Introductory Econometrics

Exercises for tutorials (Fall 2014)

Jan Zouhar
Dept. of Econometrics, Uni. of Economics, Prague, zouharj@vse.cz

September 23, 2014

Tutorial 1: Review of basic statistical concepts

Exercise 1.1 (*Three M's.*) Assign each term in the 1–4 list a meaning from the *a*–*d* list.
Note: two of the 1–4 terms are actually synonymous.

1) Mean.
2) Median.
3) Mode.
4) Expected value.

a) The most likely/frequent value in a population.
b) The long-run average of the results of many independent draws from a population.
c) The value separating the higher half of a population from the lower half.
d) The weighted average of possible values, with the weights being the probabilities of the respective values.

Exercise 1.2 (*Wages: mean vs. median.*)

a) "The average monthly wage in a population is €1,000." Does the *average wage* relate to the *mean*, *median* or *mode* of the population's wage distribution?
b) Which is typically greater, the *mean wage* or the *median wage*? (Or, do most people earn *more than* or *less than* the average wage?)

Exercise 1.3 (*Calculating the expectations.*)

a) Let x be a random variable (RV). Is it possible that $\Pr\{x = \mathsf{E}x\} = 0$? (For instance, is it possible that nobody actually earns the average wage?)
b) What is the expected value of a single die roll (for a six-sided die)?
c) Consider a "loaded" die with uneven probabilities of the possible results, stated in Table 1. What is the expected value of a roll with such a die?

Table 1: A loaded die

k	1	2	3	4	5	6
$\Pr\{x = k\}$	.1	.1	.1	.2	.2	.3

Exercise 1.4 (*Variance and standard deviation.*)

a) How would you describe the term *variance of an* RV to somebody who doesn't know anything about statistics?
b) Figure 1 shows distributions of two RVs. Which one has greater variance?
c) Compare the variances of die-roll results for an ordinary six-sided die and the loaded die from Table 1.

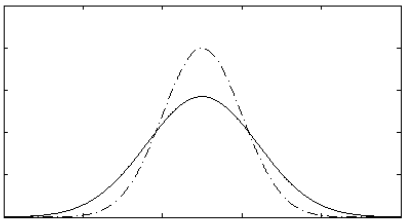

Figure 1: Two distributions, different variances

d) Assume that *height* of a person is approximately *normally distributed* with a mean of 180 cm and variance σ^2. What percentage of the population falls within $\pm\sigma$ from the population average (i.e., in the interval $[180-\sigma, 180+\sigma]$)? And how about the $\pm 2\sigma$ or $\pm 3\sigma$ range? Draw a plot that illustrates this.

e) RV x has the following characteristics: $\mathsf{E}x = 10$, $\mathsf{var}x = 0$. Is there anything more we can say about x?

Exercise 1.5 (*Calculations with expected value and variance.*) Let x and y be independent RVs with

$$\begin{aligned} \mathsf{E}x &= 10, & \mathsf{E}y &= 5, \\ \mathsf{var}\,x &= 1, & \mathsf{var}\,y &= 2. \end{aligned}$$

Calculate:

a) $\mathsf{E}[4x]$.
b) $\mathsf{E}[4x+5]$.
c) $\mathsf{E}[x+y]$.
d) $\mathsf{E}[x-y]$.
e) $\mathsf{E}[4x-3y+5]$.
f) $\mathsf{var}[4x]$.
g) $\mathsf{var}[4x+5]$.
h) $\mathsf{var}[x+y]$.
i) $\mathsf{var}[x-y]$.
j) $\mathsf{var}[4x-3y+5]$.

Exercise 1.6 (*Multiple dice.*)

a) Imagine we roll two six-sided dice and add up the results. What are the possible outcomes? What are their probabilities? Draw a plot of the probability function.

b) What is the variance of the RV from *a*? (*Hint*: the variance of a die roll is $\frac{35}{12}$.)

c) Suppose we take the arithmetic average of 10 die rolls. What is the expected value and variance of the result?

Exercise 1.7 (*Random sample and sample mean.*) The population distribution of the number of teeth (x) has a mean of 20 with a variance of 100. Assume we draw (at random) a sample of 10 people, measure the value of x for each one of them (thus obtaining values $x_1, x_2, \ldots, x_{10}$), and then calculate the arithmetic average $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$. Due to random sampling, $\bar{x}$ is a random variable.

a) What is the expected value of $\bar{x}$? What is its variance?

b) (*Law of large numbers.*) Instead of 10 people, we take n now. What happens to $\mathsf{E}\bar{x}$ and $\mathsf{var}\bar{x}$ if we gradually raise n above all limits?

c) (*Central limit theorem.*) Again, consider a sample of n people, only that now we study the result of

$$y = \sqrt{n}(\bar{x} - 20) = \frac{\sum_{i=1}^{n}(x_i - 20)}{\sqrt{n}}.$$

As n grows, what happens to the distribution of y?

Exercise 1.8 (*Visualizing the central limit theorem.*) Download the `CLT.zip` archive from my website and run the `CLT.m` file in *Matlab*. You should obtain a figure similar to Figure 2. It shows probability distributions of standardized means (see below) of samples with different sizes, drawn from the population stored in vector `x` in the *Matlab* code, line 4. Here, a *standardized mean* is the

expression $\sqrt{n}(\bar{x} - \mu)$, where μ is the population average. Therefore, the figure shows the effect of the *central limit theorem* (CLT). The starting population is {1,2,3,4,5,6}, which means that each observation is in fact a die roll. Notice how fast the distributions converge to the bell-shaped curve of *normal* (or *Gaussian*) distribution. Try changing the population, perhaps repeating some of the numbers, and see how the convergence process changes.

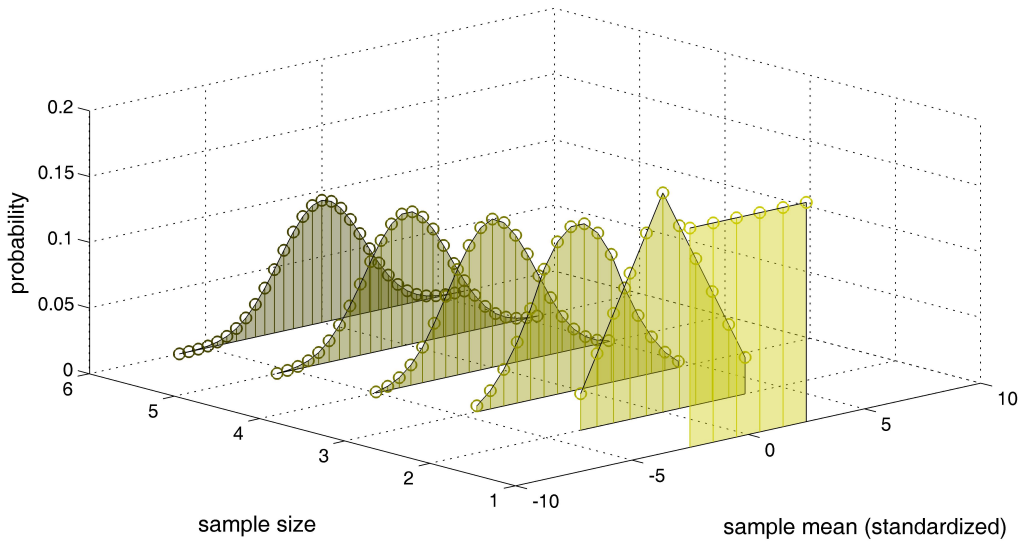

Figure 2: Die-roll standardized means (CLT demonstration)

Exercise 1.9 (*Correlation & covariance.*)

a) Would you say that *wages* and *education* are *positively correlated*, *negatively correlated* or *uncorrelated*? And how about *wages* a person's *height*?

b) Find an example of *negatively correlated* economic variables.

c) If we know that two RVs are negatively correlated, what does it tell us about their *covariance*?

d) What are the possible values of a covariance of two RVs?

e) Let x and y be independent. Is it possible that $\mathsf{cov}(x, y) = 0.58$? Why?

f) We know that $\mathsf{cov}(x, y) = 0$. Do x and y have to be independent? (If not, find an example of RVs that are *uncorrelated* despite *not* being *independent*.)

g) What are the possible values of a *correlation coefficient* of two RVs?

h) Which of the following can happen:

1) $\mathsf{corr}(x, y) = -1.56$.
2) $\mathsf{corr}(x, y) = 0.28, \quad \mathsf{cov}(x, y) = 0$.
3) $\mathsf{corr}(x, y) = 0.28, \quad \mathsf{cov}(x, y) = -0.5$.
4) $\mathsf{corr}(x, y) = 0.28, \quad \mathsf{cov}(x, y) = 0.5$.

Why? What is the relationship between the covariance and correlation coefficient of two RVs?

Exercise 1.10 (*Sharpening your eyes.*) In your web browser, type in the address

`http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/`.

Follow the instructions on the website: first press the begin button in the upper-left corner, then look at the scatterplot of two RVs (we'll denote them x and y here), and guess which of the five suggested numbers for $\mathsf{corr}(x, y)$ is correct. To see the correct value, click on the Show r button. Repeat the procedure (using the New Data button) until you manage to guess the right answer three times in a row.

Exercise 1.11 (*Conditional expectations.*)

a) What is the average monthly wage in your country, expressed in €? (Make a rough guess.)

b) Imagine you meet a person on the street and he/she tells you he/she had only finished elementary school before getting employed. Does this information change your expectation about the person's wage?

c) Estimate the following: $\mathsf{E}[wage|\, educ = 9], \mathsf{E}[wage|\, educ = 13], \mathsf{E}[wage|\, educ = 18]$.[1]

d) Based on *c*, try to approximate $\mathsf{E}[wage|\, educ]$ using a linear relationship

$$\mathsf{E}[wage|\, educ] = \beta_0 + \beta_1\, educ.$$

e) Based on *d*, what is the expected difference in wages for two people with a gap of 1 year in their education? In other words, what is the population average of $\frac{\Delta wage}{\Delta educ}$? (Or: what is $\frac{\Delta \mathsf{E}[wage|\, educ]}{\Delta educ}$?)

Exercise 1.12 (*Calculations with conditional expectations.*)

a) Let x and y be *independent* RVs, $\mathsf{E}y = 12.5$. Find $\mathsf{E}[y|x]$.

b) Let x and y be RVs, $\mathsf{E}[y|x] = 2 + 5x$. Find $\mathsf{E}[4y + 3xy + x^2|x]$ and $\mathsf{E}[4y + 3xy + x^2|x = 5]$.

Exercise 1.13 (*Conditional expectations II.*) Suppose that at a large university, college grade point average, *GPA*, and SAT score, *SAT*, are related by the conditional expectation

$$\mathsf{E}[GPA|\, SAT] = .70 + .002\, SAT.$$

a) Find the expected *GPA* when $SAT = 800$. Find $\mathsf{E}[GPA|\, SAT = 1{,}400]$. Comment on the difference.

b) If the average *SAT* in the university is 1,100, what is the average *GPA*?

Exercise 1.14 (*Conditional variance.*) Do you think the variance of *wages* varies among groups of people with different levels of *education*? E.g., is there a difference between $\mathsf{var}[wage|\, educ = 9]$ and $\mathsf{var}[wage|\, educ = 18]$?

Tutorial 2: Simple regression

Exercise 2.1 (*Gretl practice.*) Import data from the *MS Excel* file `simplereg.xls` into *Gretl* (use the *drag-and-drop* trick). (This is a fictitious dataset, the numbers don't have any real meaning.)

a) Regress y on x; i.e., estimate (using Model → Ordinary least squares) the model

$$\mathsf{E}[y|x] = \beta_0 + \beta_1 x.$$

b) Write down the estimated regression function. (*Note*: once we've estimated a model, we typically write $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.)

c) From the *Gretl* output, read the value of R^2. What does it tell us about the model?

d) Find (in *Gretl* or by calculation) the values of *SST*, *SSR* and *SSE*.

e) Draw the scatterplot of x_i's and y_i's with the estimated regression line in it (Graphs → Fitted, actual plot → Against x).

f) Save the residuals ($\hat{u}_i$) from the estimated model as a new variable (Save → Residuals). Next, find the sample mean of $\hat{u}$ (View → Summary Statistics) and sample correlation between $\hat{u}$ and x (View → Correlation Matrix). Is the result unexpected, or can we generalize it to other regression models? Explain why.

Exercise 2.2 (*Campaign expenditures.*) Load the data `voting.gdt`. The data describe election outcomes and campaign expenditures for 173 two-party races (A,B) for the U.S. House of Representatives in 1988.

[1] Here *educ* is expressed in years, i.e. 9 years typically represent elementary education and 18 years a master's degree.

a) Consider the following regression model:

$$\mathsf{E}[voteA\,|\,expenA] = \beta_0 + \beta_1\, expenA.$$

Does it make sense to use a model like this to describe the relationship between campaign expenditures and the eventual vote share? How would you interpret β_0 and β_1? In a two-party race, do you think it makes sense to look at the campaign expenditures of party A alone?

b) Consider the following regression model:

$$\mathsf{E}[voteA\,|\,shareA] = \beta_0 + \beta_1 shareA$$

where *shareA* is A's percentage share in the total campaign expenditures ("total" meaning the sum across all parties). Generate *shareA* in *Gretl* (use Add → Define new variable), estimate the model and interpret the estimates.

c) Find a story for the association between *voteA* and *shareA* supporting each of the three causation schemes.

Exercise 2.3 (*Constant elasticity model.*) Load the data `ceosal1.gdt` (the CEOs' salaries data from Lecture 2). This time, we relate *salary* to *sales*.

a) Consider the following population regression model:

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u.$$

Can you express the elasticity of *salary* with respect to *sales* in terms of the regression coefficients β_0 and β_1?

b) Generate $\log(salary)$ and $\log(sales)$ in *Gretl* (use Add → Logs of selected variables) and estimate the regression model.

c) Regress *salary* on *sales* without logarithms and look at the R^2's in both models. Does a comparison of the two R^2's tell us something meaningful?

Exercise 2.4 (*Monte Carlo.*) In the lectures, we study the *linear regression model* (LRM) using analytical means. There's another way to study the linear regression model, and that is using a computer simulation. Consider a population model

$$y = \beta_0 + \beta_1 x + u, \quad \beta_0 = 5,\ \beta_1 = 10 \tag{1}$$

and a random sample consisting of 15 observations. Carry out the following simulation in *MS Excel*. You can use the `MonteCarlo.xls` file if you like.

a) Generate x and u values at random and write them down in two columns. Use the RANDBETWEEN function to do this (the function generates a random integer within the specified bounds). You can pick any range for x; however, note that in a CLRM, $\mathsf{E}u$ has to be zero. Therefore, the lower and upper bounds for u have to be opposite numbers; i.e., use RANDBETWEEN$(-u_{\max}, u_{\max})$.

b) Create columns with both y and $\mathsf{E}[y|x]$; these will be calculated based on (1).

c) Draw a scatterplot of y vs. x and include the PRF in it (i.e., the line $\mathsf{E}[y|x] = \beta_0 + \beta_1 x$). Press F9 repeatedly to see how the random sampling procedure looks like. What does u represent in the plot?

d) Calculate the OLS-estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using INTERCEPT and SLOPE functions. Next, calculate the values of $\hat{y}$ and $\hat{u}$, and add the SRF line (i.e., the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$) into your plot. Press F9 repeatedly to see how accurately the OLS procedure estimates β_0 and β_1. Which one is more accurate (on average), $\hat{\beta}_0$ or $\hat{\beta}_1$?

e) Press F9 ten times, write down the results for $\hat{\beta}_0$ and $\hat{\beta}_1$, and then make the arithmetic average of the 10 trials for both $\hat{\beta}_0$ and $\hat{\beta}_1$. What results would you expect if we did 1000 trials instead of 10?

f) Open the `MonteCarlo2.xls` file. The experiment from *e* is automated here, only with 1000 trials instead of 10. The 1000 values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are shown in columns W and AC. In the same columns, you can see the mean and (sample) standard deviation of the 1000 trials. Compare the standard deviations for $\hat{\beta}_0$ and $\hat{\beta}_1$. Does the difference in the two values reflect your conclusions about the accuracy of the estimates?

g) The histogram plots on the right show the frequencies of $\hat{\beta}_0$ (green) and $\hat{\beta}_1$ (blue) values among the 1000 trials in the experiment. These plots tell us something about the shape of the distribution of RVs $\hat{\beta}_0$ and $\hat{\beta}_1$. Does the shape of the plots remind you of one of the well-known distributions?

Exercise 2.5 (*Conditional variance of $\hat{\beta}_1$ using Monte Carlo simulation.*) In the lectures, we derived a formula for conditional variance of $\hat{\beta}_1$ given $\boldsymbol{x}$ by analytical means. In this exercise, you'll verify the result using Monte Carlo simulation. In principle, this will be the same simulation as in the previous exercise, only that now we study the *conditional* distribution of $\hat{\beta}_1$. In order to do this, we need to select particular values of the explanatory variable and keep these values fixed in the repeated samples (i.e., only the values of u are sampled each time, x remains the same). Proceed as follows:

a) Open the `MonteCarlo3.xls` file.

b) Select specific values for x; e.g., replace the formulas in the green column I with odd numbers going from 1 to 29.

c) Look at the sample variance of the resulting 1000 trials for $\hat{\beta}_1$ in cell `X1006`. Compare this number to the analytic result, which says

$$\mathsf{var}[\hat{\beta}_1 | \boldsymbol{x}] = \frac{\sigma^2}{s_x^2}.$$

Note that $s_x^2 = \sum_{i=1}^{15}(x_i - \bar{x})^2$ and σ^2 is the variance of u. We're using RANDBETWEEN to generate u, which means u has *discrete uniform distribution*, the variance of which is

$$\frac{(b-a-1)^2 - 1}{12},$$

where a and b are the lower and upper bound, respectively. (Fill the formulas for σ^2, s_x^2 and $\mathsf{var}[\hat{\beta}_1 | \boldsymbol{x}]$ in cells `Q5`, `Q6` and `Q7`, respectively.) Press F9 repeatedly and comment on the difference between the analytical and simulation results.

Exercise 2.6 (*Factors affecting $\hat{\beta}_1$ variance.*) We'll continue working with the `MonteCarlo3.xls` file. From the lectures, you know that ...

i) ... the less variance in the disturbances,

ii) ... the more variance in the explanatory variable,

the more accurate estimates we obtain. Verify this using the simulation model.

a) Try changing the range of x-values (e.g., pick the numbers 10–24 or 0–70) and see how the variance of the estimates varies.

b) Try changing the $u_{\max}$ value and watch the resulting change in the variance of the estimates.

Tutorial 3: Multiple regression I

Exercise 3.1 (*Sleep vs. work*). The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years.

a) If adults trade off sleep for work, what is the sign of β_1?

b) What signs do you think β_2 and β_3 will have?

c) Using the data in `sleep.gdt`, estimate the model and write out the results in equation form.

d) If someone works five more hours per week, by how many minutes is *sleep* predicted to fall? Is this a large tradeoff?

e) Discuss the sign and magnitude of the estimated coefficient on *educ*.

f) Would you say *totwrk*, *educ*, and *age* explain much of the variation in *sleep*? What other factors might affect the time spent sleeping? Are these likely to be correlated with *totwrk*?

Exercise 3.2 (*Housing prices and pollution*). The following equation describes the median housing price in a community in terms of amount of pollution (*nox* for nitrous oxide) and the average number of rooms in houses in the community (*rooms*):

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u. \tag{2}$$

a) What are the probable signs of β_1 and β_2? What is the interpretation of β_1? Explain.

b) Why might *nox* (or the log of *nox*) and *rooms* be negatively correlated? If this is the case, does the simple regression of $\log(price)$ on $\log(nox)$ produce an upward or downward biased estimator of β_1?

c) Using the data in `houses.gdt`, estimate (2) and the following model:

$$\log(price) = \beta_0 + \beta_1 \log(nox) + u.$$

Is the relationship between the simple and multiple regression estimates of the elasticity of *price* with respect to *nox* what you would have predicted, given your answer in part *b*? Does this mean that $\hat{\beta}_1$ from (2) is definitely closer to the true elasticity than $\hat{\beta}_1$ from the simple regression model?

Exercise 3.3 (*Building up an econometric model*). You were asked to carry out empirical research in order to quantify the so-called *returns to schooling*, i.e., the effect of additional education on a person's wage. In lecture 1, we discussed the steps to be carried out in empirical analysis:

Step 1: Formulate the *question* of interest.
Step 2: Find a suitable *economic model*.
Step 3: Turn it into an *econometric model*.
Step 4: Obtain suitable *data*.
Step 5: Use econometric methods to *estimate* the econometric model.
Step 6: If needed, use *hypothesis tests* to answer the question of interest.

Step 1 has already been made for you: the question of interest is stated above. Your task here is to discuss steps 2, 3 and 4 in detail.

a) Put up a list of all thinkable factors that shape a person's wage.

b) Try to argue the causal link from each of the factors to a person's wage *from the standpoint of an economic theory*. Arguments such as "we all know that clever people earn more money" do not count here.

c) Explain how you would *quantify* the factors you identified. First, decide whether a factor is directly *measurable*, or whether we need to find a suitable *proxy variable*. Second, explain what units you'd use use for quantification.

d) Write down the econometric model you would use in order to estimate the effect of wages on education. Is it necessary to include all the factors (or their proxies) in the regression model? Are some of them more important than others?

e) Is it possible to drop one of the variables you included in the model without violating the $\mathsf{E}[u|educ] = 0$ assumption?

f) Based on your economic argumentation from *b*, what values of the β parameters do you expect? (For each jth variable, give at least the expected sign of β_j.)

g) Imagine you've collected the data you need and saved them into an *MS Excel* file. Sketch a structure of the *MS Excel* file (think up arbitrary data for the first two observations).

Tutorial 4: Multiple regression II

Exercise 4.1 (*Partialling out*). Using the 526 observations on workers in `wage.gdt`, regress the log of *wage* (hourly wage in $) on *educ* (years of education), *exper* (years of labor market experience), and *tenure* (years with the current employer).

a) Formulate the population regression model you are estimating.

b) Write down the estimated equation and interpret the values of the estimated coefficients.

c) Confirm the partialling out interpretation of the OLS estimates by explicitly doing the partialling out. This first requires regressing *educ* on *exper* and *tenure*, and saving the residuals, $\hat{r}_1$. Then, regress log(*wage*) on $\hat{r}_1$. Compare the coefficient on $\hat{r}_1$ with the coefficient on *educ* in the regression from *b*.

Exercise 4.2 (*Working with categories*). In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student the sum of hours in the four activities must be 168.

a) Consider the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 work + \beta_3 leisure + \beta_4 sleep + u.$$

Interpret the coefficient β_1. Does it make sense to hold *work*, *leisure*, and *sleep* fixed, while changing *study*?

b) Explain why this model violates Assumption MLR.3.

c) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?

Exercise 4.3 (*"Harmless" multicollinearity*). Suppose you postulate a model explaining final exam score in terms of class attendance. Thus, the dependent variable is final exam score, and the key explanatory variable is number of classes attended. To control for student abilities and efforts outside the classroom, you include among the explanatory variables cumulative GPA, SAT score, and measures of high school performance. Someone says, "You cannot hope to learn anything from this exercise because cumulative GPA, SAT score, and high school performance are likely to be highly collinear." What should be your response?

Exercise 4.4 (*Variable selection*). The causal links between variables y, x, z_1, z_2, z_3 and z_4 are shown in Figure 3. Your task is to quantify the causal effect of x on y. Which of the variables will you include in your equation?

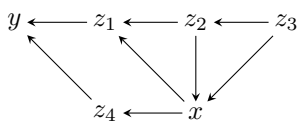

Figure 3: Causal links between variables y, x, z_1, z_2, z_3 and z_4

Exercise 4.5 (*Used cars*). The file `used_cars_original.xls` contains data on used *Škoda*s that I collected in January 2004. At that time, I owned an old *Škoda Felicia* and was thinking about selling it; I didn't sell it in the end, and later in 2009, it suddenly caught fire when my dad was driving it (accidentaly, this happened very close to our university premises), see

`http://www.pozary.cz/clanek/20375-obrazem-u-bulhara-horela-skodovka/`.

When you open the file in *MS Excel*, you'll notice that the data format is not suitable for econometric work (why?). Compare `used_cars_original.xls` with `used_cars.xls` and notice the way the *qualitative* variables were encoded into *dummies*.

a) Regress *price* on *km* and *age*; interpret all the estimated coefficients. Is there any meaningful interpretation of the intercept? Do you find its level reasonable?

b) Regress *price* on *km* and *year* now. Compare the results with your previous model. How would you interpret the intercept in this case?

c) In the data, I created the variable *age* from the year of manufacture in the following way: $age = 2004 - year$. Notice the impact this relationship had on the coefficients you estimated in *a* and *b*.

d) Regress *price* on all available explanatory variables. Why were some of the variables omitted by *Gretl*? Interpret the coefficients for the dummy variables that were retained.

e) Try and find the best regression model for the price of a used *Škoda*. Consider (and estimate) various function forms.

f) How much value does a used car lose (on average) with each additional kilometre? Discuss the various function shapes you used in *e*.

g) What price would you ask (in 2004) for *Škoda Felicia*, which has 100.000 km on the clock, the engine 1.9D and was manufactured in 1998?

h) Would be a version with petrol engine be cheaper? By how much?

i) What is the price difference between used *Octavia*s and *Felicia*s? What data will you use to find this out?

j) Find out whether the extra charge for the *combi* version varies for *Octavia*s and *Felicia*s.

k) Find out whether the extra charge for a diesel engine varies for *Octavia*s and *Felicia*s.

l) Find out whether the average value loss per km varies for *Octavia*s and *Felicia*s.

Tutorial 5: Hypothesis testing

Exercise 5.1 (*Theory check*). Which of the following can cause the usual OLS t statistics to be invalid (that is, not to have `t` distributions under H_0)?

a) Heteroskedasticity.

b) A sample correlation coefficient of .95 between two independent variables that are in the model.

c) Omitting an important explanatory variable.

Exercise 5.2 (*Practical vs. statistical significance*). Consider an equation to explain salaries of CEOs in terms of annual firm sales, return on equity (*roe*, in percent form), and return on the firm's stock (*ros*, in percent form):

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u.$$

a) In terms of the model parameters, state the null hypothesis that, after controlling for *sales* and *roe*, *ros* has no effect on CEO salary. State the alternative that better stock market performance increases a CEO's salary.

b) Esitmate the equation using the data in `ceosal1.gdt`. By what percent is salary predicted to increase, if *ros* increases by 50 points? Does *ros* have a *practically* large effect on salary?

c) Test the null hypothesis that *ros* has no effect on salary, against the alternative that *ros* has a positive effect. Carry out the test at the 10% significance level.

d) Would you include *ros* in a final model explaining CEO compensation in terms of firm performance? Explain.

Exercise 5.3 (*Individual vs. joint significance*). Using the data in `sleep.gdt`, estimate

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u$$

and report the results in equation form.

a) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? Show your work.

b) Drop *educ* and *age* from the equation and report the results in equation form. Are *educ* and *age* jointly significant in the original equation at the 5% level? Justify your answer.

c) Does including *educ* and *age* in the model greatly affect the estimated tradeoff between sleeping and working?

d) Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts *a* and *b*?

Exercise 5.4 (*Using confidence intervals to test hypotheses*). The variables in `GPA.gdt` include college grade point average (*colGPA*), high school GPA (*hsGPA*), achievement test score (*ACT*), and the average number of lectures missed per week (*skipped*) for a sample of 141 students from a large university; both college and high school GPAs are on a four-point scale. Estimate the following equation, which can be used to study the effects of skipping class on college GPA:

$$colGPA = \beta_0 + \beta_1 hsGPA + \beta_2 ACT + \beta_3 skipped + u.$$

a) Using the standard normal approximation, find the 95% confidence interval for *hsGPA*.

b) Can you reject the hypothesis H_0: $hsGPA = .4$ against the two-sided alternative at the 5% level?

c) Can you reject the hypothesis H_0: $hsGPA = 1$ against the two-sided alternative at the 5% level?

Exercise 5.5 (*Linear restrictions*). Use the data in `wages2.gdt` for this exercise.

a) Consider the standard wage equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

State the null hypothesis that another year of general workforce experience has the same effect on $\log(wage)$ as another year of tenure with the current employer.

b) Test the null hypothesis in part *a* against a two-sided alternative, at the 5% significance level, by constructing a 95% confidence interval. What do you conclude?