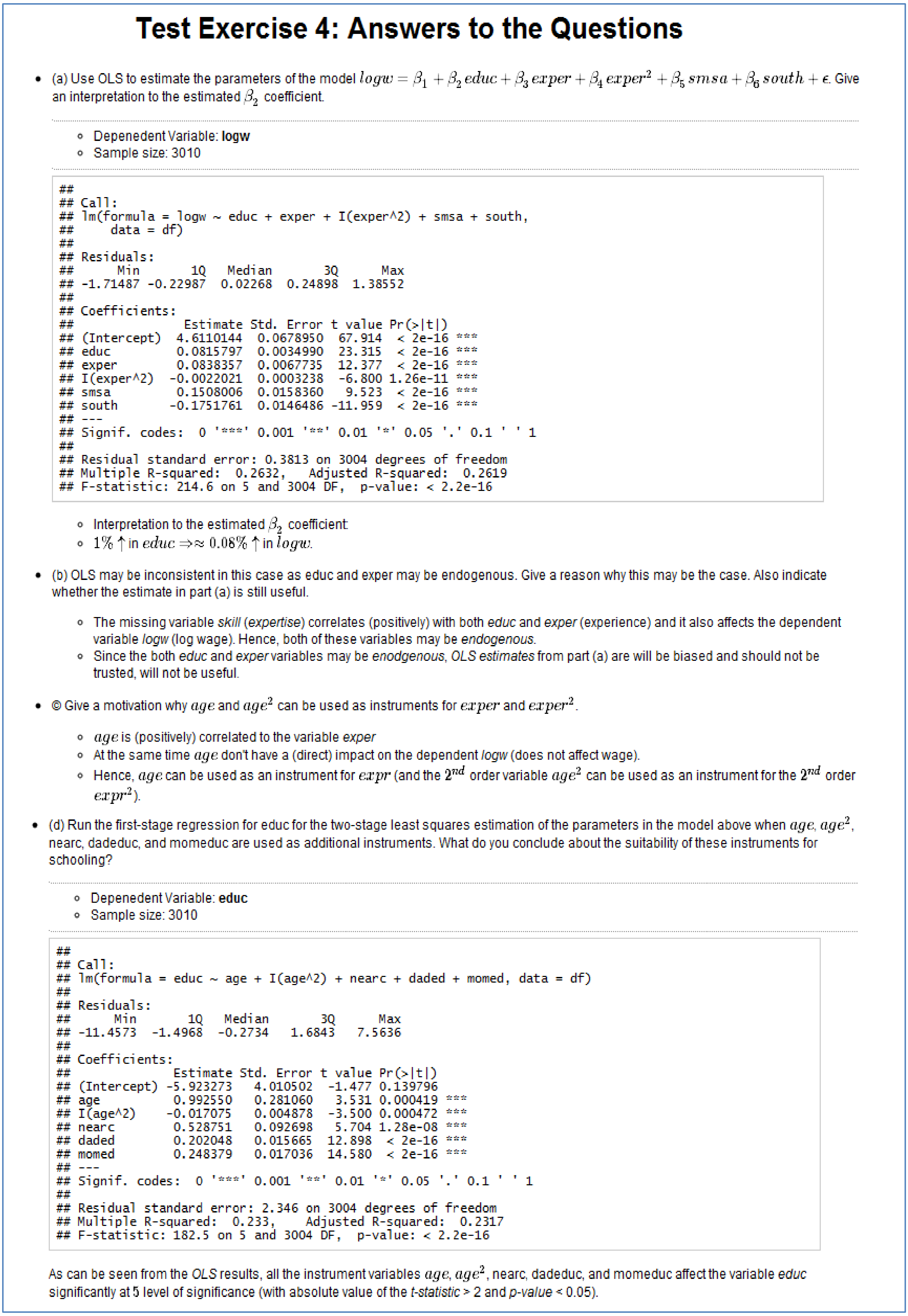


Test Exercise 4: Answers to the Questions

- (a) Use OLS to estimate the parameters of the model $logw = \beta_1 + \beta_2\,educ + \beta_3\,exper + \beta_4\,exper^2 + \beta_5\,smsa + \beta_6\,south + \epsilon$. Give an interpretation to the estimated β_2 coefficient.

    - Depenedent Variable: **logw**
    - Sample size: 3010

```
##
## Call:
## lm(formula = logw ~ educ + exper + I(exper^2) + smsa + south,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71487 -0.22987  0.02268  0.24898  1.38552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.6110144  0.0678950  67.914  < 2e-16 ***
## educ         0.0815797  0.0034990  23.315  < 2e-16 ***
## exper        0.0838357  0.0067735  12.377  < 2e-16 ***
## I(exper^2)  -0.0022021  0.0003238  -6.800 1.26e-11 ***
## smsa         0.1508006  0.0158360   9.523  < 2e-16 ***
## south       -0.1751761  0.0146486 -11.959  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3813 on 3004 degrees of freedom
## Multiple R-squared:  0.2632,	Adjusted R-squared:  0.2619
## F-statistic: 214.6 on 5 and 3004 DF,  p-value: < 2.2e-16
```

    - Interpretation to the estimated β_2 coefficient:
    - $1\% \uparrow$ in $educ \Rightarrow\approx 0.08\% \uparrow$ in $logw$.

- (b) OLS may be inconsistent in this case as educ and exper may be endogenous. Give a reason why this may be the case. Also indicate whether the estimate in part (a) is still useful.

    - The missing variable *skill* (*expertise*) correlates (positively) with both *educ* and *exper* (experience) and it also affects the dependent variable *logw* (log wage). Hence, both of these variables may be *endogenous*.
    - Since the both *educ* and *exper* variables may be *enodgenous*, *OLS estimates* from part (a) are will be biased and should not be trusted, will not be useful.

- © Give a motivation why age and age^2 can be used as instruments for $exper$ and $exper^2$.

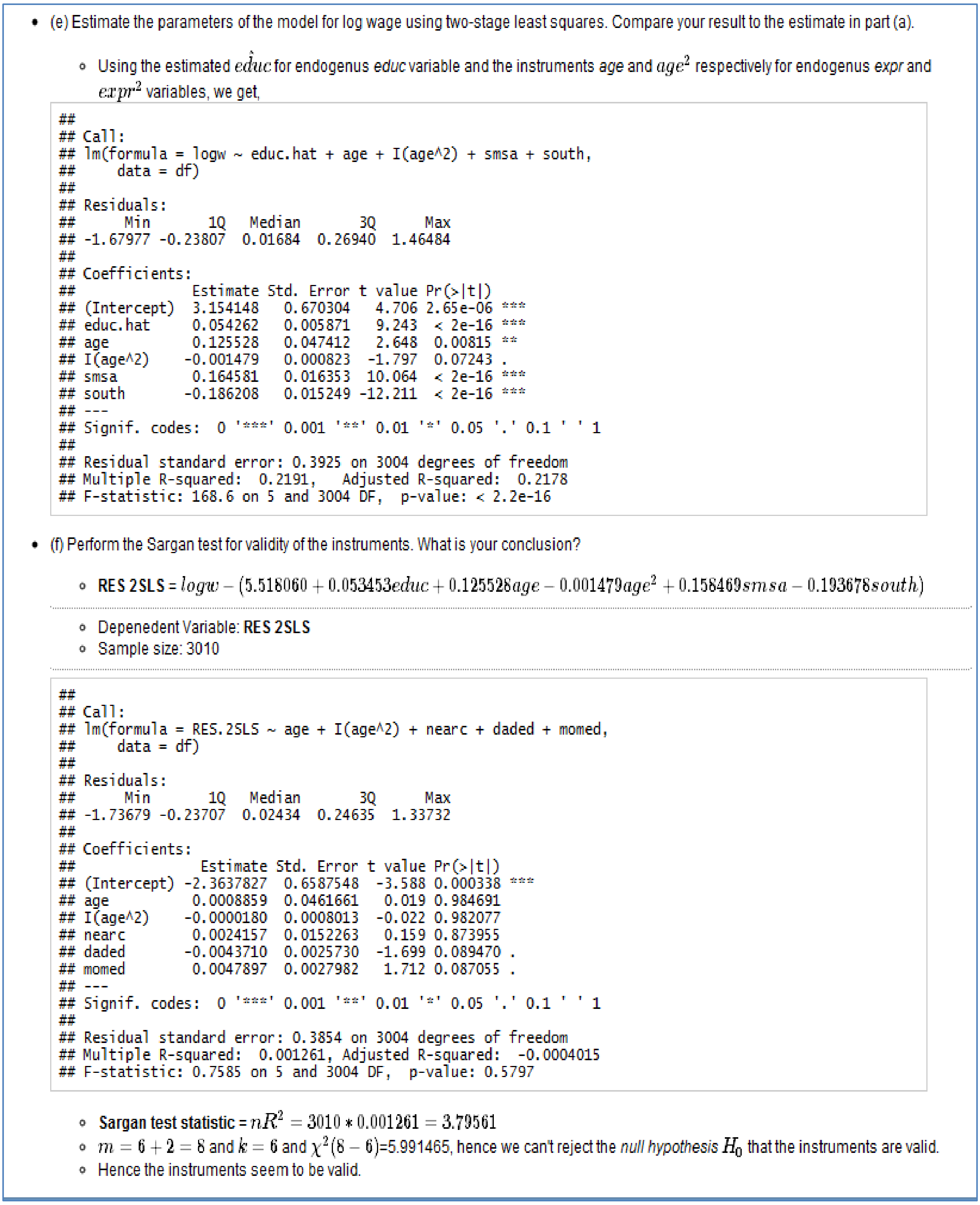
    - age is (positively) correlated to the variable *exper*
    - At the same time age don't have a (direct) impact on the dependent *logw* (does not affect wage).
    - Hence, age can be used as an instrument for $expr$ (and the 2^{nd} order variable age^2 can be used as an instrument for the 2^{nd} order $expr^2$).

- (d) Run the first-stage regression for educ for the two-stage least squares estimation of the parameters in the model above when age, age^2, nearc, dadeduc, and momeduc are used as additional instruments. What do you conclude about the suitability of these instruments for schooling?

    - Depenedent Variable: **educ**
    - Sample size: 3010

```
##
## Call:
## lm(formula = educ ~ age + I(age^2) + nearc + daded + momed, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4573  -1.4968  -0.2734   1.6843   7.5636
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.923273   4.010502  -1.477 0.139796
## age          0.992550   0.281060   3.531 0.000419 ***
## I(age^2)    -0.017075   0.004878  -3.500 0.000472 ***
## nearc        0.528751   0.092698   5.704 1.28e-08 ***
## daded        0.202048   0.015665  12.898  < 2e-16 ***
## momed        0.248379   0.017036  14.580  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.346 on 3004 degrees of freedom
## Multiple R-squared:  0.233,	Adjusted R-squared:  0.2317
## F-statistic: 182.5 on 5 and 3004 DF,  p-value: < 2.2e-16
```

As can be seen from the *OLS* results, all the instrument variables age, age^2, nearc, dadeduc, and momeduc affect the variable *educ* significantly at 5 level of significance (with absolute value of the *t-statistic* > 2 and *p-value* < 0.05).



- (e) Estimate the parameters of the model for log wage using two-stage least squares. Compare your result to the estimate in part (a).
  - Using the estimated $\hat{educ}$ for endogenus *educ* variable and the instruments *age* and age^2 respectively for endogenus *expr* and $expr^2$ variables, we get,

```
##
## Call:
## lm(formula = logw ~ educ.hat + age + I(age^2) + smsa + south,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67977 -0.23807  0.01684  0.26940  1.46484
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.154148   0.670304   4.706 2.65e-06 ***
## educ.hat     0.054262   0.005871   9.243  < 2e-16 ***
## age          0.125528   0.047412   2.648  0.00815 **
## I(age^2)    -0.001479   0.000823  -1.797  0.07243 .
## smsa         0.164581   0.016353  10.064  < 2e-16 ***
## south       -0.186208   0.015249 -12.211  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3925 on 3004 degrees of freedom
## Multiple R-squared:  0.2191,    Adjusted R-squared:  0.2178
## F-statistic: 168.6 on 5 and 3004 DF,  p-value: < 2.2e-16
```

- (f) Perform the Sargan test for validity of the instruments. What is your conclusion?
  - **RES 2SLS** = $logw - (5.518060 + 0.053453educ + 0.125528age - 0.001479age^2 + 0.158469smsa - 0.193678south)$

  - Depenedent Variable: **RES 2SLS**
  - Sample size: 3010

```
##
## Call:
## lm(formula = RES.2SLS ~ age + I(age^2) + nearc + daded + momed,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73679 -0.23707  0.02434  0.24635  1.33732
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.3637827  0.6587548  -3.588 0.000338 ***
## age          0.0008859  0.0461661   0.019 0.984691
## I(age^2)    -0.0000180  0.0008013  -0.022 0.982077
## nearc        0.0024157  0.0152263   0.159 0.873955
## daded       -0.0043710  0.0025730  -1.699 0.089470 .
## momed        0.0047897  0.0027982   1.712 0.087055 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3854 on 3004 degrees of freedom
## Multiple R-squared:  0.001261, Adjusted R-squared:  -0.0004015
## F-statistic: 0.7585 on 5 and 3004 DF,  p-value: 0.5797
```

  - **Sargan test statistic** = $nR^2 = 3010 * 0.001261 = 3.79561$
  - $m = 6 + 2 = 8$ and $k = 6$ and $\chi^2(8 - 6)$=5.991465, hence we can't reject the *null hypothesis* H_0 that the instruments are valid.
  - Hence the instruments seem to be valid.