

BOSTON COLLEGE

Department of Economics

EC 771: Econometrics

Spring 2008

Prof. Baum, Mr. Alva

PROBLEM SET 1: SOLUTIONS

Point Distribution:

1) to 8): 10 points each

9) a), c), d): 3 points each

9) b), e), f): 2 points each

9) g): 5 points

1) Model: $y = \alpha + \beta x + \epsilon$

$$\text{a) } \mathbf{y} \equiv \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} \equiv \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \mathbf{b} \equiv \begin{pmatrix} a \\ b \end{pmatrix}$$

The normal equations for this model are given by

$$(\mathbf{X}'\mathbf{X})\mathbf{b} - \mathbf{X}'\mathbf{y} = \mathbf{0},$$

which implies that

$$\mathbf{X}'(\mathbf{X}\mathbf{b} - \mathbf{y}) = \mathbf{0} \implies \mathbf{X}'(-\mathbf{e}) = \mathbf{0}$$

Thus, $\sum_i x_i e_i = 0$. Also, since the first column consists of 1s, $\sum_i e_i = 0$.

b) Since the first normal equation is

$$na + \sum_i x_i b = \sum_i y_i$$

we immediately have that

$$a = \bar{y} - b\bar{x}$$

c) The second normal equation is

$$\sum_i x_i a + \sum_i x_i^2 b = \sum_i x_i y_i.$$

Substituting a from above, we have

$$\bar{y} \sum_i x_i - b\bar{x} \sum_i x_i + b \sum_i x_i^2 = \sum_i x_i y_i \implies b = \frac{\sum_i x_i y_i - \bar{y} \sum_i x_i}{\sum_i x_i^2 - \bar{x} \sum_i x_i}$$

Then,

$$\begin{aligned} b &= \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum_i x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} = \frac{\sum_i x_i y_i - \bar{x} \sum_i y_i - \bar{y} \sum_i x_i + n\bar{x}\bar{y}}{\sum_i x_i^2 - 2\bar{x} \sum_i x_i + n\bar{x}^2} \\ &= \frac{\sum_i (x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y})}{\sum_i (x_i^2 - 2\bar{x} x_i + \bar{x}^2)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \end{aligned}$$

d) $S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$ and so

$$\frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b}\mathbf{b}'} = 2\mathbf{X}'\mathbf{X} = 2 \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}$$

Now, $n > 0$ and $\sum_i x_i^2 > 0$, since the full rank condition requires that $x_i \neq x_j \forall i \neq j$. Then,

$$\begin{aligned} \left| \frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b}\mathbf{b}'} \right| &= 4n \sum_i x_i^2 - 4 \left(\sum_i x_i \right)^2 \\ &= 4n \left(\sum_i x_i^2 - n\bar{x}^2 \right) \\ &= 4n \left[\sum_i (x_i^2 - 2\bar{x}x_i + \bar{x}^2) + 2 \sum_i \bar{x}x_i - n\bar{x}^2 - n\bar{x}^2 \right] \\ &= 4n \left[\sum_i (x_i - \bar{x})^2 \right] \end{aligned}$$

2) Prove: $(\mathbf{y} - \mathbf{X}\mathbf{c})'(\mathbf{y} - \mathbf{X}\mathbf{c}) - (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = (\mathbf{c} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{c} - \mathbf{b})$

Proof:

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\mathbf{c})'(\mathbf{y} - \mathbf{X}\mathbf{c}) - (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{c} - \mathbf{c}'\mathbf{X}\mathbf{y} + \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} - \mathbf{y}'\mathbf{y} + \mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{b}'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{c}) + (\mathbf{b}' - \mathbf{c}')\mathbf{X}'\mathbf{X}\mathbf{b} + \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= -\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{c} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{b} + \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} \\ &= (\mathbf{c}' - \mathbf{b}')\mathbf{X}'\mathbf{X}\mathbf{c} - (\mathbf{c}' - \mathbf{b}')\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= (\mathbf{c} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{c} - \mathbf{b}) \end{aligned}$$

where we use $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$ in the third line.

3) Proof: Let $M_Z = (I - Z(Z'Z)^{-1}Z') \implies M_Z y = e_Z$. Similarly, define M_X , so that $M_X y = e_X$. Now, notice that $M_Z = I - Z(Z'Z)^{-1}Z' = I - XP(P'X'XP)^{-1}P'X'$. But, since P, P' , and $(X'X)$ are invertible, $(P'X'XP)^{-1} = P^{-1}(X'X)^{-1}(P')^{-1}$. Then, $M_Z = I - PP^{-1}(X'X)^{-1}(P')^{-1}P'X' = I - X(X'X)^{-1}X' = M_X$. Thus, $e_Z = M_Z y = M_X y = e_X$.

Thus, we can conclude that changing the units of measurement of the independent variables has no effect on the fit.

4) The matrix $M^0 = I - \iota(\iota'\iota)^{-1}\iota$ subtracts the means from the observations. Suppose only X has the means subtracted. Then, $b \equiv (X'M^0M^0X)^{-1}X'M^0y$. But, since M^0 is symmetric and idempotent, we have $b = (X'M^0X)^{-1}X'(M^0y)$, which is the same as subtracting the means from both X and y . Thus, coefficients of the regressors are not affected. Now, suppose only y is "de-meanned". Then, $\tilde{b} = (X'X)^{-1}X'M^0y$. But $M^0y = y - \bar{y}\iota$, so $\tilde{b} = (X'X)^{-1}X'(y - \bar{y}\iota) = b - (X'X)^{-1}X'\bar{y}\iota$. Thus, in general, $\tilde{b} \neq b$, and so we will not get the same coefficients if only y is transformed, unless the mean of the dependent variable in the sample is 0.

5) Define $X_{n,s} = \begin{pmatrix} X_n \\ x'_s \end{pmatrix}$. Then, $X'_{n,s}X_{n,s} = X'_nX_n + x_sx'_s$. In Appendix A, we have the updating formula to calculate the inverse of a matrix with a row appended. Using this formula to find $(X'_{n,s}X_{n,s})^{-1}$, we obtain that

$$(X'_{n,s}X_{n,s} + x_sx'_s)^{-1} = (X'_nX_n)^{-1} - \frac{(X'_nX_n)^{-1}x_sx'_s(X'_nX_n)^{-1}}{1 + x'_s(X'_nX_n)^{-1}x_s}$$

Thus,

$$\begin{aligned} b_{n,s} &= (X'_{n,s}X_{n,s})^{-1}X'_{n,s}y_{n,s} \\ &= \left((X'_nX_n)^{-1} - \frac{(X'_nX_n)^{-1}x_sx'_s(X'_nX_n)^{-1}}{1 + x'_s(X'_nX_n)^{-1}x_s} \right) (X'_ny_n + x_sy_s) \end{aligned}$$

since $X'_{n,s}y_{n,s} = \begin{pmatrix} X'_n & x_s \end{pmatrix} \begin{pmatrix} y_n \\ y_s \end{pmatrix} = X'_ny_n + x_sy_s$. Then,

$$\begin{aligned} b_{n,s} &= (X'_nX_n)^{-1}X'_ny_n + (X'_nX_n)^{-1}x_sy_s - \frac{(X'_nX_n)^{-1}x_sx'_s(X'_nX_n)^{-1}(X'_ny_n + x_sy_s)}{1 + x'_s(X'_nX_n)^{-1}x_s} \\ &= b_n + (X'_nX_n)^{-1}x_sy_s - \frac{(X'_nX_n)^{-1}x_sx'_sb_n + (X'_nX_n)^{-1}x_sx'_s(X'_nX_n)^{-1}x_sy_s}{1 + x'_s(X'_nX_n)^{-1}x_s} \\ &= b_n + \frac{(X'_nX_n)^{-1}x_s(y_s - x'_sb_n)}{1 + x'_s(X'_nX_n)^{-1}x_s} + \frac{(X'_nX_n)^{-1}x_sy_sx'_s(X'_nX_n)^{-1}x_s - (X'_nX_n)^{-1}x_sx'_s(X'_nX_n)^{-1}x_sy_s}{1 + x'_s(X'_nX_n)^{-1}x_s} \\ &= b_n + \frac{(X'_nX_n)^{-1}x_s(y_s - x'_sb_n)}{1 + x'_s(X'_nX_n)^{-1}x_s} \end{aligned}$$

where the second-to-last line uses the fact that y_s is a scalar.

Notice that $y_{s,predicted} = x'_s b_n$. If $y_s = y_{s,predicted}$, then $b_{n,s} = b_n$.

6) Let $x_i = (1 \ Y_i \ P_{d,i} \ P_{n,i} \ P_{s,i}) \implies X = (1 \ Y \ P_d \ P_n \ P_s)$. Then, $E_j = Xb_j + e_j$, where $j \in \{d, n, s\}$, so $b_j = (X'X)^{-1}X'E_j$. Now, $Y = E_d + E_n + E_s$. Therefore, $\sum_j b_j = \sum_j (X'X)^{-1}X'E_j = (X'X)^{-1}X' \sum_j E_j = (X'X)^{-1}X'Y$. But since Y is a column in X , we will have an exact fit if we ran a regression of Y on X i.e.

$$b_Y = (X'X)^{-1}X'Y = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \text{ where the coefficient for the regressor } Y \text{ is } 1 \text{ and all the others are } 0. \text{ Then, since}$$

$b_Y = \sum_j b_j$, the sum of the expenditure coefficients is 1 and all other coefficients sum to 0.

7) For the estimator to be unbiased, it must be that $c_1 + c_2 = 1$, since $E[\hat{\theta}] = c_1E[\hat{\theta}_1] + c_2E[\hat{\theta}_2] = (c_1 + c_2)\theta$, where θ is the true parameter value.

Thus, we need to minimize the variance of $c_1\hat{\theta}_1 + (1 - c_1)\hat{\theta}_2$. Now, $v \equiv \text{Var}[\hat{\theta}] = \text{Var}[c_1\hat{\theta}_1 + (1 - c_1)\hat{\theta}_2] = c_1^2v_1 + (1 - c_1)^2v_2 + c_1(1 - c_1)\text{cov}(\hat{\theta}_1, \hat{\theta}_2)$, where $v_i = \text{Var}[\hat{\theta}_i]$. Since, $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, the covariance term equivalent to 0. Thus, $v = c_1^2v_1 + (1 - c_1)^2v_2$. Then, $\frac{\partial v}{\partial c_1} = 2c_1v_1 - 2(1 - c_1)v_2 = 0$, which implies that $c_1 = \frac{v_2}{v_1 + v_2}$ and $c_2 = \frac{v_1}{v_1 + v_2}$.

8) Let $q = E[Q|P]$. Then, the expected profit $\Pi = Pq - Cq = P(a + bP) - C(a + bP)$, where C is the constant marginal cost. Profit is maximized when $\frac{\partial \Pi}{\partial P} = 0$ i.e. $a + 2bP - bC = 0$. Thus, $P^* = \frac{C}{2} - \frac{a}{2b}$. Given that $C = 10$, we have $P^* = 5 - \frac{a}{2b}$ and so the optimal quantity is given by $\frac{a}{2} + 5b$.

. regress Q P

Source	SS	df	MS	Number of obs =	15
Model	197.088735	1	197.088735	F(1, 13) =	12.52
Residual	204.644598	13	15.7418922	Prob > F =	0.0036
				R-squared =	0.4906
				Adj R-squared =	0.4514
Total	401.733333	14	28.6952381	Root MSE =	3.9676

Q	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]

P		-.8405832	.2375627	-3.54	0.004	-1.353806	-.3273602
_cons		20.76912	2.821568	7.36	0.000	14.67349	26.86475

```
. lincom _cons / 2 + 5 * P
```

```
( 1) 5 P + .5 _cons = 0
```

	Q	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)		6.181644	.5276531	11.72	0.000	5.041719 7.32157

Thus, the expected value of the profit-maximizing output is 6.18, with the 95confidence interval [5.042, 7.322].

9) a)

```
. tsset Year,yearly
      time variable: Year, 1953 to 2004

. // per capita gas consump, income
. gen gaspc = GasExp/(Gasp*(Pop/1e6))

. // logs
. gen lngaspc = log(gaspc)

. local allreg Income Gasp PNC PUC PPT PD PN PS

. // reg of part a
. reg gaspc 'allreg' Year
```

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

gaspc	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

One would expect the coefficient of the price of gasoline (Gasp) to be negatively correlated, since demand should be downward sloping, which it is. For income (Income) one would expect a positive coefficient because of the income effect; the regression produces this expected result. One might expect that the coefficient of the price of new cars (PNC) to be negative, since cars and gasoline are complements, but the regressions suggests otherwise (note however that the coefficient isn't significantly different from 0). It is possible that better fuel efficiency of newer cars more than offset the increased price of the newer cars. The coefficient of the price of public transportation (PPT) is sensible, since public transportation and gasoline are substitutes. Cars are durables, so (PD) poses the same puzzle as (PNC); the same explanation above for this puzzle might apply.

b) Notice that the 95% confidence interval for PUC is a subset of the 95% confidence interval of PNC. Thus, the null hypothesis that the true parameter value of the coefficients of PUC and PNC are the same cannot be rejected.

```
. test PNC = PUC
```

```
( 1) PNC - PUC = 0
```

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

c)

```
. est store a
```

```
. // elasticities: compute at t=2004
. mean 'allreg' Year if Year==2004
```

```
Mean estimation                Number of obs    =        1
```

	Mean	Std. Err.	[95% Conf. Interval]
Income	27113	0	. .
Gasp	123.901	0	. .
PNC	133.9	0	. .
PUC	133.3	0	. .
PPT	209.1	0	. .
PD	114.8	0	. .
PN	172.2	0	. .
PS	222.8	0	. .
Year	2004	0	. .

```
. mat x2004 = e(b)
```

```
. est restore a
(results a are active now)
```

```
. mfx compute, eyex at(x2004)
```

Elasticities after regress

```
y = Fitted values (predict)
= 6.1726971
```

variable	ey/ex	Std. Err.	z	P> z	[95% C.I.]	X
Income	.9476599	.2263	4.19	0.000	.504127 1.39119	27113
Gasp	-.2224796	.08093	-2.75	0.006	-.381102 -.063857	123.901
PNC	.0125245	.2786	0.04	0.964	-.533521 .55857	133.9
PUC	-.1268632	.10488	-1.21	0.226	-.332432 .078706	133.3
PPT	.2339837	.16441	1.42	0.155	-.08826 .556228	209.1
PD	.0228545	.22098	0.10	0.918	-.410256 .455965	114.8
PN	.3540265	.35281	1.00	0.316	-.337474 1.04553	172.2
PS	-1.011648	.29332	-3.45	0.001	-1.58654 -.436759	222.8

Year | 23.53872 4.63929 5.07 0.000 14.4459 32.6316 2004

The own price elasticity is -0.2225 (and significantly different from 0), the income elasticity is 0.9477 (and significantly different from 0) and the cross- price elasticity with PPT is 0.2340 (but not significantly different from 0).

d)

```
. foreach v of local allreg {
2.     gen ln'v' = log('v')
3.     local logreg "'logreg' ln'v'"
4. }
```

```
. reg lngaspc 'logreg'
```

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

lngaspc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnIncome	1.883045	.223034	8.44	0.000	1.433254	2.332836
lnGasp	.0735984	.0676117	1.09	0.282	-.0627536	.2099504
lnPNC	.3772717	.30747	1.23	0.226	-.2428007	.997344
lnPUC	-.334021	.0996132	-3.35	0.002	-.5349102	-.1331318
lnPPT	.1404593	.1683464	0.83	0.409	-.1990435	.4799621
lnPD	.6422717	.1817908	3.53	0.001	.2756555	1.008888
lnPN	-.492239	.3269502	-1.51	0.139	-1.151597	.167119
lnPS	-.6288652	.4383016	-1.43	0.159	-1.512785	.2550542
_cons	-15.79148	2.35185	-6.71	0.000	-20.53443	-11.04852

The own price elasticity is 0.07360, the income elasticity is 1.883, and the cross- price elasticity with PPT is 0.1405.

The own price elasticity is quite different from part c) and positive, implying an upward sloping demand curve. However, the parameter estimate is not statistically significantly different from 0. The income elasticity estimate is

almost twice as high as in part c). It is likely that the elimination of the time-trend from the log-log regression has resulted in the income growth rate "bleeding" into the estimate for the effect of income. The cross-price elasticity with the price of public transportation (PPT), while somewhat similar in value from the linear regression above, is also not significantly different from 0.

The log model tries to fit a constant elasticity function to the data, whereas the previous calculation of the elasticities was carried out at the mean point of the graph assuming a linear structural equation. If the elasticity varies with the dependent variable, then one should not expect that the two models produce the same elasticities.

It isn't clear which specification is appropriate.

e)

```
. corr 'logreg' Year
(obs=52)
```

	lnIncome	lnGasp	lnPNC	lnPUC	lnPPT	lnPD	lnPN	lnPS	Year
lnIncome	1.0000								
lnGasp	0.9448	1.0000							
lnPNC	0.9473	0.9667	1.0000						
lnPUC	0.9599	0.9674	0.9940	1.0000					
lnPPT	0.9790	0.9665	0.9891	0.9910	1.0000				
lnPD	0.9536	0.9776	0.9932	0.9945	0.9864	1.0000			
lnPN	0.9754	0.9839	0.9900	0.9902	0.9942	0.9923	1.0000		
lnPS	0.9809	0.9742	0.9902	0.9912	0.9985	0.9886	0.9979	1.0000	
Year	0.9923	0.9471	0.9631	0.9683	0.9878	0.9571	0.9809	0.9885	1.0000

It appears that there is a large degree of positive correlation among all the variables. One cannot however conclude that we have a multicollinearity problem, not without further investigation. That the log-log regression produces a positive own-price elasticity is particularly concerning. One can estimate the Variance Inflation Factor (VIF) by regressing the suspect variable (lnGasp) on the other regressors used in the original log-log regression.

```
. regress lnGasp lnIncome lnPNC lnPUC lnPPT lnPD lnPN lnPS Year
```

Source	SS	df	MS	Number of obs =	52
Model	23.2012964	8	2.90016205	F(8, 43) =	400.72
Residual	.311204941	43	.007237324	Prob > F =	0.0000
				R-squared =	0.9868

-----+-----					Adj R-squared = 0.9843	
Total		23.5125013	51	.461029438	Root MSE	= .08507
-----+-----						
lnGasp		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----						
lnIncome		-1.7125	.6569362	-2.61	0.013	-3.037339 - .3876623
lnPNC		-2.27155	.6694628	-3.39	0.001	-3.621651 - .92145
lnPUC		.1843208	.2389148	0.77	0.445	-.2974968 .6661385
lnPPT		-.368804	.3811388	-0.97	0.339	-1.137444 .3998356
lnPD		.5337328	.7292676	0.73	0.468	-.9369755 2.004441
lnPN		2.218312	.6676095	3.32	0.002	.8719495 3.564675
lnPS		.7517911	.9898355	0.76	0.452	-1.244402 2.747985
Year		.0066669	.0211907	0.31	0.755	-.0360683 .0494021
_cons		3.023167	37.03858	0.08	0.935	-71.67225 77.71858
-----+-----						

The R^2 for the regression is very close to 1, implying a very high VIF. However, there is no critical VIF that allows one to classify a regression as suffering from multicollinearity or not. The easy to use command `vif` does this for you for all the regressors.

`. vif`

Variable		VIF	1/VIF
-----+-----			
lnPS		4902.30	0.000204
lnPN		1566.09	0.000639
lnPPT		790.87	0.001264
lnPNC		645.15	0.001550
lnPD		305.77	0.003270
lnIncome		216.20	0.004625
lnPUC		192.91	0.005184
lnGasp		75.38	0.013266
-----+-----			
Mean VIF		1086.83	

The high VIFs strongly suggest that multicollinearity "problems" might exist, which is to say that the estimates are highly sensitive to particular data points. See Greene's discussion for more on this topic.

f) As figured out in problem 3 of this Problem Set, the units of measurement do not affect the fit of the re-

gression, but only the value of the relevant coefficients, which are scaled by the conversion factor between the two units.

$$b_X \equiv (X'X)^{-1}X'y$$

$$b_Z \equiv (Z'Z)^{-1}Z'y = (P'X'XP)^{-1}P'X'y = P^{-1}(X'X)^{-1}(P')^{-1}P'X'y = P^{-1}(X'X)^{-1}X'y = P^{-1}b_X$$

However, the log model will have the same coefficients for the regressors regardless of the unit of measurement; only the constant term will be altered by such a change of units, but not the fit. This follows from the simple algebraic fact that $\ln(sx) = \ln(s) + \ln(x)$, where s is some scaling factor (a scalar). Thus, the change in the units will change the constant term in the log-log regression.

g)

```
. gen break = tin(1974,2004)
```

```
. ttest lngaspc, by(break)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	21	1.334769	.04365	.2000295	1.243717	1.425822
1	31	1.730146	.012755	.0710169	1.704097	1.756195
combined	52	1.570475	.0331172	.2388115	1.503989	1.63696
diff		-.3953765	.0389887		-.4736877	-.3170653
diff = mean(0) - mean(1)					t = -10.1408	
Ho: diff = 0					degrees of freedom = 50	
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) = 0.0000		Pr(T > t) = 0.0000		Pr(T > t) = 1.0000		

The average value of log per capita gas consumption for period 1 is 1.3348 and for period 2 is 1.7301, with a statistically significant increase in the value from period 1 to period 2 of 0.3954.

```
. // regs for each subset
```

```
. gen iota = 1
```

```
. reg lngaspc 'logreg' Year if ~break
```



```

Year | .0007693 .0052182 0.15 0.884 -.0100827 .0116212
_cons | -4.904748 9.733856 -0.50 0.620 -25.14741 15.33791

```

```

. mat bpost = e(b)'
. mat vpost = e(V)
. qui predict ypost if e(sample)
. est store post
. mean ypost

```

```

Mean estimation              Number of obs   =      31

```

	Mean	Std. Err.	[95% Conf. Interval]	
ypost	1.730146	.0126148	1.704383	1.755909

```

. mean 'logreg' Year iota if e(sample)

```

```

Mean estimation              Number of obs   =      31

```

	Mean	Std. Err.	[95% Conf. Interval]	
lnIncome	9.918829	.031305	9.854896	9.982762
lnGasp	4.2413	.0585685	4.121687	4.360913
lnPNC	4.692742	.0483805	4.593936	4.791548
lnPUC	4.637867	.0770827	4.480443	4.795291
lnPPT	4.765984	.0959479	4.570032	4.961936
lnPD	4.616158	.0479135	4.518305	4.71401
lnPN	4.709391	.0602159	4.586413	4.832368
lnPS	4.783979	.0866703	4.606975	4.960984
Year	1989	1.632993	1985.665	1992.335
iota	1	0	.	.

```

. mat xpost = e(b)

.
. // first term B-0 decomp: take m as post
. mat t1 = xpost*(bpost-bpre)

. // cov mtx for first term
. mat vd = vpre+vpost

. // std error for first term
. mat t1var = xpost*vd*xpost'

. scalar t1se = sqrt(t1var[1,1])

. // second term
. mat t2 = (xpost-xpre)*bpre

. // total effect
. mat t3 = t1 + t2

.
. mat list t1, ti("Differential due to change in coeffs")

symmetric t1[1,1]: Differential due to change in coeffs
      y1
y1  -.50270686

. di "Std error " t1se " approx c.i." t1[1,1]-1.96*t1se " , " t1[1,1]+1.96*t1se
Std error .24585864 approx c.i.-.98458979 , -.02082394

. mat list t2, ti("Differential due to change in regressors")

symmetric t2[1,1]: Differential due to change in regressors
      y1
y1  .89808339

. mat list t3, ti("Total differential")

symmetric t3[1,1]: Total differential
      y1
y1  .39537652

```

You can verify the above results by using the decomposition command `oaxaca` written by Ben Jann.

```
. oaxaca post pre, weight(0)
(high estimates: post; low estimates: pre)
```

```
Mean prediction 1 = 1.730146
Mean prediction 2 = 1.334769
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
difference	.3953765	.0455803	8.67	0.000	.3060407	.4847124

Linear decomposition

Total	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
W=0						
explained	.8980834	.2564011	3.50	0.000	.3955465	1.40062
unexplained	-.5027069	.2503842	-2.01	0.045	-.9934509	-.0119628