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.)

Answer Key

1 This is mostly mechanical. Remember that for simple linear regression, the slope parameter and the (not causally interpetable!) reverse slope parameter are related by their product equalling

$$\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)} \cdot \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(Z)} = \left(\frac{\operatorname{cov}(X,Y)}{\operatorname{sd}(X)\operatorname{sd}(Y)}\right)^2 = \rho^2,$$

i.e. the product is the squared correlation coefficient or "population R^2".

2.1 Recalling that, in OLS, the fitted line must pass through the sample averages, we have

$$\overline{X} = \hat{\beta}_0 + \hat{\beta}_1 \overline{Y} \implies \hat{\beta}_0 = \overline{X} - \hat{\beta}_1 \overline{Y} = 500 - 3/5 \cdot 500 = 200.$$

Unfortunately, this gets much messier if you don't spot, e.g. from symmetry considerations, that the sample averages are 500.

2.2 Here and later, it is important to spot that X and Y by construction have the same sample variance. It is not recommended to evaluate that variance!

$$R^2 = \frac{(\hat{\operatorname{cov}}(Y,X))^2}{\hat{\operatorname{var}}(Y)\hat{\operatorname{var}}(X)} = \frac{(\hat{\operatorname{cov}}(Y,X))^2}{(\hat{\operatorname{var}}(Y))^2} = \hat{\beta}_1^2 = \frac{9}{25}.$$

2.3 For the same reason (i.e., sample variances are the same), all numbers are the exact same in the reverse regression.

3.1 Assuming all possible values of X are taken in sample, the projection on dummies is well-defined with a constant only if one value of X is dropped ("indicator variable trap").

Recall that the appropriate interpretation of a coefficient in a regression on categorical indicators therefore is "effect relative to the omitted value."

3.2 This is "saturated regression." I can generate the sample averages as fitted values by setting the intercept equal to the sample average conditional on the omitted value of X and all other coefficients equal to the corresponding differences in sample averages. Since these fitted values solve the unconstrained (by linearity of predictors) least squares prediction problem, they also solve the constrained one.

3.3 The saturated regression can always produce the same fitted values (by having no constant and coefficients $(\hat{\beta}_0 + \hat{\beta}_1, \hat{\beta}_0 + 2\hat{\beta}_1, \ldots)$) and so necessarily has higher R^2.

4.1 We need X to take two distinct values so that it is not collinear with the constant.

4.2 For this we need that none of $(1, X, X^2)$ are collinear. This requires X to take at least 3 values (with 2 values, X and X^2 are collinear), but 3 values are also sufficient because values of $(1, X, X^2)$ that are collinear must solve the same deterministic equation, and that is going to be one quadratic equation in X that therefore has at most 2 distinct real solutions. (See the pattern? What is the answer for $(1, X, X^2, X^3)$?)

4.3 That will be true if X is uncorrelated with X^2, which is true when its distribution is symmetric about 0.