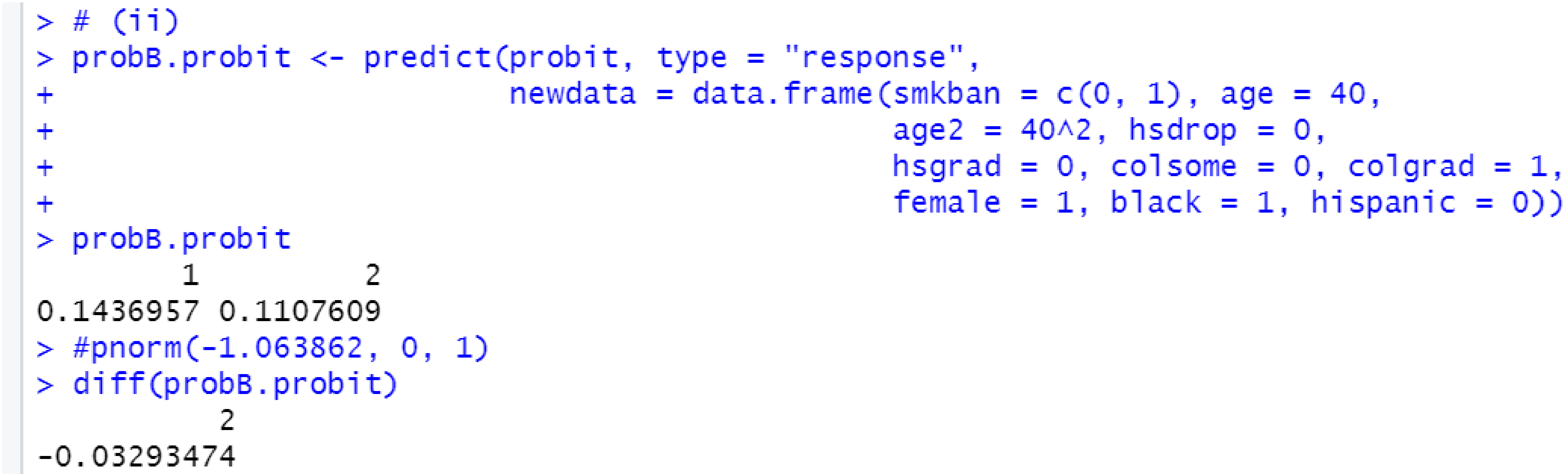


```
> # (i)
> # computation and results
> # predict probability of binary responses
> probA.probit <- predict(probit, type = "response",
+                         newdata = data.frame(smkban = c(0, 1), age = 20,
+                                              age2 = 20^2, hsdrop = 1,
+                                              hsgrad = 0, colsome = 0, colgrad = 0,
+                                              female = 0, black = 0, hispanic = 0))
> probA.probit
        1         2 
0.4641020 0.4017831 
> # compute difference in response probabilities
> diff(probA.probit)
          2 
-0.06231886 
```

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

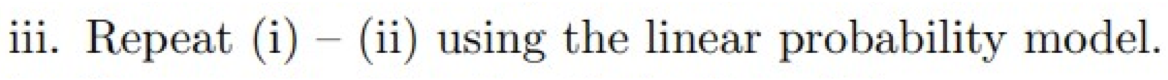


ii. Repeat (i) for Ms. B, a female, black, 40-year-old college graduate.

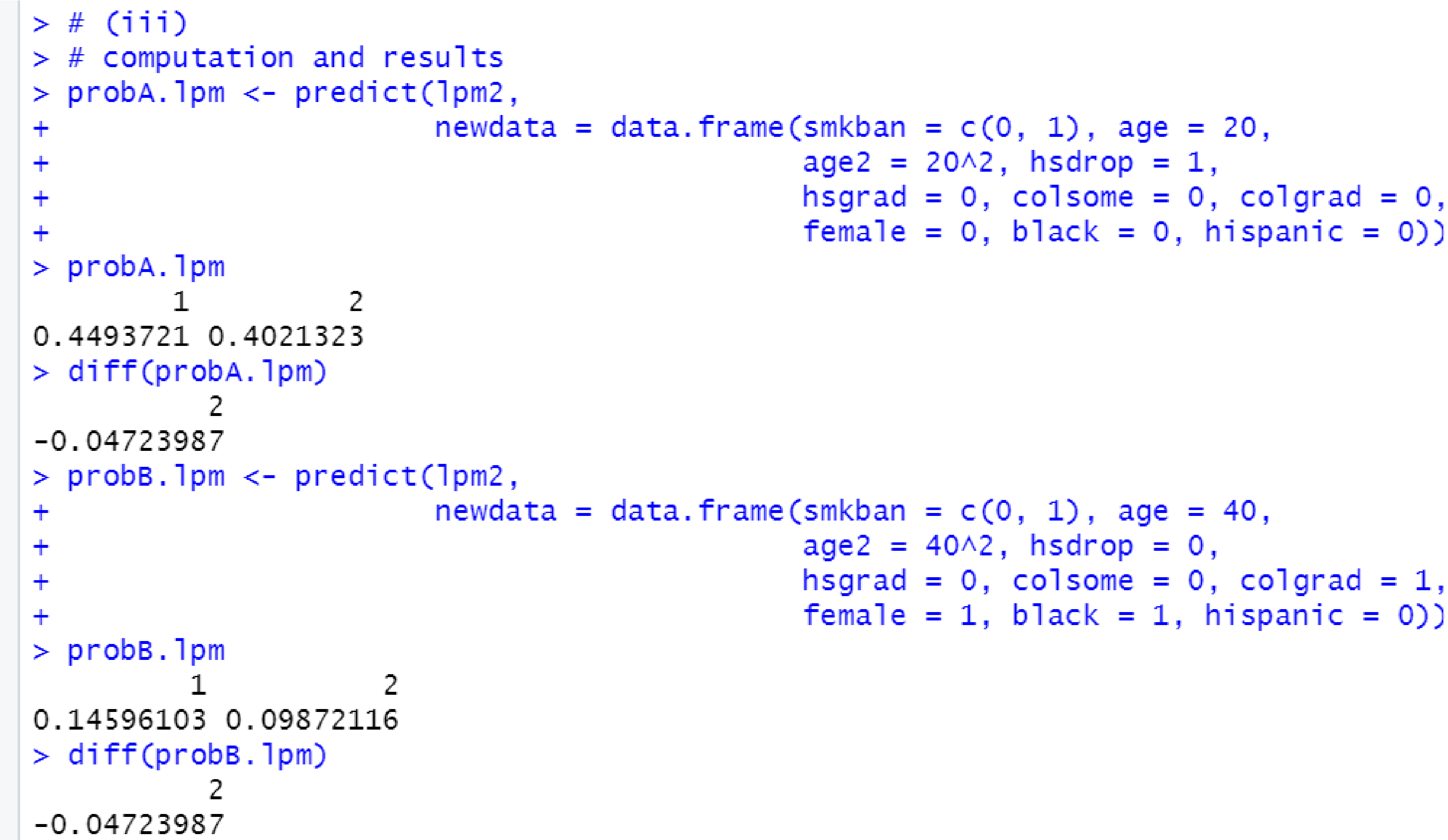


```
> # (ii)
> probB.probit <- predict(probit, type = "response",
+                         newdata = data.frame(smkban = c(0, 1), age = 40,
+                                              age2 = 40^2, hsdrop = 0,
+                                              hsgrad = 0, colsome = 0, colgrad = 1,
+                                              female = 1, black = 1, hispanic = 0))
> probB.probit
        1         2 
0.1436957 0.1107609 
> #pnorm(-1.063862, 0, 1)
> diff(probB.probit)
          2 
-0.03293474 
```

Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h

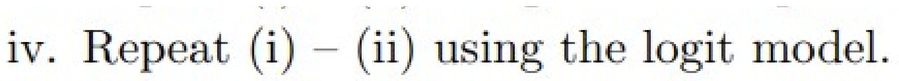


iii. Repeat (i) – (ii) using the linear probability model.

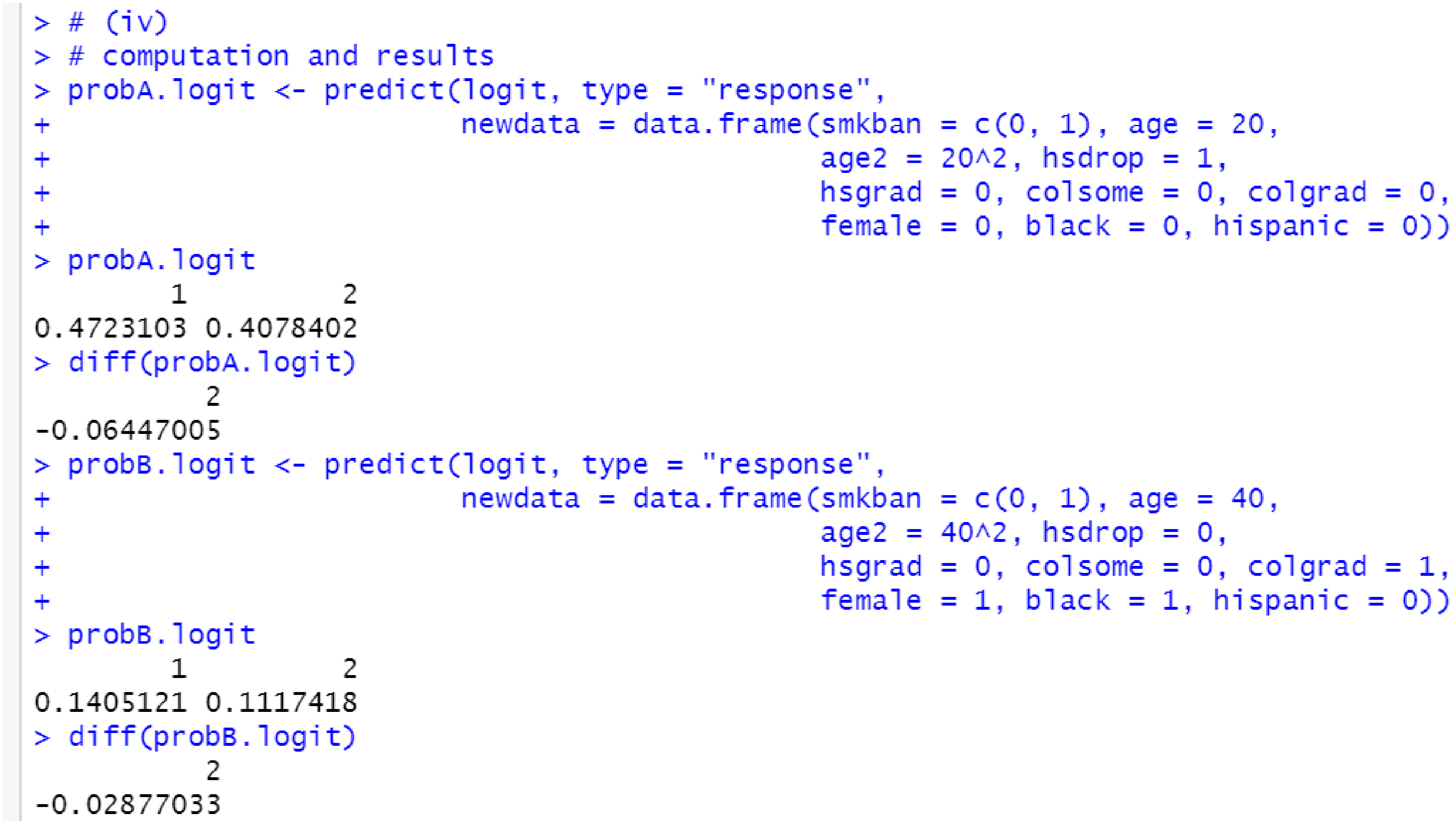


```
> # (iii)
> # computation and results
> probA.lpm <- predict(lpm2,
+                       newdata = data.frame(smkban = c(0, 1), age = 20,
+                                            age2 = 20^2, hsdrop = 1,
+                                            hsgrad = 0, colsome = 0, colgrad = 0,
+                                            female = 0, black = 0, hispanic = 0))
> probA.lpm
        1         2 
0.4493721 0.4021323 
> diff(probA.lpm)
          2 
-0.04723987 
> probB.lpm <- predict(lpm2,
+                       newdata = data.frame(smkban = c(0, 1), age = 40,
+                                            age2 = 40^2, hsdrop = 0,
+                                            hsgrad = 0, colsome = 0, colgrad = 1,
+                                            female = 1, black = 1, hispanic = 0))
> probB.lpm
         1          2 
0.14596103 0.09872116 
> diff(probB.lpm)
          2 
-0.04723987 
```

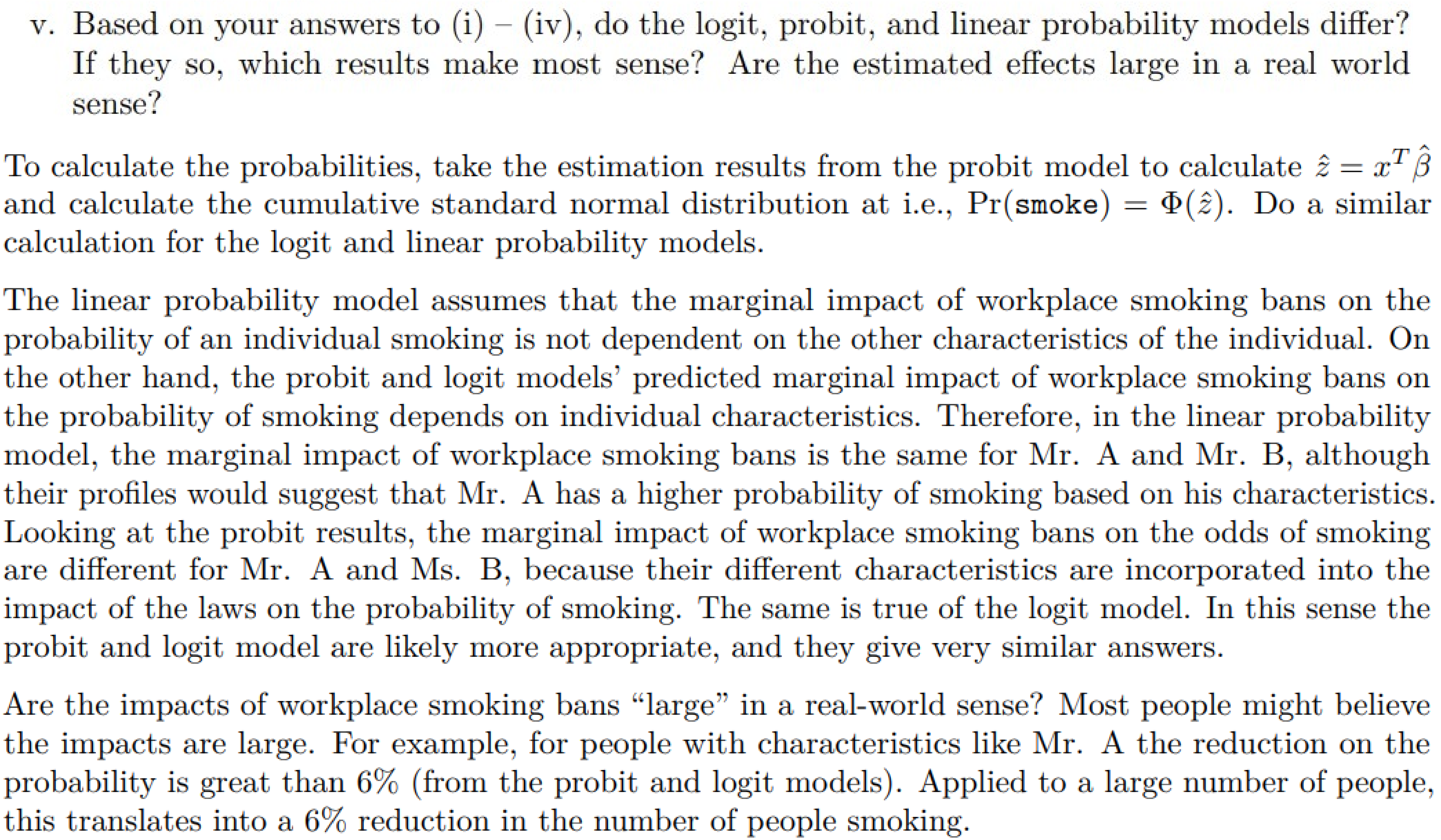
Adm · Tut 09 · E11.2 · a · b · c · d · e · f · g · h



iv. Repeat (i) – (ii) using the logit model.



```
> # (iv)
> # computation and results
> probA.logit <- predict(logit, type = "response",
+                         newdata = data.frame(smkban = c(0, 1), age = 20,
+                                              age2 = 20^2, hsdrop = 1,
+                                              hsgrad = 0, colsome = 0, colgrad = 0,
+                                              female = 0, black = 0, hispanic = 0))
> probA.logit
        1         2 
0.4723103 0.4078402 
> diff(probA.logit)
          2 
-0.06447005 
> probB.logit <- predict(logit, type = "response",
+                         newdata = data.frame(smkban = c(0, 1), age = 40,
+                                              age2 = 40^2, hsdrop = 0,
+                                              hsgrad = 0, colsome = 0, colgrad = 1,
+                                              female = 1, black = 1, hispanic = 0))
> probB.logit
        1         2 
0.1405121 0.1117418 
> diff(probB.logit)
          2 
-0.02877033 
```



v. Based on your answers to (i) – (iv), do the logit, probit, and linear probability models differ? If they so, which results make most sense? Are the estimated effects large in a real world sense?

To calculate the probabilities, take the estimation results from the probit model to calculate $\hat{z} = x^T\hat{\beta}$ and calculate the cumulative standard normal distribution at i.e., $\Pr(\texttt{smoke}) = \Phi(\hat{z})$. Do a similar calculation for the logit and linear probability models.

The linear probability model assumes that the marginal impact of workplace smoking bans on the probability of an individual smoking is not dependent on the other characteristics of the individual. On the other hand, the probit and logit models' predicted marginal impact of workplace smoking bans on the probability of smoking depends on individual characteristics. Therefore, in the linear probability model, the marginal impact of workplace smoking bans is the same for Mr. A and Mr. B, although their profiles would suggest that Mr. A has a higher probability of smoking based on his characteristics. Looking at the probit results, the marginal impact of workplace smoking bans on the odds of smoking are different for Mr. A and Ms. B, because their different characteristics are incorporated into the impact of the laws on the probability of smoking. The same is true of the logit model. In this sense the probit and logit model are likely more appropriate, and they give very similar answers.

Are the impacts of workplace smoking bans "large" in a real-world sense? Most people might believe the impacts are large. For example, for people with characteristics like Mr. A the reduction on the probability is great than 6% (from the probit and logit models). Applied to a large number of people, this translates into a 6% reduction in the number of people smoking.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Thank you

Francisco Tavares Garcia
Academic Tutor | School of Economics
tavaresgarcia.github.io

Reference

Stock, J. H., & Watson, M. W. (2019). Introduction to Econometrics, Global Edition, 4th edition. Pearson Education Limited.

CRICOS code 00025B