Section 6 Functional Form and Nonlinearities

This is a good place to remind ourselves of Assumption #0: That all observations follow the same model.

Nonlinearity in variables vs. nonlinearity in parameters

- Solving for the OLS estimator required that we differentiate the LS or likelihood function with respect to the parameters.
- In a model that is linear in parameters, the LS objective function will be quadratic, so that the least-squares normal equations based on setting the first derivatives to zero are linear in the coefficient estimator.
  - This means that we can use linear algebra to solve for the coefficient estimator.
- If the model is nonlinear in parameters, then the LS objective function will not be quadratic and the normal equations will not be linear in parameters, so numerical search methods must be used for solution.
  - This is called **nonlinear LS** and is much more computationally difficult and potentially problematic than the linear model. (Covered in S&W appendix to Ch. 8.)
  - There are times when nonlinear LS is necessary, but we try to avoid it whenever possible.
- There are many models that are nonlinear in variables but linear in parameters. These models are easy to deal with: we can transform the variables and use linear OLS methods.
- If a model is nonlinear in its regressors (or with a nonlinear dependent variable), then the coefficient on the variable is no longer $\partial Y / \partial X_j$.
  - Instead, we have to calculate $\partial Y / \partial X_j$ as a function of the coefficients and the values of X.
  - This will vary according to the functional form, so we'll talk about the partial effects for individual forms as we discuss them.
- The choice of functional form should be guided by theory, but theory rarely provides a unique specification.
  - It is often necessary to try various functional forms to see which one seems to fit the best.
  - Plotting actual and fitted values against each regressor can often be helpful in seeing nonlinearities. (S&W's Figures 8.2, 8.3)
- One way to explore nonlinearities (if you have a large enough sample) is to create a battery of dummy variables with different levels of a regressor. Looking at the pattern of coefficients for the different levels can tell you whether the relationship is approximately linear.

- For example, we could examine math SAT score effects by looking at dummies for 500 ≤ SATM < 600, 600 ≤ SATM < 700, and SATM ≥ 701, leaving out the bottom category below 500.
- This will give us four points on a general response function (with zero implicit for the omitted group, below 500).
- If the four points seem to lie on a straight line, then the linear specification is probably fine. One may also see evidence of quadratic or cubic behavior and can use more than four categories if you have enough data and want to be more discriminating.

Quadratic and higher-order polynomial models

- One easy way of incorporating curvature into a model is to introduce quadratic terms. (For the moment, we will assume only one regressor is nonlinear, so we'll ignore others.)
    - $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$
    - Possible shapes for the relationship:
        - Upward sloping at an increasing rate ($\beta_1 > 0$, $\beta_2 > 0$)
        - Upward sloping at a decreasing rate or downward sloping but flattening out ($\beta_1 > 0$, $\beta_2 < 0$)
            - Note that this curve *always* turns downward (upward) after a peak (trough) at $X = -\beta_1/2\beta_2$, so it is critical to evaluate which part(s) of the curve the sample lies in. (Are most/all of the X values of interest < or > $-\beta_1/2\beta_2$?)
            - This non-monotonicity may be good or bad depending on theory.
            - If you want a universally monotonic but diminishing effect, using ln X may be a good alternative specification.
        - Downward sloping and getting steeper as X increases ($\beta_1 < 0$, $\beta_2 < 0$)
    - Always include a graph of the response function so that your reader can understand the shape of the effect.
        - The coefficients don't tell the story in a transparent way.
    - Partial effect
        - $\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$.
        - The sign of the partial effect will change at $X = -\beta_1/2\beta_2$ if sgn(β_1) ≠ sgn(β_2), as discussed above.
    - Estimating the standard error of the partial effect
        - Conditional on X,

          $\text{var}\left(\hat{\beta}_1 + 2\hat{\beta}_2 X\right) = \text{var}\left(\hat{\beta}_1\right) + 4X^2 \text{var}\left(\hat{\beta}_2\right) + 4X \text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right).$

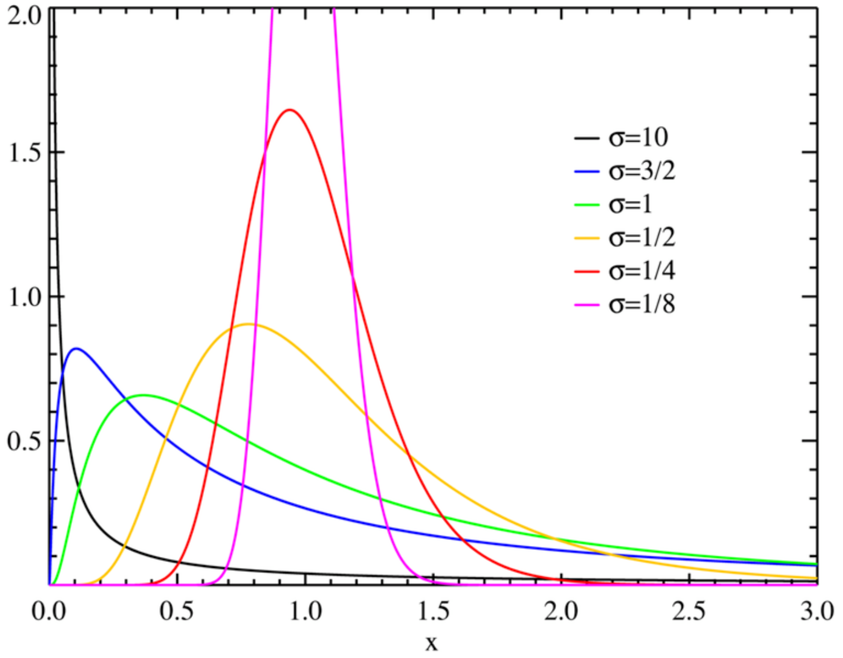
          The estimated values of the variances and covariances can be obtained from the output of your regression package. (They are the diagonal and off-diagonal elements of the estimated covariance matrix of the coefficient vector. This is obtained by estat vce after a regression command. (As usual, it will be the classical estimated covariance estimator unless you use the robust option in the regression.)

        - S&W point out two other ways of estimating the standard error of a linear combination of coefficients:
            - Do a test command that the partial effect is zero to get an *F* statistic, then an estimate of the standard error will be the absolute value of the partial effect at that *X* divided by the square root of the *F* value.
            - Transform the model into one where the desired effect is directly estimated and get the standard error from the regression table.
    - Relevant significance tests in the quadratic model:
        - Does *X* affect *Y*?
            - This is a test of the joint null hypothesis $H_0 : \beta_1 = 0, \beta_2 = 0.$ It is a standard *F* test.
        - Is the relationship quadratic rather than linear?
            - This is a *t* test of H_0: $\beta_2 = 0$, given that β_1 is assumed to be nonzero (null hypothesis is linear model).
            - This is an example of a **nested specification test** because the linear model is a special case of (nested within) the quadratic specification.
            - Note that the *t* test is preferred to comparing R^2 or $\bar{R}^2$ values.
                - The former will always be higher for the quadratic specification.
                - The latter will be higher if the *t* value exceeds one, which is well below conventional critical values.
- Higher-order polynomials
    - Do cubic, quartic, etc. relationships ever occur in economic data?
        - Yes, but they can be hard to sell.
        - Example of SAT scores and Reed GPA.
    - Same procedures apply for estimated partial effects and tests.
        - What to do if 3rd-order term is significant and 2nd-order term is not?
            - Don’t leave out the 2nd-order term.
            - Test both jointly to try to reject the linear model in favor of the cubic. If significant, retain both.

Log-based models

- Many econometric models are specified in log term.
  - Most economic variables are non-negative, so we don't need to worry about negative values. (Though many can be zero.)
  - $d(\ln x) = dx / x$ = the percentage change in *x*, so the interpretation of coefficients and effects is useful and easy.
  - The log-log model is a constant-elasticity specification with the coefficient being read directly as an elasticity.
  - Shape of log functions is often reasonable:
    - Shies away from axes
    - Monotonic with diminishing returns
- **Log of regressor only** ("linear-log" model)
  - $Y_i = \beta_0 + \beta_1 \ln X_i + u_i$.
  - Change of 1% in *X* changes ln *X* by about 0.01 and thus leads to about a $0.01\beta_1$ unit absolute change in *Y*.
    - If *X* increases by *x*%, this means it is 1 + *x*/100 times as large, which means that its log is $\ln X_0$ + ln(1 + *x*/100). If *x* is small (say, less that 20%) then the approximation is reasonable close. However, you may want to do exact calculations for formal work.
  - Partial effect in levels is $\frac{\partial Y}{\partial X} = \beta_1 \frac{1}{X}$, which is monotonically increasing or decreasing (depending on sign of β_1) but slope goes to zero as *X* gets large.
- **Log of dependent variable only** ("log-linear" model)
  - $\ln Y_i = \beta_0 + \beta_1 X_i + u_i$
    - Note that $Y_i = e^{\beta_0 + \beta_1 X_i + u_i}$, so this is clearly a different error term than when *Y* is not in log terms.
  - Change of *x* units in *X* changes ln *Y* by $\beta_1 x$ units, so it changes *Y* by about $100\beta_1 x$ percent.
    - The same approximation issues applies here. The increase of $\beta_1 x$ units in ln *Y* means that *Y* increases by a factor of $e^{\beta_1 x}$, which is approximately 1 + $\beta_1 x$ for small values of *x*. For larger values of *x* and for more formal work, it is best to calculate the exponential directly.
  - Partial effect in levels is $\frac{\partial Y}{\partial X} = \frac{\partial \ln Y}{\partial X} \frac{\partial Y}{\partial \ln Y} = \beta_2 / \frac{\partial \ln Y}{\partial Y} = \beta_2 Y$.
    - Alternatively, $\frac{\partial Y}{\partial X} = \beta_1 e^{\beta_0 + \beta_1 X + u} = \beta_1 Y$.
    - Partial effect is increasing in absolute value as *Y* increases. (Note that *Y* must always be positive in this model.)

- **Log of both regressor and dependent variable** (“log-log model”)
    - $\ln Y_i = \beta_0 + \beta_1 \ln X_i + u_i$.
        - Also implies that $Y_i = e^{\beta_0 + \beta_1 \ln X_i + u_i} = e^{\beta_0} X_i^{\beta_1} e^{u_i} = \alpha_0 X_i^{\beta_1} v_i$, where $\alpha_0 \equiv e^{\beta_0}, \quad v_i \equiv e^{u_i}$.
        - The Cobb-Douglas function takes this form (with a multiplicative error *v*, usually assumed to be log-normally distributed).
    - Change of 1% in *X* changes ln *X* by about 0.01, which changes ln *Y* by about 0.01β$_1$, which changes *Y* by about β$_1$%. (Both of the approximation caveats above apply here.)
        - Thus, **β$_1$ is the point elasticity of *Y* with respect to *X*.**
        - This makes log-log a popular function form.
    - Partial effect in levels is $\frac{\partial Y}{\partial X} = \frac{\partial Y}{\partial \ln Y} \frac{\partial \ln Y}{\partial \ln X} \frac{\partial \ln X}{\partial X} = \beta_1 \frac{Y}{X}$.
        - Alternatively, $\frac{\partial Y}{\partial X} = e^{\beta_0 + \beta_1 \ln X + u} \beta_1 \frac{\partial \ln X}{\partial X} = \beta_1 \frac{Y}{X}$.
        - Partial effect is constant in elasticity terms, but varies with *Y* and *X* in level terms.
- Which log model to choose?
    - Theory may suggest that percentage changes are more important than absolute changes for one or both variables.
        - Income is often logged if we think that a doubling of income from $50,000 to $100,000 would be associated with the same change in other variables as a doubling from $100,000 to $200,000 (rather than half as much).
        - As suggested by the previous example, logging a variable scales down extreme values. If most of the sample variation is between $20,000 and $100,000 (with mean $50,000 and standard deviation $30,000), but you have a few values of $500,000 for income, these are going to be 15 standard deviations above the mean in level terms but much less in log terms.
            - The log of 500,000 is only ln(10)=2.3 units larger than the log of 50,000. The standard deviation of the log would probably be in the range of 0.6 or so, so the highly deviant observations would be less than 4 standard deviations above the mean instead of 15.

    - Since we often want our variables to be normally distributed, we might try to decide whether the variable is more likely to be normally or log-normally distributed.
    - 

    - Note that the various log models are **not nested** with one another or with the linear or polynomial models, so *t* tests cannot discriminate between them.
    - We can use R^2 to compare models if only if the dependent variable is the same:
        - Linear model with linear-log model
        - Log-linear model with log-log model
- Box-Cox model nests log and linear terms for both dependent and independent variables in a nonlinear model.
    - Can estimate Box-Cox model and test hypothesis that the relationship is linear or log.
    - Box-Cox transformation is $B(X,\lambda)=\begin{cases}\dfrac{X^{\lambda}-1}{\lambda}, \text{if } \lambda \neq 0,\\ \ln X, \text{if } \lambda\text{=0.}\end{cases}$
        - This is a continuous function that equals $X-1$ if $\lambda = 1$ and ln*X* if $\lambda = 0$.
    - We can do a nonlinear regression of $B(Y, \lambda_Y)$ on $B(X, \lambda_X)$ and test the two λ values to see whether they are zero or one to determine whether a linear or log specification is preferred for both variables.

- Prediction of *Y* in ln *Y* models
  - If we estimate a model in which the dependent variable is in logs, prediction is a problem
  - We can predict ln *Y* by $\widehat{\ln Y} = \hat{\beta}_0 + \hat{\beta}_1 X + E(u) = \hat{\beta}_0 + \hat{\beta}_1 X$. But $E(Y) = E\left(e^{\ln Y}\right) > e^{E \ln Y} = e^{\widehat{\ln Y}}$.
  - The problem is that even if $E(u) = 0$, $E(e^u) \neq 1$.
  - If *u* is normally distributed with variance σ^2, then $E\left(e^u\right) = e^{\frac{\sigma^2}{2}}$.
    - In that case, we can predict *Y* by $\hat{Y} = e^{\frac{s_{\hat{u}}^2}{2}} e^{\widehat{\ln Y}}$. This is a consistent prediction if the error term is normal.
  - In the non-normal case, we can use a simple regression to calculate the appropriate adjustment factor:
    - Run a regression of $Y_i = \gamma e^{\widehat{\ln Y_i}}$, which is a bivariate regression without a constant term.
    - Then adjust the predictions to get $\hat{Y} = \hat{\gamma} e^{\widehat{\ln Y}}$, which, for the sample observations, are just the predicted values from the auxiliary regression.

Interaction effects

Sometimes the effect of one variable depends on the level of another. This is particularly common with dummy variables, where we might expect a regressor to have a different effect for males and females, for example.

This is an example of dealing with violations of Assumption #0. It allows subsets of the sample to have different coefficients.

- **Interactions between two dummy variables**
  - Suppose we have two sets of qualitative characteristics, sex and ethnicity. We expect different values of *Y* for males vs. females and different values for white vs. nonwhite.
    - We can model this as $Y_i = \beta_0 + \beta_1 male_i + \beta_2 white_i + \ldots + u_i$. This allows whites to have a different intercept than nonwhites and males to have a different intercept than females.
    - However, this model insists that white males have the same differential with respect to white females that nonwhite males have with respect to nonwhite females.
  - If we want to allow the sex differential to differ by ethnicity, we can include an interaction term *male*white*: $Y_i = \beta_0 + \beta_1 male_i + \beta_2 white_i + \beta_3 male_i \times white_i + \ldots + u_i$

        - In this model the intercept term is
            - β_0 for nonwhite females (all dummies are zero)
            - $\beta_0 + \beta_1$ for nonwhite males (only *male* is one)
            - $\beta_0 + \beta_2$ for white females (only *white* is one)
            - $\beta_0 + \beta_1 + \beta_2 + \beta_3$ for white males (all three dummies are one)
        - Thus,
            - β_1 measures the effect of being male for nonwhites
            - β_2 measures the effect of being white for females
            - $\beta_2 + \beta_3$ measures the effect of being white for males
            - $\beta_1 + \beta_3$ measures the effect of being male for whites
        - We can test whether any of the differences in these pairs are statistically significant by testing the null hypothesis that the corresponding effect is zero.

- **Interactions between a dummy and a continuous variable**
    - We can interact a dummy with a continuous variable to allow the effect of the continuous variable to differ depending on whether the dummy is one or zero.
        - Note that having the dummy itself in the equation allows the *intercept* to differ according to the dummy, but not the slopes.
        - Any time we have a dummy interacted with a continuous variable, we generally have a dummy for the intercept as well.
    - Consider the model $Y_i = \beta_0 + \beta_1 male_i + \beta_2 X_i + \beta_3 male_i \times X_i + \ldots + u_i$.
        - For female observations, *male* = 0 and the model is $Y_i = \beta_0 + \beta_2 X_i + \ldots + u_i$, so the constant term is β_0 and the effect of *X* on *Y* is β_2.
        - For male observations, *male* = 1 and the model is $Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i + \ldots + u_i$, so the intercept is $\beta_0 + \beta_1$ and the effect of *X* on *Y* is $\beta_2 + \beta_3$.
        - We can test whether the intercept differs for males and females by testing the null hypothesis $\beta_1 = 0$ (which is the reported *t* statistic on *male*).
        - We can test whether the effect of *X* on *Y* differs for males and females by testing the null hypothesis $\beta_3 = 0$ (which is the reported *t* statistic on the interaction variable).
        - Note that the total effect of being male in this model is $\beta_1 + \beta_3 X_i$, which varies with *X*.

- **Interactions between two continuous variables**
    - Finally, we might expect that the effect of one variable might be larger or smaller depending on the value of another continuous variable. For example, we might expect the consumption expenditures of households with large amounts of wealth to be less sensitive to income than those with lower wealth.

- If *Y* is consumption, *X* is income, and *W* is wealth, then we could model this as $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \ldots + u_i$.
  - The partial effect of *X* on *Y* is $\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 W$.
  - If wealthy households have smaller effects of income on consumption, then $\beta_3 < 0$, which is a testable hypothesis.
- Note that it rarely makes sense to have *XW* in the equation unless both *X* and *W* are in the equation alone.

- **Interactions and the Chow test**
  - Sometimes we want to know if two sub-samples have the same coefficients. This is traditionally known as the Chow test and, under classical assumptions, is easily done as an *F* test.
    - The restricted model forces the coefficients to be the same in both sub-samples, so in the *F* test formula SSR_r is the sum of squared residuals from the single regression on the full sample.
    - The unrestricted model allows all coefficients to vary across sub-samples, which is easily done by running separate regressions on the sub-samples and adding together the sums of squared residuals to get SSR_u. The number of restrictions is $k + 1$, because there are $2k + 2$ coefficients in the unrestricted model and $k + 1$ in the restricted.
    - The standard *F* formula can then be used: $F = \frac{\left(SSR_r - SSR_u\right)/\left(k+1\right)}{SSR_u/\left(n-2k-2\right)}$,

      where *q* has been replaced by the number of restrictions $k + 1$ and the denominator of the denominator is the number of degrees of freedom in the unrestricted model, the total number of observations minus the total number of coefficients estimated for the two sub-samples.
  - The Chow test above relies on classical assumptions, including homoskedasticity (even across sub-samples). We can do this test another way with interaction terms between a dummy and a continuous variable.
    - We simply include a dummy for the second sub-sample (to allow the constant to be different) and interact that dummy with *every* regressor.
    - We then test the joint null hypothesis that the coefficients on the dummy and all interaction terms are zero.
    - This test can use the robust covariance matrix and the *F* statistic formula $F = \frac{1}{q}\left(R\hat{\beta} - r\right)'\left(R\hat{\Sigma}_{\hat{\beta}}R'\right)^{-1}\left(R\hat{\beta} - r\right)$ that does not require homoskedasticity.

Other specification issues

- Data scaling
  - Suppose that Y is the interest rate. Does it matter whether we measure an interest rate of 5 percent as 5 or 0.05?
    - No. All that will happen is that the $\hat{\beta}$ vector (and its standard error) will be 100 times as large when we use 5 as when we use 0.05.
  - Suppose that X is the interest rate.
    - Doesn't matter here either. The single coefficient on the interest rate will be 100 times as large when we use 0.05 as when we use 5. Its standard error will also be 100 times as large. The other coefficients will be unaffected.
  - In log models, any change in scale of the level of the variable that is logged adds or subtracts a constant to/from the log of the variable. This affects the estimated intercept term, but not any of the slopes.
- Standardized (beta) coefficients
  - Sometimes it can be informative to ask the question: "If X increases by one standard deviation, how many standard deviations does Y change?"
    - This can be valuable in assessing how much practical importance X has on Y in the context of the overall variation in both variables.
    - These measures are (confusingly) usually called "beta" coefficients, and are calculated as $\widehat{beta}_j = \hat{\beta}_j \frac{s_{X_j}}{s_Y}$, where s_{X_j} is the sample standard deviation of X_j and s_Y is the sample standard deviation of Y.
- Testing for specification errors
  - Ramsey has proposed the "regression specification error test" or RESET to determine if the equation has the correct functional form
    - Adding all possible squares and cross-product terms and testing them eats up a lot of degrees of freedom, not to mention the possibility of cubic terms.
    - If there are missing higher-order terms, then they may be related to the powers of $\hat{Y}$.
    - The RESET test runs the regression $Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \gamma_1 \hat{Y}_i^2 + \gamma_2 \hat{Y}_i^3 + v_i$, using the squared (and perhaps cubed) predicted values from the regular regression.
    - We can then test whether the γ coefficients are zero as a test of whether the specification needs to be augmented.

Nonlinear least squares

- For models that are nonlinear in the parameters, we must generally use nonlinear search methods to find the least-squares (or maximum-likelihood) estimates.
    - Linearity in parameters depends crucially on the specification of the error term.
    - The error term in the model *must* be additive.
        - Consider the model $Y = e^{\beta_0} X^{\beta_1}$.
            - If the appropriate error term specification is $Y = e^{\beta_0} X^{\beta_1} e^{u}$, then we can take logs and get $\ln Y = \beta_0 + \beta_1 \ln X + u,$ which is linear in the parameters and can easily be estimated by linear OLS.
            - If the appropriate error term specification is $Y = e^{\beta_0} X^{\beta_1} + u$, then the model cannot be make linear in parameters with an additive error term and must be estimated nonlinearly. (I don't know why this error specification would be better, but just suppose…)
- Nonlinear estimation usually requires you to (at minimum) provide a formula for the deterministic part of the function.
    - To estimate the above model in Stata you could type
      ```
      nl (y = exp({b0}) * x^{b1}), initial (b0 5 b1 0)
      ```
    - Nonlinear search algorithms can be very slow and unreliable. It is generally very helpful to provide starting values near the optimal parameter values.
        - In this case, we might run the log-log regression (using the wrong error term specification) to get preliminary estimates of the coefficients, then insert those values in the "initial" option of the nl statement.
- Nonlinear estimation is a directed search over the parameter space to find the best combination. It is generally guided by taking numerical derivatives of the objective (LS or likelihood) function with respect to the parameters, then following the direction of greatest improvement (the gradient).
    - Some nonlinear-optimization packages allow you to enter analytic (algebraic) partial derivatives of the model with respect to the parameters. This generally speeds up convergence.
- Some objective functions may have multiple local optima. Starting far from the global optimum can cause the algorithm to become trapped at a global optimum that is inferior to the global one. Good initial values can help avoid this problem.
    - To assure that your optimum is a global one, try starting from several different sets of initial values and see if you converge to the same optimum.
- Some objective functions are badly behaved, having ridges (or valleys) where the objective function is very flat in one direction. This is particularly true if multicollinearity is a problem. If two variables are highly, positively correlated, then increasing the coefficient of one by a lot and simultaneously decreasing the coefficient of the other will have very little effect on the predicted values and the residuals, hence on the objective

function. This leads to a ridge in the likelihood function (valley in the least-squares function) at a diagonal in the space of these two variables.

- Nonlinear estimation is not as computationally problematic as in the old days, but it is still subject to these numerical difficulties.
    - Avoid it when possible by using specifications that are linear in the parameters.
    - There will be times when we need to use it for maximum-likelihood estimators such as probit and logit, but these likelihood functions are often well-behaved.