1. Theory

Question 1

a) Given by question:

- True model: $y_i^* = x_i'\beta + \epsilon_i^*$ with $E(\epsilon_i^*|x_i) = 0$ and $V(\epsilon_i^*|x_i) = \sigma_*^2$
- Estimated model: $y_i = x_i'\beta + \epsilon_i$
- Measurement error in y_i: $y_i = y_i^* + \eta_i$ with $\eta_i \sim (0, \sigma_\eta^2)$
- Further assumption: It is a "classical measurement error", i.e. it is uncorrelated with everything: $E(\eta_i|x_i) = 0$ and $E(\eta_i|\epsilon_i^*) = 0$

One can transform the true model as following (by knowing that $y_i = y_i^* + \eta_i$ holds):

$$y_i^* = x_i'\beta + \epsilon_i^*$$

Adding on both sides η_i:

$$y_i^* + \eta_i = x_i'\beta + \epsilon_i^* + \eta_i$$
$$y_i = x_i'\beta + \epsilon_i^* + \eta_i$$
$$y_i = x_i'\beta + \epsilon_i$$

Where ϵ_i is the composite error term of the true error (ϵ_i^*) and the measurement error (η_i).

<u>Mean:</u>

The mean of ϵ_i is:

$$E(\epsilon_i|x_i) = E(\epsilon_i^* + \eta_i) = E(\epsilon_i^*) + E(\eta_i)$$

From the (conditional-)mean-zero-error assumption one knows that $E(\epsilon_i^*|x_i) = 0$ holds and therefore:

$$E(\epsilon_i^*) = E_{x_i}E(\epsilon_i^*|x_i) = E_{x_i} \cdot 0 = 0$$

And from $\eta_i \sim (0, \sigma_\eta^2)$, it follows that:

$$E(\eta_i) = 0$$

Tarik Benli 15-719-818, Ramon Gmür 16-705-220

And so, one can say that the mean of the error ϵ_i is zero:

$$\boldsymbol{E(\epsilon_i|x_i)} = E(\epsilon_i^* + \eta_i) = E(\epsilon_i^*) + E(\eta_i) = 0 + 0 = \mathbf{0}$$

Variance:

The variance of ϵ_i is:

$$V(\epsilon_i) = V(\epsilon_i^* + \eta_i) = V(\epsilon_i^*) + V(\eta_i) + 2Cov(\epsilon_i^*, \eta_i)$$
$$= E\left(\left(\epsilon_i^* - \mu_{\epsilon_i^*}\right)^2\right) + E\left(\left(\eta_i - \mu_{\eta_i}\right)^2\right) + 2Cov(\epsilon_i^*, \eta_i)$$

As shown before, the means of $\epsilon_i^*, \mu_{\epsilon_i^*}$ and η_i, μ_{η_i} are zero and from the assumption $E(\eta_i|\epsilon_i^*) = 0$, one can say that $Cov(\epsilon_i^*, \eta_i) = 0$ must hold.

Therefore:

$$\boldsymbol{V(\epsilon_i)} = E((\epsilon_i^* - 0)^2) + E((\eta_i - 0)^2) + 2 \cdot 0 = E((\epsilon_i^*)^2) + E((\eta_i)^2) = \boldsymbol{\sigma^2_{\epsilon_i^*} + \sigma^2_{\eta_i}}$$

b)

$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$ (Estimator of β) is unbiased if $E(\hat{\beta}) = \beta$.

➔ $E(\hat{\beta}) = E(\beta + (X'X)^{-1}X'\epsilon) = E(\beta) + E((X'X)^{-1}X'\epsilon)) = \beta + E((X'X)^{-1}X'\epsilon))$

As shown before: $E(\epsilon_i|x_i) = E(\epsilon_i^* + \eta_i) = E(\epsilon_i^*) + E(\eta_i) = 0$

Therefore:

$$E(\hat{\beta}) = \beta + E((X'X)^{-1}X'0)) = \beta + 0 = \beta$$

One can say that $\hat{\beta}$ is unbiased.

c)

The variance of $\hat{\beta}$ without the measurement error is:

$$V(\hat{\beta}) = V(\epsilon^*|X)(X'X)^{-1} = \sigma_*^2(X'X)^{-1}$$

Where σ_*^2 refers to the variance of the error term of the true model (ϵ^*). And so, this is equal to $\sigma_{\epsilon_i^*}^2$. The variance of the error term of the model with the measurement error is: As shown already in a:

$$\sigma^2 = V(\epsilon) = \sigma_{\epsilon_i^*}^2 + \sigma_{\eta_i}^2 = \sigma_*^2 + \sigma_{\eta_i}^2$$

And therefore:

$$V(\hat{\beta}) = V(\epsilon|X)(X'X)^{-1} = (\sigma_*^2 + \sigma_{\eta_i}^2)(X'X)^{-1}$$

So, one can say that the variance of the estimator $\hat{\beta}$ is larger for the model with the measurement error in y than without.

d)

They would be right by saying it is not a big deal. A measurement error in the dependent variable y does not bias the estimator $\hat{\beta}$ but increases its variance ($V(\hat{\beta})$) as shown in c) and therefor a measurement error in the dependent variable is indeed not a big deal. But a measurement error in an independent variable biases the estimator and therefor would end up in a big deal.

2. Empirical Question

Question 1 – **IV Regression**

a) See results for columns (2) and (3) below:

VARIABLES	(2) OLS	(3) IV
highqua	0.0768***	0.0874***
	(0.0106)	(0.0166)
age	0.0778***	0.0765***
	(0.0214)	(0.0215)
agesq	-0.000968***	-0.000943***
	(0.000266)	(0.000268)
Constant	-0.428	-0.568
	(0.435)	(0.467)
Observations	428	428
R-squared	0.149	0.147

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

i) The OLS-results in column (2) show the exact same results as in BCHHS (the small discrepancy is due to rounding). The IV-results in column (3) show slightly different results than in BCHHS. Their coefficient for Education is 0.002 smaller and for Age^2 the difference is 0.1 (after rounding and multiplying Age^2 by 100). Very small discrepancies can also be found in the standard errors. However, the discrepancies are always minuscule and do not change the significance, so we consider them not serious.

ii) In this regression, looking at the constant makes no sense. On one hand, looking at constants when log-variable are involved is pointless because these regressions focus on percental changes. On the other hand (and even with normal earnings as a variable), the constant would report the (log) average earnings with 0 years of education, which is not relevant (the minimum years of education is 10).

iii) Assuming our IV-results are consistent, we interpret Education as follows: In the OLS-regression, our results show an increase in earnings of 7.7% with one year of additional education, controlling for age-effects. In the IV-regressions, our results show an 8.7% increase in earnings with one year of additional education, controlling for age-effects, which suggests a negative bias.

b) Biases

i) There are three ways in which we think education could be endogenous:

- OVB: Ability is definitively correlated with education and earnings, but it is not included in the model. We therefore have an OVB, which causes years of schooling to be endogenous.
- Simultaneous causality: Earnings may increase education in the form of "adult education". For example: With higher earnings, the chance of being able to afford education next to (or to some part instead of) working is higher. It also could be that only wealthy families can afford a higher education (private schools) for their children. The causality between education and earnings runs in both ways, which causes years of schooling to be endogenous.
- Measurement error: People might overstate their years of education.

ii) Bias signs:

- OVB: Ability has definitively a positive effect on earnings and the covariance between ability and education is positive as well (people with high abilities go to school longer, see universities). Therefore, the likely sign of this bias is <u>positive</u>.
- Simultaneous causality: If we consider education in this dataset only to include education during childhood and young adulthood, we can assume that there is no simultaneous causality and the likely sign of this bias is <u>0</u>.
- Measurement error: This would be a case of the attenuation bias (bias towards 0). With people systematically overstating their education and a positive effect of education on earnings, the likely sign of this bias is <u>negative</u>.

iii) Relevance and exogeneity (without simultaneous causality):

- OVB: The instrument of is definitively correlated with the true years of education. However, assuming that the twin's report is also correlated with the omitted ability, this does not resolve the endogeneity issue.
- Measurement error: The instrument is definitively correlated with the true years of education and therefore relevant. Furthermore, we suggest that this instrument "corrects" some part of the overstated education.

iv) The difference from the IV-results to the OLS-results is -1%, therefore we have a negative bias. Even though this is suggested by bias through measurement error, we suspected the positive OVB to weigh stronger than the measurement error bias.

v) See the results from the first stage below:

VARIABLES	(1) 1S
twihigh	0.631***
	(0.0371)
age	0.0531
	(0.0756)
agesq	-0.000930
	(0.000938)
Constant	4.835***
	(1.535)
Observations	428
R-squared	0.446

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The F-value is 290.17 (roughly the t-value squared). This is much higher than 10, so we can consider this instrument not to be weak.

vi) We have a strong first stage, significant effect and good (well evaluated) assumptions. Therefore, we believe those results.

c) See results for columns (4) and (5) below:

VARIABLES	(4) OLS	(5) IV
dhigh	0.0394*	0.0774**
	(0.0226)	(0.0331)
Observations	214	214
R-squared	0.014	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

i) Twins have the same age (difference is always 0). The difference in age drops out.

ii) See the results in the table below:

VARIABLES	(4) OLS	(5) IV
dhigh	0.0392*	0.0778**
	(0.0226)	(0.0330)
Constant	0.0142	0.0126
	(0.0471)	(0.0474)
Observations	214	214
R-squared	0.014	0.000

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Including a constant makes no sense because earnings is again in logs. Additionally, a constant shows the average difference in earnings for 0 years difference in education, which can be expected to 0. The results in this table support the exclusion of the constant (the standard errors for example are 4 times larger than the constant itself).

iii) Now the effect of a one-year difference in education (which is the same as one additional year of education) only leads to a 3.9% difference in earnings. The advantage of this procedure is that the OVB disappears. Assuming both twins have the same abilities (genetics and family), the difference between them is 0 and this model only shows the difference in log-earnings on the difference in education and the difference in the error terms. This is also consistent with the assumed positive OVB from b), since the OLS-results in c) are smaller. The standard errors also increase due to the smaller sample size. However, there is still a measurement error problem.

iv) Now the effect of a one-year difference in education (which is the same as one additional year of education) only leads to a 7.8% difference in earnings, which is rather close to the results in b). The advantage of this instrument is the following: For twins with fitting differences in years of education and years of one's twin reported years of education, their difference in earnings correlate perfectly. The ones from the twins who do not fit, do not correlate perfectly. With an IV-regression, we still allow for this type of variation, but without overestimating due to twins who misreport their education. Comparing column (5) to (4), the increasing results suggest measurement error in (4). Also, the standard errors increase.

v) We do believe these results. As in b) the results from the OLS to the IV regression shift according to our priors. The assumptions seem believable. The increased standard errors can be explained with the smaller sample size, therefore the decreased significance compared to b) does not worry us.

d) See the results in the table below:

VARIABLES	(4) OLS	(5) IV
dhigh	0.0283	0.0358
	(0.0189)	(0.0272)
Observations	210	210
R-squared	0.011	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

i) Both the magnitude and the significance in column (5) are heavily affected by dropping the outliers. The effect is halved, but the standard error decrease by only a little bit. This leads to non-significant results.

ii) Now the results in (5) can be interpreted as follows: A one-year difference in education leads to a 3.8% difference in earnings.

iii) On one hand, these decrease in the results suggest that we are overestimating the effect from education on earnings with those outliers included. However, with the standard errors nearly staying the same, we suggest that excluding these 4 observations is not a good idea.

3. Log-file

See attachment

```
                                                       ___  ____  ____  ____  ____(R)
                                                      /__    /   ____/   /   ____/
                                                     ___/   /   /___/   /   /___/
                                                       Statistics/Data Analysis

      name:  <unnamed>
       log:  C:\Users\ramon\Desktop\UZH\Empirical Methods\Problem Sets\Problem Set 3\Stata\log_gm
  log type:  smcl
 opened on:   2 Dec 2019, 16:01:51

  1 .
  2 . insheet using "C:\Users\ramon\Desktop\UZH\Empirical Methods\Problem Sets\Problem Set 3\Stata\BC
    (18 vars, 428 obs)

  3 .
  4 . *1)
  5 .
  6 . gen lnearn = log(earning)

  7 . gen agesq = age^2

  8 .
  9 .
 10 . **a)
 11 .
 12 . reg lnearn highqua age agesq

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(3, 424)       =     24.72
       Model |  20.7258534         3   6.9086178   Prob > F        =    0.0000
    Residual |  118.492426       424  .279463268   R-squared       =    0.1489
-------------+----------------------------------   Adj R-squared   =    0.1429
       Total |  139.218279       427  .326038124   Root MSE        =    .52864

------------------------------------------------------------------------------
      lnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     highqua |   .0767543   .0105917     7.25   0.000     .0559355    .0975731
         age |   .0778154   .0213949     3.64   0.000     .0357622    .1198687
       agesq |  -.0009675   .0002658    -3.64   0.000    -.0014899   -.0004451
       _cons |  -.4282208   .4347756    -0.98   0.325    -1.282805    .4263631
------------------------------------------------------------------------------

 13 . outreg2 using "regressiona.doc", replace ctitle(OLS)
    regressiona.doc
    dir : seeout

 14 . ivreg lnearn age agesq (highqua = twihigh)

    Instrumental variables (2SLS) regression

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(3, 424)       =     16.40
       Model |  20.4445064         3  6.81483547   Prob > F        =    0.0000
    Residual |  118.773773       424  .280126822   R-squared       =    0.1469
-------------+----------------------------------   Adj R-squared   =    0.1408
       Total |  139.218279       427  .326038124   Root MSE        =    .52927

------------------------------------------------------------------------------
      lnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     highqua |   .0873817   .0166363     5.25   0.000     .0546818    .1200815
         age |   .0764781   .0214809     3.56   0.000     .0342558    .1187005
       agesq |  -.0009428   .0002677    -3.52   0.000    -.0014691   -.0004165
       _cons |  -.5684209   .4669861    -1.22   0.224    -1.486317    .3494751
------------------------------------------------------------------------------
Instrumented:  highqua
Instruments:   age agesq twihigh
------------------------------------------------------------------------------
```

```
15 . outreg2 using "regressiona.doc", append ctitle(IV)
   regressiona.doc
   dir : seeout

16 .
17 . **b)
18 .
19 . ***v)
20 . reg highqua twihigh age agesq

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(3, 424)       =    113.80
       Model |  1190.87218         3  396.957394   Prob > F        =    0.0000
    Residual |   1478.9666       424  3.48812878   R-squared       =    0.4460
-------------+----------------------------------   Adj R-squared   =    0.4421
       Total |  2669.83879       427  6.25254985   Root MSE        =    1.8677

------------------------------------------------------------------------------
     highqua |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     twihigh |   .6312721   .0370589    17.03   0.000     .5584302    .7041141
         age |   .0531199   .0755603     0.70   0.482    -.0953996    .2016394
       agesq |  -.0009302   .0009385    -0.99   0.322    -.0027749    .0009144
       _cons |    4.83493   1.535057     3.15   0.002     1.817661    7.852198
------------------------------------------------------------------------------

21 . outreg2 using "regressionb.doc", replace ctitle(1S)
   regressionb.doc
   dir : seeout

22 . test _b[twihigh]=0

    ( 1)  twihigh = 0

          F(  1,   424) =  290.17
               Prob > F =    0.0000

23 .
24 . **c)
25 .
26 . drop schyear lnandse part full self married own_exp bweight exp_par parted sm16 sm18

27 . reshape wide lnearn highqua twihigh earning, i(family) j(twinno)
   (note: j = 1 2)

   Data                               long   ->   wide
   -----------------------------------------------------------------------------
   Number of obs.                      428   ->     214
   Number of variables                   8   ->      11
   j variable (2 values)            twinno   ->   (dropped)
   xij variables:
                                    lnearn   ->   lnearn1 lnearn2
                                   highqua   ->   highqua1 highqua2
                                   twihigh   ->   twihigh1 twihigh2
                                   earning   ->   earning1 earning2
   -----------------------------------------------------------------------------

28 .
29 . gen dlnearn = lnearn1 - lnearn2

30 . gen dhigh = highqua1 - highqua2
```

```
31 . gen dtwihigh = twihigh1 - twihigh2

32 .
33 . gen dearn = earning1 - earning2

34 . *This one is for d)
35 .
36 . reg dlnearn dhigh, nocons

      Source |       SS           df       MS      Number of obs   =       214
-------------+----------------------------------   F(1, 213)       =      3.04
       Model |  1.43564569         1  1.43564569   Prob > F        =    0.0826
    Residual |   100.55228       213  .472076434   R-squared       =    0.0141
-------------+----------------------------------   Adj R-squared   =    0.0094
       Total |  101.987926       214  .476579094   Root MSE        =    .68708

------------------------------------------------------------------------------
     dlnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dhigh |   .0393535   .0225666     1.74   0.083    -.0051289     .083836
------------------------------------------------------------------------------

37 . outreg2 using "regressionc.doc", replace ctitle(OLS)
   regressionc.doc
   dir : seeout

38 . ivreg dlnearn (dhigh = dtwihigh), nocons

   Instrumental variables (2SLS) regression

      Source |       SS           df       MS      Number of obs   =       214
-------------+----------------------------------   F(1, 213)       =         .
       Model |  .096383507         1  .096383507   Prob > F        =         .
    Residual |  101.891543       213   .47836405   R-squared       =         .
-------------+----------------------------------   Adj R-squared   =         .
       Total |  101.987926       214  .476579094   Root MSE        =    .69164

------------------------------------------------------------------------------
     dlnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dhigh |   .0773631   .0330598     2.34   0.020     .0121968    .1425294
------------------------------------------------------------------------------
Instrumented:  dhigh
Instruments:   dtwihigh
------------------------------------------------------------------------------

39 . outreg2 using "regressionc.doc", append ctitle(IV)
   regressionc.doc
   dir : seeout

40 .
41 . ***ii)
42 .
43 . reg dlnearn dhigh

      Source |       SS           df       MS      Number of obs   =       214
-------------+----------------------------------   F(1, 212)       =      3.01
       Model |   1.4249932         1   1.4249932   Prob > F        =    0.0844
    Residual |  100.508895       212   .47409856   R-squared       =    0.0140
-------------+----------------------------------   Adj R-squared   =    0.0093
       Total |  101.933888       213  .478562854   Root MSE        =    .68855

------------------------------------------------------------------------------
     dlnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dhigh |   .0392153   .0226195     1.73   0.084    -.0053727    .0838032
       _cons |   .0142415   .0470778     0.30   0.763    -.0785591     .107042
------------------------------------------------------------------------------
```

```
44 . outreg2 using "regressionc2.doc", replace ctitle(OLS)
   regressionc2.doc
   dir : seeout

45 . ivreg dlnearn (dhigh = dtwihigh)

   Instrumental variables (2SLS) regression

         Source |       SS           df       MS      Number of obs   =       214
   -------------+----------------------------------   F(1, 212)       =      5.54
          Model |  .04558248         1  .04558248   Prob > F        =    0.0195
       Residual |  101.888306       212  .480605215   R-squared       =    0.0004
   -------------+----------------------------------   Adj R-squared   =   -0.0043
          Total |  101.933888       213  .478562854   Root MSE        =    .69326

   ------------------------------------------------------------------------------
        dlnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   -------------+----------------------------------------------------------------
          dhigh |   .0777982   .0330489     2.35   0.019     .0126517    .1429448
          _cons |   .0126188   .0474104     0.27   0.790    -.0808375    .1060751
   ------------------------------------------------------------------------------
   Instrumented:  dhigh
   Instruments:   dtwihigh
   ------------------------------------------------------------------------------

46 . outreg2 using "regressionc2.doc", append ctitle(IV)
   regressionc2.doc
   dir : seeout

47 .
48 . **d)
49 .
50 . gen absearn = abs(dearn)

51 . preserve

52 . drop if absearn > 60
   (4 observations deleted)

53 .
54 . reg dlnearn dhigh, nocons

         Source |       SS           df       MS      Number of obs   =       210
   -------------+----------------------------------   F(1, 209)       =      2.24
          Model |  .736676732         1  .736676732   Prob > F        =    0.1361
       Residual |  68.7836569       209  .329108406   R-squared       =    0.0106
   -------------+----------------------------------   Adj R-squared   =    0.0059
          Total |  69.5203336       210  .331049208   Root MSE        =    .57368

   ------------------------------------------------------------------------------
        dlnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   -------------+----------------------------------------------------------------
          dhigh |   .0282666   .0188931     1.50   0.136     -.008979    .0655121
   ------------------------------------------------------------------------------

55 . outreg2 using "regressiond.doc", replace ctitle(OLS)
   regressiond.doc
   dir : seeout

56 . ivreg dlnearn (dhigh = dtwihigh), nocons

   Instrumental variables (2SLS) regression

         Source |       SS           df       MS      Number of obs   =       210
   -------------+----------------------------------   F(1, 209)       =         .
          Model |  .684658057         1  .684658057   Prob > F        =         .
       Residual |  68.8356756       209  .329357299   R-squared       =         .
   -------------+----------------------------------   Adj R-squared   =         .
          Total |  69.5203336       210  .331049208   Root MSE        =     .5739
```

dlnearn	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dhigh	**.0357778**	**.0272474**	**1.31**	**0.191**	**-.0179371**	**.0894928**

```
Instrumented:  dhigh
Instruments:   dtwihigh
```

```
57 . outreg2 using "regressiond.doc", append ctitle(IV)
   regressiond.doc
   dir : seeout

58 .
59 . restore

60 .
   end of do-file
```