SS154

Jasen Lo

Homework 2

Q1) Using the link (http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataSets.htm), download the data used in Koop and Tobias's (2004) 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.

```
. import delimited "/Users/jasenlo/Documents/Minerva/Hyderabad 19/SS154/Assign-
ment 3/Koop-Tobias.csv"
(10 vars, 17,919 obs)

. bysort personid: keep if _n==1
(15,741 observations deleted)

. drop if personid > 15
(2,163 observations deleted)
```

Let X1 equal a constant, education, experience, and ability (the individual's own characteristics). Let X2 contain the mother's education, the father's education, and the number of siblings (the household characteristics). Let y be the log wage.

```
. global X1 educ potexper ability

. global X2 mothered fathered siblings

. global y logwage
```

a) Compute the least squares regression coefficients in the regression of y on X1. Report and interpret the coefficients.

```
. reg $y $X1

      Source |       SS           df       MS      Number of obs   =        15
-------------+----------------------------------   F(3, 11)        =      0.82
       Model |  .171377058         3  .057125686   Prob > F        =    0.5080
    Residual |  .763316463        11  .069392406   R-squared       =    0.1834
-------------+----------------------------------   Adj R-squared   =   -0.0394
       Total |  .934693521        14  .066763823   Root MSE        =    .26342

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .014539   .0490215     0.30   0.772    -.0933566    .1224346
    potexper |     .07103   .0480342     1.48   0.167    -.0346924    .1767525
     ability |   .0266154   .0991173     0.27   0.793    -.1915404    .2447711
       _cons |    1.66364   .6185532     2.69   0.021     .3022133    3.025066
------------------------------------------------------------------------------
```

All the coefficients of this log-linear model are positive. The values of the coefficient represent the percentage change in wages per unit change in the explanatory variables. For example, one more year of education leads to a 14% increase in wages. None of the coefficients are statistically significant at α level of 0.1.

b) Compute the least squares regression coefficients in the regression of y on X1 and X2. Report and interpret the coefficients.

```
. reg $y $X1 $X2, noconstant

      Source |       SS           df       MS      Number of obs   =        15
-------------+----------------------------------   F(6, 9)         =    212.51
       Model |  64.0950841         6   10.682514   Prob > F        =    0.0000
    Residual |  .452417043         9   .05026856   R-squared       =    0.9930
-------------+----------------------------------   Adj R-squared   =    0.9883
       Total |  64.5475011        15  4.30316674   Root MSE        =    .22421

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0272969   .0324092     0.84   0.421    -.0460178    .1006116
    potexper |   .1041425   .0424856     2.45   0.037     .0080334    .2002515
     ability |    .028621   .1075159     0.27   0.796    -.2145969    .2718389
    mothered |   .1039024   .0515562     2.02   0.075    -.0127259    .2205307
    fathered |   .0017645   .0420454     0.04   0.967    -.0933489    .0968778
    siblings |   .0589675   .0649773     0.91   0.388    -.0880214    .2059563
------------------------------------------------------------------------------
```

All the coefficients of this log-linear model are positive. The values of the coefficient represents the percentage change in wages per unit change in the explanatory variables. For example, one more year of education leads to a 2% increase in wages. Coefficients of *potexper* (at α level of 0.05) and *mothered* (at α level of 0.1) are statistically signifiant. None of others are statistically significant at α level of 0.1.

c) Compute the R-squared for the the regression of y on X1 and X2 manually using the SSE and SST from the output. Repeat the computation for the case in which the constant term is omitted from X1. What happens to R-squared?

```
. reg $y $X1 $X2

      Source |       SS           df       MS      Number of obs   =        15
-------------+----------------------------------   F(6, 8)         =      1.42
       Model |  .482427232         6  .080404539   Prob > F        =    0.3140
    Residual |   .45226629         8  .056533286   R-squared       =    0.5161
-------------+----------------------------------   Adj R-squared   =    0.1532
       Total |  .934693521        14  .066763823   Root MSE        =    .23777

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0258221   .0446859     0.58   0.579    -.0772238    .1288681
    potexper |   .1033913   .0473454     2.18   0.061    -.0057875      .21257
     ability |   .0307435   .1212013     0.25   0.806    -.2487472    .3102343
   mothered  |   .1016307    .070175     1.45   0.186    -.0601932    .2634546
   fathered  |   .0016443   .0446491     0.04   0.972    -.1013167    .1046054
   siblings  |   .0591692    .069018     0.86   0.416    -.0999866     .218325
       _cons |   .0489959   .9488077     0.05   0.960    -2.138959     2.23695
```

$$R^2 = 1 - SSE/SST$$

$$= 1 - 0.4522/0.9345$$

$$= 0.5161$$

```
. reg $y $X1 $X2, noconstant

      Source |       SS           df       MS      Number of obs   =        15
-------------+----------------------------------   F(6, 9)         =    212.51
       Model |  64.0950841         6   10.682514   Prob > F        =    0.0000
    Residual |  .452417043         9   .05026856   R-squared       =    0.9930
-------------+----------------------------------   Adj R-squared   =    0.9883
       Total |  64.5475011        15  4.30316674   Root MSE        =    .22421

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0272969   .0324092     0.84   0.421    -.0460178    .1006116
    potexper |   .1041425   .0424856     2.45   0.037     .0080334    .2002515
     ability |    .028621   .1075159     0.27   0.796    -.2145969    .2718389
    mothered |   .1039024   .0515562     2.02   0.075    -.0127259    .2205307
    fathered |   .0017645   .0420454     0.04   0.967    -.0933489    .0968778
    siblings |   .0589675   .0649773     0.91   0.388    -.0880214    .2059563
```

$$R^2 = 1 - SSE/SST$$

$$= 1 - 0.4524/64.5475$$

$$= 0.9930$$

The R^2 value (0.9930) of the <reg $y $X1 $X2, noconstant> is much higher than the R^2 value (0.5161) of <reg $y $X1 $X2>. This implies that the regression line of <reg $y $X1 $X2, noconstant> fits the data closer than the regression line of <reg $y $X1 $X2>. However, this is not true since the SSR of both models are similar (0.452). When a regression has no constant, the SSE of *logwage* accounted for by the intercept are not included in the SST. As a result, the SSE of the no constant regression is inflated compared to the regression with the constant.

d) Compute the adjusted R-squared for the full regression including the constant term. Interpret your results. Do we need the constant term?

$$
\begin{aligned}
R^2 &= 1 - (1 - R^2)((n-1)/(n-k+1)) \\
&= 1 - (1 - 0.5161)((15-1)/(15-6+1)) \\
&= 0.1532
\end{aligned}
$$

The adjusted R^2 (0.1532) is a much lower compared to R^2 (0.5161). This means that at least one of the explanatory variables used in <reg $y $X1 $X2> do not improve the model, as the adjusted R^2 value was heavily penalised.

The constant term should be included from the perspective of economic theory. Having no constant term would mean a person with zero in the explanatory variables, such as education and prior experience would earn $0 in wages. This is unlikely, since fresh graduate students and self-employed people in the labour force earn more than $0 despite having no prior experience, years of education or educated parents. The model <reg $y $X1 $X2, noconstant> makes little economic sense.

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

- MR1: Relationship between the independent and dependent variables is linear[1]

```
. quietly scatter $y educ || lfit $y educ, name(scatter_educ)
. quietly scatter $y potexper || lfit $y potexper, name(scatter_potexper)
. quietly scatter $y ability || lfit $y ability, name(scatter_ability)
. quietly scatter $y mothered || lfit $y mothered, name(scatter_mothered)
. quietly scatter $y fathered || lfit $y fathered, name(scatter_fathered)
. quietly scatter $y siblings || lfit $y siblings, name(scatter_siblings)
. graph combine scatter_educ scatter_potexper scatter_ability scatter_mothered
scatter_fathered scatter_siblings, col(3) row(2)
```

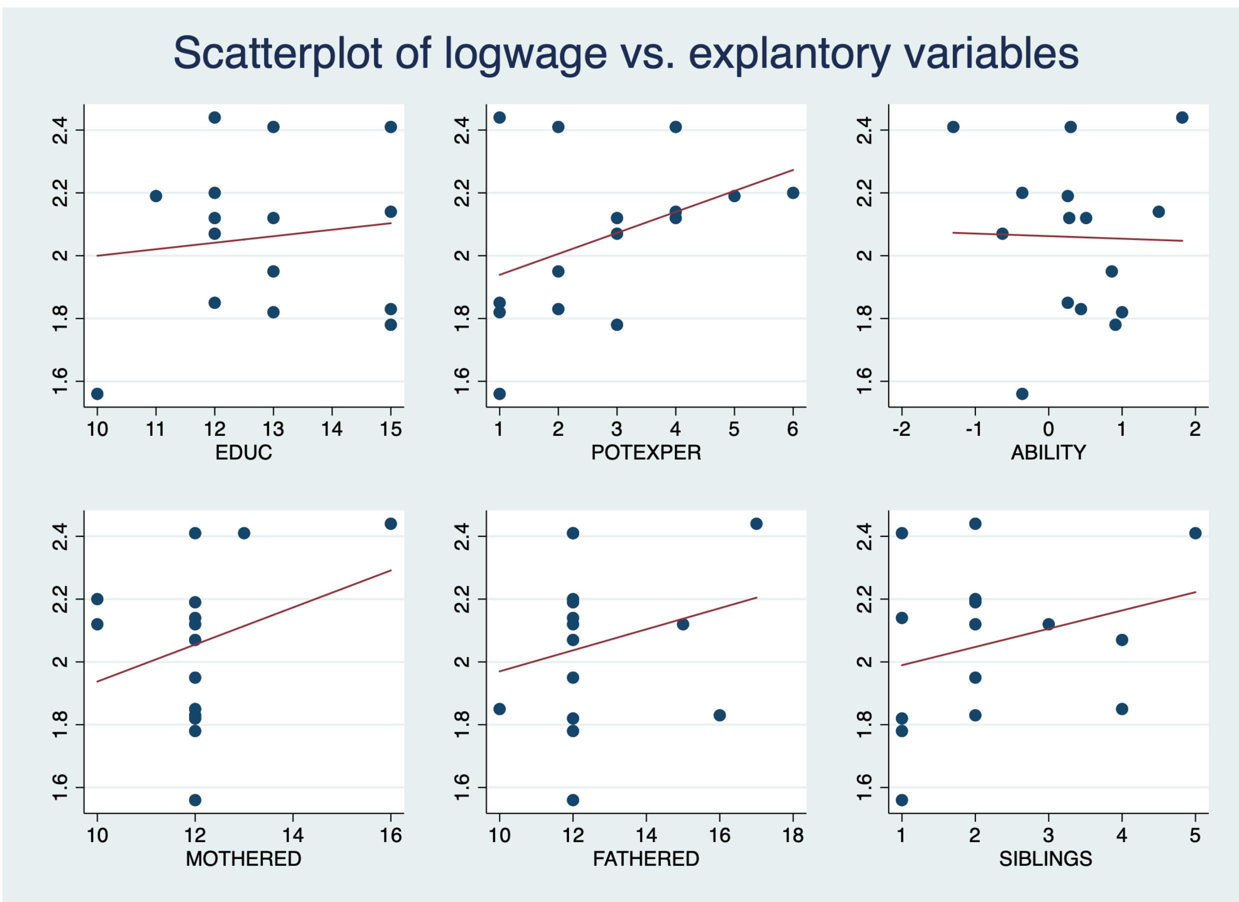

Figure 1 - Scatterplots of dependent variable *logwage* plotted against explanatory variables *educ, potexper, ability, mothered, fathered* and *siblings*. All scatterplots show linear relationships. *potexper* could be interpreted as having a curvilinear or linear relationship.

[1] dataviz - I used the graph combine command in Stata to concisely convey the presence of linearity in all of the independent variables, instead of using multiple graphs for each variable.

MR1 is satisfied since all the explanatory variables have a linear relationship with the dependent variable. The scatterplot of *logwage* vs. *potexper* could be interpreted as both non-linear or linear, but it is difficult to interpret due to the small sample of 15. However, the linear fit of the scatterplot fits well, so MR1 is conclusively satisfied.

- MR2: Residuals of the regression are be normally distributed.[2]

```
. quietly reg $y $X1 $X2
. predict res, resid
. hist res, kdensity normal
```

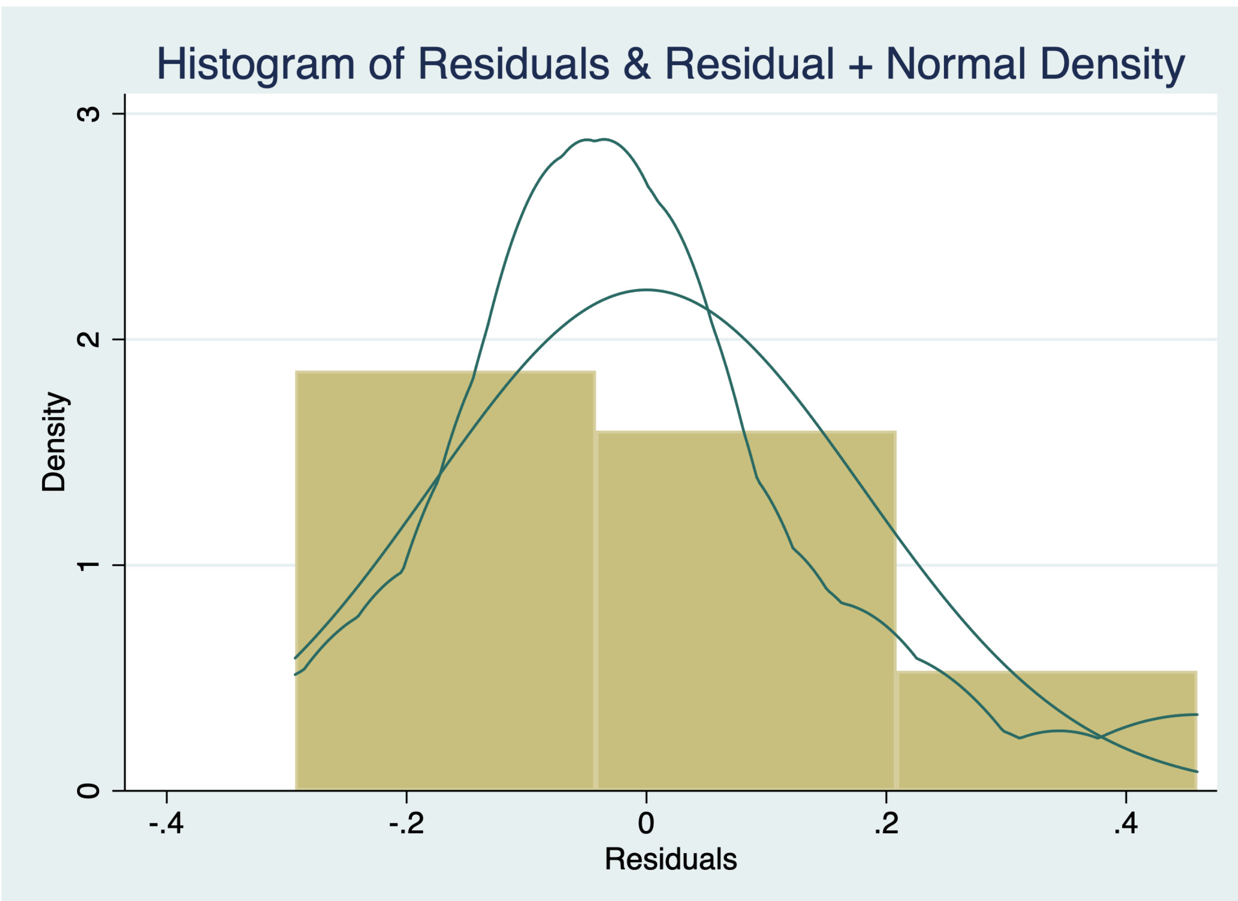

Figure 2 - Histogram of the regression residuals (olive bars) , accompanied by normal and residual density curves. Although the residuals are not perfectly normally distributed, it is approaching normality, likely with a larger sample size, they would be normally distributed.

---

[2] #distributions - I use my knowledge of the normal distributions in the context of the assumptions of the multilinear regression assumptions. Since the residuals of the regression approximated a normal distributed, I concluded that the residuals were normally distributed and thus fulfilled MR2, the second assumption of the multilinear regression.

```
. jb res

Jarque-Bera normality test:  3.234 Chi(2)  .1985
Jarque-Bera test for Ho: normality:

. quietly reg $y $X1 $X2, noconstant
. predict res2, resid

. jb res2
Jarque-Bera normality test:  3.076 Chi(2)  .2148
Jarque-Bera test for Ho: normality:
```

MR2 is satisfied after plotting a histogram of the residuals and the Jarque-Bera (JB) normality test on the residuals. *Figure 2* shows that the residuals are approximately normally distributed. The JB normality test of the residuals also reveal that the residuals are normally distributed. The p-value of the JB normality test is 0.1985, thus we fail to reject the null hypothesis that the residuals are normally distributed at α level of 0.1. The regression without the constant also has normally distributed residuals according to the JB normality test (p-value = 0.2148).

- MR5: There is no multicollinearity in the data.[3]

```
. corr $X1 $X2
(obs=15)

             |     educ potexper  ability mothered fathered siblings
-------------+------------------------------------------------------
        educ |   1.0000
    potexper |   0.0176   1.0000
     ability |   0.3527  -0.3252   1.0000
    mothered |   0.0046  -0.4371   0.3807   1.0000
    fathered |   0.1585  -0.1239   0.3929   0.5628   1.0000
    siblings |  -0.2765   0.1297  -0.6128   0.0339  -0.1843   1.0000


. quietly reg $y $X1 $X2                . quietly reg $y $X1 $X2, noconstant
. estat vif                             . estat vif, uncentered


    Variable |       VIF       1/VIF        Variable |       VIF       1/VIF
-------------+----------------------      -------------+----------------------
     ability |      2.40    0.417218        mothered |    116.81    0.008561
    mothered |      2.17    0.460425        fathered |     86.30    0.011587
```

[3] #induction - In determining whether or not there is multicollinearity in the data, I used two types of evidence to support my induced conclusion: variance inflation factor and correlation matrices to support my conclusion that there is no multicollinearity in the regression with the constant.

```
    siblings |      1.89    0.529824            educ |     52.59    0.019014
   fathered |      1.67    0.599117        siblings |      7.98    0.125328
   potexper |      1.36    0.733145        potexper |      5.46    0.183219
       educ |      1.19    0.839273         ability |      2.58    0.387111
-------------+----------------------    -------------+----------------------
    Mean VIF |      1.78                    Mean VIF |     45.29
```

According to the correlation matrix and the variance inflation factor (VIF), there is no multicollinearity for **Part a)**, but there seems to be for **Part b)**. The correlation matrix of the independent variables shows that there is no serious collinearity, as the highest correlation between two variables *siblings* and *ability* is -0.6128. The mean VIF of <reg $y $X1 $X2> is a low 1.78, however, the mean VIF of <reg $y $X1 $X2, noconstant> is 45.29. This means that serious collinearity exists between the constant term and some independent variables, such as *mothered* and *fathered*.

- MR6: homoscedasticity - variance is constant for all values of explanatory variables.[4]

```
. quietly reg $y $X1 $X2
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of logwage

         chi2(1)      =     1.25
         Prob > chi2  =   0.2628
```

MR6 is satisfied according to the Breusch-Pagan / Cook-Weisberg test for heteroskedasticity, as the data is homoscedastic. The p-value for the test is 0.2628, so we fail to reject the null hypothesis that the variance is constant for the regression model.

[4] #significance - I apply understanding of statistical significance in the context of the Breusch-Pagan / Cook-Weisberg test for heteroskedasticity, concluding that there is no heteroskedasticity since the p-value is above the alpha level of 0.1 and therefore the null hypothesis of homoscedasticity can not be rejected.

Conclusion: Classical multilinear regression assumptions of MR1, MR2, MR5 and MR6 are satisfied for **Part a)**, but only MR1, MR2 and MR6 is satisfied for **Part b)**. [5]

Q2) Data on U.S. gasoline consumption for the years 1953 to 2004 are given in Table F2.2 (http://pages.stern.nyu.edu/~wgreene/Text/Edition7/tablelist8new.htm). **To obtain the per capita quantity variable, divide GASEXP (total U.S. gas expenditure) by GASP (price index for gasoline) times Pop (U.S. population in thousands).**

```
. import delimited "/Users/jasenlo/Documents/Minerva/Hyderabad 19/SS154/Assign-
ment 3/TableF2-2.csv"
(11 vars, 52 obs)

. gen gasexp_per_cap = gasexp*10^9/(gasp*pop*1000)

. global X income gasp pnc puc ppt pd pn ps year

. global y gasexp_per_cap
```

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?

```
. reg $y $X

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =    530.82
       Model |  56.7083042         9  6.30092268   Prob > F        =    0.0000
    Residual |   .49854905        42  .011870215   R-squared       =    0.9913
-------------+----------------------------------   Adj R-squared   =    0.9894
       Total |  57.2068532        51    1.121703   Root MSE        =    .10895

------------------------------------------------------------------------------
gasexp_per~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0002157   .0000518     4.17   0.000     .0001113    .0003202
        gasp |  -.0110838   .0039781    -2.79   0.008     -.019112   -.0030557
         pnc |   .0005774   .0128441     0.04   0.964    -.0253432    .0264979
         puc |  -.0058746   .0048703    -1.21   0.234    -.0157033    .0039541
         ppt |   .0069073   .0048361     1.43   0.161    -.0028524     .016667
          pd |   .0012289   .0118818     0.10   0.918    -.0227495    .0252072
          pn |   .0126905    .012598     1.01   0.320    -.0127333    .0381142
          ps |  -.0280278   .0079962    -3.51   0.001    -.0441649   -.0118907
        year |   .0725037   .0141828     5.11   0.000     .0438816    .1011257
       _cons |  -140.4213   27.19985    -5.16   0.000    -195.3128    -85.5298
------------------------------------------------------------------------------
```

[5] Not sure how to interpret estat via, uncentered for reg …, noconstant

All coefficients except *pnc, puc, ppt, pd* and *pn* are statistically significant at 0.1 α level. Most of the signs agree with expectations except from *pd, pn* and *ps*. For the aggregate price index of consumer durables, non-durables and services, there are too many different types of consumables that may be correlated with a decrease or increase in per capita gasoline consumption. For example, there are many services that use gasoline (aviation, logistics), and many services that do not (finance, tech). However, in general, per-capita consumption of gasoline can be an indicator of economic health since one is likely to consume more gasoline when the economy is doing well. In this sense, the signs of coefficients *pd* and *pn* make sense, while *ps* does not, since one would expect service prices to increase in a healthy economy.

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.

```
. test pnc = puc

 ( 1)  pnc - puc = 0

       F(  1,    42) =    0.24
            Prob > F =    0.6233
```

A joint hypothesis test that the coefficients of *pnc* and *puc* are equivalent, gives a p-value of 0.6233, which is statistically insignificant at α level of 0.1. This means that we fail to reject the null hypothesis that the consumers do not differentiate between changes in the prices of new and used cars.

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.

```
. margins, dydx(*) at(year = 2004)

Average marginal effects                        Number of obs     =         52
```

```
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : income gasp pnc puc ppt pd pn ps year
at           : year            =        2004

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0002157   .0000518     4.17   0.000     .0001113    .0003202
        gasp |  -.0110838   .0039781    -2.79   0.008     -.019112   -.0030557
         pnc |   .0005774   .0128441     0.04   0.964    -.0253432    .0264979
         puc |  -.0058746   .0048703    -1.21   0.234    -.0157033    .0039541
         ppt |   .0069073   .0048361     1.43   0.161    -.0028524     .016667
          pd |   .0012289   .0118818     0.10   0.918    -.0227495    .0252072
          pn |   .0126905    .012598     1.01   0.320    -.0127333    .0381142
          ps |  -.0280278   .0079962    -3.51   0.001    -.0441649   -.0118907
        year |   .0725037   .0141828     5.11   0.000     .0438816    .1011257
------------------------------------------------------------------------------
```

- Price elasticity of demand: dy/dx *gasp* = -0.110838
    - A negative PED between -1 and 0 corresponds to economic theory of a inelastic demand curve, which makes sense since people are likely to continue to consume gasoline regardless of price changes. For example, it is difficult to use alternatives fuels if one already has a vehicle with a gasoline engine.
- Income elasticity of demand: dy/dx *income* = 0.0002157
    - A positive YED between 0 and 1 means the gasoline is a normal and inelastic good. This conforms to economic theory since one is likely to consume gasoline despite increasing income levels due to lifespan of vehicles and limited alternatives.
- Cross-price elasticity with public transportation: dy/dx *ppt* = 0.0069073
    - A positive XED between 0 and 1 means that that public transport and gasoline are weak substitutes of one another. This also makes economic sense since gasoline price changes wouldn't immediately prompt one to take public transport,

especially in the less urban areas of the United States which may not have extensive public transport coverage.

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?

```
. gen log_gasexp_per_cap = log(gasexp_per_cap)

. gen log_income = log(income)

. gen log_gasp = log(gasp)

. gen log_pnc = log(pnc)

. gen log_puc = log(puc)

. gen log_ppt = log(ppt)

. gen log_pd = log(pd)

. gen log_pn = log(pn)

. gen log_ps = log(ps)

. global X2 log_income log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps

. global y2 log_gasexp_per_cap

. reg $y2 $X2

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(8, 43)        =    249.60
       Model |  2.84726323         8  .355907904   Prob > F        =    0.0000
    Residual |  .061313662        43  .001425899   R-squared       =    0.9789
-------------+----------------------------------   Adj R-squared   =    0.9750
       Total |   2.9085769        51   .05703092   Root MSE        =    .03776

------------------------------------------------------------------------------
log_gasexp~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_income |   1.883045    .223034     8.44   0.000     1.433254    2.332836
    log_gasp |   .0735984   .0676117     1.09   0.282    -.0627536    .2099504
     log_pnc |   .3772717     .30747     1.23   0.226    -.2428007     .997344
     log_puc |   -.334021   .0996132    -3.35   0.002    -.5349102   -.1331318
     log_ppt |   .1404593   .1683464     0.83   0.409    -.1990435    .4799621
      log_pd |   .6422717   .1817908     3.53   0.001     .2756555    1.008888
      log_pn |   -.492239   .3269502    -1.51   0.139    -1.151597     .167119
      log_ps |  -.6288652   .4383016    -1.43   0.159    -1.512785    .2550542
       _cons |  -15.79148    2.35185    -6.71   0.000    -20.53443   -11.04852
------------------------------------------------------------------------------
```

Estimate signs of variables such as *gasp* and *pn* changed in the log-log model. The magnitudes of some variables estimates such as *income, pnc, puc, ppt* and *pd* increased between 1000% to 10000% (*log_pnc* = 0.3773, *pnc* = 0.0005).

I prefer the linear-linear model much better for the following reasons:

1. Signs of the elasticity estimates linear-linear model conform to economic theory, while the log-log model elasticities do not. Gasoline is an inelastic, normal good. However, the log-log model shows that gasoline is a elastic, luxury good because the PED is positive (0.0735984) and the YED is much higher than 1 (1.883045).
2. The linear-linear model allows finding instantaneous elasticity, while the log-log model finds the different elasticities of demand between 1953 to 2004. Elasticities over such a long period of time are not as useful, since they are vulnerable to con-founding effects of historical events (i.e: 1973 oil crisis) on price elasticises or techno-logical advances in alternative sources of fuel.

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?

```
. corr gasp pnc puc ppt pd pn ps
(obs=52)

             |     gasp      pnc      puc      ppt       pd       pn       ps
-------------+---------------------------------------------------------------
        gasp |   1.0000
         pnc |   0.9361   1.0000
         puc |   0.9228   0.9939   1.0000
         ppt |   0.9270   0.9807   0.9824   1.0000
          pd |   0.9389   0.9933   0.9878   0.9585   1.0000
          pn |   0.9627   0.9885   0.9822   0.9899   0.9773   1.0000
          ps |   0.9394   0.9785   0.9769   0.9975   0.9563   0.9936   1.0000

. corr log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps
(obs=52)

             | log_gasp  log_pnc  log_puc  log_ppt   log_pd   log_pn   log_ps
-------------+---------------------------------------------------------------
```

```
    log_gasp |   1.0000
     log_pnc |   0.9667   1.0000
     log_puc |   0.9674   0.9940   1.0000
     log_ppt |   0.9665   0.9891   0.9910   1.0000
      log_pd |   0.9776   0.9932   0.9945   0.9864   1.0000
      log_pn |   0.9839   0.9900   0.9902   0.9942   0.9923   1.0000
      log_ps |   0.9742   0.9902   0.9912   0.9985   0.9886   0.9979   1.0000
```

```
. quietly reg $y $X

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------

          pn |   1614.88    0.000619
          ps |   1229.94    0.000813
         pnc |    974.93    0.001026
          pd |    820.65    0.001219
         ppt |    481.06    0.002079
      income |    354.84    0.002818
         puc |    265.78    0.003762
        year |    198.49    0.005038
        gasp |     64.62    0.015476
-------------+----------------------
    Mean VIF |    667.24
```

```
. quietly reg $y2 $X2

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------

      log_ps |   4902.30    0.000204
      log_pn |   1566.09    0.000639
       log_ppt |    790.87    0.001264
       log_pnc |    645.15    0.001550
       log_pd  |    305.77    0.003270
       log_income | 216.20    0.004625
       log_puc |    192.91    0.005184
       log_gasp |   75.38     0.013266
-------------+----------------------
    Mean VIF |   1086.83
```

Multicollinearity is problematic for the regressions in **part a)** and **part d)** according to correlation matrix of independent variables and the VIF test. Both regressions' correlation matrices show correlations between variables to be all above 0.9, and the VIF test for both regressions have mean VIFs in the hundreds and thousands (667.24, 1086.83).

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?

```
. gen gasp_index = 100 * gasp / gasp[52]

. gen pnc_index = 100 * pnc / pnc[52]

. gen puc_index = 100 * puc / puc[52]

. gen ppt_index = 100 * ppt / ppt[52]

. gen pd_index = 100 * pd / pd[52]
```

```
. gen pn_index = 100 * pn / pn[52]

. gen ps_index = 100 * ps / pn[52]

. global X3 income gasp_index pnc_index puc_index ppt_index pd_index pn_index
ps_index year

. reg $y $X3

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =    530.82
       Model |  56.7083042         9  6.30092269   Prob > F        =    0.0000
    Residual |  .498548983        42  .011870214   R-squared       =    0.9913
-------------+----------------------------------   Adj R-squared   =    0.9894
       Total |  57.2068532        51    1.121703   Root MSE        =    .10895

------------------------------------------------------------------------------
gasexp_per~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0002157   .0000518     4.17   0.000     .0001113    .0003202
  gasp_index |   -.013733   .0049289    -2.79   0.008      -.02368   -.003786
   pnc_index |    .000773   .0171983     0.04   0.964    -.0339345    .0354806
   puc_index |  -.0078309   .0064921    -1.21   0.234    -.0209325    .0052708
   ppt_index |   .0144431   .0101123     1.43   0.161    -.0059645    .0348506
    pd_index |   .0014108   .0136402     0.10   0.918    -.0261164    .0289379
    pn_index |    .021853   .0216937     1.01   0.320    -.0219267    .0656328
    ps_index |  -.0482639   .0137695    -3.51   0.001    -.0760519   -.0204758
        year |   .0725037   .0141828     5.11   0.000     .0438817    .1011257
       _cons |  -140.4214   27.19983    -5.16   0.000    -195.3129   -85.52988
------------------------------------------------------------------------------

. reg $y $X

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =    530.82
       Model |  56.7083042         9  6.30092268   Prob > F        =    0.0000
    Residual |   .49854905        42  .011870215   R-squared       =    0.9913
-------------+----------------------------------   Adj R-squared   =    0.9894
       Total |  57.2068532        51    1.121703   Root MSE        =    .10895

------------------------------------------------------------------------------
gasexp_per~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0002157   .0000518     4.17   0.000     .0001113    .0003202
        gasp |  -.0110838   .0039781    -2.79   0.008     -.019112   -.0030557
         pnc |   .0005774   .0128441     0.04   0.964    -.0253432    .0264979
         puc |  -.0058746   .0048703    -1.21   0.234    -.0157033    .0039541
         ppt |   .0069073   .0048361     1.43   0.161    -.0028524     .016667
          pd |   .0012289   .0118818     0.10   0.918    -.0227495    .0252072
          pn |   .0126905    .012598     1.01   0.320    -.0127333    .0381142
          ps |  -.0280278   .0079962    -3.51   0.001    -.0441649   -.0118907
        year |   .0725037   .0141828     5.11   0.000     .0438816    .1011257
       _cons |  -140.4213   27.19985    -5.16   0.000    -195.3128    -85.5298
------------------------------------------------------------------------------
```

```
. gen log_gasp_index = log(gasp_index)

. gen log_pnc_index = log(pnc_index)

. gen log_puc_index = log(puc_index)
```

```
. gen log_ppt_index = log(ppt_index)

. gen log_pd_index = log(pd_index)

. gen log_pn_index = log(pn_index)

. gen log_ps_index = log(ps_index)

. global X4 log_income log_gasp_index log_pnc_index log_puc_index log_ppt_index
log_pd_index log_pn_index log_ps_index

. reg $y2 $X4

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(8, 43)        =    249.60
       Model |  2.84726325         8  .355907906   Prob > F        =    0.0000
    Residual |  .061313646        43  .001425899   R-squared       =    0.9789
-------------+----------------------------------   Adj R-squared   =    0.9750
       Total |   2.9085769        51   .05703092   Root MSE        =    .03776

------------------------------------------------------------------------------
log_gasexp_p~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
    log_income |   1.883046   .2230338     8.44   0.000     1.433255    2.332836
log_gasp_index |   .0735977   .0676117     1.09   0.282    -.0627544    .2099497
 log_pnc_index |    .377272   .3074698     1.23   0.226       -.2428    .9973439
 log_puc_index |  -.3340209   .0996132    -3.35   0.002    -.5349102   -.1331317
 log_ppt_index |   .1404582   .1683462     0.83   0.409    -.1990443    .4799607
  log_pd_index |   .6422717   .1817904     3.53   0.001     .2756563    1.008887
  log_pn_index |  -.4922384   .3269503    -1.51   0.139    -1.151596    .1671197
  log_ps_index |  -.6288642    .438301    -1.43   0.159    -1.512782    .2550539
         _cons |  -16.17863   2.368256    -6.83   0.000    -20.95468   -11.40259
------------------------------------------------------------------------------

. reg $y2 $X2

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(8, 43)        =    249.60
       Model |  2.84726323         8  .355907904   Prob > F        =    0.0000
    Residual |  .061313662        43  .001425899   R-squared       =    0.9789
-------------+----------------------------------   Adj R-squared   =    0.9750
       Total |   2.9085769        51   .05703092   Root MSE        =    .03776

------------------------------------------------------------------------------
log_gasexp~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_income |   1.883045    .223034     8.44   0.000     1.433254    2.332836
    log_gasp |   .0735984   .0676117     1.09   0.282    -.0627536    .2099504
     log_pnc |   .3772717     .30747     1.23   0.226    -.2428007     .997344
     log_puc |   -.334021   .0996132    -3.35   0.002    -.5349102   -.1331318
     log_ppt |   .1404593   .1683464     0.83   0.409    -.1990435    .4799621
      log_pd |   .6422717   .1817908     3.53   0.001     .2756555    1.008888
      log_pn |   -.492239   .3269502    -1.51   0.139    -1.151597     .167119
      log_ps |  -.6288652   .4383016    -1.43   0.159    -1.512785    .2550542
       _cons |  -15.79148    2.35185    -6.71   0.000    -20.53443   -11.04852
------------------------------------------------------------------------------
```

The regression in **part a)** changed the regression coefficients' magnitude slightly, while the regression in **part d)** barely changed at all. *Figure 3* shows why the regression coefficients for the linear-linear regression model changed, while the log-log regression model's coefficients did not.

```
. scatter gasexp_per_cap gasp || scatter gasexp_per_cap gasp_index, name(e1)

. scatter log_gasexp_per_cap log_gasp || scatter log_gasexp_per_cap log_gasp_in-
dex, name(e2)

. graph combine e1 e2
```

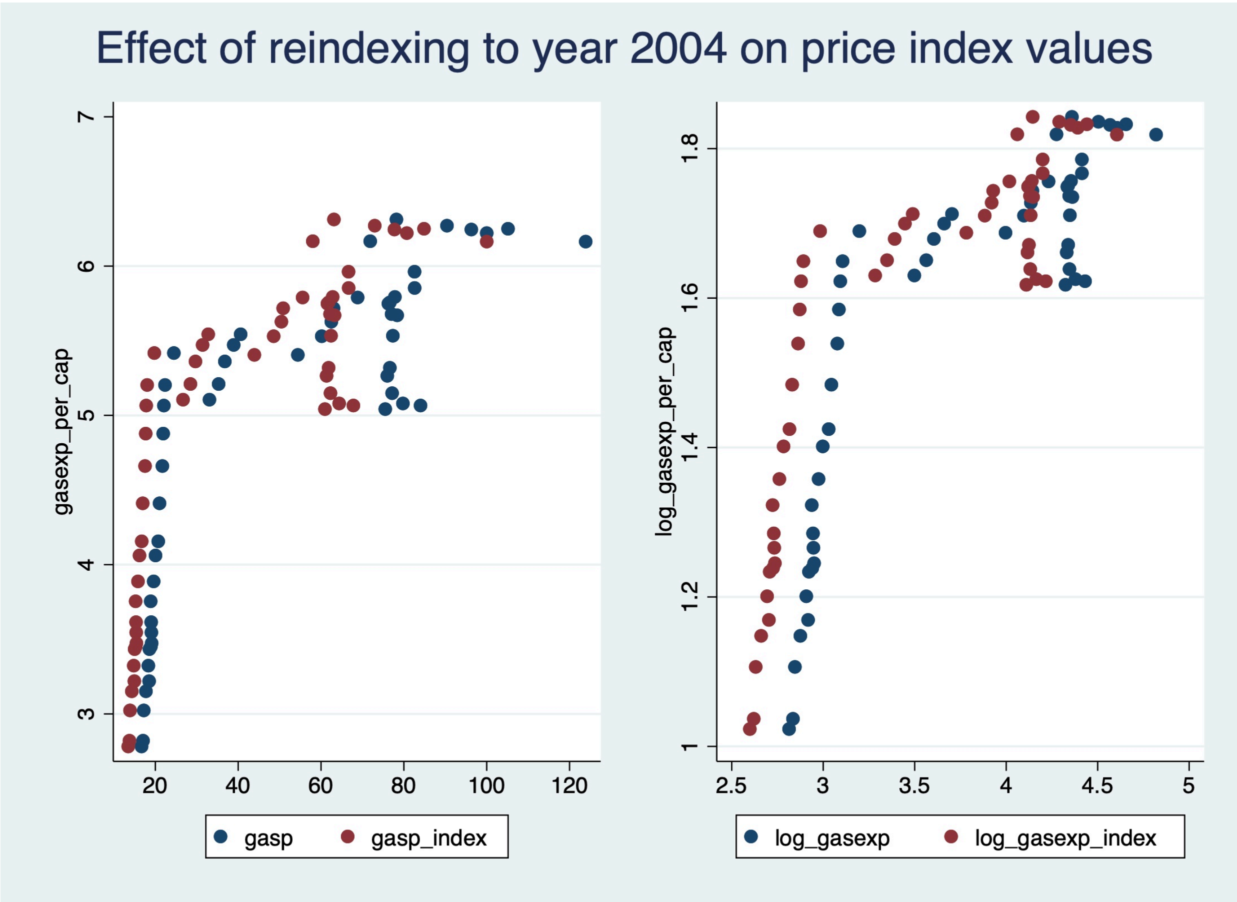

Figure 3 - Scatterplots showing the contrast between the effect of reindexing on the linear-linear model and the log-log. The first graph of the linear-linear model shows how distance between the indexation at 2000 versus at 2004 for the increases as the price index value approaches 100, whereas the distance in the log-log model remains constant regardless of the price index value.

The log-log model of **part d)** normalised the effect of the re-indexation, whereas the linear-linear model of **part a)** did not. As a result, the log-log regression model remained the same regardless of the base year indexation, while the linear-linear regression model was not the same.

Appendix

```
     *Q1
     clear
     import delimited "/Users/jasenlo/Documents/Minerva/Hyderabad 19/SS154/Assignment 
3/Koop-Tobias.csv"
     bysort personid: keep if _n==1
     drop if personid > 15
     global X1 educ potexper ability
     global X2 mothered fathered siblings
     global y logwage

     *a)
     reg $y $X1

     *b)
     reg $y $X1 $X2, noconstant

     *c)
     reg $y $X1 $X2
     di 1 - 0.4522/.9345
     reg $y $X1 $X2, noconstant
     di 1 - 0.4524/64.5475

     *d)
     reg $y $X1 $X2
     di 1-(1-0.5161)*((14)/8)

     *e)
     quietly reg $y $X1 $X2
     quietly scatter $y educ || lfit $y educ, name(scatter_educ)
     quietly scatter $y potexper || lfit $y potexper, name(scatter_potexper)
     quietly scatter $y ability || lfit $y ability, name(scatter_ability)
     quietly scatter $y mothered || lfit $y mothered, name(scatter_mothered)
     quietly scatter $y fathered || lfit $y fathered, name(scatter_fathered)
     quietly scatter $y siblings || lfit $y siblings, name(scatter_siblings)
     graph combine scatter_educ scatter_potexper scatter_ability scatter_mothered 
scatter_fathered scatter_siblings, col(3) row(2)

     predict res, resid
     hist res, kdensity normal
     jb res

     quietly reg $y $X1 $X2, noconstant
     predict res2, resid
     jb res2

     corr $X1 $X2

     quietly reg $y $X1 $X2
     estat vif
     quietly reg $y $X1 $X2, noconstant
     estat vif, uncentered

     quietly reg $y $X1 $X2
     estat hettest

     *Q2
     clear
     import delimited "/Users/jasenlo/Documents/Minerva/Hyderabad 19/SS154/Assignment 
3/TableF2-2.csv"
     gen gasexp_per_cap = gasexp*10^9/(gasp*pop*1000)
     global X income gasp pnc puc ppt pd pn ps year
```

```
     global y gasexp_per_cap

     *a)
     reg $y $X

     *b)
     test pnc = puc

     *c)
     margins, dydx(*) at(year = 2004)

     *d)
     gen log_gasexp_per_cap = log(gasexp_per_cap)
     gen log_income = log(income)
     gen log_gasp = log(gasp)
     gen log_pnc = log(pnc)
     gen log_puc = log(puc)
     gen log_ppt = log(ppt)
     gen log_pd = log(pd)
     gen log_pn = log(pn)
     gen log_ps = log(ps)
     global X2 log_income log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps
     global y2 log_gasexp_per_cap
     reg $y2 $X2

     *e)
     corr gasp pnc puc ppt pd pn ps
     quietly reg $y $X
     estat vif

     corr log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps
     quietly reg $y2 $X2
     estat vif

     *f)
     gen gasp_index = 100 * gasp / gasp[52]
     gen pnc_index = 100 * pnc / pnc[52]
     gen puc_index = 100 * puc / puc[52]
     gen ppt_index = 100 * ppt / ppt[52]
     gen pd_index = 100 * pd / pd[52]
     gen pn_index = 100 * pn / pn[52]
     gen ps_index = 100 * ps / pn[52]
     global X3 income gasp_index pnc_index puc_index ppt_index pd_index pn_index
ps_index year
     reg $y $X3
     reg $y $X

     gen log_gasp_index = log(gasp_index)
     gen log_pnc_index = log(pnc_index)
     gen log_puc_index = log(puc_index)
     gen log_ppt_index = log(ppt_index)
     gen log_pd_index = log(pd_index)
     gen log_pn_index = log(pn_index)
     gen log_ps_index = log(ps_index)


     global X4 log_income log_gasp_index log_pnc_index log_puc_index log_ppt_index
log_pd_index log_pn_index log_ps_index
     reg $y2 $X4
     reg $y2 $X2
```

```
     scatter gasexp_per_cap gasp || scatter gasexp_per_cap gasp_index, name(e1)
     scatter log_gasexp_per_cap log_gasp || scatter log_gasexp_per_cap log_gasp_index,
name(e2)
     graph combine e1 e2
```