

MULTIVARIATE LINEAR REGRESSION

Herman J. Bierens

Pennsylvania State University

November 20, 2008

1. Missing variables

Suppose you assume that the relationship between a dependent variable Y_j and an explanatory variable X_j for observations $j = 1, \dots, n$ is given by

$$Y_j = \alpha + \beta \cdot X_j + U_j, \quad j = 1, 2, \dots, n, \quad (1)$$

whereas in reality

$$Y_j = \alpha + \beta \cdot X_j + \gamma \cdot Z_j + U_j, \quad j = 1, 2, \dots, n, \quad (2)$$

where Z_j is a missing explanatory variable, and U_j is the error term. Recall that the OLS estimator of β , assuming that model (1) is correct, is

$$\hat{\beta} = \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} = \frac{\sum_{j=1}^n (X_j - \bar{X})Y_j}{\sum_{j=1}^n (X_j - \bar{X})^2}. \quad (3)$$

Substituting (2) in (3) yields

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{j=1}^n (X_j - \bar{X})(\alpha + \beta \cdot X_j + \gamma \cdot Z_j + U_j)}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \alpha \frac{\sum_{j=1}^n (X_j - \bar{X})}{\sum_{j=1}^n (X_j - \bar{X})^2} + \beta \frac{\sum_{j=1}^n (X_j - \bar{X})X_j}{\sum_{j=1}^n (X_j - \bar{X})^2} + \gamma \frac{\sum_{j=1}^n (X_j - \bar{X})Z_j}{\sum_{j=1}^n (X_j - \bar{X})^2} + \frac{\sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \beta + \gamma \frac{\sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})}{\sum_{j=1}^n (X_j - \bar{X})^2} + \frac{\sum_{j=1}^n (X_j - \bar{X})U_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \end{aligned} \quad (4)$$

where \bar{Z} is the sample mean of the Z_j 's. Note that the last equality in (4) follows from the fact that $\sum_{j=1}^n (X_j - \bar{X}) = 0$. Assuming that the variables involved are independent across the observations j and that model (2) is correct, in the sense that $E[U_j | X_j, Z_j] = 0$, it follows from (4) that

$$\begin{aligned}
E[\hat{\beta}|X_1, \dots, X_n, Z_1, \dots, Z_n] &= \beta + \gamma \frac{\sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})}{\sum_{j=1}^n (X_j - \bar{X})^2} + \frac{\sum_{j=1}^n (X_j - \bar{X}) E[U_j | X_j, Z_j]}{\sum_{j=1}^n (X_j - \bar{X})^2} \\
&= \beta + \gamma \frac{\sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})}{\sum_{j=1}^n (X_j - \bar{X})^2}.
\end{aligned} \tag{5}$$

Therefore, in general,

$$E[\hat{\beta}] = \beta + \gamma \cdot E \left[\frac{\sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})}{\sum_{j=1}^n (X_j - \bar{X})^2} \right] \neq \beta. \tag{6}$$

In other words, the OLS estimator $\hat{\beta}$ of β in model (1) is no longer unbiased, due to the missing explanatory variable Z_j , except if the sample covariance $(1/n)\sum_{j=1}^n (X_j - \bar{X})(Z_j - \bar{Z})$ is exactly zero.

To demonstrate the effect of missing explanatory variables, I have drawn X_j for $j = 1, \dots, n = 500$ independently from the $N(0,1)$ distribution, then generated the Z_j 's by $Z_j = X_j + V_j$ for $j = 1, \dots, 500$, where the V_j 's have been drawn independently from the $N(0,1)$ distribution, and next I have generated the Y_j 's by $Y_j = 1 + X_j + Z_j + U_j$ for $j = 1, \dots, 500$, where the U_j 's have been drawn independently from the $N(0,1)$ distribution. Thus, model (2) is applicable to this artificial data set, with $\alpha = \beta = \gamma = 1$.

The EasyReg output of the regression according to model (1) is:

Dependent variable:

Y (= 1+X+Z+U)

X variables:

X(1) = X

X(2) = 1

Model:

Y = b(1)X(1) + b(2)X(2) + U, where U is the error term satisfying

$E[U|X(1), X(2)] = 0$.

OLS estimation results

Parameters	Estimate	t-value (S.E.) [p-value]	H.C. t-value (H.C. S.E.) [H.C. p-value]
b(1)	2.08383	32.704 (0.06372) [0.00000]	34.040 (0.06122) [0.00000]
b(2)	0.97896	15.475 (0.06326) [0.00000]	15.496 (0.06317) [0.00000]

Notes:

- 1: S.E. = Standard error
- 2: H.C. = Heteroskedasticity Consistent. These t-values and standard errors are based on White's heteroskedasticity consistent variance matrix.
- 3: The two-sided p-values are based on the normal approximation.

Effective sample size (n):	500
Variance of the residuals:	1.996027
Standard error of the residuals (SER):	1.412808
Residual sum of squares (RSS):	994.021241
(Also called SSR = Sum of Squared Residuals)	
Total sum of squares (TSS):	3128.841499
R-square:	0.682304

Breusch-Pagan test = 0.048472

Null hypothesis: The errors are homoskedastic

Null distribution: Chi-square(1)

p-value = 0.82574

Significance levels: 10% 5%

Critical values: 2.71 3.84

Conclusions: accept accept

Thus, the OLS estimator of β in model (1) is $\hat{\beta} = 2.08383$, which is more than 100% larger than the true value 1.

Note that in this case

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma \sqrt{\sum_{j=1}^n w_{1j}^2}} \sim N[0, 1]$$

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$\frac{\hat{\beta}_k - \beta_k}{\sigma \sqrt{\sum_{j=1}^n w_{kj}^2}} \sim N[0, 1]$$
(22)

The error variance σ^2 can be estimated similar to the case of the two-variable linear regression model, namely using the sum of squared residuals

$$SSR = \sum_{j=1}^n \hat{U}_j^2, \tag{23}$$

where

$$\hat{U}_j = Y_j - \sum_{i=1}^k \hat{\beta}_i X_{ij} \tag{24}$$

is the OLS residual.

It can be shown that under Assumptions 1-3,

$$\frac{\sum_{j=1}^n \hat{U}_j^2}{\sigma^2} \sim \chi_{n-k}^2. \tag{25}$$

Because the expected value of a χ_{n-k}^2 distributed random variable is $n-k$, the result (25) suggests to estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{j=1}^n \hat{U}_j^2. \tag{26}$$

Due to (25), this estimator is unbiased: $E[\hat{\sigma}^2] = \sigma^2$. Moreover, it can be shown that under Assumptions 1-3, $\sum_{j=1}^n \hat{U}_j^2$ is independent of the $\hat{\beta}_i$'s, hence it follows from (22) and (25) and the definition of the t distribution that under Assumptions 1 and 2,

$$\begin{aligned}
\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} \sqrt{\sum_{j=1}^n w_{1,j}^2}} &\sim t_{n-k} \\
&\vdots \quad \quad \quad \vdots \\
\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{\sum_{j=1}^n w_{k,j}^2}} &\sim t_{n-k}
\end{aligned} \tag{27}$$

The denominators involved are the standard errors of the corresponding OLS estimators:

$$\hat{\sigma}_i = \hat{\sigma} \sqrt{\sum_{j=1}^n w_{i,j}^2} \quad (= \text{standard error of } \hat{\beta}_i). \tag{28}$$

The results (22), (25) and (27) do not hinge on the assumption that the explanatory variables $X_{i,j}$ are nonrandom, though. They also hold if we replace Assumption 2 by

Assumption 2*: *The model variables $Y_j, X_{1,j}, \dots, X_{k-1,j}$ are independent and identically distributed across the observations $j = 1, \dots, n$,*

and Assumption 3 by

Assumption 3*: *Conditionally on $X_{1,j}, \dots, X_{k-1,j}$ the errors U_j are $N(0, \sigma^2)$ distributed.*

Proposition 1: *Under Assumptions 1, 2* and 3* the results (22), (25) and (27) carry over.*

Furthermore, if instead of Assumption 3*,

Assumption 3**: *$E[U_j | X_{1,j}, \dots, X_{k-1,j}] = 0$, $E[U_j^2 | X_{1,j}, \dots, X_{k-1,j}] = \sigma^2 < \infty$ and $E[X_{i,j}^2] < \infty$ for $i = 1, \dots, k-1$,*

then it can be shown that

Proposition 2: Under Assumptions 1, 2* and 3** ,

$$\begin{aligned} \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\sqrt{\sum_{j=1}^n w_{1,j}^2}} &\sim N(0,1) \\ \vdots &\quad \quad \quad \vdots \\ \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{\sum_{j=1}^n w_{k,j}^2}} &\sim N(0,1) \end{aligned} \tag{29}$$

provided that n is large.

3. Testing parameter hypotheses

The results (27) and (29) can be used to test whether a particular coefficient β_i is zero or not, similar to the case of the two-variable linear regression model. The test statistic involved is the corresponding t-value,

$$\hat{t}_i = \frac{\hat{\beta}_i}{\hat{\sigma}_i} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{\sum_{j=1}^n w_{i,j}^2}}. \tag{30}$$

Proposition 3: Under the null hypothesis $\beta_i = 0$ and the conditions of Proposition 1, $\hat{t}_i \sim t_{n-k}$, and under the null hypothesis involved and the conditions of Proposition 2, $\hat{t}_i \sim N(0,1)$. Moreover, if $\beta_i > 0$ then \hat{t}_i converges in probability to ∞ if $n \rightarrow \infty^1$, and if $\beta_i < 0$ then \hat{t}_i converges in probability to $-\infty$ if $n \rightarrow \infty^2$.

The test can now be conducted in the same way as in the case of the two-variable linear regression model, either left-sided, right-sided or two-sided. The only difference is the degrees of freedom, which is $n-k$ instead of $n-2$ in the two-variable linear regression case.

Now suppose you want to test the joint hypothesis that, for example, $\beta_i = 0$ for $i =$

¹ This means that for any constant $K > 0$, $\lim_{n \rightarrow \infty} P(\hat{t}_i > K) = 1$.

² This means that for any constant $K > 0$, $\lim_{n \rightarrow \infty} P(\hat{t}_i < -K) = 1$.

1,...,m, where $m \leq k - 1$, against the alternative hypothesis that the null hypothesis is false: $\beta_i \neq 0$ for at least one index $i \leq m$. One possible way of testing this hypothesis is to conduct m separate two-sided t tests for $i = 1, \dots, m$. However, the problem is that the left-hand side random variables in (27) are in general not independent, hence under the null hypothesis $\beta_1 = \dots = \beta_m = 0$ the test statistics $\hat{t}_1, \dots, \hat{t}_m$ are in general not independent. In particular, it is impossible to select a critical value t_* such that for a given significance level $\alpha \times 100\%$, $P[|\hat{t}_1| > t_*, |\hat{t}_2| > t_*, \dots, |\hat{t}_m| > t_*] = \alpha$, because we do not know the joint distribution of $\hat{t}_1, \dots, \hat{t}_m$.

The solution of this problem is the following. Consider the restricted regression model

$$Y_j = \beta_{m+1}X_{m+1,j} + \beta_{m+2}X_{m+2,j} + \dots + \beta_{k-1}X_{k-1,j} + \beta_k + U_j, \quad j = 1, 2, \dots, n. \quad (31)$$

Then it can be shown that:

Proposition 4: *Under the null hypothesis $\beta_1 = \dots = \beta_m = 0$ and the conditions of Proposition 2,*

$$\hat{F} = \frac{(SSR_0 - SSR)/m}{SSR/(n-k)} \sim F_{m, n-k}, \quad (32)$$

and under the conditions of Proposition 2,

$$\hat{W} = m \cdot \hat{F} = \frac{SSR_0 - SSR}{SSR/(n-k)} \sim \chi_m^2, \quad (33)$$

where SSR is the sum of squared residuals of the unrestricted model (8) and SSR_0 is the sum of squared residuals of the restricted model (31). Moreover, under the alternative hypothesis that for at least one index $i \leq m$, $\beta_i \neq 0$, the test statistics \hat{F} and \hat{W} converge in probability to ∞ as $n \rightarrow \infty$ ³.

The test based on \hat{F} is called, for obvious reasons, the F test, and the test based on \hat{W} is called the Wald test, named after the statistician with that name who proposed this test. The tests involved are conducted right-sided. In particular in the case of the Wald test the null hypothesis

³ Again, this means that for any constant $K > 0$, $\lim_{n \rightarrow \infty} P(\hat{F} > K) = 1$ and $\lim_{n \rightarrow \infty} P(\hat{W} > K) = 1$.

involved is rejected at say the 5% significance level if $\hat{W} > c$, where the critical value c is chosen such that for a χ_m^2 distributed random variable W , $P[W > c] = 0.05$.

If $m = k - 1$ then the restricted model (31) takes the form

$$Y_j = \beta_k + U_j, \quad j = 1, 2, \dots, n. \quad (34)$$

The sum of squares residuals of this model, SSR_0 , is then equal to the total sum of squares of model (8),

$$TSS = \sum_{j=1}^n (Y_j - \bar{Y})^2, \quad (35)$$

where $\bar{Y} = (1/n)\sum_{j=1}^n Y_j$. The F test involved then has test statistic

$$\tilde{F} = \frac{(TSS - SSR)/(k-1)}{SSR/(n-k)}, \quad (36)$$

which has an $F_{k-1, n-k}$ distribution under the null hypothesis that $\beta_1 = \dots = \beta_{k-1} = 0$ and the conditions of Proposition 2. This test is called the overall F test. Its null hypothesis amounts to the hypothesis that none of the explanatory variables $X_{i,j}$, $i = 1, \dots, k-1$, have an effect on the dependent variable Y_j .

4. The adjusted R^2

The R^2 in the multiple regression case is defined the same as in the two-variable regression case:

$$R^2 = 1 - \frac{SSR}{TSS}. \quad (37)$$

The problem with the R^2 is that it can be inflated towards 1 by including more explanatory variables in the model, because

$$\min_{\hat{\beta}_1, \dots, \hat{\beta}_k} \sum_{j=1}^n (Y_j - \sum_{i=1}^k \hat{\beta}_i X_{i,j})^2 > \min_{\hat{\beta}_1, \dots, \hat{\beta}_k, \hat{\beta}_{k+1}} \sum_{j=1}^n (Y_j - \sum_{i=1}^{k+1} \hat{\beta}_i X_{i,j})^2. \quad (38)$$

The extreme case is where $k = n$. Then $SSR = 0$, hence $R^2 = 1$. To penalize this, the R^2 is adjusted as:

$$\bar{R}^2 = 1 - \frac{SSR/(n-k)}{TSS/(n-1)}. \quad (39)$$

The reason for this particular adjustment is that under the conditions of Proposition 1, SSR

$\sim \chi_{n-k}^2$, whereas under the null hypothesis $\beta_1 = \dots = \beta_{k-1} = 0$, $TSS \sim \chi_{n-1}^2$.

5. *Multicollinearity*

Multicollinearity is the phenomenon that (some of) the explanatory variables are highly correlated. The effect of multicollinearity is that the t-values are deflated. To demonstrate this, I have generated artificial data Y_j, X_{1j}, X_{2j} for $j = 1, \dots, n = 500$ as follows. The explanatory variables X_{1j} have been drawn independently from the $N(0,1)$ distribution. Next, I have drawn random variables V_j independently from the $N(0,1)$ distribution, and have set $X_{2j} = X_{1j} + 0.01 \cdot V_j$. Due to this construction the explanatory variables X_{1j} and X_{2j} are highly correlated. In particular, the R^2 of the regression of X_{2j} on X_{1j} is 0.999892. Finally, I have drawn the errors U_j independently from the $N(0,1)$ distribution, and have generated the dependent variables by

$$Y_j = X_{1j} + X_{2j} + 1 + U_j, \quad j = 1, \dots, n = 500. \quad (40)$$

This is model (8) with $k = 3$ and $\beta_1 = \beta_2 = \beta_3 = 1$.

The EasyReg output involved is below.

OLS estimation results

Parameters	Estimate	t-value (S.E.)	H.C. t-value (H.C. S.E.)
		[p-value]	[H.C. p-value]
b(1)	0.45258	0.101 (4.48811)	0.104 (4.35703)
		[0.91968]	[0.91727]
b(2)	1.57385	0.351 (4.48634)	0.362 (4.35350)
		[0.72573]	[0.71772]
b(3)	0.95879	21.386 (0.04483)	21.410 (0.04478)
		[0.00000]	[0.00000]

Notes:

1: S.E. = Standard error

2: H.C. = Heteroskedasticity Consistent. These t-values and

standard errors are based on White's heteroskedasticity consistent variance matrix.

3: The two-sided p-values are based on the normal approximation.

Effective sample size (n):	500
Variance of the residuals:	1.00464
Standard error of the residuals (SER):	1.002317
Residual sum of squares (RSS):	499.305951
(Also called SSR = Sum of Squared Residuals)	
Total sum of squares (TSS):	2396.64478
R-square:	0.791665
Adjusted R-square:	0.790826

Overall F test: $F(2,497) = 944.29$

p-value = 0.00000

Significance levels:	10%	5%
Critical values:	2.31	3.01
Conclusions:	reject	reject

Breusch-Pagan test = 3.286972

Null hypothesis: The errors are homoskedastic

Null distribution: Chi-square(2)

p-value = 0.19331

Significance levels:	10%	5%
Critical values:	4.61	5.99
Conclusions:	accept	accept

Note that $b(1)$, $b(2)$ and $b(3)$ are the OLS estimators of β_1 , β_2 and β_3 , respectively.

Observe that not only the OLS estimators of β_1 and β_2 are way off from the true value 1, but that also the corresponding t-values are deflated towards levels where the null hypotheses $\beta_1 = 0$ and $\beta_2 = 0$ cannot be rejected by the separate t tests at any reasonable significance level. On the other hand, the overall F test strongly rejects the joint null hypothesis that $\beta_1 = \beta_2 = 0$. These contradictory results are due to multicollinearity.

There is no cure for multicollinearity. The only thing you can do is be aware of it, and always test joint hypotheses using the F or Wald test rather than using separate t tests.

Assumption 1 does not rule out multicollinearity, but only its extreme form, where one explanatory variable is an exact linear function of other explanatory variables. For example, consider model (2), and suppose that $Z_j = \eta + \delta X_j$ without error. Then model (2) becomes

$$\begin{aligned} Y_j &= \alpha + \beta \cdot X_j + \gamma \cdot Z_j + U_j = \alpha + \beta \cdot X_j + \gamma \cdot (\eta + \delta X_j) + U_j \\ &= (\alpha + \gamma \cdot \eta) + (\beta + \gamma \cdot \delta) \cdot X_j + U_j, \quad j = 1, 2, \dots, n, \end{aligned} \quad (41)$$

Clearly, only $\alpha + \gamma \cdot \eta$ and $\beta + \gamma \cdot \delta$ can be estimated by OLS, but not α , β and γ separately without knowing η and δ . This case is ruled out by Assumption 1.

6. *Heteroscedasticity*

Recall that the errors U_j of a regression model are heteroscedastic if the conditional variance of U_j given the explanatory variables is not constant, but a function of the explanatory variables. In particular, the error terms in model (8) are heteroscedastic if there exists a non-constant function ψ such that

$$E[U_j^2 | X_{1,j}, X_{2,j}, \dots, X_{k-1,j}] = \psi(X_{1,j}, X_{2,j}, \dots, X_{k-1,j}). \quad (42)$$

Heteroscedasticity often occurs in practice. It is actually the rule rather than the exception. One of problems of heteroscedasticity is that the standard errors and t-values of the OLS parameter estimators are no longer valid. However, this problem can easily be cured by replacing the standard errors (28) with the heteroscedasticity consistent (H.C.) standard errors:

$$\tilde{\sigma}_i = \sqrt{\sum_{j=1}^n w_{i,j}^2 \hat{U}_j^2} \quad (= \text{H.C. standard error of } \hat{\beta}_i). \quad (43)$$

and the t-values with the heteroscedasticity consistent (H.C.) t-values

$$\tilde{t}_i = \frac{\hat{\beta}_i}{\tilde{\sigma}_i} \quad (= \text{H.C. t-value of } \hat{\beta}_i).. \quad (44)$$

The F and Wald tests in Proposition 4 are also no longer valid under heteroscedasticity, but the cure for this is difficult to explain at the undergraduate level. To test joint hypotheses under heteroscedasticity with EasyReg you have to increase the econometrics level to

“Intermediate”. Then after running your regression you will get the option to conduct Wald tests of linear parameter restrictions. This option gives you two versions of the Wald test, one for the homoscedastic case and one for the heteroscedastic case. See the guided tour on OLS estimation.

To decide whether the errors of model (8) are homoscedastic or heteroscedastic, use the Breusch-Pagan⁴ test. Given that

$$E[U_j^2 | X_{1j}, X_{2j}, \dots, X_{k-1j}] = g\left(\sum_{i=1}^{k-1} \gamma_i X_{ij} + \gamma_k\right) \text{ for some unknown function } g(\cdot). \quad (45)$$

the Breusch-Pagan test tests the null hypothesis

$$H_0: \gamma_1 = \dots = \gamma_{k-1} = 0 \Leftrightarrow E[U_j^2 | X_{1j}, X_{2j}, \dots, X_{k-1j}] = g(\gamma_k) = \sigma^2, \quad (46)$$

against the alternative hypothesis

$$H_0: E[U_j^2 | X_{1j}, X_{2j}, \dots, X_{k-1j}] = g\left(\sum_{i=1}^{k-1} \gamma_i X_{ij} + \gamma_k\right) \neq g(\gamma_k) \quad (47)$$

Under the null hypothesis (46) of homoscedasticity the test statistic of the Breusch-Pagan test has a χ_{k-1}^2 distribution, and the test is conducted right-sided.

⁴ Breusch, T. and A. Pagan (1979), "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica* 47, 1287-1294.