

Chapter 7: Multiple Regression Analysis with Qualitative Information

Introductory Econometrics: A Modern Approach

7.1 Describing Qualitative Information


Examples: gender, race, industry, region, rating grade...

A way to incorporate qualitative information is to use dummy variables.


They may appear as the dependent or as independent variables

A single dummy independent variable

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$



= the wage gain/loss if the person is a woman rather than a man
(holding other things fixed)



Dummy variable:
= 1 if the person is a woman
= 0 if the person is a man

7.2 A Single Dummy Independent Variable


$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

$$\delta_0 = E(wage \mid female = 1, educ) - E(wage \mid female = 0, educ)$$


$$\delta_0 = E(wage \mid female, educ) - E(wage \mid male, educ)$$

The key here is that the level of education is the same in both expectations; the difference, δ_0 , is due to gender only.

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$



= the wage gain/loss if the person is a woman rather than a man
(holding other things fixed)



Dummy variable:
= 1 if the person is a woman
= 0 if the person is a man

Dummy variable trap

Why don't we include a dummy variable, say *male*, which is one for males and zero for females?

$$wage = \beta_0 + \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

This model cannot be estimated due to perfect collinearity.

The **dummy variable trap** refers to the problem that not all categories can be included in the regression and one category needs to be left out, which is called a base or reference category.

When using dummy variables, one category always has to be omitted:

- Base category: men

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

- Base category: women

$$wage = \beta_0 + \gamma_0 male + \beta_1 educ + u$$

Alternatively, one could omit the intercept

$$wage = \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

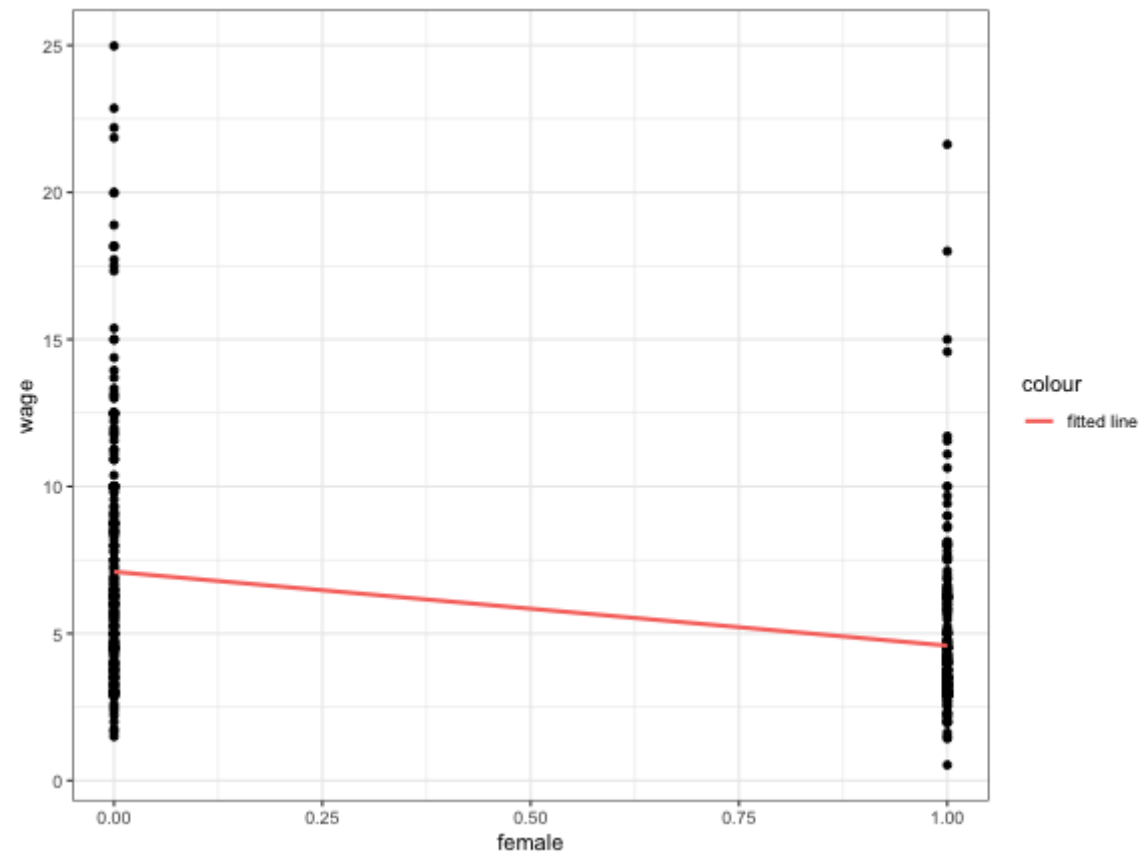
1. More difficult to test for differences between the parameters γ_0 and δ_0 .
2. R-squared formula invalid without an intercept

```
data(wage1, package='wooldridge')

# Average wage
#lm(wage ~ 1, wage1) %>% summary
#wage1 %>% select(wage) %>% stargazer(type = "text")

# Regression of wage on female
#lm(wage ~ female, wage1) %>% summary

# Graph of wage on female
ggplot(wage1, aes(x = female, y = wage)) +
  theme_bw() +
  geom_point() +
  geom_smooth(aes(col = 'fitted line'), method = "lm", se = F)
```



```
data(wage1, package='wooldridge')
#created the dummy variables (male) using the ifelse() function.
# we assign value 0 to male, if female variable is 1; or
# we assign value 1 to male, if female variable is 0
wage1$male<- ifelse(wage1$female == 1, 0, 1)

# Regression of wage on female
res1<-feols(wage ~ female, wage1)
res2<-feols(wage ~ male, wage1)
res3<-feols(wage ~ male+female-1, wage1)

models<- list(res1,res2,res3)
modelsummary(models,stars = TRUE,fmt = 2,gof_omit = "R2 | R2 Within |AIC|BIC|Log.Lik.|R2 Pseudo",
```


	Model 1	Model 2	Model 3
(Intercept)	7.10***	4.59***	
	(0.21)	(0.22)	
female	-2.51***		4.59***
	(0.30)		(0.22)
male		2.51***	7.10***
		(0.30)	(0.21)
Num.Obs.	526	526	526
R2	0.116	0.116	0.114
Std.Errors	IID	IID	IID
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001			

Comparing means of subpopulations described by dummies

$$\widehat{wage} = \underset{(0.21)}{7.10} - \underset{(0.26)}{2.51}female$$

$n = 526, R^2 = 0.116$

Not holding other factors constant,
women earn \$2.51 less than men; i.e.
the difference between the mean
wages of men and women is \$2.51

- Regression 1: Females have \$2.51 lower wages than males. The reference category is male.
- Regression 2: Males have \$2.51 higher wages than females. The reference category is female.
- Regression 1: The intercept or average wage for females is $4.59 = \hat{\beta}_0 + \hat{\delta}_0 = 7.10 - 2.51$.
- Regression 2: The intercept or average wage for males is $7.10 = \hat{\beta}_0 + \hat{\delta}_0 = 4.59 + 2.51$.
- Regression 3: Both **female** and **male** are included but there is no constant. The coefficients are the **average wage for females and males**.

Example: Effects of training grants on hours of training

Hours training per employee
↓

Dummy variable indicating whether firm received a training grant
↓

$$\widehat{hrsemp} = \underset{(43.41)}{46.67} + \underset{(5.59)}{26.25}grant - \underset{(3.54)}{0.98}sales - \underset{(3.88)}{6.07}\log(employ)$$

$n = 105, R^2 = 0.237$

This is an example of program evaluation

- Treatment group (= grant receivers) vs. control group (= no grant)
- Is the effect of treatment on the outcome of interest causal?
 - Zero conditional mean assumption

Using dummy explanatory variables in equations for log(y)

$$\widehat{\log(\text{price})} = -1.35_{(0.65)} + 0.168_{(0.038)} \log(\text{lotsize}) + 0.707_{(0.093)} \log(\text{sqrft}) \\ + 0.027_{(0.029)} \text{bdrms} + 0.054_{(0.045)} \text{colonial}$$

$n = 88, R^2 = 0.649$

Dummy indicating whether house is of colonial style

$$\frac{\Delta \log(\text{price})}{\Delta \text{colonial}} = \frac{\% \Delta \text{price}}{\% \Delta \text{colonial}} = 5.4\%$$

The house price is 5.4% higher for colonial style houses than the non-colonial houses.

7.3 Dummy Variables for Multiple Categories

- 1) Define membership in each category by a dummy variable
- 2) Leave out one category (which becomes the base category)

$$\begin{aligned}\log(\widehat{wage}) = & 0.321 + 0.213marrmale - 0.198marrfem \quad \leftarrow \text{Holding other things fixed, married women} \\ & \quad (0.100) \quad (0.055) \quad (0.058) \quad \text{earn 19.8\% less than single men (the base} \\ & -0.110lsingfem + 0.079educ + 0.027exper - 0.00054exper^2 \quad \text{category)} \\ & \quad (0.056) \quad (0.007) \quad (0.005) \quad (0.00023) \\ & \quad +0.079tenure - 0.00053tenure^2 \\ & \quad (0.007) \quad (0.00023)\end{aligned}$$

$$n = 2,725, R^2 = 0.0422$$

Interaction terms

Interaction terms for variables *female* and *married* can be done in two different ways.

1) Create four categories: *female and single*, *male and single*, *female and married*, and *male and married* and include 3 of them in the regression model (the fourth/omitted category serves as a base/reference category).

$$wage = \beta_0 + \beta_1 single_{female} + \beta_2 married_{female} + \beta_3 married_{male} + u$$

Reference category: single male

2) Include *female* and *married* and *female*married* in the regression.

$$wage = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 female * married + u$$

Incorporating ordinal information using dummy variables


Example: City credit ratings and municipal bond interest rates

Municipal bond rate Credit rating from 0 to 4 (0=worst, 4=best)


$$MBR = \beta_0 + \beta_1 CR + \text{other factors}$$

This specification would probably not be appropriate as the credit rating only contains ordinal information. A better way to incorporate this information is to define dummies:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + \text{other factors}$$



Dummies indicating whether the particular rating applies, e.g. $CR_1=1$ if $CR=1$, and $CR_1=0$ otherwise. All effects are measured in comparison to the worst rating (= base category).

Interaction terms with indicator variables

```
data(wage1, package='wooldridge')
# Generate indicator variables
wage1 %<>% mutate(single = 1 - married)
wage1 %<>% mutate(male = 1 - female)
# Categories: female*single, male*single, female*married, male*married
wage1 <-wage1 %>%
  mutate(female_single = female*single,
         male_single   = male*single,
         female_married = female*married,
         male_married  = male*married)

wage1 %>%
  select(female, male, single, married, female_single, male_single,
         female_married, male_married) %>%
  head(5)
```


##	female	male	single	married	female_single	male_single	female_married
## 1	1	0	1	0	1	0	0
## 2	1	0	0	1	0	0	1
## 3	0	1	1	0	0	1	0
## 4	0	1	0	1	0	0	0
## 5	0	1	0	1	0	0	0

##	male_married
## 1	0
## 2	0
## 3	0
## 4	1
## 5	1

```
# Regression with male_single as reference category
model_1 <- lm(wage ~ female_single + female_married + male_married, wage1)

# Regression with interaction term
model_2 <- lm(wage ~ female + married + female*married, wage1)

models <- list(model_1, model_2)
modelsummary(models,output = "markdown")
```

	Model 1	Model 2
(Intercept)	5.17***	5.17***
	(0.36)	(0.36)
female_single	-0.56	
	(0.47)	
female_married	-0.60	
	(0.46)	
male_married	2.82***	
	(0.44)	
female		-0.56
		(0.47)
married		2.82***
		(0.44)
female × married		-2.86***
		(0.61)
Num.Obs.	526	526
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

Incorporating ordinal information using dummy variables

Example: City credit ratings and municipal bond interest rates

Municipal bond rate Credit rating from 0 to 4 (0=worst, 4=best)

$$MBR = \beta_0 + \beta_1 CR + \text{other factors}$$

This specification would probably not be appropriate as the credit rating only contains ordinal information. A better way to incorporate this information is to define dummies:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + \text{other factors}$$

Dummies indicating whether the particular rating applies, e.g. $CR_1=1$ if $CR=1$, and $CR_1=0$ otherwise. All effects are measured in comparison to the worst rating (= base category).

Interactions involving dummy variables and continuous variables

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u$$

β_0 = intercept for men

β_1 = slope for men

$\beta_0 + \delta_0$ = intercept for women

$\beta_1 + \delta_1$ = slope for women

Interesting hypothesis

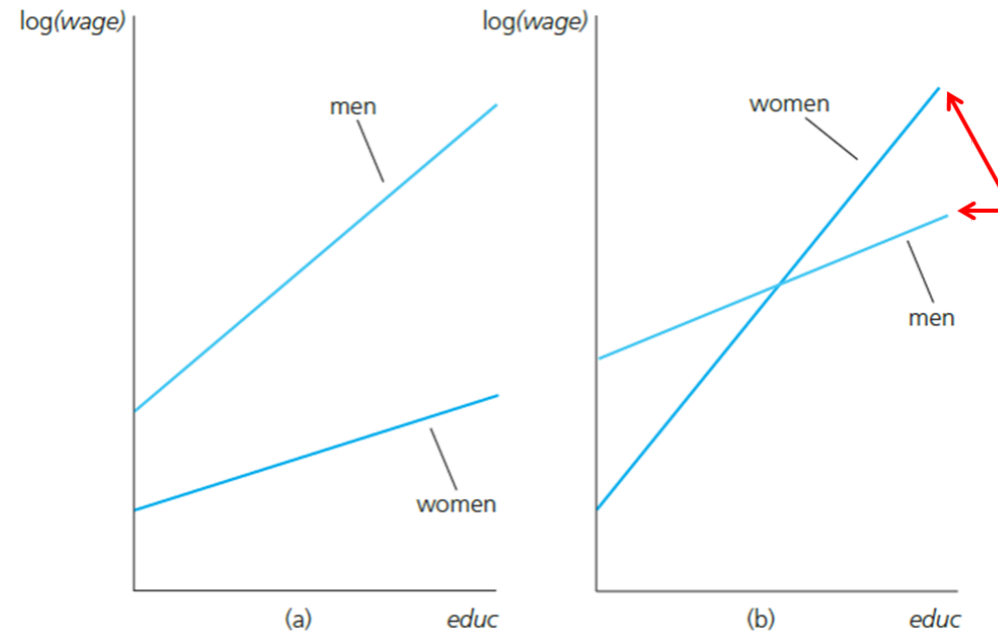
$$H_0: \delta_1 = 0$$

The return to education is the same for men and women

$$H_0: \delta_0 = 0, \delta_1 = 0$$

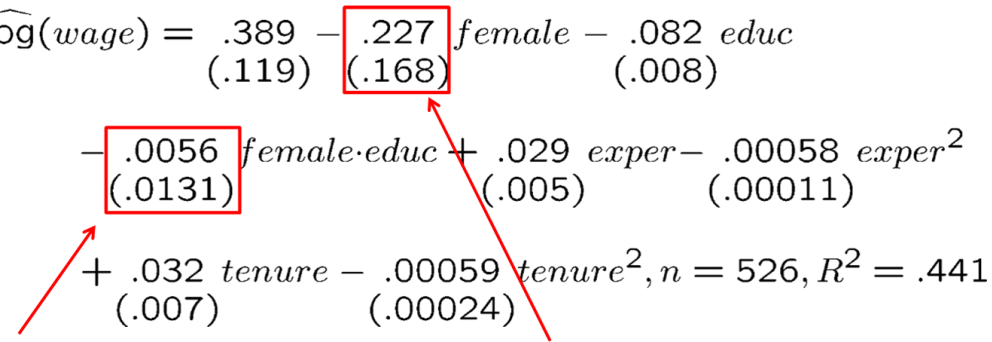
The whole wage equation is the same for men and women

Graphical illustration



Interacting both the intercept and the slope with the female dummy enables one to model completely independent wage equations for men and women.

Example: Estimated wage equation with interaction term

$$\begin{aligned}\widehat{\log(wage)} = & .389 - .227 \text{ female} - .082 \text{ educ} \\ & (.119) \quad (.168) \quad (.008) \\ & - .0056 \text{ female} \cdot \text{educ} + .029 \text{ exper} - .00058 \text{ exper}^2 \\ & (.0131) \quad (.005) \quad (.00011) \\ & + .032 \text{ tenure} - .00059 \text{ tenure}^2, n = 526, R^2 = .441 \\ & (.007) \quad (.00024)\end{aligned}$$


- Coefficient on **femal**e. Does this mean that there is no significant evidence of lower pay for women at the same levels of **educ**, **exper**, and **tenure**?
 - No. This is only the effect for $\text{educ} = 0$. To answer the question one has to recenter the interaction term, e.g. around $\text{educ} = 12.5$ (= average education).
 - Coefficient on the interaction term **femal**e***educ**, provide no evidence against hypothesis that the return to education is the same for men and women.

Testing for Differences in Regression Functions across Groups

Unrestricted model (contains full set of interactions)

$$\begin{aligned} cumgpa = & \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female * sat + \beta_2 hsperc + \delta_2 female * hsperc + \\ & + \beta_3 tothrs + \delta_3 female * tothrs + u \end{aligned}$$

Restricted model (same regression for both groups)

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

Where:

- **sat** stadardized aptitude test score
- **hsperc** high school rank percentile
- **tothrs** total hours spent in college courses

Null hypothesis: All interaction effects are zero i.e. the same regression coefficients apply to men and women

$$H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$$

Estimation of the unrestricted model

$$\begin{aligned} \widehat{cumgpa} = & 1.48 - .353 \text{ female} + .0011 \text{ sat} + .00075 \text{ female} \cdot \text{sat} \\ & (.21) \quad (.411) \quad (.0002) \quad (.00039) \\ & - .0085 \text{ hisperc} - .00055 \text{ female} \cdot \text{hisperc} \\ & (.0014) \quad (.00316) \\ & + .0023 \text{ tothrs} - .00012 \text{ female} \cdot \text{tothrs} \\ & (.0009) \quad (.00163) \end{aligned}$$

$n = 366, R^2 = .406, \overline{R^2} = .394$

Tested individually,
the hypothesis that
the interaction effects
are zero cannot be
rejected

- Estimation of the restricted model (dropping all female terms) gives us the R^2 of .352
- The F- statistic is 8.14 (with a p value of 0.000) which means that we reject the null hypothesis
- Thus, men and women athletes do follow different GPA models

Chow test

Alternative way to compute F-statistic in the given case

- Run separate regressions for men and for women; the unrestricted SSR is given by the sum of the SSR of these two regressions ($SSR_1 + SSR_2$).
- Run regression for the restricted model and store SSR (SSR_p pooled).
- If the test is computed in this way it is called the Chow-Test.
- Important: Test assumes a constant error variance across groups (homoskedasticity).

$$F = \frac{[SSR_p - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k+1)]}{k+1}$$

where k represents the number of interaction terms.

Chow test

Joint test with F-statistic (Chow test)

$$F = \frac{(SSR_P - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(85.515 - 78.355)/4}{78.355/(366-7-1)} \approx 8.18$$

Null hypothesis is rejected.

7.5 A Binary Dependent Variable: the Linear Probability Model

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

$$E(y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

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

$$E(y \mid \mathbf{x}) = 1 \cdot P(y = 1 \mid \mathbf{x}) + 0 \cdot P(y = 0 \mid \mathbf{x})$$

Linear Probability Model (LPM)

$$P(y = 1 \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

$$\beta_j = \Delta P(y = 1 \mid \mathbf{x}) / \Delta x_j$$

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

Example: Labor force participation of married women

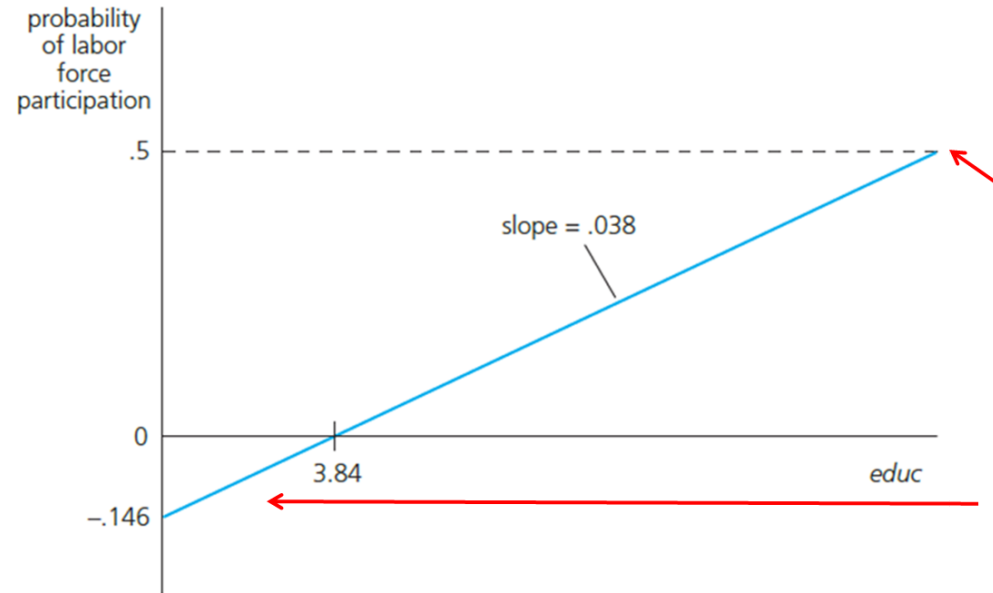
=1 if in labor force, =0 otherwise

Non-wife income (in thousand dollars per year)

$$\begin{aligned}\widehat{inlf} = & .586 - .0034 \text{ nwifeinc} + .038 \text{ educ} + .039 \text{ exper} \\ & (.154) \quad (.0014) \quad (.007) \quad (.006) \\ & - .00060 \text{ exper}^2 - .016 \text{ age} - .262 \text{ kidslt6} \\ & (.00018) \quad (.002) \quad (.034) \\ & + .0130 \text{ kidsge6}, n = 753, R^2 = .264 \\ & (.0132)\end{aligned}$$

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

Graph for $nwifeinc=50$, $exper=5$, $age=30$, $kinds1t6=1$, 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%.

There is a negative predicted probability, but no problem because no woman in the sample has $educ < 5$.

Disadvantages of the linear probability model

- Predicted probabilities may be larger than one or smaller than zero.
- Marginal probability effects sometimes logically impossible.
- The linear probability model is necessarily heteroskedastic.
- Thus, heteroskedasticity consistent standard errors need to be computed.

$$\text{Var}(y \mid \mathbf{x}) = P(y = 1 \mid \mathbf{x})[1 - P(y = 1 \mid \mathbf{x})]$$

Advantages of the linear probability model

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