SPECIFICATION OF ECONOMETRIC MODELS

Herman J. Bierens

Pennsylvania State University

March 26, 2009

1 *Introduction*

Most econometric models link an observable dependent variable Y to observable explanatory variables $X_1,...,X_m$, an unobservable variable U (the error term if the model is a linear regression), and parameters $\beta_1,....,\beta_k$, via some function f:

$$Y = f\left(X_1,...,X_m,U,\beta_1,...,\beta_k\right). \tag{1}$$

For example, in the case of a linear regression model with intercept this function f is specified as

$$f\left(X_1,...,X_m,U,\beta_1,...,\beta_k\right) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1}X_{k-1} + \beta_k + U, \quad m = k-1. \tag{2}$$

The question is: "How should we specify this function f?" There is no single answer to this question, of course, but there are a few rules to narrow down the set of possible choices for f.

2 *Feasibility rule*

The first rule is the feasibility rule:

Rule 1: **(Feasibility)** *The function f should be specified such that the equality* (1) *can hold for all possible values of the dependent variable Y, the explanatory variables* $X_1,...,X_m$ *and the error term U.*

As an example of a violation if this rule, consider the linear regression model

$$Y = \beta_1 X_1 + \beta_2 + U, \quad E[U|X_1] = 0. \tag{3}$$

where Y is the demand for a particular good or service, and X_1 is the price of that good or service. For instance, let Y be the demand for cruises to the Bahamas during the peak season, measured as the total number of people taking cruises to the Bahamas during the peak season times the average duration of a cruise, and let X_1 be the cost of a cruise for one person per day. The law of

demand predicts that if the price X_1 goes up then the demand Y goes down. Therefore, $\beta_1 < 0$. But then $\beta_1 X_1 + \beta_2 < 0$ if $X_1 > -\beta_2/\beta_1$, so that $E[Y|X_1 > -\beta_2/\beta_1] = \beta_1 X_1 + \beta_2 < 0$. However, this is impossible because demand Y cannot be negative.

At first sight one could think of fixing this problem by specifying the demand model as

$$\ln(Y) = \beta_1 X_1 + \beta_2 + U, \quad E[U|X_1] = 0.$$

so that

$$Y = f(X_1, U, \beta_1, \beta_2) = \exp(\beta_1 X_1 + \beta_2)\exp(U).$$

Assume for sake of the argument that U and X_1 are independent and that $U \sim N(0,\sigma^2)$. Then it can be shown (but that is beyond the level of this course) that $E[\exp(U)|X_1] = \exp(\sigma^2/2)$, hence

$$E[Y|X_1] = \exp(\beta_1 X_1 + \beta_2)\exp(\sigma^2/2).$$

Again, the law of demand predicts that $\beta_1 < 0$. But now this model predicts that if the price X_1 decreases towards zero the log of demand Y increases towards an upper bound. In particular,

$$E[Y|X_1] \uparrow \exp(\beta_2)\exp(\sigma^2/2) < \infty \text{ if } X_1 \downarrow 0.$$

This means that even if the good or service involved is completely free the expected demand will be bounded. This may be reasonable for some goods or services, but not for cruises to the Bahamas!

Therefore, it is better to specify this demand model as

$$\ln(Y) = \beta_1 \ln(X_1) + \beta_2 + U, \quad E[U|X_1] = 0. \tag{4}$$

Because $\beta_1 < 0$ we now have that $Y \uparrow \infty$ if $X_1 \downarrow 0$ and $Y \downarrow 0$ if $X_1 \uparrow \infty$. Note that the parameter β_1 in model (4) can be interpreted as the elasticity of demand:

$$\beta_1 = \frac{d\ln(Y)}{d\ln(X_1)} = \frac{(dY)/Y}{(dX_1)/X_1},$$

where the second equality follows from the fact that $d\ln(x)/dx = 1/x$, hence $d\ln(x) = (dx)/x$.

Another violation of rule 1 is the so-called linear probability model, which is discussed in most undergraduate econometrics textbooks. Consider the case where the dependent variable Y is a binary variable: it takes only two values, 0 or 1. For example, let $Y = 1$ if a household owns the home it lives in, and $Y = 0$ if not, and let X_1 be household income. The linear probability model is then the regression model (3) for this case.

Because $E[Y|X_1] = 0\times P[Y = 0|X_1] + 1\times P[Y = 1|X_1] = P[Y = 1|X_1]$ the linear

probability model states that

$$P[Y = 1|X_1] = \beta_1 X_1 + \beta_2. \tag{5}$$

If we would believe this to be true, then we should expect $\beta_1 > 0$, because the higher the household income X_1 the more likely the household will own the home it lives in. But then

$$P[Y = 1|X_1] > 1 \text{ if } X_1 > (1-\beta_2)/\beta_1,$$

which is impossible. The only case where the linear probability model is valid is the case where the explanatory variable X_1 is a binary variable itself: $X_1 = 0$ or $X_1 = 1$, and this explanatory variable is the only one in the model. Then it follows from (5) that

$$P[Y=1|X_1=0] = \beta_2, \; P[Y=1|X_1=1] = \beta_1 + \beta_2,$$

which is possible if $0 < \beta_2 < 1$ and $0 < \beta_1 + \beta_2 < 1$.

The problem of how to specify the function f in (1) in the case of a binary dependent variable Y will be addressed in a separate lecture note.

The problem with the linear probability model also applies to the case where the dependent variable Y is a fraction, so that $0 < Y < 1$. For example, let Y be the share of the expenditures on food and clothing in total expenditures of a household, and let X_1 be household income or the log of household income. The linear regression model (3) is not appropriate in this case, because it will be impossible to force $\beta_1 X_1 + \beta_2$ between zero and one for all possible values of X_1 if $\beta_1 \neq 0$. However, in this case there is an easy solution, namely, let

$$Y = f(X_1, U, \beta_1, \beta_2) = \frac{\exp(\beta_1 X_1 + \beta_2 + U)}{1 + \exp(\beta_1 X_1 + \beta_2 + U)}.$$

Then $0 < Y < 1$. It is easy to verify that this model can be rewritten as

$$\ln\left(\frac{Y}{1-Y}\right) = \beta_1 X_1 + \beta_2 + U, \tag{6}$$

which is a valid regression model if $E[U|X_1] = 0$.

If in model (6) Y is the share of the expenditures on food and clothing in total expenditures of a household and X_1 is household income, one may expect that $\beta_1 < 0$ because the higher income, the more the household will spend on other items than food and clothing

relative to total expenditures. Then Y is maximal for $X_1 = 0$, but this maximum is less than 1 if X_1 is household income itself rather that the log of income, because

$$\max Y = \lim_{X_1 \downarrow 0} \frac{\exp\left(\beta_1 X_1 + \beta_2 + U\right)}{1 + \exp\left(\beta_1 X_1 + \beta_2 + U\right)} = \frac{\exp(\beta_2 + U)}{1 + \exp(\beta_2 + U)} < 1.$$

This is not realistic. However, this problem can easily be cured by replacing X_1 with the log of X_1:

$$\ln\left(\frac{Y}{1-Y}\right) = \beta_1 \ln(X_1) + \beta_2 + U. \tag{7}$$

Then

$$\max Y = \lim_{X_1 \to 0} \frac{\exp\left(\beta_1 \ln(X_1) + \beta_2 + U\right)}{1 + \exp\left(\beta_1 \ln(X_1) + \beta_2 + U\right)} = \lim_{z \to \infty} \frac{\exp(z)}{1 + \exp(z)} = \lim_{z \to \infty} \frac{1}{1 + \exp(-z)} = 1,$$

which is more plausible.

3 *Scale invariance rule*

In January 2002, twelve members of the European union (Belgium, Germany, Greece, Spain, France, Ireland, Italy, Luxembourg, the Netherlands, Austria, Portugal and Finland) changed their local currencies to a common currency, the Euro. Each local currency was exchanged for Euros at a fixed exchange rate. For example, the exchange rate in the Netherlands was 2.20371 Dutch guilders for one Euro.

Consider the model $Y = f(X_1, U, \beta_1, \beta_2)$, where again Y is the demand for a good or service, measured in some non-monetary unit, and X_1 is the price of this good or service, in the Netherlands before the introduction of the Euro. Thus, the price X_1 was initially measured in Dutch guilders. After the introduction of the Euro the price became X_1/2.20371 Euros. Did this conversion to the Euro have an effect on demand Y? In January 2002 and a few months thereafter it had! All prices in Euros where suddenly about 45% of what they were before in Dutch guilders, and it took a while for the public to realize that their income was also about 45% less in Euros than what it was before in Dutch guilders. But once the public got used to the new Euro

the effect on demand of the conversion to the Euro vanished.

For the demand model $Y = f(X_1, U, \beta_1, \beta_2)$ to be valid before and after the introduction of the Euro, it must be possible to adjust the parameters β_1 and β_2 to β_1^* and β_2^* such that $Y = f(X_1, U, \beta_1, \beta_2) = f(X_1/2.20371, U, \beta_1^*, \beta_2^*)$. In other words, changes in the unit of measurement of the explanatory variables should not affect the functional form *f* of the model, because this functional form represents the behavior of economic agents that should not be affected by units of measurements. This leads to our second rule:

Rule 2: **(Scale invariance)** *The function f in* (1) *should be specified such that changes in the unit of measurements of the explanatory variables* $X_1,...,X_m$ *can be compensated by changes in the parameters* $\beta_1, \ldots, \beta_k$: *If* $Y = f(X_1, \ldots, X_m, U, \beta_1, \ldots, \beta_k)$ *then for arbitrary positive numbers* $\lambda_1, \ldots, \lambda_m$ *there exist parameters* $\beta_1^*, \ldots, \beta_k^*$ *such that* $Y = f(\lambda_1 X_1, \ldots, \lambda_m X_m, U, \beta_1^*, \ldots, \beta_k^*)$.

Clearly this rule holds for the linear regression model (2):

$$f(\lambda_1 X_1, \ldots, \lambda_m X_m, U, \beta_1^*, \ldots, \beta_k^*) = \beta_1^* \lambda_1 X_1 + \beta_2^* \lambda_2 X_2 + \cdots + \beta_{k-1}^* \lambda_{k-1} X_{k-1} + \beta_k^* + U, \quad m = k-1.$$

where $\beta_i^* = \beta_i / \lambda_i$ for $i = 1,...,k-1$ and $\beta_k^* = \beta_k$. It also holds for a log-linear model, provided that there is an intercept. For example, in the case (4),

$$\ln(Y) = \beta_1 \ln(X_1) + \beta_2 + U = \beta_1 \ln(\lambda_1 X_1) + \beta_2 - \beta_1 \ln(\lambda_1) + U,$$

so that $\beta_1^* = \beta_1$ and $\beta_2^* = \beta_2 - \beta_1 \ln(\lambda_1)$. Therefore, if one or more variables enter the model in log form you have to include an intercept, as otherwise rule 2 will be violated.

4 *Use economic theory*

In the example of the demand for cruises to the Bahamas we have combined rule 1 with the law of demand. Thus, we have already applied the following third specification rule:

Rule 3: **(Use economic theory)** *Base your model specification on economic theory, if possible.*

However, often economic theory tells you more about which variables to select, and the direction

of the effect of the independent variables on the dependent variable, than about the functional form of the model. Nevertheless, there are a few cases where the functional form can be derived from economic theory. An example is the Mincer wage equation.

The basic idea of Mincer's[1] theory is that wages increase with experience on the job up to a certain point, after which wages decrease with experience. The underlying economic theory is human capital theory: The productivity of a worker increases with on-the-job training, hence it is advantageous for firms to invest in on-the job training of their workers. The training may not be a formal training. Just by having more experience in doing a particular job one can do the job better and faster. However, human capital depreciates over time. For example, a particular job may become obsolete due to change in technology. Anyhow, the human capital theory predicts that the marginal product of a worker first increases with experience on the job but after a certain point will decrease. Consequently, wages follow the same pattern because the optimal[2] wage of a worker is equal to his or her marginal product.

A functional form of the wage-experience relationship that can mimic this pattern is a quadratic function:

$$Y = \alpha + \beta X + \gamma X^2 + U, \tag{8}$$

where Y is some measurement of wage (to be discussed below), X is experience on the job and U is an error term. The quadratic function involved is maximal at say X_0 years of experience if the slope $\beta + 2\gamma X$ is positive for $X < X_0$ and negative for $X > X_0$. A necessary condition for this is that $\beta > 0$ and $\gamma < 0$. Then $X_0 = -0.5\beta/\gamma > 0$.

As to the dependent variable Y, we cannot choose for Y the wage itself, because the right-hand side of (8) can become negative for large X, whereas the wage cannot be negative, so that then rule 1 will be violated. Therefore, let $Y = \ln(Wage)$. Then the basic Mincer wage equation takes the form

$$\ln(Wage) = \alpha + \beta.Experience + \gamma.Experience^2 + U, \quad \beta > 0,\ \gamma < 0. \tag{9}$$

---

[1] Mincer, J. (1974), *Schooling, Experience and Earnings*, New York: Columbia University Press.

[2] Optimal from the point of view of the firm.

Often this model is augmented with a variety of other explanatory variables, such as gender and race indicators, years or level of schooling, location variables, etc.

5 *Testing for misspecification of functional form*

5.1 *Ramsey's RESET test*[3]

Consider the general linear regression model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k + U. \tag{10}$$

Recall that the crucial condition for correctness of this model is that the conditional expectation of the error term U given the explanatory variables is zero:

$$E[U|X_1, X_2, \cdots, X_{k-1}] = 0 \tag{11}$$

because then

$$E[Y|X_1, X_2, \cdots, X_{k-1}] = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k. \tag{12}$$

The conditions (11) and (12) are equivalent: If (12) holds then condition (11) holds for the error term U in model (10), and vice versa.

The question is how to test the null hypothesis (11) (or equivalently, the null hypothesis (12)). Before we can answer this question, we need to formulate an alternative hypothesis. The most general alternative hypothesis is that (12) is not true, but for practical purposes we have to narrow down this alternative, as follows. Assume that for some function g and parameters $\beta_1, \cdots, \beta_{k-1}, \beta_k$,

$$E[Y|X_1, X_2, \cdots, X_{k-1}] = g\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k\right). \tag{13}$$

Then the null hypothesis (12) boils down to the hypothesis that $g(x) = x$.

If the function g in (13) is p times differentiable we can approximate it by a polynomial of order p:

$$g(x) \approx \gamma_0 + \gamma_1 x + \ldots \gamma_p x^p. \tag{14}$$

Often this approximation is already pretty close for $p = 2$, so let us focus on that case. Thus, consider the alternative hypothesis

---

[3] Ramsey, J. (1974), "Classical Model Selection Through Specification Error Tests", in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press.

$$E[Y|X_1, X_2, \cdots, X_{k-1}] = \gamma_0 + \gamma_1(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k)$$
$$+ \gamma_2(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k)^2 \quad (15)$$
$$= \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_{k-1} X_{k-1} + \delta_k + \gamma_2(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k)^2,$$

where $\delta_k = \gamma_0 + \beta_k$ and $\delta_i = \gamma_1 \beta_i$ for $i = 1,...,k-1$. Now the null hypothesis (12) corresponds to the null hypothesis that $\gamma_2 = 0$.

This suggests the following testing procedure. First, estimate model (10) by OLS, and compute

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_{k-1} X_{k-1} + \hat{\beta}_k, \quad (16)$$

where the $\hat{\beta}_i$'s are the OLS estimates of the β_i's. You can do that in EasyReg in two steps. Once your estimation results are displayed in the "What to do next?" window, open "Options" and click "Write residuals to the input file". Then the OLS residuals

$$\hat{U} = Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \cdots - \hat{\beta}_{k-1} X_{k-1} - \hat{\beta}_k = Y - \hat{Y}$$

are added to the input file, as new variable OLS Residual of "Y", where "Y" is the actual name of the dependent variable *Y*. Next, open "Menu > Input > Transform variables" in the EasyReg main window, click "Linear combination of variables", select the variables *Y* and $\hat{U}$ (= OLS Residual of "Y"), and make the linear combination $Y - \hat{U} = \hat{Y}$.

Second, include the variable $\hat{Y}^2$ in the augmented regression model:

$$Y = \gamma_2 \hat{Y}^2 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_{k-1} X_{k-1} + \delta_k + U, \quad (17)$$

and re-estimate it by OLS. Of course, you have to make the variable $\hat{Y}^2$ first, which can be done in EasyReg as follows. Open "Menu > Input > Transform variables" in the EasyReg main window again, click "Multiplicative combination of variables", select the new variable $\hat{Y}$ and choose power 2. This creates the variable $\hat{Y}^2$.

Finally, test the null hypothesis $\gamma_2 = 0$ against the alternative hypothesis $\gamma_2 \neq 0$, using the t-test.

A more general test is to estimate the augmented regression model

$$Y = \gamma_2 \hat{Y}^2 + \cdots + \gamma_p \hat{Y}^p + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_{k-1} X_{k-1} + \delta_k + U, \quad (18)$$

by OLS and test the joint null hypothesis $\gamma_2 = 0,...,\gamma_p = 0$ against the alternative that $\gamma_i \neq 0$ for some index *i* (=2,...,p), using the Wald test. This test is known as Ramsey's

Regression **S**pecification **E**rror **T**est (RESET).

5.2 *An application*

The data set for this application of the RESET test is random sample[4] of size $n = 2000$ from the Dutch Wage Structure Survey 1997, containing the following variables:

ln(Wage), where Wage is the hourly wage in Dutch guilders, times 100

Gender (Female = 1, Male = 0)

College (1 if education level ≥ 5, 0 if not)[5]

Experience (Experience with the present employer, in years)

Age (in years)

The initial model is model (9), augmented with the variables Gender, College, and Age. The EasyReg output involved is:

```
Y = LN[Wage]
X variables:
X(1) = experience
X(2) = experience^2
X(3) = Gender
X(4) = College
X(5) = age
X(6) = 1

Model: Y = b(1)X(1) +.....+ b(6)X(6)  + U, where U is the error term,
satisfying E[U|X(1),...,X(6)] = 0.

OLS estimation results
Parameters  Estimate    t-value    H.C. t-value
                         (S.E.)     (H.C. S.E.)
                       [p-value]  [H.C. p-value]
b(1)         0.01772      7.779          7.921
                        (0.00228)      (0.00224)
                        [0.00000]      [0.00000]
b(2)        -0.00045     -7.139         -7.626
                        (0.00006)      (0.00006)
                        [0.00000]      [0.00000]
```

[4] See Bierens, H.J. and J. Hartog (1988), "Non-Linear Regression with Discrete Explanatory Variables", *Journal of Econometrics* 38, 269-299, for a description of this sample. This data set is included in the EasyReg database.

[5] The original data set contains the variable "level of education", ranging from 1 to 7. Level 5 is "higher general" which is comparable with a BA degree.

```
b(3)        -0.16947    -10.322         -11.420
                       (0.01642)       (0.01484)
                       [0.00000]       [0.00000]
b(4)         0.46154     28.765          24.445
                       (0.01605)       (0.01888)
                       [0.00000]       [0.00000]
b(5)         0.01024     14.493          12.659
                       (0.00071)       (0.00081)
                       [0.00000]       [0.00000]
b(6)         6.85378    289.287         270.929
                       (0.02369)       (0.02530)
                       [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):                          2000
Variance of the residuals:                      0.073569
Standard error of the residuals (SER):          0.271236
Residual sum of squares (RSS):                146.697047
(Also called SSR = Sum of Squared Residuals)
Total sum of squares (TSS):                   290.506519
R-square:                                         0.4950
Adjusted R-square:                                0.4938

Breusch-Pagan test = 143.749687
Null hypothesis:   The errors are homoskedastic
Null distribution: Chi-square(5)
p-value = 0.00000
Significance levels:        10%         5%
Critical values:           9.24      11.07
Conclusions:             reject     reject
```

Next, augment this model with $\hat{Y}^2$ = (LN[Wage]-OLS Residual of LN[Wage])^2. Then the results become:

```
X variables:
X(1) = experience
X(2) = experience^2
X(3) = Gender
X(4) = College
X(5) = age
X(6) = (LN[Wage]-OLS Residual of LN[Wage])^2
X(7) = 1

Model: Y = b(1)X(1) +.....+ b(7)X(7)  + U, where U is the error term,
satisfying E[U|X(1),...,X(7)] = 0.
```

```
OLS estimation results
Parameters  Estimate    t-value    H.C. t-value
                         (S.E.)     (H.C. S.E.)
                      [p-value]  [H.C. p-value]
b(1)        -0.06223     -2.770          -2.392
                        (0.02247)       (0.02602)
                        [0.00561]       [0.01677]
b(2)         0.00159      2.766           2.390
                        (0.00057)       (0.00066)
                        [0.00567]       [0.01684]
b(3)         0.57380      2.753           2.373
                        (0.20843)       (0.24179)
                        [0.00591]       [0.01764]
b(4)        -1.66944     -2.801          -2.411
                        (0.59594)       (0.69237)
                        [0.00509]       [0.01590]
b(5)        -0.03613     -2.783          -2.399
                        (0.01298)       (0.01506)
                        [0.00539]       [0.01644]
b(6)         0.30645      3.577           3.068
                        (0.08567)       (0.09988)
                        [0.00035]       [0.00215]
b(7)        -7.46569     -1.865          -1.599
                        (4.00313)       (4.67026)
                        [0.06219]       [0.10992]

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):                          2000
Variance of the residuals:                      0.073137
Standard error of the residuals (SER):          0.270438
Residual sum of squares (RSS):                145.761204
(Also called SSR = Sum of Squared Residuals)
Total sum of squares (TSS):                   290.506519
R-square:                                         0.4983
Adjusted R-square:                                0.4967

Breusch-Pagan test = 129.195866
Null hypothesis:   The errors are homoskedastic
Null distribution: Chi-square(6)
p-value = 0.00000
Significance levels:        10%         5%
Critical values:          10.64      12.59
Conclusions:             reject     reject
```

Note that the parameter γ_2 in (17) corresponds to b(6). Thus, the test statistic of the RESET test is the t-value corresponding to b(6). Because there is strong evidence of heteroskedasticity, the appropriate t-value is the H.C. t-value (`3.068`). Under the null hypothesis $\gamma_2 = 0$ this t-value is

a random drawing from the standard normal distribution (because the sample size *n* is large). Moreover, recall that the 5% critical value of the two-sided standard normal test is 1.96. Clearly, the null hypothesis is rejected at the 5% significance level, and in view of the p-value involved even at the 0.1% significance level!

But now the problem arises how to fix the misspecification. Sometimes you can do that by allowing for interactions between variables. In this case it is conceivable that the effect of experience and age on the log of the wage is different for females and males, and college graduates and non-college graduates. Also, may be we need higher powers of experience as well. To test this, regress LN(Wage) on

```
X(1) = experience
X(2) = experience^2
X(3) = experience^3
X(4) = Gender
X(5) = College
X(6) = age
X(7) = Gender x College
X(8) = Gender x experience
X(9) = Gender x experience^2
X(10) = Gender x experience^3
X(11) = College x experience
X(12) = College x experience^2
X(13) = College x experience^3
X(14) = Gender x age
X(15) = College x age
X(16) = College x Gender x experience
X(17) = College x Gender x experience^2
X(18) = College x Gender x experience^3
X(19) = College x Gender x age
X(20) = 1


Model:
Y = b(1)X(1) +.....+ b(20)X(20)  + U, where U is the error term, satisfying
E[U|X(1),...,X(20)] = 0.


OLS estimation results
Parameters  Estimate    t-value    H.C. t-value
                         (S.E.)     (H.C. S.E.)
                       [p-value]  [H.C. p-value]
b(1)         0.02369      4.792           4.457
                       (0.00494)       (0.00531)
                       [0.00000]       [0.00001]
b(2)        -0.00079     -2.579          -2.397
                       (0.00030)       (0.00033)
                       [0.00991]       [0.01652]
b(3)         0.00001      1.315           1.222
                       (0.00001)       (0.00001)
                       [0.18843]       [0.22153]
```

b(4)	-0.41656	-6.512	-6.033
		(0.06397)	(0.06904)
		[0.00000]	[0.00000]
b(5)	-0.22960	-2.816	-2.238
		(0.08154)	(0.10259)
		[0.00487]	[0.02522]
b(6)	0.00650	8.260	7.454
		(0.00079)	(0.00087)
		[0.00000]	[0.00000]
b(7)	-0.00628	-0.031	-0.030
		(0.20386)	(0.21032)
		[0.97541]	[0.97617]
b(8)	0.02803	1.573	1.873
		(0.01782)	(0.01497)
		[0.11581]	[0.06106]
b(9)	-0.00279	-1.565	-1.835
		(0.00179)	(0.00152)
		[0.11750]	[0.06651]
b(10)	0.00005	1.108	1.357
		(0.00005)	(0.00004)
		[0.26781]	[0.17492]
b(11)	-0.00814	-0.578	-0.560
		(0.01407)	(0.01454)
		[0.56305]	[0.57564]
b(12)	-0.00034	-0.344	-0.329
		(0.00099)	(0.00103)
		[0.73082]	[0.74189]
b(13)	0.00001	0.378	0.388
		(0.00002)	(0.00002)
		[0.70525]	[0.69776]
b(14)	0.00776	3.592	3.135
		(0.00216)	(0.00247)
		[0.00033]	[0.00172]
b(15)	0.02142	9.964	7.651
		(0.00215)	(0.00280)
		[0.00000]	[0.00000]
b(16)	0.07313	1.147	1.423
		(0.06376)	(0.05141)
		[0.25141]	[0.15488]
b(17)	-0.01026	-1.394	-1.512
		(0.00736)	(0.00678)
		[0.16323]	[0.13056]
b(18)	0.00037	1.595	1.639
		(0.00023)	(0.00023)
		[0.11069]	[0.10125]
b(19)	-0.00844	-1.235	-1.097
		(0.00683)	(0.00769)
		[0.21683]	[0.27246]
b(20)	6.95949	243.701	216.759
		(0.02856)	(0.03211)
		[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):                          2000
Variance of the residuals:                      0.067767
Standard error of the residuals (SER):          0.260321
Residual sum of squares (RSS):                134.178399
(Also called SSR = Sum of Squared Residuals)
Total sum of squares (TSS):                   290.506519
R-square:                                         0.5381
Adjusted R-square:                                0.5337

Breusch-Pagan test = 144.599224
Null hypothesis:   The errors are homoskedastic
Null distribution: Chi-square(19)
p-value = 0.00000
Significance levels:        10%         5%
Critical values:          27.2       30.14
Conclusions:            reject     reject
```

Indeed, quite a few interaction variables are significant. None of the interaction variables involving Gender × College are significant, but they are jointly significant. However, if we augment this model with $\hat{Y}^2$ = (LN[Wage]-OLS Residual of LN[Wage])^2, the coefficient involved is still significant at the 5% level. Thus, despite the interaction variables the model is still misspecified. As shown by Bierens and Hartog (1988) [see footnote 4], the non-linearity of the model involved is much more complicated than can be captured by interaction variables and higher powers. Thus, *at the undergraduate econometrics level* the model is beyond repair.