ECO410H: Practice Questions 5

1. This question helps refresh some of your simple regression analysis skills (i.e. a single explanatory, right-hand-side variable). Practice reading this STATA regression output for the data from Collins and Preston (1966) "Concentration and price-cost margins in food manufacturing industries." An Excel spreadsheet with the original data is posted on the course site next to this document. You may use statistical software that you are more familiar with to run the same regression as below and in class.

```
regress lerner_index cr4
```

```
      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  1,    30) =   20.16
       Model |  1391.98951     1  1391.98951           Prob > F      =  0.0001
    Residual |  2071.26427    30  69.0421424           R-squared     =  0.4019
-------------+------------------------------           Adj R-squared =  0.3820
       Total |  3463.25378    31  111.717864           Root MSE      =  8.3092

------------------------------------------------------------------------------
lerner_index |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cr4 |   .3104375   .0691375     4.49   0.000       .16924    .4516351
       _cons |   6.048936   3.192605     1.89   0.068    -.4712335     12.5691
------------------------------------------------------------------------------
```

2. This question helps refresh some of your multiple regression analysis skills (i.e. more than one explanatory, right-hand-side variables). Because it presents an analysis of hypothetical cross-sectional data, it gives no information about the units of measurement or variable definitions. In all cases, the linearity assumptions underlying the analysis of these data are reasonable.

   (a) The table below summarizes these data:

```
summarize y x1-x3
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |       100    79.81715     5.91038   64.78792   93.73011
          x1 |       100    7.913418    1.177951   4.315967   10.97175
          x2 |       100    7.881578    1.145646   4.396366   10.39832
          x3 |       100    4.119784     1.01848   1.724855   6.396665
```

   i. How many cross-sectional units have been observed (number of rows)?

   ii. How many variables are observed (number of columns)?

   iii. Which of variables exhibits the greatest variation?

   iv. Can we tell whether these variables are correlated with each other by inspection of this table?

   (b) The matrix below summarizes the correlations among the variables:

```
correlate y x1 - x3
(obs=100)
```

```
             |        y       x1       x2       x3
-------------+------------------------------------
           y |   1.0000
          x1 |  -0.5150   1.0000
          x2 |  -0.7225   0.8704   1.0000
          x3 |   0.1726   0.0736   0.0137   1.0000
```

i. Which of the variables are positively correlated with each other? Of these, which are strong correlations, moderately strong correlations, or weak correlations?

ii. Which of the variables are negatively correlated with each other? Of these, which are strong correlations, moderately strong correlations, or weak correlations?

(c) Suppose y is affected by x1, x2, and x3 and that a linear regression model is appropriate. Assume that all of the underlying assumptions of the linear regression model hold.

```
regress y x1 x2 x3
```

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =   47.88
       Model |  2072.86204     3  690.954014           Prob > F      =  0.0000
    Residual |  1385.46437    96  14.4319206           R-squared     =  0.5994
-------------+------------------------------           Adj R-squared =  0.5869
       Total |  3458.32642    99  34.9325901           Root MSE      =  3.7989

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.158004    .663641     3.25   0.002     .8406867    3.475321
          x2 |  -5.669571   .6805654    -8.33   0.000    -7.020483   -4.318659
          x3 |   .9049954   .3778971     2.39   0.019     .1548755    1.655115
       _cons |   103.6967   3.076337    33.71   0.000     97.59026    109.8032
------------------------------------------------------------------------------
```

i. Write down the formal regression model behind the regression output above. Include the parameters to be estimated and an appropriate observation index.

ii. What is k? (To refresh your memory, recall that the degrees of freedom (df) of a multiple regression are $n - k - 1$).

iii. What is the R-squared of this regression? What is the interpretation?

iv. Is the overall model statistically significant?

v. What is the meaning of "Root MSE," which stands for root mean squared error? What does 3.7989 measure? Why is it relevant to the regression analysis?

vi. Which of the slope estimates are "statistically significant"?

vii. Can we reject the hypothesis that the parameter on x1 is 3?

viii. How is it possible that the slope coefficient on x1 is positive while the coefficient of correlation between x1 and y is negative? Is there a bias?

ix. Given that none of the underlying assumptions of the linear regression model are violated, what is the interpretation of each of the estimated slope coefficients?

x. What is the relationship between y and x1 if the average values of the other variables are plugged into the estimated equation?

xi. Describe what will happen to the relationship between y and x1 if the maximum values of the other variables are plugged into the estimated equation instead?

(d) The regression below is estimated with the same data as above:

```
regress y x1 x2
```

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   65.74
       Model |  1990.09267     2  995.046334           Prob > F      =  0.0000
    Residual |  1468.23375    97  15.1364304           R-squared     =  0.5754
-------------+------------------------------           Adj R-squared =  0.5667
       Total |  3458.32642    99  34.9325901           Root MSE      =  3.8906

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.357277   .6742826     3.50   0.001     1.019012    3.695541
          x2 |  -5.836894   .6932961    -8.42   0.000    -7.212895   -4.460894
       _cons |    107.167   2.779132    38.56   0.000     101.6512    112.6828
------------------------------------------------------------------------------
```

i. Write down the formal regression model behind the regression output above. Include the parameters to be estimated and an appropriate observation index.

ii. Suppose that x3 does have an effect on y but is not observed in the available data. Where is x3 in the formal regression model you wrote in the previous part?

iii. Do we see evidence of "endogeneity bias"/"omitted variable bias"? Why or why not?

(e) The regression below is estimated with the same data as above:

```
regress y x1 x3
```

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   21.77
       Model |    1071.284     2     535.642           Prob > F      =  0.0000
    Residual |  2387.04242    97  24.6086847           R-squared     =  0.3098
-------------+------------------------------           Adj R-squared =  0.2955
       Total |  3458.32642    99  34.9325901           Root MSE      =  4.9607

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -2.662194   .4244046    -6.27   0.000    -3.504519   -1.819868
          x3 |   1.228192   .4908569     2.50   0.014     .2539773    2.202407
       _cons |   95.82432   3.822908    25.07   0.000      88.2369    103.4117
------------------------------------------------------------------------------
```

i. Write down the formal regression model behind the regression output above. Include the parameters to be estimated and an appropriate observation index.

ii. Suppose that x2 does have an effect on y but is not observed in the available data. Where is x2 in the formal regression model you wrote in the previous part?

iii. Do we see evidence of "endogeneity bias"/"omitted variable bias"? Why or why not?

iv. Explain the direction of any observed biases.

v. How would you expect the regression results to change if x1 and x2 were less strongly positively correlated?

vi. How would you expect the regression results to change if x1 and x2 were not correlated?

3. Consider panel (longitudinal) data for 8 firms (Firms A, B, C, D, E, F, G and H) for 10 years (1990 - 1999). It includes three quantitative variables. For example these could measure capital, employment, profits, mark-ups, advertising, sales, prices, investment, etc.

(a) `summarize y var1 var2 firm_A - firm_H yr_1990 - yr_1999`

Variable	Obs	Mean	Std. Dev.	Min	Max
y	80	2437.859	116.8473	2108.965	2745.814
var1	80	96.39869	6.467301	80.58432	109.8849
var2	80	203.7253	14.06065	173.2735	242.0598
firm_A	80	.125	.3328055	0	1
firm_B	80	.125	.3328055	0	1
firm_C	80	.125	.3328055	0	1
firm_D	80	.125	.3328055	0	1
firm_E	80	.125	.3328055	0	1
firm_F	80	.125	.3328055	0	1
firm_G	80	.125	.3328055	0	1
firm_H	80	.125	.3328055	0	1
yr_1990	80	.1	.3018928	0	1
yr_1991	80	.1	.3018928	0	1
yr_1992	80	.1	.3018928	0	1
yr_1993	80	.1	.3018928	0	1
yr_1994	80	.1	.3018928	0	1
yr_1995	80	.1	.3018928	0	1
yr_1996	80	.1	.3018928	0	1
yr_1997	80	.1	.3018928	0	1
yr_1998	80	.1	.3018928	0	1
yr_1999	80	.1	.3018928	0	1

i. Which are dummy variables?

ii. What values can the variable firm_A take? How many observations in the data will there be of each value?

iii. If you created a new variable new_var = firm_A + firm_B + firm_C + firm_D + firm_E + firm_F + firm_G + firm_H what values would it take? How many observations in the data will there be of each value?

iv. What values can the variable yr_1993 take? How many observations in the data will there be of each value?

(b) The regression below includes a full set of firm and year dummies.

```
regress y var1 var2 firm_B - firm_H yr_1991 - yr_1999
```

```
      Source |       SS       df       MS              Number of obs =      80
-------------+------------------------------           F( 18,    61) =   43.32
       Model |  1000346.55    18  55574.8082           Prob > F      =  0.0000
    Residual |   78262.584    61  1282.99318           R-squared     =  0.9274
-------------+------------------------------           Adj R-squared =  0.9060
       Total |  1078609.13    79  13653.2801           Root MSE      =  35.819

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var1 |   1.831333   .8765363     2.09   0.041     .0785905    3.584076
        var2 |   10.33615   .4141192    24.96   0.000     9.508065    11.16423
      firm_B |    79.5159   16.78651     4.74   0.000     45.94919    113.0826
      firm_C |  -159.9315   18.30179    -8.74   0.000    -196.5282   -123.3348
      firm_D |  -25.38237   16.03622    -1.58   0.119    -57.44876    6.684027
      firm_E |   144.0202    19.1368     7.53   0.000     105.7538    182.2866
      firm_F |   95.71041   16.70022     5.73   0.000     62.31625    129.1046
      firm_G |   28.27754   16.33934     1.73   0.089    -4.394986    60.95006
      firm_H |   62.22079   16.57106     3.75   0.000     29.08491    95.35667
     yr_1991 |   12.76498   18.05589     0.71   0.482       -23.34    48.86996
     yr_1992 |   .8356397    18.2818     0.05   0.964    -35.72107    37.39235
     yr_1993 |   10.11642   18.46658     0.55   0.586    -26.80978    47.04262
     yr_1994 |  -6.959535   18.76467    -0.37   0.712    -44.48181    30.56274
     yr_1995 |   9.810427   18.35784     0.53   0.595    -26.89833    46.51919
     yr_1996 |   11.45358   18.06902     0.63   0.529    -24.67766    47.58482
     yr_1997 |   23.44013   17.95578     1.31   0.197    -12.46468    59.34494
     yr_1998 |   13.14311   18.09421     0.73   0.470    -23.03849    49.32471
     yr_1999 |   15.98167   18.09208     0.88   0.381    -20.19568    52.15902
       _cons |   118.4732   114.8936     1.03   0.307    -111.2707    348.2171
------------------------------------------------------------------------------
```

i. Write down the formal regression model behind the regression output above. Include the parameters to be estimated and an appropriate observation index.

ii. Write down an alternate formal regression model that shows the inclusion of a simple time trend. Compare and contrast this with the previous specification.

iii. What is the omitted category for the firm dummies?

iv. What is the interpretation of the coefficient on the firm_B variable?

v. What is the interpretation of the coefficient on the firm_C variable?

vi. What is the relationship for Firm A in 1990 according to this regression?

vii. What is the relationship for Firm A in 1996 according to this regression?

viii. What is the relationship for Firm B in 1996 according to this regression?

ix. Does it look like it is important to control for differences across firms? Across years?

x. What does the statistically insignificance of the firm_G fixed effect mean?

(c) The regression below includes firm fixed effects, but excludes year fixed effects.

```
regress y var1 var2 firm_B - firm_H
```

```
      Source |       SS       df       MS              Number of obs =      80
-------------+------------------------------           F(  9,    70) =   92.72
       Model |   995133.62     9  110570.402           Prob > F      =  0.0000
    Residual |   83475.512    70  1192.50731           R-squared     =  0.9226
-------------+------------------------------           Adj R-squared =  0.9127
       Total |  1078609.13    79  13653.2801           Root MSE      =  34.533

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var1 |   1.458586   .7582898     1.92   0.058     -.053775    2.970947
        var2 |   10.25862   .3677995    27.89   0.000     9.525072    10.99218
      firm_B |    80.6616   16.09547     5.01   0.000      48.5602     112.763
      firm_C |  -160.5668   17.41796    -9.22   0.000    -195.3058   -125.8278
      firm_D |  -25.03298   15.45627    -1.62   0.110    -55.85955    5.793584
      firm_E |   145.8896   18.13192     8.05   0.000     109.7267    182.0526
      firm_F |   97.04291   16.01392     6.06   0.000     65.10416    128.9817
      firm_G |   29.58329   15.69643     1.88   0.064     -1.72225    60.88882
      firm_H |   63.28781   15.90963     3.98   0.000     31.55706    95.01856
       _cons |   178.4531   97.79112     1.82   0.072    -16.58509    373.4914
------------------------------------------------------------------------------
```

i. Given these results, is it reasonable to drop the year fixed effects?

ii. Can you compare the R-squared statistic across these two specifications?

iii. How would you modify the specification if you thought that the impact of var1 on y depends on the level of var2? Write down the formal regression model. What would be the interpretation of the coefficient on var1 if you implemented this?

iv. If instead, you thought that the impact of var1 and var2 on y depend on the firm, how would you modify the specification? Write down the formal regression model.

(d) The regression below excludes both the firm fixed effects and the year fixed effects.

```
regress y var1 var2
```

```
      Source |       SS       df       MS              Number of obs =      80
-------------+------------------------------           F(  2,    77) =   82.16
       Model |  734442.056     2  367221.028           Prob > F      =  0.0000
    Residual |  344167.076    77  4469.70229           R-squared     =  0.6809
-------------+------------------------------           Adj R-squared =  0.6726
       Total |  1078609.13    79  13653.2801           Root MSE      =  66.856

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var1 |   7.414375   1.221905     6.07   0.000     4.981251    9.847499
        var2 |   7.086002   .5620243    12.61   0.000     5.966869    8.205136
```

```
       _cons |   279.5247   187.9096     1.49   0.141    -94.65116    653.7005
------------------------------------------------------------------------------
```

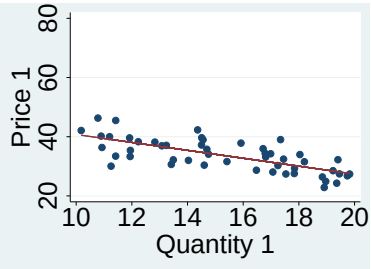
i. Why is it a terrible idea to drop the firm fixed effects?

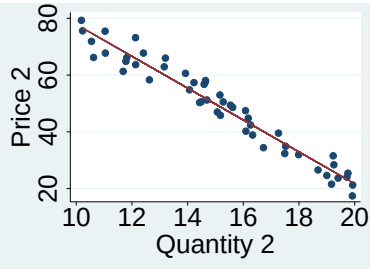
ii. Explain the biases apparent in the remaining parameter estimates.

4. TRUE/FALSE/EXPLAIN Consider data measuring the HHI and the Lerner Index for a cross section of 30 different industries characterized by oligopoly and consider the following STATA summary of that data and OLS regression. Antitrust enforcers should interpret these results to mean that a merger between two firms that increases the HHI from 0.5 to 0.8 will on average lead to an increase in the Lerner Index from 0.56 to 0.75.

```
summarize HHI Lerner_index

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         HHI |        30    .4735512    .2386455   .1053862   .8862145
Lerner_index |        30    .5470414    .2289869   .0370163   .9684148

regress Lerner_index HHI

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   21.15
       Model |  .654329277     1  .654329277           Prob > F      =  0.0001
    Residual |  .866285143    28  .030938755           R-squared     =  0.4303
-------------+------------------------------           Adj R-squared =  0.4100
       Total |  1.52061442    29   .05243498           Root MSE      =  .17589

------------------------------------------------------------------------------
Lerner_index |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         HHI |   .6294278   .1368671     4.60   0.000     .3490682    .9097874
       _cons |   .2489751   .0723332     3.44   0.002     .1008073     .397143
------------------------------------------------------------------------------
```

5. TRUE/FALSE/EXPLAIN Consider two unrelated goods: Good 1 and Good 2. For each good suppose you collect cross sectional data (different geographic areas) on price and quantity that is summarized in the following scatter diagrams and OLS regression lines. We *cannot* infer that the demand for Good 2 is less elastic than the demand for Good 1.





6. Suppose you attempted to estimate demand for satellite TV using 200 observations of different local areas (cross-sectional data). You specified the following linear functional form:

$$Q_i = \alpha + \beta p\_sat_i + \delta pop_i + \phi p\_cab_i + \gamma ave\_inc_i + \lambda urban_i + \varepsilon_i$$

where Q is the number of households in local area subscribed to satellite TV in 100's, *pop* is the population of local area in 1,000's of households, *p_cab* is the average monthly price of cable in local area in dollars, *p_sat* is the average monthly price of satellite TV in local area in dollars, *ave_inc* is the average income in the local area in 1,000's of dollars, and *urban* is a dummy variable = 1 if local area is urban and = 0 otherwise. Here are descriptive statistics and the regression results (from STATA software):

```
regress Q p_sat p_cab pop ave_inc urban
```

```
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  5,   194) =   54.93
       Model |  104613.999     5  20922.7999           Prob > F      =  0.0000
    Residual |  73896.7253   194  380.910955           R-squared     =  0.5860
-------------+------------------------------           Adj R-squared =  0.5754
       Total |  178510.725   199  897.038817           Root MSE      =  19.517

------------------------------------------------------------------------------
           Q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       p_sat |  -1.964699   .4947822    -3.97   0.000    -2.940542   -.9888566
       p_cab |   .9463366   .3392216     2.79   0.006     .2773008    1.615372
         pop |   .3653705   .0249318    14.65   0.000     .3161984    .4145426
     ave_inc |   1.150198   .4675301     2.46   0.015     .2281035    2.072292
       urban |   22.34265   3.094329     7.22   0.000     16.23981     28.4455
       _cons |    190.472   35.88504     5.31   0.000     119.6971    261.2469
------------------------------------------------------------------------------
```

```
summarize Q p_sat p_cab pop ave_inc urban
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           Q |       200    253.7439    29.95061     150.584   323.4377
       p_sat |       200    40.04437    2.830241    32.44048   47.93978
       p_cab |       200    50.22235    4.177157    39.61116   61.76898
         pop |       200    98.84983    56.10408    .5190096   199.8317
     ave_inc |       200    45.25052    2.982745     37.4399   54.58408
       urban |       200         .28    .4501256           0          1
```

(a) For a particular local area, determine the price and quantity for satellite TV if competition were perfect and cost of an additional satellite subscriber were \$10 per month. This particular local area has a population of 100,000 households, an average cable price of \$50 per month, an average income of \$40,000, and is not an urban area. Make your calculation based on the estimation results rounding the parameter estimates to the nearest tenth (0.1). Indicate the units of price and quantity with your answer.

(b) Calculate the loss of consumer surplus per month (in \$) in the local area described in part (a) if the price of satellite TV were set at \$40 per month rather than the perfectly

competitive price.

(c) Calculate the elasticity of demand at a price of \$60 and indicate whether demand is elastic or inelastic.

(d) When estimating the demand parameters with OLS, which variables are endogenous in the above specification of demand? For each variable you identify as endogenous, explain why it will be endogenous.

7. Consider a peer-reviewed academic journal article "Selling a cheaper mousetrap: Wal-Mart's effect on retail prices" published in the *Journal of Urban Economics* in 2005 by Emek Basker. Here is the abstract from Basker (2005):

> I quantify the price effect of a low-cost entrant on retail prices using a case-study approach. I consider the effect of Wal-Mart entry on average city-level prices of various consumer goods by exploiting variation in the timing of store entry. The analysis combines two unique data sets, one containing opening dates of all US Wal-Mart stores and the other containing average quarterly retail prices of several narrowly-defined commonly-purchased goods over the period 1982-2002. I focus on 10 specific items likely to be sold at Wal-Mart stores and analyze their price dynamics in 165 US cities before and after Wal-Mart entry. An instrumental-variables specification corrects for measurement error in Wal-Mart entry dates. I find robust price effects for several products, including shampoo, toothpaste, and laundry detergent; magnitudes vary by product and specification, but generally range from 1.5-3% in the short run to four times as much in the long run.

As indicated by the abstract, Basker (2005) seeks to estimate how much retail prices change in local markets (cites) after Wal-Mart enters (i.e. opens a store). For each quarter from 1982 - 2002 the data tracks prices of 10 specific products such as 11oz bottle of Johnson's Baby shampoo and 100-tablet bottle of Bayer brand aspirin for 165 different cities in the U.S. During the sample period, Wal-Mart entered many of those cities (25 already had a Wal-Mart at the start of the sample period).

(a) To address the research question in a simplistic way, describe the most basic OLS (Ordinary Least Squares) estimation approach and explain why it would suffer from an endogeneity bias. Make sure your answer includes equations showing the key parts of your empirical model (including observation indices and parameters) with any variables that you have defined. Some, but not all, of the variables you will need are: AVE_P1_{it} the average price of product 1 in market i in quarter t, AVE_P2_{it} the average price of product 2 in market i in quarter t, ... AVE_P10_{it}.

(b) What kind of data is described: time series, cross-sectional or panel data? Could fixed effects for each city and fixed effects for each quarter be added to your empirical specification in part (a)? Explain.

(c) The OLS regression model Basker (2005) presents on page 211 is a version of the solution to part (a) – please review the solution to part (a) before going on – but instead of just one RHS (right-hand side) variable, the regression also includes city dummies and quarter

dummies. Should the inclusion of these variables *help* address the endogeneity issue with the Wal-Mart dummy variable?

8. Write an empirical model of demand for a homogeneous good that would capture the following.

   A. It has a constant elasticity functional form.
   B. It will be estimated using panel data with variation over time and across markets that contain quantities, prices and some demand shifters.
   C. National advertising is an important demand shifter but it is unobserved. However, you know that it only varies over time and not across markets.
   D. The elasticity of demand may depend on the age of the head of household.

   Explain how your specification addresses each requirement: A, B, C and D.

9. Consider a peer-reviewed academic journal article "The Effect of File Sharing on Record Sales: An Empirical Analysis" published in the *Journal of Political Economy* in 2007 by Felix Oberholzer-Gee and Koleman Strumpf. Here is the abstract from Oberholzer-Gee and Strumpf (2007):

   > For industries ranging from software to pharmaceuticals and entertainment, there is an intense debate about the appropriate level of protection for intellectual property. The Internet provides a natural crucible to assess the implications of reduced protection because it drastically lowers the cost of copying information. In this paper, we analyze whether file sharing has reduced the legal sales of music. While this question is receiving considerable attention in academia, industry, and Congress, we are the first to study the phenomenon employing data on actual downloads of music files.We match an extensive sample of downloads to U.S. sales data for a large number of albums. To establish causality, we instrument for downloads using data on international school holidays. Downloads have an effect on sales that is statistically indistinguishable from zero. Our estimates are inconsistent with claims that file sharing is the primary reason for the decline in music sales during our study period.

   Further, consider this excerpt from page 12 of Oberholzer-Gee and Strumpf (2007):

   > We observe sales and downloads at the album-week level for 17 weeks. These panel data allow us to estimate a model with album fixed effects,
   >
   > $$S_{it} = X_{it}\beta + \gamma D_{it} + \omega_s t^s + \nu_i + \mu_{it}$$
   >
   > where i indicates the album, t denotes the time in weeks, S_{it} is observed sales, X_{it} is a vector of time-varying album characteristics that includes a measure of the title's popularity in the United States. D_{it} is the number of downloads for all songs on an album, and ω_s controls for time trends (a flexible polynomial or week fixed effects).

   (a) What is the primary question the paper seeks to answer? Which parameter in the above equation is directly relevant answering the question?

(b) What is ν_i? Why is it included in the equation?

(c) Oberholzer-Gee and Strumpf (2007) address the previous question on pages 12 - 14:

> The key concern in our empirical work is that the number of downloads is likely to be correlated with unobserved album-level heterogeneity. As the descriptive statistics suggest, the popularity of an album is likely to drive both file sharing and sales, implying that the parameter of interest γ will be estimated with a positive bias. The album fixed effects ν_i control for some aspects of popularity, but only imperfectly so because the popularity of many releases in our sample changes quite dramatically during the study period.

Hence, do they think that the inclusion of fixed effects will be a sufficient solution to the endogeneity problem?