Practicum Two

Josh Virene

2023-10-12

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 pi = Pi/mi, where Pi is the actual price and Mi 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, pi = B0 + 0.13826 x greeni - 1.53394 x NOSTALKSi - 0.27554 x DISPERSEi + ei 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?

Table 1: WAUGH Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40.7612635	5.3278365	7.650622	0.000000
GREEN	0.1375982	0.0070994	19.381808	0.000000
NOSTALKS	-1.3572564	0.1508215	-8.999089	0.000000
DISPERSE	-0.3452828	0.1296563	-2.663063	0.008387

Waugh's model:

$$pi = B0 + 0.13826xGREEN_i - 1.53394xNOSTALKS_i - 0.27554xDISPERSE_i + e_i$$

Practicum model:

$$p_i = \underset{(5.327837)}{40.761264} + \underset{(0.007099)}{0.137598}xGREEN_i - \underset{(0.150822)}{1.357256}xNOSTALKS_i - 0.345283x\underset{(0.129656)}{DISPERSE_i} + e_i$$

Comparing these two models, the parameter estimate that differs the most compared to what Waugh estimated is the parameter on NOSTALKS. My model calculated this as B2 = -1.357256, while Waugh's model calculated this as B2 = -1.53394. The parameter that differed the second-most is B3; my model calculated this as B3 = -0.345283, while Waugh's model calculated this as B3 = - 0.27554. The coefficient estimates for B1 are nearly identical (my model B1 = 0.137598, Waugh's model B1 = 0.13826)

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?

The mean for PRICE = 90.095, the mean for GREEN = 588.75, the mean for NOSTALKS = 19.555, and the mean for DISPERSE = 14.875. Because the means that I calculated match the means that Waugh calculated, this is not the source of inability to replicate Waugh's findings. Looking back to the correlation table, the correlations do not match those that Waugh reported, so this is likely the source of inability to replicate his findings.

c. Waugh's Appendix also provides statistics on the product moments (variances and covariances) of the four variables, as follows: (table not included)

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?

	GREEN	NOSTALKS	DISPERSE	PRICE
GREEN	24439.38	-17.09171	-180.05653	3448.18467
NOSTALKS	NA	60.73063	24.92400	-93.38465
DISPERSE	NA	NA	83.48681	-87.43028
PRICE	NA	NA	NA	868.73967

Comparing the VC Matrix provided in the assignment (Waugh's VC matrix) to the one that I computed, the GREEN-DISPERSE sample variances are not as close as the problem would have us believe (-180.05653 versus -154.54 in Waugh's) the GREEN-NOSTALKS sample variances are very close (-17.09171 in my matrix, versus -17.01 in Waugh's), and last the GREEN-PRICE sample variances are much further apart (24439.38 in my matrix versus 24317.19 in Waugh's). *Based on the results provided, there is no discernible pattern, so the analysis up to this point does not help explain why the models are different.*

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 PMi was $2.782. Comment also on the statistical significance of the parameter estimates.

Waugh's principal qualitative findings concerning the effects of variations in GREEN, NOSTALKS, and DISPERSE are not affected despite the inability to reconcile the findings in Waugh's data with the regression that was run in this Practicum. To support this, we can revisit the regression output in the Practicum model that I ran, and compare this to Waugh's model. Despite having different values, the interpretation of the marginal effects of each variable on the market price of Asparagus are the same, so qualitatively, there is no need to reconcile this difference. Quantitatively, it may be important to reconcile this difference- I believe that this has to do with differences in the moment table computed in this Practicum; it has larger sample variances across the variables compared to Waugh's model, which has implications for the regression output.

Waugh's model:

$$pi = B0 + 0.13826xGREEN_i - 1.53394xNOSTALKS_i - 0.27554xDISPERSE_i + e_i$$

Practicum model:

$$p_i = \underset{(5.327837)}{40.761264} + \underset{(0.007099)}{0.137598xGREEN_i} - \underset{(0.150822)}{1.357256}xNOSTALKS_i - \underset{(0.129656)}{0.345283xDISPERSE_i} + e_i$$

Scaled Practicum model:

$$p_i = \underset{(5.327837)}{113.3978} + \underset{(0.007099)}{0.3827976xGREEN_i} - \underset{(0.150822)}{3.7758861}xNOSTALKS_i - \underset{(0.129656)}{0.9605773xDISPERSE_i} + e_i$$

Unscaled confidence intervals:

$$40.761264 \pm 2(5.327837) = (30.10559, 51.416938)$$

$$0.137598 \pm 2(0.007099) = (0.1234, 0.151796)$$

$$-1.357256 \pm 2(0.150822) = (-1.055612, -1.6589)$$

$$-0.345283 \pm 2(0.129656) = (-0.604595, -0.085971)$$

Scaled confidence intervals:

$$113.3978 \pm 2(5.327837) = (102.7421, 124.0535)$$

$$0.3827976 \pm 2(0.007099) = (0.3685996, 0.151796)$$

$$-3.775886 \pm 2(0.150822) = (-4.07753, -3.474242)$$

$$-0.9916528 \pm 2(0.129656) = (-1.2509648, -0.7323408)$$

In both models, an increase of one hundredth of an inch of green corresponds to a $0.13 increase in price, a one-stalk increase in the number of stalks of asparagus corresponds to a decrease in the price of asparagus (though quantitatively these differ by (|1.53394| - |1.357256| = 0.176684)), and more variation in size (a one unit increase in dispersion) causes a decrease in the price (again, quantitatively these differ by (|0.27554|-|0.345283| = -0.069743)).

Examining the statistical significance of the parameter estimates, using the 95% confidence level, all the independent variables GREEN, NOSTALKS, DISPERSE are statistically significant. This can be seen from the summary statistics tables from above, as well as the fact that Waugh's beta coefficient estimates lie within the confidence intervals.

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

The calculations for the least squares estimates on each parameter are as follow: Formula:

$$\beta_i = cov(x,y)/var(x)$$

$$\beta_1 = (3430.89)/(24317.19)$$
$$\beta_1 = 0.141$$

$$\beta_2 = (-100.92)/(61.33)$$
$$\beta_2 = -1.64$$

$$\beta_3 = (-87.430)/(83.486)$$
$$\beta_3 = -1.047$$

Based on the OLS estimates using the variances and covariances from the table provided in the assignment, again there are different coefficient estimates for each independent variable, which indicates that the inability to reconcile the difference in coefficient estimates comes from the variability in these variables, which is given by variance / covariance.

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, R2, 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: Which variables are most highly correlated? Which variables are almost orthogonal?

	GREEN	NOSTALKS	DISPERSE	PRICE
GREEN	1	-0.0140293	-0.1260535	0.7483428
NOSTALKS	NA	1.0000000	0.3500297	-0.4065620
DISPERSE	NA	NA	1.0000000	-0.3246445
PRICE	NA	NA	NA	1.0000000

The variables that are most highly correlated are: GREEN and PRICE, cor(GREEN, PRICE) = 0.74834 The variables that are nearly orthogonal are: GREEN and NOSTALKS, cor(GREEN, NOSTALKS) = -0.01403

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 R2 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 R2 measures would you have obtained? Why do they equal those from the ”correct” regressions?

```
## [1] "The R^2 for the model PRICE on GREEN is 0.560016878307452"
```

```
## [1] "The R^2 for the model PRICE on NOSTALKS is 0.165292651436427"
```

```
## [1] "The R^2 for the model PRICE on DISPERSE is 0.105394072228457"
```

The square root of R^2 for the model PRICE on GREEN is 0.74 The square root of R^2 for the model PRICE on NOSTALKS is 0.40 The square root of R^2 for the model PRICE on DISPERSE is 0.32

Comparing these values computed above with the correlation coefficients reported in the first row of the above table, we see that these are equal. The reason that these are equal except in sign is that R^2 = cor(PRICE,VARIABLE)^2, so least squares regression will estimate these as either positive or negative, while this R^2 formula will always compute these as positive. The reason that if we inadvertently ran the “reverse” regressions we’d find the equal R^2 measures is because of the nature of how R^2 is calculated. R^2 = (cor(x,y))^2, in our regular regression model PRICE~VARIBLE, so we can think of the reverse as R^2 = (cor(y,x))^2 in the regression model VARIABLE~PRICE, and this will give the same value in either case. (See the correlation matrix, the lower diagonal is eliminated because it is redundant, the correlations can all be shown in the upper diagonal)

c. Notice the value of the R2 measure from the simple regression of PRICE on GREEN, computed in part (b). What do you expect to happen to this value of R2 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 R2 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 R2 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 R2 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?

Comparing the SLR models, PRICE ~ GREEN & PRICE ~ NOSTALKS to the MLR model PRICE ~ GREEN + NOSTALKS: The R^2 for the SLR model PRICE on GREEN is 0.56 The R^2 for the SLR model PRICE on NOSTALKS is 0.165 The R^2 for the MLR model PRICE on GREEN and NOSTALKS is 0.7169.

Given the correlation between GREEN and NOSTALKS of -0.01, I would expect the R^2 of the MLR model to be higher than that of each individual SLR model. The reason for this is that we can think of each variable as orthogonal, they are not related and in theory both have explanatory power to explain the variation in price. Based on this correlation, drawing on the Valentine diagrams, we can think of there being no overlap between the two independent variables, so the change in R^2 would be small.

Comparing the SLR models, PRICE ~ GREEN & PRICE ~ DISPERSE to the MLR model PRICE ~ GREEN + DISPERSE: The R^2 for the SLR model PRICE on GREEN is 0.56 The R^2 for the SLR model PRICE on DISPERSE is 0.105 The R^2 for the MLR model PRICE ~ GREEN + DISPERSE is 0.614

Given the correlation between GREEN and DISPERSE of -0.12, I again would expect the Rˆ2 of the MLR model to be higher than that of each individual SLR model. The reason for this is that we can think of each variable as very independent from each other (though at -0.12, I would not say orthogonal), they are not related and in theory both have explanatory power to explain the variation in price.

Comparing the SLR models, PRICE ~ NOSTALKS & PRICE ~ DISPERSE to the MLR model PRICE ~ NOSTALKS + DISPERSE: The Rˆ2 for the SLR model PRICE on NOSTALKS is 0.165 The Rˆ2 for the SLR model PRICE on DISPERSE is 0.105 The Rˆ2 for the SLR model PRICE on NOSTALKS + DISPERSE is 0.203

Given the correlation between NOSTALKS and DISPERSE of 0.35, I again would expect the Rˆ2 of the MLR model to be only slightly higher than that of each individual SLR model. The reason for this is that we can think of each these variables as still independent from each other (less so than the previous two cases), so our model is still incorporating a new variable with new information, but this higher correlation means that the increase of the fit of the data will increase, but not as much as the previous two cases had increased.

d. In all three cases considered in part (c), the R2 from the multiple regression (with two regressors in addition to the constant) is less than the sum of the R2's from the corresponding two simple regressions. Is the R2 from the multiple regression equation with all three regressors (GREEN, NOSTALKS, and DISPERSE) greater than or less than the sum of the R2 from the three simple regressions? Note: It might be tempting to conclude from this that the R2 from a multiple regression with a constant term and K regressors is always less than or equal to the sum of the R2 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].

The Rˆ2 for the SLR model PRICE on GREEN is 0.56 The Rˆ2 for the SLR model PRICE on NOSTALKS is 0.165 The Rˆ2 for the SLR model PRICE on DISPERSE is 0.105 The Rˆ2 for the MLR model PRICE ~ GREEN + NOSTALKS + DISPERSE is 0.7268

The sum of the Rˆ2 for each individual simple linear regression models = 0.56 + 0.165 + 0.105 = 0.83. This is greater than the sum of the MLR model, whose Rˆ2 is 0.7268. (i.e., the Rˆ2 of the multiple regression is less than the sum of the Rˆ2 from the three simple regressions)

e. 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?

Calculating the coefficient of determination for each of the variables: Formula:

$$d^2_{xj} = b_j \Sigma_i (x_{ij-\bar{x}})(y_i - \bar{y})/(\Sigma_i (y_i - \bar{y})^2)$$

GREEN:

$$= 0.13826(3430.89/(1063.64))$$

$$= 0.4459731$$

NOSTALKS:

$$= -1.53394(-100.92/(1063.64))$$

$$= 0.1455429$$

DISPERSE:

$$= -0.27554(-82.35/(1063.64))$$
$$= 0.02133308$$

I was able to replicate the coefficients of determination for NOSTALKS and DISPERSE, though for GREEN, the coefficients differ by 0.04. Waugh's answer was nearly correct, though he should not have stated "squared variability". Waugh should have instead stated: "The coefficient of determination (for a model with the three variables- GREEN, NOSTALKS, and DISPERSE) is 0.57524, indicating that 57.524 percent of the variability in price is accounted for by the three factors studied.

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 R2 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?

In running these two regressions, the Rˆ2 does not change between the two models (the original model versus the new model that regresses price on fitted values). Both models have an Rˆ2 of 0.7268. The reason that these have the same Rˆ2 is because both models are effectively assessing the same relationship between price and explanatory variables GREEN, NOSTALKS, and DISPERSE. In the second model, fitted gives the predicted value for price based on the data given in each variable, though this is the same value that would have been predicted had we ran the initial regression of PRICE on each variable in the model. The reason that the estimated intercept term is zero is that the regression model no longer has to move the intercept up and down the y-axis to try and achieve the best fit (i.e., to minimize the sum of square residuals), it can achieve this by making the intercept = 0. The reason that the slope coefficient is unity is that the model is now predicting values for price using the predicted values for price from the regression as the predictor. In effect, the model is now a "perfect fit".

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 = B0 + B1LNMEM + B2LNMULT + B3LNACCESS + 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.

```
## [1] "F = 0.743188712170374"
```

```
## [1] "p-value = 0.731239220867676"
```

Based on the result of the F test conducted above, the correct statistical decision is to fail to reject the null hypothesis “the three”slope” coefficients were equal over the 1960-1965 time period”. In replicating Chow’s experiment, the conclusion of this Practicum has the same F-test value of 0.74 and the same statistical decision to fail to reject the null.

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.

```
## [1] "F = 0.743188712170374"
```

```
## [1] "p-value = 0.731239220867676"
```

Based on the results from the unrestricted regression model, the F statistic is 0.743, so as was the case with the model analyzing the years 1960 - 1965, the same statistical decision to reject the null hypothesis that “the three”slope” coefficients were equal over the 1954-1959 time period” is reached.

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?

```
## [1] "F = 266.346539025998"

## [1] "p-value = 0"

## [1] "F = 42.0512768525628"

## [1] "p-value = 0"
```

Interpreting the model that assumes equality of the slope parameters within the first and second generations: Based on the F test provided above, we reject the null hypothesis that the slope parameters within the first generation are equal to those in the second generation; the F statistic = 266.346

Interpreting the model that relaxes the assumption of slope parameter equality within each generation: In the second model, a different conclusion than that of previous models is reached; the F test from this regression that tests the equality of marginal effects across the entire timeframe of the dataset from 1954-1965: F = 42.05. The statistical decision to reject the null hypothesis that "the three"slope" coefficients were equal over the 1954-1965 time period" is reached.

Interpretation (for each model and comparing results across models):

Revisiting the first model, we can see that we reject the null that the slope parameters are equal across periods. This result is not surprising; this draws comparison over two different timeframes, and although these periods are right next to each other, this result shows that the 11 year timeframe was enough time for technology to advance to a point at which the marginal effect of an increase in the factors that make up a computer (as specified in the hedonic regression model) will exhibit diminishing marginal returns to the (rental) price of computers. Going to the second model (relaxing the assumption of slope parameter equality), this result does not surprise me either because although this is more of a test of year-to-year variation, when we expanded the time extent of data fed into the regression model, the test will exhibit different results than the previous tests ran in parts a and b.

These results are mutually consistent. Statistically they support each other. In the first test (that assumes equality of slope parameters within generations) we are testing if they are still equal across periods, we reject this null hypothesis, and this makes sense because the F test is now examining equality in marginal effects when technology has had more years to advance. Supporting this, the second regression model tests if parameters are equal across years. The first test had a much larger F-test value at over 200, the practical interpretation is that with more time for technology to advance, the diminishing marginal returns of technological improvement (that occurs over a span of time) are apparent when comparing across periods as there is more time for technology to advance, but are less apparent (F = 42.05, so still significant) across years that are next to each other.

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 = B0 + B_tDUM_it + B1LNMFM_i + B2LNMULT_i + B3LNACCFSSi

where DUMit 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 of Bt 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 Bt. denoted as B1955, B1956, B1957, ,B1965,. 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 D1955 to D1965, LNMEM, LNMULT, and LNACCESS. Compare year-to-year changes in the estimated coefficients of these 11dummy variables with the levels of the 11Bt estimates. Why is it appropriate to compare year-to-year changes in the estimated dummy variable coefficients with levels of the estimated Bt? Comment on and interpret any differences that appear to be substantial.

Table of the beta coefficient estimates measuring the pairs of adjacent years:

year	beta_coefficient
1954-1955:	B_DUM = -0.067
1955-1956:	B_DUM = -0.131
1956-1957:	B_DUM = -0.126
1957-1958:	B_DUM = -0.257
1958-1959:	B_DUM = -0.202
1959-1960:	B_DUM = -0.503
1960-1961:	B_DUM = -0.085
1961-1962:	B_DUM = -0.286
1962-1963:	B_DUM = -0.130
1963-1964:	B_DUM = -0.316
1964-1965:	B_DUM = -0.218

Each coefficient is interpreted as: The percent change in price due to the model being introduced in the current year versus the adjacent previous year.

Table of the beta coefficient estimates in the more traditional hedonic regression model:

year	beta_coefficient
1955:	B_DUM = -0.055
1956:	B_DUM = -0.212
1957:	B_DUM = -0.284
1958:	B_DUM = -0.476
1959:	B_DUM = -0.694
1960:	B_DUM = -1.139
1961:	B_DUM = -1.239
1962:	B_DUM = -1.622
1963:	B_DUM = -1.733
1964:	B_DUM = -2.028
1965:	B_DUM = -2.306

Comparing the year to year changes given by the eleven dummy variables to those of the levels of the 11 Bt estimates we can see that the marginal effect of year is increasing by years in the latter, whereas in the dummy variable case, the marginal effect remains consistent throughout the entire time frame for the most part.

It is important to compare the year-to-year changes in the estimated dummy variable coefficients with the levels of the estimated Bt because this allows us to control for the issue that question four highlighted. Recall, that question highlighted the concept that the marginal effect of changes in technology on price early on in the period were not the same as the marginal effect of changes in technology on price later on in the period. Comparing the beta coefficients between these two sets of models the dummy variable model gives further support to the conclusion that I drew in question four. Notice that over this time period, the marginal effect of year on price is becoming larger in magnitude, so the decrease in price attributed to the year increases as more time from the initial year 1954 has elapsed. Conversely, looking at the model measuring adjacent years, the regression again does not highlight this change because it is drawing a comparison between the data over a two year span rather than the entire 11 year span.

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, D1955 to D1965. Then construct a chained price index, using the following sequential procedure: For 1955, exponentiate B1955; for 1956, exponentiate the sum B1955+B1956; for 1957, exponentiate the sum B1955 + B1956 + B1957. Continue this for each year, until for 1965 the qualityadjusted price index is computed as the antilogarithm of the sum B1955 + B1956 + B1957 + . . . + B1965. 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?

.

.

.

.

.

Reporting the coefficient estimates for the traditional hedonic price index

year	beta_coefficient
1955:	exp(Year_55) -0.03178
1956:	exp(Year_56) -0.12355
1957:	exp(Year_57) -0.16557
1958:	exp(Year_58) -0.27701
1959:	exp(Year_59) -0.40393
1960:	exp(Year_60) -0.66283
1961:	exp(Year_61) -0.72099
1962:	exp(Year_62) -0.94384
1963:	exp(Year_63) -1.00831
1964:	exp(Year_64) -1.18035
1965:	exp(Year_65) -1.34190

Reporting the coefficient estimates for the chained price index:

year	beta_coefficient
1955:	exp(Year_55) 0.28401
1956:	exp(sum to Year_56) 0.29392
1957:	exp(sum to Year_57) 0.37803
1958:	exp(sum to Year_58) 0.45119
1959:	exp(sum to Year_59) 0.494290
1960:	exp(sum to Year_60) 0.45077
1961:	exp(sum to Year_61) 0.42425
1962:	exp(sum to Year_62) 0.352837
1963:	exp(sum to Year_63) 0.30229
1964:	exp(sum to Year_64) 0.19460
1965:	exp(sum to Year_65) -0.42125

Answering the question- Why is such an index called a chained price index? The reason that this index is called a chained price index is because the marginal effect of each additional year is measured while accounting for the sum of the marginal effects across years up to the year of interest.

Empirical comparison of the chained price index with the traditional hedonic price index Comparing the beta coefficients in the traditional price index with the beta coefficients in the chained price index, the data provided in the above tables can be interpreted in the following manner. Examining the traditional index, we see that the marginal effect on price for an elapsed time of one-year is negative, meaning that there is a percentage decrease in price across single year timespans. In the chained price index, the marginal effects are positive (though very slightly positive). There could be a data issue, or something with how I specified this model because I would not expect this marginal effect to be positive (in fact it should be increasingly negative as the cumulative effects over time should cause an increasingly large decrease in price). The only way I can reconcile this is that the dataset is small, so there aren't enough observations to see the expected trend (or I just made a mistake!)

Which index do you prefer and why? I prefer the traditional hedonic price index. This index is very intuitive; nearly anyone with familiarity in introductory statistics can understand how to interpret this, and finally, it is very widely used, which likely speaks to its efficacy as an economic measure. Furthermore, this hedonic price index can be applied in a wide variety of settings (i.e., house prices, computer parts, and many others).

Appendix

```
library(knitr)
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 4,
    fig.height = 4, tidy = TRUE)

# obtaining all of the needed packages
library(tidyverse)

# establish working directory:
mydir <- getwd()

# Loading data into the r environment
WAUGH <- read.csv("~/Desktop/Desktop/Grad School (MS)/Q1/Applied Econometrics I/Assignments/Assignment T
    sep = "")
model_1 <- lm(PRICE ~ GREEN + NOSTALKS + DISPERSE, data = WAUGH)
sum_1 <- summary(model_1)
kable(sum_1$coefficients, caption = "WAUGH Summary Statistics")
# code for problem 2b
M_price = mean(WAUGH$PRICE)
M_green = mean(WAUGH$GREEN)
M_nostalks = mean(WAUGH$NOSTALKS)
M_disperse = mean(WAUGH$DISPERSE)
# print('The means for each variable PRICE, GREEN, NOSTALKS, and DISPERSE
# respectively are:') print(cat(M_price,M_green,M_nostalks,M_disperse)) code
# for problem 2c
vc_matrix <- cov(WAUGH)
vc_matrix[lower.tri(vc_matrix)] <- NA  # Set lower triangular values to NA, as these are redundant
kable(vc_matrix)
# code for problem 3a
cor_matrix <- cor(WAUGH)
cor_matrix[lower.tri(cor_matrix)] <- NA  # Set lower triangular values to NA, as these are redundant
kable(cor_matrix)
# code for problem 3b
model_3 <- lm(PRICE ~ GREEN, data = WAUGH)
sum_3 <- summary(model_3)
correl_3 <- sqrt(sum_3$r.squared)
model_4 <- lm(PRICE ~ NOSTALKS, data = WAUGH)
sum_4 <- summary(model_4)
correl_4 <- sqrt(sum_4$r.squared)
model_5 <- lm(PRICE ~ DISPERSE, data = WAUGH)
sum_5 <- summary(model_5)
correl_5 <- sqrt(sum_5$r.squared)

print(paste0("The R^2 for the model PRICE on GREEN is ", sum_3$r.squared))
print(paste0("The R^2 for the model PRICE on NOSTALKS is ", sum_4$r.squared))
print(paste0("The R^2 for the model PRICE on DISPERSE is ", sum_5$r.squared))
# print(paste0('The R^2 for the model PRICE on GREEN is ', sum_3$r.squared, '
# and the square root of this is ', correl_3)) print(paste0('The R^2 for the
# model PRICE on NOSTALKS is ', sum_4$r.squared, ' and the square root of this
# is ', correl_4)) print(paste0('The R^2 for the model PRICE on DISPERSE is ',
# sum_5$r.squared, ' and the square root of this is ', correl_5))
```

```r
# code for problem 3c first portion of this question
model_6 <- lm(PRICE ~ GREEN + NOSTALKS, data = WAUGH)
sum_6 <- summary(model_6)
# print(sum_6$r.squared)

# second portion of this question
model_7 <- lm(PRICE ~ GREEN + DISPERSE, data = WAUGH)
sum_7 <- summary(model_7)
# print(sum_7$r.squared)

# third poriton of this question
model_8 <- lm(PRICE ~ DISPERSE + NOSTALKS, data = WAUGH)
sum_8 <- summary(model_8)
# print(sum_8$r.squared)

# code for problem 3d

# code for problem 3e



# code for problem 3f

# re-running the same linear model as before, for easy access of the fitted
# values needed in this question
model_1 <- lm(PRICE ~ GREEN + NOSTALKS + DISPERSE, data = WAUGH)
sum_1 <- summary(model_1)
sum_1R2 <- sum_1$r.squared
model1_fitted <- fitted(model_1)

WAUGH2 <- WAUGH
WAUGH2$fitted <- model1_fitted
model_9 <- lm(PRICE ~ fitted, data = WAUGH2)
sum_9 <- summary(model_9)
sum_9R2 <- sum_9$r.squared

# code for problem 4 setup

# load needed libraries
library(readxl)

# import dataset
CHOW <- read_excel("Practicum 2 Chow data (1).xlsx")


# preparing the data:

# creating variables:
CHOW$LNMEM <- (log(CHOW$WORDS * CHOW$BINARY * CHOW$DIGITS))
CHOW$LNMULT <- (log(CHOW$MULT))
CHOW$LNACCESS <- (log(CHOW$ACCESS))
CHOW$LNADD <- (log(CHOW$ADD))
CHOW$LNRENT <- (log(CHOW$RENT))
```

```r
# rewriting the CHOW dataset to have dummy variables for each year

# Creating the dataframe with dummy variable: create a copy of chow
CHOWDUM <- CHOW

# Vector of years you want to create columns for
years_to_create <- 54:65

# Create columns in a for loop
for (year in years_to_create) {
    col_name <- paste0("Year_", year)  # Create a column name
    CHOWDUM[[col_name]] <- ifelse(CHOWDUM$YEAR == year, 1, 0)
}

# code for problem 4a

# running the restricted model (pooled regression):
model_10 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + factor(YEAR), data = subset(CHOW,
    YEAR >= 60 & YEAR <= 65))
sum_10 <- summary(model_10)

# running the unrestricted model:
model_11 <- lm(LNRENT ~ factor(YEAR) * LNMEM + factor(YEAR) * LNMULT + factor(YEAR) *
    LNACCESS, data = subset(CHOW, YEAR >= 60 & YEAR <= 65))

# Unrestricted model, to get the summary statistics year by year
P2_interacted <- list()
for (year in 60:65) {
    response_var <- "LNRENT"
    independent_vars <- c("LNMEM", "LNMULT", "LNACCESS")

    interaction_terms <- lapply(independent_vars, function(var) {
        paste0("Year_", year, " * ", var)
    })

    all_terms <- c(paste0("Year_", year), interaction_terms)
    formula <- as.formula(paste(response_var, "~", paste(all_terms, collapse = " + ")))

    data_subset <- subset(CHOWDUM, YEAR >= (year - 1) & YEAR <= year)
    model <- lm(formula, data = data_subset)
    model_summary <- summary(model)
    P2_interacted[[as.character(year)]] <- list(model = model, summary = model_summary)
}


# F Test
r2_model_10 <- summary(model_10)$r.squared
r2_model_11 <- summary(model_11)$r.squared

q <- length(model_11$coefficients) - length(model_10$coefficients)  # extra dfs needed
n <- nrow(subset(CHOW, YEAR >= 60 & YEAR <= 65))  # number of observations
k <- length(model_11$coefficients) - 1  # # vars in ur
```

```
FStat = ((r2_model_11 - r2_model_10)/q)/((1 - r2_model_11)/(n - k - 1))
print(paste0("F = ", FStat))

print(paste0("p-value = ", 1 - pf(FStat, q, n - k - 1)))

# code for problem 4b

# repeating the analysis for the years 1954-1959 running the restricted model
# (pooled regression):
model_13 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + factor(YEAR), data = subset(CHOW,
    YEAR >= 54 & YEAR <= 59))
sum_13 <- summary(model_13)

# Unrestricted model
model_14 <- lm(LNRENT ~ factor(YEAR) * LNMEM + factor(YEAR) * LNMULT + factor(YEAR) *
    LNACCESS, data = subset(CHOW, YEAR >= 54 & YEAR <= 59))

# unrestricted model, to get the summary statistics in the P1_interacted by
# year:
P1_interacted <- list()
for (year in 54:60) {
    response_var <- "LNRENT"
    independent_vars <- c("LNMEM", "LNMULT", "LNACCESS")

    interaction_terms <- lapply(independent_vars, function(var) {
        paste0("Year_", year, " * ", var)
    })

    all_terms <- c(paste0("Year_", year), interaction_terms)
    formula <- as.formula(paste(response_var, "~", paste(all_terms, collapse = " + ")))

    data_subset <- subset(CHOWDUM, YEAR >= (year - 1) & YEAR <= year)
    model <- lm(formula, data = data_subset)
    model_summary <- summary(model)
    P2_interacted[[as.character(year)]] <- list(model = model, summary = model_summary)
}


# F Test
r2_model_13 <- summary(model_10)$r.squared
r2_model_14 <- summary(model_11)$r.squared

q <- length(model_14$coefficients) - length(model_13$coefficients)  # extra dfs needed
n <- nrow(subset(CHOW, YEAR >= 60 & YEAR <= 65))  # number of observations
k <- length(model_11$coefficients) - 1  # # vars in ur

FStat = ((r2_model_11 - r2_model_10)/q)/((1 - r2_model_11)/(n - k - 1))
print(paste0("F = ", FStat))

print(paste0("p-value = ", 1 - pf(FStat, q, n - k - 1)))

# code for problem 4c
# in order to assume equality of the slope parameters within the first and
# second generations, we must create dummy variables over each period

CHOW$PERIOD <- ifelse(CHOW$YEAR < 60, 1, 0)

# model where the slope is held constant within each period, the argument is
# period

# running the restricted model (pooled regression):
model_19 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + factor(PERIOD), data = subset(CHOW,
    YEAR >= 54 & YEAR <= 65))
sum_19 <- summary(model_19)

# unrestricted model;
model_20 <- lm(LNRENT ~ factor(PERIOD) * LNMEM + factor(PERIOD) * LNMULT + factor(PERIOD) *
    LNACCESS, data = subset(CHOW, YEAR >= 54 & YEAR <= 65))
sum_20 <- summary(model_20)

# F - test:
r2_mod19 <- summary(model_19)$r.squared
r2_mod_20 <- summary(model_20)$r.squared
q <- length(model_20$coefficients) - length(model_19$coefficients)  # extra dfs needed
n <- nrow(subset(CHOW, YEAR >= 54 & YEAR <= 65))  # number of observations
k <- length(model_20$coefficients) - 1  # # vars in ur
FStat = ((r2_mod19)/q)/((1 - r2_mod_20)/(n - k - 1))
print(paste0("F = ", FStat))

print(paste0("p-value = ", 1 - pf(FStat, q, n - k - 1)))

# model where slope is allowed to vary across periods, the argument is year

# running the restricted model (pooled regression):
model_17 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + factor(YEAR), data = subset(CHOW,
    YEAR >= 54 & YEAR <= 65))
sum_17 <- summary(model_17)

# unrestricted model;
model_18 <- lm(LNRENT ~ factor(YEAR) * LNMEM + factor(YEAR) * LNMULT + factor(YEAR) *
    LNACCESS, data = subset(CHOW, YEAR >= 54 & YEAR <= 65))
sum_18 <- summary(model_18)

# F - test:
r2_mod15 <- summary(model_17)$r.squared
r2_mod_16 <- summary(model_18)$r.squared
q <- length(model_18$coefficients) - length(model_17$coefficients)  # extra dfs needed
n <- nrow(subset(CHOW, YEAR >= 54 & YEAR <= 65))  # number of observations
k <- length(model_18$coefficients) - 1  # # vars in ur
FStat = ((r2_mod15)/q)/((1 - r2_mod_16)/(n - k - 1))
print(paste0("F = ", FStat))

print(paste0("p-value = ", 1 - pf(FStat, q, n - k - 1)))
```

```r
# code for problem 5a


# Running the dummy variable regression models- done manually to establish the
# pattern needed in the for loop: mod_1 <- lm(LNRENT ~ Year_55 + LNMEM + LNMULT
# + LNACCESS, data=subset(CHOWTEST, YEAR >= 54 & YEAR <= 55)) mod_1_sum <-
# summary(mod_1) mod_2 <- lm(LNRENT ~ Year_56 + LNMEM + LNMULT + LNACCESS,
# data=subset(CHOWTEST, YEAR >= 55 & YEAR <= 56)) mod_2_sum <- summary(mod_2)

# Loop through the years and fit the regression models
model_list <- list()
for (year in 54:65) {
    response_var <- "LNRENT"
    independent_vars <- c(paste0("Year_", year), "LNMEM", "LNMULT", "LNACCESS")

    formula <- as.formula(paste(response_var, "~", paste(independent_vars, collapse = " + ")))

    data_subset <- subset(CHOWDUM, YEAR >= (year - 1) & YEAR <= year)
    model <- lm(formula, data = data_subset)
    model_summary <- summary(model)

    model_list[[as.character(year)]] <- list(model = model, summary = model_summary)
}

# Running the regression with dummy variables for each year:
mod_12 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + factor(YEAR), data = CHOW)
sum_mod_12 <- summary(mod_12)


# code for problem 5b

# creating the chained price index (by exponentiating the sum of beta
# coefficients for each year)
model_20_5 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55) + exp(Year_56) +
    exp(Year_57) + exp(Year_58) + exp(Year_59) + exp(Year_60) + exp(Year_61) + exp(Year_62) +
    exp(Year_63) + exp(Year_64) + exp(Year_65), data = CHOWDUM)

mod_list21_30 <- list()
# chained price index: # another case for a for-loop
model_21 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_21))

model_22 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_22))

model_23 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57),
    data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_23))

model_24 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57 +
    Year_58), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_24))
```

```r
model_25 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57 +
    Year_58 + Year_59), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_25))

model_26 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57 +
    Year_58 + Year_59 + Year_60), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_26))

model_27 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57 +
    Year_58 + Year_59 + Year_60 + Year_61), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_27))

model_28 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57 +
    Year_58 + Year_59 + Year_60 + Year_61 + Year_62), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_28))

model_29 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57 +
    Year_58 + Year_59 + Year_60 + Year_61 + Year_62 + Year_63), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_29))

model_30 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57 +
    Year_58 + Year_59 + Year_60 + Year_61 + Year_62 + Year_63 + Year_64), data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_30))

model_31 <- lm(LNRENT ~ LNMEM + LNMULT + LNACCESS + exp(Year_55 + Year_56 + Year_57 +
    Year_58 + Year_59 + Year_60 + Year_61 + Year_62 + Year_63 + Year_64 + Year_65),
    data = CHOWDUM)
mod_list21_30 <- append(mod_list21_30, summary(model_31))
```