Intermediate Econometrics

Multiple Regression Analysis with Qualitative Information

Text: Wooldridge, Chapter 7

July 10, 2011

Types of Variables

- As the two main types of variables, we distinguish between
  - *categorial*, and
  - *quantitative*

  variables.

- **Categorial variables** may come as *nominal* and *ordinal* variables.

- **Nominal variables** allow only for qualitative classification, i.e., we can only decide whether an individual in our sample belongs to certain distinct categories, but we cannot rank these categories, i.e., there is no natural ordering. Typical examples are
  - Gender, or
  - Marital Status, or
  - Religion, or
  - Industry of a Firm (Manufacturing, Retail, ...)

- An **ordinal variable** is a categorial variable where the different categories can be ordered in a meaningful sequence, but the "differences" between the categories cannot be interpreted.

- For example, data on education may consist of the highest level of education attained, such as
  - High School
  - Bachelor
  - Master
  - Ph.D.

- A **quantitative variable** is a variable with a natural ordering of the observations and where numerical differences between values have a meaning: The difference in age between a 5–year old an a 10–year old is the same as that between a 10–year old and a 15–year old.

- Until now, we have considered examples of regression analysis using quantitative variables.

- Now we use *binary* or *dummy variables* to include categorial information.

- A **binary** or **dummy** variable
  - takes the value 1 for some observations to indicate the presence of an effect or group membership,
  - and takes the value 0 for the remaining observations.

- For example, in a study of individual wage determination, we may define a variable *female* that takes on the value 1 for females and the value 0 for males.

- Same for a variable *married*.

- We may also define a variable *marrfem* that takes on the value 1 for married females and the value 0 for single females and males (married or not).

- A typical data matrix with binary or dummy variables is shown in Table 1.

Table 1: Partial listing of data used to analyze individual wage determination

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Here,

- *wage* is hourly wage
- *educ* is years of education
- *exper* is years of experience
- *female* and *married* are as indicated above

Models with a Single Dummy Independent Variable

- Consider the case with only a single dummy independent variable, which we just add as an independent variable in the equation.

- For example, consider model

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + u, \tag{1}$$

  where

  - *wage* is hourly wage,
  - $female = 1$ for females and $female = 0$ for males,
  - *educ* is years of education.

- Interpretation of β_1: The difference in expected hourly wage between females and males, given the same amount of education (and holding the factors in u likewise fixed).

- For example, economically, $\beta_1 < 0$ may suggest some form of discrimination against women, since, for the same level of other factors, they earn less than men on average.

- That is,

$$\begin{aligned}\beta_1 &= \mathsf{E}(wage|female = 1, educ) - \mathsf{E}(wage|female = 0, educ) \\ &= \mathsf{E}(wage|female, educ) - \mathsf{E}(wage|male, educ).\end{aligned}$$

- Should we have a dummy variable for males, too?

- No.

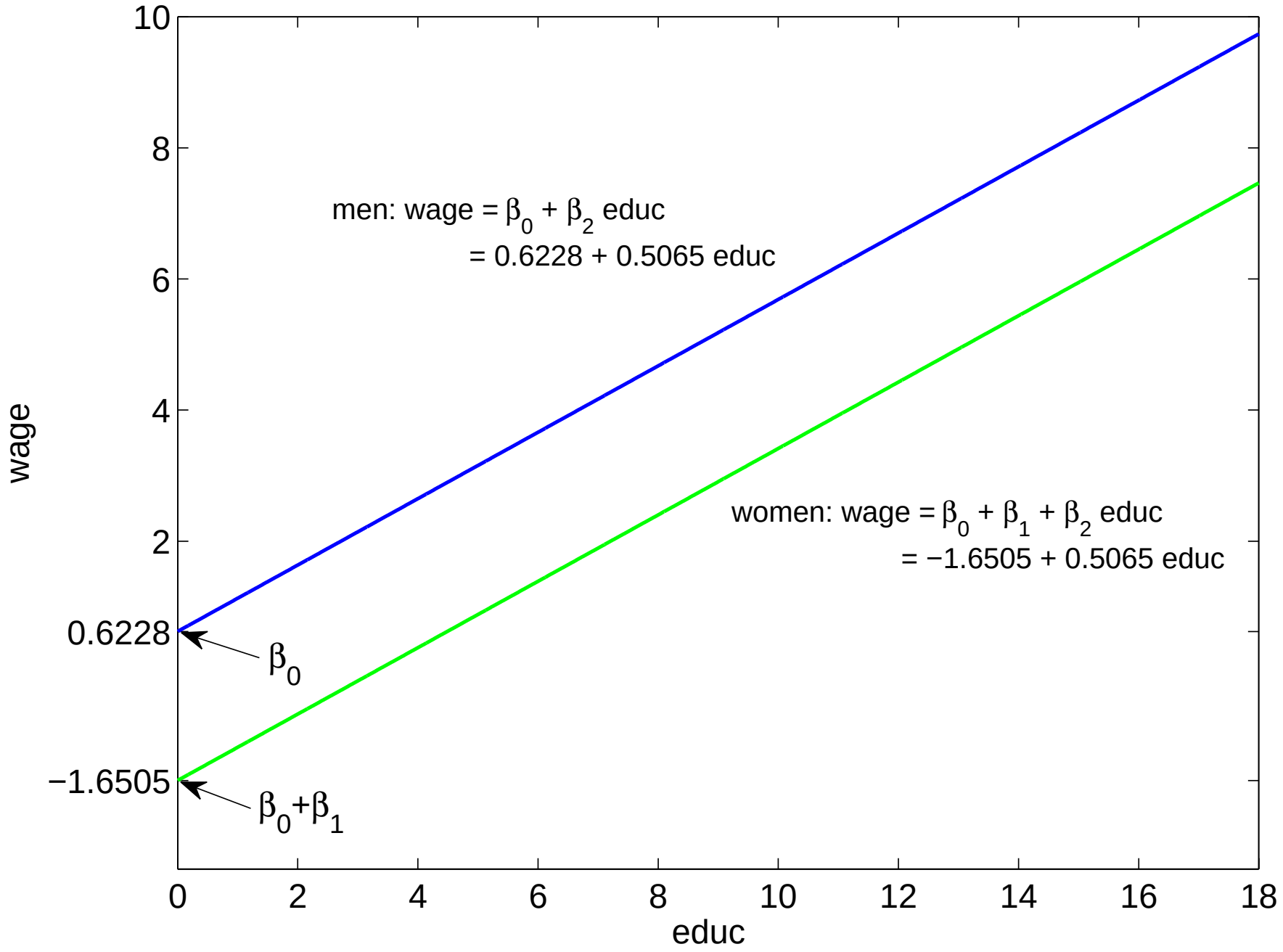
- In model 1, there are gender–specific intercepts of the regression line:
  - The intercept for males is β_0.
  - The intercept for females is $\beta_0 + \beta_1$.

- Estimating Equation (1), we get[1]

$$\begin{aligned}\widehat{wage} &= \underset{(0.6725)}{0.6228} - \underset{(0.2790)}{2.2734}\ female + \underset{(0.0504)}{0.5065}\ educ \qquad (2) \\ n &= 526, \quad R^2 = 0.2588.\end{aligned}$$

[1] The data are from 1976.

- According to this equation, if we take a woman and a man with the same levels of education, the woman earns, on average, $ 2.3 less per hour than the man.


10
8
6
4
2
0.6228
−1.6505
wage
men: wage = β0 + β2 educ
= 0.6228 + 0.5065 educ
women: wage = β0 + β1 + β2 educ
= −1.6505 + 0.5065 educ
β0
β0+β1
0
2
4
6
8
10
12
14
16
18
educ

- How would we test for "wage discrimination"?

- Answer: We use the usual t–statistic.

- Regarding the statistical theory, nothing changes when some of the independent variables are dummy variables.

- For example, in Equation (2), the t–statistic of $\widehat{\beta}_1$ is

$$t_{\widehat{\beta}_1} = \frac{-2.2734}{0.2790} = -8.1470.$$

- In model (1), *males* is the **base group** or **benchmark group**: the group against which comparisons are made.

- If we would have specified

$$wage = \beta_0^\star + \beta_1^\star male + \beta_2 educ + u,$$

then $\widehat{\beta}_0^\star$ would have been the estimated intercept for women and $\widehat{\beta}_0^\star + \widehat{\beta}_1^\star$ would have been the estimated intercept for men, with

$$\widehat{\beta}_0^\star = \widehat{\beta}_0 + \widehat{\beta}_1, \quad \widehat{\beta}_1^\star = -\widehat{\beta}_1.$$

- Clearly we could also drop the overall intercept and estimate the model

$$wage = \delta_1 male + \delta_2 female + \beta_3 educ + u,$$

but then testing for gender–specific effects would require testing for significance of the difference $\delta_1 - \delta_2$, which is more difficult.

- The methodology does not change if further independent variables are added.

- For example, we may extend (1) by controlling for additional variables, e.g.,

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 exper + \beta_4 tenure + u, \quad (3)$$

where

  - *exper* is years of potential experience,
  - *tenure* is years with current employer.

- Upon estimating, we get

$$\begin{aligned} \widehat{wage} &= \underset{(0.7246)}{-1.5679} - \underset{(0.2648)}{1.8109}\, female + \underset{(0.0493)}{0.5715}\, educ \quad (4) \\ &\quad + \underset{(0.0116)}{0.0254}\, exper + \underset{(0.0212)}{0.1410}\, tenure \\ n &= 526, \quad R^2 = 0.3635. \end{aligned}$$

- Now the coefficient on *female* measures the average difference in hourly wage between a woman and a man, given the same levels of education, experience, and tenure.

- The differential of \$ 1.81 is then due to gender or gender–related factors that have not been controlled for.

- Note that, in (4), the magnitude of $\widehat{\beta}_1$ is somewhat smaller than in (2).

- This is due to the fact that in (2), we did not control for *exper* and *tenure*, and these are negatively correlated with *female* (*tenure* in particular), i.e., they are lower on average for women.

- Thus, as we estimate β_3 and β_4 to be positive, we expect $\widehat{\beta}_1$ to be negatively biased in (1).

When the Dependent Variable is in Logarithmic Form

- Equation (3) becomes

$$log(wage) = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 exper + \beta_4 tenure + u, \quad (5)$$

and we get estimates

$$\begin{aligned} \widehat{log(wage)} &= \underset{(0.1019)}{0.5013} - \underset{(0.0372)}{0.3011}\ female + \underset{(0.0069)}{0.0875}\ educ \\ &\quad + \underset{(0.0016)}{0.0046}\ exper + \underset{(0.0030)}{0.0174}\ tenure \\ n &= 526, \quad R^2 = 0.3923. \end{aligned} \quad (6)$$

- Using our earlier approximation, the coefficient of *female* would imply that, for the same levels of *educ*, *exper*, and *tenure*, women earn, on average, 30.1% less than men.

- However, the approximation $100\Delta \log y \approx \%\Delta y$ becomes less accurate as the percentage change increases, and in such situations an exact calculation may be desirable.

- Equation (5) implies

$$\begin{aligned}
\widehat{log(wage}_F) - \widehat{log(wage}_M) &= -0.297 \\
\widehat{\log(wage_F/wage_M)} &= -0.297 \\
\frac{\widehat{wage}_F}{\widehat{wage}_M} - 1 &= \frac{\widehat{wage}_F - \widehat{wage}_M}{\widehat{wage}_M} \\
&= \exp\{\widehat{\beta}_1\} - 1 \\
&= \exp\{-0.297\} - 1 = -0.26
\end{aligned}$$

  so our estimate of the percentage wage differential is

$$100 \times (\exp\{\widehat{\beta}_1\} - 1)\% = 26\%. \tag{7}$$

- If $\widehat{\beta}_1$ is rather small in magnitude, we again get the approximation

$$\exp\{\widehat{\beta}_1\} - 1 = \left(1 + \widehat{\beta}_1 + \frac{\widehat{\beta}_1^2}{2} + \cdots\right) - 1 \approx \widehat{\beta}_1. \tag{8}$$

- But note that $\exp\{\widehat{\beta}_1\} - 1$ is a biased estimator of $\exp\{\beta_1\} - 1$.

- For $x \sim \mathsf{N}(\mu, \sigma^2)$, we have

$$\mathsf{E}(e^x) = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

  and since $\widehat{\beta}_1$ is asymptotically normal, a closer approximation is

$$100 \times \left[\exp\left(\widehat{\beta}_1 - \frac{\widehat{\sigma}^2_{\widehat{\beta}_1}}{2}\right) - 1\right],$$

  which is still not unbiased since $\widehat{\sigma}^2_{\widehat{\beta}_1}$ is estimated.[2]

[2]For detailed analysis and development of an unbiased (and efficient) estimator, see Kees Jan van Garderen and Chandra Shah (2002): Exact interpretation of dummy variables in semilogarithmic equations, *Econometrics Journal*, 5, 149–159.

Using Dummy Variables for Multiple Categories

- We can use several dummy independent variables in the same equation.
- Let us add the dummy variable *married* as well as quadratic terms of *exper* and *tenure* to model (5).
- This gives

$$\begin{aligned} \widehat{log(wage)} &= \underset{(0.0989)}{0.4178} - \underset{(0.0361)}{0.2902}\, female + \underset{(0.0408)}{0.0529}\, married \\ &+ \underset{(0.0068)}{0.0792}\, educ + \underset{(0.0053)}{0.0270}\, exper - \underset{(0.0001)}{0.0005}\, exper^2 \\ &+ \underset{(0.0068)}{0.0313}\, tenure - \underset{(0.0002)}{0.0006}\, tenure^2 \\ n &= 526, \quad R^2 = 0.4426. \end{aligned} \tag{9}$$

- The coefficient on *married* gives the (approximate) proportional differential in wages between those who are and those who are not married, holding gender, education, experience, and tenure fixed.

- Thus, the "marriage premium" appears to be positive, but its t–statistic

$$t_{married} = \frac{0.0529}{0.0408} = 1.2985,$$

  which is not significant.

- However, this may be due to the fact that the marriage premium in this specification is assumed to be the same for women and men.

- This assumption can be relaxed by allowing wage differences among four groups: married women, single women, married men, and single men.

- As a **base group**, single men are chosen.

- This gives rise to the model

$$\begin{aligned}
\widehat{log(wage)} &= \underset{(0.1000)}{0.3214} + \underset{(0.0554)}{0.2127}\, marrmale - \underset{(0.0578)}{0.1983}\, marrfem \\
&\quad - \underset{(0.0557)}{0.1104}\, singfem + \underset{(0.0067)}{0.0789}\, educ \\
&\quad + \underset{(0.0052)}{0.0268}\, exper - \underset{(0.0001)}{0.0005}\, exper^2 \\
&\quad + \underset{(0.0068)}{0.0291}\, tenure - \underset{(0.0002)}{0.0005}\, tenure^2 \\
n &= 526, \quad R^2 = 0.4609.
\end{aligned} \tag{10}$$

- As the base group are single males, married men, for example, are estimated to earn (approximately) 21.3% more than single men, holding education, experience, and tenure fixed.

- A married woman, on the other hand, earns a predicted 19.8% less than a single man with the same level of the other variables.

- If we are interested in the wage differential between single and married women, we would calculate

$$-0.1104 - (-0.1983) = 0.0879. \tag{11}$$

- For testing the significance of this quantity, however, it is easier to reestimate the model with *marrfem* as the base group, which results in

$$\begin{aligned}
\widehat{log(wage)} &= \underset{(0.1058)}{0.1231} + \underset{(0.0458)}{0.4109}\, marrmale + \underset{(0.0578)}{0.1983}\, singmale \\
&\quad + \underset{(0.0523)}{0.0879}\, singfem + \underset{(0.0067)}{0.0789}\, educ \\
&\quad + \underset{(0.0052)}{0.0268}\, exper - \underset{(0.0001)}{0.0005}\, exper^2 \\
&\quad + \underset{(0.0068)}{0.0291}\, tenure - \underset{(0.0002)}{0.0005}\, tenure^2 \\
n &= 526, \quad R^2 = 0.4609.
\end{aligned} \tag{12}$$

- This gives the standard error and t–statistic of (11) as 0.05239 and

$$t_{singfem} = \frac{0.0879}{0.0523} = 1.6795,$$

with a p–value (two–sided and relying on asymptotic normality) of 0.0932, so we would reject at the 10% level the null that there are no differences, on average, between married and single women.

- As in the previous example, if we have different intercepts for g groups or categories, we include $g - 1$ dummy variables in the model plus an intercept.

- Then the overall intercept is the intercept for the base group,

- and the dummy variable coefficient for a particular group represents the estimated difference in intercept between that group and the base group.

- An alternative way to estimate the previous model with multiple categories is by means of using *interactions*.

- For example, we can estimate

$$\begin{aligned}\widehat{\log wage} &= \underset{(0.1000)}{0.3214} - \underset{(0.0557)}{0.1104}\, female + \underset{(0.0554)}{0.2127}\, married \\ &\quad - \underset{(0.0718)}{0.3006}\, female \cdot married + \cdots\end{aligned} \tag{13}$$

- In this specification, the base group is again single men, which is obtained by setting $female = married = 0$.

- The previous parametrization (10) is more convenient when the goal is to test wage differentials between the base group and any other group.

- Equation (13) can be used to directly test whether the gender differential does depend on martial status (or, equivalently, if marriage premium does not depend on gender).

- That is, if the coefficient of $female \cdot married$ is zero in (13), then the difference between married and unmarried women is the same as that between married and unmarried men.

- This also implies that the difference between married men and married women is the same as that between unmarried men and unmarried women.

- For example, in

  $$\log(wage) = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 female \cdot married + \cdots,$$

  the intercepts are

$$\begin{aligned} \beta_{marrfem} &= \beta_0 + \beta_1 + \beta_2 + \beta_3, \\ \beta_{singfem} &= \beta_0 + \beta_1, \\ \beta_{marrmale} &= \beta_0 + \beta_2, \\ \beta_{singmale} &= \beta_0. \end{aligned}$$

- $\beta_3 = 0$ then means $\beta_{marrfem} - \beta_{marrmale} = \beta_1 = \beta_{singfem} - \beta_{singmale}$.

- The estimate in (13) shows, however, that this hypothesis can be rejected.

Incorporating Ordinal Information by Using Dummy Variables

- Suppose we want to estimate the effect of education on wage.

- We just observe the highest level of education attained instead of years of education,[3] namely,
  - High School (HS)
  - Bachelor (B)
  - Master (M)
  - Ph.D. (P).

- We could define a variable *educ* that is 0 for the first group (HS), 1 for the second group (B), 2 for the third group (M), and 3 for the fourth group (P), i.e.,

$$wage = \beta_0 + \beta_1 educ + \text{other variables} + u. \tag{14}$$

[3] One could argue that the highest level of education attained is more appropriate than using years of education.

- However, this is unsatisfactory since it assumes that the (expected) increase in wage at each "threshold" is the same.

- That is, β_1 in (14) measures the (expected) difference between a person with a Ph.D. and a Master and between a person with a Bachelor and a High School degree.

- A more flexible model would be to use binary (dummy) variables for each level at education.

- Thus, we would write

$$wage = \beta_0 + \beta_B B + \beta_M M + \beta_P P + \cdots + u,$$

and thus have

$$\begin{aligned}
\mathsf{E}(wage|HS, \ldots) &= \beta_0 + \cdots \\
\mathsf{E}(wage|B, \ldots) &= \beta_0 + \beta_B + \cdots \\
\mathsf{E}(wage|M, \ldots) &= \beta_0 + \beta_M + \cdots \\
\mathsf{E}(wage|P, \ldots) &= \beta_0 + \beta_P + \cdots
\end{aligned}$$

Allowing for Different Slopes

- The previous examples showed how to use dummy variables to allow for different intercepts for any number of groups in a multiple regression model.

- We can also allow for **differences in slopes** by considering interactions of dummy variables with other explanatory variables.

- For example, in the wage example, we may wish to test whether the return to education is gender–specific.

- If there is only gender and education in the model, we may write

$$\log(wage) = \beta_0 + \beta_1 female + (\beta_2 + \beta_3 female)educ + u. \qquad (15)$$

- The intercept for males is β_0, and the slope on education for males is β_2.

- For females, the numbers are $\beta_0 + \beta_1$ and $\beta_2 + \beta_3$, respectively.

- Testing whether the return to education is the same for men and women amounts to testing

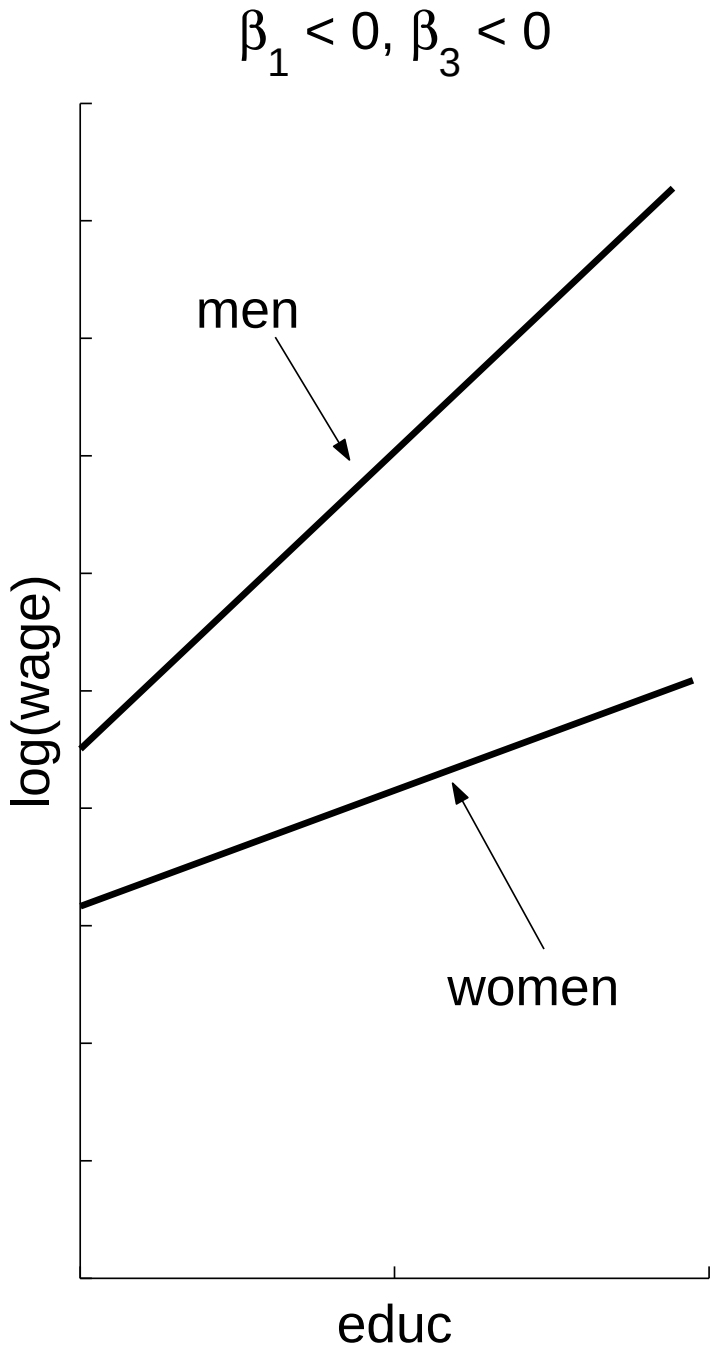
$$H_0 : \beta_3 = 0$$

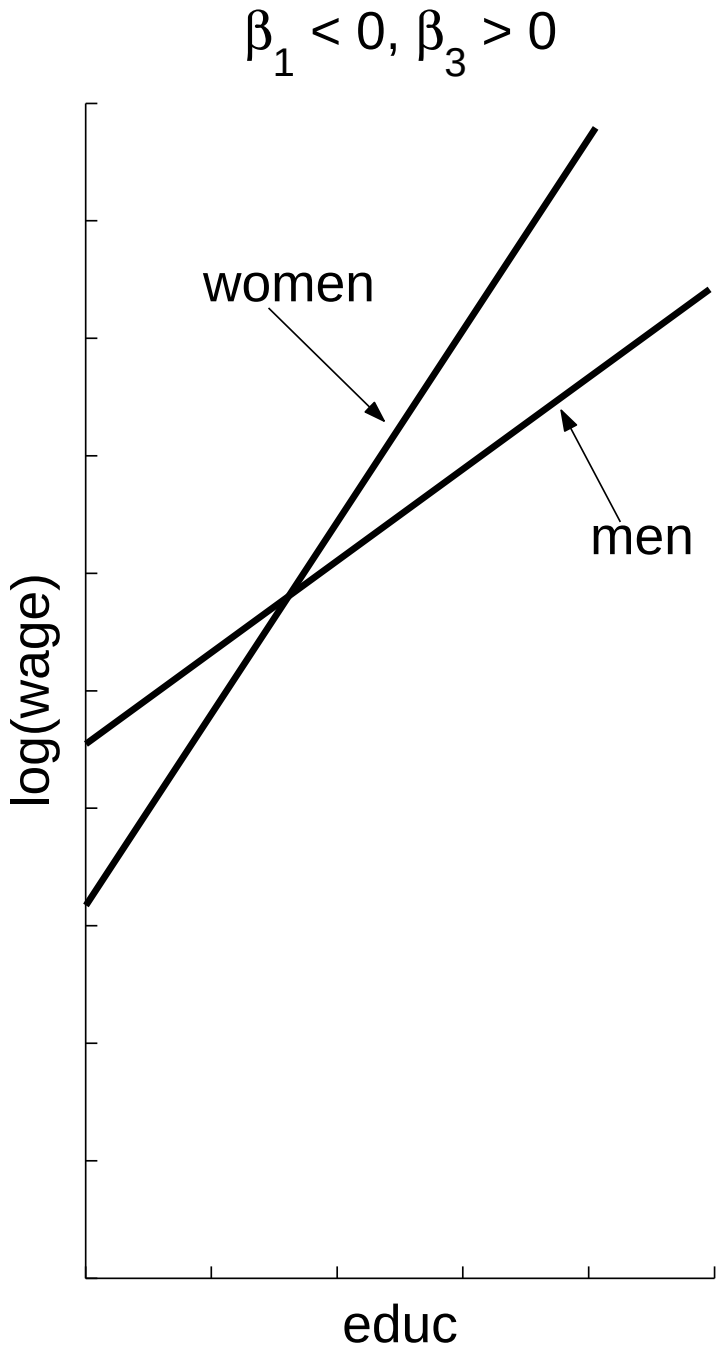
in (15).

- Likewise, testing whether expected wages are the same for men and women with the same levels of education requires testing

$$H_0 : \beta_1 = \beta_3 = 0,$$

using an F test.


$\beta_1 < 0, \beta_3 < 0$
men
women
log(wage)
educ


$\beta_1 < 0, \beta_3 > 0$
women
men
log(wage)
educ

Testing Differences in Regression Functions

- Suppose we want to test whether two populations have the same regression function.

- In the wage example, this amounts to testing

$$H_0 : \beta_1 = \beta_3 = \beta_5 = \beta_7 = \beta_9 = \beta_{11} = 0 \tag{16}$$

in

$$\begin{aligned} \log(wage) \quad = \quad & \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 educ \cdot female + \beta_4 exper \\ & + \beta_5 exper \cdot female + \beta_6 exper^2 + \beta_7 exper^2 \cdot female \\ & + \beta_8 tenure + \beta_9 tenure \cdot female + \beta_{10} tenure^2 \\ & + \beta_{11} tenure^2 \cdot female. \end{aligned} \tag{17}$$

- In general, suppose we have two groups, $g = 1$ and $g = 2$.

- Write the model as

$$y = \beta_{0,g} + \beta_{1,g}x_1 + \cdots + \beta_{k,g}x_k + u, \quad g = 1, 2.$$

- The hypothesis

$$H_0 : \beta_{j,1} = \beta_{j,2}, \quad j = 0, \ldots, k,$$

imposes $k + 1$ restrictions on the $2(k + 1)$ parameters of the full model allowing the intercept and all the slopes to be different.

- The F–statistic (**Chow statistic**) is therefore

$$F = \frac{\mathsf{SSR}_r - \mathsf{SSR}_{ur}}{\mathsf{SSR}_{ur}} \times \frac{n - 2(k + 1)}{k + 1} = \frac{\mathsf{SSR}_r - (\mathsf{SSR}_1 + \mathsf{SSR}_2)}{\mathsf{SSR}_1 + \mathsf{SSR}_2} \times \frac{n - 2(k + 1)}{k + 1},$$

where SSR_1 and SSR_2 are the residual sums of squares from group–wise regressions.

- E.g., in the wage example, estimating equation

$$\log(wage) \quad = \quad \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2$$

separately for women ($g = 1$) and mean ($g = 2$).

Table 2: Dependent variable: $\log(wage)$, $n = 526$; this has the same effects for women and men

variable	coefficient	std. error	t–stat.	p–value
$const$	0.2016	0.1015	1.9865	0.0470
$educ$	0.0845	0.0072	11.8030	0.0000
$exper$	0.0293	0.0053	5.5405	0.0000
$exper^2$	−0.0006	0.0001	−5.1890	0.0000
$tenure$	0.0371	0.0072	5.1251	0.0000
$tenure^2$	−0.0006	0.0002	−2.4677	0.0136
$R^2 = 0.3669$, SSR $= \sum_i \widehat{u}_i^2 = 93.9113$				

Table 3: Dependent variable: $\log(wage)$, $n = 252$; estimates for **women**

variable	coefficient	std. error	t–stat.	p–value
$const$	0.3230	0.1399	2.3083	0.0210
$educ$	0.0726	0.0104	7.0018	0.0000
$exper$	0.0173	0.0067	2.5812	0.0098
$exper^2$	−0.0004	0.0001	−2.5293	0.0114
$tenure$	0.0392	0.0117	3.3492	0.0008
$tenure^2$	−0.0014	0.0005	−2.8013	0.0051
$R^2 = 0.2596$, SSR $= 36.6751$				

Table 4: Dependent variable: $\log(wage)$, $n = 274$; estimates for **men**

variable	coefficient	std. error	t–stat.	p–value
$const$	0.2148	0.1319	1.6286	0.1034
$educ$	0.0868	0.0089	9.6998	0.0000
$exper$	0.0404	0.0072	5.6027	0.0000
$exper^2$	−0.0008	0.0002	−4.8640	0.0000
$tenure$	0.0325	0.0090	3.6278	0.0003
$tenure^2$	−0.0006	0.0003	−1.9448	0.0518
$R^2 = 0.4462$, SSR $= 43.2453$				

Table 5: Dependent variable: $\log(wage)$, $n = 526$; gender–specific intercept and slopes

variable	coefficient	std. error	t–stat.	p–value
$const$	0.2148	0.1295	1.6591	0.0971
$female$	0.1082	0.1928	0.5612	0.5746
$educ$	0.0868	0.0088	9.8814	0.0000
$female \cdot educ$	−0.0142	0.0138	−1.0337	0.3013
$exper$	0.0404	0.0071	5.7076	0.0000
$female \cdot exper$	−0.0231	0.0099	−2.3402	0.0193
$exper^2$	−0.0008	0.0002	−4.9551	0.0000
$female \cdot exper^2$	0.0004	0.0002	1.7939	0.0728
$tenure$	0.0325	0.0088	3.6957	0.0002
$female \cdot tenure$	0.0067	0.0148	0.4514	0.6517
$tenure^2$	−0.0006	0.0003	−1.9812	0.0476
$female \cdot tenure^2$	−0.0008	0.0006	−1.4428	0.1491
$R^2 = 0.4612$, SSR $= 79.9204$				

- Note that the parameter estimates are the same in Tables 3 and 4 and in Table 5.

- The standard errors differ somewhat due to the homoskedasticity assumption implicitly imposed in the joint model in Table 5.

- In our example, the F–statistic is therefore

$$\begin{aligned} F &= \frac{93.9113 - (43.2453 + 36.6751)}{43.2453 + 36.6751} \times \frac{526 - 2(5+1)}{5+1} \\ &= \frac{93.9113 - 79.9204}{79.9204} \times \frac{514}{6} \\ &= 14.9968, \end{aligned}$$

  which is significant at the 5% level (and indeed on any reasonable level).

- However, instead of allowing for no differences between the groups under H_0, it may be more interesting to allow for a group–specific intercept and then test for group–specific slopes.

- If we do so, the sum of squared residuals under the null is reduced to SSR $= 82.9506$, so the F statistic

$$\begin{aligned} F &= \frac{82.9506 - (43.2453 + 36.6751)}{43.2453 + 36.6751} \times \frac{526 - 2 \times 6}{5} \\ &= \frac{82.9506 - 79.9204}{79.9204} \times \frac{514}{5} \\ &= 3.8977. \end{aligned}$$

Table 6: 95% Quantiles of the F distribution (= 5% critical values) (ν_1 numerator degrees of freedom; ν_2 denominator degrees of freedom)

ν_2/ν_1	2	3	4	5	6	7	8	9	10
10	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782
15	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437
20	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479
25	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365
30	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646
35	3.2674	2.8742	2.6415	2.4851	2.3718	2.2852	2.2167	2.1608	2.1143
40	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772
45	3.2043	2.8115	2.5787	2.4221	2.3083	2.2212	2.1521	2.0958	2.0487
50	3.1826	2.7900	2.5572	2.4004	2.2864	2.1992	2.1299	2.0734	2.0261
60	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926
70	3.1277	2.7355	2.5027	2.3456	2.2312	2.1435	2.0737	2.0166	1.9689
80	3.1108	2.7188	2.4859	2.3287	2.2142	2.1263	2.0564	1.9991	1.9512
90	3.0977	2.7058	2.4729	2.3157	2.2011	2.1131	2.0430	1.9856	1.9376
100	3.0873	2.6955	2.4626	2.3053	2.1906	2.1025	2.0323	1.9748	1.9267
∞	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307

The Linear Probability Model (LPM)

- Now suppose that the *dependent variable* y is a binary variable that takes on only the two values zero and one.

- How can we then interpret the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u? \tag{18}$$

- If y is either zero or one, then, if x_j changes, y either does not change or changes from zero to one or vice versa, and β_j cannot be interpreted as the change in y given a unit increase in x_j.

- To give a meaning to (18), we may appeal to assumption

$$\mathsf{E}(u|\underbrace{x_1, \ldots, x_k}_{=\boldsymbol{x}}) = 0, \tag{19}$$

so that

$$\mathsf{E}(y|\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j. \tag{20}$$

- If y takes on only the values zero or one, then

$$p(\boldsymbol{x}) := \mathsf{P}(y = 1|\boldsymbol{x}) = \mathsf{E}(y|\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j, \tag{21}$$

  which says that the conditional probability of $y = 1$ (**response probability**) is a linear function of the independent variables $x_1, \ldots, x_k$.

- In the **linear probability model** (LPM), β_j measures the change in the probability of $y = 1$ when x_j changes, holding other factors fixed,

$$\Delta\mathsf{P}(y = 1|\boldsymbol{x}) = \beta_j \Delta x_j. \tag{22}$$

- Drawback 1: We may get probabilities larger than unity or smaller than zero.

- For example, consider a model for the probability of labor market participation of married women (1975), measured by variable $inlf$, which is one if the woman works for a wage outside the home.[4]

- The fitted model is

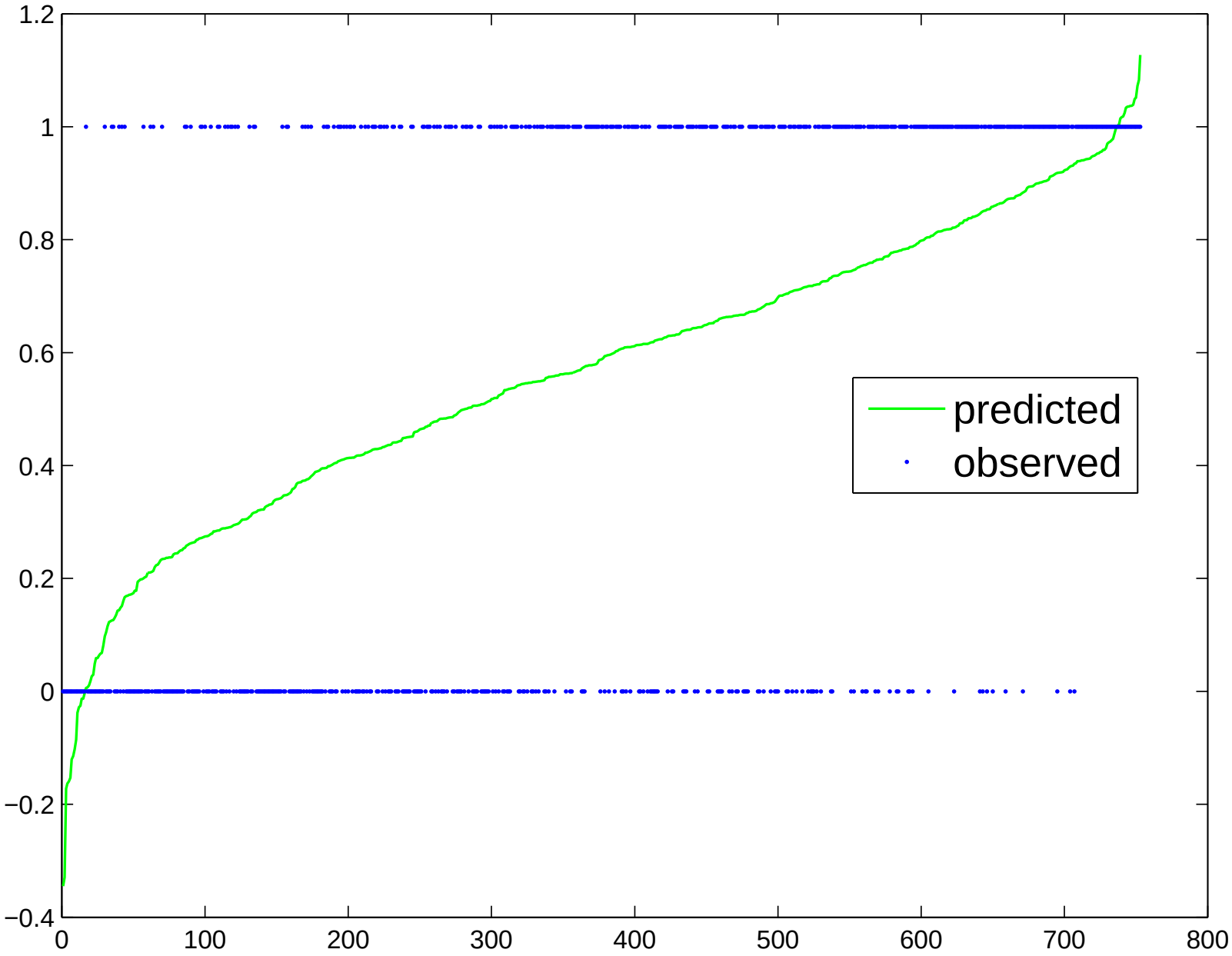
$$\begin{aligned}\widehat{inlf} &= \underset{(0.1542)}{0.5855} - \underset{(0.0014)}{0.0034}\, nwifeinc + \underset{(0.0074)}{0.0380}\, educ + \underset{(0.0057)}{0.0395}\, exper \\ &\quad - \underset{(0.0002)}{0.0006}\, exper^2 - \underset{(0.0025)}{0.0161}\, age - \underset{(0.0335)}{0.2618}\, kidslt6 + \underset{(0.0132)}{0.0130}\, kidsge6 \\ n &= 753,\end{aligned}$$

  where

  – *nwifeinc* is a measure of other sources of income (e.g., husband's earnings)
  – *kidslt6* is number of kids under 6 years
  – *kidsge6* is number of kids 6–18.

[4] Cf. Wooldridge p. 247.

- For example, we would interpret this equation in the sense that, everything else being fixed, another 5 years of education increase the probability of being in the labor force (i.e., $inlf = 1$) by $5 \times 0.038 = 0.19$.


predicted "probabilities" and observed values of inlf
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
0
100
200
300
400
500
600
700
800
predicted
observed

- We observe 16 (17) cases where the implied probability is actually less than zero (greater than unity), with minimum and maximum given by –0.3451 and 1.1272, respectively.

- Drawback 2: There is heteroskedasticity, since

$$\mathsf{Var}(y|\boldsymbol{x}) = \mathsf{E}(u^2|\boldsymbol{x}) = p(\boldsymbol{x})(1 - p(\boldsymbol{x})).$$

- With probability $p(\boldsymbol{x})$, y will be one and u will be $1 - p(\boldsymbol{x})$.

- With probability $1 - p(\boldsymbol{x})$, y will be zero and u will be $-p(\boldsymbol{x})$.

- Thus,

$$\begin{aligned}
\mathsf{E}(u^2|\boldsymbol{x}) &= p(\boldsymbol{x})(1 - p(\boldsymbol{x}))^2 + (1 - p(\boldsymbol{x}))p(\boldsymbol{x})^2 \\
&= p(\boldsymbol{x})(1 - p(\boldsymbol{x}))[1 - p(\boldsymbol{x}) + p(\boldsymbol{x})] \\
&= p(\boldsymbol{x})(1 - p(\boldsymbol{x})).
\end{aligned}$$