ECON3150/4150: Introductory Econometrics – Exam Spring 2023

1. (80%) Suppose you have the following data on mother's smoking and children's birth weight. Descriptive statistics of your data stored in data frame `df` are as follows:

```
##                  mean         SD  min max    N
## bwght     118.7080029 20.3588787 23.0 271 1387
## faminc     29.0421774 18.7371174  0.5  65 1387
## motheduc   12.9358327  2.3767284  2.0  18 1387
## parity      1.6330209  0.8941882  1.0   6 1387
## male        0.5212689  0.4997276  0.0   1 1387
## cigs        2.0886806  5.9745789  0.0  50 1387
## cigtax     19.5598414  7.7941839  2.0  38 1387
```

   where

   1. bwght birth weight, ounces (1 ounce = 28.35 gr.)
   2. faminc 1988 family income, $1000s
   3. motheduc mother's yrs of educ
   4. parity birth order of child
   5. male =1 if male child
   6. cigs cigs smked per day while preg
   7. cigtax cig. tax in home state, 1988

   You perform the following analysis:

```
reg1 = feols(bwght ~ cigs, dt)
reg2 = feols(log(bwght) ~ cigs, dt)
reg3 = feols(bwght ~ cigs + faminc, dt)
reg4 = feols(bwght ~ cigs + faminc + motheduc + parity + male, dt)
etable(reg1, reg2, reg3, reg4, signif.code = NA)
```

```
##                            reg1           reg2           reg3           reg4
## Dependent Var.:           bwght      log(bwght)          bwght          bwght
##
## Constant         119.8 (0.575)    4.77 (0.005)   117.0 (1.04)   111.8 (3.21)
## cigs            -0.514 (0.088) -0.004 (0.0008) -0.464 (0.089) -0.473 (0.090)
## faminc                                          0.092 (0.029)  0.099 (0.031)
## motheduc                                                       0.053 (0.238)
## parity                                                          1.65 (0.601)
## male                                                            3.15 (1.07)
```

```
## ______________ ______________ _______________ ______________ ______________
## S.E. type      Heterosk.-rob. Heteroske.-rob. Heterosk.-rob. Heterosk.-rob.
## Observations            1,387           1,387          1,387          1,387
## RMSE                   20.118         0.18875         20.046         19.933
```

a. Interpret the estimates and their standard errors in the 1st column (`reg1`).

   ANSWER HINT: Smoking one cigarette per day while pregnant is *associated* with a 0.5 ounce lower birth weight. This is precisely estimated and statistically significant at conventional levels. The standard error of 0.09 implies a 95% CI of about (-0.69,-0.34). The intercept estimates the average birth weight for those who do not smoke at about 120 ounces and is precisely estimated with a standard error of 0.6 which implies a 95% CI of ca. (118.7,120.9).

b. What is the t-value corresponding to the null-hypothesis that the intercept equals 100?

   ANSWER HINT: (119.8 - 100) / 0.575 = 34.43

c. Construct (briefly explain your steps) and interpret the 80 percent confidence interval for the estimate on `cigs` in 1.a.

   ANSWER HINT: We need to find the critical level which puts 20% in the tails i.e. z for which P(Z<z) = 0.9. The closed nr in the attached table is 0.8997 which implies z = 1.28 or ca. 1.3. and the 80% CI is therefore (-0.514 - 1.28 * 0.088, -0.514 + 1.28 * 0.088) = (-0.6266,-0.4014).

d. Someone argues that cigarettes cannot have a beneficial effect on birth weight and suggests you do one-sided testing. What is the critical value for the null hypothesis that cigarettes do not affect birth weight at the 10% significance level?

   ANSWER HINT: We rule out positive effects which means that we pick the critical level that puts 10% mass in the negative tail, ie the z for which P(Z< -z) = 0.1. The attached table shows only probabilities for positive z but we know that the normal distribution is symmetric, so that P(Z< -z) = 1 - P(Z <z) and we therefore have P(Z<z)=1-.1 = 0.9. The closest probability is 0.8997 which means that the critical level is -1.23 (ie we reject if the t stat is more negative than -1.28).

e. Compute the R-squared of the regression in the first column.

   ANSWER HINT: We have R-squared = 1 - SSR/TSS = 1 - mean squared error / variance of the outcome = 1 - RMSEˆ2 / SD(bwgt)ˆ2 = 1 - (20.118 / 20.3589)ˆ2 = 0.02353. Alternatively one can use $R^2 = var(\hat{\beta}_0 + \hat{\beta}_1 cigs)/var(bwgt) = var(119.8 - 0.514 cigs)/var(bwgt) = (-0.514)^2 var(cigs)/var(bwgt) = (-0.514)^2 5.9746^2/20.3589^2 = 0.02275$. (For completeness note that both approaches ignore degrees of freedom corrections which do not matter substantively here.)

f. 1 ounce is about 28.35 grams. If you would measure birth weight in grams, how would this affect your estimates in the first column?

ANSWER HINT: Instead of estimating $y = b0 + b1 \cdot cigs$ we estimate $y \cdot c = b0 \cdot c + b1 \cdot c \cdot cigs$ which shows that all estimates are multiplied by c=28.35.

g. Interpret the coefficients in the 2nd column (`reg2`).

ANSWER HINT: The coefficient on `cigs` implies that one cigarette per day during pregancy is associated with a 0.4% lower birth weight. the intercept is the estimate of birth weight without smoking (cigs=0) which equals exp(4.77) = 117.9 ounces (ca 3343 grams).

h. 1 ounce is about 28.35 grams. If you would measure birth weight in grams, how would this affect your estimates in column 2?

ANSWER HINT: We estimate $\log(c \cdot y) = log(c) + log(y) = log(c) + b0 + b1cigs$ which shows that the coefficient on `cigs` is unaffected but that the intercept increases with log(28.35).

i. Can we give the estimate on `cigs` in column 3 a causal interpretation? Does the specification in column 4 that adds more regressors change your mind? Motivate your answer.

ANSWER HINT: Here we see that keeping family income constant a cigarette per day during pregnancy is associated with 0.464 ounce lower birth weight (very similar to the first column). We believe that it is causal if we can convincingly rule out omitted variables that i) may be correlated with maternal smoking and ii) that contribute to birth weight. Here one can think of alcohol consumption, nutrition, general health, sleep, stress, etc. none of which is kept constant in column 3. While column 4 adds some controls which may partially account for this (perhaps in particular maternal education), it therefore seems hard to be confident that the estimate is indeed causal.

j. The third column (`reg3`) adds family income to the specification. Use the omitted variable bias formula to compute the correlation between family income (`faminc`) and cigarette smoking (`cigs`).

ANSWER HINT: Compared to `reg3`, $bwgt = b0 + b1 \cdot cigs + b2 \cdot faminc$, the omitted variable bias formula for `reg1` is $\widehat{b1} = b1 + b2 \cdot cov(cigs, faminc)/var(cigs)$. We have $\widehat{b1} = -0.514$ from `reg1`, $b1 = -0.464$, $b2 = 0.092$ from `reg3`, and var(cigs) = 5.9746^2=35.7 from the descriptive statistics. This gives cov(cigs, faminc) = (-0.514 - -0.464) * 5.9746^2 / 0.092 = -19.4. the correlation is therefore -19.4 / (18.7371 * 5.9746) = -0.1734.

k. Perform an F-test (you will need to assume homoskedasticity) that tests the null hypothesis that the coefficients on `motheduc`, `parity`, and `male` are jointly zero in the final regression `reg4`.

ANSWER HINT: We can use the R-squared formulation of the F-test. the R2 in `reg3` equals 1 - (20.046 / 20.3589)^2 = 0.0305, and in `reg4` : 1 - (19.933 / 20.3589)^2 = 0.0414. the F-statistic is therefore: ((0.0414 - 0.0305)/3) / ((1 -

0.0414)/(1387 -6)) = 5.234335 and follows an F(3, 1381) distribution under the null.

l. You want to test whether cigarettes have the same effect on birthweight for kids with mothers who have completed high school (motheduc>=12) compared to children with mothers who do not have a high school diploma (motheduc<12). What is the specification the regression that you will estimate and the null hypothesis you will be testing?

ANSWER HINT: Add a dummy I(motheduc>=12) and an interaction of the dummy and cigs: I(motheduc>=12)*cigs to your regression. The null hypothesis involves a zero coefficient on the interaction.

A friend suggests to use instrumental variable estimation rather than OLS, and proposes to use the cigarette tax in the home state as the instrumental variable. She also provides the following OLS regression results:

```
reg5 = feols(cigtax ~   cigs + faminc + motheduc + parity + male, dt)
reg6 = feols(cigs   ~ cigtax + faminc + motheduc + parity + male, dt)
reg7 = feols(bwght  ~ cigtax + faminc + motheduc + parity + male, dt)
etable(reg5, reg6, reg7, signif.code = NA)
```

```
##                            reg5           reg6          reg7
## Dependent Var.:          cigtax           cigs         bwght
##
## Constant            15.9 (1.32)    7.42 (1.06)  106.3 (3.27)
## cigs                0.056 (0.038)
## faminc             -0.002 (0.013) -0.030 (0.008) 0.114 (0.031)
## motheduc            0.238 (0.101) -0.424 (0.071) 0.228 (0.237)
## parity              0.116 (0.234)  0.275 (0.228)  1.51 (0.611)
## male                0.609 (0.422) -0.073 (0.313)   3.11 (1.08)
## cigtax                             0.031 (0.021) 0.106 (0.070)
## ________________ ______________ ______________ _____________
## S.E. type        Heterosk.-rob. Heterosk.-rob. Heteros.-rob.
## Observations             1,387          1,387         1,387
## RMSE                    7.7637         5.8016        20.104
```

m. Compute and interpret the IV estimate of the effect of cigarettes on birth weight.

ANSWER HINT: IV = reduced form / first stage = 0.106 / 0.031 = 3.42. Smoking one cigarette per day while pregnant is estimated to increase birth weight with 3.42 ounces. this has the unexpected sign.

n. Do you think this instrument satifies the exclusion restriction? Motivate your answer.

ANSWER HINT: No. Cigarette taxes vary across states but the regression does not control for differences across states that may correlate with birth weight (like things noted in answer hint 1.i. above). The instrument is therefore probably not

independent from the error term in the outcome (2nd stage) equation.

o. Is the instrument relevant? Motivate your answer.

ANSWER HINT: No. The instrument is not statistically significant in the first-stage reg6: it has a t = 0.031 / 0.021 = 1.476 or an F of 1.476ˆ2 = 2.179.

2. (20%) In 1980, Kentucky raised its cap on weekly earnings that were covered by worker's disability compensation program. We want to know if this new policy caused workers to spend more time unemployed. The cap increase did not affect low-earnings workers, but did affect high-earnings workers. You have data on the following:

    1. durat duration of benefits
    2. afchnge =1 if after change in benefits
    3. highearn =1 if high earner
    4. male =1 if male
    5. married =1 if married
    6. ky =1 for Kentucky
    7. mi =1 for Michigan
    8. ldurat log(durat)
    9. afhigh afchnge * highearn
    10. head =1 if head injury
    11. neck =1 if neck injury
    12. upextr =1 if upper extremities injury
    13. trunk =1 if trunk injury
    14. lowback =1 if lower back injury
    15. lowextr =1 if lower extremities injury
    16. occdis =1 if occupational disease
    17. manuf =1 if manufacturing industry
    18. construc =1 if construction industry

```
injury = fread("injury.csv")
t(sapply(injury, dstat))
```

```
##                   mean            SD         min             max    N
## durat         9.92220280    24.4975417    0.250000      182.000000 7150
## afchnge       0.47328671     0.4993208    0.000000        1.000000 7150
## highearn      0.39888112     0.4897025    0.000000        1.000000 7150
## male          0.78062798     0.4138501    0.000000        1.000000 7134
## married       0.69225157     0.4615955    0.000000        1.000000 6853
## hosp          0.26209790     0.4398064    0.000000        1.000000 7150
## indust        2.29249123     0.8767738    1.000000        3.000000 7125
## injtype       4.45090909     1.5169241    1.000000        8.000000 7150
## age          34.70584943    12.5902519   12.000000       98.000000 7146
## prewage     329.72848184   182.7989394   81.780602     1583.099976 7150
## totmed     1714.42189230 27853.3720862    0.000000  2323376.500000 7150
## injdes     4384.61034965  1332.2382953 1007.000000     9052.000000 7150
## benefit     162.92344510    61.4193557   14.869200      742.220886 7150
## ky            0.78685315     0.4095592    0.000000        1.000000 7150
## mi            0.21314685     0.4095592    0.000000        1.000000 7150
## ldurat        1.33271222     1.3085423   -1.386294        5.204007 7150
## afhigh        0.19300699     0.3946861    0.000000        1.000000 7150
## lprewage      5.65397907     0.5359215    4.404040        7.367140 7150
```

```
## lage       3.48346979    0.3548254    2.484907      4.584968 7146
## ltotmed    5.92723504    1.7438252    0.000000     14.658532 7150
## head       0.03636364    0.1872064    0.000000      1.000000 7150
## neck       0.01706294    0.1295150    0.000000      1.000000 7150
## upextr     0.29524476    0.4561846    0.000000      1.000000 7150
## trunk      0.11412587    0.3179863    0.000000      1.000000 7150
## lowback    0.26181818    0.4396549    0.000000      1.000000 7150
## lowextr    0.23160839    0.4218896    0.000000      1.000000 7150
## occdis     0.01076923    0.1032218    0.000000      1.000000 7150
## manuf      0.28084211    0.4494421    0.000000      1.000000 7125
## construc   0.14582456    0.3529550    0.000000      1.000000 7125
## highlpre   2.48413655    3.0532814    0.000000      7.367140 7150
```

```
injury[ky==1, mean(log(durat)), by = c("highearn", "afchnge")]
```

```
##    highearn afchnge       V1
##       <int>   <int>    <num>
## 1:        1       1 1.580352
## 2:        0       1 1.133273
## 3:        1       0 1.382094
## 4:        0       0 1.125615
```

You decide to estimate the impact on the logarithm of benefit duration.

a. Use the data in the output above to compute the difference-in-differences estimate.

   ANSWER HINT: (1.580352 - 1.382094) - (1.133273 - 1.125615) = 0.1906. So the policy led to a 19% increase in the duration of benefits.

b. Explain in the context of this application what you need to assume to give the estimate in 2.a a causal interpretation.

   ANSWER HINT: The common trend assumption must hold. In this context it means that without the increase in the cap the high-earnings workers would have experienced the same *change* in the log(duration) of benefits as the low earnings workers.

c. Use the variables in the dataset above to write out the regression equation you would estimate to recover the answer in 2.a.

   ANSWER HINT: ldurat = b0 + b1 * afhigh + b2 * highearn + b3 * afchnge + residual where b1 is the difference-in-differences estimate.

The Cumulative Standard Normal Distribution Function, P(Z <= z)

```
## Second decimal value of z:
##              0      1      2      3      4      5      6      7      8      9
##
##  0.0 :  0.5000 0.5040 0.5080 0.5120 0.5160 0.5199 0.5239 0.5279 0.5319 0.5359
##  0.1 :  0.5398 0.5438 0.5478 0.5517 0.5557 0.5596 0.5636 0.5675 0.5714 0.5753
##  0.2 :  0.5793 0.5832 0.5871 0.5910 0.5948 0.5987 0.6026 0.6064 0.6103 0.6141
##  0.3 :  0.6179 0.6217 0.6255 0.6293 0.6331 0.6368 0.6406 0.6443 0.6480 0.6517
##  0.4 :  0.6554 0.6591 0.6628 0.6664 0.6700 0.6736 0.6772 0.6808 0.6844 0.6879
##  0.5 :  0.6915 0.6950 0.6985 0.7019 0.7054 0.7088 0.7123 0.7157 0.7190 0.7224
##  0.6 :  0.7257 0.7291 0.7324 0.7357 0.7389 0.7422 0.7454 0.7486 0.7517 0.7549
##  0.7 :  0.7580 0.7611 0.7642 0.7673 0.7704 0.7734 0.7764 0.7794 0.7823 0.7852
##  0.8 :  0.7881 0.7910 0.7939 0.7967 0.7995 0.8023 0.8051 0.8078 0.8106 0.8133
##  0.9 :  0.8159 0.8186 0.8212 0.8238 0.8264 0.8289 0.8315 0.8340 0.8365 0.8389
##  1.0 :  0.8413 0.8438 0.8461 0.8485 0.8508 0.8531 0.8554 0.8577 0.8599 0.8621
##  1.1 :  0.8643 0.8665 0.8686 0.8708 0.8729 0.8749 0.8770 0.8790 0.8810 0.8830
##  1.2 :  0.8849 0.8869 0.8888 0.8907 0.8925 0.8944 0.8962 0.8980 0.8997 0.9015
##  1.3 :  0.9032 0.9049 0.9066 0.9082 0.9099 0.9115 0.9131 0.9147 0.9162 0.9177
##  1.4 :  0.9192 0.9207 0.9222 0.9236 0.9251 0.9265 0.9279 0.9292 0.9306 0.9319
##  1.5 :  0.9332 0.9345 0.9357 0.9370 0.9382 0.9394 0.9406 0.9418 0.9429 0.9441
##  1.6 :  0.9452 0.9463 0.9474 0.9484 0.9495 0.9505 0.9515 0.9525 0.9535 0.9545
##  1.7 :  0.9554 0.9564 0.9573 0.9582 0.9591 0.9599 0.9608 0.9616 0.9625 0.9633
##  1.8 :  0.9641 0.9649 0.9656 0.9664 0.9671 0.9678 0.9686 0.9693 0.9699 0.9706
##  1.9 :  0.9713 0.9719 0.9726 0.9732 0.9738 0.9744 0.9750 0.9756 0.9761 0.9767
##  2.0 :  0.9772 0.9778 0.9783 0.9788 0.9793 0.9798 0.9803 0.9808 0.9812 0.9817
##  2.1 :  0.9821 0.9826 0.9830 0.9834 0.9838 0.9842 0.9846 0.9850 0.9854 0.9857
##  2.2 :  0.9861 0.9864 0.9868 0.9871 0.9875 0.9878 0.9881 0.9884 0.9887 0.9890
##  2.3 :  0.9893 0.9896 0.9898 0.9901 0.9904 0.9906 0.9909 0.9911 0.9913 0.9916
##  2.4 :  0.9918 0.9920 0.9922 0.9925 0.9927 0.9929 0.9931 0.9932 0.9934 0.9936
##  2.5 :  0.9938 0.9940 0.9941 0.9943 0.9945 0.9946 0.9948 0.9949 0.9951 0.9952
##  2.6 :  0.9953 0.9955 0.9956 0.9957 0.9959 0.9960 0.9961 0.9962 0.9963 0.9964
##  2.7 :  0.9965 0.9966 0.9967 0.9968 0.9969 0.9970 0.9971 0.9972 0.9973 0.9974
##  2.8 :  0.9974 0.9975 0.9976 0.9977 0.9977 0.9978 0.9979 0.9979 0.9980 0.9981
##  2.9 :  0.9981 0.9982 0.9982 0.9983 0.9984 0.9984 0.9985 0.9985 0.9986 0.9986
```

So for example, P(Z <= 0.22) = 0.5871