Exam

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

Problem 1: (20 points)
A sample of data consists of n observations on two variables, Y and X. The true model is

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i,\ (1.1)$$

where β_1 and β_2 are parameters and ε is a disturbance term that satisfies the usual regression model assumptions.

In view of the true model, $\bar{Y} = \beta_1 + \beta_2 \bar{X} + \bar{\varepsilon}$, (1.2) where $\bar{Y}$, $\bar{X}$, and $\bar{\varepsilon}$ are the sample means of Y, X, and ε.

Subtracting the second equation from the first, one obtains

$$(1.3)\ Y_i^* = \beta_2 X_i^* + \varepsilon_i^*,$$

where $Y_i^* = Y_i - \bar{Y}$, $X_i^* = X_i - \bar{X}$, and $\varepsilon_i^* = \varepsilon_i - \bar{\varepsilon}$.
Note that, by construction, the sample means of Y^*, X^*, and ε^* are all equal to zero.

One researcher fits

$$(1.4)\ \hat{Y} = b_1 + b_2 X$$

A second researcher fits

$$(1.5)\ \hat{Y^*} = b_1^* + b_2^* X^*$$

[Note: The second researcher included an intercept in the specification.]
In the following items, you must give mathematical proofs. Unsupported intuitive guesses will not earn credit.

1. (8 points) Comparing regressions (1.4) and (1.5), and making use of the expressions for the OLS estimators of the intercept and slope coefficient in a simple regression model, demonstrate that $b_2^* = b_2$ and that $b_1^* = 0$.

2. (4 points) Comparing regressions (1.4) and (1.5), demonstrate that $\hat{Y}_i^* = \hat{Y}_i - \bar{Y}$.

3. (3 points) Demonstrate that the residuals in (1.5) are identical to the residuals in (1.4).

4. (5 points) Explain why, theoretically, the specification (1.5) of the second researcher is incorrect and he should have fitted (1.6) $\hat{Y}^* = b_2 X^*$ not including a constant in his specification. If the second researcher had fitted (1.6) instead of (1.5), how would this have affected his estimator of β_2? Would dropping the unnecessary intercept lead to a gain in efficiency?

Solution:

1. Using e.g. the formula of Assignment 1, Problem 1

$$\begin{aligned} b_2^* &= \frac{\sum (X_i^* - \bar{X}^*)(Y_i^* - \bar{Y}^*)}{\sum (X_i^* - \bar{X}^*)^2} && \text{using now that } \bar{X}^* = 0 = \bar{Y}^* \\ &= \frac{\sum X_i^* Y_i^*}{\sum (X_i^*)^2} && \text{using the definition of } X^*, Y^* \\ &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = b_2 \\ b_1^* &= \bar{Y}^* - b_2^* \bar{X}^* = 0 \qquad \text{as } \bar{X}^* = 0 = \bar{Y}^* \end{aligned}$$

2. E.g.

$$\hat{Y}_i^* = b_1^* + b_2^* X_i^* = b_2^* X_i^* = b_2(X_i - \bar{X}) = b_2 X_i - (\bar{Y} - b_1) = \hat{Y}_i - \bar{Y}$$

3. $e_i^* = Y_i^* - \hat{Y}_i^* = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y}) = e_i$

4. The theoretical Equation (1.3) does not include an intercept. Using (1.6), the researcher should have estimated $b_2 = \frac{\sum X_i^* Y_i^*}{\sum (X_i^*)^2}$. However, (1) demonstrates that he has effectively done exactly this. Hence, the estimator will be the same. Consequently, there will be no gain in efficiency.

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 liters 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

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

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		

1. Test the hypothesis that log(LAND) has no significant effect on milk production. Describe which model you use, what is the null hypothesis, the test statistics and its distribution. Explain how you used the displayed output to form your conclusion.

   The variable LAND is time invariant and hence we can only use the pooled or random effects models. In both cases, we can use the displayed p-values from the t-statistics and conclude that we cannot reject the null hypothesis of no effect of log(LAND) on milk production (at say 10% significance level). In plain English – log(LAND) does not seem to have any significant effect on log(MILK).

2. Why is the variable log(LAND) not included in Table 2?

   The variable LAND is time invariant and hence the individual dummies capture its effect; i.e. it is perfectly collinear with the cross-sectional dummies and cannot be included in the fixed effect regression.

3. Compare the models in Table 1 and 2. What is the difference between them? Which is more restrictive? Explain how would you test for these restrictions and if possible perform the test.

   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.

4. Compare the models in Table 2 and 3. What is the difference between them? Which is more restrictive? Explain how would you test for these restrictions and if possible perform the test.

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:

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

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. The Ψ matrix is not provided in the tables and hence we cannot calculate the test. 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.

Problem 3: (10 points)
Suppose that the joint distribution of the two random variables x and y is

$$f(x,y) = \frac{\theta e^{-(\beta+\theta)y}(\beta y)^x}{x!},$$

where $\beta, \theta > 0, y \geq 0, x = 0, 1, 2, ...$ and $x!$ denotes $1 \cdot 2 \cdot ... \cdot x$.

1. (8 points) Find the maximum likelihood estimators for β and θ.
2. (2 points) Find the maximum likelihood estimator of $\theta/(\beta+\theta)$.

Solution:

1.

$$\begin{aligned}
f(x,y) &= \frac{\theta e^{-(\beta+\theta)y}(\beta y)^x}{x!} \\
L(\theta,\beta|y_1,...,y_n,x_1,...,x_n) &= \prod_{i=1}^{n} \frac{\theta e^{-(\beta+\theta)y_i}(\beta y_i)^{x_i}}{x_i!} \\
\ln L(\theta,\beta|...) &= n\ln\theta - (\beta+\theta)\sum y_i + \ln\beta\sum x_i + \sum x_i \ln y_i \\
&\quad -\sum \ln(x_i!) \\
\frac{\partial \ln L}{\partial \theta} &= \frac{n}{\theta} - \sum y_i = 0 \rightarrow \hat{\theta} = \frac{n}{\sum y_i} = \frac{1}{\bar{y}} \\
\frac{\partial \ln L}{\partial \beta} &= -\sum y_i + \frac{1}{\beta}\sum x_i = 0 \rightarrow \hat{\beta} = \frac{\sum x_i}{\sum y_i} = \frac{\bar{x}}{\bar{y}}
\end{aligned}$$

2.

$$\widehat{\frac{\theta}{\beta+\theta}} = \frac{1}{\bar{y}} \Big/ \left(\frac{\bar{x}}{\bar{y}} + \frac{1}{\bar{y}}\right) = \frac{1}{1+\bar{x}}$$

Problem 4: (10 points)
Consider the stochastic processes given below, where ε_t is a 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. $X_t = 0.7X_{t-1} + \varepsilon_t$

   Covariance stationary and strictly covariance stationary.

2. $X_t = X_{t-1} + \varepsilon_t + 8\varepsilon_{t-1}$

   Integrated of order one.

3. $X_t = 0.3X_{t-1} + \varepsilon_t$ for $t \leq T_0$ and $X_t = 0.8X_{t-1} + \varepsilon_t$ for $t > T_0$.

   None of these (the autocovariances are changing around the break so it is not covariance stationary).