Problem Set 03

Lorenzo Bartolo

10/18/2021

Setup

EMPIRICAL /COMPUTER WORK

5. [50 points (all questions here are worth 5points each except part (a) which is 10 points)] Important: As usual, your answer should include a printout (can cut and paste into a file. I will show you how to do this) of relevant calculations on the computer (R or other software output) AND a write up of final answers following the sub parts of the question. The data again are the same as for Problem Set 1. Use these data to answer the following questions.

(a) It is quite common in econometrics to model income variables nonlinearly. Construct a new variable called loginc where loginc=ln(PerCapitaInc). Provide summary statistics for this new variable. (Hint: Think back to how you constructed summary statistics in problem set 1.)

```
rural_atlas_merged$logPerCapitaInc <- log(rural_atlas_merged$PerCapitaInc)
summary(rural_atlas_merged$logPerCapitaInc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   8.690   9.982  10.135  10.121  10.278  11.149       5
```

(b) Now run a model similar to the one that you ran as the final specification in problem set 2 except this time use loginc as the dependent variable (and we are also going to start controlling for metro status in addition to the other controls). Specifically, run a regression of loginc with unemployment rate in 2013 (UnempRate2013) as the main regressor, while also including the other regressors: percentage college educated (Ed5CollegePlusPct), percentage non-Hispanic black in 2010 (BlackNonHispanicPct2010), percentage Hispanic in 2010, and metro status in 2013 (Metro2013). Now, what is the estimated effect of UnempRate2013 in words? Also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

```
rural_atlas_model1 <- lm(formula = rural_atlas_merged$logPerCapitaInc ~
    rural_atlas_merged$UnempRate2013 + rural_atlas_merged$Ed5CollegePlusPct +
        rural_atlas_merged$BlackNonHispanicPct2010)

coefs <- names(coef(rural_atlas_model1))

summary(rural_atlas_model1)
```

```
##
## Call:
## lm(formula = rural_atlas_merged$logPerCapitaInc ~ rural_atlas_merged$UnempRate2013 +
##     rural_atlas_merged$Ed5CollegePlusPct + rural_atlas_merged$BlackNonHispanicPct2010)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86850 -0.08024  0.01711  0.09923  0.74723
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                             10.1337158  0.0120559 840.559
## rural_atlas_merged$UnempRate2013           -0.0412292  0.0010044 -41.047
## rural_atlas_merged$Ed5CollegePlusPct        0.0143310  0.0003306  43.346
## rural_atlas_merged$BlackNonHispanicPct2010 -0.0005672  0.0002111  -2.687
##                                           Pr(>|t|)
## (Intercept)                                < 2e-16 ***
## rural_atlas_merged$UnempRate2013           < 2e-16 ***
## rural_atlas_merged$Ed5CollegePlusPct       < 2e-16 ***
## rural_atlas_merged$BlackNonHispanicPct2010  0.00724 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1666 on 3267 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.6212, Adjusted R-squared:  0.6208
## F-statistic:  1786 on 3 and 3267 DF,  p-value: < 2.2e-16
```

```
coeftest(rural_atlas_model1, vcov = vcovHC(rural_atlas_model1,
    type = "HC0"))
```

```
##
## t test of coefficients:
##
##                                            Estimate  Std. Error  t value
## (Intercept)                             10.13371577  0.01795140 564.5084
## rural_atlas_merged$UnempRate2013           -0.04122924  0.00198116 -20.8107
## rural_atlas_merged$Ed5CollegePlusPct        0.01433103  0.00039787  36.0195
## rural_atlas_merged$BlackNonHispanicPct2010 -0.00056720  0.00022986  -2.4675
##                                           Pr(>|t|)
## (Intercept)                                < 2e-16 ***
## rural_atlas_merged$UnempRate2013           < 2e-16 ***
## rural_atlas_merged$Ed5CollegePlusPct       < 2e-16 ***
## rural_atlas_merged$BlackNonHispanicPct2010  0.01366 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# To get robust F-statistics I am testing against the null
# model (i.e., only the intercept is not set to 0)
linearHypothesis(model = rural_atlas_model1, coefs[-1])
```

```
## Linear hypothesis test
##
## Hypothesis:
## rural_atlas_merged$UnempRate2013 = 0
## rural_atlas_merged$Ed5CollegePlusPct = 0
## rural_atlas_merged$BlackNonHispanicPct2010 = 0
##
## Model 1: restricted model
## Model 2: rural_atlas_merged$logPerCapitaInc ~ rural_atlas_merged$UnempRate2013 +
##     rural_atlas_merged$Ed5CollegePlusPct + rural_atlas_merged$BlackNonHispanicPct2010
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   3270 239.212
## 2   3267  90.623  3    148.59 1785.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# heteroskedasticity-robust F-test
linearHypothesis(model = rural_atlas_model1, coefs[-1], white.adjust = "hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## rural_atlas_merged$UnempRate2013 = 0
## rural_atlas_merged$Ed5CollegePlusPct = 0
## rural_atlas_merged$BlackNonHispanicPct2010 = 0
##
## Model 1: restricted model
## Model 2: rural_atlas_merged$logPerCapitaInc ~ rural_atlas_merged$UnempRate2013 +
##     rural_atlas_merged$Ed5CollegePlusPct + rural_atlas_merged$BlackNonHispanicPct2010
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   3270
## 2   3267  3 1183.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# The effect of the UnempRate2013 on PerCapitaInc is
# statistically significant at the 10%, 5%, and 1% levels
# TODO: What about the other coefficients?
```

(c) What is the null hypothesis corresponding to the F-statistic as reported in the output for the regression in part (b)?

- Null hypothesis: UnempRate2013 = PerCapitaInc = Ed5CollegePlusPct = BlackNonHispanicPct2010 = 0

(d) What is the conclusion of the F-test as reported in the output for the regression in part (b)? Explain. (i.e. Do you reject or fail to reject the null hypothesis above and how do you know this?)

- Testing the joint hypothesis on two or more coefficients using the linear hypothesis test above, we reject the null hypothesis at the 10%, 5%, and 1% levels

(e) What are the degrees of freedom (the specific numbers reported) for the F-test as specified in your regression output? Where are these numbers coming from for this case? (i.e., how is the computer calculating these numbers specifically given the dataset here?)

- The degrees of freedom are 3 for the F-test, given we specified 3 restrictions for this model

(f) Discuss what the standard error of the regression (SER), R^2 and $\bar{R}^2$ in part (b) are telling you in terms of the numbers that you have found.

- The R^2 and $\bar{R}^2$ tell us whether the regressors (UnempRate2013, Ed5CollegePlusPct, BlackNonHispanic2010 in this case) are good at predicting, or "explaining," the values of the dependent variable (per capita income) in the sample of data
- An R^2 and $\bar{R}^2$ near 1 means that the regressors are good at predicting the values of the dependent variable (PerCapitaInc) in the sample, and an R^2 and $\bar{R}^2$ near 0 means that they are not.
- It is evident that the relatively high R^2 and $\bar{R}^2$ CANNOT be used to the conclude that the estimated relation between UnempRate2013 and PerCapitaInc is causal. There could be other determinants and/or control variables.
- The SER is a measure of the spread of the observations around the regression line, in this case measured in dollars of per-capita income

(g) Using what you know about the difference between the two formulas, explain specifically why the R^2 and $\bar{R}^2$ statistics so similar for this case.

- TODO: The R^2 and the $\bar{R}^2$ are similar because...

(h) Use an F-test to test the joint significance of the additional regressors: Ed5CollegePlus, BlackNonHispanicPct2010, HispanicPct2010, and Metro2013. Find this test statistic and clearly indicate the conclusions of the test.

```
rural_atlas_model_h <- lm(formula = rural_atlas_merged$logPerCapitaInc ~
    rural_atlas_merged$UnempRate2013 + rural_atlas_merged$Ed5CollegePlusPct +
        rural_atlas_merged$BlackNonHispanicPct2010 + rural_atlas_merged$HispanicPct2010 +
        rural_atlas_merged$Metro2013)

coefs_model_h <- names(coef(rural_atlas_model_h))

summary(rural_atlas_model_h)
```

```
##
## Call:
## lm(formula = rural_atlas_merged$logPerCapitaInc ~ rural_atlas_merged$UnempRate2013 +
##     rural_atlas_merged$Ed5CollegePlusPct + rural_atlas_merged$BlackNonHispanicPct2010 +
##     rural_atlas_merged$HispanicPct2010 + rural_atlas_merged$Metro2013)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84074 -0.07798  0.00409  0.08283  0.71550
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             10.0947358  0.0107733 937.018
## rural_atlas_merged$UnempRate2013        -0.0292432  0.0009628 -30.373
## rural_atlas_merged$Ed5CollegePlusPct     0.0141150  0.0003211  43.961
## rural_atlas_merged$BlackNonHispanicPct2010 -0.0023356  0.0001914 -12.205
## rural_atlas_merged$HispanicPct2010      -0.0049064  0.0001491 -32.904
## rural_atlas_merged$Metro2013             0.0475008  0.0058635   8.101
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## rural_atlas_merged$UnempRate2013         < 2e-16 ***
## rural_atlas_merged$Ed5CollegePlusPct     < 2e-16 ***
## rural_atlas_merged$BlackNonHispanicPct2010  < 2e-16 ***
## rural_atlas_merged$HispanicPct2010       < 2e-16 ***
## rural_atlas_merged$Metro2013            7.66e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1443 on 3213 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.7157, Adjusted R-squared:  0.7153
## F-statistic:  1618 on 5 and 3213 DF,  p-value: < 2.2e-16
```

```
coeftest(rural_atlas_model_h, vcov = vcovHC(rural_atlas_model_h,
    type = "HC0"))
```

```
##
## t test of coefficients:
##
##                                           Estimate  Std. Error  t value
## (Intercept)                             10.09473580  0.01384841 728.9454
## rural_atlas_merged$UnempRate2013        -0.02924318  0.00128119 -22.8251
## rural_atlas_merged$Ed5CollegePlusPct     0.01411501  0.00042802  32.9775
```

```
## rural_atlas_merged$BlackNonHispanicPct2010 -0.00233558  0.00018130 -12.8825
## rural_atlas_merged$HispanicPct2010         -0.00490639  0.00020533 -23.8956
## rural_atlas_merged$Metro2013                0.04750081  0.00582742   8.1513
##                                             Pr(>|t|)
## (Intercept)                                < 2.2e-16 ***
## rural_atlas_merged$UnempRate2013           < 2.2e-16 ***
## rural_atlas_merged$Ed5CollegePlusPct       < 2.2e-16 ***
## rural_atlas_merged$BlackNonHispanicPct2010 < 2.2e-16 ***
## rural_atlas_merged$HispanicPct2010         < 2.2e-16 ***
## rural_atlas_merged$Metro2013               5.106e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# heteroskedasticity-robust F-test
linearHypothesis(model = rural_atlas_model_h, coefs_model_h[-1],
    white.adjust = "hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## rural_atlas_merged$UnempRate2013 = 0
## rural_atlas_merged$Ed5CollegePlusPct = 0
## rural_atlas_merged$BlackNonHispanicPct2010 = 0
## rural_atlas_merged$HispanicPct2010 = 0
## rural_atlas_merged$Metro2013 = 0
##
## Model 1: restricted model
## Model 2: rural_atlas_merged$logPerCapitaInc ~ rural_atlas_merged$UnempRate2013 +
##     rural_atlas_merged$Ed5CollegePlusPct + rural_atlas_merged$BlackNonHispanicPct2010 +
##     rural_atlas_merged$HispanicPct2010 + rural_atlas_merged$Metro2013
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   3218
## 2   3213  5 968.78 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# The value of the F-statistic computed from the data,
# 1618, exceeds 3.02, thus, we can reject the null
# hypothesis that the coefficients are zero at any level of
# significance

# as summarized in this F-statistic, we can reject the null
# hypothesis that the UnempRate2013 has no effect on
# PerCapitaInc (holding constant: Ed5CollegePlusPct,
# BlackNonHispanicPct2010, HispanicPct2010, and Metro2013).
```

(i) If you had more time to study this question and/or more/different data, what would you suggest doing next? Propose additional variables to add and/or different specifications to try and give specific reasons why you are suggesting these. Answers will vary for this part of the problem.

- I would propose analyzing a coefficient that represents the number of discouraged workers. Since discouraged workers are not counted as "unemployed" in economic data.
- Since we know the unemployment rate is correlated to per capita income, it would be interesting to see how the unemployment rate effects per capita income while controlling for the number of discouraged workers.