CHAPTER 2.

REVIEW OF THE LINEAR REGRESSION MODEL

SOLUTIONS

by
Wei Lin and Yingying Sun
(University of California, Riverside)

Exercise 1

a. Conditional sample mean

$$\bar{y}_{Y|X=4} = \frac{0+15+6+5+2+2+2+2+2+6}{10} = 4.2$$

b. To calculate the conditional sample variance $\hat{\sigma}^2_{Y|X=5}$ given $X=5$, first calculate the conditional sample mean $\bar{y}_{Y|X=5}$,

$$\bar{y}_{Y|X=5} = \frac{20+16+6+7+5+5+4+6+8+16}{10} = 9.3.$$

Then, the conditional sample variance is

$$\begin{aligned}
\hat{\sigma}^2_{Y|X=5} &= [(20-9.3)^2 + (16-9.3)^2 + (6-9.3)^2 + (7-9.3)^2 + (5-9.3)^2 \\
&\quad +(5-9.3)^2 + (4-9.3)^2 + (6-9.3)^2 + (8-9.3)^2 + (16-9.3)^2]/9 \\
&\approx 33.1
\end{aligned}$$

c. Figure 1 shows the conditional histogram (in the vertical axis, read the number of counts) when $X=2$.

d. To calculate the unconditional mean and standard deviation of Y, just calculate the sample mean and standard deviation for the whole sample of Y (regardless of the values of X). Therefore,

$$\begin{aligned}
\bar{y} &= 4.232 \\
\hat{\sigma}_Y &= \sqrt{\hat{\sigma}^2_Y} \approx 4.825
\end{aligned}$$

Exercise 2

The data for U.S. quarterly GDP and Standard & Poor's (SP) 500 Index are available on FRED and YAHOO! Finance websites. The sample ranges from 1950Q2 to 2012Q1. Let $GRGDP$ and $RETURN$ denote the quarterly growth rate of GDP and S&P500 index returns respectively. Figure 2 and Figure 3 show the histograms and descriptive statistics of the two series respectively. The contemporaneous sample correlation coefficient of these two series is approximately 0.270. The positive correlation between the two series indicates co-movements in the same direction between macroeconomic activity and the performance of financial markets such that when the economy grows, the stock market tends to be bullish, or when the growth is sluggish, the stock market tends to be bearish.

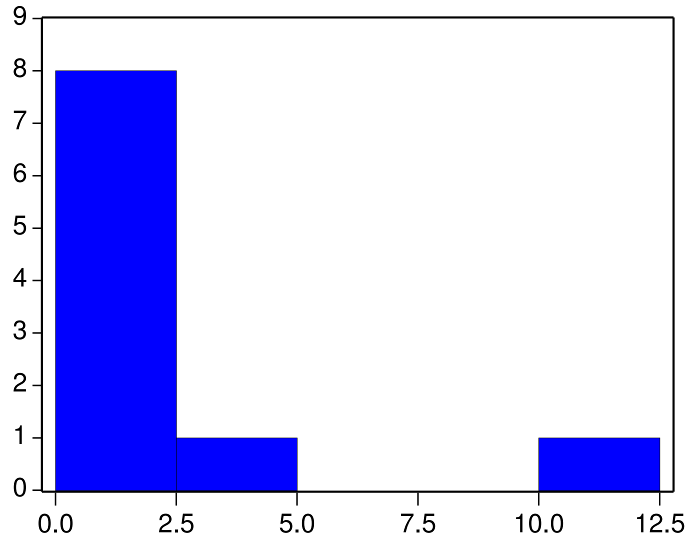

Figure 1: Conditional Histogram (counts) of Y when $X = 2$

Exercise 3

Refer to Tables 1, 2, 3, 4 and 5. R-squared is a measure of goodness of fit and it indicates the proportion of the total sample variation of Y (dependent variable) that is explained by X (independent variables). The adjusted R-squared is also a measure of goodness of fit but it penalizes the introduction of irrelevant regressors in the model. We prefer model (d) because it has the largest adjusted R^2. When the regression includes several regressors, we assess the goodness of fit with the adjusted R^2.

Model	R^2	Adjusted R^2
Model (a)	0.073031	0.069263
Model (b)	0.164929	0.161521
Model (c)	0.206594	0.193315
Model (d)	0.247690	0.231885

Table 1: R-squared and adjusted R-squared

Exercise 4

Refer to Tables 2 and 3

Model (3a) From Table II (Appendix B), we obtain the 5% critical value for $df = \infty$ given that the number of observations is quite large. For the two-tailed test, the critical value is 1.960. The t-statistic for $H_0 : \beta_1 = 0$ is 3.426827, which is larger than 1.960. Therefore, we reject $H_0 : \beta_1 = 0$ in favor of the alternative $H_1 : \beta_1 \neq 0$. For the one-tailed test with $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 > 0$, the critical value is 1.645. The t-statistic is larger than the critical value, thus we reject the null hypothesis. For the one-tailed test with $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 < 0$, the critical value is -1.645. Now, the t-statistic falls in the acceptance region, thus we fail to reject the null. In the latter case, observe the role of the alternative hypothesis; in order to reject the null, the sample information needs to provide strong evidence for a negative β_1, which is not the case, thus 'fail to reject' is the most that we should expect from this test.

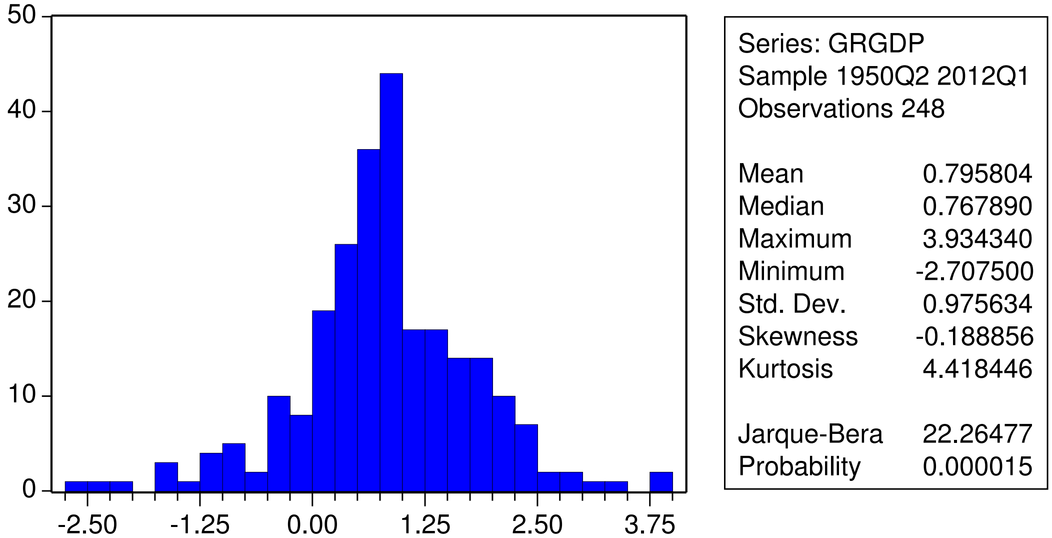

Figure 2: Histogram and Descriptive Statistics for DGP Growth Rate

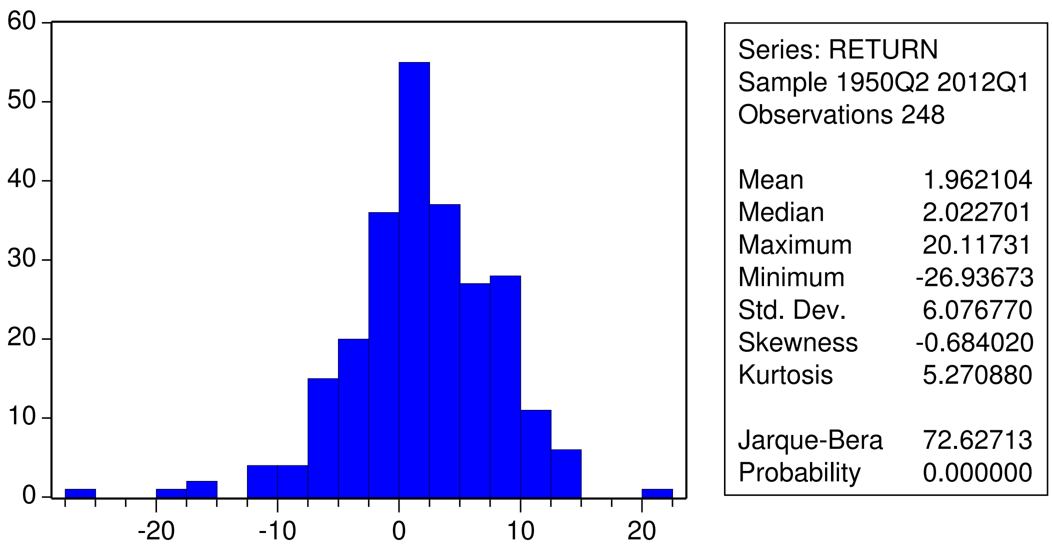

Figure 3: Histogram and Descriptive Statistics for S&P500 Returns

Dependent Variable: GRGDP				
Method: Least Squares				
Sample: 1950Q2 2012Q1				
Included observations: 248				
Newey-West HAC Standard Errors & Covariance (lag truncation=4)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.710672	0.087362	8.134837	0.0000
RETURN	0.043388	0.012661	3.426827	0.0007
R-squared	0.073031	Mean dependent var		0.795804
Adjusted R-squared	0.069263	S.D. dependent var		0.975634
S.E. of regression	0.941241	Akaike info criterion		2.724796
Sum squared resid	217.9397	Schwarz criterion		2.75313
Log likelihood	-335.8747	F-statistic		19.38108
Durbin-Watson stat	1.342014	Prob(F-statistic)		0.000016

Table 2: Estimation Results. Model (3a)

Dependent Variable: GRGDP				
Method: Least Squares				
Sample (adjusted): 1950Q3 2012Q1				
Included observations: 247 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=4)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.661818	0.0792	8.356244	0.0000
RETURN(-1)	0.064728	0.010678	6.062054	0.0000
R-squared	0.164929	Mean dependent var		0.786709
Adjusted R-squared	0.161521	S.D. dependent var		0.967023
S.E. of regression	0.885488	Akaike info criterion		2.602709
Sum squared resid	192.1019	Schwarz criterion		2.631125
Log likelihood	-319.435	F-statistic		48.38834
Durbin-Watson stat	1.475295	Prob(F-statistic)		0.000000

Table 3: Estimation Results. Model (3b)

Dependent Variable: GRGDP				
Method: Least Squares				
Sample (adjusted): 1951Q2 2012Q1				
Included observations: 244 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=4)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.571948	0.080907	7.069182	0.0000
RETURN(-1)	0.056594	0.010447	5.417203	0.0000
RETURN(-2)	0.018011	0.008314	2.166345	0.0313
RETURN(-3)	0.015672	0.008257	1.898017	0.0589
RETURN(-4)	0.011948	0.00855	1.397432	0.1636
R-squared	0.206594	Mean dependent var		0.767957
Adjusted R-squared	0.193315	S.D. dependent var		0.94932
S.E. of regression	0.852638	Akaike info criterion		2.539316
Sum squared resid	173.7511	Schwarz criterion		2.61098
Log likelihood	-304.797	F-statistic		15.5582
Durbin-Watson stat	1.549082	Prob(F-statistic)		0.000000

Table 4: Estimation Results. Model (3c)

Dependent Variable: GRGDP				
Method: Least Squares				
Sample (adjusted): 1951Q2 2012Q1				
Included observations: 244 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=4)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.44449	0.075037	5.923574	0.0000
RETURN(-1)	0.050535	0.010033	5.037049	0.0000
RETURN(-2)	0.007459	0.009295	0.802474	0.4231
RETURN(-3)	0.011149	0.008294	1.344201	0.1802
RETURN(-4)	0.007133	0.00834	0.855278	0.3933
GRGDP(-1)	0.230396	0.067098	3.433746	0.0007
R-squared	0.24769	Mean dependent var		0.767957
Adjusted R-squared	0.231885	S.D. dependent var		0.94932
S.E. of regression	0.832005	Akaike info criterion		2.494326
Sum squared resid	164.7513	Schwarz criterion		2.580322
Log likelihood	-298.308	F-statistic		15.67178
Durbin-Watson stat	2.026326	Prob(F-statistic)		0.000000

Table 5: Estimation Results. Model (3d)

Model (3b) Using the same critical values as in Model (3a), the t-statistic for $H_0 : \beta_1 = 0$ is 6.062054, which is larger than 1.960. Therefore, we reject $H_0 : \beta_1 = 0$ in favor of the alternative $H_1 : \beta_1 \neq 0$. For the one-tailed test with $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 > 0$, the 5% critical value is 1.645. The t-statistic is larger than the critical value, thus we reject the null hypothesis. For the one-tailed test with $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 < 0$, the critical value is -1.645. The t-statistic falls in the acceptance region and we fail to reject the null. However, see the comment above for the interpretation of this decision.

The strong significance of β_1 in Model (3b) indicates that the stock market is at least one-quarter leading indicator for economic growth but we should ask whether we could find additional leading time in the data. This is the objective of the next two exercises.

Exercise 5
Refer to Table 4
Model (3c) Using the same critical values as in Model (3a), the t-statistic for β_1 is 5.417203 and for β_2 is 2.166345, which are larger than 1.960. We reject the null for $\beta_1 = 0$ and for $\beta_2 = 0$ at the 5% significance level. The t-ratio for β_3 is 1.898017 and for β_4 is 1.397432, which are smaller than 1.960. We fail to reject the null for $\beta_3 = 0$ and for $\beta_4 = 0$ at the 5% significance level. This means that there is some evidence for claiming that the stock market leads for about two quarters output growth.

The F statistic for overall significance of the regression is 15.5582; this test has 4 degrees of freedom (number of restrictions) in the numerator and 239 (244-5) degrees of freedom in the denominator. The 5% critical value is about 2.37. Consequently, we reject the null hypothesis because 15.5582 $>$ 2.37, thus the overall set of regressors are informative to explain output growth.

Exercise 6
Refer to Table 5
Model (3d) Following the same guidelines as in the previous exercises, we only reject the null $H_0 : \beta_1 = 0$ at the 5% significance level but we fail to reject the null for $H_0 : \beta_2 = 0$, $H_0 : \beta_3 = 0$ and $H_0 : \beta_4 = 0$. However, β_5 is very significant; the regressor GRGDP(-1), which measures the inertia of GDP growth, is most relevant to explain growth in the next period.

The F statistic for overall significance of the regression is 15.6717, and as before, it is very significant. This is expected because of the strong significance of several regressors. We conclude that, once we control for the inertia of output growth, the stock market seems to be a leading indicator of real activity with a lead time of one quarter.

Exercise 7

The data for 'number of unemployed workers' and 'number of people in poverty' are available on FRED and U.S. Census Bureau websites. Table 6 reports the descriptive statistics for the growth rates of unemployed workers (G_UNEM) and number of people in poverty (G_POV). The correlation coefficient of the two series is 0.71. This large positive correlation means that when the number of unemployed workers increase (decrease), the number of people in poverty also tends to increase (decrease). Observe that the growth rate of unemployed people has a large dispersion compared to the growth rate of poor people. The Great Recession of 2008 was particularly sanguine by producing an increase of 59.76% in the number of unemployed people.

Sample: 1959-2010		
	G_POV (%)	G_UNEM (%)
Mean	0.448861	4.003674
Median	-0.55958	-2.37613
Maximum	12.2737	59.76986
Minimum	-14.0877	-20.245
Std. Dev.	5.37858	17.3895
Skewness	0.027338	1.425472
Kurtosis	2.885234	4.739782
Jarque-Bera	0.034341	23.70378
Probability	0.982976	0.000007
Observations	51	51

Table 6: Descriptive Statistics

Exercise 8

Let G_POV_t and G_UNEM_t denote the growth rates of number of people in poverty and unemployed persons respectively. We specify the following three regression models,

$$\begin{aligned}
&\text{Model (8a):} \quad G_POV_t = \beta_0 + \beta_1 G_UNEM_{t-1} + u_t;\\
&\text{Model (8b):} \quad G_POV_t = \beta_0 + \beta_1 G_UNEM_{t-1} + \beta_2 G_UNEM_{t-2} + \beta_3 G_UNEM_{t-3} + u_t;\\
&\text{Model (8c):} \quad G_POV_t = \beta_0 + \beta_1 G_UNEM_{t-1} + \beta_2 G_UNEM_{t-2} +\\
&\qquad\qquad\qquad + \beta_3 G_UNEM_{t-3} + \beta_4 G_POV_{t-1} + u_t.
\end{aligned}$$

Model (8a) claims that unemployment growth is one-year leading indicator for poverty growth. Model (8b) examines a more extensive dynamic relation between unemployment and poverty growth by including the past three years of unemployment growth. Model (8c) expands Model (8b) by adding previous growth in poverty, which captures the inertia of the poverty growth rate.

Tables 7, 8 and 9 report the estimation results of the three models. When there is more than one regressor in the model, we should examine the adjusted R-squared instead of the R-squared. Comparing the three models, we prefer model (8c) because it has the largest adjusted R-squared, about 32%. Models (8a) and (8b) inform about unemployment being somewhat a leading indicator of poverty but the better fit provided by Model (8c) is due to the effect of the poverty inertia.

That is, poverty growth tends to be persistent over time, positive (negative) growth is followed by positive (negative) growth. When this regressor is included, the effect of unemployment growth is greatly diminished.

Dependent Variable: G_POV				
Method: Least Squares				
Sample (adjusted): 1961 2010				
Included observations: 50 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=3)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.002186	1.016725	-0.00215	0.9983
G_UNEM(-1)	0.110109	0.042274	2.604665	0.0122
R-squared	0.126749	Mean dependent var		0.439555
Adjusted R-squared	0.108556	S.D. dependent var		5.432771
S.E. of regression	5.129423	Akaike info criterion		6.147041
Sum squared resid	1262.927	Schwarz criterion		6.223522
Log likelihood	-151.676	F-statistic		6.966994
Durbin-Watson stat	1.190826	Prob(F-statistic)		0.011165

Table 7: Estimation Results. Model (8a)

Dependent Variable: G_POV				
Method: Least Squares				
Sample (adjusted): 1963 2010				
Included observations: 48 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=3)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.257569	1.194321	0.215661	0.8302
G_UNEM(-1)	0.131781	0.047236	2.789818	0.0078
G_UNEM(-2)	-0.06389	0.036097	-1.76992	0.0837
G_UNEM(-3)	-0.01437	0.037591	-0.38223	0.7041
R-squared	0.175445	Mean dependent var		0.522257
Adjusted R-squared	0.119226	S.D. dependent var		5.527656
S.E. of regression	5.187682	Akaike info criterion		6.210106
Sum squared resid	1184.13	Schwarz criterion		6.36604
Log likelihood	-145.043	F-statistic		3.120714
Durbin-Watson stat	1.276532	Prob(F-statistic)		0.035395

Table 8: Estimation Results. Model (8b)

Exercise 9

The estimation results in Tables 7, 8 and 9 report the t-ratios corresponding to the null hypothesis of each regression coefficient to be zero as well as the F-tests for overall significance of the regressions. Let us choose a significance level of 5% and perform two-tailed t-tests, so that the critical values are -1.96 and 1.96. Models (a) and (b) show that the coefficients of G_UNEM_{t-1} are statistically significant but those of G_UNEM_{t-2} and G_UNEM_{t-3} are not. Thus, this is evidence to claim that unemployment growth leads to poverty growth with a lead time of one year. However, once we account for the inertia effect in Model (c), that is, we include a very significant regressor G_POV_{t-1} (one lag of dependent variable), unemployment growth is less relevant on leading poverty growth.

The F-tests for overall significance in the three models all reject the null hypothesis, so that the regressors considered are informative to explain poverty growth. Observe that Model (c) provides the F-test with the lowest p-value. Overall, t-ratios and F-tests point towards a relation between changes in unemployment and poverty. Given these results, the reader may be interested in estimating the following models: $G\_POV_t = \beta_0 + \beta_1 G\_UNEM_{t-1} + \beta_2 G\_POV_{t-1} + u_t$ and

Dependent Variable: G_POV				
Method: Least Squares				
Sample (adjusted): 1963 2010				
Included observations: 48 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=3)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.627246	0.813263	0.771271	0.4448
G_UNEM(-1)	-0.0131	0.056898	-0.23014	0.8191
G_UNEM(-2)	-0.08934	0.033879	-2.63715	0.0116
G_UNEM(-3)	-0.01451	0.033168	-0.43738	0.6640
G_POV(-1)	0.674105	0.141986	4.74767	0.0000
R-squared	0.379635	Mean dependent var		0.522257
Adjusted R-squared	0.321927	S.D. dependent var		5.527656
S.E. of regression	4.551759	Akaike info criterion		5.967237
Sum squared resid	890.8958	Schwarz criterion		6.162154
Log likelihood	-138.214	F-statistic		6.57852
Durbin-Watson stat	2.082441	Prob(F-statistic)		0.000319

Table 9: Estimation Results. Model (8c)

$G\_POV_t = \beta_0 + \beta_1 G\_POV_{t-1} + u_t$, and assess whether unemployment plays any role in poverty growth.

Exercise 10
Note to the instructor: This exercise is designed to bring some warnings on regression between non-stationary stochastic processes. The student does not have knowledge yet of stationarity and non-stationarity, and it will be premature to put much weight on the results of Tables 10, 11, and 12. It could serve as an illustration to the concepts to be explained in Chapter 3. However, the students will need to understand the regression output of Tables 7, 8, and 9.

Let POV_t and $UNEM_t$ denote the number of people in poverty and the number of unemployed persons respectively. We specify three similar models to those in Exercise 8. Observe that now we are modeling the relation between the levels of the series and not their growth.

$$
\begin{aligned}
\text{Model (d):} &\quad POV_t = \beta_0 + \beta_1 UNEM_{t-1} + u_t;\\
\text{Model (e):} &\quad POV_t = \beta_0 + \beta_1 UNEM_{t-1} + \beta_2 UNEM_{t-2} + \beta_3 UNEM_{t-3} + u_t;\\
\text{Model (f):} &\quad POV_t = \beta_0 + \beta_1 UNEM_{t-1} + \beta_2 UNEM_{t-2} + \beta_3 UNEM_{t-3} + \beta_4 POV_{t-1} + u_t.
\end{aligned}
$$

Tables 10, 11 and 12 report the regression results. The R-squared statistics are larger than those of the models in Exercise 8. Models (10d) and (10e) support the claim that unemployment is a one-year leading indicator of poverty. Observe that the Durbin-Watson statistic is very low and it is pointing out towards serial correlation in the residuals. Model (10f) corrects this serial correlation by introducing $POV(-1)$. Pay attention to the estimate of the coefficient attached to $POV(-1)$, which is very large, and to the increase in the adjusted R-squared, which jumped from 37% (Model 10e) to 93%. At face value, all these numbers are impressive but we should exercise some caution in interpreting any of the results of these tables because we need to consider the statistical properties of the time series POV and $UNEMP$. These are important issues that will be explained in the forthcoming chapters.

Dependent Variable: POV				
Method: Least Squares				
Sample (adjusted): 1960 2010				
Included observations: 51 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=3)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24661.7	3672.889	6.714523	0.0000
UNEM(-1)	1.234397	0.458921	2.689783	0.0097
R-squared	0.267241	Mean dependent var		32854.71
Adjusted R-squared	0.252287	S.D. dependent var		5612.197
S.E. of regression	4852.89	Akaike info criterion		19.85096
Sum squared resid	1.15E+09	Schwarz criterion		19.92672
Log likelihood	-504.2	F-statistic		17.87056
Durbin-Watson stat	0.124404	Prob(F-statistic)		0.000103

Table 10: Estimation Results. Model (10d)

Dependent Variable: POV				
Method: Least Squares				
Sample (adjusted): 1962 2010				
Included observations: 49 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=3)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	21934.14	3883.059	5.648674	0.0000
UNEM(-1)	1.846762	0.556685	3.317427	0.0018
UNEM(-2)	-1.37004	0.864561	-1.58467	0.1200
UNEM(-3)	1.108042	0.786683	1.408499	0.1659
R-squared	0.408648	Mean dependent var		32573.69
Adjusted R-squared	0.369225	S.D. dependent var		5545.536
S.E. of regression	4404.341	Akaike info criterion		19.69668
Sum squared resid	8.73E+08	Schwarz criterion		19.85111
Log likelihood	-478.569	F-statistic		10.3656
Durbin-Watson stat	0.238175	Prob(F-statistic)		0.000026

Table 11: Estimation Results. Model (10e)

Dependent Variable: POV				
Method: Least Squares				
Sample (adjusted): 1962 2010				
Included observations: 49 after adjustments				
Newey-West HAC Standard Errors & Covariance (lag truncation=3)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	788.0351	1105.691	0.712708	0.4798
UNEM(-1)	0.909875	0.18758	4.850593	0.0000
UNEM(-2)	-1.03858	0.2534	-4.09859	0.0002
UNEM(-3)	0.501	0.216324	2.315968	0.0253
POV(-1)	0.900449	0.047469	18.96932	0.0000
R-squared	0.935503	Mean dependent var		32573.69
Adjusted R-squared	0.929639	S.D. dependent var		5545.536
S.E. of regression	1470.986	Akaike info criterion		17.5217
Sum squared resid	95207132	Schwarz criterion		17.71475
Log likelihood	-424.282	F-statistic		159.5499
Durbin-Watson stat	1.550375	Prob(F-statistic)		0.000000

Table 12: Estimation Results. Model (10f)

