ARE 256a: Practicum 2
Due November 13th

Fall 2025

Questions 2-5, are drawn *The Practice of Econometrics: Classic and Contemporary* by Ernst Berndt.

1 Bookwork – Wooldridge

3.6, 3.11, 3.13, 3.18.

2 Examining Waugh's 1927 Asparagus Data

The purpose of this exercise is to involve you in an important part of the scientific method, namely, to attempt to replicate others' empirical findings. Many journals now require researchers to make their data and code available for replication purposes. In general, you should be able to replicate successfully previously reported results. In some cases, however, it will not be possible to achieve a complete replication or reconciliation of findings, and this will require you to dig further and examine the underlying data more closely. That is what we ask you to do in this exercise.

On Canvas, you will find a file called WAUGH, which contains 200 data points on four variables: (1) the *relative* price per bunch of asparagus, named PRICE, defined as $p_i = P_i/M_i$, where P_i is the actual price and M_i is the average market price for that day; (2) the number of inches of green color on the asparagus (in hundredths of inches), called GREEN; (3) the number of stalks of asparagus per bunch, denoted NOSTALKS; and (4) the variation in size (the interquartile coefficient) of the stalks, denoted DISPERSE.

a

Using these data, estimate the parameters of the multiple regression equation in which PRICE is regressed on a constant term, GREEN, NOSTALKS, and DISPERSE. Compare the parameter estimates that you obtain with those reported by Waugh,

$$p_i = \beta_0 + 0.13826 \times \text{GREEN}_i - 1.53394 \times \text{NOSTALKS}_i - 0.27554 \times \text{DISPERSE}_i + e_i.$$

Note that Waugh did not provide an estimate for the intercept or the standard errors. Which parameter estimates differ the most from those of Waugh?

b

As the results differ, further investigation appears to be warranted. Waugh [1929,Table 4,p.144] reports summary statistics of his underlying data. In particular, he reports the arithmetic means of the variables PRICE, GREEN. NOSTALKS, and DISPERSE to be 90.095. 5.8875, 19.555, and 14.875, respectively. Compute means of these variables. Are his statistics consistent with those based on the data in your file WAUGH? Do you have any hunches yet on the source of the inability to replicate Waugh's findings?

c

Waugh's Appendix also provides statistics on the product moments (variances and covariances) of the four variables, as follows:

VC Matrix	PRICE	GREEN	NOSTALKS	DISPERSE
PRICE	1063.64	3430.89	-100.92	-82.35
GREEN		24317.19	-17.01	-154.54
NOSTALKS			61.33	25.51
DISPERSE				83.07

Using your computer software and the data provided in the file WAUGH, compute the moment matrix and compare it to Waugh's, as reproduced above. Notice that the sample variances for the variables GREEN and DISPERSE are very similar to those reported by Waugh, they are not quite as close for NOSTALKS, and are very different for PRICE. Are all your covariances larger than those reported by Waugh, or does the relative size vary? Does there appear to be any pattern to the differences that might help to reconcile the findings?

d

Even though it does not appear to be possible to reconcile Waugh's data with his reported estimates of regression coefficients, are any of his principal qualitative findings concerning the effects of variations in GREEN. NOSTALKS, and DISPERSE affected? How different are your findings from his concerning the quantitative effects of one-unit changes in each of the regressors on the absolute price per bunch of asparagus? To do this calculation, you will need to know that the average market quotation PM_i was \$2.782. Comment also on the statistical significance of the parameter estimates.

e

Do you have any final thoughts on why results differ? (Hint: Compute least squares estimates using his estimated variances and covariances, reproduced above.

3 Exploring Relationships among R^2, Coefficients of Determination, and Correlation Coefficients

The purpose of this exercise is to gain an understanding of relationships among the various coefficients of determination, R^2, and correlation coefficients, as well as to comprehend better the implications of the extent of correlation among regressors.

a

Using the same Waugh data, compute the simple correlations between each of the variables. The correlation matrix that you obtain should be the following:

VC Matrix	PRICE	GREEN	NOSTALKS	DISPERSE
PRICE	1.00000	0.74834	-0.40656	-0.32464
GREEN		1.00000	-0.01403	-0.12605
NOSTALKS			1.00000	0.35003
DISPERSE				1.00000

Which variables are most highly correlated? Which variables are almost orthogonal?

b

Run three simple regressions, PRICE on GREEN, PRICE on NOSTALKS, and PRICE on DISPERSE, where each regression also includes a constant term. Take the R^2 from each of these three simple regressions, and compute its square root. Then compare its value with the appropriate correlation coefficient reported in the first row of the above table. Why are they equal (except for sign)? Now suppose you had messed up and had inadvertently run the "reverse" regressions. GREEN on PRICE, NOSTALKS on PRICE, and DISPERSE on PRICE. What R^2 measures would you have obtained? Why do they equal those from the "correct" regressions?

c

Notice the value of the R^2 measure from the simple regression of PRICE on GREEN, computed in part (b). What do you expect to happen to this value of R^2 if you now add the regressor NOSTALKS, that is, run a multiple regression equation with PRICE on a constant. GREEN, and NOSTALKS? Why? Given the correlation between the GREEN and NOSTALKS variables shown in the above table, do you expect the change in R^2 to be large or small? Why? Run this regression equation, and check to see whether your intuition is validated. Then comment on the change in the R^2 value from the simple regression of PRICE on GREEN or PRICE on DISPERSE when PRICE is regressed on both GREEN and DISPERSE; is this change consistent with the sample correlation between GREEN and DISPERSE? Similarly, what is the change in the R^2 value from the simple regression of PRICE on NOSTALKS or PRICE on DISPERSE when PRICE is regressed on both NOSTALKS and DISPERSE? Is this change consistent with the sample correlation coefficient between NOSTALKS and DISPERSE? Why?

d

In all three cases considered in part (c), the R^2 from the multiple regression (with two regressors in addition to the constant) is less than the sum of the R^2's from the corresponding two simple regressions. Is the R^2 from the multiple regression equation with all three regressors (GREEN, NOSTALKS, and DISPERSE) greater than or less than the sum of the R^2 from the three simple regressions? Note: It might be tempting to conclude from this that the R^2 from a multiple regression with a constant term and K regressors is always less than or equal to the sum of the R^2 values from the K simple regressions. However, this is not always the case, as has been shown in an interesting theoretical counter example by Harold Watts [1965].

e

Use,

$$d^2_{y,x_j} = b_j \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})}$$

to compute separate coefficients of determination, based on the regression coefficient estimates reported by Waugh and reproduced in Question 2 above, see whether you can replicate Waugh's reported coefficient of determination values for the GREEN, NOSTALKS, and DISPERSE variables as 0.40837, 0.14554, and 0.02133, respectively. You should be able to replicate Waugh for NOSTALKS and DISPERSE but not for GREEN. Waugh [1929, p.113] states: The sum of the coefficients of determination is .57524, indicating that 57.524 per cent of the squared variability in the percentage prices is accounted for by the three factors studied. Is this correct? What should Waugh have stated instead? Why?

f

As in part (a) of Exercise 1, estimate parameters in the multiple regression equation of PRICE on a constant, GREEN, NOSTALKS, and DISPERSE. Note the value of the R2 from this regression, and then compute and retrieve the fitted or predicted values. Now run a simple regression equation in which the dependent variable is PRICE and the regressors include a constant and the fitted value from the previous regression equation. Compare the R^2 from this regression to that from the first regression. Why does this result occur? Why is the value of the estimated intercept term zero and the estimated slope coefficient unity in this regression?

4 Assessing the Stability of the Hedonic Price Equation for First- and Second-Generation Computers

In this exercise we assess the stability of the hedonic price equation for computers over the period 1955–1965. One implicit hypothesis underlying the hedonic method is that goods such as computers can be viewed as the aggregate of a number of characteristics. Since firms supply various computer models embodying alternative combinations of characteristics and consumers demand them, the relationship between price and characteristics reflects the outcome of a market process. When dramatic technological changes occur, factor prices vary, or if consumer preferences change, the

relationship between the overall price of the bundle and the individual characteristics might also change. We'll use data in the file CHOW.xlsx on Canvas;

The variables in the file are:

- Volume: Number of new installations of that computer in a year
- Rent: The monthly rental of copmuts
- Words: The number of words in main memory (in thousands)
- Binary: The number of binary digits per word.
- Digits: The number of equivalent binary digits.
- Mult: Time to obtain and complete multiplication instructions.
- Add: Time to obtain and complete addition instructions.
- Access: Average time to access information from memory.
- Year: Year in which the model was introduced.
- IBMdum: A dummy equal to one if the computer was made by IBM.

From this construct:

- The natural logarithms of RENT, MULT, ACCESS and ADD, use the prefix LN.
- MEM, the product of Words×Binary×Digits. Take the log and rename as above.

a

Chow estimated a model of the form:

$$LNRENT = \beta_0 + \beta_1 LNMEM + \beta_2 LNMULT + \beta_3 LNACCESS + u$$

Conventional wisdom in the computer industry dates the first generation of computers as occurring from 1954 to about 1959 and the second generation as taking place between 1960 and about 1965. Chow [1967. p. 1123] reports that he tested the null hypothesis that the three "slope" coefficients were equal over the 1960-1965 time period and could not reject the null hypothesis; his F-test statistic was 0.74, much less than the critical value at any reasonable level of significance. Construct the appropriate variables as in part (a) of Exercise 3, and then estimate parameters in two models, one in which the slope coefficients are constrained to be the same in all years 1960-1965 (a pooled regression) and the other in which these coefficients are allowed to differ (separate, year-by-year regressions).

Based on the sums of squared residuals from these individual regressions, test the null hypothesis that the slope coefficients are equal over the 1960-1965 time period. Be particularly careful in calculating the appropriate degrees of freedom for the F-test.

b

Form appropriate dummy variables for each of the years from 1955 to 1959, and then repeat part (a) and test the null hypothesis that the slope coefficients are equal over the 1954-1959 era, first by running a pooled regression over the 1954-1959 data and then by doing year-by-year regressions, 1954 to 1959.

c

In essence, parts (a) and (b) tested for slope parameter stability within the first and the second generations of computers, respectively. To test whether the hedonic relationship changed between the first and second generations, it will be useful to run one additional regression covering the entire 1954-1965 time period, namely, a specification in which LNRENT is regressed on a constant, year-specific dummy variables for 1955 through 1965, LNMEM, LNMULT, and LNACCESS. Having run this regression, and initially assuming equality of the slope parameters within the first (1954-1959) and the second (1960- 1965) generations, test the null hypothesis that the slope coefficients of the first generation equal those of the second generation. Does this result surprise you? Why or why not? Next. relax the assumption of slope parameter equality within each generation, and test the null hypothesis that slope parameters are equal over the entire 1954-1965 time span against the alternative hypothesis that these slope coefficients varied from year to year. Note that calculation of the appropriate F-statistic requires comparing the sums of squared residuals from the 12 separate year-by-year regressions with that from the pooled 1954-1965 regression and then adjusting by the appropriate degrees of freedom. Interpret your results. Are the two test results of part (c) mutually consistent? Why or why not?

5 Using Time-Varying Hedonic Price Equations to Construct Chained Price Indexes for Computers

The procedures for constructing quality-adjusted price indexes for computers based on estimated hedonic price equations discussed in this chapter assumed that the slope coefficients were constant over time. In this exercise we relax the assumption of constant parameters over the entire data sample and instead employ adjacent year regression procedures to construct chained price indexes. The data used in this exercise are the same as in the previous question.

a

Consider the following regression equation, based on data from two adjacent years, for example, 1954 and 1955:

$$LNRENT_i = \beta_0 + \beta_t DUM_{it} + \beta_1 LNMFM_i + \beta_2 LNMULT_i + \beta_3 LNACCFSS_i$$

where DUM_{it} is a dummy variable taking on the value of 1 if model 1 was introduced in the current year (say, 1955) and o if it was introduced in the adjacent previous year (1954). The estimate ofβ_t indicates the change in the natural logarithm of the price from 1954 to 1955, holding

quality fixed. Such a regression equation could be specified for each pair of adjacent years, such as 1954-1955, 1955-1956, 1956-1957,, 1964-1965. An attractive feature of the adjacent year regression approach is that the slope coefficients are allowed to vary over time. Using the data in the file CHOW, construct the appropriate variables, estimate the 11 adjacent year regression equations by ordinary least squares. and then retrieve the 11 estimates of β_t. denoted asβ_{1955}, β_{1956}, β_{1957}, ,β_{1965},. Next, using data covering the entire 1954-1965 time period, estimate the more traditional hedonic regression equation in which LNRENT is regressed on a constant, 11 dummy variables D_{1955} to D_{1965}, LNMEM, LNMULT, and LNACCESS. Compare year-to-year changes in the estimated coefficients of these 11dummy variables with the levels of the 11β_t estimates. Why is it appropriate to compare year-to-year changes in the estimated dummy variable coefficients with levels of the estimated β_t? Comment on and interpret any differences that appear to be substantial.

b

Calculate a traditional hedonic price index for computers over the 1954-1965 time period, normalized to unity in 1954, by simply exponentiating values of the estimated coefficients on the 11 dummy variables, D_{1955} to D_{1965}. Then construct a chained price index, using the following sequential procedure: For 1955, exponentiate β_{1955}; for 1956, exponentiate the sum $\beta_{1955} + \beta_{1956}$; for 1957, exponentiate the sum $\beta_{1955} + \beta_{1956} + \beta_{1957}$. Continue this for each year, until for 1965 the quality-adjusted price index is computed as the antilogarithm of the sum $\beta_{1955} + \beta_{1956} + \beta_{1957} + ... + \beta_{1965}$. Why is such an index called a chained price index? Empirically compare this chained price index with the traditional hedonic price index. Do they differ in any substantial or systematic manner? Which index do you prefer, and why?