Home-Assignment

Problem 1:
Fixed Effects.
Consider the fixed effects model $y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it}$. The LSDV approach yields the same estimate of β as by transforming the data and estimating $(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)'\beta + \epsilon_{it} - \bar{\epsilon}_i$. Write down the matrices $M_D y$ and $M_D X$ for the LSDV approach for $i = 1, 2$ and $t = 1, 2$ and show that you get the same transformed model as with the within transformation.

Problem 2:
Use the dataset grunfeld. This model is an investment equation

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \epsilon_{it}$$

where
I_{it}=real gross investment for firm i in year t
F_{it}=real value of the firm-shares outstanding
C_{it}=real value of the capital stock

1. Obtain the coefficients for the pooled dataset through least squares.
2. Transform y and x by their variation from the group mean and obtain the within estimator with the new variables. (The group variable is the company number)
3. Calculate the fixed effects model with the estimator provided in your econometric software and compare it to your estimation results in 2. (Depending on the software you use you maybe have to declare the dataset before as a panel). Calculate the individual effect a_i for the company with group number 1.
4. Compare the standard errors of the slope coefficients of the regression in point 2 with the s.e. of the regression in point 3. Why do you have to correct the standard errors of the within estimator of point 2?

5. Calculate the random effects model with the estimator provided in your econometric software.
6. Would you prefer the fixed effects or random effects model for this dataset? Use a Hausman Test. (Do not use the test provided by your software, show how to calculate the test statistic by hand!)

Problem 3:
Consider the model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i.$$

Performing least squares estimation on a dataset with 20 observations yields the following results:

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0.96587 \\ 0.69914 \\ 1.7769 \end{pmatrix}$$

$$\widehat{cov}(b) = \begin{pmatrix} 0.21812 & 0.019195 & -0.050301 \\ 0.019195 & 0.048526 & -0.031223 \\ -0.050301 & -0.031223 & 0.037120 \end{pmatrix}$$

$$\widehat{\sigma^2} = 2.5193$$
$$R^2 = 0.9466$$

1. Find the total variation, unexplained variation and explained variation for this model.
2. Find 95% interval estimates for β_2 and β_3.
3. Test the two null hypothesis that $\beta_2 = 0$ and that $\beta_3 = 0$.
4. Test the joint hypothesis that both slope coefficients are equal to zero.

Problem 4:
Omitted variables.
Suppose the correctly specified model is $y = X\beta + Z\alpha + \epsilon$. The ϵ are uncorrelated with X and Z. $E(\epsilon|X) = E(\epsilon|Z) = 0$.

1. Show that your estimated coefficients are biased when you regress y only on X.
2. Obtain the variance of this biased estimator b of point 1 and compare it to the variance of b_f when you estimate the true model which includes Z.
   (Hint: Use Frisch-Waugh-Lovell to obtain the estimator b_f of the full model.)

Problem 5:
Suppose that the classical regression model $y = X\beta + \epsilon$ applies but that the true value of the constant is zero. In order to answer the following questions assume just one independent variable.

1. Give the formulae for the two least squares slope estimators (the one with and the one without the constant).
2. Calculate their variances.
3. Compare the variance of the least squares slope estimator computed without a constant term with that of the estimator computed with an unnecessary constant term.

Problem 6:
Suppose you estimate the following model where you want to analyze the influence of the gender of a person on the college grade point average $colgpa$. $female$ and $male$ are dummy variables which take on the value 1 when a person is female/male.

$$colgpa = \beta_0 + \beta_1 female + \beta_2 male + \epsilon$$

1. What could be a problem when you estimate this regression model?
2. What are your suggestions to solve this problem?

3. Use the dataset gpa2 and estimate the model according to your suggestions in point 2. Report the results and give an interpretation of your estimated coefficients.

Problem 7:
The variance of an estimator in the regression model with two explanatory variables $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ can be represented as

$$Var(b_k) = \frac{\sigma^2}{(1 - r_{12}^2) \sum_i (x_{ik} - \bar{x_k})^2}$$

with r_{12} as the correlation coefficient between x_1 and x_2.

1. One determinant of the variance is the correlation between the regressors. What happens when the correlation is very high/very low/exactly one? Which Gauss-Markov assumption is not fulfilled when the correlation is exactly 1?
2. What are the determinants which influence the variance according to the formula above? How do they influence the variance?

Problem 8:
The file CEOSAL1 contains information on the average annual salary of CEO's in the year 1990 and various firm characteristics. You should determine the influence of the provided variables on the salary of the CEO's by a linear regression.

1. Plot the dependent variable against each independent variable to check if the linear relationship is appropriate. If the relationship is not clear try to transform the independent variable or the dependent variable or both and check again the plots. Explain why you use a certain specification.
2. Estimate the model according to your specification and check which coefficients are significant.
3. Check for multicollinearity and make a Q-Q Plot. What do you conclude?

4. Include the squared *roe* in your regression model. Is this necessary? Explain why or why not.
5. Reestimate your final model after your considerations of point 2 and 3 and check for heteroscedasticity. (Plot the residuals vs. the fitted values and perform an adequate test statistics)

Problem 9:
Heteroscedasticity.
Consider the model $y = X\beta + \epsilon$ with $E(\epsilon|X) = 0$ and an error variance of $E(\epsilon\epsilon'|X) = \sigma^2\Omega$ where $\Omega = (I + AA')$. A is an n x m matrix with $k < m < n$. Assume that σ^2 and A are known.

1. Obtain the variance of the OLS estimator for β and compare it to the standard least squares variance of $\sigma^2(X'X)^{-1}$.
2. Demonstrate the derivation of the variance of the GLS estimator b for this model. Use the result that $(I + AA')^{-1} = (I_N - A(I_M + A'A)^{-1}A')$.

(Hint: Start with the definition of variance $Var(b) = E((b - E(b)(b - E(b)')$ and $b_{OLS} = \beta + (X'X)^{-1}X'\epsilon$ and $b_{GLS} = \beta + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\epsilon$.)

Problem 10:
Use the dataset on housing prices (hprice2) to estimate the following model:

$$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + \epsilon$$

This model relates the median housing price (in USD) in a community to certain characteristics: *nox* is the amount of nitrogen oxide in the air in parts per million, *dist* is a weighted distance of the community from five employment centers in miles, *rooms* is the average number of rooms in houses in the community and *stratio* is the average student-teacher ratio of schools in this community.

1. Obtain the estimation coefficients through least squares. Provide an interpretation of each coefficient e.g. what happens to the house price if it has one more room.

2. Check for multicollinearity and for the normality of the residuals.
3. Plot the residuals (Q-Q Plot, residuals vs. fitted) and perform adequate test statistics, do we have a problem here? If yes, take adequate actions and reestimate the regression model.
4. Now include the additional variables *radial* (index for the accessory to highways), *crime* (crimes committed per capita) and *proptax* (property tax per 1000 USD). Compare this regression to the regression in 1 by obtaining an F-Test which compares the restricted vs. the unrestricted model. Which of these two models would you choose?