Homework 2

Jacob Puthipiroj

March 2, 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.

A table for the first 15 individuals in the sample is given in the Appendix. It is also possible to click on links to be taken directly to the corresponding graphic/table in the Appendix.

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.**

   The full regression table is available here. 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-log 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.

b. **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 here. 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.

c. **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?**

The Stata output from the previous table gives SSE = 4126.17535, SST = 4999.72601. The R^2 formula is

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

The Stata output for the no-intercept model is given here, where SSE = 4316.7689. However, an incorrect value of SST = 99529.3551 is given, as a result of SST being implicitly calculated in the Stata backend assuming an intercept exists. The mathematical definition of SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$ means it is constant, and so we can use the previous SST.

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

d. **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

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

Whether or not we need the constant term depends on whether the model is better with or without it. There is some economic sense to the intercept model. Even if someone has no education, potential experience or ability, they should still be able to make minimum wage, which is greater than 0. From a statistical point of view, including the intercept gives nicer properties, such as guaranteeing that the mean of the least squares residual is zero, in line with assumption MR2 of the multiple regression model. Conversely, forcing the regression line through the origin will usually result in the residuals having nonzero mean, as this is almost always inconsistent with the best fit line. Finally, we have evidence that the intercept term was found statistically significant. Including it should increase explanatory power, and we see this in the increase in the adjusted R^2.[1]

[1] #modeltesting: I choose between the form of $f_1(y) = \beta_0 + \mathbf{B}_1 X_1 + \mathbf{B}_2 X_2$ and $f(y) = \mathbf{B}_1 X_1 + \mathbf{B}_2 X_2$ (no intercept), by answering the question of whether we need the constant term. arguing that while a no constant makes economic, this model cannot be used for extrapolation in any case. I also arguing that the intercept was found statistically significant, and helps in creating a better model from the perspective of the adjusted R^2. Furthermore, residuals will usually have a nonzero mean, because forcing the regression line through the origin is generally inconsistent with best fit. This in turn violates the Gauss-Markov assumptions, and the obtained coefficients are no longer the best linear unbiased estimators (BLUE).

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

- MR1 states that the model linearly depends on the values of the explanatory variables, and the unknown parameters. If there is omitted variable bias, or otherwise some significant specification error, this assumption is violated. Specifically, omitted variable bias can be detected through the Ramsay RESET test, which gives $p = 0.019 < \alpha = 0.05$ and $p = 0.0056 < \alpha = 0.05$ respectively. Thus there is likely some misspecification and this assumption is violated.
- MR2 states that each random errors has a probability distribution with zero mean.
- MR5 states that there is no exact collinearity or multicollinearity among the explanatory variables. We can also check the VIF tables, given in the appendix. The mean VIF for the two regressions are 1.3 and 1.54, indicating that there is not a problemtic amount of multicollinearity.
- MR6 states that the residuals are normally distributed, in other words $e_i \sim N(0, \sigma^2)$. While a histogram of the residuals would suggests the assumption of normality to be more or less reasonable, we can calculate the Jarque-Bera statistic, taking the values of skewness and kurtosis from Stata's summarize.

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$
$$JB_1 = \frac{17919}{6}\left((-0.4579588)^2 + \frac{(4.392981-3)^2}{4}\right) = 2075.09 > 5.99$$
$$JB_2 = \frac{17919}{6}\left((-0.4593643)^2 + \frac{(4.394376-3)^2}{4}\right) = 2081.85 > 5.99$$

With 5.99 being the critical value for the corresponding χ^2_2 distribution, we can conclude at the 5% significance level that the residuals do not show follow a normal distribution. However, to what extent this statistical significance achieves a practical significance is to be determined.

Overall, the most concerning violation of the classical assumptions is MR2, or the strict exogeneity assumption. There is possible problems such simultaneity (reverse causality in the form of the dependent variable, wages, having an effect on the independent variables (e.g. education, which could be prohibitively expensive), which makes it difficult to define and infer causality in this example. Another concern is MR1, which deals with model specification. There is likely some misspecification error in the form of omitted variable bias, which can similarly impact our coefficient estimates.[2]

[2] #statisticalinference: I assess, using the Ramsay RESET test for MR1, VIF for MR5, and the Jarque-Bera for MR6, whether some of the Gauss-Markov assumptions for the coefficient estimates from the two regressions to be BLUE. I use appropriate statistics, such as an explicit calculation of the Jarque-Bera statistic, which follows the chi-squared distribution, or the Ramsay RESET, which uses the F-test to support my argument. Finally, I justify the possible violations to the exogeneity assumption (MR2) by referring to the dataset.

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:

$$H_0 : \frac{\partial demand}{\partial P_{nc}} = \frac{\partial demand}{\partial P_{uc}} \iff \beta_{pnc} - \beta_{puc} = 0, \quad H_A : \frac{\partial demand}{\partial P_{nc}} \neq \frac{\partial demand}{\partial P_{uc}} \iff \beta_{pnc} - \beta_{pnc} \neq 0$$

Because we are testing only one set of linear combination of coefficients, we can use the t-test. We obtain the test statistic as follows, and obtain estimates for variances and covariance from Stata:

$$t = \frac{\bar{x} - \mu}{SE} = \frac{\hat{\beta}_{pnc} - \hat{\beta}_{puc}}{\sqrt{\text{Var}(\hat{\beta}_{pnc} - \hat{\beta}_{p}uc)}} = \frac{\hat{\beta}_{pnc} - \hat{\beta}_{puc}}{\sqrt{\text{Var}(\hat{\beta}_{pnc}) + \text{Var}(\hat{\beta}_{puc}) + 2\text{Cov}(\hat{\beta}_{pnc}, \hat{\beta}_{puc})}} \sim t_{42}(0, \text{Var}(\hat{\beta}_{pnc} - \hat{\beta}_{p}uc))$$
$$= \frac{(0.0005774) - (-0.0058746)}{\sqrt{.00016497 + .00002372 + 2(0.000009359)}} = 0.448 < t^*_{42} = 2.02$$

Thus we fail at 5% significance level ($p = 0.656 > 0.05$) to reject the null hypothesis that there is a difference between the two coefficients.[3]

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. computer-generated 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

[3] #econometrictheory: I develop the t-test for a linear combination of parameters, arguing for a manual calculation which is more exact than Stata, which might use an approximation of the variance via the delta method. I use the correct formula for estimating the variance, clearly show my work, and conclude that the test is not significant, justifying the analysis.

transportation in 2004 can also be computed manually via:

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

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: When using a log-log model, as follows

$$\ln \text{demand} = \beta_0 + \beta_{\text{GasP}} \ln(\text{GasP}) + \beta_{\text{income}} \ln(\text{Income}) + \beta_{\text{ppt}} \ln(\text{PPT}) + \cdots$$

Partial-deriving both sides with respect to GasP, Income and PPT gives

$$\begin{aligned}
\frac{\partial \text{ demand}}{\partial \text{ GasP}}\frac{1}{\text{demand}} = \frac{\beta_{\text{GasP}}}{\text{GasP}} &\iff \beta_{\text{GasP}} = \frac{\partial \text{ demand}}{\partial \text{ GasP}}\frac{\text{GasP}}{\text{demand}} = \underset{\text{(SE)}}{\widehat{\text{PED}}} = \underset{(.0540101)}{0.0605177}\\
\frac{\partial \text{ demand}}{\partial \text{ Income}}\frac{1}{\text{demand}} = \frac{\beta_{\text{Income}}}{\text{Income}} &\iff \beta_{\text{Income}} = \frac{\partial \text{ demand}}{\partial \text{ Income}}\frac{\text{Income}}{\text{demand}} = \underset{\text{(SE)}}{\widehat{\text{YED}}} = \underset{(0.2503763)}{0.9929907}\\
\frac{\partial \text{ demand}}{\partial \text{ PPT}}\frac{1}{\text{demand}} = \frac{\beta_{\text{PPT}}}{\text{Income}} &\iff \beta_{\text{PPT}} = \frac{\partial \text{ demand}}{\partial \text{PPT}}\frac{\text{PPT}}{\text{demand}} = \underset{\text{(SE)}}{\widehat{\text{XED}}} = \underset{(0.136449)}{.0192726}
\end{aligned}$$

Thus the coefficient estimates can similarly be used to estimate the elasticities. The tables for the regression output is given in the appendix. However, a key issue with doing this is assumes constant elasticity (isoelasticity), which is not necessarily the case, and has clear implications, such as in the calculation for PED or YED. The price elasticity of demand should be negative (increases in price shower lower demand), and was previously estimated to be -0.22 in 2004, but has difficulty establishing even the non-positiveness of the price elasticity - most likely because of the prominence of oil crises, such as the oil crises of 1973 and 1979. During such crises, the shortage of oil, real or perceived, led to concerns (and sometimes panic). The price of oil endogenously affects, and is affected by demand, creating a feedback loop in which higher prices cause panic and greater demand, and thus there were periods in which the price elasticity of demand was positive.

Similarly, regression did not find the estimate of the cross-price elasticity significant from 0. In c, the cross-price elasticity with respect to the price of transport in 2004 is positive, as would be expected of substitutes. Thus the specification in d. fails to capture and take into account fluctuations within the elasticities themselves over time, and the specification in a. is preferred, as it allows for more accurate point-price elasticities to be computed.[4]

[4] #specification: I use economic reasoning to evaluate if an isoelastic model is reasonable in the context of oil prices, arguing that the model does not accurately estimate the price elasticity of demand, because of unusual economic patterns in the dataset. I refer to a time series of gasoline prices, which showed a different trend in oil crises, such as during 1973 and 1979. I thoroughly explain the economic mechanisms (mania and panic) that violate the traditional demand and supply model.

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): 1200

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

personid	educ	logwage	potexper	timetrnd	ability	mothered	fathered	brknhome	siblings
1	13	1.82	1	0	1	12	12	0	1
2	15	2.14	4	6	1.5	12	12	0	1
3	10	1.56	1	2	-0.36	12	12	1	1
4	12	1.85	1	3	0.26	12	10	1	4
5	15	2.41	2	3	0.3	12	12	1	1
6	15	1.83	2	2	0.44	12	16	0	2
7	15	1.78	3	8	0.91	12	12	0	1
8	13	2.12	4	0	0.51	12	15	1	2
9	13	1.95	2	0	0.86	12	12	0	2
10	11	2.19	5	3	0.26	12	12	0	2
11	12	2.44	1	2	1.82	16	17	1	2
12	13	2.41	4	2	-1.3	13	12	0	5
13	12	2.07	3	0	-0.63	12	12	1	4
14	12	2.2	6	4	-0.36	10	12	1	2
15	12	2.12	3	4	-0.28	10	12	1	3

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 and X_2

```
      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 y on X_1 and X_2, no intercept

```
      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
------------------------------------------------------------------------------
```

Ramsay RESET in the Regression of y on X_1

```
Ramsey RESET test using powers of the fitted values of logwage
       Ho:  model has no omitted variables
               F(3, 17912) =      3.32
                  Prob > F =      0.0190
```

Ramsay RESET in the Regression of y on X_1 and X_2

```
Ramsey RESET test using powers of the fitted values of logwage
       Ho:  model has no omitted variables
               F(3, 17909) =      4.20
                  Prob > F =      0.0056
```

Variance Inflation Factor in the Regression of y on X_1

```
             |     educ potexper  ability
-------------+---------------------------
        educ |   1.0000
    potexper |  -0.2187   1.0000
     ability |   0.5279  -0.2222   1.0000
```

Variance Inflation Factor in the Regression of y on X_1 and X_2

```
    Variable |       VIF       1/VIF
-------------+----------------------
    mothered |      2.00    0.500127
    fathered |      1.98    0.505565
     ability |      1.63    0.614889
        educ |      1.46    0.686681
    siblings |      1.12    0.889541
    potexper |      1.07    0.933196
-------------+----------------------
    Mean VIF |      1.54
```

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
------------------------------------------------------------------------------
```

Full .Do File

```
clear all
import delimited Koop-Tobias.csv

egen first_obs = min(timetrnd), by(personid)
list personid educ logwage potexper timetrnd ability mothered fathered brknhome siblings if first_obs == t
global x1 educ potexper ability
global x2 mothered fathered siblings
global y logwage
regress $y $x1
regress $y $x1 $x2
regress $y $x1 $x2, noconstant

// MR1, specification (omitted variable) bias
qui reg $y $x1
estat ovtest
qui reg $y $x1 $x2
estat ovtest

// MR2, exogeneity
qui reg $y $x1
predict res1, resid
reg $y $x1 res1
test res1
qui reg $y $x1 $x2
predict res2, resid
qui reg $y $x1 $x2 res2
test res2

// MR5, collinearity
corr $x1
qui reg logwage educ potexper ability
estat vif
corr $x1 $x2
qui reg logwage educ potexper ability mothered fathered siblings
```

```
estat vif


// MR6, normality of residuals
qui reg $y $x1
predict residuals1, resid
hist residuals1, normal
summarize residuals1, detail

qui reg $y $x1 $x2
predict residuals2, resid
hist residuals2, normal
summarize residuals2, detail


clear
import delimited TableF2-2.csv
gen demand = (gasexp*10^9) / (gasp * pop*10^3)
list in -10/l
scatter gasp year

regress demand income gasp pnc puc ppt pd pn ps year
predict resid, resid

matrix list e(V) // for manual calculation
test pnc = puc // automatic test

summarize gasp income ppt demand if year == 2004 // for manual calculation
margins if year == 2004, eyex(gasp income ppt) // automatic calculation

gen ldemand = ln(demand)
gen lgasp = ln(gasp)
gen lincome = ln(income)
gen lpnc = ln(pnc)
gen lpuc = ln(puc)
gen lppt = ln(ppt)
gen lpd = ln(pd)
gen lpn = ln(pn)
gen lps = ln(ps)

regress ldemand lincome lgasp lpnc lpuc lppt lpd lpn lps year
predict log_resid, resid

pwcorr pnc puc ppt pd pn ps

qui regress demand income gasp pnc puc ppt pd pn ps year
```

```
vif
qui regress ldemand income gasp pnc puc ppt pd pn ps year
vif

gen gasp_new = (gasp * 100) / gasp[52]
gen pnc_new = (pnc * 100) / pnc[52]
gen puc_new = (puc * 100) / puc[52]
gen ppt_new = (ppt * 100) / ppt[52]
gen pd_new = (pd * 100) / pd[52]
gen pn_new = (pn * 100) / pn[52]
gen ps_new = (ps * 100) / ps[52]

regress demand income gasp pnc puc ppt pd pn ps year
regress demand income gasp_new pnc_new puc_new ppt_new pd_new pn_new ps_new year

gen lgasp_new = log(gasp_new)
gen lpnc_new = log(pnc_new)
gen lpuc_new = log(puc_new)
gen lppt_new = log(ppt_new)
gen lpd_new = log(pd_new)
gen lpn_new = log(pn_new)
gen lps_new = log(ps_new)

regress ldemand lincome lgasp lpnc lpuc lppt lpd lpn lps year
regress ldemand lincome lgasp_new lpnc_new lpuc_new lppt_new lpd_new lpn_new lps_new year
```