Three Basic Theorems for Regression

Georges Monette

2023-10-03

Contents

These are three basic theorems that help to elucidate the relationship between various models that aim to estimate the effect of a variable or group of variables controlling for other variables.

The first theorem, the projection theorem, simply states that if you express a vector Y as a sum of a linear combination of columns of a matrix X plus a vector that is orthogonal to the columns of X, then you have the unique decomposition corresponding to the least-squares regression or Y on X.

The second theorem, which formally states the equivalence between the simple regression in the 'Added Variable Plot' – also known as the 'Partial Regression Leverage Plot' in PROC REG in SAS – and the estimated coefficient for the corresponding variable in a multiple regression. The theorem shows how the AVP is sufficient assuming the validity of the model and also provides a useful visual diagnostic to detect influential points and other model violations.

The third theorem concerns the properties of a linear version of the 'propensity score' that has come to play such an important role in what is known as the Rubin Causal Model for inference. The idea is that, if you want to estimate the 'effect' of a variable X controlling for a number of other variables, $Z_1, Z_2, ..., Z_k$ then the coefficient for X in the full multiple regression will be the same as the coefficient in the regression with only two predictors: X and $\hat{X}$, where $\hat{X}$ is the predicted value of X based on the variables $Z_1, Z_2, ..., Z_k$. This implies that you can estimate the 'effect' of X controlling for a possibly large number of confounding factors by controlling only for a subset of the factors that provide the same prediction of X that would be provided by the entire set. For example, if $Z_{h+1}, ..., Z_k$ are related to X only through $Z_1, ..., Z_h$, then, in order to get an unbiased estimate of the coefficient of X controlling for $Z_1, ..., Z_k$, it suffices to perform a multiple regression on X

and $Z_1, ..., Z_h$.

This is a simple linear version of the more general ideas developed in graphical causal models by Judea Pearl and others. The idea that you only need to control for a subset of confounding factors corresponds to the 'back-door' criterion in causal models.

There is both a gain and a price to pay for the use of a simpler model. The model to predict X may be much better understood than the model that generates Y which can result in an estimate of the effect of X in whose validity you have greater confidence. The price that you pay is that the model for Y may have a greater residual standard error so that the estimate of the effect of X has a greater standard error. One sacrifices statistical precision for greater confidence in accuracy.

1 The Projection Theorem

The projection theorem says that if you express a vector Y as a sum of two components: a linear combination of a matrix X of full column rank and a vector that is orthogonal to the space spanned by X, then the coefficients of the linear combinations of columns of X are the unique least-squares regression coefficients and the second vector is the unique least-squares residual vector.

This theorem is useful because we often manipulate variables and matrices into this form and the theorem allows us to immediately conclude that we have identified the least-square coefficients and the residual vector.

Theorem: *(Projection theorem)* Let X be a matrix of full column rank and let

$$Y = Xb + e$$

where e is orthogonal to span(X)

Then b is the vector of estimated coefficients for the least-squares regression of Y on X, e is the least-squares residual vector and $||e||^2 = e'e$ is the least-squares error sum of squares, SSE.

Proof:

Let $\hat{\beta}$ be the least-squares regression coefficients.

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
&= (X'X)^{-1}X'Xb + (X'X)^{-1}X'e \\
&= b + (X'X)^{-1}0 \quad \text{since } e \perp \text{span}(X), \text{ i.e. } X'e = 0 \\
&= b
\end{aligned}$$

and thus $e = Y - X\hat{\beta}$, the least-squares residual vector. □

Corollary: Let X_1 and X_2 be matrices such that the partitioned matrix $[X_1X_2]$ is of full column rank. Let

$$Y = X_1b_1 + X_2b_2 + e \text{ with } e \perp \text{span}(X_1, X_2)$$

Then b_1 and b_2 are equal to the coefficients of the least-squares regression of Y on X_1 and X_2 and e is the least-squares residual.

2 The Added Variable Plot Theorem

Our next theorem is the 'Added Variable Plot Theorem' (AVP) more formally known as the Frisch-Waugh-Lovell Theorem. It took 30 years to prove it but here's an easy proof based on the projection theorem.

In terms of the AVP, the theorem states that a simple regression in the AVP for the regression of Y on X_1 controlling for X_2 provides the same inferences (except for the number of degrees of freedom for error) for the effect of X_1 as the multiple regression of Y on X_1 and X_2. Note that X_1 and X_2 can each contain one or more columns. In most applications with an intercept term, the column of 1's will be in X_2.

Let's first define the AVP. We define it so it can be used for more than one variable in X_1 although in typical applications X_1 has only one column.

Definition: *(Added Variable Plot)* Consider a regression of Y on two blocks of predictors X_1 and X_2. The AVP for X_1 adjusted for X_2 is the plot of the residuals of Y regressed on X_2 against the residuals of X_1 regressed on X_2.

Theorem: *(AVP or Frisch-Waugh-Lovell)* Consider the regression of Y on two blocks of predictors, X_1 and X_2, where the partitioned matrix, $[X_1 X_2]$ is of full column rank. Suppose

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + e, \text{ with } e \perp \text{span}(X_1, X_2)$$

i.e. $\hat{\beta}_1$ is the regression coefficient for X_1 in the least-squares multiple regression of Y on X_1 and X_2. Then the (residual of Y regressed on X_2) regressed on the (residual of X_1 on X_2) has

1. least-squares regression coefficient $\hat{\beta}_1$,
2. least-squares residual equal to e, and
3. $SSE = e'e$.
4. $\text{Var}_{AVP}(\hat{\beta}_1) = \text{Var}_{MR}(\hat{\beta}_1)$ where Var_{AVP} and Var_{MR} denote the variances calculated from the Added-Variable-Plot regression and the Multiple Regression, respectively.

Proof: The residual of Y in the regression on X_2 is obtained by pre-multiplying Y by

$$Q_2 = I - P_2 = I - X_2(X_2'X_2)^{-1}X_2'$$

and similarly for X_1. We obtain

$$Q_2Y = Q_2X_1\hat{\beta}_1 + Q_2X_2\hat{\beta}_2 + Q_2e$$

Now, $Q_2X_2 = 0$, so that $Q_2X_2\hat{\beta}_2 = 0$. Also, since $e \perp \text{span}(X_2)$ it follows that $Q_2e = e$. Thus

$$Q_2Y = Q_2X_1\hat{\beta}_1 + 0 + e$$

Moreover,

$$\begin{aligned} e'Q_2X_1 &= e'X_1 \text{ since} e'Q_2 = e' \\ &= 0' \text{ since } e \in \text{span}^{\perp}(X_1, X_2) \subset \text{span}^{\perp}(X_2) \end{aligned}$$

Thus, by the Projection Theorem, $\hat{\beta}_1$ is the regression coefficient of Q_2Y on Q_2X_1 and has $SSE = e'e$.

Now,

$$\begin{aligned} \text{Var}_{AVP}(\hat{\beta}_1) &= \sigma^2((Q_2X_1)'(Q_2X_1))^{-1} \\ &= \sigma^2(X_1'Q_2X_1)^{-1} \end{aligned}$$

and, using the Schur complement for the inverse of a partitioned matrix, we get:

$$\text{Var}_{MR}(\hat{\beta}) = \sigma^2 \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1}$$

and

$$\begin{aligned} \text{Var}_{MR}(\hat{\beta}_1) &= \sigma^2(X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1)^{-1} \\ &= \sigma^2(X_1'(I - X_2(X_2'X_2)^{-1}X_2')X_1)^{-1} \\ &= \sigma^2(X_1'Q_2X_1)^{-1} \\ &= \text{Var}_{AVP}(\hat{\beta}_1) \end{aligned}$$

□

Corollary: Consider a regresion of Y on X_1 and X_2 where the partitioned matrix $[X_1X_2]$ is of full column rank. Let Z be a matrix of full column rank such that $\text{span}(Z) = \text{span}(X_2)$. Then the regression coefficient(s) of Y on X_1 in the regression of Y on X_1 and X_2 is the same as the regression coefficient(s) of Y on X_1 in the regression of Y on X_1 and Z.

3 The 'Linear Propensity Score' Theorem

Finally we show that, the partial coefficient for the regression of Y on X_1 adjusting for X_2 is the same as the partial coefficient for the regression of Y on X_1 adjusting for the predictor of X_1 based on X_2, i.e. $P_2X_1 = X_2(X_2'X_2)^{-1}X_2'X_1$. However, the SSE of this regresion may be larger than that of the full multiple regression which is equal to that of the AVP regression.

Theorem: *(Linear Propensity Score)* Consider the regression of Y on two blocks of predictors, X_1 and X_2, where the partitioned matrix $[X_1X_2]$ is of full rank.

Suppose

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + e, \text{ with } e \perp \text{span}(X_1, X_2)$$

Consider, also, the regression of Y on X_1 and the predicted value of X_1 based on X_2.

Then the regression coefficients on X_1 are the same for both regressions and

$$Y = X_1\hat{\beta}_1 + \hat{X}_{1|2}\hat{\gamma}_2 + d + e$$

with $d \in \subset \text{span}(X_2) \cap \text{span}(\hat{X}_{1|2})^{\perp}$. Thus $d \perp e$, $d + e \perp \text{span}(X_1, \hat{X}_{1|2})$ and the SSE for the second regression is at least as large as that of the first regression and is equal to $d'd + e'e$.

Proof: To simplify the proof, we assume that X_1 has only one column but the more general result can also be proven.

Let $\hat{X}_{1|2} = P_2 X_1$ where $P_2 = X_2(X_2'X_2)^{-1}X_2'$. Let Z be a basis for the orthogonal complement of $\hat{X}_{1|2}$ in $\text{span}(X_2)$.

Then by the corollary to the AVP theorem, the coefficient of X_1 in the regression of Y on $X_1, \hat{X}_{1|2}$ and Z is $\hat{\beta}_1$ so that

$$\begin{aligned} Y &= X_1\hat{\beta}_1 + \hat{X}_{1|2}\hat{\gamma}_2 + Z\hat{\gamma}_3 + e \\ &= X_1\hat{\beta}_1 + \hat{X}_{1|2}\hat{\gamma}_2 + d + e \end{aligned}$$

Now we show that Z, hence d, is orthogonal to X_1. Consider that $\hat{X}_{1|2} = P_2 X_1$. Since $\text{span}(Z) \perp \text{span}(\hat{X}_{1|2})$ by construction, we have $Z'P_2X_1 = 0$. But $\text{span}(Z) \subset \text{span}(X_2)$ so that $P_2 Z = Z$.

Thus, $0 = Z'P_2X_1 = Z'X_1$ and $d \in \text{span}(Z)$ is orthogonal to $\text{span}(X_1, \hat{X}_{1|2})$. Since $d \in \text{span}(X_2)$, we have $d \perp e$ and the result follows. □