Tutorials Week 1

Position of Empirical Economics in the curriculum

Period 1	Period 2		Period 3, 4
Empirical Economics	Research Project		Thesis
	B&F	Fintech	
	B&SI	Frontiers of Business and Social Impact	
	BD&E	Frontiers of Entrepreneurship	
	EP	Policy Evaluation Skills	
	IM	Frontiers of International Management	
	FM	Next Generation Finance	
	SC&R	SC&R	
	SF&I	Sustainability Risk	

Some practical information:

In Blackboard, you can find under **Course Content**, the following:

- The datasets we use for Blackboard
- There is a PDF with the compilation of the tutorial exercises
- Stata commands
- Papers used for the datasets
- Tutorials for Stata
- Statistical tables
- doc.files
- In case you need more support in Econometrics, the slides of Dr. Anna Salomon for Bachelor are helpful,
- For support in statistics, the slides of Dr. Adriaan Kalwij from the Bachelor course are also helpful.

Regression Analysis: a recapitulation of Econometrics and Statistics in Bachelor

Empirical Economics

Period 1

Pdf file on Blackboard	Dataset on Blackboard	Papers related to the datasets	Description
C 3.4	attend.dta	Leslie Papke(2005): The Effects of Spending on Test Pass Rates: Evidence from Michigan, Journal of Public Economics 89, 821-839	OLS mechanics to write estimated models. Interpretation of the coefficients $\beta_0; \beta_1;\ and\ \beta_2$. (further explanation in the book, page 199). Basic Stata commands. Data structure, variables, interpretation regression parameters (check lecture – unit 1 page 44)
C 4.10	elem94_95	Leslie Papke(2005): The Effects of Spending on Test Pass Rates: Evidence from Michigan, Journal of Public Economics 89, 821-839	interpretation of coefficients, log variables, changes in standard error, t-test, and rejection areas. Write an economic conclusion.
6.3	wage2.dta	Blackburn McK. and Neumark, D. (1992): Unobserved ability, efficiency wages, and interindustry wage differentials. The Quarterly Journal of Economics,	Marginal effects, use, and meaning of interaction terms: meaning and how to generate them in Stata. Assess the statistical significance of the interaction term and compare the coefficient of determination with and without the interaction term. (further explanation on the book page 218)
7.14	sleep75.dta	J.E. Biddle and D.S. Hamermesh (1990): Sleep and the Allocation of Time, Journal of Political Economy 98, 922-943.	use of dummy variables, interaction terms, F-test
C 8.1	sleep75.dta	J.E. Biddle and D.S. Hamermesh (1990): Sleep and the Allocation of Time; Journal of Political Economy 98, 922-943.	heteroskedasticity, robust standard errors, the variance of the error term, heteroskedasticity testing (Breush-Pagan). See Chapter 8.3

C 3.4. Use the data in ATTEND.RAW for this exercise.

i) Obtain the minimum, maximum, and average values for the variables atndrte, priGPA, and ACT

Variables: atndrte: percent classes attended ; priGPA: cumulative GPA prior to term; ACT: ACT score

```
. use "U:\Stata\Empirical Economics Data Sets\Week 1\ATTEND.DTA"

. sum atndrte priGPA ACT

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     atndrte |        680    81.70956    17.04699       6.25        100
      priGPA |        680    2.586775    .5447141       .857       3.93
         ACT |        680    22.51029    3.490768         13         32
```

ii) Estimate the model (basic OSL):

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u$$

```
reg atndrte priGPA ACT

      Source |       SS           df       MS      Number of obs   =       680
-------------+----------------------------------   F(2, 677)       =    138.65
       Model |  57336.7612         2  28668.3806   Prob > F        =    0.0000
    Residual |  139980.564       677  206.765974   R-squared       =    0.2906
-------------+----------------------------------   Adj R-squared   =    0.2885
       Total |  197317.325       679   290.59989   Root MSE        =    14.379

------------------------------------------------------------------------------
     atndrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   17.26059   1.083103    15.94   0.000     15.13395    19.38724
         ACT |  -1.716553    .169012   -10.16   0.000    -2.048404   -1.384702
       _cons |    75.7004   3.884108    19.49   0.000     68.07406    83.32675
------------------------------------------------------------------------------
```

Write the results in equation form.

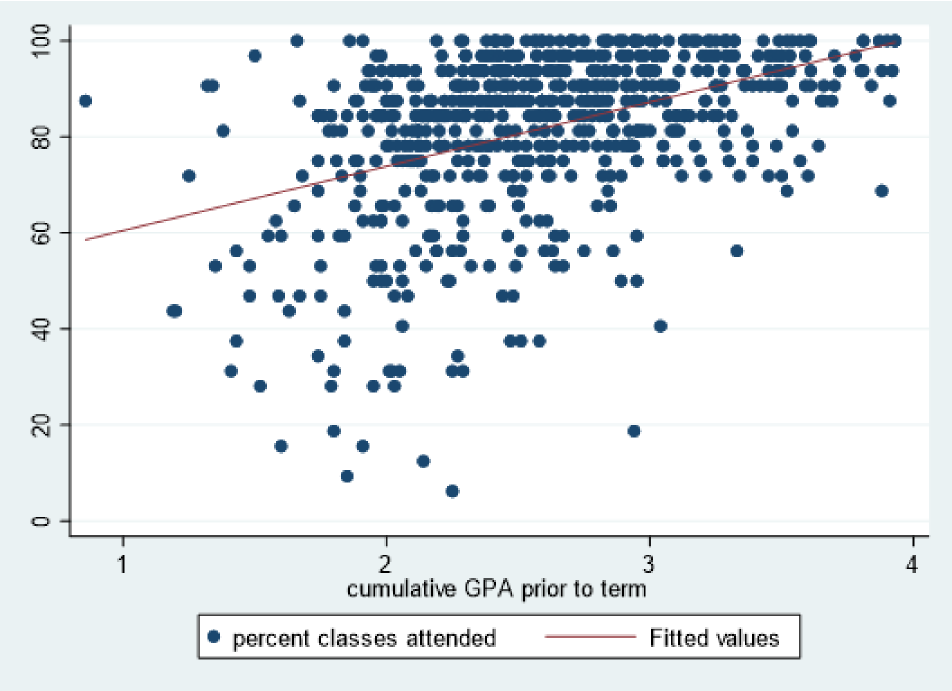
$$\widehat{atndrte} = \underset{(19.49)}{75.70} + \underset{(15.94)}{17.26} priGPA - \underset{(-10.16)}{1.72} ACT$$

Utrecht University

```
. reg atndrte priGPA ACT

      Source |       SS           df       MS      Number of obs   =       680
-------------+----------------------------------   F(2, 677)       =    138.65
       Model |  57336.7612         2  28668.3806   Prob > F        =    0.0000
    Residual |  139980.564       677  206.765974   R-squared       =    0.2906
-------------+----------------------------------   Adj R-squared   =    0.2885
       Total |  197317.325       679   290.59989   Root MSE        =    14.379

------------------------------------------------------------------------------
     atndrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   17.26059   1.083103    15.94   0.000     15.13395    19.38724
         ACT |  -1.716553    .169012   -10.16   0.000    -2.048404   -1.384702
       _cons |    75.7004   3.884108    19.49   0.000     68.07406    83.32675
------------------------------------------------------------------------------
```



Interpret the intercept. Does it have a useful meaning?

R-squared: 0.29

Students with 0 priGPA and 0 ACT would attend 75.7% of classes on average. Since this is unlikely to happen, the intercept does not have practical meaning.

iii) Discuss the estimated slope coefficients.


```
. reg atndrte priGPA ACT

      Source |       SS           df       MS      Number of obs   =       680
-------------+----------------------------------   F(2, 677)       =    138.65
       Model |  57336.7612         2  28668.3806   Prob > F        =    0.0000
    Residual |  139980.564       677  206.765974   R-squared       =    0.2906
-------------+----------------------------------   Adj R-squared   =    0.2885
       Total |  197317.325       679   290.59989   Root MSE        =    14.379

------------------------------------------------------------------------------
     atndrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   17.26059   1.083103    15.94   0.000     15.13395    19.38724
         ACT |  -1.716553    .169012   -10.16   0.000    -2.048404   -1.384702
       _cons |    75.7004   3.884108    19.49   0.000     68.07406    83.32675
------------------------------------------------------------------------------
```

atndrte: percent classes attended
priGPA: cumulative GPA prior to term
ACT: ACT score

$$atndrte = 75.70 + 17.26 priGPA - 1.72 ACT$$

Interpretation of β_i Level-level: $\Delta y = \beta_i \Delta x$

Interpretation:

- For every point increase in the prior GPA (priGPA), the attendance rate (atndrte)is expected to increase by 17,26 percentage points, holding all other variables constant.
- For every point increase in ACT score, the attendance rate (atndrte) is expected to decrease by 1.72 percentage points, holding all other variables constant.

What if two students have the same ACT?

- In that case, check the priGPA for those students. The one student with a higher priGPA will attend 17.26% more classes compared to the other.

iv) What is the predicted atndrte if priGPA=3.65 and ACT = 20?

```
The predicted attendance rate atndrte when priGPA=3.65 and ACT=20 equals 104.3705
. display _b[_cons] + _b[priGPA]*3.65 + _b[ACT]*20
104.3705
```

```
What you do in the exam, you just replace:
atndrte = 75.70+17.26(3.65)-1.72(20)=104.40%
```

What do you make of this result? Are there any students in the sample with these values of the explanatory variables?

This result is not possible. We can not predict that the attendance will increase in 104.40%. A student with a priGPA of 3.65 and ACT of 20 will for sure attend to 100% the class but not to 104.40%. It exceeds predictions

Utrecht University

v) If student A has priGPA = 3.1 and ACT = 21 and student B has priGPA = 2.1 and ACT = 26, what is the predicted difference in their attendance rates?

$$\widehat{atndrte} = 75.70 + 17.26priGPA - 1.72ACT$$

- **Student A:**

$$atndrte = 75.70 + 17.26 * (3.1) - 1.72(21)$$

$$atndrte = 75.70 + 53.51 - 36.12$$

$$\widehat{atndrte} = 93.09$$

- **Student B:**

$$atndrte = 75.70 + 17.26 * (2.1) - 1.72(26)$$

$$atndrte = 75.70 + 36.25 - 44.72$$

$$\widehat{atndrte} = 67.23$$

Predicted difference in their attendance rates= 93.09 - 67.23= 25.86

STATA Commands:

```
. di _b[_cons] + _b[priGPA]*3.1 + _b[ACT]*21
93.160625

. di _b[_cons] + _b[priGPA]*2.1 + _b[ACT]*26
67.31727

. di _b[priGPA]*(3.1-2.1) + _b[ACT]*(21-26) "%"
25.843356%
```

C.4.10 Use the data in ELEM94_95 to answer this question.

The dependent variable *lavgsal* is the log of average teacher salary and *bs* is the ratio of average benefits to average salary (by school).

i) Run the simple regression of *lavgsal* on *bs*.

```
reg lavgsal bs

      Source |       SS           df       MS      Number of obs   =     1,848
-------------+----------------------------------   F(1, 1846)      =     28.23
       Model |   1.5088834         1   1.5088834   Prob > F        =    0.0000
    Residual |  98.6724955     1,846  .053452056   R-squared       =    0.0151
-------------+----------------------------------   Adj R-squared   =    0.0145
       Total |  100.181379     1,847  .054240054   Root MSE        =     .2312

------------------------------------------------------------------------------
     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bs |  -.7951249   .1496545    -5.31   0.000    -1.088635   -.501615
       _cons |    10.7479   .0516622   208.04   0.000     10.64657    10.84922
------------------------------------------------------------------------------
```

Variables:

- bs: avgben/avgsal
- lavgsal: log(avgsal)

Is the estimated slope statistically different from zero?

- Ho: β_1 statistically significant $= 0$.
- H1: β_1 statistically significant $\neq 0$ → (two-sided test : Table G.2)
- If |t-value| > t critical value, then reject Ho.
- 5.31 > 1.96, which means that in absolute terms the t-statistics is greater than the c.v. at 5% significance level. Therefore, we reject Ho.
- We accept H1: β_1 is statistically significant different from 0.

Utrecht University

Is bs statistically different from -1?

- We need to test this:

```
. test bs=-1

 ( 1)  bs = -1

       F(  1,  1846) =    1.87
            Prob > F =    0.1712
```

Write all statistical steps for an F-test:

- $Ho = \beta_1 = -1$
- $H1 = \beta_1 \neq -1$ → Table G.3a); b) for F-Distribution
- If F value > F critical value, then reject Ho.
- If 1.87 > 2.71 (at 10% level; 0.10 **--> Table 3a**) or if 1.87> 3.84 (at 5% level, 0.05 → Table G.3b), then reject Ho.
- In both cases, it is not the case that the F-values are greater than the F-critical values. Hence, **we can not reject Ho.**
- We accept Ho: *bs* is **not** statistically different from -1.

Check table G.3.a); b) for the F-Distribution

For the exam, you have to write all the steps for a T-test or F-test:

Formulate the Ho, t statistics, and F-statistics, write the critical values at the significance level that is asked (mostly 5%), formulate the rejection area, reject or fail to reject the Ho, and write the interpretation of the coefficient and the economic meaning.

The interpretation of the regression parameter (coefficient) for bs: It is a log-level because log(y) and x.
$\%\Delta y = (100\beta_i) * \Delta x$

Calculate the percentage change: $(e^{coefficient} - 1) * 100$=
$(e^{-0.7951249t} - 1) * 100 = -54.8475$

A one-unit increase in the ratio of average benefits to average salary by school *bs*, suggests approximately a 54.85% decrease in the average salary, holding other variables constant.

Economic interpretation: This finding might suggest that schools with higher benefit ratios relative to salaries may offer lower average salaries for teachers, potentially reflecting budgetary priorities or compensation structures within those schools.

Utrecht University

ii) Add the variable *lenrol* and *lstaff* to the regression from part (i). What happens to the coefficient on *bs*?

12-9-2024

```
ii) Add the variables lenrol and lstaff to the regression from part i)

. reg lavgsal bs lenrol lstaff

      Source |       SS           df       MS      Number of obs   =     1,848
-------------+----------------------------------   F(3, 1844)      =    572.03
       Model |  48.2908776         3  16.0969592   Prob > F        =    0.0000
    Residual |  51.8905013     1,844  .028140185   R-squared       =    0.4820
-------------+----------------------------------   Adj R-squared   =    0.4812
       Total |  100.181379     1,847  .054240054   Root MSE        =    .16775

------------------------------------------------------------------------------
     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bs |  -.6050611   .1087429    -5.56   0.000    -.8183333   -.3917889
      lenrol |  -.0315853   .0084769    -3.73   0.000    -.0482106    -.01496
      lstaff |  -.7137195   .0177902   -40.12   0.000    -.7486105   -.6788285
       _cons |   13.95305   .1072337   130.12   0.000     13.74274    14.16336
------------------------------------------------------------------------------
```

To compare against i)

```
i) run a simple regression of lavgsal on bs

reg lavgsal bs

      Source |       SS           df       MS      Number of obs   =     1,848
-------------+----------------------------------   F(1, 1846)      =     28.23
       Model |  1.5088834         1   1.5088834   Prob > F        =    0.0000
    Residual |  98.6724955     1,846  .053452056   R-squared       =    0.0151
-------------+----------------------------------   Adj R-squared   =    0.0145
       Total |  100.181379     1,847  .054240054   Root MSE        =     .2312

------------------------------------------------------------------------------
     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bs |  -.7951249   .1496545    -5.31   0.000    -1.088635    -.501615
       _cons |    10.7479   .0516622   208.04   0.000     10.64657    10.84922
------------------------------------------------------------------------------
```

What happens to the coefficient on *bs*?

- It increased.
- Now it is -0.61 against -0.79. You can also calculate the percentage change for the average salaries (~-45.39%) This implies that for each one-unit increase in the ratio of average benefits to average salary –bs-, the average teachers' salary is expected to decrease by about 45.39%, holding other factors constant. This suggests a negative relationship between the benefits-to-salary ratio and teacher salaries, but a lower decrease than in i).
- This is normal when adding more variables and when controlling others.

Additional questions:

What can you say about the R-squared in both regressions and what do the standard errors say? What about the SSR?

ii) How does the situation compare with that in table 4.1?

- This is a smaller sampler. n=408
- Compared to Table 4.1. *bs* now increased. Now it is -0.589
- (It went from -0.825 to -0.605, when adding more independent variables.)

TABLE 4.1 Testing the Salary-Benefits Tradeoff

	Dependent Variable: log(*salary*)		
Independent Variables	**(1)**	**(2)**	**(3)**
b/s	−.825 (.200)	−.605 (.165)	−.589 (.165)
log(*enroll*)	——	.0874 (.0073)	.0881 (.0073)
log(*staff*)	——	−.222 (.050)	−.218 (.050)
droprate	——	——	−.00028 (.00161)
gradrate	——	——	.00097 .00066)
intercept	10.523 (0.042)	10.884 (0.252)	10.738 (0.258)
Observations *R*-squared	408 .040	408 .353	408 .361

- Is bs statistically different from zero? Yes, it is.
- Is it statistically different from -1?

```
. test bs=-1

 ( 1)  bs = -1

       F(  1,  1844) =   13.19
            Prob > F =    0.0003
```

- Ho: β_1 = -1
- H1: $\beta_1: \neq -1$ → (this is a double-sized test)
- Fvalue > Fcv, then reject Ho
- 13.9 > 2.71 (at 0.10) or > 3.84 (0.05)
- We reject Ho at the 10% and 5% significance levels level. *bs* is statistically different from -1.
- The coefficients of the other variables are different from Table 4.1. They might have been taken from another example.

Check table G.3.b for the F critical value

iii) How come the standard error on the BS coefficient is smaller in part (ii) than in part i)?

The s.e. in i) is 0.200. It is larger than the ii) and iii) (each 0.165) because there are fewer controlled variables in i). In ii) and iii) we included more regressors. This can be observed with the variance of the OLS regressors.

What happens to the error variance versus multicollinearity when *lenrol* and *lstaff* are added?

$$Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{(1-R_j^2)\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{\sigma}^2}{(1-R_j^2)SST_j}$$

$se(\hat{\beta}_j) = \sqrt{(Var(\hat{\beta}_j)}$: standard error of $\hat{\beta}_j$

- R^2 can also increase (it increased in ii) >0) and leads to lowering the standard error.
- If R^2 comes close to 1, this can lead to a multicollinearity problem (*a good fit of the model represents high variability of the explanatory and dependent variables).*
- If SST is small, multicollinearity can arise. Larger n are mainly related to higher SST.
- If there is multicollinearity, the standard error will increase substantially.

iv) How come the coefficient of lstaff is negative? Is it large in magnitude?

```
. reg lavgsal bs lenrol lstaff

      Source |       SS           df       MS      Number of obs   =     1,848
-------------+----------------------------------   F(3, 1844)      =    572.03
       Model |  48.2908776         3  16.0969592   Prob > F        =    0.0000
    Residual |  51.8905013     1,844  .028140185   R-squared       =    0.4820
-------------+----------------------------------   Adj R-squared   =    0.4812
       Total |  100.181379     1,847  .054240054   Root MSE        =    .16775

------------------------------------------------------------------------------
     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bs |  -.6050611   .1087429    -5.56   0.000    -.8183333   -.3917889
      lenrol |  -.0315853   .0084769    -3.73   0.000    -.0482106     -.01496
      lstaff |  -.7137195   .0177902   -40.12   0.000    -.7486105   -.6788285
       _cons |   13.95305   .1072337   130.12   0.000     13.74274    14.16336
------------------------------------------------------------------------------
```

This is a log-log level: $\%\Delta y = \beta_j \%\Delta x$

A 1% increase in the staff number, is associated with approximately a 0.71% decrease in the average salary while keeping the other variable fixed. Yes, it is large in magnitude.

v) Now add the variable *lunch* to the regression. Holding other factors fixed, are teachers being compensated for teaching students from disadvantaged backgrounds? Explain.

```
. reg lavgsal bs lenrol lstaff lunch

      Source |       SS       df       MS              Number of obs =    1848
-------------+------------------------------           F(  4,  1843) =  439.43
       Model |  48.904075     4  12.2260187           Prob > F      =  0.0000
    Residual | 51.2773039  1843  .027822737           R-squared     =  0.4882
-------------+------------------------------           Adj R-squared =  0.4870
       Total | 100.181379  1847  .054240054           Root MSE      =   .1668

------------------------------------------------------------------------------
     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bs |   -.516129   .1097747    -4.70   0.000    -.7314248   -.3008332
      lenrol |  -.0284092    .008456    -3.36   0.001    -.0449936   -.0118247
      lstaff |  -.6906322   .0183604   -37.62   0.000    -.7266416   -.6546228
       lunch |  -.0007581   .0001615    -4.69   0.000    -.0010747   -.0004414
       _cons |   13.83149   .1097259   126.06   0.000     13.61629    14.04669
------------------------------------------------------------------------------
```

- The number of lunches offered in a school increases because the number of more disadvantaged children in classes increases, too.
- It is a log-level: $\%\Delta y = (100\beta_i)\Delta x$: the %change of average salaries is -0.07581%, because -0.0007581*100%
- For each additional lunch provided, the average salary of teachers is expected to decrease by about 0.07581%.
- Teachers are not being compensated for teaching students from disadvantaged backgrounds.
- The school would need to reconsider how to allocate resources better, so the trade-off does not affect teacher.

Utrecht University

vi) Overall, is the pattern of results that you find with ELEM94_95.RAW consistent with the pattern in table 4.1?


```
reg lavgsal bs

      Source |       SS           df       MS      Number of obs   =     1,848
-------------+----------------------------------   F(1, 1846)      =     28.23
       Model |   1.5088834         1   1.5088834   Prob > F        =    0.0000
    Residual |  98.6724955     1,846  .053452056   R-squared       =    0.0151
-------------+----------------------------------   Adj R-squared   =    0.0145
       Total |  100.181379     1,847  .054240054   Root MSE        =     .2312

------------------------------------------------------------------------------
     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bs |  -.7951249   .1496545    -5.31   0.000    -1.088635    -.501615
       _cons |    10.7479   .0516622   208.04   0.000     10.64657    10.84922
------------------------------------------------------------------------------
```

```
. reg lavgsal bs lenrol lstaff

      Source |       SS           df       MS      Number of obs   =     1,848
                                                   F(3, 1844)      =    572.03
       Model |  48.2908776         3  16.0969592   Prob > F        =    0.0000
    Residual |  51.8905013     1,844  .028140185   R-squared       =    0.4820
                                                   Adj R-squared   =    0.4812
       Total |  100.181379     1,847  .054240054   Root MSE        =    .16775

     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

          bs |  -.6050611   .1087429    -5.56   0.000    -.8183333   -.3917889
      lenrol |  -.0315853   .0084769    -3.73   0.000    -.0482106     -.01496
      lstaff |  -.7137195   .0177902   -40.12   0.000    -.7486105   -.6788285
       _cons |   13.95305   .1072337   130.12   0.000     13.74274    14.16336
```

```
. reg lavgsal bs lenrol lstaff lunch

      Source |       SS           df       MS      Number of obs   =     1,848
                                                   F(4, 1843)      =    439.43
       Model |   48.904075         4  12.2260187   Prob > F        =    0.0000
    Residual |  51.2773039     1,843  .027822737   R-squared       =    0.4882
                                                   Adj R-squared   =    0.4870
       Total |  100.181379     1,847  .054240054   Root MSE        =     .1668

     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

          bs |   -.516129   .1097747    -4.70   0.000    -.7314248   -.3008332
      lenrol |  -.0284092    .008456    -3.36   0.001    -.0449936   -.0118247
      lstaff |  -.6906322   .0183604   -37.62   0.000    -.7266416   -.6546228
       lunch |  -.0007581   .0001615    -4.69   0.000    -.0010747   -.0004414
       _cons |   13.83149   .1097259   126.06   0.000     13.61629    14.04669
```

TABLE 4.1 Testing the Salary-Benefits Tradeoff

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
b/s	−.825 (.200)	−.605 (.165)	−.589 (.165)
log(*enroll*)	——	.0874 (.0073)	.0881 (.0073)
log(*staff*)	——	−.222 (.050)	−.218 (.050)
droprate	——	——	−.00028 (.00161)
gradrate	——	——	.00097 .00066)
intercept	10.523 (0.042)	10.884 (0.252)	10.738 (0.258)
Observations	408	408	408
R-squared	.040	.353	.361

vi) Overall, is the pattern of results that you find with ELEM94_95.RAW consistent with the pattern in table 4.1?

- Not all the coefficients have the same sign; hence, there is a qualitative difference between the regression results and Table 4.1.

- In Table 4.1. we find that with bs and *log staff* fixed, additional students would increase the average salary of the teaching staff. (the logenroll is positive in ii) and iii))

- Compared to the regressions, we found that we have lower wages on average in schools with more enrollments, holding other factors fixed (the logenroll is negative in the regressions).

- The difference in the sample in the regressions versus Table 4.1. can be due to differences in the sample. The characteristics of the schools or teachers in the two samples may differ significantly, which can lead to different relationships between enrollment and teacher salaries. Also, selection bias: smaller samples can lead to specific patterns due to the different types of schools included, e.g., schools with different demographic levels are overrepresented in one sample; this can affect the sign and magnitude of the coefficient.

- Measurement error: if the data quality is poorer due to e.g. missings in the data, can also lead to different estimates.

6.3) The following model allows the return to education to depend upon the total amount of both parent's education, called *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ * pareduc + \beta_3 exper + \beta_4 tenure + u$$

To keep in mind:

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

- For the interaction term, we have to generate two new variables. The first is *pareduc* and the other, the own education under the influence of the *pareduc*, that means: *educ*pareduc*

```
gen pareduc = meduc+feduc
(213 missing values generated)

Then generate an interaction term = educ*pareduc

gen educ_pareduc=educ*pareduc
(213 missing values generated)
```

i) Show that, in decimal form, the return to another year of education in this model is:

$$\Delta \log(wage)/\Delta educ = \beta_1 educ + \beta_2 pareduc$$

What sign do you expect for $\boldsymbol{\beta_2}$? Why?

- We expect the sign to be positive. If parents' education is higher (pareduc), the coefficient for return to education will also increase.
- The interaction term "*educ_pareduc*" is positive since we believe that it is more likely that children of better-educated parents tend to have more access to studies and become more productive workers.

```
. reg lwage educ educ_pareduc exper tenure

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(4, 717)       =     36.44
       Model |  21.4253649         4  5.35634121   Prob > F        =    0.0000
    Residual |  105.386551       717  .146982637   R-squared       =    0.1690
-------------+----------------------------------   Adj R-squared   =    0.1643
       Total |  126.811916       721  .175883378   Root MSE        =    .38338

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0467522   .0104767     4.46   0.000     .0261835     .067321
educ_pareduc |    .000775   .0002107     3.68   0.000     .0003612    .0011887
       exper |    .018871   .0039429     4.79   0.000     .0111299     .026612
      tenure |   .0102166   .0029938     3.41   0.001     .0043391    .0160942
       _cons |   5.646519   .1295593    43.58   0.000     5.392158     5.90088
------------------------------------------------------------------------------
```

ii) Using the data in WAGE2.RAW, the estimated equation is:

$$\widehat{\log(wage)} = \underset{(0.13)}{5.65} + \underset{(0.010)}{0.047}educ + \underset{(0.00021)}{0.00078}educ * pareduc + \underset{(0.004)}{0.019}exper + \underset{(0.003)}{0.010}tenure$$

$n = 722, R^2 = 0.169$ (Only 722 observations contain full information on parents' education.)

Interpret the coefficient on the interaction term. It might help to choose two specific values for pareduc – for example, pareduc = 32 if both parents have a college education, or pareduc = 24 if both parents have a high school education – and to compare the estimated return to educ:

The return to another year of education:

$$\Delta \log(wage)/\Delta educ = \beta_1 \cancel{educ} + \beta_2 pareduc$$

for example, pareduc = 32 if both parents have a college education

= 0.0468 + 0.000775(32)= 0.072

for example, pareduc = 24 if both parents have a high school education

= 0.0468 + 0.000775(24)= 0.065

The rate of returns is 6.5% and 7.2%, respectively.

Variables: wage: monthly earnings; educ: years of education; exper: years of work experience; tenure: years with current employer

ii) Using the data in WAGE2.RAW, the estimated equation is:

$$\widehat{\log(wage)} = \underset{(0.13)}{5.65} + \underset{(0.010)}{0.047}educ + \underset{(0.00021)}{0.00078}educ * pareduc + \underset{(0.004)}{0.019}exper + \underset{(0.003)}{0.010}tenure$$

Calculate this with Stata:

```
nlcom _b[educ] + _b[educ_pareduc]*32

       _nl_1:  _b[educ] + _b[educ_pareduc]*32

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .0715513   .0072101     9.92   0.000     .0574198    .0856828
------------------------------------------------------------------------------
```

```
nlcom _b[educ] + _b[educ_pareduc]*24

       _nl_1:  _b[educ] + _b[educ_pareduc]*24

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .0653515   .0076094     8.59   0.000     .0504374    .0802656
------------------------------------------------------------------------------
```

Interpretation:

If you come from parents who have a college education, your wage will increase by 7.2%. If you come from parents who only have a high school education, the rate of return on education is 6.5%.

iii) When *pareduc* is added as a separate variable to the equation, we get:

$$\log(\widehat{wage}) = \underset{(0.38)}{4.94} + \underset{(0.027)}{0.097}educ + +\underset{(0.017)}{0.033}\,pareduc \;-\underset{(0.0012)}{0.0016}educ * pareduc + \underset{(0.004)}{0.020}exper + \underset{(0.003)}{0.010}tenure$$

$$n = 722, R^2 = 0.1735$$

```
. reg lwage educ pareduc educ_pareduc exper tenure

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(5, 716)       =     30.07
       Model |  22.0046475         5  4.4009295   Prob > F        =    0.0000
    Residual |  104.807268       716  .146378866   R-squared       =    0.1735
-------------+----------------------------------   Adj R-squared   =    0.1678
       Total |  126.811916       721  .175883378   Root MSE        =    .38259

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0971133   .0273897     3.55   0.000     .0433397     .150887
     pareduc |   .0332144   .0166963     1.99   0.047     .0004348    .0659939
educ_pareduc |  -.0015683   .0011966    -1.31   0.190    -.0039175    .0007809
       exper |   .0195568   .0039499     4.95   0.000     .0118021    .0273116
      tenure |   .0103082    .002988     3.45   0.001      .004442    .0161744
       _cons |   4.937661   .3790621    13.03   0.000     4.193455    5.681867
------------------------------------------------------------------------------
```

Does the estimated return to education now depend positively on parent education?

- It does not depend positively. The return on education has a negative relationship to parental education.
- The interaction term is not significant anymore. This can have been caused by omitting the variable *pareduc in ii)* (this caused omitted variable bias).

Test the null hypothesis that the return to education does not depend on parent education.

- Test if pareduc has *an effect* on the rate of returns to education. That means, pareduc has a relationship to wage, and it can not be 0
  - Ho: $\beta_2 = 0$
  - H1: $\beta_2 \neq 0$ → (two-sided test : Table G.2)
  - If |t-stats| > tcritical value, then reject Ho.
  - 1.99 > 1.960, reject Ho at 5% significance level.
  - Pareduc has an effect on lwage.
  - Pareduc yields a positive coefficient, which is significant at 5%. So it has an effect on the average wage, but not via the rate of returns (not via education). The interaction term is not even significant at the 10% level against a two-sided alternative.

This exercise provides a good example of how omitting a level effect (*pareduc in this case) can lead to a* biased estimation of the interaction effect.

Additional task: quadratic function and new interaction term.

gen educ2=educ^2

```
reg lwage educ educ2 pareduc educ_pareduc exper tenure

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(6, 715)       =     25.51
       Model |  22.3579761         6  3.72632934   Prob > F        =    0.0000
    Residual |   104.45394       715  .146089426   R-squared       =    0.1763
-------------+----------------------------------   Adj R-squared   =    0.1694
       Total |  126.811916       721  .175883378   Root MSE        =    .38222

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .2329935   .0915571     2.54   0.011     .0532405    .4127464
       educ2 |   -.005388   .0034646    -1.56   0.120      -.01219    .0014139
     pareduc |   .0215393   .0182913     1.18   0.239    -.0143718    .0574505
educ_pareduc |   -.000748   .0013066    -0.57   0.567    -.0033132    .0018172
       exper |   .0200912   .0039609     5.07   0.000     .0123148    .0278676
      tenure |   .0105341   .0029885     3.52   0.000     .0046667    .0164014
       _cons |   4.111514    .652382     6.30   0.000     2.830701    5.392328
------------------------------------------------------------------------------
```

- The relationship between education and log wages is now non-linear because of the quadratic functional form and will depend on education and parental education.
- For this reason, we will estimate the effect of education for an individual with average education and average *pareduc.* In the next slide, we calculate the rate of return to education and parental education.

Calculate the rate of return again:

- To calculate the rate of return manually:
  = logwage = 4.11+0.23edu-0.0053educ^2 + 0.02153pareduc – 0.000748educ_pareduc …

```
The commands in Stata for the rate of returns:

 nlcom _b[educ] + 2*_b[educ2]*13.46845+ _b[educ_pareduc]*21.06094

       _nl_1:  _b[educ] + 2*_b[educ2]*13.46845+ _b[educ_pareduc]*21.06094

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .0721034   .0093799     7.69   0.000     .0537192    .0904877
------------------------------------------------------------------------------
```

- Even though the coefficients of educ2 and educ_pareduc were not significant statistically, the estimated rate of returns to education is significant at 1% and is 7.2% per school year at the mean.

- Important: when using quadratic forms, there is no linear relationship.

Hence the rate of returns to education changes with the educ. Since educ^2 has a negative coefficient, we find a decreasing rate of returns to education in line with theoretical expectations and many empirical studies.

Utrecht University

7.14 Use the data in SLEEP75.RAW for the exercise. The equation of interest is:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u$$

i) Estimate this equation separately for men and women and report the results in the usual form. Are there notable differences between the two estimated equations?

The variables are:

- The total number of working hours (totwrk) ,
- education level (educ),
- age and,
- the number of young kids (yngkid) that are younger than 3 years old and,
- the relationship to the total number of sleep hours (sleep) a man or woman gets.
- We will consider a dummy variable for this regression: Male = 1 ; Male = 0

- First we run the regression without considering the binary variable of male = 1

```
. reg sleep totwrk educ age yngkid

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(4, 701)       =     22.46
       Model |  15818709.4         4  3954677.35   Prob > F        =    0.0000
    Residual |   123421126       701  176064.374   R-squared       =    0.1136
-------------+----------------------------------   Adj R-squared   =    0.1085
       Total |   139239836       705  197503.313   Root MSE        =     419.6

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1483157   .0167037    -8.88   0.000    -.1811109   -.1155205
        educ |  -11.19025   5.889365    -1.90   0.058    -22.75315    .3726611
         age |   2.402862   1.518646     1.58   0.114    -.5787773    5.384502
      yngkid |   21.84079   49.75161     0.44   0.661    -75.83923    119.5208
       _cons |    3628.15   114.6692    31.64   0.000     3403.014    3853.286
------------------------------------------------------------------------------
```

- We need to generate age^2 and run the regression again:

$$\boxed{sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u}$$

```
. gen age2=age^2

. reg sleep totwrk educ age age2 yngkid

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(5, 700)       =     18.14
       Model |  15972384.7         5  3194476.94   Prob > F        =    0.0000
    Residual |   123267451       700  176096.359   R-squared       =    0.1147
-------------+----------------------------------   Adj R-squared   =    0.1084
       Total |   139239836       705  197503.313   Root MSE        =    419.64

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1460463   .0168809    -8.65   0.000    -.1791896   -.1129031
        educ |  -11.13772   5.890168    -1.89   0.059    -22.70223    .4267914
         age |  -8.123949   11.37049    -0.71   0.475     -30.4483     14.2004
        age2 |    .126287    .135186     0.93   0.351    -.1391317    .3917057
      yngkid |   17.15441   50.00839     0.34   0.732    -81.02999    115.3388
       _cons |   3825.375   240.2585    15.92   0.000     3353.661    4297.088
------------------------------------------------------------------------------
```

Utrecht University

Now the equations are separately estimated. For that, we need to consider the dummy variables. Male = 1; Male = 0

```
. 
reg sleep totwrk educ age age2 yngkid if male==1

      Source |       SS           df       MS      Number of obs   =       400
-------------+----------------------------------   F(5, 394)       =     14.59
       Model |  11806161.6         5  2361232.32   Prob > F        =    0.0000
    Residual |    63763979       394   161837.51   R-squared       =    0.1562
-------------+----------------------------------   Adj R-squared   =    0.1455
       Total |  75570140.6       399  189398.849   Root MSE        =    402.29

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1821232   .0244855    -7.44   0.000    -.2302618   -.1339846
        educ |  -13.05238   7.414218    -1.76   0.079    -27.62876    1.523996
         age |   7.156591   14.32037     0.50   0.618    -20.99731    35.31049
        age2 |  -.0447674   .1684053    -0.27   0.791    -.3758528    .2863181
      yngkid |   60.38021   59.02278     1.02   0.307    -55.65877    176.4192
       _cons |   3648.208   310.0393    11.77   0.000      3038.67    4257.747
------------------------------------------------------------------------------
```

```
reg sleep totwrk educ age age2 yngkid if male==0

      Source |       SS           df       MS      Number of obs   =       306
-------------+----------------------------------   F(5, 300)       =      6.50
       Model |  6201576.18         5  1240315.24   Prob > F        =    0.0000
    Residual |  57288575.9       300   190961.92   R-squared       =    0.0977
-------------+----------------------------------   Adj R-squared   =    0.0826
       Total |  63490152.1       305  208164.433   Root MSE        =    436.99

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1399495   .0276594    -5.06   0.000    -.1943806   -.0855184
        educ |  -10.20514   9.588848    -1.06   0.288    -29.07506    8.664787
         age |  -30.35657   18.53091    -1.64   0.102    -66.82361    6.110464
        age2 |   .3679406   .2233398     1.65   0.101    -.0715705    .8074516
      yngkid |  -118.2826   93.18757    -1.27   0.205    -301.6667    65.10154
       _cons |   4238.729   384.8923    11.01   0.000     3481.299     4996.16
------------------------------------------------------------------------------
```

- For males the *education* coefficient is stast. significant at 10% (0.079 < 0.10); for women, it is not significant at 10% (0.288 > 0.10).
- The relationship between *age* and weekly minutes sleep at night (*sleep*) is also different:
  - ✓ For males, it has an inverse U shape, while for women, it is U-shaped. The P-value is not statistically significant at the 10% level. We do not have enough evidence to reject the Ho at the 10% significance level.
- The coefficient of *yngkids* dummy (if there is at least one kid in the family younger than 3):
  - For men , it is positive: 60.38 and negative for women: -118.28. However, it is not statistically significant at 10%.

Utrecht University

(ii) Compute the Chow test for equality of the parameters in the sleep equation for men and women. Use the form of the test that adds *male* and the interaction terms *male*towrk*, …, *male*yngkid* and uses the full set of observations. What are the relevant *df* for the test? Should you reject the null at the 5% level?

a. What are the relevant *df* for the test?

b. Should you reject the null at the 5% level?

First, we have to generate interaction terms:

```
. gen totwrk_male=totwrk * male

. gen educ_male=educ*male

. gen age_male = age*male

. gen age2_male= age2*male

. gen yngkid_male = yngkid*male
```

Why do we generate interaction terms?

They generate a wider understanding of the variables in the model. More hypotheses can be tested about the relationship between the independent variables and the dependent ones.

Utrecht University

Then run the regression using the interaction terms:

```
reg sleep totwrk educ age age2 yngkid male totwrk_male educ_male age_male age2_male
yngkid_male
```

```
      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(11, 694)      =      9.48
       Model |  18187280.8        11  1653389.17   Prob > F        =    0.0000
    Residual |   121052555       694  174427.313   R-squared       =    0.1306
-------------+----------------------------------   Adj R-squared   =    0.1168
       Total |   139239836       705  197503.313   Root MSE        =    417.64

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1399495   .0264349    -5.29   0.000    -.1918514   -.0880476
        educ |  -10.20514   9.164321    -1.11   0.266    -28.19826    7.787983
         age |  -30.35657   17.71049    -1.71   0.087    -65.12914    4.415998
        age2 |   .3679406   .2134519     1.72   0.085    -.0511483    .7870294
      yngkid |  -118.2826   89.06187    -1.33   0.185    -293.1456    56.58047
        male |  -590.5211   488.7916    -1.21   0.227    -1550.209    369.1665
 totwrk_male |  -.0421737    .036674    -1.15   0.251     -.114179    .0298317
   educ_male |  -2.847243   11.96795    -0.24   0.812    -26.34497    20.65048
    age_male |   37.51316   23.12332     1.62   0.105    -7.886888    82.91321
   age2_male |  -.4127079   .2759136    -1.50   0.135    -.9544333    .1290175
 yngkid_male |   178.6628   108.1051     1.65   0.099     -33.5895     390.915
       _cons |   4238.729   367.8519    11.52   0.000     3516.493    4960.965
------------------------------------------------------------------------------
```

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + \beta_7 totwrk\_male + \beta_8 educ\_male + \beta_9 age\_male + \beta_{10} age2\_male + \beta_{11} yngkid\_male + u$$

Utrecht University

Now run the F-test on all coefficients that involve the male dummy (so consider the interaction terms and also the dummy male)

```
. test male totwrk_male educ_male age_male age2_male yngkid_male

 ( 1)  male = 0
 ( 2)  totwrk_male = 0
 ( 3)  educ_male = 0
 ( 4)  age_male = 0
 ( 5)  age2_male = 0
 ( 6)  yngkid_male = 0

       F(  6,   694) =    2.12
            Prob > F =    0.0495
```

$$H_0 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$$

$$H_1 = no\ true$$

$$If\ F_{value} > F_{CV}, then\ reject\ Ho$$

2.12 > 2.10, then reject Ho at 5% significant level.

We reject Ho at 5%;
Gender has an effect on the total amount of hours of sleep.

a. **What are the relevant *df* for the test?**

The relevant df for the tests are 6 and 694. 6 is the d.f for the numerator and 694 the d.f. for the denominator.

b. **Do you reject the null hypothesis at the 5% level?** Yes, we reject the Ho. Gender has an effect on sleep.

iii) Given the results from parts (ii) and (iii), what would be your final model?

It means: test the interaction terms for joint significance.

```
test totwrk_male educ_male age_male age2_male yngkid_male

 ( 1)  totwrk_male = 0
 ( 2)  educ_male = 0
 ( 3)  age_male = 0
 ( 4)  age2_male = 0
 ( 5)  yngkid_male = 0

       F(  5,   694) =    1.26
            Prob > F =    0.2814
```

$H_0 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$

$H_1\!: \; H_0 \, is \; not \; true \; or \; at \; least \; one \; is \; different \; than \; 0$

$If \;\; F_{value} > F_{CV}, then \; reject \; Ho$

$1.26 < 2.21, we \; do \; not \; reject \; Ho \; at \; 5\% \; signficance \; level.$

$The \; data \; suggests \; that \; being \; male \; does \; not \; have \; an \; effect \; on \; sleep.$

The final model: $sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u$

Unbiasedness of OLS: assumptions

OLS – the ordinary least squares – delivers unbiased estimator parameters $\boldsymbol{\beta_k}$ if the following assumptions hold:

1. Population model is linear in parameters (and the terror term is additive)
2. Error term has a zero population mean : $E(\varepsilon_i) = 0$
3. All independent variables are uncorrelated with the error term $Corr\ (\varepsilon_i, X_i) = 0$
4. No perfect (multi)collinearity between independent variables.

OLS is unbiased estimator of Var ($\widehat{\boldsymbol{\beta}_k}$) if assumptions 1-4 hold, as well as the following assumptions:

5. No serial correlation: $Corr\ \left(\varepsilon_i, \varepsilon_j\right) = 0$
6. No heteroskedasticity (homoskedasticity) : the variance of the error term is constant. $Var(\varepsilon_i) = \sigma^2$ where σ^2 is constant.

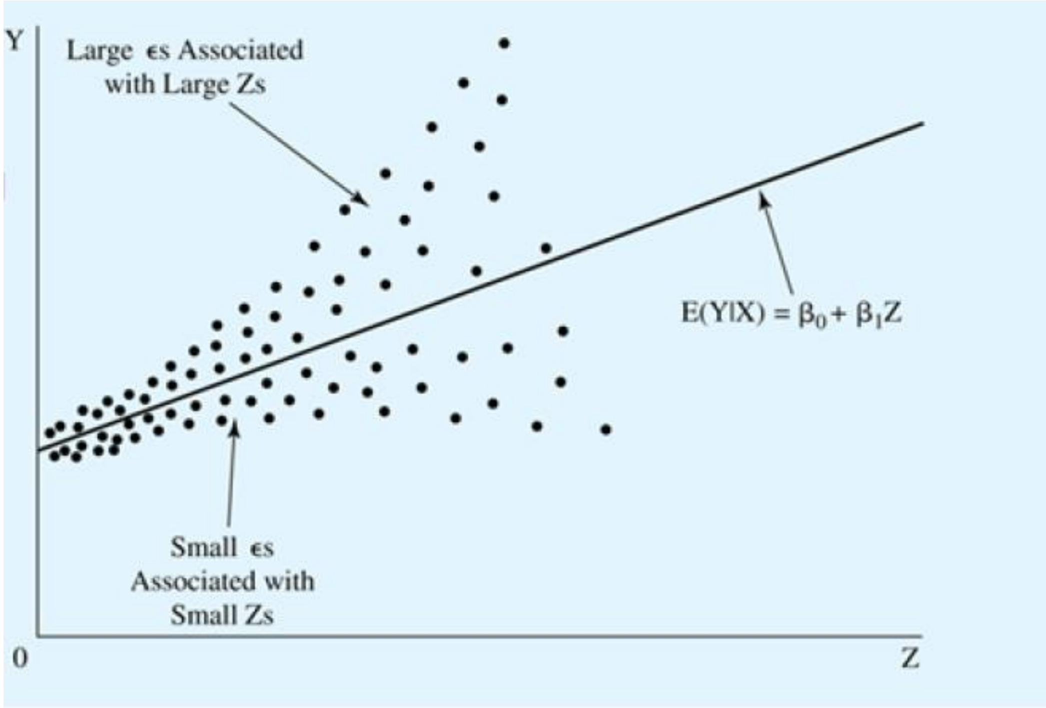
If all assumptions hold, then:

- Perform hypothesis tests about a single population parameter using the t-test
- Perform hypothesis tests about multiple population parameters using the F-test

Heteroskedasticity

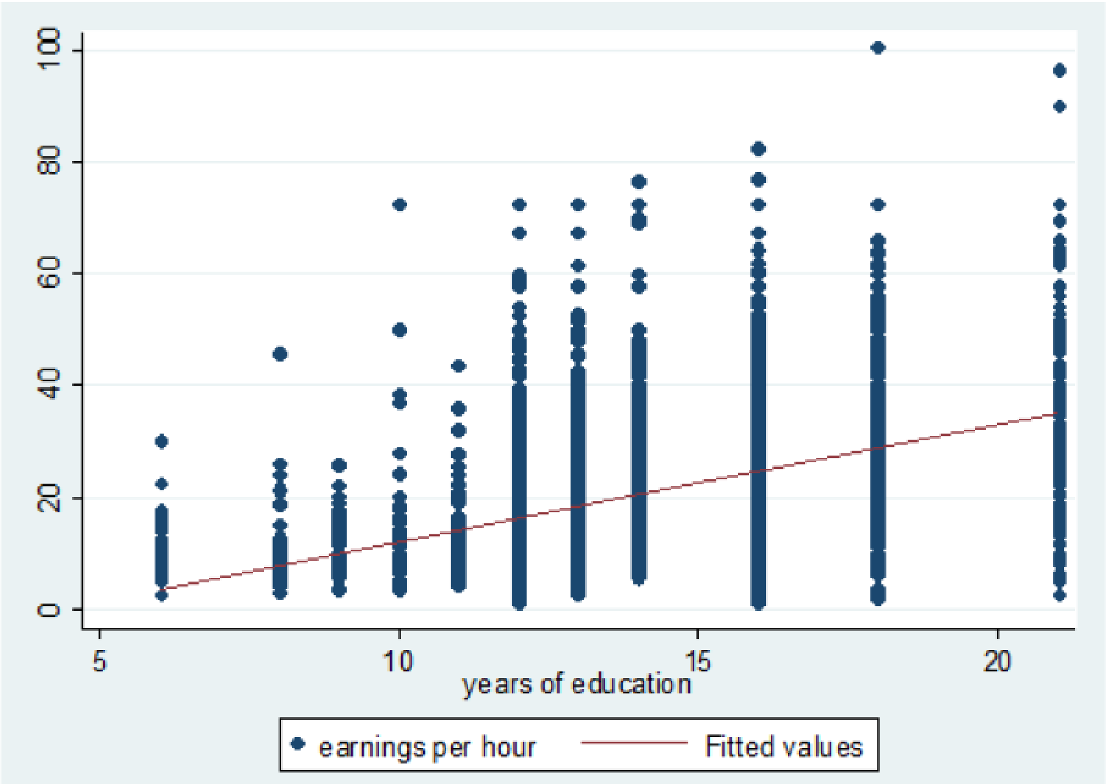
12-9-2024

Heteroskedasticity: error term does not have a constant variance $Var\ (\varepsilon_i) \neq \sigma^2$



An Error Term Whose Variance Increases as Z Increases (Heteroskedasticity)

Heteroskedasticity: violation of assumption 6



Consequences of Heteroskedasticity

- If assumptions 1-4 are not violated, then OLS is still an unbiased estimator of β_k
- But, since Var $(\widehat{\beta_k})$ depends on σ^2, it is a biased estimator of Var $(\widehat{\beta_k})$
- T-statistics are incorrect since these depend on σ^2
- F-statistics are incorrect since these depend on σ^2
- If t- and F-statistics are incorrect, we can not perform hypothesis tests.
- Without hypothesis tests, we can not perform inference about the population from a sample, which is the aim of applied econometric analysis.
- Therefore, we need to know how to diagnose heteroskedasticity (apply the Breusch Pagan test) and then solve the problem if we find any.

Breusch-Pagan Test

Steps for diagnosis for heteroskedasticity

1. Estimate the model: $Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 x_{2i} + \varepsilon_i$
2. Predict residuals: e_i from the estimated model: $Y_i = \hat{\beta}_o + \hat{\beta}_1 X_{1i} + \hat{\beta} x_{2i} + e_i$
3. Square the residuals: e_i^2
4. Regress squared residuals e_i^2 on independent variables from the original model: $e_i^2 = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + v_i$
5. Test for joint significance of the independent variables on e_i^2: if they do, then the Ho is rejected and heteroskedasticy exists.

$H_o\!: \ \delta_1 = \delta_2 = 0$ Homoskedasticity

$H_1\!: H_0 \ not \ true$ Heteroskedasticity

Remember

If there is no heteroskedasticity, then the error term does not have a constant variance.

$Var\ (\varepsilon_i) = \sigma^2 (\ where\ \sigma^2$ is a constant)

Utrecht University

Example for Heteroskedasticity

Step 1, 2 and 3:

- **Estimate the model**
- **Predict the residual**
- **Square the residual:** e_i^2

We want to examine the **relationship between economic development**, measured as log gdp per capita, **workers' education level,** and **entrepreneurship** (measured as the fraction of the working population in self-employment).

Because we expect entrepreneurship to be nonlinearly related to development, we estimate the following model:

$$\ln gdp_i = \beta_0 + \beta_1 educ_i + \beta_2 selfemp_i + \beta_3 selfemp_i^2 + \varepsilon_i$$

Estimates of the model:

```
. reg lnreggdp yearsed self_emp self_emp2

      Source |       SS           df       MS      Number of obs   =       547
-------------+----------------------------------   F(  3,    543)  =    750.25
       Model |  600.742478         3  200.247493   Prob > F        =    0.0000
    Residual |  144.930118       543  .266906294   R-squared       =    0.8056
-------------+----------------------------------   Adj R-squared   =    0.8046
       Total |  745.672595       546  1.36570072   Root MSE        =    .51663

------------------------------------------------------------------------------
    lnreggdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsed |   .3447389   .0093812    36.75   0.000      .326311    .3631669
    self_emp |   .0235793   .0038133     6.18   0.000     .0160886    .0310699
   self_emp2 |  -.0005308    .000059    -9.00   0.000    -.0006467   -.0004149
       _cons |   6.495673   .0993675    65.37   0.000     6.300481    6.690865
------------------------------------------------------------------------------

. predict uhat, resid

. gen uhat2=uhat^2
```

We want to test for heteroskedasticity, so we **predict the residuals** (e_i), and then **obtain the squared residuals** (e_i^2).

Steps 4 and 5

Regress squared residuals e_i^2 on independent variables from the original model:
Test for joint significance of the independent variables on e_i^2: if they do, then the Ho is rejected and heteroskedasticy exists:

We now **regress the squared residuals onto the explanatory variables from the original model**:

```
. reg uhat2 yearsed self_emp self_emp2

      Source |       SS       df       MS              Number of obs =     547
-------------+------------------------------           F(  3,   543) =    6.49
       Model |   3.9418781     3  1.31395937           Prob > F      =  0.0003
    Residual |  109.888897   543  .202373659           R-squared     =  0.0346
-------------+------------------------------           Adj R-squared =  0.0293
       Total |  113.830775   546  .208481273           Root MSE      =  .44986

------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsed |  -.0325959   .0081688    -3.99   0.000    -.0486422   -.0165496
    self_emp |  -.0018068   .0033205    -0.54   0.587    -.0083294    .0047157
   self_emp2 |   .0000143   .0000514     0.28   0.781    -.0000866    .0001152
       _cons |   .5186509   .0865251     5.99   0.000     .3486859    .6886158
------------------------------------------------------------------------------
```

The **explanatory variables are jointly significant**, as seen from the model F-test (p-value=0.0003<0.05). This means we **reject the null hypothesis of homoskedasticity: the errors are heteroskedastic**!

$$H_0 \ : \ \delta_1 = \delta_2 = 0 \qquad \textit{(homoskedasticity)}$$
$$H_A \ : \ H_0 \text{ not true} \qquad \textit{(heteroskedasticity)}$$

- The solution for heteroskedasticity **does not require changing the estimates** $\widehat{\beta}_k$ (since OLS is still an unbiased estimator of β_k).
- However, we do **need new standard errors** since the $\widehat{Var}(\widehat{\beta}_k)$ are incorrect.
- We calculate the **heteroskedasticity-robust standard error** in Stata.
- Caveat: this robust standard error is **only valid in large samples**!

Heteroskedasticity - Solution

Heteroskedasticity-robust standard errors in Stata

```
. reg lnreggdp yearsed self_emp self_emp2, robust

Linear regression                               Number of obs =        547
                                                F(  3,    543) =   1081.16
                                                Prob > F       =    0.0000
                                                R-squared      =    0.8056
                                                Root MSE       =    .51663

------------------------------------------------------------------------------
             |               Robust
    lnreggdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsed |   .3447389   .0115737    29.79   0.000     .3220042    .3674736
    self_emp |   .0235793   .0047807     4.93   0.000     .0141884    .0329702
   self_emp2 |  -.0005308    .000067    -7.92   0.000    -.0006624   -.0003992
       _cons |   6.495673   .1396055    46.53   0.000      6.22144    6.769906
------------------------------------------------------------------------------
```

Heteroskedasticity-robust standard errors can obtained easily by typing *,robust* at the end of the *reg* command.

Comparing regular and robust standard errors

- **Robust standard errors are typically higher** than the regular ones- although they may also be lower.
- Higher standard errors means the t-statistics become smaller (in absolute value), and **estimates become less significant**.
- **In our example, the standard errors increase somewhat**, but all coefficients are still individually significant.

Heteroskedasticity - Summary

- **Problem** = heteroskedastic errors
- **Consequence** = coefficient estimates $\widehat{\beta}$ remain unbiased (since OLS assumptions 1-4 have not been violated), but the variance estimates $\widehat{Var}(\widehat{\beta})$ (and hence also the std errors $\sqrt{\widehat{Var}(\widehat{\beta})}$) are biased (since OLS assumption 6 has been violated). This means we cannot perform hypothesis tests (t- or F-tests).
- **Diagnosis** = Breusch-Pagan test, which involves regressing the squared residuals on all explanatory variables (there is heteroskedasticity if the p-value for the model F-test is smaller than the chosen significance level).
- **Solution** = estimate the equation with heteroskedasticity-robust standard errors (Stata command *reg y x1 x2, robust*)

C.8.1 Consider the following model to explain sleeping behavior:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u$$

i) Write down a model that allows the variance of u to differ between men and women. The variance should not depend on other factors.

To solve this exercise, we need some knowledge about the **variance of the error term**:

Homoskedasticity:
- The variance of the error term is constant: $Var(\varepsilon_i) = \sigma^2$ where σ^2 is constant.
- The variance of the error term does not depend on the explanatory variables $Var(u \mid x_1, \dots x_k) = \sigma^2$

What this exercise asks us is to specify a simple linear model for the conditional variance of the error term:

$$Var(u \mid male) = \gamma_0 + \gamma_1 male_1$$

Since we do not know u but only the residuals, the empirical model is:

$$\widehat{u_i^2} = \hat{\gamma}_0 + \hat{\gamma}_1 male_i + e_i$$

Where e is an error term since we can not explain all the residual variance.

Utrecht University

ii) Use the data in SLEEP75.RAW to estimate the parameters of the model for heteroskedasticity. (You have to estimate the sleep equation by OLS, first, to obtain the OLS residuals). Is the estimated variance of u higher for men or for women?

- First, we need to generate the squared for age.
- Then run the regression,
- Then we need to predict the residuals

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u$$

```
gen age2 = age^2

. reg sleep totwrk educ age age2 yngkid male

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(6, 699)       =     16.30
       Model |  17092058.6         6  2848676.43   Prob > F        =    0.0000
    Residual |   122147777       699  174746.462   R-squared       =    0.1228
-------------+----------------------------------   Adj R-squared   =    0.1152
       Total |   139239836       705  197503.313   Root MSE        =    418.03

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1634235   .0181634    -9.00   0.000    -.1990848   -.1277622
        educ |  -11.71327   5.871952    -1.99   0.046    -23.24205   -.1844947
         age |  -8.697402   11.32909    -0.77   0.443    -30.94053    13.54572
        age2 |   .1284415   .1346696     0.95   0.341    -.1359638    .3928469
      yngkid |  -.0228006   50.27641    -0.00   1.000    -98.73367    98.68807
        male |   87.75455   34.66794     2.53   0.012     19.68877    155.8203
       _cons |   3840.852   239.4139    16.04   0.000     3370.795    4310.909
------------------------------------------------------------------------------
```

```
 predict u, res

. gen u2=u^2

reg u2 male

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(1, 704)       =      1.12
       Model |  1.4430e+11         1  1.4430e+11   Prob > F        =    0.2909
    Residual |  9.0942e+13       704  1.2918e+11   R-squared       =    0.0016
-------------+----------------------------------   Adj R-squared   =    0.0002
       Total |  9.1086e+13       705  1.2920e+11   Root MSE        =   3.6e+05

------------------------------------------------------------------------------
          u2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -28849.63   27296.51    -1.06   0.291    -82441.94    24742.69
       _cons |   189359.2   20546.36     9.22   0.000     149019.8    229698.7
------------------------------------------------------------------------------
```

Because the coefficient for male is negative, the estimated variance is higher for women.

iii) Is the variance of the u statistically different from men and for women?

- No, because the Pvalue is 0.291 > 0.10.
- The t-statistics on male is only -1.06, which is not significant at even the 20% level against a two-sided alternative. (See Table G.2)
- Note that this is not the official Breusch Pagan test that tests whether heteroskedasticity exists in the complete empirical model.
- Here, we are only concerned with the issue of whether the variance of u differs between men and women.
- We could not argue that the error variance differs by gender, and we do not have to use “robust” standard errors.

Additional material: Test for heteroskedasticity: Breusch-Pagan Test


```
reg u2 totwrk educ age age2 yngkid male

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(6, 699)       =      1.85
       Model |  1.4229e+12         6  2.3715e+11   Prob > F        =    0.0872
    Residual |  8.9663e+13       699  1.2827e+11   R-squared       =    0.0156
-------------+----------------------------------   Adj R-squared   =    0.0072
       Total |  9.1086e+13       705  1.2920e+11   Root MSE        =    3.6e+05

------------------------------------------------------------------------------
          u2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |   18.45399   15.56184     1.19   0.236    -12.09955    49.00754
        educ |  -10181.35   5030.915    -2.02   0.043    -20058.86   -303.8341
         age |  -9019.993    9706.43    -0.93   0.353    -28077.24    10037.26
        age2 |   72.79544   115.3809     0.63   0.528    -153.7392    299.3301
      yngkid |   5100.806   43075.34     0.12   0.906    -79471.74    89673.36
        male |   -37435.2   29702.47    -1.26   0.208    -95751.94    20881.54
       _cons |   515599.8   205122.8     2.51   0.012     112869.2    918330.4
------------------------------------------------------------------------------
```

```
. hettest, rhs fstat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: totwrk educ age age2 yngkid male

         F(6 , 699)    =      1.42
         Prob > F      =    0.2037
```

Ho: $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$ (homoskedasticity)

H1: Ho is not true (heteroskedasticity)

If Ftest > Fcv, then reject Ho.

1.42 < 2.10, we can not reject Ho, at 5% significance level.

Homoskedasticity holds.

If heteroskedasticity holds, estimate the equation with the robust standard error.


```
. reg sleep totwrk educ age age2 yngkid male, rob

Linear regression                               Number of obs     =        706
                                                F(6, 699)         =      14.29
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1228
                                                Root MSE          =     418.03

------------------------------------------------------------------------------
             |               Robust
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1634235    .020683    -7.90   0.000    -.2040317   -.1228154
        educ |  -11.71327   5.747549    -2.04   0.042     -22.9978   -.4287441
         age |  -8.697402   11.78685    -0.74   0.461    -31.83928    14.44447
        age2 |   .1284415   .1360228     0.94   0.345    -.1386206    .3955036
      yngkid |  -.0228006   53.90532    -0.00   1.000    -105.8585    105.8129
        male |   87.75455   35.54252     2.47   0.014     17.97166    157.5374
       _cons |   3840.852   259.1258    14.82   0.000     3332.094     4349.61
------------------------------------------------------------------------------
```

However, the findings do not change relative to the version without robust standard errors.

Utrecht University

Sharing science, *shaping tomorrow*