ECO410H: Practice Questions 5 – SOLUTIONS

1. Make sure you can identify and *explain* the following in the STATA output: the sample size ($n = 32$), the F statistic ($F = 20.16$), the P-value for the F test of the overall statistical significance of the regression (P-value = 0.0001), the R^2 ($R^2 = 0.4019$), the slope coefficient estimate ($\hat{\beta} = 0.31$), the intercept coefficient estimate ($\hat{\alpha} = 6.05$), the standard error of the slope coefficient estimate ($s.e.(\hat{\beta}) = 0.07$), the t statistic for the statistical significance of the slope coefficient estimate ($t = 4.49$), the P-value for the statistical significance of the slope coefficient estimate (P-value < 0.0005). If you have trouble *explaining* these, please review your notes/textbook from your second-year statistics course and work through these practice questions, which helps you identify what you really need to remember and understand from that course.

2. (a) i. 100 observations
   ii. 4 variables
   iii. y has the largest standard deviation and range, but these comparisons depend on units of measurement. The coefficient of variation (cv = standard deviation / mean) is a unit-free measure of variability. According to the coefficient of variation, x3 is the most variable.
   iv. Absolutely not

   (b) i. x1 and x2 are strongly positively correlated with each other. The coefficient of correlation ranges between -1 (perfect negative correlation) to 0 (no correlation) to 1 (perfect positive correlation). A value of 0.8704 indicates that x1 and x2 are strongly positively correlated. There is a weak positive correlation between x3 and y. There is a very weak positive correlations between x1 and x3 and x2 and x3: they are basically uncorrelated.
   ii. y and x2 are strongly negatively correlated with a coefficient of correlation equal to -0.7225. y and x1 are moderately negatively correlated.

   (c) i. $y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \beta_3 x3_i + \varepsilon_i$
   ii. It is 3 because there are 3 independent (explanatory) variables: x1, x2, and x3.
   iii. R-squared is 0.5994, which means that 59.94% of the variation in y is explained by variation in x1, x2, and x3.
   iv. To test the overall statistical significance of the model, do an F test. The F statistic is 47.88 with numerator degrees of freedom of 3 and denominator degrees of freedom of 96, which we could compare to the critical value found from a statistical table. However, there is no need because STATA also reports the P-value of this test: Prob $>$ F = 0.0000. Hence, we would clearly reject the null hypothesis that the model has no explanatory power (i.e. that all of the slope coefficients are zero) and conclude that this linear regression model is statistically significant. (A low P-value like this means that it is virtually impossible that we could have observed results like these for a random sample of data if the null hypothesis were true.)
   v. The Root MSE is 3.7989, which reflects the importance of the unobservables (ε) in determining y. (You may have learned it as the "standard error of estimate" or the

"standard deviation of the residuals.") The larger the variability of the error term relative to the variability of the dependent variable, the more the dependent variable is being determined by the unobservables. Hence, a relatively large Root MSE means that our model is failing to explain much about the dependent variable: most of explanation lies with the error. We see that relative to the s.d. of y, which is 5.91, the variance of the error is quite large. Of course this fact is conveniently summarized in the R-squared statistic already discussed.

vi. All three. t statistics in each case are greater than 2 in absolute value: a quick rule of thumb useful when a Student t table is not handy.

vii. No. The t statistic would be -1.27 ($= \frac{2.158-3}{0.664}$), which does not pass the rule of thumb. Hence we have insufficient evidence in these data to reject the hypothesis that the true slope parameter on x1 is 3 ($H_0 : \beta_3 = 3$).

viii. There is no bias. Correlations are pairwise: this means that we look at the correlations between two variables *without* holding any other variables constant. The advantage of the multiple regression analysis is that it allows estimation of the marginal effect of each variable while *holding the other variables constant.* In this particular case, once we controlled for the separate effects of x2 and x3, the marginal effect of an increase in x1 on y is actually positive and not negative.

ix. A one unit increase in x1 on average results in a 2.2 unit increase in y holding x2 and x3 constant. A one unit increase in x2 on average results in a 5.7 unit decrease in y holding x1 and x3 constant. A one unit increase in x3 on average results in a 0.9 unit increase in y holding x1 and x2 constant. These are causal interpretations because we were told to assume the underlying conditions hold, which includes the assumption that the right-hand-side variables (x1, x2, and x3) are exogenous: i.e. not correlated with ε, which represents unobserved variables.

x. Plugging in we obtain $\hat{y} = 103.6967 + 2.158004 * x1 - 5.669571 * (7.881578) + 0.9049954 * (4.119784)$, which simplifies to $\hat{y} = 62.73992 + 2.158004 * x1$.

xi. This just shifts the relationship: $\hat{y} = 50.531639 + 2.158004 * x1$.

(d) i. $y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \varepsilon_i$

ii. It will be incorporated into the error term: ε_i.

iii. We do not see evidence. The coefficients on x1 and x2 do not (statistically) differ from those estimated in the original regression that included x3. The reason is that x3 is uncorrelated with x1 and x2. Hence, omitting x3 does not result in omitted variable bias because it does not cause a violation of the underlying assumption that the explanatory variables and the error term are uncorrelated. Omitted variables only cause bias when they are correlated with included explanatory variables. I prefer the term "endogeneity bias" because sometimes students get confused by the term "omitted variable bias" (which does *not* arise simple because a relevant variable is omitted).

(e) i. $y_i = \beta_0 + \beta_1 x1_i + \beta_3 x3_i + \varepsilon_i$

ii. It will be incorporated into the error term: ε_i.

iii. We see an extreme case of endogeneity bias (omitted variable bias). x2 is strongly positively correlated with x1. Hence when x2 is left in the error term it creates a

violation of the assumption that the x variables are exogenous (uncorrelated with the error). However, in this specification x1 is endogenous. The coefficient on x1 went from positive and statistically significant to negative and statistically significant. A huge bias.

iv. In this case x1 will tend to be negatively correlated with the error term, which leads the coefficient on x1 to be biased downward (in the negative direction). The reason is that x1 and x2 are positively correlated. However, the coefficient on x2 is negative (we learned this from the original full regression where it was stated that there are no violations of the assumptions). Hence when x1 is above average the error tends to be negative (below the average error of 0) because the error in this model includes x2. Of course a negative error leads to a small y. Hence the regression model incorrectly attributes the low values of y to the high values of x1 when really the low values of y are being caused by the error that tends to be negative when x1 is high. Given the positive correlation between x1 and x2, x1 is getting blamed for the negative effect of x2 because this model has failed to control for the effect of x2.

v. This would tend to lead to a smaller omitted variable bias. We would expect to still observe a coefficient estimate for x1 that was biased downward, but the bias would be less severe. We may get a positive estimate, however it would still be too small.

vi. This would remove the violation of exogeneity assumption and hence remove the endogeneity bias. We would expect to see a coefficient estimate statistically near 2.15 from the original unbiased regression.

3. (a) i. $y_{it} = \beta_0 + \beta_1 var1_{it} + \beta_2 var2_{it} + \delta_i + \xi_t + \varepsilon_{it}$. The it observation index shows the data are panel data. The δ_i term is common notation for showing that a fixed effect (dummy) is included for each firm (the i index). Similarly, ξ_t shows that a fixed effect for each year (the t index) is included.

ii. $y_{it} = \beta_0 + \beta_1 var1_{it} + \beta_2 var2_{it} + \delta_i + \gamma t + \varepsilon_{it}$. Notice how there is no subscript on γ, which indicates that a single constant time trend slope will be estimated, which is less flexible than including a separate fixed effect for each year. In contrast, including ξ_t allows things to go up and down over time at differing rates, which makes it more flexible (but less parsimonious).

iii. firm_A - firm_H and yr_1990 - yr_1999 are the dummy variables.

iv. It can take two values: 0 or 1. It will be 1 for all observations related to firm_A and 0 otherwise. Given that the question stated that each firm was followed for 10 years there will be 10 ones and 70 zeros for this variable.

v. It would be 1 for all observations in the data. Each observation will be associated with one, and only one firm. Hence only one firm dummy variable will be turned on for each row of the data (each observation).

vi. It can take two values: 0 or 1. It will be 1 for all observations from year 1993 and 0 otherwise. Given that the question stated that there are 8 firms there will be 8 ones and 72 zeros for this variable.

(b) i. Firm A

ii. The dependent variable y tends to be 79.5159 units bigger for Firm B compared to Firm A.

iii. The dependent variable y tends to be 159.9315 units smaller for Firm C compared to Firm A.

iv. $\hat{y} = 118.4732 + 1.831333var1 + 10.33615var2$

v. Plugging in $\hat{y} = 118.4732 + 11.45358 + 1.831333var1 + 10.33615var2$, so relationship is $\hat{y} = 129.93 + 1.83var1 + 10.34var2$.

vi. Plugging in $\hat{y} = 118.4732 + 79.5159 + 11.45358 + 1.831333var1 + 10.33615var2$, so relationship is $\hat{y} = 209.44 + 1.83var1 + 10.34var2$.

vii. There are clearly large firm fixed effects, which means that it is important to control for these. There do not appear to be substantial year fixed effects and none of the coefficients on these are statistically significant. While technically one could do a joint F test of the significance of the year dummies it does not seem unreasonable to exclude them given their small magnitude and statistical insignificance. Of course if you had more information about this problem that should be used as well: if for example there was reason to believe that things were changing over time then you might consider the decision to drop them more carefully.

viii. The data provide insufficient evidence to reject the notion that firms A and G have the same constant term.

(c) i. Yes. The other parameter estimates are statistically unchanged and the overall fit of the model does not suffer. Note: It is very important that the coefficients on var1 and var2 are statistically unchanged. The reason fixed effects are included is to control for factors – in this case, anything that is changing over time across all firms – that may be correlated with the included explanatory variables. For example, if var1 is advertising this may well be changing over time due to macroeconomic factors. If these macroeconomic factors also affect y then failing to control for them with these year fixed effects will cause the coefficient estimate for var1 to be biased.

ii. Yes, because the dependent variable is the same in both models a comparison is reasonable. Here we see that about 92% of the variation in y is explained whether or not we include the time dummies. Hence, these time dummies are not adding to our model and we could drop them.

iii. Add an interaction terms between var1 and var2. $y_{it} = \beta_0 + \beta_1 var1_{it} + \beta_2 var2_{it} + \beta_3(var1_{it} * var2_{it}) + \delta_i + \varepsilon_{it}$. The interpretation of the coefficient on var1 would now be the effect of var1 on y when the level of var2 is zero.

iv. Add interaction terms between var1 and the firm dummies and between var2 and the firm dummies. This would be a total 14 new variables. However, it doesn't make writing the formal model cumbersome because we can just tack a subscript on these parameters to indicated that each is estimated separately for each firm: $y_{it} = \beta_0 + \beta_{1i} var1_{it} + \beta_{2i} var2_{it} + \delta_i + \varepsilon_{it}$.

(d) i. Clearly we see that the coefficients on var1 and var2 have been affected. By moving any firm-specific effects into the error it seems we have created an endogeneity bias: it looks like var1 and var2 also determined by firm-specific factors.

ii. Firm fixed effects are often included in empirical IO models because often there are differences across firms that are hard to measure. Factors such as management skills, corporate philosophy, work conditions, technology, etc. are difficult to fully capture

in any objective measure. Hence these would end up in the error term. However this is undesirable because these unobserved firm-specific factors may also affect the observed choices and behaviors of the firm we are trying to model. In other words, these unobserved factors often make our observed explanatory variable endogenous. One solution is to get these firm-specific factors out of the error term by controlling for them. This is easily done through the inclusion of firm fixed effects (i.e. dummies). In the above case it appears as if firms are choosing levels of var1 and var2 based on some of the unobserved factors. Hence when we've excluded the dummies we've created an endogeneity bias / omitted variables bias in the remaining parameter estimates.

4. FALSE. One cannot determine causality from this analysis. What this shows is only that the HHI and the Lerner Index are positively correlated. The HHI in an industry is NOT exogenously given. There are many factors in the regression error term – such as demand conditions, advertising, entry barriers – that affect both the HHI and market power, which means the HHI is endogenous. This simple regression does not even attempt to hold constant *anything* that varies across these 30 industries and attributes all differences in market power to differences in the HHI. Hence this regression is useless in helping to set antitrust policy.

5. TRUE. If the red lines were in fact the demand curves then we could say that demand for Good 2 is in general steeper (less elastic) than the demand for Good 1. (Remembering of course that elasticity is not constant along a linear demand.) However, the red lines are almost certainly not reasonable estimates of the demand for Goods 1 and 2. The problem is the usual one of endogeneity. These regression lines do not take into account ANY demand shifters. However demand shifters affect both equilibrium price and quantity. What we are observing in the above graphs are the intersection points between supply and demand in the different markets. We are NOT observing an experiment that shows OTHER THINGS EQUAL what would happen to quantity demanded if the price EXOGENOUSLY changed. If that were what we are observing then the red lines would be demand curves. But what we actually have are simply equilibrium prices and quantities for a cross-section of markets.

6. (a) Plugging in obtain:

$$Q = -2.0 * P + 0.4 * 100 + 0.9 * 50 + 1.2 * 40 + 22.3 * 0 + 187.0$$

$$Q = -2.0 * P + 40 + 45 + 48 + 0 + 187$$

$$Q = 320 - 2P$$

Write inverse demand: $P = 160 - 1/2Q$. Set P = mc to obtain perfectly competitive outcome: $160 - 1/2 * Q = 10$, $Q = 300$. So perfect competition would predict that P=\$10 per month and 30,000 households would be satellite TV subscribers in this particular local area.

(b) CS at P=\$10 per month: CS = 1/2 * 30,000 * (160-10) = \$2,250,000/month. CS at P=\$40 per month: CS = 1/2 * 24,000 * (160-40) = \$1,440,000/month. So, the price increase from \$10 to \$40 per month for satellite TV causes a \$810,000 loss in monthly consumer surplus.

(c) $\frac{\partial Q}{\partial P}\frac{P}{Q} = -2(60/200) = -0.6$. So demand is inelastic at a price of \$60.

(d) The right-hand-side variables *p_cab*, *ave_inc*, *city_dum*, and *pop* are standard demand shifters. (The price of cable is the price of a substitute good.) The error term in the regression includes the unobserved demand shifters. It is often reasonable to assume that the observed and unobserved demand shifters are not correlated. The clear problem is the price variable: this will certainly be endogenous as price is definitely correlated with all demand shifters including those in the error term.

7. (a) The key independent variable (right hand side variable, explanatory variable):

$WALMART_{it}$: dummy variable = 1 if Wal-Mart is in market i in quarter t and = 0 otherwise.

The most basic empirical specification would be a system of 10 equations (one for each of the different products):

$$AVE\_P1_{it} = \alpha_1 + \beta_1 * WALMART_{it} + \varepsilon_{1it}$$

$$AVE\_P2_{it} = \alpha_2 + \beta_2 * WALMART_{it} + \varepsilon_{2it}$$

...

$$AVE\_P10_{it} = \alpha_10 + \beta_10 * WALMART_{it} + \varepsilon_{10it}$$

The answer to the research question hinges on the estimation of the beta coefficients. If Wal-Mart's presence results in lower prices for a particular product then we would expect beta to be negative for that product. The endogeneity problem is that Wal-Mart does not enter a random sample of markets. Wal-Mart chooses which markets it will enter. This decision will be driven by conditions in that market and those local market conditions will also affect prices. The key problem is that there are factors in the error that affect BOTH average prices and Wal-Mart's entry decision. This will lead to bias in the beta parameter estimates.

(b) These data are panel (longitudinal) because there is variation across cities and over time. Yes, you could add city and quarter fixed effects because you would still have variation across time to estimate the parameter on WALMART. There would be 164 city dummies and 83 quarter dummies. That may sound like a lot of variables on the RHS (a huge multiple regression) but there are a large number of observations in the data (165 cities times 84 quarters = 13,860 observations for each of the 10 products) and this means all of these parameters can be estimated.

(c) Yes, absolutely those dummies should *help* as now any unobserved variables that are city or quarter specific that may affect Wal-Mart's entry decision would no longer be in the error term: they would be controlled for by the included dummy variables. Fixed effects (i.e. dummy variables, indicator variables) are often used to help address an endogeneity problem. In this example, it gets out of the error term anything that is city-specific or quarter-specific and correlated with Wal-Mart's decision to enter.

8.

$$ln(Q_{jt}) = \alpha + \beta ln(P_{jt}) + \lambda Age * ln(P_{jt}) + \phi_1 X_1 + ... + \phi_k X_k + \delta_t + \varepsilon_{jt}$$

A. The logarithms on quantity and price imply the constant elasticity functional form. B. The subscripts correctly indicate that panel data will be used. The X's represent the demand shifters mentioned. C. Since national advertising is an important demand shifter that is not in our data and because it varies only over time and not across markets we can include time dummies to capture it (and everything else that varies only over time). These are represented by δ_t in the above equation. D. If the elasticity of demand may depend on the age of the head of household this suggests that we need to include an interaction term between age and the logarithm of price. Hence the elasticity would be $(\beta + \lambda Age)$.

9. (a) The paper seeks to answer the question: how much has file sharing affected sales of records? The parameter in the equation that is directly relevant to the answer is γ. It measures how much downloading has impacted sales of records.

   (b) The term ν_i represents a full set of album fixed effects (save one for the omitted category). The authors included all of these dummy variables to try to control for some of the unobserved characteristics of albums. However, this does not completely fix the endogeneity problem because an album's popularity is not constant over time.

   (c) No, they argue that fixed effects are not a sufficient solution for the bias caused by downloads being correlated with the error term. In fact, we know from the abstract that they use an instrumental variable (IV) approach to address endogeneity. On page 14 Oberholzer-Gee and Strumpf (2007) explain:

> We address this issue by instrumenting for D_{it} in a two-stage least-squares model. Valid instruments Z_{it} predict file sharing but are uncorrelated with the second-stage error μ_{it}. As in the differentiated products literature, where the problem is correlation between prices and unobserved product quality, we use cost shifters to break the link between unobserved popularity, downloads, and sales.