Correct specification of a regression model requires that we determine the regressors, the functional form, and the stochastic error. Today we focus on specification of the functional form. Given the regressors, we are led to ask: How is $Y_t$ related to the regressors? Do we expect a graph of the two to be linear or curved? Does the impact of a regressor peak at one value and then decline? As in selecting regressors, we look to economic theory to guide our choice of functional form.

The baseline regression model is linear in both the coefficients and the regressors

$$Y_t = \beta_1 + \beta_2 X_{t,2} + \cdots + \beta_K X_{t,K} + U_t.$$  

Each coefficient is given by

$$\beta_k = \frac{\partial Y_t}{\partial X_{t,k}},$$

that is, the change in $Y_t$ brought about by a small change in $X_{t,k}$ holding all other regressors constant. If $\beta_k = 4$, then an increase of 1 unit in $X_{t,k}$ leads to a 4 unit increase in $Y_t$. For such a model, the partial derivative (or slope) is constant. A scale invariant measure is obtained directly from the slope coefficient by standardizing all variables, that is

$$\frac{Y_t - \bar{Y}}{S_Y} = \beta_1 + \beta_2 \frac{X_{t,2} - \bar{X}_2}{S_{X_2}} + \cdots + \beta_K \frac{X_{t,K} - \bar{X}_K}{S_{X_K}} + U_t.$$

**Logarithmic Forms**

Often economic theory tells us that the elasticity of $Y_t$ with respect to $X_{t,k}$ is constant. Because this elasticity is

$$\frac{\partial Y_t}{\partial X_{t,k}} = \beta_k \frac{X_{t,k}}{Y_t},$$

the elasticity depends on the values of the variables in the baseline model. To obtain a constant elasticity, we work with the logarithms of the variables

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{t,2} + \cdots + \beta_K \ln X_{t,K} + U_t.$$
Each coefficient is now given by

$$\beta_k = \frac{\frac{\partial Y_t}{\partial X_{t,k}}}{\frac{\partial Y_t}{X_{t,k}}}$$

that is, the elasticity of $Y_t$ with respect to $X_{t,k}$ holding all other regressors constant. If $\beta_k = 4$, then a 1 percent increase in $X_{t,k}$ leads to a 4 percent increase in $Y_t$. Because the elasticity is constant, the slope varies with the level of the variables.

(Include Graph from Page 2 of Attached Notes)

Because the logarithm is not defined for 0 or negative values, the logarithm is not appropriate for many regressors. A natural alternative is to use the logarithm for only some of the variables in the regression model,

$$Y_t = \beta_1 + \beta_2 \ln X_{t,2} + \beta_3 X_{t,3} + U_t.$$  

The coefficients $\beta_1$ and $\beta_3$ are partial derivatives, as in the baseline model. The coefficient $\beta_2$ is

$$\beta_2 = \frac{\partial Y_t}{\partial X_{t,2}},$$

that is, the change in $Y_t$ brought about by a percentage change in $X_{t,2}$ holding all other regressors constant. If $\beta_2 = 4$, then a 1 percent increase in $X_{t,2}$ leads to a .04 unit increase in $Y_t$ (to see this, $\partial Y_t = \beta_4 \frac{\partial X_{t,2}}{X_{t,2}} = 4 \cdot .01 = .04$). The dependent variable changes in proportion to percentage changes in the regressor (as in the chicken consumption example from the last lecture). Neither the slope nor the elasticity is constant.

(Include Graph from Page 3 of Attached Notes)

Alternatively

$$\ln Y_t = \beta_1 + \beta_2 X_{t,2} + \beta_3 X_{t,3} + U_t.$$  

Each coefficient is given by

$$\beta_k = \frac{\frac{\partial Y_t}{\partial X_{t,k}}}{\frac{\partial Y_t}{X_{t,k}}}.$$
that is, the percentage change in $Y_t$ brought about by a small change in $X_{t,k}$ holding all other regressors constant. If $\beta_2 = 4$, then a .1 unit increase in $X_{t,2}$ leads to a .4 percent increase in $Y_t$ (to see this, $\frac{\partial Y_t}{Y_t} = \beta_4 \partial X_{t,k} = 4 \cdot .1 = .4$). (We use a .1 unit increase here because the function is nonlinear and the linear approximation is poor for a change as large as 1 unit.)

**Polynomial Forms**

The first extension of a regression model with linear regressors is a model with a quadratic regressor

$$Y_t = \beta_1 + \beta_2 X_{t,2} + \beta_3 X_{t,3} + \beta_4 X_{t,3}^2 + U_t.$$ 

The coefficients $\beta_1$ and $\beta_2$ are partial derivatives, as in the baseline model. The coefficients $\beta_3$ and $\beta_4$ are defined together as

$$\frac{\partial Y_t}{\partial X_{t,3}} = \beta_3 + 2\beta_4 X_{t,3},$$

which depends on the level of $X_{t,3}$.

**Include Cost and Earnings Examples from Page 4 of Attached Notes**

One can continue with higher-order terms. Yet a cubic rarely fits economic data. Further, $n$ sample points between $Y_t$ and $X_t$ can be fit exactly by a polynomial of order $n - 1$ in $X_t$, so theory must guide us.

An alternative extension is to use an inverse regressor

$$Y_t = \beta_1 + \beta_2 X_{t,2} + \beta_3 \frac{1}{X_{t,3}} + U_t.$$ 

It must be the case that $X_{t,3}$ never equals zero. The coefficients $\beta_1$ and $\beta_2$ are partial derivatives, as in the baseline model. The coefficient $\beta_3$ is determined from

$$\frac{\partial Y_t}{\partial X_{t,3}} = -\beta_3 \frac{1}{X_{t,3}^2},$$

which depends on the level of $X_{t,3}$.

**Include Graphs from Page 5 of Attached Notes**

**Spline Forms**

3
If the impact of a regressor changes direction or sign at a specific value, a spline may be more appropriate than a polynomial. (Splines are piecewise linear functions that are close relatives to polynomial functions.) Consider a model with only one varying regressor, $X_{t,2}$. Suppose that at the value $X_{2}^{*}$, the sign of the relation changes.

\[(Include \ Graph \ from \ Page \ 9 \ of \ Attached \ Notes)\]

A spline regression model captures the change

$$Y_t = \beta_1 + \beta_2 X_{t,2} + \beta_3 D_t + \beta_4 D_t X_{t,2} + U_t,$$

where

$$D_t = \begin{cases} 
1 & \text{if } X_{t,2} > X_{2}^{*} \\
0 & \text{if } X_{t,2} \leq X_{2}^{*} 
\end{cases}.$$

The coefficients $\beta_1$ and $\beta_2$ are the slope coefficients from the baseline model for all values of the regressor below $X_{2}^{*}$. For values above $X_{2}^{*}$, the coefficients are jointly defined as

$$\frac{\partial Y_t}{\partial X_{t,2}} = \beta_2 + \beta_4.$$

It is important that we add both additional terms, to ensure that there is no jump in the regression model. Failure to include the term $\beta_3 D_t$ forces both lines to have the same intercept and results in the jump.

**Specification Testing**

Specification testing is fraught with danger as a statistical test cannot reveal the correct functional form. That said, a test may help select between two competing approximate forms. Use of the $R$-square statistic is not wise. First, the statistic only captures correlation. Second, if the dependent variable is measured in levels in one regression and logarithms in another, direct comparison is not valid as the total sum of squares (and so the $R$-square statistic) is not scale invariant.

For some cases, a simple coefficient estimator test statistic is appropriate. For example, the spline form presented could be tested by examining the joint significance of $\beta_3$ and $\beta_4$. 

4