

- ## Assignment - 2: BUAN 6312 Harikrishna Dev HXD220000

Source	SS	df	MS	Number of obs	=	660
				F(6, 653)	=	13.43
Model	17.0019785	6	2.83366308	Prob > F	=	0.0000
Residual	137.810143	653	.211041566	R-squared	=	0.1098
				Adj R-squared	=	0.1016

Total		154.812121	659	.234919759	Root MSE	=	.45939
ecobuy		Coefficient	Std. err.	t	P> t	[95% conf. interval]	
ecoprc		-.8026219	.1094037	-7.34	0.000	-1.017447	-.5877963
regprc		.7192675	.131639	5.46	0.000	.4607808	.9777543
faminc		.0005518	.0005295	1.04	0.298	-.000488	.0015916
hhsz		.0238227	.0125262	1.90	0.058	-.0007739	.0484193
educ		.0247849	.0083743	2.96	0.003	.008341	.0412287
age		-.0005008	.0012499	-0.40	0.689	-.0029551	.0019536
_cons		.4236865	.1649674	2.57	0.010	.099756	.747617

- Are the nonprice variables jointly significant in the LPM? (Use the usual F statistic, even though it is not valid when there is heteroskedasticity.) Which explanatory variable other than the price variables seems to have the most significant effect on the decision to buy ecolabeled apples? Does this make sense to you?

We can see that we conduct a hypothesis tests on the non price variables gives us a $p_value < 0.05$.

Therefore, we can reject the null hypothesis i.e. non-price variables are jointly significant. As $t(educ) = 2.96$ is the highest t statistic value among the non price variable, we can conclude that **education** makes most significant effect on purchase of eco-labeled apples. This makes sense that educated customers would prefer ecolabeled apples as they would be more well equipped in understanding the benefit of the consumption of them.

```
. test faminc hhsz educ age

( 1) faminc = 0
( 2) hhsz = 0
( 3) educ = 0
( 4) age = 0

F( 4, 653) = 4.43
Prob > F = 0.0015
```

- In the model from part (ii), replace $faminc$ with $\log(faminc)$. Given the R^2 , which model fits the data better? How many estimated probabilities are negative? How many are bigger than one? Should you be concerned? [Hint: Use command predict y to generate fitted values.]

```
. gen lfaminc = ln(faminc)
. reg ecobuy ecoprc regprc lfaminc hhsz educ age
```

Source		SS	df	MS	Number of obs	=	660
Model		17.278689	6	2.8797815	F(6, 653)	=	13.67
Residual		137.533432	653	.210617813	Prob > F	=	0.0000
Total		154.812121	659	.234919759	R-squared	=	0.1116
					Adj R-squared	=	0.1034
					Root MSE	=	.45893

ecobuy		Coefficient	Std. err.	t	P> t	[95% conf. interval]	
ecoprc		-.8006664	.1092981	-7.33	0.000	-1.015285	-.5860482
regprc		.721377	.1315196	5.48	0.000	.4631247	.9796294
lfaminc		.0445162	.0287239	1.55	0.122	-.0118861	.1009185
hhsz		.0227002	.012543	1.81	0.071	-.0019294	.0473297
educ		.023093	.0084508	2.73	0.006	.006499	.039687
age		-.0003865	.0012517	-0.31	0.758	-.0028444	.0020713
_cons		.3037519	.1789605	1.70	0.090	-.0476555	.6551593

We see that the **Adj-R sq** of the second model is greater in the first model. This indicates that the second model fits better. In the second model, there are two fitted probabilities are above 1 and in the range of 0.185 to 1.051. The two values aren't of concern as the source has 660 observations and the values are very close to 1. There are no negative probabilities.

2. Use the data in EZANDERS for this exercise. The data are on monthly unemployment claims in Anderson Township in Indiana, from January 1980 through November 1988. In 1984, an enterprise zone (EZ) was located in Anderson (as well as other cities in Indiana).
- Regress $\log(\text{uclms})$ on a monthly linear time trend and 11 monthly dummy variables. [Hint: Use jan as the base month for the monthly dummy variables.] What was the overall trend in unemployment claims over this period? (Interpret the coefficient on the time trend.) Is there evidence of seasonality in unemployment claims?

```
. use "C:\Users\hxd220000\Desktop\Data Sets- STATA\EZANDERS.DTA"

. regress luclms year feb mar apr may jun jul aug sep oct nov dec
```

Source	SS	df	MS	Number of obs	=	107
Model	27.0363482	12	2.25302901	F(12, 94)	=	14.36
Residual	14.7491008	94	.156905327	Prob > F	=	0.0000
				R-squared	=	0.6470
				Adj R-squared	=	0.6020
Total	41.7854489	106	.394202348	Root MSE	=	.39611

luclms	Coefficient	Std. err.	t	P> t	[95% conf. interval]
year	-.1665437	.0149503	-11.14	0.000	-.1962279 -.1368595
feb	-.0132261	.1867294	-0.07	0.944	-.3839816 .3575294
mar	-.0661643	.1867294	-0.35	0.724	-.4369198 .3045912
apr	-.3649279	.1867294	-1.95	0.054	-.7356834 .0058276
may	-.5147779	.1867294	-2.76	0.007	-.8855334 -.1440224
jun	-.5541234	.1867294	-2.97	0.004	-.9248789 -.1833679
jul	-.5191558	.1867294	-2.78	0.007	-.8899113 -.1484003
aug	-.3378477	.1867294	-1.81	0.074	-.7086032 .0329078
sep	-.7528584	.1867294	-4.03	0.000	-1.123614 -.3821029
oct	-.7867943	.1867294	-4.21	0.000	-1.15755 -.4160388
nov	-.6816665	.1867294	-3.65	0.000	-1.052422 -.310911
dec	-.3740492	.1926213	-1.94	0.055	-.7565034 .0084049
_cons	339.4264	29.66172	11.44	0.000	280.5323 398.3204

We see that coefficient of **YEAR** is **-0.1665**. This implies that the overall trend of unemployment claims decreases by **16.65%** per year. As the p-value < threshold value, we can conclude that the yearly trend is significant.

We can see that some of the monthly dummy variables are significant at a 5% level of significance, whereas some are not significant at the same threshold. This helps us understand that there is a presence of seasonal factors behind unemployment claims.

To confirm the joint significance, we perform the Wald test on the 11 monthly dummy variables.

$$H_0 : feb - dec = 0$$

$$H_1 : feb - dec \neq 0$$

```
. test feb mar apr may jun jul aug sep oct nov dec

( 1)  feb = 0
( 2)  mar = 0
( 3)  apr = 0
( 4)  may = 0
( 5)  jun = 0
( 6)  jul = 0
( 7)  aug = 0
( 8)  sep = 0
( 9)  oct = 0
(10)  nov = 0
(11)  dec = 0

      F( 11,    94) =    4.32
      Prob > F =    0.0000
```

As the $p\text{-value} < \text{threshold}$, we can reject the null hypothesis. Therefore, we can conclude that the monthly dummy variables are jointly significant.

- Add ez, a dummy variable equal to one in the months Anderson had an EZ, to the regression in part (i). Does having the enterprise zone seem to decrease unemployment claims? By how much?

```
. regress luc1ms year feb mar apr may jun jul aug sep oct nov dec ez
```

Source	SS	df	MS	Number of obs	=	107
Model	28.7422487	13	2.21094221	F(13, 93)	=	15.76
Residual	13.0432002	93	.140249465	Prob > F	=	0.0000
				R-squared	=	0.6879
				Adj R-squared	=	0.6442
Total	41.7854489	106	.394202348	Root MSE	=	.3745

	luc1ms	Coefficient	Std. err.	t	P> t	[95% conf. interval]
year		-.0811489	.0282722	-2.87	0.005	-.1372918 -.025006
feb		-.0132261	.1765405	-0.07	0.940	-.3638005 .3373484
mar		-.0661643	.1765405	-0.37	0.709	-.4167388 .2844101
apr		-.3649279	.1765405	-2.07	0.042	-.7155023 -.0143534
may		-.5147779	.1765405	-2.92	0.004	-.8653523 -.1642034
jun		-.5541234	.1765405	-3.14	0.002	-.9046978 -.203549
jul		-.5191558	.1765405	-2.94	0.004	-.8697303 -.1685814
aug		-.3378477	.1765405	-1.91	0.059	-.6884222 .0127267
sep		-.7528584	.1765405	-4.26	0.000	-1.103433 -.4022839
oct		-.7867943	.1765405	-4.46	0.000	-1.137369 -.4362198
nov		-.6816665	.1765405	-3.86	0.000	-1.032241 -.3310921
dec		-.3595756	.1821582	-1.97	0.051	-.7213057 .0021546
ez		-.5080266	.1456667	-3.49	0.001	-.7972917 -.2187614
_cons		170.2854	56.02201	3.04	0.003	59.03674 281.534

When ez is added to the regression, its coefficient is about $-.508$ (se $\approx .146$). EZ decreases the unemployment claims by:

$$100(1 - e^{-0.508}) = 39.82\%$$

3. Use the data in HSEINV for this exercise.

- Find the first order autocorrelation in $\log(\text{invpc})$ and $\log(\text{price})$ respectively. Which of the two series may have a unit root?

```
. use "C:\Users\hxd220000\Desktop\Data Sets- STATA\HSEINV.DTA"
. tsset year

Time variable: year, 1947 to 1988
Delta: 1 unit

. corr llnvpc llnvpc_1
(obs=41)
```

	llnvpc	llnvpc_1
llnvpc	1.0000	
llnvpc_1	0.6391	1.0000

The first order autocorrelation for $\log(\text{invpc})$ is 0.6391.

```
. corr lprice lprice_1
(obs=41)
```

	lprice	lprice_1
lprice	1.0000	
lprice_1	0.9492	1.0000

The first order autocorrelation for $\log(\text{price})$ is 0.9492.

As the correlation coefficient is high, we can assume they both have a unit root.

- Based on your findings in part (i), estimate the equation below and report the results in standard form. Interpret the coefficient β_1 and determine whether it is statistically significant.

$$\log(\text{invpc}_t) = \beta_0 + \beta_1 \times \Delta \log(\text{price}_t) + \beta_2 \times t + u_t$$

```
. reg linvpc gprice t
```

Source	SS	df	MS	Number of obs	=	41
Model	.575457228	2	.287728614	F(2, 38)	=	19.77
Residual	.553167094	38	.014557029	Prob > F	=	0.0000
Total	1.12862432	40	.028215608	R-squared	=	0.5099
				Adj R-squared	=	0.4841
				Root MSE	=	.12065

linvpc	Coefficient	Std. err.	t	P> t	[95% conf. interval]
gprice	3.878646	.9579971	4.05	0.000	1.939282 5.81801
t	.008037	.0015952	5.04	0.000	.0048077 .0112664
_cons	-.8532545	.040291	-21.18	0.000	-.9348193 -.7716897

We can see that the co-efficient of gprice is statistically significant. This implies that 1% growth of price results in 3.87% increase in per capita in the housing investment above it mean value.

- Now use $\Delta \log(\text{invpc}_t)$ as the dependent variable. Re-run the equation and report the results in standard form. How do your results of the coefficient β_1 change from part (ii)? Is the time trend still significant? Why or why not?

```
. reg ginvpc gprice t
```

Source	SS	df	MS	Number of obs	=	41
Model	.039000234	2	.019500117	F(2, 38)	=	0.95
Residual	.782237921	38	.020585208	Prob > F	=	0.3968
Total	.821238155	40	.020530954	R-squared	=	0.0475
				Adj R-squared	=	-0.0026
				Root MSE	=	.14348

ginvpc	Coefficient	Std. err.	t	P> t	[95% conf. interval]
gprice	1.566526	1.139214	1.38	0.177	-.7396933 3.872745
t	.000037	.001897	0.02	0.985	-.0038032 .0038772
_cons	.0059315	.0479125	0.12	0.902	-.0910623 .1029253

We see that the co-efficient is 1.567 and is not statistically significant. The time trend is not significant at 5% level of significance as the p value is 0.902.

- Recall that in the example of testing Efficient Markets Hypothesis, it may be that the expected value of the return at time t , given past returns, is a quadratic function of return_{t-1} .
- To check this possibility, use the data in NYSE to estimate

$$\text{return}_t = \beta_0 + \beta_1 \text{return}_{t-1} + \beta_2 \text{return}_{t-1}^2 + u_t$$

report the results in standard form.

```
. reg return return_1 return_1_2
```

Source	SS	df	MS	Number of obs	=	689
Model	19.2169743	2	9.60848717	F(2, 686)	=	2.16
Residual	3051.20782	686	4.4478248	Prob > F	=	0.1161
				R-squared	=	0.0063

				Adj R-squared	=	0.0034
Total		3070.42479	688	4.46282673	Root MSE	= 2.109
return		Coefficient	Std. err.	t	P> t	[95% conf. interval]
return_1		.0485723	.0387224	1.25	0.210	-.0274563 .1246009
return_1_2		-.009735	.0070296	-1.38	0.167	-.023537 .004067
_cons		.2255486	.087234	2.59	0.010	.0542708 .3968263

We can see both estimates are not statistically significant at 5%.

- State and test the null hypothesis that $E(\text{return}_t | \text{return}_{t-1})$ does not depend on return_{t-1} . [Hint: There are two restrictions to test here.] What do you conclude?

$$H_0 : \beta_1 = 0 \quad \beta_2 = 0$$

```
. test return_1 return_1_2

( 1) return_1 = 0
( 2) return_1_2 = 0

F( 2, 686) = 2.16
Prob > F = 0.1161
```

We need to satisfy the above null for our hypothesis to be satisfied. As the p value > 0.05, we cannot reject the null hypothesis.

- Drop return_{t-1}^2 from the model, but add the interaction term $\text{return}_{t-1} \times \text{return}_{t-2}$. Now test the efficient markets hypothesis. [Hint: stata can create lag (or lead) variables using subscripts conveniently. For example, you can use the command `gen return_2 = return[_n-2]` to create return_{t-2} fast.]

```
. gen return_2 = return[_n-2]
(3 missing values generated)

. gen return_2_1 = return_1*return_2
(3 missing values generated)

. reg return return_1 return_2_1
```

Source		SS	df	MS	Number of obs	=	688
Model		16.0639248	2	8.03196242	F(2, 685)	=	1.80
Residual		3053.36998	685	4.45747442	Prob > F	=	0.1658
Total		3069.4339	687	4.4678805	R-squared	=	0.0052
					Adj R-squared	=	0.0023
					Root MSE	=	2.1113

return		Coefficient	Std. err.	t	P> t	[95% conf. interval]
return_1		.0687116	.0392472	1.75	0.080	-.0083476 .1457709
return_2_1		.0113384	.0100134	1.13	0.258	-.0083222 .0309999
_cons		.1731605	.0809626	2.14	0.033	.0141959 .3321251


```
. test return_1 return_2_1

( 1) return_1 = 0
( 2) return_2_1 = 0

F( 2, 685) = 1.80
Prob > F = 0.1658
```

$$H_0 : \beta_1 = 0 \quad \beta_2 = 0$$

Since the p value of the Wald test > 0.05, we cannot reject the H0

- What do you conclude about predicting weekly stock returns based on past stock returns?

As both models look very weak when we look at the R sq and summary statistics, we cannot predict weekly stock returns from our models.

5. Use the data in KIELMC for this exercise.

- The variable *dist* is the distance from each home to the incinerator site, in feet. Consider the model

$$\log(\text{price}) = \beta_0 + \delta_0 y_{81} + \beta_1 \log(\text{dist}) + \delta_1 y_{81} \cdot \log(\text{dist}) + u.$$

If building the incinerator reduces the value of homes closer to the site, what is the sign of δ_1 ? What does it mean if $\beta_1 > 0$?

Assuming all the other variables remain constant, we can conclude that cost of home is positively correlated to the distance from the incinerator. Therefore,

$$\delta_1 > 0$$

Assuming $\beta_1 > 0$, We can assume the distance between the expensive houses and the incinerator is large.

- Estimate the model from part (i) and report the results in the usual form. Interpret the coefficient on $y_{81} \cdot \log(\text{dist})$. What do you conclude?

```
. use "C:\Users\hxd220000\Desktop\Data Sets- STATA\KIELMC.DTA"
. reg lprice y81 ldist y81ldist
```

Source	SS	df	MS	Number of obs	=	321
Model	24.3172548	3	8.10575159	F(3, 317)	=	69.22
Residual	37.1217306	317	.117103251	Prob > F	=	0.0000
				R-squared	=	0.3958
				Adj R-squared	=	0.3901
Total	61.4389853	320	.191996829	Root MSE	=	.3422

	lprice	Coefficient	Std. err.	t	P> t	[95% conf. interval]
y81		-.0113101	.8050622	-0.01	0.989	-1.59525 1.57263
ldist		.316689	.0515323	6.15	0.000	.2153005 .4180775
y81ldist		.0481862	.0817929	0.59	0.556	-.1127394 .2091117
_cons		8.058468	.5084358	15.85	0.000	7.058133 9.058803

From our analysis, we get the following equation:

$$\widehat{\text{price}} = 8.06 - 0.0113 \times y_{81} + 0.317 \text{ldist} + 0.0481 \times y_{81} \times \text{ldist}$$

$$n = 321, R^2 = 0.3958, \text{Adj} R^2 = 0.3901$$

We see that $\delta_1 = 0.0481862$, but the p-value > 0.05 . So, it is not statistically significant.

- Add *age*, *age*², *rooms*, *baths*, *log(intst)*, *log(land)*, and *log(area)* to the equation. Now, what do you conclude about the effect of the incinerator on housing values?

```
. reg lprice y81 ldist y81ldist age agesq rooms baths lintst lland larea
```

Source	SS	df	MS	Number of obs	=	321
Model	48.353762	10	4.8353762	F(10, 310)	=	114.55
Residual	13.0852234	310	.042210398	Prob > F	=	0.0000
				R-squared	=	0.7870
				Adj R-squared	=	0.7802
Total	61.4389853	320	.191996829	Root MSE	=	.20545

	lprice	Coefficient	Std. err.	t	P> t	[95% conf. interval]
y81		-.2254466	.4946914	-0.46	0.649	-1.198824 .7479309
ldist		.0009226	.0446168	0.02	0.984	-.0868674 .0887125
y81ldist		.0624668	.0502788	1.24	0.215	-.036464 .1613976

age		-.0080075	.0014173	-5.65	0.000	-.0107962	-.0052187
agesq		.0000357	8.71e-06	4.10	0.000	.0000186	.0000528
rooms		.0461389	.0173442	2.66	0.008	.0120117	.0802662
baths		.1010478	.0278224	3.63	0.000	.0463032	.1557924
lintst		-.0599757	.0317217	-1.89	0.060	-.1223929	.0024414
lland		.0953425	.0247252	3.86	0.000	.046692	.143993
larea		.3507429	.0519485	6.75	0.000	.2485266	.4529592
_cons		7.673854	.5015718	15.30	0.000	6.686938	8.660769

We can see that $\delta_1 = 0.0624668$ with a p-value = 0.215. As the summary of the regression output conducts a two-tailed test, we can assume for the one tailed test

$$H_0 : \delta_1 = 0$$

$$H_1 : \delta_1 > 0$$

$$p\text{-value}_{\text{one-tailed}} = \frac{p\text{-value}_{\text{two-tailed}}}{2} = \frac{0.215}{2} = 0.107$$

As the p-value > 0.05, we can conclude that the distance from the incinerator is not affecting the price of the houses.

- Why is the coefficient on $\log(\text{dist})$ positive and statistically significant in part (ii) but not in part (iii)? What does this say about the controls used in part (iii)?

We can see that in the first model, the coefficient of dist is statistically significant, where it is insignificant in the second model. This is due to the absense of these additional factor. To ensure they are jointly significant, we can perform the Wald's test.

```
. test age agesq rooms baths lintst lland larea

( 1)  age = 0
( 2)  agesq = 0
( 3)  rooms = 0
( 4)  baths = 0
( 5)  lintst = 0
( 6)  lland = 0
( 7)  larea = 0

F( 7, 310) = 81.35
Prob > F = 0.0000
```

As the p-value of the test is lesser than the threshold, we can conclude they are jointly significant.

- Use the data in PHILLIPS for this exercise. As we mentioned in Lecture 7, instead of the static Phillips curve model, we can estimate an expectations-augmented Phillips curve of the form

$$\Delta \ln f_t = \beta_0 + \beta_1 \text{unem}_t + e_t$$

where $\Delta \ln f_t = \ln f_t - \ln f_{t-1}$

- Estimate this equation by OLS and report the results in the usual form. In estimating this equation by OLS, we assumed that the supply shock, e_t , was uncorrelated with unem_t . If this is false, what can be said about the OLS estimator of β_1 ?

```
. reg cinf unem
```

Source		SS	df	MS	Number of obs	=	55
Model		32.6324798	1	32.6324798	F(1, 53)	=	6.13
Residual		282.055894	53	5.32180932	Prob > F	=	0.0165
Total		314.688374	54	5.82756247	R-squared	=	0.1037
					Adj R-squared	=	0.0868
					Root MSE	=	2.3069

dinf		Coefficient	Std. err.	t	P> t	[95% conf. interval]

unem	-.5176487	.209045	-2.48	0.017	-.9369398	-.0983576
_cons	2.828202	1.224871	2.31	0.025	.3714212	5.284982

We obtain the following equation by running a regression as follows:

$$\Delta inf_t = 2.83 - 0.518 \times unem_t + e_t$$

If e_t is correlated with $unem_t$, then the estimator for β_1 would be biased and inconsistent.

As the p-value of estimate of unem < 0.05, the estimate is significant at 5%.

- Suppose that e_t is unpredictable given all past information: $E(e_t | inf(t-1), unem(t-1), \dots) = 0$. Explain why this makes $unem_t - 1$ a good IV candidate for $unem_t$.

Assuming e_t is unpredictable, we can choose $unem_{t-1}$ as it correlated with the endogenous variable $unem_t$, but not to e_t . therefore, it can serve as IV for $unem_t$. As it satisfies the $E(e_t/unem_{t-1})=0$, we can conclude that $unem_{t-1}$ is not correlated to e_t . By using $unem_{t-1}$ as an IV for $unem_t$ in the regression, we can obtain consistent estimates of the causal effect of $unem_t$ on inf , even if $unem_t$ is endogenous.

- Does $unem_t - 1$ satisfy the instrument relevance assumption? [Hint: You need to run a regression to answer this question.]

```
. reg unem unem_1
```

Source	SS	df	MS	Number of obs	=	55
Model	68.9295284	1	68.9295284	F(1, 53)	=	69.12
Residual	52.8515619	53	.99719928	Prob > F	=	0.0000
Total	121.78109	54	2.25520538	R-squared	=	0.5660
				Adj R-squared	=	0.5578
				Root MSE	=	.9986

unem	Coefficient	Std. err.	t	P> t	[95% conf. interval]
unem_1	.7423824	.0892927	8.31	0.000	.5632839 .9214809
_cons	1.489685	.5202033	2.86	0.006	.446289 2.53308

As we can see that p-value of the unem_1 is below the threshold, we can conclude that the $unem_{t-1}$ is strongly correlated with $unem_t$ and satisfies the assumption.

- Estimate the expectations augmented Phillips curve by 2SLS using $unem_t - 1$ as an IV for $unem_t$. Report the results in the usual form and compare them with the OLS estimates from (i).

IV Model

```
. ivreg inf (unem = unem_1), first
```

First-stage regressions

Source	SS	df	MS	Number of obs	=	55
Model	68.9295284	1	68.9295284	F(1, 53)	=	69.12
Residual	52.8515619	53	.99719928	Prob > F	=	0.0000
Total	121.78109	54	2.25520538	R-squared	=	0.5660
				Adj R-squared	=	0.5578
				Root MSE	=	.9986

unem	Coefficient	Std. err.	t	P> t	[95% conf. interval]
unem_1	.7423824	.0892927	8.31	0.000	.5632839 .9214809
_cons	1.489685	.5202033	2.86	0.006	.446289 2.53308

Instrumental variables 2SLS regression

Source	SS	df	MS	Number of obs	=	55
--------	----	----	----	---------------	---	----

-----+-----					F(1, 53)	=	2.49
Model		41.0357451	1	41.0357451	Prob > F	=	0.1205
Residual		449.281356	53	8.47700671	R-squared	=	0.0837
-----+-----					Adj R-squared	=	0.0664
Total		490.317101	54	9.07994631	Root MSE	=	2.9115
-----+-----							
inf		Coefficient	Std. err.	t	P> t	[95% conf. interval]	
-----+-----							
unem		.5533613	.3506861	1.58	0.121	-.150026	1.256749
_cons		.6712233	2.025838	0.33	0.742	-3.392094	4.734541
-----+-----							
Instrumented: unem							
Instruments: unem_1							

The co-efficient of unem for the IV model is statistically isgnificant at 5%, whereas the OLS estimate of unem is significant at 5%.