Piecewise Linear Instrumental Variable Estimation of Causal Influence

Richard Scheines
Dept. of Philosophy
Carnegie Mellon Univ.
Pittsburgh, PA 15213

Greg Cooper
Center for Biomedical Informatics
Univ. of Pittsburgh
Pittsburgh, PA 15213

Changwon Yoo
Center for Biomedical Informatics
Univ. of Pittsburgh
Pittsburgh, PA 15213

Tianjiao Chu
Dept. of Philosophy
Carnegie Mellon Univ.
Pittsburgh, PA 15213

Abstract

Instrumental Variable (IV) estimation is a powerful strategy for estimating causal influence, even in the presence of confounding. Standard IV estimation requires that the relationships between variables is linear. Here we relax the linearity requirement by constructing a piecewise linear IV estimator. Simulation studies show that when the causal influence of X on Y is non-linear, the piecewise linear is an improvement.

1 INTRODUCTION

In non-experimental settings, estimating the causal influence of one variable X on another Y is difficult primarily because of confounders (unmeasured common causes of X and Y). Several strategies for dealing with confounders have been suggested. The first is to identify the possible confounders, measure them, and then statistically control for them. This approach, although the most common in practice, is far from optimal. First, we can never be sure we have identified all the confounders, and second, we might have measured some of them badly. Either problem will leave us with a biased estimate of the effect of X on Y. Another strategy is to do a sensitivity analysis [Rosenbaum, 1995] in which the estimate of X's effect on Y is bounded relative to a parameter that expresses how much of the association between X and Y is due to confounding. This strategy supposes that one can sensibly parameterize the "amount of confounding," but, as Spirtes [1999] has pointed out, one can only do so in very limited circumstances. A third strategy [Spirtes and Cooper, 1999] is to find a "causal instrument" Z such that Z is prior to X and Y, Z is associated with X, but Z and Y are independent conditional on X. If one assumes that the causal model underlying the data is Markov and faithful to the population it describes [Spirtes, Glymour, and Scheines, 2000], then these two conditions eliminate the possibility that X and Y are confounded, thus the non-experimental, observed association between X and Y is the appropriate estimator of the causal dependence of Y on X.

The disadvantage of the Spirtes-Cooper strategy is that X and Y must be unconfounded. In cases where X and Y are confounded, their strategy will return "no estimator found." A fourth strategy that allows for X and Y to be confounded is instrumental variables [Bowden & Turkinton, 1984]. In their typical form, instrumental variable (IV) estimators require several assumptions in order to give consistent estimates of the effect of X on Y. The work we present here is aimed at generalizing the IV framework by relaxing the requirement of linearity. In what follows, we sketch linear IV-estimation and discuss the assumptions it requires. We then sketch our improvement, which involves creating a piecewise linear IV-estimator. In the final section, we describe an experiment comparing regular regression, linear IV-estimation, and piecewise linear IV-estimation on simulated data.

2 IV ESTIMATION

Classical linear IV-estimation works on causal models like the one pictured in Figure 1.

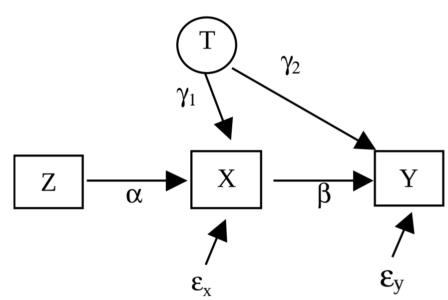

Figure 1: IV estimation model

Z, X, and Y are measured variables, T is an aggregated variable that includes all the unmeasured common causes of X and Y (the confounding), ε_x and ε_y are "error terms," and α, β, γ_1, and γ_2 are real valued linear coefficients. Z, T, ε_x, and ε_y are assumed to be pair wise independent, and the "structural equations" which define this model are:

$$X = \alpha Z + \gamma_1 T + \varepsilon_x$$
$$Y = \beta X + \gamma_2 T + \varepsilon_y$$

As Spirtes, et. al (2000) and Pearl (2000) have discussed at length, the coefficient β in this model represents the causal effect of X on Y. Assuming, without loss of generality, that Z, X, and Y are standardized to have mean 0 and variance 1, then according to this model, $\rho_{XY} = \beta + \gamma_1\gamma_2 Var(T)$, so unless γ_1 or γ_2 or Var(T) equal 0, in which case there is no confounding, the association between X and Y is not a reliable guide to the causal effect of X on Y. Because Z (the instrument) is independent of the confounder T in this model and not an effect of X, the association between Z and X and between Z and Y does not involve γ_1 or γ_2 at all: $\rho_{ZX} = \alpha$, $\rho_{ZY} = \alpha\beta$.

Thus, $\beta = \rho_{ZY} / \rho_{ZX}$, and if the model is correctly specified, then the sample correlations rho(Z,Y) and rho(Z,X) are all we need to consistently estimate β, the causal effect of X on Y.

More generally, if we have *t* observations on a vector of *K* regressors **X**, a vector of *K* instruments **Z** (one for each $X \in \mathbf{X}$), and an outcome **Y**, then the linear IV-estimator for the unbiased regression of **X** on Y is: $b^* = (\mathbf{Z'X})^{-1}\mathbf{Z'Y}$, where **Y** is a 1 x *t* matrix, **X** and **Z** are (1+*K*) x *t* matrices with the first column all "1"s, and b^* is the vector of unbiased regression coefficients [Goldberger, 1972].

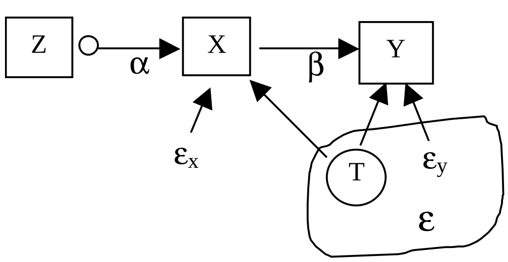

Figure 2: Generalized error term ε

Assuming that T is an aggregated variable that includes all the common causes of X and Y, and that ε is an aggregated variable that includes all the direct causes of Y except for X (Figure 2), the conditions sufficient for b^* to be an unbiased estimator of the causal influence of X on Y are:

1) Y is not a cause of X
2) Z and ε are independent
3) $\rho_{Z,X} \neq 0$
4) All effects are linear functions of their causes and error.

The IV-estimator works because the associations between Z and Y and between Z and X do not involve T, which by assumption we cannot measure. That these associations do not involve T follows from the above conditions and the assumption that causal structures are Markov and faithful [Spirtes, et al., 2000]. The independence of Z and ε entails that Z and T have no causal connection, which in turn entails that X is not a cause of Z, which we represent by drawing Z o→ X in Figure 2. If X was a cause of Z, then T would be an indirect cause of Z and thus not be independent of it. Thus X is a collider on any undirected path between Z and Y going through T.[1] If assumption 4 holds, i.e., the model is linear, then no undirected path between Z and Y going through T produces any association between Z and Y [Spirtes, et al., 2000].

The only other causal feature that is entailed by these assumptions and that is required is that Z and Y are not adjacent. This is a consequence of how ε is defined: "all other direct causes of Y except for X." If Z were a direct cause of Y, then it would be included in ε and thus not independent of ε, contrary to assumption 2. If there was an unmeasured common cause U of Z and Y, then U would be included in ε and then Z and ε would not be independent. Finally, if Y were a direct cause of Z, then ε would be an indirect cause of Z and thus not independent of it.

The fourth assumption, linearity, can be relaxed to exclude the unmeasured confounder T. That is, as the model is usually given:

$$X = \alpha Z + \gamma_1 T + \varepsilon_x$$
$$Y = \beta X + \gamma_2 T + \varepsilon_y$$

so X is a linear function of Z, T, and error , and Y is a linear function of X, T and error. It turns out that neither X nor Y must depend linearly on T. That is, for arbitrary functions f and g, if:

$$X = \alpha Z + f(T) + \varepsilon_x$$
$$Y = \beta X + g(T) + \varepsilon_y$$

[1] If the terms "collider" and "undirected path," or "Markov" and "faithful" are not familiar, see [Spirtes, et al., 2000], chapter 2 and 3.

then ρ_{ZY} / ρ_{ZX} is still a consistent IV estimator of β.[2] This is an important generalization, because we can examine the data to confirm that X depends linearly upon Z, we cannot inspect X's dependence on T. Similarly for Y's dependence on X and T.

3 PIECEWISE LINEAR IV ESTIMATION

Clearly the consistency of the IV estimator still depends on linearity in the remaining part of the model. That is, X must depend linearly on Z, and Y must depend linearly on X. Our idea is to take advantage of the fact that the functional form of the dependence of X and Y on T, aside from being additive with the other causes, is irrelevant - and to break up the X,Y,Z 3-space into regions[3] such that in each region the dependencies among X, Y, and Z *are* approximately linear.

Our strategy is a simple extension of piecewise linear regression (Figure 3), where one partitions the X,Y space with cuts in X such that the dependence of Y on X is linear within each partition. Our strategy is to search for partitions of the X,Y,Z-space defined by cuts in Z and in X such that X is a linear function of Z, and Y a linear function of X within each partition.

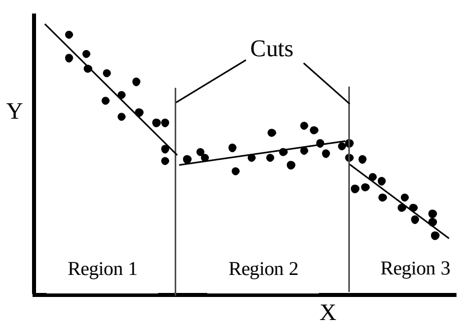

Figure 3: Piecewise linear regression

Since the dependence of X on T and of Y on T need not be linear, or even the same function, in each of these regions, the IV estimator can be applied separately in each region.

Finding the "best" set of regions is a classic problem in AI and Statistics. The smaller we make each region, the more approximately linear the relations within the region, but the smaller the sample size and the worse the statistical properties of the IV estimator in the region. Thus, the problem is how to best trade off linearity for sample size. Put another way, the complexity of the model increases with the number of regions. The trade off is thus between the complexity of the model and its fit overall.

Fortunately, this problem and others like it have been investigated by statisticians and econometricians under the nomenclature: "Change Point Analysis."[4] Assuming that there is a partition of the independent variable, possibly empty, such that each region in the partition contains at least 5 points and the dependent variable is a linear function of the independent variable plus noise in each region, Chen and Gupta [2000] give a consistent $O(N^2)$ procedure, N the sample size, for locating the change points that partition the independent variable.

If we order the sample points k=1 to n according to the independent variable Z in the regression of X on Z, for example, then the null hypothesis of no change points is:

$$H_0: \mu_{Xi} = \beta_0 + \beta_1 Z_i, \text{ for } i = 1 \text{ to } n$$

An alternative to the null is that there is one change point:

$$H_{alt}: \quad \mu_{Xi} = \beta^1{}_0 + \beta^1{}_1 Z_i, \text{ for } i = 1 \text{ to } k, \text{ and}$$
$$\mu_{Xi} = \beta^*{}_0 + \beta^*{}_1 Z_i, \text{ for } i = k+1 \text{ to } n$$

where $\beta^1{}_0 \neq \beta^*{}_0$ and $\beta^1{}_1 \neq \beta^*{}_1$.

Chen and Gupta compare H_0 and H_{alt} using the Schwarz Information Criteria (SIC).[5] $SIC(H_0)$ is computed directly from a regression of Y on X:

$$SIC(H_0) = n \log(2\pi) + n \log[\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2] + n + 3 \log(n) - n \log(n)$$

while the SIC for H_{alt} at each change point k is:

$$SIC(k) = n \log(2\pi) + n \log[\sum_{i=1}^{k} (Y_i - b^1{}_0 - b^1{}_1 X_i)^2 + \sum_{i=k+1}^{n} (Y_i - b^*{}_0 - b^*{}_1 X_i)^2] + n + 5 \log(n) - n \log(n)$$

Accept H_{alt} if, over k between 2 and n-2, $\min[SIC(k)] < SIC(H_0)$, otherwise accept the null H_0. Vostrikova (1981) proved that repeating this procedure recursively

[2] We include a proof in the Appendix.

[4] We thank Teddy Seidenfeld for pointing us to this literature.

[5] The formulas for $SIC(H_0)$ and SIC(k) given here are slight corrections of those given in Chen and Gupta [2000], pp. 113 and 114. Changwon Yoo found and corrected the errors, and Drs. Chen and Gupta confirmed Yoo's version as correct.

within the two regions defined by the change point until the null hypothesis is accepted within each region is a consistent $O(N^2)$ procedure for detecting the number of change points and their location simultaneously.

The overall piecewise linear IV-estimation procedure, **PL-IV**, is:

1. Use SIC to find the change-points in Z such that X's dependence on Z is approximately piece-wise linear.
2. Use SIC to find the change-points in X such that Y's dependence on X is approximately piece-wise linear.
3. For each region R(Z,X) defined by the change-points in Z and X, use only sample points in the region to estimate: $\mathbf{IV}_{\mathbf{R(Z,X)}} = (Z'X)^{-1}(Z'Y)$.

A slight modification of this procedure, which we will call **PL-IV*** is to leave out step 1 from **PL-IV**:

1. Use SIC to find the change-points in X such that Y's dependence on X is approximately piece-wise linear.
2. For each region R(X) defined by the change-points in X, use only sample points in the region to estimate: $\mathbf{IV}_{\mathbf{R(Z,X)}} = (Z'X)^{-1}(Z'Y)$.

4 CAUSAL PREDICTION

In our setting, the causal effect of X on Y must be estimated from training data generated by an instantiation of the causal model in Figure 1.

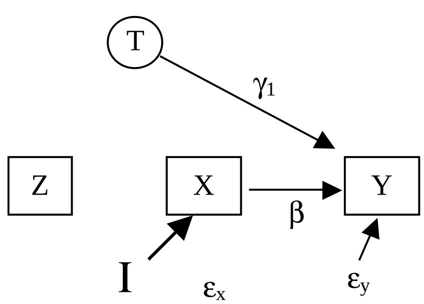

Figure 4: Ideal Intervention on X

We will consider two test contexts for causal prediction, one in which we choose a value for X at random and ideally manipulate X to take on that value, and another in which we first observe X and then randomly choose the value of some variable Δ, which we then add to the observed value of X.

In the first context, the intervention completely determines the value of X, and thus breaks the arrows into X (Figure 4), making X independent of both Z and T in the test context.

In the second the context, we do not ideally intervene on X, but only contribute some value Δ to the X observed. For example, if X was monthly income, we might run a study in which, for each subject, we picked a number between 0 and 2,000, and then added that number of dollars to a subject's monthly income.

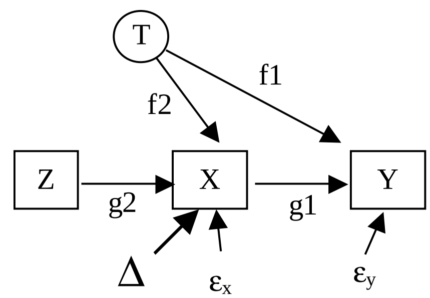

Figure 5: Changing X by Δ

In this context we do not *determine* the subject's monthly income - it is still a function of the same variables it was before our intervention - rather we contribute an extra influence Δ to it, where Δ is independent of T and Z. Thus the causal model describing this test context is Figure 5, not Figure 4. The important difference is that in the ideal intervention context of Figure 4, the X after manipulation is independent of Z and T, but in Figure 5 it is still associated with Z and T, although because of Δ not *as* associated as it was in the training context.

5 SIMULATION STUDIES

To test the performance of the estimators we have defined, we conducted a simulation study in which we generated training and both kinds of test data (Figure 4 and Figure 5) for four versions of the causal model in Figure 1.

For each version of the model in Figure 1, given below, we drew a pseudo-random sample of 2,000 training points and 500 test points. The versions differ on whether X is a linear function of Z or not, and whether Y is a linear function of X or not. Dataset 1 is from a fully linear model, dataset 2 from a model in which Y is not a linear function of X, dataset 3 from a model in which X is not a linear function of Z, and dataset 4 from a model in which Y is not a linear function of X and X is not a linear function of Z.

Dataset 1 (no-cp)

Training Data (N=2,000)
$P(Z,T,\varepsilon_x, \varepsilon_y)$ ~ normal, with means and off-diagonal covariances = 0, Var(Z) = Var(T) = 1.0, and $Var(\varepsilon_x) = Var(\varepsilon_y) = 0.5$.

$X = 0.6*Z + 0.7*T + \varepsilon_x$
$Y = 0.8*X + 0.7*T + \varepsilon_Y$

Test Data (M=500)
$Z_t, X_t, T_t, \varepsilon_{xt}, \varepsilon_{yt}$ distributed as above.
$X_{close} = X_t = X_t + \Delta$, Δ ~U[-1/2 sd(X), +1/2sd(X)] ♣
X_{far} ~ U[$X_{.05}$, $X_{.95}$] ♠
$Y_{close} = 0.8*X_{close} + 0.7*T_t + \varepsilon_{Yt}$
$Y_{far} = 0.8*X_{far} + 0.7*T_t + \varepsilon_{Yt}$

The variable X_{far} captures the test context of an ideal intervention (Figure 4). We *set* X_{far} to the draw from a uniform with width equal to the central 90% of the sample X data. We subscript it with "far" to indicate it is probably far from the X we would have observed if we had not intervened.

The variable X_{close} captures the test context of Figure 5. We drew Δ from a relatively narrow uniform and *added* it to the X observed in the test data, thus creating an X that is probably "close" to the X we would have observed if we had not intervened.

Datasets 2 through 4 are generated similarly to 1, but with different functional dependences between X, Y, and Z.

Dataset 2 (X-cp)

$$X = 0.6*Z + 0.7*T + \varepsilon_x$$
$$Y = \begin{cases} 0.8*X + 0.7*T + \varepsilon_Y & \text{if } X < 0, \\ -0.8*X + 0.7*T + \varepsilon_Y & \text{if } X \geq 0 \end{cases}$$

Dataset 3 (Z-cp)

$$X = \begin{cases} -0.6*Z + 0.7*T + \varepsilon_x & \text{if } Z < 0, \\ 1.0*Z + 0.7*T + \varepsilon_x & \text{if } Z \geq 0 \end{cases}$$
$$Y = 0.8*X + 0.7*T + \varepsilon_Y$$

Dataset 4 (X,Z -cp)

$$X = \begin{cases} 0.6*Z + 0.7*T + \varepsilon_x & \text{if } Z < 0, \\ 1.6*Z + 0.7*T + \varepsilon_x & \text{if } Z \geq 0 \end{cases}$$
$$Y = \begin{cases} 0.8*X + 0.7*T + \varepsilon_Y & \text{if } X < 0, \\ -0.8*X + 0.7*T + \varepsilon_Y & \text{if } X \geq 0 \end{cases}$$

For each of these datasets, we computed the mean-squared error for each of the following estimators of the causal effect of X on Y:

- **Reg**: multiple regression of Y on X and Z.
- **IV**: standard linear IV-estimator
- **PL-IV**: piecewise linear IV-estimator with cuts in Z and X.
- **PL-IV*:** piecewise linear IV-estimator with cuts only in X.

For each estimator, we calculated the mean-squared error on both the far and the close data. For example, for the standard linear IV-estimator, we computed:

$$MSE_{close}(IV) = \frac{\sum_{i=1 \text{ to } M} (\hat{Y}_{IV}(X_{close}) - Y_{close})^2}{M}$$

$$MSE_{far}(IV) = \frac{\sum_{i=1 \text{ to } M} (\hat{Y}_{IV}(X_{far}) - Y_{far})^2}{M}$$

where $\hat{Y}_{IV}(X)$ is the predicted value of Y for the post-manipulation value of X, and M the sample size of the test data.

6 RESULTS

Table 1 gives the mean-squared errors for each of the estimators on the test context from Figure 4, in which we ideally intervene on X, and Table 2 the test context from Figure 5, in which we contribute Δ to X but do not ideally intervene on it. The parenthetical comments in the rows refer to whether the data generating process involved change-points in X, Z, or both. The parenthetical comments on the columns refer to whether the estimator involved finding change-points in X, Z, or both.

Consider the first rows in both tables. In this row, the mean-squared error for the IV estimators within each table is the same because the estimator found no change-points and is thus equivalent to standard IV estimation. Interestingly, in the ideal intervention test context (table 1), IV estimation outperforms regular regression, but the opposite is true in the non-ideal intervention case (table 2). This is because the manipulated X and Z are independent in the data from table 1, but still dependent in table 2. If the joint distribution in the training and test set over X, Y, and Z are identical, and the model is linear, then regression will outperform any estimator. It is only when the test distribution is different than the training distribution that IV estimation offers an advantage. In table 2, the test distribution is not identical to the training because

♣sd(X) is the sample standard deviation of X in the training data.
♠$X_{.05}$ is the value of X that is greater than 5% of the training sample.

of Δ, but it is apparently close enough to give naïve regression the advantage in table 2.

Table 1: MSE_{far}

Dataset	Estimator Reg	 IV (no-cp)	 PL-IV* (X-cp)	 PL-IV (X,Z-cp)
1 (no-cp)	1.292	1.052	1.052	1.052
2 (X-cp)	1.242	1.162	1.015	1.015
3 (Z-cp)	1.189	0.970	0.970	1.111
4 (X,Z - cp)	1.496	1.316	1.063	1.328

Table 2: MSE_{close}

Dataset	Estimator Reg	 IV (no-cp)	 PL-IV* (X-cp)	 PL-IV (X,Z-cp)
1 (no-cp)	0.812	0.978	0.978	0.978
2 (X-cp)	1.268	1.589	0.930	0.930
3 (Z-cp)	0.830	0.995	0.995	0.920
4 (X,Z - cp)	1.486	1.826	0.941	0.964

The second rows in both tables correspond to a data generating process in which Y is non-linear in X, but X still linear in Z. As expected, in table 1 IV estimation outperforms naïve regression, and piecewise linear IV estimation outperforms standard IV estimation. In table 2, the PL-IV estimator affords a large advantage over both naïve regression and standard IV.

In the third and fourth rows, where X depends non-linearly on Z, the relative advantage of PL-IV vs. PL-IV* is made apparent. In the ideal intervention context of table 1, PL-IV* seems to dominate. This is because the manipulated value of X is independent of Z, and the functional dependence of Y on X (what we are trying to estimate) is also independent of Z. Using cuts in Z in the estimation reduces the sample size of the regions, and trying to use Z in the prediction introduces noise. In the third row of table 2, PL-IV slightly outperforms PL-IV*, and this seems to be because Z and the function connecting X to Y are still associated in the test context, and in a similar way to how they are associated in the training context.

Overall, PL-IV* outperforms all other estimators in table 1, where we are predicting the effect of an ideal intervention. In table 2, where we are predicting a change in X that is not an ideal intervention, the results depend on whether the functional dependence of Y on X is non-linear. If it is, then PL-IV* again dramatically outperforms both regular regression and standard IV estimation.

Acknowledgements

Research support for Greg Cooper was provided in part by grants R01-LM06696 and R01-06759 from the National Library of Medicine and by grant IIS-9812021 from the National Science Foundation.

References

Bowden, R. and Turkington, D. (1984). *Instrumental variables*. Cambridge University Press, NY

Chen, J., and Gupta, A. K. (2000). *Parametric Statistical Change Point Analysis*. Birkhauser, Boston.

Goldberger, A. (1972). *Econometric Theory*. Wiley.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press.

Rosenbaum, P. (1995) *Observational Studies*, Springer Series in Statistics, New York.

Spirtes, P. (1999). The Limits of Causal Inference from Observational Data. presentation to the American Economic Association Meetings, Boston. see: hss.cmu.edu/philosophy/people/directory/Peter_Spirtes.html

Spirtes, P., Cooper, G. (1999). An Experiment in Causal Discovery Using a Pneumonia Database, *Proceedings of AI and Statistics 99*.

Sprites, P., Glymour, C., Scheines, R. (2000). *Causation, Prediction, and Search, 2nd Edition*. MIT Press.

Vostrikov, J. (1981). Detecting "disorder" in multidimensional random processes. *Soviet Mathematics Doklady*, 24, 55-59.

Appendix

Theorem: If Z, T, ε_y, and ε_x are pairwise independent, Z, X, Y are normal with mean 0 and variance 1, and the structural equations defining the causal dependence among X, Z and T are:

$$Y = \beta X + g(T) + \varepsilon_y$$
$$X = \alpha Z + f(T) + \varepsilon_x$$

then ρ_{ZY} / ρ_{ZX} is still a consistent IV estimate of β regardless of the form of functions f and g.

Proof

1) $ZY = Z(\beta[\alpha Z + f(T) + \varepsilon_x] + g(T) + \varepsilon_y)$

2) $\quad = \alpha\beta Z^2 + Z f(T) + Z\varepsilon_x + Zg(T) + Z\varepsilon_y$

3) $E(ZY) = \alpha\beta E(Z^2) + E(Zf(T)) + E(Z\varepsilon_x) + E(Zg(T)) + E(Z\varepsilon_y)$

4) $E(Zf(T)) = E(Z\varepsilon_x) = E(Zg(T)) = E(Z\varepsilon_y) = 0$, [if A and B are independent, so are any functions of A and B.]

5) Since $E(Z^2) = Var(Z) = 1$, $\quad E(ZY) = \rho_{ZY} = \alpha\beta$

6) $ZX = Z(\alpha Z + f(T) + \varepsilon_x) = \alpha Z^2 + Z f(T) + Z\varepsilon_x$

8) $E(ZX) = \alpha\, Var(Z) = \rho_{ZX} = \alpha$

9) $\rho_{ZY} / \rho_{ZX} = \beta$