Econ 4220 Assignment 1

Everaert and Pozzi20 develop a model to examine the predictability of consumption growth in 15 OECD countries. Their data is stored in the file oecd. The variables used are growth in real per capita private consumption (CSUMPTN), growth in real per capita government consumption (GOV), growth in per capita hours worked (HOURS), growth in per capita real disposable labor income (INC), and the real interest rate (R). Using only the data for Japan, answer the following questions:

a. Estimate the following model and report the results:

$$\text{CSUMPTN} = \beta_1 + \beta_2\text{HOURS} + \beta_3\text{GOV} + \beta_4 R + \beta_5\text{INC} + e$$

Are there any coefficient estimates that are not significantly different from zero at a 5% level?

The REG Procedure
Model: MODEL1
Dependent Variable: cons

Number of Observations Read	37
Number of Observations Used	37

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.01585	0.00396	37.22	<.0001
Error	32	0.00341	0.00010646		
Corrected Total	36	0.01926			

Root MSE	0.01032	R-Square	0.8231
Dependent Mean	0.02093	Adj R-Sq	0.8010
Coeff Var	49.29978		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.00403	0.00321	1.26	0.2185
hours	1	0.42089	0.18448	2.28	0.0293
gov	1	0.15911	0.14188	1.12	0.2704
r	1	0.23748	0.07252	3.27	0.0025
inc	1	0.61023	0.13846	4.41	0.0001

Running our first regression and looking at our t and $pr > |t|$ values shows that β_1 (the intercept) and β_3GOV (growth in real per capita government consumption) are both not statistically significant from zero at the 5% level.

This is because both β_1's t value (1.26) and β_3's t value (1.12) are less than our critical threshold of (1.96) and so we do not reject our null hypothesis that:

$$H_0 : \beta_1 = 0$$

and reject our alternative hypothesis:

$$H_a : \beta_1 \neq 0$$

Similarly it is the same for β_3GOV. We can also verify this by noting that $Pr > |t|$ for β_1 and β_3 are both less than 5% and so it is NOT inside the rejection region and we do not reject our null hypothesis.

b. The coefficient β2 could be positive or negative depending on whether hours worked and private consumption are complements or substitutes. Similarly, β3 could be positive or negative depending on whether government consumption and private consumption are complements or substitutes. What have you discovered? What does a test of the hypothesis H0:β2 = 0, β3 = 0 reveal?

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.01529	0.00765	65.59	<.0001
Error	34	0.00396	0.00011658		
Corrected Total	36	0.01926			

Root MSE	0.01080	R-Square	0.7942
Dependent Mean	0.02093	Adj R-Sq	0.7820
Coeff Var	51.58930		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.00344	0.00236	1.46	0.1541
hours	1	-1.9662E-17	0	-Infty	<.0001
gov	1	-3.3734E-17	0	-Infty	<.0001
r	1	0.32450	0.06441	5.04	<.0001
inc	1	0.76242	0.07185	10.61	<.0001
RESTRICT	-1	0.00138	0.00072352	1.90	0.0554*
RESTRICT	-1	-0.00014557	0.00094074	-0.15	0.8797*

* Probability computed using beta distribution.

Unrestricted model:

$$\text{CSUMPTN} = \beta_1 + \beta_2\text{HOURS} + \beta_3\text{GOV} + \beta_4 R + \beta_5\text{INC} + e$$

Restricted model:

$$\text{CSUMPTN} = \beta_1 + \beta_4 R + \beta_5\text{INC} + e$$

Our first RESTRICT tab denotes β_2HOURS. We can see that our coefficient of the parameter estimate is positive and so hours worked and private consumption must be **complements.**

> As growth in per capita hours worked **increases**, growth in real per capita private consumption also **increases.**
> It's t value = 1.90 therefore our parameter estimate **is not** statistically significant at the **95% CL** (2.5% in each tail) but **it is** statistically significant at the **90% CL** (5% in each tail, or $t = 1.65$)

Our second RESTRICT tab (in blue) denotes β_3GOV and its coefficient is negative. Therefore government consumption and private consumption are **substitutes.**

> When growth in real per capita government consumption **increases**, growth in real per capita private consumption **decreases.**
> It's t value = -0.15 therefore our parameter estimate **is not** statistically significant at the **95% CL** (2.5% in each tail) **or** the **90% CL** (5% in each tail, or $t = 1.65$)

In our first regression in question 1 we saw that our SSE_U (Sum of Squared Errors in the Unrestricted model) was 0.00341. In our second regression we now see $SSE_R = 0.00396$. Therefore adding β_2HOURS and β_3GOV to our model reduces the sum of squared errors and increases the explanatory power of our model.

test beta 2 = beta 3

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable cons				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	0.00027846	2.62	0.0887
Denominator	32	0.00010646		

Our F-test assesses whether the change in the sum of squared errors is sufficiently large enough for our parameters to be significant or not.

From page 263 paragraph 1 of Principles of Econometrics, 5th edition by R.C Hill:

> *"If adding the extra variables has little effect on the sum of squared errors, then those variables contribute little to explaining variation in the dependent variable, and there is support for a null hypothesis that drops them.*

> *On the other hand, if adding the variables leads to a big reduction in the sum of squared errors, those variables contribute significantly to explaining the variation in the dependent variable, and we have evidence against the null hypothesis." (p. 263)*

Therefore because our F statistic $F = 2.62 > 2.58$ the change in our sum of squared errors is statistically significant at the 99% CL (and any other CL below it like 95%, t = 1.95) so we reject our null hypothesis that:

$$H_0 : \beta_2 \text{HOURS} = 0, \quad \beta_3 \text{GOV} = 0$$

and do not reject our alternate hypothesis that:

$$H_A : \beta_2 \text{HOURS} \neq 0, \quad \beta_3 \text{GOV} \neq 0 \quad \text{or both are nonzero}$$

c. Re-estimate the equation with GOV omitted and, for the coefficients of the remaining variables, comment on any changes in the estimates and their significance.

Without gov

The REG Procedure
Model: MODEL1
Dependent Variable: cons

Number of Observations Read	37
Number of Observations Used	37

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.01572	0.00524	48.82	<.0001
Error	33	0.00354	0.00010729		
Corrected Total	36	0.01926			

Root MSE	0.01036	R-Square	0.8161
Dependent Mean	0.02093	Adj R-Sq	0.7994
Coeff Var	49.49186		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.00611	0.00263	2.32	0.0266
hours	1	0.30694	0.15458	1.99	0.0554
r	1	0.25377	0.07132	3.56	0.0012
inc	1	0.74470	0.06951	10.71	<.0001

When we re-run the regression without gov we see that:

1. β_1 increases by 0.00208 (51.61%) has a lower standard error, and gains statistical significance at a 95% CL ($t = 1.26 \to t = 2.32 > 1.96$).
2. β_2HOURS decreases by 0.11395 (-27.07%) but it is still statistically significant at a 95% CL ($t = 2.28 \to t = 1.99 > 1.96$)
3. $\beta_4 r$ becomes $\beta_3 r$, increases by 0.01629 (6.86%) and is slightly more statistically significant than before ($t = 3.27 \to t = 3.56$)
4. $\beta_5$inc becomes $\beta_4$inc, increases by 0.13447 (22.04%) and has a significant increase in statistical significance. ($t = 4.41 \to t = 10.71$)
   Therefore omitting gov causes large changes in β_1, β_2HOURS and $\beta_4$inc but NOT $\beta_3 r$

What would cause our original regression to not have statistical significance initially, but gain statistical significance with the removal of $\beta_3$gov?

What about Collinearity?

> *"Using collinear data can cause estimates to be statistically significant even when the variables should be important."* (Pg 176) Using SAS for Econometrics R.C Hill

We can compute the correlations between explanatory variables, and use variance inflation to search for collinearity.

High correlations between variables are an indicator that collinearity is causing problems for the regression.

Collinearity/vif

The REG Procedure
Model: MODEL1
Dependent Variable: cons

Number of Observations Read	37
Number of Observations Used	37

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.01585	0.00396	37.22	<.0001
Error	32	0.00341	0.00010646		
Corrected Total	36	0.01926			

Root MSE	0.01032	R-Square	0.8231
Dependent Mean	0.02093	Adj R-Sq	0.8010
Coeff Var	49.29978		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.00403	0.00321	1.26	0.2185	0
hours	1	0.42089	0.18448	2.28	0.0293	1.92394
gov	1	0.15911	0.14188	1.12	0.2704	4.26172
r	1	0.23748	0.07252	3.27	0.0025	1.39470
inc	1	0.61023	0.13846	4.41	0.0001	4.08596

Gov and inc seem to have a higher variance inflation than hours and r. But all of our VIF values are lower than the textbook's benchmark of 10. This is reassuring as the example handout on eclass also shows variance inflation in the 100's for the example regressors a, $a2$ and ap. So our variance inflation isn't relatively high and collinearity is probably mild. But what about correlations?

Collinearity/vif

The CORR Procedure

4 Variables:	hours gov r inc

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
hours	37	-0.00474	0.01293	-0.17536	-0.04503	0.01139
gov	37	0.03255	0.02502	1.20419	-0.00349	0.10311
r	37	0.01283	0.02801	0.47477	-0.08672	0.06181
inc	37	0.01748	0.02511	0.64685	-0.01666	0.08993

Pearson Correlation Coefficients, N = 37 Prob > |r| under H0: Rho=0				
	hours	gov	r	inc
hours	1.00000	-0.24813 0.1387	0.49131 0.0020	0.07740 0.6489
gov	-0.24813 0.1387	1.00000	-0.12612 0.4570	0.81135 <.0001
r	0.49131 0.0020	-0.12612 0.4570	1.00000	-0.06948 0.6828
inc	0.07740 0.6489	0.81135 <.0001	-0.06948 0.6828	1.00000

Our correlation between gov and inc is at 0.81135 but it is slightly below the threshold of 0.90 (where $R_K^2 = 0.90$ leads to $VIF = 10$). So again mild collinearity.

We can also use a condition index to find the severity. SAS has a good options for assessing the severity of collinearity. You can either use PROC REG or PROC MODEL

For this we'll add the block of code (also see code.txt)

```
* condition index;
proc reg data = oecd;
model cons = hours gov r inc/collin;
```

```
title "Condition Index";
run;
```

Collinearity Diagnostics								
			Proportion of Variation					
Number	Eigenvalue	Condition Index	Intercept	hours	gov	r	inc
1	2.76052	1.00000	0.02897	0.01029	0.01058	0.01161	0.01515
2	1.28348	1.46656	0.00086304	0.17201	0.00037252	0.23243	0.00015213
3	0.65898	2.04673	0.04340	0.17499	0.00312	0.20370	0.09752
4	0.24294	3.37091	0.63217	0.30419	0.00314	0.50445	0.04799
5	0.05409	7.14374	0.29459	0.33852	0.98278	0.04781	0.83918

See Appendix 6A.4 Collinearity diagnostics for the full proof:

Exact collinearity is defined as $\lambda_i = 0$ where λ is the eigenvalue of the system. Because if this holds $Xp_i = 0$ (where p_i is the eigenvector) and there is a linear combination of the columns of X that equals zero.

The square root of the ratio of the largest eigenvalue to the i'th is called a condition index or condition number

Note: There HAS to be a unique real eigenvalue of largest magnitude by the Perron-Frobenius theorem.

> *"If the largest condition index is less than 10, then collinearity is mild, if it is between 10 and 30 the collinearity is moderate, and over 30 it is severe."* (p.187) Using SAS for Econometric R.C Hill 5e

Our largest condition index is 7.14, therefore our collinearity is again mild.

d. Estimate the equation

$$GOV = \alpha 1 + \alpha 2 HOURS + \alpha 3 R + \alpha 4 INC + v$$

and use these estimates to reconcile the estimates in part (a) with those in part (c).

Gov regressed on other explanatory

The REG Procedure
Model: MODEL1
Dependent Variable: gov

Number of Observations Read	37
Number of Observations Used	37

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.01725	0.00575	35.88	<.0001
Error	33	0.00529	0.00016026		
Corrected Total	36	0.02254			

Root MSE	0.01266	R-Square	0.7654
Dependent Mean	0.03255	Adj R-Sq	0.7440
Coeff Var	38.89657		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.01306	0.00322	4.06	0.0003
hours	1	-0.71615	0.18891	-3.79	0.0006
r	1	0.10240	0.08717	1.17	0.2485
inc	1	0.84511	0.08495	9.95	<.0001

When we run our regression we note that only $\alpha_3 r$ is not statistically significant. This implies that GOV has collinearity with both $\alpha_2$hours and $\alpha_4$inc because when we regress them onto gov they are statistically significant. Simply put, because both $\alpha_2$hours and $\alpha_4$inc have an effect on GOV when you regress against it, so when we regress everything against CNSMPTN instead, all variables except for $\alpha_3 r$ will display collinearity. Remember earlier when I said "omitting gov causes large changes in β_1, β_2HOURS and $\beta_4$inc but NOT $\beta_3 r$" well, this was caused precisely because β_3GOV in our original model is in fact collinear with $\beta_2$hours and $\beta_5 inc$ as proven (albeit quite mild).

Reconciling our estimates in part a) with those in part c) results in us concluding that our regression in part a) is probably better. Why?

Our regression in part a)'s SSE is 0.00341 and its MSS/ESS is 0.01585

Our regression in part c)'s SSE is 0.00354 and its MSS/ESS is 0.01572

Therefore the explanatory power of the model (SSE) and how well the model fits the data (MSS/ESS) is better in a) than c). (because SSE is lower and MSS/ESS is higher in a than c).

Our adjusted R^2 is also higher in model a) than model c)

e. Re-estimate the models in parts (a) and (c) with the year 2007 omitted and use each of the estimated models to find point and 95% interval forecasts for consumption growth in 2007.

Model a)

Prediction and prediction interval model a

The REG Procedure
Model: MODEL1
Dependent Variable: cons

Number of Observations Read	37
Number of Observations Used	36
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.01556	0.00389	35.64	<.0001
Error	31	0.00338	0.00010916		
Corrected Total	35	0.01895			

Root MSE	0.01045	R-Square	0.8214
Dependent Mean	0.02141	Adj R-Sq	0.7984
Coeff Var	48.79841		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.00425	0.00329	1.29	0.2052
hours	1	0.44285	0.19287	2.30	0.0286
gov	1	0.16990	0.14559	1.17	0.2521
r	1	0.23153	0.07457	3.10	0.0040
inc	1	0.59543	0.14389	4.14	0.0002

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
22	0.01981	0.0192	0.002377	-0.002675	0.0410	0.000631
23	0.00753	0.004734	0.005390	-0.0192	0.0287	0.002793
24	0.02227	0.0255	0.003960	0.002759	0.0483	-0.003278
25	0.01326	0.0227	0.001961	0.001025	0.0444	-0.009447
26	0.00402	0.0196	0.003851	-0.003085	0.0423	-0.0156
27	0.00191	-0.001607	0.003439	-0.0240	0.0208	0.003514
28	-0.01765	-0.007589	0.003779	-0.0302	0.0151	-0.0101
29	0.00547	-0.007908	0.003925	-0.0307	0.0149	0.0134
30	0.00115	0.0127	0.004522	-0.0105	0.0360	-0.0116
31	0.00994	-0.004976	0.004364	-0.0281	0.0181	0.0149
32	0.00375	0.000312	0.003146	-0.0219	0.0226	0.003438
33	-0.00543	-0.006038	0.003280	-0.0284	0.0163	0.000607
34	0.00812	0.005174	0.002587	-0.0168	0.0271	0.002945
35	0.01016	0.0126	0.002164	-0.009183	0.0343	-0.002418
36	0.00978	0.0118	0.003957	-0.0110	0.0346	-0.002007
37	.	0.008353	0.003531	-0.0141	0.0308	.

Sum of Residuals	0
Sum of Squared Residuals	0.00338
Predicted Residual SS (PRESS)	0.00514

Therefore our point estimate for model a) in 2007 is 0.008353 and our 95% interval has lower and upper bounds of -0.0141 and 0.0308.

Model c)

Prediction and prediction interval model c

The REG Procedure
Model: MODEL1
Dependent Variable: cons

Number of Observations Read	37
Number of Observations Used	36
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.01541	0.00514	46.55	<.0001
Error	32	0.00353	0.00011039		
Corrected Total	35	0.01895			

Root MSE	0.01051	R-Square	0.8136
Dependent Mean	0.02141	Adj R-Sq	0.7961
Coeff Var	49.07352		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.00632	0.00278	2.27	0.0300
hours	1	0.31530	0.15981	1.97	0.0572
r	1	0.25093	0.07311	3.43	0.0017
inc	1	0.74136	0.07157	10.36	<.0001

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
35	0.01016	0.0130	0.002152	-0.008892	0.0348	-0.002794
36	0.00978	0.0138	0.003584	-0.008820	0.0364	-0.004013
37	.	0.006579	0.003205	-0.0158	0.0290	.

Sum of Residuals	0
Sum of Squared Residuals	0.00353
Predicted Residual SS (PRESS)	0.00489

Therefore our point estimate for model c) in 2007 is 0.006579 and our 95% interval has lower and upper bounds of -0.0158 and 0.0290.

f. Which of the two models, (a) or (c), produced the more accurate forecast for 2007?

Model A) has overall better explanatory power than Model C) due to a lower SSR/SSE $0.00338 < 0.00353$ and a higher adjusted R^2, but when we actually look at the predictions for the year 2007 itself we see that model c) provides a lower Std Error Mean Predict than model a) $(0.003205 < 0.003531)$. This tells us that for the year 2007 itself model c) has the more accurate forecast.