Econometrics 1
Lecture 6: Functional form and specification
ENSAE 2014/2015

Michael Visser (CREST-ENSAE)

Outline lecture

Today's lecture is devoted to several additional issues in multiple linear regression analysis:

- Dummy explanatory variables
- Logarithmic transformations of dependent variable
- Predictions and prediction intervals
- Interactions of independent variables
- Specification tests (Chow test, RESET test, J test)
- Adjusted R-Squared

These issues are less fundamental than the material studied in the previous lectures, but are nonetheless important in understanding (and doing yourself!) empirical studies.

Dummy variables

Explanatory variables can be

- Continuous
- Discrete or categorical (ordered/unordered)

Categorical variables are very common in practice:

- gender, marital status, health status, labor market situation, etc

Example of a categorical variable taking two values:

$$female_i = \begin{cases} 1 & \text{if individual } i \text{ is female} \\ 0 & \text{if individual } i \text{ is male} \end{cases}$$

The variable *female* is a dummy variable indicating the gender of an individual. The 0/1 coding is arbitrary, but is convenient when interpreting regression coefficients.

Dummy variables (cntd)

Dummy variables can be included in MLR models as regular (continuous) variables. Example:

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + u$$

The interpretation of the two parameters is, however, slightly different

$$\begin{aligned} \beta_1 &= \partial E[wage|educ, female]/\partial educ \\ \beta_2 &= E[wage|educ, female = 1] - E[wage|educ, female = 0] \end{aligned}$$

so β_2 is an intercept shift whereas β_1 is a slope.

In the above model the males are chosen as the benchmark group or reference group. It is the group against which comparisons are made: $\beta_0 + \beta_1 educ$ is the expected wage of a man with *educ* years of education.

Dummy variables (cntd)

We can change the reference category without fundamentally changing the model:

$$\begin{aligned} wage &= \beta_0 + \beta_1 educ + \beta_2 female + u \\ &= \beta_0 + \beta_1 educ + \beta_2 (1 - male) + u \\ &= (\beta_0 + \beta_2) + \beta_1 educ - \beta_2 male + u \end{aligned}$$

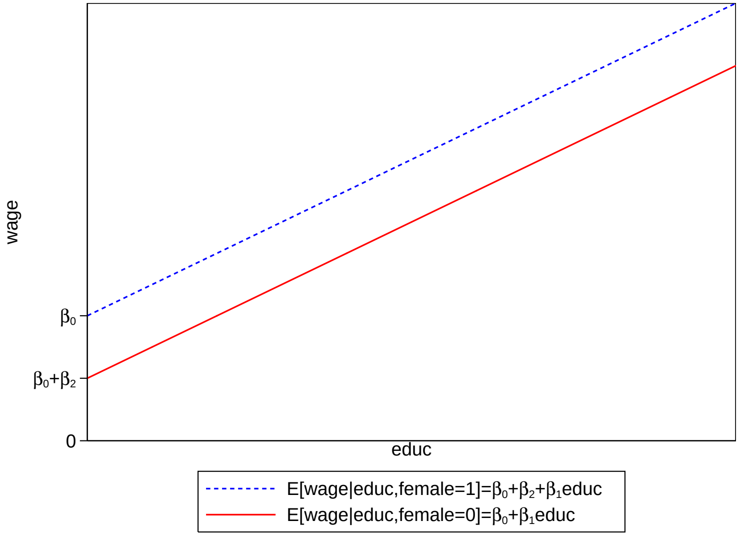
Note that *female* and *male* cannot both be included as regressors:

$$\begin{aligned} wage &= \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 man + u \\ &= (\beta_0 + \beta_3) + \beta_1 educ + (\beta_2 - \beta_3) female + u \end{aligned}$$

We cannot identify separately $\beta_0, \beta_1, \beta_2$ and β_3, but only $(\beta_0 + \beta_3)$, $\beta_1$and $(\beta_2 - \beta_3)$. This the dummy variable trap.

Dummy variables

How it looks



Dummy variables (cntd)

Next consider the case where a regressor takes on multiple discrete values. Let d represent a categorical variable assuming J distinct values.

We do not want to estimate

$$y_i = \beta_0 + \beta_1 d_i + u_i$$

as this model may be very restrictive even if d is ordered (e.g., educational degree).

Solution: create $J - 1$ dummy variables (omit one, J_{ref} to avoid DV trap)

$$d_{ij} = \begin{cases} 1 & \text{if } d_i = j \\ 0 & \text{otherwise} \end{cases}, \quad \forall j \neq J_{ref}$$

Dummy variables (cntd)

We can then estimate

$$y_i = \beta_0 + \sum_{j \neq J_{ref}} \beta_j d_{ij} + u_i$$

where the interpretation of the regression coefficients is

$$\beta_j = E[y|d = j] - E[y|d = J_{ref}]$$

and therefore relative to the omitted category!

The intercept β_0 is here the average outcome of those in the reference group

This means that to compare group k to group l we need to compare their coefficients:

$$\beta_k - \beta_l = E[y|d = k] - E[y|d = l]$$

Dummy variables - example

```
. tab nivet, generate(Dnivet)

            Niveau d'étude le plus élevé |      Freq.     Percent        Cum.
-----------------------------------------+-----------------------------------
 3ème cycle universitaire, grande école |      2,957        9.47        9.47
              2ème cycle universitaire |      2,674        8.56       18.03
               1er cycle universitaire |      1,174        3.76       21.79
                              DUT, BTS |      4,348       13.92       35.70
     Paramédical et social niveau bac+2 |        912        2.92       38.62
                    Terminale générale |      2,149        6.88       45.50
               Terminale technologique |      1,354        4.33       49.84
                     Terminale bac pro |      1,972        6.31       56.15
                   Seconde ou première |      1,026        3.28       59.44
                    Terminale CAP, BEP |      9,026       28.90       88.33
3ème seule, CAP-BEP avant l'année termi |      2,156        6.90       95.23
     4ème-6ème; enseignement spécialisé |        801        2.56       97.80
                     Classes primaires |        688        2.20      100.00
-----------------------------------------+-----------------------------------
                                  Total |     31,237      100.00
```

Dummy variables - example

```
. describe Dniv*

              storage  display
variable name   type   format    variable label
------------------------------------------------------------------------
Dnivet1         byte   %8.0g     nivet==3ème cycle universitaire, grande école
Dnivet2         byte   %8.0g     nivet==2ème cycle universitaire
Dnivet3         byte   %8.0g     nivet==1er cycle universitaire
Dnivet4         byte   %8.0g     nivet==DUT, BTS
Dnivet5         byte   %8.0g     nivet==Paramédical et social niveau bac+2
Dnivet6         byte   %8.0g     nivet==Terminale générale
Dnivet7         byte   %8.0g     nivet==Terminale technologique
Dnivet8         byte   %8.0g     nivet==Terminale bac pro
Dnivet9         byte   %8.0g     nivet==Seconde ou première
Dnivet10        byte   %8.0g     nivet==Terminale CAP, BEP
Dnivet11        byte   %8.0g     nivet==3ème seule, CAP-BEP avant l'année term.
Dnivet12        byte   %8.0g     nivet==4ème-6ème; enseignement spécialisé
Dnivet13        byte   %8.0g     nivet==Classes primaires
```

Dummy variables - example

```
// Reference: nivet==3ème cycle universitaire, grande école
. reg sal Dnivet2-Dnivet13

      Source |       SS       df       MS              Number of obs =   31237
-------------+------------------------------           F( 12, 31224) =  364.43
       Model |  4.3431e+09    12   361928300           Prob > F      =  0.0000
    Residual |  3.1009e+10 31224  993123.477           R-squared     =  0.1229
-------------+------------------------------           Adj R-squared =  0.1225
       Total |  3.5352e+10 31236  1131784.71           Root MSE      =  996.56

------------------------------------------------------------------------------
      salred |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Dnivet2 |  -812.2103   26.59427   -30.54   0.000    -864.3361   -760.0845
     Dnivet3 |  -1040.408   34.37712   -30.26   0.000    -1107.788   -973.0274
     Dnivet4 |  -912.3167   23.75425   -38.41   0.000     -958.876   -865.7574
     Dnivet5 |  -873.5381   37.74661   -23.14   0.000    -947.5229   -799.5532
     Dnivet6 |  -1064.607   28.24868   -37.69   0.000    -1119.976   -1009.239
     Dnivet7 |  -1129.504   32.70059   -34.54   0.000    -1193.598    -1065.41
     Dnivet8 |   -1214.52   28.97356   -41.92   0.000     -1271.31   -1157.731
     Dnivet9 |  -1163.297   36.10833   -32.22   0.000    -1234.071   -1092.523
    Dnivet10 |  -1230.879   21.11597   -58.29   0.000    -1272.267   -1189.491
    Dnivet11 |  -1356.431   28.22211   -48.06   0.000    -1411.748   -1301.115
    Dnivet12 |  -1422.806   39.69521   -35.84   0.000    -1500.611   -1345.002
    Dnivet13 |  -1544.543   42.18232   -36.62   0.000    -1627.222   -1461.864
       _cons |   2637.692   18.32635   143.93   0.000     2601.771    2673.612
------------------------------------------------------------------------------
```

Dummy variables - example

```
// Reference: nivet==Terminale générale
. reg sal Dnivet1-Dnivet5 Dnivet7-Dnivet13

      Source |       SS       df       MS              Number of obs =   31237
-------------+------------------------------           F( 12, 31224) =  364.43
       Model |  4.3431e+09    12   361928300           Prob > F      =  0.0000
    Residual |  3.1009e+10 31224  993123.477           R-squared     =  0.1229
-------------+------------------------------           Adj R-squared =  0.1225
       Total |  3.5352e+10 31236  1131784.71           Root MSE      =  996.56

------------------------------------------------------------------------------
      salred |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Dnivet1 |   1064.607   28.24868    37.69   0.000     1009.239    1119.976
     Dnivet2 |   252.3971   28.87097     8.74   0.000     195.8088    308.9853
     Dnivet3 |   24.19942   36.16717     0.67   0.503    -46.68968    95.08852
     Dnivet4 |   152.2907   26.27817     5.80   0.000     100.7844    203.7969
     Dnivet5 |   191.0693    39.3838     4.85   0.000     113.8755    268.2631
     Dnivet7 |  -64.89663   34.57754    -1.88   0.061      -132.67    2.876737
     Dnivet8 |  -149.9128   31.07644    -4.82   0.000    -210.8239   -89.00176
     Dnivet9 |  -98.68949   37.81652    -2.61   0.009    -172.8114   -24.56759
    Dnivet10 |  -166.2719   23.91991    -6.95   0.000    -213.1559   -119.3879
    Dnivet11 |   -291.824   30.37705    -9.61   0.000    -351.3643   -232.2838
    Dnivet12 |  -358.1991   41.25515    -8.68   0.000    -439.0608   -277.3373
    Dnivet13 |   -479.936   43.65348   -10.99   0.000    -565.4985   -394.3734
       _cons |   1573.084   21.49728    73.18   0.000     1530.949     1615.22
------------------------------------------------------------------------------
```

Logarithmic transformations

We saw in previous lectures that logarithmic transformations conveniently modify the interpretations we can give to parameters.

Logarithmic transformations have other advantages:

- Dependent variable sometimes closer to normal random variable (better approximation of MLR.6)
- Narrows the range of a variable $\Rightarrow$ reduces the impact of outliers

When to take logs and when to use levels?

- Variables that are often transformed in logs: amounts of money (wages, expenditures, sales, market value), size (city, firm)
- But no clear rules $\Rightarrow$ use common (economic) sense and look at data

Logarithmic transformations (cntd)

In the semi-log model

$$\log y = \beta_0 + \beta_1 x + u$$

we have $\Delta y/y = \beta_1 \Delta x$. This is only a good approximation when the increment $\triangle x$ is small. Consider how y changes *exactly*:

$$\begin{aligned}\Delta y &= \exp(\beta_0 + \beta_1(x + \Delta x) + u) - \exp(\beta_0 + \beta_1 x + u) \\ &= \exp(\beta_0 + \beta_1 x + u)(\exp(\beta_1 \Delta x) - 1) = y(\exp(\beta_1 \Delta x) - 1)\end{aligned}$$

and therefore

$$\widehat{\Delta y/y} = \exp(\hat{\beta}_1 \Delta x) - 1$$

This is a consistent estimator of $\Delta y/y$ (but not unbiased). It is preferable to use this exact formula when $\triangle x$ is large. No adjustment is necessary for small (infinitesimal) changes in x :

```
. di exp(0.008) - 1
.00803209

. di exp(0.08) - 1
.08328707

. di exp(0.8) - 1
1.2255409
```

Predicting y when $\log(y)$ is the outcome variable

Suppose we wish to estimate

$$\log y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

but want to use the model to predict y. We have

$$\begin{aligned} E[y|x] &= \int \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u) f(u|x) du \\ &= \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) E[\exp(u)|x] \\ &\overset{MLR.6}{=} \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \exp(\sigma^2/2) \end{aligned}$$

and therefore y can be predicted by

$$\hat{y} = \exp(\hat{\sigma}^2/2) \exp(\widehat{\log(y)})$$

where $\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$. We can also drop the normality assumption MLR.6 and estimate $\alpha_0 \equiv E[\exp(u)|x]$ by regressing y_i on $\hat{m}_i = \exp(\widehat{\log(y_i)})$ (without an intercept!).

Prediction interval

Estimating the MLR model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i \tag{1}$$

we can predict y^0 (some future, unknown, value of the dependent variable) as: $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \ldots + \hat{\beta}_k x_k^0$ (x^0 is the future value of the regressors). The prediction error is

$$\hat{e}^0 \equiv y^0 - \hat{y}^0 = \beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0 + u^0 - \hat{y}^0$$

Since the $\hat{\beta}_j$ are unbiased estimators and u^0 is an error term with mean zero, it follows that $E[\hat{e}^0]$=0 (it is actually a mean conditional on all explanatory variables in the sample and x^0).

Conditional on all explanatory variables we have

$$Var(\hat{e}^0) = Var(\hat{y}^0) + Var(u^0) = Var(\hat{y}^0) + \sigma^2$$

How can we estimate $Var(\hat{y}^0)$?

Prediction interval (cntd)

Define

$$\theta \equiv \beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0$$

Note that $\hat{\theta}$=$\hat{y}^0$. We get

$$\beta_0 = \theta - \beta_1 x_1^0 - \ldots - \beta_k x_k^0$$

Inserting into (1) gives

$$y_i = \theta + \beta_1(x_{i1} - x_1^0) + \ldots + \beta_k(x_{ik} - x_k^0) + u_i$$

So regressing y on an intercept and $(x_1 - x_1^0)$, ..., $(x_k - x_k^0)$ directly gives $\hat{\theta}$ and $\widehat{Var}(\hat{\theta}) = \widehat{Var}(\hat{y}^0)$.

Prediction interval (cntd)

So the standard error of the prediction error is

$$se(\hat{e}^0) = \sqrt{\widehat{Var}(\hat{y}^0) + \hat{\sigma}^2}$$

In practice $\widehat{Var}(\hat{y}^0)$ tends to be small (especially in large samples) compared to $\hat{\sigma}^2$. If there are many important explanatory variables missing in (1), $\hat{\sigma}^2$ tends to be large (the model is than less useful for prediction purposes).

A 95% confidence interval (or prediction interval) for y^0 is:

$$[\hat{y}^0 - t_{0.025} \times se(\hat{e}^0); \hat{y}^0 + t_{0.025} \times se(\hat{e}^0)]$$

Interactions

One important practical technique is the use of *interactions* between different explanatory variables. Consider the wage equation

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exp + u$$

but that we now wish to allow for the possibility that more highly educated individuals have steeper experience profiles:

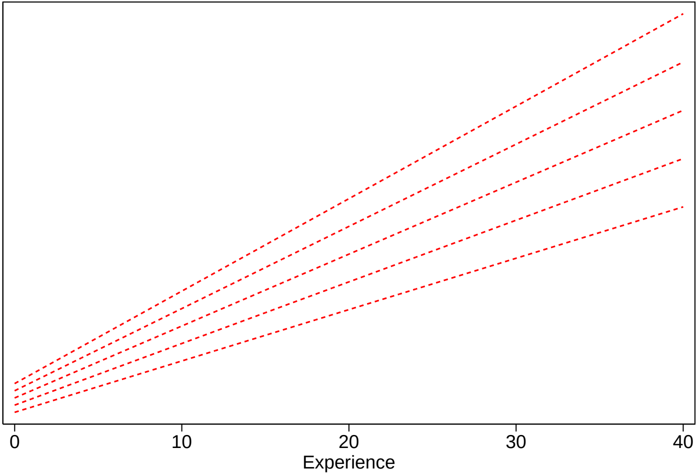
$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 exp \times educ + u$$

The partial effect of *exp* on $\ln(wage)$ is no longer β_2 but

$$\frac{\Delta \ln(wage)}{\Delta exp} = \beta_2 + \beta_3 educ$$

Interactions

Example: interaction education and experience



ln(wage)=β_0+$\beta_1$educ+$\beta_2$exp+$\beta_3$educ*exp+u

Interactions (cntd)

Interacting dummy variables with continuous variables allows the effects of the latter to vary by group or category. Example:

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 female \times educ + u \quad (2)$$

This model allows

- the intercept of the wage equation to be different for men and women
- differential return to education between men and women

The Chow test allows you to check whether the intercept and slopes are the same for both categories. Useful test in may settings and applications. Examples:

- In time series, coefficients may differ across time periods
- In labor economics, coefficients may differ by race or gender
- In cross-country analyses, coefficients may differ by government type or level of development

Chow test

Consider the model

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i, \quad i = 1, \ldots, N_1 \\
y_i &= \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_k x_{ik} + u_i, \quad i = N_1 + 1, \ldots, N_1 + N_2
\end{aligned}
$$

We would like to test $\mathcal{H}_0 : \beta_0 = \alpha_0, \beta_1 = \alpha_1, \ldots, \beta_k = \alpha_k$.

To do so we can pool all observations and estimate the model with interactions:

$$
y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \gamma_0 d_i + \gamma_1 x_{i1} d_i + \ldots + \gamma_k x_{ik} d_i + u_i
$$

where

$$
d_i = \begin{cases} 0 & \text{if } i \in \text{Group 1} \\ 1 & \text{if } i \in \text{Group 2} \end{cases}
$$

and test $\mathcal{H}_0 : \gamma_0 = \gamma_1 = \ldots = \gamma_k = 0$ against the alternative where at least one $\gamma_j \neq 0$, via a F-test.

Chow test

This amounts to estimating the restricted regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i, \quad i = 1, \ldots, N_1 + N_2$$

and then the separate regressions

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i, \quad i = 1, \ldots, N_1 \\ y_i &= \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_k x_{ik} + u_i, \quad i = N_1 + 1, \ldots, N_2 \end{aligned}$$

and use their respective SSR's to construct an F test

$$F = \frac{(SSR_r - (SSR_1 + SSR_2))/(k+1)}{(SSR_1 + SSR_2)/(n - 2(k+1))} \sim F_{k+1, n-2(k+1)}$$

where $SSR_{ur} = SSR_1 + SSR_2$.

Chow test-example

Consider again the wage model (2). The two separate wage equations are

$$\begin{aligned} ln(wage_i) &= \beta_0 + \beta_1 educ_i + u_i, \quad \text{if } i \text{ is male} \\ ln(wage_i) &= \alpha_0 + \alpha_1 educ_i + u_i, \quad \text{if } i \text{ is female} \end{aligned}$$

and our fully interacted model is as follows

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + \gamma_0 female_i + \gamma_1 female_i \times educ_i + u_i$$

The null hypothesis we wish to test is $\mathcal{H}_0 : \gamma_0 = \gamma_1 = 0$ against the alternative where at least one $\gamma_j \neq 0$. The next listings show that the F-statistic equals 516.31. Since the critical value (5% level) of the F distribution with $k + 1 = 2$ and $n - 2(k + 1)$=31233 degrees of freedom is 3, the null is strongly rejected.

Chow Test - example

```
. reg lnw edu

      Source |       SS       df       MS              Number of obs =   31237
-------------+------------------------------           F(  1, 31235) = 4498.68
       Model |  633.045256     1  633.045256           Prob > F      =  0.0000
    Residual |  4395.32675 31235  .140718001           R-squared     =  0.1259
-------------+------------------------------           Adj R-squared =  0.1259
       Total |  5028.37201 31236  .160980023           Root MSE      =  .37512

------------------------------------------------------------------------------
         lnw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        eduy |   .0528497     .000788    67.07   0.000     .0513053    .0543941
       _cons |   3.068821    .0097426   314.99   0.000     3.049725    3.087917
------------------------------------------------------------------------------
```

Chow Test - example

```
. reg lnw eduy if femme==1

      Source |       SS       df       MS              Number of obs =   15269
-------------+------------------------------           F(  1, 15267) = 2438.81
       Model |  353.227455     1  353.227455           Prob > F      =  0.0000
    Residual |  2211.20794 15267  .144835786           R-squared     =  0.1377
-------------+------------------------------           Adj R-squared =  0.1377
       Total |  2564.43539 15268  .167961448           Root MSE      =  .38057

------------------------------------------------------------------------------
         lnw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        eduy |   .0572376    .001159    49.38   0.000     .0549658    .0595094
       _cons |   2.946775   .0145747   202.18   0.000     2.918207    2.975343
------------------------------------------------------------------------------

. reg lnw eduy if femme==0

      Source |       SS       df       MS              Number of obs =   15968
-------------+------------------------------           F(  1, 15966) = 2545.46
       Model |  325.787744     1  325.787744           Prob > F      =  0.0000
    Residual |  2043.45277 15966  .127987772           R-squared     =  0.1375
-------------+------------------------------           Adj R-squared =  0.1375
       Total |  2369.24052 15967  .148383573           Root MSE      =  .35775

------------------------------------------------------------------------------
         lnw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        eduy |   .0526927   .0010444    50.45   0.000     .0506455    .0547398
       _cons |   3.135817   .0126997   246.92   0.000     3.110924     3.16071
------------------------------------------------------------------------------

. di ((4395.3 - 2043.5 -  2211.2)/2)/((2043.5 + 2211.2)/31233)
516.31
```

RESET test

In principle misspecification of the model leads to inconsistent estimates. Suppose for example that the true model is

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 exp^2 + u$$

but instead we estimate

$$\ln(wage) = \alpha_0 + \alpha_1 educ + \alpha_2 exp + v$$

then we will not get consistent estimators of the parameters of interest since v is correlated with the variable *exp* (except when $\beta_3 = 0$).

We can use the

- F test to detect a misspecified functional form: add higher order terms/interactions and perform test
- One alternative is Ramsey's RESET (Regression Equation Specification Error Test) test

RESET Test (cntd)

The idea behind the RESET test is to estimate a baseline model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u,$$

calculate $\hat{y}$ and then estimate

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + v$$

By including powers of the predicted values in the model we implicitly add nonlinear functions of the explanatory variables. In practice one usually adds only $\hat{y}^2$ and $\hat{y}^3$ (simply because it works in a satisfactory way in most applications!). The RESET test amounts to test $\mathcal{H}_0 : \delta_1 = 0, \delta_2 = 0$ via a F-test. The corresponding F statistic is approximately $F_{2,n-k-3}$ distributed.

Drawback: RESET does not tell you what to do when you reject $\mathcal{H}_0$.

RESET Test - example

```
. reg lnw eduy age

      Source |       SS       df       MS              Number of obs =   31237
-------------+------------------------------           F(  2, 31234) = 4299.80
       Model |  1085.56493     2  542.782465           Prob > F      =  0.0000
    Residual |  3942.80707 31234  .126234458           R-squared     =  0.2159
-------------+------------------------------           Adj R-squared =  0.2158
       Total |    5028.372 31236  .160980023           Root MSE      =  .35529

------------------------------------------------------------------------------
         lnw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        eduy |   .0629493   .0007651    82.27   0.000     .0614496    .0644489
         age |   .0145086   .0002423    59.87   0.000     .0140336    .0149836
       _cons |   2.417844   .0142606   169.55   0.000     2.389893    2.445795
------------------------------------------------------------------------------

. predict plnw
(option xb assumed; fitted values)

. g plnw2= plnw^2

. g plnw3= plnw^3
```

RESET Test - example

```
. reg lnw eduy age plnw?

      Source |       SS       df       MS              Number of obs =   31237
-------------+------------------------------           F(  4, 31232) = 2307.98
       Model |  1147.23391     4  286.808478           Prob > F      =  0.0000
    Residual |  3881.13809 31232  .124267997           R-squared     =  0.2282
-------------+------------------------------           Adj R-squared =  0.2281
       Total |    5028.372 31236  .160980023           Root MSE      =  .35252

------------------------------------------------------------------------------
         lnw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        eduy |    .961451   .3504928     2.74   0.006     .2744712    1.648431
         age |   .2220014   .0807778     2.75   0.006     .0636737    .3803291
       plnw2 |  -4.783435   1.507037    -3.17   0.002    -7.737288   -1.829583
       plnw3 |   .5123096    .135732     3.77   0.000     .2462695    .7783497
       _cons |   23.60608   6.622675     3.56   0.000     10.62538    36.58679
------------------------------------------------------------------------------

. test  plnw2 plnw3

 ( 1)  plnw2 = 0
 ( 2)  plnw3 = 0

       F(  2, 31232) =  248.13
            Prob > F =    0.0000
```

Non-nested alternatives

J test (Davidson-MacKinnon)

Sometimes we want to test non-nested models against each other (one model cannot be seen as a special case of the other).
Suppose, for example, that we wish to compare two models:

$$\begin{aligned} \textit{Model A}: \; y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\ \textit{Model B}: \; y &= \gamma_0 + \gamma_1 \log(x_1) + \gamma_2 \log(x_2) + v \end{aligned}$$

Neither equation is a special case of the other. The idea of the

J-test is to embed both competing models in a more general one (an artificial compound model) and then test both original models against it. Define the compound model

$$y = (1-\delta)\beta_0 + (1-\delta)\beta_1 x_1 + (1-\delta)\beta_2 x_2 + \delta\gamma_0 + \delta\gamma_1 \log(x_1) + \delta\gamma_2 \log(x_2) + e \tag{3}$$

The compound model (3) collapses to model A when $\delta = 0$ and to model B when $\delta = 1$.

Non-nested alternatives (cntd)

J test (Davidson-MacKinnon)

In general the parameters γ, β and δ cannot be identified separately. Davidson and MacKinnon suggest to replace (3) by a

model in which the unknown parameters of the model that is not being tested by estimates of these parameters that would be consistent if the model they belong to is the true one. To test

model A against (3) we thus replace γ_0, γ_1, and γ_2 by the OLS estimates $\hat{\gamma}_0$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$ obtained from estimating model B. We then estimate the following regression

$$y = \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 + \delta \hat{y} + error$$

where $\beta_0^* = (1-\delta)\beta_0$, $\beta_1^* = (1-\delta)\beta_1$, $\beta_2^* = (1-\delta)\beta_2$, and $\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 \log(x_1) + \hat{\gamma}_2 \log(x_2)$.

We can then test model A against (3) by checking whether the null hypothesis $\delta = 0$ holds via the usual t-test.

Non-nested alternatives (cntd)

J test (Davidson-MacKinnon)

Remarks:

- Idea of the test: if one model is true then the fitted value from the other model should be insignificant when added to the true model
- To test model B against (3), we need to reverse the role of models A and B, and proceed analogously.
- It is possible to reject one model, both models or neither!

Adjusted R-squared

R-squared indicates how much variation in y is explained by the regressors in the population (see previous lectures):

$$R^2 = 1 - \frac{SSR}{SST}$$

An inconvenient aspect of an R-squared is that it increases mechanically as more regressors are added. As such it is not a useful criterium two compare two nested models (one with more regressors than the other).

Recalling that $SSR/(n-k-1)$ is an unbiased estimator of the variance of the error term u, and $SST/(n-1)$ an unbiased estimator of the variance of y, a natural generalization of the R-squared is

$$\overline{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - (1-R^2)(n-1)/(n-k-1)$$

This is called the "adjusted R-squared". It's main advantage: $\overline{R}^2$ can go up <u>and down</u> when adding a regressor to the model.

Adjusted R-squared (cntd)

Remarks:

- Everything else equal, simpler models are better. Since R^2 does not penalize more complicated models it is better to use $\overline{R}^2$ in comparing different models.
- $\overline{R}^2$ can be negative. Example: $R^2 = 0.01$, $n = 51$, $k = 10$ $\Rightarrow \overline{R}^2 = -0.125$.
- If a new regressor is added to the model, $\overline{R}^2$ goes up if $|t| > 1$ (t being the t-statistic on the new regressor). If a set of variables is added, $\overline{R}^2$ goes up if $F > 1$ (F statistic for joint significance of the new variables). This shows that using the adjusted R-squared leads to the inclusion of more regressors than when using the t/F statistic.
- In general $\overline{R}^2$ cannot be used to compare models with different dependent variables (as with the regular R^2).

I understand/can apply...

- How to use
  - dummy variables
  - transformations (e.g. logarithmic)
  - Interactions
- The adjusted R-squared
- How to obtain a confidence interval for a prediction
- The Chow test and other specification tests