Exam and Solution

Please discuss each problem on a separate sheet of paper, not just on a separate page!

Problem 1: (20 points)
A health economist plans to evaluate whether screening patients on arrival or spending extra money on cleaning is more effective in reducing the incidence of infections by the MRSA bacterium in hospitals. She hypothesizes the following model:

$$MRSA_i = \beta_1 + \beta_2 S_i + \beta_3 C_i + u_i,$$

where, in hospital i, $MRSA$ is the number of infections per thousand patients, S is expenditure per patient on screening, and C is expenditure per patient on cleaning. u_i is a disturbance term that satisfies the usual regression model assumptions. In particular, u_i is drawn from a distribution with mean zero and constant variance σ^2. The researcher would like to fit the relationship using a sample of hospitals. Unfortunately, data for individual hospitals are not available. Instead she has to use regional data to fit

$$\overline{MRSA}_j = \beta_1 + \beta_2 \bar{S}_j + \beta_3 \bar{C}_j + \bar{u}_j,$$

where $\overline{MRSA}_j, \bar{S}_j, \bar{C}_j, \bar{u}_j$ are the averages of $MRSA, C, S, u$ for the hospitals in region j. There were different numbers of hospitals in the regions, there being n_j hospitals in region j.

1. Show that the variance of $\bar{u}_j$ is equal to $\frac{\sigma^2}{n_j}$ and that a regression using ordinary least squares (OLS) to fit the second equation will be subject to heteroscedasticity.

2. Assuming that the researcher knows the value of n_j for each region, explain how she could re-specify the regression model to make it homoscedastic. State the revised specification and demonstrate mathematically that it is homoscedastic.

3. Suppose that the researcher did not know the values of n_j. Explain in general terms (not mathematically) how, nevertheless, she could perform t tests relating to the regression coefficients, stating any limitations.

Solution:

1.

$$\begin{aligned} Var[\bar{u}_j] &= E[(\bar{u}_j)^2] = E\left[\left(\frac{1}{n_j}\sum_{i=1}^{n_j} u_i\right)^2\right] \\ &= \frac{1}{n_j^2}E[u_1^2 + u_2^2 + ... + u_{n_j}^2] \qquad \text{as all covariances are zero} \\ &= \frac{1}{n_j^2} n_j \cdot \sigma^2 \\ &= \frac{\sigma^2}{n_j} \end{aligned}$$

The OLS estimator remains unbiased but it is inefficient and the standard errors are invalid.

2. Multiply observation j by $\sqrt{n_j}$. The regression becomes

$$\sqrt{n_j}\overline{MRSA}_j = \sqrt{n_j}\beta_1 + \sqrt{n_j}\beta_2\bar{S}_j + \sqrt{n_j}\beta_3\bar{C}_j + \sqrt{n_j}\bar{u}_j$$

The variance of the disturbance term is now $Var[\sqrt{n_j}\bar{u}_j] = n_j Var[\bar{u}_j] = \sigma^2$ and is thus the same for all observations.

3. Use heteroscedasticity-consistent (robust/White) standard errors, a limitation being that they are valid only for large samples.

Problem 2: (20 points)
The following tables give results for an annual data set for 162 farms overs the years 1993 to 1998. The variables are:

MILK: milk output in litres per year

COWS: number of cows

LAND: land area, constant for each farm

FEED: feed input

Table 1 gives output of a pooled regression of log(MILK) on an intercept (C), log(COWS), log(LAND) and log(FEED). Table 2 gives output from a fixed effects regression where the estimated fixed effects are not reported, and Table 3 gives results from a random effects estimation:

Table 1: Pooled Estimation

Dep. Variable: LOG(MILK)

Method: Least Squares

Sample (adjusted): 1 972

Included observations: 972 after adjusting endpoints

variable	Coefficient	Std. Error	t-Statistics	Prob.
C	6,976457	0,040584	171,9009	0,0000
LOG(COWS)	0,600228	0,023564	25,49150	0,0000
LOG(LAND)	0,020668	0,014120	1,463763	0,1436
LOG(FEED)	0,455605	0,013712	33,22704	0,0000

R-squared	0,956152	Mean dep. var.	11,71364
Adjusted R-squared	0,956017	S.D. dep. var.	0,607083
S.E. of regression	0,127319	Sum squared resid	15,69139
F-statistics	7036,159	Durbin-Watson stat	0,573676

Variance covariance matrix of estimated coefficients:

variable	C	LOG(COWS)	LOG(LAND)	LOG(FEED)
C	0,001647	0,000407	-0,000121	-0,000413
LOG(COWS)	0,000407	0,000554	-0,000170	-0,000272
LOG(LAND)	-0,000121	-0,000170	0,000199	2,81E-05
LOG(FEED)	-0,000413	-0,000272	2,81E-05	0,000188

10^{-6} times the inverse of the variance covariance matrix of the estimated coefficients:

variable	C	LOG(COWS)	LOG(LAND)	LOG(FEED)
C	0,0762	0,2348	-0,1791	-0,4803
LOG(COWS)	0,2348	0,7371	0,5609	1,4984
LOG(LAND)	-0,1791	0,5609	0,4320	1,1403
LOG(FEED)	-0,4803	1,4984	1,1403	3,0579

Table 2: Fixed Effects Estimation

Dep. Variable: LOG(MILK)

Method: Pooled Least Squares

Sample: 1993 1998

Included Observations: 6

Number of cross-sections used: 162

Total panel (balanced) observations: 972

variable	Coefficient	Std. Error	t-Statistics	Prob.
LOG(COWS)	0,674705	0,032031	21,06433	0,0000
LOG(FEED)	0,396393	0,014944	26,52504	0,0000
fixed effects	...			

R-squared	0,986294	Mean dep. var.	11,71364
Adjusted R-squared	0,983529	S.D. dep. var.	0,607083
S.E. of regression	0,077914	Sum squared resid	4,905013
F-statistics	58142,43	Durbin-Watson stat	1,717733

Variance covariance matrix of estimated coefficients:

variable	LOG(COWS)	LOG(FEED)
LOG(COWS)	0,001026	-0,000375
LOG(FEED)	-0,000375	0,000223

Table 3: Random Effect Estimation

Dep. Variable: LOG(MILK)

Method: GLS (Variance Component)

Sample: 1993 1998

Included Observations: 6

Number of cross-sections used: 162

Total panel (balanced) observations: 972

variable	Coefficient	Std. Error	t-Statistics	Prob.
C	7,086916	0,057206	123,8852	0,0000
LOG(COWS)	0,657466	0,027075	24,28350	0,0000
LOG(LAND)	0,020818	0,023718	0,877749	0,3830
LOG(FEED)	0,410013	0,013630	30,08170	0,0000

GLS Transformed Regression

R-squared	0,983509	Mean dep. var.	11,71364
Adjusted R-squared	0,983458	S.D. dep. var.	0,607083
S.E. of regression	0,078080	Sum squared resid	5,901458
Durbin-Watson stat	1,445803		

Unweighted Statistics including Random Effects

R-squared	0,986028	Mean dep. var.	11,71364
Adjusted R-squared	0,985985	S.D. dep. var.	0,607083
S.E. of regression	0,071870	Sum squared resid	5,000071
Durbin-Watson stat	1,706445		

Variance covariance matrix of estimated coefficients:

variable	C	LOG(COWS)	LOG(LAND)	LOG(FEED)
C	0,003272	0,000286	-0,000377	-0,000733
LOG(COWS)	0,000286	0,000733	-0,000295	-0,000290
LOG(LAND)	-0,000377	-0,000295	0,000186	4,89E-05
LOG(FEED)	-0,000733	-0,000290	4,89E-05	0,000563

1. Test the hypothesis that the effect of log(COWS) equals the sum of the effects of log(LAND) and log(FEED) using the pooled estimation. Explain how you used the displayed output to form your conclusion.

   The hypothesis to test is whether (-cows+land + feed) (in logs) is equal to zero. We can use the F statistic to test this. There is one restriction and the degrees of freedom is 972 observations minus 4 estimated coefficients. The numerator is simply the restriction squared (plugging in the estimated coefficients) while the denominator is composed of the sum of the elements of the variance covariance matrix multiplied by their coefficients in the restriction:

$$\begin{aligned} F\left(1, 972-4\right) &= \frac{\left(-0,600228+0,020669+0,455605\right)^2}{10^{-6}\cdot(-1,1,1)\begin{pmatrix} 554 & 170 & 272 \\ 170 & 199 & 28 \\ 272 & 28 & 188 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}} \\ &= \frac{\left(-0,1240\right)^2}{10^{-6}\cdot 113} = \frac{10^{-6}\cdot 124^2}{10^{-6}\cdot 113} \doteq 136 \end{aligned}$$

   The 95% critical values for F(1,968) is somewhere between 252 and 254 (the critical values for F(1,60) and F(1,∞)) and hence we cannot reject the null hypothesis that the restriction is true.

2. Test the hypothesis that all slope coefficients of the pooled model are equal to zero.

   For this we can use the same test. The test statistics is already calculated and is equal to 7036 which is larger than the critical values (same as in part 1). Hence we conclude that at least one coefficient is significantly different from zero.

3. Which model would you prefer? Perform all the formal statistical tests to show why your model is the best one.

The pooled model (table 1) is more restrictive that the fixed effects model (table 2). In the pooled model there is a single constant that is common to all units. The fixed effect model replaces the constant with a set of individual dummies. To test the pooled restriction, we can use the F-test for the restriction that all these dummies are equal to each other. The F-test statistics is:

$$\begin{aligned} F\left(n-1, nT-n-k\right) &= \frac{\left(R^2_{fixed} - R^2_{pooled}\right)/\left(n-1\right)}{\left(1-R^2_{fixed}\right)/\left(nT-n-k\right)} \\ &= \frac{\left(0.986294 - 0.956152\right)/\left(162-1\right)}{\left(1-0.986294\right)/\left(972-162-2\right)} \\ &= \frac{0.030142/161}{0.013706/808} \doteq \frac{30}{14}\cdot\frac{808}{161} \doteq 2\cdot\frac{808}{161} > 1 \end{aligned}$$

The 95% critical value for $F\left(161, 808\right)$ is a bit bigger than the critical value of the $F\left(\infty,\infty\right)$ which is exactly one. Thus our test statistic is much larger than the critical value and we can reject the null of all the dummies are equal to each other with 95 percent confidence. Thus we should consider either the fixed or random effects model.

The fixed effect model (table 2) is more general. The random effects model (table 3) assumes that the individual effects are not correlated with the explanatory variables. To test this restriction, we could use the Hausmann test - under the null of no correlation, both the random (RE) and fixed effects (FE) estimators are consistent and the (RE) is more efficient; under the alternative RE is inconsistent, while FE is still consistent. The test statistics (which is asymptotically χ^2_K distributed) is:

$$H = \left(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}\right)' \mathbf{\Psi}^{-1} \left(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}\right),$$

where Ψ is the (asymptotic) variance covariance matrix of $\left(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}\right)$, with β being the $K \times 1$ coefficient vector excluding the constant terms. (Note that the Breusch-Pagan LM test is testing something different - whether the error components are actually present at all. Although this is a necessary condition for the use of the RE estimator, it is not a test

of the RE restriction relative to a FE model.) We have all the elements required by the test:

$$\left(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}\right) = \begin{pmatrix} 0,674705 - 0,657466 \\ 0,396393 - 0,410013 \end{pmatrix} = \begin{pmatrix} 0,017239 \\ -0,013620 \end{pmatrix},$$

and

$$\begin{aligned} \mathbf{\Psi} &= \mathbf{VC}_{\widehat{\beta}_{FE}} - \mathbf{VC}_{\widehat{\beta}_{RE}} \\ &= 10^{-6} \cdot \begin{pmatrix} 1026 & 375 \\ 375 & 223 \end{pmatrix} - 10^{-6} \cdot \begin{pmatrix} 733 & 290 \\ 290 & 49 \end{pmatrix} = 10^{-6} \cdot \begin{pmatrix} 293 & 85 \\ 85 & 174 \end{pmatrix}, \end{aligned}$$

and hence

$$\begin{aligned} \mathbf{\Psi}^{-1} &= 10^{6} \cdot \begin{pmatrix} 293 & 85 \\ 85 & 174 \end{pmatrix}^{-1} \\ &= 10^{6} \cdot 10^{-4} \cdot \begin{pmatrix} 40 & -19 \\ -19 & 67 \end{pmatrix}. \end{aligned}$$

This implies that the Hausman test statistics is

$$\begin{aligned} H &= 10^{-3} \begin{pmatrix} 17,239 \\ -13,620 \end{pmatrix}' \cdot 10^{2} \cdot \begin{pmatrix} 40 & -19 \\ -19 & 67 \end{pmatrix} 10^{-3} \begin{pmatrix} 17,239 \\ -13,620 \end{pmatrix} \\ &= 10^{-4} \cdot 15394 = 1,5394. \end{aligned}$$

The 90% critical value for chi-square with two degrees of freedom is 4,61 and hence we cannot reject the hypothesis that the random effects restriction is true. Thus the conclusion is to use the random effects model.

Problem 3: (10 points)
Humans are analyzed regarding their weight and height. The STATA output shows the results of regressing weight (WEIGHT85, measured in pounds) on height (HEIGHT, measured in inches), first with a linear specification and then with a logarithmic one (LNWEIGHT85=log(WEIGHT85), LNHEIGHT=log(HEIGHT)), including a dummy variable MALE (MALE= 1 if the observation unit is a man, otherwise 0) in both cases. Interpret all regression coefficients of both regressions.

1. reg WEIGHT85 HEIGHT MALE

Source	SS	df	MS
Model	288595.144	2	144297.572
Residual	342677.256	537	638.132692
Total	631272.4	539	1171.19184

Number of obs = 540
$F(2, 537) = 226.12$
$Prob > F = 0.0000$
R-squared = 0.4572
Adj R-squared = 0.4551
Root MSE = 25.261

WEIGHT85	Coef.	Std. Err.	t	$P > t$	[95% Conf.	Interval]
HEIGHT	4.155447	.3950937	10.52	0.000	3.379328	4.931565
MALE	15.52953	3.197231	4.86	0.000	9.24892	21.81015
cons	-133.8471	25.51672	-5.25	0.000	-183.9719	-83.72223

2. reg LNWEIGHT85 LNHEIGHT MALE

Source	SS	df	MS
Model	12.3281409	2	6.16407045
Residual	12.5598134	537	.023388852
Total	24.8879543	539	.046174312

Number of obs = 540
$F(2, 537) = 263.55$
$Prob > F = 0.0000$
R-squared = 0.4953
Adj R-squared = 0.4935
Root MSE = .15293

LNWEIGHT85	Coef.	Std. Err.	t	$P > t$	[95% Conf.	Interval]
LNHEIGHT	1.760394	.1611105	10.93	0.000	1.44391	2.076878
MALE	.1108935	.0193434	5.73	0.000	.0728955	.1488914
cons	-2.451119	.6711458	-3.65	0.000	-3.769512	-1.132726

Solution:

1. The linear regression indicates that weight tends to increase by 4.2 pounds for each inch of height and that controlling for height, males tend to weight 15.5 pounds more than females on average. The intercept has no sensible interpretation.

2. The logarithmic specification. Here you have to be a little bit careful since the height variable is logarithmic but the dummy variable is not. The coefficient of LGHEIGHT should be interpreted as an elasticity. A 1 percent increase in height tends to increase weight by 1.76 percent, controlling for sex.
   The relationship between weight and the MALE dummy variable is effectively semilogarithmic. Being male increases weight by a proportion 0.111, that is by 11.1 percent, controlling for height. The constant has no direct economic interpretation.

Problem 4: (10 points)
Consider the stochastic processes given below, where ε_t is normally distributed white noise. For each process determine whether it is covariance stationary, strictly covariance stationary, or integrated of order one [i.e. $I(1)$], or neither of these:

1. $X_t = 1 + t + \varepsilon_t$

   The mean of the process is $1 + t$ *and hence it is not covariance stationary. After first differencing we get* $\Delta X_t = \Delta\varepsilon_t$ *which is a covariance stationary process (it has a mean of zero; its variance is constant and equal* $2\sigma^2$*;the covariances* $cov(\Delta X_t, \Delta X_s)$ *is either* $-\sigma^2$ *for* $t - s = 1$*, or zero). Hence* X_t *is an* $I(1)$ *process.*

2. $(1 - 1,1L + 0,18L^2) X_t = \varepsilon_t$

   The roots of the characteristic polynomial $(1 - 1.1L + 0.18L^2)$ *are* $\lambda_{1,2} = \frac{1,1 \pm \sqrt{1.1^2 - 4 \cdot 0.18}}{2 \cdot 0.18} = \frac{1,1 \pm \sqrt{1.21 - 0.72}}{0.36} = \frac{1.1 \pm \sqrt{0.49}}{0.36} = \frac{1.1 \pm \sqrt{0.49}}{0.36} = \frac{1.1 \pm 0.7}{0.36}$*, i.e.* 5 *and* 1.1*. These are outside of the unit circle and hence the process is not covariance stationary.*

3. $X_t = \varepsilon_t \varepsilon_{t-1}$

   If we consider independent gaussian white noise we have that the mean is $E(\varepsilon_t \varepsilon_{t-1}) = 0$ *and is constant. The variance is also constant* $E(\varepsilon_t \varepsilon_{t-1} \varepsilon_t \varepsilon_{t-1}) = E(\varepsilon_t \varepsilon_t) \cdot E(\varepsilon_{t-1} \varepsilon_{t-1}) = \sigma^4$ *The covariances are always zero as* $E(\varepsilon_t \varepsilon_{t-1} \varepsilon_{t-1} \varepsilon_{t-2}) = E(\varepsilon_t) \cdot E(\varepsilon_{t-1} \varepsilon_{t-1}) \cdot E(\varepsilon_{t-2}) = 0$*. Hence the process is strictly covariance stationary. (Without the independence assumption we cannot guarantee stationarity because the variances could become time variant.)*