*Wooldridge, Introductory Econometrics, 4th ed.*

## Chapter 6: Multiple regression analysis: Further issues

What effects will the scale of the $X$ and $y$ variables have upon multiple regression? The coefficients' point estimates are $\partial y/\partial X_j$, so they are in the scale of the data–for instance, dollars of wage per additional year of education. If we were to measure either $y$ or $X$ in different units, the magnitudes of these derivatives would change, but the overall fit of the regression equation would not. Regression is based on correlation, and any linear transformation leaves the correlation between two variables unchanged. The $R^2$, for instance, will be unaffected by the scaling of the data. The standard error of a coefficient estimate is in the same units as the point estimate, and both

will change by the same factor if the data are scaled. Thus, each coefficient's $t-$ statistic will have the same value, with the same $p-$ value, irrespective of scaling. The standard error of the regression (termed "Root MSE" by Stata) is in the units of the dependent variable. The ANOVA F, based on $R^2$, will be unchanged by scaling, as will be all F-statistics associated with hypothesis tests on the parameters. As an example, consider a regression of babies' birth weight, measured in pounds, on the number of cigarettes per day smoked by their mothers. This regression would have the same explanatory power if we measured birth weight in ounces, or kilograms, or alternatively if we measured nicotine consumption by the number of packs per day rather than cigarettes per day.

A corollary to this result applies to a dependent variable measured in logarithmic form. Since

the slope coefficient in this case is an elasticity or semi-elasticity, a change in the dependent variable's units of measurement does not affect the slope coefficient at all (since $\log(cy) = \log c + \log y$), but rather just shows up in the intercept term.

## Beta coefficients

In economics, we generally report the regression coefficients' point estimates when presenting regression results. Our coefficients often have natural units, and those units are meaningful. In other disciplines, many explanatory variables are indices (measures of self-esteem, or political freedom, etc.), and the associated regression coefficients' units are not well defined. To evaluate the relative importance of a number of explanatory variables, it is common to calculate so-called beta coefficients—standardized regression coefficients, from a regression of $y^*$ on $X^*$, where the starred variables have been "z-transformed." This transformation (subtracting the mean and dividing

by the sample standard deviation) generates variables with a mean of zero and a standard deviation of one. In a regression of standardized variables, the (beta) coefficient estimates $\partial y^*/\partial X^*$ express the effect of a one standard deviation change in $X_j$ in terms of standard deviations of $y$. The explanatory variable with the largest (absolute) beta coefficient thus has the biggest "bang for the buck" in terms of an effect on $y$. The intercept in such a regression is zero by construction. You need not perform this standardization in most regression programs to compute beta coefficients; for instance, in Stata, you may just use the `beta` option, e.g. `regress lsalary years gamesyr scndbase, beta` which causes the beta coefficients to be printed (rather than the 95% confidence interval for each coefficient) on the right of the regression output.

## Logarithmic functional forms

Many econometric models make use of variables measured in logarithms: sometimes the dependent variable, sometimes both dependent and independent variables. Using the "double-log" transformation (of both $y$ and $X$) we can turn a multiplicative relationship, such as a Cobb-Douglas production function, into a linear relation in the (natural) logs of output and the factors of production. The estimated coefficients are, themselves, elasticities: that is, $\partial \log y / \partial \log X_j$, which have the units of percentage changes. The "single-log" transformation regresses $\log y$ on $X$, measured in natural units (alternatively, some columns of $X$ might be in logs, and some columns in levels). If we are interpreting the coefficient on a levels variable, it is $\partial \log y / \partial X_j$, or approximately the percentage change in $y$ resulting from a one unit change in $X$. We often use this sort of model to estimate an exponential trend–that is, a growth rate–since if the

$X$ variable is $t$, we have $\partial \log y / \partial t$, or an estimate of the growth rate of $y$. The interpretation of regression coefficients as percentage changes depends on an approximation, that $\log(1 + x) \approx x$ for small $x$. If $x$ is sizable– and we seek the effect for a discrete change in $x-$ then we must take care with that approximation. The exact percentage change, $\%\Delta y = 100 \left[\exp\left(b_j \Delta X_j\right) - 1\right]$, will give us a more accurate prediction of the change in $y$.

Why do so many econometric models utilize logs? For one thing, a model with a log dependent variable often more closely satisfies the assumptions we have made for the classical linear model. Most economic variables are constrained to be positive, and their empirical distributions may be quite non-normal (think of the income distribution). When logs are applied, the distributions are better behaved. Taking logs also reduces the extrema in the

data, and curtails the effects of outliers. We often see economic variables measured in dollars in log form, while variables measured in units of time, or interest rates, are often left in levels. Variables which are themselves ratios are often left in that form in empirical work (although they could be expressed in logs; but something like an unemployment rate already has a percentage interpretation). We must be careful when discussing ratios to distinguish between an 0.01 change and a one unit change. If the unemployment rate is measured as a decimal, e.g. 0.05 or 0.06, we might be concerned with the effect of an 0.01 change (a one per cent increase in unemployment)–which will be 1/100 of the regression coefficient's magnitude!

## Polynomial functional forms

We often make use of polynomial functional forms–or their simplest form, the quadratic–to

represent a relationship that is not likely to be linear. If $y$ is regressed on $x$ and $x^2$, it is important to note that we must calculate $\partial y/\partial x$ taking account of this form–that is, we cannot consider the effect of changing $x$ while holding $x^2$ constant. Thus, $\partial y/\partial x = b_1 + 2b_2x$, and the slope in $\{x, y\}$ space will depend upon the level of $x$ at which we evaluate the derivative. In many applications, $b_1 > 0$ while $b_2 < 0$, so that while $x$ is increasing, $y$ is increasing at a decreasing rate, or levelling off. Naturally, for sufficiently large $x$, $y$ will take on smaller values, and in the limit will become negative; but in the range of the data, $y$ will often appear to be a concave function of $x$. We could also have the opposite sign pattern, $b_1 < 0$ while $b_2 > 0$, which will lead to a U-shaped relation in the $\{x, y\}$ plane, with $y$ decreasing, reaching a minimum, and increasing–somewhat like an average cost curve. Higher-order polynomial terms may also be used, but they are not as commonly found in empirical work.

## Interaction terms

An important technique that allows for nonlinearities in an econometric model is the use of **interaction terms**–the product of explanatory variables. For instance, we might model the house `price` as a function of `bdrms`, `sqft`, and `sqft`$\cdot$ `bdrms`, which would make the partial derivatives with respect to each factor depend upon the other. For instance, $\partial price/\partial bdrms = b_{bdrms} + b_{sqft \cdot bdrms} sqft$, so that the effect of an additional bedroom on the price of the house also depends on the size of the house. Likewise, the effect of additional square footage (e.g. an addition) depends on the number of bedrooms. Since a model with no interaction terms is a special case of this model, we may readily test for the presence of these nonlinearities by examining the significance of the interaction term's estimated coefficient. If it is significant, the interaction term is needed to capture the relationship.

**Adjusted $R^2$**

In presenting multiple regression, we established that $R^2$ cannot decrease when additional explanatory variables are added to the model, even if they have no significant effect on $y$. A "longer" model will always appear to be superior to a "shorter" model, even though the latter is a more parsimonious representation of the relationship. How can we deal with this in comparing alternative models, some of which may have many more explanatory factors than others? We can express the standard $R^2$ as:

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

Since all models with the same dependent variable will have the same $SST$, and $SSR$ cannot increase with additional variables, $R^2$ is a non-decreasing function of $k$. An alternative measure, computed by most econometrics packages, is the so-called "R-bar-squared" or 'Adjusted $R^2$" :

$$\bar{R}^2 = 1 - \frac{SSR/\left(n-(k+1)\right)}{SST/\left(n-1\right)} \qquad (2)$$

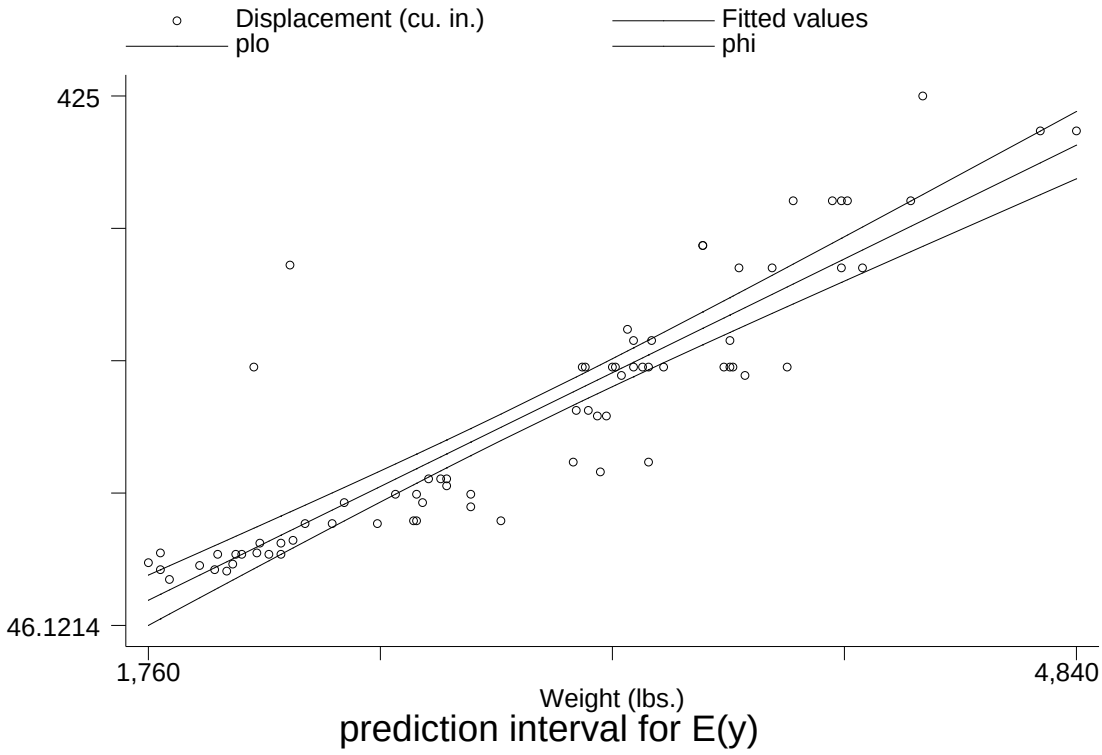
where the numerator and denominator of $R^2$ are divided by their respective degrees of freedom (just as they are in computing the mean squared measures in the ANOVA F table). For a given dependent variable, the denominator does not change; but the numerator, which is $s^2$, may rise or fall as $k$ is increased. An additional regressor uses one more degree of freedom, so $(n-(k+1))$ declines; and $SSR$ declines as well (or remains unchanged). If $SSR$ declines by a larger percentage than the degrees of freedom, then $\bar{R}^2$ rises, and vice versa. Adding a number of regressors with little explanatory power will increase $R^2$, but will decrease $\bar{R}^2-$ which may even become negative! $\bar{R}^2$ does not have the interpretation of a squared correlation coefficient, nor of a "batting average" for the model. But it may be

used to compare different models of the same dependent variable. Note, however, that we cannot make statistical judgments based on this measure; for instance, we can show that $\bar{R}^2$ will rise if we add one variable to the model with a $|t| > 1-$ but a $t$ of unity is never significant. Thus, an increase in $\bar{R}^2$ cannot be taken as meaningful (the coefficients must be examined for significance) but, conversely, if a "longer" model has a lower $\bar{R}^2$, its usefulness is cast in doubt. $\bar{R}^2$ is also useful in that it can be used to compare non-nested models– i.e. two models, neither of which is a proper subset of the other. A "subset F" test cannot be used to compare these models, since there is no hypothesis under which the one model emerges from restrictions on the other, and vice versa. $\bar{R}^2$ may be used to make informal comparisons of non-nested models, as long as they have the same dependent variable. Stata presents the $\bar{R}^2$ as the "`Adj R-squared`" on the regression output.

## Prediction and residual analysis

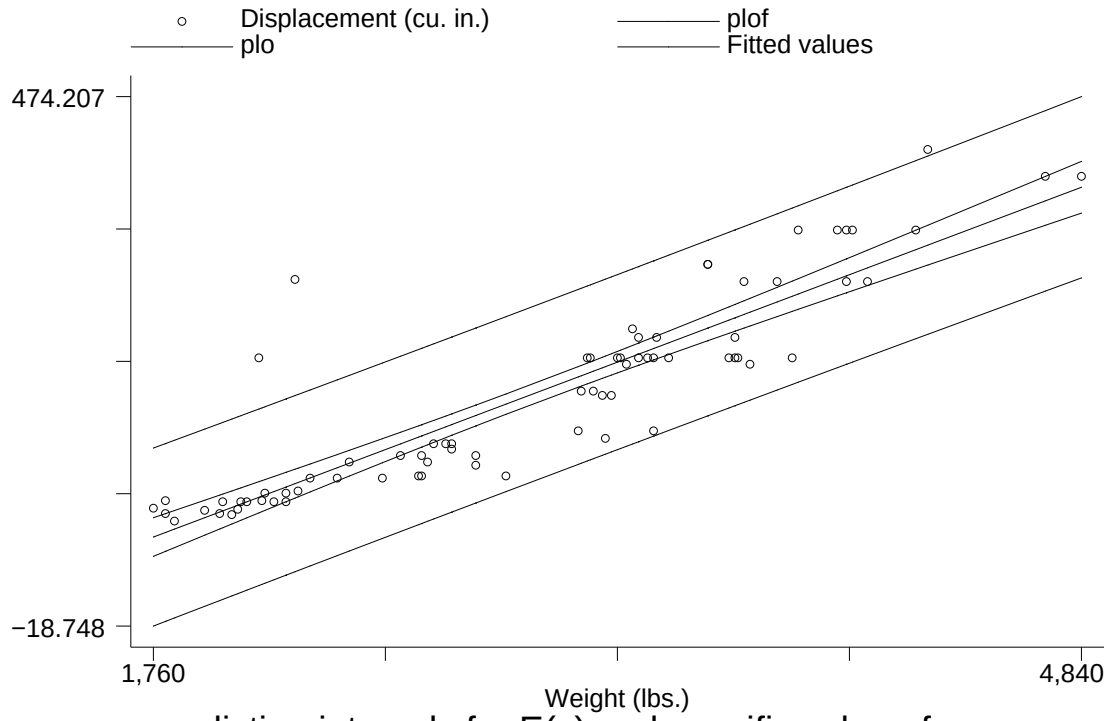
The predictions of a multiple regression are, simply, the evaluation of the regression line for various values of the explanatory variables. We can always calculate $\widehat{y}$ for each observation used in the regression; these are known as "in-sample" or "ex post" predictions. Since the estimated regression equation is a function, we can evaluate the function for any set of values $\{X_1^0, X_2^0, ..., X_k^0\}$ and form the associated point estimate $\widehat{y}^0$, which might be termed an "out-of-sample" or "ex ante" forecast of the regression equation. How reliable are the forecasts of the equation? Since the predicted values are linear combinations of the $b$ values, we can calculate an **interval estimate** for the predicted value. This is the confidence interval for $E\left(y^0\right)$ : that is, the average value that would be predicted by the model for a specific set of $X$ values. This may be calculated after

any regression in Stata using the `predict` command's `stdp` option: that is, `predict stdpred, stdp` will save a variable named "`stdpred`" containing the standard error of prediction. The 95% confidence interval will then be, for large samples, $\{\widehat{y} - 1.96 stdpred, \widehat{y} + 1.96 stdpred\}$. An illustration of this confidence interval for a simple regression is given here. Note that the confidence intervals are parabolic, with the minimum width interval at $\bar{X}$, widening symmetrically as we move farther from $\bar{X}$. For a multiple regression, the confidence interval will be narrowest at the multivariate point of means of the $X's$.




However, if we want a confidence interval for a specific value of $y-$ rather than for the mean of $y-$ we must also take into account the fact that a predicted value of $y$ will contain an error, $u$. On average, that error is assumed to be zero; that is, $E(u) = 0$. For a specific value of $y$, though, there will be an error $u_i$; we do not know its magnitude, but we have estimated

that it is drawn from a distribution with standard error $s$. Thus, the **standard error of forecast** will include this additional source of uncertainty, and confidence intervals formed for specific values of $y$ will be wider than those associated with predictions of the mean $y$. This standard error of forecast series can be calculated, after a regression has been estimated, with the `predict` command, specifying the `stdf` option. If the variable stdfc is created, the 95% confidence interval will then be, for large samples, $\{\widehat{y} - 1.96 stdfc, \widehat{y} + 1.96 stdfc\}$. An illustration of this confidence interval for a simple regression is given here, juxtaposed with that shown earlier for the standard error of prediction. As you can see, the added uncertainty associated with a draw from the error distribution makes the prediction interval much wider.



prediction intervals for E(y) and specific value of y

## Residual analysis

The OLS residuals are often calculated and analyzed after estimating a regression. In a purely technical sense, they may be used to test the validity of the several assumptions that underly the application of OLS. When plotted, do they appear systematic? Does their dispersion appear to be roughly constant, or is

it larger for some $X$ values than others? Evidence of systematic behavior in the magnitude of the OLS residuals, or in their dispersion, would cast doubt on the OLS results. A number of formal tests, as we will discuss, are based on the residuals, and many graphical techniques for examining their randomness (or lack thereof) are available. In Stata, `help regression diagnostics` discusses many of them.

The residuals are often used to test specific hypotheses about the underlying relationship. For instance, we could fit a regression of the salaries of employees of XYZ Corp. on a number of factors which should relate to their salary level: experience, education, specific qualifications, job level, and so on. Say that such a regression was run, and the residuals retrieved. If we now sort the residuals by factors not

used to explain salary levels, such as the employee's gender or race, what will we find? Under nondiscrimination laws, there should be no systematic reason for women to be paid more or less than men, or blacks more or less than whites, after we have controlled for these factors. If there are significant differences between the average residual for, e.g., blacks and whites, then we would have evidence of "statistical discrimination." Regression equations have often played an important role in investigating charges of discrimination in the workplace. Likewise, most towns' and cities' assessments of real estate (used to set the tax levy on that property) are performed by regression, in which the explanatory factors include the characteristics of a house and its neighborhood. Since many houses will not have been sold in the recent past, the regression must be run over a sample of houses that have been sold, and out-of-sample predictions used to estimate the appropriate price for a house that

has not been sold recently, based on its attributes and trends in real estate transactions prices in its neighborhood. A mechanical evaluation of the fair market value of the house may be subject to error, but previous methods used—in which knowledgeable individuals attached valuations based on their understanding of the local real estate market—are more subjective.