Task in econometrics

7. Models with qualitative variables

7.1 For 6 consumers the demand for protein shakes (y), income (x) and age group (a) were surveyed:

Consumers	1	2	3	4	5	6
Demand for protein shakes	8	12	6	15	11	9
income	10	18	8	22	14	12
ager	young	old	young	old	young	young

a) Estimate the demand function for protein shakes considering the age effect,

$$y_t = e^{\beta_1} \cdot x_t^{\beta_2} \cdot e^{\beta_3 \cdot a_t} \cdot e^{u_t}$$

with

$$a_t = \begin{cases} 0, & \text{falls Konsument jung} \\ 1, & \text{falls Konsument alt} \end{cases},$$

in linearized form with the ordinary least squares method (OLS) using the program R! Give the linearized demand function used for this purpose and the corresponding regression output!

b) Specify the linearized OLS estimated demand function for protein shakes and interpret the income effect!!

c) What are the linearized age-specific demand functions for protein shakes? Interpret the age effect on the demand for protein shakes!

solution:

- linearized demand function:

$$\ln(y_t) = \beta_1 + \beta_2 \cdot \ln(x_t) + \beta_3 \cdot a_t + u_t$$

- regression output

```
> demand1 = lm(log(y)~log(x)+a)
> summary(demand1)

Call:
lm(formula = log(y) ~ log(x) + a)
Residuals:
      1          2          3          4          5          6
 0.040459 -0.006234 -0.012954  0.006234  0.005665 -0.033170
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.37841	0.17154	-2.206	0.114540	
log(x)	1.04986	0.07187	14.607	0.000696	***
a	-0.16493	0.05199	-3.173	0.050377	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0317 on 3 degrees of freedom
Multiple R-squared: 0.9943, Adjusted R-squared: 0.9904
F-statistic: 259.5 on 2 and 3 DF, p-value: 0.0004358

b) Specify the linearized OLS estimated demand function for protein shakes and interpret the income effect!

solution:

- linearized OLS-estimated demand function:

$$\hat{\ln(y_t)} = -0{,}37841 + 1{,}04986 \cdot \ln(x_t) - 0{,}16493 \cdot a_t$$

estimated regression coefficient of income (x):

$$\hat{\beta}_2 = 1{,}04986 \quad \text{(Income elasticity)}$$

Income elasticity > 1: elastic demand for protein shakes
If income rises by 1%, the demand for
protein shakes on average by about 1.05%.

The influence of income on the demand for protein shakes is
the 1% significance level is statistically confirmed (two-sided test):

$|\ t = 14{,}406\ | > t_{3;0.995} = 5{,}841 \Rightarrow H_0$ reject (two-sided test)
or
$p = 0{,}000696 < \alpha = 0{,}01 \Rightarrow H_0$ reject (two-sided test)

c) What are the linearized age-specific demand functions for protein shakes? Interpret the age effect on the demand for protein shakes!

- linearized age-specific demand function for young consumers:

$$\hat{\ln(y_t)} = -0{,}37841 + 1{,}04986 \cdot \ln(x_t)$$

- linearized age-specific demand function for older consumers:

$$\widehat{\ln(y_t)} = -0.543341 + 1{,}04986 \cdot \ln(x_t)$$

t
- age effect

At the transition from the young age group to the older age group, the demand for protein shakes decreases by 0.16493.

The age effect on the demand for protein shakes is statistically secured at the 10% significance level (weakly significant):

$|\ t = -3.173\ | > t_{3;0.95} = 2{,}353 \Rightarrow H_0$ reject (two-sided test)
or
$p = 0.050377 < \alpha = 0{,}10 \Rightarrow H_0$ reject (two-sided test)

7.2 Data are given for 6 consumers on the demand for protein shakes (y), income (x) and age group (a) (see 7.1).

a) How must the multiplicative demand function for protein shakes given in 7.1a) be modified to reflect the age-dependent influence on income elasticity?

solution:

Demand function for protein shakes with age-dependent influence on income elasticity:

$$y_t = e^{\beta_1} \cdot x_t^{\beta_2} \cdot e^{\beta_3 \cdot \ln(x_t) \cdot a_t} \cdot e^{u_t}$$

b) Specify the linearized demand function for protein shakes, taking into account the interaction between income and age, and estimate it using the ordinary least squares method (OLS method)! Give the determined regression output of the program R!

solution:

- linearized demand function::

$$\ln(y_t) = \beta_1 + \beta_2 \cdot \ln(x_t) + \beta_3 \cdot \ln(x_t) \cdot a_t + u_t$$

- regression output

```
> demand2 = lm(log(y)~log(x)+log(x):a)
> summary(demand2)

Call:
lm(formula = log(y) ~ log(x) + log(x):a)
Residuals:
```

```
      1          2          3          4          5          6
 0.040897 -0.011868 -0.011773  0.011097  0.004984 -0.033338
Coefficients:
              Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)   -0.38650     0.18041    -2.142    0.121578
log(x)         1.05318     0.07564    13.923    0.000802 ***
log(x):a      -0.05564     0.01828    -3.044    0.055693 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03272 on 3 degrees of freedom
Multiple R-squared:  0.9939,    Adjusted R-squared:  0.9898
F-statistic: 243.5 on 2 and 3 DF,  p-value: 0.0004791
```

c) What is the numerically specified demand function for protein shakes! Interpret the income and age effect on the demand for protein shakes by specifying the age-specific demand functions!

solution:

- linearized OLS-estimated demand function:

$$\hat{\ln(y_t)} = -0{,}38650 + 1{,}05318 \cdot \ln(x_t) - 0{,}05564 \cdot \ln(x_t) \cdot a_t$$

estimated coefficient of regression of income (x):

$$\hat{\beta}_2 = 1{,}05318$$ (Income elasticity of the reference group)

estimated regression coefficient of the interaction effect between income (x) and age (a):

$$\hat{\beta}_3 = -0{,}05564$$ (Reduction of income elasticity at transition from the young age group to the older age group)

- age-specific demand function (age group young)

$$\hat{\ln(y_t)} = -0{,}38650 + 1{,}05318 \cdot \ln(x_t)$$

Income elasticity > 1: elastic demand for protein shakes
If income rises by 1%, the demand for
protein shakes on average by about 1.05%.

- age-specific demand function (age group old)

$$\hat{\ln(y_t)} = -0{,}38650 + 0.99754 \cdot \ln(x_t)$$

Income elasticity < 1: inelastic demand for protein shakes
if income rises by 1%, the demand for
Protein shakes at 0.9975% on average by a little less than 1%.

The influence of income on the demand for protein shakes is based on 1% significance level statistically secured:

$| t = 13{,}923 | > t_{3;0.995} = 5{,}841 \Rightarrow H_0$ reject (two-sided test)
or
$p = 0{,}000802 < \alpha = 0{,}01 \Rightarrow H_0$ reject (two-sided test)

The influence of the interaction effect between income and age on Demand for protein shakes is statistically at the 10% significance level secured (weakly significant):

$| t = -3{,}044| > t_{3;0.95} = 2{,}353 \Rightarrow H_0$ reject (two-sided test)
or
$p = 0.055693 < \alpha = 0{,}10 \Rightarrow H_0$ reject (two-sided test)

7.3 The production y of an industrial sector in a quarter can be traced back to the order intake x delayed by one period and to additive seasonal influences. Give the corresponding design of the observation matrix X for a period of three years!

solution:

• New orders received in the quarters of the analysis period: x1, x2, ..., xn-1, xn

Incoming orders before the start of the analysis period: x0: x0

• The seasonal variable is recorded for 4 quarters by 3 seasonal dummies

If, for example, the 1st quarter is selected as the reference quarter, seasonal dummies are calculated for the quarters

2, 3 and 4: $s_j = \begin{cases} 1, \text{falls Quartal } j \\ 0, \text{sonst} \end{cases}, \quad j=2,3,4$

Design of the observation matrix **X**: $\underset{12x5}{\mathbf{X}} = \begin{bmatrix} 1 & x_0 & 0 & 0 & 0 \\ 1 & x_1 & 1 & 0 & 0 \\ 1 & x_2 & 0 & 1 & 0 \\ 1 & x_3 & 0 & 0 & 1 \\ 1 & x_4 & 0 & 0 & 0 \\ 1 & x_5 & 1 & 0 & 0 \\ 1 & x_6 & 0 & 1 & 0 \\ 1 & x_7 & 0 & 0 & 1 \\ 1 & x_8 & 0 & 0 & 0 \\ 1 & x_9 & 1 & 0 & 0 \\ 1 & x_{10} & 0 & 1 & 0 \\ 1 & x_{11} & 0 & 0 & 1 \end{bmatrix}$

7.4 For the Keynesian consumption function, EViews produced the following estimated results for the total period 1974 -1992,

Dependent Variable: KONSUMR
Method: Least Squares Sample: 1974 1992

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	37.31676	44.21441	0.843995	0.4104
VEINKR	0.864223	0.037617	22.97439	0.0000

R-squared	0.968797	S.D. dependent var	128.5075
Adjusted R-squared	0.966962	S.E. of regression	23.35807

before

Dependent Variable: KONSUMR
Method: Least Squares Sample: 1974 1979

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-118.7550	56.85496	-2.088736	0.1050
VEINKR	0.997260	0.054577	18.27246	0.0001

R-squared	0.988162	S.D. dependent var	63.73863
Adjusted R-squared	0.985202	S.E. of regression	7.753625

and after the oil price shock 1980:

Dependent Variable: KONSUMR
Method: Least Squares Sample: 1980 1992

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	167.1905	47.73910	3.502170	0.0050
VEINKR	0.764885	0.038743	19.74249	0.0000

R-squared	0.972553	S.D. dependent var	106.1633
Adjusted R-squared	0.970057	S.E. of regression	18.37044

Use the Chow test to check whether the 1980 oil price shock could have caused a structural break in the Keynesian consumption function ($\alpha = 0.01$)!

solution:

• Null hypothesis of the Chow test:

$$H_0 : \beta_j^1 = \beta_j^2 \quad \text{für alle } j = 1,2$$

Under the null hypothesis there is no structural break, so that autonomous consumption β_1 and the marginal consumption ratio β_2 are equal before and after the oil price shock.

• Calculation of the residual square sum

- Residual square sum Q for the total observation period 1974 – 1992 (n=19, k=2):

$$Q = SSR = (n-k)\cdot SER^2 = (19-2)\cdot 23{,}35807^2 = 9275{,}2$$

- Residual square sum Q_1 for subperiod 1974 – 1979 (n=6, k=2):

$$Q_1 = SSR_1 = (n-k)\cdot SER_1^2 = (6-2)\cdot 7{,}753625^2 = 240{,}5$$

- Residual square sum Q_2 for subperiod 1980 – 1992 (n=13, k=2):

$$Q_2 = SSR_2 = (n-k)\cdot SER_2^2 = (13-2)\cdot 18{,}37044^2 = 3712{,}2$$

• Test statistic of the Chow test (n=19, k=2):

$$F = \frac{[Q-(Q_1+Q_2)]/k}{(Q_1+Q_2)/(n-2k)} = \frac{[9275{,}2-(240{,}5+3712{,}2)]/2}{(240{,}5+3712{,}2)/(19-2\cdot 2)}$$

$$= \frac{(9275{,}2-3952{,}7)/2}{3952{,}7/15} = \frac{2661{,}25}{263{,}513} = 10{,}099$$

• critical value ($\alpha = 0{,}01$):

$F_{2;15;0{,}99} = 6{,}36$

• test decision:

$(F = 10{,}099) > (F_{2;15;0{,}99} = 6{,}36)$ → H_0 reject

After the oil price shock, autonomous consumption β_1 and the marginal consumption ratio β_2 can no longer be regarded as equal, so that a structural break in the Keynesian consumption function must be concluded.

7.5 Interpret the EViews result of an OLS estimate of the Keynesian consumption function using the dummy variables D1 (1 for 1974 to 1979), 0 else and D2 (1 for 1980 to 1992, 0 else):

Dependent Variable: KONSUMR
Method: Least Squares Sample: 1974 1992

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D1	-118.7550	119.0320	-0.997673	0.3343
D2	167.1905	42.18468	3.963298	0.0012
VEINKR*D1	0.997260	0.114263	8.727736	0.0000
VEINKR*D2	0.764885	0.034235	22.34197	0.0000

solution:

- Autonomous consumption

The regression coefficients of the two dummy variables D_1 and D_2 indicate autonomous consumption before and after the 1980 oil price shock. In purely arithmetical terms, autonomous consumption after the oil price shock is 285.9455 GE [= 167.1905 - (-118.7550) GE] higher than after the oil price shock. However, it should be noted that the regression coefficient of the dummy variable D_1 does not deviate significantly from 0 (p = 0.3343). With a "true" absolute term before the oil price shock of 0, autonomous consumption after the oil price shock increases by 167.1905 Ge.

- Marginal consumption

The marginal consumption rates are given in the two subperiods by the regression coefficients of the interaction terms VEINR*D1 and VEINKR*D2. This shows exactly the opposite tendency to autonomous consumption. The marginal consumption rate in the sub-period 1974 - 1979, at a value of 0.997, is higher than in the second sub-period 1980 - 1992, when it is only 0.765. After the oil price shock, it thus falls by 23.2 percentage points [=(0.997 - 0.765)100%].

7.6 What objections can be raised to the linear probability model for qualitative dependent variables from an econometric perspective?

solution:

In the linear probability model, the expected values of the dependent qualitative variables y can be interpreted as probabilities at given values of the regression coefficients β_j and exogenous variables x_{jt}. As estimators of these probabilities, the regression values $\hat{y}_t$ would have to lie between 0 and 1 for the model to be interpreted correctly. However, this condition can be violated in a Kleist square estimate. The estimated probabilities can become negative or greater than one for individual cross-sectional units (f+r) (e.g. consumers, companies, employees), so that the estimation results of the linear probability model can no longer be interpreted meaningfully.

In the multiple linear regression model it is assumed by default that the disturbance variable is homoscedastic, i.e. has the same variance for all units. This assumption is always violated in the linear probability model, since the disturbance variance depends on the values of the

exogenous variables. It is therefore usually different for all cross-sectional units. In the linear probability model, the conditional expected value sets from the outset inadequate restrictions with regard to the disturbance variable.

7.7 X_i, i=1,2,…,n is a Bernoulli-distributed random variable,which takes the value 1 with probability p and the value 0 with probability 1-p. For the given random sample (x_1, x_2, …, x_n) from a Bernoulli distributed population,determine the maximum likelihood estimator (ML-estimator) for the probability p!

solution:

Likelihood function:

$$L(p \mid p, n) = \binom{n}{x} \cdot p^{x} \cdot (1-p)^{n-x}$$

Necessary condition for a maximum:

$$\frac{dL}{dp} = \binom{n}{x} \cdot x \cdot p^{x-1} \cdot (1-p)^{n-x} + \binom{n}{x} \cdot p^{x} \cdot (n-x) \cdot (1-p)^{n-x-1} \cdot (-1) = 0$$

$$x \cdot (1-p) - (n-x) \cdot p = 0$$

$$x - n \cdot p = 0 \Leftrightarrow n \cdot p = x$$

$$\hat{p}_{ML} = \frac{x}{n}$$ (Stichprobenanteilswert)

7.8 The choice of a means of transport V (public transport or private car) for the way to work has been estimated with EViews on the basis of the Logit model as a function of the distance E (in km) and the occupational status S using the maximum likelihood method (data see appendix):

Dependent Variable: VERKEHRSART
Method: ML - Binary Logit (Newton-Raphson)
Sample: 1 24 Included observations: 24
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-5.005772	2.122852	-2.358041	0.0184
ENTFERNUNG	0.265619	0.134538	1.974306	0.0483
STELLUNG	4.590040	1.957320	2.345063	0.0190

R-squared	0.635631	Mean dependent var	0.416667
S.D. dependent var	0.503610	S.E. of regression	0.280037
Akaike info criterion	0.744953	Sum squared resid	1.646835

Schwarz criterion	0.892210	Log likelihood	-5.939442
Restr. log likelihood	-16.30064	LR statistic	20.72239
Avg. log likelihood	-0.247477	Prob(LR statistic)	3.16E-05
Obs with Dep=0	14	Total obs	24
Obs with Dep=1	10		

a) Enter the numerically specified logit model of the means of transport selection

- with the estimated probabilities $\hat{p}_t$,

- with the estimated log odds $\ln[\hat{p}_t/(1-\hat{p}_t)]$

as regression values!

solution:

Dependent variable: Type of transport (y)

Independent variables: Distance (x_2), professional status (x_3)

- Logit model with estimated probabilities as regression values $\hat{p}_t$ as regression values:

$$\hat{p}_t = \frac{e^{-5,005772+0,265619\cdot x_2+4,590040\cdot x_3}}{1+e^{-5,005772+0,265619\cdot x_2+4,590040\cdot x_3}}$$

- Logit model with estimated log odds $\ln[\hat{p}_t/(1-\hat{p}_t)]$ as regression values:

$$\ln\left(\frac{\hat{p}_t}{1-\hat{p}_t}\right) = -5,005772 + 0,265619\cdot x_2 + 4,590040\cdot x_3$$

b) How can the effect coefficients of the variables distance (x2) and occupational status (x3) be interpreted?

solution:

Effect coefficient of distance: $e^{\hat{\beta}_2} = e^{0,265619} = 1,304$

If the distance to the place of work (x_2) increases by 1 km, the odds increase by a factor of 1.304, i.e. the odds of choosing the car instead of public transport as a means of transport for getting to work increase by 30.4% ceteris paribus.

Effect coefficient of distance: $e^{\hat{\beta}_3} = e^{4,590040} = 98,498$

The odds increase by a factor of 98.498 for an employee with a high occupational status (x3 = 1) compared to an employee with a low occupational status (x3 = 0). The ratio of chances to choose the car instead of public transport thus increases by 97.498% for a transition from an

employee in a low occupational position to an employee with a high occupational position, ceteris paribus.

c) Specify the marginal effect of distance to work (x_2) on the probability of choosing a car for employees who have the profile
- of the employee1,
- of the employee 2
own?

solution:

Marginal effect of the distance to the workplace (x_2) on the probability p for choosing a car:

$$\frac{\partial \hat{p}_t}{\partial x_{2t}} = \frac{e^{\hat{\beta}_1+\hat{\beta}_2\cdot x_{2t}+\hat{\beta}_3\cdot x_{3t}}}{\left(1+e^{\hat{\beta}_1+\hat{\beta}_2\cdot x_{2t}+\hat{\beta}_3\cdot x_{3t}}\right)^2}\cdot\hat{\beta}_2$$

- Characteristic profile of the employee 1: $\mathbf{x}'_1 = (1 \;\; 26 \;\; 1)$

$$\frac{\partial \hat{p}_1}{\partial x_{21}} = \frac{e^{\hat{\beta}_1+\hat{\beta}_2\cdot x_{21}+\hat{\beta}_3\cdot x_{31}}}{\left(1+e^{\hat{\beta}_1+\hat{\beta}_2\cdot x_{21}+\hat{\beta}_3\cdot x_{31}}\right)^2}\cdot\hat{\beta}_2 = \frac{e^{-5,00577+0,26562\cdot 26+4,59004\cdot 1}}{\left(1+e^{-5,00577+0,26562\cdot 26+4,59004\cdot 1}\right)^2}\cdot 0,26562$$

$$= \frac{658,78024}{(1+658,78024)^2}\cdot 0,26562 = \frac{658,78024}{435309,9617}\cdot 0,26562 = 0,000402$$

- Characteristic profile of the employee 1: $\mathbf{x}'_2 = (0 \;\; 5 \;\; 0)$

$$\frac{\partial \hat{p}_2}{\partial x_{22}} = \frac{e^{\hat{\beta}_1+\hat{\beta}_2\cdot x_{22}+\hat{\beta}_3\cdot x_{32}}}{\left(1+e^{\hat{\beta}_1+\hat{\beta}_2\cdot x_{22}+\hat{\beta}_3\cdot x_{32}}\right)^2}\cdot\hat{\beta}_2 = \frac{e^{-5,00577+0,26562\cdot 5+4,59004\cdot 0}}{\left(1+e^{-5,00577+0,26562\cdot 5+4,59004\cdot 0}\right)^2}\cdot 0,26562$$

$$= \frac{0,02528}{(1+0,02528)^2}\cdot 0,26562 = \frac{0,02528}{1,05120}\cdot 0,26562 = 0,00639$$

d) What is the size of the pseudo-R^2 marked „R-squared“ in EViews? Like can it be interpreted

solution:

$PR^2 = 0.635631$

L^* = -5.939442 (s. EViews-Output)

$$L_0^* = \ln\left(\frac{n_1}{n}\right)\cdot n_1 + \ln\left(\frac{n_2}{n}\right)\cdot n_2 = \ln\left(\frac{14}{24}\right)\cdot 14 + \ln\left(\frac{10}{24}\right)\cdot 10 = -0{,}538997\cdot 14 - 0{,}875469\cdot 10$$
$$= -7{,}545958 - 8{,}75469 = -16{,}300648$$

Pseudo-R^2: $PR^2 = 1 - \frac{L^*}{L_0^*} = 1 - \frac{-5{,}939442}{-16{,}300648} = 1 - 0{,}364368 = 0{,}636$

The pseudoR^2 of 0.636 indicates that the value of the likelihood function can be reduced by 63.6% by estimating the probability of choosing the car as a means of transport to get to work using the explanatory variables with the logit model instead of the mere frequency of 10/24 = 0.417.

Appendix: Data on choice of transport mode (Task 7.6)

Employee (t)	Type of transport (y)	distance (x_2)	Professional position(x_3)
1	1	26	1
2	0	5	0
3	1	17	1
4	0	1	1
5	0	14	0
6	1	28	1
7	0	19	0
8	0	2	0
9	1	22	0
10	0	8	0
11	0	6	0
12	1	18	1
13	0	4	0
14	0	12	0
15	0	3	0
16	1	15	1
17	1	8	0
18	0	10	0
19	0	2	0
20	1	10	1
21	1	16	1
22	0	11	0
23	0	4	0
24	1	5	1

Type of transport (y): 0 PT, 1 PC

Professional position (x_2): 0 low professional position, 1 high professional position