Linear models

CONTENTS:

OLS and GLS regressions

The **general regression model** with additive errors is written in vector notation as:

$$\mathbf{y} = \mathrm{E}[\mathbf{y}|\mathbf{x}] + \mathbf{u} \tag{1}$$

where $\mathrm{E}[\mathbf{y}|\mathbf{x}]$ denotes the conditional expectation of the random variable **y** given **x,** and **u** denotes a vector of unobserved random errors or disturbances.

A **linear regression model** is obtained when $\mathrm{E}[\mathbf{y}|\mathbf{x}]$ is specify to be a linear function of **x.**

In vector notation, the *i*th observation, *i*=1,…,N, is

$$y_i = \mathbf{x}_i'\beta + u_i \tag{2}$$

where $\mathbf{x}_i$is a K x 1 **regressor vector** and $\beta \equiv (\beta_1, \beta_2, \dots, \beta_j, \dots, \beta_K)'$ is a K x 1 **parameter vector.**[1]

Ordinary Least Squares (OLS)

The OLS estimator is defined to be the estimator that minimizes the sum of squared errors

$$\textstyle\sum_{i=1}^{N} u_i^2 = \mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\beta)'\,(\mathbf{y} - \mathbf{X}\beta) \tag{3}$$

Setting the derivative with respect to β equal to **0** and solving for β yields the OLS estimator:

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4}$$

Now we need some **assumptions** that ensure that the OLS estimator can consistently estimate β. The key assumption for **consistency** is the **population orthogonality condition**:

ASSUMPTION OLS.1: $\mathrm{E}(\mathbf{x}'u) = 0$ (5)

Because **x** contains a constant, Assumption OLS.1 is equivalent to saying that u has mean zero and is uncorrelated with each regressor. Sufficient for assumption OLS.1 is the zero conditional mean assumption $\mathrm{E}(u|x_1, x_2, \dots, x_K) = \mathrm{E}(u|\mathbf{x}) = 0$.

The other assumption needed for consistency of OLS is that there is no exact linear relationship among the regressors in the population. That is:

[1] In matrix notation the N observations are stacked by row to yield $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ (2.1) where **y** is an N x 1 vector of dependent variables, **X** is an N x K regressors matrix, and **u** is an N x 1 error vector. Both notations in equations (2) and (2.1) are equivalent for the linear regression model.

ASSUMPTION OLS.2: rank E($\mathbf{x}'\mathbf{x}$)= K[2] (6)

Also, under assumptions OLS.1 and OLS.2 the parameter vector β is **identified**.

Another very useful assumption is the one of **homoskedasticity**:

ASSUMPTION OLS.3: Var(u)=σ^2 (7)

This assumption is useful for deriving the limiting distribution because it implies the asymptotic validity of the usual OLS standard errors and test statistics.

The asymptotic variance of the estimator $\hat{\beta}$ then yields $var(\hat{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ where the estimator of σ^2, $\hat{\sigma}^2 \equiv \sum_{i=1}^{N} \hat{u}_i^2$ $/N – K,$ is the sum of squared residuals.

Failure of assumption OLS.3 can be solved in two ways: using **heteroskedasticity-robust standard errors** (explained in the next section) or using the **generalized least squares estimator**.

EXAMPLE 1 (Wage equation)

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 educ + \beta_3 age + \beta_4 married \qquad (8)$$

Using the data on the 935 individuals in WAGE2.RAW (from Blackburn&Neumark, 1992) we can estimate model (8) by OLS. The variables used are the monthly earnings (*wage*), years of experience (*exper*) , years of education (*educ*), age in years (*age*) and marital status (*married* =1 if married). The results suggest that one additional year of education generates an expected percentage change of 7.3% in monthly earnings, holding everything else constant at a 1% significance level. Also, we can say that, on average, married individuals are expected to earn 20.3% more than non-married individuals.

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/wage2

. regress lwage educ exper age married

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  4,   930) =   43.88
       Model |  26.2997233     4  6.57493084           Prob > F      =  0.0000
    Residual |  139.356571   930  .149845775           R-squared     =  0.1588
-------------+------------------------------           Adj R-squared =  0.1551
       Total |  165.656294   934  .177362199           Root MSE      =   .3871

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0739592   .0067421    10.97   0.000     .0607277    .0871908
       exper |   .0135097   .0038978     3.47   0.001     .0058601    .0211593
         age |   .0109904   .0048927     2.25   0.025     .0013884    .0205925
     married |   .2028898   .0412962     4.91   0.000     .1218453    .2839344
       _cons |   5.081911   .1592396    31.91   0.000       4.7694    5.394421
------------------------------------------------------------------------------
```

[2] The **rank** of a matrix is the number of linearly independent rows or columns.

Generalized least squares (GLS)

If heteroskedasticity is present the generalized least-squares (GLS) estimator is more efficient[3] than the OLS estimator. To implement this estimator is necessary to specify a model for the error variance matrix different than $\Omega = \sigma^2 \mathrm{I}$ and premultiply the linear regression model in (2.1) by $\Omega^{-1/2}$ to yield

$$\Omega^{-1/2}\mathbf{y} = \Omega^{-1/2}\mathbf{X}\beta + \Omega^{-1/2}\mathbf{u} \tag{9}$$

The errors in this model are therefore zero mean, uncorrelated and homoskedastic. So β can be efficiently estimated by OLS regression of $\Omega^{-1/2}\mathbf{y}$ on $\Omega^{-1/2}\mathbf{X}$**:**

$$\hat{\beta}_{GLS}= (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\ \Omega^{-1}\mathbf{y} \tag{10}$$

The GLS estimator cannot be directly implemented because in practice Ω is not known. Instead, we can specify that $\Omega = \Omega(\gamma)$, where γ is a finite-dimensional parameter vector. We can obtain a consistent estimate of γ ($\hat{\gamma}$) and form $\widehat{\Omega}$= $\Omega(\hat{\gamma})$. Then, the **feasible generalized least squares (FGLS)** estimator is:

$$\hat{\beta}_{FGLS}= (\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\ \widehat{\Omega}^{-1}\mathbf{y} \tag{11}$$

Note that if $\Omega(\gamma)$ is misspecified, FGLS is still consistent but we lose the efficiency gains.

EXAMPLE 2 (Heteroskedasticity in housing price equation)

Using the data on 88 houses included in HPRICE1.RAW we can perform a heteroskedasticity analysis and estimate parameters by FGLS. The database includes variables such as the price of the house in \$1000 (*price*), size of lot in square feet (*lotsize*), size of house in square feet (*sqrft*) and number of bedrooms in the house (*bdroms*).

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms \tag{12}$$

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/hprice1

. quietly reg price lotsize sqrft bdrms

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =    20.55
         Prob > chi2  =   0.0000
```

The **Breusch-Pagan /Cook-Weisberg test** tests the null hypothesis that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables. In this case we reject the null hypothesis concluding that there is presence of heterokedasticity in this data. Imagine that the error variance matrix is $\Omega = \sigma^2 h_i$. To implement a **FGLS** we need to estimate $h(x_i, \gamma)$. We can start with the assumption of Var(u|x)=$\sigma^2\exp(\gamma_0 + \gamma_1 x_1 + \cdots + \gamma_k x_k)$ and estimate $\widehat{\Omega}$= $\sigma^2 \mathrm{h}(x_i, \hat{\gamma})$.

[3] Consider $\hat{\theta}$ and $\tilde{\theta}$ as two estimators of parameter vector θ. Define the MSE ($\tilde{\theta}$)≡ $\mathrm{E}\left[(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)'\right]$, $\hat{\theta}$ is **more efficient** in terms of MSE than $\tilde{\theta}$ if MSE ($\hat{\theta}$)< MSE ($\tilde{\theta}$). If the estimator $\tilde{\theta}$ is **unbiased** then MSE ($\tilde{\theta}$)=Var($\tilde{\theta}$). Alternatively, $\tilde{\theta}$ is an unbiased estimator of θ if E($\tilde{\theta}$)= θ (Bias ($\tilde{\theta}$)=0).

The results below show how to calculate feasible generalized least squares following the specified assumption.

```
. predict e, residuals

. generate e2=e^2

. generate loge2= log(e2)

. quietly reg loge2 lotsize sqrft bdrms

. predict gamma_hat
(option xb assumed; fitted values)

. generate exp_gamma_hat= exp( gamma_hat )

. reg price lotsize sqrft bdrms[aw=1/exp_gamma_hat]
(sum of wgt is   1.3956e-01)

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   24.58
       Model |  181162.939     3  60387.6462           Prob > F      =  0.0000
    Residual |  206329.812    84  2456.30728           R-squared     =  0.4675
-------------+------------------------------           Adj R-squared =  0.4485
       Total |   387492.75    87  4453.93966           Root MSE      =  49.561

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   .0041354   .0014255     2.90   0.005     .0013006    .0069703
       sqrft |   .0924624   .0148661     6.22   0.000     .0628995    .1220253
       bdrms |   6.175452   8.893592     0.69   0.489    -11.51043    23.86133
       _cons |    45.9116   30.82353     1.49   0.140    -15.38438    107.2076
------------------------------------------------------------------------------
```

aw represents analytical weights. Those weights are inversely proportional to the variance of an observation. In this case we are using the weights calculated to correct for heteroskedasticity.

Robust and clustered standard errors

Robust standard errors

As mentioned above heteroskedasticity-robust standard errors are used to solve the potential problems of the failure of assumption OLS.3. This consists in adjusting the standard errors and test statistics when estimating β by OLS so they are valid when heteroskedasticity is suspected. This method is much easier than FGLS procedure but we sacrifice potential efficiency gains from it. The asymptotic variance of the estimator $\hat{\beta}$ with heteroskedasticity robust standard errors is

$$var(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N} \hat{u}_i^2\, \mathbf{x}_i'\mathbf{x}_i\right)(\mathbf{X}'\mathbf{X})^{-1} \equiv (\mathbf{X}'\mathbf{X})^{-1}(\hat{\Omega})(\mathbf{X}'\mathbf{X})^{-1} \qquad (13)$$

Sometimes this equation is multiplied by N/ (N - K) as a degrees of freedom correction.

Once standard errors are obtained, *t* statistics are computed in the usual way.

EXAMPLE 1 (Wage equation-Continuation)

```
. regress lwage educ exper age married, r

Linear regression                                      Number of obs =     935
                                                       F(  4,   930) =   44.46
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1588
                                                       Root MSE      =   .3871

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0739592   .0067085    11.02   0.000     .0607937    .0871248
       exper |   .0135097   .0039206     3.45   0.001     .0058154     .021204
         age |   .0109904    .004936     2.23   0.026     .0013034    .0206775
     married |   .2028898   .0429088     4.73   0.000     .1186805    .2870992
       _cons |   5.081911   .1615506    31.46   0.000     4.764865    5.398957
------------------------------------------------------------------------------
```

Comparing this results with the results in the first table we note that the coefficient estimates do not change but the standard errors do, and hence the *t* values are slightly different. If there was more heteroskedasticity in the data, we would probably observe bigger changes.

Clustered standard errors

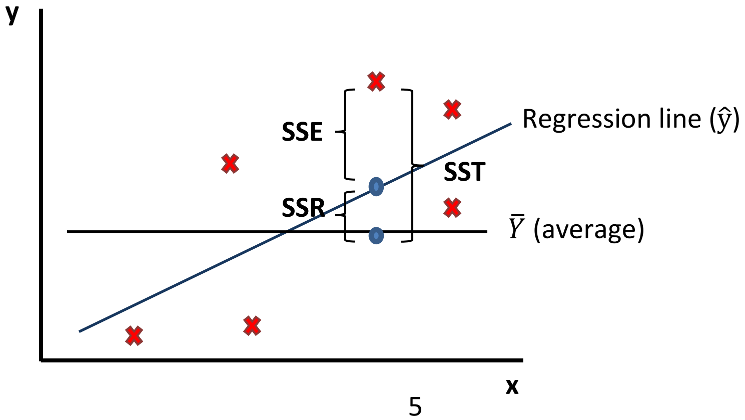
Sometimes, we may impose assumptions on the structure of the heteroskedasticity. For instance, if we suspect that the variance is homoskedastic within a group but not across groups, then we obtain residuals for all observations and calculate average residuals for each group. Then, we have $\hat{\Omega}$ which has a constant for $\hat{u}_j^2$ for group j.

In practice, we usually do not know the structure or source of heteroskedasticity. Thus, it is safe to use the robust standard errors (especially when you have a large sample size). Even if there is no heteroskedasticity, the robust standard errors will become just conventional OLS standard errors. Thus, the robust standard errors are appropriate even under homoskedasticity.

Regression analysis, prediction and specification test

Regression analysis, prediction

To understand how well **X** predicts **y** we evaluate the **variability in the y variable explained by the variable x**:



If the model includes a constant:

Sum squares total (SST): $\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})^2$

Sum squares regression (SSR): $\sum_i (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^2$ ➡ SST=SSR+SSE

Sum squares error (SSE): $\sum_i (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2$

The **coefficient of determination** ($\mathbf{R}^2$) is the proportion of the variability in **y** that is explained by the regression equation:

$$\mathbf{R}^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad \text{where } 0 \leq \mathbf{R}^2 \leq 1 \tag{14}$$

The adjusted $\mathbf{R}^2$ ($\bar{\mathbf{R}}^2$) takes into account that $\mathbf{R}^2$ automatically increases when more explanatory variables are added to the model. $\bar{\mathbf{R}}^2$ can be negative and its value will always be less than or equal to that of $\mathbf{R}^2$.

$$\bar{\mathbf{R}}^2 = 1\text{- } (1\text{- } \mathbf{R}^2)\,\frac{n-1}{n-p-1} \tag{15}$$

where *p*=total number of regressors (not counting the constant term) and *n*=sample size.

EXAMPLE 1 (Wage equation-Continuation) In the previous example, $\mathbf{R}^2$=0.1588 meaning that the 15.88% of the variability in the logarithm of the monthly earnings is explained by the specified regression. $\bar{\mathbf{R}}^2$ =0.1551 meaning than when corrected by the number of explanatory variables included in the model, 15.51% of the variability in the logarithm of the monthly earnings is explained by the regression.

Similarly, we can find the predicted values, the residuals and the relationship between the predicted and the actual values of the previous estimation of equation (8):

```
. predict plwage
(option xb assumed; fitted values)

. predict rlwage, r

. plot lwage plwage

 8.03203 +
         |                            *           *
         |                               *                      *
         |                             *        * *    *       *         *
         |                        * *      * *  *     *** *****  *** ** *
         |       *  *       **  *   *  ** **  **  **  *** ** ******  *  **
         |                  *  **** ********* ** ******************* *  *  *
         |      * *      ***** ******************************* * *  ***   *
    l    |    *       *   * ******************************  **  ***  **
    w    |   * * *       **********************************   *  *   * **
    a    |           *** ************************** ******      *    *  *
    g    | **  *** * * ** ** ****************** ** **** ***  * ***
    e    |  * **  **   * *  *********** ****** ** * * **   **  *
         |    ***  * *   ** ***********     *  ***
         |          * * * *    * *  ***  **  **    **
         |                *     *          **
         |                                    *  * *
         |               *
         |
         |
 4.74493 +                     *
          +----------------------------------------------------------------+
          6.32684                  Fitted values                   7.22284
```

Specification test

The linear regression model specifies that the conditional mean of the dependent variable equals $\mathbf{x}_i'\beta$. We can perform a test to check whether this specification is correct or not. The idea of the test is to regress the dependent variable against its predicted value and the square of the predicted value. If the model is correctly specified, the square of the predicted value should not have much explanatory power. Following the previous example we find that:

EXAMPLE 1 (Wage equation-Continuation)

```
. predict lwagep
(option xb assumed; fitted values)

. gen lwagep2=lwagep^2

. reg lwage lwagep lwagep2

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  2,   932) =   88.22
       Model |  26.3694709     2  13.1847355           Prob > F      =  0.0000
    Residual |  139.286823   932  .149449381           R-squared     =  0.1592
-------------+------------------------------           Adj R-squared =  0.1574
       Total |  165.656294   934  .177362199           Root MSE      =  .38659

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lwagep |  -2.121344   4.569685    -0.46   0.643    -11.08941     6.84672
     lwagep2 |   .2296429   .3361542     0.68   0.495    -.4300639    .8893498
       _cons |   10.59993   15.52473     0.68   0.495    -19.86755    41.06741
------------------------------------------------------------------------------
```

lwagep2 is not statistically significant . This means that we cannot reject the null hypothesis that the model is correctly specified (we can also use the command *test*).

Note that we could have done this test using a single command in Stata after we perform the original regression. The command is *linktest*.

EXERCISE 1(useful to revise t-tests not included in the notes)

A researcher is interested in the effect of school class sizes on the educational attainment of students. She estimates the linear model

$$A_i = \alpha + \beta CS_i \ + {x_i^{S}}' \gamma_S + \ {x_i^{F}}' \gamma_F + u_i \ for \ i = 1,2, \dots, N$$

where A_i is a scalar measure of attainment for student i (e.g. a standardized test score), CS_i is a measure of the average size of classes attended by student i (e.g. average number of students per class), x_i^S is a vector of observations on other characteristics of the school attended by student i (e.g. school size, school location, school type, etc.) and x_i^F is a vector of observations on family background variables for student i (e.g. number of siblings, educational attainment of parents, etc.). The model is estimated by OLS, using a random sample of 5000 students.

i) The researcher reports that the OLS estimate of the parameter β is -0.017 with standard error of 0.01. Use this information to test the null hypothesis that class size has no effect on educational attainment, against a two-sided alternative, at the 5% significance level. What do you conclude?

ii) It is suggested that the effect of class size on educational attainment may be different for boys and girls. Define the zero/one dummy variable Di to be equal to one if student i is a boy, and to be equal to zero if student i is a girl. Interpret the coefficients α^B and α^G in the model

$$A_i = \alpha^B D_i + \alpha^G (1 - D_i) + \beta CS_i \ + {x_i^{S}}' \gamma_S + \ {x_i^{F}}' \gamma_F + u_i$$

Interpret the coefficients β^B and β^G in the model

$$A_i = \alpha^B D_i + \alpha^G (1 - D_i) + \beta^B (D_i CS_i) + \beta^G ((1 - D_i) CS_i) + {x_i^{S}}' \gamma_S + \ {x_i^{F}}' \gamma_F + u_i$$

EXERCISE 2 (practical exercise)

Use the data in ATTEND.RAW ("use http://fmwww.bc.edu/ec-p/data/wooldridge/attend") to answer these questions.

a) To determine the effects of attending a lecture on a final exam performance, estimate a model relating *stndfnl* (the standardized final exam score) to *atndrte* (the percentage of lectures attended). Include the binary variables *frosh* and *soph* as explanatory variables. Interpret the coefficient on *atndrte* and discuss its significance.

b) How confident are you that the OLS estimates from part a) are estimating the causal effect of attendance? Explain.

c) Add *pri*GPA (cumulative GPA score) and *ACT* (ACT score) to the equation. What happens to the coefficient on *atndrte*? Are these additional terms statistically significant?