Econometrics II: Assignment 1

Due: Friday, January 31^{st}

1. Consider projecting Y on 1 (i.e., a constant) and X. All r.v.'s in this question are scalar valued, and we write the projection as $\hat{Y} = \hat{\alpha} + \hat{\beta}X$.

1.1 Show that the data matrix expression for $(\hat{\alpha}, \hat{\beta})$ simplifies to $\hat{\beta} = \sum(X_i - \overline{X})(Y_i - \overline{Y}))/\sum(X_i - \overline{X})^2$.

1.2 Show that R^2 is the squared sample correlation coefficient between X and Y (hence its name).

1.3 Consider the reverse projection of X on Y (and a constant). Explain how the projection coefficients and R^2 will relate.

2. (from an old exam) In rank-rank regression, covariate and outcome are expressed in terms of their rank in their respective marginal distribution. Rank-rank regression is popularly applied in analysis of intergenerational mobility, e.g. by regressing income percentiles of children (observed at adult age) on those of their parents.

Now to the actual question: I recently collected observations on $n = 999$ children and their mothers. Each child was assigned an outcome $Y_i \in \{1, \ldots, 999\}$ based on household income, where $Y_i = 1$ corresponded to the poorest child and so on. Similarly, each child's mother was assigned a rank position $X_i \in \{1, \ldots, 999\}$ corresponding to the rank position of her household income. There was no overlap in parents and no tie in household income, so that both X_i and Y_i took each value from 1 to 999 exactly once.

I wanted to regress Y on X. However, I first accidentally regressed X on Y, then poured coffee over my laptop, and then the dog ate my data. The only thing I remember is that the estimated slope was exactly 3/5.

2.1 What value did the estimated intercept take?

2.2 What value did R^2 take?

2.3 What values did intercept, slope, and R^2 take in the *intended* (i.e., reverse) regression?

3 Consider projection of Y on one categorical variable X that takes finitely many values. (Examples include gender, race, income if recorded in brackets, industry,...)

Suppose that X is recorded numerically, i.e. possible values are $x \in \{0, 1, \ldots, M\}$. We will compare direct projection of Y on X to defining indicator variables $Z_x \equiv \mathbf{1}\{X = x\}$ and projecting Y on $(Z_1, \ldots, Z_M)$.

3.1 Under what condition is the first projection well-defined with/without a constant? What about the second one?

Your answer should have implications for the appropriate interpretation of projection coefficients in the second regression. Explain. (Clarification: I here don't ask for anything deeper than you would have learned in a relevant UG class.)

3.2 For the second projection, prove that the fitted values for different values of X recover the corresponding conditional sample averages.

3.3 Can you decide ex ante which regression has the higher R^2? Can you give conditions under which R^2 will be the same?

(Hint: Attack the last question with general knowledge about constrained optimization, not with linear algebra.)

4 Consider projection of the scalar Y on (a constant and) X and potentially also on X^2. (Assume that expectations exist.)

4.1 Give a (tight) condition under which the population projection coefficient of Y on X is defined.

4.2 Give a (tight) condition under which the population projection coefficient of Y on (X, X^2) is defined. Give a simple counterexample to your condition.

4.3 Write the above projections as $\tilde{Y} = \hat{a} + \hat{b}X$ respectively $\hat{Y} = \hat{\alpha} + \hat{\beta}X + \hat{\gamma}X^2$. Give a condition on the distribution of X under which $\hat{b} = \hat{\beta}$.

(Hint/request: Please relate your answer to Frisch-Waugh decomposition.)