Department of Economics **UN3412**
Columbia University **Fall 2021**

Solutions to Problem Set 8
Introduction to Econometrics
for both sections
Seyhan Erden and Tamrat Gashaw

Question #1 [Chapter 13]

(30p) To analyze the effect of a minimum wage increase, a famous study used a quasi-experiment for two adjacent states: New Jersey and (Eastern) Pennsylvania. A $\hat{\beta}_1^{diffs-in-diffs}$ was calculated by comparing average employment changes per restaurant between the treatment group (New Jersey) and the control group (Pennsylvania). The difference-in-difference estimate $\hat{\beta}_1^{diffs-in-diffs}$ turned out to be 2.76 with a standard error of 1.36.

The authors also used a difference-in-differences estimator with additional regressors of the type

$$\Delta Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1,t} + \ldots + \beta_{1+r} W_{r,i} + u_i$$

where $i = 1, \ldots, 410$. X is a binary variable taking on the value one for the 331 observations in New Jersey. Since the authors looked at Burger King, KFC, Wendy's, and Roy Rogers fast food restaurants and the restaurant could be company owned, four W-variables were added.

(a) Given that there are four chains and the possibility of a company ownership, why did the authors not include five W-variables?

(b) OLS estimation resulted in $\hat{\beta}_1$ of 2.30 with a standard error of 1.20. Test for statistical significance and specify the alternative hypothesis.

(c) Why is this estimate different from the number calculated from $\Delta\overline{Y}^{treatment} - \Delta\overline{Y}^{control} = 2.76$? What is the advantage of employing this estimator of the simple difference-in-difference estimator?

(d) Let the vertical axis of a figure indicate the average employment fast food restaurants. There are two time periods, t = 1 and t = 2, where time period is measured on the horizontal axis. The following table presents average employment levels per restaurant for New Jersey (the treatment group) and Eastern Pennsylvania (the control group).

	PA	*NJ*
FTE Employment before	23.33	20.44
FTE Employment after	21.17	21.03

Enter the four points in the figure and label them $\overline{Y}^{treatment,before}$, $\overline{Y}^{treatment,after}$, $\overline{Y}^{control,before}$, and

$\overline{Y}^{control\ ,after}$. Connect the points. Finally calculate and indicate the value for $\hat{\beta}_1^{diffs-in-diffs}$

Answer:

(a) Including a fifth W-variable would have resulted in perfect multicollinearity.
(b) The t-statistic is +1.92. If the alternative hypothesis was $H_1 : \beta_1 < 0$, then you cannot reject the null hypothesis. If the alternative hypothesis was $H_1 : \beta_1 \neq 0$, then you cannot reject the null hypothesis at the 5% level, although you can at the 10% level. The choice of alternative hypothesis depends on prior expectations, and standard economic theory would suggest $H_1 : \beta_1 < 0$.
(c) The difference is small in terms of the standard error and may be due to sample variation. Although the difference-in-difference estimator is consistent, the difference-in-difference estimator with additional regressors can be more efficient. It is different because it stems from using the multiple regression model

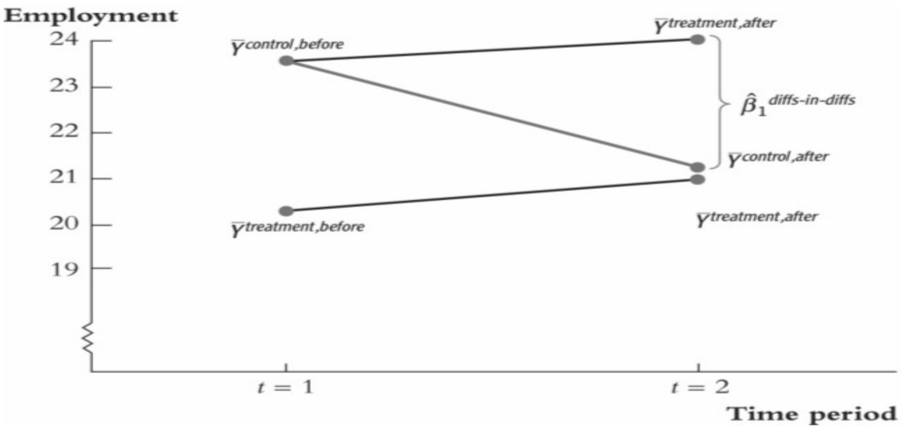
$$\Delta Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + ... + \beta_{1+r} W_{ri} + u_i,\ i = 1,..., n$$

rather than the regression with a single regressor

$$\Delta Y_i + \beta_0 + \beta_1 X_i + u_i,\ i = 1,..., n$$

and $E(u_i \,|\, X_i, W_{1i}, ..., W_{ri}) = 0$ may not hold. In that case, $\hat{\beta}_1$ is consistent as long as there is conditional mean independence. The inclusion of the characteristics also allows for testing for random receipt of treatment and random assignment using the usual F-statistic in auxiliary regressions.

(d) $\boldsymbol{\hat{\beta}_1^{diffs-in-diffs}} = \boldsymbol{\Delta\overline{Y}^{treatment}} - \boldsymbol{\Delta\overline{Y}^{control}}$= (21.03 - 20.44) - (21.17 - 23.33) = 2.75. See also the figure.



Question #2 [Chapter 14]

(40p) This question is based on an article authored by Hal R. Varian (2014) entitled "*Big Data: New Tricks for Econometrics*" published in The Journal of Economic Perspectives, Vol. 28, No. 2, pp. 3-27. Although you can download the article by searching it on CLIO, this article is posted together with this problem set. In his paper, the author describes a few of the tools for manipulating and analyzing big data. The author believes that these methods have a lot to offer and should be more widely known and used by economists.

You are required to read this article in its entirety and answer the following questions based on your reading.

a. (5p) Briefly summarize the article in your own words by focusing on the purpose of the paper and main message of the article.
b. (5p) What are the tools to manipulate Big Data?
c. (5p) What are the tools to analyze Big Data?
d. (5p) What are the general considerations for prediction?
e. (5p) What is a classification problem? How is it related to economists' explanatory variables or predictors?
f. (5p) In chapter 11, we used and discussed the HMDA data. On page 13, the author provides a data tree based on this data. Comment generally on the tree and his discussion in relation to our findings in chapter 11 (i.e., Table 11.2).
g. (5p) What is Boosting, Bagging, and Bootstrap (BBB)?
h. (5p) Discuss the variable selection section of the paper? What is the lasso and ridge regressions?

Solution:

a. This paper is on how to incorporate some data manipulating and analyzing tools into the traditional econometrics that mainly focuses on relatively small sized datasets compared to Big Data. The author has described a few of these tools in this article. He has also forwarded for economists as to how and why they need to learn about these new tools to augment their econometrics knowledge.
b. The big data manipulation tools include:
    - How to store the big data with more than a million rows?

      "If you have more than a million or so rows in a spreadsheet, you probably want to store it in a relational database, such as MySQL. Relational databases offer a flexible way to store, manipulate, and retrieve data using a Structured Query Language (SQL), which is easy to learn and very useful for dealing with medium-sized datasets."

    - How to store the big data with several gigabytes of data or several million observations?

      "Databases to manage data of this size are generically known as "NoSQL" databases. The term is used rather loosely, but is sometimes interpreted as meaning "not only

SQL." NoSQL databases are more primitive than SQL databases in terms of data manipulation capabilities but can handle larger amounts of data."

- How to store and process billions of computer-mediated transactions per day?

Students can discuss here the Google's example and the need for developing different tools to manage and analyze big data. They can also answer this question using Table 1.

c. The tools to analyze big data can include:

- The application of random sampling

"At google … random samples on the order of 0.1 percent work fine for analysis of busines."

- Conducting some exploratory data analysis and data-cleaning tasks.
- Data analysis.

"*Data analysis in statistics and econometrics can be broken down into four cate gories: 1) prediction, 2) summarization, 3) estimation, and 4) hypothesis testing. Machine learning is concerned primarily with prediction…"*

- Focusing on Prediction Problems

"*When confronted with a prediction problem of this sort an economist would think immediately of a linear or logistic regression. However, there may be better choices, particularly if a lot of data is available. These include nonlinear methods such as 1) classification and regression trees (CART); 2) random forests; and 3) penalized regression such as LASSO, LARS, and elastic nets …"*

d. There are three general (steps) considerations in prediction. These are (i) *regularization* (i.e., making the model simpler by penalizing models to get a good out-of-sample (OOS) forecasts; (ii) *dividing* the data into separate sets for the purpose of estimation, testing, and validation; and (iii) *k-fold cross validation* to reduce the complexity of the model and obtain better OOS.

e. A classification problem is a phrase used in machine learning that focuses on what characteristics of the data to use to classify the dependent variable into 0 or 1 binary variable. In economics, we use a generalized linear model like logit or probit for a classification problem using the right-hand side explanatory/predictor variables.

f. Figure 5 of this article reports the same results of our textbook's Table 11.2. The same statistically significant variables in our table are also significant in figure 5. You can use the "tabulate deny" to get the black bar of the last box in figure 5. Data tree as shown in figure 5 is another way of presenting the results of data analysis. It is a suitable method/format of reporting the result if the dataset is big data.

g. Read page 14 of the article. The BBB are ways to improve classifier performance by adding more randomness to the data.

"***Bootstrap*** *involves choosing (with replacement) a sample of size n from a dataset of size n to estimate the sampling distribution of some statistic. A variation is the "m out of n bootstrap" which draws a sample of size m from a dataset of size n> m.* ***Bagging*** *involves averaging across models estimated with several different boot strap samples in order to improve the performance of an estimator.* ***Boosting*** *involves repeated estimation where misclassified observations are given increasing weight in each repetition. The final estimate is then a vote or an average across the repeated estimates. Econometricians are well-acquainted with the bootstrap but rarely use the other two methods.*"

h. In big data, one of the main challenges is the presence of many predictors in the data. That means we are facing the problem of which variables to select that are very important without hurting our prediction performance. These two methods are ways of finding minimized MSPE by adding another restriction on our OLS methods. In the textbook, see equation 14.7 for Ridge and equation 14.9 for Lasso specification.

Question #3 [Chapter 15 and 16]

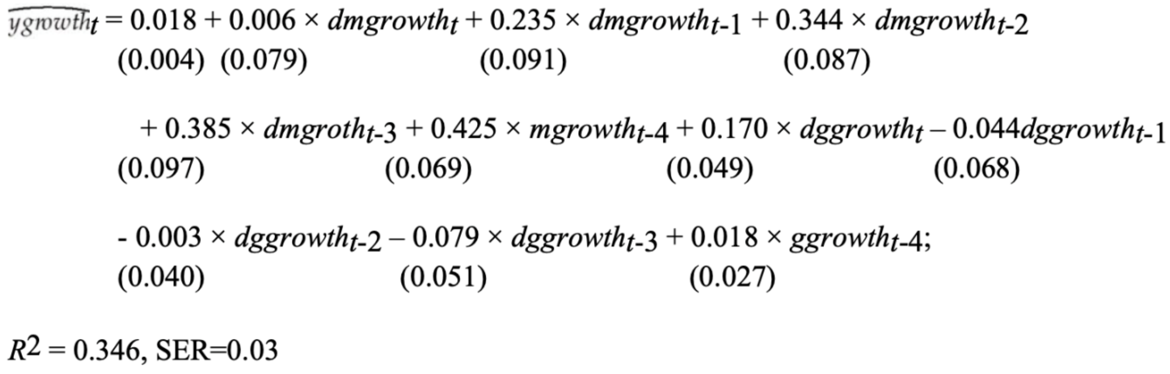
(60p) A model that attracted quite a bit of interest in macroeconomics in the 1970s was the St. Louis model. The underlying idea was to calculate fiscal and monetary policy impact and long run cumulative dynamic multipliers, by relating real output (growth) to real government expenditure (growth) and real money supply (growth). The assumption was that both government expenditures and the money supply were exogenous.

a. (6p) Visit the Federal Reserve Bank of St. Louis at https://fred.stlouisfed.org/ where you have access to the Fred Economic data and download the data for the required three variables (i.e., real GDP (GDPC1), real money supply (M2REAL), and real government expenditure (GCEC1)). The sample period should be from first quarter of 1960 to the fourth quarter of 2019 (i.e., 1960q1 – 2019q4). The real money supply (M2REAL) is available on a monthly frequency basis and don't forget to convert it into a quarterly frequency variable to match it with the other two variables. [*Hint*: Is M2REAL a stock or flow variable? Although, you may take the last month of each quarter or the three-month average as your quarterly value, here use the **first** month of each quarter.]

b. (6p) Import your data into Stata from the excel/csv files that you choose to download your data. You can also import these series directly from Fred data using a special Stata command (*freduse*) as shown https://www.youtube.com/watch?v=iiizhsX-I00. However, the data frequency should be similar not to mix a monthly data with a quarterly data in the same Stata file. You can use this approach for GDPC1 and GCEC1 but you need to add M2REAL separately. Then create a *time* variable that has a quarterly format and let STATA know that *time* is the variable you want to indicate the data is time series. And compute (i.e., generate) growth rate of these three variables after you first transform them into natural logarithm and name/label them *ygrowth*, *mgrowth*, and *ggrowth*.

In order to investigate the effect of a fiscal and monetary policies on output, you want to estimate a St. Louis type model using your quarterly data (i.e., make sure to use HAC standard errors) and report your results. That is run a distributed lag model (DLM) using your dependent variable, (i.e., *ygrowth*) on:

c. (6p) Current period *mgrowth* and see the effect of a monetary policy on current quarter's output growth. Interpret the coefficient, make sure you are obtaining the correct standard errors. Is the effect significant?
d. (6p) Current period *ggrowth* and see the effect of a fiscal policy on current quarter's output growth. Interpret the coefficient, make sure you are obtaining the correct standard errors. Is the effect significant?
e. (6p) The effect of current and next quarter's monetary policy on output growth, making sure the standard errors are correct. What is the impact multiplier? Explain the meaning. What is the cumulative multiplier? Explain the meaning.
f. (6p) The effect of current and next quarter's fiscal policy on output growth, making sure the standard errors are correct. What is the impact multiplier? What is the cumulative multiplier?
g. (6p) The **change (i.e., first difference)** in current and four lags of *mgrowth* and *ggrowth* (i.e., to mimic the original St. Louis Equation) to obtain a regression like the one shown below (i.e., your regression coefficients should be different from this one) and report your results:



$$\widehat{ygrowth}_t = \underset{(0.004)}{0.018} + \underset{(0.079)}{0.006} \times dmgrowth_t + \underset{(0.091)}{0.235} \times dmgrowth_{t-1} + \underset{(0.087)}{0.344} \times dmgrowth_{t-2}$$

$$+ \underset{(0.097)}{0.385} \times dmgroth_{t-3} + \underset{(0.069)}{0.425} \times mgrowth_{t-4} + \underset{(0.049)}{0.170} \times dggrowth_t - \underset{(0.068)}{0.044} dggrowth_{t-1}$$

$$- \underset{(0.040)}{0.003} \times dggrowth_{t-2} - \underset{(0.051)}{0.079} \times dggrowth_{t-3} + \underset{(0.027)}{0.018} \times ggrowth_{t-4};$$

$R^2 = 0.346$, SER=0.03

h. (6p) Assuming that money and government expenditures are exogenous, what do the coefficients represent? Calculate the *h*-period cumulative dynamic multipliers from these. How can you test for the statistical significance of the cumulative dynamic multipliers and the long-run cumulative dynamic multiplier?

Lag number	Monetary Dynamic Multiplier	Monetary Cumulative Multiplier	Fiscal Dynamic Multiplier	Fiscal Cumulative Multiplier
0				
1				
2				
3				

4				

i. (6p) Sketch the estimated dynamic and cumulative dynamic fiscal and monetary multipliers (i.e., something similar to Fig 16.2 in your textbook).

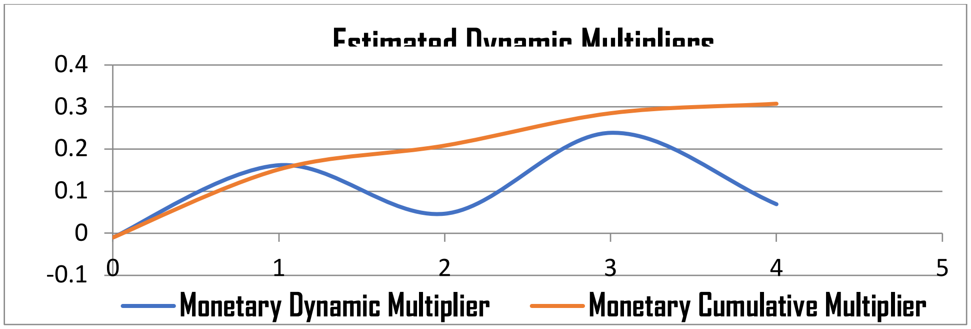
j. (6p) For these coefficients to represent dynamic multipliers, the money supply and government expenditures must be exogenous variables. Explain why this is unlikely to be the case. As a result, what importance should you attach to the above results?

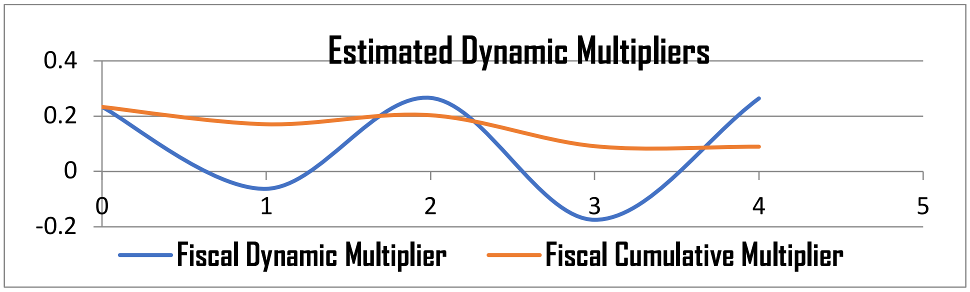
Solutions:

```
*QUESTION #2
*** 2a. Go to FRED Data website, and download the following three variables. The variables that
you are looking for are GDPC1, GEEC1, and M2REAL from 160q1 to 2019q4. The first two variables are
in quarterly frequency while the third one is in monthly frequency.***
*** You can first download them into excel or csv file format and arrange it in the requested
format before you import it into STATA. Alternatively, you can use "freduse" command and directly
import the first two into STATA and later add the third one (i.e., which is a monthly series) after
you make convert it into quarterly frequency separately.***

*** Since M2REAL is a stock variable, you can take the value of this series corresponding to the
last month of the quarters to get the quarterly values of this variable.***

*** 2b. Importing the data into Stata and creating a time variable that has a quarterly format.
Generating natural log of the three variables and compute the growth rates.***

gen time1=date(date, "YMD")
gen time = qofd(time1)
format time %tq
tsset time, quarterly
gen y=ln(gdpc1)
gen m=ln(m2real)
gen g = ln(gcec1)
gen ygrowth = D.y
*(1 missing value generated)
gen mgrowth = D.m
*(1 missing value generated)
gen ggrowth=D.g
*(1 missing value generated)

*** 2c.
*m = 0.75T1/3
gen mm=0.75*(240^(1/3))
disp mm
*4.6608486 = 5 lags
newey ygrowth mgrowth if tin(1960q1, 2019q4), lag(5)
*** 2d.
newey ygrowth ggrowth if tin(1960q1, 2019q4), lag(5)
*** 2e.
newey ygrowth mgrowth l.mgrowth if tin(1960q1, 2019q4), lag(5)
*** 2f.
newey ygrowth ggrowth l.ggrowth if tin(1960q1, 2019q4), lag(5)
*** 2g.
gen dmgrowth=D.mgrowth
*(2 missing values generated)
gen dggrowth = D.ggrowth
*(2 missing values generated)
newey ygrowth dmgrowth l.dmgrowth l2.dmgrowth l3.dmgrowth l4.mgrowth dggrowth l.dggrowth
l2.dggrowth l3.dggrowth l4.ggrowth if tin(1960q1, 2019q4), lag(5)

*** 2g. TO REPLICATE THE GIVEN RESULT IN 2g USING SUB SAMPLE.***
newey ygrowth dmgrowth l.dmgrowth l2.dmgrowth l3.dmgrowth l4.mgrowth dggrowth l.dggrowth
l2.dggrowth l3.dggrowth l4.ggrowth if tin(1960q1, 1980q4), lag(5)
* Replicating the equation in the question.

*** 2h.
```

Lag number	Monetary Dynamic Multiplier	Monetary Cumulative Multiplier	Fiscal Dynamic Multiplier	Fiscal Cumulative Multiplier
0	-0.0100708	-0.0100708	0.2334858	0.2334858
1	0.1618472	0.1517764	-0.0626559	0.1708299
2	0.0465711	0.2084183	0.2662043	0.2035484
3	0.2388277	0.2853988	-0.1746814	0.0915229
4	0.0691839	0.3080116	0.2643714	0.08969

*** 2i.





*** 2j.

*There is little reason to believe that these government instruments are exogenous. Even if the monetary base and those components of government expenditures which do not respond to business cycle fluctuations had been chosen rather than the above regressors, then these instruments respond to changes in the growth rate of GDP. As a matter of fact, government reaction functions were also estimated at the time to capture how government instruments respond to changes in target variables. As a result, the regressors will be correlated with the error term, OLS estimation is inconsistent, and inference not dependable. It is hard to imagine how useable information can be retrieved from these numbers.

Following questions will not be graded, they are for you to practice and will be discussed at the recitation:

1. SW Exercise 14.1.
2. SW Exercise 14.2
3. SW Exercise 14.5.
4. SW Empirical Exercise 15.2.
5. SW Empirical Exercise 16.1.

Solutions for Practice Questions:

SW Exercise 14.1.

14.1 (a.i) $X_{RPM}^{oos} = (0.52 - 0.60)/0.28 = -0.289$ and $X_{TExp}^{oos} = (11.1 - 13.2)/3.8 = -0.553$

(a.ii) $\hat{Y} = 750.1 - 48.7 \times (-0.289) + 8.7 \times (-0.553) = 759.4$

(b) $Y - \hat{Y} = 775.3 - 759.4 = 15.9$

(c) $\widehat{\left(TestScore^* - 750.1\right)} = -48.7 \times (RPM^* - 0.60)/0.28 + 8.7 \times (TExp^* - 13.2)/3.8$

Rearranging yields:

$\widehat{TestScore}^* = 824.2 - 173.9 \times RPM^* + 2.29 \times TExp^*$

(d) $\widehat{TestScore}^* = 824.2 - 173.9 \times 0.52 + 2.28 \times 1.11 = 759.1$, where the difference arises from rounding error.

SW Exercise 14.2.

14.2 No.

SW Exercise 14.5.

14.5 (a.i) The best predictor is $E(Y) = \mu = 2$.

(a.ii) The MSPE is the variance of Y, which is $\sigma^2 = 25$.

(b.i) $Y - \bar{Y} = (Y - \mu) + (\mu - \bar{Y})$ obtains by subtracting and adding μ to $Y - \bar{Y}$.

(b.ii) $E(Y - \mu) = E(Y) - \mu = \mu - \mu = 0$

$E(Y - \bar{Y}) = E(Y) - E(\bar{Y}) = \mu - \mu = 0$.

(b.iii) $\bar{Y}$ is computed from in-sample Ys, Y is out-of-sample and is independent of the in-sample observations. Because Y and $\bar{Y}$ are independent, they are uncorrelated.

(b.iv)

$$\begin{aligned}
MSPE &= E\left[(Y - \bar{Y})^2\right] \\
&= E\left[\left\{(Y - \mu) - (\bar{Y} - \mu)\right\}^2\right] \\
&= E\left[(Y - \mu)^2\right] + E\left[(\bar{Y} - \mu)^2\right] + 2E\left[(Y - \mu)(\bar{Y} - \mu)\right] \\
&= \operatorname{var}(Y) + \operatorname{var}(\bar{Y})
\end{aligned}$$

where the first line is the definition of the MSPE, the second follows from (b.i), the third expands the second, and the final line uses (b.iii) and the definition of variance.

(b.v) var(Y) = 25 and var($\bar{Y}$) = 25/n. The result follows because n = 10.

SW Empirical Exercise 15.2.

(a) Table 15.2 Extended Dataset (Sample Period 1932:1–2002:12)

Regressors	**(1)**	**(2)**	**(3)**
Excess Return$_{t-1}$	0.098 (0.061)	0.102 (0.061)	0.099 (0.058)
Excess Return$_{t-2}$		–0.040 (0.057)	–0.029 (0.054)
Excess Return$_{t-3}$			–0.098 (0.054)
Excess Return$_{t-4}$			0.006 (0.046)
Intercept	0.524 (0.181)	0.543 (0.186)	0.590 (0.199)
F-statistic on all coefficients (p-value)	2.61 (0.11)	1.51 (0.22)	1.41 (0.23)
$\bar{R}^2$	0.009	0.009	0.016

(b-c)

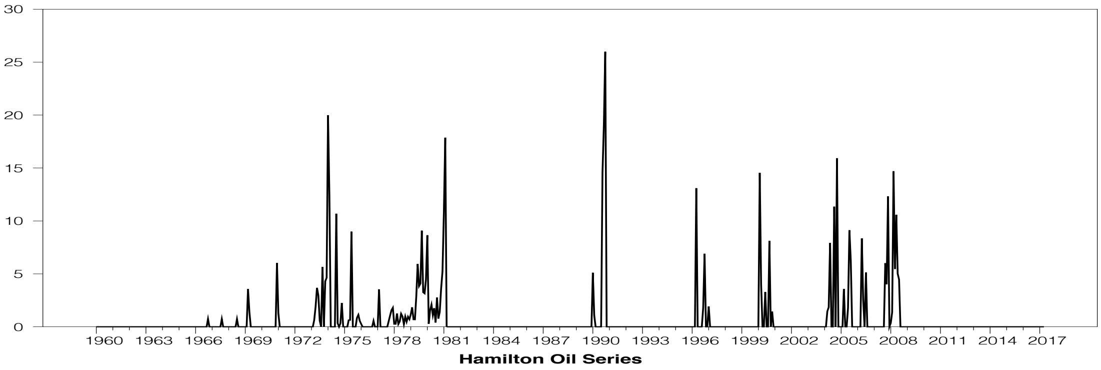
Model	**RMSFE**
Zero Forecast	4.28
Constant Forecast	4.26
AR(1)	4.28
AR(2)	4.27
AR(4)	4.29

The table shows the RMSFE for the "zero forecast" (that is, forecasting that next period's excess return will be zero), the "constant forecast" (that is, a regression that includes only a constant, equivalently the forecast from an AR(0) model), and forecasts from the AR(1), AR(2) and AR(4) models. The most accurate forecast comes from the "constant-term" forecast, consistent with the statistical significance of the coefficients in the AR models.

SW Empirical Exercise 16.1.

(a) Mean = 0.21; Standard Deviation = 0.75

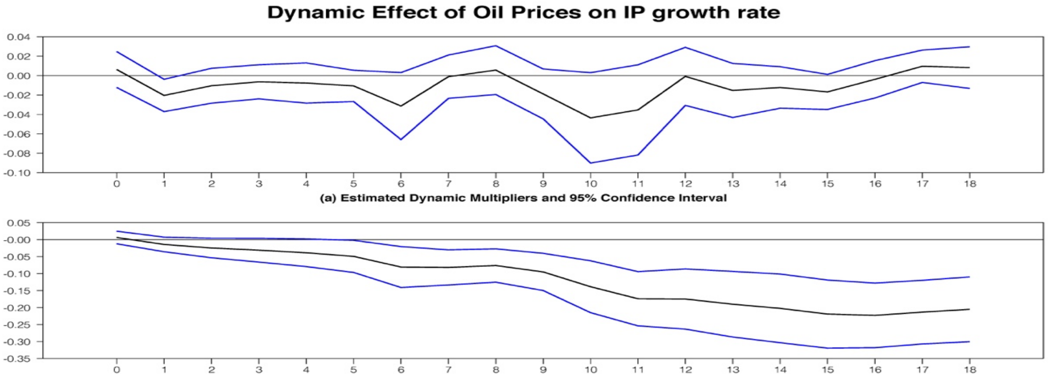
(b) O_t is the greater of zero or the percentage point difference between oil prices at date t and their maximum value during the past year. Thus $O_t \geq 0$, and $O_t = 0$ if the date t is not greater than the maximum value over the past year.



(c) m was chosen using $0.75T^{0.33}$ rounded up to the nearest integer; $m = 7$ in this case. The estimated coefficients and 95% confidence intervals are shown in the figure in part (e).

(d) The F-statistic testing that all 19 coefficients are equal to zero is 2.74, with a p-value 0.00; the coefficients are significant at the 5% but not the 1% level.

(e) The cumulative multipliers show a persistent and large decrease in industrial production following an increase in oil prices above their previous 12 month peak price. Specifically a 100% increase in oil prices is leads to an estimated 21% decline in industrial production after 18 months.



(a) Estimated Dynamic Multipliers and 95% Confidence Interval

(b) Estimated Cumulative Multipliers and 95% Confidence Interval

(f) In this case O_t is not exogenous and the results summarized in (e) are not reliable.

```
Do file for PS#8 Question 5:
import excel USMacro_Monthly.xlsx, firstrow clear
gen time = ym(Year,Month)
format time %tm
tset time
gen IP_lag1=L1.IP
gen ip_growth=100*ln(IP/IP_lag1)
summarize ip_growth if tin(1952m1, 2009m12)
tsline Oil
gen D_Oil=D.Oil
* non-cumulative effects
newey ip_growth L(0/18).Oil if tin(1947m1, 2009m12), lag(7)
testparm L(0/18).Oil
* cumulative effects
newey ip_growth L(0/17).D_Oil L(18/18).Oil if tin(1947m1, 2009m12), lag(7)
* one can plot the graphs in excel
```