Prof. Dr. S. Kranz Summer term 2018
M. Kies 2018-04-17
Sheet 3

Exercises in Empirical Industrial Organization and Consumer Choice

(Presentation: Tuesday, 2018-05-22T14:15 [2:15 p.m.], He18 R120)

Exercise 3 - Analysing the US market for broilers

This exercise has been adapted from an exercise of Glenn Harrison's IO course at the MIT

The data set *broiler.csv* (on Moodle) contains quantity, price, cost, and demographic variables on broiler chickens over 40 years in the United States. The data is taken from Dennis Epple and Bennett McCallum's paper: "Simultaneous Equation Econometrics: The Missing Example"[1]. Inspecting the data, you will see the following column headers: *year*, *q*, *y*, *pchick*, *pbeef*, *pcor*, *pf*, *cpi*, *qproda*, *pop*, and *meatex*. The cryptic names are common in empirical work, even appearing in a dataset that is intended to be used for instructional purposes. To decode them have a look at Table 1.

column header	explanation
cpi	consumer price index
meatex	exports of beef, veal, and pork in pounds
pbeef	price index of beef
pchick	price index of chicken
pcor	price index of corn
pf	price index of chicken feed
pop	U.S. population in millions
q	per-capita consumption of chicken in pounds, measured by boneless equivalent
qproda	aggregate production of chicken in pounds
y	real disposable income per capita
year	year of data

Table 1: Decoding table for dataset broiler.csv.

We are interested in using this data to estimate the demand curve for broiler chickens. Our goal is to estimate the parameters of a demand function that is linear in logs

$$\log q = \alpha + \beta \log p + \log(X)\gamma + \varepsilon$$

[1] Epple, Dennis and McCallum, Bennett T., "Simultaneous Equation Econometrics: The Missing Example" (2005). *Tepper School of Business.* Paper 111. `http://repository.cmu.edu/tepper/111`

What should enter the matrix X? The answer is everything in our data set that we think is going to shift demand for chickens.

(a) Why do we want our demand function to have such a form? Compare with Exercise 2 and the concept of elasticities.

Solution:

The form is analogous to Exercise 2 in that we assume our demand curve to have the form

$$q = e^{\alpha} \cdot p^{\beta} \cdot X_1^{\gamma_1} \cdot ... \cdot X_n^{\gamma_n} \cdot e^{\varepsilon}$$

with some $n \in \mathbb{N}$.
β is the elasticity of the price, as

$$\frac{\mathrm{d}q}{\mathrm{d}p} = \underbrace{e^{\alpha} \cdot X_1^{\gamma_1} \cdot ... \cdot X_n^{\gamma_n} \cdot e^{\varepsilon}}_{=\frac{q}{p^{\beta}}} \cdot \beta p^{\beta-1} = \beta \cdot \frac{q \cdot p^{\beta-1}}{p^{\beta}} = \beta \cdot \frac{q}{p}$$

$$\Rightarrow \quad \beta = \frac{\mathrm{d}q}{\mathrm{d}p} \cdot \frac{p}{q} = \frac{\mathrm{d}q/q}{\mathrm{d}p/p}$$

In other words, an increase of 1% of the price results in an increase of approximately β% in demand.

(b) Download the data set and load it into a data frame in R.

Solution:

First we set the working directory to the path where we saved the data. Afterwards we import the data and have a look at it to ensure, that it has been correctly imported.

```
> setwd("Your Path")
> broiler <- read.csv("broiler.csv")
> head(broiler)
  YEAR    Q     Y PCHICK PBEEF PCOR       PF  CPI  QPRODA     POP MEATEX
1 1960 19.2  9210   52.4  33.5 46.0 51.53361 29.6 4333602 180.671     50
2 1961 20.6  9361   47.4  33.0 45.1 51.86824 29.9 4944130 183.691     49
3 1962 20.6  9666   50.0  34.2 44.8 52.09133 30.2 4997189 186.538     46
4 1963 21.1  9886   49.3  33.8 49.8 50.97588 30.6 5269019 189.242     80
5 1964 21.3 10456   48.2  32.8 49.9 50.75279 31.0 5443769 191.889     78
6 1965 22.9 10965   49.8  34.4 51.8 50.97588 31.5 5871560 194.303     49
```

(c) Perform an OLS regression of log prices on log quantity, leaving out all X's. What is the interpretation of the coefficient on price, and what do you make of its sign? Do you think the OLS estimator is consistent?

Solution:

```
> ols.easy <- lm(log(Q) ~ log(PCHICK), dat=broiler)
> #note that this is identical to "with(broiler, ols.easy <- lm(log(Q) ~ log(PCHICK)))" and "ols.easy <- lm(log(broiler$Q)~log(broiler$PCHICK))"
> summary(ols.easy)

Call:
lm(formula = log(Q) ~ log(PCHICK), data = broiler)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19629 -0.05663 -0.01534  0.06771  0.17920

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.54020    0.14419   3.746 0.000595 ***
log(PCHICK)  0.65952    0.03221  20.477  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08573 on 38 degrees of freedom
Multiple R-squared:  0.9169,  Adjusted R-squared:  0.9147
F-statistic: 419.3 on 1 and 38 DF,  p-value: < 2.2e-16
```

The coefficient of log(price) is positive with a value of ≈ 0.66. We could therefore assume that with an enhancment of log(price) of 1 unit the quantity of the log of consumed chicken increases by 0.66 units, or alternatively that an increase of 1% of the price increases the quantity of consumed chickens by 0.66%.[2] We would expect a negative correlation between price and quantity. This translates to an expected negative correlation between price and quantity due to the monotony of the logarithm. We do not see this and analogous to the ice cream model we can assume that the prices are set higher depending on other observable variables. In other words, we can assume the price to be endogenous. This results in the OLS estimator not being consistent.

[2] For more information regarding the interpretation of this regression see
Boit, Kenneth., "Linear Regression Models with Logarithmic Transformations" (2011).Page 4. *Methodology Institute, London School of Economics.* `http://kenbenoit.net/assets/courses/ME104/logmodels2.pdf`.

(d) Which additional explanatory variables should enter the demand function for chicken? Which variables may not enter the demand function but might be valid instruments for the price?

Solution:

(a) Explanatory variables which might plausibly enter the demand function:

- *cpi*
  The consumer price index might be relevant for the demand of chicken. If, for example chicken is bought by poor people and the *cpi* rises, than this means, that the population is getting relatively poorer. Keep in mind, that in this scenario all other variables (especially *income*) are held constant. One might argue however, that this effects should not be very strong and that it might be more sensible to have a look at more direct substitutes. An alternative idea could be to normalize all relevant variables in regard to the cpi (see Excursus).
- *pbeef*
  The price for beef is surely very relevant for the demand of chicken, as beef can be seen as a substitute for chicken. We would expect, that when the price for beef rises, that demand for chicken increases as well. Using *pbeef* for the demand function might therefore be a good idea.
- *pchick*
  The price for chicken should enter the demand function - we assume this price to be endogeneous so we need instrumental variables to counteract this. We would expect with rising prices a lower demand for chicken.
- *pop*
  It is not obvious why the population might be relevant for the consumption of chicken per capita. We will see however, that population is highly significant in most models. This might be due to chance or might capture some sort of weird time trend. One could argue however, that the population of the US increased due to immigration and that some immigrants (which might have a higher percentage of Hindus and Muslims) have more interest in eating chicken. Using the population as an explanatory variable might capture some of those effects.
- *year* or *y*
  To capture time trends it might be necessary to use time variables within the regression. There are a lot of different ways in dealing with time-variant data, like time lags or corrections for autocorrelation, which are not fokus of this lecture.

(b) A valid instrument is a variable z, which has the following two properties:

1. z is correlated with the endogenous variable p
2. z is not correlated with the disturbance ε: $\text{cor}(z, \varepsilon) = 0$

We have thus to check those properties with the relevant variables:

- *pcor*

    1. The price for corn is surely correlated to the price of chicken. An increase in the production costs should result in an increase of the price of chicken feed and thus an increase in the price for chicken. We would expect $\text{cor}(z, p) > 0$ and indeed find $\text{cor}(z, p) \approx 0.65$
    2. Assuming that the customers do not care for corn, we would not expect a correlation to the demand-shock. Another point in favor of a neglegible correlation is that corn prices are build based on a lot of factors with chicken being a very minor one. A demand-shock for chicken should therefore not be relevant to the price of corn. One might assume $\text{cor}(z, \varepsilon) = 0$

    It might be a viable instrument.

- *pf*

    1. The price for corn is surely correlated to the price of chicken. An increase in the production costs should result in an increase of the price. We would expect $\text{cor}(z, p) > 0$ and indeed we find a very strong $\text{cor}(z, p) \approx 0.91$
    2. The correlation to the demand-shocks is a bit trickier. One could argue that a demand shock leads to a reaction in the production. More production might increase the demand for chicken feed which would increase the price of chicken feed. One might still assume $\text{cor}(z, \varepsilon) \approx 0$ but thats not so clear cut. Using the Sargan Test with different constellations of possible instrumental variables shows, that the combination of *pf* and *meatex* results in the strongest rejection of the Sargan test, giving us reasonable doubt regarding *pf* as an instrument, even though the combination *pf*/*pcor* fares better than *meatex*/*pcor*

    It might be a viable instrument but less so than the price for corn.

- *meatex*

    1. The export of other meat is probably correlated with the price - as it is relevant for the demand function of other states. We would expect $\text{cor}(z, p) > 0$, as an increase in *meatex* might be a reason for farmer to switch from chicken to other meats or in the longer term built up more infrastructure regarding other meats. A higher *meatex* might thus lead to pressure on the supply curve thus increasing the price. Indeed we find that $\text{cor}(z, p) \approx 0.88$ but this might also be due to strong time trends.
    2. It is not obvious why a demand shock for chicken should influence the export rates of other meats given constant prices. Additionally one might argue,

    that a higher export rate should not influence consumer demand so one might assume $\mathrm{cor}(z, \varepsilon) = 0$. However, using the Sargan Test with *meatex* and other possible instruments implies that the there is a realistic chance that we have an endogeniety problem with this instrument, i.e. $\mathrm{cor}(z, \varepsilon) \neq 0$. Several reasons might be brought up: A demand shock for chicken within the state might get farmers to concentrate on internal markets and not to make trade deals with international markets. A demand shock *now* might lead farmers to assume that prices might rise in the future, even given constant prices now - and thus let them concetrate more on chicken than other meat for export.

  It might be a viable instrument, but we should be wary.

- *qproda*
  1. The production of chickens influences the Supply and thus the price. A higher production should therefore result in a lower price. On the other hand a higher price should lead to a higher volume in productions so we could also expect a postive correlation. Indeed we find $\mathrm{cor}(z, p) \approx 0.96$, so this effects stomps the other one and there certainly is enough explaining power for the first condition.
  2. The customers do probably not directly react to a possible overproduction (only thorugh the price). On the other hand an increase in demand will most certainly lead to a higher production. Thus, the production of chicken should be assumed to be correlated. Indeed the Sargan Test is strongly rejected when using QPRODA as an instrumental variable.

  It might not be a viable instrument, so we should refrain from using it.

Note that these answers are not as clear cut as would be the case in a theoretical example and do not claim to be the only viable way to solve this subtask.

(e) Estimate a sensible model with instrumental variable estimation. Use the function *ivreg()* of the package *AER* and assign the output to the variable *iv*. Run *summary(iv, diagnostics=TRUE)*. What is the interpretation for your estimate of $\hat{\beta}$? What is your interpretation regarding the output of the diagnostic tests Weak Instruments, Wu-Hausmann and Sargan?

Solution:

```
> library(AER)
> iv <- ivreg(log(Q)~log(CPI) + log(PBEEF) + log(PCHICK) + log(POP) + log(Y)
    |. - log(PCHICK) + log(MEATEX) + log(PCOR),dat=broiler)
> #identical to "iv <- ivreg(log(Q)~log(CPI) + log(PBEEF) + log(PCHICK) + log
    (POP) + log(Y)|log(CPI) + log(PBEEF) + log(POP) + log(Y) + log(MEATEX) +
    log(PCOR),dat=broiler)"
> summary(iv, diagnostics=TRUE)

Call:
ivreg(formula = log(Q) ~ log(CPI) + log(PBEEF) + log(PCHICK) +
    log(POP) + log(Y) | . - log(PCHICK) + log(MEATEX) + log(PCOR),
    data = broiler)

Residuals:
      Min        1Q    Median        3Q       Max
-0.078995 -0.021924 -0.001789  0.029594  0.056467

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.82455    1.20165  -8.176 1.55e-09 ***
log(CPI)     0.15449    0.10971   1.408   0.1682
log(PBEEF)   0.05919    0.09612   0.616   0.5421
log(PCHICK) -0.35891    0.14195  -2.529   0.0163 *
log(POP)     2.77296    0.42060   6.593 1.48e-07 ***
log(Y)      -0.10746    0.16193  -0.664   0.5114

Diagnostic tests:
                 df1 df2 statistic  p-value
Weak instruments   2  33    11.393 0.000173 ***
Wu-Hausman         1  33     1.524 0.225673
Sargan             1  NA    15.479 8.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03734 on 34 degrees of freedom
Multiple R-Squared: 0.9859, Adjusted R-squared: 0.9838
Wald test: 475.7 on 5 and 34 DF,  p-value: < 2.2e-16
```

Note, that we did leave out *year*. Common ways to include time would be via a time factor variable (ie. each year gets a fixed effect), but this would result in an overfitted model. Putting time in would imply a mono-elastic relationship to time according to our calendar - which seems

not very realistic. However this might be a better idea than implying no trends at all which we are doing within this analysis.
An increase in *pbeef* results in an increase in chicken consumption, which is plausible, as beef can be seen as a substitute for chicken.
$\hat{\beta}$ shows a negative sign, which is what we would expect.
The positive sign regarding the population is not as obvious. One could argue, that this captures a time trend or the fact, that the population of the US increased due to immigrants which have a higher demand for chicken. Indeed, including a linear time trend reverses the sign to a negative estimator, throwing doubt on our "immigrant" hypothesis. However, as the time trend does not need to be linear, one might still make a time trend error in this scenario.
The effects of *cpi* and *y* are not as obvious, as well. One could argue that - given the income stays constant - an higher consumer price index means, that the people turn to less expensive food, i.e. chicken. An higher income on the other hand could mean that people buy more expensive food or save more. Both regressor are very weak though and not significant and should therefore not be overinterpreted.

Regarding the tests it is to note that

- Weak instruments:
  Strongly significant. We strongly reject the Null Hypothesis so we may assume that our instruments have indeed a relevant influence on the price. This result is good for our analysis.
- Wu-Hausmann:
  Not significant. We may not reject the Null Hypothesis. It is not obvious (based on this test), that we indeed have an endogenity problem. It might be that the price is exogeneous. This result is not very good for our analysis, as economic reasoning would lead us to believe that we should indeed have an endogenity problem. As including the *year* variable gives us much better and more expected results this result might be due to some weird time trend effects.
- Sargan:
  Strongly significant. We may strongly reject the Null Hypothesis and thus assume there to be some problems with our instrumental variables as they appear to be correlated to the ε-shock.

(f) In addition to your estimations in a) and d) estimate 2 additional models (via IV or OLS). Store each regressed model in a variable. Install and load the R package *texreg*. Use the function *screenreg()* to show a table that compares your 4 estimated models in the typical format you find in academic journals.

Solution:

```
> library(texreg)
> iv <- ivreg(log(Q)~log(CPI) + log(PBEEF) + log(PCHICK) + log(POP)
    + log(Y)|. - log(PCHICK) + log(MEATEX) + log(PCOR),dat=broiler)
> ols.easy <- lm(log(Q) ~ log(PCHICK), dat=broiler)
> iv.alt1 <- ivreg(log(Q)~log(CPI) + log(PBEEF) + log(PCHICK) + log
   (POP) + log(Y) + YEAR|. - log(PCHICK) + log(MEATEX) + log(PCOR),
   dat=broiler)
> iv.alt2 <- ivreg(log(Q)~log(CPI) + log(PBEEF) + log(PCHICK) + log
   (POP) + YEAR|. - log(PCHICK) + log(MEATEX) + log(PCOR) + log(PF)
   ,dat=broiler)
> screenreg(list(iv, ols.easy, iv.alt1, iv.alt2), custom.model.
   names=c("iv","ols.easy","iv.alt1", "iv.alt2"))
```

results in the following output of figure 1.

The first additional model is identical to our main model but with the added variable *year*. Note, that I did not logarithmise *year*, as the interpretation "Each year the chicken consumption increases by $X\%$" makes more sense than "If we increase the counter of our calendar by 1%, than the chicken consumption increases by $Y\%$".

The second additional model has several changes: We let go of y as *cpi* and y measure similar things and were not significant, and we added *pf* as an additional instrument.

Note that adding *time* results in the regressor of *pop* to be negative, which seems rather unrealistic.

```
============================================================
             iv          ols.easy    iv.alt1      iv.alt2
------------------------------------------------------------
(Intercept)  -9.82 ***    0.54 ***  -189.99 ***  -174.09 ***
             (1.20)      (0.14)      (33.66)      (31.70)
log(CPI)      0.15                    -0.30 *      -0.24 *
             (0.11)                   (0.11)       (0.10)
log(PBEEF)    0.06                     0.32 **      0.23 **
             (0.10)                   (0.09)       (0.08)
log(PCHICK)  -0.36 *      0.66 ***    -0.65 ***    -0.58 ***
             (0.14)      (0.03)       (0.13)       (0.12)
log(POP)      2.77 ***                -5.85 ***    -5.52 **
             (0.42)                   (1.61)       (1.55)
log(Y)       -0.11                    -0.21
             (0.16)                   (0.13)
YEAR                                   0.12 ***     0.11 ***
                                      (0.02)       (0.02)
------------------------------------------------------------
R^2           0.99        0.92         0.99         0.99
Adj. R^2      0.98        0.91         0.99         0.99
Num. obs.    40          40           40           40
============================================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

Figure 1: Output of *screenreg* based on the models of subtask (f).
Note: Another very useful package concerning the display of regression results is *stargazer*.

Excursus

Abstracting from the specific model and task description, one might argue, that all relevant variables have to be normalized by the price index:

```
> broiler$p.real <- broiler$PCHICK/broiler$CPI
> broiler$p.beef.real <- broiler$PBEEF/broiler$CPI
> broiler$p.corn.real <- broiler$PCOR/broiler$CPI
> broiler$p.feed.real <- broiler$PF/broiler$CPI
> broiler$y.real <- broiler$Y/broiler$CPI
>
>
> iv.alt3 <- ivreg(log(Q)~p.beef.real + p.real + POP + YEAR|. - p.real
    + log(MEATEX) + p.corn.real + p.feed.real,dat=broiler)
> summary(iv.alt3, diagnostics=TRUE)

Call:
ivreg(formula = log(Q) ~ p.beef.real + p.real + POP + YEAR |
    . - p.real + log(MEATEX) + p.corn.real + p.feed.real, data =
       broiler)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09764 -0.02644 -0.00517  0.02728  0.08078

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.940037  27.115339   1.473  0.14970
p.beef.real  0.203802   0.100417   2.030  0.05006 .
p.real      -0.421051   0.129167  -3.260  0.00249 **
POP          0.015747   0.005279   2.983  0.00517 **
YEAR        -0.020071   0.014261  -1.407  0.16811

Diagnostic tests:
                df1 df2 statistic  p-value
Weak instruments   3  33     5.438 0.003761 **
Wu-Hausman         1  34    13.208 0.000911 ***
Sargan             2  NA     4.408 0.110337
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.04374 on 35 degrees of freedom
Multiple R-Squared: 0.9801, Adjusted R-squared: 0.9778
Wald test: 435.2 on 4 and 35 DF,  p-value: < 2.2e-16
```

Note that we did not include log() transformations for our transformed variables. In this case the interpretation of p.real woud be “The demand for chicken decreases by approx. 0.42% if the price for chicken increases by one percentage point compared to other goods. It seems sensible that p.beef.real influences the demand less than the price of chicken directly.
One might still be conflicted about stationarity of the process and having a look at acf() and pacf() implies autocorrelation. On easy way to deal with this is using Differencing [3].

```

> broiler$logQ <- log(broiler$Q)
> broiler$logPop <- log(broiler$POP)
> broiler$logMeatEx <- log(broiler$MEATEX)
> broiler.diff <- broiler[-1,]-broiler[-nrow(broiler),]
> iv.alt4 <- ivreg(logQ~p.beef.real  + p.real + logPop |. - p.real +
   logMeatEx + p.feed.real,dat=broiler.diff)
> summary(iv.alt4, diagnostics=TRUE)

Call:
ivreg(formula = logQ ~ p.beef.real + p.real + logPop | . - p.real +
    logMeatEx + p.feed.real, data = broiler.diff)

Residuals:
     Min       1Q   Median       3Q      Max
-0.05588 -0.01445 -0.00544  0.01428  0.04624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02084    0.02397   0.869  0.39074
p.beef.real  0.26544    0.09000   2.949  0.00564 **
p.real      -0.20436    0.07162  -2.853  0.00722 **
logPop       0.26718    2.14395   0.125  0.90154

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   2  34     8.502 0.00101 **
Wu-Hausman         1  34     0.511 0.47961
Sargan             1  NA     0.368 0.54410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

[3] For an overview see e.g. OTexts (2016). *8.1 Stationarity and differencing.* URL: https://www.otexts.org/fpp/8/1 (visited on 04/23/2018)

```
Residual standard error: 0.02405 on 35 degrees of freedom
Multiple R-Squared: 0.3304, Adjusted R-squared: 0.273
Wald test: 3.634 on 3 and 35 DF,  p-value: 0.02207
```

The interpretation of the result is pretty complicated, but the signs of the result are what has been expected. Two other curious effects should be noted: The population effects vanishes - implying that it could very well be, that it just captured a timetrend effect. Additionally it is not obvious anymore that we have an endogenity problem (see Wu Hausmann-Test). Indeed, using OLS yields pretty much the same results, even though the strength of the price coefficent weakens a bit.

Additional Material: Table of Endogenity Tests

This table is intended to give you a better overview of the tests related to endogenity and to help you to structure your thoughts on this matter. Information missing within this table does not imply that this information is not relevant for the exam.

Name of Test	Tests for	Null-Hypothesis	Other
Wu-Hausmann	Endogenity of regressors (e.g. price p) “Do we have an endogenity problem?”	All variables are exogeneous.	
Weak Instruments	“Are the instruments sufficiently correlated with the endogenous variable?”	No Instrument has an influence on the possibly endogeneous variables	F-test on the joint significance of all explanatory variables in the first stage regression of the two stage least squares procedure
Sargan	Endogenity of instruments “Are our instruments correlated with the ε-shock?”	All Instruments are exogeneous.	We need at least one more excluded instrument than endogeneous variables.

Table 2: Overview over the tests relating to endogenity.