Problem Set 2

Lorenzo Bartolo

EMPIRICAL /COMPUTER WORK

4. [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) Run a regression to determine the impact of the 2013 unemployment rate (UnempRate2013) on the per capita income (PerCapitaInc) in a county. What is the estimated slope? Explain what this number means in words in terms of the unemployment rate and in terms of per capita income. Also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels. For this first pass, use homoskedastic standard errors.

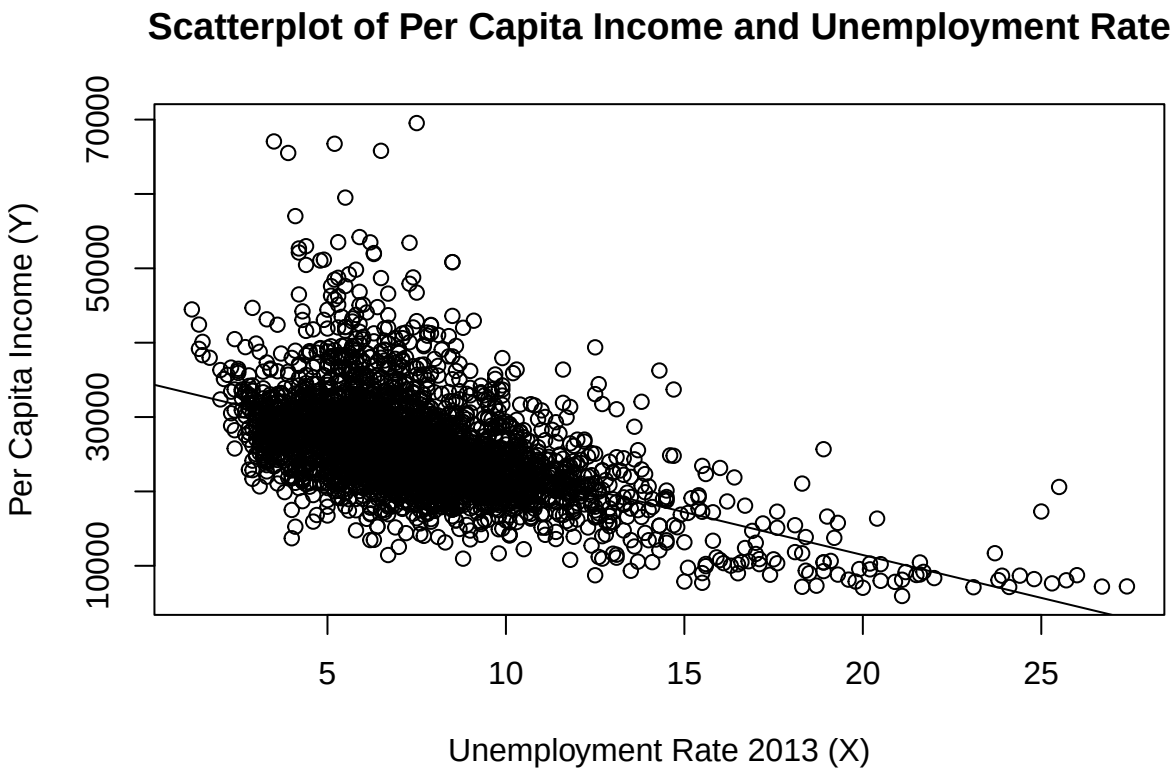
$$\widehat{PerCapitaInc} = \underset{(9.47)}{34507.06} - \underset{(0.49)}{1152.81} \times UnempRate2013 \ , \ R^2 = 0.2929 \ , \ SER = 5613$$

```
summary.lm(model_a <- lm(formula = rural_atlas_merged$PerCapitaInc ~
    rural_atlas_merged$UnempRate2013))
```

```
##
## Call:
## lm(formula = rural_atlas_merged$PerCapitaInc ~ rural_atlas_merged$UnempRate2013)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16191  -3523   -708   2327  43668
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      34507.06     257.92   133.8   <2e-16 ***
## rural_atlas_merged$UnempRate2013 -1152.81      31.33   -36.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5613 on 3269 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.2929, Adjusted R-squared:  0.2927
## F-statistic:  1354 on 1 and 3269 DF,  p-value: < 2.2e-16
```

```
# Graph
plot(PerCapitaInc ~ UnempRate2013, data = rural_atlas_merged,
    main = "Scatterplot of Per Capita Income and Unemployment Rate",
    xlab = "Unemployment Rate 2013 (X)", ylab = "Per Capita Income (Y)")
```

```
abline(model_a)
```




```
cat("Correlation between UnempRate2013 and PerCapitaInc: ", cor(rural_atlas_merged$UnempRate2013,
    rural_atlas_merged$PerCapitaInc, use = "complete.obs"))
```

```
## Correlation between UnempRate2013 and PerCapitaInc:  -0.5412059
```

```
# Robust t test
coeftest(model_a, vcov = vcovHC(model_a, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                      34507.063    258.767 133.352 < 2.2e-16 ***
## rural_atlas_merged$UnempRate2013 -1152.811     29.064 -39.665 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 1% level: Reject the hypothesis that the coefficient on
# the unemployment rate is zero at the 1% level.
confint(model_a, level = 0.99)
```

```
##                                       0.5 %    99.5 %
## (Intercept)                       33842.325 35171.801
## rural_atlas_merged$UnempRate2013 -1233.553 -1072.069
```

```
# 5% level: Reject the hypothesis that the coefficient on
# the unemployment rate is zero at the 5% level.
confint(model_a, level = 0.95)
```

```
##                                       2.5 %    97.5 %
## (Intercept)                       34001.368 35012.758
## rural_atlas_merged$UnempRate2013 -1214.235 -1091.387
```

```
# 10% level: Reject the hypothesis that the coefficient on
# the unemployment rate is zero at the 10% level.
confint(model_a, level = 0.9)
```

```
##                                         5 %      95 %
## (Intercept)                       34082.707 34931.419
## rural_atlas_merged$UnempRate2013 -1204.355 -1101.267
```

```
# Intercept = 34507.06

# Slope = The slope of -1152.81 means that when the
# unemployment rate differs by 1%, on average, the per
# capita income is lower by $1,152.81

# R-squared = On average 29.29% of the variance of the per
# capita income rate is explained by the unemployment rate

# SER = On average the deviation of the actual achieved per
# capita income rate and the regression line is $5,613
```

```
# Coefficients are the same as part (a)
summary.lm(model_b <- lm(formula = rural_atlas_merged$PerCapitaInc ~
    rural_atlas_merged$UnempRate2013))
```

(b) Re-run the regression from part (a) but this time use heteroskedastic standard errors. Are your coefficients the same as in part (a)? Why? Are your standard errors (of your betas) the same as in part (a)? Why?

```
##
## Call:
## lm(formula = rural_atlas_merged$PerCapitaInc ~ rural_atlas_merged$UnempRate2013)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16191  -3523   -708   2327  43668
##
## Coefficients:
```

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      34507.06     257.92   133.8   <2e-16 ***
## rural_atlas_merged$UnempRate2013 -1152.81      31.33   -36.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5613 on 3269 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.2929, Adjusted R-squared:  0.2927
## F-statistic:  1354 on 1 and 3269 DF,  p-value: < 2.2e-16
```

```
# Robust t test
coeftest(model_b, vcov = vcovHC(model_a, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                      34507.063    258.767 133.352 < 2.2e-16 ***
## rural_atlas_merged$UnempRate2013 -1152.811     29.064 -39.665 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.lm(model_c <- lm(formula = rural_atlas_merged$PerCapitaInc ~
    rural_atlas_merged$UnempRate2013 + rural_atlas_merged$Ed5CollegePlusPct +
        rural_atlas_merged$BlackNonHispanicPct2010 + rural_atlas_merged$HispanicPct2010))
```

(c) Run the same regression as in part (b) but now also include the additional regressors percentage of the population that is college-educated (Ed5CollegePlusPct), percentage of the population that is black (BlackNonHispanicPct2010), and percentage of the population that is Hispanic (HispanicPct2010). Now, what is the estimated impact of unemployment rate in 2013 on per capita income? Also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

```
##
## Call:
## lm(formula = rural_atlas_merged$PerCapitaInc ~ rural_atlas_merged$UnempRate2013 +
##     rural_atlas_merged$Ed5CollegePlusPct + rural_atlas_merged$BlackNonHispanicPct2010 +
##     rural_atlas_merged$HispanicPct2010)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20528.1  -1912.5    -44.4   1906.4  26878.9
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                         20905.800    272.950   76.59
## rural_atlas_merged$UnempRate2013     -500.569     24.661  -20.30
## rural_atlas_merged$Ed5CollegePlusPct  465.544      7.427   62.68
## rural_atlas_merged$BlackNonHispanicPct2010   -51.490      4.868  -10.58
```

```
## rural_atlas_merged$HispanicPct2010           -82.032      3.793  -21.63
##                                                Pr(>|t|)
## (Intercept)                                      <2e-16 ***
## rural_atlas_merged$UnempRate2013                 <2e-16 ***
## rural_atlas_merged$Ed5CollegePlusPct             <2e-16 ***
## rural_atlas_merged$BlackNonHispanicPct2010       <2e-16 ***
## rural_atlas_merged$HispanicPct2010               <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3718 on 3266 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:   0.69,  Adjusted R-squared:  0.6897
## F-statistic:  1818 on 4 and 3266 DF,  p-value: < 2.2e-16
```

```
# Robust t test
coeftest(model_c, vcov = vcovHC(model_c, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                                              Estimate Std. Error t value
## (Intercept)                                 20905.8003   371.7782  56.232
## rural_atlas_merged$UnempRate2013             -500.5690    25.9183 -19.313
## rural_atlas_merged$Ed5CollegePlusPct          465.5437    13.3529  34.864
## rural_atlas_merged$BlackNonHispanicPct2010    -51.4904     4.0809 -12.617
## rural_atlas_merged$HispanicPct2010            -82.0316     4.0502 -20.254
##                                               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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 1% level:
confint(model_c, level = 0.99)
```

```
##                                                  0.5 %      99.5 %
## (Intercept)                                 20202.31716 21609.28345
## rural_atlas_merged$UnempRate2013             -564.12970  -437.00832
## rural_atlas_merged$Ed5CollegePlusPct          446.40171   484.68577
## rural_atlas_merged$BlackNonHispanicPct2010    -64.03770   -38.94315
## rural_atlas_merged$HispanicPct2010            -91.80615   -72.25700
```

```
# 5% level:
confint(model_c, level = 0.95)
```

```
##                                                  2.5 %      97.5 %
## (Intercept)                                 20370.63025 21440.97037
## rural_atlas_merged$UnempRate2013             -548.92238  -452.21564
```

```
## rural_atlas_merged$Ed5CollegePlusPct          450.98157   480.10591
## rural_atlas_merged$BlackNonHispanicPct2010   -61.03568   -41.94517
## rural_atlas_merged$HispanicPct2010           -89.46752   -74.59563
```

```
# 10% level:
confint(model_c, level = 0.9)
```

```
##                                                    5 %        95 %
## (Intercept)                                20456.71050 21354.89012
## rural_atlas_merged$UnempRate2013            -541.14491  -459.99311
## rural_atlas_merged$Ed5CollegePlusPct         453.32384   477.76363
## rural_atlas_merged$BlackNonHispanicPct2010   -59.50036   -43.48049
## rural_atlas_merged$HispanicPct2010           -88.27147   -75.79168
```

```
# If the regressor is correlated with a variable that has
# been omitted from the analysis (for example: Black Non
# Hispanic) and that determines, in part, the dependent
# variable, then the results will change when you add in
# additional variables to the regression.
```

(d) Provide economic/econometric intuition as to why the impact of the unemployment rate's impact on per capita income changed between parts (b) and (c). Note that I am asking you to think about the context (and hence the "story" behind these data).

```
# compute 95% confidence interval for coefficients
lm_summ <- summary(model_b)

# Lower -1152.81 - 1.96 * 31.33 = -1214.235
c("lower" = lm_summ$coef[2,1] - qt(0.975, df = lm_summ$df[2]) * lm_summ$coef[2, 2],

# Upper -1152.81 + 1.96 * 31.33 = -1091.387
  "upper" = lm_summ$coef[2,1] + qt(0.975, df = lm_summ$df[2]) * lm_summ$coef[2, 2])
```

(e) Construct a 95% confidence interval for the slope coefficient on UnempRate2013 in (c). Write out your calculations. Clearly indicate how this confidence interval relates to whether UnempRate2013 is statistically significant or not in this context by relating your answer to your constructed confidence interval.

```
##     lower     upper
## -1214.235 -1091.387
```

This interval *does not* contain the value zero which leads to the rejection of the null hypothesis $\beta_{1,0} = 0$.

```
metro <- subset(rural_atlas_merged, Metro2013 == 1)

summary.lm(metro_model <- lm(formula = metro$PerCapitaInc ~ metro$UnempRate2013))
```

(f) You recall from problem set 1 that both the means of per capita income and of unemployment rate in 2013 are quite different across metro and nonmetro areas. You therefore want to explore this in more detail. Run the regression from (c) using only metro areas in 2013 (Metro2013==1). [Hint: You need to restrict the data based on a criterion before running the regression.] Now, what is the estimated effect of the 2013 unemployment rate on per capita income and also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

```
##
## Call:
## lm(formula = metro$PerCapitaInc ~ metro$UnempRate2013)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16651  -3933  -1035   2615  41298
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          39518.67     462.95   85.36   <2e-16 ***
## metro$UnempRate2013 -1505.06      54.85  -27.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6263 on 1232 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3794, Adjusted R-squared:  0.3788
## F-statistic:   753 on 1 and 1232 DF,  p-value: < 2.2e-16
```

```
# Robust t test
coeftest(metro_model, vcov = vcovHC(metro_model, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          39518.672    520.714  75.893 < 2.2e-16 ***
## metro$UnempRate2013 -1505.056     59.264 -25.396 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
non_metro <- subset(rural_atlas_merged, Metro2013 == 0)

summary.lm(non_metro_model <- lm(formula = non_metro$PerCapitaInc ~
    non_metro$UnempRate2013))
```

(g) Now, run the regression from (c) using only non-metro areas in 2013 (Metro2013==0). [Hint: You need to restrict the data based on a criterion before running the regression.] Now, what is the estimated effect of the 2013 unemployment rate on per capita income and also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

```
##
## Call:
## lm(formula = non_metro$PerCapitaInc ~ non_metro$UnempRate2013)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13943  -2789   -308   2175  40515
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              31429.14     262.26   119.8   <2e-16 ***
## non_metro$UnempRate2013  -945.33      32.27   -29.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4425 on 1983 degrees of freedom
## Multiple R-squared:  0.3021, Adjusted R-squared:  0.3017
## F-statistic: 858.3 on 1 and 1983 DF,  p-value: < 2.2e-16
```

```
# Robust t test
coeftest(non_metro_model, vcov = vcovHC(non_metro_model, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                           Estimate Std. Error t value  Pr(>|t|)
## (Intercept)              31429.141    269.813 116.485 < 2.2e-16 ***
## non_metro$UnempRate2013  -945.326     32.207 -29.352 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 1% level:
confint(non_metro_model, level = 0.99)
```

```
##                              0.5 %     99.5 %
## (Intercept)              30752.960 32105.3215
## non_metro$UnempRate2013 -1028.519  -862.1336
```

```
# 5% level:
confint(non_metro_model, level = 0.95)
```

```
##                              2.5 %     97.5 %
## (Intercept)              30914.812 31943.4695
## non_metro$UnempRate2013 -1008.606  -882.0468
```

```
# 10% level:
confint(non_metro_model, level = 0.9)
```

```
##                                5 %       95 %
## (Intercept)              30997.5643 31860.7172
## non_metro$UnempRate2013  -998.4246  -892.2281
```

```
# Because the unemployment rate and per capita income
# differ substantially from metro and non metro areas, the
# coefficients are also substantially different from the
# previous models.
```

(h) What did you learn from the comparison between results in parts (f) and (g)? Explain your answer. Note that I again am asking you to think about the context (and hence the "story" behind these data).

```
summary.lm(model_i <- lm(formula = rural_atlas_merged$PerCapitaInc ~
    rural_atlas_merged$Ed5CollegePlusPct + rural_atlas_merged$UnempRate2010 +
        rural_atlas_merged$BlackNonHispanicPct2010 + rural_atlas_merged$Metro2013))
```

(i) Return to the full sample. Now, run a regression to determine the impact of changing the percentage of the population which is college educated (Ed5CollegePlusPct) on the per capita income (PerCapitaInc) in a county. Include controls for the unemployment rate in 2010 (UnempRate2010), percentage of the population that is black (BlackNonHispanicPct2010), and now also include a dummy variable for metro status (Metro2013). Now, what is the estimated impact of percentage with a college education on per capita income? Also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

```
##
## Call:
## lm(formula = rural_atlas_merged$PerCapitaInc ~ rural_atlas_merged$Ed5CollegePlusPct +
##     rural_atlas_merged$UnempRate2010 + rural_atlas_merged$BlackNonHispanicPct2010 +
##     rural_atlas_merged$Metro2013)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19964.5  -2056.9    224.4   2263.2  25865.4
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             22065.865    328.326  67.207
## rural_atlas_merged$Ed5CollegePlusPct      435.758      9.228  47.219
## rural_atlas_merged$UnempRate2010         -584.197     22.860 -25.556
## rural_atlas_merged$BlackNonHispanicPct2010  -30.794      5.270  -5.843
## rural_atlas_merged$Metro2013                575.238    166.489   3.455
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## rural_atlas_merged$Ed5CollegePlusPct        < 2e-16 ***
## rural_atlas_merged$UnempRate2010            < 2e-16 ***
## rural_atlas_merged$BlackNonHispanicPct2010 5.64e-09 ***
## rural_atlas_merged$Metro2013               0.000557 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4109 on 3214 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.6197, Adjusted R-squared:  0.6192
## F-statistic:  1309 on 4 and 3214 DF,  p-value: < 2.2e-16
```

```
# Robust t test
coeftest(model_i, vcov = vcovHC(model_i, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                                            Estimate Std. Error  t value
## (Intercept)                               22065.8649   475.3873  46.4166
## rural_atlas_merged$Ed5CollegePlusPct        435.7577    15.4578  28.1902
## rural_atlas_merged$UnempRate2010           -584.1967    32.8980 -17.7578
## rural_atlas_merged$BlackNonHispanicPct2010  -30.7945     4.8904  -6.2970
## rural_atlas_merged$Metro2013                575.2383   172.6156   3.3325
##                                             Pr(>|t|)
## (Intercept)                               < 2.2e-16 ***
## rural_atlas_merged$Ed5CollegePlusPct      < 2.2e-16 ***
## rural_atlas_merged$UnempRate2010          < 2.2e-16 ***
## rural_atlas_merged$BlackNonHispanicPct2010 3.447e-10 ***
## rural_atlas_merged$Metro2013              0.0008705 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 1% level:
confint(model_i, level = 0.99)
```

```
##                                                0.5 %      99.5 %
## (Intercept)                               21219.64961 22912.08022
## rural_atlas_merged$Ed5CollegePlusPct        411.97262   459.54274
## rural_atlas_merged$UnempRate2010           -643.11457  -525.27887
## rural_atlas_merged$BlackNonHispanicPct2010  -44.37835   -17.21059
## rural_atlas_merged$Metro2013                146.13736  1004.33934
```

```
# 5% level:
confint(model_i, level = 0.95)
```

```
##                                                2.5 %      97.5 %
## (Intercept)                               21422.11455 22709.61527
## rural_atlas_merged$Ed5CollegePlusPct        417.66342   453.85194
## rural_atlas_merged$UnempRate2010           -629.01792  -539.37551
## rural_atlas_merged$BlackNonHispanicPct2010  -41.12828   -20.46066
## rural_atlas_merged$Metro2013                248.80379   901.67290
```

```
# 10% level:
confint(model_i, level = 0.9)
```

```
##                                                  5 %        95 %
## (Intercept)                               21525.66033 22606.06950
```

```
## rural_atlas_merged$Ed5CollegePlusPct         420.57384   450.94152
## rural_atlas_merged$UnempRate2010            -621.80853  -546.58490
## rural_atlas_merged$BlackNonHispanicPct2010   -39.46611   -22.12283
## rural_atlas_merged$Metro2013                 301.31005   849.16665
```