Methodological Note

Regression Analysis of the Determinants of Repeat Business

Regression analysis predicts the values of a dependent variable by investigating explanatory or predictor variables. Linear regression seeks to describe the relationship between the dependent (Y) and predictor (X) variables by finding the linear equation ($Y = \alpha + \beta X$) that best fits the series of ordered pairs. However, the regression will not predict the relationship exactly: there will be a residual error term (ε). Consequently, the simplest linear regression equation is $Y = \alpha + \beta X + \varepsilon$. Though the basic concept underlying regression analysis is easy to grasp, the technique possesses many complications. Linear regression will not generate a meaningful result if, for instance, the relationship between the dependent and explanatory variables is nonlinear. It is also important for ε to be distributed smoothly. If an important predictor variable is omitted, the result can be biased. Though more predictors can be added to the model, these should not be closely correlated with one another. Tests have been devised to measure the explanatory power of the regression, the significance of beta parameters, and the robustness of the model selected.[1]

The object of the present exercise is to analyze the determinants of Richard Poor's repeat business with his clients. There are three ways of measuring repeat business using the account book: by the *duration* of a client's account (the time elapsing in days from the first and to the last recorded transaction), by the total *number* of transactions with a client, and by the total *value* of a client's account. It is important to specify how these variables are related to one another. The following causal path makes good sense: the longer the duration of a customer's account, the greater the number of expected transactions; likewise, the greater the number of transactions, the higher the expected total value of a client's business. There is, however, a complication. It is only possible to observe Poor's business dealings from May 25, 1699, to July 8, 1713. Clearly, a client whose first transaction is dated (hypothetically) on June 26, 1707, has more opportunities for business than one opening an account on June 26, 1713. Therefore, it is necessary to control for the date a business relationship begins by creating a variable called "span" that records the interval of time from the first of the client's transactions to the end of the account book itself on July 8, 1713.

For regression to work effectively, it is important that the data used are not too badly skewed. Unfortunately, both the number of transactions and the value of accounts are skewed: a few clients traded with Poor many times and had large accounts. Skewness affects the error term ε, which measures residuals (the difference between predicted and actual scores). Regression only works effectively if these residuals are normally distributed. The problem of skewness can be reduced, however, by transforming the data into natural logarithms.

The account book and other sources discussed in the essay permit analysis of the following client characteristics as possible determinants of repeat business: gender, head of household, dealer in plantation produce, and Jewishness. All these characteristics are indicator variables. By definition, an indicator variable can be coded in binary form: for example, 1 = male and 0 = female. The regression estimates the effect of any one of these indicators while holding the others constant.

Returning to the causal path outlined above, three separate regressions are calculated (Tables I, II, and III). Table I uses account duration as the dependent variable (the measure of repeat business being predicted). Only indicator variables are used to predict duration. Table II uses number of transactions as the dependent variable, adding account duration as a control (since the argument is that the longer an account endures, the more transactions will be recorded). Table III uses value of transactions as the dependent variable, adding account duration and number of transactions as controls. At each stage the regression drops variables that are statistically insignificant. This technique is termed "backwards selection" because it begins by including all predictors in the model but then identifies and drops those not contributing significantly to the regression. In each of the three tables, the coefficients (the betas in the linear equation $\alpha + \beta_1 x + \beta_2 x$

$+ \varepsilon$) measure the effect of a 1 percent increase in the predictor variable on repeat business. However, where the predictor is an indicator variable that takes the value of 1 or 0, the ratios convey a better impression of the impact factor. Ratios are the natural logarithms of the coefficients. The final two columns of each table report the t-statistic and related probability that the beta coefficients are statistically different from zero. Conventionally, a p-value of less than 0.05 is used as a cutoff for significance; such a value indicates that there is less than a 5 percent probability that the conclusion beta differs from zero.

In Table I, controlling for the date an account was opened with Poor, a client's being a householder boosts duration by a factor of 2.47, and being a staples dealer boosts duration by a factor of 5.13. Overall, the regression predicts 18 percent of the variation in clients' repeat business measured by duration.

In Table II, increasing account duration by 1 percent is associated with a 0.3 percent increase in the number of transactions. Overall, the regression predicts 64 percent of the variation in clients' repeat business measured by number of transactions.

In Table III, controlling for the number of transactions, the impact factor of a client's being a staples dealer boosts the total value of an account by a factor of 1.64. Overall, the regression predicts 68 percent of the variation in repeat business measured by total value of transactions.

In none of the regressions were the indicator variables for gender or Jewishness found to be significant.

[1] For further discussion of the principles of regression and its historical applications, see Charles H. Feinstein and Mark Thomas, *Making History Count: A Primer in Quantitative Methods for Historians* (Cambridge, 2002), 93–106.

TABLE I
REGRESSION ANALYSIS OF THE DETERMINANTS OF REPEAT BUSINESS, 1707–13
DEPENDENT VARIABLE: ACCOUNT DURATION

Predictors	**Coefficient**	**Standard Error**	**Ratio**	**t-Statistic**	**Probability**
Span	1.02	0.22		4.72	0.00
Household head	0.91	0.35	2.47	2.55	0.01
Staples dealer	1.64	0.42	5.13	3.92	0.00
Constant	-3.69	1.57		-2.36	0.02

Notes: n = 172; F (3, 168) = 13.13; R^2(adj) = 0.18. "Ratio" applies only to binary variables. Dropped predictors: Jewish, gender.

Source: "Journall: No: A: Belonging: to Richard: Poor: Junr," MS Eng 7, John Carter Brown Library.

TABLE II
REGRESSION ANALYSIS OF THE DETERMINANTS OF REPEAT BUSINESS, 1707–13
DEPENDENT VARIABLE: NUMBER OF TRANSACTIONS

Predictors	**Coefficient**	**Standard Error**	**t-Statistic**	**Probability**
Account duration	0.30	0.02	17.51	0.00
Constant	0.10	0.09	1.20	0.23

Notes: n = 172; F (1, 170) = 306.59; R^2(adj) = 0.64. Dropped predictors: Jewish, span, gender, household head, staples dealer.

Source: "Journall: No: A: Belonging: to Richard: Poor: Junr," MS Eng 7, John Carter Brown Library.

TABLE III
REGRESSION ANALYSIS OF THE DETERMINANTS OF REPEAT BUSINESS, 1707–13
DEPENDENT VARIABLE: VALUE OF TRANSACTIONS

Predictors	**Coefficient**	**Standard Error**	**Ratio**	**t-Statistic**	**Probability**
Number of transactions	1.23	0.07		17.11	0.00
Staples dealer	0.50	0.16	1.64	3.08	0.00
Constant	0.84	0.12		7.22	0.00

Notes: n = 172; F (2, 169) = 180.93; R^2(adj) = 0.68. "Ratio" applies only to binary variables. Dropped predictors: Jewish, household head, account duration, span, gender.

Source: "Journall: No: A: Belonging: to Richard: Poor: Junr," MS Eng 7, John Carter Brown Library.