PLSC 503: "Multivariate Analysis for Political Research"

Multivariate Regression, I
February 14, 2017

Introduction

Multivariate least-squares linear regression is the basis for most of the "regression-type" methods in use today. Over the next few days, we'll effectively review what we've been doing for the past two weeks, but in a context in which there are multiple covariates / "independent variables." The basic model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1}$$

here:

- $\mathbf{Y}$ denotes the $N \times 1$ vector containing the response / "dependent" variable,
- $\mathbf{X}$ is a $N \times K$ matrix of covariates,
- $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters / coefficients, and
- $\mathbf{u}$ is a $N \times 1$ vector of disturbances / "errors."

We'll continue to denote $i \in \{1, ...N\}$ to index the observations, and we'll use $k \in \{1, ...K\}$ to denote the individual covariates, with K the rank of the covariate matrix (that is, the number of covariates in the model, including the constant term). This means that:

- Y_i is a scalar indicating the value of $\mathbf{Y}$ for observation i.
- $\mathbf{X}_i$ is a $1 \times K$ vector containing the K values of the independent variables $\{X_0, ...X_K\}$ for observation i, and X_{ki} is a scalar containing the value of X_k for observation i.
- β_k is a scalar containing the coefficient associated with covariate $\mathbf{X}_k$, and
- u_i is a scalar containing the value of the disturbance term for observation i.

If we "write this out" in scalar form, the model for a single observation i looks like this:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K X_{Ki} + u_i \tag{2}$$

and the full model looks like:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \cdots & X_{KN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}. \tag{3}$$

Diversion: Added Variable Plots

Weisberg discusses "added variable plots," which are a (sometimes) useful way of understanding visually how multiple variables $\mathbf{X}$ are related to Y. The basic idea is that, for a simple model

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i,$$

adding a second variable X_2 is designed to "explain" the part of Y not "explained" by X_1, *after accounting for the association between* X_1 *and* X_2. To do such a plot "by hand," we

1. Regress Y on X_1 and save the residuals $\hat{u}_i$,
2. Regress X_2 on X_1 and save the residuals (call these $\hat{e}_i$),
3. Plot $\hat{u}_i$ (conventionally on the y-axis) vs. $\hat{e}_i$ (conventionally on the x-axis).

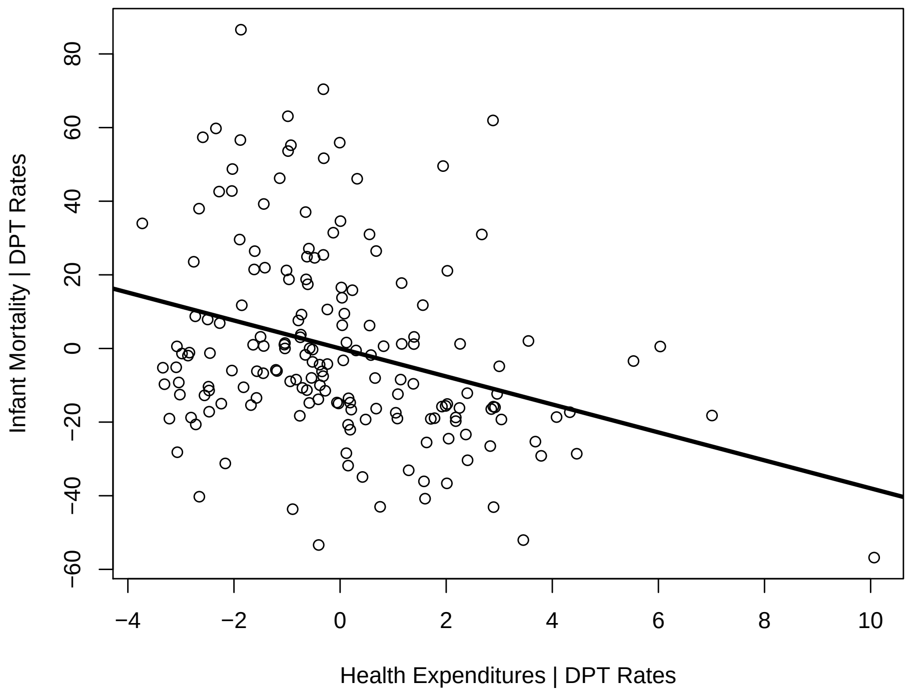
Intuitively, this yields a plot of the "part of Y unexplained by X_1" against the "part of X_2 unexplained by X_1; from it, we can get an idea of whether and how Y varies with X_2, *holding* X_1 *constant.*

An easy illustration uses the infant mortality data, and considers adding a second predictor, `healthexpGDP`, an indicator of *health expenditures* as a fraction of GDP. That plot is in Figure 1 below; the code for it is in the Appendix to these notes. Note a few things:

- Because both variables are regressions residuals, they both have means of zero. As a result,
- The resulting regression line (shown) also has an intercept of zero, and
- The slope of that regression line is exactly the same as the slope of the line we would get from estimating a model with *both* `DPTpct` and `healthexpGDP` on the right-hand side.

Note that we can use added variable plots with any number of right-hand-side variables. So, for example, to generate such a plot for a variable X_1 in a model with k right-hand-side covariates, we'd regress Y and X_1 on $X_2, X_3, ...X_k$, generate residuals from both regressions, and plot them against each other. The `avPlots` routine in the `car` package is a convenient way to do these plots.

Figure 1: Added Variable Plot: Infant Mortality and Health Expenditures Given DPT Immunization Rates



Estimation of $\hat{\boldsymbol{\beta}}$

As was the case before, the main thing we are interested in doing is estimating the parameters $\boldsymbol{\beta}$. We do so in exactly the same way as we did before: that is, by choosing a set of parameters that minimize the sum of the squared errors.

We'll start off as we did before, by rewriting (1) in terms of the errors:

$$\mathbf{u} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \tag{4}$$

The "square" of each element u_i of $\mathbf{u}$ is simply the value of u_i times itself; to multiply each element of $\mathbf{u}$ by itself, we can take the inner product of $\mathbf{u}$:

$$
\begin{aligned}
\mathbf{u}'\mathbf{u} &= \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \\
&= u_1^2 + u_2^2 + \ldots + u_N^2 \\
&= \sum_{i=1}^{N} u_i^2 \qquad (5)
\end{aligned}
$$

Now, thanks to (4), we can further rewrite (5) as:

$$
\begin{aligned}
\mathbf{u}'\mathbf{u} &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}' + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \qquad (6)
\end{aligned}
$$

where the last equality uses some simple linear algebra, including the fact that

- $(\mathbf{X}\boldsymbol{\beta})' = \boldsymbol{\beta}'\mathbf{X}'$, and
- because $\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}$ is a scalar, its is equal to its transpose $\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}$.

As before, the idea is to pick $\boldsymbol{\beta}$ so as to make (6) as small as possible. And, as before, the most straightforward way to do this is to use a little differential calculus. To that end, we first have to consider the first derivative of (6) with respect to $\boldsymbol{\beta}$:

$$
\frac{\partial \mathbf{u}'\mathbf{u}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \qquad (7)
$$

We can then set this equal to zero, and solve for $\boldsymbol{\beta}$. We'll do this in two parts. First, a bit of simple matrix algebra:

$$
\begin{aligned}
-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= 0 \\
-\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= 0 \\
\mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}'\mathbf{Y}
\end{aligned}
$$

This essentially says that the variability in $\mathbf{X}$ times $\boldsymbol{\beta}$ is equal to the covariation in $\mathbf{X}$ and $\mathbf{Y}$ (sound familiar?...). Now, in order to solve for $\boldsymbol{\beta}$, we need to "get rid" of the $\mathbf{X}'\mathbf{X}$ term. We can do this by premultiplying it times its inverse:

$$
\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
\mathbf{I}\boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
\boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (8)
\end{aligned}
$$

This is the fundamental OLS result, in matrix format.

Note that this looks an awful lot like the result in non-matrix form: $\beta = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$. In particular,

- If we think of $\mathbf{X}'\mathbf{Y}$ as the covariance of $\mathbf{X}$ and $\mathbf{Y}$, and
- $\mathbf{X}'\mathbf{X}$ as the variance of $\mathbf{X}$, then
- Premultiplying $\mathbf{X}'\mathbf{Y}$ by the inverse of $\mathbf{X}'\mathbf{X}$ is like "dividing" $\mathbf{X}'\mathbf{Y}$ by $\mathbf{X}'\mathbf{X}$

OLS Assumptions

We didn't go into a lot of detail about the assumptions of the "classical linear regression model" (CLRM) before, but it probably is worth doing so a bit more now. There are five critical assumptions, which we'll consider in turn.

1. Zero Expectation Disturbances

The first assumption is:

$$\mathrm{E}(\mathbf{u}) = \mathbf{0} \tag{9}$$

This states that the expected value of the vector of disturbances is a vector of zeros. It simply says that the expected value of each element of $\mathbf{u}$ is zero:

$$\begin{aligned}
\mathrm{E}(\mathbf{u}) &= \mathrm{E}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
&= \mathbf{0}
\end{aligned}$$

You can think of this as a sort of "necessary" condition for a good estimator – if the expectation of the errors are anything other than zero, that suggests that we can necessarily "do better" (i.e., reduce the magnitude of the errors). It's also a necessary condition for the unbiasedness of the estimator, for reasons that are (or ought to be) obvious.

2. Homoscedasticity / No Error Correlation

The second critical assumption can be written in terms of the "outer product" of the **u** matrix, as:

$$\mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I} \tag{10}$$

where σ^2 is constant $\forall i$ and $\mathbf{I}$ is an $N \times N$ identity matrix. This assumption actually encompasses two things:

1. *Homoscedasticity* (that is, constant error variance), and
2. *No Residual Autocorrelation* (that is, the covariances of the errors are all zero).

To get at this a bit more clearly, first "write out" $\mathbf{u}\mathbf{u}'$:

$$\begin{aligned}\mathbf{u}\mathbf{u}' &= \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix} \\ &= \begin{bmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_N \\ u_2u_1 & u_2^2 & \cdots & u_2u_N \\ \vdots & \vdots & \ddots & \vdots \\ u_Nu_1 & u_Nu_2 & \cdots & u_N^2 \end{bmatrix}\end{aligned} \tag{11}$$

This is often termed the matrix of "cross products" of **u**:

- Along the main diagonal are the squared errors u_i^2, which we can think of as the "variances" of the disturbances.
- Off the main diagonal are the cross-products of the errors $u_j u_\ell$, $j \neq \ell$; think of these as the "covariances" of the errors between observation j and ℓ.

The CLRM assumptions require that the errors be both uncorrelated and homoscedastic. This means that, in expectation, $\mathbf{u}\mathbf{u}'$ is required to look like:

$$\mathrm{E}(\mathbf{u}\mathbf{u}') = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \tag{12}$$

which can be written, as above, as the product of a constant σ^2 and an $N \times N$ identity matrix.

3. Fixed X

The CLRM requires that the $\mathbf{X}$s are "fixed in repeated sampling." This simply means that the $\mathbf{X}$s are not random variables – i.e., that they do not have a stochastic component. At first, this assumption seems strange to most people; and, in social-scientific settings, it is in fact more than a bit odd. In practice, however, it means that we can treat the $\mathbf{X}$s as constants in our equations, and implies two important things:

- That there is no *measurement error* in the $\mathbf{X}$s, and
- That $\text{Cov}(\mathbf{X}, \mathbf{u}) = \mathbf{0}$; that is, that there is no *model misspecification*, including no *endogeneity* in the $\mathbf{X}$s.

We'll talk about each of these a bit later on in the course.

4. No Perfect Multicollinearity

The CLRM requires that $\mathbf{X}$ be of "full column rank;" that is:

- that the rank of the $\mathbf{X}$ matrix be equal to the number of columns K (that is, the number of covariates, including the constant term) in $\mathbf{X}$, and
- that the rank K be less than the number of observations N.

This essentially means that there is no exact linear relationship among the variables in $\mathbf{X}$. It is a necessary condition for $\mathbf{X}$ to have a nonzero determinant, and thus to be invertible. Consider again:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

At a minimum, we can't calculate $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ if we can't invert $\mathbf{X}'\mathbf{X}$.

5. Normal Disturbances

For hypothesis testing, the CLRM requires that:

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{13}$$

that is, that the disturbances are distributed according to a multivariate normal distribution with mean zero (cf. assumption one) and variance σ^2 (cf. assumption two).

Under all these assumptions, the estimate obtained $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$:

- Is a linear function of $\mathbf{X}$,
- Is an unbiased estimate of the population parameter $\boldsymbol{\beta}$, and
- Is efficient (that is, has the smallest variance of all linear estimators) – in other words, it is BLUE.

We'll consider the efficiency of $\hat{\boldsymbol{\beta}}$ on Thursday; for now, let's focus on...

Unbiasedness of $\hat{\boldsymbol{\beta}}$

Recall that our model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\boldsymbol{\beta}$ is the population / "true" parameter we're after. In our equation for $\hat{\boldsymbol{\beta}}$, we can substitute this in for $\mathbf{Y}$, and see that:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \end{aligned} \tag{14}$$

and so:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{15}$$

This means that the difference between the estimate and the "true" value $\hat{\boldsymbol{\beta}}$ is equal to the covariance in $\mathbf{X}$ and $\mathbf{u}$, "divided by" the variability in $\mathbf{X}$...

- Since we assume that $\text{Cov}(\mathbf{X}, \mathbf{u}) = \mathbf{0}$, it is clear that $\text{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and therefore that the estimator is unbiased.
- As we'll see Thursday, this also allows us to derive the variances and covariances of the $\boldsymbol{\beta}$s.

A Quick Example

Consider the following two-variable regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where the data matrices are:

$$\mathbf{Y} = \begin{bmatrix} 4 \\ -2 \\ 9 \\ -5 \end{bmatrix} \tag{16}$$

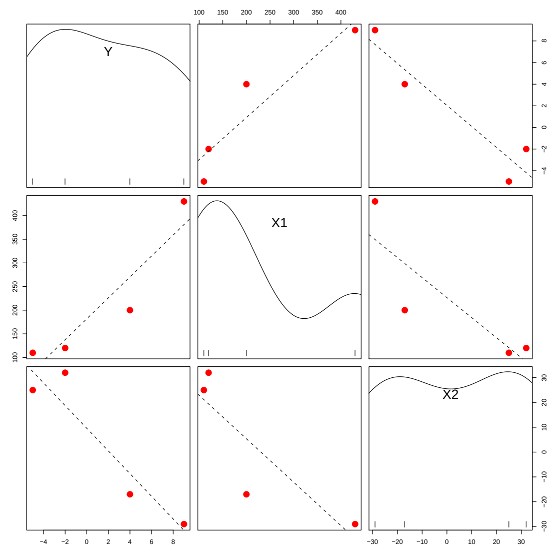
and:

$$\mathbf{X} = \begin{bmatrix} 1 & 200 & -17 \\ 1 & 120 & 32 \\ 1 & 430 & -29 \\ 1 & 110 & 25 \end{bmatrix} \tag{17}$$

The data look like this:

```
> Y<-c(4,-2,9,-5)
> X1<-c(200,120,430,110)
> X2<-c(-17,32,-29,25)
> scatterplot.matrix(~Y+X1+X2,smooth=FALSE,cex=2,pch=16)
```

Figure 2: Scatterplot Matrix of Y, X_1, and X_2



and they are correlated as:

```
> data<-cbind(Y,X1,X2)
> cor(data)
         Y      X1      X2
Y   1.0000  0.9285 -0.9425
X1  0.9285  1.0000 -0.8613
X2 -0.9425 -0.8613  1.0000
```

Now, let's estimate $\hat{\boldsymbol{\beta}}$. Remember that the formula is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. So, first we need to calculate $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$; those are equal to:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 860 & 11 \\ 860 & 251400 & -9280 \\ 11 & -9280 & 2779 \end{bmatrix} \tag{18}$$

and:

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 6 \\ 3880 \\ 518 \end{bmatrix} \tag{19}$$

Remember that $\mathbf{X}'\mathbf{X}$ is the variance-covariance matrix of $\mathbf{X}$, and $\mathbf{X}'\mathbf{Y}$ is the covariance of $\mathbf{X}$ and $\mathbf{Y}$.

Next, we need to invert $\mathbf{X}'\mathbf{X}$. We could do this "by hand," but since we all know how to do this, I'll just tell you that $|\mathbf{X}'\mathbf{X}|$ is equal to a very large, positive number (something on the order of 1.887×10^8), and

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 3.2453 & -0.0132 & -0.05694 \\ -0.0132 & 0.000058 & 0.0002468 \\ -0.0569 & 0.000247 & 0.001409 \end{bmatrix} \tag{20}$$

Doing the multiplication, we get:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{bmatrix} 3.2453 & -0.0132 & -0.05694 \\ -0.0132 & 0.000058 & 0.0002468 \\ -0.0569 & 0.000247 & 0.001409 \end{bmatrix} \begin{bmatrix} 6 \\ 3880 \\ 518 \end{bmatrix} \\ &= \begin{bmatrix} -2.264 \\ 0.0190 \\ -0.1141 \end{bmatrix} \end{aligned} \tag{21}$$

Now, compare this to the R regression output...

```
> fit<-lm(Y~X1+X2)
> summary(fit)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
     1      2      3      4
 0.531  1.639 -0.201 -1.970

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.2643     4.7284   -0.48     0.72
X1            0.0190     0.0200    0.95     0.52
X2           -0.1141     0.0985   -1.16     0.45

Residual standard error: 2.62 on 1 degrees of freedom
Multiple R-Squared: 0.941,Adjusted R-squared: 0.823
F-statistic: 7.99 on 2 and 1 DF,  p-value: 0.243
```

Viola!...

Estimation Issues

Weisberg (p. 61) says "Do not compute the least squares estimates using (21)!" His concerns stem from the fact that using what he terms "uncorrected" sums of squares and cross-products will lead to rounding error.

This is a fair point, and in fact most software (including R), replaces $\mathbf{X}$ with a QR decomposition

$$\mathbf{X} = \mathbf{QR}$$

where $\mathbf{Q}$ is an orthogonal matrix ($\mathbf{Q}'\mathbf{Q} = \mathbf{I}$) and $\mathbf{R}$ is an upper-triangular matrix.

The details of this are not terribly important. What is important is that Weisberg is right, and it's easy to show that. For example, consider these "data":

```
options(digits=16)
options(scipen=99)
z<-c(-1000000000000,0.000000000000001,1000000000000)
x<-c(-50000,0.000007,5000000)
lm(z~x)

Call:
lm(formula = z ~ x)

Coefficients:
          (Intercept)                      x
-494950994952.3740845         299970.2999707
```

Now do the same regression "by hand," using the formula in (21):

```
X<-as.matrix(x)
Z<-as.matrix(z)
beta.hat <- solve(t(X) %*% X) %*% t(X) %*% Z
beta.hat
                 [,1]
[1,] 201979.802019798
```

The difference is large; the former estimate is $\frac{299970.2999707}{201979.802019798} \times 100 = 148.515$ percent of the latter in size.

Tuesday: Hypothesis testing and inference in multivariate regression...

Appendix: R code for added variable plots

```
library(foreign)

Data<-read.dta("CountryData2000.dta")
Data<-na.omit(Data[c("infantmortalityperK","DPTpct","healthexpGDP")])

fit<-lm(infantmortalityperK~DPTpct,data=Data)
aux<-lm(healthexpGDP~DPTpct,data=Data)
plot(aux$residuals,fit$residuals,xlab="Health Expenditures | DPT Rates",
     ylab="Infant Mortality | DPT Rates")
abline(lm(fit$residuals~aux$residuals),lwd=3)

# Using avPlots from car:

library(car)

fit2<-lm(infantmortalityperK~DPTpct+healthexpGDP,data=Data)
avPlots(fit2,~healthexpGDP)
```