Multiple Regression Analysis: Further Issues

ECONOMETRICS (ECON 360)

BEN VAN KAMMEN, PHD

Introduction

Estimation (Ch. 3) and inference (Ch. 4) are the 2 fundamental skills one must perform when using regression analysis.

This chapter adds a few embellishments to OLS estimation and inference and reveals that it is not very limited by being linear in parameters.

- Sometimes this suggests that OLS is limited to estimating constant effects, which is emphatically not true.
- Here we examine cases in which the form of the relationship between *x* and *y* is more exotic, e.g., quadratic shaped or dependent on the value of another regressor.

It also critically examines the R squared statistic and its usefulness for specifying the model.

Finally it extends inference to the predicted ("fitted") values of *y* that come from the estimates.

Outline

Effects of Data Scaling on OLS Statistics.

More on Functional Form:

- Logarithmic,
- Quadratics,
- Interactions.

More on Goodness-of-Fit and Selection of Regressors.

Prediction and Residual Analysis.

Effects of data scaling on OLS statistics

Consider the following regression model.

$$lcrimes = \beta_0 + \beta_1 pop + \beta_2 pcinc + \beta_3 llawexpc + u,$$

in which the unit of observation is U.S. Cities in 1982, and the variables are defined:

crimes: number of crimes in city "i" in 1982,

pop: city population level,

pcinc: per capita income,

llawexpc: law enforcement spending, $ per capita.

Data scaling (continued)

2 of the coefficients (and their standard errors, too) are very small in magnitude, yet they are also both statistically significant.

- Also check out the positive estimate on log of law enforcement expenditure!

But the main thing is the unpleasantly small scale of the coefficients on population and per capita income.

- The former is so small STATA expresses it in scientific notation, and if you rounded the latter to 3 digits beyond the decimal, it wouldn't even round to |0.001|.

Dependent Variable: lcrimes	$\widehat{\beta}_j$	**Std. Err.**	**t**	**P value**
pop	2.48e-06	2.05e-07	12.11	0.000
pcinc	-.0001009	.0000445	-2.27	0.029
llawexpc	.3733494	.2081494	1.79	0.080
Constant	7.580081	1.44403	5.25	0.000
N=46; R^2=0.8127				

The estimates of the model using OLS and 46 cities (above).

Data scaling (continued)

Of course it isn't surprising. Increasing *population* by a single person should not make a big difference on the crime rate; nor should increasing the average income by $1.

- That's why the marginal effects are so small in magnitude.
- They are still highly useful for explaining *crime*; they just have too small a scale.

This is easily remedied without disrupting any of the statistical inference or the integrity of the regression. Merely change the units of measure to make the scale of the coefficients appropriately large.

- In this instance, it is convenient to change the scale on both "x" variables by expressing *population* in 100s of 1000s and *per capita income* in $1000s.
- This is equivalent to modifying the regression model as follows (next slide).

Data scaling (continued)

$$y = \beta_0 + \beta_1 \frac{x_1}{100000} + \beta_2 \frac{x_2}{1000} + \beta_3 x_3 + u.$$

OLS should now produce coefficients on these variables that are 100,000 and 1000 times as large, but with identical t statistics, R^2, and residual variance.

- Derivation.

The estimates using the scaled regressors is on the table on the next slide.

Data scaling (concluded)

The coefficient estimates and standard errors now have much more palatable scales, but notably different interpretations.

They represent, respectively, the effects of increasing *population* by 100,000 and of increasing *per capita income* by $1000.

The t statistics don't change because, for each variable, the interpretations of β, $\hat{\beta}$, and the standard error have all been scaled by the same constant, and this constant, e.g., 1000, just cancels out when computing the t statistic.

Dependent Variable: lcrimes	$\widehat{\boldsymbol{\beta}}_j$	**Std. Err.**	**t**	**P value**
Pop (100,000s)	.248315	.0205103	12.11	0.000
Pcinc ($1000s)	-.100899	.0445085	-2.27	0.029
llawexpc	.3733494	.2081494	1.79	0.080
Constant	7.580081	1.44403	5.25	0.000
N=46; R^2=0.8127				

Standardized coefficients

Some regressors have units of measure that are inherently vague or otherwise hard to interpret, such as scores on a standardized test like the SAT or a credit score.

For these variables, interpreting the effect (on y) of a one unit increase is non-obvious.

- It is easier to think about the effect of moving *one standard deviation* within the distribution instead.

To achieve this interpretation for the regression, simply perform OLS on *standardized* variables, i.e.,

$$\frac{y-\bar{y}}{\hat{\sigma}_y} = \beta_1 \frac{(x_1-\bar{x}_1)}{\hat{\sigma}_1} + \beta_2 \frac{(x_2-\bar{x}_2)}{\hat{\sigma}_2} + \ldots + \beta_k \frac{(x_k-\bar{x}_k)}{\hat{\sigma}_k} + u, \text{where}$$

$$\hat{\sigma} \text{ is the sample standard deviation of the variable.}$$

Standardized coefficients (continued)

But in order to preserve the equality of the model, you have to divide the RHS by the standard deviation of y and balance each term out by multiplying by standard deviation ($\hat{\sigma}_x$):

$$\frac{y-\bar{y}}{\hat{\sigma}_y} = \frac{\hat{\sigma}_1}{\hat{\sigma}_y}\beta_1\frac{(x_1-\bar{x}_1)}{\hat{\sigma}_1} + \frac{\hat{\sigma}_2}{\hat{\sigma}_y}\beta_2\frac{(x_2-\bar{x}_2)}{\hat{\sigma}_2} + \ldots + \frac{\hat{\sigma}_k}{\hat{\sigma}_y}\beta_k\frac{(x_k-\bar{x}_k)}{\hat{\sigma}_k} + \frac{u}{\hat{\sigma}_y}.$$

This changes the interpretation of the coefficients. The estimate output for standardized x_1 will now contain:

$$\frac{\hat{\sigma}_1}{\hat{\sigma}_y}\hat{\beta}_1$$

all as one quantity.

Standardized coefficients (continued)

The Wooldridge book denotes this, $\hat{b}_1$. So to estimate standardized coefficients, OLS estimates

$$z_y = \hat{b}_1 \frac{(x_1 - \bar{x}_1)}{\hat{\sigma}_1} + \hat{b}_2 \frac{(x_2 - \bar{x}_2)}{\hat{\sigma}_2} + \ldots + \hat{b}_k \frac{(x_k - \bar{x}_k)}{\hat{\sigma}_k} + \frac{u}{\hat{\sigma}_y},$$

and the effect of increasing the j^{th} regressor by one standard deviation is $\hat{b}_j$ standard deviations.

So if x_1 is an individual's (in the sample) credit score, the regression using standardized ("beta") coefficients estimates the effect of moving one standard deviation up in the distribution of credit scores.

Standardized coefficients (concluded)

STATA can estimate standardized coefficients really easily using the same syntax as a normal regression—and without actually transforming all the variables using

egen [newvar_z]=sd(oldvar), mean(0) std(1).

The syntax for doing it in the regression is:

reg [depvar] {indvars . . . }, beta.

Logarithmic functional forms

We discussed models with a (natural) logarithmically transformed variable in the simple regression chapter. So this should serve as a reminder.

The coefficient in a regression with a log-transformed y variable should be interpreted as the percentage change in y for a 1 unit increase in x.

This is an approximation, though, that is only really valid for "small" changes in x.

$$\frac{\partial \ln(y)}{\partial x_j} = \frac{\partial \ln(y)}{\partial y}\frac{\partial y}{\partial x_j} = \frac{1}{y}\frac{\partial y}{\partial x_j} = \hat{\beta}_j; \frac{\Delta y}{y} \approx \hat{\beta}_j \Delta x_j.$$

$$\Delta x_j = 1 \rightarrow \hat{\beta}_j \approx \%\Delta y.$$

Logarithmic functional forms (continued)

The approximation gets less accurate the larger the change in x_j is. To be exact about interpreting a one unit increase in x_j, consider the definition of % change:

$$(1)\ \%\Delta y = 100 * \frac{y(x_{j1}) - y(x_{j0})}{y(x_{j0})} \Leftrightarrow 1 + \frac{\%\Delta y}{100} = \frac{y(x_{j1})}{y(x_{j0})}.$$

Imagine this change was the result of differencing the regression model (with y in logs) by changing x_j one unit ($\Delta x_j = 1$):

$$(2)\ \ln\left(y(x_{j1}|x_{\neq j})\right) - \ln\left(y(x_{j0}|x_{\neq j})\right) = \hat{\beta}_j * 1.$$

Logarithmic functional forms (continued)

Take the log of (1):

$$(1) \rightarrow \ln\left(1 + \frac{\%\Delta y}{100}\right) = \ln\left(y\left(x_{j1}|x_{\neq j}\right)\right) - \ln\left(y\left(x_{j0}|x_{\neq j}\right)\right) = \hat{\beta}_j.$$

To solve for the percentage change, take the "anti-log" of both sides, i.e., make it such that taking the log of both sides results in the line above.

$$\left(1 + \frac{\%\Delta y}{100}\right) = \exp\left(\hat{\beta}_j\right) \Leftrightarrow \%\Delta y = 100 * \left[\exp\left(\hat{\beta}_j\right) - 1\right].$$

- The difference is a lot like calculating an actual elasticity in theory class—in which the starting point matters—compared to an "arc" elasticity (mid-point formula).

More about logs

1. Rescaling doesn't matter; you'll always get the same estimates and inference (except the intercept) because rescaling entails multiplying by a constant. The constant will disappear into the intercept when you take the log.

$$x_{rescale} = constant * x; \ln(x_{rescale}) = \ln(constant) + \ln(x).$$

2. Dependent variables in log form are often more likely to satisfy (at least approximate) a normal distribution, conditional on x. When y is strictly positive, the distribution is truncated and often skewed. Taking the log helps with both problems.
   - See Chapter 5, Exercise C.4.
   - It also helps with heteroskedasticity.

Even more about logs

3. Taking the log *narrows* the range of some variables, which can help reduce the power of outliers.
   - Useful when you have right-skewed variables, such as income, population, or fantasy football points.
4. Taking the log *widens* the range of other variables, e.g., ones that have a natural bound between 0 and 1, or 1 to 10.
   - For these it is inappropriate to make log transformations.
   - Obviously the same goes for variables that take negative values.
   - The same goes for values that take on *nonnegative* values, i.e., $0 \leq y$. But you can deal with that by taking the log of $(1 + y)$ instead, with little consequence.
   - **Don't just exclude observations that have a value of zero and estimate using the others though!**

Still more about logs (I promise this is the last one . . . for now)

5. Variables measured in years (any unit of time) are typically not expressed in logs.
   - It isn't very natural to think about a "percentage change" in units of time, e.g., "Alan has 8% more experience than Jasper" is a difficult statement to interpret.

If there are *a lot* of zeroes in the data for the dependent variable, a different kind of model should be applied in lieu of OLS regression. See Chapter 17.

Models with quadratics

Another simple way to enlarge the capabilities of OLS is transforming the explanatory variables to allow for non-constant, even non-monotonic, effects.

This can be accomplished by adding the square of an x variable to the regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u.$$

Models with quadratics (continued)

Though the regression only has 1 explanatory variable, its effect is not a single coefficient, and it is more appropriate to think of this as a multiple regression (instead of a simple one).

The estimates from OLS reveal the partial effect of x_1, which necessarily entails changing x_1^2 as well. Simple calculus shows that this is:

$$\frac{\partial E(y|x_1)}{\partial x_1} = \hat{\beta}_1 + 2\hat{\beta}_2 x_1 \approx \frac{\Delta E(y|x_1)}{\Delta x_1}, \text{for discrete changes in } x_1.$$

Models with quadratics (continued)

The reason for including quadratic (or higher polynomials, which work the same way) terms is to examine the form of the relationship.

- Does the effect on *y* increase in magnitude with x_1?
- Does it diminish?
- Is there a minimum or maximum, beyond which the effect changes sign?
- $\hat{\beta}_1$ and $\hat{\beta}_2$ tell you all you need to know to answer these questions.

If $\hat{\beta}_2$ is positive, for instance, the shape is an upward-opening parabola ("U") and will have a minimum; if it's negative, it is a downward-opening ("inverted U") parabola and will have a maximum.

If $\hat{\beta}_1$ and $\hat{\beta}_2$ are opposite-signed, the minimum/maximum occurs in the positive interval, as in the figure from Wooldridge below.

Models with quadratics (continued)

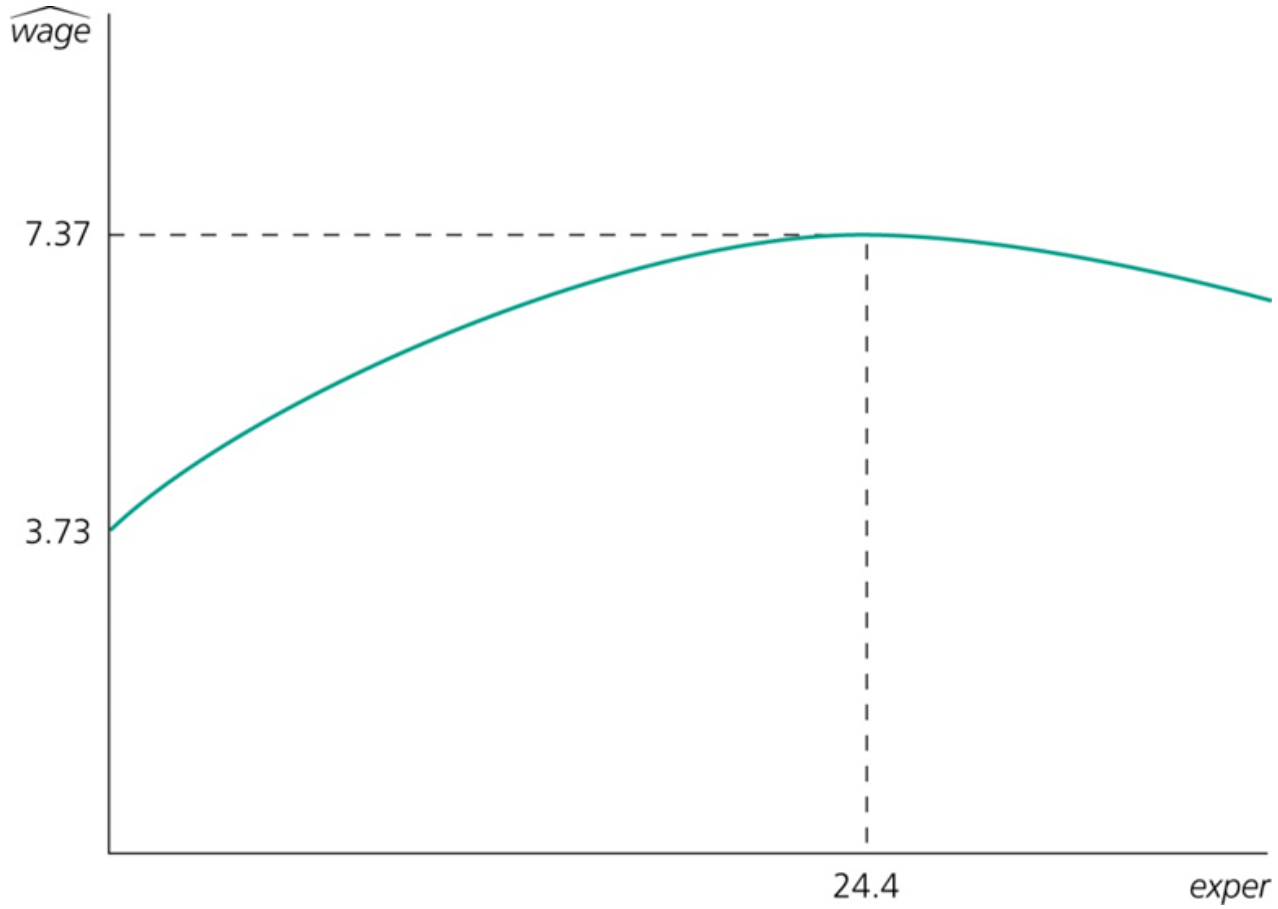
The reason for including quadratic (or higher polynomials, which work the same way) terms is to examine the form of the relationship.

- Does the effect on *y* increase in magnitude with x_1?
- Does it diminish?
- Is there a minimum or maximum, beyond which the effect changes sign?
- $\hat{\beta}_1$ and $\hat{\beta}_2$ tell you all you need to know to answer these questions.

If $\hat{\beta}_2$ is positive, for instance, the shape is an upward-opening parabola ("U"); if it's negative, it is a downward-opening ("inverted U") parabola.

If $\hat{\beta}_1$ and $\hat{\beta}_2$ are opposite-signed, the min./max. occurs in the positive interval, as in the figure from Wooldridge (next slide).

From Wooldridge: quadratic relationship between wage and experience



Models with quadratics (continued)

This is because the min./max. is the point at which the function is flat, i.e., has a slope of zero.

- $\hat{\beta}_1 + 2\hat{\beta}_2 x_1$ is the function that expresses the slope as a function of x_1, so you can solve for the point at which slope is zero by setting the partial effect equal to zero and solving for x_1.

$$\hat{\beta}_1 + 2\hat{\beta}_2 x_1 = 0 \rightarrow x_1^* = -\frac{\hat{\beta}_1}{2\hat{\beta}_2};\ x_1^* \text{ denotes the min. or max. location.}$$

$$x_1^* > 0 \text{ if } \hat{\beta}_1, \hat{\beta}_2 \text{ have opposite signs.}$$

Models with quadratics (continued)

Sometimes both sides (upward and downward sloping) of the parabola are interesting.

If the model is correctly specified (!), and you have observations that fall on both sides of the min/max, it could reveal non-monotonicity: initially x_1 is "good" for the outcome, y, but beyond some limit it is "bad," and is associated with a decrease in y.

An example is the quantity-quality trade-off for parents when they decide how many kids to have. To be concrete, think of the relationship in terms of the regression,

$$kids = \beta_0 + \beta_1 income + \beta_2 income^2 + u.$$

Models with quadratics (continued)

It is foreseeable that *income* has an inverted U-shaped relationship with the number of *kids* a couple has.

- For low levels of income, more income implies capacity to feed and clothe children and thus, more kids.
- As income increases, however, there are other opportunity costs that become greater, i.e., the parents' time is more valuable. It is theorized* that child quality (spending *per child*) responds to higher incomes as well, and can even outweigh the income elasticity for quantity. As such, child quality is a "luxury good" and well-off parents can be expected to spend much more on quality than quantity—leading to smaller families for high earning parents (with very high spending per child).
- This would be consistent with the downward-opening parabolic shape for the relationship, i.e.,

$$\hat{\beta}_2 < 0 \; and \; \hat{\beta}_1 > 0.$$

*Becker, Gary. 1960. "An Economic Analysis of Fertility." In *Demographic and Economic Change in Developing Countries*, Princeton University Press. Accessed from http://www.nber.org/chapters/c2387.pdf [7-11-2013].

Models with quadratics (concluded)

The text gives another example where, despite estimating a (U shaped) parabolic relationship, one side of the parabola is irrelevant because almost no observations lie in that range.

It is appropriate to infer a monotonic relationship (increasing in this case, at an increasing rate) over the relevant range when this is true.

- This also goes for instances in which both coefficients have the same sign and the dependent variable takes on only nonnegative values: the min/max is guaranteed to be in the irrelevant (negative) range.

Finally if one wants specific estimates of the effect on y, he can simply evaluate it for some value of x_1, e.g., the median or one or more quantiles in the distribution.

Models with interaction terms

One of the most impressive aspects of regression analysis is its capacity to estimate non-constant effects, like semi-elasticities and parabolic relationships.

Add to this list of capacities, interaction effects: in which the effect of one regressor depends on the value of another regressor.

Such a model, in the simplest case, would look like:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

Models with interaction terms (continued)

Now the partial effect of x_1 depends on x_2.

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2; \text{also } \frac{\Delta y}{\Delta x_2} = \beta_2 + \beta_3 x_1.$$

As with quadratic terms, expressing the specific effect of a variable requires evaluating it at a particular value of the other one in the interaction, e.g., the median or one or several quantiles.

Models with interaction terms (continued)

Just looking at the sign on the interaction, though, provides useful insight into the shape of the relationship.

An interaction term between *education* and *experience* in an *earnings* regression ($y = \log(earnings)$) reveals whether more schooling increases the return to experience.

- It does if the coefficient on the interaction is positive.

If the model is:

$$\log(earnings) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ * exper + u,$$

the return to experience is $R_{exper} = (\hat{\beta}_2 + \hat{\beta}_3 educ)$.

Models with interaction terms (continued)

$\hat{\beta}_2$ can be thought of as the return to a year of experience if the individual has no ($educ = 0$) education.

- If $\hat{\beta}_3$ is positive, the return increases the more education an individual has.

Formally,

$$\frac{\partial R_{exper}}{\partial educ} = \hat{\beta}_3;\ \hat{\beta}_3 > 0 \rightarrow \text{return to experience increases with education.}$$

Interactions in Stata

Adding quadratic terms and interactions is easy using the **fvvarlist** feature and without having to generate a bunch of new variables explicitly.

The syntax for an interaction is:

reg [depvar] {noninteracted indvars} c.[variable1]##c.[variable2],

and Stata will perform the interaction plus include the linear terms for both variables 1 and 2.

This is because of the double pound signs, which mean “full interaction”.

- The “c” in front of each variable is for “continuous” to distinguish the treatment from “i” for “indicators”—which is something we will discuss in a subsequent lesson.

While we're at it . . . quadratic terms in Stata

A quadratic is just a special case of an interaction, in which variable1 is identical to variable2:

reg [depvar] {noninteracted indvars} c.[variable1]##c.[variable1].

More on goodness-of-fit and selection of regressors

Interpreting the results of a regression takes the forms of testing hypotheses about the coefficients, inquiring about the validity of the Gauss-Markov Assumptions, and asking whether an estimated effect can properly be interpreted as *causal*.

None of these questions explicitly depends on having a particularly large R^2. A high R^2 does not imply a causal relationship; nor does a low R^2 mean that your hypothesis tests are invalid.

- If the regression is using experimental data, the variation in x is exogenous by design, so it doesn't matter how many other unobserved factors enter the model (driving down R^2); the evidence should be interpreted as causal (see the example in the text involving apples, page 201).

Goodness-of-fit

Low R^2 results from a large error variance, though, so this can make the standard errors from a regression large—and inference more difficult.

- But this can be overcome with a larger sample size.
- And it is also more difficult to predict precise values of the dependent variable because the errors are large and most of the predictors are not included in the model.
- Lastly *F* tests are based on changes in R^2 when variables are added, so it does have an instrumental role in inference for that reason.

But the size of R^2 is not particularly important except for these reasons.

Adjusted R squared

A technical aspect of R^2 that we have not made explicit yet is that it is a sample statistic that estimates the population R-Squared,

$$\rho^2 \equiv 1 - \frac{\sigma_u^2}{\sigma_y^2}.$$

The "sigma" terms are the variances of the error and y, respectively, which are estimated by:

$$\hat{\sigma}_u^2 = \frac{SSR}{n-k-1}, \text{s.t.}, E(\hat{\sigma}_u^2) = \sigma_u^2 \text{ and } \hat{\sigma}_y^2 = \frac{SST}{n-1}, \text{s.t.}, E\left(\hat{\sigma}_y^2\right) = \sigma_y^2.$$

Adjusted R squared (continued)

SSR is the total sum of squares of the residuals; SST is the total sum of squares of y.

- These are the unbiased estimators of both population variances.

In order for R^2 to reflect these estimates, it ought to be calculated with the degrees of freedom included.

- This expression differs from the original R^2, which does not include the degrees of freedom adjustments.

$$R^2 = 1 - \frac{SSR}{SST}, \text{whereas } \bar{R}^2 = 1 - \frac{SSR}{SST}\frac{(n-1)}{(n-k-1)}$$

Adjusted R squared (concluded)

The version that adjusts for degrees of freedom used by the (k) variables in the regression is called adjusted R-Squared.

Whereas regular R^2 always increases when more regressors are added, $\bar{R}^2$ does not.

- In both cases the SSR decreases when more (relevant) regressors are added.
- But only adjusted R^2 accounts for the increase in k. Thus there is a "penalty" for adding more explanatory variables. Adjusted R^2 will only increase if the added variable explains "enough" variation in y to justify its inclusion in the model.

STATA reports both variations by default when you run any regression, and the two statistics can readily be converted if you know n and k.

Using adjusted R squared to choose between non-nested models

One application of adjusted R^2 is to choosing between two competing models for the effect of an x variable, e.g., one with a logarithmic specification and one with a quadratic.

Comparing regular R^2 for the two models disadvantages the more "parsimonious" model (logs) with only one variable; however, comparing $\bar{R}^2$ would be a fair test because it adjusts the quadratic version's statistic based on 1 less degree of freedom.

Adjusted R squared and non-nested models (continued)

This comparison can also be useful for choosing between two highly collinear covariates, such as an individual's years of *potential* labor force experience ($age - schooling - 6$) and his actual years of experience (self-reported in a survey).

- These specifications constitute a pair of non-nested models, in which neither is a special case of the other.

The *F* test for exclusion restrictions can tell you whether the two specifications for experience *jointly* belong in the regression, but it doesn't tell you which one is better *individually*.

Comparing $\bar{R}^2$ between the two competing specifications can resolve this uncertainty in favor of the one with higher $\bar{R}^2$.

Controlling for too many factors in regression analysis

An empiricist is right to worry that he has excluded a relevant variable from his analysis—particularly one that violates Assumption MLR. 4 and biases the estimates.

It is possible to veer in to excess in the other direction as well by over controlling.

- This is the name for including too many variables in the regression, to the point at which the *ceteris paribus* interpretation of the effects is obscured.

Consider the following cautionary examples that can be classified as over controlling.

Examples of over controlling

1. Testing the effect of a policy change, e.g., different rates of beer taxation on traffic fatalities, and controlling separately for the mechanism (beer consumption) through which the policy operates.
   - "The effect of the tax, holding consumption constant" is an uninteresting result compared to "the effect of the tax, given that consumption decreases accordingly."
2. Controlling for something that is a subset of the dependent variable.
   - Doctor visits are one component of health expenditures. Regressing health expenditures on measures of health risks, and controlling for doctor visits, relegates the estimated effects to explaining only the portion of health expenditures on prescriptions and other non-doctor-visit medicine, and it would be curious why they would have such an interpretation without more elaboration.

Examples of over controlling (concluded)

3. The purpose of the estimation is another factor. In a model of house prices as a function of their attributes, the goal is to put prices on attributes that are not unbundled from the whole house and sold separately with explicit prices, e.g., there is no market for a "3rd bedroom" sold separately from a whole house and, hence, no price. To this end, there is confusion if the empiricist controls for the house's assessed value in addition to its attributes. If you're trying to see how much marginal value the attributes *actually* contribute, it is unnecessary to hold the house's value constant.
    - If the purpose of the regression was testing the accuracy of the assessments, of course, it would be necessary to include the assessed value and interesting to see if its 1:1 predicted relationship is robust to controlling for attributes of the houses.

Adding regressors to reduce error variance

There is a trade-off faced when adding regressors to a model.

- On one hand it reduces the error variance, but on the other it “partials out” more of the variation in the other regressors and adds to multicollinearity.

Unless it doesn’t. If a regressor is uncorrelated with all the others in which one is interested, it should definitely be included in the model because the multicollinearity trade-off doesn’t exist, and it just increases the precision of the estimates.

Prediction and residual analysis

Confidence Intervals for Predictions

After estimates are made using OLS, the model can be "fitted" by evaluating it at each observation using the estimated coefficients and intercept.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k, \text{where}$$

$\hat{y}$ is the "fitted value", "prediction" or "expected value of y, conditional on a set of x".

Its variance is a complex function of variances and covariances, but estimating it is not fundamentally different from a linear combination ("lincom") of estimates.

- As in chapter 4.

Prediction and residual analysis (continued)

θ_0 is the linear combination of all the estimators.

- But what is the standard error of a prediction, given a set of x?
- As with hypotheses about linear combinations of estimators, the regression model can be rewritten in term of the linear combination.

$$\beta_0 = \theta_0 - \beta_1 c_1 - \ldots - \beta_k c_k \text{ can be substituted into}$$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \text{ to get}$$

$$y = \theta_0 + \beta_1 (x_1 - c_1) + \ldots + \beta_k (x_k - c_k) + u.$$

Prediction and residual analysis (continued)

Regressing *y* on the differenced *x* variables will produce an estimate of θ_0 in the form of the intercept, along with its standard error!

This can be used along with a *t* value reflecting the desired confidence level to construct a confidence interval around the prediction of the form:

$$\hat{y} \pm se(\hat{\theta}_0) t_{\frac{\alpha}{2}}.$$

Predictions with Stata

The example (6.5) in the text can be performed using STATA (after performing the regression) using the syntax:

lincom _cons+sat*1200+hsperc*30+hsize*5+hsizesq*25.

The closer each of the values of c_j gets to its sample mean, the less variation there will be in the linear combination, which should be intuitive.

- The estimates should be the “best” and most accurate near the middle of the distribution.

Predicting *y* for individual observations

The procedure (on the last slide) predicts the conditional expectation of *y* and its standard error.

This is not the same thing as predicting the value of an individual observation of *y* in the sample.

- It's analogous to the difference between predicting a single random draw from a distribution and predicting the mean of several draws from the same distribution.
- The latter should be much more precise than the former.
- But, as with predicting a single value and predicting a mean, the expected value is the same.
- When predicting an individual value of *y*, we will use the notation, x_j^0, to distinguish it from predicting a linear combination (as before).

Fitting the model yields the prediction:

$$\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \ldots + \hat{\beta}_k x_k^0.$$

Prediction (standard) error

The standard error of an individual value, however, must account for the variation in the error term, as well.

This is derived from differencing $\hat{y}^0$ and the population value y^0, which includes the error term.

Deriving the standard error of the prediction is subtly different from the standard error of the conditional mean.

But,

$$se(\hat{y}^0) = [Var(\hat{y}) + \sigma^2]^{\frac{1}{2}}.$$

Prediction (standard) error (continued)

With the standard error (root of variance of the prediction error) in hand, you can proceed to construct confidence intervals for a prediction about a single observation of *y*.

The procedure does not differ from previous examples so we will summarize it here briefly.

The interval with confidence level $(1 - \alpha)$ and *n-k-1* degrees of freedom is:

$$\hat{y}^0 \pm se(\hat{y}^0)t_\alpha.$$

Prediction (standard) error (concluded)

In Stata the difference between the 2 predictions is equally subtle. For the standard error about the conditional mean (more precise, smaller error), the code after running the regression is:

predict {nameforse(yhat)}, stdp

For the standard error about an individual observation (less precise because of residual variance), the code after **regress** is:

predict {name for se(yhat0)}, sdtf

with the difference being the “f” rather than the “p”.

Residual analysis

Useful insights can be obtained from comparing regression-predicted values to observed values of *y*.

- Does the regression model predict a particular observation accurately, i.e., do we observe the value we *expect*, conditional on observations of the *x* variables?
- Is the observed y notably above or below the conditional expectation?
- If so the empiricist may ask what unobserved factor explains the residual.

This kind of inquiry is called residual analysis. It involves looking at one or more of the sample residuals,

$$\hat{u}_i = y_i - \hat{y}_i,$$

to learn about the validity of the model as a particular observation (or set of) is concerned.

Conclusion

OLS has much more genera applicability than stated in the introductory chapter. Being linear in parameters does not preclude:

- Non-constant effects of *x* on *y*,
- Non-monotonic, e.g., quadratic, effects of *x* on *y*,
- Effects of one *x* that depend on the value of another *x*, i.e., interaction terms.

Causality and a high R squared are not synonymous.

- It is possible to over control for *x* variables in a regression, and there is a degrees of freedom trade off (captured by adjusted R squared) for adding regressors.

The conditional expectation of y (on x) and the prediction of an individual observation of y have different standard errors, with the latter having a wider confidence interval.

Optional: data scaling derivation

If the original model is:

$$y = \beta_0 + \beta_1 x + u,$$
$$\Delta x = 1 \rightarrow \Delta y = \beta_1, \text{and } \Delta x = 1000 \rightarrow \Delta y = 1000\beta_1.$$

Re-scaling x by a constant, such as 1000, means defining a new x:

$$x' \equiv \frac{x}{1000};\ \Delta x' = 1 \Leftrightarrow \Delta x = 1000.$$

You can modify the regression model to estimate the effect of x':

$$y = \beta_0 + 1000\beta_1 x' + u, \text{such that } \Delta x' = 1 \rightarrow \Delta y = 1000\beta_1.$$

This is the same marginal effect estimated by the original regression.

- This is what you're doing when you re-scale variables. The new coefficient will be exactly 1000 times bigger than the old one, to account for the new (larger) units.

Data scaling derivation (continued)

A similar lesson applies to re-scaling y:

$$\text{when}, y = \beta_0 + \beta_1 x + u, \text{and}$$
$$y' \equiv \frac{y}{1000};\ \Delta y' = 1 \Leftrightarrow \Delta y = 1000.$$

You can estimate the effect on y':

$$y' = \frac{\beta_0}{1000} + \frac{\beta_1}{1000} x + \frac{u}{1000}, \text{such that } \Delta x = 1 \rightarrow \Delta y' = \frac{\beta_1}{1000}.$$

This is exactly $\frac{1}{1000}$ as large as the original coefficient, e.g., if y is in \$, and β_1 is 2250, y increases by \$2,250 or $\left(2.250 = \frac{\beta_1}{1000}\right)$ thousand dollars . . . which is the same amount.

- Back.

Optional: variance of fitted values

$$Var(\hat{y}) = E[(\hat{y} - E(\hat{y})]^2 = \left[(\hat{\beta}_0 - \beta_0) + \sum_{j=1}^{k}(\hat{\beta}_j - \beta_j)\, x_j\right]^2$$

$$= Var(\hat{\beta}_0) + \sum_{j=1}^{k} Var(\hat{\beta}_j)x_j^2 + 2\sum_{j=1}^{k} Cov(\hat{\beta}_0, \hat{\beta}_j)x_j + \sum_{m\neq j}^{k}\sum_{j=1}^{k} Cov(\hat{\beta}_j, \hat{\beta}_m)x_j x_m$$

So the standard error is obviously pretty tough to calculate by hand.

- That's why the clever method of parameterizing $\hat{y}$ (as the intercept like on Wooldridge 207-208) was devised.

Back.

Optional: variance of the prediction

For predicting the individual observation, the standard error is based on how far from the observed *y* the prediction is likely to be, not how far from the population model's expectation.

So we take the variance about y^0:

$$Var(\hat{y}^0) = E[(\hat{y}^0 - y^0]^2 = \left[(\hat{\beta}_0 - \beta_0) + \sum_{j=1}^{k}(\hat{\beta}_j - \beta_j)\, x_j - u\right]^2,$$

which adds another term *(u).*

$$Var(\hat{y}^0) = Var(\hat{\beta}_0) + \sum_{j=1}^{k} Var(\hat{\beta}_j)x_j^2 + 2\sum_{j=1}^{k} Cov(\hat{\beta}_0, \hat{\beta}_j)x_j + \sum_{m\neq j}^{k}\sum_{j=1}^{k} Cov(\hat{\beta}_j, \hat{\beta}_m)x_j x_m + Var(u)$$
$$= Var(\hat{y}) + Var(u) = Var(\hat{y}) + \sigma^2$$

Back.

Optional: predicting *y* when log(*y*) is the dependent variable

Fitting a model with a dependent variable in logarithmic form works the same way (fitting the model) as described previously.

(1) $\widehat{logy} \equiv \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$.

However transforming this prediction back into a level of *y* is not as simple as taking the anti-log, because

$$\ln\left(exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k\right)\right) \neq \widehat{logy}\ .$$

Optional: predicting *y* when log(*y*) is the dependent variable

This results from the inclusion of the error term when you take expectations of the population model:

$$(1)\text{estimates the model } logy = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + u$$

$$\Leftrightarrow y = \exp(\beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + u)\,.$$

Optional: predicting *y* when log(*y*) is the dependent variable

The conditional expectation of *y* on *x* is:

$$(2)\ E(y|x) = \exp(\beta_0 + \beta_1 x_1 + \ldots \beta_k x_k) * E[\exp(u)\,|x)$$

$$= \exp(\beta_0 + \beta_1 x_1 + \ldots \beta_k x_k) \exp\left(\frac{\sigma^2}{2}\right), \text{if } u \sim Normal\ (0, \sigma^2).$$

It can be shown that, $\exp\left(\frac{\sigma^2}{2}\right) = E[\exp(u)] = E[\exp(u)\,|x).$

Optional: predicting *y* when log(*y*) is the dependent variable

(2) is the consistent estimator of $\hat{y}$ when the errors are normally distributed and the unbiased estimator ($\hat{\sigma}^2$) of σ^2 is used to calculate the prediction.

If the errors are not normally distributed, which is a desirable assumption to relax, (2) can be generalized by replacing the specific form under normality $\left(\exp\left(\frac{\sigma^2}{2}\right)\right)$ with

$$\alpha_0 = E[\exp(u)], \text{such that}, E(y|x) = \alpha_0 \exp(\beta_0 + \beta_1 x_1 + \ldots \beta_k x_k)\,.$$

Optional: predicting *y* when log(*y*) is the dependent variable

The last remaining problem is estimating α_0.

This can be done consistently (but not without bias) by replacing the expectation with its sample analog,

$$\hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{u}_i)\,.$$

Then the level of *y* can be predicted from a regression with *y* in logs using the following.

$$(3)\ \hat{y} = \hat{\alpha}_0 \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_k x_k\right).$$

Optional: predicting *y* when log(*y*) is the dependent variable

In addition to merely predicting levels from a logarithmic regression, the fitted values in levels can be used to compare the log version to the levels regression.

The empiricist can obtain the R^2 from the levels regression,

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

and compare it to the R^2 calculated using fitted values calculated as in (3).

If the latter produces a higher R^2 than the regression in levels, the logarithmic specification is preferred by comparison to the regression in levels.