Homework 2

Jacob Puthipiroj

February 18, 2020

1 Heterogeneity in Returns to Schooling

Using the link, download the data used in Koop and Tobias's [1] study of the relationship between wages and education, ability, and family characteristics. Their data set is a panel of 2,178 individuals with a total of 17,919 observations. Extract the first observations for the first 15 individuals in the sample.

Let X_1 be a vector of education, experience, and ability (the individual's own characteristics). Let X_2 contain the mother's education, the father's education, and the number of siblings (the household characteristics). Let y be the log wage.

a. **Compute the least squares regression coefficients in the regression of y on X_1. Report and interpret the coefficients.**

b. The full regression table is available in the Appendix. The coefficients for *educ*, *potexp* and *ability* were all significant, and are estimated to be $\beta_{educ} = 0.0737621, \beta_{potexp} = 0.0394896$, and $\beta_{ability} = 0.0828907$ respectively.

   Given the log-linear form of the model, the coefficients can be interpreted as relative increases in the dependent variable given a unit increase in the independent variable: A 1 unit increase in education (in years), potential experience (parameterized as age - education - 5), or cognitive ability (proxied by a standardized ASVAB test score), would be expected to increase wages by 7.37% , 3.95% and 8.29% respectively, keeping all else equal.

c. **Compute the least squares regression coefficients in the regression of y on X_1 and X_2. Report and interpret the coefficients.**

   The full regression table is available in the Appendix. All coefficients except for $\beta_{mothered}$ were found to be significant at the $\alpha = 0.05$ level, at $\beta_{educ} = .0722035, \beta_{potexp} = .0395093, \beta_{ability} = 0.0774678, \beta_{fathered} = 0.0054569$ and $\beta_{siblings} = 0.0047656$ respectively. The coefficient estimate $\beta_{mothered} = -0.000117$ was not found to be statistically significant from 0.

   As above, a 1 unit increase in education, potential experience, cognitive ability, father's education (in years), and number of siblings, were expected to increase wages by 7.22%, 3.95%, 7.75%, 0.55%, and 0.48% respectively. Thus a fraction of the explanatory power of *educ* and *ability* were given to *fathered* and *siblings*, but it is unclear if this small statistical significance would translate into practical significance for decision-making.

d. **Compute the R^2 for the the regression of y on X_1 and X_2 manually using the SSE and SST from the output. Repeat the computation for the case in which the constant term is omitted. You need use the *noconstant* option, which suppresses the constant in a regression model. What happens to R^2?**

From the Stata output, we obtain SSE = 4126.17535 and SST = 4999.72601. The R^2 formula gives

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{4126.17535}{4999.72601} = 0.17471$$

In the no-intercept model, SSE = 4316.7689, while the SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$ remains the same as before.

$$R_0^2 = 1 - \frac{4316.76885}{4999.72601} = 0.13659$$

Thus R^2 is reduced by excluding the intercept term. A notable discrepancy is the given SST from an automatic, rather than manual, calculation in Stata. An incorrect value of SST = 99529.3551 is given, as a result of SST being implicitly calculated assuming an intercept exists.

e. **Compute the adjusted R^2 for the full regression with and without the constant term. Interpret your results. Do we need the constant term? (Hint: Make sure to refer to the economic theory to discuss whether one should have the constant term regardless of statistical significance)**

With the constant term, there are $n = 17919$ observations and $p = 6$, and the adjusted R^2 is

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - (1 - 0.17471)\frac{17919}{17919 - 6 - 1} = 0.17438$$

And similarly in the no-intercept model, we have

$$\bar{R}_0^2 = 1 - (1 - 0.13659)\frac{17919}{17919 - 6 - 1} = 0.13625$$

In this particular case, the intercept term is statistically significant in the full regression, and helps to explain more of the variance in the dependent variable. More generally, the intercept term helps in reducing some of the restrictions in the model. By including the intercept term, the mean of the least squares residual is always zero. This is assumption MR2 of the multiple regression model, and without the intercept term, MR2 is not guaranteed.

f. **Are any of the classical assumptions violated in part a or part b? Refer to the assumptions MR1, MR2, MR5, and MR6.**

MR1 concerns model specification, and could be violated in the case of omitted variable bias. Running the Ramsay RESET test on model in part a, and then b, gives a $p = 0.019 < \alpha = 0.05$ and $p = 0.0056 < \alpha = 0.05$ respectively. Thus there is likely some misspecification. MR2 and MR6 can be checked simultaneously via a residuals plot; and both seem to be normally distributed around 0. MR5 - multicollinearity can be checked via VIF, with a mean VIF of 1.3 and 1.54 respectively, indicating that there is not a problematic amount of multicollinearity.

2 The U.S. Gasoline Market

a. **Compute the multiple regression of per capita consumption of gasoline on per capita income, the price of gasoline, all the other prices and a time trend. Report all results. Do the signs of the estimates agree with your expectations?**

The signs of the estimates make sense. A rise in (disposable) income, for example, is unsurprisingly correlated with increased demand and therefore consumption. An increase in prices means a reduction in consumption. A positive sign for the price index of new cars could make sense; newer cars are more fuel efficient, so an increase in their prices incentivizes people to use their used cars, which requires more gasoline consumption. An increase in the price of used cars means people switch to newer, more fuel efficient cars, and thus consume less gasoline. An increase in the price of public transport means people switch to private cars and use more gasoline. I had no expectations on the impact of durables and non-durables. The final coefficient for p_s makes sense: a more expensive services sector means it is more expensive to use gasoline for cars, so the negative sign makes sense.

b. **Test the hypothesis that at least in regard to demand for gasoline, consumers do not differentiate between changes in the prices of new and used cars.**

Here, the null and alternative hypotheses are:

$$H_0 : \frac{\partial demand}{\partial P_{nc}} = \frac{\partial demand}{\partial P_{uc}}, \quad H_A : \frac{\partial demand}{\partial P_{nc}} \neq \frac{\partial demand}{\partial P_{uc}}$$

c. **Estimate the own price elasticity of demand, the income elasticity, and the cross-price elasticity with respect to changes in the price of public transportation. Do the computations at the 2004 point in the data, which means that the covariates should take the values corresponding to 2004 (i.e., use the "if" command instead of "at").**

The tables for the elasticities are given in the appendix. The estimated proportional elasticities of gasoline consumption with respect to proportion changes in price, income, and price of public transportation in 2004 can also be computed manually via:

$$\begin{aligned}
\text{PED}_{2004} &= \frac{\partial Q}{\partial P}\frac{P_{2004}}{Q_{2004}} = \beta_{\text{GasP}}\left(\frac{\text{GasP}_{2004}}{\text{Demand}_{2004}}\right) = (-.0110838)\frac{123.901}{6.164056} = -0.2225 \\
\text{YED}_{2004} &= \frac{\partial Q}{\partial I}\frac{I_{2004}}{Q_{2004}} = \beta_{\text{Income}}\left(\frac{\text{Income}_{2004}}{\text{Demand}_{2004}}\right) = (0.0002157)\frac{27113}{6.164056} = 0.9488 \\
\text{XED}_{2004} &= \frac{\partial Q}{\partial X}\frac{X_{2004}}{Q_{2004}} = \beta_{ppt}\left(\frac{ppt_{2004}}{\text{Demand}_{2004}}\right) = (0.0069073)\frac{209.1}{6.164056} = 0.2343
\end{aligned}$$

This matches the computer-generated elasticities.

d. **Reestimate the regression in logarithms so that the coefficients are direct estimates of the elasticities. (Do not use the log of the time trend). How do your estimates compare with the results in the previous question? Which specification do you prefer?**

Elasticities are closely linked with partial derivatives: they are understood to be the partial derivatives multiplied by the corresponding unit factor. When using a log-linear model, the relative change of the dependent variable with respect to a unit change in one of the independent variables (ceteris paribus), is roughly equivalent to the partial deriative, which can similarly be used to estimate the elasticities.

$$\begin{aligned}
&\log \text{consumption} = \beta_0 + \beta_{\text{GasP}}\text{GasP} + \beta_{\text{income}}\text{Income} + \beta_{\text{ppt}}\text{PPT} + \cdots \\
\iff &\text{consumption} = \exp(\beta_0 + \beta_{\text{GasP}}\text{GasP} + \beta_{\text{income}}\text{Income} + \beta_{\text{ppt}}\text{PPT} + \cdots) \\
\iff &\beta_{\text{GasP}} = -.0012406 \approx \frac{\partial Q}{\partial P}, \quad \beta_{\text{Income}} = .0000375 \approx \frac{\partial Q}{\partial I}, \quad \beta_{\text{PPT}} = .0019642 \approx \frac{\partial Q}{\partial X} \\
\iff &\text{PED}_{2004} = -0.02493, \quad \text{YED}_{2004} = 0.1649, \quad \text{XED}_{2004} = 0.06663
\end{aligned}$$

The tables for the regression output is given in the appendix. The corresponding elasticities obtained from these coefficient estimates follow the same sign as the elasticities obtained in c, though with significantly less magnitude.

- The price elasticity of demand for gasoline is negative but greater than -1 in both c and d, and so is inelastic in both models, but is closer to 0 in and hence even less elastic (more inelastic) in d.
- The income elasticity of demand for gasoline is positive and less than 1 in both c. and d. –as would be expected of a necessity good– but is closer to 0, and therefore more of a necessity good in d.
- The cross-price elasticity with respect to the price of transport is positive in both c. and d. –as would be expected of substitutes– but is closer to 0 in d, suggesting a weaker relationship between the price of oil and the price of public transport as substitutes.

Therefore the model in d. generally posits weaker relationships between the variables than in c. I would favor the specification in d. because of its greater conformity to the assumption of homoskedasticity of the multiple linear regression model. By running the Breusch-Pagan test for heteroskedasticity on the residuals of both models, the residuals of the log-linear model in d. was not found to be significant for heteroskedasticity at the $\alpha = 0.05$ level ($p = 0.0716 > 0.05$), while the linear model in c. was significant ($p = 0.0248 < 0.05$).

e. **Compute the simple correlations of the price variables. Would you conclude that multicollinearity is a "problem" for the regression in part a or part d?**

Looking at the correlation table in the appendix, there clearly is a high degree of collinearity in between the variables, with multiple pairwise correlations being above 0.9. While there is no exact collinearity, and calculations can be done, the model will have trouble allocating explanatory power amongst the independent variables.

f. **Notice that the price index for gasoline is normalized to 100 in 2000, whereas the other price indices are anchored at 1983 (roughly). If you were to renormalize the indices so that they were all 100.00 in 2004, then how would the results of the regression in part a change? How would the results of the regression in part d change?**

Price indices represent relative changes with respect to a given base year. In order to renormalize the index to a different year, divide each data point by the price of the desired base year, and multiply by 100. For example, the new index number for the price of public transport for the year 1983 (the previous base year) when using 2004 as the new base is:

$$ppt^*_{1983} = \frac{ppt_{1983}}{ppt_{2004}} \times 100 = \frac{100}{209.1} = 47.824$$

Performing the same regression in part a. using the newly normalized index gives similar results. The R^2, both normal and adjusted are identical. There is no effect on the fit of the regression, and the coefficient estimates of the other explanatory variables (income, year, and intercept) remain identical. The only difference –that of the coefficient estimates of the correpsonding price index variables– comes from an inverse proportional scaling of the original estimates in a. by the price index in 2004. For example, in the coefficient estimate for the price of public transport is originally $\beta_{ppt} = 0.0069073$ in a., and when using 2004 prices as the base, becomes

$$\beta^*_{ppt} = \beta_{ppt} \times \frac{ppt_{2004}}{100} = (0.0069073) \times \frac{(209.1)}{100} = 0.0144431$$

The results of the regression in d. change in a similar fashion. The fit of the regression, as well as the coefficient estimate for income, year, and the intercept remain identical. The estimates for the coefficient for price index variables are again scaled according to the corresponding price index for 2004. For example, in the estimate for $\beta_{ppt} = 0.0019642$ in d. and when using 2004 prices as the base, becomes

$$\beta^*_{ppt} = \beta_{ppt} \times \frac{ppt_{2004}}{100} = (0.0019642) \times \frac{(209.1)}{100} = 0.004107$$

Word Count (excluding questions, tables, charts, graphs and appendix):

References

[1] Gary Koop and Justin L Tobias. Learning about heterogeneity in returns to schooling. *Journal of Applied Econometrics*, 19(7):827–849, 2004.

3 Appendix

First 15 Observations

Regression Table for y on X_1

```
      Source |       SS           df       MS      Number of obs   =    17,919
-------------+----------------------------------   F(3, 17915)     =   1252.94
       Model |  867.088558         3  289.029519   Prob > F        =    0.0000
    Residual |  4132.63745    17,915  .230680293   R-squared       =    0.1734
-------------+----------------------------------   Adj R-squared   =    0.1733
       Total |  4999.72601    17,918  .279033709   Root MSE        =    .48029

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0737621   .0022143    33.31   0.000     .0694219    .0781022
    potexper |   .0394896   .0008984    43.96   0.000     .0377287    .0412504
     ability |   .0828907      .0046    18.02   0.000     .0738744    .0919071
       _cons |   1.027229   .0300415    34.19   0.000      .968345    1.086113
------------------------------------------------------------------------------
```

Regression Table for y on X_1

```
      Source |       SS           df       MS      Number of obs   =    17,919
-------------+----------------------------------   F(6, 17912)     =    632.02
       Model |  873.550652         6  145.591775   Prob > F        =    0.0000
    Residual |  4126.17535    17,912   .23035816   R-squared       =    0.1747
-------------+----------------------------------   Adj R-squared   =    0.1744
       Total |  4999.72601    17,918  .279033709   Root MSE        =    .47996

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0722035   .0022508    32.08   0.000     .0677918    .0766152
    potexper |   .0395093   .0008993    43.94   0.000     .0377467    .0412719
     ability |   .0774678   .0049373    15.69   0.000     .0677903    .0871453
    mothered |   -.000117   .0016963    -0.07   0.945     -.003442     .003208
    fathered |   .0054569   .0013387     4.08   0.000      .002833    .0080809
    siblings |   .0047656   .0017924     2.66   0.008     .0012523    .0082789
       _cons |   .9695096   .0337054    28.76   0.000     .9034437    1.035575
------------------------------------------------------------------------------
```

Regression Table for Gasoline Per Capita Consumption on other variables

```
      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =    530.82
       Model |  56708303.2         9  6300922.57   Prob > F        =    0.0000
    Residual |  498548.982        42  11870.2139   R-squared       =    0.9913
-------------+----------------------------------   Adj R-squared   =    0.9894
       Total |  57206852.1        51  1121702.98   Root MSE        =    108.95

------------------------------------------------------------------------------
     pc_quan |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .2157495   .0517618     4.17   0.000     .1112899    .3202091
        gasp |  -11.08386   3.978121    -2.79   0.008    -19.11203   -3.055683
         pnc |   .5773703   12.84414     0.04   0.964    -25.34316     26.4979
         puc |  -5.874637    4.87032    -1.21   0.234    -15.70334    3.954067
         ppt |   6.907265   4.836129     1.43   0.161    -2.852439    16.66697
          pd |   1.228872   11.88175     0.10   0.918    -22.74947    25.20722
          pn |   12.69048   12.59799     1.01   0.320    -12.73329    38.11425
          ps |   -28.0278   7.996249    -3.51   0.001    -44.16488   -11.89072
        year |    72.5037    14.1828     5.11   0.000     43.88165    101.1257
       _cons |  -140421.3   27199.85    -5.16   0.000    -195312.9   -85529.82
------------------------------------------------------------------------------
```

Price Elasticities for 2004

```
margins if year == 2004, eyex(gasp income ppt)

Average marginal effects                        Number of obs     =          1
Model VCE    : OLS

Expression   : Linear prediction, predict()
ey/ex w.r.t. : income gasp ppt

------------------------------------------------------------------------------
             |            Delta-method
             |      ey/ex   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .9476599   .2262966     4.19   0.000     .4909749    1.404345
        gasp |  -.2224796   .0809311    -2.75   0.009    -.3858053    -.059154
         ppt |   .2339837   .1644132     1.42   0.162    -.0978155     .565783
------------------------------------------------------------------------------
```